Discovering Information-Leaking Samples and Features

Hsiang Hsu, Shahab Asoodeh, and Flavio P. Calmon
John A. Paulson School of Engineering and Applied Sciences, Harvard University
{hsianghsu@g, shahab@seas, flavio@seas}.harvard.edu

Abstract

Discovering samples or features which leak information about correlated private data is a key challenge in designing context-aware privacy mechanisms. In this paper, we propose a framework to discover information-leaking samples and features based on an information-theoretic quantity known as the information density. We provide an estimator, TIDE, to approximate the thresholded information density with a provable sample complexity. Our framework is then validated on two real-world datasets providing evidence that the TIDE can potentially be used as a building block to design privacy mechanisms targeting only those information-leaking samples and features.

1 Introduction

Different samples and features within a dataset may leak different levels of private information. For example, not all Tweets equally reveal a user's political preference, and not all pixels in face images equally disclose emotion (see Section 3). A privacy mechanism should ideally target only those samples and features that leak excessive amount of private information if disclosed. Given the targeted set of private attributes, these types of mechanisms, known as *context-aware* mechanisms [1], improve the utility by incorporating either complete (cf. information-theoretic privacy [2–6]) or partial (cf. generative adversarial privacy [1, 7]) knowledge of the underlying data distribution.

A natural, yet mostly overlooked, first step in designing context-aware privacy mechanisms is to discover *information-leaking* samples or features for a given set of private attributes. Besides the utility improvement, discovering these features may help in improving interpretability as well. For instance, in an attempt to hide emotion in a face image one may need to add stronger noise to mouth and eyes pixels than to background pixels.

We propose a novel methodology to identify information-leaking samples and features via an information-theoretic quantity known as the *information density*[1] [9, 10]. This quantity appears in several privacy definitions including information-theoretic privacy [3–5] as well as differential privacy (under the name of *privacy loss* variable) [11–16]. We consider a dataset $\mathcal{D} = \{(\mathbf{s}_n, \mathbf{x}_n)\}_{n=1}^N$, drawn i.i.d. from $P_{S,X}$, where $\mathbf{s}_n \in \mathcal{S} = \mathbb{R}^m$ and $\mathbf{x}_n \in \mathcal{X} = \mathbb{R}^k$ are the n^{th} private attribute (e.g. emotion) and data sample (e.g. a face image), respectively. Moreover, we use $\mathbf{x}_n^j$ to denote the j^{th} feature (i.e., coordinate) of $\mathbf{x}_n$ ($j \in \{1, \dots, k\}$). The information density of the n^{th} sample is then given by

$$i(\mathbf{s}_n; \mathbf{x}_n) \triangleq \log \frac{P_{S,X}(\mathbf{s}_n; \mathbf{x}_n)}{P_S(\mathbf{s}_n) P_X(\mathbf{x}_n)} = \log \frac{P_{S|X}(\mathbf{s}_n|\mathbf{x}_n)}{P_S(\mathbf{s}_n)}. \tag{1}$$

The feature information density is analogously defined as $i(\mathbf{s}_n; \mathbf{x}_n^j)$ for any $j \in \{1, \dots, k\}$. Intuitively, $|i(\mathbf{s}_n; \mathbf{x}_n)|$ evaluates the change of belief about $\mathbf{s}_n$ upon observing $\mathbf{x}_n$. This intuition leads us to

[1] This quantity is called the pointwise mutual information in natural language processing literature [8].

Preprint. Under review.

view information density as a score for identifying information-leaking samples and features. Since we might not have access to the underlying distribution $P_{S,X}$ in practice, we need to estimate the information density. As the expected value of information density is equal to the mutual information, such estimation problem is closely connected to the estimation of mutual information which is known to be challenging [17–19] unless an adequate parametric model is assumed [20]. The main difficulty lies in the unboundedness of the information density, which leads to high complexity for precise estimation. Nevertheless, we do not need to precisely estimate information density in our framework; instead, we only need to know which samples or features have $|i(\mathbf{s}_n;\mathbf{x}_n)|$ higher than a given threshold. Therefore, it is sufficient for us to consider the *thresholded* information density, a much easier estimation problem. Inspired by [21,22], we develop the thresholded information density estimator (TIDE), based on the variational representations of f-divergences [21,23]. By trading off the estimation of those unbounded information density, we are able to implement the TIDE on two real-world datasets to discover information-leaking features by neural networks.

