Recitation Notes: Gradient Descent — Theory and Key Points

CPSC-381/581: Introduction to Machine Learning

1 Gradient Descent for Linear Regression

1.1 Loss Function and Gradient Derivation

We define the mean squared error (MSE) loss as:

$$\ell(w_0, w_1) \;=\; \frac{1}{N}\sum_{n=1}^{N}\Big(\mathbf{w}^T\mathbf{x}^{(n)} - y^{(n)}\Big)^2,$$

with

$$\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \quad \text{and} \quad \mathbf{x}^{(n)} = \begin{pmatrix} x_0^{(n)} \\ x_1^{(n)} \end{pmatrix}.$$

Often, to remove the constant factor in the derivative, we define the loss as:

$$\ell(\mathbf{w}) \;=\; \frac{1}{2N}\sum_{n=1}^{N}\Big(\mathbf{w}^T\mathbf{x}^{(n)} - y^{(n)}\Big)^2.$$

Gradient Derivation (Unscaled Loss)

Taking the derivative with respect to w_0:

$$\begin{aligned}
\frac{\partial}{\partial w_0}\,\ell(w_0, w_1) &= \frac{\partial}{\partial w_0}\left[\frac{1}{N}\sum_{n=1}^{N}\Big(w_0\,x_0^{(n)} + w_1\,x_1^{(n)} - y^{(n)}\Big)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N} 2\Big(w_0\,x_0^{(n)} + w_1\,x_1^{(n)} - y^{(n)}\Big)\frac{\partial}{\partial w_0}\Big(w_0\,x_0^{(n)} + w_1\,x_1^{(n)}\Big) \\
&= \frac{2}{N}\sum_{n=1}^{N}\Big(w_0\,x_0^{(n)} + w_1\,x_1^{(n)} - y^{(n)}\Big)x_0^{(n)}.
\end{aligned}$$

Similarly, for w_1:

$$\frac{\partial}{\partial w_1}\,\ell(w_0, w_1) \;=\; \frac{2}{N}\sum_{n=1}^{N}\Big(w_0\,x_0^{(n)} + w_1\,x_1^{(n)} - y^{(n)}\Big)x_1^{(n)}.$$

In vector form, the gradient is:

$$\nabla\ell(\mathbf{w}) \;=\; \frac{2}{N}\sum_{n=1}^{N}\Big(\mathbf{w}^T\mathbf{x}^{(n)} - y^{(n)}\Big)\mathbf{x}^{(n)}.$$

When the loss is defined as $\ell(\mathbf{w}) = \frac{1}{2N}\sum(\cdots)^2$, the gradient simplifies to:

$$\nabla\ell(\mathbf{w}) \;=\; \frac{1}{N}\sum_{n=1}^{N}\Big(\mathbf{w}^T\mathbf{x}^{(n)} - y^{(n)}\Big)\mathbf{x}^{(n)}.$$

1.2 Gradient Descent Update Rule

The standard update rule is:

$$\mathbf{w}^{(t+1)} \;=\; \mathbf{w}^{(t)} - \alpha\,\nabla\ell\big(\mathbf{w}^{(t)}\big).$$

For the unscaled loss, the update becomes:

$$\mathbf{w}^{(t+1)} \;=\; \mathbf{w}^{(t)} - \alpha\,\frac{2}{N}\sum_{n=1}^{N}\Big(\mathbf{w}^{(t)T}\mathbf{x}^{(n)} - y^{(n)}\Big)\mathbf{x}^{(n)}.$$

2 Gradient Descent for Logistic Regression

2.1 Loss Function and Gradient Derivation

The logistic regression loss (negative log-likelihood) is defined as:

$$\ell(\mathbf{w}) \;=\; \frac{1}{N}\sum_{n=1}^{N}\log\Bigl(1+\exp\bigl(-y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}\bigr)\Bigr),$$

where $y^{(n)} \in \{-1,+1\}$.

