Mean Imputation and Stochastic Coordinate Descent for Linear Systems with Missing Data

Meha Patel*, Samuel Rath†, and Chupeng Zheng‡

Projet Advisor: Anna Ma§

Abstract. As big data problems become more prevalent, the need to accurately approximate solutions to large-scale linear systems increases. Many real-world big data problems are also accompanied by the risk of missing or incomplete data, further complicating the linear models assigned to them. Current methods to address missing data involve deletion or zero imputation, which introduces bias to the model. We propose a model that adapts Stochastic Coordinate Descent (SCD) to handle missing data in linear systems and utilizes μ-imputation to retrieve a better approximation of the original data. We prove that in expectation, our proposed algorithm, μ-imputation mSCD utilizes an unbiased estimator of the gradient of the least-squares objective function when using mean imputation in the absence of data. Furthermore, we compare our algorithm's performance on synthetic data to closely related algorithms: zero-imputation mSCD and SCD. Finally, we apply μ-imputation mSCD on real-world data to demonstrate the usefulness and viability of our proposed algorithm.

1. Introduction.

Large-scale data is crucial in training algorithms that are being implemented and utilized today. However, the data collection process to train such algorithms is often imperfect and can lead to noisy and incomplete data. For example, malfunctions in physical measurement devices can cause data to become corrupt or, in extreme instances, unavailable. As another example, a person may skip questions on a questionnaire to save time resulting in incomplete survey data. While a straightforward approach for dealing with missing data is to impute the missing data with zeros or ignore missing data altogether (throwing out any and all incomplete data points), this quickly becomes impractical and inefficient.

Other methods for solving missing data problems consider imputation methods in which a new data set is created using information from the available data set. For example, in [6], Mukhopadhyay and Mukherjee propose an algorithm for imputing incomplete streaming data using a constant factor times the possibly imputed data at the previous time instance. While this method eliminates any inefficiency of solving a matrix with missing entries and provides a more accurate estimation of the full matrix, it is not independent of data being missing at random. Mukhopadhyay and Mukherjee assume that there is a previous entry from which we can draw information to determine the following entry. When we face the problem of incomplete data sets, we often find that the data is missing independently at random. This work focuses on missing data and how it arises in linear systems. In particular, we consider the following linear system of equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the measurement matrix that is not entirely known, i.e. only some of its entries are known, and we consider the over-determined case in which $m \geq n$, $\boldsymbol{x} \in \mathbb{R}^n$ is the unknown signal we wish to find, and $\boldsymbol{y} \in \mathbb{R}^m$ are the measurements. In most general cases, this system can be solved as $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{y}$,

*California State University Long Beach, Long Beach, CA (meha.patel01@student.csulb.edu)

†San Diego State University, San Diego, CA (sarath@sdsu.edu)

‡University of Chicago, Chicago, IL (chupenz@uchicago.edu)

§University of California, Irvine, Irvine, CA (anna.ma@uci.edu)

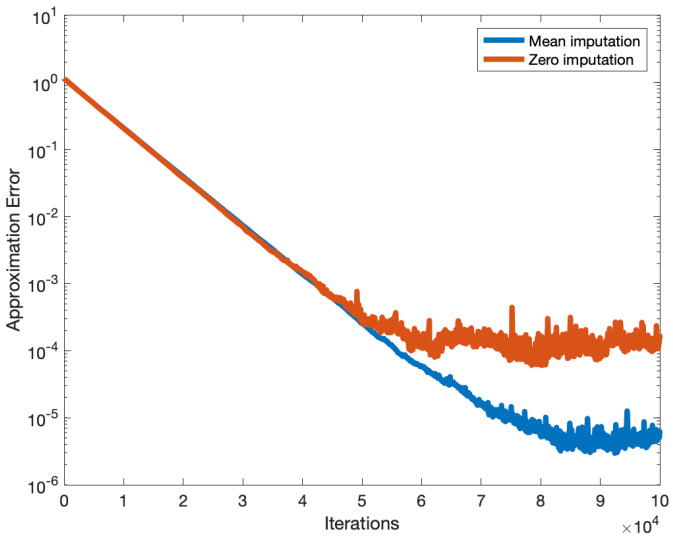

Figure 4. *Performance of μ-imputed mSCD on Garment Productivity data set from the UCI Machine Learning Repo.*

trials. We report the approximation error to the least squares solution for this data set using $\alpha = 5 \times 10^{-5}$ and $p = 0.85$. We observe that the approximation error decays linearly until the algorithm reaches a convergence horizon on the order of 10^{-5} for mean imputation and 10^{-4} for zero imputation, highlighting the benefit of using mean imputation.

5. Conclusion. We propose μ-imputed mSCD as a solution to big data problems in which missing or corrupt data issues arise. Our method utilizes an unbiased estimator of the gradient of the least-squares objective function as well as mean imputation to eliminate the issue of bias that may be introduced in the absence of data. The experimental results and theoretical proofs show that our method has less bias than current methods and, subsequently, performs better than these methods when solving large-scale linear systems with missing data. While we are able to show our algorithm's convergence capabilities through theoretical and practical experiments, we hope to analyze convergence guarantees in future work. Furthermore, for future work we believe our method can be expanded to other types of imputation methods or patterns of missing data. For example, we believe our idea of using mean imputation can be utilized to improve the cumulative information method used in [5] to create a cumulative information mean-imputed algorithm.

Acknowledgments. We thank Dr. Anna Ma from the University of California, Irvine for advising us and this work. We also thank Chelsea Huynh and Michael Strand from the University of California, Irvine for their contributions in facilitating the conversations around our research topic.

