Deja vu in Barcelona

Devdatt P. Dubhashi[1] [2]

I don't mean only the feeling on visiting wonderful Barcelona again – Las Ramblas, Gaudi, Sagrada Familia and cafe con leche. But also a *deja vu* of the intellectual kind listening to a delightful invited talk by Emo Welzl at RANDOM 98.

I once wrote a piece together with Desh Ranjan entitled "Great(er) Expectations" (BRICS Newsletter 5, July 1996). That's what triggered the feeling in Emo's talk and inspired me to write this report.

Emo's talk dealt with the technique of *random sampling*, a hugely successful paradigm in algorithms in general, and in computational geometry in particular (which was the focus of his talk). Abstractly, we have a set S of objects and we wish to find the "configuration" they form. Random sampling proceeds by picking a smaller sample $R \subseteq S$ at random and determining the "configuration" formed by R first. I'll give several concrete example from his talk below; in all of them, S consists of points in Euclidean space.

Keeping to the abstract setting, suppose there is a function φ defined on subsets of S. Given $T \subseteq S$ and $x \in S \setminus T$, we say that x *violates* T if $\varphi(T \cup \{x\}) \neq \varphi(T)$. Dually, we call $x \in T$ an *extreme* point of T if $\varphi(T) \neq \varphi(T \setminus \{x\})$. This was the terminology used by Emo. In the terminology of Mulmuley [2, 3], we would talk of the *configuration* $\sigma = \sigma(T)$ determined by T and the set

$$D(\sigma(T)) := \{x \in T \mid \varphi(T) \neq \varphi(T \setminus x)\},$$

of extreme points of T is called the *defining* set of the configuration σ. Similarly the set

$$C(\sigma(T)) := \{x \in S \setminus T \mid \varphi(T) \neq \varphi(T \cup x)\},$$

of violating points is called the *conflict set* of σ. We shall use both terminologies interchangeably and write $D(S)$ or $D(\sigma)$ etc [3]. For the analysis of various random structures and algorithms, one is interested in the expected size of $C(R)$ for a random subset R of size r.

It's high time we gave some concrete examples!

Example 1 (Smallest Number) Let S be a set of real numbers and for $T \subseteq S$, let $\varphi(T) := \min\{x \mid x \in T\}$. Then an extreme point in T is a minimum element and a conflicting point in $S \setminus T$ is an element less than all elements in T.

In this case, the expected value $a_r := \mathtt{E}[C(R) \mid |R| = r]$ of the conflict set of a random sample R of size r is at most $\frac{n-r}{r+1}$. It is instructive to see how a standard approach to showing this goes. Suppose the elements of S are all distinct so that we may as well identify S with $[n]$. For $i \in [n]$, introduce the indicator $X_i := 1[i \in C(R)]$, and noticing that $|C(R)| = \sum_i X_i$ sets up things nicely for linearity of expectation:

$$\begin{aligned}
a_r &= \sum_i \mathtt{E}[X_i] \\
&= \sum_i \mathtt{Pr}[X_i = 1] \\
&= \sum_i |\{T \subseteq [n] \mid i < T, |T| = r\}| \binom{n}{r}^{-1} \\
&= \binom{n}{r}^{-1} \sum_{1 \leq i < n} \binom{n-i}{r}
\end{aligned}$$

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi 110016, INDIA. email: ⟨dubhashi@cse.iitd.ernet.in⟩

[2] Work done while visiting **BRICS**, Basic Research in Computer Science, Centre of the Danish National Research Foundation, Department of Computer Science, University of Aarhus, Denmark.

[3] As pointed out to me by Emo subsequently, there are subtle differences between the two sets of concepts in degenerate situations, but I'll ignore that for this report.

At this point, we are forced to consult Knuth [1, p. 174], where the "upper summation" identity tells us that the binomial coefficient sum is $\binom{n}{r+1}$ and we get the desired result. But already we needed a non–trivial binomial coefficient identity – only the ominous tip of the iceberg!

As you'd imagine, the same thing goes for the kth–smallest number as well. In this case, the result is $a_r \leq k\frac{n-r}{r+1}$.

