LINEAR QUANTIZATION BY EFFECTIVE-RESISTANCE SAMPLING

Yining Wang and *Aarti Singh*

Machine Learning Department, Carnegie Mellon University, Pittsburgh PA 15213, USA

ABSTRACT

In this paper we consider the problem of allocation of measurement bits in order to reduce the statistical signal recovery error resulting from quantization error. We propose a continuous optimization problem that serves as a relaxation of the original combinatorial problem, which is amenable to classical continuous optimization solvers such as the gradient descent. We also design a "rounding" algorithm based on the idea of effective resistance sampling to turn the continuous fractional solution into a feasible bit allocation strategy with integer number of bits allocated to each design point.

Index Terms— Quantization, linear models, spectral sparsification, effective-resistance sampling

1. INTRODUCTION

Consider the noiseless linear signal model

$$y = X\beta_0 \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is an exactly known design matrix, typically generated from certain physical procedures, and $\beta_0 \in \mathbb{R}^p$ is an unknown p-dimensional signal to be recovered. We restrict ourselves to the "low-dimensional" setting $p < n$. Unlike the classical linear regression model ubiquitous in the statistics literature, the model in Eq. (1) is assumed to be *noiseless* as no noise variables are included in the measurement model $y = X\beta_0$. Such a model arises in various scenarios where the signal can be expressed or well-approximated by a small number of basis elements. We mention one specific example from the framework of signal processing on graphs [1, 2], which studies signals with an underlying complex structure that is modeled by a graph such as measurements at nodes of a network. The band-limited model for graph signals is a linear model in which the network node measurements y are well represented by a linear model where the features are the eigenvectors of the graph Laplacian or adjacency matrix corresponding to the smallest/largest eigenvalues, respectively.

The measurements of y, however, can only be made up to a total of k binary bits and hence cannot be perfectly accurate. Such measurement-constrained settings are ubiquitous in statistical signal processing and machine learning applications, such as brain signal sensing [3], Internet of Things [4] and electric power grids [5]. It is therefore important to design intelligent bit allocation algorithms such that the recovery of signal β_0 is the most accurate possible subject to given bit measurement constraints.

We formulate the bit allocation problem that will be studied in this paper as follows:

Problem 1 (passive bit allocation)**.** *Given exactly measured design $X \in \mathbb{R}^{n\times p}$ and a bit budget $k \in \mathbb{N}$, $k \geq n$, find a* bit allocation $\boldsymbol{k} = (k_1, \cdots, k_n) \in \mathbb{N}^n$, $k_1 + \cdots + k_n \leq k$ *such that the mean square error between the recovered signal $\widehat{\beta}_{\boldsymbol{k}}$ and the true signal β_0 is mimimized.*

To attack Problem 1, we first derive a *continuous relaxation* to the originally combinatorial optimization problem that is difficult to solve. Though the relaxed optimization problem has a non-convex objective and hence global optimization might be challenging, local convergence of first-order methods such as gradient descent can be expected in practice. We then borrow the idea of *effective resistance sampling* from the graph sparsification literature [6] to "round" the solution of the continuous relaxation problem.

1.1. Related work

In statistics, the question of selecting a subset of important design points so as to maximize statistical efficiency in a regression model is referred to as *experimental/optimal design* and has a long history of research [7]. Indeed, our strategy was based on existing work on computationally tractable experimental design, which involves intellectually rounding (sparsifying) a solution of certain continuously relaxed optimization problems [8, 9, 10]. One important difference, as we also remark in the final section of this paper, is the non-linearity of the budget constraint, which makes the continuously relaxed optimization problem non-convex and the subsequent effective resistance sampling algorithm difficult to analyze.

Our rounding strategy is based on the seminal work of [6] which developed the *effective resistance sampling* algorithm for the graph sparsification problem. Similar sampling strategies were also considered, under the alternative name of *leverage score sampling*, for linear regression [11, 12, 13] and matrix column/row selection problems [14].

