Learning Smooth and Fair Representations: Supplementary Materials

1 Appendix - Proofs of Results

1.1 Proof of Theorem 1

The proof of Theorem 1 uses the following lemma (from **?**) that links the demographic parity of a test function f and its balanced error rate $BER(f)$,

$$BER(f,t) = \frac{P(f(Z)=1|S=0) + P(f(Z)=0|S=1)}{2}, \tag{1}$$

where we make the dependence on the representation mapping t explicit in $BER(f,t)$.

Lemma 1.1. **?** *A representation space $(\mathcal{Z}, \mu_t)$ satisfies an $\Delta^*(t)-$ demographic parity certificate if and only if*

$$BER^*(t) \triangleq \min_{f:\mathcal{Z}\to\{0,1\}} BER(f,t) \geq \frac{1-\Delta}{2}. \tag{2}$$

Therefore, a representation space $(\mathcal{Z}, \mu_t)$ can be stamped with a $\Delta^*(t)-$ demographic parity certificate with $\Delta^*(t) \equiv 1 - 2BER^*(t)$.

To prove the result in Theorem 1, we consider a deterministic transformation t.

Lemma 1.2. *Suppose that t is a deterministic mapping from $\mathcal{X}$ to $\mathcal{Z}$. Denote K the size of $t(\mathcal{X})$ with $K \leq \infty$. Then, for all distribution μ_x over the features $\mathcal{X}$ such that for all $z \in t(\mathcal{X})$, $P(t(X)=z) > 0$, $I_{\chi^2}(X,Z) = K-1$.*

Proof. First, since t is a function, $P(Z=z|X=x)$ is equal to one if and only if $t(x)=z$. Therefore,

$$\begin{aligned} I_{\chi^2}(X,Z) &= E_x\left(1 - \frac{1}{P(Z=t(x))}\right)^2 P(Z=t(x)) \\ &= E_x\left[\frac{1}{P(Z=t(x))}\right] - 1 \\ &= \sum_{z\in t(\mathcal{X})}\left[\frac{P(X, t(X)=z)}{P(Z=z)}\right] - 1 \\ &= K-1 \end{aligned} \tag{3}$$

□

Now for a given distribution μ_x over $\mathcal{X}$ and a given transformation t, there are two cases: $I_{\chi^2}(X,Z) = \infty$ and $I_{\chi^2}(X,Z) < \infty$. Let denote $I_{\chi^2}(X,Z)$ by I_{χ^2}.

1.1.1 Case $I_{\chi^2} < \infty$

By lemma **??**, $t(\mathcal{X})$ is finite and $t(\mathcal{X}) = \{z_1, z_2, ..., z_K\}$, with $K \leq \infty$ and $z_k \neq z_{k'}$ for $k \neq k^{'}$.

For each $k \in \{1, ..., K\}$, we choose one $x_k \in \mathcal{X}$ such that $t(x_k) = z_k$. We parametrize a family of joint distributions $\mu(b)$ over $[0,1] \times \{0,1\}$ as follows: X is uniformly distributed over $\{x_1, ..., x_K\}$; and, for $b \in (0,1)$, the sensitive attribute is given by k^{th} binary expansion of b, where $X = x_k$. By Lemma **??**, the χ^2 squared mutual information between X and $t(X)$ is the same for any b and equal to $K-1$. Moreover, since the sensitive attribute is a function of $t(X)$, $\Delta_b^*(t) = 1$, where the subscript indicates that demographic parity is computed using the joint distribution $\mu(b)$ over (Z, S).

Let B denote a random variable uniformly distributed on $[0,1]$. For any auditor f_n,

$$\begin{aligned} \sup_{b\in[0,1]} E_{\mathcal{D}_n(b)} BER(f_n, t) &\overset{\text{(a)}}{\geq} E_B E_{\mathcal{D}_n(B)} BER(f_n, t) \\ &= E_{X,B} P[f_n((t(X), \mathcal{D}_n(B)) \neq \\ &S|t(X_1), ..., t(X_n), S_1, ...S_n, t(X)] \\ &\overset{\text{(b)}}{\geq} \frac{1}{2} P\left(\cap_{i=1}^n [t(X) \neq t(X_i)]\right) \\ &\overset{(c)}{=} \frac{1}{2}\left(1 - \frac{1}{K}\right)^n \\ &\overset{(d)}{=} \frac{1}{2}\left(1 - \frac{1}{I_{\chi^2}(X,Z)}\right)^n \end{aligned} \tag{4}$$

where (a) uses that the suppremum is larger than the average; (b) that for $Z \notin \{Z_1, ..., Z_n\}$, the sensitive attribute has a Bernouilli distribution with probability $1/2$; (c) that X and then Z is uniformly distributed; and, (d) that $I_{\chi^2}(X,Z) \leq K$ by Lemma **??**. Since $I_{\chi^2}(X,Z)$ is equal for all b, it follows from Lemma **??** that

$$\sup_{b\in(0,1)} \Delta^* - \Delta(f_n, t) \geq \left(1 - \frac{1}{I_{\chi^2}(Z,X)}\right)^n. \tag{5}$$

