COMMUNICATION-EFFICIENT FEDERATED NATURAL POLICY GRADIENT WITH PROVABLE CONVERGENCE PROPERTIES

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies communication-efficient federated reinforcement learning for closed-loop control of dynamical systems. In our setup, multiple agents learn a common policy through a central server without sharing raw trajectory data. Standard federated natural policy gradient (FedNPG) requires each agent to transmit both a policy gradient and a Fisher information matrix (FIM), creating a substantial communication burden as the policy dimension grows. We propose FedNPG-ED, which reduces this burden by compressing each local FIM through eigendecomposition (ED). Each agent transmits only q_k dominant eigenpairs and a scalar parameter p_k that approximates the discarded eigenvalues. The server reconstructs the local curvature approximations, aggregates them using sample-based weights, and updates the shared policy. In the linear quadratic regulator (LQR) setting, our analysis characterises linear convergence to a neighbourhood of the global optimum and conditions for maintaining closed-loop stability throughout learning. Our analysis for the LQR setup separates the roles of the two design parameters q_k and p_k. The contraction factor for convergence and the residual error depend on the step size and p_k only through their ratio, so that p_k is chosen as small as the step-size condition permits, whereas q_k fixes the per-round payload. Simulations on linear and nonlinear systems with linear and neural actors demonstrate the method's practical performance. With unequal local sample sizes, FedNPG-ED closely matches uncompressed FedNPG while transmitting substantially less data and outperforms an alternating direction method of multipliers-based FedNPG baseline at comparable communication cost.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has emerged as a powerful framework for decision-making in large-scale networked dynamical systems, with applications including IoT devices (Uprety & Rawat, 2020), autonomous vehicles (Shalev-Shwartz et al., 2016), and distributed cyber-physical systems (Dai et al., 2020). Training such agents requires large volumes of data (Ota et al., 2024), which are typically distributed among agents. Sending locally collected data to a central server (CS) is often infeasible because of limited bandwidth, latency requirements, and privacy concerns (Mo et al., 2024), which calls for distributed learning and control algorithms in which agents collaborate without sharing raw data and with minimal communication overhead.

Federated learning (FL) (McMahan et al., 2017; Elgabli et al., 2022; Dal Fabbro et al., 2024) provides such a paradigm, where agents compute on their local data and share only gradients or local model parameters with the CS. Since communication is often the main bottleneck, communication efficiency has become a central design objective. The communication overhead is particularly acute for second-order methods, which exploit curvature information to achieve faster convergence but require exchanging quantities, such as the Hessian or Fisher information matrix (FIM), whose size grows quadratically with the model dimension (Wang et al., 2017; Dal Fabbro et al., 2024).

Extending FL to control and reinforcement learning (RL) is more challenging than applying it to supervised learning because data are generated sequentially through interaction with a dynamical system and their distribution depends on the evolving policy (Bertsekas, 2019). Additional challenges arise from non-i.i.d. data, closed-loop stability requirements, and model uncertainty. Among RL methods, the actor–critic (AC) framework is widely used in control due to its ability to handle continuous state and action spaces (Konda & Tsitsiklis, 1999; Yang et al., 2019; Wang et al., 2020). For the actor update, the natural policy gradient (NPG) uses the FIM to account for the geometry of the policy space, which can improve convergence compared to the Euclidean gradient updates

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.

(Kakade, 2001). However, in a federated setting, exchanging the FIM introduces a substantial communication burden. This motivates our study of a communication-efficient federated NPG algorithm with a favourable convergence–communication trade-off.

1.1 Related Work

Federated and communication-efficient RL: FedRL combines the experience of multiple agents while supporting data privacy and improving sample efficiency. Existing work focuses mainly on learning a shared value function (Zhang et al., 2025) or optimising a shared policy (Xie & Song, 2023; Wang et al., 2024; Zhu et al., 2024; Zhang et al., 2024; Labbi et al., 2025; Lan et al., 2025) across Markov decision processes (MDPs). Efforts to reduce communication overhead (Agarwal et al., 2021; Ornia & Mazo, 2022; Gatsis, 2022; Lan et al., 2023; Chen et al., 2021) often come with trade-offs in learning performance, with communication costs increasing with the size of the state–action space and the number of agents (Agarwal et al., 2021). The event-triggered method of Ornia & Mazo (2022) introduces sampling bias, causing the Q-function to converge to a different fixed point. The approach of Gatsis (2022) addresses only value-function learning and provides guarantees for a single iteration. The alternating direction method of multipliers (ADMM)-based method of Lan et al. (2023) requires solving a local linear system involving a Hessian-like matrix at each iteration, and its performance deteriorates when agents have heterogeneous local data. The method of Chen et al. (2021) requires large batches and continuous monitoring. These communication-efficient approaches assume finite MDPs and do not address closed-loop stability. Moreover, Chen et al. (2021); Lan et al. (2023) establish only convergence to stationary points or sublinear rates, whereas we establish a linear convergence rate.

Convergence of standalone PG and NPG: These properties have been studied extensively in the linear quadratic regulator (LQR) setting, where gradient dominance and coercivity underpin global convergence and closed-loop stability despite the non-convexity of the objective in the policy parameters (Fazel et al., 2018; Hu et al., 2023). NPG-type updates further achieve linear convergence rates (Fazel et al., 2018; Kakade, 2001). Extensions include two-timescale and single-timescale AC methods (Yang et al., 2019; Zhou & Lu, 2023; Chen et al., 2025), as well as finite-time analyses (Wu et al., 2020; Tian et al., 2023). Critic convergence is typically analysed using stochastic approximation (Borkar & Borkar, 2008; Dalal et al., 2018; Mitra, 2024). These results concern the single-agent setting, while convergence guarantees for federated NPG with compressed second-order information remain largely unexplored.

Second-order methods in FL: Communication-efficient second-order schemes include GIANT (Wang et al., 2017), DONE (Dinh et al., 2022), sub-sampling (Byrd et al., 2011; Erdogdu & Montanari, 2015), and SHED (Dal Fabbro et al., 2024), which compresses the local Hessian into a few dominant eigenvalue–eigenvector pairs (EEPs). Our method is motivated by SHED, but SHED targets supervised learning with static datasets and its analysis relies on (strongly) convex losses, so its guarantees do not carry over to FedRL, where the problem is non-convex, the data distribution evolves with the policy, and closed-loop stability must be maintained. Two differences in the mechanism are also worth noting: SHED transmits eigenpairs *incrementally*, reusing those sent in earlier rounds, whereas we re-transmit q_k fresh pairs each round, and the FIM compressed here is generated by the current policy rather than fixed.

Compression and federated control. K-FAC (Martens & Grosse, 2015) offers an alternative to spectral truncation by approximating the FIM with smaller Kronecker factors. Gradient quantization (Alistarh et al., 2017) and sparsification with error feedback (Karimireddy et al., 2019) complement second-order compression. Local updates between aggregations aim to reduce communication rounds (McMahan et al., 2017; Karimireddy et al., 2020), whereas our method reduces the per-round payload. Federated system identification and LQR methods (Wang et al., 2023b;a) address heterogeneous dynamics, but leave curvature information uncompressed. Finally, Khodadadian et al. (2022) establish a linear speed-up under Markovian sampling.

1.2 Our Contributions

We leverage NPG for its convergence properties and adopt a communication-efficient strategy inspired by SHED (Dal Fabbro et al., 2024). At each round, every agent eigendecomposes its local FIM and transmits only the top-q_k EEPs together with the gradient and an approximation parameter

p_k that replaces the remaining eigenvalues. The CS reconstructs each local curvature matrix and updates the global policy by the NPG rule. To the best of our knowledge, this is the first application of such a scheme to FedRL together with a convergence analysis. We study the proposed FedNPG-ED algorithm under the LQR setting with a linear actor, which is a standard benchmark in control (Bertsekas & Tsitsiklis, 1996). Our contributions are as follows.

1. We propose FedNPG-ED, which compresses each local FIM by eigendecomposition and transmits q_k dominant EEPs together with p_k, giving explicit control over communication overhead and approximation error.
2. For the LQR problem, we show that the method preserves the linear convergence rate of vanilla FedNPG and characterise conditions under which the intermediate controllers remain stabilising during training.
3. We separate the roles of the two design parameters p_k and q_k. The contraction factor affecting the convergence speed and the residual error depend on the step size and the approximation parameter only through the ratio $\widehat{s}_k = \eta_k/p_k$, so p_k acts as a reparametrisation of the learning rate and is taken as small as the step-size condition permits (see Theorem 1). On the other hand, the truncation level q_k enters neither, and instead fixes the per-round payload and the fidelity of the transmitted curvature. We further show that the asymptotic error floor depends on neither parameter and is set by the sample budget alone (Remark 2).
4. Through simulations on linear and nonlinear systems, we show that the method outperforms ADMM-based FedNPG when the agents hold unequal numbers of local samples, at comparable communication overhead.

Scope and limitations. The analysis covers LQR with a linear Gaussian actor and homogeneous dynamics. The only source of heterogeneity addressed is the variation in local sample sizes. Additionally, the critic is treated under the assumption of an explicit oracle (refer to Assumption 4.5), rather than through an analysis of its own recursion. It is important to note that we do not claim any formal privacy guarantees, as the transmitted eigenvectors can reveal information about the local state distribution, even though no raw trajectories are shared.

2 Problem Statement

We consider a FedRL setting in which N agents, each interacting with its own environment, collaboratively learn a common optimal control policy through a CS without sharing raw data. The underlying optimisation problem is non-convex, and the standard FedNPG solution requires each agent to transmit its local FIM to the CS, incurring substantial overhead. Our goal is to design a FedNPG algorithm that reduces this overhead while retaining the convergence and stability guarantees of the vanilla method.

2.1 Background

MDP. We consider an infinite-horizon MDP $\langle \mathcal{X}, \mathcal{U}, \mathcal{P}, r \rangle$, where $\mathcal{X} \in \mathbb{R}^n$ is the state space, $\mathcal{U} \in \mathbb{R}^p$ the action space, $\mathcal{P}(x_t'|x_t, u_t)$ the transition kernel, and $r\colon \mathcal{X}\times\mathcal{U} \to \mathbb{R}$ the reward. At time t an agent takes $u_t \sim \pi_\theta(\cdot \mid x_t)$, with π_θ a stochastic policy parameterised by θ. We adopt the infinite-horizon average expected reward formulation, which is widely used in the control literature (Bertsekas & Tsitsiklis, 1996; Bertsekas, 2019),

$$\mathcal{R} = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_{x\sim\rho_x, u\sim\pi_\theta} \left[\sum_{t=1}^{T} r_t \right], \tag{1}$$

assuming the states attain a stationary distribution ρ_x under π_θ, with $r_t = r(x_t, u_t)$. The corresponding value and state-action value functions $V_{\pi_\theta}(x)$ and $Q_{\pi_\theta}(x, u)$ are defined in Appendix C. With slight abuse of notation, we write $V_k(\cdot)$ and $Q_k(\cdot, \cdot)$ for their values at the k-th training iteration, which also represents the aggregation round.

PG and NPG. By the policy gradient theorem (Sutton et al., 1999), $\nabla_\theta \mathcal{R} = \mathbb{E}_{\pi_\theta}[Q(x_t, u_t)\nabla_\theta \log \pi_\theta(u_t|x_t)]$. The sample-based estimate at iteration k is given as

$$\widehat{g}_k = \frac{1}{\tau_k} \sum_{t=1}^{\tau_k} \widehat{Q}_k\left(x_t, u_t\right) \nabla_\theta \log \pi_{\theta_k}\left(u_t|x_t\right), \tag{2}$$

where $\widehat{Q}_k$ is an estimate of true Q_k and τ_k is the number of samples. While PG follows steepest ascent in the Euclidean parameter space, NPG (Kakade, 2001) uses the FIM as a Riemannian metric, giving updates invariant to the parametrisation. Since the true FIM is unavailable in the model-free setting, it is replaced by

$$\widehat{\mathcal{F}}_k = \frac{1}{\tau_k} \sum_{t=1}^{\tau_k} \nabla_\theta \log \pi_{\theta_k}(u_t|x_t) \, \nabla_\theta \log \pi_{\theta_k}(u_t|x_t)^\top. \tag{3}$$

Therefore, the natural ascent direction is $\widehat{\mathcal{F}}_k^{-1} \widehat{g}_k$ rather than $\widehat{g}_k$; see Appendix C, which provides further details on the background of the problem under study.

2.2 Federated RL Problem Setup

We consider time-invariant dynamical systems comprising N agents,

$$x_{t+1}^{(i)} = f^{(i)}\big(x_t^{(i)}, u_t^{(i)}, w_t^{(i)}\big), \ \forall i \in \{N\}, \tag{4}$$

with $x_0^{(i)} \sim \mathcal{D}$ and i.i.d. process noise $w_t^{(i)} \sim \mathcal{D}_w$. We consider homogeneous systems, $f^{(i)} = f^{(j)}$, to focus our study on communication efficiency. Agents differ only in their local sample sizes $\tau_{i,k}$ and noise and initial-state realizations, yielding different local estimates $\widehat{g}_k^{(i)}$ and $\widehat{\mathcal{F}}_k^{(i)}$. This variability concerns local estimators rather than system dynamics or client drift from multiple local updates, as Algorithm 1 performs one local computation per round. Throughout, heterogeneity refers specifically to unequal local sample sizes.

An important special case of (4) is the linear time-invariant (LTI) system

$$x_{t+1}^{(i)} = \mathbf{A} x_t^{(i)} + \mathbf{B} u_t^{(i)} + w_t^{(i)}, \tag{5}$$

with $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$, which we adopt for the convergence analysis in Section 4 and which is extensively studied in the RL literature (Fazel et al., 2018; Hu et al., 2023; Yang et al., 2019; Chen et al., 2025). The expected average return over all agents is $\mathcal{R}_{\text{avg}} := \sum_{i=1}^N \omega_{a,k}^{(i)} \mathcal{R}^{(i)}$ subject to (4), where $\mathcal{R}^{(i)}$ is the return of agent i with one-step reward $r_t^{(i)}$ as in (1). In the LQR setting used in Section 4,

$$r_t^{(i)} = -\big(x_t^{(i)\top} \mathscr{Q} x_t^{(i)} + u_t^{(i)\top} \mathscr{R} u_t^{(i)}\big), \tag{6}$$

with $\mathscr{Q} \succ 0$ and $\mathscr{R} \succ 0$. The federated objective is for all agents to learn a globally optimal common policy maximizing $\mathcal{R}_{\text{avg}}$.

Federated natural policy gradient. At iteration k, agent i computes $\widehat{g}_k^{(i)}$ and $\widehat{\mathcal{F}}_k^{(i)}$ from $\tau_{i,k}$ local samples as in (2) and (3) and transmits them to the CS, which aggregates them by the FedAvg weights (McMahan et al., 2017)

$$\omega_{a,k}^{(i)} = \frac{\tau_{i,k}}{\sum_{i=1}^N \tau_{i,k}}, \qquad \widehat{g}_k = \sum_{i=1}^N \omega_{a,k}^{(i)} \widehat{g}_k^{(i)}, \qquad \widehat{\mathcal{F}}_k = \sum_{i=1}^N \omega_{a,k}^{(i)} \widehat{\mathcal{F}}_k^{(i)}, \tag{7}$$

and updates the global parameter along the natural ascent direction,

$$\theta_{k+1} = \theta_k + \eta_k \, \widehat{\mathcal{F}}_k^{-1} \, \widehat{g}_k, \tag{8}$$

where η_k is the learning rate. Then θ_{k+1} is broadcast to all agents. Transmitting the full FIM $\widehat{\mathcal{F}}_k^{(i)}$ every iteration incurs substantial overhead, which motivates the design proposed next. We assume full agent participation in each communication round since our focus is on communication efficiency in this study. Extending the analysis to agent dropouts is a direction for our future work.

3 Proposed Communication-Efficient FedNPG: FedNPG-ED

The standard FedNPG update (8) requires each agent to transmit its local $(d_\theta \times d_\theta)$ FIM every round, where $d_\theta = \dim(\theta_k)$. We instead transmit only the $q_k^{(i)}$ dominant EEPs together with a scalar $p_k^{(i)}$ and the gradient from agent-i to the CS, reducing the per-round communication complexity to $\mathcal{O}(d_\theta(q_k + 1) + 1)$. In other words, in round k, the CS broadcasts θ_k, then each agent collects $\tau_{i,k}$

trajectories, computes $\widehat{g}_k^{(i)}$ and $\widehat{\mathcal{F}}_k^{(i)}$, and sends $\{\widehat{g}_k^{(i)}, p_k^{(i)}, \{\widehat{\lambda}_{b,k}^{(i)}, \widehat{v}_{b,k}^{(i)}\}_{b=1}^{q_k}\}$ to the CS. Finally, the CS reconstructs each $\widehat{\mathcal{F}}_{\mathrm{ed},k}^{(i)}$, aggregates, and performs the global update. The complete algorithm, FedNPG-ED, is given in Algorithm 1 in Appendix E.

ED-based curvature approximation: For each agent i, the ED-based approximation of the FIM is constructed at the CS using the transmitted dominant EEPs and scalar, as follows

$$\widehat{\mathcal{F}}_{\mathrm{ed},k}^{(i)} = \sum_{b=1}^{q_k} \big(\lambda_{b,k}^{(i)} - p_k^{(i)}\big) v_{b,k}^{(i)} v_{b,k}^{(i)\top} + p_k^{(i)}\, I = \sum_{b=1}^{q_k} \tilde{v}_{b,k}^{(i)} \tilde{v}_{b,k}^{(i)\top} + p_k^{(i)} I, \tag{9}$$

where the eigenvalues of $\widehat{\mathcal{F}}_k^{(i)}$ are ordered $\lambda_{1,k}^{(i)} \geq \cdots \geq \lambda_{d_\theta,k}^{(i)}$ with eigenvectors $v_{b,k}^{(i)}$, the remaining $(d_\theta - q_k)$ eigenvalues are replaced by $p_k^{(i)} > 0$, and $\tilde{v}_{b,k}^{(i)} = (\widehat{\lambda}_{b,k}^{(i)} - p_k^{(i)})^{1/2} v_{b,k}^{(i)}$ are the scaled eigenvectors actually transmitted. I denotes the identity matrix of size d_θ. Note that at $q_k = d_\theta$, $p_k^{(i)} = 0$, (9) reduces to (3).

Global update rule: The policy parameter is updated using the ED-based curvature matrix, and η_k is the learning rate,

$$\theta_{k+1} = \theta_k + \eta_k\, \widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1}\, \widehat{g}_k, \;\; \text{where} \;\; \widehat{\mathcal{F}}_{\mathrm{ed},k} = \sum_{i=1}^{N} \omega_{a,k}^{(i)} \widehat{\mathcal{F}}_{\mathrm{ed},k}^{(i)}, \tag{10}$$

Q-function. Evaluating (2) requires an estimate $\widehat{Q}_k$. For our simulations, the critic parameters ϕ_k are trained by minimising the least-squares Bellman error using a Newton-like method (Schulman et al., 2015). Each agent sends its local critic gradient $\widehat{h}_k^{(i)}$ and approximate Hessian $\widehat{\mathcal{H}}_k^{(i)}$ to the CS for aggregation using the FedAvg weights. We use the same critic across all compared algorithms to isolate the effect of communication-efficient FedNPG updates. Since our focus is communication-efficient FedNPG, we omit further critic details and refer to Schulman et al. (2015) for the underlying method. A joint actor–critic analysis is left for future work.

Communication accounting. We count uplink scalars per agent per policy update, excluding downlink broadcasts and metadata. Per agent and policy update, the actor uplink payload is $\big(d_\theta(q_k+1)+1\big)$ for FedNPG-ED, counting the gradient, the q_k transmitted vectors, and the scalar p_k; the eigenvalue need not be sent separately, since the scaled eigenvector $\tilde{v}_{b,k}^{(i)}$ of (9) carries it without any loss. On the other hand, FedNPG needs to transmit $(d_\theta + d_\theta(d_\theta+1)/2)$ scalars, counting the gradient and independent FIM entries. The ADMM (Lan et al. (2023)) baseline transmits $(2d_\theta)$ scalars, where the gradient is sent along with a d_θ-dimensional directional vector at each communication round.

