Less Is More: Clustered Cross-Covariance Control for Offline RL

Nan Qiao, Sheng Yue, Shuning Wang, Yongheng Deng, Ju Ren

Paper GitHub Homepage

ABSTRACT

A fundamental challenge in offline reinforcement learning is distributional shift. Scarce data or datasets dominated by out-of-distribution (OOD) areas exacerbate this issue. Our theoretical analysis and experiments show that the standard squared error objective induces a harmful TD cross covariance. This effect amplifies in OOD areas, biasing optimization and degrading policy learning. To counteract this mechanism, we develop two complementary strategies: partitioned buffer sampling that restricts updates to localized replay partitions, attenuates irregular covariance effects, and aligns update directions, yielding a scheme that is easy to integrate with existing implementations, namely ***Clustered Cross-Covariance Control for TD (C4).*** We also introduce an explicit gradient-based corrective penalty that cancels the covariance induced bias within each update. We prove that buffer partitioning preserves the lower bound property of the maximization objective, and that these constraints mitigate excessive conservatism in ***extreme OOD areas*** without altering the core behavior of policy constrained offline reinforcement learning. Empirically, our method showcases higher stability and up to ***30% improvement*** in returns over prior methods, especially with small datasets and splits that emphasize OOD areas.

MOTIVATION

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

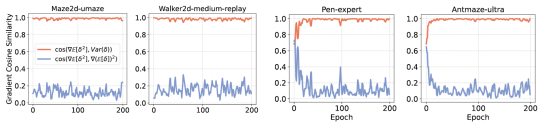
$$\delta \equiv r(\mathbf{s},\mathbf{a}) + \gamma\, \mathbb{E}_{\mathbf{a}'\sim\pi(\cdot|\mathbf{s}')} Q_{\phi'}(\mathbf{s}',\mathbf{a}') - Q_{\phi}(\mathbf{s},\mathbf{a})$$

Observation 1: TD Variance Dominates

The second-moment identity

$$\mathbb{E}[\delta^2] \;=\; \big(\mathbb{E}[\delta]\big)^2 \;+\; \mathrm{Var}[\delta],$$

shows that TD training is governed by both the mean and the variance of the residual. In difficult offline RL settings, ***the variance term is the dominant driver of optimization.***



The gradient of $\mathbb{E}[\delta^2]$ aligns much more closely with $\nabla Var[\delta]$ than with $\nabla(\mathbb{E}[\delta])^2$. This makes TD variance the right lens for analysis and control.

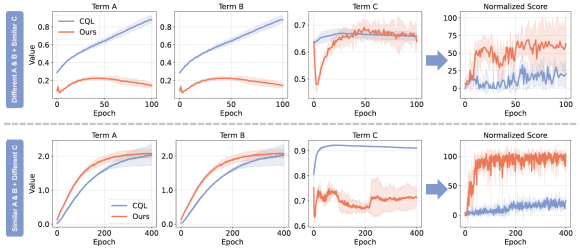
Observation 2: TD Cross Covariance Is Harmful

The variance decomposition separates helpful and **harmful** effects:

Theorem 1. *All expectations, variances, and covariances below are taken over $k, k', \mathbf{w}, \mathbf{w}'$. With the first order approximation for Q_ϕ in feature space, the variance satisfies*

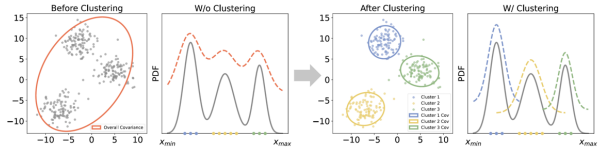
$$\mathrm{Var}[\delta] \approx \underbrace{\gamma^2 (k')^2 \,\mathrm{Var}\big(\langle \mathbf{w}', \nabla_{x'} Q_{\phi'}(x')\rangle\big) + k^2\,\mathrm{Var}\big(\langle \mathbf{w}, \nabla_x Q_\phi(x)\rangle\big)}_{\text{implicit regularizer in noisy supervised learning, denote as Term(A) and Term (B)}} \underbrace{- 2\gamma k k' \,\mathrm{Cov}\big(\langle \mathbf{w}', \nabla_{x'} Q_{\phi'}(x')\rangle, \langle \mathbf{w}, \nabla_x Q_\phi(x)\rangle\big)}_{\text{additional cross term unique to TD learning, denote as Term(C)}}. \quad (6)$$

Terms A and B behave like beneficial implicit regularizers. The TD-specific cross covariance C is the component that destabilizes updates in OOD regions.



