Exercise sheet 1: Around gradient descent

M2 MIAGE ID Apprentissage

October 1st, 2024

Exercise 1.1: One-layer neural network (Exam 2021-2022)

In this exercise, we consider the special case of a dataset with scalar labels/outputs, i.e. of the form $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ and $y_i \in \mathbb{R}$ for every $i = 1, \dots, n$. We build a simple neural network with no activation function and one homogeneous linear layer to predict the value y_i from the vector $\boldsymbol{x}_i$, resulting in the model

$$h^{lin}(\cdot; \boldsymbol{w}) : \begin{array}{ccc} \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R} \\ \boldsymbol{x} & \longmapsto & \boldsymbol{W}_1 \boldsymbol{x}, \end{array} \tag{1}$$

with $\boldsymbol{W}_1 \in \mathbb{R}^{1 \times d_x}$. Letting $d = d_x$ and $\boldsymbol{w} = \boldsymbol{W}_1^{\mathrm{T}} \in \mathbb{R}^d$, finding the best model amounts to solving

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}} \, f^{lin}(\boldsymbol{w}) := \frac{1}{2n} \sum_{i=1}^n (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i - y_i)^2. \tag{2}$$

a) What class of problems does problem (2) belong to?

b) The objective function f^{lin} is $\mathcal{C}_L^{1,1}$, i.e. its gradient is L-Lipschitz continuous. If L is known, how can its value be used in an algorithm such as gradient descent?

c) Problem (2) is convex with a $\mathcal{C}^1$ objective function.

   i) What can then be said about a point $\bar{\boldsymbol{w}}$ such that $\nabla f^{lin}(\bar{\boldsymbol{w}}) = \mathbf{0}_{\mathbb{R}^d}$?

   ii) What is the convergence rate of gradient descent on this problem?

   iii) What is the convergence rate of accelerated descent on a convex problem? Is it better or worse than that of the previous question ?

d) Suppose that the data is such that the objective f^{lin} is μ-strongly convex, in addition to the properties already mentioned above.

   i) Let $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$ be two points such that $\nabla f^{lin}(\boldsymbol{w}) = \nabla f^{lin}(\boldsymbol{v}) = \mathbf{0}_{\mathbb{R}^d}$. What can we say about $\boldsymbol{v}$ and $\boldsymbol{w}$?

   ii) What is the convergence rate of accelerated gradient on this problem?

Exercise 1.2: Two-layer linear neural networks (exam 2021-2022)

We consider a dataset $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ and $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$. We wish to learn a mapping from $\mathbb{R}^{d_x}$ to $\mathbb{R}^{d_y}$ that correctly outputs $\boldsymbol{y}_i$ when given $\boldsymbol{x}_i$ as an input. Our model will be that of a two-layer linear neural network :

$$\begin{array}{llll} \boldsymbol{h}(\cdot;\boldsymbol{w}): & \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ & \boldsymbol{x} & \longmapsto & \boldsymbol{W}_2(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)+\boldsymbol{b}_2, \end{array} \tag{3}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d_x \times m}$, $\boldsymbol{b}_1 \in \mathbb{R}^m$, $\boldsymbol{W}_2 \in \mathbb{R}^{m \times d_y}$ and $\boldsymbol{b}_2 \in \mathbb{R}^{d_y}$. We will consider $\boldsymbol{h}$ as being parameterized by $\boldsymbol{w} \in \mathbb{R}^d$, with $d = d_x m + m + m d_y + d_y$ and $\boldsymbol{w}$ concatenating all coefficients from $\boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{W}_2, \boldsymbol{b}_2$. Our goal is to determine a value of $\boldsymbol{w}$ so that $\boldsymbol{h}(\boldsymbol{x}_i;\boldsymbol{w}) \approx \boldsymbol{y}_i$, which we formalize using the squared loss $(\boldsymbol{h},\boldsymbol{y}) \mapsto \frac{1}{2}\|\boldsymbol{h}-\boldsymbol{y}\|^2$.

