# Contextual Value Iteration and Deep Approximation for Bayesian Contextual Bandits

**Kevin Duijndam**
Department of Mathematics
Vrije Universiteit Amsterdam
kevin.duijndam@klm.com

**Ger Koole**
Department of Mathematics
Vrije Universiteit Amsterdam
ger.koole@vu.nl

**Rob van der Mei**
Department of Probability and Stochastic Networks
Centrum voor Wiskunde & Informatica
mei@cwi.nl

## Abstract

We present a Bayesian value-iteration framework for contextual multi-armed bandit problems that treats the agents posterior distribution for the pay-off as the state of the Markov Decision Process. We apply finite-dimensional priors on the unknown reward parameters, and the exogenous context transition kernel. Value iteration on the belief-MDP yields an optimal policy. We illustrate the approach in an airline seat-pricing simulation. To address the curse of dimensionality, we approximate the value function with a dual-stream deep learning network and benchmark our deep value iteration algorithm on a standard contextual bandit instance.

## 1 Introduction and background

The multi-armed bandit (MAB) problem, first posed by [2], captures the tension between learning and earning when decisions must be taken sequentially. In the classical, context-free formulation, index-based optimal solutions exist. Real applications, however, are rarely independent over arms and context-free. Once such coupling between arms or context is present, the separate-index trick breaks down. To still find an optimal policy, we treat all uncertainty  arm payoffs and exogenous context dynamics  in a fully Bayesian manner, so the agents belief becomes the state; learning then reduces to an MDP solved by value iteration. We next formalize the belief-MDP and show how value iteration and its deep approximation are implemented. All code and configuration files to reproduce the experiments and figures are available at [17] (commit and configurations).

### 1.1 Background

Classical approaches - such as $\varepsilon$-greedy ([6]), Upper Confidence Bound (UCB, starting from [5], then described more specifically in [7]), and Thompson Sampling (TS, [1]) - use heuristic rules to balance exploration and exploitation and are computationally efficient, but typically do not guarantee optimal performance. Optimal policies can be obtained using methods like the Gittins index ([3, 4, 8]), which can be computationally fast, but only works for context-free MAB problems. The literature on contextual MAB spans a diverse range of approaches with many contributions in recent years, including comprehensive overviews and benchmarks of deep learning bandit algorithms ([9]), algorithms that adapt TS to the context and add optimistic bonuses ([16] and [13]), adversarial linear contextual bandits ([14]), reductions to linear bandits to achieve competitive regret bounds

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

([12]), frequentist uncertainty measures combined with deep learning ([10]), Kalman filtering with low-dimensional parameter subspaces ([11]), and tree ensembles to incorporate context alongside UCB/TS-style exploration ([15]).

### 1.2 Our contribution

1. Belief-MDP formulation including pay-off parameters and an exogenous context kernel.
2. Contextual Value Iteration (CVI): exact value iteration on a smaller belief grid.
3. Deep Value Iteration (DVI): a dual-stream deep value approximation to scale to larger problems.

## 2 Methodology

### 2.1 Value iteration in the contextual Bayesian MAB setting

We consider a contextual MAB problem modeled in a Bayesian manner. Let $A$ be a finite action set (arms), $S$ the state space where each state $s \in S$ encodes the current estimates for the parameters of the payoff distribution, and $C$ a (finite) context space. We assume $c_{t+1} \sim \mu_\xi(\cdot \mid c_t)$ (exogenous, independent of actions). So $\mu$ defines the distribution of the next context state, and does not depend on the action chosen. Define the augmented state space $X = S \times C$ with $x = (s, c)$. Let $R : X \times A \to \mathbb{R}$ be a bounded reward function and $P(s', c' \mid (s, c), a)$ the transition probability from $(s, c)$ to $(s', c')$ upon playing $a$, as determined by the Bayesian update. Note that we allow some underlying parameter $\theta$ that parametrizes the pay-off distribution as $r_t \sim \mathcal{D}_\theta(c_t, a_t)$. With discount factor $\gamma \in (0, 1)$, the value function satisfies

$$V(s, c) = \max_{a \in A} \Big\{ R(s, c, a) + \gamma \, \mathbb{E}\big[V(s', c') \mid s, c, a\big] \Big\}.$$

Value iteration converges to the unique fixed point $V^\star$; the greedy policy $a^\star(x) = \arg\max_{a \in A}\{R(x, a) + \gamma \, \mathbb{E}[V^\star(x') \mid x, a]\}$ maximizes expected discounted reward with respect to the discretized belief-MDP we solve here. The posterior update inside $V$ already optimally weighs information gain against immediate revenue, so no explicit exploration heuristic is required. This is the Contextual Value Iteration (CVI) algorithm.