Gradient Derivation (Sketch)

Using the chain rule, the partial derivative with respect to w_i is:

$$\frac{\partial}{\partial w_i}\,\ell(\mathbf{w}) \;=\; \frac{1}{N}\sum_{n=1}^{N}\frac{\exp\bigl(-y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}\bigr)(-y^{(n)})x_i^{(n)}}{1+\exp\bigl(-y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}\bigr)}.$$

This can be rewritten as:

$$\frac{\partial}{\partial w_i}\,\ell(\mathbf{w}) \;=\; \frac{1}{N}\sum_{n=1}^{N}\frac{y^{(n)}\,x_i^{(n)}}{1+\exp\bigl(y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}\bigr)}.$$

In vector form, the gradient is:

$$\nabla\ell(\mathbf{w}) \;=\; \frac{1}{N}\sum_{n=1}^{N}\frac{y^{(n)}\,\mathbf{x}^{(n)}}{1+\exp\bigl(y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}\bigr)}.$$

2.2 Gradient Descent Update Rule

The update rule for logistic regression is:

$$\mathbf{w}^{(t+1)} \;=\; \mathbf{w}^{(t)} - \alpha\,\nabla\ell\bigl(\mathbf{w}^{(t)}\bigr),$$

or explicitly,

$$\mathbf{w}^{(t+1)} \;=\; \mathbf{w}^{(t)} - \alpha\,\frac{1}{N}\sum_{n=1}^{N}\frac{y^{(n)}\,\mathbf{x}^{(n)}}{1+\exp\bigl(y^{(n)}\,\mathbf{w}^{(t)T}\mathbf{x}^{(n)}\bigr)}.$$

3 Learning Rate and Backtracking Line Search

3.1 General Update Rule Reminder

Recall the gradient descent update:

$$\mathbf{w}^{(t+1)} \;=\; \mathbf{w}^{(t)} - \alpha \,\nabla \ell\big(\mathbf{w}^{(t)}\big).$$

The learning rate α determines the step size.

3.2 Backtracking Line Search Algorithm

1. **Initialization:** Set an initial step size α_0.
2. **Candidate Update:** Compute
$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha \,\nabla \ell\big(\mathbf{w}^{(t)}\big).$$
3. **Armijo–Goldstein Condition:** Check whether
$$\ell\Big(\mathbf{w}^{(t)} - \alpha \,\nabla \ell\big(\mathbf{w}^{(t)}\big)\Big) \;\leq\; \ell\big(\mathbf{w}^{(t)}\big) - \beta \,\alpha \,\|\nabla \ell\big(\mathbf{w}^{(t)}\big)\|^2,$$
where $\beta \in (0, 1)$ is a constant.
4. **Adjust α:** If the condition is not satisfied, reduce α (e.g., set $\alpha \leftarrow \frac{1}{2}\alpha$) and repeat step 3.

Lemma 3.1 (Sufficient Decrease)**.** *Under standard smoothness assumptions on $\ell(\mathbf{w})$, there exists an $\alpha > 0$ such that the Armijo–Goldstein condition holds.*

Proof Sketch. Using the Taylor expansion around $\mathbf{w}^{(t)}$:

$$\ell\big(\mathbf{w}^{(t)} - \alpha \,\nabla \ell(\mathbf{w}^{(t)})\big) \approx \ell\big(\mathbf{w}^{(t)}\big) - \alpha \,\|\nabla \ell(\mathbf{w}^{(t)})\|^2 + \frac{L\alpha^2}{2} \|\nabla \ell(\mathbf{w}^{(t)})\|^2,$$

where L is the Lipschitz constant for $\nabla \ell$. For sufficiently small α, the quadratic term is dominated by the linear term, ensuring the condition holds if

$$\frac{L\alpha}{2} \leq \beta.$$

□

Key Points Summary

- **Gradient Descent Update Rule:**

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha \nabla \ell(\mathbf{w}^{(t)}),$$

- **Gradient Derivation:** For both models, the gradient is derived using the chain rule. For linear regression, explicit computation shows how the constant arises and how it is removed by loss re-scaling.

- **Backtracking Line Search:** This method adjusts the learning rate α to ensure a sufficient decrease in the loss function, as specified by the Armijo–Goldstein condition, and guarantees progress under smoothness assumptions.

- **Theoretical Guarantees:** Under Lipschitz continuity of the gradient, there exists a small enough α such that the sufficient decrease condition holds, ensuring that gradient descent converges.