Overall, we obtain the following problem:

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}) := \frac{1}{2n}\sum_{i=1}^n \|\boldsymbol{h}(\boldsymbol{x}_i;\boldsymbol{w}) - \boldsymbol{y}_i\|^2. \tag{4}$$

It can be shown that the function f is $\mathcal{C}^1$.

a) Give a lower bound on the objective function of problem (4).

b) In general, problem (4) is nonconvex. What does this imply about its local minima?

c) Suppose that $\boldsymbol{w}^*$ is a solution of (4). What can be said about the derivative of f at $\boldsymbol{w}^*$?

d) Write down the gradient descent iteration for problem (4) with an arbitrary stepsize.

e) Given that the problem is nonconvex, what is the theoretical convergence rate of gradient descent applied to (4)?

Exercise 1.3: Matrix completion (exam 2022-2023)

Let $\boldsymbol{X} \in \mathbb{R}^{d \times d}$ be a data matrix such that only a subset of its entries $\mathcal{S} \subset \{1, \ldots, d\}^2$ are known with $|\mathcal{S}| = n \leq d^2$. We consider the problem

$$\operatorname*{minimize}_{\boldsymbol{W} \in \mathbb{R}^{d \times d}} f(\boldsymbol{W}) := \frac{1}{2n} \sum_{(i,j) \in \mathcal{S}} ([\boldsymbol{W}]_{ij} - [\boldsymbol{X}]_{ij})^2. \tag{5}$$

a) When $\mathcal{S} = \{1, \ldots, d\}^2$, justify that $\boldsymbol{W}^* = \boldsymbol{X}$ is the unique solution of the problem.

b) Problem (5) is convex in the coefficients of $\boldsymbol{W}$. Letting $\boldsymbol{w} \in \mathbb{R}^{d^2}$ denoting the column vector formed by stacking all columns of the matrix $\boldsymbol{W}$ in order, we can reformulate the problem as

$$\operatorname*{minimize}_{\boldsymbol{w} \in \mathbb{R}^{d^2}} \hat{f}(\boldsymbol{w}) := \frac{1}{2n} \sum_{(i,j) \in \mathcal{S}} ([\boldsymbol{w}]_{i+(j-1)d} - [\boldsymbol{X}]_{ij})^2. \tag{6}$$

The function $\hat{f}$ is convex and $\mathcal{C}^1$.

i) What convergence rate guarantee can we provide on gradient descent when applied to problem (6)? What quantity does this rate apply to?

ii) What is the corresponding convergence rate for the accelerated gradient method due to Nesterov? Is it better than that of gradient descent?

iii) When $n = d^2$, the function $\hat{f}$ is a strongly convex quadratic function. Aside from Nesterov's method, what other approach can we use to obtain better convergence rates than gradient descent?

c) We now suppose that the data matrix $\boldsymbol{X}$ is symmetric, positive semidefinite and of rank $1 \ll d$. In this setting, rather than seeking an arbitrary matrix $\boldsymbol{W}$ to approximate $\boldsymbol{X}$, we can force the matrix to be rank one by writing it $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ where $\boldsymbol{u} \in \mathbb{R}^d$. Problem (5) then becomes

$$\operatorname*{minimize}_{\boldsymbol{u} \in \mathbb{R}^{d}} \tilde{f}(\boldsymbol{u}) := \frac{1}{2n} \sum_{(i,j) \in \mathcal{S}} ([\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}]_{ij} - [\boldsymbol{X}]_{ij})^2. \tag{7}$$

The objective function of problem (7) is $\mathcal{C}^2$ and nonconvex.

i) State the first-order necessary optimality conditions for problem (7).

ii) What is the convergence rate of gradient descent for this problem? What quantity does this rate apply to?

iii) Under certain assumptions on $\boldsymbol{X}$ and $\mathcal{S}$, one can show that all the local minima of this problem are global. In that case, what technique guarantees almost surely that gradient descent will converge to such a point?

