Gradient Descent Method

Backpropagation

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

Sunmook Choi

Dept. of Mathematics
Korea University

Settings

- For practical reasons, we express each data $\boldsymbol{x}$ into a row vector.
- There are different layouts to express derivatives with respect to vectors or matrices.
- In this presentation, we use the 'denominator layout'.
  For more information, you can see more details in this Wikipedia page.

Linear Regression

Gradients

For real-valued function $f : \mathbb{R}^n \to \mathbb{R}$, the derivative of f is the gradient of f

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}$$

where $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \in \mathbb{R}^n$ ($\boldsymbol{x}$ is a row vector).

Examples

- Let $\boldsymbol{a} \in \mathbb{R}^n$ and $f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{a}^\top = \sum_{i=1}^n a_i x_i$, then the gradient is $\nabla_{\boldsymbol{x}} f = \boldsymbol{a}$.
- Let $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ and $f(\boldsymbol{x}) = \boldsymbol{x}A\boldsymbol{x}^\top$, then the gradient is $\nabla_{\boldsymbol{x}} f = \boldsymbol{x}(A + A^\top)$.
  - Why? → Homework
  - Hint) $f(\boldsymbol{x}) = \boldsymbol{x}A\boldsymbol{x}^\top = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j$.

Linear Regression

Consider a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{x}_i = [x_{i1}\ x_{i2}\ \ldots\ x_{id}] \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
We assume that the relation is linear, that is, there exists $\boldsymbol{w} = [w_1\ \ldots\ w_d] \in \mathbb{R}^d$ such that

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

- $X = \left[\boldsymbol{x}_1^\top\ \boldsymbol{x}_2^\top\ \cdots\ \boldsymbol{x}_n^\top\right]^\top \in \mathbb{R}^{n\times d}$: *the design matrix* of $\mathcal{D}$ (each row is an input vector).
- $\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_n] \in \mathbb{R}^n$: 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 $L(\boldsymbol{w})$ is convex w.r.t. $\boldsymbol{w}$, the minimum is the point where $\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = 0$ (gradient!).

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

Analytic Solution for Linear Regression

MSE loss of Linear Regression:

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

and

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = \frac{1}{n}\Big(-\boldsymbol{y}X + \boldsymbol{w}(X^\top X)\Big) = 0$$

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

- The rank of $X^\top X$ is equal to the rank of X.
- Hence, $X^\top X \in \mathbb{R}^{d\times d}$ is invertible if and only if $X \in \mathbb{R}^{n\times d}$ has a full rank of d, that is, when each column vector (feature) of X is linearly independent of each other ($\because d \ll n$).

Numerical Solution for Linear Regression

MSE loss of Linear Regression:

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

and

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = \frac{1}{n} \left(-\boldsymbol{y}X + \boldsymbol{w}X^\top X \right) = \frac{1}{n} \left(\boldsymbol{w}X^\top - \boldsymbol{y} \right) X$$

so that

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \nabla_{\boldsymbol{w}} L(\boldsymbol{w}^t) = \boldsymbol{w}^t - \frac{\eta}{n} \left(\boldsymbol{w}X^\top - \boldsymbol{y} \right) X$$

or equivalently, for $j = 1, \ldots, d$,

$$w_j^{t+1} = w_j^t - \frac{\eta}{n} \sum_{i=1}^{n} (\boldsymbol{w}\boldsymbol{x}_i^\top - y_i)\, x_{ij}.$$

MLP Classification

Jacobian Matrix

For vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$, the derivative of $\boldsymbol{y} = f(\boldsymbol{x})$ is the Jacobian matrix

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

Example

- Let $A = [a_{ij}] \in \mathbb{R}^{n\times m}$, and $\boldsymbol{y} = \boldsymbol{x}A$. Then the Jacobian matrix is

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \left[\frac{\partial y_i}{\partial x_j}\right]_{ij} = A^\top.$$

Derivative w.r.t. matrices

Let $X = [x_{ij}] \in \mathbb{R}^{p\times q}$ and y be a real-valued function of X of independent variables.

