Distributional Quantization

Felix Petersen[1] and Tobias Sutter[2]

[1]*Stanford University,* [2]*University of Konstanz*

Abstract

Distributional quantization is a simple, efficient, and asymptotically error-optimal scalar quantization method for quantizing data that follows a distributional prior. For example, for data following a Gaussian distribution, we provide a simple closed-form solution for an asymptotically error-optimal quantization scheme. Quantization is particularly important to machine learning for applications such as bandwidth reduction, neural network weight compression, among many others.

This is a pre-print excerpt of a larger work. The full article will be released later and the provided excerpt is released as a supplementary for the `distquant` *package.*

https://github.com/Felix-Petersen/distquant

1 Introduction

Quantization has a long-standing history in the field of information theory starting in the 1940s [1], [2]. Since then, quantization problems have been studied with applications in various disciplines, including signal processing [3], [4], cluster analysis [5], [6], pattern and speech recognition [7], numerical integration [8], and mathematical models in economics [9], [10].

More recently, with the increasing size of neural network models, quantization has gained interest in the machine learning community [11]–[13].

In this work, we cover the scalar quantization of data following a distributional prior from a theoretical perspective. We provide an asymptotically absolute error-optimal quantization scheme for data following a known distribution, vi7., Gaussian, logistic, uniform, exponential, Gumbel, and exponential family distributions. While our theory is asymptotic, i.e., for $n \to \infty$ quantization points, we also empirically demonstrate the precision of the approach in the few-bit regime and provide bounds for the regime of finite n. In addition, we show that the absolute error-optimal quantization method is W_1 distributionally robust.

2 Distributional Quantization

We start by introducing our notation and stating our assumptions. Following this, we state our main results and provide an extensive discussion of examples in Section 2.2.

Definition 1 (k-bit or n-point Quantization). *A k-bit or n-point quantization scheme comprises an encoder e and a decoder d. The encoder is described by a mapping e from the space of real numbers $\mathbb{R}$ to a set of n ($n = 2^k$ in the k-bit case) elements denoted as $[n] = \{0, 1, \dots, n-1\}$. The decoder is a mapping d that, for each element, yields a representative number from the space of real numbers, i.e.,*

$$\mathsf{e} : \mathbb{R} \to [n] \qquad\qquad \mathsf{d} : [n] \to \mathbb{R}$$

A popular method of quantization is uniform quantization, which is typically defined via an encoder $\mathsf{e}(x) = \lfloor (x-a) \,/\, (b-a) \cdot n \rfloor$ for $x \in [a, b)$. The respective decoder or de-quantization function d is defined as $\mathsf{d}(y) = a + (b-a) \cdot (y + 0.5) \,/\, n$. Uniform quantization is optimal wrt. the expected error if the distribution of the values to be quantized is uniform with support on $[a, b)$. However, if the distribution of values to be quantized is not uniform (e.g., in neural networks), then uniform quantization is suboptimal, which motivates the proposal of distributional quantization.

Assumption 1 (Existence of a Density)**.** *Assume that inputs $x \in \mathbb{R}$, i.e., the values to be quantized, are distributed according to a probability distribution that admits a Lipschitz-continuous differentiable density g with Lipschitz constant L.*

To quantify the quality of a k-bit or n-point quantizer, we consider the *expected absolute quantization error*, which is the expectation of the absolute differences between the inputs x and a quantization round-trip ($\mathsf{d}(\mathsf{e}(x))$):

$$\mathbb{E}_{x\sim g}\big[\big|x - \mathsf{d}(\mathsf{e}(x))\big|\big]\,. \tag{1}$$

Definition 2 (Distributional Quantization)**.** *We define distributional quantizers as quantization methods in accordance with Definition 1 that are based on a probability distribution defined on $X \subset \mathbb{R}$ with a probability density function (PDF) f and a cumulative density function (CDF) F. The encoder of the distributional quantizer is defined as*

$$\mathsf{e}_f : \mathbb{R} \to \{0, 1, ..., n-1\}, \quad \mathsf{e}_f(x) = \big\lfloor F(x) \cdot n \big\rfloor, \tag{2}$$

and the decoder is respectively defined as

$$\mathsf{d}_f : \{0, 1, ..., n-1\} \to X, \quad \mathsf{d}_f(y) = F^{-1}\left(\frac{y + 1/2}{n}\right)\,. \tag{3}$$

