Beyond One Shot: Contracting for Sequential Trials

Yanwei Sun
Imperial Business School

with **Jiahua Wu** (Imperial) and **Jibang Wu** (NYU Shanghai)

IMPERIAL
BUSINESS SCHOOL

Motivation

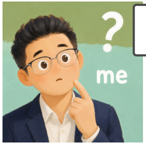
Yanwei, can you prove the conjecture before our next meeting?

Jiahua

Motivation

Can you prove the conjecture before our next meeting?

Jiahua



Me?

Yanwei

Motivation

Can you prove the conjecture before our next meeting?

Jiahua

Now I can do everything!

super Yanwei

Motivation

Codex
try

Motivation

if result not good

Codex try → Codex again

A Simple "Try Again" Can Elicit Multi-turn LLM Reasoning

Licheng Liu[1*], Zihan Wang[2*], Linjie Li[3], Chenwei Xu[2], Yiping Lu[2], Han Liu[2], Avirup Sil[4], Manling Li[2]

[1]Imperial College London [2]Northwestern University [3]University of Washington [4]IBM Research AI

unary-feedback.github.io

LG] 22 Aug 2025

Multi-turn problem solving is critical yet challenging for Large Reasoning Models (LRMs) to reflect on their reasoning and revise from feedback. Existing Reinforcement Learning with Verifiable Reward (RLVR) methods train large reasoning models on a single-turn paradigm. However, we observe that models trained with existing RL paradigms often **fail to explore alternative reasoning paths across multiple turns** and lack the capacity for self-reflection, resulting in repetitive and unadapted responses to contextual feedback. We ask: Can LRMs learn to reflect their answers in a multi-turn context? In this work, we find that training models with multi-turn RL using only unary feedback (for example, "Let's try again") after wrong answers can improve both

Motivation

if result not good

Codex
try

Codex
again

Motivation

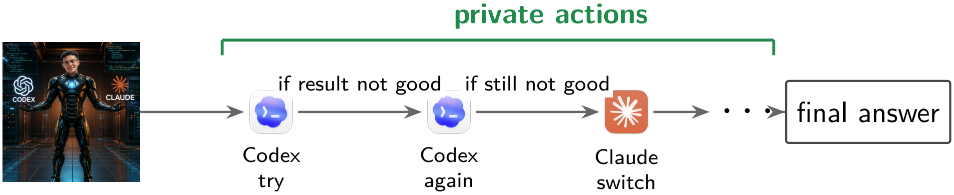
if result not good → if still not good

Codex try → Codex again → Claude switch

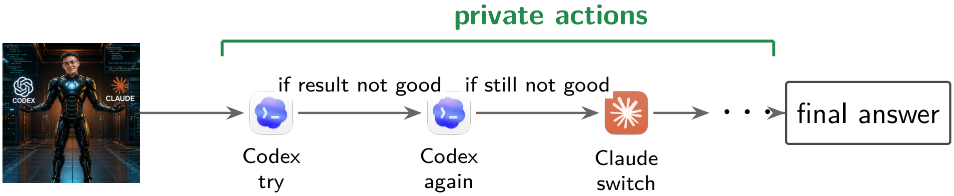
Motivation

if result not good → Codex try → if still not good → Codex again → Claude switch → ⋯ → final answer

Motivation


CODEX
CLAUDE
private actions
if result not good
if still not good
Codex
try
Codex
again
Claude
switch
final answer

Motivation



- **costly trials**: token costs.

Motivation

private actions

if result not good — if still not good

Codex try → Codex again → Claude switch → · · · → final answer

- **costly trials**: token costs.
- **resampling**: try again.

Motivation

private actions

if result not good — if still not good

Codex try → Codex again → Claude switch → · · · → final answer

- **costly trials**: token costs.
- **resampling**: try again.
- **limited trials**: each run may take half an hour, so I cannot try indefinitely before the next meeting.

Hidden Sequential Trial Is Not Just AI

Innovation funding

Resource exploration

Startup / IP deals

private path: try • retry • switch • stop $\Longrightarrow$ **contractible:** one selected “best” outcome

Image sources: Unsplash; Wikimedia Commons; Wikimedia Commons.

