Concentration Inequalities for Random Sets
— DRAFT —

December 10, 2016

Abstract

In a large finite population, each subject is randomly colored either red or green with equal probabilities, independently of the others. Then, a sub-population is selected. The goal of this paper is to bound the difference between the number of reds and the number of greens in the sub-population.

1 Introduction

In many experimental processes, a population is randomly divided to two parts, a certain measurement is done on one part and then applied to the other part. Obviously, the measurement done on one part may not be entirely accurate on the other part, due to the imbalance caused by the randomization. It is desired to have an upper bound on this imbalance. We model this process in the following way.

There is a population O with a large finite number of subjects. The population is colored randomly: for each subject, an unbiased coin-toss is used to decide whether the subject is colored red or green. Then, a sub-population $T \subseteq O$ is selected. Denote by T^R the set of red subjects in T and by T^G the set of green subjects. The difference $\big||T^R| - |T^G|\big|$ denotes the imbalance caused by the randomization process. What is high-probability upper bound on this imbalance?

There are two extreme cases:

- The easy case is when T does not depend on the coloring, i.e, T is a deterministic set defined before the coin-tosses. Then, both $|T^R|$ and $|T^G|$ are expected to be near $|T|/2$. The difference between them can easily be bounded using standard concentration inequalities; see below Lemma 2.1.

- The hard case is when T can depend on the coloring in an arbitrary way. Then, no upper bound is possible. For example, an adversary can select T to be the set of red subjects in O. In this case, $T^R = T$ and $T^L = \emptyset$ and the difference between them is as large as can be.

We are interested in an intermediate case, in which T may depend on the coloring but only in a restricted way. As an example, suppose all the subjects in O are placed on the real line, and T must be an interval. T may depend on the coloring, so it is a random variable and the standard concentration inequalities do not apply. However, the restriction to an interval means that an adversary cannot select T to be all and only the red subjects in O. Therefore we may hope to have a non-trivial upper bound on the imbalance $||T^R| - |T^G||$. Our goal in this paper is to define a family of random sets and prove high-probability upper bounds on their imbalance.

Our motivating application comes from economics. Often, to determine a price for an item, a market-research is conducted in which a random sample of the buyer population is used to calculate an 'optimal' price, p. Naturally, a price that is optimal in the sample might not be optimal in the global population. The optimality of the price depends on the set of buyers who want to buy the item in price p. Denote this set by T. Since p depends on the sampling, it is a random variable, so T is a random variable too. However, it is reasonable to assume that T is an interval, since it includes all buyers whose valuation for the item is more than p. The concentration bounds we develop in the present paper can be used to bound the imbalance in T.

2 Deterministic-set Halving Lemma

As a baseline, we repeat a known lemma for deterministic sets. We prove it in three variants that will be useful later.

Below, the shorthand "w.p. x" means "with probability of at least x".

Lemma 2.1 (Deterministic-set Halving Lemma)**.** *If T is a deterministic set, then for every constant $r \geq 1$:*

$$\text{If } |T| = t\text{:} \qquad \text{w.p. } 1 - \frac{2}{t^{2r^2}} : \quad \left||T^R| - |T^G|\right| < 2r\sqrt{t \ln t} \tag{2.1}$$

$$\text{If } |T| \geq t_{\min}\text{:} \quad \text{w.p. } 1 - \frac{2}{(t_{\min})^{2r^2}} : \quad \left||T^R| - |T^G|\right| < 2r\sqrt{|T| \ln |T|} \tag{2.2}$$

$$\text{If } |T| \leq t_{\max}\text{:} \quad \text{w.p. } 1 - \frac{2}{(t_{\max})^{2r^2}} : \quad \left||T^R| - |T^G|\right| < 2r\sqrt{t_{\max} \ln t_{\max}} \tag{2.3}$$

