Budget-Feasible Incentivized Learning in Hierarchical Principal–Agent Trees: Shadow Prices, Constrained Efficiency, and Primal–Dual No-Regret

Liz Lemma Future Detective

January 16, 2026

Abstract

We study online decision-making in hierarchical principal–agent systems where learning agents interact through local, action-contingent transfers. Prior work shows that allowing one-step transfers in a tree of bandit learners can restore social efficiency asymptotically: selfish learners behave as if they were collaborating. However, modern platforms face binding payment constraints—limited liability, budget caps, and credit constraints—that prevent principals from "overpaying to force compliance." We introduce per-node payment budgets into the tree principal–agent bandit model and ask: what is the best welfare achievable, and can decentralized learning attain it? Our first contribution is a dual characterization of the budget-feasible welfare optimum: optimal incentives correspond to local minimal inducement payments weighted by endogenous shadow prices of budget that propagate up the hierarchy. Our second contribution is Budgeted-MAIL, a decentralized primal–dual learning algorithm that combines bandit learning on shadow-price-shifted rewards with online updates of each principal's budget multiplier. We prove sublinear welfare regret relative to the optimal budget-feasible benchmark and vanishing average budget violations under mild conditions. Finally, we show an impossibility gap: when budgets fall below inducement thresholds, full efficiency is unattainable and welfare loss is necessarily linear. The framework yields policy-relevant comparative statics—how welfare and required transfers scale with depth, branching, and budget tightness—and provides a tractable foundation for designing incentive budgets in 2026-era agentic supply chains and AI service marketplaces.

Table of Contents

2. 2. Related Work: Online contract design, principal–agent bandits, limited liability/budgets in contracting, primal–dual online learning; position relative to MAIL and contract-theory ML.

3. 3. Model: Tree structure, bandit rewards, action-contingent transfers, observability; introduce per-round (or average) budget constraints; define feasible contracts and induced play.

4. 4. Budget-Feasible Benchmarks: Define constrained welfare optimum and constrained-SPNE; characterize inducibility under budgets; discuss when full efficiency is impossible; simple examples (depth 2).

5. 5. Duality and Shadow-Price Characterization: Lagrangian for budgets; define shadow-price-shifted utilities; recursive (dynamic-programming) characterization of optimal induced actions and minimal payments under multipliers.

6. 6. Algorithm (Budgeted-MAIL): Local payment estimation (or conservative upper bounds), bandit subroutine on shadow-price-shifted rewards, and online dual updates per node; implementability and decentralization.

7. 7. Main Theorems: No-regret to constrained benchmark; budget feasibility guarantees; rates vs depth and branching; conditions for strong/weak feasibility (hard vs soft budgets).

8. 8. Impossibility and Lower Bounds: Threshold budgets for implementability of unconstrained optimum; linear welfare gap when budgets are too tight; propagation phenomena under constraints.

9. 9. Comparative Statics and Interpretation: Shadow prices as marginal value of budget; monotonicity in budgets; scaling with tree size; design implications for platform credit policies.

10. 10. Experiments / Simulations (Optional but Recommended): Toy trees + constrained budgets; welfare vs budget curves; learned shadow prices; deviation frequency; sensitivity to nonstationarity.

11. 11. Discussion and Extensions: Global (horizon) budgets, stochastic budgets, multi-parent DAGs, partial observability, menus; limitations and open questions.

1 Introduction and Motivation

Delegation has become the default organizational primitive in many "agent economies" of 2026: firms deploy stacks of semi-autonomous services (procurement bots, pricing agents, compliance monitors), platforms coordinate networks of third-party sellers and advertisers, and public agencies increasingly rely on automated intermediaries to route cases, schedule inspections, or allocate scarce resources. What unifies these settings is not only that decisions are distributed across a hierarchy, but also that the parties at different layers learn and adapt over time from noisy feedback. A regional manager experiments with promotion policies while local store agents respond; a marketplace tunes ranking parameters while sellers react strategically; a security team updates detection thresholds while downstream devices decide whether to comply or free-ride. The economics is classical delegation, but the operational reality is online: reward functions are initially unknown, experimentation is unavoidable, and the principal must shape agents' behavior through contracts that are themselves part of the learning loop.

In unconstrained models of repeated principal–agent interaction with observable actions, a clean benchmark often emerges: if transfers can be chosen freely, the principal can "buy" the agent's preferred action at a payment equal to the agent's opportunity cost, thereby aligning incentives at minimal cost. This logic underlies a broad class of results in contracting theory and, more recently, in the machine-learning literature on principal–agent bandits and multi-level incentive design. It is also seductive as an engineering narrative: if we can observe downstream actions and can pay (or subsidize) accordingly, then we can restore efficiency even when each component optimizes its own objective.

The point of departure in this paper is that real delegating organizations rarely possess such unconstrained transfer instruments. Budgets are hard, local, and often binding. A team lead can offer only limited bonuses to contractors; a platform can subsidize only a small amount of promotional credit; a regulator can provide only limited assistance or reimbursement; and in many cases, payments are constrained by limited liability, compliance rules, or accounting restrictions that effectively cap per-period transfers. These constraints matter precisely in the environments where delegation is most valuable: when downstream actions create upstream externalities (e.g., effort that improves a shared model, safety investments that reduce systemic risk, or data-quality choices that affect aggregate performance), and when uncertainty makes it optimal to explore actions that may be privately unattractive. In such cases, the principal may know what it would like the agent to do, yet simply be unable to fund the required inducements.

We therefore study a tree-structured, repeated delegation problem with bandit feedback in which each node is simultaneously an agent (to its parent) and a principal (to its children), and in which each principal faces a per-round budget cap on the total transfers it can pay to its children. Conceptually, we can view each edge as a "micro-contract" that attempts to align behavior locally, while the tree structure captures the fact that alignment must propagate through multiple constrained layers. The central tension is immediate and practical: budgets limit the set of implementable downstream action profiles, but learning and adaptation require the principal to repeatedly test recommendations and adjust contracts. When payments are scarce, the principal faces a three-way trade-off among (i) inducing high-welfare actions that are privately costly to agents, (ii) conserving budget for future rounds and future contingencies, and (iii) collecting information to improve recommendations.

To build intuition before formalism, consider a principal who would like each of several children to choose an action that benefits the principal but is privately dominated for the child. With unlimited funds, the principal can compensate each child exactly for the utility gap, implement the desired profile, and do so in a way that is robust to the child's own learning dynamics (since the recommended action becomes optimal given the transfer). Under a hard cap, however, even if each individual inducement is "small," the costs add up across many children; and when a single inducement is "large," it may be infeasible outright. The principal must then ration incentives: which child actions are worth buying, which can be left to the child's self-interest, and how should these choices adjust as the principal learns the payoffs? In a tree, the problem compounds. A node that is itself budget-constrained cannot reliably implement the behavior its parent would like to see from its subtree, so upstream policies must internalize downstream scarcity.

This paper aims to illuminate that trade-off with a model and analysis that are simultaneously economic (explicit incentive constraints, implementability, welfare benchmarks) and algorithmic (unknown rewards, bandit learning, finite-horizon performance). Our guiding message is that budget constraints do not merely slow convergence or add technical nuisance; they change the implementable set and thus create an intrinsic efficiency–feasibility frontier. When budgets bind, the correct object is not the unconstrained first-best, but a constrained welfare benchmark that accounts for limited inducement capacity at every principal. Moreover, the shadow value of budget—a Lagrange multiplier in the dual—becomes an economically meaningful "internal price" that can be learned online and used to coordinate decentralized behavior.

Our contributions are threefold.

- **A budgeted delegation benchmark and its economic structure.** We formalize the welfare-maximization problem subject to local incentive compatibility and per-round budget caps. The key conceptual output is that budgets act like a local tax on transfers: when a principal's budget is scarce, it effectively inflates the cost of inducing a downstream action. This reframes delegation under scarcity as a sequence of local choices over action recommendations, each trading off direct reward against shadow-price-adjusted inducement costs.

- **A dual (shadow-price) characterization that remains decomposable on a tree.** By attaching multipliers to budget constraints, we obtain a Lagrangian representation in which each node optimizes a shadow-price-shifted objective. This delivers an inductive recursion for continuation utilities and minimal inducing payments that mirrors the unconstrained "pay-the-gap" logic, but with a crucial difference: which gaps are worth paying depends endogenously on the shadow price. In practical terms, shadow prices provide a portable summary statistic of scarcity that can be communicated upward (as an implied willingness-to-pay for inducement capacity) without requiring centralized optimization over the whole tree.
- **A decentralized primal–dual learning algorithm and an unavoidable gap under tight budgets.** We propose a simple Budgeted-MAIL procedure in which each principal combines (i) a bandit learner that chooses its own action and recommended child actions using shadow-price-shifted rewards and (ii) a dual update that raises the shadow price when spending threatens to exceed the cap. The algorithm is decentralized by construction: each node uses local observations (its own reward and observed child actions) and local budgets. At the same time, we show that if budgets fall below the inducement threshold for the unconstrained welfare-optimal profile, then no mechanism—even with full information—can avoid a linear welfare loss relative to that first-best. This impossibility result clarifies what performance guarantees are meaningful when transfers are genuinely limited.

Beyond these headline results, the analysis offers comparative statics that speak to organizational design. Increasing branching factor increases the demand for inducement capacity and tends to raise shadow prices; deeper hierarchies propagate scarcity upward, amplifying the welfare cost of tight lower-level budgets; and environments with large private-vs-social incentive gaps are precisely those in which local caps are most distortionary. These patterns align with familiar managerial prescriptions (e.g., concentrate incentive budgets where externalities are largest, simplify overly deep delegations, avoid assigning agents tasks with misaligned objectives when incentives are inelastic), but our framework makes these prescriptions operational in an online, learning setting.

We also acknowledge limitations. Our model assumes action observability along edges, which is realistic in some digital settings (logged actions, verifiable compliance events) but not in all labor or procurement environments. We focus on per-round hard caps rather than intertemporal budgeting with borrowing and carryover, an abstraction that is appropriate for many compliance-driven settings but may be too restrictive for firms with flexible compensation pools. Finally, our welfare lens abstracts from distributional concerns about who holds the budget and who bears risk; extending the framework to risk aversion, fairness constraints, or endogenous budget allocation across principals is an important direction for future work.

Roadmap. We begin by situating our contribution in the literatures on online contract design, principal–agent bandits, limited liability, and primal–dual online learning (Section 2). We then present the model and the constrained welfare benchmark (Section 3), develop the dual characterization and the recursive structure of shadow-price-adjusted inducement (Section 4), and analyze the decentralized Budgeted-MAIL algorithm and its regret properties (Section 5). We conclude with the impossibility gap and a discussion of when tight budgets should be viewed as a feature (governance and safety) rather than a bug (efficiency loss), along with implications for the design of agentic organizations and platforms (Section 6).

2 Related Work

Our paper sits at the intersection of (i) dynamic/online contract design, (ii) principal–agent learning models in which incentives shape exploration and behavior, (iii) limited-liability and budget constraints in contracting, and (iv) primal–dual methods for online learning under constraints. The closest conceptual ancestor is the emerging literature that treats contracts not as one-shot objects but as adaptive control instruments in environments with unknown payoffs and strategic agents. Our contribution is to bring a particularly stark and operational scarcity constraint—*hard, per-round payment caps at every principal in a delegation tree*—into that interface, and to show that the resulting economic structure is naturally captured by shadow prices that can be learned in a decentralized way.

Online contract design and dynamic delegation. Classical dynamic contracting studies how a principal shapes effort or information acquisition over time under hidden actions or hidden information (e.g., **?**, **?**). In these models, continuation values and intertemporal incentives play a central role, and the main difficulty is typically informational (moral hazard/adverse selection) rather than computational or statistical. A more recent line of work, motivated by platform governance and algorithmic management, emphasizes contracts that must adapt under uncertainty about the environment itself. Our setting shares the dynamic and recursive flavor of these models but differs along two dimensions. First, we adopt *observable* actions along edges, so local incentive constraints take a simple “recommendation plus transfer” form; the difficulty comes from limited transfer instruments and from bandit feedback on rewards. Second, we focus on *hierarchical* organizations (a rooted tree) in which each node is both principal and agent; recursion arises not only from intertemporal continuation values but also from delegation depth.

