Utilizing Historical Data for Neural Bandits with Domain Shift

Donovan Barcelona [1] **Lily Xu** [1]

Abstract

Integrating historical data into multi-armed bandits is a critical challenge, as indiscriminately incorporating biased offline data can lead to unwanted negative transfer. Motivated by high-stakes clinical settings where skewed training distributions have historically induced algorithmic bias (Kallus et al., 2020), we investigate how to safely bridge the offline-to-online gap. We extend the Artificial Replay meta-algorithm (Banerjee et al., 2022) to contextual neural bandits, accommodating the continuous context space by introducing closeness heuristics that enforce conservative adaptation by restricting historical replays to contexts with high spatial proximity. Empirically, we show that under severe domain shift, both full pre-training and unconstrained replay perform worse than learning entirely from scratch. Conversely, our spatial filtering approach effectively extracts the utility of biased historical data while avoiding negative transfer, consistently outperforming pure online baselines, full historical starts, and unrestricted Artificial Replay.

1. Introduction

Correctly dosing medications like Warfarin presents a significant clinical challenge due to high physiological variability among individuals. Compounding this difficulty, previous dosing algorithms have exhibited disproportionate performance across protected classes, largely driven by unbalanced training data (Kallus et al., 2020). As a motivating example, this highlights a broader challenge in offline-to-online adaptation: historically underrepresented groups frequently suffer from algorithmic bias stemming from skewed offline datasets (Suresh & Guttag, 2021).

For sequential decision-making challenges such as personalized medicine, we can formulate these problems within a contextual multi-armed bandit framework, mapping clinical and demographic features to high-dimensional context vectors. Recent advancements in neural bandits have enabled greater expressivity for such complex contexts by adaptively training networks to map high-dimensional inputs into embedding spaces for reward and uncertainty estimation (Riquelme et al., 2018). However, deep learning algorithms remain notoriously sensitive to the distributions of their training data (Zhang et al., 2021; Boopathy et al., 2023; Shwartz-Ziv et al., 2023; Yan et al., 2024; Ye et al., 2021). This sensitivity is exacerbated by domain shift, where the target distribution of online samples diverges significantly from the historical offline data.

Determining how best to effectively integrate potentially biased historical data into neural bandits remains a critical, yet underexplored, area. Recently, Artificial Replay was introduced as a meta-algorithm for finite-armed bandits which historical data as a strategic replay buffer rather than a static training set, only integrating historical samples as they become relevant during online arm selection (Banerjee et al., 2022). We hypothesize that neural bandits represent a valuable setting for Artificial Replay, allowing algorithms to mitigate representation bias by dynamically filtering historical samples rather than blindly absorbing them into network priors.

In this short paper, we investigate the offline-to-online gap for neural bandits under domain shift. First, we characterize adversarial problem settings where naively warm-starting on biased offline data leads to catastrophic negative transfer, yielding worse online performance than utilizing no history at all. Second, we extend Artificial Replay from finite-armed bandits to contextual neural bandits by replacing exact historical matches with distance-based and structural "closeness heuristics." By restricting historical replays to contexts with high spatial proximity to the online state, these heuristics enforce conservative adaptation, allowing algorithms to safely extract the utility of skewed offline data without succumbing to it. Third, we empirically show that our approach can outperform both full historical pre-training and unconstrained replay by 20.3% and 38.1%, respectively, under severe distribution mismatch.

[1]Columbia University, New York, NY, USA. Correspondence to: Donovan Barcelona <donovan.barcelona@columbia.edu>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

2. Related Work

Offline-to-online bandits under distribution shift. While the broader machine learning community has extensively studied domain shift and concept drift (Li et al., 2024), addressing the offline-to-online gap remains an emerging frontier for multi-armed bandits. Recent works have explored transfer learning for contextual bandits with non-matching contexts (Cai et al., 2024) and established theoretical bounds for non-contextual bandits under concept drift (He et al., 2026). Our work directly targets this gap by building upon the Artificial Replay (AR) meta-algorithm (Banerjee et al., 2022). While AR offers theoretical advantages in sample efficiency and reduced computational overhead for storing historical data, a limitation of the original framework is that empirically it only demonstrated the ability to *match*, rather than strictly *improve upon*, the performance of a full historical start. Furthermore, AR was originally designed for finite-armed bandits in stationary environments. We extend AR to the contextual setting under explicit domain shift. Because standard AR relies on exact historical matches—which are vanishingly rare in continuous, high-dimensional spaces—we introduce distance-based heuristics to enforce conservative behavior. This targeted spatial filtering prevents blind historical incorporation, unlocking the ability to actively outperform full pre-training when domains are misaligned.

