LEARNING WHAT TO SHARE: LATENT REPRESENTATIONS FOR FEDRL WITH OBSERVATION DISTRACTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated reinforcement learning lets agents in different environments learn a shared policy by averaging updates without exchanging trajectories. This setting becomes harder when each observation contains task-irrelevant signals: the policy must learn from returns which coordinates matter, and each client learns that lesson from a different noise realisation. We study whether the federation should share a learned representation rather than raw-observation policies. We formalise the setting as a federated exogenous block MDP, then train federated PPO with an encoder whose gradient is routed through the value loss, with or without a self-supervised objective. The objectives include a JEPA-style multi-step predictor, a behavioural metric, and a multi-step inverse model. A linear–Gaussian analysis shows why a noise-free representation is common across clients and why policy-gradient learning on raw observations pays a fixed-step-size penalty that grows with the number of distractors. On six MuJoCo tasks with clients that differ in physical parameters, the critic-trained encoder outperforms FedAvg, FedProx, and FedKL baselines under 64 appended Gaussian distractors. The gain is largely due to gradient routing rather than model width. Self-supervised objectives help on some tasks and hurt on others. For the predictive agent, the gain grows with the number of i.i.d. distractors, but its size changes with temporal correlation: it is reduced on several tasks at $\rho = 0.99$ and remains positive on most. Federation improves the critic-trained encoder on all four tasks with a single-agent comparison, although it is not the source of the basic advantage.

1 INTRODUCTION

Federated reinforcement learning (FedRL) allows agents in separate environments to learn a shared policy without pooling their trajectories (Jin et al., 2022). Each client trains locally and periodically contributes a parameter update, which the server combines using methods such as FedAvg (McMahan et al., 2017). Recent work has examined how this aggregation should account for differences between clients. FedProx constrains local models to remain close to the global model (Li et al., 2020), FedKL limits policy divergence (Xie & Song, 2023), and other methods use momentum or personalised heads (Wang et al., 2024; Collins et al., 2021; Xiong et al., 2025).

We investigate a complementary source of difficulty: clients may observe different nuisance signals even when they are solving the same task. Sensor readings often mix task-relevant state with measurement noise, background variation, or signals from unrelated subsystems. A policy that receives the full vector must infer from returns which coordinates matter for control. Distractors already slow single-agent deep RL (Nikishin et al., 2022; Grooten et al., 2023). In a federation, each client encounters a different distractor stream, so the resulting policy updates are noisy in ways that cannot benefit the other clients.

Latent-state architectures offer a natural way to separate these signals. We introduce an encoder that transforms the observation into a compact latent state before it reaches the actor and critic. The representation is learned jointly across clients and is averaged with the policy. This design shifts the central question from whether to use an encoder to how that encoder should learn which information to preserve. We use *routing* to describe which training losses are allowed to update the shared encoder.

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.

Under *critic routing*, the value loss shapes the latent state while a stop-gradient prevents the policy loss from changing it. A *jointly trained encoder* instead receives gradients from both losses. Three additional agents retain critic routing and add a representation objective: multi-step self-prediction in the style of SPR and JEPA (Schwarzer et al., 2021; LeCun, 2022; Assran et al., 2023), the SimSR behavioural metric (Zang et al., 2022), or the AC-State multi-step inverse objective (Lamb et al., 2023).

Our experiments separate aggregation, representation, and sharing. We first compare the latent agents with raw-observation PPO under FedAvg, FedProx, and FedKL. We then vary the encoder's routing and capacity, remove the distractors, change the auxiliary optimiser, and compare federated training with independent agents. The study uses six MuJoCo tasks with four physics contexts per task. The main comparison appends 64 i.i.d. Gaussian distractors. A separate predictive-agent sweep varies the number of distractors and their AR(1) correlation $\rho \in \{0.5, 0.95, 0.99\}$.

Changing the aggregation rule has a modest effect in the main comparison. FedProx changes FedAvg-PPO's return by -25% to $+13\%$, while FedKL remains within 4%. Critic routing produces a much larger change. The critic-trained encoder reaches 1.4–7.8 times the return of the best aggregation baseline on every task. The same pattern holds when the encoder is reduced to one 64-unit hidden layer, giving the complete agent approximately 1.4 times as many parameters as PPO. By contrast, allowing the policy gradient to train the encoder removes most of the gain.

The advantage is tied to the distractor setting. Separate actor and critic networks are the standard recommendation for on-policy control (Andrychowicz et al., 2021), and our clean-observation control reproduces that result: the critic-trained encoder is worse than FedAvg-PPO on four tasks and statistically indistinguishable on two (Table 12). Client-specific distractors reverse the comparison on all six tasks. This interaction, rather than the familiar architectural choice alone, is our central empirical finding. Auxiliary objectives then make task-dependent changes, and federation further raises the encoder's return on all four tasks with independent-agent controls, significantly on two.

We complement these experiments with a linear–Gaussian analysis that isolates the effect of exogenous coordinates. In this model, clients with different physics share the same noise-free encoder as a fixed point, so averaging does not introduce heterogeneity into that part of the representation. Raw-observation policy gradients also incur an extra fixed-step-size error that grows linearly with the number of distractor dimensions. The analysis gives a tractable account of the local mechanism tested by the deep-network experiments and identifies temporal correlation as a factor that can weaken denoising.

Contributions. (i) We formulate distractors in FedRL as a federated exogenous block MDP and evaluate every final client policy on every client context. (ii) We show that the signal used to train a shared encoder matters more than the aggregation rule in our setting, using controls for routing, capacity, auxiliary objectives, and shared modules. (iii) We analyse why a noise-free representation can be shared across heterogeneous clients, how distractor dimension affects raw policy gradients, and how temporal correlation changes representation learning.

2 Related work

Federated RL under heterogeneity. FedRL studies how local updates should be combined when clients face different environments. Existing approaches analyse policy or value averaging, robustness to faulty clients, and momentum (Jin et al., 2022; Fan et al., 2021; Wang et al., 2024). FedProx regularises local optimisation toward the global model (Li et al., 2020). FedKL penalises policy divergence and controls local policy drift (Xie & Song, 2023). Personalised methods share a representation while keeping client-specific heads in supervised learning and RL (Collins et al., 2021; Xiong et al., 2025). These methods change the aggregation or personalisation pattern. Our question is complementary: can the shared representation be trained to discard information that is irrelevant to control? The closest related method is FedRAG, which federates a bisimulation-style encoder for pixel control and discusses video distractors in its extended appendix (Guo et al., 2025). Among our baselines, the SimSR port is its closest state-based analogue. We study client-specific state-vector noise and its interaction with averaging directly.

Representations under exogenous noise. An Ex-BMDP augments a controllable state with variables that evolve independently of the agent (Efroni et al., 2022a;b). Such variables can slow learning (Dietterich et al., 2018). Hidden- parameter block MDPs likewise share structure across environments with different physics (Zhang et al., 2021b), while TIA and Denoised MDPs separate controllable, reward-relevant factors from other observations (Fu et al., 2021; Wang et al., 2022). Multi-step inverse models, behavioural metrics, and self-predictive objectives are three ways to seek that separation (Lamb et al., 2023; Levine et al., 2024; Zhang et al., 2021a; Castro et al., 2021; Zang et al., 2022; Grill et al., 2020; Schwarzer et al., 2021; Assran et al., 2023). SimSR performs strongly on state-based tasks with Gaussian distractors in a recent comparison (Luo et al., 2025). Work on temporal structure also matters here: self-predictive methods can need many more samples when observation noise is correlated rather than i.i.d. (Misra et al., 2024), and Ni et al. analyse the role of stop-gradients under Gaussian distractors (Ni et al., 2024). Our experiments include both the standard appended i.i.d. Gaussian protocol (Nikishin et al., 2022; Dulac-Arnold et al., 2021) and AR(1) noise. Unlike this prior work, the representation must survive parameter averaging across clients with different physics.

Routing the actor–critic representation. SAC-AE, CURL, DrQ, and SimSR let the critic train an encoder while stopping the actor's gradient at its input (Yarats et al., 2021b; Laskin et al., 2020; Yarats et al., 2021a; Zang et al., 2022). For on-policy MuJoCo agents, separate policy and value networks are already a strong standard design; PPG and IDAAC likewise separate their updates (Andrychowicz et al., 2021; Cobbe et al., 2021; Raileanu & Fergus, 2021). Our clean control recovers that result, whereas client-specific distractors reverse it. We study this reversal, and its interaction with federated averaging, rather than presenting network separation itself as the novelty.

3 Problem: federated exogenous block MDPs

We consider K clients. Client k controls an endogenous state $s_t \in \mathcal{S}$ according to

$$s_{t+1} \sim P_{c_k}(\cdot \mid s_t, a_t), \qquad r_t = R_{c_k}(s_t, a_t),$$

where c_k is the client's physics context. The contexts differ in friction or body mass, but the state and action spaces are shared. The client also observes an exogenous process $\xi_t \in \mathbb{R}^m$,

$$\xi_{t+1} \sim P_\xi(\cdot \mid \xi_t),$$

which neither affects the endogenous dynamics nor responds to actions. The processes are independent across clients. The observation is $o_t = q(s_t, \xi_t)$ and, in our experiments, is the concatenation $o_t = [s_t; \xi_t]$. Thus the endogenous state is recoverable from the full observation in principle. The learner is not given a mask or a label identifying which coordinates are safe to use. The term "distractor" refers to this task-irrelevant but observed component, not to an unobservable state. This is the Ex-BMDP construction of Efroni et al. (2022a).

We use unit-variance exogenous processes. The i.i.d. protocol draws a fresh $\xi_t \sim \mathcal{N}(0, \boldsymbol{I}_m)$ at every step (Nikishin et al., 2022). The correlated protocol is AR(1):

$$\xi_{t+1} = \rho \xi_t + \sqrt{1-\rho^2}\,\epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \boldsymbol{I}_m), \qquad \rho \in \{0.5, 0.95, 0.99\}.$$

The correlation time is $-1/\ln \rho$, about 20 steps at $\rho = 0.95$ and 100 steps at $\rho = 0.99$. Related work shows that this distinction can materially change the sample complexity of predictive representation learning (Misra et al., 2024). Our experiments test whether the federated PPO result changes in the same way.

Objective and evaluation. Clients run PPO locally and periodically average their parameters. Trajectories never leave the client. The desired policy should work across the K physics contexts,

$$J_{\text{global}}(\pi) = \frac{1}{K} \sum_{j=1}^{K} J_{c_j}(\pi), \qquad J_c(\pi) = \mathbb{E}_{\pi,c}\left[\sum_t r_t\right].$$

Local training can leave the K client copies different, so we evaluate each final policy π_k on every context c_j. The resulting matrix has one row per trained client copy and one column per evaluation context: $Y_{kj} = J_{c_j}(\pi_k)$. These K^2 entries come from cross-evaluation rather than K^2 separately

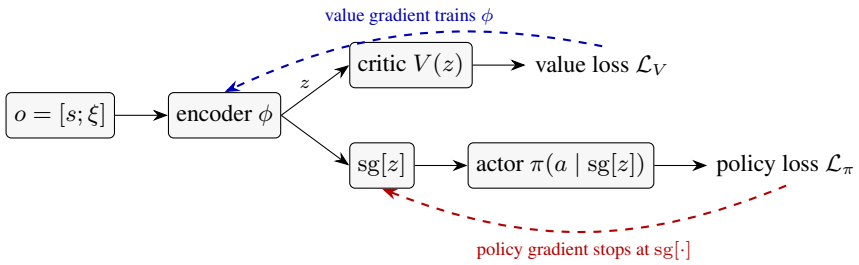

Figure 1: Critic routing. Here $\text{sg}[z]$ means a stop-gradient copy of z: its numerical value is z, but gradients from the actor loss do not flow into ϕ. The value loss still trains the encoder. Every five rounds the server averages the encoder, actor, critic, and auxiliary modules, and pools observation-normalisation statistics.

trained environments. There are K clients, each with one context, and every policy is evaluated on all K contexts. We report the mean of Y. Single agents, trained without federation on one client, are evaluated with the same cross-context protocol. Since rewards and endogenous dynamics depend on s_t alone, an optimal policy should ignore ξ_t, but raw-observation PPO must learn that structure from its own returns.

4 Latent federated PPO

PPO uses two networks. The actor maps an input to an action distribution, while the critic estimates the value of that input and supplies the return baseline used by the actor update. In a raw-observation agent, both networks receive o directly. A latent agent first computes $z = \phi(o)$ and gives z to the critic and the actor. The encoder is a tanh MLP with two hidden layers of 256 units and a LayerNorm output of dimension 64. All agents otherwise use the same federated PPO implementation (Algorithm 1 in Appendix C).

Critic routing. The actor reads $\text{sg}[z]$, so the policy loss cannot update the encoder. The value loss and any auxiliary loss can. This separation reflects the different signals available to the two updates. The distractors have no direct effect on rewards or endogenous dynamics. Value regression can therefore learn to discount them from repeated return targets. The policy gradient instead multiplies every input-dependent score by a noisy return estimate, and its variance grows with the number of distractors in the linear model of Section 5. Value updates can still move the actor's input outside PPO's clipped ratio, an effect retained by our implementation (Appendix C).

When no auxiliary loss is present, the value loss alone trains ϕ. We call this the *critic-trained encoder*. Its *narrow* version uses one hidden layer of 64 units. The resulting agent has approximately 1.4 times as many parameters as PPO, compared with six or more times for the wide latent agents. The *jointly trained encoder* has the wide architecture but lets the policy gradient update ϕ as well. This is the control that separates routing from model capacity.

Auxiliary representation objectives. We add three objectives to the critic-trained encoder. The *predictive* objective is a four-step, action-conditioned latent predictor in the style of SPR and JEPA (Schwarzer et al., 2021; LeCun, 2022; Assran et al., 2023). It predicts future latents from the current latent and intervening actions, and compares them with an exponential-moving-average (EMA) target encoder. The EMA is a slowly updated copy that supplies a more stable target:

$$\mathcal{L}_{\text{pred}} = \frac{1}{H} \sum_{j=1}^{H} \left\| \hat{z}_{t+j} - \text{sg}\left[\bar{\phi}(o_{t+j})\right] \right\|_2^2, \qquad \hat{z}_{t+j} = f(\hat{z}_{t+j-1}, a_{t+j-1}), \quad \hat{z}_t = \phi(o_t),$$

with $H = 4$. A reward head and VICReg variance and covariance terms prevent collapse (Bardes et al., 2022). All predictive terms use PPO's gradient-clipped Adam.

The *behavioural-metric* objective (SimSR) regresses cosine distances between latents onto $|r_i - r_j| + \gamma d(\hat{z}'_i, \hat{z}'_j)$, where the next latents come from a probabilistic transition ensemble with an EMA

target and an L2-normalised latent (Zang et al., 2022). The *multi-step inverse* objective (AC-State) predicts the first action from $(z_t, z_{t+\tau}, \tau)$, with $\tau \leq 8$ sampled within an episode (Lamb et al., 2023). These two ports retain their reference losses and use their own unclipped optimisers: Adam at 10^{-3} for SimSR and AdamW at 10^{-4} for AC-State. Thus the comparison tests both the representation objective and, as a secondary result, the optimiser used to apply it (Appendix D).

Federated update. Every five rounds of 2048 local steps, the server averages the encoder, EMA target, auxiliary heads, actor, and critic with equal weights. It also pools observation-normalisation statistics so that clients use one input scale. Reward normalisation remains local. Sharing all modules is the default. We later vary which modules are shared to distinguish a representation that is useful locally from one that remains useful after averaging.

5 Theory: why a noise-free latent is safe to share

We complement the empirical study with a linear–Gaussian model that makes the effect of distractors explicit. Client k has endogenous dynamics $s' = A_k s + B_k a + w$ and exogenous dynamics $\xi' = \rho\xi + \sqrt{1-\rho^2}\,\epsilon$, with observation $o = [s; \xi]$. The linear encoder $z = W_s s + E\xi$ separates the state block W_s from the noise block E. A second-moment constraint on z plays the role of LayerNorm and VICReg. The actor, critic, predictor, and reward heads are fitted to their mean-squared-error optima, and the critic uses GAE λ-returns. FedAvg averages every module. The model omits Adam and PPO clipping so that the representation dynamics can be analysed in closed form. Proposition 1 carries the formal argument: its invariance result extends to nonlinear networks and to the routing, EMA, and averaging operations used by the implemented algorithm. Theorems 2–4 are interpretive analyses of the learning signals. Assumptions (A1–A9) and complete proofs appear in Appendix A.

A shared noise-free fixed point (Proposition 1). The set $E = 0$ contains encoders whose latent state depends only on the endogenous state. At this point, exogenous coordinates are independent of the regression targets used by the value, reward, and prediction heads. Assigning them nonzero weight can therefore add variance but cannot reduce the optimal squared loss. Consequently, each loss leaves $E = 0$ unchanged. The same property holds after an EMA update, under either gradient routing, and after FedAvg for any synchronisation period or subset of shared modules. This establishes that clients with different physics agree on a noise-free fixed point. It does not require them to approach that point from every initialisation.

How temporal correlation changes the learning signal (Theorem 2). Near $E = 0$, the three regression channels act differently on a small noise component. The reward head is trained on an immediate reward that depends only on (s, a) under the Ex-BMDP assumptions. Its target contains neither a bootstrapped value nor a future latent. The target's cross-covariance with ξ is therefore zero for every ρ, so the reward channel always pulls the component toward zero with the same local force.

The value and prediction targets refer to the future, where correlated noise can reappear. For the critic, GAE bootstrapping carries a fraction $c_V = \gamma(1-\lambda)\rho/(1-\gamma\lambda\rho)$ of a noise component already present in the value estimate. The effective restoring force is multiplied by $1 - c_V$. With $\gamma = 0.99$ and $\lambda = 0.95$, this factor falls from 0.56 at $\rho = 0.95$ to 0.29 at $\rho = 0.99$. Value regression still favours ignoring the distractor, but its correction becomes weaker as the distractor persists.

The predictive target creates a second effect. If the target encoder retains a small amount of AR(1) noise, then the same coordinate remains predictable j steps later with coefficient ρ^j. Preserving that coordinate can lower the predictive loss even though it has no control value. Prediction suppresses a distractor when its persistence is weaker than the latent dynamics being modelled, but can reinforce it when the distractor is more persistent. The noise-free set remains locally stable when this target-induced reinforcement is smaller than the combined restoring forces from the value, reward, and prediction losses. The condition is sharp in the scalar case.

Critic routing alone has a more specific implication. Its local force is $\lambda_V G_V$, where G_V is the outer product of the critic's gradient and hence has rank one. It removes the component of each noise column that affects the critic, while orthogonal components receive no first-order update. The theorem therefore allows residual noise weights in the first encoder layer. Whether later nonlinear

layers suppress those components is an empirical question, which we examine through the actor's input sensitivity.

The distractor tax (Theorem 3, linear–quadratic). A raw policy must learn zero action gain for every exogenous coordinate. Although the objective $J = J_s(K_s) + q_\rho(K_\xi)$ is minimised at $K_\xi = 0$, the score-function gradient along K_ξ remains noisy. Its variance is at least linear in the number of distractors m, regardless of the baseline. For any step-size schedule α_t and each iteration,

$$\mathbb{E}J(K_t) - J^\star \geq \left(\mathbb{E}J_s(K_{s,t}) - J^\star\right) + C\,\alpha_{t-1}^2(m - d_u)_+,$$

where d_u is the action dimension and C is given in Appendix A.4. This lower bound concerns plain SGD. The extra term need not grow with m if the step size decreases as $m^{-1/2}$, and our deep agents use Adam. An exact noise-free latent avoids the term (Corollary 2). A learned latent instead incurs error proportional to its representation error, which decreases with the number of pooled samples (Theorem 4).

Why averaging helps (Theorem 4). All clients use the AR(1) process defined in Section 3, with the same ρ and stationary variance, but draw independent realisations. A noise weight fitted to one client's sample has no systematic value for another client. Averaging these independent estimation errors reduces their variance while preserving the common fixed point. With synchronisation after every step and n samples per client, the residual noise block has second moment of order $m(1+\rho)/(Kn(1-\rho))$. It therefore decreases with the pooled sample size Kn and increases with temporal persistence. Physics heterogeneity still affects the actor and critic heads, but it does not bias the shared fixed point of the noise block. Our implementation approximates the theorem's shared input scale by pooling observation statistics every five rounds.

Experimental predictions. The analysis motivates three tests. First, the latent agent's advantage should grow with m when the raw policy uses a fixed learning-rate schedule. Second, critic routing should reduce the distractor component that influences value and should yield an actor with low distractor sensitivity. Third, increasing ρ should weaken value-based denoising and make self-prediction more likely to preserve a distractor. Extending these local results to the full nonlinear training dynamics, including the interaction between encoder layers, is beyond the present analysis. We therefore test the predicted mechanisms directly rather than treating the linear model as a complete account of the deep agent.

6 Experimental setup

Tasks and clients. We use six MuJoCo tasks from Gymnasium (Towers et al., 2025; Todorov et al., 2012), with $K = 4$ clients per task. Clients share the task and action space but differ in physics. Walker2d, Hopper, and HalfCheetah use four equally spaced friction values. Ant uses four torso masses from the RRLS robustness suite (Zouitine et al., 2024). MassHopper uses the four fixed mass configurations of CausalGym (Causal Artificial Intelligence Lab, Columbia University, 2026), and InvertedDoublePendulum scales both pole masses. Each client receives its own distractor stream.

Distractors. Following the standard protocol (Nikishin et al., 2022), we append m i.i.d. $\mathcal{N}(0,1)$ coordinates to the state. The main comparison uses $m = 64$. For the predictive agent we additionally sweep $m \in \{16, 32, 48, 64\}$ and use $m = 16$ AR(1) coordinates at $\rho \in \{0.5, 0.95, 0.99\}$. These experiments separate the effect of dimension from temporal persistence. They include the predictive agent but not the critic-trained encoder.

Training and evaluation. Each client takes 749,568 environment steps (366 rounds of 2,048 steps) with PPO and synchronises every five rounds. Hyperparameters, head sizes, and budgets are identical across agents unless stated otherwise (Appendix C). At the end of training, each final policy runs 10 episodes on each client's context, with actions sampled from the policy. We report the mean of the resulting 4×4 cross-evaluation matrix. Each agent–task pair has 10 paired seeds. A seed fixes contexts, noise streams, and initial states. We use exact paired sign-flip tests with Holm correction. The implementation and analysis scripts are included in the supplement (Appendix F).

Table 1: Global return with 64 i.i.d. distractor dimensions, averaged over all 4×4 client–context cells (10 seeds, mean $\pm$ s.e.). Every agent with critic routing beats every aggregation baseline in 10/10 seeds (Holm-adjusted $p = 0.035$ within its 18 comparisons, the smallest attainable), except SimSR on Hopper (7–8/10 seeds, uncorrected $p = 0.03$–0.23). The encoder-only agents have no self-supervised objective.

	Agent	Walker2d	Ant	Hopper	MassHopper	HalfCheetah	InvDblPend.
raw obs.	FedAvg-PPO	328±3	571±18	384±15	304±39	422±21	294±6
	FedProx-PPO	371±5	537±13	345±17	339±33	316±27	281±5
	FedKL-PPO	327±4	584±27	391±14	314±42	405±23	295±6
encoder only	Jointly trained encoder	394±8	491±20	257±11	231±14	442±16	300±5
	Critic-trained encoder	1391±72	805±24	1016±48	991±83	2064±86	2308±300
with objective	Predictive (JEPA-style)	720±78	1100±77	659±30	902±64	865±66	2088±281
	Behavioural metric (SimSR)	1347±119	1752±64	430±22	771±64	2038±203	4870±184
	Multi-step inverse (AC-State)	1274±181	2356±225	802±30	880±122	2583±124	7008±284

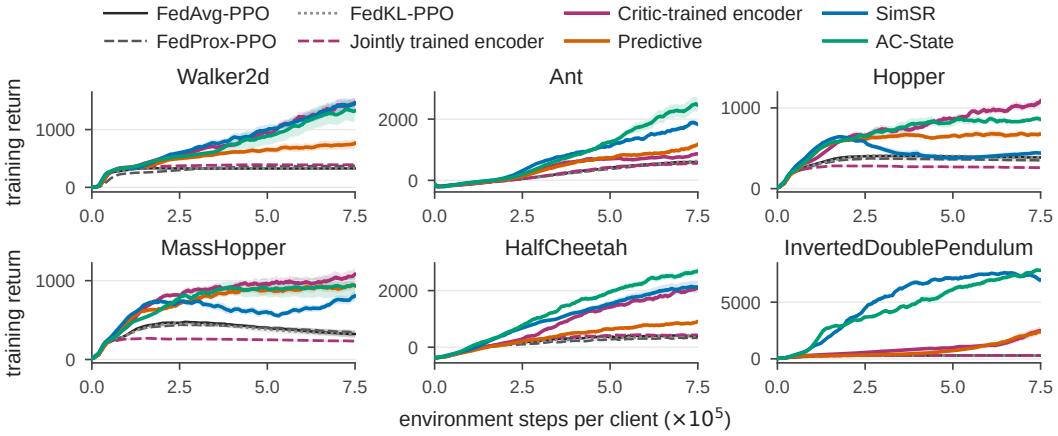

Figure 2: Training return with 64 i.i.d. distractors. Curves average each client on its own context over 10 seeds. Bands show $\pm$ s.e. after a 10-round trailing average.

Agents and calibration. FedAvg-PPO, FedProx-PPO, and FedKL-PPO run PPO on raw observations and differ only in how local updates are regularised or combined. FedProx and FedKL coefficients are tuned on three held-out seeds per task. SimSR and AC-State are ports of the authors' methods. Their representation learning rates retain the reference values and are not tuned to the reported conditions. The predictive agent and the two encoder controls use one global setting from earlier single-agent studies. On three tasks we also run two alternative FedAvg-PPO step-size/minibatch settings to measure the benefit available from baseline tuning. The comparison first changes the aggregation rule while retaining raw inputs, then changes the representation while holding the PPO budget fixed.

7 Results

7.1 Sharing a latent versus changing the aggregation rule

Changing the aggregation rule leaves both observations and actor architecture fixed (Table 1). FedKL-PPO stays within 4% of FedAvg-PPO and never differs significantly; FedProx-PPO ranges from -25% to $+13\%$. Only its Walker2d gain survives Holm correction, where calibration selected $\mu = 0.1$ rather than 10^{-3}. These rules limit policy drift but do not change what the actor can learn from its inputs: all three raw-observation actors retain roughly their initial distractor-weight share (Appendix E).

Critic-routed agents reach 1.4–24 times the best baseline's return, with SimSR on Hopper the sole exception. The critic-trained encoder alone reaches 1.4–7.8 times the best baseline, and its worst

seed beats every raw baseline's best seed on every task. Tuning FedAvg-PPO's step size or minibatch closes only 3–7% of the gap on three tasks; the encoder still wins in every seed (Table 15).

The gap opens early (Figure 2): on four tasks raw agents level off by 2.5×10^5 steps, while critic-routed agents pull away by 3.4×10^5. The critic-trained encoder is still improving at the budget's end on every task, whereas FedAvg-PPO is flat or declining on five. These are therefore end-of-budget, not asymptotic, comparisons (Appendix E).

7.2 What trains the shared latent

A latent agent differs from FedAvg-PPO in size (6 to 101 times its parameters), in routing and in objective, and paired controls separate these factors. We apply Holm correction over each comparison's six tasks (Appendix E.2).

Routing and width. The jointly trained encoder has the same latent architecture but lets the policy gradient update the encoder. It beats FedAvg-PPO significantly only on Walker2d and loses to it on Hopper (Table 16). Critic routing reverses this pattern: it raises the return on every task, by 315–2008, in all 10 seeds. The gain is not a consequence of capacity. The narrow critic-trained encoder has about 1.4 times PPO's parameters and communication, yet still wins on every task in all 10 seeds. On five tasks it retains 78–141% of the wider encoder's gain; on InvertedDoublePendulum it is substantially better (7939 versus 2308; Table 14). We therefore recommend the narrow model as the communication-conscious configuration. The decisive difference is which loss trains the representation, not how wide it is.[1]

Without distractors. The clean-observation control reverses the main pattern. The critic-trained encoder is worse than FedAvg-PPO on four tasks and statistically indistinguishable on two (Table 12). Adding distractors produces a positive interaction on all six tasks (Holm-adjusted $p = 0.012$ for each). More surprisingly, the encoder's absolute return is higher with noise than without it on Walker2d, Hopper, and MassHopper. Filtering alone cannot explain an improvement beyond clean performance, so the distractors also interact with encoder optimisation in a way we do not yet understand.

Objective. Auxiliary losses are not uniformly beneficial. All three improve on the critic-trained encoder on Ant; SimSR and AC-State also help on InvertedDoublePendulum, and AC-State on HalfCheetah. AC-State's Ant gain is 6.6 times the encoder's own gain over PPO. Yet all three hurt Hopper, and none significantly improves Walker2d, Hopper, or MassHopper. They are task-dependent additions, not replacements for critic routing.

Optimiser. The predictive agent uses PPO's clipped Adam; SimSR and AC-State use separate unclipped optimisers. Giving the predictive agent SimSR's optimiser raises every return, significantly on four (Appendix E.3), but also changes the auxiliary loss's effective weight. Exploratorily, it beats critic routing on Ant and InvertedDoublePendulum, as SimSR and AC-State do. Objective targets alone therefore do not explain their differences.

7.3 What the shared networks keep

FedAvg-PPO's two networks treat distractors differently (Table 2). Its critic's first layer sheds most of their weight, while its actor stays near its initial share; distractors account for 26–55% of actor sensitivity on the locomotion tasks. The share is only 7% on InvertedDoublePendulum, yet PPO fails there. This relative measure does not reveal how much absolute action perturbation the balance task tolerates.

Every critic-routed agent, in contrast, assigns at most 7.4% of actor sensitivity to distractors. First-layer weight is a poor guide: the critic-trained encoder retains 22–52% there, as its rank-one value channel allows (Theorem 2), but at most 6.1% at the actor, so later layers suppress what the first passes on. The jointly trained encoder resembles PPO's actor, retaining 73–80% first-layer weight on four tasks and 40–54% actor sensitivity on the locomotion tasks.

[1] The jointly and critic-trained encoders also use the same clipping groups and thresholds: policy, encoder, and auxiliary parameters are clipped separately at norm 0.5. A gradient-clipping difference therefore cannot explain the routing gain.

Table 2: Share of the distractors, in %, in each network's first-layer weight, $\|W_\xi\|_F^2/\|W\|_F^2$ (left of each cell), and in the actor's input sensitivity, $\mathbb{E}\|\partial\bar{a}/\partial o_\xi\|_F^2/\mathbb{E}\|\partial\bar{a}/\partial o\|_F^2$ for the actor mean $\bar{a}$ (right). Means over 10 seeds (Appendix E). At initialisation the weight share is $64/(64+d)$ for state dimension d: 79–88%, and 38% for Ant.

Network	Walker2d	Ant	Hopper	MassHopper	HalfCheetah	InvDblPend.
FedAvg-PPO, actor	77.6 / 49.0	38.9 / 25.8	84.0 / 41.4	83.5 / 38.2	77.5 / 55.4	77.2 / 6.9
FedAvg-PPO, critic	8.3 / –	3.9 / –	6.1 / –	10.5 / –	8.8 / –	9.9 / –
Jointly trained encoder	73.0 / 54.4	38.7 / 40.3	79.9 / 53.1	80.0 / 51.0	74.6 / 48.9	51.0 / 1.8
Critic-trained encoder	22.5 / 0.5	31.0 / 6.1	35.8 / 0.6	32.4 / 0.6	51.0 / 1.7	52.2 / 0.5
Predictive	18.2 / 6.6	20.0 / 3.2	23.9 / 7.3	25.3 / 7.4	46.7 / 7.0	46.9 / 0.2
SimSR	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.1 / 0.0	0.1 / 0.0
AC-State	9.3 / 0.2	11.7 / 0.7	10.1 / 0.1	11.7 / 0.1	28.4 / 0.5	37.8 / 0.1

Sensitivity tracks return only in part. Prediction raises sensitivity on four tasks, all with lower return, and lowers it on Ant, where return rises; SimSR ignores distractors everywhere but still collapses on Hopper. Critic routing reliably produces an actor that ignores distractors, but federation is a separate question.

7.4 Sharing and distractor structure

Federation raises the critic-trained encoder's return on all four tasks in the single-agent comparison, significantly on two (Ant and Hopper; Table 13); this effect includes the pooled observation normaliser as well as parameter averaging. The encoder already beats PPO before sharing, so federation strengthens rather than creates its advantage. For the predictive agent at $m = 16$, averaging every module wins 15 of 16 comparisons with partial splits (Table 11). The representation-only rows instantiate the FedRep-style design that shares a representation while keeping heads local, and this design collapses on Ant. Independent agents remain best on Walker2d with i.i.d. noise. Sharing helps here only when the representation and policy can adapt together.

The full sweep in Table 9 separates distractor count from persistence. As m grows from 16 to 64, the predictive agent's return relative to FedAvg-PPO increases on all four tasks, largely because PPO loses 29–60% of its return. At $\rho = 0.99$, by contrast, the gain is smaller than with 16 i.i.d. distractors on Walker2d and Ant, but not on Hopper or MassHopper. Persistent noise can be retained by the predictive target, consistent with Theorem 2, but the effect is not uniform across tasks. Moreover, this sweep includes only the predictive agent; the theorem's value-channel prediction remains untested for the critic-trained encoder.

8 Discussion

Across six tasks, the learning signal used by the shared encoder matters more than the aggregation rule. Critic routing outperforms all three raw-observation baselines under 64 distractors. Its advantage survives a large reduction in encoder capacity, disappears when the policy gradient shapes the encoder, and is absent without distractors. These controls and the sensitivity measurements show that nuisance suppression at the actor's input is necessary and capacity is not. Auxiliary losses add a task-dependent effect: lower first-layer noise weight does not always improve return, while optimiser choice and temporal correlation both matter. The objective and its optimisation should be treated as one design choice.

Scope. The evidence covers six MuJoCo tasks, four clients, and appended Gaussian distractors; pixels, realistic distractors, and larger federations remain untested. The theory captures a local linear mechanism, while the experiments test deep encoders, and the reason noise improves the encoder on three tasks remains unexplained. The wide critic-trained encoder costs about 7 times PPO's floating-point operations and 6 times its communication per round (Appendix F); the recommended narrow model reduces both parameter and communication overhead to about 1.4 times PPO's. Critic routing is intended for nuisance features that do not predict value. Value-relevant features should instead be retained.

AI USE STATEMENT

AI tools were used to assist with existing literature analysis and discovery, code implementation, and the development and verification of mathematical proofs. AI tools were also used for language editing and polishing of the manuscript. The authors independently evaluated, verified, and took responsibility for the resulting analyses, implementations, mathematical claims, and final text.

REPRODUCIBILITY STATEMENT

Sections 3 and 6 define the setting and the protocol, and Appendix C lists every hyperparameter. Appendix D documents each port, its deviations from the reference code, and its calibration. Appendix A states all assumptions and gives complete proofs, and Appendix F gives the tests. The supplementary material contains the code, the run logs and the scripts that produce every table and figure.

ETHICS STATEMENT

This work uses simulated robot-locomotion benchmarks only. It involves no human subjects and no personal data, and it makes no privacy claim about federated learning.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29304–29320, 2021.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? A large-scale study. In *International Conference on Learning Representations*, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. MICo: Improved representations via sampling-based state similarity for Markov decision processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 30113–30126, 2021.

Causal Artificial Intelligence Lab, Columbia University. CausalGym: Causal reinforcement learning playground. `https://github.com/CausalAILab/Causal-Gymnasium`, 2026. Commit c94e9a4.

Karl W. Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2020–2027, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 2089–2099, 2021.

John B. Conway. *Functions of One Complex Variable I*, volume 11 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 1978.

Thomas G. Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

Yonathan Efroni, Dylan J. Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory (COLT)*, volume 178 of *PMLR*, pp. 5062–5127, 2022a.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations (ICLR)*, 2022b.

Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-tolerant federated reinforcement learning with theoretical guarantee. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 1007–1021, 2021.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 1467–1476, 2018.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3480–3491. PMLR, 2021.

Hans Gebelein. Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

Ulf Grenander and Gábor Szegő. *Toeplitz Forms and Their Applications*. University of California Press, 1958.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Bram Grooten, Ghada Sokar, Shibhansh Dohare, Elena Mocanu, Matthew E. Taylor, Mykola Pechenizkiy, and Decebal Constantin Mocanu. Automatic noise filtering with dynamic sparse training in deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1932–1941, 2023.

Zengxia Guo, Bohui An, and Zhongqi Lu. Approximated behavioral metric-based state projection for federated reinforcement learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5298–5306, 2025. Extended version with appendix: arXiv:2505.09959.

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *PMLR*, pp. 18–37, 2022.

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 3rd edition, 2002.

Khimya Khetarpal, Zhaohan Daniel Guo, Bernardo Ávila Pires, Yunhao Tang, Clare Lyle, Mark Rowland, Nicolas Heess, Diana L. Borsa, Arthur Guez, and Will Dabney. A unifying framework for action-conditional self-predictive reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *PMLR*, pp. 181–189, 2025.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Didolkar, Dipendra Misra, Dylan Foster, Lekan Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research (TMLR)*, 2023.

H. O. Lancaster. Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, 44(1/2):289–292, 1957.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5639–5650. PMLR, 2020.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview preprint, 2022.

Alexander Levine, Peter Stone, and Amy Zhang. Multistep inverse is not all you need. *Reinforcement Learning Journal*, 2:884–925, 2024.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 2, pp. 429–450, 2020.

Michael C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

Ziyan Luo, Tianwei Ni, Pierre-Luc Bacon, Doina Precup, and Xujie Si. Understanding behavioral metric learning: A large-scale study on distracting reinforcement learning environments. *Reinforcement Learning Journal*, 6:40–96, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *PMLR*, pp. 1273–1282, 2017.

Dipendra Misra, Akanksha Saran, Tengyang Xie, Alex Lamb, and John Langford. Towards principled representation learning from videos for reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, 1982.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive RL. In *International Conference on Learning Representations (ICLR)*, 2024.

Evgenii Nikishin, Romina Abachi, Rishabh Agarwal, and Pierre-Luc Bacon. Control-oriented model-based reinforcement learning with implicit differentiation. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 7886–7894, 2022.

Roberta Raileanu and Rob Fergus. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8787–8798, 2021.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations (ICLR)*, 2021.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 12, pp. 1057–1063, 1999.

Yunhao Tang, Zhaohan Daniel Guo, Pierre Harvey Richemond, Bernardo Ávila Pires, Yash Chandak, Rémi Munos, Mark Rowland, Mohammad Gheshlaghi Azar, Charline Le Lan, Clare Lyle, András György, Shantanu Thakoor, Will Dabney, Bilal Piot, Daniele Calandriello, and Michal Valko. Understanding self-predictive learning for reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 33632–33656, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, 2025.

A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.

Claas A. Voelcker, Tyler Kastner, Igor Gilitschenski, and Amir-massoud Farahmand. When does self-prediction help? Understanding auxiliary tasks in reinforcement learning. *Reinforcement Learning Journal*, 4:1567–1597, 2024.

Han Wang, Sihong He, Zhili Zhang, Fei Miao, and James Anderson. Momentum for the win: Collaborative federated reinforcement learning across heterogeneous environments. In *International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 50530–50560, 2024.

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning world models better than the world itself. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22591–22612. PMLR, 2022.

Zhijie Xie and Shenghui Song. FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence. *IEEE Journal on Selected Areas in Communications*, 41(4): 1227–1242, 2023.

Guojun Xiong, Shufan Wang, Daniel Jiang, and Jian Li. On the linear speedup of personalized federated reinforcement learning with shared representations. In *International Conference on Learning Representations (ICLR)*, 2025.

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021a.

Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *AAAI Conference on Artificial Intelligence*, 2021b.

Hongyu Zang, Xin Li, and Mingzhong Wang. SimSR: Simple distance-based state representations for deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 8997–9005, 2022.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations (ICLR)*, 2021a.

Amy Zhang, Shagun Sodhani, Khimya Khetarpal, and Joelle Pineau. Learning robust state abstractions for hidden-parameter block MDPs. In *International Conference on Learning Representations (ICLR)*, 2021b.

Adil Zouitine, David Bertoin, Pierre Clavier, Matthieu Geist, and Emmanuel Rachelson. RRLS: Robust reinforcement learning suite. *arXiv preprint arXiv:2406.08406*, 2024.

A THEORY: SETTING, STATEMENTS AND PROOFS

This appendix gives the full statements and proofs of the results summarised in Section 5.

Reading guide. Several ingredients are standard; we prove them in full only to keep the appendix self-contained.

- *Standard.* Theorem 3(a) is least squares with irrelevant regressors (Frisch–Waugh–Lovell and the inverse-Wishart mean). Proposition 1 and Theorem 4(a1) follow from the independence and zero mean of the noise. Theorem 4(c) is a pooled-variance rate.
- *Specific to this setting.* The linearisation of the noise block with its EMA, value and reward channels, and its stability criterion (Theorem 2). The variance floor for raw policy gradient and its consequence at every iteration (Theorem 3(c,d)). The pooled-margin example (Theorem 4(b3)).

Appendix A.1 fixes the setting, the analysed idealisation of the algorithm and the assumptions, with a map to the code. Appendices A.2 to A.5 give the four results with complete proofs; Appendix A.6 summarises the numerical checks. Throughout, "S1–S3" and "A1–A9" refer to Appendix A.1.

A.1 SETTING, THE ANALYSED AGENT, AND THE ASSUMPTIONS

Setting S1 (federated exogenous block MDP). There are K clients. Client k has an endogenous MDP $(\mathcal{S} \subseteq \mathbb{R}^d, \mathcal{A} \subseteq \mathbb{R}^{d_u}, P_k, R_k, \mu_k)$; $\mathcal{S}$ and $\mathcal{A}$ are shared by all clients. An exogenous stationary Markov chain $\xi \in \mathbb{R}^m$ has a law common to all clients and is independent of everything endogenous, and of the other clients' chains. The observation is $o = [s; \xi]$ (in general $q(s, \xi)$ with s decodable). The objective is $\bar{J} = K^{-1} \sum_k J_k$; we report the cross-evaluation mean $K^{-2} \sum_{k,j} J_j(\pi_k)$.

Setting S2 (linear-Gaussian and LQ instances). $s' = A_k s + B_k a + w$ with $w \sim \mathcal{N}(0, \Sigma_{w,k})$; $\xi' = \rho\xi + \sqrt{1-\rho^2}\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, \boldsymbol{I}_m)$; behaviour policy $a = F_k s + \sigma\eta$. The LQ instance adds the cost $s^\top Q_k s + a^\top R_k a$, a discount γ, and policies $a = Ko + \sigma\eta$. The pooled normaliser makes the ξ-block of the second-moment matrix the identity in every client.

Setting S3 (the analysed agent).

- *Encoder:* $z = Wo$ with $W = [W_s, E]$, where $E \in \mathbb{R}^{k \times m}$ is the *noise block*, $k \leq d$, and $W\Sigma W^\top = \boldsymbol{I}_k$ (a second-moment idealisation of LayerNorm and VICReg).
- *EMA target:* $\bar{W} = [\bar{W}_s, \bar{E}]$; the encoder used to compute the λ-returns is the lagged $W_{\text{old}} = [\cdot, E_{\text{old}}]$.
- *Heads:* per-horizon linear predictors $\hat{z}_{t+j} = P_j z_t + Q_j a_{t:t+j-1}$ $(j \leq H)$, a reward head g, and a critic V.
- *Encoder objective (semi-gradient):* $L = \lambda_p L_{\text{pred}}(W; \bar{W}) + \lambda_V L_V(W; W_{\text{old}}) + \lambda_R L_R(W)$, with targets under stop-gradient.
- *Actor:* $a = K_z\,\text{sg}[z] + \sigma\eta$ (critic routing).
- *Federation:* FedAvg of every module, the target included, every E steps, equal weights.

Notation. $M = H^{-1} \sum_j P_j^\top P_j$, $N = H^{-1} \sum_j \rho^j P_j$, $G_V = \mathbb{E}[\nabla_z V \nabla_z V^\top]$, $G_R = \mathbb{E}[\nabla_z g \nabla_z g^\top]$, $c_V = \gamma(1-\lambda)\rho/(1-\gamma\lambda\rho)$,

$$M_{\text{tot}} = \lambda_p M + (1 - c_V)\lambda_V G_V + \lambda_R G_R, \qquad J = M_{\text{tot}}^{-1} \lambda_p N^\top, \qquad s = \left\| M_{\text{tot}}^{-1/2} \lambda_p N^\top M_{\text{tot}}^{-1/2} \right\|,$$

$\Gamma = M_{\text{tot}} - \lambda_p\,\text{sym}(N)$, $c_\rho(H) = H^{-1} \sum_j \rho^{2j}$ and $\Lambda = \lambda_{\min}(M_{\text{tot}})/\lambda_p - c_\rho(H)$.

Assumptions. Each is labelled with its status in our experiments.

- **A1** (independent exogenous block with a common law). Exact: this is how the distractors are generated.
- **A2** (endogenous behaviour policy during the representation dynamics). Self-consistent on the noise-free set, because the actor reads $\text{sg}[z]$; off it, first-order entanglement terms remain (zero for prediction at $H = 1$; $\mathbb{E}[\nabla V \otimes \partial y/\partial a] K_z E$ for the value channel).

- **A3** (linear-Gaussian dynamics; LQ cost in Theorem 3). A model.
- **A4** ($k \leq d$). Violated by the code ($k = 64$); Appendix A.3 explains how the results extend to a feature dictionary.
- **A5** (heads at their L^2 optimum given $(W, \bar{W}, W_{\text{old}})$; per-horizon predictors or $H = 1$). Approximate.
- **A6** (small steps; a slow target for Theorem 2(b)). Faithful for the target; Adam and gradient clipping are not modelled.
- **A7** (n i.i.d. segments or a stationary β-mixing trajectory per client).
- **A8** (FedAvg with equal weights, a common initialisation and all modules averaged; $E = 1$ in Theorem 4(b,c)). We run $E = 5$ rounds, which the theory does not cover exactly.
- **A9** (score-function policy gradient with states from the discounted occupancy, Monte-Carlo returns, any baseline, batch N and SGD steps). PPO clipping, GAE and Adam are not modelled.

Code map. In the code the prediction loss is a mean squared error averaged over the 64 latent dimensions and over H, so $\lambda_p = 0.03/64$ per unit of summed squared error. The value weight is $\lambda_V = 0.5 \times 0.5 = 0.25$ (the value coefficient times the factor $\frac{1}{2}$ of the squared error) and $\lambda_R = 0.1$. The EMA target is updated once per round with $\tau = 0.01$, against 320 online minibatch steps per round. The λ-returns lag by one round, and a sync every 5 rounds is 1,600 local steps. Hence $\lambda_p : \lambda_V : \lambda_R = 4.7 \cdot 10^{-4} : 0.25 : 0.1$: prediction is the weakest of the three channels wherever the critic or the reward head is sensitive, and the only one where both are blind.

A.2 Exogenous neutrality (Proposition 1)

Proposition 1 collects what can be said about exogenous noise without the linear–Gaussian model of Theorems 2–4: for general endogenous dynamics, for nonlinear networks, and for the algorithm as implemented. Parts (a), (b) and (d) are statements about the joint law of trajectories. Part (c) is a statement about the *population recursion* of the implemented algorithm (Definition 2), in which every minibatch gradient is replaced by its expectation. The finite-sample algorithm does leave the exogenous-free set; how far it goes is the subject of Theorems 2 and 4.

Model. We use a structural form of the federated Ex-BMDP of Setting S1, in which all randomness is explicit.

Definition 1 (structural federated Ex-BMDP)**.** There are K clients. For client k, measurable maps $F_{0,k}, F_k$ and an i.i.d. sequence $(w_t^k)_{t\geq 0}$ generate the endogenous variables: $s_0^k = F_{0,k}(w_0^k)$ and, for $t \geq 0$,

$$\big(r_t^k,\ b_t^k,\ s_{t+1}^{k,+},\ s_{t+1}^k\big) = F_k\big(s_t^k, a_t^k, w_{t+1}^k\big).$$

Here $r_t \in \mathbb{R}$ is the reward and $b_t = (\text{term}_t, \text{trunc}_t) \in \{0,1\}^2$ are the episode-boundary flags, with $\bar{b}_t := \max(\text{term}_t, \text{trunc}_t)$. The state reached by a_t is s_{t+1}^+, and s_{t+1} is the state at which a_{t+1} is taken: $s_{t+1} = s_{t+1}^+$ if $\bar{b}_t = 0$, and a fresh initial state otherwise (an episode clock may be part of s). The exogenous process of client k is $\xi^k = (\xi_t^k)_{t\geq 0}$, with values in $\mathbb{R}^m$. The observations are $o_t = q(s_t, \xi_t)$ and $o_{t+1}^+ = q(s_{t+1}^+, \xi_{t+1})$, for a measurable q with a measurable decoder f, $f(q(s,x)) = s$. Actions are $a_t = \pi_t(h_t, \eta_t)$ for fixed measurable maps π_t, where $h_t = (o_{0:t}, a_{0:t-1}, r_{0:t-1}, b_{0:t-1})$ and (η_t^k) is i.i.d. and independent of all other randomness. We write $\mathcal{F}^{\text{en}}$ for the σ-algebra generated by all (w_t^k), all (η_t^k) and the learner's own randomisation (the minibatch index sets).

In the code, the simulator, the action sampler, the distractor generator and the minibatch sampler use separate random-number streams. The step-indexed stream wrapper (used for the shared-versus-independent-noise experiment) matches Definition 1 within a round: the observation returned by the step that ends an episode and the reset observation that follows carry the same exogenous value. The default wrapper, used in all other runs, federated or not, redraws ξ at every reset. The remark after the proof covers it.

Assumptions used. We state A1 and A2 in exactly the form the proofs use.

(A1) The exogenous processes $(\xi^1, \ldots, \xi^K)$ are jointly independent of $\mathcal{F}^{\text{en}}$. They may be independent across clients or identical. Each ξ_t^k has a law P_ξ that does not depend on (k,t),

with $\mathbb{E}\|\xi_t\|^2 < \infty$. *Gaussian AR(1)* noise (used in part (d)) means $\xi_0 \sim \mathcal{N}(0, \boldsymbol{I}_m)$ and $\xi_{t+1} = \rho\xi_t + \sqrt{1-\rho^2}\,\epsilon_{t+1}$, where the ϵ_t are i.i.d. $\mathcal{N}(0, \boldsymbol{I}_m)$, independent of ξ_0, and $\rho \in [0, 1)$. The case $\rho = 0$ is i.i.d. noise.

(A2) The behaviour policy is endogenous: π_t depends on h_t only through $h_t^{\mathrm{en}} = (s_{0:t}, a_{0:t-1}, r_{0:t-1}, b_{0:t-1})$.

Further notation.

- A measurable g of (o, a) is *ξ-blind* if $g(o, a) = \hat{g}(f(o), a)$ for a measurable $\hat{g}$, that is, if $g(q(s, x), a)$ does not depend on x.
- For $j \geq 0$, $\mathcal{G}_t^{(j)} := \sigma(s_t, a_{t:t+j-1})$, with $\mathcal{G}_t^{(0)} = \sigma(s_t)$. One may also use $\mathcal{G}_t = \mathcal{G}_t^{(H)}$. For a j-step predictor with $j \leq H$, under (A1)–(A2), the two choices give the same ξ-average, by Lemma 1(ii).
- $\mathcal{N}_m := \mathcal{N}(0, \boldsymbol{I}_m)$, and $\|\cdot\|$ on functions is the norm of $L^2(\mathcal{N}_m)$.
- *The code's return.* Let the buffer have length n, and let one critic V give both the stored values $v_t = V(o_t)$ and the bootstraps $V(o_{t+1}^+)$. `compute_gae` sets $A_n := 0$, $A_t = \delta_t + \gamma\lambda(1-\bar{b}_t)A_{t+1}$ with $\delta_t = r_t + \gamma(1 - \mathrm{term}_t)V(o_{t+1}^+) - v_t$, and $R_t := A_t + v_t$. The return's horizon is $L_t := \min\{l \geq 1 :\ \bar{b}_{t+l-1} = 1 \text{ or } t + l - 1 = n - 1\}$, and $\tau_t := \mathrm{term}_{t+L_t-1}$.
- The clipped value loss of PPO is, for $\epsilon > 0$,

$$\ell_\epsilon(v; v_o, R) := \max\left\{(v - R)^2,\ (v_o + \mathrm{clip}(v - v_o, -\epsilon, \epsilon) - R)^2\right\}. \tag{1}$$

  The code halves this and averages it over the minibatch.

Networks, Θ_0 and the implemented algorithm (for part (c)). The networks that read observations receive the normalised observation $x_t = (x_{s,t}, x_{\xi,t})$, with $x_{s,t} = \kappa_s(s_t) \in \mathbb{R}^{d'}$ and $x_{\xi,t} = \kappa_\xi(\xi_t) \in \mathbb{R}^m$, and likewise x_{t+1}^+ for o_{t+1}^+. The maps κ_s, κ_ξ are fixed and measurable: they are the frozen observation normaliser, clipping included. Lemma 7 treats the running normaliser. Each such network reads x only through an affine first layer $h = W_{1,s}x_s + W_{1,\xi}x_\xi + c_1$, followed by an arbitrary measurable map: activations, further layers, LayerNorm and, for heads on the latent, their other inputs.

- In the latent agent, these networks are the online encoder ϕ and the EMA target $\bar{\phi}$. The parameters at the start of a round, which produce the stored values and log-probabilities, are copies of the current ones.
- In raw PPO, they are the actor and the critic.

