IE510Applied Nonlinear Programming

Lecture 2: Gradient Methods II: Convergence Analysis

Ruoyu Sun

Jan., 2018

Questions for Last Time

- **Q1:** Explain what is gradient descent.
- **Q2:** What is the relation between gradient descent and Newton method?
  1) $\nabla f(x)$, $(\nabla^2 f(x))^{-1} \nabla f(x)$. special case of $D^{-1} \nabla f(x)$
  2) Both quadratic approximation.
- **Q3:** Is it possible for GD to find the global minima?
  Yes, convex case.
- **Q4:** What are the two key ingredients of an iterative descent method?
  direction
  stepsize

Last Time summary

- Gradient descent (also called steepest descent) has the form $x \to x - \alpha \nabla f(x)$.
- **Three ways to motivate**
  - Iterative descent; hill climbing
  - Successive quadratic approximation (with identity 2nd order term)
  - Fixed point algorithm
- **Two key ingredients** of iterative descent methods:
  - Direction: $-\nabla f(x)$, $-\nabla^2 f(x)^{-1} \nabla f(x)$, general $D \nabla f(x)$
  - Stepsize
- Stepsize rules
  - Pre-fixed: constant, diminishing (requirement?) $\sum_r \alpha^r = \infty,\ \alpha^r \to 0$
  - Line search: exact/limited minimization, Armijo rule

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

1. Example of Basketball: Analysis
2. Convergence Analysis: General
3. Without Lipschitz Assumption

Today

4. Applying Gradient Descent to Regression
5. Convergence Rate Analysis

next time

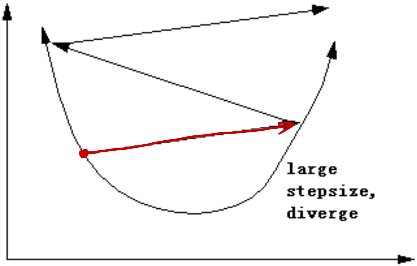
Example: Basketball again

- p is position, x is the force. $\min \frac{1}{2}(p - cx)^2$
- GD: $x^+ = x - \alpha \nabla f(x) = x - \alpha c e$, where $e = \underbrace{cx - p}_{}$, error
- 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....



Typical Convergence Analysis Types

- **Convergence to stationary solutions**
  1. Sanity check
  2. Minimal requirement of any reasonable algorithm
  3. Does not give global efficiency of the algorithm
- **Asymptotic convergence rate**: local analysis, assuming already close to a solution, let $\#$ of iterations go to infinity
  1. Linear rate/Supperlinear rate/Sublinear rate
- **Iteration complexity analysis**
  1. Measures the number of iterations required to get an optimal solution (e.g., $f(\mathbf{x}^r) - f \leq \epsilon$)
  2. Current analysis is all for the worst case and requires convexity
  3. Gives global behavior of the algorithm

Typical Convergence Analysis Types

- **Convergence to stationary solutions**
  1. Sanity check
  2. Minimal requirement of any reasonable algorithm
  3. Does not give global efficiency of the algorithm
- **Asymptotic convergence rate**: local analysis, assuming already close to a solution, let $\#$ of iterations go to infinity
  1. Linear rate/Supperlinear rate/Sublinear rate
- **Iteration complexity analysis**
  1. Measures the number of iterations required to get an optimal solution (e.g., $f(\mathbf{x}^r) - f \leq \epsilon$)
  2. Current analysis is all for the worst case and requires convexity
  3. Gives global behavior of the algorithm

Typical Convergence Analysis Types

Convergence

- **Convergence to stationary solutions**
  1. Sanity check
  2. Minimal requirement of any reasonable algorithm
  3. Does not give global efficiency of the algorithm

Convergence speed

- **Asymptotic convergence rate**: local analysis, assuming already close to a solution, let $\#$ of iterations go to infinity
  1. Linear rate/Supperlinear rate/Sublinear rate
- **Iteration complexity analysis**
  1. Measures the number of iterations required to get an optimal solution (e.g., $f(\mathbf{x}^r) - f \leq \epsilon$)
  2. Current analysis is all for the worst case and requires convexity
  3. Gives global behavior of the algorithm

Why Study Convergence Analysis?

