

A Practical and Worst-Case Efficient Algorithm for Divisor Methods of Apportionment

Raphael Reitzig* Sebastian Wild*

December 6, 2017

Proportional apportionment is the problem of assigning seats to parties according to their relative share of votes. Divisor methods are the de-facto standard solution, used in many countries.

In recent literature, there are two algorithms that implement divisor methods: one by Cheng and Eppstein [CE14] has worst-case optimal running time but is complex, while the other [Puk14] is relatively simple and fast in practice but does not offer worst-case guarantees.

We demonstrate that the former algorithm is much slower than the other in practice and propose a novel algorithm that avoids the shortcomings of both. We investigate the running-time behavior of the three contenders in order to determine which is most useful in practice.

1. Introduction

The problem of proportional apportionment arises whenever we have a finite supply of k indivisible, identical resource units which we have to distribute across n parties *fairly*, that is according to the proportional share of publicly known and agreed-upon values v_1, \dots, v_n (of the sum $V = \sum v_i$ of these values). We elaborate in this section on applications of and solutions for this problem.

*Department of Computer Science, University of Kaiserslautern; {reitzig, wild}@cs.uni-kl.de

1. Introduction

Apportionment arises naturally in politics. Here are two prominent examples:

- In a proportional-representation electoral system we have to assign seats in parliament to political parties according to their share of all votes.

The resources are seats, and the values are vote counts.

- In federal states the number of representatives from each component state often reflects the population of that state, even though there will typically be at least one representative for any state no matter how small it is.

Resources are again seats, values are the numbers of residents.

In order to use consistent language throughout this article, we will stick to the first metaphor. That is, we assign k seats to parties $[1..n]$ proportionally to their respective votes v_i , and we call k the *house size*.

A fair allocation should assign v_i/V seats to party i , where $V = v_1 + \dots + v_n$ is the total vote count of all parties. In case of electoral systems which exclude parties below a certain threshold of overall votes from seat allocation altogether, we assume they have already been removed from our list of n parties.

As seats are indivisible, this is only possible if, by chance, all v_i/V are integers; otherwise we have to come up with some rounding scheme. This is where *apportionment methods* come into play. The books by Balinski and Young [BY01] and Pukelsheim [Puk14] give comprehensive introductions into the topic with its historical, political and mathematical dimensions.

Mathematically speaking, an apportionment method is a function $f : \mathbb{R}_{>0}^n \times \mathbb{N} \rightarrow \mathbb{N}_0^n$ that maps vote counts $\mathbf{v} = (v_1, \dots, v_n)$ and house size k to a seat allocation $\mathbf{s} = (s_1, \dots, s_n) := f(\mathbf{v}, k)$ so that $s_1 + \dots + s_n = k$. We interpret \mathbf{s} as party i getting s_i seats.

There are many conceivable such methods, but there are at least three natural properties one would like apportionment systems to have:

- (P1) *Pairwise vote monotonicity*: When votes change, f should not take away seats from a party that has gained votes while at the same time awarding seats to one that has lost votes.
- (P2) *House monotonicity*: f should not take seats away from any party when the house grows (in number of seats) but votes do not change.
- (P3) *Quota rule*: The number of seats of each party should be its proportional share, rounded either up or down.

Balinski and Young have shown that

- (P1) implies (P2) [BY01, Cor. 4.3.1],
- no method can always guarantee (P1) and (P3) [BY01, Thm. 6.1], and

1. Introduction

Method	Divisor Sequence	$\delta(x)$	Sandwich
Smallest divisors	0, 1, 2, 3, ...	x	—
Greatest divisors	1, 2, 3, 4, ...	$x + 1$	—
Sainte-Laguë	1, 3, 5, 7, ...	$2x + 1$	—
Modified Sainte-Laguë	1.4, 3, 5, 7, ...	$\begin{cases} 2x+1 & x \geq 1 \\ 1.6x+1.4 & x < 1 \end{cases}$	$2x + \frac{6}{5} \pm \frac{1}{5}$
Equal Proportions	0, $\sqrt{2}$, $\sqrt{6}$, $\sqrt{12}$, ...	$\sqrt{x(x+1)}$	$x + \frac{1}{4} \pm \frac{1}{4}$
Harmonic Mean	0, $\frac{4}{3}$, $\frac{12}{5}$, $\frac{24}{7}$, ...	$\frac{2x(x+1)}{2x+1}$	$x + \frac{1}{4} \pm \frac{1}{4}$
Imperiali	2, 3, 4, 5, ...	$x + 2$	—
Danish	1, 4, 7, 10, ...	$3x + 1$	—

Table 1: Commonly used divisor methods [CE14, Table 1]. For each of the methods, we give a possible continuation δ of the respective divisor sequence (cf. Section 2) as well as linear sandwich bounds on δ , if non-trivial (cf. Lemma 2).

- (P1) holds exactly for *divisor methods* [BY01, Thm. 4.3].

Property (P1) is essential for upholding the principle of “one-person, one-vote”, an ideal pursued by electoral systems around the globe and occasionally enforced by law [Puk14, Section 2.4]. Therefore, divisor a. k. a. Huntington methods can be the only choice, for the price of (P3). Other choices can be made, of course; the aforementioned books [BY01; Puk14] discuss different trade-offs.

Divisor methods are characterized by *divisor sequences* which control the notion of “fairness” implemented by the respective method. There are many popular choices (cf. Table 1). It is not per se clear which divisor sequence is the best; there still seems to be active discussion, e. g., for the U. S. House of Representatives. One reason is that no-one has yet been able to propose a convincing, universally agreed-upon mathematical criterion that would single out one method as superior to the others. In fact, there are competing notions of fairness, each favoring a different divisor method [BY01, Section A.3]. A reasonable approach is therefore to run computer simulations of different methods and compare their outcomes empirically, for example w. r. t. the distribution of final average votes per seat v_i/s_i . For this purpose, many apportionments have to be computed, so efficient algorithms can become an issue.

We thus study the problem of computing the final seat allocation by divisor methods (given by their divisor sequences) according to vote counts and house size.

For the case of almost linear divisor sequences, the problem can be solved in time $O(n)$; this has been shown by Cheng and Eppstein [CE14] who propose a worst-case running-time-optimal algorithm which we call CHENGEPPSTEINSELECT. It is quite involved and rather difficult to implement (cf. Appendix C.3).

2. Divisor Methods Formalized

Pukelsheim [Puk14], on the other hand, proposes algorithm JUMPANDSTEP whose running time is not asymptotically optimal in the worst case but tends to perform well in practice, at least if some insight about the used divisor sequence is available and inputs are good-natured (cf. Appendix C.2).

After introducing divisor methods formally in Section 2, we propose a new algorithm in Section 3 that also attains the $O(n)$ worst-case running time bound but is straight-forward to implement *and* efficient in practice as well. It is based on a generalization of our solution for the envy-free stick-division problem [RW15b].

We finally compare the performance of the three contending algorithms with extensive running time experiments, an executive summary of which we give in Section 4.

Additional material includes an index of notation in Appendix F.

2. Divisor Methods Formalized

Let $d = (d_j)_{j=0}^{\infty}$ be an arbitrary divisor sequence, i. e. a nonnegative, strictly increasing and unbounded sequence of real numbers. We formally set $d_{-1} := -\infty$.

We require that there is a smooth continuation of d on the reals which is easy to invert. That is, we assume a function $\delta : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq d_0}$ with

- i) δ is continuous and strictly increasing,
- ii) $\delta^{-1}(x)$ for $x \geq d_0$ can be computed with a constant number of arithmetic operations, and
- iii) $\delta(j) = d_j$ (and thus $\delta^{-1}(d_j) = j$) for all $j \in \mathbb{N}_0$.

All the divisor sequences used in practice fulfill these requirements; cf. Table 1. For convenience, we continue δ^{-1} on the complete real line requiring

- iv) $\delta^{-1}(x) \in [-1, 0)$ for $x < d_0$.

Corollary 1: *Assuming i) to iv), $\delta^{-1}(x)$ is continuous and strictly increasing on $\mathbb{R}_{\geq d_0}$. Furthermore, it is the inverse of $j \mapsto d_j$ in the sense that*

$$\lfloor \delta^{-1}(x) \rfloor = \max\{j \in \mathbb{Z}_{\geq -1} \mid d_j \leq x\}$$

for all $x \in \mathbb{R}$. □

In particular, $\lfloor \delta^{-1}(x) \rfloor = j$ for $d_j \leq x < d_{j+1}$ so the floored δ^{-1} is the (zero-based) *rank* function for the set of all d_j as long as $x \geq d_0$.

Note how this reproduces what is called *d-rounding* in the literature [BY01; Puk14]; we obtain an efficient way of calculating this function via δ^{-1} .

2. Divisor Methods Formalized

Now the set of all seat assignments that are valid w. r. t. d is given by [BY01]

$$\mathcal{S}(\mathbf{v}, k) = \left\{ \mathbf{s} \in \mathbb{N}_0^n \mid \sum_{i=1}^n s_i = k \wedge \exists a > 0. \forall i \in [1..n]. s_i \in [\delta^{-1}(v_i \cdot a)] + \{0, 1\} \right\}.$$

We call a realization of a *proportionality constant* a^* ; intuitively, every seat corresponds to roughly $1/a^*$ votes.

An equivalent definition is by the set of possible results of the following algorithm [BY01, Prop. 3.3].

Algorithm 1: ITERATIVEMETHOD $_d(\mathbf{v}, k)$:

Step 1 Initialize $\mathbf{s} = 0^n$.

Step 2 While $k > 0$,

Step 2.1 Determine $I = \arg \min_{i=1}^n d_{s_i}/v_i$.

Step 2.2 Update $s_I \leftarrow s_I + 1$ and $k \leftarrow k - 1$.

Step 3 Return \mathbf{s} .

We can obtain a proportionality constant [Puk14, 59f] by

$$a^* = \max\{d_{s_i-1}/v_i \mid 1 \leq i \leq n\}, \tag{1}$$

which in turn defines the set $\mathcal{S}(\mathbf{v}, k)$.

Note that we work with d_j/v_i instead of v_i/d_j in the classical literature; Cheng and Eppstein [CE14] and we prefer the reciprocals because the case $d_0 = 0$ then handles gracefully and without special treatment. Therefore, our a^* is also the reciprocal of the proportionality constant as e. g. Pukelsheim [Puk14] defines it, we multiply by a in the definition of \mathcal{S} and we take the minimum in ITERATIVEMETHOD. It is important to note that the defined set \mathcal{S} remains unchanged by this switch.

Following the notation of Cheng and Eppstein [CE14], we furthermore define for given votes $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{Q}_{>0}^n$ the sets

$$A_i := \left\{ a_{i,j} \mid j = 0, 1, 2, \dots \right\} \quad \text{with} \quad a_{i,j} := \frac{d_j}{v_i}$$

and their multiset union

$$\mathcal{A} := \bigsqcup_{i=1}^n A_i.$$

As we will see later, the relative rank of elements in \mathcal{A} turns out to be of interest; we therefore define the *rank function* $r(x, \mathcal{A})$ which denotes the number of elements in

3. Fast Apportionment by Rank Selection

multiset \mathcal{A} that are no larger than x , that is

$$r(x, \mathcal{A}) := |\mathcal{A} \cap (-\infty, x]| = \sum_{i=1}^n |\{a_{i,j} \in \mathcal{A} \mid a_{i,j} \leq x\}|. \quad (2)$$

We write $r(x)$ instead of $r(x, \mathcal{A})$ when \mathcal{A} is clear from context.

We need two more convenient shorthands: Assuming we have $a^* \leq \bar{x}$, we denote with

$$I_{\bar{x}} := \{i \in \{1, \dots, n\} \mid v_i > d_0/\bar{x}\} \quad (3)$$

the set of parties that can hope for a seat, and with

$$\mathcal{A}^{\bar{x}} := \bigsqcup_{i \in I_{\bar{x}}} \left\{ \frac{d_j}{v_i} \in \mathcal{A} \mid \frac{d_j}{v_i} < \bar{x} \right\} = \bigsqcup_{i=1}^n \left\{ \frac{d_j}{v_i} \in \mathcal{A} \mid \frac{d_j}{v_i} < \bar{x} \right\} = \mathcal{A} \cap (-\infty, \bar{x}) \quad (4)$$

the multiset of elements from sequences of these parties that are smaller than \bar{x} , i. e. reasonable candidates for a^* .

