convex

Authors

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

Abstract

A subset of the Euclidean space is convex if it contains the line segment between any two of its points. A function is convex if its epigraph is a convex set. A convex optimization problem is the minimization of a convex function over a convex set. Empirical risk minimization (ERM) is a convex optimization problem whenever the hypothesis space is parameterized by a convex set of model parameters and the loss function is convex in those model parameters. Any local minimizer of a convex objective function is necessarily also a global minimizer, so a method that reaches a local minimum has solved the optimization problem.

Definition

Many machine learning (ML) methods learn model parameters by an iterative optimization method: each iteration updates the current model parameters to reduce an objective function within a neighborhood of these model parameters. Widely used examples are gradient descent (GD) and its variants.

These optimization methods work particularly well when the objective function is a convex function of the model parameters. Indeed, for a convex function every local minimum is necessarily a global minimum: any choice of model parameters at which no small change reduces the objective function already attains the smallest value that the objective function takes anywhere. The average squared error loss of linear regression is a convex function of the model parameters, and so is the objective function of the support vector machine (SVM). For differentiable convex objective functions and suitable step size choices, GD converges to such a minimum (Boyd and Vandenberghe, 2004).

Convexity is defined both for sets and for functions. For sets, the definition is geometric. A subset $\mathcal{C} \subseteq \mathbb{R}^d$ of the Euclidean space $\mathbb{R}^d$ is referred to as convex if it contains the line segment between any two points $\mathbf{w}, \mathbf{w}' \in \mathcal{C}$ in that set, i.e.,

$$\beta \mathbf{w} + (1-\beta)\mathbf{w}' \in \mathcal{C} \quad \text{for all } \beta \in [0,1].$$

Similarly, a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if its epigraph $\left\{ \left(\mathbf{w}^\top, t \right)^\top \in \mathbb{R}^{d+1} : t \geq f(\mathbf{w}) \right\}$ is a convex set (Boyd and Vandenberghe, 2004). Examples of a convex set and of a convex function are illustrated in Fig. 1.

Figure 1: (a) Convex set $\mathcal{C} \subseteq \mathbb{R}^d$: the line segment between any two points of $\mathcal{C}$ stays in $\mathcal{C}$. (b) Convex function $f : \mathbb{R}^d \to \mathbb{R}$: the shaded epigraph is a convex set, and the line segment connecting any two points on the graph of f lies above the graph

Two families of convex sets appear throughout ML. The first is the halfspace $\{\mathbf{w} \in \mathbb{R}^d : \mathbf{a}^\top \mathbf{w} \leq b\}$. It is determined by a nonzero normal vector $\mathbf{a} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, which fixes the direction of the bounding hyperplane, and an offset $b \in \mathbb{R}$, which fixes its position. If two points satisfy the inequality, so does every point of the segment between them (Boyd and Vandenberghe, 2004, Sect. 2.2.1). The second is the convex hull of finitely many vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, defined as the set of all their convex combinations,

$$\operatorname{conv}\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right\} := \left\{ \sum_{r=1}^{m} \beta_r \mathbf{x}^{(r)} \;:\; \beta_r \geq 0 \text{ for } r = 1, \ldots, m,\ \sum_{r=1}^{m} \beta_r = 1 \right\}.$$

It is the smallest convex set that contains these vectors (Boyd and Vandenberghe, 2004, Sect. 2.1.4).

Halfspaces are not merely one example among others: every closed convex set is given by the intersection of all halfspaces that contain it (Boyd and Vandenberghe, 2004, Sect. 2.3.1; Rockafellar, 1970, Thm. 11.5)

$$\mathcal{C} = \bigcap \left\{ \mathcal{H} \subseteq \mathbb{R}^d \;:\; \mathcal{H} \text{ is a halfspace, } \mathcal{C} \subseteq \mathcal{H} \right\}. \tag{1}$$

In general, intersecting a subset of the halfspaces containing a convex set yields a convex set that contains the original one, i.e., an outer approximation (see Fig. 2).

