Exploring Large Action Sets with Hyperspherical Embeddings using von Mises-Fisher Sampling

Walid Bendada [1 2] **Guillaume Salha-Galvan** [3] **Romain Hennequin** [1]
Théo Bontempelli [1] **Thomas Bouabça** [1] **Tristan Cazenave** [2]

Abstract

This paper introduces von Mises-Fisher exploration (vMF-exp), a scalable method for exploring large action sets in reinforcement learning problems where hyperspherical embedding vectors represent these actions. vMF-exp involves initially sampling a state embedding representation using a von Mises-Fisher distribution, then exploring this representation's nearest neighbors, which scales to virtually unlimited numbers of candidate actions. We show that, under theoretical assumptions, vMF-exp asymptotically maintains the same probability of exploring each action as Boltzmann Exploration (B-exp), a popular alternative that, nonetheless, suffers from scalability issues as it requires computing softmax values for each action. Consequently, vMF-exp serves as a scalable alternative to B-exp for exploring large action sets with hyperspherical embeddings. Experiments on simulated data, real-world public data, and the successful large-scale deployment of vMF-exp on the recommender system of a global music streaming service empirically validate the key properties of the proposed method.

1. Introduction

Exploration is a fundamental component of the reinforcement learning (RL) paradigm (Amin et al., 2021; McFarlane, 2018; Sutton & Barto, 2018). It allows RL agents to gather valuable information about their environment and identify optimal actions that maximize rewards (Amin et al., 2021; Chiappa et al., 2023; Dulac-Arnold et al., 2015; Jin et al., 2020; Ladosz et al., 2022; McFarlane, 2018; Reynolds, 2002; Slivkins et al., 2019; Sutton & Barto, 2018; Tang et al., 2017). However, as the set of actions to explore grows larger, the exploration process becomes increasingly challenging. Indeed, large action sets can lead to higher computational costs, longer learning times, and the risk of inadequate exploration and suboptimal policy development (Amin et al., 2021; Chen et al., 2021; Dulac-Arnold et al., 2015; Lillicrap et al., 2016; Sutton & Barto, 2018; Tomasi et al., 2023).

As an illustration, consider a recommender system on a music streaming service like Apple Music or Spotify, curating playlists of songs "inspired by" an initial selection to help users discover music (Bendada et al., 2023a). In practice, these services often generate such playlists all at once, using efficient nearest neighbor search systems (Johnson et al., 2019; Li et al., 2019) to retrieve songs most similar to the initial one, in a song embedding vector space learned using collaborative filtering or content-based methods (Bendada et al., 2023a;b; Bontempelli et al., 2022; Jacobson et al., 2016; Schedl et al., 2018; Zamani et al., 2019). Alternatively, one could formalize this task as an RL problem (Tomasi et al., 2023), where the recommender system (i.e., the agent) would adaptively select the next song to recommend (i.e., the next action) based on user feedback on previously recommended songs (i.e., the rewards, such as likes or skips). Using an RL approach instead of generating the playlist at once would have the advantage of dynamically learning from user feedback to identify the best recommendations (Afsar et al., 2022; Tomasi et al., 2023). However, music streaming services offer access to large catalogs with several millions of songs (Bendada et al., 2020; Jacobson et al., 2016; Schedl et al., 2018). Therefore, the agent would need to consider millions of possible actions for exploration, increasing the complexity of this task.

In particular, Boltzmann Exploration (B-exp) (Cesa-Bianchi et al., 2017; Sutton & Barto, 2018), one of the prevalent exploration strategies to sample actions based on embedding similarities, would become practically intractable as it would require computing softmax values over millions of elements (see Section 2). Furthermore, in large action sets, many actions are often irrelevant; in our example, most songs would constitute poor recommendations (Tomasi

[1]Deezer Research, Paris, France. [2]LAMSADE, Université Paris Dauphine, PSL, Paris, France. [3]SPEIT, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Walid Bendada <bendadaw@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

et al., 2023). Therefore, random exploration methods like ε-greedy (Dann et al., 2022; Sutton & Barto, 2018), although more efficient than B-exp, would also be unsuitable for production use. Since these methods ignore song similarities, each song, including inappropriate ones, would have an equal chance of being selected for exploration. This could result in negative user feedback and a poor perception of the service (Tomasi et al., 2023). Lastly, deterministic exploration strategies would also be ineffective. Systems serving millions of users often rely on batch RL (Lange et al., 2012) since updating models after every trajectory is impractical. Batch RL, unlike on-policy learning, requires exploring actions non-deterministically given a state, and deterministic exploration would result in redundant trajectories and slow convergence (Bendada et al., 2020).

In summary, exploration remains challenging in RL problems characterized by large action sets and where accounting for embedding similarities is crucial, like our recommendation example. Overall, although a growing body of scientific research has been dedicated to adapting RL models for recommendation (see, e.g., the survey by Afsar et al. (2022)), evidence of RL adoption in commercial recommender systems exists but remains limited (Chen et al., 2019; 2021; 2022; Tomasi et al., 2023). The few existing solutions typically settle for a workaround by using a truncated version of B-exp (TB-exp). In TB-exp, a small subset of candidate actions is first selected, e.g., using approximate nearest neighbor search (a framework sometimes referred to as the Wolpertinger architecture (Dulac-Arnold et al., 2015)). Softmax values are then computed among those candidates only (Chen et al., 2019; 2021; 2022). YouTube, for instance, employs this technique for video recommendation (Chen et al., 2019). TB-exp allows for exploration in the close embedding neighborhood of a given state; however, it restricts the number of candidate actions based on technical considerations rather than optimal convergence properties. Although exploring beyond this restricted neighborhood might be beneficial, finding the best way to do so in large-scale settings remains an open research question.

In this paper[1], we propose to address this important question with a focus on its theoretical foundations. Our work examines the specific setting where actions are represented by embedding vectors with unit Euclidean norm, i.e., vectors lying on a unit hypersphere. As detailed in Section 2, this setting aligns with many real-world applications. Specifically, our contributions in this paper are as follows:

- We present von Mises-Fisher exploration (vMF-exp), a scalable sampling method for exploring large sets of actions represented by hyperspherical embedding vectors. vMF-exp involves initially sampling a state embedding vector on the unit hypersphere using a von Mises-Fisher distribution (Fisher, 1953), then exploring the approximate nearest neighbors of this representation. This strategy effectively scales to large sets with millions of candidate actions to explore.
- We provide a comprehensive mathematical analysis of the behavior of vMF-exp, demonstrating that it exhibits desirable properties for effective exploration in large-scale RL problems. Notably, vMF-exp does not restrict exploration to a specific neighborhood and effectively preserves order while leveraging information from embedding vectors to assess action relevance.
- We also show that, under some theoretical assumptions, vMF-exp asymptotically maintains the same probability of exploring each action as the popular B-exp method while overcoming its scalability issues. This positions vMF-exp as a scalable alternative to B-exp.
- The primary objective of this paper is the introduction and mathematical analysis of the theoretical properties of vMF-exp. Nonetheless, as a complement to this analysis, we also report empirical validations of the method, including experiments on simulated data, on real-world publicly available data, and a discussion of the recent and successful deployment of vMF-exp on a global music streaming service to recommend music to millions of users daily. These experiments validate the key properties of the proposed method and its potential.
- We publicly release a Python implementation of vMF-exp on GitHub to enable reproducibility of our experiments and to encourage future use of the method: `https://github.com/deezer/vMF-exploration`.

