FISHER-FEATURE PROXIMAL POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Proximal Policy Optimization (PPO) loses plasticity in its trunk, and the usual diagnostic is the rank of the penultimate feature matrix. Proximal Feature Optimization (PFO) follows that diagnostic by penalizing feature drift equally in every direction. We argue that the diagnostic is the weak link. The policy sees the features only through two linear heads whose width is set by the action space, so rank, computed without the heads, cannot separate representations the policy treats very differently; and rank is a count settled at the tail of the spectrum, so it misses how variance is shared. We derive how an update on one state leaks into the clipped ratio of another: the leakage is governed by the cross-state coupling of the features, their squared inner products, which depends on how the spectrum is spread and on the feature norm, and not on a count. We measure the angular part by the participation ratio of the projected outputs, essentially the inverse mean squared cosine between states, and replace the isotropic penalty by the unique quadratic penalty that measures policy change, the Fisher information pulled back through the heads (FF-PPO). On six MuJoCo tasks the predicted mechanism is visible directly: FF-PPO lowers the coupling by more than an order of magnitude relative to PPO and several-fold relative to PFO, and wherever the penalties help, the methods order by coupling exactly as they order by trust-region excess. The gain in return grows with the width of the policy-visible subspace, where FF-PPO separates from PFO, and with the strength of the non-stationarity: with more optimization epochs per batch or under changing dynamics, PPO collapses while FF-PPO keeps learning.

1 INTRODUCTION

Proximal Policy Optimization (Schulman et al., 2017) improves sample efficiency by reusing each batch of collected trajectories for several gradient epochs. Policy learning is already non-stationary, since the states and rewards the agent sees depend on the policy that is changing, and multi-epoch optimization amplifies that: with each extra epoch the policy drifts further from the behaviour policy that produced the batch, and the surrogate objective becomes a worse approximation of the true return.

Training under such non-stationarity costs networks their *plasticity* (Abbas et al., 2023; Kumar et al., 2021), which in deep reinforcement learning shows up as feature rank collapse, capacity loss and dormant units (Sokar et al., 2023; Lyle et al., 2022). Moalla et al. (2024) gave the first systematic study of this in on-policy PPO, linked feature rank collapse to the gradient alignment that breaks PPO's trust region (Wang et al., 2020), and proposed Proximal Feature Optimization (PFO), an isotropic L_2 penalty on the drift of penultimate pre-activations between consecutive policies. PFO reduces collapse, but it treats every feature dimension alike, whereas effective learning systems compress the dimensions a task does not need while protecting the ones it does (Mack et al., 2020); a penalty that cannot tell the two apart must under-protect what matters or over-protect what does not.

PFO's design follows from its diagnostic: if the harm is lost directions, holding every feature coordinate in place stops any from disappearing, and a rise in rank is evidence that it worked. We think the first step is wrong, for a structural reason. The trunk produces a wide feature vector, but the policy sees it only through two linear heads, which emit the mean and the scale of the action distribution. Those few numbers are the whole policy at that state, so feature drift that leaves them unchanged changes nothing at all. Feature space splits into a part the policy can see, whose width is set by the action dimension, and a much larger part it cannot; rank, computed without the heads, cannot tell the

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.

two apart, and neither can a penalty built from it. Representation geometry is one channel through which PPO degrades, not the only one, and we study that channel (Section 7).

We make four contributions.

- We show the approximate rank fails as a collapse diagnostic in three independent ways: it is blind to the heads, it counts directions instead of measuring how variance is shared among them, and in a controlled task it stays fixed while learning slows by three quarters (Section 3).
- We replace it with the participation ratio of the trunk's projected outputs, which for normalized outputs is an exact function of the mean squared cosine between states, and we show that the same spectral quantity governs how far an update on one state leaks into the clipped ratio of another (Section 3).
- We show that once policy change is what we are measuring, the penalty follows: among quadratic forms on feature space, the one that measures the policy's own change is the Fisher information pulled back through the heads (Section 4).
- On six MuJoCo tasks with ten seeds, we observe the predicted mechanism directly: the cross-state coupling of eq. (8) orders the methods as the trust-region excess does, and FF-PPO reduces it by more than an order of magnitude. The return gain grows with the width $2d$ of the policy-visible subspace and with the strength of the non-stationarity, where PPO collapses and FF-PPO does not (Section 5).

2 Background

2.1 PPO and the training objective

PPO maximizes the clipped surrogate $L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t[\min(r_t\hat{A}_t,\ \mathrm{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t)]$ of Schulman et al. (2017), with $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)$ and GAE advantages $\hat{A}_t$ (Schulman et al., 2016), where ϵ is the clip width. With the unclipped value loss L^{VF} and the batch-mean entropy $\mathcal{H}$, the full objective we train, including the penalty introduced in Section 4, is

$$L(\theta) \;=\; \underbrace{L^{\mathrm{CLIP}}(\theta) - c_{\mathrm{VF}}L^{\mathrm{VF}}(\theta) + c_H\mathcal{H}(\theta)}_{\text{standard PPO}} \;-\; \beta\,\mathcal{P}(\theta), \tag{1}$$

where $\mathcal{P}$ is a penalty on feature drift and $\beta \geq 0$ its coefficient; PFO and our method differ only in $\mathcal{P}$. We set $c_H = 0$, since an entropy bonus pushes the action scale upward and the collapse of σ is part of what we observe.

2.2 Policy architecture and diagnostics

Following Haarnoja et al. (2018), the policy is a TanhNormal whose mean *and* standard deviation (scale) both depend on the state, since a state-independent scale would make half of the output heads irrelevant to feature drift. The actor and critic have separate trunks. An observation $s \in \mathbb{R}^n$ passes through the actor as

$$\begin{aligned} z_1 &= \theta_1 s + b_1, & \phi_1 &= f(z_1), & z_2 &= \theta_2\phi_1 + b_2, & \phi_2 &= f(z_2), \\ \mu &= \theta_\mu\phi_2 + b_\mu, & \psi &= \theta_\sigma\phi_2 + b_\sigma, & \sigma &= g(\psi). \end{aligned} \tag{2}$$

Here $\theta_1 \in \mathbb{R}^{D_1\times n}$ and $\theta_2 \in \mathbb{R}^{D\times D_1}$ are the trunk weights, $\theta_\mu, \theta_\sigma \in \mathbb{R}^{d\times D}$ are the mean and scale heads, and f is an elementwise activation ($\tanh$ here). z_ℓ is the pre-activation of layer ℓ and ϕ_ℓ its post-activation; PFO penalizes $z_2(s) \in \mathbb{R}^D$, and $\phi_2(s)$ is the input to the linear heads. The scale is a shifted softplus with a floor,

$$\sigma = g(\psi) = \sigma_{\min} + \mathrm{softplus}(\psi + c), \qquad u \sim \mathcal{N}\big(\mu(s),\, \mathrm{diag}(\sigma^2(s))\big), \quad a = \tanh(u), \tag{3}$$

with $c = \mathrm{softplus}^{-1}(1-\sigma_{\min})$ so that $\sigma = 1$ at initialization.

Write $\theta_{\mathrm{comb}} = [\theta_\mu; \theta_\sigma] \in \mathbb{R}^{2d\times D}$ for the two heads stacked, so that $y(s) = \theta_{\mathrm{comb}}\,\phi_2(s) \in \mathbb{R}^{2d}$ collects the policy outputs (μ, ψ) up to the fixed biases. The row space $\mathcal{R} = \mathrm{rowsp}(\theta_{\mathrm{comb}})$, of

dimension at most $2d$, is the *output-relevant subspace*: the only directions of feature space the policy depends on. Its complement $\mathcal{R}^{\perp}$ has dimension at least $D - 2d$. For a batch of N states, such as a rollout, the vectors are stacked row by row into $N \times D$ matrices Z_2 (pre-activations) and Φ_2 (post-activations), and into the $N \times 2d$ matrix of policy outputs $Y = \Phi_2 \theta_{\text{comb}}^{\top}$. The collapse diagnostic used by PFO is the approximate rank of Φ_2,

$$\text{rank}_{\delta}(\Phi_2) = \min \left\{ k : \frac{\sum_{i=1}^{k} s_i^2(\Phi_2)}{\sum_{j=1}^{D} s_j^2(\Phi_2)} \geq 1 - \delta \right\}, \tag{4}$$

where $s_1(\Phi_2) \geq s_2(\Phi_2) \geq \cdots$ are the singular values of Φ_2 (we write s_i rather than σ_i to keep σ for the policy variance). We use $\delta = 0.01$ throughout, so rank_{δ} counts the principal directions needed to recover 99% of the variance.

3 What Feature Rank Misses

Figure 1 states the problem in one picture: five representations with the same information, total variance and approximate rank, differing only in how the variance is shared among six directions, learn at visibly different speeds under ordinary PPO, in the order of their participation ratio. This section explains why, and what to measure instead.

