Project: Mathematical Analysis

Ameya Anjarlekar

February 2020

1 Time Complexity of Gradient

The time complexity of computing $\bigtriangledown_D \hat{\Psi}^{(t)}(D^{(t)})$ is calculated as follows.

$$\bigtriangledown_D \hat{\Psi}^{(t)}(D^{(t)}) = -\frac{1}{N}\sum_{1}^{N} \Phi_j^T (y_j - \Phi_j D^{(t)} \alpha_j^{(t)}) \alpha_j^{(t)^T} \tag{1}$$

Φ_j^T is nxm, $(y_j - \Phi_j D^{(t)} \alpha_j^{(t)})$ is mx1 and $\alpha_j^{(t)^T}$ is 1xp with only k elements non-zero, thus we will have summation over m elements for calculating each of the nk elements of the nxp resulting matrix. This is finally summed N times. Thus, time complexity of computing $\bigtriangledown_D \hat{\Psi}^{(t)}(D^{(t)})$ is O(Nnmk).

2 Part 1: Bounding within an envelope

2.1 Proof of theorem 1

Let $y_j = \Phi_j x_j$. Then we have,

$$
\begin{aligned}
y_j(i) =& \sum_{1}^{n} [\Phi_j]_{i,n} x_j(n) \\
E(y_j^2(i)) =& E((\sum_{1}^{n} [\Phi_j]_{i,n} x_j(n))^2) \\
E(y_j^2(i)) =& E(\sum_{1}^{n} ([\Phi_j]_{i,n} x_j(n))^2) \quad ([\Phi_j]_{i,n} \text{ are zero mean iid}) \\
E(y_j^2(i)) =& \frac{1}{m_j} \sum_{1}^{n} (x_j(n))^2 \quad ([\Phi_j]_{i,n} \sim N(0, \frac{1}{m_j}) \\
E(y_j^2(i)) =& \frac{1}{m_j} ||x_j||_2^2 \\
E(||y_j||_2^2) =& \sum_{1}^{m_j} \frac{1}{m_j} ||x_j||_2^2 \\
E(||y_j||_2^2) =& ||x_j||_2^2
\end{aligned} \tag{2}
$$

Thus,

$$
\begin{aligned}
|\sum_{1}^{N} (||y_j||_2^2 - ||x_j||_2^2| \geq \epsilon \sum_{1}^{N} ||x_j||_2^2 \\
|\sum_{1}^{N} (||y_j||_2^2 - E(||y_j||_2^2)| \geq \epsilon \sum_{1}^{N} E(||y_j||_2^2)
\end{aligned} \tag{3}
$$

Thm 1A:
$X_1, X_2, ...X_L$ are independent sub-exponential rv and K $= max_i ||X_i||_{\Psi_1}$. Then for every $t \geq 0$, $a = (a_1, a_2....a_L) \in R^L$,

$$
P(|\sum_{1}^{L} a_i X_i| \geq t) \leq 2exp\Big[- c \cdot min\big(\frac{t^2}{K^2 ||a||_2^2}, \frac{t}{K||a||_\infty}\big)\Big] \tag{4}
$$

Proof:

$$\begin{aligned}
P(\sum_1^L a_i X_i \geq t) =& P(exp(\lambda \sum_1^L a_i X_i \geq exp(\lambda t)) \quad (\lambda \geq 0) \\
\leq& exp(-\lambda t) E(exp(\lambda \sum_1^L a_i(X_i)) \quad \text{(Markov's inequality)} \\
=& exp(-\lambda t) \prod_1^L E(exp(\lambda a_i(X_i))
\end{aligned}$$

Taking,

$$\begin{aligned}
\lambda \leq& min_i \frac{c}{||a_i X_i||_{\Psi_1}} \\
=& \frac{c}{max_i ||a_i X_i||_{\Psi_1}} \\
\leq& \frac{c}{||a||_\infty K}
\end{aligned} \tag{5}$$

for some constant c. Then for an absolute constant C,

$$\begin{aligned}
P(\sum_1^L a_i X_i \geq t) \leq& exp(-\lambda t + \sum_1^L C\lambda^2 ||a_i X_i||^2_{\Psi_1}) \quad \text{(Sub-exp rv)} \\
\leq& exp(-\lambda t + \sum_1^L C\lambda^2 K^2 ||a_i||^2) \quad \text{(defn of K)} \\
=& exp(-\lambda t + C\lambda^2 K^2 ||a||_2^2)
\end{aligned} \tag{6}$$

RHS will attain minima at $\lambda = \frac{t}{2CK^2||a||_2^2}$. Substituting it in RHS, we get,

$$P(\sum_1^L a_i X_i \geq t) \leq exp\big(\frac{-t^2}{4CK^2||a||_2^2}\big) \tag{7}$$

However, minima is attained only if $\frac{t}{2CK^2||a||_2^2} \leq \frac{c}{||a||_\infty K}$. Else minima is at $\lambda = \frac{c}{||a||_\infty K}$. Substituting it, we get,

$$P(\sum_1^L a_i X_i \geq t) \leq exp\big(\frac{-ct}{K||a||_\infty} + \frac{Cc^2||a||_2^2}{||a||_\infty^2}\big) \tag{8}$$

$$\begin{aligned}
\frac{t}{2CK^2||a||_2^2} \geq& \frac{c}{||a||_\infty K} \\
\frac{t||a||_\infty}{2K||a||_2^2} \geq& Cc
\end{aligned} \tag{9}$$

$$P(\sum_1^L a_i X_i \geq t) \leq exp\big(\frac{-ct}{2K||a||_\infty}\big) \quad \text{(Using Eqs.(8,9))} \tag{10}$$

Thus, taking $\lambda = min(\frac{c}{||a||_\infty K}, \frac{t}{2CK^2||a||_2^2})$

$$P(\sum_1^L a_i X_i \geq t) \leq exp\Big[- c \cdot min\big(\frac{t^2}{K^2||a||_2^2}, \frac{t}{K||a||_\infty}\big)\Big] \tag{11}$$