Note that μ does not depend on μ_x. Therefore, for all auditors f_n,

$$\sup_{\mu} E_{\mathcal{D}_n} |\Delta^* - \Delta(f_n, t)| \geq \left(1 - \frac{1}{I_{\chi^2}}\right)^n. \tag{6}$$

1.1.2 Case $I_{\chi^2} = \infty$

By Lemma **??**, if for a distribution μ over $\mathcal{X} \times \{0,1\}$, $I_{\chi^2}(Z,X) = \infty$, then there exists an infinite countable set $\{a_k\}$ of $\mathcal{X}$ such that t takes a different value at each a_k. We choose X to take value in $\{a_k\}_{k\geq 1}$ such that $P(a_k) = p_k$ for $k \geq 0$ where the sequence $\{p_k\}_{k=1}^{\infty}$ will be chosen later on. As in the previous case, we parametrize a family of distributions over $\mathcal{X} \times \{0,1\}$ by $b \in (0,1)$ such that for $X \in \{a_1, ...\}$, the sensitive attribute S is the k^{th} term of b's binary expansion, where $X = a_k$. Because S is a deterministic function of X, $\Delta^*(t) = 1$.

Let B denote a random variable uniformly distributed on $[0,1]$. For a sample point X_i, we denote k_i such that $X_i = a_{k_i}$. For any auditor f_n,

$$\begin{aligned} \sup_{b\in[0,1]} E_{\mathcal{D}_n(b)} BER(f_n, t) &\overset{\text{(a)}}{\geq} E_B E_{\mathcal{D}_n(B)} BER(f_n, t) \\ &= E_{X,B} P[f_n((t(X), \mathcal{D}_n(B)) \neq \\ &S|t(X_1), ..., t(X_n), S_1, ...S_n, t(X)] \\ &\overset{\text{(b)}}{\geq} \frac{1}{2} P\left(\cap_{i=1}^n [k \neq k_i]\right) \\ &\overset{\text{(c)}}{=} \frac{1}{2} \sum_{k=1}^{\infty} p_k (1 - p_k)^n \end{aligned} \tag{7}$$

It remains to show that for all $\epsilon > 0$, we can choose $\{p_k\}$ such that the right hand side of inequality (**??**) is at least $1/2(1-\epsilon)$. Let $\epsilon > 0$. We choose p_k as follows. First, pick $K > \frac{1}{1-(1-\epsilon)^{1/n}}$.Then, let $p_k = 1/K$ for $1 \le k \le K$ and $p_k = 0$ elsewhere. It follows that

$$\sup_{b\in[0,1]} E_{\mathcal{D}_n(b)} BER(f_n, t) \ge \frac{1}{2}\left(1 - \frac{1}{K}\right)^n \ge \frac{1}{2}(1-\epsilon). \tag{8}$$

Therefore, using Lemma **??**, we can conclude that for all $\epsilon > 0$, there exists a distribution over $\mathcal{X} \times \{0, 1\}$ such that for all auditors f_n

$$\Delta^*(t) - \Delta(f_n, t) \ge 1 - \epsilon. \tag{9}$$

Therefore,

$$\sup_{\mu} \Delta^*(t) - \Delta(f_n, t) \ge 1 = \left(1 - \frac{1}{I_{\chi^2}}\right)^n. \tag{10}$$

1.1.3 Final Step

Therefore, by combining both cases $I_{\chi^2} < \infty$ and $I_{\chi^2} = \infty$, we have that for all distribution μ_x over the features $\mathcal{X}$,

$$\sup_{\mu} \Delta^*(t) - \Delta(f_n, t) \ge 1 = \left(1 - \frac{1}{I_{\chi^2}}\right)^n, \tag{11}$$

which implies the result in theorem 1.

