Stabilizing Dynamic Contracts in Infinite Horizon via Entropy-Regularized Subgame-Perfect Equilibrium

Liz Lemma Future Detective

January 16, 2026

Abstract

Principal–agent reinforcement learning provides an economic language for orchestrating autonomous AI agents with costly interventions (payments) rather than unrealistic centralized reward shaping. Recent work shows that in finite-horizon principal–agent MDPs, an alternating principal/agent optimization meta-algorithm converges to subgame-perfect equilibrium (SPE), but it can diverge in infinite horizon under hidden actions due to best-response discontinuities. We propose a clean stabilization: entropy-regularize both (i) the agent's best response over actions and (ii) the principal's optimization over contracts. This converts knife-edge incentive constraints into smooth, probabilistic response curves and yields a single joint soft Bellman operator. We prove this operator is a γ-contraction in sup norm for any $\gamma \in (0, 1)$, implying existence and uniqueness of a stationary regularized SPE (rSPE) and global convergence of value iteration/soft Q-learning to rSPE in infinite horizon. We further quantify how rSPE approximates unregularized SPE: the principal's value loss relative to the best unregularized SPE is bounded by $(\tau_p \log |\mathcal{B}| + \tau_a \log |\mathcal{A}|)/(1-\gamma)$, delivering an explicit stability–optimality tradeoff controlled by temperatures τ_p, τ_a. Finally, we provide cycling counterexamples showing why some smoothing is necessary and empirically stress-test the stabilized algorithm on known cycling instances. The result is a deployable 'stability knob' for always-on 2026 agent ecosystems, where institutions must be learned and robust under approximation error.

Table of Contents

with regret/bandit approaches.

3. 3. Baseline model (unregularized): hidden-action principal–agent MDP; why SPE is natural; recap of divergence in infinite horizon (cycling) and the root cause (discontinuous best response).

4. 4. Regularized model: define entropy-regularized agent and entropy-regularized principal; define rSPE; discuss interpretation (bounded rationality, smoothing, commitment/noise).

5. 5. Main theorem: contraction and uniqueness in infinite horizon; precise operator definition; statement of fixed point and policy reconstruction.

6. 6. Approximation-to-unregularized SPE: value gap bounds as $\tau_a, \tau_p \to 0$; stability–optimality frontier; when the bound is tight; discussion of what is and is not guaranteed (selection among multiple SPEs).

7. 7. Necessity and limitations: counterexamples without regularization; what breaks if $\tau_a = 0$ or $\tau_p = 0$; discussion of continuous contract spaces (need numerical integration or discretization).

8. 8. Algorithms: soft Q-learning instantiation; computational considerations; when LP/minimal-implementation is still needed vs replaced by regularized selection.

9. 9. Experiments: (i) reproduce cycling instance and show stabilization; (ii) stress tests under approximation error; (iii) optional large-MDP demonstration with function approximation (flagging reliance on numerics).

10. 10. Conclusion and open problems: partial observability, robust/DR variants, multi-agent extensions, and how to tune τ in deployment.

1 Introduction

Digital contracting is rapidly becoming an *always-on* problem. In 2026 it is increasingly common for principals—platforms, marketplaces, employers, insurers, or automated procurement systems—to interact repeatedly with adaptive agents whose effort and choices are only imperfectly observed, but whose outputs are verifiable and contractible. Recommendation partners are paid on conversions rather than on unobservable targeting effort; logistics providers are rewarded for on-time delivery rather than for internal routing decisions; content moderation vendors are paid on measured outcomes rather than on hidden screening intensity. In these settings, a contract is not a one-shot document but a policy: it is selected again and again, in response to a changing environment, and it must remain operational under continuous deployment.

This simple shift from episodic interaction to persistent interaction raises an old but consequential question: *what stabilizes incentives over time when actions are hidden*? Classical principal–agent theory teaches us how to design incentives when the informational constraint is binding. Yet much of that theory is built around a finite horizon or a static environment, where the contract is chosen once, the agent responds, and the game ends. Always-on systems do not end. They evolve, and today they evolve under algorithmic control. Contracts are updated, agents learn about the mapping from outcomes to payments, and states transition as a function of realized outcomes. A good model for these ecosystems must therefore address not only incentive compatibility at a point in time, but also the *dynamic stability* of the contracting process itself.

Finite-horizon analyses can be misleading precisely because they inherit stability from the clock. With a terminal date, backward induction often selects a well-defined equilibrium path even when incentives are brittle along the way. That discipline is useful for many applications, but it is an inadequate guide for environments where the interaction is intended to persist indefinitely and where what matters is a stationary pattern of contracting and behavior. In an infinite horizon, the principal must internalize that today's contract changes tomorrow's state distribution; the agent must internalize that today's action affects continuation payoffs through future contract offers. These feedback loops are not a technical nuance: they determine whether a deployed contracting system converges to a predictable regime or instead oscillates across qualitatively different incentive schemes.

The practical motivation is mirrored by a computational one. In modern deployments the principal does not solve a model once; it runs an update rule that iteratively improves a contract policy from data. Even if an equilibrium exists in principle, the relevant question is whether a plausible learning or planning procedure will *find* it. Hidden action is especially challenging here: small changes in a contract can induce discrete switches in the agent's best

response, which in turn can cause discontinuous jumps in outcomes and state transitions. Such discontinuities are innocuous in static comparative statics but can be fatal for iterative methods, producing cycling behavior and non-convergence. In an always-on setting, non-convergence is not merely an analytical inconvenience; it is operational instability, with real costs for welfare, safety, and compliance.

We therefore advocate a modeling move that is at once economically interpretable and algorithmically powerful: *entropy regularization* for both the agent's action choice and the principal's contract choice. Economically, entropic terms capture bounded rationality, idiosyncratic preference shocks, experimentation, or deliberate randomization to avoid being gamed. Algorithmically, they replace hard maximization with a smooth log-partition operator, turning set-valued best responses into single-valued, continuous mappings. The resulting equilibrium concept—a regularized subgame-perfect equilibrium—can be viewed as the outcome of a long-run contracting problem in which both parties trade off expected payoff against a desire for stochasticity (or, equivalently, a cost of precision in optimization).

The key conceptual payoff is that regularization becomes a *stability knob*. Temperatures govern how sharply each party concentrates on its currently best option. When these temperatures are strictly positive, the induced dynamic programming operators inherit contraction properties familiar from discounted control: continuation values are down-weighted by the discount factor, and the smooth maximization step is non-expansive. Put differently, smoothing restores the kind of global stability that discounted Markov decision processes enjoy, but that hidden-action principal–agent problems can lose when best responses are discontinuous. This stability knob has a transparent tradeoff: higher temperatures increase stability and robustness but introduce bias relative to the unregularized benchmark in which each party chooses a deterministic best response.

This tradeoff is central for policy and practice. Designers of contracting protocols often face a choice between aggressively exploiting the currently estimated best contract and maintaining exploration, robustness, and predictability. A platform that re-optimizes payments every hour based on noisy metrics may inadvertently create whiplash incentives that agents learn to exploit; a regulator may prefer a mechanism that is slightly less sharp but more reliable under misspecification. Our framework makes this tension explicit. By parameterizing smoothness, we can ask how much long-run value is sacrificed for stability and uniqueness, and how that sacrifice scales with discounting and the size of the action and contract sets.

At the same time, we are candid about what regularization does *not* solve. First, it is not a substitute for identification: if outcomes are uninformative about hidden actions, no amount of smoothing creates incentives out of thin air. Second, temperatures are not primitives in most economic environments; they summarize behavioral noise or an algorithm designer's

choice, and must be calibrated or justified. Third, stability in the tabular, finite setting does not automatically extend to large-scale function approximation, where additional sources of nonlinearity can reintroduce instability. Our goal is therefore not to claim that entropy regularization is a universal remedy, but to show that it delivers a clean and tractable baseline in which infinite-horizon contracting is well-posed and computationally meaningful.

The remainder of the paper builds on this intuition in a disciplined way. We model the principal's contract as a state-dependent policy over a finite set of outcome-contingent payment schemes, while the agent selects hidden actions after observing the state and the offered contract. Outcomes are observable and contractible; actions are not. Both parties evaluate discounted payoffs and face entropy terms that induce stochastic choice. Within this structure we obtain three properties that are particularly valuable for always-on ecosystems: (i) a unique stationary equilibrium induced by smooth best responses, (ii) global convergence of natural dynamic programming iterations to that equilibrium, and (iii) an explicit bound quantifying how close the regularized equilibrium value is to the best unregularized stationary benchmark as temperatures approach zero. These properties jointly formalize the idea that regularization is a controllable route from brittle but sharp incentives to stable and learnable contracting dynamics.

With this motivation in place, we next situate our approach within the literatures on principal–agent theory, reinforcement-learning-based contract design, entropy-regularized control, and equilibrium learning, emphasizing how the infinite-horizon hidden-action setting changes both the economic questions and the algorithmic requirements.

2 Related work

Our framework sits at the intersection of four literatures: (i) classical principal–agent theory, (ii) contract and mechanism design with reinforcement learning and dynamic optimization, (iii) entropy-regularized control and its economic interpretations, and (iv) equilibrium computation and learning in dynamic games. We highlight where our modeling choices follow established practice and where the infinite-horizon hidden-action setting forces different technical and conceptual emphases.