Example 2 (Convex Hull and Smallest Enclosing Disk) Here S is s set of points in some higher dimensional Euclidean space, and $\varphi(T)$ is the convex hull or the smallest enclosing ball of the set of points in T. For the smallest enclosing ball, $b_r \leq 3$ for all $r \geq 0$ – draw a picture, but don't fall into the trap (as I did) of thinking that this has anything to do with the fact that any three non–collinear points lie on a circle!

So let's return to the general setting: suppose we pick R to be a random sample of size $r \leq n =: |S|$. That is, R is chosen uniformly from amongst all subsets of size r. The question is: what is the expected sizes a_r and b_r of the conflict and defining sets respectively? In symbols, we're interested in:

$$a_r := \mathtt{E}[|C(R)| \mid |R| = r], \quad b_r := \mathtt{E}[|D(R)| \mid |R| = r.$$

They certainly depend on the particular situation at hand, but whatever they are, amazingly, there is always a delightfully simple relation between the two:

$$(r+1)a_r = (n-r)b_{r+1}. \tag{1}$$

Deja vu! Look at how exactly this relation took centre–stage in our newsletter contribution cited above! As Emo remarked, this simple relation lies behind a plethora of results in Computational Geometry based on random sampling. Except that this is completely hidden in the literature where all manner of binomial coefficient identities are invoked for the analysis – this is what I was referring to by the "ominous tip of the iceberg" in Example 1 above. By contrast, the only identity actually needed is the humble

$$\frac{\binom{n}{r+1}}{\binom{n}{r}} = \frac{n-r}{r+1}. \tag{2}$$

Here in fact, is the delightfully simple proof of (1) using only (2). The key point is, as alert readers would have noticed already, there is actually redundancy in the two dual sets of notions we've introduced: a point $x \in T$ is extreme iff it violates $T \setminus x$, i.e.

$$x \in D(T) \quad \leftrightarrow \quad x \in C(T \setminus x).$$

Let us therefore construct a bipartite graph G, between the sets of size r and $r+1$, connecting a set T of size r with a set T' of size $r+1$ iff $T' = T \cup x$ where, equivalently, x is extreme for T' or violates T. *Deja vu*: This is the same bipartite graph we had in "Great(er) Expectations"! Now,

$$\begin{aligned} b_{r+1} &:= \mathtt{E}[D(T') \mid |T'| = r+1] \\ &= \sum_{|T'|=r+1} |D(T')| \binom{n}{r+1}^{-1} \\ &= \sum_{T'} \sum_{x \in D(T')} \binom{n}{r+1}^{-1} \\ &= \sum_{(T,T') \in G} \binom{n}{r+1}^{-1} \end{aligned}$$

Similarly,

$$a_r := \mathtt{E}[C(R) \mid |R| = r] = \sum_{(T,T') \in G} \binom{n}{r}^{-1}.$$

So that finally,

$$\begin{aligned} \frac{a_r}{b_{r+1}} &= \frac{\sum_{(T,T') \in G} \binom{n}{r}^{-1}}{\sum_{(T,T') \in G} \binom{n}{r+1}^{-1}} \\ &= \frac{r+1}{n-r}. \end{aligned}$$

There are several variations possible on the theme. First, we notice that the proof would be the same for any probability distribution that is *symmetric*: we only used the fact that any r element subset has the same probability, $\binom{n}{r}$ of being picked. Thus the result holds for the pseudo–random distributions discussed in [3]. This is of great significance from an algorithmic point of view since it opens the way for standard derandomization techniques, such as those described in Mulmuley's book.