1.2 Proof of Corollary 1

Suppose that $\inf_{f_n\in\mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n}|\Delta^* - \Delta(f_n, t)| \le \epsilon_n$ for some $\epsilon_n > 0$. Let $f_n \in \mathcal{F}_n$ be the auditor that reaches the minimum.

We have, for any distribution μ over $\mathcal{X} \times \{0, 1\}$,

$$\begin{aligned}\left(1 - \frac{1}{I_{\chi^2}(Z, X)}\right)^n &\le \sup_{\mu}\left(1 - \frac{1}{I_{\chi^2}(Z, X)}\right)^n \\ &\overset{(a)}{\le} \sup_{\mu} E_{\mathcal{D}_n}|\Delta^* - \Delta(f_n, t)| \\ &\le \epsilon_n,\end{aligned} \tag{12}$$

where (a) uses Theorem 1. The result follows directly from equation (**??**).

1.3 Examples of Representation Mappings without Finite Sample Guarantees

Injective mappings. Suppose that t is injective from $[0, 1]^D$ to $\mathbb{R}^d$.

Consider X distributed over the countable and infinite set $\{1, 1/2, ...1/k, ...\}$ with $p_k = \kappa/k^2$ and $k^{-1} = \sum_{k=1}^{\infty} 1/k^2$. By lemma **??**, $I_{\chi^2}(X, Z) = \infty$ and thus, by Corollary 1, there exists a distribution such that $\Delta^*(t) - \Delta(f_n, t) = 1$ for all f_n.

Large $t(\mathcal{X})$. Suppose that $|\{t(x)|x \in \mathcal{X}\}| \ge n/(\ln(n))^\alpha$, for some $\alpha < 1$.

By Lemma **??**, $I_{\chi^2}(X, Z) \ge n/(\ln(n))^\alpha - 1$ and thus, by Corollary 1, if $\inf_{f_n\in\mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n}|\Delta^*(t) - \Delta(f_n, t)| = \epsilon_n$, then

$$\begin{aligned}\frac{n}{(\ln(n))^\alpha} - 1 \le I_{\chi^2}(X, Z) &\le \frac{1}{1 - \epsilon_n^{\frac{1}{n}}} \\ &\overset{(a)}{\le} \frac{n}{-\ln(\epsilon_n)},\end{aligned} \tag{13}$$

where (a) uses that $e^{-x} \ge 1 - x$. Therefore, $\epsilon_n \ge e^{-(\ln(n))^\alpha} = \omega(n^{-s})$ for $s > 0$, since $\alpha < 1$.

1.4 Proof of Theorem 2

The proof Theorem 2 relies on a upper bound of $\Delta^*(t) - \Delta(f_n, t)$ that uses the total variation distance $TV(\mu_t^s, \mu_n^s)$ between class conditional densities and their empirical counterpart:

$$TV(\mu_t^s, \mu_n^s) = \int |\mu_t^s - \mu_n^s)|dz. \tag{14}$$

Lemma 1.3. *Consider a sample $\{(z_i, s_i)\}_{i=1}^n$ from a representation distribution μ_t induced by a representation rule t. Suppose that μ_n^0 and μ_n^1 are empirical density estimators of $P(Z|S=0)$ and $P(Z|S=1)$ respectively. Denote f_n the following auditing plug-in decision: for $z \in \mathcal{Z}$, $f_n(z) = 1$ if and only if $\mu_n^1(z) > \mu_n^0(z)$. Therefore, for all n*

$$\Delta(f_n, t) \le \Delta^*(t) \le \Delta(f_n, t) + 2 \sum_{i=0,1} TV(\mu_t^i, \mu_n^i). \tag{15}$$

Proof. Let f^* denote the auditing rule that minimzes the balance error rate. Using **?** (ch 2), we show that for any auditing rule f_n

$$\begin{aligned} 2 - \int \eta_{f_n(z)}(z)\mu_t(dz) &= 2 - \sum_{i=0,1} \int_{f_n(z)=i} \eta_i(z)\mu_t(dz) \\ &= 2 - \sum_{i=0,1} \int_{f_n(z)=i} P(z|S=i)dz \\ &= 2BER(f_n), \end{aligned} \tag{16}$$

where $\eta_i(z)$ is the balanced posteriori probability $\eta_i(z) = P(S=i|Z=z)/P(S=i)$. Moreover,

$$\begin{aligned} 2BER(f^*) &= 2 - P(f^*(z) = 1|S=1] \\ &\quad - P(f^*(z) = 0|S=0) \\ &= 2 - \int_{z, \mu_t^1 > \mu_t^0} \mu_t^1(dz) - \int_{z, \mu_t^0 > \mu_t^1} \mu_t^0(dz) \\ &= 2 - \int \max_i \eta_i(z)\mu_t(dz). \end{aligned} \tag{17}$$