Performance improves when A and B remain strong while C is suppressed. This pinpoints the exact quantity that C4 should control

PRACTICAL ALGORITHM



The key identity is minimize the harmful Term C:

Theorem 2. *With cluster label Z, the cross covariance decomposes as*

$$C = \mathbb{E}\big[\,C_Z\,\big] \;+\; \mathrm{Cov}(\mu'_Z, \mu_Z). \quad (9)$$

If each minibatch is drawn from a single cluster z, the between-cluster term in Eq. (9) vanishes in that batch and, for any unit $\mathbf{w}', \mathbf{w}$,

$$\big|-2\gamma k k' \mathrm{Cov}\big(\langle \mathbf{w}', g'\rangle, \langle \mathbf{w}, g\rangle\big)\big| \;\le\; 2\gamma k k' \,\|C_z\|_2 \;\le\; 2\gamma k k' \,\sqrt{\mathrm{tr}\,\Sigma'_z}\,\sqrt{\mathrm{tr}\,\Sigma_z}. \quad (10)$$

so a single-cluster minibatch removes the between-cluster driver and leaves only the local covariance C_z. Thus, updating in offline RL is to minimize

$$\mathcal{J}(\phi, B) = \mathcal{L}_{\mathrm{TD}}(\phi, B) + \lambda\Big(\big\|\bar{C}_z(B)\big\|_F^2 + \beta(\mathrm{tr}\,\bar{C}_z(B))^2\Big).$$

Algorithm 1 C^4: Single-Cluster Offline Update for TD (EM-style)

Input offline dataset $\mathcal{D}$, number of clusters K, regularizers λ, β, iterations T.
Initialize mixture $\{p_z, \mu_z, \Omega_z\}_{z=1}^K$, critic ϕ.
for $t = 1, \ldots, T$ **do**
Compute gradients: for each sample i, form g', g and stack $y_i = [g', g]$.
E–step: $r_{iz} \propto p_z \mathcal{N}(y_i \mid \mu_z, \Omega_z)$, normalize $\sum_z r_{iz} = 1$. ①
M–step: $p_z \leftarrow \frac{1}{n}\sum_i r_{iz}$, $\mu_z \leftarrow \frac{1}{N_z}\sum_i r_{iz} y_i$, $\Omega_z \leftarrow \frac{1}{N_z}\sum_i r_{iz}(y_i-\mu_z)(y_i-\mu_z)^\top + \epsilon I$, extract C_z.
Critic minibatch: sample cluster $z \sim \mathrm{Cat}(\{p_z\})$, draw minibatch B with weights r_{iz}. ②
Penalty and update: compute $\hat{C}_z(B) = \mathrm{Cov}_B(g', g)$, minimize batch loss $\mathcal{J}(\phi, B)$ in Eq. (14). ③
end for
Output trained critic ϕ and mixture $\{p_z, \mu_z, \Omega_z\}$.

1. compute stacked gradient pairs from replay samples, update a Gaussian mixture over the stacked-gradient space
2. sample one cluster, then draw a single-cluster minibatch
3. apply the covariance-regularized TD update on that minibatch

EXPERIMENTAL RESULTS

Restricting our evaluation to a scarce-data regime of at most **10k** state-action pairs (approximately ***1% of the full dataset***), We primarily report the normalized scores for the D4RL MuJoCo locomotion tasks and summarize the main results across AntMaze, Maze2D, and Adroit.

