AUDITING ACTION-GRADIENT GEOMETRY IN ENSEMBLE COST PENALTIES

Anonymous authors
Paper under double-blind review

ABSTRACT

A common ensemble cost penalty in safe reinforcement learning adds prediction spread to the mean cost. A gradient-based actor follows the action derivative of this penalized estimate. We introduce an action-gradient audit to distinguish changes in direction from changes in signed scale and magnitude. The audit evaluates recorded behavior actions reserved from critic fitting. On matched data from three MuJoCo systems, bootstrap ensembles produce greater rotation than shared-data ensembles. The median increase exceeds 6.5° in HalfCheetah, while median parallel scales and norm ratios remain near one. Across independently seeded fits to the same data, mean gradients agree more closely than penalized gradients in every studied group. The penalized direction also has a nonpositive median advantage over the mean direction in 71 of 72 finite-step comparisons on independently fitted evaluators. These findings show that a penalty can substantially rotate the local cost gradient without yielding a direction that agrees better across fits. Evaluating both scalar estimates and their derivatives makes this distinction visible.

1 INTRODUCTION

Learning under safety constraints requires estimates of the costs associated with candidate actions. Recent methods account for uncertainty in these estimates through conservative objectives and online safety filters (Wu et al., 2024; Capone et al., 2025). For ensemble cost critics, a common construction adds prediction spread to the mean cost, $U(s, a) = \mu(s, a) + k\sigma(s, a)$, where k sets the penalty weight (Wu et al., 2024). This raises the estimated cost at uncertain actions, but a gradient-based actor consumes the derivative with respect to the action. The resulting question is whether the penalty changes the direction for reducing cost or primarily changes the strength of the mean-cost direction.

The answer depends on how disagreement changes with the action. A spread gradient parallel to the mean gradient changes its signed scale, whereas a perpendicular component turns the direction. Scalar mean and spread do not reveal these components. We give a fixed-action construction in which the member values and mean gradient stay unchanged while the spread gradient is arbitrary (Section 3.1). The derivative of the aggregate cost is therefore needed to identify the local direction supplied by the penalty.

We compare ensembles at the same recorded states and actions, reserved from critic fitting. These behavior actions provide common inputs for measuring gradient geometry. The ensembles use shared data, bootstrap samples, or bootstrap samples with randomized priors. We examine velocity and action-energy costs in Hopper, Walker2d and HalfCheetah, then vary the penalty weight and selected members.

To assess stability across fits, we train a second ensemble with independent seeds on the same data and targets. This evaluator measures agreement between corresponding gradient directions. It also evaluates small action steps along the primary mean and penalized directions.

Bootstrap sampling changes the direction most strongly in HalfCheetah. At $k = 0.5$, its paired increase over shared-data ensembles exceeds 6.5°, compared with about 1°–3° in Hopper and Walker2d. Median norm ratios remain close to one. Across independently seeded fits, mean gradients agree more closely than penalized gradients in all 18 environment–cost–construction summaries. The penalized direction has a positive median finite-step advantage in only one of 72 descriptive

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.

comparisons. Substantial rotation within one fit can therefore coexist with weaker agreement across fits.

Our contribution is a diagnostic that connects ensemble construction to the cost derivative it produces (Figure 1). The geometric decomposition separates rotation from signed scaling. Matched comparisons then reveal how construction choices change these quantities on common behavior data. Independent fits assess agreement and finite-step transfer of the resulting directions. Together, these measurements make the local effect of an uncertainty penalty inspectable before its derivative enters a policy update.

2 Related Work

Mean-plus-spread objectives. Ensembles represent predictive variation through independent training, bootstrap sampling and fixed randomized priors (Lakshminarayanan et al., 2017; Osband et al., 2016; 2018). Reinforcement learning uses this variation for weighted Bellman updates and action selection (Lee et al., 2021), pessimistic value estimation (Bai et al., 2022), and model-rollout penalties (Yu et al., 2020). Conservative Augmented Lagrangian (CAL) places mean plus spread directly in a cost objective to address underestimated cost in off-policy primal–dual learning (Wu et al., 2024). These uses establish why uncertainty enters optimization. Our question concerns the action derivative that results when the aggregate cost is differentiated.

Gradient diversity. Ensemble-Diversified Actor Critic (EDAC) encourages diversity among critic action gradients to improve uncertainty-based offline reinforcement learning (An et al., 2021). Related studies examine diversity in input-gradient or projection space (Trinh et al., 2024; Zanger et al., 2024). Diversity among members matters because they provide the derivatives from which an ensemble objective is built. The direction after aggregation also depends on their predicted values: the spread derivative weights member-gradient deviations by value deviations. We inspect this combined direction relative to the mean gradient, complementing measurements of pairwise diversity.

Constrained actor geometry. Actor–critic methods differentiate value functions through policy actions (Fujimoto et al., 2018; Haarnoja et al., 2018). Constrained algorithms combine reward and cost through trust-region, Lagrangian or feedback-control updates (Achiam et al., 2017; Tessler et al., 2019; Stooke et al., 2020). A cost action gradient is therefore an input to a larger update, whose result also depends on the reward and policy parameterization. Isolating that input at common actions lets us compare ensemble constructions without attributing the effects of a whole training algorithm to one derivative.

