The Easiest Hard Problem: Now Even Easier

Ruben Horn
Helmut Schmidt University
Hamburg, Germany
r.horn@hsu-hh.de

Sarah L. Thomson
Edinburgh Napier University
Edinburgh, United Kingdom
s.thomson4@napier.ac.uk

Daan van den Berg
Vrije Universiteit Amsterdam
University of Amsterdam
Amsterdam, Netherlands
daan@yamasan.nl

Pieter Adriaans
Institute for Logic, Language, and Computation
University of Amsterdam
Amsterdam, Netherlands
pieter@pieter-adriaans.com

ABSTRACT

We present an exponential decay function that characterizes the number of solutions to instances of the Number Partitioning Problem (NPP) with uniform distribution of bits across the integers. This function is fitted on the number of optimal solutions of random instances with lengths between 10 and 20 integers and may be used as a heuristic either directly by new algorithms for the NPP or as a benchmark to evaluate how well different Evolutionary Algorithms (EAs) cover the search space. Despite the long history of the NPP, it seems such a characterization does not yet exist.

CCS CONCEPTS

• **Mathematics of computing → Combinatorial optimization**.

KEYWORDS

The Easiest Hard Problem, Number Partitioning Problem, Combinatorial Optimization

ACM Reference Format:
Ruben Horn, Sarah L. Thomson, Daan van den Berg, and Pieter Adriaans. 2024. The Easiest Hard Problem: Now Even Easier. In *Proceedings of The Genetic and Evolutionary Computing Conference (GECCO '24).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The Number Partitioning Problem (NPP) is a well known $\mathcal{NP}$-hard problem [2, Chapter 5]. It is a special case of the also $\mathcal{NP}$-hard Subset Sum Problem (SSP), where the aim is to find a subset A of a set of positive integers S $t \in [0..\sum S]$ so that $|t - \sum A|$ is minimized. In the (two-way) NPP $t = \lceil 1/2 \sum S \rceil$ so the resulting subset have equal sums. Like other combinatorial optimization problems (like the travelling salesperson problem [2, Chapter 5] or protein-folding [1]), there can be multiple optimal solutions to a single instance. Hence, the frequency of optimal solutions may directly correlate with the computational hardness of an instance, as the probability of a search algorithm to finding a solution increases with their frequency. The peculiar property of the NPP is that it is often significantly easier than other problems in the class of $\mathcal{NP}$-hard problem. If $\sum A = \lceil 1/2 \sum S \rceil$, the algorithm can terminate immediately, because no better partition can exist. This property, as well as its *easy-hard-easy* [8], have resulted in the NPP being nicknamed "the easiest hard problem" [4, 6, 8].

As shown in the experiments by Richard Korf [5], the discrepancy of the optimal solutions $|\sum A - \lceil 1/2 \sum S \rceil|$ for NPP instances decreases as the number of integers n grows with a fixed range of values limited to m binary digits. Likewise, the number of solutions that are evaluated by a search algorithm peaks and slowly declines as perfect solutions are found (earlier). This was later also replicated by Stephan Mertens [6]. These observations are not surprising, because the number of possible partitions of n integers with m bits each $2^n/2$ grows much faster than the maximum sum of the set $n(2^m - 1)$ over n with constant m. Therefore, at some point, a perfect partition must exist and the likelihood of any search algorithm finding it increases as the number of these solutions grows even further. The relation of m/n to the frequency of optimal solutions (and their optimality) was illustrated by Brian Hayes [4] and recently replicated in [7]. To the best of our knowledge, there has not been any other attempt yet to characterize the frequency, existence, or distribution of optimal solutions to instances of the NPP.

2 CHARACTERIZATION

First, we generate a dataset by sampling instances with random integers with m bits each for $n \in [10..20]$. For each, n we generate five different instances for all possible values of the ratio $m/n \in (0, 1.5]$ with integers in $[2^{m-1}..2^m - 1]$. We enumerate all possible values of m/n by adding a single bit at the integer index that has the fewest bits. Thus, some instances do not have a perfectly uniform bit distribution. The generated instances are then solved using an exact branch and bound algorithm [9] and the number of optimal solutions counted. Since there are twice as many solutions for sets with odd sums than for equivalent sets with even sums [4], those with odd sums are discarded, leaving 8895 instances over all n. The frequency of optimal solutions for the resulting instances with even sums is visualized in Fig. 1 for $n = 10$ and $n = 20$, using a logarithmic vertical scale. Since they both appear to follow an exponential decay with a *knee* between around 0.8 to 1.0, we fit the Eq. (1) with the initial offset parameter N_0 and rate of decay $-\lambda$ over m/n. A constant offset of +1 is added, since there is always at least one optimal solution to

GECCO '24, July 14–18, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

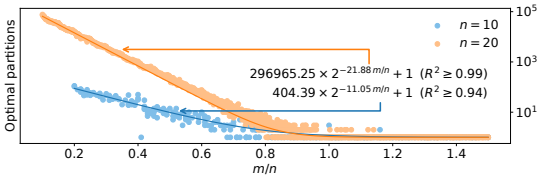

Figure 1: Characterizations of the number of optimal (mostly perfect for $m/n < 0.8$) solutions for different n and $m/n \leq 1.5$

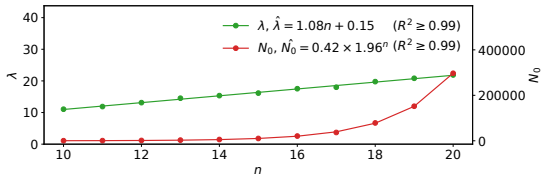

Figure 2: Fitted functions of the parameters of Eq. (1)

