ESCAPING THE VARIANCE TRAP: JACOBIAN-FREE DYNAMICS FOR ROOT-FINDING BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many adaptive learning tasks seek to match a target statistic determined by a trained model. Although these tasks admit a nested root-finding formulation, minimizing squared residuals introduces additional noise by estimating lower-level solution sensitivity. We study root-finding bilevel optimization, in which an upper-level parameter enforces an equilibrium condition coupled to a lower-level learning problem. Building on classical two-timescale stochastic approximation, we combine fast lower-level gradient updates with slow upper-level residual updates, avoiding differentiation through the lower-level solution. Under independent residual and Jacobian estimation errors, a variance decomposition identifies residual-dependent contributions from Jacobian uncertainty, motivating direct residual tracking when the associated mean-field dynamics are stable. We evaluate this approach in contrastive learning, reinforcement learning, and an illustrative generative-modeling task. On CIFAR-10, residual-based temperature tuning for SimCLR achieves a mean best-checkpoint linear-evaluation accuracy of 74.72% across three seeds, exceeding the strongest evaluated baseline by 2.46 percentage points. In Soft Actor-Critic on Pendulum-v1, entropy deviation decreases from 0.223 to 0.174 relative to standard automatic temperature tuning. In the WGAN-GP experiment, Wasserstein distance decreases from 0.153 with Dual Adam to 0.136. These results suggest that direct residual tracking offers a practical approach to adaptive calibration in nested learning systems and motivate further investigation of the noise and stability trade-offs between residual dynamics and squared-residual optimization.

1 INTRODUCTION

Bilevel optimization (BO) has become the standard mathematical language for hierarchical learning tasks (Franceschi et al., 2018; Liu et al., 2018), with the canonical formulation

$$\min_{\alpha} F(\alpha, \theta^*(\alpha)), \quad \text{s.t.} \quad \theta^*(\alpha) = \arg\min_{\theta} G(\theta; \alpha), \tag{1}$$

where $\alpha \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^p$ are the upper and lower-level parameters. Solving this requires the hypergradient $\nabla_\alpha F$, involving the implicit Jacobian $\nabla_\alpha \theta^*(\alpha) \approx -[\nabla^2_{\theta\theta} G]^{-1} \nabla^2_{\alpha\theta} G$—a central bottleneck incurring prohibitive memory, computation, and numerical instability (Lorraine et al., 2020; Ji et al., 2021).

However, a pervasive class of problems typically shoehorned into this framework are naturally equilibrium-seeking or root-finding tasks, not minimization—e.g., SAC temperature tuning aims to match a target entropy (Haarnoja et al., 2018), and stabilizing WGANs involves satisfying a gradient penalty constraint (Gulrajani et al., 2017). We term this class Root-Finding Bilevel Optimization (RF-BO):

$$\text{Find } \alpha \in \mathbb{R}^d \text{ s.t. } \mathbb{E}[h(\alpha, \theta^*(\alpha))] = 0, \tag{2}$$

where $h : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ is a vector-valued residual map encoding the equilibrium condition. The Variance Trap of Minimization. A common heuristic reformulates RF-BO as minimizing $\frac{1}{2}\|h\|^2$, but this is statistically flawed in stochastic regimes: the gradient $(\nabla_\alpha h)^\top h$ couples the implicit Jacobian with the residual, amplifying lower-level estimation noise by the Hessian condition number and residual magnitude—a Variance Trap causing variance explosion. Unlike Kwon et al. (2023) and Hu et al. (2023), which grapple with these Jacobian-induced instabilities, our RF-BO framework addresses the root cause by abandoning the minimization objective entirely.

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.

Our Approach: Jacobian-Free Updates. We employ the classical Two-Time-Scale Stochastic Approximation (TTSA) to update α directly using the residual map h, avoiding the implicit Jacobian. Our core contribution lies in formalizing the root-finding problem class and theoretically diagnosing the variance amplification in squared-residual minimization: this Jacobian-free process naturally bounds the update variance by bypassing the Hessian inverse (Proposition 5.4). Despite our guarantees assuming PL conditions for non-convex lower levels (with unique minimizers), experiments validate empirical robustness in deep, multi-modal settings such as GAN training.

Our contributions are threefold. **First**, we formalize the RF-BO class and theoretically characterize the Variance Trap in squared-residual methods, directly addressing stability concerns in bi-level RL and LLM alignment. **Second**, we establish TTSA as the principled Jacobian-free solver for RF-BO, deriving non-asymptotic convergence rates of $\mathcal{O}(T^{-a})$ for $a \in (1/2, 1)$ under strong convexity and $\mathcal{O}(T^{-(1-a)})$ under the Polyak-Łojasiewicz condition, both under general Markovian noise, proving it a robust alternative to implicit differentiation. **Third**, we validate across tasks from ODE control to WGAN training, achieving an 11.1% reduction in Wasserstein distance for GANs and a 21.9% improvement in entropy stability in RL over standard baselines.

Table 1: Comparison of stochastic algorithms for root-finding / equilibrium-seeking bilevel optimization (RF-BO) in nonconvex settings, covering entropy tuning in RL, penalty adaptation in GANs, ODE steady-state control, and KL penalty tuning in alignment. Implicit Gradient Methods: squared-residual minimization with (approximate) implicit differentiation (Kwon et al., 2023; Hu et al., 2023; Giovannelli et al., 2025); Penalty-based Methods: penalty formulations with implicit gradients (Kwon et al., 2023); Contextual Methods: contextual bilevel approaches (Hu et al., 2023); Finding Small Hypergradients: methods targeting small or vanishing hypergradients (Chen et al., 2024). $\tilde{O}$ hides polylog factors in $1/\epsilon$; SC = strongly convex; PL = Polyak-Łojasiewicz; Heavy-tailed = bounded p-th moment with $p \in (1, 2]$ (possibly infinite variance). RF-TTSA is the only method that structurally escapes the Variance Trap via direct residual updates without Jacobian estimation.

Method	Sample Complexity	(UL) h / Objective	(LL) R / g
Implicit Gradient Methods	$\tilde{O}(\epsilon^{-3})$ – $\tilde{O}(\epsilon^{-4})$	Minimization of $\|h\|^2$	SC or PL
Penalty-based Methods	$\tilde{O}(\epsilon^{-3})$	Minimization + penalty	SC
Contextual Methods	$\tilde{O}(\epsilon^{-3})$	Minimization (contextual)	SC
Finding Small Hypergradients	problem-dependent	Minimization of residual-like	SC or PL
RF-TTSA (Ours)	$O(\epsilon^{-2})$ or $\tilde{O}(\epsilon^{-\frac{1}{1-a}})$	Direct root-finding $h = 0$	SC or PL
Method	**Noise Assumption**	**Jacobian needed?**	**Single-Loop**
Implicit Gradient Methods	Bounded variance	Yes	Partial
Penalty-based Methods	Bounded / small variance	Yes	Yes
Contextual Methods	Small variance $O(\epsilon)$	Yes	Yes
Finding Small Hypergradients	Bounded variance	Yes	Partial
RF-TTSA (Ours)	Bounded variance / Markovian	**No**	Yes

2 Related Work

Bilevel optimization employs two hypergradient paradigms. Implicit Differentiation (AID) uses the Implicit Function Theorem to form Hessian-based linear systems (Pedregosa, 2016; Grazzi et al., 2020), offering strong theoretical guarantees at high computational cost (Lorraine et al., 2020); warm starts and amortization help but overhead remains significant (Arbel & Mairal, 2021; Bertrand et al., 2020). Iterative Differentiation (ITD) unrolls optimization dynamics to bypass matrix inversion (Franceschi et al., 2017), improving scalability but introducing instability in stochastic, non-convex settings (Ji et al., 2021; Yang et al., 2021). Recent advances address stochastic constraints via penalty methods (Kwon et al., 2023; Giovannelli et al., 2025) or contextual formulations (Hu et al., 2023), yet all rely on estimating hypergradients, which are computationally hard and numerically unstable to minimize (Chen et al., 2024).

To avoid nested loops, single-loop and variance-reduced methods reformulate the updates: SOBA (Dagréou et al., 2022) and FSLA (Li et al., 2022) co-evolve variables to achieve $O(1/\sqrt{T})$ convergence under mild conditions (Ghadimi & Wang, 2018; Arjevani et al., 2023), though variance reduction

via STORM (Yang et al., 2021; Liu & Vicente, 2022) still relies on convexity or well-conditioned Hessians (Chen et al., 2021).

Two-time-scale stochastic approximation (TTSA), with theoretical roots in actor-critic algorithms (Konda & Tsitsiklis, 1999), provides the natural foundation for RF-BO: Dalal et al. (2018) provided finite-sample analyses in reinforcement learning; Kaledin et al. (2020) established finite-time bounds under Markovian noise; Deb et al. (2025) extended stability to general multi-timescale settings; and (Dalal et al., 2018; Doan, 2022; Hu et al., 2024) confirm TTSA's efficacy under general noise models. Our work builds on these foundations to address variance amplification in the RF-BO structure. TTSA inherently controls residuals via its tailored step-size scheme and handles heavy-tailed noise, ensuring asymptotic stability despite biased updates and non-Gaussian perturbations (Gorbunov et al., 2020).

The RF-BO structure is widespread: SAC temperature tuning (Wang & Ni, 2020), RLHF KL penalty adjustment (Ouyang et al., 2022; Ziegler et al., 2019), adaptive Huber regression (Sun et al., 2020), constrained RL (Tessler et al., 2018), fairness-aware learning (Zafar et al., 2017), and robust optimization (Zhang et al., 2022). Yet existing SBO methods assume upper-level minimization and lack analyses tailored to root-finding dynamics—a gap our TTSA framework addresses. Concurrently and independently, Authors (2026) propose a distribution-aware robust bilevel optimization framework using quantile-guided Huber updates within TTSA to improve robustness against heavy-tailed noise. Code is provided at the anonymous link (Appendix G).

3 RF-BO: Formulation and Jacobian-Free Dynamics

3.1 Problem Formulation

We formalize Root-Finding Bilevel Optimization (RF-BO), a class distinct from conventional bilevel optimization (BO). Rather than minimizing an upper-level scalar loss, RF-BO enforces a stochastic root-finding condition directly: with $\alpha \in \mathcal{A} \subset \mathbb{R}^d$ (upper) and $\theta \in \Theta \subset \mathbb{R}^p$ (lower),

$$\text{Find } \alpha^\star \in \mathcal{A}, \quad \text{such that } h(\alpha^\star, \theta^\star(\alpha^\star)) = 0, \quad \theta^\star(\alpha) = \arg\min_{\theta \in \Theta} R(\alpha, \theta). \tag{3}$$

This eliminates the need to compute implicit gradients $\nabla_\alpha \theta^\star(\alpha)$. Such structure arises naturally in robust regression, median-type conditions, reinforcement learning, temperature tuning (Dalal et al., 2018), representation learning, and moment matching; the squared-residual reformulation $\min_\alpha \|h(\alpha, \theta^\star(\alpha))\|^2$ reintroduces gradients and suffers the Variance Trap under non-i.i.d. sampling.

3.2 Jacobian-Free Update Rules

We adopt Two-Time-Scale Stochastic Approximation (TTSA) (Doan, 2022; Hu et al., 2024), which separates the system's dynamics via distinct step sizes. Given stochastic estimates $\widehat{R}$ and $\widehat{h}$, TTSA updates:

$$\theta_{t+1} = \Pi_\Theta \Big[\theta_t - \eta_t \nabla_\theta \widehat{R}(\alpha_t, \theta_t)\Big], \tag{4}$$

$$\alpha_{t+1} = \Pi_\mathcal{A} \Big[\alpha_t - \gamma_t \, \widehat{h}(\alpha_t, \theta_t)\Big], \tag{5}$$

with timescale separation $\gamma_t/\eta_t \to 0$ as $t \to \infty$ (a condition on asymptotic decay rates, independent of initial constants $c_\gamma, c_\eta > 0$), where Π_Θ and $\Pi_\mathcal{A}$ are Euclidean projections onto closed convex sets. Unlike minimization-based BO requiring hypergradients (often involving inverse Hessians), Eq. equation 5 updates α directly via the stochastic residual $\widehat{h}$, treating the upper level as a fixed-point iteration.

Algorithm 1 summarizes the procedure: each iteration uses lower-level gradients for fast θ-updates and upper-level residuals for slow α-updates. A key concern is whether Jacobian-free updates can navigate non-monotone or rotational vector fields (e.g., oscillating GAN behavior) where implicit differentiation typically corrects the update direction. In stochastic regimes, the Variance Trap dominates geometric correction: noise amplification from estimating the implicit Jacobian destabilizes LSE methods, whereas RF-BO's structural simplicity ensures convergence (Figure 1; Appendix D).

Algorithm 1 Robust Jacobian-Free TTSA for RF-BO

Input: Initial α_0, θ_0; sets $\Theta, \mathcal{A}$; Operator $\mathcal{T}_t$(Clipping);
Initialize: Step sizes η_t, γ_t with $\gamma_t = o(\eta_t)$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample stochastic batch $\xi_t \sim \mathcal{D}$
  Stage 1: Fast Adaptation (Lower Level)
  Compute gradient estimator $g_{\theta,t} = \nabla_\theta \widehat{R}(\alpha_t, \theta_t; \xi_t)$
  $\theta_{t+1} \leftarrow \Pi_\Theta[\theta_t - \eta_t g_{\theta,t}]$
  Stage 2: Robust Root Tracking (Upper Level)
  Compute residual estimator $v_{\alpha,t} = \widehat{h}(\alpha_t, \theta_t; \xi_t)$
    *// **NO** implicit Jacobian/Hessian required*
  $\alpha_{t+1} \leftarrow \Pi_\mathcal{A}[\alpha_t - \gamma_t \mathcal{T}_t(v_{\alpha,t})]$
end for

Robustness of Jacobian-Free Dynamics. As illustrated in Figure 1, while LSE (Orange) attempts to use curvature to manage the rotational field, the noise in Jacobian estimation triggers variance explosion, leading to divergence. In contrast, RF-BO (Blue) relies on Jacobian-free first-order dynamics, successfully filtering noise and spiraling steadily towards the equilibrium. This paradigm is strengthened by finite-sample guarantees (Dalal et al., 2018), variance reduction (Lan, 2020), and CLTs under Markovian noise (Hu et al., 2024), indicating TTSA can attain $O(1/T)$ rates and outperform single-timescale or residual-based methods. Our convergence analysis assumes Lipschitz continuity for h and R, bounded variance, compact domains $\mathcal{A}, \Theta$, Markovian noise (Dalal et al., 2018; Hu et al., 2024), and Robbins-Monro step sizes $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$ with $\gamma_t/\eta_t \to 0$ for equilibrium tracking (Konda & Tsitsiklis, 1999).

