Asymptotic Distribution Theory for Shell-Sort

Jon A. Wellner

University of Washington

based on joint work with

Robert T. Smythe
Oregon State University

Talk at Oberwolfach, March 15, 2002

Email: jaw@stat.washington.edu

http: //www.stat.washington.edu/ jaw/jaw.research.html

Outline

1. **Introduction: Shell-sort**

2. **History**
   - mean behavior
   - choice of increments
   - limit distribution: (2,1) Shell sort

3. **New limit theory**
   - (3,1) Shell sort
   - (h,1) Shell sort
   - (3,2,1) Shell sort

4. **Further Problems**

1. Introduction to Shell sort

$R = 3\ \ 2\ \ 6\ \ 5\ \ 9\ \ 8\ \ 1\ \ 4\ \ 7$ $\qquad n = 9$

Linear insertion sort:

$$
\begin{array}{ccccccccc}
3 & 2 & 2 & 2 & 2 & 2 & 1 & 1 & 1 \\
 & 3 & 3 & 3 & 3 & 3 & 2 & 2 & 2 \\
 & & 6 & 5 & 5 & 5 & 3 & 3 & 3 \\
 & & & 6 & 6 & 6 & 5 & 4 & 4 \\
 & & & & 9 & 8 & 6 & 5 & 5 \\
 & & & & & 9 & 8 & 6 & 6 \\
 & & & & & & 9 & 8 & 7 \\
 & & & & & & & 9 & 8 \\
 & & & & & & & & 9
\end{array}
$$

# of inversions in $R = 15 \equiv I(R)$

# of inversions needed to sort R:

$$n + I(R) = 9 + 15 = 24$$

(2,1) - Shell sort:

$$R = 3\ \ 2\ \ 6\ \ 5\ \ 9\ \ 8\ \ 1\ \ 4\ \ 7$$

2-Stage:

3 6 9 1 7

2 5 8 4

1 3 6 7 9

2 4 5 8

1 2 3 4 6 5 7 8 9

a two-sorted list

1-Stage: Linear insertion sort:

1 2 3 4 5 6 7 8 9

(3,1) - Shell sort:

$$R = 3\ \ 2\ \ 6\ \ 5\ \ 9\ \ 8\ \ 1\ \ 4\ \ 7$$

3 5 1
2 9 4
6 8 7

1 3 5
2 4 9
6 7 8

1 2 6 3 4 7 5 9 8

a three-sorted list

1-Stage: Linear insertion sort:

1 2 3 4 5 6 7 8 9

(3,2,1) - Shell sort:

$$R = 3 \;\; 2 \;\; 6 \;\; 5 \;\; 9 \;\; 8 \;\; 1 \;\; 4 \;\; 7$$

3-Stage:

As in (3,1) - Shell sort, first make the 3 - sorted list:

1 2 6 3 4 7 5 9 8

2-Stage: Now 2 - sort this list:

1 4 5 6 8

2 3 7 9

1 2 4 3 5 7 6 9 8

1-Stage: Linear insertion sort:

1 2 3 4 5 6 7 8 9

2. History

Shell (1959), Comm. of ACM, 3pp.

Basic ingredient: Linear Insertion Sort

$$\begin{aligned} C_n &= \text{total number of comparisons needed} \\ &= \sum_{i=1}^{n} (1 + V_i) \end{aligned}$$

where

$$\begin{aligned} V_i &= \#\text{of inversions caused by } X_i \text{ in } X_1, \ldots, X_n \\ &= i - \text{SeqRank}(R_i), \quad R_i = \text{rank of } X_i \\ &\sim \text{Unif}\{0, \ldots, i-1\} \end{aligned}$$

and $V_1, \ldots, V_n$ are independent.

$$E(C_n) = \frac{n(n+3)}{4} \sim \frac{n^2}{4},$$

$$Var(C_n) = \frac{n(n-1)(2n+5)}{72} \sim \frac{n^3}{36},$$

and, by the Lindeberg-Feller CLT,

$$\frac{C_n - n^2/4}{n^{3/2}} \to_d N\left(0, \frac{1}{36}\right) .$$

(2,1) - Shell Sort:

$$S_n = \text{total \# of comparisons}$$

Knuth (1973):

$$E(S_n) = \frac{1}{8}n^2 + \sqrt{\pi}128n^{3/2} + o(n^{3/2})$$

$$Var(S_n) = \left(\frac{13}{360} - \frac{\pi}{128}\right) n^3 + o(n^3)$$

Louchard (1986):

$$\begin{aligned}\frac{S_n - n^2/8}{n^{3/2}} &\to_d N\left(0, \frac{1}{144}\right) + \frac{1}{2}\int_0^1 |B^0(t)|dt \\ &\equiv N\left(0, \frac{1}{144}\right) + \frac{1}{2}A\end{aligned}$$

where $B^0 \equiv$ standard Brownian bridge, independent of the normal random variable