Appendix. In this section we prove the major theorem of the paper, namely Theorem 3.3. To do this, we first introduce some convenient notation, π and Π, to simplify the calculation.

This notation will also be used for the proof of Theorem 3.4, and in fact one of our goals is that it conveys the essential parallels between these proofs. Finally, we also prove Theorem 3.5.

By the assumption of our model, we treat the entries of a given matrix $\tilde{\boldsymbol{A}}$ as i.i.d. Bernoulli random variables such that

$$\tilde{a}_{ij} = \begin{cases} a_{ij} & w.p. \quad p \\ \mu & w.p. \quad 1-p \end{cases},$$

for which the expected value is simply

$$\mathbb{E}_\delta\left[\tilde{a}_{ij}\right] = pa_{ij} + (1-p)\mu. \tag{5.1}$$

However, when computing expectations we often run into products of 2 entries. For these, we need to consider 2 cases:

$$\tilde{a}_{ij}\tilde{a}_{ik} = \begin{cases} a_{ij}a_{ik} & w.p. \quad p^2 \\ \mu a_{ij} & w.p. \quad p(1-p) \\ \mu a_{ik} & w.p. \quad p(1-p) \\ \mu^2 & w.p. \quad (1-p)^2 \end{cases},$$

when $j \neq k$, and

$$\tilde{a}_{ij}\tilde{a}_{ik} = \begin{cases} a_{ij}a_{ik} & w.p. \quad p \\ \mu^2 & w.p. \quad 1-p \end{cases},$$

when $j = k$. These come from the fact that when an entry is multiplied by itself, it still corresponds to one coin toss with 2 outcomes. On the other hand, when the entries are distinct, we now have 2 independent coin tosses and 4 outcomes total. Thus, we get the following expectation:

$$\mathbb{E}_\delta\left[\tilde{a}_{ij}\tilde{a}_{ik}\right] = \begin{cases} p^2 a_{ij}a_{ik} + p(1-p)\left(a_{ij}\mu + \mu a_{ik}\right) + (1-p)^2\mu^2 & j \neq k \\ pa_{ij}a_{ik} + (1-p)\mu^2 & j = k \end{cases}. \tag{5.2}$$

The 2 cases in (5.2) occur so often in our proofs that we give them labels to condense the notation. These labels will also serve to indicate how the expected value operator $\mathbb{E}_\delta[\cdot]$ *acts* on different components in our calculations, depending on whether there are products of repeated or distinct entries. Let us define the functions $\Pi : \mathbb{R}^{m\times d} \times \mathbb{R}^{d\times n} \to \mathbb{R}^{m\times n}$ and $\pi : \mathbb{R}^{m\times d} \times \mathbb{R}^{d\times n} \to \mathbb{R}^{m\times n}$ such that

$$\Pi\left[\boldsymbol{A}, \boldsymbol{B}\right] := p^2\boldsymbol{AB} + p(1-p)\left(\boldsymbol{AM}_{\scriptscriptstyle B} + \boldsymbol{M}_{\scriptscriptstyle A}\boldsymbol{B}\right) + (1-p)^2\boldsymbol{M}_{\scriptscriptstyle A}\boldsymbol{M}_{\scriptscriptstyle B} \tag{5.3}$$

$$\pi\left[\boldsymbol{A}, \boldsymbol{B}\right] := p\boldsymbol{AB} + (1-p)\boldsymbol{M}_{\scriptscriptstyle A}\boldsymbol{M}_{\scriptscriptstyle B}, \tag{5.4}$$

where $\boldsymbol{M}_{\scriptscriptstyle A} = \mu\mathbb{1}_{m\times d}$ and $\boldsymbol{M}_{\scriptscriptstyle B} = \mu\mathbb{1}_{d\times n}$, and whose dimensions are that of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively. Hence, using (5.3) and (5.4), we can now express (5.2) as

$$\mathbb{E}_\delta\left[\tilde{a}_{ij}\tilde{a}_{ik}\right] = \begin{cases} \Pi\left[a_{ij}, a_{ik}\right] & j \neq k \\ \pi\left[a_{ij}, a_{ik}\right] & j = k \end{cases}. \tag{5.5}$$

In addition to condensing the expectation, the functions $\Pi[\cdot,\cdot]$ and $\pi[\cdot,\cdot]$ have the added benefit of extending to higher dimensional expectations. For instance, suppose we need to take an expectation of the inner product of columns:

$$\begin{aligned}\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:k}\right] &= \mathbb{E}_\delta\left[\tilde{a}_{1j}\tilde{a}_{1k}+\tilde{a}_{2j}\tilde{a}_{2k}+\ldots+\tilde{a}_{mj}\tilde{a}_{mk}\right]\\ &= \mathbb{E}_\delta\left[\tilde{a}_{1j}\tilde{a}_{1k}\right]+\mathbb{E}_\delta\left[\tilde{a}_{2j}\tilde{a}_{2k}\right]+\ldots+\mathbb{E}_\delta\left[\tilde{a}_{mj}\tilde{a}_{mk}\right]\\ &= \sum_{i=1}^m \mathbb{E}_\delta\left[\tilde{a}_{ij}\tilde{a}_{ik}\right]\end{aligned}$$

