# Supplementary Material: Lower Bounds on Cross-Entropy Loss in the Presence of Test-time Adversaries

Arjun Nitin Bhagoji *[1] Daniel Cullina *[2] Vikash Sehwag [3] Prateek Mittal [3]

## A. Proof for Lemma 3 and Theorem 2

We consider the case when the data is generated from a mixture of two Gaussians with identical covariances and means that differ in their sign. Formally, we have $P_Y(1) = p_1$, $P_Y(-1) = p_{-1}$, and $P_{X|Y=y} = \mathcal{N}(y\mu, \Sigma)$. $\mathcal{X}$ is then $\mathbb{R}^d$. We set the neighborhood function $N(x) = x + \epsilon\Delta$, where $\epsilon$ is the adversarial budget and $\Delta \in \mathbb{R}^d$ is a closed, convex, absorbing and origin-symmetric set.

*Proof.* Let $c = \log \frac{p_1}{p_{-1}}$ so $yc = \log \frac{p_y}{p_{-y}}$. Let $w \in \mathbb{R}^d$ and consider the classifier

$$\begin{aligned} h(x)_y &= \frac{1}{1 + \exp(-y(w^\top x + c))} \\ &= \frac{p_y \exp(\frac{y}{2} w^\top x)}{p_y \exp(\frac{y}{2} w^\top x) + p_{-y} \exp(\frac{-y}{2} w^\top x)}. \end{aligned}$$

The output probability $h(x)_y$ is an increasing function of $yw^\top x$, so we can find $q_{(x,y)} = \inf_{\tilde{x}\in N(x)} h(\tilde{x})_y$ by computing $\inf_{\tilde{x}\in N(x)} yw^\top \tilde{x} = yw^\top x - \sup_{z\in\epsilon\Delta} w^\top z = yw^\top x - \epsilon\|w\|_\Delta^*$. Thus adversarial log loss of this classifier is

$$\sum_y p_y \mathbb{E}_{X\sim\mathcal{N}(y\mu,\Sigma)} \log(1 + \exp(-y(w^\top X + c) + \epsilon\|w\|_\Delta^*))$$

where $X \sim \mathcal{N}(y\mu, \Sigma)$ and this is an upper bound on the optimal adversarial log loss. Observe that

$$yw^\top X - \epsilon\|w\|_\Delta^* \sim \mathcal{N}(w^\top \mu - \epsilon\|w\|_\Delta^*, w^\top \Sigma w).$$

For any $z \in \Delta$, the distributions $P_{\tilde{X}|Y=y} = \mathcal{N}(y(\mu - z), \Sigma)$ are clearly feasible for the adversary. The Bayes classifier for these is

$$h(x)_y = \frac{1}{1 + \exp(-y(2(\mu - z)^\top \Sigma^{-1} x + c))}.$$

The log loss of this classifier is

$$\sum_y p_y \mathbb{E} \log(1 + \exp(-y(2(\mu - z)\Sigma^{-1} X + c)))$$

where $X \sim \mathcal{N}(y(\mu - z), \Sigma)$ and this is an lower bound on the optimal adversarial log loss. Observe that

$$\begin{aligned} &2y(\mu - z)^\top \Sigma^{-1} X \sim \\ &\quad \mathcal{N}(2(\mu - z)^\top \Sigma^{-1}(\mu - z), 4(\mu - z)^\top \Sigma^{-1}(\mu - z)). \end{aligned}$$

If we can find $w$ and $z$ such that

$$\begin{aligned} w^\top \mu - \epsilon\|w\|_\Delta^* &= 2(\mu - z)^\top \Sigma^{-1}(\mu - z) \\ w^\top \Sigma w &= 4(\mu - z)^\top \Sigma^{-1}(\mu - z), \end{aligned}$$

then these upper and lower bounds match.

