The Appendix is organized as follows:

1. Section A provides utility results useful in subsequent proofs.
2. Section B provides the proof of the lower bound described in Section 3.1
3. Section C proves helper lemmas for the results in Section 4.
4. Section D proves Theorems 1, 2 and 3.
5. Section E proves the boosting result (Lemma 3) and end to end analysis of Algorithm 1 followed by the boosting algorithm 2.
6. Section F provides details of our experimental setup
7. Section G provides more related work.

A Utlity Results

Lemma A.1. *For any $\mathbf{y} \in \mathbb{R}^d$, let $\mathsf{Q}(\mathbf{y}, \mathcal{Q})$ denote the quantization of $\mathbf{y}$ using Eq (5), where each coordinate, y_i, is quantized independently as $\mathsf{Q}(y_i, \mathcal{Q})$. Then, for any $x \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^d$,*

1. $\mathbb{E}\left[\mathsf{Q}(x, \mathcal{Q})|x\right] = x$.
2. $|\mathsf{Q}(x, \mathcal{Q}) - x| \leqslant u - l$.
3. $\mathsf{Var}\left[\mathsf{Q}(x, \mathcal{Q})|x\right] \leqslant \frac{(u-l)^2}{4}$.
4. $\mathsf{Cov}\left[\mathsf{Q}(\mathbf{v}, \mathcal{Q})|\mathbf{v}\right]$ *is a diagonal matrix.*

Proof. Throughout the proof, we condition on the fixed x and treat all randomness as coming from the independent choices made by the quantizer.

(i) *Unbiasedness.* We have

$$\mathbb{E}[\mathsf{Q}(x, \delta) \mid x] \;=\; p_i(x)u + (1 - p_i(x))\ell \;=\; x.$$

(ii) *Boundedness.* By definition, after rounding, we always round any $x \in [u, l]$ to either u or l. Therefore, $|\mathsf{Q}(x, \mathcal{Q}) - x| \leqslant u - l$.

(iii) *Variance bound.* Using the variance of a Bernoulli random variable, we have,

$$\mathrm{Var}[\mathsf{Q}(x, \mathcal{Q}) \mid x] \;=\; p_i(x)(1 - p_i(x))\,(u - l)^2 \;\leqslant\; \tfrac{(u-l)^2}{4},$$

since $t(1 - t) \leqslant \frac{1}{4}$ for all $t \in [0, 1]$.

(iv) *Covariance bound.* Since the elements of $\mathbf{v}$ are quantized independently, this follows.

□

Lemma A.2 (Choice of learning rate)**.** *Let $\eta := \frac{\alpha \log(n)}{b(\lambda_1 - \lambda_2)}$. Then, under Assumption 1, for $\theta \in (0, 1)$, η satisfies*

$$b(\eta^2\mathcal{M}^2 + \kappa^2) \leqslant \frac{0.008}{\log(d/\theta)}, \text{ and } \eta \in (0, 1)$$

for $\alpha > 1$, $b \geqslant 250\alpha^2 \log^2(n) \log\left(\frac{d}{\theta}\right) / \left(\lambda_1 - \lambda_2\right)^2$, and $\kappa^2 b \leqslant 0.004 / \log\left(\frac{d}{\theta}\right)$.

Proof. For Lemma A.8, we require,

$$4b(\eta^2\mathcal{M}^2 + \kappa^2)\left(1 + 2\log\left(d\right)\right) \leqslant 1 \tag{A.13}$$

For Theorem A.4, we require,

$$4e^2 b(\eta^2\mathcal{M}^2 + \kappa^2) \log\left(\frac{d}{\theta}\right) \leqslant \frac{1}{4} \tag{A.14}$$

where $\theta \in (0,1)$ represents the failure probability. It is not hard to see that (A.14) implies (A.13). Therefore it suffices to ensure

$$b(\eta^2\mathcal{M}^2+\kappa^2)\log\left(\frac{d}{\theta}\right) \leqslant 0.008$$

Setting each term smaller than 0.004, it suffices to have

$$b \geqslant \frac{250\alpha^2\log^2(n)\log\left(\frac{d}{\theta}\right)}{(\lambda_1-\lambda_2)^2}, \quad \kappa^2 b \leqslant \frac{0.004}{\log\left(\frac{d}{\theta}\right)}$$

which completes the proof for the first condition.

The second condition on η follows by setting $\eta \leqslant 1$ and solving for b. This yields

$$b \geqslant \max\left\{250\alpha^2\log^2(n)\log\left(\frac{d}{\theta}\right)/(\lambda_1-\lambda_2)^2, \alpha\log(n)/(\lambda_1-\lambda_2)\right\}$$

Since $\alpha > 1$, the first term is larger than the second one, which completes the proof. □

Lemma A.3. *Let* $\mathbf{w}$ *and* $\boldsymbol{\xi}$ *be vectors in* $\mathbb{R}^d$ *such that* $\|\mathbf{w}\| = 1$ *and* $\mathbf{w}+\boldsymbol{\xi} \neq 0$. *Then,*

$$\sin^2(\mathbf{w}, \mathbf{w}+\boldsymbol{\xi}) \leqslant \left(\frac{\|\boldsymbol{\xi}\|}{\|\mathbf{w}+\boldsymbol{\xi}\|}\right)^2.$$

Proof.

$$\sin^2(\mathbf{w}, \mathbf{w}+\boldsymbol{\xi}) = 1 - \left(\frac{\mathbf{w}^\top(\mathbf{w}+\boldsymbol{\xi})}{\|\mathbf{w}+\boldsymbol{\xi}\|}\right)^2 = \frac{(\mathbf{w}+\boldsymbol{\xi})^\top(\mathbf{w}+\boldsymbol{\xi}) - (1+\mathbf{w}^\top\boldsymbol{\xi})^2}{\|\mathbf{w}+\boldsymbol{\xi}\|^2}$$
$$= \frac{\boldsymbol{\xi}^\top\boldsymbol{\xi} - (\mathbf{w}^\top\boldsymbol{\xi})^2}{\|\mathbf{w}+\boldsymbol{\xi}\|^2} \leqslant \left(\frac{\|\boldsymbol{\xi}\|}{\|\mathbf{w}+\boldsymbol{\xi}\|}\right)^2.$$

□

Lemma A.4. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be unit vectors in* $\mathbb{R}^d$. *Then,*

$$\frac{1}{2}\min(\|\mathbf{x}-\mathbf{y}\|^2, \|\mathbf{x}+\mathbf{y}\|^2) \leqslant \sin^2(\mathbf{x},\mathbf{y}) \leqslant \min(\|\mathbf{x}-\mathbf{y}\|^2, \|\mathbf{x}+\mathbf{y}\|^2).$$

Proof. We express $\sin^2(\mathbf{x},\mathbf{y})$ in terms of $\|\mathbf{x}-\mathbf{y}\|$ and $\|\mathbf{x}+\mathbf{y}\|$. Since $\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x}^\top\mathbf{y} = 2 - 2\cos(\mathbf{x},\mathbf{y})$ and $\|\mathbf{x}+\mathbf{y}\|^2 = 2 + 2\cos(\mathbf{x},\mathbf{y})$,

$$\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{y}\|^2 = 4 \text{ and } \sin^2(\mathbf{x},\mathbf{y}) = 1-\cos^2(\mathbf{x},\mathbf{y}) = \frac{1}{4}\|\mathbf{x}-\mathbf{y}\|^2\|\mathbf{x}+\mathbf{y}\|^2.$$

The upper bound on $\sin^2(\mathbf{x},\mathbf{y})$ follows immediately from the above equations. For the lower bound, note that at least one of $\|\mathbf{x}-\mathbf{y}\|^2$ and $\|\mathbf{x}+\mathbf{y}\|^2$ is at least 2. If $\|\mathbf{x}+\mathbf{y}\|^2 \geqslant 2$, then $\sin^2(\mathbf{x},\mathbf{y}) \geqslant \|\mathbf{x}-\mathbf{y}\|^2/2$. Otherwise, $\sin^2(\mathbf{x},\mathbf{y}) \geqslant \|\mathbf{x}+\mathbf{y}\|^2/2$. □

Lemma A.5. *Let* $\mathbf{x}, \mathbf{y}$, *and* $\mathbf{z}$ *be non-zero vectors in* $\mathbb{R}^d$. *Then,*

$$\sin^2(\mathbf{x},\mathbf{z}) \leqslant 2\sin^2(\mathbf{x},\mathbf{y}) + 2\sin^2(\mathbf{y},\mathbf{z}).$$

B Lower Bounds

Proof of Lemma 1

Proof. Let $\mathbf{v}_1 \in \mathbb{R}^d$ be the unit vector with $\mathbf{v}_1(i) = \delta/3$ for $i \in [d-1]$ and $\mathbf{v}_1(d) = \sqrt{1-\frac{(d-1)\delta^2}{9}}$.

Consider any a vector $\mathbf{w} \in \mathcal{V}_L$, and let $\tilde{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$. Since $\mathbf{w} \in \mathcal{V}_L$, $\mathbf{w}(i) = 0$ or $|\mathbf{w}(i)| \geqslant \delta/2$. In particular, $|\mathbf{v}_1(i) - \mathbf{w}(i)| \geqslant \delta/6$ and $|\mathbf{v}_1(i) + \mathbf{w}(i)| \geqslant \delta/6$ for all $i \in [d-1]$. It follows that

$$\|\mathbf{v}_1 - \mathbf{w}\|^2 \geqslant \sum_{i=1}^{d-1}(\mathbf{v}_1(i) - \mathbf{w}(i))^2 \geqslant (d-1)\left(\frac{\delta}{6}\right)^2 = \frac{\delta^2(d-1)}{36}$$

and $\|\mathbf{v}_1 + \mathbf{w}\|^2 \geqslant \frac{\delta^2(d-1)}{36}$ similarly. The Lemma follows from A.4.

