AM 221: Advanced Optimization Spring 2016

Prof. Yaron Singer *Section 6 — Wednesday, Mar. 2nd*

1 Subdifferential calculus

Subdifferential calculus is a beautiful tool to handle convex optimization situations where the objective function is non-differentiable. The central concept in subdifferential calculus is the one of *subgradient* which generalizes the notion of gradient to non-differentiable functions. Most of the convex optimization theory can be derived with subgradients instead of gradients by making appropriate changes.

The most important property of gradients in the context of convex optimization is that they give linear lower-bounds:

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}),\ \mathbf{y} \in \mathbb{R}^n$$

when $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function differentiable at $\mathbf{x}$. This is the property that subgradients preserve as follows from this definition:

Definition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function (not necessarily differentiable). $\mathbf{g} \in \mathbb{R}^n$ is a subgradient of f at $\mathbf{x} \in \mathbb{R}^n$ iff:*

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \mathbf{g}^\intercal(\mathbf{y} - \mathbf{x}),\ \mathbf{y} \in \mathbb{R}^n$$

The set of all subgradients at $\mathbf{x}$: $\partial f(\mathbf{x}) \overset{\text{def}}{=} \{\mathbf{g} \in \mathbb{R}^n \,|\, \mathbf{g}$ is a subgradient of f at $\mathbf{x}\}$ is called the subdifferential of f at $\mathbf{x}$.

The reason why subgradients are so useful in convex optimization is that they always exist for convex functions:

Proposition 2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex continuous function, then $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x} \in \mathbb{R}^n$.*

Proof. Consider the epigraph $E(f)$ of f:

$$E(f) \overset{\text{def}}{=} \{(\mathbf{x}, y) \in \mathbb{R}^n \times \mathbb{R} \,|\, y \geq f(\mathbf{x})\}$$

□

It is easy to see that $E(f)$ is a closed convex set. Let us now consider some $\mathbf{x} \in \mathbb{R}^n$. The point $(\mathbf{x}, f(\mathbf{x}))$ is on the boundary of $E(f)$. Applying the supporting hyperplane theorem to $E(f)$ and $(\mathbf{x}, f(\mathbf{x}))$ gives the desired result.

Proposition 3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function, then $\partial f(\mathbf{x})$ is a closed and convex set.*

Proof. See Problem Set 6. □

In cases where f is differentiable, the relationship between subgradients and the gradient is given by the following proposition.

Proposition 4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex continuous function differentiable at $\mathbf{x} \in \mathbb{R}^n$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.*

Proof. It is clear that $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$. This comes from the characterization of convexity with the gradient (which motivated the definition of subgradients in the first place).

Let us now consider $\mathbf{g} \in \partial f(\mathbf{x})$. For any $\mathbf{h} \in \mathbb{R}^n$ and $\lambda > 0$, we write:

$$f(\mathbf{x} + \lambda\mathbf{h}) - f(\mathbf{x}) \geq \lambda \mathbf{g}^\intercal \mathbf{h}$$

in other words, because $\lambda > 0$:

$$\frac{f(\mathbf{x} + \lambda\mathbf{h}) - f(\mathbf{x})}{\lambda} \geq \mathbf{g}^\intercal \mathbf{h}$$

by taking the limit when λ goes to 0: $\nabla f(\mathbf{x})^\intercal \mathbf{h} \geq \mathbf{g}^\intercal \mathbf{h}$. After reordering, this is:

$$(\nabla f(\mathbf{x}) - \mathbf{g})^\intercal \mathbf{h} \geq 0$$

Choosing $\mathbf{h} = \mathbf{g} - \nabla f(\mathbf{x})$, we obtain $-\|\nabla f(\mathbf{x}) - \mathbf{g}\|^2 \geq 0$ which implies $\mathbf{g} = \nabla f(\mathbf{x})$. □

It is possible to show a reciprocal of Proposition 4 (it is slightly harder so we omit the proof).

Proposition 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and assume that for some $\mathbf{x} \in \mathbb{R}^n$, $\partial f(\mathbf{x}) = \{\mathbf{g}\}$. Then f is differentiable at $\mathbf{x}$ and $\nabla f(\mathbf{x}) = g$.*

Example. Let us consider $f : \mathbb{R} \to \mathbb{R}$ define by $f(x) = |x|, x \in \mathbb{R}$. The only point at which f is not differentiable is $x = 0$. At this point, subderivatives are characterized by the inequality $f(x) - f(0) \geq gx$, *i.e.* $|x| \geq gx$. This implies $g \in [-1, 1]$. For $x \neq 0$, the subderivatives coincide with the derivative. In summary:

$$\partial f(x) = \begin{cases} 1 & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

In the same way stationary points characterize optimal points for convex functions, we have the following proposition:

Proposition 6. *Let f be a convex continuous function, $\mathbf{x}^* \in \mathbb{R}^n$ is a global minimizer of f iff $0 \in \partial f(\mathbf{x})$.*