Using Lemma 1 from (Bhagoji et al., 2019), if we take $z$ to be the solution to optimization problem

$$\min (\mu - z)^\top \Sigma^{-1} (\mu - z) \text{ s.t. } z \in \epsilon\Delta$$

and $w = 2\Sigma^{-1}(\mu - z)$, then $\epsilon\|w\|_\Delta^* = w^\top z$, which immediately implies the desired equalities. □

## B. Proofs for Algorithm 1

### B.1. Proof of Lemma 4

*Proof.* Because each edge contains exactly one vertex in each of $\mathcal{A}$ and $\mathcal{B}$, $M\mathbf{1}_\mathcal{A} = \mathbf{1}_{\mathcal{E}\cup\mathcal{A}}$ and $M\mathbf{1}_\mathcal{B} = \mathbf{1}_{\mathcal{E}\cup\mathcal{B}}$. This gives two feasible choices for $y$: $y = \mathbf{1}_\mathcal{A}$ and $y = \mathbf{1}_\mathcal{B}$. By construction of $r$, at least one of these achieves a value of $P(\mathcal{A} \cup \mathcal{B})$. If $P(\mathcal{A}) > 0$ then

$$r^\top \mathbf{1}_\mathcal{A} = \sum_{v\in\mathcal{A}} \frac{P(\mathcal{A} \cup \mathcal{B})}{P(\mathcal{A})} p_v = P(\mathcal{A} \cup \mathcal{B})$$

and if $P(\mathcal{B}) > 0$ then $r^\top \mathbf{1}_\mathcal{B} = P(\mathcal{A} \cup \mathcal{B})$. Complementary slackness implies $(M^\top z - r)^\top y = 0$ for all optimal $y$, and thus $(M^\top z - r)_v = 0$ for all $v$ that are nonzero in some optimal $y$. If the feasible points $y = \mathbf{1}_\mathcal{A}$ and $y = \mathbf{1}_\mathcal{B}$ are optimal, then $(M^\top z)_v = r_v$ for all $v \in \mathcal{A}$ if $P(\mathcal{A}) > 0$ and for all $v \in \mathcal{B}$ if $P(\mathcal{B}) > 0$. Then Property 2 follows from the definition of $r$.

*Equal contribution [1]Department of Computer Science, University of Chicago [2]Department of Electrical and Computer Engineering, Pennsylvania State University [3]Department of Electrical Engineering, Princeton University. Correspondence to: Arjun Nitin Bhagoji <abhagoji@uchicago.edu>.

*Proceedings of the 38$^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

By strong linear programming duality, we always find $z$ and $y$ such that $\mathbf{1}^\top z = r^\top y$. If the candidate choices of $y$ described above are optimal, we satisfy the first alternative of the claim. Otherwise, we have $y$ such that $r^\top y > \mathbf{1}^\top p$. We have

$$\begin{aligned} r^\top(\mathbf{1}_{\mathcal{A}_+} + \mathbf{1}_{\mathcal{B}_+}) &> \mathbf{1}^\top p \\ \frac{P(\mathcal{A} \cup \mathcal{B})P(\mathcal{A}^+)}{P(\mathcal{A})} + \frac{P(\mathcal{A} \cup \mathcal{B})P(\mathcal{B}^+)}{P(\mathcal{B})} &> P(\mathcal{A} \cup \mathcal{B}) \\ P(\mathcal{A}^+)P(\mathcal{B}) + P(\mathcal{B}^+)P(\mathcal{A}) &> P(\mathcal{A})P(\mathcal{B}) \\ P(\mathcal{A}^+)P(\mathcal{B}^+) &> P(\mathcal{A}^-)P(\mathcal{B}^-) \end{aligned}$$

which establishes Property 1. □

### B.2. Proof of Lemma 5

*Proof.* In the base case of the induction, the output of OptProb comes from the second branch, the computation terminates, and $P(\mathcal{A}^+)P(\mathcal{B}^+) \leq P(\mathcal{A}^-)P(\mathcal{B}^-)$. We take $k = 1$, so $\mathcal{A} = \mathcal{A}_0$ and $\mathcal{B} = \mathcal{B}_0$. From Property 1 of Lemma 4, $P(\mathcal{A}^+)P(\mathcal{B}^+) = P(\mathcal{A}^-)P(\mathcal{B}^-)$ and thus from Property 2a we have $\mathbf{1}^\top z = P(\mathcal{A} \cup \mathcal{B})$ Then $q$ is specified by Line **??** and satisfies Property 3 by construction. Properties 2b and 2c of Lemma 4 implies $q_v(M^\top z)_v = P(\{v\})$, so Property 2 is established. Properties 1 and 4 hold trivially when $k = 1$.

