2024/8/24–26

連続最適化および関連分野に関する夏季学校 2024

非凸最適化アルゴリズムとその計算量解析

丸茂 直貴

東京大学 大学院情報理工学系研究科 数理情報学専攻

本日の概要

扱う非凸最適化問題: 無制約・平滑最適化

$$\min_{x \in \mathbb{R}^d} \ f(x)$$

$f\colon \mathbb{R}^d \to \mathbb{R}$ は十分滑らか，一般に**非凸**，$\inf_{x \in \mathbb{R}^d} f(x) > -\infty$

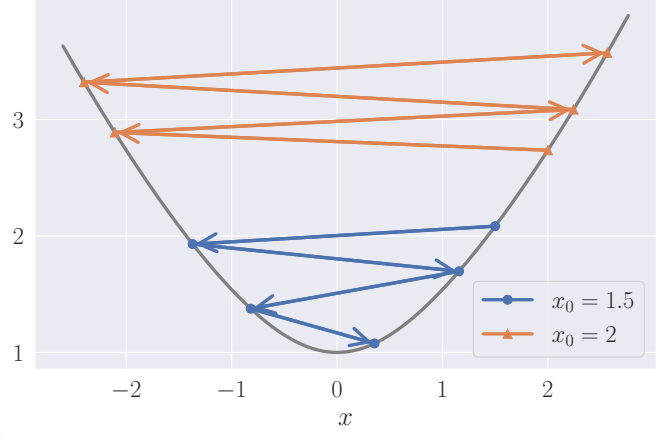
目標:
- アルゴリズムとその**計算量**保証の理解
- アルゴリズムの設計・解析における**定石**の習得

1. **最急降下法**
2. **3 次正則化 Newton 法**
3. **再始動 Heavy-ball 法**

計算量解析の動機 1

計算量の保証がないアルゴリズムは安心して使えない！

例: Newton 法は（f の性質が良くても）**収束しない**ことがある 演習 2



計算量解析の動機 2

実用的なアルゴリズムの設計に役立つ！

例: 凸関数に対する加速勾配法 [Nesterov, 1983]

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k),$$
$$y_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k)$$

- 謎の値 $\frac{k}{k+3}$ は **計算量解析** から導ける
- 加速勾配法は凸関数に対し最良の計算量を達成 😊
- 実用上も強い 😊

計算量解析の動機 3

計算量保証はアルゴリズム選択の基準として役立つ！

計算量解析をすると「どういう問題に強い / 弱いか？」がわかる

p. 51

連続最適化における計算量保証

ⓐ どのような問題クラスに対して
ⓑ どのような点を求めるために
ⓒ どのような計算を　　**最悪**でも何回すればよいか？

※ 平均ケースなどを考えることもある

例:

ⓐ 目的関数 f が線形，2 次，凸，強凸，複数の関数の和，合成関数，∇f **が Lipschitz 連続**，$\nabla^2 f$ が Lipschitz 連続，...

ⓑ $\|x - x^*\| \le \varepsilon,\ f(x) - f(x^*) \le \varepsilon,\ \|\nabla f(x)\| \le \varepsilon, \ldots$

ⓒ $f(x),\ \nabla f(x),\ \nabla^2 f(x),\ \nabla^2 f(x)v, \ldots$　　**ε-停留点**

- 最急降下法
- 3次正則化Newton法
- 再始動Heavy-ball法

非凸最適化アルゴリズムの計算量解析の定石

暫定解の列 $(x_k)_{k\in\mathbb{N}}$

計算量解析の定石

- $\|x_{k+1} - x_k\|$ が大 $\Longrightarrow$ **関数値減少量**が大 😊
- $\|x_{k+1} - x_k\|$ が小 $\Longrightarrow$ **勾配ノルム** が小 😊

} を両方示し，組み合わせる

「どちらに転んでも嬉しい」という状況を作る

非凸最適化アルゴリズムの計算量解析の定石

暫定解の列 $(x_k)_{k\in\mathbb{N}}$

計算量解析の定石

- $\|x_{k+1} - x_k\|$ が大 $\Longrightarrow$ **関数値減少量**が大 😊
- $\|x_{k+1} - x_k\|$ が小 $\Longrightarrow$ **勾配ノルム** が小 😊

} を両方示し，組み合わせる

ここからの話: 最急降下法を例に

- 上をどう示す？
- 上をどう組み合わせる？
- 上が成り立つアルゴリズムをどう設計する？ を見る

最も単純な例: 最急降下法

仮定: ∇f の L-Lipschitz 連続性

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad (\forall x, y \in \mathbb{R}^d)$$

更新式

$$x_{k+1} = x_k - \eta_k \nabla f(x_k) \qquad (\eta_k > 0)$$

計算量保証

$\Delta := f(x_0) - \inf_{x \in \mathbb{R}^d} f(x)$ とおく．$\eta_k = \frac{1}{L}$ とすると，

$$\min_{0 \leq i < k} \|\nabla f(x_i)\| \leq \sqrt{\frac{2L\Delta}{k}}$$

👉 ε-停留点を求めるための計算量: $\left\lceil \frac{2L\Delta}{\varepsilon^2} \right\rceil = O(\varepsilon^{-2})$

補足: 勾配の Lipschitz 連続性の直観

仮定: ∇f の L-Lipschitz 連続性

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \qquad (\forall x, y \in \mathbb{R}^d)$$

同値な条件　☞ 演習 1

$$\text{(上の仮定)} \qquad \Longleftrightarrow \qquad \|\nabla^2 f(x)\| \le L \quad (\forall x \in \mathbb{R}^d)$$

一次近似しにくい点があると L が大きくなる

Taylor 展開: $f(x+u) \simeq f(x) + \langle \nabla f(x), u \rangle + \frac{1}{2}\langle \nabla^2 f(x) u, u \rangle$

補足: 勾配の Lipschitz 連続性の直観

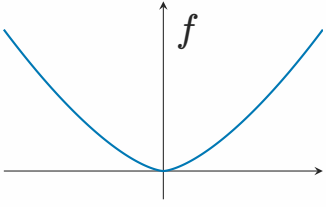
仮定: ∇f の L-Lipschitz 連続性

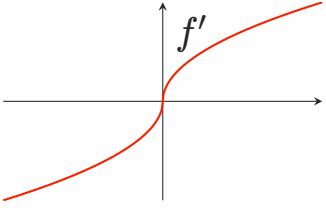
$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \qquad (\forall x, y \in \mathbb{R}^d)$$

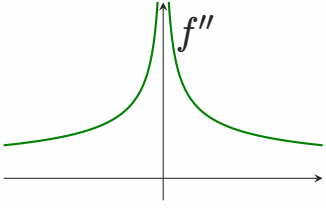
同値な条件　☞ 演習 1

(上の仮定) $\iff$ $\|\nabla^2 f(x)\| \le L \quad (\forall x \in \mathbb{R}^d)$

仮定を満たさない例: $f(x) = |x|^{3/2}$



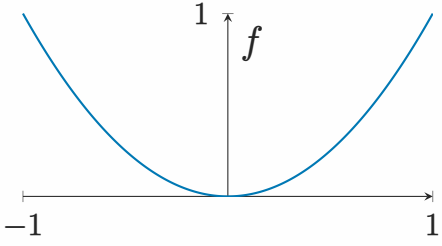


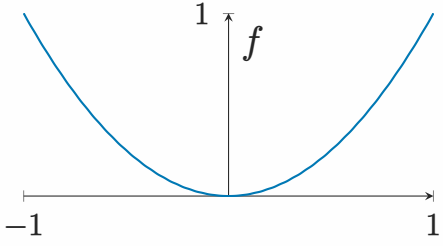


補足: 勾配の Lipschitz 連続性の直観

仮定: ∇f の L-Lipschitz 連続性

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad (\forall x, y \in \mathbb{R}^d)$$



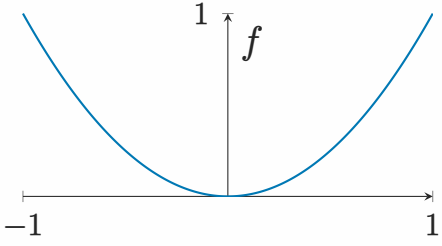


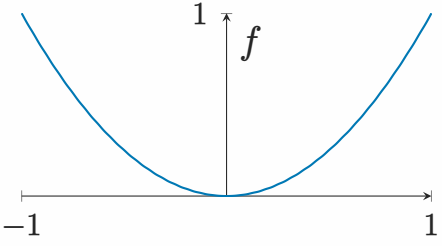
Quiz: どちらの L が大きい？

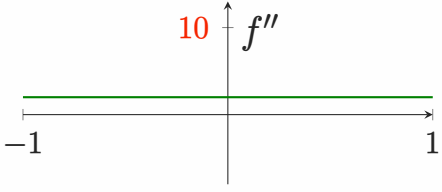
補足: 勾配の Lipschitz 連続性の直観

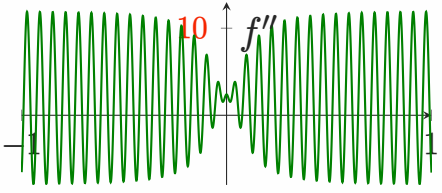
仮定: ∇f の L-Lipschitz 連続性

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \qquad (\forall x, y \in \mathbb{R}^d)$$







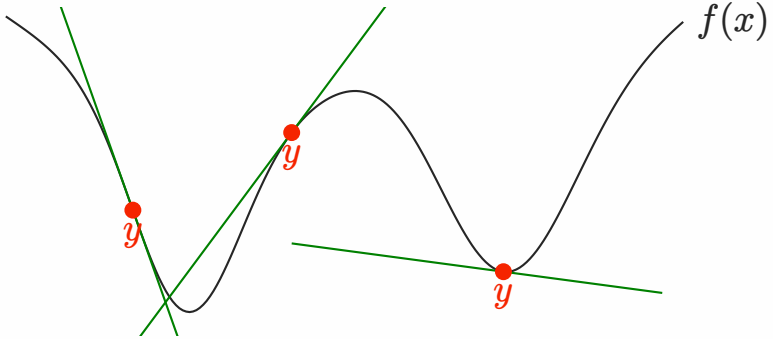


Lipschitz 連続性から従う不等式

∇f が L-Lipschitz 連続ならば，

$$\left| f(x) - \Big(f(y) + \langle \nabla f(y), x - y \rangle \Big) \right| \leq \frac{L}{2} \|x - y\|^2 \qquad (\forall x, y \in \mathbb{R}^d)$$

$f(x)$ の一次近似



Lipschitz 連続性から従う不等式

∇f が L-Lipschitz 連続ならば，

$$\left| f(x) - \Big(f(y) + \langle \nabla f(y), x - y \rangle \Big) \right| \leq \frac{L}{2} \|x - y\|^2 \qquad (\forall x, y \in \mathbb{R}^d)$$

証明: $f(x) - f(y) = \displaystyle\int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle \, \mathrm{d}t$ より，

$$\left| f(x) - f(y) - \langle \nabla f(y), x - y \rangle \right| = \left| \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), x - y \rangle \, \mathrm{d}t \right|$$

Cauchy–Schwarz

$$\leq \int_0^1 \|\nabla f(y + t(x-y)) - \nabla f(y)\| \|x - y\| \, \mathrm{d}t$$

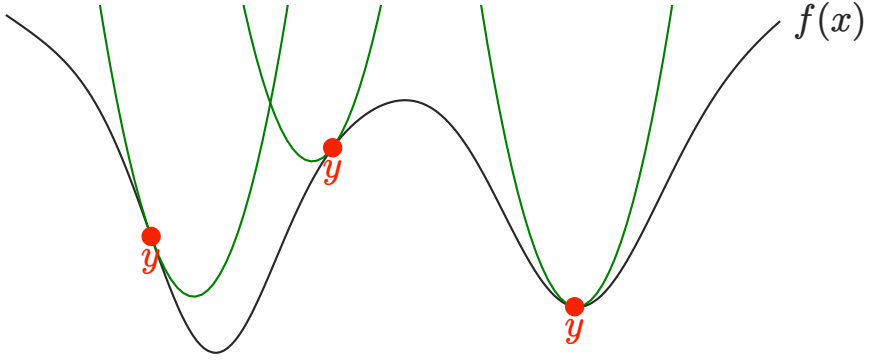
Lipschitz 連続性

$$\leq \int_0^1 Lt \|x - y\|^2 \, \mathrm{d}t = \frac{L}{2} \|x - y\|^2$$

Lipschitz 連続性から従う不等式: 上界補題

∇f が L-Lipschitz 連続ならば,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2 \qquad (\forall x, y \in \mathbb{R}^d)$$



最急降下法の解析: 関数値減少量の評価

- 上界補題:

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

- 更新式: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

より

関数値減少量の評価

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

👉 $\|x_{k+1} - x_k\|$ **が大** $\Longrightarrow$ **関数値減少量が大** 😊

最急降下法の解析: 勾配ノルムの評価

$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ より

勾配ノルムの評価

$$\|\nabla f(x_k)\| = L\|x_{k+1} - x_k\|$$

👉 $\|x_{k+1} - x_k\|$ が小 $\Longrightarrow$ 勾配ノルムが小 😊

最急降下法の解析: 2種類の評価を組み合わせる

Ⓐ $f(x_{k+1}) - f(x_k) \leq -\dfrac{L}{2}\|x_{k+1} - x_k\|^2$

Ⓑ $\|\nabla f(x_k)\| = L\|x_{k+1} - x_k\|$

Ⓐ, Ⓑ より $\dfrac{1}{2L}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$

$\Delta := f(x_0) - \inf_{x \in \mathbb{R}^d} f(x)$

k について足す: $\dfrac{1}{2L}\sum_{i=0}^{k-1}\|\nabla f(x_i)\|^2 \leq f(x_0) - f(x_k) \leq \Delta$

