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 $i = 1, \dots, n$; $X = (x_1, ..., x_n)$, $Y = (y_1, ..., y_n)$;
- 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 (empirical risk):

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

Call

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

How different is it from

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

when $\epsilon = 1/n$ and $z = z_{\bar{i}}$ for some $i \in [n]$?
To answer this question, we can re-train the whole model on $\mathcal{D}^n \setminus \{z\}$ (Leave-One-Out method), or use influence functions.
In the context where the influence functions are well defined, they are a powerful tool. However, when applied to multi-layer perceptrons, for example, their capability of approximating the effect of the LOO decreases drastically. This paper presents a new point of view: IFs do not approximate the LOO retraining, but instead the effect of another method they present, called PBRF.

Influence functions

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

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

Interpretation: It measures how much the training error changes when the data point $\bar{z}$ 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 $\bar{z}$.
Assuming that L is strongly convex in θ, 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 $\hat{\theta}$.
Remark. This part is not really clear: referring to the original reference [2], the derivation should be as follow, with our notation.
Let $L_i(\theta) = (L(\hat{\theta}; \mathcal{D}^n) - 1/n\, \ell(f(\hat{\theta}; x_i), y_i))$. Assuming that L is twice differentiable, we can use Taylor near $\hat{\theta}$:

$$L_i(\theta) = L_i(\hat{\theta}) + (\theta - \hat{\theta})^\top L_i'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^\top L_i''(\hat{\theta})(\theta - \hat{\theta}) + \mathrm{O}(\|\theta - \hat{\theta}\|^3)$$

Then, by minimizing the difference $L_i(\theta) - L_i(\hat{\theta})$ for $\theta \neq \hat{\theta}$, we get:

$$\theta = \hat{\theta} - (HL_i(\hat{\theta}))^{-1} JL_i(\hat{\theta}).$$

From this expression is also clear that the Influence Function estimator represents the parameter after one step of the Newton algorithm starting from $\hat{\theta}$ trying to get to $\hat{\theta}_{1/n, -z_i}$.

Unfortunately, there is some problems with these formulae:

1. the strong convexity is essential. If at a minimum point H has any 0-eigenvalue, we cannot invert it.

2. Even when L is strongly convex, 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. Change point of view: Influence functions are not approximators of LOO retraining, but instead of the proximal Bregman response function (PBRF).

Response functions

We define the response function as:

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

Observe that $\hat{r}_z(\varepsilon) = \hat{\theta}_{\varepsilon,-z}$ 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:

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

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:

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

where $J_{y,\hat{\theta}}$ is the Jacobian of $F(\theta) = (f(\theta;x_1),\dots,f(\theta;x_n))$ and $H_{\ell,\hat{\theta}}$ is the hessian of $\ell(f(\theta;x),y)$ in $\theta = \hat{\theta}$.
Note that the damped GNH is always positive definite, as long as $H_{\hat{\theta}}$ is SPD.
We can get the previous formula by linearizing the response function of the regularized loss:

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

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

Another issue is that, in practice, the parameter we utilise is not a minimizer of L. Thus, we want to consider a risk 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:

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

Intuitively, this quantity measures the difference between the evaluation of ℓ on y and the first order Taylor expansion of ℓ around y' computed on y.

Observation. This quantity is always non-negative as long as ℓ is convex. To exemplify, choose $\ell(y,y') = \|y-y'\|^2/2$. This yields:

$$D_{\ell^{(i)}}(y,y') = \|y-y_i\|^2/2 - \|y'-y_i\|^2/2 - \langle\nabla\|y'-y_i\|^2/2, y-y_i\rangle = \|y-y'\|^2/2.$$

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

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

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

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

As a consequence, influence functions are *not* approximating LOO retraining under the original loss L. Instead, they approximate the effect of training from θ^s under the modified objective

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

This fact makes PBRF a more suitable benchmark when testing the performances of IFs.
Concrete examples show that actually PBRF achieves good results in tasks such as mislabeled example detection, making it a viable alternative to LOO retraining.

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.

Influence Functions vs Leverage

For simplicity, let's consider the Ordinary Least Squares problem.
Let $X = (x_1|\dots|x_n)^\top \in \mathbb{R}^{(n\times d)}$, $y = (y_1,\dots,y_n) \in \mathbb{R}^n$, $e = (e_1,\dots,e_n) \in \mathbb{R}^n$, and $\theta \in \mathbb{R}^d$. In this setting, assume the dataset is generated by $y = f(\theta^\star;x) + e = X\,\theta^\star + e$ where e is the observational error. We can estimate $\theta^\star$ with $\hat\theta = (X^\top X)^{-1}X^\top y$. Consequently, the predicted data with our model will be $\hat y = X(X^\top X)^{-1}X^\top y$ and we call $P = X(X^\top X)^{-1}X^\top$, as it is an orthogonal projection on $\operatorname{ran}(X)$.

