UCLAVISIONLAB

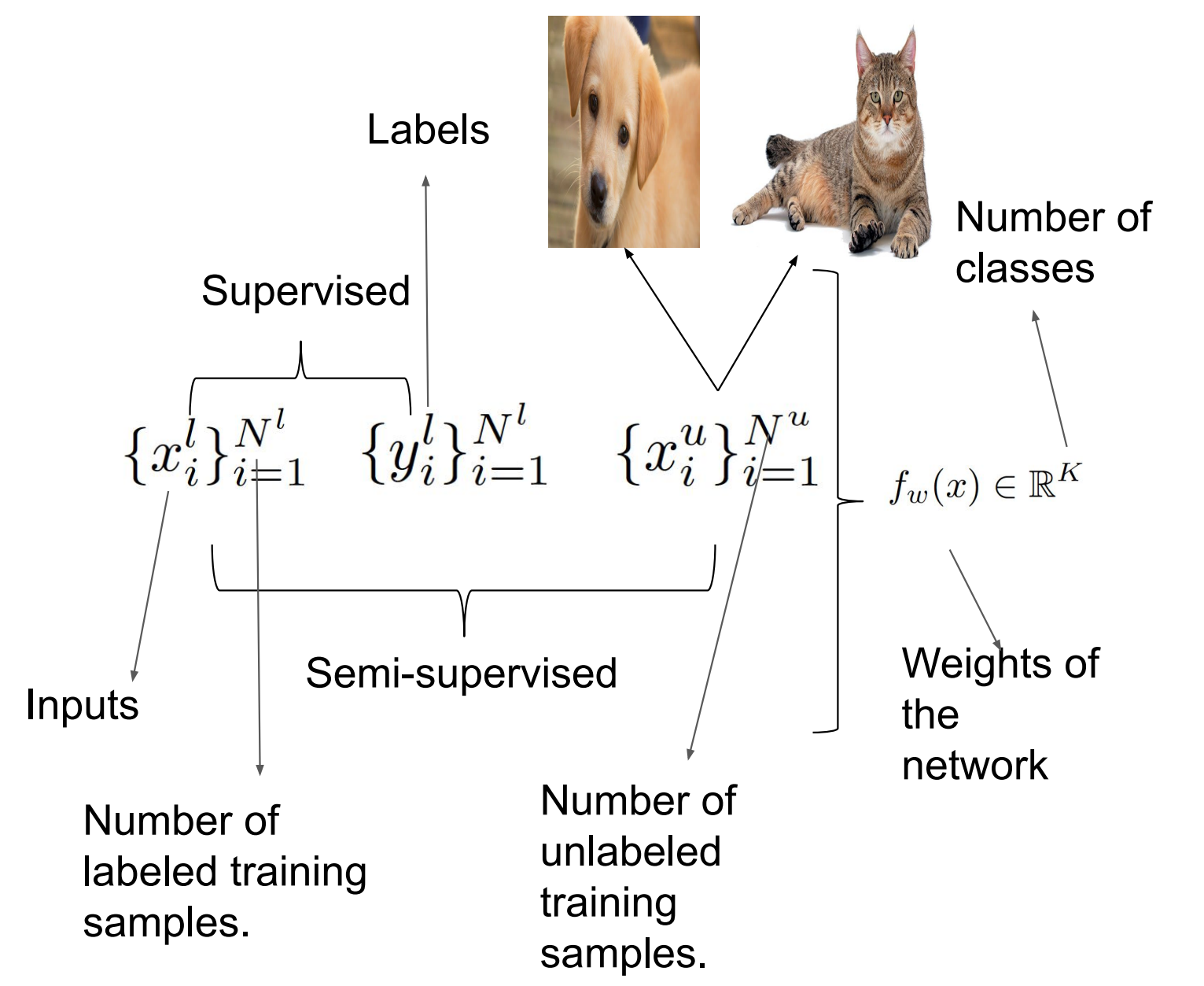
Input and Weight Space Smoothing for Semi-supervised Learning

