Exercises
Deep Learning
Fall 2018

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Fernando Pérez-Cruz
Web http://www.da.inf.ethz.ch/teaching/2018/DeepLearning/

Series Monday, Oct 15, 2018
(Deep Learning, Exercise series 3 - solutions)

Solution 1 (Activation Functions):

1. Not differentiable at 0, gradient 0 elsewhere. Thus not suited for brackprop.
2. $$\tanh(z) = \frac{1}{1+e^{-2z}} - \frac{1}{1+e^{2z}} = \frac{1}{1+e^{-2z}} - \frac{e^{-2z} \pm 1}{1+e^{-2z}} = \frac{2}{1+e^{-2z}} - 1 = 2\sigma(2z) - 1 \tag{1}$$

$$\begin{aligned} F(z) &= \tanh(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) && (2)\\ &= 2\sigma(4\mathbf{W}_2\, \sigma(2\mathbf{W}_1 \mathbf{x} + 2\mathbf{b}_1) + 2\mathbf{b}_2 - 2\mathbf{W}_1\mathbf{1}) - 1 && (3)\\ &= 2\sigma(\hat{\mathbf{W}}_2\, \sigma(\hat{\mathbf{W}}_1 \mathbf{x} + \hat{\mathbf{b}}_1) + \hat{\mathbf{b}}_2) - 1 && (4) \end{aligned}$$

Solution 2 (Elementary Logic Functions):

Goodfellow chapter 2.6. XOR problem is not linearly separable. $\mathbf{W}_2 = [1, -2]$ and $b_2 = 0$.

$$(w_1 w_2) \begin{bmatrix} 0 \\ 0 \end{bmatrix} + b_2 = 0 \rightarrow b_2 = 0 \tag{5}$$

$$(w_1 w_2) \begin{bmatrix} 2 \\ 1 \end{bmatrix} + b_2 = 0 \rightarrow w_2 = -2 \tag{6}$$

$$(w_1 w_2) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b_2 = 1 \rightarrow w_1 = 1 \tag{7}$$

$$(w_1 w_2) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b_2 = 1 \rightarrow w_1 = 1 \tag{8}$$

Solution 3 (Gradients of the Common Neural Network Layers):

For gradients of more common neural network layers see Appendix A of "Natural Language Processing (Almost) from Scratch", Collobert et al., 2011

Solution 4 (Derivation Gradient Descent):

1. The existence and uniqueness of a minimum holds for any strongly convex function [1]. Therefore, it would be enough to show that the function

$$F_k^\varepsilon := f(x) + \frac{1}{2\varepsilon}\|x - x_k^\varepsilon\|^2 \tag{9}$$

is strongly convex if $\varepsilon > 0$ is small enough. However, we will sove this problem as if we didn't know about strong convexity. In fact, the existence and uniqueness of the minimum of a strongly convex function can

[1] Don't confuse strong convexity with strict convexity, for which the existence of a minimum is not guaranteed.

be proved using the same approach.

1.1) Existence. The existence of the minimum relies on the extreme value theorem that states that if a function g is continuous on a closed and bounded (i.e. compact) set $S \in \mathbb{R}^n$, then g must attain a minimum at least once.

The plan is as follows.

- First, we present a compact set $S \subset \mathbb{R}^n$ outside which function (9) is bounded from below by its value at some point in this set, that is,

$$\exists x_0 \in S : \; F_k^\varepsilon(x_0) \leq \inf_{x \in \mathbb{R}^n \setminus S} F_k^\varepsilon(x), \text{ and therefore, } \inf_{x \in S} F_k^\varepsilon(x) \leq \inf_{x \in \mathbb{R}^n \setminus S} F_k^\varepsilon(x).$$

- Finally, as long as S is compact and function (9) is continuous, F_k^ε attains a minimum in S, that is,

$$\exists x_* \in S : \; F_k^\varepsilon(x_*) = \inf_{x \in S} F_k^\varepsilon(x),$$

  ans since

$$\inf_{x \in \mathbb{R}^n} F_k^\varepsilon(x) = \min(\inf_{x \in S} F_k^\varepsilon(x), \inf_{x \in \mathbb{R}^n \setminus S} F_k^\varepsilon(x)) = \inf_{x \in S} F_k^\varepsilon(x) = F_k^\varepsilon(x_*),$$

  this proves the existence.

