

A PROOFS

Proof of Theorem 1. [Extended from Xu et al. (2021)] By Xu et al. (2021, Lemma 2), according to the data symmetry in (4), the optimal linear classifier has the form

$$1, \dots, 1, b_\gamma = \arg \min_{\mathbf{w}, b} \mathcal{R}_\gamma(f(\cdot; \mathbf{w}, b)).$$

Recall that (6) proves that for such linear classifier, the robust error is

$$\mathcal{R}_\gamma(f) = \frac{1}{2} \Phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{1}{\sqrt{d}\sigma} \cdot b \right) + \frac{1}{2} \Phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma} \cdot b \right),$$

where Φ is the cumulative distribution function of standard normal.

The optimal b_γ to minimize $\mathcal{R}_\gamma(f)$ is achieved at the point that $\frac{\partial \mathcal{R}_\gamma(f)}{\partial b} = 0$. Thus, b_γ satisfies:

$$\phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{b_\gamma}{\sqrt{d}\sigma} \right) \cdot \frac{1}{\sqrt{d}\sigma} - \phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{b_\gamma}{K\sqrt{d}\sigma} \right) \cdot \frac{1}{K\sqrt{d}\sigma} = 0$$

where ϕ is the probability density function of standard normal. This equals to

$$\phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{b_\gamma}{\sqrt{d}\sigma} \right) = \phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{b_\gamma}{K\sqrt{d}\sigma} \right) / K$$

and

$$\begin{aligned} K &= \phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{b_\gamma}{K\sqrt{d}\sigma} \right) / \phi \left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{b_\gamma}{\sqrt{d}\sigma} \right) \\ &= e^{-\frac{1}{2} \left[\left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{b_\gamma}{K\sqrt{d}\sigma} \right)^2 - \left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{b_\gamma}{\sqrt{d}\sigma} \right)^2 \right]} \end{aligned}$$

It is not hard to see

$$-2 \log K = \left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{b_\gamma}{K\sqrt{d}\sigma} \right)^2 - \left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{b_\gamma}{\sqrt{d}\sigma} \right)^2$$

which re-arranges to a quadratic equation

$$b_\gamma^2 \frac{1}{d\sigma^2} \left(1 - \frac{1}{K^2} \right) - b_\gamma \frac{2(\theta - \gamma)}{\sigma^2} \left(1 + \frac{1}{K^2} \right) + \frac{d(\theta - \gamma)^2}{\sigma^2} \left(1 - \frac{1}{K^2} \right) = 2 \log K.$$

The solution is therefore explicit as

$$b_\gamma = \frac{K^2 + 1}{K^2 - 1} d(\theta - \gamma) - K \sqrt{\frac{4d^2(\theta - \gamma)^2}{(K^2 - 1)^2} + d\sigma^2 q(K)},$$

where $q(K) = \frac{2 \log K}{K^2 - 1}$ which is a positive constant and only depends on K . By incorporating b_γ into (6), we can get the optimal robust error $\mathcal{R}_\gamma(f_\gamma)$:

$$\mathcal{R}_\gamma(f_\gamma) = \frac{1}{2} \Phi \left(B(K, \gamma) - K \sqrt{B(K, \gamma)^2 + q(K)} \right) + \frac{1}{2} \Phi \left(-KB(K, \gamma) + \sqrt{B(K, \gamma)^2 + q(K)} \right),$$

where $B(K, \gamma) = \frac{2}{K^2 - 1} \frac{\sqrt{d}(\theta - \gamma)}{\sigma}$. □

Proof of Theorem 3. We denote the two roots of $\frac{\partial \mathcal{R}_\gamma(f(b))}{\partial b} = 0$ as b_γ^+ and b_γ^- . Here $b_\gamma \equiv b_\gamma^-$. Clearly $\mathcal{R}_\gamma(b)$ is increasing in (b_γ^-, b_γ^+) . We hope to show $b_0 \in (b_\gamma^-, b_\gamma^+) \forall \gamma > 0$, so that $\mathcal{R}_\gamma(b)$ is also increasing in (b_γ^-, b_0) .

