# 8. Proofs

## 8.1. Proof of Theorem 1

**Theorem 1** [No Free Lunch] Let $x \in \mathcal{X}$ where $\mathcal{X}$ is a finite set. Let $p(x)$ be a uniform distribution on $\mathcal{X}$. Let $q$ be any antithetic distribution $q(x_1, x_2)$. Let $\mathcal{F}$ be the set of functions $\mathcal{X} \to \mathbb{R}$ such that $\mathrm{Var}_{p(x_1)p(x_2)}[\hat{\mu}_f(x_{1:2})] \neq 0$. Then

$$\max_{f \in \mathcal{F}} \frac{\mathrm{Var}_{q(x_1,x_2)}[\hat{\mu}_f(x_{1:2})]]}{\mathrm{Var}_{p(x_1)p(x_2)}[\hat{\mu}_f(x_{1:2})]} \geq 1 - \frac{1}{|\mathcal{X}| - 1} \tag{22}$$

For sampling without replacement for any $f \in \mathcal{F}$

$$\frac{\mathrm{Var}_{q(x_1,x_2)}[\hat{\mu}_f(x_{1:2})]]}{\mathrm{Var}_{p(x_1)p(x_2)}[\hat{\mu}_f(x_{1:2})]} = 1 - \frac{1}{|\mathcal{X}| - 1} \tag{23}$$

*Proof of Theorem 1.* First we show that

$$\begin{aligned}
&\mathrm{Var}_{q(x_1,x_2)}[\hat{\mu}_f(x_{1:2})] \\
&= \frac{1}{4}\big(\mathrm{Var}_{q(x_1,x_2)}[f(x_1)] + \mathrm{Var}_{q(x_1,x_2)}[f(x_2)] + \\
&\quad 2\mathrm{Cov}_{q(x_1,x_2)}(f(x_1), f(x_2))\big) \\
&= \frac{1}{2}\mathrm{Var}_{p(x)}[f(x)] + \frac{1}{2}\mathrm{Cov}_{q(x_1,x_2)}(f(x_1), f(x_2))
\end{aligned}$$

In addition

$$\mathrm{Var}_{p(x_1)p(x_2)}[\hat{\mu}_f(x_{1:2})] = \frac{1}{2}\mathrm{Var}_{p(x)}[f(x)]$$

So

$$\frac{\mathrm{Var}_{q(x_1,x_2)}[\hat{\mu}_f(x_{1:2})]}{\mathrm{Var}_{p(x_1)p(x_2)}[\hat{\mu}_f(x_{1:2})]} = 1 + \frac{\mathrm{Cov}_{q(x_1,x_2)}(f(x_1), f(x_2))}{\mathrm{Var}_{p(x)}[f(x)]}$$

Denote $|\mathcal{X}|$ by $k$, and the elements of $\mathcal{X}$ by $v_1, v_2, \cdots, v_k$. We only have to show

$$\max_{f \in \mathcal{F}} \frac{\mathrm{Cov}_{q(x_1,x_2)}(f(x_1), f(x_2))}{\mathrm{Var}_{p(x)}[f(x)]} \geq -\frac{1}{k-1} \tag{24}$$

which is equivalent to Eq.(23).

Let $\mathcal{X} = \{v_1, \cdots, v_k\}$ be the set of $k$ values $x$ can take. Denote

$$Q = \begin{pmatrix} q(v_1,v_1) & q(v_1,v_2) & \cdots & q(v_1,v_k) \\ q(v_2,v_1) & q(v_2,v_2) & \cdots & q(v_2,v_k) \\ & & \cdots & \\ q(v_k,v_1) & q(v_k,v_2) & \cdots & q(v_k,v_k) \end{pmatrix}$$

Because $q(x_1, x_2)$ is an antithetic distribution for the uniform distribution $p(x)$, it must satisfy

$$\mathbf{1}^T Q = \frac{1}{k}\mathbf{1} \quad Q\mathbf{1} = \frac{1}{k}\mathbf{1}$$

