Novelty Assessment Report

Paper: Inverse Linear Bandits via Linear Programs
PDF URL: https://openreview.net/pdf?id=NbHuzXQT8U
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Inverse reinforcement learning (IRL) is a well-established paradigm for circumventing the need for explicit reward. In this paper, we study the problem of estimating the reward function from a single sequence of actions (i.e., a demonstration) of a stochastic linear bandit algorithm. Our main result is a unified approach for inverse linear bandits, based on the idea of formulating a linear program by tightly characterizing the confidence intervals of pulled actions. We show that the estimation error of our algorithms matches the information-theoretic lower bound, up to polynomial factors in D and $\log T$, where D is the dimensionality of the feature space and T is the length of the demonstration. Compared to prior approaches, our approach (i) gives a unified reward estimator that works when the demonstrator employs LinUCB or Phased Elimination, two popular algorithms for stochastic linear bandits, while existing estimator only works for Phased Elimination; (ii) does not require access to hyperparameters or internal states of the demonstrator algorithm as required by prior work; and (iii) works for general action sets, while existing estimator requires assumptions on the density and geometry of the action set. We further demonstrate the practicality of our new approach by validating our new algorithms on synthetic data and demonstrations constructed from real-world datasets, where our estimators significantly outperform existing ones.

Core Task Landscape

This paper addresses: **Estimating Reward Functions from Demonstrations of Stochastic Linear Bandit Algorithms**

A total of **9 papers** were analyzed and organized into a taxonomy with **10 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Inverse Reward Learning from Bandit Demonstrations**
- **Reward-Biased Maximum Likelihood Estimation for Bandits**
- **Transformer-Based In-Context Bandit Learning**
- **Preference-Based Reward Learning**

Complete Taxonomy Tree

- Estimating Reward Functions from Demonstrations of Stochastic Linear Bandit Algorithms Survey Taxonomy
- Inverse Reward Learning from Bandit Demonstrations
  - Inverse Learning for Linear Bandits
  - Linear Program-Based Inverse Estimation ★ (1 papers)
    - [0] Inverse Linear Bandits via Linear Programs (Anon et al., 2026) View paper
  - One-Shot Inverse Learning (1 papers)
    - [1] One shot inverse reinforcement learning for stochastic linear bandits (EK Guha, 2024) View paper
  - Inverse Learning for Multi-Armed Bandits
  - Exploration-Phase Reward Estimation (1 papers)
    - [5] Learning from an Exploring Demonstrator: Optimal Reward Estimation for Bandits (Guo, 2022) View paper
- Reward-Biased Maximum Likelihood Estimation for Bandits
  - RBMLE for Linear Bandits (1 papers)
  - [6] Reward-Biased Maximum Likelihood Estimation for Linear Stochastic Bandits (Yu-Heng Hung, 2021) View paper
  - Neural RBMLE for Contextual Bandits (1 papers)
  - [7] Neural Contextual Bandits via Reward-Biased Maximum Likelihood Estimation (Hung, 2022) View paper
- Transformer-Based In-Context Bandit Learning
  - Multi-Task Decision Transformer Pretraining (1 papers)
  - [3] Pretraining Decision Transformers with Reward Prediction for In-Context Multi-task Structured Bandit Learning (Mukherjee, 2024) View paper
  - Bandit-Based Prompt Tuning for Decision Transformers (1 papers)
  - [4] Prompt Tuning Decision Transformers with Structured and Scalable Bandits (Rietz, 2025) View paper
  - In-Context Learning from Suboptimal Demonstrations (1 papers)
  - [8] In-Context Reinforcement Learning From Suboptimal Historical Data (J Dong, n.d.) View paper
- Preference-Based Reward Learning
  - Optimal Design for RLHF Reward Modeling (1 papers)
  - [2] Optimal design for reward modeling in rlhf (Scheid Antoine, 2024) View paper
  - Causally Robust Preference Learning (1 papers)
  - [9] Causally Robust Preference Learning with Reasons (M Hwang, n.d.) View paper