In the inductive case, the output of OptProb comes from the first branch. Because $P(\mathcal{A}^+)P(\mathcal{B}^+) > 0$, both $\mathcal{A}^+$ and $\mathcal{B}^+$ are nonempty. Thus $|\mathcal{A}^+ \cup \mathcal{B}^-| < |\mathcal{A} \cup \mathcal{B}|$ and $|\mathcal{A}^- \cup \mathcal{B}^+| < |\mathcal{A} \cup \mathcal{B}|$, so the recursive calls both involve strictly smaller vertex sets. By induction, both recursive calls terminate. Suppose that $(q', z')$ and $(q'', z'')$ satisfy the four properties with functions $a' : \mathcal{A}^+ \to [k']$ and $b' : \mathcal{B}^- \to [k']$ and $a'' : \mathcal{A}^- \to [k'']$ and $b'' : \mathcal{B}^+ \to [k'']$ respectively. Then we take $k = k' + k''$ and define $a$ and $b$ in the following piecewise fashion:

$$a(u) = \begin{cases} a'(u) + k'' & u \in \mathcal{A}^+ \\ a''(u) & u \in \mathcal{A}^- \end{cases}$$

$$b(u) = \begin{cases} b'(u) + k'' & u \in \mathcal{B}^- \\ b''(u) & u \in \mathcal{B}^+ \end{cases}$$

Because $\mathcal{A}^+ \cup \mathcal{B}^+$ is an independent set, there are no edges $(u, v)$ with $a(u) \geq k'' > b(v)$. Along with the induction hypotheses, this established Property 1. The piecewise definitions of $q$ in line 7 and $z$ in line 8 satisfy Properties 2 and 3 because $(q', z')$ and $(q'', z'')$ do.

Property 4 requires a bit of calculation. The set $\mathcal{A}^+ \cup \mathcal{A}_{k''-1} \cup (\mathcal{B}^+ \setminus \mathcal{B}_{k''-1})$ is an independent set and from the properties of LinOpt

$$\begin{aligned} &P(\mathcal{B})P(\mathcal{A}^+) + P(\mathcal{A})P(\mathcal{B}^+) \geq \\ &P(\mathcal{B})(P(\mathcal{A}^+) + P(\mathcal{A}_{k''-1})) + P(\mathcal{A})(P(\mathcal{B}^+) - P(\mathcal{B}_{k''-1})) \end{aligned}$$

so $P(\mathcal{A})P(\mathcal{B}_{k''-1}) \geq P(\mathcal{B})P(\mathcal{A}_{k''-1})$. Similarly, $(\mathcal{A}^+ \setminus \mathcal{A}_{k''}) \cup \mathcal{B}^+ \cup \mathcal{B}_{k''}$ is an independent set and

$$\begin{aligned} &P(\mathcal{B})P(\mathcal{A}^+) + P(\mathcal{A})P(\mathcal{B}^+) \geq \\ &P(\mathcal{B})(P(\mathcal{A}^+) - P(\mathcal{A}_{k''})) + P(\mathcal{A})(P(\mathcal{B}^+) + P(\mathcal{B}_{k''})) \end{aligned}$$

so $P(\mathcal{A})P(\mathcal{B}_{k''}) \leq P(\mathcal{B})P(\mathcal{A}_{k''})$. Combining these inequalities, we have

$$\frac{P(\mathcal{A}_{k''-1})}{P(\mathcal{A}_{k''-1} \cup \mathcal{B}_{k''-1})} \leq \frac{P(\mathcal{A})}{P(\mathcal{A} \cup \mathcal{B})} \leq \frac{P(\mathcal{A}_{k''})}{P(\mathcal{A}_{k''} \cup \mathcal{B}_{k''})}.$$

Along with the induction hypotheses, this establishes Property 4. □

## C. Additional Results

In this section we present additional results that were omitted from the main body of the paper for space considerations.

### C.1. Other class pairs

In Figures 1 and 2, we present the results for the lower bound on cross-entropy loss for two other choices of class pairs, '1 vs. 9' and '2 vs. 8'. We can see that while the exact values of the lower bound differ, the trend with respect to both the adversarial budget and the number of samples is the same as in the '3 vs. 7' case.