The remainder of this paper is organized as follows. Section 2 formalizes the problem. Section 3 introduces vMF-exp, Section 4 details our theoretical analysis, Section 5 discusses our experiments, and Section 6 concludes.

2. Preliminaries

We begin this section by formally introducing the problem addressed in this paper, followed by an explanation of the limitations of existing and popular RL exploration strategies.

2.1. Problem Formulation

Notation Throughout this paper, we consider an RL agent sequentially selecting actions within a set of $n \in \mathbb{N}^\star$ actions:

$$\mathcal{I}_n = \{1, 2, \dots, n\}. \quad (1)$$

[1]Parts of the content published in this ICML 2025 conference paper were previously presented at two non-archival workshops: ICML 2024 ARLET (Bendada et al., 2024) and ICLR 2025 FPI (Bendada et al., 2025).

Each action $i \in \mathcal{I}_n$ is represented by a distinct low-dimensional vectorial representation $X_i \in \mathbb{R}^d$, i.e., by an embedding vector or simply an embedding[2], for some fixed dimension $d \in \mathbb{N}$ with $d \geq 2$ and $d \ll n$. Additionally, we assume all vectors have a unit Euclidean norm, i.e., $\|X_i\|_2 = 1, \forall i \in \mathcal{I}_n$. They form a set of embeddings noted $\mathcal{X}_n = \{X_i\}_{1 \leq i \leq n} \in (\mathcal{S}^{d-1})^n$, where $\mathcal{S}^{d-1}$ is the d-dimensional unit hypersphere (Fisher, 1953):

$$\mathcal{S}^{d-1} = \Big\{ x \in \mathbb{R}^d : \|x\|_2 = 1 \Big\}. \tag{2}$$

We also assume the availability of an approximate nearest neighbor (ANN) (Johnson et al., 2019; Li et al., 2019) search engine. Using this engine, for any vector $V \in \mathcal{S}^{d-1}$, the nearest neighbor of V among $\mathcal{X}_n$ in terms of inner product similarity (equal to the cosine similarity, for unit vectors (Tan et al., 2016)), called $X_{i^\star_V}$, can be retrieved in a sublinear time complexity with respect to n. Although ANN engines are parameterized based on a trade-off between efficiency and accuracy, we make the simplifying assumption that $X_{i^\star_V}$ is the actual nearest neighbor of V, which we later discuss in Section 4.3. Formally:

$$i^\star_V = \arg\max_{i \in \mathcal{I}_n} \langle V, X_i \rangle. \tag{3}$$

Returning to the illustrative example of Section 1, $\mathcal{X}_n$ would represent embeddings associated with each song of the catalog $\mathcal{I}_n$ of the music streaming service. In this case, n would be on the order of several millions (Bendada et al., 2020; Briand et al., 2021; Jacobson et al., 2016). The RL agent would be the recommender system sequentially recommending these songs to users. Normalizing embeddings is a common practice in both academic and industrial recommender systems (Afchar et al., 2023; Bontempelli et al., 2022; Kim et al., 2023; Schedl et al., 2018) to mitigate popularity biases, as vector norms often encode popularity information on items (Afchar et al., 2023; Chen et al., 2023). Normalizing embeddings also prevents inner products from being unbounded, avoiding overflow and underflow numerical instabilities (LeCun et al., 2015).

At time t, the agent considers a state vector $V_t \in \mathcal{S}^{d-1}$, noted V for brevity. It selects the next action in $\mathcal{I}_n$, whose relevance is evaluated by a reward provided by the environment. In our example, the agent would recommend the next song to continue the playlist, based on the previous song whose embedding V acts as the current state. In this case, the reward might be based on user feedback, such as liking or skipping the song (Bontempelli et al., 2022). The agent may select $i^\star_V$, i.e., exploit $i^\star_V$ (Sutton & Barto, 2018). Alternatively, it may rely on an exploration strategy to select another $\mathcal{I}_n$ element. Formally, an exploration strategy P is a policy function (Sutton & Barto, 2018) that, given V, selects each action $i \in \mathcal{I}_n$ with a probability $P(i \mid V) \in [0, 1]$.

[2]At this stage, we do not make assumptions regarding the specific methods used to learn these embedding vectors, nor the interpretation of proximity between vectors in the embedding space.

Objective Our goal in this paper is to develop a suitable exploration strategy for our specific setting, where hyperspherical embedding vectors represent actions, and the number of actions can reach millions. Precisely, we aim to obtain an exploration scheme meeting the following properties:

- Scalability (**P1**): we consider an exploration scalable if the time required to sample actions given a vector V is at most the time needed for the ANN engine to retrieve the nearest neighbor, which is typically achieved in a sublinear time complexity with respect to n. Scalability is a mandatory requirement for exploring large action sets with millions of elements.
- Unrestricted radius (**P2**): Radius$(P \mid V)$ is the number of actions with a non-zero probability of being explored given a state V. While exploring actions too far from V might be suboptimal (e.g., resulting in poor recommendations), it is crucial that exploration is not restricted to a specific radius by construction. Such a restriction could prevent the agent from exploring relevant actions that lie beyond this radius. An unrestricted radius ensures that the exploration strategy remains flexible and capable of adapting to various contexts, allowing for the exploration of relevant actions regardless of their embedding position.
- Order preservation (**P3**): order is preserved when the probability of selecting the action i given the state V is a strictly increasing function of $\langle V, X_i \rangle$. More formally, order preservation requires that, $\forall (i, j) \in \mathcal{I}_n^2$,

$$\langle V, X_i \rangle > \langle V, X_j \rangle \implies P(i \mid V) > P(j \mid V). \tag{4}$$

  P3 ensures that the exploration strategy properly leverages the information captured in the embedding vectors to assess the relevance of an action given a state.

2.2. Limitations of Existing Exploration Strategies

Finding an strategy that simultaneously meets the three properties **P1**, **P2** and **P3** is essential for effective exploration in RL problems with large action sets and embedding representations. Nonetheless, existing exploration methods fail to achieve this, which motivates our work in this paper.

