

484 Appendix

485 We provide a sketch of the code and additional experiments for our defensive entropy method (dent).
486 Section [A](#) includes the high-level code for dent (in PyTorch). The additional experiments cover more
487 defenses and dent ablations (Section [B](#)) and more attacks (Section [C](#)).

488 A Code Sketch

```
class DynamicModel(torch.nn.Module):
    ... # needs __init__() for optimizer, etc.

    @torch.enable_grad()
    def _update(self, inputs):
        # Perform the forward pass
        preds = self.model(inputs)
        # Compute the loss
        losses = self.loss(preds)
        # Perform the backward pass
        self.optimizer.zero_grad()
        losses.backward(retain_graph=True)
        # Update the parameters
        self.optimizer.step()

    def forward(self, x):
        # Adaptation
        self.model.train()
        for _ in range(self.max_iter):
            self._update(x)
        # Inference
        self.model.eval()
        y = self.model(x)
        return y
```

Listing 1: Sketch of dent code in PyTorch. Adaptation updates are made during testing in `forward()`.

489 Here is a sketch of our PyTorch implementation of dent. The code is simple, and self-contained, for
490 easy application to existing models and defenses. Compatibility with existing defenses is important,
491 as our experiments show that the boost from our dynamic defense compounds the robustness of static
492 defenses. This compounding improvement should continue to help as static and dynamic defenses
493 both improve.

494 B More Results

495 **Defenses, Architectures, and Datasets** Table [9](#) experiments across more defenses, architectures,
496 and datasets. These experiments need to re-train the static defenses, so we reproduce the popular
497 AT [\[31\]](#) and TRADES [\[67\]](#) defenses. We train by PGD with 10-step optimization, norm bounds of
498 $\epsilon_\infty = 8/255$ and $\epsilon_2 = 0.5$, and step sizes of $\alpha_\infty = 2/255$ and $\alpha_2 = 0.1$.

499 **Defense Objective** Dent minimizes entropy, as inspired by tent [\[60\]](#). Related work includes regu-
500 larization to instead maximize information [\[29\]](#) with a term that encourages class balance across
501 predictions. Table [10](#) ablates this regularization to show that our dynamic defense is not too sensitive
502 to it.

Table 9: Dent improves accuracy against ℓ_∞ AutoAttack across model and dataset sizes.

ACCURACY(%)	DATA	ARCH	NATURAL		ADVERSARIAL	
			STATIC	DENT	STATIC	DENT
MADRY ET AL. [31]	CIFAR-10	R-26-4	85.8	86.5	43.8	50.4
	CIFAR-100	R-26-4	59.0	60.1	20.4	23.5
	CIFAR-10	R-32-10	87.0	86.7	45.0	52.5
ZHANG ET AL. [67]	CIFAR-10	R-26-4	85.2	86.6	48.0	49.2
	CIFAR-100	R-26-4	60.1	62.4	18.0	22.5
	CIFAR-10	R-32-10	85.8	86.0	48.0	56.0

Table 10: Ablation of defense objective: entropy minimization (minent) or information maximization (maxinf) for a nominal model against $\epsilon_\infty = 1.5/255$ and robust model against $\epsilon_\infty = 8/255$. Dynamic defense is not sensitive to this choice, as both are entropic objectives, and the updates from either improve accuracy.

	NATURAL		ADVERSARIAL	
	MINENT	MAXINF	MINENT	MAXINF
NOMINAL MODEL	86.5	86.4	50.4	50.0
MADRY ET AL. [31]	92.5	92.7	45.4	45.9

503 **Steps and Computation** As dent is iterative, the amount of computation and adaptation can be
 504 balanced by choosing the number of steps. Table 11 measures adversarial accuracy across steps for
 505 nominal and adversarial training. To appreciate the computation required, we profile the time and
 506 FLOPs for dent with a ResNet-50 model on the ImageNet dataset (Table 12), with an input size
 507 of 288×288 and a batch size of 16. Our experiments show that dent updates do not immediately
 508 saturate: more steps still yield more robustness. However, these steps take more time, motivating
 509 further investigation to tune defensive optimization and reduce the necessary computation.

