Gradient Descent

CS 4820—May 2014

David Steurer

Convex optimization

- *Given:* convex function $f\colon \mathbb{R}^n \to \mathbb{R}$
- *Find:* minimizer $x^* \in \mathbb{R}^n$ of function f so that $f(x^*) = \min_{x\in\mathbb{R}^n} f(x)$

Note. This problem specification is incomplete. In particular, we did not specify how the input function f is represented. For the sake of the current discussion, we will assume that we are given an explicit formula for $f(x_1,\ldots,x_n)$ in terms of the variables $x_1,\ldots,x_n$, using standard arithmetic operations as well as max / min operations. Another issue is that exact minimizers x^* of f might have irrational coordinates. What does it mean to output x^* in this case? We resolve this issue by allowing approximation, that is, our goal is to find a point $\tilde{x} \in \mathbb{R}^n$ such that $f(\tilde{x}) \approx f(x^*)$.

Applications

Convex optimization is a very general problem. We will see two examples of problems that reduce to convex optimization.

Linear programming

Claim. Linear programming reduces to convex optimization.

Given an LP instance, we can construct a convex function such that the minimizers of this function correspond to optimal LP solutions. To illustrate this reduction, let us show that the problem of finding a solution to a system of linear inequalities reduces to convex optimization.

Let $\{a_1^\top x \geq b_1,\ldots,a_m^\top x \geq b_m\}$ be a system of linear inequalities. (Here, $a^\top x$ denotes the scalar product of the vectors a and x.) Then, a point $x \in \mathbb{R}^n$ is a solution to this system of linear inequalities if and only if $f(x) \leq 0$ for the convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) = \max\{0, b_1 - a_1^\top x, \ldots, b_m - a_m^\top x\}\,.$$

Hence, if there exists a solution to the linear system, we can find one by computing a minimizer of f.

Supervised Machine Learning

Motivation. Suppose we have a way of encoding movies as vectors in $\mathbb{R}^n$. Then, the set of all movies corresponds to some subset $Y \subseteq \mathbb{R}^n$. To each movie $y \in Y$, we can assign a label $\sigma(y) \in \{\pm 1\}$ depending on whether we like the movie or not. Further, suppose that this labeling happens to be consistent with a hyperplane H through the origin, in the sense that all points $y \in Y$ above the hyperplane are labeled $\sigma(y) = 1$ and all points $y \in Y$ below the hyperplane are labeled $\sigma(y) = -1$. However, we have seen only a small subset

$\{y_1, \ldots, y_m\} \subseteq Y$ of the set of all movies and we don't know the separating hyperplane H. Can we extrapolate such a separating hyperplane given a small number of (random) examples $y_1, \ldots, y_m$ and their labels $\sigma_1 = \sigma(y_1), \ldots, \sigma_m = \sigma(y_m)$?

Model/Problem (Support vector machine).

- *Given:* example points $y_1, \ldots, y_m \in \mathbb{R}^n$ and labels $\sigma_1, \ldots, \sigma_m \in \{\pm 1\}$
- *Find:* a vector $w \in \mathbb{R}^n$ such that the hyperplane $\{x \mid w^\top x = 0\}$ provides an "optimal separation" between positive and negative examples, where the notion of "optimal separation" is formalized as minimizing the following convex function f for some parameter $\lambda \geq 0$,
$$f(w) = \sum_{i=1}^{m} \max\{1 - \sigma_i w^\top y_i, 0\} + \lambda \cdot \|w\|^2 \,.$$

Discussion. What's the justification for the choice of f? An "ideal separation" is achieved by a vector w with very small Euclidean length such that $w^\top y_1 = \sigma_1, \ldots, w^\top y_m = \sigma_m$. However such an ideal separation might not be possible for a given set of example points and labels. The function f is some way of measuring how far way w is from an ideal separation.

Convexity

Definition. A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is *convex* if for every point $x \in \mathbb{R}^n$, there exists a *lower-bounding linear interpolation* $\ell_x\colon \mathbb{R}^n \to \mathbb{R}$ such that $\ell_x(x) = f(x)$ and $\ell_x(y) \geq f(y)$ for all $y \in \mathbb{R}^n$.