PFO pairs a diagnostic, $\text{rank}_{\delta}(\Phi_2)$, with a remedy, the drift penalty $\mathcal{P} = \mathbb{E}_s\big[\|z_2^{\theta}(s) - z_2^{\text{old}}(s)\|_2^2\big]$. We argue that each has a limitation of its own and that the two are not matched to each other, derive why the spread of the spectrum rather than a count governs the failure of the trust region, and replace the rank with a measure that tracks that spread.

3.1 The penalty and the diagnostic see different quantities

The diagnostic is computed on the post-activation features $\Phi_2 = f(Z_2)$, the penalty on the pre-activations Z_2. Since $\tanh$ is 1-Lipschitz the penalty does bound the drift of ϕ_2, but the converse fails: growth in z_2 that drives units into saturation changes Φ_2 very little, so the magnitude the penalty acts on is discarded before the diagnostic sees it, consistent with the growing pre-activation norm and stable feature norm that Moalla et al. (2024) observe. The two also differ in kind: rank is a *level* property across states at one time, the penalty constrains the per-state *increment* between consecutive policies, and small increments do not prevent slow decay. FF-PPO keeps the same penalty site so that its comparison with PFO is one of metrics alone. Finally the penalty is isotropic: the policy sees ϕ_2 only through its heads, so only $\mathcal{R}$ affects π_{θ}, yet every direction of $\mathbb{R}^D$ is charged alike, $\mathcal{R}^{\perp}$ included, and inside $\mathcal{R}$ neither the spread of variance nor the policy's sensitivity along each direction is weighed.

3.2 Rank is a counting measure

The diagnostic has two problems of its own, both visible in eq. (4). It is a *counting* measure: a direction either enters the count or it does not, and how variance is shared among the directions that do enter makes no difference. And the count is settled at the *bottom* of the spectrum, since the reported value is the index at which cumulative variance first passes 99%, which depends on how many weak directions are needed to cover the last percent. A spectrum in which one direction holds 92% of the variance and five others share the remainder equally has $\text{rank}_{0.01} = 6$, exactly like a spectrum in which all six are equal, although in the first case almost every state is represented along the same axis. Section 3.6 shows that this difference matters for learning: five spectra with identical rank produce clearly different learning speeds under ordinary PPO.

The toy analysis in Moalla et al. (2024) uses one-dimensional features, where rank one and perfect colinearity coincide; for $D > 1$ a representation can be nearly colinear while its rank stays high, which is why no principled line between helpful low rank and harmful aliasing could be drawn. On MuJoCo the methods with the lowest rank end with the highest return (Section 5).

3.3 Cross-sample coupling

The reason a count is the wrong summary comes from how clipping fails. Wang et al. (2020) show that a sample whose ratio is already clipped can still be pushed further out when its gradient aligns

with the gradients of the other samples in the batch. Write $g_t = \nabla_\theta \log \pi_\theta(a_t \mid s_t)$ for the gradient of sample t and $\Gamma_{tt'} = \langle g_t, g_{t'} \rangle$ for the inner product of two such gradients. One gradient step of size α on a batch of N samples, early in an update phase where $r_{t'} \approx 1$, moves the log-ratio of sample t by

$$\Delta \log r_t \;\approx\; \frac{\alpha}{N} \Gamma_{tt} \hat{A}_t \;+\; \underbrace{\frac{\alpha}{N} \sum_{t' \neq t} \Gamma_{tt'} \hat{A}_{t'}}_{\text{leakage}} . \tag{5}$$

The first term is sample t acting on itself, which clipping does control. The rest is *leakage*: the part of t's ratio change produced by other samples, which clipping t cannot touch, and which pushes r_t further out whenever $\hat{A}_t \hat{A}_{t'} \Gamma_{tt'} > 0$.

For linear heads taking $\phi_2(s_t)$ as input, the head part of the gradient is an outer product, $\xi_t \, \phi_2(s_t)^\top$, where $\xi_t = \nabla_y \log \pi_\theta(a_t \mid s_t) \in \mathbb{R}^{2d}$ is the score with respect to the policy outputs. The head contribution to the inner product therefore splits into a score part and a feature part,

$$\Gamma_{tt'} = \langle \xi_t, \xi_{t'} \rangle \, \big\langle \phi_2(s_t), \phi_2(s_{t'}) \big\rangle. \tag{6}$$

The factorization is standard (Bengio et al., 2020; Ren & Sutherland, 2024; Deng et al., 2025; Liu et al., 2025); the feature factor vanishes for orthogonal representations, as in a tabular policy.

At the start of an update phase the actions of different samples are independent draws, so their scores are independent with mean zero, and the expectation of eq. (6) is zero: the leakage between two samples is as likely to push one way as the other, and its sign carries no information. What accumulates over a batch is its size, so we square. Because the features are fixed once the states are drawn, only the score factor is averaged, and its second moment is set by the Fisher information $F(s)$ of the action distribution in output coordinates (Section 4, Appendix A):

$$\mathbb{E}\big[\Gamma_{tt'}^2 \,\big|\, s_t, s_{t'}\big] \;=\; \underbrace{\operatorname{tr}\big(F(s_t) F(s_{t'})\big)}_{\text{a weight}} \, \big\langle \phi_2(s_t), \phi_2(s_{t'}) \big\rangle^2. \tag{7}$$

The weight depends on how confident the policy is at the two states. With a state-dependent scale it differs from pair to pair, and that variation is the confidence weighting Section 4 builds on. To isolate the other factor, consider the case where the scale does *not* vary across states. The weight is then the same constant $c = \operatorname{tr}(F^2)$ for every pair, and summing over distinct pairs gives

$$\mathbb{E}\Big[\sum_{t \neq t'} \Gamma_{tt'}^2 \Big] \;=\; c \Big(\sum_i s_i^4(\Phi_2) \;-\; \sum_t \|\phi_2(s_t)\|^4 \Big). \tag{8}$$

For fixed total variance $\sum_i s_i^2(\Phi_2)$, the sum of fourth powers is smallest when the spectrum is flat and largest when it is concentrated, and it changes continuously as variance moves between directions whether or not any direction crosses a threshold. Expected leakage is governed by how the variance is spread, not by a count of directions. It also grows as the fourth power of the feature norm, so anything that inflates the features, such as pre-activations growing into the saturated range of $\tanh$, raises it even when the angles between states do not change. Rank enters only through the bound $\mathrm{PR} \leq \mathrm{rank}$ (eq. (9) below): a low rank forces a high sum of fourth powers and so can warn of trouble, but a high rank cannot rule it out.

The expression is in Φ_2, not the outputs, because a head update couples states through their full feature inner product, $\mathcal{R}^\perp$ included.

3.4 Projecting does not repair the measurement

The policy knows a state only through $y(s)$, so the outputs $Y = \Phi_2 \theta_{\mathrm{comb}}^\top$ are the natural place to diagnose collapse; they remove $\mathcal{R}^\perp$ and explain Moalla et al. (2024)'s observation that in Atari the collapse threshold falls at single-digit ranks, since $\mathrm{rank}(Y) \leq 2d$ (or $|\mathcal{A}|$ for discrete actions). But rank on Y is still a count: it saturates at $2d$ values, stays tail-sensitive, and one eigenvalue crossing the threshold moves it by one unit out of at most $2d$.

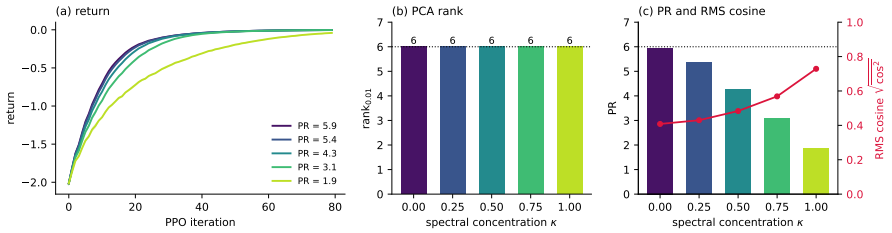

Figure 1: **Same rank, different spectrum, different learning.** Six fixed feature directions with equal total variance and a spectrum ranging from uniform ($\kappa = 0$) to concentrated ($\kappa = 1$); only the policy heads are trained, with ordinary PPO. **(a)** Return, mean over ten seeds with 95% confidence bands; the legend gives $\mathrm{PR}(\Phi_2)$ and $\mathrm{PR}(\hat{\Phi}_2)$ for each spectrum. **(b)** The approximate rank is 6 for all five spectra and cannot tell them apart. **(c)** Participation ratio of the raw features $\mathrm{PR}(\Phi_2)$ (solid) and of the row-normalized features $\mathrm{PR}(\hat{\Phi}_2)$ (hatched), with the RMS cosine (red). The RMS cosine and $\mathrm{PR}(\hat{\Phi}_2)$ are two forms of the same quantity (eq. (10)); $\mathrm{PR}(\Phi_2)$ is the one that governs leakage (eq. (8)).