- Do we really have to go through this? See Section 1.2.1, the last part (version 2 of Bertsekas's book).

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

    For different problems (convex or non-convex, smooth or non-differentiable, etc.), does the algorithm apply?

  - 1.2) Help narrow down the choice of algorithms.

    Knowing the speed helps a lot: save costly experimentation.

- 2) Help use software package.

  cvx, MOSEK, Gurobi, Tensorflow, PyTorch, Caffe2, MXNet, etc.

  "Parameter tuner" requires knowledge.

- **Example**: a student used PyTorch to solve a single-neuron network. Does not converge. Why?

Convergence Analysis of Basketball Example

GD with constant stepsize: $x^+ = x - \alpha f'(x) = (1 - \alpha c^2)x + \alpha c p.$

- Simple case: $p = 0$. Then the sequence converges if

$x^+ = (1-\alpha c^2)x$

$$|1-\alpha c^2| < 1, \qquad (1)$$

i.e., $0 < \alpha < \frac{2}{c^2}$.

- For general case $p \neq 0$?

- Use the idea of "descent" $f(x^+) < f(x)$; quantitatively,

$$\begin{aligned}
f(x^+) &= f(x) + f'(x)(x^+ - x) + \frac{1}{2}(x^+-x)^2 \cdot f''(x) + \underbrace{\cancel{o((x^+-x)^3)}}_{=0,\text{ since } f \text{ quadratic}} \\
&= f(x) - \alpha f'(x)^2 + \frac{1}{2}\alpha^2 f'(x)^2 \cdot c^2 \\
&= f(x) + \underbrace{\left(-\alpha + \frac{1}{2}\alpha^2 c^2\right)}_{\text{want } \le 0} f'(x)^2 \overset{\text{want}}{<} f(x)
\end{aligned}$$

Achieve "descent" iff (1) holds.

Convergence Analysis of Basketball Example (cont'd)

- Now we have $f(x^{r+1}) < f(x^r)$. So what?
- A decreasing sequence $\{f(x^r)\}$ must
  - either goes to $-\infty$ (impossible for this problem since $f(x) \geq 0, \forall x$)
  - or converges
- Are we done? One more thing: converge to what?
- Previous analysis: if $\{x^r\}$ converges, then $\nabla f(x^\infty) = 0$.
  Rigorously speaking, WRONG!
- **Issue**: $f(x^r)$ converges does NOT mean $\{x^r\}$ converges .

  e.g. ____________

  - Function value convergence v.s. iterate convergence

Convergence Analysis of Basketball Example (cont'd)

- Now we have $f(x^{r+1}) < f(x^r)$. So what?
- A decreasing sequence $\{f(x^r)\}$ must
  - either goes to $-\infty$ (impossible for this problem since ___________)
  - or converges
- Are we done? One more thing: converge to what?
- Previous analysis: if $\{x^r\}$ converges, then _____
  Rigorously speaking, WRONG!
- **Issue**: $f(x^r)$ converges does NOT mean $\{x^r\}$ converges .
  e.g. $f(x) = \cos(2\pi x)$, $f(x) = x^2$. $\{x^r\} = \{1, -1, 1, -1, \cdots\}$, diverge
  $f(x^r) = 1, \forall r$.
  - Function value convergence v.s. iterate convergence

Convergence Analysis of Basketball Example (cont'd)

- Correct argument:

$$\overbrace{f(x^{r+1}) - f(x^r)}^{\to 0} = \underset{\left(-\alpha+\frac{1}{2}\alpha^2c^2\right)\ \text{constant} > 0}{\beta}\, f'(x)^2$$

  $(-\alpha + \frac{1}{2}\alpha^2 c^2)$ constant >0.
  $\beta = (-\alpha + \frac{1}{2}\alpha^2 c^2)$

  $f(x^r)$ converges means $\underline{f'(x^r) \to \infty}$.

