Probability Tools for Sequential Random Projection

Yingru Li[*] YINGRULI@LINK.CUHK.EDU.CN
The Chinese University of Hong Kong, Shenzhen, China

Abstract

We introduce the first probabilistic framework for sequential random projection, an approach rooted in the challenges of sequential decision-making under uncertainty. The analysis is complicated by the sequential dependence and high-dimensional nature of random variables, a byproduct of the adaptive mechanisms inherent in sequential decision processes. This analytical difficulty is resolved by a construction of a stopped process that interconnect a series of concentration events in a sequential manner. By employing the method of mixtures within a self-normalized process, derived from the stopped process, we achieve a desired non-asymptotic probability bound. This bound represents a non-trivial martingale extension of the Johnson-Lindenstrauss (JL) lemma.

1. Introduction

The evolution of random projection from a dimensionality reduction technique to a cornerstone of sequential decision-making processes marks a significant leap in computational mathematics and machine learning. Random projection traditionally employs a matrix $\Pi = (\mathbf{z}_1, \dots, \mathbf{z}_d) \in \mathbb{R}^{M\times d}$ to transform a high-dimensional vector $\boldsymbol{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ into a lower-dimensional space, preserving the Euclidean geometry within a bounded error as guaranteed by the Johnson-Lindenstrauss (JL) lemma. This preservation of geometric relationships, crucial for the efficacy of data compression and analysis techniques, is foundational to the lemma's broad applicability, such as computer science [7, 15], signal processing [2, 3] and numerical linear algebra [19].

1.1. Sequential random projection

Recent advancements, especially in reinforcement learning, underscore the pressing need for computational models that not only accommodate but thrive on the epistemic uncertainties of sequential decision-making [12, 13]. In these dynamic environments, the application of random projection must navigate the added complexity of decisions $(x_t)_{t\geq 1}$ that are influenced by a history of previous decisions and projection vectors $(\mathbf{z}_0, x_1, \mathbf{z}_1, \dots, x_{t-1}, \mathbf{z}_{t-1})$, introducing a layer of sequential dependence absent in static models. More precisely, the sequential relationship among the random variables is

$$x_t = f_t(x_1, \mathbf{z}_1, \dots, x_{t-1}, \mathbf{z}_{t-1}), \quad t \geq 1, \tag{1}$$

where f_t describes the relationship at time t, and $\mathbf{z}_t$ is sampled from a distribution over $\mathbb{R}^M$ with independent source of randomness. This sequential relationships can be also described in fig. 1 using graphical model.

[*] The author would like to acknowledge Professor Zhi-Quan (Tom) Luo for advising this project.

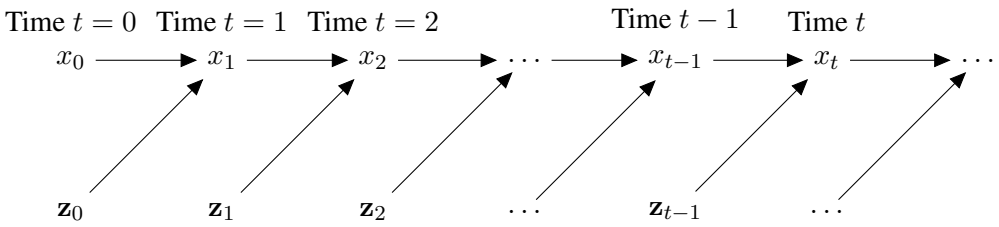

Figure 1: Sequential dependence of high-dimensional random variables due to the adaptive nature of sequential decision-making.

Unpacking the Challenges. This sequential dependence introduces significant analytical hurdles, diverging sharply from the assumptions that underpin classical random projection methods:

- *Traditional analysis of random projections* assumes the data vector $\boldsymbol{x} = (x_1, \ldots, x_d)$ is fixed before the generation of the projection matrix $\Pi = (\mathbf{z}_1, \ldots, \mathbf{z}_d)$. All existing analysis for Johnson-Lindenstrauss or the studies for extreme singular values of random matrices [18] rely on some specific distributional properties of the random matrix Π. For example, conditions on independent entries across Π [1, 8, 14, 18], independent rows or columns across Π [4, 9, 11, 18] are required to facilitate the concentration inequalities underlying the analysis.

- *Analytical difficulties in sequential settings*: In the sequential setups, the decisions x_t and projection vectors $\mathbf{z}_t$ are evolved together with sequential dependence described in eq. (1). Conditioned on the decision at time t, the data x_t, the preceding projection vectors $(\mathbf{z}_0, \ldots, \mathbf{z}_{t-1})$ lose their independence and identical distribution characteristics. This departure from independence directly challenges the foundational assumptions of analytical methods in random projection, complicating the task of maintaining accurate dimensionality reduction over sequential decisions. Specifically, without a clear understanding of the conditional distribution $P_{(\mathbf{z}_{t'})_{t'<t}|x_t}$, traditional methods that rely on the specific distributional properties of projections cannot be straightforwardly applied. This limitation underscores a critical gap in our ability to predict and control the behavior of sequential random projections.

Innovations and Contributions. In addressing the challenges inherent in sequential random projection, this research inaugurates an analytic tool, specifically designed to tackle the intricacies of sequential dependencies. Our work is distinguished by two principal innovations:

1. *Technical innovations in stopped martingale*: Central to our contributions is the construction of a stopped process, meticulously engineered to manage deviation behaviors within sequential processes. This construction crucially facilitates the precise control of concentration events over time, thereby enabling the analysis of the error bound in sequential setting.

2. *Sequential extension of the Johnson–Lindenstrauss*: Through the employment of the method of mixtures, integrated with a self-normalized process derived from the stopped process, we obtained a non-asymptotic probability bounds This bound represent a non-trivial extension of the Johnson–Lindenstrauss lemma into the realm of sequential analysis, equipping researchers and practitioners with a powerful analytical tool for the exploration of high-dimensional data in sequentially adpative processes.

