# On Natural Policy Compression

**Leonardo Cesani**[♣,†], **Davide Tenedini**[♣,†], **Matteo Papini**[♣], **Marcello Restelli**[♠]
{leonardo.cesani, davide.tenedini, marcello.restelli}@polimi.it, matteo.papini@unimi.it

♣ **Politecnico di Milano**
♠ **Università degli Studi di Milano**

† Equal contribution

## Abstract

Natural Policy Gradient is the theoretically principled approach to policy optimization: by following the geometry of the statistical manifold of policy distributions rather than the Euclidean geometry of the parameter space, it identifies the true direction of steepest ascent in behavior space. In practice, however, it is often computationally intractable, requiring online estimation and inversion of the Fisher Information Matrix at every iterate. In this paper, we pursue an alternative route: rather than fixing the optimizer, we ask whether the parameter space itself can be reparametrized offline, once and for all, so that standard Euclidean optimization becomes equivalent to natural gradient ascent. In the linear-Gaussian policy setting, we answer affirmatively. We construct an explicit Fisher-whitened embedding in which the squared Euclidean distance is proportional to the average KL divergence between policies, and show that subsequent PCA compression minimizes a Fisher-weighted reconstruction error, yielding a low-dimensional latent space in which vanilla gradient steps recover projected natural-gradient directions. We empirically validate the exactness of this construction in the linear-Gaussian setting. We then use this theoretical framework as a lens to interpret action-based neural policy compression, demonstrating through condition number analysis that its empirical success is explained by how faithfully its latent bottleneck approximates the geometric ideal identified by our analysis.

## 1 Introduction

A central challenge in policy-based Reinforcement Learning (RL, Sutton & Barto, 2018) is that the measure of progress, namely how much a change in the policy update implies a change in behavior, is fundamentally tied to the geometry of the parameter space. Standard Policy Gradient (PG, Peters & Schaal, 2008) methods, which update a parametric policy with stochastic gradient ascent, follow the direction of steepest ascent in the Euclidean space of parameters, which has no intrinsic relation to how much the induced distribution over actions changes. The Natural Policy Gradient (NPG, Kakade, 2001) update corrects this by preconditioning the gradient with the inverse Fisher Information Matrix (FIM), effectively moving in the direction of steepest ascent on the statistical manifold of policy distributions (Amari, 1998). This yields more stable and geometrically meaningful updates, and provides a principled foundation for PG methods that have proven empirically successful (Schulman et al., 2015; 2017; Shao et al., 2024).

While NPG benefits from a geometry-aware update direction, it empirically performs poorly due to statistical issues in estimating the preconditioned gradient. In fact, the FIM must be estimated from experience, stored in memory, and inverted at every iteration; these operations scale quadratically with the number of parameters and are therefore intractable for deep policy networks. While approx-

imations of NPG exist, they either sacrifice the geometric correctness that motivates NPG (Schulman et al., 2015; 2017) or introduce their own computational overhead (Grosse & Martens, 2016; Wu et al., 2017; Cesani et al., 2026). The core difficulty is that the preconditioning is inherently online: it depends on the current policy, changes at every iteration, and must be recomputed from scratch.

In this paper, we take a different perspective. Rather than approximating the FIM inversion online, we aim to find a reward-free reparametrization that learns a well-behaved policy space, in which the Euclidean distance between parameters corresponds to the proximity of the corresponding behaviors. If we can find such a coordinate system for the policy space, then standard gradient ascent in those coordinates is already natural gradient ascent, with no gradient preconditioning required. Crucially, this reparametrization depends only on the geometry of the policy family, not on any reward signal, making it compatible with the unsupervised pre-training phase of Unsupervised RL (URL, Laskin et al., 2021) frameworks, where task information is unavailable.

We analyze this idea formally in the setting of linear multivariate Gaussian policies, where the construction is mathematically tractable, and the geometry is explicit. We show that a Fisher-whitened embedding of policy parameters, evaluated against a mixture of offline occupancy measures, makes squared Euclidean distance exactly proportional to the average KL divergence between policies (Theorem 1). We then show that compressing this embedding via Principal Component Analysis (PCA) minimizes a Fisher-weighted reconstruction error, yielding a low-dimensional latent space in which gradient steps admit a projected natural-gradient interpretation (Theorem 2). The analysis also reveals the central limitation of any fixed reference distribution: as a policy departs from the training bank, the mixture geometry grows increasingly misaligned with its true local Fisher structure (Lemma 3). We close by interpreting a prior work on neural policy compression, Action-based Policy Compression (APC, Tenedini et al., 2026), through this theoretical lens, arguing that its design choices and empirical performance can be explained by how well it approximates the geometric ideal our analysis identifies.

## 2 Preliminaries

**Notation.** We denote sets with calligraphic letters $\mathcal{A}$ and their cardinality as $|\mathcal{A}|$. For a measurable set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of probability measures over $\mathcal{X}$; we write $p$ for the density of $P \in \Delta(\mathcal{X})$ and use $x \sim P$ and $x \sim p$ interchangeably. For $n \in \mathbb{N}$, $[\![n]\!] \coloneqq \{1, \dots, n\}$. Given a symmetric matrix $A \in \mathbb{R}^{d \times d}$, we denote its eigenvalues $\lambda_i$ for $i \in [\![d]\!]$ and its condition number $\kappa \coloneqq \lambda_{\max}/\lambda_{\min}$. We write $A \succeq B$ if $A - B$ is positive semidefinite, and $D(p_1 \parallel p_2)$ for a general divergence between distributions $p_1, p_2 \in \Delta(\mathcal{A})$.

**Interaction Protocol.** We model agent–environment interaction as a finite-horizon Controlled Markov Process (CMP) $\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, p, \rho_0, T)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $\rho_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and $T < \infty$ is the horizon. A stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\pi(\boldsymbol{a} \mid \boldsymbol{s})$ is the probability of taking action $\boldsymbol{a}$ in state $\boldsymbol{s}$, induces a trajectory $\tau \coloneqq (\boldsymbol{s}_0, \boldsymbol{a}_0, \boldsymbol{s}_1, \boldsymbol{a}_1, \dots, \boldsymbol{s}_T)$ and a *state occupancy measure* $d_\pi \in \Delta(\mathcal{S})$, representing the average visitation frequency over the episode:

$$d_\pi(\boldsymbol{s}) \coloneqq \frac{1}{T} \sum_{t=0}^{T-1} \Pr(\boldsymbol{s}_t = \boldsymbol{s} \mid \pi, \rho_0).$$

**Reinforcement Learning.** Augmenting the CMP with a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ yields a Markov Decision Process (MDP, Puterman, 2014) $\mathcal{M}_r \coloneqq \mathcal{M} \cup r$. The agent's goal is to find a policy maximizing the expected return:

$$\max_\pi \; J(\pi) \coloneqq \mathop{\mathbb{E}}_{\substack{\boldsymbol{s}_0 \sim \rho_0 \\ \tau \sim p(\cdot \mid \pi)}} \left[ \sum_{t=0}^{T-1} r(\boldsymbol{s}_t, \boldsymbol{a}_t) \right].$$

