KL-PGD:
A Beginner-Friendly yet Rigorous Introduction

Your Name

Abstract

KL-PGD is a variant of Projected Gradient Descent (PGD) that incorporates Kullback–Leibler (KL) divergence into its update rule to craft adversarial examples which not only maximize loss but also steer the model's output distribution toward a target or away from the true label. This article explains the motivation, mathematical formulation, and practical implementation of KL-PGD in an accessible yet academically precise manner.

1 Background: Adversarial Attacks and PGD

Adversarial attacks seek small perturbations δ within a norm constraint (e.g. $\|\delta\|_\infty \leq \epsilon$) that cause a model f_θ to misclassify an input x. A powerful and widely used white-box method is *Projected Gradient Descent* [**?**]:

$$\begin{aligned} x^{(0)} &= x, \\ x^{(t+1)} &= \Pi_{x+\mathcal{B}_\infty(\epsilon)}\Big(x^{(t)} + \alpha\,\mathrm{sign}(\nabla_x \ell(f_\theta(x^{(t)}), y))\Big), \end{aligned} \tag{1}$$

where ℓ is the classification loss (e.g. cross-entropy), α is the step size, and Π projects back into the valid ℓ_∞ ball.

Standard PGD maximizes the loss:

$$\max_{\|\delta\|_\infty \leq \epsilon} \ell(f_\theta(x+\delta), y).$$

2 Kullback–Leibler Divergence

The Kullback–Leibler (KL) divergence measures the difference between two probability distributions P and Q over classes:

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_c P(c)\,\log\frac{P(c)}{Q(c)}.$$

In classification, if $p_\theta(x)$ denotes the model's softmax output at x, one can craft adversarial objectives based on $D_{\mathrm{KL}}(p_\theta(x)\,\|\,p_\theta(x+\delta))$ to encourage the perturbed output to diverge from the original distribution.

3 KL-PGD Objective

KL-PGD augments the PGD attack by replacing (or combining) the loss-maximization objective with a KL divergence term. Two common variants are:

- **Untargeted KL-PGD:** maximize the divergence from the clean prediction

$$\max_{\|\delta\| \leq \epsilon} D_{\mathrm{KL}}\big(p_\theta(x)\,\|\,p_\theta(x+\delta)\big).$$

- **Targeted KL-PGD:** drive the perturbed output toward a specified target distribution q

$$\min_{\|\delta\|\le\epsilon} D_{\mathrm{KL}}\big(q \,\|\, p_\theta(x+\delta)\big).$$

These objectives focus on changing the entire output distribution rather than only increasing the loss for the true class.

4 KL-PGD Algorithm

Algorithm 1 Untargeted KL-PGD Attack

```
Require: input $x$, true label $y$, model $f_\theta$, radius $\epsilon$, steps $T$, step size $\alpha$
$x^{(0)} \leftarrow x$
for $t = 0$ to $T - 1$ do
    compute $p = f_\theta(x)$                                    ▷ clean softmax
    compute gradient: $g \leftarrow \nabla_{x'} D_{\mathrm{KL}}\big(p \parallel f_\theta(x')\big)\big|_{x'=x^{(t)}}$
    update: $x^{(t+1)} \leftarrow \Pi_{x+\mathcal{B}_\infty(\epsilon)}\big(x^{(t)} + \alpha\,\mathrm{sign}(g)\big)$
end for
return $x^{(T)}$
```

5 Why KL-PGD?

- **Distribution-Aware Attacks:** By optimizing KL divergence, the attack perturbs all output probabilities, not just the true-class score.

- **Sharper Decision-Boundary Exploration:** KL divergence can identify subtle vulnerabilities where the model's confidence shifts gradually.

- **Flexibility:** Targeted or untargeted, KL-PGD can craft perturbations toward arbitrary distributions (e.g. uniform or a specific wrong class).

6 Practical Tips

1. **Choice of Step Size α.** Typically $\alpha = \epsilon/T$ balances speed and stability.

2. **Number of Steps T.** More iterations yield stronger attacks but increase computation.

3. **Combination with Loss.** In practice, one may combine CE-loss and KL terms:

$$\max_\delta \big[\ell(f_\theta(x+\delta), y) + \lambda\, D_{\mathrm{KL}}(p_\theta(x) \,\|\, p_\theta(x+\delta))\big].$$

7 Conclusion

KL-PGD extends the classical PGD framework by leveraging probabilistic divergence, offering a more holistic adversarial objective. It helps expose vulnerabilities in model confidence and can be tailored to specialized threat models.

References

- Madry, A. et al. (2018). *Towards deep learning models resistant to adversarial attacks.* ICLR.

- Goodfellow, I. et al. (2015). *Explaining and harnessing adversarial examples.* ICLR.
- Kullback, S. & Leibler, R. (1951). *On information and sufficiency.* Annals of Mathematical Statistics.