The contributions of this research serve to bridge significant gaps in existing methodological frameworks, furnishing novel insights and methodologies for the application of random projection in sequential settings. By laying a robust foundation for the analysis of dependencies among projection vectors in sequential contexts, our work not only surmounts the immediate challenges posed by sequential random projection but also forges a path for future investigations and applications in this critical intersection of mathematics and computational science. In doing so, it heralds a new paradigm in the analysis and application of random projection techniques for sequential, adaptive, high-dimensional data processing.

2. Probabilistic formalism & statements

2.1. Probabilistic formalism

One of the difficulties in the analysis is to deal with the sequential dependence structure among the random variables generated from the sequential-decision making problem such as bandit and reinforcement learning problems. We define some important concept that would be useful in the analysis. Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{N}}, \mathbb{P})$ be a complete filtered probability space. We first consider the measurable properties within the filtered probability space.

Definition 1 (Adapted process) *For an index set I of the form $\{t \in \mathbb{N} : t \geq t_0\}$ for some $t_0 \in \mathbb{N}$, we say a stochastic process $(X_t)_{t\in I}$ is adapted to the filtration $(\mathcal{F}_t)_{t\in I}$ if each X_t is $\mathcal{F}_t$-measurable.*

Definition 2 ((Conditionally) σ-sub-Gaussian) *A random variable $X \in \mathbb{R}$ is σ-sub-Gaussian if*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \qquad \forall\lambda \in \mathbb{R}.$$

Let $(X_t)_{t\geq 1} \subset \mathbb{R}$ be a stochastic process adapted to filtration $(\mathcal{F}_t)_{t\geq 1}$. Let $\sigma = (\sigma_t)_{t\geq 0}$ be a stochastic process adapted to filtration $(\mathcal{F}_t)_{t\geq 0}$. We say the process is $(X_t)_{t\geq 1}$ is conditionally σ-sub-Gaussian if $\mathbb{E}[\exp(\lambda X_t) \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2\sigma_{t-1}^2}{2}\right), a.s., \forall\lambda \in \mathbb{R}$. Specifically for the index $t+1$, we can say X_{t+1} is ($\mathcal{F}_t$-conditionally) σ_t-sub-Gaussian. If σ_t is a constant σ for all $t \geq 0$, then we just say (conditionally) σ-sub-Gaussian.

For a random vector $X \in \mathbb{R}^M$ or vector process $(X_t)_{t\geq 1} \subset \mathbb{R}^M$ in high-dimension, we say it is σ-sub-Gaussian is for every fixed $v \in \mathbb{S}^{M-1}$ if the random variable $\langle v, X\rangle$, or the scalarized process $(\langle v, X_t\rangle)_{t\geq 1}$ is σ-sub-Gaussian.

Definition 3 (Almost sure unit-norm) *We say a random variable X is almost sure unit-norm if $\|X\|_2 = 1$ almost surely.*

Here, we give an example of distribution in $\mathbb{R}^M$ in example 1 that fits the above mentioned definitions: the uniform distribution $\mathcal{U}(\mathbb{S}^{M-1})$ over unit sphere $\mathbb{S}^{M-1}$. The following recent result for the moment generating function (MGF) of Beta distribution is useful for the characterization of the sub-Gaussian property of the uniform distribution over the sphere.

Lemma 4 (MGF of Beta distribution [11]) *For any $\alpha, \beta \in \mathbb{R}_+$ with $\alpha \geq \beta$, random variable $X \sim \mathrm{Beta}(\alpha, \beta)$ has variance $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ and the centered MGF $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\frac{\lambda^2\mathrm{Var}(X)}{2})$.*

For completeness, we provide the proof of lemma 4 in section C.

Example 1 (Uniform distribution over sphere $\mathcal{U}(\mathbb{S}^{M-1})$) *The random variable $\mathbf{z} \sim \mathcal{U}(\mathbb{S}^{M-1})$ is obviously unit-norm as by definition of the unit sphere $\mathbb{S}^{M-1}$. Also, according to lemma 17, the inner product between the random variable $\mathbf{z}$ and any unit vector $v \in \mathbb{S}^{M-1}$ follows a Beta distribution, i.e., $\langle \mathbf{z}, v \rangle \sim 2\,\mathrm{Beta}\left(\frac{M-1}{2}, \frac{M-1}{2}\right) - 1$. Then, from lemma 4, we could confirm $\mathbf{z} \sim \mathcal{U}(\mathbb{S}^{M-1})$ is $\sqrt{1/M}$-sub-Gaussian.*

Additionally, we characterize the boundedness on the stochastic processes.

Definition 5 (Square-bounded process) *For an index set I of the form $\{t \in \mathbb{N} : t \geq t_0\}$ for some $t_0 \in \mathbb{N}$, the stochastic process $(X_t)_{t \in I}$ is c-square-bounded if $X_t^2 \leq c$ almost surely for all $t \in I$.*

2.2. Probability tools for sequential random projection

In this section, we introduce the first probability tool for addressing sequential random projection, inspired by the challenges of sequential decision-making under uncertainty. The analysis is complicated by the sequential dependence and high-dimensionality of random variables, a consequence of the adaptive nature of sequential decision-making. Our approach leverages a novel and meticulously constructed stopped process that manages the behavior of a sequence of concentration events. By employing the method of mixtures as outlined by [6] within a self-normalized process framework, we derive a non-asymptotic probability bound in theorem 6. This bound represents a non-trivial and significant martingale-based extension of the Johnson–Lindenstrauss (JL) lemma, marking a novel contribution to the fields of random projection and sequential analysis alike. Our technical innovation offers a probability statement for sequential random projection that is unparalleled in the literature, potentially sparking independent interest in both domains.

We use short notation for $[n] = \{1, 2, \ldots, n\}$ and $\mathcal{T} = \{0, 1, \ldots, T\} = \{0\} \cup [T]$.