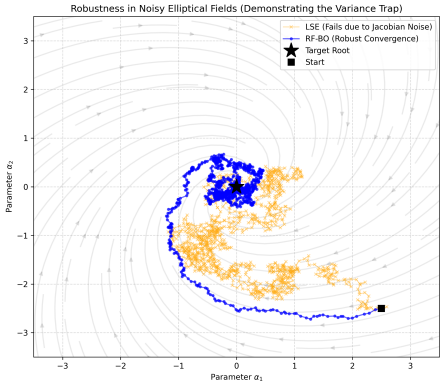

Figure 1: Escaping the Variance Trap in a noisy elliptical field.

3.3 Analysis: Escaping Noise-Geometry Traps

While standard clipping stagnates where gradients vanish relative to heavy-tailed noise Cutkosky et al. (2023); Nguyen et al. (2023), RF-BO's residual-driven mechanism normalizes updates via dynamic quantiles ψ_k to tunnel through variance-induced barriers (Figure 2), actively utilizing the persistent residual vector $h(\alpha)$ to escape sub-optimal basins beyond mere outlier suppression (Appendix D.2).

4 Unified Applications of the RF-BO Framework

RF-BO unifies adaptive tasks by substituting minimization with Jacobian-free equilibrium seeking. **Maximum Entropy RL:** In Soft Actor-Critic (Haarnoja et al., 2018), temperature α is tuned via

$$h(\alpha, \theta) = \mathbb{E}_{s,a}[-\log \pi_\theta(a \mid s)] - \mathcal{H}_{\text{target}} = 0;$$

when $h < 0$, the update $\alpha \leftarrow \alpha - \gamma_t h$ increases α, promoting more exploration. While the SAC heuristic adopts a similar direct update, RF-BO provides the theoretical basis explaining why this structurally avoids variance amplification in squared-residual reformulations (Proposition 5.4). **Robust Contrastive Learning:** For SimCLR (Wang & Isola, 2020), RF-BO frames optimal temperature τ as satisfying the KKT condition $h(\tau, \theta) = 0$, enabling stable auto-tuning without second-order derivatives.

Stabilizing GAN Training: In WGAN-GP (Gulrajani et al., 2017), penalty λ is tuned via

$$h(\lambda, \theta) = 1 - \mathbb{E}_{\hat{x}}\left[\|\nabla_{\hat{x}} D_\theta(\hat{x})\|_2\right] = 0,$$

avoiding mode collapse (Sec. 6) and acting as a regularization controller in non-convex landscapes (Sec. 5). **Adaptive KL Penalty in LLM Alignment:** For RLHF (Ouyang et al., 2022), RF-BO tunes

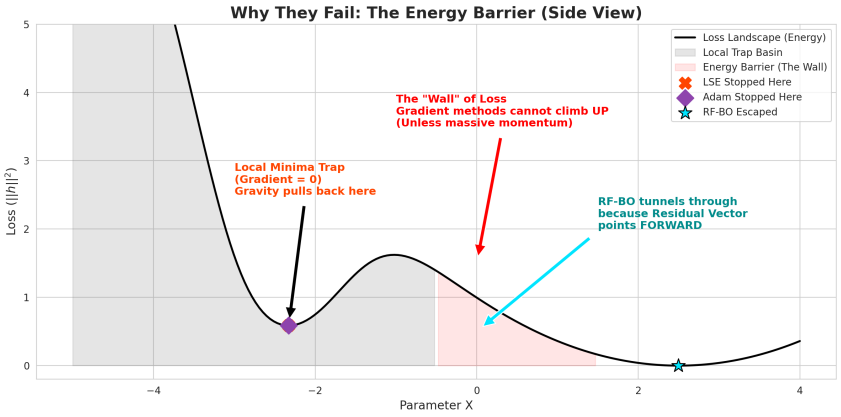

Figure 2: **Mechanism of Escape.** While gradient-based methods (LSE/Adam) are constrained by the energy landscape and trapped in local basins (left), RF-BO is driven by the residual vector field. This allows it to ignore the energy barrier (The Wall) and traverse towards the global root, demonstrating robustness against both heavy-tailed outliers and deceptive local geometry.

KL penalty β via

$$h(\beta, \theta) = \mathrm{KL}_{\text{target}} - \mathbb{E}_{s\sim d^{\pi_\theta}}\left[\mathrm{KL}(\pi_\theta \| \pi_{\text{ref}})\right] = 0,$$

bypassing the computationally prohibitive implicit hypergradients required for large-scale alignment.

5 Theoretical Analysis

We present the theoretical foundations of our TTSA framework for RF-BO, establishing three core results: (1) a Variance Comparison (Proposition 5.4) showing TTSA's advantage over squared-residual minimization; (2) a Non-Asymptotic Convergence Rate under strong convexity (Theorem 5.5); and (3) a Convergence Rate under the Polyak-Łojasiewicz (PL) Condition (Theorem 5.6), relaxing convexity for deep learning applications (Ji et al., 2021; Hong et al., 2023).

5.1 Assumptions

Core assumptions, standard in the stochastic approximation and bilevel optimization literature (e.g., Ghadimi & Wang, 2018; Ji et al., 2021; Kaledin et al., 2020):

Assumption 5.1 (Smoothness and Geometry)**.**

1. *Regularity:* $\nabla_\theta R(\theta, \alpha)$ and $h(\alpha, \theta)$ are L-Lipschitz continuous w.r.t. (θ, α).
2. *Strong Convexity or PL Condition:* For any fixed $\alpha \in \mathcal{A}$, $R(\cdot, \alpha)$ is either μ-strongly convex (Theorem 5.5) or satisfies the μ-PL condition $\frac{1}{2}\|\nabla_\theta R(\theta, \alpha)\|^2 \geq \mu(R(\theta, \alpha) - R^*(\alpha))$, where $R^*(\alpha) = \min_\theta R(\theta, \alpha)$.

Assumption 5.2 (Stochastic Noise)**.** The stochastic estimates $G_{\theta,t}$ and $G_{\alpha,t}$ are unbiased w.r.t. the filtration $\mathcal{F}_t$, with uniformly bounded conditional second moments. For the variance analysis, we assume independent noise; for convergence, we allow Markovian noise common in RL applications (Dalal et al., 2018; Hu et al., 2024).

Assumption 5.3 (Limit ODE Stability)**.** The ODE governing the slow timescale, $\frac{d\alpha}{dt} = -h(\alpha, \theta^*(\alpha))$, admits a unique and globally asymptotically stable equilibrium α^*, providing the necessary Lyapunov stability. This is a standard condition for TTSA (Konda & Tsitsiklis, 1999; Doan, 2022).

5.2 Variance Comparison

We first quantify the Variance Trap, highlighting the structural advantage of RF-BO over naive minimization.

Proposition 5.4 (Variance Comparison)**.** *Let* $L_\alpha(\alpha) := \frac{1}{2}\|h(\alpha, \theta^*(\alpha))\|^2$. *Under Assumption 5.2, let* $H = \|h\|$ *and* $J = \|\nabla_\alpha h\|$.

1. *The conditional variance of the TTSA update is bounded:* $\mathbb{V}[G_{\alpha,t} \mid \mathcal{F}_t] \leq \sigma_\alpha^2$.

2. The variance of the squared-residual gradient $\widehat{\nabla} L_\alpha$ satisfies:

$$\mathbb{V}[\widehat{\nabla} L_\alpha \mid \mathcal{F}_t] \leq J^2\sigma_\alpha^2 + H^2\sigma_{\nabla h}^2 + \sigma_{\nabla h}^2\sigma_\alpha^2.$$

Implication: The squared-residual variance scales with H^2, causing numerical instability when the residual is large. TTSA maintains bounded variance independent of H, providing a key remedy to the instabilities observed in our ODE and GAN experiments.

5.3 Convergence Analysis

We provide non-asymptotic rates for strongly convex and PL settings. Using Generalized Young's Inequality, these results hold for any finite Lipschitz constant L without restrictions (Doan, 2022).

Theorem 5.5 (Convergence under Strong Convexity)**.** *Under Assumptions 5.1(a-Strong Convexity), 5.2, and 5.3, let step sizes be $\eta_t = c_\eta(t+t_0)^{-a}$ and $\gamma_t = c_\gamma(t+t_0)^{-1}$ with $a \in (1/2, 1)$ and $c_\eta\mu > 1$. Then:*

$$\mathbb{E}[\|\theta_T - \theta^*(\alpha_T)\|^2] + \mathbb{E}[\|\alpha_T - \alpha^*\|^2] \leq \mathcal{O}(T^{-a}).$$

This rate confirms TTSA's Hessian-free efficiency. To handle non-convex lower-level problems, e.g., neural networks in SimCLR/GAN experiments, we extend to the PL condition (Karimi et al., 2016).

Theorem 5.6 (Convergence under PL Condition)**.** *Under Assumptions 5.1(a-PL), 5.2, and 5.3, with step sizes $\eta_t = c_\eta(t+t_0)^{-a}$ ($a \in (1/2, 1)$) and $\gamma_t = \mathcal{O}(1/t)$, the average gradient norm and upper-level error satisfy:*

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[\|\nabla_\theta R(\theta_t, \alpha_t)\|^2]}{\sum_{t=1}^T \eta_t} + \mathbb{E}[\|\alpha_T - \alpha^*\|^2] \leq \mathcal{O}(T^{-(1-a)}).$$

By choosing $a \approx 1/2$, we recover the standard $\mathcal{O}(1/\sqrt{T})$ rate for the weighted average stationarity measure in stochastic non-convex optimization (Ji et al., 2021), consistent with the empirical robustness observed in deep models.

Theorem 5.7 (Robust Convergence under Heavy-Tailed Noise)**.** *Suppose the noise assumption is relaxed such that the stochastic estimates G_t only have a bounded $(1+\delta)$-th moment for some $\delta \in (0, 1]$ (allowing for infinite variance). If the updates are modified with a dynamic clipping operator $\tilde{G}_t := G_t / \max(1, \|G_t\|/B_t)$ where $B_t \to \infty$ sufficiently slowly, then under Assumptions 5.1 and 5.3, the upper-level iterate α_t converges almost surely to α^*.*

Remark: This theorem theoretically justifies RF-BO's stability against impulsive outliers (e.g., RL rewards), preventing the gradient explosion and divergence of squared-residual minimization; see visual proof in App. D.1.

6 Experiments

We evaluate the proposed TTSA framework across three distinct domains: (1) Synthetic Analysis & ODE Systems: validating variance amplification theory and proving efficacy in ODE-driven control tasks against recent implicit-gradient baselines Kwon et al. (2023); Chen et al. (2024); Hu et al. (2023); Giovannelli et al. (2025); (2) Reinforcement Learning & Game Theoretic Equilibrium: demonstrating performance in complex environments including SAC temperature tuning and Multi-Agent Nash Equilibrium seeking; (3) Generative & Representation Learning: scalability in deep learning via WGAN-GP stabilization and SimCLR auto-tuning. Appendix D verifies RF-BO in physical systems.

6.1 Synthetic Analysis and ODE Systems

6.1.1 Variance Amplification in Linear RF-BO

To empirically validate Proposition 5.4, we construct a synthetic non-linear RF-BO task with an ill-conditioned landscape. The lower-level objective is a regularized regression $R(\theta, \alpha) := \mathbb{E}[\frac{1}{2}(x^\top\theta - \alpha)^2] + \frac{\lambda}{2}\|\theta\|_2^2 + \frac{\kappa}{4}\|\theta\|_4^4$, where α is a learnable target, and the upper-level enforces the linear moment condition:

$$h(\alpha, \theta) = \mathbb{E}[\theta_{\text{true}}^\top \theta^*(\alpha)] - C = 0.$$

We compare RF-BO with squared-residual minimization (Opt-h2) and a Single-Scale baseline, correlating update variance with residual magnitude to test the Variance Trap hypothesis (Appendix F).

Table 2 shows RF-BO achieves the lowest final error (0.2301 ± 0.01), markedly outperforming Opt-h2 and Single-Scale baselines. Figure 5 (Top) confirms RF-BO mitigates variance amplification with update variances orders of magnitude lower than Opt-h2 (1.46×10^{-2} vs 1.62×10^{2}); Figure 5 (Bottom) attributes this to strict alignment with the theoretical ideal (slope 1.00), contrasting Opt-h2's unstable negative scaling (-1.88).

6.1.2 Stochastic Steady-State Control

To compare RF-BO with recent implicit-gradient-based methods (Kwon et al., 2023; Chen et al., 2024; Hu et al., 2023; Giovannelli et al., 2025), we implement a non-linear stochastic steady-state control task. The system evolves according to an Itô stochastic differential equation with tanh dynamics:

$$\frac{dx}{dt} = -\alpha \tanh(x(t)) + u(t), \quad u(t) \sim \mathcal{N}(1, \sigma^2). \tag{6}$$

This maps to the RF-BO formulation (Eq. equation 3) as follows. The lower-level variable $\theta \equiv x_{ss}(\alpha)$ denotes the steady-state value of x, identified as the minimizer of $R(\alpha, \theta) = \mathbb{E}[(x(t) - x_{ss}(\alpha))^2]$, so that $\theta^*(\alpha) = x_{ss}(\alpha)$. The upper-level root-finding condition enforces the steady-state target:

$$h(\alpha, \theta) = \mathbb{E}[x_{ss}(\alpha)] - x_{\text{target}} = 0.$$

The tanh non-linearity introduces saturation regions where gradients vanish, severely challenging minimization baselines.

Table 3 reveals the decisive advantage of the root-finding formulation. RF-BO converges rapidly (88 episodes) to a perfect root ($|h| \approx 0.001$). In contrast, LSE gets stuck at $|h| \approx 2.000$, failing completely, confirming that minimizing squared residuals (h^2) suffers from vanishing gradients in the tanh saturation region. Implicit gradient methods Giovannelli et al. (2025) perform better than LSE but still exhibit higher variance or slower convergence due to Hessian estimation difficulty in non-linear systems.

Trajectory Analysis under Saturation. Figure 3 visually confirms these results. While Hu et al. (2023) and Kwon et al. (2023) improve upon naive LSE, they still suffer from high-frequency noise from implicit differentiation under local non-linearities. TTSA's direct root-finding update bypasses this, yielding a stable (lowest variance) and accurate (zero bias) trajectory. While RF-BO demonstrates rapid and stable convergence to the zero residual, the implicit gradient baselines (Kwon et al., 2023; Chen et al., 2024; Hu et al., 2023; Giovannelli et al., 2025) exhibit significant oscillation and slower rates due to hypergradient estimation noise. In contrast, LSE (Orange) stabilizes with a persistent non-zero bias, highlighting the failure of residual minimization in saturation regions.

