A Surrogate Policy Model for Auditing Black-Box Recommendation Systems: Application to Change Detection

Marouane Bazzaoui[1], Matthieu Jonckheere[1], Erwan Le Merrer[2], Gilles Trédan[1]

[1] LAAS, CNRS, Toulouse, France
[2] Inria, Rennes, France

marouane.bazzaoui@laas.fr

Abstract

Recommender systems increasingly shape information exposure. As a result, auditing them has become a growing necessity. A key challenge is to understand what can be inferred about a recommender's behaviour from black-box observations alone, i.e., without access to its internals. In this paper, we propose a method to audit recommender systems using a surrogate policy model. This surrogate policy estimator provides a local approximation of the recommender system's behaviour with a characterized approximation error. We establish the consistency and asymptotic normality of this estimator, enabling hypothesis testing. We then propose a change detection task for assessing whether or not the recommender has updated its behaviour.

Keywords

Auditing, recommender systems, surrogate models, inverse reinforcement learning.

1 Introduction

Recommendation systems play a crucial filter role in modern information flows. This role can have a dramatic impact on society [11, 14, 15, 5, 1]. Emerging regulatory frameworks, such as the EU Digital Services Act, increasingly require that such systems be subject to external scrutiny.

Auditing a recommendation system, however, is considerably more challenging than auditing a standard classifier. The space of possible items' states is huge [6]. The content evolves through complex and unknown dynamics, and the items' features are difficult to assess. Moreover, the memory effect from personalization induces a massive trajectory space impossible to explore exhaustively. Finally, external parties typically do not have access to the model's internals.

To address these challenges and provide formal guarantees, we adapt an inverse reinforcement learning (IRL) approach to the recommender auditing problem. In classifier auditing and explainability, the notion of a surrogate model is central. We here explore the construction of a surrogate recommender. We follow a two-step approach : 1) construct a local approximation (a surrogate) of the target model, and 2) conduct tests on the surrogate to obtain an audit decision. See Figure 1 for an overview of the process.

In order to illustrate a potential application of our surrogate construction approach, we consider the *change detection* problem. Change detection is arguably the simplest audit task : observing a black-box model at two distinct time intervals and deciding whether the underlying recommendation mechanism has changed. This task is particularly important for recommender systems, whose decisions meaningfully shape exposure. Detecting changes is especially crucial for sensitive topics, where a shift can signal a change in the system's neutrality or bias across competing viewpoints.

To sum up, this paper provides the following contributions :

- We consider the recommendation system a black box and we formalize a surrogate model. We show that it converges consistently to an approximation of the recommender's behaviour under a few assumptions. We show that the surrogate estimator is asymptotically normally distributed.
- We demonstrate the value of the surrogate estimator and its asymptotic normality via its application to the change detection problem. We establish a statistical test for detecting changes in the black-box model.
- We conduct experiments in a controlled environment to evaluate the effectiveness of our approach in the change detection task.

2 Previous Work

Auditing recommender systems has been an active research area for over a decade. Earlier works frame auditing from a security perspective : a (passive) recommender may be subjected to an external attack, which the audit aims to detect. Several studies have been conducted on detecting manipulations and shilling attacks in the recommendation systems context [8, 7]. While these works introduce methods to detect changes to the recommender's behaviour, they are focused only on item-level anomalies.

Explaining the decisions of a recommender from a black box perspective through local surrogate models is an approach explored in recent works such as LIME-RS [13], LIRE [4], the LIME-RS adherence and constancy study [2]. In [19] a model-agnostic framework was introduced to produce faithful post-hoc explanations. Some works ex-

plain why a particular item was recommended, *e.g.*, through counterfactual explanation methods [3]. Others focus on the recommender's broader behaviour, *e.g.*, through counterfactual audit methods [9]. Overall, this line of work is primarily focused on explainability. Compared to these works, our adapted IRL-based approach comes with formal guarantees in a favorable setting.

3 Problem Formulation

3.1 Black-Box Interaction Model

We model the target recommender as a black box that recommends items to a user from a corpus of n items indexed by $i \in \{1, 2, \dots n\}$. We consider discrete recommendation steps $t \in \{1, 2, \dots T\}$.

Each item is modeled as a discrete-time Markov chain whose state evolves at each step (whether it got recommended or not). This state represents the item features used for recommendation, for instance : number of views, age, or the intrinsic toxicity of the item. At each step, we denote the state of item $i \in \{1, 2, \dots, n\}$ by $S_i \in \mathcal{S}$ and the corpus' state vector by $\mathcal{C} = (S_1, S_2, \dots, S_n)$; the recommender then selects an action vector $A \in \{0, 1\}^n$, where $A_i = 1$ indicates that item i is recommended and $A_i = 0$ otherwise. We assume that exactly m items are recommended at each step. For readability, we omit the time index t throughout, all quantities S_i, $\mathcal{C}$, and A should be understood as referring to a generic time step.

In the reinforcement learning literature, this setting is commonly known as a restless multi-armed bandit, in which the recommender follows a policy π that, at each step, selects which m items to recommend.

In addition, we assume that the observable information of the black-box model is restricted to (i) a representation of the items' states in the form of a features matrix $\Phi(\mathcal{C}) \in \mathbb{R}^{n \times d}$ whose i-th row is the feature vector $\phi_i(S_i) \in \mathbb{R}^d$ of item i; and (ii) the agent's action $A \in \{0, 1\}^n$.

The model is observed at a time interval, during which we observe N distinct trajectories, of length T. We denote the trajectories by

$$\mathcal{T}_k := (\Phi_{k,t}, A_{k,t})_{t=1}^{T}, \quad k \in \{1, \dots N\}.$$

We therefore define the trajectories' set

$$\mathbb{T} = \{\mathcal{T}_k\}_{k=1}^{N}.$$

We write

$$\mathcal{T}_k \sim P_T$$

for some law P_T on $\left(\mathbb{R}^{n \times d} \times \{0, 1\}^n\right)^T$.

3.2 Surrogate Policy Model

In a recommender system, the decision to recommend an item depends not only on the item itself, but also on the other items in the corpus. Externally auditing such a system becomes challenging when considering all factors at play. To overcome this, we introduce a parametric surrogate model that bases its decisions on a scoring function, assigning higher recommendation probabilities to items with higher scores. Our approach consists in transferring the essential decision logic of the recommender to a simple, interpretable surrogate model with well-characterized asymptotic behaviour that enables formal hypothesis testing.

