A novel f-divergence based generative adversarial imputation method for scRNA-seq data analysis

Tong Si[1], Zackary Hopkins[2], John Yanev[2], Jie Hou[2], Haijun Gong[1*]

1 Department of Mathematics and Statistics, Saint Louis University, St. Louis, MO, USA
2 Department of Computer Science, Saint Louis University, St. Louis, MO, USA

* haijun.gong@slu.edu

Theoretical Analysis of sc-fGAIN Algorithm

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 Algorithm **??**. We adopt some notations and assumptions in Yoon *et al*'s work [**?**], 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.

Theorem 1. *Let $S_\phi(\mathbf{x}, \mathbf{h})$ be a function: $\chi \to \mathcal{R}$, where $x \in \chi$, $\mathbf{h} \in \mathcal{H}$ (hint space), and $p(\mathbf{x}, \mathbf{h}) > 0$, D be a function: $\chi \to [0, 1]^d$. If the f-divergence based objective function is defined by the Eq.* **??***, then, given a fixed generator G, there always exists one optimal discriminator $D^*(\mathbf{x}, \mathbf{h})$ if f = CE, FKL, RKL, JS, PC.*

Proof. The f-divergence based objective function Eq. **??** can be rewritten as

$$\begin{aligned}
\mathcal{L}_{DG,f}(\hat{\mathbf{X}}, \mathbf{M}, \mathbf{H}) &= \mathbb{E}_{\hat{X},M,H}[\mathbf{M}^T g_f(S_\phi(\mathbf{x}, \mathbf{h})) - (1 - \mathbf{M})^T f^*(g_f(S_\phi(\mathbf{x}, \mathbf{h})))] \\
&= \int_\chi \int_\mathcal{H} \sum_{i=1}^{d} g_f(S_\phi(x, h))_i p(\mathbf{x}, \mathbf{h}, m_i = 1) \\
&+ f^*(g_f(S_\phi(x, h)))_i p(\mathbf{x}, \mathbf{h}, m_i = 0) dh dx.
\end{aligned}$$

Given a fixed Generator G, the optimal Discriminator D^* is obtained by solving the equation $\frac{\partial \mathcal{L}_{DG,f}}{\partial S_\phi} = 0$, that is

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

After inserting the f-divergence's output activation functions and conjugate functions given in Table 1, and applying the sigmoid function $D_\phi(x) = \frac{1}{1+\exp^{-S_\phi(x)}}$ on the output of the discriminator network $S_\phi(x)$, we identified five f-divergences, including CE, FKL, RKL, JS, and PC, that always have an optimal discriminator D^* given a fixed G, for $i \in \{0, 1\}^d$,

$$D^*(\mathbf{x}, \mathbf{h})_i = \begin{cases} p(m_i = 1|\mathbf{x}, \mathbf{h}), & \text{if } f = \text{CE, RKL, JS} \\ \frac{p(m_i=1|\mathbf{x},\mathbf{h})e}{p(m_i=0|\mathbf{x},\mathbf{h}) + p(m_i=1|\mathbf{x},\mathbf{h})e}, & \text{if } f = \text{FKL} \\ \frac{1}{exp(2-2p(m_i=1|\mathbf{x},\mathbf{h})/p(m_i=0|\mathbf{x},\mathbf{h}))+1}, & \text{if } f = \text{PC}. \end{cases}$$

For a more detailed proof of Theorem 1, please refer to **??**. □

Table 1. f-divergence's output activation function, conjugate function, and the optimal discriminator D^* for a given generator G, $p = p(x, h, m_i = 1)$, and $q = p(x, h, m_i = 0)$.

f-Divergence	Output activation $g_f(s)$	Conjugate $f^*(t)$	Optimal D^*
CE	$-\log(1 + \exp(-s))$	$-\log(1 - \exp(t))$	$\frac{p}{p+q}$
FKL	s	$\exp(t-1)$	$\frac{pe}{pe+q}$
RKL	$-\exp(-s)$	$-1 - \log(-t)$	$\frac{p}{p+q}$
JS	$\log(2) - \log(1 + \exp(-s))$	$-\log(2 - \exp(t))$	$\frac{p}{p+q}$
PC	s	$\frac{1}{4}t^2 + t$	$\frac{exp(2(p-q)/q)}{1+exp(2(p-q)/q)}$

If we substitute the optimal discriminator D^* derived in Theorem 1 into the objective function Eq. **??**, we obtain the loss function of the generator G as follows:

$$\mathcal{L}_{G,f}(D^*) = \mathbb{E}_{\hat{X},M,H}[\mathbf{M}^T g_f(S_\phi(D^*)) - (1-\mathbf{M})^T f^*(g_f(S_\phi(D^*)))]. \tag{2}$$

Then, by minimizing $\mathcal{L}_{G,f}(D^*)$, we derived the second theorem.

