Supplementary Material: Obfuscation via Information Density Estimation

Hsiang Hsu
Harvard University
hsianghsu@g.harvard.edu

Shahab Asoodeh
Harvard University
shahab@seas.harvard.edu

Flavio P. Calmon
Harvard University
flavio@seas.harvard.edu

Here, we give proofs of theorems and other technical discussions omitted from Sections 2 and 3, and also provide further details about the experiment setups, the training phase, additional experiments on Gaussian synthetic data and discussion on limitation.

1 Proofs and Theoretical Backgrounds

In this section, we provide proofs omitted in the main text, as well as some discussions on the relationship between the TIDE and variational representations of f divergences and the Newey-McFadden lemma.

1.1 Proof of Theorem 1

For notational brevity, we drop y^{j-1} from the conditioning part of $P_{Y_j|s,y^{j-1}}$ and $P_{Y_j|y^{j-1}}$ and also write P and Q for $P_{Y_j|s}$ and P_{Y_j}, respectively. To prove this theorem, note that according to Definition 3, we can write

$$\mathsf{E}_{e^\varepsilon}(P\|Q) = P(i(s;Y_j)>\varepsilon) - e^\varepsilon Q(i(s;Y_j)>\varepsilon).$$

Hence, letting $\mathcal{C}$ denote the tail event $\{y : i(s;y)>\varepsilon\}$ for a given s, we have

$$\begin{aligned}
\mathsf{E}_{e^\varepsilon}(P\|Q) &= P(\mathcal{C}) - e^\varepsilon Q(\mathcal{C}) \\
&= \mathbb{E}_P\left[\mathbf{1}_{\{Y_j\in\mathcal{C}\}}\right] - e^\varepsilon \mathbb{E}_Q\left[\mathbf{1}_{\{Y_j\in\mathcal{C}\}}\right] \\
&\overset{(a)}{=} \mathbb{E}_Q\left[e^{i(s;Y_j)}\mathbf{1}_{\{Y_j\in\mathcal{C}\}}\right] - e^\varepsilon \mathbb{E}_Q\left[\mathbf{1}_{\{Y_j\in\mathcal{C}\}}\right] \\
&= \mathbb{E}_Q\left[\left(e^{i(s;Y_j)} - e^\varepsilon\right)\mathbf{1}_{\{Y_j\in\mathcal{C}\}}\right] \\
&= \mathbb{E}_Q\left[\left(e^{i(s;Y_j)} - e^\varepsilon\right)_+\right] \\
&= \mathbb{E}_Q\left[e^{i(s;Y_j)}e^{-i(s;Y_j)}\left(e^{i(s;Y_j)} - e^\varepsilon\right)_+\right] \\
&\overset{(b)}{=} \mathbb{E}_P\left[\left(1 - e^\varepsilon e^{-i(s;Y_j)}\right)_+\right] \\
&= \int_0^\infty \Pr\left(\left(1 - e^{\varepsilon - i(s;Y_j)}\right)1_{\{Y_j\in\mathcal{C}\}} \geq t\right)\mathrm{d}t,
\end{aligned}$$

where both (a) and (b) follow from the simple change-of-variable argument $\mathbb{E}_P\left[f(Y)\right] = \mathbb{E}_Q\left[e^{i(s;Y_j)}f(Y)\right]$ for any function f.

Furthermore, since $\left(1 - e^{\varepsilon - i(s;Y_j)}\right)1_{\{i(s;Y_j)>\varepsilon\}} < 1$ with probability one, we have