Proof. For every subject in T, define a random variable that equals 1 if the subject is red and -1 otherwise. These are i.i.d. random variables each of which is bounded in $[-1, 1]$. The sum of these variables is $|T^R| - |T^G|$ and the expectation of the sum is 0. For every $q \geq 0$, define the *failure probability* as:

$$P_{fail,q} := \Pr\left[\left||T^R| - |T^G|\right| > q\right]$$

By Hoeffding's inequality:

$$P_{fail,q} < 2\exp\left(\frac{-2q^2}{\sum_T (1-(-1))^2}\right) \leq 2\exp\left(\frac{-2q^2}{4 \cdot |T|}\right)$$

To get (2.1), let $q = 2r\sqrt{t \ln t}$; then $P_{fail,q} \leq 2/t^{2r^2}$.
To get (2.2), let $q = 2r\sqrt{|T| \ln |T|}$; then $P_{fail,q} \leq 2/|T|^{2r^2} \leq 2/(t_{\min})^{2r^2}$.
To get (2.3), let $q = 2r\sqrt{t_{\max} \ln t_{\max}}$; then $P_{fail,q} \leq 2/(t_{\max})^{2r^2}$. □

3 d-bounded random-sets

If the set T is not deterministic but depends on the outcomes of the random sampling, then Lemma 2.1 is not true without further restrictions. To handle such cases in a meaningful way we need to use some structure on the possible values of the set T.

Definition 3.1. A **random-set** is a random variable whose possible values are subsets of the global population O, and whose value depends on the random coloring process. The *support* of a random-set is the collection of sets that it can be with positive probability.

Definition 3.2. Given an integer $d \geq 1$, a set family W is called **d-bounded** if for every integer $j \geq 1$, the number of elements in W having cardinality j is at most $(j+1)^{d-1}$.

Definition 3.3. Given an integer $d \geq 1$, a random-set T is called **d-bounded** if its support is a d-bounded set-family.

Example 3.4. Let O be a set of real numbers. Let p be some real-valued random variable. Define $T = \{o \in O | o < p\}$. T is a random-set, since its value is a set that depends on a random variable. It is 1-bounded, because for every integer j, there is at most one possible outcome of T with cardinality j — it is the set of j smallest numbers in O. We will later generalize this example and show how to construct d-bounded random-sets. □

A d-bounded random-set is useful because of the following lemma.

Lemma 3.5 (Random-set Halving Lemma)**.** *Let T be a d-bounded random-set, for some integer $d \geq 1$. Then:*

$$\text{w.p. } 1 - 4/t_{\min} : \quad \text{If } |T| \geq t_{\min}\text{:} \qquad \left||T^R| - |T^G|\right| < 2d \cdot \sqrt{|T| \ln |T|} \tag{3.1}$$

$$\text{w.p. } 1 - 2/\sqrt{t_{\max} \ln t_{\max}} : \quad \text{If } |T| \leq t_{\max}\text{:} \quad \left||T^R| - |T^G|\right| < 2d \cdot \sqrt{t_{\max} \ln t_{\max}} \tag{3.2}$$

Proof. Denote the support of T by W (it is a collection of sets). Denote the subset of W containing sets of j elements by W^j. Every set $w_j \in W^j$ is deterministic, so it is eligible for the Deterministic-set Halving Lemma. Substituting $r = d$ in (2.1) gives, for every j, w_j:

$$\text{w.p. } 1 - \frac{2}{j^{2d^2}} : \qquad \left||w_j^R| - |w_j^G|\right| < 2d \cdot \sqrt{j \ln j} \tag{3.3}$$

Since W is d-bounded, the number of different sets in W^j is at most $(j+1)^{d-1}$. Hence, by the union bound, the above statement is true for *all* sets in W^j w.p. $1 - 2(j+1)^{d-1}/j^{2d^2} \geq 1 - 4/j^2$:

$$\text{w.p. } 1 - \frac{4}{j^2} : \quad \forall w_j \in W^j : \quad \left||w_j^R| - |w_j^G|\right| < 2d \cdot \sqrt{j \ln j} \tag{3.4}$$

Using the union bound again, the probability that inequality (3.4) is false for at least one $j \geq t_{\min}$ is upper-bounded by:

$$\sum_{j=t_{\min}}^{\infty} \frac{4}{j^2} \approx \int_{x=t_{\min}}^{\infty} \frac{4}{x^2} dx = \frac{4}{t_{\min}}$$

so w.p. $1 - 4/t_{\min}$, inequality (3.4) is true for all w_j with $|w_j| \geq t_{\min}$. This implies (3.1).

