Novel Quantization Strategies for Linear Prediction with Guarantees

Simon S. Du* SSDU@CS.CMU.EDU
Yichong Xu* YICHONGX@CS.CMU.EDU
Yuan Li LIYUANCHRISTY@GMAIL.COM
Hongyang Zhang HONGYANZ@CS.CMU.EDU
Aarti Singh AARTI@CS.CMU.EDU
Pulkit Grover PGROVER@ANDREW.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213, USA

*: **Contribute equally.**

Abstract

Quantized data is the norm in many energy constrained problems, a concrete example being brain signals recorded by distributed sensors placed around the head in Brain-Computer Interface (BCI) applications. However, machine learning algorithms typically ignore the quantized nature of such data. In this paper, we undertake a principled study of efficient quantization methods for linear classification. We propose and analyze a customized quantization scheme for diagonal linear discriminant analysis classifier including both learning and prediction steps. Experiments on synthetic and real dataset show the effectiveness of our proposed strategies.

1. Introduction

In this work, we investigate the problem of doing centralized prediction using quantized data obtained from distributed sensors. As an example, in Brain-Computer Interface (BCI) applications, hundreds or even thousands of electrodes placed around or inside the head are used to sense brain signals (Lebedev & Nicolelis, 2006). These quantized signals are then used for a specific prediction task such as classification. For example, a neuroprosthetic goal might involve predicting whether an individual is trying to move his hand towards left or right purely based on the quantized brain data to decode the patient's desired movement. Other applications include wireless sensors networks for the Internet of Things (Zhou et al., 2013) and electric power grid (Nabaee & Labeau, 2012). In these settings, sensors need to communicate data at high rates, and consequently consume large amounts of energy (Won et al., 2014). To avoid large energy consumption, data is quantized for both training and prediction. One key observation for prediction task is that all features (readings of different sensors) do not have the same relevance to the prediction goal. Thus, if we compress each feature in accordance to its relevance, we can reduce communication cost and keep prediction error low simultaneously. Formally, given communication constraints (or equivalently, energy constraints), our aim is to devise a data quantization technique that supports our prespecified task.

Traditional information-theoretic quantization techniques as those proposed by (Berger, 1979; Slepian & Wolf, 1973; Cover, 1975; Wyner & Ziv, 1976) are difficult to apply to these problems because these methods require either moving *unquantized* data to a central node prior to compression, which is not applicable in aforementioned settings, or storing and estimating parameters at each sensor, which needs complex hardware that already consumes high energy. Recently, Mahzoon et al. (Mahzoon et al., 2014) proposed rate allocation and deterministic quantization strategies for quantizing signals from m sensors and then directly used these quantized data for linear regression and linear classification, but their method needs already trained model.

In this work, we propose a two-stage active quanitization strategy for training Diagonal Linear Discriminant Analysis (DLDA) classifier. We first use initial codes based on our prior knowledge about the underlying distribution. Then after the first round sampling, we change our codes based on these data and sample again. Our final estimation of parameters of the DLDA classifier is based on the second round quantized data. Once the classifier is trained, we use a randomized dithering-noise based quantization for the testing data on which prediction is desired. Finally, Theorem 2 in Sec. 3 reveals how the number of training samples and total bits used for quantization affect the prediction accuracy. To the best of our knowledge, our pro-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

posed strategy is the first one that quantizes features in both learning and prediction steps with provable bounds. Experiments on simulated and real data demonstrate the effectiveness of our method.

1.1. Related Works

The study of quantization starts from traditional information theory, where one needs to estimate the joint distribution across all the sensors (Wyner & Ziv, 1976; Cover, 1975; Slepian & Wolf, 1973; Berger, 1979), which is hard to realize (Mahzoon et al., 2014). Recently, Zhu et al. (Zhu & Lafferty, 2014) focused on quantized estimation of Gaussian sequence models in Euclidean balls. However, there are significant differences between our work and existing ones: we search for the optimal allocated bits for the *quantized data*, rather than for quantizing the predictors, and conduct solid theoretical analysis for the behaviour of the quantized data as the input to the linear predictor.

