Supplementary Material: The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers

Mohamed El Amine Seddik[1] Cosme Louart[2,3] Romain Couillet[3] Mohamed Tamaazousti[2]
[1]LIX, Ecole Polytechnique [2]CEA List [3]GIPSA-Lab Grenoble-Alps University

February 17, 2021

Outline: This supplementary material provides the essential derivations to obtain the results of the main paper (Section 1), the extension of these results to the Softmax classifier (Section 2) and further experiments supporting our findings (Section 3). Specifically, Subsection 1.1 provides justifications about the CLT result for concentrated vectors as per Theorem 3.2. Subsection 1.2 next presents the main ingredients for obtaining Theorem 4.8. Notations about the Softmax classifier are presented In Subsection 2.1, while Subsection 2.2 provides an analog of Theorem 4.8 for the general case of the Softmax classifier. Finally, further experiments using multi-class classification with synthetic Gaussian data and MNIST data as well as CNN representations of real images are shown in Section 3.

Notation: The notation $z \in \tilde{z} \pm \mathcal{E}_q(\sigma)$ stands for $z \in \mathbb{R}$ satisfying $z \propto \mathcal{E}_q(\sigma)$ and $|\tilde{z} - \mathbb{E}[z]| \leq \mathcal{O}(\sigma)$, where $|\cdot|$ replaces Euclidean norm for vectors. $\otimes$ stands for the Kronecker product.

Reproducibility: A github link for the project will be provided.

1 Proofs of the main paper results

1.1 CLT for concentrated random vectors

In this section, we consider an asymptotic quantity $p \in \mathbb{N}$ that will represent the dimension of the random vectors we consider, and will be the quantity implicitly tending to infinity when we will employ the notations $\mathcal{O}(\sigma)$ of $\mathcal{E}_q(\sigma)$. Let us first provide two fundamental result that were already evoked in the main paper: the concentration of the Gaussian distribution $\gamma = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and the uniform distribution on the Sphere $\sqrt{p}\mathbb{S}^{p-1} \equiv \{\boldsymbol{x} \in \mathbb{R}^p, \|\boldsymbol{x}\| = \sqrt{p}\}$.

Theorem 1.1 (Concentration of the Gaussian distribution and the uniform measure on the sphere)**.** *If $\boldsymbol{x} \sim \gamma$ or $\boldsymbol{x} \sim Unif(\sqrt{p}\mathbb{S}^{p-1})$, then $\boldsymbol{x} \propto \mathcal{E}_2$.*

Those two distributions are very similar and almost equal when p is high[1]. In general one can imagine concentrated vectors $\boldsymbol{x}$ as distributed along a sphere of center $\mathbb{E}[\boldsymbol{x}]$ and radius $\sqrt{\mathrm{Tr}(\boldsymbol{C}_{\boldsymbol{x}})}$ where

$$\boldsymbol{C}_{\boldsymbol{x}} \equiv \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\intercal] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal,$$

thanks to the following lemma that is a mere consequence of the 1-Lipschitz character of the Euclidean norm and whose full proof can be found in [Led05].

Lemma 1.2 (Concentrated vectors are distributed close to the sphere)**.** *If $\boldsymbol{x} \propto \mathcal{E}_2$, then:*

$$\mathbb{P}\left(\left|\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\| - \sqrt{\mathrm{Tr}(\boldsymbol{C}_{\boldsymbol{x}})}\right|\right) \leq Ce^{-ct^2}.$$

We have the following result that can also be found in [LC20a], which controls the moments of concentrated random variables.

[1]A famous result of Poincaré states indeed that any orthogonal projection of a random vector uniformly distributed on the sphere $\sqrt{p}\mathbb{S}^{p-1}$ tends to a Gaussian vector when p tends to infinity.

Proposition 1.3 (Characterization of the concentration with a bound on the moments)**.** *A random variable z follows the concentration $z \in \tilde{z} \pm \mathcal{E}_q(\sigma)$ if and only if there exists two constants $C, c > 0$ such that for all $p \in \mathbb{N}$, for all $r > q$:*

$$\mathbb{E}\left[|z - \tilde{z}|^r\right] \leq C \left(\frac{r}{q}\right)^{\frac{r}{q}} (c\sigma)^r.$$

We provide below a proposition (similar to the result of [Kla07] provided in Theorem 3.2 in the main paper but more simple) supporting the fact that concentrated vectors share a common *concentration* behavior with Gaussian vectors. Specifically, given a random concentrated vector $\boldsymbol{x} \propto \mathcal{E}_2$, if one samples a vector $\boldsymbol{u}$ from the uniform distribution on the unit sphere sphere[2] $\mathbb{S}^{p-1}$, then most likely, the random variable $\boldsymbol{u}^\intercal \boldsymbol{x}$ behaves asymptotically (as p grows) as a Gaussian random variable. Technically, this observation allows us to compute expectations of functionals of $\boldsymbol{w}^\intercal \boldsymbol{x}$ when $\boldsymbol{x}$ is independent from $\boldsymbol{w}$ (for instance in $\boldsymbol{w}_{-i}^\intercal \boldsymbol{x}_i$). Note however, that even though $\boldsymbol{w}$ is *a priori* not guaranteed to be uniformly distributed on the unit sphere, our experiments (in the main paper and subsequently) tend to validate the Gaussianity of the random variable $\boldsymbol{w}^\intercal \boldsymbol{x}$.

Proposition 1.4. *Given an integer $p \in \mathbb{N}$ and a random vector $\boldsymbol{x} \in \mathbb{R}^p$, if $\boldsymbol{x} \propto \mathcal{E}_2$ and $\|\mathbb{E}[\boldsymbol{x}]\| \leq \mathcal{O}(1)$, then for any parameter $\kappa \geq 1$ there exists two constants $c, C > 0$ and a subset $\Theta \subset \mathbb{R}^p$ such that $\gamma(\Theta) \geq 1 - 2e^{-c\kappa^2}$, $\gamma = \mathcal{N}(\boldsymbol{0}, \frac{1}{p}\boldsymbol{I}_p)$ and for any λ-Lipschitz mapping $f : \mathbb{R} \to \mathbb{R}$:*

$$\forall \boldsymbol{u} \in \Theta : |\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})] - \mathbb{E}[f(z)]| \leq \frac{C\kappa\lambda}{\sqrt{p}},$$

where $z \sim \mathcal{N}(0, \frac{1}{p}\mathbb{E}[\|\boldsymbol{x}\|])$ or $z \sim \mathcal{N}(0, \frac{1}{p}\mathbb{E}[\mathrm{Tr}(\boldsymbol{C})])$, with $\boldsymbol{C} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\intercal] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\intercal$.

