1 SymSGD Technical Report

1.1 Variance and Covariance of $\frac{1}{r} M \cdot A \cdot A^T \cdot \Delta w$

In here, for the sake of simplicity, we use w instead of Δw and instead of k for the size of the projected space, we use r since k is used for summation indices in here, heavily. We want to estimate $v = M \cdot w$ with $\frac{1}{r} M \cdot A \cdot A^T \cdot w$, where A is a $f \times r$ matrix, where a_{ij} is a random variable with the following properties.

$$\begin{aligned}
&\mathbf{E}(a_{ij}) = 0 \\
&\mathbf{E}(a_{ij}^2) = 1 \\
&\mathbf{E}(a_{ij}^4) = \rho = 3 && \text{which makes the math simpler}
\end{aligned}$$

Let m_s^T be some row of M. Its estimation in $M \cdot w$ is $v_s = \frac{1}{r} \cdot m_s^T \cdot A \cdot A^T \cdot w$. It is easy to see that $\mathbf{E}(v_s) = m_s^T \cdot w$.

$$\begin{aligned}
\mathbf{E}(v_s) &= \mathbf{E}(\frac{1}{r} \sum_{i,j,k} m_{si} a_{ij} a_{kj} w_k) \\
&= \frac{1}{r} \sum_{i,j,k} m_{si} \mathbf{E}(a_{ij} a_{kj}) w_k \\
&= \frac{1}{r} (\sum_{i,j,k:i=k} m_{si} \mathbf{E}(a_{ij} a_{kj}) w_k + \sum_{i,j,k:i \neq k} m_{si} \mathbf{E}(a_{ij} a_{kj}) w_k) \\
&= \frac{1}{r} (\sum_{i,j} m_{si} \mathbf{E}(a_{ij} a_{ij}) w_i + \sum_{i,j,k:j \neq k} m_{si} \mathbf{E}(a_{ij}) \mathbf{E}(a_{kj}) w_k) \\
&= \frac{1}{r} \sum_{i,j} m_{si} \cdot w_i \\
&= m_s^T \cdot w
\end{aligned}$$

We will use the notation $ij = kl$ to mean $i = k \wedge j = l$, and $ij \neq kl$ to mean its negation. Let m_s, m_t be two rows of M. We want to find the covariance of the resulting v_s and v_t.

$$
\begin{aligned}
& r^2 \cdot \mathbf{E}(v_s, v_t) \\
&= r^2 \cdot \mathbf{E}\Big(\frac{1}{r^2} \sum_{i,j,k} m_{si} a_{ij} a_{kj} w_k \cdot \sum_{i',j',k'} m_{ti'} a_{i'j'} a_{k'j'} w_{k'}\Big) \\
&= \sum_{i,j,k,i',j',k'} m_{si} m_{ti'} w_k w_{k'} \mathbf{E}(a_{ij} a_{kj} a_{i'j'} a_{k'j'}) \\
&= \sum_{i,j,k,i',j',k' : ij=kj=i'j'=k'j'} m_{si} m_{ti'} w_k w_{k'} \mathbf{E}(a_{ij} a_{kj} a_{i'j'} a_{k'j'}) \\
&+ \sum_{i,j,k,i',j',k' : ij=kj \neq i'j'=k'j'} m_{si} m_{ti'} w_k w_{k'} \mathbf{E}(a_{ij} a_{kj} a_{i'j'} a_{k'j'}) \\
&+ \sum_{i,j,k,i',j',k' : ij=i'j' \neq kj=k'j'} m_{si} m_{ti'} w_k w_{k'} \mathbf{E}(a_{ij} a_{kj} a_{i'j'} a_{k'j'}) \\
&+ \sum_{i,j,k,i',j',k' : ij=k'j' \neq i'j'=kj} m_{si} m_{ti'} w_k w_{k'} \mathbf{E}(a_{ij} a_{kj} a_{i'j'} a_{k'j'}) \qquad \text{as terms with } \mathbf{E}(a_{ij}) \text{ cancel out} \\
&= \sum_{i,j} m_{si} m_{ti} w_i w_i \rho + \sum_{i,j,i',j' : ij \neq i'j'} m_{si} m_{ti'} w_i w_{i'} \\
&+ \sum_{i,j,k : i \neq k} m_{si} m_{ti} w_k w_k + \sum_{i,j,k : i \neq k} m_{si} m_{tk} w_k w_i \qquad \text{as } \mathbf{E}(a_{ij} a_{kl}) = 1 \text{ when } ij \neq kl \\
&= \rho \sum_{i,j} m_{si} m_{ti} w_i^2 \\
&+ \sum_{i,j,i',j'} m_{si} m_{ti'} w_i w_{i'} - \sum_{i,j,i',j' : ij=i'j'} m_{si} m_{ti'} w_i w_{i'} \\
&+ \sum_{i,j,k} m_{si} m_{ti} w_k^2 - \sum_{i,j,k : i=k} m_{si} m_{ti} w_k^2 \\
&+ \sum_{i,j,k} m_{si} m_{tk} w_k w_i - \sum_{i,j,k : i=k} m_{si} m_{tk} w_k w_i \\
&= (\rho - 3) \sum_{i,j} m_{si} m_{ti} w_i^2 + \sum_{i,j,i',j'} m_{si} m_{ti'} w_i w_{i'} \\
&+ \sum_{i,j,k} m_{si} m_{ti} w_k^2 + \sum_{i,j,k} m_{si} m_{tk} w_k w_i \\
&= r^2 \sum_{i,i'} m_{si} m_{ti'} w_i w_{i'} + r \sum_{i,k} m_{si} m_{ti} w_k^2 + r \sum_{i,k} m_{si} m_{tk} w_i w_k \qquad \text{as } \rho = 3 \text{ and } j \in [1 \ldots k] \\
&= (r^2 + r) \sum_{i,i'} m_{si} m_{ti'} w_i w_{i'} + r \cdot m_s^T \cdot m_t \sum_k w_k^2
\end{aligned}
$$