Let $\Xi(\theta)$ be the list of the blocks $W_{1,\xi}$ of these networks, and

$$\Theta_0 := \{\theta :\ \Xi(\theta) = 0\}.$$

A network with $W_{1,\xi} = 0$ computes $h = W_{1,s}\kappa_s(f(o)) + c_1$, so it is ξ-blind.

Definition 2 (implemented algorithm and its population recursion)**.** Every client runs rounds $r = 0, 1, \ldots$. Let $\theta^{(r,i)}$ denote all of its parameters (target included) after i minibatch steps of round r, and $\mathfrak{o}^{(r,i)}$ its optimiser state.

(i) *Collection.* Run n steps with the actor of $\theta^{(r,0)}$. Store $o_t, o_{t+1}^+, a_t, r_t, b_t, v_t = V_{\theta^{(r,0)}}(o_t)$ and $\ell_t = \log \pi_{\theta^{(r,0)}}(a_t \mid o_t)$.

(ii) *Targets.* Compute R_t and A_t with the critic of $\theta^{(r,0)}$, as above.

(iii) *Updates.* For $i = 0, \ldots, N-1$: draw a minibatch index set I_i and set $g_i = \widehat{\nabla}\mathcal{L}_{I_i}(\theta^{(r,i)}; \mathcal{D}_r)$, where $\mathcal{D}_r$ is the round's data and $\widehat{\nabla}$ is reverse-mode differentiation. At the kinks of max, clip and ReLU, $\widehat{\nabla}$ uses the implementation's fixed one-sided convention. Then set $(\theta^{(r,i+1)}, \mathfrak{o}^{(r,i+1)}) = \mathrm{Opt}_r(\theta^{(r,i)}, \mathfrak{o}^{(r,i)}, g_i)$: gradient-norm clipping per parameter group, followed by Adam with step size α_r. If the EMA runs per step, also set $\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$.

(iv) *End of round.* Apply the per-round EMA.

(v) *Federation.* Every E rounds, a subset $\mathcal{M}$ of modules is replaced, in every client, by $\sum_k \omega_k \theta^k|_{\mathcal{M}}$. The code uses $\omega_k = 1/K$. The observation normaliser may also be pooled.

The minibatch loss $\mathcal{L}_I$ is a weighted sum of the following terms:

- the clipped surrogate, with the advantages standardised within the minibatch, plus the entropy term;
- the clipped value loss $\ell_\epsilon(V(z_t); v_t, R_t)$;
- the H-step prediction loss $\frac{\lambda_p}{H}\sum_{j\le H}\mathrm{MSE}(\hat{z}_{t+j}, \bar{\phi}(o^+_{t+j}))$, over windows inside one episode, with $\hat{z}_t = \phi(o_t)$ and $\hat{z}_{t+j} = \hat{f}(\hat{z}_{t+j-1}, a_{t+j-1})$;
- the reward loss $\lambda_R\,\mathrm{MSE}(\hat{g}(z_t, a_t), r_t)$;
- the VICReg term on $\{z_t\}_{t\in I}$.

The routing (joint, critic, detached, actor-shaped or phased) decides whether each head reads $z_t = \phi(o_t)$ or $\mathrm{sg}[z_t]$. Raw PPO has the identity encoder and the policy and value terms only.

The *population recursion* is (i)–(v) with each g_i replaced by $\bar{g}_i := \mathbb{E}\big[\widehat{\nabla}\mathcal{L}_{I_i}(\theta; \mathcal{D}_r)\big]\big|_{\theta=\theta^{(r,i)}}$. The expectation is over the round's data and the index set, with the parameters, which are now deterministic, held fixed.

The remaining two conditions are:

(N) $\mathbb{E}\|\kappa_\xi(\xi_t)\| < \infty$ and $\mathbb{E}\,\kappa_\xi(\xi_t) = 0$ for all t. This holds for the exact population normaliser whenever P_ξ is symmetric, as for Gaussian noise. Then ξ has mean 0 and variance 1, so κ_ξ is, up to the 10^{-8} regulariser, the coordinatewise clip to $[-10, 10]$, which is odd.

(I) At every point visited by the population recursion, $\mathbb{E}\|\widehat{\nabla}\mathcal{L}_I\| < \infty$, and $\mathbb{E}\|\delta_u\|\,\|x_{\xi,u}\| < \infty$ for every adjoint δ_u of Lemma 5. In the code every input is clipped, the reward is clipped, and the Gaussian action samples have moments of all orders.

Proposition 1 (exogenous neutrality)**.** *Fix a client, $t \ge 0$ and $j \ge 0$, and write $\mathcal{G} = \mathcal{G}_t^{(j)}$.*

(a) Loss decomposition. *Let $Y \in L^2(\mathbb{R}^p)$ satisfy*

$$\mathbb{E}[Y \mid \mathcal{G} \vee \sigma(\xi_t)] = \mathbb{E}[Y \mid \mathcal{G}] \qquad \text{a.s.}, \tag{2}$$

let g be measurable with $g_t := g(o_t, a_{t:t+j-1}) \in L^2$, and put $\bar{g}_t := \mathbb{E}[g_t \mid \mathcal{G}]$. Then

$$\mathbb{E}\|Y - g_t\|^2 = \mathbb{E}\|Y - \bar{g}_t\|^2 + \mathbb{E}\|g_t - \bar{g}_t\|^2. \tag{3}$$

Consequences:

- *g minimises $\mathbb{E}\|Y - g_t\|^2$ if and only if $g_t = \mathbb{E}[Y \mid \mathcal{G}]$ a.s.*
- *Every minimiser agrees a.s. with a ξ-blind function of $(o_t, a_{t:t+j-1})$.*
- *Every g with $\mathbb{P}(g_t \ne \bar{g}_t) > 0$ is strictly beaten by $\bar{g}_t$, which is itself such a function.*

Under (A1)–(A2), $\bar{g}_t = \bar{g}(o_t, a_{t:t+j-1})$ for the explicit ξ-average $\bar{g}(o, a) := \int g(q(f(o), x), a)\,P_\xi(dx)$ (set to 0 where the integral does not converge absolutely). For the clipped value loss equation 1 ($p = 1$): if $v_t, R_t \in L^2$ and ξ_t is independent of $\sigma(v_t, R_t) \vee \mathcal{G}$, then $\mathbb{E}\,\ell_\epsilon(g_t; v_t, R_t) \ge \mathbb{E}\,\ell_\epsilon(\bar{g}_t; v_t, R_t)$.

(b) Where equation 2 holds. *Let (A1)–(A2) hold. For every $\mathcal{F}^{\mathrm{en}}$-measurable $Y \in L^2$, ξ_t is independent of $\sigma(Y) \vee \mathcal{G}$, and hence equation 2 holds. This covers:*

(b1) rewards r_{t+i}, $i \ge 0$;
(b2) Monte-Carlo returns $\sum_{l<L}\gamma^l r_{t+l}$ with an $\mathcal{F}^{\mathrm{en}}$-measurable horizon L;
(b3) the code's return R_t with a ξ-blind critic;
(b4) targets $\bar{\phi}(o^+_{t+i})$, $i \ge 1$, with $\bar{\phi}$ ξ-blind.

If moreover the ξ_u are i.i.d. ($\rho = 0$), then ξ_t is independent of $\sigma(Y) \vee \mathcal{G}$ also for:

(b3′) R_t with any *critic;*
(b4′) $\bar{\phi}(o^+_{t+i})$, $i \ge 1$, with any *measurable $\bar{\phi}$;*

provided these are square integrable.

(c) Invariance (MLP networks, implemented algorithm). *Let (A1), (N) and (I) hold for every client. Suppose every client starts in Θ_0, with Adam moments vanishing on the coordinates of Ξ. Then along the population recursion of Definition 2:*

- *every client stays in Θ_0 at every step, with Adam moments vanishing on Ξ;*
- *the behaviour policy of every round is endogenous, so (A2) holds throughout.*

This holds for every choice of:

- *the loss weights (zero included) and the horizon H;*
- *the routing (joint, critic, detached, actor-shaped, phased);*
- *the EMA rate and schedule (per round or per step);*
- *the clipping thresholds and the step-size schedule;*
- *the sync period E, the averaged subset $\mathcal{M}$ and the weights ω_k;*
- *whether the normaliser statistics are pooled.*

It also holds for raw PPO, with Ξ the first-layer ξ-blocks of the actor and the critic. Moreover, at every $\theta \in \Theta_0$ and for data collected by a policy in Θ_0, the Ξ-block of every minibatch gradient has conditional mean zero: $\mathbb{E}[\widehat{\nabla}_{\Xi}\mathcal{L}_I(\theta;\mathcal{D}) \mid \mathcal{F}^{\mathrm{en}}] = 0$.

(d) The two self-referential channels. *Let (A1)–(A2) hold with Gaussian AR(1) noise, let $o = (s, \xi)$, and let $j \geq 1$.*

(d1) Target. *If $\bar{\phi}(o) = \bar{\phi}_s(s) + \bar{E}\xi$ with $\bar{E} \in \mathbb{R}^{k\times m}$ and $\bar{\phi}_s(s^+_{t+j}) \in L^2$, then $\mathbb{E}[\bar{\phi}(o^+_{t+j}) \mid o_t, a_{t:t+j-1}] = \mathbb{E}[\bar{\phi}_s(s^+_{t+j}) \mid s_t, a_{t:t+j-1}] + \rho^j \bar{E}\xi_t$.*

(d2) Critic. *If $V(o) = u(s) + v^\top \xi$ with $r_{t+l}, u(s_{t+l}), u(s^+_{t+l}) \in L^2$, then the code's return satisfies $\mathbb{E}[R_t \mid o_t] = \mathbb{E}[R_t \mid s_t] + \mathbb{E}[c_V(L_t, \tau_t) \mid s_t]\, v^\top \xi_t$, where*

$$c_V(L,\tau) = \frac{\gamma(1-\lambda)\rho\left(1-(\gamma\lambda\rho)^{L-1}\right)}{1-\gamma\lambda\rho} + (1-\tau)\,\gamma^L\lambda^{L-1}\rho^L \ \in [0, \gamma\rho],$$

and $c_V(L,\tau) \to c_V := \gamma(1-\lambda)\rho/(1-\gamma\lambda\rho)$ as $L \to \infty$. Consider an infinite-horizon stream without episode boundaries, with $\sup_u \mathbb{E}\, r_u^2 < \infty$ and $\sup_u \mathbb{E}\, u(s_u)^2 < \infty$. Its untruncated return $G^\lambda_t := \sum_{l\geq 0}(\gamma\lambda)^l r_{t+l} + \gamma(1-\lambda)\sum_{l\geq 1}(\gamma\lambda)^{l-1} V(o_{t+l})$ satisfies $\mathbb{E}[G^\lambda_t \mid o_t] = \mathbb{E}[G^\lambda_t \mid s_t] + c_V\, v^\top \xi_t$ and $\mathrm{Cov}(G^\lambda_t, \xi_t) = c_V\, v$.

(d3) Nonlinear content. *Let $f : \mathcal{S} \times \mathbb{R}^m \to \mathbb{R}^p$ be measurable with $\mathbb{E}\int \|f(s^+_{t+j}, x)\|^2 \mathcal{N}_m(dx) < \infty$, and let $\tilde{f}(s,x) := f(s,x) - \int f(s,x')\,\mathcal{N}_m(dx')$ be its ξ-part. Then*

$$\begin{aligned}&\left\|\mathbb{E}[f(o^+_{t+j}) \mid o_t, a_{t:t+j-1}] - \mathbb{E}[f(o^+_{t+j}) \mid s_t, a_{t:t+j-1}]\right\|_{L^2}\\ &\qquad \leq \rho^j \left(\mathbb{E}\int \|\tilde{f}(s^+_{t+j}, x)\|^2 \mathcal{N}_m(dx)\right)^{1/2},\end{aligned}$$

with equality when $\tilde{f}(s,x) = \bar{E}x$. Likewise, for G^λ_t with any critic V, provided $\sup_u \mathbb{E}\, r_u^2 < \infty$ and $\sup_u \mathbb{E}\, V(o_u)^2 < \infty$,

$$\|\mathbb{E}[G^\lambda_t \mid o_t] - \mathbb{E}[G^\lambda_t \mid s_t]\|_{L^2} \leq c_V \sup_{l\geq 1} \left(\mathbb{E}\int \tilde{V}(s_{t+l}, x)^2\, \mathcal{N}_m(dx)\right)^{1/2}.$$

All these ξ_t-terms vanish when $\rho = 0$.

Informal. No regression loss (prediction, reward or value) pays for noise, and the expected dynamics never leave the noise-free parameters once they are there. Among the regression *targets*, and under an endogenous behaviour policy, correlated noise can enter only through the two self-referential ones, the EMA target and the bootstrap critic. In L^2 they pass on at most a fraction ρ^j and c_V of their own noise, respectively; with the code's truncated return, the critic's fraction is $c_V(L_t, \tau_t) \leq \gamma\rho$. Off Θ_0 the implemented algorithm has further channels that do not need $\rho > 0$: the value-clip anchor $v_t = V_{\mathrm{old}}(o_t)$ (it weakens the value channel's restoring force but does not reverse it), the VICReg variance and covariance terms (whose curvature along the noise block can be negative), and actor feedback when A2 fails. None of these affects the invariance of Θ_0.

LEMMAS

Lemma 1 (conditional expectations)**.** *Let $\mathcal{G}$ be a σ-algebra.*

(i) Let Y be integrable and let X be a random element independent of $\sigma(Y) \vee \mathcal{G}$. Then $\mathbb{E}[Y \mid \mathcal{G} \vee \sigma(X)] = \mathbb{E}[Y \mid \mathcal{G}]$.

(ii) (Substitution.) *Let X be $\mathcal{G}$-measurable with values in a measurable space $\mathcal{X}$, let Z be independent of $\mathcal{G}$ with values in a measurable space $\mathcal{Z}$, and let h be jointly measurable with $\mathbb{E}\|h(X,Z)\| < \infty$. Then $H(x) := \mathbb{E}\, h(x,Z)$ is defined for P_X-a.e. x, it is measurable, and $\mathbb{E}[h(X,Z) \mid \mathcal{G}] = H(X)$ a.s.*

(iii) Let δ be $\mathcal{G}$-measurable with values in $\mathbb{R}^p$, and let $x \in \mathbb{R}^m$ satisfy $\mathbb{E}\|x\| < \infty$, $\mathbb{E}\|\delta\|\|x\| < \infty$ and $\mathbb{E}[x \mid \mathcal{G}] = 0$. Then $\mathbb{E}[\delta x^\top \mathbf{1}_A] = 0$ for every $A \in \mathcal{G}$. In particular $\mathbb{E}[\delta x^\top] = 0$ and $\mathbb{E}[\delta x^\top \mid \mathcal{G}] = 0$.

Proof. We use the defining property of conditional expectation, extended from indicators to bounded measurable multipliers by linearity and uniform approximation by simple functions. Vector-valued statements are proved coordinatewise.

(i) Sets $A \cap B$ with $A \in \mathcal{G}$ and $B \in \sigma(X)$ form a π-system that contains Ω and generates $\mathcal{G} \vee \sigma(X)$. Since $\mathbf{1}_B$ is independent of the $\sigma(Y) \vee \mathcal{G}$-measurable integrable variables $Y\mathbf{1}_A$ and $\mathbb{E}[Y \mid \mathcal{G}]\mathbf{1}_A$,

$$\mathbb{E}[Y\mathbf{1}_{A\cap B}] = \mathbb{E}[Y\mathbf{1}_A]\,\mathbb{P}(B) = \mathbb{E}\big[\mathbb{E}[Y \mid \mathcal{G}]\mathbf{1}_A\big]\mathbb{P}(B) = \mathbb{E}\big[\mathbb{E}[Y \mid \mathcal{G}]\mathbf{1}_{A\cap B}\big].$$

The sets C with $\mathbb{E}[Y\mathbf{1}_C] = \mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}]\mathbf{1}_C]$ form a Dynkin system, by linearity and dominated convergence. By the π–λ theorem this system contains $\mathcal{G} \vee \sigma(X)$. Since $\mathbb{E}[Y \mid \mathcal{G}]$ is $\mathcal{G} \vee \sigma(X)$-measurable, the claim follows.

(ii) First let $h \geq 0$ (with values in $[0,\infty]$). By Tonelli, H is measurable. Fix $A \in \mathcal{G}$. The pair $(X, \mathbf{1}_A)$ is $\mathcal{G}$-measurable and Z is independent of $\mathcal{G}$, so the law of $(X, \mathbf{1}_A, Z)$ is $P_{(X,\mathbf{1}_A)} \otimes P_Z$. Tonelli then gives $\mathbb{E}[h(X,Z)\mathbf{1}_A] = \iint h(x,z)\, a\, P_Z(dz)\, P_{(X,\mathbf{1}_A)}(dx,da) = \mathbb{E}[H(X)\mathbf{1}_A]$. For integrable h, apply this to the positive and negative parts of each coordinate. Since $\mathbb{E}H^{\pm}(X) = \mathbb{E}h^{\pm}(X,Z) < \infty$, the functions $H^{\pm}$ are finite P_X-a.e.

(iii) Let $M > 0$ and $\delta_M := \delta\mathbf{1}\{\|\delta\| \leq M\}$. The multiplier $\delta_M\mathbf{1}_A$ is bounded and $\mathcal{G}$-measurable, so $\mathbb{E}[\delta_M\mathbf{1}_A x^\top] = \mathbb{E}[\delta_M\mathbf{1}_A\mathbb{E}[x \mid \mathcal{G}]^\top] = 0$. Let $M \to \infty$. Dominated convergence applies, with the integrable bound $\|\delta\|\|x\|$. □

Lemma 2 (decoupling)**.** *Assume (A2) for every client, with deterministic policies π_t (as in the population recursion). Then every $s_t^k, s_t^{k,+}, a_t^k, r_t^k, b_t^k$ is $\mathcal{F}^{\mathrm{en}}$-measurable. Under (A1), the exogenous processes are therefore independent of all endogenous processes. In particular each ξ_t is independent of $\sigma(s_{0:\infty}, s^+_{1:\infty}, a_{0:\infty}, r_{0:\infty}, b_{0:\infty})$.*

Proof. The proof is by induction on t. $s_0 = F_0(w_0)$ is $\mathcal{F}^{\mathrm{en}}$-measurable. If h_t^{en} is, then so is $a_t = \pi_t(h_t^{\mathrm{en}}, \eta_t)$, by (A2), and so is $(r_t, b_t, s^+_{t+1}, s_{t+1}) = F(s_t, a_t, w_{t+1})$. Independence then follows from (A1). □

This is the continuous-state counterpart of the decoupling of state occupancies under endogenous policies in the tabular ExoMDP (Efroni et al., 2022a, Lemma B.1). Without (A2) the induction fails at the first action that reads ξ_t. For example, if raw PPO's actor has $W_{1,\xi} \neq 0$, the later rewards depend on ξ_t. This is the "entanglement" that part (c) rules out on Θ_0.

Lemma 3 (the code's return)**.** *Suppose $o_{t+1} = o^+_{t+1}$ whenever $\bar{b}_t = 0$, as in the code. Then, for every $t < n$,*

$$R_t = \sum_{l=0}^{L_t-1} (\gamma\lambda)^l r_{t+l} + \gamma(1-\lambda) \sum_{l=1}^{L_t-1} (\gamma\lambda)^{l-1} V(o_{t+l}) + \gamma(\gamma\lambda)^{L_t-1}(1-\tau_t)\, V(o^+_{t+L_t}). \tag{4}$$

In particular, R_t does not depend on $v_t = V(o_t)$. It involves the critic only at $o_{t+1}, \ldots, o_{t+L_t-1}$ and at $o^+_{t+L_t}$.

Proof. From $v_t = V(o_t)$,

$$R_t = A_t + V(o_t) = r_t + \gamma(1 - \mathrm{term}_t)V(o^+_{t+1}) + \gamma\lambda(1 - \bar{b}_t)A_{t+1}.$$

The proof is by backward induction on t.

- If $\bar{b}_t = 1$ or $t = n-1$ (using $A_n = 0$), then $L_t = 1$, $\tau_t = \mathrm{term}_t$ and $R_t = r_t + \gamma(1-\tau_t)V(o^+_{t+1})$, which is equation 4.
- Otherwise $\mathrm{term}_t = 0$ and $o^+_{t+1} = o_{t+1}$. Using $A_{t+1} = R_{t+1} - V(o_{t+1})$,

  $$R_t = r_t + \gamma(1-\lambda)V(o_{t+1}) + \gamma\lambda R_{t+1},$$

  and $L_t = L_{t+1} + 1$, $\tau_t = \tau_{t+1}$. Substituting equation 4 for R_{t+1} and shifting the indices gives equation 4 for R_t. □

Lemma 4 (the clipped value loss is convex)**.** *For $\epsilon > 0$ and $v_o, R \in \mathbb{R}$, the map $v \mapsto \ell_\epsilon(v; v_o, R)$ of equation 1 is convex. Its right derivative satisfies $|\partial^+_v \ell_\epsilon| \le 2|v - R|$.*

Proof. Put $u := v - v_o$ and $c := R - v_o$, so that $\ell_\epsilon = \max\{(u-c)^2, (\mathrm{clip}(u, -\epsilon, \epsilon) - c)^2\}$.

- *Case $|c| \le \epsilon$.* On $|u| \le \epsilon$ the two terms agree. For $u > \epsilon$ we have $u - c > \epsilon - c \ge 0$; for $u < -\epsilon$ we have $u - c < -\epsilon - c \le 0$. In both cases the unclipped term dominates, so $\ell_\epsilon = (u-c)^2$.
- *Case $c > \epsilon$.* For $u \le \epsilon$ the same comparison gives $\ell_\epsilon = (u-c)^2$. For $u > \epsilon$, $(u-c)^2 \le (\epsilon - c)^2$ if and only if $\epsilon \le u \le 2c - \epsilon$. Hence $\ell_\epsilon = (\epsilon - c)^2$ on $[\epsilon, 2c - \epsilon]$ and $(u-c)^2$ elsewhere. This function is continuous. Its derivative is $2(u-c)$ on $(-\infty, \epsilon)$, with left limit $2(\epsilon - c) < 0$; it is 0 on $(\epsilon, 2c - \epsilon)$; and it is $2(u-c)$ on $(2c - \epsilon, \infty)$, starting at $2(c - \epsilon) > 0$. The derivative is non-decreasing, so ℓ_ϵ is convex.
- *Case $c < -\epsilon$.* This follows from the previous case by $(u, c) \mapsto (-u, -c)$, since clip is odd.

In every case the right derivative is either $2(u-c) = 2(v-R)$ or 0. □

Lemma 5 (first-layer factorisation of reverse-mode gradients)**.** *Let a network read its input only through an affine first layer $h = Wx + c_1$, with $W = [W_{1,s}, W_{1,\xi}]$. Let a scalar loss $\mathcal{L}$ be computed by a finite computation graph in which the network is evaluated at inputs $x_1, \dots, x_U$. Let δ_u be the adjoint that reverse mode assigns to the node $h_u = Wx_u + c_1$. Then*

$$\widehat{\nabla}_W \mathcal{L} = \sum_u \delta_u x_u^\top, \qquad \widehat{\nabla}_{c_1} \mathcal{L} = \sum_u \delta_u.$$

Each δ_u is a fixed measurable function of three things: the parameters; the values of all first-layer nodes of all networks in the graph; and the inputs of the graph that are not observations (actions, rewards, flags, stored values and log-probabilities, returns and advantages).

Proof. Reverse mode gives each node an adjoint: the sum over its children of the child's adjoint times the local partial derivative, evaluated at the forward values. At kinks it uses the implementation's fixed convention, which is a Borel function of the forward values. The parameters W, c_1 are parents only of the nodes h_u. Their local partials are $\partial h_u / \partial W_{ab} = e_a\, x_u[b]$ and $\partial h_u / \partial c_1 = \boldsymbol{I}$, which gives the two formulas.

Observations enter the graph only through first layers. Every node on a path from h_u to $\mathcal{L}$ is therefore a descendant of first-layer nodes and of non-observation inputs only. So its forward value, and the local partials evaluated there, are measurable functions of the listed quantities, and so is δ_u. □

Lemma 6 (zero-preserving updates)**.** *Let B be a block of coordinates, and suppose the parameters θ, the gradient g and the optimiser moments all vanish on B. Then each of the following updates leaves the parameters and the moments zero on B:*

- *a gradient step $\theta - \alpha g$;*
- *gradient-norm clipping $g \mapsto \min\{1, C/(\|g\| + 10^{-6})\}\, g$ applied to any group of coordinates;*
- *Adam (PyTorch's `Adam`, no weight decay, $\epsilon_{\mathrm{A}} > 0$).*

EMA $\tau\bar{\theta} + (1-\tau)\theta$ and averaging $\sum_k \omega_k \theta^k$ of vectors that vanish on B also vanish on B.

Proof. Clipping multiplies a group by a scalar. PyTorch's Adam updates the moments and parameters as

$$m \leftarrow \beta_1 m + (1-\beta_1) g, \qquad v \leftarrow \beta_2 v + (1-\beta_2) g \odot g,$$
$$\theta \leftarrow \theta - \frac{\alpha}{1-\beta_1^s}\, m \oslash \Big(\sqrt{v/(1-\beta_2^s)} + \epsilon_{\mathrm{A}}\Big).$$

On B these give $m_B = 0$, $v_B = 0$, and a step of $0/\epsilon_{\mathrm{A}} = 0$. The remaining maps are linear. □

Lemma 7 (normalisers)**.** *Let (A1) hold, and let $\theta \in \Theta_0$, with data collected by a policy in Θ_0. In (ii), all data so far must have been collected by policies in Θ_0.*

(i) Frozen, off-centre. *If κ_ξ is fixed with $\mathbb{E}\|\kappa_\xi(\xi_t)\| < \infty$ and $\mathbb{E}\,\kappa_\xi(\xi_t) = \bar{\mu}$ for all t, and (I) holds, then, for every network in Ξ that receives gradients,*

$$\mathbb{E}\big[\widehat{\nabla}_{W_{1,\xi}} \mathcal{L}_I\big] = \mathbb{E}\big[\widehat{\nabla}_{c_1} \mathcal{L}_I\big]\, \bar{\mu}^\top.$$

Put $G := \bar{g}_c \bar{\mu}^\top$ with $\bar{g}_c := \mathbb{E}[\widehat{\nabla}_{c_1} \mathcal{L}_I]$, and let $\varsigma \in (0,1]$ be the clipping factor. From θ, the next population iterate has the following Ξ-block:

- *under SGD, $-\alpha\varsigma G$;*
- *under Adam, with zero moments on Ξ and step count s of the weight tensor, $-\alpha a_s (\varsigma G) \oslash (\varsigma|G| + \epsilon_{\mathrm{A}}/b_s)$ coordinatewise, where $b_s := ((1-\beta_2)/(1-\beta_2^s))^{1/2}$ and $a_s := (1-\beta_1)/((1-\beta_1^s)\, b_s)$.*

Both are nonzero wherever $G \neq 0$. As $\epsilon_{\mathrm{A}}/(\varsigma b_s |G_{ab}|) \to 0$, the Adam entry tends to $-a_s \alpha \operatorname{sign}(G_{ab})$, with $a_s \in [0.48, 3.17]$ for the code's $(\beta_1, \beta_2) = (0.9, 0.999)$. So its size does not depend on how small $\bar{\mu}$ is.

(ii) Running. *The code's running normaliser (`RunningMeanStd` started at mean 0 and variance 1, followed by clipping to $[-10, 10]$) is covered here, whether it is per client, pooled at synchronisation, or a single global object. Suppose the joint law of ($\mathcal{F}^{\mathrm{en}}$-measurable variables, $\xi^{1:K}$) is invariant under $\xi^{1:K} \mapsto -\xi^{1:K}$, which holds under (A1) for centred Gaussian noise. Then $x_{s,t}$ is $\mathcal{F}^{\mathrm{en}}$-measurable and $\mathbb{E}[x_{\xi,t} \mid \mathcal{F}^{\mathrm{en}}] = 0$ for every t.*

Proof. (i) By Lemma 5, $\widehat{\nabla}_{W_{1,\xi}} \mathcal{L}_I = \sum_u \delta_u x_{\xi,u}^\top$ and $\widehat{\nabla}_{c_1} \mathcal{L}_I = \sum_u \delta_u$.

On Θ_0 every first-layer node equals $W_{1,s} x_{s,u'} + c_1$. By Lemma 2, the observations' endogenous parts, the stored quantities (which come from ξ-blind networks), the returns (Lemma 3 with a ξ-blind critic) and the target outputs are all $\mathcal{F}^{\mathrm{en}}$-measurable. So δ_u is $\mathcal{F}^{\mathrm{en}}$-measurable.

Write $x_{\xi,u} = \kappa_\xi(\xi_{t_u})$, with an $\mathcal{F}^{\mathrm{en}}$-measurable time index t_u. By (A1) and Lemma 1(ii), $\mathbb{E}[x_{\xi,u} \mid \mathcal{F}^{\mathrm{en}}] = \sum_t \mathbf{1}\{t_u = t\}\, \mathbb{E}\,\kappa_\xi(\xi_t) = \bar{\mu}$. Then Lemma 1(iii), applied to $x_{\xi,u} - \bar{\mu}$, gives $\mathbb{E}[\delta_u x_{\xi,u}^\top] = \mathbb{E}[\delta_u] \bar{\mu}^\top$. Summing over u proves the identity.

The two update formulas follow from $m = (1-\beta_1)\varsigma G$, $v = (1-\beta_2)\varsigma^2 G \odot G$ on Ξ and the bias corrections $1-\beta_1^s$ and $1-\beta_2^s$. The range of a_s over $s \geq 1$ is elementary to evaluate: the minimum 0.481 is attained at $s = 12$, and the supremum $\sqrt{1000}/10 = 3.162$ is the limit $s \to \infty$.

(ii) For one coordinate, one sample y and running count ν, the update is

$$\mu \leftarrow \mu + \frac{y-\mu}{\nu+1}, \qquad \sigma^2 \leftarrow \frac{\nu\sigma^2 + (y-\mu)^2\, \nu/(\nu+1)}{\nu+1}, \qquad \nu \leftarrow \nu+1,$$

and the output is $\mathrm{clip}((y-\mu)/\sqrt{\sigma^2 + 10^{-8}}, -10, 10)$, computed with the updated μ and σ^2.

The counts (numbers of processed observations, resets included, and the order in which clients step) are $\mathcal{F}^{\mathrm{en}}$-measurable, and an ξ-coordinate's statistics depend only on that coordinate's samples. Starting from $\mu = 0$, induction shows that, for fixed counts, μ is odd and σ^2 is even as functions of the ξ-path. Pooling ($\mu = \sum_i \nu_i \mu_i / \sum_i \nu_i$, $\sigma^2 = \sum_i \nu_i (\sigma_i^2 + (\mu_i - \mu)^2) / \sum_i \nu_i$) preserves these parities. Hence $x_{\xi,t} = \Gamma(C, \xi^{1:K})$, with C $\mathcal{F}^{\mathrm{en}}$-measurable, $|\Gamma| \leq 10$, and $\Gamma(C, -\xi) = -\Gamma(C, \xi)$. The s-coordinates depend only on endogenous samples and counts.

For $A \in \mathcal{F}^{\mathrm{en}}$, the symmetry gives $\mathbb{E}[\Gamma(C,\xi)\mathbf{1}_A] = \mathbb{E}[\Gamma(C,-\xi)\mathbf{1}_A] = -\mathbb{E}[\Gamma(C,\xi)\mathbf{1}_A]$, so this expectation is 0. □

Lemma 8 (Gaussian contraction)**.** *Let $r \in [0,1)$ and $Z \sim \mathcal{N}_m$. For $h \in L^2(\mathcal{N}_m)$ put $(T_r h)(y) := \mathbb{E}\, h(ry + \sqrt{1-r^2}\, Z)$, and $T_1 h := h$. Then:*

- *$T_r h(y)$ is finite for every y, $\|T_r h\| \le \|h\|$, and $\int T_r h \, d\mathcal{N}_m = \int h \, d\mathcal{N}_m$;*
- *if $\int h \, d\mathcal{N}_m = 0$, then $\|T_r h\| \le r\|h\|$, with equality when h is linear.*

This is the classical fact that the maximal correlation of a Gaussian pair with cross-covariance $r\boldsymbol{I}_m$ is r (Gebelein, 1941; Lancaster, 1957). We give a self-contained proof, so nothing below depends on these citations.

Proof. Let φ be the density of $\mathcal{N}_m$, and let p_y be the density of $\mathcal{N}(ry, (1-r^2)\boldsymbol{I}_m)$.

Step 1 (finiteness). The exponent of p_y^2/φ is $-\|x-ry\|^2/(1-r^2) + \|x\|^2/2$, a quadratic in x with leading coefficient $-1/(1-r^2) + 1/2 < 0$. So $\ell_y := p_y/\varphi \in L^2(\mathcal{N}_m)$, and $|T_r h(y)| \le \int |h| \ell_y \, d\mathcal{N}_m \le \|h\|\, \|\ell_y\| < \infty$.

Step 2 (symmetry). Let $X \sim \mathcal{N}_m$ and $Y := rX + \sqrt{1-r^2} Z'$ with $Z' \sim \mathcal{N}_m$ independent of X. The pair (X,Y) is centred Gaussian with covariance $\left(\begin{smallmatrix} \boldsymbol{I} & r\boldsymbol{I} \\ r\boldsymbol{I} & \boldsymbol{I} \end{smallmatrix}\right)$, so $(X,Y) \overset{d}{=} (Y,X)$. By Lemma 1(ii), $\mathbb{E}[h(Y) \mid X] = T_r h(X)$. For $g, h \in L^2(\mathcal{N}_m)$, $g(X)h(Y)$ is integrable by Cauchy–Schwarz, and

$$\langle g, T_r h \rangle = \mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(Y)h(X)] = \langle T_r g, h \rangle.$$

Moreover $\|T_r h\|^2 = \mathbb{E}[(\mathbb{E}[h(Y) \mid X])^2] \le \mathbb{E}\, h(Y)^2 = \|h\|^2$ and $\int T_r h \, d\mathcal{N}_m = \mathbb{E}\, h(Y) = \int h \, d\mathcal{N}_m$. For linear $h(x) = a^\top x$, $T_r h = r\, h$.

Step 3 (semigroup). For $r, s \in [0,1]$, $T_r T_s = T_{rs}$. Indeed, with Z, Z' i.i.d. $\mathcal{N}_m$, $W := s\sqrt{1-r^2}\, Z + \sqrt{1-s^2}\, Z' \sim \mathcal{N}(0, (1-r^2 s^2)\boldsymbol{I}_m)$, and $T_r(T_s h)(y) = \mathbb{E}\, h\big(s(ry + \sqrt{1-r^2}Z) + \sqrt{1-s^2}Z'\big) = \mathbb{E}\, h(rsy + W) = T_{rs} h(y)$. Fubini is justified by Step 1 applied to $|h|$.

Step 4 (χ^2 bound). Let $\int h \, d\mathcal{N}_m = 0$ and $u \in [0,1)$. Let (X, Y_u) be defined as in Step 2 with $r = u$, with joint density p_u, and put $L_u := p_u/(\varphi \otimes \varphi)$. For one coordinate pair, $p_u^2/(\varphi\otimes\varphi) = (2\pi)^{-1}(1-u^2)^{-1} e^{-(x,y)M(x,y)^\top}$, where $M = (1-u^2)^{-1}\left(\begin{smallmatrix} (1+u^2)/2 & -u \\ -u & (1+u^2)/2 \end{smallmatrix}\right) \succ 0$ and $\det M = 1/4$. So its integral is $(1-u^2)^{-1}$, and, by independence across the m coordinates, $\mathbb{E}_{\mathcal{N}_m \otimes \mathcal{N}_m} L_u^2 = (1-u^2)^{-m}$. With X', Y' i.i.d. $\mathcal{N}_m$ and $\mathbb{E}[h(X')h(Y')] = (\int h)^2 = 0$,

$$\langle h, T_u h \rangle = \mathbb{E}[h(X)h(Y_u)] = \mathbb{E}\big[h(X')h(Y')(L_u(X',Y') - 1)\big] \le \|h\|^2 \, \kappa(u),$$

where $\kappa(u) := \big((1-u^2)^{-m} - 1\big)^{1/2}$, by Cauchy–Schwarz and $\mathbb{E} L_u = 1$. The map $x \mapsto (1-x)^{-m} - 1$ is convex on $[0,1)$ and vanishes at 0, so it is at most $2(2^m - 1)x$ on $[0, 1/2]$. Hence $\kappa(u) \le C_m u$ for $u^2 \le 1/2$, with $C_m := (2(2^m - 1))^{1/2}$.

Step 5 (log-convexity). Let $\int h \, d\mathcal{N}_m = 0$, $h \neq 0$, $r \in (0,1)$, and $a_n := \|T_{r^n} h\|^2$ for $n \ge 0$. By Steps 2–3, for $n \ge 1$,

$$a_n = \langle h, T_{r^{2n}} h \rangle = \langle T_{r^{n-1}} h, T_{r^{n+1}} h \rangle \le \sqrt{a_{n-1} a_{n+1}}.$$

If $a_1 = 0$ we are done. Otherwise every $a_n > 0$: if $n_0 \ge 2$ were the first index with $a_{n_0} = 0$, then $a_{n_0-1}^2 \le a_{n_0-2} a_{n_0} = 0$, a contradiction. So $q_n := a_{n+1}/a_n$ is non-decreasing, and $a_n = a_0 \prod_{i<n} q_i \ge a_0 q_0^n$. By Step 4 with $u = r^{2n}$, $a_n \le \|h\|^2 C_m r^{2n}$ once $r^{4n} \le 1/2$. Hence $q_0 \le (\|h\|^2 C_m / a_0)^{1/n} r^2$ for all large n, and letting $n \to \infty$ gives $q_0 \le r^2$. That is, $\|T_r h\|^2 = a_1 = q_0 a_0 \le r^2 \|h\|^2$. For $r = 0$, $T_0 h = \int h \, d\mathcal{N}_m = 0$. □

PROOF OF PROPOSITION 1

Proof of (a). Write $S := (s_t, a_{t:t+j-1})$ and $\mathcal{H} := \mathcal{G} \vee \sigma(\xi_t)$. Since $o_t = q(s_t, \xi_t)$, the variable g_t is $\mathcal{H}$-measurable, and so is $g_t - \bar{g}_t$.

The identity. Expand

$$\|Y - g_t\|^2 = \|Y - \bar{g}_t\|^2 + \|g_t - \bar{g}_t\|^2 - 2\langle Y - \bar{g}_t,\, g_t - \bar{g}_t\rangle,$$

where all terms are integrable. Split $Y - \bar{g}_t = (Y - \mathbb{E}[Y \mid \mathcal{H}]) + (\mathbb{E}[Y \mid \mathcal{H}] - \bar{g}_t)$.

- The first part has conditional mean zero given $\mathcal{H}$. By Lemma 1(iii) (with $\delta = g_t - \bar{g}_t$), its inner product with $g_t - \bar{g}_t$ has expectation 0.
- By equation 2, the second part is $\mathbb{E}[Y \mid \mathcal{G}] - \bar{g}_t$, which is $\mathcal{G}$-measurable, while $\mathbb{E}[g_t - \bar{g}_t \mid \mathcal{G}] = 0$. By Lemma 1(iii) again, its contribution vanishes too.

This proves equation 3.

Minimisers. $\bar{g}_t$ is $\mathcal{G}$-measurable, and $Y - \mathbb{E}[Y \mid \mathcal{G}]$ has conditional mean zero given $\mathcal{G}$. So Lemma 1(iii) gives $\mathbb{E}\|Y - \bar{g}_t\|^2 = \mathbb{E}\|Y - \mathbb{E}[Y \mid \mathcal{G}]\|^2 + \mathbb{E}\|\mathbb{E}[Y \mid \mathcal{G}] - \bar{g}_t\|^2$. With equation 3,

$$\mathbb{E}\|Y - g_t\|^2 - \mathbb{E}\|Y - \mathbb{E}[Y \mid \mathcal{G}]\|^2 = \mathbb{E}\|\mathbb{E}[Y \mid \mathcal{G}] - \bar{g}_t\|^2 + \mathbb{E}\|g_t - \bar{g}_t\|^2.$$

The right side vanishes if and only if $g_t = \bar{g}_t = \mathbb{E}[Y \mid \mathcal{G}]$ a.s. Conversely, $g_t = \mathbb{E}[Y \mid \mathcal{G}]$ implies $\bar{g}_t = g_t$. By the Doob–Dynkin lemma, $\mathbb{E}[Y \mid \mathcal{G}] = \psi(S)$ and $\bar{g}_t = \psi'(S)$ for measurable ψ, ψ'. So both are the values at $(o_t, a_{t:t+j-1})$ of the ξ-blind functions $\psi(f(\cdot), \cdot)$ and $\psi'(f(\cdot), \cdot)$, and the minimum is attained. If $\mathbb{P}(g_t \neq \bar{g}_t) > 0$, the last term of equation 3 is positive.

The explicit ξ-average. Under (A1)–(A2), Lemma 2 makes ξ_t independent of $\mathcal{G}$. Since g_t is a measurable function of (S, ξ_t) and ξ_t has law P_ξ, Lemma 1(ii) gives $\bar{g}_t = \int g(q(s_t, x), a_{t:t+j-1})\, P_\xi(dx) = \bar{g}(o_t, a_{t:t+j-1})$, using $f(o_t) = s_t$.

Clipped value loss. Let $\mathcal{G}' := \sigma(v_t, R_t) \vee \mathcal{G} \supseteq \mathcal{G}$. Since ξ_t is independent of $\mathcal{G}'$, Lemma 1(ii), applied with $\mathcal{G}'$ and with $\mathcal{G}$, gives $\mathbb{E}[g_t \mid \mathcal{G}'] = H(S) = \mathbb{E}[g_t \mid \mathcal{G}] = \bar{g}_t$, where $H(s, a) := \int g(q(s, x), a)\, P_{\xi_t}(dx)$. By Lemma 4, the right derivative $d_t := \partial_v^+ \ell_\epsilon(\bar{g}_t; v_t, R_t)$ is a subgradient, so pointwise

$$\ell_\epsilon(g_t; v_t, R_t) \geq \ell_\epsilon(\bar{g}_t; v_t, R_t) + d_t(g_t - \bar{g}_t).$$

Here d_t is $\mathcal{G}'$-measurable, being a limit of difference quotients, with $|d_t| \leq 2|\bar{g}_t - R_t| \in L^2$. Also $\ell_\epsilon(\bar{g}_t) \leq \max\{(\bar{g}_t - R_t)^2, (|v_t - R_t| + \epsilon)^2\}$ is integrable. By Lemma 1(iii), $\mathbb{E}[d_t(g_t - \bar{g}_t)] = 0$, and taking expectations gives the claim. □

Proof of (b). By Lemma 2, $\mathcal{G} \subseteq \mathcal{F}^{\mathrm{en}}$ and ξ_t is independent of $\mathcal{F}^{\mathrm{en}}$. So for $\mathcal{F}^{\mathrm{en}}$-measurable Y, ξ_t is independent of $\sigma(Y) \vee \mathcal{G}$, and Lemma 1(i) gives equation 2.

The listed targets are $\mathcal{F}^{\mathrm{en}}$-measurable:

- (b1) and (b2) are immediate.
- (b3): by Lemma 3, R_t is a function of rewards, flags, L_t, τ_t and the values $\hat{V}(s_{t+l})$ and $\hat{V}(s^+_{t+L_t})$ of the ξ-blind critic $V = \hat{V} \circ f$.
- (b4): $\bar{\phi}(o^+_{t+i}) = \hat{\phi}(s^+_{t+i})$.

Now let the ξ_u be i.i.d. Then $(\mathcal{F}^{\mathrm{en}}, \xi_0, \xi_1, \dots)$ is an independent family. Finite intersections of events from disjoint groups form π-systems, so, by the π–λ theorem, ξ_t is independent of $\sigma(\mathcal{F}^{\mathrm{en}}, \xi_u : u \neq t)$.

- (b4′): for $i \geq 1$, $\bar{\phi}(o^+_{t+i}) = \bar{\phi}(q(s^+_{t+i}, \xi_{t+i}))$ is measurable with respect to that σ-algebra.
- (b3′): by Lemma 3, R_t involves the critic only at o_{t+l}, $l \geq 1$, and at $o^+_{t+L_t}$. These carry ξ_{t+l} and ξ_{t+L_t}, never ξ_t.

In both cases Lemma 1(i) applies as before. □

Proof of (c). We argue by induction over the steps of the recursion, in their order of execution, for all clients simultaneously. The induction hypothesis $\mathrm{H}(r, i)$ is: every client's $\theta^{(r,i)}$ lies in Θ_0, and its Adam moments vanish on Ξ. $\mathrm{H}(0, 0)$ holds by assumption.

Collection is endogenous. Assume $\mathrm{H}(r, 0)$. The actor reads $\phi(x)$ (latent agent) or x (raw PPO) through first layers with $W_{1,\xi} = 0$, so its action law depends on the observation only through $\kappa_s(s_t)$.

Thus (A2) holds for round r with deterministic policies, and by Lemma 2 the round's $s_t, s_t^+, a_t, r_t, b_t$ are $\mathcal{F}^{\mathrm{en}}$-measurable. So are the following:

- $x_{s,t}$;
- the stored v_t and ℓ_t (ξ-blind networks);
- R_t and A_t (Lemma 3);
- the valid-window indicators;
- the target outputs $\bar{\phi}(o^+_{t+j})$ (target in Θ_0);
- the index sets, by definition.

The expected Ξ-block vanishes. Assume $\mathrm{H}(r,i)$ and consider a network in Ξ that receives gradients: the online encoder, or raw PPO's actor and critic. The target receives none. Every first-layer node of every network in $\mathcal{L}_{I_i}$ equals $W_{1,s}x_{s,u} + c_1$, which is $\mathcal{F}^{\mathrm{en}}$-measurable, because every network reading x is in Θ_0 (each node uses that network's own parameters). The other inputs are $\mathcal{F}^{\mathrm{en}}$-measurable by the previous step. By Lemma 5, the Ξ-block of $\widehat{\nabla}\mathcal{L}_{I_i}$ is $\sum_u \delta_u x_{\xi,u}^\top$ with $\mathcal{F}^{\mathrm{en}}$-measurable δ_u. As in the proof of Lemma 7(i), with $\bar{\mu} = 0$ by (N), $\mathbb{E}[x_{\xi,u} \mid \mathcal{F}^{\mathrm{en}}] = 0$. Lemma 1(iii) and (I) then give $\mathbb{E}[\delta_u x_{\xi,u}^\top] = 0$ for each u. So $\bar{g}_i$ vanishes on Ξ. This argument holds term by term, so it covers every loss weight, every H and every routing: routing only removes edges from the graph.

It also covers the code's optional terms: the termination head, inverse dynamics (which evaluates the online encoder at o^+_{t+1} as well), PFO and the FedKL penalty. Each is computed from first-layer nodes and endogenous inputs. The FedProx penalty $\frac{\mu}{2}\|\theta - \theta_{\mathrm{glob}}\|^2$ has Ξ-gradient $\mu(0-0) = 0$ when both models are in Θ_0.

The optimiser. By Lemma 6, clipping and Adam (or SGD) map $\mathrm{H}(r,i)$ to $\mathrm{H}(r,i+1)$. The per-step EMA is a convex combination of two vectors that vanish on Ξ.

End of round and federation. The per-round EMA and the FedAvg replacement $\theta^k|_{\mathcal{M}} \leftarrow \sum_{k'} \omega_{k'} \theta^{k'}|_{\mathcal{M}}$ are linear, and they combine vectors that vanish on Ξ. Modules outside $\mathcal{M}$ and the local optimiser states are unchanged. Pooling the normaliser statistics changes κ_s, which does not matter, and leaves the frozen, centred κ_ξ unchanged. So $\mathrm{H}(r+1,0)$ holds, which completes the induction.

Raw PPO. The same argument applies, with Ξ the first-layer ξ-blocks of the actor and the critic.

Conditional mean zero. Let $\theta \in \Theta_0$, with data from a policy in Θ_0. Then $\mathbb{E}[\delta_u x_{\xi,u}^\top \mathbf{1}_A] = 0$ for every $A \in \mathcal{F}^{\mathrm{en}}$, by Lemma 1(iii). This is the statement $\mathbb{E}[\widehat{\nabla}_\Xi \mathcal{L}_I \mid \mathcal{F}^{\mathrm{en}}] = 0$. □

With the running normaliser, the proof goes through unchanged under the symmetry hypothesis of Lemma 7(ii). It used the normaliser only through two facts: $x_{s,u}$ is $\mathcal{F}^{\mathrm{en}}$-measurable, and $\mathbb{E}[x_{\xi,u} \mid \mathcal{F}^{\mathrm{en}}] = 0$.

Proof of (d). For $l \geq 1$ write $\xi_{t+l} = \rho^l \xi_t + \sigma_l Z_{t,l}$, with $\sigma_l := (1-\rho^{2l})^{1/2}$ and $Z_{t,l} := \sigma_l^{-1} \sum_{i=1}^{l} \rho^{l-i}(1-\rho^2)^{1/2}\epsilon_{t+i} \sim \mathcal{N}_m$ (for $\rho > 0$; for $\rho = 0$ take $Z_{t,l} = \epsilon_{t+l}$). The variable $Z_{t,l}$ is independent of $\mathcal{K}_t := \mathcal{F}^{\mathrm{en}} \vee \sigma(\xi_{0:t})$ by (A1). Hence $\mathbb{E}[\xi_{t+l} \mid \mathcal{K}_t] = \rho^l \xi_t$. With $\mathcal{G} \subseteq \mathcal{F}^{\mathrm{en}}$ (Lemma 2) and $\sigma(o_t) \vee \mathcal{G} = \mathcal{G} \vee \sigma(\xi_t) \subseteq \mathcal{K}_t$, the tower property gives

$$\mathbb{E}[\xi_{t+l} \mid \mathcal{G} \vee \sigma(\xi_t)] = \rho^l \xi_t. \tag{5}$$

(d1) We have $\mathbb{E}[\bar{\phi}(o^+_{t+j}) \mid \mathcal{G} \vee \sigma(\xi_t)] = \mathbb{E}[\bar{\phi}_s(s^+_{t+j}) \mid \mathcal{G}] + \rho^j \bar{E} \xi_t$. The first term uses (b) and Lemma 1(i); the second uses equation 5.

(d2) By Lemma 3, $R_t = R_t^{\mathrm{en}} + \sum_{l=1}^{L_t} w_l\, v^\top \xi_{t+l}$. Here R_t^{en} is the same expression with u in place of V, which is $\mathcal{F}^{\mathrm{en}}$-measurable. The weights are $w_l = \gamma(1-\lambda)(\gamma\lambda)^{l-1}$ for $l < L_t$ and $w_{L_t} = \gamma(\gamma\lambda)^{L_t - 1}(1-\tau_t)$, and they are bounded and $\mathcal{F}^{\mathrm{en}}$-measurable. Write $\sum_{l=1}^{L_t} w_l v^\top \xi_{t+l} = \sum_{l=1}^{n} \mathbf{1}\{l \leq L_t\} w_l v^\top \xi_{t+l}$. Pulling out the bounded $\mathcal{K}_t$-measurable weights and using $\mathbb{E}[\xi_{t+l} \mid$

$\mathcal{K}_t] = \rho^l \xi_t$ gives

$$\mathbb{E}[R_t \mid \mathcal{K}_t] = R_t^{\text{en}} + c_V(L_t, \tau_t)\, v^\top \xi_t, \qquad c_V(L, \tau) := \sum_{l=1}^{L} w_l \rho^l,$$

and summing the geometric series gives the displayed formula. Now condition on $\sigma(s_t, \xi_t) \subseteq \mathcal{K}_t$. The factor $v^\top \xi_t$ is known, $c_V(L_t, \tau_t)$ is bounded, and ξ_t is independent of $\mathcal{F}^{\text{en}}$. By Lemma 1(i),

$$\mathbb{E}[R_t \mid o_t] = \mathbb{E}[R_t^{\text{en}} \mid s_t] + \mathbb{E}[c_V(L_t, \tau_t) \mid s_t]\, v^\top \xi_t.$$

Also $\mathbb{E}[R_t \mid s_t] = \mathbb{E}[R_t^{\text{en}} \mid s_t]$. Indeed, Lemma 1(ii) with $\mathcal{G} = \sigma(s_t, L_t, \tau_t)$ and $Z = \xi_t$ gives $\mathbb{E}[c_V(L_t, \tau_t) v^\top \xi_t \mid s_t, L_t, \tau_t] = c_V(L_t, \tau_t)\, \mathbb{E}[v^\top \xi_t] = 0$.

Range of c_V. Put $y := \gamma\lambda\rho \le \lambda$. For $\tau = 0$, $c_V(L, 0) = \gamma\rho\big[\frac{1-\lambda}{1-y}(1 - y^{L-1}) + y^{L-1}\big] \le \gamma\rho$, since $(1-\lambda)/(1-y) \le 1$. Setting $\tau = 1$ only removes a non-negative term. The limit $L \to \infty$ is immediate.

Untruncated return. The two series defining G_t^λ converge in L^2 by the moment bounds. Conditional expectation is L^2-continuous, so the same computation with $L = \infty$ and all weights $\gamma(1-\lambda)(\gamma\lambda)^{l-1}$ gives $\mathbb{E}[G_t^\lambda \mid o_t] = \mathbb{E}[G_t^\lambda \mid s_t] + c_V v^\top \xi_t$. Finally, $\mathrm{Cov}(G_t^\lambda, \xi_t) = \mathbb{E}[G_t^\lambda \xi_t] = \sum_{l \ge 1} \gamma(1-\lambda)(\gamma\lambda)^{l-1} \rho^l v = c_V v$. This uses $\mathbb{E}[\xi_{t+l} \xi_t^\top] = \rho^l \boldsymbol{I}$ and $\mathbb{E}[X \xi_t] = \mathbb{E}[X]\,\mathbb{E}[\xi_t] = 0$ for the $\mathcal{F}^{\text{en}}$-measurable parts X.