3.5 What we measure instead

We replace counting with the participation ratio

$$\mathrm{PR}(A) = \frac{\left(\sum_i s_i^2(A)\right)^2}{\sum_i s_i^4(A)}, \qquad 1 \le \mathrm{PR}(A) \le \mathrm{rank}(A), \tag{9}$$

computed from the singular values $s_i(A)$ of the matrix A as given. It reports how evenly the spectrum is spread rather than how many directions are occupied (Gao et al., 2017; Litwin-Kumar et al., 2017). It is continuous, dominated by the top of the spectrum, and separates the two spectra of Section 3.2 (1.18 against 6, where $\mathrm{rank}_{0.01}$ gives 6 for both); its denominator is the sum of fourth powers in eq. (8), so at fixed total variance a smaller $\mathrm{PR}(\Phi_2)$ means proportionally more leakage.

To interpret it as an angle, normalize each row of A to unit length and write $\hat{A}$; for any matrix with unit-norm rows $\hat{a}_t$,

$$\frac{1}{\mathrm{PR}(\hat{A})} = \frac{1}{N} + \frac{N-1}{N}\,\overline{\cos^2}, \qquad \overline{\cos^2} = \frac{1}{N(N-1)} \sum_{t \neq t'} \cos^2\!\left(\hat{a}_t, \hat{a}_{t'}\right). \tag{10}$$

an identity, not an approximation (Appendix B; it is the dimension formula of Litwin-Kumar et al. (2017) on a Gram matrix with unit diagonal). When all rows share a norm ρ the raw and normalized ratios coincide and eq. (8) becomes $c\,\rho^4 N(N-1)\,\overline{\cos^2}$.

Rearranging eq. (10) gives $1/\mathrm{PR}(\hat{A}) = \overline{\cos^2} + (1 - \overline{\cos^2})/N$, so for batches large enough that $\overline{\cos^2} \gg 1/N$ the participation ratio is approximately the inverse mean squared cosine,

$$\mathrm{PR}(\hat{A}) \;\approx\; \frac{1}{\overline{\cos^2}}, \qquad \text{relative error} \;\le\; \frac{1}{N\,\overline{\cos^2}}. \tag{11}$$

With $N = 2048$ the error is under 2% on every task (Appendix B), so an average cosine translates directly into an effective number of directions the policy uses to tell states apart.

Our collapse diagnostic applies eq. (10) to the policy outputs: $\hat{Y}$ is Y with each column centred across the batch and each row then normalized (the reason for centring is given in Section 5), and since $\hat{Y}$ has $2d$ columns, $\mathrm{PR}(\hat{Y}) \le 2d$. The two matrices therefore play different roles. Φ_2 is what couples updates, through eq. (8); $\hat{Y}$ is what the policy can tell apart. We therefore report both: the angular part through $\hat{Y}$, and the full coupling of eq. (8), which adds the feature norm. The two separate exactly when norms change, which is what happens when pre-activations grow (Section 5).

The sign of the cosine carries no information, only its square does; Appendix B explains why a cosine of -1 is collapse rather than diversity.

3.6 A CONTROLLED EXAMPLE

Figure 1 comes from a one-step contextual bandit in which the answer is known in advance and only the spread of the feature variance changes: a hidden state $x \sim \mathcal{N}(0, I_6)$, reward $-\|a - x\|^2/6$, fixed features $\phi_2 = V\Lambda^{1/2}x$ with a random rotation V and a spectrum $\Lambda(\kappa)$ swept from uniform to the concentrated one of Section 3.2, and only the heads trained with ordinary PPO (construction and closed-form solution in Appendix C). Over ten seeds the approximate rank is 6 in every setting, while $\mathrm{PR}(\Phi_2)$ falls from 5.92 to 1.18, $\mathrm{PR}(\hat{\Phi}_2)$ from 5.94 to 1.87 and the RMS cosine rises from 0.41 to 0.73. The leakage factor of eq. (8) grows 5.1-fold, matching the 5.0 predicted by the fall in $\mathrm{PR}(\Phi_2)$, and the iterations needed to reach the return of the uniform spectrum at iteration 20 are 20, 21, 22, 25 and 35: curvature along direction i is proportional to λ_i, so the condition number grows from 1 to 57.5 at constant rank.

4 THE FISHER PULLBACK

A penalty on feature drift should charge nothing for drift the policy cannot see and, among the directions it can see, charge by how much the policy changes rather than how far the features moved. Every such penalty is a choice of a symmetric matrix $M(s)$,

$$\mathcal{P} = \mathbb{E}_s\big[\delta z_2(s)^\top M(s)\, \delta z_2(s)\big], \qquad \delta z_2(s) = z_2^\theta(s) - z_2^{\mathrm{old}}(s), \tag{12}$$

with PFO the case $M = I$. Neither requirement has to be imposed by hand: one condition settles M.

Proposition 1 (Uniqueness of the metric)**.** *Among symmetric matrices M, exactly one has the property that $\frac{1}{2}\,\delta^\top M\,\delta$ agrees to second order in δ with $\mathrm{KL}(\pi_y \,\|\, \pi_{y+\delta})$, the divergence caused by perturbing the policy outputs y by δ. It is the Hessian of that divergence at $\delta = 0$, which is the Fisher information $F(s)$ of the policy in output coordinates.*

The proof is three lines (Appendix D). That the Fisher information is the Hessian of the KL divergence is classical and underlies natural-gradient and trust-region methods (Kakade, 2001; Schulman et al., 2015; Martens & Grosse, 2015); we claim no novelty for it. What it buys is an interpretation of eq. (12): among quadratic penalties, the pullback of F is the one that measures the change in the policy, and any other M measures something else. The relevant Fisher is the $2d \times 2d$ one of the action distribution, not the full-parameter Fisher of TRPO, and the PPO update is unchanged. Equivalently, $\mathcal{P}_{\mathrm{FF}}$ is the second-order expansion of the per-state $\mathrm{KL}(\pi_{\mathrm{old}} \,\|\, \pi_\theta)$ with the heads detached, so the natural baseline is that divergence itself (Section 5).

4.1 THE FISHER INFORMATION AND ITS PULLBACK

The $\tanh$ squashing cancels in every log-ratio and divergence, so we work with the latent Gaussian. The Fisher information is the covariance of the score with respect to the outputs (μ, ψ) (Appendix D),

$$F(s) = \mathrm{diag}\Big(\underbrace{\tfrac{1}{\sigma_1^2}, \ldots, \tfrac{1}{\sigma_d^2}}_{\text{mean block}}, \underbrace{\tfrac{2\,g'(\psi_1)^2}{\sigma_1^2}, \ldots, \tfrac{2\,g'(\psi_d)^2}{\sigma_d^2}}_{\text{scale block}}\Big), \tag{13}$$

with 2 the variance of a one-degree chi-square. Since $y = \theta_{\mathrm{comb}}\phi_2$, the metric transports by the chain rule, and $M = \theta_{\mathrm{comb}}^\top F \theta_{\mathrm{comb}}$ in eq. (12) gives the FF-PPO penalty,

$$\mathcal{P}_{\mathrm{FF}} = \mathbb{E}_s\Big[\delta z_2(s)^\top\, \theta_{\mathrm{comb}}^\top F(s)\, \theta_{\mathrm{comb}}\, \delta z_2(s)\Big]. \tag{14}$$

θ_{comb} and $\sigma(s)$ are detached, so eq. (14) changes only the gradient reaching the trunk; its null space is exactly $\mathcal{R}^\perp$, and it is invariant to reparameterizing the features (Appendix D).

One approximation is deliberate: $\theta_{\mathrm{comb}}^\top F \theta_{\mathrm{comb}}$ is the exact metric for drift of ϕ_2, and applying it to δz_2 drops the activation Jacobian $\mathrm{diag}(f'(z_2))$. We keep it so that the regularized quantity is the same z_2 PFO regularizes and the comparison is one of metrics alone; the cost is that saturated units are over-weighted (Appendix D), and the corrected metric is an ablation.

Table 1: Final return on six MuJoCo tasks after five million steps, IQM with 95% bootstrap CI over ten seeds; best per task in bold. $2d$ is the width of the policy-visible subspace ($D = 64$). Drop from peak is in Table 5 (Appendix F).