□

Proof of Lemma 2

Proof. It suffices to construct two unit vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ such that $\inf_{\mathbf{w}\in\mathcal{V}_{NL}} \sin^2(\mathbf{w},\mathbf{v}_1) = \Omega(\zeta^2)$ and $\inf_{\mathbf{w}\in\mathcal{V}_{NL}} \sin^2(\mathbf{w},\mathbf{v}_2) = \Omega(\delta_0^2 d)$.

Let $\mathbf{v}_1$ be the vector in $\mathbb{R}^d$ with coordinates

$$\mathbf{v}_1(1) = \frac{1}{\sqrt{1+(1+\zeta/2)^2}},\ \ \mathbf{v}_1(2) = \frac{1+\zeta/2}{\sqrt{1+(1+\zeta/2)^2}},\ \ \mathbf{v}_1(i) = 0\ \forall\ i \geqslant 3.$$

For the sake of contradiction, suppose there exists $\mathbf{w}_1 \in \mathcal{V}_{NL}$ such that $\sin^2(\mathbf{w}_1,\mathbf{v}_1) \leqslant \zeta^2/100$.

Let $\tilde{\mathbf{w}}_1 := \mathbf{w}_1/\|\mathbf{w}_1\|$. By Lemma A.4,

$$\min(\|\mathbf{v}_1 - \tilde{\mathbf{w}}_1\|_2^2, \|\mathbf{v}_1 + \tilde{\mathbf{w}}_1\|_2^2) \leqslant 2\sin^2(\mathbf{v}_1,\mathbf{w}_1) \leqslant \frac{\zeta^2}{50}.$$

Flipping the sign of $\mathbf{w}_1$ if necessary, we may assume $\|\mathbf{v}_1 - \tilde{\mathbf{w}}_1\|_2^2 \leqslant \zeta^2/50$. So,

$$|\mathbf{v}_1(i) - \tilde{\mathbf{w}}_1(i)| \leqslant \zeta/7\ \forall\ i \in [d]. \tag{A.15}$$

The bound $\zeta \leqslant 0.1$ ensures $\mathbf{v}_1(1) \geqslant 20/29$ and $\mathbf{v}_1(2) - \mathbf{v}_1(1) \geqslant \zeta/3$, which also implies $\tilde{\mathbf{w}}_1(2) - \tilde{\mathbf{w}}_1(1) \geqslant \zeta/21 > 0$. It follows that

$$\begin{aligned}
\frac{\mathbf{w}_1(2)+\delta_0/\zeta}{\mathbf{w}_1(1)+\delta_0/\zeta} &= \frac{\tilde{\mathbf{w}}_1(2)+\delta_0/\zeta\cdot 1/\|\mathbf{w}_1\|}{\tilde{\mathbf{w}}_1(1)+\delta_0/\zeta\cdot 1/\|\mathbf{w}_1\|} \leqslant \frac{\mathbf{v}_1(2)+\zeta/7+\delta_0/2\zeta}{\mathbf{v}_1(1)-\zeta/7+\delta_0/2\zeta} \\
&= 1+\frac{\zeta}{2}+\frac{\delta_0/2\zeta+\zeta/7-(1+\zeta/2)(\delta_0/2\zeta-\zeta/7)}{\mathbf{v}_1(1)+\delta_0/2\zeta-\zeta/7} \\
&= 1+\frac{\zeta}{2}+\frac{2\zeta/7+\zeta^2/14-\delta_0/4}{\mathbf{v}_1(1)-\zeta/7+\delta_0/2\zeta} \\
&\leqslant 1+\frac{\zeta}{2}+\frac{2\zeta/7}{2/3} < 1+\zeta,
\end{aligned}$$

and

$$\begin{aligned}
\frac{\mathbf{w}_1(2)+\delta_0/\zeta}{\mathbf{w}_1(1)+\delta_0/\zeta} &= \frac{\tilde{\mathbf{w}}_1(2)+\delta_0/\zeta\cdot 1/\|\mathbf{w}_1\|}{\tilde{\mathbf{w}}_1(1)+\delta_0/\zeta\cdot 1/\|\mathbf{w}_1\|} \geqslant \frac{\mathbf{v}_1(2)-\zeta/7+2\delta_0/\zeta}{\mathbf{v}_1(1)+\zeta/7+2\delta_0/\zeta} \\
&= 1+\frac{\zeta}{2}+\frac{2\delta_0/\zeta-\zeta/7-(1+\zeta/2)(2\delta_0/\zeta+\zeta/7)}{\mathbf{v}_1(1)+\zeta/7+2\delta_0/\zeta} \\
&= 1+\frac{\zeta}{2}-\frac{2\zeta/7+\zeta^2/14+\delta_0}{\mathbf{v}_1(1)+\zeta/7+2\delta_0/\zeta} \\
&> 1+\frac{\zeta}{2}-\frac{\zeta\mathbf{v}_1(1)/2+\zeta^2/14+\delta_0}{\mathbf{v}_1(1)+\zeta/7+2\delta_0/\zeta} = 1.
\end{aligned}$$

Under the logarithmic quantization scheme, it can be inductively shown that $q_k+\delta_0/\zeta = (\delta_0/\zeta)\cdot(1+\zeta)^k$ for all non-negative integers k such that $q_k \in \mathcal{Q}_{NL}$. In particular, $\frac{\mathbf{w}_1(2)+\delta_0/\zeta}{\mathbf{w}_1(1)+\delta_0/\zeta}$ is a non-negative integer power of $1+\zeta$, contradicting

$$1 < \frac{\mathbf{w}_1(2)+\delta_0/\zeta}{\mathbf{w}_1(1)+\delta_0/\zeta} < 1+\zeta.$$

Therefore, $\inf_{\mathbf{w}_1\in\mathcal{V}_{NL}} \sin^2(\mathbf{w}_1,\mathbf{v}_1) \geqslant \zeta^2/100$.

For the other bound, let $\mathbf{v}_2(i) = \delta_0/3$ for $i \in d-1$ and $\mathbf{v}_2(d) = \sqrt{1-(d-1)\delta_0^2/9}$. Any $\mathbf{w}_2 \in \mathcal{V}_{NL}$ satisfies $\mathbf{w}_2(i) = 0$ or $|\mathbf{w}_2(i)| \geqslant \delta_0$ for all $i \in [d]$. Since $\|\mathbf{w}_2\| \in [1/2, 2]$, the normalized vector $\tilde{\mathbf{w}}_2 = \mathbf{w}_2/\|\mathbf{w}_2\|$ satisfies $|\tilde{\mathbf{w}}_2(i)| = 0$ or $|\tilde{\mathbf{w}}_2(i)| \geqslant \delta_0/2$ for all $i \in [d]$.

In particular $|\mathbf{v}_2(i) - \tilde{\mathbf{w}}_2(i)| \geqslant \delta_0/6$ and $|\mathbf{v}_2(i) + \tilde{\mathbf{w}}_2(i)| \geqslant \delta_0/6$ for all $i \in [d]$. By Lemma A.4,

$$\sin^2(\mathbf{w}_2,\mathbf{v}_2) \geqslant \frac{1}{2}\min\left(\|\mathbf{w}_2-\mathbf{v}_2\|^2, \|\mathbf{w}_2+\mathbf{v}_2\|^2\right) \geqslant \frac{\delta_0^2(d-1)}{72}.$$

□

C Proof of Results in Section 4

For ease of exposition, all results in this section are stated with a generic number of datapoints n. We apply these results with different choices of n (e.g. number of batches b) for proving the main theorems (Theorem 1, 2, 3). Consider Oja's Algorithm applied to the matrices $\mathbf{A}_i \in \mathbb{R}_{d\times d}$, such that $\mathbf{A}_i = \eta\mathbf{D}_i + \mathbf{\Xi}_i$ where $\mathbf{D}_i$ are independent with $\mathbb{E}[\mathbf{D}_i] = \mathbf{\Sigma}$. Let $\mathcal{S}_i$ be the set of all random vectors $\boldsymbol{\xi}$ resulting from the quantizations in the first i iterations of the algorithm, and let $\mathcal{F}_{i-}$ denote the σ-field generated by $\mathbf{D}_1, \ldots, \mathbf{D}_i$ and $\mathcal{S}_{i-1}$, and denote $\mathbb{E}_i[.] := \mathbb{E}[.|\mathcal{F}_{i-}]$. We assume the noise term $\mathbf{\Xi}_i$ is conditionally unbiased, i.e., $\mathbb{E}_i[\mathbf{\Xi}_i] = \mathbf{0}_{d\times d}$.

$$\mathcal{F}_{i-} := \sigma\left(\{\mathbf{D}_1, \ldots \mathbf{D}_i, \mathcal{S}_{i-1}\}\right), \qquad \mathcal{F}_i := \sigma\left(\{\mathbf{D}_1, \ldots \mathbf{D}_i, \mathcal{S}_i\}\right).$$

Recall the update rule

$$\mathbf{u}_i = (\mathbf{I} + \mathbf{A}_i)\mathbf{w}_{i-1}; \qquad \mathbf{w}_i = \frac{\prod_{t=i}^{1}(\mathbf{I} + \mathbf{A}_t)\mathbf{u}_0}{\|\prod_{t=i}^{1}(\mathbf{I} + \mathbf{A}_t)\mathbf{u}_0\|}. \tag{A.16}$$

We bound the numerator and denominator in (A.16) separately.

For the numerator, we will show that $\|\prod_{t=n}^{1}(\mathbf{I} + \mathbf{A}_t) - (\mathbf{I} + \eta\mathbf{\Sigma})^n\|$ is small. Let $\mathbf{Y}_i = \mathbf{I} + \mathbf{A}_i$ for $i \in [n]$, and let $\{\mathbf{Z}_i\}_{0\leqslant i\leqslant n}$ be defined as

$$\mathbf{Z}_i := \mathbf{Y}_i\mathbf{Z}_{i-1}, \qquad \mathbf{Z}_0 := \mathbf{I}. \tag{A.17}$$

Note that $\mathbf{Z}_{i-1}$ is measurable w.r.t $\mathcal{F}_{i-}$.