Similar proof follows for the less than -t case. Thus, we get

$$P(|\sum_1^L a_i X_i| \geq t) \leq 2exp\Big[- c \cdot min\big(\frac{t^2}{K^2||a||_2^2}, \frac{t}{K||a||_\infty}\big)\Big] \tag{12}$$

Since, $X_i s$ are sub-gaussian

$$\begin{aligned}
(E(|X_i|^p))^{\frac{1}{p}} \leq & \sqrt{p}||X_i||_{\psi_2} \\
(E(|X_i|^p))^{\frac{1}{p}} \leq & (E(1)E(|X_i|^{2p}))^{\frac{1}{2p}} \\
(E(|X_i|^p))^{\frac{2}{p}} \leq & (E(|X_i|^{2p}))^{\frac{1}{p}} \\
sup_p(E(|X_i|^p))^{\frac{2}{p}} \leq & sup_p(E(|X_i|^{2p}))^{\frac{1}{p}} \\
||X_i||_{\Psi_2}^2 \leq & ||X_i||_{\Psi_1}^2
\end{aligned} \tag{13}$$

$$\begin{aligned}
(E(|X_i^2|^p))^{\frac{1}{p}} = & (E(|X_i|^{2p}))^{\frac{1}{p}} \\
\leq & (\sqrt{2p}||X_i||_{\Psi_2})^2 \\
= & 2p||X_i||_{\Psi_2}^2
\end{aligned} \tag{14}$$

Thus X_i^2 is also sub exp with $||X^2||_{\Psi_1} \leq 2||X^2||_{\Psi_2}$. Therefore, we get from Eqns. (13,14)

$$||X||_{\Psi_2}^2 \leq ||X^2||_{\Psi_1} \leq 2||X||_{\Psi_2}^2 \tag{15}$$

Let $Y_i = X_i^2 - E(X_i^2)$,

$$\begin{aligned}
||Y_i||_{\Psi_1} = & sup_p p^{-1}(E|Y_i|^p)^{\frac{1}{p}} \\
= & sup_p p^{-1}(E(|X_i^2 - E(X_i^2)|^p))^{\frac{1}{p}} \\
\leq & sup_p p^{-1}(E(|X_i^2 - E(X_i^2)|^p))^{\frac{1}{p}} \qquad (16) \\
E(|X-Y|^p) \leq & E(|X|^p) + E(|Y|^p) \quad \text{(Triangle inequality)} \\
(E(|X-Y|^p))^{\frac{1}{p}} \leq & (E(|X|^p) + E(|Y|^p))^{\frac{1}{p}} \\
\leq & (E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}} \quad (p \geq 1) \qquad (17) \\
||Y_i||_{\Psi_1} \leq & p^{-1}((E(|X_i^2|^p))^{\frac{1}{p}} + (E(|E(X_i^2)|^p))^{\frac{1}{p}}) \quad \text{(Eqns. (16,17))} \qquad (18) \\
= & ||X_i^2||_{\Psi_1} + ((|E(X_i^2)|^p))^{\frac{1}{p}}) \\
\leq & ||X_i^2||_{\Psi_1} + (E(|(X_i^2)|^p))^{\frac{1}{p}}) \\
= & 2||X_i^2||_{\Psi_1} \qquad (19)
\end{aligned}$$

Thus, from Eqns. (15,19),

$$||Y_i||_{\Psi_1} \leq 4||X_i^2||_{\Psi_2} \tag{20}$$

Now using thm 1A,

$$P(|\sum_1^L a_i(X_i^2 - E(X_i^2))| \geq t) \leq 2exp\Big[-c \cdot min\big(\frac{t^2}{16T^2||a||_2^2}, \frac{t}{4T||a||_\infty}\big)\Big] \tag{21}$$

where $T = max_i||X_i^2||_{\Psi_2}$ Define $\hat{y_j} = \frac{y_j}{||y_j||_{\Psi_2}}$, thus $||y_j||_{\Psi_2} = 1$.

$$P(|\sum_1^N(||y_j||_2^2 - E(||y_j||_2^2)| \geq \epsilon \sum_1^N E(||y_j||_2^2))$$

$$P(|\sum_{j=1}^N \sum_{i=1}^{m_j}(||y_j^{(i)}||_{\Psi_2}^2(||\hat{y_j^{(i)}}||_2^2 - E(||\hat{y_j^{(i)}}||_2^2))| \geq \epsilon \sum_1^N(||x_j||_2^2)) \tag{22}$$

Comparing with thm 1A and Eq(21), T = 1,
t = $\epsilon \sum_1^N(||x_j||_2^2))$,
$||a||_2^2 = \sum_{j=1}^N \sum_{i=1}^{m_j}(||y_j^{(i)}||_{\Psi_2}^4$,
$||a||_\infty = max_{i,j}(||y_j^{(i)}||_{\Psi_2}^2$,

$$\begin{aligned} E(exp(Y_j(i)t)) =& \prod_{l=1}^{l=n} E(exp(x_j(l)[\Phi_j]_{i,n}t)) \\ =& \prod_{l=1}^{l=n} exp(Cx_j^2(l)\frac{1}{m_j}||\phi||_{\Psi_2}t^2) \\ \leq& exp(C_1||x_j||_2^2\frac{1}{m_j}||\phi||_{\Psi_2}^2t^2) \end{aligned} \tag{23}$$

$$||y_j(i)||_{\Psi_2} \leq C_1\frac{||x_j||_2||\phi||_{\Psi_2}}{\sqrt{m_j}} \tag{24}$$

Thus,

$$||a||_2^2 \leq \frac{C_1^4}{m_j}||\phi||_{\Psi_2}^4 \sum_{j=1}^N(||x_j||^4) \tag{25}$$

$$||a||_\infty \leq \frac{C_1^2}{m_j}||\phi||_{\Psi_2}^2 max_j(||x_j||^2) \tag{26}$$

