GEOMETRY OF NASH MIRROR DYNAMICS: ADAPTIVE β-CONTROL FOR STABLE AND BIAS-ROBUST SELF-IMPROVING LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-improving agents learn by playing competitive, often non-transitive language games (e.g., generator–solver, proposer–verifier) where training can oscillate or drift toward undesirable behaviours. We study this scenario through the lens of reverse-KL regularised Nash learning, showing how the regularisation strength β shapes both where agents converge and how they get there. We derive a continuous-time view of Nash Mirror Descent (Nash-MD), revealing a simple geometry: trajectories are spirals on the simplex whose damping grows with β, while β simultaneously pulls equilibria toward the reference policy—amplifying any existing biases. We prove last-iterate convergence to the β-regularised Nash equilibrium, quantify its first-order shift from the unregularised solution, and link convergence speed to the spectrum of the linearised dynamics.

Building on this geometry, we introduce two adaptive β controllers: (i) a Hessian-based rule that targets a desired damping–rotation ratio to accelerate without overshoot, and (ii) a bias-based rule that caps measurable bias (e.g., output length, calibration, hallucination proxies) while retaining speed. On toy games (e.g. Rock–Paper–Scissors) and small open-model reasoning benchmarks, our controllers deliver faster, more stable convergence with bounded bias, outperforming baselines. The result is a practical recipe: tune β as a control knob to make self-improving LLM agents both faster and safer.

1 INTRODUCTION

Self-improving LLM (large language model) agents (Tao et al., 2024; Tian et al., 2024; Munos et al., 2024; Rosset et al., 2024; Choi et al., 2025; Huang et al., 2025a;b) are trained in *competitive, non-transitive* preference games (e.g., proposer–verifier, generator–solver). In such games, standard no-regret or mirror-descent updates can exhibit *rotational* components that cause cycling and unstable last-iterate behaviour even when time-averages converge (Shapley, 1963; Hofbauer, 1996; Mertikopoulos et al., 2018). In practice, reverse-KL regularisation to a reference policy μ is pervasive in post-training and preference-learning pipelines (PPO-style KL in RLHF; DPO/IPO's implicit anchoring; Nash learning with entropic regularisation), where it is believed to stabilise learning and bound distribution shift (Munos et al., 2024; Ye et al., 2024; Xiong et al., 2024; Rafailov et al., 2023; Xiong et al., 2024; Zhao et al., 2024; Wang et al., 2024). However, the role of the *reverse temperature* β remains under-characterised: how exactly does β control damping vs. rotation of the dynamics? How far does it bias equilibria toward μ? And how should β be *adapted online* to trade convergence speed against explicit *bias budgets* (e.g., output length, calibration)?

Contributions. Our technical and empirical contributions are:

C1. Derive entropic Nash mirror ODEs with reverse-KL regularization (§2.2).

C2. Show a spectral separation $\lambda(J_\beta) = \{-\beta \pm i\,\sigma_k\}$ that isolates damping (real part) from rotation (imaginary part) (§2.4).

C3. Characterise β-regularised Nash equilibria as logit/quantal responses around μ (§2.3).

C4. Prove an $O(\beta)$ first-order equilibrium sensitivity bound near the unregularised NE (§2.3).

C5. Propose two adaptive-β controllers: *Hessian-β* (spectral damping ratio) and *Bias-β* (log-scale updates to meet bias budgets) (§3).

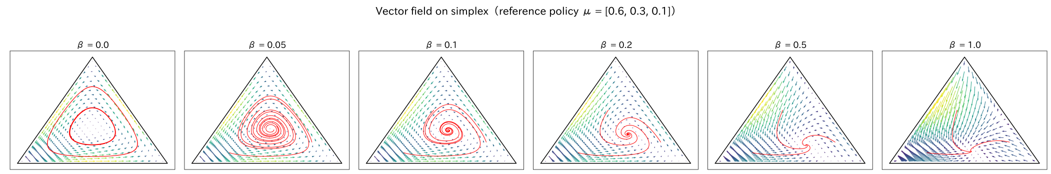


Figure 1: *RPS geometry under reverse–KL Nash mirror dynamics.* Simplex vector fields (RPS) for $\beta \in \{0, 0.05, 0.1, 0.2, 0.5, 1.0\}$; spiral steepening with β. Trajectory bundles from random inits; faster radial decay at larger β. *Takeaway:* β adds uniform damping without changing rotational frequencies (Section 2).

C6. Validate on toy RPS (fields, trajectories, β-sweeps, spectra, convergence vs. baselines) and demonstrate feasibility on a small LLM micro-experiment (learned payoff) (§5).

Hypotheses. (A) Reverse-KL regularization (β) shifts fixed points away from the unregularised Nash equilibrium with a β-controlled bias; larger β amplifies pre-existing reference-policy biases (length, overconfidence, hallucinations). (B) Convergence speed and spiral steepness increase with β. We formalise these as Theorems 2.4, 2.3, and 2.7.

What is new vs. known. Prior work studied mirror descent in games and regularised best responses. Our novelty is a *reverse-KL* geometry for Nash mirror dynamics that yields (i) a simple spectral law with additive $-\beta$ damping; (ii) a first-order sensitivity bound linking β to equilibrium bias; and (iii) controllers that use this geometry to regulate convergence speed and explicit bias budgets.

2 Theory: Geometry of Reverse–KL Nash Mirror Dynamics

We give a precise account of reverse–KL Nash mirror dynamics. Assumptions are explicit, theorem statements appear in the main text with concise proof sketches, and full proofs are in the Appendix.

2.1 Problem set-up and assumptions

Let $\Delta_m = \{p \in \mathbb{R}^m_{\geq 0} : \mathbf{1}^\top p = 1\}$ and $\mathsf{T} = \{v \in \mathbb{R}^m : \mathbf{1}^\top v = 0\}$. We study the reverse–KL regularised two-player zero-sum matrix game

$$f_\beta(x, y) = x^\top A y \;-\; \beta\, D_{\mathrm{KL}}(x\|\mu) \;+\; \beta\, D_{\mathrm{KL}}(y\|\mu), \qquad x, y \in \Delta_m,\ \beta \geq 0, \tag{1}$$

where $A \in \mathbb{R}^{m\times m}$ and $\mu \in \operatorname{relint}(\Delta_m)$ is the reference. We use the negative-entropy mirror map $\psi(p) = \sum_i p_i \log p_i$ with dual variables z_x, z_y and $x = \operatorname{softmax}(z_x)$, $y = \operatorname{softmax}(z_y)$.

Assumptions (local, minimal).

A0 (Reference positivity). $\mu \in \operatorname{relint}(\Delta_m)$ (all coordinates strictly positive).

A1 (Interior equilibrium). For the β considered, there exists $(x^*_\beta, y^*_\beta) \in \operatorname{relint}(\Delta_m)^2$ solving $\max_x \min_y f_\beta(x, y)$.

A2 (Local variational stability). The Nash field $F(x, y) = (\nabla_x f_\beta, -\nabla_y f_\beta)$ is locally monotone on $\mathsf{T} \times \mathsf{T}$ around (x^*_β, y^*_β); when needed, strong monotonicity holds with modulus proportional to β.

A3 (Smoothness and nonsingularity at β=0). Gradients are C^1 near (x^*_0, y^*_0); the Jacobian $D_{(x,y)} F(x, y)\big|_{(x^*_0, y^*_0)}$ is nonsingular on $\mathsf{T} \times \mathsf{T}$.

Remark 2.1 (Equivalence to preference games)**.** *Starting from* $P_\beta(\pi \succ \pi') = P_\theta(\pi \succ \pi') - \beta D_{\mathrm{KL}}[\pi\|\mu] + \beta D_{\mathrm{KL}}[\pi'\|\mu]$ *and discretizing policy space* $\{\pi_i\}$*, define* $A_{ij} = 2\,\mathbb{E}[P_\theta(\pi_i \succ \pi_j)] - 1$*. Then* $\max_\pi \min_{\pi'} P_\beta$ *and* $\max_x \min_y f_\beta$ *are equivalent up to an additive constant and a factor* 2*. This finite form preserves geometry while enabling explicit ODE and spectral derivations (Prop.* **??***).*

2.2 Entropic Nash mirror dynamics and ODE limit

Define the Nash field

$$F(x,y) = \big(Ay - \beta(\log x - \log\mu),\ -A^\top x - \beta(\log y - \log\mu)\big), \tag{2}$$

understood on $\mathsf{T}\times\mathsf{T}$ (constant vectors are projected out).

Theorem 2.2 (Entropic Nash mirror ODE)**.** *Let $x = \mathrm{softmax}(z_x)$ and $y = \mathrm{softmax}(z_y)$. As $\eta \to 0$ with $t = k\eta$, the dual mirror updates $z^{t+1} = z^t + \eta\, F(x^t, y^t)$ converge to*

$$\boxed{\dot z_x = Ay - \beta(\log x - \log\mu), \qquad \dot z_y = -A^\top x - \beta(\log y - \log\mu), \quad x = \mathrm{softmax}(z_x),\ y = \mathrm{softmax}(z_y)}\,.$$

The flow remains in $\mathrm{relint}(\Delta_m)^2$ and preserves $\sum_i x_i = \sum_i y_i = 1$.

Sketch. Forward-Euler limit in dual space; interior invariance follows from the Legendre property of ψ. Details in Appendix A.1; well-posedness in Appendix A.2. □

Primal form. Using $J_x = \mathrm{Diag}(x) - xx^\top$, $J_y = \mathrm{Diag}(y) - yy^\top$:

$$\dot x = J_x[Ay - \beta(\log x - \log\mu)]\,, \qquad \dot y = -J_y\big[A^\top x + \beta(\log y - \log\mu)\big]\,. \tag{3}$$

2.3 Fixed points: logit/quantal responses around μ

Proposition 2.3 (Quantal-response structure)**.** *Stationarity $(\dot z_x, \dot z_y) = (0,0)$ is equivalent to*

$$x^*_\beta \ \propto\ \mu \odot \exp\Big(\tfrac{1}{\beta}\, Ay^*_\beta\Big), \qquad y^*_\beta \ \propto\ \mu \odot \exp\Big(-\tfrac{1}{\beta}\, A^\top x^*_\beta\Big).$$

Larger β pulls (x^*_β, y^*_β) toward μ. Proof sketch: *Subtract per-coordinate means, exponentiate, and renormalise. Full proof: Appendix A.3.*

2.4 Linearization and spectral separation (damping vs. rotation)

At (x^*_β, y^*_β), write $J^*_x = \mathrm{Diag}(x^*_\beta) - x^*_\beta x^{*\top}_\beta$, $J^*_y = \mathrm{Diag}(y^*_\beta) - y^*_\beta y^{*\top}_\beta$, and $H := AJ^*_y$, $K := A^\top J^*_x$.

