MG/OPT direction is a descent direction

i.e., it forms an obtuse angle to the gradient

Andersen Ang

Department of Combinatorics and Optimization,
University of Waterloo, Waterloo, Canada

ms$\mathbf{x}$ang@uwaterloo.ca, where $\mathbf{x} = \lfloor \pi \rfloor$
Homepage: angms.science

First draft: May 6, 2021
Last update: May 18, 2021

Setting

- Unconstrained problem

$$\min_{\mathbf{x}} f(\mathbf{x}), \qquad f : \mathbb{R}^n \to \mathbb{R},\ f \in C^2$$

- Iterative update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{p}$
- 2nd-order method: update direction $\mathbf{p}$ comes from Newton's eqn.

$$\left[\nabla^2 f(\mathbf{x})\right]\mathbf{p} = -\nabla f(\mathbf{x}_k)$$

- Facts:
  - If $\mathbf{x}$ is close to a sol., Newton-type method converges fast,
  - If $\mathbf{x}$ is far from a sol., performance of Newton-type method can be bad
- Idea: Multigrid (MG)
  - Use coarse problem to compute a $\underbrace{\text{descent direction}}_{\text{“correlate” to } -\nabla f(\mathbf{x})}$ for fine problem
  - MG provides $\underbrace{\text{an effective search direction}}_{\text{cheap to compute at coarse problem}}$ in the region far from sol.
  - idea rooted from solving PDEs

Prerequisite: Multigrid notations

Common notations used in the literature / MG community

- h: level
- f_h: f at level h
- $\mathbf{I}_h^k$: level h to level k operator
  - $\mathbf{I}_h^{h-1}$: one-level coarsify
  - $\mathbf{I}_{h-1}^h$: one-level refine
  - $h = 1$: coarsest level
  - $\mathbf{I}_{h-1}^h = c(\mathbf{I}_h^{h-1})^\top, \ c \geq 0$

My notations

- Only 2 levels (easier to understand)
- $\mathbf{P}$: prolongation $(h - 1 \ \rightarrow \ h)$
- $\mathbf{R}$: restriction $(h \ \rightarrow \ h - 1)$
- $\mathbf{P} = c\mathbf{R}^\top, \ c \geq 0$

In this document: we assume we know $\mathbf{R}$ and $\mathbf{P}$ (they are given).

Prerequisite: Mean Value Theorem (MVT)

- Let $\mathbf{F}: \mathbb{R}^n \to \mathbb{R}^n$ be differentiable vector-valued vector function (e.g. gradient vector), then MVT says that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\mathbf{F}(\mathbf{a}) - \mathbf{F}(\mathbf{b}) = \nabla \mathbf{F}(\mathbf{c})(\mathbf{a} - \mathbf{b})$$

  where $\nabla \mathbf{F}$ is a matrix in $\mathbb{R}^{n \times n}$, and $\mathbf{c} \in \mathbb{R}^n$ is a point on the line $[\mathbf{a}, \mathbf{b}]$, i.e.,

$$\mathbf{c} = \mathbf{a} + t(\mathbf{a} - \mathbf{b}) \text{ for some } t \geq 0.$$

- MVT is useful to link gradient to Hessian. Suppose f is twice differentiable, then

$$\nabla f(\mathbf{a}) - \nabla f(\mathbf{b}) = \nabla^2 f(\mathbf{c})(\mathbf{a} - \mathbf{b}).$$

- Let $f: \mathbb{R}^n \to \mathbb{R}$ be differentiable scalar-valued vector function, then MVT says that for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, the following is true

$$f(\mathbf{a}) - f(\mathbf{b}) = \langle \nabla f(\mathbf{c}), \mathbf{a} - \mathbf{b} \rangle,$$

  where $\mathbf{c} = \mathbf{a} + t(\mathbf{a} - \mathbf{b})$ for some $t \geq 0$.

