AdamO: A Collapse-Suppressed Optimizer for Offline RL

Nan Qiao, Sheng Yue, Shuning Wang, Ju Ren

Paper | Project | Homepage

ABSTRACT

Offline RL critics fail spectacularly when bootstrapped temporal-difference (TD) updates amplify their own errors, driving Q-values to extreme, unusable magnitudes. Prior work blames the backup rule, the network architecture, or the squared-TD loss. **We show a more basic cause: the optimizer dynamics themselves can trigger or suppress collapse**. Treating offline TD learning as a feedback control system, we analyze Adam-based critic updates and derive a necessary and sufficient stability condition: the linearized critic is stable **iff** the spectral radius of an explicit augmented update operator stays strictly below one. A tractable sufficient form of this condition splits into a scale term (handled by standard normalization) and a geometry term governed by how far the weights deviate from isometry — motivating parameter orthogonality. Because a loss-based orthogonality penalty contaminates Adam's adaptive moments, orthogonality must instead be enforced as a **decoupled, optimizer-level correction**. We propose **AdamO**: Adam plus a budgeted orthogonality drift on selected weight blocks. AdamO provably never harms worst-case task descent and preserves Adam's dissipative continuous-time dynamics, yet is a drop-in replacement — just swap the critic's optimizer. Across D4RL it consistently improves stability and returns, with the largest gains where Adam collapses outright.

MOTIVATION

Minimizing the second moment of the temporal difference (TD) residual δ under the dataset distribution, i.e. $\min \mathbb{E}[\delta^2]$.

$$\delta \equiv r(\mathbf{s},\mathbf{a}) + \gamma\, \mathbb{E}_{\mathbf{a}'\sim\pi(\cdot|\mathbf{s}')} Q_{\phi'}(\mathbf{s}',\mathbf{a}') - Q_{\phi}(\mathbf{s},\mathbf{a})$$

Why Does Adam Trigger Value Collapse?

> *Adam is not "fast SGD."* Its momentum + variance adaptation turn TD learning into a **second-order** difference equation, so the right stability object is a spectral radius, not SGD intuition.

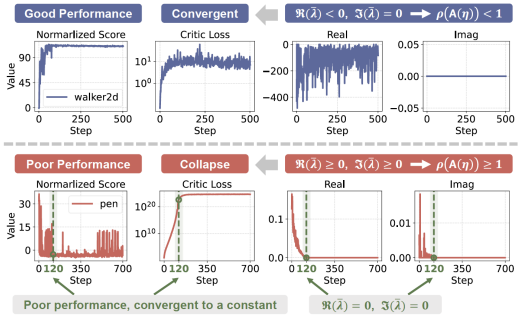
Linearized TD-error dynamics (frozen/terminal regime). With the Adam-preconditioned Gram operator $\mathsf{K}(X_1, X_2) \triangleq Z(X_1)^\top D\, Z(X_2)$, the bootstrapped update propagates through the greedy targets, giving the **TD dynamics operator**

$$\mathsf{S} \triangleq \gamma \mathsf{K}(X', X) - \mathsf{K}(X, X).$$

In **supervised learning, S is negative semidefinite**, the symmetric, marginally stable case of the Hurwitz condition, so no positive feedback, **no collapse**.

Coupling this with Adam's momentum EMA yields a closed-form second-order recurrence, summarized by the augmented operator

$$\mathsf{A}(\eta) \triangleq \begin{bmatrix} (1+\beta_1) I + \eta(1-\beta_1)\,\mathsf{S} & -\beta_1 I \\ I & 0 \end{bmatrix}.$$



Theorem 4.1 (necessary & sufficient). The frozen linearized TD error decays to 0 **iff** $\rho(\mathsf{A}(\eta)) < 1$.

Theorem 4.2 + Remark 4.3 (mechanism). If S is *Hurwitz*, a small enough stepsize guarantees $\rho < 1$. Collapse occurs exactly when bootstrapping turns into **positive feedback** that amplifies TD errors faster than the optimizer can damp them; at the boundary $(\Re(\bar\lambda) \to 0)$ a unit root appears and training **stagnates** instead of recovering.

So the entire fix reduces to one question answered next: how do we keep S Hurwitz? That is

How to Suppress It Without Breaking Adam?