In short, our main contributions include (i) designing an estimator of thresholded information density with provable guarantee (see Section 2), and (ii) experiments (see Section 3) that provide evidence that TIDE can potentially serve as a building block to design privacy mechanisms which target only those information-leaking samples and features. It is worth mentioning that context-aware privacy mechanisms, being inherently prior-dependent, have several limitations [1]. In Section 4, we address some of these limitations along with potential future directions.

Related Work The problem of balancing the competing objectives of providing meaningful information and inference, on one hand, and obfuscating sensitive information, on the other hand, has been recently investigated in [1,24,25]. Following the information-theoretic trend, these works exploit average measures (in particular mutual information) to obfuscate data to maintain privacy (of sensitive attributes) *on the average*. Our approach can be viewed as the *sample-based* version of these works in that we deal with each individual data sample separately and not on an average basis. Moreover, the approach of first discovering the information-leaking samples and then perturb those risky samples resembles, in essence, the instance-based additive mechanism of Nissim *et al*. [26] in the differential privacy setting.

As it involves estimating information density from samples, our approach is connected to the density ratio estimation problems [22,23,27], which are fundamental in various applications of machine learning and statistics, e.g. outlier detection [28], transfer learning [29], and generative adversarial nets [30]. A naïve way to estimate the density ratio is to use the plug-in estimator, that is, to estimate the empirical joint distribution $\hat{P}_{S,X}$ and marginals $\hat{P}_S$ and $\hat{P}_X$ and declare $\log \frac{\hat{P}_{S,X}(s,x)}{\hat{P}_S(s)\hat{P}_X(x)}$ as the estimated information density. However, this approach is known to perform poorly [20] unless adequate parametric models (e.g. linear [27], kernel [31], or exponential family [22] models) are assumed. The two closest approaches to thresholded information density estimation are (i) [23], which proposed using the variational representation of f-divergences to convert information density estimation into an optimization problem over finite-complexity set of functions and (ii) [22], which estimated the trimmed density ratio of variables from exponential family distributions. We adopt the idea of thresholding when solving the variational representation of f-divergences (see Section 2).

2 Estimating the Thresholded Information Density

We propose next a consistent and scalable estimator for the thresholded information density, the TIDE, and derive its sample complexity. The estimator is central to discovering information-leaking samples and features.

Thresholded Information Density Estimator (TIDE). The TIDE we proposed here is based on the variational representation of KL divergence[2], the so-called Donsker-Varadhan (DV) representation, that states

$$D(P_{S,X}\|P_SP_X) = \sup_{g:\mathcal{S}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_SP_X}[e^{g(S,X)}]. \tag{2}$$

Recall that $D(P_{S,X}\|P_SP_X)$ is equal to the mutual information $I(S;X)$ between S and X, which is in fact the expected information density $\mathbb{E}_{P_{S,X}}[i(S,X)]$. It can be shown that the maximizer g^*

[2]Other f-divergence measures could also be used by their dual representation, see Appendix A.

of the optimization problem (2) is given by the information density, i.e., $g^*(s,x) = i(s;x)$. Hence, the problem of estimating information density is equivalent to solving the optimization problem (2) given access to samples drawn from $P_{S,X}$.

Since the search space in (2) is unconstrained, directly solving the optimization by computing the empirical expectations would fail in general. One practical approach is to restrict the search space to a family $\mathcal{G}(\Theta)$ of continuous and bounded functions g_θ parameterized by θ in a compact domain $\Theta \subset \mathbb{R}^d$, where d is the number of parameters. The new constrained optimization problem corresponds to approximating the information density by a bounded function, thus the name *thresholded* information density. The thresholded information density estimator (TIDE) is then given by

$$\hat{g}_n \triangleq \underset{g_\theta \in \mathcal{G}(\Theta)}{\operatorname{argmax}} \mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[e^{g_\theta(S,X)}], \tag{3}$$

where P_{S_n,X_n} and $P_{S_n}P_{X_n}$ denote the empirical distributions of $P_{S,X}$ and $P_S P_X$, respectively.