For (3.2), consider the following two cases:

- **Case 1**: $|T| < 2\sqrt{t_{\max} \ln t_{\max}}$. Then w.p. 1:

$$\left||T^R| - |T^G|\right| \leq |T| < 2\sqrt{t_{\max} \ln t_{\max}} \leq 2d\sqrt{t_{\max} \ln t_{\max}}$$

- **Case 2**: $|T| \geq 2\sqrt{t_{\max} \ln t_{\max}}$. Use inequality (3.1) with $t_{\min} = 2\sqrt{t_{\max} \ln t_{\max}}$:

$$\text{w.p. } 1 - 4/(2\sqrt{t_{\max} \ln t_{\max}}) : \quad \left||T^R| - |T^G|\right| < 2d\sqrt{|T| \ln |T|} \leq 2d\sqrt{t_{\max} \ln t_{\max}} \quad \text{since } |T| \leq t_{\max}.$$

□

Motivated by the Random-set Halving Lemma, we now present ways to construct d-bounded random-sets.

4 d-dimensional random-sets

The property of being d-bounded is not preserved under set operations such as union. Below, we define a stronger property which is preserved under set-union.

Definition 4.1. Let W be a set-family and w' an arbitrary set. Define the following set-families:

$$W \cap w' := \{w \cap w' | w \in W\} \qquad W \setminus w' := \{w \setminus w' | w \in W\}$$

Definition 4.2. Given an integer $d \geq 1$, a set-family W is called **d-dimensional** if for every set w', the family $W \cap w'$ is d-bounded (as defined in Definition 3.3).

Note: An equivalent condition is that for every set w'', the family $W \setminus w''$ is d-bounded (apply the original definition with $w' = \overline{w''}$ = the complement of w'').

Definition 4.3. Given an integer $d \geq 1$, a random-set T is called **d-dimensional** if its support is a d-dimensional set-family.

Obviously, every d-dimensional random-set is also d-bounded, so it is eligible for the Random-Set Halving Lemma (3.5).

Below we provide three rules for constructing random sets with a bounded dimension. The first one is the *Containment-Order Rule*.

Definition 4.4. A finite set-family is called **ordered-by-containment** if the sets in the family can be indexed $\{w_1, w_2, \dots\}$ such that for all $i < j$: $w_i \subset w_j$.

Remark 4.5. In measure theory and stochastic processes theory, a set-family that is ordered-by-containment is called a *filtration*.

Lemma 4.6. *If a set-family W is ordered-by-containment, then W is 1-dimensional.*

Proof. If W is ordered-by-containment, then for every set w', $W \cap w'$ is clearly also ordered-by-containment.

In every set-family that is ordered-by-containment, $\forall i, j : i < j$: $w_i \subset w_j$. All sets are finite, so there can be at most a single w_i with any given cardinality. Hence, for every w', the family $W \cap w'$ is 1-bounded. Hence, W is 1-dimensional. □

Corollary 4.7 (Containment-Order Rule)**.** *If the support of a random-set T is ordered-by-containment, then T is 1-dimensional.*

Example 4.8. Consider a family $\{w_1, w_2, \dots\}$ where for every j, w_j is the set of j smallest elements in a finite population O of real numbers. This family is clearly ordered-by-containment. By the Containment-Order Rule, the random-set of Example 3.4 is not only 1-bounded but also 1-dimensional. □

5 Intersections and unions of random-sets

Lemma 5.1. *If W is a d-dimensional set-family and w' is any set, then the set-family $W \cap w'$ is also d-dimensional.*

Proof. We have to prove that for any set w'', the set-family $(W \cap w') \cap w''$ is d-bounded. Indeed, $(W \cap w') \cap w'' = W \cap (w' \cap w'')$, and because W is d-dimensional, by definition $W \cap (w' \cap w'')$ is d-bounded. □

Corollary 5.2. *The intersection of a d-dimensional random-set with a deterministic set yields a d-dimensional random-set.*

Lemma 5.3 (Union Rule)**.** *If T_1 is a d_1-dimensional random-set and T_2 is a d_2-dimensional random-set, then their union:*

$$T := T_1 \cup T_2$$

is a $(d_1 + d_2)$-dimensional random-set.