(d3) Let $s' := s^+_{t+j}$, which is $\mathcal{F}^{\text{en}}$-measurable, and $\mathcal{H} := \mathcal{G} \vee \sigma(\xi_t)$.

- Since $Z_{t,j}$ is independent of $\mathcal{K}_t$ and (s', ξ_t) is $\mathcal{K}_t$-measurable, Lemma 1(ii) gives $\mathbb{E}[f(s', \xi_{t+j}) \mid \mathcal{K}_t] = F(s', \xi_t)$, with $F(s', \cdot) := T_{\rho^j} f(s', \cdot)$. Integrability holds because $\mathbb{E}\|f(s', \xi_{t+j})\|^2 = \mathbb{E}\int \|f(s', x)\|^2 \mathcal{N}_m(dx)$, again by the substitution lemma.
- T preserves $\mathcal{N}_m$-means and is linear, so $F = \bar{f}(s') + \tilde{F}(s', \cdot)$, with $\bar{f}(s') = \int f(s', x) \mathcal{N}_m(dx)$ and $\tilde{F}(s', \cdot) = T_{\rho^j} \tilde{f}(s', \cdot)$.
- By the tower property and Lemma 1(i), $\mathbb{E}[f(o^+_{t+j}) \mid \mathcal{H}] = \mathbb{E}[\bar{f}(s') \mid \mathcal{G}] + \mathbb{E}[\tilde{F}(s', \xi_t) \mid \mathcal{H}]$.
- Similarly $\mathbb{E}[f(o^+_{t+j}) \mid \mathcal{G}] = \mathbb{E}[\bar{f}(s') \mid \mathcal{G}]$, because $\mathbb{E}[\tilde{F}(s', \xi_t) \mid \mathcal{F}^{\text{en}}] = \int \tilde{F}(s', y) \mathcal{N}_m(dy) = 0$ by Lemma 1(ii).

The difference in (d3) is therefore $B := \mathbb{E}[\tilde{F}(s', \xi_t) \mid \mathcal{H}]$. Hence

$$\mathbb{E}\|B\|^2 \le \mathbb{E}\|\tilde{F}(s', \xi_t)\|^2 = \mathbb{E}\,\|T_{\rho^j} \tilde{f}(s', \cdot)\|^2 \le \rho^{2j}\, \mathbb{E}\,\|\tilde{f}(s', \cdot)\|^2.$$

The steps are: conditional Jensen; Lemma 1(ii), with ξ_t independent of s'; and Lemma 8 applied componentwise for each fixed s' (each component of $\tilde{f}(s', \cdot)$ has mean zero), followed by integration over s'. For $\tilde{f}(s, x) = \bar{E}x$, $B = \rho^j \bar{E} \xi_t$, and both sides equal $\rho^{2j} \|\bar{E}\|_F^2$.

For G_t^λ, the reward terms cancel by (b). Each critic term contributes $\gamma(1-\lambda)(\gamma\lambda)^{l-1} B_l$, where B_l is the difference above with $j = l$, $\mathcal{G} = \sigma(s_t)$ and $s' = s_{t+l}$, so that $\|B_l\|_{L^2} \le \rho^l \sup_{l'} \|\tilde{V}(s_{t+l'}, \cdot)\|_{L^2}$. The series converges in L^2, and Minkowski's inequality gives the bound with $\sum_l \gamma(1-\lambda)(\gamma\lambda)^{l-1}\rho^l = c_V$. At $\rho = 0$ every bound and coefficient is 0. □

REMARKS

Terms that can pay for noise. The squared-error and convex losses of part (a) cannot reward ξ-content. Two other terms can.

- *VICReg.* Adding ξ-variance to a latent coordinate whose batch standard deviation is below the hinge target lowers the variance hinge. Adding independent ξ-content can also lower the covariance term; numerically this term dominates the hinge. At Θ_0 both effects are quadratic in Ξ, so they do not affect (c), but they make VICReg's curvature in Ξ negative. Theorem 2 idealises this channel away through the constraint $W \Sigma W^\top = \boldsymbol{I}$.
- *The PPO surrogate under joint routing.* If the policy loss reaches the encoder, ξ-content in the actor's mean acts like extra action variance and is rewarded at second order. Under critic routing (our agent) and for detached encoders this channel is absent.

A.3 THEOREM 2: WHICH NOISE THE LATENT KEEPS

This subsection proves Theorem 2 (Theorem 2 of the main text) for one client of the linear-Gaussian instance. Everything is exact for the idealised learning dynamics defined below: a continuous-time semi-gradient flow with L^2-optimal linear heads, an exponential-moving-average (EMA) target, a first-order lag for the encoder that computes the λ-returns, and a second-moment (whitening) constraint in place of LayerNorm and VICReg. The stability statements are about the linearisation transverse to the exo-free set (Definition 3); Lemma 15 gives their meaning for the nonlinear flow. What the idealisation leaves out is listed in Remark 2.5.

Setting. We use the objects of Settings S1–S3 in the following precise form.

(S1) *Endogenous process* (A2, A3). $s_{t+1} = As_t + Ba_t + w_t$ and $a_t = Fs_t + \sigma\eta_t$, with $w_t \sim \mathcal{N}(0, \Sigma_w)$, $\Sigma_w \succ 0$, $\eta_t \sim \mathcal{N}(0, \boldsymbol{I}_{d_u})$, all independent, $\sigma > 0$ and spectral radius $\rho(A+BF) < 1$. Let $\Sigma_s = \sum_{i\geq 0}(A+BF)^i(\sigma^2 BB^\top + \Sigma_w)(A+BF)^{i\top} \succ 0$, the solution of $\Sigma_s = (A+BF)\Sigma_s(A+BF)^\top + \sigma^2 BB^\top + \Sigma_w$. The initial state is $s_0 \sim \mathcal{N}(0, \Sigma_s)$, independent of $(w_t)_{t\geq 0}$ and $(\eta_t)_{t\geq 0}$. Then $(s_t, a_t)_{t\geq 0}$ is a stationary Gaussian process (a Markov chain started at its invariant law) and $s_t \sim \mathcal{N}(0, \Sigma_s)$ for all t. The reward is $r_t = R(s_t, a_t)$ for a measurable R with $\mathbb{E}r_t^2 < \infty$. We write $\mathcal{E}$ for the σ-field generated by $(s_0, (w_t)_{t\geq 0}, (\eta_t)_{t\geq 0})$; states, actions and rewards are $\mathcal{E}$-measurable.

(S2) *Exogenous process* (A1). $\xi_{t+1} = \rho\xi_t + \sqrt{1-\rho^2}\,\epsilon_t$ with $0 \leq \rho < 1$, $\epsilon_t \sim \mathcal{N}(0, \boldsymbol{I}_m)$ i.i.d., $\xi_0 \sim \mathcal{N}(0, \boldsymbol{I}_m)$, and $(\xi_0, (\epsilon_t))$ independent of $\mathcal{E}$. The observation is $o_t = [s_t; \xi_t]$.

(S3) *Whitened coordinates.* Put $x_t := \Sigma_s^{-1/2} s_t$, so that $\mathrm{Cov}(x_t) = \boldsymbol{I}_d$, and $\tilde{o}_t := [x_t; \xi_t]$, so that $\mathrm{Cov}(\tilde{o}_t) = \boldsymbol{I}_{d+m}$. A linear map $[U', E]$ of the observation is written in these coordinates as $W = [U'\Sigma_s^{1/2}, E]$, so that $[U', E]o_t = W\tilde{o}_t$. The objects in the statements below ($\mathcal{M}$, the synchronised stationary points, $\mathcal{J}$, J, s, Γ, Λ) are defined through the latent z and the heads, so they do not depend on this choice. The flow (S6) does: it is the Euclidean semi-gradient flow in these coordinates, which is part of the idealisation (Remark 2.5(3)) and is what makes the constraint set invariant (Lemma 10(iii)). The Euclidean flow in the original coordinates (s, ξ) does not keep the constraint, but it has the same synchronised stationary points and the same transverse linearisation (Remark 2.5(3)). We write $A_w := \Sigma_s^{-1/2} A \Sigma_s^{1/2}$, $A_{cw} := \Sigma_s^{-1/2}(A+BF)\Sigma_s^{1/2}$ (open- and closed-loop dynamics in these coordinates) and $B_w := \Sigma_s^{-1/2} B$.

(S4) *Encoder, target, lagged copy* (Def. 3, A4). The encoder is $z_t = W\tilde{o}_t = Ux_t + E\xi_t$ with $W = [U, E] \in \mathbb{R}^{k\times(d+m)}$, $1 \leq k \leq d$. The second-moment idealisation of LayerNorm and VICReg is the constraint set $\mathcal{W} := \{W : WW^\top = \boldsymbol{I}_k\} = \{W : \mathbb{E}z_t z_t^\top = \boldsymbol{I}_k\}$. The EMA target is $\bar{W} = [\bar{U}, \bar{E}]$; the encoder and critic that compute the λ-returns are $W_o = [U_o, E_o]$ and $w_o \in \mathbb{R}^k$. The target and the lagged copy are not constrained.

(S5) *Heads and losses* (A5). For $j = 1, \ldots, H$ let $\bar{a}_{t,j} := (a_t, \ldots, a_{t+j-1})$. The losses are

$$L_{\mathrm{pred}} = \tfrac{1}{H}\textstyle\sum_{j=1}^H \mathbb{E}\big\|P_j z_t + Q_j \bar{a}_{t,j} - \bar{W}\tilde{o}_{t+j}\big\|^2, \qquad L_R = \mathbb{E}\big(g_z^\top z_t + g_a^\top a_t - r_t\big)^2,$$
$$L_V = \mathbb{E}\big(w^\top z_t - y_t\big)^2, \qquad y_t := \textstyle\sum_{l\geq 0}(\gamma\lambda)^l r_{t+l} + \sum_{i\geq 1}\gamma(1-\lambda)(\gamma\lambda)^{i-1}\, w_o^\top W_o \tilde{o}_{t+i},$$

with $\gamma \in [0, 1)$ and $\lambda \in [0, 1]$; y_t is the λ-return of the lagged critic $w_o^\top W_o \tilde{o}$ (the GAE target $V_o(o_t) + \sum_l (\gamma\lambda)^l \delta_{t+l}$, rearranged). The objective is $L = \lambda_p L_{\mathrm{pred}} + \lambda_V L_V + \lambda_R L_R$ with $\lambda_p, \lambda_V, \lambda_R \geq 0$. The heads $(P_j, Q_j)_{j\leq H}$, w and (g_z, g_a) are the L^2-optimal ones given $(W, \bar{W}, W_o, w_o)$; they are unique (Lemma 10). The predictors are per-horizon heads; for $H = 1$ this is the code's one-step predictor, and the code's shared H-step rollout is Remark 2.1. In the *action-free variant* the predictors have no action input ($Q_j \equiv 0$); Lemma 10 and Theorem 2(a)–(e),(g) hold for it verbatim, with the same proofs (the Gram matrix of the predictor becomes $WW^\top$, and its normal equation in P_j is unchanged).

(S6) *Dynamics* (A6). With $\theta := (W, \bar{W}, W_o, w_o)$ and $G(\theta) := \nabla_W L$ evaluated at the optimal heads, with $\bar{W}, W_o, w_o$ and the heads held fixed (the semi-gradient),

$$\dot{W} = -G(\theta), \qquad \dot{\bar{W}} = \tau_1(W - \bar{W}),$$
$$\dot{W}_o = \tau_2(W - W_o), \qquad \dot{w}_o = \tau_2\big(w(\theta) - w_o\big), \qquad \tau_1, \tau_2 > 0.$$

By Lemma 10(iii) this flow leaves $\mathcal{W}$ invariant and coincides on $\mathcal{W}$ with the Riemannian semi-gradient flow.

(S7) *Exo-free set.* $\mathcal{M} := \{E = \bar{E} = E_o = 0\}$. A point $\theta^0 \in \mathcal{M}$ with $W^0 = [U^0, 0] \in \mathcal{W}$ is a *synchronised stationary point* if $\bar{W}^0 = W_o^0 = W^0$, $w_o^0 = w(\theta^0)$ and $G(\theta^0) = 0$. For $k = d$ every $U^0 \in O(d)$ gives one (Lemma 10(vii)).

Code map (verified against the supplementary code, Remark 2.4): $\lambda_p = 0.03/64$ per unit of summed squared error (the prediction loss is an MSE averaged over the 64 latent coordinates and, for $H = 4$, over the 4 rollout steps), $\lambda_V = \texttt{vf_coef} \times 0.5 = 0.25$, $\lambda_R = 0.1$, $\gamma = 0.99$, $\lambda = 0.95$; the EMA target moves by $\tau = 0.01$ once per round of $2048/64 \times 10 = 320$ online steps; the λ-returns are computed once per round with the round-start encoder and critic (a one-round lag); FedAvg synchronises every 5 rounds, i.e. every 1,600 local steps.

Notation. At a synchronised stationary point θ^0, let P_j be the z_t-coefficient of the optimal horizon-j predictor, w the optimal critic and g_z the z-part of the optimal reward head, and set

$$M = \tfrac{1}{H}\textstyle\sum_j P_j^\top P_j, \quad N = \tfrac{1}{H}\sum_j \rho^j P_j, \quad G_V = ww^\top, \quad G_R = g_z g_z^\top,$$

$$c_V = \frac{\gamma(1-\lambda)\rho}{1-\gamma\lambda\rho}, \qquad c_\rho(H) = \tfrac{1}{H}\textstyle\sum_j \rho^{2j},$$

$$M_{\mathrm{tot}} = \lambda_p M + \lambda_R G_R + (1-c_V)\lambda_V G_V, \qquad J = M_{\mathrm{tot}}^{-1}\lambda_p N^\top,$$

$$s = \big\|M_{\mathrm{tot}}^{-1/2}\lambda_p N^\top M_{\mathrm{tot}}^{-1/2}\big\|, \qquad \Gamma = M_{\mathrm{tot}} - \lambda_p \,\mathrm{sym}(N), \qquad \Lambda = \lambda_{\min}(M_{\mathrm{tot}})/\lambda_p - c_\rho(H),$$

$$R_{\mathrm{tot}} = 2(\lambda_p M + \lambda_R G_R + \lambda_V G_V),$$

where $\mathrm{sym}(X) = \frac{1}{2}(X + X^\top)$ and $\|\cdot\|$ is the spectral norm. J and s require $M_{\mathrm{tot}} \succ 0$ (and vanish when $\lambda_p = 0$); Λ requires $M_{\mathrm{tot}} \succ 0$ and $\lambda_p > 0$. Since $\gamma\rho < 1$ we always have $0 \le c_V < 1$ and $1 - c_V = (1-\gamma\rho)/(1-\gamma\lambda\rho)$. At $(\gamma, \lambda) = (0.99, 0.95)$ and $\rho = 0, .5, .95, .99$: $c_\rho(4) = 0, .083, .779, .951$ and $1 - c_V = 1, .953, .559, .289$.

Definition 3 (transverse linearisation)**.** For one noise coordinate, let $(e, \bar{e}, e_o) \in \mathbb{R}^{3k}$ be the corresponding columns of $(E, \bar{E}, E_o)$ and

$$\mathcal{J} = \begin{pmatrix} -R_{\mathrm{tot}} & 2\lambda_p N^\top & 2\lambda_V c_V G_V \\ \tau_1 \boldsymbol{I}_k & -\tau_1 \boldsymbol{I}_k & 0 \\ \tau_2 \boldsymbol{I}_k & 0 & -\tau_2 \boldsymbol{I}_k \end{pmatrix}.$$

We say that $\mathcal{M}$ is *linearly stable at* θ^0 if $\mathcal{J}$ is Hurwitz, and *linearly unstable at* θ^0 if $\mathcal{J}$ has an eigenvalue with positive real part. By Theorem 2(a), $\mathcal{J}$ is the restriction to one noise column of the derivative of the flow (S6) at θ^0, so these are statements about the linearisation of (S6) transverse to $\mathcal{M}$. Lemma 15 gives their meaning for the nonlinear flow: exponential decay of the noise blocks for as long as the endogenous blocks stay near θ^0, and instability of θ^0.

Theorem 2 (which noise the latent keeps)**.** *Assume (S1)–(S6) and let θ^0 be a synchronised stationary point (S7). "Linearly stable" and "linearly unstable" are meant in the sense of Definition 3.*

(a) Exact first-order linearisation. *The vector field of (S6) is C^∞ near θ^0, and its derivative at θ^0 is block diagonal with respect to the split into endogenous blocks $(U, \bar{U}, U_o, w_o)$ and noise blocks $(E, \bar{E}, E_o)$. On the noise blocks it acts column by column as $\mathcal{J}$:*

$$\dot{e} = -2(\lambda_p M + \lambda_R G_R + \lambda_V G_V)e + 2\lambda_p N^\top \bar{e} + 2\lambda_V c_V G_V e_o,$$
$$\dot{\bar{e}} = \tau_1(e - \bar{e}), \qquad \dot{e}_o = \tau_2(e - e_o).$$

The two corrections that could enter at first order vanish: the constraint correction $\mathrm{sym}(GW^\top)E$ is identically zero, and for fixed endogenous blocks the heads differ from their values on $\mathcal{M}$ by $O(\|(E, \bar{E}, E_o)\|^2)$, so that the noise-block velocity is linear in $(E, \bar{E}, E_o)$ up to $O(\|(E, \bar{E}, E_o)\|^3)$.

(b) Slow target. *Let $M_{\mathrm{tot}} \succ 0$.*

(i) For every $\tau_2 > 0$ there is $\bar{\tau} > 0$ such that for all $0 < \tau_1 < \bar{\tau}$: $\mathcal{M}$ is linearly stable at θ^0 if $\max \mathrm{Re}\,\mathrm{spec}(J) < 1$, and linearly unstable if $\max \mathrm{Re}\,\mathrm{spec}(J) > 1$. The k eigenvalues of $\mathcal{J}$ of smallest modulus are $\tau_1(\mu_i - 1) + o(\tau_1)$, $\mu_i \in \mathrm{spec}(J)$ with multiplicity, and $\tau_1(\mu_i - 1) + O(\tau_1^2)$ for each simple μ_i; the other $2k$ lie in the open left half-plane.

(ii) Per-round EMA. *If between two target updates the pair (e, e_o) relaxes to its equilibrium given $\bar{e}$, and the target is then updated by $\bar{e} \leftarrow (1-\tau)\bar{e} + \tau e$ with $\tau \in (0,1]$, the per-round map is $\bar{e} \mapsto ((1-\tau)\boldsymbol{I} + \tau J)\bar{e}$. It is stable iff $|1-\tau+\tau\mu| < 1$ for every $\mu \in \mathrm{spec}(J)$. If $\max \mathrm{Re}\,\mathrm{spec}(J) < 1$ this holds iff $\tau < \min_{\mu\in\mathrm{spec}(J)} 2(1-\mathrm{Re}\,\mu)/|1-\mu|^2$; if $\max \mathrm{Re}\,\mathrm{spec}(J) \geq 1$ it fails for every $\tau \in (0,1]$; at $\tau = 1$ it reads $\rho(J) < 1$.*

(iii) $\Gamma \succ 0$ *implies* $\max \mathrm{Re}\,\mathrm{spec}(J) < 1$.

(c) Every lag. *If $M_{\mathrm{tot}} \succ 0$ and $s < 1$, then $\mathcal{M}$ is linearly stable at θ^0 for all $\tau_1, \tau_2 > 0$. Quantitatively, let $\nu < 1$ and S_1, S_2 be as in Lemma 12 and $C := 2 + 5(1 + \max_{i=1,2} \|S_i\|/\tau_i)^{1/2}$. Every solution of the transverse system satisfies $\|(e, \bar{e}, e_o)(t)\| \leq C\, e^{-\kappa t/2} \|(e, \bar{e}, e_o)(0)\|$ with $\kappa = (1-\nu)\min\{\lambda_{\min}(R_{\mathrm{tot}}), \tau_1, \tau_2\} \geq (1-\nu)\min\{2\lambda_{\min}(M_{\mathrm{tot}}), \tau_1, \tau_2\}$. Moreover, if $\lambda_p > 0$, then $\Lambda > 0$ implies $s \leq (\lambda_p c_\rho(H)/\lambda_{\min}(M_{\mathrm{tot}}))^{1/2} < 1$; and $s < 1$ implies $\Gamma \succeq (1-s)M_{\mathrm{tot}} \succ 0$ and $\rho(J) \leq s < 1$.*

(d) Three channels. *The set of synchronised stationary points does not depend on ρ, and at each of them the matrices M, G_V and G_R do not depend on ρ; ρ enters $\mathcal{J}$ only through N and c_V.*

(i) Reward: restoring force $2\lambda_R G_R$, no drive; its contribution $\lambda_R G_R$ to M_{tot} is the same for every ρ.

(ii) Value: restoring force $2\lambda_V G_V$, drive $2\lambda_V c_V G_V$ through the lagged copy; it enters every criterion of (b) and (c) only through $(1-c_V)\lambda_V G_V$, with $1 - c_V = (1-\gamma\rho)/(1-\gamma\lambda\rho)$ non-increasing in ρ (strictly decreasing if $\gamma > 0$ and $\lambda < 1$), from 1 at $\rho = 0$ towards $(1-\gamma)/(1-\gamma\lambda)$ as $\rho \to 1$, and $c_V = 0$ for Monte-Carlo returns ($\lambda = 1$).

(iii) Prediction: restoring force $2\lambda_p M$, drive $2\lambda_p N^\top$. For i.i.d. noise ($\rho = 0$): $N = 0$ and $c_V = 0$, so $M_{\mathrm{tot}} = R_{\mathrm{tot}}/2$; $\mathcal{M}$ is linearly stable for all τ_1, τ_2 iff $\lambda_p M + \lambda_R G_R + \lambda_V G_V \succ 0$, and then $J = 0$ and $s = 0$.

(e) Normal open-loop case. *Let $P_j = P^j$ for $j \leq H$, with P normal and eigenvalues $\alpha_1, \ldots, \alpha_k \neq 0$, and let $\lambda_p > 0$ and $\lambda_R G_R = \lambda_V G_V = 0$ (prediction only). Then $\mathrm{spec}(J) = \{\mu_i\}_{i\leq k}$ with $\mu_i = \sum_j \rho^j \bar{\alpha}_i^j / \sum_j |\alpha_i|^{2j}$, and $s = \max_i |\mu_i|$. If P is symmetric positive definite (normal with real positive eigenvalues), then for every $H \geq 1$ and every $\tau_1, \tau_2 > 0$: $\mathcal{M}$ is linearly stable at θ^0 iff $\rho < \min_i \alpha_i$, and $\mathcal{J}$ has a positive real eigenvalue iff $\rho > \min_i \alpha_i$.*

(f) Action conditioning. *Let $k = d$, so $U^0 \in O(d)$. With the action input, $P_1 = U^0 A_w U^{0\top}$ (open-loop persistence); in the action-free variant ($Q_1 \equiv 0$), $P_1 = U^0 A_{cw} U^{0\top}$ (closed-loop persistence). Both are exact, because w_t is independent of (s_t, a_t) and $\sigma B\eta_t + w_t$ is independent of s_t. With the action input and open-loop exploration ($F = 0$), $P_j = U^0 A_w^j U^{0\top} = P_1^j$ for every j. Replacing U^0 by OU^0, $O \in O(d)$, conjugates $\mathcal{J}$ by $\mathrm{diag}(O, O, O)$; hence $\mathrm{spec}(\mathcal{J})$, $\mathrm{spec}(J)$, s and the sign of Γ do not depend on the choice of $U^0 \in O(d)$.*

(g) Sharpness ($k = 1$). *Let $k = 1$ and $M_{\mathrm{tot}} > 0$. Then for all $\tau_1, \tau_2 > 0$,*

$$\mathcal{M} \text{ is linearly stable at } \theta^0 \iff \lambda_p M + \lambda_R G_R + (1-c_V)\lambda_V G_V > \lambda_p N.$$

If $N \geq 0$ (for example if $P_j \geq 0$ for all j), this is equivalent to $s < 1$, and $\mathcal{J}$ has a real positive eigenvalue iff $s > 1$. If $N < 0$, the condition holds, so $\mathcal{M}$ is linearly stable whatever the value of s. For $H = 1$ the condition reads $\lambda_p P_1^2 + \lambda_R G_R + (1-c_V)\lambda_V G_V > \lambda_p \rho P_1$.

Corollary 1 (the latent learners are the distractor-free learners)**.** *Let $k = d$, $E = 0$ and $W = [U, 0] \in \mathcal{W}$. Then $z_t = U\Sigma_s^{-1/2} s_t$ with $U \in O(d)$. The linear policies $a = K_z z + \sigma\eta$ and linear critics $v^\top z$ on z are exactly the linear policies and critics on s, via $K_z \mapsto K_z U \Sigma_s^{-1/2}$ and $v \mapsto \Sigma_s^{-1/2} U^\top v$. For any loss that depends on the actor or critic parameters only through the map $\tilde{o} \mapsto K_z W \tilde{o}$ (resp. $v^\top W \tilde{o}$), gradient flow, and SGD or score-function policy-gradient steps computed from the same samples, on (K_z, v) coincide after the orthogonal change of variables $K_x = K_z U$, $v_x = U^\top v$ with the same algorithm run on the whitened distractor-free state x.*

Lemmas. Lemmas 9 and 10 compute the first-order terms; Lemmas 11 to 13 give stability at every lag; Lemma 14 gives the slow-target limit; Lemma 15 links the transverse linearisation to the nonlinear flow.

Lemma 9 (L2.1: exogenous moments)**.** *Under (S1)–(S5), for all $t, j \geq 0$: (i) $\mathbb{E}[X\xi_t^\top] = 0$ for every square-integrable $\mathcal{E}$-measurable random vector X; (ii) $\mathbb{E}[\xi_{t+j}\xi_t^\top] = \rho^j \boldsymbol{I}_m$; (iii) y_t is square-integrable and $\mathbb{E}[y_t\xi_t^\top] = c_V\, w_o^\top E_o$.*

Proof. By induction $\xi_t \sim \mathcal{N}(0, \boldsymbol{I}_m)$ for all t: if $\xi_t \sim \mathcal{N}(0, \boldsymbol{I}_m)$ then ξ_{t+1} is centred Gaussian with covariance $\rho^2 \boldsymbol{I}_m + (1-\rho^2)\boldsymbol{I}_m = \boldsymbol{I}_m$. Moreover ξ_t is a function of $(\xi_0, \epsilon_0, \dots, \epsilon_{t-1})$, which is independent of $\mathcal{E}$. (i) X and ξ_t are independent and $\mathbb{E}\xi_t = 0$, so $\mathbb{E}[X\xi_t^\top] = \mathbb{E}[X]\,\mathbb{E}[\xi_t]^\top = 0$. (ii) Iterating the recursion, $\xi_{t+j} = \rho^j \xi_t + \sqrt{1-\rho^2}\sum_{i=0}^{j-1} \rho^{j-1-i}\epsilon_{t+i}$, and each ϵ_{t+i} is centred and independent of ξ_t; hence $\mathbb{E}[\xi_{t+j}\xi_t^\top] = \rho^j \mathbb{E}[\xi_t\xi_t^\top] = \rho^j \boldsymbol{I}_m$. (iii) The processes $(x_t, a_t)_t$ and $(\xi_t)_t$ are stationary (by (S1) and the first sentence) and independent, hence jointly stationary. So $\|r_{t+l}\|_{L^2} = \|r_0\|_{L^2} < \infty$ and $\|w_o^\top W_o \tilde{o}_{t+i}\|_{L^2} = \|w_o^\top W_o \tilde{o}_0\|_{L^2}$, and $\gamma\lambda < 1$ because $\gamma < 1$; so both series defining y_t converge absolutely in L^2, and expectations of products with ξ_t may be taken term by term. The reward terms and the terms $w_o^\top U_o x_{t+i}$ contribute 0 by (i); the terms $w_o^\top E_o \xi_{t+i}$ contribute $\rho^i w_o^\top E_o$ by (ii). Hence $\mathbb{E}[y_t\xi_t^\top] = \sum_{i\geq 1} \gamma(1-\lambda)(\gamma\lambda)^{i-1}\rho^i\, w_o^\top E_o = \frac{\gamma(1-\lambda)\rho}{1-\gamma\lambda\rho}\, w_o^\top E_o$, since $\gamma\lambda\rho < 1$. □

Lemma 10 (L2.2: first-order terms)**.** *Assume (S1)–(S5) and let $WW^\top \succ 0$ (in particular $W \in \mathcal{W}$).*

(i) The Gram matrices of the three regressions are positive definite, so the optimal heads are unique; they are rational functions of θ with non-vanishing denominators, hence C^∞ on $\{WW^\top \succ 0\}$.

(ii) At the optimal heads, the noise block of $G = \nabla_W L$ is, exactly,

$$G_\xi = \tfrac{2\lambda_p}{H}\sum_j P_j^\top\big(P_j E - \rho^j \bar{E}\big) + 2\lambda_V\, w\big(w^\top E - c_V\, w_o^\top E_o\big) + 2\lambda_R\, g_z g_z^\top E.$$

(iii) $GW^\top = 0$. Consequently $\frac{d}{dt}(WW^\top) = 0$ along (S6), $\mathcal{W}$ is invariant, and on $\mathcal{W}$ the Riemannian gradient $G - \mathrm{sym}(GW^\top)W$ equals G.

(iv) For fixed $(U, \bar{U}, U_o, w_o)$, the optimal heads depend on the noise blocks only through $EE^\top$ and $\bar{E}E^\top$ (predictor), $EE^\top$ and $EE_o^\top w_o$ (critic), and $EE^\top$ (reward head). They are even functions of $(E, \bar{E}, E_o)$ and differ from their values at $E = \bar{E} = E_o = 0$ by $O(\|(E, \bar{E}, E_o)\|^2)$, locally uniformly in $(U, \bar{U}, U_o, w_o)$.

(v) The endogenous block G_U is even and G_ξ is odd in $(E, \bar{E}, E_o)$. On $\mathcal{M}$: $G_\xi = 0$ (so $\mathcal{M}$ is invariant), $\partial G_U/\partial(E, \bar{E}, E_o) = 0$, and $\partial G_\xi/\partial(U, \bar{U}, U_o, w_o) = 0$.

(vi) The critic's L^2-optimum on $\mathcal{M} \cap \mathcal{W}$ is $w = U(b + KU_o^\top w_o)$ with $b := \mathbb{E}[x_t \sum_{l\geq 0}(\gamma\lambda)^l r_{t+l}]$ and $K := \sum_{i\geq 1}\gamma(1-\lambda)(\gamma\lambda)^{i-1}\mathbb{E}[x_t x_{t+i}^\top]$, and $\|K\| \leq \gamma(1-\lambda)/(1-\gamma\lambda) < 1$.

(vii) If $k = d$, then for every $U^0 \in O(d)$ there is exactly one w^0 such that $\theta^0 = ([U^0, 0], [U^0, 0], [U^0, 0], w^0)$ is a synchronised stationary point, namely $w^0 = (\boldsymbol{I} - U^0 K U^{0\top})^{-1} U^0 b$.

Proof. Throughout we use Lemma 9(i)–(ii) and $\mathrm{Cov}(\tilde{o}_t) = \boldsymbol{I}$. Write $z_t = Ux_t + E\xi_t$, $\bar{z}_{t+j} = \bar{U}x_{t+j} + \bar{E}\xi_{t+j}$, $C_j := \mathbb{E}[x_t \bar{a}_{t,j}^\top]$, $A_j := \mathbb{E}[\bar{a}_{t,j}\bar{a}_{t,j}^\top]$, $C_j' := \mathbb{E}[x_{t+j}x_t^\top]$ and $D_j := \mathbb{E}[x_{t+j}\bar{a}_{t,j}^\top]$, and let b and K be as in (vi); by Lemma 9(i), $\mathbb{E}[x_t y_t] = b + KU_o^\top w_o$ for every (U_o, E_o, w_o).

(i) *Predictor.* The Gram matrix of $(z_t, \bar{a}_{t,j})$ is $\mathcal{G}_j = \left(\begin{smallmatrix} WW^\top & UC_j \\ C_j^\top U^\top & A_j \end{smallmatrix}\right)$, because $\mathbb{E}[\xi_t \bar{a}_{t,j}^\top] = 0$. Suppose $\mathbb{E}(u^\top z_t + v^\top \bar{a}_{t,j})^2 = 0$. The actions are linear in x_t and in the innovations $\eta_t, \dots, \eta_{t+j-1}$, $w_t, \dots, w_{t+j-2}$, which are independent of (x_t, ξ_t); the coefficient of η_{t+i} in a_{t+i} is $\sigma\boldsymbol{I}$ and a_{t+i} does not involve $\eta_{t+i'}$ for $i' > i$. Since independent parts of a sum add variances, the innovation part $v^\top(\cdot)$ of $v^\top \bar{a}_{t,j}$ has variance 0, and the block-triangular structure with diagonal blocks $\sigma\boldsymbol{I}$ forces $v = 0$. Then $u^\top WW^\top u = 0$, so $u = 0$. Hence $\mathcal{G}_j \succ 0$ and $[P_j, Q_j] = [\mathbb{E}\bar{z}_{t+j}z_t^\top,\ \mathbb{E}\bar{z}_{t+j}\bar{a}_{t,j}^\top]\,\mathcal{G}_j^{-1}$ with $\mathbb{E}[\bar{z}_{t+j}z_t^\top] = \bar{U}C_j'U^\top + \rho^j \bar{E}E^\top$ and $\mathbb{E}[\bar{z}_{t+j}\bar{a}_{t,j}^\top] = \bar{U}D_j$. *Critic.* Its Gram matrix is $\mathbb{E}z_t z_t^\top = WW^\top \succ 0$ and $w = (WW^\top)^{-1}\mathbb{E}[z_t y_t]$, with $\mathbb{E}[z_t y_t] = U(b + KU_o^\top w_o) + c_V EE_o^\top w_o$ by Lemma 9(i),(iii). *Reward head.* The Gram matrix of (z_t, a_t) is the case $j = 1$ of $\mathcal{G}_j$, and the right-hand side is $(U\mathbb{E}[x_t r_t], \mathbb{E}[a_t r_t])$. All entries are polynomials in θ, which proves (i).

(ii) For fixed heads each loss is a quadratic function of W. With the residual $\varrho_j := P_j z_t + Q_j \bar{a}_{t,j} - \bar{z}_{t+j}$ we get $\nabla_W \mathbb{E}\|\varrho_j\|^2 = 2P_j^\top \mathbb{E}[\varrho_j \tilde{o}_t^\top]$, whose noise block is $2P_j^\top \mathbb{E}[\varrho_j \xi_t^\top]$ and $\mathbb{E}[\varrho_j \xi_t^\top] = P_j(U \cdot 0 + E \cdot \boldsymbol{I}) + Q_j \cdot 0 - (\bar{U} \cdot 0 + \bar{E}\rho^j \boldsymbol{I}) = P_j E - \rho^j \bar{E}$. Likewise $\nabla_W \mathbb{E}(w^\top z_t - y_t)^2 = 2w\,\mathbb{E}[(w^\top z_t - y_t)\tilde{o}_t^\top]$ has noise block $2w(w^\top E - \mathbb{E}[y_t \xi_t^\top]) = 2w(w^\top E - c_V w_o^\top E_o)$ by Lemma 9(iii), and $\nabla_W \mathbb{E}(g_z^\top z_t + g_a^\top a_t - r_t)^2 = 2g_z \mathbb{E}[(g_z^\top z_t + g_a^\top a_t - r_t)\tilde{o}_t^\top]$ has noise block $2g_z(g_z^\top E + 0 - 0)$. Summing with the weights gives (ii).

(iii) From (ii)'s computation, $G = \frac{2\lambda_p}{H} \sum_j P_j^\top \mathbb{E}[\varrho_j \tilde{o}_t^\top] + 2\lambda_V w\,\mathbb{E}[(w^\top z_t - y_t)\tilde{o}_t^\top] + 2\lambda_R g_z \mathbb{E}[(g_z^\top z_t + g_a^\top a_t - r_t)\tilde{o}_t^\top]$. Multiplying on the right by $W^\top$ turns each $\tilde{o}_t^\top$ into $z_t^\top$, and each resulting expectation is zero by the normal equation of the corresponding head: $\mathbb{E}[\varrho_j z_t^\top] = 0$ (optimality in P_j), $\mathbb{E}[(w^\top z_t - y_t) z_t] = 0$ (in w), and $\mathbb{E}[(g_z^\top z_t + g_a^\top a_t - r_t) z_t] = 0$ (in g_z). Hence $GW^\top = 0$ and $\frac{d}{dt}(WW^\top) = -GW^\top - WG^\top = 0$. (This is the mechanism of Tang et al., 2023, Thm. 1, and of Ni et al., 2024, Thm. 3 for an action-conditioned predictor with an EMA target; here it covers all three heads.)

(iv) The formulas in the proof of (i) show the stated dependence; every listed quantity is even in $(E, \bar{E}, E_o)$. An even C^∞ function h has $Dh = 0$ at the origin, so $h(x) - h(0) = \int_0^1 (1-t) D^2 h(tx)[x, x]\, dt = O(\|x\|^2)$, uniformly on compact sets of the other arguments.

(v) By the computation in (ii), with $\tilde{o}_t$ replaced by its endogenous part x_t, and Lemma 9(i), $\mathbb{E}[\varrho_j x_t^\top] = P_j U + Q_j C_j^\top - \bar{U} C_j'$, $\mathbb{E}[(w^\top z_t - y_t) x_t^\top] = w^\top U - (b + K U_o^\top w_o)^\top$ and $\mathbb{E}[(g_z^\top z_t + g_a^\top a_t - r_t) x_t^\top] = g_z^\top U + g_a^\top \mathbb{E}[a_t x_t^\top] - \mathbb{E}[r_t x_t^\top]$; these depend on the noise blocks only through the heads, which are even by (iv). Hence G_U is even. By (ii), G_ξ is a sum of products of even matrices with E, $\bar{E}$ or E_o, hence odd. An odd function vanishes at 0, and so do its derivatives in the other arguments there; the derivative of an even function at 0 vanishes.

(vi) On $\mathcal{M}$ we have $E_o = 0$, so the bootstrap part of y_t is $\sum_i \gamma(1-\lambda)(\gamma\lambda)^{i-1} w_o^\top U_o x_{t+i}$, and $\mathbb{E}[x_t w_o^\top U_o x_{t+i}] = \mathbb{E}[x_t x_{t+i}^\top] U_o^\top w_o$; with $WW^\top = \boldsymbol{I}$ this gives $w = U(b + K U_o^\top w_o)$. By Cauchy–Schwarz, $|u^\top \mathbb{E}[x_t x_{t+i}^\top] v| \le \|u\|\,\|v\|$, so $\|K\| \le \sum_{i \ge 1} \gamma(1-\lambda)(\gamma\lambda)^{i-1} = \gamma(1-\lambda)/(1-\gamma\lambda)$, which is < 1 because $\gamma < 1$.

(vii) At θ^0 the lagged-copy equations vanish iff $w_o = w(\theta^0)$, i.e. iff $w_o = U^0 b + U^0 K U^{0\top} w_o$. Since $\|U^0 K U^{0\top}\| = \|K\| < 1$, this has the unique solution w^0. By (v), $G_\xi(\theta^0) = 0$; by (iii), $G_U(\theta^0) U^{0\top} = G(\theta^0) W^{0\top} = 0$, and U^0 is invertible, so $G_U(\theta^0) = 0$. □

Lemma 11 (L2.3: lagged-target Lyapunov lemma)**.** *Let $S_0, S_1, \dots, S_n \in \mathbb{R}^{k \times k}$ be symmetric positive semidefinite, $\tau_1, \dots, \tau_n > 0$, $\nu \in [0, 1)$, and let $D_1, \dots, D_n \in \mathbb{R}^{k \times k}$ satisfy $|u^\top D_i v| \le \nu \|S_i^{1/2} u\|\,\|S_i^{1/2} v\|$ for all $u, v \in \mathbb{R}^k$. Consider*

$$\dot{e} = -\Big(S_0 + \textstyle\sum_i S_i\Big) e + \sum_i D_i \bar{e}_i, \qquad \dot{\bar{e}}_i = \tau_i (e - \bar{e}_i), \qquad i = 1, \dots, n,$$

and $V := \|e\|^2 + \sum_i \tau_i^{-1} \bar{e}_i^\top S_i \bar{e}_i$. Then along every solution

$$\dot{V} \le -2 e^\top S_0 e - (1 - \nu) \textstyle\sum_i \big(e^\top S_i e + \bar{e}_i^\top S_i \bar{e}_i\big).$$

If moreover $R := S_0 + \sum_i S_i \succ 0$ and $\kappa := (1 - \nu) \min\{\lambda_{\min}(R), \tau_1, \dots, \tau_n\}$, then $V(t) \le e^{-\kappa t} V(0)$, $\|e(t)\| \le V(0)^{1/2} e^{-\kappa t/2}$ and $\|\bar{e}_i(t)\| \le (\|\bar{e}_i(0)\| + 2V(0)^{1/2}) e^{-\kappa t/2}$; in particular the system matrix is Hurwitz with spectral abscissa at most $-\kappa/2$.

Proof. Differentiating, $\dot{V} = 2e^\top \dot{e} + \sum_i 2\tau_i^{-1} \bar{e}_i^\top S_i \dot{\bar{e}}_i = -2e^\top S_0 e + \sum_i \big(-2 e^\top S_i e + 2 e^\top D_i \bar{e}_i + 2\bar{e}_i^\top S_i e - 2\bar{e}_i^\top S_i \bar{e}_i\big)$. Put $f_i = S_i^{1/2} e$ and $\bar{f}_i = S_i^{1/2} \bar{e}_i$. By the hypothesis on D_i and Cauchy–Schwarz, $2e^\top D_i \bar{e}_i + 2\bar{e}_i^\top S_i e \le 2(1+\nu)\|f_i\|\|\bar{f}_i\| \le (1+\nu)(\|f_i\|^2 + \|\bar{f}_i\|^2)$, so the i-th bracket is at most $-(1-\nu)(\|f_i\|^2 + \|\bar{f}_i\|^2)$, which is the first claim. For the second, use $2e^\top S_0 e \ge (1-\nu) e^\top S_0 e$ to get $\dot{V} \le -(1-\nu)\big(e^\top R e + \sum_i \tau_i \cdot \tau_i^{-1} \bar{e}_i^\top S_i \bar{e}_i\big) \le -\kappa V$; Grönwall's inequality gives $V(t) \le e^{-\kappa t} V(0)$ and hence the bound on $\|e(t)\| \le V(t)^{1/2}$. By variation of constants, $\bar{e}_i(t) = e^{-\tau_i t} \bar{e}_i(0) + \tau_i \int_0^t e^{-\tau_i (t-u)} e(u)\, du$; since $\kappa \le \tau_i$, $\tau_i \int_0^t e^{-\tau_i (t-u)} e^{-\kappa u/2} du \le \tau_i e^{-\kappa t/2} \int_0^\infty e^{-(\tau_i - \kappa/2)v} dv \le 2e^{-\kappa t/2}$, which gives the bound on $\bar{e}_i$. Finally, if an eigenvalue μ had $\operatorname{Re}\mu > -\kappa/2$, with eigenvector $\xi \in \mathbb{C}^{k(n+1)}$, the real solutions $\operatorname{Re}(e^{\mu t}\xi)$ and $\operatorname{Im}(e^{\mu t}\xi)$ would decay at rate $\kappa/2$, although $\|\operatorname{Re}(e^{\mu t}\xi)\|^2 + \|\operatorname{Im}(e^{\mu t}\xi)\|^2 = e^{2\operatorname{Re}\mu\, t}\|\xi\|^2$; this is a contradiction. □

Lemma 12 (L2.4: the split)**.** *Let $M_{\mathrm{tot}} \succ 0$ and $s < 1$. Put $\theta_V := \lambda_V\|G_V\|/\lambda_{\min}(M_{\mathrm{tot}})$, $\varepsilon^* := 1-c_V$ if $\theta_V = 0$ and $\varepsilon^* := \min\{1-c_V,\ (1-s)/(2\theta_V)\}$ otherwise, $\beta := 1-c_V-\varepsilon^* \in [0, 1-c_V)$, and*

$$S_0 = 0, \quad S_1 = 2(\lambda_p M + \lambda_R G_R + \beta\lambda_V G_V), \quad S_2 = 2(1-\beta)\lambda_V G_V,$$
$$D_1 = 2\lambda_p N^\top, \quad D_2 = 2\lambda_V c_V G_V.$$

Then $S_1 + S_2 = R_{\mathrm{tot}}$, $S_1 \succeq (1+s)M_{\mathrm{tot}} \succ 0$, and the hypotheses of Lemma 11 hold with

$$\nu = \max\Big\{\frac{s}{1-\varepsilon^*\theta_V},\ \frac{c_V}{c_V+\varepsilon^*}\Big\} \le \max\Big\{\frac{2s}{1+s},\ \frac{c_V}{c_V+\varepsilon^*}\Big\} < 1$$

(with the second entry omitted when $\lambda_V G_V = 0$). When θ_V is large, ε^* is small and ν can be very close to 1; the rate κ of Theorem 2(c) is then only qualitative, while the stability conclusion is unaffected.

Proof. $S_1 + S_2 = R_{\mathrm{tot}}$ is immediate. With $M_\beta := S_1/2 = M_{\mathrm{tot}} - \varepsilon^*\lambda_V G_V$ and $G_V \preceq \|G_V\|\boldsymbol{I} \preceq (\|G_V\|/\lambda_{\min}(M_{\mathrm{tot}}))M_{\mathrm{tot}}$ we get $M_\beta \succeq (1-\varepsilon^*\theta_V)M_{\mathrm{tot}}$, and $1-\varepsilon^*\theta_V \ge (1+s)/2 > 0$ by the choice of ε^*; this gives $S_1 \succeq (1+s)M_{\mathrm{tot}}$. Let $X := M_{\mathrm{tot}}^{-1/2}\lambda_p N^\top M_{\mathrm{tot}}^{-1/2}$, so $\|X\| = s$. Then $|u^\top D_1 v| = 2|\langle M_{\mathrm{tot}}^{1/2}u, XM_{\mathrm{tot}}^{1/2}v\rangle| \le 2s\,\|M_{\mathrm{tot}}^{1/2}u\|\,\|M_{\mathrm{tot}}^{1/2}v\| \le \frac{2s}{1-\varepsilon^*\theta_V}\|M_\beta^{1/2}u\|\,\|M_\beta^{1/2}v\| = \frac{s}{1-\varepsilon^*\theta_V}\|S_1^{1/2}u\|\,\|S_1^{1/2}v\|$. Since $1-\beta = c_V+\varepsilon^*$, $D_2 = \frac{c_V}{c_V+\varepsilon^*}S_2$, so $|u^\top D_2 v| \le \frac{c_V}{c_V+\varepsilon^*}\|S_2^{1/2}u\|\,\|S_2^{1/2}v\|$ by Cauchy–Schwarz. Finally $s/(1-\varepsilon^*\theta_V) \le 2s/(1+s) < 1$ because $s < 1$, and $c_V/(c_V+\varepsilon^*) < 1$ because $\varepsilon^* > 0$. □

Lemma 13 (L2.5: from Λ to s to Γ and $\rho(J)$)**.** *Let $M_{\mathrm{tot}} \succ 0$. (i) If $\lambda_p > 0$, then $s \le (\lambda_p c_\rho(H)/\lambda_{\min}(M_{\mathrm{tot}}))^{1/2}$; in particular $\Lambda > 0$ implies $s < 1$. (ii) $\langle v, \Gamma v\rangle \ge (1-s)\langle v, M_{\mathrm{tot}}v\rangle$ for all v, so $s < 1$ implies $\Gamma \succ 0$. (iii) $\mathrm{spec}(J) = \mathrm{spec}(M_{\mathrm{tot}}^{-1/2}\lambda_p N^\top M_{\mathrm{tot}}^{-1/2})$, so $\rho(J) \le s$.*

Proof. (i) Let $\|u\| = \|v\| = 1$ and $a = M_{\mathrm{tot}}^{-1/2}u$, $b = M_{\mathrm{tot}}^{-1/2}v$. Then $\langle u, M_{\mathrm{tot}}^{-1/2}\lambda_p N^\top M_{\mathrm{tot}}^{-1/2}v\rangle = \lambda_p\langle Na, b\rangle = \frac{\lambda_p}{H}\sum_j \rho^j\langle P_j a, b\rangle$. By Cauchy–Schwarz in $\mathbb{R}^k$ and then over j, $\frac{1}{H}\sum_j \rho^j|\langle P_j a, b\rangle| \le \|b\|\big(\frac{1}{H}\sum_j \rho^{2j}\big)^{1/2}\big(\frac{1}{H}\sum_j \|P_j a\|^2\big)^{1/2} = \|b\|\,c_\rho(H)^{1/2}\langle a, Ma\rangle^{1/2}$. Since $\lambda_p M \preceq M_{\mathrm{tot}}$ (the other two terms are positive semidefinite because $c_V \le 1$), $\lambda_p\langle a, Ma\rangle \le \langle a, M_{\mathrm{tot}}a\rangle = 1$; and $\|b\| \le \lambda_{\min}(M_{\mathrm{tot}})^{-1/2}$. Hence the bilinear form is at most $\lambda_p \cdot \lambda_{\min}(M_{\mathrm{tot}})^{-1/2}c_\rho(H)^{1/2}\lambda_p^{-1/2}$; taking the supremum over u, v gives the bound, and $\Lambda > 0$ is equivalent to $\lambda_p c_\rho(H)/\lambda_{\min}(M_{\mathrm{tot}}) < 1$. (ii) $\langle v, \lambda_p\,\mathrm{sym}(N)v\rangle = \langle v, \lambda_p N^\top v\rangle = \langle M_{\mathrm{tot}}^{1/2}v, XM_{\mathrm{tot}}^{1/2}v\rangle \le s\langle v, M_{\mathrm{tot}}v\rangle$ with X as in the proof of Lemma 12. (iii) $J = M_{\mathrm{tot}}^{-1/2}XM_{\mathrm{tot}}^{1/2}$ is similar to X, and $\rho(X) \le \|X\| = s$. □

Lemma 14 (L2.6: fast block and slow eigenvalues)**.** *Let $M_{\mathrm{tot}} \succ 0$, $\tau_2 > 0$, $F := \left(\begin{smallmatrix} -R_{\mathrm{tot}} & 2\lambda_V c_V G_V \\ \tau_2\boldsymbol{I} & -\tau_2\boldsymbol{I}\end{smallmatrix}\right)$, $B := \left(\begin{smallmatrix} 2\lambda_p N^\top \\ 0\end{smallmatrix}\right)$ and $C' := (\boldsymbol{I}\ \ 0)$. (i) F is Hurwitz, and $C'(-F)^{-1}B = J$. (ii) $K_0 := \sup_{\mathrm{Re}\,z\ge 0}\|C'(z\boldsymbol{I}-F)^{-1}B\| < \infty$, and every eigenvalue z of $\mathcal{J}$ with $\mathrm{Re}\,z \ge 0$ satisfies $|z| \le \tau_1(1+K_0)$. (iii) As $\tau_1 \downarrow 0$, the eigenvalues of $\mathcal{J}$ in the disk $\{|z| < \tau_1 R_0\}$, for any fixed $R_0 > \max\{1+K_0, \max_{\mu\in\mathrm{spec}J}|\mu-1|\}$, are exactly k in number (with multiplicity), and they are $\tau_1(\mu_i - 1 + o(1))$, $\mu_i \in \mathrm{spec}(J)$ with multiplicity; for a simple μ_i the $o(1)$ is $O(\tau_1)$.*

Proof. (i) Apply Lemma 11 with $n = 1$ (only the lagged copy e_o), $S_0 = 2(\lambda_p M + \lambda_R G_R)$, $S_1 = 2\lambda_V G_V$, $D_1 = 2\lambda_V c_V G_V = c_V S_1$ and $\nu = c_V < 1$; then $R = R_{\mathrm{tot}} \succeq 2M_{\mathrm{tot}} \succ 0$, so the matrix F of this system is Hurwitz. For the identity, solve $F(u_1; u_2) = B$: the second block row gives $u_1 = u_2$ and the first gives $-(R_{\mathrm{tot}} - 2\lambda_V c_V G_V)u_1 = 2\lambda_p N^\top$, i.e. $u_1 = -(2M_{\mathrm{tot}})^{-1}2\lambda_p N^\top = -J$; so $C'(-F)^{-1}B = -u_1 = J$.

(ii) Permuting the state to $(e, e_o, \bar{e})$, $\mathcal{J}$ is similar to $\mathcal{K} = \left(\begin{smallmatrix} F & B \\ \tau_1 C' & -\tau_1\boldsymbol{I}\end{smallmatrix}\right)$. Since F is Hurwitz, $z \mapsto (z\boldsymbol{I}-F)^{-1}$ is continuous on the closed right half-plane, and $\|(z\boldsymbol{I}-F)^{-1}\| \le (|z| - \|F\|)^{-1}$ for $|z| > \|F\|$; hence $K_0 < \infty$. If $\mathcal{K}(u; v) = z(u; v)$ with $\mathrm{Re}\,z \ge 0$ and $(u, v) \ne 0$, then $u = (z\boldsymbol{I}-F)^{-1}Bv$, so $v \ne 0$ and $(z+\tau_1)v = \tau_1 C'(z\boldsymbol{I}-F)^{-1}Bv$; thus $|z+\tau_1| \le \tau_1 K_0$ and $|z| \le \tau_1(1+K_0)$.