which by (5.2) is

$$\begin{aligned}&= \sum_{i=1}^m\left[p^2a_{ij}a_{ik}+p(1-p)\left(a_{ij}\mu+\mu a_{ik}\right)+(1-p)^2\mu^2\right]\\ &= p^2\sum_{i=1}^m a_{ij}a_{ik}+p(1-p)\left(\sum_{i=1}^m a_{ij}\mu+\sum_{i=1}^m\mu a_{ik}\right)+(1-p)^2\sum_{i=1}^m\mu^2\\ &= p^2\boldsymbol{A}_{:j}^T\boldsymbol{A}_{:k}+p(1-p)\left(\boldsymbol{A}_{:j}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{A}_{:k}\right)+(1-p)^2\boldsymbol{\mu}^T\boldsymbol{\mu}\end{aligned}$$

when $j\neq k$, and

$$\begin{aligned}&= \sum_{i=1}^m\left[pa_{ij}a_{ik}+(1-p)\mu^2\right]\\ &= p\sum_{i=1}^m a_{ij}a_{ik}+(1-p)\sum_{i=1}^m\mu^2\\ &= p\boldsymbol{A}_{:j}^T\boldsymbol{A}_{:k}+(1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\end{aligned}$$

when $j=k$. But these are exactly $\Pi\left[\boldsymbol{A}_{:j}^T,\boldsymbol{A}_{:k}\right]$ and $\pi\left[\boldsymbol{A}_{:j}^T,\boldsymbol{A}_{:k}\right]$, respectively.

5.1. Proof of Theorem 3.3. Here we seek to show that $s_j(\boldsymbol{x})$ is an unbiased estimator for the gradient of our loss function, i.e. $\mathbb{E}[s_j(\boldsymbol{x})]=\nabla L(\boldsymbol{x})$. To do this we use the law of iterated expectation $\mathbb{E}[s_j(\boldsymbol{x})]=\mathbb{E}_j[\mathbb{E}_\delta[s_j(\boldsymbol{x})]]$, showing one at a time that for some function $\ell_j(\boldsymbol{x})$, $\mathbb{E}_\delta[s_j(\boldsymbol{x})]=\ell_j(\boldsymbol{x})$, while $\mathbb{E}_j[\ell_j(\boldsymbol{x})]=\nabla L(\boldsymbol{x})$.

Proof. Recall that $s_j(\boldsymbol{x})=\left(c_j(\boldsymbol{x})-d_j(\boldsymbol{x})\right)\boldsymbol{e}_j$ such that

$$c_j(\boldsymbol{x}):=\frac{1}{p^2}\tilde{\boldsymbol{A}}_{:j}^T\left(\tilde{\boldsymbol{A}}\boldsymbol{x}-p\boldsymbol{y}\right)-\frac{1-p}{p^2}\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:j}x_j,$$

and

$$\begin{aligned}d_j(\boldsymbol{x}):=&\frac{1-p}{p^2}\left[\left(\tilde{\boldsymbol{A}}_{:j}^T\boldsymbol{M}+\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}-(1-p)\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x}\right.\\ &\left.-p\boldsymbol{\mu}^T\boldsymbol{y}-\left(\tilde{\boldsymbol{A}}_{:j}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{:j}-\boldsymbol{\mu}^T\boldsymbol{\mu}\right)x_j\right],\end{aligned}$$

where $c_j(\boldsymbol{x})$ addresses the missing entry without mean shift, while $d_j(\boldsymbol{x})$ accounts for mean

shift. Now, observe that

$$\mathbb{E}_\delta\left[s_j(\boldsymbol{x})\right] = \left(\mathbb{E}_\delta\left[c_j(\boldsymbol{x})\right] - \mathbb{E}_\delta\left[d_j(\boldsymbol{x})\right]\right)\boldsymbol{e}_j, \tag{5.6}$$

where the expectations of c_j and d_j will be functions of inner products of columns of $\boldsymbol{A}$. Using (5.3) and (5.4) that notation can be simplified. In particular, we can write

$$\begin{aligned}
\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}\right] &= \mathbb{E}_\delta\left[\left(\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:1}, \tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:2}, ..., \tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:j}, ..., \tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:n}\right)\right] \\
&= \left(\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:1}\right], \mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:2}\right], ..., \mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:j}\right], ..., \mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:n}\right]\right) \\
&= \left(\Pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:1}\right], \Pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:2}\right], ..., \pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:j}\right], ..., \Pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:n}\right]\right) \\
&= \Pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}\right] - \Pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:j}\right]\boldsymbol{e}_j^T + \pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:j}\right]\boldsymbol{e}_j^T, \qquad (5.7)
\end{aligned}$$

Thus, using (5.7), we compute the expectation of c_j and d_j:

