Supplementary Material

A. Auxillary Lemmas

Lemma A.1 (Lemma 14.1 (Shalev-Shwartz & Ben-David, 2014))**.** Any update of the form

$$\mathrm{P}^{(t+1)} = \Pi_{\mathcal{C}}(\mathrm{P}^{(t)} - \eta \mathrm{g}_t), \tag{13}$$

for an arbitrary sequence of matrices $\mathrm{g}_1, \mathrm{g}_2, \ldots, \mathrm{g}_T$, projection $\Pi_{\mathcal{C}}$ onto an arbitrary convex set $\mathcal{C}$, and initialization $\mathrm{P}^{(1)} = 0$ satisfies

$$\sum_{t=1}^{T} \left\langle \mathrm{P}^{(t)} - \mathrm{P}, \mathrm{g}_t \right\rangle \leq \frac{\|\mathrm{P}\|_F^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathrm{g}_t\|_F^2, \tag{14}$$

for any $\mathrm{P} \in C$.

Lemma A.2 (Lemma 4 (Warmuth & Kuzmin, 2006))**.** For any PSD A and any symmetric B, C for which $\mathrm{B} \preceq \mathrm{C}$, it holds that $\mathrm{Tr}\,(\mathrm{AB}) \leq \mathrm{Tr}\,(\mathrm{AC})$.

B. Proofs of Section 3

Proof of Theorem 2.1. Using Lemma A.1, noting $\|\mathrm{P}\|_F^2 \leq k$ for all $\mathrm{P} \in \mathcal{C}$, and $\mathrm{g}_t = -\mathrm{x}_t\mathrm{x}_t^\top$ so that $\|\mathrm{g}_t\|_F^2 = \|\mathrm{x}_t\|_2^4 \leq 1$,

$$\sum_{t=1}^{T} \mathrm{x}_t^\top \mathrm{P}\mathrm{x}_t - \sum_{t=1}^{T} \mathrm{x}_t^\top \mathrm{P}^{(t)}\mathrm{x}_t \leq \frac{k}{2\eta} + \frac{\eta T}{2}.$$

Choosing $\eta = \sqrt{\frac{k}{T}}$ completes the proof. □

Proof of Theorem 3.1. Let $\widehat{\mathrm{g}}_t = \mathrm{g}_t + \mathrm{E}_t$ be the noisy gradient. The analysis begins with bounding the distance between the t-th iterate and any candidate $\mathrm{P} \in \mathcal{C}$, $D_t = \|\mathrm{P}_t - \mathrm{P}\|_F$.

$$\begin{aligned}
D_{t+1}^2 &= \left\|\mathrm{P}^{(t+1)} - \mathrm{P}\right\|_F^2 = \left\|\Pi_{\mathcal{C}}(\mathrm{P}^{(t)} - \eta\widehat{\mathrm{g}}_t) - \mathrm{P}\right\|_F^2 \\
&\leq \left\|\mathrm{P}^{(t)} - \eta\widehat{\mathrm{g}}_t - \mathrm{P}\right\|_F^2 \\
&= \left\|\mathrm{P}^{(t)} - \mathrm{P}\right\|_F^2 + \eta^2 \|\widehat{\mathrm{g}}_t\|_F^2 \\
&\quad - 2\eta\langle \mathrm{P}^{(t)} - \mathrm{P}, \mathrm{g}_t + \mathrm{E}_t\rangle \\
&\leq D_t^2 + \eta^2 G^2 - 2\eta\langle \mathrm{P}^{(t)} - \mathrm{P}, \mathrm{g}_t\rangle \\
&\quad - 2\eta\langle \mathrm{P}^{(t)} - \mathrm{P}, \mathrm{E}_t\rangle \\
&\leq D_t^2 + \eta^2 G^2 - 2\eta\langle \mathrm{P}^{(t)} - \mathrm{P}, \mathrm{g}_t\rangle \\
&\quad + 2\eta \left\|\mathrm{P}^{(t)} - \mathrm{P}\right\|_* \|\mathrm{E}_t\|_2 \\
&\leq D_t^2 + \eta^2 G^2 - 2\eta\langle \mathrm{P}^{(t)} - \mathrm{P}, \mathrm{g}_t\rangle + 4k\eta \|\mathrm{E}_t\|_2
\end{aligned}$$

Where the first inequality is due to projection onto a convex set being non-expanding and the third one is by Holder's inequality. The last inequality follows from triangle inequality and constraints: $\left\|\mathrm{P}^{(t)} - \mathrm{P}\right\|_* \leq \left\|\mathrm{P}^{(t)}\right\|_* + \|\mathrm{P}\|_* \leq 2k$.

Noting $\mathrm{g}_t = -\mathrm{x}_t\mathrm{x}_t^\top$, rearranging and dividing both sides by 2η we get

$$\langle \mathrm{P} - \mathrm{P}^{(t)}, \mathrm{x}_t\mathrm{x}_t^\top\rangle \leq \frac{D_t^2 - D_{t+1}^2}{2\eta} + \frac{\eta}{2}G^2 + 2k\|\mathrm{E}_t\|_2 \tag{15}$$

We then sum over T iterates and use the telescopic property of the first term of the right hand side of 15, so that $\sum_{t=1}^{T} D_t^2 - D_{t+1}^2 = D_1^2 - D_{T+1}^2 \leq D_1^2$. The initial distance D_1 is bounded as follows:

$$\begin{aligned}
D_1^2 &= \left\|\mathrm{P}^{(1)} - \mathrm{P}\right\|_F^2 = \left\|\mathrm{P}^{(1)}\right\|_F^2 + \|\mathrm{P}\|_F^2 - 2\langle \mathrm{P}^{(1)}, \mathrm{P}\rangle \\
&\leq k + k + 2\left\|\mathrm{P}^{(1)}\right\|_* \|\mathrm{P}\|_2 \leq 4k
\end{aligned}$$

Plugging back the above bound in 15 and noting $\sum_{t=1}^{T} \|\mathrm{E}_t\|_2 \leq E$, we get

$$\sum_{t=1}^{T} \mathrm{x}_t^\top \mathrm{P}\mathrm{x}_t - \sum_{t=1}^{T} \mathrm{x}^\top \mathrm{P}^{(t)}\mathrm{x} \leq \frac{2k}{\eta} + \frac{\eta G^2 T}{2} + 2kE$$

By choosing optimal learning rate $\eta = \frac{2\sqrt{k}}{G\sqrt{T}}$ we get the desired results. □

Proof of Theorem 3.2. Let $\widehat{\mathrm{g}}_k = (\mathrm{x}_k + \mathrm{y}_k)(\mathrm{x}_k + \mathrm{y}_k)^\top$ denote the corrupted gradient and let $\mathrm{g}_k = \mathrm{x}_k\mathrm{x}_k^\top$ denote the unbiased estimate of C based on the k-th sample. Let v denote the top eigenvector of $\mathrm{C} := \mathbb{E}_{\mathcal{D}}[\mathrm{x}\mathrm{x}^\top]$ with associated eigenvalue λ and let $\mathrm{u} \sim \mathcal{N}(0, \mathrm{I})$ be the initialization for Oja's algorithm. We are going to assume that $\sum_{k=1}^{T} \|\mathrm{y}_k\| + \|\mathrm{y}_k\|^2 \leq \sqrt{T}$. We note that our assumption implies that $\sum_{k=1}^{T} \|\widehat{\mathrm{g}}_k - \mathrm{g}_k\| \leq O(\sqrt{T})$ and $\sum_{k=1}^{T} \|\widehat{\mathrm{g}}_k\|^n \leq O(T^{(n/2)})$. Let Φ_k^{M} and Ψ_k be defined as in the proof of Theorem 5.3. We are first going to upper bound $\mathrm{Tr}\left(\Phi_T^{\mathrm{uu}^\top}\right)$:

$$\begin{aligned}
&\mathrm{Tr}\left(\Phi_T^{\mathrm{uu}^\top}\right) = \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{uu}^\top}\right) + 2\eta\mathrm{Tr}\left(\mathrm{x}_T\mathrm{x}_T^\top\Phi_{T-1}^{\mathrm{uu}^\top}\right) \\
&+ 2\eta\mathrm{Tr}\left((\widehat{\mathrm{g}}_T - \mathrm{g}_T)\Phi_{T-1}^{\mathrm{uu}^\top}\right) + \eta^2\mathrm{Tr}\left(\widehat{\mathrm{g}}_T^2\Phi_{T-1}^{\mathrm{uu}^\top}\right) \\
&\leq \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{uu}^\top}\right)\left(1 + 2\eta\mathrm{w}_T^\top\mathrm{g}_T\mathrm{w}_T + 2\eta\|\widehat{\mathrm{g}}_T - \mathrm{g}_T\| + \eta^2\|\widehat{\mathrm{g}}_T\|^2\right) \\
&\leq \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{uu}^\top}\right)\exp\left(2\eta\mathrm{w}_T^\top\mathrm{g}_T\mathrm{w}_T + 2\eta\|\widehat{\mathrm{g}}_T - \mathrm{g}_T\| + \eta^2\|\widehat{\mathrm{g}}_T\|^2\right) \\
&\leq \ldots \\
&\leq \|\mathrm{u}\|_2^2 \exp\left(\sum_{k=1}^{T} 2\eta\mathrm{w}_k^\top\mathrm{g}_k\mathrm{w}_k + 2\eta\|\widehat{\mathrm{g}}_k - \mathrm{g}_k\| + \eta^2\|\widehat{\mathrm{g}}_k\|^2\right),
\end{aligned} \tag{16}$$

where we have used lemma 4 in (Warmuth & Kuzmin, 2006).

