S1 Appendix: Proof of Theorems

In this section, we will identify specific f-divergence functions that can be used for the generative adversarial imputation network, and provide mathematical proof for the Pseudo-code of sc-fGAIN. We adopt some notations and assumptions in Yoon *et al*'s work[1], and assume that $\mathbf{X}$ is independent of $\mathbf{M}$, where $p(\mathbf{x}, \mathbf{m}, \mathbf{h})$ denotes the joint distribution for the random variables $(\hat{\mathbf{X}}, \mathbf{M}, \mathbf{H})$, and $\hat{p}(\mathbf{x}), p(\mathbf{m}), p(\mathbf{h})$ are corresponding marginal distributions.

Proof of Theorem 1

We will use the following equation,

$$\frac{\partial}{\partial g_f(S_\phi)_i} f^*(g_f(S_\phi))_i = \frac{p(\mathbf{x}, \mathbf{h}, m_i = 1)}{p(\mathbf{x}, \mathbf{h}, m_i = 0)}. \tag{1}$$

and Table 2 to prove the Theorem 1.
Set $p_i = p(\mathbf{x}, \mathbf{h}, m_i = 1)$, $q_i = p(\mathbf{x}, \mathbf{h}, m_i = 0)$, apply the Sigmoid function on the output of the discriminator network $S_\phi(x, h)$, that is $D_i = \frac{1}{1+\exp(-S_\phi(x,h)_i)}$.

Below, we will derive the optimal discriminator D_i^* given different f-divergence functions:

- f = **Cross-entropy (CE):**

  $$\frac{\partial}{\partial g_f(S_\phi)_i} f^*(g_f(S_\phi))_i = \frac{p_i}{q_i} = \frac{\exp(-\log(1 + \exp(-S_\phi(x,h)_i))}{1 - \exp(-\log(1 + \exp(-S_\phi(x,h)_i))} = \frac{\frac{1}{1+\exp(-S_\phi(x,h)_i)}}{1 - \frac{1}{1+\exp(-S_\phi(x,h)_i)}}$$

  The above equation can be simplified as $\frac{p_i}{q_i} = \frac{D_i}{1-D_i}$.
  Finally, we get the optimal discriminator $D_i^* = \frac{p_i}{p_i+q_i}$.

- f = **Kullback-Leibler(FKL)**
  $\frac{\partial}{\partial g_f(S_\phi)_i} f^*(g_f(S_\phi))_i = \frac{p_i}{q_i} = \exp(-\log(\frac{1}{S_\phi(x,h)_i} - 1) - 1)$
  The above equation can be simplified as $\frac{p_i}{q_i} = \frac{D_i}{e(1-D_i)}$.
  Finally, we get the optimal discriminator $D_i^* = \frac{p_i e}{p_i e + q_i}$.

- f = **Reverse Kullback-Leibler (RKL)**
  $\frac{\partial}{\partial g_f(S_\phi)_i} f^*(g_f(S_\phi))_i = \frac{p_i}{q_i} = \frac{1}{-\exp(-S_\phi(x,h)_i)}$
  The above equation can be simplified as $\frac{p_i}{q_i} = \frac{D_i}{1-D_i}$.
  Finally, we get the optimal discriminator $D_i^* = \frac{p_i}{p_i+q_i}$

- f = **Jensen-Shannon (JS)**
  $\frac{\partial}{\partial g_f(S_\phi)_i} f^*(g_f(S_\phi))_i = \frac{p_i}{q_i} = \frac{\frac{2}{1+\exp(-S_\phi(x,h)_i)}}{2 - \frac{2}{1+\exp(-S_\phi(x,h)_i)}}$
  The above equation can be simplified as $\frac{p_i}{q_i} = \frac{D_i}{1-D_i}$.
  Finally, we get the optimal discriminator $D_i^* = \frac{p_i}{p_i+q_i}$.

- f = **Pearson χ^2 (PC)**
  $\frac{\partial}{\partial g_f(S_\phi)_i} f^*(g_f(S_\phi))_i = \frac{p}{q} = \frac{1}{2}S_\phi(x,h)_i + 1$
  Finally, we get the optimal discriminator $D_i^* = \frac{exp(2(p_i-q_i)/q_i)}{1+exp(2(p_i-q_i)/q_i)}$.

