A coupling time of $O(n \log n)$ for the "j-at random" shuffle

Dan Abramovich[1]

Suggested running head; The j-at random shuffle.

Abstract

Cards are shuffled in the following way: a place j in the deck is fixed, and at each time we pick the card at place j and insert it into the deck in a random place. We let n =number of cards go to infinity, and j vary in the interval $1 \leq j \leq n/2$. We will give an upper bound for the randomization time of this shuffle in the sense of [1], that specializes to $n \log n$ when $j = O(1)$ and to $n(\log n + \log \log n + O(1))$ when $j = O(n)$. The reason for the extra term is that while the case $j = O(1)$ relates directly to the "coupon collector's problem", we will give a coupling argument that in general relates to the "double coupon collector's problem", (also known as the "double dixie cup problem") which has a $\log \log n$ term.

KEY WORDS: Random walks on groups, coupling time, uniform distribution

1 INTRODUCTION

Consider the following method of mixing a deck of n cards: Fix an integer $j, 1 \leq j \leq n$. At each time, the card in position j is removed and replaced at a random position. In this paper we show that it takes $O(n \log n)$ steps to mix up the n cards.

The problem arose in a work of Aldous and Diaconis [2]. They proved a precise result for the case $j = 1$. Their method of proof (strong uniform times) breaks down for other values of j. Aldous [1] treated $j = 1$ by coupling. The problem of other values of j has been posed by Diaconis in talks.

The shuffle is treated as a Markov chain on the symmetric group $\mathbf{S}_n$. A careful description of the basic shuffle is given in section 2. Let P^k be the law of the chain, started from the identity permutation, after k steps. Let $\mathcal{U}$ be the uniform distribution on $\mathbf{S}_n$. The distance between P^k and $\mathcal{U}$ is measured by total variation:

$$\|P^k - \mathcal{U}\| = \sup_{A \in \mathbf{S}_n} |P^k(A) - \mathcal{U}(A)| \tag{1}$$

[1]Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138.

The main result shows that

$$k = n \log n + n \log \left(1 + 2j \left(1 - \frac{j}{n}\right) \frac{\log n}{n}\right) + nx \tag{2}$$

steps are sufficient.

Theorem 1 *There is a function $\phi(x)$ such that $\phi(x) \to 0$ as $x \to \infty$, such that for any $j, 1 \leq j \leq n$ and k defined by (2), $\|P^k - \mathcal{U}\| \leq \phi(x)$.*

Remark The function ϕ, along with the proof of theorem 1 is given in section 5. The rate of convergence is sharp in the sense of the following theorem:

Theorem 2 *For every $\epsilon > 0$ there is x such that for any choice of $j = j(n)$, and $k = n \log n - nx$ we have*

$$\lim_{n \to \infty} \|P^k - \mathcal{U}\| > 1 - \epsilon$$

The proof of theorem 2 is elementary, while the proof of theorem 1 is via the coupling method.

It is natural to conjecture that $n \log n + nx$ steps are sufficient, as is shown for $j = 1$ in [2]. The extra term plays a role only in the range $j > n/\log n$, in particular, when $j = n/2$.

I would like to thank Persi Diaconis for suggesting the problem and discussing and helping during the work. I would also like to thank J.F.Burnol and R.Stong for helpful discussions and remarks.

2 THE SHUFFLE

To fix ideas, cards are elements of $1, \ldots, n$ and if i is a card and $\pi \in \mathbf{S}_n$ then $\pi(i)$ is the place of the card in the deck π. We will now carefully describe the basic shuffle, and its inverse.

Definition 3 *Let n and j be fixed. Let $c_k \in \mathbf{S}_n$ be the following corresponding cycles:*

$$c_k = \begin{cases} (j, j-1, \ldots, k) & 1 \leq k \leq j \\ (j, j+1, \ldots, k) & j \leq k \leq n \end{cases}$$

Define probability measures μ, μ' on $\mathbf{S}_n$:

$$\mu(\pi) = \begin{cases} 1/n & \pi = c_k,\ k = 1, \ldots, n \\ 0 & \textit{otherwise} \end{cases}$$

$$\mu'(\pi) = \mu(\pi^{-1}).$$

We define markov chains X, X' on $\mathbf{S}_n$ generated by these probabilities:

$$P_X(\pi, \eta) = \mu(\eta\pi^{-1})$$

$$P_{X'}(\pi, \eta) = \mu'(\eta\pi^{-1}).$$

It is easy to see that $P_{X'}$ is the j-at random shuffle, and P_X is the inverse shuffle.