Next we are going to lower bound $\mathbb{E}[\mathrm{Tr}\left(\Psi_1\right)]$:

$$\begin{aligned}
&\mathbb{E}[\mathrm{Tr}\left(\Psi_1\right)] \geq \mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top\left(\mathrm{I}+2\eta\widehat{\mathrm{g}}_T\right)\Phi^{\mathrm{I}}_{T-1}\right)] \\
&= \mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top\left(\mathrm{I}+2\eta\mathrm{C}\right)\Phi^{\mathrm{I}}_{T-1}\right)] \\
&+ \mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top\left(\mathrm{I}+2\eta\mathrm{y}_T\mathrm{y}_T^\top\right)\Phi^{\mathrm{I}}_{T-1}\right)] \\
&\geq (1+2\eta\lambda)\mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top\Phi^{\mathrm{I}}_{T-1}\right)] \\
&\geq \exp\left(2\eta\lambda-2\eta^2\lambda^2\right)\mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top\Phi^{\mathrm{I}}_{T-1}\right)] \geq \cdots \\
&\geq \exp\left(T(2\eta\lambda-2\eta^2\lambda^2)\right),
\end{aligned} \tag{17}$$

where we used the fact that $\mathbb{E}[\mathrm{x}_k^\top]=0$ and that $1+2x\geq\exp\left(2x-2x^2\right)$ for $x\in[0,1]$. Finally using Lieb-Thirring's inequality we have the following upper bound on $\mathbb{E}[\mathrm{Tr}\left(\Psi_1^2\right)]$:

$$\begin{aligned}
&\mathbb{E}[\mathrm{Tr}\left(\Psi_1^2\right)] \leq \mathbb{E}[\mathrm{Tr}\left(\left(\mathrm{I}+\eta\widehat{\mathrm{g}}_1\right)^4\Psi_2^2\right)] \\
&= \mathbb{E}[\Big(\mathrm{I}+4\eta\mathrm{x}_1\mathrm{x}_1^\top+4\eta\mathrm{y}_1\mathrm{y}_1^\top+6\eta^2\widehat{\mathrm{g}}_1^2+4\eta^3\widehat{\mathrm{g}}_1^3 \\
&+\eta^4\widehat{\mathrm{g}}_1^4\Big)\mathrm{Tr}\left(\Psi_2^2\right)] \\
&\leq \exp\Big(4T\eta\lambda+4\alpha_1\eta\sqrt{T}+6\alpha_2\eta^2T+4\alpha_3\eta^3T^{(3/2)} \\
&+\alpha_4\eta^4T^2\Big),
\end{aligned} \tag{18}$$

where for the last inequality we have pushed the expectation inside the trace and used the fact that $\mathbb{E}[\|\widehat{\mathrm{g}}_1^n\|]$ is bounded above by some linear combinations of $\|\mathrm{y}_1\|^m$, where $m\in[2n]$. Next we have used the fact that $1+x\leq\exp(x)$ and induction together with $\sum_{k=1}^T\|\mathrm{y}_k\|\leq O(\sqrt{T})$ and $\sum_{k=1}^T\|\mathrm{y}_k\|^{2n}\leq O(T^{n/2})$. Let $\alpha=4\alpha_1+6\alpha_2\eta\sqrt{T}+4\alpha_3\eta^2T+\alpha_4\eta^3T^{3/2}+4\eta\lambda^2\sqrt{T}$. Using Chebyshev's inequality with equations 17 and 18 we have:

$$\begin{aligned}
&\mathbb{P}[\mathrm{Tr}\left(\Psi_1\right)\leq\exp\left(T(2\eta\lambda-2\eta^2\lambda^2)\right)(1 \\
&-\delta^{-1/2}\sqrt{\exp\left(\sqrt{T}\eta\alpha\right)-1})]\leq\delta.
\end{aligned}$$

Let $\eta=\frac{\beta}{\sqrt{T}}$ for some constant β independent of T. We claim that for small enough β it holds that with probability at least $1-\delta$, $\mathrm{Tr}\left(\Psi_1\right)\geq\frac{2}{3}\exp\left(T(2\eta\lambda-2\eta^2\lambda^2)\right)$. For the choice of η, this is equivalent to $\exp(\beta\alpha)\leq1+\frac{\delta}{9}$, where α now becomes $\alpha=4\alpha_1+6\alpha_2\beta+4\lambda^2\beta+4\alpha_3\beta^2+\alpha_4\beta^3$. Taking $\beta\to0$ we can see that $\exp(\alpha\beta)\to1<1+\delta/9$ and since $\exp(\alpha\beta)$ is a continuous function of β the desired value of $\beta>0$ exists. From the derivation in (Allen-Zhu & Li, 2017) Appendix I, it follows that $\|\mathrm{u}\|_2^2\leq O(d+\log(1/\delta))$ and $\mathrm{Tr}\left(\Phi_T^{\mathrm{uu}^\top}\right)\geq\Omega(\delta^{-2})\mathrm{Tr}\left(\Psi_1\right)$ with probability $1-\delta$. Thus with probability $1-2\delta$ it holds that

$$\begin{aligned}
&O(d+\log(1/\delta))\exp\left(\sum_{k=1}^T2\eta\mathrm{w}_k^\top\mathrm{g}_k\mathrm{w}_k+2\eta\|\widehat{\mathrm{g}}_k-\mathrm{g}_k\|\right. \\
&\left.+\eta^2\|\widehat{\mathrm{g}}_k\|^2\right)\geq\frac{2}{3\delta^2}\exp\left(T(2\eta\lambda-2\eta^2\lambda^2)\right)\implies \\
&\sum_{k=1}^T\langle\mathrm{g}_k,\mathrm{P}^{(*)}-\mathrm{P}^k\rangle\leq\frac{\log\left(O(d+\log(1/\delta))\right)}{\eta} \\
&+O(\sqrt{T})+\eta O(T)-\frac{\log\left(\frac{2}{3\delta^2}\right)}{\eta}.
\end{aligned}$$

Substituting $\eta=\frac{\beta}{\sqrt{T}}$ finishes the proof. □

C. Proofs of Section 4

Lemma C.1. The naive estimator $\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top$ is not unbiased, i.e.

$$\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]=q^2\mathrm{x}\mathrm{x}^\top+(q-q^2)\,\mathrm{diag}\,\mathrm{x}\mathrm{x}^\top,$$

where the expectation is taken with respect to Bernoulli random model.

Proof of Lemma C.1. For off-diagonal elements ($i\neq j$), $\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]_{i,j}=\mathrm{x}_i\mathrm{x}_j\mathbb{P}(i,j\in\{i_1,..,i_r\})=x_ix_j\mathbb{P}(Z_i=1)\mathbb{P}(Z_j=1)=q^2x_ix_j$. Now consider the diagonal elements. $\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]_{i,i}=\mathrm{x}_i^2\mathbb{P}(Z_i=1)=qx_i^2$. Putting together, $\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]=q^2\mathrm{x}\mathrm{x}^\top+(q-q^2)\,\mathrm{diag}\,\mathrm{x}\mathrm{x}^\top$. □

Proof of Lemma 4.1. First note that

$$\mathbb{E}_{\mathcal{S}}[\mathrm{z}\mathrm{z}^\top|\tilde{\mathrm{x}}]=\frac{r-rq}{q^2}\mathbb{E}_{\mathcal{S}}[\tilde{x}_{i_s}^2\mathrm{e}_{i_s}\mathrm{e}_{i_s}^\top|\tilde{\mathrm{x}}]=\frac{1-q}{q^2}\,\mathrm{diag}\,\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top.$$

The proof simply follows from the following equalities.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},\mathcal{R}}[\widehat{\mathrm{x}}\widehat{\mathrm{x}}^\top-\mathrm{z}\mathrm{z}^\top|\mathrm{x}] \\
&=\mathbb{E}_{\mathcal{R}}[\mathbb{E}_{\mathcal{S}}[\frac{1}{q^2}\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\tilde{\mathrm{x}}]|\mathrm{x}]-\mathbb{E}_{\mathcal{R}}[\mathbb{E}_{\mathcal{S}}[\mathrm{z}\mathrm{z}^\top|\tilde{\mathrm{x}}]|\mathrm{x}] \\
&=\frac{1}{q^2}\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]-\frac{1-q}{q^2}\mathbb{E}_{\mathcal{R}}[\mathrm{diag}\,\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}] \\
&=\mathrm{x}\mathrm{x}^\top+\frac{q-q^2}{q^2}\,\mathrm{diag}\,\mathrm{x}\mathrm{x}^\top-\frac{1-q}{q}\,\mathrm{diag}\,\mathrm{x}\mathrm{x}^\top \\
&=\mathrm{x}\mathrm{x}^\top.
\end{aligned} \tag{19}$$