**Natural Policy Gradient.** We consider parametric stochastic policies $\pi_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{d_\Theta}$. PG methods optimize the parameter vector $\boldsymbol{\theta}$ by gradient ascent on $J(\pi_{\boldsymbol{\theta}})$, using the gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ estimated from rollouts via the policy gradient theorem (Sutton et al., 1992). While this direction is the steepest ascent in the Euclidean parameter space $\Theta$, it does not account for the underlying geometry of the statistical manifold of policy distributions (Amari, 1998). The correct steepest-ascent direction in distribution space requires preconditioning by the inverse FIM:

$$F(\boldsymbol{\theta}) \coloneqq \mathop{\mathbb{E}}_{\substack{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}}} \\ \boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}}} \left[ \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s}) \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s})^\top \right],$$

yielding the NPG update $\widetilde{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = F(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. Despite its theoretical advantages, NPG is notoriously difficult to scale: estimating and inverting the FIM online at each iterate is computationally intractable for large policy networks.[1]

**Kullback-Leibler Divergence**. To quantify the behavioral discrepancy between two policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$ at a given state $\boldsymbol{s} \in \mathcal{S}$, we employ divergence measures over their conditional action distributions. The Kullback-Leibler (KL) divergence is defined as:

$$D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}) \parallel \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s})) = \int_{\mathcal{A}} \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s}) \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s})}{\pi_{\boldsymbol{\theta}'}(\boldsymbol{a} \mid \boldsymbol{s})} \, d\boldsymbol{a} \quad \forall \boldsymbol{s} \in \mathcal{S}. \tag{1}$$

Note that the second-order Taylor expansion of the KL divergence reduces to a local quadratic form defined by the FIM.

## 3 Information-Geometric Policy Embedding and Compression: The Linear-Gaussian Case

To ensure that policy parameter optimization follows the true statistical manifold of the policy space, we aim to build a universal, offline reparametrization of the policy's parameters. Rather than modifying the optimization algorithm itself, we seek to perform an update that is geometrically equivalent to NPG using standard Euclidean optimization over a pre-collected dataset of behaviors, thereby bypassing the computationally expensive online estimation and inversion of the FIM.

We accomplish this by transforming the underlying coordinates of the policy space such that Euclidean proximity between parameter vectors inherently guarantees behavioral similarity. The method proceeds in two sequential phases. First, we construct a geometry-aware embedding that maps true behavioral divergence directly into Euclidean distance. Second, because this mapping inevitably expands the parameter space and introduces significant structural redundancy, we employ a compression step to isolate the strictly informative, lower-dimensional manifold.

To establish a clean equivalence between Euclidean and Fisher geometry, we work in the setting of linear multivariate Gaussian policies $\pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s}) \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{s}), \Sigma)$, where $\mu_{\boldsymbol{\theta}}(\boldsymbol{s}) = \Phi(\boldsymbol{s})\boldsymbol{\theta}$, $\Phi(\boldsymbol{s}) \in \mathbb{R}^{d_{\mathcal{A}} \times d_\Theta}$, and $\Sigma \succ 0$ is a fixed covariance matrix.[2] The proofs for the theorems and lemma introduced in this section can be found in Appendix C.

### 3.1 Fisher-Whitened Embedding

The primary theoretical challenge is to map the original parameter space into a new geometry in which the Euclidean distance accurately reflects the KL divergence between policies. Because KL divergence is fundamentally state-dependent, building a universal mapping that preserves this geometric equivalence across all arbitrary state distributions is intractable. To overcome this, we fix a reference distribution by employing the mixture of occupancies of an offline policy bank, and show that the resulting embedding faithfully captures behavioral divergence localized to this empirical distribution.

[1] Details on NPG can be found in Appendix B.

[2] We detail the class of linear multivariate Gaussian policies in Appendix A.

Formally, given a population of policies $\Psi = \{\pi_{\boldsymbol{\theta}_i}\}_{i=1}^n$, we define the *mixture occupancy measure* $d_\Psi \in \Delta(\mathcal{S})$ as the uniform mixture of individual occupancies, $d_\Psi(\boldsymbol{s}) = \frac{1}{n}\sum_{i=1}^n d_{\pi_{\boldsymbol{\theta}_i}}(\boldsymbol{s})$. Given a set of states $\{\boldsymbol{s}_j\}_{j=1}^m$ with $\boldsymbol{s}_j \sim d_\Psi$, we define the empirical feature matrix:

$$A_\Psi \coloneqq \frac{1}{2\sqrt{m}}\left[\Sigma^{-1/2}\Phi(\boldsymbol{s}_1),\ \Sigma^{-1/2}\Phi(\boldsymbol{s}_2),\ \ldots,\ \Sigma^{-1/2}\Phi(\boldsymbol{s}_m)\right]^\top \in \mathbb{R}^{(md_{\mathcal{A}})\times d_\Theta},$$

and the corresponding *Fisher-whitened embedding*:

$$\phi_\Psi(\boldsymbol{\theta}) \coloneqq A_\Psi\boldsymbol{\theta} = \frac{1}{2\sqrt{m}}\left[\Sigma^{-1/2}\mu_{\boldsymbol{\theta}}(\boldsymbol{s}_1),\ \Sigma^{-1/2}\mu_{\boldsymbol{\theta}}(\boldsymbol{s}_2),\ \vdots,\ \Sigma^{-1/2}\mu_{\boldsymbol{\theta}}(\boldsymbol{s}_m)\right]^\top \in \mathbb{R}^{md_{\mathcal{A}}}. \quad (2)$$

This representation stacks whitened action means over a representative set of states, and can be interpreted as a *Fisher-whitened Mahalanobis embedding* of policies under $d_\Psi$. The empirical FIM evaluated over the $m$ sampled states, $\widehat{F}_\Psi \coloneqq \frac{1}{m}\sum_{j=1}^m \Phi(\boldsymbol{s}_j)^\top\Sigma^{-1}\Phi(\boldsymbol{s}_j)$, relates to the feature matrix via the identity $\widehat{F}_\Psi = 4A_\Psi^\top A_\Psi$, linking the embedding directly to the Fisher geometry.

**Theorem 1** (Information-Geometric Reparametrization)**.** *For any two linear-Gaussian policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$ with identical covariance $\Sigma$, the squared Euclidean distance between their embeddings satisfies:*

$$\|\phi_\Psi(\boldsymbol{\theta}) - \phi_\Psi(\boldsymbol{\theta}')\|_2^2 = \frac{1}{2m}\sum_{j=1}^m D_{\mathrm{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}_j) \parallel \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s}_j)\big) = \frac{1}{4}(\boldsymbol{\theta}-\boldsymbol{\theta}')^\top\widehat{F}_\Psi(\boldsymbol{\theta}-\boldsymbol{\theta}').$$

Theorem 1 establishes that $\phi_\Psi$ aligns Euclidean and Fisher geometry: the $L_2$ distance between any two embedded policies captures their behavioral divergence, averaged over the reference states. Because of this equivalence, performing a standard gradient step in the embedding space is equivalent to executing an NPG step in the original parameter space.