$$\begin{aligned}
\mathsf{E}_{e^\varepsilon}(P\|Q) &= \int_0^\infty \Pr\left(\left(1 - e^{\varepsilon - i(s;Y_j)}\right)1_{\{Y_j\in\mathcal{C}\}} \geq t\right)\mathrm{d}t \\
&= \int_0^1 \Pr\left(\left(1 - e^{\varepsilon - i(s;Y_j)}\right)1_{\{Y_j\in\mathcal{C}\}} \geq t\right)\mathrm{d}t \\
&= \int_0^1 \Pr\left(1 - e^{\varepsilon - i(s;Y_j)} \geq t\right)\mathrm{d}t \\
&= \int_0^1 \Pr\left(e^{-i(s;Y_j)} \leq (1-t)e^{-\varepsilon}\right)\mathrm{d}t \\
&= e^\varepsilon \int_0^{e^{-\varepsilon}} \Pr\left(e^{-i(s;Y_j)} \leq b\right)\mathrm{d}b \\
&= e^\varepsilon \int_\varepsilon^\infty e^{-t}\Pr\left(i(s;Y_j)\geq t\right)\mathrm{d}t.
\end{aligned}$$

1.2 Proof of Theorem 2

First assume that $m=2$. For any set $A\subset\mathcal{X}^2$ and $s\in\mathcal{S}$, we have

$$\begin{aligned}
P_{Y_1Y_2|s}(A) &= \sum_{y_1\in\mathcal{X}} P_{Y_1|s}(y_1)\Pr((y_1,Y_2)\in A|s) \\
&\leq \sum_{y_1\in\mathcal{X}} P_{Y_1|s}(y_1)\min\left\{1, e^\varepsilon\Pr((y_1,Y_2)\in A) + \delta'\right\} \\
&\leq \sum_{y_1\in\mathcal{X}} P_{Y_1|s}(y_1)\min\left\{1, e^\varepsilon\Pr((y_1,Y_2)\in A)\right\} + \delta' \\
&\leq \sum_{y_1\in\mathcal{X}} \left(e^\varepsilon P_{Y_1}(y_1) + \zeta(y_1)\right) \\
&\qquad \times \min\left\{1, e^\varepsilon\Pr((y_1,Y_2)\in A)\right\} + \delta' \\
&\leq \sum_{y_1\in\mathcal{X}} e^\varepsilon P_{Y_1}(y_1)\min\left\{1, e^\varepsilon\Pr((y_1,Y_2)\in A)\right\} \\
&\qquad + \sum_{y_1\in\mathcal{X}}\zeta(y_1) + \delta'
\end{aligned}$$

$$
\begin{aligned}
&\le e^{2\varepsilon} \sum_{y_1 \in \mathcal{X}} P_{Y_1}(y_1) \Pr((y_1, Y_2) \in A) + \sum_{y_1 \in \mathcal{X}} \zeta(y_1) + \delta' \\
&\le e^{2\varepsilon} P_{Y_1 Y_2}(A) + \sum_{y_1 \in \mathcal{X}} \zeta(y_1) + \delta' \\
&= e^{2\varepsilon} P_{Y_1 Y_2}(A) + \mathsf{E}_{e^\varepsilon}(P_{Y_1|s} \| P_{Y_1}) + \delta' \\
&\le e^{2\varepsilon} P_{Y_1 Y_2}(A) + 2\delta'
\end{aligned}
$$

where $\delta' = \frac{\delta}{2}$ and $\zeta(a) \coloneqq \left(P_{Y_1|s}(a) - e^\varepsilon P_{Y_1}(a)\right)_+$ for any $a \in \mathcal{X}$. The last step follows from the fact that $\mathsf{E}_\gamma(P\|Q) = \sum_{a\in\mathcal{X}}(P(a) - \gamma Q(a))_+$. Consequently, we obtain that

$$
P_{Y_1 Y_2|s}(A) \le e^{2\varepsilon} P_{Y_1 Y_2}(A) + \delta,
$$

for any set $A \subset \mathcal{X}^2$ for $m = 2$. Repeating this argument $(m-1)$ times, we can write

$$
P_{Y|s}(A) \le e^{m\varepsilon} P_Y(A) + \delta,
$$

for any set $A \subset \mathcal{X}^m$ and $s \in \mathcal{S}$ from which we conclude