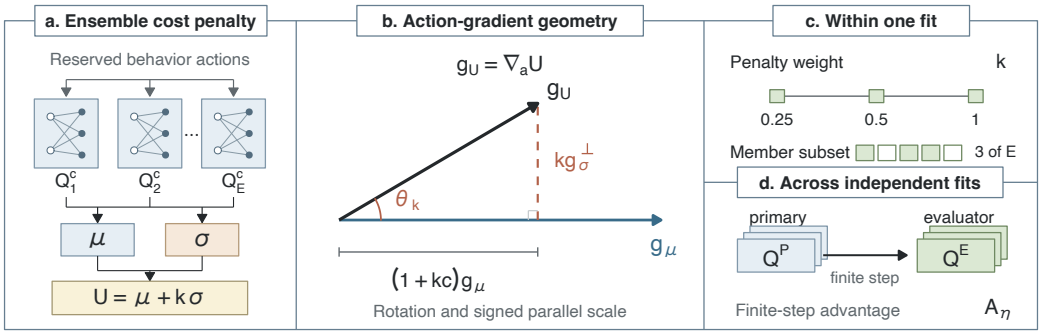


Figure 1: Action-gradient audit and evidence path. Ensemble critics define $U = \mu + k\sigma$. At reserved behavior actions, the spread gradient changes signed parallel scale $1 + kc$ and rotates the direction by θ_k. Penalty weight and member subsets test sensitivity within a fit; an evaluator fitted from independent seeds tests finite-step transfer across fits.

3 ACTION-GRADIENT AUDIT

Adding ensemble spread can change both the strength and direction of the local descent signal. At fixed states and actions, we separate these effects by comparing penalized and mean gradients. An independently fitted evaluator tests transfer through small action steps (Figure 1).

3.1 MEAN-PLUS-SPREAD DERIVATIVE

Prediction spread measures disagreement at the current action. Its effect on descent depends on how that disagreement changes nearby. Holding the state fixed, we therefore differentiate the mean and spread with respect to the action. For differentiable cost critics $Q_i^c(s,a)$, $i = 1, \dots, E$, define

$$\mu = \frac{1}{E}\sum_{i=1}^{E} Q_i^c, \qquad \sigma = \left[\frac{1}{E}\sum_{i=1}^{E}(Q_i^c - \mu)^2\right]^{1/2}, \qquad U = \mu + k\sigma. \tag{1}$$

Write $g_\mu = \nabla_a \mu$, $g_\sigma = \nabla_a \sigma$, and $g_U = g_\mu + k g_\sigma$. Spread grows along directions in which predictions above the mean move further above it and those below move further below. Its derivative thus couples value deviations with gradient deviations. For $\sigma > 0$,

$$g_\sigma = \frac{1}{E\sigma}\sum_{i=1}^{E}(Q_i^c - \mu)(\nabla_a Q_i^c - g_\mu). \tag{2}$$

At exact zero spread, the implementation sets $g_\sigma = 0$ as a computational convention. The standard deviation is generally nonsmooth there, and Eq. (2) applies only when $\sigma > 0$.

Underdetermination at a fixed action. Two ensembles can agree on every predicted value at an action while their predictions vary differently nearby. A two-critic construction makes this freedom explicit. At a chosen action a_0, fix the mean value μ_0, spread $\sigma_0 > 0$ and mean gradient h, while choosing any vector v. Define

$$Q_1^c(a) = \mu_0 + \sigma_0 + (h+v)^\top (a - a_0), \qquad Q_2^c(a) = \mu_0 - \sigma_0 + (h-v)^\top (a - a_0). \tag{3}$$

At a_0, the opposite offsets give $\mu = \mu_0$, $\sigma = \sigma_0$ and $g_\mu = h$, while $g_\sigma = v$. Varying v preserves both member values and the mean gradient but changes the spread gradient arbitrarily. For $k > 0$, the penalized gradient $g_U = h + kv$ therefore depends on slopes that the scalar statistics leave unspecified.

3.2 ROTATION AND SIGNED SCALE

The spread gradient can reinforce or oppose the mean gradient, and it can also push sideways. We separate these effects by resolving it into parallel and perpendicular components. For nonzero g_μ, write

$$g_\sigma = c g_\mu + g_\sigma^\perp, \qquad c = \frac{\langle g_\sigma, g_\mu \rangle}{\|g_\mu\|^2}, \qquad \langle g_\sigma^\perp, g_\mu \rangle = 0. \tag{4}$$

Then $g_U = (1 + kc) g_\mu + k g_\sigma^\perp$. The coefficient $1 + kc$ changes signed parallel scale, while $g_\sigma^\perp$ turns the direction away from the mean. A negative parallel coefficient can turn it beyond $90°$, so the angle must preserve orientation. For $k \geq 0$ and nonzero g_U,

$$\theta_k = \operatorname{atan2}\big(k\|g_\sigma^\perp\|, (1+kc)\|g_\mu\|\big), \tag{5}$$

$$\cos\theta_k = \frac{(1+kc)\|g_\mu\|}{\sqrt{(1+kc)^2\|g_\mu\|^2 + k^2\|g_\sigma^\perp\|^2}}. \tag{6}$$

Rotation and length can change separately. To measure both, let $s_k = 1 + kc$ be signed parallel scale, $o_k = k\|g_\sigma^\perp\|/\|g_\mu\|$ the normalized perpendicular component, and $\rho_k = \|g_U\|/\|g_\mu\|$ the norm ratio. Orthogonality gives

$$\rho_k^2 = s_k^2 + o_k^2, \qquad \cos\theta_k = \frac{s_k}{\rho_k}. \tag{7}$$

Positive s_k with $o_k = 0$ changes only scale. A nonzero o_k changes direction, even if the norm ratio remains near one. Parallel cancellation occurs when s_k is near zero; negative s_k places the penalized gradient beyond $90°$.

