CO 367 – Notes on Unconstrained Optimization I

March 7, 2011

[*Reference: Boyd and Vandenberghe – textbook and lecture slides*]

This week we consider unconstrained minimization problems: given a C^2 function $f : \mathbb{R}^n \to \mathbb{R}$, solve

$$(*) \qquad \min_{x \in \mathbb{R}^n} f(x),$$

that is, find the global minimizer(s) of f.
We assume that $p^* > -\infty$ and there exists some x^* such that $p^* = f(x^*)$.
Recall:

1. If x^* solves $(*)$, then the *first order necessary condition* holds:

$$\nabla f(x^*) = 0. \tag{1}$$

2. Suppose, in addition, that f is a convex function.

   If x^* satisfies (1), then x^* solves $(*)$.

   Hence

$$\text{solving } (*) \;\equiv\; \text{solving the equation } \nabla f(x) = 0 \text{ for } x.$$

1 How to solve $(*)$?

- Use *iterative methods*:
  - Start with an initial point $x^{(0)}$.
  - Produce a sequence $\left\{x^{(k)}\right\}_{k=1,\ldots}$ such that
    1. $f\left(x^{(k)}\right) \leq f\left(x^{(0)}\right)$ for all k.
    2. $f\left(x^{(k)}\right) \to p^*$ as $k \to \infty$.

  We would hope that the sequence $\left\{x^{(k)}\right\}$ at least has a limit point, which we would take as the solution of $(*)$.

- But we usually don't know p^* a priori — then how do we know if the objective values of our iterates $x^{(k)}$ are getting close to p^*?
- Usually we require $\left\{x^{(k)}\right\}$ to satisfy

$$\left\|\nabla f\left(x^{(k)}\right)\right\| \to 0 \tag{2}$$

  instead.
- **If $\left\{x^{(k)}\right\}$ has a limit point**, then by continuity of f and ∇f,

$$\begin{aligned} f\left(x^{(k)}\right) \to p^* \quad &\Longrightarrow \quad f(x^*) = p^*, \\ \left\|\nabla f\left(x^{(k)}\right)\right\| \to 0 \quad &\Longrightarrow \quad \nabla f(x^*) = 0. \end{aligned}$$

2 Strongly convex functions

So for what kind of function f can we be sure that, if we produce a sequence $\left\{x^{(k)}\right\}$ satisfying

1. $f\left(x^{(k)}\right) \le f\left(x^{(0)}\right)$ for all k, and
2. $\left\|\nabla f\left(x^{(k)}\right)\right\| \to 0$ as $k \to \infty$,

then

1. $\{x_k\}_{k=1,\dots}$ has a limit point x^*, and
2. The limit point x^* solves $(*)$?

One class of functions that enjoy this property is the class of *strongly convex functions*. A function is strongly convex on a set S if there exists some $m > 0$ such that

$$\nabla^2 f(x) - mI \text{ is positive semidefinite for all } x \in S,$$

or, equivalently,

$$f(y) \ge f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|\nabla f(x)\|^2$$

for all $x, y \in S$.

- Let $x^{(0)} \in \operatorname{dom} f$. If f is strongly convex on the set

$$S := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\} = f^{-1}\big((-\infty, f(x_0)]\big),$$

  which is a closed set, then

1. S is a bounded set. (So S is a compact set.)

   This ensures that the sequence $\left\{x^{(k)}\right\}$ has a limit point.
2. $p^* > -\infty$.
3. $f(x) - p^* \leq \dfrac{1}{2m} \|\nabla f(x)\|_2^2$.

   This means that even if we don't know p^*, we can look at $\|\nabla f(x^{(k)})\|$ to determine how close $f(x^{(k)})$ is to the optimal value (provided we know m).
4. The limit point x^* of $\left\{x^{(k)}\right\}$ satisfies $\nabla f(x^*)$ and therefore solves $(*)$ by convexity of f.

- Simple example/non-examples of strongly convex function:
  - $f(x, y) = x^2 + y^2$ is strongly convex.
  - $f(x, y) = x^4 + y^4$ is not strongly convex. In fact, it is not even strictly convex (at $(0, 0)$).
  - $f(x) = e^x$ is strictly convex, but not strongly convex. Note that f does not have a minimizer on $\mathbb{R}$: $\inf_{\mathbb{R}} f = 0$ but $f(x) > 0$ for all $x \in \mathbb{R}$.

3 Descent methods

An (iterative) *descent method* generates a sequence $\left\{x^{(k)}\right\}$ of point such that

$$f\left(x^{(k+1)}\right) < f\left(x^{(k)}\right).$$

Each iterate $x^{(k)}$ is obtained via the following: Three ingredients needed:

Algorithm 3.1: General descent method

```
Given( initial point x^(0) );
repeat
    1. Determine a search direction Δx.
    2. Line search. Choose a step size t > 0.
    3. x ← x + tΔx.
until stopping criterion is satisfied. ;
```

1. search direction,
2. step size, and
3. stopping criterion.

3.1 Search direction Δx

- Usually require that
$$\nabla f(x)^T \Delta x < 0. \tag{3}$$
A vector satisfying (3) is called a *descent direction*.

- Reason: f convex; if $\nabla f(x)^T \Delta x < 0$ does not hold, then for all $t > 0$
$$f(x + t\Delta x) \geq f(x) + t\nabla f(x)^T \Delta x \geq f(x),$$
so $x + t\Delta x$ does not give a lower objective value.

- **How to pick Δx?**
A natural choice if $\Delta x = -\nabla f(x)$.
This corresponds to *steepest descent method*.

3.2 Line search

Once we fix the search direction, we need to pick a point on the ray $\{x + t\Delta x : t > 0\}$. This amounts to choosing a positive parameter t, which is called the step size. There are two common types of line search:

- *Exact line search.*

  We find t such that $f(x + t\Delta x)$ is the smallest possible on the ray. To find this t is equivalent to solving the single-variate minimization problem
$$\min_{t>0} f(x + t\Delta x).$$
  It can be quite expensive to perform though. So usually we use the other alternative: backtracking.

- *Backtracking line search.*

  Backtracking is the computationally less expensive option for choosing a step size that guarantees sufficient objective value decrease at the new iterate.

  We have two parameters $\alpha \in (0, 0.5)$ and $\beta \in (0, 1)$.

  - α controls the (relative) level of objective value decrease at each iteration.
  - β is the backtracking parameter.

  The backtracking procedure is as follows:

  - We start with $t = 1$.

– Repeat

$$t \leftarrow \beta t$$

until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x. \tag{4}$$

This inequality is equivalent to

$$\frac{f(x) - f(x + t\Delta x)}{t} > \alpha \cdot \left(-\nabla f(x)^T \Delta x\right),$$

so the choice of t is such that the objective value decrease (given by the LHS) is to be at least a factor α of $-\nabla f(x)^T \Delta x$.

3.3 Stopping criterion

We usually use the first order necessary condition of optimality (1): we stop at an iterate $x^{(k)}$ if $\nabla f\left(x^{(k)}\right) \sim 0$, in the sense that

$$\left\| f\left(x^{(k)}\right)\right\| \leq \varepsilon, \tag{5}$$

where $\varepsilon > 0$ is a very small number (the user-defined zero-tolerance).