Notation. Since ℓ_x is a linear function with $\ell_x(x) = f(x)$, there exists a vector $\nabla_x f \in \mathbb{R}^n$ such that
$$\ell_x(y) = f(x) + (\nabla_x f)^\top (y - x) \,.$$

The vector $\nabla_x f$ is called a (sub-)gradient of the function f at the point $x \in \mathbb{R}^n$.

Assumption. Since we assumed that f is represented as a simple formula, there exists an efficient algorithm that, given the formula for f and a point $x \in \mathbb{R}^n$, computes $f(x)$ and $\nabla_x f$.

Gradient Descent

Parameters.

- starting point $x_0 \in \mathbb{R}^n$,
- step size $\gamma > 0$
- number of iterations $T \in \mathbb{N}$

Algorithm.

- For t from 0 to $T - 1$,
  - compute $x_{t+1} = x_t - \gamma \nabla_x f$.
- Output the best point $\tilde{x} \in \mathbb{R}^n$ among $x_0, x_1, \ldots, x_T$ (so that $f(\tilde{x}) = \min\{f(x_0), \ldots, f(x_T)\}$).

Theorem. Suppose $\|x^* - x_0\|^2 \leq D^2$ and $\|\nabla_x f\|^2 \leq L^2$ for all $x \in \mathbb{R}^n$ with $\|x^* - x\|^2 \leq D^2$. Then, if we choose $\gamma = \varepsilon/L^2$ and $T = L^2D^2/\varepsilon^2$, then Gradient Descent outputs a point $\tilde{x}$ with $f(\tilde{x}) \leq f(x^*) + \varepsilon$.

The key ingredient for the analysis is the following lemma, which shows that in each iteration either $f(x_t) \leq f(x^*) + \varepsilon$ or the distance of the current point to x^* decreases by at least $2\gamma\varepsilon - \gamma^2\|\nabla_x f\|^2$

Lemma.

$$\|x^* - x_{t+1}\|^2 \leq \|x^* - x_{t+1}\|^2 - 2\gamma \cdot (f(x_t) - f(x^*)) + \gamma^2 \cdot \|\nabla_x f\|^2$$

Proof. The following algebraic identity achieves most of the proof,

$$\begin{aligned}
\|x^* - x_{t+1}\|^2 &= \|x^* - x_t + \gamma \nabla_{x_t} f\|^2 && \text{(gradient descent iteration)} \\
&= \|x^* - x_t\|^2 + 2\gamma \cdot (\nabla_{x_t} f)^\top (x^* - x_t) + \gamma^2 \cdot \|\nabla_{x_t} f\|^2 && \text{(quadratic binomial expansion)} \\
&= \|x^* - x_t\|^2 - 2\gamma \cdot \Big(f(x_t) - f(x_t) - (\nabla_{x_t} f)^\top (x^* - x_t)\Big) + \gamma^2 \cdot \|\nabla_{x_t} f\|^2 \\
&= \|x^* - x_t\|^2 - 2\gamma \cdot \Big(f(x_t) - \ell_{x_t}(x^*)\Big) + \gamma^2 \cdot \|\nabla_{x_t} f\|^2 && \text{(definition of gradient)}
\end{aligned}$$

By convexity, $\ell_{x_t}(x^*) \leq f(x^*)$. This inequality together with the previous identity imply the inequality in the lemma.

Proof of theorem

Consider some point x_t that does not satisfy the conclusion of the theorem, i.e., $f(x_t) > f(x^*) + \varepsilon$. Then, by the choice of γ and the condition on L^2, the lemma implies that

$$\|x^* - x_{t+1}\|^2 < \|x^* - x_t\|^2 - \varepsilon^2/L^2\,.$$

Suppose that all points $x_0, \ldots, x_{k-1}$ violate the conclusion of the theorem, then

$$\|x^* - x_k\|^2 < \|x^* - x_{k-1}\|^2 - \varepsilon^2/L^2 < \|x^* - x_{k-2}\|^2 - 2 \cdot \varepsilon^2/L^2 < \cdots < \|x^* - x_0\|^2 - k \cdot \varepsilon^2/L^2$$

Since the left-hand side is nonnegative and $\|x^* - x_0\|^2 \leq D^2$, it follows that $k < D^2L^2/\varepsilon^2$. Therefore, if we run Gradient Descent for $T = D^2L^2/\varepsilon^2$, one of the points $x_0, \ldots, x_{T-1}$ satisfies the conclusion of the theorem.