vqSGD: Vector Quantized Stochastic Gradient Descent*

Venkata Gandikota[1], Daniel Kane[2], Raj Kumar Maity[1], and Arya Mazumdar[1]

[1]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01003, USA. E-mail: gandikota.venkata@gmail.com, rajkmaity@cs.umass.edu, arya@cs.umass.edu.

[2]Departments of Mathematics and CSE, University of California, San Diego, CA 92093, USA. E-mail: dakane@ucsd.edu.

February 11, 2020

Abstract

In this work, we present a family of vector quantization schemes *vqSGD* (Vector-Quantized Stochastic Gradient Descent) that provide an asymptotic reduction in the communication cost with convergence guarantees in distributed optimization. In particular, we consider a randomized scheme based on the convex hull of a point set, that returns an unbiased estimator of a d-dimensional gradient vector with bounded variance. We provide multiple efficient instances of our scheme that require only o(d) bits of communication at the expense of a reasonable increase in variance. The instances of our quantization scheme are obtained using the properties of binary error-correcting codes and provide a smooth tradeoff between the communication and the variance of quantization. Furthermore, we show that *vqSGD* also offers strong privacy guarantees.

1 Introduction

Recent surge in the volumes of available data has motivated the development of large-scale distributed learning algorithms. Synchronous Stochastic Gradient Descent (SGD) is one such learning algorithm widely used to train large models. In order to minimize the empirical loss, the SGD algorithm, in every iteration takes a small step in the negative direction of the *stochastic gradient* which is an unbiased estimate of the true gradient of the loss function.

In this work, we consider the data-distributed model of distributed SGD where the data sets are partitioned across various compute nodes. In each iteration of SGD, the compute nodes send their computed local gradients to a parameter server that averages and updates the global parameter. The distributed SGD model is highly scalable, however, with the exploding dimensionality of data and the increasing number of servers (such as in a Federated learning setup Konečný et al. [2016]), communication becomes a *bottleneck* to the efficiency and speed of learning using SGD Chilimbi et al. [2014].

In the recent years various quantization and sparsification techniques Acharya et al. [2019], Alistarh et al. [2017], Bernstein et al. [2018], Koloskova et al. [2019], Shalev-Shwartz et al. [2010], Suresh et al. [2017], Wang et al. [2018], Wen et al. [2017], Mayekar and Tyagi [2019] have been

*This research is supported in parts by NSF awards CCF 1642658, CCF 1618512, CCF 1909046, and by UMass Center for Data Science.

developed to alleviate the problem of communication bottleneck. Recently, Kalan et al. [2019] even showed the effectiveness of gradient quantization techniques for ReLU fitting. The goal of the quantization schemes is to efficiently compute either a low precision or a sparse unbiased estimate of the d-dimensional gradients. One also requires the estimates to have a bounded second moment in order to achieve guaranteed convergence.

Moreover, the data samples used to train the model often contain sensitive information. Hence, preserving privacy of the participating clients is crucial. Differential privacy Dwork et al. [2016, 2014] is a mathematically rigorous and standard notion of privacy considered in both literature and in practice. Informally, it ensures that the information from the released data (e.g. the gradient estimates) cannot be used to distinguish between two *neighboring* data sets.

Our Contribution: In this work, we present a family of *privacy-preserving vector-quantization* schemes that incur low communication costs while providing convergence guarantees. In particular, we propose quantization schemes based on convex hull of specific structured point sets in $\mathbb{R}^d$ that require $O(\log d)$ bits to communicate an unbiased gradient estimate that has bounded variance.

At a high level, our scheme is based on the idea that any vector $\boldsymbol{v} \in \mathbb{R}^d$ with bounded norm can be represented as a convex combination of a carefully constructed point set $C \subset \mathbb{R}^d$. This convex combination essentially allows us to chose a point $\boldsymbol{c} \in C$ with probability proportional to its coefficient, which makes it an unbiased estimator of $\boldsymbol{v}$. The bound on the variance is obtained from the circumradius of the convex hull of C. Moreover, communicating the unbiased estimate is equivalent to communicating the index of $\boldsymbol{c} \in C$ (according to some fixed ordering) that requires only $\log |C|$ bits.

Large convex hulls have small variation in the coefficients of the convex combination of any two points of bounded norm. This observation allows us to obtain ϵ-differential privacy (for any $\epsilon > \epsilon_0$), where ϵ_0 depends on the choice of the point set. We also propose Randomized Response (RR) Warner [1965] and RAPPOR Erlingsson et al. [2014] based mechanisms that can be used over the proposed quantization to achieve ϵ-differential privacy (for any $\epsilon > 0$) with small trade-off in the variance of the estimates.

The family of schemes described above is fairly general and can be instantiated using different structured point sets. The cardinality of the point set bounds the communication cost of the quantization scheme. Whereas, the diameter of the point set dictates the variance bounds and the privacy guarantees of the scheme.

We provide a strong characterization of the point-sets that can be used for our quantization scheme. Using the characterization, we propose construction of point-sets that allow us to attain a smooth trade-off between variance and communication of the quantization scheme. We also propose some explicit structured point sets and show tradeoff in the various parameters guaranteed by them. Our results[1] (summarized in Table 1) are the first quantization schemes in literature to achieve privacy directly through quantization.

Empirically we compare our quantization schemes to the state-of-art schemes Alistarh et al. [2017], Suresh et al. [2017]. We observe that our cross-polytope vqSGD, performs equally well in practice, while providing asymptotic reduction in the communication cost. The communication results are compared in Table 2.

Organization In Sec. 2 we start by describing some related work on communication efficiency in the federated learning setup. Subsequently in Sec. 3 we describe the settings for our results. The vqSGD quantization scheme is presented in Sec. 4. In Sec. 5 we provide a handle to test whether a

[1]Note that ϵ denotes the privacy parameter and ε refers to the packing parameter of ε-nets.

Point set	Error	Communication (bits)	Privacy	Efficiency
Random-Sampling (Theorem 6) for any $c > \log(d)$	$\frac{d}{cN}$	Nc	-	$O(\exp(c))$
Reed-Muller (C_{RM}) (Proposition 7)	$\frac{d}{N}$	$N \log 2d$	-	$O(d)$
Cross-polytope (C_{cp}) (Proposition 8)	$\frac{d}{N}$	$N \log 2d$	$\epsilon > O(\log d)$	$O(d)$
Scaled ε-Net (C_{net}) (Proposition 10)	$\frac{1}{N}$	$O_\varepsilon(Nd)^*$	-	$O\left(\left(\frac{1}{\varepsilon}\right)^d\right)$
Simplex (C_S) (Proposition 11)	$\frac{d^2}{N}$	$N \log(d+1)$	$\epsilon > \log 7$	$O(d)$
Hadamard (C_H) (Proposition 12)	$\frac{d^2}{N}$	$N \log d$	$\epsilon > \log(1+\sqrt{2})$	$O(d)$
Cross-polytope (C_{cp}) + RR (Theorem 14)	$\frac{d^2}{N}$	$N \log(2d)$	$\epsilon > 0$	$O(d)$
Cross-polytope (C_{cp}) + RAPPOR (Theorem 15)	$\frac{d^2}{N}$	$2Nd$	$\epsilon > 0$	$O(d)$

Table 1: List of results. (N: number of worker nodes, d: dimension). $^*O_\varepsilon$ hides terms involving ε

point-set is a valid vqSGD scheme, and prove existence of a point-set that achieves a communication cost equal to the dimension divided by the variance. We provide a few structured deterministic constructions of point sets in this section as well. Sec. 6 emphasizes the privacy component of vqSGD - and derives the privacy parameters of several vqSGD schemes. Finally, we provide some 'proof of the concept' experiments to support vqSGD. Missing proofs of all theorems/lemmas can be found in the appendix.

2 Related Work

The foundations of gradient quantization was laid by Seide et al. [2014] and Strom [2015] with schemes that require the compute nodes to send exactly 1-bit per coordinate of the gradient. They also suggested using local error accumulation to correct the global gradient in every iteration. While these novel techniques worked well in practice, there were no theoretical guarantees provided for convergence of the scheme. These seminal works fueled multiple research directions.

Quantization & Sparsification. Alistarh et al. [2017], Wen et al. [2017], Wang et al. [2018] propose stochastic quantization techniques to represent each coordinate of the gradient using small number of bits. The proposed schemes always return an unbiased estimator of the local gradient and require $c = \Omega(\sqrt{d})$ bits of communication to compute the global gradient with variance bounded by a multiplicative factor of $O(d/c)$. The quantization techniques for distributed SGD, can be used in the more general setting of communication efficient distributed mean estimation problem, which was the focus of Suresh et al. [2017]. The quantization schemes proposed in Suresh et al. [2017] require $O(d)$ bits of communication per compute node to estimate the global mean with a constant (independent of d) squared error (variance). Even though the tradeoff between communication and accuracy achieved by the above mentioned schemes are near optimal Zhang et al. [2013], they were unable to break the $\sqrt{d}$ barrier of communication cost. In this work, we propose quantization schemes that require about $\log d$ bits of communication and are almost optimal as well.

Gradient sparsification techniques with provable convergence (under standard assumptions) were studied in Acharya et al. [2019], Alistarh et al. [2018], Ivkin et al. [2019], Stich et al. [2018]. The main idea in these techniques is to communicate only the top-k *components* of the d-dimensional local gradients that can be accumulated globally to obtain a good estimate of the true gradient. Unlike the quantization schemes described above, gradient sparsification techniques can achieve $O(\log d)$ bits of communication, but are not usually unbiased estimates of the true gradients. Shalev-Shwartz et al. [2010] suggest randomized sparsification schemes that are unbiased, but are not known to provide any theoretical convergence guarantees in very low sparsity regimes.

See Table 2 for a comparison of our results with the state of the art quantization schemes.

Method	Error	Comm
QSGD Alistarh et al. [2017]	$\min\{\frac{d}{s^2}, \frac{\sqrt{d}}{s}\}\frac{1}{N}$	$Ns(s+\sqrt{d})$
DME Suresh et al. [2017]	$\min\{\frac{1}{Ns}, \frac{\log d}{N(s-1)^2}\}$	Nsd
vqSGD ($Q_{C_{cp}}$, $Q_{C_{RM}}$)	$\frac{d}{Ns}$	$Ns \log d$
vqSGD (Random Sampling)	$\frac{d}{Nsc}$	Nsc

Table 2: Comparison of non private quantization schemes. (N: number of worker nodes, s, c: tuning parameters (≥ 1).)

Error Feedback. Many works focused on providing techniques to reduce the error incurred due to quantization Horváth et al. [2019], Karimireddy et al. [2019] using locally accumulated errors. In this work, we focus primarily on gradient quantization techniques, and note that the variance reduction techniques of Horváth et al. [2019] can be used on top of the proposed quantization schemes.

