Goal-Oriented 1-Bit Quantization with Uncertain Goals

Malcolm Egan

***Abstract*—**

I. Introduction

II. Problem Formulation

Consider a data source $P_{\mathbf{X}}$ with data $\mathbf{X} \in \mathbb{R}^d,~d \geq 1$. A 1-bit quantizer of $\mathbf{X}$ consists of a reproduction set $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2\} \subset \mathbb{R}^d$ and the quantizer function $Q : \mathbb{R}^d \to \mathcal{Y},~\mathbf{x} \mapsto \hat{\mathbf{x}}$.

The distortion in the quantizer is measured via $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. A standard choice of the distortion function is the Euclidean square error $D_{\mathrm{SE}}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2,~\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^d$. The performance criterion for the quantizer is then defined as

$$\mathcal{E} = \mathbb{E}[D(\mathbf{X}, Q(\mathbf{X}))], \tag{1}$$

with the optimal quantizer given by

$$\min_{\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^d} \mathbb{E}[D(\mathbf{X}, Q(\mathbf{X}))]. \tag{2}$$

The optimal quantizer is then specified by the source distribution $P_{\mathbf{X}}$ and the distortion function D. The problem of coping with an imperfectly specified $P_{\mathbf{X}}$ or D is known as mismatched quantization. In the case the distortion function is unknown, a common problem is to characterize the performance loss when the quantizer is designed based on D_{SE}.

In this paper, we address the question: *is there a better design criterion for* 1*-bit quantization with an imperfectly specified distortion function?*

III. Impact of Uncertainty in the Distortion

In order to understand the impact of uncertainty on performance, we consider two models of distortion functions.

A. Data-Weighted Distortion

Given continuous data $\mathbf{X} \in \mathbb{R}^d$ with dependent elements admitting the joint density $p_{\mathbf{X}}$, the standard distortion criterion is given by

$$\begin{aligned}&\mathbb{E}[D_{\mathrm{SE}}(\mathbf{X}, Q(\mathbf{X}))] \\ &= \int_{\mathbb{R}^d} D_{\mathrm{SE}}(\mathbf{x}, Q(\mathbf{x})) p_{\mathbf{X}} \mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} D_{\mathrm{SE}}(\mathbf{x}, Q(\mathbf{x})) C(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^d p_{X_i}(x_i) \mathrm{d}\mathbf{x}.\end{aligned} \tag{3}$$

Let $K(\mathbf{x}) = c(F_1(x_1), \ldots, F_d(x_d))$. Then,

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}}[D_{\mathrm{SE}}(\mathbf{X}, Q(\mathbf{X}))] \\ &= \mathbb{E}_{\tilde{\mathbf{X}}}[K(\mathbf{X}) D_{\mathrm{SE}}(\tilde{\mathbf{X}}, Q(\tilde{\mathbf{X}}))],\end{aligned} \tag{4}$$

where $\tilde{X} \sim \prod_{i=1}^d P_{X_i}$; that is, the elements of $\tilde{\mathbf{X}}$ are independent. The function $c : [0,1]^d \to \mathbb{R}_+$ is known as the copula density function.

Uncertainty in the dependence structure of $\mathbf{X}$ is captured via this distortion function. This suggests a useful distortion criterion in this setting is

$$D_K(\mathbf{X}, Q(\mathbf{X})) = K(\mathbf{X}) D_{\mathrm{SE}}(\mathbf{X}, Q(\mathbf{X})). \tag{5}$$

B. p-Error Distortion

A common distortion function for post-training quantization of neural networks [1] is the p-error, given by

$$D_p(\mathbf{X}, \hat{\mathbf{X}}) = \sum_{i=1}^d |x_i - \hat{x}_i|^p, \tag{6}$$

where $\hat{\mathbf{X}} = Q(\mathbf{X})$.

C. Impact of Distortion Uncertainty for Symmetric Quantizers

To illustrate the impact of uncertainty in the data-weighted and p-error distortion functions, we consider the case of symmetric quantizers. We allow $\mathcal{Y} = \{-B, B\}$ for $B \in \mathbb{R}_+$. The quantization point is chosen via

$$Q(x) = \arg\min_{\hat{x} \in \{-B, B\}} (x - \hat{x})^2. \tag{7}$$

Fig. 1 plots the impact on the true distortion metric for three scenarios. Observe that the optimal quantization point magnitude (i.e., B) is significantly different between the three scenarios. This suggests that utilizing a quantizer based on D_{SE} is undesirable. If the target distortion function is known, it can be utilized to construct a new quantizer. However, ...