3. Fast Apportionment by Rank Selection

From (1) together with strict monotonicity of d , we obtain immediately that $a^* = \mathcal{A}_{(k)}$, i. e. the k th smallest element of \mathcal{A} (counting duplicates) is a suitable proportionality constant. This allows us to switch gears from the iteration-based world of Pukelsheim [Puk14] to selection-based algorithms, as previously seen by Cheng and Eppstein [CE14].

Note that even though \mathcal{A} is infinite, $\mathcal{A}_{(k)}$ always exists because the terms $a_{i,j} = d_j/v_i$ are strictly increasing in j for all $i \in \{1, \dots, n\}$.

Borrowing terminology from the field of mathematical optimization, we call a *feasible* if $r(a) \geq k$, otherwise it is *infeasible*. Feasible $a \neq a^*$ are called *suboptimal*. Our goal is to find a subset of \mathcal{A} that contains a^* but as few infeasible or suboptimal a as possible; we can then apply a rank-selection algorithm on this subset and obtain (via a^*) the solution to the apportionment problem.

Now since d is unbounded, setting *any* upper bound \bar{x} on the $a_{i,j}$ yields a finite search space $\mathcal{A}^{\bar{x}}$. By choosing any such bound that maintains $|\mathcal{A}^{\bar{x}}| \geq k$, we retain the property that a^* is the k th smallest element under consideration.

One naive way is to make sure that the party with the most votes (which should get the most seats) contributes at least k values to \mathcal{A} . This can be achieved by letting $\bar{x} = d_{k-1}/\max \mathbf{v} + \varepsilon$ (cf. the proof of Theorem 3). This alone, however, leads only to an algorithm with worst-case running time in $\Theta(kn)$, which is worse than even ITERATIVEMETHOD (with priority queues).

We can actually not improve this upper bound \bar{x} ; it is tight for the case that one party has many more votes than all others and gets (almost) all of the seats. We can, however,

3. Fast Apportionment by Rank Selection

exclude many individual elements in $\mathcal{A}^{\bar{x}}$ because they are too small to be feasible or too large to be optimal.

Towards finding suitable upper and lower bounds on a^* , we investigate its rank in the multiset \mathcal{A} of all candidates. All we know is that

$$k \leq r(a^*) \leq k + |I_{\bar{x}}|$$

since we may have any number between one and $|I_{\bar{x}}|$ parties that tie for the last seat. We can still make an ansatz with $r(\bar{a}) \geq k + |I_{\bar{x}}|$ and $r(\underline{a}) < k$, express rank function r in terms of δ^{-1} (cf. Lemma 4 in Appendix B) and derive that

$$\sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \underline{a}) \leq k - |I_{\bar{x}}| \quad \text{and} \quad \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \bar{a}) \geq k. \quad (5)$$

This pair of inequalities is indeed a sufficient condition for admissible pairs of bounds (\underline{a}, \bar{a}) ; we can conclude that $\underline{a} \leq a^* \leq \bar{a}$. For a formal proof, see Lemma 5 in Appendix B.

We now want to derive a sandwich on a^* by fulfilling the inequalities in (5) as tightly as possible. Depending on δ^{-1} , this may be hard to do analytically. However, we can make the same assumption as Cheng and Eppstein [CE14] and explicitly compute suitable bounds for divisor sequences which behave roughly linearly. This does not limit the scope of our investigation by much; see Appendix A for more on this.

Lemma 2: *Assume the continuation δ of divisor sequence d fulfills*

$$\alpha x + \underline{\beta} \leq \delta(x) \leq \alpha x + \bar{\beta}$$

for all $x \in \mathbb{R}_{\geq 0}$ with $\alpha > 0$, $\underline{\beta} \in [0, \alpha]$ and $\bar{\beta} \geq 0$. Let further some $\bar{x} > a^*$ be given. Then, the pair (\underline{a}, \bar{a}) defined by

$$\underline{a} := \max \left\{ 0, \frac{\alpha k - (\alpha - \underline{\beta}) \cdot |I_{\bar{x}}|}{V_{\bar{x}}} \right\} \quad \text{and} \quad \bar{a} := \frac{\alpha k + \bar{\beta} \cdot |I_{\bar{x}}|}{V_{\bar{x}}}$$

with $V_{\bar{x}} := \sum_{i \in I_{\bar{x}}} v_i$ fulfills the conditions of Lemma 5, that is $\underline{a} \leq a^* \leq \bar{a}$. Moreover,

$$|\mathcal{A} \cap [\underline{a}, \bar{a}]| \leq 2 \left(1 + \frac{\bar{\beta} - \underline{\beta}}{\alpha} \right) \cdot |I_{\bar{x}}|.$$

The proof consists mostly of rote calculation towards applying Lemma 5; see Appendix B for the details.

We have now derived our main improvement over the work by Cheng and Eppstein [CE14]; where they have only a one-sided bound on a^* and thus have to employ an involved search on \mathcal{A} , we have sandwiched a^* from both sides, and so tightly that the remaining search space is small enough for a simple rank selection to be efficient.

3. Fast Apportionment by Rank Selection

Building on the bounds from Lemma 2, we can improve upon the naive idea using only \bar{x} by excluding also many more elements from \mathcal{A} which are for sure not a^* . Since we remove in particular too small elements, this means that we also have to modify the rank we select; we will see that our bounds are chosen so that we can use δ^{-1} to *count* the number of elements we discard *exactly*.

Recall that we assume a fixed apportionment scheme, that is fixed d with known α , $\underline{\beta}$ and $\bar{\beta}$ as per Lemma 2.

Algorithm 2: SANDWICHSELECT(\mathbf{v}, k) $_d$:

- Step 1** Find the $v^{(1)} = \max\{v_1, \dots, v_n\}$.
- Step 2** Set $\bar{x} := d_{k-1}/v^{(1)} + \varepsilon$ for suitable¹ constant $\varepsilon > 0$.
- Step 3** Compute $I_{\bar{x}}$ as per (3).
- Step 4** Compute \underline{a} and \bar{a} as per Lemma 2.
- Step 5** Initialize $\hat{\mathcal{A}} := \emptyset$ and $\hat{k} := k$.
- Step 6** For all $i \in I_{\bar{x}}$, do:
 - Step 6.1** Compute $\underline{j} := \max\{0, \lceil \delta^{-1}(v_i \cdot \underline{a}) \rceil\}$ and $\bar{j} := \lfloor \delta^{-1}(v_i \cdot \bar{a}) \rfloor$.
 - Step 6.2** Add all d_j/v_i to $\hat{\mathcal{A}}$ for which $\underline{j} \leq j \leq \bar{j}$.
 - Step 6.3** Update $\hat{k} \leftarrow \hat{k} - j$.
- Step 7** Select and return $\hat{\mathcal{A}}_{(\hat{k})}$.

Theorem 3:

Algorithm 2 computes a^* in time $O(n)$ for any divisor sequence d that fulfills the requirements of Lemma 2.

Proof: First, we have to show that $I_{\bar{x}}$ as we compute it in Steps 1-3 is correct. We have $\bar{x} > a^* = \mathcal{A}^{\bar{x}}$ as already $r(\bar{x} - \varepsilon) = r(d_{k-1}/v^{(1)}) \geq k$; at least the k elements $\frac{d_0}{v^{(1)}}, \dots, \frac{d_{k-1}}{v^{(1)}} \in \mathcal{A}$ are no larger than $d_{k-1}/v^{(1)}$. We thus never need to consider elements $a \geq \bar{x}$, and in particular $\mathcal{A}_{(k)} = \mathcal{A}_{(k)}^{\bar{x}}$ as $\mathcal{A}^{\bar{x}} = \mathcal{A} \cap (-\infty, \bar{x})$.

So far, we have needed no additional restriction on ε in Step 2; we only need it to be positive so we do not discard a^* by accident if it is exactly $d_{k-1}/v^{(1)}$. However, the size of $\mathcal{A}^{\bar{x}}$ can be arbitrarily large – depending on the input values v_i which we do not want. Therefore, we require

$$0 < \varepsilon < \frac{d_k - d_{k-1}}{v^{(1)}}; \tag{6}$$

¹Neither correctness nor Θ -running-time is affected by the choice of ε here since it affects only the size of $I_{\bar{x}}$, which is bounded by n in any case. In particular, the size of $\hat{\mathcal{A}}$ is affected only up to a constant factor. For tweaking performance in practice, see the proof of Theorem 3.

4. Comparison of Algorithms

such exists because d is strictly increasing. Note how then $\bar{x} < d_k/v^{(1)}$ so we do not keep any additional suboptimal values.

From Step 4 on, we then construct multiset $\hat{\mathcal{A}} \subseteq \mathcal{A}$ as the subsequent union of $A_i \cap [\underline{a}, \bar{a}]$, that is

$$\begin{aligned}
\hat{\mathcal{A}} &= \biguplus_{i \in I_{\bar{x}}} \left\{ \frac{d_j}{v_i} \mid \underline{j}(i) \leq j \leq \bar{j}(i) \right\} \\
&= \biguplus_{i \in I_{\bar{x}}} \left\{ \frac{d_j}{v_i} \in \mathcal{A} \mid \delta^{-1}(v_i \cdot \underline{a}) \leq j \leq \delta^{-1}(v_i \cdot \bar{a}) \right\} \\
&= \biguplus_{i \in I_{\bar{x}}} \left\{ \frac{d_j}{v_i} \in \mathcal{A} \mid v_i \cdot \underline{a} \leq d_j \leq v_i \cdot \bar{a} \right\} \\
&= \biguplus_{i \in I_{\bar{x}}} \left\{ \frac{d_j}{v_i} \in \mathcal{A} \mid \underline{a} \leq \frac{d_j}{v_i} \leq \bar{a} \right\} \\
&= \mathcal{A} \cap [\underline{a}, \bar{a}].
\end{aligned}$$

In particular, the last step follows from (4) with $\bar{x} > a^*$. By Lemma 2, we know that $\underline{a} \leq a^* \leq \bar{a}$ for the bounds computed in Step 4, so we get in particular that $a^* \in \hat{\mathcal{A}}$.

It remains to show that we calculate \hat{k} correctly. Clearly, we discard with $(a_{i,0}, \dots, a_{i,\underline{j}-1})$ exactly \underline{j} elements in Step 6.2, that is $|A_i \cap (-\infty, \underline{a})| = \underline{j}(i)$. Therefore, we compute with

$$\hat{k} = k - \sum_{i \in I_{\bar{x}}} |A_i \cap (-\infty, \underline{a})| = r(a^*, \mathcal{A}) - |\mathcal{A} \cap (-\infty, \underline{a})| = r(a^*, \hat{\mathcal{A}})$$

the correct rank of a^* in $\hat{\mathcal{A}}$.

For the running time, we observe that the computations in steps 1 to 5 are easily done with $O(n)$ primitive instructions. The loop in Step 6 and therewith steps 6.1 and 6.3 are executed $|I_{\bar{x}}| \leq n$ times. The overall number of set operations in Step 6.2 is $|\hat{\mathcal{A}}| \in O(|I_{\bar{x}}|) \subseteq O(n)$ (cf. Lemma 2). Finally, Step 7 runs in time $O(|\hat{\mathcal{A}}|) \subseteq O(n)$ when using a (worst-case) linear-time rank selection algorithm (e. g., the median-of-medians algorithm [Blu+73]). \square

We have obtained a relatively simple algorithm that implements many divisor methods and has optimal asymptotic running time in the worst case. It remains to be seen if it is also efficient in practice.