### 3.2 Compression by PCA

While the Fisher-whitened embedding successfully maps behavioral divergence to Euclidean distance, it operates in an expanded space $\mathbb{R}^{md_{\mathcal{A}}}$ that scales multiplicatively with both the action dimension and the number of reference states, making direct learning computationally infeasible. Moreover, most directions in this embedding space are uninformative: the behavior of a policy family governed by $d_\Theta$ parameters cannot genuinely vary in all $md_{\mathcal{A}}$ dimensions. To resolve this, we compress the representation into a lower-dimensional latent space using PCA, and show that this step inherently preserves the information-geometric properties established by the reparametrization.

Specifically, given embeddings $\phi_i = \phi_\Psi(\boldsymbol{\theta}_i)$ with empirical mean $\bar{\phi}_\Psi = \frac{1}{n}\sum_{i=1}^n \phi_i$, we perform PCA on the centered embeddings, letting $U \in \mathbb{R}^{(md_{\mathcal{A}})\times k}$ contain the top $k$ principal directions. The compressed coordinates are $z_i = U^\top(\phi_i - \bar{\phi}_\Psi)$, yielding the reconstruction $\hat{\phi}_i = \bar{\phi}_\Psi + Uz_i$.

**Theorem 2** (Reconstruction Error)**.** *Let $\hat{\boldsymbol{\theta}}_i = A_\Psi^\dagger\hat{\phi}_i$ be the least-squares parameter reconstruction, where $\dagger$ denotes the Moore-Penrose pseudoinverse. Then the Euclidean reconstruction error in embedding space is equivalent to the Fisher-weighted reconstruction error in parameter space:*

$$\|\phi_\Psi(\boldsymbol{\theta}_i) - A_\Psi\hat{\boldsymbol{\theta}}_i\|_2^2 = \frac{1}{4}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^\top\widehat{F}_\Psi(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i).$$

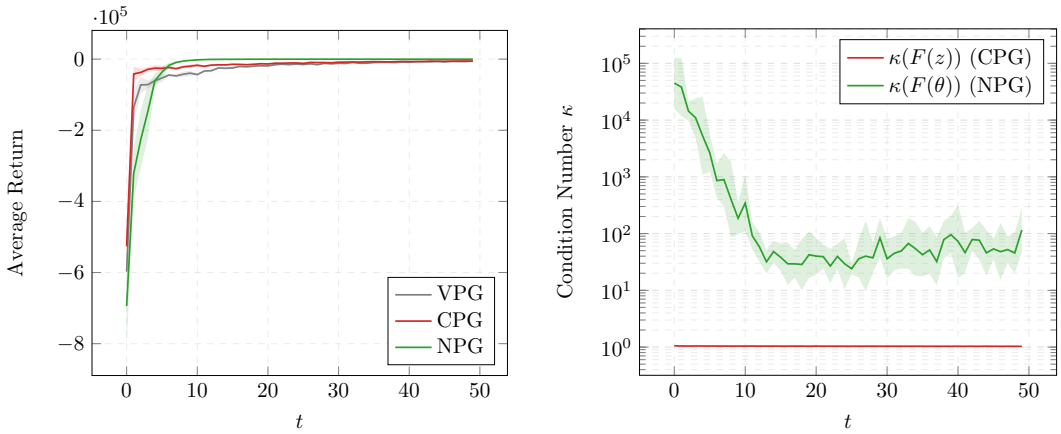
Theorem 2 shows that retaining only the top $k$ principal directions preserves the behavioral similarity established by the embedding: compression minimizes the Fisher-weighted reconstruction error, not an arbitrary Euclidean one. As a consequence, optimization in the compressed coordinates admits a clean interpretation as *projected* NPG with respect to the reference distribution $d_\Psi$.

### 3.3 Empirical Validation

We validate our theoretical framework empirically on an LQR environment with $d_{\mathcal{S}} = 6, d_{\mathcal{A}} = 3$. The complete experimental details can be found in Appendix E.

In Figure 1a, we compare the optimization performance achieved by standard Vanilla PG (VPG) and NPG performed in the original linear policy parameter space with the performance achieved by applying standard PG in the compressed parameter space, which we call Compressed PG (CPG). We show that Compressed PG achieves comparable performance with a faster convergence rate than VPG.

Most notably, in Figure 1b, we compare the condition numbers of the FIM in the original space, $F(\boldsymbol{\theta})$, with the FIM in the compressed space, which we denote as $F(z)$. While $\kappa(F(\boldsymbol{\theta}))$ varies greatly during the optimization, showing the inherent misalignment between the vanilla and natural gradient directions, $\kappa(F(z))$ stays constant around the ideal value of 1. This empirically proves that in the compressed space the Euclidean geometry is equivalent to the Fisher geometry.




(a) Optimization performance comparison.

(b) Condition number comparison.

Figure 1: Learning dynamics comparison over 50 iterations. We report the average and the 95% confidence intervals over 5 runs.

### 3.4 Limitations

The linear-Gaussian case isolates the central principle of our approach: policy compression should be driven by action-level behavior on relevant states, rather than by Euclidean proximity in parameter space. The shared reference distribution $d_\Psi$ is particularly convenient in this setting, as it yields a single explicit Fisher-whitened embedding, a common PCA geometry across all policies, and a projected natural-gradient interpretation of latent-space optimization. At the same time, this construction also reveals its main limitation: when the policy bank spans heterogeneous behaviors, a single global state distribution may fail to be equally representative of all policies. States that are central for one policy may be rarely visited by another, so compression under a fixed shared geometry can underweight policy-specific behavioral structure. We can formalize this discrepancy by examining how the true local Fisher geometry of a generic policy diverges from the pre-computed mixture geometry.

**Lemma 3** (Mixture Error)**.** *Given the exact FIM $F(\boldsymbol{\theta})$ of a parameter $\boldsymbol{\theta}$ and the mixture FIM $F_\Psi$, with $\|\Phi(\boldsymbol{s})\|_2^2 \leq M_\Phi$ and minimum covariance eigenvalue $\lambda_{\min}(\Sigma)$:*

$$\|F(\boldsymbol{\theta}) - F_\Psi\|_2 \leq \frac{2M_\Phi^{3/2}\,T}{\lambda_{\min}(\Sigma)^{3/2}} \cdot \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}\|_2.$$

Lemma 3 shows that as a policy $\pi_{\boldsymbol{\theta}}$ departs from the historical population $\{\boldsymbol{\theta}_i\}_{i=1}^n$, the error between the true Fisher geometry $F(\boldsymbol{\theta})$ and the mixture geometry $F_\Psi$ grows proportionally with the average parameter distance. Because the linear policy class constrains the structure of $A_\Psi$, no single fixed embedding can capture the true geometry simultaneously across all possible state distributions, making learned nonlinear representations a natural next step, as we discuss in the following section.

