Gradient Descent Method

Backpropagation

Derivative of Loss function w.r.t. Vectors and Matrices

Sunmook Choi

Dept. of Mathematics
Korea University

Gradients and Jacobian Matrix

For real-valued function $f\colon \mathbb{R}^n \to \mathbb{R}$,

- $\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)$ where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ ($\boldsymbol{x}$ is a row vector).

For vector-valued function $f\colon \mathbb{R}^n \to \mathbb{R}^m$,

- Let $\boldsymbol{y} = f(\boldsymbol{x})$, then

$$
\begin{aligned}
\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} &= \left[\frac{\partial y_i}{\partial x_j}\right]_{ij} \in \mathbb{R}^{m\times n} \\
&= \begin{bmatrix}
\frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\
\frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n}
\end{bmatrix}
\end{aligned}
$$

Formulas

Let $A = [a_{ij}] \in \mathbb{R}^{n\times m}$ be a matrix and $\boldsymbol{b} = (b_1, b_2, \ldots, b_m) \in \mathbb{R}^m$.

- For $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, let $\boldsymbol{y} = \boldsymbol{x}A + b$, then we have

$$\boldsymbol{y} = \begin{bmatrix} \sum_{i=1}^n a_{i1}x_i + b_1 & \sum_{i=1}^n a_{i2}x_i + b_2 & \cdots & \sum_{i=1}^n a_{im}x_i + b_m \end{bmatrix},$$

  so we have the following Jacobian matrices.

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \left[\frac{\partial y_i}{\partial x_j}\right]_{ij} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} = A^\top \quad \text{and} \quad \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}} = \left[\frac{\partial y_i}{\partial b_j}\right]_{ij} = I_m.$$

Formulas

Let $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ be a square matrix.

- For $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, let $y = \boldsymbol{x}A\boldsymbol{x}^\top \in \mathbb{R}$, then we have

$$y = \begin{bmatrix} \sum_{i=1}^n a_{i1}x_i & \sum_{i=1}^n a_{i2}x_i & \cdots & \sum_{i=1}^n a_{in}x_i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j.$$

  Then the gradient is $\frac{\partial y}{\partial \boldsymbol{x}} = \boldsymbol{x}(A + A^\top)$.

- The gradient of y w.r.t. A is $\dfrac{\partial y}{\partial A} = \boldsymbol{x}^\top \boldsymbol{x} \in \mathbb{R}^{n\times n}$, whose dimension is equal to that of A.

  Or we can compute gradients w.r.t. each row (column) vector of A and concatenate them.

Linear Regression

Consider a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, and we assume that the data is linear, that is, there exists $\boldsymbol{w} \in \mathbb{R}^d$ such that

$$y_i = \boldsymbol{w}\boldsymbol{x}_i^\top = \sum_{j=1}^{d} w_j \cdot x_{ij} \quad \text{and} \quad x_{i1} = 1.$$

- $X = \begin{bmatrix}\boldsymbol{x}_1^\top \ \boldsymbol{x}_2^\top & \cdots & \boldsymbol{x}_n^\top\end{bmatrix}^\top \in \mathbb{R}^{n\times d}$: *the design matrix* of $\mathcal{D}$.
- $\boldsymbol{y} = [y_1 \ y_2 \ \cdots \ y_n] \in \mathbb{R}^n$: the (row) vector containing the labels.

Then we want to find $\boldsymbol{w} \in \mathbb{R}^d$ that satisfies

$$\boldsymbol{y} = \boldsymbol{w}X^\top,$$

so we will try to minimize the mean squared loss function: (Ordinary Least Squares)

$$L(\boldsymbol{w}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}\boldsymbol{x}_i^\top)^2 = \frac{1}{2n}(\boldsymbol{y} - \boldsymbol{w}X^\top)(\boldsymbol{y} - \boldsymbol{w}X^\top)^\top.$$

Analytic Solution for Linear Regression

MSE loss of Linear Regression:

$$L(\boldsymbol{w}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}\boldsymbol{x}_i^\top)^2 = \frac{1}{2n}(\boldsymbol{y} - \boldsymbol{w}X^\top)(\boldsymbol{y} - \boldsymbol{w}X^\top)^\top.$$

Since the loss is convex w.r.t. $\boldsymbol{w}$, the existence of minimum is guaranteed.
Observe that the solution is $\boldsymbol{w}$ that satisfies $\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = 0$.