Theorem 6 (Sequential random projection in adaptive process) *Let $\varepsilon \in (0, 1)$ be fixed and $(\mathcal{F}_t)_{t \geq 0}$ be a filtration. Let $\mathbf{z}_0 \in \mathbb{R}^M$ be an $\mathcal{F}_0$-measurable random vector satisfies $\mathbb{E}[\|\mathbf{z}_0\|^2] = 1$ and $|\|\mathbf{z}_0\|^2 - 1| \leq (\varepsilon/2)$. Let $(\mathbf{z}_t)_{t \geq 1} \subset \mathbb{R}^M$ be a stochastic process adapted to filtration $(\mathcal{F}_t)_{t \geq 1}$ such that it is $\sqrt{c_0/M}$-sub-Gaussian and each $\mathbf{z}_t$ is unit-norm. Let $(x_t)_{t \geq 1} \subset \mathbb{R}$ be a stochastic process adapted to filtration $(\mathcal{F}_{t-1})_{t \geq 1}$ such that it is c_x-square-bounded. Here, c_0 and c_x are absolute constants. For any fixed $x_0 \in \mathbb{R}$, if the following condition is satisfied*

$$M \geq \frac{16 c_0 (1 + \varepsilon)}{\varepsilon^2} \left(\log\left(\frac{1}{\delta}\right) + \log\left(1 + \frac{c_x T}{x_0^2}\right) \right), \tag{2}$$

we have, with probability at least $1 - \delta$

$$\forall t \in \mathcal{T}, \quad (1 - \varepsilon) \left(\sum_{i=0}^{t} x_i^2 \right) \leq \| \sum_{i=0}^{t} x_i \mathbf{z}_i \|^2 \leq (1 + \varepsilon) \left(\sum_{i=0}^{t} x_i^2 \right). \tag{3}$$

Remark 7 *We say this is an "sequential random projection" argument because one can relate theorem 6 to the traditional random projection setting where $\Pi_t = (\mathbf{z}_0, \ldots, \mathbf{z}_t) \in \mathbb{R}^{M \times t+1}$ is a random projection matrix and $\boldsymbol{x}_t = (x_0, \ldots, x_t)^\top \in \mathbb{R}^{t+1}$ is the vector to be projected. The argument in eq. (3) translates to*

$$\forall t \in \mathcal{T}, \quad (1 - \varepsilon) \|\boldsymbol{x}_t\|^2 \leq \|\Pi_t \boldsymbol{x}_t\|^2 \leq (1 + \varepsilon) \|\boldsymbol{x}_t\|^2. \tag{4}$$

When assuming independence between $\boldsymbol{x}_t$ and Π_t for all $t \in \mathcal{T}$, by simply applying union bound over time index $t \in \mathcal{T}$ with existing JL analysis, we can derive that the required dimension $M = O(\varepsilon^{-2}\log(T/\delta))$ is of the same order in eq. (2). ***However****, as discussed in section 1, existing JL analytical techniques are not able to handle the sequential dependence in our setup as $\boldsymbol{x}_t$ is statistically dependent with Π_t for $t \in \mathcal{T}$. Therefore, theorem 6 is also an innovation in the literature of both random projection and sequential analysis.*

Remark 8 *The unit-norm condition in the theorem 6 can be removed. Then, more distribution of random vectors can be covered in our analytical framework. For example, the Gaussian random vector $\mathbf{z} \sim N(0, \frac{1}{M}I)$ is not unit-norm but we could rely on the centered moment generating function of $\|\mathbf{z}\|^2$ by exploiting the properties of Chi-square distribution. We leave it for the future work.*

Example 2 (Stylized stochastic process satisfying the condition in theorem 6.) *Let $(\mathbf{z}_t)_{t\geq 0}$ are mutually independent random variables, each sampled from $\mathcal{U}(\mathbb{S}^{M-1})$. Let x_0 be fixed and $(x_t)_{t\geq 1}$ be the stochastic process with the following dependence structure, interleaved with the process $(\mathbf{z}_t)_{t\geq 0}$: (1) x_t is dependent on $x_0, \mathbf{z}_0, x_1, \mathbf{z}_1, \ldots, x_{t-1}, \mathbf{z}_{t-1}$ as described in eq. (1). (2) $\mathbf{z}_t$ is independent of $x_0, \mathbf{z}_0, x_1, \mathbf{z}_1, \ldots, x_{t-1}, \mathbf{z}_{t-1}, x_t$. This sequential dependence structure is also described in fig. 1. Define the filtration $(\mathcal{F}_t)_{t\geq 0}$ where $\mathcal{F}_t = \sigma(\mathbf{z}_0, x_1, \mathbf{z}_1, \ldots, x_t, \mathbf{z}_t, x_{t+1})$. From example 1, we notice the process $(\mathbf{z}_t)_{t\geq 1}$ adapted to $(\mathcal{F}_t)_{t\geq 1}$ is $\sqrt{1/M}$-sub-Gaussian and unit-norm.*

3. Conclusions

This research has successfully established the first probabilistic framework specifically conceived for the domain of sequential random projection, addressing the complexities and challenges introduced by sequential dependencies and the high-dimensional nature of variables within sequential decision-making processes. By innovating a stopped process construction and extending the method of mixtures to include a self-normalized process, we have derived non-asymptotic probability bounds that significantly extend the Johnson–Lindenstrauss lemma into the realm of sequential analysis. These methodological advancements not only provide a robust foundation for accurately controlling concentration events and analyzing error bounds in a sequential context but also represent a seminal contribution to the intersection of computational mathematics and machine learning.

Our contributions offer a comprehensive solution to the analytical hurdles posed by the sequential dependence inherent in dynamic environments, specifically the loss of independence and identical distribution characteristics among projection vectors conditioned on sequential decisions. By addressing these challenges with precise technical innovations and extending foundational analytical tools to sequential settings, our work paves the way for future research and practical applications of random projection in sequential decision-making settings. In doing so, it heralds a paradigm shift towards a more nuanced understanding and application of random projection techniques in adaptive, high-dimensional data processing, promising to significantly influence future research directions and applications across related disciplines.

