平成 24 年度 修士論文

凸計画問題に対する劣勾配アルゴリズムの新しい枠組み提案

東京工業大学大学院
情報理工学研究科 数理・計算科学専攻

学籍番号 11M37028
伊藤 勝

指導教員
福田 光浩 准教授
山下 真 准教授

序

数理計画問題の中で, 凸計画問題は凸理論により強力な解析が可能なだけでなく, 重要な応用を数多くもち, 活発に研究がおこなわれている. 本論文では, この凸計画問題を解くための効率的な劣勾配アルゴリズムについて議論する.

Q を有限次元実ヒルベルト空間 E 上の閉凸集合, $f(x)$ を Q 上の閉凸関数として, 以下の凸計画問題を対象とする.

$$(P) \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in Q \end{array}$$

問題 (P) の最適解が存在すると仮定し, それを $x^* \in Q$ とおく. この問題に対して, 点列 $\{x_k\} \subset Q$ を逐次計算し $f(x_k)$ が最適値 $f(x^*)$ へなるべく速く収束するような反復的アルゴリズムについて考える. ニュートン法やそれにもとづく内点法といったアルゴリズムは, 最適値へ効率的な速さで収束する反復的アルゴリズムとして広く普及している. しかしこれらのアルゴリズムは各反復で多くの計算量を必要とするため, 大規模な問題に対しての適用が困難である. 本論文では, このような大規模な問題に対して有効な手段となりうる, 劣勾配アルゴリズムについて考察する. このアルゴリズムは, 最適値への収束の速さは内点法などに劣るものの, 一反復を少ない計算量で実現する. われわれは, 既存の劣勾配アルゴリズムである mirror-descent 法 [4] および Nesterov による dual-averaging 法などのアルゴリズム [5, 8, 9] を統一的に議論できるような枠組みを提案するとともに, より効率的なアルゴリズムの導出および計算量の比較をおこなう.

既存のアルゴリズム

劣勾配アルゴリズムは, 反復的に計算される $x_k \in Q$ に対して, 目的関数 $f(x)$ の劣勾配 $g_k \in \partial f(x_k) = \{g \in E \mid \forall x \in E,\ f(x) \geq f(x_k) + \langle g, x - x_k \rangle\}$ を用いた補助最適化問題を構成し, その最適解を利用して次の点 $x_{k+1} \in Q$ を更新する. 本論文で紹介する既存アルゴリズムでは, 補助問題を構成するために次の条件を満たす関数 $d(x)$ の存在を仮定する.

- $d(x)$ は Q 上で連続的微分可能な強凸関数である.
- $s \in E$ に対して, (補助) 最適化問題 $\min_{x \in Q}\{\langle s, x \rangle + d(x)\}$ は (簡単に) 解くことができる.

関数 $d(x)$ は, 補助問題の計算量が大きくならないように選択できることが望ましい. このアルゴリズムの研究は, 目的関数 $f(x)$ に微分可能性を仮定する場合としない場合によって解析方法や得られる結論が異なる.

微分可能性を仮定しない一般の目的関数に対しては, Nemirovski と Yudin [4] による Mirror-Descent (MD) アルゴリズムが知られている. このアルゴリズムは, パラメータ $\lambda_k > 0$ (ステップサイズ) を用いて次のように更新をおこなう.

$$\mathrm{MD}\ :\ x_{k+1} = \operatorname*{argmin}_{x\in Q}\left\{\lambda_k[f(x_k)+\langle g_k, x-x_k\rangle] + d(x) - d(x_k) - \langle \nabla d(x_k), x-x_k\rangle\right\}$$

MD アルゴリズムにおいて最適なステップサイズの選び方をすれば精度 ε の近似解を $O((G+A)/\varepsilon^2)$ の計算量で得ることができる. ただし, G, A はそれぞれ劣勾配の評価および補助問題を解くのに必要な計算量の上界である. しかし, このステップサイズの取り方は現実的ではない. Nesterov [9] は ステップサイズ λ_k のほかにスケーリングパラメータ β_k を導入したモデルのもとで Dual-Averaging (DA) アルゴリズムを提案した. このアルゴリズムは次のように点列 $\{x_k\}$ の更新をおこなう.

$$\mathrm{DA}\ :\ x_{k+1} = \operatorname*{argmin}_{x\in Q}\left\{\sum_{i=0}^{k}\lambda_i[f(x_i)+\langle g_i, x-x_i\rangle] + \beta_k d(x)\right\}$$

問題 (P) に対する近似解を $\widehat{x}_k = \sum_{i=0}^{k}\lambda_i x_i/\sum_{i=0}^{k}\lambda_i$ によって生成するとき, DA アルゴリズムは問題 (P) に依存しないパラメータの選び方によっても $f(\widehat{x}_k) - f(x^*) \le O(1/\sqrt{k})$ を保証できる. したがって, DA は計算量の上界 $O((G+A)/\varepsilon^2)$ を現実的に実現する. 一方 MD が現実的に保証できる効率は $f(\widehat{x}_k) - f(x^*) \le O(\log k/\sqrt{k})$ である.

目的関数 $f(x)$ が連続的微分可能である場合にはより効率的なアルゴリズムを構築することが可能である. Nesterov [5] は生成される近似解 $\{\widehat{x}_k\}$ に対して $f(\widehat{x}_k) - f(x^*) \le O(1/k^2)$ の収束率をもつアルゴリズムを初めて提案した. この収束率は精度 ε の近似解が $O((G+A)/\sqrt{\varepsilon})$ の計算量で得られることを保証する. さらに Nesterov は目的関数が微分可能ではなくても, 問題 (P) が extremal convex problem というクラスに属する場合には, このアルゴリズムが $O(1/k^2)$ の収束率をもつことを示した. Nesterov はこのアルゴリズムを [8] において, E 上の一般のノルムに対するアルゴリズムへと発展させている.