Shepp (1982); Rice (1982); Johnson and Killeen (1983):

$$P(A \le x) = \sqrt{\frac{\pi}{2}} \sum_{j=1}^{\infty} \delta_j^{-3/2} \psi(x/\delta_j^{3/2})$$

where $\delta_j = -a'_j/2^{1/3}$, $a'_j = j$-th zero of Ai',

$$\psi(t) = \left(\frac{3^2}{t}\right)^{1/3} \exp\left(-\frac{2}{27}t^2\right) Ai((3t)^{-4/3}) .$$

(h,1) - Shell Sort:

$S_n =$ total # of comparisons

Knuth (1973):

$$E(S_n) = \frac{n^2}{4h} + \frac{\sqrt{\pi}}{4}\binom{h}{2}\left(\frac{n}{h}\right)^{3/2} + o(n^{3/2})$$

Minimize by choosing $h = h_n = O(n^{1/3})$ to get

$$E(S_n) = O(n^{5/3})$$

$(t_{k_n}, t_{k_n-1}, \ldots, 1)$ - Shell Sort $\equiv \underline{t}-$Shell sort:

Take t_{k_j} from $2^a \cdot 3^b$, $a,\ b \in \{0, 1, 2, \ldots\}$

Theorem (Pratt, 1971). For $\underline{t}-$ Shell sort with increments from $2^a \cdot 3^b$,

$$E(S_n) \asymp O(n(\log n)^2)$$

$(h, k, 1)$ - Shell Sort:

Yao (1980); Janson and Knuth (1997)

3. New Limit Theory

- (2,1) - Shell Sort (n even)

Data : $X_1, Y_1, X_2, Y_2, \ldots, X_{n/2}, Y_{n/2}$

2 - Stage:

order the X_i's - requiring $C_{n/2}$ comparisons

order the Y_j's - requiring $\tilde{C}_{n/2}$ comparisons

$$S_n = C_{n/2} + \tilde{C}_{n/2} + n + I_n$$

where

$$\begin{aligned} I_n &= \text{the remaining number of inversions} \\ &= \sum_{j=1}^{n/2} V_j + \sum_{j=1}^{n/2} W_j \end{aligned}$$

where

$$\begin{aligned} V_j &= 1_{[X_{(1)} > Y_{(j)}]} + \cdots + 1_{[X_{(j)} > Y_{(j)}]} \\ W_j &= 1_{[Y_{(1)} > X_{(j)}]} + \cdots + 1_{[Y_{(j-1)} > Y_{(j)}]} \end{aligned}$$

By elementary algebra

$$\begin{aligned} I_n &= \sum_{j=1}^{n/2} \left| \sum_{i=1}^{n/2} 1_{[Y_i < X_j]} - 1_{[X_i < X_j]} \right| \\ &= \frac{n}{2} \sum_{j=1}^{n/2} \left| \hat{F}_n(X_j) - \tilde{F}_n(X_j) \right| + o_p(n) \\ &= \left(\frac{n}{2}\right)^2 \int_0^1 \left| \hat{F}_n(t) - \tilde{F}_n(t) \right| d\tilde{F}_{n/2}(t) + o_p(n) \end{aligned}$$

where

$$\begin{aligned} \hat{F}_{n/2}(t) &= \frac{1}{n/2} \sum_{i=1}^{n/2} 1_{[Y_i \le t]} \\ \tilde{F}_{n/2}(t) &= \frac{1}{n/2} \sum_{i=1}^{n/2} 1_{[X_i \le t]} \end{aligned}$$

and

$$\sqrt{\frac{n}{2}}(\tilde{F}_{n/2}(t) - t) \Rightarrow B_1(t)$$
$$\sqrt{\frac{n}{2}}(\hat{F}_{n/2}(t) - t) \Rightarrow B_2(t)$$

with B_1, B_2 independent standard Brownian bridges.

