vqSGD: Vector Quantized Stochastic Gradient Descent

Venkata Gandikota
Syracuse University
vsgandik@syr.edu

Daniel Kane
UC San Diego
dakane@ucsd.edu

Raj Kumar Maity
UMass Amherst
rajkmaity@cs.umass.edu

Arya Mazumdar
UC San Diego
arya@ucsd.edu

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 first-order distributed optimization. In the process we derive the following fundamental information theoretic fact: $\Theta(\frac{d}{R^2})$ bits are necessary and sufficient (up to an additive $O(\log d)$ term) to describe an unbiased estimator $\hat{\boldsymbol{g}}(\boldsymbol{g})$ for any $\boldsymbol{g}$ in the d-dimensional unit sphere, under the constraint that $\|\hat{\boldsymbol{g}}(\boldsymbol{g})\|_2 \leq R$ almost surely. 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 almost surely bounded norm. We provide multiple efficient instances of our scheme, that are near optimal, and require only $o(d)$ bits of communication at the expense of tolerable increase in error. The instances of our quantization scheme are obtained using the properties of binary error-correcting codes and provide a smooth trade-off between the communication and the estimation error of quantization. Furthermore, we show that *vqSGD* also offers some automatic 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 [22]), communication becomes a *bottleneck* to the efficiency and speed of learning using SGD [11].

In the recent years various quantization and sparsification techniques [2, 5, 7, 21, 26, 29, 33, 35, 37] have been developed to alleviate the problem of communication bottleneck. Recently, [19] 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 [13, 14] 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. We provide explicit and efficient quantization schemes based on convex hull of specific structured point sets in $\mathbb{R}^d$ that require $O(d \log d/R^2)$

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

bits to communicate an unbiased gradient estimate that has variance bounded above by R^2: this is within a $\log d$ factor of the optimal amount of communication that is necessary and sufficient for this purpose.

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. We provide matching upper and lower bounds on this communication cost.

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) [36] and RAPPOR [15] 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 this 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* (summarized in Table 1) are the first quantization schemes in literature to achieve privacy directly through quantization. While our randomized construction is optimal in terms of communication, the explicit schemes are within $\log d$ factor of a lower bound that we provide.

Empirically we compare our quantization schemes to the state-of-art schemes [5, 33]. 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.

While differential privacy for gradient based algorithms [1, 30] were considered earlier in literature, cpSGD [4] 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 [33] 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 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 [36] 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 [4].

Organization: In Section 2, we describe some other related work on communication efficiency in the federated learning setup. We start in Section 3 describing the settings for our results. The vqSGD quantization scheme is presented in Section 4. In Section 5 we provide a handle to test whether a 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, which matches a lower bound we prove. We provide a few structured deterministic constructions of point sets in this section as well. Section 6 emphasizes the privacy component of vqSGD - and derives the privacy parameters of several vqSGD schemes. Due to space constraint, the section containing experiments (Section A) is delegated to the appendix in the supplementary material. Missing proofs of all theorems/lemmas, and some further schemes can be found in the supplementary material.

2 Related Work

The foundations of gradient quantization was laid by [27] and [32] with schemes that require the compute nodes to send exactly 1-bit per coordinate of the gradi-

*Note that ϵ denotes the privacy parameter and ε refers to the packing parameter of ε-nets.

†O_ε hides terms involving ε

‡O_δ hides terms involving δ.

Point set	Error	Communication (bits)	Privacy	Efficiency
Gaussian-Sampling (Theorem 7) for any $c > \log(d)$	$\frac{d}{cN}$	Nc	-	$O(\exp(c))$
Reed-Muller (C_{RM}) (Proposition 8)	$\frac{d}{N}$	$N \log 2d$	-	$O(d)$
Cross-polytope (C_{cp}) (Proposition 9)	$\frac{d}{N}$	$N \log 2d$	$\epsilon > O(\log d)$	$O(d)$
Scaled ε-Net (C_{net}) (Proposition 11)	$\frac{1}{N}$	$O_\varepsilon(Nd)^\dagger$	-	$O\left(\left(\frac{1}{\varepsilon}\right)^d\right)$
Simplex (C_S) (Proposition 12)	$\frac{d^2}{N}$	$N \log(d+1)$	$\epsilon > \log 7$	$O(d)$
Hadamard (C_H) (Proposition 13)	$\frac{d^2}{N}$	$N \log d$	$\epsilon > \log(1+\sqrt{2})$	$O(d)$
Cross-polytope (C_{cp}) + RR (Theorem 15)	$\frac{d^2}{N}$	$N \log(2d)$	$\epsilon > 0$	$O(d)$
Cross-polytope (C_{cp}) + RAPPOR (Theorem 16 in supplement)	$\frac{d^2}{N}$	$2Nd$	$\epsilon > 0$	$O(d)$