Before stating our main results on the absolute expected quantization error, we first characterize the quantization error for data following a distribution with PDF g and CDF G in dependence of a distributional quantizer $\mathsf{e}_f, \mathsf{d}_f$ as characterized via a PDF f and a CDF F in accordance with Definition 2. Specifically, the expected absolute quantization error is

$$\mathbb{E}_{x\sim g}\left[|x - \mathsf{d}_f(\mathsf{e}_f(x))|\right] = \mathbb{E}_{x\sim g}\left[\left|x - F^{-1}\left(\frac{\lfloor F(x)\cdot n\rfloor + .5}{n}\right)\right|\right]\,. \tag{4}$$

Proposition 1 (Absolute Error Optimal Quantization (Continuous Case))**.** *Let the input values be distributed according to an L–Lipschitz continuous density g and assume that $\sqrt{g}$ is integrable. Then,*

$$\lim_{n\to\infty}\; n \cdot \mathbb{E}_{x\sim g}[|x - \mathsf{d}_f(\mathsf{e}_f(x))|] = \int_X \frac{g(x)}{4 \cdot f(x)}\, dx\,. \tag{5}$$

Asymptotically, the distributional quantizer $\mathsf{e}_f, \mathsf{d}_f$ that leads to the smallest absolute quantization error uses the density $f = \kappa\sqrt{g}$ where $\kappa = (\int_X \sqrt{g(x)}dx)^{-1}$, leading to an error of $1/(4\kappa^2 n)$.

In the following, in addition to the asymptotic case, we also consider the extension to the discrete case of finite n. Here, we provide an upper bound for the quantization error and find that the quantizer minimizing this upper bound also corresponds to the quantizer that is also optimal for the asymptotic case.

Proposition 2 (Absolute Error Optimal Quantization (Discrete Case))**.** *Let the input values be distributed according to an L–Lipschitz continuous density g with bounded support. Then, for any $n \in \mathbb{N}_+$*

$$\mathbb{E}_{x\sim g}[|x - \mathsf{d}(\mathsf{e}(x))|] < \sum_{i=0}^{n-1} \frac{g(y_i)}{3n^2 \cdot f(y_i)^2} \qquad \textit{where} \qquad y_i = F^{-1}\left(\frac{i + 1/2}{n}\right) \tag{6}$$

and

$$\mathbb{E}_{x\sim g}[|x - \mathsf{d}(\mathsf{e}(x))|] < \left(\sum_{i=0}^{n-1} \frac{g(y_i)}{4n^2 \cdot f(y_i)^2}\right) + \mathcal{O}\left(\frac{L}{n^2}\right) + \mathcal{O}\left(\frac{L^2}{n^3}\right) . \tag{7}$$

The distributional quantizer $\mathsf{e}_f, \mathsf{d}_f$ *that minimizes the upper bound of the absolute quantization error in Equation 6 uses a density of* $f = \omega\sqrt{g}$ *where* $\omega = (\sum_i \sqrt{g}(y_i))^{-1}$*, leading to an error bound of* $1/(3\omega^2 n)$*. Further, the distributional quantizer* $\mathsf{e}_f, \mathsf{d}_f$ *that asymptotically* $(n \to \infty)$ *minimizes the upper bound of the absolute quantization error in Equation 7 also uses a density of* $f = \omega\sqrt{g}$ *where* $\omega = (\sum_i \sqrt{g}(y_i))^{-1}$ *and asymptotically leads to an error bound of* $1/(4\omega^2 n)$*. This provides an asymptotic equivalence to Proposition 1.*

Proofs and additional statements will be provided in the full release of the article.