We introduce $I : \mathcal{S} \to \mathbb{R}$ an index function defined as a function that scores an item depending entirely on its state $S \in \mathcal{S}$. Let X_I denote an index vector, and π_I denote a soft-index-policy, such that

$$X_I(\mathcal{C}) := (0, I(S_2) - I(S_1), \dots I(S_n) - I(S_1))$$

$$\text{and} \quad \pi_I = \mathrm{st}_m(X_I)$$

where $\mathcal{C} = (S_1, \dots S_n)$, and $\mathrm{st}_m : \mathbb{R}^n \to (0, 1)^n$ denotes the soft-top-m function, a smooth approximation of the standard top-m operator. It is derived from the optimal plan of an entropic transport problem [18, 10].

By restricting the index function I to a parametric family of functions, we can define a parametric surrogate model based on the soft index policy π_I. By defining the surrogate based on a smooth approximation rather than the top-m operator itself, the model's output becomes differentiable. This enables gradient-based optimization and asymptotic analysis. Because the surrogate policy is built on the soft-top-m function [18], it is non-deterministic, and it recommends m items in expectation.

In this work, inspired by LIME [16], we consider linear models for their efficiency as well as their intrinsic interpretability. We restrict the index I to be linear in $\phi_S \in \mathbb{R}^d$, where ϕ_S denotes the features of an item whose state is $S \in \mathcal{S}$, namely

$$I_\theta(S) = \theta^\top \phi_S.$$

Hence, rather than considering an arbitrary index I, we restrict ourselves to the linear family induced by this parametrization.

Let $X_\theta = X_{\hat{I}_\theta}$, the linear surrogate model is then defined by

$$\hat{\pi}_\theta = \mathrm{st}_m(X_\theta).$$

The purpose of the surrogate model is to approximate the behaviour of the recommender's policy. A good surrogate should therefore reproduce the decision patterns of the black-box model with a controlled approximation error, while remaining simple enough to be interpretable.

4 Surrogate Policy Estimator

In order to approximate the recommender's policy, we use the maximum-likelihood approach. We first establish that the surrogate model class can asymptotically reproduce any index-policy. We then introduce an empirical estimator used to fit the surrogate from data, and we establish its consistency and asymptotic normality.

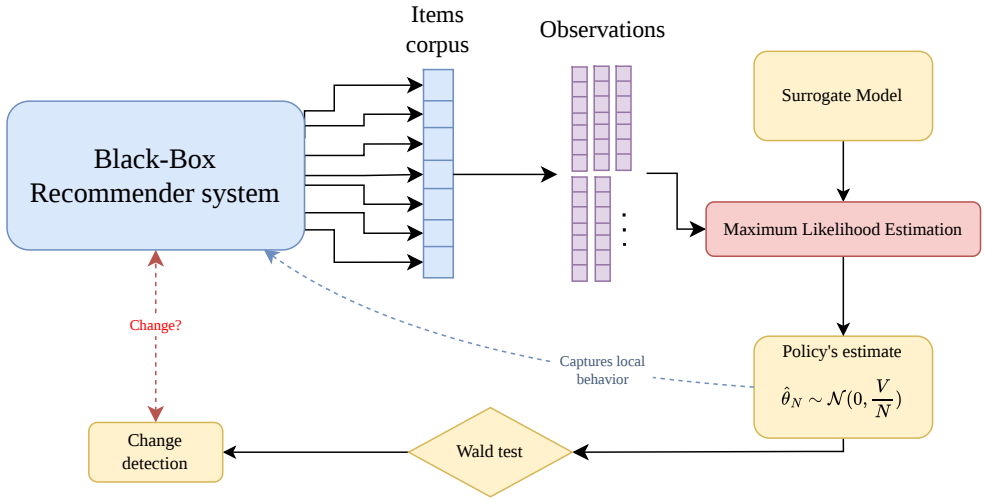

FIGURE 1 – Overview of the surrogate model auditing framework with illustrative downstream application.

4.1 Surrogate Model Asymptotic Compatibility

Under the assumption that the recommendation system is optimized with respect to some unknown objective, its policy π is deterministic, since any unconstrained MDP admits a deterministic optimal policy. We also assume that the recommender policy π admits an index representation, *i.e.*,

$$\exists I \in \mathcal{F}(\mathcal{S}, \mathbb{R}), \;\; \forall \mathcal{C} \in \mathcal{S}^n \;\; ; \;\; \pi(\mathcal{C}) = \underset{i}{\text{top-m}} \; I(S_i)$$

where $I : \mathcal{S} \to \mathbb{R}$ denotes an index, and $\mathcal{F}(\mathcal{S}, \mathbb{R})$ denotes the set of functions from $\mathcal{S}$ to $\mathbb{R}$.
We refer to any policy π that admits an index representation as an index-policy. We define the set of all index-policies Π as follows

$$\Pi = \left\{ \pi \;\middle|\; \begin{array}{l} \pi \in \mathcal{F}(\mathcal{S}^n, \{0,1\}^n), \\ \exists I \in \mathcal{F}(\mathcal{S}, \mathbb{R}) \; ; \;\; \pi(\mathcal{C}) = \text{top-m}_i \; I(S_i) \end{array} \right\}.$$

where $\mathcal{F}(\mathcal{S}^n, \{0,1\}^n)$ denotes the set of functions from $\mathcal{S}^n$ to $\{0,1\}^n$, and $\mathcal{F}(\mathcal{S}, \mathbb{R})$ denotes the set of functions from $\mathcal{S}$ to $\mathbb{R}$.

Theorem 1. *Let π denote a recommender index-policy and I its index, π_I denote a soft-index-policy with the same index I as the recommender, $\mathcal{C}$ denote the corpus' state vector, and $X_I(\mathcal{C}) = (0, I(S_2) - I(S_1), \dots I(S_n) - I(S_1))$ denote the index vector. Then*

$$\|\pi(\mathcal{C}) - \pi_I(\mathcal{C})\|_2 \leq \frac{\varepsilon(\ln n + \ln 2)}{X_I(S_{\sigma_{m+1}}) - X_I(S_{\sigma_m})}$$

where $\epsilon > 0$ is a parameter of the soft-top-m function [18], and σ the sorting permutation of the vector X_I such that $X_I(S_{\sigma_m})$ and $X_I(S_{\sigma_{m+1}})$ are the m-th and $(m+1)$-th highest values in the vector X_I.

