A Geometric Bayesian Framework for Selective Quantization

Anand Paul[1] (✉)[0000−0002−0737−2021], John Pradeed Kulanthai Samy[2], Sathiyakeerthi Madasamy[3], and

[1] Louisiana State University Health Science Center, New Orleans, LA 70112, USA
apaul4@lsuhsc.edu
[2] Intel Corp, Santa Clara, CA 95054, USA
John.Pratheep.Kulanthai.Samy@intel.com
[3] Levi Strauss Co, San Francisco, CA 94111, USA
sathiyakeerthi@gmail.com

Abstract. This research paper introduces an advanced quantization methodology that leverages both Bayesian inference and geometric insights to optimize the encoding of high-dimensional latent variables. Our approach intelligently prioritizes variables based on their informational significance, effectively balancing computational efficiency with model fidelity. By incorporating a novel Bayesian statistical model, we evaluate variable importance and model stability across various computational platforms, including NVIDIA, Intel, and AMD hardware. Empirical results demonstrate significant enhancements in processing efficiency and resource utilization, making our methodology highly suitable for deployment in resource-constrained environments. This study sets a new benchmark for quantization practices in machine learning, particularly in applications requiring real-time data processing.

Keywords: Geometric Bayesian Quantization · SRate-Distortion Optimization · Probabilistic Modeling.

1 Introduction

Traditional quantization methods primarily focus on efficiency, often overlooking the complex underlying data geometry and parameter uncertainty. This oversight can result in suboptimal performance, particularly in tasks involving high-dimensional data. To overcome these limitations, we propose a novel quantization approach that incorporates Bayesian inference principles with geometric insights from Riemannian manifolds (2; 3). This method allows for a more sophisticated quantization process that assesses not just the values but also the uncertainties and geometric structure of the model space.

Integrating geometric structures into statistical inference has been explored, yet its application in quantization remains limited. Traditional quantization focuses on computational efficiency at the expense of accuracy, often ignoring the significant impact of data geometry on model performance (4). Works

such as Variational Autoencoders (VAEs) by (2) employ variational inference for Bayesian approximation but do not address geometric quantization directly. Likewise, (5) and (1) have applied probabilistic and Bayesian methods within neural networks and matrix factorization, yet without integrating geometric insights into quantization.

Our approach is inspired by these seminal works but extends them by explicitly incorporating Riemannian geometry into the quantization process, which is particularly beneficial for managing complex data representations and large-scale Bayesian models (3; 4).

2 Methodology

The foundation of our proposed method is to assess the informational value of each latent variable z_i by evaluating its variance $\sigma_i^2(x)$ and its contribution to the posterior probability. Latent variables with low variability and lesser impact on the posterior probability can be selectively omitted from the encoding process, thus reducing the model's complexity and improving computational efficiency.

2.1 Quantization Algorithm

Evaluation of Variable Importance The importance score S_i for each latent variable is computed based on its posterior variance and the mutual information between z_i and observed data x, expressed as:

$$S_i = \frac{1}{\sigma_i^2(x)} I(z_i; x)$$

where $I(z_i; x)$ can be approximated by the dependency of x on z_i captured through Variational Inference.

Threshold Determination A threshold T is established to balance between compression efficiency and accuracy. This threshold can be dynamically adjusted or fixed based on empirical evaluation.

Selective Encoding For latent variables where $S_i < T$, these variables are not encoded directly. Instead, their most probable values or mean values, given by $\mu_i(x)$, are used during the decompression phase. This strategy modifies the traditional encoding process by integrating a decision rule based on the importance scores and threshold.

Rate-Distortion Optimization The optimization of the rate-distortion function incorporates a penalty for omitting critical variables and is formulated as:

$$\min_{\xi_i} \left\{ \frac{1}{2\sigma_i^2(x)} \left(F^{-1}(\xi_i) - \mu_i(x) \right)^2 + \lambda R(\xi_i) + \rho \mathbf{1}_{S_i < T} \right\}$$

where ρ is the penalty for skipping important variables, and $\mathbf{1}_{S_i<T}$ is an indicator function that is active when $S_i < T$. Algorithm 1 provide how the step by step process of selective encoding quantization occurs.

