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

Lecture 5: September 15

Lecturer: Lecturer: Ryan Tibshirani *Scribes: Scribes: Di Jin, Mengdi Wang, Bin Deng*

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

5.1 Canonical convex problem review

Several canonical convex problem can be summarized as below:

- Linear program (LP): $\min_x \quad c^T x$ subject to $Gx \leq h \quad Ax = b$
- Quadratic program (QP): like LP, but with a quadratic objective function
- Semidefinite program (SDP): like LP, but with matrices
- Conic program: the most general form of convex problems
- Second-order cone problem (SOCP): $\min_x \quad c^T x$ subject to $||D_i x + d_i||_2 \leq e_i^T x + f_i \, i = 1, ...p \quad Ax = b$

The relations in these convex problem are summarized as:

$$LP \subseteq QP \subseteq SOCP \subseteq SDP \subseteq Conic\ program$$

In other words, LP is a special case of QP; QP is a special case of SOCP; SOCP is a special case of SDP; all the previous problems are within conic program.

5.2 Gradient descent

5.2.1 Algorithm

Consider the following unconstrained, smooth convex optimization:

$$\min_x \ f(x)$$

where $f(x)$ is convex and differentiable with $dom(f) = \mathbb{R}^n$. Gradient descent is to gradually update x and find x which minimizes the objective function $f(x)$ Then the algorithm is written as:

- choose initial $x^{(0)}$, where $x^{(0)} \in \mathbb{R}^n$

- update $x^{(k)}$ with the following equation repeatedly:

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

  where t_k is step size at step k, $\nabla f(x^{(k-1)}$ is gradient of f plugged with $x^{(k-1)}$

- stop at some point such as decrease of $f(x)$ below some threshold

If $f(x)$ is convex and initializing $x^{(0)}$ at different points, $f(x)$ will converges to its global minimal. If $f(x)$ is strictly convex, the solution will be unique even when initializing $x^{(0)}$ at different points and update paths are different. In contrast, if $f(x)$ is non-convex, $f(x)$ does not have global minimal. When initialization of $x^{(0)}$ at different points will give different results.

5.2.2 Gradient descent interpretation

Gradient descent represent a quadratic approximation to the objective function. For convex and differentiable function $f(y)$, make Taylor expansion at x:

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x)$$

Approximate $f(y)$ by replacing $\nabla^2 f(z)$ with $\frac{1}{t}I$:

$$\begin{aligned} f(y) &\approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}(y - x)^T I (y - x) \\ &= f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}||y - x||_2^2 \end{aligned}$$

To minimize $f(y)$, take derivative against y and set to 0:

$$\nabla f(x) + \frac{1}{t}(y - x) = 0$$

$$y = x - t\nabla f(x)$$

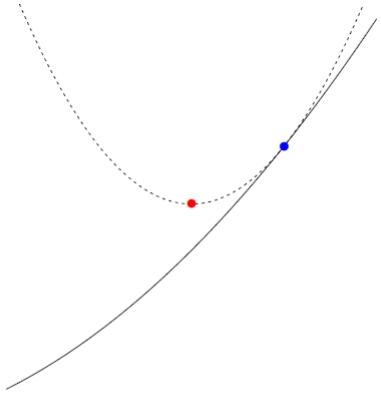
Let $x^+ = y$ denote x for next step:

$$x^+ = x - t\nabla f(x)$$

As show in Figure 5.1, at each point of x, a quadratic function is used for approximation. Then the quadratic function is minimized and x^+ is found as the solution. Set x^+ as next step of x.

5.2.3 Fixed step size

Choosing step size t can affect whether $f(x)$ will converge to its minimal as well as how fast it will converge. As shown in Figure 5.2, consider the function $f(x) = (10x_1^2 + x_2^2)/2$. If t is chosen to be too big (left panel, after 8 steps), $f(x)$ may not decrease and can diverge. In middle panel (after 100 steps), if t is too small, the gradient update moves very slow and it takes many steps to reach the minimal. In right panel it only takes about 40 steps to reach convergence for right step size. Therefore it is important to choose appropriate step size t.



