Mitigating Information Leakage in Image Representations: A Maximum Entropy Approach
(supplementary Material)

Proteek Chandan Roy and Vishnu Naresh Boddeti
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{royprote, vishnu}@msu.edu

In this supplementary material we include proof of Theorem 1 in Section 1, Corollary 1.1 in Section 2 and finally provide the numerical values of the trade-off fronts in the CIFAR-10 and CIFAR-100 experiment in Section 3.

1. Proof of Theorem 1

Theorem 1. Given a fixed encoder E, the optimal discriminator is $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) = p(s|E(\boldsymbol{x};\boldsymbol{\theta}_E))$ and the optimal predictor is $q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*) = p(t|E(\boldsymbol{x};\boldsymbol{\theta}_E))$.

Proof. Let, $\boldsymbol{z}$ be the fixed encoder output from input $\boldsymbol{x}$ i.e. $\boldsymbol{z} = E(\boldsymbol{x};\boldsymbol{\theta}_E)$. Let, $p(\boldsymbol{x},t,s)$ be the true joint distribution of the variables, i.e. input $\boldsymbol{x}$, target label t and sensitive label s. The fixed encoder is a deterministic transformation of $\boldsymbol{x}$ and generates an implicit distribution $p(\boldsymbol{z},t,s)$.

Discriminator: The objective of the discriminator is,

$$\begin{aligned}
V_1(\boldsymbol{\theta}_E,\boldsymbol{\theta}_D) &= KL\left(p\left(s|\boldsymbol{x}\right) \| q_D\left(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D\right)\right) \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{z},t,s)\sim p(\boldsymbol{z},t,s)} -\log q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D) \\
&= -\sum_{\boldsymbol{x},t,s} p(\boldsymbol{x},t,s)\log q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D) \\
\text{s.t. } & \sum_s q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D) = 1, \ \ \forall \boldsymbol{z} \\
& q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D) \geq 0, \ \ \forall \boldsymbol{z}
\end{aligned} \tag{1}$$

The Lagrangian dual of the problem can be written as

$$L = V_1(\boldsymbol{\theta}_E,\boldsymbol{\theta}_D) + \sum_{\boldsymbol{z}} \lambda(\boldsymbol{z}) \left(\sum_s q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D) - 1\right)$$

Now we take partial derivative of L w.r.t. $q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D^*)$, the distribution of optimal discriminator. Therefore, the optimal discriminator satisfies,

$$\begin{aligned}
\frac{\partial L}{\partial q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D^*)} &= 0 \\
\Rightarrow -\frac{\sum_t p(\boldsymbol{z},t,s)}{q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D^*)} + \lambda(\boldsymbol{z}) &= 0 \\
\Rightarrow \lambda(\boldsymbol{z}) q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D^*) &= p(\boldsymbol{z},s)
\end{aligned} \tag{2}$$

where we used the fact that, $\sum_t p(\boldsymbol{z},t,s) = p(\boldsymbol{z},s)$. Now summing w.r.t. to variable s on the both sides of last line and using the fact that $\sum_s q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D^*) = 1$ we get,

$$\lambda(\boldsymbol{z}) = p(\boldsymbol{z})$$

By substituting $\lambda(\boldsymbol{z})$ we obtain the solution for the optimal discriminator,

$$q_D(s|\boldsymbol{z};\boldsymbol{\theta}_D^*) = \frac{p(\boldsymbol{z},s)}{p(\boldsymbol{z})} = p(s|\boldsymbol{z}) \tag{3}$$

Therefore,

$$q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) = p(s|E(\boldsymbol{x};\boldsymbol{\theta}_E))$$

Target Predictor: The objective of the predictor is,

$$\begin{aligned}
V_2(\boldsymbol{\theta}_E,\boldsymbol{\theta}_T) &= KL\left(p\left(t|\boldsymbol{x}\right) \| q_T\left(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T\right)\right) \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{z},t,s)\sim p(\boldsymbol{z},t,s)} -\log q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T) \\
&= -\sum_{\boldsymbol{x},t,s} p(\boldsymbol{x},t,s)\log q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T) \\
\text{s.t. } & \sum_t q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T) = 1, \ \ \forall \boldsymbol{z} \\
& q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T) \geq 0, \ \ \forall \boldsymbol{z}
\end{aligned} \tag{4}$$

