Lecture 5

W. Hebisch

November 16, 2021

1 Unconstrained optimization

1.1 Unconstrained optimization, rate of convergence

To say anything about rate of convergence we need extra assumptions. We need bounds for second derivative. To use similar notation as in following lectures we introduce here Hessian:

$$\langle \nabla^2 f(x)h_1, h_2\rangle = f''(x)(h_1, h_2)$$

that is for fixed x right hand side is a (symmetric) quadratic form in h, while on left hand side $\nabla^2 f(x)$ is a linear operator uniqely defined by the equality above.

Our assumption from previous lecture can be written as

$$-MI \leq \nabla^2 f(x) \leq MI$$

where inequalite means that difference of both sides is a positve definite operator.

As long as f is smooth and domain of f is convex this is equivalent to

$$\|\nabla f(y) - \nabla f(x)\| \leq M\|y - x\|,$$

that is Lipschitz continuity of derivative of f.

Comparing f with quadratic function we get inequality

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{M}{2}\|y - x\|^2.$$

Taking $y = x - \alpha\nabla f(x)$, $\alpha = \frac{1}{M}$ we get

$$f(y) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2$$

which we call descent estimate.

Recall that in previous lecture we showed that $O(\frac{1}{\varepsilon^2})$ steps is enough to decrease gradient to magnitude ε.

Now we want to prove better estimate for convex functions. To simplify arguments we will consider constant step size $\alpha \leq \frac{2}{M}$ where as before M is Lipschitz constant of gradient of f.

First, useful estimate:

Lemma 1.1

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{2}{M}\|\nabla f(x) - \nabla f(y)\|^2 \le f(y),$$

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

Proof. To prove first inequality consider $\phi(y) = f(y) - \langle \nabla f(x), y - x\rangle$. $\nabla\phi(x) = 0$ so ϕ attains minimal value at x, so

$$f(x) = \phi(x) \le \phi(y - \frac{1}{M}\nabla\phi(y)).$$

Since gradient of ϕ has the same Lipschitz constant as f we have descent estimate

$$\phi(y - \frac{1}{M}\nabla\phi(y)) \le \phi(y) - \frac{1}{2M}\|\nabla\phi(y)\|^2.$$

Since $\nabla\phi(y) = \nabla f(y) - \nabla f(x)$ this gives first estimate.

Adding first estimate for x and y and with reversed order gives second estimate. □

Now, we will prove that under such condition distance to optimal point can not increase.

Lemma 1.2 *With assumptions as above*

$$\|x_{i+1} - x_\infty\| \le \|x_i - x_\infty\|$$

Proof.

$$\begin{aligned}\|x_{i+1} - x_\infty\|^2 &= \|x_i - x_\infty - \alpha\nabla f(x_i)\|^2 \\ &= \|x_i - x_\infty\|^2 - 2\alpha\langle \nabla f(x_i), x_i - x_\infty\rangle + \alpha^2\|\nabla f(x_i)\|^2 \\ &\le \|x_i - x_\infty\|^2 - \alpha\frac{2}{M}\|\nabla f(x_i)\|^2 + \alpha^2\|\nabla f(x_i)\|^2 \\ &\le \|x_i - x_\infty\|^2\end{aligned}$$

as long as $\alpha \le \frac{2}{M}$. □

Now, we can prove:

Lemma 1.3 *Let f be convex such that gradient of f is Lipschitz continuous with constant M. Gradient descent using constant step size $\alpha = \frac{1}{M}$ satisfies*

$$f(x_m) - f(x_\infty) \le \frac{2M\|x_0 - x_\infty\|^2}{m + 4}$$

where x_∞ is a minimizer of f.