Model

Model

principal

agent

Model

principal

contract $w(y)$

agent

Model

principal

contract $w(y)$

agent

agent side

action $i \in \mathcal{A}$: cost c_i, IID draw $Y \sim p_i$

Model

principal

contract $w(y)$

agent

agent side

action $i \in \mathcal{A}$: cost c_i, IID draw $Y \sim p_i$

private policy: which action to use; when to stop

- **resampling allowed (with cost)**
- **total trial cap K**

Model

principal

contract $w(y)$

agent

agent side

action $i \in \mathcal{A}$: cost c_i, IID draw $Y \sim p_i$

private policy: which action to use; when to stop

- **resampling allowed (with cost)**
- **total trial cap K**

report one outcome y

value $v(y)$
to the principal

Model

principal

contract $w(y)$

agent

agent side

action $i \in \mathcal{A}$: cost c_i, IID draw $Y \sim p_i$

private policy: which action to use; when to stop

- **resampling allowed (with cost)**
- **total trial cap K**

report one outcome y

value $v(y)$
to the principal

principal: value $v(y)$ − payment $w(y)$

agent: payment $w(y)$ − trial costs

Total Trial Cap K

K is a total trial budget, not a separate budget for each action.

- patience, deadlines, or resource budgets limit search

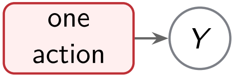
Total Trial Cap K

K is a total trial budget, not a separate budget for each action.

- patience, deadlines, or resource budgets limit search

Example:

$$K = 1$$



standard one-shot moral hazard

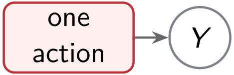
Total Trial Cap K

K is a total trial budget, not a separate budget for each action.

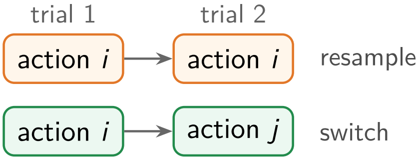
- patience, deadlines, or resource budgets limit search

Example:

$K = 1$



$K = 2$



standard one-shot moral hazard

Related Works (Condensed)

- **Moral hazard and dynamic contracts**
  - Holmström (1979); Grossman–Hart (1983); Spear–Srivastava (1987); Sannikov (2008)...
- **Delegated search and** 

  - Weitzman (1979); Armstrong–Vickers (2010); Lewis (2012); Ulbricht (2016)...
- **Contract design for Pandora-style search**
  - Ezra–Feldman–Schlesinger (2026): NP-hardness
  - Hoefer–Schecker–Schewior (2025): final box/action is contractible
  - Durandard–Vaidya–Xu (2025): min-max optimal in the sense of Carroll (2015)
  - **this work: resampling + trial cap K**

Agent's Best Response

Agent's Best Response

Proposition (Agent's Best Response)

Fix any contract w. The payoff-relevant state is (k, b), where k is # of trials left and b is the best payment already found.

Agent's Best Response

Proposition (Agent's Best Response)

Fix any contract w. The payoff-relevant state is (k, b), where k is # of trials left and b is the best payment already found.

Trial-cap dependent index: index value $z_k^w(i)$ of action i .

Agent's Best Response

Proposition (Agent's Best Response)

Fix any contract w. The payoff-relevant state is (k, b), where k is # of trials left and b is the best payment already found.

Trial-cap dependent index: index value $z_k^w(i)$ of action i .

- stop if $b > \bar{z}_k^w := \max_i z_k^w(i)$;

Agent's Best Response

Proposition (Agent's Best Response)

Fix any contract w. The payoff-relevant state is (k, b), where k is $\#$ of trials left and b is the best payment already found.

Trial-cap dependent index: index value $z_k^w(i)$ of action i .

- stop if $b > \bar{z}_k^w := \max_i z_k^w(i)$;
- if $b < \bar{z}_k^w$, continue with $i \in \arg\max_i \Psi_k^w(b, i)$, which is the *marginal sampling value*;

Agent's Best Response

Proposition (Agent's Best Response)

Fix any contract w. The payoff-relevant state is (k, b), where k is $\#$ of trials left and b is the best payment already found.