Denote

$$f = (f(v_1), f(v_2), \cdots, f(v_k))^T$$

Then because the marginal is uniform $p(v_1) = \cdots = p(v_k) = 1/k$

$$\begin{aligned}
&\mathrm{Cov}_{q(x_1,x_2)}[f(x_1), f(x_2)] \\
&= \sum_{v_1,v_2 \in \mathcal{X}} f(v_1)f(v_2)(q(v_1,v_2) - p(v_1)p(v_2)) \\
&= f^T(Q - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T)f \\
&= f^T\left(\frac{Q + Q^T}{2} - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T\right)f
\end{aligned}$$

where the last step is because

$$f^T Q f = (f^T Q f)^T = f^T Q^T f = f^T \frac{Q + Q^T}{2} f$$

Therefore for each non-symmetric $Q$, there is a symmetric joint distribution $\frac{Q+Q^T}{2}$ that achieves the same covariance. For the rest of this proof we assume that $Q$ is symmetric without loss of generality. We will use the notation

$$R \stackrel{\text{def}}{=} Q - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T$$

$R$ is a symmetric matrix

We also have

$$\begin{aligned}
\mathrm{Var}_{p(x_1)p(x_2)}[f(x)] &= \frac{1}{k}\sum_{x \in \mathcal{X}} f(x)^2 - \frac{1}{k^2}\sum_{x_1,x_2} f(x_1)f(x_2) \\
&= \frac{1}{k}f^T f - \frac{1}{k^2}f^T\mathbf{1}\mathbf{1}^T f \\
&= f^T\left(\frac{1}{k}I - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T\right)f
\end{aligned}$$

We will use the notation

$$R' \stackrel{\text{def}}{=} \frac{1}{k}I - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T$$

To briefly summarize our notation we have

$$\begin{aligned}
\mathrm{Cov}_{q(x_1,x_2)}[f(x_1), f(x_2)] &= f^T R f \\
\mathrm{Var}_{p(x)}[f(x)] &= f^T R' f
\end{aligned}$$

Now we try to find for any $R$, some $f$ such that $f^T R f / f^T R' f$ is large. In other words, we want to prove

$$\max_{f \in \mathcal{F}} \frac{f^T R f}{f^T R' f} \geq -\frac{1}{k-1} \tag{25}$$

which is equivalent to Eq.(24). As is the condition of the theorem, we require $f \in \mathcal{F}$ to satisfy $f^T R' f \neq 0$.

For any such matrix $R$, $\mathbf{1}$ must be an eigenvector with eigenvalue $0$. This is because by our definition

$$R\mathbf{1} = Q\mathbf{1} - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T\mathbf{1} = \frac{1}{k}\mathbf{1} - \frac{1}{k}\mathbf{1} = 0$$

In addition, $\mathbf{1}$ is also an eigenvector of $R'$ with eigenvalue 0 because

$$R'\mathbf{1} = \frac{1}{k}\mathbf{1} - \frac{1}{k}\mathbf{1} = 0$$

For any $f$ that is not a scalar multiple of $\mathbf{1}$, $f^TR'f > 0$. This is because

$$\text{rank}(R') \geq \text{rank}(I) - \text{rank}(\mathbf{1}\mathbf{1}^T) \geq k - 1$$

so $\mathbf{1}$ (or its scalar multiple) must be the only eigenvector with 0 as its eigenvalue. In addition $f^TR'f \geq 0$ because it is a variance.

This also implies that $f \in \mathcal{F}$, if and only if $f^TR'f \neq 0$, if and only if $f$ is not a scalar multiple of $\mathbf{1}$.

We consider two situations

**1)** $R$ has at least one positive eigenvalue. Let $f$ be the corresponding eigenvector, we have

$$f^TRf > 0 \qquad f^TR'f > 0$$

and certainly $f$ is not a scalar multiple of $\mathbf{1}$, which means that Eq.(25) must be true.