4. Comparison of Algorithms

We have implemented all algorithms mentioned above in Java [RW15a] with a focus on clarity and performance. Reviewing the algorithms resp. implementations (cf. Appendix C),

4. Comparison of Algorithms

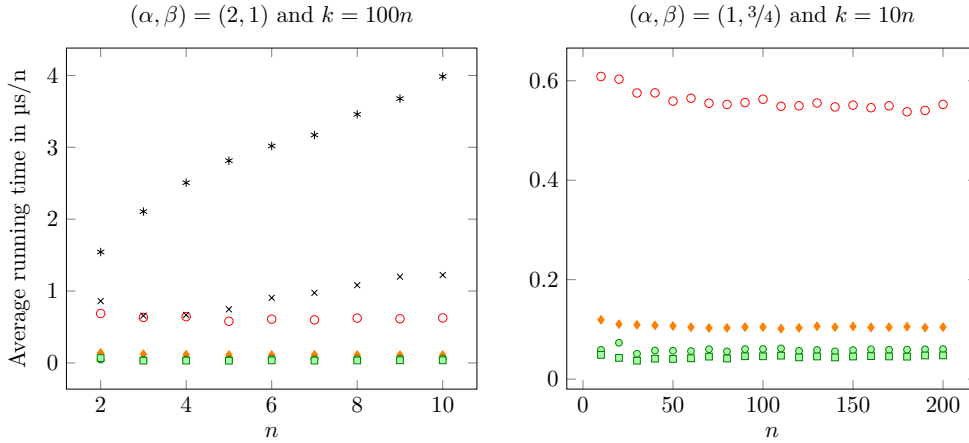


Figure 1: This figure shows average running times of SANDWICHSELECT \blacklozenge , CHENGEPSTEINSELECT \circ , JUMPANDSTEP with naive \blacksquare resp. priority-queue \bullet minimum selection, and ITERATIVEMETHOD with naive \times resp. priority-queue $*$ minimum selection, normalized by the number of parties n . The inputs are random apportionment instances with vote counts v_i drawn i. i. d. uniformly from $[1, 3]$. The numbers of parties n , house size k and method parameters (α, β) have been chosen to resemble national parliaments in Europe (left) and the U. S. House of Representatives (right), respectively.

we observe that neither ITERATIVEMETHOD nor JUMPANDSTEP are asymptotically worst-case efficient whereas CHENGEPSTEINSELECT does not seem to be practical regarding implementability. SANDWICHSELECT does not have either deficiency and is still the shortest of the non-trivial algorithms.

We evaluate relative practical efficiency by performing running time experiments on artificial instances; we fix the number of parties n , house size k and the used divisor method and draw multiple vote vectors \mathbf{v} at random according to different distributions. Where possible, we draw votes from a *continuous* distribution with fixed expectation; this ensures that vote proportions do not devolve to trivial situations as n grows.

In order to keep the parameter space manageable, we use n as free variable and fix k to a multiple of n . For ease of implementation, we restrict ourselves to divisor sequences of the form $(\alpha j + \beta)_{j \in \mathbb{N}_0}$; this still allows us to cover a range of relevant divisor methods at least approximately (cf. Table 1). We describe the machine configuration used for the experiments and further details of the setup in Appendix D.

Figure 1 shows the results of two experiments with practical parameter choices. It is clear that JUMPANDSTEP dominates the field; of the other algorithms, only SANDWICHSELECT comes close in performance. These observations are stable across many parameter choices; see also Appendix E. We will therefore restrict ourselves to JUMPANDSTEP and SANDWICHSELECT in the sequel.

Towards understanding what influences the performance of these algorithms the most,

4. Comparison of Algorithms

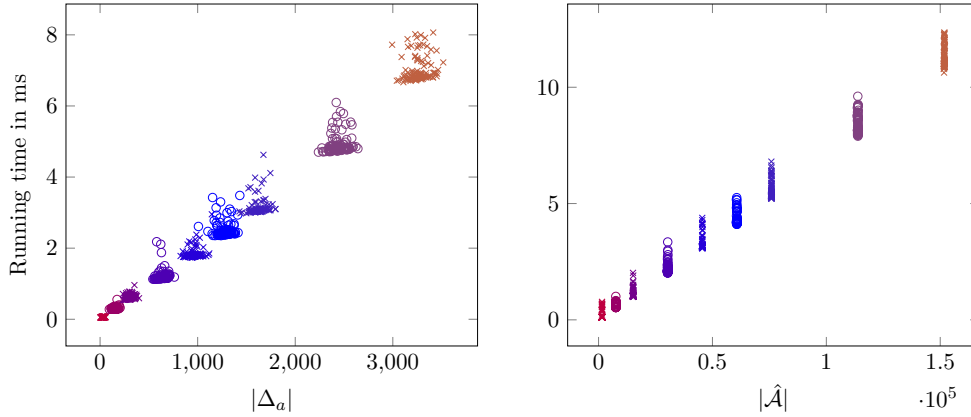


Figure 2: Running times on individual inputs plotted against $|\Delta_a|$ for JUMPANDSTEP (left) resp. $|\hat{\mathcal{A}}|$ for SANDWICHSELECT (right). Inputs are random with exponentially distributed v_i for $n \in \{1 \times, 5 \times, 10 \times, 20 \times, 30 \times, 40 \times, 50 \times, 75 \times, 100 \times\} \cdot 10^3$ and $k = 5n$; they have been apportioned w. r. t. $(\alpha, \beta) = (2, 1)$.

we have investigated how Δ_a (the number of seats JUMPANDSTEP assigns too much, i. e. $k - \sum s_i$) resp. $|\hat{\mathcal{A}}|$ (the number of candidates SANDWICHSELECT selects from) relate to the measured running times. While the connection is clear for SANDWICHSELECT, we need to look at cases where Pukelsheim’s estimators are bad; as long as $|\Delta_a| \ll n$, the $\Theta(n)$ portions of JUMPANDSTEP dominate. Figure 2 exhibits such a setting.

While JUMPANDSTEP is faster than SANDWICHSELECT in the experiments of Figure 1 and similar ones, we observe that SANDWICHSELECT is more robust against changing parameters. Figure 3 exhibits this for switching between different vote distributions: the average running times of SANDWICHSELECT are close to each other where those of JUMPANDSTEP spread out quite a bit. It may be noteworthy that each algorithm has one “outlier” distribution but they are not the same.

JUMPANDSTEP does indeed seem to outperform SANDWICHSELECT consistently so far, if not by much in some cases. We *have* found a parameterization which, even though it is admittedly rather artificial, clearly suggests that JUMPANDSTEP does indeed have $\omega(n)$ worst-case behavior and that SANDWICHSELECT can be faster; see Figure 4. The question after realistic settings for which this is the case remains open.

In summary, we have seen that SANDWICHSELECT provides good performance in a reliable way, i. e., its efficiency does not depend much on divisor sequence or input. On the other hand, JUMPANDSTEP is faster on average when good estimators are available, but can be slower in certain settings.

4. Comparison of Algorithms

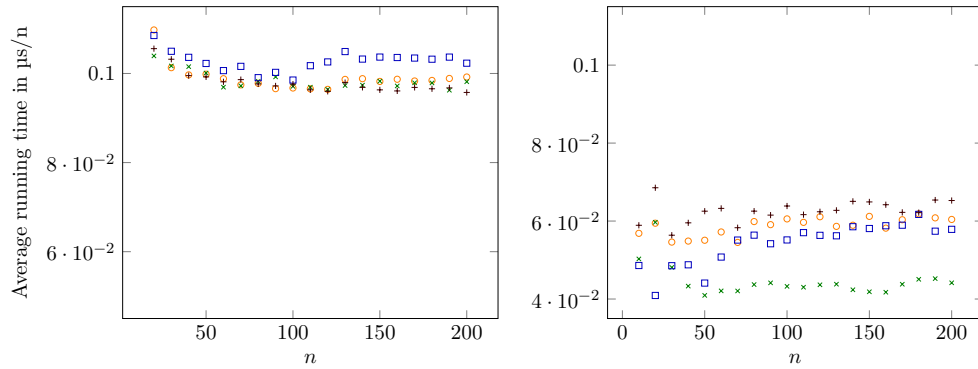


Figure 3: Normalized average runtimes of SANDWICHSELECT (left) and JUMPANDSTEP (right) on v_i drawn randomly from uniform \circ , exponential \times , Poisson \square and Pareto $+$ distributions, respectively, and with $k = 5n$ and $(\alpha, \beta) = (2, 1)$.

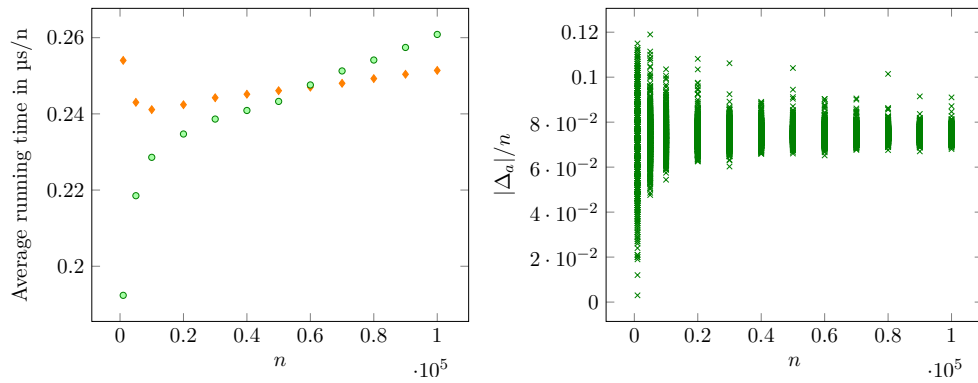


Figure 4: The left plot shows normalized running times of SANDWICHSELECT \diamond and JUMPANDSTEP \circ on instances with $k = 2n$ and Pareto-distributed v_i for $(\alpha, \beta) = (1.0, 0.001)$. The right plot shows that the average of $|\Delta_a|$ seems to converge towards a constant fraction of n in this case.

5. Conclusion

We have derived an algorithm implementing divisor methods of apportionment that is worst-case efficient, simple and practicable. As such, it does not have the shortcomings of previously known algorithms. Even though it can not usually outperform JUMPANDSTEP, its robustness against changing parameters makes it a viable candidate for use in practice.

Acknowledgments

We thank Chao Xu for pointing us towards the work by Cheng and Eppstein [CE14] and noting that the problem of envy-free stick-division [RW15b] is related to proportional apportionment as discussed there. He also observed that our approach for cutting sticks – the core ideas of which turned out to carry over to this article – could be improved to run in linear time.

Furthermore, we owe thanks to an anonymous reviewer whose constructive feedback sparked broad changes which have greatly improved the article over its first incarnation.

References

- [Blu+73] Manuel Blum et al. “Time Bounds for Selection.” In: *Journal of Computer and System Sciences* 7.4 (Aug. 1973), pp. 448–461. DOI: 10.1016/S0022-0000(73)80033-9.
- [BY01] Michel L. Balinski and H. Peyton Young. *Fair Representation. Meeting the Ideal of One Man, One Vote*. 2nd. Brookings Institution Press, 2001. ISBN: 978-0-8157-0111-8.
- [CE14] Zhanpeng Cheng and David Eppstein. “Linear-time Algorithms for Proportional Apportionment.” In: *International Symposium on Algorithms and Computation (ISAAC) 2014*. Springer, 2014. DOI: 10.1007/978-3-319-13075-0_46.
- [GKP94] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics: a foundation for computer science*. Addison-Wesley, 1994. ISBN: 978-0-20-155802-9.
- [Puk14] Friedrich Pukelsheim. *Proportional Representation. Apportionment Methods and Their Applications*. 1st ed. Springer, 2014. ISBN: 978-3-319-03855-1. DOI: 10.1007/978-3-319-03856-8.
- [RW15a] Raphael Reitzig and Sebastian Wild. *Companion Source Code*. revision db43ee7f05. 2015. URL: https://github.com/reitzig/2015_apportionment.
- [RW15b] Raphael Reitzig and Sebastian Wild. *Efficient Algorithms for Envy-Free Stick Division With Fewest Cuts*. 2015. arXiv: 1502.04048.

References

- [SW11] Robert Sedgewick and Kevin Wayne. *Algorithms*. 4th. Addison-Wesley, 2011. ISBN: 978-0-321-57351-3. URL: <http://algs4.cs.princeton.edu>.

A. Our Scope of different Methods of Apportionment

As we have seen in Section 2 there are many possible divisor sequences. For our main result (cf. page 7) we follow Cheng and Eppstein [CE14] and require the sequences to be “almost” linear; we should check that we do not unduly restrict the scope of our investigation.

We refer to the recent reference work by Pukelsheim [Puk14] and, by extension, to Balinski and Young [BY01] who classify different divisor methods of apportionment in terms of *signpost* sequences, a concept equivalent to the divisor sequences we use. They distinguish these classes of such sequences (cf. [Puk14, Sections 3.11-12]):

- *stationary* sign-posts of the form $s(n) = n - 1 + r$ with $r \in (0, 1)$;
- *power-mean* sign-posts defined by

$$\begin{aligned}\tilde{s}_p(0) &= 0, \\ \tilde{s}_p(n) &= \left(\frac{(n-1)^p + n^p}{2} \right)^{1/p},\end{aligned}$$

for $p \neq -\infty, 0, \infty$;

- and special cases $\tilde{s}_{-\infty}(n) = n - 1$, $\tilde{s}_0(n) = \sqrt{(n-1)n}$, and $\tilde{s}_{\infty}(n) = n$.