Proof of Theorem 2

Given the optimal discriminator D_i^*, for each f-divergence function, the objective function $\mathcal{L}_{G,f}(D^*)$ can be expressed as:

- f = **CE**

$$\begin{aligned}\mathcal{L}_{G,f}(D^*) &= \sum_i^d \int_\chi \int_\mathcal{H} (p_i \log(D_i^*) + q_i \log(1-D_i^*))\, dxdh \\ &= \sum_i^d \int_\chi \int_\mathcal{H} \left(p_i \log(\frac{p_i}{p_i+q_i}) + q_i \log(\frac{q_i}{p_i+q_i})\right) dxdh\end{aligned}$$

It can also be written as:

$$\begin{aligned}\mathcal{L}_{G,f}(D^*) &= \sum_{t\in\{0,1\}} \sum_{i=1}^d \int_\chi \int_\mathcal{H} p(\mathbf{x},\mathbf{h},m_i=t) \log(\frac{p(\mathbf{x},\mathbf{h},m_i=t)}{p(\mathbf{x},\mathbf{h},m_i=0)+p(\mathbf{x},\mathbf{h},m_i=1)})dhdx \\ &= \sum_{t\in\{0,1\}} \sum_{i=1}^d \int_\chi \int_\mathcal{H} p(\mathbf{x},\mathbf{h},m_i=t) \log(\frac{p(\mathbf{x},m_i=t|\mathbf{h})p(m_i=t|h)}{p(\mathbf{x}|\mathbf{h})p(m_i=t|h)})dhdx \\ &= \sum_{t\in\{0,1\}} \sum_{i=1}^d \int_\chi \int_\mathcal{H} p(\mathbf{h},m_i=t)p(\mathbf{x}|\mathbf{h},m_i=t) \log(\frac{p(\mathbf{x}|\mathbf{h},m_i=t)}{p(\mathbf{x}|\mathbf{h})})dhdx \\ &+ \sum_{t\in\{0,1\}} \sum_{i=1}^d \int_\mathcal{H} p(\mathbf{h},m_i=t)\log(p(m_i=t|\mathbf{h}))dh \\ &= \sum_{t\in\{0,1\}} \sum_{i=1}^d \int_\mathcal{H} p(\mathbf{h},m_i=t)D_{KL}(p(x|\mathbf{h},m_i=t)||p(x|\mathbf{h}))dh \\ &+ \sum_{t\in\{0,1\}} \sum_{i=1}^d \int_\mathcal{H} p(\mathbf{h},m_i=t)\log(p(m_i=t|\mathbf{h}))dh\end{aligned}$$

Since KL divergence is non-negative, so the loss function $\mathcal{L}_{G,f}(D^*)$ is minimized if and only if $\hat{p}(\mathbf{x}|\mathbf{h},m_i=t) = \hat{p}(\mathbf{x}|\mathbf{h})$ for any $i \in \{1,...,d\}$.

- f = **FKL**

$$
\begin{aligned}
\mathcal{L}_{G,f}(D^*) &= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \left(-p(\frac{1}{D^*} - 1) + q(1 + \log(\frac{1}{D^*} - 1)) \right) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \left(p \log \frac{p}{q} \right) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \left(p \log p + q \log q - (p+q) \log q \right) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \sum_{t \in \{0,1\}} p(\mathbf{x}, \mathbf{h}, m_i = t) \log \frac{P(x|h, m_i = t)}{P(x|h)} dxdh \\
&+ \sum_{t \in \{0,1\}} \sum_{i=1}^{d} \int_{\mathcal{H}} p(\mathbf{h}, m_i = t) \log(p(m_i = t|\mathbf{h}))dh \\
&- \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} p(x,h) \log p(m_i = 0|x,h) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \sum_{t \in \{0,1\}} p(\mathbf{x}|\mathbf{h}, m_i = t) p(\mathbf{h}, m_i = t) \log \frac{P(x|h, m_i = t)}{P(x|h)} dxdh \\
&+ \sum_{t \in \{0,1\}} \sum_{i=1}^{d} \int_{\mathcal{H}} p(\mathbf{h}, m_i = t) \log(p(m_i = t|\mathbf{h}))dh \\
&- \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} p(h) p(x|h) \log \frac{p(x, m_i = 0|h)}{p(x|h)} dxdh \\
&= \sum_{t \in \{0,1\}} \sum_{i=1}^{d} \int_{\mathcal{H}} p(\mathbf{h}, m_i = t) D_{KL}(p(x|\mathbf{h}, m_i = t)||p(x|\mathbf{h})) \\
&+ p(\mathbf{h}, m_i = t) \log(p_m(m_i = t|\mathbf{h}))dh \\
&+ \sum_{i=1}^{d} \int_{\mathcal{H}} p(h) D_{KL}(p(x|h)||p(x|m_i = 0, h)) - p(h) \log p(m_i = 0|h) dh
\end{aligned}
$$