Narrative

Core task: Estimating reward functions from demonstrations of stochastic linear bandit algorithms. The field addresses how to recover unknown reward structures when observing an agent's sequential decision-making behavior in bandit settings. The taxonomy reveals four main branches that capture distinct methodological perspectives. Inverse Reward Learning from Bandit Demonstrations focuses on classical inverse problems where one infers reward parameters directly from observed arm-pull sequences, often employing optimization or linear programming techniques as seen in Inverse Linear Bandits[0]. Reward-Biased Maximum Likelihood Estimation for Bandits, exemplified by Reward-Biased Maximum Likelihood[6], frames the problem through a probabilistic lens, modeling how demonstration likelihoods depend on underlying rewards. Transformer-Based In-Context Bandit Learning explores modern neural architectures that can adapt to bandit tasks in-context, with works like Pretraining Decision Transformers[3] and Prompt Tuning Transformers[4] leveraging large-scale pretraining. Preference-Based Reward Learning shifts attention to learning from comparative feedback rather than direct demonstrations, addressing robustness concerns as in Causally Robust Preference[9].

Several active themes emerge across these branches. One central question is how to handle suboptimal or exploratory demonstrators: Learning Exploring Demonstrator[5] and In-Context Suboptimal Data[8] both grapple with the reality that observed behavior may not be perfectly rational. Another contrast lies between model-free neural approaches, such as Neural Contextual Bandits[7], and model-based inverse methods that explicitly solve for reward parameters. Inverse Linear Bandits[0] sits squarely within the classical inverse learning branch, using linear programming to estimate reward vectors from bandit trajectories. Its emphasis on tractable optimization distinguishes it from likelihood-based methods like Reward-Biased Maximum Likelihood[6] and from transformer-driven in-context learners such as Pretraining Decision Transformers[3]. By focusing on linear program formulations, the original work aligns with a small cluster of optimization-centric inverse techniques, offering theoretical guarantees that complement the more flexible but less interpretable neural alternatives.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Both subtopics address the problem of inverse reinforcement learning for stochastic linear bandits from limited demonstrations. They share the constraint of working with single demonstration sequences rather than multiple trajectories or extensive exploration data. The key distinction lies in their methodological approach: the original leaf employs linear programming with confidence intervals as its core technique, while the sibling uses a broader 'one-shot inverse learning' framework without specifying the optimization formulation.

Similarities: - Both focus on estimating linear reward functions from demonstrations - Both work with stochastic linear bandit settings - Both are constrained to single demonstration sequences (not multiple trajectories) - Both exclude methods requiring exploration-phase data or multi-task settings

Differences: - Linear Program-Based Inverse Estimation explicitly uses linear programming formulations with confidence intervals as the solution method - One-Shot Inverse Learning describes the problem setting (single demonstration) but does not specify whether it uses LP-based, likelihood-based, or other optimization approaches - The original leaf's scope is more methodologically specific (LP with confidence intervals), while the sibling is more problem-setting specific (one-shot constraint)

Suggested Search Directions: - Clarify whether One-Shot Inverse Learning encompasses LP-based methods or represents alternative approaches (e.g., moment matching, Bayesian methods) - Investigate if these should be merged if One-Shot methods primarily use LP formulations, or kept separate if they represent distinct algorithmic families - Examine whether confidence interval usage is the distinguishing feature or if other statistical guarantees differentiate these approaches

Sibling Subtopics

- **One-Shot Inverse Learning** (leaves: 1, papers: 1)
- Scope: Methods estimating linear rewards from a single demonstration of stochastic linear bandit algorithms.
- Exclude: Excludes methods requiring multiple demonstrations or exploration-phase data; see sibling nodes.

Contributions Analysis

Overall novelty summary. The paper proposes a unified linear program approach to estimate reward functions from single demonstration sequences of stochastic linear bandit algorithms, specifically LinUCB and Phased Elimination. Within the taxonomy, it occupies the 'Linear Program-Based Inverse Estimation' leaf under 'Inverse Learning for Linear Bandits'. This leaf contains only the original paper itself, indicating a relatively sparse research direction. The sibling leaf 'One-Shot Inverse Learning' contains one other paper, suggesting that inverse learning for linear bandits remains an emerging area with limited prior work addressing the specific formulation of LP-based confidence interval characterization.

