Lecture 20

在这一章中，我们将 review deterministic GD-type optimization methods

§1 Plain-vanilla GD optimization methods

1. Definition: GD optimization method

令 ① objective function 的 input 维度：$d \in \mathbb{N}$

② objective function：$\mathcal{L}: \mathbb{R}^d \to \mathbb{R}$ 为 differentiable

③ 每轮迭代的 learning rate / step size：$(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$

④ initial value：$\xi \in \mathbb{R}^d$

⑤ 用于 map 迭代轮次和该轮次 input 的函数：$\Theta: \mathbb{N}_0 \to \mathbb{R}^d$

则 Θ 被称为 GD process for the objective function $\mathcal{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ，若对 $\forall\, n \in \mathbb{N}$，有

① $\Theta_0 = \xi$

② $\Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1})$

Algorithm 6.1.2: GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}, \mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}), (\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty), \xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ (cf. Definition 6.1.1)

```
Initialization: Θ ← ξ
for n = 1, ..., N do
    Θ ← Θ − γ_n(∇L)(Θ)
return Θ
```

2. Definition: GD optimization in the training of ANNs

令 ① data 的 features (input) 的维度：$d \in \mathbb{N}$

② ANN 的 hidden layers 数：$h \in \mathbb{N}$

③ ANN 的各隐藏层 neuron 数：$l_1, l_2, \cdots, l_h \in \mathbb{N}$ (input layer 维度为 d，output layer 维度为 1)

④ ANN 的总 parameters 数：$\mathfrak{d} \in \mathbb{N}$，且满足

$$\mathfrak{d} = l_1(d+1) + \left[\sum_{k=2}^{h} l_k(l_{k-1}+1)\right] + l_h + 1$$

⑤ ANN 的 activation function：$a: \mathbb{R} \to \mathbb{R}$ 为 differentiable

⑥ training data 的数量：$M \in \mathbb{N}$

⑦ training data 的 features：$x_1, x_2, \cdots, x_M \in \mathbb{R}^d$

⑧ training data 的 labels：$y_1, y_2, \cdots, y_M \in \mathbb{R}$

⑨ (MSE) loss function：$\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$，且满足对 $\forall\, \theta \in \mathbb{R}^{\mathfrak{d}}$，有

$$\mathcal{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \left| \left(\mathcal{N}^{\theta,d}_{a,l_1,a,l_2,\cdots,a,l_h,\mathrm{id}_{\mathbb{R}}}\right)(x_m) - y_m \right|^2\right]$$

⑩ 每轮迭代的 learning rate / step size：$(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$

⑪ initial value：$\xi \in \mathbb{R}^d$

⑫ 用于 map 迭代轮次和该轮次 input 的函数：$\Theta: \mathbb{N}_0 \to \mathbb{R}^d$

则 Θ 被称为 GD process for the objective function $\mathcal{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ, 若 对 $\forall\, n\in\mathbb{N}$, 有

① $\Theta_0 = \xi$

② $\Theta_n = \Theta_{n-1} - \gamma_n (\nabla\mathcal{L})(\Theta_{n-1})$

§2 GD optimization with classical momentum

1. **Definition:** Momentum GD optimization method

令 ① objective function 的 input 维度: $d\in\mathbb{N}$

② objective function: $\mathcal{L}: \mathbb{R}^d \to \mathbb{R}$ 为 differentiable

③ 每轮迭代的 learning rate / step size: $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$

④ 每轮迭代的 momentum decay factor: $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1]$

⑤ initial value: $\xi\in\mathbb{R}^d$

⑥ 用于 map 迭代轮次和该轮次 input 的函数: $\Theta: \mathbb{N}_0 \to \mathbb{R}^d$

则 Θ 被称为 GD process for the objective function $\mathcal{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ,

若 存在(每轮迭代的 momentum) $'m: \mathbb{N}_0 \to \mathbb{R}^\mathfrak{d}$, 使得对 $\forall\, n\in\mathbb{N}$, 有

① $\Theta_0 = \xi$,

$'m_0 = '0$

② $'m_n = \alpha_n\, 'm_{n-1} + (1-\alpha_n)(\nabla\mathcal{L})(\Theta_{n-1})$ (上一轮的 momentum 贡献 α 比重, 当前的负梯度贡献 $1-\alpha$ 比重)

$\Theta_n = \Theta_{n-1} - \gamma_n\, 'm_n$

Algorithm 6.3.2: Momentum GD optimization method

Input: $\mathfrak{d}, N\in\mathbb{N}, \mathscr{L}\in C^1(\mathbb{R}^\mathfrak{d},\mathbb{R}), (\gamma_n)_{n\in\mathbb{N}}\subseteq[0,\infty), (\alpha_n)_{n\in\mathbb{N}}\subseteq[0,1], \xi\in\mathbb{R}^\mathfrak{d}$

Output: N-th step of the momentum GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ (cf. Definition 6.3.1)

1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0\in\mathbb{R}^\mathfrak{d}$
2: **for** $n = 1,\dots,N$ **do**
3: $\quad \mathbf{m} \leftarrow \alpha_n\mathbf{m} + (1-\alpha_n)(\nabla\mathscr{L})(\Theta)$
4: $\quad \Theta \leftarrow \Theta - \gamma_n\mathbf{m}$
5: **return** Θ