## 5.1 Large parameters, large curvatures and loss perturbations:

In order to prove that "large parameters are also sensitive parameters" whithout using the theoretical framework we propose, one may consider a perturbation study where large parameters are perturbed, and the resulting loss variation is measured. More variation would then relate to more sensitivity.

We did in fact consider this, but found that such an experiment in general does not yield an informative quantity towards this goal (unlike our proposed Grassmannians). In the following we outline why. *TLDR: The Hessian is not PSD, so negative and positive directions of curvature may cancel out for arbitrary sets of parameters.*

At any parameter $\theta$, and under additive perturbations $\delta$, the loss can be expressed via its Taylor expansion (for gradient $g$, Hessian $H$ and Lagrange remainder $R_3$):

$$\mathcal{L}(\theta+\delta)) = \mathcal{L}(\theta) + g^T\delta + \frac{1}{2}\delta^T H\delta + R_3$$

The proposed perturbation experiment would then consist in sampling $\delta_\xi$ from a noisy distribution, additively perturbing a certain subset of parameters $\xi$ like our extracted masks, and not perturbing the rest. Then, measuring the average impact of said perturbation on the loss approximates the expectation:

$$\mathbb{E}_{\delta_\xi}[\mathcal{L}] = \mathcal{L}(\theta) + \underbrace{g^T\mathbb{E}_{\delta_\xi}[\delta_\xi] + \frac{1}{2}\langle H, \mathbb{E}_{\delta_\xi}[\delta_\xi\delta_\xi^T]\rangle + \mathbb{E}_{\delta_\xi}[R_3]}_{\varepsilon}$$

And if $\varepsilon^2$ is larger, this can be interpreted as the subset $\xi$ being more "sensitive". If we further assume that the perturbations in $\delta_\xi$ are i.i.d. samples with mean $\mu = 0$ (otherwise simply reparametrize $\theta' := \theta + \mu$ such that the mean of the samples is 0) and diagonal covariance $\sigma I_\xi$, we have:

$$\begin{aligned}\mathbb{E}_{\delta_\xi}[\mathcal{L}] &= \mathcal{L}(\theta) + g^T\mu + \frac{1}{2}\langle H, (\sigma I_\xi + \mu\mu^T)\rangle + \mathbb{E}_{\delta_\xi}[R_3] \\ &= \mathcal{L}(\theta) + 0 + \underbrace{\frac{\sigma}{2}tr_\xi(H) + \mathbb{E}_{\delta_\xi}[R_3]}_{\varepsilon}\end{aligned}$$

Since $\sigma I = Cov(\delta) = \mathbb{E}[(\delta-\mu)(\delta-\mu)^T] = \mathbb{E}[\delta\delta^T] - \mu\mu^T$. We see now how, in general, this technique cannot be relied upon. Assume a diagonal, non-PSD $H$, and a quadratic loss such that $R_3 = 0$:

1. For diagonal $H$, each diagonal entry $H_{ii}$ corresponds to the curvature of the $\theta_i$ parameter
2. Since $H$ is non-PSD, such curvatures can be negative, and $tr_\xi(H)$ could equal zero for any arbitrary subset of parameters $\xi$. Since $R_3 = 0$, this would also mean that $\varepsilon^2 = 0$
3. This also includes the cases where $\xi$ corresponds to parameters of large magnitude and curvature (examples can be arbitrarily constructed), therefore this procedure is not guaranteed to yield informative measurements to connect both phenomena.

In contrast, our proposed method is guaranteed to yield larger similarities whenever the selected parameters have larger similarity with the space containing larger directions of curvature, as shown in the paper.