Examen Master M2A
Analyse numérique et réseaux de neurones
Proposition de correction

15/04/2021

1 I

1. Pour $m = 1$, on a $R(x+d) = x+d$ pour $x+d \geqslant 0$ et $R(x+d) = 0$ pour $x+d \leqslant 0$. De même, $R(x) = x$ pour $x \geqslant 0$ et $R(x) = 0$ pour $x \leqslant 0$. Il y a quatre cas.

   - $x+d \geqslant 0$ et $x \geqslant 0$. Alors $R(x+d) - R(x) = x+d-x = d$. D'où l'inégalité.
   - $x+d \leqslant 0$ et $x \geqslant 0$. Alors $R(x+d) - R(x) = -x$ et
   $$|R(x+d) - R(x)| \leqslant |x|.$$
   Or $0 \leqslant x \leqslant -d$ d'après l'hypothèse. Donc $|x| \leqslant |d|$, d'où l'inégalité.
   - $x+d \geqslant 0$ et $x \leqslant 0$. Alors $R(x+d) - R(x) = x+d$.
   Mais $x \leqslant 0$ donc $x+d \leqslant d$ donc
   $$0 \leqslant x+d \leqslant d,$$
   Ce qui fait que $|x+d| \leqslant |d|$, d'où l'inégalité.
   - Le dernier cas est évident car $|R(x+d) - R(x)| = |0-0| = 0 \leqslant |d|$.

   Ensuite pour $m > 1$, on a
   $$\|R(x+d) - R(x)\| = \max_i |R(x_i+d_i) - R(x_i)| \leqslant \max_i |d_i| = \|d\|.$$

2. On a $f_0(x+d) - f_0(x) = W_0(x+d) + b_0 - W_0 x - b_0 = W_0 d$. Donc
   $$\|f_0(x+d) - f_0(x)\| = \|W_0 d\| \leqslant \|W_0\|\|d\|.$$
   D'où $K \leqslant \|W_0\|$.

3. On remarque que $f_1(x) = W_1 R(f_0(x)) + b_1$. Donc

$$f_1(x+d)-f_1(x) = W_1 R(f_0(x+d))+b_1-W_1 R(f_0(x+d))-b_1 = W_1\left(R(f_0(x+d)) - R(f_0(x))\right).$$

Il s'ensuit que

$$\begin{array}{rcl} \|f_1(x+d) - f_1(x)\| & \leqslant & \|W_1\| \times \|R(f_0(x+d)) - R(f_0(x))\| \\ & \leqslant & \|W_1\| \times \|f_0(x+d) f_0(x)\| \\ & \leqslant & \|W_1\| \times \|W_0\| \|d\|, \qquad \text{CQFD.} \end{array}$$

4. On remarque que $f_p(x) = W_p R(f_{p-1}(x)) + b_p$. Il s'ensuit que

$$\begin{array}{rcl} \|f_p(x+d) - f_p(x)\| & \leqslant & \|W_p\| \times \|R(f_{p-1}(x+d)) - R(f_{p-1}(x))\| \\ & \leqslant & \|W_p\| \times \|f_{p-1}(x+d) f_{p-1}(x)\| \\ & \leqslant & \|W_p\| \times \|W_{p-1}\| \times \cdots \times \|W_0\| \|d\|, \qquad \text{CQFD.} \end{array}$$

5. Evident.

6. On a la décomposition télecopique pour $\hat{x} = z$

$$f(\hat{x}) - f^{\text{obj}}(\hat{x}) = \left[f(\hat{x}) - f(x_i)\right] + \left[f(x_i) - f^{\text{obj}}(x_i)\right] + \left[f^{\text{obj}}(x_i) - f^{\text{obj}}(\hat{x})\right].$$

Or

$$\|f(\hat{x}) - f(x_i)\| \leqslant K\|\hat{x} - x_i\|,$$
$$\left\|f(x_i) - f^{\text{obj}}(x_i)\right\| = \|f(x_i) - y_i\| \leqslant \varepsilon_{\max},$$

et par la formule des accroissements finis

$$\left\|f^{\text{obj}}(x_i) - f^{\text{obj}}(\hat{x})\right\| \leqslant \|\nabla f^{\text{obj}}\| \|\hat{x} - x_i\|.$$

D'où le résultat par inégalité triangulaire et $\varepsilon_{\max} \leqslant \sqrt{N}\varepsilon$.