Trial-cap dependent index: index value $z_k^w(i)$ of action i .

- stop if $b > \bar{z}_k^w := \max_i z_k^w(i)$;
- if $b < \bar{z}_k^w$, continue with $i \in \arg\max_i \Psi_k^w(b, i)$, which is the *marginal sampling value*;
- at $b = \bar{z}_k^w$, the agent is indifferent; tie-breaking selects.

Comparing with Weitzman (1979)

Weitzman: no resampling and no trial cap K

Comparing with Weitzman (1979)

Weitzman: no resampling and no trial cap K

Weitzman index

$$\mathbb{E}_{Y \sim p_i}\left[(w(Y) - \zeta_i)^+\right] = c_i$$

depends only on c_i and p_i

Our setting

$$z_k^w(i)$$

also depends on trials left k

Comparing with Weitzman (1979)

Weitzman: no resampling and no trial cap K

Weitzman index

$$\mathbb{E}_{Y\sim p_i}\left[(w(Y)-\zeta_i)^+\right] = c_i$$

depends only on c_i and p_i

Our setting

$$z_k^w(i)$$

also depends on trials left k

More remaining trials raise the index value

$$z_{k+1}^w(i) \geq z_k^w(i) \qquad \forall i.$$

Comparing with Weitzman (1979)

Weitzman: no resampling and no trial cap K

Weitzman index

$$\mathbb{E}_{Y\sim p_i}\left[(w(Y)-\zeta_i)^+\right]=c_i$$

depends only on c_i and p_i

Our setting

$$z_k^w(i)$$

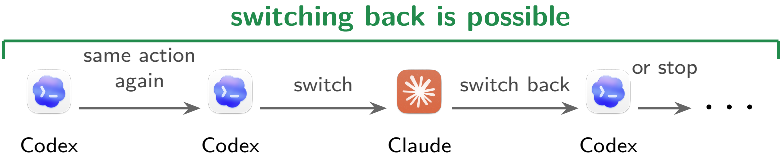
also depends on trials left k

More remaining trials raise the index value

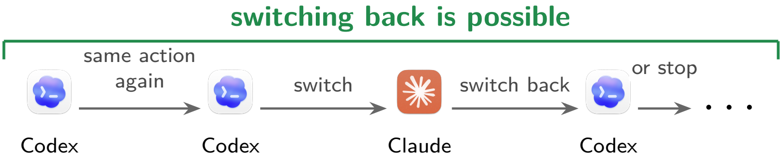
$$z_{k+1}^w(i)\geq z_k^w(i)\qquad\forall i.$$

At $K=\infty$, the state-dependent index value collapses to the Weitzman index.

Agent's Best Response Can Be Complicated



Agent's Best Response Can Be Complicated



Contract design problem: $w(y)$ moves many cutoffs $z_k^w(i)$ across states (k, b). Optimizing w means controlling a hidden adaptive policy.

Hardness of Optimal Contract

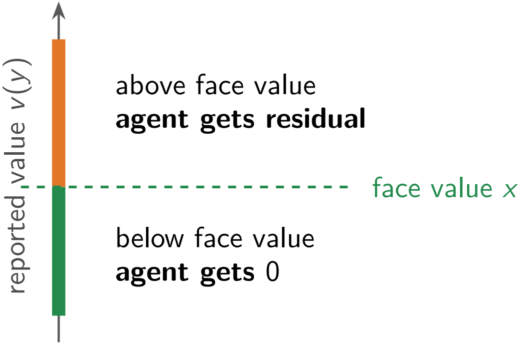
Hardness of Optimal Contract

Theorem (Computational Hardness)

For general instances, finding an optimal contract is **NP-hard**.

