Optimal Control Meets Online Mechanism: Adaptive Policy Learning with Strategic Agent Response

Anonymous authors
Paper under double-blind review

Abstract

Many platforms and protocols must adaptively choose a policy action (i.e., a price, subsidy, emission rate, or resource allocation) while heterogeneous, self-interested agents respond strategically. We formalize this as a *two-level online mechanism-design* problem: the planner commits to a state-contingent policy (mechanism network) that maps observable system state to an action; agents respond by optimizing private objectives, producing an aggregate equilibrium response. We microfound the agent layer via a heterogeneous-agent threshold equilibrium: the aggregate response is the unique solution to a monotone fixed point and increases smoothly with the planner's action (Proposition 1). With adjustment frictions, agent behavior co-adapts with the planner through a controlled diffusion whose drift is locally affine in the action, yielding an explicit continuous-time Bellman/HJB characterization of the optimal value (critic) and a closed-form greedy policy-improvement map that generalizes classical stabilization rules while internalizing the marginal value of strategic participation (Proposition 2). For deployment, we use the HJB-derived structure as an expert prior to initialize a compact mechanism network and refine it online via projected stochastic approximation with convergence guarantees (Theorem 1). We instantiate the framework on *adaptive token issuance* in blockchain protocols, where the planner sets an issuance rate and agents decide whether to stake. In comprehensive experiments (1000 Monte Carlo runs) across multiple economic regimes, the adaptive mechanism achieves significant improvements in target tracking and maintains stability relative to fixed and zero-action baselines, with all improvements statistically significant.

1 Introduction

A recurring problem across economics, operations research, and multi-agent systems is *adaptive policy design with strategic feedback*: a planner must choose actions over time (prices, subsidies, tax rates, emission schedules) while a population of self-interested agents continuously adapts its behavior in response. Classical examples include central-bank interest-rate setting (Taylor, 1993), congestion pricing, platform fee design, and emission permit allocation. In each case, the planner's action shapes the incentive landscape, agents respond strategically, and the resulting aggregate behavior feeds back into the state the planner observes.

Traditional approaches either fix the policy rule *a priori* (e.g., Taylor rules, Friedman k-percent rules) and forgo adaptivity, or rely on discretionary governance that introduces time-inconsistency (Kydland and Prescott, 1977). Recent work on differentiable mechanism design (Dütting et al., 2019; Bichler and Parkes, 2025) and two-level reinforcement learning (Zheng et al., 2021) has shown that neural-network-parameterized policies can be learned end-to-end, but these methods typically treat the agent response as a black box (learned via simulation) and offer limited structural insight into the optimal policy.

We bridge these approaches by combining *microfounded strategic response* with *continuous-time optimal control* and *online learning*. The key insight is that when agents follow a threshold participation rule with heterogeneous costs, the equilibrium aggregate response has a tractable monotone structure (a fixed-point equation) that can be linearized around an operating point. This makes the planner's Hamilton–Jacobi–Bellman (HJB) equation analytically solvable for the greedy policy-improvement step, providing an *expert prior* that dramatically structures the learning problem. We then deploy a compact neural mechanism network initialized from this prior and refine it online via projected stochastic gradient descent, with convergence guarantees from the ODE method for stochastic approximation.

The framework is general: the planner's action can be any scalar instrument, the agent response can be any participation or allocation decision with heterogeneous costs, and the system state can include multiple observables. We instantiate and evaluate it on *adaptive token issuance* in blockchain protocols; a setting where the planner (protocol) sets an issuance rate, agents decide whether to stake tokens, and the resulting staked fraction feeds back into protocol security and monetary dynamics. This application is a natural testbed because the action space, state, and agent responses are all publicly observable on-chain, and the mechanism can be deployed algorithmically.

Contributions. We make four contributions:

1. A microfounded equilibrium model for strategic agent response with heterogeneous costs, yielding a monotone, differentiable aggregate response map (Proposition 1).
2. An HJB-derived closed-form greedy policy-improvement map that provides a structured expert prior for initialization (Proposition 2).
3. An online gradient-based mechanism optimization algorithm with convergence guarantees (Theorem 1).
4. Comprehensive empirical validation (1000 Monte Carlo runs) across economic regimes and agent distributions, demonstrating statistically significant improvements over baselines ($p < 0.001$).

2 Related Work

Differentiable mechanism design. Dütting et al. (2019) introduced the use of neural networks trained via gradient descent to approximate optimal auction mechanisms; Bichler and Parkes (2025) survey the broader program of differentiable economics. These methods learn mechanisms from data but typically do not model the agent response explicitly; agents are part of the training environment.

Two-level RL for policy design. The AI Economist (Zheng et al., 2021) trains co-adapting planner and agent policies via two-level deep RL. Our work shares the two-level structure but replaces the black-box agent layer with a microfounded equilibrium model, gaining analytical tractability and an HJB-derived expert prior.

Continuous-time RL and HJB equations. Kim et al. (2021) formalize Q-learning via HJB equations for deterministic continuous-time systems. We extend this to a stochastic setting with strategic feedback and use the HJB solution for policy initialization rather than as the final policy.

