Contextual Bandits for Large-Scale Structured Discrete Constrained Optimization Problems

Pavithra Harsha[1*] **Chitra Subramanian**[1] **Naoki Abe**[1] **Shivaram Subramanian**[1]
Amadou Ba[1] **Kevin Arturo Fernández Román**[2] **Mauricio Longinos Garrido**[2]
Miao Liu[1] **Aurélie Lozano**[1] **Chandrasekhar Narayanaswami**[1]
[1]IBM Research [2]IBM CIO [*]Contact Author: pharsha@us.ibm.com

Abstract

We study contextual bandits in high-dimensional combinatorial action spaces arising in structured constrained optimization problems, such as IT resource allocation and retail assortment pricing. Key quantities in the objective or constraints must be estimated from data during the trial sequence of actions. In [12], we propose a novel, practical, and transparent approach based on general-purpose regression oracles with Inverse Gap Weighting (IGW) for seamless integration within an optimization framework. IGW sampling is efficiently managed by: (a) a column-generation reformulation of the underlying Mixed Integer Programming (MIP) model which allows for flexible lower-level predictors, causal coherence, and efficient representation of large action spaces; (b) a diverse solution pool generation to balance the exploration-exploitation trade-off in large-action spaces. To address non-smooth rewards induced by constraints, we introduce a risk-averse phased learning strategy. Experiments on an IT auto-scaling task demonstrate substantial reductions in cumulative regret, with added gains from risk-averse methods that effectively manage constraint violations. This submission summarizes [12] and sketches our extensions towards a full theoretical regret analysis.

1 Introduction

In a contextual bandit setting a learning/decision-making agent, repeatedly receives a context, selects an action, based on past observations, and receives a reward. The learner's goal is to maximize the cumulative reward of an employed policy, striking the right balance between exploration and exploitation. In [12] we propose the design of a practical contextual bandit algorithm for scenarios with a high-dimensional action space involving a structured discrete constrained optimization problem. These problems frequently arise in the operations management domain such as IT resource allocation [16, 4] and retail assortment pricing [7, 13]. Rather than treating the whole action space as an opaque entity admitting a global reward regression function, as is done in past works on contextual bandits, here we take the *transparent white box* approach in which we explicitly leverage the rich structure that is inherent in the reward optimization problem.

First of all, in many real-world settings, the overall reward can be naturally formulated as a function in terms of lower-level estimable functionals which capture the causal structure necessary to represent the action-reward relationship. These lower-level functional relationships can be modeled effectively using machine learning methods, by leveraging the inherent problem structure that restricts the causal dependencies to smaller dimensions. The response variables of these lower-level models are predicted for incoming contexts and used in a combinatorial white-box optimization formulation. These formulations are generally more tractable than their black-box counterparts due to their structured nature. Leveraging a white-box approach allows us to ensure that the underlying lower-level machine learning models are causally adequate, namely, confounders are controlled for and other causal

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

constraints can be enforced (e.g., demand decreases with price, or latency with resources). This also improves transparency and interpretability and supports predictor reuse across use cases.

The second key characteristic of these applications is the presence of operational constraints such as latency targets, sales goals, or capacity limits. While such constraints can be added as penalty terms in generic reward learning frameworks, it undermines the reward smoothness along the constraint boundaries, challenging much of the guarantees on contextual bandit methods. It makes it difficult to satisfy requirements or attain high rewards due to causal inadequacies and constraint violations.

Our novel solution to this problem setting leverages cutting edge advances in two fields: contextual bandits and discrete constrained optimization. On the bandits side, recent work have given rise to the development of general, near-optimal algorithms that are based on 'regression oracles' [8, 9, 17, 5]. A key advantage of this type of approach in our context is its ability to accommodate complex, non-standard reward models, including those we aim to consider, in which lower-level response models are composed to derive the overall reward function. The core method that is relevant to our work is the probabilistic action selection strategy referred to as 'Inverse Gap Weighting' (IGW) [9, 2, 1]. Existing extensions for large action spaces either assume linear reward embeddings [18] or reward/action-space smoothness [19] or hierarchical action structure [15] and are not applicable to our setting. We build and extend the notion of 'diversity' in the IGW sampling introduced by [18]. On the discrete constrained optimization side, we leverage the advanced technique of Column Generation (CG) in our context and develop a master-subproblem reformulation of the underlying problem which is tractably solvable by mixed-integer programming (MIP) for many practical applications. This CG approach accommodates flexible, general-purpose predictors as input, ensures causal coherence and exploits the inherent lower-dimensional problem structure for a more efficient representation of the large action space. Our extensions to effectively handle optimization constraints highlight the dichotomy between extensive early exploration in traditional bandits and more conservative, risk averse exploration under constraints, with limited violations until boundaries are better understood.