(iii) For $z \notin \operatorname{spec}(F)$ the Schur-complement formula gives $\det(z\boldsymbol{I}-\mathcal{K}) = \det(z\boldsymbol{I}-F)\det\big((z+\tau_1)\boldsymbol{I} - \tau_1 C'(z\boldsymbol{I}-F)^{-1}B\big)$. With $z=\tau_1\mu$,

$$\det(\tau_1\mu\boldsymbol{I}-\mathcal{K}) = \tau_1^k \det(\tau_1\mu\boldsymbol{I}-F)\, q_{\tau_1}(\mu), \qquad q_{\tau_1}(\mu) := \det\big((\mu+1)\boldsymbol{I} - C'(\tau_1\mu\boldsymbol{I}-F)^{-1}B\big),$$

and $q_0(\mu) = \det((\mu+1)\boldsymbol{I}-J)$ by (i). Let $D := \{|\mu| \le R_0\}$. For $\tau_1 R_0 \|F^{-1}\| < \frac{1}{2}$, a Neumann series gives $\|(\tau_1\mu\boldsymbol{I}-F)^{-1} - (-F)^{-1}\| \le 2\tau_1 R_0 \|F^{-1}\|^2$ on D, so $q_{\tau_1} \to q_0$ uniformly on D; moreover $\det(\tau_1\mu\boldsymbol{I}-F) \neq 0$ on a neighbourhood of D, where q_{τ_1} is therefore analytic. All zeros of q_0 lie in $|\mu| < R_0$ by the choice of R_0. Choose $r > 0$ so small that the closed disks of radius r around the distinct zeros of q_0 are disjoint and contained in $\{|\mu| < R_0\}$, and let $\eta > 0$ be the minimum of $|q_0|$ on their boundary circles and on $|\mu| = R_0$. For τ_1 small, $|q_{\tau_1} - q_0| < \eta \le |q_0|$ on all these circles. On each circle apply Rouché's theorem (Conway, 1978, Thm. V.3.8) to $f = q_{\tau_1}$ and $g = -q_0$: both are analytic on a neighbourhood of the closed disk, neither vanishes on the circle (for f because $|q_{\tau_1} - q_0| < |q_0|$ there), and $|f+g| = |q_{\tau_1} - q_0| < |q_0| \le |f| + |g|$. So q_{τ_1} has as many zeros as q_0 in each small disk and in $|\mu| < R_0$, namely k in total; hence it has none in $|\mu| < R_0$ outside the small disks. As $\tau_1^k \det(\tau_1\mu\boldsymbol{I}-F)$ does not vanish on D, the zeros of q_{τ_1} in $|\mu| < R_0$ are, with multiplicity, the eigenvalues of $\mathcal{K}$ in $|z| < \tau_1 R_0$ divided by τ_1. Since r is arbitrary, these are $\mu_i - 1 + o(1)$. If μ_i is simple, $\mu_i - 1$ is a simple zero of q_0; $q_{\tau_1}(\mu)$ is analytic in (τ_1, μ) near $(0, \mu_i - 1)$ with $\partial_\mu q_0(\mu_i - 1) \neq 0$, and the implicit function theorem gives a zero $\mu_i - 1 + O(\tau_1)$. □

Lemma 15 (L2.7: from the transverse linearisation to the nonlinear flow)**.** *Let θ^0 be a synchronised stationary point and write $\theta = (\theta_s, \chi)$ with endogenous blocks $\theta_s = (U, \bar{U}, U_o, w_o)$ and noise blocks $\chi = (E, \bar{E}, E_o)$. (i) If $\mathcal{M}$ is linearly stable at θ^0 ($\mathcal{J}$ Hurwitz), there are a neighbourhood $\mathcal{N}$ of θ^0_s and constants $\delta, \kappa > 0$, $C \ge 1$ such that every solution of (S6) with $\theta_s(0) \in \mathcal{N}$ and $\|\chi(0)\| \le \delta$ satisfies $\|\chi(t)\| \le Ce^{-\kappa t}\|\chi(0)\|$ for all $t < T_{\mathcal{N}} := \inf\{t \ge 0 : \theta_s(t) \notin \mathcal{N}\}$. This is conditional on the endogenous blocks: we do not prove that $\theta_s(t)$ stays in $\mathcal{N}$, so (i) is not a claim of local attraction of the nonlinear flow. (ii) If $\mathcal{M}$ is linearly unstable at θ^0, then θ^0 is an unstable equilibrium of (S6): there is $\delta_1 > 0$ such that for every $\eta > 0$ some solution starting on $\mathcal{W}$ within η of θ^0 leaves the δ_1-ball around θ^0. The same holds for the noise-block dynamics with the endogenous blocks frozen at their values at θ^0 on $\mathcal{W}$ (i.e. $U = (\boldsymbol{I} - EE^\top)^{1/2}U^0$, $\bar{U} = U_o = U^0$, $w_o = w^0$), and there the noise block itself reaches size δ_1.*

Proof. (i) By Lemma 10(ii),(iv), $\dot{\chi} = \mathcal{A}(\theta_s)\chi + \mathcal{R}(\theta_s, \chi)$, where $\mathcal{A}(\theta_s)$ acts on each noise column by the matrix $\mathcal{J}(\theta_s)$ obtained from Definition 3 with the heads evaluated at $(\theta_s, 0)$ and G_V in the drive replaced by $ww_o^\top$; $\mathcal{A}$ is continuous, $\mathcal{J}(\theta_s^0) = \mathcal{J}$, and $\|\mathcal{R}(\theta_s,\chi)\| \le C_{\mathcal{R}}\|\chi\|^3$ for θ_s in a compact neighbourhood $\mathcal{N}_0$ of θ_s^0 and $\|\chi\| \le \delta_0$ (the heads' deviations are $O(\|\chi\|^2)$ and multiply terms linear in χ; $\dot{\bar{E}}$ and $\dot{E}_o$ are exactly linear). Since $\mathcal{J}$ is Hurwitz, $P := \int_0^\infty e^{\mathcal{J}^\top t} e^{\mathcal{J}t} dt$ converges, is positive definite, and satisfies $\mathcal{J}^\top P + P\mathcal{J} = \int_0^\infty \frac{d}{dt}(e^{\mathcal{J}^\top t}e^{\mathcal{J}t})dt = -\boldsymbol{I}$. Let $\mathcal{V}(\chi) := \sum_c \chi_c^\top P \chi_c$ (sum over noise columns). Then $\dot{\mathcal{V}} \le -\|\chi\|^2 + 2\|P\|\,\|\mathcal{J}(\theta_s) - \mathcal{J}\|\,\|\chi\|^2 + 2\|P\|C_{\mathcal{R}}\|\chi\|^4$. Choose $\mathcal{N} \subset \mathcal{N}_0$ with $2\|P\|\,\|\mathcal{J}(\theta_s)-\mathcal{J}\| \le \frac{1}{4}$ on $\mathcal{N}$, and $\delta_1 \le \delta_0$ with $2\|P\|C_{\mathcal{R}}\delta_1^2 \le \frac{1}{4}$. While $\theta_s \in \mathcal{N}$ and $\|\chi\| \le \delta_1$, $\dot{\mathcal{V}} \le -\frac{1}{2}\|\chi\|^2 \le -\mathcal{V}/(2\|P\|)$, because $\mathcal{V} \le \|P\|\,\|\chi\|^2$. Put $\delta := \frac{1}{2}\delta_1(\lambda_{\min}(P)/\|P\|)^{1/2}$ and let $\theta_s(0) \in \mathcal{N}$, $\|\chi(0)\| \le \delta$. Let $T_1 \le T_{\mathcal{N}}$ be the supremum of the times $t < T_{\mathcal{N}}$ such that $\|\chi(u)\| \le \delta_1$ for all $u \in [0,t]$; $T_1 > 0$ by continuity. On $[0, T_1)$ the differential inequality gives $\mathcal{V}(t) \le e^{-t/(2\|P\|)}\mathcal{V}(0)$, hence $\|\chi(t)\|^2 \le \mathcal{V}(t)/\lambda_{\min}(P) \le \|P\|\,\|\chi(0)\|^2/\lambda_{\min}(P) \le \delta_1^2/4$. If $T_1 < T_{\mathcal{N}}$, continuity would give $\|\chi\| \le \delta_1/2 < \delta_1$ on a neighbourhood of T_1, contradicting the definition of T_1; so $T_1 = T_{\mathcal{N}}$, and on $[0, T_{\mathcal{N}})$, $\|\chi(t)\|^2 \le e^{-t/(2\|P\|)}\,\|P\|\,\|\chi(0)\|^2/\lambda_{\min}(P)$. This is the claim with $C = (\|P\|/\lambda_{\min}(P))^{1/2}$ and $\kappa = 1/(4\|P\|)$. (ii) Let x denote either the deviation $\theta - \theta^0$ in the Euclidean coordinates of (S6), where the vector field is C^∞ on $\{WW^\top \succ 0\}$ (Lemma 10(i)), or the noise blocks χ of the frozen system, which obey an autonomous C^∞ equation on $\{\|E\| < 1\}$ with equilibrium 0. In both cases $\dot{x} = \mathcal{A}x + \mathcal{R}(x)$ with $\|\mathcal{R}(x)\| \le C_{\mathcal{R}}\|x\|^2$ for $\|x\| \le \delta_0$. Here $\mathcal{A}$ is the derivative at the equilibrium: it is block diagonal with noise block acting as $\mathcal{J}$ on each column, by Theorem 2(a) and, for the frozen system, because the frozen U differs from U^0 by $O(\|E\|^2)$. So $\mathcal{A}$ has an eigenvalue with positive real part, with an eigenvector supported on the noise blocks. The rest is the instability half of Lyapunov's indirect method (Khalil, 2002, Sec. 4.3), which we prove for completeness. Choose $0 < \alpha < \max \operatorname{Re}\operatorname{spec}(\mathcal{J})$ different from the real part of every eigenvalue of $\mathcal{A}$. Let $\pi_\pm$ be the real spectral projections of $\mathcal{A}$ onto the invariant subspaces $E_\pm$ of the eigenvalues with real part $> \alpha$ and $< \alpha$, and let $\mathcal{A}_\pm$ be the restrictions; E_+ contains a real vector x_v supported on the noise blocks. Since $-(\mathcal{A}_+ - \alpha)$ and $\mathcal{A}_- - \alpha$ are Hurwitz, the integrals

$P_+ = \int_0^\infty e^{-(\mathcal{A}_+-\alpha)^\top t} e^{-(\mathcal{A}_+-\alpha)t} dt$ and $P_- = \int_0^\infty e^{(\mathcal{A}_--\alpha)^\top t} e^{(\mathcal{A}_--\alpha)t} dt$ converge, where $^\top$ denotes the adjoint with respect to the restriction of the Euclidean inner product $\langle\cdot,\cdot\rangle$ of the ambient space to $E_\pm$. They are positive definite on $E_\pm$ and satisfy $(\mathcal{A}_+ - \alpha)^\top P_+ + P_+(\mathcal{A}_+ - \alpha) = \boldsymbol{I}$ and $(\mathcal{A}_- - \alpha)^\top P_- + P_-(\mathcal{A}_- - \alpha) = -\boldsymbol{I}$ on $E_\pm$. Put $\mathcal{C}(x) := \langle \pi_+ x, P_+ \pi_+ x\rangle - \langle \pi_- x, P_- \pi_- x\rangle$. Along the flow, $\dot{\mathcal{C}} = \|\pi_+ x\|^2 + \|\pi_- x\|^2 + 2\alpha\,\mathcal{C} + 2\langle \pi_+ \mathcal{R}, P_+ \pi_+ x\rangle - 2\langle \pi_- \mathcal{R}, P_- \pi_- x\rangle$. (here $\pi_\pm$ commute with $\mathcal{A}$, and all norms are ambient). Now $x = \pi_+ x + \pi_- x$ gives $\|\pi_+ x\|^2 + \|\pi_- x\|^2 \geq \frac{1}{2}\|x\|^2$, and the last two terms are at most $c_1\|x\|^3$ with $c_1 := 2C_{\mathcal{R}}(\|\pi_+\|^2\|P_+\| + \|\pi_-\|^2\|P_-\|)$. Hence $\dot{\mathcal{C}} \geq 2\alpha\,\mathcal{C}$ while $\|x\| \leq \delta_1 := \min\{\delta_0, 1/(2c_1)\}$. Take $x(0)$ with $\|x(0)\| < \eta$ and $\mathcal{C}(x(0)) > 0$: for instance, the noise blocks $\eta' x_v$ with η' small, the other blocks at θ^0, and $U = (\boldsymbol{I} - \eta'^2 E_v E_v^\top)^{1/2} U^0$ so that $W \in \mathcal{W}$. This point differs from $\eta' x_v$ by $O(\eta'^2)$, so $\mathcal{C}(x(0)) = \eta'^2\langle x_v, P_+ x_v\rangle + O(\eta'^3) > 0$. As long as $\|x(t)\| \leq \delta_1$, $\mathcal{C}(x(t)) \geq e^{2\alpha t}\mathcal{C}(x(0)) \to \infty$, whereas $\mathcal{C}$ is bounded on that ball. So the solution leaves the ball, however small η is. In the frozen system $x = \chi$, so it is the noise block that grows to size δ_1. □

Proof of Theorem 2. *(a)* By Lemma 10(i) the heads, hence G, are C^∞ on the open set $\{WW^\top \succ 0\}$, which contains θ^0; the lagged-copy equations are linear; and $\mathcal{W}$ is invariant by Lemma 10(iii). The endogenous components $(-G_U, \tau_1(U - \bar{U}), \tau_2(U - U_o), \tau_2(w - w_o))$ have zero derivative in the noise blocks at θ^0, because G_U and w are even in $(E, \bar{E}, E_o)$ (Lemma 10(iv),(v)) and the lagged-copy terms do not involve them. The noise components $(-G_\xi, \tau_1(E - \bar{E}), \tau_2(E - E_o))$ have zero derivative in the endogenous blocks, because $G_\xi \equiv 0$ on $\mathcal{M}$ (Lemma 10(v)). So the derivative is block diagonal. By Lemma 10(ii), G_ξ is a sum of heads times E, $\bar{E}$ or E_o; when we differentiate at $\chi = 0$ only the heads' values at θ^0 survive, and $w_o^0 = w$ there. This gives

$$\dot{E} = -\tfrac{2\lambda_p}{H}\textstyle\sum_j P_j^\top P_j E - 2\lambda_V w w^\top E - 2\lambda_R g_z g_z^\top E + \tfrac{2\lambda_p}{H}\textstyle\sum_j \rho^j P_j^\top \bar{E} + 2\lambda_V c_V w w^\top E_o + (\text{higher order}),$$

which is the stated system since $\frac{1}{H}\sum_j P_j^\top P_j = M$ and $\frac{1}{H}\sum_j \rho^j P_j^\top = N^\top$. Every term is a left multiplication, so the noise columns evolve independently. The constraint correction vanishes by Lemma 10(iii). The head deviations and the cubic remainder are Lemma 10(iv).

(b)(i) By Lemma 14(ii), every eigenvalue of $\mathcal{J}$ with non-negative real part lies in $\{|z| \leq \tau_1(1 + K_0)\} \subset \{|z| < \tau_1 R_0\}$. By Lemma 14(iii), for small τ_1 the eigenvalues in this disk are $\tau_1(\mu_i - 1 + o(1))$. If $\max_i \operatorname{Re}\mu_i < 1$, choose $\bar{\tau}$ so that all of them have negative real part; then no eigenvalue has $\operatorname{Re} z \geq 0$, and $\mathcal{J}$ is Hurwitz. If $\operatorname{Re}\mu_i > 1$ for some i, the corresponding eigenvalue has positive real part for small τ_1. The remaining $2k$ eigenvalues have $|z| \geq \tau_1 R_0 > \tau_1(1 + K_0)$ and hence negative real part. The asymptotics are Lemma 14(iii). The constants K_0, R_0 and hence $\bar{\tau}$ depend on τ_2 through F.

(b)(ii) With $\bar{e}$ fixed, $f = (e, e_o)$ obeys $\dot{f} = Ff + B\bar{e}$. By Lemma 14(i), F is Hurwitz, so $f(t) \to f^* = -F^{-1}B\bar{e}$, and the computation in that proof gives $f^* = (J\bar{e}, J\bar{e})$. The per-round map is therefore $\bar{e} \mapsto (1 - \tau)\bar{e} + \tau J\bar{e}$, with eigenvalues $1 - \tau + \tau\mu$, $\mu \in \operatorname{spec}(J)$. A linear map is stable iff its spectral radius is < 1. Moreover $|1 - \tau + \tau\mu|^2 = 1 - 2\tau(1 - \operatorname{Re}\mu) + \tau^2|1 - \mu|^2$, which is < 1 iff $\tau|1 - \mu|^2 < 2(1 - \operatorname{Re}\mu)$: for $\operatorname{Re}\mu < 1$ this holds iff $\tau < 2(1 - \operatorname{Re}\mu)/|1 - \mu|^2$, and for $\operatorname{Re}\mu \geq 1$ it fails for every $\tau > 0$.

(b)(iii) If $Jv = \mu v$ with $v \in \mathbb{C}^k \setminus \{0\}$, then $\lambda_p N^\top v = \mu M_{\text{tot}} v$, so $\mu = v^* \lambda_p N^\top v / v^* M_{\text{tot}} v$ and $\operatorname{Re}\mu = v^* \lambda_p \operatorname{sym}(N) v / v^* M_{\text{tot}} v$, because $\operatorname{Re}(v^* X v) = v^* \operatorname{sym}(X) v$ for real X. If $\Gamma \succ 0$, then $v^* \lambda_p \operatorname{sym}(N) v < v^* M_{\text{tot}} v$, so $\operatorname{Re}\mu < 1$.

(c) With the split of Lemma 12, the transverse system is the system of Lemma 11 with $n = 2$, $\bar{e}_1 = \bar{e}$ and $\bar{e}_2 = e_o$, because $S_1 + S_2 = R_{\text{tot}}$, $D_1 = 2\lambda_p N^\top$ and $D_2 = 2\lambda_V c_V G_V$. Lemma 11 gives the decay at rate $\kappa/2$, with $C = 2 + 5(1 + \max_i \|S_i\|/\tau_i)^{1/2}$, since $V(0) \leq (1 + \max_i \|S_i\|/\tau_i)\|(e, \bar{e}, e_o)(0)\|^2$. Also $\lambda_{\min}(R_{\text{tot}}) \geq 2\lambda_{\min}(M_{\text{tot}})$ because $R_{\text{tot}} - 2M_{\text{tot}} = 2c_V\lambda_V G_V \succeq 0$. The last two claims are Lemma 13.

(d) On $\mathcal{M}$ all regressions defining the heads have endogenous regressors and targets: at a point of $\mathcal{M}$ with endogenous blocks $(U, \bar{U}, U_o, w_o)$, P_j regresses $\bar{U}x_{t+j}$ on $(Ux_t, \bar{a}_{t,j})$, the critic regresses $\sum_l (\gamma\lambda)^l r_{t+l} + \sum_i \gamma(1 - \lambda)(\gamma\lambda)^{i-1} w_o^\top U_o x_{t+i}$ on Ux_t, and (g_z, g_a) regresses r_t on (Ux_t, a_t). None of them involves the ξ-process, so the heads on $\mathcal{M}$ do not depend on ρ; by the formulas

in the proof of Lemma 10(v), neither do the residual moments that make up G_U on $\mathcal{M}$. The conditions defining a synchronised stationary point (S7) are $\theta^0 \in \mathcal{M}$, $W^0 \in \mathcal{W}$, $\bar{W}^0 = W_o^0 = W^0$, $w_o^0 = w(\theta^0)$ and $G(\theta^0) = 0$; on $\mathcal{M}$ the last one reduces to $G_U = 0$, because $G_\xi \equiv 0$ there (Lemma 10(v)). So none of these conditions involves ρ, and the set of synchronised stationary points does not depend on ρ (for $k = d$ it is the $O(d)$-family of Lemma 10(vii)). At each of them M, G_V and G_R are built from the heads, hence do not depend on ρ, and ρ enters $\mathcal{J}$ only through N and c_V. (i) G_R appears in R_{tot} and in no drive. (ii) The drive $2\lambda_V c_V G_V$ is read off $\mathcal{J}$; the criteria $\mathrm{spec}(J)$, s, Γ and Λ involve G_V only through M_{tot}, that is, through $(1-c_V)\lambda_V G_V$. Further, $1 - c_V = (1 - \gamma\lambda\rho - \gamma(1-\lambda)\rho)/(1-\gamma\lambda\rho) = (1-\gamma\rho)/(1-\gamma\lambda\rho)$, and its derivative in ρ is $-\gamma(1-\lambda)/(1-\gamma\lambda\rho)^2 \le 0$, with equality iff $\gamma(1-\lambda) = 0$. (iii) At $\rho = 0$, $N = 0$ and $c_V = 0$, so $M_{\mathrm{tot}} = R_{\mathrm{tot}}/2$. $\mathcal{J}$ is then block lower triangular with diagonal blocks $-R_{\mathrm{tot}}$, $-\tau_1 \boldsymbol{I}$ and $-\tau_2 \boldsymbol{I}$, and it is Hurwitz iff $R_{\mathrm{tot}} \succ 0$. In that case $M_{\mathrm{tot}} \succ 0$, and $J = 0$, $s = 0$ because $N = 0$.

(e) Write $P = \mathcal{U}\mathcal{D}\mathcal{U}^*$ with $\mathcal{U}$ unitary and $\mathcal{D} = \mathrm{diag}(\alpha_i)$. Since P is real, $P^\top = P^*$ and $(P^j)^\top = \mathcal{U}\bar{\mathcal{D}}^j\mathcal{U}^*$, so $M = \mathcal{U}\,\mathrm{diag}(m_i)\,\mathcal{U}^*$ with $m_i = \frac{1}{H}\sum_j |\alpha_i|^{2j} > 0$, and $N^\top = \mathcal{U}\,\mathrm{diag}(\bar{n}_i)\,\mathcal{U}^*$ with $n_i = \frac{1}{H}\sum_j \rho^j \alpha_i^j$. With $M_{\mathrm{tot}} = \lambda_p M$, $J = \mathcal{U}\,\mathrm{diag}(\bar{n}_i/m_i)\,\mathcal{U}^*$, which gives the eigenvalues $\mu_i = \bar{n}_i/m_i$. The matrix $M^{-1/2}N^\top M^{-1/2} = \mathcal{U}\,\mathrm{diag}(\bar{n}_i/m_i)\,\mathcal{U}^*$ is normal, so its norm s equals its spectral radius $\max_i |\mu_i|$. If P is symmetric positive definite, $\mathcal{U}$ can be taken real orthogonal and $\alpha_i > 0$. In the coordinates $\mathcal{U}^\top e$ and $\mathcal{U}^\top \bar{e}$ the transverse system splits into k systems $\frac{d}{dt}(e_i, \bar{e}_i) = \begin{pmatrix} -2\lambda_p m_i & 2\lambda_p n_i \\ \tau_1 & -\tau_1 \end{pmatrix}(e_i, \bar{e}_i)$, while e_o is a stable filter that does not feed back (eigenvalue $-\tau_2$), since $\lambda_V G_V = 0$. Each 2×2 matrix has negative trace and determinant $2\lambda_p\tau_1(m_i - n_i)$, so it is Hurwitz iff $m_i > n_i$ and has a real positive eigenvalue iff $m_i < n_i$. Finally $m_i - n_i = \frac{1}{H}\sum_j \alpha_i^j(\alpha_i^j - \rho^j)$, and every term has the sign of $\alpha_i - \rho$, since $x \mapsto x^j$ is strictly increasing on $[0, \infty)$. Part (e) is the action-conditioned, EMA-target, multi-horizon analogue of Voelcker et al. (2024, Props. 1 and 7). In their tabular, action-free, two-timescale analysis, every invariant subspace not spanned by the top-k eigenvectors of the transition matrix is asymptotically unstable when that matrix is real-diagonalisable with positive eigenvalues (Prop. 1; the stability of the top-k subspace itself is not asserted there); and when a distracting process has larger eigenvalues than the reward-relevant one, the top-k eigenspace represents the reward only through its mean (Prop. 7).

(f) On $\mathcal{M}$ with $k = d$ and $\bar{U}^0 = U^0 \in O(d)$, $z_t = U^0 x_t$, and $\bar{z}_{t+1} = U^0 x_{t+1} = (U^0 A_w U^{0\top}) z_t + (U^0 B_w) a_t + U^0 \Sigma_s^{-1/2} w_t$. The noise term w_t is independent of (s_t, η_t), hence of the regressors (z_t, a_t), and it is centred; so $(U^0 A_w U^{0\top}, U^0 B_w)$ satisfies the normal equations and is the unique optimum (Lemma 10(i)). Without the action input, $\bar{z}_{t+1} = (U^0 A_{cw} U^{0\top}) z_t + U^0 \Sigma_s^{-1/2}(\sigma B \eta_t + w_t)$, and the noise is independent of x_t. With $F = 0$, $x_{t+j} = A_w^j x_t + \sum_{i<j} A_w^{j-1-i}(B_w a_{t+i} + \Sigma_s^{-1/2} w_{t+i})$, and $a_{t+i} = \sigma\eta_{t+i}$ is independent of every w; so the residual $\sum_i A_w^{j-1-i}\Sigma_s^{-1/2} w_{t+i}$ is independent of $(x_t, \bar{a}_{t,j})$, and $P_j = U^0 A_w^j U^{0\top}$. Under $U^0 \mapsto OU^0$ with $O \in O(d)$, the synchronised stationary point of Lemma 10(vii) has $w^0 \mapsto Ow^0$ (from the formula for w^0), and every head transforms by conjugation ($P_j \mapsto OP_jO^\top$, $Q_j \mapsto OQ_j$, $w \mapsto Ow$, $g_z \mapsto Og_z$), because the regressions are unique and $\bar{U}x_{t+j}$, Ux_t are replaced by $O\bar{U}x_{t+j}$, OUx_t. Hence M, N, G_V, G_R, M_{tot}, J and Γ are conjugated by O, and $\mathcal{J}$ by $\mathrm{diag}(O, O, O)$; so $\mathrm{spec}(\mathcal{J})$, $\mathrm{spec}(J)$, s and the sign of Γ are unchanged.

(g) For $k = 1$ all quantities are scalars. Put $a = \lambda_p M + \lambda_R G_R + \lambda_V G_V$, $b = \lambda_p N$ and $c = \lambda_V c_V G_V$; then $c \ge 0$, $M_{\mathrm{tot}} = a - c > 0$, and the condition of (g) reads $a - b - c > 0$. Expanding along the first row,

$$\chi(z) := \det(z\boldsymbol{I} - \mathcal{J}) = (z+2a)(z+\tau_1)(z+\tau_2) - 2b\tau_1(z+\tau_2) - 2c\tau_2(z+\tau_1),$$

and in particular $\det(-\mathcal{J}) = \chi(0) = 2\tau_1\tau_2(a - b - c)$. *If $a - b - c > 0$:* suppose $\chi(z) = 0$ with $\mathrm{Re}\, z \ge 0$. Then $z \ne -\tau_1, -\tau_2$ and $z + 2a = 2b\tau_1/(z+\tau_1) + 2c\tau_2/(z+\tau_2)$. For $\mathrm{Re}\, z \ge 0$ and $\tau > 0$, $0 < \mathrm{Re}\,\frac{1}{z+\tau} = \frac{\mathrm{Re}\, z + \tau}{|z+\tau|^2} \le \frac{1}{\mathrm{Re}\, z + \tau} \le \frac{1}{\tau}$, so $\mathrm{Re}\,(2b\tau_1/(z+\tau_1)) \le 2\max\{b, 0\}$ and $\mathrm{Re}\,(2c\tau_2/(z+\tau_2)) \le 2c$. Taking real parts gives $2a \le \mathrm{Re}\, z + 2a \le 2\max\{b, 0\} + 2c$: for $b \ge 0$ this contradicts $a - b - c > 0$, and for $b < 0$ it contradicts $a - c > 0$. So $\mathcal{J}$ is Hurwitz. *If $a - b - c \le 0$:* then $b \ge a - c > 0$, and on $[0, \infty)$, $\varphi(z) := \chi(z)/((z+\tau_1)(z+\tau_2)) = z + 2a - 2b\tau_1/(z+\tau_1) - 2c\tau_2/(z+\tau_2)$ is continuous and strictly increasing (as $b, c \ge 0$), with $\varphi(0) = 2(a - b - c) \le 0$ and $\varphi(z) \to \infty$. So χ has a real root $z^* \ge 0$ and $\mathcal{J}$ is not Hurwitz; this proves the equivalence. If $N \ge 0$, then $b \ge 0$ and $s = b/(a - c)$, so $a - b - c > 0 \iff s < 1$;

a real positive eigenvalue $z > 0$ is a zero of φ on $(0, \infty)$, which exists iff $\varphi(0) < 0$, i.e. iff $s > 1$. If $N < 0$, then $b < 0 < a - c$ and the condition holds. For $N \geq 0$ the off-diagonal entries of $\mathcal{J}$ are non-negative, so $\mathcal{J}$ is Metzler, and the leading principal minors of $-\mathcal{J}$ are $2a$, $2\tau_1(a-b)$ and $2\tau_1\tau_2(a-b-c)$. This is consistent with the M-matrix characterisation of Hurwitz Metzler matrices, which the argument above does not need. For $H = 1$, $M = P_1^2$ and $N = \rho P_1$. □

Proof of Corollary 1. If $E = 0$ and $W \in \mathcal{W}$, then $UU^\top = \boldsymbol{I}$ and U is square, so $U \in O(d)$ and $z_t = Ux_t = U\Sigma_s^{-1/2}s_t$. The identities $K_z z = (K_z U \Sigma_s^{-1/2})s$ and $v^\top z = (\Sigma_s^{-1/2}U^\top v)^\top s$ give the bijections between the classes. Let a loss satisfy $\ell(K_z) = \tilde{\ell}(K_z U)$. Then $\nabla_{K_z}\ell = \nabla\tilde{\ell}(K_z U)\, U^\top$, so along $\dot{K}_z = -\nabla_{K_z}\ell$ we have $\frac{d}{dt}(K_z U) = -\nabla\tilde{\ell}(K_z U)U^\top U = -\nabla\tilde{\ell}(K_z U)$. The same identity holds sample by sample for stochastic gradients, and for the score $\nabla_{K_z}\log\pi(a \mid z) = \sigma^{-2}(a - K_z z)z^\top = \sigma^{-2}(a - K_x x)x^\top U^\top$. Hence the SGD and policy-gradient updates on K_z map, under $K \mapsto KU$, to the same updates on K_x computed from the same samples. The critic is identical. (This conjugacy holds for rotation-equivariant methods. Adam, which normalises per coordinate, is not rotation-equivariant.) □

Remarks.

Remark 2.1 *(Shared rollout.)* If the H-step predictor is one recursive linear map (P, Q) trained on the averaged rollout loss (so $P_j = P^j$), the per-horizon normal equations are replaced by aggregate ones. In open loop ($F = 0$) the rollout of the one-step optimum minimises every horizon, and nothing changes. Under feedback, synchronised points of $\mathcal{M}$ are in general *not* stationary for the rollout flow: the latent basis keeps rotating, and Theorem 2 does not apply. At a synchronised stationary point with a unique non-degenerate shared head, (a)–(c) hold with M_{tot} replaced by $M_{\mathrm{tot}} - \Lambda_0/2$, where $\Lambda_0 = \frac{2\lambda_p}{H}\sum_j \mathrm{sym}\big((P^j)^\top \mathbb{E}[\varrho_j z_t^\top]\big)$; the proof is that of (a) with this extra, even, constraint term. Under feedback the shared P lies between the open- and closed-loop dynamics, and a longer horizon can lower the threshold in ρ.

Remark 2.2 *(Nonlinear heads.)* For differentiable heads the computation of Lemma 10(ii) has the same structure, with $M = \frac{1}{H}\sum_j \mathbb{E}[D_j^\top D_j]$ and $N = \frac{1}{H}\sum_j \rho^j \mathbb{E}[D_j]$ built from the rollout Jacobians $D_j = \partial_z f_j$, plus a residual-curvature term that vanishes for heads linear in z. Since $\mathbb{E}[D_j^\top D_j] \succeq \mathbb{E}[D_j]^\top \mathbb{E}[D_j]$, variability of the Jacobian across states and actions adds restoring force and no drive; this is the matrix analogue of the action-variance term of Khetarpal et al. (2025, Thm. 2 and Remark 1).

Remark 2.3 *(Feature dictionary; $k > d$.)* Let $z = W\varphi(o)$ with $\varphi = (\varphi_{\mathrm{en}}(s), \varphi_{\mathrm{ex}}(o))$, $\mathbb{E}\varphi\varphi^\top = \boldsymbol{I}$ and $\mathbb{E}[\varphi_{\mathrm{ex}} \mid s] = 0$. The proofs of Lemmas 9 and 10 go through with ξ_t replaced by $\varphi_{\mathrm{ex}}(o_t)$, $\rho^j \boldsymbol{I}$ by $K_j = \mathbb{E}[\varphi_{\mathrm{ex}}(o_{t+j})\varphi_{\mathrm{ex}}(o_t)^\top]$, and $\|K_j\| \leq \lambda^{(j)} \leq (\lambda^{(1)})^j$, where $\lambda^{(j)}$ is the lag-j maximal correlation of ξ. For isotropic Gaussian AR(1) noise $\lambda^{(j)} = \rho^j$ (Mehler's formula: the lag-j conditional expectation acts as ρ^{jn} on Hermite polynomials of degree n). The Λ-criterion of (c) then holds with $c_\rho(H)$ replaced by $\frac{1}{H}\sum_j(\lambda^{(j)})^2$, for any $k \leq \dim\varphi_{\mathrm{en}}$; the sharper s-criterion is not claimed.

Remark 2.4 *(The weights used in our experiments.)* Per unit of summed squared error, $\lambda_p : \lambda_V : \lambda_R = 4.69 \cdot 10^{-4} : 0.25 : 0.1$. In a direction u with $u^\top G_V u \geq (\lambda_p/\lambda_V)u^\top M u$ the value restoring force dominates the prediction one; in directions where both critic and reward head are blind ($G_V u = G_R u = 0$) only prediction acts. The value channel keeps a fraction $1 - c_V \geq (1-\gamma)/(1-\gamma\lambda) = 0.168$ of its restoring force at every ρ.

Remark 2.5 *(What Theorem 2 does not cover.)* (1) *A2.* If the actor reads a latent with a noise block E_a, the first-order noise-block terms of the reward channel and of the one-step prediction channel are unchanged for $k = d$ (by Gaussian decomposition and the normal equations); for $k < d$, for horizons $j \geq 2$, and for the value channel, a first-order term proportional to $K_z E_a$ appears, which we do not analyse. (2) *Refresh schedule.* Refreshing the λ-returns once per round, with the online noise relaxing within the round, gives a per-round map whose eigenvalues near 1 are $1 + \tau(\mu - 1) + o(\tau)$, $\mu \in \mathrm{spec}(J)$, so for small τ it has the verdict of (b)(ii). (3) *Optimiser.* Adam, gradient and value clipping, VICReg as a soft penalty, nonlinear encoders and finite samples are not modelled. A preconditioned flow $\dot{W} = -G\,\mathrm{diag}(\Pi, \boldsymbol{I})$ has the same stationary points and transverse linearisation.

(4) *Slow target.* Part (b) is a limit statement with a non-explicit $\bar{\tau}$; only part (c) holds uniformly in the lags.

A.4 THE DISTRACTOR TAX IN A LINEAR–QUADRATIC INSTANCE

This subsection states and proves Theorem 3 and its corollaries. There is one client, and J is a discounted *cost*, to be minimised. The model makes the linear–quadratic assumption (A3) and the policy-gradient assumption (A9) explicit. Not modelled: PPO's clipping and advantage normalisation, GAE advantages in the actor, Adam, and the representation dynamics of Theorem 2; the latent encoder is held fixed here.

Setting.

Definition 4 (LQ instance with an exogenous block)**.** Let $d, d_u, m \geq 1$, $\gamma \in (0,1)$, $\rho \in [0,1)$ and $\sigma > 0$. Let $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times d_u}$, and let $Q \in \mathbb{R}^{d\times d}$, $R \in \mathbb{R}^{d_u\times d_u}$ and $\Sigma_w \in \mathbb{R}^{d\times d}$ be symmetric positive definite. For a gain $K_o = [K_s\ K_\xi]$ with $K_s \in \mathbb{R}^{d_u\times d}$ and $K_\xi \in \mathbb{R}^{d_u\times m}$, the state, the exogenous noise and the action evolve as

$$s_{t+1} = As_t + Ba_t + w_t, \qquad \xi_{t+1} = \rho\,\xi_t + \sqrt{1-\rho^2}\,\epsilon_t, \qquad a_t = K_s s_t + K_\xi \xi_t + \sigma\eta_t. \tag{6}$$

Here $s_0 \sim \mu_0$ with $\Sigma_0 := \mathbb{E}[s_0 s_0^\top]$ finite, $\xi_0 \sim \mathcal{N}(0, \boldsymbol{I}_m)$, $w_t \sim \mathcal{N}(0, \Sigma_w)$, $\epsilon_t \sim \mathcal{N}(0, \boldsymbol{I}_m)$, $\eta_t \sim \mathcal{N}(0, \boldsymbol{I}_{d_u})$, and $s_0, \xi_0, (w_t)_t, (\epsilon_t)_t, (\eta_t)_t$ are mutually independent. The observation is $o_t = [s_t; \xi_t]$ and the policy is $\pi(\cdot \mid o) = \mathcal{N}(K_o o, \sigma^2 \boldsymbol{I}_{d_u})$. With the cost $c(s,a) = s^\top Q s + a^\top R a$,

$$J(K_s, K_\xi) := \mathbb{E}\sum_{t\geq 0} \gamma^t c(s_t, a_t) \in [0, \infty].$$

$J_s(K_s)$ is the cost of the same problem without the exogenous block, under $a_t = K_s s_t + \sigma\eta_t$, so $J_s(K_s) = J(K_s, 0)$. We write spr for the spectral radius, $\|\cdot\|$ for the operator norm, $A_K := A + BK_s$, and

$$\mathcal{K}_\gamma := \{K_s :\ \mathrm{spr}(A_K) < \gamma^{-1/2}\}, \qquad J^\star := \inf_{K_s\in\mathcal{K}_\gamma} J_s(K_s), \qquad \kappa_\rho := \frac{1-\rho}{1+\rho} \in (0,1].$$

We assume $\mathcal{K}_\gamma \neq \emptyset$. For $K_s \in \mathcal{K}_\gamma$ let $P_\gamma(K_s) := \sum_{u\geq 0} \gamma^u (A_K^u)^\top (Q + K_s^\top R K_s) A_K^u$ (by Lemma 18, the unique solution of $P = Q + K_s^\top R K_s + \gamma A_K^\top P A_K$) and $M_\gamma(K_s) := R + \gamma B^\top P_\gamma(K_s) B$. Every term of the series is positive semidefinite, so $P_\gamma \succeq Q$ and $M_\gamma \succeq R$.

Exogenous response. By linearity, $s_t = s_t^{\mathrm{e}} + s_t^\xi$ and $a_t = a_t^{\mathrm{e}} + a_t^\xi$. Here $(s_t^{\mathrm{e}}, a_t^{\mathrm{e}})$ is the distractor-free trajectory, $s_{t+1}^{\mathrm{e}} = A_K s_t^{\mathrm{e}} + B\sigma\eta_t + w_t$ with $s_0^{\mathrm{e}} = s_0$ and $a_t^{\mathrm{e}} = K_s s_t^{\mathrm{e}} + \sigma\eta_t$, and the response to the exogenous block is

$$\begin{gathered} s_0^\xi = 0, \qquad s_{t+1}^\xi = A_K s_t^\xi + BK_\xi \xi_t, \qquad a_t^\xi = K_s s_t^\xi + K_\xi \xi_t, \\ q_\rho(K_s, K_\xi) := \sum_{t\geq 0} \gamma^t\, \mathbb{E}\, c(s_t^\xi, a_t^\xi) \in [0, \infty]. \end{gathered} \tag{7}$$

Definition 5 (Score-function estimator; A9)**.** Fix $K_s \in \mathcal{K}_\gamma$, K_ξ, and a batch size $N \geq 1$. Run N independent episodes of equation 6, indexed by i. Draw times $\tau_1, \ldots, \tau_N$, independent of each other and of the episodes, with $\Pr(\tau_i = t) = (1-\gamma)\gamma^t$; then $o_{\tau_i}^i$ is a draw from the normalised discounted occupancy. Let $h_t^i := (o_0^i, a_0^i, \ldots, o_t^i, a_t^i)$ be the history of episode i up to its action at time t, and let

$$G_i := \sum_{u\geq 0} \gamma^u c\big(s_{\tau_i+u}^i, a_{\tau_i+u}^i\big), \qquad b_i := b\big(\tau_i, h_{\tau_i}^i, U\big),$$

where b is any measurable function and U is a random element independent of the batch (for instance, a critic fitted on earlier data). The baseline may read ξ, and even the sampled action. It is *action-independent* if $b(t, h_t, U)$ does not depend on a_t. The estimator is

$$\hat{g} := \frac{1}{N}\sum_{i=1}^N (G_i - b_i)\, \nabla_{K_o} \log \pi\big(a_{\tau_i}^i \mid o_{\tau_i}^i\big) = \frac{1}{N}\sum_{i=1}^N \frac{G_i - b_i}{\sigma}\, \eta_{\tau_i}^i \big(o_{\tau_i}^i\big)^\top \in \mathbb{R}^{d_u\times(d+m)},$$

and $\hat{g}_s$ and $\hat{g}_\xi$ denote its first d and its last m columns.

Definition 6 (Policy-gradient SGD on raw observations)**.** Start from a deterministic gain $(K_{s,0}, K_{\xi,0})$ with $K_{s,0} \in \mathcal{K}_\gamma$. At iteration $t \geq 0$, draw a fresh batch at the current gain, form $\hat{g}_t$ as in Definition 5, and set

$$K_{\xi,t+1} = K_{\xi,t} - \alpha_t \, \hat{g}_{\xi,t}.$$

Conditionally on the past, the batch consists of N independent episodes under $(K_{s,t}, K_{\xi,t})$, and its baseline may depend on the past. The s-block may be updated by any rule that uses the same data: for example $K_{s,t+1} = K_{s,t} - \alpha_t \hat{g}_{s,t}$, a projected step, or no step. The step size $\alpha_t \geq 0$ is arbitrary. It may be deterministic or adaptive, as long as it is $\mathcal{F}_t$-measurable, where $\mathcal{F}_t$ is generated by the initial gain and by all batches and baselines used before iteration t. If $K_{s,t} \notin \mathcal{K}_\gamma$, the estimator is undefined; the run then stops, and $(K_{s,u}, K_{\xi,u}) := (K_{s,t}, K_{\xi,t})$ for all $u > t$.

Statements.

Theorem 3 (The distractor tax; LQ, one client)**.** *Consider Definitions 4 to 6.*

(a) ***Critic.*** *Let $S \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$ have any joint law with $\operatorname{rank} S = d$ almost surely. The rows may be dependent (for example, states along a trajectory), and the targets heteroskedastic or misspecified. Let $\Xi \in \mathbb{R}^{n \times m}$ have i.i.d. $\mathcal{N}(0,1)$ entries, independent of (S, y), and let $n \geq d + m + 2$. The raw critic $(\hat{\theta}_s, \hat{\theta}_\xi)$ is the least-squares fit of y on $[S\ \Xi]$, and the latent critic $\hat{\theta}_{\mathrm{lat}}$ is that of y on S; let $\mathrm{RSS}_{\mathrm{lat}} := \|y - S\hat{\theta}_{\mathrm{lat}}\|^2$. Let (s', y') be a test pair independent of the sample, with $\mathbb{E}\|s'\|^2 + \mathbb{E}y'^2 < \infty$ and $\Sigma_s := \mathbb{E}[s's'^\top]$, and let $\xi' \sim \mathcal{N}(0, \boldsymbol{I}_m)$ be independent of everything else. Define $\mathcal{R}_{\mathrm{raw}} := \mathbb{E}'(y' - \hat{\theta}_s^\top s' - \hat{\theta}_\xi^\top \xi')^2$ and $\mathcal{R}_{\mathrm{lat}} := \mathbb{E}'(y' - \hat{\theta}_{\mathrm{lat}}^\top s')^2$, where $\mathbb{E}'$ integrates over the test point only. Then, almost surely,*

$$\begin{aligned} \mathbb{E}[\mathcal{R}_{\mathrm{raw}} - \mathcal{R}_{\mathrm{lat}} \mid S, y] &= \mathbb{E}\big[\|\hat{\theta}_\xi\|^2 \mid S, y\big] \Big(1 + \operatorname{tr}\big(\Sigma_s (S^\top S)^{-1}\big)\Big) \\ &= \frac{m \, \mathrm{RSS}_{\mathrm{lat}}}{(n-d)(n-d-m-1)} \Big(1 + \operatorname{tr}\big(\Sigma_s (S^\top S)^{-1}\big)\Big). \end{aligned} \tag{8}$$

Hence, if $\mathbb{E}\mathcal{R}_{\mathrm{lat}} < \infty$,

$$\mathbb{E}\mathcal{R}_{\mathrm{raw}} - \mathbb{E}\mathcal{R}_{\mathrm{lat}} = \frac{m \, \mathbb{E}\big[\mathrm{RSS}_{\mathrm{lat}}\big(1 + \operatorname{tr}(\Sigma_s (S^\top S)^{-1})\big)\big]}{(n-d)(n-d-m-1)} \;\geq\; \frac{m \, \mathbb{E}[\mathrm{RSS}_{\mathrm{lat}}]}{(n-d)(n-d-m-1)}.$$

Suppose moreover that the rows of S are i.i.d. $\mathcal{N}(0, \Sigma_s)$ with $\Sigma_s \succ 0$, that $y = S\theta + e$ with $e \sim \mathcal{N}(0, \varsigma^2 \boldsymbol{I}_n)$ independent of S, and that the test pair follows the same model. Then $\mathbb{E}\mathcal{R}_{\mathrm{lat}} - \varsigma^2 = \varsigma^2 d/(n-d-1)$ and $\mathbb{E}\mathcal{R}_{\mathrm{raw}} - \varsigma^2 = \varsigma^2(d+m)/(n-d-m-1)$. These expected excess risks equal $\varepsilon > 0$ at the sample sizes $n_{\mathrm{lat}}(\varepsilon) = d + 1 + \varsigma^2 d/\varepsilon$ and $n_{\mathrm{raw}}(\varepsilon) = d + m + 1 + \varsigma^2(d+m)/\varepsilon$, so

$$n_{\mathrm{raw}}(\varepsilon) = n_{\mathrm{lat}}(\varepsilon) + m\,(1 + \varsigma^2/\varepsilon).$$

(b) ***Landscape.*** *For all K_s and K_ξ,*

$$J(K_s, K_\xi) = J_s(K_s) + q_\rho(K_s, K_\xi),$$

and $J(K_s, K_\xi) < \infty$ if and only if $K_s \in \mathcal{K}_\gamma$. For $K_s \in \mathcal{K}_\gamma$ there is a symmetric matrix $Z_\rho(K_s) \in \mathbb{R}^{d_u \times d_u}$ such that

$$\begin{aligned} q_\rho(K_s, K_\xi) = \operatorname{tr}\big(K_\xi^\top Z_\rho(K_s) K_\xi\big), \qquad Z_0(K_s) = \frac{M_\gamma(K_s)}{1-\gamma}, \\ \kappa_\rho \, \frac{M_\gamma(K_s)}{1-\gamma} \;\preceq\; Z_\rho(K_s) \;\preceq\; \frac{1}{\kappa_\rho} \, \frac{M_\gamma(K_s)}{1-\gamma}. \end{aligned} \tag{9}$$

In particular, for i.i.d. noise ($\rho = 0$), $J = J_s(K_s) + \operatorname{tr}(K_\xi^\top M_\gamma K_\xi)/(1-\gamma)$. Consequently:

- *$q_\rho \geq \kappa_\rho \lambda_{\min}(R) \|K_\xi\|_F^2/(1-\gamma)$ and $\nabla_{K_\xi} J = 2 Z_\rho(K_s) K_\xi$;*
- *for every $K_s \in \mathcal{K}_\gamma$, the unique minimiser of $J(K_s, \cdot)$ is $K_\xi = 0$, whatever $(A, B, Q, R, \Sigma_w, \mu_0, \gamma, \sigma, \rho)$ are;*
- *$\inf J = J^\star$.*

(c) ***Variance floor for any baseline.*** *Let $K_s \in \mathcal{K}_\gamma$ and $r := \operatorname{rank} K_\xi$. Let $\Pi_\perp$ be the orthogonal projection onto $\ker K_\xi \subseteq \mathbb{R}^m$, and let $v_{\min} := 2\gamma^2 \lambda_{\min}(Q)^2 \operatorname{tr}(\Sigma_w^2)$. For every baseline and every $c \in \mathbb{R}$,*

$$\mathbb{E}\big\|(\hat{g}_\xi - c\,\nabla_{K_\xi} J)\Pi_\perp\big\|_F^2 = \mathbb{E}\|\hat{g}_\xi \Pi_\perp\|_F^2 \;\geq\; F(K_\xi) := \frac{d_u\,(m-r)\,v_{\min}}{\sigma^2 N},$$

with both sides in $[0,\infty]$. Suppose the baseline is action-independent and $\mathbb{E}[(G_1 + |b_1|)\,\|\eta_{\tau_1}^1\|\,\|\xi_{\tau_1}^1\|] < \infty$. Then $\mathbb{E}[\hat{g}_\xi]\Pi_\perp = 0$, so F bounds the variance: $\mathbb{E}\|(\hat{g}_\xi - \mathbb{E}\hat{g}_\xi)\Pi_\perp\|_F^2 \geq F(K_\xi)$.

(d) ***SGD, at every iteration and under every schedule.*** *Let $F_0 := d_u(m-d_u)_+\, v_{\min}/(\sigma^2 N)$. For every $t \geq 1$,*

$$\mathbb{E}J(K_{s,t}, K_{\xi,t}) - J^\star \;\geq\; \big(\mathbb{E}J_s(K_{s,t}) - J^\star\big) + \frac{\kappa_\rho \lambda_{\min}(R)}{1-\gamma}\,\mathbb{E}[\alpha_{t-1}^2]\,F_0, \tag{10}$$

with all terms in $[0,\infty]$. If $K_{s,t-1} \in \mathcal{K}_\gamma$ almost surely, then also $\mathbb{E}\|K_{\xi,t}\|_F^2 \geq \mathbb{E}[\alpha_{t-1}^2]\,F_0$.

(e) ***Latent learner.*** *Let $z = W_s s + \delta\xi$ with $W_s \in \mathbb{R}^{k\times d}$ and $\delta \in \mathbb{R}^{k\times m}$. Let the latent policy be $a = K_z z + \sigma\eta$, with $K_z \in \mathbb{R}^{d_u\times k}$ and $K_z W_s \in \mathcal{K}_\gamma$. It is the raw policy with gains $(K_z W_s, K_z\delta)$. Its ξ-gain is fixed by the encoder and is not a free parameter of policy gradient on K_z, and*

$$\begin{aligned} J_s(K_z W_s) \leq J(K_z W_s, K_z\delta) &\leq J_s(K_z W_s) + \frac{\|M_\gamma(K_z W_s)\|\,\|K_z\delta\|_F^2}{\kappa_\rho(1-\gamma)} \\ &\leq J_s(K_z W_s) + \frac{\|M_\gamma(K_z W_s)\|\,\|K_z\|^2\,\|\delta\|_F^2}{\kappa_\rho(1-\gamma)}. \end{aligned} \tag{11}$$

Suppose $\delta = 0$ and $W_s \in \mathbb{R}^{d\times d}$ is invertible. Fix a nonempty compact convex set $\mathcal{C} \subset \mathbb{R}^{d_u\times d}$ with $\mathcal{C}W_s \subseteq \mathcal{K}_\gamma$ (for instance a small closed ball around a stabilising gain; $\mathcal{K}_\gamma$ is open), and let $\Pi_{\mathcal{C}}$ be the Euclidean projection onto $\mathcal{C}$. Run the projected *latent SGD*

$$K_{z,t+1} = \Pi_{\mathcal{C}}\big(K_{z,t} - \alpha_t \hat{g}_t^z\big), \qquad K_{z,0} \in \mathcal{C},$$

where $\hat{g}_t^z$ is the estimator of Definition 5 with score $\sigma^{-1}\eta z^\top$ and, as in the implemented agent, everything the learner uses is latent:

- *the baseline is $b(t, h_t^z, U)$, a function of the* latent *history $h_t^z = (z_0, a_0, c_0, \ldots, z_t, a_t)$ and of an independent U;*
- *the step size α_t is measurable with respect to the latent filtration $\mathcal{F}_t^z$ generated by $K_{z,0}$ and the latent data (latent histories, costs, baselines) of iterations before t;*
- *the maps b and α, the law of U, $K_{z,0}$ and $\mathcal{C}$ are fixed without reference to m or ρ.*

Then the joint law of $(K_{z,t}W_s)_{t\geq 0}$ does not depend on m or ρ, and $\sup_t J_s(K_{z,t}W_s) \leq \max_{K\in\mathcal{C}W_s} J_s(K) < \infty$. For orthogonal W_s it is the law of the same algorithm run on the distractor-free problem from $K_{z,0}W_s$, with $\mathcal{C}W_s$ in place of $\mathcal{C}$. (Without the projection the statement fails in a useful sense: if $B \neq 0$, a single plain step leaves $\mathcal{K}_\gamma$ with positive probability for any baseline, because given the history $G - b$ is unbounded above through the Gaussian process noise; the cost is then $+\infty$. The projection, or any rule that keeps the gain in a compact subset of $\mathcal{K}_\gamma$, is what makes the latent learner's error finite.)