	Humanoid	Ant	HalfCheetah	Walker2d	Hopper	Swimmer
$2d$	34	16	12	12	6	4
PPO	639 [595, 691]	2099 [1469, 2461]	2429 [1815, 3366]	3148 [2657, 3632]	2327 [2186, 2417]	112 [98, 120]
PFO	833 [714, 1092]	3788 [3118, 4148]	3774 [2414, 4774]	3229 [3001, 3535]	2277 [2126, 2439]	51 [51, 60]
FF-PPO-log	**1918** [1315, 2452]	3563 [3044, 4060]	3784 [2640, 4888]	3243 [3087, 3434]	2313 [2122, 2439]	51 [50, 52]
FF-PPO-exact	1020 [866, 1635]	**3857** [3457, 4307]	3455 [1954, 5210]	**3358** [2990, 3713]	2232 [2000, 2492]	51 [50, 51]
FF-PPO-mean	789 [725, 928]	3191 [2711, 3758]	2568 [1869, 4183]	3228 [2969, 3383]	**2380** [2250, 2488]	51 [50, 51]
Detached KL	858 [763, 1064]	3003 [2708, 3294]	3681 [2374, 4475]	–	2154 [1923, 2428]	52 [34, 90]
PPO + LayerNorm	752 [683, 863]	2107 [1560, 2605]	**3957** [2568, 4951]	–	2179 [2009, 2294]	**124** [123, 125]

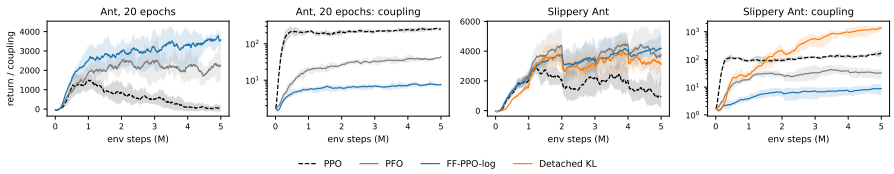

Figure 2: **Stronger non-stationarity.** Ant with twenty optimization epochs per batch, and Slippery Ant (ground friction re-sampled log-uniformly in $[0.02, 2]$ every million steps, ten epochs). Return and cross-state coupling (log scale) for each. PPO's coupling grows and its return collapses; FF-PPO keeps the coupling more than an order of magnitude lower and holds its return. IQM over ten seeds; bands are the interquartile range.

5 Experiments

Setup. We compare PPO, PFO and three variants of FF-PPO on six MuJoCo (Todorov et al., 2012) tasks, HalfCheetah, Hopper, Walker2d, Ant, Humanoid and Swimmer, using the architecture of Section 2.2. The variants differ only in the scale block of eq. (13): **FF-PPO-log** uses the constant 2, exact for an exponential scale head; **FF-PPO-exact** uses the exact weight of our softplus head; and **FF-PPO-mean** sets the block to zero (Appendix E). We add three baselines: the **detached KL** penalty $\mathbb{E}_s[\mathrm{KL}(\pi_{\mathrm{old}}(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s))]$ in closed form with the gradient stopped at the heads, which is the divergence that Proposition 1 approximates to second order; PPO with **LayerNorm** in the actor trunk; and PPO with the actor trunk **reset** every million steps (Nikishin et al., 2022). All penalized methods use $\beta = 1$ and all methods share the same hyperparameters on every task with no per-environment tuning (Appendix G); ten optimization epochs per batch as in Moalla et al. (2024), whose CleanRL variant our implementation extends. Every configuration is run with ten seeds for five million steps, and we report the interquartile mean (IQM) over seeds with 95% bootstrap confidence intervals (Agarwal et al., 2021). Final return is the mean over the last 5% of training; the drop is the difference between the peak of the smoothed curve and the final return.

Diagnostics. Every diagnostic is computed once per PPO iteration on the full rollout batch of $N = 2048$ states, under the policy that collected it and before it is updated, in float64. The rank is $\mathrm{rank}_{0.01}(\Phi_2)$. The angular diagnostic is $1/\overline{\cos^2}(\hat{Y})$, where $\hat{Y}$ is the policy output matrix $Y = \Phi_2 \theta_{\mathrm{comb}}^\top$ with columns centred and rows normalized; by eq. (11) it equals $\mathrm{PR}(\hat{Y})$ to within 2%, higher means more directions in use, and its ceiling is $2d$. The *coupling* is the feature factor of eq. (8) per pair, $\frac{1}{N(N-1)} \sum_{t \neq t'} \langle \phi_2(s_t), \phi_2(s_{t'}) \rangle^2$: by eq. (6), how strongly a head update on one state moves the policy at another, averaged over pairs. It is zero for mutually orthogonal features, as in a tabular policy, and it combines the angle between states with their length, since each term equals $\|\phi_2(s_t)\|^2 \|\phi_2(s_{t'})\|^2 \cos^2$. The *excess ratio* measures how far the trust region is exceeded: the mean probability ratio of samples above $1+\epsilon$ divided by the mean of those below $1-\epsilon$, averaged over the minibatches of an update (Moalla et al., 2024).

Under stronger non-stationarity the gap becomes an order of magnitude. Figure 2 turns the non-stationarity up in two ways. Raising the optimization epochs per batch, the tool Moalla et al. (2024) use, takes PPO on Ant from 3564 at five epochs to 2099 at ten and to 40 [−107, 390] at

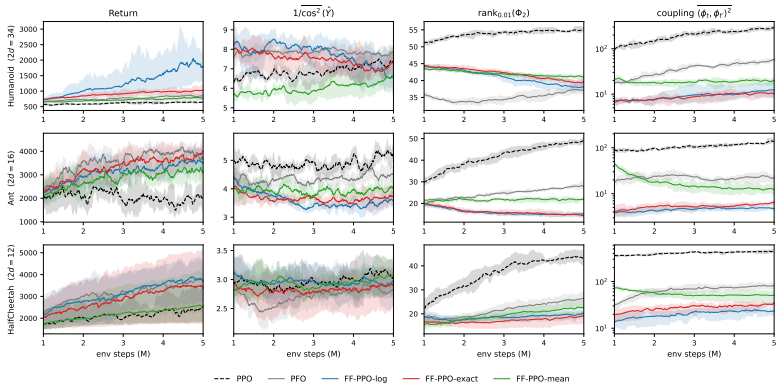

Figure 3: **Return and representation diagnostics, from 1M steps.** Humanoid, Ant and HalfCheetah, the three tasks on which PPO degrades and the penalties help, where the coupling analysis applies; all six tasks over the full five million steps, including the Swimmer limitation, are in Figure 4, and every task is in Table 1. Columns are the episode return, the angular diagnostic $1/\overline{\cos^2}(\hat{Y}) \approx \mathrm{PR}(\hat{Y})$ of the centred, row-normalized policy outputs, the approximate rank $\mathrm{rank}_{0.01}(\Phi_2)$, and the cross-state coupling of eq. (8) on a log scale. On Ant, PFO leads for most of training and reaches a higher peak, but loses twice as much of it, and the two end level (paired difference 69 $[-560, 810]$); FF-PPO's coupling is five times lower throughout, and under stronger non-stationarity the order of the returns reverses (Figure 2). All penalized methods use $\beta = 1$. IQM over ten seeds after exponential smoothing; bands are the interquartile range across seeds.

twenty, a complete collapse; PFO falls to 2274; FF-PPO-log is flat at 3612, 3563 and 3609. Hopper is insensitive to epochs for every method. On Slippery Ant, where the ground friction changes every million steps, PPO ends at 1012 [399, 2020] and FF-PPO-log at 4120 [3040, 4984], with PFO at 3702 and the detached KL at 3328.

The mechanism is visible directly. Table 3 (Appendix F) and the last column of Figure 3 measure the quantity Section 3.3 says governs the failure of the trust region. Under PPO the coupling is 131 to 446 on every task but Swimmer (1643), because the pre-activations grow into the saturated range of $\tanh$ and inflate the features; every penalty stops that growth, and the Fisher metric stops it most, leaving the coupling at least 17 times lower than under PPO and 4 to 7 times lower than under PFO on every task except Hopper. In each of Humanoid, Ant at ten and twenty epochs, Slippery Ant and HalfCheetah the methods order by coupling exactly as by excess ratio (rank correlation 1.0), and in the reverse order of return; doubling the epochs on Ant doubles PPO's coupling (131 to 261), quadruples its excess ratio (2.95 to 12.2) and collapses its return, while FF-PPO barely moves. The angular diagnostic alone does not show this: on Ant FF-PPO's features are *more* aligned than PPO's ($\overline{\cos^2}(\Phi_2)$ 0.17 against 0.10), yet its coupling is 27 times smaller because its features are three times shorter, the gap Section 3.5 anticipates; and the rank, which rises under PPO to 50, points the wrong way entirely.

Two cases qualify the picture. The detached KL on Slippery Ant keeps the excess low (2.27) with a coupling as high as PPO's (1246), suppressing the symptom without removing its cause, and ends below FF-PPO. On Swimmer the penalized methods have both the lowest coupling and the lowest return: the trunk is frozen (rank 3 to 4 against 23). For PFO this is over-regularization, and a coefficient scaled to the visible width, $\beta = 2d/D$, restores PPO's return (118). FF-PPO does not recover at any β down to 0.03: its scale collapses to $\sigma \approx 0.07$ against 0.38 for PPO, and the $1/\sigma^2$ mean block of eq. (13) then multiplies the penalty thirty-fold whatever β is. Low coupling is necessary for the trust region, not sufficient for learning, and the confidence weighting needs a floor. On Hopper, where nothing degrades, coupling does not order the excess ratio.