2 Problem Setup

We study the following discrete structured constrained optimization problem:

$$P(x): \min_{\mathbf{a}\in\mathcal{A}} \sum_{i\in\mathcal{I}} c_i(x,\mathbf{a}_i), \quad \text{subject to } L_i(x,\mathbf{a}_i)\leq 0 \quad \forall i\in\mathcal{I}. \tag{1}$$

Let $x \in \mathcal{X}$ and $\mathbf{a} \in \mathcal{A}$ refer to the context and action vector in their respective spaces. The action set $\mathcal{A}$ restricts each action dimension $j \in J$ to a discrete set $\mathcal{A}^j$, so $\mathbf{a} = (a^j)_{j\in J}$ where $a^j \in \mathcal{A}^j$. The objective components c_i and constraint terms L_i for $i \in \mathcal{I}$ are the context- and action-dependent response variables, but each depends only on a lower-dimensional partial/sub action vector $\mathbf{a}_i = (a^j)_{j\in J_i} \in \mathcal{A}_i$ where $J_i \subset J$ and $|J_i| \ll |J|$, typically at most 3-4. These partial action vectors may overlap across i, preventing trivial decomposition. The constraints L_i capture performance relative to a user-specified threshold. As the L_i's are learned over time, we also use a soft-penalized formulation with user-defined λ_i $\forall i \in \mathcal{I}$:

$$P^\lambda(x): \min_{\mathbf{a}\in\mathcal{A}} \; Z^\lambda(x,\mathbf{a}), \;\; \text{where } Z^\lambda(x,\mathbf{a}) = \sum_{i\in I} c_i(x,\mathbf{a}_i) + \lambda_i L_i(x,\mathbf{a}_i)^+. \tag{2}$$

A contextual bandit protocol of the decision problem proceeds as follows in T rounds. In each round $t \in [T]$, the learner receives a context $x_t \in \mathcal{X}$, selects an action $\mathbf{a}_t \in \mathcal{A}$ and observes the response variables $\{c'_{ti}(\mathbf{a}_{ti}), L'_{ti}(\mathbf{a}_{ti})\}$ $\forall i \in \mathcal{I}$ and based on that realizes penalized cost $Z'^\lambda_t(\mathbf{a}_t) = \sum_{i\in I} c'_{ti}(\mathbf{a}_{ti}) + \lambda_i L'_{ti}(\mathbf{a}_{ti})^+$. We assume that in each round t, although the contexts are chosen arbitrarily, the response variables conditioned on the context are sampled from fixed but unknown distributions $\mathbb{P}_{c_i}(.|x)$ and $\mathbb{P}_{L_i}(.|x)$ $\forall i \in \mathcal{I}$, respectively. We assume the learner has access to a class of regression functions or predictors $\mathcal{F}_{c_i}, \mathcal{F}_{L_i^+}$ for each $i \in \mathcal{I}$ with bounded outputs, where $L_i^+ = \max\{L_i, 0\}$ denotes the constraint violation. We work with the predictors for constraint violation, L^+, rather than L, in order to recover the true expected cost, $E[Z'^\lambda_t(\mathbf{a}_{ti})] = \sum_{i\in I} E[c'_{ti}(\mathbf{a}_i)] + \lambda_i E[L'_{ti}(\mathbf{a}_{ti})^+]$. We make the following realizability assumption [3, 8, 9, 17], where the predictors α_i^* aim to approximate the *true cost and constraint violation functions*, respectively.

Assumption 2.1 ***(Realizability)****: There exists a regression function $\alpha_i^* \in \mathcal{F}_{\alpha_i}$ such that $E[\alpha'_{ti}(\mathbf{a}_i)|x_t = x] = \alpha_i^*(x,\mathbf{a}_i)$ for all $\mathbf{a}_i \in \mathcal{A}_i$ and $t \in [T]$ where $\alpha_i \in \{c_i, L_i^+\}, \alpha'_{ti} \in \{c'_{ti}, L'^+_{ti}\}$ $\forall i \in \mathcal{I}$.*

Regret: Given the set of predictors, $f = \{\hat{\alpha}_i \in \mathcal{F}_{\alpha_i} | \alpha_i \in \{c_i, L_i^+\}\ i \in \mathcal{I}\}$, we define the induced policy as $\pi_f(x) := \arg\min_{\mathbf{a}\in\mathcal{A}} Z_f^\lambda(x, \mathbf{a})$, where Z_f^λ is obtained from Eq.(2) by substituting predictors f. The aim of the learner is to minimize the regret with respect to the optimal policy $\pi^* := \pi_{f^*}$ defined as: $\texttt{Reg}_{\texttt{CB}} = \sum_{t=1}^{T} Z_t'^\lambda(a_t) - Z_t'^\lambda(\pi^*(x_t))$.