MG/OPT

Algorithm 1: MG/OPT

Result: Sol to $\min_{\mathbf{x}} f(\mathbf{x}) \quad \ldots \quad (1)$

1 **if** $h = 1$ *(coarsest)* **then**
2   Solve (1) exactly
3 **else**
4     $\mathbf{x}_h^{(1)} = \text{Up}(\mathbf{x}_h^{(0)})$ % update, any algo. can be used;
5     $\mathbf{x}_{h-1}^{(1)} = \mathbf{I}_h^{h-1}\mathbf{x}_h^{(1)}$ % fine → coarse;
6     $\mathbf{v} := \nabla f_{h-1}(\mathbf{x}_{h-1}^{(1)}) - \mathbf{I}_h^{h-1}\nabla f_h(\mathbf{x}_h^{(1)})$ % gradient error across level;
7     $\mathbf{x}_{h-1}^{(2)} = \operatorname*{argmin}_{\mathbf{u}} \; f_{h-1}(\mathbf{u}) - \langle \mathbf{v}, \mathbf{u} \rangle$ with $\mathbf{x}_{h-1}^{(1)}$ as ini. pt.;
8     $\mathbf{x}_h^{(2)} = \mathbf{x}_h^{(1)} + \alpha \underbrace{\mathbf{I}_{h-1}^h\big(\mathbf{x}_{h-1}^{(2)} - \mathbf{x}_{h-1}^{(1)}\big)}_{\text{MG/OPT direction}}$ % line search α;
9     $\mathbf{x}_h^{(3)} = \text{Up}(\mathbf{x}_h^{(2)})$ % update, any algo.;
10 **end**

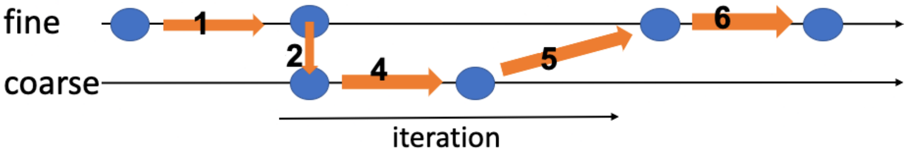
Lines 4 (9) is called pre-smoothing (post-smoothing) in MG community

2-level MG/OPT in my notation

Algorithm 2: MG/OPT, only 2 levels, in my language

Result: Sol to $\min_{\mathbf{x}} f(\mathbf{x})$

1 $\mathbf{x} = \mathrm{Up}(\mathbf{x}^0)$;
2 $\hat{\mathbf{x}} = \mathbf{R}\mathbf{x}$;
3 $\mathbf{v} = \nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R}\nabla f(\mathbf{x})$ (def of v);
4 $\hat{\mathbf{x}}^+ = \operatorname{argmin}_{\mathbf{u}} \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$;
5 $\mathbf{x}^+ = \mathbf{x} + \alpha \underbrace{\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})}_{\text{MG/OPT dir.}}$;
6 $\mathbf{x}^{++} = \mathrm{Up}(\mathbf{x}^+)$;



At convergence

Algorithm 3: MG/OPT, only 2 levels, in my language

Result: Sol to $\min_{\mathbf{x}} f(\mathbf{x})$

1 $\mathbf{x} = \mathsf{Up}(\mathbf{x}^0)$;

2 $\hat{\mathbf{x}} = \mathbf{R}\mathbf{x}$;

3 $\mathbf{v} = \nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R}\nabla f(\mathbf{x})$ (def of v);

4 $\hat{\mathbf{x}}^+ = \underset{\mathbf{u}}{\operatorname{argmin}} \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$;

5 $\mathbf{x}^+ = \mathbf{x} + \alpha \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$;

6 $\mathbf{x}^{++} = \mathsf{Up}(\mathbf{x}^+)$;

At convergence, $\mathbf{x}^+ = \mathbf{x}$ in 5, or equivalently

$$\begin{aligned} \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}) = \mathbf{0} \quad &\iff \quad \hat{\mathbf{x}}^+ = \hat{\mathbf{x}} \\ &\iff \quad \hat{\mathbf{x}}^+ = \mathbf{R}\mathbf{x} \end{aligned} \tag{FPP}$$

so $\hat{\mathbf{x}}^+_{\text{coarse}}$, the pt. after solving the coarse problem (line 4) is the same as restricted version of input $\mathbf{x}_{\text{fine}}$ (line 1)

