Adversarial Robustness of Flow-Based Generative Models – Supplementary material

1 Proof of Theorem 3.1

Theorem 1 *Let $L(\mathbf{x})$ denote the likelihood function of an input sample $\mathbf{x}$ under a Gaussian distribution $\mathcal{N}(\mu, K)$. Let $K = U\Lambda U^T$ be the Eigen-decomposition of the covariance matrix K. Let $c = [c_1, c_2, \dots, c_n] = U^T K$, and $\Lambda = diag([\lambda_1, \dots \lambda_n])$. Let η be a solution of the set of equations*

$$\sum_i \frac{c_i^2}{(1-2\eta\lambda_i)^2} = \epsilon^2$$
$$2\eta\lambda_i - 1 \geq 0 \quad \forall i$$

Then, the optimal additive perturbation δ with norm bound $\|\delta\|_2 < \epsilon$ that maximally decreases the likelihood score of sample $\mathbf{x}$ is given by

$$\delta^* = (K^{-1} - 2\eta I)^{-1} K^{-T}(\mu - \mathbf{x}) \tag{1}$$

Proof: We are interested in generating an adversarial attack on linear models trained on Gaussian input distribution. As explained in Section 2.1 of the main paper, adversarial perturbation δ on sample $\mathbf{x}$ can be obtained by solving the following optimization:

$$\min_{\delta} \quad C - \frac{1}{2}\log(|K|) - \frac{(\mathbf{x}-\mu+\delta)^T K^{-1}(\mathbf{x}-\mu+\delta)}{2}$$
$$\text{s.t. } \|\delta\|_2 < \epsilon$$

The Lagrangian L corresponding to this optimization can be written as

$$\begin{aligned} L &= -\frac{(\mathbf{x}-\mu+\delta)^T K^{-1}(\mathbf{x}-\mu+\delta)}{2} + \eta(\delta^T\delta - \epsilon^2) \\ &= -\frac{(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}{2} - (\mathbf{x}-\mu)^T K^{-1}\delta \\ &\quad - \frac{\delta^T K^{-1}\delta}{2} + \eta(\delta^T\delta - \epsilon^2) \end{aligned}$$

First-order necessary conditions (KKT) From the stationarity condition of KKT, the gradient of the Lagrangian function w.r.t the optimization variables should be 0.

$$\begin{aligned} \nabla_\delta L &= -K^{-T}(\mathbf{x}-\mu) - K^{-1}\delta + 2\eta\delta = 0 \\ (K^{-1} - 2\eta I)\delta &= K^{-T}(\mu - \mathbf{x}) \\ \delta &= (K^{-1} - 2\eta I)^{-1} K^{-T}(\mu - \mathbf{x}) \end{aligned} \tag{2}$$

From complementary slackness, we obtain

$$\eta(\delta^T\delta - \epsilon^2) = 0 \tag{3}$$

So, either $\eta = 0$ or $\delta^T\delta = \epsilon^2$. When $\eta = 0$, $\delta = \mu - \mathbf{x}$. For the other condition $\delta^T\delta = \epsilon^2$, we obtain,

$$(\mu-\mathbf{x})^T K^{-1}(K^{-1} - 2\eta I)^{-2} K^{-T}(\mu - \mathbf{x}) = \epsilon^2$$

Now, consider the Eigen-decomposition of matrix $K = U\Lambda U^T$, where $\Lambda = diag(\lambda_1, \lambda_2, \dots \lambda_n)$. Using this in the above equation, we obtain the condition

$$\sum_i \frac{c_i^2}{(1-2\eta\lambda_i)^2} = \epsilon^2 \tag{4}$$

where $c = [c_1, c_2, \dots c_n] = U^T(\mu - \mathbf{x})$. Eq. (4) can be solved numerically to obtain the value of η.

Second order sufficiency condition: The Hessian of the Lagrangian function can be written as

$$\nabla^2_{\delta\delta} L = -K^{-1} + 2\eta I$$

The above matrix should be positive semi-definite. This gives the following condition

$$2\eta\lambda_i - 1 \geq 0 \quad \forall i \tag{5}$$

We see that the solution $\delta = 0$ does not satisfy this property, hence, it can be eliminated. Hence, the optimal perturbation is the solution to Eq. (4) which satisfy Eq. (5).

