Bit-Level Probabilistic Inference for Floating-Point Model Weights

Aleksanteri Sladek[1,2] **Martin Trapp**[3] **Arno Solin**[1,2]

[1]Department of Computer Science, Aalto University, 02150 Espoo, Finland
[2]ELLIS Institute Finland, 02150 Espoo, Finland
[3]Department of Computing and Learning Systems , KTH Royal Institute of Technology, 114 28 Stockholm, Sweden

Abstract

Floating-point bit-string representations of parameters are *de facto* building blocks of neural networks, and their quantisation into lower precision formats is increasingly used to reduce their computational requirements. The information loss that quantisation induces in the model is, however, not well studied, and its detection relies on heuristics. We propose an information-theoretic approach for measuring quantisation-induced information loss. By modelling the probability distribution of the underlying floating-point bit-strings that parameters consist of, we can measure information loss as the change in differential entropy across quantisation levels. We empirically demonstrate that our method accurately captures quantisation-induced information loss in neural networks trained on small benchmark regression data sets by showing a strong correlation between our information loss metric and the observed performance degradation across quantisation levels.

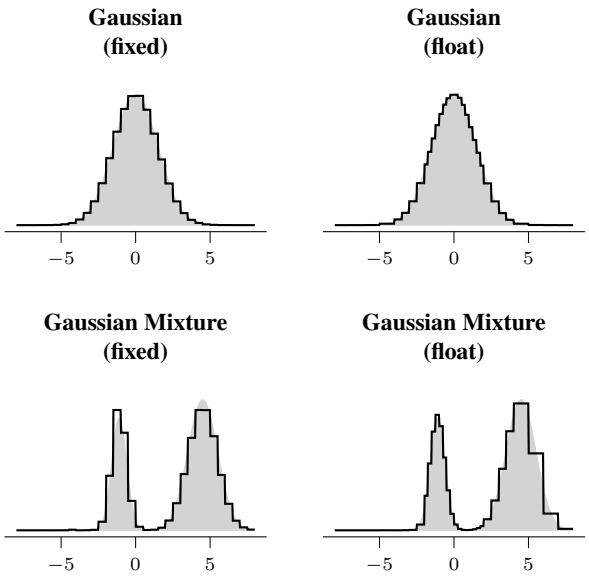


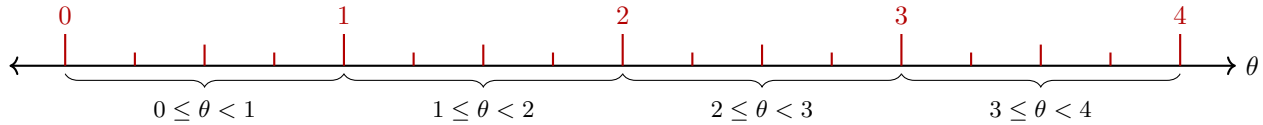
Figure 1: **BitVI with fixed- and floating-point numbers.** 5-bit approximations of two single-parameter target densities (exact grey, approximation black line). The floating-point representation captures more detail around zero and less detail further away from zero.

1 INTRODUCTION

As modern machine learning models continue to grow in parameter count, reducing memory and compute through *quantisation* has become increasingly common. Lowering the number of bits used to represent model parameters makes training and deployment substantially cheaper, and low-precision formats are now used throughout modern large-scale machine learning [Micikevicius et al., 2022, Dettmers et al., 2022, Frantar et al., 2022, Dettmers et al., 2023]. Parameter quantisation, however, is a lossy process where some information about the original parameters is lost, as it inherently involves mapping from some large discrete set of parameter values (e.g., values in float32 format) to some smaller discrete set of values (e.g., values in float8 format). Most existing work studies the effect of this loss of information through downstream performance, which is useful but indirect. Arguably, this only serves as a proxy for the information loss that occurs in model parameters from quantisation, and a more theoretically well-grounded approach would be beneficial for studying this effect.

We propose taking an information-theoretic approach for measuring the information loss which considers the posterior probability distribution of the model parameters. We hypothesize that if one could attain the posteriors of the parameters both pre- and post-quantisation, then the difference in the *entropy*, a measure of information [Shannon, 1948], of these distributions could provide a well-grounded and elegant measure for the information loss due to quantisation.