The benefit of a feature penalty grows with the visible width. Table 1 orders the tasks by $2d$, widest first, and the ratio of FF-PPO-log to PPO falls almost monotonically along it: 3.00, 1.70, 1.56, 1.03, 0.99 and 0.46. PFO follows the same trend, so it belongs to feature-drift penalties in general: the more of feature space the heads use, the more a penalty protects. On Hopper and Walker2d

all methods tie, and on HalfCheetah and Ant the penalized methods beat PPO with PFO and FF-PPO indistinguishable (paired difference on Ant 69 [-560, 810]). FF-PPO separates from PFO on Humanoid, the widest subspace: 1918 [1315, 2452] against 833 [714, 1092] for PFO and 639 [595, 691] for PPO, a paired difference over seeds of 1085 [496, 1542] against PFO; the detached KL that the Fisher metric approximates reaches only 858. The drop from peak agrees: on Ant the FF-PPO variants lose 366 to 434 against 1214 for PPO and 733 for PFO.

Ablations, coefficient and baselines. On Humanoid neither half of the Fisher metric alone reaches the full pullback (null space only 1016, confidence weighting only 1251, against 1918); on Ant the metric choice does not matter; $\beta = 1$ is untuned for both penalties (on Ant PFO peaks at $\beta = 0.3$ and FF-PPO-log at 0.1); PPO with LayerNorm is best on HalfCheetah and Swimmer; trunk resets hurt on every task (576 to 2755, Appendix F).

6 Related Work

TRPO (Schulman et al., 2015) enforces a KL constraint; PPO replaces it with clipping, which Wang et al. (2020) show fails under gradient alignment, and FR-PPO (Lascu et al., 2025) replaces it with a Fisher–Rao penalty. Rank deterioration under non-stationarity was documented in value-based methods (Kumar et al., 2021; Lyle et al., 2022), characterized through dormant units (Sokar et al., 2023) and feature norms (Abbas et al., 2023), and extended to PPO by Moalla et al. (2024), who also track dead units, pre-activation norms and capacity loss. DR3 (Kumar et al., 2022) and Representation Distinction (Ma et al., 2023) regularize features in offline RL; EWC (Kirkpatrick et al., 2017) penalizes parameter drift under a diagonal Fisher, and K-FAC (Martens & Grosse, 2015) and the natural gradient (Kakade, 2001) use it as a preconditioner. FF-PPO instead places the exact $2d \times 2d$ Gaussian Fisher, pulled back through the heads, on the features, with the policy-invisible subspace as its null space.

7 Discussion and Limitations

The analysis predicts a chain: pre-activation growth inflates the features, the coupling of eq. (8) grows with them, the trust region leaks, and the policy degrades. Each link is measured in Table 3 (Appendix F), and the chain is strongest where the policy-visible subspace is wide or the non-stationarity is strong; where the subspace is narrow, the same penalties do harm at the shared coefficient.

The coefficient, the visible width and the confidence weight. The FF-PPO penalty sums $2d$ squared output perturbations and the PFO penalty D squared feature perturbations, so at a fixed β their strength relative to the surrogate scales with the width of the space they act on. On Swimmer ($2d = 4$) that strength froze the trunk; on Ant the sweep found both methods best at β between 0.1 and 0.3. A coefficient that scales with the visible width, $\beta = \beta_0\, 2d/D$, is the natural normalization, and for PFO it repairs Swimmer. It does not repair FF-PPO, whose effective strength is set by $1/\sigma^2$ rather than by β: where the scale collapses, the penalty grows without bound and locks the trunk. The remedy is a floor on the confidence weight, or the log-coordinate scale block alone; on Humanoid, by contrast, reducing β below 1 only hurts FF-PPO, so the weighting is right where the visible subspace is wide. We leave the floor to future work.

Other channels and limits. The trust region can also fail through optimizer momentum (Wang et al., 2020), the critic adds its own error, plasticity loss has channels the spectrum does not capture (Sokar et al., 2023; Lyle et al., 2022), and a policy whose scale falls to its floor stops exploring; a method repairing the geometric channel should slow collapse rather than prevent it, and we read the results in that light. One tension could have undermined the method: eq. (8) includes the component of the features in $\mathcal{R}^\perp$, which FF-PPO leaves free to compress, and its rank does fall. The coupling measurements settle it: the compression comes with shorter features, and the full coupling, $\mathcal{R}^\perp$ included, is lowest under FF-PPO on every task. Finally, the participation ratio says that states are merged, not whether merging them is compression or aliasing, so on MuJoCo we interpret it alongside the return; the diagnostics are on-policy, so a frozen probe set is the control; and eq. (14) omits the activation Jacobian, at the cost of over-weighting saturated units.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing prose. We did not use generative AI tools to generate research ideas or experimental results. All derivations were checked by the authors, all code was reviewed and tested, and we take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The architecture is specified in Section 2.2, the penalty in eq. (14), and the three variants in Appendix E. Complete hyperparameters are in Appendix G; the derivation of the Fisher information and of the pullback is in Appendix D; the controlled example is specified in Section 3.6 and Appendix G.1, with its closed-form solution in Appendix C. Source code, job files and the analysis scripts that produce every table and figure are provided as supplementary material.

REFERENCES

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on Lifelong Learning Agents (CoLLAs)*, pp. 620–636, 2023.

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

Emmanuel Bengio, Joelle Pineau, and Doina Precup. Interference and generalization in temporal difference learning. In *International Conference on Machine Learning (ICML)*, 2020.

Wenlong Deng, Yi Ren, Muchen Li, Danica J. Sutherland, Xiaoxiao Li, and Christos Thrampoulidis. On the effect of negative gradient in group relative deep reinforcement optimization. *arXiv preprint arXiv:2505.18830*, 2025.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018.

Sham Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems (NIPS) 14*, 2001.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Aviral Kumar, Rishabh Agarwal, Tengyu Ma, Aaron Courville, George Tucker, and Sergey Levine. DR3: Value-based deep reinforcement learning requires explicit regularization. In *International Conference on Learning Representations (ICLR)*, 2022.

Andrei Lascu et al. Fisher-rao proximal policy optimization. *arXiv preprint*, 2025.

Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and L F Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164, 2017.

Jingyu Liu, Xiaopeng Wu, Jingquan Peng, Kehan Chen, Chuan Yu, Lizhong Ding, and Yong Liu. Rethinking reward miscalibration of GRPO in agentic RL. *arXiv preprint arXiv:2509.23870*, 2025.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Yi Ma, Hongyao Tang, Dong Li, and Zhaopeng Meng. Reining generalization in offline reinforcement learning via representation distinction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Michael L Mack, Alison R Preston, and Bradley C Love. Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, 11(1):46, 2020.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning (ICML)*, 2015.

Skander Moalla, Andrea Miele, Daniil Pyatko, Razvan Pascanu, and Caglar Gulcehre. No representation, no trust: Connecting representation, collapse, and trust issues in PPO. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16828–16847. PMLR, 2022.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. *arXiv preprint arXiv:2407.10490*, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 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)*, 2012.

Yuhui Wang, Hao He, Chao Wen, and Xiaoyang Tan. Truly proximal policy optimization. *arXiv preprint arXiv:1903.07940*, 2020. Preliminary version in Uncertainty in Artificial Intelligence (UAI), 2019.

Appendix

The appendix follows the order of the main text. Appendix A derives the cross-sample coupling of Section 3.3, and Appendix B proves the identity between the participation ratio and the mean squared cosine. Appendix C solves the controlled example in closed form. Appendix D derives the Fisher information, its pullback and its properties, and Appendix E compares the three scale-channel variants, Appendix F gives the full-length curves and the ablation table, and Appendix G lists all hyperparameters.

A Cross-Sample Coupling

We derive the weight in eq. (7) and then eq. (8). Let ξ_t and $\xi_{t'}$ be independent with mean zero and covariances F_t and $F_{t'}$, which differ because σ, and hence F, depends on the state. Conditioning on ξ_t,

$$\mathbb{E}\big[\langle \xi_t, \xi_{t'} \rangle^2\big] = \mathbb{E}_{\xi_t}\big[\xi_t^\top \mathbb{E}[\xi_{t'} \xi_{t'}^\top]\, \xi_t\big] = \mathbb{E}\big[\xi_t^\top F_{t'} \xi_t\big] = \operatorname{tr}\big(F_t F_{t'}\big). \tag{15}$$

Multiplying by the feature factor of eq. (6) and summing over distinct pairs,

$$\mathbb{E}\Big[\sum_{t \neq t'} \Gamma_{tt'}^2\Big] = \sum_{t \neq t'} \operatorname{tr}(F_t F_{t'})\, \big\langle \phi_2(s_t), \phi_2(s_{t'}) \big\rangle^2, \tag{16}$$

with

$$\operatorname{tr}\big(F_t F_{t'}\big) = \sum_k \frac{1}{\sigma_k(s_t)^2\, \sigma_k(s_{t'})^2} \;+\; \sum_k w_k(s_t)\, w_k(s_{t'}), \tag{17}$$

where eq. (17) uses that F is diagonal, with $w_k = 2g'(\psi_k)^2/\sigma_k^2$ the scale entry of eq. (13). The first sum makes the confidence weighting quantitative: the contribution of a pair grows as the fourth power of the inverse scale. Only the second moment of the scores is used, so the result does not depend on the shape of their distribution, which matters because the scale-channel score $\eta^2 - 1$ is strongly skewed and bounded below by -1.