Privacy. While differential privacy for gradient based algorithms Abadi et al. [2016], Shokri and Shmatikov [2015] were considered earlier in literature, cpSGD Agarwal et al. [2018] is the only work that considered achieving differential privacy for gradient based algorithms and simultaneously minimizing the gradient communication cost. The authors propose a binomial mechanism to add discrete noise to the quantized gradients to achieve communication-efficient (ϵ, δ)-differentially private gradient descent with convergence guarantees. The quantization schemes used are similar to those presented in Suresh et al. [2017] and hence require $\Omega(d)$ bits of communication per compute node. The parameters of the binomial noise are dictated by the required privacy guarantees which in turn controls the communication cost.

In this work we show that certain instantiations of our quantization schemes are ϵ-differentially private. Note that this is a much stronger privacy notion than (ϵ, δ)-privacy. Moreover, we get this privacy guarantee directly from the quantization schemes and hence the communication cost remains sublinear ($\log d$) in dimension. We also propose a Randomized Response Warner [1965] based private-quantization scheme that requires $O(\log d)$ bits of communication per compute node to get an ϵ-differential privacy while losing a factor of $O(d)$ in convergence rate. Table 3 compares the guarantees provided by our private quantization schemes with the results of cpSGD Agarwal et al. [2018].

Method	Error	Comm	DP (ϵ)
cpSGD Agarwal et al. [2018]	$O_\delta\left(\frac{d}{N}\right)^*$	$O_\delta(d)$	$\delta > 0$, $\epsilon > f(\delta)$
vqSGD ($Q_{C_{cp}}$)	$O\left(\frac{d}{N}\right)$	$O(\log d)$	$\epsilon > O(\log d)$
vqSGD (Q_{C_S})	$O\left(\frac{d^2}{N}\right)$	$O(\log d)$	$\epsilon > \log 7$
vqSGD (Q_{C_H})	$O\left(\frac{d^2}{N}\right)$	$O(\log d)$	$\epsilon > \log(2.5)$
vqSGD ($Q_{C_{cp}}$) + RR	$O\left(\frac{d^2}{N}\right)$	$O(\log d)$	$\epsilon > 0$

Table 3: Comparison of private quantization schemes. ($^*O_\delta$ hides terms involving δ)

3 Background

For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we denote the Euclidean (ℓ_2) distance between them as $\|\boldsymbol{x} - \boldsymbol{y}\|_2$. For any vector $\boldsymbol{x} \in \mathbb{R}^d$, x_i denotes its i-th coordinate. For any $\boldsymbol{c} \in \mathbb{R}^d$, and $r > 0$, let $B_d(\boldsymbol{c}, r)$ denote a d-dimensional ℓ_2 ball of radius r centered at $\boldsymbol{c}$. Also, let S^{d-1} denote the unit sphere about $\mathbf{0}_d$. Let $\boldsymbol{e_i} \in \mathbb{R}^d$ denote the i-th standard basis vector which has 1 in the i-th position and 0 everywhere else. Also, let $\mathbf{1}_d$ and $\mathbf{0}_d$ denote the all 1's vector and all 0's vector in $\mathbb{R}^d$ respectively. By $[n]$ we denote the set $\{1, 2, \dots, n\}$.

For a discrete set of points $C \subset \mathbb{R}^d$, let $\textsc{conv}(C)$ denote the convex hull of points in C, *i.e.*,

$$\textsc{conv}(C) := \left\{ \sum_{\boldsymbol{c} \in C} a_c \boldsymbol{c} \mid a_c \geq 0, \sum_{\boldsymbol{c} \in C} a_c = 1 \right\}.$$

Suppose $w \in \mathbb{R}^d$ be the parameters of a function to be learned (such as weights of a neural network). In each step of the SGD algorithm, the parameters are updated as $w \leftarrow w - \eta \hat{\boldsymbol{g}}$, where η is a possibly time-varying learning rate and $\hat{\boldsymbol{g}}$ is a stochastic unbiased estimate of $\boldsymbol{g}$, the true gradient of some loss function with respect to w. The convergence rate of the SGD algorithm depends on the variance (mean squared error) of the unbiased estimate, cf. any standard textbook such as Shalev-Shwartz and Ben-David [2014].

The goal of any gradient quantization scheme is to reduce cost of communicating the gradient, while not compromising too much on the *quality* of the gradient estimate. The quality of the gradient estimate is measured in terms the convergence guarantees it provides. Given an unbiased estimator $\hat{\boldsymbol{g}} \in \mathbb{R}^d$ of the true gradient $\boldsymbol{g}$, the convergence of SGD is known to depend on the variance of this estimate.

In distributed setting with N worker nodes, let $\boldsymbol{g_i}$ and $\hat{\boldsymbol{g}}_{\boldsymbol{i}}$ are the local true gradient and its unbiased estimate computed at the ith compute node for some $i \in \{1, \dots, N\}$. For $\boldsymbol{g} = \frac{1}{N}\sum_i \boldsymbol{g_i}$, variance of the estimate $\hat{\boldsymbol{g}} = \frac{1}{N}\sum_i \hat{\boldsymbol{g}}_{\boldsymbol{i}}$ is defined as

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{g}}) &:= \mathbf{E}\left[\| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{g_i} - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{g}}_{\boldsymbol{i}} \|_2^2 \right] \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \mathbf{E}\left[\| \boldsymbol{g_i} - \hat{\boldsymbol{g}}_{\boldsymbol{i}} \|_2^2 \right].
\end{aligned}$$

In this work, our goal is to design quantization schemes to efficiently compute unbiased estimate $\hat{\boldsymbol{g}}_{\boldsymbol{i}}$ of $\boldsymbol{g}_i$ such that $\mathrm{Var}(\hat{\boldsymbol{g}})$ is minimized.

For the privacy preserving gradient quantization schemes, we consider the standard notion of (ϵ, δ)-differential privacy (DP) as defined in Dwork et al. [2014]. Consider data-sets from a domain $\mathcal{X}$. Two data-sets $U, V \in \mathcal{X}$, are neighboring if they differ in at most one data point.

Definition 1. *A randomized algorithm $\mathcal{M}$ with domain $\mathcal{X}$ is (ϵ, δ)-differentially private (DP) if for all $S \subset$ Range$(\mathcal{M})$ and for all neighboring data sets $U, V \in \mathcal{X}$,*

$$\Pr[\mathcal{M}(U) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(V) \in S] + \delta,$$

where, the probability is over the randomness in $\mathcal{M}$. If $\delta = 0$, we say that $\mathcal{M}$ is ϵ-DP.

We will need the notion of an ε-nets subsequently.

Definition 2 (ε-net)**.** *A set of points $N(\varepsilon) \subset \mathcal{S}^{d-1}$ is an ε-net for the unit sphere $\mathcal{S}^{d-1}$ if for any point $\boldsymbol{x} \in \mathcal{S}^{d-1}$ there exists a net point $\boldsymbol{u} \in N(\varepsilon)$ such that $\|\boldsymbol{x} - \boldsymbol{u}\|_2 \leq \varepsilon$.*

There exist various constructions for ε-net over the unit sphere in $\mathbb{R}^d$ of size at most $\left(1 + \frac{2}{\varepsilon}\right)^d$ Cohen et al. [1997].

4 Quantization Scheme

We first present our quantization scheme in full generality. Individual quantization schemes with different tradeoffs are then obtained as specific instances of this general scheme.

Let $C = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c}_m\} \subset \mathbb{R}^d$ be a discrete set of points and let $\textsc{conv}(C)$ be its convex hull that satisfies

$$B_d(\mathbf{0}_d, 1) \subset \textsc{conv}(C) \subseteq B_d(\mathbf{0}_d, R), R > 1. \tag{1}$$

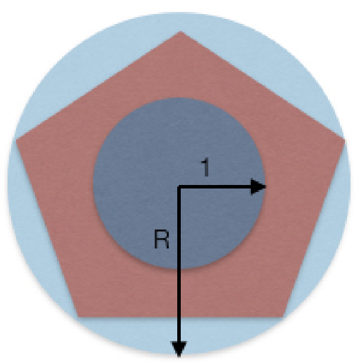


Figure 1: Convex Hull

Let $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$. Since $B_d(\mathbf{0}_d, 1) \subseteq \textsc{conv}(C)$, we can write $\boldsymbol{v}$ as a convex linear combination of points in C. Let $\boldsymbol{v} = \sum_{i=1}^{m} a_i \boldsymbol{c_i}$, where $a_i \geq 0, \sum_{i=1}^{m} a_i = 1$. We can view the coefficients of the convex combination $(a_1, \ldots, a_m)$ as a probability distribution over points in C. Define the quantization of $\boldsymbol{v}$ with respect to the set of points C as follows:

$$Q_C(\boldsymbol{v}) := \boldsymbol{c_i} \text{ with probability } a_i$$

It follows from the definition of the quantization that $Q_C(\boldsymbol{v})$ is an unbiased estimator of $\boldsymbol{v}$.

Lemma 1. $\mathbf{E}[Q_C(\boldsymbol{v})] = \boldsymbol{v}$.

Assume that the set C is fixed in advance and is known to both the compute nodes and the parameter server.

Remark 1. *Communicating the quantization of any vector $\boldsymbol{v}$, then amounts to sending a floating point number $\|v\|_2$, and the index of point $Q_C(\boldsymbol{v})$ which requires $\log |C|$ bits. For many loss functions, such as Lipschitz functions, the bound on the norm of the gradients is known to both the compute nodes and the parameter server. Therefore, we can avoid sending $\|v\|_2$ and the cost of communicating the gradients is then exactly $\log |C|$ bits.*

Any point set C that satisfies Condition (1) gives the following bound on the variance of the quantizer.

Lemma 2. *Let $C \subset \mathbb{R}^d$ be a point set satisfying Condition* (1). *For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, let $\hat{\boldsymbol{v}} := Q_C(\boldsymbol{v})$. Then,*

$$\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] \leq R^2 - 1.$$

From the above mentioned properties, we get a family of quantization schemes depending on the choice of point set C that satisfy Condition (1). For any choice of quantization scheme from this family, we get the following bound regarding the convergence of the distributed SGD.