The Lagrangian dual of the problem can be written as

$$L = V_2(\boldsymbol{\theta}_E,\boldsymbol{\theta}_T) + \sum_{\boldsymbol{z}} \lambda(\boldsymbol{z}) \left(\sum_t q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T) - 1\right)$$

Now we take partial derivative of L w.r.t. $q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T^*)$, the distribution of optimal predictor. The optimal predictor satisfies the equation.

$$\begin{aligned}\frac{\partial L}{\partial q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T^*)} &= 0 \\ \Rightarrow -\frac{\sum_s p(\boldsymbol{z},t,s)}{q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T^*)} + \lambda(\boldsymbol{z}) = 0 \\ \Rightarrow \lambda(\boldsymbol{z}) q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T^*) = p(\boldsymbol{z},t)\end{aligned} \tag{5}$$

where we used the fact that, $\sum_s p(\boldsymbol{z},t,s) = p(\boldsymbol{z},t)$. Now summing w.r.t. to variable t on the both sides of last line and using the fact that $\sum_t q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T^*) = 1$ we get,

$$\lambda(\boldsymbol{z}) = p(\boldsymbol{z})$$

By substituting $\lambda(\boldsymbol{z})$ we obtain the solution of the optimal discriminator

$$q_T(t|\boldsymbol{z};\boldsymbol{\theta}_T^*) = \frac{p(\boldsymbol{z},t)}{p(\boldsymbol{z})} = p(t|\boldsymbol{z}) \tag{6}$$

Therefore,

$$q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*) = p(t|E(\boldsymbol{x};\boldsymbol{\theta}_E))$$

□

2. Proof of Corollary 1.1

Corollary 1.1. When $s \perp\!\!\!\perp t$, let the optimum discriminator and predictor for an encoder E be $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)$ and $q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)$ respectively. The optimal encoder $E(\cdot)$ in the MaxEnt-ARL formulation induces a uniform distribution in the discriminator $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E^*);\boldsymbol{\theta}_D^*)$ over the classes of the sensitive attribute.

Proof. Here we will prove that, when discriminator is fixed, then the encoder learns a representation of data $\boldsymbol{x}$ such that $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E^*);\boldsymbol{\theta}_D^*) = 1/m$. First we note that although the discriminator is fixed, the discriminator probability $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)$ can change by changing the encoder parameters $\boldsymbol{\theta}_E$. Optimization of the encoder in MaxEnt-ARL is formulated as:

$$\begin{aligned}&\min V = \min_{\boldsymbol{\theta}_E} \mathop{\mathbb{E}}_{(\boldsymbol{x},t,s)\sim p(\boldsymbol{x},t,s)} \left[-\log q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)\right] \\ &+ \alpha \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\sum_{i=1}^m q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) \log q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)\right] \\ &+ \log m \\ &\text{s.t. } \sum_{i=1}^m q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) = 1 \\ &\qquad q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) \geq 0, \quad \forall i\end{aligned} \tag{7}$$

The Lagrangian dual of the problem can be written as,

$$L = V - \lambda\left(\sum_{i=1}^m q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) - 1\right)$$

Here λ is a Lagrangian multiplier and is assumed to be a constant in the absence of any further information. Since $s \perp\!\!\!\perp t$, we have $q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)$ is independent of $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)$ given $E(\boldsymbol{x};\boldsymbol{\theta}_E)$ from Theorem 1. Therefore, if we take derivative of L w.r.t. $q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)$ and set it to zero we have:

$$\begin{aligned}\frac{\partial L}{\partial q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)} &= 0 \\ \Rightarrow 1 + \log\left(q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*\right) - \lambda = 0 \\ \Rightarrow q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) = \exp(\lambda - 1)\end{aligned} \tag{8}$$