To present such S, let us note that the second term in Eq. 9 is positive and goes rather quickly to infinity, and f varies not too violently (its partial derivatives are bounded), which means that the infimum should be reached close to x_k^ε if ε is small enough.

In other words, if ε is small enough, then there exists some $R > 0$ such that for each x outside the closed ball $\overline{B_R(x_k^\varepsilon)} = \{x \in \mathbb{R}^n \mid \|x - x_k^\varepsilon\| \leq R\}$ the value of the function at x is guaranteed to be greater than the value of this function at some point inside the closed ball $\overline{B_R(x_k^\varepsilon)}$. Therefore, the infimum of the function over $\mathbb{R}^n$ equals to the infimum over the closed ball $\overline{B_R(x_k^\varepsilon)}$, which is always a compact set in $\mathbb{R}^n$. Thus, as a continuous function on a compact always reaches its infimum on it, *i.e.* it is a minimum, this will give the existence.

Let's prove it rigorously.

First, we review some mathematical concepts from real analysis and calculs that will be necessary for the main proof. Readers familiar with these concepts can directly go to the main proof.

Let's first review the Taylor expansion of a function $f \in C^{n+1}([a,b], \mathbb{R})$ in a neighbourhood of a point $x_0 \in (a,b)$. The notation $f \in C^{n+1}([a,b], \mathbb{R})$ means the function is differentiable $n+1$ times and is a mapping from $[a,b]$ to $\mathbb{R}$.

The Taylor-Lagrange formula

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n \; dt \,. \tag{10}$$

We will prove this formula by induction on n. Note that this is trivial for $n = 0$, as $f(x) = f(x_0) + \int_{x_0}^{x} f'(t) \; dt$.

Let's fix some integer $n \in \mathbb{N}$. Note that with an integration by parts, we have:

$$\int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n \; dt = -\frac{f^{(n+1)}(t)}{(n+1)!}(x-t)^{n+1}\Big|_{x_0}^{x} + \int_{x_0}^{x} \frac{f^{(n+2)}(t)}{(n+1)!}(x-t)^{n+1} \; dt \,. \tag{11}$$

Now assume the formula at rank n:

$$\forall x \in [a,b], \; \forall n \in \mathbb{N}, \; f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n \; dt \,. \tag{12}$$

By plugging equation (11) into (12), we get:

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \frac{f^{(n+1)}(x_0)}{(n+1)!}(x - x_0)^{n+1} + \int_{x_0}^{x} \frac{f^{(n+2)}(t)}{(n+1)!}(x-t)^{n+1} \; dt, \tag{13}$$

which is the formula at rank $n + 1$.

The Taylor expansion with the Lagrange remainder

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + r_n(f, x), \tag{14}$$

where $r_n(f, x) = \dfrac{f^{(n+1)}(x_0 + \theta(x - x_0))}{(n+1)!}(x - x_0)^{n+1}$ for some $\theta \in (0, 1)$.

Again, let us use induction on n to prove this formula. For $n = 0$, the formula follows from the mean-value theorem. Let us assume that the formula holds for $n - 1 \geq 0$, that is,

$$\forall f \in C^n([a, b]),\ x \in (a, b)\ \exists \theta \in (0, 1) :\ r_{n-1}(f, x) = \frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n.$$

Then, using the Cauchy's mean-value theorem we get

$$\begin{aligned}
\frac{r_n(f, x)}{(x - x_0)^{n+1}} &= \frac{r_n(f, x) - r_n(f, x_0)}{(x - x_0)^{n+1} - (x_0 - x_0)^{n+1}} = \frac{r_n'(f, x_0 + \eta(x - x_0))}{(n+1)(\eta(x - x_0))^n} \\
&= \frac{f'(x_0 + \eta(x - x_0)) - \sum_{k=1}^{n} \dfrac{f^{(k)}(x_0)}{(k-1)!}(\eta(x - x_0))^{k-1}}{(n+1)(\eta(x - x_0))^n} \\
&= \frac{r_{n-1}(f', x_0 + \eta(x - x_0))}{(n+1)(\eta(x - x_0))^n} = \frac{\dfrac{f^{(n+1)}(x_0 + \theta(x - x_0))}{n!}(\eta(x - x_0))^n}{(n+1)(\eta(x - x_0))^n} = \frac{f^{(n+1)}(x_0 + \theta(x - x_0))}{(n+1)!},
\end{aligned}$$

for some $0 < \theta < \eta < 1$, which concludes the proof.