2. Notation and Problem Statement

In the distributed sensor network setting, suppose we have m sensors and a sum rate of R bits that needs to be allocated across different sensors for quantization. We use bold $\mathbf{X}$ to represent a sample and X_i is the feature from i-th sensor. If there are n samples, we denote these samples by $\{\mathbf{X}(j)\}_{j=1}^n$. R_i is the number of bits we assign to i-th sensor. Thus, $\sum_{i=1}^m R_i \leq R$.

For each feature X_i, its quantized representation using R_i bits is denoted by $\widehat{X}_i$. More precisely, the i-th sensor uses an encoder function $\mathcal{E}_i : \mathbb{R} \to \{0, \cdots, 2^{R_i} - 1\}$ and sends $\mathcal{E}_i(X_i)$ to the fusion center. We assume that the communication channel is noiseless. Also, we do not use vector quantization since despite the simplicity afforded by asymptotic vector quantization analysis, we use scalar quantization strategies because we aim for designing techniques that are applicable to sensors with very small memory. The decision center uses a corresponding decoding function $\mathcal{D}_i : \{0, 1, \cdots, 2^{R_i} - 1\} \to \mathbb{R}$. Since both training and prediction are done at the fusion center, we can only use quantized data for both tasks. Our goal is to minimize prediction error.

In this paper, we consider the problem of training a linear classifier from quantized samples then doing prediction based on quantized features. We focus on a simple but widely used linear classifier, Diagonal Linear Discriminant Analysis (DLDA). DLDA is a classical classification method for continuous valued features and has been widely used in various domains (Venables & Ripley, 2013). We assume each sample $\mathbf{X}$ belongs to one of the two existing classes with equal probability, i.e., $\mathbf{Pr}[class(\mathbf{X}) = 1] = \mathbf{Pr}[class(\mathbf{X}) = 2] = 1/2$. DLDA makes the assumptions that given the class, $class(\mathbf{X}) = c$, each feature is distributed independently according to a Gaussian distribution: $X_i \sim \mathcal{N}(\mu_{ic}, \sigma_{ic}^2)$, and $\sigma_{i1} = \sigma_{i2} = \sigma_i$ for $i = 1, \cdots, m$. Without loss of generality, we also assume $-\mu_{i1} = \mu_{i2} = \mu_i$. Under these assumptions, DLDA is a linear classifier with $w_i = \mu_i/\sigma_i^2$, for $i = 1, \cdots, m$. Under quantization constraints, we can only use the quantized observations for both training and prediction. Here we want to design a training algorithm together with quantization strategies that minimize the classification error: $\mathbf{Pr}\left[\widehat{C}\left(\widehat{\mathbf{X}}\right) \neq class(\mathbf{X})\right]$, where $\widehat{C}(\cdot)$, denotes the linear classifier trained by quantized samples. While the problem of finding the optimal strategies that minimize classification error is hard, we instead relax the problem to estimating the decision variable $\mathbf{w}^\top \mathbf{X}$, and use it to obtain upper bounds on classification error.

3. Active Learning for Quantized DLDA

3.1. Quantized Training for DLDA

For DLDA, in the training phase, we need to estimate $\{\mu_i\}_{i=1}^m$ and $\{\sigma_i\}_{i=1}^m$ using quantized features and labeled data. Notice that since the two classes have symmetric means around $\mathbf{0}$ and same variance, whenever we have a sample with label 1, we can negate it and obtain a sample from class 2. Thus, equivalently, in the training phase, we are just estimating parameters of a Gaussian distribution.

Our technique has two rounds. In the first round, we use our prior knowledge about the underlying parameters to construct initial quantizers. Then we use these quantized observations to obtain a rough estimate of the underlying distributions. Based on estimated parameters from the first round, we construct new codes to quantize data from the next round. Finally, we use the quantized samples from the second round to learn parameters of underlying distribution and weight vector for DLDA classification.