2 Proof of Theorem 3.2

Lemma 2 *Let X be a $\chi^2(n)$ distribution. Then, for any $t > 1$,*

$$Pr(X \geq 2tn) \leq e^{-\frac{tn}{10}}$$

Proof: From Laurent & Massart (2000), we know that for a $\chi^2(n)$ random variable X,

$$Pr(X \geq n + 2\sqrt{nx} + 2x) \leq e^{-x}$$

Substituting $x = \frac{tn}{10}$, we get

$$Pr(X \geq n + 2n\sqrt{t/10} + 2tn/10) \leq e^{-tn/10}$$

Now, $n(1+2\sqrt{t/10}+2t/10) < n(1+2t/10+2t/10) < 2nt$ for $t > 1$. Hence,

$$\begin{aligned} Pr(X \geq 2nt) &< Pr(X \geq n + 2n\sqrt{t/10} + 2tn/10) \\ &\leq e^{-tn/10} \end{aligned}$$

Theorem 3 *Let* $\mathbf{x}$ *be an input sample drawn from* $N(\mu, \sigma^2 I)$. *Let* $L(\mathbf{x})$ *denote the log-likelihood function of the sample* $\mathbf{x}$ *estimated using a* m*- step adversarially trained model. Let* δ *be any perturbation vector having a norm-bound* $\|\delta\|_2 \leq \epsilon$. *For any* Δ, *when* $m \geq max\Big[\log\Big(\frac{1}{2\sigma^2\Delta}\Big(2\sigma\epsilon\sqrt{20\log(1/\gamma)} + \epsilon^2\Big)\Big), \log\Big(\frac{1}{2\sigma^2\Delta}\Big[2\sigma\epsilon\sqrt{2n}+\epsilon^2\Big]\Big)\Big] / \log(1+\alpha)$, *with probability greater than* $1-\gamma$,

$$L(\mathbf{x}) - L(\mathbf{x}+\delta) < \Delta$$

Proof: From Section 3.1.1 of main paper, the optimal adversarial perturbation for spherical covariance matrix is given by

$$\delta = \frac{\epsilon}{\|\mathbf{x}-\mu\|}(\mathbf{x}-\mu)$$

From Section 3.2, we know that the estimated model parameters after m steps of adversarial training is given by

$$\begin{aligned} \mu_m^{adv} &= \mu \\ K_m^{adv} &= \sigma^2(1+\alpha)^m I \end{aligned}$$

Log-likelihood drop for this model under the optimal adversarial perturbation can then be computed as

$$\begin{aligned} L(\mathbf{x}) &= C' - \frac{\|\mathbf{x}-\mu\|^2}{2\sigma^2(1+\alpha)^m} \\ L(\mathbf{x}+\delta) &= C' - \frac{\|\mathbf{x}-\mu\|^2}{2\sigma^2(1+\alpha)^m}\Big(1 + \frac{\epsilon}{\|\mathbf{x}-\mu\|}\Big)^2 \\ L(\mathbf{x}) - L(\mathbf{x}+\delta) &= \frac{1}{2\sigma^2(1+\alpha)^m}\Big(2\epsilon\|\mathbf{x}-\mu\| + \epsilon^2\Big) \end{aligned}$$

We want this likelihood difference to be less than Δ.

$$\begin{aligned} &Pr(L(\mathbf{x}) - L(\mathbf{x}+\delta) < \Delta) \\ &= Pr(\frac{\|\mathbf{x}-\mu\|}{\sigma} < \frac{2\sigma^2\Delta(1+\alpha)^m - \epsilon^2}{2\sigma\epsilon}) \end{aligned}$$

We now reparameterize $\tilde{\mathbf{x}} = \frac{\mathbf{x}-\mu}{\sigma} \sim \mathcal{N}(0, I)$. The norm vector $\|\tilde{\mathbf{x}}\|^2$ then obeys a $\chi^2(n)$ distribution with n degrees of freedom. Then,

$$\begin{aligned} &Pr(L(\mathbf{x}) - L(\mathbf{x}+\delta) < \Delta) \\ &= Pr\Big(\|\tilde{\mathbf{x}}\|^2 < \Big(\frac{2\sigma^2\Delta(1+\alpha)^m - \epsilon^2}{2\sigma\epsilon}\Big)^2\Big) \qquad (6) \\ &= 1 - Pr\Big(\|\tilde{\mathbf{x}}\|^2 \geq \Big(\frac{2\sigma^2\Delta(1+\alpha)^m - \epsilon^2}{2\sigma\epsilon}\Big)^2\Big) \end{aligned}$$