Similar to CE's proof, since KL divergence is non-negative, so $\mathcal{L}_{G,f}(D^*)$ is minimized if and only if $p(x|\mathbf{h}, m_i = t) = p(x|\mathbf{h})$ for the 1st and 2nd term.

- f **= RKL**

$$
\begin{aligned}
\mathcal{L}_{G,f}(D^*) &= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \left(p \log(\frac{D^*}{1-D^*}) - \frac{q}{e} \frac{D^*}{1-D^*} \right) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \left(p \log(\frac{pe}{q}) - \frac{q}{e} \log(\frac{pe}{q}) \right) dxdh = \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \left(q \log \frac{q}{p} \right) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \left(q \log q + p \log p - (p+q) \log(p) \right) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \sum_{t \in \{0,1\}} p(\mathbf{x}, \mathbf{h}, m_i = t) \log \frac{P(x|h, m_i = t)}{P(x|h)} dxdh \\
&+ \sum_{t \in \{0,1\}} \sum_{i=1}^{d} \int_{\mathcal{H}} p(\mathbf{h}, m_i = t) \log(p(m_i = t|\mathbf{h})) dh \\
&- \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} p(x,h) \log p(m_i = 1|x,h) dxdh \\
&= \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} \sum_{t \in \{0,1\}} p(\mathbf{x}|\mathbf{h}, m_i = t) p(\mathbf{h}, m_i = t) \log \frac{P(x|h, m_i = t)}{P(x|h)} dxdh \\
&+ \sum_{t \in \{0,1\}} \sum_{i=1}^{d} \int_{\mathcal{H}} p(\mathbf{h}, m_i = t) \log(p(m_i = t|\mathbf{h})) dh \\
&- \sum_{i=1}^{d} \int_{\chi} \int_{\mathcal{H}} p(h) p(x|h) \log \frac{p(x, m_i = 1|h)}{p(x|h)} dxdh \\
&= \sum_{t \in \{0,1\}} \sum_{i=1}^{d} \int_{\mathcal{H}} p(\mathbf{h}, m_i = t) D_{KL}(p(x|\mathbf{h}, m_i = t) || p(x|\mathbf{h})) \\
&+ p(\mathbf{h}, m_i = t) \log(p_m(m_i = t|\mathbf{h})) dh \\
&+ \sum_{i=1}^{d} \int_{\mathcal{H}} p(h) D_{KL}(p(x|h) || p(x|m_i = 1, h)) - p(h) \log p(m_i = 1|h) dh
\end{aligned}
$$

Similarly to FKL, since KL divergence is non-negative, so, to get $\mathcal{L}_{G,f}(D^*)$ minimized if and only if $p(x|\mathbf{h}, m_i = t) = p(x|\mathbf{h})$.