Define,

$$\Gamma_2 = \frac{(\sum_1^N(||x_j||_2^2))^2}{\frac{1}{m_j}\sum_{j=1}^N(||x_j||_2^4)} \tag{27}$$

$$\Gamma_\infty = \frac{\sum_1^N (||x_j||_2^2)}{\frac{1}{m_j} max_j (||x_j||_2^2)} \tag{28}$$

Substituting values in Eq(21),

$$P(|\sum_1^N (||y_j||_2^2 - E(||y_j||_2^2)| \geq \epsilon \sum_1^N (||x_j||_2^2)) \leq 2exp\Big[- c \cdot min\big(\frac{c_2^2 \epsilon^2 \Gamma_2}{||\phi||_{\Psi_2}^4}, \frac{c_2 \epsilon \Gamma_\infty}{||\phi||_{\Psi_2}^2}\big)\Big] \tag{29}$$

where $c_2 = \frac{1}{4C_1^2}$

2.2 Extending theorem 1

If $\phi \sim N(0,1)$, then $C_1 = 1$ and $||\phi||_{\Psi_2} = \sqrt{\frac{2}{\pi}}$ [41]. Thus, $c_2 = \frac{1}{4}$.

Oracle error h_t is defined for $t > 0$ as,

$$h_t = \frac{1}{nN} \sum_{j=1}^N ||x_j - D^{(t)} \alpha_j^{(t-1)}||_2^2 \tag{30}$$

Projected error (PE) is defined as,

$$g_t = \frac{1}{nN} \sum_{j=1}^N ||y_j - \Phi_j^{(2)} D^{(t)} \alpha_j^{(t-1)}||_2^2 \tag{31}$$

We replace x_j by $x_j - D^{(t)} \alpha_j^{(t-1)}$
We consider, m_js are same for all j and equal to m. Substituting and absorbing constants in c, we get,

$$P(\frac{|g_t - h_t|}{h_t} \geq \epsilon) \leq 2exp\Big[- c\epsilon min\big(\frac{\pi}{8} \epsilon \Gamma'_2, \Gamma'_\infty\big)\Big] \tag{32}$$

where

$$\Gamma'_2 = m \frac{(\sum_1^N (||x_j - D^{(t)} \alpha_j^{(t-1)}||_2^2))^2}{2 \sum_{j=1}^N (||x_j - D^{(t)} \alpha_j^{(t-1)}||_2^4)} \tag{33}$$

$$\Gamma'_\infty = m \frac{\sum_1^N (||x_j - D^{(t)} \alpha_j^{(t-1)}||_2^2)}{2 max_j (||x_j - D^{(t)} \alpha_j^{(t-1)}||_2^2)} \tag{34}$$

If error energy is spread equally among the blocks, then $||x_j - D^{(t)} \alpha_j^{(t-1)}||_2^2 = S$ for all j. Where S is a constant. Substituting, we get

$$\begin{aligned}\Gamma'_2 =& m\frac{(\sum_1^N(S^2))^2}{2\sum_{j=1}^N(S^4)}\\ =& \frac{m(NS^2)^2}{2NS^4}\\ =& \frac{mN}{2}\end{aligned} \tag{35}$$

$$\begin{aligned}\Gamma'_\infty =& m\frac{\sum_1^N(S^2)}{2max_j(S^2)}\\ =& \frac{m(NS^2)}{2S^2}\\ =& \frac{mN}{2}\end{aligned} \tag{36}$$

However, this is unlikely. We can still simplify in the following way,

$$\begin{aligned}\Gamma'_2 =& m\frac{(\sum_1^N(||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2))^2}{2\sum_{j=1}^N(||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^4)}\\ \geq& m\frac{(\sum_1^N(min_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2))^2}{2\sum_{j=1}^N(max_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^4)}\\ \geq& m\frac{(\sum_1^N(min_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2))^2}{2\sum_{j=1}^N(max_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^4)}\\ \geq& \frac{mN}{2}\frac{min_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^4}{max_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^4}\end{aligned} \tag{37}$$

$$\begin{aligned}\Gamma'_\infty =& m\frac{\sum_1^N(||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2)}{2max_j(||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2)}\\ \geq& m\frac{\sum_1^N(min_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2)}{2max_j(||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2)}\\ \geq& \frac{mN}{2}\frac{min_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2}{max_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2}\end{aligned} \tag{38}$$

Define,

$$\eta_t = \frac{min_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2}{max_j||x_j - D^{(t)}\alpha_j^{(t-1)}||_2^2} \tag{39}$$

Thus, $\Gamma'_{\infty} = \frac{mN\eta_t}{2}$ and $\Gamma'_2 = \frac{mN\eta_t^2}{2}$. Hence $\Gamma'_{\infty} = \Omega(Nm)$ and $\Gamma'_2 = \Omega(Nm)$. Hence, with a fixed probability of having

$$(1 - \epsilon_N)h_t \leq g_t \leq (1 + \epsilon_N)h_t \tag{40}$$

if $\frac{\pi}{8}\epsilon\Gamma'_2 \leq \Gamma'_{\infty}$, then $\epsilon \leq 2.55\frac{\Gamma'_{\infty}}{\Gamma'_2}$. In such a case,

$$P(\frac{|g_t - h_t|}{h_t} \geq \epsilon) \leq 2exp\bigg[- c\epsilon^2 \frac{\pi}{8}\Gamma'_2\bigg] \tag{41}$$