Hence

$$\begin{aligned}
\frac{I_n}{n^{3/2}} &= \sqrt{\frac{n}{2}} \int_0^1 |\widehat{F}_{n/2}(t) - t - (\tilde{F}_n(t) - t)|\, d\tilde{F}_{n/2}(t) \\
&\qquad + o_p(1) \\
&\to_d \int_0^1 |B_1(t) - B_2(t)|\, dt \\
&=_d \sqrt{2} \int_0^1 |B^0(t)|\, dt
\end{aligned}$$

Equivalently,

$$\frac{I_n}{n^{3/2}} \to_d \frac{1}{2} \int_0^1 |B^0(t)|\, dt \equiv \frac{1}{2} A$$

Upshot: Louchard's theorem for (2,1) - Shell sort:

$$\begin{aligned}
\frac{S_n - n^2/8}{n^{3/2}} &= \frac{C_{n/2} - n^2/16}{n^{3/2}} + \frac{\tilde{C}_{n/2} - n^2/16}{n^{3/2}} \\
&\qquad + \frac{I_n}{n^{3/2}} + o(1) \\
&\to_d N(0, \frac{1}{8 \cdot 36}) + \tilde{N}(0, \frac{1}{8 \cdot 36}) + \frac{1}{2} A \\
&= N(0, \frac{1}{144}) + \frac{1}{2} A
\end{aligned}$$

- (3,1) - Shell sort?
- (h,1) - Shell sort?

Theorem. (Smythe and Wellner, 2001). For $(h,1)-$Shell sort:

$$\frac{S_n - n^2/4h}{n^{3/2}} \to_d N(0, \frac{1}{36h^2}) + \frac{W_h}{h^{3/2}}$$

where

$$W_h = \sum_{1 \leq r < s \leq h} \int_0^1 |B_r(t) - B_s(t)| dt \,,$$

$B_1, \ldots, B_h$ independent Brownian bridge processes

Corollary. For (h,1) - Shell sort:
(a) Mean # of comparisons (Knuth):

$$E(S_n) = \frac{n^2}{4h} + \frac{\sqrt{\pi}}{4}\binom{h}{2}\left(\frac{n}{h}\right)^{3/2} + o(n^{3/2})$$

(b) Var of # of comparisons:

$$\begin{aligned} Var(S_n) \;\; &= \;\; \left\{\frac{1}{36h^2} + \frac{1}{h^3}\binom{h}{2}\left(\frac{7}{30} - \frac{\pi}{16}\right)\right. \\ &\qquad + \left. \; h(h-1)(h-2)(C - \frac{\pi}{16})\right\} n^3 \\ &\qquad + o(n^3) \end{aligned}$$

where

$$\begin{aligned} C &= E\left(\int_0^1 |B_1(t) - B_2(t)|dt \int_0^1 |B_1(t) - B_3(t)|dt\right) \\ &= .2051... \end{aligned}$$

(see Statistica Neerlandica (2002))

Distribution of W_h, $h \geq 3$, is unknown

$$\begin{aligned} W_3 &= \int_0^1 |B_1(t) - B_2(t)|dt + \int_0^1 |B_1(t) - B_3(t)|dt \\ &\quad + \int_0^1 |B_2(t) - B_3(t)|dt \end{aligned}$$

Possible via Feynman - Kac?

- (3,2,1) - Shell Sort ($n \to 3n$)

Data : $X_1, Y_1, Z_1, X_2, Y_2, Z_2, \ldots, X_n, Y_n, Z_n$

X_i's i.i.d. $U(0,1)$; Y_i's i.i.d. $U(0,1)$;
Z_i's i.i.d. $U(0,1)$

3 - Stage contributes:

$$C_n^1 + C_n^2 + C_n^3 \qquad \text{comparisons}$$

2 - Stage contributes:

$$\tilde{C}^1_{\lceil 3n/2 \rceil} + \tilde{C}^2_{\lfloor 3n/2 \rfloor} \qquad \text{comparisons}$$

where

$$\tilde{C}^j_m =_d I_m \qquad \text{from (3,1) - Shell sort}$$

1 - Stage contributes?