The following theorem establishes that the surrogate model, despite using a smooth relaxation, can approximate any index-policy arbitrarily well as the index values grow large.

Theorem 2. *Let Π be the set of all index policies. Then for $\pi \in \Pi$ an arbitrary index-policy, there exist a sequence of indexes $(I_k) \subset \mathcal{F}(\mathcal{S}, \mathbb{R})$ such that*

$$\forall \mathcal{C} \in \mathcal{S}^n \;\; \pi_{I_k}(\mathcal{C}) \underset{k}{\to} \pi(\mathcal{C})$$

$$\textit{and} \quad \forall \mathcal{C} \in \mathcal{S}^n \;\; \|I_k(\mathcal{C})\| \underset{k}{\to} +\infty$$

where $\pi_{I_k} = \text{st}_m(X_{I_k})$ a soft-index-policy, and $I_k(\mathcal{C})$ denotes a vector such that $I_k(\mathcal{C}) = (I_k(S_1), \dots I_k(S_n))$.

Theorem 1 establishes the approximation error of the surrogate policy model.
Theorem 2 implies that for any given recommender index policy, there exists a sequence of surrogate policies (I_k) realising an asymptotic approximation.

4.2 Maximum Likelihood Estimator

In this section, we define the surrogate policy estimator. We proceed as follows : first, we formalize the population's negative log-Likelihood risk, then introduce an L_2 penalty, derive its empirical counterpart, and finally define the resulting surrogate empirical estimator.
In order to have a well-defined population risk, we assume that the second moments of the feature matrices Φ_t in trajectories $\mathcal{T}$ are finite, *i.e.*, $\mathbb{E}\|\Phi_t\|_2^2 < \infty$ for all $t \in \{1, \dots T\}$. This condition prevents feature values from taking excessively large values too often, and is trivially satisfied when features are bounded.
Let $\hat{\pi}_\theta$ denote the policy defined by the surrogate model parameterized by θ. We introduce the population risk L as follows

$$L(\theta) := \mathbb{E}_{\mathcal{T}\sim P_T}\left[\sum_{(\Phi,A)\in\mathcal{T}} -\log P_\theta(A \mid \Phi)\right]$$

where $P_\theta(A \mid \Phi)$ denotes the conditional probability of the action A being played by $\hat{\pi}_\theta$ knowing Φ.
As established in Theorem 2, the minimizer of L lies on the infinite boundaries of the parameter space $\theta \in \mathbb{R}^d$, which motivates penalizing the risk. Let L^λ denote the penalized population criterion, defined by

$$L^\lambda(\theta) := L(\theta) + \lambda\|\theta\|_2^2$$

where $\lambda > 0$ denotes the regularization parameter.
The regularized population criterion L^λ is twice differentiable and strongly convex in $\theta \in \mathbb{R}^d$ and therefore admits a unique minimizer $\theta^* \in \mathbb{R}^d$,

$$\theta^* := \underset{\theta\in\mathbb{R}^d}{\arg\min}\ L^\lambda(\theta).$$

Let $\hat{L}_N^\lambda$ denote the empirical counterpart of the penalized population criterion $L^\lambda(\theta)$ on a sample of trajectories $\mathbb{T} = (\mathcal{T}_i)_{i=1}^N$. We assume the independence of these trajectories

$$\mathcal{T}_1, \mathcal{T}_2, \dots \mathcal{T}_N \overset{\text{iid}}{\sim} P_T.$$

We denote the empirical target as $\hat{\theta}_N$, then

$$\hat{\theta}_N := \underset{\theta\in\mathbb{R}^d}{\arg\min}\ \hat{L}_N^\lambda(\theta).$$

In this work, we focus primarily on the penalized estimator since the unpenalized criterion may fail to admit a finite minimizer. Therefore, from this point onward, we implicitly consider the penalized version of quantities (criterion, target) unless stated otherwise.

4.3 Asymptotic Normality

We now explicitly characterize the convergence of our surrogate estimator. First, we establish in Theorem 3 the consistency of the estimator. We then derive the asymptotic distribution using a standard Taylor expansion argument. We include key steps of the derivation that introduce the Hessian H and the covariance Σ of the score vector. Defining these two terms is essential because they appear in the asymptotic normal distribution result.

Theorem 3. *Let θ^* denote the population target parameter of the surrogate policy estimator, and let $\hat{\theta}_N$ denote the empirical counterpart from a sample of N trajectories. Then*

$$\hat{\theta}_N \xrightarrow[N\to+\infty]{p} \theta^*.$$

*Implying that the surrogate policy estimator is **consistent**.*

We write the empirical estimator as

$$\hat{\theta}_N = \underset{\theta\in\mathbb{R}^d}{\arg\min}\ \hat{L}_N^\lambda(\theta), \quad \hat{L}_N^\lambda(\theta) = \frac{1}{N}\sum_{k=1}^N l^\lambda(\mathcal{T}_k, \theta)$$

where $l^\lambda(\mathcal{T}_k, \theta)$ denotes the instantaneous loss at the trajectory $\mathcal{T}_k$, and defined by

$$l^\lambda(\mathcal{T}_k, \theta) = \left[\sum_{(\Phi,A)\in\mathcal{T}} -\log P_\theta(A \mid \Phi)\right] + \lambda\|\theta\|_2^2.$$

Let $\Psi_N : \mathbb{R}^d \to \mathbb{R}^d$ denote the score function defined by

$$\Psi_N(\theta) = \nabla_\theta \hat{L}_N^\lambda(\theta).$$

Since $\hat{L}_N^\lambda$ is twice differentiable, the score function Ψ_N and its Jacobian $\nabla\Psi_N$ are well defined on $\mathbb{R}^d$.
Using the fact that $\Psi_N(\hat{\theta}_N) = 0$ together with a Taylor expansion of Ψ_N around θ^*, we get

$$\sqrt{N}\left(\hat{\theta}_N - \theta^*\right) = -\left(\nabla\Psi_N(\tilde{\theta}_N)\right)^{-1}\sqrt{N}\Psi_N(\theta^*) \quad (1)$$

where $\tilde{\theta}_N \in \mathbb{R}^d$ lies between $\hat{\theta}_N$ and θ^*, *i.e.*, $\tilde{\theta}_N$ is a convex combination of $\hat{\theta}_N$ and θ^*.
Because the Hessian of $\hat{L}_N^\lambda$ is positive definite, and therefore invertible, the matrix $\left(\nabla\Psi_N(\tilde{\theta}_N)\right)^{-1}$ is well defined.
Let $H \in \mathbb{R}^{d\times d}$ denote the Hessian of L^λ at θ^*, then by LLN we get

$$\left(\nabla\Psi_N(\tilde{\theta}_N)\right)^{-1} \xrightarrow[N\to+\infty]{p} H^{-1}. \quad (2)$$