We are now ready to state our first result. Note that

$$\mathbf{Z}_n = \prod_{i=n}^{1}(\mathbf{I} + \mathbf{A}_i).$$

where $\mathbf{A}_i = \eta\mathbf{D}_i + \mathbf{\Xi}_i$ and $\mathbf{D}_i$ are independent $d \times d$ random matrices with mean $\mathbf{\Sigma}$.

C.1 Proof of Proposition 1

Proposition A.1. *[Proposition 1 in main paper] Under Assumption 1, for $\eta \in (0, 1)$, we have:*

$$\|\mathbf{Z}_n\|_{p,q}^2 \leqslant (\alpha + C_p\gamma)^n\|\mathbf{Z}_0\|_{p,q}^2, \text{ and}$$
$$\|\mathbf{Z}_n - (\mathbf{I} + \eta\mathbf{\Sigma})^n\|_{p,q}^2 \leqslant \alpha^n(\exp(C_pn\gamma) - 1)\|\mathbf{Z}_0\|_{p,q}^2$$

for $\mathbf{Z}_0 = \mathbf{I}$, $\alpha := (1 + \eta\lambda_1)^2$, $\gamma := 2(\eta^2\mathcal{M}^2 + \kappa^2)$ and $C_p := p - 1$.

Proof. Note that

$$\mathbb{E}[\mathbf{Y}_i|\mathcal{F}_{i-1}] = \mathbf{I} + \eta\mathbf{\Sigma} + \mathbb{E}[\mathbb{E}_{i-}[\mathbf{\Xi}_i]|\mathcal{F}_{i-1}] = \mathbf{I} + \eta\mathbf{\Sigma}$$

Note that $m_i = 1 + \eta\lambda_1$ and

$$\|\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i|\mathcal{F}_{i-1}]\| = \|\eta(\mathbf{D}_i - \mathbf{\Sigma}) + \mathbf{\Xi}_i\| \leqslant \eta\mathcal{M} + \kappa$$

The last line uses Eq 9. Thus $\sigma_i = \frac{\eta\mathcal{M}+\kappa}{1+\eta\lambda_1}$. Note that $\nu \leqslant 2(\eta^2\mathcal{M}^2 + \kappa^2)$. The same argument as in Theorem 7.4 in [HNWTW20] gives the bound. □

Lemma A.6. *Under Assumption 1, and η set according to Lemma A.2 with $b = n$,*

$$\mathbb{P}\left(\|\mathbf{Z}_n - (\mathbf{I} + \eta\mathbf{\Sigma})^n\| \geqslant t\,(1 + \eta\lambda_1)^n\right) \leqslant \max(d, e)\exp\left(-\frac{t^2}{2e^2n\gamma}\right) \quad \forall\ t \leqslant e.$$

where $\gamma := 2(\eta^2\mathcal{M}^2 + \kappa^2)$ and $e = 2.718\ldots$ is the Napier's constant.

Proof. By Proposition A.1,

$$\mathbb{P}\left(\|\mathbf{Z}_n - (\mathbf{I} + \eta\mathbf{\Sigma})^n\| \geqslant t\,(1 + \eta\lambda_1)^n\right) \leqslant \frac{\mathbb{E}\left[\|\mathbf{Z}_n - (\mathbf{I} + \eta\mathbf{\Sigma})^n\|^p\right]}{t^p\,(1 + \eta\lambda_1)^p} \leqslant \frac{\|\mathbf{Z}_n - (\mathbf{I} + \eta\mathbf{\Sigma})^n\|_{p,p}^p}{t^p\,(1 + \eta\lambda_1)^p}$$
$$\leqslant \frac{\alpha^{\frac{p}{2}}(\exp(C_pn\gamma) - 1)^{p/2}d}{t^p\,(1 + \eta\lambda_1)^p} \leqslant d\left(t^{-2}(\exp(C_pn\gamma) - 1)\right)^{p/2}.$$

If $\frac{t^2}{e^2 n\gamma} < 2$, then $e \cdot \exp\left(-\frac{t^2}{2e^2 n\gamma}\right) \geqslant 1$. Otherwise, let $p := \frac{t^2}{e^2 n\gamma} \geqslant 2$. Since $t \leqslant e$, $C_p n\gamma \leqslant pn\gamma \leqslant \frac{t^2}{e^2} \leqslant 1$. Therefore, $\exp(C_p n\gamma) - 1 \leqslant eC_p n\gamma \leqslant \frac{t^2}{e}$. Therefore,

$$\mathbb{P}\left(\|\mathbf{Z}_n - (\mathbf{I} + \eta\boldsymbol{\Sigma})^n\| \geqslant t\,(1+\eta\lambda_1)^n\right) \leqslant d.\left(t^{-2}.(\exp(C_p n\gamma) - 1)\right)^{p/2} \leqslant d\exp\left(-\frac{t^2}{2e^2 n\gamma}\right).$$

□

Lemma A.7. *Under Assumption 1 and η set according to Lemma A.2 with $b = n$,*

$$\mathbb{E}\left[\|\mathbf{Z}_n\|^2\right] \leqslant \exp\left(2\sqrt{2n\gamma \max\left\{2n\gamma, \log(d)\right\}}\right)(1+\eta\lambda_1)^{2n},$$

where $\gamma = 2(\eta^2\mathcal{M}^2 + \kappa^2)$. Moreover, if $2n\gamma\,(1 + 2\log(d)) \leqslant 1$, then

$$\mathbb{E}\left[\|\mathbf{Z}_n - \mathbb{E}[\mathbf{Z}_n]\|^2\right] \leqslant 2e^2 n\gamma\,(1+2\log(d))\,(1+\eta\lambda_1)^{2n}.$$

Proof. Using Proposition A.1 $\alpha := (1+\eta\lambda_1)^2$, and $\gamma := 2(\eta^2\mathcal{M}^2 + \kappa^2)$,

$$\mathbb{E}[\|\mathbf{Z}_n\|^2] \leqslant \|\mathbf{Z}_n\|_{p,2}^2 \leqslant (\alpha + C_p\gamma)^n \|\mathbf{Z}_0\|_{p,2}^2 \leqslant (1+\eta\lambda_1)^{2n} \exp(C_p n\gamma)\|\mathbf{Z}_0\|_{p,2}^2.$$

Set $p := \max\left(2, \sqrt{\frac{2\log d}{n\gamma}}\right)$. Then, $\|\mathbf{Z}_0\|_{p,2} = d^{\frac{1}{p}} \leqslant \exp\left(\frac{pn\gamma}{2}\right)$ and

$$\mathbb{E}[\|\mathbf{Z}_n\|^2] \leqslant (1+\eta\lambda_1)^{2n}\exp(2pn\gamma) = \exp\left(2\sqrt{2n\gamma\max\left\{2n\gamma, \log(d)\right\}}\right)(1+\eta\lambda_1)^{2n}.$$

For the second result, set $p := 2\,(1+\log(d))$. Then, $C_p n\gamma \leqslant 1$ and $\|\mathbf{Z}_0\|_p \leqslant \sqrt{e}$. By Proposition A.1,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Z}_n - \mathbb{E}[\mathbf{Z}_n]\|^2\right] \leqslant \|\mathbf{Z}_n - \mathbb{E}[\mathbf{Z}_n]\|_{p,2}^2 &\leqslant (\exp(C_p n\gamma) - 1)(1+\eta\lambda_1)^n \|\mathbf{Z}_0\|_p^2 \\ &\leqslant e^2 C_p n\gamma\,(1+\eta\lambda_1)^n \\ &\leqslant 2e^2 n\gamma\,(1+\log(d))\,(1+\eta\lambda_1)^n.\end{aligned}$$

□

C.2 Proof of Lemma 4

Lemma A.8 (Lemma 4 in main paper)**.** *Let Assumption 1 hold and η be set according to Lemma A.2 with $b = n$. Define $\gamma := 2(\eta^2\mathcal{M}^2 + \kappa^2)$. If $2n\gamma\,(1 + 2\log(d)) \leqslant 1$, then*

$$\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top \mathbf{Z}_n \mathbf{Z}_n^\top \mathbf{V}_\perp\right)\right] \leqslant \exp\left(2\eta n\lambda_1 + \eta^2 n\left(\mathcal{V}_0 + \lambda_1^2\right)\right)\left[\frac{d}{\exp\left(2\eta n\left(\lambda_1 - \lambda_2\right)\right)} + \frac{5\left(\eta^2\mathcal{V}_0 + \kappa_1\right)}{\eta\left(\lambda_1 - \lambda_2\right)}\right].$$

Proof. Let $\beta_i := \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top \mathbf{Z}_i \mathbf{Z}_i^\top \mathbf{V}_\perp\right)\right]$ for all $0 \leqslant i \leqslant n$. Then, for any $i \in [n]$,

$$\begin{aligned}\beta_i &= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top (\mathbf{I} + \mathbf{A}_i)\,\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\,(\mathbf{I} + \mathbf{A}_i^\top)\,\mathbf{V}_\perp\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top (\mathbf{I} + \mathbf{A}_i)\,\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\,(\mathbf{I} + \mathbf{A}_i^\top)\,\mathbf{V}_\perp\right)|\mathcal{F}_{i-}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top (\mathbf{I} + \eta\mathbf{Y}_i)\,\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\,(\mathbf{I} + \eta\mathbf{Y}_i)\,\mathbf{V}_\perp\right)|\mathcal{F}_{i-}\right]\right] \\ &\quad + \mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top \boldsymbol{\Xi}_i \mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top \boldsymbol{\Xi}_i^\top \mathbf{V}_\perp\right)|\mathcal{F}_{i-}\right]\right].\end{aligned}$$

