10-725/36-725: Convex Optimization **Fall 2019**

Lecture 6: September 16

Lecturer: Ryan Tibshirani *Scribes: Vishwak Srinivasan, Ziyang Wang, Arnav Choudhry*

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

6.1 Convergence Analysis

Assume f is convex and differentiable, $\text{dom}(f) = \mathbb{R}^n$ and ∇f is Lipschitz continuous with constant $L > 0$, for any x, y:

$$||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2 \tag{6.1}$$

If f is twice differentiable:

$$\nabla^2 f(x) \preceq LI \tag{6.2}$$

Theorem 6.1 *Gradient descent with fixed step size $t \leq \frac{1}{L}$ satisfies:*

$$f(x^{((k))}) - f^* \leq \frac{||x^{(0)} - x^*||_2^2}{2tk} \tag{6.3}$$

and same result holds for backtracking, with t replaced by $\frac{\beta}{L}$

To find the condition on the step size, we can use the equation $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$ from homework 1 and set $y = x^+ = x - \nabla f(x)t$. Then finding the t we can take to get a decrease in the criterion value, we get $t \leq 1/L$. Gradient descent has convergence rate of $O(\frac{1}{k})$, which means it finds ϵ-suboptimal point in $O(\frac{1}{\epsilon})$ iterations.

6.2 Analysis for strong convexity

Note that f is strongly convex means $f(x) - \frac{m}{2}||x||_2^2$ is convex for some constant $m > 0$. This impies that for a strongly convex function, its curvature is lower bounded by the curvature of the quadratic. If f is twice differentiable. $\nabla^2 f(x) \succeq mI$

Assuming Lipschitz gradient and strong convexity:

Theorem 6.2 *Gradient descent with fixed step size $t \leq \frac{2}{m+L}$ or with backtracking line search satisfies:*

$$f(x^{((k))}) - f^* \leq \gamma^k \frac{L}{2}||x^{(0)} - x^*||_2^2 \tag{6.4}$$

where $0 < \gamma < 1$

- Gradient descent with strong convexity has convergence rate of $O(\gamma^k)$, which means it finds ϵ-suboptimal point in $O(log(\frac{1}{\epsilon}))$ iterations.
- This is called linear convergence as $\frac{||x^{(k)}-x^*||_2}{||x^{(k-1)}-x^*||_2} \leq C < 1$

  [ASIDE] One of the other reasons some people may call this is linear convergence is because the plot of objective versus iteration curve looks linear on semi-log scale. However if $p = 2$ in $||x^{(k-1)} - x^*||_2^p$ in the equation above, it would be called quadratic convergence and so on.
- Important note: $\gamma = O(1 - \frac{m}{L})$, thus the convergence rate is $O(\frac{L}{m} log(\frac{1}{\epsilon}))$. This means that higher condition number $\frac{L}{m}$ results in slower convergence rate. This is due to the Hessian being ellipsoidal and not spherical, so its optimisation is slow.

A look at the conditions for $f(\beta) = \frac{1}{2}||y - X\beta||_2^2$

- Lipschitz continuity of ∇f:
  - $\nabla^2 f(x) \preceq LI$
  - As $\nabla^2 f(\beta) = X^T X, L = \lambda_{\texttt{max}}(X^T X)$
- Strong convexity of f:
  - $\nabla^2 f(x) \succeq mI$
  - As $\nabla^2 f(\beta) = X^T X, m = \lambda_{\texttt{min}}(X^T X)$
  - If X is wide (X is $n \times p$ with $p > n$), $\lambda_{\texttt{min}}(X^T X) = 0$ and f cannot be strongly convex
  - Even if $\sigma_{\texttt{min}}(X) > 0$, we can have large $\frac{L}{m} = \frac{\lambda_{\texttt{max}}(X^T X)}{\lambda_{\texttt{min}}(X^T X)}$
    * If there are correlated features, L/m increases which leads to slow convergence
    * If the features are orthogonal, $L/m = 1$ which leads to fast convergence

Claim. *Gradient Descent always finds regularised solution to the under-parametrised problem.*

Consider the least squares loss $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$. The gradient descent update would be $\beta^{(k)} = \beta^{(k-1)} + tX^T(y - X\beta^{(k-1)})$. Suppose $p > n$, $X\beta = y$ has infinitely many solutions in $\bar{\beta} + \text{null}(X)$. If we set $\beta^{(0)} = 0$, then the solution $\beta^{(k)}$ converges to $\text{argmin}\{\|\beta\|_2 : X\beta = y\}$ as k tends to ∞. The reason for this is that since we started in the row space of X, we will end in the row space of X.