Computational complexity. The Hessian H can be computed in $O(nd^2)$ time and $O(nd)$ memory, *i.e.*, linearly in the catalog size n.
Let $\Sigma \in \mathbb{R}^{d\times d}$ be the covariance matrix of $\nabla l^\lambda(\mathcal{T}, \theta^*)$, then by CLT we get

$$\sqrt{N}\Psi_N(\theta^*) \xrightarrow[N\to+\infty]{d} \mathcal{N}(0, \Sigma). \quad (3)$$

Theorem 4. *Let θ^* denote the population target parameter of the surrogate policy estimator, and let $\hat{\theta}_N$ denote the empirical counterpart from a sample of N trajectories. Using (1), (2), and (3) together with Slutsky's theorem, we get*

$$\sqrt{N}\left(\hat{\theta}_N - \theta^*\right) \xrightarrow[N\to+\infty]{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$$

where $\Sigma = \mathrm{Var}(\nabla l^\lambda(\mathcal{T}, \theta^*))$ and $H = \nabla^2 L^\lambda(\theta^*)$.

5 Change Detection Problem

Although surrogate policy estimation is our central contribution, its value is best demonstrated through a concrete downstream task. We use change detection as such a demonstration.
The model is observed at two non-overlapping episodes indexed by $e \in \{1, 2\}$. During both episodes, we observe N distinct trajectories, each of length T, each observed in a different corpus of items. In this setting, we test whether the policy of the agent changed between episode 1 and episode 2 or remained unchanged.

Let π_1 and π_2 be the agent's policies at episode 1 and episode 2, respectively. The change detection problem consists in deciding whether π_1 is different from π_2 based only on observable information : $\mathbb{T}_1$ and $\mathbb{T}_2$. Formalized as a hypothesis test, it can be written as

$$H_0' : \pi_1 = \pi_2, \quad H_1' : \pi_1 \neq \pi_2.$$

Our approach to this problem is to use the surrogate policy model, to estimate the agent's policies in the two episodes, then compare the two resulting policy estimates. Let θ_e^* be the population target parameter vector of the surrogate policy estimator for episode e. We define the following hypothesis test

$$H_0 : \theta_1^* = \theta_2^*, \quad H_1 : \theta_1^* \neq \theta_2^*.$$

Theorem 5. *Let θ_e^* be the population target parameter vector of the surrogate policy estimator for episode e. Let $\pi_e = \text{top-m}(X_e)$ be the true agent's policy at episode e, and X_e the underlying index vector. Then*

$$\pi_1 = \pi_2 \implies \theta_1^* = \theta_2^*.$$

By Theorem 5, we obtain $H_0' \implies H_0$. Therefore, if we refute H_0 we can refute the null hypothesis H_0', and consequently detect a change in the agent's policy.

5.1 Wald Test

Let $\hat{\Delta} := \hat{\theta}_2 - \hat{\theta}_1$ denote the difference in the estimators between episode 1 and episode 2. Then $\hat{\Delta}$ is asymptotically normal, and under H_0,

$$\hat{\Delta} \sim \mathcal{N}\left(0, 2\,\frac{V}{N}\right)$$

where N denotes the number of trajectories per episode. and $V = H^{-1}\Sigma H^{-1}$ denotes the asymptotic variance common to both episodes under H_0.
Let $\hat{V}_e$ denote a consistent estimator of the asymptotic variance at episode e. Let W denote the Wald test statistic defined by

$$W := \hat{\Delta}^\top \left(\frac{\hat{V}_1 + \hat{V}_2}{N}\right)^{-1} \hat{\Delta}.$$

Under H_0,

$$W \xrightarrow{d} \chi_d^2$$

where d denotes the dimension of the parameters' space. Therefore, for a given significance level $\alpha \in (0, 1)$, we reject the null-hypotheses H_0, and consequently $H_0' : \pi_1 = \pi_2$, when

$$W > q_{1-\alpha}$$

where $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ_d^2 distribution.

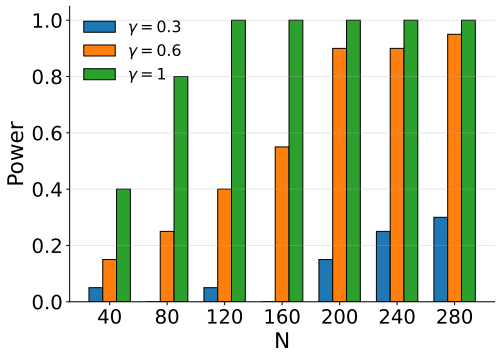

FIGURE 2 – Empirical power $(1-\beta)$ as a function of the sample size N for three separation values γ.

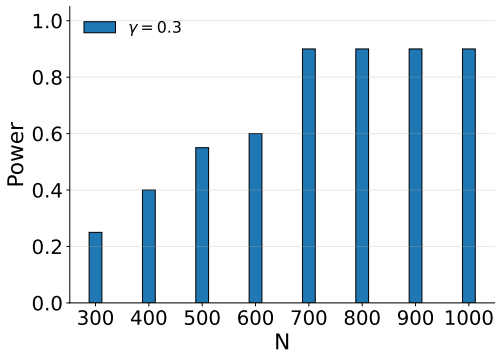

FIGURE 3 – Empirical power $(1-\beta)$ as a function of the sample size N for a small separation value $\gamma = 0.3$.

6 Experimental Study

While previous results characterize the asymptotic correctness of our approach, to be useful in practice an estimator must provide accurate results in the finite-sample regime. To that end, we evaluate change detection on synthetic data using a controlled simulated environment, which allows us to test our method's performance across a range of settings.