### C.2. Graph properties

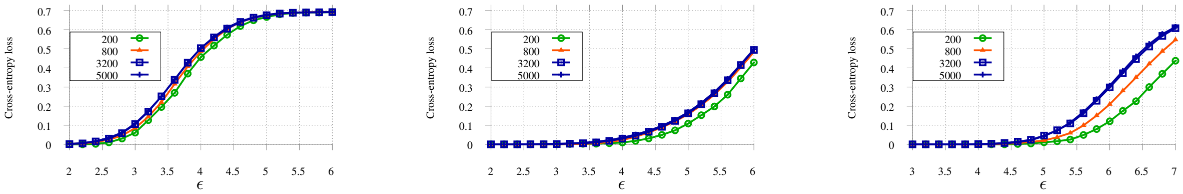
We show the variation in collision probability with the budget for different numbers of samples per class in Figure 3. This quantity can be estimated accurately even with a small number of samples, unlike the lower bound on cross-entropy.

### C.3. Runtime analysis for other datasets

In Figures 4 and 5, we show the variation in runtime for the algorithms to compute the lower bound on cross-entropy loss for the MNIST and Fashion MNIST datasets. Our custom Algorithm (Algorithm 1 in the main body) clearly outperforms the generic convex solver from CVXOPT.

### C.4. Further Gaussian results

In Table 8, we show the variation in the population- and sample-level lower bounds on the cross-entropy loss for data generated from a 2-class Gaussian mixture with $d = 2$. All other parameters are the same as in Section 4.2 of the main paper. We can see that for lower dimensional data, the gap between the bounds is small.

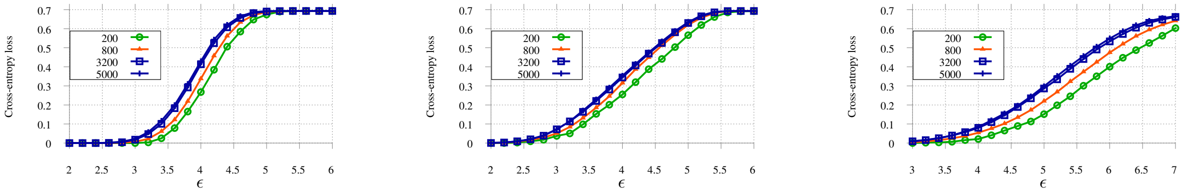


(a) MNIST

(b) Fashion MNIST

(c) CIFAR-10

*Figure 1.* **Two class problem is '1 vs. 9'**. Variation in minimum log-loss for an $\ell_2$ adversary with adversarial budget $\epsilon$ and the number of samples from each class. The maximum possible log-loss is $\ln 2$, which is around $0.693$. The total number of samples is $5000$.



(a) MNIST

(b) Fashion MNIST

(c) CIFAR-10

*Figure 2.* **Two class problem is '2 vs. 8'**. Variation in minimum log-loss for an $\ell_2$ adversary with adversarial budget $\epsilon$ and the number of samples from each class. The maximum possible log-loss is $\ln 2$, which is around $0.693$. The total number of samples is $5000$.

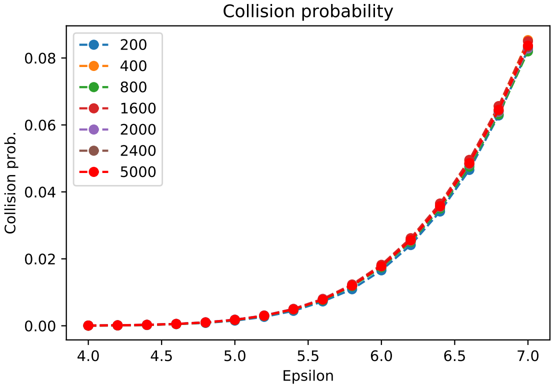
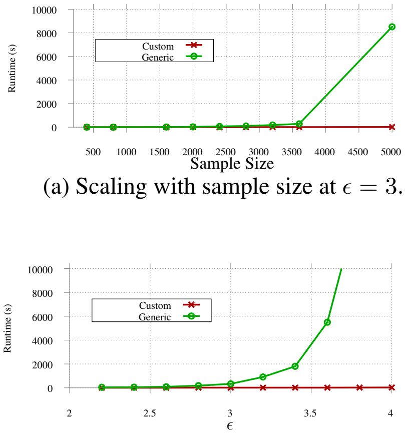