## 4 Neural Policy Compression

The linear-Gaussian analysis developed in the previous section reveals the essential design principles for behavioral compression: embed parameters into a space where Euclidean distance reflects Fisher geometry, then compress to the informative low-dimensional subspace. The tractability of this construction, however, relies on the linearity of the policy class, which allows an explicit closed-form embedding $\phi_\Psi$. Extending these principles to deep neural policies requires replacing the analytic embedding with a learned one. In this section, we show that action-based neural policy compression can be understood precisely as a nonlinear instantiation of the linear framework above, and that our theoretical analysis provides a principled explanation for both its design choices and empirical behavior.

**Action-Based Compression.** Tenedini et al. (2026) propose a two-phase unsupervised framework for compressing the policy parameter space of deep neural networks. In the pre-training phase, a behaviorally diverse dataset of policies is collected via a novelty search procedure, and a symmetric Autoencoder (AE, Hinton & Salakhutdinov, 2006) with encoder $f_\xi : \Theta \to \mathcal{Z}$ and decoder $g_\zeta : \mathcal{Z} \to \Theta$ is trained to minimize a *behavioral reconstruction loss*: the expected action-space Mean Squared Error (MSE) between the original policy and its reconstruction, evaluated over a fixed shared set of states. In the fine-tuning phase, the frozen decoder defines a differentiable low-dimensional policy manifold, over which task-specific optimization is performed via a latent variant of PGPE (Sehnke et al., 2008).

Viewed through the lens of our theory, each component of this pipeline corresponds directly to a step of the linear construction. The behavioral reconstruction loss is the natural nonlinear analog of our Fisher-weighted reconstruction error in Theorem 2: for deterministic policies, minimizing the action-space MSE over a reference state set is exactly the empirical counterpart of minimizing $(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^\top \widehat{F}_\Psi (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)$, replacing the explicit whitening matrix $A_\Psi$ with a learned nonlinear encoder. Similarly, the fixed shared state set plays the role of our mixture reference distribution $d_\Psi$, and the frozen decoder defines the latent policy manifold over which the projected optimization operates. The novelty-based dataset generation ensures that the policy bank spans a diverse range of behaviors, a prerequisite for the reference distribution to serve as a meaningful geometric proxy, precisely the assumption underlying our analysis.

Our framework also predicts where this approach is struggling. Lemma 3 shows that the error between the true Fisher geometry and the mixture geometry grows as a policy departs from the training bank. When policies are heterogeneous and the shared state set is a poor proxy for individual occupancies, the latent geometry becomes a distorted reflection of the true behavioral structure. This is consistent with the empirical findings of Tenedini et al. (2026), where the method performs well on simpler environments but shows degraded compression quality in more complex continuous control tasks, where behavioral diversity is higher and the shared reference distribution is less representative.

**Latent Geometry and Gradient Alignment.** We measure the alignment between the vanilla PG and NPG via the cosine similarity between gradient directions, defined as:

$$\cos(\vartheta) = \frac{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})^\top F(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\| \, \|F(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\|}.$$

Rather than computing this similarity directly, it is sufficient to analyze the condition number of the FIM, which establishes a strict lower bound on the alignment. We formalize this geometric relationship in the following proposition[3].

The cosine similarity between the vanilla and natural policy gradients satisfies:

$$\cos(\vartheta) \geq \frac{2\sqrt{\lambda_{\min}\lambda_{\max}}}{\lambda_{\min} + \lambda_{\max}} = \frac{2\sqrt{\kappa}}{1+\kappa}.$$

In the original parameter space $\Theta$, deep policy networks are overparametrized, resulting in singular FIMs where $\lambda_{min} \to 0$ and $\kappa \to \infty$. Standard gradient steps in this space are fundamentally misaligned with the true statistical manifold. However, by projecting into the learned manifold $\mathcal{Z}$, we isolate the informative behavioral directions. We evaluate this empirically on the Mountain Car (Moore, 1990) environment. Following the exact action-based compression pipeline of Tenedini et al. (2026), we train autoencoders with varying bottleneck dimensions $k \in \{2, 3, 5, 10\}$ on a bank of 1000 policies ($d_\Theta = 33$). The complete experimental settings can be found in Appendix E. As shown in Figure 2, the mean condition number of the latent FIM is well-behaved for low-dimensional latent spaces, but strictly increases with the bottleneck dimension. This observation validates our theoretical framework: tighter bottlenecks force the generative model to discard redundant parameter combinations, preserving only the essential, orthogonal behavioral variations. As the latent capacity grows, the representation reincorporates functional redundancies, steadily deteriorating the condition number and consequently weakening the alignment between Euclidean and natural gradients.

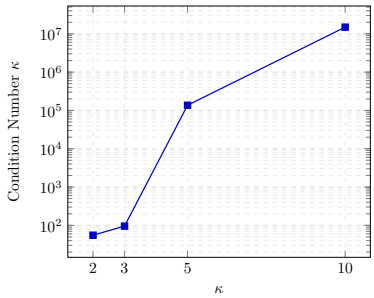



Figure 2: Evolution of latent condition numbers for different latent dimensions. We report the average and the 95% confidence intervals over 1000 policies. The CIs are smaller than the markers.

## 5 Conclusion

We have presented a theoretical analysis of information-geometric policy compression in the linear-Gaussian setting. The central result is that an offline Fisher-whitened embedding, combined with PCA compression, produces a low-dimensional latent space in which standard Euclidean optimization recovers projected natural gradient directions, entirely bypassing the online FIM estimation and inversion that make NPG intractable in practice. The key geometric insight is that behavioral compression works to the extent that the reference distribution used to define the embedding accurately reflects the local Fisher structure of the policies being compressed. Our mixture-error bound makes this precise: the further a policy lies from the training bank, the more the shared geometry underestimates its true local curvature. Still, we validate our framework by proving empirically that in the compressed space, the Euclidean and Fisher geometries are aligned.

This framework provides a principled lens through which to interpret the existing neural compression literature. Action-based compression (Tenedini et al., 2026) instantiates the core principle with a learned nonlinear encoder and a behavioral reconstruction loss that serves as an empirical proxy for Fisher-weighted error. However, it inherently suffers from the fixed-reference limitation our analysis identifies. Furthermore, our empirical evaluation of the condition numbers confirms our theoretical predictions: reducing the bottleneck dimension significantly improves the conditioning of the latent Fisher Information Matrix, ensuring that standard gradient updates faithfully approximate natural-gradient directions.

**Limitations and Future Directions.** The formal guarantees in this paper are restricted to the linear-Gaussian policy class, where the Fisher geometry is analytically tractable. Extending the equiv-

[3]The complete proof can be found in Appendix D.

alence between behavioral reconstruction and Fisher-weighted error to nonlinear neural policies, even approximately, remains an open question, and would put the neural compression literature on a firmer theoretical footing. Moreover, our mixture-error bound reveals the inherent limitations of using a fixed set of states as a reference distribution. More broadly, we hope this analysis motivates the design of compression objectives that explicitly account for the reference geometry they induce, rather than treating the choice of behavioral loss as an implementation detail.

**Acknowledgments**

This work was supported by the Italian Ministry of University and Research (MUR) under the National Recovery and Resilience Plan (NRRP), and by the European Union (EU) under the NextGenerationEU project (CUP D43C24001670008).

