Robustness via curvature regularization, and vice versa

Seyed-Mohsen Moosavi-Dezfooli*†
seyed.moosavi@epfl.ch

Alhussein Fawzi*‡
afawzi@google.com

Jonathan Uesato‡
juesato@google.com

Pascal Frossard†
pascal.frossard@epfl.ch

1. Supplementary material

1.1. Proof of Theorem 1

Lower bound. Let $\alpha := \|r^*\|$. We note that α satisfies

$$-c + \|g\|\alpha + \frac{\nu}{2}\alpha^2 \geq -c + g^T r^* + \frac{1}{2}(r^*)^T H r^* \geq 0.$$

Solving the above second-order inequality, we get $\alpha \geq \frac{\|g\|}{\nu}\left(\sqrt{1 + \frac{2\nu c}{\|g\|^2}} - 1\right)$ or $\alpha \leq -\frac{\|g\|}{\nu}\left(\sqrt{1 + \frac{2\nu c}{\|g\|^2}} + 1\right)$. However, since $\alpha \geq 0$, the first inequality holds, which precisely corresponds to the lower bound.
Upper bound. Let $\alpha \geq 0$. Define $r := \alpha u$, and let us find the minimal $|\alpha|$ such that

$$-c + g^T r + \frac{1}{2} r^T H r = -c + \alpha g^T u + \frac{\alpha^2 \nu}{2} \geq 0.$$

We note that the above inequality holds for any $|\alpha| \geq |\alpha_{\min}|$, with $|\alpha_{\min}| = \frac{|g^T u|}{\nu}\left(\sqrt{1 + \frac{2\nu c}{(g^T u)^2}} - 1\right)$. Hence, we have that $\|r^*\| \leq |\alpha_{\min}|$, which concludes the proof of the upper bound. The simplified bounds are proven using the inequality $1 + \frac{x}{2} - \frac{x^2}{2} \leq \sqrt{1+x} \leq 1 + \frac{x}{2}$.

1.2. Results of applying CURE on the SVHN dataset

We fine-tune a pre-trained ResNet-18 using our method, CURE, on SVHN dataset. The learning rate is varying between $[10^{-4}, 10^{-6}]$ for a duration of 20 epochs. The value of γ is set to 4, 8, and 12 for 10, 5, and 5 epochs respectively. Also, for SVHN, we fix $h = 1.25$.

	ResNet-18	
	Clean	Adversarial
Normal training	96.3%	0.9%
CURE	91.1%	28.4%
Adversarial training (reported in [1])	93%	33%

Table 1: Adversarial and clean accuracy for SVHN for original, regularized and adversarially trained models. Performance is reported for a ResNet-18 model, and the perturbations are computed using PGD(10) with $\epsilon = 12$.

*The first two authors contributed equally to this work.
†École Polytechnique Fédérale de Lausanne
‡DeepMind

1.3. Curvature profile of CURE

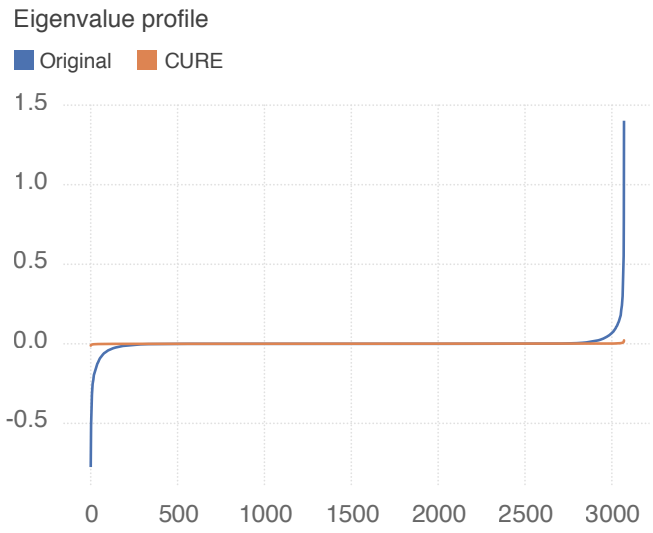
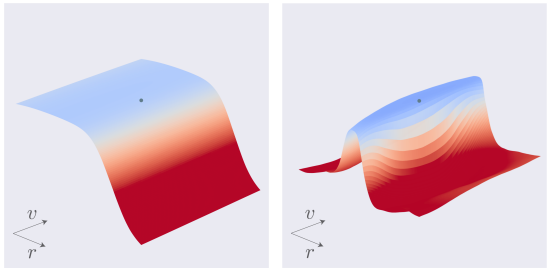


Figure 1: Curvature profiles for a ResNet-18 model trained on SVHN and its fine-tuned version using CURE.