Remark 1.5. *If $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{p}\boldsymbol{I}_p)$, then $\boldsymbol{\theta} \propto \mathcal{E}_2(1/\sqrt{p})$ and $\mathbb{P}(|\|\boldsymbol{\theta}\| - 1| \geq t) \leq 2e^{-pt^2/2}$. Therefore, considering a vector $\boldsymbol{u} \in \mathbb{R}^p$, if $\boldsymbol{u}$ is far from the sphere $\mathbb{S}^{p-1}$, although it could possibly be in the set Θ, it is highly unprobable. Our heuritic thus requires to apply Proposition 1.4 with the vector $\boldsymbol{u}/\|\boldsymbol{u}\|$ and the mapping $t \mapsto f(\|\boldsymbol{u}\|t)$.*

Proof. The set Θ appears from a concentration result where the deterministic vector $\boldsymbol{u}$ can be seen as a drawing of the random vector $\boldsymbol{\theta} \sim \gamma$ that we chose to be independent of $\boldsymbol{x}$. We already know that $\boldsymbol{\theta} \propto \mathcal{E}_2(1/\sqrt{p})$ and we know that $\phi : \boldsymbol{u} \mapsto \mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})]$ is λ'-Lipschitz with:

$$\lambda' \leq \lambda \sup_{\|\boldsymbol{v}\| \leq 1} \mathbb{E}[|\boldsymbol{v}^\intercal \boldsymbol{x}|] \leq \lambda(|\mathbb{E}[\boldsymbol{v}^\intercal \boldsymbol{x}]| + \mathcal{O}(1)) \leq \mathcal{O}(\lambda)$$

thanks to Propostion 1.3, since $\boldsymbol{v}^\intercal \boldsymbol{x} \in 0 \pm \mathcal{E}_2$. Therefore $\phi(\boldsymbol{\theta}) \in \mathbb{E}[\phi(\boldsymbol{\theta})] \pm \mathcal{E}_2(\lambda/\sqrt{p})$ thanks to the concentration of $\boldsymbol{\theta}$ and noting $\Theta = \{\boldsymbol{u} \in \mathbb{R}^p \mid \phi(\boldsymbol{u}) - \mathbb{E}[\phi(\boldsymbol{\theta})]| \leq \frac{\lambda\kappa}{\sqrt{p}}\}$, we know that there exist two constants $C, c > 0$ such that $\mathbb{P}(\boldsymbol{\theta} \notin \Theta) \leq 2e^{-c\kappa^2}$.

Besides, noting $\mathbb{E}_{\boldsymbol{x}}$ and $\mathbb{E}_{\boldsymbol{\theta}}$, respectively, the expectation on $\boldsymbol{x}$ (with $\boldsymbol{\theta}$ fixed) and the expectation on $\boldsymbol{\theta}$, we can estimate:

$$\mathbb{E}[\phi(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{\theta}^\intercal \boldsymbol{x})]\right].$$

For a fixed $\boldsymbol{x}$, $\boldsymbol{\theta}^\intercal \boldsymbol{x}$ is a Gaussian random variable with mean equal to 0 and variance equal to $\frac{1}{p}\|\boldsymbol{x}\|^2$ (since $\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal] = \frac{1}{p}\boldsymbol{I}_p$). Introducing a random variable $z \sim \mathcal{N}(0, 1)$, independent of $\boldsymbol{x}$, we have thus the identity $\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{\theta}^\intercal \boldsymbol{x})] = \mathbb{E}_z[f(\|\boldsymbol{x}\|z/\sqrt{p})]$. Finally, we have the following bound thanks to Proposition 1.3:

$$\left|\mathbb{E}\left[f\left(\frac{\|\boldsymbol{x}\|z}{\sqrt{p}}\right)\right] - \mathbb{E}_z\left[\frac{f(\mathbb{E}[\|\boldsymbol{x}\|]z)}{\sqrt{p}}\right]\right| \leq \frac{\lambda}{\sqrt{p}}\mathbb{E}\left[|\|\boldsymbol{x}\| - \mathbb{E}[\|\boldsymbol{x}\|]|\right]\mathbb{E}[|z|] \leq \mathcal{O}\left(\frac{\lambda}{\sqrt{p}}\right).$$

Therefore, choosing well the constant $C > 0$ provides the result of the proposition. □

However, the above proposition needs to be adapted for estimating quantities of the form $\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{v}^\intercal \boldsymbol{x}]$ or $\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{v}^\intercal \boldsymbol{x}\boldsymbol{x}^\intercal \boldsymbol{w}]$ for some deterministic vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^p$, as we will need subsequently. In particular, we have the following result.

[2] Which is equivalent to the Gaussian distribution $\mathcal{N}(0, \frac{\boldsymbol{I}_p}{n})$.

Proposition 1.6. *In the setting of Proposition 1.4, let us note $\nu = Unif(\mathbb{S}^{p-1})$, for any parameter $\kappa > 0$, there exists two constants $C, c > 0$ and a subset $\Theta_2 \subset (\mathbb{R}^p)^2$ such that*[3] $\nu^{\times 2}(\Theta_2) \geq 1 - 2e^{-c\kappa^2}$ *and for any λ-Lipschitz mapping $f : \mathbb{R} \to \mathbb{R}$ such that $|\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})]| \leq \lambda$, we have*

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \Theta_2 : |\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{v}^\intercal \boldsymbol{x}]| \leq \frac{C\kappa\lambda}{\sqrt{p}}.$$

Remark 1.7. *Given two independent random vectors $\boldsymbol{\theta}, \boldsymbol{\sigma} \sim \mathcal{N}(0, \frac{\boldsymbol{I}_p}{p})$, the couple $(\boldsymbol{\theta}, \boldsymbol{\sigma}) \in (\mathbb{R}^p)^2$ has the distribution $\gamma^{\otimes 2}$ and one can show that*

$$\mathbb{P}\left(|\boldsymbol{\theta}^\intercal \boldsymbol{\sigma}| \geq t\right) \leq C' e^{-c' p t^2} + C' e^{-c' p t},$$

for some constants $C', c' > 0$ (independent of p). The same result particularly holds for $(\boldsymbol{\theta}, \boldsymbol{\sigma}) \sim \nu^{\times 2}$. Thus, we see that for sufficiently large κ, the couples of Θ_2 have high probability to be quasi-orthogonal. Hence, if one considers two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$, our heuristic requires not to employ directly Proposition 1.6 as if $(\boldsymbol{u}, \boldsymbol{v}) \in \Theta_2$, but rather employ it with the couple $(\boldsymbol{u}, \boldsymbol{v} - (\boldsymbol{u}^\intercal \boldsymbol{v})\boldsymbol{u})$, we end up therefore having the identity

$$\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{v}^\intercal \boldsymbol{x}] = \boldsymbol{u}^\intercal \boldsymbol{v} \mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{u}^\intercal \boldsymbol{x}] + \mathcal{O}\left(\frac{\kappa\lambda}{\sqrt{p}}\right),$$

where the right hand side expectation can be estimated by Proposition 1.4, through the mapping $t \mapsto f(t)t$.