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

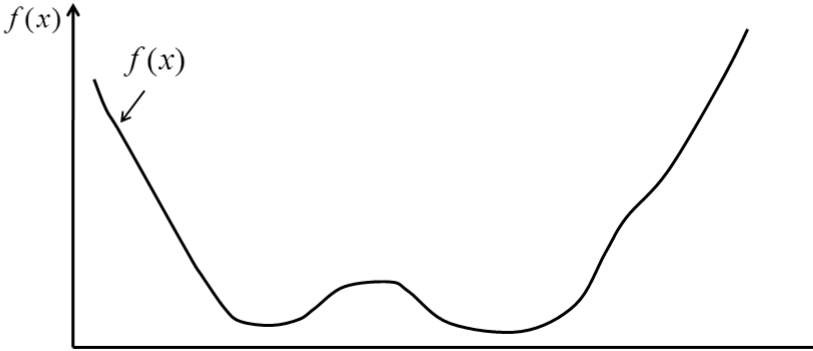
- Analysis of GD with constant stepsize for 1-dim quadratic problems

- Exercise: prove Prop. ~~1~~ 0 using fixed point theorem. 2-line proof .
  - Much easier. But hard to generalize to high-dimension case

- How much can we generalize?

Stepsize and Curvature (illustration)

Consider a general function, possibly nonconvex.

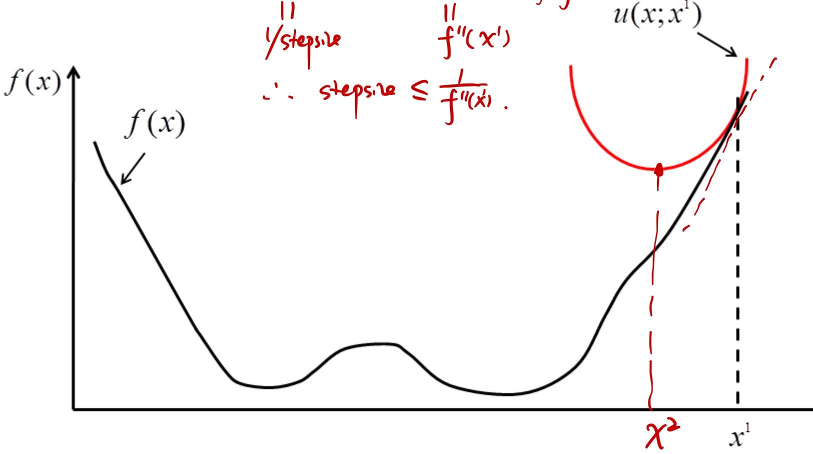


Stepsize and Curvature (illustration)

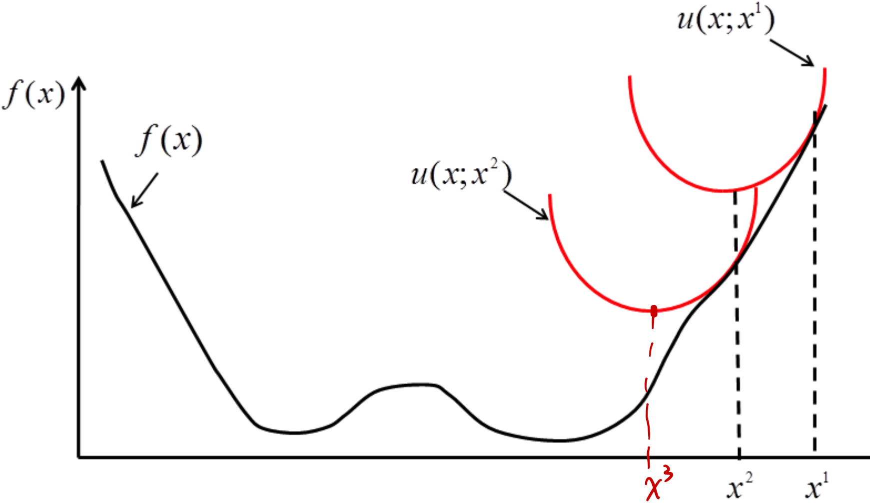
$u(x; x^1)$: quadratic approximation of f,
with curvature $\geq$ curvature of f at x^1.

$$\underbrace{\text{curvature}}_{=\ 1/\text{stepsize}} \geq \underbrace{\text{curvature of } f \text{ at } x^1}_{=\ f''(x^1)}$$

$\therefore$ stepsize $\leq \frac{1}{f''(x^1)}$.