In the theory of organizations, hierarchical contracting and delegation are standard primitives (e.g., **?**). What is less standard is to couple those primitives with online learning, where policies must be evaluated and improved while agents simultaneously best-respond to the evolving contract environment. We view our model as a step toward an economic theory of "agentic" hierarchies in which incentive design and experimentation are inseparable operational tasks.

Principal–agent bandits and incentivized learning. A parallel machine-learning literature studies bandit and reinforcement-learning problems with strategic responders. This includes work on incentivizing exploration and recommendation compliance, often in settings where a principal has a stream of users/agents with private preferences, and must trade off short-run welfare with long-run learning (see, e.g., **?** and follow-ups). Our setting differs in that the same strategic agents persist over time and learn via no-regret dynamics, which makes compliance constraints "local" and repeatedly relevant rather than a one-shot persuasion problem.

Most closely related are principal–agent bandit models with action-contingent payments and observable actions, in which the principal can align incentives by paying an agent the minimal amount needed to make a recommended action optimal. In a tree, this logic becomes a recursive mechanism: inducing an action at one level may require that the agent, acting as a principal, can in turn induce behavior downstream. The MAIL framework of **?** formalizes this recursion and shows that, absent binding payment constraints, a decentralized learning rule can achieve strong welfare guarantees by effectively implementing "pay-the-gap" inducements throughout the hierarchy. Our paper is positioned as a scarcity-aware analogue: we keep the same core observability and learning assumptions but impose binding caps on transfers at *every* principal, which breaks the unconstrained implementability benchmark and forces a new welfare benchmark, new comparative statics, and a different algorithmic architecture (primal–dual rather than purely primal).

Limited liability and budgets in contracting. Limited liability, cash constraints, and budget caps are among the most studied frictions in contract theory (e.g., **?**; see also survey discussions in **?**). In many models, limited liability alters the shape of optimal contracts by truncating punishments or restricting the principal's ability to extract surplus. In our model, the friction has a simpler but sharper interpretation: even when actions are observable and the principal would like to "buy" a privately costly action, the required inducement may be infeasible because total payments cannot exceed ρ_v in a given round. This creates an implementability constraint that is *combinatorial* across children (costs add across edges) and *propagates* up the hierarchy (a node cannot promise what it cannot fund).

We also emphasize that per-round caps capture institutional realities that are not well modeled by intertemporal borrowing: compliance regimes that limit bonuses per period, platform subsidy policies with hard caps, procurement rules, and safety governance that intentionally restrict the ability of any single manager or subsystem to "bribe" downstream components. By treating these caps as hard constraints rather than soft penalties, we obtain an explicit efficiency–feasibility frontier: some action profiles are not merely suboptimal but *infeasible* to induce, even with full information.

Learning under constraints: bandits with budgets and online Lagrangians. From an algorithmic perspective, our approach connects to online learning with resource constraints, including bandits with knapsacks and related constrained online optimization problems (e.g., **?**, **?**). In those models, a learner chooses actions that consume a global budget over time, and the key technique is often a primal–dual reduction: maintain a shadow price for the scarce resource, optimize a price-adjusted reward, and update the price based on observed consumption. Our setting shares the mathematical motif but differs economically: budgets are not "consumed" by the principal's own actions, but by transfers that must be paid *to satisfy incentive constraints* for other strategic learners. Moreover, budgets are *local* to each node and apply per round, rather than being a single global knapsack over a horizon. This locality is what makes the dual decomposition economically meaningful: λ_v is naturally interpreted as the marginal value of inducement capacity at principal v, and it enters exactly as a "tax" on the payments that v would like to make.

Our Budgeted-MAIL algorithm draws on the broad toolbox of online primal–dual methods and constrained online convex optimization (e.g., **?**, **?**, **?**). The conceptual difference is that the primal decision at each node is not a direct action but a *recommendation-and-transfer menu* subject to downstream best responses, and the regret object is *welfare* relative to an implementable benchmark rather than reward relative to a fixed arm. We therefore combine two kinds of guarantees: no-regret behavior of agents under contracts (a behavioral assumption aligned with the principal–agent bandit literature) and primal–dual control of budget violations (an algorithmic guarantee aligned with constrained OCO).

Mechanism design on trees and decentralized governance. A final connection is to mechanism design and incentives in networks and hierarchies. Tree-structured environments are attractive because they permit recursion and local message passing; this is a reason they appear both in organizational economics and in distributed systems. Our dual characterization can be read as a welfare-relevant "price system" for inducement capacity: each principal learns a local multiplier and, through the induced behavior of its subtree, effectively communicates scarcity upstream without centralized coordination. This resonates with market-design intuitions (prices summarize constraints) but in a setting where what is being priced is not a physical input but the ability to alter other learners' behavior via transfers.

Summary of our position. Relative to unconstrained principal–agent bandit models (and MAIL in particular), we ask what survives when transfers are scarce in a hard sense. The answer is that the pay-the-gap logic remains locally correct *conditional on feasibility*, but optimal recommendations become shadow-price-dependent and, in some instances, first-best welfare is unattainable regardless of learning. Relative to constrained bandit/OCO work, we contribute an economic mapping from dual variables to implementable contracts in a multi-level delegation environment, and we identify an impossibility gap that is not an artifact of finite-time learning but a genuine implementability loss induced by budgets. This sets the stage for the model in Section 3, where we formalize the tree game, define budget-feasible contracts, and state the constrained welfare benchmark that our analysis targets.

3 Model

We study a repeated principal–agent interaction on a rooted tree that captures hierarchical delegation with *observable* actions but *unknown* rewards. Time is discrete, indexed by rounds $t = 1, \ldots, T$. The organization is a rooted tree $G = (V, E)$ with depth D, where each node $v \in V$ is simultaneously (i) an *agent* to its parent $P(v)$ (undefined for the root), and (ii) a *principal* to its children $C(v)$. We write $C(v) = \emptyset$ for leaves, and we allow heterogeneous branching (a uniform branching factor B is a convenient special case).

Actions and contracts. Each node v chooses an action A_t^v every round from a finite set A with $|A| = K$ (allowing node-specific sets A_v is straightforward). In addition, each principal v can offer to each child $w \in C(v)$ a simple *action-contingent contract* consisting of a recommended action and a nonnegative transfer:

$$(B_t^w, \tau_t(w)) \in A \times \mathbb{R}_+.$$

The interpretation is deliberately minimal: principal v recommends that child w take action B_t^w, and commits to pay $\tau_t(w)$ if (and only if) w complies, i.e., if $A_t^w = B_t^w$. We impose nonnegativity (limited liability) at the edge level, so we do not allow fines.

A node v may itself receive a contract $(B_t^v, \tau_t(v))$ from its parent $P(v)$. To streamline notation, we treat $(B_t^v, \tau_t(v))$ as absent at the root and set $\tau_t(v) = 0$ when v has no parent. We emphasize that contracts are *local*: a principal can condition transfers only on the child action along that edge, not on unobserved outcomes elsewhere in the tree.

Observability and within-round timing. We adopt an informational structure tailored to recursive delegation. Along each edge $(v, w) \in E$, the parent observes the child action. Formally, after play in round t, each node v observes the realized reward X_t^v of its own node as well as the realized actions of all children $\{A_t^w : w \in C(v)\}$ and whether each promised transfer was paid. No node observes other nodes' realized rewards.

Within each round, interaction unfolds top-down. At the beginning of round t, each node v observes the contract offered by its parent (if any). Then, starting from the root and proceeding down the tree, each node chooses its action A_t^v and simultaneously offers contracts $\{(B_t^w, \tau_t(w))\}_{w\in C(v)}$ to its children, subject to the budget constraint described below. Children then choose their actions after observing the offered contracts, and so on until the leaves act. Rewards then realize, and information is revealed upward (actions are observable along edges).

This sequential timing makes the recursion operational: when v decides what to offer a child w, the contract shapes w's action directly, but it may also affect how w subsequently contracts with its own descendants and thus the induced behavior in the entire subtree rooted at w.

Bandit rewards on nodes. Each node v has an unknown mean reward function

$$\theta_v : A \times A^{|C(v)|} \to [0, 1], \qquad \theta_v(a_v, a_{C(v)}),$$

which depends on v's own action and the action profile of its children. (Leaves have $\theta_v : A \to [0, 1]$.) The realized reward in round t is

$$X_t^v \; = \; \theta_v\big(A_t^v, A_t^{C(v)}\big) \; + \; \varepsilon_t^v,$$

where $\{\varepsilon_t^v\}$ are mean-zero i.i.d. sub-Gaussian noise terms (uniformly over v and t). Thus each node faces a local bandit problem: it observes only the realized payoff of the action it actually took (and the actions of its children), but does not observe counterfactual rewards for actions not chosen, nor the mean function θ_v.

The dependence of θ_v on children actions is the key economic externality: upstream nodes may benefit from downstream behavior that downstream agents would not privately choose absent incentives. Transfers are the instrument used to internalize these local externalities along the edges of the tree.

Utilities and transfer accounting. Utilities are quasi-linear in transfers and coincide with realized rewards plus incoming payments minus outgoing payments. Specifically, node v's per-round utility is

$$U_t^v \;=\; X_t^v\big(A_t^v, A_t^{C(v)}\big) \;+\; \mathbf{1}\{A_t^v = B_t^v\}\,\tau_t(v) \;-\; \sum_{w\in C(v)} \mathbf{1}\{A_t^w = B_t^w\}\,\tau_t(w).$$

Transfers are purely internal redistribution and cancel in aggregate. Consequently, the object of interest from a system designer's perspective is *social welfare*, defined as the sum of mean rewards across nodes and time:

$$\mathrm{SW}_T \;=\; \sum_{t=1}^{T}\sum_{v\in V}\theta_v\big(A_t^v, A_t^{C(v)}\big).$$

Because rewards depend on induced actions throughout the tree, welfare is determined jointly by learning (discovering high-reward action profiles) and incentives (making those profiles behaviorally implementable).

Hard per-round budget constraints (limited inducement capacity). Our central friction is that each principal has a *hard* cap on total transfers it can promise in a round. For each node v, let $\rho_v \in [0,1]$ denote its per-round budget. Then feasibility requires

$$\sum_{w\in C(v)} \tau_t(w) \;\le\; \rho_v, \qquad \forall v\in V,\ \forall t\in\{1,\dots,T\}.$$

This constraint is local (one constraint per principal), contemporaneous (applies each round), and applies to promised transfers rather than realized transfers. It captures institutional settings in which principals cannot borrow against future budgets and cannot condition payments on future outcomes: a manager cannot promise more bonus payments than a compliance rule allows; a platform cannot subsidize beyond a daily cap; a subsystem cannot allocate more "incentive mass" than a safety policy permits.

For some arguments it is useful to compare the hard cap to an *average-budget* relaxation, in which one replaces the per-round constraint by a long-run constraint of the form

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{w\in C(v)} \tau_t(w) \;\le\; \rho_v.$$

Our main focus, however, is the hard constraint above, because it directly restricts which action profiles can be induced in a given round and therefore generates a sharp implementability frontier.

Policies, histories, and induced play. A *contracting policy* for node v specifies, at each round t, how v chooses its own action A_t^v and its outgoing contracts $\{(B_t^w, \tau_t(w))\}_{w \in C(v)}$ as a function of v's information: its past rewards $\{X_s^v\}_{s<t}$, observed past children actions $\{A_s^w\}_{w \in C(v),\, s<t}$, and past contracts and transfer realizations on incident edges. An *agent policy* for node v specifies how v responds to the current incoming contract $(B_t^v, \tau_t(v))$ and its own history when choosing A_t^v. Together, the profile of policies induces a distribution over action and reward trajectories.