It is easy to see that stationary sign-posts correspond to divisor sequences $d_j = j + \beta$ with $\beta \in (0, 1)$ (up to a shift by one); as such, Lemma 2 applies immediately with $\alpha = 1$ and $\underline{\beta} = \bar{\beta} = \beta$, and yields a particularly nice (and tight, for our choices of \underline{a} and \bar{a}) upper bound on the size of the candidate set \mathcal{A} . We cover the special cases as well; see Table 1 for the corresponding sandwich bounds.

As for the remaining power-mean sign-posts, the trivial bounds $\underline{\beta} = 0$ and $\bar{\beta} = 1$ already work. One can apply the power-mean inequality and use the slightly better bounds for $p \in \{-\infty, -1, 0, 1, \infty\}$ as given in Table 1. Even better bounds can be gleaned from observing that $\tilde{s}_p(n)$ converges to $n - 1/2$ from one side, and quickly so; $\tilde{s}_p(1)$ thus determines either $\underline{\beta}$ or $\bar{\beta}$ and the other can be chosen as $1/2$.

In summary, our algorithm SANDWICHSELECT applies to all divisor methods treated by Pukelsheim [Puk14] and Balinski and Young [BY01]

B. Lemmata and Proofs

Lemma 4: For rank function $r(x, \mathcal{A})$,

$$r(x, \mathcal{A}) = \sum_{i=1}^n \lfloor \delta^{-1}(v_i \cdot x) \rfloor + 1.$$

B. Lemmata and Proofs

Moreover, for $x < \bar{x}$ we have

$$r(x, \mathcal{A}) = \sum_{i \in I_{\bar{x}}} \lfloor \delta^{-1}(v_i \cdot x) \rfloor + 1$$

with $I_{\bar{x}} = \{i \in \{1, \dots, n\} \mid v_i > d_0/\bar{x}\}$.

B.1. Proof of Lemma 4

By eq. (2) on page 6, it suffices to show that

$$|\{a_{i,j} \mid a_{i,j} \leq x\}| = \lfloor \delta^{-1}(v_i \cdot x) \rfloor + 1$$

for each $i \in \{1, \dots, n\}$. Now, if $x \geq a_{i,j} = d_j/v_i$ for some j , then $v_i \cdot x \geq d_j$, and so $\lfloor \delta^{-1}(v_i \cdot x) \rfloor$ is the largest index j' for which $a_{i,j'} = d_{j'}/v_i \leq x$. As d_j is zero-based, there are $j' + 1 \geq 1$ such elements $a_{i,j} \leq x$ and the equation follows.

Otherwise, that is $a_{i,j} > x$ for all j , we have $j' = \lfloor \delta^{-1}(v_i \cdot x) \rfloor = -1$ by iv) and Corollary 1 and the equality holds with 0 on both sides.

For the second equality, we only have to show that the omitted summands are zero. So let $i \notin I_{\bar{x}}$ be given, that is $v_i \leq d_0/\bar{x}$. For $x < \bar{x}$, we have

$$v_i \cdot x \leq \frac{d_0}{\bar{x}} \cdot x < \frac{d_0}{\bar{x}} \cdot \bar{x} = d_0,$$

and hence $\lfloor \delta^{-1}(v_i \cdot x) \rfloor = -1$ by iv).

Lemma 5: *Let $\bar{x} > a^*$ and assume \bar{a} and \underline{a} are chosen so that they fulfill*

$$\sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \underline{a}) \leq k - |I_{\bar{x}}| \quad \text{and} \quad \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \bar{a}) \geq k.$$

Then, $\underline{a} \leq a^ \leq \bar{a}$.*

The lemma follows more or less directly; one uses the sandwich bounds on r to show that $a < \underline{a}$ are infeasible, i. e., $r(a) < k$, and that \bar{a} is feasible, and thus all $a > \bar{a}$ are suboptimal since a^* is the smallest feasible element in \mathcal{A} .

B.2. Proof of Lemma 5

As a direct consequence of Lemma 4 together with the fundamental bounds $y-1 < \lfloor y \rfloor \leq y$ on floors, we find that

$$\sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot x) < r(x, \mathcal{A}) \leq \sum_{i \in I_{\bar{x}}} (\delta^{-1}(v_i \cdot x) + 1) = |I_{\bar{x}}| + \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot x) \quad (7)$$

B. Lemmata and Proofs

for any \bar{x} and all $x < \bar{x}$. We can therewith pin down the value of r to an interval of width $|I_{\bar{x}}|$ using only δ^{-1} . We can use this to derive upper *and* lower bounds on a^* .

We show that smaller a are infeasible and larger a are clearly suboptimal, so the optimal a^* must lie in between. Let us first consider $a < \underline{a}$. There are two cases: if there is a v_i , such that $v_i a \geq d_0$, we get by strict monotonicity of δ^{-1}

$$\begin{aligned} r(a) &\stackrel{(7)}{\leq} |I_{\bar{x}}| + \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot a) \\ &< |I_{\bar{x}}| + \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \underline{a}) \\ &\leq k \end{aligned}$$

and a is infeasible. If otherwise $v_i a < d_0$, i. e., $a < d_0/v_i$, for all i , a must clearly have rank $r(a) = 0$ as it is smaller than any element $a_{i,j} \in \mathcal{A}$. In both cases we found that $a < \underline{a}$ has rank $r(a) < k$.

Now consider the upper bound, i. e., we have $a > \bar{a}$. In case $\bar{a} \geq \bar{x}$, we have $a > \bar{x} > a^*$ by assumption and any such a cannot be optimal. Otherwise, for $\bar{a} < \bar{x}$, we have

$$r(\bar{a}) \stackrel{(7)}{>} \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \bar{a}) \geq k,$$

so \bar{a} is feasible. Any element $a > \bar{a}$ can thus not be the optimal solution a^* , which is the *minimal* a with $r(a) \geq k$.

B.3. Proof of Lemma 2

We consider the linear divisor sequence continuations

$$\underline{\delta}(j) = \alpha j + \underline{\beta} \quad \text{and} \quad \bar{\delta}(j) = \alpha j + \bar{\beta}$$

for all $j \in \mathbb{R}_{\geq 0}$ and start by noting that the inverses are

$$\underline{\delta}^{-1}(x) = x/\alpha - \underline{\beta}/\alpha \quad \text{and} \quad \bar{\delta}^{-1}(x) = x/\alpha - \bar{\beta}/\alpha$$

for $x \geq \underline{\delta}(0) = \underline{\beta}$ and $x \geq \bar{\delta}(0) = \bar{\beta}$, respectively. For smaller x , we are free to choose the value of the continuation from $[-1, 0)$ (cf. iv)); noting that $x/\alpha - \bar{\beta}/\alpha < 0$ for $x < \bar{\beta}$, a choice that will turn out convenient is

$$\underline{\delta}^{-1}(x) := \max\left\{\frac{x}{\alpha} - \frac{\underline{\beta}}{\alpha}, -1\right\} \quad \text{resp.} \quad \bar{\delta}^{-1}(x) := \max\left\{\frac{x}{\alpha} - \frac{\bar{\beta}}{\alpha}, -1\right\}. \quad (8)$$

We state the following simple property for reference; it follows from $\underline{\delta}(j) \leq \delta(j) \leq \bar{\delta}(j)$ and the definition of the inverses (recall that $\underline{\beta} \leq \alpha$):

$$\frac{x}{\alpha} - \frac{\bar{\beta}}{\alpha} \leq \bar{\delta}^{-1}(x) \leq \delta^{-1}(x) \leq \underline{\delta}^{-1}(x) \leq \frac{x}{\alpha} - \frac{\underline{\beta}}{\alpha}, \quad \text{for } x \geq 0. \quad (9)$$

B. Lemmata and Proofs

Equipped with these preliminaries, we compute

$$\begin{aligned}
\bar{a} &= \frac{\alpha k + \bar{\beta}|I_{\bar{x}}|}{V_{\bar{x}}}. \\
\iff \frac{\bar{a}}{\alpha} \cdot \sum_{i \in I_{\bar{x}}} v_i &= k + \frac{\bar{\beta}}{\alpha} \cdot |I_{\bar{x}}|, \\
\iff k &= \sum_{i \in I_{\bar{x}}} \left(\frac{v_i \cdot \bar{a}}{\alpha} - \frac{\bar{\beta}}{\alpha} \right) \stackrel{(9)}{\leq} \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \bar{a}),
\end{aligned}$$

so \bar{a} satisfies the condition of Lemma 5. Similarly, we find

$$\begin{aligned}
\underline{a} &= \frac{\alpha k - (\alpha - \underline{\beta}) \cdot |I_{\bar{x}}|}{V_{\bar{x}}}, \\
\iff \frac{\underline{a}}{\alpha} \cdot V_{\bar{x}} &= k - (1 - \underline{\beta}/\alpha) \cdot |I_{\bar{x}}|, \\
\iff k &= |I_{\bar{x}}| + \sum_{i \in I_{\bar{x}}} \left(\frac{v_i \cdot \underline{a}}{\alpha} - \frac{\underline{\beta}}{\alpha} \right) \stackrel{(9)}{\geq} |I_{\bar{x}}| + \sum_{i \in I_{\bar{x}}} \delta^{-1}(v_i \cdot \underline{a}),
\end{aligned}$$

that is \underline{a} also fulfills the conditions of Lemma 5.

For the bound on the number of elements falling between \underline{a} and \bar{a} , we compute

$$\begin{aligned}
|\mathcal{A} \cap [\underline{a}, \bar{a}]| &= \sum_{i \in I_{\bar{x}}} |A_i \cap [\underline{a}, \bar{a}]| \\
&= \sum_{i \in I_{\bar{x}}} \left| \left\{ j \in \mathbb{N}_0 \mid \underline{a} \leq \frac{d_j}{v_i} \leq \bar{a} \right\} \right| \\
&= \sum_{i \in I_{\bar{x}}} \left| \left\{ j \in \mathbb{N}_0 \mid v_i \cdot \underline{a} \leq d_j \leq v_i \cdot \bar{a} \right\} \right| \\
&= \sum_{i \in I_{\bar{x}}} \left| \left\{ j \in \mathbb{N}_0 \mid \delta^{-1}(v_i \cdot \underline{a}) \leq j \leq \delta^{-1}(v_i \cdot \bar{a}) \right\} \right| \\
&\leq \sum_{i \in I_{\bar{x}}} \left(\delta^{-1}(v_i \cdot \bar{a}) - \delta^{-1}(v_i \cdot \underline{a}) + 1 \right) \\
&\stackrel{(9)}{\leq} \sum_{i \in I_{\bar{x}}} \left(\bar{\delta}^{-1}(v_i \cdot \bar{a}) - \bar{\delta}^{-1}(v_i \cdot \underline{a}) + 1 \right) \\
&\stackrel{(9)}{\leq} \sum_{i \in I_{\bar{x}}} \left(\frac{v_i \cdot \bar{a} - \bar{\beta}}{\alpha} - \frac{v_i \cdot \underline{a} - \bar{\beta}}{\alpha} + 1 \right) \\
&= \sum_{i \in I_{\bar{x}}} \left(1 + \frac{\bar{\beta} - \beta}{\alpha} + \frac{v_i \cdot \bar{a} - v_i \cdot \underline{a}}{\alpha} \right) \\
&= \left(1 + \frac{\bar{\beta} - \beta}{\alpha} \right) \cdot |I_{\bar{x}}| + (\bar{a} - \underline{a}) \cdot \frac{V_{\bar{x}}}{\alpha}
\end{aligned}$$

C. Implementing the Algorithms

$$\begin{aligned}
&= \left(1 + \frac{\bar{\beta} - \beta}{\alpha}\right) \cdot |I_{\bar{x}}| + \frac{(\alpha + \bar{\beta} - \beta) \cdot |I_{\bar{x}}|}{V_{\bar{x}}} \cdot \frac{V_{\bar{x}}}{\alpha} \\
&= 2 \left(1 + \frac{\bar{\beta} - \beta}{\alpha}\right) \cdot |I_{\bar{x}}|.
\end{aligned}$$

C. Implementing the Algorithms

In this section, we review existing algorithms for divisor methods. In particular, we elaborate on how we have implemented them for our experiments [RW15a], and on problems we have encountered in this process.