Note their Equation (17)

$$\mathcal{R}_\gamma(b) = \frac{1}{2}\Phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{1}{\sqrt{d}\sigma}b\right) + \frac{1}{2}\Phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}b\right)$$

Taking derivative w.r.t. b

$$\frac{\partial \mathcal{R}_\gamma(b)}{\partial b} = \frac{1}{2\sqrt{d}\sigma}\phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{1}{\sqrt{d}\sigma}b\right) - \frac{1}{2K\sqrt{d}\sigma}\phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}b\right)$$

Setting this derivative to 0:

$$0 = K\phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{1}{\sqrt{d}\sigma}b\right) - \phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}b\right)$$

which means

$$\frac{\phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}b\right)}{\phi\left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{1}{\sqrt{d}\sigma}b\right)} = K$$

Using the standard normal density $\phi(u) = e^{-u^2/2}$ and $\frac{\phi(u)}{\phi(v)} = e^{(v^2 - u^2)/2}$, we have

$$\begin{aligned} & \left(-\frac{\sqrt{d}(\theta - \gamma)}{\sigma} + \frac{1}{\sqrt{d}\sigma}b\right)^2 - \left(-\frac{\sqrt{d}(\theta - \gamma)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}b\right)^2 = 2\log K \\ \implies & K^2(-d(\theta - \gamma) + b)^2 - (-d(\theta - \gamma) - b)^2 = 2d\sigma^2 K^2 \log K \\ \implies & (K^2 - 1)b^2 - 2d(\theta - \gamma)(K^2 + 1)b + d^2(\theta - \gamma)^2(K^2 - 1) - 2d\sigma^2 K^2 \log K = 0 \end{aligned}$$

By $x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a} = -\frac{b}{2a} \pm \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}}$, we know

$$\begin{aligned} b_\gamma^\pm &= \frac{K^2 + 1}{K^2 - 1}d(\theta - \gamma) \pm \sqrt{\left(\frac{K^2 + 1}{K^2 - 1}d(\theta - \gamma)\right)^2 - d^2(\theta - \gamma)^2 + K^2 d\sigma^2 q(K)} \\ &= \frac{K^2 + 1}{K^2 - 1}d(\theta - \gamma) \pm K\sqrt{\frac{4d^2(\theta - \gamma)^2}{(K^2 - 1)^2} + d\sigma^2 q(K)} \end{aligned}$$

We now derive the sufficient condition that $b_0 < b_\gamma^+$:

$$\frac{K^2 + 1}{K^2 - 1}d(\theta) - K\sqrt{\frac{4d^2(\theta)^2}{(K^2 - 1)^2} + d\sigma^2 q(K)} < \frac{K^2 + 1}{K^2 - 1}d(\theta - \gamma) + K\sqrt{\frac{4d^2(\theta - \gamma)^2}{(K^2 - 1)^2} + d\sigma^2 q(K)}.$$

This is equivalent to

$$\frac{K^2 + 1}{K^2 - 1}d\gamma < K\left(\sqrt{\frac{4d^2(\theta - \gamma)^2}{(K^2 - 1)^2} + d\sigma^2 q(K)} + \sqrt{\frac{4d^2\theta^2}{(K^2 - 1)^2} + d\sigma^2 q(K)}\right).$$

Therefore, it suffices to have

$$\frac{K^2 + 1}{2K}\gamma < |\theta - \gamma| + |\theta|$$

Finally, it is easy to see the Pareto statement $\mathcal{R}_0(f) < \mathcal{R}_0(f_{\text{DP}}) \implies \mathcal{R}_\gamma(f) > \mathcal{R}_\gamma(f_{\text{DP}})$. A necessary but not sufficient condition for $\mathcal{R}_0(f) < \mathcal{R}_0(f_{\text{DP}})$ given that $b_0 > b_{\text{DP}}$ is $b > b_{\text{DP}}$, since b_0 is a minimizer which means \mathcal{R}_0 is decreasing on the interval $(-\infty, b_0)$. Similarly, \mathcal{R}_γ is increasing on the right of b_γ and thus b has higher robust error. \square

Proof of Corollary 3.2. We can characterize the robust errors based on l_2 attacks in a similar fashion to (6). We notice that

$$\begin{aligned} \mathcal{R}_\gamma(f) &= \mathbb{P}(\exists \|\mathbf{p}\|_2 \leq \epsilon \text{ s.t. } f(\mathbf{x} + \mathbf{p}) \neq y) = \max_{\|\mathbf{p}\|_2 \leq \gamma} \mathbb{P}(f(\mathbf{x} + \mathbf{p}) \neq y) \\ &= \frac{1}{2}\mathbb{P}(f(\mathbf{x} + \gamma_d/\sqrt{d}) \neq -1 \mid y = -1) + \frac{1}{2}\mathbb{P}(f(\mathbf{x} - \gamma_d/\sqrt{d}) \neq +1 \mid y = +1) \end{aligned}$$

