Measuring Data Leakage in Machine-Learning Models with Fisher Information (Supplementary material)

Awni Hannun[1] **Chuan Guo**[1] **Laurens van der Maaten**[1]

[1] Facebook AI Research

A FISHER INFORMATION OF THE GAUSSIAN MECHANISM

We provide a simple derivation of the FIM of the Gaussian mechanism applied to the empirical risk minimizer, $\boldsymbol{w}^*$. The conditional probability density of the output perturbed parameters is given by:

$$p(\boldsymbol{w}' \mid \mathcal{D}) = \int_{\boldsymbol{w}^*} p(\boldsymbol{w}' \mid \boldsymbol{w}^*, \mathcal{D}) p(\boldsymbol{w}^* \mid \mathcal{D}) d\boldsymbol{w}^* = p(\boldsymbol{w}' \mid \boldsymbol{w}^*) \tag{1}$$

where in the last step we use the fact that $\boldsymbol{w}^*$ is sufficient for $\boldsymbol{w}'$. We also assume $f(\mathcal{D})$ is deterministic, and hence $p(\boldsymbol{w}^* \mid \mathcal{D})$ is a (shifted) delta function nonzero at the optimal parameters, $\boldsymbol{w}^*$.

Using equation 1, the gradient of $\log p(\boldsymbol{w}' \mid \mathcal{D})$ with respect to $\mathcal{D}$ is given by:

$$\nabla_{\mathcal{D}} \log p(\boldsymbol{w}' \mid \mathcal{D}) = \boldsymbol{J}_f^\top \nabla_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*) \tag{2}$$

where $\boldsymbol{J}_f$ is the Jacobian of $f(\mathcal{D})$ with respect to $\mathcal{D}$. The Hessian is:

$$\begin{aligned} &\nabla^2_{\mathcal{D}} \log p(\boldsymbol{w}' \mid \mathcal{D}) = \\ &\boldsymbol{J}_f^\top \nabla^2_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*) \boldsymbol{J}_f + \mathsf{H} \nabla_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*) \end{aligned} \tag{3}$$

where H is the three-dimensional tensor of second-order derivatives (in a slight abuse of notation $\mathsf{H}_{ijk} = \frac{\partial^2 f_k}{\partial \mathcal{D}_i \mathcal{D}_j}$). Using the second-order expression for the FIM requires evaluating the expectation over $\boldsymbol{w}'$ of equation 3.

When using zero-mean isotropic Gaussian noise for the perturbation, $\mathcal{N}(0, \sigma^2 \boldsymbol{I})$, the expectation over $\boldsymbol{w}'$ of equation 3 simplifies. The gradient of $\log p(\boldsymbol{w}' \mid \boldsymbol{w}^*)$ is:

$$\nabla_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*) = \frac{\boldsymbol{w}' - \boldsymbol{w}^*}{\sigma^2}, \tag{4}$$

and hence the Hessian is:

$$\nabla^2_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*) = -\frac{1}{\sigma^2} \boldsymbol{I}. \tag{5}$$

Evaluating the expectation of equation 3 using the above expressions yields:

$$\begin{aligned} &\mathbb{E}\left[\boldsymbol{J}_f^\top \nabla^2_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*) \boldsymbol{J}_f + \mathsf{H} \nabla_{\boldsymbol{w}^*} \log p(\boldsymbol{w} \mid \boldsymbol{w}^*)\right] = \\ &\boldsymbol{J}_f^\top \mathbb{E}\left[\nabla^2_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*)\right] \boldsymbol{J}_f + \mathsf{H} \mathbb{E}\left[\nabla_{\boldsymbol{w}^*} \log p(\boldsymbol{w}' \mid \boldsymbol{w}^*)\right] = \\ &-\frac{1}{\sigma^2} \boldsymbol{J}_f^\top \boldsymbol{J}_f, \end{aligned}$$

where the second term vanishes since $\mathbb{E}[\boldsymbol{w}'] = \boldsymbol{w}^*$. Hence the FIM is given by:

$$\mathcal{I}_{\boldsymbol{w}'}(\mathcal{D}) = -\mathbb{E}\left[\nabla^2_{\mathcal{D}} \log p(\boldsymbol{w}' \mid \mathcal{D})\right] = \frac{1}{\sigma^2} \boldsymbol{J}_f^\top \boldsymbol{J}_f. \tag{6}$$