Proof. Let W_i be the support of T_i (for $i = 1, 2$) and W the support of T. Let w' by any deterministic set. We have to prove that $W \cap w'$ is a $(d_1 + d_2)$-bounded set-family.

We know that for each i, W_i is d_i-dimensional. By Lemma 5.1, $W_i \cap w'$ is d_i-dimensional. Suppose we want to construct a set in the family $W \cap w'$, and we want it to have cardinality j. The choices we can make are as follows:

- First, we choose a set w_1 from the family $W_1 \cap w'$. The size of w_1 must be between 0 and j, so we have at most $(j+1)$ choices for the size of w_1 and then at most $(j+1)^{d_1-1}$ for the set w_1 itself (because $W_1 \cap w'$ is d_1-dimensional).

- Next, we choose a set w_2 from the family $(W_2 \cap w') \setminus w_1$. The size of w_2 must be exactly $j - |w_1|$. Since the family $(W_2 \cap w') \setminus w_1$ is d_2-dimensional, we have at most $(j+1)^{d_2-1}$ choices for w_2.

All in all, the number of choices is at most $(j+1) \cdot (j+1)^{d_1-1} \cdot (j+1)^{d_2-1} = (j+1)^{d_1+d_2-1}$.

Hence the set-family $W \cap w'$ is (d_1+d_2)-bounded. Since this is true for every set w', the set-family W is (d_1+d_2)-dimensional. □

Corollary 5.4. *For every d, the union of d one-dimensional random-sets is a d-dimensional random-set (hence it is also d-bounded).*

Proof. By induction on d, using Lemma 5.3 as the induction step. □

Example 5.5. Let O be a finite set of points in the plane, $O \subseteq \mathbb{R}^2$.

Let $T := \{(x,y) \in O | x > p_x \text{ or } y > p_y\}$, where p_x and p_y are random variables. Then, T is a 2-dimensional random-set, since it is a union of the two random-sets: $T_x := \{(x,y)|x > p_x\}$ and $T_y := \{(x,y)|y > p_y\}$, which are 1-dimensional by the Order-Containment Rule. □

The analogue of Lemma 5.3 for intersections of random-sets is not true.

Example 5.6. Let $T = \{(x,y) \in O | x > p_x \text{ and } y > p_y\}$, where p_x and p_y are random variables. Then, $T = T_x \cap T_y$, where T_x, T_y are 1-dimensional random-sets defined as in the previous example. However, T is not d-bounded for any finite d. This is illustrated in Figure 1. The points represent the elements of O. Each quarter-plane represents a possible value of T. The cardinality of each such value is 1. Therefore, the number of sets of cardinality 1 in the support of T can be as high as $|O|$ (the size of the global population). This is not bounded by $(1+1)^{d-1}$ for and constant d, since $|O|$ can be arbitrarily large. Similarly, for every $j \geq 1$, the number of sets of cardinality j in the support of T is not bounded by $(j+1)^{d-1}$ for any constant d.

Moreover, if $|O|$ is sufficiently large, with high probability there will be a large number of adjacent elements of O colored red. An adversary can select a quarter-plane that contains all and only these red elements. This quarter-plane will have an arbitrarily large imbalance. □

