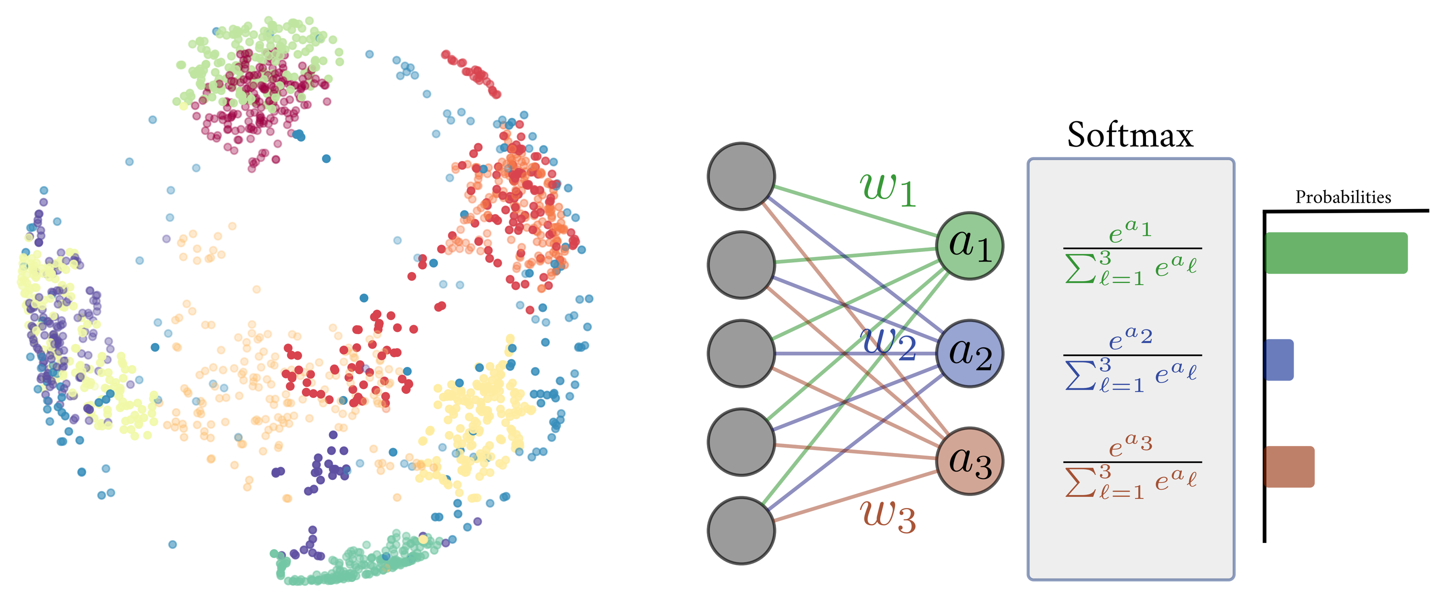
The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers

Mohamed El Amine Seddik [1,2] Cosme Louart [1,3] Mohamed Tamaazousti [1] Romain Couillet [2,3]

[1]CEA List [2]CentraleSupélec [3]GIPSA-Lab University of GrenobleAlpes

Abstract



$x_1, \ldots, x_n \in \mathbb{R}^p$

- RMT analysis of Softmax with high-dimensional **concentrated** inputs.
- Softmax weights depend only on data **means** and **covariances**.
- Asymptotic **performance** of Softmax derived based on first data moments.

Notion of Concentrated Vectors

Definition 1. $\mathcal{X} \ni \boldsymbol{x}$ is q-exponentially **concentrated** if for all $\varphi : \mathcal{X} \to \mathbb{R}$ 1-**Lipschitz**, there exists $C \geq 0$ independent of $\dim(\mathcal{X})$ and $\sigma > 0$ such that,

$$\forall t \geq 0, \quad \mathbb{P}\left(|\varphi(\boldsymbol{x}) - \mathbb{E}\varphi(\boldsymbol{x})| > t\right) \leq C\, e^{-(t/\sigma)^q},$$

denoted $\boldsymbol{x} \propto \mathcal{E}_q(\sigma)$ or $\boldsymbol{x} \propto \mathcal{E}_q$ if σ independent of $\dim(\mathcal{X})$.