Proof. This proof follows the same steps as the proof of Proposition 1.4. Considering $(\boldsymbol{\theta}, \boldsymbol{\sigma}) \sim \nu^{\times 2}$, we know that $(\boldsymbol{\theta}, \boldsymbol{\sigma}) \propto \mathcal{E}_2(1/\sqrt{p})$ ($\nu^{\times 2}$ is similar to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{2p}/2p)$), and the mapping $(\boldsymbol{u}, \boldsymbol{v}) \mapsto \mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{v}^\intercal \boldsymbol{x}]$ is λ'-Lipschitz for the Euclidean norm of $(\mathbb{R}^p)^2$, where we can bound thanks to Proposition 1.3 and Hölder inequality:

$$\begin{aligned}
\lambda' &= \sqrt{\|\mathbb{E}[\lambda \boldsymbol{v}^\intercal \boldsymbol{x}\boldsymbol{x}]\|^2 + \|\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{x}]\|^2} \\
&\leq \sqrt{\sup_{\|\boldsymbol{a}\| \leq 1} \lambda^2 \left|\mathbb{E}[(\boldsymbol{v}^\intercal \boldsymbol{x})^2]\mathbb{E}[(\boldsymbol{a}^\intercal \boldsymbol{x})^2]\right| + \left|\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})^2]\mathbb{E}[(\boldsymbol{a}^\intercal \boldsymbol{x})^2]\right|} \leq \mathcal{O}(\lambda).
\end{aligned}$$

Therefore, thanks again to Proposition 1.3, if we note:

$$\Theta_2 = \left\{ (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^p \;\middle|\; |\mathbb{E}[f(\boldsymbol{u}^\intercal \boldsymbol{x})\boldsymbol{v}^\intercal \boldsymbol{x}] - \mathbb{E}[f(\boldsymbol{\theta}^\intercal \boldsymbol{x})\boldsymbol{\sigma}^\intercal \boldsymbol{x}]| \leq \frac{\kappa\lambda}{\sqrt{p}} \right\},$$

then, there exist two constants $C, c > 0$ such that $\mathbb{P}((\boldsymbol{\theta}, \boldsymbol{\sigma}) \in \Theta_2) \leq Ce^{-c\kappa^2}$ and we retrieve the result of the proposition since (with the same notations as in the proof of Proposition 1.4):

$$\mathbb{E}[f(\boldsymbol{\theta}^\intercal \boldsymbol{x})\boldsymbol{\sigma}^\intercal \boldsymbol{x}] = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{\theta}^\intercal \boldsymbol{x})]\mathbb{E}_{\boldsymbol{\sigma}}[\boldsymbol{\sigma}^\intercal \boldsymbol{x}]\right] = 0.$$

□

1.2 Estimation of the weight statistics

The complete justifications leading to Theorem 4.8 are quite long and already provided in [LC20b] in the Gaussian setting. We thus only provide here the main ingredients to have some intuitions on the results. Typically, Propositions 1.4 and 1.6 will allow us to employ Stein-like identities as the ones given in the up coming proposition. They are specifically provided for Gaussian vectors $\boldsymbol{x}$, but, as we saw from the previous subsection, they can be extended to any concentrated vector, as long as $\boldsymbol{x} \propto \mathcal{E}_2$ and $\|\mathbb{E}[\boldsymbol{x}]\| \leq \mathcal{O}(1)$.

Proposition 1.8 (Stein identities)**.** *Given $\boldsymbol{x} \in \mathbb{R}^p$, a Gaussian random vector satisfying $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ and, $f : \mathbb{R}^p \to \mathbb{R}$ some three times differentiable function such that f, f' and f'' are all λ-Lipschitz with $\lambda \leq \mathcal{O}(1)$, then for any $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^p$ and any $A \in \mathcal{M}_p$, we have*

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{w}^\intercal \boldsymbol{x})\boldsymbol{v}^\intercal \boldsymbol{x}] &= \mathbb{E}[f(\boldsymbol{w}^\intercal \boldsymbol{x})]\boldsymbol{v}^\intercal \boldsymbol{\mu} + \mathbb{E}[f'(\boldsymbol{w}^\intercal \boldsymbol{x})]\boldsymbol{v}^\intercal \boldsymbol{C}\boldsymbol{w}, \\
\mathbb{E}[f(\boldsymbol{w}^\intercal \boldsymbol{x})\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\intercal \boldsymbol{A})] &= \mathbb{E}[f(\boldsymbol{w}^\intercal \boldsymbol{x})]\operatorname{Tr}\left(\boldsymbol{A}\left(\boldsymbol{\mu}\boldsymbol{\mu}^\intercal + \boldsymbol{C}\right)\right) + \mathbb{E}[f'(\boldsymbol{w}^\intercal \boldsymbol{x})]\boldsymbol{\mu}^\intercal (\boldsymbol{A} + \boldsymbol{A}^\intercal)\boldsymbol{C}\boldsymbol{w} + \mathbb{E}[f''(\boldsymbol{w}^\intercal \boldsymbol{x})]\boldsymbol{w}^\intercal \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{w}.
\end{aligned}$$

[3] $\nu^{\times 2} = \nu \times \nu$ stands for the measure product.

Proof. The proposition is a straightforward result of applying the Stein's identity; for $z \sim \mathcal{N}(0, \sigma^2)$ and $f : \mathbb{R} \to \mathbb{R}$ differentiable, $\mathbb{E}[f(z)z] = \sigma^2[f'(z)]$. □

Remark 1.9. *In the case where $\boldsymbol{x}$ is not Gaussian but rather only concentrated, such that $\boldsymbol{x} \propto \mathcal{E}_2$ and $\|\boldsymbol{\mu}\| \leq \mathcal{O}(1)$[4], the formulas in Proposition 1.8 become estimations with a vanishing error of order $\mathcal{O}(n^{-\frac{1}{2}})$[5] when $\|\boldsymbol{v}\| \leq \mathcal{O}(1)$ and $\|\boldsymbol{A}\| \leq \mathcal{O}(1/n)$. In particular, normalizing the second inequality with n and considering $\boldsymbol{A} \in \mathcal{M}_p$ such that $\|\boldsymbol{A}\| \leq 1$, we have*

$$\frac{1}{n}\mathbb{E}[f(\boldsymbol{w}^\intercal \boldsymbol{x})\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x}] = \frac{1}{n}\mathbb{E}[f(\boldsymbol{w}^\intercal \boldsymbol{x})] \operatorname{Tr}(\boldsymbol{A}\boldsymbol{C}) + \mathcal{O}\left(n^{-\frac{1}{2}}\right),$$

since $\frac{1}{n}\boldsymbol{\mu}^\intercal \boldsymbol{A}\boldsymbol{\mu} \leq \mathcal{O}(\frac{1}{n})$, $|\frac{1}{n}\boldsymbol{\mu}^\intercal(\boldsymbol{A}+\boldsymbol{A}^\intercal)\boldsymbol{C}\boldsymbol{w}| \leq \mathcal{O}(\frac{1}{n})$ and $|\frac{1}{n}\mathbb{E}[f''(\boldsymbol{w}^\intercal \boldsymbol{x})]\boldsymbol{w}^\intercal \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{w}| \leq \mathcal{O}(\frac{1}{n})$.

We turn now to the estimation of the weight statistics, i.e., the quantities $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\boldsymbol{C}_{\boldsymbol{w}}$. Starting from the following estimation

$$\left\| \boldsymbol{\mu}_{\boldsymbol{w}} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i})\boldsymbol{x}_i] \right\| \leq \mathcal{O}(n^{-\frac{1}{2}}), \tag{1}$$

we apply Proposition 1.8 (more precisely Remark 1.9) which yields to; for most of vectors $\boldsymbol{v} \in \mathbb{R}^p$, such that $\|\boldsymbol{v}\| \leq \mathcal{O}(1)$,

$$\mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i})\boldsymbol{v}^\intercal \boldsymbol{x}_i] = \mathbb{E}[\xi_{\delta_{k(i)}}(\tilde{z}_{k(i)})]\boldsymbol{v}^\intercal \boldsymbol{\mu}_{k(i)} + \mathbb{E}[\xi'_{\delta_{k(i)}}(\tilde{z}_{k(i)})]\boldsymbol{v}^\intercal \boldsymbol{C}_{k(i)}\mathbb{E}_{\mathcal{A}_{\boldsymbol{w}}}[\boldsymbol{w}_{-i}] + \mathcal{O}\left(n^{-\frac{1}{2}}\right),$$

where, this time, $\tilde{z}_{k(i)}$ is a Gaussian random variable having the same mean and variance as $z_i \equiv \boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i}$. Moreover, to show that $\boldsymbol{w}_{-i}$ has the same statistics as $\boldsymbol{w}$, we need a supplementary result (Theorem 1.10) from [LC20b], which in turn leads to Theorem 4.4.