Theorem 3. *Let $C \subset \mathbb{R}^d$ be a point set satisfying Condition* (1). *Let $\boldsymbol{g_i} \in \mathbb{R}^d$ be the local gradient computed at the i-th compute node, Define $\hat{\boldsymbol{g}} := \frac{1}{N}\sum_{i=1}^N \hat{\boldsymbol{g}}_{\boldsymbol{i}}$, where $\hat{\boldsymbol{g}}_{\boldsymbol{i}} := \|\boldsymbol{g_i}\| \cdot Q_C(\boldsymbol{g_i}/\|\boldsymbol{g_i}\|)$. Then,*

$$\mathbf{E}[\hat{\boldsymbol{g}}] = \boldsymbol{g} \quad \text{and} \quad \mathbf{E}\left[\|\boldsymbol{g} - \hat{\boldsymbol{g}}\|_2^2\right] = \frac{R^2-1}{N^2}\sum_i \|\boldsymbol{g_i}\|^2.$$

Remark 2. *In order to compute the quantization $Q_C(\boldsymbol{v})$, we have to first compute the convex combination of $\boldsymbol{v}$ with respect to the point set C. This requires us to solve a system of $|C|$ linear equations in $\mathbb{R}^d$. For general point sets C, this takes about $O(|C|^3)$ time (since $|C| \geq d$). We will show that there exist certain structured point sets for which we can compute these probabilities in linear time.*

From Theorem 3 we observe that the communication cost of the quantization scheme depends on the cardinality of C while the convergence is dictated by the circumradius R of the convex hull of C. In the Section 5, we present several constructions of point sets which provide varying tradeoffs between communication and variance of the quantizer.

Reducing Variance: In this section, we propose a simple repetition technique to reduce the variance of the quantization scheme. For any $s > 1$, let $Q_C(s, \boldsymbol{v}) := \frac{1}{s}\sum_{i=1}^s Q_C(\boldsymbol{v})$ be the average over s independent applications of the quantization $Q_C(\boldsymbol{v})$. Note that even though $Q_C(s, \boldsymbol{v})$ is not a point in C, we can communicate $Q_C(s, \boldsymbol{v})$ using an equivalent representation as a tuple of s independent applications of $Q_C(\boldsymbol{v})$ that requires $s \log |C|$ bits.

$$Q_C(s, \boldsymbol{v}) \equiv (\overbrace{Q_C(\boldsymbol{v}), \ldots, Q_C(\boldsymbol{v})}^{s\text{-times}}).$$

Using this repetition technique described above, we see that the variance reduces by factor of s while the communication increases by the exact same factor.

Proposition 4. *Let $C \subset \mathbb{R}^d$ be a point set satisfying Condition* (1). *For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, and any $s \geq 1$, let $\hat{\boldsymbol{v}}$ be the average of the s repetitions in $Q_C(s, \boldsymbol{v})$. Then, $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = \frac{(R^2-1)}{s}$.*

5 Constructions of Point Sets

In this section, we propose constructions of various point sets that satisfy Condition (1). The cardinality and circumradius of these point sets provide varying tradeoffs between communication and variance of the quantization scheme.

First, we show a strong characterization of the point sets that satisfy Condition (1), and later use this characterization to construct point sets with optimal tradeoffs.

Theorem 5. *Let $C = \{c_1, \ldots, c_m\} \subseteq \mathbb{R}^d$ be a discrete set of points. The unit ball $B_d(\mathbf{0}_d, 1) \subseteq$ $\textsc{conv}(C)$ if and only if for all points $\boldsymbol{x} \in S^{d-1}$, there exists a point $\boldsymbol{c} \in C$ such that $\langle \boldsymbol{x}, \boldsymbol{c} \rangle \geq 1$.*

Proof. Assume that for some $x \in S^{d-1}$, $\boldsymbol{x}^T\boldsymbol{c} < 1$ for all $c \in C$. Which implies that all points of C, and therefore the $\textsc{conv}(C)$, are separated from x by the hyperplane $H_w := \{w \in \mathbb{R}^d | x^T w = 1\}$. Therefore $x \notin \textsc{conv}(C)$.

To prove the other side, assume $B_d(\mathbf{0}_d, 1) \not\subset \textsc{conv}(C)$. Let $H_w := \{z \in \mathbb{R}^d | w^T z = 1\}$ be the separating hyperplane that partitions $B(0,1)$ such that $\textsc{conv}(C)$ lies on one side of the hyperplane. Without loss of generality, we assume $\textsc{conv}(C) \subset H_w^- := \{z \in \mathbb{R}^d | w^T z < 1\}$. Since H_w partitions the unit ball, the distance of H_w from the origin is $1/\|w\| \leq 1$.

Now consider the point $x := w/\|w\| \in S^{d-1}$. For this point, $x^T c = w^T c/\|w\| \leq 1$ for all $c \in C$. □

5.1 Random Sampling

We have the following randomized construction of point set using the characterization defined above.

Theorem 6. *Let $R > 2$. There exists a set C of $\exp(O(d/R^2))$ points of ℓ_2 norm at most R each, that satisfy Condition (1).*

Proof. Consider a set of p points $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_p\} \subset \mathbb{R}^d$ where each $\boldsymbol{c}_i$ is sampled randomly from a ball of radius R. The theorem basically shows that if we choose $|C|$ and R appropriately, then C satisfies the characterization of Theorem 5 with high probability.

For any fixed $\boldsymbol{x} \in S^{d-1}$, and $\boldsymbol{c} \in C$, define

$$p_{\boldsymbol{x},\boldsymbol{c}} := \Pr[\boldsymbol{x}^T\boldsymbol{c} \geq 2].$$

Let us choose the random set C in the following way. Each coordinate of any $\boldsymbol{c} \in C$ is chosen i.i.d. uniformly according to a zero-mean Gaussian distribution with variance $\sigma^2 := \frac{R^2}{9d}$. Then $\|\boldsymbol{c}\|^2$ is distributed according to a χ^2-distribution with variance $2d\sigma^4$. Since χ^2-distribution is subexponential Wainwright [2019][Eq. 2.18], for any $\boldsymbol{c} \in C$, we have, for any $t \geq 1$, $\Pr(\|\boldsymbol{c}\|^2 > d\sigma^2(t+1)) \leq e^{-dt/8}$. This implies,

$$\Pr(\|\boldsymbol{c}\|^2 > R^2) \leq e^{-\frac{1}{8}(R^2/\sigma^2 - d)} \leq e^{-d},$$

substituting the value of σ^2.

On the other hand note that, $\boldsymbol{x}^T\boldsymbol{c}$ is $\mathcal{N}(0, \sigma^2)$. For any $\boldsymbol{c} \in C$ we can therefore bound $p_{\boldsymbol{x},\boldsymbol{c}}$ as Borjesson and Sundberg [1979]:

$$p_{\boldsymbol{x},\boldsymbol{c}} \geq \frac{2\sigma}{(\sigma^2+4)\sqrt{2\pi}} e^{-\frac{2}{\sigma^2}}$$

Let $\mathbb{1}_{\boldsymbol{x},\boldsymbol{c}}$ denote the indicator random variable that takes a value of 1 if and only if $\boldsymbol{x}^T\boldsymbol{c} \geq 2$. Since $\mathbb{1}_{\boldsymbol{x},\boldsymbol{c}}$ is a Bernoulli random variable, we have

$$p_{\boldsymbol{x}} := \Pr[\sum_{\boldsymbol{c} \in C} \mathbb{1}_{\boldsymbol{x},\boldsymbol{c}} = 0] = (1 - p_{\boldsymbol{x},\boldsymbol{c}})^{|C|} \leq e^{-|C| \cdot p_{\boldsymbol{x},\boldsymbol{c}}}.$$

$p_{\boldsymbol{x}}$ is the probability that $\boldsymbol{x} \in S^{d-1}$ is a witness to the fact that $\textsc{conv}(C)$ does not contain the unit ball. Now, in order to complete the proof, we need to show that this bad event $\bigcup_{\boldsymbol{x}\in S^{d-1}}(\sum_{\boldsymbol{c}\in C} \mathbb{1}_{\boldsymbol{x},\boldsymbol{c}} = 0)$ happens rarely. We first show that it is sufficient to take a union bound over all the points of an ε-net, for any $\varepsilon < 1/R$.

Consider an ε-net for the unit sphere $N(\varepsilon)$ for any $\varepsilon < 1/R$. We know that such a set exists with $|N(\varepsilon)| \leq \left(1+\frac{2}{\varepsilon}\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d$ Cohen et al. [1997]. Let $\boldsymbol{x} \in N(\varepsilon)$ be a net-point, and $\boldsymbol{c} \in C$ be a point such that $\boldsymbol{x}^T c \geq 2$, then for all points $\boldsymbol{y} \in S^{d-1}$ in the ε-neighborhood of $\boldsymbol{x}$ can be written as $\boldsymbol{y} = \boldsymbol{x} + \tilde{\boldsymbol{x}}$, where $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$ has norm at most ϵ. Therefore, $\boldsymbol{y}^T\boldsymbol{c} = \boldsymbol{x}^T\boldsymbol{c} + \tilde{\boldsymbol{x}}^T\boldsymbol{c} \geq 2 - \|\tilde{\boldsymbol{x}}\|\|\boldsymbol{c}\| > 1$. Therefore if for all points in $N(\varepsilon)$ there exists a point $\boldsymbol{c} \in C$ such that $\boldsymbol{x}^T\boldsymbol{c} \geq 2$, then for all points on the unit sphere, there will be a $\boldsymbol{c} \in C$ such that $\boldsymbol{x}^T\boldsymbol{c} \geq 1$.

From union bound it then follows that

$$\begin{aligned}\Pr\left[\bigcup_{\boldsymbol{x}\in N(\varepsilon)}\left(\sum_{\boldsymbol{c}\in C} 1_{\boldsymbol{x},\boldsymbol{c}} = 0\right)\right] &\leq \sum_{\boldsymbol{x}\in N(\varepsilon)} e^{-|C|\cdot p_{\boldsymbol{x},\boldsymbol{c}}} \\ &\leq e^{d\ln 3R} e^{-|C|\cdot p_{\boldsymbol{x},\boldsymbol{c}}} \leq e^{d\ln 3R - |C|\cdot O(\min(\sigma,\sigma^{-1}))e^{-\frac{18d}{R^2}}}.\end{aligned}$$

It then follows for $|C| = \exp(\Theta(d/R^2))$ with probability at least $1-e^{-d}$, the point set C satisfies the characterization of Theorem 5.

It now remains to be shown that for all $\boldsymbol{c} \in C$, $\|\boldsymbol{c}\| \leq R$. By using a union bound on the size of C, this happens with probability at least $1 - |C| \cdot e^{-d} \geq 1 - e^{-\Theta(d)}$. □

The above stated theorem provides a randomized algorithm to generate a point set of size $\exp(\Theta(d/R^2))$ such that the quantization scheme defined in Section 4 instantiated with this point set achieves a variance of $O(R^2)$ while communicating $O(d/R^2)$ bits. In particular, there exists a quantization scheme that achieves $O(1)$ variance with $O(d)$ bits of communication (see Section 5.3.2 for a deterministic construction). Also, at a cost of communicating only $O(\log d)$ bits, our quantization scheme can achieve a variance of $O(d/\log d)$. The deterministic constructions we provide (in Section 5.2, and also Section 5.3.1), meet this bound up to a factor of $\log d$.