Blue point is x, red point is

$$x^+ = \underset{y}{\operatorname{argmin}} \; f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}\|y - x\|_2^2$$

Figure 5.1: Quadratic approximation in gradient descent

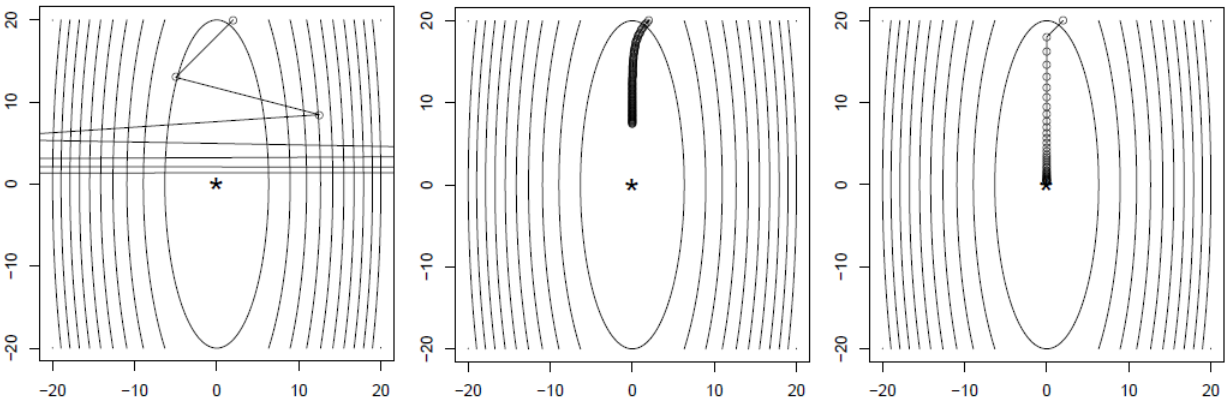
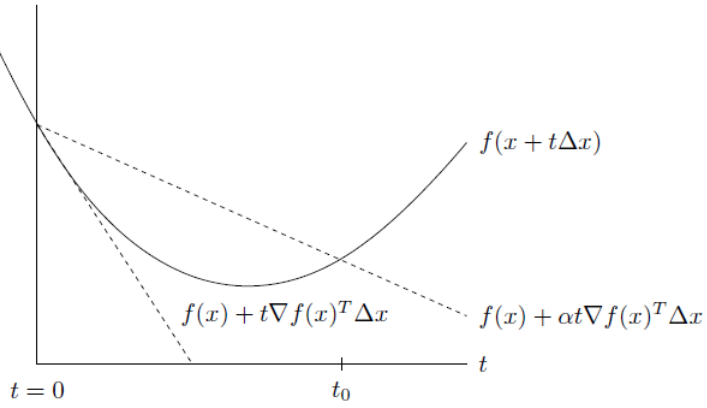


Figure 5.2: Different step size



For us $\Delta x = -\nabla f(x)$

Figure 5.3: Backtracking line search

5.3 Backtracking line search

5.3.1 Algorithm

Backtracking line search allows us to adaptively choose the right step size at each step of gradient update. The method is summarized as:

1. At each iteration, start with step size $t = 1$
2. While
$$f(x - t\nabla f(x)) > f(x) - \alpha t ||\nabla f(x)||_2^2$$
shrink $t = \beta t$, where $0 < \beta < 1$, $0 < \alpha \leq 1/2$ are some fixed parameters
3. Perform gradient update of x:
$$x^+ = x - t\nabla f(x)$$
4. Go to 1 and repeat the following steps until convergence