Safa Cicek and Stefano Soatto

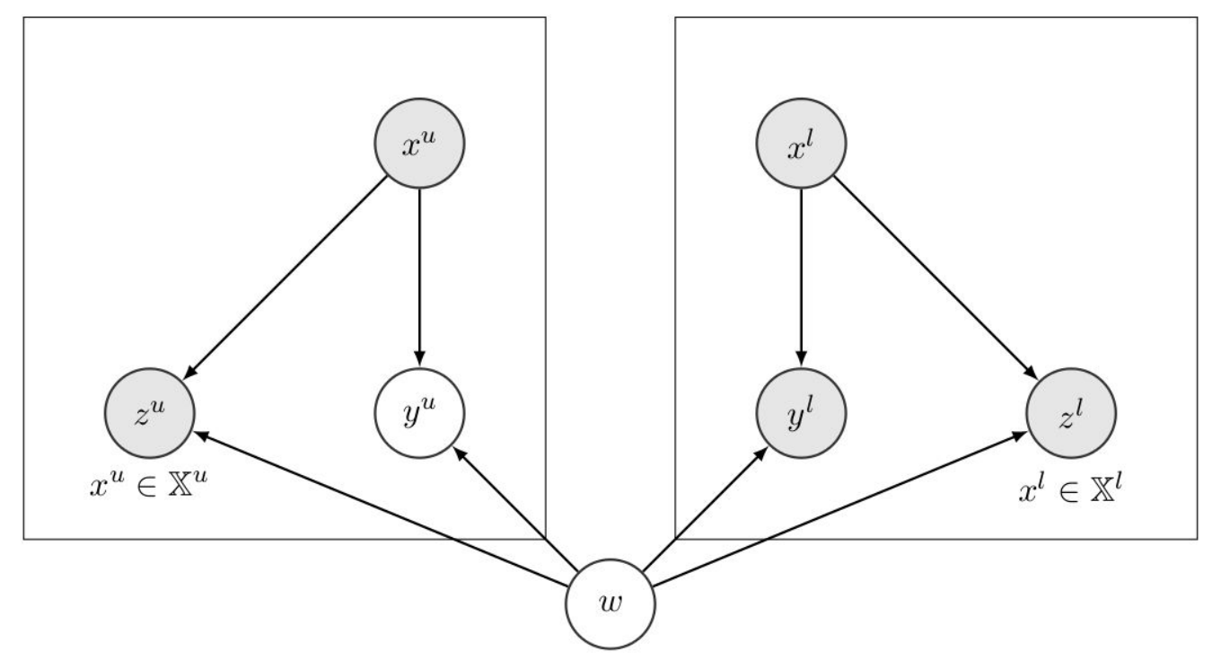
UCLAVISIONLAB

Semi-supervised Learning (SSL)

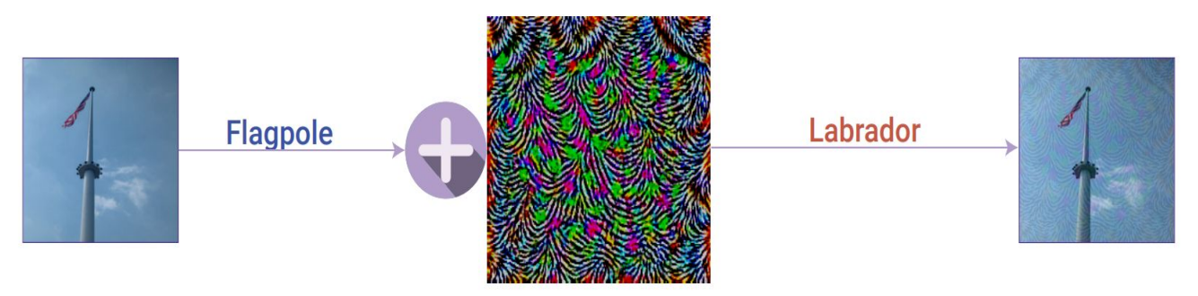
- In SSL, one is given some labeled and unlabeled data.
- Goal: train a classifier, in hope of it performing better than if trained on the labeled data alone.