IV. Risk-Averse Quantization

A. Risk-Averse Criteria

Suppose that $\mathcal{D}$ is a set of distortion functions $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ and P_D is a prior distribution on $\mathcal{D}$. In this case, we seek to construct a quantizer solving

$$\min_Q \mathbb{E}_{D, \mathbf{X}}[D(\mathbf{X}, Q(\mathbf{X}))], \tag{8}$$

This is intractable in general. A standard approach is to first choose a reference distortion function D_{ref}; e.g.,

$$D_{\mathrm{ref}}(\mathbf{X}, Q(\mathbf{X})) = \|\mathbf{X} - \hat{\mathbf{X}}\|^2. \tag{9}$$

M. Egan is with Univ Lyon, INRIA, INSA Lyon, CITI, France (email: malcolm.egan@inria.fr)

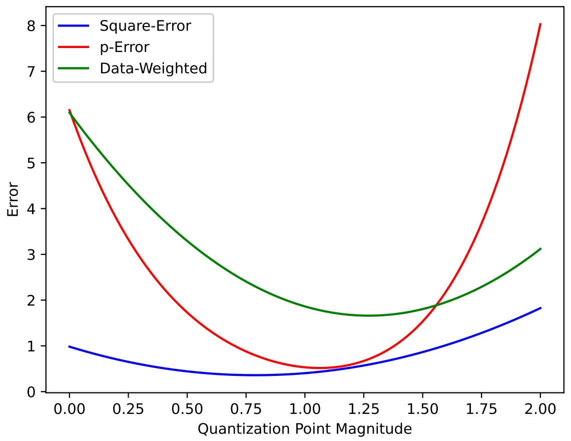

Fig. 1. Impact of distortion functions. $X \sim \mathcal{N}(0,1)$, $p=5$, $K(x)=\exp(|x|)$.

The quantizer is then constructed via

$$\min_{Q} \mathbb{E}[D_{\text{ref}}(\mathbf{X},Q(\mathbf{X}))]. \tag{10}$$

While standard, this approach ignores any uncertainty in the distortion function.

Fix the quantizer Q and note that

$$\begin{aligned}\mathbb{E}_{D,\mathbf{X}}[D(\mathbf{X},Q(\mathbf{X}))] &= \mathbb{E}_D\left[\int_0^\infty \mathbb{P}(D(\mathbf{X},Q(\mathbf{X}))>u)\mathrm{d}u\right]\\ &= \int_0^\infty \mathbb{E}_D\left[\mathbb{P}(D(\mathbf{X},Q(\mathbf{X}))>u)\right]\mathrm{d}u.\end{aligned} \tag{11}$$

In order to account for uncertainty, we seek a conservative estimate of $\mathbb{E}_D\left[\mathbb{P}(D(\mathbf{X},Q(\mathbf{X}))>u)\right]$ for each $u \in [0,\infty)$. Let $h:[0,1]\to[0,1]$ be a concave function satisfying $h(0)=0$ and $h(1)=1$. We then have

$$\mathbb{E}_D\left[\mathbb{P}(D(\mathbf{X},Q(\mathbf{X}))>u)\right] \le \mathbb{E}_D\left[h\left(\mathbb{P}(D(\mathbf{X},\hat{\mathbf{X}}))\right)\right]. \tag{12}$$

The final step (as applied in the expected distortion case) is to utilize the approximation

$$\mathbb{E}_D\left[h\left(\mathbb{P}(D(\mathbf{X},\hat{\mathbf{X}}))\right)\right] \approx h\left(\mathbb{P}(D_{\text{ref}}(\mathbf{X},\hat{\mathbf{X}}))\right). \tag{13}$$

The resulting quantity

$$\rho(D_{\text{ref}}(\mathbf{X},Q(\mathbf{X}))) = \int_0^\infty h\left(\mathbb{P}(D_{\text{ref}}(\mathbf{X},Q(\mathbf{X})))\right)\mathrm{d}u \tag{14}$$

is known as a *distortion risk measure*. By increasing the nonlinearity of h, a greater level of uncertainty in D can be accounted for.

B. Optimal Risk-Averse Quantization

An optimal risk-averse quantizer is defined as

$$\min_{\mathbf{y}_1,\mathbf{y}_2\in\mathbb{R}^d} \rho_h(D(\mathbf{X},Q(\mathbf{X}))). \tag{15}$$

In the special case $h(w)=w$, we recover the quantizer in (2). On the other hand, for general h, the solution to (15) differs from (2).