Formally, we assign R_i^{init} bits to the i-th sensor and use the following code in the first round:

$$\begin{aligned} &\mathcal{E}_i^{init}(X_i) \qquad (1)\\ &= \underset{k=0,\cdots,2^{R_i^{init}}-1}{\arg\min} \left|-\mu_i^{init} - c_i^{init}\sigma_i^{init} + k d_i^{init} - X_i\right| \end{aligned}$$

where μ_i^{init} and σ_i^{init} are our initial guess on mean and variance. $c_i^{init} = 2\max\left(\log\left(\sigma_i^{init}/\mu_i^{init}\right) R_i^{init}, 1\right)$ controls the range of quantization region and $d_i^{init} = 2(\mu_i^{init} + c_i^{init}\sigma_i^{init})/(2^{R_i^{init}} - 1)$ is the quantization unit. The corresponding decoder is, for $k = 0, \cdots, 2^{R_i^{init}} - 1$

$$\mathcal{D}_i^{init}(k) = -\mu_i^{init} - c_i^{init}\sigma_i^{init} + k d_i^{init}. \qquad (2)$$

Let n_1 be the number of samples in the first round. We estimate mean and variance by

$$\tilde{\mu}_i = \frac{\sum_{j=1}^{n_1} \widehat{X}_i(j)}{n_1}, \qquad \tilde{\sigma}_i^2 = \frac{\sum_{j=1}^{n_1}\left(\widehat{X}_i(j) - \tilde{\mu}_i\right)^2}{n_1}.$$

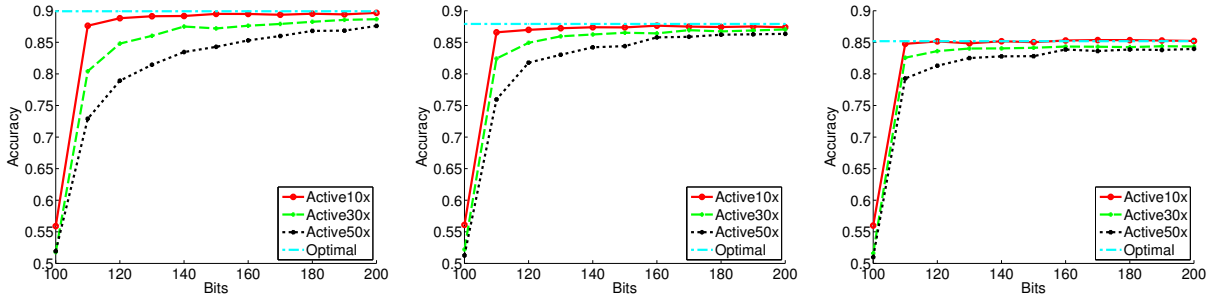

Figure 1. Classification accuracy of proposed quantization scheme on synthesized data. **Optimal** is the optimal Bayes classification rule applied to unquantized samples.

where $\widehat{X}_i(j) = \mathcal{D}_i^{init}\left(\mathcal{E}_i^{init}\left(X_i\left(j\right)\right)\right)$ is the quantized representation of $X_i(j)$.

In the second round, we assign $\widetilde{R}_i$ bits to the i-th sensor and we sample another set of data points using uniform quantization scheme informed by the first round estimation on mean and variance:

$$\widetilde{\mathcal{E}}_i\left(X_i\right) = \underset{k=0,\cdots,2^{\widetilde{R}_i}-1}{\arg\min} \left|-\tilde{\mu}_i - \tilde{c}_i\tilde{\sigma}_i + k\tilde{d}_i - X_i\right|, \quad (3)$$

$$\widetilde{\mathcal{D}}_i\left(k\right) = -\tilde{\mu}_i - \tilde{c}_i\tilde{\sigma}_i + k\tilde{d}_i. \quad (4)$$