Proof of theorem 2: The theorem can be proved directly from the definition of the total variation distance in formula (1), once we produce an event $A \in \mathbf{S}_n$ with $\|P^k_{X'}(A) - \mathcal{U}(A)\|$ large. Following [1] we fix m and look at the event

$$A_m = \{\pi \in \mathbf{S}_n | \pi(s) \geq n/2 \text{ for all } n - m < s \leq n\},$$

that is, the event that the last m cards end up in the lower half of the deck. It is clear that $\mathcal{U}(A_m) \cong 1/2^m$ for large n. Let us show that $P^k_{X'}(A_m)$ is close to 1 for suitable x. In fact, in order that any card from the last m cards end up in the upper half, the card which starts in position $n - m + 1$ has to get to the half at some time before k. This is a version of the coupon collector's problem: the probability that a card at place $n - k + 1 > n/2$ moves one place upward at a given time is k/n, and we sum up the waiting times for these events for $m \leq k \leq n/2$. These waiting times are independant, of mean n/k and variance of order n^2/k^2, and thus, by Chebyshef's inequality $P^k_{X'}(\mathbf{S}_n - A) < a/(x - \log m)$ for some constant a.

■

3 THE COUPLING

For backgroung on coupling see e.g. [1]. First, a technical definition.

Definition 4 *Let $\pi_1, \pi_2 \in \mathbf{S}_n$. Let k be a card. We say that k is* matched *if*

$$\pi_1^{-1}(k) < j \text{ and } \pi_2^{-1}(k) < j \text{ or } \pi_1^{-1}(k) = j \text{ and } \pi_2^{-1}(k) = j$$

$$\text{or } \pi_1^{-1}(k) > j \text{ and } \pi_2^{-1}(k) > j$$

We say that k is coupled *if all cards are matched and $\pi_1^{-1}(k) = \pi_2^{-1}(k)$. We denote by l the number of unmatched cards for π_1 and π_2.*

That is, the card is matched if it is on the same side of j in the two decks π_1, π_2.

Our coupling works on the process X' and is designed first to match all cards and then to couple them. An intuitive, but rather precise description of the coupling is as follows:

1. Pick a card at random from deck 1.

2. If the card is matched, or if the cards at place j are different, take the same card from deck 2 and insert them at place j of the decks.

3. If the card is not matched and the cards at place j are the same, pick an unmatched card from the same side of j at uniform distribution (There are exactly $l/2$ such cards in deck 2) and insert the cards at place j.

More formally:

Definition 5 *Define the following Markov chain Y on $\mathbf{S}_n \times \mathbf{S}_n$:*

$$P_Y((\pi_1,\pi_2),(c_{k_1}\pi_1,c_{k_2}\pi_2)) = \begin{cases} 1/n & \pi_1^{-1}(k_1) = \pi_2^{-1}(k_2) \text{ matched}; \\ & \pi_1^{-1}(j) = \pi_2^{-1}(j) \\ 2/nl & \pi_1^{-1}(j) = \pi_2^{-1}(j); \\ & \text{neither } \pi_1^{-1}(k_1) \text{ nor } \pi_2^{-1}(k_2) \\ & \text{is matched, and} \\ & \{k_1 < j \text{ and } k_2 < j \\ & \text{ or } k_1 > j \text{ and } k_2 > j\} \\ 1/n & \pi_1^{-1}(k_1) = \pi_2^{-1}(k_2);\ \pi_1^{-1}(j) \neq \pi_2^{-1}(j) \end{cases}$$

and 0 otherwise. One easily checks that the "rows" sum to 1, and thus give transition probabilities.

