Lecture 6: Gradient descent and its analysis

Yudong Chen

1 Basic descent methods

Take the form

$$x_{k+1} = x_k + \alpha_k p_k, \qquad k = 0, 1, \ldots$$

Definition 1. $p \in \mathbb{R}^d$ is a *descent direction* for f at x if

$$f(x + tp) < f(x)$$

for all sufficiently small $t > 0$.

Proposition 1. *If f is continuously differentiable (in a neighborhood of x), then any p such that $\langle \nabla f(x), p \rangle < 0$ is a descent direction.*

Proof. By Taylor's theorem:

$$f(x + tp) = f(x) + t \langle \nabla f(x + \gamma t p), p \rangle$$

for some $\gamma \in (0, 1)$. We know that $\langle \nabla f(x), p \rangle < 0$. As ∇f is continuous, for all sufficiently small $t > 0$,

$$\langle \nabla f(x + \gamma t p), p \rangle < 0,$$

hence $f(x + tp) < f(x)$. □

2 Gradient descent

What would be a good descent direction? Could try to move in the direction of $-\nabla f(x)$, since

$$-\frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \arg \max_{\|p\|_2 = 1} \langle -\nabla f(x), p \rangle .$$

"Simplest" descent algorithm:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

where α_k is the step size. Ideally, choose α_k small enough so that

$$f(x_{k+1}) < f(x_k)$$

when $\nabla f(x_k) \neq 0$.

Known as "gradient method", "gradient descent", "steepest descent" (w.r.t. the ℓ_2 norm).

3 Analysis of Gradient descent

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

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

Assumptions for this part:

(A1) f is L-smooth for $L < \infty$ (thus also continuously differentiable.)

(A2) $\mathcal{X} = \mathbb{R}^d$, i.e., the problem is unconstrained.

Note: we do *not* assume f is convex, until explicitly stated otherwise.

From properties of L-smooth functions (Lemma 1 in Lecture 4):

$$\forall y : f(y) \leq \underbrace{f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2} \|y - x_k\|_2^2}_{\text{RHS}}.$$

Set

$$x_{k+1} = \arg\min_{y \in \mathbb{R}^d} \{\text{RHS}\} = x_k - \frac{1}{L} \nabla f(x_k). \tag{1}$$

Here, the argmin can be found by setting the gradient of RHS to zero: $\nabla f(x_k) + L(x_{k+1} - x_k) = 0$. Moreover,

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

More generally, we have

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

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

Proof. Exercise. □

Remark 1. Eq. (1) gives an alternative way of deriving GD: we minimize a upper bound of f, where the upper bound is constructed using the local information $\nabla f(x_k)$.

3.1 The case of general smooth functions

We only assume f is L-smooth; f is potentially non-convex.

Repeatedly using Descent Lemma 1, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2 \\ &\leq f(x_{k-1}) - \frac{\alpha}{2} \|\nabla f(x_{k-1})\|_2^2 - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2 \\ &\vdots \\ &\leq f(x_0) - \frac{\alpha}{2} \sum_{i=0}^{k} \|\nabla f(x_i)\|_2^2. \end{aligned}$$

Rearranging terms:

$$\frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 \le f(x_0) - f(x_{k+1}).$$

Let's assume $f_* := \inf_x f(x) > -\infty$. We can bound the LHS and RHS above as

$$f(x_0) - f(x_{k+1}) \le f(x_0) - f_*$$

and

$$\frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 \ge \frac{\alpha}{2}(k+1)\min_{0\le i\le k}\|\nabla f(x_i)\|_2^2.$$

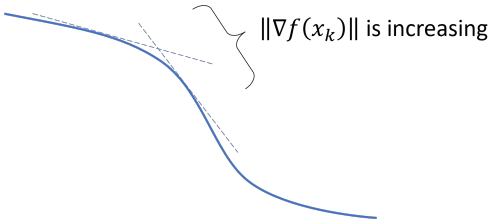
Combining last three equations:

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

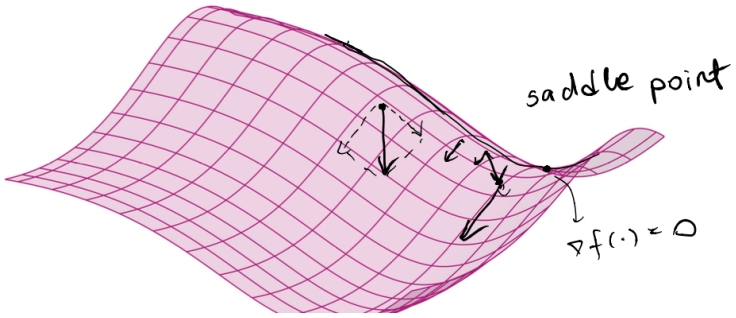
Equivalently, for any target error $\epsilon > 0$, GD finds an ϵ-near stationary point in roughly $\frac{C}{\epsilon^2}$ iterations:

$$\min_{0\le i\le k}\|\nabla f(x_i)\|_2 \le \epsilon \qquad \text{for } k+1 \ge \frac{2\left(f(x_0)-f_*\right)}{\alpha\epsilon^2}.$$

Remark 2. While function value $f(x_k)$ is decreasing in k, the gradient $\nabla f(x_k)$ need not.