Monetary policy and token issuance. Classical monetary-policy rules (Taylor, 1993; Friedman, 1960; Kydland and Prescott, 1977) are fixed-form feedback rules. Madrigal-Cianci and Breakey (2025) study state-dependent token issuance in the context of cryptocurrencies. Our framework generalizes these by learning the policy form from data while preserving structural guarantees.

3 General Framework: Two-Level Strategic Mechanism Design

We first present the framework in general terms, then specialize to token issuance in Section 4.

3.1 Setting

A planner observes a public system state $x(t) \in \mathcal{X} \subseteq \mathbb{R}^n$ and chooses an action $a(t) \in \mathcal{A} \subseteq \mathbb{R}$ at each time $t \in [0, T]$. A continuum of agents, indexed by a private cost parameter $c \geq 0$ drawn from a continuous CDF F with density f, observe the action and respond by participating (or not) in an activity whose per-unit reward depends on the action and the aggregate participation level.

Agent layer. Let $q(t) \in (0, 1]$ denote the aggregate participation rate. Each agent receives reward $R(a, q)$ from participating and participates if $R(a, q) \geq c$. We focus on the canonical case $R(a, q) = a/q$, where the planner's action is shared among participants (e.g., subsidies, rewards, or resource flows divided among participants).

Planner layer. The planner seeks to track target values for system-level KPIs (e.g., a target net flow rate $\pi^\star$, a target participation rate $q^\star$) while minimizing action cost, formalized as a quadratic running loss.

3.2 Microfounded Equilibrium Response

The threshold participation rule $R(a, q) \geq c$ with $R(a, q) = a/q$ yields a fixed-point condition for the equilibrium participation rate.

Proposition 1 (Equilibrium participation map and monotone response)**.** *For any $a \in (0, a_{\max}]$, the equilibrium participation rate $q^{\mathrm{eq}}(a) \in (0, 1]$ is the unique solution of*

$$q = F\left(\frac{a}{q}\right). \tag{1}$$

Moreover q^{eq} is differentiable and strictly increasing with derivative

$$\eta(a) := \frac{d}{da} q^{\mathrm{eq}}(a) = \frac{\frac{1}{q^{\mathrm{eq}}(a)} f\left(\frac{a}{q^{\mathrm{eq}}(a)}\right)}{1 + \frac{a}{(q^{\mathrm{eq}}(a))^2} f\left(\frac{a}{q^{\mathrm{eq}}(a)}\right)} > 0. \tag{2}$$

Proof. Define $h(q; a) = q - F(a/q)$. As $q \downarrow 0$, $F(a/q) \to 1$; at $q = 1$, $h(1; a) = 1 - F(a) \geq 0$. Uniqueness: $\partial_q h = 1 + \frac{a}{q^2} f(a/q) > 0$. Result (2) follows from implicit differentiation. □

Remark 1 (Generality). Proposition 1 holds for any reward function $R(a, q)$ that is increasing in a, decreasing in q, and satisfies $R(a, q) \to \infty$ as $q \to 0$. The specific form a/q is natural for divisible reward pools but the monotone fixed-point structure extends broadly.

3.3 Coadaptation Dynamics with Frictions

In practice, agents do not instantaneously reach equilibrium, but rather, participation adjusts with frictions. Let $x(t)$ denote a stock variable (e.g., circulating supply, resource level) and $q(t)$ the participation rate. We model:

$$dx(t) = \mu_x(x, q, a)\, dt + \sigma_x(x)\, dW_x(t), \tag{3}$$

$$dq(t) = \kappa\big(q^{\mathrm{eq}}(a(t)) - q(t)\big)\, dt + \sigma_q\, q(t)(1 - q(t))\, dW_q(t), \tag{4}$$

with $\kappa > 0$ the adjustment speed and $\sigma_q \geq 0$ participation volatility. The mean-reversion in (4) captures that participation gravitates toward the static equilibrium implied by the current action, while noise reflects idiosyncratic entry/exit decisions.

For analytical tractability, we linearize $q^{\mathrm{eq}}(a)$ around an operating point $\bar{a}$:

$$q^{\mathrm{eq}}(a) \approx q_0 + \eta a, \qquad \eta = \frac{d}{da} q^{\mathrm{eq}}(a)\Big|_{a=\bar{a}} > 0, \quad q_0 = q^{\mathrm{eq}}(\bar{a}) - \eta \bar{a}, \tag{5}$$

yielding a controlled diffusion with drift affine in a:

$$dq(t) = \kappa\big(q_0 - q(t) + \eta\, a(t)\big)\, dt + \sigma_q\, q(t)(1 - q(t))\, dW_q(t). \tag{6}$$

3.4 Planner Objective

The planner minimizes a quadratic loss around target KPIs:

$$r(t) = -\Big(w_1\big(g(x,q,a) - \pi^\star\big)^2 + w_2\big(q(t) - q^\star\big)^2 + w_3\, a(t)^2\Big), \tag{7}$$

where $g(x,q,a)$ is a system-level flow rate (e.g., net inflation, net resource change), $\pi^\star$ and $q^\star$ are targets, and $w_1, w_2, w_3 > 0$ are weights. The planner maximizes $\mathbb{E}[\int_0^T r(t)\,dt - \beta_q(q(T) - q^\star)^2]$.