For maintaining same probability, when n,M increases, leading to Γ'_2 increasing, to maintain $P(\frac{|g_t - h_t|}{h_t} \geq \epsilon_N)$ below the bound, ϵ_N should decrease proportional to $\sqrt{\frac{1}{\Gamma'_2}}$. Thus ϵ_N is of the order $O(m^{\frac{-1}{2}}N^{\frac{-1}{2}})$.
if $\frac{\pi}{8}\epsilon\Gamma'_2 \geq \Gamma'_{\infty}$, then $\epsilon \geq 2.55\frac{\Gamma'_{\infty}}{\Gamma'_2}$. In such a case,

$$P(\frac{|g_t - h_t|}{h_t} \geq \epsilon) \leq 2exp\bigg[- c\epsilon\Gamma'_{\infty}\bigg] \tag{42}$$

For maintaining same probability, when n,M increases, leading to Γ'_{∞} increasing, to maintain $P(\frac{|g_t - h_t|}{h_t} \geq \epsilon_N)$ below the bound, ϵ_N should decrease proportional to $\frac{1}{\Gamma'_{\infty}}$. Thus ϵ_N is of the order $O(m^{-1}N^{-1})$.
Hence, g_t is bounded within an envelope around h_t which shrinks as N grows.

3 Part 2: Calculating Decay Rate

We define

$$h_t(D) = \frac{1}{nN}\sum_{j=1}^{N} ||x_j - D\alpha_j^{(t-1)}||_2^2 \tag{43}$$

$$g_t(D) = \frac{1}{nN}\sum_{j=1}^{N} ||y_j - \Phi_j^{(2)} D\alpha_j^{(t-1)}||_2^2 \tag{44}$$

The PE sequence, $g_0, g_1....g_T$ is decreasing due to $\gamma < 1$ and gradient descent : **I was not able to prove this part. I also feel that the mentioned statement is incorrect.**

$$\begin{aligned}\frac{g_t - g_T}{g_{t-1} - g_T} \leq & \frac{g_t}{g_{t-1}} \quad \text{(Cross-multiply and use } g_t \leq g_{t-1}) \\ = & \frac{g_t(D^{(t)})}{g_t(D^{(t-1)})} \cdot \frac{g_t(D^{(t-1)})}{g_{t-1}(D^{(t-1)})} \\ \leq & \frac{g_t(D^{(t)})}{g_t(D^{(t-1)})} \cdot \gamma \end{aligned} \tag{45}$$

Was not able to prove the last step and I do feel it is incorrect. If that is proved then PE sequence is decreasing can also be proved. Change in $g_t(D)$ depends on the condition number of the Hessian matrix which is essentially ratio of maximum to minimum change in $g_t(D)$.

$$\kappa_t = \frac{\lambda_{max}(H_t)}{\lambda_{min}(H_t)} \tag{46}$$

3.1 Calculating Hessian matrix

$$\begin{aligned} g_t(D) = & \frac{1}{nN}\sum_{j=1}^{N} ||y_j - \Phi_j^{(2)} D\alpha_j^{(t-1)}||_2^2 \\ = & \frac{1}{nN}\sum_{j=1}^{N} y_j^T y_j + \frac{1}{nN}\sum_{j=1}^{N} \alpha_j^{(t-1)^T} D^T \Phi_j^{(2)^T} \Phi_j^{(2)} D\alpha_j^{(t-1)} \\ & - \frac{2}{nN}\sum_{j=1}^{N} y_j^T \Phi_j^{(2)} D\alpha_j^{(t-1)} \end{aligned} \tag{47}$$

($||a||_2^2 = a^T a$ and $a^T b = b^T a = K$ for 1D matrix K)

$$\alpha_j^{(t-1)^T} D^T \Phi_j^{(2)^T} \Phi_j^{(2)} D\alpha_j^{(t-1)} \tag{48}$$

$$\begin{aligned} = & Tr(\alpha_j^{(t-1)^T} D^T \Phi_j^{(2)^T} \Phi_j^{(2)} D\alpha_j^{(t-1)}) \quad \text{(1D matrix)} \\ = & Tr(D^T \Phi_j^{(2)^T} \Phi_j^{(2)} D\alpha_j^{(t-1)} \alpha_j^{(t-1)^T}) \quad (\ tr(AB) = tr(BA)) \end{aligned} \tag{49}$$

$$
\begin{aligned}
(AB)_{i,j} &= \sum_n A_{i,n} B_{n,j} \\
tr(AB) &= \sum_i \sum_n A_{i,n} B_{n,i} \\
&= vec(A)^T vec(B) \qquad (50)
\end{aligned}
$$

where $vec(A)^T = [A_{11}, A_{12}....A_{1n}, A_{21}....A_{mn}]$ and $vec(B)^T = [A_{11}, A_{21}....A_{m1}, A_{12}....A_{mn}]$. vec(.)is in column major vectorized format. Thus, D^T is replaced by D. Using result(50) in Eq(49),

$$
\alpha_j^{(t-1)^T} D^T \Phi_j^{(2)^T} \Phi_j^{(2)} D \alpha_j^{(t-1)} = vec(D)^T vec(\Phi_j^{(2)^T} \Phi_j^{(2)} D \alpha_j^{(t-1)} \alpha_j^{(t-1)^T}) \qquad (51)
$$

Consider $vec(\Phi_j^{(2)^T} \Phi_j^{(2)} D \alpha_j^{(t-1)} \alpha_j^{(t-1)^T}) = \mathrm{V}$ and $\Phi_j^{(2)^T} \Phi_j^{(2)} D \alpha_j^{(t-1)} \alpha_j^{(t-1)^T} = \mathrm{A}$. Also $\Phi_j^{(2)^T} \Phi_j^{(2)} = \mathrm{C}$ and $\alpha_j^{(t-1)} \alpha_j^{(t-1)^T} = \mathrm{E}$. C is nxn, D is nxp and E is pxp.

$$
\begin{aligned}
A_{i,j} &= \sum_{l=1}^{p} (CD)_{i,l} E_{l,j} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{p} C_{i,k} D_{k,l} E_{l,j} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{p} (C)_{i,k} E_{l,j} D_{k,l} \qquad (52)
\end{aligned}
$$