劣勾配アルゴリズムの統一的な議論

MD ではステップサイズ $\{\lambda_k\}$ が, DA ではさらにスケーリングパラメータ $\{\beta_k\}$ が用いられている. 本論文では, MD にもスケーリングパラメータを導入して以下の拡張的な更新を提案する.

$$\text{拡張 MD}\ :\ x_{k+1} = \operatorname*{argmin}_{x\in Q}\left\{\lambda_k[f(x_k)+\langle g_k, x-x_k\rangle] + \beta_k d(x) - \beta_{k-1}d(x_k) - \langle \beta_{k-1}\nabla d(x_k), x-x_k\rangle\right\}$$

$\beta_k \equiv 1$ とすればもともとの MD と一致する. この拡張された MD は DA とともにある共通の性質を満たすことを示す. われわれはこの事実にもとづき, MD や DA を統一的に議論できるような枠組みを提案し, その中でアルゴリズムの解析をおこなう. その結果, Nesterov によるものより簡単な方法で DA の解析が可能となり, さらに拡張された MD は DA よりも良い収束率の見積もりをもつことが証明できる.

劣勾配アルゴリズムは目的関数の微分可能性がある場合とない場合によって解析手段が大きく異なるために, これらを同時に対象とするアルゴリズムの体系は少ない. MD や DA はもともと微分可能性を仮定しない目的関数に対してのアルゴリズムであるが, 本論文ではわれわれの枠組みにもとづき, 微分可能な目的

関数に対しても効率的なアルゴリズムを提案する．この結果, MD や DA はこのような凸計画問題に対しても (さらには extremal convex problem に対しても), [8] などの既存アルゴリズムと同じ収束率をもつことが示される．[8] におけるアルゴリズムは一反復に解くべき補助最適化問題が 2 回であるのに対して, われわれが提案する MD や DA にもとづくアルゴリズムではその回数が 1 回である．

われわれの提案するアルゴリズムと既存アルゴリズムとの位置づけは, 下の表のようにまとめることができる．

目的関数	既存アルゴリズム	$f(\widehat{x}_k)-f(x^*)$	計算量	提案アルゴリズム	$f(\widehat{x}_k)-f(x^*)$	計算量
non-smooth	MD	$O\left(\frac{\log k}{\sqrt{k}}\sqrt{\xi(x^*,x_0)}\right)$	$O\left(\frac{G_f+G_d+A}{\varepsilon^2}\right)$	MD	$\boldsymbol{O\left(\sqrt{\frac{d_k(x^*)}{k}}\right)}$	$O\left(\frac{G_f+G_d+A}{\varepsilon^2}\right)$
	DA	$O\left(\sqrt{\frac{d(x^*)}{k}}\right)$	$O\left(\frac{G_f+A}{\varepsilon^2}\right)$	DA	$O\left(\sqrt{\frac{d(x^*)}{k}}\right)$	$O\left(\frac{G_f+A}{\varepsilon^2}\right)$
smooth	Nesterov [8]	$O\left(\frac{d(x^*)}{k^2}\right)$	$O\left(\frac{G_f+G_d+2A}{\sqrt{\varepsilon}}\right)$	MD	$\boldsymbol{O\left(\frac{d_k(x^*)}{k^2}\right)}$	$\boldsymbol{O\left(\frac{G_f+G_d+A}{\sqrt{\varepsilon}}\right)}$
				DA	$O\left(\frac{d(x^*)}{k^2}\right)$	$\boldsymbol{O\left(\frac{G_f+A}{\sqrt{\varepsilon}}\right)}$
extremal convex problem	Nesterov [5]	$O\left(\frac{\|x_0-x^*\|_2^2}{k^2}\right)$	$O\left(\frac{H+A}{\sqrt{\varepsilon}}\right)$	MD	$\boldsymbol{O\left(\frac{d_k(x^*)}{k^2}\right)}$	$O\left(\frac{H+G_d+A}{\sqrt{\varepsilon}}\right)$

“ $f(\widehat{x}_k)-f(x^*)$ ”：アルゴリズムが生成する近似解 $\{\widehat{x}_k\}$ に対する $f(\widehat{x}_k)-f(x^*)$ の上界.

“計算量”：精度 ε の近似解を得るのに十分な計算量.

G_f：目的関数 $f(x)$ の (劣) 勾配の評価に必要な計算量の上界.

G_d：$\nabla d(x)$ の評価に必要な計算量の上界.

A：各反復で構成する補助問題を解くのに必要な計算量の上界.

H：関数値 $(f_1(x),\ldots,f_m(x))$ および勾配 $(\nabla f_1(x),\ldots,\nabla f_m(x))$ の評価に必要な計算量の上界.

$\xi(z,x)=d(x)-d(z)-\langle\nabla d(z),x-z\rangle$.

$d_k(x)=d(z_k)+\langle\nabla d(z_k),x-z_k\rangle \quad (\le d(x)), \quad z_k$：各反復で解く補助問題の解.

目次

第 1 章

凸解析の基礎

まずはじめに, われわれが対象とする凸計画問題の議論をおこなう際に用いられる凸解析の基礎を述べる. これらの内容は主に Nesterov [6] および Rockafellar [10] にもとづく.

E を有限次元実ヒルベルト空間とする. われわれは E 上の内積を $\langle \cdot, \cdot \rangle$ で表す.
$\| \cdot \|$ を E 上のノルムとする. とくに, E の内積が定めるノルムは $\| \cdot \|_2$ と表す ($\|x\|_2 = \langle x, x \rangle^{1/2}$).
ノルム $\| \cdot \|$ に対する双対ノルム $\| \cdot \|_* : E \to \mathbb{R}$ を以下のように定義する.

$$\|x\|_* = \sup_{\|y\| \leq 1} \langle x, y \rangle$$

1.1 凸関数

関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ に対して $\mathrm{dom}\, f = \{x \in E \ : \ f(x) < +\infty\}$ と定める. ここで $\mathrm{dom}\, f \neq \phi$ であり, かつ次の条件が成り立つとき, $f(x)$ は凸関数であるという.