In other words

$$\mathbf{E}(v_s v_t) = (1 + \frac{1}{r}) \sum_{i,i'} m_{si} m_{ti'} w_i w_{i'} + \frac{1}{r} \cdot m_s^T \cdot m_t \sum_k w_k^2$$

The covariance $\mathrm{Cov}(a, b) = \mathbf{E}(a \cdot b) - \mathbf{E}(a)\mathbf{E}(b)$. Using this we have

$$\begin{aligned}
&\mathrm{Cov}(v_s, v_t) \\
&= (1 + \frac{1}{r}) \sum_{i,i'} m_{si} m_{ti'} w_i w_{i'} + \frac{1}{r} \cdot m_s^T \cdot m_t \sum_k w_k^2 - \mathbf{E}(v_s)\mathbf{E}(v_t) \\
&= (1 + \frac{1}{r}) \sum_{i,i'} m_{si} m_{ti'} w_i w_{i'} + \frac{1}{r} \cdot m_s^T \cdot m_t \sum_k w_k^2 - \mathbf{E}(v_s)\mathbf{E}(v_t) \\
&= (1 + \frac{1}{r})\mathbf{E}(v_s)\mathbf{E}(v_t) + \frac{1}{r} \cdot m_s^T \cdot m_t \sum_k w_k^2 - \mathbf{E}(v_s)\mathbf{E}(v_t) \\
&= \frac{1}{r}\mathbf{E}(v_s)\mathbf{E}(v_t) + \frac{1}{r} \cdot m_s^T \cdot m_t \sum_k w_k^2 \\
&= \frac{1}{r}\mathbf{E}(v_s)\mathbf{E}(v_t) + \frac{1}{r} \cdot (M \cdot M^T)_{st} \left\| w \right\|_2^2 \\
&= \frac{1}{r}(M \cdot w)_s (M \cdot w)_t + \frac{1}{r} \cdot (M \cdot M^T)_{st} \left\| w \right\|_2^2 \\
&= \frac{1}{r}((M \cdot w) \cdot (M \cdot w)^T)_{st} + \frac{1}{r} \cdot (M \cdot M^T)_{st} \left\| w \right\|_2^2
\end{aligned}$$

Let $\mathbb{C}(v)$ be the covariance matrix of v. That is, $\mathbb{C}(v)_{ij} = \mathrm{Cov}(v_i, v_j)$. So, we have

$$\mathbb{C}(v) = \frac{1}{r}(M \cdot w) \cdot (M \cdot w)^T + \frac{1}{r}(M \cdot M^T) \left\| w \right\|_2^2$$

