Modelling and Quantifying Membership Information Leakage in Machine Learning

Theorem 3. *Assume that*

(A1) Θ *is compact;*
(A2) $\lambda g(\theta) + \mathbb{E}_{\mathcal{P}}\{\ell(\mathfrak{M}(x;\theta), y)\}$ *is continuous and finite everywhere;*
(A3) $\ell(\mathfrak{M}(x;\theta), y)$ *is almost surely Lipschitz continuous with Lipschitz constant* L *on* Θ*;*
(A4) $g(\theta)$ *is strictly convex and* $\mathbb{E}\{\ell(\mathfrak{M}(x;\theta), y)\}$ *is convex.*

Then,

$$\lim_{\lambda\to\infty} \rho_{\mathrm{MI}}(\theta^*) = 0. \tag{1}$$

Consider a family of fitness functions $\ell(\mathfrak{M}(x;\theta), y)$ *parameterized by the the Lipschitz constant* $L \in [0, c)$ *for some* $c > 0$*, then*

$$\lim_{L\to 0} \rho_{\mathrm{MI}}(\theta^*) = 0. \tag{2}$$

Proof. The Danskin's theorem (see (Bertsekas, 1971, Proposition A.22) for a more general statement and proof) implies that θ^* is a continuous function of λ. Thus, $\lim_{\lambda\to\infty} \theta^* = \bar{\theta} := \arg\min_{\theta\in\Theta} g(\theta)$ and, as a result, $\lim_{\lambda\to\infty} I(\theta^*; z_i|x_i, y_i) = I(\bar{\theta}; z_i|x_i, y_i) = 0$. Again, the Danskin's theorem, implies that θ^* is a continuous function of L. For $L = 0$, the fitness function $\ell(\mathfrak{M}(x;\theta), y)$ is independent of θ because $0 \leq \|\ell(\mathfrak{M}(x;\theta), y) - \ell(\mathfrak{M}(x;\theta'), y)\| \leq L\|\theta - \theta'\| = 0$ for all $\theta, \theta' \in \Theta$. Thus $\lim_{\lambda\to\infty} \theta^* = \bar{\theta}$ and, similarly, $\lim_{\lambda\to\infty} I(\theta^*; z_i|x_i, y_i) = 0$. $\square$

References

Bertsekas, D. P. Control of uncertain systems with a set-membership description of uncertainty, 1971. Cambridge, MA: PhD Thesis, MIT.