The taxonomy reveals that inverse reward learning from bandit demonstrations branches into linear bandits and multi-armed bandits without linear structure. The original work sits within the linear bandits branch, which contrasts with multi-armed bandit approaches like 'Exploration-Phase Reward Estimation' that leverage demonstrator exploration behavior. Neighboring branches include 'Reward-Biased Maximum Likelihood Estimation for Bandits', which frames inverse problems probabilistically rather than through optimization, and 'Transformer-Based In-Context Bandit Learning', which uses neural architectures for policy adaptation. The taxonomy's scope notes clarify that the original work excludes likelihood-based methods and multi-task settings, focusing instead on optimization-centric inverse techniques for linear reward recovery.

Among the three contributions analyzed, the literature search examined 15 candidates total. The 'Unified linear program approach' examined 3 candidates with 0 refutations, suggesting novelty in the LP formulation. The 'Information-theoretic lower bound' examined 10 candidates and found 2 refutable matches, indicating that lower bound analysis for inverse reward estimation has some prior coverage. The 'Optimal unified reward estimator' examined 2 candidates with 0 refutations. These statistics reflect a limited search scope (15 papers, not exhaustive), and the presence of 2 refutable pairs for the lower bound contribution suggests that theoretical guarantees in this space have been partially explored, though the unified LP approach itself appears less contested.

Based on the limited search of 15 candidates, the work appears to introduce a novel optimization framework for inverse linear bandits, particularly in its unified treatment of multiple demonstrator algorithms without requiring hyperparameter access. The taxonomy structure confirms this sits in a sparse research direction, with only one sibling paper in the broader inverse linear bandits category. However, the information-theoretic lower bound contribution shows overlap with prior theoretical work, suggesting that while the LP methodology is distinctive, the fundamental limits of inverse reward estimation have received some attention in the examined literature.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Unified linear program approach for inverse linear bandits

Description: The authors propose a unified framework that formulates inverse linear bandits as a linear program by characterizing confidence intervals of actions selected by the demonstrator. This approach works for both LinUCB and Phased Elimination algorithms, does not require access to hyperparameters or internal states, and applies to general action sets without density or geometry assumptions.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness

URL: View paper

Brief Assessment

Combinatorial Linear Constraints[18] addresses combinatorial bandits with linear constraints for forward optimization, not inverse reinforcement learning or reward estimation from demonstrations. The technical focus is entirely different.

---

2. Online Decision Making Via Linear Programming: Resource Allocation, Bandit Feedback, and Inverse Optimization

URL: View paper

Brief Assessment

Online Linear Programming[19] focuses on online resource allocation and sequential decision-making problems, not on inverse reinforcement learning or inverse linear bandits. The candidate's scope does not overlap with the original paper's contribution of formulating inverse linear bandits as a linear program.

---

3. 12 Mathematics of Reinforcement

URL: View paper

Brief Assessment

Mathematics of Reinforcement[20] appears to discuss LP formulations and multi-armed bandits, but the provided context is too fragmentary to establish whether it presents a unified LP framework for inverse linear bandits that would refute the original paper's novelty claims.

Contribution 2: Information-theoretic lower bound for inverse reward estimation

Description: The authors establish a fundamental lower bound showing that estimation error for any inverse learner is lower bounded by a quantity involving the inverse expected feature covariance matrix. This lower bound serves as a baseline for analyzing specific reward estimators and generalizes prior hardness results from multi-armed bandits to linear bandits.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Contextual bandits with linear payoff functions

URL: View paper

Brief Assessment

Contextual Linear Payoff[15] studies forward learning (contextual bandits) where the learner selects actions to minimize regret, not inverse learning where rewards are estimated from demonstrations. The lower bound in [15] concerns regret in the forward problem, not estimation error in inverse reward recovery.

---

2. Optimal regret is achievable with bounded approximate inference error: An enhanced bayesian upper confidence bound framework

URL: View paper

Brief Assessment

Bayesian Upper Confidence[16] focuses on regret bounds for Bayesian bandit algorithms with approximate inference error, not on information-theoretic lower bounds for inverse reward estimation in linear bandits.

---

3. Low-rank bandits via tight two-to-infinity singular subspace recovery

URL: View paper

Brief Assessment

Low-rank Bandits[17] addresses a different problem domain (contextual low-rank bandits with matrix completion) rather than inverse reinforcement learning from demonstrations. The paper's lower bounds concern policy evaluation and best policy identification in low-rank bandits, not inverse reward estimation from demonstrator trajectories.