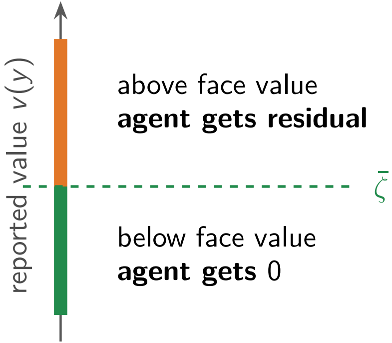
Asymptotically Optimal Contract

Debt Contract



$$w^{\text{debt}}(y;x) = (v(y) - x)^+$$

Calibrating the Face Value

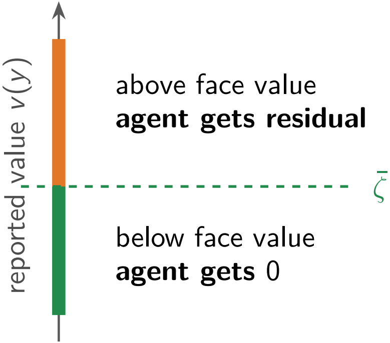


$$w^{\text{debt}}(y;\bar{\zeta}) = (v(y) - \bar{\zeta})^+$$

Weitzman index for action i

$$\mathbb{E}_{Y\sim p_i}[(v(Y) - \zeta_i)^+] = c_i$$

Calibrating the Face Value



$$w^{\text{debt}}(y;\bar{\zeta}) = (v(y) - \bar{\zeta})^+$$

Weitzman index for action i

$$\mathbb{E}_{Y\sim p_i}[(v(Y) - \zeta_i)^+] = c_i$$

Calibrate debt at the largest index

$$\bar{\zeta} = \max_i \zeta_i, \quad \bar{i} \in \arg\max_i \zeta_i$$

Asymptotic Optimal Contract

One-trial hit probability $\bar{q} = \Pr_{Y \sim p_{\bar{i}}}\{v(Y) \geq \bar{\zeta}\}$

Asymptotic Optimal Contract

One-trial hit probability $\bar{q} = \Pr_{Y \sim p_{\bar{i}}}\{v(Y) \geq \bar{\zeta}\}$

Theorem (Asymptotic Optimality)

Under $w^{\text{debt}}(y; \bar{\zeta}) = (v(y) - \bar{\zeta})^{+}$, for every $K \geq 1$,

$$\frac{U_P^{\text{debt}}(K; \bar{\zeta})}{\text{SB}_P(K)} \geq 1 - (1 - \bar{q})^K.$$

Asymptotic Optimal Contract

One-trial hit probability $\bar{q} = \Pr_{Y \sim p_{\bar{i}}}\{v(Y) \geq \bar{\zeta}\}$

Theorem (Asymptotic Optimality)

Under $w^{\text{debt}}(y; \bar{\zeta}) = (v(y) - \bar{\zeta})^+$, for every $K \geq 1$,

$$\frac{U_P^{\text{debt}}(K; \bar{\zeta})}{\text{SB}_P(K)} \geq 1 - (1 - \bar{q})^K.$$

- **At $K = \infty$, the contract is exactly optimal and first best.**

Asymptotic Optimal Contract

One-trial hit probability $\bar{q} = \Pr_{Y \sim p_{\bar{i}}}\{v(Y) \geq \bar{\zeta}\}$

Theorem (Asymptotic Optimality)

Under $w^{\text{debt}}(y;\bar{\zeta}) = (v(y) - \bar{\zeta})^{+}$, for every $K \geq 1$,

$$\frac{U_P^{\text{debt}}(K;\bar{\zeta})}{\mathrm{SB}_P(K)} \geq 1 - (1-\bar{q})^K.$$

- **At $K = \infty$, the contract is exactly optimal and first best.**
- **Agent utility is zero for every trial cap K.**

Proof Sketch: No Switching, Then Hitting

Generic contracts

best action can depend on (k, b); switching is normal.

Proof Sketch: No Switching, Then Hitting

Generic contracts
best action can depend on (k, b); switching is normal.

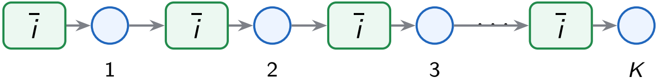
Calibrated $\bar{\zeta}$-debt
continuing $\Rightarrow$ sample a maximal-index action $\bar{i}$ every time.

Proof Sketch: No Switching, Then Hitting

Generic contracts
best action can depend on (k, b); switching is normal.