3.5 HJB Critic and Greedy Policy Improvement

Let $V(t,x,q)$ be the optimal value function. Under standard regularity conditions, V satisfies the HJB equation (Fleming and Soner, 2006):

$$0 = \max_{a\in[0,a_{\max}]} \Big\{\partial_t V + \mathcal{L}^a V + r(t)\Big\}, \qquad V(T,x,q) = -\beta_q(q - q^\star)^2, \tag{8}$$

where under (3) and (6),

$$\mathcal{L}^a V = \mu_x(x,q,a)\,\partial_x V + \kappa(q_0 - q + \eta a)\,\partial_q V + \tfrac{1}{2}\sigma_x^2(x)\,\partial_{xx}^2 V + \tfrac{1}{2}\sigma_q^2\, q^2(1-q)^2\,\partial_{qq}^2 V. \tag{9}$$

Because the Hamiltonian is quadratic in a (given the affine approximation (5) and quadratic loss (7)), the greedy improvement map has closed form.

Proposition 2 (HJB-greedy action (critic-derived policy improvement))**.** *Assume V is differentiable and that μ_x is affine in a with coefficient $\partial_a\mu_x$. Define*

$$M(t,x,q) := (\partial_a\mu_x)\,\partial_x V + \kappa\eta\,\partial_q V. \tag{10}$$

Then the HJB-greedy action is

$$a^\star(t,x,q) = \Pi_{[0,a_{\max}]}\left(\frac{w_1\big(\pi^\star + e(t)\big) - \tfrac{1}{2}M(t,x,q)}{w_1 + w_3}\right), \tag{11}$$

where $e(t)$ captures exogenous flows that the action must offset.

Remark 2 (Interpretation). The first term in (11) offsets exogenous flows to track the target; M internalizes the marginal value of the stock variable and strategic participation via $\partial_q V$ and equilibrium sensitivity $\eta > 0$. This generalizes Taylor-style rules by adding a forward-looking value-of-participation term.

4 Application: Adaptive Token Issuance

We instantiate the framework on adaptive token issuance in blockchain protocols—a domain where the planner (protocol), action (issuance rate), agent response (staking), and system state are all publicly observable, making it a natural testbed.

The public state consists of circulating supply $S(t) \in \mathbb{R}_+$ and staked fraction $p(t) \in (0,1]$. The protocol chooses an issuance rate $i(t) \in [0, \iota_{\max}]$. Net locking $L(t)$, unlocking $u(t)$, and burn $b(t)$ define an observable contraction rate $e(t) := (L(t) - u(t) + b(t))/S(t)$, so that net inflation is $i(t) - e(t)$.

The stock variable is $x = S$, the participation rate is $q = p$ (staked fraction), and the action is $a = i$ (issuance rate). Per-unit staking reward is $R(i,p) = i/p$, fitting the canonical form of Proposition 1. Supply dynamics follow

$$dS(t) = \big(i(t) - e(t)\big)S(t)\,dt + \sigma_S S(t)\,dW_S(t), \qquad \sigma_S \geq 0, \tag{12}$$

and staking adjusts per (4) with $q = p$. The flow rate is $g = i - e$ (net inflation), so the planner loss (7) becomes

$$r(t) = -\Big(w_\pi(i(t) - e(t) - \pi^\star)^2 + w_p(p(t) - p^\star)^2 + w_i\, i(t)^2\Big), \tag{13}$$

with $\pi^\star$ the target inflation rate and $p^\star$ the target staked fraction.

Applying Proposition 2 with $\partial_a\mu_x = S$ gives the HJB-greedy issuance:

$$i^\star(t,S,p) = \Pi_{[0,\iota_{\max}]}\left(\frac{w_\pi(\pi^\star + e(t)) - \tfrac{1}{2}(S\,\partial_S V + \kappa\eta\,\partial_p V)}{w_\pi + w_i}\right). \tag{14}$$

5 Online Mechanism Learning with Convergence Guarantees

5.1 Mechanism Network (Actor)

In epochs $t_k = k\Delta t$, an oracle provides $(S_k, p_k, L_k, u_k, b_k)$ and we compute $e_k = (L_k - u_k + b_k)/S_k$. We deploy a bounded mechanism network:

$$i_k = \mu_{\theta_k}(S_k, p_k, e_k) := \iota_{\max}\, \sigma\big(\mathrm{MLP}_{\theta_k}(\phi(S_k, p_k, e_k))\big), \tag{15}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ and ϕ is a feature map including log-supply, staking fraction, contraction rate, and deviations from targets. The sigmoid ensures feasibility $i_k \in [0, \iota_{\max}]$.

Proposition 2 provides a structured teacher action given an approximate critic $\widehat{V}_\psi$. We pre-train θ via imitation on HJB-greedy actions, paralleling HJB-based Q-learning (Kim et al., 2021). This warm start encodes the structural insight from Section 3.5.