3 Proposed Solution

Our proposed solution integrates several key ideas described briefly below (refer to [12] for details).

Generating lower-level predictors: We assume access to regression oracles for lower-level predictors $\alpha_i \in \{c_i, L_i^+\}\ \forall i \in \mathcal{I}$ that are trained on sequence of context–action–response data. Using only the lower-dimensional action vector as input to the regression oracles, we effectively incorporate the inherent problem structure. The predictors can be general parametric and non-parametric models, such as neural networks, regression trees, random forests, and kernels, optionally with causal constraints, to generalize effectively across contexts and actions.

Optimizing MIP via column generation (CG): We employ CG, a decomposition technique compatible with any state-of-art regression model for lower-level functions, to efficiently solve the discrete optimization problem to near-optimality. Our CG approach exploits the sparsity in the correlations between decisions to decompose the model into a high-dimensional 'restricted' master program (RMP) and several low-dimensional subproblems. CG iterates between these levels till convergence and produces an RMP that can be solved directly as a MIP using off-the-shelf optimization packages such as CPLEX and Gurobi to obtain not just an optimal solution but a range of feasible solutions.

Generating a diverse solution pool: MIP solvers can generate and store multiple solutions beyond the optimal one in a configurable solution pool while traversing the combinatorial search tree. Users can specify the pool capacity, optimality gap (relative or absolute), and a diversity-based replacement strategy to maintain varied solution pool, all without affecting solver performance (see [6, 11]).

IGW sampling: The original sampling procedure targeted small, fixed action spaces, with regret scaling linearly in action-space size, making it unsuitable for large spaces. For large spaces [18] introduced reward-function approximations and diverse action pool-based IGW sampling. We devise a heuristic IGW method that samples only from the diverse solution pool, improving the exploration and exploitation balance in large spaces. The underlying equation are given in the proposed algorithm.

Proposed algorithm - OptCB: We now introduce an initial version of our proposed algorithm, referred to as *OptCB* (Optimization-based Contextual Bandit), with the details outlined in Algorithm 1. It seamlessly integrates key components: (a) general offline regression oracles for learning (steps 4,5,10), (b) MIP optimization with CG at the core and built-in diverse pool generation (step 6), and (c) IGW sampling (steps 7 and 8). Its practical design enables easy integration into standard workflows that already combine ML and discrete optimization.

Risk-averse learning with constraints: The exploration–exploitation principle, supported by theory [17, 18], suggests extensive early exploration followed by gradual reduction. Under constraints, however, violations heavily impact the penalized objective $Z^\lambda(x, \mathbf{a})$, and risk considerations discourage excessive early exploration due to cost discontinuities. While some boundary violations are necessary for learning, large violations are prohibitively costly and best avoided. This motivates a risk-averse strategy: prioritize exploration within the "known" boundary early on, then gradually expand around its vicinity as it is better understood. Algorithm 2 illustrates a simple two-phase variant (see [12] for another variant), where the transition is triggered by a significant change in the gradient of the observed cumulative squared error series $Z_t^\lambda(x, \mathbf{a})$, a proxy for regression regret.

4 Empirical Validation

We validate our approach in synthetic environments using real-world IT resource management data. Our method reduces cumulative regret by 80% over the vanilla IGW baseline in small action spaces (256 actions) and 90% in large spaces (10,000 actions), with further gains from risk-averse strategies (8–10% and 57–65%, respectively). See plots in appendix. At scale (100M actions), we achieve 1000× speedups (35 min/iteration vs. 2 sec). Statistical significance tests and ablation studies confirm that each component of our design is critical to maximizing cumulative rewards.