How many inversions remaining
in the resulting 3-sorted and
2-sorted list?

for each point X_i, Y_i, or Z_i, associate a triple giviing the parity of the numbers of X's, Y's, and Z's that precede the point; e.g. "EOE" means:

of X-predecessors is Even
of Y-predecessors is Odd
of Z-predecessors is Even

Proposition. An inersion in the 3-sorted and 2-sorted list ocurs when the point causing the inversion is of type EOE or OEO. That is,

$I_{3n} = N_4 \equiv$ # of points of type EOE or OEO

Also let

$$
\begin{aligned}
N_1 &= \text{\# of points of types } OOE \text{ or } EEO \\
N_2 &= \text{\# of points of types } OEE \text{ or } EOO \\
N_3 &= \text{\# of points of types } OOO \text{ or } EEE
\end{aligned}
$$

$$N_1 + N_2 + N_3 + N_4 = 3n$$

Symmetry yields $E(N_j) = 3n/4$, $j = 1, \ldots, 4$,

$$\begin{aligned} 0 &= Cov(N_1, N_1 + N_2 + N_3 + N_4) \\ &= Var(N_1) + 3\,Cov(N_1, N_2) \end{aligned}$$

Example:

$$\begin{array}{ccccccccccc} \text{R} & = & 3 & 2 & 6 & 5 & 9 & 8 & 1 & 4 & 7 \\ & & \text{E} & \text{O} & \text{O} & \text{O} & \text{O} & \text{O} & \text{O} & \text{E} & \text{O} \\ & & \text{O} & \text{O} & \text{E} & \text{E} & \text{O} & \text{E} & \text{E} & \text{E} & \text{E} \\ & & \text{E} & \text{O} & \text{O} & \text{E} & \text{O} & \text{O} & \text{E} & \text{E} & \text{E} \end{array}$$

so $N_4 = 3$

3 - sorted and 2 - sorted:

$$R' \;=\; 1 \quad 2 \quad 4 \quad 3 \quad 5 \quad 7 \quad 6 \quad 9 \quad 8$$

Theorem (Smythe and Wellner, 2002).

$$\begin{aligned}\frac{I_n - 3n/4}{\sqrt{n}} &= \frac{N_4 - 3n/4}{\sqrt{n}} \\ &\to_d N(0, \frac{9}{32})\end{aligned}$$

Proof:

- Poissonize!
- Prove CLT via Markov chain.
- De-Poissonize - this is the hard part.

Let $\tilde{N}_4 \equiv \#$ of EOE's and OEO's in Poissonized process.

$$\frac{\tilde{N}_4 - 3n/4}{\sqrt{n}} \to_d N(0, 9/32)$$

fairly easily, but we need a conditional CLT:

$$\left(\frac{\tilde{N}_4 - 3n/4}{\sqrt{n}}\middle|\underline{M}_{3n} = (n, n, n)\right) \to_d N(0, 9/32)$$

with

$$\begin{aligned}\underline{M}_{3n} &= (M^X_{3n}, M^Y_{3n}, M^Z_{3n}) \\ &= (\#\text{ of } X_i\text{'s}, Y_i\text{'s}, Z_i\text{'s in } 3n \text{ steps})\end{aligned}$$

To prove the conditional CLT, consider a 12-state Markov chain recording X, Y, Z state and e_1, e_2, e_3, e_4 where

$$\begin{aligned} e_1 &= \{OOE, EEO\}, \quad e_2 = \{OEE, EOO\}, \\ e_3 &= \{OOO, EEE\}, \quad e_4 = \{EOE, OEO\} \end{aligned}$$

Let

$$\begin{aligned}
\underline{C}_{3n} &= (C_{3n,1}, \ldots, C_{3n,12}) \\
&= (\#\text{ of visits to } e_1 \cap X, && (1) \\
&\quad \#\text{ of visits to } e_1 \cap Y, && (2) \\
&\quad \#\text{ of visits to } e_1 \cap Z, && (3) \\
&\quad . \\
&\quad . \\
&\quad . \\
&\quad \#\text{ of visits to } e_4 \cap X, && (10) \\
&\quad \#\text{ of visits to } e_4 \cap Y, && (11) \\
&\quad \#\text{ of visits to } e_4 \cap Z) && (12)
\end{aligned}$$