We have taken care not to render the algorithm unnecessarily inefficient in order to perform a fair comparison of running times; the result is to the best of our abilities conditioned on a limited time budget. In particular, all of our implementations have been refined on the programming level to roughly the same degree.

For the purpose of a fair comparison, all implementation have to conform to the same interface.

Parameters: A pair $(\alpha, \beta) \in \mathbb{R}^2$ with $\alpha > 0$ and $\beta > 0$.

Input: Votes \mathbf{v} and house size k .

Output: A (symbolic) representation of all seat assignments valid w. r. t. divisor sequence $(\alpha j + \beta)_{j \geq 0}$, as well as proportionality constant a^* .

More specifically, the output is encoded as a vector of undisputed seats and a binary vector indicating which parties are tied for the remaining seats. We skip the step from a^* resp. a valid seat assignment to this representation in the pseudo code since it is elementary: all parties with “current” resp. “next” value v_i/d_{s_i-1} resp. v_i/d_{s_i} equal a^* are tied. A simple $\Theta(n)$ -time post-processing identifies these in all cases.

We have established confidence in the correctness of our implementations by extensive random testing [RW15a, `TestMain.java`]; every implementation has been run on thousands of instances. The correctness of the results has been confirmed, besides rudimentary sanity checks such as matching vector dimensions, by checking Pukelsheim’s *Max-Min Inequality* [Puk14, Theorem 4.5].

All implementations share the same numerical weakness, though: using fixed-precision arithmetics, two computations that should lead to the same result (say, a^*) yield different numbers. We compensate for that by using fuzzy comparisons: we identify numbers if they are within some constant ϵ of each other. Thus, we can reliably identify tied parties, for instance.

There is a drawback, though: if distinct values v_i/d_j are closer than ϵ (or, even without the adaption, the resolution of the chosen fixed-precision number representation), we may identify them and thus compute wrong seat assignments.

C. Implementing the Algorithms

This issue can not be circumvented on the algorithmic level. The only robust resort we know of is using arbitrary-precision arithmetics, inevitably slowing down all the algorithms.

C.1. Iterative Divisor Method

Implementing ITERATIVEMETHOD is straight-forward. An implementation using a priority queue implementation from the standard library runs in time $\Theta(n + k \log n)$. Since we expect overhead for the queue to be significant for small n , we also implement a variant which determines I using a simple linear scan, resulting in a total running time in $\Theta(kn)$.

Shared code aside, ITERATIVEMETHOD takes about 50 resp. 65 lines of code with resp. without priority queues.

C.2. Jump-and-Step

The *jump-and-step* algorithm [Puk14, Section 4.6] can be formulated using our notation as follows:

Algorithm 3: JUMPANDSTEP $_d(\mathbf{v}, k)$:

Step 1 Compute an estimator a for a^* .

Step 2 Initialize $s_i = \lfloor \delta^{-1}(v_i \cdot a) \rfloor + 1$.

Step 3 Iterate similarly to ITERATIVEMETHOD until $\sum s_i = k$ with

$$I = \begin{cases} \arg \max_{i=1}^n v_i/d_{s_i}, & \sum s_i < k; \\ \arg \min_{i=1}^n v_i/d_{s_i-1}, & \sum s_i > k. \end{cases}$$

The performance of this algorithm clearly depends on $\Delta_a := \sum s_i - k$ after Step 2; the running time is in $\Theta(n + |\Delta_a| \cdot \log n)$ when using priority queues for Step 3 (which may *not* be advisable in practice if $|\Delta_a|$ can be expected to be very small). As such, the running time is not per se bounded in n and k .

We follow the recommendations of Pukelsheim and use the estimator [Puk14, Section 6.1]

$$a := \frac{\alpha}{V} \cdot \begin{cases} k + n \cdot (\beta/\alpha - 1/2), & 0 \leq \beta/\alpha \leq 1; \\ k + n \cdot \lfloor \beta/\alpha \rfloor, & \text{else.} \end{cases}$$

The first case corresponds to Pukelsheim's *recommended* estimator for *stationary* signpost sequences, the second to his *good universal* estimator generalized to divisor sequences that are not signpost sequences in the strict sense. The additional factor α rescales the value appropriately; Pukelsheim only considers $\alpha = 1$.

C. Implementing the Algorithms

Given that these estimators guarantee $|\Delta_a| \leq n$ in the worst case, we can assume that JUMPANDSTEP runs in time $O(n \log n)$. Furthermore, Pukelsheim claims that the recommended estimator is good in practice in the sense that $|\Delta_a| \in O(1)$ in expectation, so JUMPANDSTEP may be efficient in practice for large n as well. Since their proof is limited to uniformly distributed votes and $k \rightarrow \infty$, we investigate this in Section 4.

Shared code aside, JUMPANDSTEP takes about 120 lines of code, with or without priority queues.

C.3. The Algorithm of Cheng and Eppstein

Cheng and Eppstein [CE14] do not give pseudocode for the main procedure of their algorithm which would combine the individual steps to compute $\mathcal{A}_{(k)}$. For the reader's convenience and for clarity concerning our running-time comparisons we give this top-level procedure as we have inferred it.

Algorithm 4: CHENGEPSTEINSELECT $_d(\mathbf{v}, k)$:

Step 1 Compute a suitable finite representation of \mathcal{A} .

Step 2 $\mathcal{C} := \text{FINDCONTRIBUTINGSEQUENCES}(\mathcal{A}, k)$.

Step 3 $\xi := s^{-1}(k)$ [CE14, (3)].

Step 4 If $r(\xi, \mathcal{A}) \geq k$ then
 $\xi := \text{LOWERRANKCOARSE SOLUTION}(\mathcal{A}, k, \xi)$.

Step 5 Return COARSETOEXACT (\mathcal{A}, k, ξ) .

The subroutines are given in sufficient detail in their Algorithms 1 to 3, respectively. The pseudo code given uses some high-level set operations which we did not implement naively due to performance concerns; we compute several steps during a single iteration over the respective sets of sequences.

Note that we have (hopefully) fixed an off-by-one mistake in the text. The definition of rank $r(x, A)$ is, “the number of elements of A less than or equal to x ”; that is, the rank of $A(j)$ is $j + 1$ since A is zero-based (the first element is $A(0)$). However, the authors continue to say that $r(x, A)$ “is the index j such that $A(j) \leq x < A(j + 1)$.”

Regarding performance, Cheng and Eppstein show that their algorithm runs in time $\Theta(n)$ in the worst case. Since CHENGEPSTEINSELECT computes a linear number of medians and requires a linear number of evaluations of rank function $r(x, \mathcal{A})$ (with geometrically shrinking $|\mathcal{A}|$ – otherwise the algorithm would not run in linear time), it is unclear whether the algorithm is efficient in practice.

Shared code aside, CHENGEPSTEINSELECT take about 300 lines of code. By this measure, it is the most complex of the algorithms we consider.

Additional Issues with Numerics

In addition to the concerns expressed above, there are additional numerical issues when implementing CHENGEPSTEINSELECT using fixed-precision floating-point arithmetics. In short, we have to compute certain floors and ceilings of real numbers *exactly* or we may compute a *wrong result*.

More specifically, we evaluate $r(x, \mathcal{A})$ several times by computing terms of the form $\lfloor \delta^{-1}(_) \rfloor$ (cf. Lemma 4). The problem is that the result of $\delta^{-1}(_)$ is non-integral in general, but *is* integral when the argument evaluates exactly to a d_j . With the usual floating-point arithmetic the result might be slightly smaller, though. We then erroneously round down to the next smaller integer – a critical error!

In practice, we can add a small constant to the mantissa before taking the floor. This constant has to be chosen large enough to cover potential rounding errors, but also *small enough* so as to not change subsequent calculations; CHENGEPSTEINSELECT may compute a wrong answer otherwise. This is a very delicate requirement we do not know how to fulfill in general.

C.4. SandwichSelect

We already discuss our algorithm at length in Section 3. Since we want to investigate *practical* performance, we implement rank-selection using average-case efficient Quickselect as opposed to using a linear-time algorithm with large constant factors.

We want to emphasize that our final algorithm SANDWICHSELECT is conceptually simple in the sense that there is little hidden complexity. We need exactly one call to a rank selection algorithm on a linear-size list which takes five additional linear-time operations to come up with: finding the maximal value $v^{(1)}$, constructing index set $I_{\bar{x}}$, computing $V_{\bar{x}}$, constructing multiset $\hat{\mathcal{A}}$ and computing \hat{k} . These are all quite elementary tasks in that they use one `for`-loop each which run for at most n iterations with only few operations in each. We therefore think that we can outperform CHENGEPSTEINSELECT in practice, and should not be far behind JUMPANDSTEP, either.

Regarding implementation, the delicate part was to get the bounds on j (cf. Step 6.1) right. We use floor and ceiling functions on real numbers, so rounding errors that occur in fixed-precision floating-point arithmetic can cause harm. We can circumvent this by adding (subtracting) a conservatively large constant to the mantissa of the floats before taking floors (ceilings). If this constant is larger than necessary for covering rounding errors, we might add slightly more candidates to $\hat{\mathcal{A}}$ (at most two per party) which would slightly degrade performance. Correctness, however, is *not* affected (in contrast to CHENGEPSTEINSELECT).

We also remark here that the code [RW15a] for the experimental results discussed in Section 4 is based on an earlier version of Lemma 2 with slightly weaker bounds (cf.

D. Experimental Setup

Appendix G). Experiments with the updated code are to follow, and might yield slight improvements for SANDWICHSELECT.

Shared code aside, SANDWICHSELECT takes about 100 lines of code. By this measure, it is the least complex of the non-trivial algorithms we consider.

D. Experimental Setup

We have run the experiments with Java 7 on Ubuntu 14.04 LTS running kernel 3.13.0-34-generic x86_64 GNU/Linux. The hardware platform is a ThinkPad T430s Tablet with the following core parameters according to `lshw`.

CPU: Intel[®] Core[™] i5-3320M CPU @ 2.60GHz

Cache: L1 32KiB, L2 256KiB, L3 3MiB

RAM: 4+4GiB SODIMM DDR3 Synchronous 1600 MHz (0.6 ns)

As our code is written in Java, we include a warm-up phase to trigger just-in-time compilation of the relevant methods. All times are measured using the built-in method `System.nanoTime()`. We use the same set of inputs for all algorithms, all of which have to construct the full set $\mathcal{S}(\mathbf{v}, k)$ for each input (\mathbf{v}, k) during the measurement.

In order to increase accuracy, we repeat the execution of each algorithm on each input several times and measure the total time; we then report the average time per execution.

For the selection-based algorithms, we use the randomized Quickselect-based implementation by Sedgewick and Wayne [SW11] as published on the book website. We use the (pseudo) random number generators for several distributions from the same library (download of `stdlib-package.jar` on August 11th, 2015).

For reproducing our running time experiments, make sure you have working GNU/Linux² installation with Ruby, Java 7 and Ant; then execute

```
ruby run_experiments.rb arxiv.experiment
```

for the data represented in Section 4 and Appendix E. Be warned: this may run for a long time, and it *will* create lots of images (provided you have `gnuplot` installed).