## References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. Pmlr, 2017.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Leonardo Cesani, Matteo Papini, and Marcello Restelli. How log-barrier helps exploration in policy optimization, 2026. URL https://arxiv.org/abs/2603.15001.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *International Conference on Machine Learning*, pp. 573–582. PMLR, 2016.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Sham M Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised Reinforcement Learning benchmark. 2021.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Andrew William Moore. Efficient memory-based learning for robot control. Technical report, University of Cambridge, 1990.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697, 2008.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

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

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Policy gradients with parameter-based exploration for control. In *International Conference on Artificial Neural Networks*, pp. 387–396. Springer, 2008.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Richard S Sutton, Andrew G Barto, and Ronald J Williams. Reinforcement learning is direct adaptive optimal control. *IEEE control systems magazine*, 12(2):19–22, 1992.

Davide Tenedini, Riccardo Zamboni, Mirco Mutti, and Marcello Restelli. From parameters to behaviors: Unsupervised compression of the policy space. In *International Conference on Learning Representations*, 2026.

Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, 30, 2017.

# Supplementary Materials

*The following content was not necessarily subject to peer review.*

## A Linear Multivariate Gaussian Policies

Assume the policy mean is parameterized by a state-dependent feature matrix $\Phi(\boldsymbol{s}) \in \mathbb{R}^{d_{\mathcal{A}} \times d_{\Theta}}$, with actions $\boldsymbol{a} \in \mathbb{R}^{d_{\mathcal{A}}}$ and parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{d_{\Theta}}$. The linear-Gaussian policy is:

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s}) = \frac{1}{\sqrt{(2\pi)^{d_{\mathcal{A}}}|\Sigma|}} \exp\left(-\frac{1}{2}\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big)^{\top}\Sigma^{-1}\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big)\right),$$

where $\Sigma$ is the fixed, positive-definite covariance matrix. The log-likelihood is:

$$\log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s}) = -\frac{1}{2}\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big)^{\top}\Sigma^{-1}\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big) + C,$$

where $C$ is a constant independent of $\boldsymbol{\theta}$. The score function is:

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s}) = \Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big).$$

Given the score, we can directly compute the FIM as the expected outer product of the score:

$$\begin{aligned}
F(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}}},\, \boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}}\Big[\big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s})\big)\big(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s})\big)^{\top}\Big] \\
&= \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}}\Big[\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big)\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big)^{\top}\Sigma^{-1}\Phi(\boldsymbol{s})\Big] \\
&= \mathbb{E}_{\boldsymbol{s}}\Big[\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\,\mathbb{E}_{\boldsymbol{a}}\Big[\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big)\big(\boldsymbol{a} - \Phi(\boldsymbol{s})\boldsymbol{\theta}\big)^{\top}\Big]\,\Sigma^{-1}\Phi(\boldsymbol{s})\Big] \\
&= \mathbb{E}_{\boldsymbol{s}}\big[\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\Sigma\Sigma^{-1}\Phi(\boldsymbol{s})\big] \\
&= \mathbb{E}_{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}}}}\big[\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\Phi(\boldsymbol{s})\big].
\end{aligned}$$

The simplification follows because, given $\boldsymbol{s}$, the feature matrix $\Phi(\boldsymbol{s})$ and the inverse covariance $\Sigma^{-1}$ are deterministic and factor out of the inner expectation, leaving it equal to the action covariance $\Sigma$.

### A.1 Mixture FIM

Given a population of $n$ policies with parameters $\{\boldsymbol{\theta}_i\}_{i=1}^{n}$, the mixture occupancy measure is $d_{\Psi}(\boldsymbol{s}) = \frac{1}{n}\sum_{i=1}^{n} d_{\pi_{\boldsymbol{\theta}_i}}(\boldsymbol{s})$, and the corresponding mixture FIM is $F_{\Psi} = \mathbb{E}_{\boldsymbol{s}\sim d_{\Psi}}\big[\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\Phi(\boldsymbol{s})\big]$.

**Lemma 4.** *The mixture FIM decomposes as the uniform average of individual FIMs:*

$$F_{\Psi} = \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{\theta}_i), \qquad F(\boldsymbol{\theta}_i) \coloneqq \mathbb{E}_{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}_i}}}\big[\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\Phi(\boldsymbol{s})\big].$$

*Proof.*

$$\begin{aligned}
F_{\Psi} &= \int_{\mathcal{S}} d_{\Psi}(\boldsymbol{s})\,\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\Phi(\boldsymbol{s})\,d\boldsymbol{s} = \int_{\mathcal{S}} \frac{1}{n}\sum_{i=1}^{n} d_{\pi_{\boldsymbol{\theta}_i}}(\boldsymbol{s})\,\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\Phi(\boldsymbol{s})\,d\boldsymbol{s} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}_i}}}\big[\Phi(\boldsymbol{s})^{\top}\Sigma^{-1}\Phi(\boldsymbol{s})\big] = \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{\theta}_i).
\end{aligned}$$

□

## B Natural Policy Gradient

The standard policy gradient relies on the Euclidean geometry of the parameter space, which can lead to suboptimal optimization paths when parameters encode a probability distribution (Kakade, 2001; Martens, 2020). The objective is to maximize the expected return $J(\boldsymbol{\theta})$, whose gradient is given by the Policy Gradient Theorem (Sutton et al., 1992):

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\substack{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}}} \\ \boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid \boldsymbol{s})\, Q^{\pi_{\boldsymbol{\theta}}}(\boldsymbol{s}, \boldsymbol{a})\right].$$

The steepest ascent direction given by this gradient implicitly assumes an identity metric on the parameter space. This is problematic because the gradient scales inversely with parameter units, and its behavior depends on the arbitrary choice of parametrization (Kakade, 2001).

Information geometry addresses this by measuring distance with the KL divergence (Amari, 1998; Martens, 2020), which provides a natural measure of discrepancy between a policy at $\boldsymbol{\theta}$ and a perturbed policy at $\boldsymbol{\theta} + \mathbf{d}$. A Taylor expansion around $\boldsymbol{\theta}$ yields the local geometry:

$$D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}+\mathbf{d}} \parallel \pi_{\boldsymbol{\theta}}) = \frac{1}{2}\mathbf{d}^\top F(\boldsymbol{\theta})\mathbf{d} + \mathcal{O}(\|\mathbf{d}\|^3),$$

where the FIM acts as the metric tensor of the Riemannian manifold of policy distributions (Amari, 1998; Kakade, 2001):

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

NPG arises from constraining the optimization step by the statistical curvature of the policy rather than by Euclidean distance. Replacing the Euclidean step-size constraint $\mathbf{d}^\top\mathbf{d} \le \epsilon$ with the Fisher constraint $\mathbf{d}^\top F(\boldsymbol{\theta})\mathbf{d} \le \epsilon$ changes the geometry of the steepest ascent problem. Mapping the standard gradient into this constrained space requires multiplication by the inverse metric tensor, yielding the natural gradient:

$$\widetilde{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = F(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

Because the KL divergence metric is independent of the coordinate system, this direction is invariant to any smooth and invertible reparametrization of the policy (Amari, 1998; Kakade, 2001; Martens, 2020).