What is $\mathbf{v}$?

$$\mathbf{v} = \nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R}\nabla f(\mathbf{x}) = \nabla \hat{f}(\mathbf{R}\mathbf{x}) - \mathbf{R}\nabla f(\mathbf{x})$$

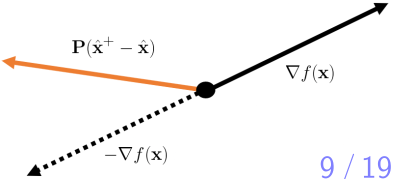
- At convergence, if $\mathbf{x}^*$ is a sol of $\min_{\mathbf{x}} f(\mathbf{x}) \overset{\text{FOC}}{\iff} \nabla f(\mathbf{x}^*) = \mathbf{0}$, so
$$\mathbf{v} = \nabla \hat{f}(\mathbf{R}\mathbf{x}^*) \qquad (\star)$$
- At $\mathbf{x}^*$, line 4 becomes
$$\hat{\mathbf{x}}^+ = \underset{\mathbf{u}}{\operatorname{argmin}} \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle \overset{(\star)}{=} \underset{\mathbf{u}}{\operatorname{argmin}} \underbrace{\hat{f}(\mathbf{u}) - \left\langle \mathbf{u}, \nabla \hat{f}(\mathbf{R}\mathbf{x}^*) \right\rangle}_{\mathcal{F}(\mathbf{u})}.$$
- $\hat{\mathbf{x}}^+$ is the minimizer of line 4 $\overset{\text{FOC}}{\iff} \nabla_{\mathbf{u}} \mathcal{F}(\mathbf{u})\Big|_{\mathbf{u}=\hat{\mathbf{x}}^+} = \mathbf{0}$, i.e.
$$\nabla \hat{f}(\hat{\mathbf{x}}^+) = \nabla \hat{f}(\mathbf{R}\mathbf{x}^*) \overset{[\nabla \hat{f}]^{-1}}{\Longrightarrow} \hat{\mathbf{x}}^+ = \mathbf{R}\mathbf{x}^*. \qquad \text{(FPP)}$$
- So $\mathbf{v}$ is defined in such a way such that the Fix-Point Property (FPP) holds.

FOC = First-order Optimality Condition.

Assumptions in MG/OPT

- Each $\min\limits_{\mathbf{x}} f_h(\mathbf{x})$ is cvx
- "Up" applied to $\min\limits_{\mathbf{x}} f_h(\mathbf{x})$ at any h is globally convergent
  - i.e., $\lim_{k\to\infty} \left\| \nabla f_h(\mathbf{x}_k) \right\| = 0$
- $\hat{\mathbf{x}}^+ = \operatorname*{argmin}\limits_{\mathbf{u}} \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$ is solved "accurate enough"
- You spend some iterations on each level: $\{N_0, N_1\}$ at least 1 is positive
- Search direction $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ is a descent direction for $\mathbf{x}$, which means

$$\begin{aligned} & \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}) \text{ correlates to } -\nabla f(\mathbf{x}) \\ \iff \quad & \left\langle \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), -\nabla f(\mathbf{x}) \right\rangle \geq 0 \\ \iff \quad & \left\langle \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \nabla f(\mathbf{x}) \right\rangle \leq 0 \end{aligned}$$



On $\hat{\mathbf{x}}^+ = \underset{\mathbf{u}}{\operatorname{argmin}}\ \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$, $\quad \mathbf{u}^* = \hat{\mathbf{x}}^+$

- If $F(\mathbf{u}) = \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$ is minimized exactly

$$\text{FOC: } \nabla_{\mathbf{u}} F(\mathbf{u}) = \mathbf{0} \iff \nabla \hat{f}(\mathbf{u}) = \mathbf{v} \overset{\text{def of } \mathbf{v}}{=} \nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R} \nabla f(\mathbf{x}) \qquad (\nabla)$$

- If $F(\mathbf{u})$ is not minimized exactly, add an error vector $\mathbf{z}$ on (∇):

$$\nabla \hat{f}(\mathbf{u}) = \nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R} \nabla f(\mathbf{x}) + \mathbf{z} \qquad (*)$$

- Next two slides show how to check $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ is a descent direction in these two cases.