$\min_{0 \leq i < k}\|\nabla f(x_i)\|^2 \leq \dfrac{1}{k}\sum_{i=0}^{k-1}\|\nabla f(x_i)\|^2$ を使う: $\min_{0 \leq i < k}\|\nabla f(x_i)\| \leq \sqrt{\dfrac{2L\Delta}{k}}$

最急降下法の解析: 2種類の評価を組み合わせる

定石

- $\cdots \leq f(x_k) - f(x_{k+1})$ を足して $\cdots \leq f(x_0) - f(x_k) \leq \Delta$ を得る
- 自明な関係 (最小値) $\leq$ (平均値) を使う

Ⓐ, Ⓑ より $\dfrac{1}{2L}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$

$\Delta := f(x_0) - \inf_{x \in \mathbb{R}^d} f(x)$

k について足す: $\dfrac{1}{2L}\sum_{i=0}^{k-1}\|\nabla f(x_i)\|^2 \leq f(x_0) - f(x_k) \leq \Delta$

$\min_{0 \leq i < k}\|\nabla f(x_i)\|^2 \leq \dfrac{1}{k}\sum_{i=0}^{k-1}\|\nabla f(x_i)\|^2$ を使う: $\min_{0 \leq i < k}\|\nabla f(x_i)\| \leq \sqrt{\dfrac{2L\Delta}{k}}$

非凸最適化アルゴリズムの計算量解析の定石（再掲）

暫定解の列 $(x_k)_{k\in\mathbb{N}}$

計算量解析の定石

- $\|x_{k+1}-x_k\|$ が大 $\Longrightarrow$ 関数値減少量が大 😊
- $\|x_{k+1}-x_k\|$ が小 $\Longrightarrow$ 勾配ノルム　が小 😊

} を両方示し，組み合わせる

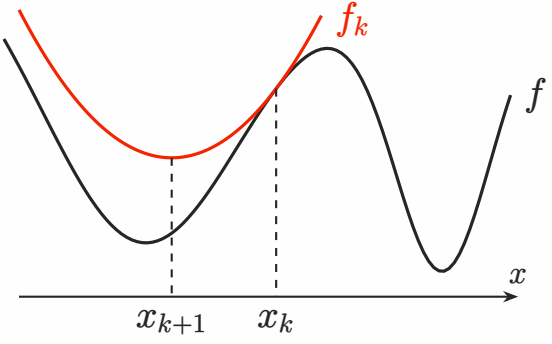
ここからの話: 最急降下法を例に

- 上をどう示す？
- 上をどう組み合わせる？
- **上が成り立つアルゴリズムをどう設計する？**　を見る

アルゴリズムの設計指針

アルゴリズム設計の定石

1. x_k 付近で f をよく近似する「シンプルな」関数 $\bar{f}_k$ を構成
2. $\bar{f}_k$ を（近似的に）最小化する点を x_{k+1} とする

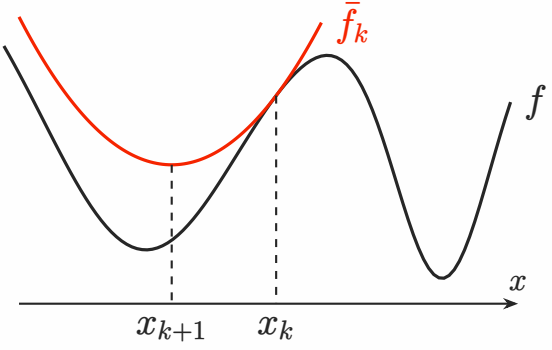


アルゴリズムの設計指針: 近似関数の選び方

定石: 近似関数に課す条件

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　Ⓑ $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

※ さらに $\bar{f}_k$ を強凸にすることも多い



アルゴリズムの設計指針: 近似関数の選び方

定石: 近似関数に課す条件

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　Ⓑ $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

※ さらに $\bar{f}_k$ を強凸にすることも多い

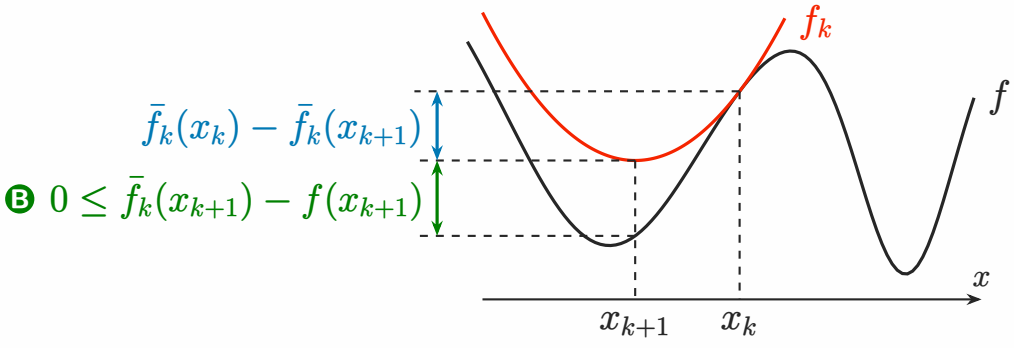
- このように $\bar{f}_k$ をとると，
  関数値減少量 & **勾配ノルム** の評価が容易 😊
- 最急降下法もこのような $\bar{f}_k$ を使う手法と見なせる

アルゴリズムの設計指針: 近似関数の選び方

定石: 近似関数に課す条件

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　Ⓑ $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

※ さらに $\bar{f}_k$ を強凸にすることも多い
→ $\bar{f}_k(x_k) - \bar{f}_k(x_{k+1})$ が評価しやすい 😊



アルゴリズムの設計指針: 最急降下法の場合

定石: 近似関数に課す条件

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　Ⓑ $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

$$
\begin{aligned}
x_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\
&= \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \bar{f}_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 \right\}
\end{aligned}
$$

- Ⓐ は明らか
- Ⓑ は上界補題 $f(x) \leq \bar{f}_k(x)\ \ (\forall x \in \mathbb{R}^d)$ より成立

アルゴリズムの設計指針: 最急降下法の場合

定石: 近似関数に課す条件

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　Ⓑ $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

$$
\begin{aligned}
x_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\
&= \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \bar{f}_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 \right\}
\end{aligned}
$$

定石

- $\bar{f}_k$ を「Taylor 展開 + 正則化項」で構成
- Lipschitz 連続性を使って Ⓑ を示す

ここまでのまとめ

計算量解析の定石

- $\|x_{k+1} - x_k\|$ が大 $\Longrightarrow$ **関数値減少量**が大 😊
- $\|x_{k+1} - x_k\|$ が小 $\Longrightarrow$ **勾配ノルム** が小 😊

} を両方示し，組み合わせる

アルゴリズム設計の定石

❶ x_k 付近で f をよく近似する「シンプルな」関数 $\bar{f}_k$ を構成

❷ $\bar{f}_k$ を（近似的に）最小化する点を x_{k+1} とする

定石: 近似関数に課す条件

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$

Ⓑ $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

ここから少し脇道に逸れて
最急降下法のステップサイズ選択の話をします

実用上重要 かつ
最急降下法以外にも共通する 話題

Lipschitz 定数 L が未知の場合への対応

最急降下法の計算量（再掲）

最急降下法でステップサイズを $\eta_k = \frac{1}{L}$ と定めると，

$$\min_{0 \leq i < k} \|\nabla f(x_i)\| \leq \sqrt{\frac{2L\Delta}{k}}$$

L が未知の場合どうする？

Lipschitz 定数 L が未知の場合への対応

最急降下法の計算量（再掲）

最急降下法でステップサイズを $\eta_k = \frac{1}{L}$ と定めると，

$$\min_{0 \le i < k} \|\nabla f(x_i)\| \le \sqrt{\frac{2L\Delta}{k}}$$

定石

- アルゴリズムと解析に登場する真値 L を推定値 ℓ で置換
- 解析で使った不等式が不成立なら，ℓ を大きくしてやり直し

$$f(x_{k+1}) - f(x_k) \le \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{\ell}{2} \|x_{k+1} - x_k\|^2$$

最急降下法 + Lipschitz 定数の推定

1: $\ell \leftarrow \ell_{\text{init}},\ \ k \leftarrow 0$ ▷ $\ell_{\text{init}} > 0$: 初期推定値
2: **loop**
3: $x \leftarrow x_k - \frac{1}{\ell}\nabla f(x_k)$
4: **if** $f(x) - f(x_k) \leq \langle \nabla f(x_k), x - x_k \rangle + \frac{\ell}{2}\|x - x_k\|^2$ **:**
5: $x_{k+1} \leftarrow x,\ \ k \leftarrow k+1$
6: **else**
7: $\ell \leftarrow \alpha\ell$ ▷ $\alpha > 1$: 定数

アルゴリズムから従う不等式:

- $f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \dfrac{\ell}{2}\|x_{k+1} - x_k\|^2$
- $\ell \leq \max\{\ell_{\text{init}},\ \alpha L\} =: \ell_{\max}$

最急降下法の解析: Lipschitz 定数を推定する場合

L が既知の場合と同様に $\dfrac{1}{2\ell}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$ が成立

$$\ell \leq \max\{\ell_{\text{init}}, \alpha L\} =: \ell_{\max}$$

$$\frac{1}{2\ell_{\max}}\|\nabla f(x_k)\|^2 \leq \frac{1}{2\ell}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$$

を k について足すと，

$$\min_{0\leq i<k}\|\nabla f(x_i)\| \leq \sqrt{\frac{2\ell_{\max}\Delta}{k}}$$

☞ k が更新される反復は高々 $\left\lceil \dfrac{2\ell_{\max}\Delta}{\varepsilon^2} \right\rceil$ 回

※ k が更新されない反復もあることに注意

最急降下法の解析: Lipschitz 定数を推定する場合

1: $\ell \leftarrow \ell_{\mathrm{init}},\ k \leftarrow 0$ ▷ $\ell_{\mathrm{init}} > 0$: 初期推定値
2: **loop**
3: $x \leftarrow x_k - \frac{1}{\ell}\nabla f(x_k)$
4: **if** $f(x) - f(x_k) \leq \langle \nabla f(x_k), x - x_k \rangle + \frac{\ell}{2}\|x - x_k\|^2$ **:**
5: $x_{k+1} \leftarrow x,\ k \leftarrow k+1$
6: **else**
7: $\ell \leftarrow \alpha\ell$ ▷ $\alpha > 1$: 定数

k が更新されない反復:

7 行目が n 回実行されるとすると，$\ell_{\mathrm{init}}\alpha^n \leq \ell_{\mathrm{max}}$ より

$$n \leq \log_\alpha \left(\frac{\ell_{\mathrm{max}}}{\ell_{\mathrm{init}}} \right)$$

最急降下法の解析: Lipschitz 定数を推定する場合

$\ell \leftarrow \ell_{\mathrm{init}},\ \ k \leftarrow 0$ ▷ $\ell_{\mathrm{init}} > 0$: 初期推定値

loop

  $x \leftarrow x_k - \frac{1}{\ell}\nabla f(x_k)$

  if $f(x) - f(x_k) \leq \langle \nabla f(x_k), x - x_k \rangle + \frac{\ell}{2}\|x - x_k\|^2$ **:**

    $x_{k+1} \leftarrow x,\ \ k \leftarrow k+1$

  else

    $\ell \leftarrow \alpha\ell$ ▷ $\alpha > 1$: 定数

上の最急降下法の反復数の上界

$$\left\lceil \frac{2\ell_{\max}\Delta}{\varepsilon^2} \right\rceil + \log_\alpha\left(\frac{\ell_{\max}}{\ell_{\mathrm{init}}}\right) \qquad (\ell_{\max} := \max\{\ell_{\mathrm{init}},\ \alpha L\})$$

最急降下法の解析: Lipschitz 定数を推定する場合

1: $\ell \leftarrow \ell_{\text{init}},\ \ k \leftarrow 0$ ▷ $\ell_{\text{init}} > 0$: 初期推定値
2: **loop**
3: $x \leftarrow x_k - \frac{1}{\ell}\nabla f(x_k)$
4: **if** $f(x) - f(x_k) \leq \langle \nabla f(x_k), x - x_k \rangle + \frac{\ell}{2}\|x - x_k\|^2$ **:**
5: $x_{k+1} \leftarrow x,\ \ k \leftarrow k + 1$
6: **else**
7: $\ell \leftarrow \alpha\ell$ ▷ $\alpha > 1$: 定数

- ∇f の Lipschitz 定数 L が未知でも計算量保証が得られた 😊
- ステップサイズ $\frac{1}{\ell}$ が単調に減少するのは実用上よくなさそう... ☹
  - 👍 実は ℓ を小さくするステップを入れても大丈夫

最急降下法の解析: Lipschitz 定数を推定する場合

1: $\ell \leftarrow \ell_{\text{init}},\ \ k \leftarrow 0$ ▷ $\ell_{\text{init}} > 0$: 初期推定値
2: **loop**
3: $x \leftarrow x_k - \frac{1}{\ell}\nabla f(x_k)$
4: **if** $f(x) - f(x_k) \leq \langle \nabla f(x_k), x - x_k \rangle + \frac{\ell}{2}\|x - x_k\|^2$ **:**
5: $x_{k+1} \leftarrow x,\ \ k \leftarrow k+1,\ \ \ell \leftarrow \beta\ell$ ▷ $0 < \beta \leq 1$: 定数
6: **else**
7: $\ell \leftarrow \alpha\ell$ ▷ $\alpha > 1$: 定数

上の最急降下法の反復数の上界　演習 3

$$\left\lceil \frac{2\ell_{\max}\Delta}{\varepsilon^2} \right\rceil \log_\alpha\left(\frac{\alpha}{\beta}\right) + \log_\alpha\left(\frac{\ell_{\max}}{\ell_{\text{init}}}\right)$$