Examples:

- $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z} \in \mathbb{R}^p$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and $\|\boldsymbol{\Sigma}\| < \infty$, then $\boldsymbol{x} \propto \mathcal{E}_2$.
- If $\mathcal{Z} \ni \boldsymbol{z} \propto \mathcal{E}_q$ and $\mathcal{G} : \mathcal{Z} \to \mathcal{X}$ L-Lipschitz, then $\mathcal{G}(\boldsymbol{z}) \propto \mathcal{E}_q(L)$.

Figure 1. Images generated by the BigGAN model [3].

Model

(M) **Data matrix** (distributed in k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$):

$$\mathbb{R}^{p\times n} \ni \boldsymbol{X} = \Big[\underbrace{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}}_{\sim\mathcal{L}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}, \underbrace{\boldsymbol{x}_{n_1+1}, \ldots, \boldsymbol{x}_{n_2}}_{\sim\mathcal{L}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}, \ldots, \underbrace{\boldsymbol{x}_{n-n_k+1}, \ldots, \boldsymbol{x}_n}_{\sim\mathcal{L}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}\Big] \propto \mathcal{E}_2$$

Model statistics: $\boldsymbol{\mu}_\ell = \mathbb{E}_{\boldsymbol{x}_i\in\mathcal{C}_\ell}[\boldsymbol{x}_i], \quad \boldsymbol{\Sigma}_\ell = \mathbb{E}_{\boldsymbol{x}_i\in\mathcal{C}_\ell}[\boldsymbol{x}_i\boldsymbol{x}_i^\intercal] - \boldsymbol{\mu}_\ell\boldsymbol{\mu}_\ell^\intercal.$

Softmax Classifier

Minimize:

$$\mathcal{L}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{k} y_{i\ell}\log p_{i\ell} + \frac{1}{2}\sum_{\ell=1}^{k}\lambda_\ell\|\boldsymbol{w}_\ell\|^2$$

$$p_{i\ell} = \frac{\exp(\boldsymbol{w}_\ell^\intercal \boldsymbol{x}_i)}{\sum_{j=1}^{k}\exp(\boldsymbol{w}_j^\intercal \boldsymbol{x}_i)}, \quad \boldsymbol{W} \equiv [\boldsymbol{w}_1^\intercal, \ldots, \boldsymbol{w}_k^\intercal]^\intercal \in \mathbb{R}^{pk}$$

Implicit Equation (scalar case for some $f : \mathbb{R} \to \mathbb{R}$)

$$\mathbb{R}^p \ni \boldsymbol{w} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i \quad \Rightarrow \quad \boxed{\boldsymbol{w} = \Psi(\boldsymbol{w}) \equiv \frac{1}{n}\boldsymbol{X} f(\boldsymbol{X}^\intercal \boldsymbol{w})}$$

Ψ is requested to be $(1-\varepsilon)$-Lipschitz for some $\varepsilon > 0$ or equivalently

$$\mathcal{A}_{\boldsymbol{w}} = \left\{\frac{1}{n}\|f'\|_\infty \|\boldsymbol{X}\boldsymbol{X}^\intercal\| \geq 1 - \varepsilon\right\} \quad \text{has low probability.}$$

Assumptions

(A) **Growth rate assumptions:** As $p \to \infty$,

1. $p/n \to c \in (0, \infty)$ and $|\mathcal{C}_\ell|/n \to \gamma_\ell \in (0, 1)$.
2. k fixed.
3. $\|\boldsymbol{\mu}_\ell\| = \mathcal{O}(1)$ for each $\ell \in [k]$.
4. $\exists\varepsilon > 0$ independent of p, n s.t. $\frac{1}{n}\|f'\|_\infty\|\boldsymbol{X}\boldsymbol{X}^\intercal\| \leq 1 - 2\varepsilon$.

Main Result

Evaluate $\boldsymbol{\mu}_w = \mathbb{E}[\boldsymbol{w}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(\boldsymbol{x}_i^\intercal \boldsymbol{w})\boldsymbol{x}_i]$ and $\boldsymbol{\Sigma}_w = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\intercal] - \boldsymbol{\mu}_w\boldsymbol{\mu}_w^\intercal$
Under (M-A), $\mathbb{P}(\mathcal{A}_{\boldsymbol{w}}) \propto e^{-n}$ and $\boldsymbol{w} \propto \mathcal{E}_2\left(n^{-\frac{1}{2}}\right) \mid \mathcal{A}_{\boldsymbol{w}}$, and there exists $(\boldsymbol{\delta}, \boldsymbol{m}, \sigma) \in (\mathbb{R}^k)^3$ satisfying