$$\begin{aligned}
\mathbb{E}_\delta\left[c_j(\boldsymbol{x})\right] &= \frac{1}{p^2}\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}\right]\boldsymbol{x} - \frac{1}{p}\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\right]\boldsymbol{y} - \frac{1-p}{p^2}\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\tilde{\boldsymbol{A}}_{:j}\right]x_j \\
&= \frac{1}{p^2}\left(\Pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}\right]\boldsymbol{x} - \Pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:j}\right]x_j + \pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:j}\right]x_j\right) \\
&\quad - \frac{1}{p}\left(p\boldsymbol{A}_{:j}^T + (1-p)\boldsymbol{\mu}^T\right)\boldsymbol{y} - \frac{1-p}{p^2}\left(\pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{A}_{:j}\right]\right)x_j \\
&= \frac{1}{p^2}\left(p^2\boldsymbol{A}_{:j}^T\boldsymbol{A} + p(1-p)\left(\boldsymbol{A}_{:j}^T\boldsymbol{M} + \boldsymbol{\mu}^T\boldsymbol{A}\right) + (1-p)^2\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x} \\
&\quad - \frac{1}{p^2}\left(p^2\boldsymbol{A}_{:j}^T\boldsymbol{A}_{:j} + p(1-p)\left(\boldsymbol{A}_{:j}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{:j}\right) + (1-p)^2\boldsymbol{\mu}^T\boldsymbol{\mu}\right)x_j \\
&\quad + \frac{1}{p^2}\left(p\boldsymbol{A}_{:j}^T\boldsymbol{A}_{:j} + (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)x_j - \frac{1}{p}\left(p\boldsymbol{A}_{:j}^T + (1-p)\boldsymbol{\mu}^T\right)\boldsymbol{y} \\
&\quad - \frac{1-p}{p^2}\left(p\boldsymbol{A}_{:j}^T\boldsymbol{A}_{:j} + (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)x_j \\
&= \left(\boldsymbol{A}_{:j}^T\boldsymbol{A} + \frac{1-p}{p}\left(\boldsymbol{A}_{:j}^T\boldsymbol{M} + \boldsymbol{\mu}^T\boldsymbol{A}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x} \\
&\quad - \left(\boldsymbol{A}_{:j}^T\boldsymbol{A} + \frac{1-p}{p}\left(\boldsymbol{A}_{:j}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{:j}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)x_j \\
&\quad + \left(\frac{1}{p}\boldsymbol{A}_{:j}^T\boldsymbol{A} + \frac{1-p}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)x_j - \left(\boldsymbol{A}_{:j}^T + \frac{1-p}{p}\boldsymbol{\mu}^T\right)\boldsymbol{y} \\
&\quad - \left(\frac{1-p}{p}\boldsymbol{A}_{:j}^T\boldsymbol{A} + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)x_j \\
&= \boldsymbol{A}_{:j}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}_{:j}^T\boldsymbol{y} + \left(\frac{1-p}{p}\left(\boldsymbol{A}_{:j}^T\boldsymbol{M} + \boldsymbol{\mu}^T\boldsymbol{A}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x} \qquad (5.8)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \left(\frac{1-p}{p}\left(\boldsymbol{A}_{:j}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{:j}\right) + \frac{2(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j \\
&\quad + \left(\frac{1-p}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j - \frac{1-p}{p}\boldsymbol{\mu}^T\boldsymbol{y} \\
\mathbb{E}_\delta\left[d_j(\boldsymbol{x})\right] &= \frac{1-p}{p^2}\left[\left(\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\boldsymbol{M}\right] + \mathbb{E}_\delta\left[\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}\right] - (1-p)\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x}\right. \\
&\quad \left. -p\boldsymbol{\mu}^T\boldsymbol{y} - \left(\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{:j}^T\boldsymbol{\mu}\right] + \mathbb{E}_\delta\left[\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{:j}\right] - \boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j\right] \\
&= \frac{1-p}{p^2}\left[\left(\pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{M}\right] + \pi\left[\boldsymbol{\mu}^T, \boldsymbol{A}\right] - (1-p)\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x}\right. \\
&\quad \left. -p\boldsymbol{\mu}^T\boldsymbol{y} - \left(\pi\left[\boldsymbol{A}_{:j}^T, \boldsymbol{\mu}\right] + \pi\left[\boldsymbol{\mu}^T, \boldsymbol{A}_{:j}\right] - \boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j\right] \\
&= \frac{1-p}{p^2}\left[\left(p\boldsymbol{A}_{:j}^T\boldsymbol{M} + p\boldsymbol{\mu}^T\boldsymbol{A} + 2(1-p)\boldsymbol{\mu}^T\boldsymbol{M} - (1-p)\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x}\right. \\
&\quad \left. -p\boldsymbol{\mu}^T\boldsymbol{y} - \left(p\boldsymbol{A}_{:j}^T\boldsymbol{\mu} + p\boldsymbol{\mu}^T\boldsymbol{A}_{:j} + 2(1-p)\boldsymbol{\mu}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j\right] \\
&= \left(\frac{1-p}{p}\boldsymbol{A}_{:j}^T\boldsymbol{M} + \frac{1-p}{p}\boldsymbol{\mu}^T\boldsymbol{A} + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x} - \frac{1-p}{p}\boldsymbol{\mu}^T\boldsymbol{y} \\
&\quad - \left(\frac{1-p}{p}\boldsymbol{A}_{:j}^T\boldsymbol{\mu} - \frac{1-p}{p}\boldsymbol{\mu}^T\boldsymbol{A}_{:j} - \frac{(1-p)(1-2p)}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j \\
&= \left(\frac{1-p}{p}\left(\boldsymbol{A}_{:j}^T\boldsymbol{M} + \boldsymbol{\mu}^T\boldsymbol{A}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{M}\right)\boldsymbol{x} \qquad (5.9) \\
&\quad - \left(\frac{1-p}{p}\left(\boldsymbol{A}_{:j}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{:j}\right) + \frac{2(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j \\
&\quad + \left(\frac{1-p}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) x_j - \frac{1-p}{p}\boldsymbol{\mu}^T\boldsymbol{y}
\end{aligned}
$$