最急降下法の解析: Lipschitz 定数を推定する場合

1: $\ell \leftarrow \ell_{\text{init}},\ \ k \leftarrow 0$ ▷ $\ell_{\text{init}} > 0$: 初期推定値
2: **loop**
3: $\quad x \leftarrow x_k - \frac{1}{\ell}\nabla f(x_k)$
4: $\quad$ **if** $f(x) - f(x_k) \leq \langle \nabla f(x_k), x - x_k \rangle + \frac{\ell}{2}\|x - x_k\|^2$ **:**
5: $\quad\quad x_{k+1} \leftarrow x,\ \ k \leftarrow k+1,\ \ \ell \leftarrow \beta\ell$ ▷ $0 < \beta \leq 1$: 定数
6: $\quad$ **else**
7: $\quad\quad \ell \leftarrow \alpha\ell$ ▷ $\alpha > 1$: 定数

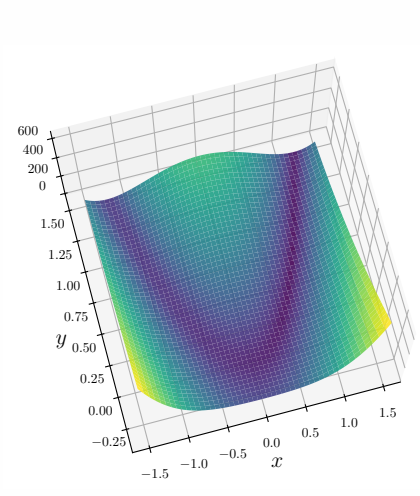
補足: Armijo 条件 [Armijo, 1966]

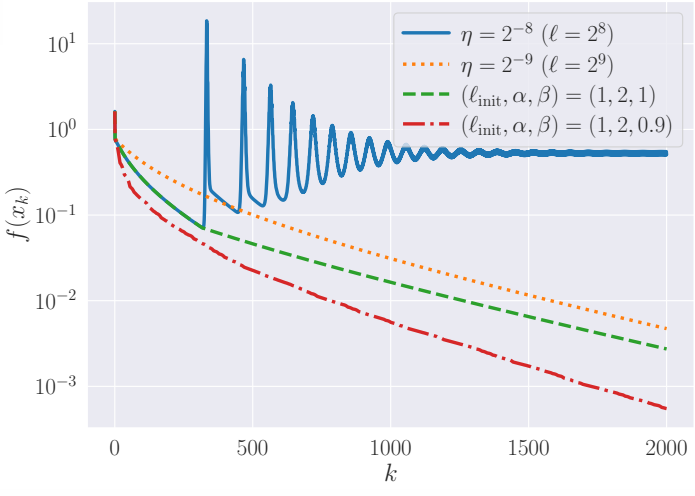
$$f(x) - f(x_k) \leq c\langle \nabla f(x_k), x - x_k \rangle \qquad (0 < c < 1\text{: 定数})$$

☞ $c = \frac{1}{2}$ とすると 4 行目の条件に一致

数値例: Rosenbrock 関数 [Rosenbrock, 1960]

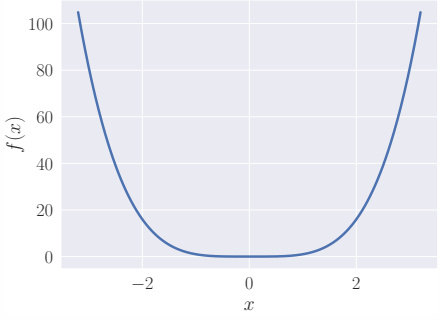
$$\min_{(x,y)\in\mathbb{R}^2} \quad (x-1)^2 + 100(y-x^2)^2$$

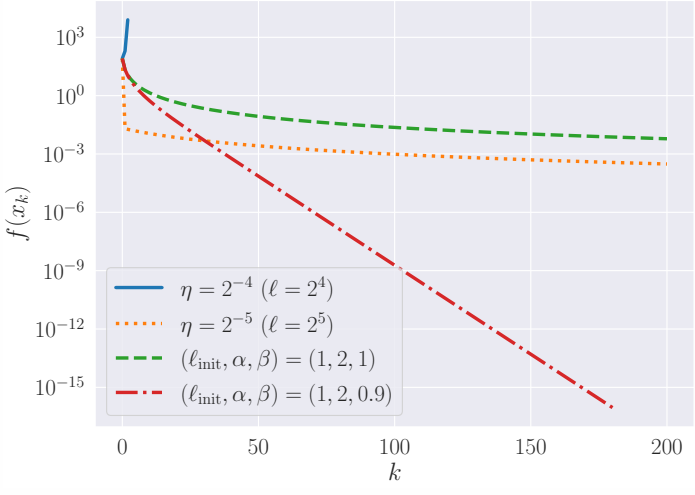




数値例: $\beta < 1$ の利点

$$\min_{x \in \mathbb{R}} \quad x^4$$





最急降下法の計算量 $\mathrm{O}\left(\frac{L\Delta}{\varepsilon^2}\right)$ は良い？ 悪い？

最急降下法の最適性

一次法: $f(x)$ と $\nabla f(x)$ の計算を通してのみ f の情報にアクセス

一次法の計算量下界 [Carmon et al., 2020]

任意の $L, \Delta, \varepsilon > 0$ と任意の一次法 $\mathcal{A}$ に対し，ある関数 f が存在し，

- ∇f は L-Lipschitz 連続
- $f(x_0) - \inf_x f(x) \leq \Delta$　（x_0: $\mathcal{A}$ の初期点）
- $\mathcal{A}$ は $c\frac{L\Delta}{\varepsilon^2}$ 回以下の関数値・勾配計算では $\|\nabla f(x)\| \leq \varepsilon$ なる x を見つけられない　（c: 定数）

最急降下法の $\mathrm{O}(\frac{L\Delta}{\varepsilon^2})$ は最良

最急降下法の最適性

一次法: $f(x)$ と $\nabla f(x)$ の計算を通してのみ f の情報にアクセス

一次法の計算量下界 [Carmon et al., 2020]

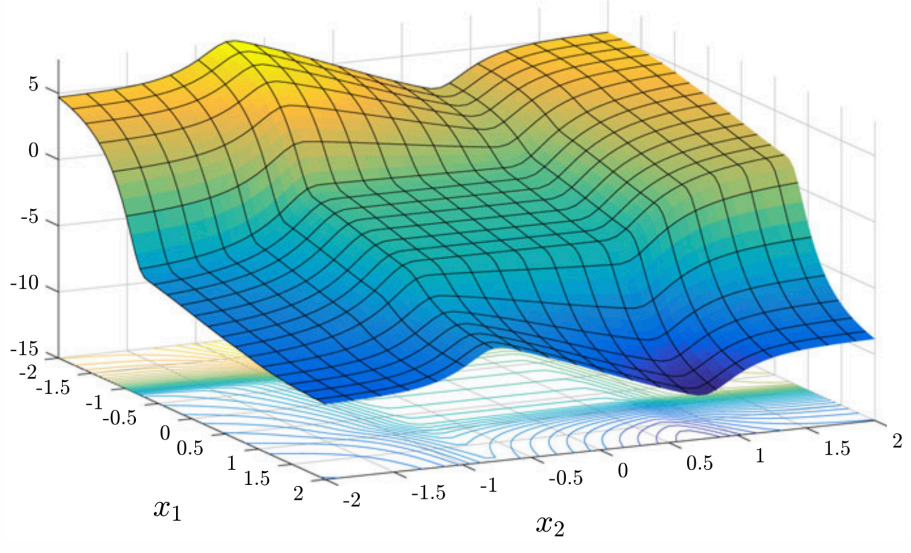
任意の $L, \Delta, \varepsilon > 0$ と任意の一次法 $\mathcal{A}$ に対し，ある関数 f が存在し，

- ∇f は L-Lipschitz 連続
- $f(x_0) - \inf_x f(x) \leq \Delta$　（x_0: $\mathcal{A}$ の初期点）
- $\mathcal{A}$ は $c\frac{L\Delta}{\varepsilon^2}$ 回以下の関数値・勾配計算では $\|\nabla f(x)\| \leq \varepsilon$ なる x を見つけられない　（c: 定数）

上の証明は難しいが，アルゴリズムを最急降下法に限れば
似たことが比較的容易に示せる 👉 演習 4

一次法にとって最悪の問題例

$\bar{f}_T(x_1, x_2, 0, ..., 0)$



図は [Carmon et al., 2020] より

- 最急降下法
- **3次正則化Newton法**
- 再始動Heavy-ball法

3 次正則化 Newton 法: 導入

定石（再掲）

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　Ⓑ $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

- $\bar{f}_k$ を「Taylor 展開 + 正則化項」で構成
- Lipschitz 連続性を使って Ⓑ を示す

- 最急降下法では 1 次の Taylor 展開を使った
- 2 次の Taylor 展開を使うと？　👉 **3 次正則化 Newton 法**

3 次正則化 Newton 法: 導入

ここで用いる上界補題

$\nabla^2 f$ が M-Lipschitz 連続ならば,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|^2_{\nabla^2 f(y)} + \frac{M}{6}\|x - y\|^3$$

$$\|x\|_A^2 := \langle Ax, x \rangle$$

比較: 最急降下法で用いた上界補題

∇f が L-Lipschitz 連続ならば,

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

3 次正則化 Newton 法: 導入

ここで用いる上界補題

$\nabla^2 f$ が M-Lipschitz 連続ならば,

$$f(x) \le f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|^2_{\nabla^2 f(y)} + \frac{M}{6}\|x - y\|^3$$

$$\|x\|_A^2 := \langle Ax, x \rangle$$

$$\bar{f}_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\|x - x_k\|^2_{\nabla^2 f(x_k)} + \frac{M}{6}\|x - x_k\|^3$$

とすると

Ⓐ $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　Ⓑ $f(x_{k+1}) \le \bar{f}_k(x_{k+1})$ 😊

仮定: $\nabla^2 f$ の M-Lipschitz 連続性

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq M\|x - y\| \qquad (\forall x, y \in \mathbb{R}^d)$$

3 次正則化 (Cubic Regularized) Newton 法 [Griewank, 1981]

$$x_{k+1} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \bar{f}_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\|x - x_k\|^2_{\nabla^2 f(x_k)} + \frac{M}{6}\|x - x_k\|^3 \right\}$$

計算量保証 [Nesterov and Polyak, 2006]

$$\min_{1 \leq i \leq k} \|\nabla f(x_i)\| \leq \left(\frac{12\sqrt{M}\Delta}{k}\right)^{2/3}$$

☞ ε-停留点を求めるための計算量: $\left\lceil \frac{12\sqrt{M}\Delta}{\varepsilon^{3/2}} \right\rceil = \mathrm{O}(\varepsilon^{-3/2})$

CRN 法: 子問題の非凸性

$$\min_{x\in\mathbb{R}^d} \quad f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2}\|x - x_k\|^2_{\nabla^2 f(x_k)} + \frac{M}{6}\|x - x_k\|^3$$

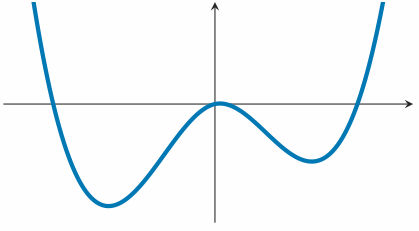
$$\downarrow \quad u := x - x_k, \quad g := \nabla f(x_k), \quad H := \nabla^2 f(x_k)$$

$$\min_{u\in\mathbb{R}^d} \quad \langle g, u\rangle + \frac{1}{2}\|u\|^2_H + \frac{M}{6}\|u\|^3$$

CRN 子問題は一般に非凸

👉 問題の特殊性を利用して解く

例: $u - 4u^2 + |u|^3$



CRN 法: 子問題の最適解の性質

$$\min_{u\in\mathbb{R}^d} \ \langle g, u\rangle + \frac{1}{2}\|u\|_H^2 + \frac{M}{6}\|u\|^3$$

補題: 子問題の最適解の性質 👍 演習 5

$u \in \mathbb{R}^d$ が上の問題の最適解ならば，

$$\sigma := \frac{M}{2}\|u\| \qquad \text{とおくと} \qquad (H+\sigma I)u = -g, \qquad H+\sigma I \succeq O$$

※ 実は $\Longleftarrow$ も成立

CRN 法: 子問題の最適解の性質

$$\min_{u\in\mathbb{R}^d} \ \langle g, u\rangle + \frac{1}{2}\|u\|_H^2 + \frac{M}{6}\|u\|^3$$

補題: 子問題の最適解の性質 👍 演習 5

$u \in \mathbb{R}^d$ が上の問題の最適解ならば,

$$\sigma := \frac{M}{2}\|u\| \qquad \text{とおくと} \qquad (H+\sigma I)u = -g, \qquad H+\sigma I \succeq O$$

- 直交行列で対角化: $H = V\Lambda V^\top$, 固有値 $\lambda_1 \le \cdots \le \lambda_d$
- $\tilde{u} := V^\top u, \ \ \tilde{g} := V^\top g$

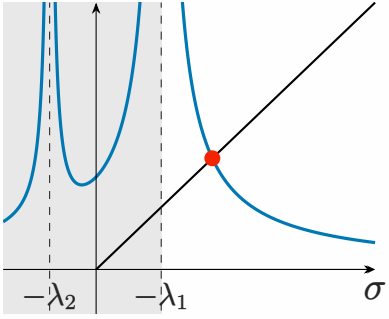
$$\sigma = \frac{M}{2}\|\tilde{u}\|, \qquad (\Lambda+\sigma I)\tilde{u} = -\tilde{g}, \qquad \sigma \ge -\lambda_1$$

$$\sigma = \frac{M}{2}\|\tilde{u}\|, \qquad (\Lambda + \sigma I)\tilde{u} = -\tilde{g}, \qquad \sigma \geq -\lambda_1$$