Theorem 1.10. *Under Assumption 1-4, for any $i \in [n]$: $\|\boldsymbol{w} - \boldsymbol{w}_{-i} - \frac{1}{n}\boldsymbol{Q}_{-i}\boldsymbol{x}_i f(\boldsymbol{x}_i^\intercal \boldsymbol{w})\| \in 0 \pm \mathcal{E}_2(n^{-\frac{1}{2}})$.*

Thus, since for any deterministic vector $\boldsymbol{u} \in \mathbb{R}^p$ such that $\|\boldsymbol{u}\| \leq \mathcal{O}(1)$, one can show that $\frac{1}{n}\boldsymbol{u}^\intercal \boldsymbol{Q}_{-i}\boldsymbol{x}_i \in 0 \pm \mathcal{E}_2(n^{-\frac{1}{2}})$, one can then deduce that $\boldsymbol{u}^\intercal \boldsymbol{w} - \boldsymbol{u}^\intercal \boldsymbol{w}_{-i} \in 0 \pm \mathcal{E}_2(n^{-\frac{1}{2}})$, which implies in particular that $\|\mathbb{E}_{\mathcal{A}_{\boldsymbol{w}}}[\boldsymbol{w}_{-i}] - \boldsymbol{\mu}_{\boldsymbol{w}}\| \leq \mathcal{O}(n^{-\frac{1}{2}})$. And studying now functionals of the form $\boldsymbol{w}^\intercal \boldsymbol{A}\boldsymbol{w}$, with $\boldsymbol{A} \in \mathcal{M}_p$ such that $\|\boldsymbol{A}\| \leq \mathcal{O}(1)$, one can also show that $\|\mathbb{E}_{\mathcal{A}_{\boldsymbol{w}}}[\boldsymbol{w}_{-i}\boldsymbol{w}_{-i}^\intercal] - \boldsymbol{C}_{\boldsymbol{w}}\| \leq \mathcal{O}(n^{-\frac{1}{2}})$.

Hence, setting $\boldsymbol{K} \equiv \sum_{\ell=1}^{k} \gamma_\ell \mathbb{E}[\xi'_{\delta_\ell}(\tilde{z}_\ell)]\boldsymbol{C}_\ell$ and $\tilde{\boldsymbol{\mu}} \equiv \sum_{\ell=1}^{k} \gamma_\ell \mathbb{E}[\xi_{\delta_\ell}(\tilde{z}_\ell)]\boldsymbol{\mu}_\ell$, we obtain from (1),

$$\|\boldsymbol{\mu}_{\boldsymbol{w}} - \tilde{\boldsymbol{\mu}} - \boldsymbol{K}\boldsymbol{\mu}_{\boldsymbol{w}}\| \leq \mathcal{O}\left(n^{-\frac{1}{2}}\right). \tag{2}$$

Now looking for an estimation of $\boldsymbol{C}_{\boldsymbol{w}}$, we start from the identity:

$$\left\| \boldsymbol{C}_{\boldsymbol{w}} + \boldsymbol{\mu}_{\boldsymbol{w}}\boldsymbol{\mu}_{\boldsymbol{w}}^\intercal - \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i})^2 \boldsymbol{x}_i\boldsymbol{x}_i^\intercal] - \frac{1}{n^2}\sum_{\substack{1\leq i,j\leq n \\ i\neq j}} \mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i})\xi_{\delta_{k(j)}}(\boldsymbol{x}_j^\intercal \boldsymbol{w}_{-j})\boldsymbol{x}_i\boldsymbol{x}_j^\intercal] \right\|_* \leq \mathcal{O}(n^{-\frac{1}{2}}). \tag{3}$$

Then, we first deduce as per Remark 1.9 that for any $\ell \in [k]$:

$$\left\|\mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(j)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i})^2 \boldsymbol{x}_i\boldsymbol{x}_i^\intercal] - \mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\tilde{z}_{k(i)})^2]\boldsymbol{C}_{k(i)}\right\| \leq \mathcal{O}(n^{-\frac{1}{2}}).$$

For any $i, j \in [n]$, $i \neq j$, let us define $\boldsymbol{w}_{-i,j}$ as the unique solution $\boldsymbol{w}_{-i,j} = \frac{1}{n}\boldsymbol{X}_{-i,j} f(\boldsymbol{X}_{-i,j}^\intercal \boldsymbol{w}_{-i,j})$, where $\boldsymbol{X}_{-i,j}$ is the matrix $\boldsymbol{X}_{-i}$ with a zero vector in the j^{th} column. Then we can show from Theorem 1.10 that

[4]The concentration $\boldsymbol{x} \propto \mathcal{E}_2$ automatically implies that $\|\boldsymbol{C}\| \leq \mathcal{O}(1)$. Indeed for all $\boldsymbol{v} \in \mathbb{R}^p$, $\boldsymbol{v}^\intercal \boldsymbol{x} \in \boldsymbol{v}^\intercal \boldsymbol{\mu} \pm \mathcal{E}_2$ and thus, Proposition 1.3 implies $\|\boldsymbol{C}\| = \sup_{\|\boldsymbol{v}\|\leq 1} \boldsymbol{v}^\intercal \boldsymbol{C}\boldsymbol{v} = \sup_{\|\boldsymbol{v}\|\leq 1} \mathbb{E}[(\boldsymbol{v}^\intercal \boldsymbol{x} - \boldsymbol{v}^\intercal \boldsymbol{\mu})^2] \leq \mathcal{O}(1)$.

[5]Recall that p (the dimension of the data) and n (the number of data) are of the same order i.e $p = \mathcal{O}(n)$ and $n = \mathcal{O}(p)$.