Task	TD3+BC	CQL	IQL	DOGE	BPPO	TSRL	A2PR	Ours
Ant-me	52.0±18.2	74.0±25.0	66.0±10.5	82.0±16.4	85.5±13.7	83.6±12.4	66.7±10.2	**100.9±5.0**
Ant-m	46.0±17.4	62.0±22.1	56.0±9.3	69.0±15.2	78.0±11.9	72.2±10.6	64.0±9.3	**84.5±6.1**
Ant-mr	31.0±15.5	36.0±16.3	41.0±10.8	46.0±12.9	52.0±10.8	49.4±13.1	44.0±9.1	**65.8±6.9**
Ant-e	72.0±25.0	94.0±29.4	82.0±15.0	98.0±20.3	103.0±11.2	100.7±12.6	88.0±10.6	**109.6±2.7**
Ant-fr	70.0±24.0	92.0±26.5	80.0±15.0	96.0±20.0	102.0±11.7	99.8±12.0	86.0±10.4	**107.6±3.2**
Hopper-m	30.7±13.2	50.1±22.3	61.0±6.2	55.6±8.3	55.0±7.8	60.9±4.1	55.9±8.4	**69.2±12.7**
Hopper-mr	11.3±4.7	13.2±2.0	16.2±3.0	19.1±3.3	45.1±8.7	23.5±8.8	12.5±5.9	**45.9±8.4**
Hopper-me	22.6±13.9	43.2±6.9	51.7±7.0	36.8±34.5	27.9±15.2	56.6±13.9	49.7±10.7	**81.3±6.0**
Hopper-e	53.6±17.1	56.1±26.4	60.9±9.6	62.2±21.7	85.0±17.9	76.7±20.4	80.0±16.8	**107.0±2.8**
Hopper-fr	32.0±13.5	45.0±22.0	56.0±6.3	54.0±8.4	60.0±9.7	53.4±11.3	55.0±8.9	**65.3±9.4**
Walker2d-m	11.2±19.2	54.1±15.5	34.2±5.2	53.7±12.6	54.7±11.4	47.3±10.1	5.9±5.2	**65.9±7.8**
Walker2d-mr	9.3±6.6	13.8±5.3	17.7±8.9	15.5±9.2	29.5±8.7	27.6±12.4	34.4±8.9	**55.4±5.9**
Walker2d-me	12.4±15.7	26.0±14.0	38.0±12.2	42.5±11.4	61.3±12.2	50.9±26.4	56.5±11.5	**96.3±10.4**
Walker2d-e	29.5±23.5	56.0±29.4	16.2±3.2	81.2±18.6	102.0±9.7	104.9±10.6	98.0±7.9	**109.5±0.3**
Walker2d-fr	14.2±19.5	55.0±16.0	36.0±5.6	55.5±12.3	52.0±10.9	44.3±10.4	48.0±9.6	**77.3±7.1**
Halfcheetah-m	25.9±8.4	41.7±2.2	35.6±2.9	42.8±2.9	28.5±3.2	43.3±2.8	37.1±2.7	**46.3±3.1**
Halfcheetah-mr	29.1±8.3	16.3±4.9	34.1±6.3	26.3±3.1	34.4±4.2	27.7±3.8	23.6±4.7	**43.1±5.3**
Halfcheetah-me	23.5±13.6	39.7±6.4	14.3±7.3	33.1±8.8	22.3±9.6	37.2±14.9	32.4±8.3	**46.0±3.5**
Halfcheetah-e	26.4±4.2	5.8±1.3	-1.1±3.8	1.4±3.1	6.5±3.4	42.0±26.4	36.0±7.8	**75.8±5.2**
Halfcheetah-fr	28.0±8.6	45.0±2.4	33.0±3.0	43.0±3.1	37.0±3.6	41.0±3.0	39.0±4.1	**58.1±3.4**
Locomotion-Avg.	31.5	46.0	41.4	50.7	56.1	57.2	50.6	**75.7**
AntMaze-Avg.	6.3	10.7	21.4	20.5	16.5	22.0	16.2	**27.0**
Maze2D-Avg.	49.7	57.4	103.6	106.1	120.6	115.7	111.4	**126.9**
Adroit-Avg.	1.4	7.3	15.2	8.4	**23.1**	15.7	-0.1	21.6