Let denote $\eta_{n,i}$ the empirical estimate of η_i. Using equations (**??**) and (**??**), the proof of lemma **??** relies on the fact that

$$\begin{aligned} BER(f_n) - BER(f^*) &= \int \max_i \eta_i(z)\mu_t(dz) \\ &- \int \eta_{f_n(z)}(z)\mu_t(dz) \\ &= \int (\max_i \eta_i(z) - \max_i \eta_{n,i}(z))\mu_t(dz) \\ &+ \int (\eta_{n,f_n(z)}(z) - \eta_{f_n(z)}(z))\mu_t(dz) \\ &\overset{(a)}{\le} \sum_{i=0,1} \int |\eta_i(z) - \eta_{n,i}(z)|\mu_t(dz) \\ &= \sum_{i=0,1} \int |\mu_t^i(z) - \mu_n^i(z)|dz, \end{aligned} \tag{18}$$

The inequality (a) comes from the following observation. If the maxima are attained for the same $i \in \{0,1\}$, then the right hand side integrand is equal to 0. Otherwise, suppose without loss of generality that $\max \eta_i(z)$ is reached for $i = 0$, then the right hand side integrand is

$$\begin{aligned}\eta_0(z) - \eta_{n,1}(z) + \eta_{n,1}(z) - \eta_1(z) &= \eta_0(z) - \eta_{n,0}(z) \\ &\quad + \eta_{n,1}(z) - \eta_1(z) \\ &\quad + \eta_{n,0}(z) - \eta_{n,1}(z) \\ &\leq |\eta_0(z) - \eta_{n,0}(z)| \\ &\quad + |\eta_1(z) - \eta_{n,1}(z)|,\end{aligned} \tag{19}$$

where the inequality follows $\max_i \eta_{n,i}(z) = \eta_{n,1}(z)$. The same argument can be applied when $\max \eta_i(z) = \eta_1(z)$. The result in lemma **??** follows from (**??**). □

The second part of the proof of theorem 2 is to show that the total variation distance between μ_n^s and μ_t^s is $O(1/\sqrt{n_s})$ for some empirical estimate of μ_t^s:

Lemma 1.4. *Consider a representation mapping $t : \mathcal{X} \to \mathcal{Z}$ and its induced distribution μ_t. Assume that $I_2(Z, X) < \infty$. Then, for $s = 0, 1$, define μ_n^s as*

$$\mu_n^s(z) = \frac{1}{n_s} \sum_{i=1, s_i = s}^{n} P(z|X = x_i) \tag{20}$$

The total variation between μ_t^s and μ_n^s can be bounded as follows:

$$E_{\mathcal{D}\sim\mathcal{X}^n}\left[TV(\mu_t^s, \mu_n^s)\right] \leq \sqrt{\frac{I_2(Z, X)}{n_s}}.$$

The upper bound of the total variation distance uses a Monte Carlo integration argument. For a sample $\mathcal{D}_n = \{x_i\}_{i=1}^n$, denote $\phi(z, x_i)$ the probability $P(Z = z|X = x_i)$. Therefore, $\mu_t(z) = E_{x\sim\mathcal{X}}[\phi(z, x)]$ and if μ_n^s is defined as in (**??**), $\mu_t^s(z) = E_{\mathbf{X}, S=s}[\mu_n^s]$, where $\mathbf{X} = \{x_i\}_{i=1}^n \sim \mathcal{X}^n$. Denote

$$\mathcal{E}^s(\mathbf{X}) = \int \left| \mu_t(z) - \frac{1}{n_s} \sum_{i=1, s_i = s}^{n} \phi(z, x_i) \right| dz, \tag{21}$$