$\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i} - \boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i,j} \in 0 \pm \mathcal{E}_2(1/\sqrt{n})$ (since $\boldsymbol{x}_i$ is mutually independent with $\boldsymbol{w}_{-i}$ and $\boldsymbol{w}_{-i,j}$), thus

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i})\xi_{\delta_{k(j)}}(\boldsymbol{x}_j^\intercal \boldsymbol{w}_{-j})\boldsymbol{x}_i^\intercal \boldsymbol{A}\boldsymbol{x}_j] \\
&\quad = \mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i,j})\xi_{\delta_{k(j)}}(\boldsymbol{x}_j^\intercal \boldsymbol{w}_{-i,j})\boldsymbol{x}_i^\intercal \boldsymbol{A}\boldsymbol{x}_j] + \mathcal{O}(n^{-\frac{1}{2}}) \\
&\quad = \left(\mathbb{E}[\xi_{\delta_{k(i)}}(\tilde{z}_{k(i)})]\boldsymbol{\mu}_{k(i)} + \mathbb{E}[\xi'_{\delta_{k(i)}}(\tilde{z}_{k(i)})]\boldsymbol{C}_{k(i)}\boldsymbol{\mu}_{\boldsymbol{w}}\right)^\intercal \boldsymbol{A} \left(\mathbb{E}[\xi_{\delta_{k(j)}}(\tilde{z}_{k(j)})]\boldsymbol{\mu}_{k(j)} + \mathbb{E}[\xi'_{\delta_{k(j)}}(\tilde{z}_{k(j)})]\boldsymbol{C}_{k(j)}\boldsymbol{\mu}_{\boldsymbol{w}}\right) \\
&\qquad + \mathbb{E}[\xi'_{\delta_{k(i)}}(\tilde{z}_{k(i)})]\mathbb{E}[\xi'_{\delta_{k(j)}}(\tilde{z}_{k(j)})] \operatorname{Tr}(\boldsymbol{C}_{k(i)}\boldsymbol{A}\boldsymbol{C}_{k(j)}\boldsymbol{C}_{\boldsymbol{w}}) + \mathcal{O}(n^{-\frac{1}{2}}).
\end{aligned}$$

Therefore, summing up these estimations for $1 \leq i, j \leq n$ and $i \neq j$, we obtain:

$$\frac{1}{n^2} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_{k(i)}}(\boldsymbol{x}_i^\intercal \boldsymbol{w}_{-i})\xi_{\delta_{k(j)}}(\boldsymbol{x}_j^\intercal \boldsymbol{w}_{-j})\boldsymbol{x}_i^\intercal \boldsymbol{A}\boldsymbol{x}_j] = \boldsymbol{\mu}_{\boldsymbol{w}}^\intercal \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{w}} + \frac{1}{n}\operatorname{Tr}(\boldsymbol{C}_{\boldsymbol{w}}\boldsymbol{K}\boldsymbol{A}\boldsymbol{K}) + \mathcal{O}(n^{-\frac{1}{2}}).$$

Moreover, if we note $\tilde{\boldsymbol{C}} \equiv \sum_{\ell=1}^k \gamma_\ell \mathbb{E}_{\mathcal{A}_w}[\xi_{\delta_\ell}(\tilde{z}_\ell)^2]\boldsymbol{C}_\ell$, equation (3) implies that,

$$\left\| \boldsymbol{C}_{\boldsymbol{w}} - \tilde{\boldsymbol{C}} - \boldsymbol{K}\boldsymbol{C}_{\boldsymbol{w}}\boldsymbol{K} \right\|_* \leq \mathcal{O}(n^{-\frac{1}{2}}). \tag{4}$$

Since $\|\boldsymbol{K}\| < 1 - \varepsilon'$ (see [LC20b] for details), for some constant $\varepsilon' > 0$, the matrix $\boldsymbol{R}_1 = (\boldsymbol{I}_p - \boldsymbol{K})^{-1}$ and the linear mapping $\boldsymbol{R}_2$ introduced in the main paper are well defined, and we have from (2) and (4), the estimations,

$$\|\boldsymbol{\mu}_{\boldsymbol{w}} - \boldsymbol{R}_1\tilde{\boldsymbol{\mu}}\| \leq \mathcal{O}\left(n^{-\frac{1}{2}}\right), \quad \left\| \boldsymbol{C}_{\boldsymbol{w}} - \boldsymbol{R}_2(\tilde{\boldsymbol{C}}) \right\|_* \leq \mathcal{O}\left(n^{-\frac{1}{2}}\right).$$

We can then naturally approximate the quantities $\delta_\ell = \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{C}_\ell\left(\boldsymbol{I}_p - \boldsymbol{K}\right)^{-1}\right)$, $m_\ell = \mathbb{E}[\tilde{z}_\ell]$ and $\sigma_\ell = \mathbb{E}[\tilde{z}_\ell^2] - m_\ell^2$ as the solutions to the fixed point equation presented in Theorem 4.8 of the main paper. Note however that the proof of the existence and uniqueness of the solutions is quite an elaborate problem, but the simulations strongly confirm this conjecture.

2 Extension to the Softmax classifier

2.1 Position of the Problem and first properties

The main encountered difficulty with the Softmax classifier setting is that the mapping f goes from $\mathbb{R}^k$ to $\mathbb{R}^k$ instead of being scalar, thus its derivative is now a differential which brings some purely formal complexities. For any $\ell \in [k]$, let us denote

$$\begin{array}{rccc}
f_\ell : \mathbb{R}^k & \longrightarrow & \mathbb{R}^k & \\
\boldsymbol{v} & \longmapsto & \left[\dfrac{\phi(v_a)/\lambda_a}{\sum_{b=1}^k \phi(v_b)} \displaystyle\sum_{b=1}^k \tilde{Y}_{\ell,b}\psi(v_b) - \tilde{Y}_{\ell,a}\psi(v_a)\right]_{1 \leq a \leq k}, &
\end{array} \tag{5}$$

where $\tilde{\boldsymbol{Y}} = \boldsymbol{I}_k$ contains in the ℓ^{th} column the label of the class ℓ. Then, the fixed point equation satisfied by the Softmax weights $\boldsymbol{W} = (\boldsymbol{w}_1^\intercal, \ldots, \boldsymbol{w}_k^\intercal)^\intercal \in \mathbb{R}^{pk}$ (all stacked in a vector) writes in the simpler form,

$$\boldsymbol{W} = \frac{1}{n}\sum_{i=1}^n \tilde{\boldsymbol{x}}_i f_{k(i)}(\tilde{\boldsymbol{x}}_i^\intercal \boldsymbol{W}),$$

where $\tilde{\boldsymbol{x}}_i \equiv \boldsymbol{I}_k \otimes \boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_i & & \\ & \ddots & \\ & & \boldsymbol{x}_i \end{pmatrix} \in \mathcal{M}_{pk,k}\ \forall i \in [n]$. Let us introduce a constant $\varepsilon > 0$ and define the event $\mathcal{A}_{\boldsymbol{W}}$ as,

$$\mathcal{A}_{\boldsymbol{W}} \equiv \left\{ \sup_{\substack{1 \leq \ell \leq k \\ \boldsymbol{z} \in \mathbb{R}^k}} \frac{\|df_\ell|_{\boldsymbol{z}}\|}{n} \|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\| \leq 1 - \varepsilon \right\},$$

where $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n) \in \mathcal{M}_{pk,nk}$. Then, replacing Assumptions 3 by its analog as follows,

Assumption 3 bis. $\sup_{\substack{1\le \ell\le k\\ \boldsymbol{z}\in\mathbb{R}^k}} \frac{\|df_\ell|_{\boldsymbol{z}}\|}{n}\mathbb{E}[\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\|] \le 1-2\varepsilon,$

we have the following analog to Lemma 4.2:

Lemma 2.1. *Under Assumption 1, 2 and 3 bis, there exists two constants $C, c > 0$ such that: $\mathbb{P}(\mathcal{A}_{\boldsymbol{W}}^c) \le Ce^{-cn}$.*

