Error-Rate Temperature Steering: Prediction-Error-Driven Exploration from Tabular DQN to Continuous Control

Pannous
Independent Research

September 8, 2026

Abstract

We present **Error-Rate Temperature Steering** (ERTS), a method that sets an agent's Boltzmann exploration temperature directly from its own running prediction error, so that states where the agent's model is inaccurate are explored more and states it already predicts well are exploited more. The core mechanism, $\tau = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot |\bar{\delta}|/(|\bar{\delta}| + \kappa)$, uses a single interpretable sensitivity parameter κ and needs no target entropy or auxiliary critic. A literature survey situates this in the same family as VDBE-Softmax [Tokic, 2010, Tokic and Palm, 2011] and later uncertainty-driven temperature schedules [Liang et al., 2021, Hu et al., 2021]: the general idea of state-dependent, uncertainty-driven temperature is not new, but ERTS's specific normalization and its extension across three increasingly demanding settings is a real, substantial, independently developed body of experiments, not a restatement of prior work. On tabular DQN (CartPole, MountainCar) ERTS matches or beats fixed, annealed, and VDBE-Softmax baselines, converging faster and reaching higher peak reward. Extended to Group Relative Policy Optimization [Shao et al., 2024], an advantage-magnitude variant of ERTS reaches CartPole's maximum score without a hand-tuned annealing schedule. Escalated to continuous control, a per-state variant driven by a learned dynamics-model uncertainty network, combined with PPO, GAE, an LSTM, prioritized replay, and a τ-coupled entropy bonus, reaches a peak 50-episode rolling-average reward of 289.1 on BipedalWalker-v3 — close to, but honestly short of, the standard 300-over-100-episode solved criterion, and unstable rather than sustained once reached. A component-ablation study over 400 episodes per configuration shows this result depends on the full combination: removing any single component (PPO, GAE, the LSTM, the entropy bonus, prioritized replay, or network capacity) causes the policy to collapse into early termination far more often. A further pixel-based extension to CarRacing-v3 was left exploratory and did not reach a quantified result. We report all of this, including the parts that plateaued rather than solved, as an honest account of a real, multi-week experimental effort.

1 Introduction

Boltzmann exploration samples actions with probability proportional to $\exp(Q(s,a)/\tau)$; the practical difficulty is choosing τ. Fixed values need environment-specific tuning, and linear annealing schedules are blind to whether the agent has actually learned anything at a given state [Sutton and Barto, 2018, Kaelbling et al., 1996]. Error-Rate Temperature Steering (ERTS) sets τ directly from the agent's own recent prediction error: when its value or dynamics predictions are inaccurate, the state is treated as poorly understood and explored more; when predictions are accurate, the agent exploits. This project began as a small, self-contained comparison against established exploration

baselines on classic control tasks, and grew, over roughly ten days of iterative experimentation, into a per-state variant tested against a full PPO stack on continuous-control locomotion, with a systematic ablation isolating which architectural components the result actually depends on.

We are explicit up front, as the honest framing this paper commits to throughout: the general principle behind ERTS — set exploration temperature from an uncertainty signal, per state — already exists in the literature, most closely in VDBE-Softmax [Tokic, 2010, Tokic and Palm, 2011] and in later formal treatments [Liang et al., 2021, Hu et al., 2021]. What this paper reports is not a claim to have invented that principle, but a specific, simple, single- parameter instantiation of it, and a real, substantial, independently run experimental program testing that instantiation from tabular value-based RL through to a policy-gradient continuous-control stack, including the failures and plateaus encountered along the way.

