A. Proof of Proposition 1

Proof. Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x})$. Since the elements of $\mathbf{X}$ are independent, the entropy of the random vector $\mathbf{X}$ equals the sum of the entropy of each individual element

$$\begin{aligned} H(\mathbf{X}) &= -\int_{\mathbf{x}} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\ &= -\sum_i \int_{\mathbf{x}} \left[\prod_j f(x_j)\right] \log f(x_i) d\mathbf{x} \\ &= -\sum_i \int_{x_i} f(x_i) \log f(x_i)\, dx_i \\ &= \sum_i H(X_i) \end{aligned} \tag{1}$$

and that

$$\begin{aligned} H(X_i) &= \int_{x_i} f(x_i) \log \frac{1}{f(x_i)}\, dx_i \\ &= \int_{x_i} (x_i - \mu_i)^2 f(x_i) \log \left(e^{\frac{1}{(x_i-\mu_i)^2}} + \frac{1}{f(x_i)} \right) dx_i. \end{aligned}$$

Since by assumption each X_i is bounded in a compact interval, we have $\log \left(e^{1/(x_i-\mu_i)^2} + 1/f(x_i) \right) > 1$. It also achieves its maximum and minimum values on this compact interval which we denote as k and h, respectively. We then have

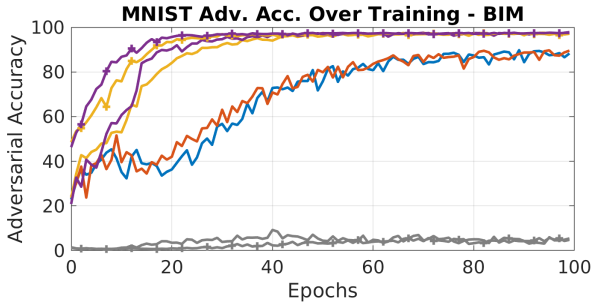
$$h \cdot \mathbb{V}\text{ar}[X_i] \leq H(X_i) \leq k \cdot \mathbb{V}\text{ar}[X_i]$$

indicating that maximizing the entropy of $\mathbf{X}$ is equivalent to maximizing $\sum_i \text{Var}(X_i)$. □

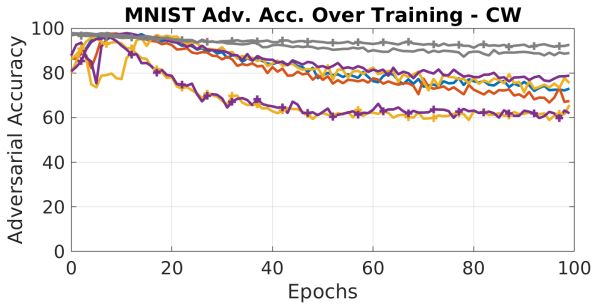
B. Additional Results

Method	BIM	CW	FGM	PGD
CNN	7.1	30.5	36.8	2.6
BNN	5.5	88.9	55.2	0.9
BNN-Off	4.8	-	40.0	2.1
M	88.2	73.0	93.0	55.2
M-V	89.5	67.4	94.0	70.0
M-S	97.1	75.4	96.3	94.0
M-V-S	97.8	78.8	97.2	95.8
M-S-Off	97.2	65.6	97.6	95.8
M-V-S-Off	97.5	61.9	97.4	94.5

Table 1. Adversarial accuracy on MNIST with various combinations of diversity promotion against L_∞ attacks.



(a) art attack



(b) art attack

Figure 1. White box attack accuracy after each epoch.

C. Synthetic Dataset

We consider the synthetic dataset constructed by (Tsipras et al., 2018) In this dataset, adversarial examples are constructed analytically (without gradients) such that one feature is adversarially robust but all others are not.

Tsipras et al. prove that any classifier that achieves arbitrarily high standard accuracy on this dataset necessarily has poor adversarial accuracy. See the original work for a detailed discussion.

The dataset is constructed as a binary classification problem over y with input features $\mathbf{x}$ such that

$$\begin{aligned} y &\sim \{-1, +1\} \\ x_1 &= \begin{cases} +y, & \text{w.p. } p \\ -y, & \text{w.p. } 1-p \end{cases} \\ x_2, ..., x_D &\sim \mathcal{N}(\eta y, 1), \end{aligned}$$

for the standard dataset. Adversarial examples are constructed by sampling $x_2, ..., x_D$ from a distribution that is inversely correlated with the label y, i.e.

$$x_2, ..., x_D \sim \mathcal{N}(-\eta y, 1). \tag{2}$$

To better match this toy dataset to our assumptions, we construct the data (including adversarial examples) as described

and then orthonormally project the features into a new space. We construct the orthonormal matrix by choosing the ones vector for the first column and the remaining columns as any orthogonal space and then normalizing. It is apparent that the bounds given in (Tsipras et al., 2018) hold through this process and that the importance of features in this new space is more evenly distributed.

If we consider a linear classifier followed by a sigmoid activation, then in the original space, the adversarially-robust weight matrix corresponds to a the standard basis vector in the first dimension. Therefore, the ideal weight matrix in the pojected space corresponds to the (normalized) ones vector.

Proof. If we let $\mathbf{v}$ be the robust classification weights, W be the orthonormal transformation matrix, and u be the transformed space, then

$$\begin{aligned} \mathbf{u} &= \mathbf{W}\mathbf{x} \\ y &= \mathbf{v}^T\mathbf{x} \\ &= \mathbf{v}^T\mathbf{W}^T\mathbf{W}\mathbf{x} \\ &= \mathbf{v}^T\mathbf{W}^T\mathbf{u} \\ \mathbf{v_u} &= \mathbf{W}\mathbf{v}. \end{aligned}$$

And since $\mathbf{v}$ is the standard basis for the first dimension, $\mathbf{v_u} = \mathbf{w}_1$, which was selected as the ones vector. □

The choice of W can be generalized to any invertible matrix.