5-bit Fixed-Point Representation



5-bit Floating-Point Representation (Excl. NaN/Inf)

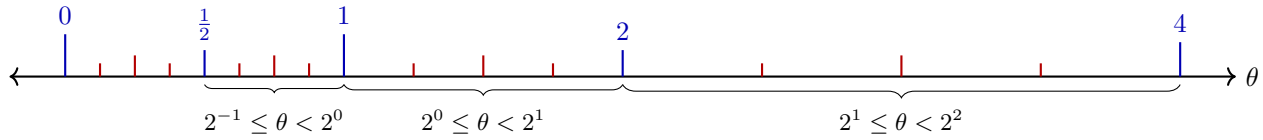


Figure 2: Comparison of how space is iteratively partitioned by 5-bit fixed-point vs. floating-point bit-string representations. For fixed-point bit-strings, the number line is partitioned into fixed-width intervals via an **arithmetic partitioning** function. For floating-point, exponent bits divide the number line via a **geometric partitioning** function, whereas mantissa bits perform arithmetic partitioning. In this example, the 5-bit fixed-point bit-string illustrated has 2 integer bits and 2 fractional bits. The floating-point bit-string has 2 exponent bits, 2 mantissa bits, a bias of 2, and skips reserving an exponent for NaN/Infinity to maximize the representable range up to the 4.0 bound.

Furthermore, we challenge the common assumption that the parameter posteriors are distributions over continuous random variables (RVs), which is particularly difficult to justify in the context of parameters quantised to very low-precisions. To accurately model posterior distributions of quantised parameters, it is necessary to recall that the underlying representations of all model parameters manipulated on a computer are finite-length bit-strings (sequences of 0s and 1s) under a chosen number system, most commonly floating-point arithmetic [IEEE, 2019, Knuth, 1997, Sterbenz, 1974]. A recently published method, BitVI [Sladek et al., 2025], builds directly on this observation and proposes modelling the posterior distribution of parameters over their underlying bit-string representations rather than over idealized continuous values. The motivation is simple: if learning and inference are ultimately carried out on discrete representations anyway, why not perform approximate inference in that same space?

Taking this view to quantised model weights (see Fig. 1), we propose an information-theoretic measure of *quantisation-induced information loss*, which leverages the tractability and hierarchical structure of the bit-string posterior approximation provided by BitVI. This allows for computing the differential entropy of the approximate parameter posterior at different bit-widths exactly and efficiently. The information loss from the quantisation process is then elegantly quantified as the change in this entropy across the quantisation levels. A central contribution of this work is the extension of BitVI to floating-point bit-string representations [IEEE, 2019], which are the standard representation format for modern machine learning systems. We also consider our method in the context of neural networks in regression tasks, which were not studied in the original work.

2 METHOD

Floating-point representations Let $\theta \in \mathbb{R}$ represent a continuous parameter of a neural network f_θ, and let $\mathbf{b} \in \mathbb{B}^B$ be the B-bit *floating-point bit-string* $\mathbf{b} = \{b_{B-1}, b_{B-2}, \ldots, b_0\}, b_i \in \{0, 1\}$ that represents θ. An approximation of θ from $\mathbf{b}$ is given by the mapping function $f_{\text{float}} : \mathbb{B} \to \mathbb{R}$:

$$\theta = \underbrace{(-1)^{b_{B-1}}}_{\text{sign}} \cdot \underbrace{(2^{(\sum_{i=0}^{E} b_{i+1} 2^i - \kappa)})}_{\text{exponent}} \cdot \underbrace{(1 + 2^{\sum_{j=1}^{M} b_{(E+j)} 2^{-j}})}_{\text{mantissa}},$$

where E and M represent the number of bits allocated for the exponent and mantissa, respectively, and κ is a bias term. This mapping function is the standard IEEE 754 floating-point representation (excluding special tokens) [IEEE, 2019], which is widely used in machine learning frameworks to represent model parameters.

Bit-strings as partitionings While Sladek et al. [2025] focused on *fixed-point* bit-string representations due to their simpler structure, we observe that the method can be extended to the more widely used floating-point bit-string representation. Similar to fixed-point bit-strings, floating-point bit-strings can also be viewed as a hierarchical binary partitioning of parameter space $\mathbb{R}$ into non-overlapping sub-intervals. The key difference is that instead of partitioning each sub-interval into equal-width sub-intervals via the arithmetic mean of the interval endpoints, the floating-point representation partitions space into unequal-width sub-intervals via a geometric mean of the endpoints. More specifically, exponential bits denote a geometric partitioning, whilst man-

Table 1: Average testing set RMSE scores ($\pm$ s.d.) from a 5-fold CV experiment on UCI benchmark datasets. Methods are bolded based on a paired t-test ($p = 5\%$) with the best performing method, and show that floating-point BitVI performs similarly or better than fixed-point BitVI.