- Here, independent variables of X means the matrix X has no special structure, e.g., not symmetric nor positive definite, etc.
- Then, the derivative of y with respect to X is given by

$$\frac{\partial y}{\partial X} = \left[\frac{\partial y}{\partial x_{ij}}\right]_{ij} \in \mathbb{R}^{p\times q}.$$

- Example: Define $f(X) = \boldsymbol{a}X\boldsymbol{b}^\top$ for $\boldsymbol{a} \in \mathbb{R}^{1\times p}$ and $\boldsymbol{b} \in \mathbb{R}^{1\times q}$, we have

$$\frac{\partial f}{\partial X} = \left[\frac{\partial f}{\partial x_{ij}}\right]_{ij} = \boldsymbol{a}^\top \boldsymbol{b} \in \mathbb{R}^{p\times q}.$$

Derivative w.r.t. matrices

Let $\boldsymbol{y} = \boldsymbol{x}W + \boldsymbol{b}$ where $W \in \mathbb{R}^{n\times m}$, $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{b} \in \mathbb{R}^m$.

- Then, the derivative $\dfrac{\partial \boldsymbol{y}}{\partial W}$ of $\boldsymbol{y}$ with respect to W should be three dimensional!
- If L is a real-valued function of $\boldsymbol{y}$, then the derivative of L w.r.t. W is

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial W} = \frac{\partial L}{\partial \boldsymbol{y}} \frac{\partial(\boldsymbol{x}W + \boldsymbol{b})}{\partial W} = \boldsymbol{x}^\top \frac{\partial L}{\partial \boldsymbol{y}}$$

  by the chain rule.

  * $\boldsymbol{x}^\top \in \mathbb{R}^{n\times 1}$, $\dfrac{\partial L}{\partial \boldsymbol{y}} \in \mathbb{R}^{1\times m}$ so that $\dfrac{\partial L}{\partial W} \in \mathbb{R}^{n\times m}$.

MLP Classification Forward Pass

Let $\mathcal{D} = \{(\boldsymbol{x}^n, y^n)\}_{n=1}^N$ be a dataset of K classes. Consider 2-layer Multi-layer Perceptron.

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

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

- Here, the parameteres are $W^1 \in \mathbb{R}^{d \times h}$, $\boldsymbol{b}^1 \in \mathbb{R}^h$, $W^2 \in \mathbb{R}^{h \times K}$, $\boldsymbol{b}^2 \in \mathbb{R}^K$.
- The activation function $\sigma(\cdot)$ is the sigmoid function, that is, $\sigma(z) = \dfrac{1}{1 + e^{-z}}$.
- Notice that $\dfrac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z))$.

MLP Classification Loss

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

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

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

- We use one-hot encoding so that if the data $\boldsymbol{x}^n$ belongs to j_*th class, then

$$y^n = [y_1^n \ \ y_2^n \ \ \cdots \ \ y_K^n] \quad \text{where} \quad y_j^n = \begin{cases} 1 & \text{if } j = j_* \\ 0 & \text{if } j \neq j_* \end{cases}.$$

- Notice that the summation over j is actually not a summation due to one-hot encoding.

Then gradients are...

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

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 first find that

$$\frac{\partial L}{\partial W^2} = \sum_n \frac{\partial L}{\partial \hat{\boldsymbol{y}}^n} \frac{\partial \hat{\boldsymbol{y}}^n}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial W^2} \quad \text{and} \quad \frac{\partial L}{\partial \boldsymbol{b}^2} = \sum_n \frac{\partial L}{\partial \hat{\boldsymbol{y}}^n} \frac{\partial \hat{\boldsymbol{y}}^n}{\partial \boldsymbol{z}^{n,2}} \frac{\partial \boldsymbol{z}^{n,2}}{\partial \boldsymbol{b}^2}.$$

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

For $L(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{K}\left[-y_j^n \log(\hat{y}_j^n)\right]$, we have $\frac{\partial L}{\partial \hat{\boldsymbol{y}}^n} = \frac{1}{N}\left[-y_1^n/\hat{y}_1^n \quad -y_2^n/\hat{y}_2^n \quad \cdots \quad -y_K^n/\hat{y}_K^n\right]$ and

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

Using the fact that $\boldsymbol{y}^n$ is one-hot encoded, we have $\frac{\partial L}{\partial \boldsymbol{z}^{n,2}} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}^n}\frac{\partial \hat{\boldsymbol{y}}^n}{\partial \boldsymbol{z}^{n,2}} = \frac{1}{N}(\hat{\boldsymbol{y}}^n - \boldsymbol{y}^n) \in \mathbb{R}^{1\times K}$.