In short, the same analysis is in place except $\gamma \rightarrow \gamma/\sqrt{d}$ when we switch from l_∞ to l_2 attacks. \square

B ABLATION STUDIES

B.1 CELEBA

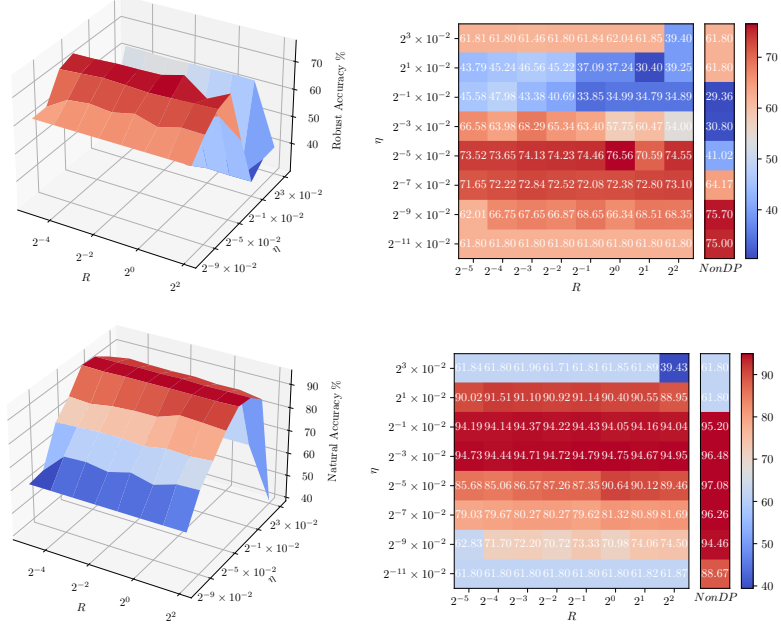


Figure 8: Robust and natural accuracy of η and R on CeleBA with label ‘Male’. We train a 2-layer CNN using DP-Adam and attack by $l_\infty(2/255)$ PGD attack. Same as in Figure 7. Here $\epsilon = 2$, batch size = 512, epochs = 10.

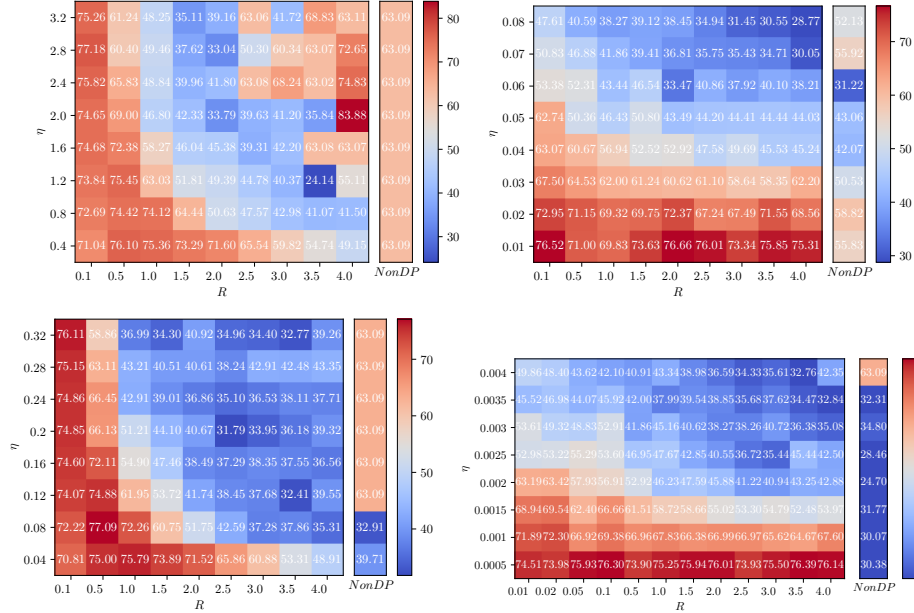


Figure 9: Robust accuracy of CeleBA with label ‘Male’ under different optimizer, trained with a 2-layer CNN and attacked by $l_\infty(2/255)$ PGD attack. Top left: SGD. Top right: Adagrad. Bottom left: SGD momentum. Bottom right: Adam. Here $\epsilon = 2$, batch size = 512, epochs = 10.