Appendix A. Technical ideas & details

Before digging into the proof, we identify some important sequential structure and also clarity our proof idea in a intuitive level. For each time $t \in \mathcal{T}$, let the short notation for the centered variable be

$$Y_t = \|\sum_{i=0}^{t} x_i \mathbf{z}_i\|^2 - \sum_{i=0}^{t} x_i^2 \quad \text{and} \quad S_t = \sum_{i=0}^{t} x_i^2. \tag{5}$$

Our *key observation* is that for any $t \in [T]$

$$\begin{aligned} \|\sum_{i=0}^{t} x_i \mathbf{z}_i\|^2 &= \|\sum_{i=0}^{t-1} x_i \mathbf{z}_i + x_t \mathbf{z}_t\|^2 \\ &= \|\sum_{i=0}^{t-1} x_i \mathbf{z}_i\|^2 + 2\left(\sum_{i=0}^{t-1} x_i \mathbf{z}_i\right)^\top x_t \mathbf{z}_t + x_t^2 \|\mathbf{z}_t\|^2 \end{aligned} \tag{6}$$

and thus we have the following relationship between Y_t and Y_{t-1} derived from eq. (6),

$$Y_t - Y_{t-1} = 2x_t \mathbf{z}_t^\top (\sum_{i=0}^{t-1} x_i \mathbf{z}_i) + x_t^2 \left(\|\mathbf{z}_t\|^2 - 1\right).$$

Since $\mathbf{z}_t$ is unit-norm, we can further simplify the exposition

$$Y_t - Y_{t-1} = 2x_t \mathbf{z}_t^\top (\sum_{i=0}^{t-1} x_i \mathbf{z}_i). \tag{7}$$

Another key observation is that the difference term in eq. (7) is a function of on the $(\sum_{i=0}^{t-1} x_i \mathbf{z}_i)$ that is $\mathcal{F}_{t-1}$-measurable. This implies, the difference term $(Y_t - Y_{t-1})$ can be controlled according to information in the history-dependent term

$$\sum_{i=0}^{t-1} x_i \mathbf{z}_i = Y_{t-1} + S_{t-1}.$$

Intuitively, once the concentration behavior is bad, i.e., Y_{t-1} has large deviation, it is highly possible to exhibit large deviation for $Y_{t'}$ in the later time index $t' \geq t$.

A.1. Stopped process and exponential supermartingale

To mathematically formalize this intuition, we introduce a definition of good event for concentration behavior and stopping time for analysis.

Definition 9 (Good event) *For each time $t \in \mathcal{T}$, we introduce the good event E_t under which the strongly concentration behavior is guaranteed, suppose $\varepsilon \in (0, 1)$,*

$$E_t(\varepsilon) = \left\{ (1-\varepsilon)\left(\sum_{i=0}^{t} x_i^2\right) \leq \|\sum_{i=0}^{t} x_i \mathbf{z}_i\|^2 \leq (1+\varepsilon)\left(\sum_{i=0}^{t} x_i^2\right) \right\}. \tag{8}$$

With short notation defined in eq. (5),

$$E_t(\varepsilon) = \{|Y_t| \leq \varepsilon S_t\}.$$

We also define the stopping time as the first time the bad event happens, i.e. the good event $E_t(\varepsilon)$ defined in eq. (8) violates.

Definition 10 (Stopping time) *For any fixed ε, we define the stopping time*

$$\tau(\varepsilon) = \min\{t \in \mathcal{T} : \neg E_t(\varepsilon)\}. \tag{9}$$

Based on the stopping time, we construct a **stopped process** to interconnect the sequence of good concentration event. For $t \in [T]$, define the stopped difference term

$$X_t^\tau = (Y_t - Y_{t-1})\mathbb{1}_{t\le\tau} \tag{10}$$

such that the process $(X_t^\tau)_{t\ge1}$ is adapted to the filtration $(\mathcal{F}_t)_{t\ge1}$.

Claim 11 *Let τ be the stopping time $\tau(\varepsilon)$ defined in eq. (9). Let $(X_t^\tau)_{t\ge1}$ be the stochastic process defined in eq. (10) which is adapted to the filtration $(\mathcal{F}_t)_{t\ge1}$. Let $A_t^\tau = \sum_{i=1}^t X_i^\tau$. Further denote $(B_t^\tau)^2 = \sum_{i=1}^t (C_i^\tau)^2$ with*

$$(C_t^\tau)^2 := \frac{4c_0}{M} x_t^2 (1+\varepsilon) S_{t-1} \mathbb{1}_{t\le\tau}.$$

If the $(\mathcal{F}_t)_{t\ge1}$-adapted process $(\mathbf{z}_t)_{t\ge1}$ is $\sqrt{c_0/M}$-sub-Gaussian and each $\mathbf{z}_t$ is unit-norm, then for any fixed $\lambda \in \mathbb{R}$

$$\left\{ M_t^\tau(\lambda) = \exp\left(\lambda A_t^\tau - \frac{\lambda^2}{2}(B_t^\tau)^2\right), \mathcal{F}_t, t \ge 1 \right\}$$

*is a **supermartingale** with mean ≤ 1.*

Proof [Proof of theorem 11] Note $\mathbb{1}_{t\le\tau} = 1 - \mathbb{1}_{\tau\le t-1}$ is $\mathcal{F}_{t-1}$-measurable. Thus, the random vector

$$(\sum_{i=0}^{t-1} x_i \mathbf{z}_i)\mathbb{1}_{t\le\tau} x_t \quad \text{is } \mathcal{F}_{t-1}\text{-measurable.}$$

By the condition that the process $(\mathbf{z}_t)_{t\ge1}$ is $\sqrt{c_0/M}$-sub-Gaussian, we conclude from the definition of conditionally sub-Gaussian in definition 2,