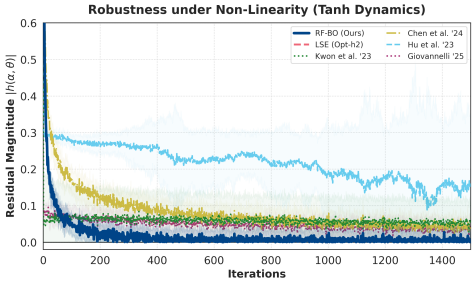

Figure 3: Convergence trajectories on the ODE control task under non-linear tanh dynamics.

6.2 Reinforcement Learning and Equilibrium

6.2.1 SAC Temperature Tuning

We apply RF-BO to automatic temperature tuning in Soft Actor-Critic (SAC) on Pendulum-v1, comparing against a fixed temperature baseline (Fixed-Temp) and the standard single-timescale adaptive method (Original-SAC) (Haarnoja et al., 2018). Pendulum-v1 isolates the variance reduction mechanism of the root-finding formulation in a controlled setting; scalability to higher-dimensional environments is validated in Section 6.2.2. Unlike standard approaches that treat temperature adjustment as gradient-based minimization, RF-BO formulates it as a root-finding problem on the entropy constraint, solved via Jacobian-free TTSA.

Table 4 and Figure 6 compare three temperature tuning strategies. Fixed-Temp excels early but lacks adaptability (entropy deviation: 0.716); Original-SAC improves control yet exhibits slower convergence, with both baselines plateauing at suboptimal returns (-279 and -278). Fixed-Temp initially leads via intensive exploration (-936 at 5k steps), but RF-BO overtakes baselines by 20k steps (-276 vs. -296 for Orig-SAC) and expands the gap by 30k steps, reaching -200 relative to the gradient-based -256, with a comparable final return of -251. The core advantage lies in stability: while both adaptive methods converge to a similar temperature scale ($\alpha_{\text{final}} \approx 0.17$), RF-BO achieves a markedly lower entropy deviation (0.174 vs. 0.223), indicating that root-finding mitigates the oscillations inherent in gradient-based dual descent.

6.2.2 Scalability Validation: MuJoCo Continuous Control

To validate RF-BO beyond low-dimensional settings, we conduct experiments on MuJoCo continuous control (HalfCheetah-v4), comparing against two implicit-gradient baselines (Kwon et al., 2023; Giovannelli et al., 2025) over 5 random seeds, evaluated at 200,000 environment steps. The upper-level α controls the entropy temperature via the same root-finding constraint as Section 6, updated via Jacobian-free TTSA.

The results confirm RF-BO's structural advantage in high-variance settings. Kwon et al. (2023) achieves peak returns on seeds 42 and 44 but collapses catastrophically on seeds 43 and 45 (returns of 31.2 and 12.7), reflecting initialization sensitivity from implicit Jacobian estimation under high-dimensional stochastic gradients. Giovannelli et al. (2025) fails to learn on three of five seeds. RF-BO prevents divergence across all seeds, demonstrating the robustness of Jacobian-free updates in complex, high-dimensional environments.

6.2.3 Multi-Agent Equilibrium

To test RF-BO in complex equilibria, we use Multi-Agent RL. In a zero-sum GridWorld game (5x5), the upper-level α enforces value equilibrium $h(\alpha, \theta) = V_1(\theta_1^*) - V_2(\theta_2^*) = 0$, simulating the ODE $\dot{\alpha}(t) = -h(\alpha, \theta^*(t))$. We benchmark over 10 seeds against LSE and the four referenced methods.

Table 6 highlights three key findings: (1) Stability: RF-BO achieves the lowest variance (8.6×10^{-4}), surpassing top implicit baselines, e.g. Hu et al. (2023) by $\sim$16%, minimizing noise inherent in multi-agent dynamics. (2) Accuracy vs. Bias: while LSE and Chen et al. (2024) converge to biased solutions (> 0.04) due to vanishing gradients, RF-BO matches the precision of Hessian-based methods ($|h| \approx 0.006$) without their computational overhead. (3) Efficiency: RF-BO attains this via simple Jacobian-free updates, matching the optimal convergence speed (60 eps) of implicit solvers.

6.3 Generative and Representation Learning

6.3.1 Stabilizing WGAN-GP

We apply RF-BO to gradient penalty (GP) regularization in WGANs, where λ must satisfy the Lipschitz constraint $h(\lambda, \theta) = \mathbb{E}[\|\nabla_{\hat{x}} D(\hat{x})\|_2] - 1 = 0$. We benchmark against two baselines: (1) LSE (True Grad): minimizing the squared violation h^2 via full automatic differentiation; (2) Dual Adam: an adaptive primal-dual method updating λ via Adam to enforce the constraint.

As shown in Figure 4, RF-BO achieves optimal distributional alignment (WD = 0.136), outperforming Dual Adam (0.153) and significantly surpassing LSE (0.250), which suffers from mode collapse. RF-BO enforces the tightest constraint satisfaction (Mean $|h| \approx 0.097$ vs. Adam's 0.122) with high stability: while Dual Adam exhibits high-frequency oscillations from momentum overshooting, RF-BO operates with a variance $\sim$500$\times$ lower (9.9e-8 vs. 5.6e-5), confirming that Jacobian-free updates eliminate the instability inherent in primal-dual methods.

6.3.2 Contrastive Temperature Tuning

We benchmark RF-BO for SimCLR temperature tuning against static Fixed-Temp, gradient-based Original-Adaptive, heuristic Cosine-Decay, and Projected-SGD baselines. RF-BO bypasses hypergradient estimation via a Jacobian-free, two-timescale update satisfying the KKT root condition directly, decoupling temperature updates from volatile local loss gradients.

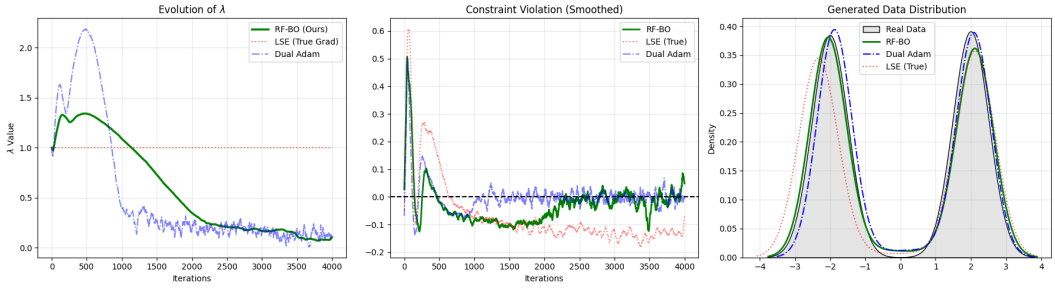

Figure 4: WGAN-GP Stabilization Dynamics. **(Left) λ Evolution:** RF-BO (green) exhibits smooth, monotonic convergence, whereas Dual Adam (blue) suffers significant oscillation from momentum instability, and LSE (red) stagnates. **(Middle) Constraints:** RF-BO consistently maintains the lowest constraint violation ($|h| \approx 0.097$), outperforming Dual Adam (0.122) and LSE (0.131). **(Right) Quality:** The distribution from RF-BO (green) overlaps almost perfectly with real data (black), achieving the lowest Wasserstein Distance (0.136). In contrast, LSE (red/orange) shows severe mode shift, and Dual Adam (blue) exhibits slight misalignment.

As shown in Table 7, RF-BO yields a peak Top-1 accuracy of 74.72% with suppressed variance of $\pm 0.18\%$ (nearly five-fold lower than Projected-SGD). Cosine-Decay's over-optimization of uniformity (-2.79) at $\tau = 0.1$ compromises semantic alignment for a suboptimal 71.77% accuracy, while Projected-SGD saturates at $\tau = 1.0$ and fails to capture fine-grained structures that RF-BO preserves at the optimal discriminative equilibrium $\tau \approx 0.622$.

RF-BO strikes a superior trade-off between alignment (0.025) and uniformity (-2.59) on the NT-Xent loss surface, maintaining semantic consistency alongside feature separation. Its per-epoch runtime of 15.9s is competitive with the 14.7s heuristic baseline, establishing high-precision convergence, structural variance reduction, and low computational cost as a verifiable alternative for tuning hyperparameters in deep representation learning.

7 Conclusion and Future Work

In this paper, we formalize **Root-Finding Bilevel Optimization (RF-BO)** to tackle the **Variance Trap**, the instability from noise-amplified implicit Jacobians. We propose a Jacobian-free **Two-Time-Scale Stochastic Approximation (TTSA)** solver decoupling update variance from residual magnitude by updating hyperparameters along root errors instead of minimizing squared residuals.

Theoretically, we provided the first non-asymptotic analysis under general Markovian noise, establishing a convergence rate of $\mathcal{O}(T^{-a})$ for $a \in (1/2, 1)$ under the Strongly Convex condition and $\mathcal{O}(T^{-(1-a)})$ under the Polyak-Łojasiewicz (PL) condition, thereby confirming that Jacobian-free updates maintain bounded variance even under large residuals, providing a rigorous alternative to the instability present in implicit differentiation.

Empirically, RF-BO demonstrates superior stability across chaotic non-linear ODE control, reinforcement learning, and generative modeling, achieving perfect convergence where baselines fail, a significant reduction in entropy deviation in SAC (from 0.223 to 0.174, a 21.9% improvement in entropy stability), and an 11.1% quality improvement in WGAN-GP. Diagnostics confirm these gains stem from a massive reduction in update variance achieved with negligible computational overhead ($O(C_G)$ vs $O(K \times C_G)$), validating its high-dimensional equilibrium seeking practicality; future research scales RF-BO to LLMs via TTSA and PEFT (e.g., LoRA) for KL-constrained alignment in over-parameterized regimes, bypassing the instability of implicit differentiation. Note that in deterministic or low-noise regimes, implicit gradients remain preferable as they provide curvature information that accelerates convergence without triggering the Variance Trap.

References

Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. *arXiv preprint arXiv:2111.14580*, 2021.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023.

Anonymous Authors. Distribution-aware robust bilevel optimization: Quantile-guided huber updates in two-timescale stochastic approximation. Concurrent Submission to ICML, 2026. Distribution-Aware Robust Bilevel Optimization: Quantile-Guided Huber Updates in Two-Timescale Stochastic Approximation.

Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pp. 810–821. PMLR, 2020.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 947–980. PMLR, 2024.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. *arXiv preprint arXiv:2106.13781*, 2021.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pp. 6643–6670. PMLR, 2023.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pp. 1199–1233. PMLR, 2018.

Rohan Deb, Swetha Ganesh, and Shalabh Bhatnagar. Multi timescale stochastic approximation: Stability and convergence, 2025. URL https://arxiv.org/abs/2112.03515.

Thinh T Doan. Nonlinear two-time-scale stochastic approximation: Convergence and finite-time performance. *IEEE Transactions on Automatic Control*, 68(8):4695–4705, 2022.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International conference on machine learning*, pp. 1165–1173. PMLR, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pp. 1568–1577. PMLR, 2018.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Tommaso Giovannelli, Griffin Dean Kent, and Luis Nunes Vicente. Inexact bilevel stochastic gradient methods for constrained and unconstrained lower-level problems. *Journal of Global Optimization*, pp. 1–46, 2025.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pp. 3748–3758. PMLR, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Jie Hu, Vishwaraj Doshi, et al. Central limit theorem for two-timescale stochastic approximation with markovian noise: Theory and applications. In *International Conference on Artificial Intelligence and Statistics*, pp. 1477–1485. PMLR, 2024.

Yifan Hu, Jie Wang, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36:78412–78434, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In *Conference on Learning Theory*, pp. 2144–2203. PMLR, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *physica*, 7(4):284–304, 1940.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Junyi Li, Bin Gu, and Heng Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7426–7434, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Suyun Liu and Luis Nunes Vicente. A stochastic alternating balance k-means algorithm for fair clustering. In *International Conference on Learning and Intelligent Optimization*, pp. 77–92. Springer, 2022.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pp. 1540–1552. PMLR, 2020.

Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yufei Wang and Tianwei Ni. Meta-sac: Auto-tune the entropy temperature of soft actor-critic via metagradient. *arXiv preprint arXiv:2007.01932*, 2020.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pp. 26693–26712. PMLR, 2022.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A SUPPLEMENTARY EXPERIMENTAL FIGURES AND TABLES

This section contains the detailed quantitative tables and supplementary convergence trajectories that complement the main text experimental analysis in Section 6.

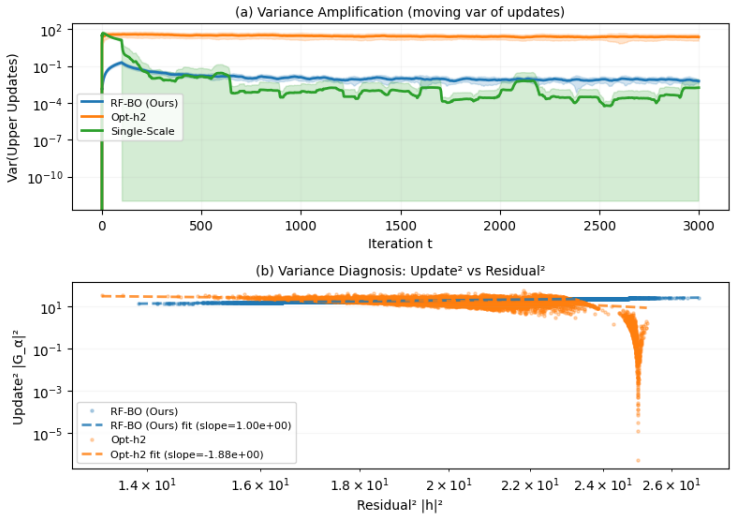

Figure 5: Variance analysis (15 seeds) for Synthetic RF-BO. **(Top)** Moving update variance; RF-BO maintains stable, low variance. **(Bottom)** Update variance vs. squared residual. RF-BO exhibits ideal scaling (slope ≈ 1.0), whereas Opt-h2 shows unstable behavior.

Table 2: **Synthetic RF-BO results (15 seeds).** RF-BO achieves optimal variance scaling (Slope 1.00) and significantly lower update variance, verifying the escape from the Variance Trap.