D is getting summed over all elements. $Vec(A) = V$. Then $A_{i,j} = V_{(j-1)n+i}$. Consider a matrix M with size pnxpn with $M_{(j-1)n+i,(l-1)n+k} = C_{i,k} E_{l,j} = C_{i,k} E_{j,l}$, since E is also symmetric. We can observe that $V = Mvec(D)$ and $M = E \otimes C$. Thus we obtain,

$$
\alpha_j^{(t-1)^T} D^T \Phi_j^{(2)^T} \Phi_j^{(2)} D \alpha_j^{(t-1)} = vec(D)^T (\alpha_j^{(t-1)} \alpha_j^{(t-1)^T} \otimes \Phi_j^{(2)^T} \Phi_j^{(2)}) vec(D) \qquad (53)
$$

$$
\begin{aligned}
y_j^T \Phi_j^{(2)} D \alpha_j^{(t-1)} &= tr(y_j^T \Phi_j^{(2)} D \alpha_j^{(t-1)}) \quad \text{(1D array)} \\
&= tr(\alpha_j^{(t-1)} y_j^T \Phi_j^{(2)} D) \quad \text{(tr(AB)=tr(BA))} \\
&= vec(\Phi_j^{(2)^T} y_j \alpha_j^{(t-1)^T})^T vec(D) \quad \text{(Result(50))} \qquad (54)
\end{aligned}
$$

Let d = vec(D), standard form of $g_t(D)$ can be written as

$$
g(d) = \frac{1}{2} d^T H_d d + f_t^T d + c \qquad (55)
$$

where

$$
H_t = \frac{2}{nN} \sum_{j=1}^{N} (\alpha_j^{(t-1)} \alpha_j^{(t-1)^T} \otimes \Phi_j^{(2)^T} \Phi_j^{(2)}) \qquad (56)
$$

$$f_t = \frac{-2}{nN}\sum_{j=1}^{N} vec(\Phi_j^{(2)^T} y_j \alpha_j^{(t-1)^T}) \tag{57}$$

$$c = \frac{1}{nN}\sum_{j=1}^{N} y_j^T y_j \tag{58}$$

The convergence rate of g_t is bounded as [32]

$$\frac{g_t(D^{(t)}) - g_t^*}{g_t(D^{(t-1)}) - g_t^*} \leq 1 - \kappa_t^{-1} \tag{59}$$

For fastest decay we should have $\kappa_t \approx 1$. However, it is data dependent. Thus, we compare it with respect to the Oracle's condition number. The hessian matrix of the oracle can be obtained by substituting Φ_j as I_n since that would mean the observed samples are same as original samples. Thus, we obtain,

$$\hat{H}_t = \frac{2}{nN}\sum_{j=1}^{N} (\alpha_j^{(t-1)} \alpha_j^{(t-1)^T} \otimes I_n) \tag{60}$$

The condition number for the Oracle is defined as $\hat{\kappa_t}$

3.2 Proof of Lemma 3

First we prove

$$E(e^{\theta X}) \preceq exp\left(\frac{e^{R\theta} - 1}{R} E(X)\right) \tag{61}$$

For a function $f(x) = e^{\theta x}$, $\theta > 0$ it is convex. Thus for $x \in [0, R]$,

$$\begin{aligned} \frac{f(x) - f(0)}{x} &\leq \frac{f(R) - f(0)}{R} \\ e^{\theta x} &\leq 1 + \frac{e^{R\theta} - 1}{R} x \end{aligned} \tag{62}$$

if $f(a) \leq g(a) \forall a$ then $f(\Lambda) \preceq g(\Lambda)$. This is because the diagonal elements of the matrix $g(\Lambda) - f(\Lambda)$ will be non-negative and for a diagonal matrix these are the eigenvalues. Thus, $g(\Lambda) - f(\Lambda)$ will be positive semi-definite which proves that $f(\Lambda) \preceq g(\Lambda)$.
We define the matrix $g(\Lambda) - f(\Lambda)$ as λ. For any row matrix v, $v\lambda v^* \geq 0$. We can replace v by vB which is also a vector. Thus, $vB\lambda B^* v^* \geq 0$. Therefore, $B\lambda B^*$ is also positive semidefinite for any B. Hence, we proved that $g(X) - f(X)$ will be positive semidefinite for X. Thus, $f(X) \preceq g(X)$.

Therefore, we obtain

$$e^{\theta X} \preceq I + \frac{e^{R\theta} - 1}{R} X \tag{63}$$

$$E(e^{\theta X}) \preceq I + \frac{e^{R\theta} - 1}{R} E(X)$$

$$E(e^{\theta X}) \preceq exp\left(\frac{e^{R\theta} - 1}{R} E(X) \right) \tag{64}$$

for $\theta > 0$. We use the below result mentioned in [37] (without proof)

$$A \preceq H \implies tr(e^A) \leq tr(e^H) \tag{65}$$

Consider $Y = \sum_k X_k$.

$$
\begin{aligned}
p(\lambda_{max}(Y) \geq t) =& P(e^{\theta\lambda_{max}(Y)} \geq e^{\theta t}) \quad (\theta > 0) \\
\leq& e^{-\theta t} E(e^{\theta\lambda_{max}(Y)}) \quad \text{(Markov's inequality)} \\
=& e^{-\theta t} E(e^{\lambda_{max}(\theta Y)}) \\
=& e^{-\theta t} E(\lambda_{max}(e^{\theta Y})) \\
\leq& e^{-\theta t} E(tr(e^{\theta Y})) \qquad (66) \\
\leq& inf_{\theta\geq 0} e^{-\theta t} E(tr(e^{\theta Y})) \\
\leq& inf_{\theta\geq 0} e^{-\theta t} tr(E(e^{\theta Y})) \\
=& inf_{\theta\geq 0} e^{-\theta t} tr(E(exp(\theta \sum_k X_k))) \\
\leq& inf_{\theta\geq 0} e^{-\theta t} tr(exp(E(\sum_k \theta X_k))) \quad \text{(Jensen's inequality)} \\
=& inf_{\theta\geq 0} e^{-\theta t} tr(exp(E(\sum_k log(exp(\theta X_k))))) \\
=& inf_{\theta\geq 0} e^{-\theta t} tr(exp(\sum_k E(log(exp(\theta X_k))))) \\
\leq& inf_{\theta\geq 0} e^{-\theta t} tr(exp(\sum_k log(E(exp(\theta X_k))))) \quad \text{(Jensen's inequality)} \\
& \qquad (67)
\end{aligned}
$$