$$z_\ell \sim \mathcal{N}(m_\ell, \sigma_\ell^2); \quad \delta_\ell = \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}_\ell\,(\boldsymbol{I}_p - \boldsymbol{K})^{-1}\right);$$

$$\tilde{\boldsymbol{\mu}} \equiv \sum_{\ell=1}^{k}\gamma_\ell\mathbb{E}[g_\ell(z_\ell)]\boldsymbol{\mu}_\ell; \quad \tilde{\boldsymbol{\Sigma}} \equiv \sum_{\ell=1}^{k}\gamma_\ell\mathbb{E}[g_\ell(z_\ell)^2]\boldsymbol{\Sigma}_\ell; \quad \boldsymbol{K} \equiv \sum_{\ell=1}^{k}\gamma_\ell\mathbb{E}[g_\ell'(z_\ell)]\boldsymbol{\Sigma}_\ell;$$

$$\boldsymbol{R}_1 \equiv (\boldsymbol{I}_p - \boldsymbol{K})^{-1}; \quad \boldsymbol{R}_2(\boldsymbol{M}) = \boldsymbol{M} + \boldsymbol{K}\boldsymbol{R}_2(\boldsymbol{M})\boldsymbol{K};$$

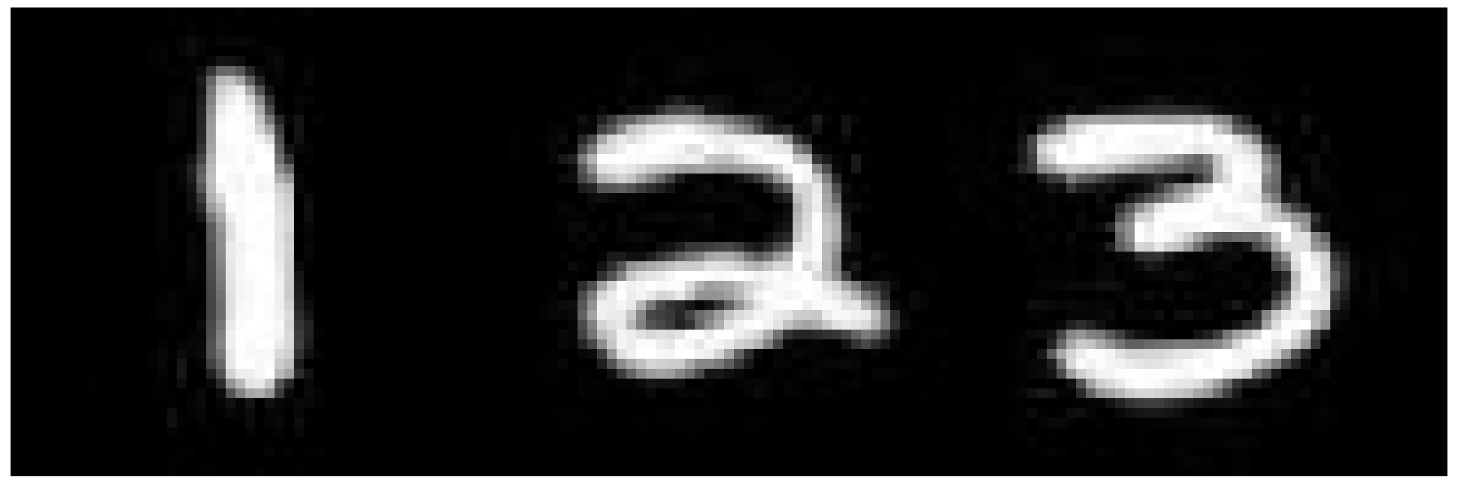
$$m_\ell \equiv \boldsymbol{\mu}_\ell^\intercal\boldsymbol{R}_1\tilde{\boldsymbol{\mu}}; \quad \sigma_\ell^2 \equiv \frac{1}{n}\operatorname{Tr}(\boldsymbol{\Sigma}_\ell\boldsymbol{R}_2(\tilde{\boldsymbol{\Sigma}})) + \tilde{\boldsymbol{\mu}}^\intercal\boldsymbol{R}_1\boldsymbol{\Sigma}_\ell\boldsymbol{R}_1\tilde{\boldsymbol{\mu}};$$