## C Proofs for Section 3

We work with $m$ states $\{\boldsymbol{s}_j\}_{j=1}^m$ sampled from the mixture occupancy measure $d_\Psi$. The empirical feature matrix is:

$$A_\Psi \coloneqq \frac{1}{2\sqrt{m}} \begin{bmatrix} \Sigma^{-1/2}\Phi(\boldsymbol{s}_1) \\ \Sigma^{-1/2}\Phi(\boldsymbol{s}_2) \\ \vdots \\ \Sigma^{-1/2}\Phi(\boldsymbol{s}_m) \end{bmatrix} \in \mathbb{R}^{(md_{\mathcal{A}})\times d_\Theta},$$

and the corresponding embedding $\phi_\Psi(\boldsymbol{\theta}) \coloneqq A_\Psi\boldsymbol{\theta} \in \mathbb{R}^{md_{\mathcal{A}}}$ as defined in Equation (2).

### C.1 Proof of Theorem 1

**Theorem 1** (Information-Geometric Reparametrization)**.** *For any two linear-Gaussian policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$ with identical covariance $\Sigma$, the squared Euclidean distance between their embeddings satisfies:*

$$\|\phi_\Psi(\boldsymbol{\theta}) - \phi_\Psi(\boldsymbol{\theta}')\|_2^2 = \frac{1}{2m}\sum_{j=1}^m D_{\mathrm{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}_j) \parallel \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s}_j)\big) = \frac{1}{4}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \widehat{F}_\Psi(\boldsymbol{\theta} - \boldsymbol{\theta}').$$

*Proof.* For two linear-Gaussian policies with identical covariance $\Sigma$, the KL divergence at state $\boldsymbol{s}_j$ is the Mahalanobis distance between their means:

$$D_{\mathrm{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}_j) \parallel \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s}_j)\big) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \Phi(\boldsymbol{s}_j)^\top \Sigma^{-1} \Phi(\boldsymbol{s}_j)(\boldsymbol{\theta} - \boldsymbol{\theta}').$$

The squared Euclidean distance between embeddings is:

$$\|\phi_\Psi(\boldsymbol{\theta}) - \phi_\Psi(\boldsymbol{\theta}')\|_2^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top A_\Psi^\top A_\Psi (\boldsymbol{\theta} - \boldsymbol{\theta}').$$

Expanding $A_\Psi^\top A_\Psi$ from the block structure of $A_\Psi$:

$$A_\Psi^\top A_\Psi = \frac{1}{4m} \sum_{j=1}^{m} \Phi(\boldsymbol{s}_j)^\top \Sigma^{-1} \Phi(\boldsymbol{s}_j).$$

Substituting:

$$\begin{aligned}
\|\phi_\Psi(\boldsymbol{\theta}) - \phi_\Psi(\boldsymbol{\theta}')\|_2^2 &= \frac{1}{4m} \sum_{j=1}^{m} (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \Phi(\boldsymbol{s}_j)^\top \Sigma^{-1} \Phi(\boldsymbol{s}_j)(\boldsymbol{\theta} - \boldsymbol{\theta}') \\
&= \frac{1}{2m} \sum_{j=1}^{m} D_{\mathrm{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}_j) \parallel \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s}_j)\big).
\end{aligned}$$

Finally, the empirical FIM $\widehat{F}_\Psi = \frac{1}{m} \sum_{j=1}^{m} \Phi(\boldsymbol{s}_j)^\top \Sigma^{-1} \Phi(\boldsymbol{s}_j) = 4 A_\Psi^\top A_\Psi$, so:

$$\|\phi_\Psi(\boldsymbol{\theta}) - \phi_\Psi(\boldsymbol{\theta}')\|_2^2 = \frac{1}{4}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \widehat{F}_\Psi (\boldsymbol{\theta} - \boldsymbol{\theta}').$$

□

*Remark* 5 (Invariance to Divergence Measure). For linear-Gaussian policies with fixed covariance $\Sigma$, the KL divergence at state $\boldsymbol{s}$ equals the Mahalanobis distance between policy means:

$$D_{\mathrm{KL}}\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}) \parallel \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s})\big) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \Phi(\boldsymbol{s})^\top \Sigma^{-1} \Phi(\boldsymbol{s})(\boldsymbol{\theta} - \boldsymbol{\theta}').$$

As established in Theorem 1, the Euclidean distance in embedding space recovers the Fisher-weighted parameter distance. For the squared Hellinger distance, the exact formula gives:

$$H^2\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}), \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s})\big) = 1 - \exp\left(-\frac{1}{8}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \Phi(\boldsymbol{s})^\top \Sigma^{-1} \Phi(\boldsymbol{s})(\boldsymbol{\theta} - \boldsymbol{\theta}')\right),$$

which for small parameter separations reduces via Taylor expansion to:

$$H^2\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s}), \pi_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{s})\big) \approx \frac{1}{8}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \Phi(\boldsymbol{s})^\top \Sigma^{-1} \Phi(\boldsymbol{s})(\boldsymbol{\theta} - \boldsymbol{\theta}').$$

Consequently, the KL divergence, the squared Hellinger distance, and the Euclidean distance in embedding space all induce the same local quadratic form governed by the FIM, up to constant scaling factors.

### C.2 Proof of Theorem 2

**Theorem 6** (Reconstruction Error)**.** *Let $\hat{\boldsymbol{\theta}}_i = A_\Psi^\dagger \hat{\phi}_i$ be the least-squares parameter reconstruction, where $\dagger$ denotes the Moore-Penrose pseudoinverse. Then the Euclidean reconstruction error in embedding space is equivalent to the Fisher-weighted reconstruction error in parameter space:*

$$\|\phi_\Psi(\boldsymbol{\theta}_i) - A_\Psi \hat{\boldsymbol{\theta}}_i\|_2^2 = \frac{1}{4}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^\top \widehat{F}_\Psi (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i).$$

*Proof.* Given the orthogonal projection matrix $U$ with $U^\top U = I_k$, the compressed code and its embedding-space reconstruction are:

$$z_i = U^\top(\phi_\Psi(\boldsymbol{\theta}_i) - \bar{\phi}_\Psi) \in \mathbb{R}^k,$$
$$\hat{\phi}_i = \bar{\phi}_\Psi + U z_i \in \mathbb{R}^{m d_{\mathcal{A}}}.$$

The least-squares parameter reconstruction $\hat{\boldsymbol{\theta}}_i = A_\Psi^\dagger \hat{\phi}_i$ minimizes $\|A_\Psi \boldsymbol{\theta} - \hat{\phi}_i\|_2^2$ over $\boldsymbol{\theta} \in \mathbb{R}^{d_\Theta}$. The Euclidean reconstruction error is:

$$\|\phi_\Psi(\boldsymbol{\theta}_i) - A_\Psi \hat{\boldsymbol{\theta}}_i\|_2^2 = (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^\top A_\Psi^\top A_\Psi (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i).$$

Substituting the identity $A_\Psi^\top A_\Psi = \frac{1}{4}\widehat{F}_\Psi$ from Theorem 1:

$$\|\phi_\Psi(\boldsymbol{\theta}_i) - A_\Psi \hat{\boldsymbol{\theta}}_i\|_2^2 = \frac{1}{4}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^\top \widehat{F}_\Psi (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i).$$