Computational complexity. Compression shifts cost from communication to local computation. A full eigendecomposition of a dense $d_\theta \times d_\theta$ FIM costs $\mathcal{O}(d_\theta^3)$ operations, but FedNPG-ED needs only the $q_k \ll d_\theta$ dominant EEPs, which restarted Lanczos or randomized subspace methods approximate at far lower cost (Lehoucq et al., 1998; Halko et al., 2011). The saving is largest when the FIM is never formed. From (3), FIM is the Gram matrix $\tau_{i,k}^{-1} \Psi_k^{(i)\top} \Psi_k^{(i)}$ of the score matrix $\Psi_k^{(i)} \in \mathbb{R}^{\tau_{i,k} \times d_\theta}$, so its dominant eigenpairs are the dominant right singular pairs of $\Psi_k^{(i)}$ and are obtained from FIM–vector products alone in $\mathcal{O}(d_\theta \tau_{i,k} q_k)$ operations, linear rather than cubic in d_θ. Warm-starting from the previous round's eigenpairs may reduce this further. Additionally, convergence of the iterative methods depends on the spectrum and the target accuracy.

4 Theoretical Results

We analyse the convergence and closed-loop stability of FedNPG-ED in the LQR setting, whose tractable structure allows us to quantify the effects of curvature compression on learning and control. Note that we have used the LQR setup for theoretical analysis only; the proposed algorithm works for nonlinear systems with neural AC architectures, as demonstrated in our numerical simulations. In our setup, each agent evolves according to (5) with one-step reward (6), and following the standard LQR policy-optimisation convention (Fazel et al., 2018; Hu et al., 2023). Under this setup, we equivalently minimise the average cost $\mathcal{J}_{\mathrm{avg}} := -\mathcal{R}_{\mathrm{avg}}$.

Policy structure: We consider the stochastic Gaussian policy

$$u_t^{(i)} \sim \pi_\theta(\cdot \mid x_t^{(i)}) = \mathcal{N}\big(-Kx_t^{(i)}, \Sigma_\sigma\big), \quad i.e. \quad u_t^{(i)} = -Kx_t^{(i)} + \sigma_t^{(i)}, \;\; \sigma_t^{(i)} \sim \mathcal{N}\big(0, \Sigma_\sigma\big), \tag{11}$$

with common gain K and fixed exploration covariance Σ_σ, motivated by the linearity of the optimal LQR policy, since agent-specific $\Sigma_\sigma^{(i)}$ would add notation without further insight. The closed-loop dynamics are

$$x_{t+1}^{(i)} = \big(\mathbf{A} - \mathbf{B}K\big)x_t^{(i)} + \bar{w}_t^{(i)}, \qquad \bar{w}_t^{(i)} \sim \mathcal{N}\big(0, \Sigma_w^{(i)} + \mathbf{B}\Sigma_\sigma \mathbf{B}^\top\big), \tag{12}$$

and we write $\mathcal{K} := \{K : \rho(\mathbf{A} - \mathbf{B}K) < 1\}$ for the stabilising set and K_* for the globally optimal gain.

Assumption 4.1. *The pair $(\boldsymbol{A}, \boldsymbol{B})$ is stabilisable, and $(\boldsymbol{A}, \mathscr{Q}^{1/2})$ is detectable.*

Assumption 4.2. *There exist $\mu_\xi, \mu_g > 0$ such that $\mathbb{E}_{x_0^{(i)} \sim \mathcal{D}}[x_0^{(i)} x_0^{(i)\top}] \succeq \mu_\xi I$ and $\|x_0^{(i)}\| \le \mu_g$ almost surely, $\forall i$.*

Assumption 4.3. *An initial common stabilising controller $K_0 \in \mathcal{K}$ is available.*

Assumption 4.4 (Ergodicity and bounded state)**.** *For every $K \in \mathcal{K}$, the closed-loop process (12) is ergodic and admits a unique stationary distribution $\rho_{x,K}$, to which it converges from any initial condition. Moreover, the physical constraints of the plant confine the states to a compact set, so that $\|x_t^{(i)}\| \le c_x < \infty$ $\forall t$ and all $i \in \{N\}$.*

Assumption 4.5 (Oracle critic)**.** *The critic returns the exact action-value function of the current policy, that is, $\widehat{Q}_k(x,u) = Q_{K_k}(x,u)$ for every (x,u) and every k.*

Assumptions 4.1–4.3 are standard for LQR policy optimisation (Fazel et al., 2018; Hu et al., 2023). Assumption 4.4 ensures a unique stationary state distribution and uniformly bounded states. Assumption 4.5 provides exact policy evaluation, eliminating critic approximation error and allowing the analysis to focus on actor updates and curvature compression. The critic recursion is therefore kept outside the scope of the analysis.

Theorem 1 (Linear convergence of FedNPG-ED)**.** *Let Assumptions 4.1–4.3 hold. Consider the proposed FedNPG-ED algorithm (Algorithm 1 in Appendix E) with the actor given in (11). Let $p_k^{(i)} = p_k > 0$ and the step size η_k satisfies*

$$0 < \frac{\eta_k}{p_k} < \min\left\{\frac{1}{\mathcal{L}}, \frac{\|\Sigma_{K_*}\|}{4\mu_\xi^2\, \lambda_{\min}(\mathscr{R})}\right\}, \tag{13}$$

where $\mathcal{L}$ is the smoothness constant of $\mathcal{J}_{\text{avg}}$. Then, at the k-th iteration for any $\varepsilon_{g_1} > 0$, with probability at least $(1 - \delta_g)$,

$$\mathcal{J}_{\text{avg}}(K_{k+1}) - \mathcal{J}_{\text{avg}}(K_*) \le \left(1 - \widehat{\beta}\right)\left(\mathcal{J}_{\text{avg}}(K_k) - \mathcal{J}_{\text{avg}}(K_*)\right) + \widehat{\varepsilon}_L, \tag{14}$$

where the contraction factor $\widehat{\beta} \in (0,1)$ and the residual error $\widehat{\varepsilon}_L$ are given by

$$\widehat{\beta} := \frac{4\mu_\xi^2\, \lambda_{\min}(\mathscr{R})}{\|\Sigma_{K_*}\|}\left(\frac{\eta_k}{p_k} - \frac{\mathcal{L}}{2}\frac{\eta_k^2}{p_k^2}\right), \qquad \widehat{\varepsilon}_L := r\left(\frac{\eta_k}{p_k} - \frac{\mathcal{L}}{2}\frac{\eta_k^2}{p_k^2}\right)\varepsilon_{g_1}^2. \tag{15}$$

Here, ε_{g_1} bounds the deviation of the sample-based policy gradient (G'_{K_k}) from its exact counterpart G_{K_k}, i.e., $\|G_{K_k} - G'_{K_k}\| \le \varepsilon_{g_1}$, which holds with probability at least $(1 - \delta_g)$ with

$$\delta_g := (n+p)\exp\left(\frac{-\sum_{i=1}^N \tau_{i,k}\, \varepsilon_{g_1}^2/2}{\sigma_Q^2 + Q_{\max}\, \varepsilon_{g_1}/3}\right), \tag{16}$$

where σ_Q^2 and $Q_{\max}$ denote the matrix variance proxy and an almost-sure bound on the spectral norm of the centred score-weighted Q-function (that is, the centred stochastic policy-gradient contribution), and $r := \operatorname{rank}\big(G_{K_k}\big) \le \min\{n,p\}$ is the rank of the policy-gradient matrix.

Proof. The proof of Theorem 1 is provided in Appendix G.3. ■

Remark 1 (Step-size conditions). The bound $1/\mathcal{L}$ in (13) may be relaxed to $2/\mathcal{L}$ without affecting the convergence, since $\widehat{\beta} \in (0,1)$ still holds. The tighter bound is what makes p_k a usable tuning parameter. In other words, under (13), $\widehat{\beta}$ increases monotonically as p_k decreases, whereas beyond $1/\mathcal{L}$ a further reduction of p_k slows the certified rate. Shrinking p_k is nevertheless limited by (13) itself, which $p_k \le \mathcal{L}\eta_k/2$ violates, see Appendix G.5 for more details.

Remark 2 (Effect of the approximation parameter $p_k^{(i)} = p_k$ and the truncation level $q_k^{(i)} = q_k$). For clarity, we will use the same p_k and q_k across all agents, allowing us to gain better insight into their impact on the learning dynamics without unnecessary notational complexity. Let

$$\widehat{s}_k := \frac{\eta_k}{p_k}, \qquad f(\widehat{s}_k) := \widehat{s}_k - \frac{\mathcal{L}}{2}\widehat{s}_k^2, \qquad c := \frac{4\mu_\xi^2\,\lambda_{\min}(\mathscr{R})}{\|\Sigma_{K_*}\|}. \tag{17}$$

Therefore, using (13) and (15), we can rewrite $\widehat{\beta} = c\,f(\widehat{s}_k)$, $\widehat{\varepsilon}_L = r\,f(\widehat{s}_k)\,\varepsilon_{g_1}^2$, and $\widehat{s}_k \in (0, s^\star)$, where $s^\star := \min\{1/\mathcal{L}, 1/c\}$.

Two consequences follow, with derivations provided in Appendix G.5. *(i)* Both $\widehat{\beta}$ and $\widehat{\varepsilon}_L$ depend on (η_k, p_k) only through the ratio $\widehat{s}_k$, and f is strictly increasing and concave on $(0, s^\star)$, so $\partial\widehat{\beta}/\partial p_k < 0$ and $\widehat{\beta}$ can be increased by taking p_k as small as (13) permits, *i.e.*, up to the ceiling $\widehat{\beta} < c\,f(s^\star)$. However, shrinking p_k without limit is counter-productive, since $f(\widehat{s}_k) \le 0$ once $\widehat{s}_k \ge 2/\mathcal{L}$; too small a p_k over-relaxes the ascent direction along the eigen-subspace for which no curvature is transmitted, and (13) is exactly what excludes that regime. *(ii)* The truncation level q_k enters neither $\widehat{\beta}$ nor $\widehat{\varepsilon}_L$, because $\lambda_{\min}(\widehat{\mathcal{F}}_{\mathrm{ed},k}) = p_k$ for every q_k. q_k controls the payload, *i.e.*, $d_\theta(q_k+1)+1$ scalars per agent per round against $d_\theta + d_\theta(d_\theta+1)/2$ for the full FedNPG. In other words, q_k controls how many directions use the same curvature scaling $\widehat{\lambda}_{b,k}^{-1}$ as uncompressed FedNPG. The remaining directions are scaled uniformly by $1/p_k$. Increasing q_k therefore preserves more curvature information and reduces the distortion $D_k := \frac{\|\widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1} - \widehat{\mathcal{F}}_k^{-1}\|_2}{\|\widehat{\mathcal{F}}_k^{-1}\|_2} \le \frac{\lambda_{q+1,k}}{p_k} - 1$, see Appendix G.5. D_k is non-increasing in q_k and increasing as p_k decreases. Therefore, a higher p_k must be compensated with a larger q_k.

Guideline. Choose $p_k^{(i)}$ satisfying $0 < p_k^{(i)} \le \lambda_{d_\theta,k}^{(i)}$ and select η_k to satisfy the convergence conditions. For a prescribed distortion tolerance $D_{\max}$, choose the smallest q_k such that $\lambda_{q_k+1,k}^{(i)} \le (1+D_{\max})p_k^{(i)}$ for every agent. This retains only the eigenpairs needed to satisfy the local distortion bound. Thus, q_k controls the communication–fidelity trade-off, while p_k and η_k jointly determine the update scaling.

Corollary 1 (Unrolled convergence and sample complexity)**.** *Let the hypotheses of Theorem 1 hold with $\widehat{\beta}$ and $\widehat{\varepsilon}_L$ held constant over C communication rounds, write $\Delta_k := \mathcal{J}_{\mathrm{avg}}(K_k) - \mathcal{J}_{\mathrm{avg}}(K_*)$ for the optimality gap at round k, so that Δ_0 is the initial gap, and let δ_{f_1} denote the failure probability of the covariance concentration bound (77) in Appendix G.3. Then, with probability at least $1 - C(\delta_g + \delta_{f_1})$, simultaneously for all $k \le C$,*

$$\Delta_k \;\le\; \big(1-\widehat{\beta}\big)^k \Delta_0 + \Delta_\infty, \qquad \Delta_\infty := \widehat{\varepsilon}_L/\widehat{\beta} = r\,\varepsilon_{g_1}^2/c. \tag{18}$$

Hence for any $\epsilon > 0$ and $\delta \in (0,1)$, running $C(\epsilon) = \lceil \widehat{\beta}^{-1}\ln(2\Delta_0/\epsilon)\rceil$ rounds with $\varepsilon_{g_1}^2 = c\,\epsilon/(2r)$ and per-round budget $\sum_i \tau_{i,k} \ge 2(\sigma_Q^2 + Q_{\max}\varepsilon_{g_1}/3)\varepsilon_{g_1}^{-2}\ln[(3n+p)C(\epsilon)/\delta]$ gives $\Delta_{C(\epsilon)} \le \epsilon$ with probability at least $1-\delta$, the dimension factor $(3n+p)$ collecting the two matrix-Bernstein union bounds.

Proof. See Appendix G.4. ∎

The union bound is what turns the one-step contraction (14) into a statement about the whole trajectory. Otherwise, the per-round guarantee alone does not imply linear convergence over C rounds, since the favourable event must hold at every round simultaneously.

Closed-loop stability during training: Stability follows from coercivity of the average cost. As shown in Appendix G.1, $\mathcal{J}_{\mathrm{avg}}(K) \to \infty$ as K approaches the boundary of $\mathcal{K}$ or as $\|K\| \to \infty$, so the cost acts as a barrier. The residual error makes the cost sequence non-monotone, so a claim of monotone decrease would be incorrect. On the event of Corollary 1, $(1-\widehat{\beta})^k \le 1$ gives the uniform bound $\mathcal{J}_{\mathrm{avg}}(K_k) \le \mathcal{J}_{\mathrm{avg}}(K_0) + \Delta_\infty =: k_\epsilon$ for all $k \le C$, so every K_k lies in the sublevel set $\mathcal{S}(k_\epsilon) := \{K \in \mathcal{K} \mid \mathcal{J}_{\mathrm{avg}}(K) \le k_\epsilon\}$, which is a compact subset of $\mathcal{K}$ by Corollary 2, and therefore remains stabilising with no projection or constraint enforcement. Note that the confidence degrades linearly in C, and a large initial gap Δ_0 enlarges $\mathcal{S}(k_\epsilon)$ and weakens the implied margin to the boundary of $\mathcal{K}$.

5 NUMERICAL RESULTS

We first evaluate the proposed algorithm on an LQR system and compare its performance with the FedNPG and FedNPG-ADMM (Lan et al. (2023)) baselines. We then extend the evaluation to non-linear benchmark systems using neural actor and critic architectures. The federation consists of N agents with randomly heterogeneous local sampling budgets $\tau_{i,k}$, reflecting the form of heterogeneity considered in our setting. Each configuration is run for C aggregation rounds using the same set of hyperparameter values, and repeated over S independent seeds. All performance curves report the mean and 95% confidence interval (CI) over S independent trials. Note that in our numerical study, we have used an adaptive learning rate commonly used for NPG (Rajeswaran et al., 2017) with a user-defined $\alpha_a > 0$ as $\eta_k = \sqrt{\alpha_a / \left(\widehat{g}_k^\top \widehat{\mathcal{F}}_k^{-1} \widehat{g}_k\right)}$. In FedNPG-ED, $\widehat{\mathcal{F}}_k$ is replaced by $\widehat{\mathcal{F}}_{\mathrm{ed},k}$.

5.1 LQR SYSTEM

We evaluate FedNPG-ED against FedNPG and communication efficient FedNPG-ADMM (Lan et al. (2023)) variant on a fourth-order UAV LTI model (Zhao et al., 2023). All three methods use the same architectures, *i.e.*, a linear actor and a fully connected neural critic. The complete parameters protocol is supplied in (23) and Table 3 in Appendix F.1. Note that for a linear actor with gain matrix $K \in \mathbb{R}^{p \times n}$, the parameter dimension is $d_\theta = np$. When exploiting the Kronecker structure, we compress the FIM to the $n \times n$ state covariance, so the relevant curvature dimension is n.

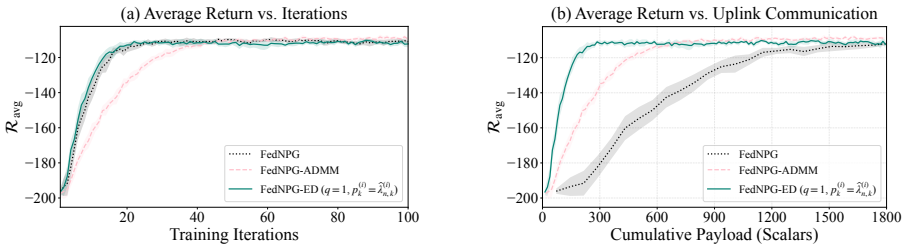

Figure 1: Learning (mean and 95% CI) and communication performance comparison for the linear control benchmark.

Figure 1(a) compares the learning behaviour of all methods against the training time steps or aggregation rounds, where $q_k = q = 1$ is fixed, and $p_k^{(i)} = \hat{\lambda}_{n,k}^{(i)}$ is a time-dependent and agent-dependent quantity for the FedNPG-ED. Note that $\hat{\lambda}_{n,k}^{(i)}$ denotes the smallest eigenvalue of agent i's local FIM at aggregation round k. Initialised with the same stabilising controller, all methods eventually approach a similar performance regime and maintain closed-loop stability throughout training. FedNPG and FedNPG-ED exhibit substantially faster convergence than FedNPG-ADMM. Note that, since $q = 1$ is used in this experiment, FedNPG-ED and the FedNPG-ADMM baseline transmit nearly the same number of scalars per round. Furthermore, even with a single retained eigenpair, FedNPG-ED achieves a return close to that of FedNPG. The small gap in the return is consistent with the approximation distortion D_k in (132) remaining small when the local curvature spectra decay rapidly, thereby preserving the dominant curvature directions in the transmitted spectral information.

Figure 1(b) plots the average return against cumulative actor-training uplink communication payload from one agent to the CS. All methods reach similar returns $\mathcal{R}_{\mathrm{avg}}$, but FedNPG-ED requires the fewest transmitted scalars, using only $q = 1$ and $p_k^{(i)} = \widehat{\lambda}_{n,k}^{(i)}$. These results demonstrate improved actor-side communication efficiency without an evident loss in performance in this experiment.

Additional numerical experiments with different choices of q and $p_k^{(i)}$, together with the corresponding analyses, are provided in Appendix D.1. In addition, Appendix D.2 reports results for the same LQR system using neural actor–critic architectures following the parameter protocol given in Table 4 in Appendix F.2. The larger neural actor increases the dimension of the FIM, providing a more demanding test of communication efficiency.

5.2 MUJOCO BENCHMARKS

Our convergence analysis focuses on LQR systems with a linear actor, for which the tractable structure permits explicit theoretical guarantees. The proposed compression scheme, as given in

Algorithm 1, also applies to FIMs generated by neural policies. We therefore evaluate its empirical performance beyond the theoretical setting on the nonlinear Hopper-v5 MuJoCo environment (Todorov et al., 2012), using neural actor–critic architectures. These experiments compare the learning performance and communication cost of FedNPG, FedNPG-ADMM, and FedNPG-ED against the high complexity benchmarks. Additionally, we performed a numerical study using a higher-dimensional Humanoid-v5 environment from the MuJoCo framework, and the results are provided in Appendix D.3. The details of the Hopper-v5 environment and the network configurations are summarised in Tables 5 and 6 in Appendix F.2.