The last line used $\mathbb{E}[\boldsymbol{\Xi}_i|\mathcal{F}_{i-}] = \mathbf{0}$ and that $\mathbf{Z}_{i-1}$ is measurable with respect to $\mathcal{F}_{i-}$. In other words,

$$\beta_i = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top (\mathbf{I} + \eta\mathbf{Y}_i)\,\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\,(\mathbf{I} + \eta\mathbf{Y}_i)\,\mathbf{V}_\perp\right)\right] + \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{Z}_{i-1}^\top \mathbb{E}\left[\boldsymbol{\Xi}_i^\top \mathbf{V}_\perp \mathbf{V}_\perp{}^\top \boldsymbol{\Xi}_i|\mathcal{F}_{i-}\right]\mathbf{Z}_{i-1}\right)\right].$$

For the first term, following the analysis of Lemma 10 from [JJK+16],

$$\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top\left(\mathbf{I}+\eta\mathbf{Y}_i\right)\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\left(\mathbf{I}+\eta\mathbf{Y}_i\right)\mathbf{V}_\perp\right)\right] \leqslant \left(1+2\eta\lambda_2+\eta^2\left(\mathcal{V}_0+\lambda_1^2\right)\right)\beta_{i-1}+\eta^2\mathcal{V}_0\left\|\mathbb{E}\left[\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\right]\right\|_2$$
$$\leqslant \left(1+2\eta\lambda_2+\eta^2\left(\mathcal{V}_0+\lambda_1^2\right)\right)\beta_{i-1}+\eta^2\mathcal{V}_0\mathbb{E}\left[\left\|\mathbf{Z}_{i-1}\right\|_2^2\right]. \tag{A.18}$$

The second term can be bounded as follows:

$$\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{Z}_{i-1}^\top\mathbb{E}\left[\mathbf{\Xi}_i^\top\mathbf{V}_\perp\mathbf{V}_\perp{}^\top\mathbf{\Xi}_i|\mathcal{F}_{i-}\right]\mathbf{Z}_{i-1}\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{\Xi}_i^\top\mathbf{V}_\perp\mathbf{V}_\perp{}^\top\mathbf{\Xi}_i|\mathcal{F}_{i-}\right]\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\right)\right]$$
$$\leqslant \mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{\Xi}_i^\top\mathbf{V}_\perp\mathbf{V}_\perp{}^\top\mathbf{\Xi}_i\right)|\mathcal{F}_{i-}\right]\left\|\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\right\|_2\right]$$
$$\leqslant \kappa_1\mathbb{E}\left[\left\|\mathbf{Z}_{i-1}\mathbf{Z}_{i-1}^\top\right\|_2\right]. \tag{A.19}$$

Combining (A.18) and (A.19), we obtain the recurrence

$$\beta_i \leqslant \left(1+2\eta\lambda_2+\eta^2\left(\mathcal{V}_0+\lambda_1^2\right)\right)\beta_{i-1}+\left(\eta^2\mathcal{V}_0+\kappa_1\right)\mathbb{E}[\|\mathbf{Z}_{i-1}\|_2^2].$$

By Lemma A.7, we have for $\gamma := 2(\eta^2\mathcal{M}^2+\kappa^2)$,

$$\beta_i \leqslant \left(1+2\eta\lambda_2+\eta^2\left(\mathcal{V}_0+\lambda_1^2\right)\right)\beta_{i-1}+\left(\eta^2\mathcal{V}_0+\kappa_1\right)\exp\left(2\sqrt{2n\gamma\log d}\right)\left(1+\eta\lambda_1\right)^{2(i-1)}$$
$$\leqslant \exp\left(2\eta\lambda_2+\eta^2\left(\mathcal{V}_0+\lambda_1^2\right)\right)\beta_{i-1}+s\exp\left(2\eta\lambda_1\left(i-1\right)+\eta^2(\mathcal{V}_0+\lambda_1^2)\left(i-1\right)\right),$$

where $s = \left(\eta^2\mathcal{V}_0+\kappa_1\right)\exp\left(2\sqrt{2n\gamma\log d}\right)$. Unrolling the recursion,

$$\beta_n \leqslant \exp\left(2\eta n\lambda_1+\eta^2 n\left(\mathcal{V}_0+\lambda_1^2\right)\right)\left[\exp\left(-2\eta n\left(\lambda_1-\lambda_2\right)\right)\beta_0+s.\sum_{t=0}^{n-1}\left(\frac{\exp\left(2\eta\lambda_2+\eta^2\left(\mathcal{V}_0+\lambda_1^2\right)\right)}{\exp\left(2\eta\lambda_1+\eta^2\left(\mathcal{V}_0+\lambda_1^2\right)\right)}\right)^{2(n-1-t)}\right]$$
$$\leqslant \exp\left(2\eta n\lambda_1+\eta^2 n\left(\mathcal{V}_0+\lambda_1^2\right)\right)\left[\exp\left(-2\eta n\left(\lambda_1-\lambda_2\right)\right)\beta_0+\frac{s}{1-\exp(-2\eta\left(\lambda_1-\lambda_2\right))}\right]$$
$$\leqslant \exp\left(2\eta n\lambda_1+\eta^2 n\left(\mathcal{V}_0+\lambda_1^2\right)\right)\left[\exp\left(-2\eta n\left(\lambda_1-\lambda_2\right)\right)\beta_0+\frac{2.35s}{2\eta\left(\lambda_1-\lambda_2\right)}\right].$$

where the last line follows since $\frac{1}{1-e^{-x}} \leqslant \frac{2.35}{x}$ for $x \leqslant 2$. The proof follows since $\beta_0 \leqslant d$ and $\exp\left(2\sqrt{2n\gamma\log d}\right) \leqslant \exp(\sqrt{2}) < 4.12$. □

D Proofs of Theorems 1, 2, and 3

D.1 Proof of Theorem 1

We are now ready to present the proof of Theorem 1, which follows from the following Theorem A.4 with $\theta = 0.9$ and $n = b$.

Theorem A.4. *Fix $\theta \in (0,1)$. Then, for $\mathbf{w}_b$ being the output of Algorithm 1, under assumption 1, η set as $\frac{\alpha\log n}{b(\lambda_1-\lambda_2)}$, α set as in Lemma A.2, $\kappa_1 \leqslant 1/2$, and*

$$\sqrt{2e^2b\gamma\log\left(d/\theta\right)} \leqslant \frac{1}{2},$$

for $\gamma := 2(\eta^2\mathcal{M}^2+\kappa^2)$. Then, with probability at least $1-3\theta$,

$$\sin^2(\mathbf{w},\mathbf{v}_1) \leqslant \frac{24\log\left(1/\theta\right)}{\theta^3}\left[\frac{d}{\exp\left(2\alpha\log(n)\right)}+\frac{5\left(\eta^2\mathcal{V}_0+\kappa_1\right)}{\eta\left(\lambda_1-\lambda_2\right)}\right]+8\kappa_1.$$

Proof. Note that by Algorithm 1 and the definition of $\mathbf{Z}$ in (A.17),

$$\mathbf{u}_b = \frac{\mathbf{Z}_b\mathbf{u}_0}{\|\mathbf{Z}_b\mathbf{u}_0\|}.$$

Since $\mathbf{v}_1\mathbf{v}_1^\top + \mathbf{V}_\perp \mathbf{V}_\perp{}^\top = \mathbf{I}_d$,

$$\sin^2(\mathbf{u}_b, \mathbf{v}_1) = 1 - \left(\mathbf{u}_b^\top \mathbf{v}_1\right)^2 = \left\| \frac{\mathbf{V}_\perp \mathbf{V}_\perp{}^\top \mathbf{Z}_n \mathbf{u}_0}{\|\mathbf{Z}_b \mathbf{u}_0\|} \right\|^2.$$

By Lemma 6 from [JJK+16], with probability at least $1-\theta$,

$$\sin^2(\mathbf{u}_b, \mathbf{v}_1) \leqslant \frac{2.5\log(1/\theta)}{\theta^2} \frac{\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top \mathbf{Z}_b \mathbf{Z}_b^\top \mathbf{V}_\perp\right)}{\mathbf{v}_1^\top \mathbf{Z}_b \mathbf{Z}_b^\top \mathbf{v}_1}.$$

We apply Lemma A.7 with $q = 2$ and $p = 2(1+\log(d))$. Since $b\gamma(1+2\log(d)) \leqslant 1$,

$$\mathbb{E}\left[\|\mathbf{Z}_b - (\mathbf{I}+\eta\mathbf{\Sigma})^b\|\right] \leqslant \|\mathbf{Z}_b - (\mathbf{I}+\eta\mathbf{\Sigma})^b\|_{p,2} \leqslant \sqrt{e^2 b\gamma(1+2\log(d))}\,(1+\eta\lambda_1)^b, \tag{A.20}$$

For the numerator, we use Lemma A.8 and Markov's inequality to get

$$\mathrm{Tr}\left(\mathbf{V}_\perp{}^\top \mathbf{Z}_b \mathbf{Z}_b^\top \mathbf{V}_\perp\right) \leqslant \frac{1}{\theta}\exp\left(2\eta b\lambda_1 + \eta^2 b\left(\mathcal{V}_0 + \lambda_1^2\right)\right)\left[\frac{d}{\exp\left(2\eta b(\lambda_1-\lambda_2)\right)} + \frac{5\left(\eta^2\mathcal{V}_0 + \kappa_1\right)}{\eta(\lambda_1-\lambda_2)}\right]. \tag{A.21}$$

with probability at least $1-\theta$. The denominator can be bounded as

$$\left\|\mathbf{Z}_b^\top \mathbf{v}_1\right\| \geqslant \left\|(\mathbf{I}+\eta\mathbf{\Sigma})^b \mathbf{v}_1\right\| - \left\|\left(\mathbf{Z}_b - (\mathbf{I}+\eta\mathbf{\Sigma})^b\right)^\top \mathbf{v}_1\right\| \geqslant (1+\eta\lambda_1)^n - \left\|\mathbf{Z}_b - (\mathbf{I}+\eta\mathbf{\Sigma})^b\right\|.$$