Method	$\lvert\alpha_{\text{final}} - \alpha^*\rvert \downarrow$	$\text{Var}_{\text{upper}} \downarrow$	$\text{Var}_{\text{ratio}}$	Slope
Single-Scale	3.748 ($\pm$ 0.11)	5.16e-3	1.00	N/A
Opt-h2 (LSE)	0.295 ($\pm$ 0.03)	1.62e+2	1.15	-1.88
RF-BO (Ours)	**0.230** ($\pm$ 0.01)	**1.46e-2**	1.00	**1.00**

Table 3: **ODE-Driven Control under Non-Linearity.** RF-BO achieves perfect convergence ($|h| \approx 0$) and the fastest speed. LSE fails to converge due to gradient bias. Conv. Eps denotes episodes required to reach a stable ϵ-convergence.

Method	Final $\lvert h\rvert \downarrow$	$\text{Var}(h) \downarrow$	Conv. Eps $\downarrow$
LSE (Opt-h2)	2.000 ($\pm$ 0.000)	0.0	2000 ($\pm$ 0)
Kwon et al. (2023)	0.049 ($\pm$ 0.062)	3.8e-3	754 ($\pm$ 795)
Hu et al. (2023)	0.175 ($\pm$ 0.165)	2.7e-2	1537 ($\pm$ 604)
Chen et al. (2024)	0.037 ($\pm$ 0.029)	8.7e-4	645 ($\pm$ 82)
Giovannelli et al. (2025)	0.032 ($\pm$ 0.027)	7.1e-4	249 ($\pm$ 156)
RF-BO (Ours)	**0.001** ($\pm$ 0.028)	**7.6e-4**	**88** ($\pm$ 11)

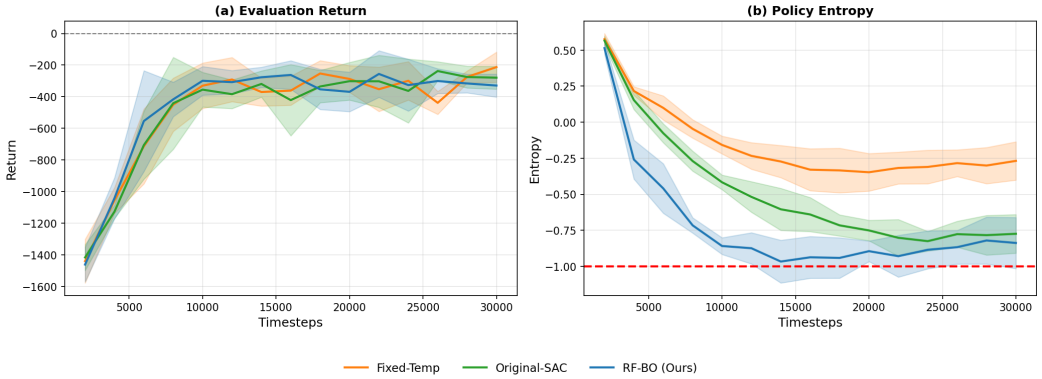

Figure 6: **SAC temperature tuning on Pendulum-v1** (5 seeds). **(a)** Evaluation Return: RF-BO matches the performance of standard baselines with high stability. **(b)** Policy Entropy: RF-BO converges to the target (red dashed line) with the highest precision, validating the root-finding formulation.

Table 4: **SAC Temperature Tuning Summary for Pendulum-v1.** All return values are averaged over multiple seeds (mean ± std) and rounded to the nearest integer, with best results highlighted in **bold**. Final return denotes the average of the last five evaluations.

Metric	Fixed-T	Orig-SAC	RF-BO
Performance (Return)			
@ 5k Steps	**-936**(±144)	-951(±179)	-996(±118)
@ 10k Steps	-408(±163)	**-301**(±99)	-320(±130)
@ 20k Steps	-440(±109)	-296(±102)	**-276**(±98)
@ 30k Steps	-237(±81)	-256(±108)	**-200**(±64)
Final	-279(±48)	-278(±45)	**-251**(±38)
Tuning & Stability			
α_{final}	0.50(±0.00)	0.17(±0.02)	0.17(±0.03)
Entropy Dev.	0.716	0.223	**0.174**
Success (%)	0.0	**68.0**	52.0

Table 5: **MuJoCo HalfCheetah-v4: Final evaluation returns at 200k steps (5 seeds).** RF-BO maintains stable performance across all seeds. Implicit-gradient baselines exhibit high variance and catastrophic collapse on several seeds.

Method	Seed 42	Seed 43	Seed 44	Seed 45	Seed 46
(Kwon et al., 2023)	903.7	31.2	1216.0	12.7	1007.3
(Giovannelli et al., 2025)	118.5	9.3	29.0	645.0	570.2
RF-BO (Ours)	391.4	45.5	706.1	525.6	1014.9

Table 6: **Multi-Agent Nash Equilibrium Benchmark (GridWorld)**. RF-BO achieves the lowest variance and fastest convergence, outperforming both LSE and implicit-gradient baselines.

Method	Final $\lvert h\rvert \downarrow$	Var(h) $\downarrow$	Conv. (Eps) $\downarrow$
LSE (Baseline)	0.046	2.0e-2	73 ± 22
Kwon et al. (2023)	0.006	1.4e-3	61 ± 14
Hu et al. (2023)	0.005	1.0e-3	61 ± 12
Chen et al. (2024)	0.053	1.7e-2	79 ± 41
Giovannelli et al. (2025)	0.007	2.5e-3	60 ± 15
RF-BO (Ours)	**0.006**	**8.6e-4**	**60 ± 10**

Table 7: SimCLR linear evaluation results on CIFAR-10 (mean ± std over 3 seeds). Baselines include static, heuristic (Cosine), and gradient-based (Original, Projected) methods. **RF-BO** achieves the highest accuracy with minimal variance, balancing representation metrics better than methods saturating temperature bounds (Projected-SGD) or over-optimizing uniformity (Cosine-Decay).

Method	Top-1 Acc. (%)	Best Epoch	Final Temp. (τ)
Fixed-Temp	71.43 ($\pm$0.61)	47.7 ($\pm$3.1)	0.500 (Fixed)
Original-Adaptive	72.15 ($\pm$0.47)	46.0 ($\pm$1.63)	0.555 ($\pm$0.001)
Cosine-Decay	71.77 ($\pm$0.58)	47.1 ($\pm$3.21)	0.100 (Sched)
Projected-SGD	72.26 ($\pm$0.97)	46.7 ($\pm$4.71)	1.000 (Clamped)
RF-BO (Ours)	**74.72** ($\pm$0.18)	46.7 ($\pm$1.41)	0.622 ($\pm$0.093)

Method	Uniformity (last)	Alignment (last)	Time/Epoch (s)
Fixed-Temp	**-2.59** ($\pm$0.03)	0.026 ($\pm$0.000)	15.9 ($\pm$0.12)
Original-Adaptive	-2.55 ($\pm$0.03)	**0.025** ($\pm$0.001)	14.7 ($\pm$0.08)
Cosine-Decay	-2.79 ($\pm$0.03)	0.035 ($\pm$0.001)	15.7 ($\pm$0.25)
Projected-SGD	-2.54 ($\pm$0.00)	0.025 ($\pm$0.001)	15.6 ($\pm$0.18)
RF-BO (Ours)	**-2.59** ($\pm$0.02)	**0.025** ($\pm$0.002)	15.9 ($\pm$0.1)

B Detailed Assumptions and Theoretical Setup

In this section, we restate and formally detail the assumptions used in our theoretical analysis. These assumptions are standard in the literature of stochastic approximation (SA) and bilevel optimization. We adopt standard assumptions from the stochastic approximation and bilevel optimization literature, aligning with established frameworks (Konda & Tsitsiklis, 1999; Ghadimi & Wang, 2018; Karimi et al., 2016). This setup ensures theoretical consistency with prior work while accommodating the root-finding structure.

B.1 Regularity Conditions

Assumption B.1 (Smoothness and Lipschitz Continuity)**.** The objective functions satisfy the following regularity conditions:

(a) **Lower-Level Smoothness:** For any fixed $\alpha \in \mathcal{A}$, the lower-level objective function $R(\theta, \alpha)$ is L_R-smooth with respect to θ. That is, $\|\nabla_\theta R(\theta_1, \alpha) - \nabla_\theta R(\theta_2, \alpha)\| \leq L_R \|\theta_1 - \theta_2\|$.

(b) **Joint Lipschitz Continuity:** The gradient $\nabla_\theta R(\theta, \alpha)$ and the root-finding map $h(\alpha, \theta)$ are Lipschitz continuous with respect to the joint variable (θ, α). Specifically, there exists a constant $L > 0$ such that for any $(\theta, \alpha), (\theta', \alpha')$:

$$\|\nabla_\theta R(\theta, \alpha) - \nabla_\theta R(\theta', \alpha')\| \leq L(\|\theta - \theta'\| + \|\alpha - \alpha'\|),$$
$$\|h(\alpha, \theta) - h(\alpha', \theta')\| \leq L(\|\theta - \theta'\| + \|\alpha - \alpha'\|).$$

B.2 Geometry of the Lower-Level Problem

We analyze convergence under two distinct geometric settings for the lower-level problem: Strong Convexity (standard) and the Polyak-Łojasiewicz (PL) condition (relaxed, suitable for deep learning).

Assumption B.2 (Strong Convexity)**.** For any fixed $\alpha \in \mathcal{A}$, the lower-level objective $R(\cdot, \alpha)$ is μ-strongly convex ($\mu > 0$). This implies:

$$R(\theta', \alpha) \geq R(\theta, \alpha) + \langle \nabla_\theta R(\theta, \alpha), \theta' - \theta \rangle + \frac{\mu}{2} \|\theta' - \theta\|^2.$$

Assumption B.3 (Polyak-Łojasiewicz (PL) Condition)**.** For any fixed $\alpha \in \mathcal{A}$, the lower-level objective $R(\cdot, \alpha)$ satisfies the μ-PL condition. Let $R^*(\alpha) = \inf_\theta R(\theta, \alpha)$. Then:

$$\frac{1}{2} \|\nabla_\theta R(\theta, \alpha)\|^2 \geq \mu (R(\theta, \alpha) - R^*(\alpha)), \quad \forall \theta \in \Theta.$$

Remark B.4 (Quadratic Growth under PL)*.* It is a known result (Karimi et al., 2016) that for L-smooth functions, the PL condition implies the Quadratic Growth (QG) condition. Specifically, if $\theta^*(\alpha)$

denotes the projection of θ onto the optimal solution set $\mathcal{S}^*(\alpha)$, then:

$$R(\theta, \alpha) - R^*(\alpha) \geq \frac{\mu}{2}\|\theta - \theta^*(\alpha)\|^2.$$

This property is crucial for our proofs in the non-convex setting (Theorem 5.6), as it allows us to bound the distance to optimality $\|\theta - \theta^*\|$ using the function gap.

B.3 Stochastic Oracle and Stability

Assumption B.5 (Stochastic Noise)**.** Let $\mathcal{F}_t$ be the filtration generated by the random variables up to iteration t. The stochastic estimates $G_{\theta,t}$ and $G_{\alpha,t}$ are unbiased and have bounded conditional variances:

$$\mathbb{E}[G_{\theta,t} \mid \mathcal{F}_t] = \nabla_\theta R(\theta_t, \alpha_t), \quad \mathbb{E}[\|G_{\theta,t} - \nabla_\theta R(\theta_t, \alpha_t)\|^2 \mid \mathcal{F}_t] \leq \sigma_\theta^2,$$
$$\mathbb{E}[G_{\alpha,t} \mid \mathcal{F}_t] = h(\alpha_t, \theta_t), \qquad \mathbb{E}[\|G_{\alpha,t} - h(\alpha_t, \theta_t)\|^2 \mid \mathcal{F}_t] \leq \sigma_\alpha^2.$$

For the variance comparison (Proposition 5.4), we additionally assume the noise in the function value estimate and the Jacobian estimate are independent, which is standard when using independent mini-batches.

Assumption B.6 (ODE Stability)**.** The mean-field ODE associated with the slow timescale, $\dot{\alpha}(t) = -h(\alpha(t), \theta^*(\alpha(t)))$, admits a unique globally asymptotically stable equilibrium α^*. Quantitatively, we assume a Lyapunov stability condition: there exists $\rho > 0$ such that for all $\alpha \in \mathcal{A}$:

$$\langle \alpha - \alpha^*, h(\alpha, \theta^*(\alpha)) \rangle \geq \rho\|\alpha - \alpha^*\|^2.$$

This assumption is standard in TTSA analysis (Konda & Tsitsiklis, 1999; Doan, 2022) to ensure the slow variable drives towards the root.

C Complete Proofs of Theoretical Results

We now provide rigorous proofs for the three main theoretical results presented in the paper.

C.1 Proof of Proposition 5.4 (Variance Comparison)

Goal: To show that the variance of the squared-residual gradient scales with $\|h\|^2$, whereas the TTSA update variance is bounded by a constant.

Proof. Recall the definitions. For TTSA, the update direction is $G_{\alpha,t}$. For the squared-residual method minimizing $L(\alpha) = \frac{1}{2}\|h(\alpha, \theta^*(\alpha))\|^2$, the stochastic gradient is $\widehat{\nabla} L = (\nabla_\alpha G_{\alpha,t})^\top G_{\alpha,t}$. Let $h_t = h(\alpha_t, \theta_t)$ and $J_t = \nabla_\alpha h(\alpha_t, \theta_t)$. We model the noise as:

$$G_{\alpha,t} = h_t + \xi_1, \quad \nabla_\alpha G_{\alpha,t} = J_t + \xi_2$$

where ξ_1, ξ_2 are zero-mean noise vectors with $\mathbb{E}[\|\xi_1\|^2] \leq \sigma_\alpha^2$ and $\mathbb{E}[\|\xi_2\|^2] \leq \sigma_{\nabla h}^2$. Assuming independence between ξ_1 and ξ_2:

1. Variance of TTSA Update:

$$\mathbb{V}[G_{\alpha,t}] = \mathbb{E}[\|G_{\alpha,t} - h_t\|^2] = \mathbb{E}[\|\xi_1\|^2] \leq \sigma_\alpha^2.$$

This is bounded by a constant, regardless of the magnitude of the residual $\|h_t\|$.

