Gradient Methods II: Convergence Analysis

Ruoyu Sun

Questions for Last Time

- **Q1:** Explain what is gradient descent.
- **Q2:** Provide at least two ways to derive gradient descent.
- **Q3:** What is the **similarity** of gradient descent and Newton method?
- **Q4:** What are the two key ingredients of an iterative descent method?
- **Q5:** What stpesize rules are there?

Today

- Convergence Analysis of GD and iterative descent methods
- After today's course, you will be able to
  - Describe the convergence results for GD with various stepsize rules
  - Show the proof of convergence for GD with constant stepsize
  - Point out whether the results apply to a real world example
  - Advanced: Distinguish different kinds of convergence

Outline

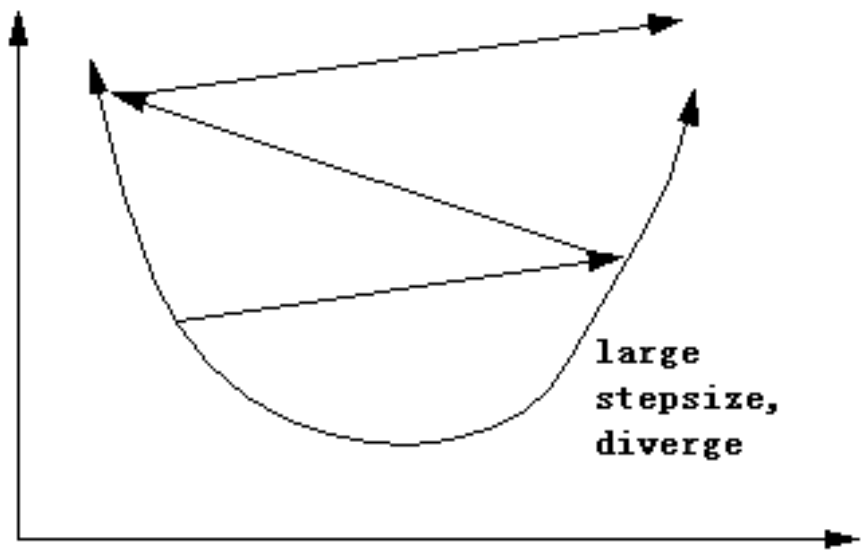
Example of Basketball (1D Quadratic Function Minimization)

- p is position, x is the force. $\min \frac{1}{2}(p - cx)^2$
- GD: $x^+ = x - \alpha \nabla f(x) = x - \alpha c(cx - p)$.
- Stepsize α represents how aggressively you adjust your strength

  What should be your strategy?

Strategy of Picking Stepsize

- **Line search** rules: too complex for this task
- **Constant**: how to pick the constant? (yeah, to achieve descent, but how? trial?)
- If not too carefully....



1D Convergence Analysis: Method 1

GD with constant stepsize:

$$x_{r+1} = x_r - \alpha f'(x_r) = (1 - \alpha c^2)x_r + \alpha c p. \tag{1}$$

- Simple case $p = 0$.
  Then $x_{r+1} = (1 - \alpha c^2)x_r$, or $x_r = (1 - \alpha c^2)^r x_0$.
  Then the sequence converges to $x^* = 0$ if

$$|1 - \alpha c^2| < 1, \text{ i.e., } 0 < \alpha < 2/c^2. \tag{2}$$

- For general case $p \neq 0$? Similar proof.

Claim 1: When using GD with stepsize α to solve $\min_x (p - cx)^2$, if $0 < \alpha < 2/c^2$, then $x^r \to x^*$, where x^* is the global-min.
Proof: Note $x^* = p/c$ is the global-min. Then (1) implies $x_{r+1} - x^* = (1 - \alpha c^2)(x_r - x^*)$.
$0 < \alpha < 2/c^2$ implies $|1 - \alpha c^2| < 1$, thus $x_r - x^* \to 0$, as $r \to \infty$.

Limitation of Method 1

Can we extend to multi-dim case?

The key of Method 1 seems to be: gradient update equation can be written as a linear update:

$$x_{r+1} = x_r - \alpha f'(x_r) = (1 - \alpha c^2)x_r + \alpha cp.$$

For general functions, $\nabla f(x)$ is not linear.