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \forall x, y \in \mathrm{dom}\, f,\ \alpha \in [0,1]$$

$f(x)$ が凸関数であるとき, その定義により $\mathrm{dom}\, f$ は凸集合になる.

凸関数の基本的かつ重要な性質として Jensen の不等式がある.

定理 1.1.1 ([6], Lemma 3.1.1)**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ が凸関数であるための必要十分条件は,

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \leq \sum_{i=1}^{m} \alpha_i f(x_i), \quad \forall m \in \mathbb{N}, \quad \forall x_i \in \mathrm{dom}\, f, \quad \alpha_i \geq 0, \quad \sum_{i=1}^{m} \alpha_i = 1$$

が成り立つことである.

以下にもうひとつの凸関数の特徴付けを述べよう. 関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ に対して, f の *epigraph* を

$$\mathrm{epi}\, f = \{(x, \mu) \in \mathrm{dom}\, f \times \mathbb{R} \mid f(x) \leq \mu\}$$

と定義する. この epigraph に関して次の定理が成り立つ.

定理 1.1.2 ([6], Theorem 3.1.2)**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ が凸関数であることと, $\mathrm{epi}\, f$ が凸集合であることは同値である.

1.1.1　閉凸関数

関数の epigraph について次の性質が成り立つ.

定理 1.1.3 ([10], Theorem 7.1)**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ に関して以下は同値である.
(1) $\mathrm{epi}\, f$ は閉集合である.
(2) 任意の $\alpha \in \mathbb{R}$ に対して $\{x \in E \mid f(x) \leq \alpha\}$ は閉集合である.
(3) $f(x)$ は E 上で下半連続である. すなわち, 任意の $x \in E$ に対して以下が成り立つ.

$$x_i \to x \;\Rightarrow\; f(x) \leq \liminf_{i \to \infty} f(x_i)$$

この事実に関して, われわれは閉凸関数の定義を与える.

定義 1.1.4 ([6])**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸関数とする. $\mathrm{epi}\, f$ が閉集合であるとき, $f(x)$ は閉凸関数であるという.

閉凸関数の演算に関しての性質を述べよう.

命題 1.1.5 ([6], Theorem 3.1.5)**.** $f_1, f_2 : E \to \mathbb{R} \cup \{+\infty\}$ を閉凸関数とする.

(1) 任意の $\alpha_1, \alpha_2 > 0$ に対して $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ とおくと, $f(x)$ は閉凸関数であり, $\mathrm{dom}\, f = \mathrm{dom}\, f_1 \cap \mathrm{dom}\, f_2$ を満たす.
(2) $f(x) = \max\{f_1(x), f_2(x)\}$ とおくと, $f(x)$ は閉凸関数であり, $\mathrm{dom}\, f = \mathrm{dom}\, f_1 \cap \mathrm{dom}\, f_2$ を満たす.

命題 1.1.6 ([6], Theorem 3.1.7)**.** 集合 Δ の各元 $y \in \Delta$ に対して関数 $\phi(y, \cdot) : E \to \mathbb{R} \cup \{+\infty\}$ がそれぞれ閉凸関数であるとする. このとき $f(x) = \sup_{y \in \Delta} \phi(y, x)$ とおくと, $f(x)$ は閉凸関数であり,

$$\mathrm{dom}\, f = \left\{ x \in \bigcap_{y \in \Delta} \mathrm{dom}\, \phi(y, \cdot) \;\middle|\; \exists \gamma > 0 \text{ s.t. } \forall y \in \Delta,\; \phi(y, x) \leq \gamma \right\}$$

が成り立つ.

1.2　方向微分と劣微分

1.2.1　方向微分

凸関数の性質を調べるために, 方向微分の概念を用意する.

定義 1.2.1 ([10])**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ とする. $x \in \mathrm{dom}\, f,\ y \in E$ に対して, 極限

$$\lim_{\alpha \downarrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha}$$

が存在するとき, これを点 x における f の方向 y への方向微分であるといい, $f'(x; y)$ と表す.

f が x において微分可能であるときは, 上記の両側極限が存在し, $f'(x;y) = \langle \nabla f(x), y \rangle$ が成り立つ. 次に凸関数に対する方向微分の性質を述べよう.

定理 1.2.2 ([10], Theorem 23.1)**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸関数, $x \in \operatorname{dom} f$ とする.
(1) 任意の $y \in E$ に対して, $f'(x;y)$ の定義における極限が $[-\infty, +\infty]$ 上に存在して, 以下が成り立つ.

$$f'(x;y) = \inf_{\alpha>0} \frac{f(x+\alpha y) - f(x)}{\alpha}$$

(2) 任意の $y \in E$, $\alpha > 0$ に対して $f'(x;\alpha y) = \alpha f'(x;y)$ が成り立つ.
(3) 任意の $y \in E$ に対して $f'(x;y) \geq -f'(x;-y)$ が成り立つ.

定理 1.2.3. $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸関数, $x \in \operatorname{dom} f$ とする. このとき,

$$f(y) \geq f(x) + f'(x; y-x), \quad \forall y \in E$$

が成り立つ.

証明. 定理 1.2.2 より,

$$f'(x; y-x) = \inf_{\alpha>0} \frac{f(x+\alpha(y-x)) - f(x)}{\alpha} \overset{\alpha=1}{\leq} f(y) - f(x)$$

□

1.2.2 劣微分

凸関数に対する劣微分の概念を述べよう.

定義 1.2.4 ([6])**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸関数とする. $x \in \operatorname{dom} f$ に対して, $g \in E$ が以下の条件を満たすとき, g を f の点 x における劣勾配という.

$$f(y) \geq f(x) + \langle g, y-x \rangle, \quad \forall y \in \operatorname{dom} f$$

f の点 $x \in \operatorname{dom} f$ における劣勾配全体の集合を, f の点 x における劣微分といい, $\partial f(x)$ と表す.

