Representation Focused Algorithm for Deep Networks (R., Jones & Harchaoui 2022)

Reduced objective

- Consider learning a feature representation $\phi(\cdot, \theta)$ and a linear predictor W on top of $\phi(\cdot, \theta)$,

$$\min_{\theta, W} \frac{1}{n} \sum_{i=1}^{n} \ell(W^\top \phi(x_i, \theta), y_i) + \Omega(\theta, W)$$

- For squared loss ℓ, penalty Ω, can define

$$f(\theta) := \min_{W} \frac{1}{n} \sum_{i=1}^{n} \ell(W^\top \phi(x_i, \theta), y_i) + \Omega(\theta, W)$$

$\rightarrow$ pseudo-likeli. Besag (1975) or Wiberg algo. Wiberg (1976)

Algorithm Idea

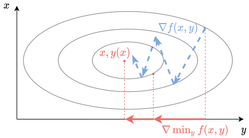
- Stochastic gradient descent on reduced objective $f(\theta)$

$\rightarrow$ Biased oracles with bias controlled by mini-batch size

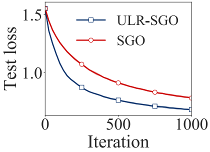
- Generalized to approx. minimizers for non-quad. losses

$\rightarrow$ Gradient of $f(\theta)$ obtained by implicit diff. Blondel et al. (2021)

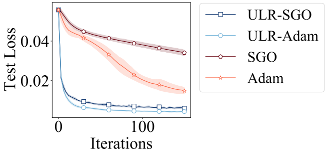
$\rightarrow$ Can be plugged into e.g. Adam



Potentially circumvent oscillations



All-CNN on CIFAR10 multinomial loss



LeNet5 on MNIST squared loss

SGO: Stochastic Gradient Optimization

ULR-X: Proposed oracle with optim. algo. X