B.2 CIFAR10

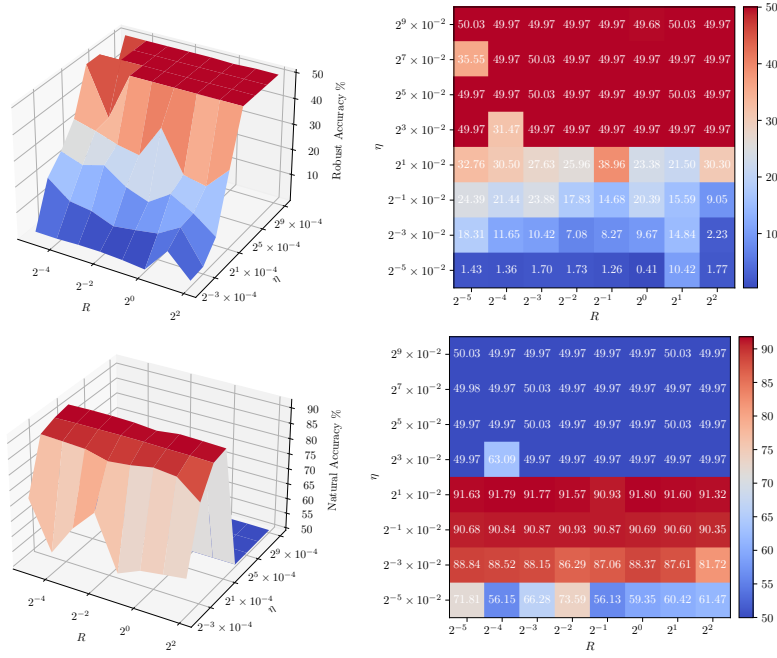


Figure 10: Robust and natural accuracy of η and R on CelebA with label ‘Smiling’. We train ViT-tiny using DP-RMSprop and attack by $l_\infty(2/255)$ PGD attack. Here $\epsilon = 2$, batch size = 1024, epoch = 1.

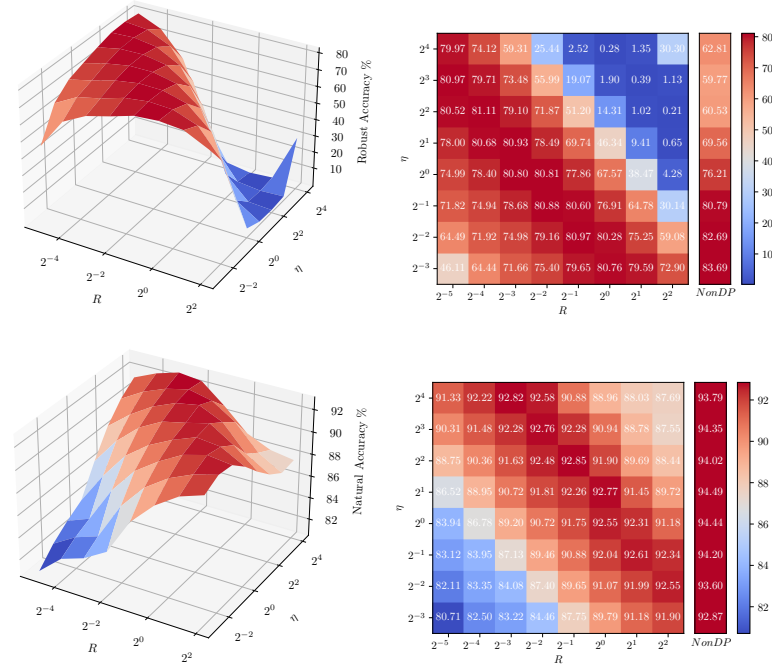


Figure 11: Robust and clean accuracy of η and R on CIFAR10, transferred from SimCLRv2 pre-trained on unlabelled ImageNet. We use DP-SGD and attack by $l_\infty(2/255)$ PGD attack. Here $\epsilon = 2$, batch size = 1024, epochs = 50.