The Taylor expansion of a multi-variable function

The Taylor formulas can be easily generalized for the functions of multiple variables. Let us do that for $f \in C^{n+1}(\mathbb{R}^d \to \mathbb{R})$, in a neighborhood of $y \in \mathbb{R}^d$.

Let us introduce an auxiliary function $g(t) := f(y + t(x - y))$ and write the Taylor expansion formula with the Lagrange remainder for it in a neighborhood of $t = 0$.

$$g(t) = \sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!} t^k + \frac{g^{(n+1)}(\theta t)}{(n+1)!} t^{n+1}$$

for some $\theta \in (0, 1)$. The derivatives can be computed using the chain rule:

$$g^{(k)}(\xi) = \frac{\partial^k}{\partial t^k} f(y + t(x - y))\Big|_{t=\xi} = \sum_{\substack{\alpha_1, \ldots, \alpha_d \geq 0 \\ \sum_{i=1}^{d} \alpha_i = k}} \frac{k!}{\alpha_1! \ldots \alpha_d!} \frac{\partial^k f(y + \xi(x - y))}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} (x_1 - y_1)^{\alpha_1} \ldots (x_d - y_d)^{\alpha_d}.$$

Thus,

$$\begin{aligned}
f(x) = g(1) = &\sum_{k=0}^{n} \sum_{\substack{\alpha_1, \ldots, \alpha_d \geq 0 \\ \sum_{i=1}^{d} \alpha_i = k}} \frac{1}{\alpha_1! \ldots \alpha_d!} \frac{\partial^k f(y)}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} (x_1 - y_1)^{\alpha_1} \ldots (x_d - y_d)^{\alpha_d} \\
&+ \sum_{\substack{\alpha_1, \ldots, \alpha_d \geq 0 \\ \sum_{i=1}^{d} \alpha_i = n+1}} \frac{1}{\alpha_1! \ldots \alpha_d!} \frac{\partial^{n+1} f(y + \theta(x - y))}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} (x_1 - y_1)^{\alpha_1} \ldots (x_d - y_d)^{\alpha_d}.
\end{aligned}$$

For $n = 1$ this formula can be written as follows:

$$f(x) = f(y) + \langle f'(y), x - y \rangle + \frac{1}{2} \langle f''(y + \theta(x - y))(x - y), x - y \rangle$$

for some $\theta \in (0, 1)$, where $\langle \cdot, \cdot \rangle$ denotes the inner product on $\mathbb{R}^d$, $f'(y) = (\frac{\partial f}{\partial x_1} f(y), \ldots, \frac{\partial f}{\partial x_d} f(y))$ and $f''(y) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(y)\right]$ are the gradient and the Hessian of function f at the given point y.

A few more definitions

Def. *A set $S \subset \mathbb{R}^n$ is called compact in the n-dimensional Euclidean space if it is closed (i.e. contains its boundary) and bounded.*

Def. *A function $f : U \to \mathbb{R}$ is called continuous at a point $x_0 \in U \subset \mathbb{R}^n$ if*

$$\forall \varepsilon > 0\ \exists \delta = \delta(\varepsilon) > 0 :\ |f(x) - f(x_0)| < \varepsilon\ \forall x \in B_\delta(x_0) \cap U,$$

where $B_\delta(x_0) = \{x \in \mathbb{R}^n \mid \|x - x_0\| < \delta\}$ is the open ball of radius δ around x_0.

Def. *A function f is called continuous in a set $U \subset \mathbb{R}^n$ if it is continuous at each point of that set.*

Back to the main proof. To derive a lower bound of the function F_k^ε, let us write the Taylor expansion of $F_k^\varepsilon(x)$ in the neighborhood of $x = x_k^\varepsilon$ with the Lagrange's form of the remainder:

$$\begin{aligned} F_k^\varepsilon(x) &= F_k^\varepsilon(x_k^\varepsilon) + \langle F_k^{\varepsilon\prime}(x_k^\varepsilon), x - x_k^\varepsilon\rangle + \frac{1}{2}\langle F_k^{\varepsilon\prime\prime}(x_k^\varepsilon + \theta(x - x_k^\varepsilon))(x - x_k^\varepsilon), x - x_k^\varepsilon\rangle \\ &= F_k^\varepsilon(x_k^\varepsilon) + \langle f'(x_k^\varepsilon), x - x_k^\varepsilon\rangle + \frac{1}{2}\langle f''(x_k^\varepsilon + \theta(x - x_k^\varepsilon))(x - x_k^\varepsilon), x - x_k^\varepsilon\rangle + \frac{1}{2\varepsilon}\|x - x_k^\varepsilon\|^2, \end{aligned}$$

for some $0 < \theta < 1$.