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

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

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

We now compute $\dfrac{\partial L}{\partial W^1} = \sum_n \dfrac{\partial L}{\partial z^{n,1}} \dfrac{\partial z^{n,1}}{\partial W^1}, \quad \dfrac{\partial L}{\partial \boldsymbol{b}^1} = \sum_n \dfrac{\partial L}{\partial z^{n,1}} \dfrac{\partial z^{n,1}}{\partial \boldsymbol{b}^1}.$

We have $\dfrac{\partial L}{\partial \boldsymbol{a}^n} = \dfrac{\partial L}{\partial z^{n,2}} \dfrac{\partial z^{n,2}}{\partial \boldsymbol{a}^n} = \dfrac{\partial L}{\partial z^{n,2}}(W^2)^\top, \quad \dfrac{\partial L}{\partial z^{n,1}} = \dfrac{\partial L}{\partial \boldsymbol{a}^n} \dfrac{\partial \boldsymbol{a}^n}{\partial z^{n,1}},$ and $\dfrac{\partial \boldsymbol{a}^n}{\partial z^{n,1}} = \mathrm{diag}[a_i^n(1 - a_i^n)].$

Then $\dfrac{\partial L}{\partial z^{n,1}} = \dfrac{\partial L}{\partial z^{n,2}} (W^2)^\top \,\mathrm{diag}[a_i^n(1 - a_i^n)] \in \mathbb{R}^{1\times h}.$

$$\therefore \frac{\partial L}{\partial W^1} = \sum_n \frac{\partial L}{\partial z^{n,1}} \frac{\partial z^{n,1}}{\partial W^1} = \sum_n (\boldsymbol{x}^n)^\top \frac{\partial L}{\partial z^{n,1}} = \sum_n (\boldsymbol{x}^n)^\top \frac{\partial L}{\partial z^{n,2}} (W^2)^\top \,\mathrm{diag}[a_i^n(1 - a_i^n)] \in \mathbb{R}^{d\times h}$$

$$\frac{\partial L}{\partial \boldsymbol{b}^1} = \sum_n \frac{\partial L}{\partial z^{n,1}} \frac{\partial z^{n,1}}{\partial \boldsymbol{b}^1} = \sum_n \frac{\partial L}{\partial z^{n,1}} = \sum_n \frac{\partial L}{\partial z^{n,2}} (W^2)^\top \,\mathrm{diag}[a_i^n(1 - a_i^n)].$$

Gradient Descent Methods

Finally, we apply the gradient descent method to update the parameters.

- Recall that $W^1 \in \mathbb{R}^{d\times h}$, $\boldsymbol{b}^1 \in \mathbb{R}^{1\times h}$, $W^2 \in \mathbb{R}^{h\times K}$, $\boldsymbol{b}^2 \in \mathbb{R}^{1\times K}$ and

$$\frac{\partial L}{\partial W^1} \in \mathbb{R}^{d\times h}, \quad \frac{\partial L}{\partial \boldsymbol{b}^1} \in \mathbb{R}^{1\times h}, \quad \frac{\partial L}{\partial W^2} \in \mathbb{R}^{h\times K}, \quad \frac{\partial L}{\partial \boldsymbol{b}^2} \in \mathbb{R}^{1\times K}.$$

- Hence, the update equations should be

$$W^1 \leftarrow W^1 - \eta \frac{\partial L}{\partial W^1}, \quad \boldsymbol{b}^1 \leftarrow \boldsymbol{b}^1 - \eta \frac{\partial L}{\partial \boldsymbol{b}^1}$$

$$W^2 \leftarrow W^2 - \eta \frac{\partial L}{\partial W^2}, \quad \boldsymbol{b}^2 \leftarrow \boldsymbol{b}^2 - \eta \frac{\partial L}{\partial \boldsymbol{b}^2}$$