Because $A_\Psi \hat{\boldsymbol{\theta}}_i$ is the orthogonal projection of $\hat{\phi}_i$ onto the column space of $A_\Psi$, minimizing the embedding-space error is equivalent to minimizing the Fisher-weighted parameter reconstruction error under $d_\Psi$. □

### C.3 Proof of Lemma 3

**Lemma 7** (Mixture Error)**.** *Given the exact FIM $F(\boldsymbol{\theta})$ of a parameter $\boldsymbol{\theta}$ and the mixture FIM $F_\Psi$, with $\|\Phi(\boldsymbol{s})\|_2^2 \le M_\Phi$ and minimum covariance eigenvalue $\lambda_{\min}(\Sigma)$:*

$$\|F(\boldsymbol{\theta}) - F_\Psi\|_2 \le \frac{2M_\Phi^{3/2} T}{\lambda_{\min}(\Sigma)^{3/2}} \cdot \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{\theta}_i - \boldsymbol{\theta}\|_2.$$

*Proof.* Using the FIM definition for linear-Gaussian policies and linearity of integration:

$$\|F(\boldsymbol{\theta}) - F_\Psi\|_2 = \left\| \int_{\mathcal{S}} \big(d_{\boldsymbol{\theta}}(\boldsymbol{s}) - d_\Psi(\boldsymbol{s})\big) \Phi(\boldsymbol{s})^\top \Sigma^{-1} \Phi(\boldsymbol{s})\, d\boldsymbol{s} \right\|_2.$$

Applying the triangle inequality and sub-multiplicativity of the spectral norm, then using $\|\Phi(\boldsymbol{s})\|_2^2 \le M_\Phi$ and $\|\Sigma^{-1}\|_2 = 1/\lambda_{\min}(\Sigma)$:

$$\|F(\boldsymbol{\theta}) - F_\Psi\|_2 \le \frac{M_\Phi}{\lambda_{\min}(\Sigma)} \int_{\mathcal{S}} |d_{\boldsymbol{\theta}}(\boldsymbol{s}) - d_\Psi(\boldsymbol{s})|\, d\boldsymbol{s} = \frac{2M_\Phi}{\lambda_{\min}(\Sigma)} D_{\mathrm{TV}}(d_\Psi \parallel d_{\boldsymbol{\theta}}).$$

By convexity of total variation:

$$D_{\mathrm{TV}}(d_\Psi \parallel d_{\boldsymbol{\theta}}) \le \frac{1}{n}\sum_{i=1}^n D_{\mathrm{TV}}(d_{\pi_{\boldsymbol{\theta}_i}} \parallel d_{\boldsymbol{\theta}}).$$

By the finite-horizon occupancy bound of Achiam et al. (2017):

$$D_{\mathrm{TV}}(d_\Psi \parallel d_{\boldsymbol{\theta}}) \le \frac{2T}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}_i}}}\left[D_{\mathrm{TV}}\big(\pi_{\boldsymbol{\theta}_i}(\cdot \mid \boldsymbol{s}) \parallel \pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s})\big)\right].$$

Applying Pinsker's inequality ($D_{\mathrm{TV}} \le \sqrt{D_{\mathrm{KL}}/2}$) and Jensen's inequality for the square root:

$$D_{\mathrm{TV}}(d_\Psi \parallel d_{\boldsymbol{\theta}}) \le \sqrt{2}\, T \frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}_{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}_i}}}\left[D_{\mathrm{KL}}\big(\pi_{\boldsymbol{\theta}_i}(\cdot \mid \boldsymbol{s}) \parallel \pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{s})\big)\right]}.$$

For linear-Gaussian policies, the expected KL divergence under $d_{\pi_{\boldsymbol{\theta}_i}}$ is:

$$\mathbb{E}_{\boldsymbol{s}\sim d_{\pi_{\boldsymbol{\theta}_i}}}\big[D_{\mathrm{KL}}\big(\pi_{\boldsymbol{\theta}_i} \parallel \pi_{\boldsymbol{\theta}}\big)\big] = \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta})^\top F(\boldsymbol{\theta}_i)(\boldsymbol{\theta}_i - \boldsymbol{\theta}) \le \frac{M_\Phi}{2\lambda_{\min}(\Sigma)}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}\|_2^2,$$

where the inequality uses $\|F(\boldsymbol{\theta}_i)\|_2 \le M_\Phi/\lambda_{\min}(\Sigma)$. Substituting:

$$D_{\mathrm{TV}}(d_\Psi \parallel d_{\boldsymbol{\theta}}) \le \frac{\sqrt{M_\Phi}\,T}{\sqrt{\lambda_{\min}(\Sigma)}} \cdot \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}\|_2.$$

Combining with the earlier spectral norm bound:

$$\|F(\boldsymbol{\theta}) - F_\Psi\|_2 \le \frac{2M_\Phi}{\lambda_{\min}(\Sigma)} \cdot \frac{\sqrt{M_\Phi}\,T}{\sqrt{\lambda_{\min}(\Sigma)}} \cdot \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}\|_2 = \frac{2M_\Phi^{3/2}\,T}{\lambda_{\min}(\Sigma)^{3/2}} \cdot \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}\|_2.$$

□

# D Condition Number and Gradient Alignment

We characterize how the condition number of the FIM controls the alignment between the vanilla policy gradient and the natural policy gradient. Let $F(\boldsymbol{\theta}) \in \mathbb{R}^{d_\Theta \times d_\Theta}$ be symmetric positive definite with eigenvalues $\lambda_{\min} \le \cdots \le \lambda_{\max}$, and let $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \neq 0$. The vanilla gradient is $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ and the natural gradient is $\widetilde{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = F(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. Their alignment is measured by the cosine similarity

$$\cos(\vartheta) = \frac{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})^\top F(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\|\,\|F(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\|}.$$

**Lemma 8.** *The minimum of* $\cos(\vartheta)$ *over all* $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \neq 0$ *is achieved by a vector lying in the two-dimensional subspace spanned by the eigenvectors associated with* $\lambda_{\min}$ *and* $\lambda_{\max}$.

*Proof.* Since $F(\boldsymbol{\theta})$ admits an eigendecomposition $F(\boldsymbol{\theta}) = Q\Lambda Q^\top$ with $Q$ orthogonal, let $x = Q^\top \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. Then:

$$\cos(\vartheta) = \frac{\sum_i x_i^2/\lambda_i}{\sqrt{\sum_i x_i^2}\,\sqrt{\sum_i x_i^2/\lambda_i^2}}.$$

The expression is homogeneous in $x$, so we set $\sum_i x_i^2 = 1$ and let $w_i = x_i^2 \ge 0$ with $\sum_i w_i = 1$. Setting $s_i := 1/\lambda_i \in [\alpha, \beta]$ with $\alpha := 1/\lambda_{\max}$ and $\beta := 1/\lambda_{\min}$, we have $\cos(\vartheta) = m/\sqrt{v}$ where $m = \sum_i w_i s_i$ and $v = \sum_i w_i s_i^2$ are the first and second moments of $S \sim w$.