---

4. Balanced linear contextual bandits

URL: View paper

Brief Assessment

Balanced Linear Contextual[13] focuses on online contextual bandits with balancing methods for bias reduction in reward estimation, not on inverse reward estimation from demonstrations or establishing information-theoretic lower bounds for inverse learners.

---

5. Learning from an Exploring Demonstrator: Optimal Reward Estimation for Bandits

URL: View paper

Prior Art Analysis

Learning Exploring Demonstrator[5] establishes an information-theoretic lower bound for inverse reward estimation in multi-armed bandits that directly addresses the same problem. The candidate paper proves that estimation error for any inverse learner is lower bounded by a quantity involving the inverse expected feature covariance matrix (or equivalently, the number of pulls for each arm in the MAB case). This work predates the original paper and demonstrates that similar fundamental limits were already established in the bandit setting, which the original paper explicitly acknowledges as a generalization.

Evidence

Evidence 1 - **Rationale**: Learning Exploring Demonstrator[5] establishes a fundamental lower bound showing estimation error is inversely proportional to the square root of the number of pulls (which relates to the inverse expected feature covariance in linear bandits). The original paper explicitly states this is a generalization: 'when specializing to the special case of multi-armed bandits, the lower bound in theorem 1 is equivalent to the hardness result in guo et al. (2021)'. This demonstrates prior work exists on the same fundamental limit. - **Original**: we first show that when the demonstrator employs linucb or phased elimination, for any actionain the action set, the estimation error of any reward estimator is lower bounded by a quantity related toaand the inverse expected feature covariance matrix of the demonstrator. this lower bound serves as a... - **Candidate**: theorem 1. (proof in appendix a) for every k-armed bernoulli bandit instance msatisfying maxi∈[k] |μi -1/2|⩽1/4 and for each suboptimal arm𝐼= i∗, the following is true. suppose that the demonstrator employs algorithm a, and let e[na i,t] denote the expected number of times arm i is pulled by awhen...

Evidence 2 - **Rationale**: The original paper explicitly acknowledges that their lower bound is a generalization of Learning Exploring Demonstrator[5]'s result from multi-armed bandits to linear bandits. This direct acknowledgment, combined with the mathematical equivalence shown, demonstrates that the fundamental insight about information-theoretic limits was already established in the candidate paper. - **Original**: intuitively,‖a‖ ¯v -1 quantifies the amount of information the demonstrator has provided for a specific actiona. when specializing to the special case of multi-armed bandits, the lower bound in theorem 1 is equivalent to the hardness result in guo et al. (2021) (which works only for multiarmed bandi... - **Candidate**: note that in addition to applying to any reward estimation procedure, theorem 1 provides a fundamental limit for any choice of demonstrator algorithm in terms of the degree of exploration in that algorithm. its proof utilizes information-theoretic lower bounds on the demonstrator's regret (kaufmann ...

---

6. Impact of representation learning in linear bandits

URL: View paper

Brief Assessment

Representation Learning Impact[11] studies multi-task linear bandits with shared representations, not inverse reward estimation from demonstrations. The lower bounds in [11] concern regret for forward bandit problems, not estimation error for inverse learners recovering reward parameters from action sequences.

---

7. Linear Contextual Bandits with Interference

URL: View paper

Brief Assessment

Linear Bandits Interference[14] focuses on interference effects in contextual bandits where multiple units' actions affect each other's rewards, not on inverse reward estimation or information-theoretic lower bounds for learning from demonstrations.

---

8. Variance-Dependent Regret Bounds for Non-stationary Linear Bandits

URL: View paper

Brief Assessment

Variance-Dependent Regret[10] focuses on non-stationary linear bandits with variance-based regret bounds, not on inverse reward estimation or information-theoretic lower bounds for learning reward parameters from demonstrations.

---

9. Stochastic bandits with linear constraints

URL: View paper

Brief Assessment

Linear Constraints Bandits[12] studies constrained bandits with reward/cost parameters, not inverse reward estimation from demonstrations. The lower bound in this candidate concerns regret bounds for constrained optimization, not estimation error for inverse learners.

