WONI: WEIGHT-SPACE ORBIT STRUCTURED NATURAL GRADIENT FOR SCALABLE POLICY OPTIMIZATION VIA INVERSE-FISHER PRECONDITIONER

Anonymous authors
Paper under double-blind review

ABSTRACT

In deep RL with multi-objective and constrained settings, the natural gradient provides a principled way to handle the correlations among gradients within the policy geometry, while its computation, which grows with the number of gradients, remains a bottleneck. In this paper, we study curvature approximations for inverting the Fisher information matrix, which can be applied to a set of gradients to obtain multiple natural gradients efficiently. We propose a curvature approximation structured by the weight-space symmetry of the policy network and calibrated with second moments estimated from samples. Empirical results show that applying the proposed method in place of the conjugate gradient method used in various second-order algorithms achieves comparable or better performance with substantially less computation.

1 INTRODUCTION

Over the past decade, efforts to deploy deep reinforcement learning (RL) beyond simulated environments into real-world applications have increased substantially (Dulac-Arnold et al., 2021). However, a single reward function often struggles to capture the requirements of real-world decision-making tasks, which involve multiple objectives and constraints. For example, in autonomous driving, a vehicle may consider multiple objectives, including passenger comfort, energy efficiency, and travel time, beyond reaching the destination, while satisfying constraints on collision avoidance, speed limits, and various traffic rules (Kiran et al., 2022). To address this gap, research has expanded into multi-objective RL (MORL) (Hayes et al., 2022), constrained/safe RL (Achiam et al., 2017), and constrained multi-objective RL (CMORL) (Kim et al., 2025).

Difficulties arise from interactions among multiple gradients, as conflicting or aligned components can cancel or amplify each other, while mismatched scales let a single direction dominate the update (Yu et al., 2020; Liu et al., 2021). Accordingly, linear scalarization in MORL can lead to convergence to local optima and limit the search for Pareto-optimal policies (Roijers et al., 2013; Hayes et al., 2022), whereas Lagrangian methods in constrained settings can amplify gradient interactions, leading to frequent constraint violations and difficult recovery from infeasible regions (Yao et al., 2024). Similar issues have been addressed in multi-task learning (MTL) through Pareto optimization (Sener & Koltun, 2018), conflict projection (Yu et al., 2020), and gradient balancing (Liu et al., 2021), offering promising approaches for RL. In principle, these approaches should be considered under the geometry of the policy distribution with *natural gradients*, the steepest directions under the metric given by the Fisher information matrix (FIM) (ichi Amari, 1998; Kakade, 2002). Handling gradient interactions in Euclidean parameter space instead introduces a geometric mismatch that can lead to sensitivity to high curvature and reliance on hyperparameter tuning. Therefore, second-order methods based on natural gradients can handle multiple gradients with respect to the policy geometry (Kim et al., 2023; 2025), but obtaining natural gradients typically relies on Hessian-free conjugate gradient (HF) (Martens, 2010; Schulman et al., 2015), which must be applied to each gradient, leaving a computational bottleneck that scales with the number of objectives and constraints.

We focus on a property of the FIM induced by the policy, which is defined as the expected outer product of the score function (Kakade, 2002; Martens, 2020) and is therefore independent of any objective or constraint, allowing a single inverse to precondition all gradients. Accordingly, methods

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.

that directly approximate the FIM can be leveraged, including Kronecker-Factored Approximate Curvature (K-FAC) and its variants (Martens & Grosse, 2015; George et al., 2018; Gao et al., 2021; 2023). In the RL literature, however, curvature approximations remain largely unexplored in multi-gradient settings, while existing studies focus on convergence speed and stability in single-objective RL (Wu et al., 2017; Song & Wu, 2018; Gebotys et al., 2022; Huo et al., 2026).

In this paper, we propose a **W**eight-space **O**rbit-structured **N**atural gradient approximation with an **I**nverse FIM preconditioner (WONI) to make curvature-aware policy updates practical for increasing numbers of objectives and constraints. Inspired by the curvature constraints derived from the weight-space symmetry analyzed in recent work (Artemev et al., 2026), our method approximates the FIM with a symmetry-structured form and rescales the curvature in its eigenbasis. Our contributions are

- We propose WONI, a weight-space orbit-structured curvature approximation for the policy FIM that replaces the isotropic structure induced by orbit averaging with weighted second moments, yielding an efficiently invertible inverse-Fisher preconditioner.
- We conduct extensive experiments on curvature approximations that can be shared across multiple gradients, including K-FAC and its variants, and systematically evaluate their performance across different actor architectures and configurations.
- We evaluate WONI against HF, K-FAC variants, and Euclidean updates in single-objective, multi-constrained, and multi-objective RL. The empirical results show that WONI achieves competitive policy performance across the evaluated settings while substantially reducing the cost of natural-gradient computation relative to HF in constrained multi-objective tasks.

2 BACKGROUND

Notation. Throughout, we consider a stochastic policy π_θ parameterized by an L-layer neural network, where θ denotes the vector of all network parameters. The parameters consist of the weight matrices $\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and biases $\boldsymbol{b}_\ell \in \mathbb{R}^{d_\ell}$ for $\ell = 1, \dots, L$, where d_ℓ is the number of nodes in layer ℓ. The input layer is indexed as layer 0 with dimension d_0. The activation function and layer input are denoted by ϕ and $\boldsymbol{a}_\ell = \phi(\boldsymbol{z}_{\ell-1})$, respectively, where $\boldsymbol{z}_\ell = \boldsymbol{W}_\ell \boldsymbol{a}_\ell + \boldsymbol{b}_\ell$ denotes the pre-activation. A constant input is used for the bias.

2.1 MDP WITH MULTIPLE OBJECTIVES AND CONSTRAINTS

Standard RL is formulated on a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$ with a state space $\mathcal{S}$ of states s, an action space $\mathcal{A}$ of actions a, a transition kernel P, a reward function r, an initial state distribution ρ, and a discount factor $\gamma \in [0, 1)$. With the extended MDP $(\mathcal{S}, \mathcal{A}, P, \{r_n\}_{n=1}^N, \{c_m\}_{m=1}^M, \rho, \gamma)$, MORL considers N reward functions, safe RL introduces M cost functions with corresponding thresholds b_m, and CMORL combines both (Roijers et al., 2013; Altman, 1999; Kim et al., 2025). For trajectories τ generated by π_θ from $s_0 \sim \rho$, the n-th reward return follows $J_{r_n}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[\sum_{t=0}^{\infty} \gamma^t r_n(s_t, a_t)]$, and the cost return $J_{c_m}(\theta)$ is defined analogously. Depending on the MDP variants, the resulting problem is

$$\max_\theta \ \big(J_{r_1}(\theta), \dots, J_{r_N}(\theta)\big) \quad \text{s.t.} \quad J_{c_m}(\theta) \le b_m \quad \forall m. \tag{1}$$

For $N > 1$, maximization is understood in the Pareto sense over feasible policies. Preference weights or min-max formulations can be used to select among the resulting trade-offs.

2.2 NATURAL GRADIENT AND THE FISHER INFORMATION MATRIX

For brevity, we drop the subscripts r_n and c_m and write $J(\theta)$ for the return of a reward or a cost, with $A(s, a)$ denoting its corresponding advantage function. With $A(s, a)$ estimated by the corresponding critic, the gradient of $J(\theta)$ is given by $\nabla_\theta J(\theta) = \mathbb{E}_{s,\, a \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a \mid s)\, A(s, a)]$. Since the policy π_θ is the action distribution parameterized by θ, the change of this policy should be measured by a distance between distributions, while this gradient is typically computed under the Euclidean distance on the parameter space and therefore does not reflect the change of the policy clearly. Instead, the FIM defines the Riemannian metric on π_θ that measures this change locally, defined as

$$\boldsymbol{F} = \mathbb{E}_{s,\, a \sim \pi_\theta}\left[\nabla_\theta \log \pi_\theta(a \mid s)\, \nabla_\theta \log \pi_\theta(a \mid s)^\top\right]. \tag{2}$$

The FIM is an invariant metric on π regardless of the choice of the parameterization, representing the sensitivity of the policy to a change of each parameter. Therefore, applying the inverse of the FIM to a gradient equalizes the influence of each parameter coordinate on the change of the policy, which gives the natural gradient $\boldsymbol{F}^{-1}\nabla_\theta J(\theta)$, the steepest direction under this metric (ichi Amari, 1998; Kakade, 2002). Accordingly, the interaction between natural gradients $\tilde{\boldsymbol{g}}_1$ and $\tilde{\boldsymbol{g}}_2$ should be measured under the Fisher metric, with $\langle \tilde{\boldsymbol{g}}_1, \tilde{\boldsymbol{g}}_2 \rangle_{\boldsymbol{F}} = \tilde{\boldsymbol{g}}_1^\top \boldsymbol{F} \tilde{\boldsymbol{g}}_2$, $\|\tilde{\boldsymbol{g}}_1\|_{\boldsymbol{F}} = \sqrt{\tilde{\boldsymbol{g}}_1^\top \boldsymbol{F} \tilde{\boldsymbol{g}}_1}$, and $\cos \angle_{\boldsymbol{F}}(\tilde{\boldsymbol{g}}_1, \tilde{\boldsymbol{g}}_2) = \langle \tilde{\boldsymbol{g}}_1, \tilde{\boldsymbol{g}}_2 \rangle_{\boldsymbol{F}} / (\|\tilde{\boldsymbol{g}}_1\|_{\boldsymbol{F}} \|\tilde{\boldsymbol{g}}_2\|_{\boldsymbol{F}})$. These quantities can differ substantially from their Euclidean counterparts depending on $\boldsymbol{F}$.

For convenience, we hereafter write the natural gradients of the n-th reward and m-th cost returns as $\tilde{\boldsymbol{g}}_{r_n} \coloneqq \boldsymbol{F}^{-1}\nabla_\theta J_{r_n}(\theta)$ and $\tilde{\boldsymbol{g}}_{c_m} \coloneqq \boldsymbol{F}^{-1}\nabla_\theta J_{c_m}(\theta)$. The score function is written as $\boldsymbol{g}_s \coloneqq \nabla_\theta \log \pi_\theta(a \mid s)$.

2.3 Kronecker-factored approximate curvature

Since explicitly forming and inverting the FIM is infeasible for deep neural networks, various approximations have been proposed in the DNN literature, among which K-FAC and its variants estimate the FIM layer-wise using Kronecker products (Martens & Grosse, 2015). This Kronecker-factored approximation also applies to the policy $\pi_\theta(a \mid s)$, a conditional distribution over actions, as demonstrated by ACKTR (Wu et al., 2017). As shown in equation 2, the FIM is constructed from the score function $\boldsymbol{g}_s$, which is obtained by backpropagation. The score function with respect to the weight matrix $\boldsymbol{W}_\ell$ can be written as

$$\frac{\partial \log \pi_\theta(a \mid s)}{\partial \boldsymbol{W}_\ell} = \frac{\partial \log \pi_\theta(a \mid s)}{\partial \boldsymbol{z}_L} \frac{\partial \boldsymbol{z}_L}{\partial \boldsymbol{z}_{L-1}} \cdots \frac{\partial \boldsymbol{z}_{\ell+1}}{\partial \boldsymbol{z}_\ell} \frac{\partial \boldsymbol{z}_\ell}{\partial \boldsymbol{W}_\ell} = \boldsymbol{\delta}_\ell \boldsymbol{a}_\ell^\top. \tag{3}$$

Here $\boldsymbol{\delta}_\ell = \partial \log \pi_\theta(a \mid s) / \partial \boldsymbol{z}_\ell$ denotes the gradient of $\log \pi_\theta(a \mid s)$ with respect to $\boldsymbol{z}_\ell$. K-FAC first applies a block-diagonal approximation $\boldsymbol{F} \approx \operatorname{diag}(\boldsymbol{F}_1, \ldots, \boldsymbol{F}_L)$, where each block is expressed as $\mathbb{E}[\boldsymbol{a}_\ell \boldsymbol{a}_\ell^\top \otimes \boldsymbol{\delta}_\ell \boldsymbol{\delta}_\ell^\top]$. Assuming statistical independence between $\boldsymbol{a}_\ell \boldsymbol{a}_\ell^\top$ and $\boldsymbol{\delta}_\ell \boldsymbol{\delta}_\ell^\top$, K-FAC further approximates each block as

$$\boldsymbol{F}_\ell \approx \mathbb{E}\left[\boldsymbol{a}_\ell \boldsymbol{a}_\ell^\top\right] \otimes \mathbb{E}\left[\boldsymbol{\delta}_\ell \boldsymbol{\delta}_\ell^\top\right]. \tag{4}$$

2.4 Weight-space symmetries

A weight-space symmetry is a transformation of a network's parameters that leaves its function unchanged (Zhao et al., 2025). Permuting hidden nodes is a representative example. Reordering the nodes of layer ℓ by a permutation matrix $\boldsymbol{T}$ transforms the adjacent $\boldsymbol{W}_\ell$, $\boldsymbol{b}_\ell$, and $\boldsymbol{W}_{\ell+1}$ as

$$\boldsymbol{W}_\ell \mapsto \boldsymbol{T}\boldsymbol{W}_\ell, \qquad \boldsymbol{b}_\ell \mapsto \boldsymbol{T}\boldsymbol{b}_\ell, \qquad \boldsymbol{W}_{\ell+1} \mapsto \boldsymbol{W}_{\ell+1}\boldsymbol{T}^\top. \tag{5}$$

The network output remains unchanged because the next layer cancels the permutation of $\boldsymbol{a}_{\ell+1}$ through $\boldsymbol{T}^\top \boldsymbol{T} = \boldsymbol{I}$. More generally, this symmetry extends to the class of invertible transformations admitted by the activation function.

A differentiable loss $\mathcal{L}$ that depends on the parameters through the network output inherits the same symmetry, $\mathcal{L}(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell, \boldsymbol{W}_{\ell+1}) = \mathcal{L}(\boldsymbol{T}\boldsymbol{W}_\ell, \boldsymbol{T}\boldsymbol{b}_\ell, \boldsymbol{W}_{\ell+1}\boldsymbol{T}^\top)$. Hence the gradients transform as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_\ell} \mapsto \boldsymbol{T} \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_\ell}, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_\ell} \mapsto \boldsymbol{T} \frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_\ell}, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{\ell+1}} \mapsto \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{\ell+1}} \boldsymbol{T}^\top. \tag{6}$$

Recent studies extended this property to curvature approximation based on the weight-space orbit (Bernacchia, 2025; Artemev et al., 2026). For a finite group $\mathcal{G} = \{\boldsymbol{T} \mid \pi_{\boldsymbol{T}\theta} = \pi_\theta\}$ of orthogonal parameter transformations, let $\theta^\star = |\mathcal{G}|^{-1} \sum_{\boldsymbol{T} \in \mathcal{G}} \boldsymbol{T}\theta$ denote the orbit-averaged parameters. Assuming that the loss is twice differentiable, they apply a second-order Taylor approximation around $\theta^\star$, which yields the secant condition $\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^\star) \approx \nabla^2\mathcal{L}(\theta^\star)(\theta - \theta^\star)$. Averaging the outer products of this relation over the orbit yields a matrix relation between the parameter and gradient variations.

By Lemma 3.1 of Artemev et al. (2026), the orbit-averaged outer product of these gradient differences is invariant under every group transformation and belongs to the commutant of $\mathcal{G}$, the space

of matrices that commute with all group elements. For the groups considered, this matrix can be expressed using a small set of group-dependent basis elements and their coefficients. Under an isotropic approximation of the parameter variations, this matrix approaches the square of $\nabla^2\mathcal{L}(\theta^\star)$. Its positive semidefinite square root therefore approximates $\nabla^2\mathcal{L}(\theta^\star)$, hence its inverse serves as the preconditioner.