Principal–agent theory and dynamic incentives. The core economic friction we study—unobserved actions with contractible outcomes—is the canonical moral hazard problem, developed in static and repeated settings in foundational work such as **??**. A large subsequent literature studies dynamic moral hazard and relational contracting, where continuation values provide incentives and contracts can depend on histories or promised utilities (e.g., **?**). Those models emphasize commitment, history dependence,

and the rich structure of optimal contracts under persistent private information. Our focus is deliberately different. We restrict attention to Markov (state-dependent) contracts and policies in a finite MDP, not because history dependence is unimportant, but because modern "contracting policies" deployed in platforms and automated procurement are often implemented as stationary rules that map observables to transfers. This restriction allows us to treat the contracting problem as a dynamic system and to ask when natural iterative procedures converge to a well-defined stationary regime.

A second difference concerns equilibrium selection. Many economic treatments of dynamic moral hazard appeal to optimality within a class of incentive-compatible contracts, often derived via a principal's Bellman equation over promised utilities. Here, because actions are hidden and contracts are chosen repeatedly, the relevant object is a *subgame-perfect* solution concept in the underlying stochastic game induced by the contract choice and the agent's subsequent action. In finite horizons, backward induction provides discipline; in infinite horizons, stationary equilibria may be multiple and best-response maps can be discontinuous. Our contribution is to show that a minimal and economically interpretable smoothing device turns this potentially ill-posed equilibrium selection problem into a globally stable fixed-point problem.

Contract design in reinforcement learning and dynamic decision systems. A growing literature in computer science and adjacent areas studies contract design when the principal interacts with a learning or adaptive agent, often motivated by markets, crowdsourcing, and platform incentive design. One branch formulates *bilevel* problems in which the principal optimizes a reward or payment rule anticipating an agent that solves an MDP or a bandit problem (e.g., "reinforcement learning from incentives" and "policy design" formulations). These approaches often assume a Stackelberg structure: the principal commits to a contract (or a parameterized reward function), the agent computes a best-response policy, and the principal evaluates outcomes. This is analytically convenient and captures settings where commitment is credible and contract changes are infrequent.

In contrast, our environment is intrinsically *always-on*: the principal selects contracts repeatedly as a function of state, and the agent's continuation payoff depends on the principal's future contract policy. This pushes the appropriate equilibrium notion closer to Markov perfect equilibrium or subgame-perfect equilibrium in stochastic games than to one-shot Stackelberg commitment. Put differently, the principal cannot be treated as choosing a single contract parameter once; instead, the principal is itself a dynamic optimizer whose policy is part of equilibrium. This distinction matters because it changes the shape of the principal's optimization: the objective is not simply to pick a contract that induces a desired stationary agent policy under fixed continuation values, but to pick a *contract policy* whose induced

state distribution feeds back into future incentives.

We also differ from regret-minimization and bandit-based contracting models that emphasize learning transfers under uncertainty about the agent's type or response function. Such models often study myopic environments (no state transitions) or treat each round as independent conditional on the contract, yielding regret bounds for exploration–exploitation in contract space. Those tools are valuable when the principal must learn the mapping from payments to outcomes, but they typically abstract from the equilibrium feedback loop created by state dynamics and strategic adaptation over time. Our analysis instead takes the dynamic structure seriously and asks for conditions under which the natural dynamic programming iterations are stable. The resulting value-loss bound in temperatures complements regret-style bounds by quantifying a different tradeoff: not exploration for identification, but smoothness for equilibrium uniqueness and computational robustness.

Entropy-regularized control, bounded rationality, and logit choice. Entropy regularization has become a standard device in modern reinforcement learning, particularly in maximum-entropy RL and "soft" dynamic programming (e.g., **??**). Technically, replacing a hard max by a log-sum-exp yields a smooth Bellman operator with strong contraction properties in discounted settings, and the induced policies take a Gibbs (softmax) form. Economically, the same structure appears in quantal response and logit choice models, where entropy terms arise from additive i.i.d. extreme-value shocks or bounded rationality **?**. It also connects to rational inattention and control-cost interpretations, where entropy captures the cost of precision in selecting actions **?**.

We build directly on these insights, but apply them in a two-player principal–agent setting with hidden action. The novelty is not the softmax formula per se; it is the way regularization resolves a specific pathology of dynamic moral hazard: small contract changes can induce discontinuous changes in the agent's optimal action, which can in turn create discontinuous changes in the induced outcome distribution and state transitions. In such a setting, the principal's effective objective can be non-smooth even with discounting, and alternating best-response dynamics can fail to converge. By regularizing *both* the agent's response and the principal's contract selection, we obtain a single-valued joint operator whose fixed point defines a regularized subgame-perfect equilibrium. The temperatures τ_a and τ_p then play a dual role: they are interpretable as noise or optimization frictions, and they serve as explicit stability parameters in the equilibrium mapping.

Equilibrium learning and computation in dynamic games. Finally, our analysis relates to the literature on learning equilibria in Markov games and stochastic games, including value-iteration-like methods, actor–critic

schemes, and two-timescale stochastic approximation. In general-sum Markov games, equilibrium computation is challenging: Bellman operators need not be contractions, equilibria may be non-unique, and naive best-response dynamics can cycle. A parallel stream shows that entropy regularization can improve stability in multi-agent RL by smoothing best responses and enabling convergence guarantees under appropriate conditions (e.g., regularized Markov perfect equilibria and entropy-regularized policy gradients).

Our contraction-based result can be viewed as an instance of this broader principle specialized to the principal–agent timing and information structure. The hidden-action feature is not merely decorative: because the principal does not observe actions, its continuation value depends on the agent's *policy* only through outcomes, which introduces an additional layer where discontinuities can arise. The payoff from our formulation is that, in the finite tabular setting, we recover the familiar discounted-control geometry: the regularized joint Bellman operator is a γ-contraction in $\|\cdot\|_\infty$, implying a unique fixed point and global convergence of iterates. This provides a clean benchmark for "equilibrium learning" in contracting problems and clarifies what breaks when the regularization vanishes.

These comparisons motivate the baseline model we study next. We first present the unregularized hidden-action principal–agent MDP and the natural stationary subgame-perfect equilibrium notion, and then illustrate why infinite-horizon dynamics can be unstable when best responses are discontinuous. The regularized model can then be read as the minimal modification that restores well-posedness while keeping the contracting primitives economically transparent.

3 Baseline model (unregularized): hidden-action principal–agent MDP

We begin from a minimal infinite-horizon contracting environment in which incentives must be provided through outcomes rather than actions. Time is discrete and the economy is summarized by a finite state $s \in \mathcal{S}$. In each period, the principal first chooses a contract $b \in \mathcal{B} \subset \mathbb{R}_{\geq 0}^{|\mathcal{O}|}$, where $b(o)$ is the transfer paid if outcome $o \in \mathcal{O}$ is realized. After observing (s, b), the agent privately chooses an action $a \in \mathcal{A}$. The outcome is then drawn according to $o \sim O(\cdot \mid s, a)$, the transfer $b(o)$ is paid, and the next state evolves as $s' \sim T(\cdot \mid s, o)$. The key informational friction is that the principal never observes a; the only contractible and publicly observed signal of effort is o (and, through T, the induced state transition).

Per-period payoffs are

$$R_a(s, b, a, o) = r(s, a) + b(o), \qquad R_p(s, b, o) = r_p(s, o) - b(o),$$

with bounded primitives and common discount factor $\gamma \in (0, 1)$. This timing

captures an "always-on" relationship: contracts are not committed once and for all, but are selected repeatedly as a function of observable conditions, and the agent understands that today's action affects future states and therefore future contracts. Because outcomes are observed and payments executed each period, the public history is well-defined even though actions are hidden. The substantive question is therefore not whether incentives can be provided at all, but whether the dynamic feedback between future promised transfers and current hidden actions yields a well-posed stationary equilibrium object.

Stationary Markov strategies and induced values. We restrict attention to stationary Markov rules. A (possibly mixed) principal strategy is a kernel $\rho(\cdot \mid s) \in \Delta(\mathcal{B})$ mapping states to distributions over contracts, and the agent strategy is $\pi(\cdot \mid s, b) \in \Delta(\mathcal{A})$ mapping observed (s, b) into distributions over actions. Given (ρ, π), the resulting controlled Markov chain on $\mathcal{S}$ is determined by the composition of ρ, π, O, and T. The principal's discounted value from state s is

$$V_p^{\rho,\pi}(s) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \Big(r_p(s_t, o_t) - b_t(o_t)\Big) \,\middle|\, s_0 = s\right],$$

with an analogous definition $V_a^{\rho,\pi}(s)$ for the agent, replacing $r_p(s_t, o_t) - b_t(o_t)$ by $r(s_t, a_t) + b_t(o_t)$. We emphasize that even under the Markov restriction, the principal's continuation payoff depends on π through the induced outcome distribution and hence the induced state visitation, while the agent's continuation payoff depends on ρ through the distribution of future contracts. This mutual dependence is the source of strategic feedback that is absent in one-shot Stackelberg formulations.

Why subgame-perfect equilibrium is the appropriate benchmark. Because the principal chooses b each period after observing the current state, and the agent chooses a after observing (s, b), a natural solution concept is subgame perfection: after every publicly observed history, each player's continuation strategy is optimal given the other's continuation strategy. In finite-horizon models, backward induction selects such behavior uniquely (up to tie-breaking). In the infinite horizon, however, the same logic must be expressed as a fixed point in continuation values. Under our stationarity restriction, the relevant equilibrium is a *stationary* subgame-perfect equilibrium (SPE): a pair (ρ^*, π^*) such that (i) the agent's action rule π^* is optimal in every (s, b) given the principal's contract policy ρ^* going forward, and (ii) the principal's contract rule ρ^* is optimal in every s anticipating the induced response π^* both today and in the future.