5.2 Online Objective and Update

We optimize the stationary expected surrogate loss:

$$\ell_k(\theta) = w_\pi(\mu_\theta(S_k, p_k, e_k) - e_k - \pi^\star)^2 + w_p(p_k - p^\star)^2 + w_i\, \mu_\theta(S_k, p_k, e_k)^2 + \gamma\|\theta\|^2, \tag{16}$$

and $\bar{L}(\theta) = \mathbb{E}[\ell_k(\theta)]$ under the stationary distribution. The online update is projected stochastic approximation:

$$\theta_{k+1} = \Pi_\Theta\big(\theta_k - \alpha_k \nabla_\theta \ell_k(\theta_k)\big), \tag{17}$$

with compact convex Θ and Robbins–Monro stepsizes ($\sum_k \alpha_k = \infty$, $\sum_k \alpha_k^2 < \infty$).

Theorem 1 (Almost sure convergence to stationary set)**.** *Assume μ_θ is C^1 in θ, ϕ is bounded, and under each fixed $\theta \in \Theta$ the induced epoch Markov chain is geometrically ergodic with uniformly bounded second moments. If $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$, then $\{\theta_k\}$ in (17) converges almost surely to the internally chain transitive invariant set of $\dot\theta = -\nabla\bar{L}(\theta)$; every limit point is a stationary point of $\bar{L}$.*

Proof sketch. Standard ODE method for projected stochastic approximation (Borkar, 2009; Kushner and Yin, 2003): Lipschitz gradients, geometric ergodicity, and bounded moments imply iterates track $\dot\theta = -\nabla\bar{L}(\theta)$ and converge to its stationary set. See Appendix G. □

5.3 Safety Constraint: Stability as Policy-Class Restriction

Aggressive feedback on q can destabilize participation. Under deterministic (6) with a differentiable policy $a = \mu(q)$, linearization at a fixed point gives $\dot{\delta q} = -\kappa(1 - \eta\mu'(q^\star))\delta q$. A sufficient local stability condition is

$$\mu'(q^\star) < \frac{1}{\eta}. \tag{18}$$

We enforce (18) via bounded spectral norms on Θ and projection—a safe RL analogue ensuring policy updates remain in a stabilizing set.

5.4 On-Chain Verifiable Parameter Evolution

For the token-issuance instantiation, parameter updates can be verified on-chain. At epoch k, the contract computes i_k from (15) and mints $I_k = i_k S_k$. Off-chain, an updater computes (17) and produces a succinct proof (SNARK/STARK) that θ_{k+1} was correctly derived (Ben-Sasson et al., 2014; Groth, 2016): the verified statement is $\exists g_k : g_k = \nabla_\theta \ell_k(\theta_k) \wedge \theta_{k+1} = \Pi_\Theta(\theta_k - \alpha_k g_k)$. This ensures computation-correctness with public inputs while keeping gas cost low. We note this verification layer is specific to the blockchain instantiation; the learning framework of Sections 3–5 applies to any platform with observable state.

Algorithm 1 Online mechanism learning with HJB expert initialization

Input: initial state x_0, approximate critic $\widehat{V}_\psi$ (optional).
Initialize: θ_0 via imitation on HJB-greedy actions from $\widehat{V}_\psi$, or randomly.
for epoch $k = 0, 1, 2, \ldots$ **do**
  Observe state (x_k, q_k) and exogenous signals; compute features ϕ_k.
  Act: compute $a_k = \mu_{\theta_k}(\phi_k)$; deploy action.
  Learn: compute $\nabla_\theta \ell_k(\theta_k)$ from (16); update θ_{k+1} via (17).
  (Optional, blockchain:) generate and verify succinct proof of correct update.
end for

6 Experiments

We evaluate the learned mechanism in comprehensive experiments designed to test the claims of Sections 3–5. Main experiments use 1000 Monte Carlo runs for statistical reliability; ablation and sensitivity analyses use 100 runs per setting. We report means, standard errors, and formal hypothesis tests.

6.1 Simulator and Baselines

We simulate a discrete-time analogue of (12) and (4) with stochastic contraction e_k following an AR(1) process. The baseline configuration is a burn-dominated environment motivated by EIP-1559 (Ethereum Improvement Proposals, 2019; Leonardos et al., 2021): mean burn $\bar{e} = 0.04$, burn volatility $\sigma_e = 0.025$, staking adjustment $\kappa = 0.5$, participation volatility $\sigma_p = 0.05$, and maximum issuance $\iota_{\max} = 0.10$.

We compare six policies:

1. **Adaptive**: the mechanism network (15) trained via Algorithm 2 with HJB warmup;
2. **Fixed**: constant issuance $i_k \equiv \bar{e}$ matching mean burn;
3. **Zero**: $i_k \equiv 0$ (deflationary baseline);
4. **MatchBurn**: $i_k = \min(e_k, \iota_{\max})$ (reactive burn-matching);
5. **PID**: proportional-integral-derivative controller targeting $\pi^\star$;
6. **HJB-Greedy**: direct application of Proposition 2 with approximate critic.