with $n_s = |\{i | s_i = s\}|$. We have

$$\begin{aligned} E_{\mathbf{X}}[\mathcal{E}^s(\mathbf{X})] &\overset{(a)}{\leq} E_{\mathbf{X}}\left[\sqrt{\int \left(\frac{\mu_t(z) - \mu_n^s(z)}{\mu_t(z)}\right)^2 \mu_t(z) dz}\right] \\ &\overset{(b)}{=} \frac{1}{n_s} E_{\mathbf{X}}\left[\sqrt{\int \sum_{i=1, s_i=s}^{n} \left(\frac{\mu_t(z) - \phi(z, x_i)}{\mu_t(z)}\right)^2 \mu_t(z) dz}\right] \\ &\overset{(c)}{\leq} \frac{1}{n_s} \sqrt{E_{\mathbf{X}}\left[\int \sum_{i=1, s_i=s}^{n} \left(\frac{\mu_t(z) - \phi(z, x_i)}{\mu_t(z)}\right)^2 \mu_t(z) dz\right]} \\ &\overset{(d)}{=} \frac{1}{n_s} \sqrt{\sum_{i=1, s_i=s}^{n} E_{\mathbf{X}}\left[\int \left(\frac{\mu_t(z) - \phi(z, x_i)}{\mu_t(z)}\right)^2 \mu_t(z) dz\right]} \\ &\overset{(e)}{=} \sqrt{\frac{I_2(Z, X)}{n_s}}, \end{aligned} \tag{22}$$

where (a) applies Cauchy-Schwarz inequality; (b) uses the fact that the samples are independently drawn and that $E_{x_i}[\phi(z, x_i)] = \mu_t(z)$; (c) that the squared-root is concave; (d) that expectation and integral can be interchange; and, (e) the definition of the chi-squared mutual information between Z and X.

Putting lemma **??** and **??** together, we get the upper bound in theorem 2.

1.5 χ^2 versus Classic Mutual Information

Features are uniformly distributed over $[0,1]$ and $t(x) = i$ for $x \in [1/(i+1), 1/i))$ and $i > 0$. For each $i > 0$, the sensitive attribute is constant over $[1/(i+1), 1/i))$ and equal to 1 with probability 1/2.

Form Lemma **??**, it is clear that $I_{\chi^2}(X,Z) = \infty$. On the other hand, we can show that the classic mutual information between X and Z, $I_{Sh}(X,Z)$ is bounded. Since t is deterministic,

$$\begin{aligned} I_{Sh}(X,Z) &= \sum_{i=1}^{\infty} \frac{\ln(i(i+1))}{i(i+1)} \\ &\leq \frac{\ln(2)}{2} + \int_1^{\infty} \frac{\ln(x(x+1))}{x^2} dx \\ &\overset{(a)}{=} \frac{\ln(2)}{2} + 1 + \int_1^{\infty} \frac{1}{x(x+1)} dx \\ &\overset{(b)}{\leq} \frac{\ln(2)}{2} + 2 < \infty, \end{aligned} \tag{23}$$

where (a) and (b) use integration by part and (b) the fact that $1/x \geq 1/(x+1)$.

1.6 Proof if Theorem 3

We only prove the upper bound on the χ^2 mutual information since the remaining results in Theorem 3 follow directly from Theorem 2.

Since the mapping $(p,q) \to q(p/q-1)^2$ is convex and since Z is an infinite mixtures of Gaussians, we have that for $x \in \mathcal{X}$

$$\begin{aligned} &\int \left(\frac{\mu_{t*\sigma}(z|X=x)}{\mu_{t*\sigma}(z)} - 1 \right)^2 \mu_{t*\sigma}(z) dz \\ \leq &\int \int \left(\frac{\mu_{t*\sigma}(z|X=x)}{\mu_{t*\sigma}(z|X=x^{'})} - 1 \right)^2 \mu_{t*\sigma}(z|X=x^{'}) dz \mu(dx^{'}) \\ &, \overset{(a)}{=} \int \chi^2(z|X=x)||z|X=x^{'})\mu(dx^{'}), \end{aligned} \tag{24}$$

where we use Fubini Theorem to invert the summation over z and $x^{'}$ and (a) uses the definition of the χ^2 divergence between $p(z|X=x)$ and $p(z|X=x^{'}$. Since both $p(z|X=x)$ and $p(z|X=x^{'}$ are Gaussians with variance σ^2 and mean $t(x)$ and $t(x^{'})$, respectively, the integrand in the right hand side of (**??**) can be computed analytically as

$$\begin{aligned} &\chi^2(z|X=x)||z|X=x^{'}) = \\ &\frac{1}{2}\left[\exp\left(\frac{||t(x)-t(x^{'})||_2}{\sigma^2} \right) - 1 \right]. \end{aligned} \tag{25}$$