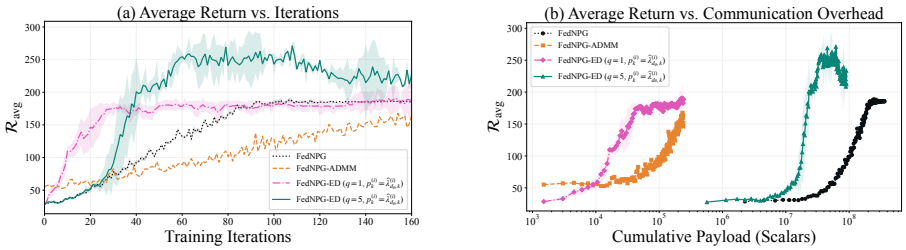

Figure 2: Learning (mean and 95% CI) and communication performance comparison for Hopper-v5 benchmark with neural actor.

Figure 2(a) provides $\mathcal{R}_{\text{avg}}$ as a function of training iterations. FedNPG-ED with $q = 1$ exhibits the fastest initial improvement and subsequently maintains a performance comparable to the FedNPG baseline, which requires considerably more iterations to reach a similar return. Increasing the retained eigenspace to $q = 5$ is associated with a substantial improvement in learning performance, with $\mathcal{R}_{\text{avg}}$ increasing over a considerable portion of training, although with greater variability and a moderate decline at later iterations; meanwhile, FedNPG-ADMM improves more gradually and attains a lower return than FedNPG baseline.

Figure 2(b) demonstrates a complementary view of $\mathcal{R}_{\text{avg}}$ against cumulative communication overhead on a logarithmic scale. Aggressive compression with $q = 1$ provides FedNPG-ED to achieve a return comparable to FedNPG much earlier at substantially lower communication cost. Increasing q to 5 preserves more spectral information in the communicated natural-gradient representation and is associated empirically with a markedly higher return, albeit at increased communication cost and greater variability. Thus, the choice of q provides a tunable parameter governing spectral fidelity, communication overhead, and empirical learning performance. FedNPG-ADMM operates in a considerably lower communication regime than FedNPG, although its attained return is lower over the displayed training horizon.

6 Conclusions and Future Works

We developed FedNPG-ED, a communication-efficient FedRL method for closed-loop control of multi-agent systems. The method combines FedAvg with eigendecomposition-based FIM compression, retaining dominant eigenpairs and approximating the remaining spectrum with a scalar parameter. Under Assumptions 4.1–4.5, our LQR analysis characterises linear convergence to a neighbourhood of the global optimum and conditions for maintaining closed-loop stability without explicit constraints or projection. We further distinguished the roles of the approximation parameter and the truncation level, showing that the former reparametrises the learning rate, while the latter determines the communication payload and curvature fidelity, and that neither affects the asymptotic accuracy, which is determined solely by the sample budget. Simulation studies on linear and nonlinear systems, with both linear and neural actor structures, support these findings. We further observe that FedNPG-ED outperforms the ADMM baseline under heterogeneous local sample sizes at comparable communication cost. In some tested settings, smaller $p_k^{(i)}$ also yields faster convergence than uncompressed FedNPG, even when only a single EEP is retained.

Limitations. Theoretical analysis is restricted to LQR with a linear Gaussian actor, homogeneous dynamics, and an exact critic. Heterogeneity is limited to unequal local sample sizes. Relaxing these restrictions is left for future work. Further directions for future research are outlined in the Appendix A.

Reproducibility Statement

The assumptions underlying every theoretical claim are stated in Section 4 (Assumptions 4.1–4.5), and complete proofs of Theorem 1 and Corollary 1 are given in Appendix G.3 and Appendix G.4, with the supporting lemmas in Appendix G.1. The algorithm is stated in full as Algorithm 1 in Appendix E, and the rules for selecting p_k and q_k are given in Appendix G.5. The system matrices, cost weights, federation parameters, sample budgets, number of rounds and number of seeds are listed in Appendix F. The per-round communication payload of every method, which is the quantity the comparison turns on, is defined in Section 3, together with the cost of computing the transmitted eigenpairs. The code implementing algorithm 1 is provided with the submission.

Ethics Statement

This work studies federated policy optimisation using simulated linear and nonlinear dynamical systems. It involves no human participants, personal data, or sensitive data. Raw trajectories remain with the agents, but this does not provide a formal privacy guarantee because transmitted curvature information may reveal properties of local state distributions. Deployment in safety-critical systems requires further validation, as our stability guarantees are probabilistic and depend on the stated assumptions.

Statement on the Use of Large Language Models

In this work, we utilised generative AI tools primarily for drafting and editing text, as well as suggesting the manuscript structure. Additionally, we received assistance with coding the proposed algorithm and baseline algorithms. We also sought help in formulating mathematical claims, drafting proofs, reviewing research methodology and experiments, and interpreting results. Specifically, two large language models helped critically review the manuscript: one provided by the conference organisers and another obtained through a paid subscription.

We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, or propose or refine research hypotheses. Tasks such as translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work.

All theoretical statements, proofs, algorithmic contributions, and experiments are the work of the authors, who have verified every derivation presented here and take full responsibility for the content of the paper.

References

Mridul Agarwal, Bhargav Ganguly, and Vaneet Aggarwal. Communication efficient parallel reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 247–256. PMLR, 2021.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2020.

Dimitri Bertsekas. *Reinforcement learning and optimal control*, volume 1. Athena Scientific, 2019.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996. ISBN 978-1-886529-10-6.

Vivek S Borkar and Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

Richard H. Byrd, Gillian M. Chin, Will Neveitt, and Jorge Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21 (3):977–995, 2011.

Tianyi Chen, Kaiqing Zhang, Georgios B Giannakis, and Tamer Başar. Communication-efficient policy gradient methods for distributed reinforcement learning. *IEEE Transactions on Control of Network Systems*, 9(2):917–929, 2021.

Xuyang Chen, Jingliang Duan, and Lin Zhao. Global optimality of single-timescale actor-critic under continuous state-action space: A study on linear quadratic regulator. *arXiv preprint arXiv:2505.01041*, 2025.

Pengcheng Dai, Wenwu Yu, He Wang, Guanghui Wen, and Yuezu Lv. Distributed reinforcement learning for cyber-physical system with multiple remote state estimation under dos attacker. *IEEE Transactions on Network Science and Engineering*, PP:1–1, 08 2020.

Nicolò Dal Fabbro, Subhrakanti Dey, Michele Rossi, and Luca Schenato. SHED: A newton-type algorithm for federated learning based on incremental hessian eigenvector sharing. *Automatica*, 160, 2024.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Canh T Dinh, Nguyen H Tran, Tuan Dung Nguyen, Wei Bao, Amir Rezaei Balef, Bing B Zhou, and Albert Y Zomaya. DONE: distributed approximate newton-type method for federated edge learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(11):2648–2660, 2022.

Anis Elgabli, Chaouki Ben Issaid, Amrit S. Bedi, Ketan Rajawat, Mehdi Bennis, and Vaneet Aggarwal. FedNew: A communication-efficient and privacy-preserving newton-type method for federated learning. In *International Conference on Machine Learning*, pp. 5861–5877, 2022.

Murat A Erdogdu and Andrea Montanari. Convergence rates of sub-sampled newton methods. *Advances in Neural Information Processing Systems*, 28, 2015.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 1467–1476, 2018.

Konstantinos Gatsis. Federated reinforcement learning at the edge: Exploring the learning-communication tradeoff. In *European Control Conference*, pp. 1890–1895, 2022.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):123–158, 2023.

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

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International conference on machine learning*, pp. 3252–3261. PMLR, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: linear speedup under markovian sampling. In *International Conference on Machine Learning*, pp. 10997–11057, 2022.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

Safwan Labbi, Paul Mangold, Daniil Tiapkin, and Eric Moulines. On global convergence rates for federated policy gradient under heterogeneous environment. *https://arxiv.org/abs/2505.23459*, 2025.

Guangchen Lan, Han Wang, James Anderson, Christopher Brinton, and Vaneet Aggarwal. Improved communication efficiency in federated natural policy gradient via admm-based gradient updates. In *Conference on Neural Information Processing Systems*, 2023.

Guangchen Lan, Dong-Jun Han, Abolfazl Hashemi, Vaneet Aggarwal, and Christopher G Brinton. Asynchronous federated reinforcement learning with policy gradient updates: Algorithm design and convergence analysis. *International Conference on Learning Representations*, 2025.

Richard B Lehoucq, Danny C Sorensen, and Chao Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM, 1998.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Aritra Mitra. A simple finite-time analysis of td learning with linear function approximation. *IEEE Transactions on Automatic Control*, 2024.

Kanghua Mo, Peigen Ye, Xiaojun Ren, Shaowei Wang, Wenjun Li, and Jin Li. Security and privacy issues in deep reinforcement learning: Threats and countermeasures. *ACM Comput. Surv.*, 56(6), February 2024.

Daniel Jarne Ornia and Manuel Mazo. Event-based communication in distributed Q-learning. In *61st Conference on Decision and Control (CDC)*, pp. 2379–2386. IEEE, 2022.

Kei Ota, Devesh K Jha, and Asako Kanezaki. A framework for training larger networks for deep reinforcement learning. *Machine Learning*, 113(9):6115–6139, 2024.

Aravind Rajeswaran, Kendall Lowrey, Emanuel V Todorov, and Sham M Kakade. Towards generalization and simplicity in continuous control. *Advances in Neural Information Processing Systems*, 30, 2017.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Haoxing Tian, Alex Olshevsky, and Yannis Paschalidis. Convergence of actor-critic with multi-layer neural networks. *Advances in neural information processing systems*, 36:9279–9321, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Aashma Uprety and Danda B Rawat. Reinforcement learning for IoT security: A comprehensive survey. *IEEE Internet of Things Journal*, 8(11):8693–8706, 2020.

Han Wang, Aritra Mitra, Hamed Hassani, George J Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023a.

Han Wang, Sihong He, Zhili Zhang, Fei Miao, and James Anderson. Momentum for the win: Collaborative federated reinforcement learning across heterogeneous environments. In *International Conference on Machine Learning*, 2024.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *International Conference on Learning Representations*, 2020.

Shusen Wang, Farbod Roosta-Khorasani, Peng Xu, and Michael W. Mahoney. GIANT: globally improved approximate newton method for distributed optimization. *Advances in neural information processing systems*, 31, 2017.

Zhaolin Wang et al. FedSysID: A federated approach to sample-efficient system identification. In *Proceedings of International Conference on Machine Learning*, 2023b.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Zhijie Xie and Shenghui Song. FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence. *IEEE Journal on Selected Areas in Communications*, 41 (4):1227–1242, 2023.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Conference on Neural Information Processing Systems*, 2019.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *International Conference on Learning Representations*, 2024.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Gap-dependent bounds for federated Q-learning. *arXiv preprint arXiv:2502.02859*, 2025.

Feiran Zhao, Keyou You, and Tamer Başar. Global convergence of policy gradient primal–dual methods for risk-constrained lqrs. *IEEE Transactions on Automatic Control*, 68(5):2934–2949, 2023.

Mo Zhou and Jianfeng Lu. Single timescale actor-critic method to solve the linear quadratic regulator with convergence guarantees. *Journal of Machine Learning Research*, 24(222):1–34, 2023.

Feng Zhu, Robert W. Heath, and Aritra Mitra. Towards fast rates for federated and multi-task reinforcement learning. In *Conference on Decision and Control*, pp. 2658–2663, 2024.

A Future Work

Extensions of the analysis. An important direction is to consider agents with different dynamics $(\mathbf{A}^{(i)}, \mathbf{B}^{(i)})$, following Wang et al. (2023b;a). In this setting, agents may have different optimal controllers, and the analysis must account for their competing objectives and differences in local curvature. Another direction is to replace the exact critic in Assumption 4.5 with a learned critic and analyse the joint actor–critic updates, extending Wu et al. (2020); Tian et al. (2023) to the federated setting. A differentially private version would also be valuable. This requires calibrating the noise added to transmitted information and quantifying its effects on convergence and stability.

Extensions of the algorithm and evaluation. Compressing the critic curvature $\widehat{\mathcal{H}}_k^{(i)}$, which is currently transmitted in full, could further reduce communication. Its effect on critic learning would require additional analysis. Another option is to reuse previously transmitted eigenpairs and send

updates incrementally, as in Dal Fabbro et al. (2024). When curvature changes slowly, this could reduce communication and support warm starts for the eigen-solver, although outdated curvature information must be accounted for. Performing several local policy updates between aggregations (Karimireddy et al., 2020) could also reduce communication rounds but would introduce client drift.

B NOTATION

$\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ denote the sets of $n \times 1$ real vectors and $m \times n$ real matrices, respectively, and X^T denotes the transpose of a matrix or vector X. $\mathcal{N}(\mu, \mathbf{\Sigma})$ is the Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$, and $\mathbf{\Sigma} \geq 0$ (resp. > 0) indicates a positive semi-definite (resp. definite) matrix. $\|\cdot\|$ denotes the Frobenius norm of a matrix or the Euclidean norm of a vector, and $\mathrm{tr}(\cdot)$ the trace of a matrix. $\mathbb{E}[\cdot]$ and $\mathbb{P}\{\cdot\}$ denote the expectation operator and probability measure, respectively, $\widehat{(\cdot)}$ indicates an estimated or approximated value, and $\lambda(\cdot)$ symbolises an eigenvalue.

C BACKGROUND DETAILS

We have relocated some standard equations from Subsection 2.1 of the main manuscript here to reduce the length of the main document due to page-limit constraints.

Value functions. Under the infinite-horizon average-reward formulation (1), the state value function and the state-action value function are

$$V_{\pi_\theta}(x) = \lim_{T\to\infty} \sum_{t=0}^{T} \left\{\mathbb{E}_{x\sim\rho_x, u\sim\pi_\theta}\left[r_t \mid x_0 = x\right] - \mathcal{R}\right\}, \tag{19}$$

$$Q_{\pi_\theta}(x,u) = \lim_{T\to\infty} \sum_{t=0}^{T} \left\{\mathbb{E}_{x\sim\rho_x, u\sim\pi_\theta}\left[r_t \mid x_0 = x, u_0 = u\right] - \mathcal{R}\right\}. \tag{20}$$

Natural policy gradient. The true FIM at θ_k is

$$\mathcal{F}_k = \mathbb{E}_{x\sim\rho_{x,k}, u\sim\pi_{\theta_k}}\left[\nabla_\theta \log \pi_{\theta_k}(u_t|x_t)\, \nabla_\theta \log \pi_{\theta_k}(u_t|x_t)^\top\right], \tag{21}$$

and the natural ascent direction solves the trust-region problem

$$\underset{\theta}{\text{maximize}}\ \ g_k^\top(\theta - \theta_k) \quad \text{subject to } (\theta-\theta_k)^\top \mathcal{F}_k(\theta-\theta_k) \leq \delta, \quad \delta > 0. \tag{22}$$

Replacing $\mathcal{F}_k$ by the sample estimate (3) shows that the practical natural ascent direction is $\widehat{\mathcal{F}}_k^{-1}\widehat{g}_k$, in contrast to the PG direction $\widehat{g}_k$.

D SUPPLEMENTARY RESULTS

D.1 LINEAR ACTOR FOR LQR PROBLEM

We investigate the two design components of the proposed spectral approximation: (i) the number of retained eigendirections $q \in \{1, 4\}$ and (ii) the scalar approximation $p_k^{(i)}$ used for the discarded spectral subspace. The experiments are conducted under heterogeneous local sampling budgets $\tau_{i,k}$, such that the amount of local data varies across agents and federated iterations. Figure 3 reports both iteration-wise convergence and return as a function of cumulative communication, and Table 1 summarises the return and communication cost.

Mechanism used for heterogeneous local sampling budgets. To emulate heterogeneity in local data availability, the number of local batches used by agent i is randomised at every federated iteration. Specifically, for agent i at iteration k, we independently draw

$$\tau_{i,k} \sim \mathrm{Unif}\{2, 3, 4, 5\},$$

where $\tau_{i,k}$ denotes the number of local batches used to construct the local gradient and curvature estimates.

Effect of fixed q and different $p_k^{(i)}$. For each fixed $q \in \{1, 4\}$, we compare FedNPG-ED with three residual spectral approximations, $p_k^{(i)} = \widehat{\lambda}_{q+1,k}^{(i)}$, $p_k^{(i)} = \widehat{\lambda}_{\mathrm{avg},k}^{(i)}$, and $p_k^{(i)} = \widehat{\lambda}_{n,k}^{(i)}$ to the FedNPG baseline. Note that $\widehat{\lambda}_{q+1,k}$ and $\widehat{\lambda}_{n,k}$ denote the largest and smallest discarded eigenvalues, respectively, and $\widehat{\lambda}_{\mathrm{avg},k} := (\widehat{\lambda}_{q+1,k} + \widehat{\lambda}_{n,k})/2$ denotes their mean. For $q = 1$, all three FedNPG-ED variants closely track full FedNPG and converge to a similar return, despite retaining only one dominant eigendirection. The choice of $p_k^{(i)}$ mainly affects the transient behaviour: $p_k^{(i)} = \widehat{\lambda}_{n,k}^{(i)}$ exhibits faster initial improvement in learning in this experiment, whereas the differences become small near convergence. This observation is reflected in Table 1.

Table 1: Ablation of the eigenspace dimension q and spectral-tail approximation $p_k^{(i)}$ for the linear actor on the LQR problem.

Method	q	$p_k^{(i)}$	$\mathcal{R}_{\mathrm{avg}}$	Comm. Volume (Scalars)	Reduction (%)
FedNPG	–	–	≈ -112	72	0.00
FedNPG-ADMM	–	–	≈ -110	16	77.78
FedNPG-ED	1	$\widehat{\lambda}_{q+1,k}^{(i)}$	≈ -111	13	81.94
FedNPG-ED	1	$\widehat{\lambda}_{\mathrm{avg},k}^{(i)}$	≈ -111	13	81.94
FedNPG-ED	1	$\widehat{\lambda}_{n,k}^{(i)}$	≈ -112	13	81.94
FedNPG-ED	4	$\widehat{\lambda}_{q+1,k}^{(i)}$	≈ -111	25	65.28
FedNPG-ED	4	$\widehat{\lambda}_{\mathrm{avg},k}^{(i)}$	≈ -111	25	65.28
FedNPG-ED	4	$\widehat{\lambda}_{n,k}^{(i)}$	≈ -112	25	65.28

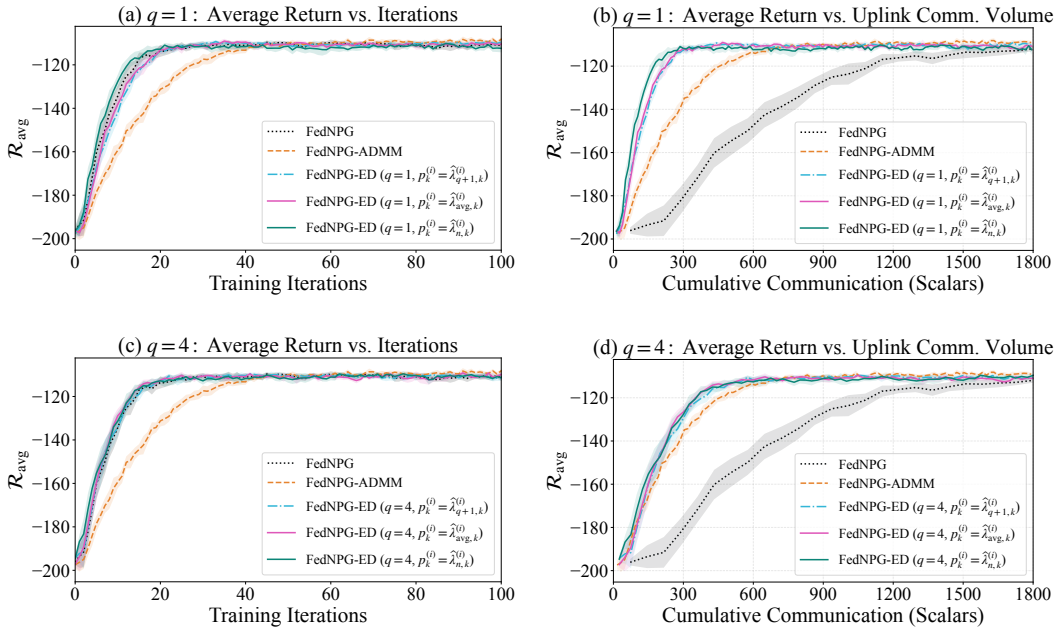

Figure 3: Effect of ablations of the eigenspace dimension q on learning performance and communication efficiency for a LQR system with linear actor.

