

# Concentration Inequalities for Multinoulli Random Variables

**Jian Qian**

*Sequel Team - Inria Lille*

JIAN.QIAN@ENS.FR

**Ronan Fruit**

*Sequel Team - Inria Lille*

RONAN.FRUIT@INRIA.FR

**Matteo Pirotta**

*Sequel Team - Inria Lille*

MATTEO.PIROTTA@INRIA.FR

**Alessandro Lazaric**

*Facebook AI Research*

LAZARIC@FB.COM

**Editor:** XXXX

## 1. Problem Formulation

We analyse the concentration properties of the random variable  $Z_n \geq 0$  defined as:

$$Z_n := \max_{v \in [0, D]^S} \left\{ (\hat{p}_n - p)^\top v \right\} \quad (1)$$

where  $\hat{p}_n \in \Delta^S$  is a random vector,  $p \in \Delta^S$  is deterministic and  $\Delta^S = \{x \in \mathbb{R}^S : \sum_{i=1}^S x_i = 1 \wedge x_i \geq 0\}$  is the  $(S-1)$ -dimensional simplex. It is easy to show that the maximum in Eq. 1 is equivalent to computing the (scaled)  $\ell_1$ -norm of the vector  $\hat{p}_n - p$ :

$$Z_n = \max_{u \in [-\frac{D}{2}, \frac{D}{2}]} \left\{ (\hat{p}_n - p)^\top \left( u + \frac{D}{2} e \right) \right\} = \frac{D}{2} \|\hat{p}_n - p\|_1 \quad (2)$$

where we have used the fact that  $\frac{D}{2}(\hat{p}_n - p)^\top e = 0$ . As a consequence,  $Z_n$  is a bounded random variable in  $[0, D]$ . While the following discussion apply to Dirichlet distributions, we focus on  $\hat{p}_n \sim \frac{1}{n} \text{Multinomial}(n, p)$ . The results previously available in the literature are summarized in the following.

The literature has analysed the concentration of the  $\ell_1$ -discrepancy of the true distribution and the empirical one in this setting.

**Proposition 1** (*Weissman et al., 2003*) *Let  $p \in \Delta^S$  and  $\hat{p} \sim \frac{1}{n} \text{Multinomial}(n, p)$ . Then, for any  $S \geq 2$  and  $\delta \in [0, 1]$ :*

$$\mathbb{P} \left( \|\hat{p} - p\|_1 \geq \sqrt{\frac{2S \ln(2/\delta)}{n}} \right) \leq \mathbb{P} \left( \|\hat{p} - p\|_1 \geq \sqrt{\frac{2 \ln((2^S - 2)/\delta)}{n}} \right) \leq \delta \quad (3)$$

This concentration inequality is at the core of the proof of UCRL, see (Jaksch et al., 2010, App. C.1). Another inequality is provided in (Devroye, 1983, Lem. 3).

**Proposition 2** (Devroye, 1983) Let  $p \in \Delta^S$  and  $\hat{p} \sim \frac{1}{n} \text{Multinomial}(n, p)$ . Then, for any  $0 \leq \delta \leq 3 \exp(-4S/5)$ :

$$\mathbb{P} \left( \|\hat{p}_n - p\|_1 \geq 5 \sqrt{\frac{\ln(3/\delta)}{n}} \right) \leq \delta \quad (4)$$

While Prop. 1 shows an explicit dependence on the dimension of the random variable, such dependence is hidden in Prop. 2 by the constraint on  $\delta$ . Note that for any  $0 \leq \delta \leq 3 \exp(-4S/5)$ ,  $\sqrt{\frac{\ln(3/\delta)}{n}} > \sqrt{\frac{4S}{5n}}$ . This shows that the  $\ell_1$ -deviation always scales proportionally to the dimension of the random variable, i.e., as  $\sqrt{S}$ .