If the scale is state-independent then $F_t \equiv F$ and the weight is the constant $c = \operatorname{tr}(F^2)$. The feature factor summed over all pairs, including $t = t'$, is $\|\Phi_2 \Phi_2^\top\|_F^2 = \|\Phi_2^\top \Phi_2\|_F^2 = \sum_i s_i^4(\Phi_2)$, which by eq. (9) equals $\operatorname{tr}(\Phi_2^\top \Phi_2)^2 / \operatorname{PR}(\Phi_2)$. Removing the N diagonal terms $\|\phi_2(s_t)\|^4$ gives eq. (8). When every row has the same norm ρ, $\langle \phi_2(s_t), \phi_2(s_{t'}) \rangle^2 = \rho^4 \cos^2 \vartheta_{tt'}$, so the sum becomes $c\, \rho^4 N(N-1)\, \overline{\cos^2}$, and eq. (10) gives

$$\overline{\cos^2} = \frac{1}{N-1}\Big(\frac{N}{\operatorname{PR}(\hat{\Phi}_2)} - 1\Big). \tag{18}$$

As a check, $\operatorname{PR} = 1$ (all states aligned) gives $\overline{\cos^2} = 1$.

B The Participation Ratio and the Cosine Identity

Let $\hat{A} \in \mathbb{R}^{N \times m}$ have unit-norm rows and let $G = \hat{A}\hat{A}^\top$ be its Gram matrix, so $G_{tt'} = \cos(\hat{a}_t, \hat{a}_{t'})$ and $\operatorname{tr}(G) = N$. The squared singular values of $\hat{A}$ are the nonzero eigenvalues of G, so $\sum_i s_i^2(\hat{A}) = \operatorname{tr}(G) = N$ and $\sum_i s_i^4(\hat{A}) = \|G\|_F^2$. By eq. (9),

$$\frac{1}{\operatorname{PR}(\hat{A})} = \frac{\|G\|_F^2}{\operatorname{tr}(G)^2} = \frac{1}{N^2} \sum_{t,t'} \cos^2(\hat{a}_t, \hat{a}_{t'}) = \frac{1}{N} + \frac{N-1}{N}\, \overline{\cos^2}, \tag{19}$$

separating the N diagonal terms, each equal to one, from the $N(N-1)$ off-diagonal ones. This is eq. (10). The diagonal says nothing about how states relate; its only effect is the $1/N$ term, which is the ceiling $\operatorname{PR} \leq N$. The same algebra appears in the dimension formula of Litwin-Kumar et al. (2017, Eq. 9), written there for the covariance of neurons; applied to the Gram matrix of normalized rows, the diagonal is exactly one and the formula becomes an identity. In the MuJoCo diagnostics $\hat{A}$ is $\hat{Y}$, the centred policy outputs with each row normalized; centring is done before normalization. The implementation evaluates $\|G\|_F^2$ as $\|\hat{A}^\top \hat{A}\|_F^2$, an $m \times m$ product with the same Frobenius norm, so the whole batch can be used without forming the $N \times N$ Gram matrix.

Accuracy of eq. (11)**.** The relative error of $\mathrm{PR} \approx 1/\overline{\cos^2}$ is at most $2d/N$: 0.3% on Hopper and 1.7% on Humanoid at $N = 2048$. In the controlled example, RMS cosines of 0.41 and 0.73 give $1/\overline{\cos^2} = 5.9$ and 1.9, the values reported in Section 3.6. The approximation saves no computation, since both sides come from the same $2d \times 2d$ matrix, and it needs a full batch: on a minibatch of 64 states the $1/N$ term is not negligible.

The sign of the cosine. The square also settles a case that looks paradoxical. A cosine of -1 means two states lie on one line through the origin, pointing in opposite directions. That is a rank-one pair, so it is collapse, not diversity. In eq. (5) the size of the leakage is the same for $\cos = \pm 1$; only its direction changes, through the sign of $\hat{A}_t \hat{A}_{t'} \Gamma_{tt'}$. Two opposed states with opposite advantages push each other's ratios outward exactly as two aligned states with matching advantages do, and the one-dimensional toy of Moalla et al. (2024) exhibits both cases. Advantage signs are not determined by the feature geometry at collection time, so if they are treated as independent of it, the expected damage depends on $\cos^2$ alone. The target is $\cos = 0$, where updating one state leaves another state's ratio untouched, as in a tabular policy; this is the condition under which per-sample clipping enforces the trust region exactly.

C Exact Solution of the Controlled Example

Construction. We want a problem whose answer we know in advance and in which the only thing that changes is how the feature variance is spread. A one-step contextual bandit gives us this: no bootstrapping, so a difference in learning speed cannot come from critic error; no episodes; fixed features and trained heads, so the spectrum is chosen rather than learned; and a closed-form optimum (Appendix C). Each round draws a hidden state $x \sim \mathcal{N}(0, I_D)$ with $D = 6$, the agent chooses $a \in \mathbb{R}^D$, and the reward is $-\|a - x\|^2/D$. The agent sees only

$$\phi_2 = V \Lambda^{1/2} x, \tag{20}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ sets how evenly the variance is spread and V is a random rotation, which leaves rank and participation ratio unchanged and keeps weak directions off the coordinate axes where Adam would compensate for them. Every condition carries the same information and optimal policy, has total variance D, $\mathcal{R} = \mathbb{R}^D$, and a state-independent scale, so the weight in eq. (8) is constant; we report the spectrum of the fixed Φ_2, since Y moves with the head. The spectrum is swept from uniform to concentrated,

$$\Lambda(\kappa) \;\propto\; (1-\kappa)\,\tfrac{1}{D}\mathbf{1} \;+\; \kappa\,(0.92,\, 0.016,\, \ldots,\, 0.016), \qquad \kappa \in \{0, 0.25, 0.5, 0.75, 1\}, \tag{21}$$

with $\kappa = 1$ the spectrum of Section 3.2, and the heads are trained with ordinary PPO (Appendix G.1).

With features $\phi_2 = V\Lambda^{1/2}x$, a linear mean head $\mu = \Theta\phi_2$, a state-independent scale σ, and reward $r = -\frac{1}{D}\|a - x\|^2$ with $a \sim \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$, the expected return is

$$\begin{aligned} J(\Theta, \sigma) &= -\frac{1}{D}\,\mathbb{E}_x\Big[\big\|\Theta V \Lambda^{1/2} x - x\big\|^2\Big] - \frac{1}{D}\sum_k \sigma_k^2 \\ &= -\frac{1}{D}\big\|\Theta V \Lambda^{1/2} - I\big\|_F^2 - \frac{1}{D}\sum_k \sigma_k^2, \end{aligned} \tag{22}$$

using $\mathbb{E}[xx^\top] = I$. The optimum is $\Theta^\star = \Lambda^{-1/2} V^\top$ with $J^\star = 0$ in the limit $\sigma \to 0$, so every condition shares the same optimal return and the same optimal policy, and the conditions differ only in how quickly they reach it. The first term of eq. (22) is quadratic in Θ with Hessian proportional to $\mathbb{E}[\phi_2\phi_2^\top] = V\Lambda V^\top$. Curvature along feature direction i is therefore proportional to λ_i, and the condition number of the problem is

$$\varkappa = \frac{\max_i \lambda_i}{\min_i \lambda_i}, \tag{23}$$

which equals 1 for the uniform spectrum and $0.92/0.016 = 57.5$ for the concentrated one, at identical $\mathrm{rank}_{0.01} = 6$ and identical total variance $\sum_i \lambda_i = D$. Because V is a random rotation, the ill-conditioned directions are not aligned with the coordinate axes, so the per-coordinate preconditioning performed by Adam cannot remove them. This is the exact sense in which the spectrum, and not the rank, determines the difficulty of the problem.