**2)** $R$ does not have any positive eigenvalues. Because $Q$ is a matrix with no negative entries, $\text{tr}(Q) \geq 0$. In addition $\text{tr}(\frac{1}{k^2}\mathbf{1}\mathbf{1}^T) = \frac{1}{k}$, so

$$\text{tr}(R) = \text{tr}(Q) - \text{tr}(\frac{1}{k^2}\mathbf{1}\mathbf{1}^T) \geq -\frac{1}{k} \tag{26}$$

We know that $R$ must have a zero eigenvalue, and all other eigenvalues are non-positive. We arrange them in non-ascending order

$$0 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_k$$

It is easy to see that $\lambda_2 \geq -\frac{1}{k(k-1)}$ because otherwise

$$tr(R) = \sum_i \lambda_i < -\frac{1}{k(k-1)}(k-1) < -\frac{1}{k}$$

which violates Eq.(26). Suppose the eigenvector corresponding to $\lambda_2$ is $g$. Because $R$ is symmetric, we can always select $g$ orthogonal to the other eigenvectors. In particular, $g^T\mathbf{1} = 0$. The $f$ we will choose is $f_{\text{bad}} = g - \mathbf{1}$. We know that $f_{\text{bad}} \in \mathcal{F}$ as it is not a scalar multiple of $\mathbf{1}$. For $f_{\text{bad}}$, we have

$$\begin{aligned} f_{\text{bad}}^T R f_{\text{bad}} &= (g-\mathbf{1})^T R (g - \mathbf{1}) \\ &= g^TRg \geq -\frac{1}{k(k-1)}g^Tg \end{aligned}$$

where the above inequalities come from the fact that $R\mathbf{1} = \mathbf{1}^TR = 0$, and $g^T\mathbf{1} = 0$.

Similarly we have

$$\begin{aligned} & f_{\text{bad}}^T R' f_{\text{bad}} = (g-\mathbf{1})^T R'(g-\mathbf{1}) \\ &= \frac{1}{k}(g-\mathbf{1})^T(g-\mathbf{1}) - \frac{1}{k^2}(g-\mathbf{1})^T\mathbf{1}\mathbf{1}^T(g-\mathbf{1}) \\ &= \frac{1}{k}(g^Tg + k) - \frac{1}{k^2}k^2 = \frac{1}{k}g^Tg \end{aligned}$$

This means that for this choice of $f_{\text{bad}} = g - \mathbf{1}$

$$\frac{f_{\text{bad}}^T R f_{\text{bad}}}{f_{\text{bad}}^T R' f_{\text{bad}}} \geq -\frac{1}{k-1}$$

which proves Eq.(25).

Finally we show that sampling without replacement achieves equality. For sampling without replacement

$$\begin{aligned} Q &= \begin{pmatrix} 0 & \frac{1}{k(k-1)} & & \frac{1}{k(k-1)} \\ \frac{1}{k(k-1)} & 0 & & \frac{1}{k(k-1)} \\ & & \cdots & \\ \frac{1}{k(k-1)} & \frac{1}{k(k-1)} & & 0 \end{pmatrix} \\ &= \frac{1}{k(k-1)}(\mathbf{1}\mathbf{1}^T - I) \end{aligned}$$

Then

$$R = Q - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T = \frac{1}{k^2(k-1)}\mathbf{1}\mathbf{1}^T - \frac{1}{k(k-1)}I$$

Note that the set of eigenvalues for $\mathbf{1}\mathbf{1}^T$ is

$$k, 0, \cdots, 0$$

so the eigenvalues for $R$ must be

$$0, -\frac{1}{k(k-1)}, \cdots, -\frac{1}{k(k-1)}$$

Denote this eigen-decomposition as $R = H^T\Lambda H$. As before let $R' = \frac{1}{k}I - \frac{1}{k^2}\mathbf{1}\mathbf{1}^T$. Because $R'$ is a scalar multiple of $R$, $R'$ must have the same eigenvectors as $R$, with eigenvalues

$$0, \frac{1}{k}, \cdots, \frac{1}{k}$$

Denote the eigen-decomposition as $R' = H^T\Lambda'H$. Choose any $f$, we compute $g = Hf$. If $g = (*, 0, \cdots, 0)$ ($*$ denotes any real number) we will have $f^TR'f = g^T\Lambda'g = 0$ and our theorem excludes this degenerate situation. When $g \neq (*, 0, \cdots, 0)$, we have

$$\begin{aligned} \frac{\text{Cov}_{q(x_1,x_2)}(f(x_1), f(x_2))}{\text{Var}_{p(x)}[f(x)]} &= \frac{f^TRf}{f^TR'f} \\ &= \frac{g^T\Lambda g}{g^T\Lambda' g} = -\frac{1}{k-1} \end{aligned}$$