Plugging (5.8) and (5.9) into (5.6), we get

$$
\mathbb{E}_\delta\left[s_j(\boldsymbol{x})\right] = \left(\mathbb{E}_\delta\left[c_j(\boldsymbol{x})\right] - \mathbb{E}_\delta\left[d_j(\boldsymbol{x})\right]\right)\boldsymbol{e}_j = \left(\boldsymbol{A}_{:j}^T\left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right)\right)\boldsymbol{e}_j, \tag{5.10}
$$

for which we refer to the right-hand side expression as "$\ell_j(\boldsymbol{x})$". Recalling our loss function $L(\boldsymbol{x}) = \frac{1}{2n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2$, it can be shown after some expansions that

$$
\frac{\partial L(\boldsymbol{x})}{\partial x_j} = \frac{1}{n}\boldsymbol{A}_{:j}^T\left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right).
$$

Thus, when we go to compute the column-wise expectation of $\ell_j(\boldsymbol{x})$ (assuming columns of $\boldsymbol{A}$ are selected uniformly at random) we get

$$
\mathbb{E}_j[\ell_j(\boldsymbol{x})] = \sum_{j=1}^{n}\frac{1}{n}\ell_j(\boldsymbol{x}) = \sum_{\mathrm{j}=1}^{n}\frac{1}{n}\left(\boldsymbol{A}_{:j}^T\left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right)\right)\boldsymbol{e}_j = \sum_{\mathrm{j}=1}^{n}\frac{\partial L(\boldsymbol{x})}{\partial x_j}\boldsymbol{e}_j. \tag{5.11}
$$

Finally, combining (5.10) and (5.11) gives us the desired gradient,

$$
\mathbb{E}\left[s_j(\boldsymbol{x})\right] = \mathbb{E}_j\left[\mathbb{E}_\delta\left[s_j(\boldsymbol{x})\right]\right] = \mathbb{E}_j\left[\ell_j(\boldsymbol{x})\right] = \nabla L(\boldsymbol{x}).
$$

■

5.2. Proof of Theorem 3.4. Similar to the last proof, here we show that $t_i(\boldsymbol{x})$ is an unbiased estimator for the gradient of the loss function, $\mathbb{E}\left[t_i(\boldsymbol{x})\right] = \nabla F(\boldsymbol{x})$. Again, we use an iterated expectation $\mathbb{E}\left[t_i(\boldsymbol{x})\right] = \mathbb{E}_i\left[\mathbb{E}_\delta\left[t_i(\boldsymbol{x})\right]\right]$, showing first that for the function $f_i(\boldsymbol{x})$, $\mathbb{E}_\delta\left[t_i(\boldsymbol{x})\right] = \nabla f_i(\boldsymbol{x})$ holds, then finally that $\mathbb{E}_i\left[\nabla f_i(\boldsymbol{x})\right] = \nabla F(\boldsymbol{x})$.

Proof. Recall that $t_i(\boldsymbol{x}) = g_i(\boldsymbol{x}) - h_i(\boldsymbol{x})$ such that

$$g_i(\boldsymbol{x}) := \frac{1}{p^2}\tilde{\boldsymbol{A}}_{i:}^T\left(\tilde{\boldsymbol{A}}_{i:}^T\boldsymbol{x} - py_i\right) - \frac{1-p}{p^2}\mathrm{diag}\left(\tilde{\boldsymbol{A}}_{i:}^T\tilde{\boldsymbol{A}}_{i:}\right)\boldsymbol{x},$$

and

$$\begin{aligned} h_i(\boldsymbol{x}) := \frac{1-p}{p^2}\Big[&\left(\tilde{\boldsymbol{A}}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{i:} - (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\ &-p\boldsymbol{\mu}^T y_i - \mathrm{diag}\left(\tilde{\boldsymbol{A}}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{i:} - \boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x}\Big], \end{aligned}$$

where $g_i(\boldsymbol{x})$ plays the analogous role to $c_j(\boldsymbol{x})$ in the previous proof (missing entry without mean shift), and $h_i(\boldsymbol{x})$ the role of $d_j(\boldsymbol{x})$ (mean shift). Now, observe that

$$\mathbb{E}_\delta\left[t_i(\boldsymbol{x})\right] = \mathbb{E}_\delta\left[g_i(\boldsymbol{x})\right] - \mathbb{E}_\delta\left[h_i(\boldsymbol{x})\right], \tag{5.12}$$

where the expectations of g_i and h_i will be functions of inner products of columns of $\boldsymbol{A}$. Using (5.3) and (5.4) that notation can be simplified. In particular, we can write

$$\begin{aligned} \mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{i:}^T\tilde{\boldsymbol{A}}_{i:}\right] &= \mathbb{E}_\delta\left[\begin{pmatrix} \tilde{a}_{i1}\tilde{a}_{i1}, & \tilde{a}_{i1}\tilde{a}_{i2}, & \dots & \tilde{a}_{i1}\tilde{a}_{in} \\ \tilde{a}_{i2}\tilde{a}_{i1}, & \tilde{a}_{i2}\tilde{a}_{i2}, & \dots & \tilde{a}_{i2}\tilde{a}_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{in}\tilde{a}_{i1}, & \tilde{a}_{in}\tilde{a}_{i2}, & \dots & \tilde{a}_{in}\tilde{a}_{in} \end{pmatrix}\right] \\ &= \begin{pmatrix} \mathbb{E}_\delta\left[\tilde{a}_{i1}\tilde{a}_{i1}\right], & \mathbb{E}_\delta\left[\tilde{a}_{i1}\tilde{a}_{i2}\right], & \dots & \mathbb{E}_\delta\left[\tilde{a}_{i1}\tilde{a}_{in}\right] \\ \mathbb{E}_\delta\left[\tilde{a}_{i2}\tilde{a}_{i1}\right], & \mathbb{E}_\delta\left[\tilde{a}_{i2}\tilde{a}_{i2}\right], & \dots & \mathbb{E}_\delta\left[\tilde{a}_{i2}\tilde{a}_{in}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}_\delta\left[\tilde{a}_{in}\tilde{a}_{i1}\right], & \mathbb{E}_\delta\left[\tilde{a}_{in}\tilde{a}_{i2}\right], & \dots & \mathbb{E}_\delta\left[\tilde{a}_{in}\tilde{a}_{in}\right] \end{pmatrix} \end{aligned} \tag{5.13}$$