$\tilde{g}_1 \neq 0$ を仮定すると $\sigma > -\lambda_1$

このとき $\tilde{u}_i = -\dfrac{\tilde{g}_i}{\lambda_i + \sigma}$

$$\rightarrow \quad \sigma = \frac{M}{2}\sqrt{\sum_{i=1}^{d} \frac{\tilde{g}_i^2}{(\lambda_i + \sigma)^2}}$$



右辺は $\sigma > -\lambda_1$ で単調減少 **二分法で解ける** 😊

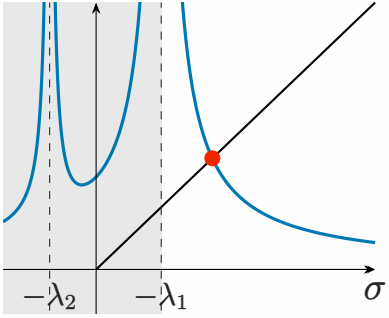
※ $\tilde{g}_1 = 0$ の場合は Hard case と呼ばれ，少々厄介 （割愛）

$$\sigma = \frac{M}{2}\|\tilde{u}\|, \qquad (\Lambda + \sigma I)\tilde{u} = -\tilde{g}, \qquad \sigma \geq -\lambda_1$$

$\tilde{g}_1 \neq 0$ を仮定すると $\sigma > -\lambda_1$

このとき $\tilde{u}_i = -\dfrac{\tilde{g}_i}{\lambda_i + \sigma}$

$$\rightarrow \quad \sigma = \frac{M}{2}\sqrt{\sum_{i=1}^{d} \frac{\tilde{g}_i^2}{(\lambda_i + \sigma)^2}}$$



- σ が十分大ならば，$\tilde{u}$ は $-\tilde{g}$ 方向に近い
- σ が $-\lambda_1$ に近づくにつれ，$\tilde{u}$ の $-\boldsymbol{e}_1$ 成分が増えていく

CRN 法: 解の更新方向の直観

CRN 法は，解の更新方向 $u := x - x_k$ として

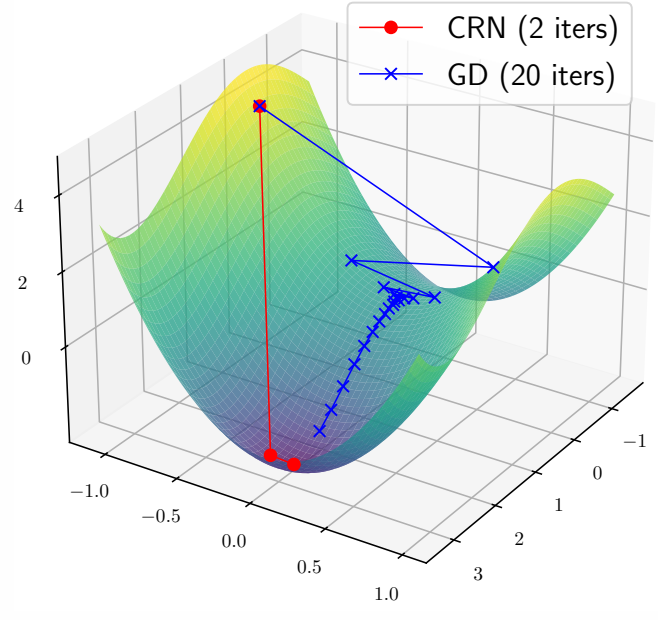
- $-\nabla f(x_k)$: **最急降下方向**
- $\nabla^2 f(x_k)$ の**最小**固有値に対応する**固有ベクトル方向**

を上手く組み合わせている！

固有ベクトル方向を活用することで**鞍点**を回避しやすい 😊

- σ が十分大ならば，$\tilde{u}$ は $-\tilde{g}$ 方向に近い
- σ が $-\lambda_1$ に近づくにつれ，$\tilde{u}$ の $-\boldsymbol{e}_1$ 成分が増えていく

数値例: CRN 法は鞍点を回避しやすい



ここからCRN法の**計算量解析**をします

最急降下法で学んだ定石通り

計算量解析の定石（再掲）

- $\|x_{k+1} - x_k\|$ が大 $\Longrightarrow$ **関数値減少量**が大 😊
- $\|x_{k+1} - x_k\|$ が小 $\Longrightarrow$ **勾配ノルム** が小 😊

} を両方示し，組み合わせる

CRN 法の解析: 関数値減少量の評価

既に示したこと

$u_k := x_{k+1} - x_k$, $\sigma_k := \frac{M}{2}\|u_k\|$ とおくと,

$$(H + \sigma_k I)u_k = -g, \qquad H + \sigma_k I \succeq O$$

CRN 法の解析: 関数値減少量の評価

既に示したこと

$u_k := x_{k+1} - x_k$, $\sigma_k := \frac{M}{2}\|u_k\|$ とおくと,

$$(H + \sigma_k I)u_k = -g, \qquad H + \sigma_k I \succeq O$$

上界補題

$$f(x_{k+1}) - f(x_k) \leq \langle g, u_k \rangle + \frac{1}{2}\|u_k\|_H^2 + \frac{M}{6}\|u_k\|^3$$

$(H + \sigma_k I)u_k = -g$

$$= -\|u_k\|_{H+\sigma_k I}^2 + \frac{1}{2}\|u_k\|_H^2 + \frac{M}{6}\|u_k\|^3$$

$$= -\frac{1}{2}\|u_k\|_{H+\sigma_k I}^2 - \frac{\sigma_k}{2}\|u_k\|^2 + \frac{M}{6}\|u_k\|^3$$

$H + \sigma_k I \succeq O$

$$\leq -\frac{\sigma_k}{2}\|u_k\|^2 + \frac{M}{6}\|u_k\|^3$$

$\sigma_k = \frac{M}{2}\|u_k\|$

$$= -\frac{M}{12}\|u_k\|^3$$

ここで使う補題

$\nabla^2 f$ が M-Lipschitz 連続ならば，

$$\left\|\nabla f(x) - \Big(\nabla f(y) + \nabla^2 f(y)(x-y)\Big)\right\| \le \frac{M}{2}\|x-y\|^2 \qquad (\forall x, y \in \mathbb{R}^d)$$

最急降下法で用いた補題（再掲）

∇f が L-Lipschitz 連続ならば，

$$\left|f(x) - \Big(f(y) + \langle \nabla f(y), x-y\rangle\Big)\right| \le \frac{L}{2}\|x-y\|^2 \qquad (\forall x, y \in \mathbb{R}^d)$$

と全く同様に示せる

CRN 法の解析: 勾配ノルムの評価

ここで使う補題

$\nabla^2 f$ が M-Lipschitz 連続ならば，

$$\left\|\nabla f(x) - \Big(\nabla f(y) + \nabla^2 f(y)(x-y)\Big)\right\| \le \frac{M}{2}\|x-y\|^2 \qquad (\forall x, y \in \mathbb{R}^d)$$

上の補題と三角不等式

$$\begin{aligned}
\|\nabla f(x_{k+1})\| &\le \left\|\nabla f(x_k) + \nabla^2 f(x_k) u_k\right\| + \frac{M}{2}\|u_k\|^2 \\
&= \|g + H u_k\| + \frac{M}{2}\|u_k\|^2 \\
(H + \sigma_k I)u_k = -g \quad &= \sigma_k \|u_k\| + \frac{M}{2}\|u_k\|^2 \\
\sigma_k = \tfrac{M}{2}\|u_k\| \quad &= M\|u_k\|^2
\end{aligned}$$

CRN 法の解析: 2 種類の評価を組み合わせる

Ⓐ $f(x_{k+1}) - f(x_k) \leq -\dfrac{M}{12}\|u_k\|^3$

Ⓑ $\|\nabla f(x_{k+1})\| \leq M\|u_k\|^2$

Ⓐ, Ⓑ より $\dfrac{\|\nabla f(x_{k+1})\|^{3/2}}{12\sqrt{M}} \leq f(x_k) - f(x_{k+1})$

これを k について足し，最急降下法のときと同様に評価すると，

$$\min_{1\leq i\leq k}\|\nabla f(x_i)\| \leq \left(\frac{12\sqrt{M}\Delta}{k}\right)^{2/3}$$

👉 計算量は $\left\lceil \frac{12\sqrt{M}\Delta}{\varepsilon^{3/2}} \right\rceil = \mathrm{O}(\varepsilon^{-3/2})$

CRN 法の解析: 2 種類の評価を組み合わせる

Ⓐ $f(x_{k+1}) - f(x_k) \leq -\dfrac{M}{12}\|u_k\|^3$

Ⓑ $\|\nabla f(x_{k+1})\| \leq M\|u_k\|^2$

Ⓐ, Ⓑ より $\|\nabla f(x_{k+1})\| \leq \cdots \|u_k\|^{3/2}$

実はもう少し嬉しいことが示せる！

これを k について足し，最急降下法のときと同様に評価すると，

$$\min_{1 \leq i \leq k} \|\nabla f(x_i)\| \leq \left(\frac{12\sqrt{M}\Delta}{k}\right)^{2/3}$$

👉 計算量は $\left\lceil \frac{12\sqrt{M}\Delta}{\varepsilon^{3/2}} \right\rceil = O(\varepsilon^{-3/2})$

CRN 法の解析: 2 次の停留点への収束

Ⓐ $f(x_{k+1}) - f(x_k) \leq -\dfrac{M}{12}\|u_k\|^3$

Ⓑ $\|\nabla f(x_{k+1})\| \leq M\|u_k\|^2$

Ⓒ $\nabla^2 f(x_{k+1}) \succeq \nabla^2 f(x_k) - M\|u_k\| I \succeq -\dfrac{3}{2}M\|u_k\| I$

Lipschitz 連続性

$\nabla^2 f(x_k) \succeq -\sigma_k I = -\frac{M}{2}\|u_k\| I$

Ⓐ, Ⓑ, Ⓒ を組み合わせると，$\|\nabla f(x)\| \leq \varepsilon$ に加えて

$$\nabla^2 f(x) \succeq -\frac{3}{2}\sqrt{M\varepsilon} I$$

も満たす点 x が $\left\lceil \frac{12\sqrt{M}\Delta}{\varepsilon^{3/2}} \right\rceil = \mathrm{O}(\varepsilon^{-3/2})$ 反復で見つかることがわかる

CRN 法の解析: 2 次の停留点への収束

$\varepsilon \to 0$ とすると，2 次の停留点:

$$\nabla f(x) = \mathbf{0}, \qquad \nabla^2 f(x) \succeq O$$

👉 **CRN 法は鞍点を回避しやすい**

Ⓐ, Ⓑ, Ⓒ を組み合わせると，$\|\nabla f(x)\| \le \varepsilon$ に加えて

$$\nabla^2 f(x) \succeq -\frac{3}{2}\sqrt{M\varepsilon}\, I$$

も満たす点 x が $\left\lceil \frac{12\sqrt{M}\Delta}{\varepsilon^{3/2}} \right\rceil = \mathrm{O}(\varepsilon^{-3/2})$ 反復で見つかることがわかる

ここからCRN法のその他の話題をいくつか紹介

- 子問題を近似的に解く
- Lipschitz定数 M の推定
- Lazy CRN法
- Hesse行列・ベクトル積の利用

子問題を近似的に解く [Cartis et al., 2012]

- 正則化を強めに設定:

$$\bar{f}_k(x) \coloneqq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\|x - x_k\|^2_{\nabla^2 f(x_k)} + \frac{M}{3}\|x - x_k\|^3$$

- 近似解の条件: Ⓐ $\|\nabla \bar{f}_k(x_{k+1})\| \leq \frac{M}{2}\|u_k\|^2$, Ⓑ $\bar{f}_k(x_{k+1}) \leq \bar{f}_k(x_k)$

$$u_k \coloneqq x_{k+1} - x_k$$

正則化を強めると子問題の**1次の停留点**を求めるだけで十分に 😊

※ 元問題の2次の停留点を求めたい場合はもう少し頑張る必要あり

子問題を近似的に解く：関数値減少量の評価

- 正則化を強めに設定:

$$\bar{f}_k(x) \coloneqq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\|x - x_k\|^2_{\nabla^2 f(x_k)} + \frac{M}{3}\|x - x_k\|^3$$

- 近似解の条件: Ⓐ $\|\nabla \bar{f}_k(x_{k+1})\| \leq \frac{M}{2}\|u_k\|^2$, Ⓑ $\bar{f}_k(x_{k+1}) \leq \bar{f}_k(x_k)$

$$u_k \coloneqq x_{k+1} - x_k$$

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_{k+1}) - \bar{f}_k(x_k) \\ &\overset{\text{上界補題}}{\leq} \left(\bar{f}_k(x_{k+1}) - \frac{M}{6}\|u_k\|^3\right) - \bar{f}_k(x_k) \\ &\overset{\text{Ⓑ}}{\leq} -\frac{M}{6}\|u_k\|^3 \end{aligned}$$

子問題を近似的に解く：勾配ノルムの評価

- 正則化を強めに設定:

$$\bar{f}_k(x) \coloneqq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\|x - x_k\|^2_{\nabla^2 f(x_k)} + \frac{M}{3}\|x - x_k\|^3$$