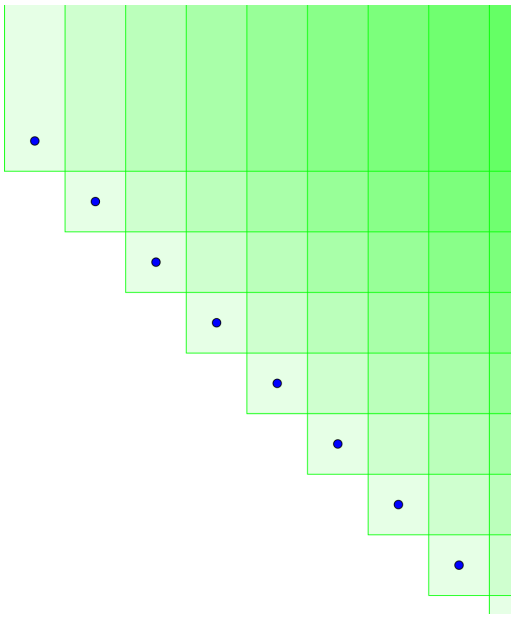

Figure 1: An intersection of two 1-dimensional random-sets may not be d-bounded.

Intersections of random-sets have a bounded dimension if one of the elements in the intersection has a bounded cardinality.

Lemma 5.7 (Intersection Rule)**.** *Given integers $d, n, d'' \geq 1$ and $k \geq 1$, if:*

- *$T_{k,d}$ is a d-dimensional random-set which is bounded by: $|T_{k,d}| < k$ w.p. 1.*
- *For every $i \in 1, \ldots, n$, T_i is a d''-dimensional random-set.*

Then their intersection T, defined as:

$$T = T_{k,d} \cap T_1 \cap \cdots \cap T_n$$

has a dimension of at most $((d + n \cdot d'') \lg k)$.

Proof. Let $W_{k,d}$ be the support of $T_{k,d}$; for each $i \in 1, \ldots, n$ let W_i be the support of T_i; let W be the support of T. Let w' by any deterministic set. We have to prove that $W \cap w'$ is $((d + n \cdot d'') \lg k)$-bounded set-family.

Every set $w \in W \cap w'$ can be constructed in the following way:

- Select a set $w_0 \in W_{k,d} \cap w'$, having j_0 items.
- Select a set $w_1 \in (W_1 \cap w_0) \cap w'$, having j_1 items;
- Select a set $w_2 \in ((W_2 \cap w_1) \cap w_0) \cap w'$, having j_2 items;
- ... Select a set $w_n \in (W_n \cap ... \cap w_0) \cap w'$, having j_n items.

By definition of $T_{k,d}$, $|T_{k,d}| < k$ so $j_0 \leq k - 1$. Since $T_{k,d}$ is d-dimensional, given j_0, the number of choices for w_0 is at most $(j_0 + 1)^{d-1} \leq k^{d-1}$. Since there are at most k choices for j_0, the total number of choices for w_0 is at most k^d.

For every $i \geq 1$, $j_i \leq j_{i-1}$ and the final set w is equal to w_n. Hence, the number of elements in w is j_n. So we have to select a weakly-decreasing sequence of non-negative integers, $j_1, \ldots, j_{n-1}$, such that $k > j_0 \geq j_1 \geq \cdots \geq j_{n-1} \geq j$. For each j_i there are at most $k - j < k$ choices, so the total number of sequences is at most k^{n-1}.

The set-families used in each of the following steps are intersections of a d''-dimensional set with deterministic sets. Hence they are all d''-dimensional. For every selection of j_i, there are at most $(j_i + 1)^{d''-1} \leq k^{d''-1}$ choices for w_i. The total number of choices for all the w_i, for $i = 1, \ldots, n$, is thus at most $k^{n \cdot (d''-1)}$.

Multiplying the three numbers of choices gives that the total number of ways to construct w is at most $k^{d+(n-1)+n(d''-1)} = k^{d+n \cdot d''-1} \leq (j + 1)^{((d+n \cdot d'') \lg k)-1}$. □

Remark 5.8. If the set $T_{k,d}$ is deterministic and $|T_{k,d}| = k$, then the set $W_{k,d}$ is a singleton and there is only one way to choose w_0. Since $1 = k^0$, the proof is still valid if we take $d = 0$, so the resulting random-set is $(n \cdot d'' \cdot \lg k)$-dimensional.

Effectively, in the Intersection Rule, a deterministic set is equivalent to a zero-dimensional random-set. The same is true in the Union Rule.