Remark 3. When $\nabla f(x) = 0$, x may be a local min or a saddle point. Without further assumption, finding a stationary point is the best we can hope for (recall the hard case mentioned at the end of Lecture 4). Under certain assumptions (which exclude the hard case), we can show that randomly initialized GD usually converges to a local min.[1] [2]



[1] "Gradient Descent Converges to Minimizers", Jason Lee, Max Simchowitz, Michael Jordan, Benjamin Recht, 2016.
[2] Plot by Jelena Diakonikolas

3.2 The convex case

How does convexity help? Let $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x)$. (We assume the minimum is attained. The minimizer may not be unique.) Convexity gives lower bounds on $f_* = f(x^*)$:

$$\forall x : f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle .$$

Goal is to bound the *optimality gap* $f(x_{k+1}) - f(x^*)$. We have

$$\begin{aligned}
f(x^*) &\geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle \qquad \text{by convexity} \\
&= f(x_k) + \frac{1}{\alpha} \langle x_k - x_{k+1}, x^* - x_k \rangle \\
&= f(x_k) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2 - \frac{1}{2\alpha} \| \underbrace{x_k - x_{k+1}}_{-\alpha \nabla f(x_k)} \|_2^2 \\
&\qquad \text{using the Law of Cosines, a generalization of } (a-b)(c-a) = \frac{1}{2}(c-b)^2 - \frac{1}{2}(a-b)^2 - \frac{1}{2}(c-a)^2 \\
&= f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2 + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2 \\
&\geq f(x_{k+1}) + \frac{1}{2\alpha} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2\alpha} \|x_k - x^*\|_2^2 \qquad \text{by descent lemma.}
\end{aligned}$$

1) Distance to minimizer: We have

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2 &\leq 2\alpha \left(f(x^*) - f(x_{k+1}) \right) \\
&\leq 0
\end{aligned}$$

with strict inequality whenever $f(x_{k+1}) \neq f(x^*)$. So GD never moves further away from the set of minimizers.

2) Bound on optimality gap: We have

$$\begin{aligned}
f(x_{k+1}) - f(x^*) &\leq \frac{1}{2\alpha} \left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 \right) \\
\implies \sum_{k=0}^{K} \left[f(x_{k+1}) - f(x^*) \right] &\leq \sum_{k=0}^{K} \frac{1}{2\alpha} \left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 \right) \quad \longleftarrow \text{"telescoping sum"} \\
&\leq \frac{1}{2\alpha} \left(\|x_0 - x^*\|_2^2 - \|x_{K+1} - x^*\|_2^2 \right) \leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 .
\end{aligned}$$

But $f(x_1) \geq f(x_2) \geq \cdots$, so

$$\sum_{k=0}^{K} \left[f(x_{k+1}) - f(x^*) \right] \geq (K+1) \left[f(x_{K+1}) - f(x^*) \right] .$$

Combining,

$$f(x_{k+1}) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\alpha(k+1)}.$$

Equivalently, for each $\epsilon > 0$, we have $f(x_k) - f(x^*) \leq \epsilon$ after at most

$$k = \left\lceil \frac{\|x_0 - x^*\|_2^2}{2\alpha\epsilon} \right\rceil \text{ iterations.}$$

Compare with the general case.

Remark 4 (Telescoping Sum). We just saw a pattern that will appear many times in the proofs this semester. We summarize this argument below:

Lemma 2. *Let $\{a_k\}_{k\geq 0}$ and $\{D_k\}_{k\geq 0}$ be sequences of real numbers, with D_k non-negative. If*

$$a_k \leq D_k - D_{k+1} \quad \text{for all } k,$$

then

$$\min_{0\leq i\leq k} a_i \leq \frac{D_0}{k+1} \quad \text{for all } k.$$

If in addition a_k is non-increasing in k, then

$$a_k \leq \frac{D_0}{k+1} \quad \text{for all } k.$$

Proof. Observe that

$$(k+1)\cdot \min_{0\leq i\leq k} a_i \leq \sum_{i=0}^{k} a_i \leq \sum_{i=0}^{k}(D_i - D_{i+1}) = D_0 - D_{k+1} \leq D_0.$$

Moreover, when a_i is non-increasing in i, we have $\min_{0\leq i\leq k} a_i \geq a_k$. □

Formalizing the proof of this lemma in the theorem prover Lean 4 was posted as a challenge by Damek Davis and later taken up by Terry Tao.

3.3 The strongly convex case

Assume f is m-strongly convex. For all k:

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

Rearranging:

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

so

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

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

$$\|x_{k+1} - x^*\|_2^2 \leq (1 - m\alpha)^{k+1}\|x_0 - x^*\|_2^2.$$

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

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

Compare with previous two cases.

Exercise 1. Show that we also have

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

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