5.3.2 Backtracking interpretation

For convex and differentiable function $f(x)$, take Taylor expansion at $x - t\nabla f(x)$:

$$f(x - t\nabla f(x)) = f(x) - t\nabla f(x)^T \nabla f(x) + \frac{t^2}{2} \nabla f(x)^T \nabla^2 f(z) \nabla f(x) \geq f(x) - t||\nabla f(x)||_2^2$$

where $\nabla^2 f(z)$ is positive semi-definite matrix due to convexity of $f(x)$. Thus for a convex function it always lies about its tangent line. When t is very small, according to the Taylor expansion, there is:

$$f(x - t\nabla f(x)) \approx f(x) - t\nabla f(x)^T \nabla f(x) = f(x) - t||\nabla f(x)||_2^2 < f(x) - \alpha t||\nabla f(x)||_2^2 \ (0 < \alpha \leq 1/2)$$

By shrinking step size $t = t\beta$, there exists t to make:

$$f(x - t\nabla f(x)) \leq f(x) - \alpha t||\nabla f(x)||_2^2 \ (0 < \alpha \leq 1/2)$$

This is the stopping condition for step size search. The fix parameter α can be considered as acceptable fraction decrease in f predicted by linear exploration. The backtracking line search can be represented with Figure 5.3.

As for selections of parameters α and β, if α is too small, which means we only accept small step size, it may make gradient update very slow. For simplicity, in reality, α can be chosen to be 0.5. β determine how fast and accurate to get the acceptable step size. Big β can cost less rounds of search, but the search can be crude and can shrink step size too much. Small β represent fine search for step size, but may need more rounds of search. It is worth noting that backtracking line search contains inner loop and outer loop. Compared to the gradient descent with right step size, it dose not save steps.

5.4 Exact line search

Exact line search chooses step size to do the best along direction of negative gradient. The algorithm can be summarized as:

1. Initialize $x^{(0)}$
2. Compute the gradient $\nabla f(x^{(k)})$, where $k = 0, 1, 2, 3, ...$
3. For each step, find t such as:
$$t = \underset{s \geq 0}{argmin}\ f(x - s\nabla f(x))$$
4. Update with the following equation:
$$x^{(k+1)} = x^{(k)} - t\nabla f(x^{(k)})$$
5. Go to 2 and repeat until convergence

In practice, it may not be possible to do the minimization in step 4 exactly. In addition, approximations to exact line search are often not much more efficient than backtracking. Thus it is usually not worth the procedures.

5.5 Convergence analysis

Assume that f is convex and differentiable, with $\mathbf{dom(f)} = \mathbb{R}^n$, and additionally for any x and y

$$\|\nabla f(x) - \nabla f(y) \leq L\|x - y\|_2$$

i.e., ∇f is Lipschitz continuous with constant $L > 0$.
Then we have the theorem as following.

Theorem 5.1 *Gradient descent with fixed step size $t \leq 1/L$ satisfies*

$$f(x^{(k)}) - f^* \leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk}) \tag{5.1}$$

Proof: ∇f is Lispschitz continuous with constant $L \Rightarrow$

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2, \quad \text{all } x, y \tag{5.2}$$

(Note: Proof of the inequality in 5.2 is a part of Homework 1).
Assume that y is Gradient descent update and we could get the following equations:

$$y = x^+ = x - t\nabla f(x) \Leftrightarrow \tag{5.3}$$

$$\nabla f(x) = \frac{x - x^+}{t} \tag{5.4}$$

Plugging 5.4 in 5.2, we have

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^T(-t\nabla f(x)) + \frac{L}{2}t^2\|\nabla f(x)\|_2^2 \\ &= f(x) - t\|\nabla f(x)\|_2^2 + \frac{Lt^2}{2}\|\nabla f(x)\|_2^2 \\ &= f(x) - (1 - \frac{Lt}{2})t\|\nabla f(x)\|_2^2 \end{aligned} \tag{5.5}$$

Taking that $0 < t \leq 1/L$, we could obtain the upper bound of 5.5, when $t = 1/L$

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{t}{2}\|\frac{x - x^+}{t}\|_2^2 \\ &= f(x) - \frac{1}{2t}\|x - x^+\|_2^2 \end{aligned} \tag{5.6}$$