Now, we use Lemma 2 in (6). Set

$$t = \frac{1}{2n}\Big(\frac{2\sigma^2\Delta(1+\alpha)^m - \epsilon^2}{2\sigma\epsilon}\Big)^2 \qquad (7)$$

Then,

$$\begin{aligned} &Pr\Big(\|\tilde{\mathbf{x}}\|^2 \geq \Big(\frac{2\sigma^2\Delta(1+\alpha)^m - \epsilon^2}{2\sigma\epsilon}\Big)^2\Big) \\ &\leq \exp\Big(\frac{-1}{20}\Big(\frac{2\sigma^2\Delta(1+\alpha)^m - \epsilon^2}{2\sigma\epsilon}\Big)^2\Big) \\ &\leq \gamma \end{aligned}$$

Simplifying the above expression, we obtain,

$$m \geq \frac{\log\Big(\frac{1}{2\sigma^2\Delta}\Big(2\sigma\epsilon\sqrt{20\log(1/\gamma)} + \epsilon^2\Big)\Big)}{\log(1+\alpha)}$$

Also, in condition (7), we require $t > 1$. This gives,

$$m > \frac{\log\Big(\frac{1}{2\sigma^2\Delta}\Big[2\sigma\epsilon\sqrt{2n} + \epsilon^2\Big]\Big)}{\log(1+\alpha)}$$

Hence, when $m \geq max\Big[\log\Big(\frac{1}{2\sigma^2\Delta}\Big(2\sigma\epsilon\sqrt{20\log(1/\gamma)} + \epsilon^2\Big)\Big), \log\Big(\frac{1}{2\sigma^2\Delta}\Big[2\sigma\epsilon\sqrt{2n}+\epsilon^2\Big]\Big)\Big] / \log(1+\alpha)$, $Pr(L(\mathbf{x}) - L(\mathbf{x}+\delta) < \Delta) \geq 1-\gamma$. This concludes the proof.

3 Extension of linear attacks to non-linear models

We now present how the linear attacks discussed in the previous section can be used to attack non-linear flow-based models. Let $G : \mathbb{R}^d \to \mathbb{R}^d$ denote a non-linear flow-based generative model. We can locally linearize the generator function using a first order Taylor approximation as

$$\begin{aligned} G(\mathbf{z}) &\approx G(\mathbf{z_0}) + \nabla G^T|_{\mathbf{z}=\mathbf{z_0}}(\mathbf{z}-\mathbf{z_0}) \\ &= \big(G(\mathbf{z_0}) - \nabla G^T|_{\mathbf{z}=\mathbf{z_0}}\mathbf{z_0}\big) + \nabla G^T|_{\mathbf{z}=\mathbf{z_0}}\mathbf{z} \end{aligned}$$

Since $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$, $G(\mathbf{z})$ which is an affine transformation of $\mathbf{z}$ also obeys a Gaussian distribution with mean and covariance given by

$$\begin{aligned} \mu &= G(\mathbf{z_0}) - \nabla G^T|_{\mathbf{z}=\mathbf{z_0}}\mathbf{z_0} \\ K &= \nabla G \nabla G^T|_{\mathbf{z}=\mathbf{z_0}} \end{aligned}$$

Using this local linear approximation, we can use Theorem 1 to create an adversarial attack. Given a perturbation bound ϵ, Theorem 1 provides an efficient form for computing adversarial attack. However, the local linear approximation might not hold true for larger perturbation radii. Hence, we propose an iterative version of adversarial attack, called Iterated Gaussian

Iteration	NLL
Clean	2.075
1	2.077
2	2.093
3	2.123
4	2.141
5	2.180
6	2.204
7	2.249
8	2.292
9	2.317
10	2.360

Table 1: Sample results of iterated linear Gaussian attack on a CIFAR-10 image. Attack strength is weaker than PGD-1 (2.92) or PGD-10 (3.65).

attack, as given in Algorithm 1. Sample experimental results on Iterated Gaussian attack are reported in Table 1. We observed that the attack is weaker than one or ten step PGD with the maximum epsilon, which obtain NLL scores 2.92 and 3.65 respectively.