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X_t^\tau) \mid \mathcal{F}_{t-1}] &= \mathbb{E}[\exp(2\lambda x_t \langle \mathbf{z}_t, \sum_{i=0}^{t-1} x_i \mathbf{z}_i\rangle \mathbb{1}_{t\le\tau}) \mid \mathcal{F}_{t-1}] \\
&\le \exp\left(\frac{\lambda^2}{2}(4c_0/M) x_t^2 \| \sum_{i=0}^{t-1} x_i \mathbf{z}_i\|^2 \mathbb{1}_{t\le\tau}\right) \\
&\le \exp\left(\frac{\lambda^2}{2}(4c_0/M) x_t^2 (1+\varepsilon) S_{t-1} \mathbb{1}_{t\le\tau}\right) \\
&= \exp\left(\frac{\lambda^2}{2}(C_t^\tau)^2\right)
\end{aligned} \tag{11}$$

where the last inequality is because of the stopping time argument. Thus, the claim holds as

$$\mathbb{E}[M_t^\tau(\lambda) \mid \mathcal{F}_{t-1}] = M_{t-1}^\tau(\lambda) \mathbb{E}[\exp(\lambda X_t^\tau - \frac{\lambda^2}{2}(C_t^\tau)^2) \mid \mathcal{F}_{t-1}] \le M_{t-1}^\tau(\lambda),$$

where the inequality is due to eq. (11). ■

The following de la Peña et al. [6]-type self-normalized bound would be useful to prove our main theoretical contribution of sequential random projection in theorem 6.

Theorem 12 (Any-time self-normalized concentration bound)
Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration and $\{(A_t, B_t), t \geq 1\}$ be a sequence of pairs of random variables satisfying that for all $\lambda \in \mathbb{R}$

$$\left\{\exp\left(\lambda A_t - \frac{\lambda^2}{2}B_t^2\right), \mathcal{F}_t, t \geq 1\right\} \text{ is a supermartingale with mean} \leq 1.$$

Then, for any fixed positive sequence $(L_t)_{t\geq 1}$, with probability at least $1-\delta$

$$\forall t \geq 1, \quad |A_t| \leq \sqrt{2\left(B_t^2 + L_t\right)\log\left(\frac{1}{\delta}\frac{\left(B_t^2+L_t\right)^{1/2}}{L_t^{1/2}}\right)}$$

The proof of theorem 12 can be found in section B.

We also need the following trigger lemma for the initial preparation of the proof in theorem 6.

Lemma 13 (Trigger lemma) *For any sequence of event $(\mathcal{E}_t, t \in \mathcal{T})$, define the stopping time τ as the first time t the event $\mathcal{E}_t$ is violated, i.e.*

$$\tau = \min\{t \in \mathcal{T} : \neg\mathcal{E}_t\}.$$

Then, the following equality holds for all $t \in \mathcal{T}$,

$$\{\tau \leq t\} = \neg\mathcal{E}_{t\wedge\tau}.$$

A.2. Proof of theorem 6

Now we are ready to provide the details of the proof, completing the intuition and mathematical construction.

Proof [Proof of theorem 6] We apply lemma 13 for $\mathcal{E}_t = E_t(\varepsilon)$ and it follows

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathcal{T}, \neg E_t(\varepsilon)\right) &= \mathbb{P}(\tau \leq T) \\
&= \mathbb{P}\left(\neg E_{T\wedge\tau}(\varepsilon)\right) \\
&= \mathbb{P}\left(|Y_{T\wedge\tau}| \geq \varepsilon S_{T\wedge\tau}\right) \\
&= \mathbb{P}\left(|Y_0 + \sum_{t=1}^{T}(Y_t - Y_{t-1})\mathbb{1}_{t\leq\tau}| \geq \varepsilon S_{T\wedge\tau}\right) \qquad (12)
\end{aligned}$$

By the construction of stopped process $Y_{T\wedge\tau} - Y_0 = \sum_{t=1}^{T} X_t^\tau = A_T^\tau$. Then, *our goal*, from eq. (12), becomes to upper bound the RHS of eq. (13),

$$\mathbb{P}\left(\exists t \in \mathcal{T}, (\neg E_t)\right) = \mathbb{P}\left(|Y_0 + A_T^\tau| \geq \varepsilon S_{T\wedge\tau}\right) \qquad (13)$$

By theorem 11, the pair of processes $(A_t^\tau, B_t^\tau)_{t\geq 1}$ with

$$A_t^\tau = \sum_{i=1}^{t} X_i^\tau = \sum_{i=1}^{t} (Y_t - Y_{t-1}) \mathbb{1}_{t\leq\tau}$$

and

$$(B_t^\tau)^2 = \sum_{i=1}^{t} (4c_0/M) x_t^2 (1+\varepsilon) S_{t-1} \mathbb{1}_{t\leq\tau}$$

satisfies the conditions in theorem 12. Then applying the theorem 12 on the pair of processes $(A_t^\tau, B_t^\tau)_{t\geq 1}$ yields that, with probability at least $1-\delta$,

$$\forall t \geq 1, |A_t^\tau| \leq \sqrt{2\left((B_t^\tau)^2 + L_t\right) \log\left(\frac{1}{\delta}\frac{\left((B_t^\tau)^2 + L_t\right)^{1/2}}{L_t^{1/2}}\right)}$$

Since by the condition in theorem 6, we have $|Y_0| \leq (\varepsilon/2)x_0^2$. Now we want to argue that for any fixed $\varepsilon \in (0,1)$, with suitable choice of L_T and M, we have with probability at least $1-\delta$

$$|Y_0 + A_T^\tau| \leq \underbrace{\sqrt{2\left((B_T^\tau)^2 + L_T\right) \log\left(\frac{1}{\delta}\frac{\left((B_T^\tau)^2 + L_T\right)^{1/2}}{L_T^{1/2}}\right)} + (\varepsilon/2)x_0^2}_{(I)} \leq \varepsilon S_{T\wedge\tau}. \tag{14}$$

Claim 14 *The following configuration suffices for eq. (14):*

$$L_T \leq \frac{2c_0(1+\varepsilon)x_0^4}{M} \quad \textit{and} \quad M \geq (16c_0(1+\varepsilon)/\varepsilon^2)\left(\log\left(\frac{1}{\delta}\right) + \log\left(1 + \frac{c_x T}{x_0^2}\right)\right).$$