[15] also considered the problem of intelligent quantization for learning problems. However, the setting in [15] is the

design matrix X being imperfectly measured, which differs from our setting where X is exactly known and the response y can only be measured with non-negligible quantization error.

The bit allocation problem has also been well-studied in the signal processing society as resource management research, e.g., in [16].

2. METHOD AND ANALYSIS

We present the main bit allocation algorithm and its associated analysis. In Sec. 2.1, we show how the mean square error of the recovered signal $\hat{\beta}_{\boldsymbol{k}}$ behaves when the bit allocation strategy $\boldsymbol{k} = (k_1, \cdots, k_n)$ is given. We then derive a continuous relaxation of Problem 1 in Sec. 2.2, whose solution is then rounded by an effective resistance algorithm in Sec. 2.3. Finally, we give a theorem establishing that when the bit budget k is not too small, the rounded solution is within a $(1+\varepsilon)$-relative approximation error compared to the solution of the continuous relaxed problem.

2.1. Weighted OLS and its mean square error

Suppose a bit allocation strategy $\boldsymbol{k} = (k_1, \cdots, k_n) \in \mathbb{N}_+^n$ is given, such that $\sum_{i=1}^n k_i \le n$. Let $\mathrm{round}(\cdot)$ be the rounding operator towards the closest integer and $U[a,b]$ be the uniform distribution on interval $[a,b]$. The observed quantized value of $y_i = x_i^\top \beta_0$ with $k_i \in \mathbb{N}_+$ binary bits of measurement can then be expressed as

$$\widetilde{y}_i = 2^{-(k_i-1)} \cdot \mathrm{round}\left[2^{k_i-1}\left(\frac{y_i}{M} + \delta_i\right)\right] \tag{2}$$

where $M := \max_{1\le i\le n} |x_i^\top \beta_0|$ is a known bounded constant and $\delta \sim U[-2^{-k_i}M, 2^{-k_i}M]$ is a *dithering* variable that introduces additional stochasticity to the deterministic model (1). The dithering step de-couples the statistical dependency in the quantized error and is an important concept in the signal processing literature [17]. Note also that the most significant bit in $\widetilde{y}_i$ indicates the sign of y_i, and hence only $(k_i - 1)$ bits are available to measure the absolute value of y_i.

As the number of measure bits differ for different design points x_i, the rounding (quantization) error of each $\tilde{y}_i$ also differs, making the quantized linear model (2) similar to a linear regression model with heteroscedastic noise. Because the noise levels are known (controlled by the bit allocation strategy $\boldsymbol{k}$ directly), a *weighted* Ordinary Least Squares (OLS) estimator is reasonable for the recovery of β_0 which we define as follows:

$$\widehat{\beta}_{\boldsymbol{k}} \in \mathrm{argmin}_{\beta\in\mathbb{R}^p} \sum_{i:k_i>0} 4^{k_i+1}(\tilde{y}_i - x_i^\top \beta)^2. \tag{3}$$

The following Proposition upper bounds the mean square error of $\hat{\beta}_{\boldsymbol{k}}$. Its proof is a standard analysis of weighted OLS estimators for heteroscedastic linear models.

Proposition 1. *The weighted OLS estimator $\widehat{\beta}_{\boldsymbol{k}}$ satisfies*

$$\mathbb{E}\|\widehat{\beta}_{\boldsymbol{k}} - \beta_0\|_2^2 \le M^2 \cdot \mathrm{tr}\left[\left(\sum_{i:k_i>0} 4^{k_i+1} x_i x_i^\top\right)^{-1}\right]. \tag{4}$$