A convenient way to express these conditions is via (unregularized) Bellman equations. Define the agent's action-value function given a continuation

principal policy ρ by

$$Q_a(s,b,a) = \mathbb{E}_{o\sim O(\cdot|s,a),\, s'\sim T(\cdot|s,o)}\Big[r(s,a) + b(o) + \gamma\, \mathbb{E}_{b'\sim\rho(\cdot|s')} V_a(s',b')\Big], \quad (1)$$

and the corresponding value

$$V_a(s,b) = \max_{a\in\mathcal{A}} Q_a(s,b,a). \quad (2)$$

The agent's best-response correspondence is then

$$\mathrm{BR}_a(s,b) \in \arg\max_{a\in\mathcal{A}} Q_a(s,b,a),$$

with $\pi(\cdot \mid s,b)$ placing probability one on any selected maximizer (or mixing over maximizers if desired).

Similarly, given an agent policy π, the principal's contract-value function is

$$Q_p(s,b) = \mathbb{E}_{a\sim\pi(\cdot|s,b),\, o\sim O(\cdot|s,a),\, s'\sim T(\cdot|s,o)}\Big[r_p(s,o) - b(o) + \gamma V_p(s')\Big], \quad (3)$$

with value

$$V_p(s) = \max_{b\in\mathcal{B}} Q_p(s,b). \quad (4)$$

An SPE in stationary Markov strategies is thus a fixed point in which π^* selects maximizers of (1)–(2) given ρ^*, while ρ^* selects maximizers of (3)–(4) given π^*.

The infinite-horizon pathology: discontinuous best responses and cycling. Although the hard-max Bellman equations resemble familiar dynamic programming objects, the *joint* equilibrium mapping can be ill-behaved. The crux is that each player's optimal choice depends on the other's continuation behavior through arg max operations. When multiple actions (or contracts) yield nearly identical continuation values, an arbitrarily small change in beliefs about future play can flip the identity of the maximizer. In a hidden-action environment, such flips have first-order consequences: switching the agent's action changes the entire distribution of outcomes $O(\cdot \mid s,a)$, and therefore changes both current transfers and the transition kernel over future states. Consequently, the principal's objective as a function of b is typically only piecewise smooth, with kinks at the boundaries where the agent switches actions. Because the principal's future policy determines the agent's continuation payoff, those kinks move endogenously with ρ, creating a nontrivial feedback loop.

This observation matters for both equilibrium selection and computation. Consider the natural "alternating best response" procedure: fix a principal policy ρ, compute an agent best response $\pi \in \mathrm{BR}_a(\rho)$; then, fixing π,

compute a principal best response $\rho \in \mathrm{BR}_p(\pi)$; iterate. In a single-agent discounted MDP, the Bellman optimality operator is a γ-contraction even with a hard max, guaranteeing convergence of value iteration. Here, by contrast, the composition of two set-valued best-response correspondences need not be a contraction and need not even be single-valued. As a result, iterates can oscillate between qualitatively different regimes rather than settling to a fixed point.

One can see the mechanism in a simple two-by-two intuition. Suppose that in some state s the agent has two candidate actions, a^ℓ and a^h, and the principal has two candidate contracts, b^ℓ and b^h. For a range of continuation values, b^h makes a^h strictly optimal, while b^ℓ makes a^ℓ optimal. If a^h induces outcomes that move the state distribution toward regions where the principal later prefers b^ℓ (because, say, future rents become expensive), then the principal's best response to the induced π can switch from b^h to b^ℓ. But once the principal switches to b^ℓ, the agent optimally reverts to a^ℓ, which changes the induced state distribution back toward the region where b^h is again attractive. The alternation can therefore generate a deterministic two-cycle $(b^h, a^h) \to (b^\ell, a^\ell) \to (b^h, a^h)$, even though the underlying primitives are time-invariant and $\gamma < 1$. The discount factor controls how much future payoffs matter, but it does not remove the discontinuity created by hard best responses.

Implications and motivation for smoothing. The practical implication is that, in the unregularized infinite-horizon model, stationary SPE can be difficult to compute and can be sensitive to seemingly innocuous perturbations (e.g., rounding in a discretized contract set, numerical error in value estimates, or sampling noise in learned outcome models). Economically, this sensitivity reflects an equilibrium selection problem: when incentives are provided through a coarse outcome signal, the set of contracts that approximately implement a given action can be large, and the agent may be close to indifferent across actions over substantial regions of the state space. Computationally, these near-indifferences translate into discontinuous policy updates under arg max, which can prevent convergence of iterative procedures that would be reliable in standard MDPs.

These considerations motivate the regularized model we study next. By replacing hard best responses with entropy-regularized choice, we obtain a smooth, single-valued response mapping for both players, restoring contraction-like stability while preserving the economic primitives of hidden action and outcome-contingent transfers.

4 Regularized model: entropy smoothing and the rSPE concept

We now modify the baseline environment in one targeted way: instead of modeling each player as making an exact (hard) best response via an arg max, we model choice as *entropy-regularized* optimization. This change preserves the economic primitives of hidden action and outcome-contingent transfers, but replaces discontinuous best-response correspondences with smooth, single-valued maps. The resulting equilibrium notion will be a *regularized subgame-perfect equilibrium* (rSPE).

Entropy-regularized agent problem. Fix a principal stationary contract policy $\rho(\cdot \mid s)$. The agent observes (s, b) and selects a mixed action $\pi(\cdot \mid s, b)$. We assume the agent maximizes a discounted objective that includes an entropy term each period,

$$U_a(\pi; \rho) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \left(r(s_t, a_t) + b_t(o_t) + \tau_a H(\pi(\cdot \mid s_t, b_t))\right)\right], \tag{5}$$

where $\tau_a > 0$ indexes the strength of regularization and $H(p) = -\sum_i p_i \log p_i$ is Shannon entropy. The entropy term makes randomization directly valuable to the agent, which has two consequences that will be crucial for well-posed dynamics: (i) the maximizer exists and is unique, and (ii) the induced policy is everywhere interior (full support) whenever $\tau_a > 0$.

As in standard soft control, it is convenient to express the agent's problem through soft Bellman objects. Given ρ, define the agent soft Q-function by

$$Q_a(s, b, a) = \mathbb{E}_{o\sim O(\cdot|s,a),\, s'\sim T(\cdot|s,o)}\left[r(s, a) + b(o) + \gamma\, \mathbb{E}_{b'\sim\rho(\cdot|s')}\, V_a(s', b')\right], \tag{6}$$

and the corresponding soft value as the log-sum-exp aggregate

$$V_a(s, b) = \mathrm{LSE}_{\tau_a}((Q_a(s, b, a))_{a\in\mathcal{A}}) = \tau_a \log \sum_{a\in\mathcal{A}} \exp\left(\frac{Q_a(s, b, a)}{\tau_a}\right). \tag{7}$$

The associated optimal policy is recovered by a Gibbs (softmax) map,

$$\pi(a \mid s, b) = \frac{\exp(Q_a(s, b, a)/\tau_a)}{\sum_{a'\in\mathcal{A}} \exp(Q_a(s, b, a')/\tau_a)}. \tag{8}$$

Relative to the hard-max benchmark, (7) replaces the discontinuous operator $x \mapsto \max_i x_i$ with the smooth operator $x \mapsto \mathrm{LSE}_{\tau_a}(x)$, and (8) replaces selection from a (potentially set-valued) arg max with a unique, continuous mapping from continuation values to choice probabilities.

Entropy-regularized principal problem. We impose an analogous regularization for the principal. Fix an agent action rule $\pi(\cdot \mid s, b)$. The principal chooses a (possibly mixed) contract policy $\rho(\cdot \mid s)$ to maximize

$$U_p(\rho; \pi) = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \left(r_p(s_t, o_t) - b_t(o_t) + \tau_p H(\rho(\cdot \mid s_t))\right)\right], \tag{9}$$

where $\tau_p > 0$ plays the same role for contract choice. Define the principal soft Q-function

$$Q_p(s, b) = \mathbb{E}_{a \sim \pi(\cdot|s,b),\, o \sim O(\cdot|s,a),\, s' \sim T(\cdot|s,o)}\left[r_p(s, o) - b(o) + \gamma V_p(s')\right], \tag{10}$$

the soft value

$$V_p(s) = \mathrm{LSE}_{\tau_p}((Q_p(s, b))_{b \in \mathcal{B}}) = \tau_p \log \sum_{b \in \mathcal{B}} \exp\left(\frac{Q_p(s, b)}{\tau_p}\right), \tag{11}$$

and the induced contract policy

$$\rho(b \mid s) = \frac{\exp(Q_p(s, b)/\tau_p)}{\sum_{b' \in \mathcal{B}} \exp(Q_p(s, b')/\tau_p)}. \tag{12}$$

Two features are worth highlighting. First, because the principal's action is itself a contract (a vector of transfers across outcomes), the hard-max formulation tends to generate kinks precisely where small changes in continuation values switch the preferred contract. The soft aggregator in (11) smooths those kinks, which is useful even when $\mathcal{B}$ arises from discretizing an underlying continuous contract space. Second, τ_p has a transparent interpretation as controlling the degree of randomization in contracting: when τ_p is small, $\rho(\cdot \mid s)$ concentrates on near-maximizers; when τ_p is large, contracts are chosen more diffusely.