1.4. Loss surface visualization



(a) CURE (b) Original network

Figure 2: Illustration of the negative of the loss surface of the original and the fine-tuned networks trained on SVHN; i.e., $-\ell(s)$ for points s belonging to a plane spanned by a normal direction r to the decision boundary, and random direction v. The original sample is illustrated with a blue dot. The light blue part of the surface corresponds to low loss (i.e., corresponding to the classification region of the sample), and the red part corresponds to the high loss (i.e., adversarial region).

1.5. Evolution of curvature and robustness

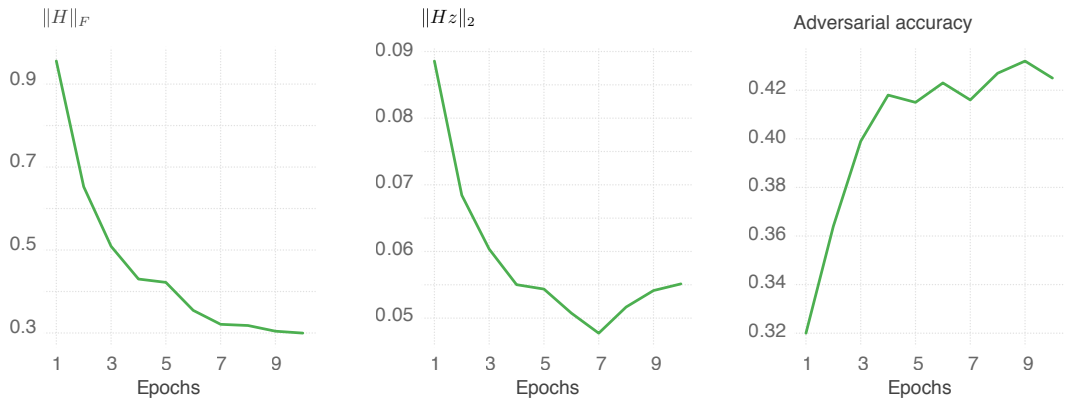


Figure 3: Evolution throughout the course of our CURE fine-tuning for a ResNet-18 on SVHN. The curves are averaged over 1000 datapoints. **Left:** estimate of Frobenius norm , **Middle:** $\|Hz\|$, where $z = \text{sign}(\nabla \ell(x))/\|\text{sign}(\nabla \ell(x))\|_2$ and **Right:** adversarial accuracy computed using PGD(10) with $\epsilon = 8$. The Frobenius norm is estimated with $\|H\|_F^2 = \mathbb{E}_{z\sim\mathcal{N}(0,I)}\|Hz\|^2$, where the expectation is approximated with an empirical expectation over 100 samples $z_i \sim \mathcal{N}(0, I)$.

1.6. Adversarial accuracy

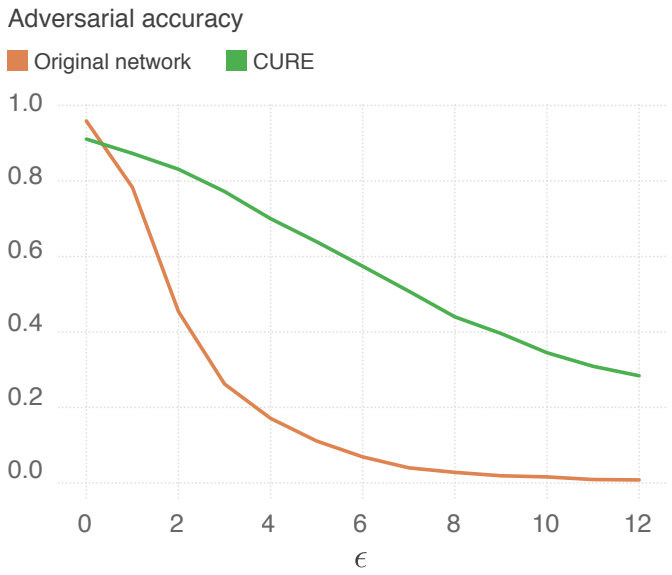


Figure 4: Adversarial accuracy versus perturbation magnitude ϵ computed using PGD(10), for ResNet-18 trained with CURE on SVHN. Curve generated for 2000 random test points.

References

[1] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 1