---

10. One shot inverse reinforcement learning for stochastic linear bandits

URL: View paper

Prior Art Analysis

One Shot Inverse RL[1] establishes an information-theoretic lower bound for inverse reward estimation in stochastic linear bandits that predates the original paper. The candidate paper proves that for any inverse estimator, the estimation error is lower bounded by Ω(√(d/t)), using a Le Cam binary testing approach. This demonstrates that similar hardness results for inverse linear bandits existed prior to the original paper's submission, directly challenging the novelty claim that the authors were first to establish such fundamental limits.

Evidence

Evidence 1 - **Rationale**: The candidate paper explicitly states a lower bound of Ω(√(d/t)) for any inverse estimator in linear bandits, which is the same fundamental limit discussed in the original paper, showing prior establishment of such bounds. - **Original**: our main result is a unified approach for inverse linear bandits, based on the idea of formulating a linear program by tightly characterizing the confidence intervals of pulled actions. we show that the estimation error of our algorithms matches the information-theoretic lower bound, up to polynomia... - **Candidate**: theorem 5.1. for a bandit instance m characterized by reward parameter θ∗ 1 and action set a, there exists a bandit instance m′ with parameter θ∗ 2 and the same action set a such that any inverse estimator incurs error max{‖ˆθ - θ∗ 2‖2, ‖ˆθ - θ∗ 1‖2} = eω r d t

---

Contribution 3: Optimal unified reward estimator matching lower bound

Description: The authors develop a unified reward estimator that achieves estimation error matching the information-theoretic lower bound up to polynomial factors in dimension d and log T. The estimator requires only an approximation of the best reward, works without access to demonstrator hyperparameters or internal states, and places no assumptions on action set density or geometry.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Learning from an Exploring Demonstrator: Optimal Reward Estimation for Bandits

URL: View paper

Brief Assessment

Learning Exploring Demonstrator[5] addresses multi-armed bandits with discrete arms, while the original paper tackles linear bandits with continuous action sets using a unified LP-based approach that doesn't require hyperparameter access.

---

2. One shot inverse reinforcement learning for stochastic linear bandits

URL: View paper

Brief Assessment

One Shot Inverse RL[1] develops an estimator specifically for phased elimination with assumptions on action set density and geometry, requiring access to optimal reward and optimal arm. The original paper claims a unified estimator working for both LinUCB and phased elimination without such restrictive assumptions.

Appendix: Text Similarity Detection

Textual similarity detection checked 13 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. One shot inverse reinforcement learning for stochastic linear bandits

Detected in: Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Inverse Linear Bandits via Linear Programs View paper
- [1] One shot inverse reinforcement learning for stochastic linear bandits View paper
- [2] Optimal design for reward modeling in rlhf View paper
- [3] Pretraining Decision Transformers with Reward Prediction for In-Context Multi-task Structured Bandit Learning View paper
- [4] Prompt Tuning Decision Transformers with Structured and Scalable Bandits View paper
- [5] Learning from an Exploring Demonstrator: Optimal Reward Estimation for Bandits View paper
- [6] Reward-Biased Maximum Likelihood Estimation for Linear Stochastic Bandits View paper
- [7] Neural Contextual Bandits via Reward-Biased Maximum Likelihood Estimation View paper
- [8] In-Context Reinforcement Learning From Suboptimal Historical Data View paper
- [9] Causally Robust Preference Learning with Reasons View paper
- [10] Variance-Dependent Regret Bounds for Non-stationary Linear Bandits View paper
- [11] Impact of representation learning in linear bandits View paper
- [12] Stochastic bandits with linear constraints View paper
- [13] Balanced linear contextual bandits View paper
- [14] Linear Contextual Bandits with Interference View paper
- [15] Contextual bandits with linear payoff functions View paper
- [16] Optimal regret is achievable with bounded approximate inference error: An enhanced bayesian upper confidence bound framework View paper
- [17] Low-rank bandits via tight two-to-infinity singular subspace recovery View paper
- [18] Combinatorial bandits with linear constraints: Beyond knapsacks and fairness View paper
- [19] Online Decision Making Via Linear Programming: Resource Allocation, Bandit Feedback, and Inverse Optimization View paper
- [20] 12 Mathematics of Reinforcement View paper