Using the convexity of f, we could get

$$\begin{aligned} f(x^*) \geq f(x) - \nabla f(x)^T(x^* - x) \Leftrightarrow \\ f(x) \geq f(x^*) + \nabla f(x)^T(x^* - x) \end{aligned} \tag{5.7}$$

where x^* is the optimal of convex function f.
Taking 5.6 and 5.7, we could get

$$\begin{aligned} f(x^+) &\leq f(x^*) - \nabla f(x)^T(x - x^*) - \frac{1}{2t}\|x^+ - x\|_2^2 \qquad (\nabla f(x) = \frac{x - x^+}{t}) \\ &= f(x^*) - (\frac{x - x^+}{t})^T(x - x^*) - \frac{1}{2t}\|x^+ - x\|_2^2 \\ &= f(x^*) - \frac{1}{t}(x - x^+)^T(x - x^*) - \frac{1}{2t}\|x^+ - x\|_2^2 \\ &= f(x^*) - \frac{1}{2t}(2(x - x^+)^T(x - x^*) - \|x^+ - x\|_2^2) \\ &= f(x^*) - \frac{1}{2t}(2(x - x^+)^T(x - x^*) - \|x^+ - x\|_2^2 + \|x - x^*\|_2^2 - \|x - x^*\|_2^2) \\ &= f(x^*) - \frac{1}{2t}(\|x - x^*\|_2^2 - (\|x - x^*\|_2^2 - 2(x - x^+)^T(x - x^*) + \|x^+ - x\|_2^2) \\ &= f(x^*) + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - x^*\|_2^2) \end{aligned} \tag{5.8}$$

In 5.8, move $f(x^*)$ to the left hand of the equation and obtain

$$f(x^+) - f(x^*) \leq \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - x^*\|_2^2) \tag{5.9}$$

Summing over iterations:

$$\begin{aligned} \sum_{i=1}^{k}(f(x^{(i)} - f(x^*))) &\leq \frac{1}{2t}(\|x^{(0)} - x^*\| - \|x^{(k)} - x^*\|_2^2) \\ &\leq \frac{1}{2t}\|x^{(0)} - x^*\|_2^2 \end{aligned} \tag{5.10}$$

Since $f(x^{(k)})$ is nonincreasing, we have:

$$\begin{aligned} f(x^{(k)}) - f(x^*) &\leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f(x^*)) \\ &\leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk} \end{aligned} \tag{5.11}$$

■

Note: We say gradient descent have convergence rate $\mathcal{O}(1/k)$. In other words, to get $f(x^{(k)}) - f(x^*) \leq \varepsilon$, we need $\mathcal{O}(1/\varepsilon)$ iterations.

5.5.1 Convergence analysis for Backtracking line search

With the same assumptions, f is convex and differentiable, with $\mathbf{dom(f)} = \mathbb{R}^n$, and additionally for any x and y

$$\|\nabla f(x) - \nabla f(y) \leq L\|x - y\|_2$$

i.e., ∇f is Lipschitz continuous with constant $L > 0$.
Then we have the theorem as following.

Theorem 5.2 *Gradient descent with backtracking line search satisfies*

$$f(x^{(k)}) - f^* \leq \frac{\|x^{(0)} - x^*\|_2^2}{2t_{min}k}) \tag{5.12}$$

where $t_{min} = min\{1, \beta/L\}$

Proof: If we do line search with $\alpha = 1/2$ and start with $t = 1$.Selected step size satisfies $t_k \geq t_{min} = min\{1, \beta/L\}$.
From 5.9

$$\begin{aligned} f(x^{(i)}) &\leq f^* + \frac{1}{2t_i}(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2) \\ &\leq f^* + \frac{1}{2t_{min}}(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2) \end{aligned} \tag{5.13}$$

add the upper bounds to get

$$f(x^{(k)}) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^*) \leq \frac{1}{2kt_{min}}\|x^{(0)} - x^*\|_2^2 \tag{5.14}$$

Conclusion: Backtracking line search has the same $1/k$ bound as with constant step size. ■

Note: If β is not too small, then we do not lose much compared to fixed step size. $\beta = 1$ returns the exact rate of fixed step size.