where $\tilde{c}_i = 2\log\left(\frac{m}{\epsilon}\right)\max\left(\log\frac{\tilde{\sigma}_i}{\tilde{\mu}_i}, 1\right)$ and $\tilde{d}_i = 2\left(\tilde{\mu}_i + \tilde{c}_i\tilde{\sigma}_i\right)/(2^{\widetilde{R}_i}-1)$. Let n_2 be the number of observations from the second round we use to estimate the mean, the variance and the weight vector for DLDA:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_2}\widehat{X}_i(j)}{n_2}, \hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n_2}\left(\widehat{X}_i(j) - \hat{\mu}_i\right)^2}{n_2}, \hat{w}_i = \frac{\hat{\mu}_i}{\hat{\sigma}_i^2},$$

where $\widehat{X}_i(j) = \widetilde{\mathcal{D}}_i\left(\widetilde{\mathcal{E}}_i\left(X_i\left(j\right)\right)\right)$ is the quantized representation of $X_i(j)$ in the second round.

3.2. Quantized Prediction for DLDA

In the previous section, we have good estimations on underlying distribution of features ($\{\hat{\mu}_i\}_{i=1}^m$ and $\{\hat{\sigma}_i\}_{i=1}^m$) and weight vector $\widehat{\mathbf{w}}$ for DLDA. In this section, we discuss how to use these estimations for prediction. First, we assign bits to each sensor according to (9).

As the first step of our quantization, we pick b_i for each sensor such that $|X_i| \le b_i$ holds with high probability. Then for the i-th sensor, we place 2^{R_i} quantization points uniformly in the region $[-b_i, b_i]$, i.e., the quantization points are $\{-b_i + kd_i | k = 0, \cdots, 2^{R_i} - 1\}$ where $d_i = 2b_i/(2^{R_i}-1)$ is a unit quantization region. For the feature from the i-th sensor, X_i, we first add dithering noise γ_i uniformly distributed within $[-d_i/2, d_i/2]$, then we map this value to the nearest quantization point. Formally, our

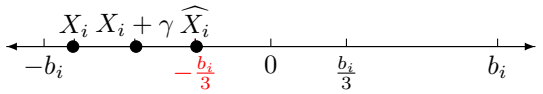

Figure 2. An illustration of dithering based quantization strategy. We use $R_i = 2$ bits for quantizing X_i, so $d_i = 2b_i/(2^{2^{R_i}-1}) = b_i/3$. In this scenario, feature X_i is quantized to $-\frac{1}{3}b_i$ because after adding dithering noise, the nearest quantization point is $-\frac{1}{3}b_i$.

encoding and decoding functions are

$$\mathcal{E}_i(x) = \underset{k\in\{0,\cdots,2^{R_i}-1\}}{\arg\min} |-b_i + kd_i - x - \gamma|, \quad (5)$$

$$\mathcal{D}_i(k) = -b_i + kd_i. \quad (6)$$

Fig. 2 provides an example of such a quantization strategy. By adding dithering noise, we now show that the correlation between quantization error from different sensors is removed (consistent with (Schuchman, 1964)). Specifically, with we derived the following result:

Theorem 1. *Suppose for $i = 1, \cdots, m$, $|X_i| \le b_i$, with dithering noise quantization strategy, we have*

$$\mathbf{E}\left[\left(\mathbf{w}^\top\mathbf{X} - \mathbf{w}^\top\widehat{\mathbf{X}}\right)^2\right] \le 4\sum_{i=1}^m w_i^2 b_i^2 \cdot 2^{-2R_i}. \quad (7)$$

Now we can optimize bits assignment for the test data to minimize Eqn. (7):

$$\min_{R_i, i=1,\cdots,m} \sum_{i=1}^m w_i^2 b_i^2 \cdot 2^{-2R_i} \quad (8)$$

$$\text{s.t} \quad \sum_{i=1}^m R_i = R, R_i \ge 1 \text{ for } i = 1, \cdots, m$$

Routine algebra shows that the optimal bits assignment for the i-th sensor is:

$$R_i = \left[\frac{1}{2}\log\frac{8\ln 2 \cdot w_i^2 b_i^2}{\lambda} - 1\right]_+ + 1, \quad (9)$$

where $[x]_+ = \max\{x, 0\}$ and λ is selected such that $\sum R_i = R$. The rates of each source are then rounded to the nearest integer to ensure feasibility of quantization.