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) &= \frac{1}{2n}\frac{\partial}{\partial \boldsymbol{w}}(\boldsymbol{y}\boldsymbol{y}^\top - \boldsymbol{y}X\boldsymbol{w}^\top - \boldsymbol{w}X^\top\boldsymbol{y}^\top + \boldsymbol{w}X^\top X\boldsymbol{w}^\top) \\
&= \frac{1}{2n}\frac{\partial}{\partial \boldsymbol{w}}(\boldsymbol{y}\boldsymbol{y}^\top - 2\boldsymbol{w}X^\top\boldsymbol{y}^\top + \boldsymbol{w}X^\top X\boldsymbol{w}^\top) \\
&= \frac{1}{n}\left(-\boldsymbol{y}X + \boldsymbol{w}(X^\top X)\right) = 0
\end{aligned}$$

Hence, if $X^\top X$ is invertible, then we have an optimal solution: $\hat{\boldsymbol{w}} = \boldsymbol{y}X(X^\top X)^{-1}$.

Computing Gradients in Classification

Consider 2-Layer MLP with dataset $\mathcal{D} = \{(\boldsymbol{x}^n, y^n)\}$.

- Input dim: d, Hidden dim: h, Output dim: K (# of classes)

$$\boldsymbol{x}^n \mapsto \boldsymbol{z}^{n,1} = \boldsymbol{x}^n W^1 + \boldsymbol{b}^1 \mapsto \boldsymbol{a}^n = \sigma(\boldsymbol{z}^{n,1}) \mapsto \boldsymbol{z}^{n,2} = \boldsymbol{a}^n W^2 + \boldsymbol{b}^2 \mapsto \hat{\boldsymbol{y}}^n = Softmax\,(\boldsymbol{z}^{n,2})$$

- Categorical Cross Entropy loss: $\theta = (W^1, \boldsymbol{b}^1, W^2, \boldsymbol{b}^2)$

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{K} \left[-y_j^n \log(\hat{y}_j^n) \right]$$

Using one-hot-encoding, if the data x^n belongs to j_*th class, then

$$y_j^n = \begin{cases} 1 & \text{if } j = j_* \\ 0 & \text{if } j \neq j_* \end{cases}.$$

Gradients are..

We want to find the following gradients: $\dfrac{\partial L}{\partial W^1}, \quad \dfrac{\partial L}{\partial \boldsymbol{b}^1}, \quad \dfrac{\partial L}{\partial W^2}, \quad \dfrac{\partial L}{\partial \boldsymbol{b}^2}.$

Using the chain rule, we find that

$$\frac{\partial L}{\partial W^2} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial W^2}, \qquad \frac{\partial L}{\partial \boldsymbol{b}^2} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial \boldsymbol{b}^2},$$

$$\frac{\partial L}{\partial W^1} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,1}} \frac{\partial \boldsymbol{z}^{n,1}}{\partial W^1}, \qquad \frac{\partial L}{\partial \boldsymbol{b}^1} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,1}} \frac{\partial \boldsymbol{z}^{n,1}}{\partial \boldsymbol{b}^1}$$

where $\dfrac{\partial L}{\partial \boldsymbol{z}^{n,2}} = \dfrac{\partial L}{\partial \hat{\boldsymbol{y}}^n} \dfrac{\partial \hat{\boldsymbol{y}}^n}{\partial \boldsymbol{z}^{n,2}}$ and $\dfrac{\partial L}{\partial \boldsymbol{z}^{n,1}} = \dfrac{\partial L}{\partial \boldsymbol{z}^{n,2}} \dfrac{\partial \boldsymbol{z}^{n,2}}{\partial \boldsymbol{a}^n} \dfrac{\partial \boldsymbol{a}^n}{\partial \boldsymbol{z}^{n,1}}.$