Stepsize and Curvature (illustration)



Stepsize and Curvature: Relation

- From the above graph illustration (2nd interpretation of GD), we obtain important intuition.

- **Intuition**: stepsize should be inversely proportional to the curvature of the function

- **Example**: $f(x) = \frac{1}{2}(p - cx)^2$. Curvature is $f''(x) = c$, stepsize

$$0 < \alpha < 2/c.$$

Typical choice $\alpha = 1/c$.

Non-convex: $\alpha < \dfrac{2}{\text{max curvature}}$

Extension to High Dimension

- The key ingredient in the proof: decrease of function value.
- For twice-differentiable function $f(\mathbf{x})$, let $\mathbf{x}^+ = \mathbf{x} - \alpha \nabla f(\mathbf{x})$, then

Compare: $\|v\|^2$ v.s. $v^T A v$?
$v^T A v \le \lambda_{max}(A) \|v\|^2$, for symmetric A.

$$
\begin{aligned}
f(\mathbf{x}^+) &\approx f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x} \rangle + \frac{1}{2}(\mathbf{x}^+ - \mathbf{x})^T \nabla^2 f(x) (\mathbf{x}^+ - \mathbf{x}) \\
&= f(x) - \alpha \|\nabla f(x)\|^2 + \frac{1}{2} \underbrace{\nabla f(x)^T \nabla^2 f(x) \nabla f(x)}_{} \cdot \alpha^2 \\
&\overset{(i)}{\le} f(x) + \underbrace{(-\alpha + \tfrac{1}{2} L \alpha^2)}_{\text{want } < 0} \|\nabla f(x)\|^2 \qquad \le L \|\nabla f(x)\|^2 \cdot \alpha^2 \\
&\overset{(ii)}{<} f(\mathbf{x}),
\end{aligned}
$$

where in (i) we assumed

$$\nabla^2 f(\mathbf{y}) \preceq LI, \quad \forall y \in \mathbb{R}^n. \tag{2}$$

i.e. $\lambda_{max}(\nabla^2 f(y)) \le L$.

and (ii) holds when

$$0 < \alpha < \frac{2}{L}. \tag{3}$$

The Descent Lemma

- We are almost done... except one math improvement.
- Mathematicians want a weak condition: no need to assume twice-differentiable, but continuously-differentiable. (i.e. $\nabla f(x)$ is continuous)
- **Assumption 1**: $f : \mathbb{R}^n \to \mathbb{R}$ has L-Lipschitz gradient, i.e.,

  $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- **The Descent Lemma**: Under Assumption 1, we have

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

  Remark: Sometimes directly assume f satisfies this lemma, called L-smooth.
- With this lemma, the argument in the last page still holds.

Proof of Descent Lemma (skip in class)

- 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

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

Converg. Result 1: Constant Stepsize

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

  then we have
  - Either $f(x^r) \to -\infty$,
  - or $\nabla f(x^r) \to 0$.
- This is the first formal result of this course.
- It provides strong guidance on how to pick stepsize.
- Its limitation?
  - Too conservative? Not adaptive.
  - L may not exist?

Converg. Result 1: Constant Stepsize

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

  then we have
  - Either
  - or

- This is the first formal result of this course.
- It provides strong guidance on how to pick stepsize.
- Its limitation?
  - Too conservative? Not adaptive.
  - L may not exist? More discussion later.

Subtleties 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: diverge. $\|x^r\| \to \infty$ (even if f converges)
  - Possibility 2: jump around (non-isolated stationary points);
- **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 mine is a boy. I might have no child.
  - Mathematically, $\underbrace{\text{set of my child}}_{\text{possibly } \phi} \subseteq$ set of boys. Empty set possible.