□

Proof of Theorem 4.2. We first bound $\|\widehat{\mathrm{g}}_t\|_F^2$ as follows:

$$\begin{aligned}\|\widehat{\mathrm{g}}_t\|_F^2 &= \left\|\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top - \mathrm{z}_t\mathrm{z}_t^\top\right\|_F^2 \\ &= \left\|\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top\right\|_F^2 + \left\|\mathrm{z}_t\mathrm{z}_t^\top\right\|_F^2 - 2\langle\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top, \mathrm{z}_t\mathrm{z}_t^\top\rangle \\ &\le \left\|\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top\right\|_F^2 + \left\|\mathrm{z}_t\mathrm{z}_t^\top\right\|_F^2 \\ &\le \|\widehat{\mathrm{x}}_t\|^4 + \|\mathrm{z}_t\|^4 \\ &\le \frac{1}{q^2}\|\tilde{\mathrm{x}}\|^4 + \frac{r_t^2(1-q)^2}{q^4}\|\mathrm{x}\|_\infty^4 \\ &\le \frac{1}{q^2} + \frac{r_t^2(1-q)^2}{q^4}\end{aligned}$$

where the first inequality follows from the fact that inner product of PSD matrices is non-negative, the third inequality is by definition of $\widehat{\mathrm{x}}$ and z, and the last inequality holds since $\|\mathrm{x}\|_\infty \le \|\mathrm{x}\| \le 1$. Taking the expectation of both sides, noting the for Binomial distribution $\mathbb{E}_{\mathcal{R}}[r_t^2] = \mathrm{var}(r_t) + \mathbb{E}_{\mathcal{R}}[r_t]^2 = dq(1-q) + d^2q^2$, we get $\mathbb{E}_{\mathcal{R}}[\|\widehat{\mathrm{g}}\|_F^2] \le \frac{1}{q^2} + \frac{dq(1-q)^3 + d^2q^2(1-q)^2}{q^4}$. Now, using Lemma A.1 with $\{\widehat{\mathrm{g}}_t\}_{t=1}^T$, noting $\|\mathrm{P}\|_F^2 \le k$ for all $\mathrm{P} \in C$

$$\begin{aligned}&\mathbb{E}[\sum_{t=1}^T\left\langle \mathrm{P}^* - \mathrm{P}^{(t)}, -\widehat{\mathrm{g}}_t\right\rangle] \\ &\le \frac{k}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T\frac{q^2 + dq(1-q)^3 + d^2q^2(1-q)^2}{q^4} \\ &\le \frac{k}{2\eta} + \frac{\eta}{2}T\frac{q^2 + dq(1-q)^3 + d^2q^2(1-q)^2}{q^4}.\end{aligned}$$

Setting $\eta = \frac{q^2}{\sqrt{q^2 + dq(1-q)^3 + d^2q^2(1-q)^2}}\sqrt{\frac{k}{T}}$ and taking expectation with respect to the internal randomization $\mathcal{S}$ and the distribution of the missing data $\mathcal{R}$ finishes the proof. □

Proof of Theorem Oja with missing entries 4.3. We begin by bounding $\mathbb{E}_{\mathcal{S},\mathcal{R}}[\|\widehat{\mathrm{g}}_k\|^2], \mathbb{E}_{\mathcal{S},\mathcal{R}}[\|\widehat{\mathrm{g}}_k\|^3]$ and $\mathbb{E}_{\mathcal{S},\mathcal{R}}[\|\widehat{\mathrm{g}}_k\|^4]$. Let $a := \left\|\widehat{\mathrm{x}}_k\widehat{\mathrm{x}}_k^\top\right\|$ and $b := \left\|\mathrm{z}_k\mathrm{z}_k^\top\right\|$. Using Jensen's inequality we have $\|\widehat{\mathrm{g}}_k\|^2 \le 2a^2 + 2b^2, \|\widehat{\mathrm{g}}_k\|^3 \le 4a^3 + 4b^3, \|\widehat{\mathrm{g}}_k\|^4 \le 8a^4 + 8b^4$. Since $\|\mathrm{x}\|_2 \le 1$ it holds that $a^n \le \frac{1}{q^n}$ and $b^n \le \frac{r_k^n(1-q)^n}{q^{2n}}$. Let

$$\begin{aligned}\mu_2 &= dq(1-q+dq) \\ \mu_3 &= dq(1-3q+3dq+2q^2-3dq^2+d^2q^2) \\ \mu_4 &= dq(1-7q+7dq+12q^2-18dq^2+6d^2q^2 \\ &\quad -6q^3+11dq^3-6d^2q^3+d^3q^3).\end{aligned}$$

These are the second, third and fourth moment of the binomial random variable r_k, respectively. Also let $\tilde{\alpha}_2^k = \frac{2}{q^2} + \frac{2r_k^2(1-q)^2}{q^4}$.Thus we have

$$\mathbb{E}_{\mathcal{S},\mathcal{R}}[\|\widehat{\mathrm{g}}_k\|^n] \le \alpha_n = \frac{2^{n-1}}{q^n} + \frac{2^{n-1}\mu_n(1-q)^n}{q^{2n}}.$$

Define Φ_k^{M}, Ψ_k, w_k, λ,v and u as in the proof of theorem 5.3. We have:

$$\begin{aligned}&\mathrm{Tr}\left(\Phi_T^{\mathrm{uu}^\top}\right) \\ &= \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{uu}^\top}\right) + 2\eta\mathrm{Tr}\left(\widehat{\mathrm{g}}_T\Phi_{T-1}^{\mathrm{uu}^\top}\right) + \eta^2\mathrm{Tr}\left(\widehat{\mathrm{g}}_T^2\Phi_{T-1}^{\mathrm{uu}^\top}\right) \\ &\le \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{uu}^\top}\right) + 2\eta\mathrm{Tr}\left(\widehat{\mathrm{g}}_T\Phi_{T-1}^{\mathrm{uu}^\top}\right) + \eta^2\tilde{\alpha}_2^T\mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{uu}^\top}\right) \\ &\le \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{uu}^\top}\right)\exp\left(\eta^2\tilde{\alpha}_2^T + 2\eta\mathrm{w}_T^\top\widehat{\mathrm{g}}_T\mathrm{w}_T\right) \\ &\cdots \le \|\mathrm{u}\|_2^2\exp\left(\sum_{t=1}^T\eta^2\tilde{\alpha}_2^t + 2\eta\sum_{t=1}^T\mathrm{w}_t^\top\widehat{\mathrm{g}}_t\mathrm{w}_t\right).\end{aligned} \tag{20}$$

Next we need a lower bound on $\mathbb{E}[\Psi_1]$. This is done in the same way as in Theorem 5.3:

$$\mathbb{E}[\mathrm{Tr}(\Psi_1)] \ge \exp\left(T(2\eta\lambda - 2\eta^2\lambda^2)\right), \tag{21}$$

Finally, we need to bound $\mathbb{E}[\Psi_1^2]$. It holds that:

$$\begin{aligned}&\mathbb{E}[\mathrm{Tr}\left(\Psi_1^2\right)] = \mathbb{E}[\mathrm{Tr}\left((I+\eta\widehat{\mathrm{g}}_1)^2\Psi_2(I+\eta\widehat{\mathrm{g}}_1)^2\Psi_2\right)] \\ &\le \mathbb{E}[\mathrm{Tr}\left((\mathrm{I}+\eta\widehat{\mathrm{g}}_1)^4\Psi_2^2\right)] \\ &= \mathbb{E}[\mathrm{Tr}\left((\mathrm{I}+4\eta^3\widehat{\mathrm{g}}_1^3 + 6\eta^2\widehat{\mathrm{g}}_1^2 + 4\eta\widehat{\mathrm{g}}_1 + \eta^4\widehat{\mathrm{g}}_1^4)\Psi_2^2\right)] \\ &\le \mathrm{Tr}\left((\mathrm{I}+\eta^4\alpha_4 + 4\eta^3\alpha_3 + 6\eta^2\alpha_2 + 4\eta\lambda)\Psi_2^2\right) \\ &\le \mathrm{Tr}\left((\mathrm{I}+\eta^2\alpha + 4\eta\lambda)\mathbb{E}[\Psi_2^2]\right) \\ &\le \exp\left(\alpha\eta^2 + 4\eta\lambda\right)\mathbb{E}[\mathrm{Tr}\left(\Psi_2^2\right)],\end{aligned} \tag{22}$$

where $\alpha = \alpha_4 + 4\alpha_3 + 6\alpha_2$. Using Chebyshev's inequality it holds that

$$\begin{aligned}&\mathbb{P}[\mathrm{Tr}(\Psi_1) \le \exp\left(T(2\eta\lambda - 2\eta^2\lambda^2)\right)(1 \\ &- \delta^{-1/2}\sqrt{\exp\left(\alpha\eta^2 + 4\eta^2\lambda^2\right)})] \le \delta.\end{aligned}$$

As long as $\eta \le \frac{\log(1+\delta/9)}{\sqrt{T}(\alpha+4\lambda^2)}$, this implies that with probability $1-\delta$, $\mathrm{Tr}(\Psi_1) \ge \frac{2}{3}\exp\left(T(2\eta\lambda - 2\eta^2\lambda^2)\right)$. From the derivation in (Allen-Zhu & Li, 2017) Appendix I, it follows that $\|\mathrm{u}\|_2^2 \le O(d + \log(1/\delta))$ and $\mathrm{Tr}\left(\Phi_T^{\mathrm{uu}^\top}\right) \ge \Omega(\delta^{-2})\mathrm{Tr}(\Psi_1)$ with probability $1-\delta$. Thus with probability $1-2\delta$ it holds that

$$\begin{aligned}&O(d+\log(1/\delta))\exp\left(\sum_{t=1}^T\eta^2\tilde{\alpha}_2^t + 2\eta\sum_{t=1}^T\mathrm{w}_t^\top\widehat{\mathrm{g}}_t\mathrm{w}_t\right) \\ &\ge \frac{2}{3\delta^2}\exp\left(T(2\eta\lambda - 2\eta^2\lambda^2)\right).\end{aligned}$$