Representation bias in neural bandits. Deep learning models are highly prone to overfitting when trained on narrow or skewed data distributions (Zhang et al., 2021), resulting in brittle representations that generalize poorly to out-of-distribution data. This fragility is exacerbated in sequential online environments, where the hypothesis space must adapt without access to a perfectly representative historical dataset (Boopathy et al., 2023). In high-stakes applications like personalized healthcare, initializing a model on biased data can disproportionately harm historically underrepresented groups. We advance the Neural Linear framework (Riquelme et al., 2018) by demonstrating how this representation bias manifests as harmful negative transfer under domain shift, and how conservative replay mechanisms can successfully mitigate it.

3. Model and Approach

Problem formulation. We consider contextual bandit problems with K discrete actions. We assume access to a static, historical offline dataset $\mathcal{H} = \{(x_h, a_h, r_h)\}_{h=1}^{H}$. A historical datapoint h includes contexts $x_h \in \mathbb{R}^d$ drawn from an offline distribution $\mathcal{D}_{\text{off}}$, historical action $a_h \in [K]$, and observed reward $r_h \in \mathbb{R}$. During online deployment, at each timestep t, the environment reveals a new context x_t drawn from an online distribution $\mathcal{D}_{\text{on}}$. Crucially, we consider settings with *domain shift* ($\mathcal{D}_{\text{off}} \neq \mathcal{D}_{\text{on}}$). To handle complex context-reward mappings, we utilize the Neural Linear framework (Riquelme et al., 2018), learning a neural network to map raw contexts x_t to dense representations z_t for Bayesian linear regression of the form

$$\hat{r}_a = z_t^\top \beta_a$$

to estimate both the reward and uncertainty for each arm. While highly expressive, deep neural networks—and by extension, the Neural Linear algorithm—are prone to overfitting and poor generalization when naively initialized on heavily biased offline data (Zhang et al., 2021).

Artificial Replay for conservative adaptation. To bridge the offline-to-online gap, we extend the neural linear framework with Artificial Replay. Rather than universally warm-starting the algorithm by pre-training on the entirety of the historical data $\mathcal{H}$ ("Full Start")—which risks embedding the biases of $\mathcal{D}_{\text{off}}$ into the model priors—we treat the offline dataset as a dynamic episodic memory buffer.

During online interaction, if the neural linear algorithm proposes an action a_t for context x_t, it dynamically queries $\mathcal{H}$. We enforce *conservative behavior* by replaying a historical sample only if its logged action matches a_t and its context satisfies a defined "closeness heuristic". We introduce three distinct heuristics to filter these approximate matches:

- **Baseline AR:** Greedily selects the nearest, unused historical context that shares the proposed action, utilizing a chosen distance metric $d(\cdot, \cdot)$:

$$h^\star = \arg \min_{\{h \in [H] | a_h = a_t\}} d(x_t, x_h).$$

- **AR (Match Signs):** Acts as a lightweight structural prior. A historical sample is only eligible for replay if its context vector lies within the exact same orthant as the online context x_t (i.e., identical feature signs).

- **AR (ϵ-matching):** Enforces strict spatial proximity. It only permits replays if a matching-action historical context falls within an absolute distance threshold ϵ of the online context:

$$h^\star = \arg \min_{\{h \in [H] | a_h = a_t,\ d(x_t, x_h) \leq \epsilon\}} d(x_t, x_h).$$

The complete, mathematically formal execution sequence of the Neural Linear framework augmented with this conservative adaptation mechanism is detailed in Appendix A.

4. Synthetic Experiments

We design a 2D synthetic benchmark to isolate the effects of distribution shift between historical datasets and online deployment, evaluating how safely methods leverage biased priors.

4.1. Environment Setup

Contexts, Actions, and Rewards. Contexts $x \in \mathbb{R}^3$ (including an intercept) are sampled from a 2D unit ball. We evaluate $K = 4$ arms, where each arm's weight vector θ_a is constrained to a distinct quadrant (Figure 1). We primarily utilize a cubic reward function, $r(x, a) = 10(x^\top \theta_a)^3 - 2(x^\top \theta_a)$, creating complex, nonlinear relationships between context geometry and expected reward.