5.2 Derandomizing with Reed Muller Codes

In this section, we propose a deterministic construction of point set based on first order Reed-Muller codes that satisfy the characterization of Theorem 5. Let us assume that d is a power of 2, *i.e.*, $d = 2^p$ for some $p \geq 1$. This assumption can be made without loss of generality by padding the gradients with zeros to make the dimension a power of 2.

Our quantization scheme is based on the first order Reed-Muller codes, $\mathrm{RM}(1,p)$ MacWilliams and Sloane [1977]. Each codeword of $\mathrm{RM}(1,p)$ is given as the evaluation of a degree 1, p-variate polynomial over all points in $\mathbb{F}_2^p$. Mapping these codewords to reals using the coordinate-wise map $\phi : \mathbb{F}_2 \to \mathbb{R}$ defined as $\phi(b) = (-1)^b$ will give us a set of $2d$ points in $\{\pm 1\}^d$. Let RM denote this set of mapped codewords.

We show that the set of points in RM satisfy the characterization of Theorem 5, and therefore will give us a quantization scheme with the following guarantees:

Proposition 7. *For any* $\boldsymbol{v} \in B_d(\boldsymbol{0}_d, 1)$*, let* $\hat{\boldsymbol{v}} := Q_{C_{RM}}(\boldsymbol{v})$*. Then,* $\mathbf{E}[\hat{v}] = \boldsymbol{v}$ *and* $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = O(d)$*.*

Proof. We prove this theorem by showing that the point set C_{RM} satisfies the characterization of Theorem 5. Since all points in C_{RM} have squared norm exactly d, from Lemma 1 and Lemma 2, the proof follows.

First note that the matrix with the points in RM its rows has the following structure:

$$H := \begin{bmatrix} H_p \\ -H_p \end{bmatrix}$$

where, H_p is the $2^p \times 2^p$ Hadamard matrix.

For any fixed $\boldsymbol{x} \in S^{d-1}$, consider the sum $S(\boldsymbol{x}) := \sum_{\boldsymbol{c} \in C_{RM}} (\boldsymbol{x}^T \boldsymbol{c})^2$. We first show that $S(\boldsymbol{x}) \geq 2(d+1)$.

$$\begin{aligned} S(\boldsymbol{x}) &= \sum_{\boldsymbol{c} \in C_{RM}} (\boldsymbol{x}^T \boldsymbol{c})^2 = 2 \sum_{\boldsymbol{h_i} \in H_p} (\boldsymbol{x}^T \boldsymbol{h_i})^2 \\ &= 2\|H_p \boldsymbol{x}\|^2 = 2(\boldsymbol{x}^T H_p^T)(H_p \boldsymbol{x}) \\ &\overset{(i)}{=} 2d \cdot \|x\|^2 = 2d. \end{aligned}$$

(i) follows from the fact that the columns of the Hadamard matrix are mutually orthogonal and therefore, $H_p^T H_p = d \cdot I_d$, where, I_d denotes the $d \times d$ identity matrix.

By an averaging argument, it then follows that there exists at least one $\boldsymbol{c} \in C_{RM}$ such that $|\boldsymbol{x}^T \boldsymbol{c}| \geq 1$. Since for every $\boldsymbol{c} \in C_{RM}$, there exists $-\boldsymbol{c} \in C_{RM}$, we get that $x^T \boldsymbol{c} \geq 1$ for some $\boldsymbol{c} \in C_{RM}$. □

Remark 3. *Instead of first order Reed-Muller codes, one can choose the point set using any binary linear code $C \subseteq \mathbb{F}_2^d$ as follows. First map all the codewords from $\mathbb{F}_2^d$ to $\mathbb{R}^d$ using the coordinate wise mapping ϕ described above. The point set containing all such mapped codewords, and their complements will give a quantization scheme with variance $O(d)$. The communication will however be $\log(2|C|)$, where $|C|$ denotes the number of codewords in C. In this regard, the first order Reed-Muller codes described above provide the best communication guarantees and is also efficiently computatble.*

5.3 Other Deterministic Constructions

We now present several explicit constructions of point sets that give quantization schemes with varying tradeoffs. On one end of the spectrum, the cross-polytope scheme requires only $O(\log d)$ bits to communicate an unbiased estimate of a vector in $\mathbb{R}^d$ with variance $O(d)$. On the other end, the ε-net based scheme achieves a constant variance at the cost of $O(d)$ bits of communication.

5.3.1 Cross Polytope Scheme

Consider the following point set of $2d$ points in $\mathbb{R}^d$:

$$C_{cp} := \{\pm\sqrt{d}\ \boldsymbol{e_i} \mid i \in [d]\},$$

The convex hull $\textsc{conv}(C_{cp})$ is a scaled cross polytope that satisfies Condition (1) with $R = \sqrt{d}$ (see Proposition 8 for the proof). Let $Q_{C_{cp}}$ be the instantiation of the quantization scheme described in Section 4 with the point set C_{cp}.

To compute the convex combination of any point $\boldsymbol{v} \in \textsc{conv}(C_{cp})$, we need a non-negative solution to the following system of equations

$$\begin{bmatrix} \sqrt{d} I_d & -\sqrt{d} I_d \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_{2d} \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} \quad \text{such that } \sum_{i=1}^{2d} a_i = 1, \tag{2}$$

where, I_d is the $d \times d$ identity matrix. Equation 2 leads to the following closed form solution that can be computed in $O(d)$ time:

$$a_i = \begin{cases} \frac{v_i}{\sqrt{d}} + \frac{\gamma}{2d} & \text{if } v_i > 0 \text{ and } i \leq d \\ -\frac{v_i}{\sqrt{d}} + \frac{\gamma}{2d} & \text{if } v_i \leq 0 \text{ and } i > d \\ \frac{\gamma}{2d} & \text{otherwise} \end{cases} \tag{3}$$

where, $\gamma := 1 - \frac{\|\boldsymbol{v}\|_1}{\sqrt{d}}$, is a non-negative quantity for every $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$.

The bound on the variance of the quantizer follows directly from Lemma 2.

Proposition 8. *For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, let $\hat{\boldsymbol{v}} := Q_{C_{cp}}(\boldsymbol{v})$. Then, $\mathbf{E}[\hat{\boldsymbol{v}}] = \boldsymbol{v}$ and $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = O(d)$.*

Moreover, using the variance reduction technique described in Section 4 with $s = O(\frac{d}{\log d})$, the cross polytope based quantization scheme $Q_{C_{cp}}$ achieves a variance of $O(\log d)$ at the cost of communicating $O(d)$ bits.

We note that the cross-polytope quantization scheme described above when used along with the variance reduction technique (by repetition), is in essence similar to Maurey sparsification Acharya et al. [2019].

5.3.2 Scaled ε-nets

On the other end of the spectrum, we now show the existence of points sets of exponential size that are contained in a constant radius ball. This point set allows us to obtain a gradient quantization scheme with $O(d)$ communication and $O(1)$ variance. Recall the definition of the ε-net. We now show that appropriate constant scaling of the net points satisfies Condition (1).

Lemma 9. *For any $0 < \varepsilon < 1$, let $R = \frac{1}{1-\varepsilon^2/2}$. The point set $C_{net} := \{R \cdot \boldsymbol{u} \mid \boldsymbol{u} \in N(\varepsilon)\}$ satisfies Condition (1).*

Let Q_{net} be the instantiation of the quantization scheme with point set C_{net}. From Lemma 2, we then directly get the following guarantees for the quantization scheme obtained from scaled ε-nets, C_{net} for some constant $\varepsilon < 1$.

Proposition 10. *For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, let $\hat{\boldsymbol{v}} := Q_{C_{net}}(\boldsymbol{v})$. Then, $\mathbf{E}[\hat{\boldsymbol{v}}] = \boldsymbol{v}$ and $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = \frac{1}{(1-\varepsilon^2/2)^2} - 1$.*

Moreover, Q_{net} requires $O(d \log \frac{1}{\varepsilon})$ bits to represent the unbiased gradient estimate.

6 Private Quantization

In this section we show that under certain conditions the quantization scheme $Q_C(.)$ obtained from the point set C is also ϵ-differentially private. We first see why certain quantization schemes described in Section 4 are not privacy preserving in general.

Let C be any point set with $|C| > d + 1$. For any point $\boldsymbol{x} = \sum_{i=1}^{|C|} a_i \boldsymbol{c_i} \in \textsc{conv}(C)$, let $\textsc{supp}(\boldsymbol{x}, C) = \{\boldsymbol{c}_i \in C \mid a_i \neq 0\}$ denote the indices of the points in C that are in the range of $Q_C(\boldsymbol{x})$.

In order for Q_C to be differentially private for any $\epsilon > \epsilon_0$, we need to show that for any two gradient vectors, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ of neighboring datasets and any $\boldsymbol{z} \in \textsc{supp}(\boldsymbol{x}, C) \cup \textsc{supp}(\boldsymbol{y}, C)$,

$$\frac{\Pr[Q_C(\boldsymbol{x}) = \boldsymbol{z}]}{\Pr[Q_C(\boldsymbol{y}) = \boldsymbol{z}]} \leq e^{\epsilon_0}. \tag{4}$$

Since $|C| > d+1$, there may exist two gradient vectors $\boldsymbol{x}, \boldsymbol{y} \in \textsc{conv}(C)$ such that $\textsc{supp}(\boldsymbol{x}, C) \neq \textsc{supp}(\boldsymbol{y}, C)$. Therefore, for any $\boldsymbol{z} \in \textsc{supp}(\boldsymbol{x}, C) \setminus \textsc{supp}(\boldsymbol{y}, C)$, there may not exist any finite ϵ_0 for which Equation (4) holds.

The discussion above establishes a sufficient condition for the quantization scheme Q_C to be differentially private. Essentially, we want all points in $B_d(\mathbf{0}_d, 1)$ to have full support on all the points in C. This is definitely possible when $|C| = d+1$. Therefore if the point set satisfying Condition (1) has size $|C| = d+1$, then the quantization scheme Q_C is ϵ-differentially private, for some $\epsilon > \epsilon(C)$.

We now present two constructions of point sets C of size exactly $d+1$ satisfying Condition (1) that give an ϵ-differentially private quantization scheme. Both the schemes achieve a communication cost of $\log(d+1)$, but the variance is a factor d larger than the non-private scheme, $Q_{C_{cp}}$.

6.1 Simplex Scheme

Consider the following set of $d+1$ points

$$C_S = \{2d\ \boldsymbol{e_i} \mid i \in [d]\} \cup \{-4\mathbf{1}_d\}.$$

The convex hull of C_S satisfies Condition (1) with $R = O(d)$ (see Proposition 11 for proof). Since the size of the set is exactly $d+1$, every point in the unit ball can be represented as a convex combination of all the points in C_S (*i.e.*, all coefficients of the convex combination are non zero). This fact will be used crucially to show that this scheme is also differentially private.