E 上の内積の連続性と線形性により, 劣微分は常に閉凸集合である. 方向微分と劣微分の間には次のような関係がある.

定理 1.2.5 ([10], Theorem 23.2)**.** f を凸関数, $x \in \operatorname{dom} f$ とする. このとき $g \in E$ に対して, $g \in \partial f(x)$ であるための必要十分条件は,

$$f'(x;y) \geq \langle g, y \rangle, \quad \forall y \in E$$

が成り立つことである.

定理 1.2.6 ([10], Theorem 23.4)**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸関数とする. このとき, 任意の $x \in \operatorname{ri}(\operatorname{dom} f)$ に対して $\partial f(x) \neq \phi$ であり, さらに以下が成り立つ.

$$f'(x;y) = \sup\{\langle g, y \rangle \mid g \in \partial f(x)\}, \quad y \in E$$

とくに, $x \in \operatorname{int}(\operatorname{dom} f)$ のとき, $\partial f(x)$ は有界閉凸集合となり, このとき任意の $y \in E$ に対して $f'(x;y)$ は有限である.

1.3　リプシッツ連続性

リプシッツ連続性は本論文の解析に有益な概念である.

定義 1.3.1 ([6])**.** $f : E \to \mathbb{R} \cup \{+\infty\}$, $Q \subset \operatorname{dom} f$, $L \geq 0$ とする. 以下の条件が成り立つとき, f は Q 上で定数 L に対してリプシッツ連続であるという.

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in Q$$

このとき, L を f の Q 上のリプシッツ定数と呼ぶ.

ここでは, リプシッツ連続な凸関数の性質について調べよう.

定理 1.3.2. 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$, 開集合 $Q \subset \operatorname{dom} f$ および $L \geq 0$ に対して以下は同値である.
(i)　f は Q 上で定数 L に対してリプシッツ連続である.
(ii)　任意の $x \in Q$, $g \in \partial f(x)$ に対して, $\|g\|_* \leq L$ が成り立つ.

証明. (i) $\Rightarrow$ (ii). $x \in Q$, $g \in \partial f(x)$ とする. $\|g\|_* = \sup\{\langle g, y\rangle : \|y\| \leq 1\}$ より, われわれの目標は次の不等式を示すことである.

$$\langle g, y\rangle \leq L, \quad \forall y \in E,\ \|y\| \leq 1$$

任意の $y \in E$, $\|y\| \leq 1$ をとる. $x \in Q$ について Q は開集合であるから, $x + \alpha y \in Q$ となる $\alpha > 0$ が存在する. このとき, $g \in \partial f(x)$ に注意すれば,

$$\langle g, y\rangle = \frac{1}{\alpha}\langle g, (x + \alpha y) - x\rangle \leq \frac{f(x + \alpha y) - f(x)}{\alpha} \leq \frac{L\|(x + \alpha y) - x\|}{\alpha} = L\|y\| \leq L$$

となる.
(ii) $\Rightarrow$ (i). 任意の $x, y \in Q \subset \operatorname{dom} f$ に対して, Q は開集合なので $x \in \operatorname{int}(\operatorname{dom} f)$. したがって定理 1.2.6 より $\partial f(x) \neq \phi$ なので, $g \in \partial f(x)$ をとれば,

$$f(x) - f(y) \leq \langle g, x - y\rangle \leq \|g\|_*\|x - y\| \leq L\|x - y\|$$

である. 同様にして $f(y) - f(x) \leq L\|x - y\|$ も示される. □

1.4　微分可能性と勾配のリプシッツ連続性

凸関数の微分可能性について述べる.

定理 1.4.1 ([10], Theorem 25.1)**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸関数, $x \in \operatorname{dom} f$ とする. このとき, f が x において微分可能であることと, $\partial f(x)$ が一点集合であることは同値である. さらに, このとき $x \in \operatorname{int}(\operatorname{dom} f)$ かつ $\partial f(x) = \{\nabla f(x)\}$ が成り立つ, とくに次の不等式が成り立つ.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle, \quad \forall y \in \operatorname{dom} f$$

以下では, リプシッツ連続な勾配をもつ微分可能な関数の性質について調べる. 次の定理は本論文の解析において重要な役割を果たす.

定理 1.4.2. 関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸集合 $Q \subset \mathrm{dom}\, f$ 上で連続的微分可能とする. さらに, ∇f は Q 上で定数 L に対してリプシッツ連続であるとする, すなわち,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in Q$$

が成り立つと仮定する. このとき, 任意の $x, y \in Q$ に対して,

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

が成り立つ.

証明. $x, y \in Q$ をとり $\phi(\tau) = f(x + \tau(y - x))$, $\tau \in [0, 1]$ とおく. Q の凸性より, $x + \tau(y - x) \in Q$, $\forall \tau \in [0, 1]$ なので, $\phi(\tau)$ は $[0, 1]$ 上連続的微分可能であり, $\phi'(\tau) = \langle \nabla f(x + \tau(y - x)), y - x \rangle$ が成り立つ. したがって,

$$\int_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle \, d\tau = \int_0^1 \phi'(\tau) d\tau = \phi(1) - \phi(0) = f(y) - f(x)$$

である. ゆえに,

$$\begin{aligned}
\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| &= \left| \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle \, d\tau - \langle \nabla f(x), y - x \rangle \right| \\
&= \left| \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle \, d\tau \right| \\
&\leq \int_0^1 \left| \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle \right| d\tau \\
&\leq \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_* \times \|y - x\| d\tau \\
&\leq \int_0^1 L\|\tau(y - x)\| \times \|y - x\| d\tau \\
&= \frac{L}{2}\|y - x\|^2
\end{aligned}$$

□

1.5 強凸関数

本論文において重要な概念である強凸関数について述べよう.

