THE UNIVERSITY OF TEXAS
AT AUSTIN

EE381V LARGE SCALE OPTIMIZATION

Problem Set 3

Edited by LaTeX

Department of Computer Science

STUDENT
Jimmy Lin
xl5224

COURSE COORDINATOR
Sujay Sanghavi

UNIQUE NUMBER
17350

RELEASE DATE
October 08, 2014

DUE DATE
October 16, 2014

TIME SPENT
5 hours

October 16, 2014

Table of Contents

List of Figures

Part I
Written Problems

1 Gradient descent with diminising step size

Proof. Start from analyzing gradient descent with step size t.

$$\begin{aligned} f(x^+) &= f(x + t\Delta x) && (1)\\ &= f(x - t\nabla f(x)) && (2)\\ &= f(x) - t||\nabla f(x)||_2^2 + \frac{\nabla^2 f(x)}{2} t^2 ||\nabla f(x)||_2^2 && (3)\\ &\leq f(x) - t||\nabla f(x)||_2^2 + \frac{L}{2} t^2 ||\nabla f(x)||_2^2 && (4) \end{aligned}$$

The last step is natural derivation of L-Lipschitz gradient of $f(\cdot)$.
Set gradient of RHS to zero and get the minimum of RHS when $t = \frac{1}{L}$, then we have

$$f(x^+) \leq f(x) - \frac{1}{2L}||\nabla f(x)||_2^2 \tag{5}$$

which can be rewritten as

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\frac{1}{2L}||\nabla f(x^{(k)})||_2^2 \tag{6}$$

Then we recursively apply (6) for infinity many step with summation,

$$f(x^{(+\infty)}) \leq f(x^{(0)}) - \sum_{k=0}^{+\infty} \frac{1}{2L}||\nabla f(x^{(k)})||_2^2 \tag{7}$$

Since $f(\cdot)$ has finite minimum f_{min}, and then

$$f_{min} \leq f(x^{(+\infty)}) \leq f(x^{(0)}) - \sum_{k=0}^{+\infty} \frac{1}{2L}||\nabla f(x^{(k)})||_2^2 \tag{8}$$

That is

$$\sum_{k=0}^{+\infty} \frac{1}{2L}||\nabla f(x^{(k)})||_2^2 \leq f(x^{(0)}) - f_{min} \tag{9}$$

Since $\forall k,\ t^{(k)} < \frac{2}{L}$, then

$$\sum_{k=0}^{+\infty} \frac{1}{4} t^{(k)} ||\nabla f(x^{(k)})||_2^2 < \sum_{k=0}^{+\infty} \frac{1}{2L}||\nabla f(x^{(k)})||_2^2 \leq f(x^{(0)}) - f_{min} \tag{10}$$

Then in terms of the given fact, we have

$$\frac{1}{4}||\nabla f(x^{(k)})||_2^2 \to 0 \text{ a.k.a } \nabla f(x^{(k)}) \to 0 \tag{11}$$

which indicates gradient descent eventually converges to a stationary point. □

2 Gradient descent and non-convexity

Apply eigenvalue decomposition for $f(x) = x^T Q x$, then we have $g(y) = f(x)$, where $g(y) = y^T \Lambda y$ and $\Lambda = diag(\lambda_0, ..., \lambda_n)$ are diagonal matrix with eigen values. Then we just need to analyze the convergence on $g(y)$ since $g(y)$ is just a coordinate change of $f(x)$.
Start from gradient descent update for $g(y)$,

$$\nabla g(y) = \Lambda y \tag{12}$$
$$y^{(k+1)} = y^{(k)} - \eta \Lambda y^{(k)} \tag{13}$$

Let us look at update on each dimension individually,

$$y_i^{(k+1)} = y_i^{(k)} - \eta \lambda_i y_i^{(k)} \tag{14}$$
$$= \underbrace{(1 - \eta\lambda_i)}_{c_i} y_i^{(k)} \tag{15}$$

Let $c_i = (1 - \eta\lambda_i)$, it is obvious that for non-optimal y_i,

$$\text{when } \lambda_i = 0,\ c_i = 1,\ g(y) \text{ does neither converge and diverge.} \tag{16}$$
$$\text{when } \lambda_i > 0,\ c_i < 1,\ g(y) \text{ diverge if } |c_i| < 1, \text{ and converge if } c_i < -1. \tag{17}$$
$$\text{when } \lambda_i < 0,\ c_i > 1,\ g(y) \text{ always diverge whatever step size } \eta \text{ is.} \tag{18}$$