2.1 W_1 Distributionally Robust Quantization

In many practical situations, the true distribution of the input signal $x \sim \mathbb{P}$ may be unknown. Instead, one may have access to a prior belief or guess of what $\mathbb{P}$ could be, described by $\widehat{\mathbb{P}}$. In such settings, it is natural to consider a quantization scheme that minimizes the expected absolute quantization error under the worst-case input distribution that is consistent with the prior $\widehat{\mathbb{P}}$, i.e.,

$$\max_{\mathbb{Q}\in\mathbb{B}_\rho^{\mathsf{W}_1}(\widehat{\mathbb{P}})} \mathbb{E}_{x\sim\mathbb{Q}}\big[|x - \mathsf{d}(\mathsf{e}(x))|\big] \tag{8}$$

where $\mathbb{B}_\rho^{\mathsf{W}_1}(\widehat{\mathbb{P}})$ denotes a W_1 ball of probability distributions centered around the prior $\widehat{\mathbb{P}}$:

$$\mathbb{B}_\rho^{\mathsf{W}_1}(\widehat{\mathbb{P}}) = \left\{\mathbb{Q} \in \mathcal{P}(X) \ : \ \mathsf{W}_1(\mathbb{Q}, \widehat{\mathbb{P}}) \leq \rho\right\} . \tag{9}$$

We assume the prior distribution $\widehat{\mathbb{P}}$ admits an L-Lipschitz continuous density $\widehat{g}$. Our goal is then to find the best distributional encoder/decoder pair (e, d) to minimize the expected absolute quantization error (8) under the worst-case distribution $\mathbb{Q}$ lying in the ball $\mathbb{B}_\rho^{\mathsf{W}_1}(\widehat{\mathbb{P}})$. See [14], [15] for a comprehensive discussion on distributionally robust modelling and optimization.

Proposition 3 (W_1 Distributionally Robust Quantization)**.** *Let the distance in (9) be the Wasserstein 1-distance and* $\rho \geq 0$*. Assume that the prior distribution* $\widehat{\mathbb{P}}$ *satisfies Assumptions 1 with density denoted by* $\widehat{g}$ *and additionally* $\sqrt{\widehat{g}}$ *is integrable. Then, asymptotically, the distributional quantizer* $\mathsf{e}_f, \mathsf{d}_f$ *that leads to the smallest worst-case expected absolute quantization error uses the density* $f = \kappa\sqrt{\widehat{g}}$ *where* $\kappa = (\int_X \sqrt{\widehat{g}(x)}dx)^{-1}$*, leading to a quantization error of* $1/(4\kappa^2 n) + \rho$*.*

Proposition 3 states that the proposed distributional quantization scheme (from Proposition 1) is robust with respect to the input distribution, i.e., choosing the distributional quantization scheme induced by the prior belief distribution $\widehat{\mathbb{P}}$ is an optimal choice if the true input distribution is unknown and approximated by a prior belief.

2.2 Examples

In the presented distributional quantization, the k-bit encoder and decoder are parameterized via the CDF F. In other words, different CDFs induce different quantization schemes. The aim of this paper is to study what is the statistically optimal choice of encoder and decoder (i.e., leading to the smallest absolute quantization error). We illustrate this by presenting different distributional quantization schemes:

Example 1 (Uniform Quantization)**.** *The popular vanilla uniform quantization appears as a natural example of distributional quantization, where the distribution is uniform.*

In the following, we show 3 quantization schemes induced by CDFs of popular distributions.

Example 2 (Gaussian CDF Quantization). *An example of distributional quantization is the Gaussian CDF quantization, i.e., $F(x)$ is a Gaussian CDF.*

Example 3 (Logistic CDF Quantization). *Another interesting example of is the logistic CDF quantization. For this, the vanilla form of the logistic CDF is $F(x) = \frac{1}{1+e^{-x}}$, aka. the (logistic) sigmoid. Accordingly, $F^{-1}(y) = \log(\frac{y}{1-y})$ corresponds to the logit function.*

Example 4 (Cauchy CDF Quantization). *A special example of distributional quantization is Cauchy quantization, i.e., quantization based on a Cauchy distribution. This case is special because the Cauchy distribution is a distribution with infinite variance and extreme outliers, leading to a substantial robustness of Cauchy quantization against outliers.*

For a given input distribution, Proposition 1 states how to select the distributional quantizers leading to the information-theoretically minimal expected absolute quantization error. We explicitly highlight this for the following prominent input distributions, which we also summarize in Table 1.

