Lecture 3 : Generalized Quadratics

EE227C . Lecturer: Professor Martin Wainwright . Scribe: Alvin Wan

It doesn't always make sense to talk about distance to solutions. In this case, we can transform between

$$\|x^l - x^*\|_2 \Longleftrightarrow f(x^l) - f(x^*)$$

We call $\|x^l - x^*\|_2$ the solution approximation. We call $f(x^l) - f(x^*)$ the cost approximation. We can show that if we hav ea good solution in terms of $l-2$ distance, we have a good solution up to a constant M. This is left as an exercise to the reader.

$$f(x^l) - f(x^*) \leq \frac{M}{2}\|x^l - x^*\|_2^2$$

We lose a factor of M, but the function value is not too far off.

1 Complexity

We discuss the Iteration complexity v. Flop (floating point operations) complexity:

1. We found last time that iteration complexity is $O(\kappa \log(\frac{\|x^0 - x^*\|_2}{\epsilon})$. What are some issues with this step size $\frac{2}{m+M}$? We probably don't know it. In principle, we can compute this using Q, but diagonalization is most likely as expensive as solving the objective.

2. Flop complexity is iteration complexity multiplied by the number of flops required to compute the gradient.

Consider $\nabla f(x) = Qx - b$. For a $d \times d, Q$ without structure, we have the following complexities:

1. For an arbitrary matrix: $O(d^2)$

2. sparse Q: $nnz(Q)$

3. Least squares: $\|y - Ax\|_2^2$, where $b = A^T y, Q = A^T A$: $O(nd^2)$ (to compute gradient naively). For a fixed x, we can compute $Qx = A^T(Ax)$ in $O(nd)$. We pay $O(nd)$ every time, but with the previous, pre-computation is $O(nd^2)$.

We should definitely exploit structure of Q when we can but here, those details are nonessential in our analysis.

2 Motivation

Considering behavior for quadratics gives us intuition for a broader class of algorithms. So, we look at quadratics more generally, building off of last lecture's bounds.

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ that is convex and differentiable. Suppose as well that we have two other interesting conditions:

a) M-Lipschitz gradient/M-smoothness: $\|\nabla f(x) - \nabla f(y)\|_2 \leq M\|x-y\|_2, \forall x, y \in \mathbb{R}^d$

b) m-strong convexity/monotonicity: $\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq m\|x-y\|_2^2, \forall x, y \in \mathbb{R}^d$

The gradient $\|\nabla f(x) - \nabla f(y)\|_2 = \|Q(x-y)\|_2 \leq M\|x-y\|_2$.

3 Sub-Optimal Bound

Consider our example, and plug in the gradient.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle = (x-y)^T Q(x-y) \geq m\|x-y\|_2^2$$

This notion is saying that our function is sandwiched between quadratics of M and m. Let's consider gradient descent for such a function. We can do this because we assume the function is differentiable.

$$x^{l+1} = x^l - \alpha \nabla f(x^l), l = 0, 1, 2 \ldots$$

The result we get will not be optimal, but we'll have some more machinery to obtain the optimal result. The two conditions smoothness and convexity allow us to analyze gradient descent.

$$\|x^{l+1} - x^*\|_2^2$$

Say $m > 0$. Is the optimal unique? Consider distinct x^* and y^*, where the gradient must vanish at both. We see then that 0 is then grater than a value that's strictly positive. Contradiction, so the optimal is indeed unique.

$$\begin{aligned}
\|x^{l+1} - x^*\|_2^2 &= \|x^{l+1} - x^l + x^l - x^*\|_2^2 \\
&= \|x^{l+1} - x^l - \alpha(\nabla f(x^l) - \nabla f(x^*)) + x^l - x^*\|_2^2 \\
&= \alpha^2\|\nabla f(x^l) - \nabla f(x^*)\|_2^2 - 2\alpha\langle \nabla f(x^l) - \nabla f(x^*), x^l - x^*\rangle + \|x^l - x^*\|_2^2 \\
&\leq \alpha^2 M^2\|x^l - x^*\|_2^2 - 2\alpha m\|x^l - x^*\|_2^2 \\
&= (1 - 2\alpha m + \alpha^2 M^2)\|x^l - x^*\|_2^2
\end{aligned}$$

In step 2, note that the gradient is 0. For the upper bound, we invoke our two assumptions of smoothness and convexity. Combine like terms for the last step. Taking this, we then compute α^*.

$$\alpha^* = \frac{m}{M^2}$$

Plugging in $\alpha = \alpha^*$, we then have that

$$(1 - \frac{m^2}{M^2}) < 1$$

Fitting into the form, we have $(1 - \frac{1}{\kappa^2})$, where $\kappa = \frac{M}{m} \geq 1$. These conditions and some algebra produce a non-trivial value, but we see a slower convergence rate that isn't quite optimal.

