Lecture 7–8: Other Basic Descent Methods

Yudong Chen

1 Analysis of gradient descent, cont'd

Consider the gradient descent (GD) iteration with *constant* stepsize:

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \qquad \forall k = 0, 1, \ldots$$

where f is L-smooth for $L < \infty$ and $\mathcal{X} = \mathbb{R}^d$.

Lemma 1 (Descent Lemma). *If $x_{k+1} = x_k - \alpha \nabla f(x_k), \alpha \in (0, \frac{1}{L}]$, then*

$$f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2 .$$

1.1 The strongly convex case

Assume f is m-strongly convex and L-smooth, and $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x)$. For all k:

$$\begin{aligned}
f(x^*) &\ge f(x_k) + \left\langle \underbrace{\nabla f(x_k)}_{\frac{1}{\alpha}(x_k - x_{k+1})}, x^* - x_k \right\rangle + \frac{m}{2} \|x^* - x_k\|_2^2 && \text{by strong convexity} \\
&\ge f(x_{k+1}) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2 + \frac{m}{2} \|x^* - x_k\|_2^2 && \text{same argument as Lec 6 convex case} \\
&= f(x_{k+1}) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \left(\frac{1}{2\alpha} - \frac{m}{2} \right) \|x_k - x^*\|_2^2 .
\end{aligned}$$

Rearranging:

$$\frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 \le \left(\frac{1}{2\alpha} - \frac{m}{2} \right) \|x_k - x^*\|_2^2 + \underbrace{f(x^*) - f(x_{k+1})}_{\le 0},$$

so

$$\|x_{k+1} - x^*\|_2^2 \le (1 - m\alpha) \|x_k - x^*\|_2^2 .$$

When $\alpha \le \frac{1}{L}$, we know that $m\alpha \in (0, 1]$ since $m \le L$. Therefore, we have

$$\|x_{k+1} - x^*\|_2^2 \le (1 - m\alpha)^{k+1} \|x_0 - x^*\|_2^2 , \qquad \longleftarrow \text{convergence rate} \tag{1}$$

i.e., geometric convergence (a.k.a. "linear convergence" in optimization literature.)

Equivalently, $\|x_{k+1} - x^*\|_2 \le \epsilon$ after at most

$$O\left(\frac{1}{m\alpha} \log \left(\frac{\|x_0 - x^*\|_2}{\epsilon} \right) \right) \text{ iterations.} \qquad \longleftarrow \text{iteration complexity} \tag{2}$$

Compare with previous two cases.

Exercise 1. For a simple quadratic function $f(x) = \|x - x^*\|_2^2$, all measures of optimality (optimality gap $f(x) - f(x^*)$, gradient norm $\|\nabla f(x)\|_2^2$ and distance to optimum $\|x - x^*\|_2^2$) are equivalent up to constants. Try to prove that the same is true for a function that is both strongly convex and smooth, as such a function is sandwiched between two quadratics. With this in mind, you can try to further prove geometric convergence in terms of the function value:

$$f(x_{k+1}) - f(x^*) \leq (1 - m\alpha)^{k+1} (f(x_0) - f(x^*)) .$$

How about $\|\nabla f(x_{k+1})\|_2$?

Remark 1. The bounds in (1) and (2) depend on $m\alpha$, which equals $\frac{m}{L}$ if we take $\alpha = \frac{1}{L}$. Note that $\frac{L}{m}$ is (an upper bound of) the condition number of the Hessian $\nabla^2 f$. Fast convergence if $\nabla^2 f$ is well-conditioned.

1.2 Unknown L

All previous analysis is valid when we use a stepsize $\alpha \leq \frac{1}{L}$, which requires knowing L, or at least an upper bound of L. How to choose α if we don't know L?

1.2.1 Trial and error

For example:

- Choose the largest α for which GD does not diverge.
- Use your lucky number as the initial value of α. Adjust and see if it works better.

The second option is popular among machine learning practitioners. For example, PyTorch, a popular package for training neural networks, implements several variants of GD with default stepsizes like 0.01 or 0.001, which is the starting point for most users.