Experimental setup. We simulate the behaviour of a recommender by a linear index-policy. We observe its decisions across different sets of items, each of size $n = 10$. Its interactions with each set of items yield a trajectory over $T = 5$ consecutive recommendation steps. Each item is characterized by three features : popularity, toxicity, and profitability. The popularity value evolves in time while toxicity and profitability stay unchanged, but all three are randomly initialized. We conduct the change detection test between two generated trajectory sets, one by a *toxic* policy, and the other by a *neutral* policy; both use the same weights for popularity and profitability, $\theta_{pop} = 10$ and $\theta_{prf} = 5$. The toxic policy additionally assigns a positive weight $\theta_{tox} = \gamma$ to toxicity, whereas the neutral policy assigns a null weight to toxicity $\theta_{tox} = 0$. Both simulated policies prefer recommending popular and profitable items, but the *toxic* recommender has systematic preferences for toxic items, while the *neutral* recommender is indifferent

to toxicity.
In this setup, the change detection hypothesis test, introduced in Section 5, is formulated as

$$H_0 : \theta^*_{tox} = \theta^*_{neut} \ , \quad H_1 : \theta^*_{tox} \neq \theta^*_{neut}$$

where θ^*_{tox} and θ^*_{neut} denote the surrogate model's estimations of the weights vector of the *toxic* recommender and the *neutral* recommender, respectively. We set the confidence level in our experiments at $(1 - \alpha) = 95\%$.
The primary question in this section is the trade-off between (i) cost, *i.e.*, the number of observed trajectories required to detect a change ; (ii) separation, *i.e.*, the distance between the two tested policies ; (iii) reliability, *i.e.*, the ability to detect a change when one occurs. We quantify the trade-off between these three axes by the following metrics :

- **Cost :** N, the number of trajectories generated under each policy.
- **Separation :**γ, the difference in the toxicity weights between the two policies.
- **Reliability :** $1 - \beta$ (power), the probability of detecting a change when one occurs.

The estimates are obtained via a gradient descent algorithm by minimizing the empirical criterion.
We vary the cost and separation parameters in a grid where N takes the values $\{40, 80, 120, 160, 200, 240, 280\}$, and γ takes the values $\{0.3, 0.6, 1.0\}$. We repeat the test 20 times for each pair (N, γ), and we measure the empirical power $1 - \beta$. See Figure 2.
In order to investigate the trade-off between the cost and the reliability in the task of detecting a small change $\gamma = 0.3$, we evaluate the power $1 - \beta$ at a bigger scale of $N \in \{300, 400, 500, 600, 700, 800, 900, 1000\}$. See Figure 3.

7 Discussion

Limitations of the modeling. While our black-box modeling choices simplify reality, we argue that they remain reasonably aligned with it. First, assuming that the recommender's policy is index-based is not overly restrictive, since in this type of sequential allocation problem the optimal policy admits an index-based representation under mild conditions [17]. Second, a linear surrogate may not adequately represent abrupt policy shifts or strongly non-linear dynamics ; extending the framework to non-linear surrogate models lies outside the scope of this paper. Finally, the i.i.d. trajectory assumption can be justified in practice by collecting each trajectory from a different item set, or by separating trajectories sufficiently to eliminate temporal dependence.

Feature observability. The items' states in the present work are represented by features that we assume capture all the information relevant to the recommender. However, it is usually impossible for an external party to observe that representation in its entirety. In practice, auditors can observe a subset of the features. In that case, a change in the recommender's preferences with respect to the observed features can be detected, while a change orthogonal to the observed information is invisible to the test. Moreover, unobserved features that correlate with observed ones can cause omitted-variable bias in our estimates. The estimated coefficients thus represent the recommender's behavior projected onto the observed features space, which is sufficient for auditing a change, though not for causal interpretation of the features' effects.

Beyond change detection. In the present work, we construct a surrogate model that locally approximates a recommender's policy and that has statistical characteristics that facilitate rigorous analysis. While we focus on the change detection task in this work, the proposed framework addresses a broader range of questions that we leave for future work : What is the system optimizing ? Are there viewpoint biases ? Can the system be manipulated ?

8 Conclusion

We proposed a framework for auditing black-box recommender systems using a surrogate policy model. We derived a simple surrogate policy estimator that provides a local approximation of the recommender system's behaviour with a characterized approximation error. We established consistency and asymptotic normality of the surrogate policy estimator, enabling hypothesis testing and further analysis of the local approximation. We demonstrated the auditing capability of the framework through a change detection task, both theoretically and experimentally in a controlled simulation environment.
This work opens several directions for future research. Applying the change detection method to real-world recommender systems is an important next step. Moreover, the proposed surrogate policy model can serve as a foundation for a broader range of auditing tasks, such as bias identification and fairness assessment.

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv :1907.13286*, 2019.

[2] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Francesco Maria Donini, Vincenzo Paparella, and Claudio Pomo. Adherence and constancy in lime-rs explanations for recommendation. *arXiv preprint arXiv :2109.00818*, 2021.

[3] Oren Barkan, Veronika Bogina, Liya Gurevitch, Yuval Asher, and Noam Koenigstein. A counterfactual framework for learning and evaluating explanations for recommender systems. In *Proceedings of the ACM Web Conference 2024*, pages 3723–3733, 2024.

[4] Léo Brunot, Nicolas Canovas, Alexandre Chanson, Nicolas Labroche, and Willème Verdeaux. Preference-based and local post-hoc explanations for recommender systems. *Information Systems*, 108 :102021, 2022.

[5] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and

decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.

[6] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[7] Min Gao, Renli Tian, Junhao Wen, Qingyu Xiong, Bin Ling, and Linda Yang. Item anomaly detection based on dynamic partition for time series in recommender systems. *PloS one*, 10(8) :e0135155, 2015.

[8] Min Gao, Quan Yuan, Bin Ling, and Qingyu Xiong. Detection of abnormal item based on time intervals for recommender systems. *The Scientific World Journal*, 2014(1) :845897, 2014.

[9] Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J Watts. Causally estimating the effect of youtube's recommender system using counterfactual bots. *Proceedings of the national academy of sciences*, 121(8) :e2313377121, 2024.

[10] Gauri Jain, Pradeep Varakantham, Haifeng Xu, Aparna Taneja, Prashant Doshi, and Milind Tambe. Irl for restless multi-armed bandits with applications in maternal and child health. In *Pacific Rim International Conference on Artificial Intelligence*, pages 165–178. Springer, 2024.

[11] Erwan Le Merrer, Gilles Trédan, and Ali Yesilkanat. Modeling rabbit-holes on youtube. *Social network analysis and mining*, 13(1) :100, 2023.

[12] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4 :2111–2245, 1994.

[13] Caio Nóbrega and Leandro Marinho. Towards explaining recommendations through local surrogate models. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, pages 1671–1678, 2019.

[14] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. Down the (white) rabbit hole : The extreme right and online recommender systems. *Social Science Computer Review*, 33(4) :459–478, 2015.