- f = **JS**

$$
\begin{aligned}
\mathcal{L}_{G,f}(D^*) &= \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\mathcal{H}} \log(2D^*) + q\log(2-2D^*) \\
&= \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\mathcal{H}} p\log(\frac{2p}{p+q}) + q\log(\frac{2q}{p+q}) \\
&= \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\mathcal{H}} p(\mathbf{x},\mathbf{h},m_i=t)\log(\frac{p(\mathbf{x},\mathbf{h},m_i=t)}{p(\mathbf{x},\mathbf{h},m_i=0)+p(\mathbf{x},\mathbf{h},m_i=1)})dhdx \\
&+ \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\mathcal{H}} p(\mathbf{x},\mathbf{h},m_i=t)\log(2)dxdh \\
&= \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\mathcal{H}} p(\mathbf{x},\mathbf{h},m_i=t)\log(\frac{p(\mathbf{x},m_i=t|\mathbf{h})p_m(m_i=t|h)}{p(\mathbf{x}|\mathbf{h})p_m(m_i=t|h)})dhdx + C \\
&= \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\mathcal{H}} p(\mathbf{h},m_i=t)p(\mathbf{x}|\mathbf{h},m_i=t)\log(\frac{p(\mathbf{x}|\mathbf{h},m_i=t)}{p(\mathbf{x}|\mathbf{h}))}dhdx \\
&+ \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{H}} p(\mathbf{h},m_i=t)p(\mathbf{x}|\mathbf{h},m_i=t)\log(p_m(m_i=t|\mathbf{h}))dh + C \\
&= \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{H}} p(\mathbf{h},m_i=t)D_{KL}(p(\cdot|\mathbf{h},m_i=t)||p(\cdot|\mathbf{h}))dh \\
&+ \sum_{t\in\{0,1\}}\sum_{i=1}^{d}\int_{\mathcal{H}} p(\mathbf{h},m_i=t)\log(p_m(m_i=t|\mathbf{h}))dh + C
\end{aligned}
$$

C denotes a constant, therefore, the loss function is minimized if $D_{KL}(p||q)=0$, that is, $p(x|\mathbf{h},m_i=t)=p(x|\mathbf{h})$

Proof of Theorem 3

Theorem 2 has proved that, $\hat{p}(\mathbf{x}|\mathbf{h},m_i=1)=\hat{p}(\mathbf{x}|\mathbf{h},m_i=0)=\hat{p}(\mathbf{x}|\mathbf{h})$ is valid for f = CE, FKL, RKL, and JS. If $\mathbf{H}$ is independent of $\mathbf{M}$, and $\mathbf{H}$ is conditionally independent of $\mathbf{X}$ given $\mathbf{M}$, it is easy to verify that $\hat{p}(\mathbf{x}|m_i=1)=\hat{p}(\mathbf{x}|m_i=0)$, for all $i\in\{1,...,d\}$. Follow the same argumentation as[1] for the cross-entropy case, there are more parameters than the number of equations, so the density $\hat{p}$ is not unique.

To get a unique density solution, a hinting mechanism is needed such that $\mathbf{H}$ reveals some information of $\mathbf{M}$ to the discriminator D, which means that they are not independent. In the last section, we adopt the method proposed in[1] to sample the hint variable using the Eq. 2, and assume $\mathbf{B}$ and $\mathbf{M}$ are independent. This hinting mechanism can ensure that the generator is capable of replicating the desired distribution of the data, that is the Theorem 4.

Proof of Theorem 4

The proof is similar to the CE scenario[1]. Theorem 2 has shown that $\hat{p}(\mathbf{x}|\mathbf{h}, m_i = 1) = \hat{p}(\mathbf{x}|\mathbf{h}, m_i = 0)$ holds for the f-divergence of CE, FKL, RKL and JS. Because Hint matrix is defined as

$$\mathbf{H} = \mathbf{B} \odot \mathbf{M} + 0.5(1 - \mathbf{B}). \tag{2}$$

, $\hat{p}(\mathbf{x}|\mathbf{h}, m_i = 1) = \hat{p}(\mathbf{x}|\mathbf{b}, m_i = 1) = \hat{p}(\mathbf{x}|\mathbf{h}, m_i = 0) = \hat{p}(\mathbf{x}|\mathbf{b}, m_i = 0)$ is valid. Since $\mathbf{B}$ and $\mathbf{M}$ are independent, it is easy to prove $\hat{p}(\mathbf{x}|m_i = 1) = \hat{p}(\mathbf{x}|m_i = 0)$. It means, for any two vectors $\mathbf{m_1}, \mathbf{m_2} \in \{0, 1\}^d$ that differ only on one component, we have $\hat{p}(\mathbf{x}|\mathbf{m_1}) = \hat{p}(\mathbf{x}|\mathbf{m_2})$.
This equation also holds true for any two vectors $\mathbf{m_1}$ and $\mathbf{m_2}$ in $\{0, 1\}^d$, because we can always find a sequence of vectors between $\mathbf{m_1}$ and $\mathbf{m_2}$, such that all the adjacent vectors differ from each other in only one component. Consequently, the imputed data distribution $\hat{p}(\mathbf{x}|\mathbf{m})$ is the same for all possible vectors $\mathbf{m} \in \{0, 1\}^d$. This unique imputed data density, denoted by $\hat{p}(\mathbf{x}|\mathbf{1})$, corresponds to the true data $\mathbf{X}$'s density $p(\mathbf{x})$, that is, $\hat{p}(\mathbf{x}|\mathbf{m}) = \hat{p}(\mathbf{x}|\mathbf{1}) = p(\mathbf{x})$. The proof is based on the Theorem 2, so it is true for f = CE, FKL, RKL, JS.
Theorem 1-4 theoretically confirm that the generative adversarial imputation network method remains valid if and only if the loss function is defined using four f-divergence, including CE, FKL, RKL, and JS divergence. The flexibility offered by the f-divergence formulation allows sc-fGAIN to accommodate various types of data and distributions, making it a more universal approach for imputing missing values.