It seems hard to extend the proof to general functions.

1D Convergence Analysis: Method 2

Step 1: Following "descent" analysis: use Taylor expansion (now up to 2nd order; previously only up to 1st order)

$$f(x_{r+1}) = f(x_r) + f'(x_r)(x_{r+1} - x_r) + \frac{1}{2}(x_{r+1} - x_r)^2 \underbrace{f''(x_r)}_{c^2}$$

$$= f(x_r) - \alpha f'(x_r)^2 + \frac{1}{2}c^2\alpha^2 f'(x_r)^2$$

$$= f(x_r) + (-\alpha + \frac{1}{2}\alpha^2 c^2) f'(x_r)^2$$

If $0 < \alpha < 2/c^2$, we have $f(x_{r+1}) < f(x_r)$.

Step 2: $\{f(x_r)\}$ is decreasing and lower bounded by 0 (since $f(x) = (cx - p)^2$), thus $\{f(x_r)\}$ converges to a value f_∞.

Step 3: Let $\beta = -\alpha + \frac{1}{2}\alpha^2 c^2$, then $\beta f'(x_r)^2 = f(x_{r+1}) - f(x_r) \to 0$, as $r \to \infty$. Thus $f'(x_r) \to 0$.

Claim 2: When using GD with stepsize α to solve $\min_x (p - cx)^2$, if $0 < \alpha < 2/c^2$, then $f'(x^r) \to 0$.

Remarks on Two Proof Methods

Remark 1: Different starting point of two methods:

- Method 1 analyzes iterate x_r
- Method 2 analyzes function $f(x_r)$.

Remark 2: Extension:

- Method 1 can be somehow generalized to strongly convex functions (later lectures);
- Method 2 can be easily generalized to non-convex functions (next section)

Remarks on Meaning of "Convergence"

From Step 2 of Method 2, we already obtain:

Type-1: $f(x_r)$ converges to a value f_∞.

- Better than nothing (in some algorithm, $f(x_r)$ diverges to ∞)
- But not very useful! f_∞ can be anything: $0.5, 1.5, 100, 0, ...$

From Step 4 of Method 2, we have:

Type-2: $\nabla f(x_r) \to 0$.

- Implies: for r large enough, x_r is "approximate stat-pt"
- Can be extended to nonconvex functions

In Method 2, we do not mention:

Type-3: $\{x_r\}$ converges to a stationary point.

- Type-2 convergence does NOT imply Type-3 convergence.

In Method 1, we proved:

Type-4: $\{x_r\}$ converges to a global-min.

- Actually holds for strongly convex problems (by GD).
- Does NOT hold for general functions

Type of "Convergence"

Type-1: $f(x_r)$ converges to a value f_∞.

Type-2: $\nabla f(x_r) \to 0$.

Type-3: $\{x_r\}$ converges to a stationary point.

Type-4: $\{x_r\}$ converges to a global-min.

Relation: **Type-4** $\Rightarrow$ **Type-3** $\Rightarrow$ $\begin{cases} \textbf{Type-2} \\ \textbf{Type-1} \end{cases}$

Outline

The Descent Lemma

Assumption 1: $f : \mathbb{R}^n \to \mathbb{R}$ has L-Lipschitz gradient, i.e.,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$

Remark: we say f is L-smooth if f satisfies this assumption.

The Descent Lemma: Under Assumption 1, we have

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{3}$$

Intuition of the descent lemma:

$$f(\mathbf{x}) \approx f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + (\mathbf{x} - \mathbf{y})^T \nabla^2 f(\mathbf{y})(\mathbf{x} - \mathbf{y}).$$

Assumption-1 implies $\lambda_{\max}(\nabla^2 f(\mathbf{z})) \leq L, \forall \mathbf{z}$,

thus $(\mathbf{x} - \mathbf{y})^T \nabla^2 f(\mathbf{y})(\mathbf{x} - \mathbf{y}) \leq L\|\mathbf{x} - \mathbf{y}\|^2$.

Plugging into the expansion, done.

See formal proof next.