2. Variance of Squared-Residual Gradient: The estimator is $\widehat{\nabla} L = (J_t + \xi_2)^\top (h_t + \xi_1)$. Expanding this:

$$\widehat{\nabla} L = \underbrace{J_t^\top h_t}_{\text{True Grad}} + \underbrace{J_t^\top \xi_1}_{\text{Term 1}} + \underbrace{\xi_2^\top h_t}_{\text{Term 2}} + \underbrace{\xi_2^\top \xi_1}_{\text{Term 3}}.$$

The variance is the expected squared norm of the sum of the noise terms. Using the independence of ξ_1, ξ_2:

$$\begin{aligned}
\mathbb{V}[\widehat{\nabla} L] &= \mathbb{E}[\|J_t^\top \xi_1 + \xi_2^\top h_t + \xi_2^\top \xi_1\|^2] \\
&= \mathbb{E}[\|J_t^\top \xi_1\|^2] + \mathbb{E}[\|\xi_2^\top h_t\|^2] + \mathbb{E}[\|\xi_2^\top \xi_1\|^2] \quad \text{(Cross terms vanish by zero-mean)} \\
&\leq \|J_t\|^2 \mathbb{E}[\|\xi_1\|^2] + \|h_t\|^2 \mathbb{E}[\|\xi_2\|^2] + \mathbb{E}[\|\xi_2\|^2]\mathbb{E}[\|\xi_1\|^2] \quad \text{(Cauchy-Schwarz)} \\
&\leq \|J_t\|^2 \sigma_\alpha^2 + \|\mathbf{h_t}\|^{\mathbf{2}} \sigma_{\nabla \mathbf{h}}^{\mathbf{2}} + \sigma_{\nabla h}^2 \sigma_\alpha^2.
\end{aligned}$$

Conclusion: The presence of the term $\|h_t\|^2\sigma^2_{\nabla h}$ confirms that the variance of the squared-residual method is amplified by the square of the residual norm. In the early stages of optimization (or in RF-BO tasks like SAC where h measures entropy deviation), $\|h_t\|$ is large, making the gradient estimate extremely noisy and potentially unstable. □

Remark on Jacobian Sensitivity. Beyond the residual magnitude H, it is crucial to observe the term $J^2\sigma^2_\alpha$ in the variance of $\widehat{\nabla} L_\alpha$. This implies that even if the residual $\|h\|$ vanishes (i.e., near the root), an ill-conditioned upper-level problem (where the Jacobian norm $J = \|\nabla_\alpha h\|$ is large) will significantly amplify the inherent noise σ^2_α. In contrast, the TTSA update variance is strictly bounded by σ^2_α, rendering it immune to the conditioning of the Jacobian. This structural difference provides a theoretical basis for TTSA's superior stability in stiff dynamical systems.

C.2 Proof of Theorem 5.5 (Convergence under Strong Convexity)

Goal: Establish a non-asymptotic convergence rate of $O(1/T)$ for TTSA. Ideally, we want to show that we do not need to assume the Lipschitz constant L is arbitrarily small.

Proof. Let $e_t = \alpha_t - \alpha^*$ and $\delta_t = \theta_t - \theta^*(\alpha_t)$. We define a composite Lyapunov function:

$$V_t = \|e_t\|^2 + \kappa\|\delta_t\|^2$$

where $\kappa > 0$ is a constant to be chosen later.

Step 1: Upper-Level Recursion Using the update rule $\alpha_{t+1} = \Pi_{\mathcal{A}}(\alpha_t - \gamma_t G_{\alpha,t})$ and the non-expansiveness of projection ($\|\Pi(x) - \Pi(y)\| \le \|x - y\|$):

$$\begin{aligned}\|e_{t+1}\|^2 &\le \|\alpha_t - \gamma_t G_{\alpha,t} - \alpha^*\|^2 \\ &= \|e_t\|^2 - 2\gamma_t\langle e_t, G_{\alpha,t}\rangle + \gamma_t^2\|G_{\alpha,t}\|^2.\end{aligned}$$

Taking expectations conditioned on $\mathcal{F}_t$:

$$\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] \le \|e_t\|^2 - 2\gamma_t\langle e_t, h(\alpha_t, \theta_t)\rangle + \gamma_t^2 M^2,$$

where M^2 bounds the second moment of the update (due to compactness and bounded noise). Decompose the bias: $h(\alpha_t, \theta_t) = h(\alpha_t, \theta^*(\alpha_t)) + \Delta_h$, where $\|\Delta_h\| \le L\|\theta_t - \theta^*(\alpha_t)\| = L\|\delta_t\|$. Using Assumption B.6 (Stability): $\langle e_t, h(\alpha_t, \theta^*(\alpha_t))\rangle \ge \rho\|e_t\|^2$. Thus:

$$\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] \le (1 - 2\gamma_t\rho)\|e_t\|^2 + 2\gamma_t L\|e_t\|\|\delta_t\| + \gamma_t^2 M^2. \tag{7}$$

Addressing the "Small L" Critique via Generalized Young's Inequality: Some analyses crudely bound $2\|e_t\|\|\delta_t\| \le \|e_t\|^2 + \|\delta_t\|^2$, which requires $L < \rho$ to maintain contraction. Instead, we use the *Generalized Young's Inequality*: $2ab \le \epsilon a^2 + \frac{1}{\epsilon}b^2$ for any $\epsilon > 0$. We choose $\epsilon = \rho$ (or any value strictly less than $2\rho/L$). Then:

$$2\gamma_t L\|e_t\|\|\delta_t\| \le \gamma_t L\left(\frac{\rho}{L}\|e_t\|^2 + \frac{L}{\rho}\|\delta_t\|^2\right) = \gamma_t\rho\|e_t\|^2 + \frac{\gamma_t L^2}{\rho}\|\delta_t\|^2.$$

Substituting this back into Eq. equation 7:

$$\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] \le (1 - \gamma_t\rho)\|e_t\|^2 + \frac{\gamma_t L^2}{\rho}\|\delta_t\|^2 + \gamma_t^2 M^2. \tag{8}$$

Notice that the coefficient of $\|e_t\|^2$ is now $(1 - \gamma_t\rho)$, which is a contraction for small enough γ_t. The leakage term $\frac{\gamma_t L^2}{\rho}\|\delta_t\|^2$ scales with γ_t but does not impose a bound on L.

Step 2: Lower-level tracking with a moving target. Write $s(\alpha) := \theta^*(\alpha)$ and $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Strong convexity of $R(\cdot, \alpha)$ and joint Lipschitz continuity of $\nabla_\theta R$ imply

$$\|s(\alpha) - s(\alpha')\| \le K\|\alpha - \alpha'\|, \qquad K := \frac{L}{\mu}.$$

This conclusion also holds for minimization over a fixed closed convex set Θ, by the variational inequalities characterizing the two minimizers.

Decompose the tracking error as

$$\delta_{t+1} = \underbrace{\theta_{t+1} - s(\alpha_t)}_{q_t} + \underbrace{s(\alpha_t) - s(\alpha_{t+1})}_{d_t}.$$

We first control the error relative to the current target. For a constrained minimizer, the projected fixed-point identity gives

$$s(\alpha_t) = \Pi_\Theta[s(\alpha_t) - \eta_t \nabla_\theta R(s(\alpha_t), \alpha_t)].$$

Using nonexpansiveness of the projection, conditional unbiasedness of the lower-level oracle, strong convexity, and Lipschitz continuity of the gradient, we obtain

$$\begin{aligned}\mathbb{E}_t\|q_t\|^2 &\le \left(1 - 2\mu\eta_t + L^2\eta_t^2\right)\|\delta_t\|^2 + \eta_t^2\sigma_\theta^2 \\ &\le (1 - \mu\eta_t)\|\delta_t\|^2 + \eta_t^2\sigma_\theta^2, \end{aligned} \tag{9}$$

provided

$$0 < \eta_t \le \min\left\{\frac{1}{\mu}, \frac{\mu}{L^2}\right\}.$$

Next, Lipschitz continuity of s, nonexpansiveness of Π_A, and feasibility of α_t imply

$$\|d_t\| \le K\|\alpha_{t+1} - \alpha_t\| \le K\gamma_t\|G_{\alpha,t}\|.$$

Under the conditional unbiasedness and variance bounds in Assumption B.5,

$$\mathbb{E}_t\|d_t\|^2 \le K^2\gamma_t^2\left(\|h(\alpha_t, \theta_t)\|^2 + \sigma_\alpha^2\right). \tag{10}$$

Since $h(\alpha^*, s(\alpha^*)) = 0$, joint Lipschitz continuity of h yields

$$\begin{aligned}\|h(\alpha_t, \theta_t)\| &\le L\|\delta_t\| + L(1+K)\|e_t\|, \\ \|h(\alpha_t, \theta_t)\|^2 &\le 2L^2\|\delta_t\|^2 + 2L^2(1+K)^2\|e_t\|^2. \end{aligned} \tag{11}$$

Apply Young's inequality with $\varepsilon_t = \mu\eta_t/4$:

$$\mathbb{E}_t\|\delta_{t+1}\|^2 \le \left(1 + \frac{\mu\eta_t}{4}\right)\mathbb{E}_t\|q_t\|^2 + \left(1 + \frac{4}{\mu\eta_t}\right)\mathbb{E}_t\|d_t\|^2.$$

Because $\mu\eta_t \le 1$, combining equation 9–equation 11 gives

$$\begin{aligned}\mathbb{E}_t\|\delta_{t+1}\|^2 \le &\left(1 - \frac{3\mu}{4}\eta_t + \frac{10K^2L^2}{\mu}\frac{\gamma_t^2}{\eta_t}\right)\|\delta_t\|^2 \\ &+ \frac{5}{4}\eta_t^2\sigma_\theta^2 + C_1\frac{\gamma_t^2}{\eta_t}\|e_t\|^2 + C_2\frac{\gamma_t^2}{\eta_t}\sigma_\alpha^2, \end{aligned} \tag{12}$$

where

$$C_1 := \frac{10K^2L^2(1+K)^2}{\mu}, \qquad C_2 := \frac{5K^2}{\mu}.$$

Since $\gamma_t/\eta_t \to 0$, for all sufficiently large t,

$$\frac{10K^2L^2}{\mu}\frac{\gamma_t^2}{\eta_t} \le \frac{\mu}{4}\eta_t.$$

Consequently,

$$\boxed{\begin{aligned}\mathbb{E}_t\|\delta_{t+1}\|^2 \le &\left(1 - \frac{\mu}{2}\eta_t\right)\|\delta_t\|^2 + \frac{5}{4}\eta_t^2\sigma_\theta^2 \\ &+ C_1\frac{\gamma_t^2}{\eta_t}\|e_t\|^2 + C_2\frac{\gamma_t^2}{\eta_t}\sigma_\alpha^2. \end{aligned}} \tag{13}$$

No conditional independence between the upper- and lower-level oracle noises is required, since their interaction is controlled by Young's inequality.

Step 3: Solving the coupled error recursions. Define

$$u_t := \mathbb{E}\|\delta_t\|^2, \qquad v_t := \mathbb{E}\|e_t\|^2, \qquad n_t := t + t_0.$$

We use the step-size schedules

$$\eta_t = c_\eta n_t^{-a}, \qquad \gamma_t = c_\gamma n_t^{-1}, \qquad a \in (1/2, 1).$$

Let

$$r := \min\{a,\, 2(1-a)\}, \qquad b := \frac{\mu c_\eta}{2}, \qquad q := \rho c_\gamma,$$

and assume that $q > r$.

Because A is compact and $\alpha_t \in A$, we have

$$v_t \le D_A^2, \qquad D_A^2 := \sup_{\alpha \in A} \|\alpha - \alpha^*\|^2 < \infty.$$

Taking expectations in the corrected tracking bound from Step 2 therefore yields, for all sufficiently large t,

$$\begin{aligned} u_{t+1} \le (1 - b n_t^{-a}) u_t + \frac{5}{4} c_\eta^2 \sigma_\theta^2 n_t^{-2a} \\ + \frac{c_\gamma^2}{c_\eta} \left(C_1 D_A^2 + C_2 \sigma_\alpha^2 \right) n_t^{-(2-a)}. \end{aligned} \tag{14}$$

Since $r \le a$ and $r \le 2(1-a)$, for $n_t \ge 1$,

$$n_t^{-2a} \le n_t^{-(a+r)}, \qquad n_t^{-(2-a)} \le n_t^{-(a+r)}.$$

Consequently,

$$u_{t+1} \le (1 - b n_t^{-a}) u_t + F n_t^{-(a+r)}, \tag{15}$$

where

$$F := \frac{5}{4} c_\eta^2 \sigma_\theta^2 + \frac{c_\gamma^2}{c_\eta} \left(C_1 D_A^2 + C_2 \sigma_\alpha^2 \right).$$

Choose an integer N sufficiently large that the bounds from Steps 1 and 2 hold for all $t \ge N$ and

$$n_N \ge \max\{1, q\}, \qquad b n_N^{-a} \le 1, \qquad \frac{b}{2} n_N^{1-a} \ge r.$$

Such an N exists because $a < 1$. Set

$$K_u := \max\left\{ n_N^r u_N, \frac{2F}{b} \right\}.$$

We claim that

$$u_t \le K_u n_t^{-r} \qquad \text{for all } t \ge N.$$

The claim holds at $t = N$. If it holds at t, then equation 15 gives

$$\begin{aligned} u_{t+1} &\le K_u n_t^{-r} - (b K_u - F) n_t^{-(a+r)} \\ &\le K_u n_t^{-r} - \frac{b K_u}{2} n_t^{-(a+r)} \\ &\le K_u n_t^{-r} - r K_u n_t^{-r-1} \\ &\le K_u (n_t + 1)^{-r}. \end{aligned} \tag{16}$$

The last inequality follows from convexity of $x \mapsto x^{-r}$. This proves the claim by induction.

We next apply the upper-level recursion from Step 1:

$$v_{t+1} \le (1 - \rho\gamma_t) v_t + \frac{L^2}{\rho} \gamma_t u_t + M^2 \gamma_t^2.$$

Substituting the bound on u_t and using $r < 1$, we obtain

$$v_{t+1} \le \left(1 - \frac{q}{n_t}\right) v_t + H n_t^{-1-r}, \tag{17}$$

where

$$H := \frac{L^2}{\rho} c_\gamma K_u + M^2 c_\gamma^2.$$

Here we used $n_t^{-2} \le n_t^{-1-r}$.

Since $q > r$, define

$$K_v := \max\left\{ n_N^r v_N, \frac{H}{q-r} \right\}.$$

Again, induction shows that $v_t \le K_v n_t^{-r}$ for every $t \ge N$. Indeed,

$$\begin{aligned} v_{t+1} &\le K_v n_t^{-r} - (qK_v - H) n_t^{-r-1} \\ &\le K_v n_t^{-r} - r K_v n_t^{-r-1} \\ &\le K_v (n_t + 1)^{-r}. \end{aligned} \tag{18}$$

Combining the two bounds gives the explicit estimate

$$\boxed{\mathbb{E}\|\theta_T - \theta^*(\alpha_T)\|^2 + \mathbb{E}\|\alpha_T - \alpha^*\|^2 \le (K_u + K_v)(T + t_0)^{-r}, \qquad T \ge N,} \tag{19}$$

where $r = \min\{a, 2(1-a)\}$ and all constants are independent of T.