The coefficients of the convex combination of any point $\boldsymbol{v} \in \textsc{conv}(C_S)$ can be computed from the following system of linear equations:

$$\begin{bmatrix} -4\mathbf{1}_d^T \\ 2\sqrt{d} I_d \end{bmatrix}^T \begin{bmatrix} a_0 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} \text{ such that } \sum_{i=0}^{d} a_i = 1. \tag{5}$$

Equation 5 leads to the following closed form solution that can be computed in linear-time:

$$a_0 = \frac{2d - \sum_{i=1}^{d} v_i}{6d} \qquad a_i = \frac{v_i + 4a_0}{2d} \qquad \forall i \geq 1. \tag{6}$$

Proposition 11. *For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, let $\hat{\boldsymbol{v}} := Q_{C_S}(\boldsymbol{v})$. Then, $\mathbf{E}[\hat{\boldsymbol{v}}] = \boldsymbol{v}$ and $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = O(d^2)$. Moreover, Q_{C_S} is ϵ-DP for any $\epsilon > \log 7$.*

6.2 Hadamard Scheme

We now propose another quantization scheme with same communication cost, but provides better privacy guarantees. This quantization scheme is similar to the one presented in Section 5.2 and is based on the columns of a Hadamard matrix.

Let us assume that $d+1$ is a power of 2, *i.e.*, $d+1 = 2^p$ for some $p \geq 1$. For any $i \in [d+1]$, let $\boldsymbol{h_i} \in \mathbb{R}^d$ denote the i-th column of H_p with the first coordinate punctured. Consider the following set of $d+1$ points obtained from the punctured columns of H_p:

$$C_H = \{2\sqrt{d}\ \boldsymbol{h_i} \mid i \in [d+1]\}$$

The quantization scheme Q_{C_H} can be implemented in linear time since computing the probabilities requires computing a matrix vector product,

$$\begin{bmatrix} a_1 & \cdots & a_{d+1} \end{bmatrix}^T = \frac{1}{d+1} \cdot H_p^T \begin{bmatrix} 1 \\ \frac{\boldsymbol{v}}{2\sqrt{d}} \end{bmatrix}$$

that has closed form solution for each a_i as:

$$a_i = \frac{1}{d+1}(1 + \frac{\boldsymbol{h_i}^T \boldsymbol{v}}{2\sqrt{d}}) \tag{7}$$

Proposition 12. *For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, let $\hat{\boldsymbol{v}} := Q_{C_H}(\boldsymbol{v})$. Then, $\mathbf{E}[\hat{\boldsymbol{v}}] = \boldsymbol{v}$ and $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = O(d^2)$. Moreover, Q_{C_H} is ϵ-DP for any $\epsilon > \log(1+\sqrt{2})$.*

Remark 4. *Finally, we remark that even though the point set C_{cp} in the cross-polytope scheme (in Section 5.3.1) has more than $d+1$ points, it still gives us differential privacy although with slightly worse privacy parameter.*

Proposition 13. *$Q_{C_{cp}}$ is ϵ-DP for any $\epsilon > \log d$.*

We now show a Randomized Response scheme that can be used on top of any of our quantization schemes to make it ϵ-DP. This scheme incurs the same communication as the original quantizer, however, the price of privacy is paid by factor of d increase in the variance. We also propose a weaker version using Rappor , that incurs a higher communication cost depending on the point set of choice (see Supplementary Material for details).

6.3 Randomized Response

Recall that the quantization scheme Q_C for general point sets C need not be differentially private. In this section, we present a Randomized Response (RR) Warner [1965] mechanism that can be used over the output of Q_C to make it ϵ-differentially private (for any $\epsilon > 0$). This modified scheme retains the original communication cost of Q_C, but the cost for privacy is paid by a factor of $O(d)$ in the variance term.

Recall that the quantization scheme described in Section 4, $Q_C(\boldsymbol{v})$, takes a vector $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$ and returns a point $\boldsymbol{c_i} \in C$. The RR scheme takes the output of $Q_C(\boldsymbol{v})$ and returns a another random vector from C.

For any $\epsilon > 0$, define $p := p(\epsilon) = \frac{e^\epsilon}{e^\epsilon + |C| - 1}$ and $q := \frac{1-p}{|C|-1} = \frac{1}{e^\epsilon + |C| - 1}$. We define the private quantization of a vector $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$ as

$$\hat{\boldsymbol{v}} = PQ_{C,\epsilon}(\boldsymbol{v}) = \frac{1}{p-q} \sum_{i=1}^{|C|} (\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}} - q)\boldsymbol{c_i},$$

where, $\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}}$ is an indicator of the event $\boldsymbol{y} = \boldsymbol{c_i}$ and $\boldsymbol{y} := \text{RR}_p(Q_C(\boldsymbol{v}), C)$ is defined as

$$\text{RR}_p(Q_C(\boldsymbol{v}), C) = \begin{cases} Q_C(\boldsymbol{v}) & \text{w.p. } p \\ \boldsymbol{z} \in C \setminus \{Q_C(\boldsymbol{v})\} & \text{w.p. } q \end{cases}$$

We claim that the quantization scheme $PQ_{C,\epsilon}$ is ϵ-differentially private.

Theorem 14. *Let $C \subset \mathbb{R}^d$ be any point set satisfying Condition (1). For any $\epsilon > 0$, let $p = \frac{e^\epsilon}{e^\epsilon + |C| - 1}$ and $q = \frac{1}{e^\epsilon + |C| - 1}$. For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, let $\hat{\boldsymbol{v}} = PQ_{C,\epsilon}(\boldsymbol{v}) = \frac{1}{p-q} \sum_{i=1}^{|C|} (\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}} - q)\boldsymbol{c_i}$, where, $\boldsymbol{y} := \text{RR}_p(Q_C(\boldsymbol{v}), C)$. Then, $\mathbf{E}[\hat{\boldsymbol{v}}] = \boldsymbol{v}$ and $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = O(|C|R^2)$, where the expectation is taken over the randomness in both Q_C and RR_p. Moreover, the scheme is ϵ-differentially private.*

Proof. First we show that $\hat{\boldsymbol{v}} = PQ_{C,\epsilon}(\boldsymbol{v}) = \frac{1}{p-q}\sum_{i=1}^{|C|}(\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}} - q)\boldsymbol{c_i}$ is an unbiased estimator of v. From linearity of expectations, we have

$$\mathbf{E}[\hat{\boldsymbol{v}}] = \frac{1}{p-q}\sum_{i=1}^{|C|}(\Pr[\boldsymbol{y} = \boldsymbol{c_i}] - q)\boldsymbol{c_i}, \tag{8}$$

where, the expectation is taken over the randomness of both the quantization and RR scheme. Recall that

$$\boldsymbol{y} := \text{RR}_{\ p}(Q_C(\boldsymbol{v}), C) \in C,$$

where $p = \frac{e^\epsilon}{e^\epsilon + |C| - 1}$. Therefore,

$$\begin{aligned}\Pr(\boldsymbol{y} = \boldsymbol{c_i}) &= \sum_{j=1}^{|C|}\Pr[\boldsymbol{y} = \boldsymbol{c_i}|Q_C(\boldsymbol{v}) = \boldsymbol{c_j}] \cdot \Pr[Q_C(\boldsymbol{v}) = \boldsymbol{c_j}] \\ &= (p-q)a_i + q.\end{aligned}$$

Therefore $\mathbf{E}[\hat{\boldsymbol{v}}] = \frac{1}{p-q}\sum_{i=1}^{|C|} a_i\boldsymbol{c_i} = \boldsymbol{v}$.

Now we bound the variance of the estimator

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|^2] &= \mathbf{E}[\|\sum_{i=1}^{|C|}(\frac{1}{p-q}(\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}} - q) - a_i)\boldsymbol{c_i}\|^2] \\ &\leq \sum_{i=1}^{|C|}\mathbf{E}[\|(\frac{1}{p-q}(\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}} - q) - a_i)^2\|\boldsymbol{c_i}\|^2] \\ &= \sum_{i=1}^{|C|} Var[\|(\frac{1}{p-q}(\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}} - q))\|\boldsymbol{c_i}\|^2] \\ &= (\frac{1}{p-q})^2\sum_{i=1}^{|C|} Var(\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}})\|\boldsymbol{c_i}\|^2 \\ &= O(|C|R^2)\end{aligned}$$

As $\|c_i\|^2 \leq R^2$ and $Var(\mathbf{1}_{\{\boldsymbol{y}=\boldsymbol{c_i}\}}) \leq 1/4$.

Privacy Now we show that our scheme is ϵ differentially private where ϵ is the input parameter to the RR algorithm. For any two points $\boldsymbol{v}, \boldsymbol{w} \in B_d(\mathbf{0}_d, 1)$,

$$\frac{PQ_{C,\epsilon}(\boldsymbol{v}) = y}{PQ_{C,\epsilon}(\boldsymbol{w}) = y} = \frac{\sum_{i=1}^{|C|}\Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i})\Pr(Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\sum_{j=1}^{|C|}\Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j})\Pr(Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \tag{9}$$

$$\leq \frac{\max_i \Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i})\sum_{i=1}^{|C|}\Pr(Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\min_j \Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j})\sum_{i=1}^{|C|}\Pr(Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \tag{10}$$

$$= \frac{\max_i \Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\min_j \Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \leq e^\epsilon \tag{11}$$

we are using the following privacy property of Randomized Rounding Warner [1965] mechanism in Equation (11)

$$\sup_{i,j}\frac{\Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \leq e^\epsilon \quad \forall \boldsymbol{v}, \boldsymbol{w}$$

□

6.4 Privacy using Rappor

In this section, we present an alternate mechanism to make the quantization scheme ϵ-DP (for any $\epsilon > 0$). The main idea is to use the Rappor Erlingsson et al. [2014] mechanism over a 1-hot encoding of the indices of vertices in C. Though in doing so, we have to tradeoff on the communication a bit. Instead of sending $\log |C|$ bits, this scheme now requires one to send $O(|C|)$ bits to achieve privacy.

Recall that the quantization scheme described in Section 4, $Q_C(\boldsymbol{v})$, takes a vector $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$ and returns a point $\boldsymbol{c_i}$ in C. We can interpret the output as the bit string $\boldsymbol{b} \in \{0,1\}^{|C|}$ which is the indicator of the point $\boldsymbol{c_i}$ in C (according to some fixed arbitrary ordering of C). Note that this is essentially the 1-hot encoding of $\boldsymbol{c_i}$. In the RAPPOR scheme each bit of the 1-hot bit string $\boldsymbol{b}$ is flipped independently with probability $p := p(\epsilon) = \frac{1}{(e^{\epsilon/2}+1)}$.

For any $\epsilon > 0$, let $p = \frac{1}{(e^{\epsilon/2}+1)}$. Define, the private quantization of a vector $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$ as

$$\hat{\boldsymbol{v}} := PQ_{C,\epsilon}(\boldsymbol{v}) = \frac{1}{(1-2p)} \sum_{j=1}^{|C|} (y_j - p)\, \boldsymbol{c_j}$$

where, $\boldsymbol{y} := \text{Rappor}_{\ p}(\text{1-hot } (Q_C(\boldsymbol{v}), C)) \in \{0,1\}^{|C|}$.