Proof. It is a consequence of the result of concentration of product of concentrated variables provided in [LC20a] and that sets that since $\|\tilde{\boldsymbol{X}}\|/\sqrt{n} \propto \mathcal{E}_2(n^{-\frac{1}{2}})$ and $\mathbb{E}[\|\tilde{\boldsymbol{X}}\|/\sqrt{n}] \le \mathcal{O}(1)$, then $\frac{1}{n}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\| = \frac{1}{n}\|\tilde{\boldsymbol{X}}\|^2 \propto \mathcal{E}_2(1/\sqrt{n})+\mathcal{E}_1(1/n)$. This signifies that there exist two constants $C, c > 0$ such that for any $p \in \mathbb{N}$:

$$\mathbb{P}\left(\left|\frac{1}{n}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\| - \mathbb{E}\left[\frac{1}{n}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\|\right]\right| \ge t\right) \le Ce^{-cnt^2} + Ce^{-cnt}.$$

Thus, in particular, Assumption 3 bis allows us to bound:

$$\mathbb{P}(\mathcal{A}_{\boldsymbol{W}}^c) \le \mathbb{P}\left(\left|\frac{1}{n}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\| - \mathbb{E}\left[\frac{1}{n}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\|\right]\right| \ge \varepsilon\right) \le Ce^{-cn\varepsilon^2} + Ce^{-cn\varepsilon},$$

which gives us the result of the Lemma, modifying slightly the constants $C, c > 0$. □

Hence, in the case of the multiclass softmax classification, we obtain also the concentration of $\boldsymbol{W}$.

Theorem 2.2. *Under Assumption 1, 2 and 3 bis: $(\boldsymbol{W} \mid \mathcal{A}_{\boldsymbol{W}}) \propto \mathcal{E}_2(n^{-\frac{1}{2}})$.*

2.2 Analog to Theorem 4.8

Basically, the fixed point equation of Theorem 4.8 remains the same in the (multi-class) Softmax setting, we rather just need to adapt the objects $\boldsymbol{Q}, \boldsymbol{\delta}, \xi_\ell, \tilde{\boldsymbol{C}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{K}, \boldsymbol{R}_1, \boldsymbol{R}_2$ to the vectorial mapping introduced in equation (5). Hence, we are looking at first to an analog of Theorem 4.4 to link $\boldsymbol{W}$ with the random matrix $\boldsymbol{W}_{-i}$ defined as the only solution to $\boldsymbol{W}_{-i} = \frac{1}{n}\sum_{\substack{1\le j\le n\\ j\neq i}}^n \tilde{\boldsymbol{x}}_j f_{k(j)}(\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{W}_{-i})$. For that an adaptation of Assumption 4 is needed.

Assumption 4 bis. $\forall \ell \in [k], \sup_{\boldsymbol{z}\in\mathbb{R}^k} \|df_\ell^2|_{\boldsymbol{z}}\| \le \infty$[6].

Then introducing the notations $\tilde{\boldsymbol{X}}_{-i} = (\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_{i-1}, \boldsymbol{0}, \tilde{\boldsymbol{x}}_{i+1}, \dots \tilde{\boldsymbol{x}}_n) \in \mathcal{M}_{pk,nk}$ and:

$$\boldsymbol{Q}_{-i} \equiv \left(\boldsymbol{I}_{kp} - \frac{1}{n}\tilde{\boldsymbol{X}}_{-i}\boldsymbol{D}^{(i)}\tilde{\boldsymbol{X}}_{-i}^\intercal\right)^{-1} \in \mathcal{M}_{kp}, \quad \boldsymbol{D}^{(i)} = \mathrm{Diag}(\boldsymbol{D}_j^{(i)}) \in \mathcal{M}_{kn}, \quad \boldsymbol{D}_j^{(i)} \equiv df_{k(j)}\big|_{\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{W}_{-i}} \in \mathcal{M}_k.$$

We have the following proposition which is at the core of the Softmax classifier analysis.

Proposition 2.3 (Analog to Theorem 4.4)**.** *Under Assumptions 1, 2, 3 bis and 4 bis, for any $i \in [n]$:*

$$\forall t > 0 \,:\, \mathbb{P}_{\mathcal{A}_X}\left(\left\|\tilde{\boldsymbol{x}}_i^\intercal \boldsymbol{W} - \tilde{\boldsymbol{x}}_i^\intercal \boldsymbol{W}_{-i} + \frac{1}{n}\tilde{\boldsymbol{x}}_i^\intercal \boldsymbol{Q}_{-i}\tilde{\boldsymbol{x}}_i f_{k(i)}(\tilde{\boldsymbol{x}}_i^\intercal \boldsymbol{W})\right\| \ge t\right) \le Ce^{-cnt^2}.$$

Consequently, we retrieve an analog to Theorem 4.8 for the Softmax classifier.

Theorem 2.4 (Analog to Theorem 4.8)**.** *Under Assumptions 1, 2, 3 bis and 4 bis, there exists three unique tuples $(\boldsymbol{m}_\ell)_{1\le\ell\le k} \in (\mathbb{R}^k)^{[k]}$, $(\boldsymbol{\sigma}_\ell)_{1\le\ell\le k} \in (\mathcal{M}_k)^{[k]}$ and $(\boldsymbol{\Delta}_\ell)_{1\le\ell\le k} \in (\mathcal{M}_k)^{[k]}$ satisfying the conditions ($\forall \ell \in [k]$):*

- $\tilde{\boldsymbol{z}}_\ell \sim \mathcal{N}(\boldsymbol{m}_\ell, \boldsymbol{\sigma}_\ell^2)$;
- $\xi_\ell : \mathbb{R}^k \to \mathbb{R}^k$ *satisfying for any $\boldsymbol{z} \in \mathbb{R}^k$:*
$$\xi_\ell(\boldsymbol{z}) = f_\ell(\boldsymbol{z} + \boldsymbol{\Delta}_\ell \xi_\ell(\boldsymbol{z}));$$
- $\boldsymbol{K} \equiv \sum_{a=1}^k \gamma_a \mathbb{E}[d\xi_a|_{\tilde{\boldsymbol{z}}_a}] \otimes \boldsymbol{C}_a \in \mathcal{M}_{kp}$;
- $\bar{\boldsymbol{Q}} \equiv (\boldsymbol{I}_{kp} - \boldsymbol{K})^{-1} \in \mathcal{M}_{pk}$;
- $\boldsymbol{\Delta}_\ell = \left[\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{C}_\ell \bar{\boldsymbol{Q}}_{\mathcal{I}_a,\mathcal{I}_b}\right)\right]_{1\le a,b\le k} \in \mathcal{M}_k$;
- $\boldsymbol{R} : \mathcal{M}_{pk} \to \mathcal{M}_{pk}$ *defined, for $\boldsymbol{M} \in \mathcal{M}_{pk}$, as*
$$\boldsymbol{R}(\boldsymbol{M}) = \boldsymbol{M} + \boldsymbol{K}(\boldsymbol{R}(\boldsymbol{M}))\boldsymbol{K};$$
- $\tilde{\boldsymbol{\mu}} = \sum_{a=1}^k \gamma_a \mathbb{E}[\xi_a(\tilde{\boldsymbol{z}}_a)] \otimes \boldsymbol{\mu}_a \in \mathcal{M}_{kp}$;
- $\tilde{\boldsymbol{C}} = \sum_{a=1}^k \gamma_a \mathbb{E}[\xi_a(\tilde{\boldsymbol{z}}_a)\xi_a(\tilde{\boldsymbol{z}}_a)^\intercal] \otimes \boldsymbol{C}_a \in \mathcal{M}_{kp}$;
- $\boldsymbol{m}_\ell = \boldsymbol{\mu}_\ell^\intercal \bar{\boldsymbol{Q}} \tilde{\boldsymbol{\mu}}$;
- $\boldsymbol{\sigma}_\ell^2 = \left[\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{C}_\ell \boldsymbol{R}(\tilde{\boldsymbol{C}})_{\mathcal{I}_a,\mathcal{I}_b}\right) + \tilde{\boldsymbol{\mu}}\bar{\boldsymbol{Q}}_{:,\mathcal{I}_a}\boldsymbol{C}_\ell \bar{\boldsymbol{Q}}_{\mathcal{I}_b,:}\tilde{\boldsymbol{\mu}}\right]_{1\le a,b\le k}$.