$$
\begin{aligned}
E(\lambda_{max}(Y)) =& \frac{1}{\theta} E(\lambda_{max}(\theta \sum_k X_k)) \\
=& \frac{1}{\theta} E(log(exp(\lambda_{max}(\theta \sum_k X_k)))) \\
=& \frac{1}{\theta} E(log(\lambda_{max}(exp(\theta \sum_k X_k)))) \\
\leq& \frac{1}{\theta} E(log(tr(exp(\theta \sum_k X_k)))) \\
\leq& \frac{1}{\theta} log(E(tr(exp(\theta \sum_k X_k)))) \quad \text{(Jensen's inequality)} \\
\leq& inf_{\theta\geq 0} \frac{1}{\theta} log(tr(exp(\sum_k log(E(e^{\theta X_k}))))) \qquad (68) \\
\leq& inf_{\theta\geq 0} \frac{1}{\theta} log(tr(exp(g(\theta) \sum_k E(X_k))))) \text{(Eq.(64,65)} \\
\leq& inf_{\theta\geq 0} \frac{1}{\theta} log(d \cdot \lambda_{max}(exp(g(\theta) \sum_k E(X_k))))) \\
\leq& inf_{\theta\geq 0} \frac{1}{\theta} log(d \cdot exp(\lambda_{max}(g(\theta) \sum_k E(X_k))))) \\
\leq& inf_{\theta\geq 0} \frac{1}{\theta} log(d \cdot exp(g(\theta) \lambda_{max}(\sum_k E(X_k))))) \\
=& inf_{\theta\geq 0} \frac{1}{\theta} \left[logd + g(\theta)\mu_{max} \right] \qquad (69)
\end{aligned}
$$

where $\mu_{max} = \lambda_{max}(E(Y))$ On similar terms, we can also obtain the bound for $E(\lambda_{min}(Y))$. Using Eq(64,65,67) we obtain,

$$\begin{aligned} p(\lambda_{max}(Y) \geq t) \leq & inf_{\theta \geq 0} e^{-\theta t} tr(exp(\sum_k g(\theta) E(X_k))) \\ = & inf_{\theta \geq 0} e^{-\theta t} tr(exp(g(\theta) \sum_k E(X_k))) \\ = & inf_{\theta \geq 0} e^{-\theta t} tr(exp(g(\theta) \mu_{max})) \end{aligned} \tag{70}$$

Similar bounds can be calculated for $p(\lambda_{max}(Y) \geq t)$.

3.3 Lemma 4

We substitute S as H_t in lemma 3.
$E({\Phi_j^{(2)}}^T \Phi_j^{(2)}) = I_n$ since $\phi_{i,j}$ are iid and thus non-diagonal elements are zero while diagonal elements are $\frac{1}{m_j} m_j = 1$.
ξ_N is calculated as

$$\begin{aligned} \xi_n = & \lambda_{max}(\alpha_j^{(t-1)} {\alpha_j^{(t-1)}}^T \otimes {\Phi_j^{(2)}}^T \Phi_j^{(2)}) \\ = & \lambda_{max}(\alpha_j^{(t-1)} {\alpha_j^{(t-1)}}^T) \lambda_{max}({\Phi_j^{(2)}}^T \Phi_j^{(2)}) \end{aligned} \tag{71}$$

The last step is proved as follows, suppose for eigenvectors v_1, v_2 and corresponding eigenvalues λ_1, λ_2, Consider A to be pxp and B to be nxn.

$$\begin{aligned} Av_1 = & \lambda_1 v_1 \\ Bv_2 = & \lambda_2 v_2 \\ Av_1 \otimes Bv_2 = & \lambda_1 v_1 \otimes \lambda_2 v_2 \\ Av_1 \otimes Bv_2 = & \lambda_1 \lambda_2 (v_1 \otimes v_2) \end{aligned} \tag{72}$$

$$\begin{aligned} (Av_1 \otimes Bv_2)_{(i-1)n+j} = & (Av_1)_i (Bv_2)_j \\ = & (\sum_{r=1}^{p} A_{i,r} v_{1_{r,1}}) (\sum_{z=1}^{n} B_{j,z} v_{2_{z,1}}) \\ = & \sum_{r=1}^{p} \sum_{z=1}^{n} A_{i,r} B_{j,z} v_{1_{r,1}} v_{2_{z,1}} \end{aligned} \tag{73}$$

$$\begin{aligned} [(A \otimes B)(v_1 \otimes v_2)]_{(i-1)n+j,1} = & \sum_{k=1}^{k=pn} (A \otimes B)_{(i-1)n+j,k} (v_1 \otimes v_2)_{k,1} \\ = & \sum_{r=1}^{p} \sum_{z=1}^{n} (A \otimes B)_{(i-1)n+j,(r-1)n+z} (v_1 \otimes v_2)_{(r-1)n+z,1} \\ = & \sum_{r=1}^{p} \sum_{z=1}^{n} A_{i,r} B_{j,z} v_{1_{r,1}} v_{2_{z,1}} \end{aligned} \tag{74}$$

Thus, using Eq(72,73,74), we prove Eq(71).
From [34], for a Gaussian Φ_j, we get (**I was not able to prove this.**)

$$P(\lambda_{max}(\Phi_j^T \Phi_j) \geq (1+\eta)) \leq 2e^{-m(\frac{\eta^2}{4}-\frac{\eta^3}{6})} \tag{75}$$