Definition (regularized subgame-perfect equilibrium). An *rSPE* is a pair of stationary Markov policies (ρ^*, π^*) together with bounded functions $(Q_a^*, V_a^*, Q_p^*, V_p^*)$ such that (i) (Q_a^*, V_a^*, π^*) satisfy (6)–(8) given ρ^*, and (ii) (Q_p^*, V_p^*, ρ^*) satisfy (10)–(12) given π^*. Operationally, rSPE replaces the mutual arg max optimality requirements of stationary SPE with mutual *Gibbs consistency*: each player's mixed action is exactly the softmax of its own continuation Q-values, and those Q-values are themselves computed under the other player's induced policy.

Interpretation and what regularization does (and does not) assume. We view entropy regularization as a disciplined way to capture forces that are present in applications but absent from the knife-edge hard-max model. One interpretation is *bounded rationality*: the agent and principal optimize subject to informational or computational frictions that make

perfectly sharp best responses implausible, and τ_a, τ_p quantify the severity of these frictions. A second, closely related interpretation is *private payoff shocks* (or "trembles"): if, at the moment of choice, each action carries an i.i.d. Type-I extreme value perturbation, then the optimal choice probabilities take the logit form (8)–(12), with τ proportional to the shock scale. In either case, the key modeling move is not that players are ex ante committed to randomize, but that their *effective* behavior is smooth in continuation values, eliminating knife-edge switching.

At the same time, we emphasize a limitation: the entropy term is not derived from first principles of contracting, and τ_a, τ_p are not primitives of technology or preferences in the same way as O or T. Rather, they parameterize a regularization that trades off faithfulness to the hard-max benchmark against stability of equilibrium selection and computation. In the limit $\tau_a, \tau_p \to 0$, the soft operators approach the hard-max operators, and the induced policies concentrate on best responses; for positive temperatures, the model selects a unique smooth equilibrium object. This is precisely the sense in which regularization can be read as an equilibrium selection device: it refines a potentially ill-behaved correspondence into a well-defined mapping while keeping the strategic feedback structure intact.

Finally, regularization has a direct computational implication. Because LSE_τ is smooth and Lipschitz, small errors in estimated continuation values translate into small changes in ρ and π, rather than discrete jumps. This property is especially valuable in hidden-action environments, where a small change in incentives can induce large changes in behavior and therefore in state visitation. The next section formalizes this stability by showing that, under $\tau_a, \tau_p > 0$, the induced joint soft Bellman operator is a γ-contraction and therefore admits a unique fixed point.

5 Main theorem: contraction, uniqueness, and computation in the infinite horizon

We now formalize the stability claim implicit in the preceding discussion: once both players use entropy-regularized optimization, the equilibrium conditions can be written as a single discounted fixed-point problem. The key step is to define an operator that (i) reconstructs the players' mixed policies from candidate Q-functions via the Gibbs maps (8) and (12), and then (ii) applies the corresponding soft Bellman expectations (6) and (10). Because the only intertemporal feedback enters through discounted continuation values, this operator inherits the standard γ-contraction structure familiar from entropy-regularized control, despite the hidden-action strategic interaction.

State space for the fixed-point problem. Let $\mathcal{Q}_a$ denote the space of bounded real functions on $\mathcal{S} \times \mathcal{B} \times \mathcal{A}$, and $\mathcal{Q}_p$ the space of bounded real

functions on $\mathcal{S} \times \mathcal{B}$. We endow the product space $\mathcal{Q}_a \times \mathcal{Q}_p$ with the sup norm

$$\|(Q_a, Q_p)\|_\infty \;=\; \max\Big\{ \sup_{s,b,a} |Q_a(s,b,a)|,\; \sup_{s,b} |Q_p(s,b)| \Big\}.$$

Boundedness is immediate under our standing assumptions (finite sets and bounded per-period rewards), and it ensures that all log-sum-exp expressions are finite.

Policy and value reconstruction from candidate Q-functions. Given any pair $(Q_a, Q_p) \in \mathcal{Q}_a \times \mathcal{Q}_p$, we define the induced soft values and Gibbs policies pointwise as follows. First, for the agent,

$$V_a^{Q_a}(s,b) \;=\; \tau_a \log \sum_{a \in \mathcal{A}} \exp\!\left(\frac{Q_a(s,b,a)}{\tau_a} \right), \qquad \pi^{Q_a}(a \mid s,b) \;=\; \frac{\exp(Q_a(s,b,a)/\tau_a)}{\sum_{a' \in \mathcal{A}} \exp(Q_a(s,b,a')/\tau_a)}.$$

Second, for the principal,

$$V_p^{Q_p}(s) \;=\; \tau_p \log \sum_{b \in \mathcal{B}} \exp\!\left(\frac{Q_p(s,b)}{\tau_p} \right), \qquad \rho^{Q_p}(b \mid s) \;=\; \frac{\exp(Q_p(s,b)/\tau_p)}{\sum_{b' \in \mathcal{B}} \exp(Q_p(s,b')/\tau_p)}.$$

These reconstructions are single-valued and yield full-support mixed strategies whenever $\tau_a, \tau_p > 0$, which is precisely what removes the discontinuities present in the hard-max benchmark.

The joint soft Bellman operator. We define $\mathcal{T} : \mathcal{Q}_a \times \mathcal{Q}_p \to \mathcal{Q}_a \times \mathcal{Q}_p$ by $\mathcal{T}(Q_a, Q_p) = (\mathcal{T}_a(Q_a, Q_p), \mathcal{T}_p(Q_a, Q_p))$, where the two components are the natural soft Bellman back-ups computed under the reconstructed policies. For the agent, for every (s, b, a),

$$(\mathcal{T}_a(Q_a, Q_p))(s,b,a) = \mathbb{E}_{o \sim O(\cdot|s,a),\, s' \sim T(\cdot|s,o)} \left[r(s,a) + b(o) + \gamma\, \mathbb{E}_{b' \sim \rho^{Q_p}(\cdot|s')}\, V_a^{Q_a}(s', b') \right]. \tag{13}$$

For the principal, for every (s, b),

$$(\mathcal{T}_p(Q_a, Q_p))(s,b) = \mathbb{E}_{a \sim \pi^{Q_a}(\cdot|s,b),\, o \sim O(\cdot|s,a),\, s' \sim T(\cdot|s,o)} \left[r_p(s,o) - b(o) + \gamma\, V_p^{Q_p}(s') \right]. \tag{14}$$

By construction, fixed points of $\mathcal{T}$ are exactly the objects required by the equilibrium conditions: if $(Q_a^*, Q_p^*) = \mathcal{T}(Q_a^*, Q_p^*)$, then the induced $(V_a^{Q_a^*}, \pi^{Q_a^*})$ and $(V_p^{Q_p^*}, \rho^{Q_p^*})$ satisfy (6)–(12) simultaneously.

Theorem (contraction and uniqueness of rSPE). Fix finite $\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{B}$, bounded rewards, and $\gamma \in (0,1)$. If $\tau_a > 0$ and $\tau_p > 0$, then the joint operator $\mathcal{T}$ defined in (13)–(14) is a γ-contraction on $(\mathcal{Q}_a \times \mathcal{Q}_p, \|\cdot\|_\infty)$. Consequently:

1. **(Unique fixed point)** there exists a unique pair (Q_a^*, Q_p^*) with $(Q_a^*, Q_p^*) = \mathcal{T}(Q_a^*, Q_p^*)$;

2. **(Unique policies)** the reconstructed Gibbs policies $\pi^* = \pi^{Q_a^*}$ and $\rho^* = \rho^{Q_p^*}$ are uniquely determined and have full support;

3. **(Unique rSPE)** the pair (ρ^*, π^*) forms the unique stationary Markov rSPE.

Why the contraction is economically natural. At a high level, the argument mirrors the one-player case. The discounted Bellman step is affine in continuation values, and stochastic averaging (expectations under O, T, and the reconstructed mixed policies) is non-expansive under $\|\cdot\|_\infty$. Entropy regularization enters through the log-sum-exp aggregator: for any two vectors x, y of equal dimension,

$$|\mathrm{LSE}_\tau(x) - \mathrm{LSE}_\tau(y)| \leq \|x - y\|_\infty,$$

so replacing hard maximization by LSE_τ preserves the Lipschitz modulus 1. Intuitively, the soft value is a smooth "certainty equivalent" of future payoffs, and changing the continuation payoff profile by at most ε changes that certainty equivalent by at most ε. The only systematic shrinkage comes from discounting by γ, which is exactly what yields the γ-contraction.

Computational corollary (global convergence of joint value iteration). For any initialization $(Q_a^{(0)}, Q_p^{(0)})$, define iterates $(Q_a^{(k+1)}, Q_p^{(k+1)}) = \mathcal{T}(Q_a^{(k)}, Q_p^{(k)})$. The contraction property implies geometric convergence:

$$\left\|(Q_a^{(k)}, Q_p^{(k)}) - (Q_a^*, Q_p^*)\right\|_\infty \leq \gamma^k \left\|(Q_a^{(0)}, Q_p^{(0)}) - (Q_a^*, Q_p^*)\right\|_\infty.$$

Thus, unlike alternating hard best responses (which may cycle in hidden-action settings), the entropy-regularized mapping admits a globally stable computation: iterating the soft Bellman operator converges from any starting point to the unique rSPE objects. This is the sense in which $\tau_a, \tau_p > 0$ provide not only an equilibrium selection device but also an algorithmic regularization: equilibrium behavior becomes a fixed point of a contraction, rather than a potentially ill-behaved correspondence.

From uniqueness to approximation. Having obtained a unique and well-behaved infinite-horizon equilibrium for every $(\tau_a, \tau_p) > 0$, we can now ask how this selected rSPE relates to the (possibly multiple) stationary SPE of the unregularized model. The next section quantifies the value loss induced by smoothing and characterizes the resulting stability–optimality trade-off as $\tau_a, \tau_p \to 0$.

