Ⓔ SAM SAM

Training Dataset $S = \bigcup_{i=1}^{n} \{(x_i, y_i)\}$ from $\mathcal{D}$ i.i.d

model parameters. $w \in W \subseteq \mathbb{R}^d$

Data point loss function $l : W \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$

Training loss : $L_S(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} l(w, x_i, y_i)$

Population loss : $L_D(w) \triangleq \mathbb{E}_{(x,y)\sim D}[l(w, x, y)]$

Theorem : $\rho > 0$, w.h.p. over training set S, from $\mathcal{D}$

hyper param.

Regularization $\lambda\|w\|_2$ can somehow do it !!

$$L_D(w) \leq \max_{\|\epsilon\|_2 \leq \rho} L_S(w+\epsilon) + \underbrace{h(\|w\|_2^2/\rho^2)}$$

strictly increasing function $(\mathbb{R}_+ \to \mathbb{R}_+)$ conditioned on L_D

Rewriting

$$\underbrace{\left[\max_{\|\epsilon\|_2 \leq \rho} L_S(w+\epsilon) - L_S(w)\right]}_{\text{Sharpness}} + L_S(w) + h\left(\|w\|_2^2/\rho\right)$$

connected to the bound

So SAM optimization problem.

$$\min_{w} L_S^{SAM}(w) + \lambda\|w\|_2^2 \quad // \quad L_S^{SAM}(w) = \max_{\|\epsilon\|_p \leq \rho} L_S(w+\epsilon)$$

↳ p norm $= [1, \infty]$

Finding L_S^{SAM} via optimization

$$\epsilon^*(w) \triangleq \arg\max_{\|\epsilon\|_p \leq \rho} L_S(w+\epsilon) \approx \arg\max_{\|\epsilon\|_p \leq \rho} L_S(w) + \epsilon^T \nabla_w L_S(w)$$

$$= \arg\max_{\|\epsilon\|_p \leq \rho} \epsilon^T \nabla_w L_S(w)$$

Solution revolves math:

Elementwise operation

$$\hat{\epsilon}(w) = \rho\, \text{sign}\left(\nabla_w L_s(w)\right) \left|\nabla_w L_s(w)\right|^{q-1} \Big/ \underbrace{\left(\|\nabla_w L_s(w)\|_q^q\right)}_{\text{norm}}^{1/p}$$

where $1/p + 1/q = 1$

Now,

as max value

$$\nabla_w L_s^{SAM}(w) \approx \nabla_w L_s(w + \hat{\epsilon}(w))$$

at the point $\underline{w + \hat{\epsilon}(w)}$

$$= \frac{d(w + \hat{\epsilon}(w))}{dw} \nabla_w L_s(w)\Big|_{w+\hat{\epsilon}(w)}$$

$$= \nabla_w L_s(w)\Big|_{w+\hat{\epsilon}(w)} + \underbrace{\frac{\partial \hat{\epsilon}(w)}{\partial w} \nabla_w L_s(w)\Big|_{w+\hat{\epsilon}(w)}}_{\text{Dropped !!}}$$

For Acceleration: Dropping the Second term.

$$\nabla_w L_s^{SAM}(w) \approx \nabla_w L_s(w)\Big|_{w+\hat{\epsilon}(w)}$$