Random and ε-greedy Exploration The most straightforward example of an action exploration strategy would be the random (uniform) policy, where:

$$\forall i \in \mathcal{I}_n, P_{\text{rand}}(i \mid V) = \frac{1}{n}. \tag{5}$$

A popular variant is the ε-greedy strategy (Sutton & Barto, 2018). With a probability $\varepsilon \in [0, 1]$, the agent would choose the next action uniformly at random. With a probability $1 - \varepsilon$, it would exploit the most relevant action based on its knowledge. Random and ε-greedy exploration strategies are scalable (**P1**), as elements of $\mathcal{I}_n$ can be uniformly sampled in $\mathcal{O}(1)$ time (Cormen et al., 2022). Additionally, they verify **P2**. Indeed, $\text{Radius}(P_{\text{rand}}|V) = n$ since every action can be selected. However, these strategies ignore embeddings at the sampling phase and do not achieve order preservation (**P3**). This is a significant limitation, reinforced by the fact that these policies have a maximal radius. As explained in Section 1, in large action sets, many actions are often irrelevant, e.g., most songs from the musical catalog would constitute poor recommendations given an initial state (Tomasi et al., 2023). Exploring each action/song with equal probability, including inappropriate ones, could result in negative user feedback and a poor perception of the service (Tomasi et al., 2023).

Boltzmann Exploration To overcome the limitations of random exploration, actions can be sampled based on their embedding similarity with V, typically measured using dot products. The prevalent approach in RL is Boltzmann Exploration (B-exp) (Amin et al., 2021; Cesa-Bianchi et al., 2017; Chen et al., 2021; Sutton & Barto, 2018), which employs the Boltzmann distribution for action sampling:

$$\forall i \in \mathcal{I}_n, P_{\text{B-exp}}(i \mid V, \mathcal{X}_n, \kappa) = \frac{\mathrm{e}^{\kappa\langle V, X_i\rangle}}{\sum_{j=1}^{n} \mathrm{e}^{\kappa\langle V, X_j\rangle}}, \quad (6)$$

where the hyperparameter $\kappa \in \mathbb{R}^+$ controls the entropy of the distribution. B-exp samples actions according to a strictly increasing function of their inner product similarity with V for $\kappa > 0$, guaranteeing order preservation (**P3**). By carefully tuning κ, one can ensure that irrelevant actions are practically never selected while maintaining a non-zero probability of recommending actions with less than maximal similarity, thereby indirectly controlling the radius of the policy (**P2**). Unfortunately, B-exp does not satisfy **P1**, i.e., it is not scalable to large action sets. Indeed, evaluating Equation (6) requires explicitly computing the probability of sampling each individual action before actually sampling from them, which is prohibitively expensive for large values of n (Chen et al., 2021). Note that, while we focus on B-exp, these concerns would remain valid for any other sampling distribution requiring explicitly computing similarities and probabilities for each of the n actions (Amin et al., 2021).

Truncated Boltzmann Exploration Due to these scalability concerns, previous work sometimes settled for a workaround consisting in sampling actions from a Truncated version of B-exp (Chen et al., 2021), which we refer to as TB-exp. A small number $m \ll n$ of candidate actions, usually around hundreds or thousands, is first retrieved using the ANN search engine, leading to a candidate action set $\mathcal{I}_m(V)$. The sampling step is subsequently performed only within $\mathcal{I}_m(V)$. More formally, for all $i \in \mathcal{I}_m(V)$:

$$P_{\text{TB-exp},m}(i \mid V, \mathcal{X}_n, \kappa) = \frac{\mathrm{e}^{\kappa\langle V, X_i\rangle}}{\sum_{j\in\mathcal{I}_m(V)} \mathrm{e}^{\kappa\langle V, X_j\rangle}}. \quad (7)$$

TB-exp performs action selection in a time that depends on m instead of n, and has been successfully deployed in production environments involving millions of actions (Chen et al., 2019; 2021; 2022). While it still satisfies **P3**, TB-exp also meets **P1** for small values of m. However, it no longer satisfies **P2**. This method restricts the radius, i.e., the number of candidate actions, based on technical considerations rather than exploration efficiency. This restriction can potentially hinder model convergence by neglecting the exploration of relevant actions beyond this fixed radius. This highlights the difficulty of designing a method that simultaneously satisfies **P1**, **P2**, and **P3** – ideally, one with properties akin to B-exp but with greater scalability.

3. From Boltzmann to vMF Exploration

In this section, we present our proposed solution for exploring large action sets with hyperspherical embeddings.

3.1. von Mises–Fisher Exploration

The inability of B-exp to scale arises from its need to compute all n sampling probabilities explicitly. In this paper, we propose von Mises-Fisher Exploration (vMF-exp), an alternative strategy that overcomes this constraint. Specifically, given an initial state vector V, vMF-exp consists in:

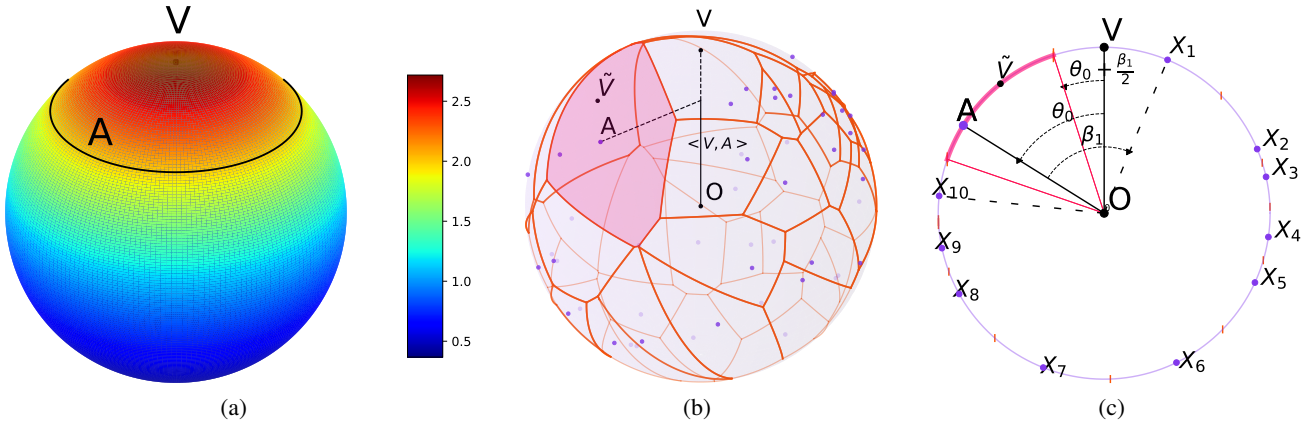
- Firstly, sampling an hyperspherical vector $\tilde{V}$ on the unit hypersphere $\mathcal{S}^{d-1}$, according to a von Mises-Fisher distribution (Fisher, 1953) centered on V.
- Secondly, selecting $\tilde{V}$'s nearest neighbor action in the embedding space for exploration.

In directional statistics, the vMF distribution (Fisher, 1953) is a continuous vector probability distribution defined on the unit hypersphere $\mathcal{S}^{d-1}$. It has recently been used in RL to assess the uncertainty of gradient directions (Zhu et al., 2024). For all $\tilde{V} \in \mathcal{S}^{d-1}$, its probability density function (PDF) is defined as follows:

$$f_{\text{vMF}}(\tilde{V} \mid \kappa, V, d) = C_d(\kappa) e^{\kappa\langle V, \tilde{V}\rangle}, \quad (8)$$

with:

$$C_d(\kappa) = \frac{1}{\int_{\tilde{V}\in\mathcal{S}^{d-1}} e^{\kappa\langle V, \tilde{V}\rangle}\, \mathrm{d}\tilde{V}} = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)}. \quad (9)$$



(a) (b) (c)

Figure 1. (a) Probability density function (PDF) of a 3-dimensional vMF distribution. (b) vMF-exp explores the action a represented by the embedding vector A when the sampled vector $\tilde{V}$ lies in A's Voronoï cell, shown in red in 3D. (c) Same as (b) in a 2-dimensional setting.

and where $\kappa \in \mathbb{R}^+$. The function $I_{\frac{d}{2}-1}$ designates the modified Bessel function of the first kind (Baricz, 2010) at order $d/2-1$. Figure 1(a) illustrates the PDF of a vMF distribution on the 3-dimensional unit sphere. For any $\tilde{V} \in \mathcal{S}^{d-1}$, $f_{\text{vMF}}(\tilde{V} \mid \kappa, V, d)$ is proportional to $e^{\kappa\langle V,\tilde{V}\rangle}$, which is reminiscent of the B-exp sampling probability of Equation (6). The hyperparameter κ controls the entropy of the distribution. In particular, for $\kappa = 0$, the vMF distribution boils down to the uniform distribution on $\mathcal{S}^{d-1}$.

3.2. Properties

P1 vMF-exp only requires sampling a d-dimensional vector instead of handling a discrete distribution with n parameters, allowing $\tilde{V}$ to be sampled in constant time with respect to n. Therefore, vMF-exp is a scalable exploration strategy. Efficient sampling algorithms for vMF distributions have been well-studied (Kang & Oh, 2024; Pinzón & Jung, 2023) (see Appendix F for practical details). As shown in the following sections, we successfully explored sets of millions of actions without scalability issues, using the Python vMF sampler from Pinzón & Jung (2023) for simulations in Section 4 and a custom implementation for the recommendation application in Section 5.

P2 The probability of sampling $i \in \mathcal{I}_n$ given V for exploration is the probability that X_i is the nearest neighbor of $\tilde{V}$ within $\mathcal{X}_n$, i.e., that $\tilde{V}$ lies in $\mathcal{S}_{\text{Voronoï}}(X_i \mid \mathcal{X}_n)$, the Voronoï cell of X_i in the Voronoï tessellation of $\mathcal{S}^{d-1}$ defined by $\mathcal{X}_n$ (Du et al., 1999; 2010) (see Figures 1(b), 1(c)). We have:

$$\mathcal{S}_{\text{Voronoï}}(X_i \mid \mathcal{X}_n) = \{\tilde{V} \in \mathcal{S}^{d-1}, \forall j \in \mathcal{I}_n, \langle \tilde{V}, X_i\rangle \geq \langle \tilde{V}, X_j\rangle\}, \tag{10}$$

and:

$$\bigcup_{i \in \mathcal{I}_n} \mathcal{S}_{\text{Voronoï}}(X_i \mid \mathcal{X}_n) = \mathcal{S}^{d-1}. \tag{11}$$

Thus, vMF-exp's sampling probabilities can be written as:

$$\forall i \in \mathcal{I}_n,\ P_{\text{vMF-exp}}(i \mid V, \mathcal{X}_n, \kappa) = \int_{\tilde{V} \in \mathcal{S}_{\text{Voronoï}}(X_i \mid \mathcal{X}_n)} f_{\text{vMF}}(\tilde{V} \mid \kappa, V, d)\, \mathrm{d}\tilde{V}, \tag{12}$$

which is always strictly positive. Therefore, vMF-exp satisfies the unrestricted radius property (**P2**). Similar to B-exp, adjusting κ ensures that actions with low similarity have negligible sampling probabilities in practice.

P3 $P_{\text{vMF-exp}}(i \mid V, \mathcal{X}n, \kappa)$ increases due to two factors: the average $f_{\text{vMF}}(\tilde{V} \mid \kappa, V, d)$ value for $\tilde{V} \in \mathcal{S}_{\text{Voronoï}}(X_i \mid \mathcal{X}_n)$, correlated to $\langle X_i, V\rangle$ and contributing to order preservation, and the surface area of $\mathcal{S}_{\text{Voronoï}}(X_i \mid \mathcal{X}_n)$, which measures X_i's dissimilarity to other $\mathcal{X}_n$ elements. Actions in a low-density subspace of $\mathcal{S}^{d-1}$ have larger Voronoï cells and may be selected more often than those near V but in high-density regions. Thus, vMF-exp favors actions similar to V and dissimilar to others, with order preservation being dependent on $\mathcal{X}_n$'s distribution. Section 4 focuses on a setting where B-exp and vMF-exp asymptotically share similar probabilities. Consequently, vMF-exp, like B-exp, will verify order preservation (**P3**). In conclusion, in this setting, vMF-exp will verify **P1**, **P2**, and **P3** simultaneously.

4. Theoretical Comparison: vMF-exp vs B-exp

We now provide a mathematical comparison of vMF-exp and B-exp. We focus on the theoretical setting presented in Section 4.1. We show that, in this setting, vMF-exp maintains the same probability of exploring each action as B-exp, while overcoming its scalability issues. As noted above, this implies that vMF-exp verifies **P1**, **P2**, and **P3** simultaneously and, therefore, acts as a scalable alternative to the popular but unscalable B-exp for exploring large action sets with hyperspherical embeddings.

