10-725/36-725: Convex Optimization **Fall 2015**

Lecture 6: September 17

Lecturer: Ryan Tibshirani *Scribes: Scribes: Wenjun Wang, Satwik Kottur, Zhiding Yu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Intuition for Steepest descent: Close to gradient descent with a choice of norm to be used. If L_2 is chosen, it results exactly in gradient descent.

6.1 Gradient Boosting

Consider constructing a model as a weighted sum of trees (see Figure 6.1) that is used to predict measurements $x_i \in \mathbb{R}^p, i = 1, \dots n$ given the observations $y = (y_1, \dots, y_n) \in \mathbb{R}^n$,

$$\theta_i = \sum_{j=1}^{m} \beta_j . T_j(x_i), \quad i = 1, \dots n \tag{6.1}$$

Each of these trees produces an output x_i given a predictor measurement. Typically, the space of all such possible trees M is exponentially huge. For example, even if we restrict ourselves to a small set of trees of depth 5, we can observe that the number of trees is exponential in number. Therefore, the space is too huge to optimize and we can never get enough observations to solve it analytically.

Given this model, we would like to find the weights β_j that minimizes some loss function L, say squared loss $L(y_i, \theta_i) = (y_i - \theta_i)^2$. Gradient boosting is basically a version of gradient descent forced to work with trees and solve the following optimization problem:

$$\min_{\beta \in \mathbb{R}^M} \sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{M} \beta_j . T_j(x_i)\right) \tag{6.2}$$

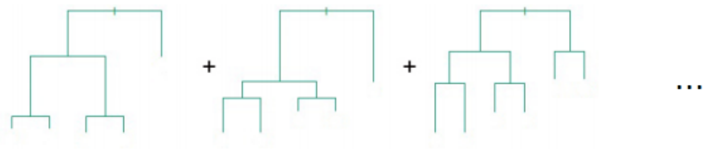

Figure 6.1: Model as a weighted sum of trees

We begin with an initial model, for example, take a single tree $\theta^{(0)} = T_0$. At the k^{th} iteration:

1. Evaluate gradient g at current prediction $\theta^{(k-1)}$,

$$g_i \left[\frac{\partial L(y_i, \theta_i)}{\partial \theta_i}\right]_{\theta_i = \theta_i^{(k-1)}}, \quad i = 1, \dots n \tag{6.3}$$

2. These gradients might not be always be a tree or sum of trees, therefore going away from the constraint space. Gradient boosting now finds the tree that is closest to the negative gradients.

   This problem is not typically hard to solve (approximately) for a single tree. This is the main novelty in gradient boosting.

3. Update the prediction with some learning rate α_k. Observe that such updates will always give predictions that are weighted sum of trees, as desired.

$$\theta^{(k)} = \theta^{(k-1)} + \alpha_k . T_k \tag{6.4}$$

We could also replace trees with any weak learner and derive an equivalent form for the boosted gradient.

Can we do better ?
We have seen that gradient descent in the case of convex, differentiable functions with Lipschitz continuous gradients have a convergence rate of $\mathcal{O}(1/\epsilon)$. However, there is a first-order method that has a convergence rate of $\mathcal{O}(1/\epsilon^2)$.

6.2 Subgradient Basics

Subgradients are very central parts of convex analysis.In non-smooth minimization, subgradients will play important roles.

6.2.1 Recap of Gradient Descent

Consider the following minimization problem:

$$min_x f(x) \tag{6.5}$$

Gradient descent can be used to minimize $f(x)$ iterating the following steps:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, ... \tag{6.6}$$

t_k is the step size at the kth iteration. It can be chosen to be fixed and small, or chosen by backtracking line search to guarantee the reduction of cost function value.

Downsides:

- f has to be differentiable. (Can be fixed by subgradient)
- Slow convergence. (Can be improved by introducing acceleration)

6.2.2 Concept of Subgradient

Subgradient is a generalization of the concept of gradients to non-smooth functions. Given a convex and differentiable function f, its first order approximation using gradient is always an underestimate of f:

$$f(y) \geq f(x) + \nabla^\top (y - x) \quad \text{for all } x, y \tag{6.7}$$

Similarly, a subgradient of a convex function f at x is the set of $g \in R^n$ such that:

$$f(y) \geq f(x) + g^\top (y - x) \quad \text{for all } y \tag{6.8}$$

- For convex functions, such g always exists.
- If f is differentiable at x, then f has a unique subgradient at x which is exactly $\nabla f(x)$.
- Although the same definition of subgradients can also work for nonconvex functions, subgradients may not exist at certain locations, even if they may be smooth.
- Two examples of nonconvex with no subgradients everywhere: $f(x) = -x^2$ and $f(x) = x^3$.