Theorem 2. *The f-divergence based loss function $\mathcal{L}_{G,f}(D^*)$ has a global minimum if and only if the density p satisfies:*

$$\hat{p}(\mathbf{x}, \mathbf{h}, m_i = 1) = \hat{p}(\mathbf{x}, \mathbf{h}, m_i = 0), \tag{3}$$

$$\hat{p}(\mathbf{x}|\mathbf{h}, m_i = 1) = \hat{p}(\mathbf{x}|\mathbf{h}, m_i = 0) = \hat{p}(\mathbf{x}|\mathbf{h}), \tag{4}$$

for each $i \in \{1, \ldots, d\}$, $x \in \mathbf{X}$ and $h \in \mathcal{H}$ such that $p(\mathbf{h}|m_i = t) > 0$. And this theorem is true only if f = CE, FKL, RKL, JS.

Yoon *et al*'s work [**?**] proved the validity of this theorem for the cross-entropy based loss function. We will prove that this theorem is also valid for the forward KL, reverse KL, and Jensen-Shannon divergence based loss functions described by Eq. 2, but it does not hold for the Pearson χ^2 divergence.

Proof. We will present a concise proof of this theorem, focusing on the KL-divergence case, which is more intricate compared to the cross-entropy scenario. After substituting D^*, using the Eq. 2 and objective function in the Table **??**, the KL-divergence based loss function can be simplified as

$$\mathcal{L}_{G,f}(D^*) = \int_{\mathcal{X}} \int_{\mathcal{H}} \sum_{i=1}^{d} p(\mathbf{x}, \mathbf{h}, m_i = 1) \log \frac{p(\mathbf{x}, \mathbf{h}, m_i = 1)}{p(\mathbf{x}, \mathbf{h}, m_i = 0)} dh dx.$$

It follows that $\mathcal{L}_{G,f}(D^*)$ is minimized if and only if $p(\mathbf{x}, \mathbf{h}, m_i = 1) = p(\mathbf{x}, \mathbf{h}, m_i = 0)$ for any $i \in \{1, ..., d\}$.

The above loss function can also be rewritten as

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

Since KL divergence D_{KL} 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 $i \in \{1, ..., d\}$. The detailed proof for different f-divergence cases are given in the **??**. $\square$

In comparison to [**?**], our work in Theorem 1-2 offers a more general proof based on the f-divergence functions, establishing that the optimal discriminator and generator can be attained using the sc-fGAIN algorithm when the loss function is formulated using four distinct f-divergence functions: cross-entropy, KL, reverse KL, and JS divergence. Theorem 2 demonstrates the independence of $\mathbf{x}$ from the mask variable $\mathbf{M}$ given the hint variable $\mathbf{H}$. The amount of information contained in $\mathbf{H}$ directly influences the learning capability of the generator G. If $\mathbf{H}$ contains less informative hints or lacks important information, the learning ability of the generator may be compromised, which is discussed in the Theorem 3.

Theorem 3. *In the sc-fGAIN algorithm, for f = CE, FKL, RKL, and JS, if the hint variable $\mathbf{H}$ is independent of mask variable $\mathbf{M}$, then the density $\hat{p}$ in the Theorem 2 is not unique.*

Proof. 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 [**?**] for the cross-entropy case, there are more parameters than the number of equations, so the density $\hat{p}$ is not unique. $\square$

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 [**?**] to sample the hint variable using the Eq. **??**, 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.

Theorem 4. *If the hint variable $\mathbf{H}$ is sampled according to Eq.* **??**, *then the density $\hat{p}$ in Theorem 2 is unique and satisfies $\hat{p}(\mathbf{x}|\mathbf{m}) = \hat{p}(\mathbf{x}|\mathbf{1})$ for any vector $\mathbf{m} \in \{0, 1\}^d$ and f = CE, FKL, RKL, JS, where $\hat{p}(\mathbf{x}|\mathbf{1})$ is the density of $\mathbf{X}$. That is, the distribution of imputed data is same as the distribution of original data.*

Proof. The proof is similar to the CE scenario [**?**]. 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 of Eq. **??**, $\hat{p}(\mathbf{x}|\mathbf{h}, m_i = 1) = \hat{p}(\mathbf{x}|\mathbf{b}, m_i = 1) = \hat{p}(\mathbf{x}|\mathbf{h}, m_i = 0) = \hat{p}(\mathbf{x}|\mathbf{b}, m_i = 0)$ 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. $\square$

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.