- 近似解の条件: Ⓐ $\|\nabla \bar{f}_k(x_{k+1})\| \le \frac{M}{2}\|u_k\|^2$, Ⓑ $\bar{f}_k(x_{k+1}) \le \bar{f}_k(x_k)$

$$\begin{aligned}
\|\nabla f(x_{k+1})\| &\le \left\|\nabla \bar{f}_k(x_{k+1})\right\| + \left\|\nabla f(x_{k+1}) - \nabla \bar{f}_k(x_{k+1})\right\| \\
&\overset{\text{Ⓐ}}{\le} \frac{M}{2}\|u_k\|^2 + \left\|\nabla f(x_{k+1}) - \nabla \bar{f}_k(x_{k+1})\right\| \\
&= \frac{M}{2}\|u_k\|^2 + \left\|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)u_k - M\|u_k\|u_k\right\| \\
&\overset{\|\text{緑}\| \le \frac{M}{2}\|u_k\|^2}{\le} \frac{M}{2}\|u_k\|^2 + \frac{M}{2}\|u_k\|^2 + M\|u_k\|^2 = 2M\|u_k\|^2
\end{aligned}$$

子問題を近似的に解く

- 正則化を強めに設定:

$$\bar{f}_k(x) \coloneqq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\|x - x_k\|^2_{\nabla^2 f(x_k)} + \frac{M}{3}\|x - x_k\|^3$$

- 近似解の条件: Ⓐ $\|\nabla \bar{f}_k(x_{k+1})\| \leq \frac{M}{2}\|u_k\|^2$, Ⓑ $\bar{f}_k(x_{k+1}) \leq \bar{f}_k(x_k)$

- 正則化を強めたおかげで証明がシンプルに 😊
- 「子問題の最適解の性質」の補題も不使用

Lipschitz 定数 M の推定

最急降下法と同様に Lipschitz 定数 M の推定が可能 😊

M の推定値 m に対し，解析で用いた不等式

- $f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), u_k \rangle + \frac{1}{2}\|u_k\|^2_{\nabla^2 f(x_k)} + \frac{m}{6}\|u_k\|^3$
- $\|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k) u_k\| \leq \frac{m}{2}\|u_k\|^2$

の成立をチェック

成立していなければ m を大きくしてやり直す

Lazy CRN 法 [Doikov et al., 2023]

$\nabla^2 f(\mathrm{x}_0)$	reuse Hessian $\longrightarrow$			$\nabla^2 f(\mathrm{x}_m)$
$\nabla f(\mathrm{x}_0)$	$\nabla f(\mathrm{x}_1)$	$\dots$	$\nabla f(\mathrm{x}_{m-1})$	$\nabla f(\mathrm{x}_m)$

（図は [Doikov et al., 2023] より）

定理: Lazy CRN 法の反復数 [Doikov et al., 2023]

- Hesse 行列を m 反復再利用
- 正則化項 $\frac{M}{6}\|u\|^3$ を $mM\|u\|^3$ に

$\longrightarrow$ 必要な反復数は $\mathrm{O}(\sqrt{m})$ 倍

計算量: $\mathrm{O}\left(\frac{\sqrt{M}\Delta}{\varepsilon^{3/2}}\right)(\mathsf{G}+\mathsf{H}) \quad \longrightarrow \quad \mathrm{O}\left(\frac{\sqrt{M}\Delta}{\varepsilon^{3/2}}\right)\left(\sqrt{m}\mathsf{G}+\frac{\mathsf{H}}{\sqrt{m}}\right)$

👉 $m \simeq \frac{\mathsf{H}}{\mathsf{G}}$ と設定すれば，計算量 $\mathrm{O}\left(\frac{\sqrt{M}\Delta}{\varepsilon^{3/2}}\right)\sqrt{\mathsf{G}\cdot\mathsf{H}}$ 😊

Hesse行列・ベクトル積の利用

$$\min_{u \in \mathbb{R}^d} \ \langle g, u \rangle + \frac{1}{2}\|u\|_H^2 + \frac{M}{6}\|u\|^3$$

dが大きいとき，$H := \nabla^2 f(x_k) \in \mathbb{R}^{d \times d}$を分解するのは非現実的... ☹

Hesse行列・ベクトル積 (HVP) $\nabla^2 f(x)v$をオラクルとして利用

- $x, v \in \mathbb{R}^d$に対し，$\nabla^2 f(x)v$の計算が容易な場合がある
  例: $\nabla^2 f(x)$が疎 or 低ランク，高速自動微分
- 勾配で近似も可能: $\nabla^2 f(x)v \simeq \frac{\nabla f(x+\delta v) - \nabla f(x)}{\delta}$ （非推奨）
- メモリが節約でき，$d \simeq 10^6$程度でも動く 😊

HVP が使える CRN 子問題の解法いろいろ

- **Lanczos 法** [Carmon and Duchi, 2018, 2020; Cartis et al., 2011; Gould and Simoncini, 2020] 等
- 共役勾配法 + 最小固有値計算 [Royer et al., 2020]
- $2(d+1)$ 次の一般化固有値問題に帰着 [Lieder, 2020]
  上の基となった信頼領域法版: [Adachi et al., 2017]
- $(d+1)$ 次の 2 次固有値問題に帰着 [Jia et al., 2022]
- 制約付き凸最適化に帰着 $\rightarrow$ 加速射影勾配法 [Jiang et al., 2021]
- 局所最適化とジャンプを繰り返す [Cristofari et al., 2019]
- 小さい固有値のみで方程式を近似 [Gao et al., 2022]
- ...

CRN 子問題のための Lanczos 法

$$\min_{u \in \mathbb{R}^d} \ \langle g, u \rangle + \frac{1}{2}\|u\|_H^2 + \frac{M}{6}\|u\|^3$$

- **Krylov 部分空間**に制限: $u \in \mathrm{span}\{g, Hg, \ldots, H^{m-1}g\}$
- $[g, Hg, \ldots, H^{m-1}g] = QR$ と **QR 分解**し，$u = Qv$ とおく

$$\min_{v \in \mathbb{R}^m} \ \langle Q^\top g, v \rangle + \frac{1}{2}\|v\|_T^2 + \frac{M}{6}\|v\|^3$$

- $T := Q^\top H Q$ は**三重対角** 😊
- HVP の計算回数は $\mathrm{O}(m)$
- Hard case 対策に乱数を使うことも（例: [Carmon and Duchi, 2018]）

部分空間の次元 m の選び方 [Carmon and Duchi, 2020, Lemma 6.1]

$$m \geq \left\lceil 24 \cdot 2^{3/4} \frac{\sqrt{L}}{(M\varepsilon)^{1/4}} \left(1 + \frac{1}{8} \log^2\left(\frac{4d}{\delta^2}\right)\right) \right\rceil = \Theta\left(\varepsilon^{-1/4}\right)$$

ならば，確率 $1 - \delta$ で，**十分な精度**の子問題の解が得られる

CRN 子問題のための Lanczos 法: m の設定

部分空間の次元 m の選び方 [Carmon and Duchi, 2020, Lemma 6.1]

$$m \geq \left\lceil 24 \cdot 2^{3/4} \frac{\sqrt{L}}{(M\varepsilon)^{1/4}} \left(1 + \frac{1}{8}\log^2\left(\frac{4d}{\delta^2}\right)\right) \right\rceil = \Theta\left(\varepsilon^{-1/4}\right)$$

ならば，確率 $1-\delta$ で，**十分な精度**の子問題の解が得られる

↓ $\times\, \mathrm{O}\left(\frac{\sqrt{M}\Delta}{\varepsilon^{3/2}}\right)$ （CRN 法の反復数）

CRN 法全体での HVP 計算回数 [Carmon and Duchi, 2020, Prop 6.3]

$$\mathrm{O}(1) \cdot \frac{\sqrt{L}M^{1/4}\Delta}{\varepsilon^{7/4}} \left(1 + \log^2\left(\frac{d}{\delta^2}\right) + \log\left(\frac{\sqrt{L}M^{1/4}\Delta}{\varepsilon^{7/4}}\right)\right) = \tilde{\mathrm{O}}\left(\varepsilon^{-7/4}\right)$$

ここまでのアルゴリズムの計算量の比較

	Lipschitz 仮定	計算量 /Δ	2 次停留点
❶ 最急降下法	∇f	$\mathrm{O}(L\varepsilon^{-2})$G	
❷ CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})$(G + H)	✓
❸ Lazy CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\sqrt{\mathrm{G}\cdot\mathrm{H}}$	✓
❹ CRN + Lanczos	∇f & $\nabla^2 f$	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})$HVP + $\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})$G	✓

ここまでのアルゴリズムの計算量の比較

	Lipschitz 仮定	計算量 /Δ	2 次停留点
❶ 最急降下法	∇f	$\mathrm{O}(L\varepsilon^{-2})\mathsf{G}$	
❷ CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})(\mathsf{G}+\mathsf{H})$	✓
❸ Lazy CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\sqrt{\mathsf{G}\cdot\mathsf{H}}$	✓
❹ CRN + Lanczos	∇f & $\nabla^2 f$	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})\mathsf{HVP}$ $+\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\mathsf{G}$	✓

- Lipschitz 定数 L, M
- 精度 ε
- オラクルコスト G, H, HVP
- 子問題の計算コスト

} などを考慮してアルゴリズム選択

👉 演習 1(c), (d)

ここまでのアルゴリズムの計算量の比較

	Lipschitz 仮定	計算量 /Δ	2 次停留点
❶ 最急降下法	∇f	$\mathrm{O}(L\varepsilon^{-2})\mathsf{G}$	
❷ CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})(\mathsf{G}+\mathsf{H})$	✓
❸ Lazy CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\sqrt{\mathsf{G}\cdot\mathsf{H}}$	✓
❹ CRN + Lanczos	∇f & $\nabla^2 f$	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})\mathsf{HVP}$ $+\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\mathsf{G}$	✓

例:

- 「d が大」のとき，❶ or ❹ が有力
- さらに「ε が小」or「$M \ll L$」で，HVP が使えるなら ❹ が有力

ここまでのアルゴリズムの計算量の比較

	Lipschitz 仮定	計算量 /Δ	2 次停留点
❶ 最急降下法	∇f	$\mathrm{O}(L\varepsilon^{-2})\mathsf{G}$	
❷ CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})(\mathsf{G}+\mathsf{H})$	✓
❸ Lazy CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\sqrt{\mathsf{G}\cdot\mathsf{H}}$	✓
❹ CRN + Lanczos	∇f & $\nabla^2 f$	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})\mathsf{HVP}$ $+\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\mathsf{G}$	✓

例:

- 「dが小」で H がそれほど大きくないなら，❷ or ❸ が有力
- さらに「H $\simeq$ G」なら ❷ が有力 （定数倍の都合）

- 最急降下法
- 3次正則化Newton法
- **再始動Heavy-ball法**

より良い計算量の追求

	Lipschitz 仮定	計算量 /Δ	2 次停留点
❶ 最急降下法	∇f	$\mathrm{O}(L\varepsilon^{-2})\mathsf{G}$	
❷ CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})(\mathsf{G}+\mathsf{H})$	✓
❸ Lazy CRN	$\nabla^2 f$	$\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\sqrt{\mathsf{G}\cdot\mathsf{H}}$	✓
❹ CRN + Lanczos	∇f & $\nabla^2 f$	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})\mathsf{HVP}$ $+\mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})\mathsf{G}$	✓
❺ これから説明	∇f & $\nabla^2 f$	$\mathrm{O}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})\mathsf{G}$	

∇f と $\nabla^2 f$ の Lipschitz 連続性を仮定する手法（一部）

	計算量 /Δ	2 次停留点	決定的
❹	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})$HVP $+ \mathrm{O}(\sqrt{M}\varepsilon^{-\frac{3}{2}})$G	✓	
ⓐ	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})$G	✓	
ⓑ	$\tilde{\mathrm{O}}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})$G		✓
❺	$\mathrm{O}(\sqrt{L}M^{\frac{1}{4}}\varepsilon^{-\frac{7}{4}})$G		✓

❹ CRN 系 [Agarwal et al., 2017; Carmon and Duchi, 2020]
加速勾配法 + 特殊な正則化 + 最小固有値近似計算 [Carmon et al., 2018]
Damped Newton-CG + 最小固有値近似計算 [Royer et al., 2020]

ⓐ HVP を使わず最小固有値近似計算 [Allen-Zhu and Li, 2018; Xu et al., 2017]
加速勾配法 + 摂動 [Jin et al., 2018]

ⓑ Convex until proven guilty [Carmon et al., 2017]

❺ Heavy-ball 法 or 加速勾配法 + 再始動 [Li and Lin, 2022, 2023; Marumo and Takeda, 2024a,b]

余談: 初期のアルゴリズムの例 [Agarwal et al., 2017]

Algorithm 1 FastCubic$(f, x_0, \varepsilon, L, L_2)$

Input: $f(x)$ that satisfies (2.1) with L_2 and L; a starting vector x_0; a target accuracy ε.