4 Equivalent Formulations of M-smoothness

First, let us consider functions from $\mathbb{R}^d \to \mathbb{R}$ that are convex and differentiable. We will examine equivalent characterizations of M-smoothness.

(a) $\|\nabla f(x) - \nabla f(y)\|_2 \leq M\|x - y\|_2, \forall x, y \in \mathbb{R}^d$

(b) $\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq M\|x - y\|_2^2, \forall x, y \in \mathbb{R}^d$

(c) tangent approximation, Taylor series error upper-bound:
$f(y) - \{f(x) + \langle \nabla f(x), y - x\rangle\} \leq \frac{M}{2}\|x - y\|_2^2, \forall x, y \in \mathbb{R}^d$

(d) Taylor series error lower-bound:
$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \geq \frac{1}{2M}\|\nabla f(x) - \nabla f(y)\|_2^2$

(e) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{2M} \|\nabla f(x) - \nabla f(y)\|_2^2$

The list above is not exhaustive, but we'll then proceed by working through some of these proofs. This will introduce several techniques useful in other scenarios. For equivalent formulations of M-strong convexity, flip all inequalities and change $M \to m$.

(a) $\implies$ **(b)**: This is a one line proof.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\|_2 \|x - y\|_2 \leq M \|x - y\|_2^2$$

(b) $\implies$ **(c)**: Use the interpolation method, and set one endpoint to be what you're after. So we'll interpolate along a line between y and x.

$$G(t) = f(x + t(y - x)) - f(x) - \langle \nabla f(x), t(y - x)$$

,where $G(1) =$ LHS of (c), and $G(0) = 0$. So, apply the Fundamental Theorem of Calculus.

$$\begin{aligned} G(1) - G(0) &= \int_0^1 G'(t) dt \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), t(y - x) \rangle \frac{dt}{t} \quad = \int_0^1 (Mt^2 \|y - x\|_2^2) \frac{dt}{t} \\ &= \frac{M}{2} \|x - y\|_2^2 \end{aligned}$$

In the last step, we applied (b).

(c) $\implies$ **(d)**: We will begin by proving a lemma.

Lemma: Let $g : \mathbb{R}^d \to \mathbb{R}$ be differentiable, satisfy condition (c) with optimal value at x^*. Then,

$$g(x) - g(x^*) \geq \frac{1}{2M}\|\nabla g(x)\|_2^2, \forall x \in \mathbb{R}^d$$

Proof: Take $g(x^*) = \min_{y\in\mathbb{R}^d} g(y) \leq \min_{y\in\mathbb{R}^d}\{g(x) + \langle \nabla g(x), y - x\rangle + \frac{M}{2}\|x - y\|_2^2$. This is another way of seeing gradient descent, which regularizes some linear approximation.

$$y^* = x - \frac{1}{M}\nabla g(x) = \inf_{y\in\mathbb{R}^d}\{g(x) + \langle \nabla g(x), y - x\rangle + \frac{M}{2}\|y - x\|_2^2$$

This is telling us that the minimizer is a gradient step, where the step size is $\frac{1}{M}$. Plug in and do calculus to obtain our bound.

Finally, let us use the lemma to prove our bound. For a fixed $x \in \mathbb{R}^d$, define $g_x(z) = f(z) - \langle \nabla f(x), z\rangle$. This is minimized when $\nabla f(z^*) = \nabla f(x)$. In other words, z^* is a minimizer. It may not be unique, but since f is convex, we can attain one of them. Apply the lemma. We can apply, because we simply shifted f by a linear term. In other words, if f satisfies the above conditions and we add a linear term, the conditions still hold. They are unaffected as the bounds are quadratic.

$$g_x(y) - g_x(x) \geq \frac{1}{2M}\|\nabla g_x(y)\|_2^2$$

Rewriting, we have

$$\begin{aligned} g_x(y) - g_x(x) &= f(y) - f(x) - \langle \nabla f(x), y - x\rangle \\ \frac{1}{2M}\|\nabla g_x(y)\|_2^2 &= \frac{1}{2M}\|\nabla f(y) - \nabla f(x)\|_2^2 \end{aligned}$$

(d) $\implies$ **(e)**: Start with the original statement, then flip the roles of x and y. Sum the two together.

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x\rangle &\geq \frac{1}{2M}\|\nabla f(x) - \nabla f(y)\|_2^2 \\ f(x) - f(y) - \langle \nabla f(y), x - y\rangle &\geq \frac{1}{2M}\|\nabla f(y) - \nabla f(x)\|_2^2 \end{aligned}$$

Sum the two and we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{M} \|\nabla f(x) - \nabla f(y)\|_2^2$$

(e) $\implies$ **(a)**: Use Cauchy Schwartz.

Takeaway: Smoothness is upper bounding and convexity is lower-bounding. So we are sandwiching between a quadratic of M and a quadratic of m.

Takeaway: The above lemma has a dual that we will see again. This technique in general is useful as it allows us to relate differences in function values and gradients.