Algorithm 1 OptCB

Input:
Epoch schedule $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{T'} = T$.
MIP constraints target penalty vector: λ.
MIP pool parameters (gap, size, capacity): g, s, l.
IGW sampling parameters: γ_0, ρ.
for epoch $m = 0, 1, 2, \cdots, T' - 1$ **do**
  for round $t = \tau_m + 1, \cdots, \tau_{m+1}$ **do**
    Observe context x_t.
    Receive predictions $\hat{c}_i, \hat{L}_i^+$ from regression oracles.
    Solve the MIP reformulation $P^\lambda(x_t)$ with pool parameters (g, s, l) to obtain a diverse solution pool S.
    Set IGW sampling probability $p_t(.)$: (i) Evaluate penalized objective $Z^\lambda(\mathbf{a}) \ \forall \mathbf{a} \in S$. (ii) Define $S^* = \text{argmin}_{a \in S} Z^\lambda(\mathbf{a})$. (iii) Let $Z^{\lambda *}$ be the corresponding optimal objective. (iv) Set $\gamma_t = \gamma_0 \tau_m^\rho$ (see [12] for alternative approaches). (v) Determine sampling probability as follows:

$$p_t(\mathbf{a}) = \frac{1}{[|S| + \gamma_t \left(Z^\lambda(\mathbf{a}) - Z^{\lambda *}\right)]} \ \forall \ \mathbf{a} \in S \setminus S^*, \text{ and } p_t(\mathbf{a}) = \frac{\left[1 - \sum_{\mathbf{a} \in S \setminus S^*} p_t(\mathbf{a})\right]}{|S^*|} \ \forall \ \mathbf{a} \in S^*,$$

    Sample $\mathbf{a}_t \sim p_t(.)$, execute $\mathbf{a}_t$ and observe resultant response variables $c_{it}, L_{it} \ \forall i \in I$
  end for
  Re-train regression predictors $\hat{c}_i, \hat{L}_i^+$ with all data including that from epoch m using offline least squares.
end for

Algorithm 2 Two-phase RiskAverse OptCB
Assumption: Monotonic Predictors $L_i(x, .)$
Same as OptCB barring following changes

Input (additional):
Initialize power-law exponent: $\theta_0 > 1$.
Target phase transition power-law exponent: $\bar{\theta}$.
Additional IGW sampling parameter: $\bar{\rho} > \rho$.
In step 6, solve reformulation $P(x_t)$ with hard constraints if $\theta_{\tau_m} > \bar{\theta}$ and its soft counterpart $P^\lambda(x_t)$ otherwise. To ensure hard problem is not infeasible, if for any $i \in \mathcal{I}$, $\min L_i(x_t, .) > 0$, we modify $L_i(x_t, .)$ to $L_i(x_t, .) - \min L_i(x_t, .)$ to obtain a solution in round t.
In step 7(iv), set $\gamma_t = \gamma_0 \tau_m^\rho$ if $\theta_{\tau_m} > \bar{\theta}$, else set $\gamma_t = \gamma_0 \tau_m^{\bar{\rho}}$ (see [12] for alternative approaches).
After step 8, compute regression regret $\text{Reg}_{\text{Sq}}(.)$ or its proxy, cumulative squared error of $Z^\lambda(x, \mathbf{a})$.
After step 10, perform log-log regression on all the computed $\text{Reg}_{\text{Sq}}(.)$ values to obtain the power-law coefficient $\theta_{\tau_{m+1}}$.

5 Regret Analysis

IGW sampling, combined with an estimation–decision meta-algorithm, unifies several prior regret guarantees in contextual bandits through the decision-to-estimation (DEC) coefficient [10]. Bounded DEC, together with bounded squared-loss regression regret under the realizability assumption, yields optimal regret guarantees in contextual information settings with finite, linear and certain continuous action settings [9, 10]. However, existing works for bounding DEC, including that in [18], do not extend naturally to the discrete, large-scale combinatorial settings considered here.

We introduce a refined DEC notion that captures the interplay between the decomposed estimation error and its impact on decision quality in our formulation. Recall that our approach leverages the MIP solver's ability to generate diverse near-optimal solutions during branch-and-bound search while maintaining performance and ensuring the solutions are within a user-specified optimality gap. The IGW sampling scheme over the diverse solution pool allows us to visit the ϵ-independent dimensions in finite time t_ϵ, with the number of such dimensions upper bounded by the Eluder dimension [14]. Using the optimality gap together with the Eluder dimension, we upper bound the refined DEC, in the penalized reward setting, to obtain $\sqrt{T}$ regret when T exceeds t_ϵ. On the other hand when $T < t_\epsilon$ the regret is controlled by the radius of the ball covering the action space leveraging the diverse solution pool. Extending this theoretical treatment to capture the trade-off between early exploration and the risk of large constraint violations, and to develop alternative risk-aware estimation-to-decision algorithms, remains an open and promising direction for future research.