$$
\mathsf{E}_{e^\varepsilon}(P_{Y|s} \| P_Y) \le \delta.
$$

1.3 Proof of Theorem 3

For any $\gamma \ge 1$ and $y^{j-1} \in \mathcal{X}^{j-1}$, we have

$$
\begin{aligned}
&\mathsf{E}_\gamma(P_{Y_j|s,y^{j-1}} \| P_{Y_j|y^{j-1}}) && \text{(S.1)} \\
&\quad \le \sup_{x^{j-1}} \mathsf{E}_\gamma(P_{Y_j|s,x^{j-1},y^{j-1}} \| P_{Y_j|x^{j-1},y^{j-1}}) \\
&\quad = \sup_{x^{j-1}} \mathsf{E}_\gamma(P_{Y_j|s,x^{j-1}} \| P_{Y_j|x^{j-1}}), && \text{(S.2)}
\end{aligned}
$$

where the inequality follows from the convexity of E_γ-divergence in each of its arguments (see, e.g., Sason and Verdú (2016)). Notice that for any given $x^{j-1} \in \mathcal{X}^{j-1}$, we can write (with an abuse of notation)

$$
\begin{aligned}
&\mathsf{E}_\gamma(P_{Y_j|s,x^{j-1}} \| P_{Y_j|x^{j-1}}) \\
&= \int_B [P(\mathrm{d}x_j|s,x^{j-1})\mathcal{N}(x_j,\lambda) - e^\varepsilon P(\mathrm{d}x_j|x^{j-1})\mathcal{N}(x_j,\lambda)]_+ \\
&\quad + \int_{B^\mathsf{c}} [P(\mathrm{d}x_j|s,x^{j-1}) - e^\varepsilon P(\mathrm{d}x_j|x^{j-1})]_+ \\
&= \int_B [P(\mathrm{d}x_j|s,x^{j-1})\mathcal{N}(x_j,\lambda) - e^\varepsilon P(\mathrm{d}x_j|x^{j-1})\mathcal{N}(x_j,\lambda)]_+
\end{aligned}
$$

where we use B and B^c to write $B_j^\varepsilon(x^{j-1})$ and its complement. This demonstrates that the mass points corresponding to the event B^c do not contribute in the E_γ-divergence.

Letting $P = P_{X_j|s,x^{j-1}}$ and $Q = P_{X_j|x^{j-1}}$ for a given x^{j-1}, it follows from above that

$$
\begin{aligned}
&\sup_{x^{j-1}} \mathsf{E}_\gamma(P_{Y_j|s,x^{j-1}} \| P_{Y_j|x^{j-1}}) \\
&\qquad \le \mathsf{E}_\gamma(P * \mathcal{N}(0,\lambda) \| Q * \mathcal{N}(0,\lambda)) \\
&\qquad = \mathbb{E}\left[\mathsf{E}_\gamma(\mathcal{N}(A,\lambda)\|\mathcal{N}(B,\lambda)\right], && \text{(S.3)}
\end{aligned}
$$

where $*$ denotes the convolution operator and $A \sim P$ and $B \sim Q$ and the expectation is taken over any arbitrary coupling of P and Q (e.g., their product). It can be shown that

$$
\begin{aligned}
&\mathsf{E}_\gamma(\mathcal{N}(\mu_1,\lambda^2\mathrm{I}_r)\|\mathcal{N}(\mu_2,\lambda^2\mathrm{I}_r)) \\
&\qquad = \mathsf{Q}\left(\frac{\log\gamma}{\beta} - \frac{1}{2}\beta\right) - \gamma\mathsf{Q}\left(\frac{\log\gamma}{\beta} + \frac{1}{2}\beta\right), && \text{(S.4)}
\end{aligned}
$$

where $\mathsf{Q}(v) = \Pr(\mathcal{N}(0,1) \ge v) = \int_v^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\mathrm{d}t$ and $\beta = \frac{\|\mu_1 - \mu_2\|}{\lambda}$. Notice that the E_γ-divergence between two Gaussian distributions depends on their means only through their differences.