Algorithm 1 Iterated Gaussian attack

Require: Input sample $\mathbf{x}$, ℓ_2 perturbation radius ϵ

1: Choose a small perturbation ball ϵ_s
2: Initialize $\mathbf{x}^{att} = \mathbf{x}$
3: **while** $\|\mathbf{x}^{att} - \mathbf{x}\|_2 < \epsilon$ **do**
4: Set $\mathbf{z_0} = G^{-1}(\mathbf{x}^{att})$
5: Find $\mathbf{x}^{att}$ using Theorem 1 with perturbation radius ϵ_l
6: **end while**

4 Visualization of Attack Distributions

In addition the tables reported in the main text, we visualize the distributions of the attack likelihoods for CIFAR-10. We show these results for the three model types (1) clean (2) adversarially trained and (3) hybrid adversarially trained, against the two attack models (1) in-distribution attacks at various ϵ compared against out-of-distribution samples (unattacked uniform noise) and (2) out-of-distribution attacks at various ϵ compared against "plain" in-distribution samples.

The trends observed here are consistent with what we previously reported in Tables 1 and 2 of the main paper. For in-distribution attacks, we observe that clean model (which is not adversarially trained) is susceptible to adversarial attacks even at perturbations as low as $\epsilon = 1$. This can be seen as the $\epsilon = 1$ attack distribution falls clearly to the right of the clean $\epsilon = 0$ distribution (Figure 4). In addition, we observe that the likelihoods for $\epsilon = 4$ exceeds that for the out-of-distribution attacks, thus confirming that in-distribution samples can made less likely than out-of-distribution samples. For adversarially trained and hybrid models (Figures 2 and 3), NLL distributions of unperturbed and perturbed samples all overlap, showing that models are robust against in-distribution attacks.

We observe that out-of-distribution attacks are successful against all models - clean models and in-distribution adversarially and hybrid trained models (Figures 4, 5, and 6). Thus we do not provide such robustness. As the attack ϵ increases, the likelihoods of out-of-distribution samples are pushed towards the in-distribution likelihoods. Although the distributions do not overlap as in the in-distribution case, the trend is clear: higher ϵ push the distribution towards the clean distribution.

5 Are generated samples from adversarially trained model adversarial?

Generative models trained on adversarial examples provide a unique opportunity to ask the question of whether the samples generated by this model has an adversarial nature. To do this, we generate samples from an adversarially trained model, and evaluate their likelihood on model trained on unperturbed samples. We find that samples generated by adversarially trained model are indeed adversarial with respect to the unperturbed model, at a strength comparable to that on which the model was trained. These results are shown in in Figure 7.

6 Experimental Details

All model architectures for GLOW and RealNVP were trained with default values given in their respective implementations. Adversarial and Hybrid models were trained with $\epsilon = 8$, and $m = 10$ attack iterations. GLOW test sizes for CIFAR-10 and LSUN Bedroom test size were $N = 10,000$ and $N = 1200$ (default) respectively. For GLOW robustness evaluations, adversaries were trained with $m = 32$ and $m = 40$ for CIFAR-10 and LSUN Bedroom respectively.

For the random noise baseline, random images were generated as Unif$[-\epsilon, \epsilon]$ (centered) and then clipped to $[0, 255]$.

Out-of-distribution attacks were performed with a Unif[0,255] random image, and trained trained with $m = 100$ iterations.

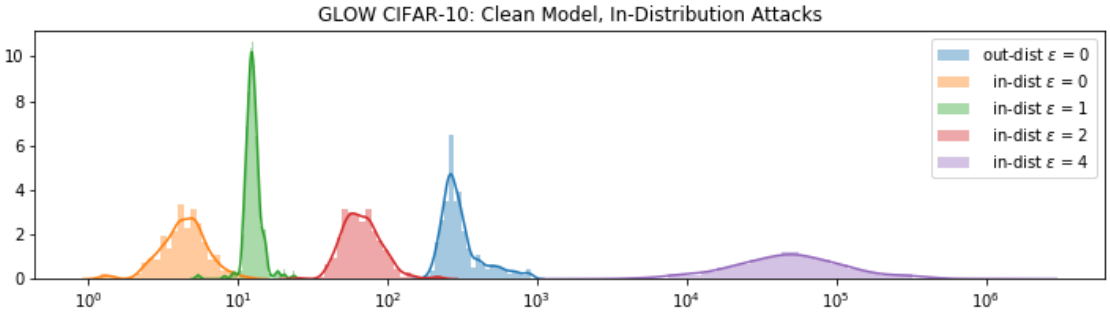

Figure 1: In-distribution attack distributions for clean model.