$\kappa \leftarrow (\frac{900}{\varepsilon L})^{1/2}$.
for $t = 0$ **to** ∞ **do**
  $m_t(h) \triangleq \nabla f(x_t)^\top h + \frac{h^\top \nabla^2 f(x_t) h}{2} + \frac{L}{6}\|h\|^3$
  $(\lambda, v, v_{\min}) \leftarrow$ FastCubicMin$\left(\nabla f(x_t),\ \nabla^2 f(x_t),\ L, L_2, \kappa\right)$
  $h' \leftarrow$ either v or $\frac{\lambda v_{\min}}{2L}$ whichever gives smaller value for $m_t(h)$;
  Set $x_{t+1} \triangleq x_t + h'$
  if $m_t(h') > -\frac{\varepsilon^{3/2}}{c\sqrt{L}}$ **then return** x_{t+1}. ◇ *c is a constant; we proved $c = 2.4 * 10^6$ works*
end for

$$c = 2.4 \cdot 10^6$$

Algorithm 2 FastCubicMin$(g, \mathbf{H}, L, L_2, \kappa)$ (main algorithm for cubic minimization)

Input: g a vector, $\mathbf{H}$ a symmetric matrix, parameters κ, L and L_2 which satisfies $-L_2\mathbf{I} \preceq \mathbf{H} \preceq L_2\mathbf{I}$.
Output: $(\lambda, v, v_{\min})$

$B \leftarrow L_2 + \sqrt{L\|g\| + \frac{1}{\kappa}}$.
$\tilde{\varepsilon} \leftarrow 1/\left(10000\left(\max\left\{L, \|g\|, \frac{3\kappa}{10}, B, 1\right\}\right)^{20}\right)$
$\lambda_0 \leftarrow 2B$.
for $i = 0$ **to** ∞ **do**
  Compute v such that $\|v + (\mathbf{H} + \lambda_i \mathbf{I})^{-1} g\| \le \tilde{\varepsilon}$.
  if $L\|v\| \in [2\lambda_i - L\tilde{\varepsilon}, 2\lambda_i + L\tilde{\varepsilon}]$ **then**
    return $(\lambda_i, v, \emptyset)$.
  else if $L\|v\| > 2\lambda_i + L\tilde{\varepsilon}$ **then**
    return BinarySearch$(\lambda_1 = \lambda_{i-1}, \lambda_2 = \lambda_i, \tilde{\varepsilon})$.
  else if $L\|v\| < 2\lambda_i - L\tilde{\varepsilon}$ **then**
    Let Power Method find vector w that is 9/10-appx leading eigenvector of $(\mathbf{H} + \lambda_i \mathbf{I})^{-1}$:

$$\tfrac{9}{10}\lambda_{\max}((\mathbf{H} + \lambda_i \mathbf{I})^{-1}) \le w^\top (\mathbf{H} + \lambda_i \mathbf{I})^{-1} w \le \lambda_{\max}((\mathbf{H} + \lambda_i \mathbf{I})^{-1}) \ .$$

    Compute a vector $\tilde{w}$ such that $\|\tilde{w} - (\mathbf{H} + \lambda_i \mathbf{I})^{-1} w\| \le \hat{\varepsilon} \triangleq \frac{1}{60B}$.
    $\Delta \leftarrow \frac{1}{2}\frac{1}{\tilde{w}^\top w - \hat{\varepsilon}}$.
    if $\Delta > \frac{1}{2\kappa}$ **then**
      $\tilde{\lambda}_{i+1} \leftarrow \lambda_i - \frac{\Delta}{2}$.
      if $\tilde{\lambda}_{i+1} > 0$ **then** $\lambda_{i+1} \leftarrow \tilde{\lambda}_{i+1}$ **else** $\lambda_{i+1} \leftarrow 0$
    else
      Use AppxPCA to find any unit vector $v_{\min}$ such that $v_{\min}^\top \mathbf{H} v_{\min} \le \lambda_{\min}(\mathbf{H}) + \frac{1}{10\kappa}$.
      Flip the sign of $v_{\min}$ so that $g^\top v_{\min} \le 0$.
      return $(\lambda_i, v, v_{\min})$.
    end if
  end if
end for

$$10000(\cdots)^{20}$$

Algorithm 3 BinarySearch$(\lambda_1, \lambda_2, \tilde{\varepsilon})$ (binary search subroutine)

Input: $\lambda_1 \ge \lambda_2$, $L\|(\mathbf{H} + \lambda_1 \mathbf{I})^{-1} g\| \le 2\lambda_1$, $L\|(\mathbf{H} + \lambda_2 \mathbf{I})^{-1} g\| \ge 2\lambda_2$, $\lambda_2 + \lambda_{\min}(\mathbf{H}) > 0$
Output: $(\lambda, v, \emptyset)$

for $t = 1$ **to** ∞ **do**
  $\lambda_{\mathrm{mid}} \leftarrow \frac{\lambda_1 + \lambda_2}{2}$
  Compute vector v such that $\|v + (\mathbf{H} + \lambda_{\mathrm{mid}} \mathbf{I})^{-1} g\| \le \tilde{\varepsilon}/2$
  if $L\|v\| \in [2\lambda_{\mathrm{mid}} - L\tilde{\varepsilon}, 2\lambda_{\mathrm{mid}} + L\tilde{\varepsilon}]$ **then**
    return $(\lambda_{\mathrm{mid}}, v, \emptyset)$
  else if $L\|v\| + L\tilde{\varepsilon} \le 2\lambda_{\mathrm{mid}}$ **then**
    $\lambda_1 \leftarrow \lambda_{\mathrm{mid}}$
  else if $L\|v\| - L\tilde{\varepsilon} \ge 2\lambda_{\mathrm{mid}}$ **then**
    $\lambda_2 \leftarrow \lambda_{\mathrm{mid}}$
  end if
end for

最新のアルゴリズムはよりシンプル & 実用的 😊

Heavy-ball 法と加速勾配法

- どちらも**慣性**を利用する手法
- 近似関数 $\bar{f}_k$ の最小化としては解釈しにくい（？）

Heavy-ball 法と加速勾配法

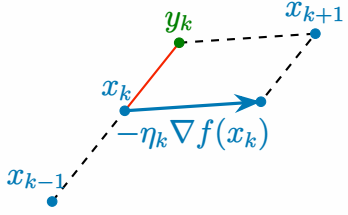
Heavy-ball 法 (HB) [Polyak, 1964]

$$x_{k+1} = x_k - \eta_k \nabla f(x_k) + \theta_k (x_k - x_{k-1})$$

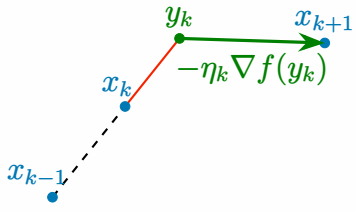
加速勾配法 (Accelerated gradient descent; AGD) [Nesterov, 1983]

$$x_{k+1} = x_k - \eta_k \nabla f\big(x_k + \theta_k (x_k - x_{k-1})\big) + \theta_k (x_k - x_{k-1})$$

- $\theta_k \in [0, 1]$: 慣性係数
- $x_{-1} := x_0$ と初期化



HB



AGD

Heavy-ball 法と加速勾配法

Heavy-ball 法 (HB) [Polyak, 1964]

$$x_{k+1} = x_k - \eta_k \nabla f(x_k) + \theta_k (x_k - x_{k-1})$$

加速勾配法 (Accelerated gradient descent; AGD) [Nesterov, 1983]

$$x_{k+1} = x_k - \eta_k \nabla f\big(x_k + \theta_k (x_k - x_{k-1})\big) + \theta_k (x_k - x_{k-1})$$

↓

$\eta_k = \eta,\ \theta_k = 1 - \gamma\sqrt{\eta}$ と定め，$\eta \to 0$　　($\gamma \geq 0$: 定数)

HB 微分方程式

$$\ddot{x}(t) = -\gamma \dot{x}(t) - \nabla f(x(t))$$

Heavy-ball 法と加速勾配法

Heavy-ball 法 (HB) [Polyak, 1964]

$$x_{k+1} = x_k - \eta_k \nabla f(x_k) + \theta_k (x_k - x_{k-1})$$

加速勾配法 (Accelerated gradient descent; AGD) [Nesterov, 1983]

$$x_{k+1} = x_k - \eta_k \nabla f\big(x_k + \theta_k (x_k - x_{k-1})\big) + \theta_k (x_k - x_{k-1})$$

凸の場合:

- HB は強凸 **2 次**関数に対し最良の計算量 [Polyak, 1964]
  η_k, θ_k が k に非依存だと**一般**の強凸関数に対し**非最適** [Goujaud et al., 2023]
- AGD は**一般**の凸関数 / 強凸関数に対し最適 [Nesterov, 1983, 2004]

Heavy-ball 法と加速勾配法

Heavy-ball 法 (HB) [Polyak, 1964]

$$x_{k+1} = x_k - \eta_k \nabla f(x_k) + \theta_k (x_k - x_{k-1})$$

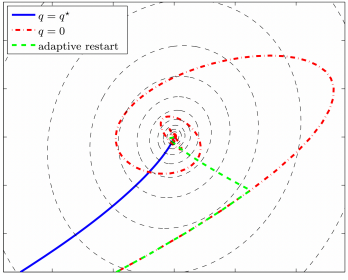
加速勾配法 (Accelerated gradient descent; AGD) [Nesterov, 1983]

$$x_{k+1} = x_k - \eta_k \nabla f\big(x_k + \theta_k (x_k - x_{k-1})\big) + \theta_k (x_k - x_{k-1})$$

非凸の場合:

- ∇f と $\nabla^2 f$ が Lipschitz 連続ならば，HB も AGD も **再始動** と組み合わせると強い

途中で $x_{k-1} = x_k$ と再設定



図は [O'Donoghue and Candès, 2015] より

最新の HB/AGD [Marumo and Takeda, 2024a,b] の利点

- Lipschitz 定数 L, M の自動推定が可能で，真の値は**入力不要**
- ε も**入力不要**（ε を停止条件のみに使用）

最新の HB/AGD [Marumo and Takeda, 2024a,b] の利点

- Lipschitz 定数 L, M の自動推定が可能で，真の値は**入力不要**
- ε も**入力不要**（ε を停止条件のみに使用）

参考: それ以前の一次法は L, M, ε の入力が必要

- **例:** [Li and Lin, 2022] では慣性係数 $\theta_k = 1 - \dfrac{2(M\varepsilon)^{1/4}}{\sqrt{L}}$
- 計算量解析で以下の不等式を使用．M の推定は難しそう ☹

- $f(x) - f(y) \leq \langle \nabla f(y), x - y \rangle + \dfrac{1}{2}\|x - y\|^2_{\nabla^2 f(y)} + \dfrac{M}{6}\|x - y\|^3$
- $\left\| \nabla f(x) - \nabla f(y) - \nabla^2 f(y)(x - y) \right\| \leq \dfrac{M}{2}\|x - y\|^2$

最新の HB/AGD [Marumo and Takeda, 2024a,b] の利点

- Lipschitz 定数 L, M の自動推定が可能で，真の値は**入力不要**
- ε も**入力不要**（ε を停止条件のみに使用）

参考 $\nabla^2 f(y)$ を含まない不等式のみを使って解析する

- **例:** [Li and Lin, 2022] では慣性係数 $\theta_k = 1 - \dfrac{2(M\varepsilon)^{1/4}}{\sqrt{L}}$
- 計算量解析で以下の不等式を使用．M の推定は難しそう ☹

- $f(x) - f(y) \leq \langle \nabla f(y), x - y \rangle + \dfrac{1}{2}\|x - y\|^2_{\nabla^2 f(y)} + \dfrac{M}{6}\|x - y\|^3$
- $\left\| \nabla f(x) - \nabla f(y) - \nabla^2 f(y)(x - y) \right\| \leq \dfrac{M}{2}\|x - y\|^2$

計算量解析に用いる Hessian-free 不等式 1

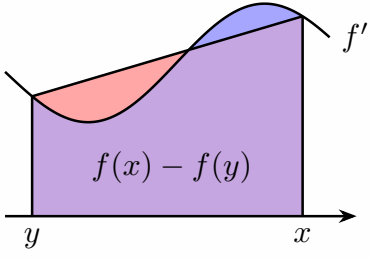
$$f(x) - f(y) \leq \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|^2_{\nabla^2 f(y)} + \frac{M}{6}\|x - y\|^3$$

の代わりに

$$f(x) - f(y) \leq \frac{1}{2}\langle \nabla f(x) + \nabla f(y), x - y \rangle + \frac{M}{12}\|x - y\|^3$$

👉 証明は演習 6

- **台形則**の誤差評価の多次元版
- $\nabla^2 f(y)$ なしでも
  $f(x) - f(y) \leq \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2$
  より良いオーダー



計算量解析に用いる Hessian-free 不等式 2

$$\left\|\nabla f(x) - \nabla f(y) - \nabla^2 f(y)(x-y)\right\| \leq \frac{M}{2}\|x-y\|^2$$

の代わりに

$$\left\|\nabla f(\bar{x}) - \sum_{i=1}^{n} \lambda_i \nabla f(x_i)\right\| \leq \frac{M}{2} \sum_{i=1}^{n} \lambda_i \|x_i - \bar{x}\|^2$$

👉 証明は演習 6

- $\bar{x} \coloneqq \sum_{i=1}^{n} \lambda_i x_i$: $x_1, \ldots, x_n$ の重みつき平均
- **直観:** M が小 $\Longrightarrow$ ∇f は affine に近い $\Longrightarrow$ 左辺が小
- **Jensen の不等式** $f(\bar{x}) - \sum_{i=1}^{n} \lambda_i f(x_i) \leq 0$ に類似

凸最適化で $f(\bar{x})$ が小さいことを示すのによく使われる

ここからHB法の**計算量解析**をします

- 簡単のため，L, M は既知と仮定
- ステップサイズ $\eta_k = \frac{1}{L}$，慣性係数 $\theta_k = 1$ と設定

計算量解析の定石（再掲）

- $\|x_{k+1} - x_k\|$ が大 $\Longrightarrow$ **関数値減少量**が大 😊
- $\|x_{k+1} - x_k\|$ が小 $\Longrightarrow$ **勾配ノルム** が小 😊

} を両方示し，組み合わせる

HB 法の解析: 関数値減少量 & 勾配ノルム の評価

これ以降 $S_k \coloneqq \sum_{i=0}^{k-1} \|x_{i+1} - x_i\|^2$, $\bar{x}_k \coloneqq \frac{1}{k} \sum_{i=0}^{k-1} x_i$, $x_k^\star \coloneqq \operatorname*{argmin}_{x \in \{x_1, \dots, x_k\}} f(x)$