Dataset	(n, d)	10-bit BITVI (fixed)	10-bit BITVI (float)	6-bit BITVI (fixed)	6-bit BITVI (float)
YACHT	(308, 6)	$\mathbf{1.5752 \pm 0.5322}$	$\mathbf{1.7018 \pm 0.6175}$	2.0008 ± 0.5833	1.9141 ± 0.6273
KIN8NM	(8192, 8)	0.0964 ± 0.0056	$\mathbf{0.0781 \pm 0.0022}$	0.1177 ± 0.0037	0.1190 ± 0.0026
BOSTON	(506, 13)	$\mathbf{3.4388 \pm 0.4204}$	$\mathbf{3.4529 \pm 0.3189}$	$\mathbf{3.4500 \pm 0.2917}$	3.7481 ± 0.4802
ENERGY	(768, 8)	$\mathbf{0.7726 \pm 0.0376}$	0.9750 ± 0.0575	1.3458 ± 0.0858	1.5155 ± 0.1149
PROTEIN	(45730, 9)	4.7017 ± 0.0442	$\mathbf{4.4250 \pm 0.0395}$	4.8528 ± 0.0394	4.9367 ± 0.0308
POWER	(9568, 4)	4.1349 ± 0.0883	$\mathbf{4.0783 \pm 0.0889}$	4.3936 ± 0.0860	4.3988 ± 0.1080
CONCRETE	(1030, 8)	$\mathbf{6.0720 \pm 0.3952}$	$\mathbf{6.0300 \pm 0.3611}$	$\mathbf{6.6486 \pm 0.5015}$	6.7065 ± 0.4417
WINERED	(1599, 11)	$\mathbf{0.6409 \pm 0.0203}$	$\mathbf{0.6321 \pm 0.0404}$	$\mathbf{0.6322 \pm 0.0371}$	$\mathbf{0.6364 \pm 0.0346}$
NAVAL	(11934, 14)	0.0054 ± 0.0002	$\mathbf{0.0018 \pm 0.0002}$	0.0095 ± 0.0003	0.0096 ± 0.0003
WINEWHITE	(4898, 11)	$\mathbf{0.7097 \pm 0.0139}$	$\mathbf{0.7049 \pm 0.0106}$	$\mathbf{0.7132 \pm 0.0127}$	$\mathbf{0.7255 \pm 0.0126}$

tissa bits denote an arithmetic partitioning. See Algorithm 1 and Fig. 2 for more details.

Distributions of floats Let $C_\phi(\cdot)$ be a probability distribution over a length $\mathbf{b}$ floating-point bit-string which encodes parameter θ. We construct $C_\phi(\cdot)$ as a *deterministic* probabilistic circuit (PC) [Choi et al., 2020], which consists of a binary tree structure of computational nodes. The non-leaf nodes represent weighted summations with a set of weights $w_0, w_1 \in \phi$, and leaf nodes represent uniform density functions $\mathcal{U}(a_{\mathbf{b}_i}, b_{\mathbf{b}_{i+1}})$. The binary partitioning tree structure of the floating-point bit-string representation informs this structure, with endpoints $a_{\mathbf{b}_i}, b_{\mathbf{b}_{i+1}}$ of each of the 2^B input nodes corresponding to the intervals that two neighbouring[1] floating-point bit-strings $\mathbf{b}_i, \mathbf{b}_{i+1}$ define. Furthermore, the weights ϕ can be interpreted as corresponding to the conditional probabilities $p(b_i = 1 \mid b_{i+1}, \ldots, b_{B-1})$ of a bit given its preceding bits. A key advantage of using this formulation is that deterministic PCs allow for the computation of entropy exactly and efficiently [Vergari et al., 2021], which is a key aspect that our method leverages.