$$= \begin{pmatrix} \pi\left[a_{i1}, a_{i1}\right], & \Pi\left[a_{i1}, a_{i2}\right], & \dots & \Pi\left[a_{i1}, a_{in}\right] \\ \Pi\left[a_{i2}, a_{i1}\right], & \pi\left[a_{i2}, a_{i2}\right], & \dots & \Pi\left[a_{i2}, a_{in}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \Pi\left[a_{in}, a_{i1}\right], & \Pi\left[a_{in}, a_{i2}\right], & \dots & \pi\left[a_{in}, a_{in}\right] \end{pmatrix}$$

$$= \Pi\left[\boldsymbol{A}_{i:}^T, \boldsymbol{A}_{i:}\right] - \mathrm{diag}\left(\Pi\left[\boldsymbol{A}_{i:}^T, \boldsymbol{A}_{i:}\right]\right) + \mathrm{diag}\left(\pi\left[\boldsymbol{A}_{i:}^T, \boldsymbol{A}_{i:}\right]\right).$$

Thus, taking each expectation separately and using (5.13), we get:

$$
\begin{aligned}
\mathbb{E}_\delta\left[g_i(\boldsymbol{x})\right] &= \frac{1}{p^2}\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{i:}^T\tilde{\boldsymbol{A}}_{i:}\right]\boldsymbol{x} - \frac{1}{p}\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{i:}^T\right]y_i - \frac{1-p}{p^2}\mathrm{diag}\left(\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{i:}^T\tilde{\boldsymbol{A}}_{i:}\right]\right)\boldsymbol{x} \\
&= \frac{1}{p^2}\left(\Pi\left[\boldsymbol{A}_{i:}^T,\boldsymbol{A}_{i:}\right] - \mathrm{diag}\left(\Pi\left[\boldsymbol{A}_{i:}^T,\boldsymbol{A}_{i:}\right]\right) + \mathrm{diag}\left(\pi\left[\boldsymbol{A}_{i:}^T,\boldsymbol{A}_{i:}\right]\right)\right)\boldsymbol{x} \\
&\quad - \frac{1}{p}\left(p\boldsymbol{A}_{i:}^T + (1-p)\boldsymbol{\mu}^T\right)y_i - \frac{1-p}{p^2}\left(\mathrm{diag}\left(\pi\left[\boldsymbol{A}_{i:}^T,\boldsymbol{A}_{i:}\right]\right)\right)\boldsymbol{x} \\
&= \frac{1}{p^2}\left(p^2\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + p(1-p)\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + (1-p)^2\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad - \frac{1}{p^2}\mathrm{diag}\left(p^2\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + p(1-p)\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + (1-p)^2\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad + \frac{1}{p^2}\mathrm{diag}\left(p\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} - \frac{1}{p}\left(p\boldsymbol{A}_{i:}^T + (1-p)\boldsymbol{\mu}^T\right)y_i \\
&\quad - \frac{1-p}{p^2}\mathrm{diag}\left(p\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&= \left(\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + \frac{1-p}{p}\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad - \mathrm{diag}\left(\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + \frac{1-p}{p}\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad + \mathrm{diag}\left(\frac{1}{p}\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + \frac{1-p}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} - \left(\boldsymbol{A}_{i:}^T + \frac{1-p}{p}\boldsymbol{\mu}^T\right)y_i \\
&\quad - \mathrm{diag}\left(\frac{1-p}{p}\boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:} + \frac{(1-p)^2}{p^2}\ \boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&= \boldsymbol{A}_{i:}^T\boldsymbol{A}_{i:}\boldsymbol{x} - \boldsymbol{A}_{i:}^Ty_i + \left(\frac{1-p}{p}\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad - \mathrm{diag}\left(\frac{1-p}{p}\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + \frac{2(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad + \mathrm{diag}\left(\frac{1-p}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} - \frac{1-p}{p}\boldsymbol{\mu}^Ty_i,
\end{aligned}
\tag{5.14}
$$