### 2.2 Deep value iteration via function approximation

To address the curse of dimensionality inherent in tabulating $V$ on large belief grids, we approximate the value function using a neural network that closely follows the theoretical structure. We employ a dual-stream architecture that explicitly separates the contribution from arm uncertainty and contextual information: the base stream embeds each arms belief parameters and aggregates across arms before passing through an MLP to yield a base value; in parallel, the context stream projects the context through its own MLP to produce a context-dependent correction. The two outputs are concatenated and passed through a final fusion block that learns a non-linear combination, yielding the overall value estimate. Offline, the network is pre-trained using randomly sampled states (with zero context) to robustly learn the base value function; afterwards, the base stream is frozen. In the online phase, a warm-up stage with actual transitions (chosen arm, context, reward) learns the context corrections, after which experience replay with mini-batch updates stabilizes temporal-difference learning. This is the Deep Value Iteration (DVI) algorithm.

## 3 Results

### 3.1 Airline seat pricing (Poisson GLM).

We evaluate CVI on a small but complete pricing problem to assess the approach in 2.1. We simulate an airline that needs to decide the price for a seat, where there are multiple types of flights (the context). The price used can be seen as the action, or arm, chosen.

**Model** At each round $t$ we observe a discrete context $c_t \in \{0, 1\}$, with $c = 0$ denoting leisure and $c = 1$ business. We choose a price $a_t \in \mathcal{P}$ from a small finite grid and realize seats sold

$$y_t \sim \text{Poisson}(\lambda_\theta(c_t, a_t)), \qquad \lambda_\theta(c, a) \;=\; \exp\big(\theta^\top \phi(c, a)\big),$$

with revenue $r_t = a_t\, y_t$. Context evolves exogenously via a Markov kernel $\mu(c' \mid c)$.

To capture that business is less price elastic, we use the 4-dimensional feature map

$$\phi(c, a) \;=\; \big[\,1,\; a,\; \mathbf{1}\{c{=}1\},\; \mathbf{1}\{c{=}1\} \cdot a\,\big]^\top,$$

with corresponding parameter names $\theta = \big[\theta_0, \theta_p, \theta_{b0}, \theta_{bp}\big]^t$. This gives the log-intensity $\log \lambda_\theta(c, a) = \theta_0 + \theta_p \cdot a + \theta_{b0}\mathbf{1}\{c{=}1\} + \theta_{bp}\mathbf{1}\{c{=}1\} \cdot a$ so that leisure has logintensity $\theta_0 + \theta_p a$ and business $\theta_0 + \theta_{b0} + (\theta_p + \theta_{bp})a$.

**Belief and planning** We maintain a Gaussian belief on $\theta$: $\theta \sim \mathcal{N}(m, \Sigma)$ and update the mean with a one-step Laplace/EKF correction, while keeping $\Sigma$ fixed to the prior covariance $\Sigma_0$ to keep the belief state finite. We discretize $m$ on a tensor grid $\mathcal{M}$ centered at the prior mean and define the belief state as $(m, c) \in \mathcal{M} \times \{0, 1\}$. Running value iteration on this finite state space yields a table $V^*$ and a greedy policy $\pi^*(m, c)$ that we use online by look-up.

**Experimental setup** True parameters are chosen to yield realistic elasticities. The agents prior is close but miscalibrated on the slopes to permit learning. We grid each coordinate of $m$ with 5 points over $\pm 1.5$ prior standard deviations (a $5{\times}5{\times}5{\times}5$ grid), set $\gamma = 0.99$, truncate the Poisson support at $y_{\max} = 12$, and jitter the initial belief to avoid identical early actions across episodes. We compare against contextual $\varepsilon$-greedy, contextual Thompson sampling, LinUCB, LinTS, and Tree-UCB using the same GLM and features. Each method is run for 2500 episodes of 500 rounds. We report total expected regret and mean cumulative expected regret, where we compare against an oracle that knows all parameters and selects $a_t^*(x_t)$, so results and confidence intervals reflect expectations under the data-generating process and leave out realized noise.