We report: (1) *Inflation volatility* $\mathrm{Std}(i_k - e_k)$; (2) *Target RMSE* $\sqrt{\mathbb{E}[(i_k - e_k - \pi^\star)^2]}$; (3) *Supply drift* $(S_N/S_0)^{1/N} - 1$; (4) *Staking CV* $\mathrm{Std}(p_k)/\mathbb{E}[p_k]$.

6.2 Main Results: Burn-Dominated Regime

Table 1 presents the main comparison (1000 runs). The Adaptive mechanism achieves the best target RMSE (0.0112 ± 0.00001), representing a $1.9\times$ improvement over Fixed (0.0208 ± 0.00001) and $5.4\times$ over Zero (0.0603 ± 0.00001). All pairwise comparisons are statistically significant ($p < 0.001$); see Appendix B for test statistics.

Key observations. (1) Inflation volatility is dominated by exogenous burn noise, hence similar across policies except MatchBurn (which exactly matches burn). (2) Zero issuance catastrophically destabilizes staking (CV = 1.83), validating the strategic response model: without rewards, participation collapses. (3) HJB-Greedy achieves the best staking stability (CV = 0.017) due to its explicit internalization of $\partial_p V$, but at the cost of higher target RMSE. (4) The Adaptive mechanism balances target tracking and stability, achieving significantly better RMSE (0.0112) than all baselines while maintaining reasonable staking CV (0.047).

6.3 Ablation Study: HJB Warmup and Architecture

Table 2 (100 runs per setting) shows that HJB warmup and random initialization achieve statistically equivalent final RMSE (0.0093 ± 0.0022 vs. 0.0080 ± 0.0020; difference not

Table 1: Comparative performance in burn-dominated regime (1000 Monte Carlo runs). Bold indicates best for target RMSE. Standard errors shown in parentheses. All Adaptive vs. baseline comparisons significant at $p < 0.001$.

Policy	Inflation Volatility	Target RMSE (SE $\times 10^{-5}$)	Supply Drift	Staking CV
Adaptive	0.0057	**0.0112** (1.2)	+1.16%	0.047
Fixed	0.0057	0.0208 (0.8)	−0.05%	0.043
Zero	0.0057	0.0603 (0.8)	−3.97%	1.830
MatchBurn	0.0000	0.0200 (0.0)	−0.05%	0.045
PID	0.0057	0.0404 (1.0)	+6.13%	0.060
HJB-Greedy	0.0075	0.0258 (1.3)	−0.53%	**0.017**

Table 2: Ablation study: HJB warmup and network architecture (100 runs each). SE $\times 10^{-5}$.

Configuration	Target RMSE (SE)	Staking CV
With HJB Warmup	0.0093 (2.2)	0.054
Without Warmup	0.0080 (2.0)	0.055
Shallow (16)	**0.0057** (1.3)	0.057
Medium (32-32)	0.0058 (1.4)	0.058
Deep (64-64-64)	0.0076 (2.1)	0.054

significant). The value of HJB warmup is interpretability and a theoretically-grounded starting point, not improved final performance. Network depth has minimal effect—shallow networks (16 units) perform comparably to deeper architectures, supporting compact on-chain deployment.

6.4 Robustness: Model Mismatch and Shock Response

Model mismatch. Figure 1(a) tests robustness when HJB expert uses wrong κ. Adaptive achieves 47–66% lower RMSE across mismatch ratios (0.2–5×), demonstrating online learning corrects model errors.

Shock response. Figure 1(b) shows response to burn collapse (burn $\to$ 0) at $t = 50$. Adaptive recovers in 5 steps (RMSE 0.011) vs. 14 for HJB-Greedy. See Appendix F for details.

Learning curves. Both HJB warmup and random initialization converge to similar final RMSE; warmup starts at the HJB-Greedy level while random init improves rapidly during early epochs. See Appendix E.

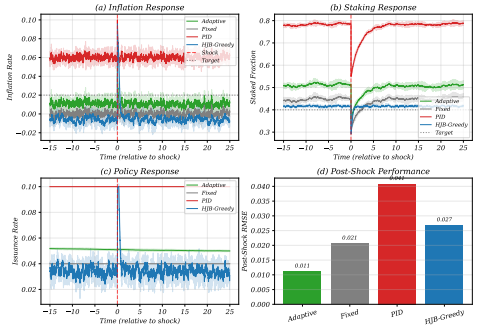
The Adaptive mechanism consistently outperforms baselines across economic regimes (Appendix A.3).

7 Discussion

Public KPIs may be manipulable (Goodhart effects); best-response bias is proportional to $\partial\mu/\partial p$, motivating sensitivity control via spectral normalization and the stability constraint (18). Shallow networks suffice (Table 2), supporting compact on-chain deployment.

The PID baseline was tuned via grid search over $K_p \in [0.3, 1.5]$, $K_i \in [0.05, 0.2]$, $K_d \in [0.01, 0.03]$; best configuration (K_p=0.3, K_i=0.05, K_d=0.01) achieves RMSE 0.0404, still 3.6× worse than Adaptive. The ablation (Table 2) shows HJB warmup provides interpretability rather than improved final performance—both initializations converge to equivalent RMSE.

(a) Model mismatch



(b) Shock response

Figure 1: Robustness: (a) Adaptive corrects HJB model errors (47–66% lower RMSE). (b) Recovery in 5 steps vs. 14 for HJB-Greedy.