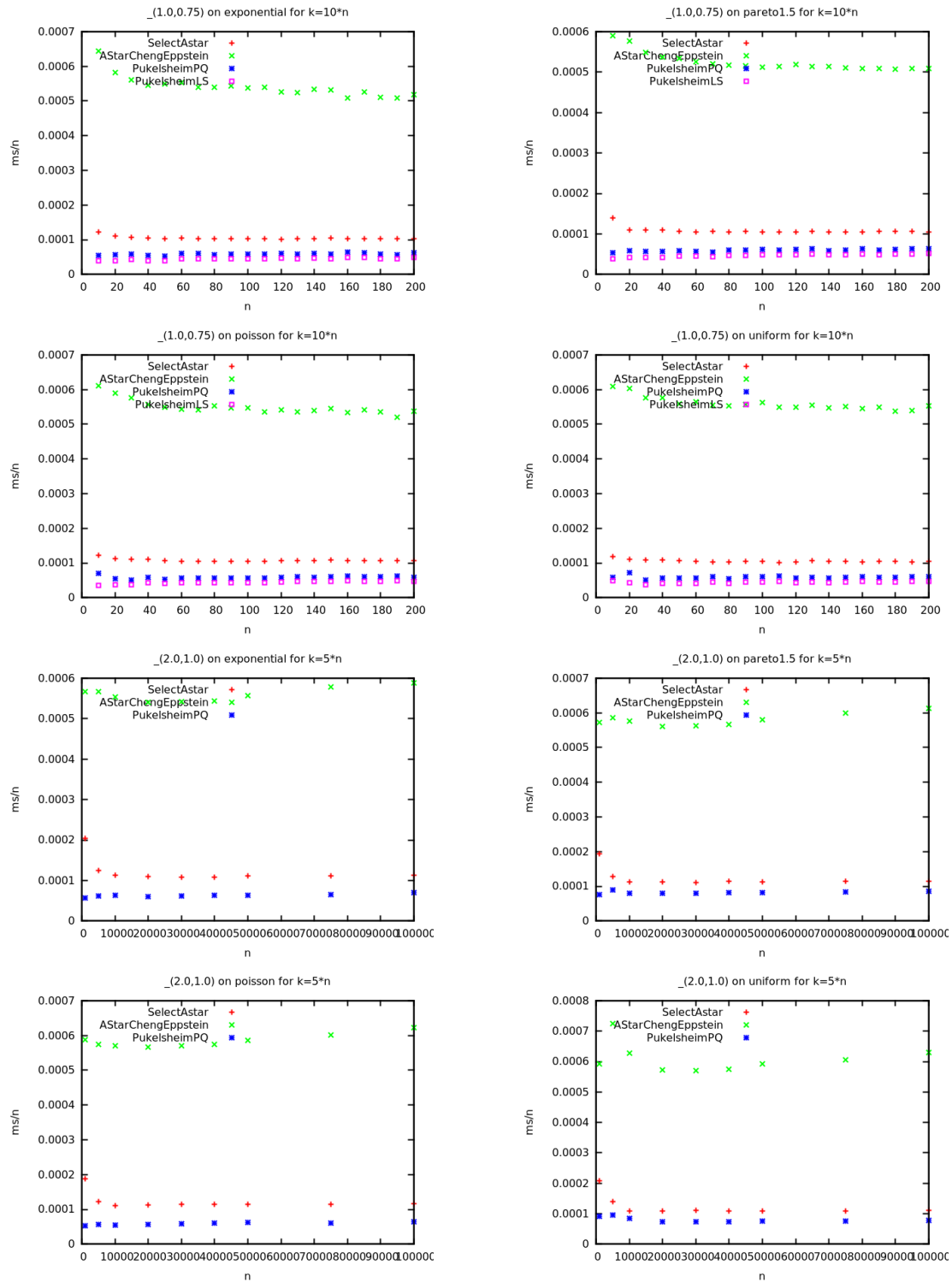
E. More Running-Time Experiments

We apologize to only offer draft graphics without commentary for the time being.

²Our framework *may* work on other platforms, maybe with small adjustments to the Ruby code, but we have not tried. See `README.md` for a workaround.

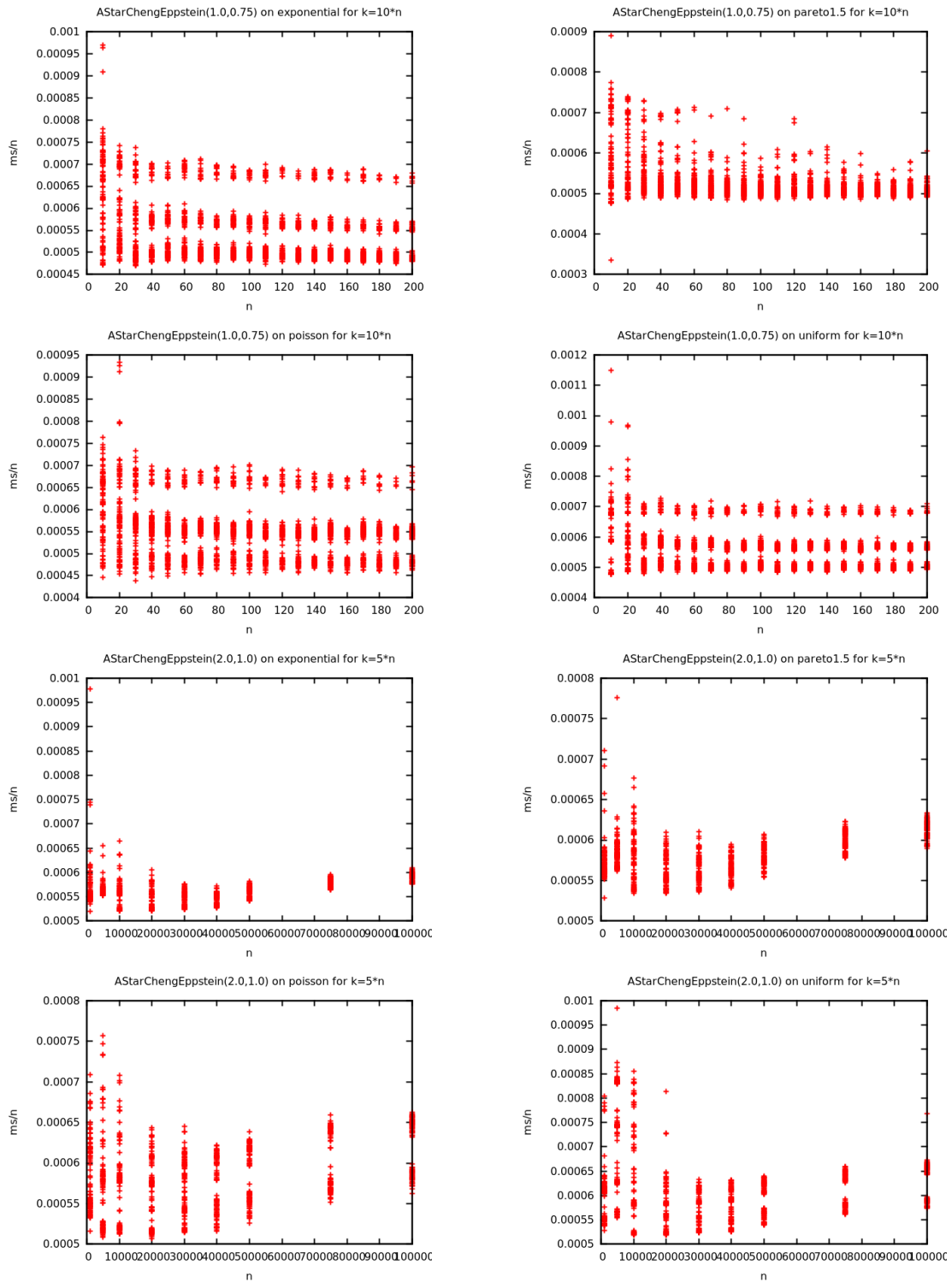
E. More Running-Time Experiments

Average normalized runtimes for several input distributions and across several orders of magnitudes of n .



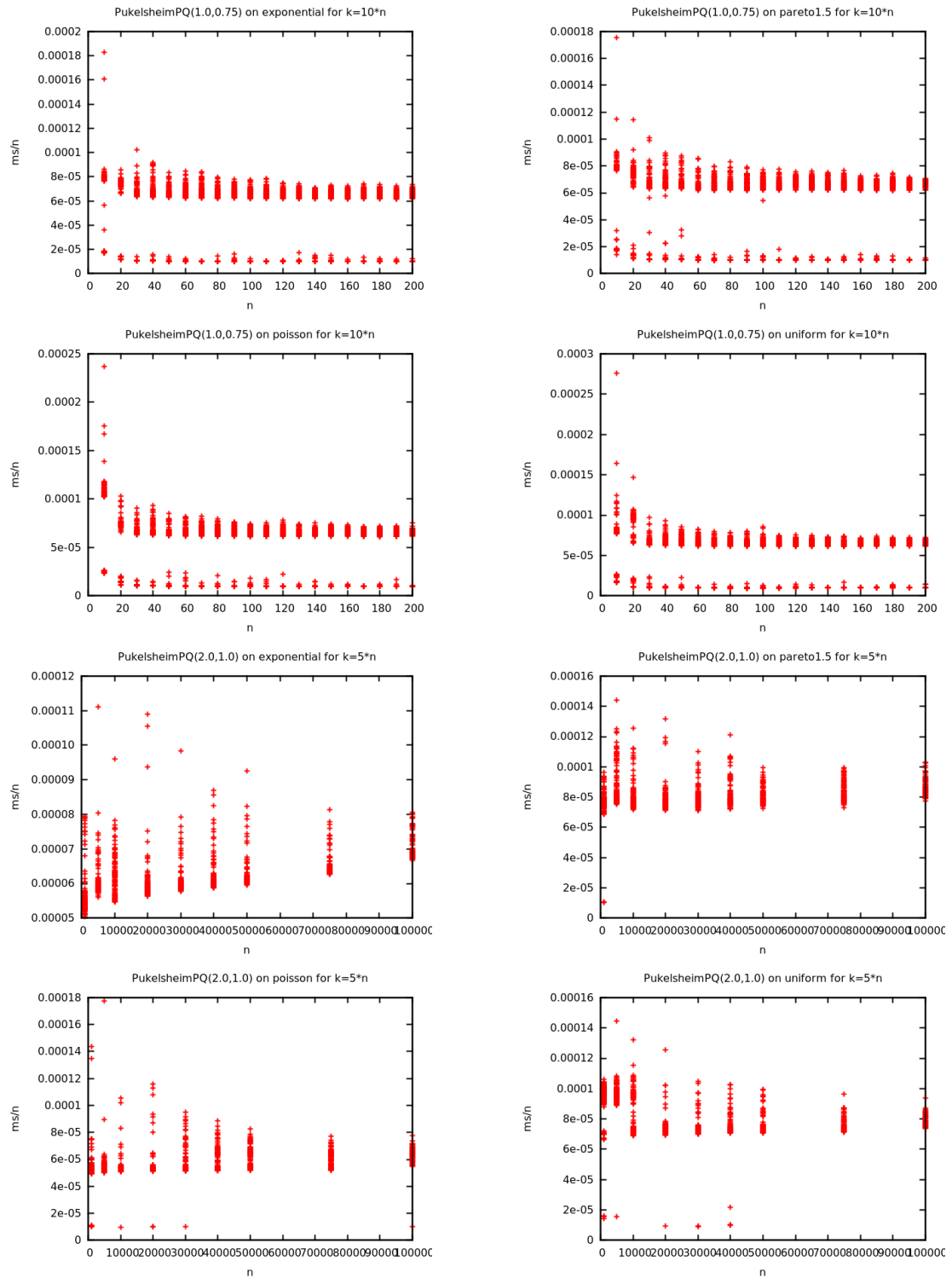
E. More Running-Time Experiments

Normalized runtimes of CHENGEPSTEINSELECT for several input distributions and across several orders of magnitudes of n .