Taking logarithms on both sides implies that

$$2\eta\lambda T - 2\eta\sum_{t=1}^{T}\mathrm{Tr}\left(\mathbb{E}_{\mathcal{S},\mathcal{R}}[\mathrm{w}_t\mathrm{w}_t^\top]\mathbb{E}_{\mathcal{S},\mathcal{R}}[\widehat{\mathrm{g}}_t]\right) \leq$$
$$\log(O(d+\log(1/\delta))) - \log\left(\frac{2}{3\delta^2}\right) + \eta^2\sum_{t=1}^{T}\mathbb{E}_{\mathcal{S},\mathcal{R}}[\tilde{\alpha}_2^t],$$

where we have used the independence between $\widehat{\mathrm{g}}_t$ and w_t. This implies

$$\sum_{t=1}^{T}\langle \mathrm{g}_t, \mathrm{P}^{(*)} - \mathbb{E}_{\mathcal{S},\mathcal{R}}[\mathrm{P}^{(t)}]\rangle$$
$$\frac{\log(O(d+\log(1/\delta))) - \log\left(\frac{2}{3\delta^2}\right)}{2\eta} + T\eta\alpha_2$$

Setting $\eta = \frac{\log(1+\delta/9)}{\sqrt{T}(\alpha+4\lambda^2)}$ finishes the proof.

□

D. Proofs of Section 5

Lemma D.1. The naive estimator $\tilde{\mathrm{x}}_i\tilde{\mathrm{x}}_i^\top$ is not unbiased, i.e.

$$\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}] = \frac{r(r-1)}{d(d-1)}\mathrm{x}\mathrm{x}^\top + \frac{r(d-r)}{d(d-1)}\operatorname{diag}\mathrm{x}\mathrm{x}^\top,$$

where expectation is taken with respect to the uniform sampling model.

Proof of Lemma D.1. For off-diagonal elements ($i \neq j$), $\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]_{i,j} = \mathrm{x}_i\mathrm{x}_j\mathbb{P}(i,j \in \{i_1,..,i_r\})$ so we need to compute the probability of the event that two fixed element $i,j \in [1\cdots d]$ are both in the subset $\{i_1,..,i_r\}$ sampled uniformly at random from all subsets of size r. The number of sets of r elements which contain i and j is exactly $\binom{d-2}{r-2}$ and the total number of r sets is $\binom{d}{r}$ thus $\mathbb{P}(i,j \in \{i_1,..,i_r\}) = \frac{\binom{d-2}{r-2}}{\binom{d}{r}} = \frac{r(r-1)}{d(d-1)}$ and so $\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]_{i,j} = \mathrm{x}_i\mathrm{x}_j\frac{r(r-1)}{d(d-1)}$. Now consider the diagonal elements. $\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]_{i,i} = \mathrm{x}_i^2\mathbb{P}(i \in \{i_1,..,i_r\})$ so we need to compute the probability of the event $i \in \{i_1,..,i_r\}$. The number of sets with size r containing i is $\binom{d-1}{r-1}$ and thus $\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}]_{i,i} = \mathrm{x}_i^2\frac{r}{d}$. Putting together,

$$\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}] = \frac{r(r-1)}{d(d-1)}\mathrm{x}\mathrm{x}^\top + (\frac{r}{d} - \frac{r(r-1)}{d(d-1)})\operatorname{diag}\mathrm{x}\mathrm{x}^\top$$
$$= \frac{r(r-1)}{d(d-1)}\mathrm{x}\mathrm{x}^\top + \frac{r(d-r)}{d(d-1)}\operatorname{diag}\mathrm{x}\mathrm{x}^\top.$$

□

Proof of Lemma 5.1. First note that

$$\mathbb{E}_{\mathcal{S}}[\mathrm{z}\mathrm{z}^\top|\tilde{\mathrm{x}}] = \frac{dr-r^2}{r-1}\mathbb{E}_{\mathcal{S}}[\tilde{x}_{i_s}^2\mathrm{e}_{i_s}\mathrm{e}_{i_s}^\top|\tilde{\mathrm{x}}] = \frac{d-r}{r-1}\operatorname{diag}\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top.$$

The proof simply follows from the following equalities.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},\mathcal{R}}[\widehat{\mathrm{x}}\widehat{\mathrm{x}}^\top - \mathrm{z}\mathrm{z}^\top|\mathrm{x}] \\
&= \mathbb{E}_{\mathcal{R}}[\mathbb{E}_{\mathcal{S}}[\widehat{\mathrm{x}}\widehat{\mathrm{x}}^\top|\tilde{\mathrm{x}}]|\mathrm{x}] - \mathbb{E}_{\mathcal{R}}[\mathbb{E}_{\mathcal{S}}[\mathrm{z}\mathrm{z}^\top|\tilde{\mathrm{x}}]|\mathrm{x}] \\
&= \frac{d(d-1)}{r(r-1)}\mathbb{E}_{\mathcal{R}}[\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top|\mathrm{x}] - \frac{d-r}{r-1}\mathbb{E}_{\mathcal{R}}[\operatorname{diag}\tilde{\mathrm{x}}\tilde{\mathrm{x}}^\top] \\
&= \mathrm{x}\mathrm{x}^\top + \frac{d-r}{r-1}\operatorname{diag}\mathrm{x}\mathrm{x}^\top - \frac{d-r}{r-1}\operatorname{diag}\mathrm{x}\mathrm{x}^\top \\
&= \mathrm{x}\mathrm{x}^\top.
\end{aligned} \tag{23}$$

where $\operatorname{diag}\mathrm{A}, \mathrm{A} \in \mathbb{R}^{d\times d}$ is the $d \times d$ matrix consisting of the diagonal of A. □

Proof of Theorem 5.2. We first bound $\|\widehat{\mathrm{g}}_t\|_F^2$ as follows:

$$\begin{aligned}
\|\widehat{\mathrm{g}}_t\|_F^2 &= \left\|\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top - \mathrm{z}_t\mathrm{z}_t^\top\right\|_F^2 \\
&= \left\|\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top\right\|_F^2 + \left\|\mathrm{z}_t\mathrm{z}_t^\top\right\|_F^2 - 2\langle\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top, \mathrm{z}_t\mathrm{z}_t^\top\rangle \\
&\leq \left\|\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top\right\|_F^2 + \left\|\mathrm{z}_t\mathrm{z}_t^\top\right\|_F^2 \\
&\leq \|\widehat{\mathrm{x}}_t\|^4 + \|\mathrm{z}_t\|^4 \\
&\leq \frac{d^2(d-1)^2}{r^2(r-1)^2} + \frac{r^2(d-r)^2}{(r-1)^2}\|\mathrm{x}\|_\infty^4 \\
&\leq \frac{d^2(d-1)^2 + r^4(d-r)^2}{r^2(r-1)^2}.
\end{aligned}$$

where the first inequality follows from $\langle\widehat{\mathrm{x}}_t\widehat{\mathrm{x}}_t^\top, \mathrm{z}_t\mathrm{z}_t^\top\rangle \geq 0$, since this is an inner product of PSD matrices, the third inequality follows the definition of $\widehat{\mathrm{x}}$ and z, and the forth inequality holds since $\|\tilde{\mathrm{x}}\| \leq \|\mathrm{x}\| \leq 1$.Now, using Lemma A.1 with $\{\widehat{\mathrm{g}}_t\}_{t=1}^T$, noting $\|\mathrm{P}\|_F^2 \leq k$ for all $\mathrm{P} \in C$, we have that

$$\sum_{t=1}^{T}\left\langle \mathrm{P}^* - \mathrm{P}^{(t)}, -\widehat{\mathrm{g}}_t\right\rangle$$
$$\leq \frac{k}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\frac{d^2(d-1)^2 + r^4(d-r)^2}{r^2(r-1)^2}$$
$$= \frac{k}{2\eta} + \frac{\eta}{2}T\frac{d^2(d-1)^2 + r^4(d-r)^2}{r^2(r-1)^2}.$$

Since $\mathbb{E}_{\widehat{\mathrm{g}}_t}[=]\mathrm{g}_t = \mathrm{x}_t\mathrm{x}_t^\top$, setting $\eta = \frac{r(r-1)}{\sqrt{d^2(d-1)^2+r^4(d-r)^2}}\sqrt{\frac{k}{2T}}$ and taking expectation with respect to the internal randomization $\mathcal{S}$ and the distribution of the missing data $\mathcal{R}$ finishes the proof. □

Proof of Theorem 5.3. First we bound $\|\widehat{\mathrm{g}}_k\|_2$. Since $\widehat{\mathrm{g}}_k = \widehat{\mathrm{x}}_k\widehat{\mathrm{x}}_k^\top - \mathrm{z}_k\mathrm{z}_k^\top$ is the difference of two PSD matrices it holds that

$$\|\widehat{\mathrm{g}}_k\| = \max\left(\left\|\widehat{\mathrm{x}}_k\widehat{\mathrm{x}}_k^\top\right\|, \left\|\mathrm{z}_k\mathrm{z}_k^\top\right\|\right) \leq \alpha. \tag{24}$$