2.1 Consistency and Sample Complexity of the TIDE

The TIDE obtained by solving (3) belongs to a broader class of *extremum estimators* [32] of the form $\hat{a} = \operatorname{argmax}_{a\in\mathcal{A}} \Lambda_n(a)$, where $\Lambda_n(a)$ is an objective function and $\mathcal{A}$ is a parameter space. The consistency of extremum estimators is guaranteed by the Newey-McFadden Lemma (see Appendix B), which in turn implies the consistency of the TIDE (see [33, Proposition 3]). In this paper, we further turn our attention to deriving the sample complexity of the TIDE. We make further assumption that functions in $\mathcal{G}(\Theta)$ are Lipschitz, and use the concentration inequality[3] [34] of the information density to prove the following theorem (see proof in Appendix C). To avoid technical complications, we assume that $\mathbb{E}_{P_{S,X}}[g(S,X)]$ and $\mathbb{E}_{P_S P_X}[e^{g(S,X)}]$ are finite for all functions g in $\mathcal{G}(\Theta)$.

Proposition 1 (Sample Complexity)**.** *Assume that functions in $\mathcal{G}(\Theta)$ are bounded by M and Lipschitz with respect to θ, and $\Theta \subset \mathbb{R}^d$ is compact. Then we have $|\hat{g}_n(s,x) - g^*(s,x)| \leq \eta$ with probability at least $1 - e^{-M}$, for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$, where $n = O(\frac{M^2 d(\log d - \log \eta + M)}{\eta^2})$.*

Observe that thresholding the information density is crucial for the bound in the previous theorem to hold: if $M \to \infty$ (i.e., estimating the true information density), the sample complexity of the TIDE grows to infinity and the result is vacuous.

2.2 Implementation

For practitioners' purpose, we use the set of functions representable by a neural network with output clipped to $[-M, M]$ to approximate the set of continuous and bounded functions in $\mathcal{G}$. By sampling $(\mathbf{s}_n, \mathbf{x}_n)$ from $P_{S,X}$ and $(\mathbf{s}_n, \mathbf{x}_n)$ from $P_S \times P_X$ for the first and second expectations in (3), we can back-propagate on the neural network, and after training, the TIDE outputs the estimate of the thresholded information density of samples $|i(\mathbf{s}_n;\mathbf{x}_n)| \leq M$ and of features $|i(\mathbf{s}_n;\mathbf{x}_n^j)| \leq M$.

3 Experiments

We validate the implementation of the TIDE in Section 2.2 with a focus on discovering information-leaking features on two real-world datasets, (i) detecting emotion-leaking pixels in GENKI-4K dataset [35], and (ii) discovering politically-charged terms in the Tweets of online media [36]. Throughout the experiments, we set $M = 2.00$. For experiments on identifying information-leaking samples, see [33].

GENKI-4K Smiling Dataset. This dataset contains 2400 images for training and 600 for testing, where each image $(\mathbf{x}_n)$ is a 64×64 pixels (each pixel is a feature $\mathbf{x}_n^j$) face that is smiling ($S = 1$) or not ($S = 0$), We train the TIDE and achieve $I(S;X) = 0.594$ bits. We select 10 faces from the test set for illustration in Figure 1 row (a). In order to estimate $i(\mathbf{s}_n;\mathbf{x}_n^j)$ for each pixels $\mathbf{x}_n^j$, we create an artificial image in which we keep the values of the original image within a patch of size 3×3 pixels, say $\mathbf{x}_n^r, \mathbf{x}_n^{r+1}, \ldots, \mathbf{x}_n^{r+8}$, and set the rest pixels to be zero. Running r from 1 to $k - 8$, we scan the entire image with the patch, and feed each artificial image into the TIDE to estimate the thresholded information density of the patch, i.e. $i(\mathbf{s}_n;\mathbf{x}_n^r, \mathbf{x}_n^{r+1}, \ldots, \mathbf{x}_n^{r+8})$ for all r. Then $i(\mathbf{s}_n;\mathbf{x}_n^j)$

[3] In other words, $\Pr\{i(S;\mathbf{x}_n) > t\} \leq e^{-t}, \forall \mathbf{x}_n$.