Therefore,

$$\begin{aligned} I_{\chi^2}(X,Z) &\leq \frac{1}{2} E_{x,x^{'}} \left[\exp\left(\frac{||t(x)-t(x^{'})||_2^2}{\sigma^2} \right) \right] \\ &\leq \frac{1}{2} \exp\left(\frac{2||t||_{\infty}^2}{\sigma^2} \right). \end{aligned} \tag{26}$$

1.7 Proof of Theorem 4

By **?**, we know that the balanced error rate of the optimal auditor f^* is given by

$$\begin{aligned} BER(f^*) &= \frac{1}{2}\int \min(\eta(z,0),\eta(z,1))\mu_{t*\sigma}(dz) \\ &= \frac{1}{4}\int(\eta(z,0)+\eta(z,1))\mu_{t*\sigma}(dz) \\ &- \frac{1}{4}\int|\eta(z,0)-\eta(z,1)|\mu_{t*\sigma}(dz) \\ &\overset{(a)}{=} \frac{1}{2}-\frac{1}{4}\int|\eta(z,0)-\eta(z,1)|\mu_{t*\sigma}(dz), \end{aligned} \tag{27}$$

where (a) uses the definition of $\eta(z,s)=P(Z=z|S=s)/P(z)$. Therefore, by Lemma **??**,

$$\mathcal{L}_{DP}(\mu_{t,\sigma}) = \frac{1}{2}\int|\mu^0_{t,\sigma}(z)-\mu^1_{t,\sigma}(z)|dz \tag{28}$$

and that

$$\mathcal{L}_{DP}(\mu_{n,\sigma}) = \frac{1}{2}\int|\mu^0_{n,\sigma}(z)-\mu^1_{n,\sigma}(z)|dz. \tag{29}$$

Therefore, for any t and any features distribution μ over the features $\mathcal{X}$,

$$\begin{aligned} |\mathcal{L}_{DP}(\mu_{n,\sigma})-\mathcal{L}_{DP}(\mu_{t,\sigma})| &\overset{\text{(a)}}{\leq} \int|(\mu^0_{t,\sigma}(z)-\mu^1_{t,\sigma}(z)) \\ &\quad -(\mu^0_{n,\sigma}(z)-\mu^1_{n,\sigma}(z))|dz \\ &\overset{\text{(b)}}{\leq} \int|(\mu^0_{t,\sigma}(z)-\mu^0_{n,\sigma}(z))|dz \\ &+\int|(\mu^1_{t,\sigma}(z)-\mu^1_{n,\sigma}(z))|dz \\ &\overset{\text{(c)}}{\leq} \exp\left(\frac{||t||^2_\infty}{\sigma^2}\right)\left(\sqrt{\frac{1}{n_0}}+\sqrt{\frac{1}{n_1}}\right), \end{aligned} \tag{30}$$

where (a) and (b) are consequences of triangular inequalities; and (c) follows from the definition of total variation distance, the upper bound in lemma **??** and theorem 3.

1.8 Monte Carlo Approximation

Lemma 1.5. *Let $m>0$ and $n>0$. Consider a sample $\{(x_i,s_i)\}$ and a noise vector $\{noise_{ji}\}$ of $n\times m$ draws from a d-dimensional Gaussian $\mathcal{N}(0,\sigma I_d)$. Denote $\mu_{n,\sigma}$ the empirical density as in (**??**) and for $i=1,...,n$ and $j=1,...,m$ $z_{ij}=t(x_i)+noise_{ij}$. If*

$$\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|\eta_n(z_{ij},1)-\eta_n(z_{ij},0)| \tag{31}$$

then $\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma})$ is an unbiased estimator of $\mathcal{L}_{DP}(\mu_{n,\sigma})$ and

$$E_{noise}\left[(\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma})-\mathcal{L}_{DP}(\mu_{n,\sigma}))^2\right] \leq \frac{8||t||^2_\infty+4\sigma^2}{\sigma^2}\frac{1}{nm}. \tag{32}$$

Proof. First, $\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma})$ is an unbiased estimator of $\mathcal{L}_{DP}(\mu_{n,\sigma})$ because

$$\begin{aligned} E_{noise}\left[\hat{\mathcal{L}}_{DP}\right] &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}E_{noise}[|\eta_n(z_{ij},1)-\eta_n(z_{ij},0)|] \\ &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathcal{L}_{DP}(\mu_{n,\sigma}) \\ &= \mathcal{L}_{DP}(\mu_{n,\sigma}). \end{aligned} \tag{33}$$