$$
\theta_\gamma(a,\lambda) \triangleq \mathsf{E}_\gamma(\mathcal{N}(\mu,\lambda^2\mathrm{I}_r)\|\mathcal{N}(0,\lambda^2\mathrm{I}_r)),
$$

where $\|\mu\| = a$. According to (S.3), we can now write

$$
\sup_{x^{j-1}} \mathsf{E}_\gamma(P_{Y_j|s,x^{j-1}} \| P_{Y_j|x^{j-1}}) \le \sup_{a\in C} \theta_\gamma(\|a\|,\lambda) = \theta_\gamma(K,\lambda),
$$

where the equality is due to the fact that $a \mapsto \theta_\gamma(a,\lambda)$ is increasing for a fixed λ. This, together with (S.2), implies

$$
\mathsf{E}_\gamma(P_{Y_j|s,y^{j-1}} \| P_{Y_j|y^{j-1}}) \le \theta_\gamma(K,\lambda),
$$

and hence (14) is satisfied if $\theta_{e^\varepsilon}(K,\lambda) \le \frac{\delta}{m}$.

1.4 Estimating Information Density using f-Divergences

Other f-divergence measures could also be used to estimate the information density by leveraging their dual representation (Nguyen et al., 2010). Given a convex function f with $f(1) = 0$, the f-divergence $D_f(P\|Q) = \mathbb{E}_Q f\left(\frac{P}{Q}\right)$ can be expressed as

$$
D_f(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))], \quad \text{(S.5)}
$$

where $f^*(t) \triangleq \sup_{x\in\mathbb{R}}\{xt - f(t)\}$ is the Fenchel convex conjugate of f. It can be shown that the optimizer is the subdifferential $\partial f(\frac{P}{Q})$ which, in turn, is a non-decreasing function of $\frac{P}{Q}$. Thus, $D_f(P\|Q)$ is also a candidate loss function in density ratio estimation problems.

1.5 Newey-McFadden Lemma

Lemma 1 ((Newey and McFadden, 1994, Theorem 2.1)). *Given the extremum estimator* $\hat{a} = \operatorname{argmax}_{a\in\mathcal{A}} \Lambda_n(a)$, *if (i)* $\mathcal{A}$ *is compact; (ii) there exists a limiting function* $\Lambda(a)$ *such that* $\Lambda_n(a)$ *converges to* $\Lambda(a)$ *in probability over* $\mathcal{A}$*; (iii)* $\Lambda(a)$ *is continuous and has unique maximum at* $a = a^*$*, then* $\hat{a}$ *is a consistent estimator of* a^*.

1.6 Proof of Theorem 4

Let the objective function of the extremum estimator be

$$\Lambda_n(g) \triangleq \mathbb{E}_{P_{S_n,X_n}}[g(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[e^{g(S,X)}]. \quad \text{(S.6)}$$

We prove this theorem by checking the properties of $\Lambda_n(g)$ according to Lemma 1. First, since Θ is compact and the mappings g_θ are continuous, the images $\mathcal{G}(\Theta)$ is also compact. Second, by triangular inequality, for $g \in \mathcal{G}(\Theta)$, we have

$$\begin{aligned}
&|\Lambda_n(g) - (\mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{p_S p_X}[e^{g(S,X)}])| \\
&\le \sup_{g\in\mathcal{G}(\Theta)} |\mathbb{E}_{P_{S,X}}[g(S,X)] \\
&\qquad - \mathbb{E}_{P_{S_n,X_n}}[g(S,X)]| \qquad \text{(S.7)} \\
&+ \sup_{g\in\mathcal{G}(\Theta)} |\log \mathbb{E}_{P_S P_X}[g(S,X)] \\
&\qquad - \log \mathbb{E}_{P_{S_n}P_{X_n}}[g(S,X)]|.
\end{aligned}$$

Since the function g is uniformly bounded by M, i.e. $|g| \le M$ for all θ, s and x, and logarithm is Lipschitz continuous with constant e^M in the interval $[e^{-M}, e^M]$, we have

$$\begin{aligned}
&|\log \mathbb{E}_{P_S P_X}[g(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[g(S,X)]| \\
&\le e^M |\mathbb{E}_{P_S P_X}[g(S,X)] - \mathbb{E}_{P_{S_n}P_{X_n}}[g(S,X)]|. \qquad \text{(S.8)}
\end{aligned}$$