Thus, with a probability of atleast $1-p_o$.

$$\begin{aligned} p_o \leq & 2e^{-m(\frac{\eta^2}{4}-\frac{\eta^3}{6})} \\ \frac{1}{m}ln(\frac{2}{p_o}) \geq & (\frac{\eta^2}{4}-\frac{\eta^3}{6}) \\ \frac{1}{m}ln(\frac{2}{p_o}) \geq & (\frac{\eta^2}{4}-\frac{\eta^2}{6})(eta \leq 1) \\ \sqrt{\frac{12}{m}ln(\frac{2}{p_o})} \geq & \eta \\ \lambda_{max} \leq & 1+\sqrt{\frac{12}{m}ln(\frac{2}{p_o})} \end{aligned} \tag{76}$$

$\alpha_j^{(t-1)}\alpha_j^{(t-1)^T}$ is a rank 1 matrix. Thus its maximum eigenvalue is its only non-zero eigenvalue. We can observe that the vector $\alpha_j^{(t-1)}$ is an eigenvector of the matrix with the eigenvalue $||\alpha_j^{(t-1)}||_2^2$.

$$\frac{1}{n}\lambda_{max}(\alpha_j^{(t-1)}\alpha_j^{(t-1)^T}) = \frac{1}{n}||\alpha_j^{(t-1)}||_2^2 \tag{77}$$

$$\leq \frac{1}{n}||x||_2^2 \tag{78}$$

$$\leq 1(\text{Each element in X is} \leq 1) \tag{79}$$

The paper doesn't mention any constraint on D. However, I feel there should be a constraint on the Frobenius norm of D for eqn (78) to be satisfied. Thus, with probability $1-p_o$,

$$\xi_N \leq \frac{2}{N}(1+\sqrt{\frac{12}{m}ln(\frac{2}{p_o})}) \tag{80}$$

There appears to be a typo in the paper. Paper mentions 24 instead of 12 for the equation above. However, 12 is consistent.
Thus, substituting above quantities, bounds are established as follows,

$$P(\lambda_{max}(H_t) \geq (1+\delta)\lambda_{max}(\hat{H}_t)) \leq np\left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\frac{\lambda_{max}(\hat{H}_t)}{\xi_N}} \tag{81}$$

$$P(\lambda_{min}(H_t) \leq (1-\delta)\lambda_{min}(\hat{H}_t)) \leq np\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\lambda_{min}(\hat{H}_t)}{\xi_N}} \tag{82}$$

3.4 Calculating the order of δ_N

We define,

$$p_1 = np\left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\frac{\lambda_{max}(\hat{H}_t)}{\xi_N}} \tag{83}$$

$$p_2 = np\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\lambda_{min}(\hat{H}_t)}{\xi_N}} \tag{84}$$

Dividing by np and taking log we get,

$$\frac{\xi_N}{\lambda_{max}(\hat{H}_t)} ln(\frac{np}{p_1}) = (-\delta + (1+\delta)ln(1+\delta)) \tag{85}$$

$$\frac{\xi_N}{\lambda_{min}(\hat{H}_t)} ln(\frac{np}{p_2}) = (\delta + (1-\delta)ln(1-\delta)) \tag{86}$$

Expanding $ln(1+\delta)$ with $\delta < 1$ and ignoring higher terms. We get RHS of both Eq.(85,86) as $\frac{\delta_n^2}{2}$. For getting surer probabilistic bound, we take the higher value of δ_N out of the two. Thus,

$$\delta_N = \sqrt{2\xi_N} max\left(\sqrt{\frac{ln(\frac{np}{p_1})}{\lambda_{max}(\hat{H}_t)}}, \sqrt{\frac{ln(\frac{np}{p_2})}{\lambda_{min}(\hat{H}_t)}}\right) \tag{87}$$

In the paper there is not a factor of 2. However, the order of δ_N remains the same. Moreover, I am sure that there should be a factor of 2.

$\sqrt{\xi_N}$ is of the order $O(N^{\frac{-1}{2}} m^{\frac{-1}{4}})$ from Eq(80).
Thus, δ_N is of the order $O(m^{\frac{-1}{4}} N^{\frac{-1}{2}} [ln(np)]^{\frac{1}{2}})$

3.5 Finding the convergence rate

$$\begin{aligned}
P\left(\kappa_t \geq \frac{1+\delta_t}{1-\delta_t}\hat{\kappa_t}\right) \leq & P(\left[\lambda_{max}(H_t) \geq (1+\delta_t)\lambda_{max}(\hat{H}_t)\right] \cup \left[\lambda_{min}(H_t) \leq (1-\delta_t)\lambda_{min}(\hat{H}_t)\right]) \\
= & 1 - P(\left[\lambda_{max}(H_t) \leq (1+\delta_t)\lambda_{max}(\hat{H}_t)\right] \cap \left[\lambda_{min}(H_t) \geq (1-\delta_t)\lambda_{min}(\hat{H}_t)\right]) \\
= & 1 - P(\lambda_{max}(H_t) \leq (1+\delta_t)\lambda_{max}(\hat{H}_t)) P(\lambda_{min}(H_t) \geq (1-\delta_t)\lambda_{min}(\hat{H}_t)) \\
& \textbf{(Considering both the events are independent.)} \\
= & 1 - (1-p_1)(1-p_2) \\
= & p_1 + p_2 - p_1 p_2
\end{aligned} \tag{88}$$