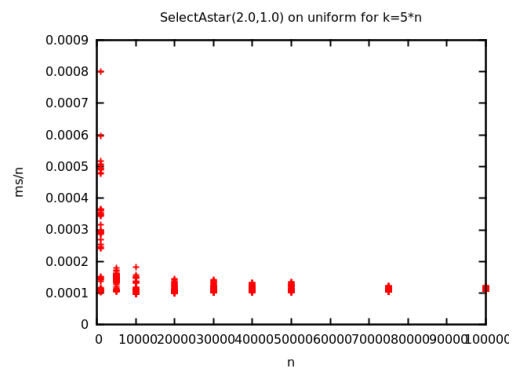
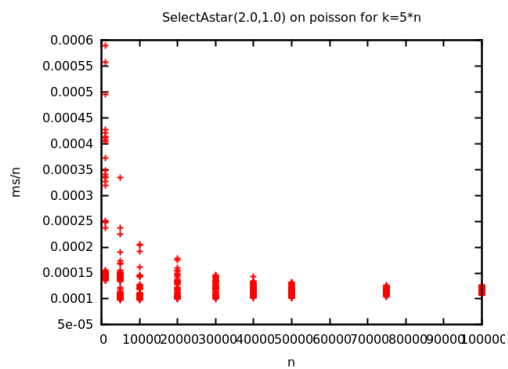
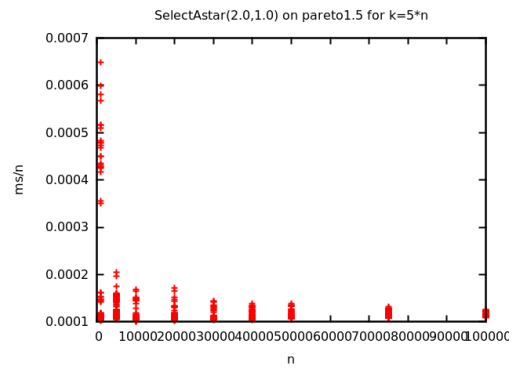
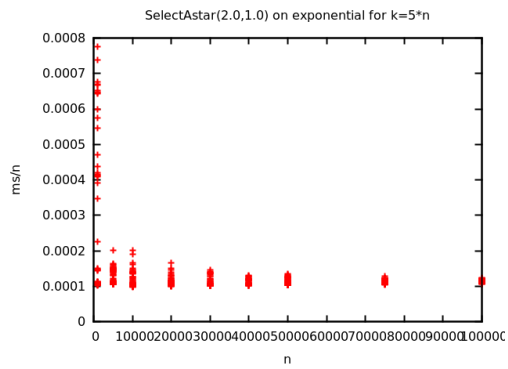
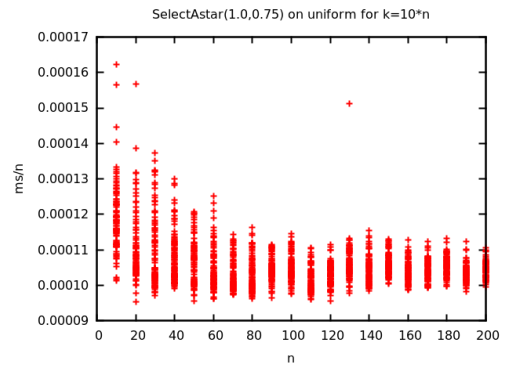
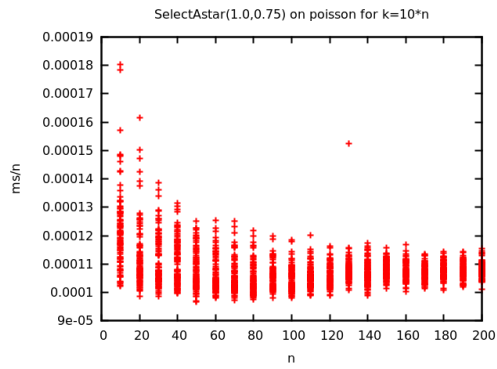
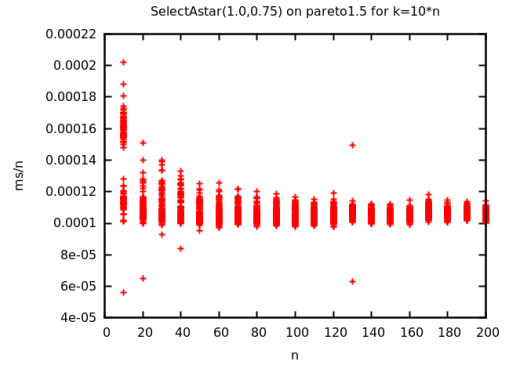
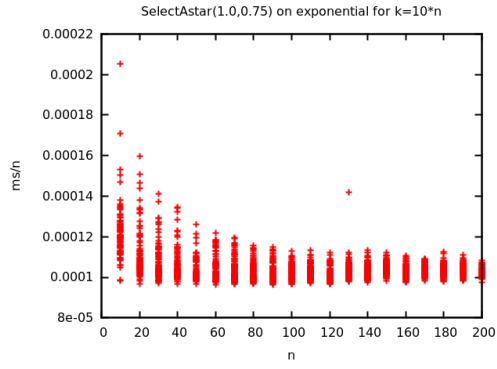
E. More Running-Time Experiments

Normalized runtimes of JUMPANDSTEP for several input distributions and across several orders of magnitudes of n .



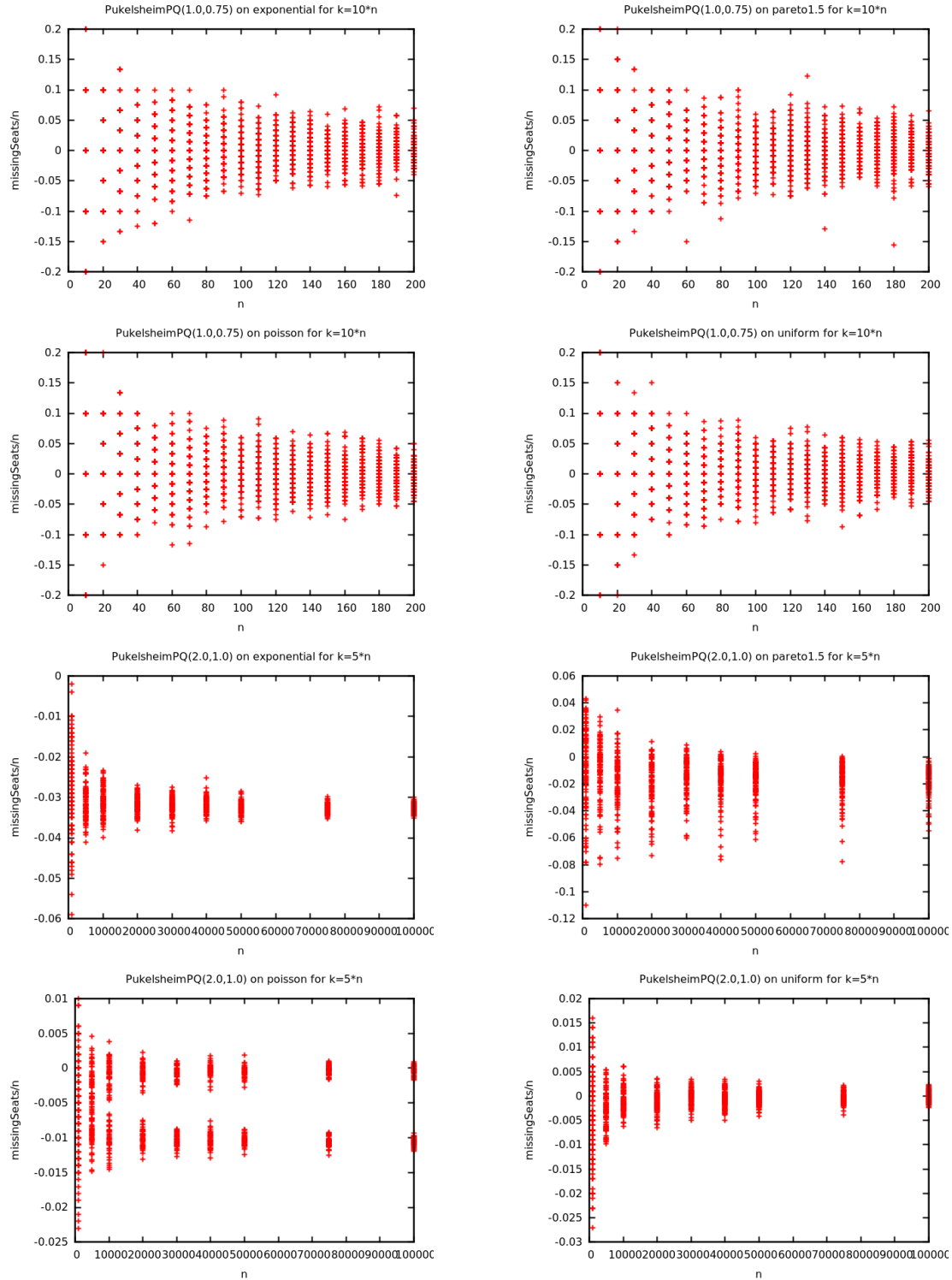
E. More Running-Time Experiments

Normalized runtimes of SANDWICHSELECT for several input distributions and across several orders of magnitudes of n .



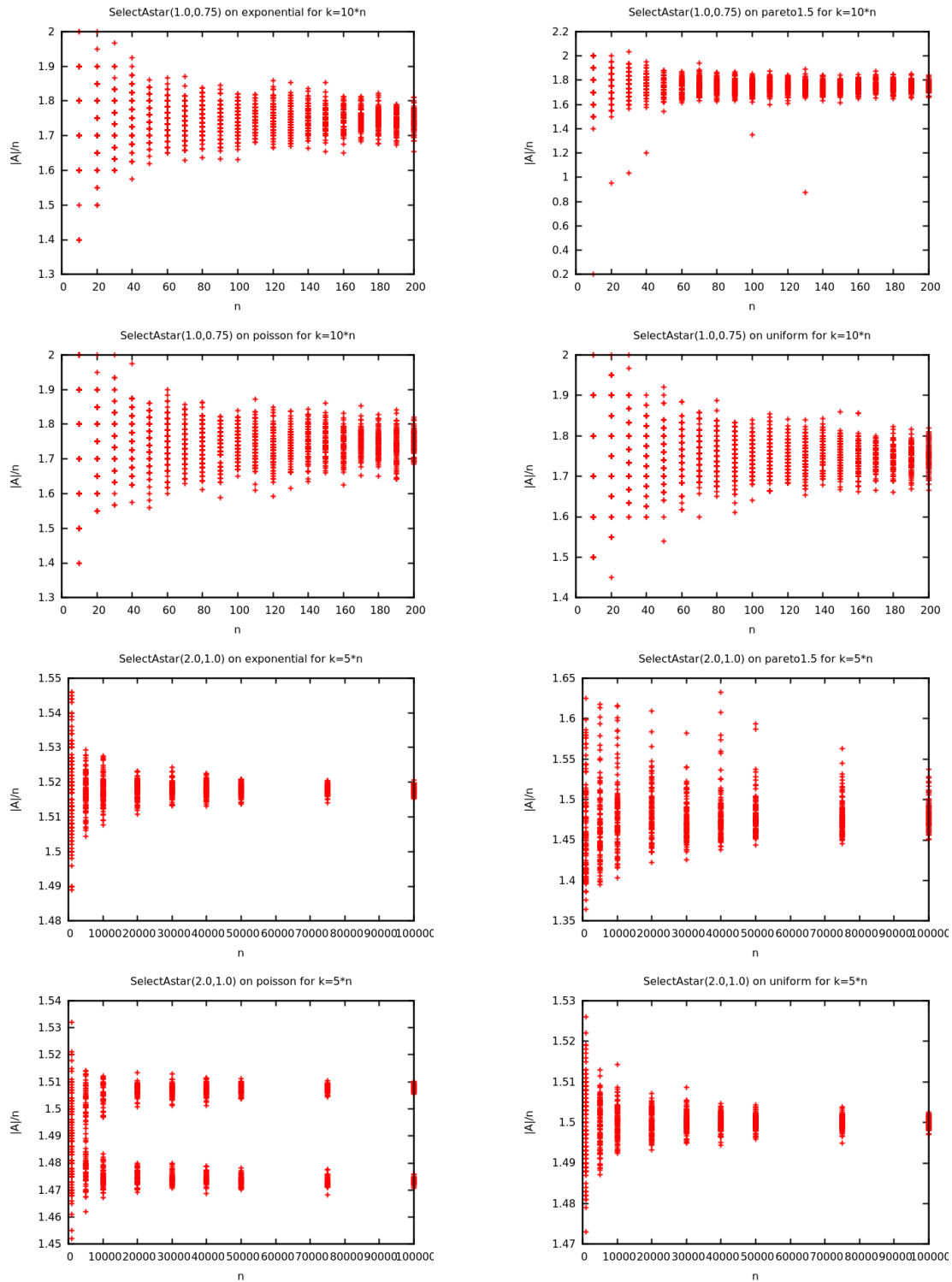
E. More Running-Time Experiments

Normalized Δ_a of JUMPANDSTEP for several input distributions and across several orders of magnitudes of n .