Moreover, since $\mathcal{G}$ is compact and g is continuous, the functions g and e^g satisfy the uniform law of large numbers (van de Geer, 2000). Thus, Given $\eta > 0$, there exists an integer N such that for all $n \ge N$ and with probability one,

$$\sup_{g\in\mathcal{G}(\Theta)} |\mathbb{E}_{P_{S,X}}[g(S,X)] - \mathbb{E}_{P_{S_n,X_n}}[g(S,X)]| \le \frac{\eta}{2}, \quad \text{(S.9)}$$

and

$$\begin{aligned}
&\sup_{g\in\mathcal{G}(\Theta)} |\log \mathbb{E}_{P_S P_X}[g(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[g(S,X)]| \\
&\le \frac{\eta}{2} e^{-M}. \qquad \text{(S.10)}
\end{aligned}$$

Summarizing (S.7)-(S.10), we have with probability one

$$\begin{aligned}
&|\Lambda_n(g) - (\mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}])| \\
&\le \eta. \qquad \text{(S.11)}
\end{aligned}$$

In other words, there exists a limiting function $\Lambda(g) = \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}]$ such that $\Lambda_n(g)$ converges to $\Lambda(g)$ in probability.

Third, since $\Lambda(g) = \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}]$ consists of linear combinations (expectations) and continuous mappings (logarithm and exponential) of the continuous function g, $\Lambda(g)$ is continuous. Moreover, $\Lambda(g)$ has a unique optimizer g^*. Therefore, by Lemma 1, we know that with probability one,

$$|\hat{g}_n(s,x) - \hat{g}_\theta(s,x)| \le \eta, \ \forall s \in \mathcal{S}, x \in \mathcal{X}, \quad \text{(S.12)}$$

giving the consistency of the information density estimator.

Note that a constant function g results in 0 for the objective (20); therefore, the constant function will not be selected as a possible solution for the optimization unless S and X are independent.

1.7 Proof of Theorem 5

By Hoeffding's inequality (Hoeffding, 1994), for all functions g bounded by M, i.e. $|g| \le M$, we have

$$\begin{aligned}
&\Pr\{|\mathbb{E}_{P_{S_n,X_n}}[g(S,X)] - \mathbb{E}_{P_{S,X}}[g(S,X)]| > \frac{\eta}{4}\} \\
&\le 2\exp\left(-\frac{2n^2(\frac{\eta}{2})^2}{(2M)^2 n}\right) = 2\exp\left(-\frac{n\eta^2}{32M^2}\right). \qquad \text{(S.13)}
\end{aligned}$$

Moreover, since g_θ is parameterized by θ, we utilize the union bound (Shalev-Shwartz and Ben-David, 2014, Lemma 2.2) to extend (S.13) for the parameters θ. For this purpose, recall that $\Theta \subset \mathbb{R}^d$ is compact and bounded by C, by the exterior covering number of bounded subspace (Shalev-Shwartz and Ben-David, 2014, pp. 337), we know the r-covering number $N(r,\Theta)$ of Θ is upper bounded by

$$N(r,\Theta) \le \left(\frac{2C\sqrt{d}}{r}\right)^d. \quad \text{(S.14)}$$

By (S.13) and (S.14), we have

$$\begin{aligned}
&\Pr\{\exists \theta_l \in \Theta \ s.t. \ \sup_{g_\theta} |\mathbb{E}_{P_{S_n,X_n}}[g_{\theta_l}(S,X)] \\
&- \mathbb{E}_{P_{S,X}}[g_{\theta_l}(S,X)]| > \frac{\eta}{4}\} \qquad \text{(S.15)} \\
&\le 2N(r,\Theta)\exp\left(-\frac{n\eta^2}{32M^2}\right).
\end{aligned}$$

where θ_l is in the r-cover of Θ. Since $\mathcal{G}(\Theta)$ is compact, we can replace the supremum by maximum. To make $2N(r,\Theta)\exp\left(-\frac{n\eta^2}{32M^2}\right) < \delta$, we have

$$n > \frac{32M^2(\log N(r,\Theta) + \log\frac{2}{\delta})}{\eta^2}. \quad \text{(S.16)}$$