Corollary 2 (What "converges faster" means)**.** *Fix, without reference to m, the endogenous problem $(A, B, Q, R, \Sigma_w, \mu_0)$, the constants γ, σ, ρ, the batch size N, an initial gain $K_{s,0} \in \mathcal{K}_\gamma$ and a deterministic schedule $(\alpha_t)_{t\geq 0}$. For each m, let $(K_{s,t}^{\mathrm{raw}}, K_{\xi,t}^{\mathrm{raw}})$ be a raw SGD run with m distractors (Definition 6, with any $K_{\xi,0}$ and any baselines). Let $K_{s,t}^{\mathrm{lat}}$ be the gains of the exact latent learner, i.e. the projected latent SGD of Theorem 3(e) with $z = s$ ($W_s = \boldsymbol{I}_d$, $\delta = 0$), a compact convex $\mathcal{C} \subset \mathcal{K}_\gamma$ containing $K_{s,0}$, the same schedule, and latent baselines fixed without reference to m or ρ. Its suboptimality $e_t^{\mathrm{lat}} := \mathbb{E}J_s(K_{s,t}^{\mathrm{lat}}) - J^\star$ is finite and does not depend on m or ρ. Let $C := \kappa_\rho \lambda_{\min}(R)\, d_u\, v_{\min}/\big((1-\gamma)\sigma^2 N\big) > 0$.*

(i) *For every $t \geq 1$,*

$$\mathbb{E}J(K_{s,t}^{\mathrm{raw}}, K_{\xi,t}^{\mathrm{raw}}) - J^\star \;\geq\; \big(\mathbb{E}J_s(K_{s,t}^{\mathrm{raw}}) - J^\star\big) + C\,\alpha_{t-1}^2\,(m-d_u)_+.$$

So at every iteration and under every schedule, the raw learner carries an additive term linear in m, and the exact latent learner carries none.

(ii) Hence the raw learner is strictly worse, $\mathbb{E}J(K^{\mathrm{raw}}_{s,t}, K^{\mathrm{raw}}_{\xi,t}) > \mathbb{E}J_s(K^{\mathrm{lat}}_{s,t})$, for every $t \geq 1$ with $\alpha_{t-1} > 0$ and every $m > m_0(t) := d_u + e^{\mathrm{lat}}_t/(C\alpha^2_{t-1})$.

(iii) In the Gaussian homoskedastic model of Theorem 3(a), the raw critic needs $m(1+\varsigma^2/\varepsilon)$ more samples than the latent critic to reach expected excess risk ε.

(iv) Stationary floor under a constant step. *Suppose also that:*

- $\alpha_t \equiv \alpha > 0$;
- *the baselines are action-independent, with $\mathbb{E}[b^2_{t,i} \mid \mathcal{F}_t] < \infty$;*
- $\mathbb{E}\|s_0\|^3 < \infty$;
- *for some $\bar{L} < \infty$ with $\alpha\bar{L} \leq 1$, almost surely for every t, $\mathrm{spr}(A + BK_{s,t}) < 1$ and $2(1-\gamma)\|Z_\rho(K_{s,t})\| \leq \bar{L}$.*

The last condition holds, for instance, if the s-block is frozen or projected onto a compact set of stable gains; by equation 9, $\bar{L} = 2\kappa_\rho^{-1} \sup_t \|M_\gamma(K_{s,t})\|$ is admissible. Then, for every $t \geq 1$,

$$\mathbb{E}\|K_{\xi,t}\|^2_F \;\geq\; \alpha^2 F_0\, \frac{1-(1-\alpha\bar{L})^{2t}}{1-(1-\alpha\bar{L})^2}, \qquad \liminf_{t\to\infty} \mathbb{E}J(K_{s,t}, K_{\xi,t}) - J^\star \;\geq\; \frac{\kappa_\rho \lambda_{\min}(R)\, \alpha F_0}{2\bar{L}(1-\gamma)}.$$

This is the only sense in which we claim that learning on the latent state converges faster. We prove no convergence rate for either learner, and nothing about the raw learner's s-block. If the schedule may depend on m (for instance, $\alpha_t \propto m^{-1/2}$), the term in (i) need not grow with m; we do not analyse that trade-off.

Corollary 3 (Learned latent)**.** *In Theorem 3(e), let the encoder error δ and the latent gain K_z be random (for example, outputs of the learning algorithm), with $K_z W_s \in \mathcal{K}_\gamma$, $\|K_z\| \leq \bar{k}_z$ and $\|M_\gamma(K_z W_s)\| \leq \bar{\mu}$ almost surely. Then*

$$0 \;\leq\; \mathbb{E}\big[J(K_z W_s, K_z\delta) - J_s(K_z W_s)\big] \;\leq\; \frac{\bar{\mu}\,\bar{k}_z^2}{\kappa_\rho(1-\gamma)}\, \mathbb{E}\|\delta\|^2_F.$$

So the latent's m-dependent cost is proportional to $\|K_z\|^2\mathbb{E}\|\delta\|^2_F$, where $\mathbb{E}\|\delta\|^2_F$ is the encoder's regression error on the noise block. It carries no factor σ^{-2}, unlike Theorem 3(c,d). With K clients and n samples per client, δ is the noise block $\hat{E}$ of Theorem 4(c). There $\|\hat{E}\|_F = \|\tilde{\Gamma}^{-1}\hat{g}_0\|_F + O_p((Kn)^{-1})$, and the leading term satisfies $\mathbb{E}\|\tilde{\Gamma}^{-1}\hat{g}_0\|^2_F \leq \frac{m}{Kn} \cdot \frac{1+\rho}{1-\rho} \cdot \frac{\bar{\sigma}^2_u}{(1-\bar{s})^2\lambda_{\min}(\bar{M}_{\mathrm{tot}})^2}$. The bound applies to $\mathbb{E}\|\delta\|^2_F$ only to leading order: the remainder is controlled in probability, not in mean. The raw learner instead pays at least $C\alpha^2_{t-1}(m-d_u)_+$ at iteration t (Corollary 2(i)).

A.4.1 Lemmas

Lemma 16 (First inverse moments of a Gaussian Gram matrix)**.** *Let $G \in \mathbb{R}^{\nu\times m}$ have i.i.d. $\mathcal{N}(0,1)$ entries with $\nu \geq m+2$. Then $G^\top G$ is almost surely invertible, and*

$$\mathbb{E}\big[(G^\top G)^{-1}\big] = \frac{\boldsymbol{I}_m}{\nu-m-1}, \qquad \mathbb{E}\big[G(G^\top G)^{-2}G^\top\big] = \frac{m}{\nu(\nu-m-1)}\,\boldsymbol{I}_\nu.$$

The first identity is the mean of the inverse of a $\mathcal{W}_m(\nu, \boldsymbol{I}_m)$ Wishart matrix; see, e.g., Muirhead (1982, Ch. 3). We include a proof.

Proof. Write $g_1, \ldots, g_m$ for the columns of G. Given $g_1, \ldots, g_{j-1}$, the column $g_j \sim \mathcal{N}(0, \boldsymbol{I}_\nu)$ lies in their span, a subspace of dimension at most $j-1 < \nu$, with probability 0. Hence G has full column rank and $W := G^\top G \succ 0$ almost surely. Let $G_{-1} := [g_2, \ldots, g_m]$ and $\Pi^{\mathrm{c}} := \boldsymbol{I}_\nu - G_{-1}(G_{-1}^\top G_{-1})^{-1}G_{-1}^\top$ (for $m = 1$, $\Pi^{\mathrm{c}} := \boldsymbol{I}_\nu$). This is an orthogonal projection of rank $\nu - m + 1$ and a function of G_{-1}. By the Schur-complement formula for the inverse of a partitioned positive definite matrix,

$$(W^{-1})_{11} = \big(g_1^\top g_1 - g_1^\top G_{-1}(G_{-1}^\top G_{-1})^{-1}G_{-1}^\top g_1\big)^{-1} = \big(g_1^\top \Pi^{\mathrm{c}} g_1\big)^{-1}.$$

Since g_1 is independent of G_{-1}, conditionally on G_{-1} we have $g_1^\top \Pi^{\mathrm{c}} g_1 \sim \chi^2_{\nu-m+1}$. For $k \geq 3$,

$$\mathbb{E}\big[1/\chi^2_k\big] = \int_0^\infty \frac{1}{x}\,\frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}\,dx = \frac{\Gamma(k/2-1)}{2\,\Gamma(k/2)} = \frac{1}{k-2}.$$

With $k = \nu - m + 1 \geq 3$ this gives $\mathbb{E}(W^{-1})_{11} = 1/(\nu - m - 1)$. Permuting the columns of G does not change its law, so every diagonal entry of W^{-1} has this mean. As $W^{-1} \succ 0$, we have $|(W^{-1})_{ij}| \leq \frac{1}{2}\big((W^{-1})_{ii} + (W^{-1})_{jj}\big)$, so $\mathbb{E}W^{-1}$ is finite. For every orthogonal $O \in \mathbb{R}^{m\times m}$, GO has the law of G, so $\mathbb{E}W^{-1} = O^\top(\mathbb{E}W^{-1})O$. A matrix that commutes with every reflection $\boldsymbol{I} - 2uu^\top$ maps u to a multiple of u for every unit vector u, and is therefore a multiple of the identity. Hence $\mathbb{E}W^{-1} = \boldsymbol{I}_m/(\nu - m - 1)$.

For the second identity let $F := GW^{-2}G^\top \succeq 0$. Then $\|F\| \leq \operatorname{tr} F = \operatorname{tr}(W^{-1})$, which is integrable. For orthogonal $O \in \mathbb{R}^{\nu\times\nu}$, OG has the law of G and $F(OG) = OF(G)O^\top$, so as before $\mathbb{E}F = c\,\boldsymbol{I}_\nu$. The constant is fixed by $c\,\nu = \mathbb{E}\operatorname{tr} F = \mathbb{E}\operatorname{tr}(W^{-1}) = m/(\nu - m - 1)$. □

Lemma 17 (Toeplitz sandwich)**.** *For $n \geq 1$ and $\rho \in [0,1)$, the matrix $\mathcal{T}_n := (\rho^{|i-j|})_{i,j=1}^n$ satisfies $\kappa_\rho \boldsymbol{I}_n \preceq \mathcal{T}_n \preceq \kappa_\rho^{-1}\boldsymbol{I}_n$.*

This is the classical bound for Toeplitz forms, Gray (2006, Lemma 4.1), applied to the symbol f_ρ below; see also Grenander & Szegő (1958). The coefficients $\rho^{|k|}$ of f_ρ are absolutely summable, as that lemma requires. We include the short proof.

Proof. Let $f_\rho(\omega) := \sum_{k\in\mathbb{Z}} \rho^{|k|}e^{ik\omega} = \frac{1-\rho^2}{1-2\rho\cos\omega+\rho^2}$. The series converges uniformly, so $(2\pi)^{-1}\int_{-\pi}^{\pi} e^{i(j-l)\omega} f_\rho(\omega)\,d\omega = \rho^{|j-l|}$. As $\rho \geq 0$, f_ρ is minimal at $\omega = \pi$ and maximal at $\omega = 0$, so $\kappa_\rho \leq f_\rho \leq \kappa_\rho^{-1}$. For $x \in \mathbb{R}^n$,

$$x^\top \mathcal{T}_n x = \sum_{j,l} x_j x_l\,\rho^{|j-l|} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Big|\sum_j x_j e^{ij\omega}\Big|^2 f_\rho(\omega)\,d\omega,$$

and $(2\pi)^{-1}\int_{-\pi}^{\pi} |\sum_j x_j e^{ij\omega}|^2 d\omega = \|x\|^2$. The two bounds on f_ρ give the claim. □

Lemma 18 (Closed forms)**.** *Let $K_s \in \mathcal{K}_\gamma$ and $K_\xi \in \mathbb{R}^{d_u\times m}$. Put $x_t := o_t = [s_t; \xi_t]$,*

$$\mathcal{A} := \begin{bmatrix} A_K & BK_\xi \\ 0 & \rho\boldsymbol{I}_m \end{bmatrix}, \quad \mathcal{A}^0 := \begin{bmatrix} A & 0 \\ 0 & \rho\boldsymbol{I}_m \end{bmatrix}, \quad \mathcal{B} := \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad \bar{Q} := \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix},$$

$\Sigma_\nu := \operatorname{diag}(\Sigma_w, (1-\rho^2)\boldsymbol{I}_m)$, $\Sigma_\zeta := \Sigma_\nu + \sigma^2\mathcal{B}\mathcal{B}^\top$ and $\Sigma_0^x := \operatorname{diag}(\Sigma_0, \boldsymbol{I}_m)$.

(i) $x_{t+1} = \mathcal{A}x_t + \zeta_t$, where $\zeta_t := \sigma\mathcal{B}\eta_t + \nu_t$ and $\nu_t := [w_t; \sqrt{1-\rho^2}\,\epsilon_t]$. The ζ_t are i.i.d. $\mathcal{N}(0, \Sigma_\zeta)$, and ζ_t is independent of $x_0, \dots, x_t$. Moreover $\operatorname{spr}(\mathcal{A}) = \max\{\operatorname{spr}(A_K), \rho\} < \gamma^{-1/2}$.

(ii) The series $\mathcal{P} := \sum_{u\geq 0}\gamma^u(\mathcal{A}^u)^\top(\bar{Q} + K_o^\top R K_o)\mathcal{A}^u$ converges. $\mathcal{P}$ is the unique solution of $\mathcal{P} = \bar{Q} + K_o^\top R K_o + \gamma\mathcal{A}^\top\mathcal{P}\mathcal{A}$, and its upper-left $d\times d$ block is $P_\gamma(K_s)$. The same argument with $\mathcal{A}$ replaced by A_K shows that P_γ is well defined and is the unique solution of its Lyapunov equation.

(iii) For every t, almost surely, $\mathbb{E}\big[\sum_{u\geq 0}\gamma^u c(s_{t+u}, a_{t+u}) \;\big|\; h_t\big] = \mathcal{Q}(x_t, a_t) < \infty$, where

$$\mathcal{Q}(x,a) := x^\top\bar{Q}x + a^\top R a + \gamma\Big[(\mathcal{A}^0x + \mathcal{B}a)^\top\mathcal{P}(\mathcal{A}^0x + \mathcal{B}a) + \operatorname{tr}(\mathcal{P}\Sigma_\nu) + v\Big],$$
$$v := \frac{\sigma^2\operatorname{tr} R + \gamma\operatorname{tr}(\mathcal{P}\Sigma_\zeta)}{1-\gamma}.$$

The value function is $V(x) := \mathbb{E}\big[\sum_u \gamma^u c(s_u, a_u) \;\big|\; x_0 = x\big] = x^\top\mathcal{P}x + v$.

(iv) $J(K_s, K_\xi) = \operatorname{tr}(\mathcal{P}\Sigma_0^x) + v < \infty$. On the open set $\mathcal{K}_\gamma \times \mathbb{R}^{d_u\times m}$, J is infinitely differentiable, and

$$\nabla_{K_o} J = 2E_K\Sigma_\gamma, \qquad E_K := RK_o + \gamma\mathcal{B}^\top\mathcal{P}\mathcal{A},$$
$$\Sigma_\gamma := \sum_{t\geq 0}\gamma^t\mathbb{E}[x_t x_t^\top] = \sum_{u\geq 0}\gamma^u\mathcal{A}^u\Big(\Sigma_0^x + \frac{\gamma}{1-\gamma}\Sigma_\zeta\Big)(\mathcal{A}^u)^\top.$$

Part (iv) is the discounted, noisy, Gaussian-policy analogue of the policy-gradient expression of Fazel et al. (2018, Lemma 1), which is stated for $u_t = -Kx_t$ without noise or discounting.

Proof. (i) Substitute $a_t = K_o x_t + \sigma\eta_t$ into equation 6; note $\mathcal{A} = \mathcal{A}^0 + \mathcal{B}K_o$. The noises $(\eta_t, w_t, \epsilon_t)$ at time t are independent of $x_0, \dots, x_t$, which are functions of s_0, ξ_0 and earlier noises. $\mathcal{A}$ is block upper triangular, so its spectrum is that of A_K together with ρ, and $\rho < 1 \le \gamma^{-1/2}$.

(ii) By Gelfand's formula there are $C < \infty$ and $\theta < 1$ with $\|(\sqrt{\gamma}\mathcal{A})^u\| \le C\theta^u$, so the series converges. The linear map $X \mapsto \gamma\mathcal{A}^\top X\mathcal{A}$ has spectral radius $\gamma\,\mathrm{spr}(\mathcal{A})^2 < 1$, so the Lyapunov equation has exactly one solution, and the series solves it. By induction, the first d columns of $\mathcal{A}^u$ are $[A_K^u; 0]$. Hence, for any C, the upper-left block of $(\mathcal{A}^u)^\top C\mathcal{A}^u$ is $(A_K^u)^\top C_{ss} A_K^u$, and $(\bar{Q} + K_o^\top R K_o)_{ss} = Q + K_s^\top R K_s$.

(iii) Let $C := \bar{Q} + K_o^\top R K_o$. Since η_u is independent of x_u, $\mathbb{E}[c(s_u, a_u) \mid x_0 = x] = \mathrm{tr}(C\,\mathbb{E}[x_u x_u^\top \mid x_0 = x]) + \sigma^2 \,\mathrm{tr}\, R$. By (i), $\mathbb{E}[x_u x_u^\top \mid x_0 = x] = \mathcal{A}^u x x^\top (\mathcal{A}^u)^\top + \sum_{i<u} \mathcal{A}^i \Sigma_\zeta (\mathcal{A}^i)^\top$. The costs are nonnegative, so sums and expectations may be exchanged (Tonelli), and $\sum_u \gamma^u \sum_{i<u} f(i) = \sum_i f(i)\gamma^{i+1}/(1-\gamma)$ for $f \ge 0$. This gives $V(x) = x^\top \mathcal{P} x + \frac{\gamma}{1-\gamma}\,\mathrm{tr}(\mathcal{P}\Sigma_\zeta) + \frac{\sigma^2 \,\mathrm{tr}\, R}{1-\gamma} = x^\top \mathcal{P} x + v$. Given h_t, the cost at time t is $x_t^\top \bar{Q} x_t + a_t^\top R a_t$, and $x_{t+1} = \mathcal{A}^0 x_t + \mathcal{B} a_t + \nu_t$ with $\nu_t \sim \mathcal{N}(0, \Sigma_\nu)$ independent of h_t. From x_{t+1} on, the process is the chain of (i) started at x_{t+1} and driven by noise independent of (h_t, ν_t). Hence the conditional expected return is $x_t^\top \bar{Q} x_t + a_t^\top R a_t + \gamma\,\mathbb{E}[V(\mathcal{A}^0 x_t + \mathcal{B} a_t + \nu_t) \mid h_t] = \mathcal{Q}(x_t, a_t)$.

(iv) $J = \mathbb{E}V(x_0)$, and $\mathbb{E}[x_0 x_0^\top] = \Sigma_0^x$ because ξ_0 is independent of s_0 with mean 0 and covariance $\boldsymbol{I}_m$. The solution of the Lyapunov equation is $\mathrm{vec}\,\mathcal{P} = (\boldsymbol{I} - \gamma\mathcal{A}^\top \otimes \mathcal{A}^\top)^{-1} \mathrm{vec}(\bar{Q} + K_o^\top R K_o)$. This is a rational function of K_o whose denominator does not vanish on $\mathcal{K}_\gamma \times \mathbb{R}^{d_u \times m}$, since $\gamma\,\mathrm{spr}(\mathcal{A})^2 < 1$ by (i). The set is open because the spectral radius is continuous, so J is C^∞ there. Differentiate the Lyapunov equation in the direction dK_o, with $d\mathcal{A} = \mathcal{B}\,dK_o$:

$$d\mathcal{P} = dK_o^\top E_K + E_K^\top dK_o + \gamma\mathcal{A}^\top d\mathcal{P}\,\mathcal{A}, \qquad \text{so} \qquad d\mathcal{P} = \sum_{u \ge 0} \gamma^u (\mathcal{A}^u)^\top \big(dK_o^\top E_K + E_K^\top dK_o\big)\mathcal{A}^u.$$

Since Σ_ζ does not depend on K_o, $dJ = \mathrm{tr}\,\big(d\mathcal{P}\,(\Sigma_0^x + \frac{\gamma}{1-\gamma}\Sigma_\zeta)\big) = \mathrm{tr}\,\big((dK_o^\top E_K + E_K^\top dK_o)\Sigma_\gamma\big) = 2\,\mathrm{tr}(dK_o^\top E_K \Sigma_\gamma)$, which is the claim. The two expressions for Σ_γ agree because $\mathbb{E}[x_t x_t^\top] = \mathcal{A}^t \Sigma_0^x (\mathcal{A}^t)^\top + \sum_{i<t} \mathcal{A}^i \Sigma_\zeta (\mathcal{A}^i)^\top$ by (i). □

Lemma 19 (Conditional variance of the return)**.** *Let $K_s \in \mathcal{K}_\gamma$ and fix $t \ge 0$. Let $G_t := \sum_{u \ge 0} \gamma^u c(s_{t+u}, a_{t+u})$, let U be independent of the episode, and let $\mathcal{H}_t := \sigma(U, h_t)$. For every $\mathcal{H}_t$-measurable b, almost surely,*

$$\mathbb{E}\big[(G_t - b)^2 \;\big|\; \mathcal{H}_t\big] \;\ge\; 2\gamma^2 \,\mathrm{tr}\,\big((P_\gamma(K_s)\Sigma_w)^2\big) \;\ge\; v_{\min} = 2\gamma^2 \lambda_{\min}(Q)^2 \,\mathrm{tr}(\Sigma_w^2).$$

Proof. Let $\mathcal{H}_t' := \sigma(\mathcal{H}_t, w_t)$, and partition $\mathcal{P}$ into blocks $\mathcal{P}_{ss}$, $\mathcal{P}_{s\xi}$, $\mathcal{P}_{\xi\xi}$, with $\mathcal{P}_{ss} = P_\gamma := P_\gamma(K_s)$ by Lemma 18(ii). Given $\mathcal{H}_t'$, the next state $s_{t+1} = \mu_t + w_t$ is known, where $\mu_t := As_t + Ba_t$. Moreover $\xi_{t+1} = \rho\xi_t + \sqrt{1-\rho^2}\epsilon_t$ with ϵ_t independent of $\mathcal{H}_t'$, and the process after $t+1$ is driven by noise independent of $(\mathcal{H}_t', \epsilon_t)$. By Lemma 18(iii), expanding $V(x_{t+1}) = x_{t+1}^\top \mathcal{P} x_{t+1} + v$ and integrating out ϵ_t,

$$\mathbb{E}[G_t \mid \mathcal{H}_t'] = c(s_t, a_t) + \gamma\,\mathbb{E}\big[V(x_{t+1}) \;\big|\; \mathcal{H}_t'\big] = C_t + \gamma\big(w_t^\top P_\gamma w_t + 2 w_t^\top h_t\big).$$

Here $h_t := P_\gamma \mu_t + \rho\,\mathcal{P}_{s\xi}\xi_t$ and C_t are $\mathcal{H}_t$-measurable. This conditional mean is finite. Hence, for the real random variable $G_t \ge 0$ and the $\mathcal{H}_t'$-measurable b, $\mathbb{E}[(G_t - b)^2 \mid \mathcal{H}_t'] \ge (\mathbb{E}[G_t \mid \mathcal{H}_t'] - b)^2$. To see this, pass to a regular conditional distribution given $\mathcal{H}_t'$: there b is a constant and G_t has a finite mean. For any real Y with finite mean and any constant β, $\mathbb{E}(Y - \beta)^2 \ge (\mathbb{E}Y - \beta)^2$; this is trivial if $\mathbb{E}Y^2 = \infty$ and follows from the variance decomposition otherwise. Now take $\mathbb{E}[\cdot \mid \mathcal{H}_t]$ and use that the conditional mean minimises the conditional mean squared deviation over $\mathcal{H}_t$-measurable b:

$$\mathbb{E}[(G_t - b)^2 \mid \mathcal{H}_t] \;\ge\; \mathrm{Var}\big(\mathbb{E}[G_t \mid \mathcal{H}_t'] \;\big|\; \mathcal{H}_t\big) = \gamma^2\big(2\,\mathrm{tr}((P_\gamma\Sigma_w)^2) + 4h_t^\top \Sigma_w h_t\big).$$

The equality holds because $w_t \sim \mathcal{N}(0, \Sigma_w)$ is independent of $\mathcal{H}_t$, and because of three moment identities for $w \sim \mathcal{N}(0, \Sigma_w)$:

- $\mathrm{Var}(w^\top P_\gamma w) = 2\,\mathrm{tr}((P_\gamma \Sigma_w)^2)$: write $w = \Sigma_w^{1/2} g$ and $\Sigma_w^{1/2} P_\gamma \Sigma_w^{1/2} = \sum_j \lambda_j u_j u_j^\top$, so that $w^\top P_\gamma w = \sum_j \lambda_j (u_j^\top g)^2$ is a sum of independent terms with variances $2\lambda_j^2$;
- $\mathrm{Var}(2w^\top h) = 4h^\top \Sigma_w h$;
- $\mathrm{Cov}(w^\top P_\gamma w, w^\top h) = 0$, by the symmetry $w \mapsto -w$.

Dropping $4h_t^\top \Sigma_w h_t \geq 0$ gives the first bound. For the second, $P_\gamma \succeq Q \succeq \lambda_{\min}(Q)\boldsymbol{I}$ gives $X := \Sigma_w^{1/2} P_\gamma \Sigma_w^{1/2} \succeq Y := \lambda_{\min}(Q)\Sigma_w \succeq 0$. Then $\mathrm{tr}(X^2) - \mathrm{tr}(Y^2) = \mathrm{tr}\big((X-Y)(X+Y)\big) = \mathrm{tr}\big((X-Y)^{1/2}(X+Y)(X-Y)^{1/2}\big) \geq 0$, and $\mathrm{tr}((P_\gamma \Sigma_w)^2) = \mathrm{tr}(X^2)$. □

Lemma 20 (Unbiasedness)**.** *Suppose* $\mathrm{spr}(A + BK_s) < 1$ *and* $\mathbb{E}\|s_0\|^3 < \infty$*, and let the baseline be action-independent with* $\mathbb{E}[b_1^2 \mid U] < \infty$*. Then* $\mathbb{E}[\|\hat{g}\|_F \mid U] < \infty$ *and* $\mathbb{E}[\hat{g} \mid U] = (1-\gamma)\nabla_{K_o} J(K_s, K_\xi)$*; in particular* $\mathbb{E}[\hat{g}_\xi \mid U] = 2(1-\gamma) Z_\rho(K_s) K_\xi$.

This is the instance of the policy-gradient theorem (Sutton et al., 1999, Thm. 1) that we need. We prove it because the state space is unbounded.

Proof. Work conditionally on U, which is independent of the batch. The N summands of $\hat{g}$ are then i.i.d., so take $N = 1$ and drop the index i. Write $b_t := b(t, h_t, U)$ and let G_t be the return from time t. Since τ is independent of the episode, it suffices to show two things: that $\mathbb{E}[(G_t - b_t)\sigma^{-1}\eta_t x_t^\top] = 2E_K \mathbb{E}[x_t x_t^\top]$ for each fixed t, and that $\sum_t (1-\gamma)\gamma^t\, \mathbb{E}\big[(G_t + |b_t|)\|\eta_t\|\|x_t\|\big] < \infty$. Summing against $(1-\gamma)\gamma^t$ then gives $\mathbb{E}\hat{g} = 2(1-\gamma)E_K \Sigma_\gamma = (1-\gamma)\nabla_{K_o} J$ by Lemma 18(iv), and the ξ-block follows from Theorem 3(b), whose proof does not use this lemma.

Moments. Here $\mathrm{spr}(\mathcal{A}) = \max\{\mathrm{spr}(A_K), \rho\} < 1$, so $\|\mathcal{A}^u\| \leq C\theta^u$ for some $C < \infty$ and $\theta < 1$. By Lemma 18(i), $x_t = \mathcal{A}^t x_0 + \sum_{i<t} \mathcal{A}^{t-1-i}\zeta_i$. The sum is centred Gaussian with covariance $\preceq C^2\|\Sigma_\zeta\|(1-\theta^2)^{-1}\boldsymbol{I}$, and $\mathbb{E}\|x_0\|^3 \leq 4(\mathbb{E}\|s_0\|^3 + \mathbb{E}\|\xi_0\|^3) < \infty$. Hence $\sup_t \mathbb{E}\|x_t\|^3 < \infty$.

Baseline term. Let $h_t^- := (h_{t-1}, o_t)$. The baseline b_t is a function of (t, h_t^-, U), and η_t is independent of (h_t^-, U) with mean zero. So $\mathbb{E}[b_t\, \eta_t x_t^\top] = \mathbb{E}\big[b_t\, \mathbb{E}[\eta_t \mid h_t^-, U]\, x_t^\top\big] = 0$. Integrability and summability follow from Cauchy–Schwarz over t and ω jointly:

$$\sum_t (1-\gamma)\gamma^t\, \mathbb{E}\big[|b_t|\,\|\eta_t\|\|x_t\|\big] \leq \Big(\sum_t (1-\gamma)\gamma^t \mathbb{E} b_t^2\Big)^{1/2} \Big(\sum_t (1-\gamma)\gamma^t\, d_u \mathbb{E}\|x_t\|^2\Big)^{1/2}.$$

The first factor is $\mathbb{E}[b_1^2 \mid U]^{1/2} < \infty$, and the second is finite by the moment bound, because $\mathbb{E}[\|\eta_t\|^2\|x_t\|^2] = d_u \mathbb{E}\|x_t\|^2$.

Return term. By Lemma 18(iii), $\mathbb{E}[G_t \mid h_t] = \mathcal{Q}(x_t, a_t)$, and $0 \leq \mathcal{Q}(x,a) \leq c_1(1 + \|x\|^2 + \|a\|^2)$ for a constant c_1. With $a_t = K_o x_t + \sigma\eta_t$ and the moment bound, $\mathbb{E}[G_t\|\eta_t\|\|x_t\|] = \mathbb{E}[\mathcal{Q}(x_t, a_t)\|\eta_t\|\|x_t\|]$ is bounded uniformly in t, so its weighted sum over t is finite. Hence $\mathbb{E}[G_t \eta_t x_t^\top] = \mathbb{E}[\mathcal{Q}(x_t, K_o x_t + \sigma\eta_t)\, \eta_t x_t^\top]$. Condition on x_t: then $\eta_t \sim \mathcal{N}(0, \boldsymbol{I}_{d_u})$ is independent of x_t, and $f(\eta) := \mathcal{Q}(x_t, K_o x_t + \sigma\eta)$ is a quadratic polynomial in η. Stein's identity (Stein, 1981, Lemma 1) holds for a standard normal variable and an absolutely continuous function with integrable derivative. Applied to each coordinate $\eta_{t,j}$, conditionally on the other coordinates, it gives $\mathbb{E}[f(\eta_t)\eta_t \mid x_t] = \mathbb{E}[\nabla f(\eta_t) \mid x_t] = \sigma\, \mathbb{E}[\nabla_a \mathcal{Q}(x_t, a_t) \mid x_t]$. Since $\nabla_a \mathcal{Q}(x,a) = 2Ra + 2\gamma\mathcal{B}^\top \mathcal{P}(\mathcal{A}^0 x + \mathcal{B}a)$ and $\mathcal{A}^0 + \mathcal{B}K_o = \mathcal{A}$, we get $\mathbb{E}[\nabla_a \mathcal{Q}(x_t, a_t) \mid x_t] = 2(RK_o + \gamma\mathcal{B}^\top \mathcal{P}\mathcal{A})x_t = 2E_K x_t$. Therefore $\mathbb{E}[G_t \sigma^{-1}\eta_t x_t^\top] = 2E_K \mathbb{E}[x_t x_t^\top]$. □

A.4.2 Proof of Theorem 3

(a) Critic. Write $\Pi_S := S(S^\top S)^{-1} S^\top$ and $\Pi_S^{\mathsf{c}} := \boldsymbol{I}_n - \Pi_S$.

Partialling out. Fix a measurable map $S \mapsto U_S \in \mathbb{R}^{n \times (n-d)}$ whose columns form an orthonormal basis of the range of Π_S^{c} (for example, Gram–Schmidt applied to the columns of Π_S^{c}). Then $\Pi_S^{\mathsf{c}} = U_S U_S^\top$, and conditionally on (S, y) the matrix $\tilde{G} := U_S^\top \Xi \in \mathbb{R}^{(n-d)\times m}$ has i.i.d. $\mathcal{N}(0,1)$ entries, because $\mathrm{vec}\,\tilde{G} = (\boldsymbol{I}_m \otimes U_S^\top)\,\mathrm{vec}\,\Xi$. As $n - d \geq m + 2$, $\Xi^\top \Pi_S^{\mathsf{c}} \Xi = \tilde{G}^\top \tilde{G} \succ 0$ almost surely (Lemma 16), so $[S\ \Xi]$ has full column rank and both least-squares fits are unique. The normal equations of the raw fit are $S^\top S\hat{\theta}_s + S^\top \Xi \hat{\theta}_\xi = S^\top y$ and $\Xi^\top S\hat{\theta}_s + \Xi^\top \Xi \hat{\theta}_\xi = \Xi^\top y$. The first gives

$\hat\theta_s = \hat\theta_{\mathrm{lat}} + \Delta$ with $\Delta := -(S^\top S)^{-1} S^\top \Xi \hat\theta_\xi$. Substituting into the second gives $\Xi^\top \Pi_S^{\mathrm{c}} \Xi\, \hat\theta_\xi = \Xi^\top \Pi_S^{\mathrm{c}} y$, that is, $\hat\theta_\xi = (\tilde G^\top \tilde G)^{-1} \tilde G^\top \tilde y$ with $\tilde y := U_S^\top y$ and $\|\tilde y\|^2 = y^\top \Pi_S^{\mathrm{c}} y = \mathrm{RSS}_{\mathrm{lat}}$. This is the Frisch–Waugh–Lovell theorem (Frisch & Waugh, 1933; Lovell, 1963).

The moment. By Lemma 16 with $\nu = n - d$,

$$\mathbb{E}\big[\|\hat\theta_\xi\|^2 \,\big|\, S, y\big] = \tilde y^\top\, \mathbb{E}\big[\tilde G (\tilde G^\top \tilde G)^{-2} \tilde G^\top \,\big|\, S, y\big]\, \tilde y = \frac{m\, \mathrm{RSS}_{\mathrm{lat}}}{(n-d)(n-d-m-1)}.$$

Risk difference. For fixed coefficients, $\mathbb{E}'(y' - \theta^\top s' - \vartheta^\top \xi')^2 = \mathbb{E}'(y' - \theta^\top s')^2 + \|\vartheta\|^2$, because ξ' is independent of (s', y') with mean 0 and covariance $\boldsymbol{I}_m$. For fixed θ and Δ, $\mathbb{E}'(y' - (\theta+\Delta)^\top s')^2 - \mathbb{E}'(y' - \theta^\top s')^2 = -2\ell(\theta)^\top \Delta + \Delta^\top \Sigma_s \Delta$, where $\ell(\theta) := \mathbb{E}'[(y' - \theta^\top s') s']$. Hence

$$\mathcal{R}_{\mathrm{raw}} - \mathcal{R}_{\mathrm{lat}} = \|\hat\theta_\xi\|^2 - 2\ell(\hat\theta_{\mathrm{lat}})^\top \Delta + \Delta^\top \Sigma_s \Delta. \tag{12}$$

Conditioning. Conditionally on (S, y), $\operatorname{vec}\Xi \sim \mathcal{N}(0, \boldsymbol{I})$. So $\Pi_S \Xi$ and $\Pi_S^{\mathrm{c}} \Xi$ are jointly Gaussian and uncorrelated, since $\operatorname{Cov}(\operatorname{vec}\Pi_S\Xi, \operatorname{vec}\Pi_S^{\mathrm{c}}\Xi) = \boldsymbol{I}_m \otimes \Pi_S \Pi_S^{\mathrm{c}} = 0$; hence they are independent. Now $\hat\theta_\xi = ((\Pi_S^{\mathrm{c}}\Xi)^\top \Pi_S^{\mathrm{c}}\Xi)^{-1} (\Pi_S^{\mathrm{c}}\Xi)^\top y$ is a function of $(S, y, \Pi_S^{\mathrm{c}}\Xi)$, while $S^\top \Xi = S^\top \Pi_S \Xi$ is a function of $(S, \Pi_S \Xi)$. So, conditionally on $\mathcal{G} := \sigma(S, y, \Pi_S^{\mathrm{c}}\Xi)$, the columns of $S^\top \Xi$ are i.i.d. $\mathcal{N}(0, S^\top S)$. Therefore $\mathbb{E}[\Delta \mid \mathcal{G}] = 0$. With $C := (S^\top S)^{-1} \Sigma_s (S^\top S)^{-1}$, the (j,l) entry of $\mathbb{E}[(S^\top \Xi)^\top C (S^\top \Xi) \mid \mathcal{G}]$ is $\delta_{jl} \operatorname{tr}(C S^\top S)$, so

$$\mathbb{E}[\Delta^\top \Sigma_s \Delta \mid \mathcal{G}] = \hat\theta_\xi^\top\, \mathbb{E}\big[(S^\top \Xi)^\top C (S^\top \Xi) \,\big|\, \mathcal{G}\big]\, \hat\theta_\xi = \operatorname{tr}\left(\Sigma_s (S^\top S)^{-1}\right) \|\hat\theta_\xi\|^2.$$

All terms of equation 12 are conditionally integrable given (S, y). The first and the last are nonnegative with finite conditional means by the two displays. The middle one is bounded by $2\|\ell(\hat\theta_{\mathrm{lat}})\|\, \|\Delta\|$, where $\ell(\hat\theta_{\mathrm{lat}})$ is $\sigma(S,y)$-measurable and finite, and $\mathbb{E}[\|\Delta\|^2 \mid S, y] = \operatorname{tr}((S^\top S)^{-1})\, \mathbb{E}[\|\hat\theta_\xi\|^2 \mid S, y] < \infty$ by the same computation with Σ_s replaced by $\boldsymbol{I}_d$. Taking $\mathbb{E}[\cdot \mid \mathcal{G}]$ in equation 12, and then $\mathbb{E}[\cdot \mid S, y]$, gives the first equality of equation 8; the second is the moment display. If $\mathbb{E}\mathcal{R}_{\mathrm{lat}} < \infty$, then $\mathbb{E}\mathcal{R}_{\mathrm{raw}} = \mathbb{E}\mathcal{R}_{\mathrm{lat}} + \mathbb{E}\big[\mathbb{E}[\mathcal{R}_{\mathrm{raw}} - \mathcal{R}_{\mathrm{lat}} \mid S, y]\big]$, where the last term lies in $[0, \infty]$. The lower bound follows from $\operatorname{tr}(\Sigma_s (S^\top S)^{-1}) \geq 0$.

Gaussian homoskedastic case. Consider a design $X \in \mathbb{R}^{n \times p}$ with i.i.d. rows $\mathcal{N}(0, \Sigma)$, $\Sigma \succ 0$, $n \geq p + 2$. Let $y = X\beta + e$ with $e \sim \mathcal{N}(0, \varsigma^2 \boldsymbol{I}_n)$ independent of X, and take a test pair from the same model. The least-squares fit $\hat\beta$ has $\mathbb{E}'(y' - \hat\beta^\top x')^2 - \varsigma^2 = (\hat\beta - \beta)^\top \Sigma (\hat\beta - \beta)$ and $\mathbb{E}[(\hat\beta - \beta)^\top \Sigma (\hat\beta - \beta) \mid X] = \varsigma^2 \operatorname{tr}(\Sigma (X^\top X)^{-1})$. Writing $X = X_0 \Sigma^{1/2}$ with X_0 standard Gaussian, $\mathbb{E}(X^\top X)^{-1} = \Sigma^{-1/2} \mathbb{E}(X_0^\top X_0)^{-1} \Sigma^{-1/2} = \Sigma^{-1}/(n - p - 1)$ by Lemma 16 with $\nu = n$. So the expected excess risk is $\varsigma^2 p/(n - p - 1)$. The latent critic is the case $p = d$. The raw critic is the case $p = d + m$, with design $[S\ \Xi]$ (rows $\mathcal{N}(0, \operatorname{diag}(\Sigma_s, \boldsymbol{I}_m))$), $\beta = (\theta, 0)$ and test point (s', ξ', y'). Solving $\varsigma^2 p/(n - p - 1) = \varepsilon$ gives $n = p + 1 + \varsigma^2 p/\varepsilon$, and subtracting the two cases gives $m(1 + \varsigma^2/\varepsilon)$.

(b) Landscape. *Decomposition.* For each t, (s_t^ξ, a_t^ξ) is a linear function of $(\xi_0, \ldots, \xi_t)$. It is therefore centred and square-integrable, and independent of $(s_t^{\mathrm{e}}, a_t^{\mathrm{e}})$, which is a function of $(s_0, (w_u), (\eta_u))$ with finite second moments. Expanding the quadratic cost, the cross term $2\big((s_t^{\mathrm{e}})^\top Q s_t^\xi + (a_t^{\mathrm{e}})^\top R a_t^\xi\big)$ has mean zero. So $\mathbb{E}c(s_t, a_t) = \mathbb{E}c(s_t^{\mathrm{e}}, a_t^{\mathrm{e}}) + \mathbb{E}c(s_t^\xi, a_t^\xi)$ for every t. Summing with weights γ^t over nonnegative terms gives $J = J_s(K_s) + q_\rho$, since $(s^{\mathrm{e}}, a^{\mathrm{e}})$ is the distractor-free trajectory.

Finiteness. If $K_s \in \mathcal{K}_\gamma$, then $J_s(K_s) = J(K_s, 0) < \infty$ by Lemma 18(iv), and $q_\rho < \infty$ by the sandwich below. If $K_s \notin \mathcal{K}_\gamma$, write $s_t^{\mathrm{e}} = A_K^t s_0 + \sum_{i<t} A_K^{t-1-i}(B\sigma\eta_i + w_i)$. The w-sum is centred and independent of the rest, so $\mathbb{E}[s_t^{\mathrm{e}}(s_t^{\mathrm{e}})^\top] \succeq \sum_{i<t} A_K^i \Sigma_w (A_K^i)^\top$ and $\mathbb{E}c(s_t^{\mathrm{e}}, a_t^{\mathrm{e}}) \geq \lambda_{\min}(Q)\lambda_{\min}(\Sigma_w) \sum_{i<t} \|A_K^i\|_F^2$. As $\|A_K^i\|_F \geq \operatorname{spr}(A_K)^i \geq \gamma^{-i/2}$, we get $J_s(K_s) \geq \lambda_{\min}(Q)\lambda_{\min}(\Sigma_w) \sum_{t \geq 1} \sum_{i<t} \gamma^{t-i} = \infty$. Hence $J \geq J_s = \infty$.

I.i.d. noise. Let $\rho = 0$, $K_s \in \mathcal{K}_\gamma$, $u_t := K_\xi \xi_t$ (i.i.d. $\mathcal{N}(0, S_u)$ with $S_u := K_\xi K_\xi^\top$), and $\tilde Q := Q + K_s^\top R K_s$. Then $s_t^\xi = \sum_{j<t} A_K^{t-1-j} B u_j$ and $c(s_t^\xi, a_t^\xi) = (s_t^\xi)^\top \tilde Q s_t^\xi + 2(s_t^\xi)^\top K_s^\top R u_t + u_t^\top R u_t$. The middle term has mean zero, since s_t^ξ depends only on $u_0, \ldots, u_{t-1}$. The independent centred u_j

give $\mathbb{E}(s_t^\xi)^\top \tilde{Q} s_t^\xi = \sum_{l<t} \operatorname{tr}\left(B^\top (A_K^l)^\top \tilde{Q} A_K^l B S_u\right)$. Summing over t with weights γ^t, and using $\sum_t \gamma^t \sum_{l<t} f(l) = \sum_l f(l)\gamma^{l+1}/(1-\gamma)$,

$$q_0 = \frac{\operatorname{tr}(RS_u) + \gamma \operatorname{tr}(B^\top P_\gamma B S_u)}{1-\gamma} = \frac{\operatorname{tr}(K_\xi^\top M_\gamma K_\xi)}{1-\gamma}.$$

AR(1) noise: structure and sandwich. Let $K_s \in \mathcal{K}_\gamma$, fix $T \geq 1$, and write $q_\rho^{(T)} := \sum_{t<T} \gamma^t \mathbb{E} c(s_t^\xi, a_t^\xi)$. Let $\mathbf{u} := (u_0; \ldots; u_{T-1}) \in \mathbb{R}^{Td_u}$ with $u_t := K_\xi \xi_t$. Each (s_t^ξ, a_t^ξ), $t < T$, is a fixed linear function of $\mathbf{u}$, and the cost is a nonnegative quadratic form. So $\sum_{t<T} \gamma^t c(s_t^\xi, a_t^\xi) = \mathbf{u}^\top \mathbf{C}_T \mathbf{u}$ for a symmetric $\mathbf{C}_T \succeq 0$ that depends on $(A, B, Q, R, K_s, \gamma)$ but not on K_ξ or ρ. Stationarity gives $\mathbb{E}[u_i u_j^\top] = \rho^{|i-j|} S_u$, so $\operatorname{Cov}(\mathbf{u}) = \mathcal{T}_T \otimes S_u$ and $q_\rho^{(T)} = \operatorname{tr}\left(\mathbf{C}_T(\mathcal{T}_T \otimes S_u)\right)$.

First, this is linear in the symmetric matrix S_u. So there is a symmetric $Z_\rho^{(T)}$ with $q_\rho^{(T)} = \operatorname{tr}(Z_\rho^{(T)} S_u) = \operatorname{tr}(K_\xi^\top Z_\rho^{(T)} K_\xi)$.

Second, $\mathcal{T}_T \otimes S_u = (\boldsymbol{I} \otimes S_u^{1/2})(\mathcal{T}_T \otimes \boldsymbol{I})(\boldsymbol{I} \otimes S_u^{1/2})$, so $q_\rho^{(T)} = \operatorname{tr}\left(\tilde{\mathbf{C}}_T(\mathcal{T}_T \otimes \boldsymbol{I})\right)$ with $\tilde{\mathbf{C}}_T := (\boldsymbol{I} \otimes S_u^{1/2})\mathbf{C}_T(\boldsymbol{I} \otimes S_u^{1/2}) \succeq 0$. By Lemma 17, $\kappa_\rho \boldsymbol{I} \preceq \mathcal{T}_T \otimes \boldsymbol{I} \preceq \kappa_\rho^{-1} \boldsymbol{I}$. For $X \succeq 0$ and $\lambda \boldsymbol{I} \preceq Y \preceq \Lambda \boldsymbol{I}$, $\lambda \operatorname{tr} X \leq \operatorname{tr}(XY) \leq \Lambda \operatorname{tr} X$ (apply $\operatorname{tr}(X^{1/2}(Y - \lambda \boldsymbol{I})X^{1/2}) \geq 0$). Since $\operatorname{tr} \tilde{\mathbf{C}}_T = \operatorname{tr}(\mathbf{C}_T(\boldsymbol{I} \otimes S_u)) = q_0^{(T)}$,

$$\kappa_\rho \, q_0^{(T)} \;\leq\; q_\rho^{(T)} \;\leq\; \kappa_\rho^{-1} q_0^{(T)}.$$

Letting $T \to \infty$ (monotone convergence) gives $\kappa_\rho q_0 \leq q_\rho \leq \kappa_\rho^{-1} q_0$ for every K_ξ. In particular $q_\rho < \infty$.

For $v \in \mathbb{R}^{d_u}$, take $K_\xi = v e_1^\top$. Then $v^\top Z_\rho^{(T)} v = q_\rho^{(T)}(K_s, v e_1^\top)$ is nonnegative, nondecreasing in T, and bounded by $\kappa_\rho^{-1} v^\top M_\gamma v/(1-\gamma)$. By polarisation, $Z_\rho^{(T)}$ converges to a symmetric $Z_\rho \succeq 0$, and $q_\rho = \lim_T q_\rho^{(T)} = \operatorname{tr}(K_\xi^\top Z_\rho K_\xi)$. With $K_\xi = v e_1^\top$, the scalar sandwich and the i.i.d. formula give $\kappa_\rho v^\top M_\gamma v/(1-\gamma) \leq v^\top Z_\rho v \leq \kappa_\rho^{-1} v^\top M_\gamma v/(1-\gamma)$, which is the Loewner sandwich in equation 9. For $\rho = 0$, both sides coincide and $Z_0 = M_\gamma/(1-\gamma)$.

Consequences. $q_\rho \geq \kappa_\rho \operatorname{tr}(K_\xi^\top M_\gamma K_\xi)/(1-\gamma) \geq \kappa_\rho \lambda_{\min}(R) \|K_\xi\|_F^2/(1-\gamma)$ because $M_\gamma \succeq R$. The map $K_\xi \mapsto J_s(K_s) + \operatorname{tr}(K_\xi^\top Z_\rho K_\xi)$ has gradient $2 Z_\rho K_\xi$. It exceeds its value at $K_\xi = 0$ whenever $K_\xi \neq 0$. The minimiser $K_\xi = 0$ does not depend on the constants of the model. Finally, $J = \infty$ off $\mathcal{K}_\gamma$, and on $\mathcal{K}_\gamma$, $J \geq J_s$ with equality at $K_\xi = 0$; so $\inf J = J^\star$.

(c) Floor. By (b), $\nabla_{K_\xi} J \, \Pi_\perp = 2 Z_\rho K_\xi \Pi_\perp = 0$, which gives the equality for every c. Write $\hat{g}_\xi \Pi_\perp = N^{-1} \sum_i X_i$ with $X_i := \sigma^{-1}(G_i - b_i)\, \eta_{\tau_i}^i (\Pi_\perp \xi_{\tau_i}^i)^\top$. Since $\|uv^\top\|_F = \|u\|\|v\|$,

$$\mathbb{E}\|X_1\|_F^2 = \sigma^{-2} \sum_{t \geq 0} (1-\gamma)\gamma^t \, \mathbb{E}\big[(G_t - b(t, h_t, U))^2 \, \|\eta_t\|^2 \, \|\Pi_\perp \xi_t\|^2\big],$$

because τ_1 is independent of the episode. Here the index 1 is dropped, and G_t is the return from time t as in Lemma 19. Both η_t and ξ_t are $\mathcal{H}_t$-measurable, with $\mathcal{H}_t = \sigma(U, h_t)$ ($\eta_t = \sigma^{-1}(a_t - K_o o_t)$). By Lemma 19, $\mathbb{E}[(G_t - b)^2 \mid \mathcal{H}_t] \geq v_{\min}$, so each expectation is at least $v_{\min} \mathbb{E}[\|\eta_t\|^2 \|\Pi_\perp \xi_t\|^2] = v_{\min}\, d_u (m - r)$. The last equality holds because $\eta_t \sim \mathcal{N}(0, \boldsymbol{I}_{d_u})$ and $\xi_t \sim \mathcal{N}(0, \boldsymbol{I}_m)$ are independent (the latter by stationarity) and $\operatorname{tr} \Pi_\perp = m - r$. Hence $\mathbb{E}\|X_1\|_F^2 \geq d_u (m-r) v_{\min}/\sigma^2$.

Conditionally on U, the X_i are i.i.d. If $\mathbb{E}\|X_1\|_F^2 < \infty$, then $\mathbb{E}\|N^{-1}\sum_i X_i\|_F^2 = N^{-1}\mathbb{E}\|X_1\|_F^2 + (1 - N^{-1})\|\mathbb{E} X_1\|_F^2 \geq N^{-1}\mathbb{E}\|X_1\|_F^2$. Otherwise both sides are infinite: conditionally on $X_2, \ldots, X_N$, $\mathbb{E}\|X_1 + Y\|_F^2 = \infty$ for every fixed Y. This proves the floor.

Zero mean. Fix t. Let $\chi_u := (\boldsymbol{I} - \Pi_\perp)\xi_u$ and $\varphi_u := \Pi_\perp \xi_u$. Since $K_\xi \xi_u = K_\xi \chi_u$, the trajectory $(s_u, a_u)_u$, the return G_t and η_t are functions of $(s_0, (w_u), (\eta_u), (\chi_u))$. The Gaussian processes (χ_u) and (φ_u) are uncorrelated, since $\operatorname{Cov}(\varphi_u, \chi_{u'}) = \rho^{|u-u'|} \Pi_\perp (\boldsymbol{I} - \Pi_\perp) = 0$. Hence they are independent, and both are independent of $(s_0, (w_u), (\eta_u), U)$. So $\mathbb{E}[G_t \eta_t \varphi_t^\top] = \mathbb{E}[G_t \eta_t]\, \mathbb{E}[\varphi_t]^\top = 0$. For an action-independent baseline $b = b(t, h_t^-, U)$ with $h_t^- = (h_{t-1}, o_t)$, $\mathbb{E}[b\, \eta_t \varphi_t^\top] = \mathbb{E}\big[b\, \mathbb{E}[\eta_t \mid h_t^-, U]\, \varphi_t^\top\big] = 0$. Both products are integrable by the hypothesis, since $\|\varphi_t\| \leq \|\xi_t\|$. The hypothesis also dominates the sum over t, so $\mathbb{E}[\hat{g}_\xi \Pi_\perp] = \mathbb{E} X_1 = 0$.