Labels
Number of classes
Supervised
$\{x_i^l\}_{i=1}^{N^l}$ $\{y_i^l\}_{i=1}^{N^l}$ $\{x_i^u\}_{i=1}^{N^u}$
$f_w(x) \in \mathbb{R}^K$
Inputs
Semi-supervised
Weights of the network
Number of labeled training samples.
Number of unlabeled training samples.

What are the Possible Ways to Exploit Unlabeled Data with Discriminative Models?

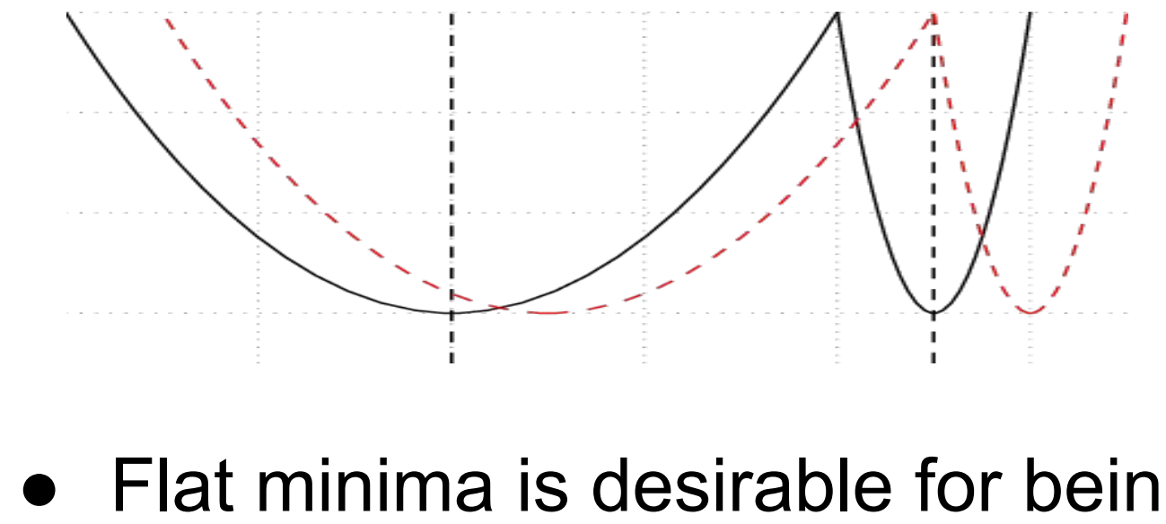


Motivation for Input and Weight Space Smoothing



Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

- For many tasks like image classification, very small adversarial perturbation is nuisance for the task and deep neural networks are vulnerable to such perturbations.

