# gradient

## Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

## Abstract

The gradient of a real-valued function is a vector that determines the local linear approximation of the function. The entries of the gradient are the partial derivatives of the function. However, the existence of partial derivatives does not imply the existence of a gradient, unless the function is convex. Geometrically, a nonzero gradient is orthogonal to the level sets and points in the direction of steepest ascent. In machine learning (ML), gradients of the empirical risk minimization (ERM) objective drive gradient descent (GD) methods and are computed for deep networks by backpropagation.

## Definition

The gradient of a real-valued function $f : \mathbb{R}^d \to \mathbb{R} : \mathbf{w} \mapsto f(\mathbf{w})$ determines a local linear approximation of $f$. Formally, the gradient of $f$ at a point $\mathbf{w}' \in \mathbb{R}^d$ is a vector $\mathbf{g} \in \mathbb{R}^d$ such that

$$\lim_{\mathbf{w} \to \mathbf{w}'} \frac{f(\mathbf{w}) - \big(f(\mathbf{w}') + \mathbf{g}^\top (\mathbf{w} - \mathbf{w}')\big)}{\|\mathbf{w} - \mathbf{w}'\|_2} = 0.$$

If such a vector exists, it is unique and denoted by $\nabla f(\mathbf{w}')$ or $\nabla f(\mathbf{w})\big|_{\mathbf{w}'}$ (Rudin, 1976, Ch. 9). The function $f(\mathbf{w}') + \big(\nabla f(\mathbf{w}')\big)^\top (\mathbf{w} - \mathbf{w}')$ is the local linear approximation of $f$ at $\mathbf{w}'$ (see Fig. 1). The entries of the gradient are the partial derivatives of $f$, $\nabla f(\mathbf{w}) = \big(\partial f/\partial w_1, \ldots, \partial f/\partial w_d\big)^\top$.

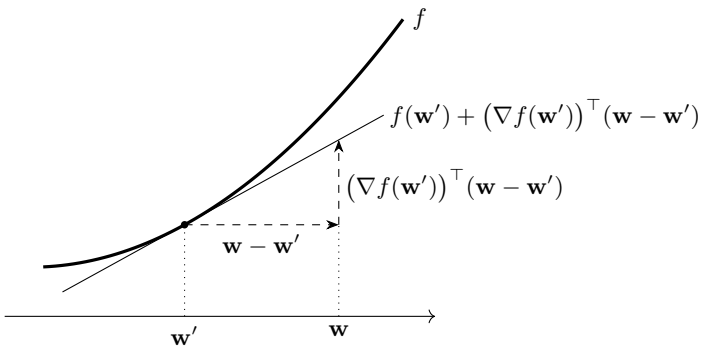


Figure 1: A differentiable function $f : \mathbb{R} \to \mathbb{R}$ and its local linear approximation at a point $\mathbf{w}'$. The slope triangle shows the displacement $\mathbf{w} - \mathbf{w}'$ and the resulting change $\big(\nabla f(\mathbf{w}')\big)^\top (\mathbf{w} - \mathbf{w}')$ of the local linear approximation

In general, the existence of all partial derivatives of $f$ at a point does not guarantee that the gradient exists there (Rudin, 1976, Ch. 9). For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, however, it does: if the partial derivatives $\partial f / \partial w_j$, for $j = 1, \ldots, d$, exist at a point $\mathbf{w}'$, then $f$ is differentiable at $\mathbf{w}'$, and the gradient $\nabla f(\mathbf{w}')$ is the vector of these partial derivatives (Rockafellar, 1970, Sect. 25).

The definition carries over to a real-valued function $f : \mathcal{H} \to \mathbb{R}$ on a Hilbert space $\mathcal{H}$: the inner product $\langle \mathbf{g}, \mathbf{w} - \mathbf{w}' \rangle$ of $\mathcal{H}$ replaces the term $\mathbf{g}^\top (\mathbf{w} - \mathbf{w}')$, and the norm of $\mathcal{H}$ replaces the Euclidean norm (Bauschke and Combettes, 2011).

The gradient has a geometric interpretation. At every point where $f$ is differentiable and $\nabla f \neq \mathbf{0}$, the gradient is orthogonal to the level set of $f$ through that point, and it points in the direction of steepest ascent of $f$. The negative gradient $-\nabla f(\mathbf{w}')$ points in the direction of steepest descent (see Fig. 2; Boyd and Vandenberghe, 2004, Sect. 9.4). At a local minimum of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, there cannot be any direction of descent and consequently the gradient must vanish (see zero-gradient condition).

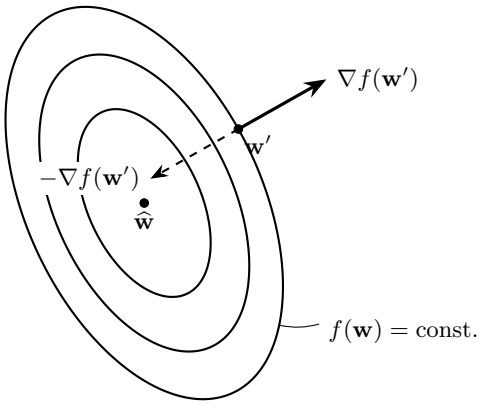


Figure 2: Level sets $f(\mathbf{w}) = \text{const.}$ of a differentiable function $f$ of two variables, with a minimum at $\widehat{\mathbf{w}}$. The gradient $\nabla f(\mathbf{w}')$ (solid) is orthogonal to the level set through $\mathbf{w}'$ and points in the direction of steepest ascent. The negative gradient $-\nabla f(\mathbf{w}')$ (dashed) points in the direction of steepest descent, toward smaller values of $f$. Gradient descent (GD) repeatedly steps along this direction. Data generated by `pythondemos/gradient.py`

Gradients are instrumental for the training of machine learning (ML) models. They guide the update of model parameters in order to minimize the incurred loss on a training set $\mathcal{D}^{(\text{train})} = \big\{\big(\mathbf{x}^{(r)}, y^{(r)}\big)\big\}_{r=1}^{m}$. As a case in point, linear regression minimizes the objective function

$$f(\mathbf{w}) := \frac{1}{m}\sum_{r=1}^{m}\big(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\big)^2.$$

This function is convex and differentiable, with gradient

$$\nabla f(\mathbf{w}) = -\frac{2}{m}\sum_{r=1}^{m}\big(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\big)\mathbf{x}^{(r)}.$$

For a deep artificial neural network (ANN) with differentiable activation functions, the empirical risk minimization (ERM) objective function is also differentiable. The gradient of this objective function can be computed by backpropagation (Goodfellow et al., 2016, Sect. 6.5; Rumelhart et al., 1986).

## Cross References

function, vector, differentiable, partial derivative, GD, zero-gradient condition, convex, Hilbert space

## References


1. Bauschke and Combettes (2011). Convex Analysis and Monotone Operator Theory in Hilbert Spaces. 1st ed., Springer Science+Business Media, New York, NY, USA.

2. Boyd and Vandenberghe (2004). Convex Optimization. Cambridge Univ. Press, Cambridge, U.K.

3. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

4. Rockafellar (1970). Convex Analysis. Princeton Univ. Press, Princeton, NJ, USA.

5. Rudin (1976). Principles of Mathematical Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.

6. Rumelhart et al. (1986). Learning Representations by Back-Propagating Errors. Nature, 323, pp. 533–536.