S1 Appendix: Tables

References

[1] J. Yoon, J. Jordon and M. Schaar, Gain: Missing data imputation using generative adversarial nets, in *International conference on machine learning*, 2018.

Table 1: Objective functions of the Discriminator in the sc-fGAIN models based on different f-divergence function.

Divergence	Objective function
CE	$\mathcal{L}_D = \mathbb{E}_{\hat{X},M,H}\mathbf{M}^T\left(\log D(\hat{\mathbf{X}},\mathbf{H})\right) + (1-\mathbf{M})^T\left(\log(1-D(\hat{\mathbf{X}},\mathbf{H}))\right)$
FKL	$\mathcal{L}_D = \mathbb{E}_{\hat{X},M,H}\mathbf{M}^T\left(\log \frac{D(\hat{\mathbf{X}},\mathbf{H})}{1-D(\hat{\mathbf{X}},\mathbf{H})}\right) + (1-\mathbf{M})^T\left(-\frac{D(\hat{\mathbf{X}},\mathbf{H})}{e(1-D(\hat{\mathbf{X}},\mathbf{H}))}\right)$
RKL	$\mathcal{L}_D = \mathbb{E}_{\hat{X},M,H}\mathbf{M}^T\left(1-\frac{1}{D(\hat{\mathbf{X}},\mathbf{H})}\right) + (1-\mathbf{M})^T\left(\log \frac{e(1-D(\hat{\mathbf{X}},\mathbf{H}))}{D(\hat{\mathbf{X}},\mathbf{H})}\right)$
JS	$\mathcal{L}_D = \mathbb{E}_{\hat{X},M,H}\mathbf{M}^T\left(\log 2D(\hat{\mathbf{X}},\mathbf{H})\right) + (1-\mathbf{M})^T\left(\log(2-2D(\hat{\mathbf{X}},\mathbf{H}))\right)$
PC	$\mathcal{L}_D = \mathbb{E}_{\hat{X},M,H}\mathbf{M}^T\left(-\log \frac{1-D(\hat{\mathbf{X}},\mathbf{H})}{D(\hat{\mathbf{X}},\mathbf{H})}\right)$ $+ (1-\mathbf{M})^T\left(-\frac{1}{4}(\log \frac{1-D(\hat{\mathbf{X}},\mathbf{H})}{D(\hat{\mathbf{X}},\mathbf{H}})^2 + \log \frac{1-D(\hat{\mathbf{X}},\mathbf{H})}{D(\hat{\mathbf{X}},\mathbf{H})}\right)$

Table 2: Comparison of RMSE score (2000 genes) in the A549 cell line

Method \ Missing rate	0.1	0.2	0.3	0.4	0.5
MAGIC	0.10309	0.09844	0.10168	0.11060	0.12283
scImpute	0.15627	0.15312	0.15401	0.15424	0.15479
PBLR	0.10545	0.13518	0.15233	0.16398	0.17242
sc-fGAIN(CE)	0.15206	0.15113	0.14939	0.15550	0.14953
sc-fGAIN(FKL)	0.12846	0.11287	0.11960	0.13058	0.12606
sc-fGAIN(RKL)	0.18133	0.17231	0.14557	0.14608	0.17078
sc-fGAIN(JS)	0.14958	0.15509	0.15329	0.15407	0.14922