We claim that the quantization scheme $PQ_{C,\epsilon}$ is ϵ-differentially private. Moreover, adding the noise over the 1-hot encoding maintains the unbiasedness of the gradient estimate but incurs a factor of $|C|$ in variance term while the communication cost is $O(|C|)$.

Theorem 15. *Let $C \subset \mathbb{R}^d$ be any point set satisfying Condition (1). For any $\epsilon > 0$, let $p = \frac{1}{(e^{\epsilon/2}+1)}$. For any $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$, let $\hat{\boldsymbol{v}} := \frac{1}{1-2p} \sum_{j=1}^{|C|} (y_j - p)\, \boldsymbol{c_j}$, where, $\boldsymbol{y} := \text{Rappor}_{\ p}(\text{1-hot } (Q_C(\boldsymbol{v}), C))$. Then, $\mathbf{E}[\hat{\boldsymbol{v}}] = \boldsymbol{v}$ and $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = O(|C|R^2)$. Moreover, the scheme is ϵ-differentially private.*

Proof of Theorem 15. First we show that $\hat{\boldsymbol{v}} = \frac{1}{(1-2p)} \sum_{j=1}^{|C|} (y_j - p)\, \boldsymbol{c_j}$ is an unbiased estimator of v. From linearity of expectations, we have

$$\mathbf{E}[\hat{\boldsymbol{v}}] = \frac{1}{(1-2p)} \sum_{j=1}^{|C|} (\mathbf{E}[y_j] - p)\, \boldsymbol{c_j}, \tag{12}$$

where, the expectation is taken over the randomness of both the quantization and RAPPOR scheme.

Recall that

$$\boldsymbol{y} := \text{Rappor}_{\ p}(\text{1-hot } (Q_C(\boldsymbol{v}), C)) \in \{0,1\}^{|C|}.$$

Each entry of the vector $\boldsymbol{y}$ is an independent binary random variable and

$$\begin{aligned}
\mathbf{E}[y_j] = \Pr(y_j = 1) &= \sum_{i=1}^{|C|} \Pr(y_j, Q_C(\boldsymbol{v}) = \boldsymbol{c_i}) \\
&= \sum_{i=1}^{|C|} Pr(y_j | Q_C(\boldsymbol{v}) = \boldsymbol{c_i}) Pr(Q_C(\boldsymbol{v}) = \boldsymbol{c_i}) \\
&= Pr(y_j | Q_C(\boldsymbol{v}) = \boldsymbol{c_j}) Pr(Q_C(\boldsymbol{v}) = \boldsymbol{c_j}) \\
&+ \sum_{i \neq j}^{|C|} Pr(y_j | Q_C(\boldsymbol{v}) = \boldsymbol{c_i}) Pr(Q_C(\boldsymbol{v}) = \boldsymbol{c_i}) \\
&= (1-p)a_j + p(1-a_j) = p + (1-2p)a_j. \qquad (13)
\end{aligned}$$

Plugging Equation (13) in Equation (12) , we get

$$\mathbf{E}(\hat{\boldsymbol{v}}) = \frac{1}{(1-2p)} \sum_{j=1}^{|C|} (p + (1-2p)a_j - p)\, \boldsymbol{c_j} = \sum_{j=1}^{|C|} a_j \boldsymbol{c_j} = \boldsymbol{v} \tag{14}$$

Now we show a bound on the variance of the estimate

$$\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = \mathbf{E}\left[\| \sum_{j=1}^{|C|} a_j \boldsymbol{c_j} - \frac{1}{(1-2p)} \sum_{j=1}^{|C|} (y_j - p)\, \boldsymbol{c_j} \|_2^2\right] \tag{15}$$

$$= \sum_{j=1}^{|C|} \mathbf{E}\left(a_j - \frac{(y_j - p)}{(1-2p)}\right)^2 |\boldsymbol{c_j}|^2 \tag{16}$$

(all the cross terms are 0 as they are mutually independent and $\mathbf{E}\left(a_j - \frac{y_j - p}{1-2p}\right) = 0$)

$$= \sum_{j=1}^{|C|} \mathrm{var}\left(\frac{y_j - p}{1-2p}\right) |\boldsymbol{c_j}|^2 \tag{17}$$

$$= \left(\frac{1}{1-2p}\right)^2 \sum_{j=1}^{|C|} \mathrm{var}(y_j)\, |\boldsymbol{c_j}|^2 = O(|C|R^2) \tag{18}$$

Equation (18) comes form the fact that y_j is a binary random variable and $Var(y_j) = Pr(y_j)(1 - Pr(y_j)) \leq \frac{1}{4}$ and $|\boldsymbol{c_j}|^2 \leq R^2$.

Privacy Now we show that our scheme is ϵ differentially private where ϵ is the input parameter to the RAPPOR algorithm. For any two points $\boldsymbol{v}, \boldsymbol{w} \in B_d(\mathbf{0}_d, 1)$,

$$\frac{PQ_{C,\epsilon}(\boldsymbol{v}) = y}{PQ_{C,\epsilon}(\boldsymbol{w}) = y} = \frac{\sum_{i=1}^{|C|} \Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i}) \Pr(Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\sum_{j=1}^{|C|} \Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j}) \Pr(Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \tag{19}$$

$$\leq \frac{\max_i \Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i}) \sum_{i=1}^{|C|} \Pr(Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\min_j \Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j}) \sum_{i=1}^{|C|} \Pr(Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \tag{20}$$

$$= \frac{\max_i \Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\min_j \Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \leq e^{\epsilon} \tag{21}$$

By the privacy property of RAPPOR Erlingsson et al. [2014] mechanism , we are using the following fact in equation (21)

$$\sup_{i,j} \frac{\Pr(y|Q_C(\boldsymbol{v}) = \boldsymbol{c_i})}{\Pr(y|Q_C(\boldsymbol{w}) = \boldsymbol{c_j})} \leq e^{\epsilon} \quad \forall \boldsymbol{v}, \boldsymbol{w}$$

Communication : Now we show that for the RAPPOR based scheme the expected communication is linear in $|C|$. Say y is the output when RAPPOR is applied to one hot encoded binary string. Without loss of generality say the the bit string is $\boldsymbol{e_i}$. The output y is generated as follows

$$\Pr(y_j = 1) = \begin{cases} p & \text{if } j \neq i \\ (1-p) & \text{if } j = i \end{cases}$$

So the expected sparsity (l_0 norm) of the output is

$$\mathbf{E}[\|y\|_0] = \sum_{i}^{|C|} y_i = (|C| - 1)p + (1 - p)$$
$$= |C|p + (1 - 2p) = O(|C|)$$

□

7 Experiments

We use our gradient quantization scheme to train a fully connected ReLU activated network with 1000 hidden nodes using the MNIST dataset LeCun and Cortes [2010] (60000 data points with label $[0 - 9]$). We use the *cross-entropy loss* function for the training the neural network with a total of $d = 795010$ parameters.

The dataset is divided equally among 100 workers. Each worker computes the local gradients and communicates the quantized gradient to the master which then aggregates and send the updated parameters. We plot the error at each iteration (Figure 2) and compare our results with QSGD quantization.

We use vqSGD with cross polytope scheme, $Q_{C_{cp}}$, along with the variance reduction technique with repetition parameter $s = 100$. Therefore, each local machine sends about $2060 = 100 \cdot \log(2d)$ bits whereas, QSGD requires about 8554 bits of communication per machine (computed by averaging over the total bits of communication over 3000 iterations) to communicate the quantized gradient. Our results indicate that vqSGD converges slightly faster compared to QSGD while communicating lesser bits.

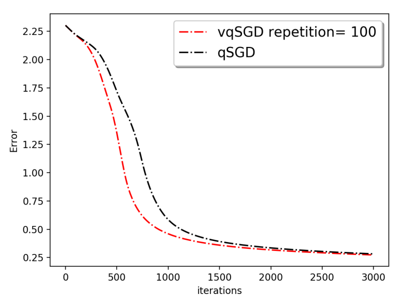


Figure 2: Convergence for fully connected ReLU network with 100 workers with QSGD and vqSGD.

Further experiments: We also use our gradient quantization scheme to solve the least squares problem and logistic regression for binary classification. We refer the reader to the supplementary material for further details regarding the performance of our schemes in comparison to the state of the art gradient quantization techniques.

8 Conclusion

We propose a general framework of convex-hull based private vector quantization schemes for distributed SGD that can be instantiated with any point set satisfying certain properties. The communication, variance and privacy tradeoffs for these mechanisms depend on the choice of point set. The proposed cross-polytope quantization scheme with low communication overhead is shown experimentally, to achieve convergence rates similar to the existing state-of-the-art quantization schemes which use orders of magnitudes more communication.

To further improve the tradeoffs between these parameters, we require constructions of point sets of size $\exp(o(d))$ with diameter $o(\sqrt{d})$. Information theoretically, we are asking the question of computing the variance of an unbiased estimator in terms of its unconditional entropy. While the Cramér-Rao bound provides a limit in terms of Fisher information, what we want is a rate-distorton type result for unbiased estimator Cover and Thomas [2012]. An information theoretic lower bound of this kind will be providing a better understanding of the limitations of the vqSGD approach.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

Jayadev Acharya, Chris De Sa, Dylan Foster, and Karthik Sridharan. Distributed learning with sublinear communication. In *International Conference on Machine Learning*, pages 40–50, 2019.

Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: Communication-efficient and differentially-private distributed SGD. In *Advances in Neural Information Processing Systems*, pages 7564–7575, 2018.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5973–5983, 2018.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimization for non-convex problems. In *International Conference on Machine Learning*, pages 560–569, 2018.

P Borjesson and C-E Sundberg. Simple approximations of the error function q(x) for communications applications. *IEEE Transactions on Communications*, 27(3):639–643, 1979.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 571–582, 2014.

Gérard Cohen, Iiro Honkala, Simon Litsyn, and Antoine Lobstein. *Covering codes*, volume 54. Elsevier, 1997.

Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51, 2016.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic Distributed Learning with Gradient Quantization and Variance Reduction. *arXiv e-prints*, art. arXiv:1904.05115, Apr 2019.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed SGD with sketching. In *Advances in Neural Information Processing Systems*, pages 13144–13154, 2019.

Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A Salman Avestimehr. Fitting ReLus via SGD and quantized SGD. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2469–2473. IEEE, 2019.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signSGD and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261, 2019.

Anastasiia Koloskova, Sebastian Urban Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. *Proceedings of Machine Learning Research*, 97(CONF), 2019.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error-correcting codes*, volume 16. Elsevier, 1977.