3 Proposed Method

3.1 Orbit-averaged FIM and its structure

Unlike the Hessian, which requires second-order derivatives of the loss, the FIM of π_θ can be directly obtained with the score function, therefore our method starts with applying the weight-space symmetry to the score function $\boldsymbol{g}_s$. For an orthogonal symmetry transformation $\boldsymbol{T} \in \mathcal{G}$, the score function satisfies $\boldsymbol{g}_s(\boldsymbol{T}\theta) = \boldsymbol{T}\boldsymbol{g}_s(\theta)$, and hence the FIM at the transformed parameters satisfies $\boldsymbol{F}(\boldsymbol{T}\theta) = \boldsymbol{T}\boldsymbol{F}\boldsymbol{T}^\top$. Averaging these symmetry-equivalent FIMs over the group gives the orbit-averaged FIM

$$\boldsymbol{F}^\star = \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{T}\in\mathcal{G}} \boldsymbol{T}\boldsymbol{F}\boldsymbol{T}^\top. \tag{7}$$

Lemma 1. *The orbit-averaged FIM $\boldsymbol{F}^\star$ is the orthogonal projection of $\boldsymbol{F}$ onto the invariant matrix subspace under the Frobenius inner product.*

Lemma 2. *The orbit-averaged FIM $\boldsymbol{F}^\star$ is symmetric positive semidefinite whenever the FIM is.*

Proofs are in Appendix A.1 and A.2, respectively.

Lemma 1 implies that $\boldsymbol{F}^\star$ is invariant under every $\boldsymbol{T} \in \mathcal{G}$. Pointwise activations allow $\boldsymbol{T}$ to permute hidden nodes, while odd activations additionally allow sign flips. These symmetries determine $\mathcal{G}$ as the *permutation group* or *signed permutation group*, which fixes the relations among the entries of $\boldsymbol{F}^\star$ and its degrees of freedom.

Following the invariant forms derived in Appendix D.5 of Artemev et al. (2026), we express $\boldsymbol{F}^\star$ for an L-layer network in terms of the blocks $\boldsymbol{F}_{\ell\ell'} = \mathbb{E}[\boldsymbol{a}_\ell\boldsymbol{a}_{\ell'}^\top \otimes \boldsymbol{\delta}_\ell\boldsymbol{\delta}_{\ell'}^\top]$, where $\boldsymbol{F}_{\ell\ell}$ coincides with $\boldsymbol{F}_\ell$ in equation 4. Since the activation is set for each layer separately, $\boldsymbol{T}$ and $\mathcal{G}$ can be written layer-wise as $\boldsymbol{T}_\ell \in \mathcal{G}_\ell$, which map the score function of $\boldsymbol{W}_\ell$ to $\boldsymbol{T}_\ell\boldsymbol{\delta}_\ell\boldsymbol{a}_\ell^\top\boldsymbol{T}_{\ell-1}^\top$. We focus on the diagonal blocks, where $\boldsymbol{T}_{\ell-1}$ acts only on $\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top$ and $\boldsymbol{T}_\ell$ only on $\boldsymbol{\delta}_\ell\boldsymbol{\delta}_\ell^\top$. The group average therefore separates over the nodes of the two layers,

$$\boldsymbol{F}^\star_{\ell\ell} = \mathbb{E}\left[\frac{1}{|\mathcal{G}_{\ell-1}|}\sum_{\boldsymbol{T}_{\ell-1}\in\mathcal{G}_{\ell-1}} \boldsymbol{T}_{\ell-1}\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top\boldsymbol{T}_{\ell-1}^\top \otimes \frac{1}{|\mathcal{G}_\ell|}\sum_{\boldsymbol{T}_\ell\in\mathcal{G}_\ell} \boldsymbol{T}_\ell\boldsymbol{\delta}_\ell\boldsymbol{\delta}_\ell^\top\boldsymbol{T}_\ell^\top\right], \tag{8}$$

where $\mathcal{G}_0$ and $\mathcal{G}_L$ contain only the identity because the input and output nodes carry no symmetry. We examine the resulting diagonal block structure for each of the two groups below, with the detailed calculations in Appendix B.1 and the off-diagonal blocks in Appendix B.2.

Signed permutation group. A sign flip of one hidden node maps every entry of $\boldsymbol{F}^\star$ in which that node appears once to its own negative, and these entries therefore vanish. For $L > 2$ every block between the weights of two different layers also vanishes, and the only remaining block between layers pairs the bias of layer $L-1$ with $\boldsymbol{W}_L$. In each averaged outer product in equation 8, sign flips eliminate the off-diagonal entries on the hidden nodes, and permutations make the diagonal entries equal. This gives

$$\boldsymbol{F}^\star_{11} = \frac{1}{d_1}\mathbb{E}\big[\boldsymbol{a}_1\boldsymbol{a}_1^\top \otimes \|\boldsymbol{\delta}_1\|^2\boldsymbol{I}_{d_1}\big] = \frac{1}{d_1}\mathbb{E}\big[\|\boldsymbol{\delta}_1\|^2\boldsymbol{a}_1\boldsymbol{a}_1^\top\big] \otimes \boldsymbol{I}_{d_1}, \tag{9}$$

$$\boldsymbol{F}^\star_{\ell\ell} = \frac{1}{d_{\ell-1}d_\ell}\mathbb{E}\big[\|\boldsymbol{a}_\ell\|^2\boldsymbol{I}_{d_{\ell-1}} \otimes \|\boldsymbol{\delta}_\ell\|^2\boldsymbol{I}_{d_\ell}\big] = s_\ell\,\boldsymbol{I}_{d_{\ell-1}} \otimes \boldsymbol{I}_{d_\ell}, \tag{10}$$

$$\boldsymbol{F}^\star_{LL} = \frac{1}{d_{L-1}}\mathbb{E}\Big[\|\boldsymbol{a}_L\|^2\boldsymbol{I}_{d_{L-1}} \otimes \boldsymbol{\delta}_L\boldsymbol{\delta}_L^\top\Big] = \frac{1}{d_{L-1}}\boldsymbol{I}_{d_{L-1}} \otimes \mathbb{E}\Big[\|\boldsymbol{a}_L\|^2\boldsymbol{\delta}_L\boldsymbol{\delta}_L^\top\Big]. \tag{11}$$

The hidden nodes keep only the scalar s_ℓ given in Appendix B.1, while the input and output nodes keep full $d_0 \times d_0$ and $d_L \times d_L$ matrices.

Permutation group. Without sign flips, the blocks between two layers and the off-diagonal entries within each block survive the group average. Permutations make the diagonal entries equal and the off-diagonal entries equal within each averaged outer product in equation 8. Writing a vector over these nodes as $\boldsymbol{x} = \bar{x}\,\mathbf{1} + \boldsymbol{x}_\perp$, where $\bar{x}$ is the mean of its entries and $\mathbf{1}$ is the all-ones vector, we call $\bar{x}\,\mathbf{1}$ the mean component and $\boldsymbol{x}_\perp$ the zero-mean component. The average of $\boldsymbol{x}\boldsymbol{x}^\top$ retains $\bar{x}^2\mathbf{1}\mathbf{1}^\top$ on the mean component and spreads $\|\boldsymbol{x}_\perp\|^2$ evenly over the zero-mean component. This gives

$$\boldsymbol{F}^\star_{11} = \mathbb{E}\big[\bar{\delta}_1^2 \boldsymbol{a}_1\boldsymbol{a}_1^\top\big] \otimes \mathbf{1}\mathbf{1}^\top + \tfrac{1}{d_1-1}\mathbb{E}\big[\|\boldsymbol{\delta}_{1\perp}\|^2 \boldsymbol{a}_1\boldsymbol{a}_1^\top\big] \otimes \big(\boldsymbol{I}_{d_1} - \tfrac{1}{d_1}\mathbf{1}\mathbf{1}^\top\big), \tag{12}$$

$$\begin{aligned}\boldsymbol{F}^\star_{\ell\ell} &= s_\ell^{\mathbf{1}}\,\mathbf{1}\mathbf{1}^\top \otimes \mathbf{1}\mathbf{1}^\top + s_\ell^{\mathbf{1},\perp}\,\mathbf{1}\mathbf{1}^\top \otimes \big(\boldsymbol{I}_{d_\ell} - \tfrac{1}{d_\ell}\mathbf{1}\mathbf{1}^\top\big) \\ &\quad + s_\ell^{\perp,\mathbf{1}}\,\big(\boldsymbol{I}_{d_{\ell-1}} - \tfrac{1}{d_{\ell-1}}\mathbf{1}\mathbf{1}^\top\big) \otimes \mathbf{1}\mathbf{1}^\top + s_\ell^{\perp}\,\big(\boldsymbol{I}_{d_{\ell-1}} - \tfrac{1}{d_{\ell-1}}\mathbf{1}\mathbf{1}^\top\big) \otimes \big(\boldsymbol{I}_{d_\ell} - \tfrac{1}{d_\ell}\mathbf{1}\mathbf{1}^\top\big),\end{aligned} \tag{13}$$

$$\boldsymbol{F}^\star_{LL} = \mathbf{1}\mathbf{1}^\top \otimes \mathbb{E}\Big[\bar{a}_L^2 \boldsymbol{\delta}_L\boldsymbol{\delta}_L^\top\Big] + \tfrac{1}{d_{L-1}-1}\big(\boldsymbol{I}_{d_{L-1}} - \tfrac{1}{d_{L-1}}\mathbf{1}\mathbf{1}^\top\big) \otimes \mathbb{E}\Big[\|\boldsymbol{a}_{L\perp}\|^2 \boldsymbol{\delta}_L\boldsymbol{\delta}_L^\top\Big]. \tag{14}$$

The block of $\boldsymbol{W}_\ell$ with hidden nodes on both sides retains the four scalars $s_\ell^{\mathbf{1}}$, $s_\ell^{\mathbf{1},\perp}$, $s_\ell^{\perp,\mathbf{1}}$, and $s_\ell^{\perp}$ given in Appendix B.1 for each component pair from layers $\ell-1$ and ℓ. The blocks of $\boldsymbol{W}_1$ and $\boldsymbol{W}_L$ retain two full matrices on the input or output nodes, one for each component of their hidden nodes.

3.2 WONI CURVATURE

The orbit-averaged blocks in equations 9–14 are computed from sample estimates of weighted second moments. While computing $\boldsymbol{\delta}_\ell$ with actions sampled from the current policy to ensure that these estimates target the FIM, WONI replaces the identity and projection matrices to recover anisotropy lost through orbit averaging, as described below.

Second moments on the hidden nodes. For the signed permutation group, the replacements are

$$\boldsymbol{I}_{d_{\ell-1}} \to \frac{d_{\ell-1}\,\mathbb{E}\big[\|\boldsymbol{\delta}_\ell\|^2 \boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top\big]}{\mathbb{E}[\|\boldsymbol{a}_\ell\|^2\|\boldsymbol{\delta}_\ell\|^2]}, \qquad \boldsymbol{I}_{d_\ell} \to \frac{d_\ell\,\mathbb{E}\Big[\|\boldsymbol{a}_\ell\|^2 \boldsymbol{\delta}_\ell\boldsymbol{\delta}_\ell^\top\Big]}{\mathbb{E}[\|\boldsymbol{a}_\ell\|^2\|\boldsymbol{\delta}_\ell\|^2]}, \tag{15}$$

where the squared norm of the opposite side weights each moment. For $\boldsymbol{W}_1$ and $\boldsymbol{W}_L$, we replace only the hidden-side identity and retain the weighted second moment on the input or output side. Each weight block is then a single Kronecker product of weighted second moments.

For the permutation group, we replace the zero-mean projections by

$$\begin{aligned}\boldsymbol{I}_{d_{\ell-1}} - \frac{1}{d_{\ell-1}}\mathbf{1}\mathbf{1}^\top &\to \frac{(d_{\ell-1}-1)\,\mathbb{E}\big[w\,\boldsymbol{a}_{\ell\perp}\boldsymbol{a}_{\ell\perp}^\top\big]}{\mathbb{E}[w\,\|\boldsymbol{a}_{\ell\perp}\|^2]}, \\ \boldsymbol{I}_{d_\ell} - \frac{1}{d_\ell}\mathbf{1}\mathbf{1}^\top &\to \frac{(d_\ell-1)\,\mathbb{E}\Big[w\,\boldsymbol{\delta}_{\ell\perp}\boldsymbol{\delta}_{\ell\perp}^\top\Big]}{\mathbb{E}[w\,\|\boldsymbol{\delta}_{\ell\perp}\|^2]},\end{aligned} \tag{16}$$

where w is the squared norm of the opposite component in equations 12–14, using the full vector on the input or output side. The pair of zero-mean components becomes a Kronecker product of weighted second moments. A pair with one mean component retains a full moment on the other side, weighted by the squared mean. The pair of mean components retains $s_\ell^{\mathbf{1}}$.

Under either group the replacements preserve $\operatorname{tr}\boldsymbol{F}^\star_{\ell\ell}$, which equals $\operatorname{tr}\boldsymbol{F}_{\ell\ell} = \mathbb{E}[\|\boldsymbol{a}_\ell\|^2\|\boldsymbol{\delta}_\ell\|^2]$ since orthogonal averaging leaves the trace unchanged. The weighted moments retain information about the dependence between $\boldsymbol{a}_\ell$ and $\boldsymbol{\delta}_\ell$ that K-FAC removes through its independence assumption in equation 4. For the signed permutation group, the resulting weight blocks coincide with the TKFAC blocks computed from the same vectors, as derived in Appendix B.3.

Bias. Besides the weights in equations 15 and 16, we treat each bias as the weight of an invariant constant input. In the first layer, the constant input is treated as an additional node, with the bias included in the $\boldsymbol{W}_1$ block. In later layers, signed permutations flip hidden inputs but leave the constant input unchanged, separating the bias from the weights. For the intermediate layers ($1 <$

$\ell < L$), the hidden-output block follows the same identity replacement as in equation 15, while the layer-L bias retains $\mathbb{E}[\boldsymbol{\delta}_L \boldsymbol{\delta}_L^\top]$. Under permutations, the constant input and the mean component of the hidden input nodes remain invariant, keeping the bias with the weights and adding one coordinate to the input-side mean component. Each intermediate-layer pair containing the input-side mean component gains one coordinate and remains a full moment, while at layer L the same applies to the pair of the input-side mean component and the full output vector.

Output and standard deviation. Under a Gaussian policy, the joint block of $\boldsymbol{W}_L$ and $\boldsymbol{b}_L$ requires no orbit averaging. If the mean $\boldsymbol{\mu}$ is linear in the nodes of layer $L-1$ and the standard deviation $\boldsymbol{\sigma}$ is state-independent, an action $a = \boldsymbol{\mu} + \mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, gives $\boldsymbol{\delta}_L = \mathrm{diag}(\boldsymbol{\sigma})^{-1}\boldsymbol{\epsilon}$, independent of $\boldsymbol{a}_L$. With $\boldsymbol{a}_L^+$ denoting $\boldsymbol{a}_L$ augmented by the constant input, the block is exactly $\mathbb{E}[\boldsymbol{a}_L^+ \boldsymbol{a}_L^{+\top}] \otimes \mathrm{diag}(\boldsymbol{\sigma}^{-2})$. The remaining policy parameter, $\log \boldsymbol{\sigma}$, has an uncoupled block $2\boldsymbol{I}$, as derived in Appendix B.4.

We denote by $\hat{\boldsymbol{F}}$ the curvature estimated by WONI, with these blocks placed along its diagonal. Omitting cross-layer blocks discards only the block between the bias of layer $L-1$ and $\boldsymbol{W}_L$ under signed permutations for $L > 2$, and the nonzero blocks in Appendix B.2 under permutations.

3.3 Natural gradients of a set of gradients

Each weight block of $\hat{\boldsymbol{F}}$ other than the full moments is a Kronecker product of two weighted second moments. Let the columns $\boldsymbol{u}_j$ of $\boldsymbol{U}$ and $\boldsymbol{v}_i$ of $\boldsymbol{V}$ be the eigenvectors of the input- and output-side moments, and let $\boldsymbol{S}$ contain the block eigenvalues in the basis $\{\boldsymbol{v}_i \boldsymbol{u}_j^\top\}$. A gradient block $\nabla_{\boldsymbol{W}_\ell} J$ is mapped to its natural-gradient block as

$$\boldsymbol{V}\Big[\big(\boldsymbol{V}^\top \nabla_{\boldsymbol{W}_\ell} J\, \boldsymbol{U}\big) \oslash \big(\boldsymbol{S} + \lambda\big)\Big]\boldsymbol{U}^\top, \tag{17}$$

where $\oslash$ denotes entrywise division and $\lambda > 0$ is a small damping constant. Every block of $\hat{\boldsymbol{F}}$ is positive semidefinite, as shown in Appendix A.3. Under permutations, the gradient block is first split into the component pairs of equation 13. Pairs containing the mean component use their full moments, inverted by Cholesky factorization with the same damping, while the pair of zero-mean components uses equation 17. Replacing division by multiplication with $\boldsymbol{S}$ gives the product of $\hat{\boldsymbol{F}}$ with any vector. When exact Fisher-vector products are not used, inner products, norms, and angles between natural gradients are evaluated under the same $\hat{\boldsymbol{F}}$.