(d) SGD. Let $t \geq 1$ and $\Pi_\perp^{(t-1)}$ be the projection onto $\ker K_{\xi,t-1}$. Suppose first that $K_{s,t-1} \in \mathcal{K}_\gamma$ almost surely. As $K_{\xi,t-1}\Pi_\perp^{(t-1)} = 0$, the update gives $K_{\xi,t}\Pi_\perp^{(t-1)} = -\alpha_{t-1}\hat{g}_{\xi,t-1}\Pi_\perp^{(t-1)}$. Pathwise, therefore,

$$\|K_{\xi,t}\|_F^2 \;\geq\; \|K_{\xi,t}\Pi_\perp^{(t-1)}\|_F^2 = \alpha_{t-1}^2\,\|\hat{g}_{\xi,t-1}\Pi_\perp^{(t-1)}\|_F^2.$$

Conditionally on $\mathcal{F}_{t-1}$, the batch of iteration $t-1$ is as in Definition 5 at the $\mathcal{F}_{t-1}$-measurable gain $(K_{s,t-1}, K_{\xi,t-1})$, with U generated by $\mathcal{F}_{t-1}$. So by (c), $\mathbb{E}[\|\hat{g}_{\xi,t-1}\Pi_\perp^{(t-1)}\|_F^2 \mid \mathcal{F}_{t-1}] \geq F(K_{\xi,t-1}) \geq F_0$, because $\operatorname{rank} K_{\xi,t-1} \leq \min\{d_u, m\}$. Since α_{t-1} is $\mathcal{F}_{t-1}$-measurable, $\mathbb{E}\|K_{\xi,t}\|_F^2 \geq \mathbb{E}[\alpha_{t-1}^2]F_0$, with no unbiasedness or martingale argument.

For equation 10, first note the pathwise bound $J(K_{s,t}, K_{\xi,t}) - J^\star \;\geq\; \big(J_s(K_{s,t}) - J^\star\big) + \kappa_\rho\lambda_{\min}(R)\|K_{\xi,t}\|_F^2/(1-\gamma)$ in $[0,\infty]$. It holds by (b) on $\{K_{s,t} \in \mathcal{K}_\gamma\}$. On the complement both sides are $+\infty$, since $J \geq J_s = \infty$ there by (b). Take expectations. If $K_{s,t-1} \in \mathcal{K}_\gamma$ almost surely, use the previous paragraph. Otherwise the stopped run has $K_{s,t} = K_{s,t-1} \notin \mathcal{K}_\gamma$ with positive probability, so $\mathbb{E}J(K_{s,t}, K_{\xi,t}) = \infty$ and equation 10 holds trivially.

(e) Latent learner. The latent action is $a = K_z(W_s s + \delta\xi) + \sigma\eta = (K_z W_s)s + (K_z\delta)\xi + \sigma\eta$, so the latent policy is the raw policy $(K_z W_s, K_z\delta)$. By (b), $0 \leq J - J_s(K_z W_s) = \operatorname{tr}((K_z\delta)^\top Z_\rho(K_z\delta)) \leq \kappa_\rho^{-1}(1-\gamma)^{-1}\operatorname{tr}((K_z\delta)^\top M_\gamma(K_z\delta)) \leq \kappa_\rho^{-1}(1-\gamma)^{-1}\|M_\gamma\|\,\|K_z\delta\|_F^2$. Also $\|K_z\delta\|_F \leq \|K_z\|\,\|\delta\|_F$. This proves equation 11.

Now let $\delta = 0$ and W_s be invertible, and write $K_{s,t} := K_{z,t}W_s$. We show by induction on t that $(K_{z,0}, \ldots, K_{z,t})$, together with all latent data of iterations $< t$, is a fixed measurable function (fixed without reference to m, ρ) of the endogenous randomness $\omega_{<t}^{\mathrm{en}} := (s_0^i, w^i, \eta^i, \tau_i, U)$ of those iterations; the case $t = 0$ holds because $K_{z,0}$ is fixed. At iteration t, $K_{z,t} \in \mathcal{C}$, so the policy is $a = K_{s,t}s + \sigma\eta$ with $K_{s,t} \in \mathcal{K}_\gamma$, and it ignores ξ. Hence the iteration's endogenous trajectories are generated from ω^{en} by the distractor-free dynamics under $K_{s,t}$, whatever ξ is; the latent history h^z ($z_u = W_s s_u$, actions and costs) is a fixed function of them; so are the baselines $b(t, h^z, U)$, the latent estimator $\hat{g}_t^z = N^{-1}\sum_i (G_i - b_i)\sigma^{-1}\eta_{\tau_i}^i (W_s s_{\tau_i}^i)^\top$, the $\mathcal{F}_t^z$-measurable step α_t, and $K_{z,t+1} = \Pi_{\mathcal{C}}(K_{z,t} - \alpha_t\hat{g}_t^z)$. This completes the induction. Since the law of ω^{en} does not involve m or ρ and the functions are fixed, the joint law of $(K_{z,t})_t$, hence of $(K_{s,t})_t$, does not depend on m or ρ. The bound $\sup_t J_s(K_{s,t}) \leq \max_{\mathcal{C}W_s} J_s$ holds because $K_{s,t} \in \mathcal{C}W_s$, a compact subset of $\mathcal{K}_\gamma$ on which J_s is finite and continuous (it is the trace of the solution of a discrete Lyapunov equation, which depends continuously on the gain on $\mathcal{K}_\gamma$). For orthogonal W_s, $\hat{g}_t^z W_s = \hat{g}_t^s$ (the distractor-free estimator with baseline $b(t, W_s h, U)$), and $K_{s,t+1} = \Pi_{\mathcal{C}}(K_{z,t} - \alpha_t\hat{g}_t^z)W_s = \Pi_{\mathcal{C}W_s}(K_{s,t} - \alpha_t\hat{g}_t^s)$, since right multiplication by an orthogonal matrix is an isometry; this is the distractor-free projected SGD.

Why the projection. Suppose $B \neq 0$ and $\alpha_t > 0$, and take any baseline. Given the history up to the sampled action, $G_i - b_i \geq \gamma\, c(s_{\tau_i+1}, a_{\tau_i+1}) - b_i$ is unbounded above, because s_{τ_i+1} contains the Gaussian process noise w_{τ_i} and $c \geq \lambda_{\min}(Q)\|s\|^2$. Choosing η and s along directions u, v with $v^\top Bu \neq 0$ (an event of positive probability), the step $-\alpha_t\hat{g}$ moves K along $-c\,uv^\top$ with arbitrarily large $c > 0$, and $\operatorname{spr}(A + B(K - c\,uv^\top)) \geq |\operatorname{tr}(A + BK) - c\,v^\top Bu|/d \to \infty$. So a plain step leaves $\mathcal{K}_\gamma$ with positive probability, and then $\mathbb{E}J_s = +\infty$. □

A.4.3 Proofs of the corollaries

Corollary 2. (i) is Theorem 3(d) with a deterministic schedule, since $d_u(m - d_u)_+ v_{\min}/(\sigma^2 N) = F_0$. The claim about the latent learner is Theorem 3(e). Its policy has no ξ-gain, so its cost is $J_s(K_{s,t}^{\mathrm{lat}})$, and the law of $K_{s,t}^{\mathrm{lat}}$, hence e_t^{lat}, does not depend on m or ρ; $e_t^{\mathrm{lat}} \leq \max_{\mathcal{C}} J_s - J^\star < \infty$ by (e).

(ii) Drop the nonnegative first term in (i). For $m > m_0(t)$, $C\alpha_{t-1}^2(m - d_u) > e_t^{\mathrm{lat}}$, so $\mathbb{E}J(K_{s,t}^{\mathrm{raw}}, K_{\xi,t}^{\mathrm{raw}}) - J^\star > e_t^{\mathrm{lat}} = \mathbb{E}J_s(K_{s,t}^{\mathrm{lat}}) - J^\star$.

(iii) is the last display of Theorem 3(a).

(iv) Let $e_t := \mathbb{E}\|K_{\xi,t}\|_F^2$, $\bar{g}_t := \mathbb{E}[\hat{g}_{\xi,t} \mid \mathcal{F}_t]$, and let $\Pi_\perp^{(t)}$ project onto $\ker K_{\xi,t}$. By Lemma 20 applied conditionally on $\mathcal{F}_t$, whose hypotheses are those of (iv), $\bar{g}_t = 2(1-\gamma)Z_\rho(K_{s,t})K_{\xi,t}$.

Suppose $\mathbb{E}[\|\hat{g}_{\xi,t}\|_F^2 \mid \mathcal{F}_t] < \infty$. Then the variance decomposition gives

$$\mathbb{E}\big[\|K_{\xi,t+1}\|_F^2 \;\big|\; \mathcal{F}_t\big] = \|K_{\xi,t} - \alpha\bar{g}_t\|_F^2 + \alpha^2\,\mathbb{E}\big[\|\hat{g}_{\xi,t} - \bar{g}_t\|_F^2 \;\big|\; \mathcal{F}_t\big].$$

If instead that conditional second moment is infinite, the left side is infinite as well, since $\alpha > 0$. Bound the two terms separately.

- $K_{\xi,t} - \alpha\bar{g}_t = (\boldsymbol{I} - 2\alpha(1-\gamma)Z_\rho)K_{\xi,t}$. Here $\boldsymbol{I} - 2\alpha(1-\gamma)Z_\rho$ is symmetric with eigenvalues in $[1-\alpha\bar{L}, 1] \subseteq [0,1]$, because $0 \preceq 2(1-\gamma)Z_\rho \preceq \bar{L}\boldsymbol{I}$. Hence $\|K_{\xi,t} - \alpha\bar{g}_t\|_F \geq (1-\alpha\bar{L})\|K_{\xi,t}\|_F$.
- $\bar{g}_t\Pi_\perp^{(t)} = 0$, so $\|\hat{g}_{\xi,t} - \bar{g}_t\|_F^2 \geq \|(\hat{g}_{\xi,t} - \bar{g}_t)\Pi_\perp^{(t)}\|_F^2 = \|\hat{g}_{\xi,t}\Pi_\perp^{(t)}\|_F^2$, whose conditional mean is at least F_0 by Theorem 3(c).

So $e_{t+1} \geq (1-\alpha\bar{L})^2 e_t + \alpha^2 F_0$. With $e_0 \geq 0$, induction gives $e_t \geq \alpha^2 F_0 \sum_{j<t}(1-\alpha\bar{L})^{2j}$, which is the first claim. As $t \to \infty$, the right side tends to $\alpha^2 F_0/(1-(1-\alpha\bar{L})^2) = \alpha F_0/(\bar{L}(2-\alpha\bar{L})) \geq \alpha F_0/(2\bar{L})$. Finally, by Theorem 3(b), $\mathbb{E}J(K_{s,t}, K_{\xi,t}) - J^\star \geq \kappa_\rho \lambda_{\min}(R) e_t/(1-\gamma)$.

Corollary 3. By equation 11, pathwise, $0 \leq J(K_z W_s, K_z\delta) - J_s(K_z W_s) \leq \bar{\mu}\bar{k}_z^2\|\delta\|_F^2/(\kappa_\rho(1-\gamma))$. Take expectations. □

A.5 Federation pools the latent's data; the heads carry the bias (Theorem 4)

Setting and notation. Clients are indexed by $k \in [K]$. Client k has an endogenous process $(s_{k,t}, a_{k,t}, r_{k,t})_{t\in\mathbb{Z}}$ (state, action, reward) that is strictly stationary with finite fourth moments and is otherwise arbitrary; clients may differ arbitrarily (in the LG/LQ instance of Setting S2: $(A_k, B_k, F_k, \Sigma_{w,k}, Q_k, R_k)$). Only parts (b2), (b3) and (d) below, and the (TRAJ) rate in (e2), use the linear-Gaussian (LG) or linear-quadratic (LQ) structure (Assumption A3). The exogenous process of client k is the stationary Gaussian AR(1) chain of Setting S2, $\xi_{k,t+1} = \rho\,\xi_{k,t} + \sqrt{1-\rho^2}\,\varepsilon_{k,t}$ with $\rho \in [0,1)$, so that

$$\mathbb{E}\,\xi_{k,t} = 0, \qquad \mathbb{E}\big[\xi_{k,t}\xi_{k,t'}^\top\big] = \rho^{|t-t'|}\boldsymbol{I}_m. \tag{13}$$

Assumptions A1–A2 make the family of all endogenous processes (all clients, all times, including the behaviour policy's own randomness) independent of the family of all exogenous processes. Under *independent streams* (A1) the chains $\xi_1, \ldots, \xi_K$ are independent; under *shared streams*, $\xi_{k,t} = \xi_t$ for every k. Different clients run separate environments; with the behaviour policies fixed (A2), their endogenous processes are independent of one another. Observations are written in pooled-whitened coordinates (Setting S2): $o_t = [s_t; \xi_t]$ with $\Sigma_{o,k} := \mathbb{E}_k[o_t o_t^\top]$ and $\bar{\Sigma}_o := K^{-1}\sum_k \Sigma_{o,k} = \boldsymbol{I}_{d+m}$. The ξ-block of every $\Sigma_{o,k}$ is $\boldsymbol{I}_m$ and $\mathbb{E}_k[s_t\xi_t^\top] = 0$ in every client, so $\Sigma_{o,k} = \mathrm{diag}(\Sigma_{s,k}, \boldsymbol{I}_m)$ with $\Sigma_{s,k} := \mathbb{E}_k[s_t s_t^\top]$ and $K^{-1}\sum_k \Sigma_{s,k} = \boldsymbol{I}_d$. Until the remark on episode resets at the end of this subsection, the exogenous chains run without resets.

To avoid clashes with symbols used elsewhere, in this subsection we write d_z for the latent dimension (k in Setting S3), S for the number of local steps between synchronisations (E in Assumption A8; our runs synchronise every 5 rounds, i.e. $S = 1{,}600$ gradient steps), and $\mathrm{sr}(\cdot)$ for the spectral radius. The letter K alone is the number of clients; K_s, K_ξ and K_z are policy gains, as in Theorem 3. We write $\mathrm{sym}(X) = \frac{1}{2}(X + X^\top)$, $\|\cdot\|$ for the Euclidean or operator norm, and $\|\cdot\|_F$ for the Frobenius norm.

The analysed agent (Setting S3, linear version). The online encoder, the EMA target and the encoder that computed the λ-returns are $W = [W_s, E]$, $\bar{W} = [\bar{W}_s, \bar{E}]$ and $W_{\mathrm{old}} = [W_{\mathrm{old},s}, E_{\mathrm{old}}]$ in $\mathbb{R}^{d_z\times(d+m)}$, with $z_t = Wo_t = W_s s_t + E\xi_t$ and $\bar{z}_t = \bar{W}o_t$. The heads are $h = \big((P_j, Q_j)_{j\leq H}, g_z, g_a, v, v_{\mathrm{old}}\big)$. Client k's loss is

$$\begin{aligned} L_k(W; \bar{W}, W_{\mathrm{old}}, h) = \frac{\lambda_p}{H}\sum_{j=1}^{H}\mathbb{E}_k\big\|P_j z_t + Q_j a_{t:t+j-1} - \mathrm{sg}[\bar{z}_{t+j}]\,\big\|^2 \\ + \lambda_V\,\mathbb{E}_k\big(v^\top z_t - \mathrm{sg}[y_t]\,\big)^2 + \lambda_R\,\mathbb{E}_k\big(g_z^\top z_t + g_a^\top a_t - r_t\big)^2, \end{aligned} \tag{14}$$

where $a_{t:t+j-1}$ stacks $a_t, \dots, a_{t+j-1}$ and y_t is the λ-return of the code (the GAE advantage plus $V_{\mathrm{old}}(o_t)$, in which $V_{\mathrm{old}}(o_t)$ cancels),

$$y_t = \sum_{i\geq 0} (\gamma\lambda)^i r_{t+i} + \sum_{i\geq 1} c_i \, v_{\mathrm{old}}^\top W_{\mathrm{old}} o_{t+i},$$
$$c_i := \gamma(1-\lambda)(\gamma\lambda)^{i-1}, \qquad \sum_{i\geq 1} c_i \rho^i = \frac{\gamma(1-\lambda)\rho}{1-\gamma\lambda\rho} = c_V. \tag{15}$$

The encoder follows the semi-gradient: $\nabla_W L_k$ differentiates equation 14 through z_t only. The pooled loss is $\bar{L} := K^{-1}\sum_k L_k$. We call $x := (E, \bar{E}, E_{\mathrm{old}})$ the *noise state*, $\mathcal{M} := \{x = 0\}$ the exo-free set, and $\theta := (W_s, \bar{W}_s, W_{\mathrm{old},s}, h)$ an *endogenous configuration*. *Pooled heads* $\bar{h}$ are the minimisers of $\bar{L}$ over $\big((P_j, Q_j)_j, g_z, g_a, v\big)$ given the encoders and v_{old} (Assumption A5 for the pooled loss); a stationary point of the dynamics has $v_{\mathrm{old}} = \bar{v}$. Each head may carry an intercept (a constant regressor). Intercepts never multiply ξ: they add constants to the endogenous residuals, and hence to the scores $u_{k,t}$ of equation 17 below, and are otherwise omitted from the notation.

Normalisation: the pooled idealisation.

Assumption 1 (N-pool)**.** *The federated encoder satisfies the pooled second-moment constraint $W\bar{\Sigma}_o W^\top = WW^\top = \boldsymbol{I}_{d_z}$, that is, $K^{-1}\sum_k \mathbb{E}_k[z_t z_t^\top] = \boldsymbol{I}_{d_z}$, and its dynamics are the semi-gradient dynamics projected onto this constraint in pooled-whitened coordinates.*

This transfers the second-moment idealisation of LayerNorm and VICReg (Setting S3; Theorem 2) to a shared encoder. At $S = 1$ there is a single W, and per-client constraints $W\Sigma_{o,k}W^\top = \boldsymbol{I}_{d_z}$ for all k are in general infeasible: if $d_z = d$ and W_s is invertible they give $W_s(\Sigma_{s,k} - \Sigma_{s,l})W_s^\top = 0$, so $\Sigma_{s,k} = \Sigma_{s,l}$ for all k, l (homogeneous clients). At pooled-optimal heads the constraint exerts no force (Lemma 23), so the projection is inactive. The code, however, computes VICReg on each client's own minibatch, that is, it normalises per client. A smooth per-client second-moment penalty adds a first-order heterogeneity term to the noise dynamics, which we do not analyse. **Parts (b) and (c) of Theorem 4 are therefore statements about the pooled idealisation Assumption 1**; part (a) holds under either normalisation.

Pooled margin quantities. For heads h let $M = H^{-1}\sum_j P_j^\top P_j$, $N = H^{-1}\sum_j \rho^j P_j$, $G_V = vv^\top$, $G_R = g_z g_z^\top$,

$$M_{\mathrm{tot}} = \lambda_p M + (1 - c_V)\lambda_V G_V + \lambda_R G_R, \qquad \tilde{\Gamma} = M_{\mathrm{tot}} - \lambda_p N^\top,$$
$$J = M_{\mathrm{tot}}^{-1}\lambda_p N^\top, \qquad s = \big\| M_{\mathrm{tot}}^{-1/2}\, \lambda_p N^\top M_{\mathrm{tot}}^{-1/2} \big\| \tag{16}$$

(the last two when $M_{\mathrm{tot}} \succ 0$). A bar ($\bar{M}, \bar{N}, \bar{G}_V, \bar{G}_R, \bar{M}_{\mathrm{tot}}, \tilde{\Gamma}, \bar{J}, \bar{s}$) means evaluation at the pooled heads of the configuration under discussion; a subscript k ($\tilde{\Gamma}_k$, J_k, s_k) means evaluation at client k's own optimal heads when client k trains alone.

Federated averaging (A8). All clients start from a common initialisation. Between synchronisations each client runs S local steps of its own algorithm (semi-gradient steps on equation 14, the EMA update of $\bar{W}$, the refresh of $W_{\mathrm{old}}, v_{\mathrm{old}}$, and the head updates); at a synchronisation every averaged module is replaced by its client mean with equal weights. "FedAvg of all modules" averages $(W, \bar{W}, W_{\mathrm{old}}, h)$. At $S = 1$, in population, FedAvg of all modules is the semi-gradient dynamics of the pooled loss $\bar{L}$ with shared parameters.

Sampling (A7). Client k contributes n samples at times $t_1, \dots, t_n$; a sample at time t is the segment of the process from t on that the losses read (states, actions and rewards from t, and the targets' future observations). Under (IID) the n segments are independent draws of the stationary law (so ξ_{k,t_i} and $\xi_{k,t_{i'}}$ are independent for $i \neq i'$); under (TRAJ) they start at consecutive times $t_i = i$ of one stationary trajectory. Let $C_\xi \in \mathbb{R}^{n\times n}$ be the correlation matrix of the exogenous chain at the sample times: $C_\xi = \boldsymbol{I}_n$ under (IID) and $C_\xi = T_\rho := [\rho^{|i-i'|}]_{i,i'\leq n}$ under (TRAJ). Under shared streams we assume common sample times: the i-th samples of all clients read the same exogenous values (automatic under (TRAJ) with the code's step index within a round). Empirical means are $\hat{\mathbb{E}}_k[f] = n^{-1}\sum_{i\leq n} f_{k,t_i}$ and $\hat{\mathbb{E}} = K^{-1}\sum_k \hat{\mathbb{E}}_k$. We write $(\kappa_1, \kappa_2) := \big(\frac{1+\rho}{1-\rho}, \frac{1+\rho^2}{1-\rho^2}\big)$ under (TRAJ) and $(\kappa_1, \kappa_2) := (1, 1)$ under (IID).

Asymptotic regime (R). Every O_p and every "with probability tending to one" below refers to $Kn \to \infty$ with the pooled population objective held fixed: either K is fixed and $n \to \infty$, or the federation consists of r replicas of a fixed set of K_0 environments ($K = rK_0$, independent data) and $r \to \infty$ at any $n \geq 1$. Along either sequence the pooled population stationary point, its pooled heads, $\tilde{\Gamma}$ and $\bar{\sigma}_u^2$ (defined below) do not change, so all constants are fixed. Under shared streams the limit must include $n \to \infty$. Wherever a non-asymptotic bound is available we state it; see Theorem 4(c2).

Scores. For an endogenous configuration θ, the *endogenous score* of client k at time t is

$$\begin{aligned} u_{k,t}(\theta) = \frac{\lambda_p}{H} \sum_{j=1}^{H} P_j^\top \big(P_j W_s s_t + Q_j a_{t:t+j-1} - \bar{W}_s s_{t+j} \big) \\ + \lambda_V \, v \big(v^\top W_s s_t - y_t^0 \big) + \lambda_R \, g_z \big(g_z^\top W_s s_t + g_a^\top a_t - r_t \big), \end{aligned} \tag{17}$$

with $y_t^0 = \sum_{i \geq 0} (\gamma\lambda)^i r_{t+i} + \sum_{i \geq 1} c_i \, v_{\mathrm{old}}^\top W_{\mathrm{old},s} s_{t+i}$, all evaluated on client k's endogenous process. If the heads carry intercepts, each of the three residuals in equation 17 also contains its head's intercept, and y_t^0 contains the constant $\beta_{\mathrm{old}} \sum_{i \geq 1} c_i$ contributed by the intercept β_{old} of the old critic. In every case $u_{k,t}(\theta)$ is a measurable function of client k's endogenous process alone. The *noise score* is

$$\hat{g}_{0,k}(\theta) = \hat{\mathbb{E}}_k \big[u_{k,t}(\theta) \, \xi_{k,t}^\top \big] \in \mathbb{R}^{d_z \times m}, \qquad \hat{g}_0(\theta) = \frac{1}{K} \sum_{k=1}^{K} \hat{g}_{0,k}(\theta). \tag{18}$$

By Lemma 21 below, $\hat{g}_0(\theta)$ is one half of the empirical pooled E-gradient at $x = 0$.

Regularity for Theorem 4(c3).

Assumption 2 (Reg)**.** *Let $(\theta^\star, x = 0)$ be a pooled population stationary point with $\tilde{\Gamma}$ invertible, and consider independent streams in the regime (R). With probability tending to one, an empirical pooled stationary point $(\hat{\theta}, \hat{x})$ exists with $\hat{x} = (\hat{E}, \hat{\bar{E}}, \hat{E})$ and pooled-optimal empirical heads, and there are orthogonal $O \in \mathbb{R}^{d_z \times d_z}$ (gauge) with $\|O \cdot \hat{\theta} - \theta^\star\| = O_p\big((Kn)^{-1/2}\big)$, where $O\cdot$ acts by $W \mapsto OW$, $P_j \mapsto O P_j O^\top$, $Q_j \mapsto O Q_j$, $g_z \mapsto O g_z$, $v \mapsto Ov$, $v_{\mathrm{old}} \mapsto O v_{\mathrm{old}}$ (predictor intercepts map to $O\beta_j$; the scalar intercepts and the reward head's action part g_a are unchanged).*

This is the $\sqrt{Kn}$-consistency of the endogenous part of a Z-estimator. For (IID) with K fixed and $n \to \infty$ it would follow from Theorem 5.21 of van der Vaart (1998) (Section 5.3: i.i.d. observations, estimating equations with a Lipschitz envelope, a nonsingular derivative, and consistency), applied to the i.i.d. vectors that stack the K clients' i-th samples, once the gauge is fixed and consistency and finite fourth moments are assumed. That theorem does not cover $K \to \infty$ or (TRAJ) data, and we do not verify consistency modulo the gauge; so we assume (Reg) rather than prove it. It is used only in (c3) and in the learned-encoder statement of (e3).

Theorem 4 (federation pools the latent's data; the heads carry the bias)**.** *Assume A1, A2, A5 and A8 with heterogeneous clients and a common $\rho \in [0, 1)$. In addition: parts (b) and (c) assume the pooled normalisation Assumption 1; parts (b2), (b3) and (d) assume A3, and part (d) the policy-gradient model A9 (Definitions 5 and 6); parts (c) and (e2) use the sampling schemes of A7 and the regime (R). Part (a2) and part (d3) do not use A2.*

(a) ***No heterogeneity bias in the noise-block gradient on $\mathcal{M}$.***

- *(a1) For every client k, every endogenous configuration θ (heads not necessarily optimal) and every $(\bar{E}, E_{\mathrm{old}})$, the E-gradient of L_k at $E = 0$ is $-2\lambda_p N_h^\top \bar{E} - 2\lambda_V c_V \, v v_{\mathrm{old}}^\top E_{\mathrm{old}}$ with $N_h := H^{-1} \sum_j \rho^j P_j$; in particular it vanishes on $\mathcal{M}$. Hence $\mathcal{M}$ is invariant under every client's population dynamics (any step sizes, EMA rate and refresh schedule) and under FedAvg with any number S of local steps and any subset of averaged modules. This holds with the projection of Assumption 1, with any smooth per-client penalty $K^{-1} \sum_k \psi(W \Sigma_{o,k} W^\top)$, and with no normalisation. For MLP agents, under the hypotheses of Proposition 1(c) ((A1), (N) and (I) in every client, a start in Θ_0, and Adam moments vanishing on the first-layer ξ-blocks), the set Θ_0 of agents whose first-layer ξ-weights (online, target, old encoder) vanish is invariant under the population recursion of the implemented algorithm, for every S and every subset of averaged modules.*

(a2) (Exact sign symmetry, finite samples.) *Let the training algorithm be flip-equivariant (Definition 7; the predictive agent and raw PPO are, by Remark 1). Suppose the exogenous innovations ε are independent of (Θ_0, ω) and $-\varepsilon \overset{d}{=} \varepsilon$, and the initial state Θ_0 is independent of the other randomness ω and $\iota\Theta_0 \overset{d}{=} \Theta_0$. Then for every step t the joint law of the whole algorithm state and (ω, ε) is invariant under $(\Theta_t, \omega, \varepsilon) \mapsto (\iota\Theta_t, \omega, -\varepsilon)$, which negates the ξ-odd coordinates. In particular $\mathbb{E}\, E_{k,t} = 0$ for the linear agent, and the expectation of every first-layer ξ-column of an MLP agent (and, for raw PPO, of the actor's and critic's) is 0, for every client k and every t, whenever it exists. This holds for any heterogeneity, any S, any subset of averaged modules, SGD or Adam, gradient clipping, the running normaliser, VICReg on each client's minibatch, and independent or shared streams, and it holds without A2 (the actor may read the retained noise).*

(b) ***Pooled margin*** *($S = 1$). Consider FedAvg of all modules at $S = 1$ with pooled heads.*

(b1) For every noise state x the pooled E-gradient equals the right side of equation 26 (twice), evaluated at the common heads. Linearise at a point of $\mathcal{M}$ that is stationary for the pooled dynamics, so $\bar{W}_s = W_{\mathrm{old},s} = W_s$, $v_{\mathrm{old}} = \bar{v}$ and the heads are pooled-optimal. The noise dynamics are then, column by column,

$$\begin{aligned} \dot{e} &= -2\big(\lambda_p \bar{M} + \lambda_R \bar{G}_R + \lambda_V \bar{G}_V\big) e + 2\lambda_p \bar{N}^\top \bar{e} + 2\lambda_V c_V \bar{G}_V e_{\mathrm{old}}, \\ \dot{\bar{e}} &= \tau_1 (e - \bar{e}), \qquad \dot{e}_{\mathrm{old}} = \tau_2 (e - e_{\mathrm{old}}), \end{aligned} \tag{19}$$

with an $O(\|x\|^3)$ remainder, and the Jacobian of the full pooled field at the point is block diagonal with respect to the split into endogenous blocks and noise blocks. This is the system of Theorem 2(a) with pooled heads, so the conclusions of Theorem 2(b,c) hold with $\bar{J}$ and $\bar{s}$. Here and in (b3), "exponentially stable" and "unstable" are meant in the transverse sense of Definition 3, for the pooled system equation 19; we do not re-derive the link to the nonlinear flow (Lemma 15) for the pooled dynamics. If $d_z = d$, every point of $\mathcal{M}$ with W_s orthogonal, $\bar{W}_s = W_{\mathrm{old},s} = W_s$, pooled-optimal heads and $v_{\mathrm{old}} = \bar{v}$ is stationary.

(b2) (Scalar state.) *Let $d = d_z = H = 1$, $\lambda_V = \lambda_R = 0$, and let exploration be open-loop and centred (a_t independent of s_t, with $\mathbb{E}_k a_t = 0$, in every client) with $s_{t+1} = \alpha_k s_t + b_k a_t + w_t$ and $\mathbb{E} w_t = 0$. Then $\bar{P}_1 = \bar{\alpha} := \sum_k c_k \alpha_k / \sum_k c_k$ with $c_k = \mathbb{E}_k[s_t^2]$, and $\bar{J} = \rho/\bar{\alpha}$ if $\bar{\alpha} \neq 0$. If all $\alpha_k > 0$, then $\bar{J}$ lies between $\min_k J_k$ and $\max_k J_k$, where $J_k = \rho/\alpha_k$; in particular $\bar{J} > 1$ whenever every $J_k > 1$.*

(b3) ($d \geq 2$.) *This is false for $d \geq 2$. Take $d = d_z = 2$, $H = 1$, $\lambda_V = \lambda_R = 0$, $B_k = \boldsymbol{I}_2$, $F_k = 0$, $\sigma = 1$, $\Sigma_{w,k} = 0.1\, \boldsymbol{I}_2$, $\rho = \frac{1}{2}$, $A_1 = \mathrm{diag}(0.95, 0.3)$ and $A_2 = \mathrm{diag}(0.3, 0.95)$. Then $\bar{P}_1 = \bar{\alpha}\, \boldsymbol{I}_2$ with $\bar{\alpha} = (0.95 c_{\mathrm{hi}} + 0.3 c_{\mathrm{lo}})/(c_{\mathrm{hi}} + c_{\mathrm{lo}}) = 0.8871$, where $c_{\mathrm{hi}} = 1.1/(1 - 0.95^2)$ and $c_{\mathrm{lo}} = 1.1/(1 - 0.3^2)$. So $\bar{J} = \bar{s}\, \boldsymbol{I}_2$ with $\bar{s} = 0.5636 < 1$ and the pooled noise dynamics are exponentially stable for all $\tau_1, \tau_2 > 0$. Each client alone has $\mathrm{spec}(J_k) = \{0.526, 1.667\}$, and its transverse matrix (Definition 3) has a positive real eigenvalue for all $\tau_1, \tau_2 > 0$. The same holds in every dimension $d \geq 2$, by padding with coordinates that are stable for both clients.*

(c) ***Contamination*** *($S = 1$). Let $(\theta^\star, x = 0)$ be a pooled population stationary point with $\tilde{\Gamma}$ invertible, and let $\hat{g}_0 = \hat{g}_0(\theta^\star)$.*

(c1) (No bias.) *$\mathbb{E}\, \hat{g}_{0,k} = 0$ for every k. Under independent streams the $\hat{g}_{0,k}$ are independent across clients.*

(c2) (Fixed configuration; non-asymptotic.) *With $\theta = \theta^\star$ held fixed, the empirical pooled fixed-point equation for the noise block ($E = \bar{E} = E_{\mathrm{old}}$) is affine, $\hat{g}_0 + \hat{\mathcal{T}}(E) = 0$, with $\mathbb{E}\hat{\mathcal{T}}(E) = \tilde{\Gamma} E$. Let $\mathcal{A} := \{\|\hat{\mathcal{T}} - \tilde{\Gamma}\| \leq (2\|\tilde{\Gamma}^{-1}\|)^{-1}\}$, where $\|\cdot\|$ is the operator norm of linear maps of $(\mathbb{R}^{d_z \times m}, \|\cdot\|_F)$. On $\mathcal{A}$ the equation has a unique solution $\hat{E}_{\theta^\star}$, and $\|\hat{E}_{\theta^\star} + \tilde{\Gamma}^{-1}\hat{g}_0\|_F \leq 2\|\tilde{\Gamma}^{-1}\|^2\, \|\hat{\mathcal{T}} - \tilde{\Gamma}\|\, \|\hat{g}_0\|_F$. Put*

$$b_{\mathcal{T}} := \big\|\lambda_p \bar{M} + \lambda_R \bar{G}_R + \lambda_V \bar{G}_V\big\| + \frac{\lambda_p}{H} \sum_{j=1}^{H} \|\bar{P}_j\| + \lambda_V \|\bar{v}\|^2\, \frac{\gamma(1-\lambda)}{1 - \gamma\lambda},$$

$$C_{\mathcal{T}} := 2m^2 \kappa_2\, b_{\mathcal{T}}^2, \qquad C_g := m\, \kappa_1\, \bar{\sigma}_u^2,$$

with $\bar\sigma_u^2$ as in (c4). Under independent streams,

$$\Pr(\mathcal{A}^c) \le \frac{4\|\tilde\Gamma^{-1}\|^2 C_{\mathcal{T}}}{Kn}, \qquad \mathbb{E}\Big[\big\|\hat E_{\theta^\star} + \tilde\Gamma^{-1}\hat g_0\big\|_F \mathbf{1}_{\mathcal{A}}\Big] \le \frac{2\|\tilde\Gamma^{-1}\|^2\sqrt{C_{\mathcal{T}}C_g}}{Kn}, \tag{20}$$

so $\hat E_{\theta^\star} = -\tilde\Gamma^{-1}\hat g_0 + O_p\big((Kn)^{-1}\big)$ in the regime (R). Under shared streams equation 20 holds with $C_{\mathcal{T}}/(Kn)$ replaced by $C_{\mathcal{T}}/n$ and $C_g/(Kn)$ by $C_g/(Kn) + m\kappa_1\|\bar\mu\|^2/n$, where $\bar\mu := K^{-1}\sum_k \mathbb{E}u_{k,t}$. The remainder is then $O_p\big(K^{-1/2}n^{-1}\big)$ if $\bar\mu = 0$ and $O_p(n^{-1})$ otherwise, and $\Pr(\mathcal{A}) \to 1$ requires $n \to \infty$. (The factor m^2 in $C_{\mathcal{T}}$ comes from bounding operator norms by Frobenius norms.)

(c3) (Joint fixed point.) *Under Assumption 2 (independent streams, regime (R)), $\|\hat E\|_F = \|\tilde\Gamma^{-1}\hat g_0\|_F + O_p\big((Kn)^{-1}\big)$. If the empirical stationary point is unique up to the gauge, and the gauge is fixed by a rule that depends on $\hat\theta$ only, then $\hat E$ and $-\hat E$ have the same law.*

(c4) (Exact second moment of the leading term.) *Under independent streams,*

$$\begin{aligned}\mathbb{E}\big\|\tilde\Gamma^{-1}\hat g_0\big\|_F^2 &= \frac{m}{K^2n^2}\sum_{k=1}^{K}\mathbb{E}\Big[\tilde U_k^\top\big(C_\xi \otimes \boldsymbol{I}_{d_z}\big)\tilde U_k\Big],\\ \tilde U_k &:= \big(\tilde\Gamma^{-1}u_{k,t_1};\dots;\tilde\Gamma^{-1}u_{k,t_n}\big) \in \mathbb{R}^{nd_z},\end{aligned} \tag{21}$$

with $u_{k,t} = u_{k,t}(\theta^\star)$. Consequently, with $\bar\sigma_{\tilde u}^2 := \frac{1}{Kn}\sum_{k,i}\mathbb{E}\|\tilde\Gamma^{-1}u_{k,t_i}\|^2$ and $\bar\sigma_u^2 := \frac{1}{Kn}\sum_{k,i}\mathbb{E}\|u_{k,t_i}\|^2$,

$$\frac{1-\rho}{1+\rho}\,\frac{m\,\bar\sigma_{\tilde u}^2}{Kn} \;\le\; \mathbb{E}\big\|\tilde\Gamma^{-1}\hat g_0\big\|_F^2 \;\le\; \frac{1+\rho}{1-\rho}\,\frac{m\,\bar\sigma_{\tilde u}^2}{Kn} \tag{22}$$

under (TRAJ); under (IID) both sides are replaced by $m\bar\sigma_{\tilde u}^2/(Kn)$, with equality. If $\bar M_{\mathrm{tot}} \succ 0$ and $\bar s < 1$, then $\mathrm{sym}(\tilde\Gamma) \succeq (1-\bar s)\bar M_{\mathrm{tot}}$ and

$$\mathbb{E}\big\|\tilde\Gamma^{-1}\hat g_0\big\|_F^2 \le \frac{m}{Kn}\cdot\frac{1+\rho}{1-\rho}\cdot\frac{\bar\sigma_u^2}{(1-\bar s)^2\lambda_{\min}(\bar M_{\mathrm{tot}})^2}. \tag{23}$$

(c5) (A single client.) *Suppose client k, training alone (the setting of Theorem 2, with its own heads and normalisation), has a population stationary point with $x = 0$ at which its operator $\tilde\Gamma_k$ is invertible, and let $u^{(k)}$ be its scores and $\hat g_{0,k}$ its noise score there. Then $\mathbb{E}\|\tilde\Gamma_k^{-1}\hat g_{0,k}\|_F^2 \ge \mathbb{E}\|\hat g_{0,k}\|_F^2/\|\tilde\Gamma_k\|^2 \ge \frac{1-\rho}{1+\rho}\,\frac{m\,\sigma_{u,k}^2}{n\,\|\tilde\Gamma_k\|^2}$, where $\sigma_{u,k}^2 = \frac{1}{n}\sum_i\mathbb{E}\|u_{t_i}^{(k)}\|^2$ and the factor $\frac{1-\rho}{1+\rho}$ is 1 under (IID).*

(c6) (Shared versus independent streams; common sample times.) *If the $u_{k,t}$ have time-constant means $\mu_k := \mathbb{E}u_{k,t}$ and $\tilde\mu_k := \tilde\Gamma^{-1}\mu_k$, then*

$$\mathbb{E}_{\mathrm{shared}}\big\|\tilde\Gamma^{-1}\hat g_0\big\|_F^2 - \mathbb{E}_{\mathrm{indep}}\big\|\tilde\Gamma^{-1}\hat g_0\big\|_F^2 = \frac{m\,\mathbf{1}^\top C_\xi \mathbf{1}}{K^2n^2}\Big(\big\|\textstyle\sum_k\tilde\mu_k\big\|^2 - \sum_k\|\tilde\mu_k\|^2\Big). \tag{24}$$

The two coincide when every score is zero-mean; shared streams are never worse when $\sum_k\mu_k = 0$.

(d) ***Raw noise weights also pool.*** *Let client k be the LQ instance of Definition 4 with its own $(A_k, B_k, Q_k, R_k, \Sigma_{w,k}, \mu_{0,k})$, common $(d, d_u, m, \gamma, \rho, \sigma)$ and stabilising set $\mathcal{K}_{\gamma,k}$, and let the raw policies be $a = K_s s + K_\xi\xi + \sigma\eta$. Under FedSGD ($S = 1$) every client computes the estimator $\hat g_k$ of Definition 5 at the common gain from its own batch of N episodes, and the server applies the update of Definition 6 to $\hat g := K^{-1}\sum_k\hat g_k$.*

(d1) For every client k and every $K_s \in \mathcal{K}_{\gamma,k}$, the map $K_\xi \mapsto J_k(K_s, K_\xi)$ is uniquely minimised at $K_\xi = 0$ (Theorem 3(b)); for $K_s \in \bigcap_k\mathcal{K}_{\gamma,k}$, so is $\bar J := K^{-1}\sum_k J_k$.

(d2) Let $K_s \in \bigcap_k\mathcal{K}_{\gamma,k}$, and let the batches be independent across clients (independent streams). For every fixed $\Pi \in \mathbb{R}^{m\times m}$ with $\mathbb{E}\|\hat g_{\xi,k}\Pi\|_F^2 < \infty$ for all k, $\mathbb{E}\|(\hat g_\xi - \mathbb{E}\hat g_\xi)\Pi\|_F^2 = K^{-2}\sum_k\mathbb{E}\|(\hat g_{\xi,k} - \mathbb{E}\hat g_{\xi,k})\Pi\|_F^2$. If moreover the baselines are action-independent and satisfy the integrability condition of Theorem 3(c) in every client, then, with $\Pi_\perp$ the

orthogonal projector onto $\ker K_\xi$, $\mathbb{E}[\hat{g}_\xi]\Pi_\perp = 0$ *and* $\mathbb{E}\|\hat{g}_\xi\Pi_\perp\|_F^2 \geq K^{-1}\cdot\frac{1}{K}\sum_k F_k$, *where* $F_k := d_u(m-\operatorname{rank}K_\xi)v_{\min,k}/(\sigma^2 N)$ *with* $v_{\min,k} := 2\gamma^2\lambda_{\min}(Q_k)^2\operatorname{tr}(\Sigma_{w,k}^2)$ *is client* k*'s floor from Theorem 3(c). (Under the hypotheses of Lemma 20 in every client,* $\mathbb{E}\hat{g} = (1-\gamma)\nabla\bar{J}$.*)*

(d3) (Sign symmetry of the noise gain.) *Suppose that (i)* $K_{\xi,0}$ *is independent of all episodes and of* $K_{s,0}$, *with* $K_{\xi,0} \overset{d}{=} -K_{\xi,0}$ *(for instance* $K_{\xi,0}=0$*); (ii) the step sizes, the rule for the* s*-block and every baseline are* flip-invariant*: they are measurable functions of the data of the run (histories, sampled times, costs and earlier gains) and of auxiliary randomness independent of all episodes, whose values do not change when every exogenous coordinate* ξ *in the data and every earlier* K_ξ *are negated; and (iii)* $\mathbb{E}\|K_{\xi,t}\|_F < \infty$. *Then for every* t *the sequences* $(K_{s,u}, K_{\xi,u})_{u\leq t}$ *and* $(K_{s,u}, -K_{\xi,u})_{u\leq t}$ *have the same law, so* $\mathbb{E}K_{\xi,t} = 0$. *This holds for independent or shared streams and does not use A2. Conditions (i) and (ii) cannot be dropped: Definition 6 allows a deterministic* $K_{\xi,0}\neq 0$, *and from* $K_{\xi,0}=0$ *the admissible baseline* $b_i = c\,\eta^i_{\tau_i,1}\xi^i_{\tau_i,1}$ *(*$c\neq 0$*) gives* $\mathbb{E}K_{\xi,1} = (\alpha_0 c/\sigma)\,e_1 e_1^\top \neq 0$. *For raw PPO the conclusion is part of (a2).*

(e) ***Head bias (linear critics,*** $S=1$***).*** *Let a linear critic* w *regress an endogenous target* y *on features* ϕ*: either raw,* $\phi = o = (s,\xi)$*, or exo-free latent,* $\phi = W_s s$ *with* W_s *invertible. Let* $C_k = \mathbb{E}_k\phi\phi^\top \succ 0$, $w_k^\star = C_k^{-1}\mathbb{E}_k[\phi y]$ *and* $\mathcal{R}_k(w) = \mathbb{E}_k(y-\phi^\top w)^2 - \mathbb{E}_k(y-\phi^\top w_k^\star)^2$.

(e1) The pooled critic is $\bar{w} = (\sum_k C_k)^{-1}\sum_k C_k w_k^\star$ *and* $\mathcal{R}_k(w) = \|w-w_k^\star\|_{C_k}^2$. *The head bias* $b_k := \|\bar{w}-w_k^\star\|_{C_k}^2$ *does not depend on* n *and is the same for the raw and the exo-free agent. Exactly,* $\mathcal{R}_k^{\mathrm{raw}}(w_s,w_\xi) = \|w_s - w_{k,s}^\star\|_{\Sigma_{s,k}}^2 + \|w_\xi\|^2$, *and a latent critic on* $z = W_s s + E\xi$, *whose excess risk is measured from the exo-free optimum,* $\mathcal{R}_k^{\mathrm{lat}}(w;E) := \mathbb{E}_k(y-w^\top z)^2 - \mathbb{E}_k(y-s^\top w_{k,s}^\star)^2$, *has* $\mathcal{R}_k^{\mathrm{lat}}(w;E) = \|w-w_k^{z\star}\|_{A_k}^2 + \|E^\top w\|^2$, *where* $A_k = W_s\Sigma_{s,k}W_s^\top$ *and* $w_k^{z\star} = W_s^{-\top}w_{k,s}^\star$.

(e2) For the pooled least-squares critic from n *samples per client, the leading term* $\bar{C}^{-1}\hat{h}$ *(*$\bar{C} = K^{-1}\sum_k C_k$, $\hat{h} = \hat{\mathbb{E}}[\phi(y-\phi^\top\bar{w})]$*) has mean zero. The first-order errors* $\mathrm{err}_k := b_k + \mathbb{E}\|\bar{C}^{-1}\hat{h}\|_{C_k}^2$ *satisfy* $\mathrm{err}_k^{\mathrm{raw}} = \mathrm{err}_k^{\mathrm{lat}} + V_\xi$ *exactly, with* $V_\xi = \mathbb{E}\|\hat{h}_\xi\|^2$. *Under independent streams* V_ξ *obeys the sandwich equation 22 (with* $d_z=1$ *and* $\tilde{\Gamma}=1$*) with* $\bar{\sigma}_u^2$ *replaced by* $\bar{\sigma}_e^2 := \frac{1}{Kn}\sum_{k,i}\mathbb{E}\,e_{k,t_i}^2$, $e = y - s^\top\bar{w}_s$*; under shared streams it gains the term equation 24 with* $\mu_k = \mathbb{E}_k e$. *Under independent streams the remainder* $\hat{w}-\bar{w}-\bar{C}^{-1}\hat{h}$ *is* $O_p((Kn)^{-1})$ *in the regime (R) whenever* $\hat{C}-\bar{C}$ *and* $\hat{h}$ *are* $O_p((Kn)^{-1/2})$*, in particular under (IID) and under replication (*$K = rK_0$*).*

(e3) Let the federation replicate K_0 *fixed environments* r *times (*$K = rK_0$*, independent data). Then* $\bar{w}$ *and* b_k *do not depend on* r*, and the realised excess risks of the raw and of the exo-free agent converge to* b_k *almost surely as* $r\to\infty$. *If* $b_k > 0$,

$$\frac{\mathcal{R}_k^{\mathrm{raw}}(\hat{w}^{\mathrm{raw}})}{\mathcal{R}_k^{\mathrm{lat}}(\hat{w}^{\mathrm{lat}};\hat{E})} \;\longrightarrow\; 1$$

in probability, also for a learned latent encoder $(\hat{W}_s,\hat{E}) \to (W_s,0)$ *in probability, as in (c3). For the exo-free latent the first-order errors satisfy, exactly,*

$$\frac{\mathrm{err}_k^{\mathrm{raw}}}{\mathrm{err}_k^{\mathrm{lat}}} = 1 + \frac{V_\xi(1)}{r\,b_k + V_s(1)}, \qquad V_s := \mathbb{E}\|\bar{\Sigma}_s^{-1}\hat{h}_s\|_{\Sigma_{s,k}}^2,$$

where $V(1)$ *denotes the value at* $r=1$. *So the ratio falls to* 1 *at rate* $1/r$ *when* $b_k > 0$*, and does not depend on* r *when* $b_k = 0$.

Roadmap. Lemma 21 is the engine: given the heads, the noise-block gradient is the same in every client. Lemmas 22 and 23 make the pooled linearisation exact to first order. Lemmas 24 to 26 give the second moments. Lemma 27 is the finite-sample symmetry. Parts (d) and (e) are then linear algebra.

Lemma 21 (The noise-block gradient does not see the client)**.** *Fix a client* k*, an endogenous configuration* θ *(arbitrary heads) and a noise state* $x = (E,\bar{E},E_{\mathrm{old}})$. *Let* $\hat{\Upsilon}_{k,i} := \hat{\mathbb{E}}_k[\xi_{k,t+i}\xi_{k,t}^\top]$ *for*

$i \geq 0$. *Then the empirical half-gradient is*

$$\begin{aligned}\tfrac{1}{2}\nabla_E \hat{L}_k = \hat{g}_{0,k}(\theta) + \frac{\lambda_p}{H}\sum_{j=1}^{H} P_j^\top \left(P_j E \hat{\Upsilon}_{k,0} - \bar{E}\hat{\Upsilon}_{k,j}\right) \\ + \lambda_R\, g_z g_z^\top E \hat{\Upsilon}_{k,0} + \lambda_V\, v\Big(v^\top E \hat{\Upsilon}_{k,0} - \sum_{i\geq 1} c_i\, v_{\mathrm{old}}^\top E_{\mathrm{old}} \hat{\Upsilon}_{k,i}\Big),\end{aligned} \tag{25}$$

where $\hat{L}_k$ is equation 14 with $\mathbb{E}_k$ replaced by $\hat{\mathbb{E}}_k$. In population ($\hat{\mathbb{E}}_k \to \mathbb{E}_k$),

$$\tfrac{1}{2}\nabla_E L_k = \frac{\lambda_p}{H}\sum_{j=1}^{H} P_j^\top \left(P_j E - \rho^j \bar{E}\right) + \lambda_R\, g_z g_z^\top E + \lambda_V\, v\big(v^\top E - c_V\, v_{\mathrm{old}}^\top E_{\mathrm{old}}\big), \tag{26}$$

which depends on the heads, on x and on (ρ, γ, λ), but not on the client's endogenous law. (With episode resets this last property needs common reset statistics; see the remark on resets at the end of this subsection.)