Let v be the top eigenvector of C with associated eigenvalue λ. Also let $\mathrm{g}_k = \mathrm{x}_k\mathrm{x}_k^\top$ and w_k be the k-th iterate of Oja's algorithm, such that $\mathrm{P}^{(k)} = \mathrm{w}_k\mathrm{w}_k^\top$. Define

$$\Phi_k^{\mathrm{M}} = (\mathrm{I}+\eta\widehat{\mathrm{g}}_k)\cdots(\mathrm{I}+\eta\widehat{\mathrm{g}}_1)\mathrm{M}(\mathrm{I}+\eta\widehat{\mathrm{g}}_1)\cdots(\mathrm{I}+\eta\widehat{\mathrm{g}}_k)$$
$$\Psi_k = (\mathrm{I}+\eta\widehat{\mathrm{g}}_k)\cdots(\mathrm{I}+\eta\widehat{\mathrm{g}}_T)\mathrm{v}\mathrm{v}^\top(\mathrm{I}+\eta\widehat{\mathrm{g}}_T)\cdots(\mathrm{I}+\eta\widehat{\mathrm{g}}_k).$$

Following the calculations in (Allen-Zhu & Li, 2017) Appendix I, we have:

$$\begin{aligned}
&\mathrm{Tr}\left(\Phi_T^{\mathrm{u}\mathrm{u}^\top}\right) = \mathrm{Tr}\left((\mathrm{I}+\eta\widehat{\mathrm{g}}_T)\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}(\mathrm{I}+\eta\widehat{\mathrm{g}}_T)\right)\\
&= \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right) + 2\eta\mathrm{Tr}\left(\widehat{\mathrm{g}}_T\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right) + \eta^2\mathrm{Tr}\left(\widehat{\mathrm{g}}_T\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\widehat{\mathrm{g}}_T\right)\\
&= \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right) + 2\eta\mathrm{Tr}\left(\widehat{\mathrm{g}}_T\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right) + \eta^2\mathrm{Tr}\left(\widehat{\mathrm{g}}_T^2\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right)\\
&\le \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right) + 2\eta\mathrm{Tr}\left(\widehat{\mathrm{g}}_T\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right) + \eta^2\alpha^2\mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right)\\
&= \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right)(1+\eta^2\alpha^2+2\eta\mathrm{w}_T^\top\widehat{\mathrm{g}}_T\mathrm{w}_T)\\
&\le \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right)\exp\left(\eta^2\alpha^2+2\eta\mathrm{w}_T^\top\widehat{\mathrm{g}}_T\mathrm{w}_T\right)\\
&\cdots \le \|\mathrm{u}\|_2^2\exp\left(T\eta^2\alpha^2+2\eta\sum_{k=1}^T\mathrm{w}_k^\top\widehat{\mathrm{g}}_k\mathrm{w}_k\right),
\end{aligned} \tag{25}$$

where for the first inequality we have used the fact that $\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}$ is PSD and $\|\widehat{\mathrm{g}}_k\|_2 \le \alpha$ and for the second inequality we have used $1+x \le \exp(x)$. We have also used the fact that $\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top} = \mathrm{Tr}\left(\Phi_{T-1}^{\mathrm{u}\mathrm{u}^\top}\right)\mathrm{w}_T\mathrm{w}_T^\top$. Next we have

$$\begin{aligned}
&\mathbb{E}[\Psi_1] = \mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top(\mathrm{I}+\eta\widehat{\mathrm{g}}_T)\Phi_{T-1}^{\mathrm{I}}(\mathrm{I}+\eta\widehat{\mathrm{g}}_T)\right)]\\
&= \mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top(\mathrm{I}+\eta\widehat{\mathrm{g}}_T)\Phi_{T-1}^{\mathrm{I}}\right) + \eta^2\mathrm{v}^\top\widehat{\mathrm{g}}_T\Phi_{T-1}^{\mathrm{I}}\widehat{\mathrm{g}}_T\mathrm{v}]\\
&\ge \mathbb{E}[\mathrm{Tr}\left(\mathrm{v}\mathrm{v}^\top(\mathrm{I}+\eta\mathrm{C})\Phi_{T-1}^{\mathrm{I}}\right)]\\
&= (1+2\eta\lambda)\mathbb{E}[\mathrm{v}^\top\Phi_{T-1}^{\mathrm{I}}\mathrm{v}]\\
&\ge \exp\left(2\eta\lambda-2\eta^2\lambda^2\right)\mathbb{E}[\mathrm{v}^\top\Phi_{T-1}^{\mathrm{I}}\mathrm{v}] \ge \cdots\\
&\ge \exp\left((2\eta\lambda-2\eta^2\lambda^2)T\right),
\end{aligned} \tag{26}$$

where the expectation is both with respect to $\mathcal{R}$ and $\mathcal{D}$ and we have used $1+2x \ge \exp\left(2x-2x^2\right)$, for $0 \le x \le 1$. Finally we have:

$$\begin{aligned}
&\mathbb{E}[\mathrm{Tr}\left(\Psi_1^2\right)] = \mathbb{E}[\mathrm{Tr}\left((I+\eta\widehat{\mathrm{g}}_1)^2\Psi_2(I+\eta\widehat{\mathrm{g}}_1)^2\Psi_2\right)]\\
&\le \mathbb{E}[\mathrm{Tr}\left((\mathrm{I}+\eta\widehat{\mathrm{g}}_1)^4\Psi_2^2\right)]\\
&= \mathbb{E}[\mathrm{Tr}\left((\mathrm{I}+4\eta^3\widehat{\mathrm{g}}_1^3+6\eta^2\widehat{\mathrm{g}}_1^2+4\eta\widehat{\mathrm{g}}_1+\eta^4\widehat{\mathrm{g}}_1^4)\Psi_2^2\right)]\\
&\le \mathbb{E}[\mathrm{Tr}\left((\mathrm{I}+11\eta^2\alpha^4\mathrm{I}+4\eta\mathrm{g}_1)\Psi_2^2\right)]\\
&\le \exp\left(11\eta^2\alpha^4+4\eta\lambda\right)\mathbb{E}[\mathrm{Tr}\left(\Psi_2^2\right)] \le \cdots\\
&\le \exp\left(T(11\eta^2\alpha^4+4\eta\lambda)\right).
\end{aligned} \tag{27}$$

Combining equations 26 and 27 with Chebyshev's inequality we get:

$$\begin{aligned}
&\mathbb{P}[\mathrm{Tr}(\Psi_1) \le \exp\left(T(2\eta\lambda-2\alpha^2\eta^2)\right) -\\
&\exp\left(T(2\eta\lambda-2\alpha^2\eta^2)\right)\sqrt{\frac{(\exp\left(T(11\eta^2\alpha^4+4\eta^2\lambda^2)\right)-1)}{\delta}}]\\
&\le \delta.
\end{aligned}$$

Thus with probability $1-\delta$ it holds

$$\begin{aligned}
\mathrm{Tr}(\Psi_1) &\ge \exp\left(T(2\eta\lambda-2\alpha^2\eta^2)\right)(1-\\
&\delta^{-1/2}\sqrt{\exp\left(T(11\eta^2\alpha^4+4\eta^2\lambda^2)\right)-1}\Big)\\
&\ge \frac{2}{3}\exp\left(T(2\eta\lambda-2\alpha^2\eta^2)\right)
\end{aligned}$$

as long as $\eta \le \frac{\log(1+\delta/9)}{(11\alpha^2+4\lambda^2)\sqrt{T}}$. Following the derivations (Allen-Zhu & Li, 2017), we have that with probability $1-\delta$ it holds that $\mathrm{Tr}\left(\Phi_T^{\mathrm{u}\mathrm{u}^\top}\right) \ge \Omega(\delta^{-2})\mathrm{Tr}(\Psi_1)$ and that $\|u\|_2^2 \le O(d+\log(1/\delta^2))$. Union bound, together with inequality 25 implies that with probability $1-2\delta$ we have:

$$\begin{aligned}
&O(d+\log(1/\delta^2))\exp\left(T\eta^2\alpha^2+2\eta\sum_{k=1}^T\mathrm{w}_k^\top\widehat{\mathrm{g}}_k\mathrm{w}_k\right)\\
&\ge \frac{2}{3\delta^2}\exp\left(T(2\eta\lambda-2\alpha^2\eta^2)\right) \implies\\
&\log(O(d+\log(1/\delta^2))) + 2\eta\sum_{k=1}^T\mathrm{w}_k^\top\widehat{\mathrm{g}}_k\mathrm{w}_k\\
&\ge \log(2/3\delta^2) + 2T\eta\lambda - 2T\alpha^2\eta^2 \implies\\
&\sum_{k=1}^T\lambda - \mathrm{w}_k^\top\widehat{\mathrm{g}}_k\mathrm{w}_k \le\\
&\frac{\log(O(d+\log(1/\delta^2))) - \log(2/3\delta^2)/2}{\eta} + 2T\alpha^2\eta \implies\\
&\sum_{k=1}^T\langle\mathrm{g}_k, \mathrm{P}^* - \mathbb{E}_{\mathcal{S},\mathcal{R}}[\mathrm{P}_k]\rangle \le\\
&\frac{\sqrt{T}(11\alpha^2+4\lambda^2)}{\log(1+\delta/9)}\left(\log(O(d+\log(1/\delta^2)))-\log(2/3\delta^2)/2\right)\\
&+ 2\frac{\sqrt{T}}{\log(1+\delta/9)},
\end{aligned}$$

where in the last implication we have substituted $\eta = \frac{\log(1+\delta/9)}{(11\alpha^2+4\lambda^2)\sqrt{T}}$ and used the fact that $\alpha^2 < 11\alpha^2+4\lambda^2$. □