This means that sampling without replacement achieves our theoretical upper bound on minimax performance. $\square$

### 8.2. Proof of Proposition 1

**Proposition 1** Let $q_\theta(\boldsymbol{x}_{1:m})$ be a Gaussian-reparameterized antithetic of order $m$ for $p(\boldsymbol{x})$. Then for any $k$:

1. For any $\Sigma_\theta \in \boldsymbol{\Sigma}_{\text{unbiased}}$, the estimator (10) is unbiased
$$\mathbb{E}_{q_\theta(\boldsymbol{x}_{1:m})}[\hat{\mu}_f(\boldsymbol{x}_{1:m})] = \mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})]$$
2. If $\Sigma_\theta = I$, the Gaussian-reparameterized antithetic is equivalent to i.i.d sampling.
3. Given a Cholesky decomposition $\Sigma_\theta = L_\theta L_\theta^T$, we can sample from $q_\theta(\boldsymbol{x}_{1:m})$ by drawing $m$ i.i.d. samples $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m)^T$ from $\mathcal{N}(0, I_d)$, and $\boldsymbol{x}_{1:m} = L_\theta \boldsymbol{\delta}$.

*Proof of Proposition 1.* Part 1: Because $\Sigma_\theta \in \boldsymbol{\Sigma}_{\text{unbiased}}$, each component $\boldsymbol{\epsilon}_i$ of $(\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_m) \sim \mathcal{N}(0, \Sigma_\theta)$ is marginally $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I_d)$. By assumption, this means that $g(\epsilon_i) \sim p(x)$. Combined with Eq. (3 ) thsi finishes the proof.

Part 2: By construction, if $\Sigma_\theta = I$ then $(\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_m) \sim \mathcal{N}(0, \Sigma_\theta)$ are i.i.d. Thus $g(\epsilon_i)$ are also i.i.d.

Part 3: Given a Cholesky decomposition $\Sigma_\theta = L_\theta L_\theta^T$, we can sample $(\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_m)$ via $(\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_m) = L_\theta(\mathbf{z}_1, \cdots, \mathbf{z}_m)$ where $(\mathbf{z}_1, \cdots, \mathbf{z}_m) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. □

### 8.3. Proof of Theorem 2

**Theorem 2** For any $\epsilon > 0$, the map $\psi$ defined in Eq.(13) is a surjection from $\mathbb{M}^{m\times m}$ into $\boldsymbol{\Sigma}_{\text{unbiased}}$.

*Proof of Theorem 2.* We first check that $\psi$ is well defined.

For any $\theta \in \mathbb{M}^{m\times m}$, denote $\tilde{\Sigma} = \epsilon I + \theta\theta^T \in \mathbb{M}^{m\times m}$. When $\epsilon > 0$, this must be positive definite as a matrix in $\mathbb{R}^{md\times md}$. Because $\tilde{\Sigma}$ is positive definite as a matrix $\mathbb{R}^{md\times md}$, each element of $\text{diag}(\tilde{\Sigma})$ as a matrix $\mathbb{M}$ must be a positive definite element of $\mathbb{M}$, and must have an inverse. This means that $\text{diag}(\tilde{\Sigma})^{-1/2}$ is also well defined. Therefore $\psi(\theta)$ is well defined.

It is obvious that $\psi(\theta)$ has identity diagonal. It is also positive semi-definite, so $\psi(\theta) \in \boldsymbol{\Sigma}_{\text{unbiased}}$.

Now we prove that the map is a surjection. Choose any $\Sigma \in \boldsymbol{\Sigma}_{\text{unbiased}}$, let

