MDL - TD6

27/02/2024

Exercise 1: Lazy Training in Deep Learning

Consider the minimization using gradient methods, of an objective function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^+$ defined as

$$\mathcal{L}(\theta) = R(h(\theta)),$$

where $\mathbb{R}^d$ is the parameter space, $h : \mathbb{R}^d \to \mathcal{H}$ maps smoothly parameters to some bounded function space $\mathcal{H}$ and $R : \mathcal{H} \to \mathbb{R}$ is a smooth loss. We assume that h is differentiable with a Lipschitz differential ($\|Dh(\theta) - Dh(\theta')\| \leq L_{Dh}\|\theta - \theta'\|$), and similarly, R is L_R-smooth.

Lazy training refers to the case where, while performing gradient steps on iterates $(\theta^t)_{t\geq 0}$, the loss $\mathcal{L}(\theta^t)$ drastically decreases while the differentials $Dh(\theta^t)$ do not sensibly change. We initialize a gradient-based method in a point $\theta_0 \in \mathbb{R}^d$ that is neither a minimizer of $\mathcal{L}$ (i.e. $\mathcal{L}(\theta_0) > 0$) nor a critical point (i.e. $\nabla\mathcal{L}(\theta_0) \neq 0$). We consider a single gradient descent step:

$$\theta_1 = \theta_0 - \eta\nabla\mathcal{L}(\theta_0).$$

Q1: Approximate the relative change in objective $\Delta(\mathcal{L}) = \frac{|\mathcal{L}(\theta_1) - \mathcal{L}(\theta_0)|}{\mathcal{L}(\theta_0)}$ in terms of $\mathcal{L}(\theta_0)$, $\nabla\mathcal{L}(\theta_0)$ and $\eta > 0$.

Q2: Approximate the relative change in differential of h, $\Delta(Dh) = \frac{\|Dh(\theta_1) - Dh(\theta_0)\|}{\|Dh(\theta_0)\|}$, in terms of $\nabla\mathcal{L}(\theta_0)$, $D^2h(\theta_0)$, $Dh(\theta_0)$ and η.

Q3: Show that lazy training occurs when

$$\frac{\|D^2h(\theta_0)\|}{\|Dh(\theta_0)\|} \ll \frac{\|\nabla\mathcal{L}(\theta_0)\|}{\mathcal{L}(\theta_0)}.$$

Q4: For the square loss $R(f) = \frac{1}{2}\|f - f^\star\|^2$ for some fixed target function $f^\star$, this leads to

$$\kappa_h(\theta_0) := \|h(\theta_0) - f^\star\| \times \frac{\|D^2h(\theta_0)\|}{\|Dh(\theta_0)\|^2} \ll 1.$$

For $\alpha > 0$, derive an expression for $\kappa_{\alpha h}$, and deduce how to choose the scaling α and $h(\theta_0)$ the initialization in order to favor lazy training.

Q5: Assume now that h is a L-layer MLP: $\theta = (W_1, \ldots, W_L)$ and

$$h(\theta) = W_L\sigma(W_{L-1}\sigma(W_{L-2}\ldots\sigma(W_1 z)\ldots),$$

where W_l are the weight matrices, σ is homogeneous activation function ($\sigma(\lambda z) = \lambda\sigma(z)$). For $\lambda > 0$, express $h(\lambda\theta)$ in terms of λ and $h(\theta)$.

Q6: Deduce an expression for $\kappa_h(\lambda\theta)$. When does lazy training occur in these networks?