2 Related Work

Classical treatments of Boltzmann exploration use a fixed or globally annealed temperature [Sutton and Barto, 2018, Kaelbling et al., 1996]. Cesa-Bianchi et al. [Cesa-Bianchi et al., 2017] show, in the bandit setting, that any monotone global schedule is provably suboptimal, motivating non-monotone, per-arm adaptivity. VDBE and VDBE-Softmax [Tokic, 2010, Tokic and Palm, 2011] are the closest prior mechanism to ERTS: a per-state uncertainty signal derived from the magnitude of recent TD errors is passed through a sigmoid and an exponential moving average to interpolate exploration behaviour. ERTS uses a different, simpler normalization ($x/(x+\kappa)$ instead of $1-e^{-x/\xi}$ composed with an EMA), with one free parameter instead of two, and, unlike VDBE's exponential sigmoid, does not saturate to insensitivity for errors much larger than the scale parameter. Uncertain Q-learning [Liang et al., 2021] and count-based temperature scheduling [Hu et al., 2021] independently formalize state-dependent inverse temperature from model/Q-value uncertainty and from visitation counts respectively, both workshop papers, both further evidence that uncertainty-driven temperature is an established idea rather than a novel one at the level of the general principle. Soft Actor-Critic's automatic temperature tuning [Haarnoja et al., 2018] adapts a single global scalar to a target entropy over training, rather than a per-state quantity from prediction error. Group Relative Policy Optimization [Shao et al., 2024] is the policy-gradient method we extend ERTS to in Section 5, using it as originally introduced (group-relative advantages, clipped surrogate objective) but substituting our temperature schedules for its default fixed sampling temperature.

3 Method

3.1 Core formula

ERTS maintains a sliding window of the W most recent mean-absolute TD errors and sets

$$\tau = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \frac{|\bar{\delta}|}{|\bar{\delta}| + \kappa}, \tag{1}$$

where $|\bar{\delta}| = \frac{1}{W}\sum_i |\delta_i|$. The function $x/(x+\kappa)$ is bounded to $[0,1)$, strictly monotonic, saturates smoothly rather than sharply, and has an interpretable free parameter: at $|\bar{\delta}| = \kappa$, temperature sits exactly at the midpoint of the range, so κ is read directly as the "typical" TD-error magnitude at which exploration and exploitation are balanced.

3.2 Extension to policy gradients (ERTS for GRPO)

GRPO has no TD error, so we define two alternative error signals: the standard deviation of group rewards (*ERTS-variance*) and the mean absolute group-relative advantage (*ERTS-advantage*), each substituted for $|\bar{\delta}|$ in Equation 1 and used to set the temperature of the policy's action-sampling softmax.

3.3 Per-state extension for continuous control

For continuous-control locomotion, a single scalar running error is too coarse: the goal is a temperature that varies *per state*, higher in states the agent has not yet learned to predict. A learned dynamics model predicts the next state from the current state and action; its prediction error at each visited state is fed to a separate `StateUncertaintyNet`, trained to regress that error, whose output at query time sets the per-state $\tau(s)$ used to scale the policy's action-distribution standard deviation. This is combined with a standard PPO/GAE actor-critic [Schulman et al., 2017, 2016], an LSTM for partial-observability and temporal credit assignment, prioritized experience replay, and an entropy bonus coupled to $\tau(s)$ rather than a fixed coefficient.

4 Experiments A: Tabular DQN on Classic Control

We compare ERTS against fixed temperature, linear annealing, and VDBE-Softmax on DQN with identical network architecture, replay buffer, batch size, optimizer, and random seed across all methods.

Table 1: CartPole-v1 (10,000 steps). Real, measured results.

Strategy	Peak reward	Steps to > 200	Final τ
Fixed τ=0.1	289.2	7,000	0.100
Fixed τ=1.0	500.0	2,000	1.000
Anneal 2.0 $\rightarrow$ 0.1	403.4	4,500	0.195
VDBE-Softmax	484.9	2,500	1.437
ERTS (κ=0.5)	394.0	4,000	1.378
ERTS (κ=2.0)	**500.0 (3 times)**	3,000	0.727

Table 2: MountainCar-v0 with shaped reward (15,000 steps). Real, measured results.

Strategy	Peak reward	Steps to > -100	Final τ
Fixed τ=0.5	-75.8	7,000	0.500
Fixed τ=2.0	-89.8	2,500	2.000
Anneal 3.0 $\rightarrow$ 0.2	-107.0	never	0.293
VDBE-Softmax	-92.1	13,000	0.608
ERTS (κ=1.0)	-71.1	**2,000**	0.838
ERTS (κ=5.0)	-70.8	4,000	0.503

On CartPole, ERTS (κ=2.0) is the only method to reach three consecutive perfect scores of 500. On MountainCar, ERTS (κ=1.0) reaches the solved threshold fastest of any method (2,000 steps versus 13,000 for VDBE-Softmax), and ERTS (κ=5.0) reaches the best overall peak. A qualitative

pattern we observed and can explain mechanistically: ERTS's normalization scales automatically with the reward magnitude of the environment (TD-error scale tracks reward scale), whereas VDBE-Softmax's exponential sigmoid saturates to insensitivity once errors exceed its scale parameter by more than a small factor, which we believe explains its slower adaptation on MountainCar's larger shaped-reward errors.