A similar behaviour is observed for $q = 4$. Retaining $q = 4$ eigendirections makes the three choices of $p_k^{(i)}$ nearly indistinguishable in terms of iteration-wise learning behaviour, and all the ED variants converge to approximately the same return region as FedNPG. This reduced sensitivity to $p_k^{(i)}$ is consistent with the fact that a larger retained eigenspace leaves a smaller region of the spectrum

to be represented by the scalar residual approximation. Thus, within the tested regime, the precise choice of $p_k^{(i)}$ appears less consequential as q increases.

Linear actor communication payload. In Table 1, Comm. Volume (Scalars) denotes the number of scalars in the communication payload that the agent i transmits to the CS in each aggregation round for actor training. For FedNPG, each agent i communicates the policy-gradient vector together with the full FIM, resulting in $L_{\text{FedNPG}} = np + n^2$. Although the FIM is symmetric with its upper (or lower) triangular entries, we do not exploit FIM symmetry of FedNPG for communication payload in the reported results. Rather, our simulations for FedNPG transmit and count the full $n \times n$ FIM when computing the communication volume. For FedNPG-ED, each agent communicates the policy-gradient vector, q retained eigenvalues and their corresponding q eigenvectors, along with a scalar $p_k^{(i)}$ to approximate the discarded spectral tail. Hence, $L_{\text{FedNPG-ED}} = np + nq + 1$. For FedNPG-ADMM, each i communicates two d_θ-dimensional quantities: the local policy-gradient vector and the corresponding global direction vector, resulting in a per-round communication payload of $L_{\text{FedNPG-ADMM}} = 2np$. Consequently, the communication reduction of a method M relative to full FedNPG is computed as $\text{Reduction}(\%) = \left(1 - \frac{L_M}{L_{\text{FedNPG}}}\right) \times 100, \quad M \in \{\text{FedNPG-ED}, \text{FedNPG-ADMM}\}$.

Effect of the retained eigenspace dimension q. Increasing q improves the fidelity of the compressed spectral representation because a larger number of eigendirections are communicated explicitly. However, this improvement comes at a direct communication cost. As reported in Table 1, increasing q from 1 to 4 raises the per-round communication volume of FedNPG-ED from 13 to 25 scalars, reducing the communication saving relative to FedNPG from 81.94% to 65.28%. Despite this substantial increase in payload, the returns obtained with $q = 4$ are not markedly better than those obtained with $q = 1$. In particular, the $q = 1$ configurations achieve performance close to full FedNPG. These results expose the central design trade-off: $q \uparrow \implies$ higher spectral fidelity but larger communication payload.

Performance–communication trade-off. The return versus communication plots in Figure 3 provide a different perspective from aggregation round-wise convergence. The compressed ED methods enter the high-return regime using substantially less communication compared to the FedNPG. This advantage is particularly pronounced for $q = 1$, for which the ED variants retain only one dominant eigenpair and therefore progress rapidly along the return axis while remaining within a comparatively small communication budget.

For $q = 4$, FedNPG-ED continues to reduce communication relative to full FedNPG, but its communication trajectories extend farther along the horizontal axis because four eigendirections are transmitted. Consequently, the comparison between $q = 1$ and $q = 4$ illustrates that matching the iteration-wise convergence of full FedNPG does not require transmitting the largest tested eigenspace. In this experiment, $q = 1$ provides the more aggressive compression regime, whereas $q = 4$ provides a higher-fidelity but more communication-intensive approximation.

Communication cost at full eigenspace. When $q = 4 = n$, FedNPG-ED retains the entire eigenspace of the state-covariance matrix, and hence the benefit of spectral compression is reduced. Nevertheless, its per-round communication payload remains smaller than that of FedNPG: FedNPG-ED communicates 25 scalars per agent, compared with 72 scalars for FedNPG, corresponding to a communication reduction of 65.28%. This reduction is substantially smaller than the 81.94% achieved with $q = 1$, for which only 13 scalars are transmitted. These results illustrate the intended operating regime of FedNPG-ED: retaining a small dominant eigenspace ($q \ll n$) provides the greatest communication saving, whereas the advantage progressively diminishes as q approaches the full curvature dimension. Importantly, the experiments show that this reduction in communication is achieved while maintaining a steady-state return comparable to that of FedNPG.

Comparison with FedNPG-ADMM. FedNPG-ADMM exhibits a noticeably slower transient when performance is measured against training iterations. In particular, during the intermediate training phase, its average return remains below those of FedNPG and the FedNPG-ED variants, although it eventually approaches a comparable steady-state performance region. In contrast, the return vs. communication results highlight the communication efficiency of FedNPG-ADMM: its per-round payload is only 16 scalars, corresponding to a 77.78% reduction relative to the 72 scalars communicated by FedNPG. The iteration-wise and communication-wise comparisons therefore reveal a

clear trade-off. FedNPG-ED preserves the curvature information of FedNPG through a compressed spectral representation and achieves a faster transient in iterations, while FedNPG-ADMM communicates a smaller payload than FedNPG-ED with $q = 4$, but requires more training iterations to reach the same performance region in this experiment.

Overall ablation finding. The ablation of q suggests that FedNPG-ED is relatively robust to the specific residual spectral approximation $p_k^{(i)}$ under the considered sampling heterogeneity. More importantly, the retained eigenspace dimension q controls the dominant communication–fidelity trade-off. The $q = 1$ setting preserves the learning behaviour of full FedNPG while reducing the per-round communication from 72 to 13 scalars, whereas $q = 4$ increases the payload to 25 scalars without a commensurate improvement in return. These observations indicate that a small retained eigenspace can be sufficient when the dominant spectral directions capture the curvature information most relevant to the natural-gradient update.

D.2 Neural Actor for LQR Problem

To evaluate the proposed method beyond the linear policy parametrisation, we next consider a neural actor for the same LQR problem. Unlike the linear actor, where the Kronecker structure enables curvature compression in the n-dimensional state space, the neural actor induces the FIM in the d_θ-dimensional parameter space and the relevant curvature dimension is the total number of trainable actor parameters. Thus, such a Kronecker reduction is generally unavailable for a neural actor. This setting therefore provides a more challenging and practically relevant test of the communication–performance trade-off achieved by the proposed ED-based compression scheme.

In addition, we adopt the same heterogeneous local sampling mechanism as in the linear-actor setting in Appendix D.1. Specifically, at each federated iteration k, the local sampling budget of agent i is independently randomised as $\tau_{i,k} \sim \mathrm{Unif}\{2, \cdots, 5\}$, thereby capturing variability in local data availability across agents and iterations.

Learning behaviour with a neural actor. Figure 4 examines the effect of the retained eigenspace dimension q and the spectral-tail approximation $p_k^{(i)}$ when the linear actor is replaced by a neural policy. In contrast to the low-dimensional linear actor considered in Appendix D.1, the neural actor induces a substantially higher-dimensional FIM, making the spectral approximation more aggressive relative to the full curvature representation. Nevertheless, all FedNPG-ED variants exhibit a clear improvement from their initial return and approach a similar higher return region as training progresses. The results therefore indicate that the dominant eigenspace retains sufficient curvature information to support effective natural-gradient updates even when only a small fraction of the full FIM is communicated.

Effect of the retained spectral rank q. Increasing q from 1 to 10 and 20 does not yield a corresponding improvement in return. As reported in Table 2, with $p_k^{(i)} = \widehat{\lambda}_{q+1,k}^{(i)}$, the average returns are approximately -57, -57, and -58 for $q = 1$, 10, and 20, respectively. A similar behaviour is observed for $p_k^{(i)} = \widehat{\lambda}_{\mathrm{avg},k}^{(i)}$, for which the corresponding returns are approximately -56, -58, and -59. Thus, increasing the retained rank provides little performance benefit over the aggressively compressed $q = 1$ representation in these cases. This suggests that the curvature information most relevant to the policy update is concentrated predominantly in a relatively small dominant eigenspace. Consequently, increasing q improves the fidelity of the FIM approximation, but the additional spectral information does not necessarily translate into a better policy once the principal curvature directions have been captured.

Influence of the spectral-tail approximation. The choice of $p_k^{(i)}$ has a more visible influence on performance. Across all three values of q, using either $\widehat{\lambda}_{q+1,k}^{(i)}$ or $\widehat{\lambda}_{\mathrm{avg},k}^{(i)}$ yields a higher return than using the smallest eigenvalue $\widehat{\lambda}_{d_\theta,k}^{(i)}$. At $q = 1$, the corresponding returns are approximately -57, -56, and -65, respectively. The same qualitative behaviour is observed for $q = 10$, where the returns are approximately -57, -58, and -64, and for $q = 20$, where they are approximately -58, -59, and -62.

This behaviour can be interpreted through the approximation of the discarded spectral tail. Using $\widehat{\lambda}_{d_\theta,k}^{(i)}$ assigns the smallest curvature value uniformly to all discarded directions and may therefore

Table 2: Ablation of the retained eigenspace dimension q and spectral-tail approximation $p_k^{(i)}$ for the neural actor on the LQR problem.

Method	q	$p_k^{(i)}$	$\mathcal{R}_{\text{avg}}$	Comm. Volume (Scalars)	Reduction (%)
FedNPG	–	–	≈ -59	493,506	0.00
FedNPG-ADMM	–	–	≈ -69	1,404	99.72
FedNPG-ED	1	$\widehat{\lambda}_{q+1,k}^{(i)}$	≈ -57	1,405	99.72
FedNPG-ED	1	$\widehat{\lambda}_{\text{avg},k}^{(i)}$	≈ -56	1,405	99.72
FedNPG-ED	1	$\widehat{\lambda}_{d_\theta,k}^{(i)}$	≈ -65	1,405	99.72
FedNPG-ED	10	$\widehat{\lambda}_{q+1,k}^{(i)}$	≈ -57	7,723	98.44
FedNPG-ED	10	$\widehat{\lambda}_{\text{avg},k}^{(i)}$	≈ -58	7,723	98.44
FedNPG-ED	10	$\widehat{\lambda}_{d_\theta,k}^{(i)}$	≈ -64	7,723	98.44
FedNPG-ED	20	$\widehat{\lambda}_{q+1,k}^{(i)}$	≈ -58	14,743	97.01
FedNPG-ED	20	$\widehat{\lambda}_{\text{avg},k}^{(i)}$	≈ -59	14,743	97.01
FedNPG-ED	20	$\widehat{\lambda}_{d_\theta,k}^{(i)}$	≈ -62	14,743	97.01

under-represent the curvature associated with the neglected eigenspace. In comparison, $\widehat{\lambda}_{q+1,k}^{(i)}$ and $\widehat{\lambda}_{\text{avg},k}^{(i)}$ provide richer scalar representations of the discarded spectrum. The neural-actor experiment therefore indicates that, under aggressive spectral compression, the scalar approximation assigned to the discarded subspace can be as important as the number of explicitly retained eigendirections.

Communication–curvature fidelity trade-off. The communication advantage of spectral compression becomes prominant when the retained rank is varied. Full FedNPG communicates $493{,}506$ scalars per aggregation round, whereas FedNPG-ED requires only $1{,}405$, $7{,}723$, and $14{,}743$ scalars for $q = 1$, 10, and 20, respectively. These correspond to communication reductions of approximately 99.72%, 98.44%, and 97.01% relative to FedNPG. Hence, even the largest tested rank retains a substantial communication advantage, while $q = 1$ provides the most aggressive compression. Importantly, the increase in communication from $q = 1$ to $q = 10$ or $q = 20$ does not produce a commensurate improvement in return. Figure 4 therefore demonstrates that, for the LQR system experiment with neural actor, a very small retained eigenspace can provide an effective communication–curvature fidelity trade-off.

Communication trade-off for neural actor. The communication behaviour is notably different for the neural actor, for which the curvature matrix is defined in the d_θ-dimensional policy-parameter space rather than the n-dimensional state space. Consequently, FedNPG requires transmitting a $d_\theta \times d_\theta$ curvature matrix, whereas FedNPG-ED communicates only q dominant eigenpairs together with the residual spectral approximation p_k. Since d_θ can be substantially larger than the state dimension, choosing $q \ll d_\theta$ yields a considerably wider regime in which eigendecomposition-based compression remains communication-efficient. Nevertheless, the same fundamental trade-off persists: increasing q improves the fidelity of the curvature approximation at the expense of a larger communication payload, and the compression benefit eventually vanishes as q approaches the full curvature dimension d_θ.

Neural actor communication payload. For the neural actor, the per-agent communication payloads of FedNPG is $L_{\text{FedNPG}} = d_\theta + d_\theta^2$, for FedNPG-ED is $L_{\text{FedNPG-ED}} = d_\theta(q+1) + 1$, and for FedNPG-ADMM is $L_{\text{FedNPG-ADMM}} = 2d_\theta$, respectively. The reduction in communication by each method is computed as in Appendix D.1.

Comparison with FedNPG and FedNPG-ADMM. Table 2 further highlights the distinction between optimisation performance and communication efficiency. Full FedNPG reaches the return of approximately -59, but requires $493{,}506$ communicated scalars per round. FedNPG-ADMM reduces this payload to $1{,}404$ scalars, although its return is approximately -69. In comparison, FedNPG-ED with $q = 1$ requires only $1{,}405$ scalars—almost the same communication payload as FedNPG–ADMM while attaining approximately -57 and -56 with the $\widehat{\lambda}_{q+1,k}^{(i)}$ and $\widehat{\lambda}_{\text{avg},k}^{(i)}$ tail approximations, respectively.

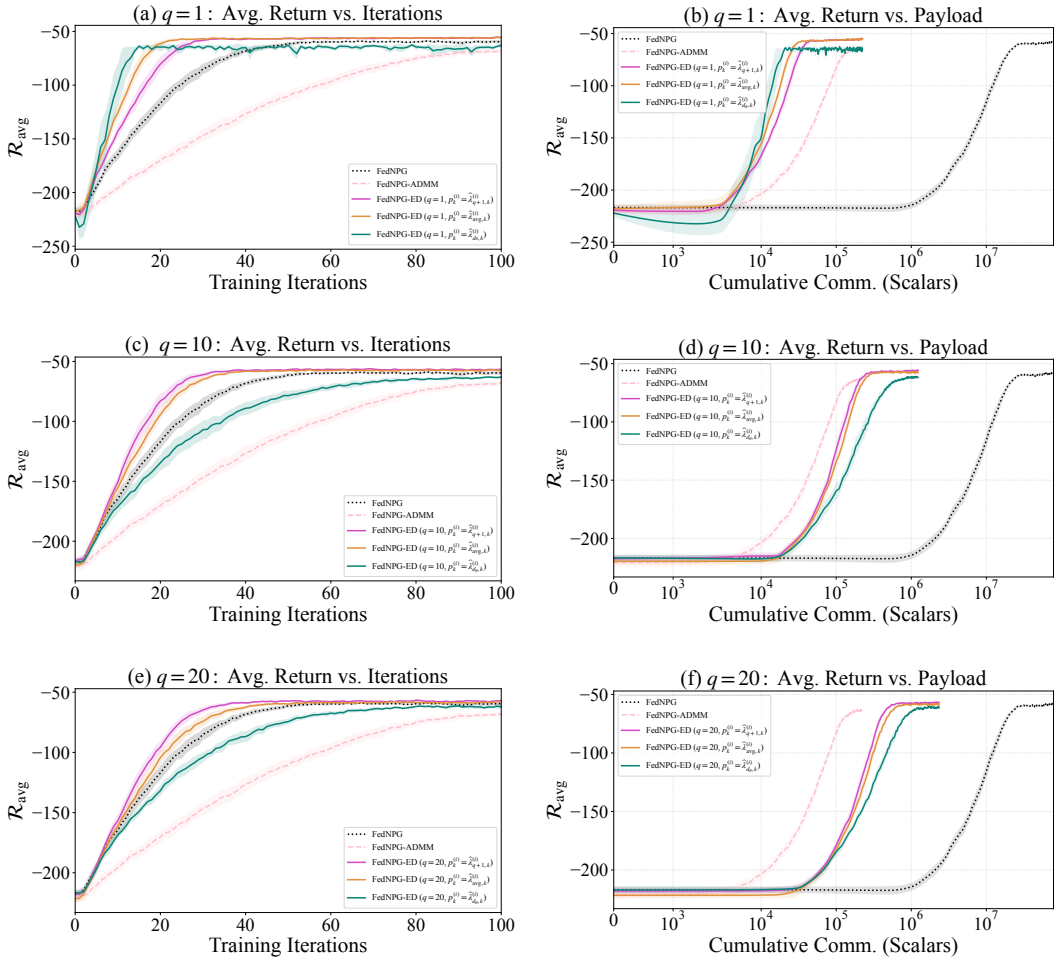

Figure 4: Effect of the retained eigenspace dimension q and spectral-tail approximation $p_k^{(i)}$ on learning and communication efficiency for an LQR system with the neural actor.

Thus, in this experiment, the compressed eigenspace representation retains useful curvature information at a communication cost comparable to the ADMM baseline. This illustrates an important distinction between the two communication-efficient approaches: FedNPG-ADMM avoids transmitting the full curvature matrix through its splitting mechanism, whereas FedNPG-ED explicitly preserves selected dominant curvature directions.

Joint effect of the retained eigenspace and spectral-tail approximation. Taken together, Figure 4 and Table 2 show that increasing q alone is not sufficient to improve performance; the approximation assigned to the discarded eigenspace also plays an important role. In particular, the $q = 1$ configurations with $p_k^{(i)} = \widehat{\lambda}_{q+1,k}^{(i)}$ or $\widehat{\lambda}_{\text{avg},k}^{(i)}$ attain returns comparable to those obtained with $q = 10$ and $q = 20$, while requiring substantially fewer transmitted scalars. Hence, for the considered neural LQR setting, an effective spectral approximation is determined jointly by the dominant eigendirections retained explicitly and the scalar used to represent the discarded spectral subspace. The results favour aggressive spectral compression when the dominant curvature directions are adequately captured.

D.3 Neural Actor for Humanoid-v5 Benchmark

We further evaluate the proposed FedNPG-ED framework on the Humanoid-v5 benchmark, which provides a substantially higher-dimensional and more challenging continuous-control setting. Un-

like the preceding LQR experiments, the policy is parameterised by a neural actor, resulting in a considerably larger parameter space and, consequently, a much higher communication cost for transmitting the full FIM. This setting therefore provides a useful benchmark for assessing whether the proposed ED-based compression can retain effective policy learning while substantially reducing the communication burden. We compare FedNPG with FedNPG-ED using $q \in \{1, 25\}$ and $p_k^{(i)} = \widehat{\lambda}_{d_\theta,k}^{(i)}$, and evaluate their performance in terms of both training iterations and cumulative communication overhead. The environmental details of Humanoid-v5 and the set of experimental parameters are provided in Table 5 and 7 in Appendix F.2.

Learning performance and convergence behaviour. Figure 5(a) illustrates that all three methods start from a comparable initial performance level, with $\mathcal{R}_{\text{avg}}$ around 95. FedNPG subsequently improves more gradually and relatively smoothly, reaching approximately $115-120$ towards the end of training. FedNPG-ED also exhibits a clear improvement in average return, demonstrating that the compressed curvature information remains sufficient to support effective policy updates in the Humanoid-v5 environment.