B.3 MNIST

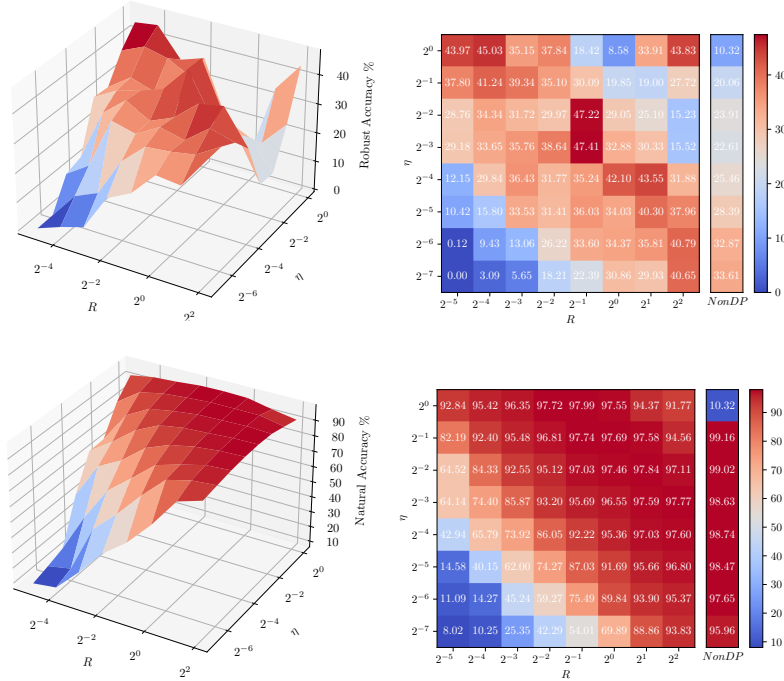


Figure 12: Robust and clean accuracy of η and R on MNIST. We train the CNN from Tramer & Boneh (2020) using DP-SGD and attack by $l_\infty(32/255)$ PGD attack. Here $\epsilon = 2$, batch size = 512, epochs = 40.

B.4 FASHION MNIST

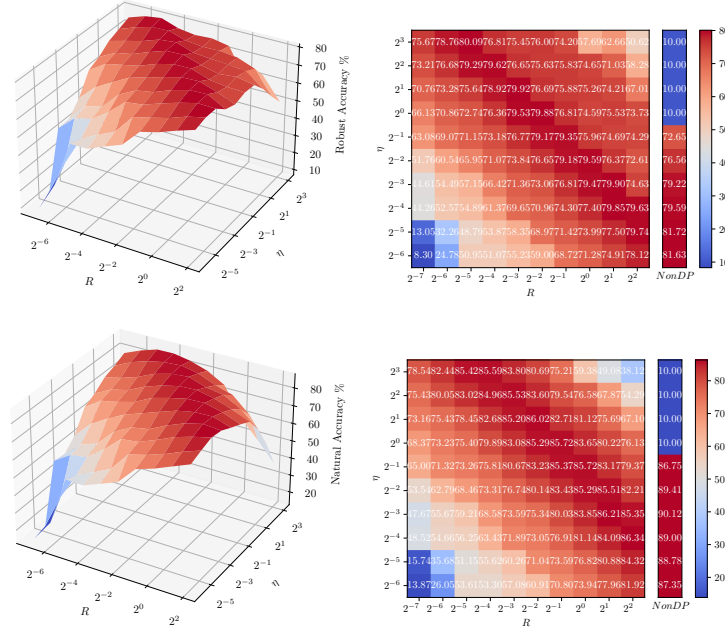


Figure 13: Robust and clean accuracy of η and R on Fashion MNIST. We train the CNN from Tramer & Boneh (2020) using DP-SGD and attack by $l_\infty(2/255)$ PGD attack. Here $\epsilon = 2$, batch size = 2048, epochs = 40.