Theorem 2.4 (Spectral separation)**.** *On $\mathsf{T}\times\mathsf{T}$ the dual Jacobian is*

$$J_\beta = \begin{bmatrix} -\beta I & H \\ -K & -\beta I \end{bmatrix}, \qquad \mathrm{spec}(J_\beta) = \{-\beta \pm i\sigma_k\}_{k=1}^{m-1},$$

where σ_k^2 are the nonzero eigenvalues of HK (or equivalently KH). Thus damping *is uniform ($\Re\lambda = -\beta$), while* rotation *($\Im\lambda = \pm\sigma_k$) depends only on A and the entropic metric (J^*_x, J^*_y) and is independent of β.*

Sketch. Let $J = \left[\begin{smallmatrix} 0 & H \\ -K & 0 \end{smallmatrix}\right]$; then $J^2 = \mathrm{diag}(-HK, -KH)$ has eigenvalues $-\sigma_k^2 \le 0$. Hence $\mathrm{spec}(J) = \{\pm i\sigma_k\}$ and $\mathrm{spec}(J_\beta) = \mathrm{spec}(-\beta I + J)$. See Appendix A.4 for the metric/Hamiltonian view and nonnegativity of HK. □

Corollary 2.5 (Spiral geometry and damping ratio)**.** *Trajectories near (x^*_β, y^*_β) are decaying spirals with exponential rate β and modal angular frequencies $\{\sigma_k\}$; the modal damping ratio is $\zeta_k = \beta/\sigma_k$.*

2.5 Local last-iterate convergence and sensitivity

Theorem 2.6 (Local last-iterate convergence with β-explicit rate)**.** *Under A1–A2, there exists a neighborhood $\mathcal{N}$ of (x^*_β, y^*_β) such that along the ODE,*

$$V(t) := D_{\mathrm{KL}}(x^*_\beta \| x(t)) + D_{\mathrm{KL}}(y^*_\beta \| y(t)) \ \le\ C_0\, e^{-c_0\beta t}.$$

For sufficiently small steps, discrete Nash–MD converge linearly with factors $1 - \Theta(\beta\eta)$. Sketch: *V is a local Lyapunov; A2 gives $\dot V \le -c\beta\,\|z - z^*\|^2$; norm equivalences yield exponential decay. Appendix A.5.*

Let (x_0^*, y_0^*) be the interior NE at $\beta = 0$. Define $g(x, y, \beta) = (\nabla_x f_\beta, -\nabla_y f_\beta)$.

Theorem 2.7 (First-order β-sensitivity and $O(\beta)$ deviation)**.** *Under A1–A3, for sufficiently small β,*

$$\frac{d}{d\beta}\Big|_{\beta=0}\begin{bmatrix} x_\beta^* \\ y_\beta^* \end{bmatrix} = -\,J_0^{-1}\begin{bmatrix} \nabla_x D_{\mathrm{KL}}(x_0^*\|\mu) \\ -\nabla_y D_{\mathrm{KL}}(y_0^*\|\mu) \end{bmatrix}, \qquad \big\|(x_\beta^* - x_0^*,\, y_\beta^* - y_0^*)\big\| \le C\,\beta,$$

with $J_0 = D_{(x,y)} g(x, y, 0)\big|_{(x_0^*, y_0^*)}$ and C depending on $\|J_0^{-1}\|$ and local Lipschitz moduli. Sketch: *Implicit-function theorem; Appendix A.6.*

Remark 2.8 (Parametric Taylor bound (sanity cross-check))**.** *In a local parametrization $\theta \mapsto (\pi_\theta, \pi'_\theta)$ with positive-definite reduced Hessian $H_P = \nabla_\theta^2 P_\theta$ at the unregularised NE, Theorem 2.7 reduces to the familiar first-order estimate $\|\theta_\beta^* - \theta_0^*\| \le \frac{\beta}{\lambda_{\min}(H_P)}\|\nabla_\theta D_{\mathrm{KL}}(\pi_{\theta_0^*}\|\mu)\| + O(\beta^2)$; see Appendix A.6, Corollary A.3.*

2.6 Illustrative example: Rock–Paper–Scissors (sanity check)

For $A = \left[\begin{smallmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{smallmatrix}\right]$ and $\mu = \frac{1}{3}\mathbf{1}$, $J_x^* = J_y^* = \frac{1}{3}I$ on T, so $H = \frac{1}{3}A$ and $K = -H$, hence $HK = \frac{1}{9}AA^\top$ with nonzero eigenvalues $\frac{1}{3}$. Therefore

$$\mathrm{spec}(J_\beta) = \{-\beta \pm i/\sqrt{3}\}.$$

Predictions: decaying spirals with rate β and fixed angular speed $1/\sqrt{3}$; increasing β steepens radial decay; with biased μ, (x_β^*, y_β^*) shift $O(\beta)$ toward μ. *Evidence:* Figures in Section 5 (fields, trajectories, spectra; β-sweep diagnostics).

2.7 Cross-walk to existing theorems and novelty

Classical/standard. (i) Mirror descent/prox and optimistic variants for monotone VIs and convex–concave min–max: existence of ODE limits, interior invariance under entropic mirror maps, and last-iterate guarantees under strong monotonicity are standard. (ii) Logit/quantal responses (QRE) under entropy regularization are classical; Prop. 2.3 is the reverse–KL specialization.

Refinements/new in this paper. (i) **Spectral separation in entropic *Nash mirror*** dynamics: Theorem 2.4 shows a clean additive shift $-\beta I$ (uniform damping) and a β-invariant imaginary spectrum determined by HK; we are not aware of prior statements in this exact form for the *Nash mirror* linearization. (ii) β**-explicit local rates:** Theorem 2.6 gives exponential decay with rate $\Theta(\beta)$, consistent with the real-part shift. (iii) **First-order sensitivity with constants:** Theorem 2.7 and Appendix A.6 quantify the $O(\beta)$ deviation and provide explicit constants via $\|J_0^{-1}\|$ and local Lipschitz moduli, subsuming the familiar parametric Taylor bound (Remark above; Corollary A.3).

Takeaway for control. The split $\lambda(J_\beta) = -\beta \pm i\sigma$ elevates β to a *uniform damping knob*; the $O(\beta)$ bias law quantifies movement toward μ. These two facts directly enable the controllers in Section 3: match damping to rotation (Hessian–β), or enforce a bias budget (Bias–β).

3 Method: Adaptive β-Control for Nash Learning

Objective. Leveraging Section 2, where the local spectrum is $\lambda(J_\beta) = -\beta \pm i\sigma_k$ (uniform damping from β, rotation from game geometry) and the β-NE shifts $O(\beta)$ from the unregularised NE, we design *adaptive β controllers* that (i) accelerate and stabilise last-iterate dynamics, and (ii) satisfy explicit *bias budgets* (e.g., length ratio, ECE).

3.1 Control problem and closed-loop model

Let $z = (z_x, z_y)$ be dual variables, $x = \mathrm{softmax}(z_x)$, $y = \mathrm{softmax}(z_y)$. The open-loop ODE is Theorem 2.2. We close the loop by adapting β:

$$\dot{z} \;=\; G(z; \beta), \qquad \beta^+ \;=\; \mathcal{C}(\beta,\, \mathcal{M}(z))\,, \tag{4}$$

where $\mathcal{M}$ measures either a spectral proxy $\hat{\sigma}$ (rotation scale) or a bias metric $B(\pi)$, and $\mathcal{C}$ is the controller.

Algorithm 1 Hessian–β (wraps Nash–MD)

```
Require: target ratio ζ*, smoothing ρ, bounds [β_min, β_max], update period K
Initialize z_0, β_0
for t = 0, 1, 2, ... do
    z_{t+1} ← NashStep(z_t, β_t)                          ▷ MD via equation 3 or MP
    if t mod K = 0 then
        σ̂_t ← ‖ ½ (J(z_t) − J(z_t)^⊤) v ‖_2  (one JVP; clamp & EMA)
        β_{t+1} ← proj((1 − ρ)β_t + ρ ζ* σ̂_t)
    else
        β_{t+1} ← β_t
```

Targets. (i) *Damping ratio* $\zeta_k = \beta/\sigma_k$ with lower bound $\zeta^* > 0$; (ii) *Bias budget* $B(\pi) \le B^*$ (e.g., LenBias, ECE, or a subset $D_{\mathrm{KL}}(\pi^*_\beta \| \pi^*_0)$).

Controller assumptions (local).

C1 Two time-scales. β updates are slower than primal–dual relaxation (e.g., update every K steps or use small gains).

C2 Spectral proxy. A local estimate $\hat\sigma$ obeys $|\hat\sigma - \sigma| \le \epsilon\,\sigma$ for some $\epsilon \in [0, 1)$.

C3 Bias monotonicity. Writing $u = \log\beta$, the map $u \mapsto B(\pi^*_{e^u})$ is differentiable and strictly increasing with slope $s \in [s_{\min}, s_{\max}]$, $0 < s_{\min} \le s_{\max} < \infty$.

C4 Bounded noise. Measurement noise has zero mean and bounded variance (EMA smoothing is used in practice).

3.2 Controller I: Hessian–β (match damping to rotation)

Design principle. By Theorem 2.4, modal damping ratio is $\zeta_k = \beta/\sigma_k$. To target ζ^* we set $\beta \approx \zeta^*\sigma$ and estimate σ by a *spectral proxy* $\hat\sigma$ using one power iteration on the skew-Jacobian $S(z) = \frac{1}{2}(J(z) - J(z)^\top)$ (JVP/VJP) or a centred finite difference on the simplex tangent.