定義 1.5.1. 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$, 凸集合 $Q \subset \mathrm{dom}\, f$ および $\sigma \geq 0$ に対して, 以下の条件が成り立つとき, f は Q 上で定数 σ に対して強凸であるという.

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \alpha(1 - \alpha)\frac{\sigma}{2}\|x - y\|^2, \quad \forall x, y \in Q, \ \alpha \in [0, 1] \tag{1.5.1}$$

このとき, σ を f の Q 上の強凸定数と呼ぶ.

定理 1.5.2. 凸関数 $f, g : E \to \mathbb{R} \cup \{+\infty\}$ は凸集合 $Q \subset \operatorname{dom} f \cap \operatorname{dom} g$ 上でそれぞれ定数 $\sigma, \tau \geq 0$ に対して強凸であるとする.
(1) $\beta > 0$ に対して, $h(x) = \beta f(x)$ とおくと, h は Q 上で定数 $\beta\sigma$ に対して強凸である.
(2) $h(x) = f(x) + g(x)$ とおくと, h は Q 上で定数 $\sigma + \tau$ に対して強凸である.

証明. (1) f の強凸性の定義 (1.5.1) の両辺に β をかければわかる.
(2) f および g の強凸性の定義 (1.5.1) を足し合わせればわかる. □

定理 1.5.3. 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$, 凸集合 $Q \subset \operatorname{dom} f$ および $\sigma \geq 0$ に対して以下は同値である.
(1) f は Q 上で定数 σ に対して強凸である.
(2) 任意の $x, y \in Q$ に対して次の不等式が成り立つ.

$$f(y) \geq f(x) + f'(x; y - x) + \frac{\sigma}{2}\|y - x\|^2$$

証明. (1) $\Rightarrow$ (2). $x, y \in Q$ とする. $\alpha \in (0, 1)$ に対して, 強凸性の定義 (1.5.1) は以下のように変形できる.

$$\begin{aligned} f(y) &\geq \frac{f(x + (1-\alpha)(y-x)) - \alpha f(x)}{1-\alpha} + \alpha\frac{\sigma}{2}\|x - y\|^2 \\ &= f(x) + \frac{f(x + (1-\alpha)(y-x)) - f(x)}{1-\alpha} + \alpha\frac{\sigma}{2}\|x - y\|^2 \end{aligned}$$

したがって, $\alpha \uparrow 1$ とすれば, (2) の不等式が得られる.
(2) $\Rightarrow$ (1). $x, y \in Q$, $\alpha \in (0, 1)$ をとる. $z = \alpha x + (1-\alpha)y \in Q$ とおくと, $x - z = (1-\alpha)(x - y)$, $y - z = \alpha(y - x)$ であるから,

$$\begin{aligned} f(x) &\overset{(2)}{\geq} f(z) + f'(z; x - z) + \frac{\sigma}{2}\|x - z\|^2 \\ &= f(z) + (1-\alpha)f'(z; x - y) + (1-\alpha)^2\frac{\sigma}{2}\|x - y\|^2 \\ &\geq f(z) - (1-\alpha)f'(z; y - x) + (1-\alpha)^2\frac{\sigma}{2}\|x - y\|^2 \quad (\because \text{ 定理 1.2.2 より}) \end{aligned}$$

および

$$\begin{aligned} f(y) &\overset{(2)}{\geq} f(z) + f'(z; y - z) + \frac{\sigma}{2}\|y - z\|^2 \\ &= f(z) + \alpha f'(z; y - x) + \alpha^2\frac{\sigma}{2}\|y - x\|^2 \end{aligned}$$

が得られる. したがって,

$$\alpha f(x) + (1-\alpha)f(y) \geq f(z) + \left[\alpha(1-\alpha)^2 + (1-\alpha)\alpha^2\right]\frac{\sigma}{2}\|x - y\|^2 = f(z) + \alpha(1-\alpha)\frac{\sigma}{2}\|x - y\|^2$$

を得る. これは, 強凸性の定義における不等式 (1.5.1) である. □

系 1.5.4. 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ は凸集合 $Q \subset \operatorname{dom} f$ 上で定数 $\sigma \geq 0$ に対して強凸であるとする. このとき, 以下が成り立つ.

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\sigma}{2}\|y - x\|^2, \quad x, y \in Q,\ g \in \partial f(x)$$

証明. 定理 1.5.3 と定理 1.2.5 より,

$$f(y) \geq f(x) + f'(x; y-x) + \frac{\sigma}{2}\|y-x\|^2 \geq f(x) + \langle g, y-x \rangle + \frac{\sigma}{2}\|y-x\|^2$$

□

系 1.5.5. 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ と凸集合 $Q \subset \mathrm{dom}\, f$ について, f は Q 上で定数 $\sigma \geq 0$ に対して強凸かつ微分可能とする. このとき,

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\sigma}{2}\|y-x\|^2, \quad x, y \in Q$$

が成り立つ.

証明. 系 1.5.4 および定理 1.4.1 より. □

1.5.1 Bregman distance

$f : E \to \mathbb{R} \cup \{+\infty\}$ を閉凸集合 $Q \subset \mathrm{dom}\, f$ 上で定数 $\sigma > 0$ に対して強凸かつ連続的微分可能な凸関数とする. 次のような *Bregman distance* [3] を考察しよう.

$$B_f(z, x) = f(x) - f(z) - \langle \nabla f(z), x - z \rangle, \quad x, z \in Q$$

系 1.5.5 より,

$$B_f(z, x) \geq \frac{\sigma}{2}\|x - z\|^2, \quad \forall x, z \in Q$$

が成り立つ. したがって, $x, z \in Q$ に対して $x = z \iff B_f(z, x) = 0$ である.

例えば, $f(x) = \frac{1}{2}\|x\|_2^2 = \frac{1}{2}\langle x, x \rangle$ のときは $\nabla f(x) = x$ であるから,

$$B_f(z, x) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|z\|_2^2 - \langle z, x - z \rangle = \frac{1}{2}\|x - z\|_2^2$$

Bregman distance の利用は, われわれが提案するアルゴリズムを構築するのに役立つ.