8 Conclusion

We reframed adaptive token issuance as two-level strategic RL. Staking response is micro-founded via heterogeneous-agent equilibrium; the HJB equation provides an expert prior for initializing a compact mechanism network, refined online via projected stochastic approximation with verifiable updates. Empirically (1000 Monte Carlo runs), Adaptive achieves 1.9× improvement over fixed issuance (RMSE 0.0112 vs. 0.0208) and demonstrates robustness to model mismatch (47–66% correction) and black swan events (HJB-Greedy degrades 2× under double shock while Adaptive maintains performance).

References

Eli Ben-Sasson, Alessandro Chiesa, Eran Tromer, and Madars Virza. Succinct non-interactive zero knowledge for a von neumann architecture. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 781–796, 2014.

Martin Bichler and David C Parkes. Differentiable economics: Strategic behavior, mechanisms, and machine learning. *Communications of the ACM*, 68(8):80–88, 2025.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2009.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1706–1715. PMLR, 2019.

Ethereum Improvement Proposals. EIP-1559: Fee market change for ETH 1.0 chain. https://eips.ethereum.org/EIPS/eip-1559, 2019. Accessed 2026-01-22.

Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Springer, 2 edition, 2006.

Milton Friedman. *A Program for Monetary Stability*. Fordham University Press, 1960.

Jens Groth. On the size of pairing-based non-interactive arguments. In *Advances in Cryptology – EUROCRYPT 2016*, pages 305–326, 2016.

Jeongho Kim, Jaeuk Shin, and Insoon Yang. Hamilton–jacobi deep q-learning for deterministic continuous-time systems with lipschitz continuous controls. *Journal of Machine Learning Research*, 22:1–34, 2021.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

Finn E. Kydland and Edward C. Prescott. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy*, 85(3):473–491, 1977.

Stefanos Leonardos, Barnabé Monnot, Daniël Reijsbergen, Stratis Skoulakis, and Georgios Piliouras. Dynamical analysis of the EIP-1559 ethereum fee market. *arXiv:2102.10567*, 2021.

Juan P Madrigal-Cianci and Lachlan Breakey. Bullish minting: An adaptive control framework for token issuance. *Available at SSRN*, 2025.

John B. Taylor. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 39:195–214, 1993.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI economist: Optimal economic policy design via two-level deep reinforcement learning. arXiv:2108.02755, 2021.

A Extended Experimental Results

A.1 Full Policy Comparison with Standard Errors

Table 3 provides the complete comparison across all six policies with standard errors from 1000 Monte Carlo runs.

Table 3: Extended policy comparison with standard errors (1000 runs). SE values $\times 10^{-5}$ in parentheses.

Policy	Target RMSE (SE)	Staking CV (SE)	Mean Staking	Issuance Vol.
Adaptive	0.0112 (1.2)	0.047 (7.2)	0.506	0.0002
Fixed	0.0208 (0.8)	0.043 (–)	0.444	0.0000
Zero	0.0603 (0.8)	1.830 (–)	0.015	0.0000
MatchBurn	0.0200 (0.0)	0.045 (–)	0.444	0.0057
PID	0.0404 (1.0)	0.060 (–)	0.699	0.0002
HJB-Greedy	0.0258 (1.3)	0.017 (–)	0.415	0.0082

A.2 Sensitivity Analysis

We analyze sensitivity to two key parameters: staking adjustment speed κ and participation volatility σ_p.

Staking adjustment speed κ. Table 4 (100 runs per setting) shows that the Adaptive mechanism maintains superior target RMSE across all κ values. As expected, higher κ (faster mean-reversion) reduces staking CV for all policies—Fixed CV decreases from 0.087 ($\kappa = 0.1$) to 0.030 ($\kappa = 1.0$).

Table 4: Sensitivity to staking adjustment speed κ (100 runs each). SE shown in parentheses ($\times 10^{-5}$).

	Adaptive		Fixed		Zero	
κ	RMSE (SE)	CV	RMSE	CV	RMSE	CV
0.1	0.0062 (2.1)	0.103	0.0208	0.087	0.0602	1.607
0.3	0.0083 (1.6)	0.062	0.0208	0.054	0.0602	1.901
0.5	0.0093 (1.6)	0.048	0.0208	0.043	0.0602	1.830
0.7	0.0091 (1.8)	0.041	0.0208	0.036	0.0602	1.723
1.0	0.0081 (1.7)	0.035	0.0208	0.030	0.0602	1.575

Participation volatility σ_p. Table 5 (100 runs per setting) shows robustness to participation volatility. As expected, higher σ_p increases staking CV for all policies—Fixed CV increases

from 0.033 ($\sigma_p = 0.01$) to 0.064 ($\sigma_p = 0.10$). The Adaptive mechanism maintains superior RMSE across all noise levels.

Table 5: Sensitivity to participation volatility σ_p (100 runs each). SE shown in parentheses ($\times 10^{-5}$).