Proof of Descent Lemma (reading)

- See Prop. A. 24 of Bertsekas for proof of this lemma; also given below.
- Let t be a scalar and let $g(t) = f(x + ty)$
- Chain rule: $g'(t) = y'\nabla f(x + ty)$
- We have the following

$$
\begin{aligned}
& f(x+y) - f(x) \\
&= g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 y'\nabla f(x+ty)dt \\
&\leq \int_0^1 y'\nabla f(x)dt + \left| \int_0^1 y'(\nabla f(x+yt) - \nabla f(x))dt \right| \\
&\leq \int_0^1 y'\nabla f(x)dt + \int_0^1 \|y\| \|\nabla f(x+yt) - \nabla f(x)\| dt \\
&\leq y'\nabla f(x) + \|y\| \int_0^1 Lt\|y\| dt \qquad \text{(Lipschitz continuity)} \\
&= y'\nabla f(x) \frac{L}{2} \|y\|^2
\end{aligned}
$$

Alternative Way of Deriving Descent Lemma (reading)

There is a slightly different way to obtain the inequality in the descent lemma. This method is closer to the intuition in an earlier slide.

Assumption 2: f is twice differentiable and $\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq L, \forall \mathbf{x}$.
Descent lemma (variant): Under Assumption 2, we have

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2. \tag{4}$$

Proof: It directly follows from (2.6) of Taylor's theorem below (snapshot from book "Numerical Optimization").

Theorem 2.1 (Taylor's Theorem).
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that $p \in \mathbb{R}^n$. Then we have that

$$f(x + p) = f(x) + \nabla f(x + tp)^T p, \tag{2.4}$$

for some $t \in (0, 1)$. Moreover, if f is twice continuously differentiable, we have that

$$\nabla f(x + p) = \nabla f(x) + \int_0^1 \nabla^2 f(x + tp) p \, dt, \tag{2.5}$$

and that

$$f(x + p) = f(x) + \nabla f(x)^T p + \tfrac{1}{2} p^T \nabla^2 f(x + tp) p, \tag{2.6}$$

for some $t \in (0, 1)$.

Convergence Analysis of General Functions

- The starting point of Method 2: decrease of function value.
- **Step 1**: Assume f is L-smooth. Let $\mathbf{x}_{r+1} = \mathbf{x}_r - \alpha \nabla f(\mathbf{x}_r)$. Then

$$\begin{aligned} f(\mathbf{x}_{r+1}) - f(\mathbf{x}_r) &\leq \langle \nabla f(\mathbf{x}_r), \mathbf{x}_{r+1} - \mathbf{x}_r \rangle + \frac{L}{2} \|\mathbf{x}_{r+1} - \mathbf{x}_r\|^2 \\ &= -\alpha \|\nabla f(\mathbf{x}_r)\|^2 + \frac{L}{2} \alpha^2 \|\nabla f(\mathbf{x}_r)\|^2 \qquad (5) \\ &= \alpha(-1 + \frac{L}{2}\alpha) \|\nabla f(\mathbf{x}_r)\|^2 \overset{(i)}{<} 0. \end{aligned}$$

  where (i) holds when $0 < \alpha < \frac{2}{L}$.

- **Step 2**: $\{f(\mathbf{x}_r)\}$ is decreasing; there are two possibilities:
  - **Case 1:** $f(\mathbf{x}_r) \to -\infty$.
  - **Case 2:** There is a lower bound of the sequence $\{f(\mathbf{x}_r)\}$.

    Then $f(\mathbf{x}_r) \to f_\infty$. This implies $f(\mathbf{x}_{r+1}) - f(\mathbf{x}_r) \to 0$.

    Then by (5), we get $\alpha(-1 + \frac{L}{2}\alpha) \|\nabla f(\mathbf{x}_r)\|^2 \to 0$.

    Since $\alpha(-1 + \frac{L}{2}\alpha)$ is a positive constant, we have $\|\nabla f(\mathbf{x}_r)\| \to 0$.

Converg. Result 1: Constant Stepsize

- **Proposition** 1: Suppose we use GD with constant stepsize α to solve $\min_{\mathbf{x}\in R^n} f(\mathbf{x})$. Suppose
  - (i) f has L-Lipschitz gradient;
  - (ii) $0 < \alpha < 2/L$,

  then we have
  - Either $f(\mathbf{x}_r) \to -\infty$.
  - or $\|\nabla f(\mathbf{x}_r)\| \to 0$.
- This is the first formal result of this course.
- It provides strong guidance on how to pick stepsize: $\alpha < 2/L$.
- Its limitation?
  - Too conservative? Not adaptive.
  - L may not exist?