[6] Here, given $\boldsymbol{z} \in \mathbb{R}^k$, $df_\ell^2|_{\boldsymbol{z}}$ is a bilinear form on $\mathbb{R}^k$, thus $\|df_\ell^2|_{\boldsymbol{z}}\| = \sup_{\|\boldsymbol{a}\|,\|\boldsymbol{b}\|\le 1} df_\ell^2|_{\boldsymbol{z}} \cdot (\boldsymbol{a}, \boldsymbol{b})$.

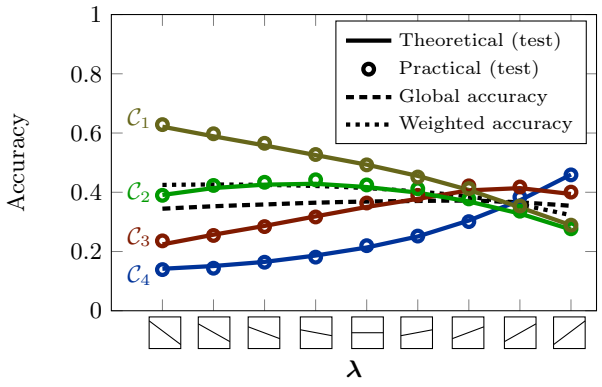

Figure 1: Theoretical and practical test accuracies on synthetic Gaussian data varying the regularizing vector $\boldsymbol{\lambda}$. The considered parameters are $n = p = 200$, $k = 4$, $\gamma_1 = 2/5$, $\gamma_2 = 3/10$, $\gamma_3 = 1/5$ and $\gamma_4 = 1/10$. The statistics $\boldsymbol{\mu}_\ell$ and $\boldsymbol{C}_\ell$ for $\ell \in [k]$ are sampled randomly as $\boldsymbol{\mu}_\ell \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$ and $\boldsymbol{C}_\ell \sim \mathrm{Diag}(\boldsymbol{u}) + \sqrt{\boldsymbol{v}\boldsymbol{v}^\intercal}$ with $\boldsymbol{u} \sim \mathrm{Unif}([0,1]^p)$ and $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$.

Where $\forall a \in [k]$, $\mathcal{I}_a \equiv \{ap+1, \ldots, (a+1)p\}$ *and we define for any matrix* $\boldsymbol{T} \in \mathcal{M}_{pk}$ *and any* $a, b \in [k]$*:*

$$\boldsymbol{T}_{\mathcal{I}_a, \mathcal{I}_b} \equiv [T_{i,j}]_{i \in \mathcal{I}_a, j \in \mathcal{I}_b}, \quad \boldsymbol{T}_{\mathcal{I}_a, :} \equiv [T_{i,j}]_{i \in \mathcal{I}_a, j \in [pk]}, \quad \boldsymbol{T}_{:, \mathcal{I}_b} \equiv [T_{i,j}]_{i \in [pk], j \in \mathcal{I}_b}.$$

Then one has the following estimations,

$$\|\boldsymbol{\mu}_W - \bar{\boldsymbol{Q}}\tilde{\boldsymbol{\mu}}\| \leq \mathcal{O}(n^{-\frac{1}{2}}), \quad \|\boldsymbol{C}_W - \boldsymbol{R}\tilde{\boldsymbol{C}}\|_* \leq \mathcal{O}(n^{-\frac{1}{2}}),$$

and for a new datum $\boldsymbol{x}$ *in class* $\mathcal{C}_\ell$*, independent of the training data* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$*, one has*

$$\|\mathbb{E}_{\mathcal{A}_{\boldsymbol{W}}}[\tilde{\boldsymbol{x}}^\intercal \boldsymbol{W}] - \boldsymbol{m}_\ell\| \leq \mathcal{O}(n^{-\frac{1}{2}}), \quad \|\mathbb{E}_{\mathcal{A}_{\boldsymbol{W}}}[\tilde{\boldsymbol{x}}^\intercal \boldsymbol{W}\boldsymbol{W}^\intercal \boldsymbol{x}] - \boldsymbol{\sigma}_\ell^2 - \boldsymbol{m}_\ell \boldsymbol{m}_\ell^\intercal\| \leq \mathcal{O}(n^{-\frac{1}{2}}).$$

3 Further experiments

3.1 Synthetic Gaussian data

This subsection provides experiments using synthetic Gaussian data. Recalling the setting of Figure 1, which depicts the theoretical versus the practical accuracies (of a four-class classification problem with Softmax) for different choices of $\boldsymbol{\lambda}$. One can first observe that the practical accuracies are perfectly estimated by their theoretical counterparts as per Theorem 2.4. We have also depicted in this figure the global accuracy of the classifier along with its weighted accuracy (computed as $\sum_\ell \gamma_\ell a_\ell$ with a_ℓ being the accuracy of class $\mathcal{C}_\ell$). These quantities notably highlight the existence of an optimal choice of $\boldsymbol{\lambda}$ for maximising the classifier accuracy (for the less representative classes), which in turn might be anticipated through our present investigation.

3.2 GAN-generated MNIST images

We provide in this subsection experiments with a multi-class classification of data issued from a GAN[7]-generated version of the MNIST digits "1", "2", and "3" with the following settings:

1. We samples 50000 images of each class;
2. We multiplied each of the data by a randomly chosen matrix $\boldsymbol{A} \in \mathcal{M}_{p, p_{\mathrm{MNIST}}}$, where $p = 200$ and $p_{\mathrm{MNIST}} = 784$;
3. For each class $\mathcal{C}_\ell$, $\ell \in [3]$, we divided each set of indexes $\mathcal{I}_\ell = [50000]$ into three subset: $\mathcal{I}_\ell^{\mathrm{tr}}$, $\mathcal{I}_\ell^{\mathrm{tst}}$ and $\mathcal{I}_\ell^{\mathrm{pop}}$ such that $\#\mathcal{I}_\ell^{\mathrm{tr}} = \#\mathcal{I}_\ell^{\mathrm{tst}} = 5000$ and $\#\mathcal{I}_\ell^{\mathrm{pop}} = 40000$;
4. For all $\ell \in [3]$, we computed the empirical mean $\boldsymbol{\mu}_\ell$ and covariance $\boldsymbol{C}_\ell$ with the data indexed by I_ℓ^{pop};

[7] We used a standard DC-GAN model [MO14] trained on the whole MNIST dataset.