E. Proofs of Section 6

Proof of Theorem 6.1. Denoting $(\mathrm{P}^{(t)} - \mathrm{I})\mathrm{x}_t$ by y_t, we can rewrite g_t as $\mathrm{g}_t = -\frac{\mathrm{x}_t\mathrm{y}_t^\top + \mathrm{y}_t\mathrm{x}_t^\top}{2\|\mathrm{y}_t\|_2}$. Thus we have

$$\begin{aligned}\|g_t\|_F^2 &= \frac{1}{4\|y_t\|_2^2}\mathrm{Tr}\left((\mathrm{x}_t\mathrm{y}_t^\top + \mathrm{y}_t\mathrm{x}_t^\top)^\top(\mathrm{x}_t\mathrm{y}_t^\top + \mathrm{y}_t\mathrm{x}_t^\top)\right)\\ &\le \frac{4\|\mathrm{x}_t\|_2^2\|\mathrm{y}_t\|_2^2}{4\|\mathrm{y}_t\|_2^2} = \|\mathrm{x}_t\|_2^2 \le 1.\end{aligned}$$

By convexity of norms, $\|\mathrm{x} - \mathrm{P}^{(t)}\mathrm{x}\|_2 - \|\mathrm{x} - \mathrm{P}^*\mathrm{x}\|_2 \le \langle \mathrm{P}^{(t)} - \mathrm{P}^*, \mathrm{g}_t\rangle$. Using Lemma A.1, noting $\|\mathrm{P}\|_F^2 \le k$ for all $\mathrm{P} \in C$, and $\|\mathrm{g}_t\|_F^2 \le 1$,

$$\sum_{t=1}^T \|\mathrm{x}_t - \mathrm{P}^{(t)}\mathrm{x}_t\|_2 - \sum_{t=1}^T \|\mathrm{x}_t - \mathrm{P}^*\mathrm{x}_t\|_2 \le \frac{k}{2\eta} + \frac{\eta T}{2}.$$

Choosing $\eta = \sqrt{\frac{k}{T}}$ completes the proof. □

F. Proofs of Section 8

Proof of Theorem 8.1. Our proof follows (Zinkevich, 2003). We start the analysis by bounding the distance of our iterates at time t from any dynamic competitor $\mathrm{P}_*^{(t-1)}$ at time $t-1$

$$\begin{aligned}&\left\|\mathrm{P}^{(t+1)} - \mathrm{P}_*^{(t)}\right\|_F^2 = \left\|\Pi_C\left(\mathrm{P}^{(t)} - \eta\mathrm{g}_t\right) - \mathrm{P}_*^{(t)}\right\|_F^2\\ &\le \left\|\mathrm{P}^{(t)} - \eta\mathrm{g}_t - \mathrm{P}_*^{(t)}\right\|_F^2\\ &= \left\|\mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)} + \mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)} - \eta\mathrm{g}_t\right\|_F^2\\ &= \left\|\mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)}\right\|_F^2 + \left\|\mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)}\right\|_F^2\\ &+ 2\langle \mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)}, \mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)}\rangle + \eta^2\|\mathrm{g}_t\|_F^2\\ &- 2\eta\langle \mathrm{g}_t, \mathrm{P}^{(t)} - \mathrm{P}_*^{(t)}\rangle\\ &\le \left\|\mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)}\right\|_F^2 + \left\|\mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)}\right\|_F^2\\ &+ 2\left\|\mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)}\right\|_F \cdot \left\|\mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)}\right\|_F + \eta^2\\ &+ 2\eta\left(\mathrm{x}_t^\top\mathrm{P}^{(t)}\mathrm{x}_t - \mathrm{x}_t^\top\mathrm{P}_*^{(t)}\mathrm{x}_t\right)\\ &\le \left\|\mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)}\right\|_F^2 + 2\sqrt{k}\left\|\mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)}\right\|_F\\ &+ 4\sqrt{k}\left\|\mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)}\right\|_F + \eta^2\\ &+ 2\eta\left(\mathrm{x}_t^\top\mathrm{P}^{(t)}\mathrm{x}_t - \mathrm{x}_t^\top\mathrm{P}_*^{(t)}\mathrm{x}_t\right)\end{aligned}$$

Rearranging and dividing both sides by 2η we get

$$\begin{aligned}\mathrm{x}_t^\top\mathrm{P}_*^{(t)}\mathrm{x}_t - \mathrm{x}_t^\top\mathrm{P}^{(t)}\mathrm{x}_t &\le \frac{3\sqrt{k}}{\eta}\left\|\mathrm{P}_*^{(t-1)} - \mathrm{P}_*^{(t)}\right\|_F + \frac{\eta}{2}\\ &+ \frac{\left\|\mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)}\right\|_F^2 - \left\|\mathrm{P}^{(t+1)} - \mathrm{P}_*^{(t)}\right\|_F^2}{2\eta}\end{aligned}$$

Summing over t, observe the telescopic sum on the right hand side is bounded by

$$\begin{aligned}&\sum_{t=1}^T \left\|\mathrm{P}^{(t)} - \mathrm{P}_*^{(t-1)}\right\|_F^2 - \left\|\mathrm{P}^{(t+1)} - \mathrm{P}_*^{(t)}\right\|_F^2\\ &= \left\|\mathrm{P}^{(1)} - \mathrm{P}_*^{(0)}\right\|_F^2 - \left\|\mathrm{P}^{(T+1)} - \mathrm{P}_*^{(T)}\right\|_F^2\\ &\le \left\|-\mathrm{P}_*^{(0)}\right\|_F^2 \le k\end{aligned}$$

Plugging the definition of the total shift S into the sum

$$\begin{aligned}\sum_{t=1}^T \mathrm{x}_t^\top\mathrm{P}_*^{(t)}\mathrm{x}_t - \sum_{t=1}^T \mathrm{x}_t^\top\mathrm{P}^{(t)}\mathrm{x}_t &\le \frac{3\sqrt{k}S}{\eta} + \frac{\eta T}{2} + \frac{k}{2\eta}\\ &= \frac{6\sqrt{k}S + k}{2\eta} + \frac{\eta T}{2}\end{aligned}$$

Choosing $\eta = \sqrt{\frac{6\sqrt{k}S+k}{T}}$ completes the proof. □

F.1. Experimental Results

We provide additional experiments in Figures 3, 4, 5 and 6 for PCA with partial observations and for PCA with missing entries. As stated in the main text, all our observations from Figure 1 and Figure 2 hold for the results presented here.

Further, in Figure 7, we present experimental results for the proposed *Absolute Subspace Deviation Model* algorithm of Section 6(referred to as *Robust GD* in the plots). The algorithms against which we compare are *Robust Online PCA* proposed in (Goes et al., 2014) (referred to as *Robust MEG* in the plots) which is an instance of online mirror descent with choice of potential function being entropy and simple batch PCA on the whole dataset. As ground truth we take the k-dimensional subspace returned by the proposed method in (Lerman et al., 2012). In the following discussion the ground truth is represented by the projection matrix $\mathrm{P}^* \in \mathbb{R}^{d\times d}$ and the estimated projection matrix returned by an algorithm is denoted by $\mathrm{P}^{(t)}$. Since our experiments are ran on a synthetic dataset of fixed size, we denote this dataset by $\mathrm{X} \in \mathbb{R}^{d\times n}$. The criteria against which we evaluate are average angle between subspaces given by $\mathrm{Tr}\left((\mathrm{P}^{(t)})^\top\mathrm{P}^*\right)/k$, reconstruction error on the whole dataset given by $\left\|(\mathrm{U}^{(t)})^\top\mathrm{X} - (\mathrm{U}^{(*)})^\top\mathrm{X}\right\|_F/n$ where a rank-k projection matrix P is decomposed as $\mathrm{P} = \mathrm{U}\mathrm{U}^\top$ for orthogonal $\mathrm{U} \in \mathbb{R}^{d\times k}$. We also evaluate on the total regret incurred. The data set is generated as in (Goes et al., 2014). First inliers are sampled from a k-dimensional Multivariate Normal distribution with 0 mean and covariance matrix $\frac{1}{k}\mathrm{I}_k$. Next the inliers are embedded in a d-dimensional space via the linear transformation $\mathrm{U} \in \mathbb{R}^{d\times k}$ with entries $U_{ii} = 1$ and $U_{i\neq j} = 0$. Finally outliers are