1.2.2 Exact line search

Choose α as the solution to the *one-dimensional* optimization problem

$$\min_{\alpha > 0} f\left(x_k - \alpha \nabla f(x_k)\right) .$$

That is, we find the exact minimum of f along the half line $\{x_k - \alpha \nabla f(x_k) : \alpha > 0\}$.

This method is most useful when f has some special structure so that the above 1-D problem can be solved efficiently at low cost.

1.2.3 Backtracking line search

Start with some initial α_0. Sequentially try stepsizes $\alpha_0, \frac{1}{2}\alpha_0, \frac{1}{4}\alpha_0, \frac{1}{8}\alpha_0, \ldots$ until the descent condition

$$f\left(x_k - \alpha \nabla f(x_k)\right) \leq f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2$$

is satisfied. Backtracking terminates before or when $\frac{1}{2^t}\alpha_0 \leq \frac{1}{L}$ is satisfied for the first time, so it requires no more than $O\left(\log\left(\alpha_0 L\right)\right)$ function evaluations of f and one gradient computation at x_k.

This method is useful when function evaluation is easy but solving the exact linear search problem is costly.

2 Other descent methods

There are other descent methods for which the conclusion of the Descent Lemma holds.

Examples:

1. **Preconditioned methods:**
$$x_{k+1} = x_k - \alpha S_k \nabla f(x_k),$$
where S_k is a symmetric positive definite matrix with all eigenvalues in $[\gamma_1, \gamma_2]$, $0 < \gamma_1 < \gamma_2 < \infty$.
From properties of L-smooth functions (Lemma 1 in Lecture 4):
$$\begin{aligned} f(x_{k+1}) &\le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) - \alpha \underbrace{\langle S_k \nabla f(x_k), \nabla f(x_k) \rangle}_{\ge \gamma_1 \|\nabla f(x_k)\|_2^2} + \frac{L}{2}\alpha^2 \underbrace{\|S_k \nabla f(x_k)\|_2^2}_{\le \gamma_2^2 \|\nabla f(x_k)\|_2^2} \\ &\le f(x_k) - \underbrace{\left(\alpha\gamma_1 - \frac{L}{2}\gamma_2^2\alpha^2\right)}_{>0 \text{ for sufficiently small } \alpha} \|\nabla f(x_k)\|_2^2. \end{aligned}$$
Newton's method uses $S_k = \left(\nabla^2 f(x_k)\right)^{-1}$; need $\nabla^2 f(x_k)$ to have positive eigenvalues for this to work.
With appropriately chosen S_k, preconditioned methods can converge substantially faster *near* x^* than GD.

2. **Gauss-Southwell (aka greedy coordinate descent):**
$$x_{k+1} = x_k - \alpha \underbrace{\nabla_{i_k} f(x_k) e_{i_k}}_{-p_k}$$
where $i_k = \arg\max_{1 \le i \le d} \{-\nabla_i f(x_k)\}$, and $e_{i_k} = [0, 0, \ldots, \underbrace{1}_{i_k \text{ position}}, \ldots, 0]$ is the i_k-th standard basis vector in $\mathbb{R}^d$. Note that
$$\|p_k\|_2^2 \ge \frac{1}{d} \|\nabla f(x_k)\|_2^2,$$
hence one can show that (exercise) for $\alpha = \frac{1}{L}$,
$$f(x_{k+1}) \le f(x_k) - \frac{1}{2Ld} \|\nabla f(x_k)\|_2^2.$$
This algorithm is most useful when i_k and $\nabla_{i_k} f(x_k)$ are much easier to compute than the full gradient $\nabla f(x_k)$.
Can be viewed as steepest descent w.r.t. ℓ_1 norm.

3. **Randomized coordinate descent.** Similar to above, except that i_k is chosen uniformly at random from $\{1, 2, \ldots, d\}$. See HW2.

4. **Stochastic gradient descent (SGD),** where

$$x_{k+1} = x_k - \alpha g(x_k, \xi_k),$$

where ξ_k's are i.i.d. random variable satisfying $\mathbb{E}_{\xi_k}[g(x_k, \xi_k)] = \nabla f(x_k)$. That is, $g(x_k, \xi_k)$ is an unbiased (but potentially very noisy) estimate of the true gradient at x_k. Under certain assumptions it satisfies the descent condition *in expectation*. We will discuss SGD later this semester.