6.2.3 Examples of Subgradients

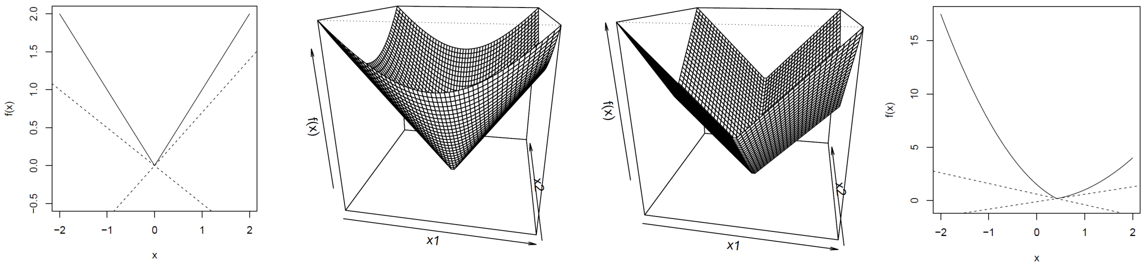

Figure 6.2: Illustration of the subgradients of three example nonsmooth functions. From left to right: 1. absolute value; 2. l-2 norm; 3. l-1 norm. 4. pointwise max of two differentiable convex functions.

Absolute value The function of absolute value has the form: $f : \mathbb{R} \rightarrow \mathbb{R}, f(x) = |x|$. The first image in Fig. 6.2 illustrates the corresponding curve. The function is not differentiable at $x = 0$. Therefore taking the definition of subgradient and substituting $x = 0$ into the definition, one gets:

$$|y| \geq gy \tag{6.9}$$

Consider both cases where $y > 0$ and $y < 0$, we can get $-1 \leq g \leq 1$. Thus:

- For $x \neq 0$, $g = \text{sign}(x)$.
- For $x = 0$, $g \in [-1, 1]$.

l-2 norm The function of l-2 norm has the form: $f : \mathbb{R}^n \to \mathbb{R}, f(x) = \|x\|_2$. The second image in Fig. 6.2 illustrates the corresponding curve. The gradient of the function is $x/\|x\|_2$ at $x \geq 0$ but not well defined at $x = 0$ since the denominator will become 0. Again taking the definition of subgradient and substituting $x = 0$ in to the definition, one gets:

$$\|y\|_2 \geq g^\top y \tag{6.10}$$

Using Cauchy-Schwarz inequality, we have $g^\top y \leq |g^\top y| \leq \|g\|_2\|y\|_2$. To ensure that $g^\top y \leq \|y\|_2$ holds, we need to make sure that $\|g\|_2\|y\|_2 \leq \|y\|_2$ holds, which gives $\|g\|_2 \leq 1$. Thus:

- For $x \neq 0$, $g = x/\|x\|_2$.
- For $x = 0$, $g \in \{g|\|g\|_2 \leq 1\}$.

l-1 norm The function of l-1 norm has the form: $f : \mathbb{R}^n \to \mathbb{R}, f(x) = \|x\|_1$. The third image in Fig. 6.2 illustrates the corresponding curve. For any x where $x_i \neq 0, \forall i$, The ith component g_i of the function gradient g is $\text{sign}(x_i)$.

When there exist certain $x_i = 0$, the function becomes nondifferentiable. The subgradient is simply a multidimensional extension of the absolute value case at each dimension. Therefore the subgradient of l-1 norm is:

- For $x_i \neq 0$, $g_i = \text{sign}(x_i)$.
- For $x_i = 0$, $g_i \in [-1, 1]$.

Pointwise max of two differentiable convex functions The function has the form: $f1, f2 : \mathbb{R}^n \to \mathbb{R}, f(x) = max\{f_1(x), f_2(x)\}$. The fourth image in Fig. 6.2 illustrates the corresponding curve. The function is differentiable at any location where $f_1(x) > f_2(x)$ or $f_1(x) < f_2(x)$. At these locations the subgradient is uniquely equal to the gradient of the larger function. However at locations where $f_1(x) = f_2(x)$, the function becomes nondifferentiable. As a result:

- For $f_1(x) > f_2(x)$, $g = \nabla f_1(x)$.
- For $f_1(x) < f_2(x)$, $g = \nabla f_2(x)$.
- For $f_1(x) = f_2(x)$, $g = t\nabla f_1(x) + (1-t)\nabla f_2(x), 0 \leq t \leq 1$.

6.2.4 Subdifferential

For convex functions, the subdifferential is defined as: $\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$. There are several associated properties:

- For $\partial f(x)$ is closed and convex for both convex and nonconvex functions (directly comes from the definition of subgradient).
- Nonempty (but can be empty for nonconvex f).
- f is differentiable indicates $\partial f(x) = \nabla f(x)$.
- $\partial f(x) = \{g\}$ indicates f is differentiable at x and $\nabla f(x) = g$ (Can use this property to prove the smoothness of some nonobvious convex functions)

6.2.5 Connection to Convex Geometry

Given a convex set $C \subset \mathbb{R}^n$, the indicator function $I_C : \mathbb{R}^n \to \mathbb{R}$ is defined as:

$$I_C(x) = I\{x \in C\} \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \text{ not in } C \end{cases} \tag{6.11}$$

Then for $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x) = \{g \in \mathbb{R}^n : g^\top x \geq g^\top y, \forall y \in C\}$.