The moments, their eigendecompositions, and $\boldsymbol{S}$ are independent of the gradient and computed once per update. Equation 17 is then applied to the $N + M$ gradients stacked along a batch dimension, adding only two small matrix multiplications per layer and gradient. HF instead performs a separate sequence of Fisher-vector products for each gradient, with each product requiring a backward pass over the batch through the retained KL-gradient graph. WONI therefore shares the dominant computation across gradients, with only a small additional cost per objective or constraint. Appendix B.5 compares the computational complexity of the methods.

Scope of the approximation. The remaining error comes from the Kronecker form of each block and from the dropped blocks between layers. The group is limited by the symmetry that the activation function has, the replacements in equations 15 and 16 relax the identity or projection that the orbit average prescribes, and the Kronecker form fixes both the eigenvectors and the eigenvalues of a block as products of its two sides. Coupled blocks between adjacent layers, which are not used in the experiments, are described in Appendix B.6.

4 Experiments

We evaluate whether WONI can serve as a practical alternative to HF in RL implementations by examining how well its curvature approximation preserves interactions among multiple gradients and how the resulting policy updates compare with HF and other approximation methods. Since the orbit-averaged structure of WONI depends on the activation function, we first compare it with baselines in single-gradient settings across different actor configurations. We then evaluate multi-constraint tasks initialized from infeasible states and extend to a multi-objective setting to assess whether WONI preserves gradient interactions under the policy geometry.

Baselines. We included HF as the standard approach for computing natural gradients in RL and K-FAC as a representative curvature approximation method (Martens, 2010; Martens & Grosse, 2015). EKFAC keeps the Kronecker eigenbasis of K-FAC and replaces its eigenvalues by the second moments of the projected samples in that basis (George et al., 2018). TKFAC rescales the two factors to make the trace of their Kronecker product equal to the trace of the block (Gao et al., 2021). TEKFAC applies both corrections (Gao et al., 2023). For each task, every baseline shares all settings of the algorithm, and only the natural gradient and the quantities derived from it are replaced for the comparison. We use two curvature modes, exact and EMA. The exact mode uses exact Fisher-vector products, while the EMA mode averages the estimated Fisher moments across updates with decay 0.95 and applies Fisher-vector products using the same estimated Fisher to maintain consistency with the natural gradients.

4.1 SINGLE OBJECTIVE SETTING

TRPO (Schulman et al., 2015) is trained on four continuous-control tasks introduced in Gymnasium (Towers et al., 2024): Hopper-v5, Walker2d-v5, HalfCheetah-v5, and LunarLanderContinuous-v3. We evaluate the natural-gradient performance of the curvature approximation methods across a range of actor settings. The layer width is varied among 64, 128, and 256. For the signed permutation group, Tanh is used with either orthogonal initialization (Saxe et al., 2014) or the initialization of Lee et al. (2025), referred to as robust. For the permutation group, ReLU or LeakyReLU is used with He initialization (He et al., 2015). The final-layer weight gain is set to 0.01 or 1. Each run consists of 500 updates using a batch size of 2048 under a KL limit of 0.01. The final return is averaged over 20 evaluation episodes for each of the last five updates.

The results are summarized in Table 1 and the full results table are provided in Appendix C.5. WONI maintains competitive performance across the evaluated conditions, with the main exception of LunarLander-v3 under the signed-permutation setting. KFAC and its variants generally benefit from the additional curvature corrections, although the gains are not uniform across tasks. HF is comparatively weaker overall, yielding the lowest mean return in six of the eight settings.

Table 1: Evaluation results in the single-objective setting across different hidden-layer sizes. Values are reported as mean $\pm$ 95% confidence interval. Bold indicates the highest mean in each row. Gray highlights the highest mean and results not significantly different from it according to a two-sided Welch t-test ($\alpha = 0.05$).

Group	Task	HF	KFAC	TKFAC	EKFAC	TEKFAC	WONI
Signed Permutation	Hopper-v5	2,909.36 ± 197.32	3,165.91 ± 92.69	3,026.20 ± 98.94	3,381.23 ± 70.42	3,302.67 ± 85.76	**3,395.08** ± 63.04
	LunarLander-v3	149.61 ± 21.17	196.44 ± 9.80	**206.00** ± 5.83	193.10 ± 6.49	191.06 ± 6.82	193.06 ± 9.68
	HalfCheetah-v5	1,850.16 ± 205.03	1,814.31 ± 160.66	**2,084.52** ± 182.68	1,978.69 ± 175.33	2,002.78 ± 186.04	1,984.22 ± 184.16
	Walker2d-v5	2,056.83 ± 151.77	2,805.61 ± 153.76	**2,979.37** ± 173.77	2,747.01 ± 172.79	2,841.92 ± 168.18	2,921.98 ± 153.47
Permutation	Hopper-v5	2,182.30 ± 261.13	2,546.56 ± 153.23	2,254.38 ± 160.17	2,990.26 ± 134.31	**3,092.82** ± 134.42	2,988.07 ± 129.90
	LunarLander-v3	200.87 ± 19.21	245.13 ± 5.19	**251.26** ± 5.94	231.36 ± 7.13	231.09 ± 6.70	246.28 ± 5.25
	HalfCheetah-v5	1,148.24 ± 180.77	1,466.82 ± 149.58	1,489.66 ± 147.28	**1,549.66** ± 155.08	1,412.46 ± 135.74	1,448.58 ± 140.58
	Walker2d-v5	998.52 ± 187.38	932.94 ± 112.93	991.17 ± 120.00	1,047.08 ± 128.68	1,132.63 ± 134.23	**1,190.68** ± 131.00

4.2 MULTI-CONSTRAINED SETTING

Two locomotion tasks with the quadruped Laikago and the biped Cassie introduced in Kim et al. (2024) are used. Each task has three constraints on body angle, height, and foot timing, denoted Cost 0, Cost 1, and Cost 2, and training starts from an infeasible state.

The policy is updated by a primal rule that follows $\tilde{\boldsymbol{g}}_{r_1}$ when every constraint holds and otherwise follows $-\sum_{m\in\mathcal{V}} \beta_m \tilde{\boldsymbol{g}}_{c_m}$ over the set $\mathcal{V}$ of violated constraints, where $\boldsymbol{\beta}$ solves the quadratic program $\min_{\boldsymbol{\beta}\geq 0,\, \sum_{m\in\mathcal{V}} \beta_m = 1} \left\| \sum_{m\in\mathcal{V}} \beta_m \tilde{\boldsymbol{g}}_{c_m} \right\|_{\boldsymbol{F}}^2$ on the Fisher inner products among the natural gradients of the violated constraints. Every run takes 2×10^6 environment steps, and the final values are the mean over 100 evaluation episodes of the last checkpoint.

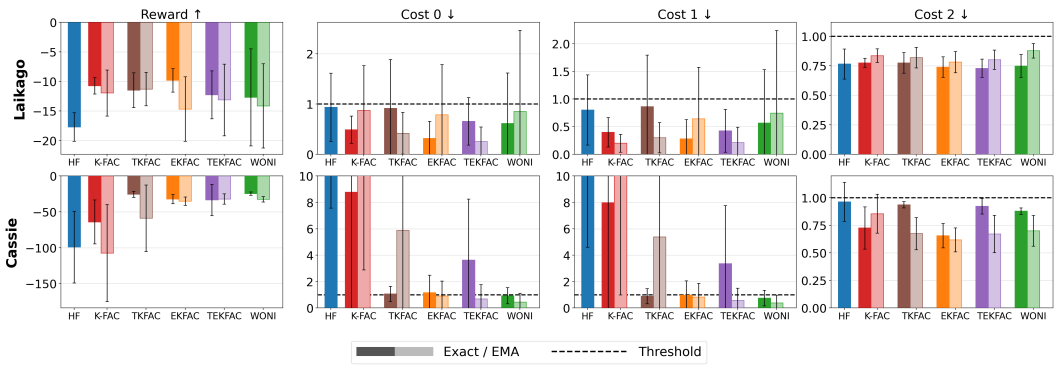


Figure 1:

Figure 1 shows the final return and the three costs normalized by their limits, and Figure 2 in Appendix C.6 shows corresponding learning curve. On Laikago, every method satisfied the constraints, whereas the results varied across methods on Cassie. Difference arises when the QP is solved from the natural gradients and their Fisher inner products with all three constraints violated. HF performs the worst because each natural gradient is estimated by a separate iterative FVP solve. Each solve builds its Krylov subspace from the corresponding gradient, so different advantage-weighted gradients emphasize different Fisher eigendirections during the iterations and can incur different approximation errors. These errors directly affect the gradient interactions used in the QP. The KFAC variants preserve these interactions better as further corrections are applied, with eigenbasis-wise correction preserving them more effectively than trace correction. WONI achieves the best performance in this setting.

4.3 Multi-objective setting

Tasks. The Laikago and Cassie tasks, introduced in their multi-objective extensions with two objectives and three constraints, are trained using CoMOGA with a policy that takes a preference weight as input (Kim et al., 2025). We further compare the original Euclidean update of CoMOGA in the multi-objective setting with its WONI-based update on the Hopper, HalfCheetah, Ant, Humanoid, Swimmer, and Walker2d tasks of MO-Gymnasium (Felten et al., 2023). CoMOGA converts each objective into an improvement target and updates the policy along $\sum_n \nu_n \tilde{\boldsymbol{g}}_{r_n}$, where $\boldsymbol{\nu}$ solves $\min_{\boldsymbol{\nu}\geq 0} \frac{1}{2}\boldsymbol{\nu}^\top \boldsymbol{\Gamma}\boldsymbol{\nu} - \boldsymbol{\xi}^\top \boldsymbol{\nu}$, with $\Gamma_{nn'} = \langle \tilde{\boldsymbol{g}}_{r_n}, \tilde{\boldsymbol{g}}_{r_{n'}} \rangle_{\boldsymbol{F}}$ and $\boldsymbol{\xi}$ determined by the preference weights. Then, trains a single preference-conditioned policy by distilling the intermediate policies computed for preferences sampled from a Dirichlet distribution. Following Kim et al. (2025), hypervolume (HV) is computed from the feasible non-dominated points using a task-specific reference point shared across methods, and sparsity (SP) is scaled depends on the task.

Table 2: Evaluation results on the legged robot tasks in the constrained multi-objective setting. The two values in CP Front indicates the number of non-dominated feasible points and number of feasible points among the 20 different preference.

	Laikago				Cassie			
Method	HV ↑ ($\times 10^3$)	SP ↓ ($\times 10^{-3}$)	CP front ↑	GFLOPs ↓	HV ↑ ($\times 10^3$)	SP ↓ ($\times 10^{-3}$)	CP front ↑	GFLOPs ↓
HF	26.88 ± 15.68	572.1 ± 827.0	6.8 | 9.2	1781.51	6.73 ± 0.34	**20.5** ± 4.6	**14.2** | **16.8**	1746.58
K-FAC	**41.21** ± 1.66	79.7 ± 53.0	9.0 | 10.6	102.14	7.05 ± 0.25	58.5 ± 49.9	10.4 | 11.0	98.92
TKFAC	36.92 ± 5.17	104.4 ± 25.2	7.8 | 11.0	102.21	**7.19** ± 0.25	22.1 ± 7.1	12.8 | 14.6	98.99
EKFAC	37.56 ± 3.12	92.2 ± 91.3	9.0 | 10.2	130.76	7.12 ± 0.29	30.6 ± 17.7	12.8 | 14.6	127.06
TEKFAC	35.78 ± 4.92	76.3 ± 72.7	**10.2** | 11.0	130.82	7.04 ± 0.36	53.4 ± 58.1	11.4 | 12.0	127.13
WONI	39.83 ± 1.62	**52.8** ± 30.4	9.8 | **12.2**	127.68	6.99 ± 0.45	33.8 ± 26.1	12.8 | 15.0	124.58

Table 3: Evaluation results on the multi-objective setting. HV is scaled by 10^9, 10^6, 10^9, 10^4, 10^3, and 10^6 for Hopper-v4, HalfCheetah-v4, Ant-v4, Humanoid-v4, Swimmer-v4, and Walker2d-v4, respectively, and SP by 10^{-3} for all tasks.

	Hopper-v4		HalfCheetah-v4		Ant-v4		Humanoid-v4		Swimmer-v4		Walker2d-v4	
Metric	Euclidean	WONI	Euclidean	WONI	Euclidean	WONI	Euclidean	WONI	Euclidean	WONI	Euclidean	WONI
HV ↑	**6.88** ± 0.43	6.80 ± 0.43	**2.14** ± 0.31	2.07 ± 0.32	**2.00** ± 2.41	1.53 ± 2.06	5.80 ± 1.01	**12.54** ± 2.30	4.70 ± 0.40	**4.76** ± 0.48	5.61 ± 2.08	**6.81** ± 0.52
SP ↓	23.9 ± 8.8	**21.2** ± 12.5	**12.3** ± 2.4	16.7 ± 10.1	21.0 ± 8.5	**12.6** ± 6.3	242.9 ± 232.2	**99.3** ± 84.9	**21.5** ± 5.4	189.4 ± 263.1	49.9 ± 18.7	**15.4** ± 0.3
Front ↑	32.8 ± 6.4	**35.4** ± 9.0	17.6 ± 1.4	**17.8** ± 1.5	25.3 ± 3.3	**31.0** ± 4.3	6.7 ± 2.9	**10.7** ± 6.0	**13.5** ± 1.5	11.8 ± 5.5	11.0 ± 1.6	**17.0** ± 0.8

Table 2 reports the final results of CoMOGA in the constrained multi-objective setting, with Figure 3 in Appendix C.6 presents the corresponding learning curves. WONI remains competitive on the CP point metric, achieving the best or second-best performance across the tasks and consequently yielding favorable HV and SP values. The performance differences among the methods are less pronounced under CoMOGA, where multiple preferences are distilled into a single model without preserving the correlations among the method-specific updates.

Table 3 reports the final results for the multi-objective tasks, and Figure 4 shows the corresponding learning curves. WONI produces better Pareto fronts than the Euclidean methods, consistent with its treatment of interactions among gradients. The final performance gap varies across tasks, however, because the Euclidean step size and the KL constraint govern update magnitudes differently, and a fixed KL limit may suit some tasks better than others.

Limitations. WONI is restricted by the symmetries available in the policy network and currently considers permutation and signed-permutation groups. Its block-diagonal implementation discards cross-layer curvature, while the Kronecker structure further restricts the within-layer approximation. Extending the method to coupled blocks, broader symmetry groups, and other policy architectures remains future work.

5 Conclusion

We proposed WONI, a curvature approximation that computes the natural gradients of multiple objectives and constraints in RL from a single structured estimate of the FIM. Starting from the orbit average of the FIM under the weight-space symmetry of the policy network, WONI replaces the identity and projection matrices on the hidden nodes with weighted second moments estimated from samples and inverts each block in closed form. The estimate is built once per update and shared across all gradients, whereas HF repeats its conjugate gradient iterations for every gradient. In single-objective TRPO, WONI obtained the lowest mean rank over the actor configurations. In the multi-constrained setting, WONI was the only method that kept all three costs of Cassie below their limits in every mode. In the constrained multi-objective setting, WONI reached hypervolumes comparable to the Kronecker-factored baselines and above HF on both legged tasks, and comparable to the Euclidean direction on the MO-Gymnasium tasks. The structure of WONI depends on the symmetry of the activation function and discards the blocks between layers, and extending it to coupled blocks and to architectures with other symmetries remains future work.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Artem Artemev, Rui Xia, Benjamin M. Boyd, Youjing Yu, Felix Dangel, Guillaume Hennequin, and Alberto Bernacchia. Exploiting weight-space symmetries for approximating curvature. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.00442.