For fixed gradients, the perpendicular component also determines how the angle responds to the penalty weight. When $r = \|g_\sigma^\perp\|/\|g_\mu\| > 0$, $d\theta_k/dk = r/[(1+kc)^2 + k^2 r^2] > 0$. Increasing the weight then increases the oriented angle; the size of that response depends on the measured components.

Changing common cost units preserves the geometry of fitted critics. For $Q_i^{c\prime} = \alpha Q_i^c + \beta$, with constants $\alpha > 0$ and β, both gradients are multiplied by α. Thus c, s_k, o_k, ρ_k, θ_k and the normalized descent directions stay the same. Retraining on rescaled targets can change the fitted functions.

The choice of action coordinates matters separately. All angles use the Euclidean metric in native action coordinates, so anisotropic coordinate rescaling can change them.

3.3 Finite-step transfer across independent fits

A rotated direction may depend on the particular fitted critics. We assess its stability using an evaluator $\widetilde{U}$ fitted with independent seeds to the same training data and targets. Let P and E denote the primary and evaluator fits. At each reserved behavior action, $C_\mu = \cos(g_\mu^P, g_\mu^E)$ measures mean-gradient agreement, and $C_U = \cos(g_U^P, g_U^E)$ measures penalized-gradient agreement. Larger cosine means closer agreement.

The finite-step comparison asks which primary direction reduces the evaluator objective more. From the same action, we take normalized descent steps using $d_\mu = g_\mu/\|g_\mu\|$ or $d_U = g_U/\|g_U\|$. Clipping to native action bounds can change the distance traveled, so we divide objective decrease by realized displacement. For nonzero displacement,

$$a_\eta(d) = \mathrm{clip}(a - \eta d), \qquad \Delta_\eta(d) = \frac{\widetilde{U}(s,a) - \widetilde{U}(s, a_\eta(d))}{\|a - a_\eta(d)\|}. \tag{8}$$

The paired advantage $A_\eta = \Delta_\eta(d_U) - \Delta_\eta(d_\mu)$ is positive when the penalized direction produces the larger gain. For each seed, we take the median advantage over 512 behavior actions. We then report the median and range of these five seed-level values. Seeds serve as the replication units.

Local interpretation of transfer. For small steps, transfer depends on alignment with the evaluator gradient. Consider an interior action where $\widetilde{U}$ is differentiable. Sufficiently small steps are unclipped, and a first-order expansion gives

$$\lim_{\eta \downarrow 0} A_\eta = \langle g_U^E, d_U - d_\mu \rangle, \qquad g_U^E = \nabla_a \widetilde{U}(s,a). \tag{9}$$

When $g_U^E \neq 0$, the limiting advantage is positive precisely when d_U has the larger cosine with that gradient. Both directions are scored against the same penalized evaluator objective. The agreement statistics use different references: C_μ uses the evaluator mean gradient, and C_U uses its penalized gradient. Their ordering alone therefore does not determine the sign of A_η.

Refitting matters because the normalized penalized gradient is already the steepest-descent direction of its own objective. Under the same local conditions, using the primary fit as evaluator gives

$$\lim_{\eta \downarrow 0} A_\eta\big|_{\widetilde{U}=U} = \|g_U\|\big(1 - \cos\theta_k\big) \geq 0. \tag{10}$$

An independently fitted evaluator tests whether this first-order advantage survives a change in fitted functions.

Finite-radius effects. Curvature and clipping can change finite-step transfer. We examine several radii and repeat the analysis on unclipped actions. For unclipped steps, assume the evaluator's Hessian operator norm is at most L along both step segments. Taylor's theorem gives

$$\big|A_\eta - \langle g_U^E, d_U - d_\mu \rangle\big| \leq L\eta. \tag{11}$$

Each normalized gain contributes at most $L\eta/2$ to the remainder, so a limiting advantage larger in magnitude than $L\eta$ retains its sign. This bound interprets the step-size sweep; the measured comparisons evaluate objectives directly without estimating L.

3.4 Relation to actor updates

The measurements describe $g_\mu(s, a_b)$ and $g_U(s, a_b)$ at reserved behavior actions a_b. To relate this to policy optimization, consider the minimization-form actor loss $L_{\text{actor}} = L^r(s, \pi_\theta(s)) + \lambda U(s, \pi_\theta(s))$. At the policy action, the actor combines the reward and cost derivatives. The policy Jacobian maps this signal to parameter space, and the update aggregates over states. Our audit isolates the cost derivative at a_b. The finite-step outcome measures change in the fitted evaluator objective at those states.