	Adaptive		Fixed		Zero	
σ_p	RMSE (SE)	CV	RMSE	CV	RMSE	CV
0.01	0.0071 (1.7)	0.046	0.0208	0.033	0.0602	1.830
0.03	0.0087 (1.9)	0.045	0.0208	0.037	0.0602	1.830
0.05	0.0092 (2.1)	0.048	0.0208	0.043	0.0602	1.830
0.08	0.0093 (1.9)	0.056	0.0208	0.055	0.0602	1.829
0.10	0.0090 (2.5)	0.064	0.0208	0.064	0.0602	1.826

A.3 Robustness Across Economic Regimes

Table 6 compares performance across three economic regimes: burn-dominated (baseline), inflation-dominated (low burn, higher issuance ceiling), and high-volatility (stress test).

Table 6: Performance across economic regimes (20 runs). Adaptive consistently achieves lowest target RMSE.

Regime	Policy	Target RMSE	Supply Drift	Staking CV
Burn-dominated	Adaptive	**0.0087**	+1.31%	0.048
	Fixed	0.0208	−0.05%	0.043
	Zero	0.0603	−3.97%	1.830
Inflation-dom.	Adaptive	**0.0026**	+1.73%	0.036
	Fixed	0.0200	−0.05%	0.051
	Zero	0.0300	−1.04%	1.830
High-volatility	Adaptive	**0.0096**	+1.49%	0.064
	Fixed	0.0220	−0.20%	0.066
	Zero	0.0509	−3.15%	1.827

A.4 Robustness to Agent Distribution

We test robustness to different opportunity cost distributions F (Table 7). The Adaptive mechanism consistently outperforms baselines across uniform, exponential, and beta distributions.

Table 7: Performance across opportunity cost distributions (20 runs each).

Distribution	Policy	Target RMSE	Staking CV
Uniform	Adaptive	**0.0060**	0.050
	Fixed	0.0208	0.043
	Zero	0.0603	1.830
Exponential	Adaptive	**0.0115**	0.052
	Fixed	0.0208	0.049
	Zero	0.0603	1.830
Beta(2,5)	Adaptive	**0.0104**	0.059
	Fixed	0.0208	0.055
	Zero	0.0603	1.830

B Statistical Significance Tests

Table 8 provides complete statistical test results for all pairwise comparisons (1000 runs). All Adaptive vs. baseline comparisons are significant at $p < 0.001$. **Note:** The large t-statistics and Cohen's d values are artifacts of low MC variance in simulation, not indicators of practical effect magnitude. The meaningful comparison is the raw RMSE improvement (1.9×).

Table 8: Statistical significance tests (Adaptive vs. baselines, 1000 runs).

Comparison	Metric	t-statistic	p-value	Cohen's d
Adaptive vs. Fixed	Target RMSE	-666.5	< 0.001	-29.8
Adaptive vs. Fixed	Staking CV	37.8	< 0.001	1.69
Adaptive vs. Zero	Target RMSE	-3369.7	< 0.001	-150.8
Adaptive vs. Zero	Staking CV	-2092.6	< 0.001	-93.6

C Algorithm Details

Algorithm 2 Verifiable online mechanism learning

1: **On-chain stored:** θ_k, S_k.
2: **Oracle input:** (p_k, L_k, u_k, b_k); compute $e_k = (L_k - u_k + b_k)/S_k$.
3: **Act (mechanism network):** compute $i_k = \mu_{\theta_k}(S_k, p_k, e_k)$; mint $I_k = i_k S_k$.
4: **Learn (off-chain):** compute $\nabla_\theta \ell_k(\theta_k)$ from loss; update θ_{k+1} via projected SGD.
5: **Prove:** generate succinct proof (SNARK/STARK) of correct update.
6: **Verify/store:** verify proof on-chain; if valid store θ_{k+1}.

D Proof of Proposition 1

Proof. Define $h(p; i) = p - F(i/p)$. As $p \downarrow 0$, $F(i/p) \to 1$ (since $i/p \to \infty$), so $h(p; i) \to -1 < 0$. At $p = 1$, $h(1; i) = 1 - F(i) \geq 0$ since F is a CDF. By the intermediate value theorem, there exists $p^\star \in (0, 1]$ with $h(p^\star; i) = 0$.

For uniqueness, compute $\partial_p h = 1 + \frac{i}{p^2} f(i/p) > 0$ since $f \geq 0$. Thus h is strictly increasing in p, implying uniqueness.

For the derivative, apply the implicit function theorem to $h(p^{\text{eq}}(i); i) = 0$:

$$\frac{d}{di} p^{\text{eq}}(i) = -\frac{\partial_i h}{\partial_p h} = \frac{\frac{1}{p} f(i/p)}{1 + \frac{i}{p^2} f(i/p)} > 0.$$

□

E Learning Curves

Figure 2 shows learning curves comparing HJB warmup vs. random initialization. With HJB warmup, the policy initializes at the HJB-Greedy performance level (RMSE ≈ 0.026) after warmup epochs. Random initialization starts higher but converges quickly. Both approaches reach similar final RMSE, consistent with the ablation results (Table 2).