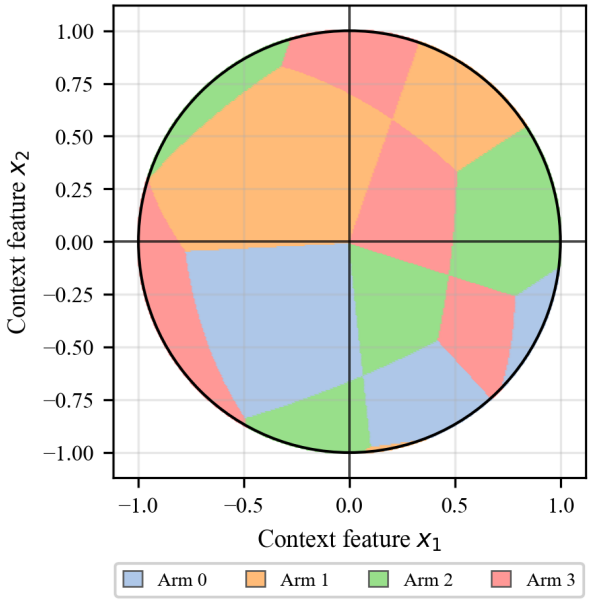

Figure 1. Optimal arm regions for the cubic reward function $r(x, a) = 10(x^\top \theta_a)^3 - 2(x^\top \theta_a)$ with $d = 2$ and $K = 4$. The arm intercept is dropped and weight vectors are constrained so each arm behaves predictably in a given quadrant: $\theta_0 = [+0.09, +0.42]$, $\theta_1 = [+0.13, -0.49]$, $\theta_2 = [-0.43, +0.36]$, $\theta_3 = [-0.61, -0.22]$.

Domain Shift and Action Bias. To simulate realistic distribution shifts, offline and online contexts are sampled from distinct distributions over the unit ball. Offline contexts are either drawn uniformly, or subjected to a spatial bias where a probability mass $p \in (0, 1]$ is strictly concentrated within a target quadrant, with the remaining $1 - p$ drawn uniformly. We independently skew the historical action distribution (e.g., an offline policy that pulls Arm 1 80% of the time) to decouple context mismatch from action coverage mismatch. This design mirrors real-world scenarios—such as clinical dosing—where historical logged data may be biased toward a specific standard-of-care intervention, reflecting a lack of past exploration across alternative treatments.

4.2. Evaluation Protocol

We evaluate over $T = 1000$ online timesteps with a historical buffer of $H = 100$, averaging across 36 trials (reporting mean $\pm$ 1 SEM). We use the Neural Linear bandit from Riquelme et al. (2018) (MLP sizes $[50, 50]$, learning rate 0.01). Baselines include **Full Start** (pre-training on all offline data), **No History** (pure online learning), and our **Artificial Replay (AR)** variants. Standard AR greedily replays the nearest matching-action historical sample, while **AR** ($\epsilon = 0.5$) restricts replays to contexts within a 0.5 Euclidean distance, and **AR (Match Signs)** requires identical feature signs.

4.3. Experiment 1: Opposing Quadrant Mismatch

We first examine an adversarial shift where offline data is drawn entirely from Q1 and the online environment draws entirely from Q4. We assume a historical policy that pulls Arm 1 80% of the time. Because the optimal arm in Q1 is suboptimal in Q4, incorporating offline data naively risks reinforcing incorrect action preferences.

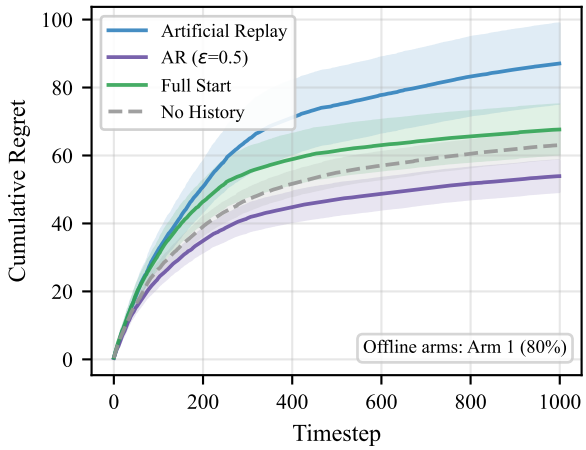

Figure 2. Cumulative regret for Q1 $\rightarrow$ Q4 domain shift ($H = 100$, $T = 1000$, 36 trials, mean $\pm$ 1 SEM). AR (ϵ=0.5): 53.9 ± 4.6; Full Start: 67.6 ± 7.3; No History: 63.0 ± 3.9; AR: 87.1 ± 11.7. AR (Match Signs) is excluded because it has identical performance to No History in this setting.

