

Lecture 1 — December 1, 2009

Lecturer: Martin Wainwright

Scribe: Martin Wainwright

1.1 Motivation

Thus far, we have explored a variety of optimization methods (e.g., generalized descent methods, Newton's method, projected gradient methods), all of which apply to differentiable functions. However, many functions that arise in practice may be non-differentiable at certain places. A common example is the absolute value function $f(x) = |x|$, or its multivariate extension, the ℓ_1 -norm $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$. In this lecture, we discuss a generalization of the gradient for non-differentiable convex functions.

1.2 Subgradients and subdifferentials

In order to motivate the definition that follows, let us recall the *tangent approximation interpretation* of the gradient for a convex function. If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and differentiable, then the tangent plane inequality guarantees that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \text{for all } y \in \text{dom}(f). \quad (1.1)$$

Geometrically, this inequality means that the tangent plane with normal vector $\nabla f(x)$ supports the epigraph of f at x .

With this intuition, we have the following:

Definition: A vector $v \in \mathbb{R}^n$ is a *subgradient* of a convex function at x if

$$f(y) \geq f(x) + \langle v, y - x \rangle \quad \text{for all } y \in \text{dom}(f).$$

The set of all subgradients at x is called the *subdifferential*, and is denoted by $\partial f(x)$.

Remarks:

- (a) It can be shown that for a convex function and an element $x \in \text{dom}(f)$, the subdifferential $\partial f(x)$ is always a non-empty set. (See Appendix B of Bertsekas [1] for details of this argument, which uses the separating hyperplane theorem that we have covered.) From the definition, we also see that $\partial f(x)$ must be a convex set, since any convex combination of subgradients is also a subgradient.
- (b) If the function f is actually differentiable at x , then we have $\partial f(x) = \{\nabla f(x)\}$; when f is not differentiable, then the subdifferential $\partial f(x)$ is a more interesting set, as we will find the examples to follow.

1.2.1 Some examples

- (a) The simplest example is the absolute value function $f(x) = |x|$. This function is differentiable for all $x \neq 0$, with $f'(x) = \text{sign}(x)$. It is non-differentiable at $x = 0$. To determine the possible subgradients at the remaining point $x = 0$, we note that the inequality

$$f(y) = |y| \geq f(0) + v \cdot y = v \cdot y$$

holds for all $y \in \mathbb{R}$ as long as $|v| \leq 1$. Therefore, we conclude that $\partial f(0) = [-1, +1]$.

- (b) Now consider the function $f(x) = \max\{0, \frac{1}{2}(x^2 - 1)\}$, which is continuous and convex. It is differentiable everywhere except at $x = +1$ or $x = -1$. At these two points, a little calculation shows that $\partial f(1) = [0, 1]$ and $\partial f(-1) = [-1, 0]$.
- (c) As a more complex example, let us consider the maximum eigenvalue function—namely, the function λ_{\max} that maps any symmetric matrix $Q \in \mathcal{S}^n$ to its maximum eigenvalue $\lambda_{\max}(Q)$. We claim that this is a convex function of the matrix Q . Indeed, from our earlier work on SDPs, we can write

$$\lambda_{\max}(Q) = \max_{\substack{Y \succeq 0 \\ \text{trace}(Y)=1}} \text{trace}(YQ).$$

This representation shows that $\lambda_{\max}(Q)$ is the maximum of a collection of linear functions, and so convex in Q (see Chapter 3 of Boyd and Vanbenberghe [2]).

Let us compute the sub-differential of $\lambda_{\max}(Q)$. In particular, we need to determine the set of symmetric matrices V such that

$$\lambda_{\max}(R) \geq \lambda_{\max}(Q) + \text{trace}(V(R - Q)) \quad \text{for all } R \in \mathcal{S}^n,$$

where we recall that the trace function defines an inner product on the space of symmetric matrices. If we let $z \in \mathbb{R}^n$ be a maximum eigenvalue of Q , normalized such that $\|z\|_2 = 1$, then we have $\lambda_{\max}(R) \geq \text{trace}(Rzz^T)$. Therefore, if we add and subtract Q , then

$$\begin{aligned} \lambda_{\max}(R) &\geq \text{trace}(Qzz^T) + \text{trace}((R - Q)zz^T) \\ &= \lambda_{\max}(Q) + \text{trace}((R - Q)zz^T). \end{aligned}$$

Thus, we have shown that the matrix $Z = zz^T$, for any unit-norm eigenvector of Q , belongs to the subdifferential $\lambda_{\max}(Q)$. By the definition of subgradient, any convex combination of sub-gradients is also a subgradient, so that we conclude that

$$\partial \lambda_{\max}(Q) = \text{convhull} \{ z_i z_i^T \mid \|z_i\|_2 = 1, \quad Qz_i = \lambda_{\max}(Q)z_i \}.$$

Note that if Q has a unique maximal eigenvector (say z with $\|z\|_2 = 1$), then the max. eigenvalue function is differentiable at Q , with $\nabla \lambda_{\max}(Q) = zz^T$.