First we compute $\dfrac{\partial L}{\partial \boldsymbol{z}^{n,2}} = \dfrac{\partial L}{\partial \hat{\boldsymbol{y}}^n} \dfrac{\partial \hat{\boldsymbol{y}}^n}{\partial \boldsymbol{z}^{n,2}}$.

$$\frac{\partial L}{\partial \hat{\boldsymbol{y}}^n} = \frac{1}{N} \begin{bmatrix} -y_1^n/\hat{y}_1^n & -y_2^n/\hat{y}_2^n & \cdots & -y_K^n/\hat{y}_K^n \end{bmatrix}$$

$$\frac{\partial \hat{\boldsymbol{y}}^n}{\partial \boldsymbol{z}^{n,2}} = \left[\frac{\partial \hat{y}_i^n}{\partial z_j^{n,2}} \right]_{ij}, \text{ where } \quad \frac{\partial \hat{y}_i^n}{\partial z_j^{n,2}} = \begin{cases} \hat{y}_i^n (1 - \hat{y}_i^n) & \text{if } i = j \\ -\hat{y}_i^n \hat{y}_j^n & \text{if } i \neq j \end{cases}.$$

Then, using the fact that $\sum_{j=1}^K y_j^n = 1$ for each n, we obtain $\frac{\partial L}{\partial \hat{\boldsymbol{y}}^n} \frac{\partial \hat{\boldsymbol{y}}^n}{\partial \boldsymbol{z}^{n,2}} = \frac{1}{N}(\hat{\boldsymbol{y}}^n - \boldsymbol{y}^n)$.

$$\therefore \frac{\partial L}{\partial W^2} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial W^2} = \frac{1}{N} \sum_n (\hat{\boldsymbol{y}}^n - \boldsymbol{y}^n) \frac{\partial (\boldsymbol{a}^n W^2 + \boldsymbol{b}^2)}{\partial W^2} = \frac{1}{N} \sum_n \boldsymbol{a}^{n\top} (\hat{\boldsymbol{y}}^n - \boldsymbol{y}^n)$$

$$\frac{\partial L}{\partial \boldsymbol{b}^2} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial \boldsymbol{b}^2} = \frac{1}{N} \sum_n (\hat{\boldsymbol{y}}^n - \boldsymbol{y}^n)$$

Likewise, we compute $\frac{\partial L}{\partial \boldsymbol{z}^{n,1}} = \frac{\partial L}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial \boldsymbol{a}^n} \frac{\partial \boldsymbol{a}^n}{\partial \boldsymbol{z}^{n,1}}$.

$$\frac{\partial \boldsymbol{z}^{n,2}}{\partial \boldsymbol{a}^n} = W^{2\top} \in \mathbb{R}^{K\times h}, \quad \frac{\partial \boldsymbol{a}^n}{\partial \boldsymbol{z}^{n,1}} = \mathrm{diag}[a_i^n(1-a_i^n)] \in \mathbb{R}^{h\times h}$$

$$\therefore \frac{\partial L}{\partial \boldsymbol{a}^n} = \frac{\partial L}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial \boldsymbol{a}^n} = \frac{\partial L}{\partial \boldsymbol{z}^{n,2}} W^{2\top} \Rightarrow \frac{\partial L}{\partial \boldsymbol{z}^{n,1}} = \frac{\partial L}{\partial \boldsymbol{a}^n} \frac{\partial \boldsymbol{a}^n}{\partial \boldsymbol{z}^{n,1}} = \left[\tfrac{\partial L}{\partial a_1^n} a_1^n(1-a_1^n) \quad \cdots \quad \tfrac{\partial L}{\partial a_h^n} a_h^n(1-a_h^n) \right]$$

$$\therefore \frac{\partial L}{\partial W^1} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,1}} \frac{\partial \boldsymbol{z}^{n,1}}{\partial W^1} = \sum_n x^{n\top} \frac{\partial L}{\partial \boldsymbol{z}^{n,1}}$$

$$\frac{\partial L}{\partial \boldsymbol{b}^1} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,1}} \frac{\partial \boldsymbol{z}^{n,1}}{\partial \boldsymbol{b}^1} = \sum_n \frac{\partial L}{\partial \boldsymbol{z}^{n,1}}$$