Table 1: List of results (per-iteration of SGD). (N: number of worker nodes, d: dimension).

Method	Error	Comm
QSGD [5]	$\min\{\frac{d}{s^2}, \frac{\sqrt{d}}{s}\}\frac{1}{N}$	$Ns(s+\sqrt{d})$
DME [33]	$\min\{\frac{1}{Ns}, \frac{\log d}{N(s-1)^2}\}$	Nsd
vqSGD $Q_{C_{cp}}$	$\frac{d}{Ns}$	$Ns \log d$
Gaussian	$\frac{d}{Nsc}$	Nsc

Table 2: Comparison of non private quantization schemes.

Method	Error	Comm	DP (ϵ)
cpSGD [4]	$O_\delta\left(\frac{d}{N}\right)^\ddagger$	$O_\delta(d)$	$\delta > 0$, $\epsilon > f(\delta)$
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. (N: number of worker nodes, s, c: tuning parameter (≥ 1))

ent. 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: [5, 35, 37] 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 per compute node 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 [33]. The quantization schemes proposed in [33] 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 [39], they were unable to break the $\sqrt{d}$ barrier of communication cost (per compute node). Moreover, the schemes proposed in [5, 33] are variable length codes that achieve low communication in expectation. The worst case communication cost could be higher. In a parallel work [26], the authors propose an efficient fixed-length quantization scheme that achieves near-optimal convergence with T-rounds of SGD. However, the goal of their work is different from ours, and the methodologies are different as well.

In this work, we propose (fixed length) quantization schemes that require $o(d)$ (as low as $\log d$) bits of communication per iteration of SGD, and are almost optimal as well. In fact for any c-bits of communication, the quantization scheme with Gaussian points achieves a variance of $O(d/c)$ that meets the lower bounds for any unbiased quantization scheme (also shown in the current work).

Gradient sparsification techniques with provable convergence (under standard assumptions) were studied in [2, 6, 18, 31]. 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. [29] 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.

Error Feedback: Many works [17, 20] focused on providing techniques to reduce the error incurred due to quantization using locally accumulated errors. In this work, we focus on gradient quantization techniques, and note that the variance reduction techniques of [17] can be used on top of the proposed quantization.

3 Preliminaries

Let $[n]$ denote the set $\{1, 2, \ldots, n\}$ and let $\mathbf{1}_d$, $\mathbf{0}_d$ denote the all 1's vector and all 0's vector in $\mathbb{R}^d$ respectively. 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, for any prime power q, let $\mathbb{F}_q$ denote a finite field with q elements.

For a discrete set of points $C \subset \mathbb{R}^d$, let $\text{CONV}(C)$ denote the convex hull of points in C, *i.e.*, $\text{CONV}(C) := \left\{\sum_{\boldsymbol{c}\in C} a_c \boldsymbol{c} \mid a_c \geq 0, \sum_{\boldsymbol{c}\in C} a_c = 1\right\}$.

Suppose $\boldsymbol{w} \in \mathbb{R}^d$ 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 $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta\hat{\boldsymbol{g}}$, 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 $\boldsymbol{w}$. The assumption of unbiasedness is crucial here, that implies $\mathbf{E}\hat{\boldsymbol{g}} = \boldsymbol{g}$.

The goal of any gradient quantization scheme is to reduce cost of communicating the gradient, i.e., to act as an first-order oracle, 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. In this work, we will develop a scheme that is an *almost surely bounded oracle* for gradients, i.e., $\|\hat{\boldsymbol{g}}\|_2^2 \leq B$ with probability 1, for some $B > 0$. The convergence rate of the SGD algorithm for any convex function f depends on the upper bound of the norm of the unbiased estimate, i.e., B, cf. any standard textbook such as [28].