Proposition:

$$\frac{\underline{C}_{3n} - (3n/12)\underline{1}}{\sqrt{n}} \to_d N(0, \Sigma)$$

where

$$\Sigma = \frac{1}{96}\Sigma_0$$

and

$$\Sigma_0 =$$

$$\begin{array}{rrrrrrrrrrrr}
19 & -5 & -5 & -5 & 1 & -5 & -5 & -5 & 1 & 7 & 1 & 1 \\
-5 & 19 & -5 & 1 & 7 & 1 & -5 & -5 & 1 & 1 & -5 & -5 \\
-5 & -5 & 19 & -5 & 1 & -5 & 1 & 1 & 7 & 1 & -5 & -5 \\
-5 & 1 & -5 & 19 & -5 & -5 & 7 & 1 & 1 & -5 & -5 & 1 \\
1 & 7 & 1 & -5 & 19 & -5 & 1 & -5 & -5 & -5 & -5 & 1 \\
-5 & 1 & -5 & -5 & -5 & 19 & 1 & -5 & -5 & 1 & 1 & 7 \\
-5 & -5 & 1 & 7 & 1 & 1 & 19 & -5 & -5 & -5 & 1 & -5 \\
-5 & -5 & 1 & 1 & -5 & -5 & -5 & 19 & -5 & 1 & 7 & 1 \\
1 & 1 & 7 & 1 & -5 & -5 & -5 & -5 & 19 & -5 & 1 & -5 \\
7 & 1 & 1 & -5 & -5 & 1 & -5 & 1 & -5 & 19 & -5 & -5 \\
1 & -5 & -5 & -5 & -5 & 1 & 1 & 7 & 1 & -5 & 19 & -5 \\
1 & -5 & -5 & 1 & 1 & 7 & -5 & 1 & -5 & -5 & -5 & 19
\end{array}$$

Proof of the conditional CLT:

- local limit theorem for Markov chains
  Kolmogorov (1949)

Upshot for (3,2,1) - Shell sort:

$$\begin{aligned} S_n &= C_n^1 + C_n^2 + C_n^3 && (\mathit{stage}\ 3) \\ &\quad + \tilde{C}^1_{3n/2} + \tilde{C}^2_{3n/2} && (\mathit{stage}\ 2) \\ &\quad + 3n + I_{3n} \end{aligned}$$

where

$$\frac{C_n^j - n^2/4}{n^{3/2}} \to_d T_j \sim N(0, 1/36), \quad j = 1,2,3$$

$$\frac{\tilde{C}^j_{3n/2}}{(3n/2)^{3/2}} \to_d \frac{W_{3,j}}{3^{3/2}}, \quad j = 1,2,$$

$$\frac{I_{3n} - 3n/4}{\sqrt{n}} \to_d N(0, 9/32)$$

Hence

$$\frac{S_{3n} - 3n^2/4}{n^{3/2}} \to_d N(0,1/12) + 2^{3/2}(W_{3,1} + W_{3,2})$$

Thus I_{3n} doesn't contribute to the limit distribution!

4. Problems

A. Limit distribution for (h,2,1) - Shell sort for $h = 5$ or $h = 7$?

B. Distribution of W_h for $h \geq 3$?

C. Limit distributions for $(h_n, 1)-$Shell sort with $h_n = O(n^{1/3})$?

D. Choice of h, k for (h,k,1) - Shell sort to minimize combinations of $E(S_n)$ and $Var(S_n)$.

E. Relationship of sorting to shuffling? In the RSK correspondence, we can go from sorting to shuffling and vice-versa via the bijective correspondence of permutations with pairs of Young tableaux. What additional knowledge is needed to turn a sorting method into a method of shuffling? What additional knowledge is needed to turn a method of shuffling into a method of sorting?

Selected References

- Johnson, B. McK. and Killeen, T. (1983). An explicit formula for the cdf of the L_1 norm of the Brownian bridge. *Ann. Prob.* **11**, 807 - 808.

- Kolmogorov, A. N. (1962). A local limit theorem for Markov chains. In *Select. Transl. Math. Statist. and Probability* **2**, 109-129. American Mathematical Society, Providence, R. I. Translation of a Russian article Izv. Aka. Nauk SSSR Ser. Mat. **13** (1949), 281 - 300.

- Louchard, G. (1986). Brownian motion and algorithm complexity. *BIT* **26**, 17 - 24.

- Mahmoud, H. M (2000). *Sorting: A Distribution Theory*. Wiley, New York

- Rice, S. O. (1982). The integral of the absolute value of the pinned Wiener process - calculation of its probability density by numerical integration. *Ann. Prob.* **10**, 240 - 243.

- Shepp, L. A. (1982). On the integral of the absolute value of the pinned Wiener process. *Ann. Prob.* **10**, 234 - 239.

- Smythe, R. T. and Wellner, J. A. (2001). Stochastic analysis of Shell sort. *Algorithmica* **31**, 442 - 457.

- Smythe, R. T. and Wellner, J. A. (2002). Asymptotic analysis of (3,2,1)-Shell sort. *Random Structures and Algorithms*, to appear.