Proof: By convexity

$$f(x_\infty) - f(x_i) \geq \langle \nabla f(x_i), x_\infty - x_i \rangle \geq -\|\nabla f(x_i)\| \|x_\infty - x_i\|$$

so

$$\|\nabla f(x_i)\| \geq \frac{f(x_i) - f(x_\infty)}{\|x_i - x_\infty\|} \geq \frac{f(x_i) - f(x_\infty)}{\|x_0 - x_\infty\|}$$

Combining this with descent estimate we get

$$f(x_{i+1}) - f(x_i) \leq -\frac{(f(x_i) - f(x_\infty))^2}{2M\|x_0 - x_\infty\|^2}$$

Writing $\delta_i = f(x_i) - f(x_\infty)$ this means

$$\delta_{i+1} - \delta_i \leq -\frac{\delta_i^2}{2M\|x_0 - x_\infty\|^2}$$

so

$$\delta_{i+1} \leq \delta_i - \frac{\delta_i^2}{2M\|x_0 - x_\infty\|^2}$$

Now the we get result by induction. It is clearly true for $i = 0$. The right hand side is a quadratic in δ_i, which attains maximal value at $\delta_{max} = M\|x_0 - x_\infty\|^2$. By inductive assumption $\delta_i \leq \frac{2M\|x_0 - x_\infty\|^2}{i+4}$, which is smaller than δ_{max}, so our quadratic is increasing for relevant δ_i and we get estimate from above plugging in upper estimate for δ_i.

Consequently

$$\delta_{i+1} \leq \frac{2M\|x_0 - x_\infty\|^2}{i+4} - \frac{2M\|x_0 - x_\infty\|^2}{(i+4)^2}$$

$$= 2M\|x_0 - x_\infty\|^2 (\frac{1}{i+4} - \frac{1}{(i+4)^2})$$

$$\leq 2M\|x_0 - x_\infty\|^2 (\frac{1}{i+4} - \frac{1}{(i+4)(i+5)})$$

$$= \frac{2M\|x_0 - x_\infty\|^2}{i+5}.$$

□

Remark. We get similar result for smaller steps. However, when $\alpha > \frac{2}{M}$ our proof that $\|x_i - x_\infty\|$ is nonincreasing no longer works. For $\frac{1}{M} < \alpha \leq \frac{2}{M}$ we get worse estimate of descent. In practice larger steps are likely to give faster convergence, but theory suggests small steps. So there is a discrepancy. We will see that similar discrepancy appears in different situations.

Lemma 1.4 *With assumptions as above we have*

$$\min_{m/2\leq i<m} \|\nabla f(x_i)\| \leq \frac{4M\|x_0 - x_\infty\|}{m+1}.$$

Proof: Applying our general gradient estimate from previous lecture with x_0 replaced by x_{m_0} where $m_0 = m/2$ (rounded up) we get

$$\min_{m_0\leq i<m} \|\nabla f(x_i)\|^2 \leq \frac{2M(f(x_{m_0}) - f(x_\infty))}{m - m_0 + 1}.$$

Using estimate for $f(x_{m_0}) - f(x_\infty)$ this is

$$\leq \frac{2M(2M\|x_0 - x_\infty\|^2)}{(m_0+4)(m-m_0+1)}$$

$$\leq \frac{(4M\|x_0 - x_\infty\|)^2}{(m+1)^2}$$

which gives the claim. □

Clearly, this is better then non-convex gradient estimate.

Now we come to lower bound.

We will say that a method is first order method if

$$x_{i+1} \in x_0 + \mathrm{lin}\{\nabla f(x_0), \dots, \nabla f(x_i)\}.$$

Intuitively, the condition above means that we use only information obtained from first derivatives to choose direction. Clearly gradient descent is a first order method, but we will see that most methods that we study are first order methods.

Lemma 1.5 *For any k, $1 \leq k \leq \frac{n-1}{2}$, and any $x_0 \in \mathbb{R}^n$ there exists a convex function $f : \mathbb{R}^n \to \mathbb{R}$ such that gradient of f is Lipschitz continuous with constant M and for any first-order method we have*

$$f(x_k) - f(x_\infty) \geq \frac{3M\|x_0 - x_\infty\|^2}{32(k+1)^2},$$

$$\|x_k - x_\infty\|^2 \geq \frac{1}{8}\|x_0 - x_\infty\|^2$$

where x_∞ is the minimum of f.

There is a method (Nesterow acceleration) for which we have corresponding upper bound.

Proof. To prove the lemma we need to construct appropriate couterexample. It is enough to do this for $x_0 = 0$.

