Weekly meetings

If Influence Functions are the Answer, Then What is the Question?

Setting

Consider a prediction task (regression problem) with:

- Input space $\mathcal{X}$;
- Output space $\mathcal{Y}$;
- Training set $\mathcal{D}^n = \{z_i\}_{i=1}^n$ where $z_i = (x_i, y_i)$ for all $i = 1, \dots, n$;
- Parameter $\theta \in \Theta := \mathbb{R}^d$;
- $f(\theta; x)$ estimator of $\mathcal{Y} \mid \mathcal{X}$;
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g., $\ell(y', y) \mapsto \|y' - y\|^2$).

We aim to minimize the training error:

$$L(\theta; \mathcal{D}^n) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(\theta; x_i),\, y_i\big).$$

Call

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} L(\theta; \mathcal{D}^n).$$

How different is it from

$$\hat{\theta}_{\varepsilon, -z} = \arg\min_{\theta \in \Theta} \big(L(\theta; \mathcal{D}^n) - \varepsilon\, \ell(f(\theta; x), y)\big)\,?$$

We can re-train the whole model on $\mathcal{D}^n \backslash \{z\}$ (Leave-One-Out method), or use influence functions.

Influence functions

Definition 1. Given $(\bar{x}, \bar{y}) = \bar{z} \in \mathcal{D}^n$, the *influence loss difference* relative to $\bar{z}$ is

$$\mathcal{Q}(\bar{z}; \hat{\theta}) = \frac{d}{d\varepsilon} \Big[L(\hat{\theta}; \mathcal{D}^n) - \varepsilon\, \ell(f(\hat{\theta}; \bar{x}), \bar{y}) \Big] \Big|_{\varepsilon = 1/n}.$$

Interpretation: It measures how much the training error changes when the training datum $\bar{z}$ is removed.

Definition 2. Given $(\bar{x}, \bar{y}) = \bar{z} \in \mathcal{D}^n$, the *influence function* relative to $\bar{z}$ is

$$\mathcal{I}(\bar{z}; \hat{\theta}) = \lim_{\varepsilon \to 0} \frac{\hat{\theta}_{\varepsilon+1/n,-\bar{z}} - \hat{\theta}_{1/n,-\bar{z}}}{\varepsilon}.$$

Interpretation: It represents the direction in which the optimal parameter moves when the training objective is perturbed by removing $\bar{z}$.

Assume that L is strongly convex. Then we can rewrite the previous quantities in the closed forms:

$$\mathcal{I}(z; \hat{\theta}) = H_{\hat{\theta}}^{-1} \nabla \ell(f(\hat{\theta}; x), y), \qquad \mathcal{Q}(z; \hat{\theta}) = \nabla \ell(f(\hat{\theta}; x), y)^\top H_{\hat{\theta}}^{-1} \nabla \ell(f(\hat{\theta}; x), y),$$

where $H_{\hat{\theta}}$ is the Hessian of L at $\hat{\theta}$, which is difficult to compute.

Unfortunately, there is some problems with these formulae.
Indeed, the strong convexity assumption is essential if we want the LOO method to be well approximated by influence functions. Even under such assumption, the problem is not trivial because computing the inverse of the Hessian and the matrix-vector product are heavy computations.

Solution 1. For iHVP, there exist efficient approximations requiring $O(nd)$ flops instead of $O(n^3)$.

Solution 2. Influence functions do not approximate LOO retraining, but the proximal Bregman response function (PBRF).

Response functions

We define the response function as:

$$\hat{r}_z(\varepsilon) = \arg\min_{\theta \in \Theta} \left(L(\theta; \mathcal{D}^n) - \varepsilon \, \ell(f(\theta; x), y) \right).$$

Then $\hat{r}_z(\varepsilon) = \hat{\theta}_{(\varepsilon,-z)}$ and $\hat{r}_z(0) = \hat{\theta}$.

Since $\hat{r}$ is differentiable at 0 we can expand with Taylor at first order and we get:

$$\hat{r}_{z,\mathrm{lin}}(1/n) = \hat{\theta} + \frac{1}{n} H_{\hat{\theta}}^{-1} \nabla \ell(f(\hat{\theta}; x), y).$$