Calibrated $\bar{\zeta}$-debt
continuing $\Rightarrow$ sample a maximal-index action $\bar{i}$ every time.

no switching: $\bar{i}, \bar{i}, \bar{i}, \ldots$ until the threshold is reached



continue after failures; stop at the first $v(Y_t) \geq \bar{\zeta}$, or when the cap K is exhausted

Proof Sketch: No Switching, Then Hitting

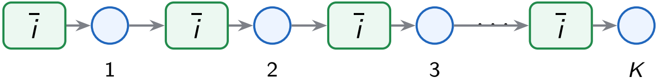
Generic contracts

best action can depend on (k, b); switching is normal.

Calibrated $\bar{\zeta}$-debt

continuing $\Rightarrow$ sample a maximal-index action $\bar{i}$ every time.

no switching: $\bar{i}, \bar{i}, \bar{i}, \ldots$ until the threshold is reached



continue after failures; stop at the first $v(Y_t) \geq \bar{\zeta}$, or when the cap K is exhausted

a hit pays the principal $\bar{\zeta}$

agent keeps $v(y) - \bar{\zeta}$

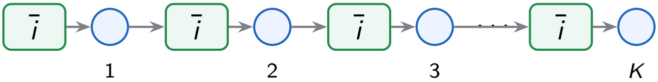
principal keeps $\bar{\zeta}$

Proof Sketch: No Switching, Then Hitting

Generic contracts
best action can depend on (k, b); switching is normal.

Calibrated $\bar{\zeta}$-debt
continuing $\Rightarrow$ sample a maximal-index action $\bar{i}$ every time.

no switching: $\bar{i}, \bar{i}, \bar{i}, \ldots$ until the threshold is reached



continue after failures; stop at the first $v(Y_t) \geq \bar{\zeta}$, or when the cap K is exhausted

a hit pays the principal $\bar{\zeta}$
agent keeps $v(y) - \bar{\zeta}$
principal keeps $\bar{\zeta}$

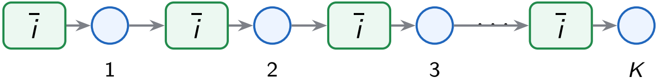
hit probability
$\Pr(\text{hit by } K) = 1 - (1 - \bar{q})^K$

Proof Sketch: No Switching, Then Hitting

Generic contracts
best action can depend on (k, b); switching is normal.

Calibrated $\bar{\zeta}$-debt
continuing $\Rightarrow$ sample a maximal-index action $\bar{i}$ every time.

no switching: $\bar{i}, \bar{i}, \bar{i}, \ldots$ until the threshold is reached



continue after failures; stop at the first $v(Y_t) \geq \bar{\zeta}$, or when the cap K is exhausted

a hit pays the principal $\bar{\zeta}$
agent keeps $v(y) - \bar{\zeta}$
principal keeps $\bar{\zeta}$

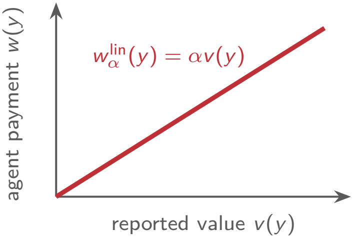
hit probability
$\Pr(\text{hit by } K) = 1-(1-\bar{q})^K$

second best bound
$\mathrm{SB}_P(K) \leq \mathrm{FB}_P(\infty) = \bar{\zeta}$

$$U_P^{\text{debt}}(K; \bar{\zeta}) \;\geq\; \bar{\zeta} \cdot \left[1 - (1 - \bar{q})^K\right] \;\geq\; \left[1 - (1 - \bar{q})^K\right] \cdot \mathrm{SB}_P(K)$$