Alberto Bernacchia. Global curvature for second-order optimization of neural networks. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/bernacchia25a.html.

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

Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*, 2023.

Kai-Xin Gao, Xiao-Lei Liu, Zheng-Hai Huang, Min Wang, Zidong Wang, Dachuan Xu, and Fan Yu. A trace-restricted kronecker-factored approximation to natural gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7519–7527, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16921.

Kai-Xin Gao, Xiao-Lei Liu, Zheng-Hai Huang, Min Wang, Shuangling Wang, Zidong Wang, Dachuan Xu, and Fan Yu. Eigenvalue-corrected natural gradient based on a new approximation. *Asia-Pacific Journal of Operational Research*, 40(1):2340005, 2023. doi: 10.1142/S0217595923400055.

Brennan Gebotys, Alexander Wong, and David A. Clausi. Understanding the effects of second-order approximations in natural policy gradient reinforcement learning. *arXiv preprint arXiv:2201.09104*, 2022. URL https://arxiv.org/abs/2201.09104.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a kronecker-factored eigenbasis. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/48000647b315f6f00f913caa757a70b3-Abstract.html.

Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022. doi: 10.1007/s10458-022-09552-y. URL https://link.springer.com/article/10.1007/s10458-022-09552-y.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Yingxiao Huo, Satya Prakash Dash, Radu Stoican, Samuel Kaski, and Mingfei Sun. Rank-1 approximation of inverse Fisher for natural policy gradients in deep reinforcement learning. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=ko8Kn7TS6m.

Shun ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1531–1538, 2002. URL https://proceedings.neurips.cc/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html.

Dohyeong Kim, Kyungjae Lee, and Songhwai Oh. Trust region-based safe distributional reinforcement learning for multiple constraints. In *Advances in Neural Information Processing Systems*, 2023.

Dohyeong Kim, Taehyun Cho, Seungyub Han, Hojun Chung, Kyungjae Lee, and Songhwai Oh. Spectral-risk safe reinforcement learning with convergence guarantees. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=9JFSJitKC0.

Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Conflict-averse gradient aggregation for constrained multi-objective reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogXkmugNZw. arXiv v1 title: "Scale-Invariant Gradient Aggregation for Constrained Multi-Objective Reinforcement Learning".

B. Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022.

Hyunwoo Lee, Hayoung Choi, and Hyunju Kim. Robust weight initialization for tanh neural networks with fixed point analysis. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02242.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

James Martens. Deep learning via Hessian-free optimization. In *International Conference on Machine Learning*, pp. 735–742. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/458.pdf.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In Francis Bach and David Blei (eds.), *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. JMLR.org, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 527–538, 2018. URL https://arxiv.org/abs/1810.04650.

Jiaming Song and Yuhuai Wu. An empirical analysis of proximal policy optimization with Kronecker-factored natural gradients. *arXiv preprint arXiv:1801.05566*, 2018. URL https://arxiv.org/abs/1801.05566.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024. URL https://arxiv.org/abs/2407.17032.

Yuhuai Wu, Elman Mansimov, Roger B. Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1708.05144.

Yihang Yao, Zuxin Liu, Zhepeng Cen, Peide Huang, Tingnan Zhang, Wenhao Yu, and Ding Zhao. Gradient shaping for multi-constraint safe reinforcement learning. In *Learning for Dynamics and Control Conference*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020. URL https://arxiv.org/abs/2001.06782.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces. *arXiv preprint arXiv:2506.13018*, 2025.

AI Use Disclosure We used generative AI tools to translate and refine English expressions and for the visualization of the manuscript. We also used them to remove identifying information from our experiment codes for anonymous review and to check that the code is anonymized.

A Proofs

Notation. To distinguish from the inner product and the norm under the FIM, we write the Frobenius inner product as $\langle \boldsymbol{M}_1, \boldsymbol{M}_2 \rangle_{\mathcal{F}} = \mathrm{tr}(\boldsymbol{M}_1^\top \boldsymbol{M}_2)$ and the Frobenius norm as $\|\boldsymbol{M}\|_{\mathcal{F}}^2 = \langle \boldsymbol{M}, \boldsymbol{M} \rangle_{\mathcal{F}}$, where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are matrices of equal size.

A.1 Proof of Lemma 1

Lemma 1. *The orbit-averaged FIM $\boldsymbol{F}^\star$ is the orthogonal projection of $\boldsymbol{F}$ onto the invariant matrix subspace under the Frobenius inner product.*

Proof. Let $\mathcal{C}$ be the set of matrices $\boldsymbol{M}$ with $\boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^\top = \boldsymbol{M}$ for every $\boldsymbol{T} \in \mathcal{G}$. This set is a linear subspace. For a square matrix $\boldsymbol{M}$ of the size of $\boldsymbol{F}$, let $\mathcal{P}(\boldsymbol{M}) = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{T} \in \mathcal{G}} \boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^\top$ denote its orbit average, with $\boldsymbol{F}^\star = \mathcal{P}(\boldsymbol{F})$. We show that $\mathcal{P}$ is the orthogonal projection onto $\mathcal{C}$ under the Frobenius inner product, which gives the lemma.

First, $\mathcal{P}(\boldsymbol{M}) \in \mathcal{C}$ for every $\boldsymbol{M}$. For $\boldsymbol{T}' \in \mathcal{G}$,

$$\boldsymbol{T}' \, \mathcal{P}(\boldsymbol{M}) \, \boldsymbol{T}'^\top = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{T} \in \mathcal{G}} (\boldsymbol{T}'\boldsymbol{T}) \, \boldsymbol{M} \, (\boldsymbol{T}'\boldsymbol{T})^\top = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{T}'' \in \mathcal{G}} \boldsymbol{T}'' \boldsymbol{M} \boldsymbol{T}''^\top = \mathcal{P}(\boldsymbol{M}), \tag{18}$$

where the second equality substitutes $\boldsymbol{T}'' = \boldsymbol{T}'\boldsymbol{T}$, which runs over $\mathcal{G}$ exactly once because $\mathcal{G}$ is a group.

Second, $\mathcal{P}$ fixes $\mathcal{C}$. For $\boldsymbol{M} \in \mathcal{C}$ every term of the average equals $\boldsymbol{M}$, hence

$$\mathcal{P}(\boldsymbol{M}) = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{T} \in \mathcal{G}} \boldsymbol{M} = \boldsymbol{M}. \tag{19}$$

Together with equation 18 this shows that $\mathcal{P}$ is idempotent, $\mathcal{P}(\mathcal{P}(\boldsymbol{M})) = \mathcal{P}(\boldsymbol{M})$, and that its range is exactly $\mathcal{C}$.

Third, $\mathcal{P}$ is self-adjoint under the Frobenius inner product. For matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$,

$$\langle \boldsymbol{M}_1, \mathcal{P}(\boldsymbol{M}_2)\rangle_{\mathcal{F}} = \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{T}\in\mathcal{G}} \operatorname{tr}\Big(\boldsymbol{M}_1^\top \boldsymbol{T}\boldsymbol{M}_2\boldsymbol{T}^\top\Big) = \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{T}\in\mathcal{G}} \operatorname{tr}\Big((\boldsymbol{T}^\top\boldsymbol{M}_1\boldsymbol{T})^\top \boldsymbol{M}_2\Big) \tag{20}$$

$$= \Big\langle \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{T}\in\mathcal{G}} \boldsymbol{T}^\top\boldsymbol{M}_1\boldsymbol{T},\ \boldsymbol{M}_2\Big\rangle_{\mathcal{F}} = \Big\langle \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{T}\in\mathcal{G}} \boldsymbol{T}\boldsymbol{M}_1\boldsymbol{T}^\top,\ \boldsymbol{M}_2\Big\rangle_{\mathcal{F}} \tag{21}$$

$$= \langle \mathcal{P}(\boldsymbol{M}_1), \boldsymbol{M}_2\rangle_{\mathcal{F}}.$$

The second equality in equation 20 is the cyclicity of the trace, and the second equality in equation 21 replaces the sum over $\boldsymbol{T}$ by the sum over $\boldsymbol{T}^\top = \boldsymbol{T}^{-1}$, which runs over $\mathcal{G}$ exactly once because each $\boldsymbol{T}$ is orthogonal and $\mathcal{G}$ is closed under inverses.

A linear map that is idempotent and self-adjoint is the orthogonal projection onto its range. Hence $\mathcal{P}$ is the orthogonal projection onto $\mathcal{C}$, and $\boldsymbol{F}^\star = \mathcal{P}(\boldsymbol{F})$ is the orthogonal projection of $\boldsymbol{F}$ onto the invariant matrix subspace. In particular, for every $\boldsymbol{M} \in \mathcal{C}$ the residual is orthogonal to $\boldsymbol{M}$ by equations 21 and 19,

$$\langle \boldsymbol{F} - \boldsymbol{F}^\star, \boldsymbol{M}\rangle_{\mathcal{F}} = \langle \boldsymbol{F}, \boldsymbol{M}\rangle_{\mathcal{F}} - \langle \boldsymbol{F}, \mathcal{P}(\boldsymbol{M})\rangle_{\mathcal{F}} = 0, \tag{22}$$

which gives $\|\boldsymbol{F} - \boldsymbol{M}\|_{\mathcal{F}}^2 = \|\boldsymbol{F} - \boldsymbol{F}^\star\|_{\mathcal{F}}^2 + \|\boldsymbol{F}^\star - \boldsymbol{M}\|_{\mathcal{F}}^2 \geq \|\boldsymbol{F} - \boldsymbol{F}^\star\|_{\mathcal{F}}^2$, with equality only for $\boldsymbol{M} = \boldsymbol{F}^\star$. The orbit-averaged FIM is therefore the invariant matrix closest to the FIM in the Frobenius norm. □

A.2 Positive semidefiniteness of the orbit-averaged FIM

Lemma 2. *The orbit-averaged FIM $\boldsymbol{F}^\star$ is symmetric positive semidefinite whenever the FIM is.*

Proof. Let $\boldsymbol{F}$ be positive semidefinite and let $\boldsymbol{x}$ be any vector. Then

$$\boldsymbol{x}^\top \boldsymbol{F}^\star \boldsymbol{x} = \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{T}\in\mathcal{G}} \boldsymbol{x}^\top \boldsymbol{T}\boldsymbol{F}\boldsymbol{T}^\top\boldsymbol{x} = \frac{1}{|\mathcal{G}|}\sum_{\boldsymbol{T}\in\mathcal{G}} (\boldsymbol{T}^\top\boldsymbol{x})^\top \boldsymbol{F}(\boldsymbol{T}^\top\boldsymbol{x}) \ \geq\ 0, \tag{23}$$

since each summand is a quadratic form of $\boldsymbol{F}$. Symmetry follows from the symmetry of each $\boldsymbol{T}\boldsymbol{F}\boldsymbol{T}^\top$. □

A.3 Positive semidefiniteness of the WONI curvature

A weighted moment $\mathbb{E}[w\,\boldsymbol{x}\boldsymbol{x}^\top]$ with $w \geq 0$ satisfies $\boldsymbol{e}^\top\mathbb{E}[w\,\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{e} = \mathbb{E}[w\,(\boldsymbol{e}^\top\boldsymbol{x})^2] \geq 0$ for every vector $\boldsymbol{e}$ and is therefore positive semidefinite. Every weight block other than the full moments is the Kronecker product of two weighted moments of this kind divided by a positive scalar, as given in Section 3.3. Each such block is therefore positive semidefinite, because the eigenvalues of a Kronecker product are the products of the eigenvalues of its factors. The bias blocks of the signed permutation group are nonnegative multiples of weighted moments of the same kind, and under the permutation group the bias lies inside the full moments. The full moments of the component pairs that contain the mean component are Gram matrices of per-sample vectors, the exact output block is the Kronecker product of two positive semidefinite matrices, and the block of the log standard deviation is $2\boldsymbol{I}$. Every block of $\hat{\boldsymbol{F}}$ is therefore positive semidefinite, and $\hat{\boldsymbol{F}} + \lambda\boldsymbol{I}$ is positive definite for $\lambda > 0$, which makes equation 17 and the Cholesky factorizations well defined.

B Method Details

B.1 Diagonal blocks of the orbit-averaged FIM

The group $\mathcal{G}$ is the product of the groups $\mathcal{G}_k$ of the nodes of each layer k. For a vector $\boldsymbol{x}$ and a matrix $\boldsymbol{M}$ over the nodes of layer k we write the first- and second-order averages over these nodes as

$$\mathcal{R}_1^k(\boldsymbol{x}) = \frac{1}{|\mathcal{G}_k|}\sum_{\boldsymbol{T}_k\in\mathcal{G}_k} \boldsymbol{T}_k\boldsymbol{x}, \qquad \mathcal{R}_2^k(\boldsymbol{M}) = \frac{1}{|\mathcal{G}_k|}\sum_{\boldsymbol{T}_k\in\mathcal{G}_k} \boldsymbol{T}_k\boldsymbol{M}\boldsymbol{T}_k^\top. \tag{24}$$

On the input and output nodes the groups contain only the identity, and both averages leave their argument unchanged. On d hidden nodes under the signed permutation group, a sign flip of node i changes the sign of x_i and of every entry of $\boldsymbol{M}$ in row i or column i except M_{ii}, and permutations make the remaining diagonal constant, which gives

$$\mathcal{R}_1^k(\boldsymbol{x}) = \mathbf{0}, \qquad \mathcal{R}_2^k(\boldsymbol{M}) = \frac{\operatorname{tr} \boldsymbol{M}}{d}\, \boldsymbol{I}_d. \tag{25}$$

Under the permutation group the nodes form one orbit and the entries of $\boldsymbol{M}$ form two orbits, the diagonal and the off-diagonal entries. Each average replaces the entries of an orbit by their mean, which gives

$$\mathcal{R}_1^k(\boldsymbol{x}) = \bar{x}\, \mathbf{1}, \qquad \mathcal{R}_2^k(\boldsymbol{M}) = \frac{\operatorname{tr} \boldsymbol{M}}{d}\, \boldsymbol{I}_d + \frac{\mathbf{1}^\top \boldsymbol{M} \mathbf{1} - \operatorname{tr} \boldsymbol{M}}{d(d-1)}\, (\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}_d), \tag{26}$$

where $\bar{x}$ is the mean of the entries of $\boldsymbol{x}$.