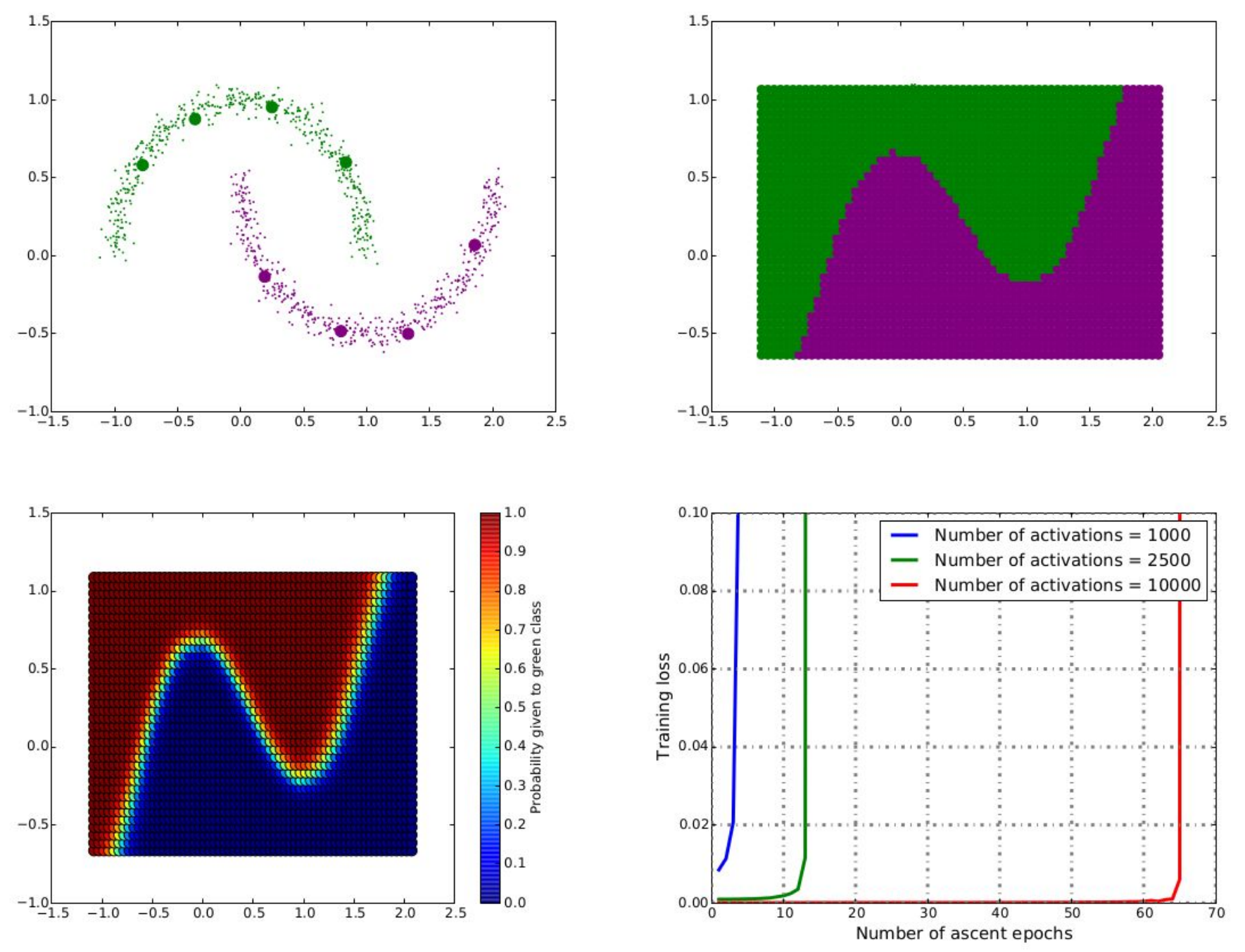


Keskar, N. S., et. al.. (2016). On large-batch training for deep learning: Generalization gap and sharp minima.

- Flat minima is desirable for being robust to domain shift for minimality of the learned representation [1].

[1] Hochreiter, Sepp, and Jürgen Schmidhuber. "Flat minima." *Neural Computation* 9.1 (1997): 1-42.

Input Smoothing and Weight Smoothing do not Imply Each Other.



- Over-parameterized networks are more robust to adversarial noises in the weight space even when they have the same decision boundary (i.e. the same input smoothness).

Input Smoothing

$$\min_w \sum_{x_i \in X} \ell(f(x_i;w), f(x_i+\Delta x_i;w))$$
$$\text{subject to } \Delta x_i = \underset{||\Delta x_i||<\epsilon_x}{\arg\max}\, \ell(f(x_i;w), f(x_i+\Delta x_i;w))\, \forall x_i \in X$$

Supervised Setting

$$\Delta x \approx \epsilon_x \frac{g}{||g||_2}$$
$$\text{subject to } g = \nabla_x \ell(P(y|x), f(x;w))$$

$$\min_w \sum_{x_i \in X} \ell(f(x_i;w), f(x_i+\Delta x_i;w))$$
$$\text{subject to } \Delta x_i = \underset{||\Delta x_i||<\epsilon_x}{\arg\max}\, \ell(P(y_i|x_i), f(x_i+\Delta x_i;w))\, \forall x_i \in X$$

Would this work in SSL setting?