6.3 Subgradient Calculus

Subgradients are important for two reasons:

- Convex analysis: Optimality characterization of convex functions via subgradient,monotonicity, close relationship to duality
- Convex optimization: Minimize (almost) any convex function via subgradients

Here, we provide some basic rules for convex functions:

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$ (if $a < 0$, it will turn the function into a concave function)
- Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- Affine composition: if $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$
- Finite pointwise maximum: if $f(x) = max_{i=1,...m} f_i(x)$, then $\partial f(x) = conv(\bigcup_{i:f_i(x)=f(x)} \partial f_i(x))$
  i.e., convex hull of union of subdifferentials of 'active' functions at x
  specialcase : if f_i differentiable, use its differential instead.
- General pointwise maximum: if $f(x) = nax_{s \in S}$, then $\partial f(x) \subseteq cl\{conv(\bigcup_{s:f_s(x)=f(x)} \partial f_s(x))\}$ and under some regularity conditions (on S, f_s), we get an equality above
- Norms: Important special case, $f(x) = ||x_p||$. Let q be such that $1/p + 1/q = 1$, then $||x||_p = max_{||z||_q \leq 1} z^T x$. Hence, $\partial f(x) = argmax_{||z||_q \leq 1} z^T x$

6.4 Optimality Condition

For any f (convex or not),

$$f(x*) = min_x f(x) \Longleftrightarrow 0 \in \partial(x*)$$

i.e., $x*$ is a minimizer iff 0 is a subgradient of f at $x*$. The reason is that $g = 0$ being a subgradient means that for all y,

$$f(y) \leq f(x*) + 0^T(y - x*) = f(x*)$$

Note: The implication for a convex and differentiable function f, with $\partial f(x) = \{\nabla f(x)\}$

6.5 Derivation of First-order Optimality

Recall that for f convex and differentiable, the problem

$$min_x f(x) s.t. x \in C$$

is solved at x iff

$$\nabla f(x)^T (y-x) \leq 0 for all y \in C$$

Intuiviely says that gradient increases as we move away from x.

Proof: First recast problem as: $min_x f(x) + I_C(x)$ Now apply subgradient optimality: $0 \in \partial(f(x) + I_C(x))$ But

$$\begin{gathered} 0 \in \partial(f(x) + I_C(x)) \\ \iff 0 \in \{\nabla f(x)\} + \mathcal{N}_c(x) \\ \iff -\nabla f(x) \in \mathcal{N}_c(x) \\ \iff -\nabla f(x)^T x \leq -\nabla f(x)^T y \text{ for all } y \in C \\ \iff \nabla f(x)^T (y-x) \leq 0 \text{ for all } y \in C \end{gathered}$$

■

Note: the condition $0 \in \{\nabla f(x)\} + \mathcal{N}_c(x)$ is a fully general condition for optimality in a convex problem. But this is not always easy to work with (KKT conditions, later, are easier)

6.6 Example: Distance to a Convex Set

The problem can be formalized as follows:

$$dist(x, C) = min_{y \in C} ||y - x||_2$$

Write $dist(x, C) = ||x - P_C(x)||_2$, where $P_C(x)$ is the projection of x onto C. Then when $dist(x, C) > 0$,

$$\partial dist(x, C) = \left\{ \frac{x - P_C(x)}{||x - P_C(x)||_2} \right\}$$

Only has one element, so in fact $dist(x, C)$ is differentiable and this is its gradient. Thus, we will only show one direction, i.e., that

$$\frac{x - P_C(x)}{||x - P_C(x)||_2} \in \partial dist(x, C)$$

Write $u = P_C(x)$. Then by first-order optimality conditions for a projection,

$$(u - x)^T (y - u) \leq 0 \text{ for all } y \in C$$

Hence

$$C \subseteq H = \{y : (u-x)^T(y-u) \geq 0\}$$

Claim: for any y,

$$dist(y, C) \geq \frac{(x-u)^T(y-u)}{||x-u||_2}$$

Check: first, for $y \in H$, the right-hand side is ≤ 0. Now for $y \notin H$, we have $(x-u)^T(y-u) = ||x-u||_2||y-u||_2 cos\theta$ where θ is the angle between $x-u$ and $y-u$. Thus

$$\frac{(x-u)^T(y-u)}{||x-u||_2} = ||y-x||_2 cos\theta = dist(y, H) \leq dist(y, C)$$

as desired.
Using the claim, we have for any y

$$dist(y, C) \geq \frac{(x-u)^T(y-x+x-u)}{||x-u||_2} = ||x-u||_2 + \left(\frac{x-u}{||x-u||_2}\right)^T(y-x)$$

Hence $g = (x-u)/||x-u||_2$ is a subgradient of $dist(x, C)$ at x.