Thus, with a probability of atleast $(1-p_1)(1-p_2)$ we have $\kappa_t \geq \frac{1+\delta_t}{1-\delta_t}\hat{\kappa_t}$. Cross-multiplying in Eq(59)

$$\begin{aligned}
g_t(D^{(t)}) - g_t^* \leq & (1-\kappa_t^{-1})(g_t(D^{(t-1)}) - g_t^*) \\
\frac{g_t(D^{(t)}) - g_t^*}{g_t(D^{(t-1)})} \leq & (1-\kappa_t^{-1})\left(\frac{g_t(D^{(t-1)}) - g_t^*}{g_t(D^{(t-1)})}\right) \\
\frac{g_t(D^{(t)})}{g_t(D^{(t-1)})} \leq & 1-\kappa_t^{-1}\left(\frac{g_t(D^{(t-1)}) - g_t^*}{g_t(D^{(t-1)})}\right) \qquad (89)
\end{aligned}$$

Using Eq(40), with a high probability, we get

$$\begin{aligned}
\kappa_t^{-1}\left(\frac{g_t(D^{(t-1)}) - g_t^*}{g_t(D^{(t-1)})}\right) \geq & \left(\frac{1+\delta_N}{1-\delta_N}\right)\hat{\kappa_t^{-1}}\left(1 - \frac{(1+\epsilon_N)}{(1-\epsilon_N)}\frac{h_t^*}{h_t(D^{(t-1)})}\right) \\
\approx & (1+2\delta_N)\hat{\kappa_t}^{-1}\left(1-(1+2\epsilon_N)\frac{h_t^*}{h_t(D^{(t-1)})}\right) \text{(Using binomial expansion)} \\
\approx & \left(1-\frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1} - 2\epsilon_N\frac{h_t^*}{h_t(D^{(t-1)})}\hat{\kappa_t}^{-1} \\
& - 2\delta_N\left(1-\frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1} \quad (\epsilon_N\delta_N \approx 0) \qquad (90) \\
= & \left(1-\frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1} - 2\rho_N \qquad (91)
\end{aligned}$$

where

$$\rho_N = \epsilon_N\frac{h_t^*}{h_t(D^{(t-1)})}\hat{\kappa_t}^{-1} + \delta_N\left(1-\frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1} \qquad (92)$$

ρ_N is a convex combination of δ_N and ϵ_N and thus it is of the order $O(m^{\frac{-1}{4}}N^{\frac{-1}{2}}[ln(np)]^{\frac{1}{2}})$. Using Eq.(45,89,91),

$$\frac{g_t - g_T}{g_{t-1} - g_T} \leq \gamma\left[1-\left(1-\frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1}\right] + 2\gamma\rho_N \qquad (93)$$

The first term will correspond to the convergence rate if the measurements were complete since it contains oracle values while the second term can be considered as the penalty associated with using incomplete measurements which is of the order $O(\gamma m^{\frac{-1}{4}}N^{\frac{-1}{2}}[ln(np)]^{\frac{1}{2}})$.

4 Decay rate of h_t

$$\begin{aligned}\frac{h_t - h_T}{h_{t-1} - h_T} \leq & \frac{h_t}{h_{t-1}} \quad \text{(Cross-multiply and use } h_t < h_{t-1}\text{)} \\ \leq & \left(\frac{1+\epsilon_N}{1-\epsilon_N}\right)\frac{g_t}{g_{t-1}} \quad \text{(Using Eq.(40))} \end{aligned} \tag{94}$$

To ensure $h_t < h_{t-1}$,

$$\begin{aligned}\left(\frac{1+\epsilon_N}{1-\epsilon_N}\right)\frac{g_t}{g_{t-1}} \leq & 1 \\ \frac{g_t}{g_{t-1}} \leq & \left(\frac{1-\epsilon_N}{1+\epsilon_N}\right) \end{aligned} \tag{95}$$

Using Eq.(45) we know that

$$\begin{aligned}\frac{g_t}{g_{t-1}} \leq & \frac{g_t(D^{(t)})}{g_t(D^{(t-1)})} \cdot \gamma \\ \leq & \gamma \quad \text{(Since, gradient descent is used)} \end{aligned} \tag{96}$$

Thus, taking $\gamma \leq \left(\frac{1-\epsilon_N}{1+\epsilon_N}\right)$ ensures $h_t < h_{t-1}$.
A very small γ would result in a quick convergence but the dictionary might not be able to adapt itself to the data. Also, if the data is insufficient then $\epsilon_N \approx 1$ and thus, an appropriate choice of γ wont be possible.
Using Eq(45,94),

$$\begin{aligned}\frac{h_t}{h_{t-1}} \leq & \left(\frac{1+\epsilon_N}{1-\epsilon_N}\right)\frac{g_t(D^{(t)})}{g_t(D^{(t-1)})}\gamma \\ \leq & (1+2\epsilon_N)\gamma\left[1 - \left(1 - \frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1} + 2\rho_N\right] \text{(Eq(45,93) and using binomial expansion on } \epsilon_N\text{)} \\ \approx & \gamma\left[1 - \left(1 - \frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1}\right] + 2\gamma\epsilon_N\left[1 - \left(1 - \frac{h_t^*}{h_t(D^{(t-1)})}\right)\hat{\kappa_t}^{-1}\right] \\ & + 2\gamma\rho_N \quad (\epsilon_N\rho_N \approx 0, \text{ since } \rho_N \text{ is a convex combination of } \epsilon_N \text{ and } \delta_N) \end{aligned} \tag{97}$$

Similarly as before, the first term will correspond to the convergence rate if the measurements were complete since it contains oracle values while the second term can be considered as the penalty associated with using incomplete measurements which is of the order $O(\gamma m^{\frac{-1}{4}} N^{\frac{-1}{2}} [ln(np)]^{\frac{1}{2}})$.
γ is a regularizer and should be selected depending on the noise in the signal. This can also be decided based on the validation error.
ϵ_N depends on the sampling matrix. For a fixed N,m, having a matrix with highly random values would reduce ϵ_N more compared to having a binary sampling matrix which is there in the case of image inpainting. Also, having a large non-overlapping blocks, N means more data and thus the dictionary matrix can adjust more to make the oracle error and the projected error similar. Thus, this will ensure a smaller ϵ_N.