Proof. The semi-gradient of equation 14 with respect to E is $\nabla_E \hat{L}_k = 2\hat{\mathbb{E}}_k[\delta_t \xi_t^\top]$, where $\delta_t = \frac{\lambda_p}{H}\sum_j P_j^\top \varepsilon_{j,t} + \lambda_V v\, \varepsilon_{V,t} + \lambda_R g_z \varepsilon_{R,t}$ collects the residuals $\varepsilon_{j,t} = P_j z_t + Q_j a_{t:t+j-1} - \bar{z}_{t+j}$, $\varepsilon_{V,t} = v^\top z_t - y_t$ and $\varepsilon_{R,t} = g_z^\top z_t + g_a^\top a_t - r_t$ (the targets $\bar{z}_{t+j}$ and y_t are not differentiated). Split each residual into its endogenous part and its part linear in the noise. With $z_t = W_s s_t + E\xi_t$, $\bar{z}_{t+j} = \bar{W}_s s_{t+j} + \bar{E}\xi_{t+j}$ and $y_t = y_t^0 + \sum_{i\geq 1} c_i v_{\mathrm{old}}^\top E_{\mathrm{old}} \xi_{t+i}$ (from equation 15),

$$\begin{aligned}\varepsilon_{j,t} = \varepsilon_{j,t}^0 + P_j E\xi_t - \bar{E}\xi_{t+j}, \qquad \varepsilon_{R,t} = \varepsilon_{R,t}^0 + g_z^\top E\xi_t, \\ \varepsilon_{V,t} = \varepsilon_{V,t}^0 + v^\top E\xi_t - \textstyle\sum_{i\geq 1} c_i v_{\mathrm{old}}^\top E_{\mathrm{old}} \xi_{t+i}.\end{aligned}$$

Here the ε^0 are the residuals at $x = 0$, which are functions of client k's endogenous process, and the endogenous half of δ_t is exactly $u_{k,t}(\theta)$ of equation 17. Multiplying by $\xi_t^\top$ and averaging gives equation 25. The series over i converges in L^1 because $\sum_i c_i < 1$ and $\mathbb{E}\|\xi_{t+i}\xi_t^\top\|_F \leq m$. In population, $\mathbb{E}_k[u_{k,t}\xi_{k,t}^\top] = \mathbb{E}_k[u_{k,t}]\,(\mathbb{E}\xi_{k,t})^\top = 0$ by the independence of the endogenous and exogenous families (A1–A2) and equation 13. Also $\mathbb{E}\hat{\Upsilon}_{k,i} = \rho^i \boldsymbol{I}_m$, and $\sum_{i\geq 1} c_i \rho^i = c_V$. Substituting yields equation 26. □

Lemma 22 (Pooled heads move at second order)**.** *Fix the endogenous blocks $(W_s, \bar{W}_s, W_{\mathrm{old},s}, v_{\mathrm{old}})$. The pooled optimal predictor (P_j, Q_j), reward head (g_z, g_a) and critic v solve linear systems whose coefficient matrices and right-hand sides are affine functions of the bilinear forms $EE^\top$, $\bar{E}E^\top$ and $EE_{\mathrm{old}}^\top v_{\mathrm{old}}$. If these systems are nonsingular at $x = 0$, the pooled heads are real-analytic functions of x near 0 and even in x, so $\bar{h}(x) = \bar{h}(0) + O(\|x\|^2)$.*

Proof. The pooled normal equations involve only the pooled second moments $K^{-1}\sum_k \mathbb{E}_k[\cdot]$ of $(z_t, a_{t:t+j-1})$, of $(\bar{z}_{t+j}, (z_t, a_{t:t+j-1}))$, of $(r_t, (z_t, a_t))$ and of (y_t, z_t). By independence and equation 13, $\mathbb{E}_k[z_t z_t^\top] = W_s \mathbb{E}_k[s_t s_t^\top] W_s^\top + EE^\top$, $\mathbb{E}_k[z_t a_{t:t+j-1}^\top] = W_s \mathbb{E}_k[s_t a_{t:t+j-1}^\top]$, $\mathbb{E}_k[\bar{z}_{t+j} z_t^\top] = \bar{W}_s \mathbb{E}_k[s_{t+j} s_t^\top] W_s^\top + \rho^j \bar{E}E^\top$, $\mathbb{E}_k[\bar{z}_{t+j} a_{t:t+j-1}^\top] = \bar{W}_s \mathbb{E}_k[s_{t+j} a_{t:t+j-1}^\top]$, $\mathbb{E}_k[z_t r_t] = W_s \mathbb{E}_k[s_t r_t]$, $\mathbb{E}_k[a_t r_t]$ is free of x, and $\mathbb{E}_k[z_t y_t] = W_s \mathbb{E}_k[s_t y_t^0] + c_V\, EE_{\mathrm{old}}^\top v_{\mathrm{old}}$. The solution of a nonsingular linear system is a rational, hence real-analytic, function of its entries. The entries are affine in the three bilinear forms, which are even in x and $O(\|x\|^2)$. □

Lemma 23 (No constraint force at optimal heads)**.** *Let $\mathcal{L}$ be L_k, $\bar{L}$ or an empirical version of either, and let the heads be $\mathcal{L}$-optimal given the encoders. Then the semi-gradient $G = \nabla_W \mathcal{L}$ satisfies $GW^\top = 0$; this identity does not depend on the coordinates in which o is written. Consequently the Euclidean flow $\dot{W} = -G$, run in fixed coordinates, conserves $WW^\top$ in those coordinates, and the projection onto $\{WW^\top = \boldsymbol{I}_{d_z}\}$ is inactive: the projected and the free flows coincide. For $\bar{L}$ (or its empirical version) in pooled-whitened coordinates this is the constraint of Assumption 1; for L_k in client-k-whitened coordinates it is client k's own constraint (the setting of Theorem 2).*

The lemma does not make G tangent to a constraint $WAW^\top = \boldsymbol{I}$ with $A \neq \boldsymbol{I}$. For instance, for $\mathcal{L} = L_k$ in pooled-whitened coordinates, G is in general not tangent to client k's constraint $W\Sigma_{o,k}W^\top = \boldsymbol{I}$: numerically, at random points on that constraint $GW^\top \approx 10^{-16}$ while $\|\operatorname{sym}(G\Sigma_{o,k}W^\top)\|/\|G\| = 0.07$–$0.26$.

Proof. As in the proof of Lemma 21, $G = 2\,\mathbb{E}[\delta_t o_t^\top]$ with $\mathbb{E}$ the expectation defining $\mathcal{L}$. So $GW^\top = 2\,\mathbb{E}[\delta_t z_t^\top] = 2\big(\frac{\lambda_p}{H}\sum_j P_j^\top \mathbb{E}[\varepsilon_{j,t} z_t^\top] + \lambda_V v\,\mathbb{E}[\varepsilon_{V,t} z_t^\top] + \lambda_R g_z\,\mathbb{E}[\varepsilon_{R,t} z_t^\top]\big)$. Each expectation vanishes by the normal equation of the corresponding optimal head: $\mathbb{E}[\varepsilon_{j,t} z_t^\top] = 0$ for P_j, $\mathbb{E}[\varepsilon_{V,t} z_t] = 0$ for v, and $\mathbb{E}[\varepsilon_{R,t} z_t] = 0$ for g_z. For the pooled loss these are the pooled normal equations $K^{-1}\sum_k \mathbb{E}_k[\cdot] = 0$. Then $\frac{d}{dt}WW^\top = -(GW^\top + WG^\top) = 0$. □

Lemma 24 (AR(1) correlation sections)**.** *For $\rho \in [0,1)$ and every n, the eigenvalues of $T_\rho = [\rho^{|i-i'|}]_{i,i'\le n}$ lie in $\big[\frac{1-\rho}{1+\rho}, \frac{1+\rho}{1-\rho}\big]$. Hence $\frac{1-\rho}{1+\rho}\|U\|^2 \le U^\top(T_\rho \otimes \boldsymbol{I}_q)U \le \frac{1+\rho}{1-\rho}\|U\|^2$ for every $U \in \mathbb{R}^{nq}$.*

Proof. Let $f(\omega) = \sum_{l\in\mathbb{Z}} \rho^{|l|} e^{\mathrm{i}l\omega}$. Summing the two geometric series, $f(\omega) = \frac{1}{1-\rho e^{\mathrm{i}\omega}} + \frac{\rho e^{-\mathrm{i}\omega}}{1-\rho e^{-\mathrm{i}\omega}} = \frac{1-\rho^2}{1-2\rho\cos\omega+\rho^2}$. Since $(1-\rho)^2 \le 1 - 2\rho\cos\omega + \rho^2 \le (1+\rho)^2$, we get $\frac{1-\rho}{1+\rho} \le f \le \frac{1+\rho}{1-\rho}$. The series converges absolutely, so $\rho^{|l|} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega) e^{-\mathrm{i}l\omega} d\omega$. For $x \in \mathbb{R}^n$, $x^\top T_\rho x = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega)\big|\sum_i x_i e^{\mathrm{i}i\omega}\big|^2 d\omega$, and Parseval gives $\frac{1}{2\pi}\int |\sum_i x_i e^{\mathrm{i}i\omega}|^2 d\omega = \|x\|^2$. This proves the eigenvalue bounds. The eigenvalues of $T_\rho \otimes \boldsymbol{I}_q$ are those of T_ρ, each repeated q times. □

Lemma 25 (Exact second moments of noise scores)**.** *Let $(u_i)_{i\le n}$ and $(u'_i)_{i\le n}$ be random vectors in $\mathbb{R}^q$ with finite second moments, and let $(\xi_i)_{i\le n}$ and $(\xi'_i)_{i\le n}$ be random vectors in $\mathbb{R}^m$. Suppose the pair (u,u') is independent of the pair (ξ,ξ'), and $\mathbb{E}[\xi'_{i'}\xi_i^\top] = D_{ii'}\boldsymbol{I}_m$ with $\mathbb{E}\xi_i = 0$. Let $G = n^{-1}\sum_i u_i\xi_i^\top$ and $G' = n^{-1}\sum_i u'_i\xi_i'^\top$. Then $\mathbb{E}G = 0$ and*

$$\mathbb{E}\langle G, G'\rangle_F = \frac{m}{n^2}\sum_{i,i'} D_{ii'}\,\mathbb{E}[u_i^\top u'_{i'}].$$

In particular, for $u' = u$ and $\xi' = \xi$ with $\mathbb{E}[\xi_{i'}\xi_i^\top] = (C_\xi)_{ii'}\boldsymbol{I}_m$, $\mathbb{E}\|G\|_F^2 = \frac{m}{n^2}\mathbb{E}[U^\top(C_\xi \otimes \boldsymbol{I}_q)U]$ with $U = (u_1;\dots;u_n)$.

Proof. $\mathbb{E}G = n^{-1}\sum_i \mathbb{E}[u_i](\mathbb{E}\xi_i)^\top = 0$ by independence. Next, $\langle G, G'\rangle_F = n^{-2}\sum_{i,i'}(u_i^\top u'_{i'})(\xi_i^\top \xi'_{i'})$. Both factors are integrable by Cauchy–Schwarz, and they are independent, so each term has expectation $\mathbb{E}[u_i^\top u'_{i'}]\,\mathbb{E}[\xi_i^\top \xi'_{i'}]$. Finally $\mathbb{E}[\xi_i^\top \xi'_{i'}] = \operatorname{tr}\mathbb{E}[\xi'_{i'}\xi_i^\top] = mD_{ii'}$. □

Lemma 26 (Margin to conditioning)**.** *If $M_{\mathrm{tot}} \succ 0$ and $s < 1$, then $\operatorname{sym}(\tilde\Gamma) \succeq (1-s)M_{\mathrm{tot}}$. Consequently $\tilde\Gamma$ is invertible, and $\|\tilde\Gamma^{-1}Y\|_F \le \|Y\|_F/\big((1-s)\lambda_{\min}(M_{\mathrm{tot}})\big)$ for every $Y \in \mathbb{R}^{d_z\times m}$.*

Proof. For $x \in \mathbb{R}^{d_z}$ let $y = M_{\mathrm{tot}}^{1/2}x$. Then $|x^\top\lambda_p N^\top x| = |y^\top M_{\mathrm{tot}}^{-1/2}\lambda_p N^\top M_{\mathrm{tot}}^{-1/2}y| \le s\|y\|^2 = s\,x^\top M_{\mathrm{tot}}x$, so $x^\top\tilde\Gamma x \ge (1-s)x^\top M_{\mathrm{tot}}x \ge (1-s)\lambda_{\min}(M_{\mathrm{tot}})\|x\|^2$. By Cauchy–Schwarz, $\|\tilde\Gamma x\|\,\|x\| \ge x^\top\tilde\Gamma x$, hence $\|\tilde\Gamma x\| \ge (1-s)\lambda_{\min}(M_{\mathrm{tot}})\|x\|$. Therefore $\|\tilde\Gamma^{-1}\| \le ((1-s)\lambda_{\min}(M_{\mathrm{tot}}))^{-1}$. Apply this column by column to Y. □

Definition 7 (Flip-equivariant algorithm)**.** Let the full state of a federated training run after t steps be $\Theta_t = (\Theta_t^{\mathrm{e}}, \Theta_t^{\mathrm{o}})$, where Θ_t^{o} is a real vector (the ξ-*odd* coordinates), and let $\iota(\Theta^{\mathrm{e}}, \Theta^{\mathrm{o}}) = (\Theta^{\mathrm{e}}, -\Theta^{\mathrm{o}})$. The state contains the agents' parameters and optimiser and normaliser statistics, and the environments' current states and the stored data. The randomness of the run has three parts: the initial state Θ_0 (the agents' initialisation; the environments are reset at the first step); the *exogenous innovations* ε, the Gaussian draws of the distractor generators (initial values at resets and AR(1) innovations) in the order in which they are consumed; and all other randomness ω (environment transitions and resets, action noise, minibatch indices). The run is $\Theta_{t+1} = \Psi_t(\Theta_t, \omega, \varepsilon)$ for measurable maps Ψ_t. It is *flip-equivariant* if $\Psi_t(\iota\Theta, \omega, -\varepsilon) = \iota\Psi_t(\Theta, \omega, \varepsilon)$ for all $t, \Theta, \omega, \varepsilon$.

Remark 1 (The implemented agents are flip-equivariant)**.** For the MLP agents of our experiments (the predictive agent and raw PPO), take as ξ-odd coordinates the first-layer ξ-columns of every network that reads the observation (online, target and old encoder, or raw actor and critic) with their Adam first moments, the ξ-entries of the observation normaliser's running mean, and the distractor values held by the run. Flipping these together with the exogenous innovations leaves every first-layer pre-activation unchanged: the first layer is linear in the normalised observation, and the normaliser's ξ-map is odd (its clipping at ± 10 is symmetric). Every network output, loss, target, return and action distribution is therefore unchanged, the gradient of each odd coordinate changes sign, and every other gradient is unchanged. Adam's second moments, the running variances and the gradient-norm clipping (per parameter group) depend on squares and are even, so every optimiser step, EMA update, normaliser update and FedAvg average maps the flipped state to the flipped successor, which is Definition 7.

Lemma 27 (Sign symmetry)**.** *Let the run be flip-equivariant, and suppose that*

$$(\iota\Theta_0, \omega, -\varepsilon) \overset{d}{=} (\Theta_0, \omega, \varepsilon). \tag{27}$$

Then for every t, $(\iota\Theta_t, \omega, -\varepsilon) \overset{d}{=} (\Theta_t, \omega, \varepsilon)$. In particular $(\Theta_t^{\mathrm{e}}, -\Theta_t^{\mathrm{o}}) \overset{d}{=} (\Theta_t^{\mathrm{e}}, \Theta_t^{\mathrm{o}})$, so $\mathbb{E}[f(\Theta_t^{\mathrm{e}})\,\Theta_t^{\mathrm{o}}] = 0$ for every bounded measurable f whenever $\mathbb{E}\|\Theta_t^{\mathrm{o}}\| < \infty$. Hypothesis equation 27 holds if Θ_0, ω and ε are independent, $\iota\Theta_0 \overset{d}{=} \Theta_0$ and $-\varepsilon \overset{d}{=} \varepsilon$.

The marginal conditions alone do not suffice. If $\Theta_0^{\mathrm{o}} = g(\varepsilon_0)$ for an even g with values ± 1 and $g(\varepsilon_0) \overset{d}{=} -g(\varepsilon_0)$ (say $g = +1$ iff $|\varepsilon_0|$ exceeds its median), the flip-equivariant update $\Theta_1^{\mathrm{o}} = \Theta_0^{\mathrm{o}}\,\varepsilon_0^2$ gives $\mathbb{E}\Theta_1^{\mathrm{o}} = \mathbb{E}[g(\varepsilon_0)\varepsilon_0^2] > 0$.

Proof. Let F_t be the t-fold composition, $\Theta_t = F_t(\Theta_0, \omega, \varepsilon)$. By induction on t, $F_t(\iota\Theta_0, \omega, -\varepsilon) = \iota F_t(\Theta_0, \omega, \varepsilon)$: the case $t = 0$ is trivial, and $F_{t+1}(\iota\Theta_0, \omega, -\varepsilon) = \Psi_t(\iota F_t(\Theta_0, \omega, \varepsilon), \omega, -\varepsilon) = \iota F_{t+1}(\Theta_0, \omega, \varepsilon)$. Put $\mathcal{H}_t(a, b, c) := (F_t(a, b, c), b, c)$. Then $(\iota\Theta_t, \omega, -\varepsilon) = \mathcal{H}_t(\iota\Theta_0, \omega, -\varepsilon)$, which by equation 27 has the law of $\mathcal{H}_t(\Theta_0, \omega, \varepsilon) = (\Theta_t, \omega, \varepsilon)$. Taking the $(\Theta^{\mathrm{e}}, \Theta^{\mathrm{o}})$-marginal gives $\mathbb{E}[f(\Theta_t^{\mathrm{e}})\Theta_t^{\mathrm{o}}] = -\mathbb{E}[f(\Theta_t^{\mathrm{e}})\Theta_t^{\mathrm{o}}]$. For the sufficient condition: under independence the law of $(\Theta_0, \omega, \varepsilon)$ is the product of the three marginal laws, and the map $(\iota, \mathrm{id}, -)$ acts factor by factor and preserves each of them. □

Proof of Theorem 4(a). *(a1).* Put $E = 0$ in equation 26. The gradient is $2 \cdot \big(-\frac{\lambda_p}{H} \sum_j \rho^j P_j^\top \bar{E} - \lambda_V c_V v v_{\mathrm{old}}^\top E_{\mathrm{old}} \big)$, for arbitrary heads and endogenous blocks, and it vanishes when also $\bar{E} = E_{\mathrm{old}} = 0$. On $\mathcal{M}$ the acting encoder carries no noise, so A2 is self-consistent there. Take any population update of the algorithm: a semi-gradient step on (W, h) with any step size, the EMA $\bar{W} \leftarrow (1-\tau)\bar{W} + \tau W$, the refresh $(W_{\mathrm{old}}, v_{\mathrm{old}}) \leftarrow (W, v)$, and head steps. On $\mathcal{M}$ each maps $x = 0$ to $x = 0$: the E-step is zero and the other two are linear. The normalisation does not change this. The projection of Assumption 1 replaces G by $G - \mathrm{sym}(GW^\top)W$, whose E-block $G_E - \mathrm{sym}(GW^\top)E$ equals G_E at $E = 0$; a per-client penalty adds $2K^{-1}\sum_k \nabla\psi(W\Sigma_{o,k}W^\top)E$ to the E-gradient (differentiate $W \mapsto \psi(W\Sigma_{o,k}W^\top)$ and use $\Sigma_{o,k} = \mathrm{diag}(\Sigma_{s,k}, \boldsymbol{I}_m)$), which vanishes at $E = 0$. A synchronisation replaces each averaged noise block by a mean of zeros and leaves non-averaged ones untouched. So $\mathcal{M}$ is invariant under every client's population dynamics and under FedAvg with any S and any subset of averaged modules. Continuous-time flows behave the same way: their noise component vanishes on $\mathcal{M}$, so by uniqueness of solutions of locally Lipschitz ODEs a solution started in $\mathcal{M}$ stays there. For MLP agents the statement is Proposition 1(c), whose hypotheses are those listed in (a1): on Θ_0 every network output and every target is a function of the endogenous process, and the gradient of the first-layer ξ-columns is $\mathbb{E}[\delta_1\kappa_\xi(\xi_t)^\top] = \mathbb{E}[\delta_1]\,\mathbb{E}[\kappa_\xi(\xi_t)]^\top = 0$ by (A1) and (N), where κ_ξ is the (clipped) normaliser's ξ-map and δ_1 the back-propagated first-layer error. The averaging argument is unchanged.

(a2). The hypotheses of (a2) are the sufficient condition of Lemma 27: ε is independent of (Θ_0, ω) and Θ_0 of ω, so the three are independent. Apply Lemma 27 with the odd coordinates of Remark 1. For the linear agent of this subsection, run with any finite-sample gradient algorithm on the empirical losses (with the projection of Assumption 1, a per-client penalty, or neither), the odd coordinates are $(E, \bar{E}, E_{\mathrm{old}})$ of every client together with the exogenous values held by the run. The empirical losses and the empirical second moments $\hat{\mathbb{E}}_k[z_t z_t^\top]$ depend on these only through the products $E\xi_t$, $\bar{E}\xi_{t+j}$ and $E_{\mathrm{old}}\xi_{t+i}$, which are invariant under the flip. So every gradient with respect to a noise block is odd

and every other gradient is even (and so is the projection term $\mathrm{sym}(GW^\top)W$ block by block), and the algorithm is flip-equivariant by the argument of Remark 1. Hence $E_{k,t} \overset{d}{=} -E_{k,t}$ and $\mathbb{E}E_{k,t} = 0$ whenever the expectation exists. None of the hypotheses of Lemma 27 involves A2, the heterogeneity of the clients, S, the set of averaged modules, or whether the streams are shared: sharing means that one sequence of innovations is used by all clients, and its law is still sign-symmetric. □

Proof of Theorem 4(b). *(b1).* At $S = 1$ with all modules averaged, all clients hold the same parameters before every step, so the pooled update is the client average of updates evaluated at common heads. By Lemma 21 each client's noise half-gradient is the same function equation 26 of (h, x), so the pooled noise update equals that function too. By Lemma 23 applied to $\bar{L}$ in pooled-whitened coordinates, $GW^\top = 0$ at pooled-optimal heads, so the projection of Assumption 1 is inactive and the constraint correction $-\mathrm{sym}(GW^\top)E$ of Theorem 2(a) is identically zero here. With pooled-optimal heads, $\bar{h} = \bar{h}(\theta_{\mathrm{en}}, x)$, where θ_{en} are the endogenous blocks. The noise field is $F(x) = \mathcal{A}(\bar{h}(\theta_{\mathrm{en}}, x))\, x$, with $\mathcal{A}(h)$ the linear map equation 26 (times -2), together with the linear EMA and refresh fields. By Lemma 22, $F(x) = \mathcal{A}(\bar{h}(\theta_{\mathrm{en}}, 0))x + O(\|x\|^3)$.

The endogenous field depends on x only through $\bar{h}$. Indeed $\nabla_{W_s}\bar{L} = 2K^{-1}\sum_k \mathbb{E}_k[\delta_t s_t^\top]$, and the noise-linear part of δ_t multiplies ξ, which is independent of s_t with mean zero. So the endogenous field is even in x and its x-derivative vanishes at $x = 0$. Conversely $F(0; \theta_{\mathrm{en}}) = 0$ for every θ_{en} by (a1), so $\partial_{\theta_{\mathrm{en}}} F = 0$ at $x = 0$. Hence the Jacobian at a stationary point of $\mathcal{M}$ is block diagonal, and its noise block is $\mathcal{A}(\bar{h}(\theta_{\mathrm{en}}, 0))$. Writing equation 26 at the pooled heads with $v_{\mathrm{old}} = \bar{v}$, column by column, gives equation 19.

Theorem 2(b,c) are statements about the linear system equation 19 for arbitrary coefficient matrices of the form $M = H^{-1}\sum_j P_j^\top P_j$, $N = H^{-1}\sum_j \rho^j P_j$, $G_V = vv^\top$, $G_R = g_z g_z^\top$. The pooled matrices have this form, so those statements apply verbatim with $\bar{J}$ and $\bar{s}$.

For the last claim ($d_z = d$) take $x = 0$, W_s orthogonal, $\bar{W}_s = W_{\mathrm{old},s} = W_s$, pooled-optimal heads and $v_{\mathrm{old}} = \bar{v}$. Lemma 23 gives $\nabla_{W_s}\bar{L}\, W_s^\top + \nabla_E \bar{L}\, E^\top = 0$, and $E = 0$, so $\nabla_{W_s}\bar{L}\, W_s^\top = 0$ and $\nabla_{W_s}\bar{L} = 0$. Also $\nabla_E \bar{L} = 0$ by (a1). The EMA and the refresh are at rest because $W = W_{\mathrm{old}} = W$, the heads are optimal, and $v_{\mathrm{old}} = \bar{v}$ makes the critic refresh a rest point.

(b2). Take $W_s = 1$ (the sign is a gauge), so $z_t = x_t := s_t/\bar{\sigma}$ with $\bar{\sigma}^2 = K^{-1}\sum_k c_k$, and $\bar{z}_{t+1} = x_{t+1}$. The pooled normal equations for (P_1, Q_1) are $\sum_k \mathbb{E}_k[(P_1 x_t + Q_1 a_t - x_{t+1})x_t] = 0$ and $\sum_k \mathbb{E}_k[(P_1 x_t + Q_1 a_t - x_{t+1})a_t] = 0$. Centred open-loop exploration gives $\mathbb{E}_k[x_t a_t] = \mathbb{E}_k[x_t]\,\mathbb{E}_k[a_t] = 0$, so the first equation alone reads $P_1 \sum_k \mathbb{E}_k[x_t^2] = \sum_k \mathbb{E}_k[x_{t+1}x_t]$. Since $x_{t+1} = \alpha_k x_t + (b_k a_t + w_t)/\bar{\sigma}$ with a_t and w_t independent of x_t and centred, the right side is $\sum_k \alpha_k \mathbb{E}_k[x_t^2]$. Hence $\bar{P}_1 = \sum_k \alpha_k c_k / \sum_k c_k = \bar{\alpha}$, $\bar{M} = \bar{\alpha}^2$ and $\bar{N} = \rho\bar{\alpha}$, so $\bar{M}_{\mathrm{tot}} = \lambda_p \bar{\alpha}^2 \succ 0$ when $\bar{\alpha} \neq 0$, and then $\bar{J} = \rho/\bar{\alpha}$. A single client regresses on its own data and gets $P_1 = \alpha_k$ and $J_k = \rho/\alpha_k$. If every $\alpha_k > 0$, then $\bar{\alpha} \in [\min_k \alpha_k, \max_k \alpha_k]$ is a convex combination, so $\bar{J} \in [\rho/\max_k \alpha_k, \rho/\min_k \alpha_k] = [\min_k J_k, \max_k J_k]$.

(b3). Here $B_k = \boldsymbol{I}$, $\sigma = 1$ and $\Sigma_{w,k} = 0.1\boldsymbol{I}$, so the stationary covariance of the raw state solves $\Sigma_{s,k}^{\mathrm{raw}} = A_k \Sigma_{s,k}^{\mathrm{raw}} A_k^\top + 1.1\,\boldsymbol{I}$. For diagonal A_k it is $\mathrm{diag}\left(1.1/(1 - a_{k,i}^2)\right)_i$. Centred open-loop exploration decouples the pooled normal equations as in (b2), giving $\bar{P}_{\mathrm{raw}} \sum_k \Sigma_{s,k}^{\mathrm{raw}} = \sum_k A_k \Sigma_{s,k}^{\mathrm{raw}}$, and all matrices are diagonal. Each coordinate therefore has $\bar{\alpha} = (0.95\, c_{\mathrm{hi}} + 0.3\, c_{\mathrm{lo}})/(c_{\mathrm{hi}} + c_{\mathrm{lo}})$, with $c_{\mathrm{hi}} = 1.1/0.0975 = 11.2821$ and $c_{\mathrm{lo}} = 1.1/0.91 = 1.2088$, so $\bar{\alpha} = 0.8871$. The pooled whitening $(K^{-1}\sum_k \Sigma_{s,k}^{\mathrm{raw}})^{-1/2}$ is diagonal and commutes with $\bar{P}_{\mathrm{raw}}$, so $\bar{P}_1 = \bar{\alpha}\boldsymbol{I}_2$ (at $W_s = \boldsymbol{I}$, and at every $W_s = w\boldsymbol{I}$), $\bar{J} = \rho \bar{P}_1^{-1} = 0.5636\,\boldsymbol{I}_2$ and $\bar{s} = \rho/\bar{\alpha} = 0.5636$.

With $\lambda_V = \lambda_R = 0$ and $H = 1$, equation 19 decouples into one 2×2 system per entry of $(E, \bar{E})$, $\dot{e} = -2\lambda_p \bar{\alpha}^2 e + 2\lambda_p \rho \bar{\alpha}\, \bar{e}$, $\dot{\bar{e}} = \tau_1(e - \bar{e})$, plus the e_{old}-block, which is driven by e but does not feed back ($\lambda_V = 0$) and contributes the eigenvalue $-\tau_2$. The 2×2 matrix has trace $-2\lambda_p\bar{\alpha}^2 - \tau_1 < 0$ and determinant $2\lambda_p\tau_1\bar{\alpha}(\bar{\alpha} - \rho) > 0$. A real 2×2 matrix is Hurwitz iff its trace is negative and its determinant positive, so the pooled noise dynamics are exponentially stable for all $\tau_1, \tau_2 > 0$.

Client k alone, in its own whitened coordinates, has $P_1 = (\Sigma_{s,k}^{\mathrm{raw}})^{-1/2} A_k (\Sigma_{s,k}^{\mathrm{raw}})^{1/2} = A_k$ by the normal equations (w_t is independent of (s_t, a_t); this is Theorem 2(e)). So $J_k = \rho A_k^{-1}$ has spectrum $\{0.5/0.95, 0.5/0.3\} = \{0.526, 1.667\}$. In the coordinate with $a = 0.3$ the 2×2 determinant is

$2\lambda_p\tau_1 \cdot 0.3 \cdot (0.3 - 0.5) < 0$, which gives a positive real eigenvalue of the client's transverse matrix for all $\tau_1, \tau_2 > 0$. For $d > 2$, append coordinates with $a_{k,i} = 0.9 > \rho$ for both clients. Everything stays diagonal, and those coordinates are stable for both clients and for the pool. □

Proof of Theorem 4(c). *(c1).* By Lemma 25, with $u_i = u_{k,t_i}$ and $\xi_i = \xi_{k,t_i}$ (independent of the u_i by A1–A2), $\mathbb{E}\hat{g}_{0,k} = 0$. Under independent streams $\hat{g}_{0,k}$ is a function of client k's endogenous and exogenous processes only, and these families are independent across k.

(c2). Put $E = \bar{E} = E_{\mathrm{old}}$ in equation 25 and average over k at the pooled heads of $\theta^\star$ (with $v_{\mathrm{old}} = \bar{v}$). This gives $\frac{1}{2}\nabla_E \hat{L} = \hat{g}_0 + \hat{\mathcal{T}}(E)$, where

$$\hat{\mathcal{T}}(E) = \frac{\lambda_p}{H}\sum_j \bar{P}_j^\top\left(\bar{P}_j E\hat{\Upsilon}_0 - E\hat{\Upsilon}_j\right) + \lambda_R\,\bar{g}_z\bar{g}_z^\top E\hat{\Upsilon}_0 + \lambda_V\,\bar{v}\bar{v}^\top E\Big(\hat{\Upsilon}_0 - \sum_{i\geq 1} c_i\hat{\Upsilon}_i\Big),$$
$$\hat{\Upsilon}_i := \frac{1}{K}\sum_k \hat{\Upsilon}_{k,i}.$$

Since $\mathbb{E}\hat{\Upsilon}_i = \rho^i \boldsymbol{I}_m$ and $\sum_i c_i\rho^i = c_V$, we have $\mathbb{E}\hat{\mathcal{T}}(E) = \tilde{\Gamma}E =: \mathcal{T}(E)$.

Next we bound the fluctuations of $\hat{\Upsilon}_i$. Under (TRAJ), Isserlis' theorem for the Gaussian chain gives, for entries (a, b),

$$\mathrm{Var}\left(\tfrac{1}{n}\sum_t \xi_{a,t+i}\xi_{b,t}\right) = \frac{1}{n^2}\sum_{t,t'}\left(\rho^{2|t-t'|} + \delta_{ab}\,\rho^{|t-t'+i|+|t-t'-i|}\right) \leq \frac{2}{n}\,\frac{1+\rho^2}{1-\rho^2},$$

using $|l+i|+|l-i| \geq 2|l|$ and $\sum_{l\in\mathbb{Z}}\rho^{2|l|} = \frac{1+\rho^2}{1-\rho^2}$. Under (IID) the n segments are independent, and the variance is $n^{-1}\,\mathrm{Var}(\xi_{a,t+i}\xi_{b,t}) = n^{-1}(1+\delta_{ab}\rho^{2i}) \leq 2/n$. So $\mathbb{E}\|\hat{\Upsilon}_{k,i} - \rho^i\boldsymbol{I}\|_F^2 \leq 2m^2\kappa_2/n$ uniformly in i. Under independent streams the K summands of $\hat{\Upsilon}_i - \rho^i\boldsymbol{I}$ are independent and centred, so $\mathbb{E}\|\hat{\Upsilon}_i - \rho^i\boldsymbol{I}\|_F^2 \leq 2m^2\kappa_2/(Kn)$; under shared streams with common sample times $\hat{\Upsilon}_i = \hat{\Upsilon}_{k,i}$ for every k, and the bound is $2m^2\kappa_2/n$.

Because $\|AYB\|_F \leq \|A\|\,\|Y\|_F\,\|B\|$ and $\|\cdot\| \leq \|\cdot\|_F$,

$$\|\hat{\mathcal{T}} - \mathcal{T}\| \leq \left\|\lambda_p\bar{M} + \lambda_R\bar{G}_R + \lambda_V\bar{G}_V\right\|\,\|\hat{\Upsilon}_0 - \boldsymbol{I}\|_F + \frac{\lambda_p}{H}\sum_j \|\bar{P}_j\|\,\|\hat{\Upsilon}_j - \rho^j\boldsymbol{I}\|_F + \lambda_V\|\bar{v}\|^2\sum_{i\geq 1} c_i\|\hat{\Upsilon}_i - \rho^i\boldsymbol{I}\|_F.$$

By Minkowski's inequality in L^2, $\sum_{i\geq 1} c_i = \gamma(1-\lambda)/(1-\gamma\lambda)$ and the uniform bound, $(\mathbb{E}\|\hat{\mathcal{T}} - \mathcal{T}\|^2)^{1/2} \leq b_{\mathcal{T}}\,(2m^2\kappa_2/(Kn))^{1/2}$, that is, $\mathbb{E}\|\hat{\mathcal{T}} - \mathcal{T}\|^2 \leq C_{\mathcal{T}}/(Kn)$ under independent streams, and $\leq C_{\mathcal{T}}/n$ under shared streams. Next, (c4) below with $\tilde{\Gamma}$ replaced by $\boldsymbol{I}$ (Lemma 25 and the upper bound of Lemma 24, with equality under (IID)) gives $\mathbb{E}\|\hat{g}_0\|_F^2 \leq m\kappa_1\bar{\sigma}_u^2/(Kn) = C_g/(Kn)$ under independent streams. Under shared streams (c6), again with $\tilde{\Gamma} = \boldsymbol{I}$, adds $\frac{m\mathbf{1}^\top C_\xi\mathbf{1}}{K^2n^2}\big(\|\sum_k \mu_k\|^2 - \sum_k\|\mu_k\|^2\big) \leq \frac{m\,\mathbf{1}^\top C_\xi\mathbf{1}}{n^2}\|\bar{\mu}\|^2 \leq \frac{m\kappa_1}{n}\|\bar{\mu}\|^2$, since $\mathbf{1}^\top C_\xi\mathbf{1} \leq n\,\lambda_{\max}(C_\xi) \leq n\kappa_1$ by Lemma 24.

On $\mathcal{A}$ write $\hat{\mathcal{T}} = \mathcal{T}(\mathcal{I} + \mathcal{T}^{-1}(\hat{\mathcal{T}} - \mathcal{T}))$, where $\mathcal{T}^{-1}$ is left multiplication by $\tilde{\Gamma}^{-1}$, so $\|\mathcal{T}^{-1}(\hat{\mathcal{T}} - \mathcal{T})\| \leq \frac{1}{2}$. The Neumann series shows that $\hat{\mathcal{T}}$ is invertible with $\|\hat{\mathcal{T}}^{-1}\| \leq 2\|\tilde{\Gamma}^{-1}\|$. So the unique solution is $\hat{E}_{\theta^\star} = -\hat{\mathcal{T}}^{-1}\hat{g}_0$, and, since $\mathcal{T}^{-1} - \hat{\mathcal{T}}^{-1} = \hat{\mathcal{T}}^{-1}(\hat{\mathcal{T}} - \mathcal{T})\mathcal{T}^{-1}$,

$$\left\|\hat{E}_{\theta^\star} + \tilde{\Gamma}^{-1}\hat{g}_0\right\|_F = \left\|\hat{\mathcal{T}}^{-1}(\hat{\mathcal{T}} - \mathcal{T})\mathcal{T}^{-1}\hat{g}_0\right\|_F \leq 2\|\tilde{\Gamma}^{-1}\|^2\,\|\hat{\mathcal{T}} - \mathcal{T}\|\,\|\hat{g}_0\|_F.$$

Markov's inequality gives $\Pr(\mathcal{A}^c) = \Pr\left(\|\hat{\mathcal{T}} - \mathcal{T}\|^2 > (4\|\tilde{\Gamma}^{-1}\|^2)^{-1}\right) \leq 4\|\tilde{\Gamma}^{-1}\|^2\,\mathbb{E}\|\hat{\mathcal{T}} - \mathcal{T}\|^2$, and the Cauchy–Schwarz inequality gives $\mathbb{E}\big[\|\hat{E}_{\theta^\star} + \tilde{\Gamma}^{-1}\hat{g}_0\|_F\mathbf{1}_{\mathcal{A}}\big] \leq 2\|\tilde{\Gamma}^{-1}\|^2(\mathbb{E}\|\hat{\mathcal{T}} - \mathcal{T}\|^2)^{1/2}(\mathbb{E}\|\hat{g}_0\|_F^2)^{1/2}$. Inserting the bounds gives equation 20 and its shared-stream version. In the regime (R) all constants are fixed, so $\Pr(\mathcal{A}) \to 1$ and, by Markov's inequality applied on $\mathcal{A}$, the remainder is $O_p((Kn)^{-1})$ under independent streams; under shared streams it is $O_p(K^{-1/2}n^{-1})$ if $\bar{\mu} = 0$ and $O_p(n^{-1})$ otherwise.

(c3). After applying the gauge O of Assumption 2 we may assume $\|\hat\theta-\theta^\star\|=O_p((Kn)^{-1/2})$. This is legitimate: the empirical losses are invariant under the gauge, so $O\cdot(\hat\theta,\hat x)$ is again an empirical stationary point, with noise block $O\hat E$ and $\|O\hat E\|_F=\|\hat E\|_F$. The noise-block stationarity reads $\hat g_0(\hat\theta)+\hat{\mathcal{T}}_{\hat\theta}(\hat E)=0$, where $\hat{\mathcal{T}}_\theta$ is $\hat{\mathcal{T}}$ with the heads of θ (by equation 25).

First, $\|\hat{\mathcal{T}}_{\hat\theta}-\mathcal{T}_{\theta^\star}\|\le\|\hat{\mathcal{T}}_{\hat\theta}-\mathcal{T}_{\hat\theta}\|+\|\mathcal{T}_{\hat\theta}-\mathcal{T}_{\theta^\star}\|$. By the display in the proof of (c2), $\|\hat{\mathcal{T}}_\theta-\mathcal{T}_\theta\|\le b(\theta)\,Z$ for every θ, where $b(\theta)$ is a continuous function of the heads of θ and $Z:=\|\hat\Upsilon_0-\boldsymbol{I}\|_F+\sum_{j\le H}\|\hat\Upsilon_j-\rho^j\boldsymbol{I}\|_F+\sum_{i\ge1}c_i\|\hat\Upsilon_i-\rho^i\boldsymbol{I}\|_F$ does not depend on θ and has $\mathbb{E}Z^2=O((Kn)^{-1})$. Since $\hat\theta\to\theta^\star$ in probability, the first term is $b(\hat\theta)Z=O_p((Kn)^{-1/2})$. The second is $O(\|\hat\theta-\theta^\star\|)$ because $\mathcal{T}_\theta$ is polynomial in θ.

Second, $u_{k,t}(\theta)=\sum_\alpha p_\alpha(\theta)X_{\alpha,k,t}$, where X_α ranges over the finitely many endogenous vectors s_t, $a_{t:t+H-1}$, s_{t+j} ($j\le H$), r_t, $\sum_{i\ge0}(\gamma\lambda)^i r_{t+i}$, $\sum_{i\ge1}c_i s_{t+i}$ and the constant 1 (for intercepts), all with finite second moments, and the p_α are matrix polynomials. So $\hat g_0(\hat\theta)-\hat g_0(\theta^\star)=\sum_\alpha(p_\alpha(\hat\theta)-p_\alpha(\theta^\star))\,\hat{\mathbb{E}}[X_\alpha\xi^\top]$. By Lemmas 24 and 25, $\mathbb{E}\|\hat{\mathbb{E}}[X_\alpha\xi^\top]\|_F^2\le m\kappa_1\,\bar{\mathbb{E}}\|X_\alpha\|^2/(Kn)$, where $\bar{\mathbb{E}}$ averages over clients, so $\hat{\mathbb{E}}[X_\alpha\xi^\top]=O_p((Kn)^{-1/2})$ and this difference is $O_p((Kn)^{-1})$.

Combining, as in (c2), $\hat E+\tilde\Gamma^{-1}\hat g_0(\theta^\star)=-(\hat{\mathcal{T}}_{\hat\theta}^{-1}-\mathcal{T}_{\theta^\star}^{-1})\hat g_0(\hat\theta)-\mathcal{T}_{\theta^\star}^{-1}\big(\hat g_0(\hat\theta)-\hat g_0(\theta^\star)\big)=O_p((Kn)^{-1})$. Taking norms and undoing the gauge gives the claim for $\|\hat E\|_F$.

For the law statement, suppose the stationary point is unique up to the gauge, and fix the gauge by a rule depending on $\hat\theta$ only. Negating the exogenous data maps $(\hat\theta,\hat x)$ to $(\hat\theta,-\hat x)$, because each $\hat L_k$ depends on $(E,\bar E,E_{\mathrm{old}},\xi)$ only through the products $E\xi_t$, $\bar E\xi_{t+j}$ and $E_{\mathrm{old}}\xi_{t+i}$. The data law is invariant under this negation, because the exogenous family is independent of the endogenous one (A1–A2) and the centred Gaussian AR(1) law is sign-symmetric. So $\hat E\overset{d}{=}-\hat E$.

(c4). Apply Lemma 25 with $u_i=\tilde\Gamma^{-1}u_{k,t_i}$ and $\xi_i=\xi_{k,t_i}$ to get $\mathbb{E}\|\tilde\Gamma^{-1}\hat g_{0,k}\|_F^2=\frac{m}{n^2}\mathbb{E}[\tilde U_k^\top(C_\xi\otimes\boldsymbol{I})\tilde U_k]$. Expand $\mathbb{E}\|\tilde\Gamma^{-1}\hat g_0\|_F^2=K^{-2}\sum_{k,l}\mathbb{E}\langle\tilde\Gamma^{-1}\hat g_{0,k},\tilde\Gamma^{-1}\hat g_{0,l}\rangle$. For $k\ne l$ the two factors are independent with mean zero by (c1), so the cross terms vanish, and equation 21 follows. Under (TRAJ), Lemma 24 bounds each quadratic form between $\frac{1-\rho}{1+\rho}\mathbb{E}\|\tilde U_k\|^2$ and $\frac{1+\rho}{1-\rho}\mathbb{E}\|\tilde U_k\|^2$, and $\sum_k\mathbb{E}\|\tilde U_k\|^2=Kn\,\bar\sigma_{\tilde u}^2$. This is equation 22; under (IID), $C_\xi=\boldsymbol{I}$ gives equality. Finally, $\bar\sigma_{\tilde u}^2\le\|\tilde\Gamma^{-1}\|^2\bar\sigma_u^2$, and Lemma 26 (with pooled heads) gives $\|\tilde\Gamma^{-1}\|\le((1-\bar s)\lambda_{\min}(\bar M_{\mathrm{tot}}))^{-1}$ and $\mathrm{sym}(\tilde\Gamma)\succeq(1-\bar s)\bar M_{\mathrm{tot}}$. This is equation 23.

(c5). For any matrix Y, $\|Y\|_F=\|\tilde\Gamma_k\tilde\Gamma_k^{-1}Y\|_F\le\|\tilde\Gamma_k\|\,\|\tilde\Gamma_k^{-1}Y\|_F$, which gives the first inequality. The second is Lemma 25 for $u^{(k)}$ together with the lower bound of Lemma 24 under (TRAJ) (equality with factor 1 under (IID)).

(c6). With shared streams the $k=l$ terms are unchanged. For $k\ne l$, Lemma 25 with $D=C_\xi$ (the same chain, read at common sample times) gives $\mathbb{E}\langle\tilde\Gamma^{-1}\hat g_{0,k},\tilde\Gamma^{-1}\hat g_{0,l}\rangle=\frac{m}{n^2}\sum_{i,i'}(C_\xi)_{ii'}\mathbb{E}[\tilde u_{k,t_i}^\top\tilde u_{l,t_{i'}}]$. Since clients' endogenous processes are independent, $\mathbb{E}[\tilde u_{k,t_i}^\top\tilde u_{l,t_{i'}}]=\tilde\mu_k^\top\tilde\mu_l$. Summing over $k\ne l$ and dividing by K^2 gives equation 24, because $\sum_{k\ne l}\tilde\mu_k^\top\tilde\mu_l=\|\sum_k\tilde\mu_k\|^2-\sum_k\|\tilde\mu_k\|^2$. If $\sum_k\mu_k=0$, then $\sum_k\tilde\mu_k=0$ and the difference is $-\frac{m\mathbf{1}^\top C_\xi\mathbf{1}}{K^2n^2}\sum_k\|\tilde\mu_k\|^2\le0$, since $C_\xi\succ0$. □

Proof of Theorem 4(d). *(d1).* Let $K_s\in\mathcal{K}_{\gamma,k}$. By Theorem 3(b) for client k, $J_k(K_s,K_\xi)=J_{s,k}(K_s)+q_{\rho,k}(K_s,K_\xi)$ with $J_{s,k}(K_s)<\infty$, $q_{\rho,k}(K_s,K_\xi)=\mathrm{tr}(K_\xi^\top Z_{\rho,k}(K_s)K_\xi)$, so $q_{\rho,k}(K_s,0)=0$, and $q_{\rho,k}(K_s,K_\xi)\ge\kappa_\rho\lambda_{\min}(R_k)\|K_\xi\|_F^2/(1-\gamma)>0$ for $K_\xi\ne0$. So $K_\xi=0$ is the unique minimiser of $J_k(K_s,\cdot)$. For $K_s\in\bigcap_k\mathcal{K}_{\gamma,k}$ every $J_k(K_s,0)$ is finite, and averaging over k preserves both properties. (For $K_s\notin\mathcal{K}_{\gamma,k}$, $J_k(K_s,\cdot)\equiv\infty$ by Theorem 3(b), so the restriction is needed.)

(d2). At the common gain, $\hat g_k$ is a function of client k's batch and baseline, and the batches are independent across clients. So $\hat g_\xi\Pi-\mathbb{E}[\hat g_\xi]\Pi=K^{-1}\sum_k(\hat g_{\xi,k}\Pi-\mathbb{E}[\hat g_{\xi,k}]\Pi)$ is a sum of independent, centred, square-integrable terms, the cross terms of its squared norm have expectation 0, and the

identity follows. Under the further hypotheses, Theorem 3(c) applied to client k at $K_s \in \mathcal{K}_{\gamma,k}$ gives $\mathbb{E}[\hat{g}_{\xi,k}]\Pi_\perp = 0$ and $\mathbb{E}\|\hat{g}_{\xi,k}\Pi_\perp\|_F^2 \geq F_k$ (both sides in $[0,\infty]$). Hence $\mathbb{E}[\hat{g}_\xi]\Pi_\perp = 0$ and $\mathbb{E}\|\hat{g}_\xi\Pi_\perp\|_F^2 = K^{-2}\sum_k \mathbb{E}\|\hat{g}_{\xi,k}\Pi_\perp\|_F^2 \geq K^{-2}\sum_k F_k$; when some term is infinite, so is the left side. The parenthetical claim is Lemma 20 applied to each client.

(d3). Write the run as a deterministic function of $(K_0, \omega, \varepsilon)$. Here ε collects the exogenous innovations of all episodes of all clients at all iterations (the initial values ξ_0 and the innovations ϵ_u of Definition 4; under shared streams one sequence is used by all clients), and ω collects everything else: initial states, w, η, the sampled times τ, and the auxiliary randomness of the baselines, step sizes and s-rule. We show by induction on t that replacing $(K_{\xi,0}, \varepsilon)$ by $(-K_{\xi,0}, -\varepsilon)$ leaves $(K_{s,u})_{u\leq t}$ and all endogenous variables of iterations $< t$ unchanged, and negates $(K_{\xi,u})_{u\leq t}$ and all exogenous variables of iterations $< t$. The case $t = 0$ is trivial. Suppose the claim holds up to t. In every episode of iteration t the exogenous path is linear in the innovations, so it is negated. The action $a_u = K_{s,t}s_u + K_{\xi,t}\xi_u + \sigma\eta_u$ is unchanged because $K_{\xi,t}\xi_u$ is, so by induction over u ($s_{u+1} = A_k s_u + B_k a_u + w_u$) the states, actions and costs are unchanged. So the returns G_i are unchanged, and by (ii) so are the baselines, the step size and the s-rule, whose arguments change only by negating exogenous coordinates and earlier K_ξ's. The score $\sigma^{-1}\eta\, o^\top$ has s-columns $\sigma^{-1}\eta s^\top$, which are unchanged, and ξ-columns $\sigma^{-1}\eta\xi^\top$, which are negated. Hence every $\hat{g}_{s,k,t}$ is unchanged, every $\hat{g}_{\xi,k,t}$ is negated, $K_{\xi,t+1} = K_{\xi,t} - \alpha_t\hat{g}_{\xi,t}$ is negated and $K_{s,t+1}$ is unchanged; the stopping rule of Definition 6 depends on $K_{s,t}$ only. This completes the induction. By (i), (ii) and Definition 4 (with A1), $K_{\xi,0}$, $(K_{s,0}, \omega)$ and ε are independent, $K_{\xi,0} \overset{d}{=} -K_{\xi,0}$, and ε consists of independent centred Gaussians, so $-\varepsilon \overset{d}{=} \varepsilon$. Hence $(K_{s,0}, K_{\xi,0}, \omega, \varepsilon) \overset{d}{=} (K_{s,0}, -K_{\xi,0}, \omega, -\varepsilon)$, and the induction gives $(K_{s,u}, K_{\xi,u})_{u\leq t} \overset{d}{=} (K_{s,u}, -K_{\xi,u})_{u\leq t}$; with (iii), $\mathbb{E}K_{\xi,t} = -\mathbb{E}K_{\xi,t} = 0$. For the counterexample, let $K_{\xi,0} = 0$ and $b_i = c\,\eta^i_{\tau_i,1}\xi^i_{\tau_i,1}$, which is a function of $(\tau_i, h^i_{\tau_i})$ and the current gain. At $K_\xi = 0$ the return G_i and $\eta^i_{\tau_i}$ are independent of $\xi^i_{\tau_i}$, so $\mathbb{E}[G_i\eta^i_{\tau_i}(\xi^i_{\tau_i})^\top] = 0$, while $\mathbb{E}[b_i\eta^i_{\tau_i}(\xi^i_{\tau_i})^\top] = c\,\mathbb{E}[\eta_1\eta]\,\mathbb{E}[\xi_1\xi]^\top = c\,e_1e_1^\top$. So $\mathbb{E}\hat{g}_{\xi,k,0} = -(c/\sigma)e_1e_1^\top$ for every client, and $\mathbb{E}K_{\xi,1} = (\alpha_0 c/\sigma)e_1e_1^\top$. □

Proof of Theorem 4(e). *(e1).* The pooled loss $K^{-1}\sum_k \mathbb{E}_k(y - \phi^\top w)^2$ has normal equation $\sum_k(C_k w - C_k w_k^\star) = 0$, which gives $\bar{w}$. By client k's normal equation, $\mathbb{E}_k[\phi(y - \phi^\top w_k^\star)] = 0$, so $\mathbb{E}_k(y - \phi^\top w)^2 = \mathbb{E}_k(y - \phi^\top w_k^\star)^2 + \|w - w_k^\star\|_{C_k}^2$.

For raw features, $C_k = \mathrm{diag}(\Sigma_{s,k}, \boldsymbol{I}_m)$ and $\mathbb{E}_k[\phi y] = (\mathbb{E}_k[sy], 0)$, because ξ is independent of (s, y), centred and whitened. So $w_k^\star = (w_{k,s}^\star, 0)$ with $w_{k,s}^\star = \Sigma_{s,k}^{-1}\mathbb{E}_k[sy]$, $\bar{w} = (\bar{w}_s, 0)$ with $\bar{w}_s = (\sum_k \Sigma_{s,k})^{-1}\sum_k \Sigma_{s,k}w_{k,s}^\star$, and $\mathcal{R}_k^{\mathrm{raw}}(w_s, w_\xi) = \|w_s - w_{k,s}^\star\|_{\Sigma_{s,k}}^2 + \|w_\xi\|^2$. In particular $b_k^{\mathrm{raw}} = \|\bar{w}_s - w_{k,s}^\star\|_{\Sigma_{s,k}}^2$.

For $\phi = W_s s$, $C_k = A_k$, $w_k^{z\star} = W_s^{-\top}w_{k,s}^\star$ and $\bar{w}^z = W_s^{-\top}\bar{w}_s$. Then $b_k^{\mathrm{lat}} = (\bar{w}_s - w_{k,s}^\star)^\top W_s^{-1}A_k W_s^{-\top}(\bar{w}_s - w_{k,s}^\star) = b_k^{\mathrm{raw}}$. Neither involves n.

For a latent critic on $z = W_s s + E\xi$, $\mathbb{E}_k(y - w^\top z)^2 = \mathbb{E}_k(y - w^\top W_s s)^2 + \|E^\top w\|^2$, because the cross term $\mathbb{E}_k[(y - w^\top W_s s)\xi^\top]E^\top w$ vanishes and $\mathbb{E}\xi\xi^\top = \boldsymbol{I}$. Subtracting $\mathbb{E}_k(y - s^\top w_{k,s}^\star)^2$ and applying the Pythagorean identity in z-coordinates gives $\mathcal{R}_k^{\mathrm{lat}}(w; E)$.

(e2). The empirical normal equation $\hat{C}\hat{w} = \hat{\mathbb{E}}[\phi y]$, with $\hat{C} := \hat{\mathbb{E}}[\phi\phi^\top]$, gives $\hat{C}(\hat{w} - \bar{w}) = \hat{h}$. Also $\mathbb{E}\hat{C} = \bar{C}$ and $\mathbb{E}\hat{h} = K^{-1}\sum_k(\mathbb{E}_k[\phi y] - C_k\bar{w}) = 0$ by the pooled normal equation. On the event $\|\hat{C} - \bar{C}\| \leq (2\|\bar{C}^{-1}\|)^{-1}$ a Neumann series gives $\|\hat{C}^{-1} - \bar{C}^{-1}\| \leq 2\|\bar{C}^{-1}\|^2\|\hat{C} - \bar{C}\|$, so $\hat{w} - \bar{w} - \bar{C}^{-1}\hat{h} = (\hat{C}^{-1} - \bar{C}^{-1})\hat{h} = O_p((Kn)^{-1})$ whenever $\hat{C} - \bar{C}$ and $\hat{h}$ are $O_p((Kn)^{-1/2})$. Under independent streams and (IID) this holds by Chebyshev's inequality, because $\hat{C}$ and $\hat{h}$ are averages of Kn independent terms with finite variances (finite fourth moments) and means $\bar{C}$ and 0; under replication it holds for either sampling scheme, because $\hat{C} - \bar{C}$ and $\hat{h}$ are averages of r i.i.d. centred copies with finite variances. The remaining claims of (e2) concern the leading term $\bar{C}^{-1}\hat{h}$ only. For raw features $\bar{w}_\xi = 0$, so $\hat{h} = (\hat{h}_s, \hat{h}_\xi) = (\hat{\mathbb{E}}[se], \hat{\mathbb{E}}[\xi e])$ with $e = y - s^\top\bar{w}_s$. Both $\bar{C} = \mathrm{diag}(\bar{\Sigma}_s, \boldsymbol{I})$ and C_k are block diagonal, so $\mathbb{E}\|\bar{C}^{-1}\hat{h}\|_{C_k}^2 = \mathbb{E}\|\bar{\Sigma}_s^{-1}\hat{h}_s\|_{\Sigma_{s,k}}^2 + \mathbb{E}\|\hat{h}_\xi\|^2$. For the exo-free latent, $\bar{C}^{-1}\hat{h} = W_s^{-\top}\bar{\Sigma}_s^{-1}\hat{h}_s$, whose A_k-norm is the same first term. Together with (e1)

this gives $\mathrm{err}_k^{\mathrm{raw}} = \mathrm{err}_k^{\mathrm{lat}} + V_\xi$. Under independent streams the sandwich for $V_\xi = \mathbb{E}\|\hat{\mathbb{E}}[e\,\xi^\top]\|^2$ is (c4) with $d_z = 1$, $\tilde{\Gamma} = 1$ and $u = e$, since e is endogenous; under shared streams (c6) with the same choices adds the term equation 24 with $\mu_k = \mathbb{E}_k e$.