6.3 Practicalities

Stopping rule: stop when $||\nabla f(x)||_2$ is small

- $\nabla f(x^*) = 0$ at solution x^*
- If f is strongly convex with m, $||\nabla f(x)||_2 \leq \sqrt{2m\epsilon} \Rightarrow f(x) - f^* \leq \epsilon$

Pros and Cons of gradient descent:

- Pros:
  - Simple idea, and each iteration is cheap (usually)
  - Fast for well-conditioned, strongly convex problems
- Cons:
  - Can often be slow, because many interesting problems are not strongly convex or well-conditioned
  - Cannot handle nondifferentiable functions

6.4 Nesterov acceleration

Gradient descent has $O(\frac{1}{\epsilon})$ convergence rate over problem class of convex, differentiable functions with Lipschitz gradients.

First-order method: updates $x^{(k)}$ iteratively

$$x^{(0)} + \texttt{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), ..., \nabla f(x^{(k-1)})\} \tag{6.5}$$

Theorem 6.3 (Nesterov) *For any $k \leq \frac{n-1}{2}$ and any starting point $x^{(0)}$, there is a function f in the problem class such that any first-order method satisfies:*

$$f(x^{(k)}) - f^* \geq \frac{3L||x^{(0)} - x^*||_2^2}{32(k+1)^2} \tag{6.6}$$

Can attain convergence rate $O(\frac{1}{k^2})$. Gradient Descent is a type of first-order method, which can be proved using induction. Since Gradient Descent converges at $O(\frac{1}{\epsilon})$, Theorem 6.3 shows that there are more optimal methods than Gradient Descent, which converge at a rate of $O(\frac{1}{\sqrt{\epsilon}})$.

6.5 Analysis for nonconvex case

Assume f is differentiable with Lipschitz gradient, nonconvex. Instead of optimality, we settle for a ϵ-substationary point solution, $||\nabla f(x)||_2 \leq \epsilon$

Theorem 6.4 *Gradient descent with fixed step size $t \leq \frac{1}{L}$ satisfies:*

$$min_{i=0,...,k}||\nabla f(x^{(i)})||_2 \leq \sqrt{\frac{2(f(x^{(0)}) - f^*)}{t(k+1)}} \tag{6.7}$$

- The gradient descent has convergence rate $O(\frac{1}{\sqrt{k}})$, or $O(\frac{1}{\epsilon^2})$
- This rate cannot be improved (over class of differentiable functions with Lipschitz gradients) by any deterministic algorithm.

6.6 Introduction to subgradients

For a convex and differentiable functions f, the first order criterion states that for all x, y:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{6.8}$$

Subgradients are motivated for the case when f is non-differentiable, and are used to define the tighest affine function that underestimates f.

Definition 6.5 (Subgradient) *g is a* ***subgradient*** *of a convex function f at x if*

$$f(y) \geq f(x) + g^T (y - x) \qquad \forall y$$

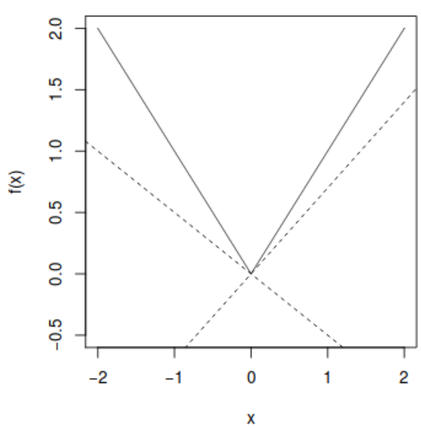
Some properties of subgradients:

- Always exists in the relative interior of the dom(f).
- If f is indeed differentiable at x, then $g = \nabla f(x)$ uniquely.
- This definition is universal - can hold for non-convex functions too. However, it could be possible that g doesn't exist.

6.6.1 Examples

The following examples elucidate the differences about subgradients at points of differentiability and non-differentiability.

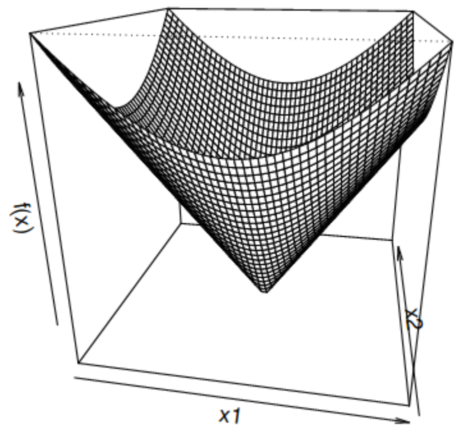
- Consider $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = |x|$. It has one point of non-differentiability, namely at $x = 0$.



  - For $x \neq 0$, the subgradient is unique and is $g = \text{sign}(x)$
  - For $x = 0$, the subgradient is any element of $[-1, 1]$, which can be arrived at by using the definition.