4 Matched Experimental Design

We compare ensembles at identical recorded states and actions so that changes in sampled behavior do not confound the construction comparison. Training and evaluation use disjoint complete episodes. Within each environment and cost, all constructions share each seed's buffer and targets.

Ensemble comparisons. The shared-data construction trains independently initialized members on the same fitting pairs. Bootstrap gives each member a separate transition resample while matching its initialization seed to the shared-data member. Randomized-prior critics reuse the bootstrap resamples and trainable-network seeds, adding fixed prior networks at scale 0.35 (Osband et al., 2018). This pairing lets us compare the effect of resampling and then the added prior while holding the underlying behavior data fixed.

Systems and behavior data. We use Hopper-v5, Walker2d-v5, and HalfCheetah-v5 from Gymnasium with MuJoCo dynamics (Towers et al., 2025; Todorov et al., 2012). Each of five evaluation seeds supplies one behavior buffer with 8,000 transitions per environment. At every step, the action is independently uniform with probability 0.5; otherwise it follows an AR(0.8) process with uniform innovations, clipped to the native bounds. Complete training episodes provide 7,488 fitting pairs. The remaining 512 behavior actions are balanced across at least five disjoint evaluation episodes and never enter critic fitting.

The episode split is fixed before critic fitting. We compute cost-to-go targets through the end of each training episode, then retain the first 7,488 training pairs. Evaluation actions are evenly spaced within separate episodes. Their allocation is balanced across episodes up to the available lengths. Completing the episodes can require collecting more than the 8,000 retained transitions. Evaluation seeds are 211, 307, 401, 503 and 607; the separate calibration buffer uses seed 101. All constructions reuse these selections.

Costs and targets. The costs penalize excessive speed or action effort. The velocity cost is $c_t^v = \max\{0, |v_t| - v_{\max}\}$. We fix $v_{\max}$ at the 75th percentile of $|v_t|$ in the training portion of the independent calibration buffer. The thresholds are 0.454 for Hopper, 1.113 for Walker2d and 0.646 for HalfCheetah. The action-energy cost is $c_t^e = \max\{0, \|a_t\|^2/d_a - 0.25\}$, where d_a is action dimension. Monte Carlo cost-to-go targets sum discounted costs within training episodes using $\gamma = 0.97$.

Fitting and evaluation. Each ensemble has five SiLU critics with two hidden layers of width 128. We minimize squared error on standardized targets for 1,200 Adam steps, using batch size 256 and learning rate 3×10^{-4}. Both costs use the same behavior buffer within each environment and evaluation seed.

The primary analysis uses all five members ($E = 5$) and $k = 0.5$. On each fitted ensemble we also evaluate $k \in \{0.25, 0.5, 1\}$ and all ten three-member subsets. The geometry study contains 90 environment, cost, construction and seed combinations, with 46,080 action evaluations. Each combination receives a five-member evaluator trained with seed offset $1{,}000{,}003$, separating initialization, bootstrap draws and prior seeds from the primary fit. Finite-step sizes are $\eta \in \{0.01, 0.025, 0.05, 0.1\}$.

The numerical mask guards against unstable angles near zero gradients. For each gradient norm, we take the median over its positive values within the setting. Both $\|g_\mu\|$ and $\|g_U\|$ must exceed 10^{-6} times their respective medians. Sensitivity checks use thresholds 0, 10^{-4} and 10^{-2}.

5 RESULTS

5.1 CONSTRUCTION EFFECTS ON ROTATION

Bootstrap sampling increases rotation most in HalfCheetah (Figure 2). We measure this effect by subtracting the shared-data median angle from the bootstrap median within each matched seed, then taking the median of those differences. At $k = 0.5$, the increases are $6.53°$ for velocity cost and $6.73°$ for energy cost, compared with roughly $1°$–$3°$ in Hopper and Walker2d (Table 1). Both HalfCheetah differences are positive in all five seeds. Adding a randomized prior changes the paired bootstrap median by at most $0.11°$ in magnitude across the six environment–cost groups.