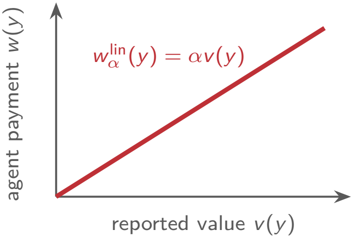
Linear Contracts Can Be Arbitrarily Bad

Linear contract

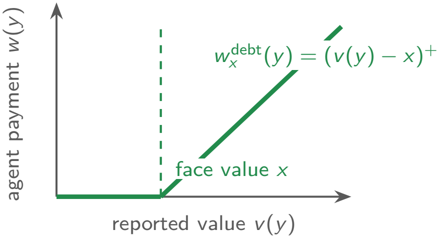


Linear Contracts Can Be Arbitrarily Bad

Linear contract

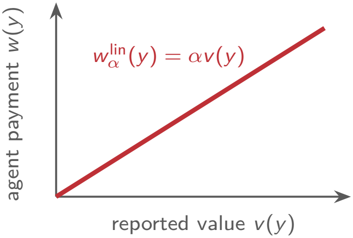


Debt contract

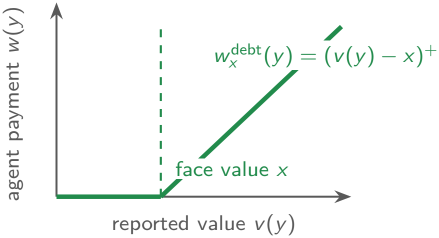


Linear Contracts Can Be Arbitrarily Bad

Linear contract



Debt contract

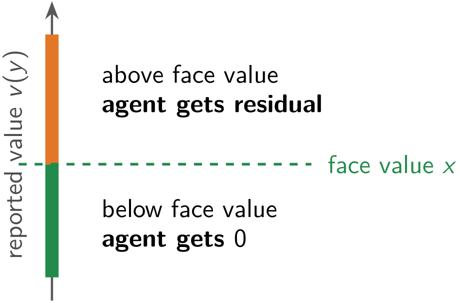


Linear: pays every report; a moderate outcome can stop search early

Debt: pays zero below x; after a hit, the principal keeps face value x.

Optimal Debt Contract

Earlier: the simple calibrated choice $x = \bar{\zeta}$. **Now: optimize x for finite K.**



$$\boxed{w^{\text{debt}}(y; x) = (v(y) - x)^+}$$

Computing the Optimal Debt Contract

Proposition (Finite-K Debt)

For any fixed trial cap K, an optimal debt face value

$$x^*_{\text{debt}}(K) \in \arg\max_x U_P^{\text{debt}}(K; x)$$

can be computed exactly in time **polynomial** in the number of actions and states.

Computing the Optimal Debt Contract

Proposition (Finite-K Debt)

For any fixed trial cap K, an optimal debt face value

$$x^*_{\text{debt}}(K) \in \arg\max_x U_P^{\text{debt}}(K; x)$$

can be computed exactly in time **polynomial** in the number of actions and states.

Key idea: as the face value x moves, the principal's payoff is **piecewise linear**; enumerate the breakpoints.

Optimal Debt Face Value

Optimal Debt Face Value

Optimal Debt Face Value

$$0 \le \bar{\zeta} - x^*_{\text{debt}}(K) \le \bar{\zeta}(1-\bar{q})^K$$

$$0 \le U^{\text{debt}}_{P,*}(K) - U^{\text{debt}}_{P}(K;\bar{\zeta}) \le \bar{\zeta}(1-\bar{q})^K$$

both gaps vanish exponentially in K

The Impact of Trial Cap

Should a Larger Trial Cap K Help?

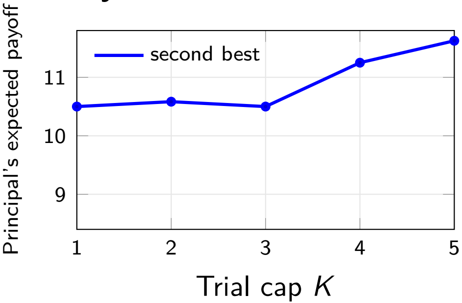
Natural intuition: more trials should help the principal.

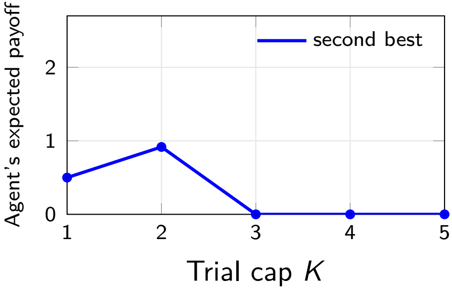
more freedom for hidden search → deeper exploration → better final report

A larger cap K expands the agent's feasible search policy, so one might expect the principal's payoff to rise.