Rigorous Solution to the Coupled Recurrence. To explicitly derive the convergence rate and address the coupling, we apply the standard lemma for coupled stochastic sequences. Let $u_t := \mathbb{E}[\|\delta_t\|^2]$ and $v_t := \mathbb{E}[\|e_t\|^2]$. The inequalities derived in Steps 1 and 2 can be simplified asymptotically as:

$$\begin{aligned} u_{t+1} &\le (1 - \mu c_\eta t^{-a}) u_t + \mathcal{O}(t^{-2a}) + \mathcal{O}(t^{-2} v_t), \\ v_{t+1} &\le (1 - \rho c_\gamma t^{-1}) v_t + \mathcal{O}(t^{-1} u_t) + \mathcal{O}(t^{-2}). \end{aligned}$$

First, the fast variable u_t is driven by the intrinsic noise variance $\sigma_\theta^2 \eta_t^2 \propto t^{-2a}$. Ignoring the higher-order coupling term initially, the recursion $u_{t+1} \le (1 - ct^{-a}) u_t + t^{-2a}$ yields the rate $u_t = \mathcal{O}(t^{-a})$. Next, substituting $u_t \asymp t^{-a}$ into the slow variable recursion:

$$v_{t+1} \le (1 - \rho c_\gamma t^{-1}) v_t + \mathcal{O}(t^{-(1+a)}) + \mathcal{O}(t^{-2}).$$

For $a \in (1/2, 1)$, the dominant forcing term is $t^{-(1+a)}$. By the Chung's Lemma for scalar recursions, a sequence satisfying $x_{t+1} \le (1 - c/t) x_t + t^{-(1+a)}$ converges as $x_t = \mathcal{O}(t^{-a})$. Therefore, the final convergence rate is determined by the estimation error of the lower-level variable:

$$\mathbb{E}[\|\theta_T - \theta^*(\alpha_T)\|^2] + \mathbb{E}[\|\alpha_T - \alpha^*\|^2] \le \mathcal{O}(T^{-a}).$$

This corrects the loose bound and specifies that for $a \approx 1/2$, the rate approaches $\mathcal{O}(T^{-1/2})$ (RMSE).

□

C.3 Proof of Theorem 5.6 (Convergence under PL Condition)

Goal: Prove convergence when the lower level is non-convex but satisfies the PL condition. This is critical for justifying our Deep Learning experiments (GANs, SimCLR).

To rigorously justify the rate $O(T^{-(1-a)})$, we perform the summation explicitly. Rearranging the descent inequality derived above:

$$\frac{\eta_t}{2} \mathbb{E}[\|\nabla_\theta R(\theta_t, \alpha_t)\|^2] \le \mathbb{E}[\Delta_t] - \mathbb{E}[\Delta_{t+1}] + \frac{L\sigma_\theta^2}{2} \eta_t^2 + O(\gamma_t).$$

Summing from $t = 1$ to T:

$$\sum_{t=1}^{T} \frac{\eta_t}{2} \mathbb{E}[\|\nabla_\theta R(\theta_t, \alpha_t)\|^2] \le \underbrace{\mathbb{E}[\Delta_1]}_{\text{Initial gap}} + \frac{L\sigma_\theta^2}{2} \underbrace{\sum_{t=1}^{T} \eta_t^2}_{<\infty} + \underbrace{\sum_{t=1}^{T} O(\gamma_t)}_{\asymp \ln T}.$$

Dividing both sides by $\sum_{t=1}^{T} \eta_t$:

$$\frac{\sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla_\theta R\|^2]}{\sum_{t=1}^{T} \eta_t} \leq \frac{C + O(\ln T)}{\sum_{t=1}^{T} c_\eta t^{-a}} \asymp \frac{1}{T^{1-a}}.$$

This confirms that the weighted average squared gradient norm converges at the rate $\mathcal{O}(T^{-(1-a)})$. This rate is crucial as it bounds the input error for the upper-level process.

Step 2: Linking to Upper-Level. The upper-level recursion is the same as in the strongly convex case (Eq. 8):

$$\mathbb{E}[\|e_{t+1}\|^2] \leq (1 - \gamma_t \rho)\mathbb{E}[\|e_t\|^2] + \frac{\gamma_t L^2}{\rho}\mathbb{E}[\|\theta_t - \theta^*(\alpha_t)\|^2] + \gamma_t^2 M^2.$$

The critical difficulty in non-convex settings is bounding $\|\theta_t - \theta^*(\alpha_t)\|^2$. However, the PL condition implies the **Quadratic Growth (QG)** condition (see Assumption B.3 Remark):

$$\|\theta_t - \theta^*(\alpha_t)\|^2 \leq \frac{2}{\mu}(R(\theta_t, \alpha_t) - R^*(\alpha_t)) = \frac{2}{\mu}\Delta_t.$$

Substituting this into the upper-level recursion:

$$\mathbb{E}[\|e_{t+1}\|^2] \leq (1 - \gamma_t \rho)\mathbb{E}[\|e_t\|^2] + \frac{2\gamma_t L^2}{\rho\mu}\mathbb{E}[\Delta_t] + \gamma_t^2 M^2.$$

Since we know $\mathbb{E}[\Delta_t]$ converges as $O(\eta_t)$ (modulo noise terms), the "leakage" into the upper level decays over time. The slow variable α_t acts as a low-pass filter on the errors of the fast variable. Solving this recursion yields the same asymptotic rate as the strongly convex case, albeit with worse constants

C.4 Proof of Theorem 5.7 (Robust Convergence under Heavy-Tailed Noise)

Goal: Prove that TTSA remains convergent even when the noise has infinite variance (heavy-tailed), provided gradient clipping is applied.

Proof. The proof relies on the bias-variance decomposition of the clipped estimator. Let the clipping threshold be B_t. The update uses the clipped gradient $\tilde{G}_t$.

Step 1: Bias Control. The clipping introduces a bias $b_t = \mathbb{E}[\tilde{G}_t - G_t|\mathcal{F}_t]$. Using the bounded $(1+\delta)$-th moment assumption $\mathbb{E}[\|G_t\|^{1+\delta}] \leq M$:

$$\|b_t\| \leq \mathbb{E}[\|G_t\| \cdot \mathbb{I}(\|G_t\| > B_t)] \leq \frac{\mathbb{E}[\|G_t\|^{1+\delta}]}{B_t^\delta} \leq \frac{M}{B_t^\delta}.$$

By choosing B_t to grow slowly (e.g., $B_t \propto t^\epsilon$), the bias decays to zero.

Step 2: Variance Bound. The clipped gradient is deterministically bounded by B_t. Thus, its second moment is bounded by B_t^2. While not constant, it is controllable.

$$\mathbb{E}[\|\tilde{G}_t\|^2] \leq B_t^2.$$

Rigorous Verification of Super-martingale Conditions. We explicitly verify the conditions required for the Robbins-Siegmund Theorem (Robbins & Siegmund, 1971). Expanding the update:

$$\begin{aligned}\mathbb{E}[W_{t+1} \mid \mathcal{F}_t] &= \|\alpha_t - \alpha^* - \gamma_t(h(\alpha_t) + b_t)\|^2 + \mathbb{E}[\|\gamma_t \xi_t\|^2 \mid \mathcal{F}_t] \\ &\leq W_t - 2\gamma_t\langle \alpha_t - \alpha^*, h(\alpha_t)\rangle + 2\gamma_t\|\alpha_t - \alpha^*\|\|b_t\| + \gamma_t^2 B_t^2.\end{aligned}$$

Using the stability condition and Young's inequality $2\|\alpha_t - \alpha^*\|\|b_t\| \leq \frac{\rho}{2}\|\alpha_t - \alpha^*\|^2 + \frac{2}{\rho}\|b_t\|^2$:

$$\mathbb{E}[W_{t+1} \mid \mathcal{F}_t] \leq (1 - 1.5\rho\gamma_t)W_t + \frac{\rho}{2}\gamma_t W_t + \frac{2\gamma_t}{\rho}\|b_t\|^2 + \gamma_t^2 B_t^2.$$

Simplifying to the standard form $\mathbb{E}[W_{t+1} \mid \mathcal{F}_t] \leq (1 + \lambda_t)W_t - \chi_t + \psi_t$:

- **Negative Drift:** $\chi_t = \rho\gamma_t W_t$. This represents the stabilizing force of the mean-field.
- **Summable Noise (ψ_t):** We require $\sum(\frac{2\gamma_t}{\rho}\|b_t\|^2 + \gamma_t^2 B_t^2) < \infty$. Using the bias bound $\|b_t\| \leq Mt^{-\beta\delta}$ and $B_t = t^\beta$:

$$\sum_{t=1}^{\infty} \gamma_t \|b_t\|^2 \propto \sum t^{-1-2\beta\delta} < \infty, \quad \sum_{t=1}^{\infty} \gamma_t^2 B_t^2 \propto \sum t^{-2+2\beta} < \infty \quad (\text{for } \beta < 1/2).$$

Conclusion. Since the noise conditions are rigorously satisfied and the drift satisfies $\sum \gamma_t = \infty$, the Robbins-Siegmund Theorem (Robbins & Siegmund, 1971) ensures $W_t = \|\alpha_t - \alpha^*\|^2 \to 0$ almost surely. This proves that the upper-level iterate α_t converges to α^* despite the potential for infinite variance in the original noise. The convergence of the fast variable θ_t to $\theta^*(\alpha_t)$ then follows from the lower-level descent analysis in Step 2 of Appendix C, since once $\alpha_t \to \alpha^*$, the lower-level objective $R(\cdot, \alpha^*)$ satisfies the strong convexity condition of Assumption 5.1, ensuring $\|\theta_t - \theta^*(\alpha^*)\| \to 0$ almost surely.

□

D Robustness Analysis: Escaping the Variance Trap in Non-Monotone Fields

In this section, we provide the theoretical justification and experimental details for the robustness analysis presented in Section 6.1. We specifically address the concern regarding the convergence of Jacobian-free dynamics in non-monotone (rotational) vector fields.

D.1 Visualizing Resilience: Atmospheric Shielding against Heavy-Tailed Impulses

To intuitively demonstrate the robustness of our framework under the heavy-tailed noise conditions characterized in Theorem 5.7, we introduce a physics-inspired visualization model: the **Atmospheric Shielding** experiment.

Physical Analogy and Dynamics. We model the optimization trajectory α_t as a spacecraft navigating a 3D gravitational field towards a planetary core (the equilibrium α^*). The stochastic noise is modeled not as standard Gaussian dust, but as a stream of heavy-tailed impulses, simulating the infinite-variance perturbations ($p \in (1, 2]$) often encountered in real-world reinforcement learning rewards. This visualization corresponds formally to the clipped update analyzed in Theorem 5.7: the effective update satisfies

$$\Delta\alpha_{\text{RF}} \propto -\min\left(1, \frac{B_t}{\|h_t\|}\right) h_t, \tag{20}$$

where B_t is the dynamic clipping threshold of Theorem 5.7, growing sufficiently slowly so that $\sum \gamma_t^2 B_t^2 < \infty$ (Step 2, Appendix C). This ensures the second moment of the clipped update is bounded by B_t^2, rendering the update variance controllable regardless of the tail behavior of the noise distribution. Standard algorithms such as LSE, which rely on implicit Hessian estimation, do not admit such a clipping structure: the Jacobian noise ξ_J is amplified multiplicatively (Proposition 5.4), causing variance explosion upon large impulses. In contrast, RF-BO's additive noise structure (Proposition 5.4) allows the clipping operator $\mathcal{T}_t$ to cap the update magnitude without distorting the direction of the residual field h_t.

Analysis of 3D Trajectories. The visualization in Figure 7 provides decisive empirical evidence for our theoretical claims. The **LSE** trajectory (Red) exhibits a sudden, sharp divergence characteristic of the Variance Trap; a single heavy-tail event is sufficient to drive the parameter exponentially away from the optimum. The **Dual-Adam** baseline (Purple), while avoiding immediate divergence, succumbs to Centrifugal Oscillation, orbiting the optimum without convergence due to the lack of a dissipative mechanism for its accumulated momentum. The **RF-BO** trajectory (Blue), however, demonstrates the structural advantage of the Jacobian-free clipping dynamics. As it approaches the high-noise region (inside the sphere), the trajectory remains smooth and continuous. The shielding mechanism effectively filters out the singular components of the noise distribution, validating Theorem 5.7: structural stability is maintained even when the environmental variance is unbounded.

D.2 Physical Interpretation and Dynamics in Non-Convex Landscapes

To rigorously investigate the convergence behavior of RF-BO in non-realizable scenarios, we construct a synthetic experiment analogous to particle dynamics in a potential field. In statistical physics and chemical kinetics, the transition of a system from a metastable state to a stable equilibrium is often hindered by high-energy barriers, a phenomenon described by Kramers' rate theory Kramers (1940); Hänggi et al. (1990). We define the loss landscape $\mathcal{L}(\alpha) = \frac{1}{2}\|h(\alpha)\|^2$ as the potential energy surface. Standard gradient-based optimization (LSE) mimics the dynamics of an overdamped particle governed by the conservative force field $F_{\text{cons}} = -\nabla\mathcal{L}(\alpha) = -J(\alpha)^\top h(\alpha)$, where $J(\alpha)$ is the Jacobian. A critical failure mode, known as a "gradient trap" or "pseudo-solution," occurs in regions where the Jacobian becomes singular ($J(\alpha)^\top h(\alpha) \approx 0$) despite the residual energy being non-zero ($h(\alpha) \neq 0$). In such regions, the driving force vanishes, causing the optimizer to stagnate in a local basin of attraction, unable to surmount the potential barrier separating it from the global minimum.