Hence, the characterization of those initial points, for which gradient with any step size will diverge, is those points that are already optimal on dimension i (x_i is already optimal) for eigen value on dimension i is negative ($\lambda_i < 0$). Briefly, initial points x where x_i is optimal for all i, $\lambda_i < 0$.
For other initial points, they will diverge with inappropriate choice of step size. Starting from these initial points can still converge to the global optima with arbitrary step size if and only if for one initial point x, all x_i satisfy: (a) already optimal if $\lambda_i \leq 0$. (b) $-1 < 1 - \eta\lambda_i < 1$ if $\lambda_i > 0$

3 Jacobi Method

Prove that, for a convex continuously differentiable f, and a step size $\alpha = 1/n$ where n is the number of coordinates, the next iterates of the Jacobi method produces a lower function value than x, provided x does not already minimize the function.

Proof. Let $x_i^* = (x_1, ..., x_{i-1}, \bar{x}_i, x_{i+1}, ..., x_n)$. Then we attempt to represent x^+ as convex combination of n points $(x_i^*,\ i = 1, ..., n)$.

$$x^+ = x + \alpha(\bar{x} - x) \tag{19}$$
$$= x + \frac{1}{n}(\bar{x} - x) \tag{20}$$
$$= (1 - \frac{1}{n})x + \frac{1}{n}\bar{x} \tag{21}$$
$$= (\frac{n-1}{n})x + \frac{1}{n}\bar{x} \tag{22}$$
$$= \left(\frac{n-1}{n}x_1 + \frac{1}{n}\bar{x}_1, \frac{n-1}{n}x_2 + \frac{1}{n}\bar{x}_2, ..., \frac{n-1}{n}x_n + \frac{1}{n}\bar{x}_n\right) \tag{23}$$
$$= \sum_{i=1}^{n} \frac{1}{n} x_i^* \tag{24}$$

which presents us the convex combination. Then

$$f(x^+) \leq f(\sum_{i=1}^{n} \frac{1}{n} x_i^*) \tag{25}$$
$$\leq \sum_{i=1}^{n} \frac{1}{n} f(x_i^*) \qquad f \text{ is convex} \tag{26}$$
$$\leq \sum_{i=1}^{n} \frac{1}{n} f(x) \qquad \forall i, f(x_i^*) \leq f(x) \tag{27}$$
$$= f(x) \tag{28}$$

where equality holds when $\forall i,\ x_i = \bar{x}_i$ $(x^+ = x)$, that is, point x does already minimize the function. And in other cases, the next iterate of the Jacobi method produces a lower function value than x. □

4 Step size in Newton

(a) Values of t obtain global convergence

For update with Newton step $\Delta x = \nabla^2 f(x)^{-1}\nabla f(x)$

$$\begin{aligned} f(x^+) &= f(x + t\Delta x) && (29)\\ &= f(x) + \nabla f(x)^T \cdot (t\Delta x) + (t\Delta x)^T \nabla^2 f(x)(t\Delta x) && (30)\\ &= f(x) - t\nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x) + t^2 \nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x) && (31)\\ &= f(x) + (t^2 - t)\lambda^2(x) && (32)\\ &= f(x) + t(t-1)\lambda^2(x) && (33) \end{aligned}$$

where $\lambda^2(x) = \nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x)$. Then in order to obtain global convergence to $x^* = 0$,

$$f(x^+) - f(x) = t(t-1)\lambda^2(x) < 0 \tag{34}$$

Since $f(x) = ||x||^3$ is strongly convex (obvious) with unique minima, then

$$\lambda^2(x) = \nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x) \geq \frac{1}{M}||\nabla f(x)||_2^2 \geq 0 \tag{35}$$

Then since the fixed step size $t > 0$,

$$t - 1 < 0 \tag{36}$$

That is

$$0 < t < 1 \tag{37}$$

Hence, for $0 < t < 1$, we can obtain global convergence to the minimum of $f(x) = ||x||^3$.