510 C More Attacks

511 We evaluate dent against attacks with more iterations and higher norm bounds. In the same vein,
 512 we evaluate against the expanded benchmark of AutoAttack Plus: this applies the same four attack
 513 types as AutoAttack but with higher computational budgets. As AutoAttack only includes one
 514 black-box attack (Square), we also evaluate against the Boundary attack [5], for broader coverage of
 515 the black-box setting.

516 **Attacks with More Iterations** It is important to evaluate defenses against sufficiently strong at-
 517 tacks. We ablate the number of steps for APGD-CE, an attack used by AutoAttack, to check its
 518 effectiveness (Table 13). Results indicate that 100 iterations are sufficient, with diminishing returns
 519 for more iterations. Therefore, standard AutoAttack’s configuration is sufficient for evaluating dent’s
 520 robustness.

521 **Attacks with Higher Norm Bounds** Sufficiently large norm bounds should allow attacks to reach a
 522 high success rate. Figure 4 shows that dent’s robust accuracy with a nominal model decreases as we
 523 increase the norm bounds for both ℓ_∞ and ℓ_2 attacks. Specifically, our attacks for evaluating dent’s
 524 ℓ_∞ and ℓ_2 robustness can successfully find adversarial examples with sufficiently large norm bounds.
 525 Meanwhile, Figure 4 demonstrates that dent consistently improves the nominal model’s robustness
 526 against attacks of various strength.

527 **Attacks with AutoAttack Plus** To further analyze dent’s robustness against AutoAttack, we bench-
 528 mark dent against AutoAttack Plus, an extended version of AutoAttack. Table 14 confirms that
 529 dent’s improves the static model’s adversarial accuracy against various attacks. Furthermore, dent’s

Table 11: Ablation of optimization iterations per defense update. More steps deliver more accuracy across models and attacks.

ACCURACY(%)	0	5	10	20	30
RESNET-26-4 [BARE MODEL]					
$\epsilon_\infty = 1.5/255$	8.8	36.1	45.4	49.6	51.0
$\epsilon_2 = 0.2$	9.2	28.0	36.5	39.8	41.7
RESNET-26-4 [MADRY ET AL. [31]]					
$\epsilon_\infty = 8/255$	43.8	46.3	50.4	56.0	58.9
$\epsilon_2 = 0.5$	47.3	48.8	53.0	56.4	57.7
RESNET-32-10 [$\epsilon_\infty = 8/255$]					
MADRY ET AL. [31]	45.0	47.7	52.5	57.1	58.7
ZHANG ET AL. [67]	48.0	48.8	56.0	64.1	67.1

Table 12: Profiling dent computation in time (seconds) and operations (FLOPs) for the dynamic defense of a ResNet-50 on ImageNet. The batch size is 16, and the computation includes all operations for forward, backward, and optimization.

	0	1	5	10	20	30	40	50
ABSOLUTE (S)	0.1	0.3	1.1	2.2	4.2	6.5	8.6	10.8
RELATIVE (\times)	1.0	3.4	12.8	25.3	49.1	75.9	99.9	125.3

530 adversarial accuracy reported in Table 14 is comparable to the standard AutoAttack, indicating that
 531 our evaluation of dent’s robustness is sufficient.

532 **Boundary Attack** For breadth, we evaluate dent against the Boundary black-box decision-based
 533 attack [5]. Our main experiments measure dent’s robustness to AutoAttack, including its black-box
 534 Square attack [2]. Square is a score-based attack, which relies on the confidence of predictions. As
 535 dent optimizes confidence by entropy minimization, it may interfere with such score-based attacks.
 536 We experiment with Boundary as an alternative, because decision-based attacks rely only on the
 537 classification and not the confidence.

538 We attack an adversarially-trained model [13] equipped with dent, and compare Boundary with
 539 AutoAttack in Table 15. The Boundary attack is weaker than the AutoAttack ensemble with or
 540 without dent. By default, Boundary is initialized with an unbounded perturbation by adding noise,
 541 but this is not effective against dent. We attempted to strengthen the attack by nearest neighbor
 542 initialization from misclassifications in the validation set. Our Boundary evaluation is based on the
 543 implementation in the Foolbox toolkit [39].

Table 13: Checking attack effectiveness against one iteration of dent. For $\epsilon_\infty = 8/255$ APGD-CE attacks Madry et al. [31] 100 steps sufficiently reduce adversarial accuracy to evaluate dent.