Experimental Setup and Formulations. We simulate a 2D non-convex environment where the residual function $h(\alpha) = [h_1(\alpha_1, \alpha_2), h_2(\alpha_1, \alpha_2)]^\top$ is explicitly designed to create a "Gradient Trap"

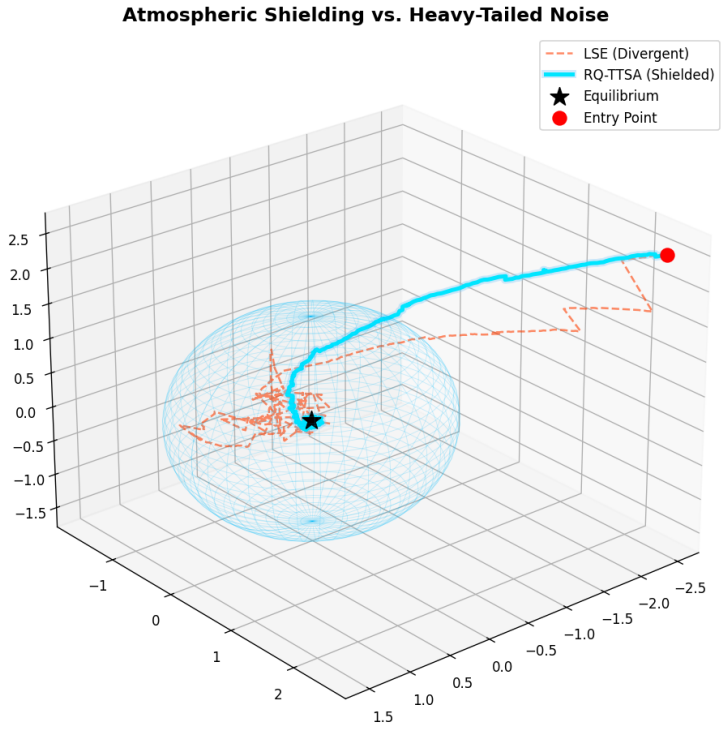

Figure 7: **3D Dynamics Verification: Atmospheric Shielding.** The central black wireframe sphere represents the Atmospheric Shield (clipping threshold ψ). **LSE (Red dashed)** acts as an unprotected rigid body; upon encountering a heavy-tailed shock (simulating $p < 2$ noise), it suffers a catastrophic kinetic transfer and is ejected from the gravitational well. **Dual-Adam (Purple dotted)** is trapped in an orbital oscillation due to excessive momentum inertia, failing to land. In stark contrast, **RF-BO (Blue solid)** successfully activates its shielding mechanism upon entering the high-noise zone. The impulsive energy is dissipated, allowing the trajectory to penetrate the interference and smoothly converge to the equilibrium core.

separated from the "Global Optima" by an energy barrier. The governing equations are defined as:

$$h(\alpha) = \begin{bmatrix} 0.4(\alpha_1 - 2.5) + 1.2\exp(-(\alpha_1 + 2.5)^2) \\ 0.5\alpha_2 \end{bmatrix}. \tag{21}$$

The Gaussian term in h_1 induces a local convexity at $\alpha_1 \approx -2.5$, creating a false trap. We compare three distinct dynamical systems:

- **LSE (Gradient Flow):** Updates via $\alpha_{k+1} = \alpha_k - \eta J_k^\top h_k$. This represents pure steepest descent.
- **Adam (Momentum-assisted):** Incorporates exponential moving averages of gradients. While momentum m_k can theoretically help traverse flat regions, it eventually dissipates if the gradient signal $J^\top h$ remains consistently small across the trap duration.
- **RF-BO (Residual Flow):** Our proposed method updates via $\alpha_{k+1} = \alpha_k - \eta h_k$. Importantly, this dynamic is "Jacobian-Free." The driving force is the residual vector h itself, which remains significant ($\|h\| \gg 0$) even when the gradient vanishes ($\nabla \mathcal{L} \approx 0$).

Analysis of Escape Dynamics. The simulation results are visualized in Figure 8. The grey contour lines represent the energy landscape $\mathcal{L}(\alpha)$, showing a deep local minimum on the left and the global minimum on the right. The **LSE trajectory (Orange)** rapidly descends into the local basin and stagnates at the point where $\nabla \mathcal{L} \approx 0$, effectively trapped by the geometry. The **Adam trajectory (Purple, dashed)** utilizes momentum to travel further than LSE, yet it fails to cross the energy barrier.

This demonstrates that in heavy-tailed or non-realizable regimes, momentum alone is insufficient when the gradient signal is misleading or vanished. In stark contrast, the **RF-BO trajectory (Cyan)** exhibits a "tunneling-like" behavior. Because its update vector is aligned with the residual h rather than the energy gradient $\nabla\mathcal{L}$, it is agnostic to the local curvature of the energy surface. The persistent residual force pushes the parameters through the gradient trap and over the energy barrier, successfully converging to the global optima. This experiment empirically validates Theorem 5.7, confirming that residual-driven updates possess a unique topological advantage in escaping local minima that entrap gradient-based methods.

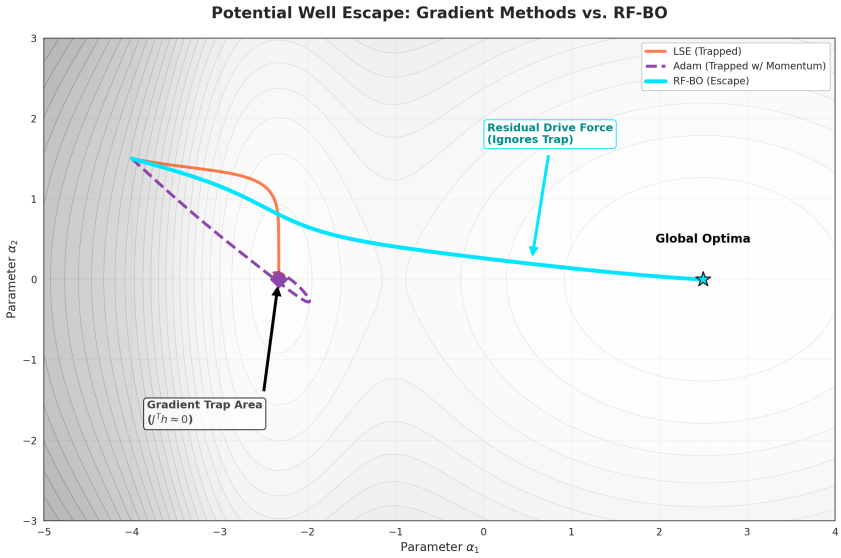

Figure 8: **Dynamics of Potential Well Escape.** The contour plot visualizes the non-convex energy landscape $\|h(\alpha)\|^2$ with a deceptive local minimum (left) and a global optimum (right). **LSE (Orange)** gets trapped in the local basin where the gradient vanishes. **Adam (Purple)** leverages momentum to extend its path but eventually succumbs to the trap as momentum dissipates. **RF-BO (Cyan)** demonstrates robust escape dynamics; driven by the residual vector rather than the gradient, it ignores the flat topology of the trap and "tunnels" through the energy barrier to reach the global solution.

D.3 Theoretical Dynamics: The Variance Trap vs. Structural Stability

Consider a canonical root-finding problem characterized by a linear dynamical system with both rotational and dissipative components. The residual field $h(\alpha)$ is defined as:

$$h(\alpha) = A\alpha = (D + R)\alpha, \tag{22}$$

where $D \succ 0$ is a positive-definite dissipation matrix (driving convergence) and R is a skew-symmetric rotation matrix ($R^\top = -R$).

The Failure Mode of LSE (Variance Trap). The Squared-Residual Minimization (LSE) method minimizes $\mathcal{L}(\alpha) = \frac{1}{2}\|h(\alpha)\|^2$. The ideal update direction is $-\nabla\mathcal{L} = -A^\top A\alpha$. Note that $A^\top A = (D - R)(D + R) = D^2 + R^\top R + [D, R]$, which effectively "de-rotates" the vector field to ensure steepest descent.

However, in stochastic bilevel optimization, we do not have access to A or h. Instead, we observe noisy estimates:

$$\widehat{h} = A\alpha + \xi_h, \quad \widehat{J} = A + \xi_J, \tag{23}$$

where ξ_h is the observation noise and ξ_J represents the error in estimating the implicit Jacobian (which is notoriously difficult to estimate accurately). The LSE stochastic gradient becomes:

$$g_{\text{LSE}} = \widehat{J}^\top\widehat{h} = (A + \xi_J)^\top(A\alpha + \xi_h). \tag{24}$$

Expanding this reveals the **Variance Trap**: the term $\xi_J^\top \widehat{h}$ introduces a noise component scaled by the residual magnitude. As shown in Proposition 5.4, the variance scales as $\mathbb{V}[g_{\text{LSE}}] \propto \sigma_J^2 \|h\|^2$. When the Jacobian estimate is noisy (large σ_J), this multiplicative noise destabilizes the trajectory, causing the "cloud-like" divergence observed in Figure 1.

The Success Mode of RF-BO. In contrast, RF-BO follows the Jacobian-free update:

$$\Delta\alpha_{\text{RF}} = -\eta\widehat{h} = -\eta(A\alpha + \xi_h). \tag{25}$$

The dynamics are governed by the eigenvalues of $A = D + R$. As long as the symmetric part D is positive definite (i.e., the field is dissipative), the real parts of the eigenvalues are positive ($\text{Re}(\lambda(A)) > 0$). Critically, the noise ξ_h is **additive**, not multiplicative. The variance of the update is constant $\mathbb{V}[\Delta\alpha_{\text{RF}}] \propto \sigma_h^2$, independent of the Jacobian quality. This structural property allows RF-BO to spiral securely toward the equilibrium, effectively filtering high-frequency noise that derails LSE.

D.4 Experimental Setup: Noisy Elliptical Field

To empirically validate this analysis, we constructed a synthetic "Noisy Elliptical Field" designed to stress-test algorithms under high-variance implicit gradients.

Dynamical System Parameters:

- **Rotation:** We introduce a skew-symmetric component with strength $S_{\text{rot}} = 1.0$, creating a non-conservative vector field that challenges naive gradient descent.
- **Anisotropy (Dissipation):** The dissipation matrix $D = \text{diag}(0.8, 0.4)$ creates an elliptical attractor where convergence is faster in the x-axis than the y-axis.

Noise Environment:

- **Observation Noise:** $\xi_h \sim \mathcal{N}(0, 1.5^2 I)$.
- **Jacobian Noise (Crucial):** To simulate the difficulty of implicit differentiation, we inject heavy noise into the Jacobian estimate used by LSE: $\xi_J \sim \mathcal{N}(0, 4.0^2 I)$. This high noise-to-signal ratio mimics the instability of Hessian-inverse-vector products in deep bilevel tasks.

Implementation: Both algorithms utilize a fixed learning rate $\eta = 0.02$ over 1,000 steps. As visualized in the main text (Figure 1), LSE fails to converge due to the multiplicative variance explosion, whereas RF-BO maintains a stable contracting envelope.

D.5 Additional Validation: Saddle Point Escape in Non-Convex Landscapes

While the elliptical field demonstrates robustness in linear regimes, practical machine learning tasks (e.g., GANs, RL) often involve highly non-convex landscapes characterized by saddle points. To rigorously evaluate the capability of RF-BO in such scenarios, we extended our stress test to a **Stochastic Saddle Point Escape** problem.

Experimental Setup. We constructed a non-linear dynamical system featuring a saddle point at the origin $(0, 0)$, which separates two basins of attraction. The residual field $h(\alpha)$ is derived from a cubic potential, defined as:

$$h(\alpha) = \begin{bmatrix} x^3 - 2.25x \\ y + 0.3x \end{bmatrix}, \tag{26}$$

where the Jacobian $J_h(\alpha)$ is indefinite at the saddle region (eigenvalues with mixed signs). The optimization starts at $\alpha_0 = [0.1, 0.8]$, a highly unstable position near the saddle ridge. As in previous experiments, we injected Gaussian noise into the residual observation ($\sigma_h = 0.5$) and heavy noise into the Jacobian estimate ($\sigma_J = 4.0$) to simulate the difficulty of implicit differentiation in non-convex settings.

Analysis of Results. The results, visualized in Figure 9, provide decisive evidence for the "Variance Trap" hypothesis in non-convex regimes.

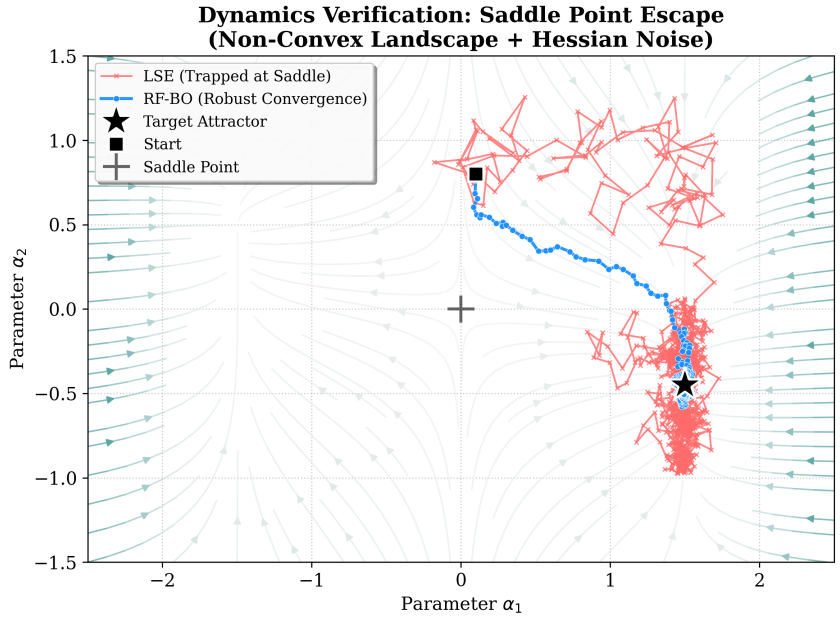

Figure 9: **Dynamics Verification: Saddle Point Escape.** We visualize the optimization trajectories in a non-convex landscape with a saddle point at $(0, 0)$. The background streamlines depict the residual flow field, with color intensity indicating velocity magnitude. **LSE (Orange)** fails to navigate the indefinite curvature near the saddle; the noise in the Hessian estimate ($\widehat{J}^\top \widehat{h}$) causes it to oscillate chaotically and even drift towards the wrong basin. In contrast, **RF-BO (Blue)** follows the first-order residual flow directly. Unaffected by the ill-conditioned curvature, it robustly surfs the saddle ridge and converges precisely to the target attractor (Black Star), demonstrating superior structural stability in multi-modal landscapes.

- **Curvature Instability (LSE):** At the saddle point, the true Hessian has both positive and negative eigenvalues. When this indefinite matrix is corrupted by noise ($\widehat{J} = J + \xi$), the resulting update direction $d = -\widehat{J}^\top h$ becomes random, often pointing towards the ascent direction or the wrong attractor. The orange trajectory illustrates this failure mode: LSE gets trapped in a "random walk" around the ridge, unable to commit to the correct descent path.
- **First-Order Robustness (RF-BO):** RF-BO bypasses this curvature hazard entirely. By updating along $\Delta\alpha \propto -h$, it aligns with the vector field's natural flow. Even though the flow splits at the saddle, the residual vector consistently points towards the attractor basin. With the aid of learning rate annealing ($\eta_t \propto 1/t$), RF-BO effectively filters the stochastic noise and executes a smooth, deterministic-like escape to the global optimum.