1.3 Subgradients and Lagrangian duality

One setting in which subgradients often arise in the context of dual optimization methods that try to maximize the Lagrangian dual function q , known as *dual methods*. In this setting, we often obtain a subgradient “for free”, namely as a by-product of computing the dual function. In particular, suppose that we wish to optimize the function f subject to the constraint $g(x) \leq 0$. For $\mu \geq 0$, the associated dual function takes the form

$$q(\mu) = \inf_{x \in \text{dom}(f)} \{f(x) + \mu^T g(x)\} = \inf_{x \in \text{dom}(f)} L(x, \mu).$$

As we have seen, the function q is always concave in μ . Recalling that a function q is concave if and only if $-q$ is convex, the subgradient relation translates as follows: the vector $z \in \partial q(\mu)$ if and only if

$$q(\nu) \leq q(\mu) + z^T(\nu - \mu) \quad \text{for all } \nu \in \text{dom}(q).$$

The following result shows how to obtain subgradients of the dual function:

Proposition: For a given $\mu \geq 0$, let x_μ be any element of $\arg \min_{x \in \text{dom}(f)} L(x, \mu)$. Then $g(x_\mu)$ is an element of $\partial q(\mu)$.

Proof: For any $\nu \in \text{dom}(q)$, since $q(\nu)$ is obtained by minimizing $L(x, \nu)$ over $x \in \text{dom}(f)$, we have $q(\nu) \leq L(x_\mu, \nu)$. Moreover, since x_μ achieves the minimum, we have $q(\mu) = L(x_\mu, \mu)$. Combining the pieces, we obtain

$$\begin{aligned} q(\nu) &\leq L(x_\mu, \nu) \\ &= L(x_\mu, \mu) + [L(x_\mu, \nu) - L(x_\mu, \mu)] \\ &= q(\mu) + \langle g(x_\mu), \nu - \mu \rangle. \end{aligned}$$

□

1.4 Subgradient method for dual optimization

Let us now consider a natural extension of a gradient method for maximizing the Lagrangian dual function q . In particular, define the convex set

$$M := \{\mu \in \mathbb{R}^n \mid \mu \geq 0, q(\mu) > -\infty\},$$

corresponding to the set of dual feasible vectors. This method generates a sequence $\{\mu^k\}_{k=0}^\infty$ contained within M as follows:

- (1) Initialize at some $\mu^0 \in M$.
- (2) Given a sequence $\{s_k\}$ of positive step sizes:

- (a) Find a subgradient $g^k \in \partial q(\mu^k)$.
 (b) Update

$$\mu^{k+1} = \Pi_M(\mu^k + s^k g^k)$$

where Π_M denotes the projection onto M .

We assume that M is a closed set so that the projection onto M is well-defined.

In contrast to a gradient method, the subgradient method is *not guaranteed* to be an ascent method. That is, it is possible that $q(\Pi_M(\mu^k + s g^k)) < q(\mu^k)$ for all $s > 0$, as we saw in the illustration from class.

However, the following result guarantees that μ^{k+1} becomes closer to an optimal dual solution μ^* .

Proposition: If μ^k is not optimal, then for every dual optimal solution μ^* , we have

$$\|\mu^{k+1} - \mu^*\|_2 < \|\mu^k - \mu^*\|_2$$

for all stepsizes s^k such that

$$0 < s^k < \frac{2(q(\mu^*) - q(\mu^k))}{\|g^k\|_2^2} = s^*.$$

Proof: By definition, we have

$$\|\mu^k + s^k g^k - \mu^*\|_2^2 = \|\mu^k - \mu^*\|_2^2 - 2s^k (\mu^* - \mu^k)^T g^k + (s^k)^2 \|g^k\|_2^2.$$

Since $g^k \in \partial q(\mu^k)$, we have $(\mu^* - \mu^k)^T g^k \geq q(\mu^*) - q(\mu^k)$, and hence

$$\|\mu^k + s^k g^k - \mu^*\|_2^2 \leq \|\mu^k - \mu^*\|_2^2 - 2s^k (q(\mu^*) - q(\mu^k)) + (s^k)^2 \|g^k\|_2^2.$$

Now define the function

$$h(t) := -2t (q(\mu^*) - q(\mu^k)) + t^2 \|g^k\|_2^2.$$

We see that $h(0) = h(s^*) = 0$ and moreover that $h(t) < 0$ for all $t \in (0, s^*)$, which establishes that

$$\|\mu^k + s^k g^k - \mu^*\|_2 < \|\mu^k - \mu^*\|_2.$$

Finally, by non-expansiveness of projection, we have

$$\begin{aligned} \|\mu^{k+1} - \mu^*\|_2 &= \|\Pi_M(\mu^k + s^k g^k) - \Pi_M(\mu^*)\|_2 \\ &\leq \|\mu^k + s^k g^k - \mu^*\|_2 < \|\mu^k - \mu^*\|_2 \end{aligned}$$

which completes the proof. □

Bibliography

- [1] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.
- [2] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.