As shown in Figure 2, both Full Start (67.6 ± 7.3) and standard AR (87.1 ± 11.7) suffer from the heavily biased initialization, performing worse or no better than the pure online No History baseline (63.0 ± 3.9). This suggests that indiscriminately incorporating out-of-domain historical data can delay genuine exploration and hinder online adaptation.

However, the Artificial Replay framework remains highly effective when properly constrained. By enforcing spatial proximity before a replay is accepted, **AR (ϵ=0.5)** achieves the lowest regret at 53.9 ± 4.6, outperforming No History by a significant margin. The ϵ-threshold successfully filters out misleading offline data, enabling the agent to safely extract value from a smaller fraction of historically relevant samples while avoiding the negative transfer that harms Full Start and standard AR.

4.4. Experiment 2: Biased Offline to Uniform Online

We next test a softer mismatch: offline data is entirely from Q4 under a historical policy that pulls Arm 0 80% of the time, while online contexts are drawn uniformly across the unit ball. This evaluates whether our methods can gracefully adapt to broader online exploration when partial domain overlap exists.

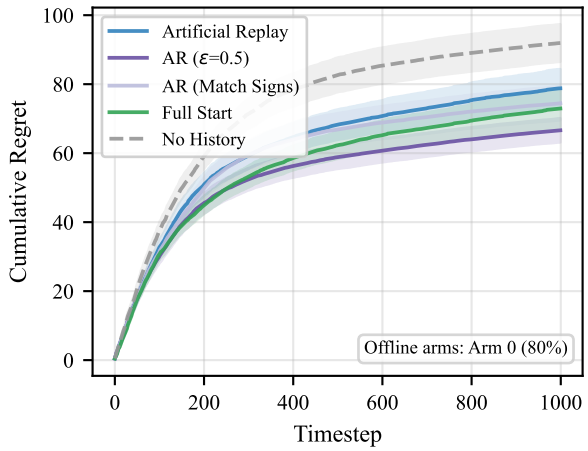

Figure 3. Cumulative regret for Q4 → Uniform domain shift ($H = 100$, $T = 1000$, 36 trials, mean ± 1 SEM). AR (ϵ=0.5): 66.6 ± 3.5; Full Start: 72.9 ± 5.5; AR (Match Signs): 74.5 ± 4.9; AR: 78.8 ± 5.6; No History: 91.9 ± 5.4.

As shown in Figure 3, all history-aware methods substantially outperform the No History baseline (91.9 ± 5.4), confirming that even highly biased offline data can carry value when the domains partially overlap. However, **AR (ϵ=0.5)** again achieves the best performance (66.6 ± 3.5). Standard AR (78.8 ± 5.6) noticeably underperforms its ϵ-filtered counterpart, consistent with our findings from Experiment 1 that unconstrained replay dilutes useful signals by indiscriminately injecting unrelated samples.

These pathological examples suggest that dynamically gating data reuse via spatial proximity offers a promising alternative to either wholesale pre-training or unrestricted replay. Rather than claiming universal reliability, these findings showcase how spatial filtering can provide a compelling mechanism for conservative adaptation, effectively mitigating negative transfer even under severe distributional mismatch.

5. Conclusion and Next Steps

Integrating historical data into sequential decision-making is a double-edged sword under domain shift. As we have shown, indiscriminately incorporating out-of-domain historical data can delay exploration and trigger negative transfer. However, our findings suggest that the Artificial Replay meta-algorithm can serve as a promising framework when properly constrained. By extending it with spatial closeness heuristics, algorithms can encourage conservative offline-to-online adaptation. Dynamically gating replays based on spatial proximity provides a potential mechanism for models to extract the utility of skewed offline datasets while mitigating their representation bias.

Next Steps. While our ϵ-matching heuristic successfully filters misleading data in our low-dimensional benchmarks, it highlights two key methodological challenges. First, determining the optimal distance threshold ϵ *a priori* without online interaction remains an open problem; future work will explore adaptive thresholding mechanisms that adjust ϵ based on online reward uncertainty. Second, as these methods scale to higher-dimensional context spaces, Euclidean distance may suffer from the curse of dimensionality. We plan to evaluate alternative distance metrics, such as cosine similarity, to preserve the robustness of our spatial filtering. Finally, we intend to validate these conservative adaptation techniques on real-world clinical benchmarks, specifically targeting the IWPC Warfarin dosing dataset, to assess their practical efficacy in mitigating algorithmic bias in precision medicine.