A transformation $\boldsymbol{T}$ maps the score function $\boldsymbol{\delta}_\ell \boldsymbol{a}_\ell^\top$ of $\boldsymbol{W}_\ell$ to $\boldsymbol{T}_\ell \boldsymbol{\delta}_\ell \boldsymbol{a}_\ell^\top \boldsymbol{T}_{\ell-1}^\top$, and $\operatorname{vec}(\boldsymbol{T}_\ell \boldsymbol{\delta}_\ell \boldsymbol{a}_\ell^\top \boldsymbol{T}_{\ell-1}^\top) = (\boldsymbol{T}_{\ell-1} \boldsymbol{a}_\ell) \otimes (\boldsymbol{T}_\ell \boldsymbol{\delta}_\ell)$. The diagonal block of $\boldsymbol{T}\boldsymbol{F}\boldsymbol{T}^\top$ is therefore

$$\mathbb{E}\Big[\boldsymbol{T}_{\ell-1} \boldsymbol{a}_\ell \boldsymbol{a}_\ell^\top \boldsymbol{T}_{\ell-1}^\top \otimes \boldsymbol{T}_\ell \boldsymbol{\delta}_\ell \boldsymbol{\delta}_\ell^\top \boldsymbol{T}_\ell^\top\Big]. \tag{27}$$

Since $\mathcal{G}$ is this product, $\boldsymbol{T}_{\ell-1}$ and $\boldsymbol{T}_\ell$ are drawn independently, and the average of a Kronecker product of two independent factors is the Kronecker product of their averages, which gives equation 8. For every orthogonal $\boldsymbol{T}_k$, $\operatorname{tr}(\boldsymbol{T}_k \boldsymbol{M} \boldsymbol{T}_k^\top) = \operatorname{tr} \boldsymbol{M}$, and $\operatorname{tr}(\boldsymbol{M}_1 \otimes \boldsymbol{M}_2) = \operatorname{tr} \boldsymbol{M}_1 \operatorname{tr} \boldsymbol{M}_2$, which gives

$$\operatorname{tr} \boldsymbol{F}^\star_{\ell\ell} = \mathbb{E}\big[\|\boldsymbol{a}_\ell\|^2 \|\boldsymbol{\delta}_\ell\|^2\big] = \operatorname{tr} \boldsymbol{F}_{\ell\ell}. \tag{28}$$

Under the signed permutation group the average of a matrix $\boldsymbol{M}$ on d hidden nodes is $(\operatorname{tr} \boldsymbol{M}/d)\, \boldsymbol{I}_d$ by equation 25, and substituting $\boldsymbol{M} = \boldsymbol{a}_\ell \boldsymbol{a}_\ell^\top$ or $\boldsymbol{\delta}_\ell \boldsymbol{\delta}_\ell^\top$ gives equations 9–11 with $s_\ell = \mathbb{E}[\|\boldsymbol{a}_\ell\|^2 \|\boldsymbol{\delta}_\ell\|^2]/(d_{\ell-1} d_\ell)$. Under the permutation group, substituting $\boldsymbol{M} = \boldsymbol{x}\boldsymbol{x}^\top$ into equation 26, with $\operatorname{tr} \boldsymbol{M} = \|\boldsymbol{x}\|^2 = d\bar{x}^2 + \|\boldsymbol{x}_\perp\|^2$ and $\mathbf{1}^\top \boldsymbol{M} \mathbf{1} = d^2 \bar{x}^2$, gives

$$\frac{\|\boldsymbol{x}\|^2}{d} \boldsymbol{I}_d + \frac{d^2\bar{x}^2 - \|\boldsymbol{x}\|^2}{d(d-1)} \big(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}_d\big) = \bar{x}^2\, \mathbf{1}\mathbf{1}^\top + \frac{\|\boldsymbol{x}_\perp\|^2}{d-1} \Big(\boldsymbol{I}_d - \frac{1}{d} \mathbf{1}\mathbf{1}^\top\Big), \tag{29}$$

where the coefficient of $\boldsymbol{I}_d$ on the left is $(d\|\boldsymbol{x}\|^2 - d^2\bar{x}^2)/(d(d-1)) = \|\boldsymbol{x}_\perp\|^2/(d-1)$. Substituting it into equation 8 gives equations 12–14, where expanding the product for a layer with hidden nodes on both sides gives

$$s_\ell^{\mathbf{1}} = \mathbb{E}\big[\bar{a}_\ell^2 \bar{\delta}_\ell^2\big], \quad s_\ell^{\mathbf{1},\perp} = \frac{\mathbb{E}\big[\bar{a}_\ell^2 \|\boldsymbol{\delta}_{\ell\perp}\|^2\big]}{d_\ell - 1}, \quad s_\ell^{\perp,\mathbf{1}} = \frac{\mathbb{E}\big[\|\boldsymbol{a}_{\ell\perp}\|^2 \bar{\delta}_\ell^2\big]}{d_{\ell-1} - 1}, \quad s_\ell^{\perp} = \frac{\mathbb{E}\big[\|\boldsymbol{a}_{\ell\perp}\|^2 \|\boldsymbol{\delta}_{\ell\perp}\|^2\big]}{(d_{\ell-1}-1)(d_\ell - 1)}. \tag{30}$$

B.2 Off-diagonal blocks of the orbit-averaged FIM

This section derives the blocks $\boldsymbol{F}^\star_{\ell\ell'}$ with $\ell < \ell'$, and the blocks with $\ell > \ell'$ follow from the symmetry of $\boldsymbol{F}^\star$. The blocks of the weights are derived first and the blocks of the biases at the end. As in Appendix B.1, the transformations of different layers are drawn independently.

Blocks between two layers. We index an entry of $\boldsymbol{F}_{\ell\ell'} = \mathbb{E}[\boldsymbol{a}_\ell \boldsymbol{a}_{\ell'}^\top \otimes \boldsymbol{\delta}_\ell \boldsymbol{\delta}_{\ell'}^\top]$ by i, p for $\boldsymbol{a}_\ell, \boldsymbol{\delta}_\ell$ and by j, q for $\boldsymbol{a}_{\ell'}, \boldsymbol{\delta}_{\ell'}$, with $[\boldsymbol{F}_{\ell\ell'}]_{(i,p),(j,q)} = \mathbb{E}[a_{\ell,i}\, \delta_{\ell,p}\, a_{\ell',j}\, \delta_{\ell',q}]$. The indices i, p, j, q run over the nodes of layers $\ell - 1, \ell, \ell' - 1, \ell'$. For $\ell' \geq \ell + 2$ these four sets of nodes are distinct, and the independent averages over them give

$$\boldsymbol{F}^\star_{\ell\ell'} = \mathbb{E}\Big[\mathcal{R}_1^{\ell-1}(\boldsymbol{a}_\ell)\, \mathcal{R}_1^{\ell'-1}(\boldsymbol{a}_{\ell'})^\top \otimes \mathcal{R}_1^{\ell}(\boldsymbol{\delta}_\ell)\, \mathcal{R}_1^{\ell'}(\boldsymbol{\delta}_{\ell'})^\top\Big]. \tag{31}$$

For $\ell' = \ell + 1$ the indices p and j run over the nodes of layer ℓ and are moved by the same $\boldsymbol{T}_\ell$, which gives

$$[\boldsymbol{F}^\star_{\ell,\ell+1}]_{(i,p),(j,q)} = \mathbb{E}\big[\mathcal{R}_1^{\ell-1}(\boldsymbol{a}_\ell)_i\, \mathcal{R}_2^{\ell}(\boldsymbol{\delta}_\ell \boldsymbol{a}_{\ell+1}^\top)_{pj}\, \mathcal{R}_1^{\ell+1}(\boldsymbol{\delta}_{\ell+1})_q\big]. \tag{32}$$

Signed permutation group. For $\ell < \ell'$ the nodes of layer ℓ are hidden, hence $\mathcal{R}_1^\ell(\boldsymbol{\delta}_\ell) = \mathbf{0}$ by equation 25 and every block with $\ell' \geq \ell + 2$ vanishes. A block with $\ell' = \ell + 1$ vanishes whenever the nodes of layer $\ell - 1$ or $\ell + 1$ are hidden, which holds for $L > 2$. For $L = 2$ these two sets are the input and output nodes, and equation 32 gives

$$\boldsymbol{F}_{12}^\star = \mathbb{E}\left[\frac{\boldsymbol{\delta}_1^\top \boldsymbol{a}_2}{d_1}\,(\boldsymbol{a}_1 \otimes \boldsymbol{I}_{d_1})(\boldsymbol{I}_{d_1} \otimes \boldsymbol{\delta}_2^\top)\right], \tag{33}$$

which does not vanish in general.

Permutation group. On hidden nodes $\mathcal{R}_1^k$ keeps the mean component $\bar{x}\,\mathbf{1}$. By equation 31 a block with $\ell' \geq \ell + 2$ therefore keeps only the mean component of every set of hidden nodes. When all four sets of nodes involved are hidden, it reduces to a single scalar,

$$\boldsymbol{F}_{\ell\ell'}^\star = \mathbb{E}\big[\bar{a}_\ell\,\bar{a}_{\ell'}\,\bar{\delta}_\ell\,\bar{\delta}_{\ell'}\big]\,\mathbf{1}\mathbf{1}^\top \otimes \mathbf{1}\mathbf{1}^\top. \tag{34}$$

For $\ell' = \ell + 1$ the shared nodes of layer ℓ enter through equation 26 with $\boldsymbol{M} = \boldsymbol{\delta}_\ell \boldsymbol{a}_{\ell+1}^\top$, for which $\operatorname{tr} \boldsymbol{M} = \boldsymbol{\delta}_\ell^\top \boldsymbol{a}_{\ell+1}$ and $\mathbf{1}^\top \boldsymbol{M} \mathbf{1} = d_\ell^2\,\bar{\delta}_\ell\,\bar{a}_{\ell+1}$, and the other two sets of nodes enter through $\mathcal{R}_1^k$. These blocks do not vanish, and keeping only the diagonal blocks drops them under the permutation group.

Biases. The bias of layer ℓ has the score function $\boldsymbol{\delta}_\ell$, which lies on the nodes of layer ℓ only. An entry between $\boldsymbol{b}_\ell$ and another parameter therefore survives the signed permutation group only if every hidden node in it appears an even number of times. For $L > 2$ this leaves a single block between layers, the block between $\boldsymbol{b}_{L-1}$ and $\boldsymbol{W}_L$, in which the index of $\boldsymbol{\delta}_{L-1}$ and the input index of $\boldsymbol{W}_L$ run over the nodes of layer $L - 1$ and the output index runs over the output nodes. By equation 25 with $\boldsymbol{M} = \boldsymbol{\delta}_{L-1}\boldsymbol{a}_L^\top$,

$$\boldsymbol{F}_{\boldsymbol{b}_{L-1}\boldsymbol{W}_L}^\star = \mathbb{E}\left[\frac{\boldsymbol{\delta}_{L-1}^\top \boldsymbol{a}_L}{d_{L-1}}\,\big(\boldsymbol{I}_{d_{L-1}} \otimes \boldsymbol{\delta}_L^\top\big)\right]. \tag{35}$$

The same argument removes the coupling between the bias and the weights of a layer with hidden nodes on both sides, and between the bias and the weights of layer L, which gives the separate bias blocks under the signed permutation group. Under the permutation group $\mathcal{R}_1^k$ keeps the mean component $\bar{x}\,\mathbf{1}$ of the hidden input nodes, hence the coupling between $\boldsymbol{b}_\ell$ and $\boldsymbol{W}_\ell$ survives through $\bar{a}_\ell$ and places the bias with the mean component on the input side.

B.3 Relation to Kronecker-factored approximations

Consider a block with per-sample vectors $\boldsymbol{x}$ on its input side and $\boldsymbol{y}$ on its output side, the exact block $\boldsymbol{B} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \otimes \boldsymbol{y}\boldsymbol{y}^\top]$, with $\operatorname{tr} \boldsymbol{B} = \mathbb{E}[\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2]$. Let

$$\boldsymbol{X} = \mathbb{E}\big[\|\boldsymbol{y}\|^2 \boldsymbol{x}\boldsymbol{x}^\top\big], \qquad \boldsymbol{Y} = \mathbb{E}\big[\|\boldsymbol{x}\|^2 \boldsymbol{y}\boldsymbol{y}^\top\big], \qquad \operatorname{tr} \boldsymbol{X} = \operatorname{tr} \boldsymbol{Y} = \operatorname{tr} \boldsymbol{B}. \tag{36}$$

K-FAC approximates $\boldsymbol{B}$ by $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] \otimes \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$, which equals $\boldsymbol{B}$ when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent. TKFAC takes the trace-one factors $\boldsymbol{X}/\operatorname{tr} \boldsymbol{B}$ and $\boldsymbol{Y}/\operatorname{tr} \boldsymbol{B}$ and scales their product by $\operatorname{tr} \boldsymbol{B}$, which gives $\operatorname{tr} \boldsymbol{B}\,(\boldsymbol{X}/\operatorname{tr} \boldsymbol{B}) \otimes (\boldsymbol{Y}/\operatorname{tr} \boldsymbol{B}) = \boldsymbol{X} \otimes \boldsymbol{Y}/\operatorname{tr} \boldsymbol{B}$.

Under the signed permutation group the replacement equation 15 turns every weight block of equations 9–11 into $\boldsymbol{X} \otimes \boldsymbol{Y}/\operatorname{tr} \boldsymbol{B}$ with $\boldsymbol{x} = \boldsymbol{a}_\ell$ and $\boldsymbol{y} = \boldsymbol{\delta}_\ell$. The WONI block therefore coincides with the TKFAC block computed from the same vectors.

The two constructions differ in the parts that the Kronecker form does not fix. The bias of a layer with hidden input nodes forms a separate block under sign flips, whereas TKFAC keeps it as one more coordinate of $\boldsymbol{x}$. The output block of a Gaussian policy with a linear mean is exact in WONI. The damping is added to the entries of $\boldsymbol{S}$, whereas TKFAC adds it to each factor. Under the permutation group the orbit average separates the mean component of the hidden nodes. A single Kronecker product in the eigenbasis of $\boldsymbol{X}$ and $\boldsymbol{Y}$ mixes this component with the zero-mean component and treats it with diagonal eigenvalues, whereas WONI keeps the full second moments of the component pairs that contain the mean component and inverts them by a Cholesky factorization.

B.4 GAUSSIAN OUTPUT AND STANDARD DEVIATION

Let the policy be $\mathcal{N}(\boldsymbol{\mu}(s), \mathrm{diag}(\boldsymbol{\sigma}^2))$, where $\boldsymbol{\mu}(s) = \boldsymbol{W}_L \boldsymbol{a}_L + \boldsymbol{b}_L$ and the standard deviation $\boldsymbol{\sigma}$ does not depend on the state. Resampling an action as $a = \boldsymbol{\mu} + \mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ gives the two score functions

$$\boldsymbol{\delta}_L = \frac{\partial \log \pi_\theta(a \mid s)}{\partial \boldsymbol{\mu}} = \mathrm{diag}(\boldsymbol{\sigma}^{-2})(a - \boldsymbol{\mu}) = \mathrm{diag}(\boldsymbol{\sigma})^{-1}\boldsymbol{\epsilon}, \qquad \frac{\partial \log \pi_\theta(a \mid s)}{\partial \log \sigma_i} = \epsilon_i^2 - 1. \quad (37)$$

Since $\boldsymbol{\epsilon}$ does not depend on s, $\boldsymbol{\delta}_L$ is independent of $\boldsymbol{a}_L$, and the joint block of $\boldsymbol{W}_L$ and $\boldsymbol{b}_L$ factorizes exactly as $\mathbb{E}[\boldsymbol{a}_L^+ \boldsymbol{a}_L^{+\top}] \otimes \mathbb{E}[\boldsymbol{\delta}_L \boldsymbol{\delta}_L^\top] = \mathbb{E}[\boldsymbol{a}_L^+ \boldsymbol{a}_L^{+\top}] \otimes \mathrm{diag}(\boldsymbol{\sigma}^{-2})$ with $\boldsymbol{a}_L^+ = (\boldsymbol{a}_L, 1)$. The score function of every weight and bias is linear in $\boldsymbol{\epsilon}$, because $\boldsymbol{\delta}_\ell$ follows from $\boldsymbol{\delta}_L$ through matrices that depend on s only. Its product with $\epsilon_i^2 - 1$ therefore has zero mean by the symmetry of $\boldsymbol{\epsilon}$, which leaves $\log \boldsymbol{\sigma}$ uncoupled from the weights and the biases. The block of $\log \boldsymbol{\sigma}$ is $\mathbb{E}[(\epsilon_i^2 - 1)(\epsilon_j^2 - 1)]$, which is 2 for $i = j$ and 0 otherwise, hence $2\boldsymbol{I}$. When the mean passes through a squashing function, $\boldsymbol{\delta}_L$ depends on s through the derivative of this function, and the output block is then treated like any other block.

B.5 COMPUTATIONAL COMPLEXITY

Table 4 lists the leading floating-point operations of one update for a block with d_{in} inputs including the bias and d_{out} outputs, n_s samples, and $K = N + M$ gradients. A product of an $a \times b$ and a $b \times c$ matrix counts $2abc$ operations, a Cholesky factorization of size d counts $d^3/3$, and $E(d)$ denotes a symmetric eigendecomposition of size d. HF is counted for the whole network with n_{cg} conjugate gradient iterations, each a Fisher-vector product of cost $C_{\mathrm{FVP}} = O(n_s n_\theta)$ for n_θ weights. Lower-order terms such as the norms and the elementwise scaling are omitted.

Table 4: Total floating-point operations of one update per block, with WONI under the signed permutation group (signed perm.) and the permutation group (perm.).