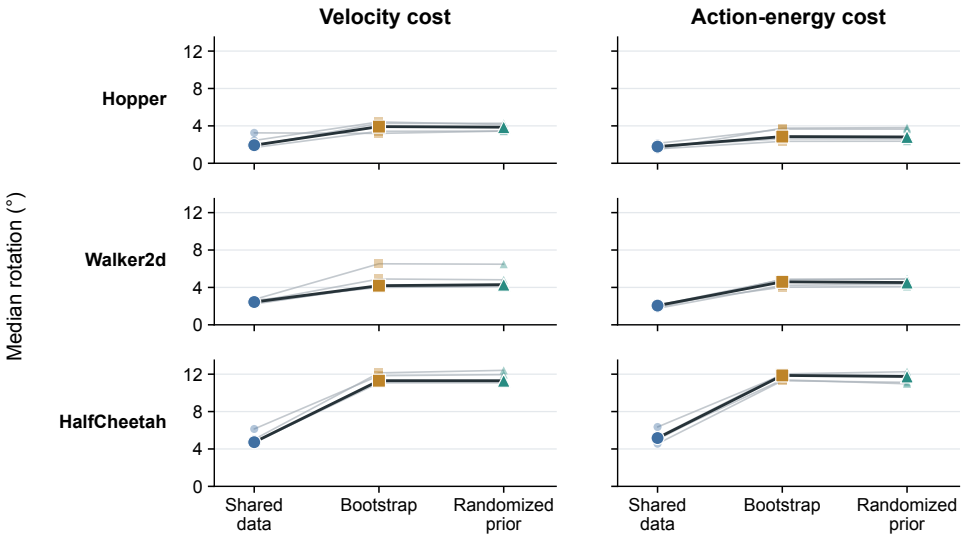


Figure 2: Matched construction trajectories at $E = 5$ and $k = 0.5$. Pale lines connect constructions evaluated on the same behavior buffer for one seed; dark lines show medians over five seeds. Bootstrap produces the largest paired rotation in HalfCheetah, while the fixed randomized prior adds little to the bootstrap angle under this configuration.

Table 1: Rotation, scale and member sensitivity at $k = 0.5$. Bootstrap produces the largest rotation increase in HalfCheetah, while median parallel scales remain close to one.

		Five-member geometry			Member selection
Cost	Environment	Bootstrap increase (°)	Signed parallel scale	Angles $> 10°$ (%)	Three-member subset span (°)
Velocity	Hopper	+1.99	1.020	7.2	3.87–5.63
	Walker2d	+1.96	1.014	3.9	4.80–5.73
	HalfCheetah	+6.53	1.018	61.5	12.06–14.60
Energy	Hopper	+1.07	1.025	3.7	2.83–4.23
	Walker2d	+2.55	1.005	3.5	4.67–6.24
	HalfCheetah	+6.73	1.005	63.3	11.63–15.36

Entries are medians across five seeds. Each seed supplies a paired angle difference, median scale and fraction above $10°$. Differences compare bootstrap with shared data; other columns describe bootstrap. For each seed, the subset span runs from the smallest to largest median angle over ten three-member subsets. We aggregate the lower and upper endpoints separately.

This directional change occurs alongside modest changes in gradient size. Across the six bootstrap groups, median signed parallel scales are 1.005–1.025 and median norm ratios are 1.009–1.043.

For each seed, we measure the fraction of actions whose rotation exceeds $10°$. Median fractions across seeds are 61.5% and 63.3% for HalfCheetah velocity and energy, versus at most 7.2% elsewhere (Table 1). Only one of 15,360 bootstrap action evaluations has negative parallel scale. The observed rotation thus mainly occurs with a positive component along the mean gradient.

5.2 Penalty weight and subset sensitivity

Increasing the penalty weight produces the largest angular response in HalfCheetah (Figure 3). Holding the critics fixed isolates the effect of k. The geometry in Section 3 explains the monotonic increase; the fitted gradients determine its magnitude. At $k = 1$, bootstrap median angles reach about $22°$ for both HalfCheetah costs, compared with less than $10°$ in Hopper and Walker2d. Shared-data ensembles remain lower in every matched group.

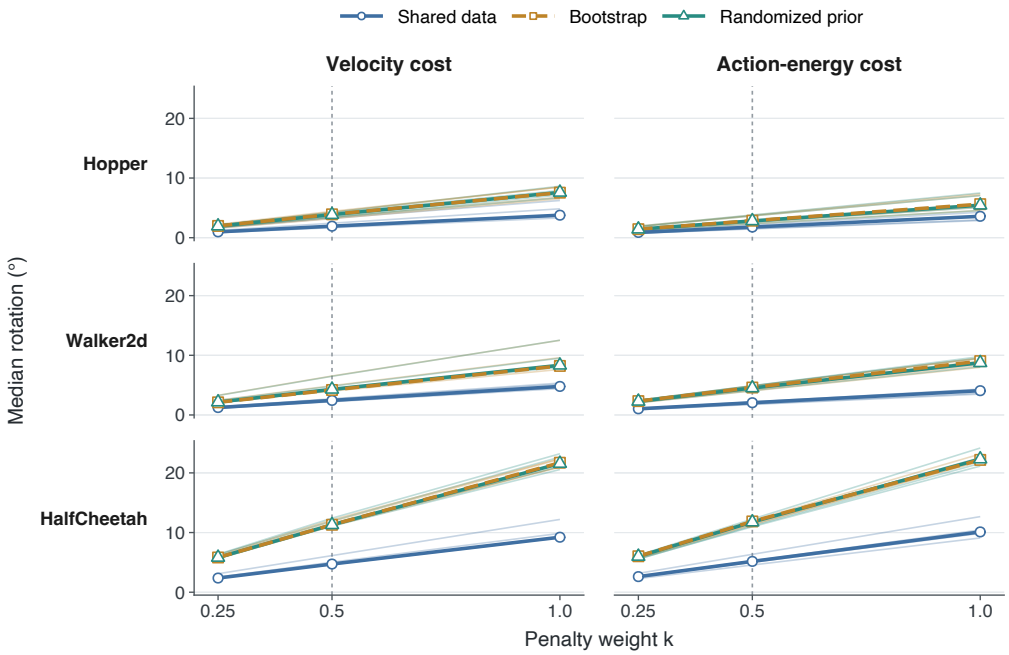


Figure 3: Sensitivity to the penalty weight in matched ensembles with five members. Pale lines connect weights evaluated on the same fitted ensemble; dark lines show medians over five seeds. The vertical rule marks the primary $k = 0.5$. Increasing k produces the largest angular change in HalfCheetah bootstrap ensembles and ensembles with randomized priors.