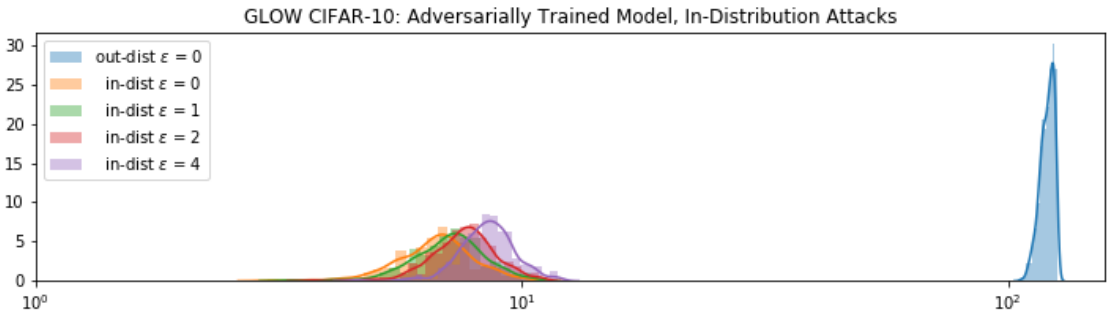

Figure 2: In-distribution attack distributions for adversarially trained model.

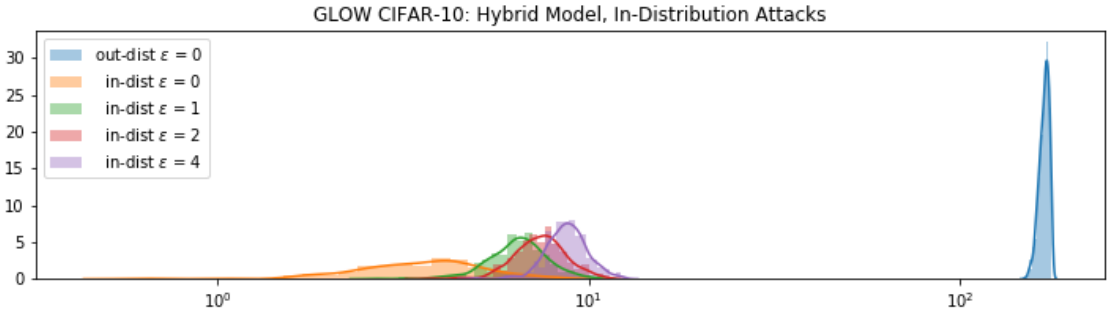

Figure 3: In-distribution attack distributions for hybrid adversarially trained model.

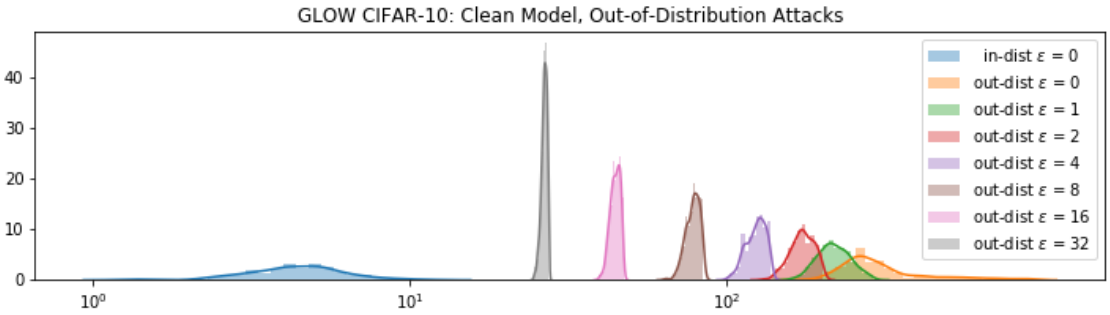

Figure 4: Out-of-distribution attack distributions for clean model.

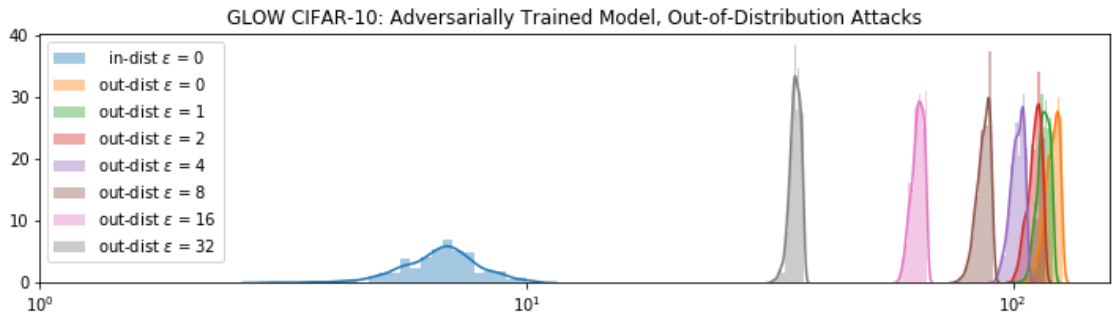

Figure 5: Out-of-distribution attack distributions for adversarially trained model.

