Appendix: Understanding Black-box Predictions via Influence Functions

Pang Wei Koh\textsuperscript{1} Percy Liang\textsuperscript{1}

\textbf{A. Deriving the influence function $I_{\text{up, params}}$}

For completeness, we provide a standard derivation of the influence function $I_{\text{up, params}}$ 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:

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

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

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

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

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

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

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

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

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

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

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

$$0 \approx \left[ \nabla R(\hat{\theta}) + \epsilon \nabla L(z, \hat{\theta}) \right] + \left[ \nabla^2 R(\hat{\theta}) + \epsilon \nabla^2 L(z, \hat{\theta}) \right] \Delta_{\epsilon} + o(1) \Delta_{\epsilon}$$

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

Solving for $\Delta_{\epsilon}$, we get:

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

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

$$\Delta_{\epsilon} \approx - \nabla^2 R(\hat{\theta})^{-1} \nabla L(z, \hat{\theta}) \epsilon.$$  \hfill (7)

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

$$\left. \frac{d \hat{\theta}_{\epsilon,z}}{d \epsilon} \right|_{\epsilon=0} = -H_{\theta}^{-1} \nabla L(z, \hat{\theta}) \overset{\text{def}}{=} I_{\text{up, params}}(z). $$

\textbf{B. Influence at non-convergence}

Consider a training point $z$. When the model parameters $\hat{\theta}$ are close to but not at a local minimum, $I_{\text{up, params}}(z)$ 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 $\hat{\theta}$. The high-level idea is that even though the gradient of the empirical risk at $\hat{\theta}$ is not 0, the Newton step from $\hat{\theta}$ 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 $I_{\text{up, params}}(z)$ tracks).

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

$$N_{\epsilon,z} \overset{\text{def}}{=} - \left[ H_{\theta} + \epsilon \nabla^2_{\theta} L(z, \hat{\theta}) \right]^{-1} \left[ g + \epsilon \nabla_{\theta} L(z, \hat{\theta}) \right].$$

\hfill (11)
Ignoring terms in $\epsilon g$, $\epsilon^2$, and higher, we get $N_{\epsilon,z} \approx -H_\theta^{-1} \left( g + \epsilon \nabla_\theta L(z, \hat{\theta}) \right)$. Therefore, the actual change due to a Newton step $N_{\epsilon,z}$ is equal to a constant $-H_\theta^{-1} g$ (that doesn’t depend on $z$) plus $\epsilon$ times $I_{\text{up, params}}(z) = -H_\theta^{-1} \nabla_\theta L(z, \hat{\theta})$ (which captures the contribution of $z$).

References