Variational approximation of bit-strings Given a data point $\mathbf{x} \in \mathcal{X}$, we aim to find the posterior distribution $p(\theta \mid \mathcal{X})$ of the neural network model $f_\theta(\mathbf{x})$ parameters θ. To this end, we use *variational inference* [Blei et al., 2017] to approximate this posterior with $C_\phi(\theta)$ as the variational distribution $q_\phi(\theta)$. Specifically, we optimize the parameters ϕ by maximizing the evidence lower bound (ELBO). We decompose the ELBO into an expected log-likelihood and an entropy term,

$$\begin{aligned}\mathcal{L}(\phi) &= \underbrace{\mathbb{E}_{q_\phi(\theta)}[\log p(\mathcal{X}, \theta)]}_{\text{Expected log-joint}} + \underbrace{\mathbb{H}[q_\phi(\theta)]}_{\text{Entropy}} \\ &\approx \sum_{i=1}^{K} \log p(\mathcal{X}_i, \theta_i) + \mathbb{H}[q_\phi(\theta)], \quad \theta_i \sim q_\phi(\theta).\end{aligned} \tag{1}$$

[1] By a "neighbouring" bit-string, we mean the bit-string which encodes the next nearest parameter value that can be represented by $\mathbf{b}$

Since the expected log-likelihood is intractable, we use Monte Carlo sampling to estimate it.

A measure of quantisation-induced uncertainty With an approximation of the parameter posterior distribution over bit-strings in hand, we can then compute our proposed metric for quantisation-induced information loss between bit-strings $\mathbf{b}$ and $\mathbf{b}'$, of lengths B and B' respectively where $B' < B$. This computation is defined as:

$$\underset{B \to B'}{\Delta I} = \mathbb{H}[q_\phi^B(\theta)] - \mathbb{H}[q_\phi^{B'}(\theta)], \tag{2}$$

where the hierarchical structure and tractability of the variational distribution $q_\phi = C_\phi(\theta)$ allows for computing the differential entropy $\mathbb{H}[\cdot]$ of the parameter posterior at the different bit-widths B and B'. This metric directly ties into the information-theoretic concepts of *information gain* and *mutual information* [Cover and Thomas, 2006, MacKay, 2002] of bit-strings $\mathbf{b}$ and $\mathbf{b}'$; We can effectively compute how many units of information of the original B-bit representation are lost in the B'-bit representation by computing the mutual information between bit-string $\mathbf{b}'$ and the discarded bits $\mathbf{b} \setminus \mathbf{b}'$. Alternatively, this can be stated as how many units of information would be *gained* if one added the discarded bits $\mathbf{b} \setminus \mathbf{b}'$ back to the B'-bit representation $\mathbf{b}'$. As such, the proposed metric directly gives the information lost when going from a B-bit representation to a B'-bit representation, subject to the posterior approximation. Intuitively, a larger change in entropy indicates a larger information loss induced by the quantisation operation and vice versa. This can then be used to identify parameters that are more sensitive to quantisation than others, and incorporated into decision-making about which parameters and what precision to quantise to.

3 EXPERIMENTS

The proposed method was empirically validated with two experiments conducted on 10 benchmark regression data sets from the UCI machine learning repository [Kelly et al.,