Solutions

Solutions to Exercise 1.1

a) The function $f(\boldsymbol{W})$ is always nonnegative (as a sum of squares, i.e. nonnegative numbers). When $n = d^2$, we have that

$$f(\boldsymbol{W}) = 0 \quad \Leftrightarrow \quad ([\boldsymbol{W}]_{ij} - [\boldsymbol{X}]_{ij})^2 = 0 \ \forall (i,j) \in \{1, \dots, d\}^2 \quad \Leftrightarrow \quad \boldsymbol{W} = \boldsymbol{X}.$$

As a result, the problem has a single global minimum given by $\boldsymbol{W}^* = \boldsymbol{X}$.

b) Convex formulation

i) Since the problem is convex, we know that after $K \geq 1$ iterations of gradient descent, the iterate $\boldsymbol{w}_K$ satisfies

$$\hat{f}(\boldsymbol{w}_K) - \min_{\boldsymbol{w} \in \mathbb{R}^{d^2}} \hat{f}(\boldsymbol{w}) \leq \mathcal{O}\left(\frac{1}{K}\right).$$

Gradient descent thus converges at a rate $\frac{1}{K}$.

ii) The rate for accelerated gradient on such a problem is $\frac{1}{K^2}$, which is a better rate as it converges more quickly to 0.

iii) When $\hat{f}$ is a strongly convex quadratic function, the heavy-ball method (aka Polyak's method) attains the optimal rate of convergence for strongly convex functions, which is better than gradient descent. *NB: The value of that rate is not required to answer the question.*

c) (Nonconvex case)

i) If $\bar{\boldsymbol{u}} \in \mathbb{R}^d$ is a local minima of problem (7), then $\nabla \tilde{f}(\bar{\boldsymbol{u}}) = \mathbf{0}$.

ii) For this problem, after $K \geq 1$ iterations of gradient descent, we have

$$\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right),$$

hence the convergence rate of gradient descent is in $\frac{1}{\sqrt{K}}$.

iii) Initializing gradient descent with a random point guarantees almost surely that it will converge to a local minima under the assumptions of this question.

Solutions to Exercise 1.2

a) The value 0 is a lower bound on this objective function, since it is always nonnegative. Any value less than or equal to 0 also works.

b) The local minima of a nonconvex problem are not necessarily global minima.

c) By the first-order necessary conditions, if $\boldsymbol{w}^*$ is a solution of (4), then its gradient is zero, that is $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$.

d) Using an arbitrary stepsize $\alpha_k > 0$, the kth iteration of gradient descent can be written as

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k).$$

e) For a nonconvex problem such as (4), it can be guaranteed that, after $K \geq 1$ iterations of gradient descent, one has

$$\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\| \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

Solutions to Exercise 1.3

a) Problem (2) is a linear least-squares problem.

b) If a Lipschitz constant L for the gradient is known, the stepsize can be chosen as the constant value $\alpha = \frac{1}{L}$. *NB: Other values less than $\frac{2}{L}$ would also guarantee decrease of the function value at every iteration.*

c)

i) Since the problem is convex, any point $\bar{\boldsymbol{w}}$ such that $\nabla f^{lin}(\bar{\boldsymbol{w}}) = \mathbf{0}_{\mathbb{R}^d}$ is a global minimum.

ii) On such a convex problem, after $K \geq 1$ iterations of gradient descent, one obtains that

$$f(\boldsymbol{w}_k) - \min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) \leq \mathcal{O}\left(\frac{1}{K}\right).$$

iii) On a convex problem, after $K \geq 1$ iterations of accelerated gradient, one obtains that

$$f(\boldsymbol{w}_k) - \min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) \leq \mathcal{O}\left(\frac{1}{K^2}\right),$$

which is better than the rate for gradient descent since it converges more rapidly towards 0.

d)

i) Since the function is strongly convex and continuously differentiable, it has a unique global minimum, which is the unique solution of the equation $\nabla f^{lin}(\boldsymbol{w}) = \mathbf{0}_{\mathbb{R}^d}$. Therefore, if $\boldsymbol{w}$ and $\boldsymbol{v}$ satisfy $\nabla f^{lin}(\boldsymbol{w}) = \nabla f^{lin}(\boldsymbol{v}) = \mathbf{0}_{\mathbb{R}^d}$, then we must have $\boldsymbol{v} = \boldsymbol{w}$.

ii) On a strongly convex problem, after $K \geq 1$ iterations of accelerated gradient, one obtains that

$$f(\boldsymbol{w}_k) - \min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) \leq \mathcal{O}\left((1 - \sqrt{\mu} L)^K\right).$$