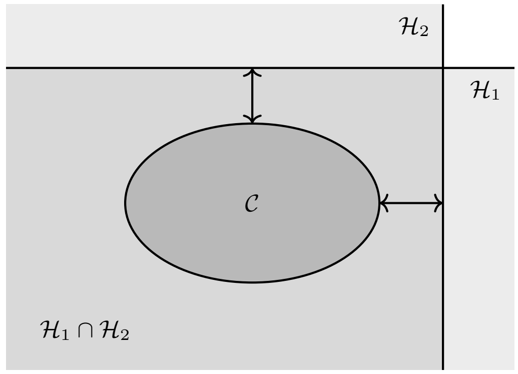

Figure 2: A convex set $\mathcal{C}$ and two halfspaces $\mathcal{H}_1, \mathcal{H}_2$ that contain it, in three shades of gray: lightest where only one halfspace covers the plane, mid-gray for their intersection $\mathcal{H}_1 \cap \mathcal{H}_2$, darkest for $\mathcal{C}$ itself. Neither halfspace is tight: its bounding hyperplane does not touch $\mathcal{C}$, leaving the gaps marked by the double arrows. The intersection is therefore a strictly larger convex set than $\mathcal{C}$ – an outer approximation

Approximate inference in probabilistic models defined over a graph uses this construction: the set of achievable marginal distributions is a convex set described by a large number of linear inequalities, and message-passing methods optimize over an outer approximation of it obtained from a subset of those inequalities (Wainwright and Jordan, 2008, Sect. 3.4.1; Sontag and Jaakkola, 2007).

Returning to (1), that intersection can be narrowed: it needs to consider only those halfspaces which are bounded by a supporting hyperplane. Every boundary point of a non-empty convex set carries such a hyperplane (Boyd and Vandenberghe, 2004, Sect. 2.5.2), and each point outside the set is already excluded by one of the halfspaces that these supporting hyperplanes bound: for a point $\mathbf{x} \notin \mathcal{C}$ and $\mathcal{C}$ closed, there is a unique nearest point $\mathbf{p} \in \mathcal{C}$ to $\mathbf{x}$. That nearest point is a boundary point of $\mathcal{C}$, and therefore one of the points that carry a supporting hyperplane: a point of the interior could be moved a little toward $\mathbf{x}$ and would still lie in $\mathcal{C}$, so it would not be the nearest one. The hyperplane through $\mathbf{p}$ with normal vector $\mathbf{x}-\mathbf{p}$ supports $\mathcal{C}$ at $\mathbf{p}$, and the halfspace it bounds contains $\mathcal{C}$ but not $\mathbf{x}$ (see Fig. 3). Fig. 4 shows eight supporting hyperplanes of an ellipse.

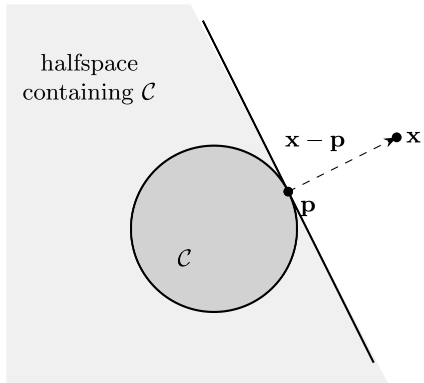

Figure 3: Why the supporting hyperplanes already suffice to recover a closed convex set $\mathcal{C}$, here a disk (darker shading). For a point $\mathbf{x}$ outside $\mathcal{C}$, let $\mathbf{p}$ be the point of $\mathcal{C}$ nearest to $\mathbf{x}$. The hyperplane through $\mathbf{p}$ orthogonal to $\mathbf{x} - \mathbf{p}$ (solid line) touches $\mathcal{C}$ at $\mathbf{p}$, so it is a supporting hyperplane, and the halfspace it bounds (lighter shading) contains $\mathcal{C}$ but not $\mathbf{x}$. Every point outside $\mathcal{C}$ is excluded by one such halfspace, so intersecting these halfspaces recovers $\mathcal{C}$

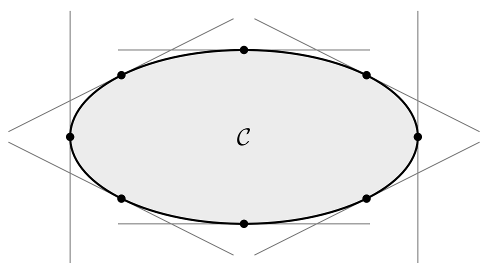

Figure 4: A convex set $\mathcal{C}$ (shaded ellipse) and eight hyperplanes that touch it (gray lines). Each of them bounds a halfspace that contains $\mathcal{C}$, and the intersection of all halfspaces containing $\mathcal{C}$ is $\mathcal{C}$ itself