Even for probability distributions that are not symmetric, one can get similar results. A natural and widely occurring distribution on subsets is the *product measure*:

$$\mu(T) := \prod_{x \in T} p_x \prod_{x \notin T} q_x,$$

where $p_x, q_x, x \in S$ are arbitrary positive reals. *Deja vu*! This is the measure we were concerned with in "Great(er) Expectations". It is also the standard measure for the random graph and it crops up in many other situations like LP–based methods of *randomised rounding*. Reexamining the proof above shows that for such a product measure μ,

$$\frac{a_r}{b_{r+1}} = \frac{\sum_T \mu(T)}{\sum_{T'} \mu(T')},$$

and so,

$$\min_x \frac{q_x}{p_x} \leq \frac{a_r}{b_{r+1}} \leq \max_x \frac{q_x}{p_x}.$$

Another fruitful line is to extend the relations "violate" and "extreme" that we had between a point and a subset to relations between subsets. Say that a disjoint subset U violates T if $U \cap C(T) \neq \emptyset$ i.e. U contains a point that violates T. Dually, say that $U \subseteq T$ is extreme in T if $U \cap D(T) \neq \emptyset$ i.e. U contains an extreme point of T. How are these two related? Unfortunately, its is *not true* that U violates T iff U is extreme for $T \bigcup U$.

Example 3 (Closest pair) Let $\varphi(T)$ denote the distance between a closest pair in T. Let us take a pair of points p, q very close to each other (in particular closer than any pair in T) but far away from all points in T. Now, $\{p, q\}$ is extreme in $T \bigcup \{p, q\}$ but neither p nor q by itself violates T and so neither does the set $\{p, q\}$.

Emo defined a property to satisfy *locality* if in fact this is true: U violates T iff U is extreme for $T \bigcup U$. This is true of all the other examples listed above except the example of closest pairs.

Let us rephrase locality equivalently as follows: U does *not* violate T iff U is *not* extreme for $T \bigcup U$. This prompts us to construct a bipartite graph between sets of size r and sets of size $r + k$ connecting a set T of size r with a set T' of size $r + k$ if $T' = T \bigcup U$ and, equivalently, U does not violate T or U is not extreme for T'. What does the same argument for (1) give us?

$$\binom{n}{r} a_r^k = \binom{n}{r+k} b_{r+k}^k, \qquad (3)$$

where

$$a_r^k := \mathtt{E}\left[\binom{n - r - |C(R)|}{k} \mid |R| = r\right], \quad r+k \leq n,$$

and

$$b_r^k := \mathtt{E}\left[\binom{r - |D(R)|}{k} \mid |R| = r\right] \quad k \leq r.$$

Finally, let us see what this implies in *regular* structures which Emo defined as follows: φ is regular for S of *dimension* δ if $|D(T)| = \delta$ for all $T \subseteq S$ with $|T| \geq \delta$. In the smallest number example, we have a local property which is regular of dimension 1 while in the smallest enclosing ball example, we have a local property of dimension at most 3 (nothing to do with three non–collinear points being co–cyclic!). Then since $b_r^k = \binom{r-\delta}{k}$ for all $r \geq \delta$, from (3), we get a *higher moment* formula:

$$\mathtt{E}\left[\binom{n - r - |C(R)|}{k} \mid |R| = r\right] = \frac{\binom{n}{r+k}\binom{r+k}{k}}{\binom{n}{r}}.$$

Notice that this depends only on the locality and regularity properties – these two alone completely determine the distribution of $|C(R)|$, at least upto the applicable moments! Hence, one can apply a tail estimate obtained from such an applicable kth moment for one concrete simple problem, say, the smallest number, and apply it without change to any other problem sharing the same locality and regularity properties! Notice also that the analysis is valid without any independence assumptions which will open the avenues for derandomization.

There are lots of other variations on the theme that I'm sure occur to you as you read this – for instance, a weaker version of locality that would be satisfied by the closet pair example. Emo and his co–author Bernd Gärtner have promised that they will soon write up a fuller account.

References

[1] R. Graham, D. Knuth and O. Patshnik, *Concrete Mathematics*, Addison Wesley, 1989.

[2] K. Mulmuley, "Randomized Geometric Algorithms and Pseudorandom Generators", *Algorithmica*, 16 pp. 450–463, 1996.

[3] K. Mulmuley, *Computational geometry: An Introduction through Randomized Algorithms*, Prentice Hall.

Extracted from the IARCS Newsletter Vol 4 (1), April 1999.
See the entire Newsletter at `http://www.imsc.ernet.in/ iarcs`
Indian Association for Research in Computing Science.