Therefore, the mean squared error can be written as

$$\begin{aligned} &E_{noise}\left[(\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) - \mathcal{L}_{DP}(\mu_{n,\sigma}))^2\right] \\ &= \frac{1}{n^2 m}\sum_{i=1}^{n} var_{noise}\left[k(x_i + noise)\right], \end{aligned} \tag{34}$$

where $k(z) = |\eta_n(z,1) - \eta_n(z,0)|$. Moreover, by Gaussian Poincare inequality,

$$\begin{aligned} var_{noise}\left[k(x_i + noise)\right] &\overset{(a)}{\leq} \sigma^2 E_{noise}||\nabla k(x_i + noise)||_2^2 \\ &\overset{(b)}{=} 2\sigma^2 \sum_s E_{noise}\left[||\nabla \log(\mu_{n,\sigma}^s(z,s))||_2^2\right] \end{aligned} \tag{35}$$

where (a) uses the fact that the noise is Gaussian with standard deviation σ; (b) that $z = x_i + noise$ and that $\nabla\eta_n(z,s) = \eta_n(z,s)\nabla\log(\mu_{n,\sigma}^s(z,s)) + (1 - \eta_n(z,s)\nabla\log(\mu_{n,\sigma}^s(z,1-s))$. Moreover, for $s = 0,1$

$$\begin{aligned} \nabla\log(\mu_{n,\sigma}^s(z,s)) &\overset{(a)}{=} \sum_{i=1}^{n} \nabla\log(\phi(z,x_i))P(X = x_i|z) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z - t(x_i))P(X = x_i|z), \end{aligned} \tag{36}$$

where (a) denotes the Gaussian density with mean $t(x)$ and standard deviation σ as $\phi(z,x)$. Therefore,

$$||\nabla\log(\mu_{n,\sigma}^s(z,s))||_2 \leq \frac{||z||_2 + ||t||_\infty}{\sigma^2}. \tag{37}$$

□

Moreover, $z \sim \mu_{n,\sigma}$, which is a mixture of n Gaussians, each with a non-central second moment equal to $\sigma^2 + ||t(x_i)||^2$. Therefore,

$$E_{noise}||z||_2^2 \leq \sigma^2 + ||t||_\infty^2. \tag{38}$$

By combining (**??**), (**??**) (**??**) and (**??**), we obtain that

$$var_{noise}\left[k(x_i + noise)\right] \leq 4\frac{2||t||_\infty^2 + \sigma^2}{\sigma^2}, \tag{39}$$

and thus that

$$E_{noise}\left[(\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) - \mathcal{L}_{DP}(\mu_{n,\sigma}))^2\right] \leq 4\frac{2||t||_\infty^2 + \sigma^2}{\sigma^2 nm}. \tag{40}$$

2 Appendix - Experimental Details

2.1 Dataset

Swiss Roll. For the Swiss Roll synthetic data, we use $20,000$ samples for training the autoencoder (t,g) and $10,000$ fresh samples to train the downstream processors and $10,000$ to evaluate their demogrpahic disparity.

DSprites. For the DSprites dataset, we follow the setting from **?**. The DSprites dataset has six independent factors of variation: color (black or white); shape (square, heart, ellipse), scales (6 values), orientation (40 angles in $[0,2\pi]$); x- and y- positions (32 values each). We adapt the sampling to generate a source of potential unfairness. We consider shape as the sensitive attribute. We assign to each possible combination of attributes a weight proportional to $\frac{i_{shape}}{3} + \left(\frac{i_X}{32}\right)^3$, where $i_{shape} \in \{0,1,2\}$ and $i_X = \{0,1,...,21\}$. we sample $600,000$ combinations of latent factors to train the encoder-decoder; $20,000$ to train the downstream processors; and, $20,000$ to evaluate the disparity of the downstream processors.

Adults. The Adults dataset [1] contains $49K$ individuals and includes information on 10 features related to professional occupation, education attainment, race, capital gains, hours worked and marital status. The sensitive attribute is the gender to which individuals self-identify to. The data is split into a $32K$ set to train the auto-encoder; a $13K$ set to train the downstream processors; and, a $3K$ test set to evaluate the disparity of the processors.

Heritage. The Health Heritage dataset [2] contains $220K$ individuals with 66 features related to age, clinical diagnoses and procedure, lab results, drug prescriptions and claims payment aggregated over 3 years. The sensitive attribute is the gender to which individuals self-identify to. After removing individuals with missing records, we split the data into a $142K$ set to train the auto-encoder; a $17K$ set to train the downstream processors; and, a $17K$ test set to evaluate the disparity of the processors.