We first consider the optimal decision rule. Let $\mathcal{Y}=\{\mathbf{y}_1,\mathbf{y}_2\}$ be fixed quantization points and $Q:\mathbb{R}^d\to\mathcal{Y}$ be an arbitrary decision rule. Observe that

$$\begin{aligned}&\int_0^\infty h\left(\mathbb{P}\left(D(\mathbf{X},Q(\mathbf{X}))>u\right)\right)\mathrm{d}u\\ &\ge \int_0^\infty h\left(\mathbb{P}\left(\min_{\hat{\mathbf{x}}\in\{\mathbf{y}_1,\mathbf{y}_2\}} D(\mathbf{X},\hat{\mathbf{x}})>u\right)\right)\mathrm{d}u\end{aligned} \tag{16}$$

since for each $u\ge 0$,

$$\mathbb{P}\left(D(\mathbf{X},Q(\mathbf{X}))>u\right) \ge \mathbb{P}\left(\min_{\hat{\mathbf{x}}\in\{\mathbf{y}_1,\mathbf{y}_2\}} D(\mathbf{X},\hat{\mathbf{x}})>u\right). \tag{17}$$

As such, the optimal decision rule is given by

$$Q^*(\mathbf{x}) = \min_{\hat{\mathbf{x}}\in\{\mathbf{y}_1,\mathbf{y}_2\}} D(\mathbf{x},\hat{\mathbf{x}}), \tag{18}$$

which is the same as for the expected distortion criteria. The optimal quantization points can therefore be obtained via

$$\min_{\mathbf{y}_1,\mathbf{y}_2} \int_0^\infty h\left(\mathbb{P}\left(\min_{\hat{\mathbf{x}}\in\{\mathbf{y}_1,\mathbf{y}_2\}} D(\mathbf{X},\hat{\mathbf{x}})>u\right)\right)\mathrm{d}u. \tag{19}$$

This problem can be solved via the cross-entropy algorithm detailed in Alg. 1.

Algorithm 1 Quantizer Optimization Algorithm

```
Input: Maximum number of iterations $T_{\max}$, samples from $P_X$ $\{\mathbf{x}_i\}_{i=1}^S$, samples for CEM $N$, number of elite samples $N_e$, smoothing parameter $\alpha$, reference distortion function $D_{ref}$, risk distortion function $h$, initial search parameters $\mu$, $\sigma$.
Initialize $t=0$, $Y^*$, and $\rho_{\text{best}}=\infty$.
for $t<T_{\max}$ do
    $t \leftarrow t+1$
    Sample quantizers $Y_i \sim \mathcal{N}(\mu,\sigma^2)$, $i=1,\ldots,N$.
    Estimate risk measure for each $Y_i$, $i=1,\ldots,N$.
    if $\rho(Y_i)<\rho_{best})$ then
        $\rho_{\text{best}} \leftarrow \rho(Y_i)$, $Y^* \leftarrow Y_i$.
    end if
    Update $\mu,\sigma$.
end for
Output: $Y^*$
```

C. Choice of Risk Measure

Distorting the probabilities via the function h allows us to account for uncertainty in the distortion function. However, the choice of h remains. In order to make this decision, we need to relate h to the uncertainty in D. Clearly if D_{ref} is perfectly known, then we can choose $h(w)=w$. This is also the case if the quantizer minimizing $\mathbb{E}[D_{ref}]$ is also the optimum for all $D\in\mathcal{D}$.

The need for a more general risk measure arises when there is a distortion function in $\mathcal{D}$ which has a very different quantizer and the resulting distortion is much higher using the quantizer for D_{ref}. Ideally we should choose h such that for all u, $h(\mathbb{P}(D_{ref}(X, Q(X)) > u)) = \mathbb{P}(D(X, Q(X)) > u)$, where D is a "worst case distortion function". Of course it is unreasonable to expect we can do this. But it gives an idea: choose a test distortion function D and choose h such that, for a given quantizer, we have equality in h. As h is often parameterized by a small number of parameters, this will give us these parameters.

V. Numerical Results

The main thing to show is that if we optimize with a risk measure, we can get a solution that behaves like the solution to the expected nonstandard distortion. This is possible to a certain extent with the square error distortion. I want to show that the risk measure can increase and decrease the quantization point locations.

k = 1 new distort [-1.01419832 0.8437102] 0.14479253483548768 k = 0.1 new distort [-1.3904106 1.60435738] 0.5774578661307037

k = 1 se [-0.72524994 0.55889982] 0.1871454056077626 0.19905564064955272 k = 0.1 se [-0.93519153 1.12675629] 0.9032207311548344 0.16920261344585585

k = 1 se [-0.7373108 0.55402933] 0.18713752387147375 0.19905575152095342 k = 0.01 se [-0.93847094 1.13123067] 1.0155872930497436 0.1702801881449852

k = 1 se [-0.73045476 0.55232584] 0.18711365708716696 k = 0.001 se [-0.94311303 1.12859039] 1.0249175412759495

k = 1 se [-0.72600127 0.55217849] 0.18712239007108994 k = 0.00001 [-0.941351 1.1255944] 1.0323499972928063

An observation is that there seems to be a limit on how much the solution can be perturbed.

References

[1] Y. Nahshan *et al.*, "Loss aware post-training quantization," *Machine Learning*, vol. 110, no. 11, pp. 3245–3262, 2021.