1	2	3	6	13	25	50	100	200	400	800
63.2	59.6	56.6	53.1	50.8	49.9	49.5	49.4	49.0	49.1	49.0

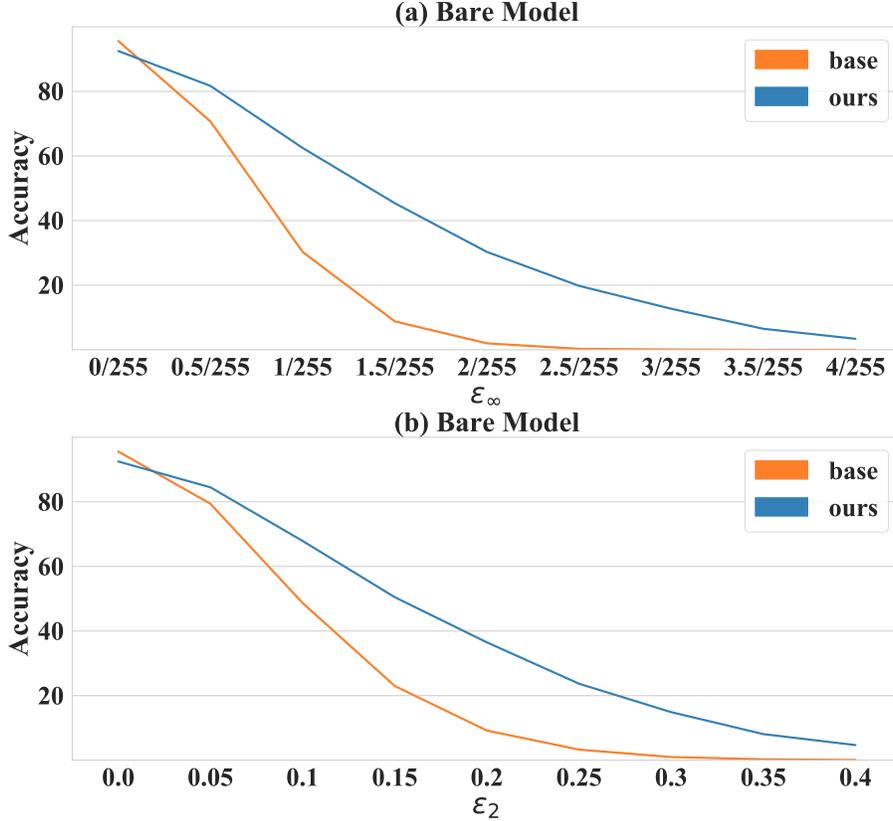


Figure 4: Adversarial accuracy of a nominal model against attacks with varied norm bounds on CIFAR-10. Our dynamic defense consistently improves the robustness of the static model. With sufficiently high bounds however, the attacks succeed in breaking dent’s defense.

Table 14: Benchmark of dent against ℓ_∞ and ℓ_2 norm-bounded attacks on CIFAR-10 by AutoAttack and AutoAttack Plus. AutoAttack Plus only reduces dent’s adversarial accuracy a little, and so the standard AutoAttack is sufficient for evaluation.

ACCURACY(%)	NATURAL	AUTOATTACK	AUTOATTACK+
NOMINAL MODEL ($\epsilon_\infty = 1.5/255$)			
STATIC	95.6	8.8	8.6
DENT	92.5	45.4	38.3
MADRY ET AL. [31] ($\epsilon_\infty = 8/255$)			
STATIC	85.8	43.8	43.8
DENT	86.5	50.4	48.0
DING ET AL. [13] ($\epsilon_\infty = 8/255$)			
STATIC	87.5	41.4	35.2
DENT	87.6	47.6	45.1

Table 15: Dent is robust to black-box attacks, including AutoAttack (Square) and Boundary under $\epsilon_2 = 1.5$. Square is score-based while Boundary is decision-based. The AutoAttack ensemble is the more effective attack overall, so we choose it for our primary evaluation.

ACCURACY(%)	NATURAL	AUTOATTACK	BOUNDARY
[13]	88.0	41.4	72.8
+ DENT	87.9	47.6	70.8