Although we provide an almost surely bounded oracle as our quantization scheme, previous quantization schemes, such as [5], provides a *mean square bounded oracle*, i.e., an unbiased estimate $\hat{\boldsymbol{g}}$ of $\boldsymbol{g}$ such that $\mathbf{E}\|\hat{\boldsymbol{g}}\|_2^2 \leq B$ for some $B > 0$. It is known that, even with a mean square bounded oracle, SGD algorithm for a convex function converges with dependence on the upper bound B (see [9]). As discussed in [5], one can also consider the variance of $\hat{\boldsymbol{g}}$ without any palpable difference in theory or practice. Therefore, below we consider the variance of the estimate $\hat{\boldsymbol{g}}$ as the main measure of error.

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 $i \in \{1, \ldots, N\}$. For $\boldsymbol{g} = \frac{1}{N}\sum_i \boldsymbol{g_i}$, the variance of the estimate $\hat{\boldsymbol{g}} = \frac{1}{N}\sum_i \hat{\boldsymbol{g}}_{\boldsymbol{i}}$ is defined as

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

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 $\text{Var}(\hat{\boldsymbol{g}})$ is minimized.

For the privacy preserving gradient quantization schemes, we consider the standard notion of (ϵ, δ)-differential privacy (DP) as defined in [14]. 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 $(1 + 2/\varepsilon)^d$ [12].

Definition 3 (Hadamard Matrix)**.** *A Hadamard matrix H_n of order n is a $n \times n$ square matrix with entries from ± 1 whose rows are mutually orthogonal. Therefore, it satisfies $HH^T = nI_n$, where I_n is the $n \times n$ identity matrix.*

Sylvester's construction [16] provides a recursive technique to construct Hadamard matrices for orders that are powers of 2 which can be defined as follows. Let $H_1 = [1]$ be the Hadamard matrix of order 2^0, and let H_p denote the Hadamard matrix of order 2^p, then a Hadamard matrix of order 2^{p+1} can be constructed as

$$H_{p+1} = \begin{bmatrix} H_p & H_p \\ H_p & -H_p \end{bmatrix}$$

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 $C = \{\boldsymbol{c_1}, \ldots, \boldsymbol{c}_m\} \subset \mathbb{R}^d$ be a discrete set of points such that its convex hull, $\text{CONV}(C)$ satisfies

$$B_d(\mathbf{0}_d, 1) \subset \text{CONV}(C) \subseteq B_d(\mathbf{0}_d, R), R > 1. \quad (1)$$

Let $\boldsymbol{v} \in B_d(\mathbf{0}_d, 1)$. Since $B_d(\mathbf{0}_d, 1) \subseteq \text{CONV}(C)$, 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. $\mathbf{E}[Q_C(\boldsymbol{v})] = \boldsymbol{v}$.

We assume that C is fixed in advance and is known to the compute nodes and the parameter server.

Remark 1. *Communicating the quantization of any vector $\boldsymbol{v}$, amounts to sending a floating point number $\|\boldsymbol{v}\|_2$, 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. In such settings 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, $\|\hat{\boldsymbol{v}}\|_2^2 \leq R^2$ almost surely, and $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] \leq R^2$.*

Remark 2. *If, for any vector $\boldsymbol{v}$, we send the floating point number $\|\boldsymbol{v}\|_2$ separately, instead of there being a known upper bound on gradient, we can just assume without loss of generality that $\boldsymbol{v} \in S^{d-1}$. In this case, the subsequent bounds on variance $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = \mathbf{E}\left[\|\hat{\boldsymbol{v}}\|_2^2\right] - \|\boldsymbol{v}\|_2^2$ can be replaced by $R^2 - 1$.*

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 node, Define $\hat{\boldsymbol{g}} := \frac{1}{N}\sum_{i=1}^N \hat{\boldsymbol{g}}_i$, where $\hat{\boldsymbol{g}}_i := \|\boldsymbol{g_i}\| \cdot Q_C(\boldsymbol{g_i}/\|\boldsymbol{g_i}\|)$. Then,*

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

Remark 3. *Computing the quantization $Q_C(.)$ amounts to solving a system of $d+1$ linear equations in $\mathbb{R}^{|C|}$. For general point sets C, this takes about $O(|C|^3)$ time (since $|C| \geq d+1$). However, we show that for certain structured point sets, the quantization $Q_C(.)$ can be computed in linear time.*

Note that Theorem 3 holds even when $\boldsymbol{g_i}$ is an unbiased stochastic gradient and not the true local gradient. The quantization error is then computed with respect to this estimate of the local gradient.