第 2 章

凸計画問題

この章では, 本論文が対象とする凸計画問題に関する理論を述べる.

2.1　凸計画問題

われわれが対象とする凸計画問題とは, 凸集合上での凸関数の最小値や最小解を求める最適化問題である. 形式的には, 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ と凸集合 $Q \subset \operatorname{dom} f$ に対して, 以下のような問題 (P) として表される.

$$\begin{array}{lll} (P) & \min & f(x) \\ & \text{s.t.} & x \in Q \end{array}$$

問題 (P) の最適解全体の集合をしばしば $\operatorname*{Argmin}_{x \in Q} f(x)$ とかく. すなわち,

$$\operatorname*{Argmin}_{x \in Q} f(x) = \left\{ x^* \in Q \;\middle|\; f(x^*) = \min_{x \in Q} f(x) \right\}$$

とくに問題 (P) の最適解が一意に存在するとき, それを $\operatorname*{argmin}_{x \in Q} f(x)$ とあらわす.

凸計画問題 (P) の最適値 $f^* = \min_{x \in Q} f(x)$ を正確に求めることは困難である. そこで, $f(x_k)$ が f^* に収束するような点列 $\{x_k\} \subset Q$ を求めることを考える. x_k を計算する反復的な手続きが与えられれば, 十分大きな k に対して得られた x_k を問題 (P) に対する近似解とみなすことができる. このような手続きは反復的アルゴリズムと呼ばれる. われわれの主たる目標は, 凸計画問題 (P) に対する効率的な反復的アルゴリズムを求めることである.

反復的アルゴリズムの効率の良さは, 与えられた許容誤差 $\varepsilon > 0$ に対して

$$f(x_k) - f^* \le \varepsilon$$

となる $x_k \in Q$ を得るために十分な計算量を判断基準に用いる.

2.1.1　最適性条件

$Q = E$ である凸計画問題には以下の最適性条件がある.

定理 2.1.1 ([6], Theorem 3.1.15)**.** 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ および $x^* \in \operatorname{dom} f$ に対して, $f(x^*) = \min_{x \in \operatorname{dom} f} f(x)$ であるための必要十分条件は $0 \in \partial f(x^*)$ が成り立つことである.

一般の実行可能領域 Q に対しては, 以下の最適性条件がある ([2], Propositon 2.1.1 および 2.1.2 を参照).

定理 2.1.2 ([2])**.** $f : E \to \mathbb{R} \cup \{+\infty\}$ を凸関数, $Q \subset \operatorname{dom} f$ を凸集合とする. $x^* \in Q$ に対して以下は同値である.
(i) x^* は最適化問題 $\min_{x \in Q} f(x)$ の最適解である.
(ii) 任意の $x \in Q$ に対して $f'(x^*; x - x^*) \geq 0$ が成り立つ.

この定理から, Q 上で微分可能な目的関数に対する最適性条件が得られる.

系 2.1.3. 凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ と凸集合 $Q \subset \operatorname{dom} f$ に対して, f は Q 上で連続的微分可能とする. このとき, $x^* \in Q$ に対して, 以下は同値である.
(i) x^* は最適化問題 $\min_{x \in Q} f(x)$ に対する最適解である.
(ii) $\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in Q$
とくに $x^* \in \operatorname{int} Q$ のとき, (ii) は $\nabla f(x^*) = 0$ であることと同値である.

2.1.2　強凸関数の最小化問題

目的関数が強凸である凸計画問題の重要な性質を示そう. われわれは次の補題を用いる ([11], 定理 1.3.1).

補題 2.1.4. $f : E \to \mathbb{R} \cup \{+\infty\}$ を E 上の下半連続な関数, $Q \subset \operatorname{dom} f$ を有界閉集合とする. このとき, f の Q 上の最小点が存在する.

この補題を用いて, 強凸関数の最小化問題についての性質を示そう.

定理 2.1.5. 閉凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ と空でない閉凸集合 $Q \subset \operatorname{dom} f$ に対して, $f(x)$ は Q 上で定数 $\sigma > 0$ に対して強凸であるとする. このとき, 最適化問題 $\min_{x \in Q} f(x)$ は唯一の最適解 $x^* \in Q$ をもつ. さらに, 以下が成り立つ.

$$f(x) \geq f(x^*) + \frac{\sigma}{2} \|x - x^*\|^2, \quad x \in Q$$

証明. 一般性を失うことなく $\operatorname{dom} f = Q$ としてよい[1]. まず, 最適解の存在を示す. 凸集合 Q は空でないから $\bar{x} \in \operatorname{ri} Q$ がとれる ([10], Theorem 6.2 を見よ). この $\bar{x}$ に対して凸集合 $S = \{x \in Q \mid f(x) \leq f(\bar{x})\} \ni \bar{x}$ を考える. このとき, 最適化問題 $\min_{x \in Q} f(x)$ は $\min_{x \in S} f(x)$ に等しい. f が閉凸関数であることから, 定理 1.1.3 より S は閉集合である. つぎに, S が有界であることを示そう. $\bar{x} \in \operatorname{ri} Q = \operatorname{ri}(\operatorname{dom} f)$ であるから,

[1] $\phi(x) = \begin{cases} f(x) & : \quad x \in Q \\ +\infty & : \quad x \notin Q \end{cases}$ は $\operatorname{dom} \phi = Q$ を満たす強凸な閉凸関数であり, この $\phi(x)$ に対して定理の主張を示せばよい.