This markov chain has the followig rather obvious properties, after the first step was done:

1. The number of unmatched cards is even; matching occures in pairs; no unmatching occurs.

2. If $\pi_1^{-1}(j) \neq \pi_2^{-1}(j)$ the corresponding cards are on the same side of j:

$$\pi_2\pi_1^{-1}(j) < j \text{ and } \pi_1\pi_2^{-1}(j) < j, \text{ or}$$

$$\pi_2\pi_1^{-1}(j) > j \text{ and } \pi_1\pi_2^{-1}(j) > j$$

3. If at some t, $\pi_1^{-1}(j) \neq \pi_2^{-1}(j)$, then at time $t+1$ we have $\pi_1^{-1}(j) = \pi_2^{-1}(j)$

4. The relative orders in the two decks of the cards that have been coupled at j at some time before, are the same, that is, if the previously coupled cards are in deck 1 in places $i_1 < i_2 < \ldots < j < \ldots < i_h$ and in deck 2 in places $i'_1 < i'_2 < \ldots < j < \ldots < i'_h$ then for every k we have $\pi_1^{-1}(i_k) = \pi_2^{-1}(i'_k)$

The last property deserves a definition:

Definition 6 *A card k is said to be* almost coupled *in the markov chain at time T, if at some $t < T$ we have $Y_t(k,k) = (j,j)$*

We still need to verify

Lemma 7 *The process $(P_Y, \mathcal{U}, \delta)$ is a coupling for the process (P_X, δ).*

Proof: Since both P_X and P_Y are Markovian, it is enogh to show that

$$\forall \tau \; P_X(\pi, \pi') = \sum_{\sigma} P_Y((\pi,\tau),(\pi',\sigma)) = \sum_{\sigma} P_Y((\tau,\pi),(\sigma,\pi'))$$

The process is completely symmeteric, therefore it is enough to consider only the first equality. This equality is obvious enough, once one is convinced by the intuitive description of the coupling in the beginning of the section.

■

The basic result on coupling that we use here is the *coupling lemma* (see [1]):

Lemma 8 *Let $T = \min\{t|$ the two components of Y_t are equal $\}$ be the coupling time. Then $\|P^k - \mathcal{U}\| \leq Prob\{T > k\}$.*

4 ESTIMATION OF MATCHING TIME

Let T be the coupling time for Y. Write $T = T_1 + T_2$, where T_1 is the time untill we have complete matching, and $T_2 = T - T_1$.

Now $T_1 \leq \sum_{s=1}^{\lceil n/2 \rceil} T_1^{(2s)}$, where $T_1^{(l)}$ is the time for getting an extra matching given we start with l unmatched cards and the cards at j are the same. (In fact, the sum goes up to the number of unmatches cards at time 0, call it $2s_0$ where s_0 is a random variable with mean around $(j-1)(n-j)/n$ and variance of the same order of magnitude.)

Define $\overline{h}$ to be the number of matched cards in places r, with $r < j$; $\underline{h}$ to be the number of matched cards in places r, with $r > j$.

Whenever the cards at j are the same, we may pick an unmatched card in probability l/n, say it is in place k. Given this, $k < j$ in probability $1/2$. Two new matchings will occure in the following step, unless the next card is matched in place $< j$. Similarly for $k > j$.

Thus, the probability for new matchings in the following two steps is

$$\frac{l}{2n} \cdot \frac{n - \underline{h} - 1}{n} + \frac{l}{2n} \cdot \frac{n - \overline{h} - 1}{n} = \frac{l(n-2+l)}{2n^2} \geq \frac{l}{2n}$$

This already shows that $E(T_1) = n \log n + O(n)$. Let us study this more carefully. Write $\alpha = (n+l-2)/(2n)$ and $p = l/n$. Let Z be a geometric random variable with parameter p, that is $\text{Prob}(Z = k) = p(1-p)^{k-1}$. The generating

function of Z is $z(t) = \frac{pt}{1-(1-p)t}$. Write $X = T_1^{(l)}$ We have the following renewal equation:

$$\text{Prob}\{X = n\} = \sum_{m=1}^{n-2}(1-\alpha)\text{Prob}\{Z = m\}\text{Prob}\{X = n-m-1\}$$

$$+\alpha\text{Prob}\{Z = n-1\}.$$

Writing $x(t)$ for the generating function of X, this translates into

$$x(t) = t(1-\alpha)z(t)x(t) + \alpha t z(t)$$

or

$$x(t) = \frac{\alpha t z(t)}{1-(1-\alpha)tz(t)} = \frac{\alpha t^2 p}{1-(1-p)t-(1-\alpha)t^2 p}$$