A tractable **sufficient** condition for Hurwitzness (Prop. 5.1) cleanly separates two effects. With $\Phi = D^{1/2} Z(X)$, $\Phi_* = D^{1/2} Z(X')$:

$$\underbrace{\gamma\, \|\Phi\|_2\, \|\Phi_*\|_2}_{\text{(i) bootstrapped scale}} + \underbrace{\|\Phi^\top \Phi - I\|_2}_{\text{(ii) feature geometry}} < 1 \implies \mathsf{S} \text{ is Hurwitz.}$$

- **(i) Scale term** → already controlled by ubiquitous practice: input normalization / clipping + spectral-norm constraints.
- **(ii) Geometry term** → reduces to a *parameter-only* near-isometry defect $\varepsilon = \|\Psi^\top \Psi - I\|_2$, i.e. how far the weight blocks deviate from orthonormal. **This is the missing knob.**

Why not just add an orthogonality penalty to the loss? Running Adam on $\nabla \widetilde{L}_t = g_t + \lambda r_t$ feeds the orthogonality gradient into Adam's *second-moment* recursion

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)\,(g_t + \lambda r_t)^{\odot 2},$$

so λr_t leaks into (m_t, v_t) and **distorts future task steps even after it vanishes** (the same pathology AdamW fixes for weight decay).

PRACTICAL ALGORITHM

AdamO keeps Adam's moments driven only by the task gradient, and adds a per-layer orthogonality correction that is (a) scale-matched to the Adam step, (b) capped by a task-alignment budget, and (c) exactly Adam when κ=0.

For each constrained weight block W (regularizer $R(W) = \frac{1}{4}\|W^\top W - I\|_F^2$, closed-form gradient r_t):

$$\textbf{scale-match:}\quad \delta_{t,0} = \kappa \frac{\|u_t\|_F}{\|r_t\|_F + \varepsilon_r} r_t, \qquad \|\delta_t\| \leq \kappa \|u_t\|.$$

$$\textbf{budget:}\quad \langle g_t, \delta_t \rangle_F \geq -\tau \big(\langle g_t, u_t \rangle_F\big)_+, \qquad \tau \in [0,1),$$

with the largest feasible step taken in closed form (Eq. 19), and finally

$$\omega_{t+1} = \omega_t - \eta\,(u_t + \delta_t)$$

Algorithm 1 AdamO: Adam with Orthogonality Correction

Init $\omega_0, m_0 = v_0 = 0, \eta, \kappa, \tau$ and Adam params
for $t = 0, 1, \ldots$ **do**
  $g_t \leftarrow \nabla L_t(\omega_t)$
  Compute standard Adam update u_t — *Adam*
  Compute δ_t by applying (17)–(19) layer-wise — *Orthogonality Correction*
  $\omega_{t+1} \leftarrow \omega_t - \eta\,(u_t + \delta_t)$
end for

EXPERIMENTAL RESULTS

Scarce-data regime: D4RL with 10k state–action pairs. AdamO is dropped into six offline RL algorithms, TD3+BC, IQL, ReBRAC, ACTIVE, PARS, SQOG, replacing only the critic optimizer.