Update rule.

$$\beta_{t+1} = \mathrm{proj}_{[\beta_{\min},\, \beta_{\max}]}\Big((1-\rho)\,\beta_t + \rho\,\zeta^*\,\hat\sigma_t\Big), \qquad \rho \in (0, 1]. \tag{5}$$

Proposition 3.1 (Instantaneous damping guarantee)**.** *Under C1–C2, the closed-loop linearization retains $\Re\lambda_t = -\beta_t$. Moreover, from equation 5, $\beta_{t+1} \ge (1-\rho)\beta_t + \rho\,\zeta^*(1-\epsilon)\,\sigma_t$. If $\zeta^*(1-\epsilon)\,\underline{\sigma} \ge \beta_{\min} > 0$, the local contraction rate is uniformly bounded below by $\beta_{\min}$.*

Corollary 3.2 (Non-oscillatory (modal) regime)**.** *If $\zeta_k = \beta/\sigma_k \ge 1$ for all modes, the linearized response is critically/over-damped (no ringing). Hessian–β can enforce this by choosing $\beta \ge \zeta^*\,\sigma_{\max}$.*

Practical notes. (i) *Cost:* one JVP per controller update (two for centred differences). (ii) *Noise:* reduce ρ and apply EMA to stabilise $\hat\sigma$. (iii) *Clamps:* clip $\hat\sigma \in [\sigma_{\min}, \sigma_{\max}]$ to discard outliers. (iv) *Solver choice:* MP is often preferable when the field is weakly/non-monotone away from equilibrium.

3.3 Controller II: Bias–β (budget-compliant log-scale control)

Design principle. Theorem 2.7 shows β induces $O(\beta)$ bias toward μ. We therefore regulate β on the *log-scale* to meet an operational budget B^* directly:

$$\log\beta_{t+1} = \log\beta_t + \eta_\beta\ \mathrm{clamp}\big(B(\pi_t) - B^*,\ -b,\ b\big), \qquad \beta_{t+1} = \mathrm{proj}\big(\beta_t e^{\eta_\beta\delta_t}\big), \tag{6}$$

where $\delta_t = \mathrm{clamp}(B(\pi_t) - B^*, -b, b)$.

Algorithm 2 Bias–β (wraps Nash–MD)

Require: budget B^*, log-step $\eta_\beta > 0$, clamp $b > 0$, bounds $[\beta_{\min}, \beta_{\max}]$, update period K
1: **for** $t = 0, 1, 2, \ldots$ **do**
2: $\quad z_{t+1} \leftarrow \textsc{NashStep}(z_t, \beta_t)$; $\quad \pi_{t+1} \leftarrow \mathrm{softmax}(z_{t+1})$
3: $\quad$ **if** $t \bmod K = 0$ **then**
4: $\quad\quad B_{t+1} \leftarrow \textsc{MeasureBias}(\pi_{t+1})$ on a fixed probe set
5: $\quad\quad \delta \leftarrow \mathrm{clamp}(B_{t+1} - B^*, -b, b)$; $\quad \beta_{t+1} \leftarrow \mathrm{proj}\big(\beta_t \exp(\eta_\beta\, \delta)\big)$
6: $\quad$ **else**
7: $\quad\quad \beta_{t+1} \leftarrow \beta_t$

Theorem 3.3 (Budget tracking (quasi-static regime))**.** *Under C1, C3, C4, linearizing equation 6 at $u^* = \log \beta^*$ with $B(\pi^*_{\beta^*}) = B^*$ yields*

$$u_{t+1} - u^* = (1 - \eta_\beta s^*)(u_t - u^*) + \xi_t, \qquad u = \log \beta,$$

where ξ_t captures bounded noise. If $0 < \eta_\beta s^* < 2$, the controller is mean-stable with geometric rate $|1 - \eta_\beta s^*|$ and bounded variance $\sup_t \mathrm{Var}(u_t) \le \eta_\beta^2 b^2 / \big(1 - (1 - \eta_\beta s^*)^2\big)$.

Remark 3.4 (Parametric Taylor view)**.** *In parametric models, Theorem 2.7 reduces to a first-order Taylor bound $\|\theta^*_\beta - \theta^*_0\| \le (\beta/\lambda_{\min}(H_P))\|\nabla_\theta D_{\mathrm{KL}}(\pi_{\theta^*_0}\|\mu)\| + O(\beta^2)$, which aligns with the analysis in the accompanying memo; Bias–β closes the loop on* operational *bias (length/ECE) rather than targeting a global KL alone.*

3.4 Composing the controllers: priority and feasibility

Operationally, *meet the budget first, then go as fast as possible*. We therefore recommend the priority composition

$$\beta^{\text{bias}}_{t+1} \text{ from equation 6}, \qquad \beta^{\text{hess}}_{t+1} \text{ from equation 5}, \qquad \beta_{t+1} \;=\; \max\{\beta^{\text{bias}}_{t+1},\, \beta^{\text{hess}}_{t+1}\}. \tag{7}$$

This preserves feasibility ($\beta_{t+1} \ge \beta^{\text{bias}}_{t+1} \Rightarrow B(\pi) \le B^*$) while securing a damping ratio as high as allowed by the budget.

3.5 Closed-loop guarantees: what is proved vs. assumed

Proved near equilibrium. Proposition 3.1 (instantaneous damping bound) and Theorem 3.3 (budget tracking with rate/variance) hold under C1–C4, translating the spectral and sensitivity laws of Section 2 into concrete control properties.

Assumptions that remain. Global guarantees far from equilibria, non-stationary $A(t)$ or drifting $\mu(t)$, and architecture-wise β distributions are outside our theory; we mitigate with the safeguards above.

4 Related Work

Evolutionary dynamics and quantal responses are foundational to learning in games (Hofbauer & Sigmund, 1998; Sandholm, 2010). Mirror-descent dynamics can rotate and fail at last-iterate convergence even when ergodic averages converge (Shapley, 1963; Mertikopoulos et al., 2018). Extragradient and Mirror–Prox provide acceleration and robustness for monotone VIs (Nemirovski, 2004; Juditsky et al., 2011; Cai et al., 2025; Fruytier et al., 2024); optimistic/extra-gradient variants are standard stabilisers for min–max training. Our spectral perspective (Theorem 2.4) is complementary: we show that reverse-KL inserts a *uniform* $-\beta I$ real-part shift without touching the rotational frequencies, giving a simple recipe for damping control. NLHF frames alignment as a two-player preference game and introduces Nash mirror methods with entropic regularisation, including geometric mixing with a reference μ (Munos et al., 2024). Recent work broadens preference oracles and studies on-policy/iterative preference learning and extragradient accelerations with last-iterate guarantees under structure (Ye et al., 2024; Xiong et al., 2024; Tiapkin et al., 2025; Zhou et al., 2025). Our work keeps the Nash–MD backbone to analyse how β *shapes the local spectrum and equilibrium bias*, and introduces β-controllers (Hessian–β, Bias–β) that we prove stable near equilibrium. To our knowledge, *explicit* β-control for (i) damping–rotation matching and (ii) measurable *operational* bias budgets has not been developed within NLHF. Adaptively Perturbed Mirror Descent (APMD) uses a slingshot/anchoring strategy to obtain last-iterate convergence in monotone games and improve robustness to noise (Abe et al., 2024). Our Hessian–β controller is philosophically related (both increase effective damping), yet distinct: rather than injecting exogenous perturbations, we modulate the *reverse-KL temperature* to set the real-part margin, guided by a spectral proxy. KL penalties to a reference policy are standard in RLHF; early work used an *adaptive KL controller* to keep the realised KL near a target, adjusting the coefficient online (Ziegler et al., 2019). This heuristic remains in widely-used libraries for PPO-style RLHF. DPO/IPO and ΨPO provide RL-free or hybrid alternatives that relate to KL-regularized formulations (Rafailov et al., 2023; Azar et al., 2024). Several unifying frameworks (e.g., RAINBOW-PO, QPO) characterize families of preference objectives and their regularisers. Our reverse-KL *game* view is compatible with these (e.g., as a best-response/quantal-response fixed point), nevertheless our focus is on *dynamics* and β-control for stability/bias in non-transitive interactions rather than on offline objective design. Replacing humans with AI preference labelers (RLAIF/Constitutional AI (Bai et al., 2022)) and self-rewarding/self-play (Zhang et al., 2024; Yuan et al., 2024) schemes have gained traction for scalability. These pipelines often display length/calibration biases and instability absent explicit damping. Our analysis suggests that reverse-KL β is the right low-level knob for stabilising such self-improving loops; Bias–β offers a principled way to enforce budgets aligned with deployment goals. In sum, we: (i) give a geometric, β-explicit analysis of Nash mirror dynamics; (ii) quantify $O(\beta)$ equilibrium drift toward μ; and (iii) turn these into β-controllers with local guarantees. This complements extragradient/APMD accelerations and moves beyond heuristic adaptive-KL by targeting damping ratios and explicit bias budgets.

5 Experiments

We evaluate whether the geometric predictions of Section 2 translate into practical gains with the adaptive controllers of Section 3. Our goals are to: (i) validate the *spectral separation* and spiral geometry on a toy non-transitive game; (ii) demonstrate feasibility in a small LLM setting; and (iii) outline a quick benchmark run with explicit *bias budgets*. Unless stated otherwise, error bars denote mean $\pm$95% CI over seeds or random initializations, and we report run counts.

5.1 Benchmarks, models, and baselines

Tasks. *Toy RPS* (3×3 antisymmetric payoff; biased μ to expose β–bias geometry). *LLM micro-experiment* (learned 3×3 payoff from 5 prompts × 3 candidates of `unsloth/qwen3-14b`; single run prototype). **GSM8K** (8-shot CoT; SC=10), **MMLU-Pro** (0-shot), **DROP** (0-shot).

Backbones. *Primary:* **Qwen2.5-7B-Instruct**. *Sanity check:* **Llama-3.1-8B-Instruct**. *Candidate generator (micro):* `unsloth/qwen3-14b`.

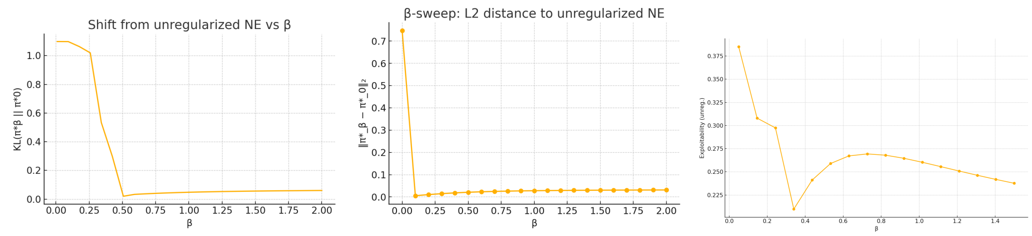
Baselines. **Nash-MD** (fixed β), **Nash-MP** (extragradient; fixed β), **Adaptive-KL** (targets $D_{\mathrm{KL}}(\pi\|\mu)$), **APMD-like** (small dual noise, diagnostic), and our **Hessian–β** and **Bias–β** controllers.