For $s \in \mathbb{R}^n$ we define

$$f(s) = \frac{1}{2}\left(s_1^2 + s_n^2 + \sum_{i=1}^{n-1}(s_{i+1} - s_i)^2\right) - s_1.$$

Clearly $\nabla^2 f$ is strictly positive definite. One can check that $\nabla f(s) = 0$ when

$$s_i = 1 - \frac{i}{n+1}$$

so optimal value is $\frac{1}{2}(-1 + \frac{1}{n+1})$.

Let V_i be subspace of $\mathbb{R}^n$ consisting from vectors such that only first i coordinates are nonzero ($V_0 = \{0\}$). One can check that when $x_i \in V_i$, then $\nabla f(x_i) \in V_{i+1}$ so for any first order method starting at $x_0 = 0$ we have $x_i \in V_i$.

At best first order metod will give optimal value on V_i. However, f restricted to V_i looks exactly like f with $n = i$, so optimal value is $\frac{1}{2}(-1 + \frac{1}{i+1})$ and therefore error

$$E(x_i) = \frac{1}{2}(-1 + \frac{1}{i+1}) - \frac{1}{2}(-1 + \frac{1}{n+1})$$
$$= \frac{1}{2}(\frac{1}{i+1} - \frac{1}{n+1})$$

and in particular when $2(i+1) \le n+1$ then

$$E(x_i) \ge \frac{1}{4}\frac{1}{i+1}.$$

Also, note that

$$\|x_i\|^2 = \sum_{j=1}^{i}(1 - \frac{j}{i+1})^2 = \sum_{j=1}^{i}\frac{j^2}{(i+1)^2}$$
$$\le \sum_{j=1}^{i}\frac{(j+1)^3 - j^3}{3(i+1)^2} = \frac{(i+1)^3}{3(i+1)^2} - \frac{1}{3(i+1)^2} \le \frac{i+1}{3}$$

so

$$E(x_i) \ge O(\frac{1}{(i+1)^2})\|x_0 - x_\infty\|^2$$

Moreover,

$$\|x_i - x_\infty\|^2 \ge c\|x_0 - x_\infty\|^2$$

□

Previous lemma means that even for quite reqular convex functions we can not expect very fast convergence. In fact, there is already problem with convex quadratic functions.

Above we have difficulty because f is convex, but second derviative may be very small in some directions (and relatively large in other directions). So we need lower bound on second derivative. Technically it is more elegant to use following condition.

We say that f is *strongly convex with constant* m when $f - m\|x\|^2$ is convex.

This condition holds when f is twice differentiable and all eigenvalues of $\nabla^2 f$ are bounded from below by m. However, as written above condition does not require existence of second deriviative.

Lemma 1.6 *When f is strongly convex with constant m, that is $f - m\|x\|^2$ is convex, gradient of f is Lipschitz continuous with constant M, then for gradient descent with constant step size $\alpha = \frac{2}{M+m}$ we have*

$$\|x_i - x_\infty\| \le C^i \|x_0 - x_\infty\|$$

where

$$C = \frac{M-m}{M+m}$$

and x_∞ is a minimal point.

Proof: We will represent ∇f using $\nabla^2 f$ (which is possible under our assumptions, in particular $\nabla^2 f$ exists almost everywhere):

$$\nabla f(x_i) = \nabla f(x_i) - \nabla f(x_\infty) = \int_0^1 \nabla^2 f(x + th_i) h_i$$

where $h_i = x_i - x_\infty$. So

$$h_{i+1} = x_{i+1} - x_\infty = x_i - x_\infty - \alpha \nabla f(x_i)$$

$$= (I - \alpha \int_0^1 \nabla^2 f(x + th_i)) h_i = Ah_i$$

By assumption $mI \le \nabla^2 f(x + th_i) \le MI$, so

$$\frac{2m}{M+m} I \le \alpha \nabla^2 f(x + th_i) \le \frac{2M}{M+m} I.$$

So

$$-\frac{M-m}{M+m} I \le A \le \frac{M-m}{M+m} I$$

and since A is symmetric

$$\|A\| \le \frac{M-m}{M+m}.$$

Consequently

$$\|x_{i+1} - x_\infty\| \le \|A\| \|x_i - x_\infty\| \le C \|x_i - x_\infty\|$$

and claim follows by induction. □

Remark: Again, slightly enlarging C we get result for steps smaller than $\frac{2}{M+m}$ and for slightly larger steps. But the proof breaks down when steps are much larger.