Selecting members creates additional variation within a fitted ensemble. For each seed, we evaluate all ten three-member subsets and record the smallest and largest median angles. We take the median of each endpoint across seeds to obtain the spans in Table 1. HalfCheetah retains larger angles and has the widest spans for both costs. Its stronger rotation persists under member selection, while the precise angle depends on which members are retained.

5.3 Finite-step transfer across independent fits

The mean gradient agrees more closely across fits than the penalized gradient throughout the study. We summarize agreement for each combination of three environments, two costs and three constructions, giving 18 groups with five seeds each. All 18 group summaries favor the mean gradient, as do 88 of the 90 individual seed settings. Table 2 reports the bootstrap groups, where the largest median gaps occur in HalfCheetah. The stronger rotation seen in that system therefore coexists with weaker agreement of the penalized direction across independently seeded fits.

We then compare the objective decrease produced by the two primary directions on the same evaluator. Evaluating four step sizes in each of the 18 groups gives 72 comparisons. The penalized direction has

nonpositive median paired advantage A_η in 71 of them (Figure 4). These are descriptive comparisons: each group reuses the same fits across step sizes, with five seeds as the replication units. The sole positive median occurs for Hopper velocity with bootstrap at $\eta = 0.05$; its advantage is 6.3×10^{-7} and its seed range crosses zero.

At individual actions, either direction can perform better. For each seed, we compute the fraction of actions favoring the penalized direction. The medians of these fractions across seeds range from 37.9% to 50.0%. The group-level result therefore summarizes a mixture of local outcomes. Table 2 reports the bootstrap comparisons at $\eta = 0.05$.

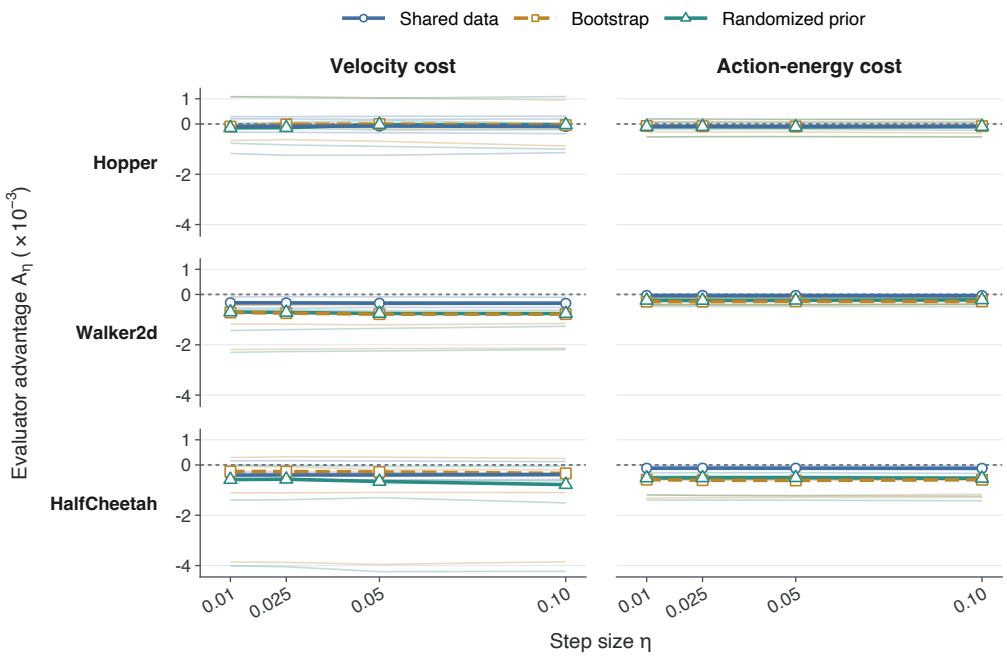


Figure 4: Transfer across independent fits at $k = 0.5$. Evaluators use the same fitting data and targets, with independent initialization, resampling and prior seeds. Gains measure evaluator objective decrease per unit realized displacement. The vertical axis shows penalized-direction gain minus mean-direction gain; positive values favor the penalized direction. Each pale trajectory reports one seed's median over 512 actions. Dark trajectories show medians across five seeds.

Clipping affects at most 5.08% of actions in a comparison. Equation (8) uses realized displacement, and we also repeat the summary after excluding every action clipped under either direction. At least 486 of 512 actions remain in each seed-level comparison, with 71 of 72 median advantages still nonpositive. The result therefore persists on actions whose nominal steps stay inside the bounds.

Table 2: Transfer to independently fitted evaluators for bootstrap ensembles at $k = 0.5$ and $\eta = 0.05$. Advantage is $10^3 A_\eta$ from Eq. (8); positive values favor the penalized direction. Cosines compare corresponding primary and evaluator gradients. Within each seed, we take the median advantage and cosines over 512 actions. Entries show the median [min, max] of these five seed-level values. A win means greater evaluator objective decrease per unit realized displacement. The final column gives the median win fraction across seeds for the penalized direction.