Method	Total cost
HF	$Kn_{\mathrm{cg}}C_{\mathrm{FVP}}$
K-FAC, TKFAC	$2n_s(d_{\mathrm{in}}^2 + d_{\mathrm{out}}^2) + E(d_{\mathrm{in}}) + E(d_{\mathrm{out}}) + 4Kd_{\mathrm{in}}d_{\mathrm{out}}(d_{\mathrm{in}} + d_{\mathrm{out}})$
EKFAC, TEKFAC	$4n_s(d_{\mathrm{in}}^2 + d_{\mathrm{out}}^2) + 2n_s d_{\mathrm{in}}d_{\mathrm{out}} + E(d_{\mathrm{in}}) + E(d_{\mathrm{out}}) + 4Kd_{\mathrm{in}}d_{\mathrm{out}}(d_{\mathrm{in}} + d_{\mathrm{out}})$
WONI (signed perm.)	$2n_s(d_{\mathrm{in}}^2 + d_{\mathrm{out}}^2) + E(d_{\mathrm{in}}) + E(d_{\mathrm{out}}) + 4Kd_{\mathrm{in}}d_{\mathrm{out}}(d_{\mathrm{in}} + d_{\mathrm{out}})$
WONI (perm.)	$n_s(4d_{\mathrm{in}}^2 + 10d_{\mathrm{out}}^2) + E(d_{\mathrm{in}}) + E(d_{\mathrm{out}}) + (d_{\mathrm{in}}^3 + 8d_{\mathrm{out}}^3)/3$ $+K\big(4d_{\mathrm{in}}d_{\mathrm{out}}(d_{\mathrm{in}} + d_{\mathrm{out}}) + 2d_{\mathrm{in}}^2 + 8d_{\mathrm{out}}^2\big)$

B.6 COUPLED BLOCKS BETWEEN ADJACENT LAYERS

The hidden nodes of layer ℓ are shared by the rows of $\boldsymbol{W}_\ell$ that write to them and by the columns of $\boldsymbol{W}_{\ell+1}$ that read from them, and under the permutation group the block $\boldsymbol{F}^\star_{\ell,\ell+1}$ of Appendix B.2 does not vanish. The parameters of $\boldsymbol{W}_\ell$ and of $\boldsymbol{W}_{\ell+1}$ attached to the hidden nodes of layer ℓ can be treated jointly as two parameter groups. Each group is expressed in the eigenbasis of its own weighted second moments, where the diagonal matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ hold the second moments of its samples along the basis directions. With $\boldsymbol{C}_{11}$ and $\boldsymbol{C}_{22}$ the full second moments of the two groups and $\boldsymbol{C}_{12}$ their cross moment, carried into the same bases, the coupling is measured by the whitened cross moment $\boldsymbol{H} = \boldsymbol{C}_{11}^{-1/2}\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1/2}$. Its spectral norm satisfies $\|\boldsymbol{H}\|_2 \le 1$ by the Cauchy–Schwarz inequality. The coupled block

$$\begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{D}_1^{1/2}\boldsymbol{H}\boldsymbol{D}_2^{1/2} \\ \boldsymbol{D}_2^{1/2}\boldsymbol{H}^\top\boldsymbol{D}_1^{1/2} & \boldsymbol{D}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{D}_1^{1/2} & \\ & \boldsymbol{D}_2^{1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{H} \\ \boldsymbol{H}^\top & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_1^{1/2} & \\ & \boldsymbol{D}_2^{1/2} \end{bmatrix} \quad (38)$$

is positive semidefinite if $\|\boldsymbol{H}\|_2 \le 1$, because the eigenvalues of the middle factor are one plus or minus the singular values of $\boldsymbol{H}$, together with further eigenvalues equal to one when $\boldsymbol{H}$ is not square. The whitening makes $\|\boldsymbol{H}\|_2 \le 1$ hold by construction, and the coupled block therefore remains positive semidefinite. The experiments use the independent blocks of $\hat{\boldsymbol{F}}$.

C Experimental Details

C.1 Curvature approximations

All curvature approximations use the same damping constant 0.01 and recompute their moments and eigendecompositions at every update, except that the EMA mode averages the moments over updates. K-FAC uses its factored Tikhonov damping, EKFAC and TEKFAC add the damping to the corrected eigenvalues, and TKFAC adds its square root to each trace-scaled factor. WONI adds the damping to $\boldsymbol{S}$ as in equation 17 and to the diagonal of the full moments of the component pairs that contain the mean component. K-FAC and its variants append the bias to the layer input as a constant coordinate. HF solves each natural gradient by conjugate gradient on the exact Fisher-vector product with the same damping inside the solve.

C.2 Single objective setting

The four tasks are used as released in Gymnasium without modification. TRPO follows the implementation of Stable-Baselines3 (Raffin et al., 2021), in which only the natural gradient is replaced. The KL divergence is summed over the action dimensions before the batch mean for the curvature, the step size, and the line search. The actor has a state-independent log standard deviation and a linear mean. The step is sized by an exact Fisher-vector product along the direction and accepted by the backtracking line search of Table 5, and the methods therefore differ in the search direction alone.

Table 5: TRPO hyperparameters.

Actor hidden layers	$(64, 64)$, $(128, 128)$, $(256, 256)$
Critic hidden layers	$(256, 256)$
Critic learning rate (Adam)	0.0003
Target KL	0.01
Rollout length per update	2048
Updates	500
Discount γ	0.99
GAE parameter	0.95
Advantage normalization	True
Line search iterations and decay	10, 0.8
Critic iterations and minibatch size	10, 128
CG iterations (HF)	15
Damping	0.01
Initial log standard deviation	0

C.3 Multi-constrained setting

Laikago and Cassie are the locomotion tasks of Kim et al. (2024) with an episode length of 500 steps and five parallel environments. The costs measure the body angle, the height of the center of mass, and the foot timing, with limits 0.025, 0.025, and 0.4 per step, which are 12.5, 12.5, and 200 per episode. The distributional critics of 25 quantiles are used for the reward and for the costs. Table 6 lists the settings. The final values are the mean over 100 deterministic evaluation episodes of the checkpoint at 2×10^6 steps.

C.4 Multi-objective setting

On Laikago and Cassie, CoMOGA uses the reward terms of Kim et al. (2025) as two objectives and the three constraints above, with an actor of two LeakyReLU layers of width 512, a bounded mean, and a fixed log standard deviation of -2. On MO-Hopper with three objectives and MO-HalfCheetah with two, CoMOGA uses a replay buffer of 10^6 transitions, a batch size of 256, and an entropy constraint. Both groups of tasks use a trust region of size 0.05 and the damping 0.01, and HF uses 10 conjugate gradient iterations. The Pareto front is evaluated with 20 preference weights on Laikago, Cassie, and MO-HalfCheetah and with 50 on MO-Hopper. The hypervolume uses the non-dominated points that satisfy every constraint and a reference point equal to the componentwise

Table 6: Hyperparameters of the multi-constrained setting.

Actor and critic hidden layers	$(512, 512)$, LeakyReLU
Actor output	squashed Gaussian, global learned log standard deviation, initial value -2
Environment steps	$2M$
Steps per update	1000, on a batch of the latest 10000 transitions
Discount γ and GAE parameter	0.99, 0.97
Critic learning rate and iterations	0.0003, 40
Target KL	0.001
Linesearch iterations and decay	10, 0.8
CG iterations (HF)	10
Damping	0.01

minimum of the non-dominated set of the union of these points over all methods, seeds, and checkpoints of a task. The sparsity sums over the objectives the squared gaps between consecutive values of the front normalized to $[0, 1]$ and divides the sum by the number of gaps (Kim et al., 2025).

C.5 Full results

Table 7: Evaluation results in single objective setting across different settings under hidden layer size 64.

Task	Act.	Init.	Gain	HF	KFAC		EKFAC		TKFAC		TEKFAC		WONI	
					Base	EMA	Base	EMA	Base	EMA	Base	EMA	Base	EMA
Hopper -v5	Tanh	Orth.	0.01	2,461.96 ± 861.85	3,246.41 ± 228.46	3,265.23 ± 312.90	3,292.59 ± 463.30	3,573.53 ± 35.30	2,978.44 ± 536.99	2,765.68 ± 651.48	3,533.09 ± 83.72	3,541.65 ± 84.62	3,444.80 ± 142.00	3,460.78 ± 204.71
			1	2,321.35 ± 793.09	3,073.10 ± 421.77	2,947.98 ± 583.77	3,251.80 ± 281.76	3,422.90 ± 115.34	3,125.73 ± 480.11	3,082.34 ± 688.76	3,107.62 ± 520.28	3,428.35 ± 173.32	3,570.28 ± 47.76	3,483.84 ± 131.60
		Robust	0.01	2,754.78 ± 664.37	3,015.12 ± 627.21	3,377.14 ± 451.61	2,947.78 ± 787.29	3,408.69 ± 375.95	2,991.78 ± 443.82	2,254.32 ± 646.50	2,592.05 ± 581.95	2,514.14 ± 471.20	3,279.47 ± 435.91	2,973.66 ± 566.02
			1	2,489.41 ± 1,043.21	2,879.77 ± 435.74	2,857.42 ± 628.61	3,564.59 ± 29.50	3,065.37 ± 283.43	3,185.06 ± 294.68	2,906.68 ± 697.56	2,707.19 ± 791.19	3,125.75 ± 608.89	3,222.77 ± 519.03	3,407.25 ± 204.31
	ReLU	He	0.01	2,933.98 ± 271.03	3,137.38 ± 292.05	2,459.14 ± 1,069.03	3,466.50 ± 269.64	3,301.19 ± 538.94	2,243.33 ± 585.10	2,829.19 ± 512.74	3,329.31 ± 382.18	3,584.45 ± 58.61	3,197.94 ± 579.19	3,202.34 ± 185.42
			1	2,294.63 ± 1,107.27	2,626.13 ± 688.39	2,388.66 ± 1,023.83	2,499.40 ± 1,087.70	2,967.83 ± 203.43	2,032.94 ± 880.00	1,866.51 ± 828.85	2,816.70 ± 800.75	2,431.12 ± 1,228.19	2,757.85 ± 546.35	3,137.96 ± 286.09
	Leaky ReLU		0.01	2,606.86 ± 779.36	2,487.90 ± 434.50	3,055.89 ± 416.40	3,527.17 ± 62.42	3,348.00 ± 201.01	2,964.88 ± 655.61	2,737.22 ± 796.23	3,434.17 ± 324.78	3,393.29 ± 243.74	3,334.19 ± 413.88	2,970.38 ± 773.41
			1	1,414.72 ± 784.76	2,188.04 ± 1,046.82	2,386.04 ± 771.35	2,769.32 ± 888.85	2,188.82 ± 1,099.93	2,193.67 ± 709.32	1,812.15 ± 1,050.96	3,237.83 ± 142.70	2,039.35 ± 1,305.56	2,700.40 ± 1,097.33	2,426.00 ± 1,122.59
LunarLander -v3	Tanh	Orth.	0.01	129.55 ± 96.59	204.25 ± 23.44	199.32 ± 19.03	183.41 ± 18.38	180.76 ± 16.56	225.43 ± 14.95	226.56 ± 19.21	158.65 ± 33.52	180.77 ± 27.52	225.89 ± 18.61	213.93 ± 20.12
			1	194.75 ± 63.16	186.62 ± 30.06	220.29 ± 28.28	158.78 ± 37.54	213.34 ± 16.41	212.44 ± 12.59	211.38 ± 23.27	194.39 ± 36.70	204.19 ± 14.42	194.53 ± 25.29	217.73 ± 16.64
		Robust	0.01	140.22 ± 90.61	231.64 ± 11.55	173.38 ± 77.99	199.76 ± 30.09	207.36 ± 27.79	216.31 ± 12.01	199.17 ± 17.61	203.99 ± 30.92	215.21 ± 10.47	200.37 ± 36.00	194.14 ± 40.34
			1	176.39 ± 73.64	164.37 ± 87.07	204.60 ± 36.70	197.32 ± 23.92	223.90 ± 16.32	221.79 ± 41.33	180.73 ± 48.78	189.27 ± 37.58	199.98 ± 37.60	177.41 ± 50.25	207.49 ± 11.35
	ReLU	He	0.01	190.89 ± 39.40	244.00 ± 20.56	251.57 ± 23.62	206.83 ± 36.60	244.28 ± 12.28	196.13 ± 91.54	255.51 ± 16.39	215.20 ± 14.53	231.92 ± 15.11	230.41 ± 33.15	260.22 ± 12.70
			1	173.88 ± 141.94	240.35 ± 10.11	259.30 ± 13.10	238.31 ± 20.56	261.08 ± 12.93	220.40 ± 58.32	253.65 ± 19.62	215.11 ± 36.55	239.28 ± 21.00	239.14 ± 25.29	245.16 ± 21.80
	Leaky ReLU		0.01	191.25 ± 11.80	234.37 ± 21.24	242.70 ± 42.26	199.05 ± 76.88	222.48 ± 25.68	263.17 ± 16.18	269.93 ± 10.39	237.63 ± 19.36	214.18 ± 74.78	257.95 ± 15.52	245.46 ± 10.77
			1	202.78 ± 44.85	244.30 ± 21.82	255.00 ± 21.81	196.72 ± 47.22	210.39 ± 11.29	244.39 ± 13.83	244.27 ± 27.99	240.42 ± 31.67	245.54 ± 7.02	241.80 ± 21.70	215.87 ± 77.86
HalfCheetah -v5	Tanh	Orth.	0.01	1,586.77 ± 253.61	1,600.70 ± 427.87	1,248.08 ± 115.96	1,385.97 ± 182.01	1,422.60 ± 297.76	1,439.64 ± 86.72	1,423.97 ± 23.88	1,430.34 ± 431.34	1,258.06 ± 58.20	1,955.00 ± 611.21	1,321.02 ± 219.52
			1	1,875.64 ± 758.82	1,854.88 ± 929.20	2,462.26 ± 844.11	2,330.55 ± 1,040.68	2,261.98 ± 1,004.84	2,890.00 ± 817.34	2,863.27 ± 1,043.26	2,061.47 ± 696.71	2,533.62 ± 863.76	2,649.67 ± 1,019.47	1,851.18 ± 792.08
		Robust	0.01	2,194.25 ± 1,080.12	1,896.81 ± 791.34	1,469.50 ± 59.32	2,784.63 ± 1,065.25	1,626.50 ± 307.81	1,567.23 ± 75.14	1,869.15 ± 896.34	2,001.36 ± 619.76	1,374.39 ± 98.00	2,084.78 ± 820.25	2,033.43 ± 833.04
			1	1,336.12 ± 98.26	1,426.43 ± 70.95	2,021.29 ± 1,271.27	1,476.12 ± 105.50	2,070.80 ± 1,030.00	1,472.56 ± 61.32	2,607.71 ± 1,215.32	1,447.57 ± 87.98	1,473.07 ± 188.01	1,831.81 ± 930.80	1,980.14 ± 1,345.35
	ReLU	He	0.01	1,343.57 ± 682.32	2,276.57 ± 916.03	1,195.18 ± 315.80	2,352.68 ± 434.07	1,548.62 ± 799.84	1,898.78 ± 962.63	1,081.36 ± 72.82	1,719.57 ± 689.80	1,813.92 ± 985.48	1,440.32 ± 825.06	1,271.14 ± 382.57
			1	965.86 ± 740.37	1,672.83 ± 1,085.75	1,391.57 ± 743.69	1,465.67 ± 916.19	1,161.68 ± 857.27	1,530.51 ± 985.79	1,668.75 ± 979.08	1,619.37 ± 668.16	1,533.49 ± 1,056.03	1,535.32 ± 857.91	1,563.65 ± 883.61
	Leaky ReLU		0.01	1,641.76 ± 357.45	1,704.26 ± 627.63	1,945.74 ± 757.71	1,499.55 ± 586.18	1,724.32 ± 897.92	1,672.80 ± 697.47	1,389.01 ± 674.36	1,256.58 ± 598.31	1,455.60 ± 389.99	2,028.00 ± 775.32	1,759.51 ± 911.50
			1	1,034.76 ± 717.21	1,391.72 ± 984.24	1,517.97 ± 959.72	1,512.62 ± 1,029.41	1,481.89 ± 925.50	1,360.08 ± 757.20	1,308.56 ± 823.95	1,493.75 ± 877.87	1,468.39 ± 1,061.56	1,261.19 ± 809.44	1,239.20 ± 685.89
Walker2d -v5	Tanh	Orth.	0.01	1,964.55 ± 236.18	2,229.78 ± 683.53	2,129.34 ± 1,087.07	2,715.98 ± 310.80	2,726.21 ± 943.25	2,270.24 ± 933.04	1,912.62 ± 584.65	2,876.55 ± 247.52	3,567.02 ± 256.63	2,802.72 ± 436.84	2,699.24 ± 786.88
			1	1,629.02 ± 469.60	1,731.29 ± 570.70	2,334.92 ± 958.07	2,585.71 ± 480.26	2,097.18 ± 654.37	2,739.38 ± 181.15	3,259.99 ± 538.53	2,338.02 ± 293.59	2,426.73 ± 553.95	2,364.00 ± 713.26	3,026.24 ± 907.37
		Robust	0.01	1,935.59 ± 882.74	3,051.79 ± 419.09	2,591.58 ± 1,226.86	3,011.12 ± 897.89	3,353.03 ± 348.48	2,842.45 ± 839.01	2,270.79 ± 923.49	3,529.12 ± 281.61	3,623.19 ± 409.37	3,180.96 ± 510.87	3,161.26 ± 833.86
			1	2,178.57 ± 637.21	3,637.19 ± 730.94	2,916.99 ± 516.98	3,425.49 ± 677.16	3,900.65 ± 370.35	3,414.16 ± 586.96	3,855.33 ± 476.58	3,007.95 ± 808.20	3,633.99 ± 708.35	3,073.21 ± 454.20	3,445.25 ± 625.60
	ReLU	He	0.01	1,727.05 ± 266.34	1,357.10 ± 373.17	1,154.79 ± 453.10	2,180.80 ± 431.40	1,869.51 ± 364.64	1,141.04 ± 780.29	1,532.48 ± 364.98	1,939.24 ± 277.17	2,154.12 ± 260.39	1,817.53 ± 571.53	1,920.82 ± 307.32
			1	501.09 ± 345.45	485.31 ± 261.42	798.82 ± 475.16	935.66 ± 671.21	737.17 ± 377.15	796.63 ± 541.05	726.41 ± 348.37	1,061.52 ± 859.00	632.43 ± 290.06	1,015.54 ± 695.41	912.99 ± 674.10
	Leaky ReLU		0.01	1,499.89 ± 620.52	1,499.49 ± 506.25	1,428.02 ± 373.58	1,698.65 ± 423.84	1,550.21 ± 360.29	1,236.30 ± 546.16	1,090.14 ± 415.83	1,998.33 ± 317.53	1,710.25 ± 350.78	1,897.15 ± 329.62	1,910.37 ± 502.93
			1	501.10 ± 117.25	549.09 ± 380.46	725.93 ± 551.67	635.88 ± 539.48	468.99 ± 169.59	1,096.03 ± 523.23	815.50 ± 623.51	806.08 ± 773.89	714.54 ± 449.23	1,250.68 ± 807.83	973.93 ± 715.71