Table 1: Fitted values and R^2 metric for Eq. (1)

n	N_0	λ	R^2	RMSE	RMSPE
10	404.388	11.051	0.941	4.477	1.053%
11	714.169	11.887	0.977	5.088	0.549%
12	1427.166	13.128	0.982	7.971	0.614%
13	2920.215	14.526	0.978	16.487	0.379%
14	5597.876	15.347	0.990	21.939	0.329%
15	10216.711	16.143	0.994	32.009	0.434%
16	20825.808	17.533	0.992	72.013	0.299%
17	36584.847	18.002	0.995	82.399	0.387%
18	78474.330	19.770	0.992	262.836	0.294%
19	151955.578	20.829	0.992	439.552	0.264%
20	296965.253	21.877	0.993	800.628	0.262%

the NPP, even if it is not perfect.

$$N_0 \times 2^{-\lambda m/n} + 1 \tag{1}$$

The coefficient of determination R^2 is very high, with over 0.99 for lengths $n \geq 14$ in Table 1, which indicates that this model quite accurately describes the observations in our dataset. While the Root Mean Square Error (RMSE) appears at first glance to be large for increasing n, the Root Mean Square Percentage Error (RMSPE) is below 0.5% for $n \geq 13$ and the characterization in Fig. 1 is visually sound. Since the values in Table 1 increase monotonically, we attempt to generalize Eq. (1) by characterizing its parameters as functions of n. While λ increases linearly, N_0 grows exponentially. The combined model is therefore given by Eq. (2) and has an R^2 score of 0.993 and RMSE of 379.831 or RMSPE of 0.384%. Since λ is almost equal to n, we can simplify the model significantly into Eq. (3) and still achieve a reasonable R^2 score of 0.956 and RMSE of 945.625 or RMSPE of 0.718%.

$$(0.418583 \times 1.961385^n) \times 2^{-(1.081473n + 0.149822)\, m/n} + 1 \tag{2}$$

$$(0.42 \times 1.96^n) \times 2^{-m} + 1 \tag{3}$$

3 DISCUSSION

The rationale behind fitting an exponential decay function with a base of two comes from the binary exponential nature of the possible partitions: the choice of including or excluding each integer. Since the distribution of bits over the integers and the sampling of their concrete values is uniform, the histogram of all possible subset sums probably form something similar to an Irwin–Hall distribution [3]. The solution to the NPP is at the center of this distribution, and the parameter N_0 predicts the maximum height of its peak. As m and therefore $\sum S$ grows, this distribution is stretched out and flattens as described by Eq. (1) because the number of solutions stays the same for constant n. Since this may also cause *gaps* in the distribution and no partition yields a perfect partition, the algorithm cannot terminate early while the search tree cannot be pruned (much) due to the uniformity of the integers. This likely explains why Van den Berg and Adriaans [9] also find that the hardest and easiest instances both have a uniform distribution of bits over the integers. This complexity of the NPP might be an indication of *fractal* properties: A non-integer dimension and self-similarity of subset sum frequencies.

4 CONCLUSION

In this paper, we proposed a model that predicts the number of optimal solutions for the NPP reasonably well. In fact, this model can be simplified without sacrificing a lot of accuracy by eliminating one of its parameters (λ) but maintaining an R^2 score of 0.956, making the NPP *even easier*. Investigating the underlying distribution further might allow an extension to the SSP. In the realm of Evolutionary Algorithms (EAs) the NPP might be an interesting benchmark to evaluate how well/quickly EAs discover all optimal solutions which can now be predicted reasonably accurately. Instances with $m/n \leq 0.8$ appear to have perfect solutions [4, 7]. Perhaps the discrepancy for $m/n > 0.8$ can be characterized similarly as a heuristic of the optimal solution discrepancy for non-exact/metaheuristic algorithms. In the future we will further investigate the potential fractal properties of the NPP and SSP.

REFERENCES

[1] Aviezri S. Fraenkel. 1993. Complexity of protein folding. *Bulletin of Mathematical Biology* 55, 6 (Nov. 1993), 1199–1210. https://doi.org/10.1007/bf02460704

[2] Michael R. Garey and David S. Johnson. 1990. *Computers and Intractability; A Guide to the Theory of NP-Completeness.* W. H. Freeman & Co., USA.

[3] Philip Hall. 1927. The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika* 19, 3-4 (12 1927), 240–244. https://doi.org/10.1093/biomet/19.3-4.240

[4] Brian Hayes. 2002. Computing Science: The Easiest Hard Problem. *American Scientist* 90, 2 (2002), 113–117. http://www.jstor.org/stable/27857621

[5] Richard E. Korf. 1998. A complete anytime algorithm for number partitioning. *Artificial Intelligence* 106, 2 (Dec. 1998), 181–203. https://doi.org/10.1016/s0004-3702(98)00086-1

[6] Stephan Mertens. 2003. The Easiest Hard Problem: Number Partitioning. https://doi.org/10.48550/ARXIV.COND-MAT/0310317

[7] Nikita Sazhinov, Ruben Horn, Pieter Adriaans, and Daan van den Berg. 2023. The Partition Problem, and How The Distribution of Input Bits Affects the Solving Process. In *Proceedings of the 15th International Joint Conference on Computational Intelligence.* SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0012143600003595

[8] Ethan L. Schreiber, Richard E. Korf, and Michael D. Moffitt. 2018. Optimal Multi-Way Number Partitioning. *J. ACM* 65, 4, Article 24 (jul 2018), 61 pages. https://doi.org/10.1145/3184400

[9] Daan Van den Berg and Pieter Adriaans. 2021. Subset Sum and the Distribution of Information. In *Proceedings of the 13th International Joint Conference on Computational Intelligence.* 134–140. https://doi.org/10.5220/0010673200003063