(e3). Under replication the population moments are averages over the multiset of environments, so $\bar{C}$, $\bar{w}$ and b_k do not depend on r. The pooled empirical moments are averages over r i.i.d. copies of the K_0-client data set, each with finite mean. By the strong law of large numbers, $\hat{C} \to \bar{C}$ and $\hat{\mathbb{E}}[\phi y] \to K_0^{-1} \sum_k \mathbb{E}_k[\phi y]$ almost surely. Hence $\hat{w} \to \bar{w}$ almost surely, and by continuity $\mathcal{R}_k(\hat{w}) \to b_k$ for the raw and the exo-free agents.

For a learned encoder $(\hat{W}_s, \hat{E})$ with $\hat{W}_s \to W_s$ and $\hat{E} \to 0$ in probability (Assumption 2 and (c3)), the latent regression moments converge in probability to those of the exo-free latent. Its realised risk, computed as in (e1), is $\|\hat{W}_s^\top \hat{w} - w^\star_{k,s}\|^2_{\Sigma_{s,k}} + \|\hat{E}^\top \hat{w}\|^2$, which tends to b_k in probability. If $b_k > 0$, the ratio therefore tends to 1 in probability.

For the first-order formula: the pooled score is the average of r i.i.d. centred copies, $\hat{h}(r) = r^{-1} \sum_c \hat{h}^{(c)}$, so $V_s(r) = V_s(1)/r$ and $V_\xi(r) = V_\xi(1)/r$ exactly. Substituting into $\mathrm{err}_k^{\mathrm{raw}}/\mathrm{err}_k^{\mathrm{lat}} = (b_k + V_s + V_\xi)/(b_k + V_s)$ gives the display. □

A.6 Numerical checks

Every identity and inequality stated in this appendix was also checked numerically on random instances of the model class, including edge cases (ρ near 0 and near 1, ill-conditioned and non-normal predictor matrices, $k < d$, heterogeneous clients):

- exact identities: Theorem 2(a) and the moment identities of Theorem 3(a,b) matched finite-difference and closed-form computations to 10^{-12};
- inequalities and stability criteria: Theorem 2(b)–(g) and the bounds of Theorems 3 and 4 were never violated over more than a thousand random instances per result;
- Monte-Carlo checks: the variance floor of Theorem 3(c) and the contamination bound of Theorem 4(c) were consistent with their stated values within sampling error.

The scripts are included in the supplementary code.

B Environments, clients and distractors

Physics contexts. Each task has four clients at equally spaced points of the task's one-dimensional uncertainty set.

Table 3: Client physics per task.

Task	Parameter	Clients	Source
Walker2d-v5	world friction	0.1, 1.4, 2.7, 4.0	RRLS (Zouitine et al., 2024)
Hopper-v5	world friction	0.1, 1.07, 2.03, 3.0	RRLS
Ant-v5	torso mass	0.1, 1.07, 2.03, 3.0	RRLS
HalfCheetah-v5	world friction	0.1, 1.07, 2.03, 3.0	RRLS
InvertedDoublePendulum-v5	pole masses	$\times$0.5, 0.83, 1.17, 1.5	CARL-style $\pm 50\%$
MassHopper	body masses	four fixed configurations	CausalGym (Causal Artificial Intelligence Lab, Columbia University, 2026)

Disclosed implementation details.

- *Friction.* RRLS sets only the floor's friction. MuJoCo takes the larger of a contact's two geom frictions, so values below the bodies' own friction would be inert. We therefore write the value to every geom's sliding friction, which makes the whole published interval effective.
- *MassHopper body indices.* CausalGym's MassHopper wrapper indexes bodies as thigh, leg and foot = 1, 2, 3. In Hopper-v5, however, body 0 is the world. The four configurations therefore scale the *torso* ($\times 4$, 0.5, 3, 0.7) and the *thigh* ($\times 3$, 0.7, 0.7, 3). We use the upstream code unchanged, so the clients' physics still differ as intended; only the names of the scaled bodies differ from the original description.

Distractors. The appended block evolves as

$$\xi_{t+1} = \rho\xi_t + \sqrt{1-\rho^2}\,\epsilon_t, \qquad \xi_0 \sim \mathcal{N}(0, \boldsymbol{I}),$$

redrawn at every reset. $\rho = 0$ is the i.i.d. protocol (Nikishin et al., 2022). Each client has an independent stream, seeded from the run seed and the client index. Observations, including the distractor block, pass through the same running normaliser.

C Implementation and hyperparameters

All agents share one PPO implementation. A raw-observation agent is the latent agent with the encoder replaced by the identity ($z = o$) and no auxiliary losses. Its actor and critic are then the standard two-layer, 64-unit tanh MLPs of PPO on MuJoCo. Every linear layer is initialised orthogonally, and the policy is a Gaussian with a state-independent log standard deviation. Algorithm 1 (Figure 3) summarises one client's round, and Table 4 lists the settings; each value was checked against the configuration of the runs.

Algorithm 1: latent federated PPO (one client's view)

for round $n = 1, \ldots, N$:

1. Collect 2048 steps with $\pi(\cdot \mid \mathrm{sg}[\phi(o)])$; normalise observations with the pooled statistics.
2. For 10 epochs of minibatches: minimise the PPO clipped loss (actor on $\mathrm{sg}[z]$) $+ c_v\mathcal{L}_V$, plus $\mathcal{L}_{\mathrm{rep}}$ in the same step (predictive) or in a separate step with its own optimiser (ports); update the target encoder, if any.
3. **if** $n \bmod E = 0$: send all parameters and normalisation moments; receive their average.

Figure 3: The federated PPO loop shared by all agents. Raw-observation agents use $z = o$ and $\mathcal{L}_{\mathrm{rep}} = 0$; FedProx adds its proximal term to the actor and critic, FedKL its KL penalties to the actor. Hyperparameters are in Table 4.

Table 4: Hyperparameters, identical for every agent unless stated.

Group	Parameter	Value
PPO	rollout length per round	2048 steps
	epochs / minibatch size	10 / 64 (no KL early stopping)
	learning rate	$3 \cdot 10^{-4}$, Adam ($\epsilon = 10^{-5}$), linearly annealed to 0
	γ / GAE λ	0.99 / 0.95
	clip / value clip	0.2 / yes
	value coef. / entropy coef.	0.5 / 0
	gradient clipping	norm 0.5, separately for policy, encoder and auxiliary parameters
	advantages	normalised per minibatch
	observation normalisation	running mean/variance, clipped to ± 10; pooled at every synchronisation
	reward normalisation	running return scale, clipped to ± 10; per client
Latent agents	encoder	tanh MLP $o \to 256 \to 256 \to z$
	latent	$z \in \mathbb{R}^{64}$; LayerNorm without affine parameters (SimSR: L2 normalisation instead)
	actor / critic heads	tanh MLP $z \to 64 \to 64$
	routing	critic routing (the actor reads $\mathrm{sg}[z]$)
Predictive	predictor	$f(z,a) = z + g(z) + B(z)a + r(z,a)$, hidden width 256, learning rate $\times 3$
	reward head	tanh MLP $(z,a) \to 64 \to 64 \to 1$
	prediction horizon H	4, open loop
	λ_{pred} / λ_{rew} / λ_{reg}	0.03 / 0.1 / 0.003 (VICReg target s.d. 1)
	target encoder	EMA with retention 0.99, updated once per round
	optimiser for $\mathcal{L}_{\mathrm{rep}}$	PPO's (one Adam, one backward pass)
Federation	clients K / synchronisation	4 / every 5 rounds
	steps per client	749,568 (366 rounds)
	aggregation	equal-weight average of every module and of the observation-normaliser moments

The SimSR and AC-State settings are given with their ports in Appendix D, and the matched-optimiser variant of the predictive agent in Appendix E.3.

Three implementation details. Under critic routing the actor reads $\mathrm{sg}[\phi(o)]$ from the current encoder at every minibatch, so value-loss updates to the encoder within one PPO update also change

the policy, and PPO's clipped ratio does not see that change. The code also offers a variant in which the actor reads a snapshot of the encoder taken at the start of each update (`actor_snapshot`), but no reported result uses it. Raw PPO clips the joint actor–critic gradient at norm 0.5; latent agents instead clip the policy, encoder, and auxiliary groups separately, also at 0.5. The matched clipping used by the routing control is noted in the main text. Finally, FedAvg-PPO and the predictive agent on Walker2d, Ant, Hopper and MassHopper, federated and single-agent, come from an earlier version of the runner whose logs lack configuration fields added later; with the current code, a FedAvg-PPO seed from that version (Walker2d, without distractors) reproduces its logged training bit for bit over its first two rounds.

Origin of the settings. The predictive agent's settings, including the routing, the prediction horizon and the loss weights, come from earlier single-agent experiments on overlapping environments. They are global across tasks and were not tuned for the conditions reported here.

D PORTED METHODS: SOURCES, DEVIATIONS AND CALIBRATION

Sources. We ported each method from its authors' code, at a pinned commit:

- *SimSR*: the official implementation (commit `9ecf13e`);
- *multi-step inverse*: the AC-State objective as released with Levine et al. (2024) (commit `5eb0d51`);
- *FedKL*: the official TRPO-style implementation (commit `540db76`);
- *FedProx*: the authors' perturbed-gradient update, in the copy vendored by the FedKL repository at the same commit.

None of the references runs as released in our setting. SimSR is SAC on pixels. AC-State trains no control policy: it discovers representations in tabular and grid-world settings with discrete actions. FedKL is written in TensorFlow 1 for Reacher and SUMO.

Common to all. Every ported method uses the same setup as the other agents: PPO on state vectors, four heterogeneous clients, FedAvg every 5 rounds, $749{,}568$ steps per client, and the same seeds, evaluation and statistics.

SimSR.

- *Kept from the reference:*
  - the SimSR loss, a Huber regression of pairwise cosine distances onto $|\Delta r| + \gamma(1 - \langle \hat{p}_i, \hat{p}_j \rangle)$;
  - the probabilistic transition ensemble (5×512), with weight decay 10^{-7};
  - the EMA target (retention 0.975 per minibatch) and the L2-normalised latent;
  - its own unclipped Adam at 10^{-3} for the encoder and the ensemble.
- *Different in our pipeline:* a state MLP instead of the pixel encoder, a PPO V-critic instead of SAC's twin Q, on-policy minibatches of 64 instead of replay batches of 128, and the scaled reward.
- *As in the official code but not in the paper's text:* predicted next latents are not re-normalised.

Multi-step inverse (AC-State objective).

- *The model:* a head predicts a_t from $[z_t, z_{t+k}, \mathrm{emb}(k)]$, with $k \sim U\{1, \ldots, \min(8, \text{steps left in the episode})\}$.
- *Different in our pipeline:*
  - a Gaussian negative log-likelihood, because our actions are continuous;
  - no vector-quantisation bottleneck, because the policy needs a continuous latent;
  - its own AdamW at 10^{-4} without weight decay, as in the reference.
- The original work trains no control policy, so this agent is our PPO instantiation of its objective.

FedProx. The proximal gradient $\mu(w - w_{\text{global}})$ is added to the actor and critic gradients after PPO's gradient clipping, and passes through Adam. The reference uses plain SGD, so a given μ acts differently here (Table 8 shows one consequence).

FedKL.

- *Added to PPO's clipped surrogate:* $\sigma\,\mathbb{E}_s\sqrt{\mathrm{KL}(\pi_{\text{global}}\|\pi)}$ and $\beta\,\mathbb{E}_s\mathrm{KL}(\pi_{\text{old}}\|\pi)$.
- *Adaptation:* σ and β adapt towards their targets with the reference's clamps.
- *Not ported:* the reference's local learner is the adaptive-KL variant of PPO, and its code retries a client's local training when the reward drops. We did not port the retry.

Representation optimisers. The SimSR and AC-State optimisers anneal with training, like every other optimiser in the pipeline, whereas the references use constant rates. We anneal them because, at a constant rate, their encoders kept drifting late in training and moved the policy two to three times as much as the predictive agent's did; the constant-rate versions were not run.

Calibration. The coefficients of the ported methods were tuned on seeds 900–902, which the comparison does not use. Each setting is scored by the clients' mean training return over the last 25 rounds, with ties broken by the final cross-evaluation, and a method keeps its default setting unless another setting beats it by more than one standard error of the paired difference. FedKL's base targets follow the authors' heuristic, 1.1 times the global and local divergences observed under FedAvg. Table 5 lists the grids and the chosen settings.

Table 5: Calibration grids and chosen settings (seeds 900–902).

Method	Grid	Chosen
FedProx, μ (per task)	$\{10^{-3}, 10^{-2}, 3\cdot10^{-2}, 10^{-1}, 1\}$	10^{-1} on Walker2d, 10^{-3} elsewhere
FedKL, global target (per task)	$\{0.5, 1, 2\}\times \mathrm{nm}^\star$, and $1\times$ without the local penalty	$2\times$ on HalfCheetah, $1\times$ elsewhere
SimSR, representation rate (global)	$\{\frac{1}{3}, 1, 3\}\times10^{-3}$	10^{-3} (the reference)
AC-State, representation rate (global)	$\{\frac{1}{3}, 1, 3\}\times10^{-4}$	10^{-4} (the reference)

The FedKL base targets $(\mathrm{nm}^\star, \mathrm{kl}^\star)$ were:

- MassHopper, (0.748, 0.137); Walker2d, (0.407, 0.0353); Hopper, (0.722, 0.127);
- Ant, (0.455, 0.0472); HalfCheetah, (0.413, 0.0375); InvertedDoublePendulum, (0.907, 0.416).

For SimSR the best rate was $\frac{1}{3}\times$, with an advantage of $+0.21 \pm 0.41$ in normalised return. That is below the rule's threshold, so the reference rate was kept.

E ADDITIONAL RESULTS

End of the training budget. Over the last 100 rounds, the critic-trained encoder's training return is still rising on every task, by 111 to 1605 per 100 rounds (the mean over 10 seeds of each seed's least-squares slope; `analysis/late_slopes.py`), whereas FedAvg-PPO's changes by -63 to $+13$ per 100 rounds on five tasks and rises by 132 on Ant. The ratios in Table 1 are therefore values at the end of the budget rather than asymptotes.

Own return. Table 6 repeats Table 1 with each client's final policy evaluated on its own context only. Every agent with critic routing stays above every aggregation baseline on every task; within each group the order changes on some tasks.

Aggregation rules against FedAvg.

Noise weight in every network. FedProx's critic keeps far more noise weight than FedAvg's or FedKL's: 20–75% against 4–11%. This holds even on the five tasks where $\mu = 10^{-3}$. Its proximal term is added after gradient clipping and passes through Adam, whose per-parameter normalisation can let a small term dominate weights whose PPO gradient is small. We have not isolated this cause. It does not change FedProx's return much (Table 7), and the actor, which reads the raw observation, keeps its initial share under all three rules.

Distractor count and correlation (predictive agent). With 16 distractors the predictive agent is not always ahead: it trails FedAvg-PPO on MassHopper and is level with it on Hopper. At $m = 16$

Table 6: Own return (mean over the diagonal of the cross-evaluation matrix), 64 i.i.d. distractor dimensions, 10 seeds, mean ± s.e.

	Agent	Walker2d	Ant	Hopper	MassHopper	HalfCheetah	InvDblPend.
raw obs.	FedAvg-PPO	328±2	567±22	385±16	304±39	434±23	293±6
	FedProx-PPO	375±6	534±16	343±18	342±34	315±29	282±5
	FedKL-PPO	326±4	552±21	393±15	311±42	441±40	289±7
encoder only	Jointly trained encoder	389±7	507±26	260±14	231±13	452±19	289±5
	Critic-trained encoder	1380±82	831±36	1097±63	1006±80	2137±83	2419±344
with objective	Predictive (JEPA-style)	724±68	1113±101	648±26	917±62	824±74	2084±268
	Behavioural metric (SimSR)	1332±112	1728±93	427±20	779±69	2125±219	5476±253
	Multi-step inverse (AC-State)	1274±183	2391±250	832±36	897±121	2559±135	7408±209

Table 7: FedProx-PPO and FedKL-PPO against FedAvg-PPO: difference in mean global return, seeds in which the rule is higher, exact sign-flip p, and Holm-adjusted p over these 12 comparisons.

Rule	Task	Δ vs FedAvg	Seeds higher	p	p_{Holm}
FedProx	Walker2d	$+13.2\%$	10/10	0.002	0.023
	Ant	-5.9%	4/10	0.195	1
	Hopper	-10.1%	1/10	0.016	0.172
	MassHopper	$+11.5\%$	6/10	0.369	1
	HalfCheetah	-25.1%	2/10	0.016	0.172
	InvDblPend.	-4.3%	3/10	0.135	1
FedKL	Walker2d	-0.3%	4/10	0.822	1
	Ant	$+2.3\%$	7/10	0.486	1
	Hopper	$+1.8\%$	4/10	0.740	1
	MassHopper	$+3.3\%$	4/10	0.863	1
	HalfCheetah	-4.0%	4/10	0.252	1
	InvDblPend.	$+0.3\%$	4/10	0.941	1

AC-State beats FedAvg-PPO on Walker2d, Ant and Hopper, and the predictive agent on MassHopper (Table 17), so the shortfall is specific to the predictive agent. Figures 5 and 6 show the absolute global returns behind Figure 4, for federated and single agents, and Table 9 gives every cell with its test.

Federation versus single agents (predictive agent). Table 10 compares each agent trained alone on its own client with the same agent federated, at 64 dimensions.

What to share (predictive agent). Averaging only the world model, the world model without its predictor, or only the actor and critic is worse than averaging everything in all four settings. Averaging the representation alone collapses on Ant. Averaging only the predictor is mixed: better on Walker2d i.i.d. ($+239$, $p = 0.08$) and much worse on Ant. On Walker2d with i.i.d. noise, single agents reach a higher return than every federated variant, matching the negative federation effect in Table 10. We ran this study for the predictive agent only.

Clean-observation reference. Without distractors, federated PPO beats the predictive agent on three of the four tasks we ran. Its global return is higher on Walker2d (1375 ± 229 against 805 ± 193), MassHopper (1461 ± 74 against 577 ± 91) and HalfCheetah (2183 ± 113 against 1929 ± 338). On Ant the two are level (1282 ± 82 against 1341 ± 118). These runs come from an earlier batch with the same runner, contexts and budget, and have 5 seeds (on MassHopper, 6 for PPO and 2 for the predictive agent). With 64 i.i.d. distractor dimensions, PPO's global return falls by 55–81% on these tasks, and the predictive agent's changes by -55% to $+56\%$. The predictive agent's advantage in Table 1 exists only with distractors. On HalfCheetah with 16 AR(1) distractor dimensions the two fall to 986 and 790.

Clean-observation control. FedAvg-PPO and the critic-trained encoder were run without distractors on the six tasks, with 10 seeds and every other setting of Table 1. The analysis is `analysis/clean_analysis.py` in the supplementary material. Without distractors the en-

Table 8: First-layer noise-weight share of every network after training (mean over clients, then seeds), including the FedProx and FedKL agents.

Network	Walker2d	Ant	Hopper	MassHopper	HalfCheetah	InvDblPend.
Initialisation, $64/(64+d)$	79.0 / –	37.9 / –	85.3 / –	85.3 / –	79.0 / –	87.7 / –
FedAvg-PPO, actor	77.6 / 49.0	38.9 / 25.8	84.0 / 41.4	83.5 / 38.2	77.5 / 55.4	77.2 / 6.9
FedAvg-PPO, critic	8.3 / –	3.9 / –	6.1 / –	10.5 / –	8.8 / –	9.9 / –
FedProx-PPO, actor	78.5 / 54.3	38.8 / 25.2	84.4 / 43.5	83.9 / 41.4	77.8 / 57.6	78.9 / 8.0
FedProx-PPO, critic	75.0 / –	19.7 / –	60.2 / –	62.0 / –	44.6 / –	37.7 / –
FedKL-PPO, actor	77.6 / 49.1	38.7 / 25.1	83.9 / 40.9	83.6 / 38.8	77.4 / 54.9	76.9 / 6.9
FedKL-PPO, critic	8.3 / –	4.0 / –	5.3 / –	9.8 / –	8.9 / –	9.8 / –
Jointly trained encoder	73.0 / 54.4	38.7 / 40.3	79.9 / 53.1	80.0 / 51.0	74.6 / 48.9	51.0 / 1.8
Critic-trained encoder	22.5 / 0.5	31.0 / 6.1	35.8 / 0.6	32.4 / 0.6	51.0 / 1.7	52.2 / 0.5
Predictive	18.2 / 6.6	20.0 / 3.2	23.9 / 7.3	25.3 / 7.4	46.7 / 7.0	46.9 / 0.2
SimSR	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.1 / 0.0	0.1 / 0.0
AC-State	9.3 / 0.2	11.7 / 0.7	10.1 / 0.1	11.7 / 0.1	28.4 / 0.5	37.8 / 0.1

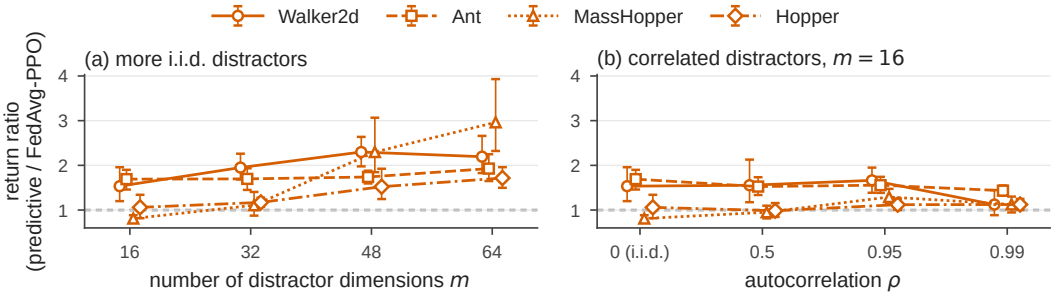

Figure 4: Federated global return of the predictive agent over FedAvg-PPO's, as i.i.d. distractors grow in number (a) and as 16 distractors become correlated (b); 10 seeds, 95% bootstrap intervals, dashed line at parity, categorical ρ axis. Absolute returns, single agents and tests follow.

coder is worse than FedAvg-PPO on Walker2d, Ant, MassHopper and InvertedDoublePendulum, and not significantly different on Hopper and HalfCheetah. The interaction is positive on every task (Table 12). Its gain is therefore specific to the distractors on the first four tasks. On Hopper and HalfCheetah its effect without distractors is unresolved. On Walker2d, Hopper and MassHopper the encoder's return is higher with 64 noise dimensions than without (1391 against 693 on Walker2d). Since a pure filter could at best recover the clean return, part of the interaction on these tasks must come from the distractors helping the encoder rather than from the encoder removing them; we have not found the cause.

E.1 Controls for the critic-trained encoder

Three controls concern the critic-trained encoder at $m = 64$ (analysis: `analysis/headline_analysis.py` in the supplementary material). Single agents, run on the four tasks of the single-agent comparison, test whether federation matters for it. The narrow critic-trained encoder has one hidden layer of 64 units, so every layer on its critic's path is as wide as PPO's; on Walker2d it has 26,509 parameters, against 19,277 for PPO and 120,333 for the critic-trained encoder. FedAvg-PPO's headroom is measured on the three tasks where it loses the largest share of its clean return, with one setting that lowers the step size from 3×10^{-4} to 10^{-4} and another that raises the minibatch from 64 to 256. This check uses seeds 0–4 and selects the better setting on the same seeds, which favours FedAvg-PPO, so it reports no tests.

Single agents. Trained alone, the critic-trained encoder still beats single-agent PPO on all four tasks, in 10 of 10 seeds (Table 13). Federation raises its return on all four, by 207 to 331, an effect that is significant after Holm correction on Ant and Hopper and unresolved on Walker2d and MassHopper. Only on Hopper does federation help the encoder significantly more than it helps PPO.

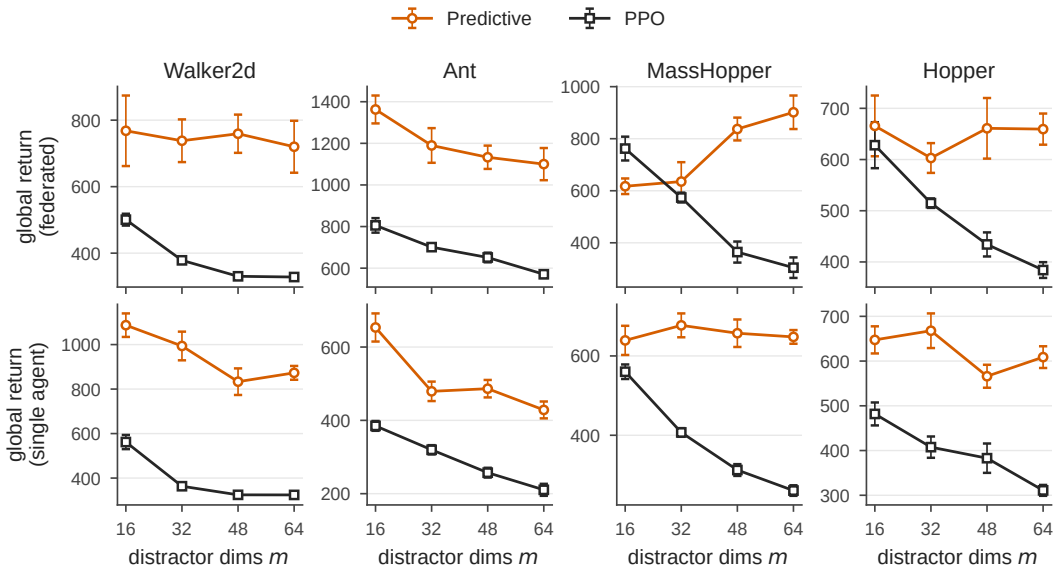

Figure 5: Global return of the predictive agent and PPO as the number of i.i.d. distractor dimensions grows, federated (top; PPO is FedAvg-PPO) and single agents (bottom). 10 seeds, $\pm$ s.e.

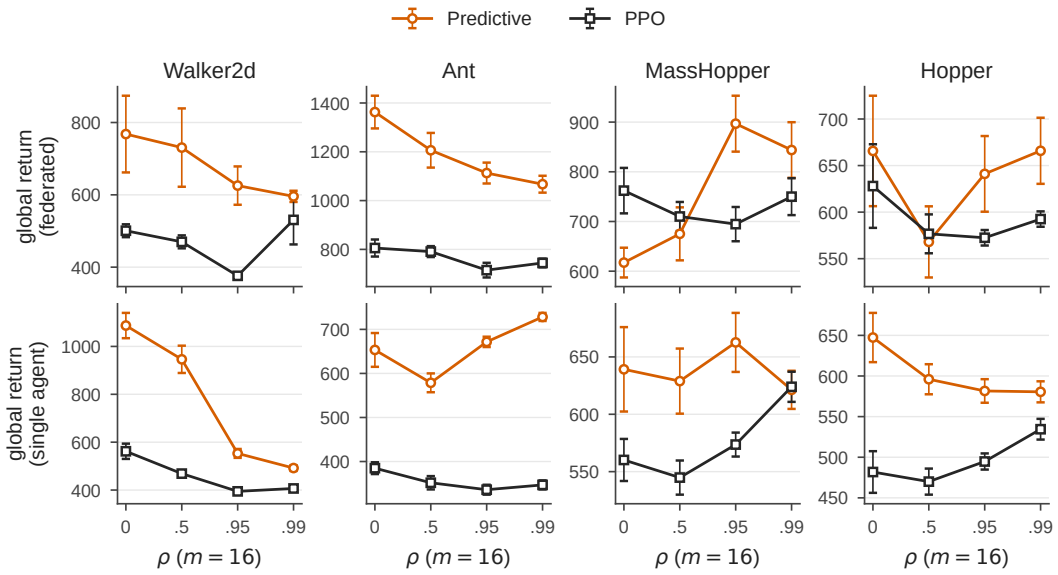

Figure 6: As Figure 5, for 16 distractor dimensions with autocorrelation ρ ($\rho = 0$ is i.i.d.).

For both agents, the federation effect includes the pooled observation normaliser as well as parameter averaging.

Critic routing at PPO's width. The narrow critic-trained encoder beats FedAvg-PPO on every task, in 10 of 10 seeds (Table 14). On five tasks the wider encoder adds nothing significant, the narrow one reaching 78–141% of its gain, and on InvertedDoublePendulum the wider encoder is significantly worse (-5631): there the narrow one reaches 7939, above FedAvg-PPO's clean return of 7532.

Table 9: Full grid: global return, 10 seeds, single agent vs. federated, every distractor setting. Gain = latent agent over PPO (seeds in which it is higher, out of 10). Fed. effect = federated over single for the same agent. Bold: $p \leq 0.05$, exact paired sign-flip, uncorrected. After Holm correction within each family, the gain is significant in 14 of the 16 federated distractor-count cells (one of them the negative MassHopper cell at 16 dimensions), 5 of the 12 federated correlated cells, 15 of 16 single-agent distractor-count cells and 9 of 12 single-agent correlated cells.

Task	Distractors	Single: PPO → predictive	Gain	Fed.: PPO → predictive	Gain	Fed. eff. PPO	Fed. eff. predictive
Walker2d	16 i.i.d.	562 → 1087	**+93%** (10/10)	501 → 768	**+53%** (9/10)	−11% (3/10)	−29% (1/10)
	32 i.i.d.	363 → 994	**+173%** (10/10)	378 → 738	**+95%** (10/10)	+4% (7/10)	−26% (1/10)
	48 i.i.d.	325 → 833	**+157%** (10/10)	330 → 759	**+130%** (10/10)	+2% (5/10)	−9% (4/10)
	64 i.i.d.	324 → 873	**+169%** (10/10)	328 → 720	**+119%** (10/10)	+1% (6/10)	−18% (3/10)
	16 ρ=.5	469 → 946	**+102%** (10/10)	470 → 731	**+55%** (9/10)	+0% (7/10)	−23% (1/10)
	16 ρ=.95	395 → 553	**+40%** (10/10)	376 → 626	**+66%** (10/10)	−5% (5/10)	+13% (6/10)
	16 ρ=.99	407 → 492	**+21%** (10/10)	531 → 596	+12% (6/10)	+30% (5/10)	+21% (10/10)
Ant	16 i.i.d.	385 → 653	**+70%** (10/10)	806 → 1363	**+69%** (9/10)	+109% (10/10)	+109% (10/10)
	32 i.i.d.	320 → 479	**+50%** (8/10)	701 → 1190	**+70%** (10/10)	+119% (10/10)	+148% (10/10)
	48 i.i.d.	257 → 486	**+89%** (10/10)	651 → 1133	**+74%** (10/10)	+153% (10/10)	+133% (10/10)
	64 i.i.d.	211 → 428	**+104%** (10/10)	571 → 1100	**+93%** (10/10)	+171% (10/10)	+157% (10/10)
	16 ρ=.5	352 → 579	**+65%** (10/10)	791 → 1206	**+53%** (10/10)	+125% (10/10)	+108% (10/10)
	16 ρ=.95	336 → 672	**+100%** (10/10)	715 → 1113	**+56%** (10/10)	+113% (10/10)	+66% (10/10)
	16 ρ=.99	347 → 728	**+110%** (10/10)	744 → 1067	**+43%** (10/10)	+115% (10/10)	+47% (10/10)
MassHopper	16 i.i.d.	560 → 639	+14% (8/10)	762 → 617	**−19%** (0/10)	+36% (9/10)	−3% (3/10)
	32 i.i.d.	407 → 677	**+66%** (10/10)	574 → 635	+11% (6/10)	+41% (10/10)	−6% (4/10)
	48 i.i.d.	312 → 657	**+110%** (10/10)	364 → 837	**+130%** (10/10)	+17% (8/10)	+27% (8/10)
	64 i.i.d.	261 → 648	**+149%** (10/10)	304 → 902	**+196%** (10/10)	+17% (6/10)	+39% (8/10)
	16 ρ=.5	545 → 629	**+15%** (9/10)	710 → 675	−5% (5/10)	+30% (9/10)	+7% (4/10)
	16 ρ=.95	574 → 663	**+16%** (8/10)	695 → 897	**+29%** (9/10)	+21% (10/10)	+35% (10/10)
	16 ρ=.99	624 → 621	−0% (5/10)	750 → 844	+13% (7/10)	+20% (8/10)	+36% (10/10)
Hopper	16 i.i.d.	482 → 647	**+34%** (9/10)	628 → 666	+6% (6/10)	+30% (7/10)	+3% (5/10)
	32 i.i.d.	408 → 668	**+64%** (10/10)	515 → 603	**+17%** (8/10)	+26% (10/10)	−10% (3/10)
	48 i.i.d.	383 → 566	**+48%** (9/10)	434 → 661	**+52%** (10/10)	+13% (6/10)	+17% (6/10)
	64 i.i.d.	311 → 609	**+96%** (10/10)	384 → 659	**+72%** (10/10)	+24% (8/10)	+8% (6/10)
	16 ρ=.5	470 → 596	**+27%** (10/10)	577 → 568	−2% (3/10)	+23% (8/10)	−5% (4/10)
	16 ρ=.95	495 → 582	**+18%** (9/10)	572 → 641	+12% (7/10)	+16% (10/10)	+10% (7/10)
	16 ρ=.99	534 → 581	+9% (7/10)	593 → 666	+12% (6/10)	+11% (9/10)	+15% (8/10)

Table 10: Global return with 64 i.i.d. distractor dimensions, single agents (each client alone, no sharing) versus federation (all modules averaged), 10 seeds. Gain = predictive latent agent over PPO in the same mode. Fed. effect = federated over single for the same agent. All gains are significant at $p \leq 0.05$ (paired sign-flip).

	Single agent: PPO → predictive		Federated: PPO → predictive		Fed. effect	
Task	return	gain	return	gain	PPO	predictive
Walker2d	324 → 873	+169%	328 → 720	+119%	+1%	−18%
Ant	211 → 428	+104%	571 → 1100	+93%	+171%	+157%
Hopper	311 → 609	+96%	384 → 659	+72%	+24%	+8%
MassHopper	261 → 648	+149%	304 → 902	+196%	+17%	+39%

Headroom of FedAvg-PPO. The better of the two settings raises FedAvg-PPO's return by 39 to 127, which closes only 3–7% of the gap to the critic-trained encoder (Table 15), and on every task the critic-trained encoder's worst seed still beats the better setting's best seed.

E.2 Encoder controls: architecture, routing and objective

Design. Two agents without any self-supervised objective separate the ingredients of the latent agents. Both run with the settings of Table 1.

- *Critic-trained encoder.* The predictive agent without its objective: no predictor, reward head, VICReg or EMA target. The encoder is trained by the value loss alone, and the actor reads $\mathrm{sg}[z]$.
- *Jointly trained encoder.* The same architecture with joint routing, so the policy gradient trains the encoder too.

Each difference isolates one ingredient, with caveats:

Table 11: Averaging different modules of the predictive agent (16 distractor dimensions, 10 seeds; global return). "World model" is the encoder, target, predictor and reward head.

Averaged modules	Walker2d i.i.d.	Walker2d AR(1)	Ant i.i.d.	Ant AR(1)
everything	768	626	1363	1113
world model only	459	361	44	−140
world model without predictor	451	347	−30	−198
predictor only	1007	548	580	676
actor and critic only	329	311	796	836
nothing (single agents)	1087	553	653	672
PPO, everything (FedAvg)	501	376	806	715

Table 12: Clean-observation control: global return of FedAvg-PPO and the critic-trained encoder without distractors ($m = 0$) and with 64 i.i.d. dimensions (10 paired seeds each). Δ_0: encoder − PPO at $m = 0$; interaction: $\Delta_{64} - \Delta_0$ per seed; Holm-adjusted p over the six tasks in parentheses. Last column: share of PPO's drop from $m = 0$ to $m = 64$ that the encoder recovers at $m = 64$.

Task	PPO, $m = 0$	encoder, $m = 0$	PPO, $m = 64$	encoder, $m = 64$	Δ_0	interaction	recovered
Walker2d	1491	693	328	1391	−798 (0.012)	+1861 (0.012)	91%
Ant	1265	774	571	805	−491 (0.012)	+726 (0.012)	34%
Hopper	922	798	384	1016	−124 (0.574)	+756 (0.012)	117%
MassHopper	1324	763	304	991	−561 (0.016)	+1248 (0.012)	67%
HalfCheetah	2276	2544	422	2064	+268 (0.574)	+1374 (0.012)	89%
InvDblPend.	7532	6726	294	2308	−805 (0.018)	+2820 (0.012)	28%

- jointly trained encoder − FedAvg-PPO: capacity, the LayerNorm latent and an actor–critic encoder shared by both heads;
- critic-trained − jointly trained encoder: routing;
- predictive − critic-trained encoder: the objective;
- SimSR − critic-trained encoder: the objective together with SimSR's L2-normalised latent and its own optimiser;
- AC-State − critic-trained encoder: the objective together with AC-State's own optimiser.

Tests are exact paired sign-flip tests. Holm is applied over each row's 6 tasks.

Reporting rule. With 10 seeds the smallest attainable p is 0.00195, so a non-significant difference is weak evidence. A Holm-significant difference is reported with its size. A non-significant one is attributed to the critic-trained encoder only if the upper end of its 95% interval is below a tenth of the critic-trained encoder's own gain over FedAvg-PPO. Otherwise it is reported as unresolved; the interval comes from inverting the sign-flip test. The same rule applies to the routing and capacity rows. Intervals for every cell are in the supplementary material (`docs/baselines/SCAFFOLD_RESULTS.md`).

AC-State with 16 distractor dimensions. With few distractors the predictive agent is not always ahead of FedAvg-PPO (Section 7.4). To see whether this holds for the family, AC-State was run at $m = 16$ on the four dose tasks. Its runs are paired with the existing 16-dimension cells (Table 17).

E.3 Matched-optimiser control

Design. The ports differ from the predictive agent in the objective and in the optimiser that applies it. To separate the two for the predictive objective, we trained a variant of the predictive agent that keeps every loss term and weight. The only change is that the representation losses leave PPO's optimiser. They get their own unclipped Adam at SimSR's rate (10^{-3}) on encoder, predictor and reward head. It steps after every PPO minibatch on the updated encoder and anneals like PPO's rate. This is how the SimSR port applies its loss. PPO's optimiser then trains the encoder through the value loss only, as in the ports. AC-State's optimiser (10^{-4}) was not transplanted, and we did not run a variant at that rate.

Table 13: Single agents against federation, critic-trained encoder and FedAvg-PPO (10 paired seeds). Federation effect: federated − single-agent return of the encoder; single agents: encoder − PPO, both trained alone; difference: the encoder's federation effect minus PPO's. Holm-adjusted p over four tasks in parentheses.

Task	encoder, fed.	encoder, single	PPO, fed.	PPO, single	federation effect	single agents	difference
Walker2d	1391	1148	328	324	+243 (0.070)	+824 (0.008)	+239 (0.158)
Ant	805	496	571	211	+310 (0.008)	+285 (0.008)	−50 (0.188)
Hopper	1016	685	384	311	+331 (0.008)	+374 (0.008)	+258 (0.008)
MassHopper	991	784	304	261	+207 (0.070)	+523 (0.008)	+163 (0.195)

Table 14: Critic routing at PPO's width (federated, 10 paired seeds). Narrow: the critic-trained encoder with one hidden layer of 64 units. Holm-adjusted p over six tasks in parentheses. Last column: share of the critic-trained encoder's gain over FedAvg-PPO that the narrow encoder reaches, with a 95% bootstrap interval.

Task	FedAvg-PPO	narrow	critic-trained	narrow − PPO	critic-trained − narrow	share reached
Walker2d	328	1301	1391	+972 (0.012)	+91 (1.000)	91% [76, 110]
Ant	571	902	805	+331 (0.012)	−96 (0.312)	141% [107, 190]
Hopper	384	878	1016	+494 (0.012)	+138 (0.367)	78% [60, 99]
MassHopper	304	1018	991	+714 (0.012)	−27 (1.000)	104% [79, 141]
HalfCheetah	422	2055	2064	+1633 (0.012)	+9 (1.000)	99% [87, 114]
InvDblPend.	294	7939	2308	+7645 (0.012)	−5631 (0.012)	380% [296, 521]

The change is larger than its description suggests.

- *Effective weight.* A separate Adam normalises the representation gradient by itself, so the objective's effective weight on the encoder rises sharply. Over the first 60–90 rounds of seed 0 on each task, the representation step moved the encoder 1.2–2.3 times as far per round as the PPO step, and partly against it (cosine -0.2 to -0.5). In the original agent the objective's gradient was a 2–7% share of the encoder's.
- *Incidental changes.* The predictor's rate changes from $9 \cdot 10^{-4}$, three times the encoder's, to 10^{-3}, and the reward head's from $3 \cdot 10^{-4}$. Adam's ϵ is 10^{-8} instead of 10^{-5}, and there is no clipping.
- *Unchanged.* The EMA target is still updated once per round.

Protocol. Its primary tests compare it with the original agent, with Holm over the 6 tasks and 95% intervals by inverting the sign-flip test. The variant uses the seeds, flags and budget of Table 1. Everywhere else, the predictive agent is the original one.

Results. The variant's return is higher than the original agent's on every task (Table 18). The difference is significant on Hopper, Ant, HalfCheetah and InvertedDoublePendulum, and not on Walker2d or MassHopper, where the intervals still include gains of up to 425 and 379. Where the SimSR port was ahead of the original agent (Walker2d, Ant, HalfCheetah, InvertedDoublePendulum), the variant closes 29–110% of the gap. On those tasks it closes 33–62% of the gap to AC-State, and on Hopper it overtakes AC-State. The encoder's first-layer noise share falls to 0–6%, below the critic-trained encoder's on every task and 0.2% or less on four. The variant's secondary comparisons with the aggregation baselines are all significant (18 of 18, Holm within 18).

The comparison with the critic-trained encoder is exploratory (exact sign-flip tests, Holm over 6). Against it, the variant adds on Ant ($+742$) and InvertedDoublePendulum ($+2843$) and is lower on Walker2d (-491). It is not significantly different on Hopper (-65), MassHopper ($+87$) or HalfCheetah (-306). SimSR and AC-State show the same pattern on Ant and InvertedDoublePendulum (Table 16). With a matched optimiser, the predictive objective adds on the tasks where the two ported ones do, although it still trails AC-State on Ant, HalfCheetah and InvertedDoublePendulum. The control changes the optimiser and, with it, the objective's effective weight; it does not separate the two.

Table 15: Headroom of FedAvg-PPO at $m = 64$ (federated, seeds 0–4; descriptive). Each setting's mean, with its paired difference to FedAvg-PPO in parentheses. Gap closed: (better setting − FedAvg-PPO) / (critic-trained encoder − FedAvg-PPO), with a 95% bootstrap interval; the better setting is chosen on the same seeds. Last column: seeds on which the critic-trained encoder beats the better setting.

Task	FedAvg-PPO	step size 10^{-4}	minibatch 256	critic-trained	gap closed	seeds
InvDblPend.	287	402 (+115)	414 (+127)	2166	7% [4, 12] (minibatch)	5/5
HalfCheetah	443	370 (−73)	488 (+45)	2083	3% [-2, 7] (minibatch)	5/5
Walker2d	330	369 (+39)	352 (+21)	1282	4% [2, 7] (step size)	5/5

Table 16: Differences in global return (mean over 10 paired seeds; Holm-adjusted p in parentheses), 64 i.i.d. distractor dimensions.

Difference	Walker2d	Ant	Hopper	MassHopper	HalfCheetah	InvDblPend.
Jointly trained encoder − FedAvg-PPO	+66 (0.012)	−80 (0.117)	−127 (0.012)	−74 (0.258)	+20 (0.844)	+6 (0.844)
Critic-trained − jointly trained encoder	+997 (0.012)	+315 (0.012)	+759 (0.012)	+760 (0.012)	+1622 (0.012)	+2008 (0.012)
Predictive − critic-trained encoder	−671 (0.018)	+295 (0.012)	−357 (0.012)	−89 (0.965)	−1198 (0.012)	−220 (0.965)
SimSR − critic-trained encoder	−44 (1.000)	+947 (0.012)	−586 (0.012)	−220 (0.123)	−26 (1.000)	+2562 (0.012)
AC-State − critic-trained encoder	−117 (1.000)	+1550 (0.012)	−214 (0.012)	−111 (1.000)	+519 (0.029)	+4700 (0.012)

F Statistics and compute

Pairing and tests. A seed fixes the clients' network initialisation, environment seeds and distractor streams. Every method runs on seeds 0–9, so comparisons are paired by seed. For paired differences $\delta_1, \ldots, \delta_n$ we use the exact two-sided sign-flip test, whose p-value is computed over all 2^n sign assignments. With $n = 10$ the smallest attainable p is 0.002. Holm's step-down correction is applied within the family each claim refers to:

- each latent agent's 18 comparisons with the aggregation baselines (Table 1);
- the 12 comparisons of FedProx and FedKL with FedAvg;
- the 16 distractor-count and 12 correlated-noise cells of the dose study, separately for federated and single agents;
- each control of Appendix E over its own tasks: six for the encoder, clean-observation, width and matched-optimiser controls, and four for the single-agent and AC-State controls.

Tables report means ± standard errors over seeds.

Aggregate scores. As a robustness check on the per-task tests we follow Agarwal et al. (2021). Each run's global return is divided by FedAvg-PPO's mean global return on the same task. The 60 normalised scores of each agent (6 tasks × 10 seeds) are summarised by their interquartile mean (IQM), median and mean. The 95% intervals are percentile intervals from a stratified bootstrap that resamples seeds within each task (10,000 replicates). The probability of improvement $P(X > Y)$ is the chance that a random run of X beats a random run of Y on the same task. It is averaged over tasks, with 2,000 replicates for its interval.

The probability of improvement over every aggregation baseline is at least 0.99 for the predictive agent and AC-State, and 0.95–0.97 for SimSR, whose lower value comes from Hopper. Among the latent agents it is 0.85 [0.79, 0.91] for AC-State over the predictive agent, 0.75 [0.67, 0.83] for AC-State over SimSR and 0.70 [0.65, 0.75] for SimSR over the predictive agent.

Parameters, FLOPs and communication. Table 20 counts parameters and, per synchronisation, the number of values each client sends. FedAvg of all modules sends every parameter, including the EMA target. The table also gives the matrix-multiply FLOPs (forward and backward) of one update round, which is 10 epochs on a 2048-step buffer. PyTorch's FLOP counter measured them on Walker2d-sized inputs (17 state and 64 noise dimensions, 6 actions). Environment simulation and action selection are not included.

Table 17: AC-State at 16 i.i.d. distractor dimensions: differences in global return (mean over 10 paired seeds; Holm-adjusted p over the four tasks in parentheses).

Difference	Walker2d	Ant	MassHopper	Hopper
AC-State − FedAvg-PPO	+468 (0.020)	+1509 (0.008)	+144 (0.119)	+244 (0.012)
AC-State − predictive	+200 (0.436)	+952 (0.008)	+289 (0.012)	+207 (0.086)

Table 18: Matched-optimiser control: the predictive agent with its representation losses in their own unclipped Adam at 10^{-3} (the variant), against the original agent; 64 i.i.d. distractor dimensions, 10 paired seeds. Δ: mean paired difference in global return, with a 95% interval from inverting the sign-flip test; p: Holm over the 6 tasks; gap closed: (variant − original)/(port − original), given where the port's mean exceeds the original's. Last two columns, in %: first-layer noise-weight share of the encoder, and the actor's noise-sensitivity share (exploratory, as in Table 2); the original agent's value in parentheses.

Task	Variant	Original	Δ [95% CI]	seeds >	p	gap closed (SimSR / AC-State)	weight share	sensitivity
Walker2d	901±79	720	+181 [−82, +425]	9/10	0.170	0.29 / 0.33	0.2 (18.2)	0.1 (6.6)
Ant	1547±50	1100	+447 [+208, +676]	9/10	0.016	0.69 / 0.36	0.1 (20.0)	0.0 (3.2)
Hopper	952±53	659	+292 [+135, +437]	9/10	0.018	– / 2.04	0.0 (23.9)	0.0 (7.3)
MassHopper	1078±45	902	+177 [−25, +379]	7/10	0.168	– / –	0.0 (25.3)	0.0 (7.4)
HalfCheetah	1758±100	865	+893 [+585, +1202]	10/10	0.012	0.76 / 0.52	3.9 (46.7)	0.7 (7.0)
InvDblPend.	5151±158	2088	+3063 [+2476, +3641]	10/10	0.012	1.10 / 0.62	6.3 (46.9)	0.0 (0.2)

SimSR's cost is dominated by its transition ensemble (five networks of width 512, 1.82M parameters). The predictive agent's largest module is its predictor (183k parameters).

Hardware and total compute. Every run used one CPU core on a 24-core workstation (Intel Core Ultra 9 285K) shared with other jobs. Wall-clock times were measured under full load and are indicative only. The results in this paper come from 2,160 finished federated runs, including calibration, and took about 3,200 core-hours. The matched-optimiser control adds 60 runs, the encoder controls 120, AC-State at 16 dimensions 40, the clean-observation control 120, and the controls for the critic-trained encoder 130 (about 70 core-hours).

G Extended related work

Table 21 places our study next to the closest federated and representation-learning work. The closest federated study, FedRAG (Guo et al., 2025), shares a bisimulation-style encoder across clients whose bodies differ, on DeepMind Control from pixels. It tests natural-video distractors on CartPole only, in the appendix of its extended version.

Table 19: FedAvg-normalised aggregate scores over 6 tasks $\times$ 10 seeds, with 95% stratified bootstrap intervals.

Agent	IQM	Median	Mean
FedAvg-PPO	0.99 [0.96, 1.03]	1.00 [0.96, 1.02]	1.00 [0.95, 1.05]
FedProx-PPO	0.97 [0.94, 1.01]	0.95 [0.93, 0.99]	0.97 [0.92, 1.01]
FedKL-PPO	1.01 [0.98, 1.04]	1.01 [0.98, 1.04]	1.01 [0.95, 1.05]
Jointly trained encoder	0.94 [0.91, 0.97]	0.95 [0.90, 0.97]	0.93 [0.90, 0.95]
Critic-trained encoder	3.64 [3.42, 3.87]	3.59 [3.18, 4.10]	4.05 [3.71, 4.39]
Predictive	2.22 [2.06, 2.40]	2.06 [1.96, 2.40]	2.99 [2.67, 3.31]
Behavioural metric (SimSR)	3.47 [3.17, 3.79]	3.33 [3.00, 3.49]	5.37 [5.09, 5.64]
Multi-step inverse (AC-State)	4.18 [3.76, 4.66]	4.00 [3.33, 4.80]	7.16 [6.77, 7.55]

Table 20: Model size, communication and update cost (Walker2d-sized inputs), with measured wall-clock per federated run.

Agent	Parameters	Values sent per sync	GFLOPs per update round	Wall-clock per run (h)
FedAvg / FedProx / FedKL-PPO	19,277	19,277	1.95	0.3–1.1
Critic-trained / jointly trained encoder	120,333	120,333	14	0.6–0.9
Predictive	312,142	415,374	120	1.7–2.5
Predictive, own optimiser (Appendix E.3)	312,142	415,374	132	–
SimSR	1,943,693	2,046,925	348	7.3–9.7
AC-State	237,657	237,657	52	1.4–1.8

Table 21: Positioning against the closest work. "–": not studied or not reported.

Work	Venue	FedRL	Distractors studied	Heterogeneous clients	Learned latent	Clients
FedProx (Li et al., 2020)	MLSys 2020	(supervised)	–	statistical	–	30–1000
FedKL (Xie & Song, 2023)	IEEE JSAC 2023	✓	–	goal region / action bias	–	≈60 / 7
QAvg / PAvg (Jin et al., 2022)	AISTATS 2022	✓	–	transition dynamics	–	–
PFedRL-Rep (Xiong et al., 2025)	ICLR 2025	✓	–	transition dynamics	shared representation	–
FedRAG (Guo et al., 2025)	IJCAI 2025	✓	video, extended version only	body dimensions	✓	2–4
SimSR (Zang et al., 2022)	AAAI 2022	–	video	–	✓	1
AC-State (Lamb et al., 2023)	TMLR 2023	–	✓	–	✓	1
This work		✓	none, i.i.d. and AR(1), up to 64 dims	friction, masses (RRLS, CausalGym)	value-trained, three objectives	4