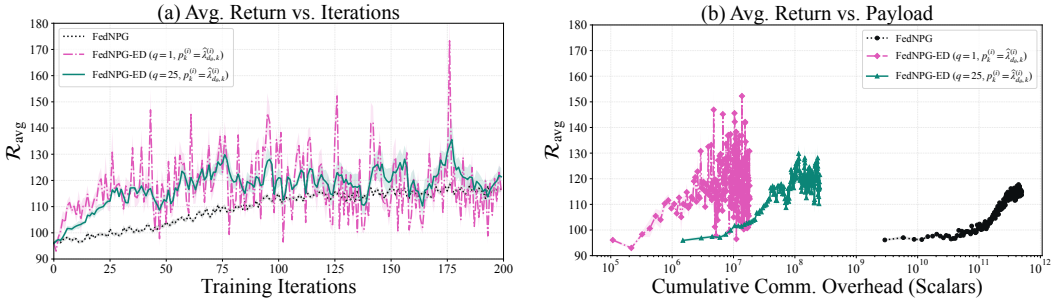

Figure 5: Effect of the retained eigenspace dimension q and spectral-tail approximation $p_k^{(i)}$ on learning and communication efficiency for the Humanoid-v5 benchmark.

Effect of the retained eigenspace dimension q. The choice of q reveals a trade-off between aggressive curvature compression and the stability of the resulting policy updates. With $q = 1$, FedNPG-ED retains only the dominant eigendirection and exhibits substantially larger fluctuations in $\mathcal{R}_{\text{avg}}$, including pronounced peaks and drops throughout training. In contrast, $q = 25$ produces a noticeably smoother learning trajectory while maintaining returns broadly above the FedNPG baseline over substantial portions of training. This suggests that retaining a larger eigenspace provides a more informative approximation of the FIM and consequently more stable natural-gradient updates.

Performance–communication trade-off. In Figure 5, the communication-domain results highlight the main advantage of FedNPG-ED. FedNPG requires cumulative communication on the order of 10^{11} scalars to reach its higher-return region, whereas the ED variants operate at substantially smaller communication volumes. In particular, $q = 1$ achieves strong returns with communication concentrated around 10^6 to 10^7 scalars, while $q = 25$ trades some of this communication saving for smoother learning and operates roughly in the 10^7 to 10^8 range. Thus, increasing q improves the fidelity of the curvature approximation at the expense of a larger payload.

Overall effect of eigendecomposition-based compression. Taken together, the results show that FedNPG-ED can preserve meaningful policy improvement while substantially reducing the communication required to transmit curvature information. The $q = 1$ configuration represents the more aggressive compression regime but introduces greater variability, whereas $q = 25$ provides a more balanced compromise between communication efficiency and learning stability. The comparison with full FedNPG therefore illustrates the central benefit of retaining only a low-dimensional spectral representation of the FIM rather than communicating the full curvature matrix.

FedNPG-ADMM showed limited learning progress with the same neural actor–critic architecture and learning rates under unequal local sample sizes. We therefore omit its curve for clarity and focus on FedNPG and FedNPG-ED, both of which showed clear improvement under this configuration.

E ALGORITHM 1

Algorithm 1: FedNPG-ED based AC algorithm

Data: MDP $\langle \mathcal{X}, \mathcal{U}, \mathcal{P}, r \rangle$; number of communication rounds C; number of agents N; fixed number of retained EEPs $q_k < d_\theta$

Initialize: actor parameter $\theta_0 \in \mathbb{R}^{d_\theta}$

for $k = 0, 1, \ldots, C-1$ **do**

  Server broadcast: send θ_k to all agents

  Agent level computation:

  for $i = 1, \ldots, N$ **do**

    Collect local trajectories $\{x_t, u_t, r_t\}_{t=1}^{\tau_{i,k}}$ using the policy π_{θ_k}.

    Compute local actor quantities: $\widehat{g}_k^{(i)}$ from (2) and $\widehat{\mathcal{F}}_k^{(i)}$ from (3).

    Compute the top-$q_k^{(i)}$ EEPs $\{(\widehat{\lambda}_{b,k}^{(i)}, \widehat{v}_{b,k}^{(i)})\}_{b=1}^{q_k^{(i)}}$ of $\widehat{\mathcal{F}}_k^{(i)}$ in (9).

    Form the compressed actor payload $\{\widehat{g}_k^{(i)}, p_k^{(i)}, \{\widehat{\lambda}_{b,k}^{(i)}, \widehat{v}_{b,k}^{(i)}\}_{b=1}^{q_k}\}$.

    Send compressed actor payload to CS.

  end

  Server level computation:

  Reconstruct $\left\{\widehat{\mathcal{F}}_{\mathrm{ed},k}^{(i)}\right\}_{i=1}^{N}$ using (9).

  Compute FedAvg weight $\{\omega_{a,k}^{(i)}\}_{i=1}^{N}$ using (7).

  Aggregate actor quantities: $\widehat{g}_k$ and $\widehat{\mathcal{F}}_{\mathrm{ed},k}$ from (7) and (10), respectively.

  Update global parameter:

$$\theta_{k+1} \leftarrow \theta_k + \eta_k \, \widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1} \, \widehat{g}_k.$$

end

F SIMULATION PARAMETERS

F.1 LINEAR POLICY PARAMETERS

The parameters for the linear policy experiments are given in Table 3, and the system model parameters are given as follows.

$$\mathbf{A} = \begin{bmatrix} 1.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.125 & 0.0 \\ 0.5 & 0.0 \\ 0.0 & 0.125 \\ 0.0 & 0.5 \end{bmatrix}, \quad \mathscr{Q} = \mathrm{diag}(1.0,\, 0.1,\, 2.0,\, 0.2),$$

$$\Sigma_w^{(i)} = \begin{bmatrix} 1.25 & 5.0 & 0.0 & 0.0 \\ 5.0 & 20.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.25 & 5.0 \\ 0.0 & 0.0 & 5.0 & 20.0 \end{bmatrix}, \quad \mathscr{R} = I \in \mathbb{R}^{2\times 2}, \quad \Sigma_\sigma^{(i)} = 5 I_2. \tag{23}$$

F.2 NONLINEAR POLICY PARAMETERS

The neural-policy parameters for linear system and MuJoCo benchmarks are summarized in Tables 4, 5, 6, and 7.

Table 3: Simulation parameters for the linear-actor LQR experiments.

Parameter	Value
State dimension, n	4
Action dimension, p	2
Number of participating federated agents, N	5
Communication rounds, C	100
Trajectory length	1000
Batch size	200
Number of seeds, S	10
Actor architecture	$[0]$
Actor parameter for adaptive learning rate, α_a	5×10^{-3}
Critic architecture	$[10, 50]$
Activation function for the hidden nodes	$\tanh(\cdot)$
Fixed learning rate for the critic	10^{-2}
Number of truncation levels in FedNPG-ED, q	$1, 4$

Table 4: Simulation parameters for the neural-actor LQR experiments.

Parameter	**Value**
State dimension, n	4
Action dimension, p	2
Number of participating federated agents, N	5
Number of global federated updates, C	100
Number of time steps per trajectory	1000
Number of samples used for training	200
Number of independent experimental runs, S	10
Hidden-layer dimensions of the neural actor	$[100]$
Actor parameter for adaptive learning rate, α_a	5×10^{-3}
Hidden-layer dimensions of the neural critic	$[10, 50]$
Activation function for the hidden nodes	$\tanh(\cdot)$
Critic parameter for fixed learning rate	10^{-2}
Standard deviation of the Gaussian policy	$10I_2$
Actor FIM eigenpairs retained in FedNPG-ED, q	$1, 10, 20$

Table 5: Summary of the MuJoCo environments used in the experiments.

Environment	State Dim.	Action Dim.	Action Space	DoF	Max. Steps	Task
Hopper-v5	11	3	$[-1, 1]^3$	3	1000	Forward locomotion
Humanoid-v5	348	17	$[-0.4, 0.4]^{17}$	17	1000	Bipedal locomotion

G Theoretical Proofs

G.1 Technical Lemmas

Under a fixed policy $K \in \mathcal{K}$ and the control input in (11), the average LQR cost admits

$$\begin{aligned} \mathcal{J}_{\text{avg}}(K) &= \sum_{i=1}^{N} \omega_{a,k}^{(i)} \ \text{tr}\left[\left(\mathscr{Q} + K^\top \mathscr{R} K\right)\Sigma_K^{(i)} + \mathscr{R}\Sigma_\sigma^{(i)}\right] \\ &= \sum_{i=1}^{N} \omega_{a,k}^{(i)} \ \text{tr}\left(P_K^{(i)}\Sigma_{\bar{w}}^{(i)} + \mathscr{R}\Sigma_\sigma^{(i)}\right), \end{aligned} \tag{24}$$

where $\Sigma_{\bar{w}}^{(i)}$ denotes the covariance of the effective additive disturbance entering the closed-loop system dynamics. It is straightforward to derive (24) from the single agent case as given in Yang et al. (2019). $\Sigma_K^{(i)}$ denotes the closed-loop steady-state state covariance for agent i. Since we assume homoginity in the system dynamics across agents, we will drop the agent index i from the variables

Table 6: Parameter settings on Hopper-v5 experiments.

Parameter	Value
State dimension, n	11
Action dimension, p	3
Number of federated agents, N	3
Number of communication rounds, C	160
Local sample budget, $\tau_{i,k}$	$\mathrm{Unif}\{5, \cdots, 9\}$
Trajectory length	500
Batch size	500
Number of seeds, S	10
Actor architecture	$[100]$
Actor parameter, α_a	10^{-5}
Critic architecture	$[64, 64]$
Activation function	$\tanh(\cdot)$
Critic parameter	5×10^{-2}
Standard deviation for the actor's NN Weight initialization	0.02
Initial bias of the actor NN layers	0
Gaussian policy standard deviation	$0.1 I_3$
Retained EEPs, q	$1, 5$

Table 7: Experimental parameters on Humanoid-v5.

Parameter	Value
State dimension, n	348
Action dimension, p	17
Number of agents, N	3
Communication rounds, C	200
Trajectory length	500
Batch size	500
Number of seeds, S	10
Actor architecture	$[128, 64]$
Actor parameter, α_a	5×10^{-4}
Critic architecture	$[128, 64, 32]$
Activation function	$\tanh(\cdot)$
Critic parameter, α_c	5×10^{-2}
Weight initialization std., ν	0.07
Initial bias	0
Policy standard deviation, $\Sigma_\sigma^{(i)}$	$0.05 I_{17}$
Retained EEPs, q	$1, 25$

which do not change with the agent in the subsequent analysis for simplicity. Therefore, Σ_K is the unique positive definite solution to the Lyapunov equation

$$\Sigma_K = \Sigma_{\bar{w}} + \big(\mathbf{A} - \mathbf{B}K\big)\Sigma_K\big(\mathbf{A} - \mathbf{B}K\big)^\top, \tag{25}$$

and P_K is the unique positive solution to the Bellman equation (see Yang et al. (2019))

$$P_K = \big(\mathscr{Q} + K^\top \mathscr{R} K\big) + \big(\mathbf{A} - \mathbf{B}K\big)^\top P_K \big(\mathbf{A} - \mathbf{B}K\big), \tag{26}$$

respectively. Therefore, the dependence of $\mathcal{J}_{\text{avg}}(K)$ on the policy gain K is characterised through both the quadratic term $\mathscr{Q} + K^\top \mathscr{R} K$ and the K-dependent state covariance Σ_K, which is equivalent to the Bellman solution P_K.

Now, we present several important results that are useful for closed-loop stability during training and convergence analysis of the proposed algorithm, given in the following:

Lemma 1. *The average cost function $\mathcal{J}_{\text{avg}}$ is coercive on the stabilising set $\mathcal{K}$, i.e., for any sequence $\{K_m\}_{m=1}^{\infty} \subset \mathcal{K}$,*

$$\|K_m\| \to +\infty \quad \textit{or} \quad K_m \to \partial\mathcal{K} \quad \Longrightarrow \mathcal{J}_{\text{avg}}(K_m) \to +\infty. \tag{27}$$

Proof. Since each individual LQR cost $\mathcal{J}^{(i)}(K)$ is coercive on the stabilising set $\mathcal{K}$ (Hu et al., 2023), it follows that for any sequence $\{K_m\}_{m\in\mathbb{N}} \subset \mathcal{K}$ satisfying $\|K_m\| \to +\infty$ or $K_m \to \partial\mathcal{K}$, we have $\mathcal{J}^{(i)}(K_m) \to +\infty \; \forall i \in \{N\}$. Hence,

$$\mathcal{J}_{\text{avg}}(K_m) \to +\infty. \tag{28}$$

Therefore, $\mathcal{J}_{\text{avg}}$ is coercive on $\mathcal{K}$. This completes the proof. ■

Corollary 2. *If the average cost function $\mathcal{J}_{\text{avg}}(K)$ is coercive in the stabilising set $\mathcal{K}$, then the sublevel set $\mathcal{S}(k_\epsilon)$ defined in Section 4 is bounded.*

Proof. The stabilising set

$$\mathcal{S}(k_\epsilon) = \{K \in \mathcal{K} \mid \mathcal{J}_{\text{avg}}(K_0) \le k_\epsilon\} \tag{29}$$

is a sublevel set of $\mathcal{J}_{\text{avg}}$ on $\mathcal{K}$, and K_0 is an initial stabilising gain. Suppose, for contradiction, that $\mathcal{S}(k_\epsilon)$ is unbounded. Then there exists a sequence $\{K_m\}_{m=1}^{\infty} \subset \mathcal{S}(k_\epsilon)$ such that $\|K_m\| \to \infty$. Since $\mathcal{J}_{\text{avg}}$ is coercive on $\mathcal{K}$, it follows that

$$\mathcal{J}_{\text{avg}}(K_m) \to +\infty. \tag{30}$$

However, because $K_m \in \mathcal{S}(k_\epsilon) \; \forall m$, we also have

$$\mathcal{J}_{\text{avg}}(K_m) \le k_\epsilon \quad \forall m, \tag{31}$$

which is a contradiction. Therefore, $\mathcal{S}(k_\epsilon)$ must be bounded. This completes the proof. ■

Lemma 2. *The average cost function $\mathcal{J}_{\text{avg}} : \mathcal{K} \to \mathbb{R}$ is twice continuously differentiable in $\mathcal{K}$.*

Proof. Let $A_K := \mathbf{A} - \mathbf{B}K$. By Assumption 4.3, A_K is Schur stable $\forall K \in \mathcal{K}$. Hence, closed-loop dynamics (12) admits a unique stationary distribution with finite second moment, and $\Sigma_K := \mathbb{E}\left[x_t^{(i)} x_t^{(i)\top}\right]$ is the unique solution of (25). Under stationarity, the ergodic cost of each agent i is

$$\mathcal{J}^{(i)}(K) = \operatorname{tr}\big((\mathscr{Q} + K^\top \mathscr{R} K)\Sigma_K\big) + \operatorname{tr}(\mathscr{R}\Sigma_\sigma)\,, \tag{32}$$

where the cross-term vanishes due to the $x_t^{(i)}$ independence and zero mean of the exploration noise. Thus, the regularity of $\mathcal{J}^{(i)}(K)$ reduces to that of the mapping $K \mapsto \Sigma_K$. We define

$$F(K, \Sigma_K) := \Sigma_K - A_K \Sigma_K A_K^\top - \Sigma_{\bar{w}}.$$

Then Σ_K satisfies $F(K, \Sigma_K) = 0, \forall K \in \mathcal{K}$. Since A_K depends affinely on K, the mapping $(K, \Sigma_K) \mapsto F(K, \Sigma_K)$ is twice continuously differentiable. The Fréchet derivative with respect to Σ_K is

$$D_{\Sigma_K} F(K, \Sigma_K)[X] = X - A_K X A_K^\top. \tag{33}$$

Because A_K is Schur stable, the linear operator $X \mapsto X - A_K X A_K^\top$ is invertible for every $K \in \mathcal{K}$. Therefore, by the implicit function theorem, the mapping is $K \mapsto \Sigma_K \in C^2(\mathcal{K})$.

Finally, the mapping $K \mapsto \mathscr{Q} + K^\top \mathscr{R} K$ is polynomial in K, and hence C^2. Combining this with (32) implies $\mathcal{J}^{(i)} \in C^2(\mathcal{K}), \; \forall i \in \{N\}$ (Hu et al., 2023). Since $\mathcal{J}_{\text{avg}}(K)$ is a finite weighted sum of $\mathcal{J}^{(i)}(K)$, it follows that $\mathcal{J}_{\text{avg}} \in C^2(\mathcal{K})$. This completes the proof. ■

Lemma 3. *The sublevel set $\mathcal{S}(k_\epsilon)$, defined in (29), is compact.*

Proof. By Lemma 2, the map $K \mapsto \mathcal{J}_{\text{avg}}(K)$ is continuous on $\mathcal{K}$. Therefore, $\mathcal{S}(k_\epsilon)$ is a sublevel set of a continuous function and, therefore, it is closed relative to $\mathcal{K}$, which implies that $\mathcal{S}(k_\epsilon)$ is closed.

Next, by Corollary 2, $\mathcal{J}_{\text{avg}}$ is coercive on $\mathcal{K}$. Hence, every sublevel set of $\mathcal{J}_{\text{avg}}$ is bounded. In particular, $\mathcal{S}(k_\epsilon)$ is bounded (see Proposition 11.12 in (Bauschke & Combettes, 2020)).

Therefore, $\mathcal{S}(k_\epsilon)$ is closed and bounded $\implies$ $\mathcal{S}(k_\epsilon)$ is compact. This completes the proof.

■

G.2 KEY LEMMAS

Note that for the Gaussian linear policy (11), the score function is $\nabla_K \log \pi_K(u \mid x) = \Sigma_\sigma^{-1}(u + Kx)x^\top$, and the FIM with respect to $\mathrm{vec}(K)$ factorizes exactly as

$$\mathcal{F}_k = \Sigma_{K_k} \otimes \Sigma_\sigma^{-1}, \tag{34}$$

where $\Sigma_{K_k} = \mathbb{E}[x_t x_t^\top]$ is the stationary state covariance under K_k. Consequently the natural-gradient step in $\mathrm{vec}(K)$ coordinates is equivalent to the matrix update $K_{k+1} = K_k - \eta_k \nabla_K \mathcal{J}_{\mathrm{avg}}(K_k)\Sigma_{K_k}^{-1}$ (Yang et al. (2019)), which is used in the proofs.

Lemma 4. *The cost function $\mathcal{J}_{\mathrm{avg}}(K)$ is $\mathcal{L}$-smooth in the sublevel set $\mathcal{S}(k_\epsilon)$, where the constant $\mathcal{L} > 0$ depends on ϵ and the problem parameters. More precisely,*

$$\left\|\nabla^2 \mathcal{J}_{\mathrm{avg}}(K)\right\| \le \mathcal{L}, \tag{35}$$

where $\|\cdot\|$ denotes the Frobenius norm for a matrix or the Euclidean norm of a matrix treated as a vector. Then $\mathcal{L} < \infty$, and for any (K, K') such that

$$K_c := (1-c)K + cK' \in \mathcal{S}(k_\epsilon), \qquad \forall c \in [0,1], \tag{36}$$

and the inequalities hold in the following

$$\mathcal{J}_{\mathrm{avg}}(K') - \mathcal{J}_{\mathrm{avg}}(K) \le \mathrm{tr}\big((\nabla \mathcal{J}_{\mathrm{avg}}(K))^\top (K' - K)\big) + \frac{\mathcal{L}}{2}\|K' - K\|_F^2, \tag{37}$$

$$\|\nabla \mathcal{J}_{\mathrm{avg}}(K') - \nabla \mathcal{J}_{\mathrm{avg}}(K)\|_F \le \mathcal{L}\|K' - K\|_F. \tag{38}$$

Proof. The set $\mathcal{S}(k_\epsilon)$ is compact by Lemma 3. Moreover, according to Lemma 2, the mapping $K \mapsto \mathcal{J}_{\mathrm{avg}}(K) \in C^2(\mathcal{K}) \implies K \mapsto \nabla^2 \mathcal{J}_{\mathrm{avg}}(K)$ is continuous in $\mathcal{S}(k_\epsilon)$.

