Appendix: Understanding Black-box Predictions via Influence Functions

Pang Wei Koh [1] **Percy Liang** [1]

A. Deriving the influence function $\mathcal{I}_{\text{up,params}}$

For completeness, we provide a standard derivation of the influence function $\mathcal{I}_{\text{up,params}}$ in the context of loss minimization (M-estimation). This derivation is based on asymptotic arguments and is not fully rigorous; see van der Vaart (1998) and other statistics textbooks for a more thorough treatment.

Recall that $\hat{\theta}$ minimizes the empirical risk:

$$R(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta). \tag{1}$$

We further assume that R is twice-differentiable and strictly convex in θ, i.e.,

$$H_{\hat{\theta}} \stackrel{\text{def}}{=} \nabla^2 R(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta}) \tag{2}$$

exists and is positive definite. This guarantees the existence of $H_{\hat{\theta}}^{-1}$, which we will use in the subsequent derivation.

The perturbed parameters $\hat{\theta}_{\epsilon,z}$ can be written as

$$\hat{\theta}_{\epsilon,z} = \arg\min_{\theta \in \Theta} \left\{ R(\theta) + \epsilon L(z, \theta) \right\}. \tag{3}$$

Define the parameter change $\Delta_\epsilon = \hat{\theta}_{\epsilon,z} - \hat{\theta}$, and note that, as $\hat{\theta}$ doesn't depend on ϵ, the quantity we seek to compute can be written in terms of it:

$$\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} = \frac{d\Delta_\epsilon}{d\epsilon}. \tag{4}$$

Since $\hat{\theta}_{\epsilon,z}$ is a minimizer of (3), let us examine its first-order optimality conditions:

$$0 = \nabla R(\hat{\theta}_{\epsilon,z}) + \epsilon \nabla L(z, \hat{\theta}_{\epsilon,z}). \tag{5}$$

Next, since $\hat{\theta}_{\epsilon,z} \to \hat{\theta}$ as $\epsilon \to 0$, we perform a Taylor expansion of the right-hand side:

$$\begin{aligned} 0 \approx &\left[\nabla R(\hat{\theta}) + \epsilon \nabla L(z, \hat{\theta})\right] + \\ &\left[\nabla^2 R(\hat{\theta}) + \epsilon \nabla^2 L(z, \hat{\theta})\right] \Delta_\epsilon, \end{aligned} \tag{6}$$

where we have dropped $o(\|\Delta_\epsilon\|)$ terms.

Solving for Δ_ϵ, we get:

$$\begin{aligned} \Delta_\epsilon \approx -&\left[\nabla^2 R(\hat{\theta}) + \epsilon \nabla^2 L(z, \hat{\theta})\right]^{-1} \\ &\left[\nabla R(\hat{\theta}) + \epsilon \nabla L(z, \hat{\theta})\right]. \end{aligned} \tag{7}$$

Since $\hat{\theta}$ minimzes R, we have $\nabla R(\hat{\theta}) = 0$. Keeping only $O(\epsilon)$ terms, we have

$$\Delta_\epsilon \approx -\nabla^2 R(\hat{\theta})^{-1} \nabla L(z, \hat{\theta})\epsilon. \tag{8}$$

Combining with (2) and (4), we conclude that:

$$\left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla L(z, \hat{\theta}) \tag{9}$$

$$\stackrel{\text{def}}{=} \mathcal{I}_{\text{up,params}}(z). \tag{10}$$

B. Influence at non-convergence

Consider a training point z. When the model parameters $\tilde{\theta}$ are close to but not at a local minimum, $\mathcal{I}_{\text{up,params}}(z)$ is approximately equal to a constant (which does not depend on z) plus the change in parameters after upweighting z and then taking a single Newton step from $\tilde{\theta}$. The high-level idea is that even though the gradient of the empirical risk at $\tilde{\theta}$ is not 0, the Newton step from $\tilde{\theta}$ can be decomposed into a component following the existing gradient (which does not depend on the choice of z) and a second component responding to the upweighted z (which $\mathcal{I}_{\text{up,params}}(z)$ tracks).

Let $g \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta L(z_i, \tilde{\theta})$ be the gradient of the empirical risk at $\tilde{\theta}$; since $\tilde{\theta}$ is not a local minimum, $g \neq 0$. After upweighting z by ϵ, the gradient at $\tilde{\theta}$ goes from $g \mapsto g + \epsilon \nabla_\theta L(z, \tilde{\theta})$, and the empirical Hessian goes from $H_{\tilde{\theta}} \mapsto H_{\tilde{\theta}} + \epsilon \nabla_\theta^2 L(z, \tilde{\theta})$. A Newton step from $\tilde{\theta}$ therefore changes the parameters by:

$$N_{\epsilon,z} \stackrel{\text{def}}{=} -\left[H_{\tilde{\theta}} + \epsilon \nabla_\theta^2 L(z, \tilde{\theta})\right]^{-1} \left[g + \epsilon \nabla_\theta L(z, \tilde{\theta})\right]. \tag{11}$$

[1]Stanford University, Stanford, CA. Correspondence to: Pang Wei Koh <pangwei@cs.stanford.edu>, Percy Liang <pliang@cs.stanford.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Ignoring terms in ϵg, ϵ^2, and higher, we get $N_{\epsilon,z} \approx -H_{\tilde{\theta}}^{-1}\left(g + \epsilon \nabla_\theta L(z, \tilde{\theta})\right)$. Therefore, the actual change due to a Newton step $N_{\epsilon,z}$ is equal to a constant $-H_{\tilde{\theta}}^{-1} g$ (that doesn't depend on z) plus ϵ times $\mathcal{I}_{\text{up,params}}(z) = -H_{\tilde{\theta}}^{-1} \nabla_\theta L(z, \tilde{\theta})$ (which captures the contribution of z).

References

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, 1998.