Using the first (non-trivial) constraint, we have

$$\begin{aligned}\sum_{i=1}^m q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) = 1 \\ \sum_{i=1}^m \exp(\lambda-1) = 1 \\ \exp(\lambda-1)\sum_{i=1}^m 1 = 1 \\ m(\exp(\lambda-1)) = 1 \\ \lambda = \log(1/m) + 1\end{aligned}$$

Hence, the probability distribution of the discriminator after the encoder's parameters $\boldsymbol{\theta}_E$ are optimized is $q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E^*);\boldsymbol{\theta}_D^*) = 1/m$. Thus, when the optimum discriminator parameters are fixed, the encoder optimizes the representation such that the discriminator does not leak any information, i.e., it induces a uniform distribution. □

3. CIFAR Trade-Off

We report the numerical values of the target accuracy and adversary accuracy trade-off results on the CIFAR-10 and CIFAR-100 experiments in Table 1 and Table 3, respectively. Similarly, we report the numerical values of the target accuracy and adversary entropy trade-off results on the CIFAR-10 and CIFAR-100 experiments in Table 2 and Table 4, respectively.

Target Accuracy (%)	97.75	97.73	97.68
Adversary Accuracy (%)	23.44	23.09	22.68

(a) No Privacy

Target Accuracy (%)	97.52	97.44	97.35	91.52	91.15	60.00
Adversary Accuracy (%)	20.83	20.77	20.64	19.68	14.27	10.00

(b) ML-ARL

Target Accuracy (%)	97.78	97.74	97.53	96.79	95.01	92.34	61.17
Adversary Accuracy (%)	23.44	22.91	21.17	21.14	19.05	12.00	10.64

(c) MaxEnt-ARL

Table 1: CIFAR-10: Target Accuracy (%) vs Adversary Accuracy

Target Accuracy (%)	97.75	97.73	97.71
Adversary Entropy (nats)	1.65	1.65	1.67

(a) No Privacy

Target Accuracy (%)	97.52	97.50	96.58	95.97	60.00
Adversary Entropy (nats)	1.65	1.66	1.80	2.16	2.30

(b) ML-ARL

Target Accuracy (%)	97.78	97.74	97.58	97.53	97.14	96.79	95.76	92.34	61.17
Adversary Entropy (nats)	1.65	1.66	1.78	2.11	2.26	2.26	2.27	2.27	2.29

(c) MaxEnt-ARL

Table 2: CIFAR-10: Target Accuracy (%) vs Adversary Entropy

Target Accuracy (%)	71.99	71.56
Adversary Accuracy (%)	30.69	30.59

(a) No Privacy

Target Accuracy (%)	71.32	70.52	70.43	69.98	69.42	24.66	22.22	5.00
Adversary Accuracy (%)	15.43	15.09	14.84	14.60	14.41	6.81	6.72	1.00

(b) ML-ARL

Target Accuracy (%)	71.17	70.80	70.50	67.63	63.81	61.98	60.03	59.11	5.37	5.00
Adversary Accuracy (%)	16.88	16.60	16.43	13.23	8.38	5.02	3.80	2.81	1.23	1.00

(c) MaxEnt-ARL

Table 3: CIFAR-100: Target Accuracy (%) vs Adversary Accuracy

Target Accuracy (%)	71.99
Adversary Entropy (nats)	2.09

(a) No Privacy

Target Accuracy (%)	71.32	64.90	56.99	54.46	24.66	22.22	5.00
Adversary Entropy (nats)	2.50	2.51	2.68	2.88	3.77	3.88	4.60

(b) ML-ARL

Target Accuracy (%)	71.17	71.05	70.80	67.63	67.38	65.71	63.81	61.98	59.11	56.32	5.37	5.00
Adversary Entropy (nats)	2.27	2.28	2.31	2.91	3.01	3.24	4.14	4.56	4.57	4.57	4.59	4.60

(c) MaxEnt-ARL

Table 4: CIFAR-100: Target Accuracy vs Adversary Entropy