and

$$
\begin{aligned}
\mathbb{E}_\delta\left[h_i(\boldsymbol{x})\right] &= \frac{1-p}{p^2}\left[\left(\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{i:}^T\boldsymbol{\mu}\right] + \mathbb{E}_\delta\left[\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{i:}\right] - (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x}\right. \\
&\qquad \left. -p\boldsymbol{\mu}^Ty_i - \mathrm{diag}\left(\mathbb{E}_\delta\left[\tilde{\boldsymbol{A}}_{i:}^T\boldsymbol{\mu}\right] + \mathbb{E}_\delta\left[\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{i:}\right] - \boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x}\right] \\
&= \frac{1-p}{p^2}\left[\left(\pi\left[\tilde{\boldsymbol{A}}_{i:}^T\boldsymbol{\mu}\right] + \pi\left[\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{i:}\right] - (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x}\right. \\
&\qquad \left. -p\boldsymbol{\mu}^Ty_i - \mathrm{diag}\left(\pi\left[\tilde{\boldsymbol{A}}_{i:}^T\boldsymbol{\mu}\right] + \pi\left[\boldsymbol{\mu}^T\tilde{\boldsymbol{A}}_{i:}\right] - \boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x}\right] \\
&= \frac{1-p}{p^2}\left[\left(p\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + p\boldsymbol{\mu}^T\boldsymbol{A}_{i:} + 2(1-p)\boldsymbol{\mu}^T\boldsymbol{\mu} - (1-p)\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x}\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad -p\boldsymbol{\mu}^T y_i - \operatorname{diag}\left(p\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + p\boldsymbol{\mu}^T\boldsymbol{A}_{i:} + 2(1-p)\boldsymbol{\mu}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x}\Big] \\
&= \left(\frac{1-p}{p}\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \frac{1-p}{p}\boldsymbol{\mu}^T\boldsymbol{A}_{i:} + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} - \frac{1-p}{p}\boldsymbol{\mu}^T y_i \\
&\quad - \operatorname{diag}\left(\frac{1-p}{p}\boldsymbol{A}_{i:}^T\boldsymbol{\mu} - \frac{1-p}{p}\boldsymbol{\mu}^T\boldsymbol{A}_{i:} - \frac{(1-p)(1-2p)}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&= \left(\frac{1-p}{p}\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + \frac{(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad - \operatorname{diag}\left(\frac{1-p}{p}\left(\boldsymbol{A}_{i:}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}_{i:}\right) + \frac{2(1-p)^2}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} \\
&\quad + \operatorname{diag}\left(\frac{1-p}{p^2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\boldsymbol{x} - \frac{1-p}{p}\boldsymbol{\mu}^T y_i.
\end{aligned} \tag{5.15}
$$

Plugging (5.14) and (5.15) into (5.12), we get

$$
\mathbb{E}_\delta\left[t_i(\boldsymbol{x})\right] = \mathbb{E}_\delta\left[g_i(\boldsymbol{x})\right] - \mathbb{E}_\delta\left[h_i(\boldsymbol{x})\right] = \boldsymbol{A}_{i:}^T\left(\boldsymbol{A}_{i:}\boldsymbol{x} - y_i\right). \tag{5.16}
$$

Now, recall our objective function $F(\boldsymbol{x}) = \frac{1}{2m}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2$, which can be rewritten as $F(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\boldsymbol{x})$, where $f_i(\boldsymbol{x}) = \frac{1}{2}\left(\boldsymbol{A}_{i:}\boldsymbol{x} - y_i\right)^2$. After some expansions, it is not too difficult to show that

$$
\nabla f_i(\boldsymbol{x}) = \boldsymbol{A}_{i:}^T\left(\boldsymbol{A}_{i:}\boldsymbol{x} - y_i\right), \tag{5.17}
$$

and notice this is exactly the right-hand side of (5.16). Thus, assuming rows i are uniformly selected from $[m]$, we take the row-wise expectation of $\nabla f_i(\boldsymbol{x})$ to get

$$
\mathbb{E}_i\left[\nabla f_i(\boldsymbol{x})\right] = \sum_{i=1}^m \frac{1}{m}\nabla f_i(\boldsymbol{x}) = \nabla\left[\frac{1}{m}\sum_{i=1}^m f_i(\boldsymbol{x})\right] = \nabla F(\boldsymbol{x}). \tag{5.18}
$$

Finally, combining (5.16), (5.17), and (5.18), we get

$$
\mathbb{E}\left[t_i(\boldsymbol{x})\right] = \mathbb{E}_i\left[\mathbb{E}_\delta\left[t_i(\boldsymbol{x})\right]\right] = \mathbb{E}_i\left[\nabla f_i(\boldsymbol{x})\right] = \nabla F(\boldsymbol{x}).
$$

the desired gradient. ∎

5.3. Proof of Theorem 3.5. In this proof, we first define the μ, then we consider the different cases, such as 0-imputation and μ-imputation for the matrix A. By calculating the expectation for both cases, $\mathbb{E}_\delta\|\boldsymbol{A} - \tilde{\boldsymbol{A}}_0\|_F^2 \geq \mathbb{E}_\delta\|\boldsymbol{A} - \tilde{\boldsymbol{A}}_\mu\|_F^2$ holds.

Proof. Given $\tilde{\boldsymbol{A}}_0$ is matrix $\boldsymbol{A}$ with missing entries $\tilde{a}_{0,ij}$ in which 0 is imputed and $\tilde{\boldsymbol{A}}_\mu$ is matrix $\boldsymbol{A}$ with entries $\tilde{a}_{\mu,ij}$ in which the mean value of $\boldsymbol{A}$, a fixed $\mu = \frac{1}{mn}\sum_{ij} a_{ij}$, is imputed, we want to find $\mathbb{E}_\delta\left[\|\boldsymbol{A} - \tilde{\boldsymbol{A}}_0\|_F^2\right]$ and $\mathbb{E}_\delta\left[\|\boldsymbol{A} - \tilde{\boldsymbol{A}}_\mu\|_F^2\right]$ and compare them.