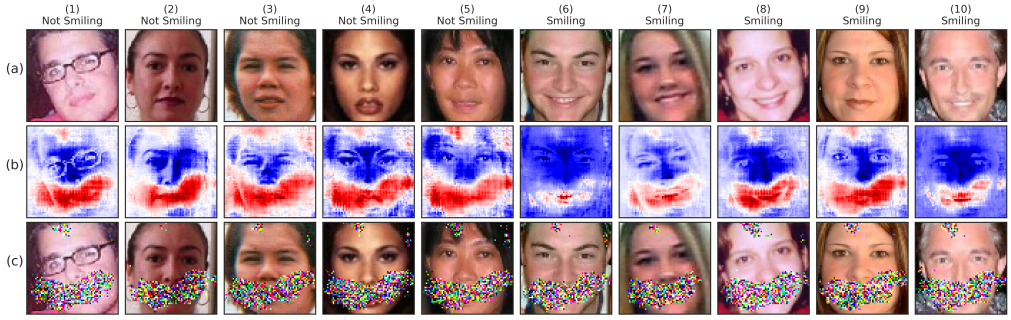

Figure 1: The TIDE on the GENKI-4K smiling dataset. Row (a): original images, row (b): information-leaking pixels (features) (red parts indicate higher thresholded information density), row (c): applying standard Gaussian noise only on pixels with high thresholded information density clearly hides the private information while preserving utility (e.g. gender).

is set to be the average of $i(\mathbf{s}_n; \mathbf{x}_n^r, \mathbf{x}_n^{r+1}, \ldots, \mathbf{x}_n^{r+8})$ for all patches which contain $\mathbf{x}_n^j$. We report $i(\mathbf{s}_n; \mathbf{x}_n^j)$ for each pixel of an image in Figure 1 row (b) which indicate pixel with high privacy risk. Note that the TIDE can not only reveal the pixels informative to smiling (mostly pixels that compose the mouths), but also captures the contour of faces. We add Gaussian noise to those pixels with thresholded information density higher than 0.9, and show the resulting images in Figure 1 row (c), which hides the private attribute while preserving other useful information in the image that is irrelevant of smiling. For example, we can still identify the gender of the people in Figure 1 row (c).

Political Preferences of Tweets. We collect 75946 Tweets from more than 20 online publishers (e.g. CNN, Bloomberg, New York Times), and determine their private attribute S as the political preference of being right-wing ($\mathbf{s}_n = 0$) and left-wing ($\mathbf{s}_n = 1$) according to [36], where the numbers of samples with each political bias are equivalent. We pre-process the Tweets to keep only meaningful terms (i.e. pieces of words) and use bags-of-words model [37] representation to tokenize all the pieces of words for each Tweet according to term frequency, ending up with 24657 terms (i.e. features $\mathbf{x}_n^j$, $j \in \{1, \cdots, 24657\}$). We train the TIDE using the tokenized Tweets as $\mathbf{x}_n$ and achieve $I(S; X) = 0.645$ bits. In Figure 2, we show the estimate of thresholded information density of each terms $i(\mathbf{s}_n; \mathbf{x}_n^j)$, and some politically-charged terms that would reveal the political bias.

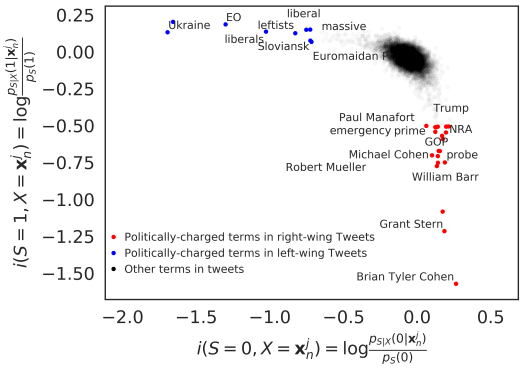

Figure 2: Left: Outputs from TIDE for terms in Tweets. GOP: Grand Old Party (i.e. the Republican Party), NRA: National Rifle Association, EO: Entrepreneurs' Organization, Euromaidan Pr.: Euromaidan Press.

4 Final Remark

We discuss the limitations and future directions in the following.

Limitations. In order to estimate the information density, we make two key assumptions: (i) we know *a priori* private attributes that we wish to hide (e.g., political preference), and (ii) we have access to a reference dataset from which we can train machine learning models (though this is difficult to avoid as discussed in [38]). Although these assumptions are restrictive in practice, they allow us to develop a systematic machinery to discover information-leaking samples and features in an entirely data-driven and automated manner.