As the partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ are bounded, the norm of the Hessian $f''(x)$ is also bounded. Let us denote

$$M := \sup_{x\in\mathbb{R}^n} \|f''(x)\|.$$

Then we have

$$\begin{aligned} F_k^\varepsilon(x) &= F_k^\varepsilon(x_k^\varepsilon) + \langle f'(x_k^\varepsilon), x - x_k^\varepsilon\rangle + \frac{1}{2\varepsilon}\|x - x_k^\varepsilon\|^2 + \frac{1}{2}\langle f''(x_k^\varepsilon + \theta(x - x_k^\varepsilon))(x - x_k^\varepsilon), x - x_k^\varepsilon\rangle \\ &\geq F_k^\varepsilon(x_k^\varepsilon) - \|f'(x_k^\varepsilon)\|\|x - x_k^\varepsilon\| + \left(\frac{1}{2\varepsilon} - \frac{M}{2}\right)\|x - x_k^\varepsilon\|^2. \end{aligned}$$

Here we used the following inequalities:

(a) the Schwarz's inequality $|\langle x, y\rangle| \leq \|x\|\|y\|$, which implies $\langle x, y\rangle \geq -\|x\|\|y\|$ $\forall x,\, y \in \mathbb{R}^n$,

(b) the definition of a matrix norm: $\|A\| := \sup_{x\in\mathbb{R}^n,\, \|x\|=1} \|Ax\|$, which implies $\|Ax\| \leq \|A\|\|x\|$ $\forall x \in \mathbb{R}^n$.

Using both inequalities, one can write $|\langle Ax, x\rangle| \leq \|Ax\|\|x\| \leq \|A\|\|x\|^2$ and hence $\langle Ax, x\rangle \geq -\|A\|\|x\|^2$ $\forall x \in \mathbb{R}^n$.

Thus,

$$F_k^\varepsilon(x) - F_k^\varepsilon(x_k^\varepsilon) \geq \left(\left(\frac{1}{2\varepsilon} - \frac{M}{2}\right)\|x - x_k^\varepsilon\| - \|f'(x_k^\varepsilon)\|\right)\|x - x_k^\varepsilon\|.$$

If $\frac{1}{2\varepsilon} - \frac{M}{2} > 0$, the term on the right-hand side goes to infinity as $\|x - x_k^\varepsilon\| \to \infty$, hence we choose $\varepsilon \in (-\infty, \frac{1}{M})$. Now, as the expression on the left-hand side goes to infinity as well, there exists $R > 0$ such that

$$F_k^\varepsilon(x) > F_k^\varepsilon(x_k^\varepsilon) \text{ for all } x \in \mathbb{R}^n \text{ such that } \|x - x_k^\varepsilon\| > R.$$

Hence

$$\inf_{x\in\mathbb{R}^n} F_k^\varepsilon(x) = \inf_{x:\, \|x - x_k^\varepsilon\| \leq R} F_k^\varepsilon(x). \tag{15}$$

As the closed ball $\{x \in \mathbb{R}^n \mid \|x - x_k^\varepsilon\| \leq R\}$ of radius R and centered on x_k^ε in $\mathbb{R}^n$ is compact, and as the function F_k^ε is continuous, it reaches its global minimum on this compact set, which is also its global minimum on $\mathbb{R}^n$ due to (15), as F_k^ε is greater than its value at x_k^ε everywhere outside this ball.

1.2) Uniqueness. We just need to show that we can choose ε such that F_k^ε is strictly convex, or that its Hessian matrix is positive definite everywhere on $\mathbb{R}^n$. As $\|f''(x)\| \leq M$, hence by choosing $\varepsilon < \frac{1}{M}$, we can ensure that $F_k^{\varepsilon\prime\prime}(x) = f'' + \frac{1}{\varepsilon}I_n \succeq (\frac{1}{\varepsilon} - M)I_n \succ 0$, where I_n is the identity matrix.