謎の仮定 $\sqrt{(k-1)^5 S_{k-1}} \leq \frac{3L}{4M}$ の下，

- S_k が大 $\Longrightarrow$ 関数値減少量が大 😊: $f(x_k^\star) - f(x_0) \leq -\frac{L S_k}{4k}$
- S_k が小 $\Longrightarrow$ 勾配ノルム が小 😊: $\min_{1 \leq i \leq k} \|\nabla f(\bar{x}_i)\| \leq 3L \sqrt{\frac{S_k}{k^3}}$

👉 証明は演習 7, 8

これまでとの違い:

- k 反復分をまとめて評価．$\|x_{k+1} - x_k\|$ の代わりに S_k が登場
- 平均解 $\bar{x}_k$ で勾配ノルムを評価

HB 法の解析: 関数値減少量 & 勾配ノルム の評価

これ以降 $S_k \coloneqq \sum_{i=0}^{k-1} \|x_{i+1} - x_i\|^2$, $\bar{x}_k \coloneqq \frac{1}{k} \sum_{i=0}^{k-1} x_i$, $x_k^\star \coloneqq \operatorname*{argmin}_{x \in \{x_1, \ldots, x_k\}} f(x)$

謎の仮定 $\sqrt{(k-1)^5 S_{k-1}} \leq \frac{3L}{4M}$ の下，

- S_k が大 $\Longrightarrow$ 関数値減少量が大 😊: $f(x_k^\star) - f(x_0) \leq -\frac{L S_k}{4k}$
- S_k が小 $\Longrightarrow$ 勾配ノルム　が小 😊: $\min_{1 \leq i \leq k} \|\nabla f(\bar{x}_i)\| \leq 3L \sqrt{\frac{S_k}{k^3}}$

👉 証明は演習 7, 8

- 謎の仮定の動機は？　→ 次ページ
- 謎の仮定の正当化は？　→ **再始動**の活用

2 種類の不等式

- $f(x_{i+1}) - f(x_i) \leq \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \dfrac{L}{2} \|x_{i+1} - x_i\|^2$
- $f(x_{i+1}) - f(x_i) \leq \dfrac{1}{2} \langle \nabla f(x_{i+1}) + \nabla f(x_i), x_{i+1} - x_i \rangle + \dfrac{M}{12} \|x_{i+1} - x_i\|^3$

に重みを付けて $0 \leq i < k$ について足し，頑張ると，

$$f(x_k^\star) - f(x_0) \leq \frac{1}{2k-1} \left(-\frac{LS_k}{2} + \frac{M}{6}(k-1) \sum_{i=0}^{k-2} \|x_{i+1} - x_i\|^3 \right)$$

☞ 厄介な $\sum_i \|x_{i+1} - x_i\|^3$ を仮定 $\sqrt{(k-1)^5 S_{k-1}} \leq \dfrac{3L}{4M}$ で処理

$$f(x_k^\star) - f(x_0) \leq -\frac{LS_k}{4k}$$

もう一つの Hessian-free 不等式

$$\left\| \nabla f(\bar{x}_k) - \frac{1}{k} \sum_{i=0}^{k-1} \nabla f(x_i) \right\| \leq \frac{M}{2k} \sum_{i=0}^{k-1} \|x_i - \bar{x}_k\|^2$$

を使って頑張ると，

$$\min_{1 \leq i \leq k} \|\nabla f(\bar{x}_i)\| \leq 2L \sqrt{\frac{S_k}{k^3}} + \frac{M}{6}(k-1)S_{k-1}$$

☞ $\frac{M}{6}(k-1)S_{k-1}$ を仮定 $\sqrt{(k-1)^5 S_{k-1}} \leq \frac{3L}{4M}$ で処理

$$\min_{1 \leq i \leq k} \|\nabla f(\bar{x}_i)\| \leq 3L \sqrt{\frac{S_k}{k^3}}$$

再始動 Heavy-ball 法

1: $(x_0, x_{-1}) \leftarrow (x_{\mathrm{init}}, x_{\mathrm{init}}),\ \ k \leftarrow 0$
2: **loop**
3: $\quad k \leftarrow k+1$
4: $\quad x_k \leftarrow 2x_{k-1} - x_{k-2} - \frac{1}{L}\nabla f(x_{k-1})$ ▷ $\theta_k = 1$ の HB 法
5: $\quad$ **if** $\sqrt{k^5 S_k} > \frac{3L}{4M}$ **:**
6: $\quad\quad (x_0, x_{-1}) \leftarrow (x_k^\star, x_k^\star),\ \ k \leftarrow 0$

再始動により常に $\sqrt{(k-1)^5 S_{k-1}} \le \dfrac{3L}{4M}$ を保証

👉 $\begin{cases} f(x_k^\star) - f(x_0) \le -\dfrac{LS_k}{4k} \\ \min\limits_{1\le i\le k} \|\nabla f(\bar{x}_i)\| \le 3L\sqrt{\dfrac{S_k}{k^3}} \end{cases}$ が常に成立 😊

再始動 Heavy-ball 法

1: $(x_0, x_{-1}) \leftarrow (x_{\text{init}}, x_{\text{init}}),\ \ k \leftarrow 0$
2: **loop**
3: $\quad k \leftarrow k + 1$
4: $\quad x_k \leftarrow 2x_{k-1} - x_{k-2} - \frac{1}{L}\nabla f(x_{k-1})$ $\quad \triangleright\ \theta_k = 1$ の HB 法
5: $\quad$ **if** $\sqrt{k^5 S_k} > \frac{3L}{4M}$ **:**
6: $\quad\quad (x_0, x_{-1}) \leftarrow (x_k^\star, x_k^\star),\ \ k \leftarrow 0$

とても大雑把な説明:

- HB 法は，強凸 2 次関数に対しては最適
- 再始動 HB は，2 次関数でないことの悪影響を再始動でリセット

再始動 HB 法の解析: 2 種類の評価を組み合わせる

Ⓐ $f(x_k^\star) - f(x_0) \leq -\dfrac{LS_k}{4k}$

Ⓑ $\min\limits_{1 \leq i \leq k} \|\nabla f(\bar{x}_i)\| \leq 3L\sqrt{\dfrac{S_k}{k^3}}$

- $\varepsilon < \min\limits_{1 \leq i \leq k} \|\nabla f(\bar{x}_i)\|$ と Ⓑ より，$S_k > \left(\dfrac{\varepsilon}{3L}\right)^2 k^3$
- 第 k 反復で再始動時，条件 $\sqrt{k^5 S_k} > \dfrac{3L}{4M}$ より $S_k > \left(\dfrac{3L}{4M}\right)^2 k^{-5}$

よって，

$$S_k = S_k^{7/8} S_k^{1/8} > \left(\left(\frac{\varepsilon}{3L}\right)^2 k^3\right)^{7/8} \left(\left(\frac{3L}{4M}\right)^2 k^{-5}\right)^{1/8} = \frac{\varepsilon^{7/4}}{3\sqrt{6L^3} M^{1/4}} k^2$$

再始動 HB 法の解析: 2 種類の評価を組み合わせる

- **A** $f(x_k^\star) - f(x_0) \leq -\dfrac{LS_k}{4k}$
- **B** $\min\limits_{1 \leq i \leq k} \|\nabla f(\bar{x}_i)\| \leq 3L\sqrt{\dfrac{S_k}{k^3}}$
- **C** $S_k > \dfrac{\varepsilon^{7/4}}{3\sqrt{6L^3}M^{1/4}}k^2$

A, **C** を組み合わせると，$\dfrac{\varepsilon^{7/4}}{12\sqrt{6L}M^{1/4}}k \leq f(x_0) - f(x_k^\star)$

(最後の再始動までの合計反復数) $= \sum k \leq \dfrac{12\sqrt{6L}M^{1/4}}{\varepsilon^{7/4}}\Delta$

最後の再始動以降の反復数は $\dfrac{6\sqrt{L\Delta}}{\varepsilon} = \mathrm{O}(\varepsilon^{-1})$ （容易，詳細割愛）

再始動 HB 法の計算量保証

$\bar{x}_k := \frac{1}{k}\sum_{i=0}^{k-1} x_i$ が $\|\nabla f(\bar{x}_k)\| \le \varepsilon$ を初めて満たすまでの合計反復数は

$$\frac{12\sqrt{6L}M^{1/4}\Delta}{\varepsilon^{7/4}} + \mathrm{O}(\varepsilon^{-1}) \qquad \text{以下}$$

アルゴリズム（再掲）

1: $(x_0, x_{-1}) \leftarrow (x_{\mathrm{init}}, x_{\mathrm{init}}),\ \ k \leftarrow 0$
2: **loop**
3: $\quad k \leftarrow k+1$
4: $\quad x_k \leftarrow 2x_{k-1} - x_{k-2} - \frac{1}{L}\nabla f(x_{k-1})$ ▷ $\theta_k = 1$ の HB 法
5: $\quad$ **if** $\sqrt{k^5 S_k} > \frac{3L}{4M}$ **:** ▷ $S_k := \sum_{i=0}^{k-1}\|x_{i+1} - x_i\|^2$
6: $\quad\quad (x_0, x_{-1}) \leftarrow (x_k^\star, x_k^\star),\ \ k \leftarrow 0$

ここからHB法の発展的話題を紹介

- L, M が未知の場合
- Universal HB法
- 再始動なしHB法？

L, M が未知の場合: L の推定

最急降下法のときと似た方法で推定したいが，
各反復で L の推定値 ℓ が変わると解析が困難に ☹

方針:

- 不等式チェックにより ℓ が小さすぎると判明したら再始動
- 再始動時以外には ℓ を更新しない

L, M が未知の場合: L の推定

1: $(x_0, x_{-1}) \leftarrow (x_{\text{init}}, x_{\text{init}}),\ \ell \leftarrow \ell_{\text{init}},\ k \leftarrow 0$
2: **loop**
3: $\quad k \leftarrow k + 1$
4: $\quad x_k \leftarrow 2x_{k-1} - x_{k-2} - \frac{1}{\ell}\nabla f(x_{k-1})$
5: $\quad$ **if** $f(x_k) - f(x_{k-1}) > \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle + \frac{\ell}{2}\|x_k - x_{k-1}\|^2$ **:**
6: $\quad\quad (x_0, x_{-1}) \leftarrow (x_k^\star, x_k^\star),\ k \leftarrow 0,\ \ell \leftarrow \alpha\ell$ $\quad \triangleright\ \alpha > 1$: 定数
7: $\quad$ **else if** $\sqrt{k^5 S_k} > \frac{3\ell}{4M}$ **:**
8: $\quad\quad (x_0, x_{-1}) \leftarrow (x_k^\star, x_k^\star),\ k \leftarrow 0$

- L が既知の場合と同様の計算量保証が成立
- 8行目に ℓ を減少させるステップを入れても OK 😊

L, M が未知の場合: M の推定

M の推定値 m は毎反復更新しても解析に影響なし

解析で使った不等式

- $f(x_i) - f(x_{i-1}) \leq \frac{1}{2}\langle \nabla f(x_i) + \nabla f(x_{i-1}), x_i - x_{i-1} \rangle + \frac{m}{12}\|x_i - x_{i-1}\|^3 \quad (1 \leq i \leq k)$
- $\|\nabla f(\bar{x}_{i+1})\| \leq \frac{\ell}{i+1}\|x_{i+1} - x_i\| + \frac{m}{4} i S_i \quad (1 \leq i \leq k)$

を満たす最小の m を第 k 反復の推定値 m_k とする

$$m_k := \max\left\{ \max_{1 \leq i \leq k} \frac{12}{\|x_i - x_{i-1}\|^3}\left(f(x_i) - f(x_{i-1}) - \frac{1}{2}\langle \nabla f(x_i) + \nabla f(x_{i-1}), x_i - x_{i-1} \rangle \right), \right.$$
$$\left. \max_{1 \leq i \leq k} \frac{4}{i S_i}\left(\|\nabla f(\bar{x}_{i+1})\| - \frac{\ell}{i+1}\left\| x_i - x_{i-1} - \frac{1}{\ell}\nabla f(x_i) \right\| \right) \right\}$$

L, M が未知の場合: M の推定

M の推定値 m は毎反復更新しても解析に影響なし

解析で使った不等式

- $f(x_i) - f(x_{i-1}) \leq \frac{1}{2}\langle \nabla f(x_i) + \nabla f(x_{i-1}), x_i - x_{i-1} \rangle + \frac{m}{12}\|x_i - x_{i-1}\|^3 \quad (1 \leq i \leq k)$
- $\|\nabla f(\bar{x}_{i+1})\| \leq \frac{\ell}{i+1}\|x_{i+1} - x_i\| + \frac{m}{4} i S_i \quad (1 \leq i \leq k)$

を満たす最小の m を第 k 反復の推定値 m_k とする

M の推定を容易にするための鍵:

- **Hessian-free 不等式**の利用
- x_k の計算に m_k 不使用

L, M が未知の場合: 完全版アルゴリズム

$(x_0, x_{-1}) \leftarrow (x_{\mathrm{init}}, x_{\mathrm{init}}),\ \ell \leftarrow \ell_{\mathrm{init}},\ k \leftarrow 0$

loop

  $k \leftarrow k+1$

  $x_k \leftarrow 2x_{k-1} - x_{k-2} - \frac{1}{\ell}\nabla f(x_{k-1})$

  M の推定値 m_k の計算

  if $f(x_k) - f(x_{k-1}) > \langle \nabla f(x_{k-1}), x_k - x_{k-1}\rangle + \frac{\ell}{2}\|x_k - x_{k-1}\|^2$ **:**

    $(x_0, x_{-1}) \leftarrow (x_k^\star, x_k^\star),\ k \leftarrow 0,\ \ell \leftarrow \alpha\ell$

  else if $\sqrt{k^5 S_k} > \frac{3\ell}{4m_k}$ **:**

    $(x_0, x_{-1}) \leftarrow (x_k^\star, x_k^\star),\ k \leftarrow 0,\ \ell \leftarrow \beta\ell$

Universal HB 法

- ∇f が Lipschitz 連続なら，最急降下法で $\mathrm{O}(\varepsilon^{-2})$
- $\nabla f, \nabla^2 f$ が Lipschitz 連続なら，再始動 HB 法で $\mathrm{O}(\varepsilon^{-7/4})$

仮定によってアルゴリズムを使い分けるのは大変 ☹

疑問: 単一のアルゴリズムで両方の保証を達成可能か？

Universal HB 法

- ∇f が Lipschitz 連続なら，最急降下法で $\mathrm{O}(\varepsilon^{-2})$
- $\nabla f, \nabla^2 f$ が Lipschitz 連続なら，再始動 HB 法で $\mathrm{O}(\varepsilon^{-7/4})$

仮定によってアルゴリズムを使い分けるのは大変 ☹

疑問: 単一のアルゴリズムで両方の保証を達成可能か？

- **可能！** [Marumo and Takeda, 2024b]
- しかも「中間」の仮定にも対応可能

Universal HB 法: 計算量の比較

仮定: $\nabla^2 f$ の Hölder 連続性

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq H_\nu \|x - y\|^\nu \qquad (\forall x, y \in \mathbb{R}^d)$$

※ ∇f が L-Lipschitz 連続 $\implies$ $H_0 \leq 2L$

ν	0	$[0,1]$	1
p in $\mathrm{O}(\varepsilon^{-p})$	2	$\dfrac{4+3\nu}{2+2\nu}$	$\dfrac{7}{4}$
	最急降下法	[M–Takeda, 2024b]	[Li and Lin, 2022]

より強い仮定，より良い計算量

Hölder 連続性を使う計算量解析の例: [Cartis et al., 2017, 2019; Devolder et al., 2014; Dvurechensky, 2017; Ghadimi et al., 2019; Grapiglia and Nesterov, 2017, 2019, 2020; Lan, 2015; Nesterov, 2015]...