6 Approximation to unregularized SPE: value-gap bounds and the stability–optimality frontier

Entropy regularization selects a unique stationary Markov rSPE for every $(\tau_a, \tau_p) > 0$. The natural economic question is how far this selected outcome can be from the unregularized benchmark in which both players use hard best responses. In this section we quantify the worst-case *value loss* induced by smoothing, clarify when the bound is informative (and when it is not), and interpret the resulting stability–optimality trade-off as $\tau_a, \tau_p \to 0$.

A generic "softmax vs. max" inequality. The basic ingredient is the standard comparison between hard maximization and the log-sum-exp aggregator. For any finite set I and any vector $(x_i)_{i\in I}$,

$$\max_{i\in I} x_i \;\le\; \mathrm{LSE}_\tau\big((x_i)_{i\in I}\big) \;\le\; \max_{i\in I} x_i \;+\; \tau \log |I|. \tag{15}$$

The left inequality says regularization never *undershoots* the best available continuation value when measured in the soft value units; the right inequality says it can overshoot by at most $\tau \log |I|$. Economically, $\tau \log |I|$ is the maximal "randomization rent" one can extract from the entropy term when the decision maker is indifferent among many alternatives.

Value-gap bound for the principal. Fix an initial state s_0. Let $V_p^{\tau_p,\tau_a}(s_0)$ be the principal's value at the unique rSPE under temperatures (τ_p, τ_a). Let $\sup_{\mathrm{SPE}} V_p^{\mathrm{SPE}}(s_0)$ be the principal's maximal value across (possibly multiple) stationary Markov SPE in the unregularized game. Then the rSPE value satisfies the bound

$$V_p^{\tau_p,\tau_a}(s_0) \;\ge\; \sup_{\mathrm{SPE}} V_p^{\mathrm{SPE}}(s_0) \;-\; \frac{\tau_p \log |\mathcal{B}| + \tau_a \log |\mathcal{A}|}{1-\gamma}. \tag{16}$$

The logic is directly analogous to one-player entropy-regularized control, with two layers of smoothing. The principal's own soft value replaces a hard maximization over contracts $b \in \mathcal{B}$, generating a per-period discrepancy of at most $\tau_p \log |\mathcal{B}|$. In addition, because the agent uses a soft best response, the principal faces a softened mapping from contracts to induced actions; this introduces an additional discrepancy controlled by $\tau_a \log |\mathcal{A}|$. Discounting then converts a per-period bound into an infinite-horizon bound through the geometric series $\sum_{t\ge 0} \gamma^t = (1-\gamma)^{-1}$.

Two remarks help interpret (16). First, the bound is *uniform in the primitives*: it does not depend on the magnitudes of r, r_p, b, nor on the details of O and T, beyond boundedness. This is exactly why it must scale with the worst-case entropy gap $\log |\cdot|$: without additional curvature or margin assumptions, a decision maker can be made arbitrarily close to indifferent, so

even tiny smoothing can change the selected mixture substantially. Second, the bound isolates how patience amplifies approximation error: for fixed temperatures, the loss grows like $1/(1-\gamma)$, so near-undiscounted environments require smaller τ_a, τ_p to achieve a given approximation tolerance.

Convergence as $\tau_a, \tau_p \to 0$: what it does and does not mean. Equation (16) immediately yields

$$\lim_{\tau_a,\tau_p\to 0} V_p^{\tau_p,\tau_a}(s_0) \;=\; \sup_{\text{SPE}} V_p^{\text{SPE}}(s_0),$$

in the sense that the value gap can be made arbitrarily small by taking temperatures small. This is a *value* statement, not a statement about equilibrium *selection* among multiple unregularized SPE. When the unregularized model admits several stationary SPE with different principal payoffs, the rSPE does not in general converge to a particular equilibrium profile independent of how (τ_a, τ_p) are taken to zero. Intuitively, when the hard-max correspondence has flat regions (ties), the softmax rule breaks ties smoothly but in a way that depends on relative payoff differences at the scale of τ. Small perturbations—including the interaction of the two temperatures—can therefore determine which hard best-response branch is approached.

In particular, even if $V_p^{\tau_p,\tau_a}(s_0)$ approaches the *best* attainable SPE value, the limiting *policies* (ρ^*, π^*) may fail to converge, or may converge only along subsequences, whenever the unregularized equilibrium set contains continua or whenever multiple contracts/actions are exactly optimal in some states. Put differently: entropy regularization provides a canonical *selection for each* (τ_a, τ_p), but not necessarily a canonical selection at $\tau = 0$ without additional tie-breaking structure.

The stability–optimality frontier. From a mechanism-design perspective, τ_a and τ_p parameterize a frontier between (i) *stability*—smooth dependence of behavior on payoffs and on approximation error—and (ii) *optimality* relative to the unregularized benchmark. Larger temperatures enlarge the right-hand side of (16) linearly, but they also make behavior less brittle.

A simple way to see the stability benefit is through sensitivity of mixed strategies to perturbations in Q-values. For the agent,

$$\pi^{Q_a}(a \mid s, b) \;=\; \frac{\exp(Q_a(s,b,a)/\tau_a)}{\sum_{a'} \exp(Q_a(s,b,a')/\tau_a)},$$

so a perturbation of size ε in $Q_a(s,b,\cdot)$ changes log-odds by at most ε/τ_a. Thus, for fixed (s,b), smaller τ_a makes the mapping from estimated values to action probabilities steeper and hence more sensitive to estimation error (or to small contract changes). Analogous statements hold for the principal's contract selection as a function of $Q_p(s,\cdot)$ with scale τ_p. In applications

where Q-values are learned from finite samples or approximated with function approximation, these Lipschitz-type controls are often the difference between stable training and oscillatory behavior.

The frontier is therefore operational: a designer can interpret τ_a, τ_p as robustness knobs. If the environment is noisy, learning is approximate, or the contract set is discretized coarsely, modest regularization can improve out-of-sample performance even if it introduces some bias relative to the ideal hard-max optimum.

When is the bound tight? The inequality (16) is worst-case and can be tight (up to constants) in knife-edge cases. Tightness arises when, at many histories/states, the relevant Q-vectors are nearly flat across a large fraction of $\mathcal{A}$ and/or $\mathcal{B}$. For example, if the principal is nearly indifferent among a large menu of contracts in a given state, then LSE_{τ_p} can exceed max by close to $\tau_p \log |\mathcal{B}|$. Similarly, if the agent is nearly indifferent among several actions under the prevailing contract, then the agent's soft value can differ from the hard max by nearly $\tau_a \log |\mathcal{A}|$, and this difference propagates through continuation values.

Conversely, in environments with *margins*—i.e., a unique optimal contract and action separated by a gap from the runner-up—the effective value loss is typically much smaller than (16). In such cases, once τ_a, τ_p are below the relevant payoff gaps, the soft policies concentrate sharply on the unique maximizers and the rSPE behavior becomes close to the hard-max equilibrium. Making such "margin" statements formal is possible but necessarily depends on instance-specific constants (gaps that vary across states and along the equilibrium path), so we treat (16) as the robust baseline guarantee.

What is not guaranteed. Finally, it is worth separating three distinct claims. First, we *do* guarantee a unique rSPE for every $(\tau_a, \tau_p) > 0$ and a computable contraction-based procedure to find it. Second, we *do* guarantee that the principal's value at rSPE is within an explicit and vanishing bound of the best unregularized stationary SPE value as $\tau_a, \tau_p \to 0$. Third, we *do not* claim that rSPE selects a particular unregularized equilibrium outcome independent of the limiting path, nor that it resolves multiplicity in a way that is normatively "correct" absent further equilibrium selection criteria. In the next section we complement these positive results with necessity and limitation statements, illustrating what breaks without smoothing and what additional complications arise when the contract space is continuous rather than finite.

7 Necessity and limitations: what breaks without smoothing, and what changes with continuous contract sets

Our existence, uniqueness, and global-convergence statements rely on the fact that *both* players replace hard best responses with entropy-regularized (logit) choice. This is not merely a technical convenience. In hidden-action environments, discontinuities in best-response correspondences are the central obstruction to stable computation and to equilibrium selection, and smoothing is precisely what removes those discontinuities. Here we make this point concrete by describing what can fail when either temperature is set to zero, and by clarifying the additional complications that arise when the contract menu is naturally continuous rather than finite.

Why a positive temperature is doing real work. In the unregularized game, the agent's best response is a set-valued correspondence

$$b \;\mapsto\; \arg\max_{a \in \mathcal{A}} Q_a(s, b, a),$$

and the principal's best response is similarly set-valued in b. Both correspondences can jump as payoffs cross tie thresholds. These jumps are particularly acute in moral-hazard problems because contracts affect the agent only through the distribution of outcomes, so small changes in $b(\cdot)$ can flip incentives discretely when two actions have nearly equal expected utility. As a result, the natural "alternating best-response" mapping (principal optimizes given the agent's best response; agent best-responds given the principal's contract choice) need not be a contraction, need not be single-valued, and can exhibit cycles.