Proof. Without loss of generality assume $k_i > 0$ for all i, because for those design points with $k_i = 0$ no information is gained and therefore these points can be excluded from the analysis. Let $w_i = 4^{k_i+1}$ be the weight of design point x_i and define $W := \mathrm{diag}(w_1, \cdots, w_n)$. The weighted OLS estimator $\widehat{\beta}_{\boldsymbol{k}}$ then admits a closed-form expression

$$\widehat{\beta}_{\boldsymbol{k}} = (X^\top W X)^{-1} X^\top W \widetilde{y}, \tag{5}$$

where $\widetilde{y} = (\widetilde{y}_1, \cdots, \widetilde{y}_n) \in \mathbb{R}^n$. Define $\varepsilon := \widetilde{y} - y$. Using the linear model that $y = X\beta_0$, we have $\widehat{\beta}_{\boldsymbol{k}} - \beta_0 = (X^\top W X)^{-1} X^\top W \varepsilon$. On the other hand, by the quantized error model Eq. (2), it holds that $\mathbb{E}[\varepsilon_i | X] = 0$ and $\mathbb{E}[\varepsilon_i^2 | X] \le 4^{-(k_i+1)} M^2 = w_i^{-1} M^2$. Subsequently,

$$\begin{aligned}
&\mathbb{E}\|\widehat{\beta}_{\boldsymbol{k}} - \beta_0\|_2^2 \\
&= \mathrm{tr}\left[(X^\top W X)^{-1} X^\top W \mathbb{E}(\varepsilon\varepsilon^\top) W^\top X (X^\top W X)^{-1}\right] \\
&\le M^2 \cdot \mathrm{tr}\left[(X^\top W X)^{-1} X^\top W X (X^\top W X)^{-1}\right] \\
&= M^2 \cdot \mathrm{tr}\left[(X^\top W X)^{-1}\right].
\end{aligned}$$

□

For simplicity we define $F(\boldsymbol{k}; X) := \mathrm{tr}[\sum_{i=1}^n 4^{k_i+1} x_i x_i^\top]^{-1}$. The bit allocation problem (Problem 1) is then reduced to the combinatorial optimization problem of finding $\boldsymbol{k} = (k_1, \cdots, k_n) \in \mathbb{N}_+^n$, $\sum_{i=1}^n k_i \le k$ such that $F(\boldsymbol{k}; X)$ is minimized.

2.2. The continuous relaxation

We introduce a continuous relaxation of the combinatorial optimization problem mentioned in the previous section, which is relatively easier to optimize using conventional continuous optimization methods such as the gradient descent.

$$\begin{aligned}
\min_{\boldsymbol{\pi}=(\pi_1,\cdots,\pi_n)\in\mathbb{R}^n} \ & \mathrm{tr}\left[\left(\sum_{i=1}^n (4^{\pi_i} - 1) x_i x_i^\top\right)^{-1}\right] \\
s.t. \quad & \pi_i \ge 0, \ \ \|\boldsymbol{\pi}\|_1 \le k_0.
\end{aligned} \tag{6}$$

Let $\boldsymbol{\pi}^*$ denote the optimal solution to Eq. (6) and define $F(\boldsymbol{\pi}; X) := \mathrm{tr}[\sum_{i=1}^n 4^{\pi_i+1} x_i x_i^\top]^{-1}$. It is straightforward that Eq. (6) is a strict relaxation of the original combinatorial optimization problem, because any integral bit allocation strategy $\boldsymbol{k}$ is automatically feasible to Eq. (6). Formally, we have the following proposition:

input : $X \in \mathbb{R}^{n\times p}$, quantization budget k, support size $s < k - p$, number of repetitions B.
output: $\widehat{\boldsymbol{k}} \in \mathbb{N}^n$ satisfying $k_1 + \cdots + k_n \le k$.
1. Continuous optimization: solve for $\boldsymbol{\pi}^*$, the optimal solution of Eq. (6).
2. Pre-conditioning: $\Sigma_* = \sum_{i=1}^n 4^{\pi_i^*+1} x_i x_i^\top$; leverage scores $\ell_i = x_i^\top \Sigma_*^{-1} x_i$.
3. **for** $b \in \{1, \cdots, B\}$ **do**
   3.1. Initialization: $\{w_i^{(b)}\}_{i=1}^n = 0$.
   3.2. Repeat for s times: sample $i_t \in [n]$ from the categorical distribution $\Pr[i_t = i] = p_i \propto 4^{\pi_i^*+1}\ell_i$ and update $w_{i_t}^{(b)} \leftarrow w_{i_t}^{(b)} + 4^{\pi_i^*+1}/p_{i_t}$.
   3.3. Define allocation $\widehat{\boldsymbol{k}}^{(b)}$: $\widehat{k}_i^{(b)} = \left\lceil \frac{(k-s)\log(1+w_i^{(b)})}{\sum_{j:w_j^{(b)}>0}\log(1+w_j^{(b)})} \right\rceil$ and $\widehat{k}_i^{(b)} = 0$ if $w_i^{(b)} = 0$.
end
4. Output $\widehat{\boldsymbol{k}}$ in $\{\widehat{\boldsymbol{k}}^{(b)}\}_{b=1}^B$ with the smallest objective $F(\widehat{\boldsymbol{k}}; X)$.