We also get result for improvement of goal function:

Lemma 1.7 *With assumptions as above and step* $\alpha = \frac{1}{M}$

$$f(x_{i+1}) - f(x_\infty) \leq C(f(x_i) - f(x_\infty))$$

where $C = 1 - \frac{m}{M}$.

Proof: From decay estimate

$$f(x_{i+1}) \leq f(x_i) - \frac{1}{2M}\|\nabla f(x_i)\|^2$$

so

$$f(x_{i+1}) - f(x_\infty) \leq f(x_i) - f(x_\infty) - \frac{1}{2M}\|\nabla f(x_i)\|^2$$

By strong convexity

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|^2$$

Computing derivative we see that $z = x - \frac{1}{m}\nabla f(x)$ minimizes right hand side (with fixed x and variable y).

So

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

Using $y = x_\infty$ and $x = x_i$ we get

$$\frac{1}{2m}\|\nabla f(x_i)\|^2 \geq f(x_i) - f(x_\infty)$$

Plugging the above into decay estimate we get

$$\begin{aligned} f(x_{i+1}) - f(x_\infty) &\leq f(x_i) - f(x_\infty) - \frac{1}{2M}2m(f(x_i) - f(x_\infty)) \\ &= C(f(x_i) - f(x_\infty)) \end{aligned}$$

with $C = 1 - \frac{m}{M}$ which gives the claim. □

Remark: clearly exact line search is as good or better than this. Armijo's rule gives at least fraction of decay of fixed step, so we get similar result with slightly larger C.

Remark: When $\frac{m}{M}$ is reasonably big this is much better than result without strong convexity: we need number of steps that grows linearly with accuracy while $\frac{1}{\epsilon}$ is exponential in number of bits.

What to do when $\frac{m}{M}$ is very small? It may happen that $\frac{m}{M}$ is small due to bad scaling, this happened in quadratic example from previous lecture. Simple diagonal scaling sometimes helps, but rotating bad example we see that in general we may need to rescale by arbitrary linear mapping A. In a sense optimal rescaling would give

$$\|Ah\|^2 = \langle \nabla^2 f(x)h, h\rangle$$

We can not get this for all x simultaneously, but can do for single point x_i. Namely, we define new scalar product as $\langle h, h\rangle_x = \langle \nabla^2 f(x)h, h\rangle$. Then steepest descent direction is given by Newton formula

$$(\nabla^2 f(x_i))^{-1}\nabla f(x_i).$$

1.2 Newton method

When $x_{i+1} = x_i - (\nabla^2 f(x_i))^{-1}\nabla f(x_i)$ we say about pure Newton method.

To derive Newton formula we find steepest descent direction at x minimizing

$$\frac{\langle \nabla f(x), h\rangle}{\langle h, h\rangle_x^{1/2}} = \frac{\langle \nabla f(x), h\rangle}{\langle \nabla f^2(x)h, h\rangle^{1/2}}.$$

Computing derivative with respect to h we get

$$\nabla f(x)\langle \nabla^2 f(x)h, h\rangle^{1/2} - \langle \nabla f(x), h\rangle \nabla^2 f(x)h$$

as numerator. Comparing this to 0 gives equation

$$\langle \nabla^2 f(x)h, h\rangle^{1/2}\nabla f(x) = \langle \nabla f(x), h\rangle \nabla^2 f(x)h$$

that is

$$h = \frac{\langle \nabla^2 f(x)h, h\rangle^{1/2}}{\langle \nabla f(x), h\rangle}(\nabla^2 f(x))^{-1}\nabla f(x).$$

Note that

$$\frac{\langle \nabla^2 f(x)h, h\rangle^{1/2}}{\langle \nabla f(x), h\rangle}$$

is just a negative constant, so in fact we get $(\nabla^2 f(x))^{-1}\nabla f(x)$ as the steepest descent direction.