Proof [Proof of theorem 14] Recall the definition $S_t = \sum_{t=0}^{t} x_i^2$. We first calculate the term $(B_T^\tau)^2$ by our construction,

$$\begin{aligned}
(B_T^\tau)^2 &\leq \frac{4c_0}{M}\sum_{t=1}^{T\wedge\tau} x_t^2\left((1+\varepsilon)S_{t-1}\right) &(15)\\
&= \frac{4c_0(1+\varepsilon)}{M}\sum_{t=1}^{T\wedge\tau} x_t^2\left(S_{T\wedge\tau} - \underbrace{(S_{T\wedge\tau} - S_{t-1})}_{\geq 0}\right)\\
&\leq \frac{4c_0(1+\varepsilon)}{M}(S_{T\wedge\tau} - x_0^2)S_{T\wedge\tau}. &(16)
\end{aligned}$$

From eq. (15), the almost sure upper bound of $(B_T^\tau)^2$ assuming $x_t^2 \leq c_x$ is

$$(B_T^\tau)^2 \leq \frac{4c_0(1+\varepsilon)}{M}\sum_{t=1}^{T} c_x(x_0^2 + (t-1)c_x) \leq \frac{4c_0(1+\varepsilon)}{M}(c_x x_0^2 T + c_x^2 T^2/2)$$

Since $(a+b)^2 \leq (1+\lambda)(a^2+(1/\lambda)b^2)$ for all fixed $\lambda \geq 0$, the term (I) from eq. (14) becomes

$$(I)^2 \leq (1+\lambda)\left(2\left((B_T^\tau)^2+L_T\right)\log\left(\frac{1}{\delta}\frac{\left((B_T^\tau)^2+L_T\right)^{1/2}}{L_T^{1/2}}\right)+\frac{\varepsilon^2 x_0^4}{4\lambda}\right)$$

Let $L_T \leq 4c_0(1+\varepsilon)\ell/M$ and ℓ to be determined.

$$\begin{aligned}(I)^2 &\leq (1+\lambda)\left(2\left(B_T^2+L_T\right)\log\left(\frac{1}{\delta}\frac{\left(B_T^2+L_T\right)^{1/2}}{L_T^{1/2}}\right)+\frac{\varepsilon^2 x_0^4}{4\lambda}\right)\\ &\leq (1+\lambda)\left(\frac{8c_0(1+\varepsilon)}{M}\left((S_{T\wedge\tau}-x_0^2)S_{T\wedge\tau}+\ell\right)\log\left(\frac{1}{\delta}\sqrt{\frac{\left(c_x x_0^2 T+c_x^2T^2/2+\ell\right)}{\ell}}\right)+\frac{\varepsilon^2 x_0^4}{4\lambda}\right)\end{aligned}$$

Let $M \geq (8c_0(1+\varepsilon)/m)\log\left(\frac{1}{\delta}\sqrt{\frac{c_x x_0^2T+c_x^2T^2/2+\ell}{\ell}}\right)$ and m to be determined, we can simplify

$$(I)^2 \leq (1+\lambda)\left(m((S_{T\wedge\tau}-x_0^2)S_{T\wedge\tau}+\ell)+\frac{\varepsilon^2 x_0^4}{4\lambda}\right)$$

Let $\ell = x_0^4/2$, $m = \varepsilon^2/(1+\lambda)$ and $\lambda = 1$, we have

$$(I)^2 \leq \varepsilon^2((S_{T\wedge\tau}-x_0^2)S_{T\wedge\tau}+x_0^4/2+x_0^4/2) \leq \varepsilon^2 S_{T\wedge\tau}^2$$

where the last inequality is due to $x_0^2 = S_0 \leq S_{T\wedge\tau}$ and $x_0^4 \leq x_0^2 S_{T\wedge\tau}$. The conclusion is that we could select

$$\begin{aligned}M &\geq (16c_0(1+\varepsilon)/\varepsilon^2)\log\left(\frac{1}{\delta}\sqrt{\frac{2c_x x_0^2T+c_x^2T^2+x_0^4}{x_0^4}}\right)\\ &= (16c_0(1+\varepsilon)/\varepsilon^2)\log\left(\frac{1}{\delta}\sqrt{\frac{(c_xT+x_0^2)^2}{x_0^4}}\right)\\ &= (16c_0(1+\varepsilon)/\varepsilon^2)\left(\log\left(\frac{1}{\delta}\right)+\log\left(1+\frac{c_xT}{x_0^2}\right)\right)\end{aligned}$$

and the auxiliary variable

$$L_T \leq \frac{2c_0(1+\varepsilon)x_0^4}{M}.$$

■

■

Appendix B. Proof of theorem 12: Method of mixtures

Robbins-Siegmund method of mixtures [16] originally is developed to evaluate boundary crossing probabilities for Brownian motion. The method was further developed in the general theory for self-normalized process [5, 6, 10].

Remark 15 (Essential idea of Laplace approximation) *If we integrate the exponential of a function that has a pronounced maximum, then we can expect that the integral will be close to the exponential function of the maximum. In our case, let*

$$M_t(\lambda) = \exp\left(\lambda A_t - \frac{\lambda^2}{2} B_t^2\right)$$

Informally, with this principle of Laplace approximation, we would have

$$\max_{\lambda} M_t(\lambda) \approx \int_{\Omega} M_t(\lambda) dh(\lambda)$$

where h is some measure on Ω.

The main benefit of replacing the maximum $\max_\lambda M_t(\lambda)$ with an integral $\bar{M}_t := \int_\Omega M_t(\lambda) dh(\lambda)$ is that we can handle the expectation $\mathbb{E}[\bar{M}_t]$ easier while we don't know the upper bound on $\mathbb{E}[\max_\lambda M_t(\lambda)]$. This is formalized in the following lemma.

Lemma 16 *Let (h_t) be a sequence of probability measures on Ω. If $(M_t(\lambda), \mathcal{F}_t, t \geq 1)$ is a supermartingale with $\mathbb{E}[M_1(\lambda)] \leq 1$ for all $\lambda \in \Omega$, then for any $t \geq 1$, the integrated random variable $\bar{M}_t = \int_\Omega M_t(\lambda) dh_t(\lambda)$ has expectation $\mathbb{E}[\bar{M}_t] \leq 1$.*

Further, let τ be a stopping time with respect to filtration $(\mathcal{F}_t)_{t\geq 0}$, i.e. $\{\tau \leq t\} \in \mathcal{F}_t, \forall t \geq 0$. Then $M_\tau(\lambda)$ is almost surely well-defined with expectation $\mathbb{E}[M_\tau(\lambda)] \leq 1$ as well as $\mathbb{E}[\bar{M}_\tau] \leq 1$.