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.

Also note that the convex combination computed for a particular vector $\boldsymbol{v}$ with respect to the point set C need not be unique and *any* convex combination using the point set C gives an unbiased estimate of the vector $\boldsymbol{v}$. However, we will later see that the appropriate choice of convex combination dictates the privacy guarantees of the quantization scheme.

Reducing Variance: In this section, we propose a simple repetition technique to reduce the variance of the quantization scheme. For any $s > 1$, let $Q_C(s, \boldsymbol{v}) := \frac{1}{s}\sum_{i=1}^s Q_C^{(i)}(\boldsymbol{v})$ 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. Using this repetition technique, 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 $\hat{\boldsymbol{v}} := Q_C(s, \boldsymbol{v})$. Then, $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] \leq R^2/s$.*

5 Constructions of Point Sets and Lower Bound

In this section, we propose constructions of point sets that satisfy Condition (1) and provide varying tradeoffs

between communication and variance of the quantization scheme. But first, we start with a lower bound that shows that one must communicate $\Omega(\frac{d}{R^2})$ bits to achieve an error of $O(R^2)$ in the estimate of the gradient as per Condition (1).

Theorem 5. *Let $C \subseteq \mathbb{R}^d$ be a discrete set of points that satisfy Condition* (1)*. Then*

$$|C| \geq \exp(\alpha d/R^2)$$

for some absolute constant $\alpha > 0$.

To prove the lower bound, 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 6. *Let $C = \{c_1, \ldots, c_m\} \subseteq \mathbb{R}^d$ be a discrete set of points. The unit ball $B_d(\mathbf{0}_d, 1) \subseteq \text{CONV}(C)$ if and only if for any $\boldsymbol{x} \in S^{d-1}$, there exists a point $\boldsymbol{c} \in C$ such that $\langle \boldsymbol{x}, \boldsymbol{c} \rangle \geq 1$.*

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

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

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

Proof of Theorem 5. The proof of this theorem will use a packing argument for S^{d-1}. Let $\boldsymbol{c} \in C$. We will estimate the *cardinality* of the set $P(\boldsymbol{c}) := \{\boldsymbol{x} \in S^{d-1} : \langle \boldsymbol{x}, \boldsymbol{c} \rangle \geq 1\}$ under the uniform measure over S^{d-1}. Using Theorem 6, size of C must be at least

$$\frac{\text{area}(S^{d-1})}{\max_{\boldsymbol{c} \in C} \text{area}(P(\boldsymbol{c}))}$$

for it to satisfy Condition (1).

Note that, $P(\boldsymbol{c})$ is a hyperspherical cap with angle ϕ such that $\cos\phi \geq \frac{1}{\|\boldsymbol{c}\|} \geq \frac{1}{R}$, since C satisfy Condition (1). The area of a cap can be computed using the incomplete beta functions, however a probabilistic argument below will serve to lower bound this.

If we uniformly at random choose a vector $\boldsymbol{z}$ from S^{d-1}, then the probability p that it is within an angular distance ϕ of a fixed unit vector, $\boldsymbol{u}$, will exactly be the the ratio of the areas of the hyperspherical cap and the sphere. Again this probability is known to follow a shifted Beta distribution, but we can estimate it from above using concentration bounds.

