MDL - TD2

23/01/2024

Exercise 1: Cross entropy and the Łojasiewicz condition

We consider a classification problem with $C \geq 1$ classes, and a training dataset $(x_i, y_i)_{i \in [\![1,N]\!]} \in (\mathbb{R}^d \times \{0,1\}^C)^N$. For a model that output scores $g_\theta(x) \in \mathbb{R}^C$ for each class, we define the cross entropy loss $\ell_{CE} : \mathbb{R}^C \times \{0,1\}^C \to \mathbb{R}_+$ as

$$\ell_{CE}(x, y) = -\sum_{k=1}^{C} y_k \ln\left(\frac{e^{x_k}}{\sum_{l=1}^{C} e^{x_l}}\right).$$

Q1: Does cross entropy verify the Polyak-Łojasiewicz condition w.r.t. to its first input?

Q2: Show that, when $\sum_i y_i = 1$, the cross entropy loss verifies a Łojasiewicz condition (w.r.t. to its first input) of the form

$$\|\nabla_x \ell_{CE}(x, y)\| \geq 1 - e^{-\ell_{CE}(x,y)}$$

Exercise 2: Gradient noise and risk minimization

We consider a risk minimization setting in which our objective is to minimize

$$\min_\theta \mathcal{L}(\theta) = \mathbb{E}(\ell(\theta, Z)),$$

where $\theta \mapsto \ell(\theta, z)$ is a β-smooth loss function, and Z is drawn according to a certain data distribution. Moreover, let $\theta^\star \in \operatorname{argmin}_\theta \mathcal{L}(\theta)$ be an optimizer of the objective.

Q3: Show that, for any data sample z, we have

$$\ell(\theta, z) - \min_\theta \ell(\theta, z) \geq \frac{1}{2\beta}\|\nabla_\theta \ell(\theta, z)\|^2.$$

Q4: From the above equation, prove that

$$\mathbb{E}(\|\nabla_\theta \ell(\theta, Z)\|^2) \leq 2\beta\left(\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) + \Delta\right),$$

where $\Delta = \mathbb{E}(\ell(\theta^\star, Z) - \min_\theta \ell(\theta, Z))$. How can we interpret this inequality, given that our stochastic gradients are $\nabla_\theta \ell(\theta, Z_t)$ for a data sample Z_t?

Q5: Show that $\mathcal{L}$ is also β-smooth.

Q6: We now assume that $\mathcal{L}$ satisfies the μ-PL assumption. Derive a bound on the approximation error $\mathbb{E}(\mathcal{L}(\theta_t) - \mathcal{L}(\theta^\star))$ for SGD with $\theta_{t+1} = \theta_t - \eta\nabla_\theta \ell(\theta, Z_t)$.

Q7: What is the optimal step size? Compare the convergence rate with the setting in which the noise on the gradient has a bounded variance.