Proof Using Fubini's theorem and the fact that $M_t(\lambda)$ is a supermartingale with $\mathbb{E}[M_t(\lambda)] \leq \mathbb{E}[M_1(\lambda)] = 1$, we have

$$\mathbb{E}[\bar{M}_t] = \int_\Omega \mathbb{E}[M_t(\lambda)] dh_t(\lambda) \leq 1.$$

For the expectation of stopped version $M_\tau(\lambda)$ and $\bar{M}_\tau$, we apply (supermartingale) optional sampling theorem. ■

Finally, we are comfortable to drive the proof of the self-normalized concentration bounds.

Proof [Proof of theorem 12] Let $\Lambda = (\Lambda_t)$ be a sequence of independent Gaussian random variable with densities

$$f_{\Lambda_t}(\lambda) = c(L_t) \exp(-\frac{1}{2} L_t \lambda^2)$$

where $c(A) = \sqrt{A/2\pi}$ is a normalizing constant. We explicitly calculate $\bar{M}_t$ for any $t \geq 1$,

$$\begin{aligned}
\bar{M}_t &= \int_{\mathbb{R}} \exp\left(\lambda A_t - \frac{\lambda^2}{2} B_t^2\right) f_{\Lambda_t}(\lambda) d\lambda \\
&= \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\lambda - \frac{A_t}{B_t^2}\right)^2 B_t^2 + \frac{1}{2}\frac{A_t^2}{B_t^2}\right) f_{\Lambda_t}(\lambda) d\lambda \\
&= \exp\left(\frac{1}{2}\frac{A_t^2}{B_t^2}\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\lambda - \frac{A_t}{B_t^2}\right)^2 B_t^2\right) f_{\Lambda_t}(\lambda) d\lambda \\
&= c\left(L_t\right) \exp\left(\frac{1}{2}\frac{A_t^2}{B_t^2}\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\left(\lambda - A_t/B_t^2\right)^2 B_t^2 + \lambda^2 L_t\right)\right) d\lambda.
\end{aligned}$$

Completing the square yields

$$\left(\lambda - \frac{A_t}{B_t^2}\right)^2 B_t^2 + \lambda^2 L_t = \left(\lambda - \frac{A_t}{L_t + B_t^2}\right)^2 \left(L_t + B_t^2\right) + \frac{A_t^2}{B_t^2} - \frac{A_t^2}{L_t + B_t^2}.$$

By the change of variables $\lambda' = \lambda - A_t/(L_t + B_t^2)$ in the following (i),

$$\begin{aligned}
\bar{M}_t &= c\left(L_t\right) \exp\left(\frac{1}{2}\frac{A_t^2}{L_t + B_t^2}\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\lambda - \frac{A_t}{L_t + B_t^2}\right)^2 \left(L_t + B_t^2\right)\right) d\lambda \\
&\overset{(i)}{=} c\left(L_t\right) \exp\left(\frac{1}{2}\frac{A_t^2}{L_t + B_t^2}\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\lambda^2 \left(L_t + B_t^2\right)\right)\right) d\lambda \\
&= \frac{c\left(L_t\right)}{c\left(L_t + B_t^2\right)} \exp\left(\frac{1}{2}\frac{A_t^2}{L_t + B_t^2}\right).
\end{aligned}$$

A final application of Markov's inequality yields

$$\begin{aligned}
&\mathbb{P}\left[|A_\tau| \geq \sqrt{2\left(L_\tau + B_\tau^2\right) \log\left(\frac{1}{\delta}\frac{\left(L_\tau + B_\tau^2\right)^{1/2}}{L_\tau^{1/2}}\right)}\right] \\
&= \mathbb{P}\left[\frac{c\left(L_\tau\right)}{c\left(L_\tau + B_\tau^2\right)} \exp\left(\frac{1}{2}\frac{A_\tau^2}{L_\tau + B_\tau^2}\right) \geq \frac{1}{\delta}\right] \\
&\leq \delta \cdot \mathbb{E}\left[\frac{c\left(L_\tau\right)}{c\left(L_\tau + B_\tau^2\right)} \exp\left(\frac{1}{2}\frac{A_\tau^2}{L_\tau + B_\tau^2}\right)\right] \\
&\overset{\text{i)}}{\leq} \delta \cdot \mathbb{E}\left[\bar{M}_\tau\right] \overset{\text{(ii)}}{\leq} \delta,
\end{aligned}$$

where (i) uses the inequality for $\bar{M}_\tau$ derived above, and (ii) follows from lemma 16.

To get the anytime result in theorem 12, we define the stopping time

$$\tau = \min\left\{t \geq 1 : |A_t| \geq \sqrt{2\left(L_t + B_t^2\right) \log\left(\frac{1}{\delta}\frac{\left(L_t + B_t^2\right)^{1/2}}{L_t^{1/2}}\right)}\right\}$$

With an application of extended version of lemma 13, and applying the previous inequality yields

$$\begin{aligned}
&\mathbb{P}\left[\exists t \geq 1, |A_t| \geq \sqrt{2\left(L_t + B_t^2\right)\log\left(\frac{1}{\delta}\frac{\left(L_t + B_t^2\right)^{1/2}}{L_t^{1/2}}\right)}\right] \\
&= \mathbb{P}\left[\tau < \infty, |A_\tau| \geq \sqrt{2\left(L_\tau + B_\tau^2\right)\log\left(\frac{1}{\delta}\frac{\left(L_\tau + B_\tau^2\right)^{1/2}}{L_\tau^{1/2}}\right)}\right] \\
&\leq \mathbb{P}\left[|A_\tau| \geq \sqrt{2\left(L_\tau + B_\tau^2\right)\log\left(\frac{1}{\delta}\frac{\left(L_\tau + B_\tau^2\right)^{1/2}}{L_\tau^{1/2}}\right)}\right] \\
&\leq \delta.
\end{aligned}$$

This completes the proof. ■

Appendix C. Additional lemmas

For the completeness, we provide the full details of lemma 4, which is adapted from [11].