D GAUSSIAN FISHER INFORMATION AND THE PULLBACK

Scores and Fisher information. For a diagonal Gaussian with mean μ_k and scale σ_k, writing $\eta_k = (u_k - \mu_k)/\sigma_k$ for the standardized latent action,

$$\frac{\partial \log \pi}{\partial \mu_k} = \frac{\eta_k}{\sigma_k}, \qquad \frac{\partial \log \pi}{\partial \sigma_k} = \frac{\eta_k^2 - 1}{\sigma_k}. \tag{24}$$

At collection time the action was drawn from this Gaussian, so η_k is a standard normal draw; this is a property of the sampling step, not an assumption. Using $\mathbb{E}[\eta_k] = 0$, $\mathbb{E}[\eta_k^2] = 1$, $\mathbb{E}[\eta_k^3] = 0$ and $\mathbb{E}[\eta_k^4] = 3$,

$$\mathbb{E}\Big[\Big(\frac{\eta_k}{\sigma_k}\Big)^2\Big] = \frac{1}{\sigma_k^2}, \qquad \mathbb{E}\Big[\Big(\frac{\eta_k^2 - 1}{\sigma_k}\Big)^2\Big] = \frac{2}{\sigma_k^2}, \qquad \mathbb{E}\Big[\frac{\eta_k}{\sigma_k} \cdot \frac{\eta_k^2 - 1}{\sigma_k}\Big] = 0, \tag{25}$$

so F is diagonal, and the vanishing cross term is what makes the mean and scale channels separable. The heads emit ψ with $\sigma = g(\psi)$, and since the Fisher transforms as a metric under a change of coordinates, the scale entry becomes $g'(\psi_k)^2 \cdot 2/\sigma_k^2$, which is eq. (13).

Uniqueness (Proposition 1). Let $K(\delta) = \mathrm{KL}(\pi_y \,\|\, \pi_{y+\delta})$. Then $K(0) = 0$ and $K(\delta) \geq 0$ for all δ, so $\delta = 0$ is a global minimum and $\nabla K(0) = 0$. Taylor's theorem gives $K(\delta) = \frac{1}{2}\delta^\top \nabla^2 K(0)\delta + o(\|\delta\|^2)$. If a symmetric M satisfies $\frac{1}{2}\delta^\top M\delta = K(\delta) + o(\|\delta\|^2)$ then $\delta^\top(M - \nabla^2 K(0))\delta = o(\|\delta\|^2)$ for all δ; taking $\delta = tv$ and letting $t \to 0$ gives $v^\top(M - \nabla^2 K(0))v = 0$ for every v, and a symmetric matrix with vanishing quadratic form is zero. Hence $M = \nabla^2 K(0) = F$.

Pullback and kernel. Write $M_F = \theta_{\mathrm{comb}}^\top F \theta_{\mathrm{comb}}$. With $y = \theta_{\mathrm{comb}}\phi_2$ linear in ϕ_2, the second-order form transports as $\delta\phi_2^\top M_F\, \delta\phi_2$, giving eq. (14). For $v \in \mathbb{R}^D$, $v^\top M_F v = (\theta_{\mathrm{comb}} v)^\top F(\theta_{\mathrm{comb}} v)$, which vanishes if and only if $\theta_{\mathrm{comb}} v = 0$ because $F \succ 0$; hence $\ker M_F = \ker \theta_{\mathrm{comb}} = \mathcal{R}^\perp$. Movement along $\mathcal{R}^\perp$ leaves μ and ψ unchanged and costs nothing, while $M = I$ charges it in full.

Properties of the pullback. On the mean block the weighting is exact rather than approximate: the term of the Gaussian divergence involving the means is $(\mu - \mu')^2/2\sigma'^2$, so $1/\sigma_k^2$ is that term itself. The penalty is also a property of the policy rather than of its parameterization. Replacing ϕ_2 by $A\phi_2$ and θ_{comb} by $\theta_{\mathrm{comb}} A^{-1}$, for any invertible A, leaves the policy and every trajectory unchanged; under it the drift becomes $A\delta$ and the metric $A^{-\top} M_F A^{-1}$, so $\delta^\top M_F \delta$ is unchanged while $\|\delta\|^2$ is not. A uniform penalty therefore assigns different values to two descriptions of the same policy.

The activation Jacobian. The exact second-order metric for pre-activation drift is $J_f M_F J_f$ with $J_f = \mathrm{diag}(f'(z_2))$, since $\delta\phi_2 = J_f\, \delta z_2 + o(\|\delta z_2\|)$. For $\tanh$, $f'(z_{2,i}) = 1 - \phi_{2,i}^2 \in (0, 1]$, so the exact form down-weights saturated coordinates relative to eq. (14). Whenever no unit is exactly saturated, J_f is invertible, and the kernel of the corrected metric, $J_f^{-1} \ker M_F$, has the same dimension as $\ker M_F$.

E THE THREE SCALE-CHANNEL VARIANTS

The mean block of eq. (13) does not depend on the scale map g; the scale block does, and it is clearest written as

$$\frac{2\, g'(\psi_k)^2}{\sigma_k^2} \;=\; 2\Big(\frac{\partial \log \sigma_k}{\partial \psi_k}\Big)^2, \tag{26}$$

twice the squared sensitivity of $\log \sigma$ to the coordinate the penalty acts on. For an exponential head, $\sigma = \exp(\psi)$, the sensitivity is one at every state and the weight is exactly 2: the Fisher with respect to $\log \sigma$. That constant has a meaning. The divergence between two Gaussians with the same mean depends only on the ratio of their scales, so doubling σ costs the same from 0.1 to 0.2 as from 1 to 2, and in a logarithmic coordinate the metric is flat. For our softplus head, using

$\mathrm{sigmoid}(x) = 1 - e^{-\,\mathrm{softplus}(x)}$, the derivative is $g'(\psi) = 1 - e^{-(\sigma - \sigma_{\min})}$, and the exact weight is a function of σ alone,

$$w(\sigma) = \frac{2\big(1 - e^{-(\sigma - \sigma_{\min})}\big)^2}{\sigma^2}. \tag{27}$$

FF-PPO-log uses the constant 2, FF-PPO-exact uses $w(\sigma)$, and FF-PPO-mean sets the block to zero, which halves $\dim \mathcal{R}$ from $2d$ to d.

Table 2: Scale-channel Fisher weight for the softplus head, eq. (27), against the constant 2 of a log-standard-deviation head. The ratio is the factor by which FF-PPO-log over-penalizes the scale channel.

σ	0.05	0.1	0.2	0.5	1.0	2.0	3.0
exact weight $w(\sigma)$	1.23	1.48	1.50	1.20	0.79	0.37	0.20
constant weight (log-std head)	2.00	2.00	2.00	2.00	2.00	2.00	2.00
ratio	1.63	1.35	1.34	1.67	2.53	5.37	9.98

The constant is a strict upper bound. For the softplus head the sensitivity in eq. (26) is $\mathrm{sigmoid}(x)/(\sigma_{\min} + \mathrm{softplus}(x))$ with $x = \psi + c$. Let $q(x) = \mathrm{softplus}(x) - \mathrm{sigmoid}(x)$. Then $q'(x) = \mathrm{sigmoid}(x)^2 > 0$ and $q(x) \to 0$ as $x \to -\infty$, so $q > 0$ everywhere and $\mathrm{sigmoid}(x) < \mathrm{softplus}(x) \le \sigma$. The sensitivity is therefore strictly below one and the weight strictly below 2 at every state. Numerically its supremum is 1.52 for $\sigma_{\min} = 0.01$, so the constant over-penalizes by at least a factor 1.3 everywhere. Table 2 shows that the ratio is not constant: it is smallest near $\sigma \approx 0.2$ and grows in both directions, so no choice of β reproduces the exact weighting, and since σ varies across the batch and shrinks over training, the distortion drifts in both dimensions.

Behaviour at the floor. As $\sigma \to \sigma_{\min}$ the softplus saturates, the sensitivity goes to zero and $w(\sigma)$ follows it: at $\sigma = 0.011$ the weight is 0.017 against the constant 2. This is correct about the divergence, since moving ψ then barely moves σ, but it means the exact metric stops penalizing scale drift exactly where the policy is sharpest, which is where a small change in σ does most to the divergence. The constant has the opposite profile: wrong everywhere, but uniform in its protection. Which trade matters is empirical, which is why all three variants are run.

F Additional Results

Ablations and the coefficient in detail. On Humanoid neither half of the Fisher metric reaches the full pullback: the null space alone ($M = \theta_{\mathrm{comb}}^{\top}\theta_{\mathrm{comb}}$) gives 1016, a paired difference of 901 [112, 1555], and confidence weighting alone ($M = \bar{w}(s)\, I$) gives 1251, 667 [−18, 1292]; on Ant the metric choice does not matter (Table 4). The coefficient $\beta = 1$ is untuned for both penalties: on Ant PFO peaks at $\beta = 0.3$ and FF-PPO-log at 0.1, and both degrade at $\beta = 3$. PPO with LayerNorm is the best method on HalfCheetah and Swimmer and no worse than PPO elsewhere; trunk resets hurt on every task.

G Hyperparameters

G.1 Controlled example

Table 6 lists the settings for the controlled example of Section 3.6. The five spectra of eq. (21) share every setting; only $\Lambda(\kappa)$ changes.

G.2 MuJoCo

Table 7 lists the hyperparameters used in all MuJoCo experiments. No hyperparameter is tuned per environment.