¿From this we can immediately compute the mean and variance of X:

$$x'(t) = \frac{\alpha z(t) + \alpha t z'(t)}{1-(1-\alpha)tz(t)} + \frac{\alpha t z(t)((1-\alpha)z(t) + (1-\alpha)tz'(t))}{(1-(1-\alpha)tz(t))^2}$$

Thus

$$E(X) = x'(1) = \frac{1}{\alpha}(\frac{p+1}{p})$$

Similarly we get after computation

$$Var(X) = p^{-2}O(1)$$

Summing up we get $Var(T_1) = O(n^2)$ and $E(T_1) = n\log s_0 + O(n)$ with $s_0 = (j-1)(n-j)/n$, if j varies such that the limit exists.

Following [3] we may do the following computation of the limit of the normalized characteristic function of T_1:

$$E(e^{i\theta(T_1 - n\log s_0(n))/n}) = s_0^{-i\theta}\prod_{s=1}^{s_0}\frac{e^{2i\theta/n}s(n+2s-2)/n^2}{1-e^{i\theta/n}(n-2s)/n - e^{2n\theta/n}s(n-2s+2)/n^2}$$

$$= s_0^{-i\theta}\prod_{s=0}^{s_0}\frac{1}{(1-\frac{i\theta}{s})}\frac{e^{i\theta/n}}{(1-\frac{i\theta}{n})}\frac{1}{(1+O(1/(sn)))}$$

The last two terms go to 1 by easy bounds. The main term is interesting only for $s_0 \to \infty$, and then the limit is given by the product formula for the gamma function:

$$\lim s_0^{i\theta}\prod_{s=1}^{s_0}(1-i\theta/s)^{-1} = \Gamma(1-i\theta)$$

Taking inverse Fourier transform of the gamma function (using the integral formula of gamma and a change of variable) we get

$$\lim \text{Prob}\{(T_1 - n \log s_0)/n < t) = e^{-e^{-t}}.$$

The reason we may assume that the number of initially unmatched cards is a constant $2s_0$ is that by Chebyshev's inequality, it contributes only an $n^{1/2}$ term which goes to zero when divided by n.

Remark: From this section we immediately get a bound on coupling time of $2n \log n + O(n)$. In the following sections we improve this considerably.

5 ESTIMATION OF COUPLING TIME

In order to improve our estimate we use the property of the preserved relative order of almost coupled cards. The main point is to compare the coupling time to the time of getting two complete sets of n coupons, when at each given time a new coupon is drawn uniformly and independently, from the possible n coupons. The limit distribution of this time S was given in [3]. They show that $\lim \text{Prob}\{S_n/n - (\log n + \log \log n) < x\} = e^{-e^{-x}}$.

Diaconis gives the following heuristic argument for an easy estimation of this problem: Suppose that we need to get n coupons, s_0 of them are needed twice. Let us estimate how many are left after $n \log n$ trials: the number of times we got a given coupon is approximately a Poisson random variable with parameter $\lambda = \log n$. The probability of having exactly one copy is $\log n/n$ and for no copy is $1/n$. The expected number of coupons we still need is $r = s_0 \log n/n + 1$, and the expected time for getting them is approximately $n \log(r)$. If we repeat the argument when we wait $n(\log n + \log r + x)$ for big x, we are quite sure to have a complete set, just by Markow's inequality. A direct repetition of the argument in [3] shows that for $s_0(n) \to \infty$ we have

$$\lim \text{Prob}\{S_n/n - (\log n + \log r) < x\} = e^{-e^{-x}}.$$

Proof of theorem 1: Let $N = n(\log n + \log(1 + s_0 \log n/n) + x)$, where s_0 is the number of cards not matched at the beginning, which has expected value around $2j(1 - j/n)$. We will show that $\text{Prob}\{(T - N)/n > 0\} < 100(1 + x)e^{-x}$.