Prathamesh Mayekar and Himanshu Tyagi. Ratq: A universal fixed-length quantizer for stochastic optimization. *arXiv preprint arXiv:1908.08200 (a shorter version to appear at AISTATS 2020)*, 2019.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321. ACM, 2015.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3329–3337. JMLR. org, 2017.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, pages 9850–9861, 2018.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in neural information processing systems*, pages 1509–1519, 2017.

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.

A Further Experiments

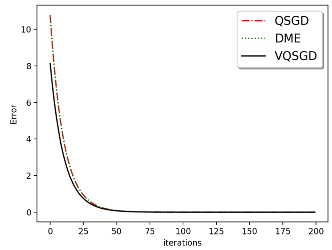
We experimentally show the performance of vqSGD using the cross polytope Q_{cp}, to solve the least squares problem and logistic regression for binary classification.

Least Squares: In the least square problem, we solve for $\boldsymbol{\theta^*} = \arg\min_{\boldsymbol{\theta}} \|A\boldsymbol{\theta} - \boldsymbol{b}\|_2^2$, where the matrix $A \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\theta^*} \in \mathbb{R}^d$ are generated by sampling each entry from $\mathcal{N}(0,1)$ and we set $\boldsymbol{b} = A\boldsymbol{\theta^*}$.

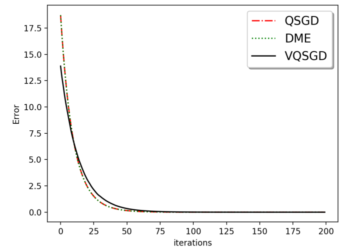
In order to show the performance of vqSGD, we simulate the iterations of distributed SGD with $n = 10000$ data samples distributed equally among $N = 500$ worker nodes. In every iteration of SGD, each worker node computes the local gradient on individual data batch and communicates the quantized version of the local gradient to the parameter server. The parameter server on receiving all the quantized gradients averages them and broadcasts the updated model to all the workers. The convergence of SGD is measured by the error term $\|\boldsymbol{\theta^*} - \boldsymbol{\theta_t}\|_2$, where $\boldsymbol{\theta_t}$ is the computed parameter at the end of t-th iteration of distributed SGD.

We compare the convergence of the least square problem for $d = 100, 200, 500$ against the state-of-the-art quantization schemes - DME Suresh et al. [2017] and QSGD Alistarh et al. [2017]. The results are presented in Figure 3.

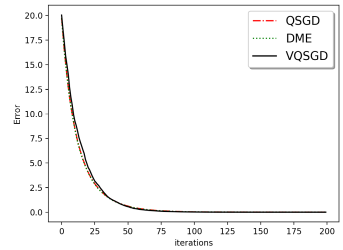
The results indicate that vqSGD achieves the same rate of convergence and accuracy as DME and QSGD while communicating only $\log(2d)$ bits and one real (l_2 norm of the vector form each server), whereas, DME (one bit stochastic quantization) and QSGD both require communication of about $\sqrt{d}$ bits and one real.



(a) d=100



(b) d=200



(c) d=500

Figure 3: Comparison of convergence for the least square problem with $d = 100, 200, 500$.

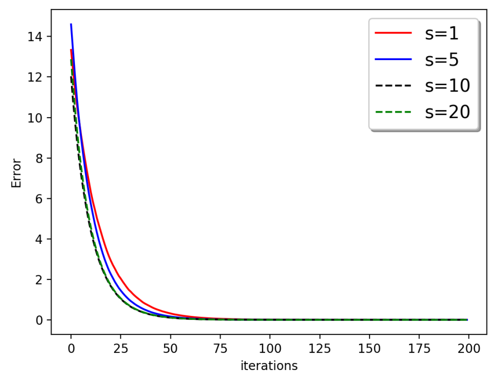


Figure 4: Convergence of $\boldsymbol{\theta_t}$ for $s = 1, 5, 10, 20$.
for least square problem

We show the improvement in the performance of vqSGD using the repetition technique for variance reduction. Recall that using repetition technique, each worker now sends s different indices instead of 1 which increases the communication to $s \log(2d)$ bits and 1 real. In Figure 4 we plot the convergence of the lease square problem with $d = 200$ with different values of $s = 1, 5, 10, 20$. We see the evident improvement in the convergence of vqSGD using this repetition scheme with increasing s.

Binary Classification: We compared the performance of vqSGD against DME and QSGD for the binary classification problem with logistic regression using various datasets from the UCI repository Chang and Lin [2011]. The logistic regression objective is defined as

$$\frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-b_i \boldsymbol{a_i}^T \boldsymbol{\theta}) + \frac{1}{2n}\|\boldsymbol{\theta}\|_2^2, \tag{22}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter, $\boldsymbol{a_i} \in \mathbb{R}^d$ is the feature data and $b_i \in \{-1, +1\}$ is its corresponding label.

We partition the data into 20 equal-sized batches, each assigned to a different worker node. We calculate the classification error for different (test) datasets after training the parameter in the distributed settings (same as described in least square problem). Results of the experiments are presented in Table 4, where each entry is averaged over 20 different runs.

Method	DME	QSGD	vqSGD
a5a ($d = 122$)	0.238 ± 0.0003	0.238 ± 0.0002	0.2368 ± 0.0029
a9a ($d = 123$)	0.234 ± 0.0003	0.234 ± 0.00017	0.234 ± 0.0015
gisset-scale ($d = 5000$)	0.0947 ± 0.00384	0.10475 ± 0.006	0.1480 ± 0.0174
splice ($d = 60$)	0.467 ± 0.017	0.4505 ± 0.0352	0.16618 ± 0.0054

Table 4: Comparison in classification error (mean$\pm$ standard deviation) for various UCI datasets

We note that for most datasets, with the exception of gisset-scale, vqSGD with $O(N \log d)$ bits of communication per iteration performs equally well or sometimes even better than QSGD and DME with $O(Nd)$ bits of communication per iteration.

B Missing Proofs from Section 4

Proof of Lemma 1.

$$\mathbf{E}[Q_C(\boldsymbol{v})] = \sum_{i=1}^{|C|} a_i \cdot \boldsymbol{c_i} = \tilde{\boldsymbol{v}} = \boldsymbol{v}.$$

□

Proof of Lemma 2. From the definition of the quantization function,

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{v} - Q_C(\boldsymbol{v})\|_2^2] &= \mathbf{E}[\|Q_C(\boldsymbol{v})\|^2] - \|\boldsymbol{v}\|^2 \\ &\leq R^2 - 1.\end{aligned}$$

This is true as C satisfies Condition (1) and therefore, each point $\boldsymbol{c_i} \in C$ has a bounded norm, $\|\boldsymbol{c_i}\| \leq R$. □

Proof of Theorem 3. Since $\hat{\boldsymbol{g}}$ is the average of N unbiased estimators, the fact that $\mathbf{E}[\hat{\boldsymbol{g}}] = \boldsymbol{g}$ follows from Lemma 1. For the variance computation, note that

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{g} - \hat{\boldsymbol{g}}\|_2^2] &= \frac{1}{N^2}\left(\sum_{i=1}^{N} \mathbf{E}[\|\boldsymbol{g_i} - \hat{\boldsymbol{g}}_{\boldsymbol{i}}\|_2^2]\right) && \text{(since } \hat{\boldsymbol{g}}_{\boldsymbol{i}} \text{ is an unbiased estimator of } \boldsymbol{g} \text{)} \\ &\leq \frac{(R^2 - 1)}{N^2} \sum_{i=1}^{N} \|g_i\|^2 && \text{(from Lemma 2).}\end{aligned}$$

□

Proof of Proposition 4. The proof follows simply by linearity of expectations.

$$\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = \mathbf{E}\left[\|\frac{1}{s}\sum_{i=1}^{s}(\boldsymbol{v} - Q_C(\boldsymbol{v}))\|^2\right] = \frac{1}{s}\cdot(R^2 - 1) \qquad \text{(from Lemma 2).}$$

□

C Missing Proofs from Section 5

Proof of Proposition 8. The proof of Proposition 8 follows directly from Lemma 2 provided the point set C_{cp} satisfies Condition (1) with $R = \sqrt{d}$. We will now prove this fact.

Since each vertex is of the form $\pm\sqrt{d}\boldsymbol{e_i}$, it follows that all the vertices of $\textsc{conv}(C_{cp})$, and hence the entire convex hull lies inside a ball of radius $\sqrt{d}$, *i.e.*, , $\textsc{conv}(C_{cp}) \subset B_d(\mathbf{0}_d, \sqrt{d})$.

To prove that the unit ball is contained in the convex hull $\textsc{conv}(C_{cp})$, we pick any arbitrary point $v \in B_d(\mathbf{0}_d, 1)$ and show that it can written as a convex combination of points in C_{cp}. The fact follows from the solution to the system of linear equations (2) given in Equation (3). Note that the solution satisfies $a_i \geq 0$ and $\sum_i a_i = 1$ for any point $\boldsymbol{v} \in B_d(0, 1)$. □

Proof of Lemma 9. Let $K := \textsc{conv}(N(\varepsilon))$ be the convex hull of the ε-net points of the unit sphere. Let $B_d(\mathbf{0}_d, r)$ be the inscribed ball in K for some $r < 1$. We show that $r \geq 1 - \varepsilon$.

Consider the face of K that is tangent to $B_d(\mathbf{0}_d, r)$ at point $\boldsymbol{z}$. We will show that $\|\boldsymbol{z}\|_2 \geq 1 - \varepsilon$. Extend the line joining $(\mathbf{0}_d, \boldsymbol{z})$ to meet $\mathcal{S}^{d-1}$ at point $\boldsymbol{x}$. Since $\boldsymbol{x} \in \mathcal{S}^{d-1}$, we know that there exists a net point $\boldsymbol{u}$ at a distance of at most ε from it. Therefore, the distance of $\boldsymbol{x}$ from K is upper bounded by ε, *i.e.*, $\|\boldsymbol{x} - \boldsymbol{z}\| \leq \|\boldsymbol{x} - \boldsymbol{u}\| \leq \varepsilon$. Therefore $\|\boldsymbol{z}\| = 1 - \|\boldsymbol{x} - \boldsymbol{z}\| \geq 1 - \varepsilon$.

Therefore scaling all the points of $N(\varepsilon)$ by any $R \geq \frac{1}{1-\varepsilon}$ we see that $B_d(\mathbf{0}_d, 1) \subseteq \textsc{conv}(C)$. □

D Missing Proofs from Section 6

Proof of Proposition 11. First we show that the point set C_S satisfies Condition (1) with $R = 2d$. The fact that $\textsc{conv}(C_S) \subset B_d(\mathbf{0}_d, 2d)$ follows trivially from the observation that each point in $C_S \in B_d(\mathbf{0}_d, 2d)$.