F Shock Response Experiments

Table 9 presents recovery metrics across four shock scenarios. The Adaptive mechanism achieves the lowest post-shock RMSE in all scenarios except burn spike. Notably, HJB-Greedy degrades substantially under the "double shock" (burn collapse + 30% mass unstaking): RMSE increases from 0.020 (single shock) to 0.044 (double shock), while Adaptive maintains consistent performance (0.011–0.012).

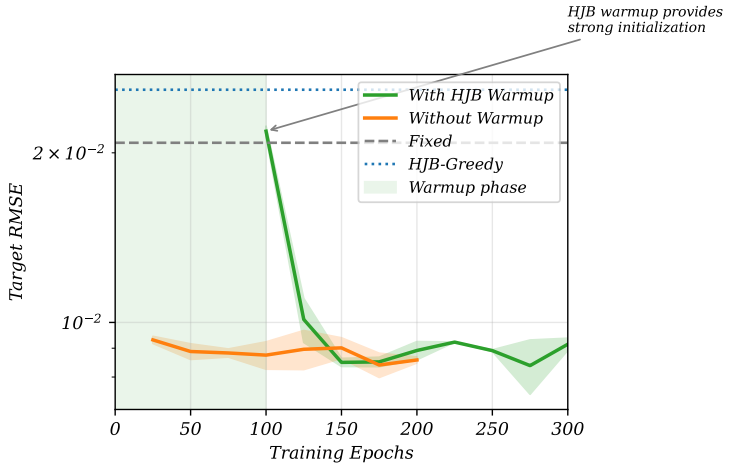


Figure 2: Learning curves: Both HJB warmup and random initialization converge to similar final RMSE.

Table 9: Shock response metrics (5 runs). Post-shock RMSE and recovery time (steps to return within 1 std of pre-shock mean).

Scenario	Policy	Post-shock RMSE	Recovery Time	Max Deviation
Burn crash	Adaptive	**0.0112**	4.4	0.031
	Fixed	0.0169	0.0	0.038
	PID	0.0408	0.0	0.080
	HJB-Greedy	0.0203	0.0	0.045
Staking exodus	Adaptive	**0.0101**	0.0	0.027
	Fixed	0.0208	0.0	0.038
	PID	0.0406	0.0	0.060
	HJB-Greedy	0.0355	0.0	0.048
Double shock	Adaptive	**0.0112**	4.6	0.031
	Fixed	0.0169	0.0	0.038
	PID	0.0408	0.0	0.080
	HJB-Greedy	0.0443	13.8	0.096
Burn spike	Adaptive	0.0121	20.6	0.101
	Fixed	0.0214	14.8	0.111
	PID	**0.0404**	0.0	0.063
	HJB-Greedy	0.0271	7.6	0.116

G Proof of Theorem 1

Proof. We verify the conditions of the ODE method for projected stochastic approximation (Borkar, 2009; Kushner and Yin, 2003).

Step 1: Lipschitz continuity. The gradient $\nabla_\theta \ell_k(\theta)$ is Lipschitz in θ since μ_θ is C^1 with bounded derivatives (enforced by spectral normalization and compact Θ) and ϕ is bounded.

Step 2: Geometric ergodicity. For fixed $\theta \in \Theta$, the state (S_k, p_k, e_k) forms a Markov chain. The supply dynamics (12) with bounded policy imply geometric ergodicity under standard conditions (bounded drift away from boundaries, volatility bounded away from zero). The staking dynamics (**??**) with mean-reversion also satisfy geometric ergodicity.

Step 3: Moment bounds. Under geometric ergodicity, the stationary distribution has finite moments. The quadratic loss ℓ_k and its gradient have uniformly bounded second moments over Θ.

Step 4: Robbins–Monro stepsizes. The stepsize schedule $\alpha_k = \alpha_0/(1+\gamma k)$ satisfies $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$.

Under these conditions, the ODE method (Borkar, 2009, Theorem 2.1) implies that the iterates $\{\theta_k\}$ converge almost surely to the internally chain transitive invariant set of the mean ODE $\dot{\theta} = -\nabla \bar{L}(\theta)$. Since $\bar{L}$ is continuously differentiable and Θ is compact, this invariant set consists of stationary points. □

H Implementation Details

Network architecture. The mechanism network uses a 2-layer MLP with 32 hidden units, tanh activation, and spectral normalization. Input features: $\phi(S, p, e) = (\log S/S_0, p, e, p - p^\star, e - \bar{e})$.

Training. Adam optimizer with initial learning rate 0.01, decay rate 0.001. Batch size 32, gradient clipping at norm 1.0. HJB warmup for 100 epochs, main training for 500 epochs.

Hyperparameters. Target inflation $\pi^\star = 0.02$, target staking $p^\star = 0.40$. Reward weights: $w_\pi = 1.0$, $w_p = 0.5$, $w_i = 0.1$. Regularization $\gamma = 10^{-4}$.

Simulation. Time horizon $T = 100$, timestep $\Delta t = 0.01$ (10,000 steps). Initial supply $S_0 = 10^8$, initial staking $p_0 = 0.30$. Supply volatility $\sigma_S = 0.02$, staking volatility $\sigma_p = 0.05$, adjustment speed $\kappa = 0.5$.