We want to estimate the probability that there is card which is not almost coupled, and thus it is enogh to bound the expected number of these cards. For each particular unmatched card we bound the probability that it is not almost coupled by the probability of this event in a slightly delayed process, in which the term $(n + l - 2)/(2n)$ determining the probability that it is matched, given that it was chosen in one of the decks in the previous step is replaced by $1/2$. Whith a further delay in case the cards at j are different, we can easily cook up a process in which the waiting time t for matching this card has the following

renewal equation:

$$\text{Prob}\{t=k\} = \sum_{l=1}^{k-2} \frac{1}{2}\text{Prob}\{u=l\}\text{Prob}\{t=k-l-1\}$$

$$+\frac{1}{2}\text{Prob}\{u=k-1\}$$

where u is an exponential random variable with parameter $2/n$. Namely, in this delayed process we choose our card in probability $2/n$, and in the next step we match it in probability $1/2$.

Now using this renewal equation it can be easily shown that the waiting time t for matching the card has the same distribution as the following random variable:

Let X_i be independent Boolean random variable with probability $2/n$. Let Y_i be independent, and independent of the X_i, Boolean random variables with probability $1/2$. Let $s = \min\{i : X_i Y_i = 1\}$ and let $v = X_1 + \cdots + X_s$. Then $s + v$ has the distribution of our matching time.

It may be easily verified that

$$s \sim Exponential(1/n)$$

and

$$v|(s=k) \sim Binomial(k, 1/(n-1)) + 1.$$

We have

$$\begin{aligned}
\mathcal{P} = &\ \text{Prob}\{\text{ some card is not coupled at time } N\} \\
\leq &\ s_0 \text{Prob}\left\{ \begin{array}{l} \text{a given card is not almost coupled} \\ \text{in the modified process at time } N \end{array} \right\} \\
&\ +(n-s_0)(1-1/n)^N \\
\leq &\ s_0 \sum_{k=1}^{N} \left(\text{Prob}\{s=k\} \sum_{m=1}^{k} \text{Prob}\{v=m|s\}(1-1/n)^{N-k-m} \right) + e^{-x} \\
= &\ \frac{s_0}{n} \sum_{k=1}^{N} \left(1-\frac{1}{n}\right)^{N-2} \sum_{m=1}^{k} \binom{k-1}{m-1} \left(\frac{n}{(n-1)^2}\right)^{m-1} \left(\frac{n-2}{n-1}\right)^{k-m} + e^{-x} \\
= &\ \frac{s_0}{n} \sum_{k=1}^{N} \left(1-\frac{1}{n}\right)^{N-2} \left(1+\frac{1}{(n-1)^2}\right)^{k-1} + e^{-x}.
\end{aligned}$$

The second term in the summand is less than 3 independently of $k < 2n\log n$. We get a bound

$$\mathcal{P} \leq 12\frac{s_0}{n} N(1-\frac{1}{n})^N \leq 12\frac{s_0}{n}\frac{N}{n\frac{s_0}{n}\log n}e^{-x} + e^{-x} \leq 100(1+x)e^{-x}.$$

Again, a simple use of Chebyshev's inequality allows using a constant $2j(1-j/n)$ for s_0 rather than a random variable describing the initially non-matched cards, with a little loss in the constants

■

References

[1] Aldous,D. (1983). Random walks on groups and rapidly mixing Markov chains. *Seminar on probability XVII, Lecture notes in mathematics* 986, pp.243-297,Springer Berlin.

[2] Aldous,D., Diaconis,P. (1986). Shuffling cards and stopping times. *American Mathematical Monthly* 93 no.5, pp.333-348

[3] Erdös,P., Rényi,A. (1961). On a classical problem of probability thoery *Magyar tud. acad. mat. kutató int. közl.* 6 no. 1-2, pp.215-220.