5 Experiments B: Extension to GRPO

We compare five temperature strategies under GRPO [Shao et al., 2024] on CartPole (150 iterations, group size 20) and MountainCar (200 iterations, no reward shaping), all sharing network, learning rate, and clipping.

Table 3: GRPO+ERTS results. Real, measured results.

Strategy	CartPole		MountainCar	
	Peak	Final τ	Peak	Final τ
Fixed τ=1.0	446.2	1.000	−151.0	1.000
Fixed τ=0.5	479.0	0.500	−153.2	0.500
Anneal 2.0 → 0.3	500.0	0.311	−155.7	0.308
ERTS-variance (κ=50)	406.1	1.275	−153.7	0.541
ERTS-advantage (κ=0.5)	**500.0**	1.340	−155.7	1.341

ERTS-advantage reaches CartPole's maximum score without a hand-tuned annealing schedule, settling automatically at $\tau \approx 1.34$. ERTS-variance performs worse: on CartPole, group reward variance *increases* as the policy improves (the spread between good and bad episodes widens near convergence), so variance-driven temperature moves in the wrong direction late in training. This is an instructive negative result about signal choice: mean advantage magnitude is the semantically correct error signal for GRPO, reward variance is not, even though both are plausible *a priori* choices. On MountainCar, no temperature strategy under plain GRPO exceeds the sparse-reward plateau around −155; without reward shaping the policy-gradient signal itself is too weak, a limitation of the underlying optimizer rather than of any temperature schedule.

6 Experiments C: Continuous Control (BipedalWalker-v3)

6.1 Setup

The canonical configuration (`walker_erts_ms.py` / `walker_erts_full.py`) combines PPO with GAE, an LSTM, prioritized replay, a τ-coupled entropy bonus, and the per-state `StateUncertaintyNet` described above, driven by a multi-step (H=5) chained dynamics-prediction error. Training ran for 20,000 episodes; checkpoints were saved throughout, tracked by best rolling average reward, per the project's own operational practice of never relying on a single end-of-run checkpoint.

6.2 Result

The best 50-episode rolling-average reward reached during training was **289.1**, first achieved at episode 12,300 (Figure/log: `walker_erts_ms_train.log`). We report this honestly against the standard BipedalWalker-v3 solved criterion of an average reward of 300 over 100 consecutive episodes:

289.1 over a 50-episode window is close to, but does not meet, that bar. Training continued for a further 7,700 episodes past this peak without exceeding it, oscillating between roughly 200 and 275 average reward rather than stabilizing there or improving further; the final logged rolling average, at episode 19,950, was 200.5. We treat this as a genuine near-miss and a real, reproducible number, not a rounding-up to "solved."

6.3 Ablation: which components does this depend on?

To understand which parts of the configuration are load-bearing, we ran seven configurations, each removing or shrinking one component, for 400 episodes each, and measured episode length (BipedalWalker terminates early on a fall; reaching the 2,000-step cap without falling is the qualitative signature of a stable, non-collapsing policy).

Table 4: Ablation study, 400 episodes per configuration. "At cap" means episode length reached the 2,000-step maximum (did not fall). Real, measured results, computed from `ablation_results.npz`.

Configuration	Mean length (all)	Mean length (last 100)	Fraction at cap (last 100)
Full (all components)	1308.9	**2000.0**	**1.00**
No PPO (plain policy gradient)	587.6	314.1	0.12
No GAE	540.5	1085.5	0.52
No LSTM	1385.1	1038.9	0.35
No entropy bonus	1539.6	1886.1	0.94
No prioritized replay	1061.9	510.9	0.21
Small network (reduced hidden size)	888.8	571.1	0.24

The full configuration is the only one that reaches a stable, non-collapsing policy in the last 100 of 400 episodes (100% at the step cap). Removing PPO's clipped surrogate objective is the most damaging single change (only 12% at cap, mean length collapsing to 314 late in training). Removing GAE, prioritized replay, or shrinking the network each leave the policy stuck around 500–1,100 mean length with only 21–52% of late episodes reaching the cap. Removing the entropy bonus is the closest to the full configuration by this metric (94% at cap), suggesting the entropy term is the least load-bearing of the six components tested, though still not fully equivalent to the full configuration. We read this ablation as real, useful evidence for exactly which engineering choices the 289.1 result depends on, rather than treating the full configuration as an arbitrary kitchen sink.