B JACOBIAN OF THE MINIMIZER

Let $\ell(\boldsymbol{w}^\top \boldsymbol{x}, y)$ be a convex, twice-differentiable loss function. Let $f_i(\boldsymbol{x}, y)$ denote the minimizer of the regularized empirical risk as a function of $(\boldsymbol{x}, y)$ at the i-th example:

$$f_i(\boldsymbol{x}, y) = \arg\min_{\boldsymbol{w}} \sum_{j \neq i} \ell(\boldsymbol{w}^\top \boldsymbol{x}_j, y_j) + \ell(\boldsymbol{w}^\top \boldsymbol{x}, y) + \frac{n\lambda}{2} \|\boldsymbol{w}\|_2^2. \tag{7}$$

We aim to derive an expression for $\boldsymbol{J}_{f_i}\big|_{x_i, y_i}$, the Jacobian of $f_i(\boldsymbol{x}, y)$ with respect to $(\boldsymbol{x}, y)$ evaluated at $(\boldsymbol{x}_i, y_i)$. Taking the gradient of equation 7 with respect to $\boldsymbol{w}$ and setting it to 0 gives an implicit function for $\boldsymbol{w}^* = f_i(\boldsymbol{x}, y)$:

$$0 = \sum_{j \neq i} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_j, y_j) + \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}, y) + n\lambda \boldsymbol{w}^*. \tag{8}$$

Implicit differentiation of equation 8 with respect to $(\boldsymbol{x}, y)$ gives:

$$0 = \sum_{j \neq i} \nabla^2_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_j, y_j) \boldsymbol{J}_{f_i} + \nabla_{\boldsymbol{x},y} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}, y) + n\lambda \boldsymbol{J}_{f_i}. \tag{9}$$

The second term can be computed using the product rule:

$$\begin{aligned} &\nabla_{\boldsymbol{x},y} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}, y) = \\ &\qquad \nabla^2_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}, y) \boldsymbol{J}_{f_i} + \nabla_{\boldsymbol{x},y} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^\top \boldsymbol{x}, y)\Big|_{\boldsymbol{w}=\boldsymbol{w}^*}. \end{aligned} \tag{10}$$

Evaluating equation 10 at $(\boldsymbol{x}_i, y_i)$ and substituting into equation 9 yields:

$$0 = \left[\sum_{j=1}^{n} \nabla^2_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top}\boldsymbol{x}_j, y_j)\boldsymbol{J}_{f_i} + \nabla_{\boldsymbol{x},y}\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}^{\top}\boldsymbol{x}, y) + n\lambda\boldsymbol{J}_{f_i}\right]_{\boldsymbol{w}^*,\boldsymbol{x}_i,y_i}$$

$$= \left[\boldsymbol{H}_{\boldsymbol{w}^*}\boldsymbol{J}_{f_i} + \nabla_{\boldsymbol{x},y}\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}^{\top}\boldsymbol{x}, y)\right]_{\boldsymbol{w}^*,\boldsymbol{x}_i,y_i}, \qquad (11)$$

where the Hessian $\boldsymbol{H}_{\boldsymbol{w}^*} = \sum_{j=1}^{n} \nabla^2_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top}\boldsymbol{x}_j, y_j) + n\lambda\boldsymbol{I}$. Solving for $\boldsymbol{J}_{f_i}$ yields:

$$\boldsymbol{J}_{f_i}\Big|_{\boldsymbol{x}_i,y_i} = -\boldsymbol{H}^{-1}_{\boldsymbol{w}^*}\nabla_{\boldsymbol{x},y}\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}^{*\top}\boldsymbol{x}_i, y_i). \qquad (12)$$

References