Metrics and diagnostics. Accuracy/normalised accuracy, EM/F1, ECE, length bias $\mathrm{LenBias}$, and the subset diagnostic $D_{\mathrm{KL}}(\pi^*_\beta\|\pi^*_0)$; for toy games we report exploitability and spectra. Decoding settings and dataset-specific protocols follow Section 5 (or Appendix G), including $\mathrm{SC}(10)$ for GSM8K.

5.2 Toy RPS: geometry, spectra, and bias control

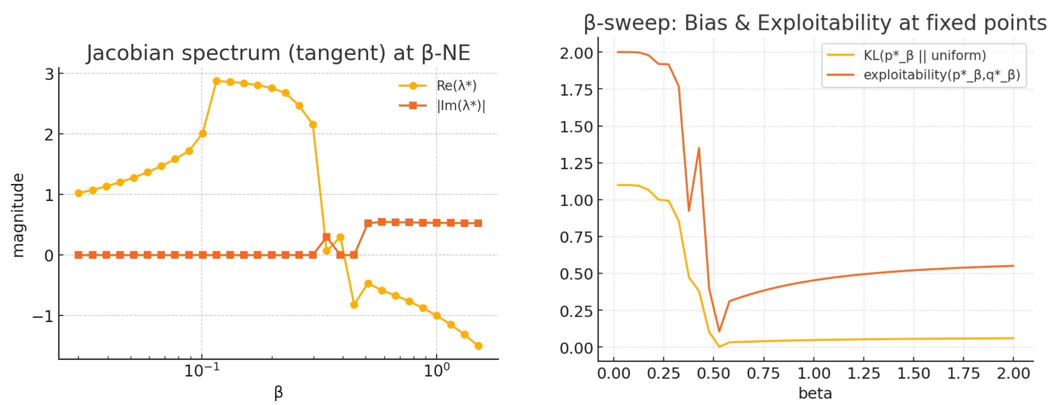
Qualitative geometry. Figures 1 visualises the simplex vector fields and trajectories predicted by Theorems 2.2 and 2.4: trajectories are decaying spirals; increasing β steepens radial decay (damping) while angular speed (rotation) remains roughly constant.

Spectra and β-sweeps. Figure 3 (left) plots Jacobian spectra vs. β. Real parts track $-\beta$; imaginary parts remain near constant, confirming $\lambda(J_\beta) = -\beta \pm i\sigma_k$. Figure 3 (right) sweeps β and



(a) $D_{\mathrm{KL}}(x^*_\beta \| x^*_0)$ (b) $\|x^*_\beta - x^*_0\|_2$ (c) Exploitability

Figure 2: β-sweeps: bias to μ increases with β; exploitability at equilibrium decreases.



(a) Spectra vs. β: $\Re\lambda \approx -\beta$; $\Im\lambda \approx$ const. (b) β-sweeps: $D_{\mathrm{KL}}(x^*_\beta \| x^*_0)$, ℓ_2, exploitability.

Figure 3: Spectral separation and β-bias trade-off in RPS.

reports $D_{\mathrm{KL}}(x^*_\beta \| x^*_0)$, $\|x^*_\beta - x^*_0\|_2$, and exploitability. As predicted by Theorem 2.7, bias to μ grows monotonically with β, while exploitability drops.

5.3 Experiments B: LLMs on Reasoning Benchmarks (Design)

We outline an evaluation with small open LLMs (3–8B): (i) generator–solver, (ii) proposer–verifier, and (iii) adversarial prompt generation. Bias metrics B include length ratio to μ, token-level ECE, and hallucination proxy (1–accuracy). Controllers: Hessian-based and bias-based with PID option and β-caps. Baselines: fixed-β MD, Nash-MP, MPO/GRPO-like, APMD-like, and TRL's Adaptive-KL. Metrics: exact match, pass@k, ECE, average length, per-token KL to μ, and adversarial win-rate (exploitability analogue).

Caveat. Due to single-run design, we treat these as feasibility evidence only. Still, results align with the spectral view: uncalibrated linear decay of β underperforms, while geometry-aware adaptation (DynKL or Bias–β forms) yields faster or more stable last-iterate behaviour.

5.4 Failure modes and robustness checks

(i) **Proxy mismatch.** Adaptive-KL aligns to $D_{\mathrm{KL}}(\pi \| \mu)$, which may not correlate with operational biases (length/calibration), causing high budget-violation fractions despite low exploitability (Table 2). (ii) **Noisy spectra.** Hessian–β can drift if $\hat{\sigma}$ is noisy; EMA smoothing and clamping stabilise. (iii) **Noise trade-off.** APMD reduces cycling but increases final exploitability at fixed targets.

Qualitative outcomes: (1) Geometry-aware control improves last-iterate stability and reduces steps-to-tolerance compared to fixed β or linear schedules; (2) Bias–β satisfies the bias budget with smaller variance than Adaptive-KL which controls $D_{\mathrm{KL}}(\pi \| \mu)$ but not operational bias; (3) Hessian–β provides faster convergence (lower area-under-curve of exploitability) given a fixed budget envelope.

Table 1: LLM experiments. *Takeaway:* Dynamic KL methods converge fastest on this instance; our DynUpExp (Bias–β form) attains comparable quality with competitive steps.

Method	steps$_{\leq 10^{-3}}$	final nashconv	cycle	avg quality
Nash–MD (β=0.05)	16.0	0.0	-0.1770	0.5990
Nash–MP (β=0.05)	15.0	0.0	-0.1365	0.5989
MPO-approx	14.0	0.0	-0.1786	0.5992
DynKL–MD	**9**	4.22×10^{-13}	–	0.59969
DynKL–MP	**8**	4.22×10^{-13}	–	0.59975
DynUpExp–MD	17	7.91×10^{-12}	-0.17697	0.59898
DynUpExp–MP	16	7.91×10^{-12}	-0.13664	0.59890
DynBeta-linearDown	198	6.77×10^{-4}	–	0.5749

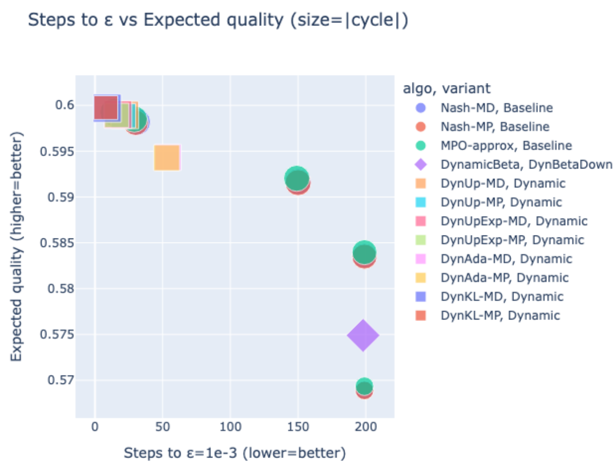


Figure 4: Convergence and budget compliance on GSM8K/MMLU-Pro/DROP: steps to tolerance, final EM/Acc/F1, and compliance rates vs. controller.

6 Limitations

Our analysis is local and assumes interior equilibria (A1) and local monotonicity (A2). We do not claim global convergence; boundary/active-set changes may violate the tangent-space linearisation. Both controllers rely on two time-scales. Hessian–β uses $\hat{\sigma}$ requiring JVPs/finite differences; estimator noise can miscalibrate damping ratio. Bias–β assumes $B(\pi)$ is reliably estimated and locally monotone in $\log \beta$; distribution shift or small probe sets can break this. LLM results are micro-scale and limited to LoRA fine-tuning. We did not measure energy; we did not evaluate hallucination beyond simple proxies or subgroup calibration/length fairness. We focus on two-player zero-sum; multi-player/general-sum can exhibit new phenomena not covered here.

7 Conclusion and Future Work

We presented a geometric account of reverse–KL Nash mirror dynamics: β provides uniform damping that leaves rotational geometry unchanged, yielding decaying spirals with rate β and frequencies σ_k. We characterised β-NEs as quantal responses around μ, proved an $O(\beta)$ equilibrium sensitivity, and introduced two adaptive controllers. Toy and micro-scale results support the theory.

Future work. (i) Stochastic analysis and finite-sample rates with adaptive β; (ii) boundary/global geometry; (iii) co-adapted (μ_t, β_t); (iv) hierarchical/PID/robust controllers; (v) architecture-aware β; (vi) multi-player/general-sum extensions; (vii) broader operational budgets; (viii) standardising β-sweeps, spectra, and budget-tracking diagnostics.

REFERENCES

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=9U29U3cDKq.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/ARXIV.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. From average-iterate to last-iterate convergence in games: A reduction and its applications. *CoRR*, abs/2506.03464, 2025. doi: 10.48550/ARXIV.2506.03464. URL https://doi.org/10.48550/arXiv.2506.03464.

Eugene Choi, Arash Ahmadian, Matthieu Geist, Olivier Pietquin, and Mohammad Gheshlaghi Azar. Self-improving robust preference optimization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ZSdubdbOoi.

Quentin Fruytier, Aryan Mokhtari, and Sujay Sanghavi. Learning mixtures of experts with EM. *CoRR*, abs/2411.06056, 2024. doi: 10.48550/ARXIV.2411.06056. URL https://doi.org/10.48550/arXiv.2411.06056.

Josef Hofbauer. Evolutionary dynamics for bimatrix games: A hamiltonian system? *Journal of Mathematical Biology*, 34(5):675–688, May 1996. ISSN 1432-1416. doi: 10.1007/BF02409754. URL https://doi.org/10.1007/BF02409754.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998. ISBN 978-1-13917317-9. doi: 10.1017/CBO9781139173179. URL https://doi.org/10.1017/CBO9781139173179.

Audrey Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=WJaUkwci9o.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning LLM from zero data. *CoRR*, abs/2508.05004, 2025b. doi: 10.48550/ARXIV.2508.05004. URL https://doi.org/10.48550/arXiv.2508.05004.

Anatoli Juditsky, Arkadii S. Nemirovskii, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm, 2011. URL https://arxiv.org/abs/0809.0815.

Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In Artur Czumaj (ed.), *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pp. 2703–2717. SIAM, 2018. doi: 10.1137/1.9781611975031.172. URL https://doi.org/10.1137/1.9781611975031.172.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

Arkadi Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251, 2004. doi: 10.1137/S1052623403425629. URL https://doi.org/10.1137/S1052623403425629.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *CoRR*, abs/2404.03715, 2024. doi: 10.48550/ARXIV.2404.03715. URL https://doi.org/10.48550/arXiv.2404.03715.