In our experiment, we set $p = 0.95$ and $\eta = \frac{2}{\sqrt{D-1}}$. A robust classifier, which only depends on x_1, would achieve 95% standard and adversarial accuracy whereas a classifier which depends on all covaraites can achieve arbitrarility good accuracy but with low adversarial accuracy. (see Eq. 4 in (Tsipras et al., 2018)).

We train a simple model on this dataset and test the standard and adversarial accuracies throughout the training procedure.

Figure 2 illustrates the standard and adversarial accuracy with respect to the adversarial upper bound for three different Bayesian networks: with no defense, with $M = 100, V = 120, S = 10$ (Case 1), and with $M = 0, V = 0, S = 40$ (Case 2). Points in the figure are collected at different epochs to assess how the adversarial accuracy compares to the upper bound throughout the training process.

As expected, the Bayesian network achieves strong standard accuracy with poor adversarial accuracy and does not even achieve the adversarial upper bound. Both models penalized by our methods consistently achieve the upper bound and influence the model so that it is less prone to achieving arbitrary accuracy. Case 1 showcases good standard accuracy and strong adversarial accuracy, corresponding to the theoretical upper bound, after a warm up period. Case 2 achieves the maximum adversarial accuracy possible. Additional details can be found in Fig. 3 and Fig. 4.

These two cases are representative of the behavior of a variety of different hyperparameter choices. We observe that our penalties cause the model to either achieve and sustain maximum adversarial accuracy or increase adversarial accuracy to a point before decaying to arbitrary standard accuracy.

We take this as encouragement that, since adversarial examples in this setting are constructed analytically (without model gradients), this indicates that our method is capable of creating general adversarial robustness and does not simply obfuscate gradients. The tendency towards the upper bound is striking because we have not directly informed the model about the adversarial problem — there is no adversarial training we only require a diverse ensembling based on inherent properties of Bayesian networks and some diversity encouraging penalties.

Figure 3 provides additional context of the accuracies over the training process. As expected, the undefended model converges rapidly to a high standard accuracy but with adversarial robustness well below the upper bound. In the first case, once the model has achieved the robust, standard accuracy, the adversarial accuracy gradually increases before it ultimately peaks and degrades. While this degradation is unfortunate, it is noteworthy that the adversarial accuracy stays near the upper bound throughout the decay process. In the other case, the model appears to undergo a phase change and quickly converges to and maintains the maximum possible adversarial accuracy.

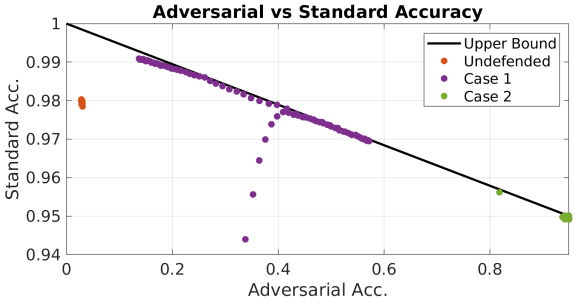


Figure 2. Standard and adversarial accuracy for various models compared to the upper bound. Note, some negative jitter was introduced to Case 2 to enhance visualization.

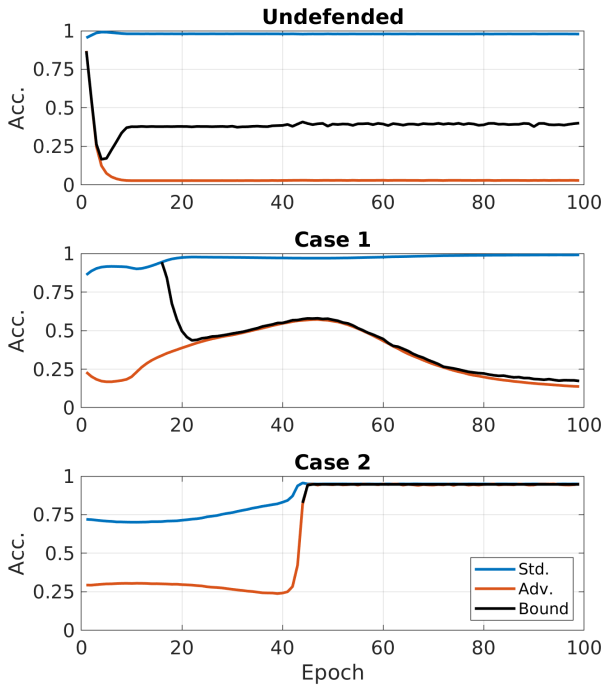


Figure 3. Standard and adversarial accuracy evolution for the test set with various models.

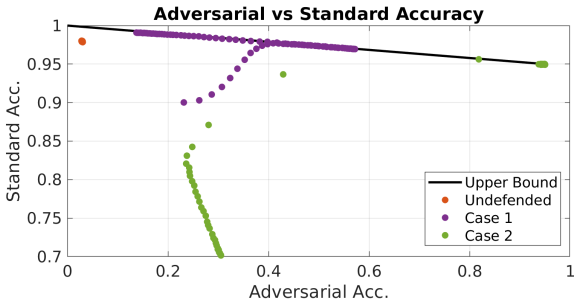


Figure 4. Standard and adversarial accuracy for various models compared to the upper bound with increased plotting range. Note, some negative jitter was introduced to Case 2 to enhance visualization.