If $\hat{\mathbf{x}}^+ = \underset{\mathbf{u}}{\operatorname{argmin}}\ \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$ is solved exactly

- If $F(\mathbf{u}) = \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$ is minimized exactly

$$\text{FOC: } \nabla_{\mathbf{u}} F(\mathbf{u}) = \mathbf{0} \iff \nabla \hat{f}(\mathbf{u}) = \mathbf{v} \overset{\text{def of } \mathbf{v}}{=} \nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R} \nabla f(\mathbf{x}) \qquad (\nabla)$$

- To check $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ is a descent direction, evaluate

$$\begin{aligned}
\Big\langle \nabla f(\mathbf{x}), \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}) \Big\rangle &= \Big\langle \mathbf{P}^\top \nabla f(\mathbf{x}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{\mathbf{P} = c\mathbf{R}^\top}{=} c\Big\langle \mathbf{R} \nabla f(\mathbf{x}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{(\nabla)}{=} c\Big\langle \nabla \hat{f}(\hat{\mathbf{x}}) - \nabla \hat{f}(\mathbf{u}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{\mathbf{u} = \mathbf{x}^+}{=} c\Big\langle \nabla \hat{f}(\hat{\mathbf{x}}) - \nabla \hat{f}(\hat{\mathbf{x}}^+), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&= -c\Big\langle \nabla \hat{f}(\hat{\mathbf{x}}^+) - \nabla \hat{f}(\hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{\text{MVT}}{=} -c\Big\langle \nabla^2 \hat{f}(\xi_1)(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle
\end{aligned}$$

where ξ_1 is a point on the line segment $[\hat{\mathbf{x}}^+, \hat{\mathbf{x}}]$

If $\hat{\mathbf{x}}^+ = \underset{\mathbf{u}}{\operatorname{argmin}}\ \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$ is not solved exactly

- If $F(\mathbf{u})$ is not minimized exactly, add an error vector $\mathbf{z}$ on (∇):

$$\nabla \hat{f}(\mathbf{u}) = \nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R} \nabla f(\mathbf{x}) + \mathbf{z} \qquad (*)$$

- To check $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ is a descent direction, evaluate

$$\begin{aligned}
\Big\langle \nabla f(\mathbf{x}), \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}) \Big\rangle &= \Big\langle \mathbf{P}^\top \nabla f(\mathbf{x}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{\mathbf{P}=c\mathbf{R}^\top}{=} c \Big\langle \mathbf{R} \nabla f(\mathbf{x}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{(*)}{=} c \Big\langle \nabla \hat{f}(\hat{\mathbf{x}}) - \nabla \hat{f}(\mathbf{u}) + \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{\mathbf{u}=\mathbf{x}^+}{=} c \Big\langle \nabla \hat{f}(\hat{\mathbf{x}}) - \nabla \hat{f}(\hat{\mathbf{x}}^+) + \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&= -c \Big\langle \nabla \hat{f}(\hat{\mathbf{x}}^+) - \nabla \hat{f}(\hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle + c \Big\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle \\
&\overset{\text{MVT}}{=} -c \Big\langle \nabla^2 \hat{f}(\xi_1)(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle + c \Big\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle
\end{aligned}$$

where ξ_1 is a point on the line segment $[\hat{\mathbf{x}}^+, \hat{\mathbf{x}}]$

On $\hat{\mathbf{x}}^+ = \underset{\mathbf{u}}{\operatorname{argmin}}\ \hat{f}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{v} \rangle$, $\quad \mathbf{u}^* = \hat{\mathbf{x}}^+$

$$\Big\langle \nabla f(\mathbf{x}), \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}) \Big\rangle = -c\Big\langle \nabla^2 \hat{f}(\xi_1)(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle + c\Big\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle$$

- Recall what we want to show

  $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ correlates to $-\nabla f(\mathbf{x}) \iff \Big\langle \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \nabla f(\mathbf{x}) \Big\rangle \leq 0$

- the first term $-c\Big\langle \nabla^2 \hat{f}(\xi_1)(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle$ is nonpositive
  $\because \nabla^2 \hat{f} \succeq \mathbf{0}$ as each $\min_{\mathbf{x}} f_h(\mathbf{x})$ is cvx (by assumption).
  Note there here we don't care about where exactly is ξ_1.