By the Weierstrass theorem, $\left\|\nabla^2 \mathcal{J}_{\mathrm{avg}}(K)\right\|$ is bounded on $\mathcal{S}(k_\epsilon)$, i.e. it attains its maximum on $\mathcal{S}(k_\epsilon)$ and is finite, as given in (35), so $\mathcal{L} < \infty$.

Now we fix any (K, K') such that the entire line segment is $K_c = (1-c)K + cK' \in \mathcal{S}(k_\epsilon)$, $c \in [0,1]$. Let $\Delta := K' - K$, and we define the scalar function

$$\varphi(c) := \mathcal{J}_{\mathrm{avg}}(K + c\Delta), \quad c \in [0,1]. \tag{39}$$

Since $\mathcal{J}_{\mathrm{avg}} \in C^2(\mathcal{K})$, so is φ. By the chain rule,

$$\varphi'(c) = \mathrm{tr}\Big(\big(\nabla \mathcal{J}_{\mathrm{avg}}(K + c\Delta)\big)^\top \Delta\Big), \tag{40}$$

$$\varphi''(c) = \mathrm{tr}\big(\Delta^\top \nabla^2 \mathcal{J}_{\mathrm{avg}}(K + c\Delta)[\Delta]\big). \tag{41}$$

Applying Taylor's theorem to (39) with integral remainders at $c = 0$ and $c = 1$ gives the following

$$\mathcal{J}_{\mathrm{avg}}(K') = \mathcal{J}_{\mathrm{avg}}(K) + \mathrm{tr}\big((\nabla \mathcal{J}_{\mathrm{avg}}(K))^\top \Delta\big) + \int_0^1 (1-c)\,\varphi''(c)\,dc. \tag{42}$$

Using (35) and the Frobenius norm,

$$|\varphi''(c)| = \left|\mathrm{tr}\left(\Delta^\top \nabla^2 \mathcal{J}_{\mathrm{avg}}(K + c\Delta)[\Delta]\right)\right| \le \left\|\nabla^2 \mathcal{J}_{\mathrm{avg}}(K + c\Delta)\right\| \|\Delta\|_F^2 = \mathcal{L}\|\Delta\|_F^2. \tag{43}$$

Hence, substituting (43) into (42) leads to (37).

Next, by the fundamental theorem of calculus $\varphi'(1) - \varphi'(0) = \int_0^1 \varphi''(c)dc$, applied to the gradient along the segment $\varphi'(1) = \nabla \mathcal{J}_{\mathrm{avg}}(K')$ and $\varphi'(0) = \nabla \mathcal{J}_{\mathrm{avg}}(K)$, we get $\nabla \mathcal{J}_{\mathrm{avg}}(K') - \nabla \mathcal{J}_{\mathrm{avg}}(K) = \int_0^1 \nabla^2 \mathcal{J}_{\mathrm{avg}}(K + c\Delta)[\Delta]\,dc$. Taking the Frobenius norm and using the triangle inequality yields (38). This completes the proof. ■

Lemma 5. *$\forall K \in \mathcal{K}$ the state covariance matrix Σ_K satisfies*

$$\lambda_{\min}(\Sigma_K) \ge \mu_\xi > 0. \tag{44}$$

Then the average cost $\mathcal{J}_{\mathrm{avg}}(K)$ is gradient dominant in $\mathcal{K}$, i.e. for any $K \in \mathcal{K}$,

$$\mathcal{J}_{\mathrm{avg}}(K) - \mathcal{J}_{\mathrm{avg}}(K_*) \le \frac{\|\Sigma_{K_*}\|}{4\mu_\xi^2\, \lambda_{\min}(\mathscr{R})} \|\nabla \mathcal{J}_{\mathrm{avg}}(K)\|_F^2, \tag{45}$$

where K_* is an optimal policy.

Proof. For any $K \in \mathcal{K}$ and A_K is Schur stable by Assumption 4.3. Hence, the solutions P_K and Σ_K given by (25) and (26), respectively, are unique. Then, using the LQR policy-gradient $\nabla_K \mathcal{J}(K) = 2E_K\Sigma_K$, $i \in \{N\}$, we have

$$\nabla_K \mathcal{J}_{\text{avg}}(K) = 2\sum_{i=1}^{N} \omega_{a,k}^{(i)} E_K \Sigma_K, \tag{46}$$

where

$$E_K := \big(\mathscr{R} + B^\top P_K B\big)K - B^\top P_K A. \tag{47}$$

By Lemma 11 in (Fazel et al., 2018), for $\{i \in N\}$,

$$\mathcal{J}^{(i)}(K) - \mathcal{J}^{(i)}(K_*) \le \|\Sigma_{K_*}\| \operatorname{tr}\Big(E_K^\top \big(\mathscr{R} + B^\top P_K B\big)^{-1} E_K\Big). \tag{48}$$

Since $B^\top P_K B \succeq 0$, we have $\mathscr{R} + B^\top P_K B \succeq \mathscr{R}, \mathscr{R} + B^\top P_K B \succeq \mathscr{R}$, and thus $\big(\mathscr{R} + B^\top P_K B\big)^{-1} \preceq \frac{1}{\lambda_{\min}(\mathscr{R})} I$. Hence,

$$\mathcal{J}^{(i)}(K) - \mathcal{J}^{(i)}(K_*) \le \frac{\|\Sigma_{K_*}\|}{\lambda_{\min}(\mathscr{R})} \operatorname{tr}\left(E_K^\top E_K\right). \tag{49}$$

Averaging over agent i yields

$$\begin{aligned}\mathcal{J}_{\text{avg}}(K) - \mathcal{J}_{\text{avg}}(K_*) &= \sum_{i=1}^{N} \omega_{a,k}^{(i)} \big(\mathcal{J}^{(i)}(K) - \mathcal{J}^{(i)}(K_*)\big) \\ &\le \sum_{i=1}^{N} \omega_{a,k}^{(i)} \frac{\|\Sigma_{K_*}\|}{\lambda_{\min}(\mathscr{R})} \|E_K\|_F^2. \end{aligned} \tag{50}$$

Now, $E_K = \frac{1}{2}\nabla_K \mathcal{J}(K)\Sigma_K^{-1}$ and $\frac{1}{\lambda_{\min}(\Sigma_K)} \le \frac{1}{\mu_\xi}$ from (44), so that

$$\begin{aligned}\sum_{i=1}^{N} \omega_{a,k}^{(i)} \|E_K\|_F^2 &\le \frac{1}{4}\sum_{i=1}^{N} \omega_{a,k}^{(i)} \|\nabla_K \mathcal{J}^{(i)}(K)\|_F^2 \big\|\Sigma_K^{-1}\big\|^2 \\ &\le \frac{1}{4\mu_\xi^2}\sum_{i=1}^{N} \omega_{a,k}^{(i)} \|\nabla_K \mathcal{J}^{(i)}(K)\|_F^2. \end{aligned} \tag{51}$$

Substituting (51) into (50) concludes (45). This completes the proof. ■

Lemma 6. *For each iteration k, the approximation error between the covariance matrices of the standard FedNPG and the ED-based FedNPG satisfies the following with $p_k^{(i)} = p_k$ and $q_k^{(i)} = q_k$ for all agents i:*

$$\big\|\widehat{\Sigma}_{K_k} - \bar{\Sigma}_{K_k}\big\|_2 \le \sum_{i=1}^{N} \omega_{a,k}^{(i)} \left(\widehat{\lambda}_{q_k+1,k}^{(i)} - p_k\right) =: \bar{\rho}_k. \tag{52}$$

With FedAvg, the covariance matrix of the standard FedNPG is

$$\bar{\Sigma}_{K_k} := \sum_{i=1}^{N} \omega_{a,k}^{(i)} \bar{\Sigma}_{K_k}^{(i)}, \tag{53}$$

where $\omega_{a,k}^{(i)}$ as given in (7), and its local covariance matrix for each agent i admits the representation

$$\bar{\Sigma}_{K_k}^{(i)} = \sum_{b=1}^{n} \widehat{\lambda}_{b,k}^{(i)}\, \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top}, \tag{54}$$

where the eigenvalues of the empirical covariance matrix $\bar{\Sigma}_{K_k}^{(i)}$ are ordered as

$$\widehat{\lambda}_{1,k}^{(i)} \ge \widehat{\lambda}_{2,k}^{(i)} \ge \cdots \ge \widehat{\lambda}_{q_k+1,k}^{(i)} \ge \cdots \ge \widehat{\lambda}_{n,k}^{(i)} \ge 0.$$

In addition, the approximation using the EEPs in the ED-based FedNPG is defined as

$$\widehat{\Sigma}_{K_k} := \sum_{i=1}^{N} \omega_{a,k}^{(i)} \left(\sum_{b=1}^{q_k} (\widehat{\lambda}_{b,k}^{(i)} - p_k) \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top} + p_k I \right) \tag{55}$$

from (9)*, where* $p_k^{(i)}$ *lies between* $0 < p_k \leq \widehat{\lambda}_{q_k+1,k}^{(i)}, \quad \forall (k,i)$.

Proof. Using (53) with (54) yields

$$\bar{\Sigma}_{K_k} = \sum_{i=1}^{N} \left[\omega_{a,k}^{(i)} \sum_{b=1}^{n} \widehat{\lambda}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top} \right]. \tag{56}$$

Now we define, for each agent i, the orthogonal projector onto the dominant q_k-dimensional eigenspace as

$$P_{q_k}^{(i)} := \sum_{b=1}^{q_k} \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top}. \tag{57}$$

Then the approximation of the covariance matrix of the ED-based FedNPG in (55) can be rewritten as

$$\widehat{\Sigma}_{K_k} = \sum_{i=1}^{N} \omega_{a,k}^{(i)} \left[\sum_{b=1}^{q_k} \widehat{\lambda}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top} + p_k \big(I - P_{q_k}^{(i)} \big) \right]. \tag{58}$$

Since $I - P_{q_k}^{(i)} = \sum_{b=q_k+1}^{n} \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top}$, subtracting (58) from (56) yields

$$\bar{\Sigma}_{K_k} - \widehat{\Sigma}_{K_k} = \sum_{i=1}^{N} \left[\omega_{a,k}^{(i)} \sum_{b=q_k+1}^{n} (\widehat{\lambda}_{b,k}^{(i)} - p_k) \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top} \right]. \tag{59}$$

Taking the spectral norm on both sides, we obtain the following expression

$$\big\| \widehat{\Sigma}_{K_k} - \bar{\Sigma}_{K_k} \big\|_2 \leq \sum_{i=1}^{N} \left[\omega_{a,k}^{(i)} \left\| \sum_{b=q_k+1}^{n} (\widehat{\lambda}_{b,k}^{(i)} - p_k) \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top} \right\|_2 \right]. \tag{60}$$

Since the matrices $\widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top}$ are mutually orthogonal rank-one projectors, for each i, we can represent (60) as

$$\left\| \sum_{b=q_k+1}^{n} (\widehat{\lambda}_{b,k}^{(i)} - p_k) \widehat{v}_{b,k}^{(i)} \widehat{v}_{b,k}^{(i)\top} \right\|_2 = \max_{q_k+1 \leq b \leq n} \big| \widehat{\lambda}_{b,k}^{(i)} - p_k \big|. \tag{61}$$

Moreover, by the condition of p_k and because the eigenvalues are ordered, we obtain

$$\max_{q_k+1 \leq b \leq n} \left| \widehat{\lambda}_{b,k}^{(i)} - p_k \right| = \widehat{\lambda}_{q_k+1,k}^{(i)} - p_k.$$

Hence,

$$\big\| \widehat{\Sigma}_{K_k} - \bar{\Sigma}_{K_k} \big\|_2 \leq \sum_{i=1}^{N} \omega_{a,k}^{(i)} \left(\widehat{\lambda}_{q_k+1,k}^{(i)} - p_k \right) = \bar{\rho}_k. \tag{62}$$

This completes the proof. ■

Lemma 7. *For a stable policy* π_{K_k} *of the true FedNPG, the following inequality*

$$\big\| \Sigma'_{K_k} \big\| \leq \mathcal{J}_{\text{avg}}(K_0) / \lambda_{\min}(\mathcal{Q}) \tag{63}$$

holds.

Proof. It follows from the single-agent case [see, Lemma C.1 in (Yang et al., 2019)] and extends to the federated setting. Using (24), we can write

$$\mathcal{J}_{\text{avg}}(K_0) \geq \sum_{i=1}^{N} \omega_{a,k}^{(i)} \operatorname{tr}\left[\left(\mathscr{Q} + K^\top \mathscr{R} K\right)\Sigma'_{K_k}\right] = \operatorname{tr}\left[\left(\mathscr{Q} + K^\top \mathscr{R} K\right)\Sigma'_{K_k}\right] \geq \lambda_{\min}(\mathscr{Q}) \operatorname{tr}\left(\Sigma'_{K_k}\right)$$
$$\geq \lambda_{\min}(\mathscr{Q}) \left\|\Sigma'_{K_k}\right\|. \tag{64}$$

This completes the proof. ■

G.3 Proof of Theorem 1

Proof. The proof proceeds in three steps as follows:

Step 1: We consider FedNPG updates using the ED-based FIM, the full FIM, and the exact FIM. At k-th iteration, the ED Fisher matrix, the full Fisher matrix, and the true Fisher matrix-based policy updates from the same K_k value are given by

$$K_{k+1} = K_k - \eta_k \, \widehat{\nabla}_K \mathcal{J}_{\text{avg}}(K_k) \, \widehat{\Sigma}_{K_k}^{-1}, \tag{65}$$
$$\bar{K}_{k+1} = K_k - \eta_k \, \bar{\nabla}_K \mathcal{J}_{\text{avg}}(K_k) \, \bar{\Sigma}_{K_k}^{-1}, \tag{66}$$
$$K'_{k+1} = K_k - \eta_k \, \nabla'_K \mathcal{J}_{\text{avg}}(K_k) \, \Sigma'^{-1}_{K_k}, \tag{67}$$

respectively, with step size $0 < \eta_k < 2/\mathcal{L}$ (Hu et al., 2023). We define $\widehat{G}_{K_k} := \widehat{\nabla}_K \mathcal{J}_{\text{avg}}(K_k)$, $\bar{G}_{K_k} := \bar{\nabla}_K \mathcal{J}_{\text{avg}}(K_k)$, $G'_{K_k} := \nabla'_K \mathcal{J}_{\text{avg}}(K_k)$. Using (65), we can write

$$K_{k+1} = K_k - \eta_k \widehat{G}_{K_k} \widehat{\Sigma}_{K_k}^{-1}, \tag{68}$$

where $\widehat{\Sigma}_{K_k}$ denotes the covariance approximation using the FedNPG-ED. Similarly, from (66) and (67), we have

$$\bar{K}_{k+1} = K_k - \eta_k \bar{G}_{K_k} \bar{\Sigma}_{K_k}^{-1}, \tag{69}$$
$$K'_{k+1} = K_k - \eta_k G'_{K_k} \Sigma'^{-1}_{K_k}, \tag{70}$$

where $\bar{\Sigma}_{K_k}$ and Σ'_{K_k} are the full covariance matrix without compression estimated using sample data and the true covariance matrix, from the standard FedNPG and the true FedNPG, respectively.

Since the Fisher matrix admits the Kronecker representation with respect to $\text{vec}(K)$, the inverses of the covariance matrix form of the natural gradient step in (68), (69), and (70) are equivalent to the inverse-Fisher forms. Now, subtracting (70) from (69) and (69) from (68) gives

$$\bar{K}_{k+1} - K'_{k+1} = -\eta_k \left[\bar{G}_{K_k}\left(\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\right) + \left(\bar{G}_{K_k} - G'_{K_k}\right)\Sigma'^{-1}_{K_k}\right], \tag{71}$$
$$K_{k+1} - \bar{K}_{k+1} = -\eta_k \left[\widehat{G}_{K_k}\left(\widehat{\Sigma}_{K_k}^{-1} - \bar{\Sigma}_{K_k}^{-1}\right) + \left(\widehat{G}_{K_k} - \bar{G}_{K_k}\right)\bar{\Sigma}_{K_k}^{-1}\right]. \tag{72}$$

Note that since the gradients $\bar{G}_{K_k}$ and $\widehat{G}_{K_k}$ are the same in the standard FedNPG and ED-based FedNPG cases, we consider $\bar{G}_{K_k} = \widehat{G}_{K_k} = G_{K_k}$. Furthermore, taking the spectral norm in (71) and (72) reduces to

$$\|\bar{K}_{k+1} - K'_{k+1}\|_2 \leq \eta_k \left[\|G_{K_k}\|_2 \|\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\|_2 + \|G_{K_k} - G'_{K_k}\|_2 \|\Sigma'^{-1}_{K_k}\|_2\right], \tag{73}$$
$$\|K_{k+1} - \bar{K}_{k+1}\|_2 \leq \eta_k \|G_{K_k}\|_2 \|\widehat{\Sigma}_{K_k}^{-1} - \bar{\Sigma}_{K_k}^{-1}\|_2. \tag{74}$$

Step 2(a): We compute the matrix concentration (Tropp, 2015) bounds for (73) and (74). Since the concentration inequality involves a probabilistic quantity, the empirical covariance matrix of the standard FedNPG is the sample mean, i.e.,

$$\bar{\Sigma}_{K_k} = \sum_{i=1}^{N} \omega_{a,k}^{(i)} \bar{\Sigma}_{K_k}^{(i)} = \sum_{i=1}^{N} \omega_{a,k}^{(i)} \left(\frac{1}{\tau_{i,k}} \sum_{t=1}^{\tau_{i,k}} x_t^{(i)} x_t^{(i)\top}\right)$$
$$= \frac{1}{\sum_i \tau_{i,k}} \sum_{t=1}^{\sum_i \tau_{i,k}} x_t x_t^\top. \tag{75}$$

Note that states along a closed-loop trajectory are generally dependent. Our analysis adopts an idealized sampling model in which samples are drawn independently from the stationary distribution of the current policy. In practice, random sampling from a replay buffer can reduce temporal dependence but does not guarantee independence. Independent samples can be obtained from independently generated roll-outs, taking one sample from each. Alternatively, sufficiently separated samples from a geometrically mixing trajectory can be treated as approximately independent. Nevertheless, the iid sampling assumption is standard in the similar literature. From the above sample mean, we can write the following form using the concentration inequality:

$$\Pr\Big(\|\bar{\Sigma}_{K_k} - \Sigma'_{K_k}\|_2 \geq \varepsilon_{f_1}\Big) \leq \delta_{f_1}, \tag{76}$$

where

$$\delta_{f_1} := 2n \exp\left(\frac{-\sum_i \tau_{i,k}\, \varepsilon_{f_1}^2}{2c_x\left(\left\|\Sigma'_{K_k}\right\|_2 + \frac{2\varepsilon_{f_1}}{3}\right)}\right), \tag{77}$$

$\|x_t\| \leq c_x < \infty$, $c_x > 0$, and the scalar $\varepsilon_{f_1} > 0$ is very small. Therefore, the quantity $\|\bar{\Sigma}_{K_k} - \Sigma'_{K_k}\|_2$ can be made small as $\tau_{i,k}$ becomes large, with a very high probability, i.e.,

$$\|\bar{\Sigma}_{K_k} - \Sigma'_{K_k}\|_2 \leq \varepsilon_{f_1}, \quad w.p.\,(1-\delta_{f_1}). \tag{78}$$

Now, the perturbation theory of matrix inverse admits the representation of

$$\|\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\|_2 \leq \|\Sigma'^{-1}_{K_k}\|_2 \frac{\|\bar{\Sigma}_{K_k} - \Sigma'_{K_k}\|_2}{1 - \|\Sigma'^{-1}_{K_k}\|_2 \|\bar{\Sigma}_{K_k} - \Sigma'_{K_k}\|_2}, \tag{79}$$

implying that

$$\|\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\|_2 \leq \frac{\|\Sigma'^{-1}_{K_k}\|_2^2\, \varepsilon_{f_1}}{1 - \|\Sigma'^{-1}_{K_k}\|_2\, \varepsilon_{f_1}}, \quad w.\,p.\,(1-\delta_{f_1}), \tag{80}$$

This holds provided that

$$\|\Sigma'^{-1}_{K_k}\|_2\, \|\bar{\Sigma}_{K_k} - \Sigma'_{K_k}\|_2 \leq 1. \tag{81}$$

Since ε_{f_1} is very small as $\tau_{i,k}$ goes large, $\|\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\|_2$ is small with a high probability.