A Appendix: Regret Plots

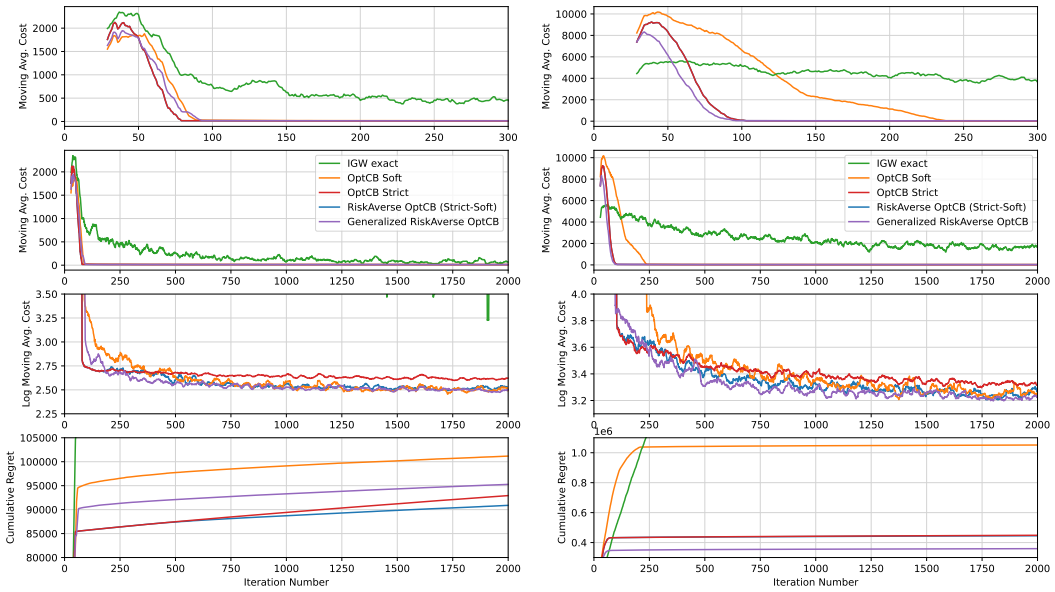

Figure 1: Instantaneous cost and cumulative regret metrics for: (left plot) small-scale environment with 256 combinatorial actions; and (right plot) large-scale environment with 10K combinatorial actions.

References

[1] Naoki Abe, Alan W Biermann, and Philip M Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37:263–293, 2003.

[2] Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

[3] Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

[4] Romil Bhardwaj, Kirthevasan Kandasamy, Asim Biswal, Wenshuo Guo, Benjamin Hindman, Joseph Gonzalez, Michael Jordan, and Ion Stoica. Cilantro: Performance-Aware resource allocation for general objectives via online feedback. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 623–643, 2023.

[5] Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *Journal of Machine Learning Research*, 22(133):1–49, 2021.

[6] CPLEX. Cplex solution pool. `https://www.ibm.com/docs/en/icos/22.1.2?topic=solutions-what-is-solution-pool`.

[7] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & service operations management*, 18(1):69–88, 2016.

[8] Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

[9] Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[10] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[11] Gurobi. Gurobi solution pool. `https://docs.gurobi.com/projects/optimizer/en/current/features/solutionpool.html#solution-pool`.

[12] Pavithra Harsha, Chitra Subramanian, Naoki Abe, Shivaram Subramanian, Amadou Ba, Kevin Arturo Fernández Román, Mauricio Longinos Garrido, and Chandrasekhar Narayanaswami. Practical contextual bandits for large-scale structured discrete constrained optimization problems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 850–859, 2025.

[13] Pavithra Harsha, Shivaram Subramanian, and Markus Ettl. A practical price optimization approach for omnichannel retailing. *INFORMS Journal on Optimization*, 1(3):241–264, 2019.

[14] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[15] Rajat Sen, Alexander Rakhlin, Lexing Ying, Rahul Kidambi, Dean Foster, Daniel N Hill, and Inderjit S Dhillon. Top-k extreme contextual bandits with arm hierarchy. In *International Conference on Machine Learning*, pages 9422–9433. PMLR, 2021.

[16] Jiuchen Shi, Hang Zhang, Zhixin Tong, Quan Chen, Kaihua Fu, and Minyi Guo. Nodens: Enabling resource efficient and fast {QoS} recovery of dynamic microservice applications in datacenters. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, pages 403–417, 2023.

[17] David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

[18] Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pages 27428–27453. PMLR, 2022.

[19] Yinglun Zhu and Paul Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *International Conference on Machine Learning*, pages 27574–27590. PMLR, 2022.