Example 5 (Uniform input distribution). *If the input distribution is uniform on the interval $[a, b]$, then the optimal distributional quantization method according to Proposition 1 uses CDF F that corresponds to a uniformly distributed random variable on $[a, b]$, i.e., $F(x) = G(x)$.*

Example 6 (Gaussian input distribution). *Assume that the input distribution is Gaussian with mean μ and variance σ^2. Then, the absolute error optimal distributional quantization method according to Proposition 1 uses a CDF F that corresponds to a Gaussian random variable with mean μ and variance $2\sigma^2$, i.e., $F(x) = \frac{1}{2}\left(1 + \mathsf{erf}\left(\frac{x-\mu}{2\sigma}\right)\right)$.*

This means that under the mentioned modification of the variance, Gaussian CDF quantization is the information theoretically optimal quantization for data following a Gaussian distribution.

Example 7 (Logistic input distribution). *If the input is distributed according to the logistic distribution with centered at zero and with unit scale parameter, i.e., $g(x) = \frac{e^{-x}}{(1+e^{-x})^2}$, then we get $\kappa = 1/\pi$ and according to Proposition 1 the CDF F of the optimal distributional quantizer is $F(x) = \frac{2}{\pi} \cdot \tan^{-1}(e^{x/2})$.*

Example 8 (Gumbel input distribution). *If the input is distributed according to the Gumbel distribution, i.e., $g(x) = e^{-x-e^{-x}}$, then $\kappa = 1/\sqrt{2\pi}$ and according to Proposition 1 the CDF F of the optimal distributional quantizer is $F(x) = 1-\mathsf{erf}\left(\frac{e^{-x/2}}{\sqrt{2}}\right)$, where erf denotes the error function.*

Example 9 (Exponential family input distribution). *A convenient and remarkably general setting is to consider an input distribution belonging to an exponential family of distributions [16], i.e.,*

$$g(x) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x),$$

where $\eta \in \mathbb{R}^s$ denotes the (natural) parameter, T is the sufficient statistic, $A(\eta)$ the log-partition function and h the base measure. It can be shown that $f(x) = \kappa\sqrt{g(x)}$ then is also the density of belonging to the exponential family with same sufficient statistic updated natural parameter $\frac{1}{2}\eta$, base measure $\sqrt{h(\cdot)}$, i.e.,

$$f(x) = \exp\left(\sum_{i=1}^{s} \frac{\eta_i}{2} T_i(x) - C(\eta)\right) \sqrt{h(x)}, \tag{10}$$

where $C(\eta)$ is a constant ensuring that $\int_X f(x)dx = 1$. If h is constant then the distribution f describes the same distribution as g (for different natural parameters). We note that Example 6 is a special case of the exponential family distribution.

Table 1: This table shows the CDFs corresponding to different distributional quantization schemes moment-matched such that the location parameter is zero and the scale is one. Given is the CDF of the distribution as well as the CDF of the *absolute error optimal* quantization scheme according to Proposition 1. The quantizers are efficiently computable as well as invertable.

Distribution		CDF	Abs. Error Optimal F
Gaussian	μ=0 σ=1	$\frac{1}{2}\left(1+\mathsf{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$	$\frac{1}{2}\left(1+\mathsf{erf}\left(\frac{x}{2}\right)\right)$
Logistic	μ=0 s=1	$\frac{1}{1+e^{-x}}$	$\frac{2}{\pi}\tan^{-1}\left(e^{x/2}\right)$
Exponential	λ=1	$1-e^{-x}$	$1-e^{-x/2}$
Gumbel	μ=0 s=1	$e^{-x-e^{-x}}$	$\mathsf{erfc}\left(\frac{e^{-x/2}}{\sqrt{2}}\right)$

Example 10 (Cauchy input distribution)**.** *Assume that the input is distributed according to the simplest Cauchy distribution, i.e.,* $g(x) = \frac{1}{\pi(1+x^2)}$*. This example does not satisfy the assumptions of Proposition 1 as* κ *is unbounded. Moreover, in this setting, expected absolute quantization error for any quantization method is unbounded too, i.e.,* $\mathbb{E}_{x\sim g}[|x-\mathsf{d}(\mathsf{e}(x))|]=\infty$.