$\bar{g} \in \partial f(\bar{x})$ が存在する (定理 1.2.6). このとき, 系 1.5.4 より,

$$\begin{aligned} x \in S \Longrightarrow f(\bar{x}) &\geq f(x) \geq f(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle + \frac{\sigma}{2}\|x - \bar{x}\|^2 \\ \Longrightarrow & \frac{\sigma}{2}\|x - \bar{x}\|^2 \leq -\langle \bar{g}, x - \bar{x} \rangle \leq \|\bar{g}\|_* \times \|x - \bar{x}\| \\ \Longrightarrow & \|x - \bar{x}\| \leq \frac{2}{\sigma}\|\bar{g}\|_* \\ \Longrightarrow & \|x\| \leq \|\bar{x}\| + \frac{2}{\sigma}\|\bar{g}\|_* \end{aligned}$$

となる. ゆえに, S は有界閉集合である. また, 定理 1.1.3 より $f(x)$ は E 上で下半連続なので S 上に最小点をもつ (補題 2.1.4).

次に, 最適解の一意性を示そう. $x_1,\ x_2 \in Q$ に対して $f(x_1) = f(x_2) = \min_{x \in Q} f(x)$ であるとする. このとき, 凸結合 $(x_1 + x_2)/2 \in Q$ に対して f の強凸性より

$$\min_{x \in Q} f(x) \leq f(\frac{x_1 + x_2}{2}) \leq \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2) - \frac{1}{2}(1 - \frac{1}{2})\frac{\sigma}{2}\|x_1 - x_2\|^2 = \min_{x \in Q} f(x) - \frac{\sigma}{8}\|x_1 - x_2\|^2$$

したがって, $\frac{\sigma}{8}\|x_1 - x_2\|^2 \leq 0$ ゆえ $x_1 = x_2$ を得る. 最後に, $x^* = \operatorname*{argmin}_{x \in Q} f(x)$ に対して, 強凸性 (定理 1.5.3) と最適性 (定理 2.1.2) より,

$$\forall x \in Q, \quad f(x) \geq f(x^*) + f'(x^*; x - x^*) + \frac{\sigma}{2}\|x - x^*\|^2 \geq f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2$$

□

以下の定理は本論文において重要である.

定理 2.1.6. $f, d : E \to \mathbb{R} \cup \{+\infty\}$ を閉凸関数, $Q \subset \operatorname{dom} f \cap \operatorname{dom} d$ を閉凸集合する. さらに $d(x)$ は Q 上で定数 $\sigma > 0$ に対して強凸であるとする. このとき, $\beta > 0$ に対して, $\phi(x) = f(x) + \beta d(x)$ とおくと, 最適化問題 $\min_{x \in Q} \phi(x)$ は唯一の最適解 $x^* \in Q$ をもち, 以下が成り立つ.

$$\phi(x) \geq \phi(x^*) + \beta\xi(x^*, x), \quad \forall x \in Q$$

ただし, $\xi(z, x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$

証明. 命題 1.1.5 と定理 1.5.2 より, $\phi(x)$ は Q 上で定数 $\beta\sigma > 0$ に対して強凸な閉凸関数である. したがって, 定理 2.1.5 より, 一意な最小点 $x^* = \operatorname*{argmin}_{x \in Q} \phi(x)$ が存在する. この最小点に対して, 最適性条件 (定理 2.1.2) より, 任意の $x \in Q$ に対して,

$$\begin{aligned} 0 \leq \phi'(x^*; x - x^*) &= \lim_{\alpha \downarrow 0} \frac{\phi(x^* + \alpha(x - x^*)) - \phi(x^*)}{\alpha} \\ &= \lim_{\alpha \downarrow 0} \left(\frac{f(x^* + \alpha(x - x^*)) - f(x^*)}{\alpha} + \beta \times \frac{d(x^* + \alpha(x - x^*)) - d(x^*)}{\alpha} \right) \\ &= \lim_{\alpha \downarrow 0} \frac{f(x^* + \alpha(x - x^*)) - f(x^*)}{\alpha} + \beta \times \lim_{\alpha \downarrow 0} \frac{d(x^* + \alpha(x - x^*)) - d(x^*)}{\alpha} \\ &= f'(x^*; x - x^*) + \beta \langle \nabla d(x^*), x - x^* \rangle \end{aligned}$$

が成り立つ. ゆえに, 任意の $x \in Q$ に対して,

$$
\begin{aligned}
&f'(x^*; x - x^*) \geq -\beta \langle \nabla d(x^*), x - x^* \rangle \\
\Longrightarrow\ &f'(x^*; x - x^*) + \beta d(x) - \beta d(x^*) \geq \beta [d(x) - d(x^*) - \langle \nabla d(x^*), x - x^* \rangle] = \beta \xi(x^*, x) \\
\Longrightarrow\ &f'(x^*; x - x^*) + \beta d(x) \geq \beta d(x^*) + \beta \xi(x^*, x)
\end{aligned}
$$

が得られる. 以上より, $x \in Q$ に対して以下を得る.

$$
\begin{aligned}
\phi(x) &= f(x) + \beta d(x) \\
&\geq [f(x^*) + f'(x^*; x - x^*)] + \beta d(x) \quad (\because\ \text{定理 1.2.3 より}) \\
&\geq f(x^*) + [\beta d(x^*) + \beta \xi(x^*, x)] \\
&= \phi(x^*) + \beta \xi(x^*, x)
\end{aligned}
$$

□

2.2 Extremal convex problems

この節で扱う extremal convex problem は Nesterov [5, 8] にもとづく. 以下のような最適化問題を *extremal convex problem* という.

$$
\begin{aligned}
(P) \qquad &\min \quad f(x) = \max_{u \in U} \left\{ \sum_{j=1}^{m} u^{(j)} f_j(x) - \phi(u) \right\} \\
&\text{s.t.} \quad x \in Q
\end{aligned}
\tag{2.2.1}
$$

ただし, $\phi(u) : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ は連続凸関数, $U \subset \operatorname{dom} \phi$ は有界閉凸集合, 各 $f_j : E \to \mathbb{R} \cup \{+\infty\}$ は凸関数, $Q \subset \bigcap_{j=1}^{m} \operatorname{dom} f_j$ は閉凸集合であり, 各 $f_j(x)$ は Q 上で連続的微分可能であるとする. また, 目的関数 $f(x)$ の閉凸性を保証するために, 以下も仮定する.