Task Name	TD3+BC		IQL		ReBRAC		ACTIVE		PARS		SQOG	
	Adam	AdamO	Adam	AdamO	Adam	AdamO	Adam	AdamO	Adam	AdamO	Adam	AdamO
AntMaze-umaze	72.2	**92.5**$_{\uparrow 28\%}$	80.5	**83.8**$_{\uparrow 4\%}$	94.3	**96.5**$_{\uparrow 2\%}$	92.4	**95.1**$_{\uparrow 3\%}$	**97.3**	93.5$_{\downarrow 4\%}$	89.6	**93.1**$_{\uparrow 4\%}$
AntMaze-umaze-div	47.0	**82.2**$_{\uparrow 75\%}$	55.8	**68.2**$_{\uparrow 22\%}$	87.2	**91.5**$_{\uparrow 5\%}$	75.9	**78.2**$_{\uparrow 3\%}$	**93.2**	92.4$_{\downarrow 1\%}$	72.8	**87.1**$_{\uparrow 20\%}$
AntMaze-med-play	0.3	**28.5**$_{\uparrow \infty}$	**70.4**	70.1$_{\downarrow 0.4\%}$	85.2	**89.4**$_{\uparrow 5\%}$	73.7	**86.5**$_{\uparrow 17\%}$	**91.5**	92.8$_{\uparrow 1\%}$	60.9	**65.4**$_{\uparrow 7\%}$
AntMaze-med-div	0.2	**16.4**$_{\uparrow \infty}$	66.9	**75.5**$_{\uparrow 13\%}$	79.8	**85.1**$_{\uparrow 7\%}$	77.8	**82.4**$_{\uparrow 6\%}$	**87.1**	90.6$_{\uparrow 4\%}$	65.8	**81.2**$_{\uparrow 23\%}$
AntMaze-large-play	0.0	**13.7**$_{\uparrow \infty}$	38.5	**41.8**$_{\uparrow 9\%}$	55.4	**61.8**$_{\uparrow 12\%}$	50.9	**56.5**$_{\uparrow 11\%}$	46.6	**52.9**$_{\uparrow 14\%}$	52.6	**57.4**$_{\uparrow 9\%}$
AntMaze-large-div	0.0	**16.5**$_{\uparrow \infty}$	32.4	**36.1**$_{\uparrow 11\%}$	50.5	**56.2**$_{\uparrow 11\%}$	46.6	**61.8**$_{\uparrow 33\%}$	50.8	**56.5**$_{\uparrow 11\%}$	47.5	**52.8**$_{\uparrow 11\%}$
AntMaze-ultra-div	0.0	**2.4**$_{\uparrow \infty}$	19.0	**31.1**$_{\uparrow 64\%}$	5.7	**9.4**$_{\uparrow 65\%}$	10.0	**19.8**$_{\uparrow 98\%}$	42.1	**48.6**$_{\uparrow 15\%}$	0.0	**4.2**$_{\uparrow \infty}$
AntMaze-ultra-play	0.0	**1.8**$_{\uparrow \infty}$	21.0	**29.9**$_{\uparrow 42\%}$	20.6	**26.8**$_{\uparrow 30\%}$	13.2	**11.5**$_{\downarrow 13\%}$	55.9	**62.4**$_{\uparrow 12\%}$	0.0	**2.1**$_{\uparrow \infty}$
AntMaze Avg.	15.0	**31.8**$_{\uparrow 112\%}$	48.1	**54.6**$_{\uparrow 14\%}$	59.8	**64.6**$_{\uparrow 8\%}$	55.1	**61.4**$_{\uparrow 11\%}$	**70.6**	73.7$_{\uparrow 4\%}$	48.7	**55.4**$_{\uparrow 14\%}$
HalfCheetah-m	35.9	**53.5**$_{\uparrow 49\%}$	29.7	**35.2**$_{\uparrow 19\%}$	42.8	**49.4**$_{\uparrow 15\%}$	42.9	**48.5**$_{\uparrow 13\%}$	44.9	**51.2**$_{\uparrow 14\%}$	36.5	**41.8**$_{\uparrow 15\%}$
HalfCheetah-mr	39.1	**46.2**$_{\uparrow 18\%}$	32.7	**38.4**$_{\uparrow 17\%}$	40.7	**47.1**$_{\uparrow 16\%}$	34.9	**40.6**$_{\uparrow 16\%}$	45.4	**50.8**$_{\uparrow 12\%}$	30.2	**35.5**$_{\uparrow 18\%}$
HalfCheetah-me	33.5	**60.1**$_{\uparrow 79\%}$	48.1	**54.6**$_{\uparrow 14\%}$	63.9	**71.5**$_{\uparrow 12\%}$	55.7	**62.3**$_{\uparrow 12\%}$	75.7	**82.9**$_{\uparrow 10\%}$	58.9	**65.4**$_{\uparrow 11\%}$
Hopper-m	40.7	**75.5**$_{\uparrow 86\%}$	38.9	**45.1**$_{\uparrow 16\%}$	78.6	**86.2**$_{\uparrow 10\%}$	26.2	**31.8**$_{\uparrow 21\%}$	73.7	**80.4**$_{\uparrow 9\%}$	45.8	**51.3**$_{\uparrow 12\%}$
Hopper-mr	21.3	**55.8**$_{\uparrow 162\%}$	46.6	**52.9**$_{\uparrow 14\%}$	64.2	**70.8**$_{\uparrow 10\%}$	62.5	**68.7**$_{\uparrow 10\%}$	67.5	**74.1**$_{\uparrow 10\%}$	61.4	**67.2**$_{\uparrow 9\%}$