C MORE TABLES

attack magnitude	SimCLRv2 pre-trained on unlabelled ImageNet								ResNet50	
	DP $\epsilon = 2$	DP $\epsilon = 2$	DP $\epsilon = 4$	DP $\epsilon = 4$	DP $\epsilon = 8$	DP $\epsilon = 8$	Non-DP $\epsilon = \infty$	Non-DP $\epsilon = \infty$	Non-DP $\epsilon = \infty$	Non-DP $\epsilon = \infty$
	robust	accurate	robust	accurate	robust	accurate	robust	accurate	adv 0.5	accurate
$\gamma = 0$	89.69%	92.87%	90.91%	93.41%	91.22%	93.64%	94.29%	94.55%	90.83%	95.25%
$\gamma = 0.25$	82.12%	59.91%	83.35%	74.10%	83.77%	79.03%	82.91%	72.63%	82.34%	8.66%
$\gamma = 0.5$	71.99%	12.76%	72.79%	40.97%	73.08%	54.53%	63.32%	35.95%	70.17%	0.28%
$\gamma = 1.0$	46.30%	9.49%	44.46%	8.97%	44.65%	9.68%	18.81%	0.98%	40.47%	0.00%
$\gamma = 2.0$	4.82%	9.49%	3.12%	8.97%	2.97%	9.63%	0.00%	0.07%	5.23%	0.00%

Table 5: Natural and robust accuracy of SimCLRv2 (Chen et al., 2020) and ResNet50 (Engstrom et al., 2019) on CIFAR10 under 20 steps l_2 PGD attack. Here *robust* parameters are obtained by grid search over η and R against $l_2(0.25)$, and *accurate* parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. See detailed hyperparameters in Appendix D.

	Natural	FGSM	BIM	PGD $_{\infty}$	APGD $_{\infty}$	PGD $_2$	APGD $_2$
Non-DP	94.55%	18.71%	15.97%	15.96%	16.04%	35.95%	35.89%
DP, $\epsilon = 2$	92.73%	10.35%	0.03%	0.03%	0.03%	12.76%	12.68%
DP, $\epsilon = 4$	93.49%	30.10%	9.10%	9.09%	9.12%	40.97%	41.01%
DP, $\epsilon = 8$	93.74%	31.86%	28.08%	28.09%	28.09%	54.53%	54.54%

Table 6: Natural and robust accuracy of models transferred from unlabelled ImageNet pre-trained SIMCLRv2 on CIFAR10 under general adversarial attacks with $\gamma_{\infty} = 4/255$ and $\gamma_2 = 0.5$. Attack steps are 20 if applicable. Model hyper-parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. DP models are trained using DP-SGD, $R = 0.1$, $\eta_{DP} = 4$, momentum = 0.9, batch size = 1024. Non-DP models are trained using SGD with the same hyper-parameters except $\eta_{non-DP} = 0.4$.

attack magnitude	Non-DP $\epsilon = \infty$	DP $\epsilon = 2$	DP $\epsilon = 4$	DP $\epsilon = 8$
$\gamma = 0.0$	99.24%	98.01%	98.32%	98.50%
$\gamma = 0.25$	97.57%	95.29%	95.94%	96.65%
$\gamma = 0.5$	93.32%	90.28%	91.71%	92.97%
$\gamma = 1.0$	66.58%	63.95%	73.32%	77.08%
$\gamma = 2.0$	36.28%	39.88%	51.48%	52.74%

Table 7: Robust accuracy on MNIST under 20 steps l_2 PGD attack. Model hyper-parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. DP models are trained using DP-SGD, $R = 0.1$, $\eta_{DP} = 0.5$, momentum = 0.9, batch size = 512. Non-DP models are trained using SGD with the same hyper-parameters except $\eta_{non-DP} = 0.05$.

attack magnitude	Non-DP $\epsilon = \infty$	DP $\epsilon = 2$	DP $\epsilon = 4$	DP $\epsilon = 8$
$\gamma = 0.0$	99.24%	98.01%	98.32%	98.50%
$\gamma = 2/255$	98.73%	97.12%	97.43%	97.84%
$\gamma = 4/255$	97.88%	95.78%	96.32%	97.13%
$\gamma = 8/255$	95.32%	92.31%	93.51%	94.74%
$\gamma = 16/255$	82.06%	77.67%	80.28%	85.82%

Table 8: Robust accuracy on MNIST under 20 steps l_{∞} PGD attack. Model hyper-parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. DP models are trained using DP-SGD, $R = 0.1$, $\eta_{DP} = 0.5$, momentum = 0.9, batch size = 512. Non-DP models are trained using SGD with the same hyper-parameters except $\eta_{non-DP} = 0.05$.