The next theorem reveals how the number of training samples and the number of bits for quantization affect the prediction accuracy:

Theorem 2. *Assume for* $i = 1, \cdots, m$, $\mu_i \leq \mu_i^{init}$, $\sigma_i \leq \sigma_i^{init}$, *in the first stage* [1] $R_i^{init} = \widetilde{\Omega}\left(\log\left(\frac{\mu_i^{init}}{\mu_i} + \frac{\sigma_i^{init}}{\mu_i}\right)\right)$, *and* [2] $n_1 = \widetilde{\Omega}\left(\log\left(\frac{m}{\delta}\right)\left[\left(\frac{\mu_i^{init}}{\mu_i}\right)^2 + \left(\frac{\sigma_i^{init}}{\mu_i}\right)^2 + \left(\frac{\mu_i^{init}}{\sigma_i}\right)^4 + \left(\frac{\sigma_i^{init}}{\sigma_i}\right)^4\right]\right)$ *and in the second stage* $\widetilde{R}_i = \widetilde{\Omega}\left(\log\left(\frac{1}{\epsilon}\left(\frac{\mu_i}{\sigma_i} + \frac{\sigma_i}{\mu_i}\right)\right)\right)$, $n_2 = \widetilde{\Omega}\left(\frac{1}{\epsilon^2}\log^2\left(\frac{m}{\epsilon}\right)\log\left(\frac{m}{\delta}\right)\left(\frac{\mu_i^4}{\sigma_i^4} + \frac{\sigma_i^4}{\mu_i^4}\right)\right)$ *then with probability at least* $1 - \delta$, *for all* $i = 1, \cdots, m$,

$$\mathbf{Pr}\left(\widehat{C}\left(\widehat{\mathbf{X}}\right) \neq class\left(\mathbf{X}\right)\right) = opt + O\left(\epsilon\right),$$

where opt denotes the classification error of the best possible classifier.

Theorem 2 shows the prediction error comes from two sources: one from quantization, the other from the statistical inference. For a given target accuracy parameter ϵ, the number of bits required for each sensor $\tilde{R}_i$ depends logarithmically on $1/\epsilon$. Therefor, totally we need $O\left(m \log\left(1/\epsilon\right)\right)$ bits to make error induced by quantization be at the order of ϵ. The number of samples required depends quadratically on $1/\epsilon$ up to logarithm factor. Thus, if we have infinite bits (no quantization error), we recover the same sample complexity for parametric model for inference and prediction (Wasserman, 2013).

4. Experiments

4.1. Simulated Data

We first test our quantization strategies on synthesized data. Data is generated according to DLDA assumptions: for $i = 1, \cdots, m$, μ_i is set to 1 and σ_i is set to be i, $i^{1.2}$ and i^2 respectively for left, middle and right plots of Fig. 1. We use $m = 100$ sensors and number of total bits R varies from 100 to 200. We use $n_1 = 1000$ samples in the first round and $n_2 = 10000$ samples in the second round for training and 10000 samples for testing. The initial guesses of parameters are set to be 10 to 50 times of the true values. Fig. 1 shows that the more accurate initial guesses are, the fewer bits needed to achieve certain classification accuracy. Also notice that if signal-to-noise ratio (μ_i/σ_i) of some sensors are much larger than that of others, we need fewer bits to reach optimal classification accuracy.

[1] We omit $\log(\log(\cdot))$ terms.

[2] We omit $\log(\cdot)$ dependencies on μ_i and σ_i.

4.2. Real Data

In this section, we test our quantization scheme on EEG data. We use the brain signals of the first subject in experiment of data set 1 from BCI Competition IV (Blankertz et al., 2007). In the experiment, there are total 200 trials. Each trial corresponds to a motor imagery of either *left hand* or *foot* and lasts 8s. There total $m = 59$ sensors and signals were sampled at 100Hz. See (Blankertz et al., 2007) for the details. For each trial, raw EEG time series are band-pass filtered with a butterworth IIR filter of order 5. Then variance is calculated for each channel (band power) and the logaithm is applied to the normalized variance to yield a feature vector for that trial. Thus, we generate 200 instances each with 59 features. Then we randomly select 40 samples for testing and the remaining for training. For training, 40 samples are used in the first round and 120 samples are used in the second round. We use 10 times of true mean and variance of training samples as initial guesses.