4.1. Setting and Assumptions

We focus on the setting where embeddings are independent and identically distributed (i.i.d.) and follow a uniform distribution on the unit hypersphere, i.e.,

$$\mathcal{X}_n \sim \mathcal{U}(\mathcal{S}^{d-1}). \tag{13}$$

For convenience in our proofs, we consider the action set to be the union of $\mathcal{I}_n$, the set of n actions, and another action a with a known embedding $A \in \mathcal{S}^{d-1}$. The resulting entire action set $\mathcal{I}_{n+1}$ and embedding set $\mathcal{X}_{n+1}$ are defined as $\mathcal{I}_{n+1} = \mathcal{I}_n \cup \{a\}$ and $\mathcal{X}_{n+1} = \mathcal{X}_n \cup \{A\}$. In this section, we are interested in the probability of each exploration scheme, B-exp and vMF-exp, to sample a among all actions of $\mathcal{I}_{n+1}$ given a state embedding vector $V \in \mathcal{S}^{d-1}$. These probabilities are defined respectively as:

$$P_{\text{B-exp}}(a \mid n, d, V, \kappa) = \mathbb{E}_{\ \mathcal{X}_n \sim \mathcal{U}(\mathcal{S}^{d-1})}\Big[P_{\text{B-exp}}(a \mid V, \mathcal{X}_{n+1}, \kappa)\Big], \tag{14}$$

and:

$$P_{\text{vMF-exp}}(a \mid n, d, V, \kappa) = \mathbb{E}_{\ \mathcal{X}_n \sim \mathcal{U}(\mathcal{S}^{d-1})}\Big[P_{\text{vMF-exp}}(a \mid V, \mathcal{X}_{n+1}, \kappa)\Big]. \tag{15}$$

4.2. Results

We now present and discuss our main theoretical results. For brevity, we report all intermediary lemmas and mathematical proofs in the Appendices A, B, C and D of this paper. Our first and most general result links the asymptotic behavior of B-exp and vMF-exp as the action set grows.

Proposition 4.1. *In the setting of Section 4.1, we have:*

$$\lim_{n \to +\infty} \frac{P_{\textit{B-exp}}(a \mid n, d, V, \kappa)}{P_{\textit{vMF-exp}}(a \mid n, d, V, \kappa)} = 1. \tag{16}$$

Proposition 4.1 states that, for large values of n, the probability of sampling the action a for exploration is asymptotically the same using either B-exp or vMF-exp. This result follows from the respective asymptotic characterizations of $P_{\text{B-exp}}$ and $P_{\text{vMF-exp}}$, detailed below. Importantly, it implies that, for large values of n, vMF-exp shares the same properties as B-exp (**P2**, **P3**), including order preservation. However, as noted in Section 3, vMF-exp offers greater scalability since its implementation only requires sampling a vector of a fixed size d, an operation independent of n (**P1**).

Next, we present a common approximate analytical expression for both methods, denoted P_0 and defined as follows:

$$P_0(a \mid n, d, V, \kappa) = \frac{f_{\text{vMF}}(A \mid V, \kappa)\mathcal{A}(\mathcal{S}^{d-1})}{n}, \tag{17}$$

with $\mathcal{A}(\mathcal{S}^{d-1})$ denoting the surface area of the hypersphere $\mathcal{S}^{d-1}$. The following two propositions describe the rate at which this asymptotic behavior is reached by B-exp and vMF-exp, respectively, as n grows.

Proposition 4.2. *In the setting of Section 4.1, we have:*

$$P_{\textit{B-exp}}(a \mid n, d, V, \kappa) = P_0(a \mid n, d, V, \kappa) + o(\frac{1}{n\sqrt{n}}). \tag{18}$$

Proposition 4.3. *In the setting of Section 4.1, we have:*

$$P_{\textit{vMF-exp}}(a \mid n, d, V, \kappa) = P_0(a \mid n, d, V, \kappa) + \begin{cases} \mathcal{O}(\frac{1}{n^2}) & \text{if } d = 2, \\ \mathcal{O}(\frac{1}{n^{1+\frac{2}{d-1}}}) & \text{if } d > 2. \end{cases} \tag{19}$$

In essence, when n is large, the probability of sampling a can be approximated by the PDF of the vMF distribution evaluated at A multiplied by the average surface area of A's Voronoï cell, for both methods. As n grows, this Voronoï cell shrinks until f_{vMF} becomes nearly constant across its entire surface. Figure 2(f) illustrates this interpretation.

The rate at which the two exploration methods reach their asymptotic behavior differs. Specifically, the shrinking rate of the Voronoi cell depends on the dimension of the hypersphere, explaining why the second term in Equation (19) depends on d. This dependency does not occur with B-exp. Consequently, for large values of d, one may require a higher number of actions n before the asymptotic behavior of Equation (16) is observed. For this reason, it is useful to obtain a more precise approximation of $P_{\text{vMF-exp}}(a \mid n, d, V, \kappa)$ when d increases, which we provide in the next section.

4.3. Discussion

High Dimension Building on the above discussion, Proposition 4.4 provides a more precise expression for $P_{\text{vMF-exp}}(a \mid n, V, \kappa)$ when d increases (approximately $d \geq 20$ in our experiments). This expression is derived by examining the first two terms of the Taylor expansion (Abramowitz & Stegun, 1948) of f_{vMF} near A, rather than relying solely on the zero-order term. The second term becomes increasingly significant as d grows. Despite its apparent complexity, the expression has a straightforward interpretation: the negative sign before $\langle V, A \rangle$ indicates that, when A is similar to V, it is sampled less often than with B-exp for the same κ and d. Conversely, when A is on the opposite side of the hypersphere, the term positively contributes to $P_{\text{vMF-exp}}(a \mid n, V, \kappa)$. In summary, for larger d, vMF-exp is expected to explore more extensively than B-exp with the same κ.

Proposition 4.4. *Let* $B : (z_1, z_2) \mapsto \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,dt$ *denote the Beta function, and* $\Gamma : z \mapsto \int_0^\infty t^{z-1}e^{-t}\,dt$