Now, let $r = \frac{\eta}{8L}$, and recall that g_θ is L-Lipschitz continuous with respect to θ, then for any $\theta \in \Theta$, we have with probability one

$$|g_\theta - g_{\theta_l}| \le L|\theta - \theta_l| \le Lr = L \times \frac{\eta}{8L} = \frac{\eta}{8}. \quad \text{(S.17)}$$

By triangular inequality, for any $\theta \in \Theta$, whenever $n > \frac{32M^2(d\log\frac{16LC\sqrt{d}}{\eta}+\log\frac{2}{\delta})}{\eta^2}$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{g_\theta}|\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] - \mathbb{E}_{P_{S,X}}[g_\theta(S,X)]| \\
\leq\ &\max_{g_\theta}|\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] - \mathbb{E}_{P_{S_n,X_n}}[g_{\theta_l}(S,X)]| \\
&+\max_{g_\theta}|\mathbb{E}_{P_{S_n,X_n}}[g_{\theta_l}(S,X)] - \mathbb{E}_{P_{S,X}}[g_{\theta_l}(S,X)]| \\
&+\max_{g_\theta}|\mathbb{E}_{P_{S,X}}[g_\theta(S,X)] - \mathbb{E}_{P_{S,X}}[g_{\theta_l}(S,X)]| \\
\leq\ &\frac{\eta}{8}+\frac{\eta}{4}+\frac{\eta}{8} = \frac{\eta}{2}
\end{aligned} \tag{S.18}$$

Therefore, we have

$$\begin{aligned}
&\Pr\{\max_{g_\theta}|\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] \\
&-\mathbb{E}_{P_{S,X}}[g_\theta(S,X)]| \leq \frac{\eta}{2}\} \geq 1-\delta.
\end{aligned} \tag{S.19}$$

Similarly, starting from

$$\begin{aligned}
&\Pr\{\exists\theta_l \in \Theta\ s.t.\ |\log\mathbb{E}_{P_{S_n}P_{X_n}}[e^{g_{\theta_l}(S,X)}] \\
&-\log\mathbb{E}_{P_S P_X}[e^{g_{\theta_l}(S,X)}]| \geq \frac{\eta}{4}\} \\
&\leq 2N(r,\Theta)\exp\left(-\frac{n\eta^2}{32M^2}\right),
\end{aligned} \tag{S.20}$$

we also conclude that for any $\theta \in \Theta$, whenever $n > \frac{32M^2(d\log\frac{16LC\sqrt{d}}{\eta}+\log\frac{2}{\delta})}{\eta^2}$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\Pr\{\max_{g_\theta}|\log\mathbb{E}_{P_{S_n,X_n}}[E^{g_\theta(S,X)}] \\
&-\log\mathbb{E}_{P_{S,X}}[e^{g_\theta(S,X)}]| \leq \frac{\eta}{2}\} \geq 1-\delta.
\end{aligned} \tag{S.21}$$

Summarizing (S.19) and (S.21), whenever $n > \frac{32M^2(d\log\frac{16LC\sqrt{d}}{\eta}+\log\frac{2}{\delta})}{\eta^2}$, for any $\theta \in \Theta$, we have

$$\begin{aligned}
&\Pr\{|\max\Lambda_n(\hat{g}_n(s,x)) - \max\Lambda(g(s,x))| \leq \eta\} \\
&\geq \Pr\{\max_{g_\theta}|\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] \\
&\qquad -\mathbb{E}_{P_{S,X}}[g_\theta(S,X)]| \\
&\qquad +\max_{g_\theta}|\log\mathbb{E}_{P_{S_n,X_n}}[E^{g_\theta(S,X)}] \\
&\qquad -\log\mathbb{E}_{P_{S,X}}[e^{g_\theta(S,X)}]| \leq \eta\} \\
&\geq 1-\delta.
\end{aligned} \tag{S.22}$$