To show that $B_d(\mathbf{0}_d, 1) \subset \textsc{conv}(C_S)$, consider any face of the convex hull, $F_{\boldsymbol{c}} := \textsc{conv}(C_S \setminus \{\boldsymbol{c}\})$, for some $\boldsymbol{c} \in C_S$. We show that $F_{\boldsymbol{c}}$ is at an ℓ_2 distance of at least 1 from $\mathbf{0}_d$. This in turn shows that any point outside the convex hull must be outside the unit ball as well.

First consider the case when $\boldsymbol{c} = -4\mathbf{1}_d$. We observe that the face $F_{\boldsymbol{c}}$ is contained in the hyperplane $H_{\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^d \mid \frac{1}{\sqrt{d}}\mathbf{1}_d^T\boldsymbol{x} = 2\sqrt{d}\}$, and therefore is at a distance of $O\left(\sqrt{d}\right)$ from the origin.

Now consider the case when $\boldsymbol{c} = 2d\ \boldsymbol{e_1}$. Let $\boldsymbol{w} = \frac{2}{\sqrt{\frac{1}{4}+\frac{1}{d}}}(-\frac{1}{4} + \frac{1}{2d}, \frac{1}{2d}, \ldots, \frac{1}{2d})^T \in \mathbb{R}^d$ be a unit vector. We note that $F_{\boldsymbol{c}} \subset H_{\boldsymbol{c}}$, where $H_{\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^d \mid \boldsymbol{w}^T\boldsymbol{x} = \frac{2}{\sqrt{\frac{1}{4}+\frac{1}{d}}}\}$ is the hyperplane defined by the unit normal vector $\boldsymbol{w}$ that is at a distance of at least 1 from $\mathbf{0}_d$.

Since all other faces are symmetric, the proof for the case $\boldsymbol{c} = 2d\ \boldsymbol{e_i}, i \in [d]$ follows similarly.

Privacy: We now show that the quantization scheme is ϵ-differentially private for any $\epsilon > \log 7$. From the definition of ϵ-DP, it is sufficient to show that for any $\boldsymbol{x}, \boldsymbol{y} \in B_d(\mathbf{0}_d, 1)$, and any $\boldsymbol{c} \in C_S$,

$$\frac{\Pr[Q_{C_S}(\boldsymbol{x}) = \boldsymbol{c}]}{\Pr[Q_{C_S}(\boldsymbol{y}) = \boldsymbol{c}]} \leq 7.$$

Since $\boldsymbol{x}, \boldsymbol{y} \in \textsc{conv}(C_S)$, we can express them as the convex combination of points in C_S. Let $\boldsymbol{x} = \sum_{\boldsymbol{c} \in C_S} a_c^{(\boldsymbol{x})}$. Similarly, let $\boldsymbol{y} = \sum_{\boldsymbol{c} \in C_S} a_c^{(\boldsymbol{y})}$. Then, from the construction of the quantization function Q_{C_S}, we know that

$$\frac{\Pr[Q_{C_S}(\boldsymbol{x}) = \boldsymbol{c}]}{\Pr[Q_{C_S}(\boldsymbol{y}) = \boldsymbol{c}]} = \frac{a_c^{(\boldsymbol{x})}}{a_c^{(\boldsymbol{y})}}.$$

We now show that the ratio $\frac{a_c^{(\boldsymbol{x})}}{a_c^{(\boldsymbol{y})}}$ is at most 7 for any pair $\boldsymbol{x}, \boldsymbol{y} \in B_d(\mathbf{0}_d, 1)$ and any $\boldsymbol{c} \in C_S$. The privacy bound follows from this observation.

First, consider the case $\boldsymbol{c} = -4\mathbf{1}_d$. From the closed form solution for any $\boldsymbol{x} \in \operatorname{CONV}(C_S)$ described in Equation (6), we know that $a_c^{(\boldsymbol{x})} = \frac{1}{3} - \frac{\sum_{i=1}^d x_i}{6d}$. For any $\boldsymbol{x} \in B_d(\mathbf{0}_d, 1)$, $\sum_{i=1}^d x_i \in [-\|\boldsymbol{x}\|_1, \|\boldsymbol{x}\|_1] \subseteq \left[-\sqrt{d}, \sqrt{d}\right]$. Therefore, $a_c^{(\boldsymbol{x})} \in \left[\frac{1}{3} - \frac{1}{6\sqrt{d}}, \frac{1}{3} + \frac{1}{6\sqrt{d}}\right]$. It then follows that for any $\boldsymbol{x}, \boldsymbol{y} \in B_d(\mathbf{0}_d, 1)$ and $\boldsymbol{c} = -4\mathbf{1}_d$,

$$\frac{a_c^{(\boldsymbol{x})}}{a_c^{(\boldsymbol{y})}} \leq \frac{\frac{1}{3} + \frac{1}{6\sqrt{d}}}{\frac{1}{3} - \frac{1}{6\sqrt{d}}} = 1 + \frac{2}{2\sqrt{d} - 1} \leq 3$$

Now we consider the case when $\boldsymbol{c} = 2d\ \boldsymbol{e_1}$. Then from the closed from solution in Equation (6), we get that for any $\boldsymbol{x} \in \operatorname{CONV}(C_S)$ the coefficient $a_c^{(\boldsymbol{x})} = \frac{x_1}{2d}\left(1 - \frac{2}{3d}\right) - \frac{\sum_{i=2}^d x_i}{3d^2} + \frac{2}{3d}$. Note that this quantity is maximized for $\boldsymbol{x} = \boldsymbol{e_1}$ and minimized for $\boldsymbol{x} = -\boldsymbol{e_1}$. Therefore the ratio for any $\boldsymbol{x}, \boldsymbol{y} \in B_d(\mathbf{0}_d, 1)$ and $\boldsymbol{c} = 2d\ \boldsymbol{e_1}$ is at most

$$\frac{a_c^{(\boldsymbol{x})}}{a_c^{(\boldsymbol{y})}} \leq \frac{7d - 2}{d + 2} \leq 7$$

The ratio for all other vertices can be computed in a similar fashion and is bounded by the same quantity. □

Proof of Proposition 12. First, we show that C_H satisfies Condition (1) with $R = 2d$. The fact that $\operatorname{CONV}(C_H) \subset B_d(\mathbf{0}_d, 2d)$ is trivial and follows since every point in C_H is contained in $B_d(\mathbf{0}_d, 2d)$.

To show that $B_d(\mathbf{0}_d, 1) \subset C_H$, consider any $\boldsymbol{x} \in B_d(\mathbf{0}_d, 1)$, and the closed form solution for the coefficients a_i given by Equation (7). We now show that these coefficients indeed give a convex combination. Note that $a_i := \frac{1}{d+1}\left(1 + \frac{\boldsymbol{c}^T\boldsymbol{x}}{4d}\right) \geq 0$. This holds since $\boldsymbol{c^T x} \geq \|\boldsymbol{c}\|\|\boldsymbol{x}\| \geq -2d$. Moreover, from the property of Hadamard matrices,

$$\sum_{i=1}^{d+1} a_i = \frac{1}{d+1}\begin{bmatrix}1 & \dots & 1\end{bmatrix} H_p^T \begin{bmatrix} 1 \\ \frac{\boldsymbol{x}}{2\sqrt{d}} \end{bmatrix} = 1.$$

The last equality follows from the following property of the Hadamard matrices that can be proved using induction.

$$\begin{bmatrix}1 & \dots & 1\end{bmatrix} H_p^T = \begin{bmatrix}2^p & 0 & \dots & 0\end{bmatrix}.$$

Therefore, any $\boldsymbol{x} \in B_d(\mathbf{0}_d, 1)$ can be expressed as a convex combination of the points in C_H, *i.e.*, , $\boldsymbol{x} = \sum_{i=1}^{d+1} a_i \boldsymbol{c_i}$, for $\boldsymbol{c_i} \in C_H$.

Privacy: We now show that the quantization scheme is ϵ-differentially private for any $\epsilon > \frac{1}{\sqrt{d}}$. From the definition of ϵ-DP, it is sufficient to show that for any $\boldsymbol{x}, \boldsymbol{y} \in B_d(\mathbf{0}_d, 1)$, and any $\boldsymbol{c} \in C_H$,

$$\frac{\Pr[Q_{C_H}(\boldsymbol{x}) = \boldsymbol{c}]}{\Pr[Q_{C_S}(\boldsymbol{y}) = \boldsymbol{c}]} \leq 1 + \sqrt{2}$$

Since $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{CONV}(C_S)$, we can express them as the convex combination of points in C_H. Let $\boldsymbol{x} = \sum_{\boldsymbol{c} \in C_H} a_c^{(\boldsymbol{x})}$. Similarly, let $\boldsymbol{y} = \sum_{\boldsymbol{c} \in C_H} a_c^{(\boldsymbol{y})}$. Then, from the construction of the quantization function Q_{C_H}, we know that

$$\frac{\Pr[Q_{C_H}(\boldsymbol{x}) = \boldsymbol{c}]}{\Pr[Q_{C_H}(\boldsymbol{y}) = \boldsymbol{c}]} = \frac{a_c^{(\boldsymbol{x})}}{a_c^{(\boldsymbol{y})}}. \tag{23}$$

From the closed form solution in Equation (7), we know that for any $\boldsymbol{x} \in \textsc{conv}(C_H)$, the coefficient of $\boldsymbol{c}$ in the convex combination of $\boldsymbol{x}$ is given by $a_c^{(\boldsymbol{x})} = \frac{1}{d+1}\left(1 + \frac{\boldsymbol{c}^T\boldsymbol{x}}{4d}\right)$. Plugging this in Equation (23), we get

$$\frac{\Pr[Q_{C_H}(\boldsymbol{x}) = \boldsymbol{c}]}{\Pr[Q_{C_H}(\boldsymbol{y}) = \boldsymbol{c}]} = \frac{a_c^{(\boldsymbol{x})}}{a_c^{(\boldsymbol{y})}} = \frac{1 + \frac{\boldsymbol{c}^T\boldsymbol{x}}{4d}}{1 + \frac{\boldsymbol{c}^T\boldsymbol{y}}{4d}} = 1 + \frac{\frac{\boldsymbol{c}^T(\boldsymbol{x}-\boldsymbol{y})}{4d}}{1 + \frac{\boldsymbol{c}^T\boldsymbol{y}}{4d}} \tag{24}$$

$$\leq 1 + \frac{\|\boldsymbol{c}\|_2\|\boldsymbol{x}-\boldsymbol{y}\|_2}{4d - \|\boldsymbol{c}\|_2} \qquad \text{for } \boldsymbol{y} = -\frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2} \tag{25}$$

$$\leq 1 + \frac{2\sqrt{2}d}{4d - 2d} \tag{26}$$

$(mbox since \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \sqrt{2}$ and $\|\boldsymbol{c}\|_2 = 2d.)$

$$= 1 + \sqrt{2} \tag{27}$$

This concludes the proof of Proposition 12. □