Table 8: Evaluation results in single objective setting across different settings under hidden layer size 128.

Task	Act.	Init.	Gain	HF	KFAC		EKFAC		TKFAC		TEKFAC		WONI	
					Base	EMA	Base	EMA	Base	EMA	Base	EMA	Base	EMA
Hopper -v5	Tanh	Orth.	0.01	3,202.74 ± 421.37	3,424.91 ± 116.68	3,248.04 ± 526.05	3,544.77 ± 74.78	3,554.26 ± 74.11	3,176.01 ± 426.00	2,951.20 ± 233.41	3,571.06 ± 64.89	3,576.90 ± 40.25	3,547.76 ± 45.36	3,555.57 ± 73.68
			1	3,310.82 ± 272.29	3,101.21 ± 850.97	3,259.33 ± 292.09	3,526.48 ± 39.14	3,495.17 ± 141.23	3,326.64 ± 194.85	3,247.62 ± 160.01	3,525.70 ± 76.72	3,614.64 ± 50.46	3,435.84 ± 172.85	3,456.16 ± 182.88
		Robust	0.01	3,121.42 ± 523.29	3,486.10 ± 111.22	2,739.34 ± 604.25	3,603.61 ± 27.84	2,941.20 ± 512.59	2,720.71 ± 599.65	2,895.87 ± 426.51	3,114.63 ± 540.27	3,163.09 ± 470.56	3,001.37 ± 817.40	3,527.20 ± 169.34
			1	3,093.26 ± 593.39	3,055.58 ± 280.86	2,807.94 ± 712.85	3,490.65 ± 163.75	3,029.47 ± 796.47	2,998.14 ± 353.00	3,380.14 ± 199.56	3,268.89 ± 373.04	3,277.61 ± 552.67	3,426.15 ± 283.23	3,160.04 ± 342.21
	ReLU	He	0.01	2,803.45 ± 519.15	3,023.85 ± 447.70	2,273.39 ± 321.47	3,349.91 ± 216.09	3,135.54 ± 612.08	2,861.79 ± 676.34	2,117.32 ± 553.36	3,053.81 ± 803.87	3,519.24 ± 65.74	3,267.75 ± 220.02	3,359.60 ± 261.69
			1	2,572.33 ± 647.05	2,295.71 ± 1,271.27	2,386.33 ± 692.77	3,276.49 ± 146.94	3,242.34 ± 268.74	3,021.84 ± 391.54	2,135.21 ± 992.18	3,261.19 ± 117.31	2,959.25 ± 732.73	3,061.86 ± 248.69	3,015.08 ± 431.18
	Leaky ReLU		0.01	2,514.48 ± 927.80	2,869.62 ± 615.51	2,196.78 ± 642.24	3,479.86 ± 62.26	3,423.40 ± 140.71	2,151.81 ± 654.70	2,029.78 ± 783.64	3,005.72 ± 564.80	3,441.55 ± 227.57	3,389.87 ± 181.97	3,293.24 ± 254.50
			1	2,421.55 ± 978.35	2,509.48 ± 848.45	2,352.88 ± 867.73	2,566.93 ± 686.22	3,258.61 ± 251.98	2,337.24 ± 939.86	2,204.96 ± 584.03	3,228.33 ± 303.09	3,222.95 ± 146.72	3,252.93 ± 215.54	2,684.19 ± 504.53
LunarLander -v3	Tanh	Orth.	0.01	123.54 ± 94.96	198.38 ± 11.45	185.58 ± 48.46	165.30 ± 44.25	198.24 ± 28.93	193.67 ± 30.13	216.58 ± 27.19	181.58 ± 48.84	211.10 ± 27.86	104.67 ± 121.60	207.25 ± 19.08
			1	156.99 ± 56.75	218.07 ± 23.00	232.09 ± 13.54	200.18 ± 45.57	196.37 ± 36.29	214.03 ± 16.01	209.14 ± 7.73	173.09 ± 36.56	194.09 ± 18.24	188.58 ± 18.79	194.97 ± 47.37
		Robust	0.01	134.11 ± 67.59	185.67 ± 38.43	199.96 ± 24.81	190.94 ± 23.49	204.05 ± 32.62	194.84 ± 17.35	206.76 ± 37.40	155.20 ± 54.91	207.40 ± 18.55	197.74 ± 16.52	157.62 ± 15.67
			1	170.88 ± 48.46	178.72 ± 65.05	220.13 ± 15.77	177.79 ± 34.07	209.94 ± 43.14	178.87 ± 42.82	189.02 ± 58.47	178.63 ± 51.84	223.67 ± 12.41	213.97 ± 26.99	222.32 ± 18.92
	ReLU	He	0.01	238.98 ± 17.95	233.30 ± 45.24	248.90 ± 34.07	214.33 ± 38.53	265.82 ± 12.93	242.59 ± 25.54	272.08 ± 10.11	206.56 ± 45.16	257.56 ± 15.01	236.77 ± 24.23	255.09 ± 11.20
			1	230.43 ± 34.44	241.57 ± 18.03	270.60 ± 11.71	244.72 ± 18.32	272.51 ± 5.63	239.86 ± 27.99	266.27 ± 8.97	230.43 ± 30.69	257.64 ± 14.67	239.17 ± 14.93	268.89 ± 9.98
	Leaky ReLU		0.01	153.52 ± 153.41	218.46 ± 21.35	252.21 ± 28.82	196.23 ± 24.67	260.40 ± 20.67	246.95 ± 13.37	270.25 ± 12.65	228.04 ± 26.03	246.02 ± 16.26	234.80 ± 33.17	262.88 ± 18.61
			1	174.64 ± 54.79	255.46 ± 10.95	237.35 ± 28.84	245.98 ± 13.99	201.72 ± 61.83	258.62 ± 17.38	261.37 ± 12.35	211.38 ± 26.65	244.15 ± 51.18	234.49 ± 21.23	255.75 ± 15.86
HalfCheetah -v5	Tanh	Orth.	0.01	1,384.85 ± 411.75	1,313.74 ± 119.25	1,274.79 ± 129.97	1,538.56 ± 508.64	1,250.57 ± 77.66	2,242.78 ± 951.84	1,441.74 ± 61.05	1,216.04 ± 49.81	1,211.27 ± 76.32	1,499.49 ± 634.53	1,214.37 ± 47.14
			1	1,395.14 ± 262.92	2,948.62 ± 1,070.33	2,319.48 ± 682.96	2,902.18 ± 347.38	2,546.24 ± 678.69	2,645.20 ± 957.87	2,045.49 ± 1,069.32	2,207.97 ± 617.05	2,008.25 ± 614.92	2,311.83 ± 582.76	2,249.57 ± 733.62
		Robust	0.01	1,925.44 ± 935.39	1,472.45 ± 93.78	1,437.99 ± 42.24	2,208.75 ± 883.55	1,590.41 ± 426.37	2,057.69 ± 943.29	1,524.00 ± 48.85	2,459.32 ± 1,222.63	2,228.30 ± 1,064.97	1,860.34 ± 776.75	1,701.01 ± 862.88
			1	2,688.98 ± 804.25	3,164.08 ± 1,140.57	1,747.33 ± 825.10	2,852.28 ± 1,311.55	2,668.31 ± 1,167.63	2,605.62 ± 1,013.05	2,822.99 ± 1,158.63	3,180.06 ± 1,046.22	3,592.08 ± 1,131.03	2,768.21 ± 1,375.36	3,195.02 ± 1,174.25
	ReLU	He	0.01	1,137.34 ± 301.06	1,964.19 ± 536.18	1,275.93 ± 618.88	1,404.50 ± 398.50	2,176.49 ± 1,121.34	1,623.02 ± 727.62	1,743.16 ± 835.04	1,835.49 ± 674.46	1,361.69 ± 551.97	1,714.13 ± 723.99	1,397.36 ± 509.12
			1	564.38 ± 731.45	1,107.97 ± 686.65	889.80 ± 658.18	1,035.32 ± 758.33	962.15 ± 668.79	1,234.20 ± 959.75	1,063.97 ± 716.60	990.55 ± 767.19	1,318.33 ± 743.42	954.31 ± 704.57	962.20 ± 641.31
	Leaky ReLU		0.01	1,482.71 ± 581.61	1,912.85 ± 550.60	1,312.09 ± 637.55	1,793.12 ± 671.06	1,999.08 ± 941.46	2,074.62 ± 742.07	1,414.58 ± 480.73	1,794.47 ± 724.74	1,169.52 ± 387.47	1,802.89 ± 687.39	1,909.63 ± 585.26
			1	573.30 ± 702.15	826.16 ± 744.75	1,124.26 ± 668.04	1,033.01 ± 757.93	940.24 ± 739.14	918.00 ± 730.56	1,048.40 ± 765.43	1,172.02 ± 707.68	875.82 ± 810.59	1,066.07 ± 691.42	953.63 ± 764.33
Walker2d -v5	Tanh	Orth.	0.01	1,848.16 ± 491.10	2,703.36 ± 768.01	2,906.40 ± 330.12	1,801.43 ± 1,065.12	2,654.53 ± 584.61	2,359.74 ± 760.78	2,200.62 ± 856.01	1,758.92 ± 445.03	2,757.63 ± 483.73	2,202.19 ± 429.13	2,813.79 ± 653.11
			1	2,064.54 ± 250.02	2,605.03 ± 420.20	2,281.68 ± 776.25	2,287.68 ± 460.11	2,426.41 ± 826.93	2,847.11 ± 856.95	2,982.98 ± 330.50	2,198.20 ± 806.21	2,665.86 ± 1,111.34	2,232.46 ± 318.78	2,777.46 ± 370.34
		Robust	0.01	2,392.53 ± 601.63	2,962.08 ± 474.97	3,343.83 ± 519.61	2,769.12 ± 554.56	3,079.50 ± 589.51	3,521.41 ± 1,254.06	3,439.16 ± 812.36	3,043.70 ± 421.52	3,142.04 ± 598.88	3,211.40 ± 836.40	2,752.61 ± 980.52
			1	2,270.56 ± 780.79	3,698.64 ± 339.68	3,853.91 ± 426.72	3,784.08 ± 799.43	4,089.97 ± 674.80	3,956.50 ± 704.35	3,692.47 ± 745.37	3,787.88 ± 725.47	3,535.10 ± 622.97	4,149.22 ± 957.47	3,658.93 ± 663.07
	ReLU	He	0.01	1,662.70 ± 443.33	1,019.49 ± 573.30	1,499.97 ± 156.24	1,400.88 ± 406.77	1,789.16 ± 387.73	1,176.96 ± 557.37	1,636.76 ± 187.87	1,447.42 ± 561.22	1,777.96 ± 461.01	1,705.24 ± 499.92	1,391.90 ± 323.77
			1	315.58 ± 180.75	325.18 ± 234.24	444.60 ± 384.79	295.87 ± 171.93	292.03 ± 197.11	361.70 ± 251.27	338.84 ± 334.16	380.21 ± 260.62	259.06 ± 196.09	417.28 ± 295.45	327.50 ± 223.37
	Leaky ReLU		0.01	1,575.71 ± 393.22	1,239.07 ± 518.35	1,399.63 ± 466.62	1,149.10 ± 321.47	1,738.09 ± 433.54	1,739.18 ± 213.19	1,940.97 ± 131.86	1,844.66 ± 187.90	1,483.49 ± 427.39	1,871.60 ± 469.57	1,469.53 ± 424.41
			1	313.88 ± 193.45	412.95 ± 345.29	505.61 ± 353.66	474.75 ± 520.15	542.90 ± 268.95	305.51 ± 217.53	352.56 ± 223.73	493.27 ± 428.38	376.25 ± 237.66	700.44 ± 424.60	521.00 ± 334.13

Table 9: Evaluation results in single objective setting across different settings under hidden layer size 256.

