Gradient Descent

Yutao Chen

⧖ Oct 03 2025
⧗ Oct 08 2025

Contents

In machine learning, we are often tasked with updating the model parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ to minimize some objective function $J(\boldsymbol{\theta}) : \mathbb{R}^d \mapsto \mathbb{R}$.

Gradient descent tells us to update $\boldsymbol{\theta}$ in the direction of the negative gradient, that is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla J(\boldsymbol{\theta}_t),$$

where η is commonly known as the *learning rate*.

Lagrangian Method

Directly minimizing the black-box objective $J(\boldsymbol{\theta})$ is intractable in general. We instead consider a linear approximation of $J(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_t$ using the first-order Taylor expansion:

$$J(\boldsymbol{\theta}) \approx J(\boldsymbol{\theta}_t) + \nabla J(\boldsymbol{\theta}_t)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_t).$$

We minimize the linear approximation of $J(\boldsymbol{\theta})$ with a squared Euclidean distance penalty $\|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_2^2$:

$$\arg\min_{\boldsymbol{\theta}} \; J(\boldsymbol{\theta}_t) + \nabla J(\boldsymbol{\theta}_t)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_t) + (2\eta)^{-1} \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_2^2,$$

where the learning rate η controls the *inverse* strength of the penalty. Intuitively, the penalty discourages large updates as the linear approximation might only hold in the vicinity of $\boldsymbol{\theta}_t$.

Setting the gradient of the optimization objective above to zero, we have

$$\begin{aligned} 0 &= \nabla J(\boldsymbol{\theta}_t) + \eta^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) \\ \implies \boldsymbol{\theta}^* &= \boldsymbol{\theta}_t - \eta \nabla J(\boldsymbol{\theta}_t). \end{aligned}$$

Riemannian Manifold

By far, we have assumed a Euclidean parameter space Θ such that $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, the squared distance between two parameters $D(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is

$$D(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top (\boldsymbol{\theta} - \boldsymbol{\theta}').$$

However, in general the parameter space Θ is not Euclidean. For example, one can parametrize the variance of a univariate Gaussian distribution by letting $\boldsymbol{\theta} = \sigma^2$, $\boldsymbol{\theta} = \sigma$, or $\boldsymbol{\theta} = \log \sigma$ (and more). Depending the chosen parametrization, updating $\boldsymbol{\theta}$ by the same constant c

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - c$$

will result in very different distributions.

More generally, one can consider the parameter space as a *Riemannian manifold*. The squared distance between two parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ on a Riemannian manifold[1] is

$$D(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top G(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\theta}'),$$

where the *Riemannian metric tensor* $G(\boldsymbol{\theta}) \in \mathbb{R}^{d \times d}$ is a symmetric and positive definite matrix depending on $\boldsymbol{\theta}$. Note that the Euclidean space corresponds to a special case where $G(\boldsymbol{\theta}) = \boldsymbol{I}$.

Amari (1998) showed that the steepest descent on a Riemannian manifold is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta G^{-1}(\boldsymbol{\theta}_t) \nabla J(\boldsymbol{\theta}_t).$$

Particularly when $G(\boldsymbol{\theta})$ is the Fisher information matrix (FIM), this is known as the [[Natural Gradient Descent]].

References

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*(2), 251–276. https://doi.org/10.1162/089976698300017746

[1] Strictly speaking, in the tangent space of the Riemannian manifold at $\boldsymbol{\theta}$.