*A better inequality.* The natural question is whether is possible to derive a concentration inequality independent from the dimension of  $p$  by exploiting the correlation between  $\hat{p}$  and the maximizer vector  $v^*$ . This question has been recently addressed in (Agrawal and Jia, 2017, Lem. C.2):

**Lemma 3** (Agrawal and Jia, 2017) Let  $p \in \Delta^S$  and  $\hat{p} \sim \frac{1}{n} \text{Multinomial}(n, p)$ . Then, for any  $\delta \in [0, 1]$ :

$$\mathbb{P} \left( \|\hat{p}_n - p\|_1 \geq \sqrt{\frac{2 \ln(1/\delta)}{n}} \right) \leq \delta$$

Their results resemble the one in Lem. 2 but removes the constraint on  $\delta$ . As a consequence, the implicit or explicit dependence on the dimension  $S$  is removed. In the following, we will show (empirically and theoretically) that Lem. 3 may not be correct.

## 2. Theoretical Analysis (the asymptotic case)

In this section, we provide a counter-argument to the Lem. 3 in the asymptotic regime (i.e.,  $n \rightarrow +\infty$ ). The overall idea is to show that the expected value of  $Z_n$  asymptotically grows as  $O(\sqrt{S})$  and  $Z_n$  itself is well concentrated around its expectation. As a result, we can deduce that all quantiles of  $Z_n$  grow as  $O(\sqrt{S})$  as well.

We consider the true vector  $p$  to be uniform, i.e.,  $p = (\frac{1}{S}, \dots, \frac{1}{S})$  and  $\hat{p} \sim \frac{1}{n} \text{Multinomial}(n, p)$ .<sup>1</sup> The following lemma provides a characterization of the variable  $Z_S := \lim_{n \rightarrow +\infty} \sqrt{n} Z_n$ .

**Lemma 4** Consider  $S \in \mathbb{N}$ ,  $\mathcal{S} = \{1, \dots, S\}$  and  $p = (\frac{1}{S}, \dots, \frac{1}{S})$  be the uniform distribution on  $\mathcal{S}$ . Let  $e_S$  be the vector of ones of dimension  $S$ . Define  $Y \sim \mathcal{N}(0, I_S - \frac{1}{S-1} N)$  where  $N = e_S e_S^T - I_S$  is the matrix with 0 in all the diagonal entry and 1 elsewhere, and  $Y^+ = (\max(Y_i, 0))_{i \in \mathcal{S}}$ . Then:

$$Z_S = \lim_{n \rightarrow +\infty} \sqrt{n} Z_n \sim \|Y^+\|_1 D \sqrt{\frac{S-1}{S^2}}.$$

Furthermore,

$$\mathbb{E}[Z_S] = \sqrt{\frac{S-1}{S^2}} \cdot \mathbb{E} \left[ \sum_{i=1}^S Y_i^+ \right] = \sqrt{S-1} \cdot \mathbb{E}[Y_1^+] = \sqrt{\frac{S-1}{2\pi}}.$$

1. The analysis holds also in the case  $\hat{p} \sim \text{Dirchlet}(np)$ .

While the previous lemma may already suggest that  $Z_S$  should grow as  $O(\sqrt{S})$  as its expectation, it is still possible that a large part of the distribution is concentrated around a value independent from  $S$ , with limited probability assigned to, e.g., values growing as  $O(S)$ , which could justify the  $O(\sqrt{S})$  growth of the expectation. Thus, in order to conclude the analysis, we need to show that  $Z_S$  is concentrated “enough” around its expectation.

Since the random variables  $Y_i$  are correlated, it is complicated to directly analyze the deviation of  $Z_S$  from its mean. Thus we first apply an orthogonal transformation on  $Y$  to obtain independent r.v. (recall that jointly normally distributed variables are independent if uncorrelated).

**Lemma 5** *Consider the same settings of Lem. 4 and recall that  $Y \sim \mathcal{N}(0, I_S - \frac{1}{S-1}N)$ . There exists an orthogonal transformation  $U \in O_S(\mathbb{R})$ , s.t.*

$$W = \sqrt{\frac{S-1}{S}}UY \sim \mathcal{N}\left(0, \begin{bmatrix} I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}\right).$$

By exploiting the transformation  $U$  we can write that  $Z_S \sim g(W) := \frac{1}{\sqrt{S}}e_S^\top (U^\top W)^+$ . Since  $W_i$  are i.i.d. standard Gaussian random variables and  $g$  is 1-Lipschitz, we can finally characterize the mean and the deviations of  $Z_S$  and derive the following anticoncentration inequality for  $Z_S$ .