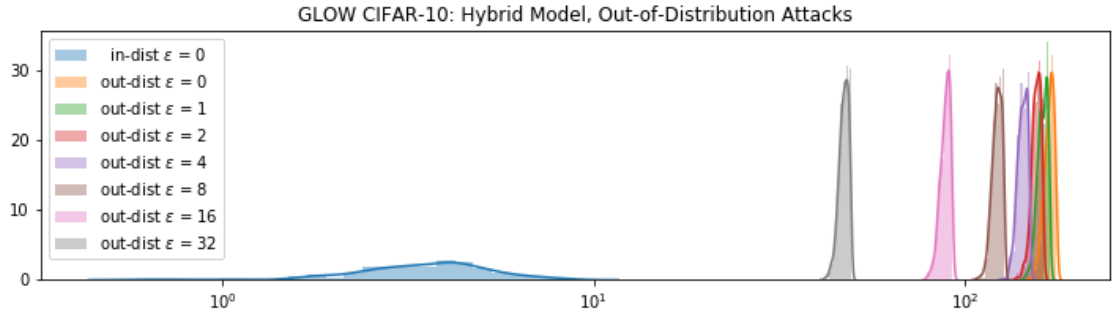

Figure 6: Out-of-distribution attack distributions for hyrbid adversarially trained model.

7 Instability in GLOW likelihood evaluations on CIFAR-10 for high ϵ

In our experiments, we observed variance in likelihood evaluations for GLOW models trained on CIFAR-10 under strong adversaries (attack strengths $\epsilon \geq 8$). On the other hand, low attack strengths ($\epsilon < 8$) had negligible variance. This adds uncertainty as to the "true" value of the attack strength, however we maintain that the trend is clear: stronger adversaries are more disruptive of the likelihood score. As this paper is primarily concerned with the existence of adversarial attacks and robust defenses, we consider this issue not germane to the present work. For completeness, we give details on our investigation of this issue below.

Sources of randomness and numerical issues were investigated. Two sources of randomness found were (1) the addition of uniform random noise in the computation of *continuous* log-likelihoods (Equation (2) in Kingma & Dhariwal (2018)) and (2) the random initialization of rotation matrices $\mathbf{W}$ in the invertible 1×1 convolution (paragraph below Equation (9) in Kingma & Dhariwal (2018)).

High NLL values correspond to very small probabilities. Since the entire computation is done in the log scale, underflow is not a problem. Computations are by default performed with float32, having the range of approximately $\pm 3.4 \times 10^{38}$, which far exceeds the highest value we observed of 10^{16}.

The authors of Kingma & Dhariwal (2018) propose LU-decomposition as a fast means of computing the determinant. In the reference implementation this option was disabled by default. We found that enabling it helped with numeric stability, with negligible drop in training speed.

References

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 10215–10224. Curran Associates, Inc., 2018.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

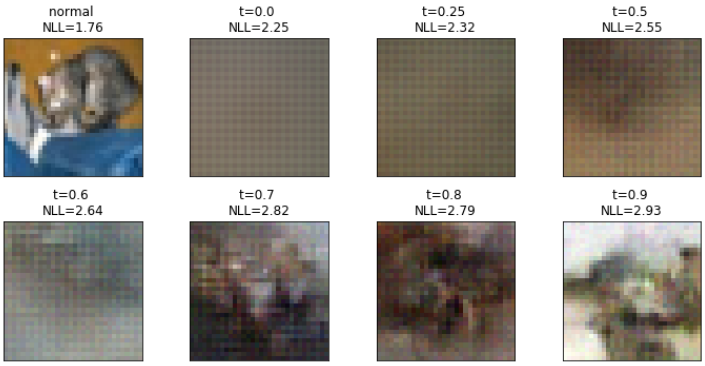

Figure 7: Samples generated at different temperatures from an adversarially trained model, evaluated against a clean model.