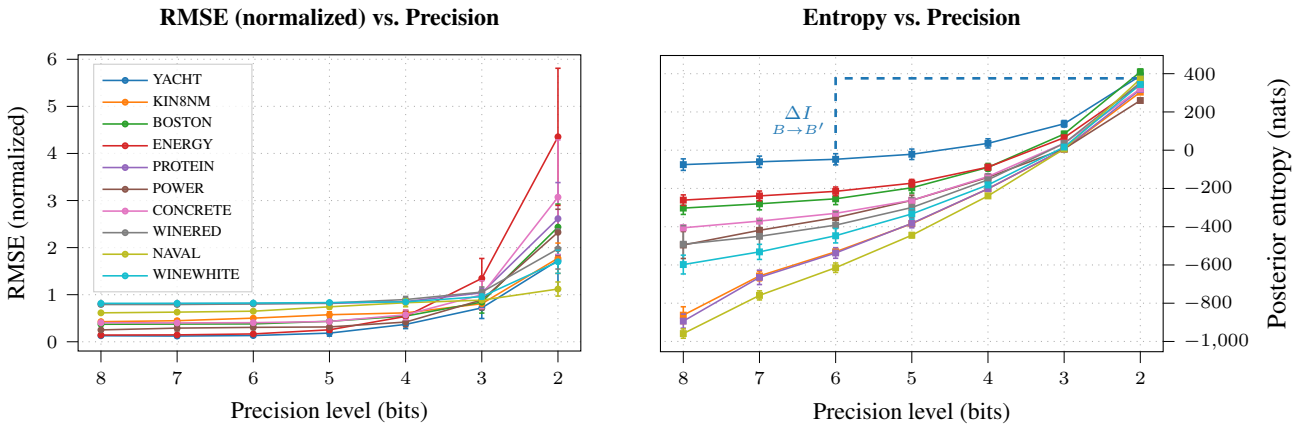


Figure 3: Full-network quantisation on UCI regression tasks. Normalised test RMSE and posterior entropy are shown as functions of parameter precision, with error bars denoting s.d. across folds. Both remain stable at moderate precision and increase below roughly 4–5 bits, indicating that the change in entropy signal tracks the onset of predictive degradation.

2025] via the *Bayesian Benchmarks*[2] testing framework. These regression tasks consist of predicting a 1D continuous value for data sets of various sizes and dimensionalities (See Table 1). We trained a small NN model consisting of two 16-unit hidden layers, with layer normalisation [Ba et al., 2016] and the $\tanh(\cdot)$ activation function applied after the hidden layers. For our posterior approximation, we used a Gaussian likelihood function $p(\mathbf{x} \mid \theta) = \mathcal{N}(f_\theta(\mathbf{x}), \sigma^2)$ with fixed variance σ^2, a Gaussian prior $p(\theta) = \mathcal{N}(\mu_p, \sigma_p^2)$, with μ_p and σ_p^2 selected such that the prior corresponds to the Kaiming initialization [He et al., 2015] and assumed that the NN parameters θ are independent of one another (mean-field assumption) in our variational family $q_\phi(\theta)$. We used 5-fold cross-validation, with a further 10% each training set reserved as a validation set used for early stopping, and report the average root mean squared error (RMSE) on the testing set across folds for each method and dataset.

Comparing fixed and floating-point BitVI We first compare the performance of the floating-point variant of BitVI with its original fixed-point variant. We trained 10-bit and 6-bit BitVI models with both the fixed-point and floating-point representations, and show the results in Table 1. The results on the 10-bit setting show that BitVI works similarly or better with floating-point representations. Reducing precision to 6-bits predictably leads to some performance degradation in both methods, but maintains surprisingly good performance given the quantisation level.

Quantisation-induced information loss We evaluated whether our proposed metric for quantisation-induced information loss accurately reflects the performance degradation observed with quantisation. The results of this experiment are shown in Fig. 3, which shows the change in RMSE with increasing quantisation (decreasing bit-width) from 8-bits to 2-bits, alongside the change in entropy. The metrics exhibit similar trends, with both RMSE and entropy remaining relatively stable until around 4–5 bits, after which they both increase rapidly.

[2]Bayesian Benchmarks GitHub repository by Hugh Salimbeni et al.

4 DISCUSSION & CONCLUSION

We proposed an approach based on the information-theoretic concept of entropy to measure the information loss that parameter quantisation induces in machine learning models. By approximating the posterior distribution over the underlying bit-string representations of parameters with a hierarchical and tractable probabilistic model, we could directly compute how the entropy of our approximation changes at different quantisation bit-widths. We also provided promising preliminary results suggesting that the proposed metric captures information loss that quantisation induces.

Our perspective is also closely related to ideas from probabilistic numerics, where numerical computation itself is treated as a source of uncertainty and numerical tasks are cast as inference problems [Hennig et al., 2015, Cockayne et al., 2019, Hennig et al., 2022]. While probabilistic numerics has mostly focused on problems such as integration and differential equations, the underlying message is directly relevant here. Finite-precision representation is not just a hardware detail, but rather a numerical constraint that shapes the uncertainty we can express about model parameters. Future work will explore scaling our method to study quantised large language models [Ma et al., 2024], investigate specialized number representations for deep learning, such as BFloats [Wang and Kanwar, 2019] and FP8 [Micikevicius et al., 2022], and study how to perform more targeted *mixed-precision* quantisation using our proposed metric to identify the safest parameters to quantise. In particular, we draw inspiration from Van Baalen et al. [2020] who also propose a Bayesian approach for quantisation, and elegantly unify quantisation and pruning.