5. **Gradient descent w.r.t. ℓ_p norm**, where

$$x_{k+1} = \arg\min_u \left\{ f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\alpha} \|u - x_k\|_p^2 \right\}.$$

See HW2.

6. **Mirror descent,** where

$$x_{k+1} = \arg\min_{u \in \mathcal{X}} \left\{ f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\alpha_k} D_\psi(u, x_k) \right\},$$

and $D_\psi(\cdot, \cdot)$ is the Bregman divergence generated by a convex function ψ. See HW2.

3 Convergence of descent methods

Consider any iterative method that generates a sequence $x_0, x_1, \ldots$ satisfying the descent condition

$$f(x_{k+1}) \le f(x_k) - \frac{\beta}{2} \|\nabla f(x_k)\|_2^2, \quad \forall k \ge 0 \tag{3}$$

for some $\beta > 0$.

3.1 General case

Assume f is bounded from below: $f(x) \ge f_* > -\infty, \forall x$. The same analysis from the previous lecture applies since the analysis only uses the descent property (3). This gives

$$\min_{0 \le i \le k} \|\nabla f(x_i)\|_2 \le \sqrt{\frac{2(f(x_0) - f_*)}{\beta(k+1)}}.$$

However, the analysis for the convex and the strongly convex cases of gradient descent does not immediately transfer to other descent methods. There, we crucially used that the update was of the form $x_{k+1} = x_k - \alpha \nabla f(x_k)$. Below, we show that it is still possible to obtain similar (though slightly weaker) guarantees as for gradient descent if we are only assuming that our method satisfies (3).

3.2 Convex case

Assume that f has a global min x^* and

$$R_0 := \max\{\|x - x^*\|_2 : f(x) \le f(x_0)\} < \infty;$$

that is, the sublevel set defined by x^0 is bounded.

Denote the optimality gap by $\Delta_k := f(x_k) - f(x^*)$. By convexity we have

$$\Delta_k = f(x_k) - f(x^*) \le \langle \nabla f(x_k), x_k - x^* \rangle \le R_0 \|\nabla f(x_k)\|_2 .$$

(Picture) Plugging into the descent condition (3), we obtain

$$\begin{aligned} & f(x_{k+1}) \le f(x_k) - \frac{\beta}{2R_0^2}\Delta_k^2 \\ \Longrightarrow\ & \Delta_{k+1} \le \Delta_k - \frac{\beta}{2R_0^2}\Delta_k^2 = \Delta_k\left(1 - \frac{\beta}{2R_0^2}\Delta_k\right). \end{aligned} \tag{4}$$

This recursion can be solved in multiple ways. Since $1 - x \le \frac{1}{1+x}, \forall x \ge 0$, (4) implies

$$\Delta_{k+1} \le \Delta_k \frac{1}{1 + \frac{\beta}{2R_0^2}\Delta_k} = \frac{1}{\frac{1}{\Delta_k} + \frac{\beta}{2R_0^2}}.$$

Inverting both sides gives

$$\frac{1}{\Delta_{k+1}} \ge \frac{1}{\Delta_k} + \frac{\beta}{2R_0^2},$$

and recursively

$$\frac{1}{\Delta_{k+1}} \ge \frac{1}{\Delta_0} + \frac{(k+1)\beta}{2R_0^2} \ge \frac{(k+1)\beta}{2R_0^2}.$$

That is,

$$f(x_{k+1}) - f(x^*) \le \frac{2R_0^2}{\beta(k+1)}.$$

Remark 2. Compare with the bound for GD in the convex case: $f(x_{k+1}) - f(x^*) \le \frac{\|x_0 - x^*\|_2^2}{2\alpha(k+1)}$.