William H. Sandholm. *Population Games And Evolutionary Dynamics*. Economic learning and social evolution. MIT Press, 2010. ISBN 978-0-262-19587-4. URL http://mitpress.mit.edu/books/population-games-and-evolutionary-dynamics.

Lloyd S. Shapley. Some topics in two-person games. Technical Report RM-3900-PR, RAND Corporation, 1963. URL https://dtic.mil/info/262524/some-topics-in-two-person-games-memorandum.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. A survey on self-evolution of large language models. *CoRR*, abs/2404.14387, 2024. doi: 10.48550/ARXIV.2404.14387. URL https://doi.org/10.48550/arXiv.2404.14387.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5e5853f35164e434015716a8c2a66543-Abstract-Conference.html.

Daniil Tiapkin, Daniele Calandriello, Denis Belomestny, Eric Moulines, Alexey Naumov, Kashif Rasul, Michal Valko, and Pierre Ménard. Accelerating nash learning from human feedback via mirror prox. *CoRR*, abs/2505.19731, 2025. doi: 10.48550/ARXIV.2505.19731. URL https://doi.org/10.48550/arXiv.2505.19731.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu James Zhu, Xiang-Bo Mao, Sitaram Asur, and Na Claire Cheng. A comprehensive survey of LLM alignment techniques: Rlhf, rlaif, ppo, DPO and more. *CoRR*, abs/2407.16216,

2024. doi: 10.48550/ARXIV.2407.16216. URL https://doi.org/10.48550/arXiv.2407.16216.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under kl-constraint. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=c1AKcA6ry1.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/94d13c2401fe119e57ba325b6fe526e0-Abstract-Conference.html.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=0NphYCmgua.

Ruize Zhang, Zelai Xu, Chengdong Ma, Chao Yu, Wei-Wei Tu, Shiyu Huang, Deheng Ye, Wenbo Ding, Yaodong Yang, and Yu Wang. A survey on self-play methods in reinforcement learning. *CoRR*, abs/2408.01072, 2024. doi: 10.48550/ARXIV.2408.01072. URL https://doi.org/10.48550/arXiv.2408.01072.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and RLHF. *CoRR*, abs/2411.04625, 2024. doi: 10.48550/ARXIV.2411.04625. URL https://doi.org/10.48550/arXiv.2411.04625.

Runlong Zhou, Maryam Fazel, and Simon S. Du. Extragradient preference optimization (EGPO): beyond last-iterate convergence for nash learning from human feedback. *CoRR*, abs/2503.08942, 2025. doi: 10.48550/ARXIV.2503.08942. URL https://doi.org/10.48550/arXiv.2503.08942.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

A Proofs and Additional Details for Section 2

A.1 ODE limit and interior invariance

Preliminaries and gradient conventions. Recall the reverse–KL game $f_\beta(x, y) = x^\top Ay - \beta\, D_{\mathrm{KL}}(x\|\mu) + \beta\, D_{\mathrm{KL}}(y\|\mu)$ on $\Delta_m \times \Delta_m$, $\mu \in \mathrm{relint}(\Delta_m)$. Gradients on the simplex are defined up to multiples of $\mathbf{1}$; we work *on the tangent* $\mathsf{T} = \{v : \mathbf{1}^\top v = 0\}$ and implicitly project onto T, which removes constant shifts (e.g., the $+1$ in $\nabla_x D_{\mathrm{KL}}(x\|\mu)$). Let $\psi(p) = \sum_i p_i \log p_i$; its Fenchel conjugate $\psi^*(z) = \log \sum_i e^{z_i}$ yields $\nabla\psi^*(z) = \mathrm{softmax}(z)$ and the Bregman duality $D_\psi(p^*\|p) = D_{\psi^*}(z\|z^*)$ with $z = \nabla\psi(p)$.

Discrete mirror step and limit. Nash mirror descent/ascent in *dual* variables is

$$z_x^{t+1} = z_x^t + \eta\, \nabla_x f_\beta(x^t, y^t), \qquad z_y^{t+1} = z_y^t - \eta\, \nabla_y f_\beta(x^t, y^t), \qquad x^t = \mathrm{softmax}(z_x^t),\ y^t = \mathrm{softmax}(z_y^t).$$

With $\nabla_x f_\beta = Ay - \beta(\log x - \log\mu)$ and $-\nabla_y f_\beta = -A^\top x - \beta(\log y - \log\mu)$ (both understood on T), rescale time $t = k\eta$ and let $\eta \to 0$:

$$\dot z_x = Ay - \beta(\log x - \log\mu), \qquad \dot z_y = -A^\top x - \beta(\log y - \log\mu).$$

This proves Theorem 2.2.

Interior invariance and normalization. For any $z \in \mathbb{R}^m$, $\mathrm{softmax}(z) \in \mathrm{relint}(\Delta_m)$; thus the ODE in z defines a flow whose pushforward $x = \mathrm{softmax}(z_x)$, $y = \mathrm{softmax}(z_y)$ always lies in $\mathrm{relint}(\Delta_m)^2$. In the *primal* form (Eq. equation 3), the Jacobians $J_x = \mathrm{Diag}(x) - xx^\top$ and $J_y = \mathrm{Diag}(y) - yy^\top$ satisfy $J_x\mathbf{1} = J_y\mathbf{1} = 0$, hence $\frac{d}{dt}\sum_i x_i = \mathbf{1}^\top \dot x = \mathbf{1}^\top J_x[\cdot] = 0$ and similarly for y, so the simplex constraints are preserved.

A.2 Well-posedness on $\mathrm{relint}(\Delta_m)^2$

Lemma 1 (Local Lipschitzness and existence/uniqueness)**.** *Fix a compact neighbourhood $\mathcal{U} \subset \mathrm{relint}(\Delta_m)^2$ around (x^*_β, y^*_β) with $\min_i x_i \ge \alpha$, $\min_j y_j \ge \alpha$ on $\mathcal{U}$ for some $\alpha > 0$. Then the primal vector field in equation 3 is locally Lipschitz on $\mathcal{U}$, and the ODE admits a unique solution for any initial condition in $\mathcal{U}$.*

Proof. On $\mathcal{U}$, $x \mapsto \log x$ and $y \mapsto \log y$ are Lipschitz with constants α^{-1} coordinatewise; $J_x(x)$ and $J_y(y)$ are smooth with operator norms bounded on $\mathcal{U}$. Thus $x \mapsto J_x[Ay - \beta(\log x - \log\mu)]$ and $y \mapsto J_y[A^\top x + \beta(\log y - \log\mu)]$ are locally Lipschitz, and Picard–Lindelöf yields existence/uniqueness. ☐

Remark A.1 (Local forward completeness)**.** *In the dual variables, the field is smooth everywhere; the softmax map keeps (x, y) interior for all finite times. On a small enough neighborhood of (x^*_β, y^*_β), solutions remain in $\mathcal{U}$, so the flow is forward complete there.*

A.3 Quantal-response characterization

[Prop. 2.3, detailed] Stationarity $(\dot z_x, \dot z_y) = (0, 0)$ is equivalent to

$$Ay^*_\beta = \beta(\log x^*_\beta - \log\mu) + c_x\mathbf{1}, \qquad -A^\top x^*_\beta = \beta(\log y^*_\beta - \log\mu) + c_y\mathbf{1}$$

for some scalars c_x, c_y. Subtracting per-coordinate means eliminates c_x, c_y, and exponentiation yields $x^*_\beta \propto \mu \odot \exp(\frac{1}{\beta}Ay^*_\beta)$, $y^*_\beta \propto \mu \odot \exp(-\frac{1}{\beta}A^\top x^*_\beta)$.

Proof. Directly from Theorem 2.2, stationarity implies the linear relations above. Let $\Pi(v) = v - \frac{1}{m}(\mathbf{1}^\top v)\mathbf{1}$ denote mean subtraction. Applying Π to both sides removes c_x, c_y, so $\Pi(\log x^*_\beta - \log\mu) = \frac{1}{\beta}\Pi(Ay^*_\beta)$, and similarly for y. Exponentiate elementwise and renormalise to the simplex. ☐

A.4 HAMILTONIAN/METRIC STRUCTURE AND SPECTRAL CALCULUS

We linearise the dual field in *log-probability* coordinates

$$u := \log x - \tfrac{1}{m}(\mathbf{1}^\top \log x)\,\mathbf{1}, \qquad v := \log y - \tfrac{1}{m}(\mathbf{1}^\top \log y)\,\mathbf{1},$$

which parametrise the quotient by additive constants and agree with z up to gauge. At an interior equilibrium (x^*_β, y^*_β), variations satisfy

$$\delta x = J^*_x\,\delta u, \qquad \delta y = J^*_y\,\delta v,$$

since for the softmax map $\delta x = (\mathrm{Diag}(x^*) - x^* x^{*\top})\,\delta(\log x) = J^*_x\,\delta u$.

Lemma 2 (Block Jacobian in (u,v))**.** *The Jacobian of the (u,v)-dynamics at (x^*_β, y^*_β) restricted to $\mathsf{T}\times\mathsf{T}$ is*

$$J_\beta = \begin{bmatrix} -\beta I & H \\ -K & -\beta I \end{bmatrix}, \qquad H := A\,J^*_y, \quad K := A^\top J^*_x.$$

Proof. Differentiate the right-hand sides of Theorem 2.2 w.r.t. (u,v) using $\delta(\log x) = \delta u$, $\delta(\log y) = \delta v$, and $\delta y = J^*_y \delta v$, $\delta x = J^*_x \delta u$. The diagonal blocks yield $-\beta I$; the off-diagonal blocks are $A\delta y$ and $-A^\top \delta x$. □

Lemma 3 (Nonnegativity of HK on T)**.** *On T, J^*_x, J^*_y are symmetric positive definite. Then*

$$HK = AJ^*_y A^\top J^*_x \sim \big(J_x^{*1/2} A J_y^{*1/2}\big)\big(J_x^{*1/2} A J_y^{*1/2}\big)^\top \succeq 0,$$

and similarly $KH \succeq 0$. Thus the eigenvalues of HK and KH are real and nonnegative.