Hopper-me	32.6	**84.2**$_{\uparrow 158\%}$	66.5	**73.2**$_{\uparrow 10\%}$	78.5	**85.9**$_{\uparrow 9\%}$	62.7	**69.1**$_{\uparrow 10\%}$	75.5	**81.6**$_{\uparrow 8\%}$	72.9	**78.5**$_{\uparrow 8\%}$
Walker2d-m	21.2	**68.4**$_{\uparrow 223\%}$	54.9	**61.5**$_{\uparrow 12\%}$	62.2	**69.4**$_{\uparrow 12\%}$	53.6	**59.2**$_{\uparrow 10\%}$	68.0	**74.5**$_{\uparrow 10\%}$	49.8	**55.6**$_{\uparrow 12\%}$
Walker2d-mr	19.3	**52.5**$_{\uparrow 172\%}$	51.4	**57.8**$_{\uparrow 12\%}$	69.4	**76.1**$_{\uparrow 10\%}$	45.8	**51.4**$_{\uparrow 12\%}$	63.4	**69.2**$_{\uparrow 9\%}$	58.5	**64.3**$_{\uparrow 10\%}$
Walker2d-me	22.4	**98.5**$_{\uparrow 340\%}$	57.3	**63.7**$_{\uparrow 11\%}$	74.0	**80.8**$_{\uparrow 9\%}$	58.6	**64.9**$_{\uparrow 11\%}$	81.6	**88.3**$_{\uparrow 8\%}$	68.7	**74.9**$_{\uparrow 9\%}$
Locomotion Avg.	29.6	**66.1**$_{\uparrow 123\%}$	47.3	**53.6**$_{\uparrow 13\%}$	63.8	**70.8**$_{\uparrow 11\%}$	49.2	**55.2**$_{\uparrow 12\%}$	66.2	**72.6**$_{\uparrow 10\%}$	53.6	**59.4**$_{\uparrow 11\%}$
Pen-human	−4.1	**83.1**$_{\uparrow \infty}$	79.1	**89.8**$_{\uparrow 14\%}$	**105.4**	102.5$_{\downarrow 3\%}$	106.2	**109.8**$_{\uparrow 3\%}$	88.1	**94.5**$_{\uparrow 7\%}$	**77.0**	75.9$_{\downarrow 1\%}$
Pen-cloned	5.6	**82.4**$_{\uparrow 1371\%}$	46.5	**82.2**$_{\uparrow 77\%}$	**98.5**	92.4$_{\downarrow 6\%}$	96.5	**100.2**$_{\uparrow 4\%}$	**107.1**	105.8$_{\downarrow 1\%}$	73.6	**87.4**$_{\uparrow 19\%}$
Door-human	−0.3	**0.2**$_{\uparrow \infty}$	3.5	**4.8**$_{\uparrow 37\%}$	−0.1	**0.1**$_{\uparrow \infty}$	0.0	**0.2**$_{\uparrow \infty}$	0.1	**0.3**$_{\uparrow 200\%}$	−0.1	**0.1**$_{\uparrow \infty}$
Door-cloned	−0.3	**0.1**$_{\uparrow \infty}$	3.3	**4.5**$_{\uparrow 36\%}$	0.1	**0.5**$_{\uparrow 400\%}$	0.1	**0.4**$_{\uparrow 300\%}$	0.1	**0.3**$_{\uparrow 200\%}$	0.1	**0.5**$_{\uparrow 400\%}$
Hammer-human	**1.1**	0.5$_{\downarrow 55\%}$	1.9	**3.1**$_{\uparrow 63\%}$	0.3	**0.8**$_{\uparrow 167\%}$	0.3	**0.9**$_{\uparrow 200\%}$	0.3	**0.8**$_{\uparrow 167\%}$	0.3	**0.8**$_{\uparrow 167\%}$
Hammer-cloned	0.3	**0.4**$_{\uparrow 33\%}$	1.7	**3.2**$_{\uparrow 88\%}$	5.4	**7.5**$_{\uparrow 39\%}$	5.9	**7.8**$_{\uparrow 32\%}$	4.6	**7.0**$_{\uparrow 52\%}$	5.1	**6.8**$_{\uparrow 33\%}$
Relocate-human	−0.3	**0.2**$_{\uparrow \infty}$	0.1	**0.5**$_{\uparrow 400\%}$	0.2	**0.6**$_{\uparrow 200\%}$	0.2	**0.7**$_{\uparrow 250\%}$	0.1	**0.7**$_{\uparrow 600\%}$	0.2	**0.6**$_{\uparrow 200\%}$
Relocate-cloned	**0.1**	0.0$_{\downarrow 100\%}$	0.0	0.0	0.0	**0.4**$_{\uparrow \infty}$	**0.2**	0.1$_{\downarrow 50\%}$	**0.1**	0.0$_{\downarrow 100\%}$	−0.2	**0.6**$_{\uparrow \infty}$
Adroit Avg.	0.3	**20.9**$_{\uparrow 6866\%}$	17.0	**23.5**$_{\uparrow 38\%}$	**26.2**	25.6$_{\downarrow 2\%}$	26.2	**27.5**$_{\uparrow 5\%}$	25.1	**26.2**$_{\uparrow 4\%}$	19.5	**21.6**$_{\uparrow 11\%}$