Algorithm 1: Bit allocation algorithm by leverage score sampling

Proposition 2. $F(\boldsymbol{\pi}^*; X) \le \min_{\boldsymbol{k}\in\mathbb{N}^n, \sum_{i=1}^n k_i \le k} F(\boldsymbol{k}; X)$.

The continuous formulation in Eq. (6) is non-convex and is therefore challenging to achieve global optimality efficiently. Nevertheless, in practice alternative approximate optimization methods can be applied. One possibility is to perform (projected) gradient descent on Eq. (6) with multiple initializations. Additionally, one may introduce auxiliary variables $y_i = 4^\pi - 1$ and re-formulate Eq. (6) as a DC-programming problem. Existing packages on DC programming [18] can then be invoked to approximately optimize Eq. (6).

2.3. Rounding by effective resistance sampling

The solution $\boldsymbol{\pi}^*$ to the continuous optimization problem in Eq. (6) is fractional and thus cannot be directly used as a bit allocation strategy. In this section, we present a rounding algorithm based on the celebrated *effective resistance sampling* method [6].

Algorithm 1 gives a pseudo-code description of our proposed rounding algorithm. At a higher level, the algorithm "samples" each data point x_i with replacement with a probability that is proportional to both the weight in the continuous optimization problem π_i^* and the *leverage score* (also known as *effective resistance* in a graph sparsification setting [6]) of x_i with respect to $\Sigma_* := \sum_{i=1}^n 4^{\pi_i^*+1} x_i x_i^\top$.

The following proposition shows that the output strategy $\widehat{\boldsymbol{k}}$ is always feasible.

Proposition 3. *With probability 1 it holds that* $\widehat{\boldsymbol{k}} \in \mathbb{N}_+^n$ *and* $\sum_{i=1}^n \widehat{k}_i \le k$.

Proof. $\widehat{\boldsymbol{k}} \in \mathbb{N}^n$ clearly holds. On the other hand, because $\widehat{k}$ has at most s elements that are not zero, we have

$$\begin{aligned}\sum_{i=1}^n \widehat{k}_i &\le \sum_{\widehat{k}_i>0} 1 + \frac{(k-s)\log(1+w_i^{(b)})}{\sum_{j:w_j^{(b)}>0}\log(1+w_j^{(b)})} \\ &\le s + k - s = k.\end{aligned}$$

□

The effective resistance sampling method is known to enjoy superior theoretical guarantees in terms of approximation of the continuous solutions spectrally [6]. Due to space constraints, we omit the formal proof of rounding performance of Algorithm 1. The readers are referred to Secs. 3.2 and 3.3 of [9] which analyzed the effective sampling method for a similar experiment selection (with replacement) problem, whose proofs can be easily adapted to show $(1+\varepsilon)$ approximation guarantee of Algorithm 1.

3. CONCLUDING REMARKS

There are many open questions along this direction of research. Below we mention three problems that we think are the most interesting.