*Figure 3.* Variation in collision probability with attacker budget $\epsilon$ for the CIFAR-10 dataset



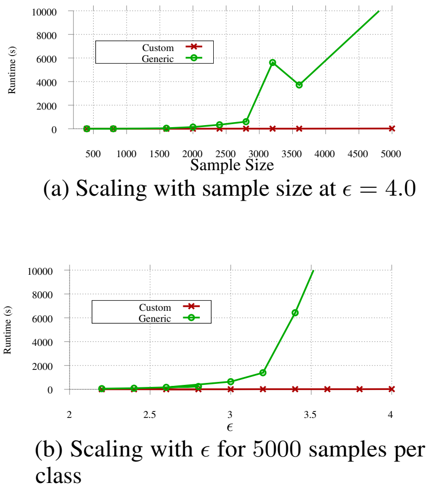
(a) Scaling with sample size at $\epsilon = 3.8$

(b) Scaling with $\epsilon$ for 5000 samples per class

*Figure 4.* Algorithm runtime comparisons for MNIST

### C.5. Minimum $0-1$ loss

We note that the optimal classifier probabilities that are obtained in the course of determining the minimum log-loss can be thresholded to obtain the classification outcomes of the optimal classifier. Care must be taken, however, for data points where the optimal probability is $\frac{1}{2}$ in the two class case. For all points of this type, we just classify them as being in class 1, which avoids any conflicts and recovers the numerical values from previous work (Bhagoji et al., 2019). These bounds are plotted in Figure 6 as the line 'Minimum loss'.



(a) Scaling with sample size at $\epsilon = 4.0$

(b) Scaling with $\epsilon$ for 5000 samples per class

*Figure 5.* Algorithm runtime comparisons for Fashion MNIST

| Activation function | Robust train loss | Robust test loss |
|---|---|---|
| ReLU | 0.106 | 0.236 |
| ELU | 1.056 | 1.060 |
| Tanh | 13.012 | 13.099 |
| Leaky ReLU | 0.103 | 0.348 |
| SELU | 0.704 | 0.706 |

*Table 1.* Variation in train and test loss for a ResNet-18 trained on MNIST with an $\ell_2$ norm adversary with $\epsilon = 3.0$

## D. More robust training results

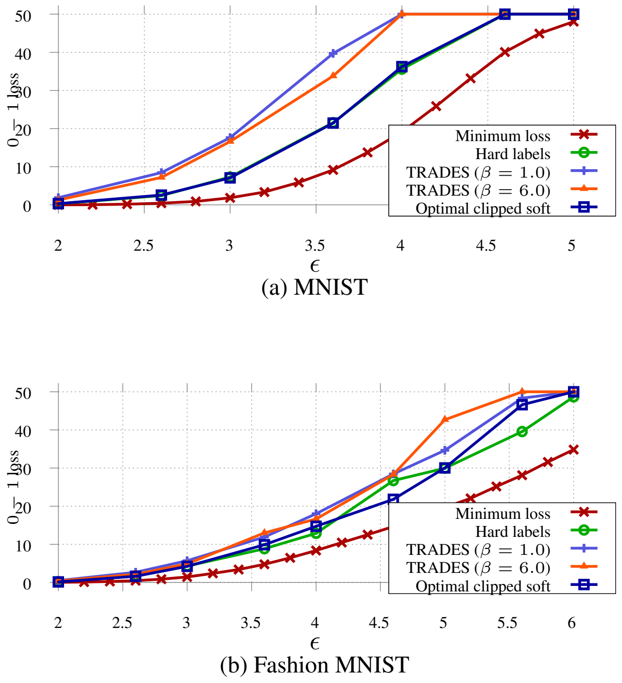
### D.1. Robust $0-1$ loss

We also compare the minimum possible $0-1$ loss to that obtained by various robust training methods using AutoAttack (Croce & Hein, 2020) for both training (Figure 6) and test (Figure 7) data. We find that robust training using optimal clipped soft labels can outperform standard hard label training, and that TRADES performs poorly at higher adversarial budgets.