Future Directions. We ideally seek to design privacy-assuring mechanisms that go beyond the indiscriminate addition of noise or uniform randomization on a whole dataset based on the TIDE. In Figure 1 row (c), we employ a primitive mechanism that applies standard Gaussian noise merely on information-leaking pixels; however, how to design "optimal" perturbations or randomization with provable privacy guarantees remains open and could be a topic worthy of further study.

References

[1] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, no. 12, p. 656, 2017.

[2] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. of IEEE Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1401–1408.

[3] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1512–1534, 2018.

[4] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *arXiv preprint arXiv:1807.07878*, 2018.

[5] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2018.

[6] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar, "On the robustness of information-theoretic privacy measures and mechanisms," *arXiv preprint arXiv:1811.06057*, 2018.

[7] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Generative adversarial privacy," *arXiv preprint arXiv:1807.05306*, 2018.

[8] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[9] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[10] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, 1993.

[11] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. of ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2003.

[12] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 635–658.

[13] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *arXiv preprint arXiv:1603.01887*, 2016.

[14] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *in Proc. of the International Conference on Machine Learning*. Proceedings of Machine Learning Research, 2018.

[15] A. D. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal processing magazine*, vol. 30, no. 5, pp. 86–94, 2013.

[16] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[17] G. Valiant and P. Valiant, "Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts," in *Proc. of ACM symposium on Theory of computing (STOC)*, 2011.

[18] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

[19] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5986–5997.

[20] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[21] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. Courville, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[22] S. Liu, A. Takeda, T. Suzuki, and K. Fukumizu, "Trimmed density ratio estimation," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[23] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[24] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, and G. Sapiro, "Adversarially learned representations for information obfuscation and inference," in *Proc. of International Conference on Machine Learning (ICML)*, 2019.

[25] X. Chen, T. Navidi, S. Ermon, and R. Rajagopal, "Distributed generation of privacy preserving data with user customization," *arXiv preprint arXiv:1904.09415*, 2019.

[26] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 2007, pp. 75–84.

[27] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, "Relative density-ratio estimation for robust distribution comparison," in *Proc. of Advances in neural information processing systems (NeurIPS)*, 2011.

[28] A. Smola, L. Song, and C. H. Teo, "Relative novelty detection," in *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

[29] M. Sugiyama, M. Krauledat, and K.-R. MÃžller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research (JMLR)*, vol. 8, no. May, pp. 985–1005, 2007.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[31] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[32] T. Amemiya, "Asymptotic properties of extremum estimators," *Advanced econometrics, Harvard university press*, 1985.

[33] H. Hsu, S. Asoodeh, and F. P. Calmon, "Information-theoretic privacy watchdogs," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2019.

[34] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes for ECE563 (UIUC) and*, vol. 6, pp. 2012–2016, 2014.

[35] "The MPLab GENKI Database, GENKI-4K Subset," http://mplab.ucsd.edu, 2009.

[36] "Predicting political bias with python," https://medium.com/linalgo/predict-political-bias-using-python-b8575eedef13, accessed: 2019-03-21.

[37] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[38] I. Žliobaitė and B. Custers, "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models," *Artificial Intelligence and Law*, vol. 24, no. 2, pp. 183–201, 2016.

[39] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," *Handbook of econometrics*, vol. 4, pp. 2111–2245, 1994.

[40] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *The Collected Works of Wassily Hoeffding*. Springer, 1994, pp. 409–426.

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

Appendices

A Estimating Information Density using f-Divergences

Other f-divergence measures could also be used to estimate the information density by leveraging their dual representation [23]. 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

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))], \tag{A.1}$$

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.

B Newey-McFadden Lemma

Lemma 1 ([39, 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^*.*

C Proof of Proposition 1

By Hoeffding's inequality [40], for all functions g bounded by M, i.e. $|g|\leq M$, we have

$$\Pr\{|\mathbb{E}_{P_{S_n,X_n}}[g(S,X)] - \mathbb{E}_{P_{S,X}}[g(S,X)]|> \frac{\eta}{4}\} \leq 2\exp\left(-\frac{2n^2(\frac{\eta}{2})^2}{(2M)^2 n}\right) = 2\exp\left(-\frac{n\eta^2}{32M^2}\right). \tag{C.1}$$