Cost	Environment	Gain advantage ($\times 10^{-3}$)	Mean cross-fit cosine	Penalized cross-fit cosine	Penalized wins (%)
Velocity	Hopper	$+0.001$ [-0.689, $+1.038$]	0.987 [0.969, 0.991]	0.983 [0.963, 0.987]	50.0
	Walker2d	-0.783 [-2.153, -0.381]	0.975 [0.954, 0.984]	0.968 [0.943, 0.980]	40.4
	HalfCheetah	-0.277 [-3.955, $+0.311$]	0.861 [0.667, 0.871]	0.834 [0.631, 0.860]	48.2
Energy	Hopper	-0.102 [-0.503, $+0.169$]	0.985 [0.981, 0.996]	0.983 [0.981, 0.992]	43.9
	Walker2d	-0.274 [-0.412, -0.120]	0.973 [0.963, 0.980]	0.966 [0.961, 0.977]	38.9
	HalfCheetah	-0.616 [-1.294, -0.485]	0.878 [0.759, 0.929]	0.835 [0.747, 0.888]	39.1

Numerical sensitivity. Varying the relative gradient norm threshold from zero through 10^{-2} retains every evaluated action, and none has exact zero ensemble spread. Direct angle calculations agree to numerical precision with values reconstructed from saved cosines and from the parallel and orthogonal decomposition.

6 Discussion

The matched study shows why an ensemble penalty needs both directional and magnitude evaluation. In HalfCheetah, near-unit norm ratios coexist with substantial rotation. Gradient size alone would conceal this change, while angle alone would leave its strength unspecified. Signed parallel scale identifies attenuation or opposition along the mean-gradient axis. Together, these quantities describe how the penalty changes the local cost signal.

An independent evaluator tests whether the changed direction remains useful after refitting. Normalized descent along the penalized gradient gives the largest first-order decrease in its own differentiable objective. Scoring on another fit tests whether that advantage transfers. In our study, substantial rotation coexists with weaker directional agreement and generally nonpositive median transfer advantage. Sensitivity within a fit and transfer across fits thus reveal different properties of the penalty. Measuring both helps distinguish a direction that changes with the penalty from one whose local benefit persists across fitted critics.

7 Limitations

The experiments cover two controlled costs, one critic architecture, and three MuJoCo systems under a synthetic mixed behavior process. Five seeds support paired ranges but not precise estimates of population uncertainty. Euclidean angles depend on the chosen action coordinates. The finite-step evaluator measures local critic transfer rather than policy improvement. Extending the analysis to actor training requires the reward gradient, constraint multiplier and policy Jacobian. It also requires accounting for state weighting and the visitation distribution induced by the policy.

8 Conclusion

A mean-plus-spread cost penalty can substantially rotate the action gradient while leaving median scale and norm near one. The bootstrap effect is largest in HalfCheetah. Across independently seeded fits to the same data, mean gradients agree more closely than penalized gradients. The penalized direction has nonpositive median paired advantage in 71 of 72 finite-step comparisons. The audit connects these findings: it measures how the penalty changes a local direction and whether that direction transfers across fitted critics.

Reproducibility Statement

The supplementary material includes the core experimental code and run instructions. Section 4 describes data generation, cost targets, model fitting and evaluation settings, including seeds and numerical thresholds. Section 3 defines the estimands and aggregation; Section 5 reports clipping and sensitivity checks. The appendix presents the historical comparison separately and identifies its unmatched construction groups.

AI Use Statement

AI was used to assist with language polishing, improving clarity and readability, and refining the organization and presentation of the manuscript. The authors reviewed and verified the final content.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7436–7447, 2021.

Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhihong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations*, 2022.

Alexandre Capone, Ryan Kazuo Cosner, Aaron Ames, and Sandra Hirche. Learning safe control via on-the-fly bandit exploration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6726–6745. PMLR, 2025. URL https://proceedings.mlr.press/v267/capone25a.html.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596. PMLR, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. SUNRISE: A simple unified framework for ensemble learning in deep reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6131–6141. PMLR, 2021.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143. PMLR, 2020.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, John Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, J. K. Terry, 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*, volume 38, 2025. doi: 10.52202/085713-4916.

Trung Trinh, Markus Heinonen, Luigi Acerbi, and Samuel Kaski. Input-gradient space particle inference for neural network ensembles. In *The Twelfth International Conference on Learning Representations*, 2024.

Zifan Wu, Bo Tang, Qian Lin, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Off-policy primal-dual safe reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142, 2020.

Moritz Akiya Zanger, Wendelin Böhmer, and Matthijs T. J. Spaan. Diverse projection ensembles for distributional reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

APPENDIX

HISTORICAL BREADTH COMPARISON

The historical archive provides a separate breadth comparison across 27 five-member ensembles: three environments, three constructions and three seeds per group. Shared-data and bootstrap runs use seeds 101, 211 and 307. Randomized-prior runs instead use seeds 401, 503 and 607 with different behavior buffers. These construction contrasts are descriptive; the main study supplies the matched comparison.

At $E = 5$ and $k = 0.5$, the 27 ensemble median cosines have median 0.9957 ($5.3°$) and range 0.9786–0.9994. The affine R^2 between σ and μ, fitted across the evaluation segment, ranges from 0.002 to 0.697. This fit pools distinct states, measuring association among their values; the local angle concerns derivatives at a fixed state and behavior action. The two summaries therefore describe different properties of the same ensemble. Table S1 and Figure S1 retain all 27 ensembles.