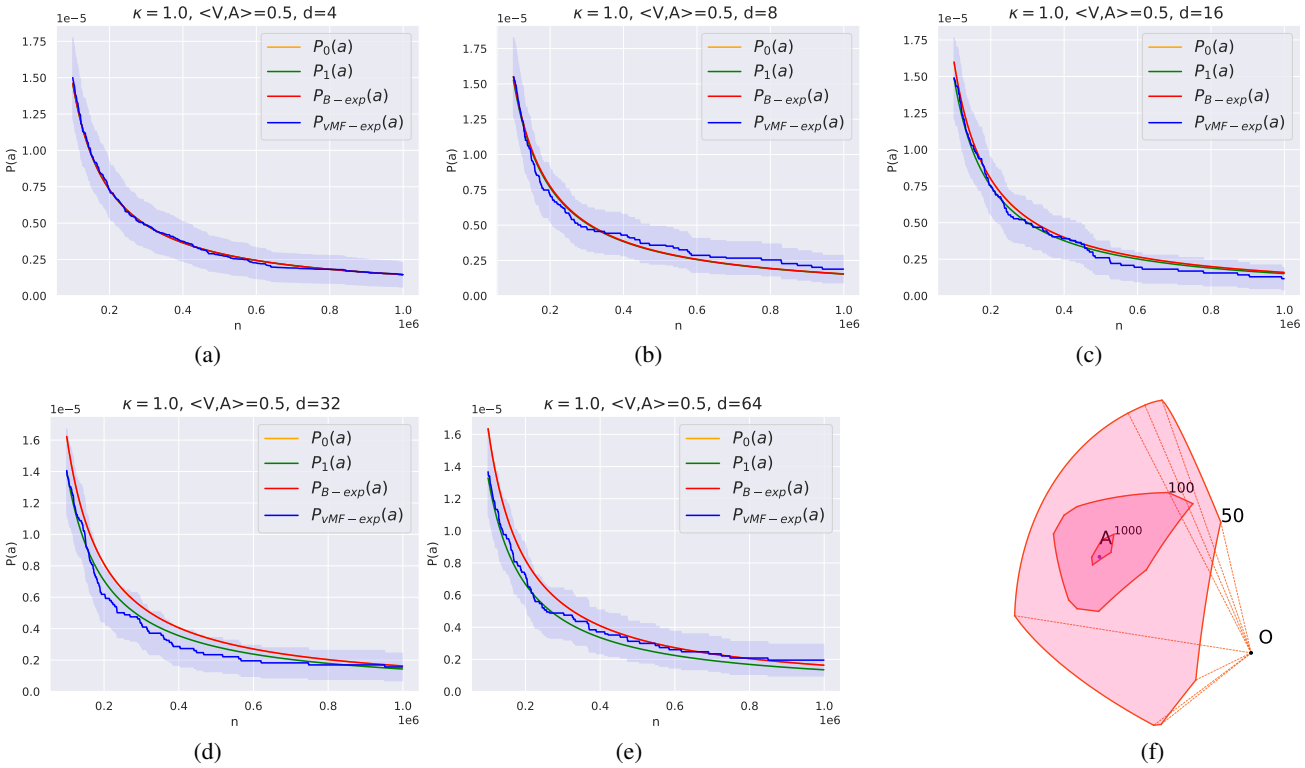


Figure 2. (a) to (e): Validation of the key properties discussed in Sections 4.2 and 4.3 using Monte Carlo simulations, as further elaborated in Section 4.3. (f) Illustration of the 3-dimensional Voronoï cell of a vector A, for action numbers $n \in \{50, 100, 1000\}$.

denote the Gamma function (Abramowitz & Stegun, 1948). In the setting of Section 4.1 with $d \geq 3$, we have:

$$P_{\text{vMF-exp}}(a \mid n, V, \kappa) = P_1(a \mid n, V, \kappa) + \mathcal{O}(\frac{1}{n^{\frac{2}{d-1}}}), \tag{20}$$

with:

$$P_1(a \mid n, V, \kappa) = P_0(a \mid n, V, \kappa) - \left[\frac{f_{\text{vMF}}(A \mid V, \kappa)\mathcal{A}(\mathcal{S}^{d-1})}{n} \frac{\kappa\langle V, A\rangle\Gamma(\frac{d+1}{d-1})}{2} \times \left(\frac{(d-1)B(\frac{1}{2}, \frac{d-1}{2})}{n}\right)^{\frac{2}{d-1}}\right]. \tag{21}$$

The case $d = 2$ In 2 dimensions, Voronoï cells are arcs of a circle and are delimited by the perpendicular bisectors of two neighboring points, as shown in Figure 1(c). Interestingly, in this specific case, $P_{\text{vMF-exp}}(a \mid n, d = 2, V, \kappa)$ can be computed using geometric arguments. A comprehensive mathematical analysis is provided in Appendix B. This analysis confirms that, when $d = 2$, vMF-exp approaches its asymptotic behavior faster than B-exp, as indicated by the $\mathcal{O}(\frac{1}{n^2})$ term in Proposition 4.3.

Validation of Theoretical Properties via Monte Carlo Simulations Using the efficient Python sampler of Pinzón & Jung (2023), we repeatedly sampled vectors $\mathcal{X}_n \sim \mathcal{U}(S^{d-1})$ and $\tilde{V} \sim \text{vMF}(V, \kappa)$, for various values of d, κ, and $\langle V, A\rangle$. Figure 2 reports, for $\kappa = 1.0$, $\langle V, A\rangle = 0.5$ and growing values of d, the $P_{\text{vMF-exp}}(a)$ sampling probability depending on the number of actions n, as well as $P_{\text{B-exp}}(a)$ with similar parameters and our approximations $P_0(a)$ and $P_1(a)$. We repeated all experiments 8 million times and reported 95% confidence intervals. The results are consistent with our key theoretical findings in this paper.

Firstly, in line with Proposition 4.2, $P_{\text{B-exp}}(a)$ and $P_0(a)$ are indistinguishable for this range of n values. Secondly, for small d values (Figures 2(a), 2(b), 2(c)), $P_{\text{vMF-exp}}$ is also tightly aligned with $P_{\text{B-exp}}(a)$ and $P_0(a)$, consistently with Proposition 4.1 and 4.3. Note that the y-axis is on a 1e-5 scale; hence, probabilities are extremely close. Thirdly, when $d \geq 16$ (Figures 2(d), 2(e)), $P_1(a)$ becomes more distinguishable from $P_0(a)$ and constitutes a better approximation of $P_{\text{vMF-exp}}(a)$ than $P_0(a)$, as per Proposition 4.4. Lastly, since $\langle V, A\rangle > 0$, Proposition 4.4 predicts that $P_{\text{B-exp}}(a) \geq P_{\text{vMF-exp}}(a)$ for large d, which our experiments confirm. We provide comparable simulations with other $(d, \kappa, \langle V, A\rangle)$ combinations in Appendix G. All simulations are reproducible using our source code (see Section 5).

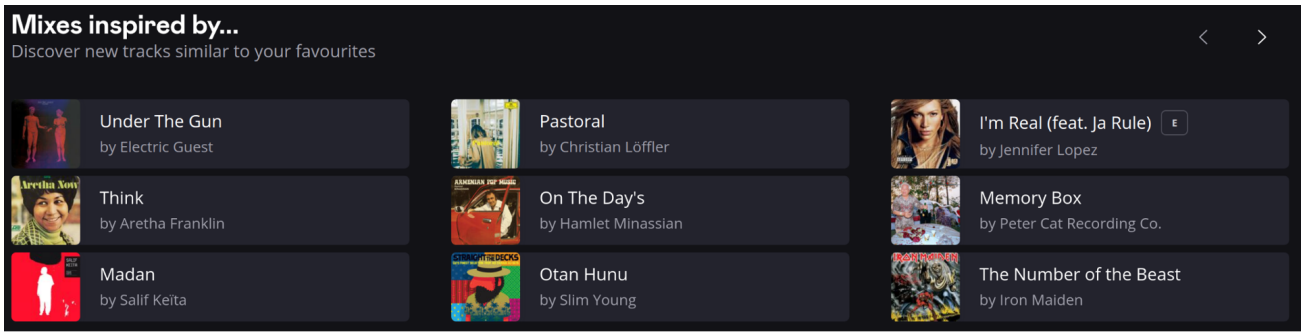


Figure 3. Interface of the "Mixes Inspired By" recommender system on the music streaming service Deezer. This system presents a personalized shortlist of songs liked by the user. Clicking on a song generates a playlist "inspired by" this song. As detailed in Section 5 and Appendix I, vMF-exp has been employed for months on the production environment of this service to generate playlists, exploring songs from a catalog containing millions of candidates.