[15] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you ?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[17] Peter Whittle. Restless bandits : Activity allocation in a changing world. *Journal of applied probability*, 25(A) :287–298, 1988.

[18] Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, and Tomas Pfister. Differentiable top-k with optimal transport. *Advances in neural information processing systems*, 33 :20520–20531, 2020.

[19] Zhichao Xu, Hansi Zeng, Juntao Tan, Zuohui Fu, Yongfeng Zhang, and Qingyao Ai. A reusable model-agnostic framework for faithfully explainable recommendation and system scrutability. *ACM Transactions on Information Systems*, 42(1) :1–29, 2023.

A Proofs

Proof of Theorem 1. Let π denote a recommender index-policy and I its index, π_I denote a soft-index-policy with the same index I as the recommender, $\mathcal{C}$ denote the vector of the items' state, and $X_I(\mathcal{C}) = (0, I(S_2) - I(S_1), \dots I(S_n) - I(S_1))$ denote the index vector.
We show that

$$\|\pi(\mathcal{C}) - \pi_I(\mathcal{C})\|_2 \leq \frac{\varepsilon(\ln n + \ln 2)}{X_I(S_{\sigma_{m+1}}) - X_I(S_{\sigma_m})}$$

where $\epsilon > 0$ is a parameter of the soft-top-m function [18], and σ the sorting permutation of the vector X_I.
For simplicity we denote $X = X_I(\mathcal{C})$
The concerned theorem, is a direct consequence of a result in [18] (Theorem 2). It states the following

$$\|\Gamma_{soft}(X) - \Gamma(X)\|_F \leq \frac{\varepsilon(\ln n + \ln 2)}{n\left(X_I(S_{\sigma_{m+1}}) - X_I(S_{\sigma_m})\right)} \tag{1}$$

where Γ_{soft} and Γ are defined by (we suppress the X notation for simplicity)

$$\Gamma_{soft} = \arg\min_{\Gamma_{soft}} \langle \Gamma_{soft}, C\rangle + \epsilon H(\Gamma_{soft}),$$
$$\text{s.t } \Gamma_{soft} 1_2 = u,\ \Gamma_{soft} 1_n = v$$

and

$$\Gamma = \arg\min_{\Gamma} \langle \Gamma, C\rangle,$$
$$\text{s.t } \Gamma 1_2 = u,\ \Gamma 1_n = v$$

where $\epsilon > 0$, $u = (\frac{1}{n}, \dots \frac{1}{n}) \in \mathbb{R}^n$, $v = (\frac{n-k}{n}, \frac{k}{n})$, and $C \in \mathbb{R}^{n\times 2}$ such that $C_{i1} = x_i^2$ and $C_{i2} = (x_i - 1)^2$.
By [18] we have

$$\pi(\mathcal{C}) = \text{top-m}(X) = n\Gamma(X)\ [0,1]^\top$$

and

$$\pi_I(\mathcal{C}) = \text{st}_m(X) = n\Gamma_{soft}(X)\ [0,1]^\top$$

Then

$$\|\pi_I(\mathcal{C}) - \pi(\mathcal{C})\|_2 \leq n\,\|\Gamma_{soft}(X) - \Gamma(X)\|_F \tag{2}$$

By (1) and (2) we conclude

$$\|\pi_I(\mathcal{C}) - \pi(\mathcal{C})\|_2 \leq \frac{\varepsilon(\ln n + \ln 2)}{X_I(S_{\sigma_{m+1}}) - X_I(S_{\sigma_m})}$$

□

Proof of Theorem 2. We show that for an arbitrary index-policy $\pi \in \Pi$ the following

$$\exists (I_k) \subset \mathcal{F}(\mathcal{S}, \mathbb{R})\ \ \forall \mathcal{C} \in \mathcal{S}^n\ \ \pi_{I_k}(\mathcal{C}) \underset{k}{\to} \pi(\mathcal{C})$$

where $\pi_{I_k} = \text{st}_m(X_{I_k})$, and $I_k(\mathcal{C})$ denotes a vector such that $I_k(\mathcal{C}) = (I_k(S_1), \dots I_k(S_n))$.
For simplification let $X_{I_k} = X_k$.
the policy π admits an index-based structure, let I be its index. Take a sequence of indexes $I_k = kI$ defined as the recommender's index times k. Since X_k is defined by

$$X_k = (0, I_k(S_2) - I_k(S_1), \dots I_k(S_n) - I_k(S_1))$$

then $X_k = kX$ where

$$X = (0, I(S_2) - I(S_1), \dots I(S_n) - I(S_1))$$

The recommender's policy $\pi = \text{top-m}\, X$ depends only on the ranking of X, then it can also be written as $\pi = \text{top-m}\, kX$.
Then we can consider π and π_{I_k} to have the same index I_k. Hence, by Theorem 1, we have

$$\|\pi_{I_k}(\mathcal{C}) - \pi(\mathcal{C})\|_2 \leq \frac{\varepsilon(\ln n + \ln 2)}{kX(S_{\sigma_{m+1}}) - kX(S_{\sigma_m})}$$

where σ denotes a sorting permutation of X.
We have

$$\frac{\varepsilon(\ln n + \ln 2)}{k\left(X(S_{\sigma_{m+1}}) - X(S_{\sigma_m})\right)} \underset{k}{\to} 0$$

Then

$$\pi_{I_k}(\mathcal{C}) \underset{k}{\to} \pi(\mathcal{C})$$

Now we show that if π_{I_k} converges to π, then

$$\forall \mathcal{C} \in \mathcal{S}^n\ \ \|I_k\| \to \infty.$$

We assume that $I_k(\mathcal{C})$ does not diverge to infinity and we show a contradiction.
Since $I_k(\mathcal{C})$ does not diverge to infinity, it is bounded and By Bolzano–Weierstrass, it has a further sub-sequence

$$I_{k_j}(\mathcal{C}) \to J \in \mathbb{R}^n.$$

Because st_m is continuous : $\pi_{I_{k_j}}(\mathcal{C}) \to \pi_J(\mathcal{C})$.
Hence $\pi(\mathcal{C}) = \pi_J(\mathcal{C})$, a contradiction because $\pi(\mathcal{C})$ is a deterministic policy and $\pi_J(\mathcal{C})$ is a non-deterministic policy knowing J finite.