Classically, Newton method was obtained looking at quadratic approximation to $f(x+h)$:

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h\rangle + o(\|h\|^2).$$

Minimizing quadratic function of h on the right hand side gives

$$h = -(\nabla^2 f(x))^{-1}\nabla f(x).$$

So we get iterative algorithm

$$x_{i+1} = x_i - \alpha(\nabla^2 f(x))^{-1}\nabla f(x).$$

$\alpha = 1$ is pure Newton method, when α may be smaller than 1 we have damped Newton method.

Newton method for optimization is closely related to Newton method for equation solving, namely Newton method for optimization of f produces the same approximations as Newton method for solving equation $\nabla f(x) = 0$. When solving equations it is easy to see that pure Newton method may diverge, that is approximations converge to a cycle or are chaotic. The same may happen when using pure Newton method for optimization.

Consider now $g(y) = f(Ay + b)$ where A is an invertible matrix. We have

$$\langle \nabla g(y), h\rangle = \langle \nabla f(Ay + b), Ah\rangle,$$

$$g''(y)(h, h) = f''(Ay + b)(Ah, Ah)$$

so

$$\nabla g(y) = A^T \nabla f(Ay + b),$$
$$\nabla^2 g(y)h = A^T(\nabla^2 f)(Ay + b)Ah$$

and

$$\begin{aligned}(\nabla^2 g(y))^{-1}\nabla g(y) &= A^{-1}(\nabla^2 f(Ay + b))^{-1}(A^T)^{-1}A^T\nabla f(Ay + b)\\ &= A^{-1}\nabla^2 f(Ay + b))^{-1}\nabla f(Ay + b)\end{aligned}$$

Rewriting, we get

$$A(\nabla^2 g(y))^{-1}\nabla g(y) = (\nabla^2 f(Ay + b))^{-1}\nabla f(Ay + b).$$

Let y_i be approximations produced by Newton method applied to g and let $w_i = (\nabla^2 g(y_i))^{-1}\nabla g(y_i)$ be corresponding search directions. Put $x_i = Ay_i + b$. Our result shows that starting Newton method for f at x_i we get $h_i = Aw_i$ as as search direction. Clearly

$$g(y_i + \alpha w_i) = f(A(y_i + \alpha w_i) + b) = f(Ay_i + b + \alpha Aw_i) = f(x_i + \alpha h_i)$$

so we use the same function of α during line search, so we will get the same α_i. Consequently, starting Newton method for f at x_0 we will get x_i at step i.

In other words, Newton method is invariant under affine transformations: changing variables in affine way changes approximations produced by Newton method in the same way. This is quite different than gradient descent, where change of variables plays much more role.

Note: to have true invariance we should have invariant stopping criterion. Good criterion is given by smallness of step h_i using our scalar product

$$\begin{aligned}\langle h_i, h_i\rangle_{x_i} = \langle \nabla^2 f(x_i) h_i, h_i\rangle = -\langle \nabla^2 f(x_i)(\nabla^2 f(x_i))^{-1}\nabla f(x_i), h_i\rangle \\ = -\langle \nabla f(x_i), h_i\rangle.\end{aligned}$$

Using our formulas we have

$$-\langle \nabla g(y_i), w_i\rangle = -\langle A^T \nabla f(Ay_i + b), A^{-1}h_i\rangle = -\langle \nabla f(x_i), h_i\rangle$$

so requirement $-\langle \nabla f(x_i), h_i\rangle < \varepsilon$ gives invariant stopping criterion.

1.3 Further reading

Stephen Boyd, Lieven Vandenberghe, Convex Optimization, chapter 9.

David G. Luenberger, Yinyu Ye, Linear and Nonlinear Programming, chapters 7 and 8.

Yurii Nesterov, Introductory lectures on convex optimization, Springer 2004, chapter 1 (despite title more advanced than other texts).

Jorge Nocedal, Stephen J. Wright, Numerical Optimization, chapter 3.