Link with Thompson Sampling One might notice interesting similarities between vMF-exp and bandit arm exploration with the Thompson Sampling method (Chapelle & Li, 2011). We refer the interested reader to our Appendix E for a detailed comparison of the two approaches.

Limitations and Future Work While we believe our study provides valuable insights into vMF-exp, several limitations must be acknowledged. Our theoretical guarantees are currently restricted to the distributions described in Section 4.1. Although vMF-exp can be applied in practice with hyperspherical embeddings from any other distributions, we have not yet extended our guarantees to such cases[3].

For instance, studying vMF-exp in clustered embedding settings, as is sometimes the case with music recommendation embeddings (where clusters can, e.g., summarize music genres (Salha-Galvan et al., 2022)), could be insightful. We believe that future research should benefit from the approach we proposed in this paper to derive the full (non-trivial) demonstration for the uniform distribution case.

Our future work will also investigate the second-order term from Proposition 4.4, which may be significant for large κ, as well as the impact of ANN errors. Although our analysis assumes exact neighbor retrieval, this assumption may break down for extremely large action sets (Johnson et al., 2019), potentially causing minor exploration perturbations.

5. vMF-exp in the Real World

As explained in Section 1, the primary objective of this paper was to introduce vMF-exp and provide a rigorous mathematical analysis of its theoretical properties for exploring large action sets represented by hyperspherical embedding vectors. Nonetheless, as an opening to our work and a complement to this mathematical analysis, we describe in this section several empirical validations of vMF-exp, further detailed in Appendices G, H, and I. Taken together, these experiments validate the claimed scalability and theoretical properties of the proposed method, as well as its potential and impact for real-world applications. Specifically:

- To begin with, we recall that, as explained in Section 4.3, a more comprehensive outline of the Monte Carlo simulations discussed in that section is provided in Appendix G. The additional results presented in this appendix are consistent with those shown in Figure 2 of Section 4.3. To ensure the reproducibility of these experiments and to encourage future use of the vMF-exp method, we have released a Python implementation of vMF-exp on GitHub with this paper: `https://github.com/deezer/vMF-exploration`.
- To go further, we recognize that some readers may wish to explore our topic through reproducible experiments on real-world data. Therefore, in Appendix H, we empirically validate the main properties of vMF-exp using a large-scale, publicly available dataset of one million GloVe word embedding vectors (Pennington et al., 2014). We demonstrate that vMF-exp simultaneously satisfies **P1**, **P2**, and **P3** on this GloVe dataset. Furthermore, this study on GloVe vectors shows the accuracy of our approximations of $P_{\text{B-exp}}(a)$ and $P_{\text{vMF-exp}}(a)$ from Propositions 4.2, 4.3, and 4.4, despite the fact that GloVe vectors do not strictly meet the i.i.d. and uniform assumptions of our theoretical study. This additional study is fully reproducible using the code provided in the GitHub repository mentioned above.
- The final appendix of this paper, Appendix I, showcases a real-world application of vMF-exp. We present its successful large-scale deployment in the private

[3]Nonetheless, results from Section 5 will tend to confirm the practical usefulness of our propositions on real-world embedding vectors that do *not* strictly satisfy assumptions from Section 4.1.

production system of the global music streaming service Deezer (Bendada et al., 2023a). On this service, vMF-exp has been employed for months to recommend playlists of songs inspired by an initial selection to millions of users (see Figure 3), exploring a catalog of millions of candidate songs. This application, validated by a positive worldwide online A/B test, confirms the practical relevance of our work. As vMF-exp was successfully deployed in production, achieving a sampling latency of just a few milliseconds, it also confirms the scalability of the method.

6. Conclusion

This paper introduced vMF-exp, a scalable method for exploring large action sets in RL problems where hyperspherical embedding vectors represent these actions. vMF-exp scales effectively to large sets with millions of actions, exhibits desirable properties, and, under theoretical assumptions, even asymptotically preserves the same exploration probabilities as B-exp, a prevalent RL exploration method that suffers from scalability limitations. This establishes vMF-exp as a scalable and practical alternative to B-exp. While the primary focus of this paper is on the theoretical foundations of vMF-exp, extensive experiments on simulated data, real-world public data, and the successful deployment of vMF-exp on a music streaming service validated the scalability and practical relevance of the proposed method.

Impact Statement

This paper introduces a scalable method for exploring large action sets in reinforcement learning problems, with applications including recommender systems. While our work is primarily methodological and theoretical, its integration into real-world systems, such as personalized music recommender systems, can influence user experience, engagement, and exposure to information. We highlight the potential for both positive outcomes (e.g., improved personalization and efficiency) and risks such as over-personalization or bias amplification if used without appropriate fairness and diversity constraints. We encourage careful deployment and further study when applying this method to sensitive domains involving social or behavioral data.

References

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government Printing Office, 1948.

Afchar, D., Hennequin, R., and Guigue, V. Of spiky svds and music recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 926–932, 2023.

Afsar, M. M., Crump, T., and Far, B. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Amin, S., Gomrokchi, M., Satija, H., van Hoof, H., and Precup, D. A survey of exploration methods in reinforcement learning. *arXiv preprint arXiv:2109.00157*, 2021.

Aumüller, M., Bernhardsson, E., and Faithfull, A. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. In *Proceedings of the 10th International Conference on Similarity Search and Applications*, pp. 34–49. Springer, 2017.

Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S., and Ridgeway, G. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Banerjee, S. and Roy, A. *Linear Algebra and Matrix Analysis for Statistics*. CRC Press, 2014.

Baricz, Á. *Generalized Bessel Functions of the First Kind*. Springer, 2010.

Bendada, W., Salha, G., and Bontempelli, T. Carousel personalization in music streaming apps with contextual bandits. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 420–425, 2020.

Bendada, W., Bontempelli, T., Morlon, M., Chapus, B., Cador, T., Bouabça, T., and Salha-Galvan, G. Track mix generation on music streaming services using transformers. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 112–115, 2023a.

Bendada, W., Salha-Galvan, G., Bouabça, T., and Cazenave, T. A scalable framework for automatic playlist continuation on music streaming services. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 464–474, 2023b.

Bendada, W., Salha-Galvan, G., Hennequin, R., Bontempelli, T., Bouabça, T., and Cazenave, T. vmf-exp: von mises-fisher exploration of large action sets with hyperspherical embeddings. In *ICML 2024 Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.

Bendada, W., Salha-Galvan, G., Hennequin, R., Bontempelli, T., Bouabça, T., and Cazenave, T. von mises-fisher sampling of glove vectors. In *ICLR 2025 Workshop on Frontiers in Probabilistic Inference: Learning meets Sampling*, 2025.