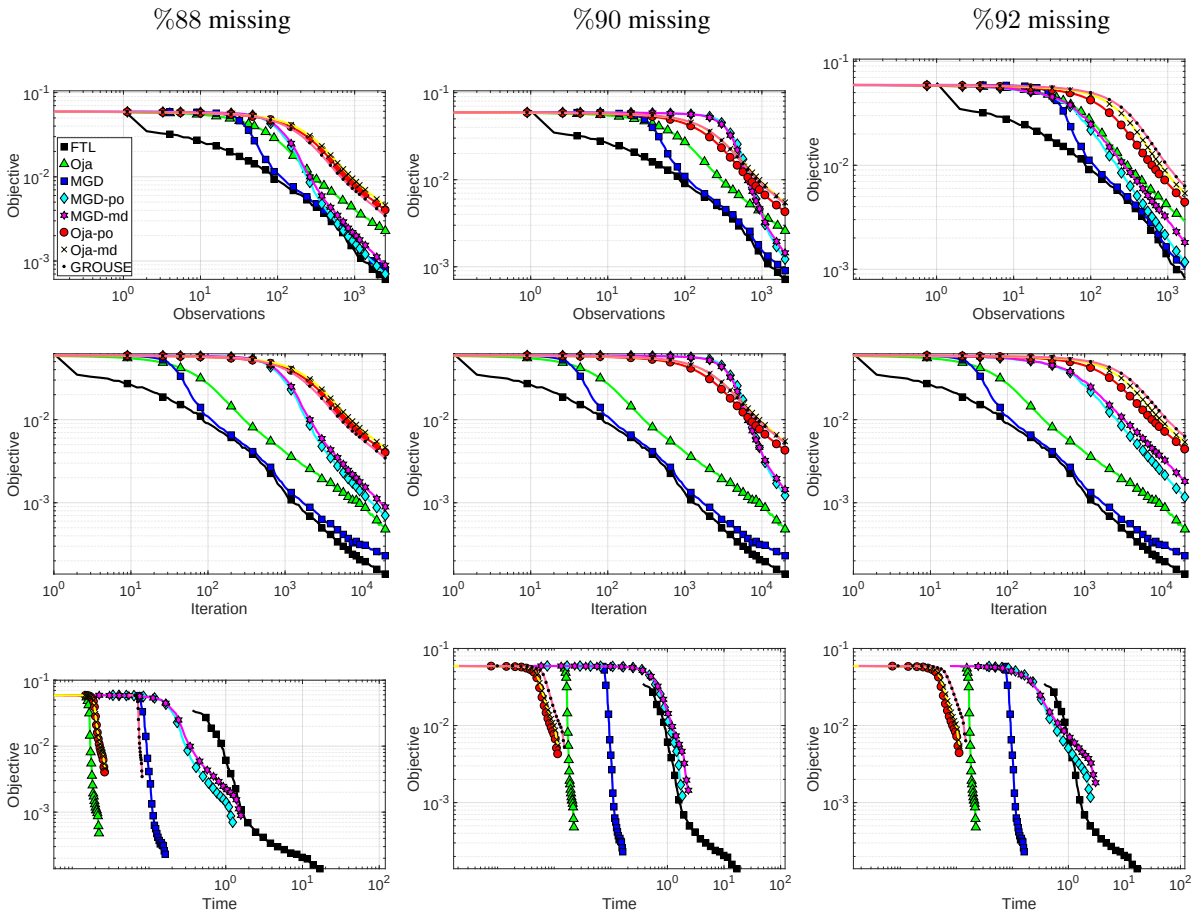

Figure 3: Comparisons of Oja, MGD, MGD-PO, MGD-MD, Oja-PO, Oja-MD, GROUSE for PCA with missing data on **XRMB** dataset, in terms of the variance captured on a test set as a function of number of observed entries for **k=2** (top), number of iterations (middle) and runtime (bottom).

sampled from a d-dimensional Multivariate Normal distribution with diagonal covariance $\frac{\lambda}{d \times \text{mean}(\Sigma_{\text{out}})}\Sigma_{\text{out}}$ where $(\Sigma_{\text{out}})_{ii} = i$, here λ is a user-specified parameter which governs the SNR. In our experiments $\lambda = 20$ the number of outliers is $40\%, 60\%, 80\%$ of the total number of points, $k = 2$ and $d = 100$. In our comparison we also include a "capped" version of our proposed method where the rank of each intermediate iterate is hard-capped to 15 by keeping the 15 directions associated with the top 15 singular values of our current iterate.

We now briefly discuss the more interesting points which should be observed from the provided Figure 7. First our algorithm clearly out-performs the *Robust Online PCA* and batch PCA in all the chosen criteria. Secondly, since *Robust Online PCA* is required to always keep full-rank iterates as it operates in the complementary space to the one we wish to recover it has to pay computational cost $O(d^3)$ per iteration while our method only pays $O(d\tilde{k}^2)$ where $\tilde{k}$ is the rank of the current iterate. As can be seen from the plots for the cases when number of outliers are 40% or 60% the intermediate rank of iterates does not grow too much so in practice our proposed method is efficient. Finally we note that the capped version of the proposed method performs as well or even better in some cases.

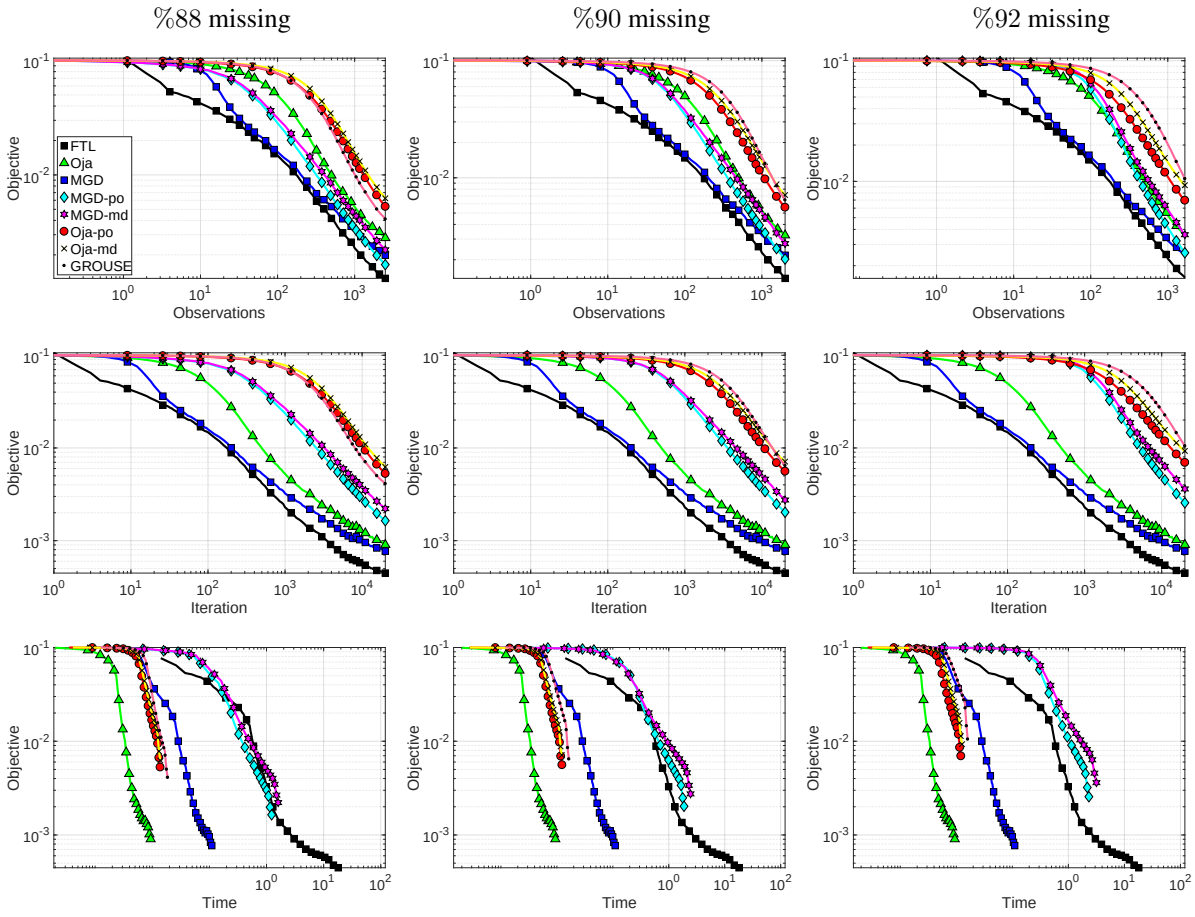

Figure 4: Comparisons of Oja, MGD, MGD-PO, MGD-MD, Oja-PO, Oja-MD, GROUSE for PCA with missing data on **XRMB** dataset, in terms of the variance captured on a test set as a function of number of observed entries for **k=4** (top), number of iterations (middle) and runtime (bottom).