Universal HB 法: 計算量の詳細

Universal 再始動 HB 法の反復数 [Marumo and Takeda, 2024b]

$$\inf_{\nu\in[0,1]} \left\{ 91\sqrt{L} H_\nu^{\frac{1}{2+2\nu}} \varepsilon^{-\frac{4+3\nu}{2+2\nu}} \Delta \right\} + \mathrm{O}(\varepsilon^{-1})$$

ポイント:

- **アルゴリズムは ν に非依存** ☞ Hölder 連続性についての知識不要
- $H_\nu = +\infty$ なる ν があっても OK
- $H_0 \leq 2L$ より，計算量は最急降下法の $\mathrm{O}\left(\frac{L\Delta}{\varepsilon^2}\right)$ 以下
- $\nu = 1$ とすると $\mathrm{O}\left(\frac{\sqrt{L} M^{1/4} \Delta}{\varepsilon^{7/4}}\right)$

再始動なし HB 法？

- HB 法の $O(\varepsilon^{-7/4})$ 収束を示すのに再始動を用いた
- $O(\varepsilon^{-7/4})$ や $\tilde{O}(\varepsilon^{-7/4})$ の他手法は再始動以上に複雑なことをする
- **凸**の場合，加速勾配法は再始動なしで最良の計算量を達成

疑問: 再始動なしのシンプルな HB 法でも $O(\varepsilon^{-7/4})$ は達成可能か？

再始動なし HB 法？

- HB 法の $O(\varepsilon^{-7/4})$ 収束を示すのに再始動を用いた
- $O(\varepsilon^{-7/4})$ や $\tilde{O}(\varepsilon^{-7/4})$ の他手法は再始動以上に複雑なことをする
- **凸**の場合，加速勾配法は再始動なしで最良の計算量を達成

疑問: 再始動なしのシンプルな HB 法でも $O(\varepsilon^{-7/4})$ は達成可能か？

- **HB 微分方程式**は再始動なしで $O(\varepsilon^{-7/4})$ 収束 [Okamura et al., 2024]
- 微分方程式を離散化して得られる HB 法の解析は
  完全解決していない

HB 微分方程式の収束解析 [Okamura et al., 2024]

$T > 0$: 定数

- $\gamma := (3M)^{\frac{2}{7}} \left(\dfrac{\Delta}{T}\right)^{\frac{1}{7}}$
- $\ddot{x}(t) = -\gamma \dot{x}(t) - \nabla f(x(t)), \qquad x(0) = x_0, \qquad \dot{x}(0) = \mathbf{0}$
- $\bar{x}(t) := \displaystyle\int_0^t \frac{\gamma \mathrm{e}^{\gamma s}}{\mathrm{e}^{\gamma t} - 1} x(s)\,\mathrm{d}s$

$$\Longrightarrow \qquad \min_{0 \le t \le T} \|\nabla f(\bar{x}(t))\| \le \frac{7}{6}(3M)^{\frac{1}{7}} \left(\frac{\Delta}{T}\right)^{\frac{4}{7}} + \mathrm{O}\left(T^{-\frac{10}{7}}\right)$$

- 再始動 HB では $\gamma = 0$ $(\theta_k = 1)$．上は摩擦 $(\gamma > 0)$ で再始動を代替
- 離散化して $\mathrm{O}(\varepsilon^{-7/4})$ の HB 法を得るには現状追加の仮定が必要

まとめ

最急降下法，**CRN**，**再始動 HB**，**Levenberg–Marquardt 法** 演習 9

アルゴリズム設計の定石

- f の近似 $\bar{f}_k$ を最小化する点を x_{k+1} とする
- **A** $\bar{f}_k(x_k) = f(x_k),\ \nabla \bar{f}_k(x_k) = \nabla f(x_k)$　　**B** $f(x_{k+1}) \leq \bar{f}_k(x_{k+1})$

他: 強凸な $\bar{f}_k$，強めの正則化，慣性，Lipschitz 定数の推定，...

計算量解析の定石

- $\|x_{k+1} - x_k\|$ が大 $\Longrightarrow$ **関数値減少量**が大
- $\|x_{k+1} - x_k\|$ が小 $\Longrightarrow$ **勾配ノルム** が小

} を両方示し，組み合わせる

他: Lipschitz 連続性から **B** を示す，(最小値) $\leq$ (平均値)，...

参考文献 I

S. Adachi, S. Iwata, Y. Nakatsukasa, and A. Takeda. Solving the trust-region subproblem by a generalized eigenvalue problem. SIAM Journal on Optimization, 27(1):269–291, 2017. URL https://doi.org/10.1137/16M1058200.

N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, pages 1195–1199, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345286. URL https://doi.org/10.1145/3055399.3055464.

Z. Allen-Zhu and Y. Li. NEON2: Finding local minima via first-order oracles. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d4b2aeb2453bdadaa45cbe9882ffefcf-Paper.pdf.

L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. Pacific Journal of Mathematics, 16(1):1–3, 1966. URL https://dx.doi.org/10.2140/pjm.1966.16.1.

Y. Carmon and J. C. Duchi. Analysis of Krylov subspace solutions of regularized non-convex quadratic problems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/349f36aa789af083b8e26839bd498af9-Paper.pdf.

Y. Carmon and J. C. Duchi. First-order methods for nonconvex quadratic minimization. SIAM Review, 62(2):395–436, 2020. URL https://doi.org/10.1137/20M1321759.

参考文献 II

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In D. Precup and Y. W. Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 654–663. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/carmon17a.html.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. SIAM Journal on Optimization, 28(2):1751–1772, 2018. URL https://doi.org/10.1137/17M1114296.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. Mathematical Programming, 184(1):71–120, 2020. URL https://doi.org/10.1007/s10107-019-01406-y.

C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results. Mathematical Programming, 127(2):245–295, 2011. URL https://doi.org/10.1007/s10107-009-0286-5.

C. Cartis, N. I. M. Gould, and P. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. SIAM Journal on Optimization, 22(1):66–86, 2012. URL https://doi.org/10.1137/100812276.

C. Cartis, N. I. M. Gould, and P. L. Toint. Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients. Optimization Methods and Software, 32(6): 1273–1298, 2017. URL https://doi.org/10.1080/10556788.2016.1268136.

参考文献 III

C. Cartis, N. I. M. Gould, and P. L. Toint. Universal regularization methods: Varying the power, the smoothness and the accuracy. SIAM Journal on Optimization, 29(1):595–615, 2019. URL https://doi.org/10.1137/16M1106316.

A. Cristofari, T. Dehghan Niri, and S. Lucidi. On global minimizers of quadratic functions with cubic regularization. Optimization Letters, 13(6):1269–1283, 2019. URL https://doi.org/10.1007/s11590-018-1316-0.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming, 146(1):37–75, 2014. URL https://doi.org/10.1007/s10107-013-0677-5.

N. Doikov, E. M. Chayti, and M. Jaggi. Second-order optimization with lazy Hessians. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 8138–8161. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/doikov23a.html.

P. Dvurechensky. Gradient method with inexact oracle for composite non-convex optimization. arXiv preprint, 2017. URL https://arxiv.org/abs/1703.09180.

Y. Gao, M.-C. Yue, and M. Ng. Approximate secular equations for the cubic regularization subproblem. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 14250–14260. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

S. Ghadimi, G. Lan, and H. Zhang. Generalized uniformly optimal methods for nonlinear programming. Journal of Scientific Computing, 79(3):1854–1881, 2019. URL https://doi.org/10.1007/s10915-019-00915-4.

参考文献 IV

B. Goujaud, A. Taylor, and A. Dieuleveut. Provable non-accelerations of the heavy-ball method. arXiv preprint arXiv:2307.11291, 2023.

N. I. M. Gould and V. Simoncini. Error estimates for iterative algorithms for minimizing regularized quadratic subproblems. Optimization Methods and Software, 35(2):304–328, 2020. URL https://doi.org/10.1080/10556788.2019.1670177.

G. N. Grapiglia and Y. Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. SIAM Journal on Optimization, 27(1):478–506, 2017. URL https://doi.org/10.1137/16M1087801.

G. N. Grapiglia and Y. Nesterov. Accelerated regularized Newton methods for minimizing composite convex functions. SIAM Journal on Optimization, 29(1):77–99, 2019. URL https://doi.org/10.1137/17M1142077.

G. N. Grapiglia and Y. Nesterov. Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives. SIAM Journal on Optimization, 30(4):2750–2779, 2020. URL https://doi.org/10.1137/19M1259432.

A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical report, Technical report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1981.

X. Jia, X. Liang, C. Shen, and L.-H. Zhang. Solving the cubic regularization model by a nested restarting Lanczos method. SIAM Journal on Matrix Analysis and Applications, 43(2):812–839, 2022. URL https://doi.org/10.1137/21M1436324.

R. Jiang, M.-C. Yue, and Z. Zhou. An accelerated first-order method with complexity analysis for solving cubic regularization subproblems. Computational Optimization and Applications, 79(2):471–506, 2021. URL https://doi.org/10.1007/s10589-021-00274-7.

C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In S. Bubeck, V. Perchet, and P. Rigollet, editors, Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 1042–1085. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/jin18a.html.

G. Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. Mathematical Programming, 149(1):1–45, 2015. URL https://doi.org/10.1007/s10107-013-0737-x.

H. Li and Z. Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $O(\varepsilon^{-7/4})$ complexity. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 12901–12916. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22o.html.

H. Li and Z. Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $O(\varepsilon^{-7/4})$ complexity. Journal of Machine Learning Research, 24(157):1–37, 2023. URL http://jmlr.org/papers/v24/22-0522.html.

F. Lieder. Solving large-scale cubic regularization by a generalized eigenvalue problem. SIAM Journal on Optimization, 30(4):3345–3358, 2020. URL https://doi.org/10.1137/19M1291388.

参考文献 VI

N. Marumo and A. Takeda. Parameter-free accelerated gradient descent for nonconvex minimization. SIAM Journal on Optimization, 34(2):2093–2120, 2024a. URL https://doi.org/10.1137/22M1540934.

N. Marumo and A. Takeda. Universal heavy-ball method for nonconvex optimization under Hölder continuous Hessians. Mathematical Programming, 2024b. URL https://doi.org/10.1007/s10107-024-02100-4.

Y. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 269(3):372–376, 1983.

Y. Nesterov. Introductory Lectures on Convex Optimization: A Basic Course. Springer, New York, 2004. URL https://doi.org/10.1007/978-1-4419-8853-9.

Y. Nesterov. Universal gradient methods for convex optimization problems. Mathematical Programming, 152(1): 381–404, 2015. URL https://doi.org/10.1007/s10107-014-0790-0.

Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. Mathematical Programming, 108(1):177–205, 2006. URL https://doi.org/10.1007/s10107-006-0706-8.

B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. Foundations of Computational Mathematics, 15(3):715–732, 2015. URL https://doi.org/10.1007/s10208-013-9150-3.

K. Okamura, N. Marumo, and A. Takeda. Primitive heavy-ball dynamics achieves $O(\varepsilon^{-7/4})$ convergence for nonconvex optimization. arXiv preprint arXiv:2406.06100, 2024.

参考文献 VII

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1–17, 1964. ISSN 0041-5553. URL https://doi.org/10.1016/0041-5553(64)90137-5.

H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. The Computer Journal, 3 (3):175–184, 01 1960. ISSN 0010-4620. URL https://doi.org/10.1093/comjnl/3.3.175.

C. W. Royer, M. O'Neill, and S. J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. Mathematical Programming, 180(1):451–488, 2020. URL https://doi.org/10.1007/s10107-019-01362-7.

Y. Xu, R. Jin, and T. Yang. NEON+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. arXiv preprint, 2017. URL https://arxiv.org/abs/1712.01033.