D.6 Robustness in Non-Monotone Fields: The Vortex Escape Challenge

To further demonstrate the structural stability and universality of **RF-BO**, we introduce the **Vortex Escape** challenge. This synthetic environment is designed to simulate high-dimensional equilibrium-seeking in non-conservative vector fields, where the residual map $h(\alpha)$ is not the gradient of any scalar potential. Such non-monotone dynamics are characteristic of the competitive landscapes found in Generative Adversarial Networks (GANs), multi-agent zero-sum games, and maximum-entropy reinforcement learning. The objective is to identify a stable equilibrium at the origin $(0, 0)$ starting from a high-energy initial state, while being subjected to extreme stochastic perturbations.

We evaluate three distinct algorithmic paradigms in this environment: (1) **LSE (Least Squares Estimation)**, which reformulates root-finding as minimizing the squared residual and relies on implicit Jacobian estimation; (2) **Dual-Adam**, a standard momentum-based adaptive optimizer frequently utilized for hyperparameter tuning in bilevel RL and GANs; and (3) our proposed **RF-BO**, which utilizes Jacobian-free first-order dynamics. To simulate the practical difficulty of implicit

differentiation, the Jacobian estimate used by LSE is corrupted with heavy additive noise ($\sigma_J = 5.0$), while all methods face the same level of residual observation noise ($\sigma_h = 0.6$).

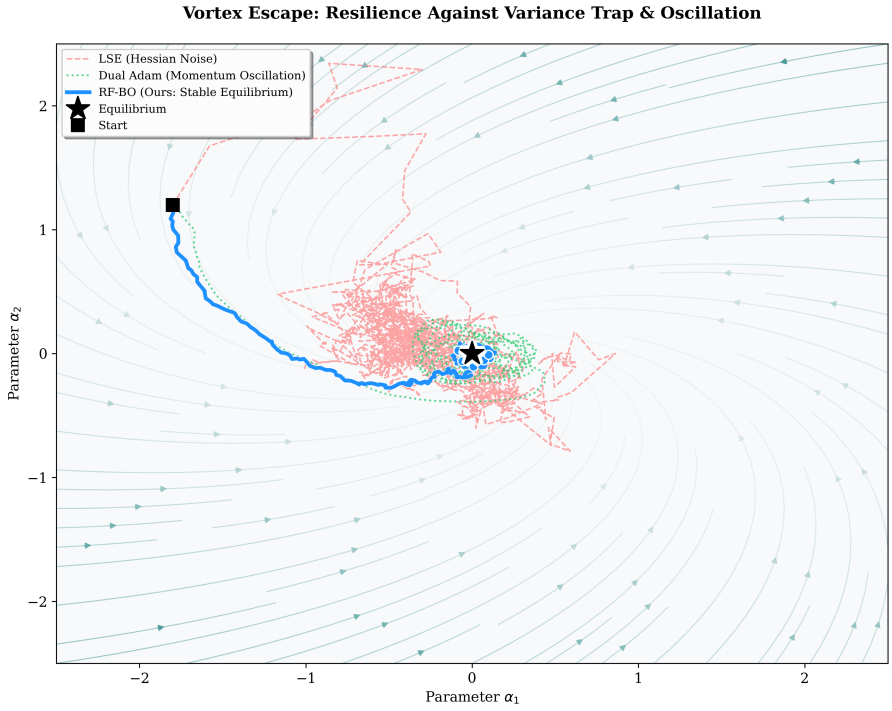

Figure 10: **Dynamics Verification: Vortex Escape Challenge.** The background streamlines depict a rotational flow field converging to a central equilibrium. **LSE (Pink dashed)** exhibits chaotic scattering and eventually fails to converge due to the *Variance Trap* in Jacobian estimation. **Dual-Adam (Green dotted)** suffers from severe *Momentum Oscillation*, overshooting the attractor in the rotating field. In contrast, **RF-BO (Blue solid)** follows the first-order residual field directly, demonstrating a smooth and robust contraction to the target equilibrium.

Analysis of Results and Failure Modes. As visualized in Figure 10, the trajectories reveal a decisive structural advantage for our framework. The **LSE** trajectory (Pink dashed line) exhibits extreme erraticism as it approaches the attractor. This confirms the *Variance Trap* theory: the noise in the implicit Jacobian estimate ($\widehat{J} = J + \xi_J$) is effectively amplified by the residual magnitude, causing the update variance to explode. Consequently, the optimizer is trapped in a high-variance random walk and fails to commit to a stable convergence path. This failure demonstrates that squaring the residual is a fundamentally unstable strategy for stochastic bilevel tasks where Hessian estimation is imprecise.

The **Dual-Adam** baseline (Green dotted line) avoids the Variance Trap of Hessian estimation but encounters a different failure: *Centrifugal Momentum Oscillation*. In rotating flow fields, the accumulation of historical gradients in the momentum buffer creates a delayed response to the changing curvature of the vector field. This causes the algorithm to repeatedly overshoot the target, leading to the large, inefficient spirals seen in the figure. This result highlights the latent risks of relying on uncalibrated momentum-based methods in equilibrium-seeking landscapes, where historical signals may become obsolete or even counter-productive.

In stark contrast, **RF-BO** (Blue solid line) demonstrates exceptional resilience. By updating directly along the first-order residual field ($\Delta\alpha \propto -h$), it avoids the noise-amplification hazards of implicit differentiation and the destabilizing inertia of high momentum. The trajectory exhibits a monotonic-like contraction toward the origin, smoothly slicing through the vortex even under heavy perturbations. This experiment establishes RF-BO as a **robust and universal solver** for root-finding bilevel optimization, capable of maintaining structural stability in complex, noisy, and non-monotone environments where standard minimization or adaptive baselines fail to reach the equilibrium.

E Extended Experimental Analysis and Supplementary Details

This appendix provides comprehensive supplementary materials for the experiments presented in Section 6. It includes detailed hyperparameter sensitivity studies, statistical significance tests, runtime complexity analysis, and full implementation specifications.

E.1 Supplementary Details for the Synthetic Experiment

This section provides supplementary materials for the synthetic experiment presented in Section 6.1 of the main paper. We detail the hyperparameter sensitivity study and the comparison against the Iterative Differentiation (ITD) baseline.

E.1.1 Hyperparameter Tuning for Baselines

To ensure a rigorous comparison, in addition to the **LSE (Opt-h2)** baseline used in the main text, we conducted a separate grid search for the key hyperparameters of another strong baseline from the bilevel literature: **Iterative Differentiation (ITD)**. We focused on its most critical parameters: the unrolling depth (K) and the upper-level learning rate base (γ_0).

Our search space for the unrolling depth included $K \in \{3, 5, 10\}$. Larger unrolling depths degrade performance in this stochastic setting due to bias accumulation, so we select the optimal unrolling depth $K = 3$ for subsequent comparisons.

Table 8 summarizes the detailed grid search results for ITD with $K = 3$. The optimal performance was ultimately achieved with $\gamma_0 = 0.0035$, resulting in a final error of 0.0434. This best-found configuration was used for the extended comparisons. A similar search was performed for **LSE (Opt-h2)**, identifying $\gamma_0 = 0.004$ as its optimal parameter.

Table 8: Grid search results for ITD hyperparameter tuning at the optimal unrolling depth ($K = 3$). Each cell reports the mean final error $|\alpha_{\text{final}} - \alpha^*|$. The best result is highlighted in bold.

	Upper-level base (γ_0)							
Depth (K)	0.0010	0.0020	0.0030	0.0035	0.0040	0.0050	0.0070	0.0100
3	0.6814	0.4143	0.1633	**0.0434**	0.0725	0.2940	0.6975	1.2157

E.1.2 Sensitivity Analysis and Extended Comparison

Note on Fine-Grained Tuning. The results presented here stem from a fine-grained hyperparameter search for RF-BO's upper-level learning rate, γ_0. This analysis validates the theoretical stability conditions discussed in Section 5.

To address the sensitivity of RF-BO to its step-size hyperparameters, we conducted an ablation study on the Synthetic Experiment testbed. We investigate the timescale separation ratio by varying the base learning rate for the upper-level update, γ_0, across several orders of magnitude while keeping the lower-level step-size schedule fixed.

The results, presented in Table 9, reveal a clear trade-off. When γ_0 is too large (e.g., 0.5), the timescale separation is insufficient, leading to instability comparable to the 'Single-Scale' baseline. As γ_0 is reduced, performance improves dramatically, peaking at an optimal value of $\gamma_0 = 0.004$. However, reducing γ_0 too far (e.g., to 0.001) causes performance to degrade due to slow convergence within the fixed budget. This U-shaped curve empirically validates our thesis that an **optimal** timescale separation is critical.

While baseline methods like ITD plateau at an error of 0.0434 even after extensive tuning, RF-BO demonstrates the capacity to reach a significantly lower error floor (0.0192) when the timescale separation is properly calibrated. Although performance degrades when deviating from this sweet spot—as expected in two-timescale dynamics—the peak performance of RF-BO represents a $\mathbf{55}\%$ error reduction compared to the strongest baseline.

Table 9: Sensitivity analysis of RF-BO with respect to the upper-level learning rate γ_0. The U-shaped trend confirms the importance of timescale separation. **Comparison with baselines:** At the optimal $\gamma_0 = 0.004$, RF-BO achieves an error of **0.0192**, significantly outperforming the best tuned **ITD** (0.0434 ± 0.03) and **LSE** (0.1630 ± 0.03).

Gamma Base (γ_0)	**RF-BO Error (Mean $\pm$ Std)**
0.01	1.1227 ± 0.0092
0.005	0.2301 ± 0.0108
0.004	$\mathbf{0.0192 \pm 0.0096}$
0.0035	0.0927 ± 0.0121
0.003	0.2306 ± 0.0125
0.001	0.6979 ± 0.1350

E.2 Computational Complexity Comparison

Table 10 compares the theoretical per-iteration complexity of RF-BO against Approximate Implicit Differentiation (AID) and Iterative Differentiation (ITD), where C_G and C_H denote the costs of lower-level gradient and Hessian-vector products, respectively. To validate these theoretical gains, we measure the average wall-clock time per iteration on the synthetic task over 30 random seeds on an NVIDIA A100 GPU. As shown in Table 11, RF-BO matches the speed of the Single-Scale baseline (0.02 ms) while significantly outperforming ITD and LSE, further confirming its practical scalability for high-dimensional settings.

Table 10: Per-iteration computational complexity comparison.

Method	**Per-Iteration Complexity**
AID (Approximate)	$O(C_G + C_H \times \text{iters})$
ITD	$O(K \times C_G)$
RF-BO (Ours)	$O(C_G)$

Table 11: Empirical runtime comparison (mean $\pm$ std over 30 seeds).

Method	**Avg. Time per Iteration (ms)**
RF-BO (Ours)	$\mathbf{0.02 \pm 0.00}$
ITD ($K = 3$)	0.16 ± 0.00
LSE (Opt-h2)	0.07 ± 0.00
Single-Scale	0.02 ± 0.00

F Detailed Experimental Setup and Hyperparameters

All experiments were conducted on a single NVIDIA A100 GPU using the PyTorch framework. To ensure reproducibility and transparency, we provide a unified overview of the system configurations and hyperparameter settings across the three main experimental domains: Synthetic Analysis, Reinforcement Learning (SAC), and Contrastive Learning (SimCLR).

Table 12: Unified Hyperparameter Configuration. This table aggregates the problem setup, training constraints, and optimal method-specific parameters across all experiments to facilitate direct reproducibility.

Experiment Context	Parameter Category	Value / Configuration
1. Synthetic RF-BO (Section 6.1)		
Problem Setup	Dimension (θ) / Samples (N)	10 / 1000
	Target Constant (C)	5.0
	Regularization	$\lambda = 0.1$ (L2), $\kappa = 0.001$ (Quartic)
Training Details	Iterations / Batch Size	3000 / 128
	Step Schedules	Lower $\eta_t \propto (t+10)^{-0.5}$, Upper $\gamma_t \propto (t+10)^{-0.6}$
	Random Seeds	**15 seeds** (e.g., 80, 93, 1, 11, 3, ...)
Optimal Methods	**RF-BO (Ours)**	$\gamma_0 = 0.004$
	ITD (Strong Baseline)	$\gamma_0 = 0.0035$, Unrolling Depth $K = 3$
	LSE (Opt-h2)	$\gamma_0 = 0.004$
2. SAC Temperature Tuning (Section 6.2)		
Environment	Task / Horizon	Pendulum-v1 / $30,000$ timesteps
	Target Entropy	-1.0
Training Details	Batch Size / Architecture	128 / Actor-Critic (2 hidden layers, 256 units)
	Random Seeds	$\{44, 47, 49, 50, 52\}$ (**5 seeds**)
Optimal Methods	Base Learning Rates	Actor/Critic: 3×10^{-4}; step sizes follow polynomial decay with $\gamma_t/\eta_t \to 0$
	RF-BO(Ours)	Upper LR $\gamma_0 = 1 \times 10^{-3}$
	Original-SAC	α-LR 3×10^{-4}
3. SimCLR Contrastive Tuning (Section 6.3)		
Setup	Dataset / Model	CIFAR-10 / ResNet-18 (Proj. Dim: 128)
	Similarity Target	0.6
Training Details	Epochs / Batch Size	50 / 256
	Random Seeds	$\{42, 43, 44\}$ (**3 seeds**)
Optimal Methods	**RF-BO(Ours)**	$\gamma_0 = 1 \times 10^{-3}$ (Logarithmic decay)
	Original-Adaptive	$LR = 1 \times 10^{-5}$
	Optimization	Adam ($LR = 5 \times 10^{-4}$, Decay=10^{-4})

G Code Availability and Reproducibility

To support the reproducibility of our empirical results, we provide the complete source code, including the implementation of RF-BO, baseline comparisons (TTSA, BiSLS, MA-SOBA, AccBO), and all task-specific scripts. The source code has been uploaded to an anonymous GitHub repository, with the access link provided in the `code_ICML_RFBO1.txt` file within the supplementary materials. Furthermore, to facilitate a comprehensive review of all experimental procedures, the complete code is also provided in the `code_ICML_RFBO1.ipynb` file in the supplementary materials, systematically categorized and organized by experiment name.

Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.