2.2 Encoder-Decoder

For the DSprites dataset, the autoencoder architecture – taken directly from **?** – includes 4 convolutional layers and 4 deconvolutional layers and uses ReLU activations. For the Swiss Roll dataset and the two real world datasets, the encoder and decoder are made of fully connected layers with ReLU activations. Table **??** shows more architectural details for each dataset. Hyperparameter values are in Table **??**.

Dataset	Encoder	Decoder	Activation
Swiss Roll	Linear(3, 64), Linear(64, 64), Linear(64, 64)	Linear(3, 64), Linear(64, 3)	ReLU/Tanh

DSprites	Conv(1, 32, 4, 2),
	Conv(32, 32, 4, 2),
	Conv(32, 64, 4, 2),
	Conv(64, 64, 4, 2),
	Linear(1024, 128)
		Linear(28, 128),
		Linear(128, 1024)
		ConvT2d(64, 64, 4, 2),
		ConvT2d(64, 32, 4, 2)
		ConvT2d(32, 32, 4, 2)
		ConvT2d(32, 61, 4, 2)
			ReLU/Tanh

Adults	Linear(10, 64), Linear(64, 10)	Linear(10, 64), Linear(64, 10)	ReLU/Tanh
Heritage	Linear(66, 128), Linear(128, 24)	Linear(24, 128), Linear(128, 66)	ReLU/Tanh

Table 1: Architecture details. $Conv2d(i, o, k, s)$ represents a 2D-convolutional layer with input channels i, output channels o, kernel size k and stride s. $ConvT2d(i, o, k, s)$ represents a 2D-deconvolutional layer with input channels i, output channels o, kernel size k and stride s. $Linear(i, o)$ represents a fully connected layer with input dimension i and output dimension o. The *tanh* activation is only applied to the last layer of the encoder.

Dataset	Number of iterations	Learning rate	σ	λ_{max}		
				AGWN	AdvCE	AdvL1
Swiss Roll	4K	10^{-3}	0.05	10	4	4
DSprites	270K	10^{-4}	0.05	0.025	0.035	0.035
Adults	55K	10^{-3}	0.02	2.6	2.8	2.8
Heritage	55K	0.5×10^{-4}	0.05	2.6	2.6	2.6

Table 2: Hyperparameter values for training encoder-decoder networks.

2.3 Comparative Methods

AdvCE. AdvCE is a fair representation learning method from **?**. The auditor is modeled as an adversarial neural network a that predicts sensitive attributes from samples of the representation distribution and minimizes

[1] https://archive.ics.uci.edu/ml/datasets/adult

[2] https://foreverdata.org/1015/index.html

the following cross-entropy loss:

$$\mathcal{L}_{CE}(a) = -\frac{1}{n}\sum_{i=1}^{n} s_i \log(a(x_i) + (1 - s_i)\log(1 - a(x_i)). \tag{41}$$

Moreover, the autoencoder is trained to minimize a loss $\mathcal{L}_{rec} - \lambda\mathcal{L}_{CE}(a)$.

AdvL1 AdvL1 (**?**) replaces the cross-entropy loss by a group L1 loss: instead of (**??**), the adversary minimizes

$$\mathcal{L}_{L1} = \frac{1}{n_0}\sum_{i,s_i=0} a(x_i) - \frac{1}{n_1}\sum_{i,s_i=1} a(x_i), \tag{42}$$

and the autoencoder minimizes $\mathcal{L}_{rec} - \lambda\mathcal{L}_{L1}(a)$.

For both AdvCE and AdvL1, the adversarial auditor is modeled as a neural network with 3 hidden layers of 64 neurons each for Adults and Swiss Roll; 3 hidden layers of 128 neurons each for Heritage; and, 3 hidden layers of 256 neurons each for DSprites.

2.4 Downstream Processors

The downstream test functions that probe the demographic parity of the representation distribution are fully connected neural networks with 2 to 4 hidden layers with 32 to 128 neurons each. Each test function is trained for 400 epochs with a learning rate of 0.001. After the autoencoder is trained, its weights are frozen, and fresh representations are generated by $10,000$ forward passes of the encoder on the test data. The generated fresh representations form the inputs of the test functions.

2.5 Figure 2 to 5

To generate Figure 2 to 4, we train an auto-encoder for a given value of the coefficient λ on the fairness component of the loss function and repeat the simulations 50 times. We vary the value of λ from 0 to λ_{max}, where λ_{max} is reported for each dataset in Table **??**.