MDL - TD5

20/02/2024

Exercise 1: Gradient clipping

The goal of this exercise is to study GD and SGD with clipping, an algorithm of choice to make the training of DNNs robust to noise, adversaries, or to make it private. We first show that the norm of the gradients for GD tend to 0, and then that for SGD classical analyses will fail.

Let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ be a β-**smooth** function *i.e.*, such that for all $\theta, \theta' \in \mathbb{R}^d$, we have:

$$\mathcal{L}(\theta') \leq \mathcal{L}(\theta) + \nabla\mathcal{L}(\theta')^\top(\theta - \theta') + \frac{L}{2}\|\theta - \theta'\|^2 \,.$$

We assume that $\mathcal{L}$ is lower bounded, and minimized over $\mathbb{R}^d$ at some point $\theta^\star \in \mathbb{R}^d$. We first consider **gradient descent** on $\mathcal{L}$, with **gradient clipping** to prevent gradients from exploding. More precisely, for some $\gamma > 0, c > 0$, and $\theta_0 \in \mathbb{R}^d$,

$$\theta_{k+1} = \theta_k - \gamma\mathcal{C}(\nabla\mathcal{L}(\theta_k))\,,$$

where for $g \in \mathbb{R}^d$, $\mathcal{C}(g) = g$ if $\|g\| \leq c$ and $\mathcal{C}(g) = c\frac{g}{\|g\|}$ if $\|g\| > c$.

Q1: Explain what clipping does, with words.

Q2: Write the recursion verified by (θ_k) in the following form:

$$\theta_{k+1} = \theta_k - \gamma_k\nabla\mathcal{L}(\theta_k)\,.$$

Q3: Using the smoothness of $\mathcal{L}$, bound $\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k)$ as:

$$\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k) \leq -\gamma_k\Big(1 - \frac{\gamma_k\beta}{2}\Big)\|\nabla\mathcal{L}(\theta_k)\|^2\,.$$

Q4: Show that for $\gamma \leq \frac{1}{\beta}$ we have

$$\sum_{k<K}\gamma_k\|\nabla\mathcal{L}(\theta_k)\|^2 \leq 2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^\star))\,.$$

Q5: Show that $\|\nabla\mathcal{L}(\theta_k)\| \to 0$, and derive a rate of convergence.

We now focus on **SGD**, the algorithm used in practice:

$$\theta_{k+1} = \theta_k - \gamma\mathcal{C}(g_k)\,,$$

where g_k is an unbiased **stochastic gradient** estimate, verifying $\mathbb{E}g_k = \nabla\mathcal{L}(\theta_k)$ and of **variance** smaller than σ^2.

Q6: Write the recursion verified by (θ_k) in the following form:

$$\theta_{k+1} = \theta_k - \gamma_k g_k\,.$$

Q7: Using the smoothness of $\mathcal{L}$, bound $\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k)$ as,

$$\mathbb{E}\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k) \leq -\mathbb{E}[\gamma_k g_k]^\top \nabla\mathcal{L}(\theta_k) + \frac{\gamma^2\beta}{2}(c^2 + \sigma^2)\,.$$

where the expectation is taken wrt g_k, conditionally on θ_k.

Q8: Construct an adversarial example of stochastic gradient showing that it is impossible to derive convergence of the gradients to 0 for SGD with clipping for any smooth function and noise distribution.

Q9: Assuming that $\mathbb{E}[\gamma_k g_k]^\top \nabla\mathcal{L}(\theta_k) \geq \frac{1}{2}\|\nabla\mathcal{L}(\theta_k)\|_2^2$ for all k, prove that $\|\nabla\mathcal{L}(\theta_k)\|_2 \to 0$.