Using Lemma A.6, with probability atleast $1-\theta$,

$$\begin{aligned}
\|\mathbf{Z}_b\mathbf{v}_1\| &\geqslant (1+\eta\lambda_1)^b - \sqrt{2e^2 b\gamma\log(d/\theta)}\,(1+\eta\lambda_1)^b \\
&= (1+\eta\lambda_1)^b\left(1 - \sqrt{2e^2 b\gamma\log(d/\theta)}\right) \\
&\geqslant \exp\left(\eta\lambda_1 b - \eta^2\lambda_1^2 b\right)\left(1 - \sqrt{2e^2 b\gamma\log(d/\theta)}\right).
\end{aligned} \tag{A.22}$$

where the last line follows since $(1+x) \geqslant \exp\left(x - x^2\right)$ for all $x \geqslant 0$. From equations (A.21), (A.22), and the assumption $\sqrt{2e^2 b\gamma\log(d/\theta)} \leqslant 1/2$, it follows that with probability $1-3\theta$,

$$\sin^2(\mathbf{u}_b, \mathbf{v}_1) \leqslant \frac{12\log(1/\theta)}{\theta^3}\left[\frac{d}{\exp(2\alpha\log(n))} + \frac{5\left(\eta^2\mathcal{V}_0 + \kappa_1\right)}{\eta(\lambda_1-\lambda_2)}\right]. \tag{A.23}$$

Since $\mathbf{w} \leftarrow \mathsf{Q}(\mathbf{u}_b, \mathcal{Q})$, by Lemma A.9 and using $\|\boldsymbol{\xi}\| \leqslant \kappa_1 \leqslant 0.5$,

$$\sin^2(\mathbf{w}, \mathbf{u}_b) \leqslant \frac{\|\boldsymbol{\xi}\|^2}{\|\mathbf{u}_b + \boldsymbol{\xi}\|^2} \leqslant \frac{\|\boldsymbol{\xi}\|^2}{(\|\mathbf{u}_b\| - \|\boldsymbol{\xi}\|)^2} \leqslant \frac{\kappa_1}{0.5^2} \leqslant 4\kappa_1. \tag{A.24}$$

The result follows by using equations (A.23), (A.24), and Lemma A.5. □

D.2 Proofs of Theorems 2 and 3

Next, we apply Theorem A.4 to analyze the quantized version of Oja's algorithm as described in Algorithm 1. The idea is to show that the error from the rounding operation can be incorporated into the noise in the iterates of Oja's algorithm, which, given an appropriately chosen σ-field, will be mean zero. For this subsection, we will use:

$$\mathbf{D}_i = \sum_{j \in B_i} \mathbf{X}_j \mathbf{X}_j^T/(n/b),$$

where $\mathbf{A}_i = \eta\left(\mathbf{D}_i + \boldsymbol{\xi}_{a,i}\mathbf{u}_{i-1}^T\right) + \boldsymbol{\xi}_{2,i}\mathbf{u}_{i-1}^T + (\mathbf{I}+\eta\mathbf{D}_i)\boldsymbol{\xi}_{1,i}\mathbf{u}_{i-1}^T$.

We first state and prove some intermediate results needed for the proof of Theorem 2.

Theorem A.5. *Let $d, n, b \in \mathbb{N}$, and let $\{\mathbf{X}_i\}_{i\in[n]}$ be a set of n IID vectors in $\mathbb{R}^d$ satisfying assumption 1. Let $\eta := \frac{\alpha \log n}{b(\lambda_1-\lambda_2)}$ be the learning rate set as in Lemma A.2. Suppose the quantization grid $\mathcal{Q} = \mathcal{Q}_L$, and $\sqrt{4e^2 b(4\eta^2 + 9\delta^2 d)\log(d/\theta)} \leqslant \frac{1}{2}$. Then, with probability at least 0.9, the output $\mathbf{w}$ of Algorithm 1 satisfies*

$$\sin^2(\mathbf{w}, \mathbf{v}_1) \leqslant \frac{24\log(1/\theta)}{\theta^3}\left[\frac{d}{n^{2\alpha}} + \frac{5\alpha\mathcal{V}\log n}{n(\lambda_1-\lambda_2)^2} + \frac{30b\delta^2 d}{\alpha \log n}\right] + 48\delta^2 d.$$

Proof. In order to apply Theorem 1, we come up with valid choices of $\mathcal{V}_0$, κ, and κ_1.

Since each $\mathbf{D}_i$ is symmetric and $\{\mathbf{X}_i\}_{i\in[n]}$ are independent,

$$\left\|\mathbb{E}[(\mathbf{D}_i - \mathbf{\Sigma})(\mathbf{D}_i - \mathbf{\Sigma})^T]\right\| = \left\|\frac{1}{n/b}\mathbb{E}[(\mathbf{X}_1\mathbf{X}_1^T - \mathbf{\Sigma})^2]\right\| \leqslant \frac{b\mathcal{V}}{n} =: \mathcal{V}_0. \tag{A.25}$$

Next,

$$\mathbf{\Xi}_i = \eta\boldsymbol{\xi}_{a,i}\mathbf{u}_{i-1}^T + \boldsymbol{\xi}_{2,i}\mathbf{u}_{i-1}^T + (\mathbf{I} + \eta\mathbf{D}_i)\boldsymbol{\xi}_{1,i}\mathbf{u}_{i-1}^T.$$

Also observe that

$$\mathbb{E}[\boldsymbol{\xi}_{1,i}|\mathcal{F}_{i-}] = 0, \qquad \mathbb{E}[\boldsymbol{\xi}_{a,i}|\mathcal{F}_{i-}] = 0, \qquad \mathbb{E}[\boldsymbol{\xi}_{2,i}|\boldsymbol{\xi}_{a,i}, \boldsymbol{\xi}_{1,i}, \mathcal{F}_{i-}] = 0, \tag{A.26}$$

By equation A.26,

$$\mathbb{E}[\mathbf{\Xi}_i^T\mathbf{\Xi}_i|\mathcal{F}_{i-}] = \mathbb{E}[\eta^2\mathbf{u}_{i-1}\boldsymbol{\xi}_{a,i}^T\boldsymbol{\xi}_{a,i}\mathbf{u}_{i-1}^T + \mathbf{u}_{i-1}\boldsymbol{\xi}_{2,i}^T\boldsymbol{\xi}_{2,i}\mathbf{u}_{i-1}^T + \mathbf{u}_{i-1}\boldsymbol{\xi}_{1,i}^T(I+\eta\mathbf{D}_i)(I+\eta\mathbf{D}_i)^T\boldsymbol{\xi}_{2,i}\mathbf{u}_{i-1}^T|\mathcal{F}_{i-}]$$
$$\implies \left\|\mathbb{E}[\mathbf{\Xi}_i^T\mathbf{\Xi}_i|\mathcal{F}_{i-}]\right\|_F \leqslant \eta^2\delta^2 d + \delta^2 d + (1+\eta)^2\delta^2 d \leqslant 6\delta^2 d =: \kappa_1.$$

As for κ, we have

$$\|\mathbf{\Xi}_i\| \leqslant 2(1+\eta)\delta\sqrt{d} \leqslant 3\delta\sqrt{d} =: \kappa$$

We are now ready to obtain the sin-squared error. Note that $\mathcal{M} \leqslant 2$, since $\|\mathbf{X}_i\| \leqslant 1$ almost surely, for all $i \in [n]$. By Theorem A.4, with probability at least $1 - 3\theta$,

$$\sin^2(\mathbf{w}, \mathbf{v}_1) \leqslant \frac{24\log(1/\theta)}{\theta^3}\left[\frac{d}{\exp(2\alpha\log(n))} + \frac{5\left(\eta^2\mathcal{V}_0 + \kappa_1\right)}{\eta(\lambda_1-\lambda_2)}\right] + 8\kappa_1.$$

as long as $\sqrt{2e^2 b\gamma\log(d/\theta)} \leqslant \frac{1}{2}$. Our parameter choices are $\mathcal{V}_0 = \frac{b\mathcal{V}}{n}$, $\kappa = 3\delta\sqrt{d}$, and $\kappa_1 = 6\delta^2 d$.

$$\sin^2(\mathbf{w}, \mathbf{v}_1) \leqslant \frac{24\log(1/\theta)}{\theta^3}\left[\frac{d}{n^{2\alpha}} + \frac{5\alpha\mathcal{V}\log n}{n(\lambda_1-\lambda_2)^2} + \frac{30b\delta^2 d}{\alpha \log n}\right] + 48\delta^2 d.$$

□

Lemma A.9. *Let $\mathbf{u} = Q(\mathbf{w}, \mathcal{Q}_{NL})$, where $\mathbf{u} \in \mathbb{R}^d$ and $\mathcal{Q}_{NL}$ is defined in equation 4. Then,*

$$\|\mathbf{w} - Q(\mathbf{w}, \mathcal{Q}_{NL})\| \leqslant \delta_0\sqrt{d} + \|\mathbf{w}\|\zeta$$

Proof. Let $\boldsymbol{\xi} = \mathsf{Q}(\mathbf{w}, \mathcal{Q}_{NL}) - \mathbf{w}$. Say $\mathbf{w}_i > 0$. Let k be the unique integer such that $\mathbf{w}_i \in [q_k, q_{k+1}]$. Equivalently for negative $\mathbf{w}_i$, say the bin is $[-q_{k+1}, -q_k]$. We have:

$$|\boldsymbol{\xi}_i| \leqslant q_{k+1} - q_k \leqslant \delta_0 + \zeta q_k \leqslant |\mathbf{w}_i|\zeta + \delta_0$$