- When we replace ground truth label P(y|x) with the current estimate for SSL case f(x; w), gradient w.r.t. x at delta x = 0 is always zero as it is minimum at zero. Hence, first order input perturbation is not enough in SSL setting.

Virtual Adversarial Training:

- 2nd order approximation of [1],

$$\Delta x \approx \epsilon_x \frac{g}{||g||_2}$$
$$\text{subject to } g = \nabla_{\Delta x} \ell(f(x;w), f(x+\Delta x;w))\Big|_{\Delta x=\xi d} \qquad d \sim N(0,1)$$

1) Approximates the loss with 2nd order Taylor,
2) Approximate the Hessian with 1 step of power iteration.

[1] Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:1704.03976 (2017)

Weight Smoothing

$$\min_w \sum_{x_i \in X} \ell(f(x_i;w), f(x_i;w+\Delta w))$$
$$\text{subject to } \Delta w = \underset{||\Delta w||<\epsilon_w}{\arg\max} \sum_{x_i \in X} \ell(f(x_i;w), f(x_i;w+\Delta w))$$

Conservative penalty [1],

$$w_t = \arg\min_w \ell(P(y|x); f(x;w)) + \gamma||w - w_{t-1}||_2^2$$

[1] Li, M., Zhang, T., Chen, Y., and Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 661–670. ACM.

Joint Input and Weight Smoothing

$$\min_w \sum_{x_i \in X} \ell(f(x_i;w), f(x_i;w+\Delta w)) + \ell(f(x_i;w), f(x_i+\Delta x_i;w))$$
$$\text{subject to } \Delta w = \underset{||\Delta w||<\epsilon_w}{\arg\max} \sum_{x_i \in X} \ell(f(x_i;w), f(x_i;w+\Delta w))$$
$$\Delta x_i = \underset{||\Delta x_i||<\epsilon_x}{\arg\max}\, \ell(f(x_i;w), f(x_i+\Delta x_i;w))\, \forall x_i \in X$$

$$\min_w \sum_x \ell_{CE}(x;w) + \lambda L(x;w)$$

where

$$L(x;w) = \ell(f(x;w), f(x; w + \underset{||\Delta w||<\epsilon_w}{\arg\max} \sum_{x_i \in X} \ell(f(x_i;w), f(x_i;w+\Delta w))))+$$
$$\ell(f(x;w), f(x + \underset{||\Delta x||<\epsilon_x}{\arg\max}\, \ell(f(x;w), f(x+\Delta x;w)); w))$$

$$\ell_{CE}(x;w) = -\langle P(y|x), \log f(x;w)\rangle$$

A New Algorithm for Weight Smoothing: Adversarial Block Coordinate Descent (ABCD)

Algorithm 1 Adversarial Block Coordinate Descent (ABCD)