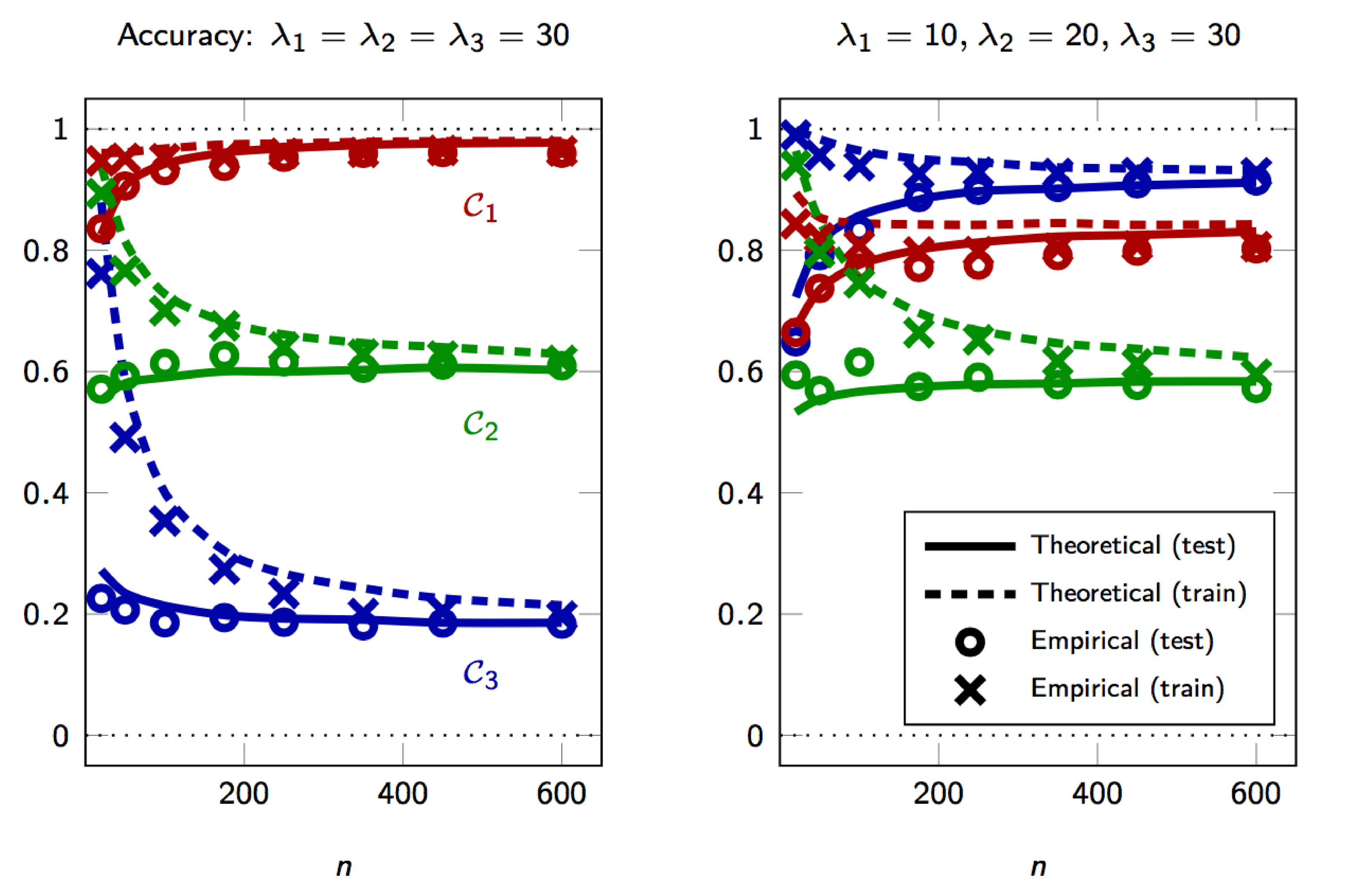
Furthermore,

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

Key Observation: Only first and second order statistics matter!

Simulations





Conclusion

- (El-Karoui+'13, Mai+'19) analyzed logistic regression under **Gaussian** data.
- We generalized these ideas to a k-class mixture of **concentrated** data.
- **Universality:** "*Softmax treats input data as Gaussian random vectors*".
- **Optimality:** Softmax is optimal for data with *strongly discriminative class-wise means* as suggested by distance-based image classification approaches (**Mensink+'13**).

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets", in NIPS 2014.

[2] Terence Tao, "Topics in random matrix theory, volume 132". American Mathematical Society Providence, RI, 2012.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan, "Large scale GAN training for high fidelity image synthesis", in ICLR 2019.

AISTATS 2021 - **Contact:** melaseddik@gmail.com