Thus we have:

$$\|\boldsymbol{\xi}\| \leqslant \delta_0\sqrt{d} + \|\mathbf{w}\|\zeta.$$

□

Theorem A.6. *Fix $\theta \in (0,1)$. Let the initial vector $\mathbf{u}_0 \sim \mathcal{N}(0, \mathbf{I})$. Let the number of batches b and quantization scale δ be such that $\sqrt{4e^2 b(4\eta^2 + 32\delta_0^2 d + 98\zeta^2)\log(d/\theta)} \leqslant 1/2$. Then, under assumption 1 with η set as $\frac{\alpha\log n}{b(\lambda_1-\lambda_2)}$, where α is set as in Lemma A.2, $\delta_0\sqrt{d} \leqslant 0.25$, and $\zeta \leqslant 0.25$, with probability at least $1 - 3\theta$, the output $\mathbf{w}_b$ of Algorithm 1 gives:*

$$\sin^2(\mathbf{w}, \mathbf{v}_1) \leqslant \frac{24\log(1/\theta)}{\theta^3}\left[\frac{d}{n^{2\alpha}} + \frac{5\alpha\mathcal{V}\log n}{n(\lambda_1-\lambda_2)^2} + \frac{5b(4\delta_0\sqrt{d} + 7\zeta)^2}{\alpha\log n}\right] + 8(4\delta_0\sqrt{d} + 7\zeta)^2.$$

Proof. In order to apply Theorem 1 we need to bound $\mathcal{V}$, κ and κ_1. We start with the first. For us, $\mathbf{D}_i$ is defined in Eq 9. Let $\mathcal{R}_i$ denote the random variables in the quantization up to and including the i^{th} update.

Our analysis is analogous to the previous theorem. Note that the $\mathcal{V}_0$ parameter is as in Eq A.25.

Now we will work out κ and κ_1 since those are the only quantities that change for the nonlinear quantization. Recall that we have,

$$\boldsymbol{\Xi}_i = \eta \boldsymbol{\xi}_{a,i} \mathbf{u}_{i-1}^T + \boldsymbol{\xi}_{2,i} \mathbf{u}_{i-1}^T + (\mathbf{I} + \eta \mathbf{D}_i) \boldsymbol{\xi}_{1,i} \mathbf{u}_{i-1}^T.$$

We have,

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{\Xi}_i^T \boldsymbol{\Xi}_i | \mathcal{F}_{i-}] \\
&= \eta^2 \mathbb{E}[\mathbf{u}_{i-1} \boldsymbol{\xi}_{a,i}^T \boldsymbol{\xi}_{a,i} \mathbf{u}_{i-1}^T | \mathcal{F}_{i-}] + \mathbb{E}[\mathbf{u}_{i-1} \boldsymbol{\xi}_{2,i}^T \boldsymbol{\xi}_{2,i} \mathbf{u}_{i-1}^T | \mathcal{F}_{i-}] + \mathbb{E}[\mathbf{u}_{i-1} \boldsymbol{\xi}_{1,i}^T (I + \eta \mathbf{D}_i)(I + \eta \mathbf{D}_i)^T \boldsymbol{\xi}_{2,i} \mathbf{u}_{i-1}^T | \mathcal{F}_{i-}]
\end{aligned}$$

Now we obtain the Frobenius norm of $\boldsymbol{\xi}_{a,i}$, $\boldsymbol{\xi}_1$, and $\boldsymbol{\xi}_2$ under the nonlinear quantization. We start with the norm of $\mathbf{w}_i$, a quantized version of a unit vector $\mathbf{u}_{i-1}$.

By Lemma A.9, $\|\mathbf{w}_i\| \leqslant 1 + \delta_0 \sqrt{d} + \zeta$. Let $\mathbf{s}_j = \mathbf{X}_j (\mathbf{X}_j^T \mathbf{w}_i)$. Then,

$$\|\mathbf{s}_j\| \leqslant \|\mathbf{w}_i\| \leqslant 1 + \delta_0 \sqrt{d} + \zeta.$$

Another application of Lemma A.9 gives:

$$\|\boldsymbol{\xi}_{a,j,i}\| = \|\mathsf{Q}(\mathbf{s}_j, \mathcal{Q}_{NL}) - \mathbf{s}_j\| \leqslant \delta_0 \sqrt{d} + (1 + \delta_0 \sqrt{d} + \zeta)\zeta \leqslant \delta_0 \sqrt{d} + 1.5\zeta$$

which implies $\|\boldsymbol{\xi}_{a,i}\| \leqslant \delta_0 \sqrt{d} + 1.5\zeta$. Next, we bound $\boldsymbol{\xi}_{1,i} = \mathsf{Q}(\mathbf{u}_{i-1}, \mathcal{Q}_{NL}) - \mathbf{u}_{i-1}$. By Lemma A.9,

$$\|\boldsymbol{\xi}_{1,i}\| \leqslant \delta_0 \sqrt{d} + \zeta \|\mathbf{u}_{i-1}\| = \delta_0 \sqrt{d} + \zeta.$$

Finally we bound $\boldsymbol{\xi}_{2,i}$. Recall that:

$$\begin{aligned}
\mathbf{y}_i &= \frac{\sum_{j \in \mathcal{B}_j} \mathbf{X}_j (\mathbf{X}_j^T \mathbf{w}_i)}{n/b} + \boldsymbol{\xi}_{a,i} \\
\boldsymbol{\xi}_{2,i} &= \mathsf{Q}(\mathbf{y}_i, \delta) - \mathbf{y}_i
\end{aligned}$$

Since each $\left\|\mathbf{X}_j \mathbf{X}_j^\top \mathbf{w}_i\right\| \leqslant 1 + \delta_0 \sqrt{d} + \zeta$,

$$\|\mathbf{y}_i\| \leqslant 1 + \delta_0 \sqrt{d} + \zeta + \|\boldsymbol{\xi}_{a,i}\| \leqslant 1 + 2\delta_0 \sqrt{d} + 2.5\zeta \leqslant 3.25.$$

By Lemma A.9,

$$\|\boldsymbol{\xi}_{2,i}\| \leqslant \delta_0 \sqrt{d} + \zeta \|\mathbf{y}_i\| \leqslant \delta_0 \sqrt{d} + 3.25\zeta.$$

In all, it follows that

$$\|\boldsymbol{\Xi}_i\| \leqslant \eta \|\boldsymbol{\xi}_{a,i}\| + \|\boldsymbol{\xi}_{2,i}\| + (1 + \eta)\|\boldsymbol{\xi}_{1,i}\| \leqslant (\delta_0 \sqrt{d} + 1.5\zeta) + (\delta_0 \sqrt{d} + 3.25\zeta) + 2(\delta_0 \sqrt{d} + \zeta) \leqslant 4\delta_0 \sqrt{d} + 7\zeta =: \kappa.$$

We are ready to obtain the sin-squared error. Note that $\mathcal{M} \leqslant 2$, since $\|\mathbf{X}_i\| \leqslant 1$ almost surely, for all $i \in [n]$. By Theorem A.4, with probability at least $1 - 3\theta$,

$$\sin^2(\mathbf{w}, \mathbf{v}_1) \leqslant \frac{24 \log(1/\theta)}{\theta^3} \left[\frac{d}{\exp(2\alpha \log(n))} + \frac{5\left(\eta^2 \mathcal{V}_0 + \kappa_1\right)}{\eta(\lambda_1 - \lambda_2)} \right] + 8\kappa_1.$$

as long as $\sqrt{2e^2 b\gamma \log(d/\theta)} \leqslant \frac{1}{2}$. Our parameter choices are $\mathcal{V}_0 = \frac{b\mathcal{V}}{n}, \kappa = 4\delta_0 \sqrt{d} + 7\zeta$, and $\kappa_1 = (4\delta_0 \sqrt{d} + 7\zeta)^2$. Therefore,

$$\sin^2(\mathbf{w}, \mathbf{v}_1) \leqslant \frac{24 \log(1/\theta)}{\theta^3} \left[\frac{d}{n^{2\alpha}} + \frac{5\alpha \mathcal{V} \log n}{n(\lambda_1 - \lambda_2)^2} + \frac{5b(4\delta_0 \sqrt{d} + 7\zeta)^2}{\alpha \log n} \right] + 8(4\delta_0 \sqrt{d} + 7\zeta)^2.$$

□

D.2.1 Finishing the Proofs of Theorems 2 and 3

Proof of Theorem 2. For the linear quantization scheme, we apply Theorem A.5 with $\theta = 1/30$ and $b = \Theta\left(\frac{\alpha^2 \log^2 n \log d}{(\lambda_1 - \lambda_2)^2}\right)$. Moreover, since $\delta = \tilde{O}\left(\frac{\lambda_1 - \lambda_2}{\alpha\sqrt{d}}\right)$, the condition $\sqrt{4e^2 b(4\eta^2 + 9\delta^2 d)\log(d/\theta)} \leqslant \frac{1}{2}$ holds. The Theorem follows by substituting these values into the bound of Theorem A.5.

The proof of non-linear quantization scheme follows analogously by using Theorem A.6. □

Proof of Theorem 3. We set $\theta = 1/30$. For the linear quantization scheme, we apply Theorem A.5 with $b = n$. Moreover, since $\delta = 2^{2-\beta} = O\left(\min\left(\frac{\lambda_1 - \lambda_2}{\alpha\sqrt{d}\log(n)}, \frac{1}{\sqrt{dn}}\right)\right)$, the condition $\sqrt{4e^2 b(4\eta^2 + 9\delta^2 d)\log(d/\theta)} \leqslant \frac{1}{2}$ holds. The Theorem follows by substituting these values into the bound of Theorem A.5.