□

Proof of Theorem 3. Let θ^* denote the population target parameter of the surrogate policy estimator, and let $\hat{\theta}_N$ denote the empirical counterpart based on a sample of N trajectories. We show that the estimator is consistent, that is,

$$\hat{\theta}_N \underset{N\to\infty}{\xrightarrow{p}} \theta^*.$$

We begin by formulating the gradient and the Hessian of one component of the instantaneous loss (at one state-action pair instead of the whole trajectory). For simplicity, we denote this component by l, without explicitly writing the state or time indices. We also denote by π the value of the

policy at a given state, viewed as a vector, and by π_i its coordinates. Let X denote the index vector, and let x_i be its i-th coordinate. We have

$$\pi = \mathrm{st}_m(X).$$

Let $\bar{R}$ and R denote a partition of $\{1, 2, \dots, n\}$, where R is the set of recommended indices and $\bar{R}$ is the set of non-recommended ones. Then

$$l = -\sum_{i \in R} \log(\pi_i) - \sum_{i \in \bar{R}} \log(1 - \pi_i).$$

The goal is to find ∇l and $\nabla^2 l$. Then, by linearity of all the criteria, we can derive any other gradient or Hessian. Since the soft-index policies are defined as the minimizers of entropic optimal transport problems [18], we know that they are twice differentiable. This implies that all the criteria based on such policies are also twice differentiable.
The KKT formula for the entropic OT problem in [18] (See Proof of Theorem 1)

$$\Gamma_{ij} = \exp\left(\frac{f_i + g_j - C_{ij}}{\epsilon}\right), \qquad \forall i, j,$$

where f and g are the Lagrange multipliers.
We consider a small perturbation :

$$\Gamma_{ij} \to \Gamma_{ij} + \delta\Gamma_{ij}, \qquad x_i \to x_i + \delta x_i,$$

$$f_i \to f_i + \delta f_i, \qquad g_j \to g_j + \delta g_j.$$

We then obtain

$$\delta\Gamma_{ij} = \Gamma_{ij}\,\frac{\delta f_i + \delta g_j - \delta C_{ij}}{\epsilon}, \qquad \forall i, j.$$

Since Γ has constant marginals, we have

$$\sum_i \delta\Gamma_{ij} = 0, \qquad \sum_j \delta\Gamma_{ij} = 0.$$

After some algebra, we obtain

$$\begin{cases} u_i\,\delta f_i + \sum_j \Gamma_{ij}\,\delta g_j = a_i, \\ \sum_i \Gamma_{ij}\,\delta f_i + v_j\,\delta g_j = b_j, \end{cases}$$

where u and v denote the marginals of Γ. The terms a_i and b_j are given by

$$a_i = 2\Gamma_{i1} x_i\,\delta x_i + 2\Gamma_{i2}(x_i - 1)\,\delta x_i,$$

$$b_1 = 2\sum_i \Gamma_{i1} x_i\,\delta x_i,$$

$$b_2 = 2\sum_i \Gamma_{i2}(x_i - 1)\,\delta x_i.$$

Knowing that $\pi_i = n\Gamma_{i2}$, and after some algebra, we get

$$\delta\pi_i = \frac{2}{\epsilon}\pi_i(1 - \pi_i)\left(\delta x_i - \frac{\sum_r \pi_r(1 - \pi_r)\delta x_r}{W}\right),$$

where

$$W = \sum_r \pi_r(1 - \pi_r).$$

Therefore,

$$\frac{\partial \pi_i}{\partial x_j} = \begin{cases} -\dfrac{2}{\epsilon W}(1 - \pi_i)\pi_j(1 - \pi_j), & \text{if } i \neq j, \\ \dfrac{2}{\epsilon}(1 - \pi_i)\left(1 - \dfrac{\pi_i(1 - \pi_i)}{W}\right), & \text{if } i = j. \end{cases}$$

Then, after some algebra, we get the gradient of l with respect to X :

$$\nabla_{X,i} l = \begin{cases} \dfrac{2}{\epsilon}(\pi_i - 1), & \text{if } i \in R, \\ \dfrac{2}{\epsilon}\pi_i, & \text{if } i \in \bar{R}. \end{cases}$$

Let $M = I_n - M'$, where I_n is the identity matrix and M' is the matrix whose entries are all zero except on the first column, which is filled with ones. Then, by definition,

$$X = M\,\Phi\,\theta,$$

where Φ denotes the feature matrix and θ denotes the parameter of the index.
Hence, the gradient of l with respect to the parameter θ is

$$\begin{aligned} \nabla l &= \Phi^\top M^\top \nabla_X l \\ &= \Phi^\top \nabla_X l. \end{aligned}$$

Using the expressions of $\nabla_X l$ and $\frac{\partial \pi_i}{\partial x_j}$, we obtain the Hessian $H^{(X)}$ of l with respect to X :

$$H^{(X)}_{ij} = \begin{cases} \dfrac{4}{\epsilon^2 W}\pi_i(1 - \pi_i)\big(W - \pi_i(1 - \pi_i)\big), & \text{if } i = j, \\ -\dfrac{4}{\epsilon^2 W}\pi_i(1 - \pi_i)\pi_j(1 - \pi_j), & \text{if } i \neq j. \end{cases}$$

We also obtain the Hessian $H^{(\theta)}$ of l with respect to θ :

$$H^{(\theta)} = \Phi^\top M^\top H^{(X)} M \Phi.$$

Notice that

$$H^{(X)}_{ii} = \sum_{j \neq i} \left|H^{(X)}_{ij}\right| \qquad \text{for every } i \in \{1, \dots, n\}.$$

Hence, $H^{(X)}$ is positive semi-definite, and consequently $H^{(\theta)}$ is also positive semi-definite. By linearity of the expectation and of the summation operator, we conclude that the unpenalized risks $L(\theta)$ and $\hat{L}_N(\theta)$ are convex. By adding the ℓ_2-penalization term, the population criterion $L^\lambda(\theta)$ and its empirical counterpart $\hat{L}^\lambda_N(\theta)$ become 2λ-strongly convex (*i.e.*, the Hessian matrices of both terms are $\succcurlyeq 2\lambda I_d$ where I_d denotes the identity matrix). In particular, L^λ admits a unique minimizer θ^*.
Since the trajectories $T_1, \dots, T_N$ are i.i.d., the law of large numbers gives, for any fixed $\theta \in \mathbb{R}^d$,

$$\hat{L}^\lambda_N(\theta) \xrightarrow[N \to \infty]{p} L^\lambda(\theta).$$

Since $\hat{L}_N^\lambda$ is convex, L^λ has the unique minimizer θ^*, and $\hat{L}_N^\lambda(\theta)$ converges pointwise to $L^\lambda(\theta)$ for every $\theta \in \mathbb{R}^d$, consistency follows from Newey and McFadden [12] (Theorem 2.7) :

$$\hat{\theta}_N \xrightarrow[N\to\infty]{p} \theta^*.$$

Hence, the surrogate policy estimator is consistent.