Since the area of the hyperspherical cap is invariant to its center, we can take $\boldsymbol{u}$ to be the first standard basis vector. It is known that if $\boldsymbol{g} = (g_1, g_2, \ldots, g_d) \in \mathbb{R}^d$ is a random vector with i.i.d. Gaussian $\mathcal{N}(0, 1)$ entries, then $\boldsymbol{z} := \boldsymbol{g}/\|\boldsymbol{g}\|$ is uniform over S^{d-1}. Therefore,

$$\begin{aligned}
&\Pr(\langle \boldsymbol{z}, \boldsymbol{u} \rangle \geq 1/R) = \Pr(g_1/\|\boldsymbol{g}\| \geq 1/R) \\
&\leq \Pr(g_1 \geq \|\boldsymbol{g}\|/R \mid \|\boldsymbol{g}\| \geq \sqrt{d}/4) + \Pr(\|\boldsymbol{g}\| < \sqrt{d}/4) \\
&\leq \Pr(g_1 \geq \sqrt{d}/4R) + \Pr(\|\boldsymbol{g}\| < \sqrt{d}/4).
\end{aligned}$$

Now since g_1 is $\mathcal{N}(0, 1)$, $\Pr(g_1 \geq \frac{\sqrt{d}}{4R}) \leq \exp(-\frac{d}{32R^2})$, from Chernoff bound. On the other hand $\|\boldsymbol{g}\|^2$ is a χ^2 distribution of d degrees of freedom. Since that is subexponential, we have $\Pr(\|\boldsymbol{g}\| < \sqrt{d}/4) \leq \Pr(\|\boldsymbol{g}\|^2 < d/16) \leq \exp(-\frac{225d}{2048}) \leq \exp(-\frac{d}{32R^2})$ for any $R \geq 1$.

This implies, $|C| \geq (2\exp(-d/(32R^2)))^{-1}$. □

5.1 Gaussian point set

We provide a randomized construction of point set using the characterization defined above, that is optimal in terms of communication.

Theorem 7. *Let $R \in [5, 6\sqrt{d}]$. There exists a set C of $\exp(O(d/R^2 + \log d))$ points of ℓ_2 norm at most R each, that satisfy Condition* (1)*.*

The above stated theorem provides a randomized algorithm to generate a point set of size about $\exp(\Theta(d/R^2))$ such that the quantization scheme defined in Section 4 instantiated with this point set achieves a variance of $O(R^2)$ while communicating $\tilde{O}(d/R^2)$ bits, hence meeting the lower bound of Theorem 5. In particular, there exists a quantization scheme that achieves $O(1)$ variance with $\tilde{O}(d)$ bits of communication (see supplementary material for a deterministic construction). Also, at the 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 Sec. 5.2, and also Q_{cp} in the supplement), 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 Condition 1. We assume d to be a power of 2, *i.e.*, $d = 2^p$ for some $p \geq 1$.

Our quantization scheme is based on the first order Reed-Muller codes, $\text{RM}(1, p)$ ([25]). Each codeword of $\text{RM}(1, p)$ is given as the evaluations 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 C_{RM} denote this set of mapped codewords.

We show that the set of points in C_{RM} satisfy the characterization of Theorem 6, and therefore will give us a quantization scheme with $\log 2d$ communication and the following guarantees:

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

Remark 4. *Instead of first order Reed-Muller codes, we can use any binary linear code $\mathrm{C} \subseteq \mathbb{F}_2^d$ to construct the point set as follows. Map all the codewords from $\mathbb{F}_2^d$ to $\mathbb{R}^d$ using 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 $\log(2|\mathrm{C}|)$, where $|\mathrm{C}|$ denotes the number of codewords in C. In this regard, the first order Reed-Muller codes described above provide the best communication guarantees and the quantization is also efficiently computable.*

5.3 Other Deterministic Constructions

We now present several explicit constructions of point sets that give quantization schemes with varying trade-offs. 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 $O(d)$. While 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 9 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

$$\begin{bmatrix}\sqrt{d}I_d & -\sqrt{d}I_d\end{bmatrix}\begin{bmatrix}a_1\\ \vdots \\ a_{2d}\end{bmatrix} = \begin{bmatrix}v_1\\ \vdots \\ v_d\end{bmatrix} \quad \text{s.t.} \sum_{i=1}^{2d} a_i = 1, \quad (2)$$

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:

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

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 9. *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 $Q_{C_{cp}}$ 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 ([2]).

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. We show that an appropriate constant scaling of an ε-net points (see Definition 2) satisfies Condition (1).

Lemma 10. *For any $0 < \varepsilon < 1$, let $R = \frac{1}{1-\varepsilon}$. 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 vqSGD with point set C_{net}. From Lemma 2, we then get the following guarantees for the quantization scheme obtained from scaled ε-nets, C_{net} for some constant $\varepsilon < 1$.