Proof. Similarity with the symmetric product follows by inserting $J_x^{*1/2} J_x^{*-1/2}$ and $J_y^{*1/2} J_y^{*-1/2}$ and regrouping. □

[Thm. 2.4, detailed] Let $\{\sigma_k^2\}_{k=1}^{m-1}$ be the nonzero eigenvalues of HK on T. Then

$$\mathrm{spec}(J_\beta) = \{-\beta \pm i\sigma_k\}_{k=1}^{m-1}.$$

Proof. With $J = \left[\begin{smallmatrix} 0 & H \\ -K & 0 \end{smallmatrix}\right]$, we have $J^2 = \mathrm{diag}(-HK, -KH)$. By Lemma 3, HK and KH are diagonalizable with nonnegative spectra and share the same nonzero eigenvalues. Hence the eigenvalues of J are $\pm i\sigma_k$, and adding $-\beta I$ shifts the real parts by $-\beta$. □

Remark A.2 (Metric viewpoint (optional))**.** *Let $G = \mathrm{diag}(J_x^{*-1}, J_y^{*-1})$ on $\mathsf{T}\times\mathsf{T}$. Then J is G-skew-adjoint in the sense that $J^\top G + GJ = 0$ when $K = H^\top$ (e.g., when A is such that $AJ^*_y = (AJ^*_y)$ and $K = H^\top$); in general the spectral result follows from J^2 without assuming skew-adjointness.*

A.5 LYAPUNOV DECAY AND DISCRETE-TIME LAST-ITERATE RATES

Define the Lyapunov function

$$V(x,y) = D_{\mathrm{KL}}(x^*_\beta \| x) + D_{\mathrm{KL}}(y^*_\beta \| y) = D_\psi(x^*_\beta \| x) + D_\psi(y^*_\beta \| y).$$

By Bregman duality,

$$\frac{d}{dt} D_{\mathrm{KL}}(x^*_\beta \| x(t)) = \langle x(t) - x^*_\beta,\, \dot z_x(t)\rangle, \qquad \frac{d}{dt} D_{\mathrm{KL}}(y^*_\beta \| y(t)) = \langle y(t) - y^*_\beta,\, \dot z_y(t)\rangle.$$

Hence, along solutions of Theorem 2.2,

$$\begin{aligned} \dot V(t) &= \langle x - x^*_\beta,\, \nabla_x f_\beta(x,y)\rangle + \langle y - y^*_\beta,\, -\nabla_y f_\beta(x,y)\rangle \\ &= \langle (x - x^*_\beta, y - y^*_\beta),\, F(x,y) - F(x^*_\beta, y^*_\beta)\rangle, \end{aligned}$$

since $F(x^*_\beta, y^*_\beta) = 0$ at equilibrium.

Lemma 4 (Local strong monotonicity). *Under Assumption A2, there exist constants $m_\beta, L > 0$ and a neighborhood $\mathcal{N}$ of (x^*_β, y^*_β) such that for all $(x, y) \in \mathcal{N}$,*

$$\langle u - u^*,\ F(u) - F(u^*)\rangle\ \ge\ m_\beta\, \|u - u^*\|^2, \qquad \|F(u) - F(u^*)\|\ \le\ L\|u - u^*\|,$$

with $u = (x, y)$, $u^* = (x^*_\beta, y^*_\beta)$ and $m_\beta \ge c\,\beta$ for some $c > 0$.

Proof. The reverse–KL terms contribute a symmetric part proportional to β in the Jacobian on $\mathsf{T}\times\mathsf{T}$. Continuity of derivatives yields the bounds locally; the proportionality $m_\beta \propto \beta$ follows from the diagonal $-\beta I$ in Lemma 2. ☐

[Thm. 2.6, detailed] There exist $C_0, c_0 > 0$ and a neighborhood $\mathcal{N}$ such that solutions starting in $\mathcal{N}$ satisfy $V(x(t), y(t)) \le C_0\, e^{-c_0\beta t}$. Moreover, for sufficiently small steps η, discrete Nash–MD and Nash–MP converge linearly to (x^*_β, y^*_β) with factors $1 - \Theta(\beta\eta)$ (MD) and $1 - \Theta(\beta\eta)$ (MP with a larger admissible step).

Proof. Strong monotonicity (Lemma 4) implies $\dot V \le -m_\beta\|u - u^*\|^2$. On a compact neighborhood where $x_i, y_j \in [\alpha, 1]$ and $\sum_i x_i = \sum_j y_j = 1$, the Hessian $\nabla^2\psi(p) = \mathrm{Diag}(1/p_i)$ is bounded above and below, so V is locally equivalent to $\|u - u^*\|^2$: $\underline{c}\|u - u^*\|^2 \le V \le \overline{c}\|u - u^*\|^2$. Thus $\dot V \le -(m_\beta/\overline{c})V$, giving exponential decay with $c_0 = m_\beta/\overline{c} = \Theta(\beta)$. For the discrete schemes, standard results for (mirror) gradient methods on strongly monotone and Lipschitz VIs yield linear factors $1 - \eta m_\beta + O(\eta^2 L^2)$; choosing η small enough gives the stated rates. Full details appear in Appendix A.5.1. ☐

A.5.1 Sketch for discrete mirror descent and mirror-prox. Let $\Phi(u) = D_\psi(u^*\|u)$ with $u = (x, y)$. One-step analysis for mirror descent with step η gives $\Phi(u^{t+1}) \le \Phi(u^t) - \eta\langle u^{t+1} - u^*, F(u^t)\rangle + \frac{\eta^2 L^2}{2}\|u^{t+1} - u^t\|^2$. Strong monotonicity and Lipschitzness bound the right-hand side by $(1 - \eta m_\beta)\Phi(u^t)$ up to $O(\eta^2)$; mirror-prox admits a larger step via an intermediate evaluation of F. (We omit routine constants; see any standard VI text for details.)

A.6 Implicit-function sensitivity and constants

Let $g(x, y, \beta) = (\nabla_x f_\beta(x, y), -\nabla_y f_\beta(x, y))$ on $\mathsf{T}\times\mathsf{T}$. At $\beta = 0$, suppose $(x^*_0, y^*_0) \in \mathrm{relint}(\Delta_m)^2$ is an interior NE and $J_0 := D_{(x,y)}g(x, y, 0)\big|_{(x^*_0, y^*_0)}$ is nonsingular on $\mathsf{T}\times\mathsf{T}$ (Assumption A3). Then $g(x^*_\beta, y^*_\beta, \beta) = 0$ defines a C^1 curve (x^*_β, y^*_β) for small β by the implicit-function theorem.

Lemma 5 (Directional derivative at $\beta = 0$).

$$\frac{d}{d\beta}\Big|_{\beta=0}\begin{bmatrix} x^*_\beta \\ y^*_\beta \end{bmatrix} = -\,J_0^{-1}\begin{bmatrix} \partial_\beta\nabla_x f_\beta(x^*_0, y^*_0) \\ -\partial_\beta\nabla_y f_\beta(x^*_0, y^*_0) \end{bmatrix} = -\,J_0^{-1}\begin{bmatrix} \nabla_x D_{\mathrm{KL}}(x^*_0\|\mu) \\ -\nabla_y D_{\mathrm{KL}}(y^*_0\|\mu) \end{bmatrix},$$

where gradients are taken on T (constants along $\mathbf{1}$ vanish).

Proof. Differentiate $g(x^*_\beta, y^*_\beta, \beta) = 0$ at $\beta = 0$: $J_0\,[\frac{d}{d\beta}(x^*_\beta, y^*_\beta)]^\top + \partial_\beta g(x^*_0, y^*_0, 0) = 0$. Since $\partial_\beta\nabla_x f_\beta = -(\log x - \log\mu)$ and $\partial_\beta(-\nabla_y f_\beta) = -(\log y - \log\mu)$, projection to T yields the stated expression, which coincides with $\nabla D_{\mathrm{KL}}(\cdot\|\mu)$ up to a constant shift along $\mathbf{1}$ that vanishes. ☐

[Thm. 2.7, constants] There exist $\delta > 0$ and $C > 0$ such that for $\beta \in [0, \delta]$, $\|(x^*_\beta - x^*_0,\ y^*_\beta - y^*_0)\| \le C\,\beta$. One admissible constant is $C = \kappa\, M$ with $\kappa = \|J_0^{-1}\|$ and $M = \sup\{\|\partial_\beta g(x, y, \beta)\| : (x, y) \in \mathcal{N},\ \beta \in [0, \delta]\}$ on a small neighborhood $\mathcal{N}$.

Proof. The implicit-function theorem gives (x^*_β, y^*_β) continuously differentiable near 0 with derivative given by Lemma 5. Local Lipschitzness of g in (x, y, β) and boundedness of J_0^{-1} imply $\|(x^*_\beta, y^*_\beta) - (x^*_0, y^*_0)\| \le \int_0^\beta \|J_0^{-1}\|\,\|\partial_\beta g(x^*_s, y^*_s, s)\|\, ds \le \kappa M\,\beta$. ☐

Bounding $\|\partial_\beta g\|$. On T, $\partial_\beta g(x, y, \beta) = (-(\log x - \log \mu), -(\log y - \log \mu))$. On a compact neighborhood with $x_i, y_j \geq \alpha > 0$, the map $p \mapsto \log p$ is $1/\alpha$–Lipschitz coordinatewise, hence $\|\partial_\beta g\| \leq c(\alpha, \mu)$.

Norm equivalences used throughout. On any compact subset of $\mathrm{relint}(\Delta_m)$, there exist constants $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ such that for all $v \in \mathsf{T}$,

$$\underline{\lambda}\,\|v\|_2^2 \;\leq\; v^\top \nabla^2 \psi(p)\, v \;\leq\; \overline{\lambda}\,\|v\|_2^2,$$

with $\nabla^2 \psi(p) = \mathrm{Diag}(1/p)$. This yields equivalence between Euclidean norms and the local information-geometric (entropic) norms and underlies the conversions between $\|u - u^*\|^2$ and V in A.5.

Summary. A.1–A.6 establish: (i) the entropic Nash mirror ODE and interior invariance; (ii) local well-posedness; (iii) logit/quantal equilibria; (iv) a block Hamiltonian linearization with spectra $-\beta \pm i\sigma_k$; (v) Lyapunov decay and discrete linear rates with constants proportional to β; and (vi) $O(\beta)$ equilibrium sensitivity with explicit constants from J_0^{-1} and Lipschitz moduli.