Table S1: Historical alignment at behavior actions for $E = 5$ and $k = 0.5$. Entries are the median and range across three seeds. Randomized prior rows use separate seeds and buffers from the other constructions.

Environment	Construction	Median cosine [range]	Affine R^2 [range]
Hopper	shared data	0.9988 [0.9985, 0.9994]	0.125 [0.002, 0.218]
Hopper	bootstrap	0.9972 [0.9965, 0.9982]	0.232 [0.008, 0.337]
Hopper	randomized prior	0.9958 [0.9955, 0.9978]	0.367 [0.338, 0.697]
Walker2d	shared data	0.9987 [0.9983, 0.9988]	0.242 [0.215, 0.385]
Walker2d	bootstrap	0.9958 [0.9957, 0.9960]	0.286 [0.271, 0.293]
Walker2d	randomized prior	0.9954 [0.9945, 0.9956]	0.195 [0.155, 0.256]
HalfCheetah	shared data	0.9926 [0.9919, 0.9940]	0.150 [0.147, 0.271]
HalfCheetah	bootstrap	0.9817 [0.9786, 0.9857]	0.285 [0.249, 0.316]
HalfCheetah	randomized prior	0.9833 [0.9798, 0.9834]	0.144 [0.052, 0.205]

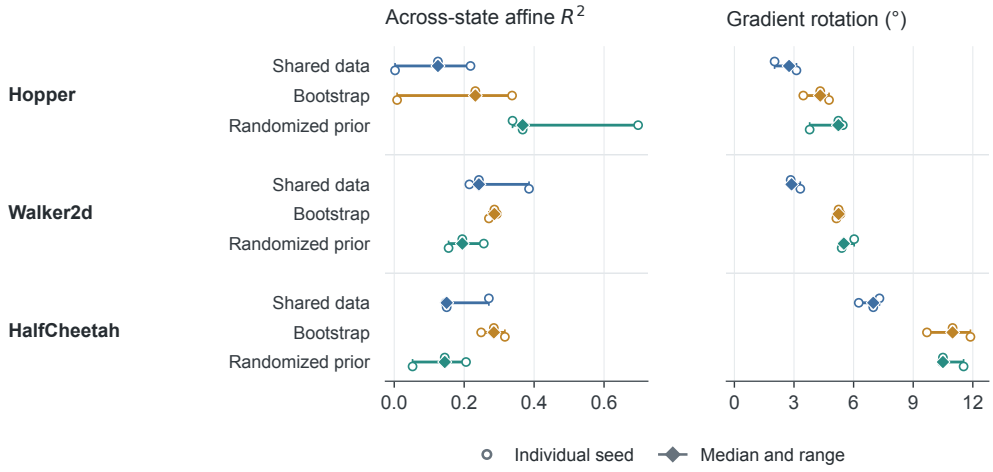


Figure S1: Historical value association and gradient rotation for $E = 5$, $k = 0.5$. Circles show individual ensembles; diamonds and bars show three-seed medians and min–max ranges. Left: affine R^2 pooled across 512 states. Right: local gradient rotation at behavior actions.

The archive also evaluates $k \in \{0.25, 0.5, 1\}$ on the same critics and one fixed nested subset of three members. Figure S2 reports weight trajectories using all five members. At $k = 1$, Hopper and Walker2d medians remain below approximately $13°$, while HalfCheetah bootstrap medians and randomized prior medians approach $20°$. The main matched study enumerates all ten subsets; this supplementary archive contains one nested subset.

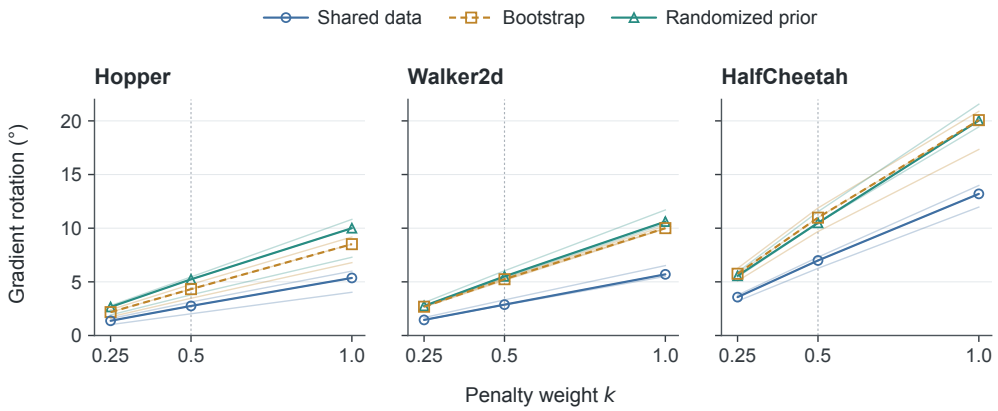


Figure S2: Historical penalty weight trajectories for ensembles with five members. Pale lines show individual seeds; marked lines convert the median cosine across three seeds to degrees. Randomized prior runs use separate buffers, so comparisons across constructions remain descriptive.