Def. *A matrix $A \in \mathbb{R}^{n\times n}$ is called positive definite (positive semidefinite) if $\langle Ax, x\rangle > 0$ ($\langle Ax, x\rangle \geq 0$) $\forall x \in \mathbb{R}^n\backslash\{0\}$.*

Above, we were using the following property.

Cor. $A \preceq \|A\|I_n$ $\forall A \in \mathbb{R}^{n\times n}$, or equivalently, $-A \succeq -\|A\|I_n$ i.e. the matrix $\|A\|I_n - A$ is positive semidefinite.

Proof

$$\langle(\|A\|I_n - A)x, x\rangle = \|A\|\|x\|^2 - \langle Ax, x\rangle \geq \|A\|\|x\|^2 - \|Ax\|\|\|x\| \geq \|A\|\|x\|^2 - \|A\|\|x\|^2 = 0 \quad \forall x \in \mathbb{R}^n.$$

2. As x_{k+1}^ε is the global minimum of the differentiable function F_k^ε defined on the open set $\mathbb{R}^n$, the gradient of F_k^ε vanishes at x_{k+1}^ε, *i.e.* $F_k^{\varepsilon\prime}(x_{k+1}^\varepsilon) = 0$ $\forall k$, and hence

$$\frac{1}{\varepsilon}(x_{k+1}^\varepsilon - x_k^\varepsilon) = -f'(x_{k+1}^\varepsilon).$$

Solution 5 (Local quadratic approximation):

Recall that we can approximate a function $f(\boldsymbol{w})$ using a Taylor expansion around a point $\bar{\boldsymbol{w}}$ as

$$f(\boldsymbol{w}) = f(\bar{\boldsymbol{w}}) + (\boldsymbol{w} - \bar{\boldsymbol{w}})^\top \nabla f(\bar{\boldsymbol{w}}) + \frac{1}{2}(\boldsymbol{w} - \bar{\boldsymbol{w}})^T H(\boldsymbol{w} - \bar{\boldsymbol{w}}) + O(\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^3) \tag{16}$$

Let's now consider the particular case of a local quadratic approximation around a point $\boldsymbol{w}^*$ that is a minimum of the error function (i.e. $\boldsymbol{w}^* \in \operatorname{argmin}_{\boldsymbol{w}} f(\boldsymbol{w})$). In this case the linear term vanishes since $\nabla f(\boldsymbol{w}^*) = 0$. We therefore get

$$\begin{aligned}
f(\boldsymbol{w}) &\approx f(\boldsymbol{w}^*) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^\top H(\boldsymbol{w} - \boldsymbol{w}^*) \\
&= f(\boldsymbol{w}^*) + \frac{1}{2}(\sum_{i=1}^d \alpha_i \boldsymbol{u}_i)^\top H(\sum_{i=1}^d \alpha_i \boldsymbol{u}_i) \\
&= f(\boldsymbol{w}^*) + \frac{1}{2}(\sum_{i=1}^d \alpha_i \boldsymbol{u}_i)^\top (\sum_{i=1}^d \alpha_i H \boldsymbol{u}_i) \\
&= f(\boldsymbol{w}^*) + \frac{1}{2}(\sum_{i=1}^d \alpha_i \boldsymbol{u}_i)^\top (\sum_{i=1}^d \alpha_i \lambda_i \boldsymbol{u}_i).
\end{aligned}$$

Recall that the eigenvectors are orthonormal, i.e. $\boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{ij}$. We then get

$$\begin{aligned}
f(\boldsymbol{w}) &\approx f(\boldsymbol{w}^*) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2 \boldsymbol{u}_i^\top \boldsymbol{u}_i \\
&= f(\boldsymbol{w}^*) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2.
\end{aligned}$$

Solution 6 (Weierstrass theorem):