For the non-linear scheme, the proof follows analogously by applying Theorem A.6. □

E Proof of Boosting Lemma (Lemma 3)

In this section, we present the proof of the boosting procedure. The novelty of our analysis of the boosting lemma is that we use a different quantization scale ϵ. For ease of exposition, we present all proofs such that $\tilde{\rho}$ uses the extended quantization grid

$$\mathcal{Q}_L^*(\epsilon) = \{k\epsilon : k \in \mathbb{Z}\}.$$

which ensures $\left|\mathsf{Q}(\sin^2(\mathbf{u}_i, \mathbf{u}_j), \mathcal{Q}_L^*(\epsilon)) - x\right| \leqslant \epsilon$ almost surely. However, this requires a larger number of bits than needed. As we note later in Remark A.1, a restricted quantization grid

$$\mathcal{Q}_L(\epsilon) = \{-2^{\beta-1}\epsilon, -(2^{\beta-1}-1)\epsilon, \ldots, -\epsilon, 0, \epsilon, \ldots, (2^{\beta-1}-1)\epsilon\}.$$

which only quantizes values between $[-2^{\beta-1}, (2^{\beta-1}-1)\epsilon]$ and projects to the ends otherwise, suffices because the comparisons made in Algorithm 2 match exactly for both quantization grids. This requires a modest assumption that $\beta \geqslant 4$ which is already assumed in Section 3.4 while optimizing the parameters.

Proof of Lemma 3

Proof. For any $i \in [r]$, define the indicator random variable $\chi_i := \mathbb{1}\left(\sin^2(\mathbf{u}_i, \mathbf{v}) \leqslant \epsilon\right)$. Then, $\Pr(\chi_i = 1) \geqslant 1 - p$, where $p = 0.1$. Let $\mathcal{S} := \{i \in [r] : \chi_i = 1\}$, and define the event

$$\mathcal{E} := \{|\mathcal{S}| > 0.6r\}.$$

The Chernoff bound for the sum of independent Bernoulli random variables gives

$$\mathbb{P}\left(|\mathcal{S}| \leqslant (1-\theta)\,\mathbb{E}\left[|\mathcal{S}|\right]\right) \leqslant \exp\left(-\frac{\theta^2 \mathbb{E}\left[|\mathcal{S}|\right]}{2}\right) \;\forall\, \theta \in (0,1).$$

By linearity of expectation, $\mathbb{E}\left[|\mathcal{S}|\right] \geqslant (1-p)r$. Setting $\theta = 1/3$,

$$\mathbb{P}\left(\mathcal{E}^c\right) \leqslant \mathbb{P}\left(|\mathcal{S}| \leqslant 0.6r\right) \leqslant e^{-r/20} \leqslant \delta.$$

It suffices to show that if the event $\mathcal{E}$ holds, then $\bar{\mathbf{u}}$ is well-defined and has small sin-squared error with $\mathbf{v}$. Recall,

$$\bar{\mathbf{u}} := \mathbf{u}_i \text{ such that } \left|\{j \in [r] : \tilde{\rho}\left(\mathbf{u}_i, \mathbf{u}_j\right) \leqslant 5\epsilon\}\right| \geqslant 0.5r,$$

Conditioned on $\mathcal{E}$, such a $\bar{\mathbf{u}}$ always exists because $\mathbf{u}_i$ such that $i \in \mathcal{S}$ is a valid choice. Indeed, for any $i, j \in \mathcal{S}$,

$$\left|\tilde{\rho}\left(\mathbf{u}_i, \mathbf{u}_j\right)\right| \leqslant \sin^2\left(\mathbf{u}_i, \mathbf{u}_j\right) + \epsilon \leqslant 2\sin^2\left(\mathbf{u}_i, \mathbf{v}\right) + 2\sin^2\left(\mathbf{v}, \mathbf{u}_j\right) + \epsilon \leqslant 5\epsilon.$$

Since $\tilde{\rho}(\bar{\mathbf{u}}, \mathbf{u}_j) \leqslant 5\epsilon$ for at least $0.5r$ indices $j \in [r]$ and $|\mathcal{S}| \geqslant 0.6r$, there exists some index $j \in \mathcal{S}$ such that $\tilde{\rho}(\bar{\mathbf{u}}, \mathbf{u}_j) \leqslant 5\epsilon$. Therefore,

$$\sin^2(\bar{\mathbf{u}}, \mathbf{v}) \leqslant 2\sin^2(\bar{\mathbf{u}}, \mathbf{u}_j) + 2\sin^2(\mathbf{u}_j, \mathbf{v}) \leqslant 2\tilde{\rho}(\bar{\mathbf{u}}, \mathbf{u}_j) + 2\epsilon + 2\sin^2(\mathbf{u}_j, \mathbf{v}) \leqslant 14\epsilon.$$

□

Theorem A.7. *Suppose $\mathcal{A}$ is the batched Oja's algorithm with the setting of Theorem 2. Let ϵ be the probability 0.9 error bound guaranteed by Theorem 2, $r = \lceil 20\log(1/\theta)\rceil$, and $m = nr$. Let $\{\mathbf{X}_i\}_{i\in[m]}$ be n IID data drawn from a distribution satisfying assumption 1, and $\mathbf{u}_j \leftarrow \mathcal{A}(\{\mathbf{X}_i\}_{(j-1)n+1\leqslant i\leqslant jn}$ for all $j \in [r]$. Define the function $\tilde{\rho}$ as $\tilde{\rho}(\mathbf{x},\mathbf{y}) = Q\left(\sin^2(\mathbf{x},\mathbf{y}), \mathcal{Q}_L^*(\epsilon)\right)$, where $\mathcal{Q}_L^*(\epsilon)$ is the linear quantization grid with spacing ϵ. Then, with probability at least $1-\theta$, $\bar{\mathbf{u}} :=$ SuccessBoost$(\{\mathbf{u}_i\}_{i\in[r]}, \tilde{\rho}, \epsilon)$ satisfies*

$$\sin^2(\bar{\mathbf{u}}, \mathbf{v}_1) \leqslant 14\epsilon.$$

Proof. The vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are mutually independent. By Theorem 2, $\Pr\left(\sin^2(\mathbf{u}_i, \mathbf{v}_1) > \epsilon\right) \leqslant 0.1\ \forall\, i \in [r]$. Finally, since the quantization grid while computing $\sin^2(\mathbf{x},\mathbf{y})$ for each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ has scale ϵ, $\left|\tilde{\rho}(\mathbf{u}_i,\mathbf{u}_j) - \sin^2(\mathbf{x},\mathbf{y})\right| \leqslant \epsilon$. Therefore, all conditions of Lemma 3 apply, and the theorem follows. □

Remark A.1. *The quantization scheme for aggregation of the $\Theta(\log 1/\theta)$ vectors in the boosting algorithm is different from that of the main algorithm (which uses either a linear scheme with parameter δ or a logarithmic scheme with parameters ζ and δ_0). The quantization grid in the boosting algorithm is linear with parameter ϵ, where ϵ is the high-probability error guarantee. If the number of bits β to represent any number is fixed,*

$$\mathcal{Q}_L(\epsilon) = \{-2^{\beta-1}\epsilon, -(2^{\beta-1}-1)\epsilon, \ldots, -\epsilon, 0, \epsilon, \ldots, (2^{\beta-1}-1)\epsilon\}.$$

Since $\sin^2(\mathbf{u}_i,\mathbf{v}_1) \leqslant \epsilon$ with probability at least 0.9, most values of $\sin^2(\mathbf{u}_i,\mathbf{u}_j)$ will be at most 4ϵ using Lemma A.5. Any j for which $\sin^2(\mathbf{u}_i,\mathbf{u}_j)$ is outside this range of representable values has a large $\sin^2$ error with $\mathbf{u}_i$ and therefore not contained in the set $\{j \in [r] : \tilde{\rho}(\mathbf{u}_i,\mathbf{u}_j) \leqslant 5\epsilon\}$. Rigorously, consider the extended quantization grid

$$\mathcal{Q}_L^*(\epsilon) = \{k\epsilon : k \in \mathbb{Z}\}.$$

Then, $\left|Q(\sin^2(\mathbf{u}_i,\mathbf{u}_j), \mathcal{Q}_L^*(\epsilon)) - x\right| \leqslant \epsilon$ almost surely. Moreover, if $\sin^2(\mathbf{u}_i,\mathbf{u}_j)$ is contained in the range of $\mathcal{Q}_L(\epsilon)$, then $Q(\sin^2(\mathbf{u}_i,\mathbf{u}_j), \mathcal{Q}_L(\epsilon)) = Q(\sin^2(\mathbf{u}_i,\mathbf{u}_j), \mathcal{Q}_L^*(\epsilon))$. Therefore,

$$\mathbb{1}(Q(\sin^2(\mathbf{u}_i,\mathbf{u}_j), \mathcal{Q}_L(\epsilon)) \leqslant 5\epsilon) = \mathbb{1}(Q(\sin^2(\mathbf{u}_i,\mathbf{u}_j), \mathcal{Q}_L^*(\epsilon)) \leqslant 5\epsilon).$$

That is, by not quantizing $\tilde{\rho}(\mathbf{u}_i,\mathbf{u}_j)$ indices j for which $\sin^2(\mathbf{u}_i,\mathbf{u}_j)$ is greater than $(2^{\beta-1}-1)\epsilon$ in magnitude but simply not counting j as a neighbor of i in the boosting procedure, the algorithm is unaffected. This requires a modest bound on β such as $\beta \geqslant 4$, (because $2^{\beta-1}-1 = 7 > 5$), which is already assumed in Section 3.4 while optimizing the parameters.