To connect the model to the principal–agent logic, we require that agents respond to the current contract according to a one-step best-response condition with respect to their own continuation value for the round. Concretely, fix a node w and a round t. Conditional on the contract $(B_t^w, \tau_t(w))$ offered by $P(w)$, node w chooses an action that maximizes its (possibly history-dependent) expected utility for that round:

$$A_t^w \in \arg\max_{a \in A} \Big\{ \mu_{w,t}^{\text{cont}}(a) + \mathbf{1}\{a = B_t^w\} \tau_t(w) \Big\},$$

where $\mu_{w,t}^{\text{cont}}(a)$ summarizes w's expected intrinsic payoff from choosing a, including the effect of how that choice changes downstream behavior through w's own contracts to $C(w)$. In the full-information version of the model, $\mu_{w,t}^{\text{cont}}$ is induced by the known θ's; in our learning environment, it is the object that w learns over time via bandit feedback while simultaneously responding to incentives.

Operationally, we assume nodes follow no-regret bandit learning dynamics conditional on the contracts they receive (a high-probability action-regret condition). This behavioral assumption is the bridge between incentives and learning: it ensures that, when a principal offers a transfer sufficient to make a recommendation locally optimal, the agent will converge to compliance up to vanishing regret.

Feasible contracts and implementability. A profile of contracts in round t is *budget-feasible* if it satisfies the per-round caps at every principal. Given budget feasibility, a recommended action profile is *inducible* in round t if there exist transfers (respecting budgets) such that every node finds it optimal to take its recommended action given its incoming contract and its own contracting problem with its children. In depth $D > 2$, inducibility is inherently recursive: to induce a child to take an action that is privately costly, the parent must pay; but that payment must itself be compatible with the child having enough budget to induce the behavior that makes the action attractive in its subtree. This recursive feasibility is precisely what hard budgets make economically salient: the hierarchy can fail not because agents are unresponsive, but because principals lack the inducement capacity to align incentives through multiple constrained layers.

The remainder of the paper uses this model to define a constrained welfare benchmark—the best achievable welfare under incentive compatibility and budget feasibility—and to design decentralized learning rules that approach it.

4 Budget-Feasible Benchmarks

Our learning goal is inherently comparative: we can only evaluate a decentralized algorithm relative to what is *achievable* under the same incentive and budget frictions. We therefore formalize a constrained welfare benchmark—the best performance attainable by any (possibly history-dependent) contracting scheme that is incentive compatible along every edge and respects the hard per-round budget caps. Doing so also clarifies the economic content of the budgets: they do not merely slow learning, but can shrink the set of implementable action profiles and thereby generate an intrinsic efficiency loss even under full information.

Budget-feasible, incentive-compatible play. Fix a round t. A (pure) *contract profile* is a collection $\{(B_t^w, \tau_t(w))\}_{(v,w)\in E}$ together with an action profile $\{A_t^v\}_{v\in V}$. We say the contract profile is *budget feasible* if each principal respects its cap,

$$\sum_{w\in C(v)} \tau_t(w) \le \rho_v \qquad \forall v \in V.$$

Given feasibility, we say the recommended actions are *locally incentive compatible* if every node w (including internal nodes) is willing to follow its recommendation given its one-step continuation payoff. Using the notation from the model, this condition takes the form

$$A_t^w \in \arg\max_{a\in A} \Big\{ \mu_{w,t}^{\text{cont}}(a) + \mathbf{1}\{a = B_t^w\}\tau_t(w) \Big\}, \qquad \forall w \in V,$$

with the understanding that $(B_t^w, \tau_t(w))$ is absent for the root. Intuitively, $\mu_{w,t}^{\text{cont}}(a)$ aggregates what w expects to earn (net of its own outgoing transfers) when it chooses a, accounting for how a affects contracting and behavior in the subtree below w. The key point is that budgets matter *twice*: they constrain a principal's ability to induce its children, and they also constrain an agent's ability to shape its own continuation payoff by inducing its descendants.

The constrained welfare optimum. To benchmark welfare, we adopt a full-information yardstick in which the mean reward functions $\{\theta_v\}$ are treated as known to the benchmark planner, but the planner is constrained to use the same local contracts, limited liability, and per-round budgets as in the actual game. Formally, let $\Pi(\rho)$ denote the set of (possibly randomized, history-dependent) policy profiles such that along every sample path: (i) all outgoing transfers satisfy the hard caps ρ_v in every round, and (ii) agents' action choices are sequentially rational given the offered contracts (equivalently, satisfy the local one-step best-response condition induced by the policies in the continuation). We define the *budget-feasible welfare benchmark* as

$$\mathrm{OPT}(\rho) \;=\; \max_{\pi \in \Pi(\rho)} \mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{v \in V} \theta_v\big(A_t^v, A_t^{C(v)}\big) \right].$$

Because transfers cancel in welfare and the environment is stationary across rounds, $\mathrm{OPT}(\rho)$ can be interpreted as the welfare achieved by an optimal *stationary* constrained contracting pattern repeated each period (randomization may still be useful to convexify implementability). The important feature of $\mathrm{OPT}(\rho)$ is that it benchmarks *efficiency subject to inducement capacity*: it already internalizes the fact that some downstream actions are too expensive to elicit when budgets are tight.

Constrained equilibria (constrained-SPNE). The decentralized game induces equilibrium restrictions beyond feasibility: principals are not coordinated by a central planner, and each node acts in its own interest given anticipated downstream responses. Accordingly, it is useful to also name the strategic benchmark: we call a policy profile a *budget-feasible subgame perfect Nash equilibrium* (constrained-SPNE) if it is a SPNE of the extensive-form game induced by our timing and observability, with the additional restriction that every principal's outgoing transfers satisfy $\sum_{w \in C(v)} \tau_t(w) \le \rho_v$ at every history. In the full-information version of the model, a constrained-SPNE exists under mild compactness assumptions (e.g., allowing mixed contracts), and its equilibrium welfare cannot exceed $\mathrm{OPT}(\rho)$. For our learning results, $\mathrm{OPT}(\rho)$ is the appropriate welfare target; constrained-SPNE is the appropriate interpretive notion for decentralization.

Inducibility under hard budgets. A central object in both benchmarks is the set of action profiles that a principal can induce from its children in a given round. The observable-action, action-contingent contract structure implies a simple logic: to make a recommendation b optimal for a child w, the parent must compensate w for the maximum utility loss w would incur from complying rather than taking its best alternative.

To make this precise, fix a node w and suppose we consider the full-information continuation problem in the subtree rooted at w, taking as given that w will optimally contract with $C(w)$ subject to its own budget ρ_w. Let $\mu_w(a)$ denote w's resulting one-round expected intrinsic utility (own reward minus optimal outgoing transfers) when w chooses action a at the top of its subtree. Then the *minimal inducing payment* needed to make b optimal for w is

$$\tau_b^\star(w) \;=\; \max_{a\in A} \mu_w(a) - \mu_w(b) \;\geq\; 0.$$

This is the smallest transfer that (weakly) closes the utility gap between b and w's best deviation. Under limited liability, there is no cheaper way to enforce b because the parent cannot punish noncompliance.

Given these edge-wise costs, a parent v can induce a recommended profile $b_{C(v)} = (b_w)_{w\in C(v)}$ in one step only if it can afford the sum of the necessary gap payments:

$$\sum_{w\in C(v)} \tau_{b_w}^\star(w) \;\leq\; \rho_v.$$

This condition highlights the economic role of the branching factor: even when each individual child is cheap to incentivize, the total inducement cost scales additively across children, so fixed ρ_v turns inducement into a knapsack-type choice over which downstream behaviors to internalize.

Depth-2 example: closed-form feasibility and "pay-the-gap." The logic above becomes fully transparent in the canonical depth-2 case where the children are leaves. Then $\mu_w(a) = \theta_w(a)$ because w has no descendants and no outgoing transfers. Hence

$$\tau_b^\star(w) = \max_{a\in A} \theta_w(a) - \theta_w(b),$$

and v can implement $b_{C(v)}$ if and only if $\sum_{w\in C(v)} \tau_{b_w}^\star(w) \leq \rho_v$. Moreover, whenever implementation is feasible, the minimum-cost contract is simply to pay each leaf exactly its gap payment. In this sense, with observable actions, budgets do not change *how* we incentivize a fixed recommendation—they change *which* recommendations are feasible.

A minimal illustration uses $K = 2$ actions, $A = \{\ell, h\}$. Let a leaf w privately prefer ℓ, say $\theta_w(\ell) = 1$ and $\theta_w(h) = 0$, while the parent v benefits from h, say $\theta_v(a_v, h) = 1$ and $\theta_v(a_v, \ell) = 0$ (holding a_v fixed). Then $\tau_h^\star(w) = 1$. If $\rho_v \geq 1$, the parent can induce h and attain welfare $1+0$ at nodes (v, w) in each round; if instead $\rho_v < 1$, inducing h is impossible even with full information, and the unique feasible recommendation is ℓ, yielding welfare $0 + 1$. The *direction* of the welfare loss depends on where the externality lies: budgets prevent the upstream node from internalizing the downstream private cost, so the organization may systematically choose actions that are privately attractive but socially misaligned.

When full efficiency is impossible (an impossibility gap). The preceding example also isolates the key impossibility phenomenon: if the unconstrained welfare-maximizing action profile requires, at some principal v, total minimal inducement exceeding ρ_v, then no mechanism in our class can implement that profile. Because the constraint is hard and contemporaneous, repeated interaction does not help: the principal cannot "make up" for a shortfall today by promising more tomorrow, nor can it use negative transfers to economize on payments. Consequently, there exist environments in which any budget-feasible contract scheme—even under full information and perfect rationality—incurs a constant per-round welfare gap relative to the unconstrained optimum, and hence a linear loss $\Omega(T)$ over horizon T.

This observation motivates two design choices in the remainder of the paper. First, we evaluate learning performance against $\mathrm{OPT}(\rho)$, not against an infeasible unconstrained benchmark. Second, we explicitly track how budgets reshape inducement incentives throughout the hierarchy: when a downstream action is expensive to elicit, the constrained optimum may rationally substitute toward cheaper-to-induce behaviors, even if they are locally less productive. The next section formalizes this substitution through a dual representation in which budgets appear as endogenous shadow prices that "tax" transfers and propagate upward through the tree.

5 Duality and Shadow-Price Characterization

Budgets fundamentally couple a principal's inducement decisions across children: even with full information, a node cannot independently choose the "best" recommendation for each child if the resulting gap payments are jointly unaffordable. A convenient way to separate this coupling—both analytically (to characterize the benchmark) and algorithmically (to design decentralized updates)—is to attach an endogenous *shadow price* to each node's budget and study the resulting priced problem.

A Lagrangian for hard per-round budgets. Fix the full-information benchmark problem (so the mean rewards are known) and consider one generic round, suppressing time indices for readability. For each principal v, the hard cap $\sum_{w \in C(v)} \tau(w) \leq \rho_v$ is a local feasibility constraint on its outgoing transfers. Introducing a multiplier $\lambda_v \geq 0$ for this constraint yields the Lagrangian in which paying one unit of transfer by v carries an additional marginal penalty λ_v. Because each node's realized utility already subtracts its outgoing transfers, the shadow price effectively inflates the cost of paying by a factor $(1+\lambda_v)$: in the priced problem, a transfer $\tau(w)$ reduces v's priced objective by $(1+\lambda_v)\tau(w)$, while $\lambda_v \rho_v$ is a constant rebate term.

Formally, for a fixed multiplier vector $\lambda = (\lambda_v)_{v \in V}$, we evaluate a contract/action profile by the priced objective

$$\sum_{v \in V} \Big(\theta_v(a_v, a_{C(v)}) - \sum_{w \in C(v)} (1+\lambda_v)\tau(w) + \lambda_v \rho_v \Big),$$

subject to the same local incentive constraints that make each recommendation sequentially rational. The key point is that λ removes the *hard* coupling from the objective (budgets are no longer constraints), replacing it with a *soft* coupling through prices. When λ_v is large, node v behaves as if it faces a steep internal "tax" on transfers and therefore substitutes toward cheaper-to-induce downstream actions.