The thresholded information density estimator, in this sense, gives a thresholded (clipped) information density, i.e. $|\hat{g}_n(s,x) - g^*(s,x)| \leq \eta$ if $g^*(s,x) \leq M$ and $|\hat{g}_n(s,x) - g^*(s,x)| \geq \eta$ otherwise. By the concentration of the information density (Polyanskiy and Wu, 2014), we also know the probability that the information density is clipped is upper bounded, i.e.

$$\Pr\{|g^*(s,x)| \geq M\} \leq e^{-M}. \tag{S.23}$$

Therefore, whenever $n > \frac{32M^2(d\log\frac{16LC\sqrt{d}}{\eta}+\log\frac{2}{\delta})}{\eta^2}$, for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$, we have

$$\begin{aligned}
&\Pr\{|\hat{g}_n(s,x) - g^*(s,x)| \leq \eta\} \\
&\geq 1-\delta \geq 1-e^{-M},
\end{aligned} \tag{S.24}$$

by choosing $\delta \geq e^{-M}$, and the desire result follows.

2 Experimental Details

In this section, we provide detailed experimental setups including architecture of the function g in TIDE, training details for the experiments shown in the main text.

2.1 GENKI-4K Smiling Dataset

The GENKI-4K smiling dataset (MPLab, 2009) contains 2400 colorful images for training and 600 for test, where each image, viewed as X, is a 64×64 pixels face that is smiling ($S = 1$) or not ($S = 0$).

Since the inputs of the encoder TIDE are images, we use adopt a convolutional neural net with three convolutional layers, two fully-connected layers, and a readout layer. The convolutional layers have kernels with dimension $(5, 5, 3, 64)$, $(5, 5, 64, 64)$, and $(3, 3, 64, 128)$ respectively. After flatting the output of the third convolutional layer, we feed the output to two fully-connected layers with 384 and 192 neurons respectively. We train for 100 epochs using `AdagradOptimizer` with learning rate 0.0001 and batch size 256, and achieve $I(S, X) = 0.594 < H(S) = 1$ bits.

The adversary we used here is also a convolutional neural net with identical structure as the TIDE with the difference that the objective is the cross-entropy loss for classification, and is trained for 150 epochs using `AdagradOptimizer` with learning rate 0.005 and batch size 256.

2.2 Celebrity Attributes (CelebA) Dataset

The CelebA dataset (Liu et al., 2015) contains 202599 colorful images, where each image is a 218×178 pixels face of a celebrity with 40 distinct binary labels, including `smiling`, `gender`, `Arched Eyebrows`, etc. We select 100000 face images as X and the private attribute S as smiling or not.

Since the inputs of the encoder TIDE are images, we use adopt a convolutional neural net with five convolutional layers, two fully-connected layers, and a readout layer. The convolutional layers have kernels with dimension $(5, 5, 3, 64)$, $(5, 5, 64, 64)$, $(3, 3, 128, 128)$, $(3, 3, 128, 128)$, and $(3, 3, 64, 128)$ respectively. After

Table 1: WMAE of the information density estimation on Gaussian synthetic data ($M = 5$).

	Empirical Plug-In Estimator				Kernel Density Estimator				TIDE			
d \ ρ	0.0	0.1	0.2	0.5	0.0	0.1	0.2	0.5	0.0	0.1	0.2	0.5
1	0.466	0.509	1.092	1.821	0.252	0.434	0.973	1.395	**0.005**	**0.007**	**0.011**	**0.057**
10	5.305	7.613	9.704	18.245	2.869	4.076	6.698	11.496	**1.010**	**1.216**	**1.884**	**2.503**

flatting the output of the third convolutional layer, we feed the output to two fully-connected layers with 384 and 192 neurons respectively. We train for 100 epochs using `AdagradOptimizer` with learning rate 0.005 and batch size 64, and achieve $I(S, X) = 0.967 \approx H(S) = 1$ bits.