The next step replaces the iterate-dependent quantity $\left\|\Sigma'^{-1}_{K_k}\right\|_2$ by a constant. By Assumption 4.2 and (44), $\Sigma'_K \succeq \mu_\xi I$ for every $K \in \mathcal{K}$, therefore,

$$\sup_{K \in \mathcal{S}(k_\epsilon)} \left\|\Sigma'^{-1}_K\right\|_2 \;\leq\; \frac{1}{\mu_\xi} =: \bar{\sigma}_{\text{inv}}. \tag{82}$$

Using (82), we rewrite (80) in the form of

$$\|\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\|_2 \leq \frac{\bar{\sigma}_{\text{inv}}^2\, \varepsilon_{f_1}}{1 - \bar{\sigma}_{\text{inv}} \varepsilon_{f_1}}, \quad w.\,p.\,(1-\delta_{f_1}). \tag{83}$$

Therefore,

$$\Pr\Big(\|\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\|_2 \geq \varepsilon_{b_1}\Big) \leq \delta_{f_1}, \tag{84}$$

where, from (77),

$$\delta_{f_1} = 2n \exp\left(\frac{-\sum_i \tau_{i,k}\, \varepsilon_{f_1}^2}{2c_x\left(\left\|\Sigma'_{K_k}\right\|_2 + \frac{2\varepsilon_{f_1}}{3}\right)}\right) \leq 2n \exp\left(\frac{-\sum_i \tau_{i,k}\, \varepsilon_{f_1}^2}{2c_x\left(\frac{\mathcal{J}_{\text{avg}}(K_0)}{\lambda_{\min}(\mathcal{Q})} + \frac{2\varepsilon_{f_1}}{3}\right)}\right) \tag{85}$$

by Lemma 7. Finally, (83) takes the form of

$$\|\bar{\Sigma}_{K_k}^{-1} - \Sigma'^{-1}_{K_k}\|_2 \leq \varepsilon_{b_1}, \quad w.p.\,(1-\delta_{f_1}), \tag{86}$$

with

$$\varepsilon_{b_1} \leq \frac{\bar{\sigma}_{\text{inv}}^2 \, \varepsilon_{f_1}}{1 - \bar{\sigma}_{\text{inv}} \varepsilon_{f_1}} \; = \; O(\varepsilon_{f_1}). \tag{87}$$

Therefore, from (73), we have

$$\|\bar{K}_{k+1} - K'_{k+1}\|_2 \leq \eta_k \Big[\|G_{K_k} - G'_{K_k}\|_2 \, \|\Sigma'^{-1}_{K_k}\|_2 + \|G_{K_k}\|_2 \, \varepsilon_{b_1} \Big], \quad w.p. \, (1 - \delta_{f_1}). \tag{88}$$

Step 2(b): Next, we bound the deviation between the estimated gradient and the exact gradient appearing in (73). The estimated gradient and the true gradient are

$$G_{K_k} = \sum_{i=1}^{N} \omega_{a,k}^{(i)} \left(\frac{1}{\tau_{i,k}} \sum_{t=1}^{\tau_{i,k}} \bar{g}_t^{(i)} \right) = \frac{1}{\sum_i \tau_{i,k}} \sum_{t=1}^{\sum_i \tau_{i,k}} \bar{g}_t, \tag{89}$$

where

$$\bar{g}_t^{(i)} := \widehat{Q}_k \Big(x_t^{(i)}, u_t^{(i)} \Big) \, \nabla_K \log \pi_{K_k} \Big(u_t^{(i)} \mid x_t^{(i)} \Big), \tag{90}$$

and

$$G'_{K_k} = \mathop{\mathbb{E}}_{(x,u)} \left[Q_k \left(x, u \right) \nabla_K \log \pi_{K_k}(u \mid x) \right]. \tag{91}$$

Since our focus is the convergence of the actor, the critic is handled through Assumption 4.5: $\widehat{Q}_k = Q_{K_k}$, so $\bar{g}_t$ is an unbiased estimator of G'_{K_k} and the only error remaining in ε_{g_1} is finite-sample. We stress that this is an assumption on the critic, not an analysis of it. Therefore, we have

$$X_Q = Q(x_t, u_t) \, \nabla_K \log \pi_{K_k}(u_t \mid x_t) - \mathop{\mathbb{E}}_{(x,u)} \left[Q \left(x, u \right) \nabla_K \log \pi_{K_k}(u \mid x) \right], \tag{92}$$

where $\|X_Q\| \leq Q_{\max}$, and $Q_{\max} > 0 \; \forall (x, u)$ almost surely. Now we define the variance

$$\sigma_Q^2 := \max \left\{ \left\| \mathbb{E}[X_Q X_Q^\top] \right\|_2, \left\| \mathbb{E}[X_Q^\top X_Q] \right\|_2 \right\} \tag{93}$$

and obtain for any $\varepsilon_{g_1} > 0$,

$$\Pr\left(\left\| G_{K_k} - G'_{K_k} \right\|_2 \geq \varepsilon_{g_1} \right) \leq \delta_g, \tag{94}$$

implying that

$$\left\| G_{K_k} - G'_{K_k} \right\|_2 \leq \varepsilon_{g_1}, \quad w.p. \; (1 - \delta_g), \tag{95}$$

where $\delta_g := (n+p) \exp\left(\frac{- \sum_i \tau_{i,k} \, \varepsilon_{g_1}^2 / 2}{\sigma_Q^2 + Q_{\max} \, \varepsilon_{g_1} / 3} \right)$, and n and p are the dimensions of x and u, respectively.

Smaller σ_Q^2 means that the score-weighted Q-samples are more concentrated around their expectation. In contrast, $Q_{\max}$ controls the largest single-sample deviation possible from the exact gradient, as given in (92). So, the Bernstein denominator $(\sigma_Q^2 + \frac{Q_{\max} \varepsilon_{g_1}}{3})$ contains both the variance information and the worst-case sample magnitude. Most importantly, because $\sum_i \tau_{i,k}$, the probability behaves as $\delta_g \sim \exp(- \sum_i \tau_{i,k} \times \text{constant})$. Therefore, increasing the total number of samples among the participating agents improves the concentration of G_{K_k} around G'_{K_k}. Now, using the uniform bound (82), the first term in the RHS of (88) can be represented as

$$\left\| G_{K_k} - G'_{K_k} \right\|_2 \left\| \Sigma'^{-1}_{K_k} \right\|_2 \leq \varepsilon_{g_1} \bar{\sigma}_{\text{inv}} = \varepsilon_g, \quad w.p. \; (1 - \delta_g). \tag{96}$$

Finally, on substitution of (96) into (88), we find the bound for the deviation of (73), i.e.,

$$\|\bar{K}_{k+1} - K'_{k+1}\|_2 \leq \eta_k \left(\varepsilon_g + \|G_{K_k}\|_2 \, \varepsilon_{b_1} \right), \tag{97}$$

with probability at least $(1 - \delta_{f_1} - \delta_g)$.

Step 3(a): We express the gradient dominance property of Lemma 5 in terms of the sampled gradient G_{K_k}, which is the quantity that drives the update (68). Throughout the remainder of the proof we write

$$c := \frac{4 \mu_\xi^2 \, \lambda_{\min}(\mathscr{R})}{\|\Sigma_{K_*}\|}, \qquad f(s) := s - \frac{\mathcal{L}}{2} s^2. \tag{98}$$

We know from Lemma 5 that the average cost deviation of the true FedNPG from the optimal cost is dominated by the true gradient G'_{K_k}, and applying $\|\cdot\|_F \leq \sqrt{r}\,\|\cdot\|_2$ for rank-r matrices produces

$$\mathcal{J}_{\text{avg}}(K_k) - \mathcal{J}_{\text{avg}}(K_*) \leq \frac{\|\Sigma_{K_*}\|\, r}{4\mu_\xi^2\, \lambda_{\min}(\mathscr{R})} \left\|G'_{K_k}\right\|_2^2. \tag{99}$$

Moreover, using (95) and the triangle inequality, we obtain the upper bound for G'_{K_k} as

$$\left\|G'_{K_k}\right\|_2 \leq \left\|G_{K_k} - G'_{K_k}\right\|_2 + \|G_{K_k}\|_2 = \varepsilon_{g_1} + \|G_{K_k}\|_2. \tag{100}$$

Now, we substitute the true gradient in (99) for its upper bound in (100),

$$\mathcal{J}_{\text{avg}}(K_k) - \mathcal{J}_{\text{avg}}(K_*) \leq \frac{\|\Sigma_{K_*}\|\, r}{4\mu_\xi^2\, \lambda_{\min}(\mathscr{R})} \left(\varepsilon_{g_1}^2 + \|G_{K_k}\|_2^2\right), \qquad \text{w.p. } (1-\delta_g), \tag{101}$$

to get

$$\frac{c}{r}\left(\mathcal{J}_{\text{avg}}(K_k) - \mathcal{J}_{\text{avg}}(K_*)\right) - \varepsilon_{g_1}^2 \leq \|G_{K_k}\|_2^2, \qquad \text{w.p. } (1-\delta_g), \tag{102}$$

which lower bounds the squared norm of the sampled gradient by the optimality gap, up to the sampling error ε_{g_1}.

Step 3(b): We demonstrate that the policy in the ED-based FedNPG converges to a neighbourhood of the global optimal policy K_* at a linear rate. Using $\mathcal{L}$-smoothness in Lemma 4 for the average cost deviation holds

$$\mathcal{J}_{\text{avg}}(K_{k+1}) \leq \mathcal{J}_{\text{avg}}(K_k) + \langle G_{K_k}, (K_{k+1} - K_k)\rangle + \frac{\mathcal{L}}{2}\|K_{k+1} - K_k\|_F^2. \tag{103}$$

Replacement with (68) results in

$$\mathcal{J}_{\text{avg}}(K_{k+1}) \leq \mathcal{J}_{\text{avg}}(K_k) - \eta_k \left\langle G_{K_k}, G_{K_k}\widehat{\Sigma}_{K_k}^{-1}\right\rangle + \frac{\mathcal{L}}{2}\eta_k^2 \|G_{K_k}\|_F^2 \left\|\widehat{\Sigma}_{K_k}^{-1}\right\|_2^2. \tag{104}$$

Using the extreme eigenvalues of $\widehat{\Sigma}_{K_k}^{-1}$, and the fact that $\lambda_{\max}(\widehat{\Sigma}_{K_k}^{-1}) = 1/p_k$ by (9),

$$\begin{aligned}
\mathcal{J}_{\text{avg}}(K_{k+1}) &\leq \mathcal{J}_{\text{avg}}(K_k) - \eta_k \|G_{K_k}\|_F^2\, \lambda_{\min}\left(\widehat{\Sigma}_{K_k}^{-1}\right) + \frac{\mathcal{L}}{2}\eta_k^2 \|G_{K_k}\|_F^2 \lambda_{\max}\left(\widehat{\Sigma}_{K_k}^{-1}\right)^2, \\
&\leq \mathcal{J}_{\text{avg}}(K_k) - \left(\frac{\eta_k}{p_k} - \frac{\mathcal{L}}{2}\frac{\eta_k^2}{p_k^2}\right)\|G_{K_k}\|_F^2 \\
&\leq \mathcal{J}_{\text{avg}}(K_k) - r\left(\frac{\eta_k}{p_k} - \frac{\mathcal{L}}{2}\frac{\eta_k^2}{p_k^2}\right)\|G_{K_k}\|_2^2 \;=\; \mathcal{J}_{\text{avg}}(K_k) - r\, f(\widehat{s}_k)\, \|G_{K_k}\|_2^2,
\end{aligned} \tag{105}$$

where $\widehat{s}_k := \eta_k/p_k$ and f is defined in (98). With the gradient dominance property of Step 3(a), that is (102), we obtain

$$\mathcal{J}_{\text{avg}}(K_{k+1}) \leq \mathcal{J}_{\text{avg}}(K_k) - r\, f(\widehat{s}_k)\left[\frac{c}{r}\left(\mathcal{J}_{\text{avg}}(K_k) - \mathcal{J}_{\text{avg}}(K_*)\right) - \varepsilon_{g_1}^2\right], \qquad \text{w.p. } (1-\delta_g). \tag{106}$$

Now, subtracting the average optimal cost from both sides and rearranging leads to the recursion of the ED-based FedNPG as

$$\mathcal{J}_{\text{avg}}(K_{k+1}) - \mathcal{J}_{\text{avg}}(K_*) \leq \left(1 - \widehat{\beta}\right)\left(\mathcal{J}_{\text{avg}}(K_k) - \mathcal{J}_{\text{avg}}(K_*)\right) + \widehat{\varepsilon}_L, \qquad \text{w.p. } (1-\delta_g), \tag{107}$$

with the contraction factor

$$\widehat{\beta} := \frac{4\mu_\xi^2\, \lambda_{\min}(\mathscr{R})}{\|\Sigma_{K_*}\|}\left(\frac{\eta_k}{p_k} - \frac{\mathcal{L}}{2}\frac{\eta_k^2}{p_k^2}\right) \;=\; c\, f(\widehat{s}_k), \tag{108}$$

and the residual error

$$\widehat{\varepsilon}_L := r\left(\frac{\eta_k}{p_k} - \frac{\mathcal{L}}{2}\frac{\eta_k^2}{p_k^2}\right)\varepsilon_{g_1}^2 \;=\; r\, f(\widehat{s}_k)\, \varepsilon_{g_1}^2. \tag{109}$$

The convergence of the FedNPG-ED to the optimal policy given in (107) can be guaranteed if $\widehat{\beta} \in (0,1)$. From (108), $\widehat{\beta} > 0$ requires

$$f(\widehat{s}_k) > 0 \implies \widehat{s}_k \left(1 - \frac{\mathcal{L}}{2}\widehat{s}_k\right) > 0 \implies \widehat{s}_k < \frac{2}{\mathcal{L}}, \tag{110}$$

as $c > 0$. This condition alone is not enough to show that $\widehat{\beta} < 1$; a sufficient condition for the step size is

$$c\,\widehat{s}_k < 1 \implies \widehat{s}_k < \frac{1}{c} = \frac{\|\Sigma_{K_*}\|}{4\mu_\xi^2\,\lambda_{\min}(\mathscr{R})}, \tag{111}$$

because $f(\widehat{s}_k) < \widehat{s}_k$. Therefore, the step size is selected so that

$$0 < \frac{\eta_k}{p_k} < \min\left\{\frac{2}{\mathcal{L}}, \frac{\|\Sigma_{K_*}\|}{4\mu_\xi^2\,\lambda_{\min}(\mathscr{R})}\right\}, \tag{112}$$

and then $0 < \widehat{\beta} < 1$, so that FedNPG-ED converges linearly to a neighbourhood of K_* of radius $\widehat{\varepsilon}_L/\widehat{\beta}$.

Dependence of $\widehat{\beta}$ on the approximation parameter. It remains to characterise how $\widehat{\beta}$ varies with p_k, for a fixed η_k. Differentiating (108),

$$\frac{\partial\widehat{\beta}}{\partial p_k} = c\left(-\frac{\eta_k}{p_k^2} + \mathcal{L}\frac{\eta_k^2}{p_k^3}\right) = c\,\frac{\eta_k}{p_k^2}\left(\mathcal{L}\frac{\eta_k}{p_k} - 1\right). \tag{113}$$

Since $c\,\eta_k/p_k^2 > 0$, the sign of (113) is that of $(\mathcal{L}\eta_k/p_k - 1)$: the derivative is strictly positive for $p_k < \mathcal{L}\eta_k$, vanishes only at $p_k = \mathcal{L}\eta_k$, and is strictly negative for $p_k > \mathcal{L}\eta_k$. Hence $\widehat{\beta}$ increases on $(0, \mathcal{L}\eta_k)$ and decreases on $(\mathcal{L}\eta_k, \infty)$, so $p_k = \mathcal{L}\eta_k$ is the unique stationary point and therefore the unique maximiser of $\widehat{\beta}$ over $p_k > 0$, with

$$\widehat{\beta}\big|_{p_k = \mathcal{L}\eta_k} = c\,f(1/\mathcal{L}) = \frac{c}{2\mathcal{L}}. \tag{114}$$

Note thatNo second-order test is needed since the sign change of (113) from positive to negative already identifies the maximum, and the uniqueness of the stationary point rules out any other local extremum on $p_k > 0$.

Two consequences follow, both visible in the illustrative Figure 6(a), which plots $\widehat{\beta}$ against p_k for a fixed η_k. First, over the admissible range (112), i.e., $\widehat{s}_k < \min\{1/\mathcal{L}, 1/c\}$, the function f is strictly increasing, so $\widehat{\beta}$ is strictly decreasing in p_k and the largest certified contraction is approached at the boundary

$$p_k^\star := \eta_k \max\{\mathcal{L}, c\}, \tag{115}$$

which coincides with the unconstrained maximiser $\mathcal{L}\eta_k$ whenever $c \le \mathcal{L}$. In Figure 6(a) the markers show $p_k^\star$ for the two regimes, it sits exactly at the peak for $c \le \mathcal{L}$, and to the right of the peak when $c > \mathcal{L}$, where the condition $c\,\widehat{s}_k < 1$ binds first. Second, p_k cannot be shrunk without limit. For $p_k \le \mathcal{L}\eta_k/2$, equivalently $\widehat{s}_k \ge 2/\mathcal{L}$, one has $f(\widehat{s}_k) \le 0$ by (110) and the bound (107) becomes vacuous; this is the shaded strip on the left of Figure 6(a). A too small p_k over-relaxes the ascent direction along the discarded eigen-subspace, for which no curvature information is transmitted, and (112) is exactly what excludes that regime. Figure 6(b) shows $f(\cdot)$ as a function of $\widehat{s}_k$, where f is the concave parabola with its peak at $\widehat{s}_k = 1/\mathcal{L}$ and its second root at $\widehat{s}_k = 2/\mathcal{L}$.

Restricting the step size to the increasing branch of f, i.e., imposing

$$0 < \frac{\eta_k}{p_k} < \min\left\{\frac{1}{\mathcal{L}}, \frac{\|\Sigma_{K_*}\|}{4\mu_\xi^2\,\lambda_{\min}(\mathscr{R})}\right\}, \tag{116}$$

which is a tightening of (112), therefore guarantees the monotone relation

$$p_k \downarrow \implies \widehat{s}_k \uparrow \implies \widehat{\beta} \uparrow, \tag{117}$$

and it is (116) that is stated as the step-size condition of Theorem 1. This completes the proof.