Shadow-price-shifted continuation utilities. To obtain a recursive characterization, we define priced continuation objects that mirror the tree structure. Fix a node v and suppose that the continuation problems in the subtrees rooted at its children are summarized by action-indexed values $\mu_w^\lambda(a)$ for each $w \in C(v)$. We interpret $\mu_w^\lambda(a)$ as the maximal *priced* expected utility attainable in the subtree rooted at w when w is induced to choose top-level action a, and then optimally contracts with its own children (recursively) in the priced sense.

Given these child values, if v chooses its own action $a \in A$ and recommends an action profile $b_{C(v)} = (b_w)_{w\in C(v)}$ to its children, then the priced one-step payoff to v (including continuation from children) is

$$\mu_v^\lambda(a, b_{C(v)}) \;=\; \theta_v\big(a, b_{C(v)}\big) + \sum_{w\in C(v)} \mu_w^\lambda(b_w) - \sum_{w\in C(v)} (1{+}\lambda_v)\, \tau_{b_w}^{\star,\lambda}(w) + \lambda_v \rho_v, \tag{1}$$

where $\tau_{b_w}^{\star,\lambda}(w)$ is the *minimal* transfer required to make b_w optimal for w given its priced continuation. The corresponding priced value of choosing a at v is

$$\mu_v^\lambda(a) \;=\; \max_{b_{C(v)}\in A^{C(v)}} \mu_v^\lambda(a, b_{C(v)}), \qquad \mu_v^{\star,\lambda} \;=\; \max_{a\in A} \mu_v^\lambda(a).$$

The root's value $\mu_r^{\star,\lambda}$ (for root r) is the priced objective of the entire tree under λ.

Two features of (1) are worth emphasizing. First, the $\mu_w^\lambda(b_w)$ terms propagate the benefits of downstream welfare upward, so that a parent trades off its own reward against improvements achievable in each child subtree. Second, $(1 + \lambda_v)$ appears *only* on transfers paid by v, capturing precisely that the scarcity is local: even if it is cheap for w to pay its children, this does not relax v's cap, and so it does not enter v's price.

"Pay the gap" remains optimal under prices. The priced recursion relies on the fact that, for any fixed recommendation b to a child w, we never want to overpay w: transfers are costly in (1), and, with observable actions and limited liability, only the event $A^w = b$ can be rewarded. Hence the minimal payment that induces b is still the relevant object.

Concretely, define the priced continuation value for w as $\mu_w^\lambda(\cdot)$ as above. Then the least transfer that makes b a best response for w (weakly) is

$$\tau_b^{\star,\lambda}(w) \;=\; \max_{a\in A}\mu_w^\lambda(a) \;-\; \mu_w^\lambda(b) \;\geq\; 0. \tag{2}$$

Necessity follows from the one-step IC inequality for w: to prevent deviation to the best alternative action, the contract must cover the maximal utility gap. Sufficiency follows because paying exactly $\tau_b^{\star,\lambda}(w)$ makes b attain the same continuation utility as the best deviation, and any tie-breaking can be handled by an arbitrarily small perturbation (or by allowing mixed actions/contracts). Importantly, (2) is *recursive*: what it costs to induce b at w depends on how valuable w's own downstream inducements are under λ.

Backward induction on the tree (dynamic programming). Equations (1)–(2) define a dynamic program indexed by λ. At leaves w, we have $C(w) = \emptyset$, so

$$\mu_w^\lambda(a) \;=\; \theta_w(a) + \lambda_w\rho_w, \qquad \tau_b^{\star,\lambda}(w) \;=\; \max_{a\in A}\theta_w(a) - \theta_w(b),$$

(where the $\lambda_w\rho_w$ term is constant in a and thus irrelevant for the argmax). For an internal node v, assume inductively that we have computed $\mu_w^\lambda(\cdot)$ and thus $\tau^{\star,\lambda}(\cdot)$ for all $w \in C(v)$. Then we compute $\mu_v^\lambda(a, b_{C(v)})$ via (1), maximize over $b_{C(v)}$ to obtain $\mu_v^\lambda(a)$, and finally form the inducing payments $\tau_b^{\star,\lambda}(v)$ via (2) when v itself is treated as an agent of $P(v)$.

This recursion delivers a transparent economic interpretation. For fixed λ, node v chooses recommendations $b_{C(v)}$ as if it faced a menu of child actions, where selecting b_w yields continuation benefit $\mu_w^\lambda(b_w)$ but carries priced cost $(1 + \lambda_v)\tau_{b_w}^{\star,\lambda}(w)$. Thus λ_v acts as an internal exchange rate between downstream utility gains and current budget consumption, and the recursion makes explicit how tightness in one layer propagates upward through μ^λ.

The dual viewpoint and its limitations. Let $g(\lambda)$ denote the optimal priced value obtained by applying the recursion above (equivalently, maximizing the Lagrangian-relaxed objective under IC). For any $\lambda \geq 0$, $g(\lambda)$ upper bounds the constrained optimum, and the dual problem is to minimize this upper bound over λ:

$$\inf_{\lambda\geq 0} g(\lambda).$$

When a strong duality argument applies (typically requiring a convexification, e.g., randomized contracts or an average-budget relaxation), a minimizing $\lambda^\star$ can be interpreted as the *shadow prices* of the original hard caps:

$\lambda_v^\star > 0$ only when v's budget is effectively binding (in the complementary-slackness sense), while $\lambda_v^\star = 0$ when v has slack and behaves as if unconstrained. Under pure, per-round hard caps, the primal problem is generally nonconvex (discrete actions, knapsack-like feasibility), so exact strong duality need not hold; nevertheless, the priced recursion remains the right organizing principle for both comparative statics and learning. In particular, Budgeted-MAIL will treat λ_v as an *online* price updated from observed spending, and will learn behavior that approximately optimizes the priced objective while driving long-run budget violations to zero.

6 Budgeted-MAIL: Decentralized Primal–Dual Learning with Budget-Feasible Inducement

The shadow-price recursion in Section 5 suggests an operational lesson: if we knew the right multipliers λ, then each node could behave *as if* it faced a stable internal price of budget, paying the (priced) gap when it wants to induce a child action and otherwise economizing on transfers. Budgeted-MAIL instantiates this idea online, under bandit feedback and without a central coordinator. Each node v runs two coupled learners: a *primal* bandit routine that selects (i) its own action and (ii) recommendations to its children to maximize a shadow-price-shifted objective, and a *dual* update that raises (resp. lowers) its local shadow price when it overspends (resp. underspends) relative to ρ_v. The only information exchanged is along edges: contracts $(B_t^w, \tau_t(w))$, observed actions, and (optionally) low-dimensional continuation summaries used to compute conservative gap payments.

Local payment estimation: conservative gap upper bounds. A distinctive difficulty relative to the unconstrained setting is that we cannot rely on "pay extra and learn" when budgets bind. If v ever offers a transfer that is too small to make b optimal for a child w, then w may deviate, corrupting the parent's learning signal; if v offers a transfer that is too large, it may violate ρ_v. Budgeted-MAIL therefore uses *conservative* estimates of the minimal inducing payment.

At a high level, we ask each child w to maintain bandit estimates of its priced continuation values $\mu_w^\lambda(\cdot)$ (for its current local multiplier λ_w), and to expose to its parent an upper-confidence proxy $\widehat{\mu}_{w,t}(\cdot)$ together with a confidence radius $\beta_{w,t}$ such that, with high probability,

$$\left|\widehat{\mu}_{w,t}(a) - \mu_w^\lambda(a)\right| \le \beta_{w,t} \qquad \forall a \in A.$$

(We discuss implementability below; in the simplest instantiation, $\widehat{\mu}_{w,t}(a)$ is an empirical mean of a shaped reward and $\beta_{w,t}$ is a standard sub-Gaussian

UCB radius.) Given this, the parent v computes, for each candidate recommendation $b \in A$, an *upper bound* on the true priced gap $\tau_b^{\star,\lambda}(w)$ from (2):

$$\widehat{\tau}_{b,t}(w) \ := \ \Big(\max_{a\in A} \widehat{\mu}_{w,t}(a)\Big) - \widehat{\mu}_{w,t}(b) \ + \ 2\beta_{w,t}. \tag{3}$$

When the confidence event holds, $\widehat{\tau}_{b,t}(w) \geq \tau_b^{\star,\lambda}(w)$, so paying $\widehat{\tau}_{b,t}(w)$ makes b a (weak) best response for w under the priced continuation. This "optimism on costs" is intentionally asymmetric: we would rather slightly overpay early (while still respecting ρ_v) than underpay and lose control of the induced action.

Two practical remarks are important. First, if v is under a *hard* cap, overpayment must be controlled by restricting which profiles $b_{C(v)}$ are ever recommended. Second, when K is large, sending $\widehat{\mu}_{w,t}(a)$ for all a can be communication-heavy; one can compress by sending only $\max_a \widehat{\mu}_{w,t}(a)$ and $\widehat{\mu}_{w,t}(b)$ for the recommended b, at the cost of limiting the parent's ability to evaluate counterfactual recommendations. Our theory is agnostic to this engineering choice; it only uses that $\widehat{\tau}_{b,t}(w)$ is a valid upper bound with high probability.

The primal bandit step: learning on shadow-price-shifted rewards. Fix a node v. In each round t, v chooses its own action $A_t^v \in A$ and a recommendation $B_t^w \in A$ for each child $w \in C(v)$. We view the pair

$$z \ = \ (a, b_{C(v)}) \in \mathcal{Z}_v \ := \ A \times A^{C(v)}$$

as a "meta-action" (an arm) available to v. Given $z = (a, b_{C(v)})$ and a current shadow price $\lambda_{v,t}$, the priced recursion motivates the following local score:

$$\widehat{S}_{v,t}(z) \ := \ \widehat{\theta}_{v,t}(a, b_{C(v)}) \ + \sum_{w\in C(v)} \widehat{\mu}_{w,t}(b_w) \ - \ (1+\lambda_{v,t}) \sum_{w\in C(v)} \widehat{\tau}_{b_w,t}(w), \tag{4}$$

where $\widehat{\theta}_{v,t}(a, b_{C(v)})$ is v's bandit estimate of $\theta_v(a, b_{C(v)})$ from its own realized rewards X_t^v. In words, v trades off (i) its immediate expected reward, (ii) the continuation value it expects to unlock in each child subtree by inducing b_w, and (iii) the shadow-price-inflated transfer needed to implement those inductions.

Budgeted-MAIL allows any standard adversarial/stochastic bandit subroutine to select $z_t \in \mathcal{Z}_v$ based on the feedback available to v. A canonical choice is an Exp3-style routine over $\mathcal{Z}_v$ using realized payoffs

$$\widetilde{X}_t^v \ := \ X_t^v(A_t^v, A_t^{C(v)}) \ - \ (1+\lambda_{v,t}) \sum_{w\in C(v)} \mathbf{1}\{A_t^w = B_t^w\}\tau_t(w),$$

augmented by the children's reported continuation summaries. A simpler (and often sharper under i.i.d. rewards) alternative is UCB on $\mathcal{Z}_v$ using that v

observes X_t^v and the realized child action profile $A_t^{C(v)}$. The cost of generality is computational: $|\mathcal{Z}_v| = K^{1+|C(v)|}$ grows exponentially in the branching factor. This is not merely an artifact of our analysis: without structure on $\theta_v(\cdot)$, v faces a genuine combinatorial bandit problem. In applications with large B, one typically imposes separability or low-order interactions (e.g., θ_v additive across children), in which case (4) decomposes and the primal step becomes tractable via per-child indices.