**Results** Figure 1a plots the total expected regret with a 95%-confidence interval, and figure 1c shows the cumulative expected regret over time. CVI outperforms benchmark algorithms by a factor of $\approx 1.7\times$ in total expected regret, with comparable variance across runs. Wall-clock on a MacBook Pro (M1) laptop: computing $V^*$ once takes $\approx$170 seconds; the subsequent online phase (all $2500{\times}500$ decisions) takes $\approx 50$ seconds, compared to Contextual $\varepsilon$-Greedy: $\approx 40\,s$, Contextual TS: $\approx 110\,s$, LinTS: $\approx 90\,s$, LinUCB: $\approx 60\,s$, Tree-UCB: $\approx 1130\,s$.

### 3.2 Logistic Contextual Bandit simulation

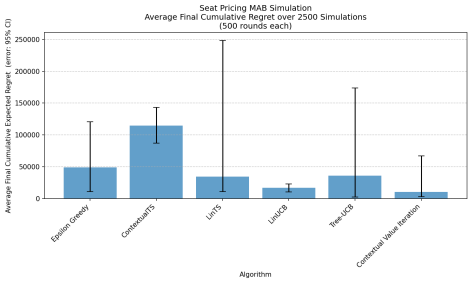
We evaluate Deep Value Iteration (DVI) on a synthetic but standardized contextual bandit in which the dominant statistical challenge is learning arm payoffs under uncertainty rather than context classification. We adopt standard linear/ensemble baselines representative of contextual bandit practice.

At each round $t$, a $d$dimensional context $x_t \sim \mathcal{N}(0, I_d)$ is observed and an arm $a_t \in \{1, \ldots, K\}$ is chosen. The Bernoulli reward obeys a logistic GLM with a shared slope and perarm intercepts:

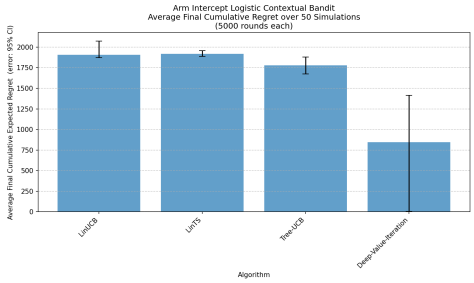
$$\Pr(r_t{=}1 \mid x_t, a_t{=}a) \;=\; \sigma\big(\alpha_a + x_t^\top \beta\big), \qquad \sigma(z) = \tfrac{1}{1+e^{-z}}.$$

Initialization follows a fixed generative recipe (same across methods): arm effects $\alpha_a \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$ with $\sigma_\alpha = 1.0$, and a shared slope $\beta$ drawn from $\mathcal{N}(0, I_d)$ and then rescaled to $\|\beta\|_2 = 0.4\,\sigma_\alpha$. Because $x \sim \mathcal{N}(0, I_d)$, the contextual term $x^\top \beta$ is $\mathcal{N}(0, \|\beta\|_2^2)$ with standard deviation 0.4, i.e., context contributes non-trivially, but is smaller than the intercept variability. This makes the problem armcentric (learning the unknown $\{\alpha_a\}$ is crucial), while still allowing the optimal arm to depend on $x$.

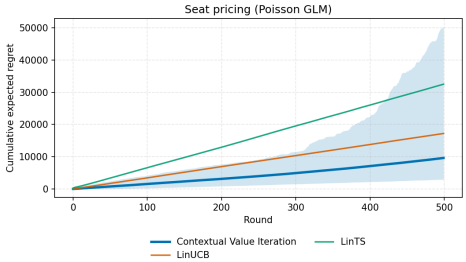
**Setup and metrics** We use $K{=}20$ arms, $d{=}10$ context dimensions, $T{=}5000$ rounds, averaged over 50 independent seeds. Performance is reported as expected cumulative regret against a contextual oracle that knows $(\alpha, \beta)$ and plays $a_t^\star(x_t) = \arg\max_a \sigma(\alpha_a + x_t^\top \beta)$ at every round. We also report end-to-end wall-clock time.




(a) Seat pricing (Poisson GLM) - expected regret over 2500 episodes, 500 rounds each. CVI shows $\approx 1.7\times$ lower expected regret vs LinUCB, LinTS and Tree-UCB.