Fig. 3 shows the classification accuracy on testing samples with different bits used. The unquantized classifier is trained directly using 160 training samples without quantization and then is applied to unquantized testing samples. Notice that even with just an average of 3 bits per sensor, full (infinite number of bits) quantization accuracy can be achieved. Another observation is that as we increase total bits, the result becomes more stable.

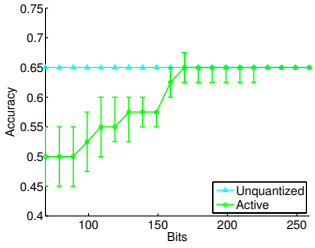

Figure 3. Classification accuracy of proposed quantization scheme on EEG data.

5. Conclusion

In this paper, we propose and analyze an active learning based quantization algorithm together with a prediction algorithm that only require quantized samples for diagonal linear discriminant analysis. Experiments on synthetic and real world data show that with a few bits, we can achieve near optimal accuracy as using un-quantized samples. In this work, we only consider DLDA classifier. How to efficiently assign bits among sensors and quantize features for nonlinear classifiers is an important problem that has both theoretical and practical implications.

References

Berger, T. Decentralized estimation and decision theory. In *IEEE Seven Springs Workshop on Information Theory, Mt. Kisco, NY*, 1979.

Blankertz, Benjamin, Dornhege, Guido, Krauledat, Matthias, Müller, Klaus-Robert, and Curio, Gabriel. The non-invasive berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

Cover, Thomas M. A proof of the data compression theorem of slepian and wolf for ergodic sources (corresp.). *Information Theory, IEEE Transactions on*, 21(2):226–228, 1975.

Gamburd, Alex, Lafferty, John, and Rockmore, Dan. Eigenvalue spacings for quantized cat maps. *Journal of Physics A: Mathematical and General*, 36(12):3487, 2003.

Lebedev, Mikhail A and Nicolelis, Miguel AL. Brain–machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546, 2006.

Mahzoon, Majid, Albalawi, Hassan, Li, Xin, and Grover, Pulkit. Using relative-relevance of data pieces for efficient communication, with an application to neural data acquisition. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pp. 160–166. IEEE, 2014.

Nabaee, Mahdy and Labeau, Fabrice. Quantized network coding for sparse messages. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pp. 828–831. IEEE, 2012.

Schuchman, Leonard. Dither signals and their effect on quantization noise. *Communication Technology, IEEE Transactions on*, 12(4):162–165, 1964.

Slepian, David and Wolf, Jack K. Noiseless coding of correlated information sources. *Information theory, IEEE Transactions on*, 19(4):471–480, 1973.

Venables, William N and Ripley, Brian D. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

Wasserman, Larry. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Won, Minho, Albalawi, Hassan, Li, Xin, and Thomas, Donald E. Low-power hardware implementation of movement decoding for brain computer interface with reduced-resolution discrete cosine transform. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pp. 1626–1629. IEEE, 2014.

Wyner, Aaron D and Ziv, Jacob. The rate-distortion function for source coding with side information at the decoder. *Information Theory, IEEE Transactions on*, 22 (1):1–10, 1976.

Zhou, Yang, Huang, Chuan, Jiang, Tao, and Cui, Shuguang. Wireless sensor networks and the internet of things: Optimal estimation with nonuniform quantization and bandwidth allocation. *Sensors Journal, IEEE*, 13(10):3568–3574, 2013.

Zhu, Yuancheng and Lafferty, John. Quantized estimation of gaussian sequence models in euclidean balls. In *Advances in Neural Information Processing Systems*, pp. 3662–3670, 2014.