Corollary A.3 (Parametric Taylor bound (specialization of Thm. 2.7))**.** *Let $\theta \mapsto (\pi_\theta, \pi'_\theta)$ be a C^2 local parametrization of policies near an interior unregularised NE θ_0^*, and define $P_\theta := P(\pi_\theta \succ \pi'_\theta)$. Suppose the reduced Hessian $H_P(\theta_0^*) := \nabla_\theta^2 P_\theta\big|_{\theta_0^*}$ is positive definite on the tangent (with smallest eigenvalue $\lambda_{\min} > 0$). Then for sufficiently small β,*

$$\|\theta_\beta^* - \theta_0^*\| \;\leq\; \frac{\beta}{\lambda_{\min}} \left\|\nabla_\theta D_{\mathrm{KL}}(\pi_{\theta_0^*} \| \mu)\right\| \;+\; O(\beta^2),$$

where θ_β^* is the β-regularised NE in parameter space.

Sketch. Write the stationarity conditions for $P_\theta - \beta\, D_{\mathrm{KL}}(\pi_\theta \| \mu)$, subtract the unregularised condition at θ_0^*, and first-order expand $\nabla_\theta P_\theta$ around θ_0^*: $H_P(\theta_0^*)(\theta_\beta^* - \theta_0^*) = \beta\, \nabla_\theta D_{\mathrm{KL}}(\pi_{\theta_0^*} \| \mu) + O(\beta^2)$. Taking norms and using $\|H_P^{-1}\| = 1/\lambda_{\min}$ yields the claim. This is a parametric specialization of Theorem 2.7; see Appendix A.6 for the general IFT-based constants. □

B Proofs for Section 3

B.1 Proof of Proposition 3.1 (Hessian–β: instantaneous damping)

Setup and notation. Fix an iteration index t and suppose the controller updates β every $K \geq 1$ steps (Assumption C1). During the t-th primal step (or the K-step mini-epoch), β is held constant at β_t; the dual/primal dynamics are those of the open-loop ODE with parameter β_t:

$$\dot{z} = G(z; \beta_t), \qquad z = (z_x, z_y), \quad x = \mathrm{softmax}(z_x),\; y = \mathrm{softmax}(z_y).$$

Let $J(z; \beta) := \nabla_z G(z; \beta)$ be the Jacobian of the dual field. At a neighborhood of the equilibrium $(x_{\beta_t}^*, y_{\beta_t}^*)$, Section 2 shows that

$$J(z^*; \beta_t) \;=\; \begin{bmatrix} -\beta_t I & H \\ -K & -\beta_t I \end{bmatrix}, \qquad \mathrm{spec}\left(J(z^*; \beta_t)\right) = \{-\beta_t \pm i\sigma_k\}_{k=1}^{m-1}, \tag{8}$$

with $H = AJ_y^*$, $K = A^\top J_x^*$, and $\{\sigma_k^2\}$ the nonzero eigenvalues of HK (Theorem 2.4).

Real parts (instantaneous damping). Because β is constant within the step, the linearization of the *closed-loop* system along the primal dynamics coincides with the *open-loop* linearization at β_t, hence equation 8 applies. Therefore, *instantaneously* the eigenvalues satisfy $\Re\lambda_t = -\beta_t$, proving the first claim.

One-step lower bound on β_{t+1}. By the controller update equation 5,

$$\beta_{t+1} \;=\; \mathrm{proj}_{[\beta_{\min}, \beta_{\max}]}\Big((1-\rho)\beta_t + \rho\, \zeta^*\, \hat{\sigma}_t\Big).$$

Assumption C2 gives $|\hat{\sigma}_t - \sigma_t| \leq \epsilon\, \sigma_t$ for a representative modal frequency σ_t. Hence $\hat{\sigma}_t \geq (1-\epsilon)\sigma_t$ and

$$\beta_{t+1} \;\geq\; (1-\rho)\beta_t + \rho\, \zeta^*(1-\epsilon)\sigma_t$$

before projection; the projection can only increase this lower bound if it clips up to $\beta_{\min}$. This proves the inequality in Proposition 3.1.

Uniform contraction rate. If $\zeta^*(1-\epsilon)\,\underline{\sigma} \geq \beta_{\min} > 0$ for a local lower bound $\underline{\sigma}$ of the representative modal frequency, then $\beta_{t+1} \geq \beta_{\min}$ holds inductively (assuming the projection box contains $\beta_{\min}$). Since $\Re\lambda_t = -\beta_t$, the local contraction rate of the linearised flow is uniformly bounded below by $\beta_{\min}$ at every step, completing the proof. $\square$

Remark B.1 (Two time-scales and mid-epoch linearization)**.** *Assumption C1 ensures β varies on a slower time-scale than the primal dynamics. For $K > 1$, the* intra-epoch *linearization uses fixed β_t; for $K = 1$, small ρ produces the same effect. Either way, the instantaneous real parts remain $-\beta_t$.*

B.2 Proof of Theorem 3.3 (Bias–β: budget tracking)

Closed-loop map in $\log\beta$. Let $u_t := \log\beta_t$. The controller equation 6 updates

$$u_{t+1} = u_t + \eta_\beta\,\delta_t, \qquad \delta_t = \operatorname{clamp}\big(B(\pi_t) - B^*,\, -b,\, b\big). \tag{9}$$

Let $u^* = \log\beta^*$ with $B(\pi^*_{\beta^*}) = B^*$.

Quasi-static approximation and linearization. By two time-scales (C1) and local convergence (Theorem 2.6), the policy iterate tracks the regularised equilibrium: $\pi_t = \pi^*_{\beta_t} + e_t$ with a small tracking error e_t that decays geometrically between controller updates. Write the budget map in the equilibrium proxy:

$$\bar{B}(u) := B(\pi^*_{e^u}), \qquad \bar{B}'(u^*) = s^* > 0 \quad \text{(Assumption C3)}.$$

Near u^* and ignoring saturation ($|\delta_t| < b$), we have

$$\delta_t = \bar{B}(u_t) - B^* + \nu_t = s^*(u_t - u^*) + r_t + \nu_t,$$

where r_t collects higher-order terms of the Taylor remainder (bounded by $c\,|u_t - u^*|^2$) and ν_t collects measurement/tracking noise from e_t and mini-batching[1]. Substituting into equation 9 yields

$$u_{t+1} - u^* = (1 - \eta_\beta s^*)\,(u_t - u^*) + \eta_\beta(r_t + \nu_t), \tag{10}$$

until clamps activate. When clamps activate ($|\delta_t| = b$), the increment is deterministically bounded by $|\eta_\beta b|$ and we can treat saturation as an additional bounded disturbance.

Mean stability and geometric rate. Taking expectations and using $\mathbb{E}\,\nu_t = 0$ and $\mathbb{E}\,r_t = O(\mathbb{E}|u_t - u^*|^2)$ gives

$$\mathbb{E}[u_{t+1} - u^*] = (1 - \eta_\beta s^*)\,\mathbb{E}[u_t - u^*] + O\big(\mathbb{E}|u_t - u^*|^2\big).$$

For $0 < \eta_\beta s^* < 2$, the linear part has contraction factor $\phi := |1 - \eta_\beta s^*| < 1$. For small enough neighbourhoods (so that the quadratic term is dominated), this yields the claimed geometric rate $|\mathbb{E}[u_t - u^*]| \leq \phi^t |u_0 - u^*|$.

Variance bound under bounded noise. From equation 10, ignoring higher-order r_t (or absorbing it into the disturbance as a bias term), we have an AR(1) recursion with bounded additive noise:

$$u_{t+1} - u^* = \phi\,(u_t - u^*) + \eta_\beta \xi_t, \qquad \phi = 1 - \eta_\beta s^*, \quad \xi_t := r_t + \nu_t,$$

where $|\xi_t| \leq b$ almost surely when clamps are active, and $\operatorname{Var}(\xi_t) \leq \sigma_\xi^2 < \infty$ by C4 otherwise. Standard AR(1) variance calculus gives the stationary bound

$$\sup_t \operatorname{Var}(u_t) \;\leq\; \frac{\eta_\beta^2\,\sigma_\xi^2}{1 - \phi^2} \;\leq\; \frac{\eta_\beta^2\, b^2}{1 - (1 - \eta_\beta s^*)^2},$$

where the second inequality uses the saturation envelope $|\xi_t| \leq b$ and thus $\sigma_\xi^2 \leq b^2$. This proves Theorem 3.3. $\square$

Remark B.2 (On the sign and size of s^*)**.** *Assumption C3 (strict increase in $u = \log\beta$) is natural when B measures drift toward μ (e.g., length ratio, ECE), because Theorem 2.7 implies $\frac{d\pi^*_\beta}{d\beta}\big|_{\beta=0}$ points in the $\nabla D_{\mathrm{KL}}(\cdot\|\mu)$ direction; if B correlates positively with this direction, then $s^* > 0$. In parametric models, this matches the first-order Taylor bound derived in the memo (Appendix A.6, Corollary A.3).*

[1] Under C4, ν_t is zero-mean with bounded variance; e_t can be folded into ν_t due to geometric decay between controller steps.

B.3 Composition: feasibility first, damping second

We justify the priority composition

$$\beta_{t+1}^{\text{bias}} \text{ from equation 6,} \qquad \beta_{t+1}^{\text{hess}} \text{ from equation 5,} \qquad \beta_{t+1} = \max\{\beta_{t+1}^{\text{bias}}, \beta_{t+1}^{\text{hess}}\}.$$

Proposition B.3 (Budget feasibility and damping lower bound)**.** *Suppose C1–C4 hold. If the Bias–β loop is mean-stable at $u^* = \log \beta^*$ (Theorem 3.3), then the composed update preserves budget feasibility in the limit: $B(\pi^*_{\beta_t}) \to B^*$ and $B(\pi_t) \le B^* + o(1)$ almost surely. Moreover, the realised damping obeys $\Re\lambda_t = -\beta_t \ge \min\{\beta_t^{bias}, \beta_t^{hess}\}$, and thus is at least the Hessian-target when the bias budget admits it (i.e., when $\beta_t^{hess} \ge \beta_t^{bias}$).*

Proof. By construction $\beta_{t+1} \ge \beta_{t+1}^{\text{bias}}$, so the composed loop cannot pick a β smaller than the bias-feasible one. Under mean stability of Bias–β (Theorem 3.3), $\beta_t^{\text{bias}} \to \beta^*$ and therefore $B(\pi^*_{\beta_t}) \to B^*$; two time-scales ensure $B(\pi_t) - B(\pi^*_{\beta_t}) \to 0$ in mean, yielding the stated feasibility. The damping statement follows from $\Re\lambda_t = -\beta_t$ (Proposition 3.1). □

Summary. Hessian–β sets the instantaneous damping via a spectral proxy, while Bias–β enforces a budget in expectation; the max-composition attains the larger of the two β values, ensuring *feasibility first, speed second.*

C Additional Method Explanations

C.1 Diagnostics, hyperparameters, and safeguards

Diagnostics. (i) *β-sweeps:* plot $D_{\text{KL}}(\pi^*_\beta \| \pi^*_0)$, ℓ_2, exploitability; (ii) *Spectral panels:* $\Re\lambda$ vs. β, $\Im\lambda$ (test Theorem 2.4); (iii) *Budget tracking:* $B(\pi_t)$ curves (overshoot and rate).