Remark 1. *The behavior observed in Example 10 for the Cauchy input distribution can be generalized to a heuristic: Except for a small set of distributions* ($g \notin \Omega(1/x^2)\cup\mathcal{O}(1/x^{2+\alpha}), \alpha>0$)*, the properties of* $\kappa=\infty$ *and an infinite expected error* $\mathbb{E}_{x\sim g}[|x-\mathsf{d}(\mathsf{e}(x))|]=\infty$ *are equivalent.*

3 Empirical Evaluation

We implemented distributional quantization in the `distquant` package. We present simulation plots at `https://github.com/Felix-Petersen/distquant/blob/main/SIMULATIONS.md`, where we also compare against float16, histogram-based quantization, Lloyd max quantization [5], etc.

In this section, we present results for quantizing the weights of off-the-shelf CIFAR-10 models. We consider Gaussian optimal quantization as well as logistic CDF quantization and also consider float16, bfloat16, and qint8 as baselines. We do not utilize any post-quantization training / fine-tuning because we want to observe the raw performance of the models after quantization. The results are displayed in Table 2. We can observe that distributional quantization provides a strong method for quantizing neural network weights. In particular, at quantization to 8 bits, our models do not suffer a loss in accuracy. When we go to the extreme level of 4 bits, the acccuracy drops by only $2-4\%$. Even at 3 bits, the models are still usable, maintainting most of their performance; with post-quantization fine-tuning strategies, the performance could be improved.

Table 2: Performance of pre-trained CIFAR-10 neural network models compressed using distributional quantization without any post-quantization training / fine-tuning.

Model	Orig.	F16	BF16	qint8	Gaussian (ours)				Logistic CDF (ours)			
k (bits)	32	16	16	8	3	4	6	8	3	4	6	8
ResNet-18 [17]	93.1	93.1	93.0	11.3	85.5	91.2	93.0	93.1	72.7	89.3	92.7	93.1
ResNet-50 [17]	93.6	93.7	93.7	16.5	74.8	89.7	93.2	93.6	14.8	90.3	93.3	93.6
MobileNet-V2 [18]	93.9	93.9	93.8	11.6	25.3	80.6	93.2	93.8	17.1	11.6	93.1	94.0
VGG-13 [19]	94.2	94.2	94.2	21.0	89.2	92.5	94.1	94.3	86.6	91.6	94.0	94.3

Conclusion

Distributional quantization provides an information theoretically optimal quantization strategy for quantizing data following a distributional prior. We provide an PyTorch package "`distquant`" for applying distributional quantization and included some empirical evaluations. We will release the full version of this article in the future, including additional statements as well as all proofs.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[2] N. J. A. Sloane and A. D. Wyner, "Coding theorems for a discrete source with a fidelity criterioninstitute of radio engineers, international convention record, vol. 7, 1959.," in *Claude E. Shannon: Collected Papers*. IEEE, 1993.

[3] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Interscience, 2006, ISBN: 0471241954.

[5] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[6] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, 1960.

[7] V. Digalakis, L. Neumeyer, and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the world wide web," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 1, pp. 82–90, 1999.

[8] G. Pagés, "A space quantization method for numerical integration," *J. Comput. Appl. Math.*, vol. 89, no. 1, pp. 1–38, Mar. 1998.

[9] G. C. Pflug, "Scenario tree generation for multiperiod financial optimization by optimal discretization," in 2, Ser. B, vol. 89, Mathematical programming and finance, 2001, pp. 251–271.

[10] G. C. Pflug and A. Pichler, "Approximations for probability distributions and stochastic optimization problems," in *Stochastic optimization methods in finance and energy*, ser. Internat. Ser. Oper. Res. Management Sci. Vol. 163, Springer, New York, 2011, pp. 343–387.

[11] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," *Computing Research Repository (CoRR) in arXiv*, 2018.

[12] M. Shkolnik, B. Chmiel, R. Banner, *et al.*, "Robust quantization: One model to rule them all," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5308–5317.

[13] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Llm.int8(): 8-bit matrix multiplication for transformers at scale," *Proc. Neural Information Processing Systems (NeurIPS)*, 2022.

[14] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *Computing Research Repository (CoRR) in arXiv*, 2019.

[15] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*, INFORMS, 2019, pp. 130–166.

[16] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. Springer, 1998.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computing Research Repository (CoRR) in arXiv*, 2014.