Hard-budget implementability: feasible recommendation sets and truncation. To guarantee $\sum_{w \in C(v)} \tau_t(w) \le \rho_v$ *for every* t, we couple the primal step with a feasibility filter based on the conservative gap bounds (3). Specifically, define the *safe* set of recommendation profiles at time t:

$$\mathcal{B}_{v,t}^{\text{safe}} := \Big\{ b_{C(v)} \in A^{C(v)} : \sum_{w \in C(v)} \widehat{\tau}_{b_w,t}(w) \le \rho_v \Big\}.$$

In the hard-budget version, the primal learner is restricted to arms $z = (a, b_{C(v)})$ with $b_{C(v)} \in \mathcal{B}_{v,t}^{\text{safe}}$, and the offered payments are set as

$$\tau_t(w) := \widehat{\tau}_{B_t^w,t}(w).$$

When the confidence event holds, these transfers both (i) induce the intended child actions (up to tie-breaking/no-regret effects) and (ii) satisfy the budget constraint by construction. If $\mathcal{B}_{v,t}^{\text{safe}}$ is empty early on due to large uncertainty (large $\beta_{w,t}$), a conservative fallback is to recommend an arbitrary profile and offer zero transfers, thereby spending nothing while still collecting data on θ_v. This behavior is economically natural: when a principal is unsure how expensive it is to incentivize its agents, it temporarily "waits and learns" rather than risking a budget blow-up.

A more aggressive variant replaces the safe-set restriction with truncation: choose $b_{C(v)}$ using (4) and then project the payment vector $(\widehat{\tau}_{b_w,t}(w))_w$ onto the ℓ_1 ball of radius ρ_v. Truncation preserves feasibility but may break IC for some children, so its analysis requires explicitly accounting for deviation regret; we therefore treat it as a practical heuristic unless additional slack conditions are imposed.

Dual updates: endogenous local prices from observed spending. Each node updates its own shadow price using only its outgoing transfers, reflecting that the constraint is local. Let

$$s_{v,t} := \sum_{w \in C(v)} \mathbf{1}\{A_t^w = B_t^w\} \tau_t(w)$$

denote realized spending (equal to offered spending in the hard-budget safe-set version, since all transfers are paid when inducement succeeds). Budgeted-

MAIL performs the online projected subgradient update

$$\lambda_{v,t+1} \;=\; \Big[\lambda_{v,t} + \eta_v\,(s_{v,t} - \rho_v)\Big]_+, \tag{5}$$

with step size $\eta_v > 0$. When v tends to hit its cap, $\lambda_{v,t}$ rises and (4) increasingly discourages expensive inducements; when v persistently underspends, $\lambda_{v,t}$ drifts down toward zero and the node behaves approximately unconstrained. From a policy perspective, λ_v can be interpreted as a revealed "marginal value of public funds" internal to the organization: high shadow prices identify which managerial layers are effectively cash-constrained and therefore where relaxing ρ_v would yield the largest welfare gains.

Decentralization and message passing. Budgeted-MAIL is decentralized in the strong sense that each node v needs only: (i) its own realized reward X^v_t, (ii) its children's realized actions (observable by assumption), (iii) its own outgoing payments, and (iv) continuation summaries produced by its children (e.g., $\widehat{\mu}_{w,t}(\cdot)$ and $\beta_{w,t}$). No node needs to know the global tree, the rewards of distant nodes, or the budgets of other principals. Computation is local: v's primal learner ranges over $\mathcal{Z}_v$, and its dual update (5) uses only $s_{v,t}$.

We emphasize a limitation that is intrinsic to any fully decentralized approach: while action observability makes inducement verifiable, continuation-value communication is not itself contractible in our model. Our theoretical algorithm therefore implicitly treats these messages as part of the mechanism implementation (cooperative computation), not as strategic reports. In settings where nodes may misreport such summaries, one would need an additional layer of incentive design (e.g., audit schemes or proper scoring rules), which is outside our scope. Subject to this caveat, Budgeted-MAIL provides a clean separation of roles: contracts enforce *behavioral* compliance (actions), while learning and dual updates determine *which* behaviors are worth purchasing under scarce budgets.

6.1 Main Theorems: Regret, Feasibility, and the Role of Depth and Branching

We now state the performance guarantees that justify Budgeted-MAIL as an economically meaningful substitute for an offline planner with full knowledge of θ and full control over downstream behavior. Throughout, we evaluate performance against the constrained welfare benchmark $\mathrm{OPT}(\rho)$ defined by incentive compatibility and the per-round hard caps ρ. Given a realized play path $(A^v_t)_{v,t}$, we define welfare regret

$$\mathrm{Reg}^{\mathrm{SW}}_T(\rho) \;:=\; T\cdot \mathrm{OPT}(\rho) \;-\; \mathbb{E}\left[\sum_{t=1}^{T}\sum_{v\in V}\theta_v\big(A^v_t, A^{C(v)}_t\big)\right],$$

and, for each principal v, the cumulative budget violation

$$\mathrm{Viol}_T(v) \;:=\; \sum_{t=1}^{T}\Big(\sum_{w\in C(v)} \tau_t(w) - \rho_v\Big)_+, \qquad \overline{\mathrm{Viol}}_T(v) := \frac{1}{T}\mathrm{Viol}_T(v).$$

Our theorems come in two flavors that correspond to two notions of feasibility. The *hard-budget* variant enforces $\sum_{w\in C(v)} \tau_t(w) \le \rho_v$ pointwise in t (strong feasibility). The *soft-budget* variant allows overspending but controls $\overline{\mathrm{Viol}}_T(v)$ via a dual update (weak feasibility). This distinction is not merely technical: in applications, a hard cap corresponds to genuine limited liability or cash-on-hand constraints, while a soft cap corresponds to budgeting on an accounting horizon (e.g., quarterly) where temporary overdrafts are possible but penalized.

Theorem 6.1 (Hard-budget feasibility and regret under slack)**.** *Fix $\delta \in (0, 1)$. Consider the hard-budget safe-set version of Budgeted-MAIL in which each principal v restricts recommendations to $\mathcal{B}^{\mathrm{safe}}_{v,t}$ and sets $\tau_t(w) = \widehat{\tau}_{B^w_t,t}(w)$. Suppose each node's estimation routine yields confidence radii $\beta_{v,t}$ such that, with probability at least $1 - \delta$, all continuation-value estimates satisfy the uniform event*

$$\big|\widehat{\mu}_{v,t}(a) - \mu^\lambda_v(a)\big| \le \beta_{v,t} \qquad \forall v \in V,\ \forall t \le T,\ \forall a \in A.$$

Assume additionally a uniform slack *(strict feasibility) condition: there exists a benchmark policy achieving $\mathrm{OPT}(\rho)$ whose induced minimal payments satisfy, for every principal v and every round, $\sum_{w\in C(v)} \tau^\star(w) \le \rho_v - \gamma_v$ for some $\gamma_v > 0$. Then, with probability at least $1 - \delta$,*

1. *(Strong feasibility) For all v and all $t \le T$, $\sum_{w\in C(v)} \tau_t(w) \le \rho_v$ (hence $\mathrm{Viol}_T(v) = 0$).*

2. *(Inducement validity on the confidence event) Every recommended child action is a best response under the priced continuation, up to the child's own no-regret deviations.*

3. *(Welfare regret) For suitable learning rates and dual stepsizes, the welfare regret satisfies*

$$\mathrm{Reg}^{\mathrm{SW}}_T(\rho) \;\le\; \sum_{v\in V}\Big(c_1 W_v \;+\; \widetilde{O}\big(\mathsf{R}_v(T)\big)\Big),$$

where $\mathsf{R}_v(T)$ is the action-regret rate of v's chosen primal routine on the meta-action set $\mathcal{Z}_v$, and $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $(T, K, |V|, 1/\delta)$.

The economic content of Theorem 6.1 is that strict feasibility (a margin γ_v) converts conservative overestimation of gap payments into a transient cost rather than a permanent distortion. Early on, $\beta_{w,t}$ is large, so $\widehat{\tau}_{b,t}(w)$

can be substantially above $\tau_b^{\star,\lambda}(w)$; without slack, this can exclude welfare-relevant profiles from $\mathcal{B}_{v,t}^{\mathrm{safe}}$, forcing the principal to behave as if it were poorer than it truly is. Slack ensures that, once $\beta_{w,t}$ falls below γ_v (after a node-dependent burn-in W_v), the safe set contains the benchmark's recommended profile and the algorithm can compete with $\mathrm{OPT}(\rho)$ without ever violating hard caps. From a policy perspective, this highlights a practical design principle: if one insists on strict per-period cash constraints, it is valuable to leave explicit "headroom" in budgets to accommodate incentive-estimation uncertainty.

When slack is absent, the appropriate comparison point for hard-feasible learning is necessarily weaker. One can either (i) benchmark against $\mathrm{OPT}(\rho-\gamma)$ for an explicit buffer γ that shrinks with T, or (ii) switch to weak feasibility and let the dual variable absorb temporary overspending. The next theorem formalizes the second path.

Theorem 6.2 (Soft budgets: sublinear welfare regret and vanishing average violation)**.** *Consider a soft-budget version of Budgeted-MAIL that allows arbitrary recommendations and uses the projected dual update* (5) *with step-size $\eta_v \propto 1/\sqrt{T}$. Assume bounded rewards, sub-Gaussian noise, and that each node's primal routine guarantees expected action regret $\mathsf{R}_v(T) = o(T)$ on $\mathcal{Z}_v$ with respect to the priced score it observes. Then there exist constants $c_2, c_3 > 0$ such that, for all T,*

$$\mathrm{Reg}_T^{\mathrm{SW}}(\rho) \;\le\; \sum_{v\in V} \Big(c_2 W_v \;+\; \widetilde{O}\big(\mathsf{R}_v(T)\big) \Big) \;+\; \widetilde{O}\left(\sum_{v\in V} \rho_v \sqrt{T} \right),$$

and, simultaneously for each principal v,

$$\mathbb{E}\big[\mathrm{Viol}_T(v)\big] \;\le\; c_3\sqrt{T}, \qquad so \qquad \mathbb{E}\big[\overline{\mathrm{Viol}}_T(v)\big] = O\left(\frac{1}{\sqrt{T}}\right).$$

Theorem 6.2 captures the canonical primal–dual tradeoff: we obtain a clean no-regret guarantee against $\mathrm{OPT}(\rho)$ without requiring strict feasibility, at the price of allowing $O(\sqrt{T})$ cumulative overspending. In organizational terms, the dual variable $\lambda_{v,t}$ plays the role of an internal accounting price: if a layer persistently overspends, the algorithm raises $\lambda_{v,t}$ and makes future inducements more expensive, pushing behavior toward cheaper-to-incentivize actions. The vanishing average violation statement ensures that, over long horizons, each layer's realized spending is asymptotically consistent with its budget cap, even though per-period strict feasibility is not enforced.

Rates and the impact of branching. The abstract form $\widetilde{O}(\mathsf{R}_v(T))$ in both theorems is deliberate: the regret rate is inherited from the chosen bandit routine and from the size/structure of $\mathcal{Z}_v = A \times A^{C(v)}$. If v treats each $z = (a, b_{C(v)})$ as an independent arm and uses an adversarial algorithm such as Exp3, then a typical guarantee is

$$\mathsf{R}_v(T) \;=\; O\Big(\sqrt{T\,|\mathcal{Z}_v|\log|\mathcal{Z}_v|}\Big) \;=\; O\Bigg(\sqrt{T\,K^{1+|C(v)|}\,(1+|C(v)|)\log K}\Bigg),$$

which is exponential in $|C(v)|$. This formalizes an economically intuitive congestion effect of span-of-control: even if budgets were ample, a manager with many agents faces a combinatorial exploration problem unless payoffs decompose. Conversely, if $\theta_v(a, b_{C(v)})$ is additive across children or has low-order interactions, then the priced score decomposes and the effective regret can scale only polynomially in $|C(v)|$ (e.g., via per-child indices), restoring tractability. Our welfare theorems remain valid under either regime; what changes is the concrete form of $\mathsf{R}_v(T)$.