Converg. Result: Extensions

Let us briefly summarize a few extensions discussed next.

Stepsize rule choice:

- Any stepsize rule we learned (constant, diminishing, line search) works; (Prop 1b, Prop 2)

  properly bounded stepsize in $[\epsilon, 2/L - \epsilon]$ also works. (Prop 1b)

- For line search, NO Lipschitz-smoothness is needed. (Prop 2)

Update direction choice: Any gradient-related update direction (correlates well with $-\nabla f(x)$) works.

Fluctuating or diminishing stepsize

- **Proposition 1b** Under Assumption 1 (L-Lipschitz gradient), using GD with either one of the following choices of stepsize:
  1. There exits a scalar $\epsilon \in (0, 2)$ such that for all r

$$\epsilon < \alpha_r \leq \frac{(2-\epsilon)}{L}$$

  2. $\alpha_r \to 0$, and $\sum_{r=1}^{\infty} \alpha_r = \infty$ (i.e., $\alpha_r = \frac{1}{r}$)

  then either $f(\mathbf{x}^r) \to -\infty$ or $\nabla f(\mathbf{x}^r) \to 0$.

- **Remark 1**: Fluctuating in a range (strictly below $2/L$ and above 0) is fine.

- **Remark 2**: **Lipschitz-cts-gradient assumption** is used for constant and diminishing stepsize.

Can be even more general by allowing more choices of descent directions. See Prop. 1.2.3 and Prop. 1.2.3 in textbook.

Diminishing Stepsize Result: Textbook Version

A snapshot of textbook result (more general than our version)

Proposition 1.2.4: (Convergence for a Diminishing Stepsize) Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$. Assume that for some constant $L > 0$, we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall\, x, y \in \Re^n, \tag{1.26}$$

and that there exist positive scalars c_1, c_2 such that for all k we have

$$c_1\|\nabla f(x^k)\|^2 \leq -\nabla f(x^k)'d^k, \qquad \|d^k\|^2 \leq c_2\|\nabla f(x^k)\|^2. \tag{1.27}$$

Suppose also that

$$\alpha^k \to 0, \qquad \sum_{k=0}^{\infty} \alpha^k = \infty.$$

Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of f.

Convergence Result 2: No Assumption

- **Proposition 2**: Suppose we minimize a differentiable function by GD with either one of the following line search rules:
  - minimization rule,
  - limited minimization rule,
  - Armijo rule,

  then either $f(\mathbf{x}^r) \to -\infty$ or $\nabla f(\mathbf{x}^r) \to 0$.
- Proof omitted. Intuition: adaptive.
- **Practical guidance**: to avoid Lipschitz gradient assumption? Use Armijo rule.

See Proposition 1.2.1 in textbook for a slightly more general result: descent direction only requires to be "gradient related" (next slide).

Gradient-related

Algorithm: $\mathbf{x}^{r+1} = \mathbf{x}^r - \alpha_r \mathbf{d}^r$.

Do we have to use the exact negative gradient direction? No!

Any gradient-related update direction works for Prop. 1, 1b and 2.

Gradient related condition: For any sequence $\{\mathbf{x}^r\}$ that converges to a nonstationary point, the corresponding direction $\{\mathbf{d}^r\}$ is bounded and satisfies

$$\lim_{r\to\infty} \langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle < 0$$

Two cases that it holds: (see discussions after eq. (1.13) of textbook)

- $\mathbf{d}^r = -\mathbf{D}^r \nabla f(\mathbf{x}^r)$, with $\mathbf{D}^r$ being a positive definite matrix where eigenvalues are upper and lower bounded by positive constants.
- For some scalars $c_1 > 0, c_2 > 0, p_1 > 0, p_2 \geq 0$, for all r,

$$\nabla f(\mathbf{x}^r)^T \mathbf{d}^r \leq -c_1 \|\nabla f(\mathbf{x}^r)\|^{p_1}, \quad \|\mathbf{d}^r\| \leq c_2 \|\nabla f(\mathbf{x}^r)\|^{p_2}. \tag{6}$$

In text book, Prop. 1.2.1 to 1.2.4 are all presented for gradient related update directions (Prop. 1.2.1 and 1.2.2 use exactly the gradient related condition; Prop. 1.2.3 and Prop. 1.2.4 used variants of (6)).