Proof We utilize the order-2 recurrence for central moments [17]: for a beta random varaiable $X \sim \mathrm{Beta}(\alpha, \beta)$, we have

$$\begin{aligned}
\mathbb{E}\left[(X - \mathbb{E}[X])^p\right] =& \frac{(p-1)(\beta-\alpha)}{(\alpha+\beta)(\alpha+\beta+p-1)} \cdot \mathbb{E}\left[(X - \mathbb{E}[X])^{p-1}\right] \\
&+ \frac{(p-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+p-1)} \cdot \mathbb{E}\left[(X - \mathbb{E}[X])^{p-2}\right]
\end{aligned}$$

Let $m_p := \frac{\mathbb{E}[(X-\mathbb{E}[X])^p]}{p!}$, when $\alpha \geq \beta$, it follows that m_p is non-negative when p is even, and negative otherwise. Thus, for even p,

$$m_p \leq \frac{1}{p} \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+p-1)} m_{p-2} \leq \frac{\mathrm{Var}\left(X\right)}{p} \cdot m_{p-2}.$$

After repeating the above recursive equation for $p/2$ times and combining with $m_p \leqslant 0$ for odd p, it yields the following relationships

$$m_p \leqslant \begin{cases} \frac{\mathrm{Var}(X)^{\frac{p}{2}}}{p!!} & p \text{ even} \\ 0 & p \text{ odd} \end{cases}.$$

With the application of $p!! = 2^{p/2}(p/2)!$ for even p, for $t \geqslant 0$ we obtain

$$\mathbb{E}[\exp(\lambda[X - \mathbb{E}[X]])] \leqslant 1 + \sum_{p=2}^{+\infty} m_p \lambda^p = 1 + \sum_{p=1}^{+\infty} (\lambda^2 \mathrm{Var}\left(X\right)/2)^p / p! = \exp\left(\frac{\lambda^2 \mathrm{Var}\left(X\right)}{2}\right)$$

■

Lemma 17 *For any fixed unit vector $u \in \mathbb{S}^{n-1}$, for any random vector $v \sim \mathcal{U}(\mathbb{S}^{n-1})$, the inner product $u^\top v$ is distributed as* $2\,\mathrm{Beta}\left(\frac{n-1}{2}, \frac{n-1}{2}\right) - 1$.

Proof By rotational invariance of $\mathrm{Uniform}(\mathbb{S}^{n-1})$, the distribution of $u^\top v$ should be identical $\forall u \in \mathbb{S}^{n-1}$. WLOG, let us look at $u = e_1 = (1, 0, \ldots, 0)$ that would project v to the first coordinate, i.e., the value of $v^\top e_1 = v_1$. Let v_1 be defined as X, which is a random variable. Probability density that $X = x \in [-1, 1]$ is proportional to the surface area occupied between x and $x + dx$ occupied by the other coordinates. The surface area is a frustum of a cone with base as a $n-1$ dimension shell of radius as $\sqrt{1-x^2}$ and a height of dx with slope of the cone as $1/\sqrt{1-x^2}$. Hence, the probability density is

$$f_X(x) \propto \frac{\sqrt{1-x^2}^{n-2}}{\sqrt{1-x^2}} \propto (1-x)^{\frac{n-3}{2}}(1+x)^{\frac{n-3}{2}}$$

We examine the transformed random variable $Y = (X+1)/2$, i.e. $X = 2Y - 1$. By the Change-of-Variable Technique,

$$f_Y(y) = 2f_X(2y-1) \propto (2-2y)^{\frac{n-3}{2}}(2y)^{\frac{n-3}{2}} \propto (1-y)^{\frac{n-3}{2}} y^{\frac{n-3}{2}}$$

Thus,

$$Y \sim \mathrm{Beta}\left(\frac{n-1}{2}, \frac{n-1}{2}\right).$$

Then, by the fact $X = 2Y - 1$,

$$X \sim 2\,\mathrm{Beta}\left(\frac{n-1}{2}, \frac{n-1}{2}\right) - 1.$$

■

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[2] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive approximation*, 28:253–263, 2008.

[3] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[4] Michael B Cohen, TS Jayram, and Jelani Nelson. Simple analyses of the sparse johnson-lindenstrauss transform. In *1st Symposium on Simplicity in Algorithms (SOSA 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[5] Victor H de la Peña, Michael J Klass, and Tze Leung Lai. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability*, pages 1902–1933, 2004.

[6] Victor H de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

[7] Piotr Indyk. Algorithmic applications of low-distortion embeddings. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, page 1, 2001.

[8] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[9] Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

[10] Tze Leung Lai. Martingales in sequential analysis and time series, 1945-1985. In *Electronic Journal for history of probability and statistics*, 2009.

[11] Yingru Li. Simple, unified analysis of johnson-lindenstrauss with applications, 2024. URL https://arxiv.org/abs/2402.10232.

[12] Yingru Li, Jiawei Xu, Lei Han, and Zhi-Quan Luo. Q-star meets scalable posterior sampling: Bridging theory and practice via hyperagent. In *Forty-first International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.10228.

[13] Ziniu Li, Yingru Li, Yushun Zhang, Tong Zhang, and Zhi-Quan Luo. HyperDQN: A randomized exploration method for deep reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=X0nrKAXu7g-.

[14] Jiří Matoušek. On variants of the johnson–lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

[15] Shanmugavelayutham Muthukrishnan et al. Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2):117–236, 2005.

[16] Herbert Robbins and David Siegmund. Boundary crossing probabilities for the wiener process and sample sums. *The Annals of Mathematical Statistics*, pages 1410–1429, 1970.

[17] Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.

[18] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, page 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

[19] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.