Subtleties of "Convergence" (optional)

- **Improvement**: To guarantee $\{\mathbf{x}^r\}$ does not diverge? Under Assumption 2a and 2b below.
- **Assumption 2a:** The level set $\{f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$ is compact;

  Assumption 2b: The algorithm is decreasing (i.e. $f(x^{r+1}) \leq f(x^r)$)

- **Improvement**: To guarantee $\{\mathbf{x}^r\}$ converge to a single point? Under Assumption 2a,2b and Assumption 3.
- **Assumption 3**: Every stationary point is isolated
  - rigorously speaking, only require the sequence falls into a neighborhood of Isolated local-min (Prop 1.2.5 Capture Theorem);

isolated

non-isolated

Application to Least Squares

- **Example**: $f(x) = \|Ax - b\|^2$.
- **Case 1**: Strictly convex case. A is nonsingular fat (overdetermined; more sample than features).

  f satisfies Assumption 2a and Assumption 3.

  So GD with stepsize

  $$0 < \alpha < \frac{2}{\lambda_{\max}(A^T A)}$$ (satisfies Assumption 2b)

  converges to the unique global-min.

- **Case 2**: Non-strictly convex case.

  A is singular; or A is tall (underdetermined; more features than samples).

  For GD with proper stepsize, we still have $\nabla f(\mathbf{x}^r) \to 0$. But Prop. 1 doesn't imply $\{\mathbf{x}^r\}$ converge (need advanced tool).

Summary: “Convergence” Means What?

- Most **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 “converge to stationary points”, but you should understand

  - It **really means**, theoretically, every limit point is stationary point
  - Without further assumptions, it doesn't even mean convergence.

  In practice, diverging and jumping around are rare. There are deeper reasons, not covered in this course.

- Sometimes we can prove convergence to a single stationary point (e.g. strictly convex).

- Philosophical question: what is knowledge? Do you know “GD converges”?

More General Convergence Result

- **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., $\alpha_r = \frac{1}{r}$)

  we have every limit point is a stationary point.

- **Remark 1**: We can pick constant stepsize $\alpha_r =$ 1/L. But fluctuating in a range is also fine.

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

- **Remark 3**: 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

A snapshot of textbook result.

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

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

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

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

condition on direction

Suppose also that

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

diminishing stepsize

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.

Contain many elements, I break them down into several pieces.

Application to Simple Example 1

- Is Lipschitz gradient common?

  Well, at least non-Lipschitz gradient is common.

- **Example 1**: $f(x) = x^4$.
- Use GD to solve it? $x^+ = x - \alpha \cdot 4x^3$.

Lipschitz gradient? No.

$$\frac{|\nabla f(x) - \nabla f(y)|}{|x-y|} = \frac{|4x^3 - 4y^3|}{|x-y|} = |4(x^2+y^2+xy)| \to \infty \text{ as } x, y \to \infty.$$

Solution: ① level set $\{f(x) \le f(x^0)\} = \{x^4 \le f(x^0)\} \triangleq \Omega$ is compact;

Suppose $|x| \le B$, $\forall x \in \Omega$, then $|\nabla f(x) - \nabla f(y)| / |x-y| \le 12B^2$.

② Pick stepsize $\alpha = \frac{1}{12B^2}$, then f is decreasing, $\{x^r\}$ stays in Ω.

③ By same argument, GD with stepsize $\alpha = \frac{1}{12B^2}$ converges to 0.

Application to Simple Example 2

- **Example 2**: $f(x, y) = (xy - 1)^2, \quad x, y \in \mathbb{R}$.
- This is 1-neuron linear neural-net, or 1-dim matrix factorization.
- Use GD to solve it?

$$\nabla f(x,y) = \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \end{pmatrix} = \begin{pmatrix} y(xy-1) \\ x(xy-1) \end{pmatrix}.$$

It's easy to verify ∇f is NOT Lipschitz continuous in $\mathbb{R}$.

Level-set? Unbounded.

Cannot directly apply Prop. 1.

Convergence Result 2: No Assumption

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

  every limit point of the sequence is a stationary point.

- 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

(skip in class; can read it yourself).

- The direction $\mathbf{d}^r$ cannot be orthogonal to $\nabla f(\mathbf{x}^r)$ [figure]
- **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

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

- Is this condition satisfied for $\mathbf{d}^r = -\mathbf{D}^r \nabla f(\mathbf{x}^r)$, with $\mathbf{D}^r$ being a positive definite matrix?

Answer: - - - - -

- If $D^r = D$, $\forall r$, where D is PD, fine. When D^r can change, it is different. Answer is no.

Homework

- 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 rate analysis