- Remaining task: show $c\big\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \big\rangle \leq 0$ for the inexact case.

On $c\left\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle$

- As $c \geq 0$, we can focus on $\left\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle$

$$
\begin{aligned}
\left\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle \quad & \overset{(*)}{=} \quad \left\langle \nabla \hat{f}(\mathbf{u}) - \left(\nabla \hat{f}(\hat{\mathbf{x}}) - \mathbf{R} \nabla f(\mathbf{x})\right), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle \\
& \overset{\text{def of } \mathbf{v}}{=} \quad \left\langle \nabla \hat{f}(\mathbf{u}) - \mathbf{v}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle \\
& = \quad \left\langle \nabla \hat{f}(\xi_1) - \mathbf{v} + \nabla \hat{f}(\mathbf{u}) - \nabla \hat{f}(\xi_1), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle \\
& = \quad \left\langle \nabla \hat{f}(\xi_1) - \mathbf{v}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle + \left\langle \nabla \hat{f}(\mathbf{u}) - \nabla \hat{f}(\xi_1), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle
\end{aligned}
$$

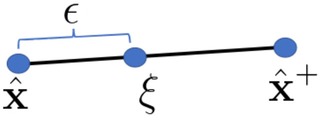
- The third step is very tricky: to create the $\left\langle \nabla \hat{f}(\mathbf{u}) - \nabla \hat{f}(\xi_1), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle$ so that we can use MVT.

- What's remain: show these terms ≤ 0

On the second term

- Fact from MVT: ξ_1 is in interval $[\hat{\mathbf{x}}^+, \hat{\mathbf{x}}]$:

$$\xi_1 = \hat{\mathbf{x}}^+ + \epsilon(\hat{\mathbf{x}} - \hat{\mathbf{x}}^+), \;\; \epsilon \geq 0 \qquad (\%)$$



- Hence the second term

$$\begin{aligned}
\left\langle \nabla \hat{f}(\mathbf{u}) - \nabla \hat{f}(\xi_1), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle &= -\left\langle \nabla \hat{f}(\xi_1) - \nabla \hat{f}(\hat{\mathbf{x}}^+), \hat{\mathbf{x}} - \hat{\mathbf{x}}^+ \right\rangle \\
&\overset{\text{MVT}}{=} -\left\langle \nabla^2 \hat{f}(\xi_2)(\xi_1 - \hat{\mathbf{x}}^+), \hat{\mathbf{x}} - \hat{\mathbf{x}}^+ \right\rangle \\
&\overset{(\%)}{=} -\left\langle \nabla^2 \hat{f}(\xi_2)\epsilon(\hat{\mathbf{x}} - \hat{\mathbf{x}}^+), \hat{\mathbf{x}} - \hat{\mathbf{x}}^+ \right\rangle \\
&= -\epsilon\left\langle \nabla^2 \hat{f}(\xi_2)(\hat{\mathbf{x}} - \hat{\mathbf{x}}^+), \hat{\mathbf{x}} - \hat{\mathbf{x}}^+ \right\rangle \\
&\leq 0 \quad \because \nabla^2 \hat{f} \succeq \mathbf{0}
\end{aligned}$$

for some ξ_2 in the interval $[\xi_1, \hat{\mathbf{x}}^+]$

On the first term

- What we have

$$\left\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle = \left\langle \nabla \hat{f}(\xi_1) - \mathbf{v}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle + \underbrace{\left\langle \nabla \hat{f}(\mathbf{u}) - \nabla \hat{f}(\xi_1), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle}_{\leq 0, \text{ see previous slide}}. \qquad \text{(z)}$$