Hyperparameters. Hessian–β: $\zeta^* \in \{0.6, 0.8, 1.0\}$, $\rho \in \{0.2, 0.5\}$, update every $K \in \{1, 2\}$ steps. Bias–β: $\eta_\beta \in \{0.2, 0.4, 0.6\}$, $b \in \{0.25, 0.5\}$, $K \in \{1, 2, 4\}$. Bounds: choose $[\beta_{\min}, \beta_{\max}]$ from task ranges.

Complexity. Hessian–β: one JVP per controller step (two for centered differences); negligible vs. LLM forward. Bias–β: one scalar measurement on a small probe set. Memory overhead is negligible.

Safeguards. Two time-scales (small ρ, η_β or larger K), projection $\beta \in [\beta_{\min}, \beta_{\max}]$ and clamping for anti-windup, EMA for noisy $\hat{\sigma}$, and periodic re-estimation of μ to avoid lock-in.

D Additional experimental settings and results

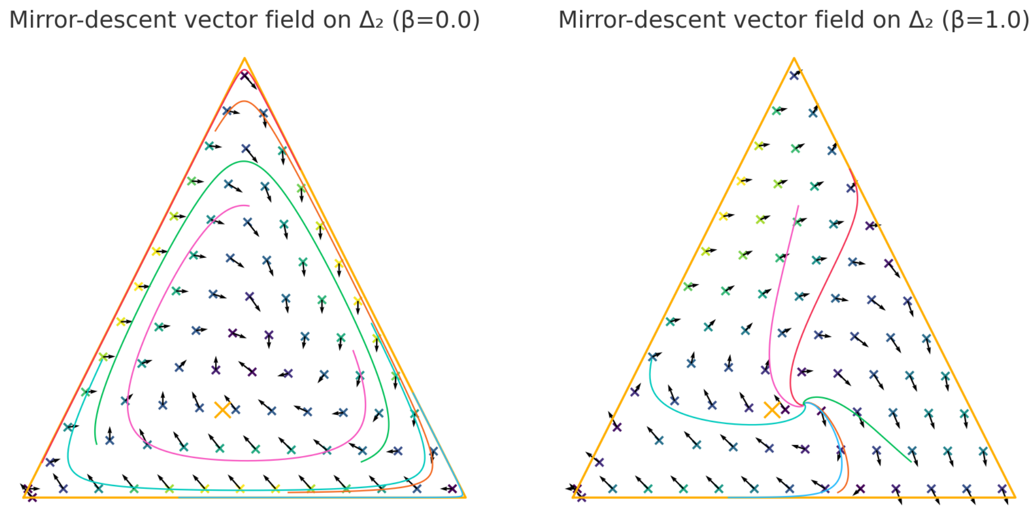
D.1 Experiments A: Toy Geometry on RPS

Setup. RPS payoff A (antisymmetric), biased reference $\mu = (0.36, 0.32, 0.32)$. We implement Nash-MD/MP, APMD-like, MPO/GRPO-like, and the two controllers. We render vector fields (symmetric restriction $x = y$), simulate trajectories, sweep β, and compute Jacobian spectra numerically at fixed points.

Hyperparameters. Stepsizes $\eta \in [0.1, 0.2]$; grid step 0.08 for fields; 600 steps for trajectories; finite difference $h = 10^{-6}$ for spectra. All seeds and code are released (see Reproducibility).

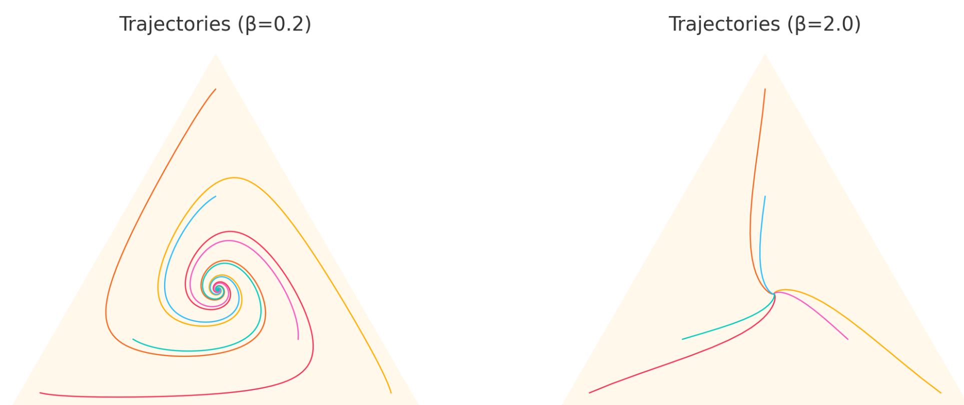
Findings. (i) Vector fields and trajectories confirm *spiral steepening* with β; (ii) $D_{\text{KL}}(x^*_\beta \| x^*_0)$ increases with β (bias amplification); (iii) real parts of spectra scale as $-\beta$, while imaginary parts remain roughly constant; (iv) Nash-MP and adaptive-β controllers converge faster than fixed-β MD while controlling bias.

Quantitative comparison. Table 2 summarises exploitability and budget-violation fractions (mean±CI over $n\!=\!6$ random initializations).



(a) Vector field ($\beta = 0$).   (b) Vector field ($\beta = 1$).

Figure 5: Simplex vector fields: larger β yields steeper spirals (increased damping).



(a) $\beta = 0.2$ (b) $\beta = 2.0$

Figure 6: Trajectory bundles: faster radial decay for larger β.

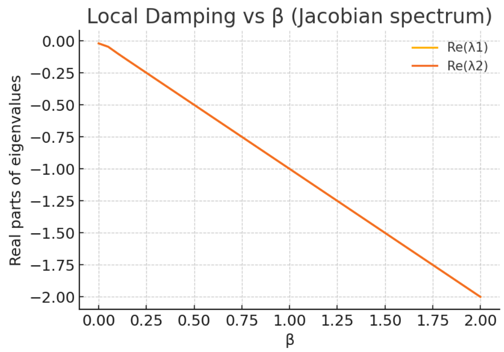
D.2 Ablations and Diagnostics

We ablate controller gains and β values to stress the speed–bias trade-off.

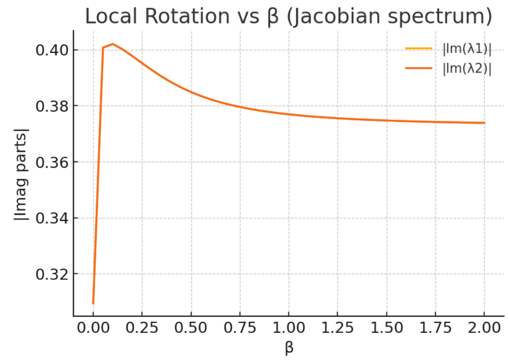
β-sweep. $\beta \in \{0, 0.1, 0.3, 1.0, 3.0\}$. *Observed:* (i) bias to μ (KL, ℓ_2) increases monotonically with β; (ii) exploitability decreases; (iii) spectral real parts follow $-\beta$ while imaginary parts are nearly constant. *Expected figures:* see Fig. 3 (right) for the toy setting; for LLMs we plan a small diagnostic panel mirroring these trends.

Bias–β gains. $\eta_\beta \in \{0.2, 0.4, 0.6\}$, clamp $b \in \{0.25, 0.5\}$. *Observed:* larger η_β improves budget tracking yet overshoots without clamps; clamps bound transients.

Hessian–β scaling. Target ratios $\zeta^\star \in \{0.6, 0.8, 1.0\}$, relaxation $\rho \leq 0.5$. *Observed:* increasing $\zeta^\star$ accelerates convergence but may drift bias if $\hat{\sigma}$ is noisy; clamping $\hat{\sigma}$ and EMA smoothing mitigate this. *Expected figure:* step vs. exploitability for different $\zeta^\star$.



(a) Real parts vs. β.



(b) Imag parts vs. β.

Figure 7: Spectra: real parts track $-\beta$ (uniform damping); rotation stays near-constant.

Table 2: Toy RPS ($n=6$ inits). *Takeaway:* **Bias–β** satisfies the bias budget with low variance while keeping exploitability competitive; Adaptive-KL meets a KL target but violates the *operational* bias budget frequently.

Method	final exploit (mean)	std	95% CI	bias viol. frac (mean)	std	95% CI
Nash–MD (β=0.5)	0.0562	0.0000	[0.0562, 0.0562]	0.138	0.021	[0.121, 0.155]
Nash–MP (β=0.5)	0.0562	0.0000	[0.0562, 0.0562]	0.164	0.041	[0.131, 0.197]
APMD-like (β=0.5, σ=0.02)	0.0606	0.0000	[0.0606, 0.0606]	0.817	0.014	[0.806, 0.829]
Adaptive–KL (KL*=0.03)	0.0398	0.0429	[0.0054, 0.0742]	0.814	0.059	[0.767, 0.861]
Bias–β (B^*=0.04)	**0.0489**	0.0005	[0.0485, 0.0493]	**0.042**	0.008	[0.036, 0.048]

APMD noise. $\sigma \in \{0, 0.01, 0.02\}$. *Observed:* noise breaks cycles but raises final exploitability at fixed targets. *Expected figure:* cycle magnitude vs. σ and final exploitability vs. σ.

D.3 Reproducibility and Artifacts

We release anonymised code/notebooks/figures/eval scripts (https://anonymous.4open.science/r/geonash-263B/). We log harness versions and commit hashes, and fix decoding settings (temperature, top-p, max tokens). All hyperparameters, data splits, and seeds are documented. Toy/micro plots are reproducible on CPU/1$\times$GPU in minutes; benchmark inference on 1$\times$A100/H100 80GB completes within one day with SC(10) on GSM8K.

E The Use of Large Language Models (LLMs)

The LLMs are used for assisting

- Manuscript structuring and editing
- Mathematical exposition (proof sketches and readability)
- LaTeX engineering

.