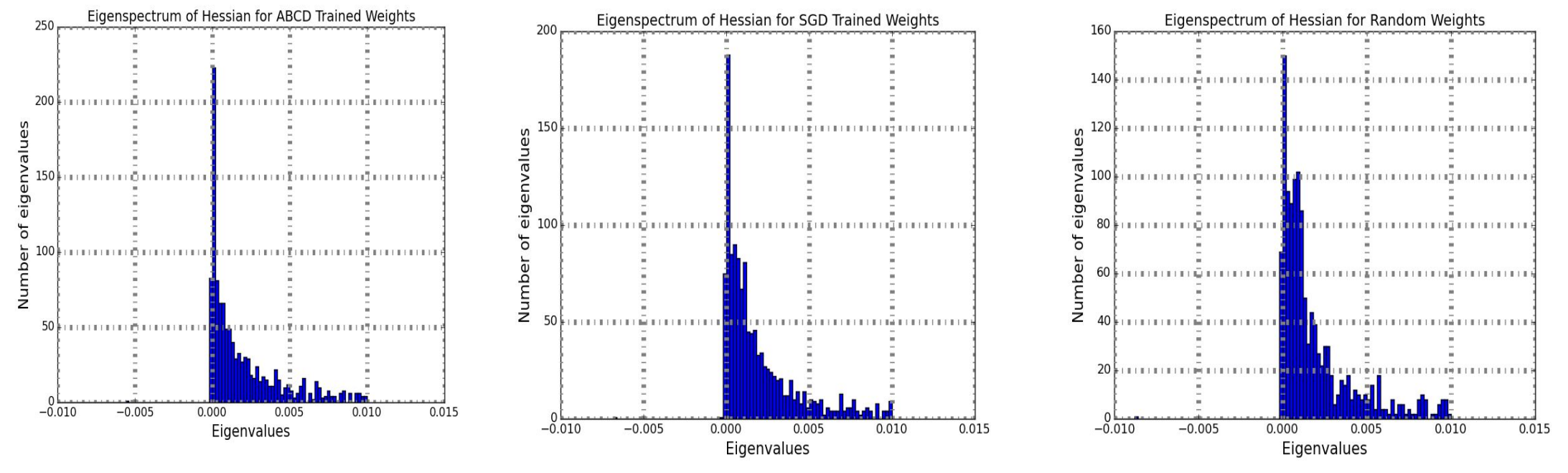
```
Input: Minibatch set $B_t$, loss function $\ell(\cdot)$, initial weights $w_0$.
Hyper-parameters: Ascent and descent learning rates $\eta_A$ and $\eta_D$. Number of inner iterations $L$.
Output: Final weights $w_L$.
for $l = 1 : L$ do
    $\Gamma_i$ sample from $\{0, -1\}$ for all $i \in \{1, \dots, |w_0|\}$.
    $\Gamma_i^a = \Gamma_i$ for all $i \in \{1, \dots, |w_0|\}$.
    $\Gamma_i^d = \Gamma_i + 1$ for all $i \in \{1, \dots, |w_0|\}$.
    // Run stochastic gradient ascent with a small learning rate $\eta_A$
    $w_{l-\frac{1}{2}} = w_{l-1} - \eta_A \Gamma^a \odot \nabla_{w_{l-1}} \left( \frac{1}{|B_t|} \sum_{i=1}^{|B_t|} \ell(x_i; w_{l-1}) \right)$
    // Run stochastic gradient descent with a learning rate $\eta_D \gg \eta_A$
    $w_l = w_{l-\frac{1}{2}} - \eta_D \Gamma^d \odot \nabla_{w_{l-\frac{1}{2}}} \left( \frac{1}{|B_t|} \sum_{i=1}^{|B_t|} \ell(x_i; w_{l-\frac{1}{2}}) \right)$
```