We know,

$$
\tilde{a}_{0,ij} = \begin{cases} 0 & w.p \quad 1-p \\ a_{ij} & w.p \quad p \end{cases} \quad \Rightarrow \quad \tilde{a}_{0,ij}^2 = \begin{cases} 0 & w.p \quad 1-p \\ a_{ij}^2 & w.p \quad p \end{cases}
$$

$$\tilde{a}_{\mu,ij} = \begin{cases} \mu & w.p \quad 1-p \\ a_{ij} & w.p \quad p \end{cases} \quad \Rightarrow \quad \tilde{a}^2_{\mu,ij} = \begin{cases} \mu^2 & w.p \quad 1-p \\ a^2_{ij} & w.p \quad p \end{cases}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_\delta \left[\|\boldsymbol{A} - \tilde{\boldsymbol{A}}_0\|_F^2 \right] &= \mathbb{E}_\delta \left[\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - \tilde{a}_{0,ij})^2 \right] \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_\delta \left[\left(a_{ij}^2 - 2a_{ij}\tilde{a}_{0,ij} + \tilde{a}_{0,ij}^2 \right) \right] \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_\delta \left[a_{ij}^2 \right] + \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_{\boldsymbol{A},\delta} \left[\tilde{a}_{0,ij}^2 \right] - \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_\delta \left[2a_{ij}\tilde{a}_{0,ij} \right] \\
&= \|\boldsymbol{A}\|_F^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} p a_{ij}^2 - \sum_{i=1}^{m} \sum_{j=1}^{n} 2p a_{ij}^2 \\
&= \|\boldsymbol{A}\|_F^2 - \sum_{i=1}^{m} \sum_{j=1}^{n} p a_{ij}^2 \\
&= \|\boldsymbol{A}\|_F^2 - p\|\boldsymbol{A}\|_F^2 = (1-p)\|\boldsymbol{A}\|_F^2
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}_\delta \left[\|\boldsymbol{A} - \tilde{\boldsymbol{A}}_\mu\|_F^2 \right] &= \mathbb{E}_\delta \left[\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - \tilde{a}_{\mu,ij})^2 \right] \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_\delta \left[\left(a_{ij}^2 - 2a_{ij}\tilde{a}_{\mu,ij} + \tilde{a}_{\mu,ij}^2 \right) \right] \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_\delta \left[a_{ij}^2 \right] + \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_\delta \left[\tilde{a}_{\mu,ij}^2 \right] - \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}_\delta \left[2a_{ij}\tilde{a}_{\mu,ij} \right] \\
&= \|\boldsymbol{A}\|_F^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} \left(p a_{ij}^2 + \mu^2(1-p) \right) - \sum_{i=1}^{m} \sum_{j=1}^{n} 2a_{ij} \left(p a_{ij} + \mu(1-p) \right) \\
&= \|\boldsymbol{A}\|_F^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} p a_{ij}^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} \mu^2(1-p) - 2\sum_{i=1}^{m} \sum_{j=1}^{n} p a_{ij}^2 - 2\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}\mu(1-p) \\
&= \|\boldsymbol{A}\|_F^2 - \sum_{i=1}^{m} \sum_{j=1}^{n} p a_{ij}^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} \mu^2(1-p) - 2\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}\mu(1-p) \\
&= \|\boldsymbol{A}\|_F^2 - \sum_{i=1}^{m} \sum_{j=1}^{n} p a_{ij}^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} \mu(1-p) \left(\mu - 2a_{ij} \right) \\
&= (1-p)\|\boldsymbol{A}\|_F^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} \mu(1-p) \left(\mu - 2a_{ij} \right)
\end{aligned}$$

For the last term we have,

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{n}\mu(1-p)\left(\mu-2a_{ij}\right) &= \mu(1-p)\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mu-2a_{ij}\right)\\ &= \mu(1-p)\left(mn\mu-\sum_{i=1}^{m}\sum_{j=1}^{n}2a_{ij}\right)\\ &= \mu(1-p)\left(mn\mu-2mn\mu\right)\\ &= -mn\mu^2(1-p),\end{aligned}$$

which is non-positive. Thus, $\mathbb{E}_\delta[\|\boldsymbol{A}-\tilde{\boldsymbol{A}}_0\|_F^2] \geq \mathbb{E}_\delta[\|\boldsymbol{A}-\tilde{\boldsymbol{A}}_\mu\|_F^2]$

REFERENCES

[1] R. Jiang, T. Sun, D. Song, and J. J. Li, *Statistics or biology: the zero-inflation controversy about scrna-seq data*, Genome biology, 23 (2022), pp. 1–24.

[2] D. Leventhal and A. S. Lewis, *Randomized methods for linear constraints: Convergence rates and conditioning*, 2008.

[3] A. Ma and D. Needell, *Stochastic gradient descent for linear systems with missing data*.

[4] A. Ma, D. Needell, and A. Ramdas, *Convergence properties of the randomized extended gauss–seidel and kaczmarz methods*, SIAM Journal on Matrix Analysis and Applications, 36 (2015), pp. 1590–1604.

[5] S. Mukhopadhyay, *Stochastic gradient descent for linear systems with sequential matrix entry accumulation*, Signal Processing, 171 (2020), p. 107494.

[6] S. Mukhopadhyay and A. Mukherjee, *Imdlms: An imputation based lms algorithm for linear system identification with missing input data*, IEEE Transactions on Signal Processing, 68 (2020), pp. 2370–2385.

[7] D. Needell, *Randomized kaczmarz solver for noisy linear systems*, BIT Numerical Mathematics, 50 (2010), pp. 395–403.

[8] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization, 22 (2012), pp. 341–362.

[9] M. S. Rahim, A. A. Imran, and T. Ahmed, *Mining the productivity data of garment industry*, International Journal of Business Intelligence and Data Mining, 1 (2021), p. 1.