3.3 Strongly convex case

Assume f is m-strongly convex and has a unique global min x^*. By strong convexity:

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|_2^2, \qquad \forall x, y.$$

We minimize both sides over y, and note that the right hand size is minimized at $x - \frac{1}{m}\nabla f(x)$. Therefore,

$$\begin{aligned} f(x^*) = \inf_y f(y) &\ge \inf_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|_2^2 \right\} \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2, \qquad \forall x. \end{aligned}$$

Equivalently,

$$\|\nabla f(x)\|_2^2 \geq 2m\left[f(x) - f(x^*)\right], \qquad \forall x. \tag{5}$$

Combining with the descent condition (3), we get

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq f(x_k) - f(x^*) - \frac{\beta}{2}\|\nabla f(x_k)\|_2^2 \\ &\leq f(x_k) - f(x^*) - \frac{\beta}{2} \cdot 2m\left[f(x_k) - f(x^*)\right] \\ &= (1 - m\beta) \cdot \left[f(x_k) - f(x^*)\right]. \end{aligned}$$

Hence we have geometric convergence

$$f(x_{k+1}) - f(x^*) \leq (1 - m\beta)^{k+1}\left(f(x_0) - f(x^*)\right).$$

Remark 3. Even if f is not strongly convex, as long as (5) holds, the above analysis goes through. The condition (5) is called the *Polyak-Łojasiewicz* (PŁ) condition or *gradient domination* condition.

Exercise 2. As a quintessential example of a function that is not strongly convex but satisfies PŁ, consider $f(x) = \frac{1}{2}x^\top Ax$, where the matrix A is p.s.d. but singular. Show that f satisfies PŁ with $m =$?

Exercise 3. Can you find a nonconvex function that satisfies PL?

4 Other generalizations of strong convexity

A strongly convex function cannot be flat near the minimum: the function value must grow when moving away from the minimizer. There are several other conditions that also control the growth of a function and hence can be viewed as generalizations of strong convexity.

Recall the definition of strong convexity:

$$f\left((1-\alpha)x + \alpha y\right) \leq (1-\alpha)f(x) + \alpha f(y) - \frac{m}{2}(1-\alpha)\alpha\|y - x\|_2^2, \qquad \forall x, y, \forall \alpha \in (0,1). \tag{6}$$

$$\iff f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|_2^2, \qquad \forall x, y. \tag{7}$$

One may replace the ℓ_2 norm on the right hand side by another norm $\|\cdot\|$, or by another polynomial of norm $\|y - x\|^r$ (*uniform convexity*).

There are further generalization that covers some nonconvex functions. We have talked about the PL condition (5). PL can be generalized further to the *Kurdyka-Łojasiewicz (KL)* condition, which is (5) with $\|\nabla f(x)\|^r$ on the LHS. Another generalization is known as the *sharpness* condition or *Holderian error bounds*: a function is called (r, m)-sharp if

$$f(x) - \min_y f(y) \geq \frac{m}{r}\min_{x^* \in \mathcal{X}^*}\|x - x^*\|^r, \quad \forall x$$

where $\mathcal{X}^* := \arg\min_{x \in \mathbb{R}^d} f(x)$ denotes the set of minimizers.

Exercise 4. Use (7) to verify that an m-strongly convex function is $(2, m)$-sharp.

These conditions enable fast convergence of iterative algorithms (faster than merely assuming smoothness).

5 Generalization of smoothness (optional)

Complementary to the above "growth" conditions, the smoothness condition stipulates that a function cannot grow/fluctuate too quickly. One may generalize smoothness by replacing Lipschitz-continuity of gradient by Holder-continuity.

Definition 1. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called (κ, L)-weakly smooth for $\kappa \in [1, 2]$ w.r.t. a norm $\|\cdot\|$ if there exists a constant $L < \infty$ such that

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|^{\kappa - 1}, \qquad \forall x, y.$$

$(2, L)$-weak smoothness is the same as the usual L-smoothness. $(1, L)$-weak smoothness means $\|\nabla f(x) - \nabla f(y)\|_* \leq L$, which implies Lipschitz continuity of f.

Example 1. Examples of (weak) smoothness:

1. The log-sum-exp (soft-max) function $f(x) = \log \sum_{i=1}^d e^{x_i}$ is 1-smooth w.r.t. $\|\cdot\|_\infty$.
2. $\frac{1}{2}\|x\|_p^2$ with $p \geq 2$ is $(p-1)$-smooth w.r.t. $\|\cdot\|_p$.
3. $\frac{1}{2}\|x\|_p^p$ with $p \in [1, 2]$ is $(p, 1)$-weakly smooth w.r.t. $\|\cdot\|_p$.