Rates and the impact of depth. Depth affects learning through two channels. First, deeper trees increase the number of active learners, so bounds that sum over $v \in V$ worsen mechanically with $|V|$, which in a B-ary tree scales like $O(B^D)$. Second, continuation values μ_v^λ are defined recursively, so estimation errors and burn-in periods propagate upward: a principal cannot reliably evaluate expensive recommendations for a child whose subtree has not yet learned its own continuation values. This is why the node-dependent time-to-learn parameters W_v appear additively. Economically, this is a dynamic version of the standard multi-layer contracting friction: upstream decisions are only as good as the downstream agents' ability to predict and implement the consequences of those decisions.

Strong versus weak feasibility: when do we need which? If limited liability is literal (no overdrafts, no intertemporal smoothing), strong feasibility is non-negotiable, and Theorem 6.1 shows that we can still achieve no-regret behavior provided there is slack. If, instead, budgets represent accounting rules or internal targets that can be averaged over time, Theorem 6.2 suggests that weak feasibility is sufficient and may be strictly more powerful, because it avoids excluding high-welfare recommendations that are only *temporarily* estimated to be expensive. In either case, the theorems clarify the central tradeoff illuminated by our model: scarce budgets are not merely a static constraint on payments, but an endogenous force shaping which parts of the delegation tree can be profitably explored, which behaviors can be reliably purchased, and how quickly shadow prices converge to their economically meaningful levels.

6.2 Impossibility and Lower Bounds: Threshold Budgets, Linear Gaps, and Propagation

The regret guarantees in Section 6.1 are intentionally stated relative to the *budget-feasible* benchmark $\mathrm{OPT}(\rho)$. This is not only an analytical convenience: when budgets are hard per-period caps, there are environments in which the *unconstrained* welfare-optimal behavior cannot be implemented by *any* contract scheme, even with full information and perfect optimization. In those cases, learning is not the bottleneck; feasibility is. We therefore separate two distinct questions: (i) can the welfare-optimal recommendations be purchased at all under ρ? and (ii) conditional on feasibility, can a decentralized learner approach $\mathrm{OPT}(\rho)$?

Budget thresholds for implementability. Fix an instance (i.e., a collection of mean reward functions θ_v) and consider the welfare-optimal (unconstrained) stationary profile that would be chosen by a planner who can recommend actions and pay transfers without any caps.[1] Denote one such optimal recommendation rule by $b^{\mathrm{uc}} = (b_v^{\mathrm{uc}})_{v\in V}$. Given observable actions and one-step IC, the minimum payment needed to induce b_w^{uc} at an edge $P(w) \to w$ is the agent's (continuation) gap between its best action and the recommended one. In the depth-2 case this is exactly $\tau^{\star}_{b_w^{\mathrm{uc}}}(w) = \max_{a\in A}\theta_w(a) - \theta_w(b_w^{\mathrm{uc}})$; in general depth, the same logic holds with continuation values, yielding the recursively defined $\tau^{\star,\lambda=0}_{b_w^{\mathrm{uc}}}(w)$.

This leads to a simple necessary condition for implementability under hard budgets:

$$S_v^{\mathrm{uc}} \;:=\; \sum_{w\in C(v)} \tau^{\star,0}_{b_w^{\mathrm{uc}}}(w) \;\le\; \rho_v \qquad \forall v \in V. \tag{6}$$

When (6) fails at some principal v, there is no sequence of contracts—adaptive, history-dependent, or randomized—that can induce the unconstrained recommendation $b_{C(v)}^{\mathrm{uc}}$ in every round while respecting the per-round cap, because (by definition of $\tau^{\star,0}$) every IC contract inducing b_w^{uc} must pay at least $\tau^{\star,0}_{b_w^{\mathrm{uc}}}(w)$ on that edge, and payments cannot be shifted across time.

We emphasize that (6) is not merely a sufficient condition that might be loosened with more sophisticated mechanisms. Under our primitives—observable actions, limited liability $\tau \ge 0$, and per-round caps—the gap-payment lower bound is tight: the only way to implement an action that is privately suboptimal is to pay at least its utility gap in that round. Thus budgets create a hard *implementability frontier* in the space of action profiles.

Linear welfare loss when budgets are too tight. When the unconstrained optimum is infeasible, the welfare comparison to the unconstrained

[1] Equivalently, $\rho_v = \infty$ for all v, or no payment constraints.

planner necessarily exhibits a linear gap in horizon T. The next statement formalizes the idea (and complements Proposition 4) that repeated interaction does not wash out a binding feasibility constraint.

Proposition 6.3 (Infeasibility implies a per-round welfare gap)**.** *There exist depth-2 instances with a principal v and a single leaf child w, action set $A = \{0, 1\}$, and budget $\rho_v < 1$ such that:*

1. *the unconstrained welfare-optimal recommendation plays $b_w^{\mathrm{uc}} = 1$ in every round and requires $\tau_1^\star(w) = 1$;*
2. *under the hard cap $\rho_v < 1$, no contract can induce $A_t^w = 1$ with probability one in any round t;*
3. *consequently, any budget-feasible policy (even with full information) satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{u\in\{v,w\}} \theta_u(A_t^u, A_t^{C(u)})\right] \leq T \cdot \mathrm{OPT}(\infty) - \Delta T$$

for some constant $\Delta \in (0, 1]$ that does not depend on T.

A concrete construction is instructive. Let the child's reward be $\theta_w(0) = 1$ and $\theta_w(1) = 0$, so the child strictly prefers action 0 by a gap of 1. Let the principal's reward be $\theta_v(1) = 1$ and $\theta_v(0) = 0$, so the principal strictly prefers the child to play 1. Then welfare is maximized by inducing 1 (total welfare $= 1$ per round), but the required payment is $\tau_1^\star(w) = 1$. If $\rho_v < 1$, the action 1 is simply not purchasable; any budget-feasible interaction induces 0 (welfare $= 1$ at the child, 0 at the principal) or some mixture that cannot reach the unconstrained welfare. The welfare shortfall is a constant Δ per round, hence linear over time.

The broader lesson is that "learning the right incentives" cannot substitute for "having enough incentives to offer." In a bandit environment, one might hope that exploration could discover states in which inducement is cheap; however, in our model the relevant gaps are properties of the agents' payoffs, and if the desired behavior is always privately dominated by at least Δ, then hard caps force a persistent distortion.

Lower bounds with branching: additive infeasibility across children. Budgets bind through sums, so branching amplifies infeasibility in a mechanically additive way. Consider B leaf children $w_1, \dots, w_B$ with identical gaps: for each child, the welfare-relevant recommended action requires a minimum payment $g > 0$. Then implementing the unconstrained profile at v requires total payment Bg. If $\rho_v < Bg$, at most $\lfloor \rho_v / g \rfloor$ children can be induced each round (in any IC, budget-feasible scheme), and the resulting welfare loss scales proportionally with the number of “unfunded” children. This is a sharp sense in which span-of-control and budgets interact: even when each bilateral misalignment is small, aggregate misalignment can exceed a manager’s cap and force rationing of incentives.

Propagation with depth: constrained continuation values cascade upward. Depth introduces a second amplification channel that is less purely accounting-based and more economic: a binding budget deep in the tree lowers feasible continuation values, which then changes what is worth inducing upstream. Formally, define the *feasible continuation value* of a node v as the optimal expected welfare obtainable in the subtree rooted at v, subject to IC and budgets ρ within that subtree. Denote this value (in per-round terms) by $\mathrm{VAL}_v(\rho)$. By construction, $\mathrm{VAL}_v(\rho)$ satisfies a backward recursion: it is the maximum, over v’s own action and inducible recommendations to children, of $\theta_v(\cdot)$ plus $\sum_{w\in C(v)} \mathrm{VAL}_w(\rho)$, subject to the payments needed to implement those recommendations fitting under ρ_v. A budget reduction at a descendant u reduces $\mathrm{VAL}_u(\rho)$; through the recursion, it weakly reduces $\mathrm{VAL}_v(\rho)$ for every ancestor v of u. In this sense, feasibility constraints “propagate” upward even if the constrained node is several layers away from the root.

This propagation can create large losses when upstream payoffs are complementary in downstream behavior. In a chain (a path graph) of depth D, one can construct instances where each node’s high-payoff action is valuable only if its child takes a particular costly action. If the lowest-level principal lacks budget to induce that costly action, then the child never takes it; the parent then finds its own high-payoff action unattractive; and so on up the chain. The result is not merely a localized loss at the constrained node, but a cascade in which each layer abandons an otherwise welfare-improving action because the downstream condition cannot be satisfied. In the dual language, a tight cap at a lower node corresponds to a high shadow price λ_u, which reduces the shadow-price-shifted continuation utility passed upward; upstream principals behave as if downstream improvements were “taxed,” and may rationally stop paying for them even when their own budgets are slack.

Implications for what our learning guarantees can and cannot promise. These lower-bound phenomena delimit the scope of any algorithmic result: without sufficient budgets, there is no policy that can approximate the unconstrained planner’s welfare, and thus one should not evaluate decentralized learning against $\mathrm{OPT}(\infty)$. Instead, $\mathrm{OPT}(\rho)$ is the correct target because it internalizes the implementability frontier induced by hard caps. Put differently, our primal–dual perspective is not only a method for computing or learning near-optimal behavior; it is also a diagnostic: when the learned shadow prices are persistently high at certain nodes, this is an endogenous certificate that the organization is operating on the boundary of feasibility, where linear welfare gaps relative to unconstrained ideals are unavoidable.

The next section turns this diagnostic into comparative statics: we interpret λ_v as the marginal value of relaxing ρ_v, establish monotonicity properties, and discuss how the budget frontier scales with tree size, offering design guidance for credit limits and incentive budgets in platform and organizational settings.

6.3 Comparative Statics and Interpretation: Shadow Prices, Monotonicity, and Budget Design

Budgets enter our model as *hard* per-round feasibility constraints, and Section 6.2 shows that these constraints can generate persistent (indeed linear-in-T) efficiency losses relative to an unconstrained planner. Here we ask a different, more design-oriented question: holding fixed the underlying payoff environment $(\theta_v)_{v\in V}$, how does achievable welfare vary with the budget vector $\rho = (\rho_v)_{v\in V}$, and how should we interpret the dual variables $(\lambda_v)_{v\in V}$ that arise in the shadow-price characterization?

Shadow prices as marginal value of budget. We start with the economic content of λ_v. In the dualized welfare problem, $\lambda_v \geq 0$ is the multiplier on v's per-round cap $\sum_{w\in C(v)} \tau_t(w) \leq \rho_v$. When strong duality holds (or when we work with the usual convexification via randomized contracts / relaxed average constraints), λ_v admits the standard envelope interpretation: it is the marginal value of relaxing v's cap. Concretely, let $\mathrm{OPT}(\rho)$ denote the optimal *budget-feasible* welfare benchmark over horizon T. Define the per-round value $V(\rho) := \frac{1}{T}\mathrm{OPT}(\rho)$. Under regularity conditions ensuring differentiability of V at ρ, we obtain the sensitivity formula

$$\frac{\partial V(\rho)}{\partial \rho_v} \;=\; \lambda_v^\star(\rho), \tag{7}$$

where $\lambda^\star(\rho)$ is an optimal dual solution. Even without differentiability, $\lambda_v^\star(\rho)$ can be interpreted as a subgradient: it upper-bounds the welfare gain from a small budget increase and is zero whenever the constraint is slack at the optimum. Economically, $\lambda_v^\star$ measures the *shadow return* (in welfare units per dollar of transfer capacity) of incremental budget at node v. This is precisely the statistic we would like a decentralized learning scheme to output if our goal is *budget design* rather than only *policy optimization.*

This interpretation connects directly to Budgeted-MAIL. The dual update

$$\lambda_{v,t+1} = \left[\lambda_{v,t} + \eta(\textstyle\sum_{w\in C(v)} \tau_t(w) - \rho_v)\right]_+$$

is a textbook "price adjustment" rule: when the realized spend at v exceeds the cap, the internal price of budget rises, and future recommendations shift toward cheaper-to-induce child actions. When average violations vanish, complementary slackness implies that persistent positive prices coincide with binding caps in the long run. Thus the learned $\lambda_{v,t}$ is not a mere proof artifact; it is an *operational* measure of scarcity of incentive capacity at v.