Algorithm 1 Selective Encoding Quantization

Input: Latent variables z, observed data x, parameters (λ, ρ, T)
Output: Quantized representation ξ of z
Compute the posterior variance $\sigma_i^2(x)$ for each dimension i of z.
Calculate the mutual information $I(z_i; x)$ for each dimension.
Determine the importance score S_i for each latent variable z_i:
for $i = 1$ to K **do**
    $S_i = \frac{1}{\sigma_i^2(x)} I(z_i; x)$
end for
Establish the threshold T for important variables.
Encode each variable:
for $i = 1$ to K **do**
    if $S_i < T$ **then**
        Skip direct encoding of z_i.
        Use $\mu_i(x)$ or mean value for decompression.
    else
        Quantize z_i using the CDF and inverse CDF:
        $\xi_i = F(F^{-1}(z_i))$
        Optimize the rate-distortion trade-off:
        Minimize $\frac{1}{2\sigma_i^2(x)}(F^{-1}(\xi_i) - \mu_i(x))^2 + \lambda R(\xi_i) + \rho \mathbf{1}_{S_i < T}$
    end if
end for
return the quantized values ξ.

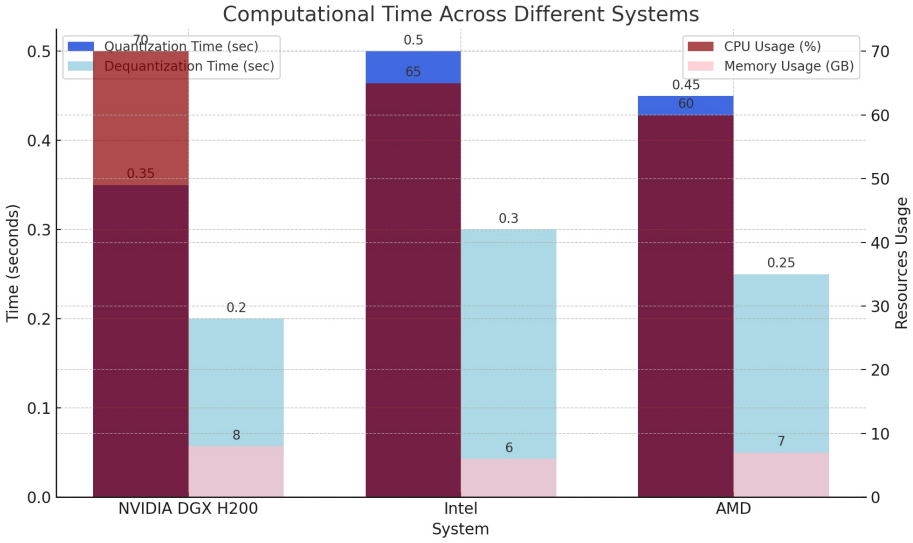

Fig. 1. Computational Time Across Different Systems

3 Experiments

This bar chart illustrates the time required for quantization and dequantization on NVIDIA DGX H200, Intel, and AMD systems (see Fig. 1). Quantization refers to time it takes to convert continuous variable values into discrete counterparts. Dequantization refers Time to revert the quantized data back to its original form.

4 Conclusion

This study introduced a novel Bayesian-Geometric quantization technique that significantly enhances the efficiency of machine learning models by intelligently prioritizing variable encoding. Future research could explore adaptive threshold mechanisms and extend this approach to real-time machine learning applications, potentially incorporating deep learning frameworks to further validate and refine the quantization process.

Author Conflict of Interest

The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Barkan, O.: Bayesian neural networks. Nature Communications **8**(15762) (2017). https://doi.org/10.1038/ncomms15762

[2] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: The International Conference on Learning Representations (2014)

[3] Li, Y., Turner, R.E.: Variational inference: A review for statisticians. Journal of the American Statistical Association **114**(527), 1845–1866 (2019). https://doi.org/10.1080/01621459.2018.1454641

[4] MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, UK (2003)

[5] Mnih, A., Salakhutdinov, R.R.: Probabilistic matrix factorization. In: Advances in Neural Information Processing Systems (2008)