**Theorem 6** *Let  $p \in \Delta^S = (\frac{1}{S}, \dots, \frac{1}{S})$  and  $\hat{p}_n \sim \frac{1}{n} \text{Multinomial}(n, p)$ . Define  $Z_n = \max_{v \in [0, D]} \{(\hat{p}_n - p)^\top v\}$  and  $Z_S = \lim_{n \rightarrow +\infty} \sqrt{n}Z_n$ . Then, for any  $\delta \in (0, 1)$ :*

$$\mathbb{P}\left[Z_S \geq \sqrt{\frac{2(S-1)}{\pi}} - \sqrt{2 \log(2/\delta)}\right] \geq 1 - \delta.$$

This result shows that every quantile of  $Z_S$  is dependent on the dimension of the random variable, i.e.,  $\sqrt{S}$ . Similarly to Lem. 2, it is possible to lower bound the quantile by a dimension-free quantity at the price of having an exponential dependence on  $S$  in  $\delta$ .

## Appendix A. Proof for the asymptotic scenario

In this section we report the proofs of lemmas and theorem stated in Sec. 2.

### A.1 Proof of Lem. 4

Let  $Y_{n,i} = \frac{1}{\sqrt{n}\sqrt{\frac{S-1}{S^2}}} \sum_{j=1}^n (X_i^j - \frac{1}{S})$  and  $Y_n = (Y_{n,i})_{i \in \mathcal{S}}$ . Then:

$$\begin{aligned} \sqrt{n}Z_n &= \sqrt{n} \max_{v \in [0,D]^S} (\hat{p} - p)^\top v = \sqrt{n} \max_{v \in [0,D]^S} \sum_{i=1}^S \frac{v_i}{n} \sum_{j=1}^n (X_i^j - \frac{1}{S}) \\ &= \max_{v \in [0,D]^S} \sum_{i=1}^S Y_{n,i} v_i \sqrt{\frac{S-1}{S^2}} = D \sqrt{\frac{S-1}{S^2}} \cdot e^\top Y_n^+, \end{aligned}$$

where we used the fact that the  $v$  maximizing  $Z_n$  takes the largest value  $D$  for all positive components  $Y_{n,i}$  and is equal to 0 otherwise. We recall that the covariance of the normalized multinoulli variable  $Y_{n,i}$  with probabilities  $p_i = 1/S$  is  $I_S - \frac{1}{S-1}N$ . As a result, a direct application of the central limit theorem gives  $Y_n \xrightarrow{\mathcal{D}} Y \sim \mathcal{N}(0, I_S - \frac{1}{S-1}N)$ . Then we can apply the functional CLT and obtain  $Z_S = \lim_{n \rightarrow \infty} \sqrt{n}Z_n = \lim_{n \rightarrow \infty} e_S^\top Y_n^+ \sqrt{\frac{S-1}{S^2}} \xrightarrow{\mathcal{D}} \sqrt{\frac{S-1}{S^2}} \cdot e^\top Y^+$ , where  $Y^+$  is a random vector obtained by truncating from below at 0 the multi-variate Gaussian vector  $Y$ . Since the marginal distribution of each random variable  $Y_i$  is  $\mathcal{N}(0, 1)$ , i.e., are identically distributed (see definition in Lem. 4),  $Y_i^+$  has a distribution composed by a Dirac distribution in 0 and a half normal distribution, and its expected value is  $\mathbb{E}[Y_i^+] = 1/\sqrt{2\pi}$ , while leads to the final statement on the expectation.