5.5.2 Convergence analysis under strong convexity

Reminder: strong convexity of f means $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for some $m > 0$. If f is twice differentiable, then this implies.

$$\nabla^2 f(x) \succeq mI \quad \text{for any } x \tag{5.15}$$

Sharper lower bound than that from usual convexity:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|_2^2 \quad \text{all } x, y \tag{5.16}$$

Under Lipschitz assumption as above, and also strong convexity, we have the following theorem,

Theorem 5.3 *Gradient descent with fixed step size $t \leq 2/(m + L)$ or with backtracking line search satisfies*

$$f(x^{(k)}) - f^* \leq c^k \frac{L}{2}\|x^{(0)-x^*}\|_2^2 \tag{5.17}$$

where $0 < c < 1$

Proof: If $x^+ = x - t\nabla f(x)$ and $0 < t < 2/(m + L)$:

$$\begin{aligned}
\|x^+ - x^*\|_2^2 &= \|x - t\nabla f(x) - x^*\|_2^2 \\
&= \|x - x^*\|_2^2 - 2t\nabla f(x)^T(x - x^*) + t^2\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m + L})\|x - x^*\|_2^2 + t(t - \frac{2}{m + L})\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m + L})\|x - x^*\|_2^2
\end{aligned} \tag{5.18}$$

Then we have,

$$\|x^{(k)} - x^*\|^2 \leq c^k\|x^{(0)} - x^*\|_2^2, \qquad c = 1 - t\frac{2mL}{m + L} \tag{5.19}$$

For $t = 2/(m + L)$, get $c = \frac{\gamma-1}{\gamma+1}^2$ with $\gamma = L/m$ Bound on function value:

$$f(x^{(k)}) - f^* \leq \frac{L}{2}\|x^{(k)} - x^*\|_2^2 \leq \frac{c^k L}{2}\|x^{(0)} - x^*\|_2^2 \tag{5.20}$$

■

This theorem concludes that the converge rate is $\mathcal{O}(c^k)$. In other words, to get $f(x^{(k)}) - f^* \leq \varepsilon$, need $\mathcal{O}(log(1/\varepsilon))$ iterations.

5.5.3 Example of conditions

Let's look at the conditions for a simple problem, $f(\beta) = \frac{1}{2}\|y - \mathbf{X}\beta\|_2^2$.
Lipschitz continuity of ∇f:

- This means $\nabla^2 f(x) \preceq LI$
- As $\nabla^2 f(\beta) = \mathbf{X}^T\mathbf{X}$

Strong convexity of f:

- This means $\nabla^2 f(x) \succeq mI$
- As $\nabla^2 f(\beta) = \mathbf{X}^T\mathbf{X}$
- If $\mathbf{X}$ is wide–i.e., $\mathbf{X}$ is $n \times p$ with $p > n$– then $\sigma_{min}(\mathbf{X}) = 0$, and f can't be strongly convex.
- Even if $\sigma_{min}(\mathbf{X}) > 0$, can have a very large condition number $L/m = \sigma_{min}(\mathbf{X})/\sigma_{min}(\mathbf{X})$

5.5.4 Practicalities

Stopping rule:stop when $\|\nabla f(x)\|^2$ is small

- Recall $\nabla f(x^*) = 0$ at solution x^*
- If f is strongly convex with parameter m, then

$$\|\nabla f(x)\|_2 \leq \sqrt{(2m\varepsilon)} \Rightarrow f(x)$$

Pros and cons of gradient descent:

- Pro: simple idea, and each iteration is cheap
- Pro: very fast for well-conditional, strong convex problems.
- Con: often slow, because interesting problems aren't strongly convex or well-conditioned.
- Con: can't handle nondifferentiable functions.

References

[L. Vandenberghe, Lecture notes for EE 236C, UCLA, Spring 2011-2012]