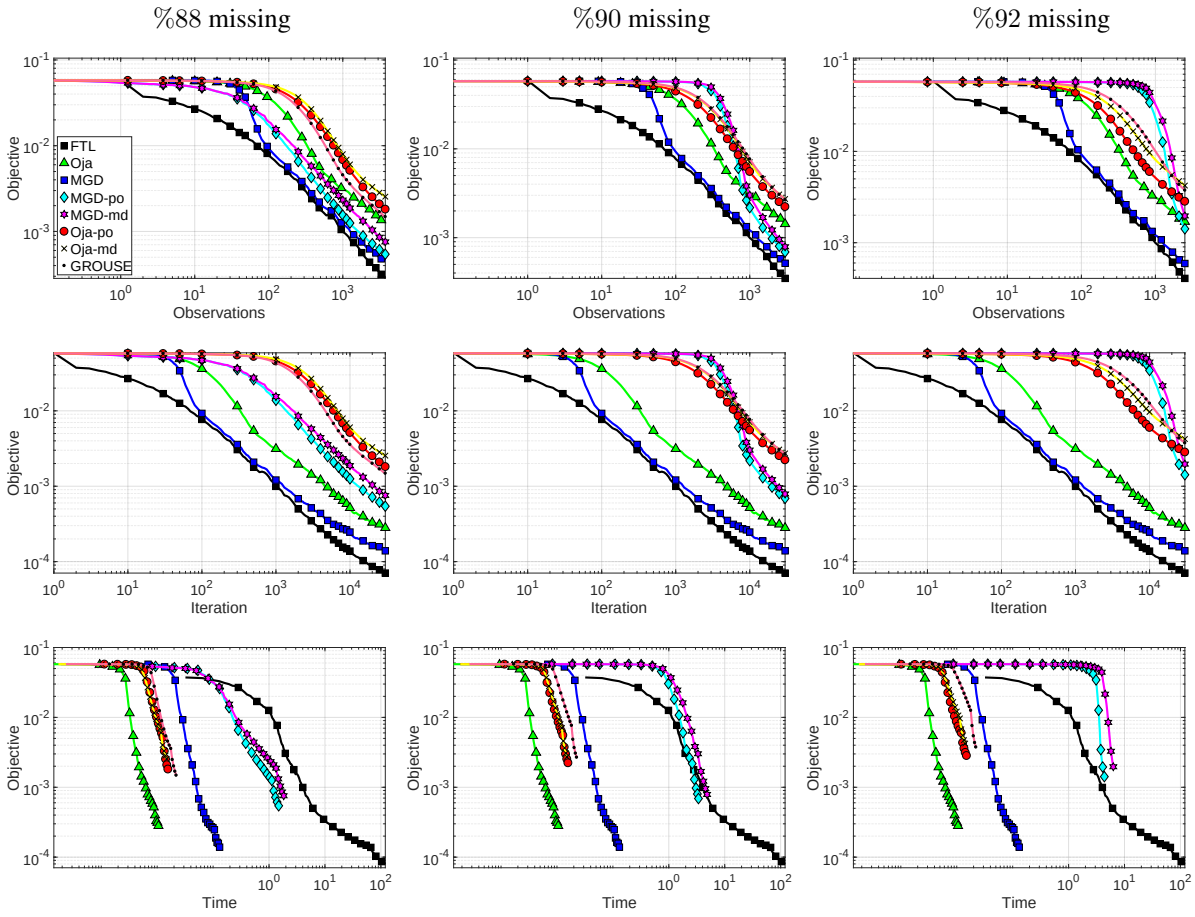

Figure 5: Comparisons of Oja, MGD, MSG-PO, MSG-MD, Oja-PO, Oja-MD and GROUSE for PCA with missing data on **MNIST** dataset, in terms of the variance captured on a test set as a function of number of observed entries for **k=2** (top), number of iterations (middle) and runtime (bottom).

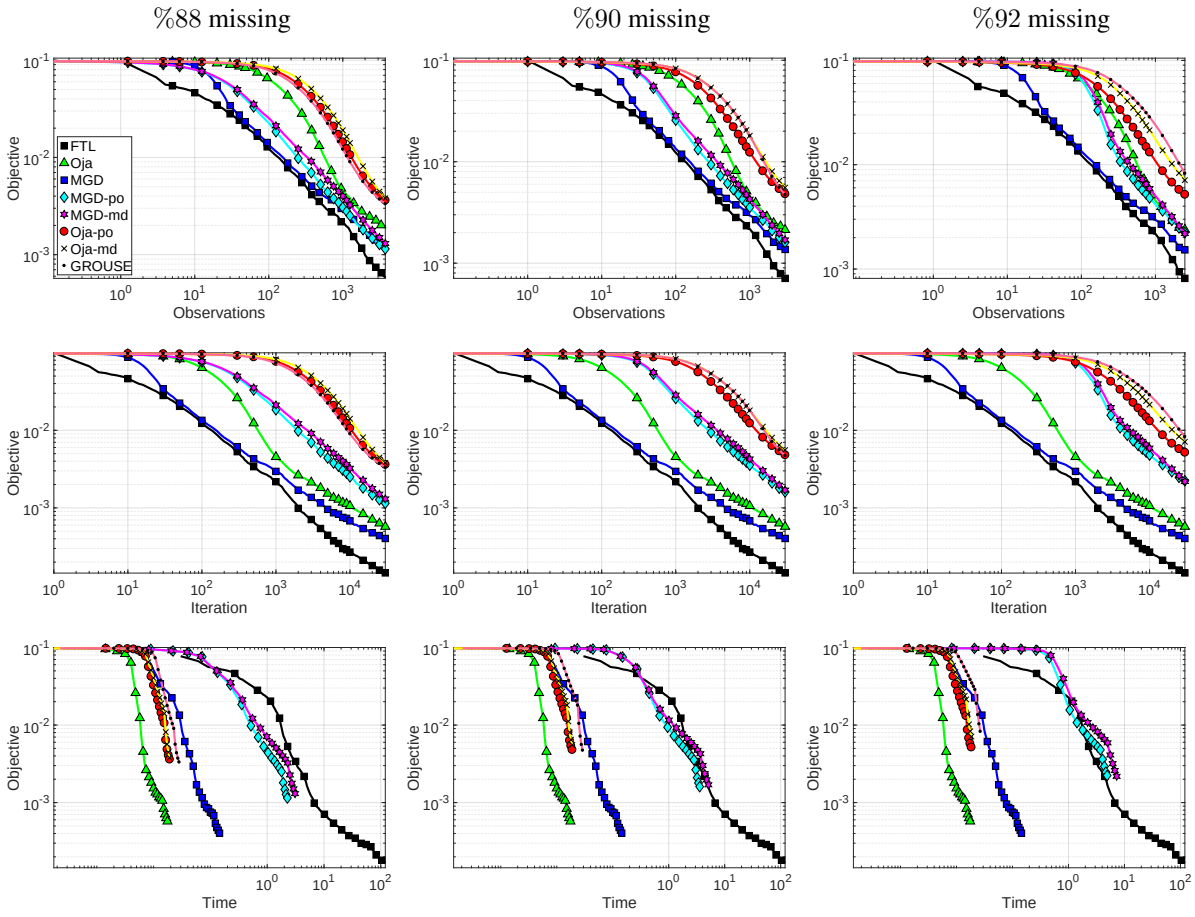

Figure 6: Comparisons of Oja, MGD, MSG-PO, MSG-MD, Oja-PO, Oja-MD and GROUSE for PCA with missing data on **MNIST** dataset, in terms of the variance captured on a test set as a function of number of observed entries for **k=4** (top), number of iterations (middle) and runtime (bottom).

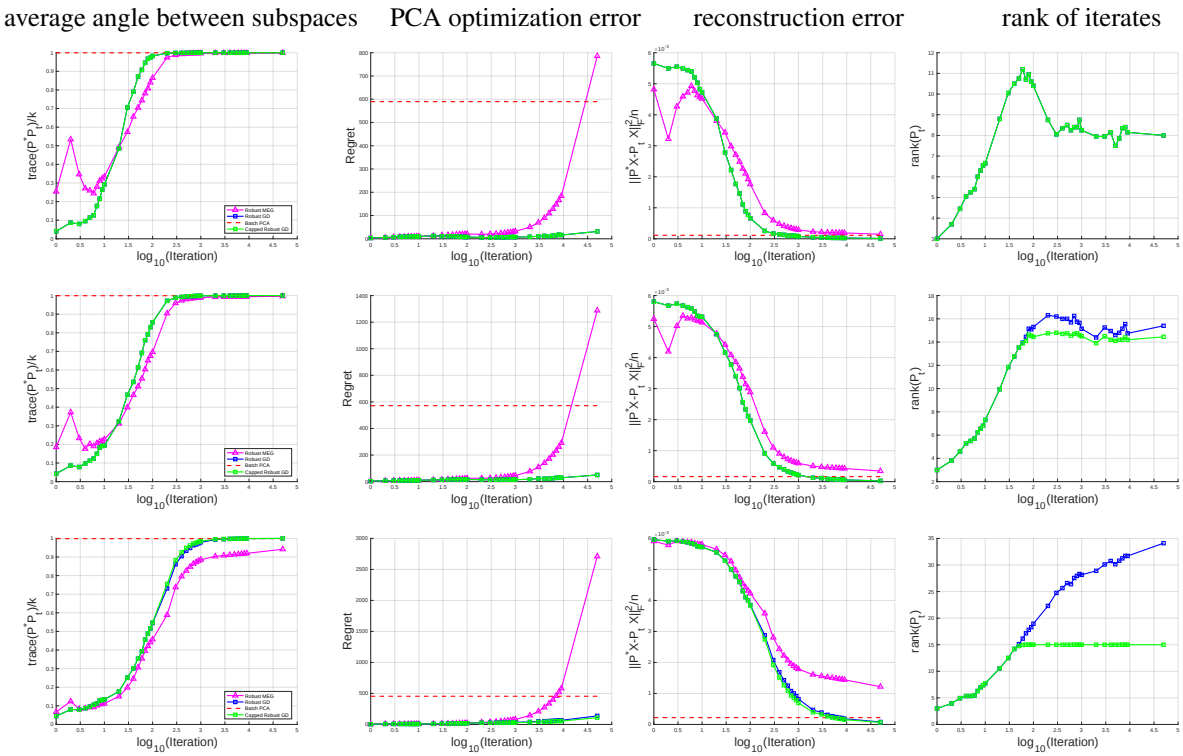

Figure 7: ($k = 2$) Comparison of batch PCA, Robust Online PCA (Robust MEG), and Absolute Subspace Deviation Model (Robust GD) with 40% outliers (top row), 60% outliers (middle row) and 80% outliers (bottom row) on synthetic dataset. Experiments are in terms of the average subspace angle (left most), variance captured on a test set (left), reconstruction error (right) and the rank of iterates (right most) as a function of number of samples.