The adversaries we used for emotion and gender detection here are also convolutional neural nets with identical structure as the TIDE with the difference that the objective is the cross-entropy loss for classification. We train the adversaries for 300 epochs using `AdagradOptimizer` with learning rate 0.001 and batch size 2000.

2.3 Politically-Biased Tweets

We collect 75946 tweets from more than 20 online publishers (e.g. CNN, Bloomberg, New York Times) using the Twitter API, and determine its private attribute S as the political bias of being right-wing ($S = 0$) and left-wing ($S = 1$) according to Rachez (2017). We clean up the tweets to only keep meaningful terms (i.e. pieces of words), and use bag-of-words representation (Manning et al., 2010) to tokenize all the pieces of words for each tweet according to term frequency, ending up with 24657 words (x_j). We order the x_j by the order it appears in the training texts of the Tweets.

The TIDE is a simple feed-forward neural network consists of three hidden layers with `ReLU` activation with 100 neurons for each hidden layer, and a readout layer with 32 neurons. We train for 50 epochs using `AdagradOptimizer` with learning rate 0.005 and batch size 128, and achieve $I(S; X) = 0.645$ bits.

3 Additional Experiments on Synthetic Data

We apply the TIDE in Section 3 on Gaussian synthetic data to estimate the trimmed information density with limited number of samples and $M = 5$. We consider two d-dimensional multivariate standard Normal random variables S and X, with pairwise correlation $\text{corr}(S_i, X_j) = \rho \mathbf{1}_{\{i=j\}}$, $\rho \in (-1, 1)$, $1 \leq i, j, \leq d$. Since the KL divergence is invariant to continuous bijective transformations of the considered variables, it is sufficient to consider S and X with standard Normal marginals. We generate 3000 samples with $70\% - 30\%$ train-test split accordingly. The TIDE is a simple feed-forward neural network consists of three hidden layers with `ReLU` activation with 100, 50, 50 neurons for each hidden layer, and a readout layer with 50 neurons. We jointly train over the entire training set for 3000 epochs using `AdagradOptimizer` with learning rate 0.005.

We compare the plug-in estimator using empirical distributions (with 30 bins for quantization), the Gaussian kernel density estimator (Bishop, 2006), and the TIDE using 3k samples, and report the Weighted Mean Absolute Error (WMAE) of the information density in Table 1, where the weights are chosen as the ground true joint distributions and each number in the table is averaged over 10 repeated experiments. Note that since the Normal random variable is continuous, quantized empirical distribution gives loose estimate. The kernel density estimator performs better than the plug-in estimator but worse than the TIDE due to limited number of samples.

4 Final Remarks

We introduced a new information obfuscation framework that first identifies information-leaking features using the trimmed information density, and then tailors the obfuscation mechanism only on these features. To our knowledge, this framework is the first formal application of information density to quantify information-leaking features, and could potentially serve as a data-driven tool for designing obfuscation mechanism for high-dimensional data.

It is worth mentioning that information obfuscation, being inherently prior-dependent, has several limitations (Huang et al., 2017). In order to estimate the information density, we make two key assumptions: (i) we know *a priori* sensitive attributes that we wish to hide (e.g., political preference), and (ii) we have access to a reference dataset from which we can fit the TIDE (though this is difficult to avoid as discussed in Žliobaitė and Custers (2016)). Although these assumptions are restrictive in practice, they allow us to develop systematic machinery to discover information-leaking samples and features in an entirely data-driven manner.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Hoeffding, W. (1994). Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer.

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2017). Context-aware generative adversarial privacy. *Entropy*, 19(12):656.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proc. of International Conference on Computer Vision (ICCV)*.

Manning, C., Raghavan, P., and Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

MPLab, T. (2009). The MPLab GENKI Database, GENKI-4K Subset. `http://mplab.ucsd.edu`.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

Polyanskiy, Y. and Wu, Y. (2014). Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6:2012–2016.

Rachez, A. (2017). Predicting political bias with python. https://medium.com/linalgo/predict-political-bias-using-python-b8575eedef13. Accessed: 2019-03-21.

Sason, I. and Verdú, S. (2016). f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

Žliobaitė, I. and Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2):183–201.