our bounds are ef ciently computable for empirical/mixture distributions via reformulation as a linear program.

Finally, we remark that analzing the D_ϵ optimal transport cost may be interesting in itself. The optimal transport cost $c_\epsilon(x, x') = \mathbb{1}\{d(x, x') > 2\epsilon\}$ is discontinuous and does not satisfy triangle inequality. This makes it hard to analyse D_ϵ using standard techniques in optimal transport literature. For instance, it would be interesting to see how fast D_ϵ between empirical distributions converges to D_ϵ between the true data-generating distributions. This may be used to obtain nite-sample lower bounds for adversarial error. Recent work (Jog, 2020) in this line of research derives sample complexity bounds for estimating D_ϵ from empirical distributions using a reverse Gaussian isoperimetric inequality for sets of the form A^ϵ. Another recent work (Yu, 2019) implies a sharp threshold for the asymptotics of D_ϵ between product distributions in terms of the 1-Wasserstein metric.

References

A., S., Huang, W. R., Studer, S., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? *International Conference on Learning Representations*, 2019.

Athalye, A., Carlini, N., and A., W. D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. *Algorithmic Learning Theory*, 2018.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. *Conference on Neural Information Processing Systems*, 2019.

Bhattacharjee, R. and Chaudhuri, K. When are non-parametric methods robust? In *International Conference on Machine Learning*, 2020. To appear.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3 14. ACM, 2017.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, 2017.

Cohen, J., Rosenfeld, E., and Kolter, J. Z. Certi ed adversar - ial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

Cullina, D., Bhagoji, A. N., and Mittal, P. PAC-learning in the presence of adversaries. In *Conference on Neural Information Processing Systems*, pp. 230 241, 2018.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. Adversarial risk and robustness: General de nitions and implications for the uniform distribution. In *Conference on Neural Information Processing Systems*, 2018.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. Lower bounds for adversarially robust PAC learning. *arXiv preprint arXiv:1906.05815*, 2019.

Fawzi, A., Fawzi, H., and Fawzi, O. Adversarial vulnerability for any classi er. In *Conference on Neural Information Processing Systems*, 2018.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT Press, 2016.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gourdeau, P., Kanade, V., Kwiatkowska, M., and Worrell, J. On the hardness of robust classi cation. *Conference on Neural Information Processing Systems*, 2019.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classi er against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pp. 2266 2276, 2017.

Holmstrom, L. and Koistinen, P. Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks*, 3(1):24 38, 1992.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Conference on Neural Information Processing Systems*, 2019.

Jog, V. Reverse Lebesgue and Gaussian isoperimetric inequalities for parallel sets with applications. *arXiv preprint arXiv:2006.09568*, 2020.

Khim, J. and Loh, P.-L. Adversarial risk bounds for binary classi cation via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.