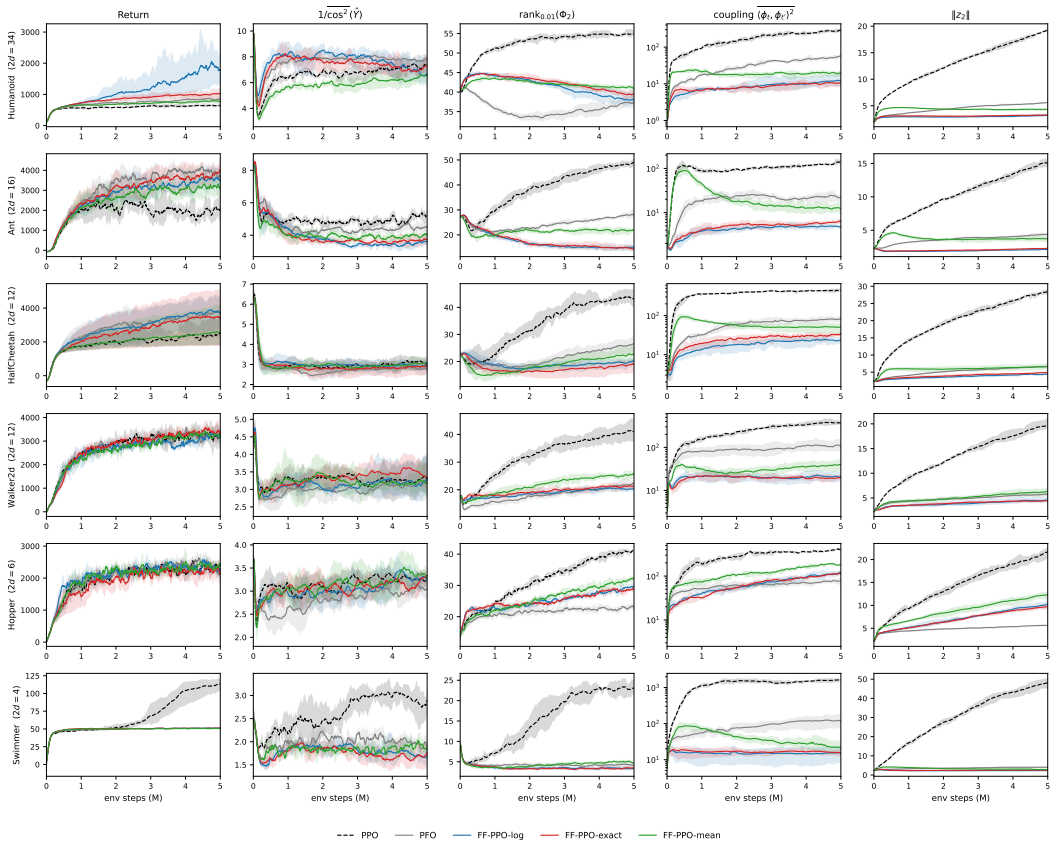

Figure 4: Figure 3 over the full five million steps.

Table 3: Mechanism. Return, cross-state coupling of eq. (8), trust-region excess ratio and feature norm, each the median over ten seeds of the last 5% of training. Within every setting where the penalties help, the methods order by coupling as they order by excess ratio; Swimmer is the counterexample discussed in Section 5.

Setting	Method	Return	Coupling	Excess	$\|\phi_2\|$
Humanoid	PPO	634	275	6.12	6.7
	PFO	850	56	4.07	4.1
	FF-PPO-log	1735	10	2.51	2.5
Ant, 10 epochs	PPO	2176	131	2.95	5.7
	PFO	4141	23	2.34	3.2
	FF-PPO-log	3583	4.9	2.10	1.8
Ant, 20 epochs	PPO	85	261	12.17	7.1
	PFO	2388	40	3.25	4.1
	FF-PPO-log	3711	7.5	2.47	2.3
Slippery Ant	PPO	719	144	3.43	6.1
	PFO	3953	33	2.45	3.6
	FF-PPO-log	3994	8.1	2.23	2.0
	Detached KL	3339	1246	2.27	6.4
HalfCheetah	PPO	2184	446	2.95	7.1
	PFO	4158	81	2.36	4.4
	FF-PPO-log	3939	22	2.26	3.1
Swimmer	PPO	116	1643	2.42	7.6
	PFO	51	120	2.05	3.4
	FF-PPO-log	51	17	1.97	2.1

Table 4: Metric ablation and coefficient sweep (feature norms $\|\phi_2\|$ at the end of training: PPO 5.7 to 7.6, PFO 3.2 to 4.4, FF-PPO-log 1.8 to 4.9 across tasks). Final return, IQM [95% bootstrap CI] over ten seeds. The paired difference is the IQM difference to FF-PPO-log over matched seeds, with its 95% bootstrap CI.

	Ant	Humanoid	paired diff. (Humanoid)
FF-PPO-log, $M = \theta_{\text{comb}}^\top F \theta_{\text{comb}}$	3563 [3044, 4060]	1918 [1315, 2452]	–
FF-PPO-exact	3857 [3457, 4307]	1020 [866, 1635]	–
$M = \theta_{\text{comb}}^\top \theta_{\text{comb}}$ (null space only)	3716 [2973, 4115]	1016 [757, 1585]	901 [112, 1555]
$M = \bar{w}(s)\, I$ (confidence only)	3629 [3097, 3939]	1251 [1074, 1705]	667 [−18, 1292]
Jacobian-corrected	2976 [2373, 3823]	1156 [863, 1736]	–
PFO, $M = I$	3788 [3118, 4148]	833 [714, 1092]	1085 [496, 1542]
Detached KL	3003 [2708, 3294]	858 [763, 1064]	–
	Ant: PFO / FF-PPO-log at β		
	$\beta = 0.1$	$\beta = 0.3$	$\beta = 3$
PFO	3200 [2624, 3851]	3881 [3432, 4140]	3491 [2892, 3857]
FF-PPO-log	3822 [3474, 4095]	3804 [3492, 4151]	2924 [2695, 3273]

Table 5: Drop from peak to final return (IQM over ten seeds), lower is better.

	Humanoid	Ant	HalfCheetah	Walker2d	Hopper	Swimmer
PPO	66	1214	237	628	626	2
PFO	81	733	167	561	708	1
FF-PPO-log	418	368	106	415	733	1
FF-PPO-exact	106	366	145	493	570	1
FF-PPO-mean	69	434	60	461	656	1
Detached KL	200	1040	169	–	702	13

Table 6: Settings for the controlled example (Section 3.6).

Hyperparameter	**Value**
Task	
Hidden state and action dim. (D)	6
Reward	$-\|a - x\|^2/D$
Spectra (κ)	$\{0, 0.25, 0.5, 0.75, 1\}$
Total variance $\sum_i \lambda_i$	D
Rotation V	random orthogonal; one per seed, same for all κ
Policy	
Features	fixed, $\phi_2 = V\Lambda^{1/2}x$ (not trained)
Mean head	linear with bias, trained
Scale	learned, state-independent ($\sigma = 1$ at start)
PPO	
Batch size per iteration	512
Iterations	100
Update epochs	10
Minibatch size	64
Clip coefficient (ϵ)	0.2
Baseline	batch-mean reward (no critic)
Advantage normalization	per minibatch
Optimizer	Adam, 3×10^{-4}
Gradient norm clip	0.5
Seeds per setting	5

Table 7: Hyperparameters used in all MuJoCo experiments.

Hyperparameter	**Value**
Environment	
Parallel environments (N_{env})	2
Episode truncation length	1000
Observation normalization	Yes (frozen after warmup)
Normalization warmup samples	4000
Time-aware observation	Yes
Network architecture	
Hidden layers ($D_1 \to D$)	$64 \to 64$
Activation f	$\tanh$
Actor–critic sharing	None (separate trunks)
Action distribution	TanhNormal (state-dependent μ, σ)
Scale map g	$\sigma_{\min} + \text{softplus}(\psi + c)$, $\sigma_{\min} = 0.01$
Weight initialization	PyTorch default (Kaiming uniform)
PPO	
Total timesteps	5×10^6
Rollout length (T)	1024
Batch size ($B = N_{\text{env}} \times T$)	2048
Number of minibatches	32
Minibatch size (M)	64
Update epochs (K)	10
Discount (γ)	0.99
GAE λ	0.95
Clip coefficient (ϵ)	0.2
Value function coefficient (c_{VF})	0.5
Entropy coefficient (c_H)	0.0
Learning rate (Adam)	3×10^{-4}
Gradient norm clip	0.5
Advantage normalization	Per-minibatch
Penalty and diagnostics	
Penalty coefficient β	1 (all penalized methods)
Diagnostic batch	full rollout batch ($B = 2048$), pre-update
Diagnostic spaces	Z_2, Φ_2, Y; centred and uncentred
Diagnostic precision	float64; identity checked at iteration 1