Impact Statement

This research is fundamentally motivated by the need to address algorithmic bias and disparate impact in high-stakes sequential decision-making, such as personalized healthcare and clinical dosing. Historically underrepresented groups frequently suffer when predictive models are trained on skewed or unrepresentative offline datasets. By developing conservative adaptation methods that prevent neural models from over-leveraging biased historical priors, our work provides a technical mechanism to mitigate representation bias. While our current empirical results are simulated, the ultimate goal of this research is to ensure that machine learning systems deployed in social welfare and medical settings adapt safely and equitably to diverse target populations.

References

Banerjee, S., Sinclair, S. R., Tambe, M., Xu, L., and Yu, C. L. Artificial replay: a meta-algorithm for harnessing historical data in bandits, 2022.

Boopathy, A., Liu, K., Hwang, J., Ge, S., Mohammedsaleh, A., and Fiete, I. R. Model-agnostic measure of generalization difficulty. In *International Conference on Machine Learning*, pp. 2857–2884. PMLR, 2023.

Cai, C., Cai, T. T., and Li, H. Transfer learning for contextual multi-armed bandits, 2024.

He, Q., Wang, M., Liu, X., Wang, Z., and Kong, F. Learning across the gap: Hybrid multi-armed bandits with hetero-

geneous offline and online data. In *Advances in Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=kThBNZTMaw.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 110. ACM, 2020. doi: 10.1145/3351095.3373154.

Li, J.-L., Hsu, C.-F., Chang, M.-C., and Chen, W.-C. A comprehensive review of machine learning advances on data change: A cross-field perspective, 2024.

Riquelme, C., Tucker, G., and Snoek, J. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *International Conference on Learning Representations*, 2018.

Shwartz-Ziv, R., Goldblum, M., Li, Y., Bruss, C. B., and Wilson, A. G. Simplifying neural network training under class imbalance. *Advances in Neural Information Processing Systems*, 36:35218–35245, 2023.

Suresh, H. and Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9. ACM, 2021. doi: 10.1145/3465416.3483305.

Yan, H., Qian, Y., Peng, F., Luo, J., Li, F., et al. Neural collapse to multiple centers for imbalanced data. *Advances in Neural Information Processing Systems*, 37: 65583–65617, 2024.

Ye, H.-J., Zhan, D.-C., and Chao, W.-L. Procrustean training for imbalanced deep learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 92–102, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A. Neural Linear Bandit with Artificial Replay

In this section, we detail the formal integration of Artificial Replay within the Neural Linear bandit architecture introduced by Riquelme et al. (2018).

Algorithm 1 Neural Linear Bandit with Artificial Replay (AR)

Require: Historical offline dataset $\mathcal{H}$, online horizon T, arm space K, distance metric $d(\cdot, \cdot)$, closeness threshold ϵ

1: Initialize Neural Network representation layers ϕ_θ
2: Initialize Bayesian linear regression heads (mean β_k, precision Σ_k) for each arm $k \in [K]$
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Observe online context $x_t \in \mathcal{D}_{\text{on}}$
5: Extract feature embedding $z_t \leftarrow \phi_\theta(x_t)$
6: Compute expected rewards $\hat{r}_k = z_t^\top \beta_k$ and select action $a_t \leftarrow \arg\max_k \text{TS}(\hat{r}_k, \Sigma_k)$
▷ **Artificial Replay Phase**
7: Query $\mathcal{H}$ for samples matching action a_t: $\mathcal{H}_{a_t} = \{(x_h, a_h, r_h) \in \mathcal{H} \mid a_h = a_t\}$
8: **if** $\mathcal{H}_{a_t}$ is not empty **then**
9: Find closest matching historical context: $h^* = \arg\min_{h \in \mathcal{H}_{a_t}} d(x_t, x_h)$
10: **if** $d(x_t, x_{h^*}) \leq \epsilon$ **then** ▷ *Enforce spatial ϵ-matching proximity constraint*
11: Extract historical embedding $z_{h^*} \leftarrow \phi_\theta(x_{h^*})$
12: Update Bayesian regression parameters $(\beta_{a_t}, \Sigma_{a_t})$ using (z_{h^*}, r_{h^*})
13: **if** network retrain interval reached **then**
14: Retrain representation layers ϕ_θ on active sample buffer
15: **end if**
16: **goto** Step 5 ▷ *Return to calculate a new feature embedding without taking an online step*
17: **end if**
18: **end if**
▷ **Online Interaction Phase**
19: Execute action a_t in environment and observe online reward r_t
20: Update Bayesian regression parameters $(\beta_{a_t}, \Sigma_{a_t})$ using (z_t, r_t)
21: **if** network retrain interval reached **then**
22: Retrain representation layers ϕ_θ on active sample buffer
23: **end if**
24: **end for**