Task	Act.	Init.	Gain	HF	KFAC		EKFAC		TKFAC		TEKFAC		WONI	
					Base	EMA	Base	EMA	Base	EMA	Base	EMA	Base	EMA
Hopper -v5	Tanh	Orth.	0.01	3,078.46 ± 863.05	3,062.14 ± 336.63	3,562.54 ± 25.22	3,522.86 ± 62.25	3,341.06 ± 366.96	3,188.17 ± 409.57	3,145.65 ± 544.18	3,317.89 ± 216.98	3,508.47 ± 107.59	3,515.60 ± 122.22	3,548.56 ± 27.93
			1	3,094.31 ± 414.69	3,529.48 ± 29.60	3,471.14 ± 145.58	3,423.50 ± 115.44	3,520.69 ± 45.53	3,424.38 ± 183.37	3,199.40 ± 364.89	3,367.50 ± 278.64	3,528.35 ± 70.09	3,490.36 ± 70.44	3,521.73 ± 70.83
		Robust	0.01	3,042.67 ± 355.30	2,713.69 ± 700.35	3,077.31 ± 326.13	3,561.44 ± 62.78	3,506.79 ± 126.46	2,751.17 ± 724.64	2,753.77 ± 606.93	3,495.23 ± 149.20	3,563.52 ± 44.04	3,547.31 ± 72.92	3,238.11 ± 343.37
			1	2,941.20 ± 953.37	3,428.43 ± 83.46	3,352.43 ± 254.41	3,553.09 ± 109.18	3,007.18 ± 450.03	3,322.29 ± 287.51	2,857.53 ± 541.64	3,481.76 ± 180.44	3,339.02 ± 226.53	3,417.05 ± 237.67	3,250.25 ± 373.70
	ReLU	He	0.01	2,114.55 ± 806.58	2,421.37 ± 917.28	2,811.98 ± 551.61	3,047.17 ± 954.06	3,339.08 ± 349.51	2,761.15 ± 648.81	2,852.35 ± 416.35	3,318.43 ± 317.32	3,253.71 ± 489.79	3,270.03 ± 498.80	3,310.19 ± 201.72
			1	1,163.32 ± 933.83	2,731.20 ± 785.88	1,974.82 ± 749.71	2,220.56 ± 762.87	3,086.64 ± 277.91	1,668.33 ± 760.99	1,634.07 ± 561.07	2,967.65 ± 490.93	2,576.49 ± 851.48	2,444.96 ± 1,230.00	1,916.89 ± 940.47
	Leaky ReLU		0.01	2,092.53 ± 775.73	3,004.61 ± 424.15	3,317.40 ± 340.05	2,864.45 ± 609.91	2,996.97 ± 782.02	2,319.72 ± 888.15	2,686.38 ± 755.79	3,510.07 ± 111.04	3,569.10 ± 50.39	3,240.26 ± 443.81	3,052.52 ± 479.59
			1	1,255.24 ± 970.29	1,794.05 ± 990.28	2,424.75 ± 859.90	2,336.80 ± 596.21	2,073.26 ± 864.37	1,436.05 ± 954.89	1,207.20 ± 438.91	2,286.70 ± 1,105.02	2,787.33 ± 835.32	3,152.38 ± 343.83	2,274.93 ± 871.09
LunarLander -v3	Tanh	Orth.	0.01	177.37 ± 61.47	178.77 ± 24.73	128.58 ± 116.10	203.19 ± 39.50	188.05 ± 20.06	185.68 ± 13.30	207.27 ± 43.18	169.05 ± 13.41	201.67 ± 45.24	153.37 ± 54.89	209.84 ± 8.24
			1	98.47 ± 102.16	176.39 ± 57.85	206.44 ± 33.78	187.46 ± 24.78	189.74 ± 45.98	214.69 ± 15.79	197.49 ± 14.05	172.89 ± 18.84	166.99 ± 45.57	105.41 ± 81.32	210.21 ± 36.55
		Robust	0.01	148.09 ± 58.57	215.30 ± 22.09	201.63 ± 28.68	157.82 ± 43.71	210.22 ± 19.42	228.05 ± 20.66	198.80 ± 13.76	195.38 ± 29.83	201.11 ± 25.93	203.97 ± 39.36	208.65 ± 36.02
			1	144.93 ± 85.97	181.38 ± 95.21	223.00 ± 14.84	175.47 ± 31.39	214.90 ± 9.11	194.20 ± 33.03	221.19 ± 30.32	195.76 ± 31.18	211.29 ± 22.06	199.55 ± 52.36	223.94 ± 25.29
	ReLU	He	0.01	239.17 ± 21.92	249.99 ± 8.92	249.72 ± 19.90	200.25 ± 43.95	257.98 ± 15.03	261.18 ± 16.43	267.96 ± 7.45	222.50 ± 29.63	248.33 ± 14.02	238.09 ± 30.15	257.84 ± 18.10
			1	184.89 ± 20.14	221.33 ± 37.80	249.84 ± 14.59	223.87 ± 13.95	254.52 ± 8.54	258.42 ± 18.54	242.38 ± 18.65	210.62 ± 27.82	240.76 ± 33.00	230.14 ± 30.34	242.64 ± 21.46
	Leaky ReLU		0.01	242.69 ± 24.12	245.53 ± 14.40	253.28 ± 32.33	236.17 ± 18.21	256.03 ± 23.70	247.83 ± 21.38	257.82 ± 20.78	189.85 ± 45.87	263.14 ± 15.36	259.80 ± 18.13	265.89 ± 14.19
			1	187.31 ± 45.01	207.45 ± 23.10	276.46 ± 5.38	214.19 ± 27.06	228.85 ± 26.98	230.83 ± 22.59	258.43 ± 7.21	201.32 ± 34.25	248.44 ± 23.42	236.75 ± 9.76	255.78 ± 13.92
HalfCheetah -v5	Tanh	Orth.	0.01	1,737.01 ± 508.91	1,291.52 ± 88.86	1,594.55 ± 704.58	1,195.51 ± 120.04	1,195.05 ± 49.83	1,470.54 ± 19.72	1,905.91 ± 921.64	1,181.96 ± 133.24	1,745.24 ± 1,059.52	1,120.64 ± 59.06	1,200.35 ± 69.52
			1	2,122.72 ± 781.95	2,150.97 ± 1,208.45	2,047.97 ± 1,177.76	2,110.77 ± 1,172.12	2,560.37 ± 1,291.37	2,137.93 ± 1,233.99	2,220.21 ± 1,336.03	2,232.93 ± 1,349.34	2,191.27 ± 1,282.55	2,227.71 ± 1,294.98	2,506.24 ± 1,176.66
		Robust	0.01	2,004.33 ± 784.88	1,462.94 ± 86.51	1,408.15 ± 38.97	1,912.43 ± 943.36	1,268.80 ± 72.57	2,150.37 ± 985.22	1,901.87 ± 895.73	2,491.80 ± 717.36	1,876.14 ± 1,118.09	1,889.15 ± 818.79	1,901.77 ± 1,060.32
			1	1,950.68 ± 814.45	1,789.22 ± 722.36	2,139.69 ± 1,048.83	2,739.11 ± 808.29	1,590.00 ± 561.44	1,923.34 ± 805.10	2,799.35 ± 1,218.32	2,467.58 ± 976.52	2,196.63 ± 1,145.59	2,189.48 ± 1,039.52	2,079.06 ± 920.89
	ReLU	He	0.01	1,345.65 ± 461.88	1,666.83 ± 802.02	1,594.77 ± 489.44	2,081.88 ± 840.63	1,675.12 ± 748.80	1,429.51 ± 599.34	1,566.83 ± 615.99	1,179.38 ± 443.81	1,422.27 ± 905.51	1,594.93 ± 859.51	1,704.47 ± 1,006.86
			1	1,348.44 ± 776.89	1,046.08 ± 520.41	1,239.21 ± 354.51	1,595.16 ± 654.20	1,487.82 ± 777.80	1,645.76 ± 622.98	1,558.73 ± 678.05	1,756.85 ± 600.45	1,633.29 ± 609.88	1,509.22 ± 727.44	1,648.23 ± 625.50
	Leaky ReLU		0.01	1,103.22 ± 269.30	1,939.09 ± 1,193.06	1,644.42 ± 777.59	1,664.81 ± 804.70	1,516.93 ± 689.75	1,989.56 ± 1,110.67	1,805.33 ± 1,035.46	1,286.71 ± 386.35	1,020.43 ± 178.12	1,307.02 ± 616.96	1,223.03 ± 365.20
			1	1,237.84 ± 758.52	1,070.65 ± 627.83	1,493.42 ± 719.01	1,426.62 ± 442.84	1,652.59 ± 651.92	1,463.79 ± 300.28	1,262.61 ± 318.11	1,135.36 ± 407.88	1,586.24 ± 639.43	1,393.13 ± 425.05	1,527.37 ± 707.05
Walker2d -v5	Tanh	Orth.	0.01	1,852.70 ± 236.75	2,240.02 ± 156.30	2,057.39 ± 281.32	2,037.31 ± 564.89	1,534.26 ± 921.05	2,843.73 ± 1,092.84	1,998.49 ± 592.60	1,899.37 ± 218.29	1,349.28 ± 925.44	1,787.88 ± 538.76	2,144.38 ± 346.71
			1	2,325.27 ± 286.35	2,425.46 ± 379.48	2,615.74 ± 704.73	1,782.88 ± 546.05	2,234.70 ± 595.91	2,987.89 ± 774.65	2,592.68 ± 415.01	2,108.23 ± 577.73	3,135.60 ± 915.72	2,401.47 ± 337.07	2,565.03 ± 604.64
		Robust	0.01	2,054.61 ± 657.47	3,050.14 ± 488.90	3,394.56 ± 309.80	2,644.17 ± 674.55	3,275.82 ± 284.38	3,587.56 ± 533.18	3,142.65 ± 845.03	2,301.09 ± 556.45	2,840.63 ± 1,072.26	3,051.28 ± 659.55	3,238.73 ± 723.49
			1	2,165.82 ± 465.76	3,373.48 ± 566.51	3,199.95 ± 842.47	2,692.38 ± 634.09	3,018.63 ± 1,131.80	2,948.70 ± 1,011.87	3,838.13 ± 235.52	3,483.67 ± 661.66	3,196.32 ± 739.60	3,790.91 ± 275.30	3,596.85 ± 238.87
	ReLU	He	0.01	1,639.46 ± 355.40	1,729.40 ± 80.78	1,464.67 ± 341.68	1,313.30 ± 552.42	1,460.02 ± 363.21	1,407.24 ± 290.50	1,777.67 ± 480.09	1,354.90 ± 335.76	1,826.76 ± 240.04	1,080.45 ± 458.63	1,734.48 ± 558.01
			1	283.52 ± 140.75	270.87 ± 140.90	398.82 ± 203.32	411.92 ± 322.17	571.49 ± 403.93	398.17 ± 218.49	361.60 ± 97.30	490.85 ± 284.23	514.03 ± 331.60	595.98 ± 242.64	554.19 ± 318.43
	Leaky ReLU		0.01	1,727.25 ± 131.75	1,315.80 ± 306.83	1,702.67 ± 345.67	1,401.01 ± 467.79	1,493.41 ± 373.41	1,428.06 ± 339.91	1,407.44 ± 345.13	1,197.18 ± 336.07	1,574.93 ± 552.24	1,563.74 ± 240.53	1,811.29 ± 175.93
			1	234.97 ± 110.73	397.14 ± 253.50	266.07 ± 199.78	306.17 ± 216.30	413.06 ± 339.73	431.23 ± 219.34	289.69 ± 75.45	346.28 ± 161.35	799.31 ± 460.45	622.52 ± 379.00	510.16 ± 397.95

C.6 LEARNING CURVES

We report learning curves for the experiments with multiple objectives and constraints in the main text.

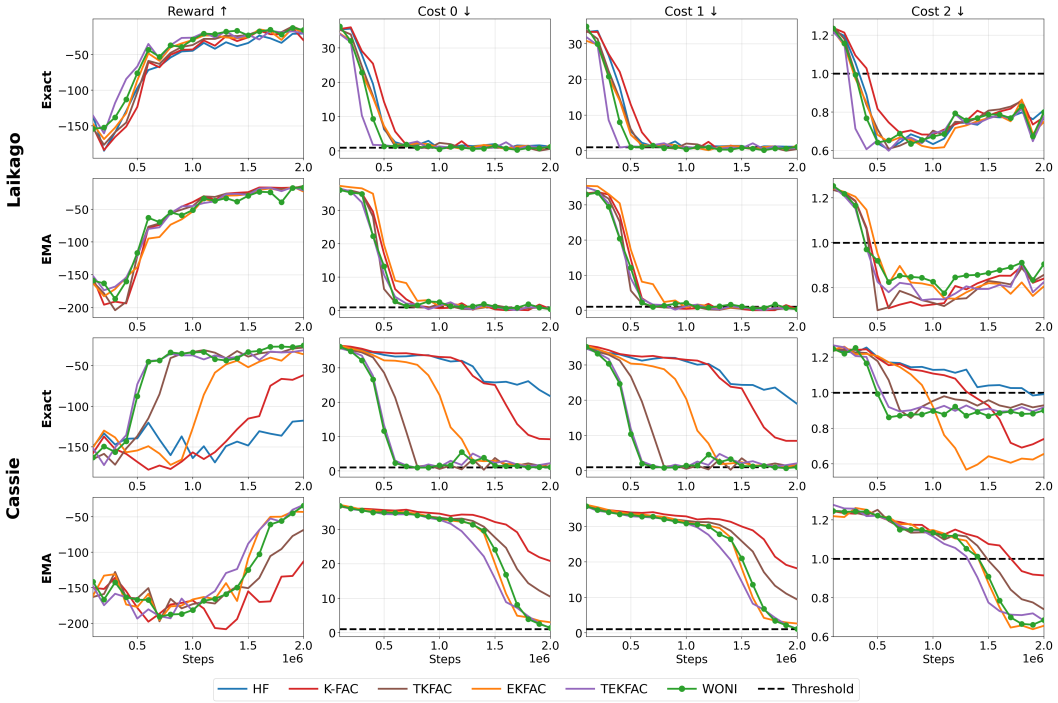


Figure 2: Learning curves on the multi-constrained setting with Cassie and Laikago under three simultaneous constraints.

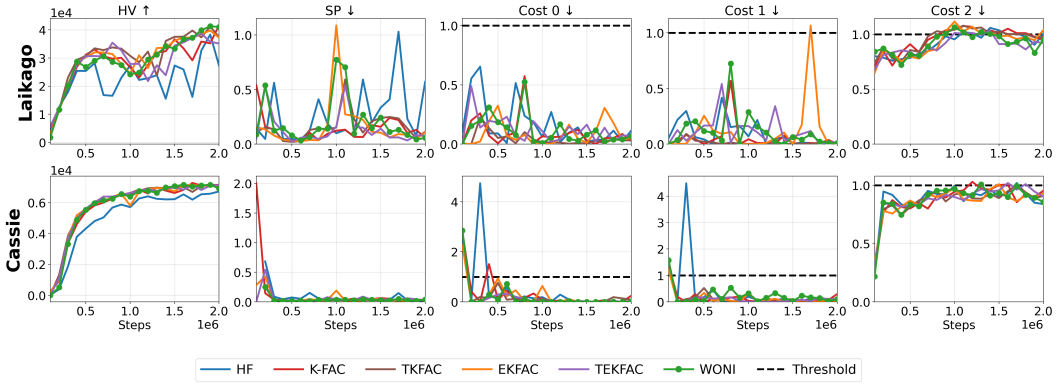


Figure 3: Learning curves on the constrained multi-objective setting with Cassie and Laikago under three simultaneous constraints.

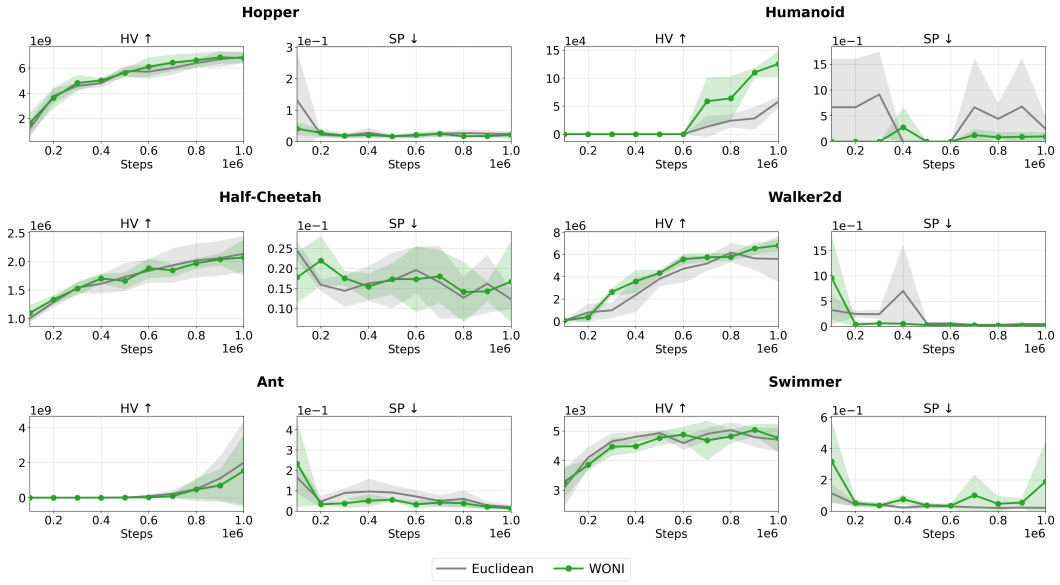


Figure 4: Learning curves on the MO-Gymnasium tasks for HV and SP.