We require $H_{\hat{\theta}}$ positive definite to invert it; thus $\hat{\theta}$ must be a minimizer.
To compute influence functions for MLPs, we can approximate $H_{\hat{\theta}}$ using the Gauss–Newton Hessian (GNH) and add damping:

$$\mathcal{I}^\dagger(z; \hat{\theta}) = \left(J_{y,\hat{\theta}}^\top H_{\hat{\theta}} J_{y,\hat{\theta}} + \lambda I \right)^{-1} \nabla \ell(f(\hat{\theta}; x), y),$$

where $J_{y,\hat{\theta}}$ is the Jacobian of $F(\theta) = (f(\theta; x_1), \ldots, f(\theta; x_n))$.
Note that the damped GNH is always positive definite, as long as $H_{\hat{\theta}}$ is SPD.
The damped response satisfies:

$$\hat{r}_{z,\mathrm{damp}}(\varepsilon) = \arg\min_{\theta\in\Theta} \Big(L(\theta; \mathcal{D}^n) - \varepsilon\, \ell(f(\theta; x), y) + \tfrac{\lambda}{2}\|\theta - \hat{\theta}\|^2 \Big),$$

$$\hat{r}_{z,\mathrm{damp,lin}}(1/n) \approx \hat{\theta} + \frac{1}{n}\mathcal{I}^{\dagger}(z; \hat{\theta}).$$

Since, in practice, the parameter we consider is not a minimizer of L, we want to consider the loss for which the early-stopped point θ^s is optimal:

$$\mathcal{L}(\theta; \theta^s, \mathcal{D}^n) = \frac{1}{n}\sum_{i=1}^{n} D_{\ell^{(i)}}\big(f(\theta; x_i), f(\theta^s; x_i)\big),$$

with $D_{\ell^{(i)}}$ being the Bregman difference:

$$D_{\ell^{(i)}}(y, y') = \ell(y, y_i) - \ell(y', y_i) - \nabla_1 \ell(y', y_i)^{\top}(y - y').$$

Consequently, we can define the *Proximal Bregman Response Function* (PBRF) as:

$$r^b_{z,\mathrm{damp}}(\varepsilon) = \arg\min_{\theta\in\Theta} \Big(\mathcal{L}(\theta; \theta^s, \mathcal{D}^n) - \varepsilon\, \ell(f(\theta; x), y) + \tfrac{\lambda}{2}\|\theta - \theta^s\|^2 \Big).$$

The interesting property of this object is that the linearized PBRF satisfies

$$r^b_{z,\mathrm{damp,lin}}(1/n) = \theta^s + \frac{1}{n}\mathcal{I}^{\dagger}(z; \theta^s).$$

Thus influence functions do *not* approximate LOO retraining under the original loss L. Instead, they approximate the effect of training from θ^s under the modified objective

$$\mathcal{L}(\theta; \theta^s, \mathcal{D}^n) - \tfrac{1}{n}\ell(f(\theta; x), y) + \tfrac{\lambda}{2}\|\theta - \theta^s\|^2.$$

Error decomposition

The approximation error of influence functions decomposes into:

- **Warm-start gap:** LOO starts from a random parameter (cold start), while IF are related to θ^s; we can then converge to another "optimal" point;
- **Proximity gap:** the factor $||theta - theta^s||$ induces the warm start not to move far away from $theta^s$;
- **Non-convergence gap:** in practice we almost never start from a fully trained network;
- **Linearization error:** produced by approximating the Taylor expansion at first order;
- **Solver error:** inexact iHVP computation.

Remark. The PBRF formulation eliminates the first three gaps.

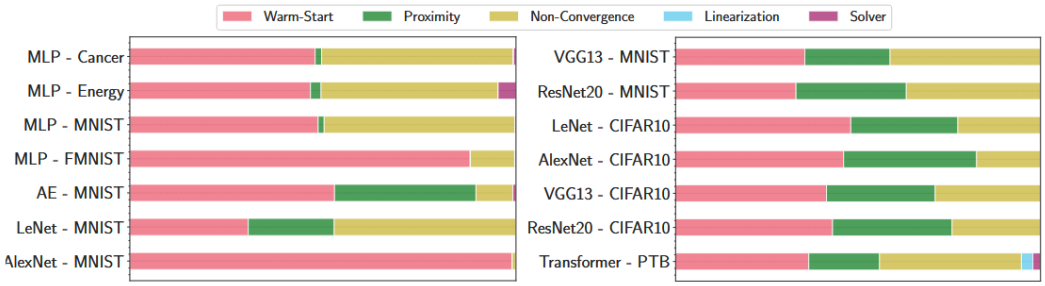

Figure 1: Visual representation of the error defomposition on different datasets and models. The main focus is that the largest components are the first three.

Bibliography

[1] Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger Grosse. If influence functions are the answer, then what is the question?

[2] Pang Wei Koh, Kai-Siang Ang, Hubert H. K. Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects.

[3] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions.