**Theorem 1.** *Suppose one of the following conditions is satisfied:*

1. *The model has a Gaussian likelihood and* $\chi(\boldsymbol{y}) = O(\exp(\|\boldsymbol{y}\|_1))$
2. *The model has a Laplace likelihood and* $\chi(\boldsymbol{y}) = O(\exp(\sum_{i=t_0}^{T} \sqrt{|y_i|}))$
3. *The model has a logistic likelihood and* $\chi(\boldsymbol{y}) = O(\exp(\sum_{i=t_0}^{T} \sqrt{|y_i|}))$

*Then the interchange of integration and differentiation for the score-function estimator is valid. In particular, all polynomially bounded statistics satisfy these conditions.*

Following Theorem 2.4.3 in (Casella & Berger, 2002), let us denote

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \chi(\boldsymbol{y}) q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z].$$

Note that we call the perturbation $\boldsymbol{\theta}$ to have consistent notations, and that we consider $\boldsymbol{x}$ to be fixed as it does not change during the attack. $f$ is differentiable and we have

$$\frac{\partial f}{\partial \boldsymbol{\theta}} = \chi(\boldsymbol{y}) \frac{\partial q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z]}{\partial \boldsymbol{\theta}}. \tag{1}$$

In order to interchange integration and differentiation, Theorem 2.4.3 requires to dominate the rate of change

$$\left| \frac{f(\boldsymbol{y}, \boldsymbol{\theta}_0 + \boldsymbol{\delta}) - f(\boldsymbol{y}, \boldsymbol{\theta}_0)}{\boldsymbol{\delta}} \right|,$$

for $\|\boldsymbol{\delta}\|_1 \leq \boldsymbol{\delta}_0$, by an integrable function. In practice, the mean-value theorem yields

$$\left| \frac{f(\boldsymbol{y}, \boldsymbol{\theta}_0 + \boldsymbol{\delta}) - f(\boldsymbol{y}, \boldsymbol{\theta}_0)}{\boldsymbol{\delta}} \right| \leq \sup_{\boldsymbol{\epsilon} \in [\boldsymbol{0}, \boldsymbol{\delta}]} \left\| \left. \frac{\partial f}{\partial \boldsymbol{\theta}} \right|_{(\boldsymbol{y}, \boldsymbol{\theta}_0 + \boldsymbol{\epsilon})} \right\|_1,$$

and allows to instead bound the quantity

$$\sup_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\|_1 \leq \boldsymbol{\delta}_0} \left\| \left. \frac{\partial f}{\partial \boldsymbol{\theta}} \right|_{(\boldsymbol{y}, \boldsymbol{\theta}_0 + \boldsymbol{\delta})} \right\|_1.$$

Equation (1) allows to express the partial derivative of $f$ as a function of $\chi$ and $q$. Hence, we need to bound

$$\sup_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\|_1 \leq \boldsymbol{\delta}_0} \left\| \chi(\boldsymbol{y}) \left. \frac{\partial q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z]}{\partial \boldsymbol{\theta}} \right|_{(\boldsymbol{y}, \boldsymbol{\theta}_0 + \boldsymbol{\delta})} \right\|_1.$$

We define $\mu_i$ to be the mean predicted by the neural network for timestep $i$. Similarly, we define $\sigma_i$ as the standard deviation predicted by the network. As $i$ goes from $t_0$ to $T$, the chain rule yields

$$\frac{\partial q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z]}{\partial \boldsymbol{\theta}} = \sum_{i=t_0}^{T} \frac{\partial q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z]}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \boldsymbol{\theta}} + \frac{\partial q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z]}{\partial \sigma_i} \cdot \frac{\partial \sigma_i}{\partial \boldsymbol{\theta}}. \tag{2}$$

Since $\mu_i$ and $\sigma_i$ are learned by a neural network, their partial derivatives $\frac{\partial \mu_i}{\partial \boldsymbol{\theta}}$ and $\frac{\partial \sigma_i}{\partial \boldsymbol{\theta}}$ can be bounded by the global Lipschitz constant $L$ of the network (it is not necessary to find the exact constant, an upper bound such as the one obtained in (Szegedy et al., 2013) is sufficient). Besides, let us denote $\psi_i(y_i, \mu_i, \sigma_i)$ the likelihood at timestep $i$. By definition, we have

$$q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z] = \prod_{j=t_0}^{T} \psi_j.$$

Since only $\psi_i$ depends on $\mu_i$ and $\sigma_i$, we get

$$\frac{\partial q[\boldsymbol{y}|\boldsymbol{x} + \boldsymbol{\theta}, z]}{\partial \mu_i} = \frac{\partial \psi_i}{\partial \mu_i} \prod_{j \neq i} \psi_j,$$

and similarly

$$\frac{\partial q[\boldsymbol{y}|\boldsymbol{x}+\boldsymbol{\theta},z]}{\partial \sigma_i} = \frac{\partial \psi_i}{\partial \sigma_i} \prod_{j \neq i} \psi_j.$$

Applied to equation (2), this yields

$$\left\| \frac{\partial q[\boldsymbol{y}|\boldsymbol{x}+\boldsymbol{\theta},z]}{\partial \boldsymbol{\theta}} \right\|_1 \leq L \sum_{i=t_0}^{T} \left( \left\| \frac{\partial \psi_i}{\partial \mu_i} \prod_{j \neq i} \psi_j \right\|_1 + \left\| \frac{\partial \psi_i}{\partial \sigma_i} \prod_{j \neq i} \psi_j \right\|_1 \right). \tag{3}$$

Combined with equation (1), we obtain

$$\left\| \frac{\partial f}{\partial \boldsymbol{\theta}} \right\|_1 \leq |\chi(\boldsymbol{y})| L \sum_{i=t_0}^{T} \left( \left\| \frac{\partial \psi_i}{\partial \mu_i} \prod_{j \neq i} \psi_j \right\|_1 + \left\| \frac{\partial \psi_i}{\partial \sigma_i} \prod_{j \neq i} \psi_j \right\|_1 \right). \tag{4}$$

Here, we consider three cases for $\psi$: Gaussian, Laplace or logistic distribution.

**Case 1** (Gaussian distribution)**.**

In the case of a Gaussian likelihood, we have

$$\psi_i(y_i, \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[ -\frac{1}{2} \left( \frac{y_i - \mu_i}{\sigma_i} \right)^2 \right],$$

After computations, we obtain

$$\frac{\partial \psi_i}{\partial \mu_i} = \frac{y_i - \mu_i}{\sigma_i} \cdot \psi_i = O\left(\exp\left(-|y_i|^{1.5}\right)\right)$$

and

$$\frac{\partial \psi_i}{\partial \sigma_i} = \left( \frac{(y_i - \mu_i)^2}{\sigma_i^3} - \frac{1}{\sigma_i} \right) \cdot \psi = O\left(\exp\left(-|y_i|^{1.5}\right)\right)$$

Besides, we have that

$$\prod_{j \neq i} \psi_j = O\left( \exp\left( -\sum_{j \neq i} |y_j|^{1.5} \right) \right).$$

Hence,

$$\left\| \frac{\partial \psi}{\partial \mu_i} \prod_{j \neq i} \psi_j \right\|_1 + \left\| \frac{\partial \psi}{\partial \sigma_i} \prod_{j \neq i} \psi_j \right\|_1 = O\left( \exp\left( -\sum_{i=t_0}^{T} |y_i|^{1.5} \right) \right).$$

Together with equation (4), this gives the following inequality

$$\left\| \frac{\partial f}{\partial \boldsymbol{\theta}} \right\|_1 \leq |\chi(\boldsymbol{y})| \cdot L \cdot \sum_{i=t_0}^{T} \left( \left\| \frac{\partial \psi}{\partial \mu_i} \prod_{j \neq i} \psi_j \right\|_1 + \left\| \frac{\partial \psi}{\partial \sigma_i} \prod_{j \neq i} \psi_j \right\|_1 \right) = |\chi(\boldsymbol{y})| \cdot L \cdot O\left( \exp\left( -\sum_{i=t_0}^{T} |y_i|^{1.5} \right) \right).$$

Using the assumption that $\chi(\boldsymbol{y}) = O(\exp(||\boldsymbol{y}||_1))$,

$$\left\| \frac{\partial f}{\partial \boldsymbol{\theta}} \right\|_1 = O(\exp(||\boldsymbol{y}||_1)) \cdot O\left( \exp\left( -\sum_{i=t_0}^{T} |y_i|^{1.5} \right) \right) = O\left( \exp\left( -\sum_{i=t_0}^{T} |y_i|(\sqrt{|y_i|} - 1) \right) \right) = O(\exp(-||\boldsymbol{y}||_1))$$

All the asymptotic majorations are valid in the vicinity of $\boldsymbol{\theta}_0$, therefore we can take the $\sup$ on $\boldsymbol{\delta}$

$$\sup_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\|_1 \leq \boldsymbol{\delta}_0} \left\| \left. \frac{\partial f}{\partial \boldsymbol{\theta}} \right|_{(\boldsymbol{y}, \boldsymbol{\theta}_0 + \boldsymbol{\delta})} \right\|_1 = O(\exp(-||\boldsymbol{y}||_1))$$

The right hand term is positive and integrable with respect to $\boldsymbol{y}$. This satisfies the domination condition of the theorem, and thus concludes the proof.

**Case 2** (Laplace distribution)**.**

In the case of a Laplace distribution, we have

$$\psi_i(y_i, \mu_i, \sigma_i) = \frac{1}{2\sigma_i} \exp\left(-\left|\frac{y_i - \mu_i}{\sigma_i}\right|\right),$$

After computations, we obtain asymptotic majorations for the partial derivatives of $\psi_i$

$$\frac{\partial \psi_i}{\partial \mu_i} = -\frac{\operatorname{sign}(y_i - \mu_i)}{\sigma_i} \cdot \psi_i = O\left(\exp\left(-|y_i|^{0.75}\right)\right)$$

and

$$\frac{\partial \psi_i}{\partial \sigma_i} = \frac{|y_i - \mu_i| - 1}{\sigma_i} \cdot \psi_i = O\left(\exp\left(-|y_i|^{0.75}\right)\right)$$

Besides,

$$\prod_{j \neq i} \psi_j = O\left(\exp\left(-\sum_{j \neq i} |y_j|^{0.75}\right)\right).$$

Using equation (4) (with a similar reasoning as for the Gaussian distribution), it follows that

$$\left\|\frac{\partial f}{\partial \boldsymbol{\theta}}\right\|_1 \leq |\chi(\boldsymbol{y})| \cdot L \cdot O\left(\exp\left(-\sum_{i=t_0}^{T} |y_i|^{0.75}\right)\right).$$

Again, using the assumption that $\chi(\boldsymbol{y}) = O\left(\exp\left(\sum_{i=t_0}^{T} \sqrt{|y_i|}\right)\right)$,

$$\begin{aligned}\left\|\frac{\partial f}{\partial \boldsymbol{\theta}}\right\|_1 &= O\left(\exp\left(\sum_{i=t_0}^{T} \sqrt{|y_i|}\right)\right) \cdot O\left(\exp\left(-\sum_{i=t_0}^{T} |y_i|^{0.75}\right)\right) \\ &= O\left(\exp\left(-\sum_{i=t_0}^{T} \sqrt{|y_i|}(|y_i|^{0.25} - 1)\right)\right) = O\left(\exp\left(-\sum_{i=t_0}^{T} \sqrt{|y_i|}\right)\right).\end{aligned}$$

The majoration being valid around $\boldsymbol{\theta}_0$, we also take the sup on $\boldsymbol{\delta}$

$$\sup_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\|_1 \leq \boldsymbol{\delta}_0} \left\| \left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{(\boldsymbol{y}, \boldsymbol{\theta}_0 + \boldsymbol{\delta})} \right\|_1 = O\left(\exp\left(-\sum_{i=t_0}^{T} \sqrt{|y_i|}\right)\right).$$

The right-hand term is integrable and satisfies the domination condition of the theorem.

**Case 3** (Logistic distribution)**.**

Finally, in the case of a logistic likelihood, we have

$$\psi_i(y_i, \mu_i, \sigma_i) = \frac{\exp\left(-\frac{y_i - \mu_i}{\sigma_i}\right)}{\sigma_i \left(1 + \exp\left(-\frac{y_i - \mu_i}{\sigma_i}\right)\right)^2}.$$

Computations realized with a formal calculator yield

$$\frac{\partial \psi_i}{\partial \mu_i} = O\left(\exp\left(-|y_i|^{0.75}\right)\right)$$

and

$$\frac{\partial \psi_i}{\partial \sigma_i} = O\left(\exp\left(-|y_i|^{0.75}\right)\right)$$

We also have

$$\prod_{j \neq i} \psi_j = O\left(\exp\left(-\sum_{j \neq i} |y_j|^{0.75}\right)\right).$$

The rest of the proof is exactly similar to the case of a Laplace distribution.

## References


Casella, G. and Berger, R. L. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.