Moreover, since g_θ is parameterized by θ, we utilize the union bound [41, Lemma 2.2] to extend (C.1) 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 [41, pp. 337], we know the r-covering number $N(r,\Theta)$ of Θ is upper bounded by

$$N(r,\Theta) \leq \left(\frac{2C\sqrt{d}}{r}\right)^d. \tag{C.2}$$

By (C.1) and (C.2), we have

$$\begin{aligned} & \Pr\{\exists\theta_l \in \Theta \; s.t. \; \sup_{g_\theta} |\mathbb{E}_{P_{S_n,X_n}}[g_{\theta_l}(S,X)] - \mathbb{E}_{P_{S,X}}[g_{\theta_l}(S,X)]|> \frac{\eta}{4}\} \\ \leq \; & 2N(r,\Theta)\exp\left(-\frac{n\eta^2}{32M^2}\right). \end{aligned} \tag{C.3}$$

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

$$n > \frac{32M^2(\log N(r,\Theta) + \log\frac{2}{\delta})}{\eta^2}. \tag{C.4}$$

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

$$|g_\theta - g_{\theta_l}| \leq L|\theta - \theta_l| \leq Lr = L \times \frac{\eta}{8L} = \frac{\eta}{8}. \tag{C.5}$$

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 $1-\delta$,

$$\begin{aligned} & \max_{g_\theta} |\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] - \mathbb{E}_{P_{S,X}}[g_\theta(S,X)]| \\ \leq \; & \max_{g_\theta} |\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] - \mathbb{E}_{P_{S_n,X_n}}[g_{\theta_l}(S,X)]| \\ & + \max_{g_\theta} |\mathbb{E}_{P_{S_n,X_n}}[g_{\theta_l}(S,X)] - \mathbb{E}_{P_{S,X}}[g_{\theta_l}(S,X)]| \\ & + \max_{g_\theta} |\mathbb{E}_{P_{S,X}}[g_\theta(S,X)] - \mathbb{E}_{P_{S,X}}[g_{\theta_l}(S,X)]| \\ \leq \; & \frac{\eta}{8} + \frac{\eta}{4} + \frac{\eta}{8} = \frac{\eta}{2} \end{aligned} \tag{C.6}$$

Therefore, we have

$$\Pr\{\max_{g_\theta}|\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)]-\mathbb{E}_{P_{S,X}}[g_\theta(S,X)]|\leq\frac{\eta}{2}\}\geq 1-\delta. \tag{C.7}$$

Similarly, starting from

$$\begin{aligned}&\Pr\{\exists\theta_l\in\Theta\ s.t.\ |\log\mathbb{E}_{P_{S_n}P_{X_n}}[e^{g_{\theta_l}(S,X)}]-\log\mathbb{E}_{P_SP_X}[e^{g_{\theta_l}(S,X)}]|\geq\frac{\eta}{4}\}\\ \leq\ & 2N(r,\Theta)\exp\left(-\frac{n\eta^2}{32M^2}\right),\end{aligned} \tag{C.8}$$

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 $1-\delta$,

$$\Pr\{\max_{g_\theta}|\log\mathbb{E}_{P_{S_n,X_n}}[E^{g_\theta(S,X)}]-\log\mathbb{E}_{P_{S,X}}[e^{g_\theta(S,X)}]|\leq\frac{\eta}{2}\}\geq 1-\delta. \tag{C.9}$$

Summarizing (C.7) and (C.9), 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

$$\begin{aligned}&\Pr\{|\max\Lambda_n(\hat{g}_n(s,x))-\max\Lambda(g(s,x))|\leq\eta\}\\ \geq\ &\Pr\{\max_{g_\theta}|\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)]-\mathbb{E}_{P_{S,X}}[g_\theta(S,X)]|\\ &\qquad+\max_{g_\theta}|\log\mathbb{E}_{P_{S_n,X_n}}[E^{g_\theta(S,X)}]-\log\mathbb{E}_{P_{S,X}}[e^{g_\theta(S,X)}]|\leq\eta\}\\ \geq\ &1-\delta.\end{aligned} \tag{C.10}$$

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 [34], we also know the probability that the information density is clipped is upper bounded, i.e.

$$\Pr\{|g^*(s,x)|\geq M\}\leq e^{-M}. \tag{C.11}$$

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

$$\Pr\{|\hat{g}_n(s,x)-g^*(s,x)|\leq\eta\}\geq 1-\delta\leq 1-e^{-M}, \tag{C.12}$$

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