(b) Reason that convergence is not quadratic

The convergence is not quadratic because for $0 < t < 1 (t \neq 1)$, the analysis on textbook for rate of convergence does not hold any more. Specifically,

$$\begin{aligned} ||\nabla f(x+)||_2 &= || \int_0^t (\nabla^2 f(x + \eta\Delta x) - \nabla^2 f(x))\Delta x d\eta||_2 && (38)\\ &\leq \frac{Lt}{2}||\Delta x||_2^2 && (39)\\ &\leq \frac{Lt}{2m^2}||\nabla f(x)||_2^2 && (40) \end{aligned}$$

which is not consistent with the formula of quadratic convergence.

5 Composite functions

(a) Run gradient descent on f and g

Show that the entire sequence of iterates will then be the same.

Proof. The gradient descent direction for $f(x)$ and $g(x)$ are respectively

$$\Delta x_{f(x)} = \nabla_x f(x) \tag{41}$$
$$\Delta x_{g(x)} = \nabla_x g(x) = \nabla_x \phi(f(x)) = \nabla_{f(x)} \phi(f(x)) \nabla_x f(x) \tag{42}$$

Apply direction to update rule, we have

$$x^+_{(f)} = x + t^*_{(f)} \nabla_x f(x) \tag{43}$$
$$x^+_{(g)} = x + t^*_{(g)} \nabla_{f(x)} \phi(f(x)) \nabla_x f(x) \tag{44}$$

where the optimal step size for $f(x)$ is

$$t^*_{(f)} = \arg\min_t f(x + t\nabla_x f(x)) \tag{45}$$

and the optimal step size for $g(x)$ is

$$\begin{aligned} t^*_{(g)} &= \arg\min_t g(x + t\nabla_{f(x)} \phi(f(x)) \nabla_x f(x)) && (46)\\ &= \arg\min_t \phi\big(f(x + t\nabla_{f(x)} \phi(f(x)) \nabla_x f(x))\big) && (47)\\ &= \arg\min_t f(x + t\nabla_{f(x)} \phi(f(x)) \nabla_x f(x)) \qquad \phi(\cdot) \text{ is increasing function} && (48) \end{aligned}$$

Now we observe that both step size $t^*_{(f)}$ and $t^*_{(g)}$ can be seen as exact line search of $f(\cdot)$ on point x towards direction $\nabla_x f(x)$. (Note that $\phi(f(x))$ is a scalar.) But two step size has different scale due to the existence of $\phi(f(x))$ on (48). Hence, we have

$$t^*_{(f)} = t^*_{(g)} \nabla_{f(x)} \phi(f(x)) \tag{49}$$

Thus,

$$\begin{aligned} x^+_{(f)} &= x + t^*_{(f)} \nabla_x f(x) && (50)\\ &= x + t^*_{(g)} \nabla_{f(x)} \phi(f(x)) \nabla_x f(x) && (51)\\ &= x^+_{(g)} && (52) \end{aligned}$$

which indicates that one iteration of gradient descent method on $f(\cdot)$ and $g(\cdot)$ starting with the same point x (arbitrarily) will go to the same point ($x^+_{(f)} = x^+_{(g)}$). Recursively applying this derivation, it is proved that the entire sequence of iterates will then be the same. □

(b) Run Newton method on f and g

Not true for Newton Method.

Proof. The Newton step for $f(x)$ and $g(x)$ are respectively

$$\Delta x_{f(x)} = \nabla_x^2 f(x)^{-1} \nabla_x f(x) \tag{53}$$

$$\Delta x_{g(x)} = \nabla_x^2 g(x)^{-1} \nabla_x g(x) \tag{54}$$

$$= \left(\nabla_{f(x)} \phi(f(x)) \nabla_x^2 f(x) + \nabla_{f(x)}^2 \phi(f(x)) I \right)^{-1} \nabla_{f(x)} \phi(f(x)) \nabla_x f(x) \tag{55}$$

where $\nabla_{f(x)} \phi(f(x))$ is a scalar. According to matrix inversion lemma,

$$\Delta x_{g(x)} = \nabla_{f(x)} \phi(f(x)) \underbrace{\left(I - \nabla_x^2 f(x)^{-1} \nabla_{f(x)}^2 \phi(f(x)) \big(I - \nabla_x^2 f(x)^{-1} \nabla_{f(x)}^2 \phi(f(x)) \big) \right)}_{S} (\nabla_x^2 f(x))^{-1} \nabla_x f(x) \tag{56}$$

which indicates that $\Delta x_{f(x)}$ and $\Delta x_{g(x)}$ have different direction since S is non-trivial. Hence, the exact line search on these two function from the same initial point will cause different result. And thus, the entire sequence of iterates will not be the same. □