### D.2. Ablation

**Activation functions:** In Table 1, we study the variation in training and test cross-entropy loss with the activation functions used in a ResNet-18. We find at a budget of 3.0 for MNIST, the standard ReLU activation function performs the best, justifying our choice of this activation function throughout. For the ELU and Tanh activation functions, the network is unable to converge, implying that not all activation functions perform well at higher budgets.

**Architecture:** We also experimented with different ResNet architectures to test if increasing the size of the network would lead to lower values of the robust cross-entropy loss.



(a) MNIST

(b) Fashion MNIST

*Figure 6.* Comparison on *training data* between the $0-1$ loss obtained by different training methods (computed using AutoAttack) versus the optimal loss.

| Architecture | Robust train loss | Robust test loss |
|---|---|---|
| ResNet-18 | 0.451 | 0.451 |
| ResNet-50 | 0.387 | 0.387 |
| ResNet-101 | 0.422 | 0.425 |

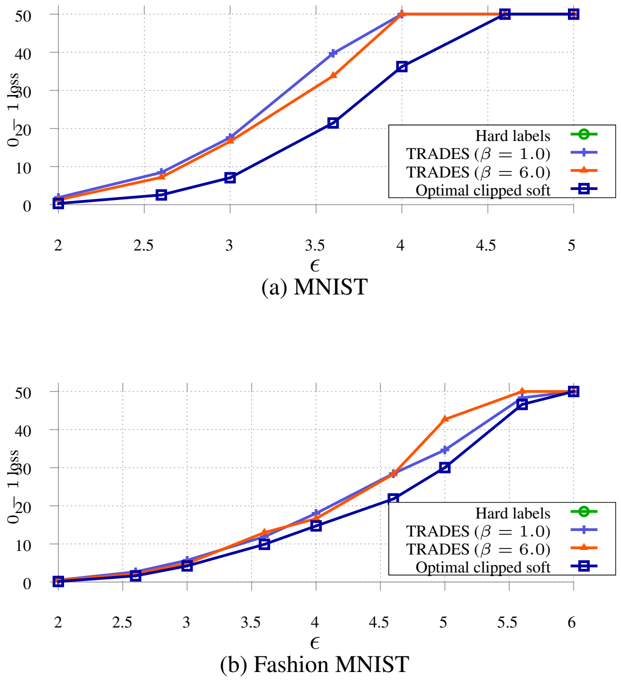
*Table 2.* Variation in train and test loss for models trained on Fashion MNIST with an $\ell_2$ norm adversary with $\epsilon = 5.0$

However, in Table 2, we find that while the loss varies across architectures, an increase in size is not guaranteed to even lower the training loss.

### D.3. CIFAR-10 robust training

We robustly train a ResNet-18 on the CIFAR-10 dataset using $\ell_2$ budgets of $\epsilon = 1.0$ and 2.0. We find that at $\epsilon = 1.0$, the training loss with both adversarial training and TRADES goes to 0, but the test loss is around 1, with a robust classification accuracy of just above 50%, implying that some robust learning is just about possible.

When the budget increases to $\epsilon = 2.0$, the network has below 50% robust classification accuracy on the test set for both training methods. Thus, the performance of current robust classifiers is very far from the optimal cross-entropy lower bound of 0.0 at both these budgets.



(a) MNIST

(b) Fashion MNIST

*Figure 7.* Comparison on *test data* between the $0 - 1$ loss obtained by different training methods (computed using AutoAttack) versus the optimal loss.

## References

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7496–7508, 2019.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020.

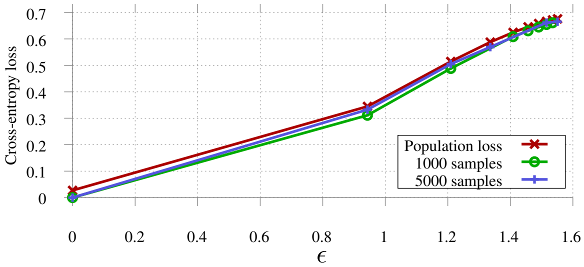


*Figure 8.* Comparing the population-level and sample-level lower bounds on cross-entropy loss for synthetic 2-class Gaussian data of dimension 2.