Note 4

04 Convex Optimization

by Alvin Wan

The goal in optimization is usually to find some parameters θ that minimize a continuous **objective function** $f(\theta)$. Note f is considered **smooth** if its gradient is also continuous.

1 Convexity

Consider a **convex set** C, which means that for any points in C we select, the line between them will lie completely in C. How is this related to, say, a convex function? As it turns out, we can relate the two by considering the **epigraph** of a function f, the set of all points that lie above f.

1.1 Epigraph

First, let us formalize the definition of a **strict epigraph**. (The standard epigraph has a weak inequality):

$$\{(x, t) : x \in \mathbb{R}^n, t \in \mathbb{R}, t > f(x)\}$$

Second, consider the formal definition of a **strongly convex** function f. A function f is convex if between any two points on f, the line between them lies entirely above the function - formally, if it satisfies the following, for $t \in [0, 1]$ and all x_1, x_2. Without loss of generality, consider $x_1 < x_2$.

$$f((1-t)x_1 + tx_2) < (1-t)f(x_1) + tf(x_2)$$

To define a **weakly convex** function, we simply convert the strict inequality into a weak one, to get $f((1-t)x_1 + tx_2) \leq (1-t)f(x_1) + tf(x_2)$.

Explanation of Definition:

Intuitively, we take the points $(x_1, f(x_1)), (x_2, f(x_2))$ and draw the line between them. Consider taking your finger, and following the line from x_1 to x_2. t represents how far along we are, on that line, where $t = 0$ is at x_1 and $t = 1$ is at x_2. To compute the y-value at a certain point along that line, we simply weight the two y-values at the end points. If we are closer to x_1, weight $f(x_1)$ more heavily. Thus, the y value on that line is

$$(1 - t)f(x_1) + tf(x_2)$$

The corresponding evaluation at the real value of f is straightforward. Instead of weighting the y values, we weight the x values, then evaluate the true y value at that point.

$$f((1 - t)x_1 + tx_2)$$

On an intuitive level, this is how the above definition captures the condition that any line lies above f, or in words, any line between two points on f will necessarily lie in the epigraph of f.

Theorem: The epigraph of a function f is a convex set if and only if f is weakly convex.

Proof:

The forward direction is simpler. Take a pair of points $(x_i, f(x_i))$. Since f is a convex set, we know for any point along the line between them, (x, y) must satisfy the condition $y \geq f(x)$. Plugging in for $x = (1 - t)x_1 + tx_2$ and $y = (1 - t)y_1 + ty_2 = (1 - t)f(x_1) + tf(x_2)$, our claim immediately follows.

We then prove the reverse direction: if f is convex, the epigraph is a convex set. Take any two points (x_i, y_i) and the line between them.

$$x = (1 - t)x_1 + tx_2$$
$$y = (1 - t)y_1 + ty_2$$

Apply the definition of an epigraph, to plug in for y_i. Then, apply the definition of convexity.

$$y = (1-t)y_1 + ty_2 \geq (1-t)f(x_1) + tf(x_2) \geq f((1-t)x_1 + tx_2) = f(x)$$

Thus, $y \geq f(x)$ and (x, y) is in the epigraph.

1.2 Solutions

If we minimize over a convex function f, we could find a **local minimum**, a value x such that there exists a ball of some radius around x where all x' in this ball satisfy $f(x') \geq f(x)$. It may be a **global minimum**, where for *all* x', $f(x') \geq f(x)$. However, the global minimum is often difficult and sometimes impossible to find. A continuous convex function has:

1. No minimum, as the function may be weakly convex and go to $-\infty$.
2. Just one local minimum
3. Connected set of local minima that are all global minima.

Perceptrons fall in the last category, and neural networks often find local not global minima.

1.3 Cases

Consider the following cases when functions are convex.

1. Constant functions: $f(x) = c$
2. Powers of x: $f(x) = x^r$ on interval $[0, \infty]$ when $r \geq 1$
3. Affine transform: $f(x)$ convex implies $f(w^T x + \beta)$ is convex.
4. Log: $-\log(x)$ is convex.
5. Reverse of Concave: If $f(x)$ is concave, $-f(x)$ is convex.
6. Combination: If f, g are convex, then $\max(f(x), g(x))$ is convex.

2 Optimization Problems

2.1 Unconstrained Optimization

For smooth functions, we can use gradient descent, the nonlinear conjugate gradient, or Newton's method. For nonsmooth functions, we can use gradient descent. A **line search** finds a local minimum along one dimension.

2.2 Constrained Optimization

Here we minimize $f(\theta)$ for some parameters θ subject to $g(\theta) = 0$ for some constraint g, where g is a smooth function. In other words, we have smooth equality constraints. To solve this, use Lagrange multipliers.

2.2.1 Lagrange Multipliers

Lagrange multipliers allow us to solve optimization problems with constraints.

1. Set up a new function $h(\theta) = f(\theta) + \lambda g(w)$, where λ is the multiplier.
2. Take partials with respect to every parameter in θ and λ. If there are d parameters, there are $d + 1$ unknowns.
3. Set each partial to 0. This makes $d + 1$ equations for $d + 1$ unknowns.
4. Solve the system.
5. Ensure that you do not divide variables x_i out without considering the possibility that $x_i = 0$.

2.3 Linear Programming

With a linear objective function with inequality constraints, our goal is to maximize or minimize $c \cdot w$ subject to $Aw \leq b$. Note the polytope or feasible region is convex. For some constraints - named the **active constraints** - the optimum achieves equality. In an SVM, vectors that likewise touch or cross the slab are called **support vectors**.

2.4 Quadratic Programming

Minimize $f(w) = w^T Q w + c^T w$ subject to $Aw \leq b$, where Q is a symmetric, positive definite matrix.