1. Using $\frac{1}{J_n}\int_{-1}^1 (1-u^2)^n \, du = 1$ it is easy to see that we can write $f(x)$ as

$$f(x) = \frac{1}{J_n}\int_{-1}^1 f(x)(1-u^2)^n \, du$$

2. We first note that we can rewrite $P_n(x) = \frac{1}{J_n}\int_0^1 f(t)[1-(t-x)^2]^n \, dt$ as

$$\begin{aligned}
P_n(x) &= \frac{1}{J_n}\int_{-1+x}^{1+x} f(t)[1-(t-x)^2]^n \, dt \\
&= \frac{1}{J_n}\int_{-1}^1 f(x+u)[1-u^2]^n \, du,
\end{aligned}$$

where the first equality uses the fact that $f(x) = 0$ outside $(0, 1)$ and the second equality is a simple substitution $t - x = u$ (note that this changed the domain of integration). Therefore,

$$P_n(x) - f(x) = \frac{1}{J_n}\int_{-1}^1 [f(x+u) - f(x)](1-u^2)^n \, du\,. \tag{17}$$

3. Since f is continuous on the compact $[a-1, b+1]$, it is uniformly continuous on it and hence

$$\forall \epsilon > 0 \; \exists \delta = \delta(\epsilon) > 0 : \; |f(x+u) - f(x)| \leq \frac{\epsilon}{2} \; \forall x \in [a,b], u \in \overline{B_\delta(0)}.$$

For the case $|u| > \delta$, we use the bound $|f(x)| \leq M$ to derive the following inequality

$$|f(x+u) - f(x)| \leq 2M \leq 2M\frac{u^2}{\delta^2}.$$

For any value of u, one of the other of the 2 quantities in the upper bound is greater than or equal to $|f(x+u)-f(x)|$, and therefore

$$|f(x+u)-f(x)| \leq \frac{\epsilon}{2} + 2M\frac{u^2}{\delta^2}. \tag{18}$$

Def. *A function $f: U \to \mathbb{R}$ is called uniformly continuous on a set $U \subset \mathbb{R}^n$ if*

$$\forall \varepsilon > 0\ \exists \delta = \delta(\varepsilon) > 0:\ \forall x, y \in U \quad \|x-y\| < \delta \implies |f(x)-f(y)| < \varepsilon.$$

Thm. (Heine-Cantor) *Every continuous function on a compact set is uniformly continuous on it.*

4. By applying integration by parts, we get

$$J_n' = \int_{-1}^{1} u \cdot u(1-u^2)^n\, du = -u\frac{(1-u^2)^{n+1}}{2(n+1)}\Bigg|_{-1}^{1} + \int_{-1}^{1} \frac{(1-u^2)^{n+1}}{2(n+1)}\, du = \frac{J_{n+1}}{2(n+1)}. \tag{19}$$

5. Combining equations (17) and (18) and substituting (19), we get

$$\begin{aligned}|P_n(x)-f(x)| &\leq \frac{1}{J_n}\int_{-1}^{1}\frac{\epsilon}{2}(1-u^2)^n du + \frac{1}{J_n}\int_{-1}^{1} 2M\frac{u^2}{\delta^2}(1-u^2)^n du \\ &= \frac{\epsilon}{2} + \frac{2M}{\delta^2 J_n}J_n' = \frac{\epsilon}{2} + \frac{MJ_{n+1}}{\delta^2 J_n(n+1)} < \frac{\epsilon}{2} + \frac{M}{\delta^2(n+1)}\end{aligned}$$

since $J_{n+1} < J_n$.
It follows that for n sufficiently large, $\frac{M}{\delta^2(n+1)} < \frac{\epsilon}{2}$, which concludes the proof.

Solution 7 (Weierstrass theorem, simplification in the $\mathcal{C}^\infty$ case):

Let $c \in (a,b)$. By the Taylor-Lagrange formula, we have

$$\forall x \in [a,b],\ \forall n \in \mathbb{N},\ f(x) = \sum_{k=0}^{n}\frac{f^{(k)}(c)}{k!}(x-c)^k + \int_c^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n\, dt\,. \tag{20}$$

If we set $P_n(x) := \sum\limits_{k=0}^{n}\dfrac{f^{(k)}(c)}{k!}(x-c)^k$, we have

$$\|f-P_n\|_\infty = \max_{x\in[a,b]}\left|\int_c^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n\, dt\right| = O\left(\frac{q^{n+1}(b-a)^n}{n!}(b-a)\right) = o(1),\ \text{when } n \to \infty. \tag{21}$$