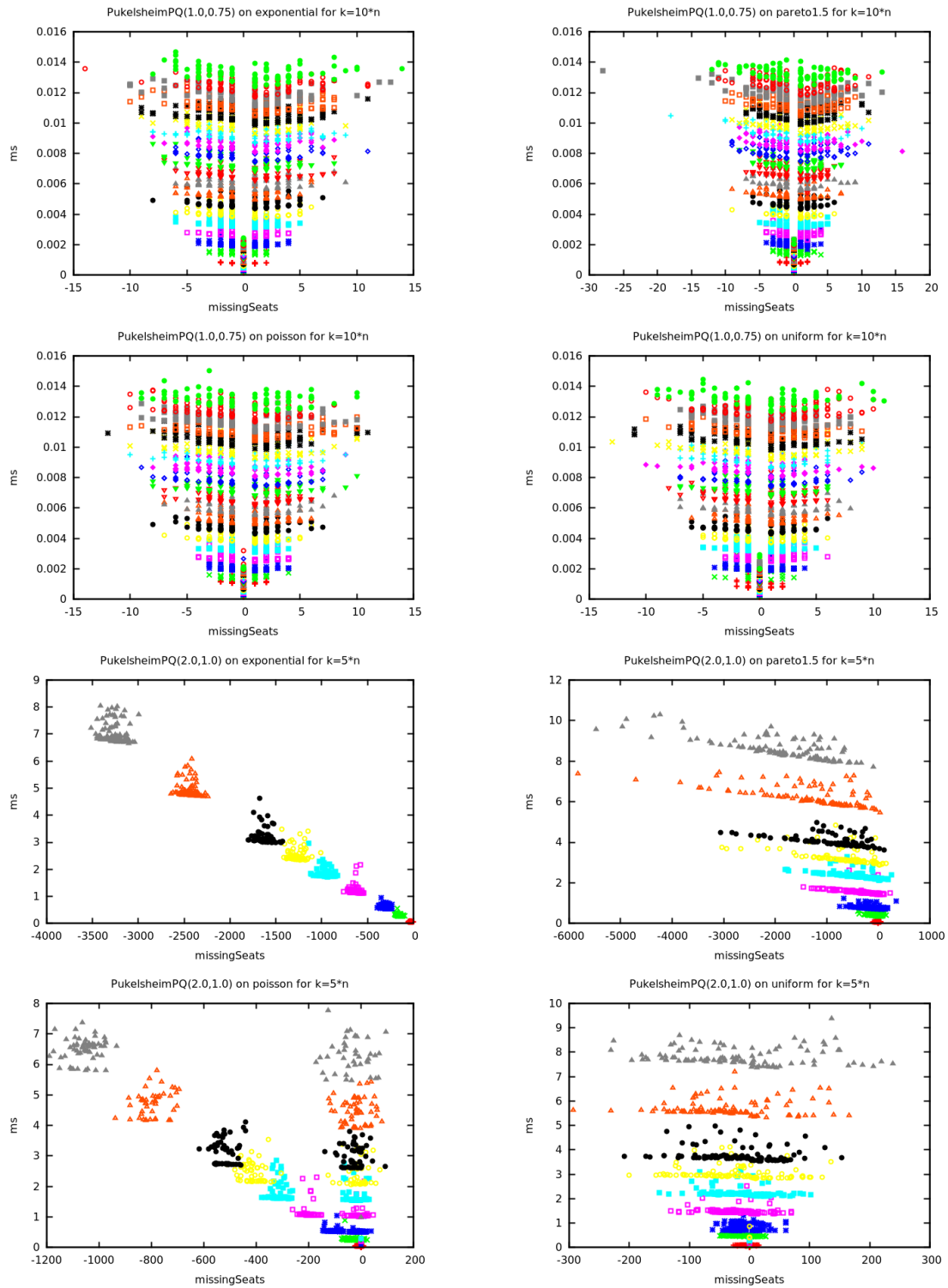
E. More Running-Time Experiments

Normalized $|\hat{\mathcal{A}}|$ of SANDWICHSELECT for several input distributions and across several orders of magnitudes of n .



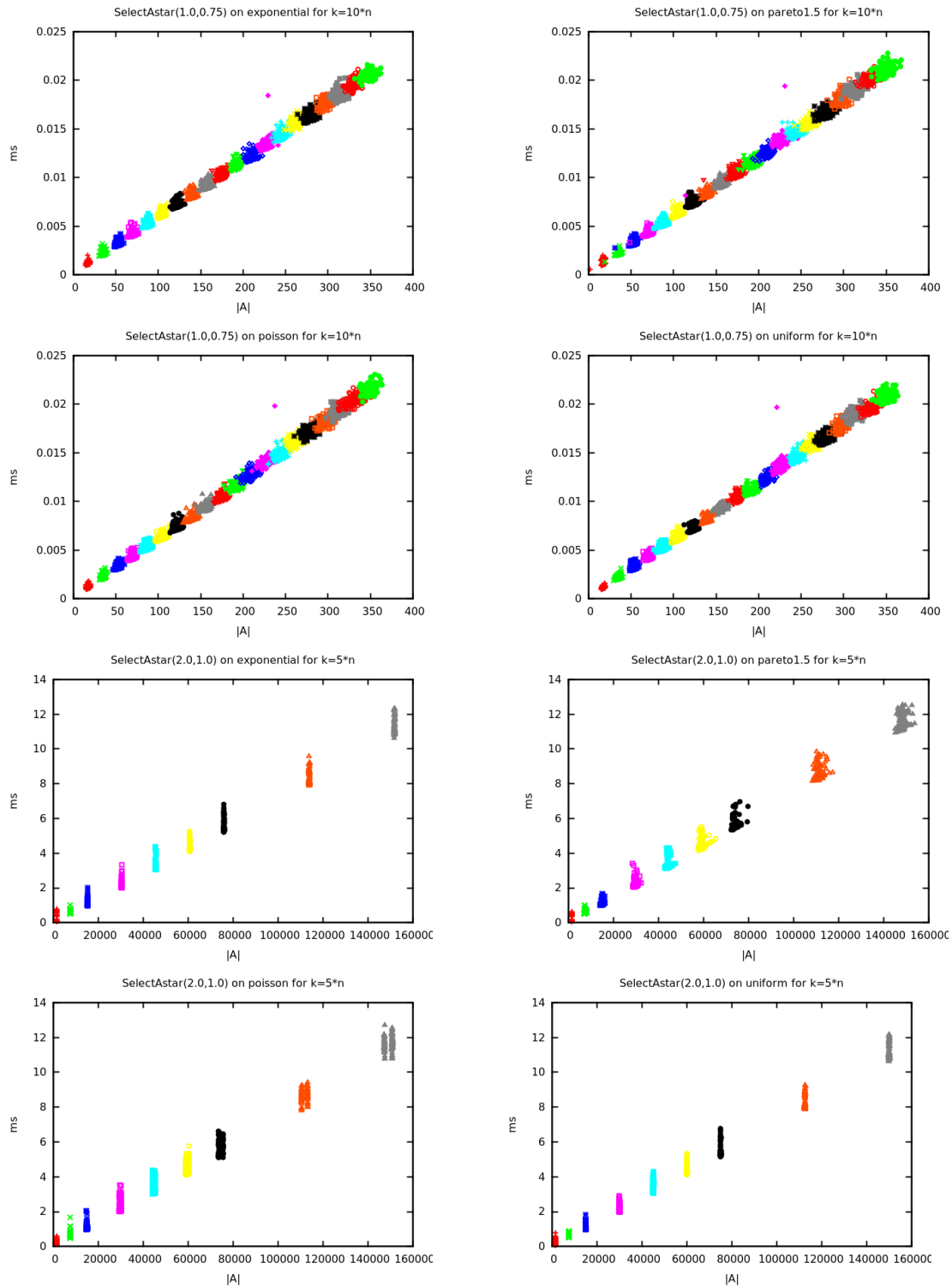
E. More Running-Time Experiments

Runtimes against Δ_a of JUMPANDSTEP for several input distributions and across several orders of magnitudes of n . Each color stands for one n .



E. More Running-Time Experiments

Runtimes against $|\hat{\mathcal{A}}|$ of SANDWICHSELECT for several input distributions and across several orders of magnitudes of n . Each color stands for one n .



F. Index of Used Notation

In this section, we collect the notation used in this paper. Some might be seen as “standard”, but we think including them here hurts less than a potential misunderstanding caused by omitting them.

Generic Mathematical Notation

- $\lfloor x \rfloor, \lceil x \rceil$ floor and ceiling functions, as used in [GKP94].
- $M_{(k)}$ The k th smallest element of (multi)set/vector M (assuming it exists); if the elements of M can be written in non-decreasing order, M is given by $M_{(1)} \leq M_{(2)} \leq M_{(3)} \leq \dots$.
Example: For $M = \{5, 8, 8, 8, 10, 10\}$, we have $M_{(1)} = 5$,
 $M_{(2)} = M_{(3)} = M_{(4)} = 8$, and $M_{(5)} = M_{(6)} = 10$.
- $M^{(k)}$ Similar to $M_{(k)}$, but $M^{(k)}$ denotes the k th *largest* element.
- $\mathbf{x} = (x_1, \dots, x_d)$. . . to emphasize that \mathbf{x} is a vector, it is written in **bold**; components of the vector are written in regular type.
- \mathcal{M} to emphasize that \mathcal{M} is a multiset, it is written in calligraphic type.
- $\mathcal{M}_1 \uplus \mathcal{M}_2$ multiset union; multiplicities add up.

Notation Specific to the Problem

- party, seat, vote (count), chamber size
 Parties are assigned seats (in parliament), so that the number of seats s_i that party i is assigned is (roughly) proportional to that party’s vote count v_i and the overall number of assigned seats equals the chamber size k .
- $d = (d_j)_{j=0}^\infty$ the divisor sequence used in the highest averages method; d must be a nonnegative, (strictly) increasing and unbounded sequence.
- δ, δ^{-1} a continuation of $j \mapsto d_j$ on the reals and its inverse, both of which can be evaluated in constant time.
- n number of parties in the input.
- \mathbf{v}, v_i $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{Q}_{>0}^n$, vote counts of the parties in the input.
- V the sum $v_1 + \dots + v_n$ of all vote counts.
- k $k \in \mathbb{N}$, the number of seats to be assigned; also called house size.
- \mathbf{s}, s_i $\mathbf{s} = (s_1, \dots, s_n) \in \mathbb{N}_0$, the number of seats assigned to the respective parties; the result.
- $a_{i,j}$ $a_{i,j} := d_j/v_i$, the ratio used to define divisor methods; i is the party, j is the number of seats i has already been assigned.

G. Changelog

A_i	For party i , $A_i := \{a_{i,0}, a_{i,1}, a_{i,2}, \dots\}$ is the list of (reciprocals of) party i 's ratios.
a	We use a as a free variable when an arbitrary $a_{i,j}$ is meant.
\mathcal{A}	$\mathcal{A} := A_1 \uplus \dots \uplus A_n$ is the multiset of all averages.
$r(x, \mathcal{A})$	the rank of x in \mathcal{A} , that is the number of elements in multiset \mathcal{A} that are no larger than x ; $r(x)$ for short if \mathcal{A} is clear from context.
a^*	the ratio $a^* = a_{i^*,j^*}$ selected for assigning the last (i. e. the k th) seat; corresponds to \mathbf{s} by $s_i = r(a^*, A_i)$; $a^* = \mathcal{A}_{(k)}$ (cf. Section 2 and Section 3).
\bar{x}	an upper bound $\bar{x} > a^*$; we use $\bar{x} = d_{k-1}/v_1 + \varepsilon$, where $\varepsilon > 0$ is a suitable constant.
$I_{\bar{x}}$	$I_{\bar{x}} := \{i \mid v_i > d_0/\bar{x}\}$; the set of parties i whose vote count is large enough, so that $a_{i,0} < \bar{x}$, i. e. so that they contribute to the rank of \bar{x} in \mathcal{A} .
$V_{\bar{x}}$	the sum of the vote counts of all parties in $I_{\bar{x}}$.
$\mathcal{A}^{\bar{x}}$	the elements in \mathcal{A} that are smaller than \bar{x} , i. e., $\mathcal{A} \cap (-\infty, \bar{x})$.
\underline{a}, \bar{a}	lower and upper bounds on candidates $\underline{a} \leq a \leq \bar{a}$ such that still $a^* \in \mathcal{A} \cap [\underline{a}, \bar{a}]$.

G. Changelog

The following (substantial) changes have been made from arXiv version 2 to 3.

- Lemma 2 has been strengthened; both \underline{a} and the upper bound on $|\mathcal{A} \cap [\underline{a}, \bar{a}]|$ have been improved. Both changes are due to the observation that we could require $\underline{\beta} \leq \alpha$ without loss of generality.

Related notation update: $(\check{\beta}, \beta) \rightsquigarrow (\underline{\beta}, \bar{\beta})$.

- We have added Appendix A in order to clarify that the assumptions we make for our main result do restrict the scope of divisor methods we cover by too much.