Machine Learning in Finance

Solution sheet 3

Exercise 3.1 (Gradient descents)

(a) If the optimization problem is strictly convex, does gradient descent always converge? Does it always converge to the global minimum?

(b) If the optimization problem is convex, step length chosen sufficiently small, does gradient descent always converge? Does it always converge to the global minimum?

(c) What is the difference between batch gradient and stochastic gradient descent? And what is mini-batch gradient descent?

Solution 3.1

(a) No because of learning rate.

(b) No because global minimum might not exist, for example linear function.

Exercise 3.2 (Backpropogation of neural network) Let $\theta = (w, b, a) \in \mathbb{R}^3$ and let σ be the activation function. We consider the shallow neural network $f_\theta \colon \mathbb{R} \to \mathbb{R}$ s.t.

$$f_\theta(x) = a\sigma(wx + b). \tag{1}$$

Then we solve the regression problem with 3 data point $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, 2, 3$ by minimizing the L^2 loss

$$\mathcal{L}_f = \sum_{i=1,2,3} \big(f_\theta(x_i) - y_i\big)^2. \tag{2}$$

(a) When solving the regression, do we compute $\nabla_{x_0} \mathcal{L}_f$ or $\nabla_\theta \mathcal{L}_f$?

(b) Compute $\partial_w f$ and $\partial_b f$ by chain rule. Do you find any intermediate value computed twice in both $\partial_w f$ and $\partial_b f$?

(c) Consider regression problem as a constrained optimization problem

$$\begin{aligned} &\min \sum_{i=1,2,3} l_i \\ &l_i = (\tilde{y}_i - y_i)^2 \\ &\tilde{y}_i = a\sigma(z_i), \qquad i = 1, 2, 3. \\ &z_i = wx_i + b \end{aligned} \tag{3}$$

Solve it by Lagrange multiplier and relate this with backpropagation.

(d) Generalize this idea to deep neural networks.

Solution 3.2

(a) $\nabla_\theta f$

(b) Let $z = wx_0 + b$ then

$$\partial_w \mathcal{L}_f = \partial_z \mathcal{L}_f \cdot x_0 = \big(a\sigma(w_0 x + b) - y_0\big)\sigma' a(wx_0 + b)x_0, \tag{4}$$
$$\partial_b \mathcal{L}_f = \partial_z \mathcal{L}_f \cdot 1 = \big(a\sigma(wx_0 + b) - y_0\big)a\sigma'(wx_0 + b) \tag{5}$$

(c) Consider the Lagrangian

$$\mathcal{L} = l - \lambda_l(l - (y - y_0)^2) - \lambda_y(y - a\sigma(z)) - \lambda_z(z - (wx_0 + b)) \tag{6}$$

Compute the gradient

$$\begin{aligned}
\partial_l \mathcal{L} &= 1 - \lambda_l \\
\partial_y \mathcal{L} &= \lambda_l \frac{\partial (y - y_0)^2}{\partial y} - \lambda_y \\
\partial_z \mathcal{L} &= \lambda_y \frac{\partial a\sigma(z)}{\partial z} - \lambda_z \\
\partial_w \mathcal{L} &= \lambda_z \frac{\partial (wx_0 + b)}{\partial w} \\
\partial_b \mathcal{L} &= \lambda_z \frac{\partial (wx_0 + b)}{\partial b}
\end{aligned}$$

Let $\nabla \mathcal{L} = 0$ we get exactly the backpropagation formula.

(d) See [2].

Exercise 3.3 (Controlled ODEs) Consider the controlled ODE: $X_0 = x \in \mathbb{R}$

$$dX_t^\theta = V^\theta(t, X_t^\theta)dt, \quad t \in [0, T]. \tag{7}$$

(a) Let

$$a_t = \frac{\partial X_T^\theta}{\partial X_t^\theta}. \tag{8}$$

Prove that

$$\frac{d}{dt} a_t = -\frac{\partial V^\theta}{\partial x}(t, X_t^\theta) \cdot a_t, \quad a_T = 1, \tag{9}$$

and relate a_t with $J_{t,T}$ in the lecture notebook.

(b) Prove that

$$\frac{d}{dt}\Big(\frac{\partial X_t^\theta}{\partial \theta} a_t\Big) = a_t \frac{\partial V^\theta}{\partial \theta}(t, X_t^\theta), \tag{10}$$

and

$$\frac{\partial X_T^\theta}{\partial \theta} = -\int_T^0 \frac{\partial X_T^\theta}{\partial X_t^\theta} \cdot \frac{\partial V^\theta}{\partial \theta}(t, X_t^\theta)dt. \tag{11}$$

(c) Is every feedforward neural network a discretization of controlled ODE?

Solution 3.3

(a) By chain rule (details see ex class recording)

(b) Solve this ODE by variation of parameters

$$\frac{d}{dt}\frac{\partial X_t^\theta}{\partial \theta} = \frac{\partial V_t^\theta}{\partial \theta}(X_t^\theta) + \frac{\partial V_t^\theta}{\partial x}(X_t^\theta) \cdot \frac{\partial X_t^\theta}{\partial \theta} \tag{12}$$

References

[1] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[2] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28, 1988.