Proof. See problem set 6. □

Example. In this example we will use Proposition 6 to minimize $f : \mathbb{R}^n \to \mathbb{R}$ given by:

$$f(\mathbf{x}) = \frac{1}{2}\|y - x\|^2 + \lambda\|x\|_1$$

Using the previous example, it is easy to see that the subgradients of f at $\mathbf{x}$ have the form:

$$\mathbf{y} - \mathbf{x} + \lambda\mathbf{s}$$

where:

$$\mathbf{s}_i = \begin{cases} 1 & \text{if } x_i > 0 \\ [-1, 1] & \text{if } x_i = 0 \\ -1 & \text{if } x_i < 0 \end{cases}$$

Then if we define $\mathbf{x}^*$ by:

$$x_i^* = S(y_i) \overset{\text{def}}{=} \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } y_i \in [-\lambda, \lambda] \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$

It is not hard to verify that $0 \in \partial f(\mathbf{x})$. Hence the optimal solution is given by $\mathbf{x}^* = S(\mathbf{y})$ where S is the function defined above; this function is known as the *soft-thresholding* operator.

There are many other results in subdifferential calculus. For the purpose of designing convex optimization algorithms, it is important to note that gradient descent can be directly generalized to *subgradient descent*: at each iteration, instead of moving in the direction of the gradient, move in the direction of any subgradient; this algorithm provably converge to an optimal solution.

A good reference on convex optimization with subdifferentials is *Convex Analysis* by R. Rockafellar.

2 Backtracking line search

In class, we analyze gradient descent with exact line search. At each iteration, the step size was determined by solving:

$$\min_{t \in \mathbb{R}} f(\mathbf{x} + t\nabla f(\mathbf{x}))$$

In scenarios where exact line search is hard to solve, a simple heuristic to find a good step size is *Backtracking Line Search*: starting from $t = 1$, the stepsize is reduced by a multiplicative factor $\beta < 1$ until the decrease in value (with respect to f) is small enough. The formal description of Backtracking Line Search is given by Algorithm 1.

Algorithm 1 Backtracking Line Search, with parameters $\alpha \in (0, \frac{1}{2})$ and $\beta \in (0, 1)$

```
Require: $\mathbf{x}$, $\nabla f(\mathbf{x})$
t ← 1
while $f(\mathbf{x} - t\nabla f(\mathbf{x})) > f(\mathbf{x}) - \alpha t \|\nabla f(\mathbf{x})\|^2$ do
    $t \leftarrow \beta t$
end while
return $t$
```

Remarkably, it is still possible to obtain convergence guarantees of gradient descent when using backtracking line search.

Theorem 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an m, M-strongly convex function. For any $\alpha \in (0, \frac{1}{2})$ and $\beta \in (0, 1)$, the solution $\mathbf{x}^{(k)}$ at the kth iteration of gradient descent with backtracking line search satisfies:*

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \left(1 - 2\alpha m \min\left(1, \frac{\beta}{M}\right)\right)^k \left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)\right)$$

Proof. As in the standard analysis of gradient descent, we have:

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) + \left(\frac{Mt^2}{2} - t\right) \|\nabla f(\mathbf{x}^{(k)})\|^2$$

this follows from the standard quadratic upper bound on strongly convex functions. When $0 \leq t \leq \frac{1}{M}$, we have $\frac{Mt^2}{2} - t \leq -\frac{t}{2}$, hence:

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) - \alpha t \|\nabla f(\mathbf{x}^{(k)})\|^2 \tag{1}$$

since $\alpha < \frac{1}{2}$. This shows that backtracking line search finishes in at most $\log_\beta \frac{1}{M}$ iterations. Furthermore when backtracking line search terminates, we have either:

- $t = 1$, if we didn't enter the while loop
- $t \geq \frac{\beta}{M}$, if we entered the while loop: at the last iteration we had $t \geq \frac{1}{M}$, after multiplying by β one last time, we get $t \geq \frac{\beta}{M}$.

Combining this fact with Equation (2), we obtain:

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) - \alpha \min\left(1, \frac{\beta}{M}\right) \|\nabla f(\mathbf{x}^{(k)})\|^2$$

We can now conclude as in the standard analysis of gradient descent:

$$f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^*) \leq f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) - \alpha \min\left(1, \frac{\beta}{M}\right) \|\nabla f(\mathbf{x}^{(k)})\|^2$$

By strong convexity, we have:

$$\|\nabla f(\mathbf{x}^{(k)})\|^2 \geq 2m \left(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)\right)$$

Hence:

$$f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^*) \leq \left(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)\right)\left(1 - 2m\alpha \min\left(1, \frac{\beta}{M}\right)\right)$$

The statement of the theorem then follows by a simple induction on k. □

Remark. Since $\alpha\beta < \frac{1}{2}$ it is interesting to note that the rate of convergence is worse than with exact line search. How to set α and β? There is a trade-off between:

- convergence rate, where choosing α and β as close as possible to $\frac{1}{2}$ and 1 respectively is better
- running time, where choosing α and β close to $\frac{1}{2}$ and 1 increases the number of iterations of the while loop in the backtracking line search algorithm.