### A.2 Proof of Lem. 5

Denote  $\lambda(A)$  the set of eigenvalues of square matrix  $A$ . Let  $B \in \mathbb{R}^{S \times S}$  such that  $B = \begin{bmatrix} 0_{S,S-1} & e_S \end{bmatrix}$ , where  $0_{S,S-1} \in \mathbb{R}^{S \times (S-1)}$  is a matrix full of zeros. Then, we can write the eigenvalues of the covariance matrix of  $Y$  as

$$\begin{aligned} \lambda(I_S - \frac{1}{S-1}N) &= \lambda\left(\frac{S}{S-1}I_S - \frac{1}{S-1}e_S e_S^\top\right) = \lambda\left(\frac{S}{S-1}I_S - \frac{1}{S-1}BB^\top\right) \\ &= \frac{S}{S-1} \lambda\left(I_S - \frac{1}{S-1}B^\top B\right) = \frac{S}{S-1} \lambda\left(I_S - \begin{bmatrix} 0_{S-1} & 0 \\ 0 & 1 \end{bmatrix}\right), \end{aligned}$$

where we use the fact that  $\lambda(I - A^\top A) = \lambda(I - AA^\top)$ . As a result, the covariance of  $Y$  has one eigenvalue at 0 and eigenvalues equal to  $\frac{S}{S-1}$  with multiplicity  $S-1$ . As a result, we can diagonalize it with an orthogonal matrix  $U \in O_S(\mathbb{R})$  (obtained using the normalized eigenvectors) and obtain

$$U(I_S - \frac{1}{S-1}N)U^\top = \begin{bmatrix} \frac{S}{S-1}I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Define  $W = \sqrt{\frac{S-1}{S}}UY$ , then:

$$\begin{aligned} \text{Cov}(W, W) &= \frac{S-1}{S} \text{Cov}(UY, UY) = \frac{S-1}{S} U \text{Cov}(Y, Y) U^T \\ &= \frac{S-1}{S} U \left( I_S - \frac{1}{S-1} N \right) U^T = \begin{bmatrix} I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Thus  $W \sim \mathcal{N}\left(0, \begin{bmatrix} I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}\right)$ .

### A.3 Proof of Thm. 6

By exploiting Lem. 4 and Lem. 5 we can write:

$$Z_S \sim e_S^T Y^+ \cdot \sqrt{\frac{S-1}{S^2}} = e_S^T \left( \sqrt{\frac{S}{S-1}} U^T W \right)^+ \cdot \sqrt{\frac{S-1}{S^2}} = e_S^T (U^T W)^+ \cdot \frac{1}{\sqrt{S}}$$

Let  $g(\cdot) = e_S^T (U^T \cdot)^+ \frac{1}{\sqrt{S}}$ . Then  $g$  is 1-Lipschitz:

$$|g(x) - g(y)| \leq \text{Lip}(e_S^T \cdot) \text{Lip}(U^T \cdot) \text{Lip}((\cdot)^+) \frac{1}{\sqrt{S}} \|x - y\|_2 = \sqrt{S} \cdot 1 \cdot 1 \cdot \frac{1}{\sqrt{S}} \|x - y\|_2$$

where  $\text{Lip}(f)$  denotes the Lipschitz constant of a function  $f$  and we exploit the fact that  $U$  is an orthonormal matrix.

We can now study the concentration of the variable  $Z_S$ . Given that  $W$  is a vector of i.i.d. standard Gaussian variables<sup>2</sup> and  $g$  is 1-Lipschitz, we can use (Wainwright, 2017, Thm. 2.4) to prove that for all  $t > 0$ :

$$\mathbb{P}(Z_S \geq \mathbb{E}[Z_S] - t) \geq 1 - \mathbb{P}(|Z_S - \mathbb{E}[Z_S]| \geq t) \geq 1 - 2e^{-\frac{t^2}{2}}.$$

Substituting the value of  $\mathbb{E}[Z_S]$  and inverting the bound gives the desired statement.

## References

- Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NIPS*, pages 1184–1194, 2017.
- Luc Devroye. The equivalence of weak, strong and complete convergence in  $\ell_1$  for kernel density estimates. *The Annals of Statistics*, 11(3):896–904, 09 1983. doi: 10.1214/aos/1176346255.
- Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.
- Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. 2017.
- Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Labs, 2003.

---

2. Note that we can drop the last component of  $W$  since it is deterministically zero.