7. Par exemple, on peut essayer d'équilibrer les différentes erreurs en prenant

$$K \approx \frac{1}{Q}\left(\|\nabla f^{\text{obj}}\| Q + \sqrt{N}\varepsilon\right), \quad Q = \max_{z\in[0,1]^m} \max_{x_i} \|z - x_i\|.$$

En insérant la loi proposée, on obtient pour l'erreur

$$\|f(z) - f^{\text{obj}}(\hat{x})\| \leqslant \left(C\varepsilon^{-\alpha} + \|\nabla f^{\text{obj}}\|\right) Q + N^{\frac{1}{2}}\varepsilon$$

En tant que fonction de ε, le membre de droite diverge en 0^+ et diverge en $+\infty$. En équilibrant les termes, on obtient une valeur optimale

$$C\varepsilon_{\text{opt}}^{-\alpha} Q = N^{\frac{1}{2}}\varepsilon_{\text{opt}} \Longrightarrow \varepsilon_{\text{opt}} = \left(\frac{CQ}{\sqrt{N}}\right)^{\frac{1}{1+\alpha}}$$

ce qui fournit une valeur minimale de l'erreur.

2 Descente de gradient continue avec LASSO

1. La dérivée de la fonction valeur absolue est constante -1 pour $x < 0$ et constante +1 pour $x > 0$. Cette fonction n'est pas dérivable une seconde fois (cela ferait apparaître une masse de Dirac). Ou encore $g(x) = \frac{d}{dx}|x|$ n'est pas Lipschitz. En effet (on prend $y = -x$)

$$g(x) - g(-x) = 2 \text{ pour tout } x > 0.$$

   Si g était Lipschitz on aurait

$$2 = |g(x) - g(-x)| \leqslant K|x - y| = 2Kx \Longrightarrow K \geqslant \frac{1}{x} \text{ pour tout } x > 0.$$

   En faisant tendre $x \to 0^+$, on aurait $K \to +\infty$ ce qui est impossible. Donc g n'est pas Lipschitz.

   Partant, ∇J_1 n'est pas Lipschitz, ni ∇J_0.

2. La première question est évidente.

   La fonction $x \mapsto \sqrt{x^2 + \mu}$ est $C^\infty(\mathbb{R})$, donc J_1^μ est même de classe $C^\infty(\mathbb{R}^p)$. On a $\frac{d}{dx}\sqrt{x^2 + \mu} = \frac{x}{(x^2+\mu)^{\frac{1}{2}}}$ et

$$\frac{d^2}{dx^2}\sqrt{x^2 + \mu} = \frac{1}{(x^2 + \mu)^{\frac{1}{2}}} - \frac{1}{2}\frac{x^2}{(x^2 + \mu)^{\frac{3}{2}}} = \frac{\frac{1}{2}x^2 + \mu}{(x^2 + \mu)^{\frac{3}{2}}} > 0.$$

   Donc cette fonction est convexe, puis J_1^μ aussi.

3. On a

$$\left\langle \frac{d}{dt}W_\mu(t) - \nabla J_0(W_\mu(t)), Y - W_\mu(t) \right\rangle = \left\langle -\nabla J_1^\mu(W_\mu(t)), Y - W_\mu(t) \right\rangle.$$

   Comme J_1^μ est convexe, on a par une formule de Taylor à l'ordre 2 pour la fonction $f(\theta) = J_1^\mu(W + \theta(Y - W))$

$$J_1^\mu(Y) = J_1^\mu(W) + \langle \nabla J_1^\mu(W), Y - W \rangle$$
$$+\frac{1}{2}\left\langle \nabla^2 J_1^\mu(W + \theta(Y - W))(Y - W), Y - W \right\rangle.$$

   Donc

$$J_1^\mu(Y) \geqslant J_1^\mu(W) + \langle \nabla J_1^\mu(W), Y - W \rangle.$$

   Reportant plus haut, on obtient

$$\left\langle \frac{d}{dt}W_\mu(t) - \nabla J_0(W_\mu(t)), Y - W_\mu(t) \right\rangle + J_1^\mu(Y) - J_1^\mu(W_\mu(t))$$
$$= \langle -\nabla J_1^\mu(W_\mu(t)), Y - W_\mu(t) \rangle + J_1^\mu(Y) - J_1^\mu(W_\mu(t)) \geqslant 0.$$

4. On "enlève" les μ.

   L'intérêt est ne plus fair apparaître la fonction ∇J_1 qui est équivoque pour $w_i = 0$ (car elle n'est pas continue).