Acknowledgements

We acknowledge funding from the Research Council of Finland (362408 and 339730). This work was a part of Finland's Ministry of Education and Culture's Doctoral Education Pilot under Decision No. VN/3137/2024-OKM-6 (The Finnish Doctoral Program Network in Artificial Intelligence, AI-DOC). We acknowledge the computational resources provided by the Aalto Science-IT project. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, abs/1607.06450, 2016.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

YooJung Choi, Antonio Vergari, and Guy Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. Technical report, University of California, Los Angeles (UCLA), 2020.

Jon Cockayne, Chris J Oates, Timothy John Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Ltd, 2006.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. `LLM.int8()`: 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 30318–30332. Curran Associates, Inc., 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLORA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 10088–10115. Curran Associates, Inc., 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 1026–1034. IEEE Computer Society, 2015.

Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179), 2015.

Philipp Hennig, Michael A Osborne, and Hans P Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

IEEE. IEEE standard for floating-point arithmetic. *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pages 1–84, 2019.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. `https://archive.ics.uci.edu`, 2025.

Donald E Knuth. *The Art of Computer Programming: Fundamental Algorithms, Volume 1*. Addison-Wesley Professional, 1997.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit LLMs: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, USA, 2002. ISBN 0521642981.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, et al. FP8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Aleksanteri Sladek, Martin Trapp, and Arno Solin. Approximate Bayesian inference via bitstring representations. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 286 of *Proceedings of Machine Learning Research*, pages 3939–3957. PMLR, 2025.

Pat H. Sterbenz. *Floating-point Computation*. Prentice-Hall, 1974.

Mart Van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, and Max Welling. Bayesian bits: Unifying quantization and pruning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 5741–5752. Curran Associates, Inc., 2020.

Antonio Vergari, YooJung Choi, Anji Liu, Stefano Teso, and Guy Van den Broeck. A compositional atlas of tractable circuit operations for probabilistic inference. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 13189–13201. Curran Associates, Inc., 2021.

Shibo Wang and Pankaj Kanwar. Bfloat16: The secret to high performance on cloud tpus. *Google Cloud Blog*, 4 (1), 2019.

Bit-Level Probabilistic Inference for Floating-Point Model Weights (Supplementary Material)

Aleksanteri Sladek[1,2] **Martin Trapp**[3] **Arno Solin**[1,2]

[1]Department of Computer Science, Aalto University, 02150 Espoo, Finland
[2]ELLIS Institute Finland, 02150 Espoo, Finland
[3]Department of Computing and Learning Systems , KTH Royal Institute of Technology, 114 28 Stockholm, Sweden

A APPENDIX A: ADDITIONAL THEORETICAL DETAILS

In this section, we provide additional material for building intuition on how the floating-point bit-string representations work. In Algorithm 1, we show an algorithm which iteratively partitions the parameter space $\mathbb{R}$. In Fig. 2, we show a visual comparison of how partitioning parameter space $\mathbb{R}$ with a fixed-point vs. a floating-point representation differs.

Algorithm 1: Floating-Point Partitioning Algorithm

Input: Bit-string $B = \{b_{B-1}, \ldots, b_0\}$, Exponent bits E, Mantissa bits M, No sign bit S
Output: Interval $[L, R]$

```
// Step 1: Constants and Bias Initialization
```
Bias $\leftarrow 2^{E-1} - 1$
$L \leftarrow 2^{-\text{Bias}}$ // Minimum non-zero magnitude
$R \leftarrow 2^{(2^E-1)-\text{Bias}+1}$ // Maximum magnitude

```
// Step 2: Exponent Partitioning (Geometric)
```
for $i \leftarrow B-1$ **to** $B-E-1$ **do**
  $G \leftarrow \sqrt{L \cdot R}$ // Partition using Geometric Mean
  if $b_i == 1$ **then**
    $L \leftarrow G$
  else
    $R \leftarrow G$
  end
end

```
// Step 3: Mantissa Partitioning (Arithmetic)
```
for $i \leftarrow B-E-1$ **to** 0 **do**
  $A \leftarrow \frac{L+R}{2}$ // Partition using Arithmetic Mean
  if $b_i == 1$ **then**
    $L \leftarrow A$
  else
    $R \leftarrow A$
  end
end
return $[L, R]$