	Natural	FGSM	BIM	PGD _∞	APGD _∞	PGD ₂	APGD ₂
Non-DP	99.24%	97.92%	97.88%	97.88%	97.77%	93.32%	93.27%
DP , $\epsilon = 2$	98.01%	95.89%	95.80%	95.79%	95.63%	90.28%	90.15%
DP , $\epsilon = 4$	98.32%	96.45%	96.32%	96.33%	96.27%	91.71%	91.68%
DP , $\epsilon = 8$	98.50%	97.19%	97.15%	97.15%	97.06%	92.97%	92.94%

Table 9: Natural and robust accuracy of CNN models on MNIST under general adversarial attacks with $\gamma_\infty = 4/255$ and $\gamma_2 = 0.5$. Attack steps are 20 if applicable. Model hyper-parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. DP models are trained using DP-SGD, $R = 0.1$, $\eta_{DP} = 0.5$, momentum = 0.9, batch size = 512. Non-DP models are trained using SGD with the same hyper-parameters except $\eta_{non-DP} = 0.05$.

attack magnitude	Non-DP $\epsilon = \infty$	DP $\epsilon = 2$	DP $\epsilon = 4$	DP $\epsilon = 8$
$\gamma = 0.0$	89.75%	85.95%	86.60%	86.74%
$\gamma = 0.25$	57.37%	69.24%	72.93%	75.35%
$\gamma = 0.5$	25.21%	46.09%	54.30%	59.23%
$\gamma = 1.0$	7.87%	16.77%	25.95%	29.08%
$\gamma = 2.0$	7.47%	11.77%	16.85%	17.00%

Table 10: Robust accuracy on Fashion MNIST under 20 steps l_2 PGD attack. Model hyper-parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. DP models are trained using DP-SGD, $R = 0.1$, $\eta_{DP} = 4$, momentum = 0.9, batch size = 2048. Non-DP models are trained using SGD with the same hyper-parameters except $\eta_{non-DP} = 0.4$.

attack magnitude	Non-DP $\epsilon = \infty$	DP $\epsilon = 2$	DP $\epsilon = 4$	DP $\epsilon = 8$
$\gamma = 0.0$	89.75%	85.95%	86.60%	86.74%
$\gamma = 2/255$	76.19%	78.29%	79.84%	81.47%
$\gamma = 4/255$	64.46%	69.75%	72.60%	74.72%
$\gamma = 8/255$	47.24%	54.62%	57.87%	60.52%
$\gamma = 16/255$	23.26%	28.51%	31.68%	30.90%

Table 11: Robust accuracy on Fashion MNIST under 20 steps l_∞ PGD attack. Model hyper-parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. DP models are trained using DP-SGD, $R = 0.1$, $\eta_{DP} = 4$, momentum = 0.9, batch size = 2048. Non-DP models are trained using SGD with the same hyper-parameters except $\eta_{non-DP} = 0.4$.

	Natural	FGSM	BIM	PGD _∞	APGD _∞	PGD ₂	APGD ₂
Non-DP	89.75%	70.41%	64.56%	64.44%	53.41%	25.21%	23.13%
DP , $\epsilon = 2$	85.95%	72.11%	69.76%	69.71%	67.13%	46.09%	45.41%
DP , $\epsilon = 4$	86.60%	73.67%	72.68%	72.69%	70.84%	54.30%	53.92%
DP , $\epsilon = 8$	86.74%	75.45%	74.75%	74.74%	73.71%	59.23%	58.98%

Table 12: Natural and robust accuracy of CNN models on Fashion MNIST under general adversarial attacks with $\gamma_\infty = 4/255$ and $\gamma_2 = 0.5$. Attack steps are 20 if applicable. Model hyper-parameters are directly adopted from Tramer & Boneh (2020) for highest natural accuracy. DP models are trained using DP-SGD, $R = 0.1$, $\eta_{DP} = 4$, momentum = 0.9, batch size = 2048. Non-DP models are trained using SGD with the same hyper-parameters except $\eta_{non-DP} = 0.4$.