Outline

Three Things

Next, we discuss three things about convergence results.

- New type of convergence: "every limit point is stationary"
- Applying convergence results to a few examples
- Why we study convergence?

New Type of "Convergence"

- In Prop. 1: either f diverges to $-\infty$, or gradient converges to zero.
- Note: $\nabla f(\mathbf{x}^r) \to 0 \not\Rightarrow \mathbf{x}^r$ converges:
  - Possibility 1: diverges, i.e., $\|\mathbf{x}^r\| \to \infty$ (even if f converges)
  - Possibility 2: jump between multiple stat-pts (non-isolated stat-pts);
- **Wrong statement**: ....then $\{\mathbf{x}^r\}$ converges to a stationary point.
- **Textbook** version: Every limit point of $\{\mathbf{x}^r\}$ is a stationary point.
  - Does it imply a limit point exists? No.
  - Example: Every child of John Snow is a boy. But he might have no child (by the logic of mathematicians).
  - Mathematically, (set of John's children) $\subseteq$ (set of boys). Empty set possible.

“Convergence” Means What?

- **Textbook** results: Every limit point of $\{\mathbf{x}^r\}$ is a stationary point.

- **Everyday language**: “converge to stationary points”.

  In the course, I may say “xx algorithm converge to stationary points”, but you shall understand it as:

  - Rigorously speaking, **means** “every limit point is a stationary point”
  - Without further assumptions, it doesn’t even mean “convergence”
  - There are indeed practical examples (e.g.logistic regression) that GD does NOT converge to any point (the sequence diverges), but we still say “converge to stationary points” since $\|\nabla f(\mathbf{x}^r)\| \to 0$

  Why? **Just for convenience.**

  - **Our expression**: For xx problem, GD converges to stationary points
  - **Precise expression** (long): For xx problem, the gradient of the sequence generated by GD converges to 0. (only apply to lower bounded problems)

The treatment of the book “Numerical optimization” avoids the iterates. Its main convergence result says: for a certain line search method, if f is bounded below, then $\nabla f(\mathbf{x}_r) \to 0$.

Type of "Convergence"

Type-1: $f(x_r)$ converges to a value f_∞.

Type-2: $\nabla f(x_r) \to 0$.

Type-2.5: Every limit point of $\{\mathbf{x}_r\}$ is a stationary point.

Type-3: $\{x_r\}$ converges to a stationary point.

Type-4: $\{x_r\}$ converges to a global-min.

Relation: **Type-4** $\Rightarrow$ **Type-3** $\Rightarrow$ $\begin{cases} \textbf{Type-2.5} \\ \textbf{Type-2} \\ \textbf{Type-1} \end{cases}$

Conclusion: GD converges as Type-2.5 (every limit point is stationary).

- If lower bounded, then GD converges as Type-2 (gradient converges to 0)

Application to Least Squares

- **Example 1**: $f(x) = \|\mathbf{X}^T\mathbf{w} - \mathbf{y}\|^2$.

  Note that the Hessian is $\mathbf{X}\mathbf{X}^T$, thus f is L-smooth with $L = \lambda_{\max}(\mathbf{X}\mathbf{X}^T)$.

  So for GD with stepsize

$$0 < \alpha < \frac{2}{\lambda_{\max}(\mathbf{X}\mathbf{X}^T)}$$

  every limit point is a stationary point (also a global-min).

Application to Simple Example 1

Is Lipschitz gradient common?

Well, non-Lipschitz gradient is common.

Example 2: $f(x) = x^4$.