Proposition 11. *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 $\mathbf{E}\left[\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2\right] = \frac{1}{(1-\varepsilon)}$.*

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. First, we see why the quantization scheme described in Section 4 is 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 points in C that are in the range of $Q_C(\boldsymbol{x})$.

In order for Q_C to be ϵ-DP for any $\epsilon > \epsilon_0$, we have to show for gradients $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ of any two neighboring datasets and for any $\boldsymbol{z} \in \text{SUPP}(\boldsymbol{x}, C) \cup \text{SUPP}(\boldsymbol{y}, C)$,

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

If $|C| > d+1$, there may exist two gradients $\boldsymbol{x}, \boldsymbol{y} \in \text{CONV}(C)$ such that $\text{SUPP}(\boldsymbol{x}, C) \neq \text{SUPP}(\boldsymbol{y}, C)$. Therefore, for any $\boldsymbol{z}$ in the symmetric difference of the sets $\text{SUPP}(\boldsymbol{x}, C)$ and $\text{SUPP}(\boldsymbol{y}, C)$, Eq. (4) will not hold for any finite ϵ_0.

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, $Q_{C_{cp}}$.

(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 12 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 ϵ-DP.

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

$$\begin{bmatrix} -4\mathbf{1}_d^T & 2dI_d \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} \text{ s.t. } \sum_{i=0}^{d} a_i = 1. \tag{5}$$

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

$$a_0 = 1/3 - \frac{\sum_{i=1}^{d} v_i}{6d}, \quad a_i = \frac{v_i}{2d} + \frac{2a_0}{d}\ \forall i \geq 1. \tag{6}$$

Proposition 12. *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 $\epsilon > \log 7$.*

(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 (see Definition 3) formed using the Sylvester construction.

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,

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

that has closed form solution for each a_i as:

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

Proposition 13. *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 $\epsilon > \log(1 + \sqrt{2})$.*

Finally, we remark that even though C_{cp} in the cross-polytope scheme (in Section 5.3.1) has more than $d+1$ points, it still gives us ϵ-DP for any $\epsilon > O(\log d)$. Note that differential privacy for such large parameters is also of interest to the community [3].

Proposition 14. *Let $\tilde{C}_{cp}$ be the set of point in the cross-polytope point set scaled by a factor of 2. Let $\tilde{C}_{cp} = \{\pm 2\sqrt{d} e_i | i \in [d]\}$, then $Q_{\tilde{C}_{cp}}$ is ϵ-DP for any $\epsilon > \log d$.*

We now show a Randomized Response (RR) scheme that can be used on top of any of our quantization schemes to achieve privacy. 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.

6.1 Randomized Response

We present a Randomized Response (RR) mechanism, introduced by [36], that can be used over the output of Q_C to make it ϵ-DP (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 15. *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 $\text{RR}_{\ p}$. Moreover, the scheme is ϵ-differentially private.*

6.2 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 mechanism ([15]) 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 16. *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.*

7 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. While the explicit efficient schemes seems to have a $\log d$-factor communication overhead, we believe it will be hard but interesting to get rid of the factor with deterministic construction.

Information theoretically, we are asking the question of computing the variance of an unbiased estimator of points in the unit sphere, in terms of its unconditional entropy. We have established the exact trade-off between variance and entropy for almost surely bounded estimators. We, in this paper, have tried to minimize the communication: but we believe our techniques will be applicable to variance reduction techniques as well - at the expense of $\Omega(d)$ communication.

Acknowledgements This research is supported in part by NSF awards CCF 1642658, CCF 1934846, and CCF 1909046.

References

[1] 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.

[2] 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.

[3] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1120–1129, 2019.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8, 2017.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] Stelios Georgiou, Christos Koukouvinos, and Jennifer Seberry. Hadamard matrices, orthogonal designs and construction algorithms. In *Designs 2002*, pages 133–205. Springer, 2003.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[24] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

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

[26] Prathamesh Mayekar and Himanshu Tyagi. Ratq: A universal fixed-length quantizer for stochastic optimization. *arXiv preprint arXiv:1908.08200*, 2019.

[27] 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.

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

[29] 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.

[30] 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.

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

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

[33] 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.

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

[35] 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.

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

[37] 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.

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[39] 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.