□

Proof of Theorem 4. Let θ^* denote the population target parameter of the surrogate policy estimator, and let $\hat{\theta}_N$ denote the empirical counterpart from a sample of N trajectories. We show that

$$\sqrt{N}\big(\hat{\theta}_N - \theta^*\big) \xrightarrow{d} \mathcal{N}\big(0,\, H^{-1}\Sigma H^{-1}\big)\,,$$

where $\Sigma = \mathrm{Var}\big(\nabla l^\lambda(T,\theta^*)\big)$ and $H = \nabla^2 L^\lambda(\theta^*)$.
We write

$$\hat{L}_N^\lambda(\theta) = \frac{1}{N}\sum_{k=1}^{N} l^\lambda(T_k,\theta), \qquad \Psi_N(\theta) = \nabla_\theta \hat{L}_N^\lambda(\theta).$$

Then

$$\Psi_N(\theta) = \frac{1}{N}\sum_{k=1}^{N} \nabla_\theta l^\lambda(T_k,\theta).$$

By Theorem 3, we have

$$\hat{\theta}_N \xrightarrow{p} \theta^*.$$

Since $\hat{L}_N^\lambda$ is differentiable and $\hat{\theta}_N$ is its unique minimizer, we have

$$\Psi_N(\hat{\theta}_N) = \nabla \hat{L}_N^\lambda(\hat{\theta}_N) = 0.$$

And since θ^* is the unique minimizer of the population criterion L^λ, we also have

$$\nabla_\theta L^\lambda(\theta^*) = 0$$
$$\mathbb{E}\big[\nabla_\theta l^\lambda(T,\theta^*)\big] = 0.$$

We now apply a Taylor expansion of Ψ_N around θ^*. Since Ψ_N is continuously differentiable, there exists $\tilde{\theta}_N$ between $\hat{\theta}_N$ and θ^* such that

$$0 = \Psi_N(\hat{\theta}_N) = \Psi_N(\theta^*) + \nabla\Psi_N(\tilde{\theta}_N)\big(\hat{\theta}_N - \theta^*\big).$$

Rearranging gives

$$\sqrt{N}\big(\hat{\theta}_N - \theta^*\big) = -\big(\nabla\Psi_N(\tilde{\theta}_N)\big)^{-1}\sqrt{N}\,\Psi_N(\theta^*). \quad (1)$$

We know that $\nabla^2 L_N^\lambda(\tilde{\theta}_N) = \nabla\Psi_N(\tilde{\theta}_N)$ is positive definite, then its inverse in (1) is well defined.
We now study the the right-hand side of (1).
First, since $\hat{\theta}_N \xrightarrow{p} \theta^*$, and $\tilde{\theta}_N$ a convex combination of $\hat{\theta}_N$ and θ^*, it also converges in probability to θ^*. Also,

$$\nabla\Psi_N(\tilde{\theta}_N) = \nabla^2 \hat{L}_N^\lambda(\tilde{\theta}_N) = \frac{1}{N}\sum_{k=1}^{N} \nabla^2 l^\lambda(T_k,\tilde{\theta}_N).$$

By the law of large numbers we obtain

$$\nabla\Psi_N(\tilde{\theta}_N) \xrightarrow{p} \mathbb{E}\big[\nabla^2 l^\lambda(T,\theta^*)\big] = \nabla^2 L^\lambda(\theta^*) = H.$$

Therefore

$$\big(\nabla\Psi_N(\tilde{\theta}_N)\big)^{-1} \xrightarrow{p} H^{-1}. \quad (2)$$

Second, we have

$$\sqrt{N}\,\Psi_N(\theta^*) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N} \nabla l^\lambda(T_k,\theta^*).$$

The trajectories $T_1,\ldots,T_N$ are i.i.d., and

$$\mathbb{E}\big[\nabla l^\lambda(T,\theta^*)\big] = 0.$$

Hence, by the multivariate central limit theorem,

$$\sqrt{N}\,\Psi_N(\theta^*) \xrightarrow{d} \mathcal{N}(0,\Sigma), \quad (3)$$

where

$$\Sigma = \mathrm{Var}\big(\nabla l^\lambda(T,\theta^*)\big).$$

Combining (1), (2), and (3) with Slutsky's theorem yields

$$\sqrt{N}\big(\hat{\theta}_N - \theta^*\big) \xrightarrow{d} -H^{-1}Z, \qquad Z \sim \mathcal{N}(0,\Sigma).$$

Since H is a Hessian matrix, it is symmetric, and since $-Z$ has the same distribution as Z, we get

$$-H^{-1}Z \sim \mathcal{N}\big(0,\, H^{-1}\Sigma H^{-1}\big)\,.$$

Therefore

$$\sqrt{N}\big(\hat{\theta}_N - \theta^*\big) \xrightarrow{d} \mathcal{N}\big(0,\, H^{-1}\Sigma H^{-1}\big)\,.$$

□

Proof of Theorem 5. Let θ_e^* be the population target parameter vector of the surrogate policy estimator for episode e. Let $\pi_e = \text{top-m}(X_e)$ be the true agent's policy at episode e, and X_e the underlying index vector. We show

$$\pi_1 = \pi_2 \implies \theta_1^* = \theta_2^*.$$

In order to show that implication, we have to prove that a well-defined mapping $F(\pi) = \theta^*$ exists.
Under the assumption that all trajectories are i.i.d, the initial distribution of the features Φ_1 is the same between episode 1 and episode 2.
Since the transition probabilities in the system depend only on the current state and action, the joint distribution of the features in a trajectory depend only on the policy that have been played. Consequently, the joint distribution of the actions in a trajectory also depends only on the policy that has been played.
Then the distribution of the trajectories P_T depends only on the policy that has been played. And we know that the population target parameter vector is derived from the distribution the trajectories. Hence a mapping $\pi \to P_T \to \theta*$.
We conclude

$$\pi_1 = \pi_2 \implies \theta_1^* = \theta_2^*.$$

□