MDL - Homework

06/02/2024

Instructions: Please send your solution to the following exercises before 20/02/2024 end of day to the email address: `kevin.scaman@inria.fr`.

Exercise 1: Parallel automatic differentiation

We consider a sequential neural network of the form $g_\theta(x) = g^{(L)}(x, \theta)$ where $g^{(0)}(x, \theta) = x$ and $\forall l \in [\![1, L]\!]$,

$$g^{(l)}(x, \theta) = f^{(l)}\left(g^{(l-1)}(x, \theta),\, \theta^{(l)}\right),$$

where $\theta = (\theta^{(1)}, \dots, \theta^{(L)})$, $\theta^{(l)} \in \mathbb{R}^{p^{(l)}}$ and $f^{(l)} : \mathbb{R}^{d^{(l-1)}} \times \mathbb{R}^{p^{(l)}} \to \mathbb{R}^{d^{(l)}}$. When the output dimension $d^{(L)} = 1$, backpropagation is very efficient w.r.t. computation time. Our objective is to propose more efficient algorithms when we have access to a large number of parallel workers. For simplicity, we will consider that all layer widths are equal $d^{(l)} = w$ for $l \in [\![1, L-1]\!]$, and $d^{(L)} = 1$.

Q1: What is the computational complexity of the backpropagation algorithm (up to a multiplicative constant)?

Q2: Write the derivative of the model $g^{(L)}$ w.r.t. the parameter $\theta^{(l)}$ for $l \in [\![1, L]\!]$ as a product of $L - l + 1$ matrices.

Q3: Show that a product of K matrices can be computed in $\lceil \ln_2(K) \rceil$ iterations using parallel workers. What is the computational complexity (up to a multiplicative constant)? How many workers are necessary?

Q4: Propose an algorithm to compute the gradient of the model $g^{(L)}$ w.r.t. any parameter $\theta^{(l)}$ in time proportional to $\lceil \ln_2(L - l + 1) \rceil$. What is the computational complexity (up to a multiplicative constant)? When is this algorithm faster than backpropagation?

Q5: If all partial derivatives w.r.t. $\theta^{(l)}$ are computed in parallel, how many workers are necessary?

Q6: Show that most computations can be reused for multiple partial derivatives (identical products appearing in the derivatives), and propose an algorithm taking advantage of this fact. Show that the number of workers needed for this algorithm is linear in L instead of quadratic, while still requiring a computation time in $O(\ln_2(L))$.

Exercise 2: Non-smooth optimization via random noise

Let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ be a differentiable objective function lower bounded by $\mathcal{L}^\star \in \mathbb{R}$. We first assume that $\mathcal{L}$ is β-smooth and μ-PL. Let $(\theta_t)_{t \in \mathbb{N}}$ be generated as follows, with $\eta < \frac{1}{\beta}$:

$$\theta_{t+1} = \theta_t - \eta G_t\,,$$

for stochastic gradients G_t verifying $\mathbb{E}(G_t) = \nabla\mathcal{L}(\theta_t)$ and $\mathbb{E}(\|G_t - \nabla\mathcal{L}(\theta_t)\|^2) \le \sigma^2$.

Q7: Provide an upper bound on the optimization error in expectation $\mathbb{E}(\mathcal{L}(\theta_t) - \mathcal{L}^\star)$. What is the optimal step size, and which convergence rate do you obtain?

Q8: Is the objective function used to train a deep learning model always smooth? If not, give a counter-example.

We now consider that $\mathcal{L}$ is no longer β-smooth, but only L-Lipschitz (and μ-PL). For $\gamma > 0$ we define the function $\mathcal{L}^\gamma$ as:

$$\mathcal{L}^\gamma(\theta) = \mathbb{E}(\mathcal{L}(\theta + \gamma\xi)) \,, \quad \text{where} \quad \xi \sim \mathcal{N}(0, I_d) \,, \quad x \in \mathbb{R}^d \,.$$

Q9: Show that $\mathcal{L}^\gamma$ is an approximation of $\mathcal{L}$, and in particular $\forall\theta \in \mathbb{R}^d$, $|\mathcal{L}^\gamma(\theta) - \mathcal{L}(\theta)| \leq L\gamma\sqrt{d}$.

Q10: Using integration by parts, show that $\nabla\mathcal{L}^\gamma(\theta) = \frac{1}{\gamma}\mathbb{E}(\mathcal{L}(\theta + \gamma X)X)$.

Q11: Use this result to show that, $\forall v, \theta, \theta' \in \mathbb{R}^d$, $|\langle\nabla\mathcal{L}^\gamma(\theta) - \nabla\mathcal{L}^\gamma(\theta'), v\rangle| \leq \frac{L\|v\|}{\gamma}$.

Q12: Conclude on the smoothness of $\mathcal{L}^\gamma$ and find its smoothness constant.

Q13: We now assume that $\operatorname{var}(\nabla\mathcal{L}(\theta + \gamma X)) \leq c\mathbb{E}(\|\nabla\mathcal{L}(\theta + \gamma X)\|^2)$ where $c \in (0, 1)$. Show that $\mathcal{L}^\gamma$ verifies the $(1 - c)\mu$-PL condition.

Q14: Can the gradient $\nabla\mathcal{L}^\gamma(\theta)$ be estimated by a fixed number K of samples? What is the variance of this estimator?

Q15: Finally, propose a minimization method for $\mathcal{L}$ using K samples of the gradient $\nabla\mathcal{L}(\theta_t + \gamma\xi_{t,k})$ at each iteration $t \geq 0$, where $(\xi_{t,k})_{k\in[\![1,K]\!]} \sim \mathcal{N}(0, I_d)$ are i.i.d. Gaussian random variables. What is the optimal step size, and which convergence rate do you obtain?