- Use GD to solve it, do we have convergence guarantee?
- Is the gradient Lipschitz-continuous? No.
- $\frac{|f'(x)-f'(y)|}{|x-y|} = \frac{|4x^3-4y^3|}{|x-y|} = |4(x^2+y^2+xy)| \to \infty$, as $x, y \to \infty$
- Cannot apply Prop. 1 to claim GD with constant or diminishing stepsize converges to stationary points
  - It does NOT mean GD with constant stepsize does NOT converge to stationary points; it just means we do NOT know yet
  - Practical use? There is a possibility of failure, but not for sure
- Anyhow, can still apply Prop. 2 to claim GD with line search converges to stationary points

Application to Simple Example 3

Example 3: $f(x) = (xy-1)^2, x, y \in \mathbb{R}$.
This is 1-neuron linear neural-net, or 1-dim matrix factorization.

- Use GD to solve it, do we have convergence guarantee?
- Is the gradient Lipschitz-continuous? No.
- Cannot apply Prop. 1 to claim GD with constant or diminishing stepsize converges to stationary points
- Anyhow, can still apply Prop. 2 to claim GD with line search converges to stationary points

Remark: Any function that grows faster than quadratic (e.g. 4-th order polynomial) is NOT Lipschitz-smooth.
So Prop. 1 is actually NOT easy to apply to practice directly.

Remark: Then what is the point of Prop. 1 (if not easy to apply?)

- Guidance on stepsize choice: let local Lipschitz constant $L_r = \lambda_{\max}(\nabla^2 f(\mathbf{x}_r))$; it is reasonable to pick $\alpha_r = \Theta(1/L_r)$, though no rigorous convergence proof for this choice

Why Study Convergence Analysis?

- 1) Help choose algorithm.
  - 1.1) Determine range of applicability.

    For different problems which algorithm to choose?
    - E.g. assume Newton method not good for nonconvex problems (no convergence guaranteee);
    - But modified Newton method (adding τI) might worth trying, since there is convergence guarantee (gradient-related direction)

  - 1.2) Help narrow down the choice of algorithms, by computation cost analysis.
    - Per-iteration cost: e.g. no need to try Newton method for 10^6-dim problem due to time cost; instead, try GD
    - Convergence speed (cost of total iterations): will discuss later

- 2) Help use software package.

  cvx, MOSEK, Gurobi, CPLEX, Tensorflow, PyTorch, etc.
  - How to tune hyperparameters?
  - How to interpret the computation results? (e.g. fails to converge, why?)

See Section 1.2.1, the last part (version 2 of Bertsekas's book).

Summary 1: Type of "Convergence"

Type-1: $f(x_r)$ converges to a value f_∞.

Type-2: $\nabla f(x_r) \to 0$.

Type-2.5: Every limit point of $\{\mathbf{x}_r\}$ is a stationary point.

Type-3: $\{x_r\}$ converges to a stationary point.

Type-4: $\{x_r\}$ converges to a global-min.

Relation: **Type-4** $\Rightarrow$ **Type-3** $\Rightarrow$ $\begin{cases} \textbf{Type-2.5} \\ \textbf{Type-2} \\ \textbf{Type-1} \end{cases}$

Conclusion: GD converges as Type-2.5 (every limit point is stationary).

- If lower bounded, then GD converges as Type-2 (gradient converges to 0)

Summary 2: Conditions for Convergence

- **Key Assumption**: f is differentiable and L-smooth.
- **Prop. 1**: Under Key Assumption, GD with constant stepsize $\in (0, 2/L)$ works; i.e.,
  - every limit point is a stationary point ;
  - if there is a lower bound to f, then $\nabla f(x^r) \to 0$.
- **Prop. 1b:** Under same condition to Prop. 1, GD with bounded stepsize in $(\epsilon, 2/L - \epsilon)$ or diminishing stepsize satisfying $\sum_r \alpha_r = \infty$ and $\alpha_r \to 0$ works.
- Main drawback of Prop 1 $\&$ 1b: L may not exist
- **Prop. 2:** (line search) GD with line search (Armijo rule, minimization, limited minimization) works.

  Remark: no assumption on L-Lipschitz gradient.

Reading Material

- Read Section 1.2, especially Section 1.2.2, especially Prop 1.2.1 - 1.2.3.
- Read Proof Prop. 1.2.4 if you are interested in analysis of diminishing stepsize (delicate proof!)
- Next time: convergence speed analysis