For the epigraph of a differentiable convex function f, the supporting hyperplane is read off the gradient: at the boundary point $\big(\mathbf{w}^\top, f(\mathbf{w})\big)^\top$, its normal vector is $\big(\nabla f(\mathbf{w})^\top, -1\big)^\top$, since

$$\big(\nabla f(\mathbf{w})^\top, -1\big) \left[\begin{pmatrix} \mathbf{w}' \\ t \end{pmatrix} - \begin{pmatrix} \mathbf{w} \\ f(\mathbf{w}) \end{pmatrix} \right] \leq 0 \quad \text{for every } \big(\mathbf{w}'^\top, t\big)^\top \in \operatorname{epi} f.$$

This is the first-order condition $f(\mathbf{w}') \geq f(\mathbf{w}) + \nabla f(\mathbf{w})^\top(\mathbf{w}' - \mathbf{w})$ written as a statement about the epigraph (Boyd and Vandenberghe, 2004, Sect. 3.1.7).

Of the two families of convex sets introduced above, the convex hull is the one that characterizes the linear separability of data points with binary label values. Consider m data points with feature vectors $\mathbf{x}^{(r)} \in \mathbb{R}^d$ and labels $y^{(r)} \in \{-1, +1\}$, for $r = 1, \ldots, m$. These data points are linearly separable if some linear classifier predicts every label correctly, i.e., there are $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $b \in \mathbb{R}$ with $y^{(r)}\big(\mathbf{w}^\top \mathbf{x}^{(r)} + b\big) > 0$ for $r = 1, \ldots, m$. This holds precisely when the convex hull of the feature vectors labeled $+1$ and the convex hull of those labeled -1 do not intersect (Boyd and Vandenberghe, 2004, Sect. 8.6.1). Separability is therefore a property of two convex sets built from the data points.

Convexity also appears in the study of probabilistic models, where it singles out a widely used family. In particular, an exponential family is constituted by the probability distributions whose probability density function (pdf) has the form

$$p(\mathbf{x}; \mathbf{w}) = h(\mathbf{x}) \exp\big(\mathbf{w}^\top \mathbf{t}(\mathbf{x}) - A(\mathbf{w})\big).$$

This pdf is parameterized by the vector $\mathbf{w}$ and the sufficient statistics $\mathbf{t}(\cdot)$. The log-partition function A is fully determined by requiring $p(\mathbf{x}; \mathbf{w})$ to integrate to one. Its Hessian is the covariance matrix of the sufficient statistics $\mathbf{t}(\mathbf{x})$ under $p(\mathbf{x}; \mathbf{w})$,

$$\nabla^2 A(\mathbf{w}) = \mathbb{E}\big\{\mathbf{t}(\mathbf{x})\mathbf{t}(\mathbf{x})^\top\big\} - \mathbb{E}\big\{\mathbf{t}(\mathbf{x})\big\}\mathbb{E}\big\{\mathbf{t}(\mathbf{x})\big\}^\top,$$

which is positive semi-definite (psd). A twice-differentiable function whose Hessian is psd is convex, so A is a convex function on its domain (Wainwright and Jordan, 2008, Prop. 3.1). The convexity of A implies that the negative log-likelihood function of a training set, up to an additive constant that does not depend on $\mathbf{w}$,

$$-\sum_{r=1}^{m} \mathbf{w}^\top \mathbf{t}\big(\mathbf{x}^{(r)}\big) + mA(\mathbf{w}),$$

is a convex function of $\mathbf{w}$. Thus, for an exponential family, maximum likelihood estimation is a convex optimization problem.

Cross References

Euclidean space, function, epigraph, convex optimization, minimum, objective function, halfspace, supporting hyperplane, Hessian

References

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

2. Wainwright and Jordan (2008). Graphical Models, Exponential Families, and Variational Inference. Found. Trends Mach. Learn., 1(1–2), pp. 1–305.

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

4. Sontag and Jaakkola (2007). New Outer Bounds on the Marginal Polytope. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 1393–1400.