Monotonicity in budgets and diminishing returns. We next formalize the basic comparative statics of $V(\rho)$. Because budgets only restrict feasible contracts, increasing ρ can never reduce achievable welfare.

Proposition 6.4 (Monotonicity and concavity in ρ)**.** *Fix the primitives (θ_v, A). Then $V(\rho)$ is coordinate-wise nondecreasing in ρ: if $\rho' \geq \rho$ componentwise, then $V(\rho') \geq V(\rho)$. Moreover, under the convexified formulation (e.g., allowing randomized contracts and considering per-round expected budgets), $V(\rho)$ is concave in ρ, and there exists an optimal dual vector $\lambda^\star(\rho)$ such that $\lambda^\star(\rho) \in \partial V(\rho)$.*

The first claim is immediate: any policy feasible under ρ remains feasible under $\rho' \geq \rho$. The concavity statement captures diminishing returns: once we can already afford the key inducement gaps, additional budget has lower marginal impact. Combining concavity with (7) yields a useful monotonicity of shadow prices:

$$\rho' \geq \rho \quad \Longrightarrow \quad \lambda_v^\star(\rho') \leq \lambda_v^\star(\rho) \ \text{(in the sense of subgradients)}. \tag{8}$$

In words: relaxing a constraint cannot increase its own scarcity price. This is the formal counterpart to the intuition that if a manager receives a higher incentive budget, the "internal tax" on transfers should fall.

Two caveats are worth flagging. First, with hard per-round caps and purely deterministic contracts, $V(\rho)$ need not be globally concave because the feasible inducement set can be combinatorial (a knapsack-like selection of which children to subsidize). Concavity is recovered in the standard way once we allow randomization across affordable recommendation profiles, or when we replace hard caps by average constraints. Second, even when $V(\rho)$ is concave, $\lambda^\star(\rho)$ need not be unique; in practice we view the learned $\lambda_{v,t}$ as converging to a *consistent* scarcity signal rather than a uniquely identified object.

Scaling with tree size: branching and the "span-of-control" effect. How large do budgets need to be as organizations or platforms scale? Our model isolates two mechanical forces.

With branching, budgets bite additively because payments sum across children. In the depth-2 case, Proposition 1 yields the exact affordability condition $\sum_{w \in C(v)} \tau^\star_{b_w}(w) \leq \rho_v$. If children are statistically similar, $\tau^\star_{b_w}(w) \approx g$ for a welfare-relevant recommendation, then a principal with B children needs ρ_v on the order of Bg to implement the same quality of downstream behavior. Holding ρ_v fixed as B grows forces rationing: the principal shifts to cheaper recommendations, induces only a subset of children, or both. In shadow-price terms, the cost inflation factor $(1 + \lambda_v)$ rises endogenously with B, because the cap becomes tight more often. This is a precise sense in which expanding span of control without commensurate incentive budgets can reduce welfare even if each individual relationship is "easy" to manage in isolation.

Scaling with depth: shadow-price propagation and effective upstream tightness. Depth creates a more subtle scaling channel. When budgets bind in the lower layers, continuation values passed upward are reduced, which changes what upper-layer principals find worthwhile to induce. In the dual recursion, this shows up as shadow-price-adjusted minimal inducing payments $\tau^{\star,\lambda}$ and shadow-price-shifted continuation utilities μ_v^λ: downstream scarcity effectively "taxes" upstream improvements. As a consequence, even if a high-level principal has a generous ρ_v, it may optimally spend little because the downstream actions that would complement its own choices are too expensive (in true or shadow costs) to implement further down the tree.

This has a practical implication for diagnosing organizational bottlenecks. A persistently high λ_u at a low-level node u is not merely a local symptom; it predicts a global distortion because it depresses the net continuation payoff of many upstream decisions. In this sense, budgets at lower levels can have outsized welfare impact, a pattern familiar from operations and platform settings where limited "credits" or "coupons" at the edge constrain the effectiveness of higher-level coordination.

Budget design as marginal-value equalization. Because λ_v measures marginal welfare per unit of budget at node v, it provides a principled rule for reallocating incentive capacity across a tree. Suppose a designer can increase total budget by a small amount ε (or reallocate budgets across nodes while holding $\sum_v \rho_v$ fixed). A first-order prescription is to allocate incremental budget to the nodes with the highest shadow prices. In an interior optimum of such a meta-problem, shadow prices would equalize across nodes that receive positive budget increments, mirroring the classic equi-marginal principle:

$$\lambda_v^\star \approx \lambda_{v'}^\star \quad \text{for nodes } v, v' \text{ that are jointly "on the margin" of investment.}$$

Budgeted-MAIL therefore does more than learn near-optimal actions for a fixed ρ: it produces the very statistics needed to *redesign* ρ. For example, a platform deciding how many subsidy credits to allocate to different regions (nodes) can interpret λ_v as the welfare value of increasing the region's credit limit by one unit. Regions with high learned λ_v are those where incentive scarcity most constrains downstream behavior and hence overall welfare.

Policy interpretation: credit limits, risk control, and robustness. Finally, we connect these comparative statics to the motivating "credit policy" interpretation. Hard per-round caps are a stylized form of limited liability and risk control: an organization may be unwilling or unable to expose any manager to arbitrarily large incentive payouts in a single period. Our analysis clarifies the tradeoff. Tight caps reduce worst-case transfer exposure but can generate predictable efficiency losses and distort behavior toward low-gap, low-impact actions. Shadow prices quantify this tradeoff locally and dynamically: if λ_v is near zero, raising ρ_v is not valuable (and may only increase risk); if λ_v is persistently large, the cap is an active bottleneck, and increasing ρ_v (or smoothing payments across time, if feasible) has high expected welfare return.

We close with a limitation that also suggests a design lever. Our strongest welfare and sensitivity conclusions align most cleanly with formulations that allow randomization or average-budget relaxations. In practice, many platforms can approximate such relaxations by smoothing credit usage (e.g., rolling budgets, credit banking, or allowing unused budget to carry over). From our perspective, these are not merely engineering tweaks: they convexify the feasible set, reduce knapsack-type discontinuities, and make shadow prices more stable and interpretable as marginal values.

6.4 Experiments and Simulations (Illustrative)

Our theoretical results characterize what is feasible under hard per-round caps and how shadow prices summarize scarcity. To complement that characterization—and to sanity-check the behavior of a fully decentralized implementation—we find it useful to study small-scale simulations in which the environment is controlled and the relevant objects (welfare benchmarks, minimal inducing payments, and dual optima) can be computed or tightly approximated. The goal of these experiments is not to claim empirical realism, but rather to make three qualitative points visible: (i) welfare exhibits sharp "budget thresholds" that align with inducement feasibility; (ii) learned shadow prices behave like scarcity signals and localize bottlenecks; and (iii) the interaction between learning noise and hard caps can be diagnosed through deviation frequencies and budget-violation statistics.

Toy tree environments. We simulate rooted trees with depths $D \in \{2, 3, 4\}$, branching factors $B \in \{2, 4, 8\}$, and a common action set $A = \{1, \dots, K\}$ with $K \in \{3, 5\}$. Rewards are generated from bounded mean functions $\theta_v \in [0, 1]$ plus i.i.d. sub-Gaussian noise, consistent with our assumptions. To make delegation nontrivial, we construct θ_v to contain (a) an "own-action" term that creates private incentives at each node and (b) an "alignment" or "externality" term that makes a principal care about the actions of its children. A convenient parametric family is

$$\theta_v(a_v, a_{C(v)}) \;=\; \underbrace{\alpha_v(a_v)}_{\text{own term}} \;+\; \underbrace{\frac{\beta_v}{|C(v)|} \sum_{w \in C(v)} \mathbf{1}\{a_w = \pi_{v \to w}(a_v)\}}_{\text{alignment term}} \quad \text{clipped to } [0,1], \tag{9}$$

where $\alpha_v(\cdot) \in [0,1]$ is drawn once at initialization (e.g., i.i.d. from a Beta distribution and then normalized), $\beta_v \in [0,1]$ controls the strength of externalities, and $\pi_{v \to w}$ is a fixed mapping (possibly identity) encoding which child action best complements v's choice. Leaves have $C(w) = \emptyset$, so their $\theta_w(a_w)$ reduces to $\alpha_w(a_w)$. This construction ensures that (i) children face private tradeoffs among their own actions, so inducement can be costly, and (ii) upstream nodes benefit from coordinating downstream actions, so transfers have social value.

Budgets are imposed per internal node, $\rho_v \in [0,1]$, and we study both homogeneous budgets ($\rho_v \equiv \rho$) and bottleneck budgets (e.g., a single low-level node u has $\rho_u \ll \rho$ while others are large). We set horizons T in the range 10^4 to 10^5, which is long enough to observe steady-state dual behavior while still showing transient learning effects.

Algorithms and baselines. Our main algorithmic object is a decentralized Budgeted-MAIL implementation: each principal v maintains (i) a bandit learner over composite decisions $(a_v, b_{C(v)})$ with payoffs shifted by the current shadow price and estimated inducement cost, and (ii) a dual update of the form

$$\lambda_{v,t+1} = \left[\lambda_{v,t} + \eta\left(\textstyle\sum_{w \in C(v)} \tau_t(w) - \rho_v\right)\right]_+.$$

Transfers $\tau_t(w)$ are set to the estimated minimal inducing payment consistent with the recommended action B_t^w, possibly inflated by a small slack $\epsilon > 0$ to reduce tie-breaking deviations. Because hard caps can be violated by naive dual methods during transients, we also evaluate a conservative variant that enforces feasibility by construction: if the estimated sum of minimal payments for the intended recommendation profile exceeds ρ_v, the principal either (a) switches to a cheaper recommendation profile or (b) drops subsidies to a subset of children until the cap is met.

We compare against three simple baselines meant to isolate what shadow pricing contributes.

1. *Unconstrained delegation:* principals ignore budgets and pay the estimated minimal inducing payments. This is infeasible under hard caps but provides an upper envelope for what one would do absent constraints.

2. *Greedy budget-myopic:* each v chooses the recommendation profile that maximizes its current empirical mean reward minus *raw* transfer cost, without a dual term (equivalently $\lambda_v \equiv 0$).

3. *Fixed-price rule:* each v uses a constant internal price $\bar{\lambda}_v$ (hand-tuned or set by a short calibration phase) and then runs only the primal learner.

All methods use the same child response model (children act as no-regret bandits conditional on contracts), so differences are attributable to how principals select recommendations and manage caps.

Outcome metrics. We track four families of statistics.

1. *Welfare:* realized social welfare $\sum_{t \le T} \sum_v X_t^v$ and its per-round average. For depth-2 trees we also compute a near-exact offline benchmark by enumerating action profiles and applying the affordability condition $\sum_w \tau^\star_{b_w}(w) \le \rho_v$ using the true means.
2. *Budget usage and violations:* $\sum_{t \le T} \sum_{w \in C(v)} \tau_t(w)$, the fraction of rounds in which $\sum_w \tau_t(w) = \rho_v$ (cap binding), and the cumulative violation $\sum_t [\sum_w \tau_t(w) - \rho_v]_+$ when feasibility is not enforced by construction.
3. *Learned shadow prices:* trajectories of $\lambda_{v,t}$ and their time averages; we also report cross-sectional summaries such as $\max_v \lambda_{v,T}$ and how prices concentrate at bottlenecks.
4. *Deviation frequency:* the event rate $\mathbf{1}\{A_t^w \neq B_t^w\}$ by depth and by time, which is a practical proxy for whether inducement payments are sufficiently estimated and sufficiently slackened to overcome learning noise and tie-breaking.