Note that we can use this computation for matrix $N = M - I$ as well since we did not assume anything about the matrix M from the beginning. Therefore, for $v' = w + \frac{1}{r} N \cdot A \cdot A^T \cdot w$, $\mathbb{C}(v') = \frac{1}{r}(N \cdot w) \cdot (N \cdot w)^T + \frac{1}{r}(N \cdot N^T) \left\| w \right\|_2^2$ since w is a constant in v' and $\mathbb{C}(a + x) = \mathbb{C}(x)$ for any constant vector a and any probabilistic vector x. Next we try to bound $\mathbb{C}(v)$.

2 Bounding $\mathbb{C}(v)$

We can bound $\mathbb{C}(v)$ by computing its trace since $tr(\mathbb{C}(v)) = \sum_i var(v_i)$, the summation of the variance of elements of v.

$$\begin{aligned} tr(\mathbb{C}(v)) &= \frac{1}{r} tr((M \cdot w) \cdot (M \cdot w)^T) + \frac{1}{r} \|w\|_2^2 \, tr(MM^T) \\ &= \frac{1}{r} \|M \cdot w\|_2^2 + \frac{1}{r} \|w\|_2^2 \left(\sum_i \lambda_i (M \cdot M^T) \right) \\ &= \frac{1}{r} \|M \cdot w\|_2^2 + \frac{1}{r} \|w\|_2^2 \left(\sum_i \sigma_i (M)^2 \right) \end{aligned}$$

where $\lambda_i M \cdot M^T$ is the i^{th} largest eigenvalue of $M \cdot M^T$ which is the square of i^{th} largest singular value of M, $\sigma_i(M)^2$. Since $\|M \cdot w\|_2^2 \leq \|w\|_2^2 \|M\|_2^2 = \|w\|_2^2 \sigma_{max}(M)^2$, we can bound $tr(\mathbb{C}(v))$ as follows:

$$tr(\mathbb{C}(v)) \leq \frac{1}{r} (\sigma_{max}(M)^2) + \frac{1}{r} \|w\|_2^2 \left(\sum_i \sigma_i (M)^2 \right)$$

It is trivial to see that:

$$\frac{1}{r} \|w\|_2^2 \left(\sum_i \sigma_i (M)^2 \right) \leq tr(\mathbb{C}(v))$$

Combining the two inequalities, we have:

$$\frac{1}{r} \|w\|_2^2 \left(\sum_i \sigma_i (M)^2 \right) \leq tr(\mathbb{C}(v)) \frac{1}{r} (\sigma_{max}(M)^2) + \frac{1}{r} \|w\|_2^2 \left(\sum_i \sigma_i (M)^2 \right)$$

The same bounds can be derived when $N = M - I$ is used.

3 Rank of Matrix M

Lemma 3.1. *For the matrix* $M_{a \to b} = \prod_{i=b}^{a} (I - \alpha X_i^T \cdot X_i)$, $\text{rank}(M_{a \to b} - I) \leq b - a$.

Proof. The proof is by induction. The base case is when $a = b$ and $M_{a \to b} = I$. It is clear that $I - I = 0$ which is of rank zero. For the inductive step, assume that $\text{rank}(M_{a \to b-1} - I) \leq b - a - 1$. We have

$$\begin{aligned} M_{a \to b} - I &= (I - \alpha X_b^T \cdot X_b) M_{a \to b-1} - I \\ &= (M_{a \to b-1} - I) - \alpha X_b^T \cdot (X_b \cdot M_{a \to b-1}) \end{aligned}$$

Term $\alpha X_b^T \cdot (X_b \cdot M_{a \to b-1})$ is a rank-1 matrix and term $(M_{a \to b-1} - I)$ is of rank $b - a - 1$ by induction hypothesis. Since for any two matrices A and B, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$, $\text{rank}(M_{a \to b} - I) \leq \text{rank}(M_{a \to b-1}) + \text{rank}(-\alpha X_b^T \cdot (X_b \cdot M_{a \to b-1})) \leq b - a - 1 + 1 = b - a$. □