1. The continuous relaxed optimization problem in Eq. (6) has a non-convex objective; therefore, it is challenging to obtain a global optimal solution using computationally tractable algorithms. It is an interesting question whether approximate optimality can be achieved, or whether convergence to local optima can be guaranteed for first-order methods such as the gradient descent;
2. Unlike the experimental design problem for linear regression [9, 10], the budget constraint for the bit allocation problem is not linear in the weights of the rescaled design points. Thus, the (expected) total weight

$\sum_{i=1}^{n} w_i^{(b)}$ in Algorithm 1 does not have an easy upper bound, making it very challenging to bound the performance gap between the continuous solution $\boldsymbol{\pi}^*$ and its rounded version $\widehat{\boldsymbol{k}}$. Currently we do not know an easy fix to this problem, and potentially rounding/sparsification algorithms other than the effective resistance sampling should be considered;

3. The problem we considered in this paper (Problem 1) is *passive*, in the sense that the bit allocation strategy $\boldsymbol{k}$ is determined only based on knowledge of X. In practice, however, it is usually feasible to measure the signal in a feedback-driven manner, where previous (partial) measurements of $\widetilde{y}$ might provide guidance in the allocation of bits for later measurements. We refer this problem as the *active* bit allocation problem, and conjecture that improvements can be made by taking previous measurements of $\widetilde{y}$ into consideration.

Acknowledgement

We thank Adams Wei Yu, Yu-Xiang Wang, Luiz Chamon and two anonymous referees for their helpful comments. This research is supported in part by NSF grants CCF-1563918 and CAREER IIS-1252412.

4. REFERENCES

[1] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[2] Aliaksei Sandryhaila and Jose MF Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 80–90, 2014.

[3] Mikhail A Lebedev and Miguel AL Nicolelis, "Brain–machine interfaces: past, present and future," *TRENDS in Neurosciences*, vol. 29, no. 9, pp. 536–546, 2006.

[4] Yang Zhou, Chuan Huang, Tao Jiang, and Shuguang Cui, "Wireless sensor networks and the internet of things: Optimal estimation with nonuniform quantization and bandwidth allocation," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3568–3574, 2013.

[5] Mahdy Nabaee and Fabrice Labeau, "Quantized network coding for sparse messages," in *Proceedings of the IEEE Statistical Signal Processing Workshop (SSP)*, 2012.

[6] Daniel A Spielman and Nikhil Srivastava, "Graph sparsification by effective resistances," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1913–1926, 2011.

[7] Friedrich Pukelsheim, *Optimal design of experiments*, vol. 50, SIAM, 1993.

[8] Siddharth Joshi and Stephen Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, 2009.

[9] Yining Wang, Adams Wei Yu, and Aarti Singh, "On computationally tractable selection of experiments in regression models," *Journal of Machine Learning Research*, vol. 18, no. 143, pp. 1–41, 2017.

[10] Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang, "Near-optimal design of experiments via regret minimization," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[11] Ping Ma, Michael W Mahoney, and Bin Yu, "A statistical perspective on algorithmic leveraging," *Journal of Machine Learning Research*, vol. 16, pp. 861–911, 2015.

[12] Ahmed Alaoui and Michael Mahoney, "Fast randomized kernel ridge regression with statistical guarantees," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

[13] Shusen Wang, Alex Gittens, and Michael W Mahoney, "Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[14] Petros Drineas, Michael W Mahoney, and S Muthukrishnan, "Relative-error CUR matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 844–881, 2008.

[15] S Du, Yichong Xu, H Zhang, C Li, P Grover, and A Singh, "Novel quantization strategies for linear prediction with guarantees," in *Proceedings of ICML 2016 Workshop on On-Device Intelligence*, 2016.

[16] Engin Masazade, Ruixin Niu, and Pramod K Varshney, "Dynamic bit allocation for object tracking in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5048–5063, 2012.

[17] Leonard Schuchman, "Dither signals and their effect on quantization noise," *IEEE Transactions on Communication Technology*, vol. 12, no. 4, pp. 162–165, 1964.

[18] Reiner Horst and Nguyen V Thoai, "DC programming: overview," *Journal of Optimization Theory and Applications*, vol. 103, no. 1, pp. 1–43, 1999.