	Natural	FGSM	BIM	PGD _∞	APGD _∞	PGD ₂	APGD ₂
Non-DP	94.29%	14.48%	12.02%	12.00%	12.03%	31.36%	31.28%
DP, $\epsilon = 2$	92.73%	15.70%	1.59%	1.61%	1.62%	28.05%	28.06%
DP, $\epsilon = 4$	93.49%	30.89%	5.23%	5.27%	5.25%	35.96%	35.98%
DP, $\epsilon = 8$	93.74%	9.66%	4.30%	4.29%	4.31%	33.21%	33.23%

Table 13: Natural and robust accuracy of models transferred from unlabelled ImageNet pre-trained SIMCLRv2 on CIFAR10 under general adversarial attacks with $\gamma_\infty = 4/255$ and $\gamma_2 = 0.5$. Attack steps are 20 if applicable. Model in each row is the most accurate model obtained by simple grid search: Non-DP: $\eta = 0.5$; DP $_{\epsilon=2}$: $\eta = 1.0, R = 0.25$; DP $_{\epsilon=4}$: $\eta = 8, R = 0.0625$, DP $_{\epsilon=8}$: $\eta = 0.5, R = 1.0$. All models are trained using SGD or DP-SGD, momentum = 0.9 and batch size = 1024.

D HYPERPARAMETER SETUP

In Table 1, SimCLRv2 models are pre-trained on unlabelled ImageNet and fine-tuned on CIFAR10. *Accurate* models are directly adopted from Tramer & Boneh (2020) for highest natural accuracy, where optimizer is DP-SGD and SGD, $R = 0.1$, $\eta_{DP} = 4$, $\eta_{non-DP} = 0.4$, momentum = 0.9, batch size = 1024. *Robust* models are obtained by grid search over η and R against $l_\infty(2/255)$, where Non-DP: $\eta = 0.0625$; DP $_{\epsilon=2}$: $\eta = 4, R = 0.0625$; DP $_{\epsilon=4}$: $\eta = 0.5, R = 0.0625$, DP $_{\epsilon=8}$: $\eta = 0.125, R = 0.25$. Other settings are the same as the *accurate* ones. Adversarial attack is l_∞ , 20 steps, alpha = 0.1.

In Table 5, SimCLRv2 models are pre-trained on unlabelled ImageNet and fine-tuned on CIFAR10. *Accurate* models are directly adopted from Tramer & Boneh (2020) for highest natural accuracy, where optimizer is DP-SGD and SGD, $R = 0.1$, $\eta_{DP} = 4$, $\eta_{non-DP} = 0.4$, momentum = 0.9, batch size = 1024. *Robust* models are obtained by grid search over η and R against $l_2(0.25)$, where Non-DP: $\eta = 0.0625$; DP $_{\epsilon=2}$: $\eta = 0.0625, R = 0.25$; DP $_{\epsilon=4}$: $\eta = 0.5, R = 0.0625$, DP $_{\epsilon=8}$: $\eta = 0.125, R = 0.25$. Other settings are the same as the *accurate* ones. Adversarial attack is l_2 , 20 steps, alpha = 0.1.

In Figure 6, models are SimCLRv2 pre-trained on unlabelled ImageNet and fine-tuned on CIFAR10 using DP-SGD, with $\epsilon = 8$, batch size = 1024. Adversarial attack is l_∞ PGD, $\gamma = 4/255$, alpha=0.1.

In Table 2, models the same as in Table 1 with *robust* parameters, where optimizer is DP-SGD and SGD, momentum = 0.9, batch size = 1024, Non-DP: $\eta = 0.0625$; DP $_{\epsilon=2}$: $\eta = 4, R = 0.0625$; DP $_{\epsilon=4}$: $\eta = 0.5, R = 0.0625$, DP $_{\epsilon=8}$: $\eta = 0.125, R = 0.25$. Adversarial attack steps = 20, alpha = 0.1 if applicable.

In Table 3, models are ResNet18 and ViT-tiny trained on CelebA, label Smiling. Images are resized to 224×224 . Optimizer is DP-RMSprop with epochs = 5, batch size = 1024, $\eta = 0.0002$, $R = 0.1$, delta=5e-6. Adversarial attack is l_∞ PGD, 20 steps, alpha = $1/255$.

In Table 4, models are ResNet18 as in Table 3, trained on CelebA, label ‘Smiling’. Images are resized to 224×224 . Optimizer is DP-RMSprop with epochs = 5, batch size = 1024, $\eta = 0.0002$, $R = 0.1$, delta=5e-6. Adversarial attack is $l_\infty(2/255)$ with $\alpha_\infty = 1/255$ and $l_2(0.25)$ with $\alpha_2 = 0.2$, 20 steps, if applicable.

In Figure 7, models are 2-layer CNN trained on CelebA label ‘Male’ using DP-Adam, where $\epsilon = 2$, batch size = 512, epochs = 10. Adversarial attack is $l_\infty(2/255)$ PGD, 20 steps, alpha = 0.1.