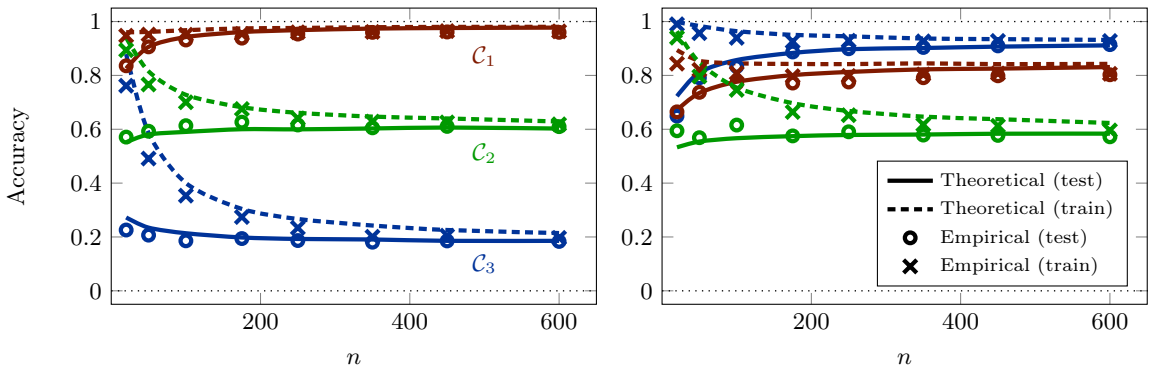

Figure 2: Classification accuracies of GAN-generated MNIST digits with unbalanced classes in the training set: $\gamma_1 = 1/2$, $\gamma_2 = 1/3$ and $\gamma_3 = 1/6$ for two choices of $\boldsymbol{\lambda}$: $[30, 30, 30]$ **(left)** $[10, 20, 30]$ **(right)**.

5. In order to make the classification problem harder, for all $\ell \in [3]$ we retrieved from each data indexed by I_ℓ^{tr} and I_ℓ^{tst} the vector $(1-\alpha)\boldsymbol{\mu}_\ell$, with $\alpha = 0.5$ and we corrected the mean ($\boldsymbol{\mu}_\ell \leftarrow \alpha\boldsymbol{\mu}_\ell$).

Figure 2 pictures with two different regularizing vector $\boldsymbol{\lambda} \in \mathbb{R}^3$:

- The theoretical performance on the training set and on the test set based on the estimated class-wise means and covariances and relying on Theorem 2.4;
- The empirical performances computed on the data indexed by the sets $\mathcal{I}_\ell^{\text{tr}}$ and $\mathcal{I}_\ell^{\text{tst}}$, for $\ell \in [k]$.

As we can notice from Figure 2, the theoretical estimations closely match the empirical ones in both settings. Moreover, it is deductible that when the classes of the training set are unbalanced, the regularizing parameter vector $\boldsymbol{\lambda}$ should be fine-tuned accordingly. Besides, note that with an optimal choice of $\boldsymbol{\lambda}$, the accuracy of classification of the less represented class is better than the accuracy of the two other classes. We therefore believe that our present investigation could be helpful for the automatic fine-tuning of $\boldsymbol{\lambda}$.

3.3 CNN features of real Imagenet images

In this subsection, we provide further experiments using real images from the Imagenet dataset [DDS+09]. Figure 3-(left) depicts the learned Softmax weights against their expected large p, n asymptotics as predicted by Theorem 2.4. As for GAN generated images, we observe a perfect match between the learned weights and the theoretical predictions. An almost perfect match is also observed for the scores (between the practical scores and their theoretical counterparts) as depicted in Figure 3-(right) which strongly suggests that the conclusions of Theorem 2.4 generalize to real data.

References

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Kla07] B. Klartag. A central limit theorem for convex sets. *Inventiones mathematicae*, 168:91–131, 2007.

[LC20a] Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. *submitted to Random Matrices: Theory and Applications*, 2020.

[LC20b] Cosme Louart and Romain Couillet. Concentration of solutions to random equations with concentration of measure hypotheses, 2020.

[Led05] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

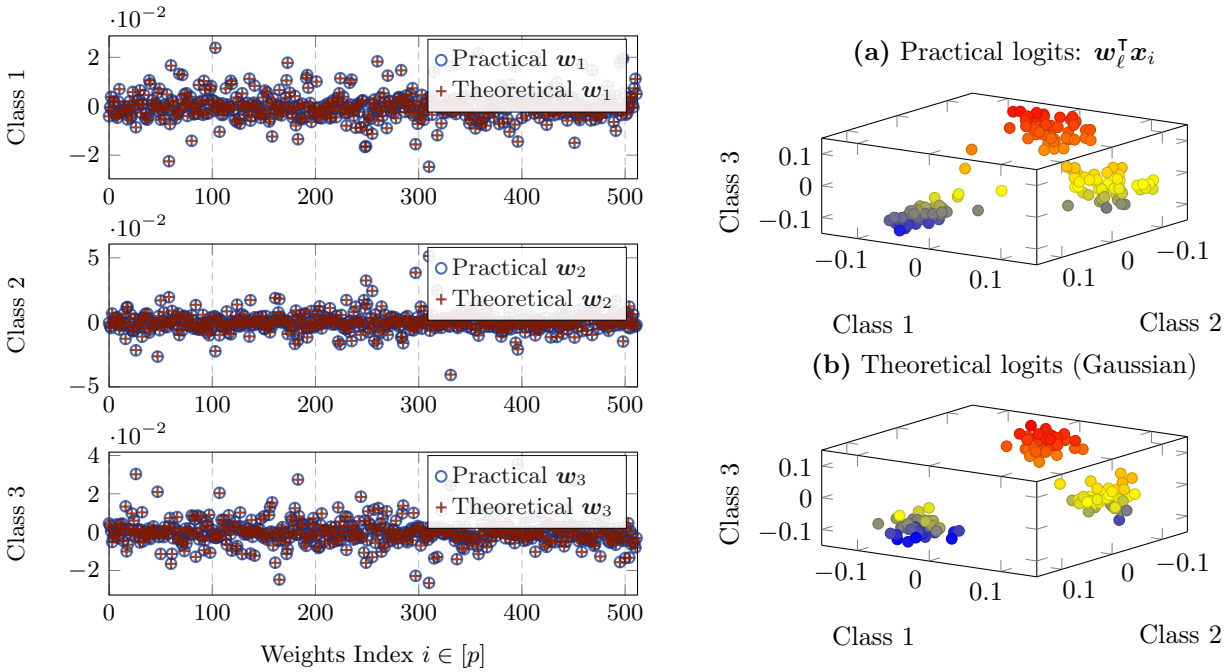

Figure 3: **(Left)** Learned weights (blue circles) versus theoretical estimates (red crosses) from Theorem 2.4. **(Right)** Practical **(a)** versus theory-predicted **(b)** logits, on a test set independent from the training set. The data are Resnet18 [SIVA17] representations ($p = 512$) of randomly selected images from the Imagenet dataset [DDS+09]; $k = 3$ classes: *hamburger*, *mushroom*, *pizza*; $n = 3000$; regularization constants $\lambda_1 = \lambda_2 = \lambda_3 = 1.5$; data normalized such that $\|\boldsymbol{x}_i\| = 0.1 \cdot \sqrt{p}$ to ensure $\mathcal{A}_{\boldsymbol{w}}$.

[MO14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[SIVA17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.