More Trials Can Hurt Both Sides

Lose-lose phenomenon: after re-optimizing an unrestricted contract, a larger cap can **lower** both the principal's and the agent's payoffs **simultaneously**.

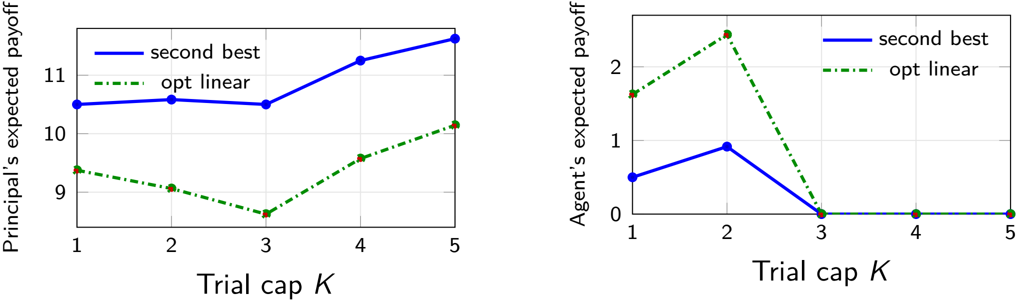




Reason: more private trials create more hidden search paths; a final-outcome contract must discipline all of them.

More Trials Can Hurt Both Sides

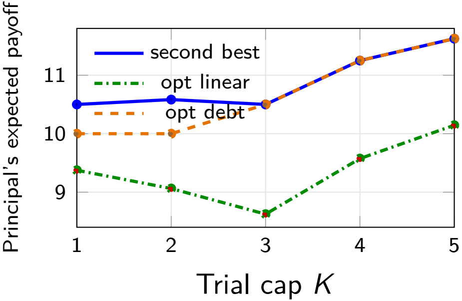
Lose-lose phenomenon: after re-optimizing an unrestricted contract, a larger cap can **lower** both the principal's and the agent's payoffs **simultaneously**.

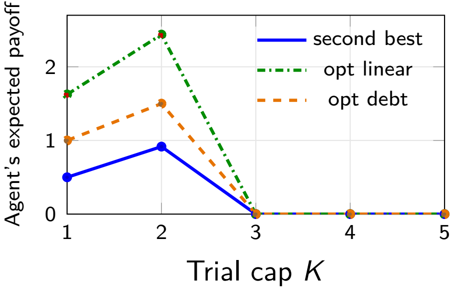


Optimal linear: lose-lose can happen when K increases.

More Trials Can Hurt Both Sides

Lose-lose phenomenon: after re-optimizing an unrestricted contract, a larger cap can **lower** both the principal's and the agent's payoffs **simultaneously**.





Optimal linear: lose-lose can happen when K increases.

Optimal debt: principal payoff $\nearrow$ with the trial cap K.

Extensions

Extensions: The Same Debt Logic Survives

Time-contingent

payments can depend
on stopping time

Extensions: The Same Debt Logic Survives

Time-contingent

payments can depend
on stopping time

Correlated samples

best so far s

$\downarrow$

sample i: $Y \sim P_i(\cdot \mid s)$

Extensions: The Same Debt Logic Survives

Time-contingent

payments can depend on stopping time

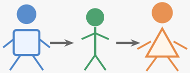
Correlated samples

best so far s

$\downarrow$

sample i: $Y \sim P_i(\cdot \mid s)$

Many agents



offer contracts sequentially

Extensions: The Same Debt Logic Survives

Time-contingent

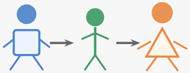
payments can depend
on stopping time

Correlated samples

best so far s

sample i: $Y \sim P_i(\cdot \mid s)$

Many agents



offer contracts
sequentially

Main result survives: calibrated
debt is still asymptotically optimal.

Takeaway

Model: Moral hazard + 

with **resampling** and a **trial cap**

Main Message: Debt contract is all you need

Thank you.