Billingsley, P. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

Bontempelli, T., Chapus, B., Rigaud, F., Morlon, M., Lorant, M., and Salha-Galvan, G. Flow moods: Recommending music by moods on deezer. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 452–455, 2022.

Boytsov, L. and Naidan, B. Engineering efficient and effective non-metric space library. In *Proceedings of the 6th International Conference on Similarity Search and Applications*, pp. 280–293. Springer, 2013.

Briand, L., Salha-Galvan, G., Bendada, W., Morlon, M., and Tran, V.-A. A semi-personalized system for user cold start recommendation on music streaming apps. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2601–2609, 2021.

Briand, L., Bontempelli, T., Bendada, W., Morlon, M., Rigaud, F., Chapus, B., Bouabça, T., and Salha-Galvan, G. Let's get it started: Fostering the discoverability of new releases on deezer. In *European Conference on Information Retrieval*, pp. 286–291. Springer, 2024.

Cesa-Bianchi, N., Gentile, C., Lugosi, G., and Neu, G. Boltzmann exploration done right. *Advances in Neural Information Processing Systems*, 30, 2017.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in Neural Information Processing Systems*, 24, 2011.

Chen, J., Wu, J., Wu, J., Cao, X., Zhou, S., and He, X. Adap-τ: Adaptively modulating embedding magnitude for recommendation. In *Proceedings of the ACM Web Conference 2023*, pp. 1085–1096, 2023.

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., and Chi, E. H. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pp. 456–464, 2019.

Chen, M., Chang, B., Xu, C., and Chi, E. H. User response models to improve a reinforce recommender system. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 121–129, 2021.

Chen, M., Xu, C., Gatto, V., Jain, D., Kumar, A., and Chi, E. Off-policy actor-critic for recommender systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 338–349, 2022.

Chiappa, A. S., Marin Vargas, A., Huang, A., and Mathis, A. Latent exploration for reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Coolidge, J. L. The story of the binomial theorem. *The American Mathematical Monthly*, 56(3):147–157, 1949.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms*. MIT Press, 2022.

Dann, C., Mansour, Y., Mohri, M., Sekhari, A., and Sridharan, K. Guarantees for epsilon-greedy reinforcement learning with function approximation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 4666–4689. PMLR, 2022.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, et al. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Du, Q., Faber, V., and Gunzburger, M. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, 1999.

Du, Q., Gunzburger, M., and Ju, L. Advances in studies and applications of centroidal voronoi tessellations. *Numerical Mathematics: Theory, Methods and Applications*, 3 (2):119–142, 2010.

Dulac-Arnold, G., Evans, R., van Hasselt, H., Sunehag, P., Lillicrap, T., Hunt, J., Mann, T., Weber, T., Degris, T., and Coppin, B. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*, 2015.

Fischer, H. *A History of the Central Limit Theorem: from Classical to Modern Probability Theory*. Springer, 2011.

Fisher, R. A. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

Gentle, J. E. *Computational Statistics*, volume 308. Springer, 2009.

Gnedenko, B. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, pp. 423–453, 1943.

Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., and Kumar, S. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3887–3896, 2020.

Iwasaki, M. and Miyazaki, D. Optimization of indexing based on k-nearest neighbor graph for proximity search. *arXiv preprint arXiv:1810.07355*, 2018.

Jacobson, K., Murali, V., Newett, E., Whitman, B., and Yon, R. Music Personalization at Spotify. pp. 373–373, 2016.

Jacod, J. and Protter, P. *Probability Essentials*. Springer Science & Business Media, 2004.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019.

Kang, S. and Oh, H.-S. Novel sampling method for the von mises–fisher distribution. *Statistics and Computing*, 34 (3):106, 2024.

Kim, D., Park, J., and Kim, D. Test-time embedding normalization for popularity bias mitigation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4023–4027, 2023.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

Koren, Y. and Bell, R. Advances in Collaborative Filtering. *Recommender Systems Handbook*, pp. 77–118, 2015.

Ladosz, P., Weng, L., Kim, M., and Oh, H. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement Learning: State-Of-The-Art*, pp. 45–73. Springer, 2012.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, W., Zhang, Y., Sun, Y., Wang, W., Li, M., Zhang, W., and Lin, X. Approximate nearest neighbor search on high dimensional data — experiments, analyses, and improvement. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1475–1488, 2019.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *Proceedings of the 4th International Conference on Learning Representation*, 2016.

Malkov, Y. A. and Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2018.

Mardia, K. V. and Jupp, P. E. *Directional Statistics*. John Wiley & Sons, 2009.

McFarlane, R. A survey of exploration strategies in reinforcement learning. *McGill University*, 3:17–18, 2018.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

Oehlert, G. W. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

Pinzón, C. and Jung, K. Fast python sampler for the von mises fisher distribution. *HAL Id: hal-04004568*, 2023.

Reynolds, S. I. Reinforcement learning with exploration. *Ph.D. Thesis, University of Birmingham*, 2002.

Salha-Galvan, G., Lutzeyer, J. F., Dasoulas, G., Hennequin, R., and Vazirgiannis, M. Modularity-aware graph autoencoders for joint community detection and link prediction. *Neural Networks*, 153:474–495, 2022.

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., and Elahi, M. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7:95–116, 2018.

Simhadri, H. V., Aumüller, M., Ingber, A., Douze, M., Williams, G., Manohar, M. D., Baranchuk, D., Liberty, E., Liu, F., Landrum, B., et al. Results of the big ann: Neurips'23 competition. *arXiv preprint arXiv:2409.17424*, 2024.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2):1–286, 2019.

Sra, S. A short note on parameter approximation for von mises-fisher distributions: And a fast implementation of is(x). *Computational Statistics*, 27:177–190, 2012.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Tan, P.-N., Steinbach, M., and Kumar, V. *Introduction to Data Mining*. Pearson Education India, 2016.

Tang, H., Houthooft, R., Foote, D., Stooke, A., Xi Chen, O., Duan, Y., Schulman, J., DeTurck, F., and Abbeel, P. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Tomasi, F., Cauteruccio, J., Kanoria, S., Ciosek, K., Rinaldi, M., and Dai, Z. Automatic music playlist generation via simulation-based reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4948–4957, 2023.

Zamani, H., Schedl, M., Lamere, P., and Chen, C.-W. An analysis of approaches taken in the acm recsys challenge 2018 for automatic music playlist continuation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–21, 2019.

Zhu, Y., Liu, J., Wei, W., Fu, Q., Hu, Y., Fang, Z., An, B., Hao, J., Lv, T., and Fan, C. vmfer: von mises-fisher experience resampling based on uncertainty of gradient directions for policy improvement of actor-critic algorithms. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2621–2623, 2024.