A canonical failure mode is a two-state, two-action, two-contract construction in which each player's unique best response at time t makes the other player's unique best response at time $t+1$ switch to the opposite extreme. When both best-response steps are hard argmax operations, this induces a deterministic two-cycle. The underlying economic mechanism is simple: the principal slightly increases incentives to induce action a_1; once a_1 is induced, the principal prefers to reduce incentives (because payments are costly), which in turn makes the agent revert to a_0; and so on. The details of such a construction can be implemented with bounded rewards and a discount factor $\gamma \in (0, 1)$ by arranging that (i) the two actions generate outcome distributions that differ only slightly, and (ii) the principal's gross reward is sufficiently sensitive to which action is taken, but the agent's intrinsic cost makes the agent indifferent at a knife edge. At the knife edge, arbitrarily small numerical error or perturbation in Q-values causes the induced action to flip, and the subsequent contract choice flips with it.

What breaks if $\tau_a = 0$ (hard agent response), even if $\tau_p > 0$. Suppose we keep the principal regularized but let the agent best-respond via a hard max:

$$V_a(s,b) \;=\; \max_{a\in\mathcal{A}} Q_a(s,b,a), \qquad \pi(\cdot \mid s,b) \in \arg\max_a Q_a(s,b,a).$$

The principal's Bellman backup still contains a smooth log-sum-exp over contracts, but the quantity being exponentiated,

$$Q_p(s,b) = \mathbb{E}_{a\sim\pi(\cdot|s,b),\, o,s'}\left[r_p(s,o) - b(o) + \gamma V_p(s')\right],$$

now depends on $\pi(\cdot \mid s,b)$, which can be discontinuous in $(Q_a(s,b,a))_{a\in\mathcal{A}}$ and thus discontinuous in b. In particular, the mapping from contract vectors b to induced outcomes (through the agent's argmax) can have jump discontinuities. This matters for computation: even if the principal takes a softmax over b, the function $b \mapsto Q_p(s,b)$ can be highly non-smooth, with large effective Lipschitz constants (indeed, no global Lipschitz bound) because a tiny change in b can change the selected a and hence the entire outcome distribution. Consequently, the joint operator need not be a γ-contraction in sup norm, and we can lose both global convergence of iteration and uniqueness of the stationary equilibrium.

Economically, $\tau_p > 0$ alone gives the principal a smooth *selection* among contracts, but it does not smooth the induced mapping from contracts to behavior when the agent's incentives are close. Thus, principal randomization by itself does not eliminate the core moral-hazard discontinuity: the agent's response to incentives.

What breaks if $\tau_p = 0$ (hard principal response), even if $\tau_a > 0$. Now suppose the agent uses a soft best response but the principal chooses a hard maximizer:

$$V_p(s) \;=\; \max_{b\in\mathcal{B}} Q_p(s,b), \qquad \rho(\cdot \mid s) \in \arg\max_b Q_p(s,b).$$

With $\tau_a > 0$, the mapping $b \mapsto \pi(\cdot \mid s,b)$ becomes continuous, and this indeed eliminates one major source of discontinuity. However, the principal's own argmax can still generate non-robust jumps as $Q_p(s,b)$ changes, especially when two contracts are close substitutes. In a dynamic setting, these jumps can propagate across time through continuation values, leading again to non-convergent iterates of naive best-response dynamics. Moreover, multiplicity reappears: if several contracts are optimal at a state, there is no canonical selection without an explicit tie-breaking rule, and different selections can support different stationary equilibria.

From a mechanism-design standpoint, this case is also troubling in applications with approximation or learning. Even when the agent's response is smooth, a principal that implements a strict argmax is maximally sensitive

to estimation error in Q_p: an ε-perturbation can flip the chosen contract entirely. Thus $\tau_p > 0$ is not only an equilibrium-selection device; it is a robustness device for the principal's optimization problem.

The knife-edge case $(\tau_a, \tau_p) = (0, 0)$: cycling and non-computability by naive iteration. When both temperatures are zero, we are back in the classical stationary Markov SPE concept with hidden action. In finite spaces, stationary SPE may exist, but the set of equilibria can be large and the correspondence from primitives to equilibrium outcomes can be discontinuous. More importantly for our purposes, the computational object changes: the natural fixed-point iteration that alternates hard best responses is not generally well-behaved. There are instances (including variants of the cycling construction described above) where iterating best responses does not converge from generic initializations. This does not contradict existence of equilibrium; it indicates that equilibrium computation requires more delicate methods (global search, mixed-integer formulations, or explicit equilibrium solvers) and that small numerical perturbations can substantially change the equilibrium selected.

Continuous contract spaces: conceptual and computational complications. We have maintained $\mathcal{B}$ finite to keep both the equilibrium operator and the value-gap bounds transparent. In many contracting environments, however, $\mathcal{B}$ is naturally continuous (e.g., payments $b(o)$ can vary continuously subject to limited liability). Extending entropy regularization to continuous $\mathcal{B}$ is conceptually straightforward but requires additional structure.

First, the principal's soft value becomes a log-partition *integral* rather than a finite log-sum:

$$V_p(s) \;=\; \tau_p \log \int_{\mathcal{B}} \exp\big(Q_p(s,b)/\tau_p\big)\, d\mu(b),$$

where μ is a chosen reference measure on $\mathcal{B}$ (e.g., Lebesgue on a compact subset). This highlights an often-overlooked modeling choice: in continuous spaces, "entropy regularization" is properly interpreted as *relative entropy* (a KL penalty) with respect to μ, and the value depends on the scale/volume induced by that reference measure. Without compactness or integrability conditions, the log-partition function may be infinite, and the regularized objective can become ill-posed.

Second, computation typically requires numerical integration or approximation. In low dimensions one may approximate the integral via quadrature; in higher dimensions one typically resorts to (i) discretization of $\mathcal{B}$ to a finite grid (reducing to our baseline model), (ii) Monte Carlo estimation of the log-partition, or (iii) restricting $\rho(\cdot \mid s)$ to a parametric family (e.g., an

exponential family over contracts) so that sampling and density evaluation are tractable. Each choice introduces a new approximation layer beyond (τ_a, τ_p): discretization error, sampling variance, or function-approximation bias. Moreover, discretization interacts with regularization in a nontrivial way: refining a grid improves the approximation to the continuous contract problem but increases the effective menu size, which can change the magnitude and interpretation of entropy terms unless one simultaneously adjusts the reference measure or rescales the regularizer.

These limitations are not defects of the framework; they clarify its domain of clean guarantees. Finite $\mathcal{B}$ delivers a sharp contraction-based theory with explicit bounds. Continuous $\mathcal{B}$ is often the economically natural benchmark, but it shifts part of the problem from equilibrium existence to numerical analysis: specifying the appropriate reference measure and building reliable approximations to the soft Bellman integrals. The next section therefore turns to algorithms and implementations that preserve the stability benefits of regularization while remaining practical in large or continuous design spaces.

8 Algorithms: soft value iteration and soft Q-learning in hidden-action contracting

The contraction result is not only a conceptual equilibrium-selection statement; it also suggests a particularly simple computational template. Because both players' operators are smooth and globally stable, we can treat the rSPE as the unique fixed point of a coupled pair of soft Bellman equations and compute it by straightforward iteration, much as in single-agent entropy-regularized control. In the finite (tabular) setting, this lets us replace brittle alternating argmax best responses with numerically well-behaved log-partition updates.

Model-based soft value iteration (tabular). When the primitives (O, T, r, r_p) are known, we can iterate the joint operator directly. Given current estimates $(Q_a^{(k)}, Q_p^{(k)})$, we reconstruct the policies by Gibbs formulas,

$$\pi^{(k)}(a \mid s, b) = \frac{\exp(Q_a^{(k)}(s, b, a)/\tau_a)}{\sum_{a'} \exp(Q_a^{(k)}(s, b, a')/\tau_a)}, \qquad \rho^{(k)}(b \mid s) = \frac{\exp(Q_p^{(k)}(s, b)/\tau_p)}{\sum_{b'} \exp(Q_p^{(k)}(s, b')/\tau_p)}.$$

We then compute the associated soft values,

$$V_a^{(k)}(s, b) = \tau_a \log \sum_{a \in \mathcal{A}} \exp(Q_a^{(k)}(s, b, a)/\tau_a), \qquad V_p^{(k)}(s) = \tau_p \log \sum_{b \in \mathcal{B}} \exp(Q_p^{(k)}(s, b)/\tau_p),$$

and apply the soft Bellman backups:

$$Q_a^{(k+1)}(s,b,a) = \mathbb{E}_{o\sim O(\cdot|s,a),\, s'\sim T(\cdot|s,o)}\Big[r(s,a) + b(o) + \gamma\, \mathbb{E}_{b'\sim\rho^{(k)}(\cdot|s')} V_a^{(k)}(s',b')\Big],$$
$$Q_p^{(k+1)}(s,b) = \mathbb{E}_{a\sim\pi^{(k)}(\cdot|s,b),\, o\sim O(\cdot|s,a),\, s'\sim T(\cdot|s,o)}\Big[r_p(s,o) - b(o) + \gamma\, V_p^{(k)}(s')\Big].$$

In implementation, these expectations are finite sums. The main computational burden is therefore combinatorial rather than conceptual: a naive update scales like

$$O(|\mathcal{S}|\,|\mathcal{B}|\,|\mathcal{A}|\,|\mathcal{O}|) \quad \text{for } Q_a, \qquad O(|\mathcal{S}|\,|\mathcal{B}|\,|\mathcal{A}|\,|\mathcal{O}|) \quad \text{for } Q_p,$$

with an additional factor for transitions if T is dense in $|\mathcal{S}|$. The contraction guarantee justifies asynchronous variants (updating one (s,b,a) triple at a time) and standard acceleration tricks (Gauss–Seidel sweeps, prioritized updates), since all inherit global stability in the tabular case.