$$\zeta(\Sigma) = \text{diag}(\Sigma)^{-1/2}\Sigma\text{diag}(\Sigma)^{-T/2}$$

then it is easy to see that $\zeta(\Sigma) = \Sigma$. In addition, for any diagonal matrix $D \in \mathbb{M}^{m\times m}$ whose diagonal elements are all positive definite elements of $\mathbb{M}$, we have $\zeta(D\Sigma D^T) = \Sigma$. We choose $D = \alpha I$, where $\alpha \in \mathbb{R}_{>0}$; $I$ is the identity matrix of $\mathbb{M}^{m\times m}$. We choose a sufficiently large $\alpha$ such that $\alpha^2\Sigma - \epsilon I$ is positive definite element of $\mathbb{R}^{md\times md}$. By the cholesky decomposition in $\mathbb{R}^{md\times md}$, there exists $\theta \in \mathbb{R}^{md\times md}$ such that $\theta\theta^T = \alpha^2\Sigma - \epsilon I$. We have, by construction, found a $\theta$ that satisfy $\psi(\theta) = \Sigma$. This is because $\epsilon I + \theta\theta^T = \alpha^2\Sigma = \alpha\Sigma\alpha$, so $\zeta(\epsilon I + \theta\theta^T) = \Sigma$.

□

## 9. Results of IWAE

| | MNIST | | Omniglot | |
|---|---|---|---|---|
| noise dimension | 5 | 10 | 5 | 10 |
| i.i.d sampling | 113.79 | 98.92 | 142.50 | 130.65 |
| negative sampling | 113.71 | 98.89 | 142.35 | 130.37 |
| Our method | **113.61** | **98.71** | **142.15** | **130.23** |

Table 1: Negative Log Likelihood of our methods compared with negative sampling and i.i.d sampling on MNIST and Omniglot dataset. Our method can achieve a tighter bound on all settings.

## 10. Results of GANs

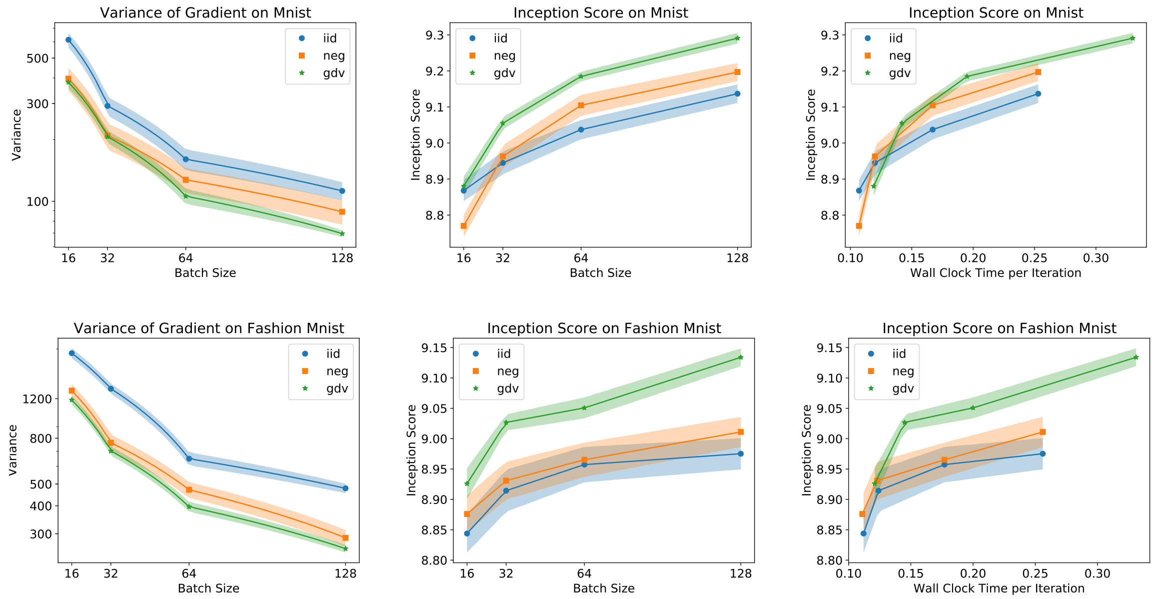


Figure 3: Variance reduction for GAN training. **Left:** Variance of gradient estimation for different batch sizes. **Middle:** Inception score after 50 epochs of training for different mini-batch batch sizes $m$. **Right:** Inception score by wall-clock time. For small batch size $m$, adaptive antithetic improves marginally compared to baselines; because of its overhead, the overall wall-clock time is worse; for larger batch size $m$, adaptive antithetic performs significantly better, the overall wall-clock time is also better.