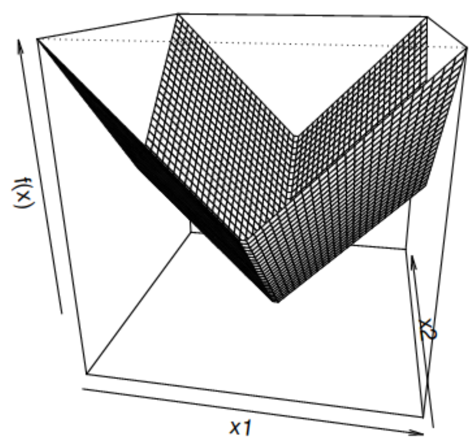
.

- Consider $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = ||x||_2$. It has one point of non-differentiabilty, namely at $x = \mathbf{0}$.



  - For $x \neq \mathbf{0}$, the subgradient is unique and is $g = \frac{x}{||x||_2}$
  - For $x = \mathbf{0}$, the subgradient is any element of $\{v : ||v||_2 \leq 1\}$, which can be arrived at by using the definition.

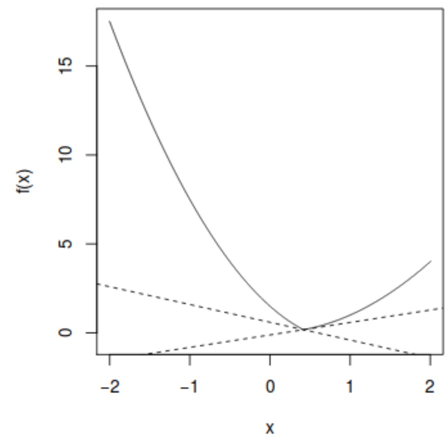
- Consider $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = ||x||_1$. It has more than one point of non-differentiabilty that is when any one of the components equal 0.



  - For $x_i \neq 0$, the i^{th} component of the subgradient is unique and is $g_i = \text{sign}(x_i)$
  - For $x_i = 0$, the i^{th} subgradient is any element of $[-1, 1]$.

  Note that this coincides with the first example when $n = 1$.

- Consider $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = \max\{f_1(x), f_2(x)\}$ where $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ and are convex and differentiable.



- If $f(x) = f_1(x)$ i.e., $f_1(x) > f_2(x)$, then g is unique and is given by $\nabla f_1(x)$.
- If $f(x) = f_2(x)$ i.e., $f_1(x) < f_2(x)$, then g is unique and is given by $\nabla f_2(x)$.
- If $f_1(x) = f_2(x)$, then g is any point on the line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$.

6.7 Subdifferentials

Definition 6.6 (Subdifferential) *The subdifferential of a convex function f at $x \in \mathrm{dom}(f)$ is the collection of all subgradients of f at x*

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x)\}$$

Some properties of the subdifferential:

- For convex f, $\partial f(x) \neq \emptyset$. However, for concave f, $\partial f(x) = \emptyset$.
- $\partial f(x)$ is closed and convex for any f.
- Since the subgradient is unique at points of differentiability, $\partial f(x) = \{\nabla f(x)\}$ when f is differentiable at x.
- $\partial f(x)$ is singleton, then f is differentiable at x and $\nabla f(x)$ is that only element of $\partial f(x)$.

Lemma 6.7 (Connection to Convex Geometry) *Let $C \subseteq \mathbb{R}^n$ be a convex set. Consider $I_C : \mathbb{R}^n \to \mathbb{R}$ such that $I_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$. Then $\partial I_C(x) = \mathcal{N}_C(x)$, where $\mathcal{N}_C(x)$ is the normal cone of C at x.*

Proof: g is a subgradient of I_C at x iff it satisfies the subgradient inequality.

$$I_C(y) \geq I_C(x) + g^T(y - x)$$

If $y \notin C$, then $I_C(y) = \infty$ and the inequality holds trivially. Otherwise if $y \in C$, $I_C(y) = 0$ and the inequality is equivalent to $g^T(y - x) \leq 0$. ■

6.8 Subgradient calculus

Some basic rules for convex functions and their subgradients / subdifferentials:

- *Positive scaling:* $\partial(\alpha f) = \alpha \cdot \partial f$ if $\alpha > 0$
- *Addition:* $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- *Affine composition:* Let $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$
- *Finite pointwise maximum:* Let $f(x) = \max_{i \in [1,m]} f_i(x)$. Then:

  $$\partial f(x) = \text{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$$

  This is a generalization of the example given earlier.
- *Norms:* To each norm $||.||$, there is a **dual norm** $||.||_*$ such that:

  $$||x|| = \max_{||z||_* \leq 1} z^T x$$

  If $f(x) = ||x||_p$, consider q satisfying the relation $\frac{1}{p} + \frac{1}{q} = 1$, then:

  $$||x||_p = \max_{||z||_q \leq 1} z^T x$$

  When $p = 2$, $q = 2$. Also, $\partial f(x) = \operatorname{argmax}_{||z||_q \leq 1} z^T x$.