Sample-based learning: coupled soft Q-learning. When (O,T) or rewards are unknown, the same structure yields a natural learning instantiation in which each side performs soft Q-learning with bootstrapping. The agent observes (s,b,a,o,s') and can update

$$Q_a(s,b,a) \leftarrow (1-\alpha)\, Q_a(s,b,a) + \alpha\Big(r(s,a) + b(o) + \gamma\, \mathbb{E}_{b'\sim\rho(\cdot|s')} V_a(s',b')\Big),$$

where $V_a(s',b') = \tau_a \log \sum_{a'} \exp(Q_a(s',b',a')/\tau_a)$. The principal observes (s,b,o,s') (but not a); importantly, it can still update Q_p from the realized outcome and next state without ever imputing the hidden action:

$$Q_p(s,b) \leftarrow (1-\beta)\, Q_p(s,b) + \beta\Big(r_p(s,o) - b(o) + \gamma\, V_p(s')\Big), \qquad V_p(s') = \tau_p \log \sum_{b'} \exp(Q_p(s',b')/\tau_p).$$

This is the key practical advantage of working with outcomes o as the contractible statistic: the principal's temporal-difference target only requires o and s'. The hidden action matters only through the data-generating process, not through observables needed for the update.

Because the policies π and ρ are deterministic functions of the current Q-tables, the learning dynamics are coupled. In the tabular setting, the contraction theorem provides a strong heuristic: simultaneous updates behave like stochastic approximation to a contraction mapping, so we should expect stable convergence under standard Robbins–Monro step-size conditions and sufficient exploration. In practice, exploration is endogenous here—the entropy terms themselves produce persistent randomization—but one may still add explicit exploration (e.g. occasional uniform mixing in ρ) when τ_p is very small.

Numerical stability and temperature schedules. Two numerical issues arise repeatedly. First, computing LSE_τ should use the standard stabilization

$$\tau \log \sum_i e^{x_i/\tau} = m + \tau \log \sum_i e^{(x_i - m)/\tau}, \qquad m = \max_i x_i,$$

to avoid overflow when τ is small. Second, while our theory treats τ_a, τ_p as fixed primitives, practitioners often want behavior close to the unregularized SPE. A common approach is annealing: start with larger temperatures (promoting exploration and smoothing) and gradually decrease them. This can work well empirically, but it reintroduces a familiar tradeoff: as $\tau \downarrow 0$, the problem becomes less smooth and the effective condition number worsens, so one typically needs slower step sizes, more samples, or both.

When the regularized selection replaces LP-based contract choice. In many finite-menu applications, the principal's contract design problem is operationally: choose $b \in \mathcal{B}$ subject to exogenous feasibility constraints (limited liability, budget caps, regulatory restrictions), where $\mathcal{B}$ is already a discrete set of admissible contracts (e.g. standardized bonus schemes). In that case, the entropy-regularized operator is a direct substitute for per-state deterministic optimization: instead of repeatedly solving $\max_{b \in \mathcal{B}} Q_p(s, b)$ (with all its tie-breaking and discontinuity problems), the principal computes $\rho(b \mid s) \propto \exp(Q_p(s, b)/\tau_p)$. The algorithmic implication is simple: *equilibrium computation becomes iterated evaluation of soft Bellman backups plus softmaxes*, with no additional inner optimization.

This contrasts with "minimal-implementation" approaches in static principal–agent models, where one often solves a linear program to implement a target action at minimum expected payment subject to incentive constraints. In our dynamic hidden-action setting, if we insist on designing contracts from a rich (effectively continuous) family $\{b(\cdot)\}$ and we impose constraints that are not captured by a finite $\mathcal{B}$, then an inner optimization problem can reappear. For example, if $\mathcal{B}$ is described by linear inequalities (limited liability $b(o) \geq 0$, budget $\sum_o b(o) \leq \bar{B}$, etc.), then evaluating either the hard maximum $\max_b Q_p(s, b)$ or its soft analogue (a log-partition integral) generally requires numerical optimization or discretization. In such cases, the regularizer does not eliminate the need for optimization; rather, it changes the nature of the inner problem from brittle argmax selection to a smoother objective (often strictly convex in suitable parametrizations), which is typically easier to solve and more robust to approximation error.

Function approximation and large state spaces. Finally, while our guarantees are sharpest in the finite tabular case, the same computational blueprint extends naturally to large $|\mathcal{S}|$ with parametric approximators $Q_a(s, b, a; \theta)$

and $Q_p(s, b; \phi)$. One can view the principal as running a soft actor–critic over contracts (actor ρ_ψ, critic Q_p), while the agent runs a soft actor–critic over actions (actor π_η, critic Q_a). The key limitation is that contraction is no longer automatic under nonlinear approximation; stability becomes empirical and depends on optimization choices, replay, target networks, and coverage of (s, b) pairs. We therefore treat large-scale implementations as a numerical extension: the rSPE structure supplies a disciplined target (the coupled soft Bellman equations) and a principled exploration mechanism (entropy), but it does not, by itself, resolve the standard instabilities of deep reinforcement learning.

9 Experiments: stabilization, stress tests, and a large-scale illustration

We use a small set of experiments to make three points that mirror the theory: (i) the entropy terms eliminate the non-convergence pathologies that arise under hard best responses in hidden-action contracting; (ii) the resulting equilibrium computation is materially more robust to approximation error in value estimates and in model estimation; and (iii) the same coupled "soft" structure can be carried to larger problems with function approximation, though we emphasize that the latter is a numerical demonstration rather than a theorem-backed claim.

(i) Reproducing a cycling instance and showing stabilization. Our first experiment revisits a canonical failure mode for alternating best responses in dynamic principal–agent problems: even in a finite tabular environment, the principal's contract update can trigger a discrete change in the agent's optimal action, which then changes the principal's continuation value enough to flip the contract choice back, producing a cycle. Concretely, we implement a small hidden-action MDP in which the agent has two actions ("safe" versus "risky") and the outcome distribution differs sharply across actions; the principal's gross reward $r_p(s, o)$ is aligned with the risky outcome only in some states, and the transition kernel $T(\cdot \mid s, o)$ makes those states persistent. We restrict the principal to a finite menu $\mathcal{B}$ containing contracts that are near substitutes in expected payment but that tilt incentives in opposite directions.

In the unregularized baseline ($\tau_a = \tau_p = 0$), we run the natural alternating scheme: given a contract policy, compute the agent's best-response policy by dynamic programming; then, given the induced agent policy, compute the principal's best-response contract policy. Starting from different initializations, we observe persistent cycling of the induced stationary policies and non-convergence of the iterates of the principal value $V_p^{(k)}$ (the iterates oscillate within a band rather than approaching a limit). Importantly, this is not

a numerical artifact of step sizes: the baseline is model-based and uses exact expectations, so the oscillation reflects the discontinuity of argmax selection.

We then repeat the same procedure with entropy regularization, setting $\tau_a > 0$ and $\tau_p > 0$ and performing soft value iteration on the coupled Bellman system. In contrast to the baseline, the iterates converge rapidly from all initializations to the same fixed point. Empirically, the convergence rate is geometric and visually consistent with the γ-contraction prediction: plotting $\|Q^{(k+1)} - Q^{(k)}\|_\infty$ on a log scale yields an approximately linear decay. The qualitative mechanism is exactly the one highlighted by the theory: the principal's policy $\rho(\cdot \mid s)$ varies smoothly with $Q_p(s, \cdot)$, and the agent's policy $\pi(\cdot \mid s, b)$ varies smoothly with $Q_a(s, b, \cdot)$, so the iterative mapping cannot "jump" across best-response correspondences. We also verify that as τ_a, τ_p are decreased toward zero, the regularized fixed point approaches one of the limiting unregularized equilibria when those equilibria exist, but the numerical condition worsens (more iterations are needed and the iterates become more sensitive to rounding), consistent with smoothing as a stability–bias tradeoff.

(ii) Stress tests under approximation error. The second set of experiments asks a practical question: in finite contracting problems, one rarely has exact Q-values, even in the tabular setting, because either the model must be estimated or the values must be learned from sample. We therefore introduce controlled perturbations and measure how much the induced policies and achieved values degrade under the unregularized and regularized operators.

We consider two perturbation modes. First, we add bounded noise directly to the Q-tables before policy reconstruction: at each iteration we use $\widetilde{Q}_a = Q_a + \epsilon_a$ and $\widetilde{Q}_p = Q_p + \epsilon_p$, where ϵ is i.i.d. across entries and clipped to $[-\bar{\epsilon}, \bar{\epsilon}]$. Second, we estimate the primitives (O, T) from a finite dataset, compute model-based backups using the empirical $(\widehat{O}, \widehat{T})$, and evaluate the resulting stationary policies in the true environment.

Across both perturbation modes, the unregularized baseline exhibits "brittle" behavior: small $\bar{\epsilon}$ frequently triggers discrete switches in either the agent's action or the principal's contract, leading to large changes in realized value. In particular, the principal's realized return can drop sharply when a contract that was almost tied in $Q_p(s, \cdot)$ becomes the argmax due to noise; because action is hidden, the downstream effect is amplified through the induced change in π and therefore in the outcome distribution. Under entropy regularization, the same perturbations translate into gradual changes in ρ and π. We quantify this by measuring (a) the average total variation distance between the perturbed and unperturbed policies, and (b) the value loss relative to the unperturbed solution. Both metrics scale smoothly with $\bar{\epsilon}$, and the slope is controlled by the temperatures: larger τ_a, τ_p produce

smaller policy deviations for a fixed Q-perturbation, at the cost of a larger asymptotic bias relative to the hard-max optimum.