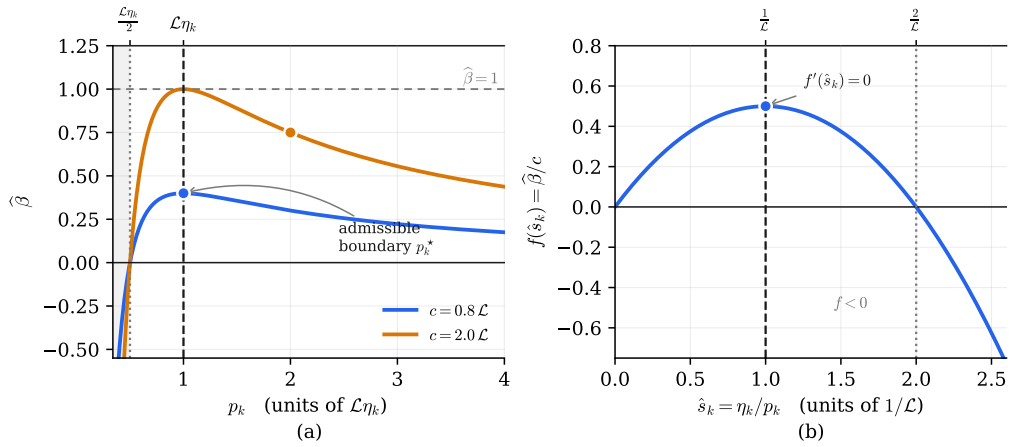

Figure 6: (a) $\widehat{\beta}$ versus p_k, (b)$f(\cdot)$ versus $\widehat{s}_k$.

Remark 3 (Exact admissibility of the maximiser). The condition $c\,\widehat{s}_k < 1$ used in (111) is sufficient but not necessary for $\widehat{\beta} < 1$; the exact requirement is $c\,f(\widehat{s}_k) < 1$. At the maximiser $\widehat{s}_k = 1/\mathcal{L}$ this reads $c < 2\mathcal{L}$ by (114). Whenever the problem constants satisfy $c < 2\mathcal{L}$, the design rule $p_k = \mathcal{L}\eta_k$ is therefore admissible and attains the fastest certified contraction $\widehat{\beta} = c/(2\mathcal{L})$.

Remark 4 (Uncompressed FedNPG as a reference). Repeating Step 3(b) with $\bar{\Sigma}_{K_k}$ in place of $\widehat{\Sigma}_{K_k}$, so that $\widehat{\lambda}_{n,k}$ replaces p_k and $\bar{s}_k := \eta_k/\widehat{\lambda}_{n,k}$ replaces $\widehat{s}_k$ throughout, yields the corresponding recursion for the uncompressed FedNPG, with

$$\bar{\beta} := c\,f(\bar{s}_k) \tag{118}$$

and

$$\bar{\varepsilon}_L := r\,f(\bar{s}_k)\,\varepsilon_{g_1}^2, \tag{119}$$

under (112) with p_k replaced by $\widehat{\lambda}_{n,k}$. Unlike p_k, the smallest eigenvalue $\widehat{\lambda}_{n,k}$ is determined by the data rather than by the designer, so no design freedom is associated with $\bar{\beta}$. These quantities serve as the reference against which the compressed rate is compared in Appendix G.5.

■

G.4 Proof of Corollary 1

Proof. Fix $k \le C$. Theorem 1 gives (14) on an event $\mathcal{E}_k$ with $\mathbb{P}(\mathcal{E}_k^c) \le \delta_g + \delta_{f_1}$. On $\mathcal{E} := \bigcap_{k \le C} \mathcal{E}_k$, which by the union bound satisfies $\mathbb{P}(\mathcal{E}^c) \le C(\delta_g + \delta_{f_1})$, the recursion holds at every round and can be unrolled,

$$\Delta_k \le (1-\widehat{\beta})^k \Delta_0 + \widehat{\varepsilon}_L \sum_{j=0}^{k-1} (1-\widehat{\beta})^j \le (1-\widehat{\beta})^k \Delta_0 + \frac{\widehat{\varepsilon}_L}{\widehat{\beta}},$$

which is (18). The geometric sum converges because $\widehat{\beta} \in (0,1)$. Since $\widehat{\beta}$ and $\widehat{\varepsilon}_L$ share the factor $f(\widehat{s}_k)$, that factor cancels in the floor and $\Delta_\infty = r\,\varepsilon_{g_1}^2/c$, which depends on neither η_k, p_k nor q. Splitting the target as $\Delta_\infty \le \epsilon/2$ and $(1-\widehat{\beta})^C \Delta_0 \le \epsilon/2$ gives, respectively, $\varepsilon_{g_1}^2 \le c\,\epsilon/(2r)$ and, using $\ln(1/(1-\widehat{\beta})) \ge \widehat{\beta}$,

$$C(\epsilon) = \left\lceil \frac{1}{\widehat{\beta}} \ln \frac{2\Delta_0}{\epsilon} \right\rceil. \tag{120}$$

Finally, requiring $C(\delta_g + \delta_{f_1}) \le \delta$ and inverting (16) and (77) for the per-round sample budget gives the budget stated in Corollary 1, in which the dimension factor $(3n+p)$ collects the two matrix-Bernstein union bounds. The budget grows only logarithmically in $C(\epsilon)$, whereas the confidence degrades linearly; this is the price of requiring the bound to hold simultaneously at every round, which the per-round statement of Theorem 1 does not provide. ■

G.5 Detailed discussion on the selection of p_k and q

This appendix provides the derivations underlying the guideline of Section 4. Throughout we use the notation of (17), together with

$$\bar{s}_k := \frac{\eta_k}{\lambda_{\min,k}}, \quad f(\bar{s}_k) := \bar{s}_k - \frac{\mathcal{L}}{2}\bar{s}_k^2, \quad \lambda_{\min,k} := \lambda_{\min}\big(\widehat{\mathcal{F}}_k\big) \equiv \widehat{\lambda}_{n,k}, \quad a_k := \mathcal{L}\bar{s}_k = \frac{\mathcal{L}\eta_k}{\lambda_{\min,k}}. \tag{121}$$

By (108) and (109), the contraction factor and the residual error of FedNPG-ED are $\widehat{\beta} = c\, f(\widehat{s}_k)$ and $\widehat{\varepsilon}_L = r f(\widehat{s}_k)\varepsilon_{g_1}^2$, whereas by (118) and (119) the corresponding quantities of the uncompressed FedNPG are $\bar{\beta} = c\, f(\bar{s}_k)$ and $\bar{\varepsilon}_L = r f(\bar{s}_k)\varepsilon_{g_1}^2$. Note also that, by (9) and the standing condition $0 < p_k^{(i)} \le \widehat{\lambda}_{n,k}^{(i)}$,

$$\operatorname{eig}\big(\widehat{\mathcal{F}}_{\mathrm{ed},k}\big) = \big\{\widehat{\lambda}_{1,k}, \ldots, \widehat{\lambda}_{q,k}, \underbrace{p_k, \ldots, p_k}_{d_\theta - q}\big\}, \quad \text{so that} \quad \lambda_{\min}\big(\widehat{\mathcal{F}}_{\mathrm{ed},k}\big) = p_k \ \text{ for every } q < d_\theta. \tag{122}$$

(a) Reparametrisation and monotonicity in p_k. Both $\widehat{\beta}$ and $\widehat{\varepsilon}_L$ are functions of the single scalar $\widehat{s}_k = \eta_k/p_k$; consequently, the rescaling $(\eta_k, p_k) \mapsto (\alpha\eta_k, \alpha p_k)$, $\alpha > 0$, leaves the guarantee of Theorem 1 unchanged. Thus, p_k is not an additional degree of freedom on top of the learning rate: it *reparametrises* it. As mentioned in the Numerical Results section (Section 5), for our numerical study, we have used an adaptive learning rate standard for NPG (Rajeswaran et al., 2017) with a user-defined $\alpha_a > 0$ as $\eta_k = \sqrt{\alpha_a / \big(\widehat{g}_k^\top \widehat{\mathcal{F}}_k^{-1} \widehat{g}_k\big)}$. For the FedNPG-ED case, $\widehat{\mathcal{F}}$ is replaced by $\widehat{\mathcal{F}}_{\mathrm{ed}}$. Therefore, this is of practical relevance because the adaptive rule of Section 5 makes η_k itself a function of p_k, so the quantity to be monitored during training is $\widehat{s}_k$ and not η_k in isolation. On $(0, 1/\mathcal{L})$ we have $f'(\widehat{s}_k) = 1 - \mathcal{L}\widehat{s}_k > 0$ and $f''(\widehat{s}_k) = -\mathcal{L} < 0$, *i.e.*, f is strictly increasing and strictly concave, therefore,

$$\frac{\partial \widehat{\beta}}{\partial p_k} = c\left(-\frac{\eta_k}{p_k^2} + \mathcal{L}\frac{\eta_k^2}{p_k^3}\right) = c\,\frac{\eta_k}{p_k^2}\left(\mathcal{L}\,\frac{\eta_k}{p_k} - 1\right) < 0 \qquad \text{for all } \widehat{s}_k \in (0, s^\star), \tag{123}$$

since $s^\star \le 1/\mathcal{L}$. Therefore, $\widehat{\beta}$ is strictly decreasing in p_k over the admissible range, and

$$\sup_{\widehat{s}_k \in (0, s^\star)} \widehat{\beta} = c\, f(s^\star), \quad \text{approached as } \ p_k \downarrow p_k^\star := \frac{\eta_k}{s^\star} = \eta_k \max\{\mathcal{L},\, c\}. \tag{124}$$

When the smoothness constant is an active constraint, *i.e.*, $c \le \mathcal{L}$, (124) gives the ceiling

$$\widehat{\beta} < c\, f(1/\mathcal{L}) = \frac{c}{2\mathcal{L}} = \frac{2\mu_\xi^2 \lambda_{\min}(\mathscr{R})}{\mathcal{L}\,\|\Sigma_{K_*}\|}, \tag{125}$$

which no choice of p_k can exceed. Finally, p_k cannot be shrunk without limit: $f(\widehat{s}_k) \le 0$ for $\widehat{s}_k \ge 2/\mathcal{L}$, in which case the bound of Theorem 1 is vacuous. This reflects the fact that a too small p_k *over-relaxes* the ascent direction along the discarded eigen-subspace, where no curvature information is transmitted; condition (13) is exactly what excludes this regime.

(b) Quantifying the speed-up over the uncompressed FedNPG. Let $p_k = \varrho_k \lambda_{\min,k}$ with $\varrho_k > 0$, so that $\widehat{s}_k = \bar{s}_k/\varrho_k$. Using $f(\bar{s}_k/\varrho_k) = \big(\bar{s}_k/(2\varrho_k^2)\big)(2\varrho_k - a_k)$ and $f(\bar{s}_k) = (\bar{s}_k/2)(2 - a_k)$, the ratio of the two contraction factors becomes

$$S(\varrho_k) := \frac{\widehat{\beta}}{\bar{\beta}} = \frac{\widehat{\varepsilon}_L}{\bar{\varepsilon}_L} = \frac{f(\bar{s}_k/\varrho_k)}{f(\bar{s}_k)} = \frac{2\varrho_k - a_k}{\varrho_k^2\,(2 - a_k)}. \tag{126}$$

The inequality $S(\varrho_k) > 1$ is equivalent to $(2 - a_k)\varrho_k^2 - 2\varrho_k + a_k < 0$, whose roots are $\varrho_k = \big[1 \pm (1 - a_k)\big]/(2 - a_k)$, *i.e.*, $\varrho_k = a_k/(2 - a_k)$ and $\varrho_k = 1$. Since $2 - a_k > 0$, the quadratic is strictly negative between the roots, so that

$$S(\varrho_k) > 1 \iff \varrho_k \in \left(\frac{a_k}{2 - a_k},\, 1\right). \tag{127}$$

Moreover, $\mathrm{d}S/\mathrm{d}\varrho_k = 2(a_k - \varrho_k)/\big(\varrho_k^3(2-a_k)\big)$, which is positive for $\varrho_k < a_k$ and negative for $\varrho_k > a_k$; hence, S is maximized at $\varrho_k = a_k$, *i.e.*, at $p_k = \mathcal{L}\eta_k$ or equivalently $\widehat{s}_k = 1/\mathcal{L}$, with

$$S_{\max} = \frac{1}{a_k(2-a_k)} \;\approx\; \frac{\lambda_{\min,k}}{2\mathcal{L}\eta_k} \qquad \text{for } a_k \ll 1. \tag{128}$$

Because (13) enforces $\varrho_k > \bar{s}_k/s^\star \geq a_k$, every admissible choice $p_k < \lambda_{\min,k}$ lies inside the window (127) and therefore *strictly* improves the contraction factor of the uncompressed FedNPG, while $p_k = \lambda_{\min,k}$ reproduces it exactly. The attainable acceleration (128) scales as $1/a_k$. In other words, the more conservative the step size of the standard FedNPG relative to the smoothness limit $1/\mathcal{L}$, the more the ED parametrization can recover.

(c) Invariance of the asymptotic accuracy and the round complexity. Let $\Delta_k := \mathcal{J}_{\mathrm{avg}}(K_k) - \mathcal{J}_{\mathrm{avg}}(K_*)$. Unrolling the recursion (14) and summing the geometric series gives

$$\Delta_k \leq \big(1-\widehat{\beta}\big)^k \Delta_0 + \widehat{\varepsilon}_L \sum_{j=0}^{k-1} \big(1-\widehat{\beta}\big)^j \leq \big(1-\widehat{\beta}\big)^k \Delta_0 + \Delta_\infty, \qquad \Delta_\infty := \frac{\widehat{\varepsilon}_L}{\widehat{\beta}}. \tag{129}$$

Since $\widehat{\beta}$ and $\widehat{\varepsilon}_L$ share the common factor $f(\widehat{s}_k)$, this factor cancels and

$$\Delta_\infty = \frac{r\,\varepsilon_{g_1}^2}{c} = \frac{r\,\|\Sigma_{K_*}\|\,\varepsilon_{g_1}^2}{4\mu_\xi^2 \lambda_{\min}(\mathscr{R})}, \tag{130}$$

which depends on *neither* η_k, p_k *nor* q: the asymptotic accuracy is dictated solely by the gradient estimation error ε_{g_1}, hence by the sample budget $\sum_i \tau_{i,k}$ through (16). Thus, although $\widehat{\varepsilon}_L \geq \bar{\varepsilon}_L$ whenever $p_k \leq \lambda_{\min,k}$, the inflated residual is exactly compensated by the enlarged contraction, and the compression parameters act purely on the transient. Requiring $\big(1-\widehat{\beta}\big)^k \Delta_0 \leq \epsilon - \Delta_\infty$ and using $\ln\big(1/(1-\widehat{\beta})\big) \geq \widehat{\beta}$, the number of communication rounds needed to reach an accuracy $\epsilon > \Delta_\infty$ satisfies

$$C(\epsilon) \leq \left\lceil \frac{1}{\widehat{\beta}} \ln \frac{\Delta_0}{\epsilon - \Delta_\infty} \right\rceil \;\propto\; \frac{1}{\widehat{\beta}}, \tag{131}$$

so the speed-up (126) translates one-to-one into a reduction of the number of rounds.

(d) The three channels through which q acts. By (122), the truncation level appears in neither $\widehat{\beta}$ nor $\widehat{\varepsilon}_L$. It acts instead as follows.

(d.1) Payload. Each agent transmits the pairs $\{\widehat{\lambda}_{b,k}^{(i)}, \widehat{v}_{b,k}^{(i)}\}_{b=1}^{q}$, the scalar $p_k^{(i)}$ and the gradient $\widehat{g}_k^{(i)}$, *i.e.*, $(d_\theta+1)(q+1)$ scalars per round, against $d_\theta(d_\theta+1)/2$ scalars for the full symmetric FIM, giving a compression ratio of $2(q+1)/d_\theta$.

(d.2) Curvature fidelity. According to Lemma 6, $\|\widehat{\mathcal{F}}_{\mathrm{ed},k} - \widehat{\mathcal{F}}_k\|_2 \leq \lambda_{q+1,k} - p_k$. Combining this with the resolvent identity $\widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1} - \widehat{\mathcal{F}}_k^{-1} = \widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1}\big(\widehat{\mathcal{F}}_k - \widehat{\mathcal{F}}_{\mathrm{ed},k}\big)\widehat{\mathcal{F}}_k^{-1}$ and with (122) yields

$$\big\|\widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1} - \widehat{\mathcal{F}}_k^{-1}\big\|_2 \leq \frac{\lambda_{q+1,k} - p_k}{p_k\,\lambda_{\min,k}}, \qquad D_k := \frac{\big\|\widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1} - \widehat{\mathcal{F}}_k^{-1}\big\|_2}{\big\|\widehat{\mathcal{F}}_k^{-1}\big\|_2} \leq \frac{\lambda_{q+1,k}}{p_k} - 1, \tag{132}$$

where we used $\|\widehat{\mathcal{F}}_k^{-1}\|_2 = 1/\lambda_{\min,k}$. The relative distortion D_k is dimensionless, non-increasing in q (as $\lambda_{q+1,k}$ decreases with q) and increasing as p_k decreases. Substituting (132) into (74) bounds the one-round deviation from the uncompressed trajectory as

$$\big\|K_{k+1} - \bar{K}_{k+1}\big\|_2 \leq \eta_k \|G_{K_k}\|\,\frac{\lambda_{q+1,k} - p_k}{p_k\,\lambda_{\min,k}}. \tag{133}$$

(d.3) Alignment of the ascent direction. For a single agent, or when the local eigenbases coincide, (9) gives $\widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1} = \sum_{b=1}^{q} \widehat{\lambda}_{b,k}^{-1} v_{b,k} v_{b,k}^\top + p_k^{-1}\big(I - P_{q,k}\big)$ with $P_{q,k} := \sum_{b=1}^{q} v_{b,k} v_{b,k}^\top$. Introducing the tail-energy fraction $\varsigma_k(q) := \|(I-P_{q,k})\widehat{g}_k\|^2/\|\widehat{g}_k\|^2 \in [0,1]$, the exact identity

$$\widehat{g}_k^\top \widehat{\mathcal{F}}_{\mathrm{ed},k}^{-1} \widehat{g}_k = \sum_{b=1}^{q} \frac{\big\langle \widehat{g}_k, v_{b,k}\big\rangle^2}{\widehat{\lambda}_{b,k}} + \varsigma_k(q)\,\frac{\|\widehat{g}_k\|^2}{p_k} \tag{134}$$

holds, showing that the $1/p_k$ amplification acts *only* on the discarded subspace, while the retained directions are scaled by their true curvature $\widehat{\lambda}_{b,k}^{-1}$. Hence q interpolates between the exact FedNPG ($q = d_\theta$, where $\varsigma_k \equiv 0$ and p_k becomes immaterial) and a p_k^{-1}-scaled federated PG method ($q = 0$, where $\widehat{\mathcal{F}}_{\mathrm{ed},k} = p_k I$ and the update reduces to $\theta_{k+1} = \theta_k + (\eta_k/p_k)\widehat{g}_k$), and the bound of Theorem 1 corresponds to the tail-dominated instance $\varsigma_k(q) \to 1$.

(e) Coupling between p_k and q, and the communication-optimal design. Imposing a distortion tolerance $D_k \le D_{\max}$ in (132) is equivalent to

$$\lambda_{q+1,k} \le (1 + D_{\max})\, p_k, \tag{135}$$

i.e., an aggressive (small) p_k must be paid for with a larger q; conversely, a per-round budget $q \le q_{\max}$ imposes the floor $p_k \ge \lambda_{q_{\max}+1,k}/(1 + D_{\max})$. By (131) and *(d.1)*, the total communication load to reach an accuracy ϵ is

$$\mathcal{C}(\epsilon) \;\approx\; N(d_\theta + 1)(q + 1)\, \frac{1}{\widehat{\beta}(p_k)}\, \ln \frac{\Delta_0}{\epsilon - \Delta_\infty}, \tag{136}$$

whose only dependence on q is the linear factor $(q + 1)$. The design problem therefore decouples: q should be taken as small as the fidelity requirement (135) allows, and p_k should then be used to maximize $\widehat{\beta}$.

In summary, q should be read off the spectral decay of the local FIM and kept as small as the distortion tolerance permits, whereas p_k should be taken as small as (13) permits, *i.e.*, slightly above $\mathcal{L}\eta_k$ and never below $\mathcal{L}\eta_k/(2 - a_k)$: the former governs what is communicated, the latter governs how fast the algorithm converges, and by (130) neither affects the asymptotic accuracy floor.