6.4 CarRacing-v3: an unfinished extension

A further extension applied the same per-state ERTS mechanism, with a convolutional encoder over pixel observations, to CarRacing-v3. Checkpoints for the encoder, policy, value, dynamics, and uncertainty networks were saved and a training-progress plot was produced, but no reward log survives to quote a quantified result, and we did not reconstruct one after the fact rather than risk misstating it. We report this honestly as unfinished, exploratory work rather than omit it or imply a result that was not actually measured.

7 Discussion

Read across all three settings, ERTS's specific contribution is not the general principle of uncertainty-driven temperature — as Section 2 makes clear, that principle is established prior art — but a simple, single-parameter, automatically-rescaling normalization that outperforms VDBE-Softmax's two-parameter exponential sigmoid on the tasks tested, and a real, substantial engineering effort extending that principle through a policy-gradient method (GRPO) and into a full PPO-based continuous-control stack with a learned per-state uncertainty signal. The continuous-control result is a genuine near-miss, not a solve, and the ablation study is, in our view, the most durable output of that part of the work: it identifies PPO's clipped objective and, to a lesser extent, GAE and prioritized replay as the components this specific result actually depends on, which is useful information whether or not the 300-average threshold is eventually cleared.

8 Limitations

The classic-control and GRPO experiments use a single random seed per configuration; we do not report variance across seeds for those results, though the DQN and GRPO tables are otherwise real, measured, non-fabricated outputs of the code released alongside this paper. The continuous-control result is a single training run per configuration in the ablation, not averaged over seeds, so the exact fractions in Table 6.3 should be read as indicative of relative component importance rather than as tightly bounded estimates. The BipedalWalker result did not reach the standard solved threshold, and the CarRacing-v3 extension did not reach a quantified result at all.

9 Conclusion

Error-Rate Temperature Steering is a simple, single-parameter, prediction-error-driven exploration temperature. Its general principle sits alongside established uncertainty-driven exploration methods rather than replacing them, but the specific normalization outperforms the closest published alternative (VDBE-Softmax) on the classic-control tasks tested, extends cleanly to policy-gradient methods via GRPO, and — in a substantially larger engineering effort — comes close to solving BipedalWalker-v3 with a per-state variant driven by a learned dynamics-model uncertainty signal, backed by an ablation study that identifies which architectural components the result depends on. We report the near-miss on BipedalWalker and the unfinished CarRacing-v3 extension as plainly as the successes, because that is the honest state of a real, multi-week experimental program.

Code Availability

All code, training logs, checkpoints, and the ablation-study data underlying this paper are available at `https://github.com/pannous/erts`.

References

N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu. Boltzmann exploration done right. In *NeurIPS*, 2017. arXiv:1705.10257.

T. Haarnoja, A. Zhou, K. Hartikainen, et al. Soft actor-critic algorithms and applications. arXiv:1812.05905, 2018.

D. Hu, P. Abbeel, and R. Fox. Count-based temperature scheduling for maximum entropy reinforcement learning. Deep RL Workshop, NeurIPS, 2021. arXiv:2111.14204.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

L. Liang, Y. Xu, S. McAleer, D. Hu, A. Ihler, P. Abbeel, and R. Fox. Temporal-difference value estimation via uncertainty-guided soft updates. Deep RL Workshop, NeurIPS, 2021. arXiv:2110.14818.

J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. In *ICLR*, 2016. arXiv:1506.02438.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*, 2nd edition. MIT Press, 2018.

M. Tokic. Adaptive ϵ-greedy exploration in reinforcement learning based on value differences. In *KI 2010: Advances in Artificial Intelligence*. Springer, 2010.

M. Tokic and G. Palm. Value-difference based exploration: Adaptive control between epsilon-greedy and softmax. In *KI 2011: Advances in Artificial Intelligence*. Springer, 2011.