A practical implication emerges from these stress tests. In contracting applications where $|\mathcal{B}|$ is large (many standardized bonus schemes) or where incentives are finely balanced, small estimation errors are unavoidable. The regularized equilibrium provides a principled way to trade off incentive sharpness against stability. In deployment language, τ plays a role analogous to a robustness knob: if the environment or the learned model is noisy, slightly higher temperatures can prevent regime-switching behavior that would otherwise appear as erratic contract changes.

(iii) A large-MDP demonstration with function approximation. Finally, we provide an optional large-scale illustration to show that the coupled structure is not confined to toy tabular settings. We construct a gridworld-style state space with stochastic outcomes and hidden actions (the agent's action affects the distribution over outcomes such as "high output" versus "low output," which in turn affect both the principal payoff and the transition). The principal's action is a discrete contract index $b \in \mathcal{B}$ drawn from a menu that encodes a few interpretable shapes (e.g. high-powered incentives, flat wage, and intermediate schemes), possibly crossed with budget caps.

We implement two soft actor–critic learners: one for the agent (critic $Q_a(s, b, a; \theta)$, actor $\pi_\eta(a \mid s, b)$) and one for the principal (critic $Q_p(s, b; \phi)$, actor $\rho_\psi(b \mid s)$). The critics are trained with temporal-difference targets that mirror the soft Bellman equations, and the actors are trained to match the Gibbs policies induced by the critics (equivalently, to maximize the entropy-regularized objectives). Because the principal does not observe a, its critic update uses only (s, b, o, s'), reinforcing the operational appeal of outcome-based contracting: the learning signal is aligned with what is contractible and observable.

We find that training is stable over a broad range of hyperparameters when τ_a, τ_p are not too small, and that learned behavior is qualitatively sensible: in regions of the state space where high effort is valuable, the principal shifts probability mass toward higher-powered contracts, while in low-return regions it economizes on incentives. We emphasize, however, that this is a numerical demonstration, not a guarantee. With nonlinear approximation, contraction need not hold, and familiar deep-RL issues arise (overestimation, distribution shift, and sensitivity to target network updates). We therefore interpret this experiment as evidence that the rSPE equations provide a coherent learning target and a stable exploration mechanism, rather than as a claim of universal convergence.

Taken together, these experiments support the central message of the paper: smoothing is not merely a technical convenience. In dynamic hidden-

action contracting, entropy regularization functions as an equilibrium-selection device, a computational stabilizer, and a practical hedge against approximation error, while preserving a clear economic interpretation as controlled randomization over contracts and actions.

10 Conclusion and open problems

We have studied a dynamic hidden-action principal–agent problem through the lens of entropy-regularized control. The substantive modeling move is to replace hard best responses—which are discontinuous in payoffs and therefore fragile in coupled dynamic systems—with soft responses induced by $\tau_a > 0$ for the agent and $\tau_p > 0$ for the principal. This yields a regularized subgame-perfect equilibrium (rSPE) characterized by coupled soft Bellman equations, and, in the finite tabular setting, a joint operator that is a γ-contraction in $\|\cdot\|_\infty$. The economic content of the regularization is not merely computational: it formalizes controlled randomization over actions and contracts, selects a unique stationary Markov equilibrium, and quantifies a transparent stability–bias tradeoff via the bound of order $(\tau_p \log|\mathcal{B}| + \tau_a \log|\mathcal{A}|)/(1-\gamma)$.

Several open problems are immediate once we move beyond the stylized assumptions needed for a clean contraction argument.

Partial observability and richer information structures. Our baseline assumes that the principal conditions only on the observed state s_t when selecting b_t, and that the outcome o_t is the only contractible signal. In practice, the principal often faces partial observability of the underlying economic state (demand, quality, macro conditions), while the agent may have private information about that state or about their own cost type. A natural extension replaces s_t with a latent state and allows the principal to condition on a history of signals, yielding a principal-side belief state and a POMDP-like object. Two challenges arise. First, the equilibrium object becomes a fixed point in policy space over beliefs rather than in finite-dimensional Q-tables, complicating existence and uniqueness results. Second, incentive constraints interact with filtering: contracts shape not only effort but also the informativeness of outcomes about hidden actions, and thus the evolution of beliefs. An appealing direction is to combine entropy regularization with belief-space dynamic programming, using softmax policies to maintain continuity while tracking approximate beliefs; however, it remains open to identify conditions under which a contraction-type argument survives (e.g., with finite belief grids or with additional regularity assumptions). Economically, such a framework would let us analyze how contract power should respond to uncertainty about the environment, and when the principal should optimally "pay for information" by inducing actions that make outcomes more diagnostic.

Robust and distributionally robust contracting. Another limitation of the baseline is correct specification of primitives (O, T, r_p). Contracting environments are often misspecified: the principal may distrust the mapping from actions to outcomes, or worry that rare events are under-sampled. This motivates robust MDP and distributionally robust (DR) variants in which the principal evaluates contracts under a worst-case model in an ambiguity set around $(\widehat{O}, \widehat{T})$. Conceptually, robustness is complementary to entropy: robustness addresses model uncertainty, while entropy addresses strategic discontinuities and computational brittleness. Technically, combining the two raises delicate questions because the principal's Bellman operator becomes a max–min–soft object. One promising route is to adopt convex ambiguity sets that preserve dynamic consistency (e.g., rectangular sets) so that the robust backup decomposes statewise, and then to study whether the resulting "robust soft" operator remains a contraction. On the economic side, robust contracting changes the interpretation of τ_p: the principal may rationally randomize not only for stability but also to hedge against misspecification, producing incentive schemes that are deliberately less sharp in environments with high ambiguity. We view the joint calibration of robustness radii and temperatures as a key deployment question: overly conservative ambiguity sets can dominate the effect of incentives, while too little robustness can resurrect the same regime-switching behavior that motivates smoothing.

Multiple agents, teams, and strategic interaction. Many organizations contract with multiple agents whose actions jointly determine outcomes, sometimes with complementarities (team production) and sometimes with competition (tournaments). Extending rSPE to these settings raises two kinds of issues. First, the informational externality of outcomes becomes more complex: an outcome informative about one agent may be confounded by others' hidden actions, affecting both incentives and learning dynamics. Second, equilibrium selection becomes more delicate because the principal's contract menu may induce multiple correlated-response patterns among agents. Entropy regularization is attractive here because it naturally induces interior mixed strategies and can smooth best-response correspondences in multi-agent games, but the fixed-point analysis becomes higher-dimensional and may require monotonicity or potential-game structure to recover uniqueness. A particularly relevant direction is networked contracting, where outcomes are local (peer effects, platform moderation, supply chains) and the principal chooses a vector of local contracts; understanding whether separability or approximate decomposability yields scalable equilibrium computation is open. From a policy perspective, these extensions matter for interpreting observed pay schemes in teams: what looks like "inefficiently low-powered incentives" may be rational once we account for

strategic spillovers, partial observability, and the value of stability.

Continuous contract spaces and mechanism design constraints. We have emphasized a finite menu $\mathcal{B}$, which is both computationally convenient and economically realistic in settings with standardized schemes. Yet many applications call for continuous contracts, additional constraints (limited liability, monotonicity, budget balance), or richer payment rules based on histories. A continuous $\mathcal{B}$ changes the nature of the principal's softmax: one would replace LSE_{τ_p} with a log-partition integral and interpret $\rho(\cdot \mid s)$ as a Gibbs density over contracts. This raises questions of existence (normalizability), computation (sampling in contract space), and economic interpretation (does the resulting randomization reflect genuine commitment or merely an approximation device?). Moreover, once $\mathcal{B}$ is large, our approximation bound highlights a tension: increasing $|\mathcal{B}|$ reduces discretization error but increases the worst-case entropy bias through $\log|\mathcal{B}|$. Developing principled discretization schemes—for example, adaptive menus that refine only where $Q_p(s,\cdot)$ is steep—is a practical open problem with direct relevance to organizational design.

How should we tune τ_a and τ_p in deployment? Temperatures are central because they govern the stability–bias frontier. In applications, we rarely observe τ directly, and we may not want to set it purely for numerical convenience. One interpretation treats τ as bounded rationality or implementation noise: τ_a captures the agent's limited optimization or unmodeled idiosyncrasies, while τ_p captures institutional frictions that prevent perfectly sharp contract selection. Under this view, τ is an estimand, to be fit to observed behavior. A second interpretation treats τ as a robustness and exploration knob chosen by the designer: higher τ yields smoother responses and reduces sensitivity to value estimation error (at an explicit bias cost). Practically, we can tune τ by validation on held-out episodes (or historical counterfactuals), selecting the smallest temperatures that avoid instability under plausible perturbations of Q or primitives. A useful operational heuristic is a homotopy schedule: start with larger (τ_a, τ_p) to obtain a stable fixed point, then gradually anneal toward smaller values while monitoring sensitivity (e.g., policy variation or worst-case value under model uncertainty). Formalizing such procedures—and linking them to guarantees under misspecification and function approximation—remains an important direction for making dynamic contracting algorithms both economically interpretable and operationally reliable.

More broadly, we view entropy-regularized equilibrium as a bridge: it retains the core economic structure of hidden-action contracting while importing the analytical and computational tools of soft dynamic programming. The remaining work lies in identifying the boundary of this bridge—

where partial observability, misspecification, and strategic complexity begin to dominate contraction-based guarantees—and in developing theory-guided heuristics that preserve the economic logic of incentives while meeting the practical demands of noisy data and large state spaces.