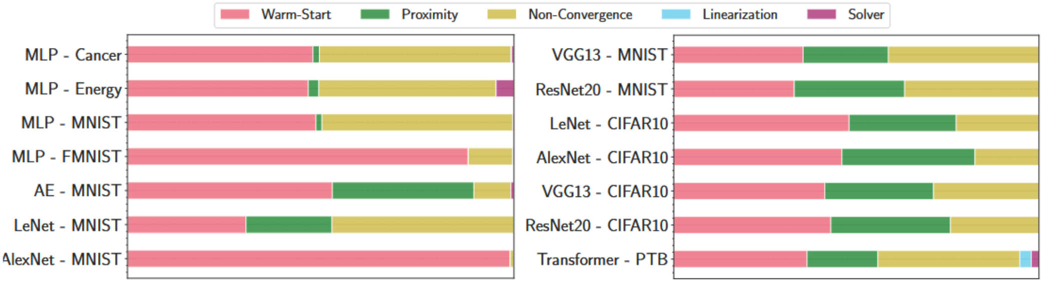

Figure 1: Visual representation of the error defompositionon on different datasets and models. The main focus is that the largest components are the first three.

Definition 3. The *leverage score* of the i-th sample data is $P_{ii} = x_i(X^\top X)^{-1}x_i^\top$.

This quantity describes how much the i-th sample data affects the i-th prediction of our model. The bigger it is, the more probable it is that the i-th sample point is an outlier.

Indeed, if we consider $X = JF(\theta)$ as the X in our case, then the influence loss difference is approximately [1] the same as the leverage. However, the interpretation of the problem would become different.

In any case, an expression of interest that involves both leverage and influence functions is the following (cf. [3]):

$$\hat{\theta} - \hat{\theta}_{1/n,-z} = \frac{(X^\top X)^{-1}x_i\hat{e}_i}{1 - P_{ii}},$$

where $\hat{e}_i = y_i - x_i^\top\hat{\theta}$.

Observe that the $1 - P_{ii}$ at denominator means that outliers also have high influence in the training.

Example: IFs for OLS

Let's compute all the quantities we defined so far in the case of OLS with the euclidean loss.

We have:

- Loss function: $\ell(f(\theta; x_i), y_i) = \frac{1}{2}(y_i - x_i^\top\theta)^2$;
- Gradient: $\nabla\ell(f(\theta; x_i), y_i) = -x_i(y_i - x_i^\top\theta)$;
- Hessian: $H_\theta = \sum_{i=1}^n x_i x_i^\top = X^\top X$;
- Influence function: $\mathcal{I}(z_i; \hat{\theta}) = (X^\top X)^{-1}x_i(y_i - x_i^\top\hat{\theta})$;

[1] we can write $HF(\theta) = JF(\theta)^\top JF(\theta) + \sum \dots$. If we omit the second term, we have the sought approximation. This is acceptable when the parameter is close to optimal, since the sum annihilates for optimal parameters.

- PBRF: $\frac{1}{2}\arg\min_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i^\top(\theta-\hat{\theta}))^2-\frac{1}{n}\|y_i-x_i^\top\theta\|^2+\lambda\|\theta-\hat{\theta}\|^2\right)$; note that in this case $\theta^s=\hat{\theta}$ since we can compute it explicitly.

Reformulation of influence functions

Remark. The object similar to what we are going to discuss, which is present in the paper, is $\hat{r}_z(\varepsilon)$.

Let $\mathcal{X}$, $\mathcal{Y}$ be two measurable spaces and fix $D_n\subset(\mathcal{X}\times\mathcal{Y})^n$ such that $D_n=(z_i)_{i=1}^n$ with $z_i=(x_i,y_i)\in\mathcal{X}\times\mathcal{Y}$.
Given two random variables X and Y respectively on $\mathcal{X}$ and $\mathcal{Y}$, we are interested in studying the distribution of $Y|X$. To do so, we choose a Banach space Θ and a parametric function $f:\Theta\times\mathcal{X}\mapsto\mathcal{Y}$ as an estimator of such distribution.
In order to evaulate the estimators, we give the following definitions:

Definition 4. Given $\ell\in C^2(\mathcal{Y}\times\mathcal{Y};\mathbb{R}_+)$ (called *loss function*), we define the *empirical risk* for the estimator $f(\theta;x)$ on the dataset D_n as:

$$R(\theta):=\sum_{i=1}^{n}\ell(f(\theta;x_i);y_i). \tag{0.1}$$

Given $\varepsilon\in[0,1]$ and $j\in[n]$, we are interested in solving the minimization problem:

$$\min_{\theta\in\Theta}\sum_{i=1}^{n}\ell(f(\theta;x_i);y_i)+(\varepsilon-1)\ell(f(\theta;x_j);y_j).$$

For simplicity, assume that for any ε the objective function has a unique minimum. We call $\theta(\varepsilon)$ such minimum. Let us introduce the notation:

$$\ell_j(\theta)=\ell(f(\theta;x_j);y_j),$$
$$R_j(\theta)=\sum_{\substack{i\in[n]\\ i\neq j}}^{n}\ell_i(\theta).$$

Thus, we can rewrite:

$$\theta(\varepsilon)=\arg\min_{\theta\in\Theta}R_j(\theta)+\varepsilon\ell_j(\theta).$$

Notice that $\theta(\varepsilon)$ symbolizes the parameter for which our model best fits the data while we change the weight of one data point in the training. It is particularly of interest to understand how

$$\theta(1)=\theta^\star=\arg\min R(\theta)$$

and

$$\theta(0)=\theta^\star_{-j}=\arg\min R_j(\theta)$$

are different, since they represent respectively the optimal parameter obtained while training on the whole data and on the dataset minus one particular point.