For every $s \in [\alpha, \beta]$, $(s-\alpha)(s-\beta) \le 0$, which gives $s^2 \le (\alpha+\beta)s - \alpha\beta$. Taking expectations: $v \le (\alpha+\beta)m - \alpha\beta$, with equality if and only if $S \in \{\alpha, \beta\}$ almost surely. Therefore:

$$\cos(\vartheta)^2 = \frac{m^2}{v} \ge \frac{m^2}{(\alpha+\beta)m - \alpha\beta} =: \psi(m), \quad m \in [\alpha, \beta].$$

A direct computation shows $\psi'(m) = 0$ at $m_* = 2\alpha\beta/(\alpha+\beta)$, giving:

$$\cos(\vartheta)^2 \ge \psi(m_*) = \frac{4\alpha\beta}{(\alpha+\beta)^2}, \qquad \cos(\vartheta) \ge \frac{2\sqrt{\alpha\beta}}{\alpha+\beta}.$$

Equality requires support only on $\{\alpha, \beta\}$, i.e., the gradient lies in the span of the eigenvectors at $\lambda_{\min}$ and $\lambda_{\max}$. □

The cosine similarity between the vanilla and natural policy gradients satisfies:

$$\cos(\vartheta) \geq \frac{2\sqrt{\lambda_{\min}\lambda_{\max}}}{\lambda_{\min} + \lambda_{\max}} = \frac{2\sqrt{\kappa}}{1+\kappa}.$$

*Proof.* By Lemma 8, the minimum cosine similarity is:

$$\cos(\vartheta) \geq \frac{2\sqrt{\alpha\beta}}{\alpha+\beta} = \frac{2\sqrt{1/(\lambda_{\max}\lambda_{\min})}}{1/\lambda_{\max} + 1/\lambda_{\min}} = \frac{2\sqrt{\lambda_{\min}\lambda_{\max}}}{\lambda_{\min} + \lambda_{\max}}.$$

Dividing numerator and denominator by $\lambda_{\min}$ and substituting $\kappa = \lambda_{\max}/\lambda_{\min}$:

$$\cos(\vartheta) \geq \frac{2\sqrt{\kappa}}{1+\kappa}.$$

□

Theorem 4 shows that a well-conditioned FIM (small $\kappa$) guarantees strong alignment between the vanilla and natural gradients: when $\kappa = 1$ the bound equals 1, meaning both directions coincide exactly. Conversely, a large condition number can make the vanilla gradient an arbitrarily poor proxy for the natural gradient. This provides a concrete geometric motivation for our compression framework: by constructing a latent space in which the empirical FIM is close to a scaled identity, Euclidean gradient steps in that space are well-aligned with the natural gradient directions.

# E Experimental Settings

In this section, we present the experimental settings of the empirical results presented in this work.

## E.1 Linear Quadratic Regulator

We evaluate the linear construction on a Linear Quadratic Regulator (LQR) task with a state space of dimension 6 and an action space of dimension 3. The environment dynamics are governed by transition matrices $A \in \mathbb{R}^{6\times 6}$ and $B \in \mathbb{R}^{6\times 3}$. The state transition matrix $A$ is defined as:

$$A = \begin{bmatrix} 0.998 & 0.22 & 0.08 & 0 & 0 & 0 \\ 0 & 0.97 & 0.18 & 0.07 & 0 & 0 \\ 0 & 0 & 0.90 & 0.16 & 0.06 & 0 \\ 0 & 0 & 0 & 0.72 & 0.12 & 0 \\ 0 & 0 & 0 & 0 & 0.45 & 0.10 \\ 0 & 0 & 0 & 0 & 0 & 0.20 \end{bmatrix}.$$

The control matrix $B$ is generated by sampling its entries from a standard normal distribution $\mathcal{N}(0,1)$. The three columns are then scaled by factors of 1.0, 0.08, and 0.01 respectively, and the entire matrix is normalized by $1/\sqrt{3}$. The reward is defined by the quadratic cost matrices $Q$ and $R$, given by:

$$Q = \mathrm{diag}(35.0, 10.0, 3.0, 0.9, 0.2, 0.05), \quad R = 0.05I,$$

where $I$ is the identity matrix.

### E.2 Policy Embedding and Compression

The policy embedding and subsequent compression are executed in two distinct phases. First, we evaluate the parameters of a behaviorally diverse policy bank over a set of reference states to build the Fisher-whitened embedding matrix $A_\Psi$. The embedding of any policy parameter vector $\theta_i$ is then computed as $\phi_i = A_\Psi \theta_i$.

In the second phase, we perform PCA to isolate the informative low-dimensional subspace. In particular, we retain only the dimension that accounts for $99\%$ of the dataset's variance. Then, we compute the empirical mean of the embedded policy bank, $\bar{\phi}_\Psi$, and center the embeddings. Applying Singular Value Decomposition (SVD) on these centered embeddings yields the orthogonal projection matrix $\mathbf{U} \in \mathbb{R}^{(md_\mathcal{A}) \times k}$ comprising the top $k$ principal directions. This operation minimizes the Fisher-weighted reconstruction error and allows any policy parameter $\theta$ to be compressed into a latent representation $\mathbf{z}$:

$$\mathbf{z} = \mathbf{U}^\top (A_\Psi \theta - \bar{\phi}_\Psi)$$

Consequently, the full parameter vector can be reconstructed from the latent space using the Moore-Penrose pseudoinverse $A_\Psi^\dagger$:

$$\hat{\boldsymbol{\theta}} = A_\Psi^\dagger (\bar{\phi}_\Psi + \mathbf{U}\mathbf{z})$$

This ensures that the retained components correctly align with the steepest behavioral variations defined by the local Fisher geometry. This geometric alignment provides the foundation for the Compressed Policy Gradient (CPG) algorithm, which leverages the low-dimensional latent space to perform projected natural gradient ascent via standard Euclidean optimization.

### E.3 Neural Policy Compression

For the neural experiments, we employ the implementation provided by (Tenedini et al., 2026). We evaluate our framework on the Continuous Mountain Car environment. The policies are parameterized by a Multi-Layer Perceptron with a single hidden layer of 8 units, ELU activations, and a Tanh output layer, resulting in $d_\Theta = 33$ parameters. We generate an initial dataset of 5000 policies by sampling weights uniformly in the range $[-2.5, 2.5]$. The bank is later cut down to the top 20% most diverse policies. The symmetric Autoencoder compresses the parameters into latent dimensions $k \in \{2, 3, 5, 10\}$. The encoder uses hidden layers of sizes 25 and 10, and the decoder mirrors this architecture. We train the networks using the Mean Squared Error loss, a learning rate of $10^{-4}$, and a batch size of 64 for 75 epochs. For a parameter vector $\boldsymbol{\theta}$, we compute the parameter space FIM $F(\boldsymbol{\theta})$. The latent space FIM $F(z)$ is then obtained by projecting $F(\boldsymbol{\theta})$ through the Jacobian of the trained decoder.