ABCD + VAT for SSL

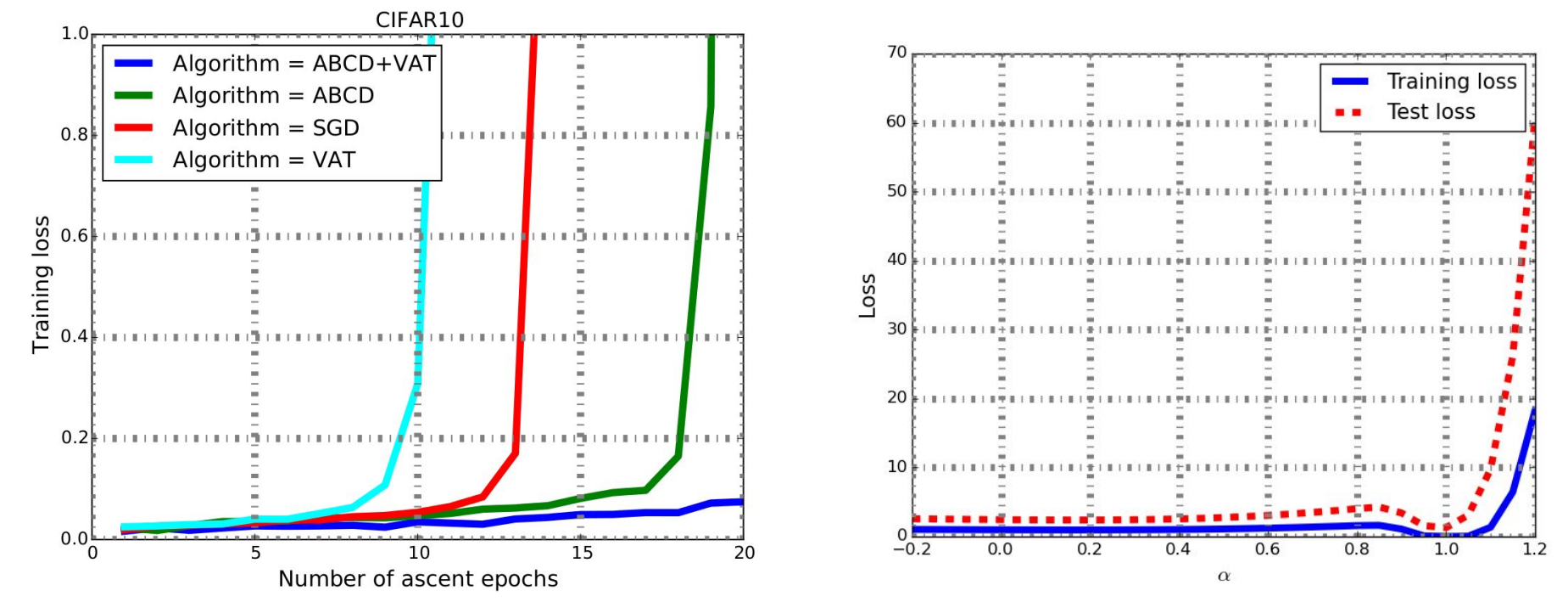
Algorithm 2 SSL algorithm using ABCD as optimizer; VAT and entropy as regularizers. $\ell_{CE}(x;w)$, $\ell_E(x;w)$, $\ell_{VAT}(x;w)$ are as defined in Eq. 11, Eq. 12, Eq. 5 respectively.

```
for $t = 1 : T$ do
    // Run ABCD on cross entropy for labeled samples:
    Sample $B_t^l$
    $w_{t'} = ABCD(B_t^l, \ell_{CE}(x;w), w_{t-1})$ // weight smoothing
    // Run ABCD on entropy and SGD on VAT loss for unlabeled samples:
    Sample $B_t^u$
    $w_{t'} = ABCD(B_t^u, \ell_E(x;w), w_{t'})$ // weight smoothing
    $w_t = SGD(B_t^u, \ell_{VAT}(x;w), w_{t'})$ // input smoothing
```

Hessian of the Solution Converged



Robustness to Perturbations in Weight Space



Comparison to State-of-the-art.

	VAT [1]	Stochastic Transformation [2]	Temporal Ensemble [3]	GAN +FM [4]	Mean Teacher [5]	VAdD [6]	Ours
CIFAR10	10.55	11.29	12.16	15.59	12.31	**9.22**	9.28 ± 0.21
SVHN	3.86	NR	4.42	5.88	3.95	3.55	**3.53 ± 0.24**

[1] Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:1704.03976 (2017)
[2] Sajjadi, M., Javanmardi, M., Tasdizen, T.: Mutual exclusivity loss for semi- supervised deep learning. In: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE (2016) 1908-1912
[3] Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
[4] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems. (2016) 2234-2242
[5] Tarvainen, A., Valpola, H.: Mean teachers are better role models:Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems. (2017) 1195-1204
[6] Park, S., Park, J.-K., Shin, S.-J., and Moon, I.-C. (2017). Adversarial dropout for supervised and semi-supervised learning. arXiv preprint arXiv:1707.03631.