Welfare–budget curves and inducement thresholds. A robust qualitative pattern is that average welfare as a function of a homogeneous budget level ρ exhibits a pronounced "knee." For small ρ, principals cannot afford to induce the actions that generate alignment benefits in (9), and welfare remains close to what would arise from largely uncoordinated play. As ρ passes the typical scale of minimal inducing payments (which depends on the dispersion of $\alpha_w(\cdot)$ at leaves and on downstream shadow costs in deeper trees), welfare increases rapidly and then saturates once the relevant high-impact recommendations become affordable. In depth-2 trees this knee aligns closely with the empirical distribution of $\sum_{w \in C(v)} \tau^\star_{b_w}(w)$ for the welfare-relevant profile b, making the feasibility logic visible in a single plot. In deeper trees the transition is smoother, but the same basic phenomenon appears: once lower-level caps allow downstream coordination to "unlock," upstream spending becomes more productive.

Shadow prices as bottleneck detectors. When budgets are homogeneous, learned prices $\lambda_{v,t}$ tend to settle into a band that decreases with ρ.

When we introduce a single bottleneck node u with a low cap, the shadow-price profile becomes highly localized: λ_u grows and remains persistently elevated, while nodes away from the bottleneck often converge to near-zero prices even if they have many children. This localization matches the economic logic that the marginal value of budget is highest where binding constraints actually block inducement. Moreover, in depth-D trees, high λ values at lower depths predict reductions in upstream willingness to pay: we observe principals above the bottleneck shifting their recommended profiles toward actions that are less complementary but cheaper to implement given downstream scarcity. This is precisely the kind of endogenous "coordination retreat" that the dual recursion is meant to encode.

Deviation events and conservative feasibility. Deviation frequencies are informative because hard caps interact with estimation error: a principal might underpay early on, inducing occasional deviations even when the intended profile is nominally affordable. In our simulations, adding a small slack ϵ to estimated minimal payments reduces deviations markedly, at the cost of consuming more budget and thus increasing the frequency with which caps bind. The conservative feasibility variant (which never exceeds ρ_v by construction) typically eliminates budget violations entirely and keeps deviations low after a short burn-in, but it can pay a welfare price when ρ_v is near the knee: by avoiding occasional over-spend, it sometimes forgoes high-value profiles that would have been implementable with slightly more aggressive exploration. This tradeoff is useful in practice because it separates two distinct failure modes—"we recommended the wrong actions" versus "we could not afford the right ones"—and suggests that tuning ϵ and the dual step size η is partly a risk-management choice.

Sensitivity to nonstationarity. Finally, we probe robustness by introducing a single change point at time $T/2$, resampling a subset of the $\alpha_v(\cdot)$ terms or changing the alignment strength β_v. Predictably, welfare drops immediately after the change for all methods, but the recovery dynamics differ. Methods with adaptive dual prices reallocate spending more quickly when the environment shifts which children are "worth subsidizing," whereas fixed-price rules can remain stuck in a mispriced regime (either overspending on low-value gaps or underspending when new profitable coordination opportunities emerge). The same experiment highlights a limitation: if the environment becomes systematically more expensive to induce (larger gaps), then with hard caps the post-change steady state may have permanently lower welfare, not because learning fails but because feasibility tightens. This is an important diagnostic role for $\lambda_{v,t}$: persistent upward drift in λ following a change point is an operational signal that the organization has moved into a more budget-scarce regime.

Taken together, these simulations make the main objects of our analysis tangible: budgets generate sharp feasibility frontiers, shadow prices track which caps are truly constraining, and deviation/budget statistics provide a practical lens on whether welfare losses are incentive-limited or information-limited.

7 Discussion and Extensions

Our focus on *hard per-round* caps $\sum_{w \in C(v)} \tau_t(w) \leq \rho_v$ isolates the canonical scarcity tension: even when downstream coordination is valuable, it may simply be unaffordable in a given round. This modeling choice also makes the dual interpretation clean—λ_v acts as a local shadow cost of spending that feeds into recursive continuation values. At the same time, organizational budgeting in practice is often intertemporal, uncertain, and mediated through richer contracting primitives than a single recommended action with a compliance bonus. We briefly discuss several extensions that we view as both natural and technically consequential.

Global (horizon) budgets and pacing. A common alternative is a *horizon budget* at each principal v,

$$\sum_{t=1}^{T} \sum_{w \in C(v)} \tau_t(w) \ \leq \ \mathcal{B}_v, \tag{10}$$

possibly with $\mathcal{B}_v = T\rho_v$ for comparability. Compared to per-round caps, (10) admits intertemporal substitution: a principal can overspend early to accelerate learning or to seize a transient high-value coordination opportunity, then underspend later to satisfy feasibility ex post. The dual now attaches a multiplier to a *single* coupling constraint across time, so the natural shadow price λ_v becomes an intertemporal "pacing" signal rather than an instantaneous feasibility signal. Algorithmically, this pushes us from "always feasible" conservative rules toward online resource allocation methods: one can interpret $\lambda_{v,t}$ as the Lagrange multiplier of a remaining-budget constraint and update it via online mirror descent or via a virtual-queue recursion that tracks remaining slack. A key conceptual change is that deviation control and budget control decouple: we can pay inducing amounts whenever desired, but must schedule *when* we do so. This is closely aligned with how many organizations manage annual budgets (front-loading versus end-of-year spending) and suggests that shadow prices can be used operationally as pacing rates.

Budget replenishment and stochastic budgets. Budgets are often uncertain or replenished stochastically (e.g., monthly replenishment, revenue-linked spending, or contingent funding). A stylized model replaces ρ_v by a random process $\rho_{v,t}$ observed at time t, retaining the per-round feasibility constraint $\sum_w \tau_t(w) \le \rho_{v,t}$. If $\rho_{v,t}$ is exogenous and bounded, then the primal decision becomes a bandit problem with time-varying action feasibility, and the dual signal $\lambda_{v,t}$ must respond not only to overspending but also to realized scarcity shocks. If $\rho_{v,t}$ is *endogenous* (e.g., replenishment depends on past performance), the incentives become more subtle: spending can affect future capacity, which in turn changes optimal exploration. A practical approach is to treat the budget process as a constrained stochastic control problem and to use drift-plus-penalty style updates (virtual queues) to guarantee stability of long-run violations while preserving no-regret learning on the primal side. The open theoretical question is whether one can obtain high-probability welfare regret bounds that scale gracefully with the variability of $\rho_{v,t}$, especially in deeper trees where downstream volatility propagates upward through continuation utilities.

Multi-parent delegation: DAGs instead of trees. Many real delegation graphs are not trees: a downstream unit may be influenced by multiple upstream stakeholders. A minimal extension replaces the tree by a DAG in which an agent w has a set of parents $\mathrm{Pa}(w)$, each offering a contract $(B_t^{w,p}, \tau_t^{w,p})$ for $p \in \mathrm{Pa}(w)$. If transfers are additive, w's one-step IC condition becomes

$$A_t^w \in \arg\max_{a \in A} \Big\{ \mu_w^{\mathrm{cont}}(a) + \sum_{p \in \mathrm{Pa}(w)} \mathbf{1}\{a = B_t^{w,p}\}\, \tau_t^{w,p} \Big\}.$$

This seemingly small change creates two conceptual difficulties. First, parents can *free ride* on each other: if several principals value the same downstream action, each has an incentive to underpay and hope others cover the gap, which can lead to coordination failure even when aggregate budgets are ample. Second, the dual decomposition used in the tree case no longer cleanly separates across edges because an agent's best response couples multiple contracts. One possible route is to impose a coordination protocol among parents (e.g., a single "lead" principal sets the recommendation and others post action-contingent subsidies), effectively turning the DAG into a tree plus side payments. Another is to move to a mechanism-design viewpoint in which the agent receives a *single* aggregated contract computed from parents' bids, but then incentive compatibility must be enforced against strategic principals as well. Understanding when shadow-price signals remain local (as bottleneck detectors) in DAGs is an open and practically relevant question.

Partial observability and moral hazard. Our baseline assumes along-edge action observability, so transfers can be conditioned directly on compliance. In many settings, the principal observes only a noisy proxy (output) or an endogenous signal that depends on both actions and shocks. If v cannot verify A_t^w, then $\tau_t(w)$ must be conditioned on an observable signal $Y_t^{v \leftarrow w}$, and the minimal inducing payment becomes a function of likelihood ratios rather than a deterministic gap. This pushes the model toward repeated moral hazard with learning: the principal must simultaneously learn the reward externalities and the signal structure well enough to design incentives, while facing limited liability and budget caps. A central limitation here is that "pay the gap" is no longer well-defined without observability; the cheapest implementable incentive may require randomized payments or score-based schemes, and hard budgets can bind precisely when noise is large (since stronger incentives require larger expected transfers). Deriving a recursion for $\tau^{\star,\lambda}$ under partial observability, together with regret guarantees, would substantially broaden applicability but likely requires new identification and concentration arguments.

Menus, lotteries, and convexification. We restricted attention to simple take-it-or-leave-it recommendations with a compliance bonus. A richer contracting language allows menus over actions, i.e., a function $\tau(\cdot)$ where the agent is paid $\tau(a)$ if it chooses action a. Menus can be useful even with observable actions because they allow the principal to economize on payments by targeting *multiple* near-optimal actions and by mitigating tie-breaking and learning noise (an issue that becomes acute under hard caps). More importantly, menus and lotteries can convexify the feasible set of induced action distributions, which is often what is needed to justify strong duality and tight primal–dual characterizations. From a learning standpoint, however, menus enlarge the action space faced by the principal (it must learn over a richer policy class), and under budget caps the menu design problem resembles a knapsack over incentive intensities. A promising intermediate step is to allow randomized recommendations with a fixed "gap-plus-slack" payment rule, which can deliver convexification benefits while keeping the induced behavior interpretable.

Strategic agents beyond no-regret responses. Our analysis treats each node as an agent that runs a no-regret bandit algorithm conditional on received contracts, which is a disciplined way to model bounded rationality and limited information. Yet in many applications agents are forward-looking and can anticipate the principal's learning and dual updates. Under hard budgets, this creates a new strategic lever: by deviating early, an agent might increase the principal's estimated inducement cost, thereby raising λ or shifting future recommendations, potentially securing higher future transfers. Capturing such "manipulation of the shadow price" requires equilibrium analysis with strategic learning on both sides (a repeated game with endogenous information). The resulting impossibility and robustness questions are largely open: which contract forms are manipulation-proof, and what welfare guarantees survive without a no-regret assumption?

Computational and statistical bottlenecks in deep hierarchies. Even with simple contracts, the backward recursion defining $\mu_v^{\star,\lambda}$ and $\tau^{\star,\lambda}$ can be computationally burdensome when K, B, or depth D is large, because each principal effectively faces a combinatorial choice over $(a_v, b_{C(v)})$. Our decentralized learning perspective mitigates this by avoiding explicit enumeration, but the regret bounds still typically degrade with the size of the composite action space and with error propagation down the tree. This suggests two complementary research directions: (i) impose structure on θ_v (e.g., linearity, separability, low-rank interactions) to obtain dimension-free learning rates; and (ii) study message-passing or factor-graph methods that exploit local dependence to approximate best responses and shadow-price updates without full enumeration.

Interpretation and design implications. From a policy or management perspective, the shadow prices λ_v provide a language for diagnosing whether poor performance is *information-limited* (insufficient learning/exploration) or *resource-limited* (budgets too tight to implement the desired coordination). Extensions such as global budgets and stochastic replenishment strengthen this interpretation: λ_v becomes an internal "cost of funds" that can guide pacing and contingency planning. At the same time, the impossibility gap reminds us that no amount of learning can overcome binding feasibility constraints when key downstream actions are simply too costly to induce. Clarifying the boundary between what can be recovered by better algorithms (menus, pacing, structure) and what requires real resource relaxation remains, in our view, the central open question raised by budgeted delegation.