One way to analyze this, is by the means of influence functions.

Definition 5. The *influence function* of z is:

$$I(z) := \frac{d}{d\varepsilon}\theta(\varepsilon)\bigg|_{\varepsilon=0} = \dot{\theta}(0). \tag{0.2}$$

Remark. The previous quantity is well defined because $\theta(\varepsilon)$ is C^1 thanks to the Implicit function theorem (applied to $\nabla\mathscr{L}(\theta,\varepsilon) = L(\theta) + \varepsilon\ell_j(\theta)$ we get that there exists $\theta(\varepsilon)$ differentiable such that $\nabla\mathscr{L}(\theta(\varepsilon),\varepsilon) = 0$ in a neighbourhood of 1 at least, since by definition $\nabla\mathscr{L}(\theta^\star,1) = \nabla R(\theta^\star) = 0$).
It is not clear yet if we are more interested in $\dot{\theta}(0)$ or $\dot{\theta}(1)$.

Proposition 0.0.1. *We can write $I(z)$ explicitly as:*

$$I(z) = -H_{R_j}^{-1}\nabla_\theta\ell_j(\theta^\star_{-j}) \tag{0.3}$$

where H_{R_j} is the Hessian of R_j in $\theta^\star_{-j}$.

Proof. By definition, $\theta(\varepsilon)$ satisfies:

$$\nabla_\theta(R_j(\theta(\varepsilon)) + \varepsilon\ell_j(\theta(\varepsilon))) = 0.$$

Taking another derivative in ε yields:

$$\begin{aligned} \nabla^2_{\theta,\varepsilon}(R_j(\theta(\varepsilon)) + \varepsilon\ell_j(\theta(\varepsilon))) &= 0 \iff \\ \nabla^2_{\theta\theta}R_j(\theta(\varepsilon))\dot{\theta}(\varepsilon) + \nabla_\theta\ell_j(\theta(\varepsilon)) + \varepsilon\nabla^2_{\theta\theta}\ell_j(\theta(\varepsilon))\dot{\theta}(\varepsilon) &= 0 \iff \\ \dot{\theta}(\varepsilon) = -(\nabla^2_{\theta\theta}(R_j(\theta(\varepsilon)) + \varepsilon\ell_j(\theta(\varepsilon))))^{-1}\nabla_\theta\ell_j(\theta(\varepsilon)). \end{aligned}$$

Evaluating in $\varepsilon = 0$ concludes the proof:

$$\dot{\theta}(0) = -(\nabla^2_{\theta\theta}R_j(\theta(0)))^{-1}\nabla_\theta\ell_j(\theta(0)).$$

■

Remark. If we wanted to compute $\dot{\theta}(1)$, it would also have a nice form:

$$\dot{\theta}(1) = -H_R^{-1}\nabla_\theta\ell_j(\theta^\star).$$

where H_R is the Hessian of R in $\theta^\star$. Indeed, if our goal is to *unlearn* a data point, this formulation does not require us to retrain the model, as we have already access to $\theta^\star$ (but not to $\theta^\star_{-j}$).

It could also be of interest to consider:

$$Q(z) := \frac{d}{d\varepsilon}(R_j(\theta(\varepsilon)) + \varepsilon\ell_j(\theta(\varepsilon)))\bigg|_{\varepsilon=0}. \tag{0.4}$$

Corollary 0.0.1. *The following formulae hold:*

$$\begin{aligned} \frac{d}{d\varepsilon}(R_j(\theta(\varepsilon)) + \varepsilon\ell_j(\theta(\varepsilon)))\bigg|_{\varepsilon=0} &= \ell_j(\theta^\star_{-j}), \\ \frac{d}{d\varepsilon}(\ell_j(\theta(\varepsilon)))\bigg|_{\varepsilon=0} &= -\nabla_\theta\ell_j(\theta^\star_{-j})^\top H_{R_j}^{-1}\nabla_\theta\ell_j(\theta^\star_{-j}), \end{aligned} \tag{0.5}$$

Proof. For the first equality, taking the derivative yields:

$$\nabla_\theta R_j(\theta(\varepsilon))\dot{\theta}(\varepsilon) + \ell_j(\theta(\varepsilon)) + \varepsilon \nabla_\theta \ell_j(\theta(\varepsilon))\dot{\theta}(\varepsilon)$$

and recalling that on $\theta(\varepsilon)$ it holds $\nabla(R_j(\theta(\varepsilon)) + \varepsilon \ell_j(\theta(\varepsilon))) = 0$ we can conclude.
For the second one it suffices to substitute the definition of $\dot{\theta}$ from the previous equation after taking the derivative.

■

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] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression.* Springer, 1983.

[3] Fumio Hayashi. *Econometrics.* Princeton University Press, 2000.

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

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