(b) Contextual Bandit ($K$=20, $d$=10, $T$=5000, 50 seeds): Error bars: 95%-confidence interval. DVI shows $\approx 2\times$ lower expected regret vs LinUCB, LinTS and Tree-UCB.




(c) Seat pricing (Poisson GLM) - mean cumulative expected regret over 2500 episodes, 500 rounds each. Shaded band (95% confidence interval) shown only for CVI for readability.

Figure 1: Results of CVI and DVI benchmarks.

**Compared methods** We compare DVI to strong contextual bandit baselines: LinUCB, LinTS, and Tree-UCB. For linear/GLM methods we supply the hybrid feature map $\phi(x, a) = [\, x;\, e_a \,]$, so that all baselines can represent perarm fixed effects (intercepts) alongside the shared slope on $x$.

**Results** Figure 1b summarizes average final expected cumulative regret with a 95%-confidence interval. DVI attains substantially lower regret than the baselines (about $2\times$ improvement on average), demonstrating that planning with a value function over belief states can efficiently trade off exploration and exploitation in this arm-uncertainty regime. DVIs across-seed variability is larger, which can follow from the nonconvex value-function training and approximate planning, but the mean performance advantage is clear. Online wall-clock on a MacBook Pro (M1): DVI $\approx 950\,s$, LinUCB $\approx 20\,s$, LinTS $\approx 30\,s$, Tree-UCB $\approx 970\,s$, so DVI has a relatively longer run-time, but remains manageable.

## 4 Conclusion and discussion

By defining the state of a contextual MAB as its posterior estimates, the problem becomes an MDP and value iteration yields an optimal policy (for the discretized belief-MDP). Context-free index policies arise as special cases. We have shown that CVI works on a practical pricing simulation, and that a dual-stream deep learning approximation makes the approach tractable at larger scale. The approach remains sensitive to priors and suffers the curse of dimensionality for large exact grids - motivating function approximation - yet offers a practical, model-based alternative to heuristic exploration with competitive empirical performance.

## References

[1] William R Thompson. “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. In: *Biometrika* 25.3-4 (1933), pp. 285–294.

[2] Herbert Robbins. “Some aspects of the sequential design of experiments”. In: (1952).

[3] John Gittins. “A dynamic allocation index for the sequential design of experiments”. In: *Progress in statistics* (1974), pp. 241–266.

[4] John C Gittins. “Bandit processes and dynamic allocation indices”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 41.2 (1979), pp. 148–164.

[5] Tze Leung Lai and Herbert Robbins. “Asymptotically efficient adaptive allocation rules”. In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

[6] Nicolo Cesa-Bianchi and Paul Fischer. “Finite-time regret bounds for the multiarmed bandit problem.” In: *ICML*. Vol. 98. Citeseer. 1998, pp. 100–108.

[7] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. “Finite-time analysis of the multiarmed bandit problem”. In: *Machine learning* 47 (2002), pp. 235–256.

[8] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[9] Carlos Riquelme, George Tucker, and Jasper Snoek. “Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling”. In: *arXiv preprint arXiv:1802.09127* (2018).

[10] Rong Zhu and Mattia Rigotti. “Deep bandits show-off: Simple and efficient exploration with deep networks”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17592–17603.

[11] Gerardo Duran-Martin, Aleyna Kara, and Kevin Murphy. “Efficient online bayesian inference for neural bandits”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 6002–6021.

[12] Osama A Hanna, Lin Yang, and Christina Fragouli. “Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms”. In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 1791–1821.

[13] Yuko Kuroki et al. “Best-of-Both-Worlds Algorithms for Linear Contextual Bandits”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 1216–1224.

[14] Haolin Liu, Chen-Yu Wei, and Julian Zimmert. “Bypassing the simulator: Near-optimal adversarial linear contextual bandits”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[15] Hannes Nilsson et al. “Tree ensembles for contextual bandits”. In: *arXiv preprint arXiv:2402.06963* (2024).

[16] Ruitu Xu, Yifei Min, and Tianhao Wang. “Noise-adaptive thompson sampling for linear contextual bandits”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[17] Kevin Duijndam. *Contextual Value Iteration and Deep Value Iteration Notebook*. Version v1.0. Jupyter notebook and scripts to reproduce all figures and tables. Commit: 7e4147e. GitHub, 2025. URL: `https://github.com/KevinDuijndam/ContextualValueIteration/blob/main/CVI-DVI-notebook.ipynb` (visited on 08/30/2025).