F Experimental Details

Given the sample size n, dimension d, and decay exponent λ in the eigenvalues, we first draw an $n \times d$ matrix Z with independent entries uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$ so that each coordinate has unit variance. We then build a kernel matrix $K \in \mathbb{R}^{d\times d}$ with entries $K_{ij} = \exp\left(-|i-j|^{0.01}\right)$ and define a variance profile $\sigma_i = 5\, i^{-\lambda}$ for $i = 1, \ldots, d$. The population covariance is formed as $\Sigma = (\sigma\sigma^\top) \circ K$, where $\circ$ denotes the Hadamard product. Computing the eigendecomposition of Σ yields its square root $\Sigma^{1/2}$, and the observed data matrix is taken as $X = \left(\Sigma^{1/2} Z^\top\right)^\top$. We then extract the largest two eigenvalues $\lambda_1 > \lambda_2$ of Σ and the associated top eigenvector v_1 for evaluation.

G Related Work

In this section, we provide some more related work on low-precision optimization. For an excellent survey and history of quantization, see [GKD+22]. Dettmers *et al.* introduced QLoRA, which back-propagates through a frozen 4-bit quantized LLM into LoRA modules, enabling efficient finetuning of 65B-parameter models on a single 48 GB GPU with full 16-bit performance retention [DPHZ23]. Earlier works [XMHK23] examined the impact of stochastic roundoff errors and their bias on gradient descent convergence under low-precision arithmetic. Yu *et al.* propose Collage, a lightweight low-precision scheme for LLM training in distributed settings, combining block-wise quantization with error-feedback to stabilize large-scale pretraining [YGG+24]. Finally, communication-efficient distributed SGD techniques, such as 1-bit SGD with error feedback [SFD+14] and randomized

sketching primitives (e.g., Johnson–Lindenstrauss projections [JL84]), further underscore the broad efficacy of low-precision computation.

Low-Precision Optimization: Reducing the bit-width of model parameters and gradient updates has proven effective for alleviating communication and memory bottlenecks in large-scale learning. QSGD [AGL+17] uses randomized rounding to compress each coordinate to a few bits while preserving unbiasedness, incurring only an $O(\sqrt{d}/2^b)$ increase in gradient noise for b bits. TernGrad [WXY+17] maps gradients to $\{-1, 0, +1\}$ plus a shared scale and demonstrates negligible accuracy loss on ImageNet and CIFAR benchmarks. Suresh *et al.* [SYKM17] achieve optimal communication–accuracy trade-offs via randomized rotations and scalar quantization. More recently, "dimension-free" analyses such as Li & De Sa [LDS19] avoid scaling the required error rate with model dimension, instead depending on a suitably defined smoothness parameter.

Low-Precision PCA: Streaming PCA methods update eigenvector estimates incrementally, using $O(dp)$ memory to maintain the top p components in d dimensions. Oja's rule [Oja82] implements a Hebbian-style one-pass update with normalization. However, to our knowledge, *no general analysis* of Oja's algorithm under limited precision exists before our work.

References

[AGL+17] 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 1707–1718, 2017.

[AZL17] Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–492. IEEE, 2017.

[CYWZ18] Minshuo Chen, Lin Yang, Mengdi Wang, and Tuo Zhao. Dimensionality reduction for stationary time series via stochastic nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[DPHZ23] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023. https://arxiv.org/abs/2305.14314.

[GKD+22] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[HCS+16] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in Neural Information Processing Systems*, 2016.

[Heb49] Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, New York, 1949.

[HNWTW20] De Huang, Jonathan Niles-Weed, Joel A. Tropp, and Rachel Ward. Matrix concentration for products, 2020.

[HW19] Amelia Henriksen and Rachel Ward. AdaOja: Adaptive Learning Rates for Streaming PCA. *arXiv e-prints*, page arXiv:1905.12115, May 2019.

[JJK+16] Prateek Jain, Chi Jin, Sham Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Proceedings of The 29th Conference on Learning Theory (COLT)*, June 2016.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In *Contemporary Mathematics*, volume 26, page 189–206, 1984.

[KLL+23] Jonathan Kelner, Jerry Li, Allen X Liu, Aaron Sidford, and Kevin Tian. Semi-random sparse recovery in nearly-linear time. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2352–2398. PMLR, 2023.

[KS24a] Syamantak Kumar and Purnamrita Sarkar. Oja's algorithm for streaming sparse pca. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[KS24b] Syamantak Kumar and Purnamrita Sarkar. Streaming pca for markovian data. *Advances in Neural Information Processing Systems*, 36, 2024.

[KWW+17] Urs Köster, Tristan J. Webb, Xin Wang, Marcel Nassar, Arjun K. Bansal, William H. Constable, Oğuz H. Elibol, Scott Gray, Stewart Hall, Luke Hornof, Amir Khosrowshahi, Carey Kloss, Ruby J. Pai, and Naveen Rao. Flexpoint: An adaptive numerical format for efficient training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 1742–1750, 2017.

[LD19] Zheng Li and Christopher M. De Sa. Dimension–free bounds for low-precision training. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 11728–11738, 2019.

[LDS19] Zheng Li and Christopher De Sa. Dimension-free bounds for low-precision training. In *Advances in Neural Information Processing Systems*, 2019.

[LDX+17] Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding. In *Advances in Neural Information Processing Systems*, pages 5813–5823, 2017.

[LSW21] Robert Lunde, Purnamrita Sarkar, and Rachel Ward. Bootstrapping the error of oja's algorithm. *Advances in Neural Information Processing Systems*, 34:6240–6252, 2021.

[MNA+18] Paulius Micikevicius, Sharan Narang, Gabriel Alben, Gregory Diamos, Erich Elsen, David Garcia, Dmitry Ginsburg, Michael Houston, Oleksii Kuchaiev, Sanjo Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

[Mon22] Jean-Marie Monnez. Stochastic approximation of eigenvectors and eigenvalues of the q-symmetric expectation of a random matrix. *Communications in Statistics-Theory and Methods*, pages 1–15, 2022.

[MP22] Nikos Mouzakis and Eric Price. Spectral guarantees for adversarial streaming pca, 2022.

[NTSW+22] Miloš Nikolić, Enrique Torres Sanchez, Jiahui Wang, Ali Hadi Zadeh, Mostafa Mahmoud, Ameer Abdelhadi, Kareem Ibrahim, and Andreas Moshovos. Schrödinger's fp: Dynamic adaptation of floating-point containers for deep learning training. *arXiv preprint arXiv:2204.13666*, 2022.

[Oja82] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15:267–273, 1982.

[OYP25] Kaan Ozkara, Tao Yu, and Youngsuk Park. Stochastic rounding for llm training: Theory and practice. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. https://arxiv.org/abs/2502.20566.

[Pea01] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[Rie67] Bernhard Riemann. *Ueber die Darstellbarkeit einer Function durch eine trigonometrische Reihe*. Dieterich, 1867. In German.

[SFD+14] 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 *Interspeech*, 2014.

[She97] William Fleetwood Sheppard. On the calculation of the most probable values of frequency-constants for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society*, 1(1):353–380, 1897.

[SLZ+18] Christopher De Sa, Megan Leszczynski, Jian Zhang, Alana Marzoev, Christopher R. Aberger, Kunle Olukotun, and Christopher Ré. High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*, 2018.

[Sti86] S. M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, 1986.

[SYK21] Heming Sun, Lu Yu, and Jiro Katto. Learned image compression with fixed-point arithmetic. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021.

[SYKM17] Ananda Theertha Suresh, Felix X. Yu, Harsha Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. *arXiv preprint arXiv:1611.00349*, 2017.

[SZOR15] Christopher M. De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild-style algorithms. In *Advances in Neural Information Processing Systems*, pages 2674–2682, 2015.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Wed72] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

[WXY+17] Wei Wen, Chunpeng Xu, Felix Yan, Chunyi 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*, 2017.

[XLY+24] Yongqi Xu, Yujian Lee, Gao Yi, Bosheng Liu, Yucong Chen, Peng Liu, Jigang Wu, Xiaoming Chen, and Yinhe Han. Bitq: Tailoring block floating point precision for improved dnn efficiency on resource-constrained devices. *arXiv preprint arXiv:2409.17093*, 2024.

[XMHK23] Lu Xia, Stefano Massei, Michiel E. Hochstenbach, and Barry Koren. On the influence of stochastic roundoff errors and their bias on the convergence of the gradient descent method with low-precision floating-point computation, 2023.

[Yat09] Randy Yates. Fixed-point arithmetic: An introduction. *Digital Signal Labs*, 81(83):198, 2009.

[YGG+24] Tao Yu, Gaurav Gupta, Karthick Gopalswamy, Amith R. Mamidala, Hao Zhou, Jeffrey Huynh, Youngsuk Park, Ron Diamant, Anoop Deoras, and Luke Huan. Collage: Lightweight low-precision strategy for llm training. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[YHW18] Puyudi Yang, Cho-Jui Hsieh, and Jane-Ling Wang. History pca: A new algorithm for streaming pca. *arXiv preprint arXiv:1802.05447*, 2018.

[YIY21] Hisakatsu Yamaguchi, Makiko Ito, and Katsuhiro Yoda. Training deep neural networks in 8-bit fixed point with dynamic shared exponent management. In *Proceedings of the 2021 Design, Automation & Test in Europe Conference (DATE)*, 2021.

[Zie03] Eric R Ziegel. Principal component analysis. *Technometrics*, 45(3):276–277, 2003.

[ZLK+17] Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4035–4043, 2017.

[ZMK22] Sai Qian Zhang, Bradley McDanel, and T. Kung, H.˙ Fast: Dnn training under variable precision block floating point with stochastic rounding. In *Proceedings of the 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 846–860, 2022.

[ZWG+23] Jiajun Zhou, Jiajun Wu, Yizhao Gao, Yuhao Ding, Chaofan Tao, Boyu Li, Fengbin Tu, Kwang-Ting Cheng, Hayden Kwok-Hay So, and Ngai Wong. Dybit: Dynamic bit-precision numbers for efficient quantized neural network inference. *arXiv preprint arXiv:2302.12510*, 2023.