- Suppose the problem in line 4, i.e.,

$$\hat{\mathbf{x}}^+ = \underset{\mathbf{u}}{\operatorname{argmin}} \; \underbrace{\hat{f}(\mathbf{u}) - \langle \mathbf{v}, \mathbf{u} \rangle}_{F(\mathbf{u})} \qquad (4)$$

  is solved with ini. guess $\hat{\mathbf{x}}$, then $F(\hat{\mathbf{x}}^+)$ should be smaller than $F(\hat{\mathbf{x}})$

$$\begin{array}{rll} & \hat{f}(\hat{\mathbf{x}}^+) - \langle \mathbf{v}, \hat{\mathbf{x}}^+ \rangle & \leq \hat{f}(\hat{\mathbf{x}}) - \langle \mathbf{v}, \hat{\mathbf{x}} \rangle \\ \iff & \hat{f}(\hat{\mathbf{x}}^+) - \hat{f}(\hat{\mathbf{x}}) - \langle \mathbf{v}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \rangle & \leq 0 \\ \overset{\text{MVT}}{\iff} & \langle \nabla \hat{f}(\xi_3), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \rangle - \langle \mathbf{v}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \rangle & \leq 0 \\ \iff & \langle \nabla \hat{f}(\xi_3) - \mathbf{v}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \rangle & \leq 0 \end{array} \qquad (\#)$$

  Then by (#), the first term of (z) ≤ 0 and we finished the proof.

- Important note
  - If (4) is solved exactly, then all the inequalities in (#) hold.
  - If (4) is not solved exactly, as long as the algorithm that solves (4) can decrease the value of F at $\hat{\mathbf{x}}$, then the inequalities in (#) hold.

What we have proved

- What we want to show: MG/PT direction $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ is a descent direction

$$\left\langle \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \nabla f(\mathbf{x}) \right\rangle \leq 0$$

- What we showed

$$\left\langle \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \nabla f(\mathbf{x}) \right\rangle \quad = \quad \underbrace{-c\left\langle \nabla^2 \hat{f}(\xi)(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle}_{\leq 0} + \underbrace{c\left\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \right\rangle}_{\leq 0}$$

- Therefore, $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ is a descent direction for $\mathbf{x}$.

- What does descent direction means: cost value will go down in the update

$$\mathbf{x}^+ = \mathbf{x} + \alpha \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$$

  where $\alpha \geq 0$ is obtained by line search

Where we used MVT

- First time: gradient-to-Hessian (p.12)

$$\nabla \hat{f}(\hat{\mathbf{x}}^+) - \nabla \hat{f}(\hat{\mathbf{x}}) = \nabla^2 \hat{f}(\xi_1)(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}),$$

  where $\xi_1 \in [\hat{\mathbf{x}}^+, \hat{\mathbf{x}}]$.

- Second time: gradient-to-Hessian (p.16)

$$\nabla \hat{f}(\xi_1) - \nabla \hat{f}(\hat{\mathbf{x}}^+) = \nabla^2 \hat{f}(\xi_2)(\xi_1 - \hat{\mathbf{x}}^+),$$

  where $\xi_2 \in [\xi_1, \hat{\mathbf{x}}^+]$.

- Third time: scalar-to-gradient (p.16)

$$\hat{f}(\hat{\mathbf{x}}^+) - \hat{f}(\hat{\mathbf{x}}) = \langle \nabla \hat{f}(\xi_3), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \rangle,$$

  where $\xi_3 \in [\hat{\mathbf{x}}^+, \hat{\mathbf{x}}]$.

- Note that we do not care about exactly where are these points.

Summary: what we have proved

- MG/PT direction $\mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}})$ is a descent direction, i.e.,

$$\Big\langle \mathbf{P}(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \nabla f(\mathbf{x}) \Big\rangle = \underbrace{-c\Big\langle \nabla^2 \hat{f}(\xi)(\hat{\mathbf{x}}^+ - \hat{\mathbf{x}}), \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle}_{\leq 0} + \underbrace{c\Big\langle \mathbf{z}, \hat{\mathbf{x}}^+ - \hat{\mathbf{x}} \Big\rangle}_{\leq 0}$$

- Reference: Stephen G. Nash, "A multigrid approach to discretized optimization problems", Optimization Methods and Software, 2000

End of document