仮定 2.2.1. 問題 (P) に対する集合 U と凸関数 $f_j(x),\ j = 1, \ldots, m$ について,

$$
f_j(x) \text{ が 一次関数でないならば, 任意の } u \in U \text{ に対して } u^{(j)} \geq 0
$$

が成り立つ.

われわれは, 問題 (P) が extremal convex problem であるというときは, 仮定 2.2.1 が成り立つものとする.

命題 2.2.2. Extremal convex problem (P) における目的関数 $f(x)$ は閉凸関数であり, $J = \{j \mid \exists u \in U \text{ s.t. } u^{(j)} \neq 0\}$ に対して,

$$
\operatorname{dom} f = \bigcap_{j \in J} \operatorname{dom} f_j
$$

が成り立つ.

証明. $u \in U$ に対して $g_u(x) = \displaystyle\sum_{j=1}^{m} u^{(j)} f_j(x) - \phi(u)$ とおくと, $f(x) = \displaystyle\max_{u \in U} g_u(x)$ とかける.

任意の $u \in U$ に対して,

$$g_u(x) = \sum_{j:u^{(j)}>0} u^{(j)} f_j(x) + \sum_{j:u^{(j)}<0} u^{(j)} f_j(x) - \phi(u)$$

であるが, $u^{(j)} < 0$ なる j に対しては, 仮定 2.2.1 より $f_j(x)$ は一次関数である. したがって, $g_u(x)$ は x に関する閉凸関数であり,

$$\mathrm{dom}\, g_u = \bigcap_{j:u^{(j)}\neq 0} \mathrm{dom}\, f_j$$

が成り立つ (命題 1.1.5). したがって, 命題 1.1.6 より, $f(x) = \max_{u\in U} g_u(x)$ は閉凸関数であり, U が有界閉集合であることに注意すると,

$$\begin{aligned}
\mathrm{dom}\, f &= \left\{ x \in \bigcap_{u\in U} \bigcap_{j:u^{(j)}\neq 0} \mathrm{dom}\, f_j \ :\ \exists \gamma \ \text{ s.t. } \ \forall u \in U,\ g_u(x) \leq \gamma \right\} \\
&= \bigcap_{u\in U} \bigcap_{j:u^{(j)}\neq 0} \mathrm{dom}\, f_j \\
&= \bigcap_{j\in J} \mathrm{dom}\, f_j
\end{aligned}$$

□

2.2.1　線形化

Extremal convex problem において, 線形化の概念が重要である.

定義 2.2.3. Extremal convex problem (P) について, 目的関数 $f(x)$ の $y \in Q$ における線形化 $f(y, x)$ を,

$$f(y, x) = \max_{u\in U} \left\{ \sum_{j=1}^{m} u^{(j)} [f_j(y) + \langle \nabla f_j(y), x - y \rangle] - \phi(u) \right\}$$

と定める.

命題 2.2.4. Extremal convex problem (P) について以下が成り立つ.

$$f(x) \geq f(y, x), \quad \forall x, y \in Q.$$

証明. $x, y \in Q$ とする. 各 $j = 1, \ldots, m$ に対して, 定理 1.4.1 より,

$$f_j(x) \geq f_j(y) + \langle \nabla f_j(y), x - y \rangle$$

である. とくに, $f_j(x)$ が一次関数のときは等号が成り立つから, 仮定 2.2.1 に注意すれば, 任意の $u \in U$ に対して,

$$\begin{aligned}
&\sum_{j=1}^{m} u^{(j)} f_j(x) \geq \sum_{j=1}^{m} u^{(j)} [f_j(y) + \langle \nabla f_j(y), x - y \rangle] \\
\Rightarrow &\sum_{j=1}^{m} u^{(j)} f_j(x) - \phi(u) \geq \sum_{j=1}^{m} u^{(j)} [f_j(y) + \langle \nabla f_j(y), x - y \rangle] - \phi(u)
\end{aligned}$$

が成り立つ. ゆえに両辺の $u \in U$ に関する最大値をとれば,

$$f(x) \geq f(y, x)$$

が得られる. □

命題 2.2.5. Extremal convex problem (P) に対して $f_j(x),\ j = 1, \ldots, m$ は Q 上で定数 $\sigma_j \geq 0$ に対して強凸であり, $\nabla f_j(x)$ が定数 $L_j \geq 0$ に対して Q 上リプシッツ連続であるとする. このとき $L, \sigma \geq 0$ を,

$$L = \max_{u \in U} \sum_{j=1}^{m} u^{(j)} L_j, \quad \sigma = \min_{u \in U} \sum_{j=1}^{m} u^{(j)} \sigma_j$$

によって定めるとき, 以下が成り立つ.

$$\frac{\sigma}{2}\|x - y\|^2 \leq f(x) - f(y, x) \leq \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in Q$$

証明. $x, y \in Q$ に対して, 仮定より $f_j(x), j = 1, \ldots, m$ は次の不等式を満たす (系 1.5.5 と定理 1.4.2 より).

$$\frac{\sigma_j}{2}\|x - y\|^2 \leq f_j(x) - [f_j(y) + \langle \nabla f_j(y), x - y \rangle] \leq \frac{L_j}{2}\|x - y\|^2$$

とくに $f_j(x)$ が一次関数であるときはこの不等式は等号を満たすので[2], 仮定 2.2.1 に注意すれば, 任意の $u \in U$ に対して,

$$\frac{u^{(j)}\sigma_j}{2}\|x - y\|^2 \leq u^{(j)} f_j(x) - u^{(j)}[f_j(y) + \langle \nabla f_j(y), x - y \rangle] \leq \frac{u^{(j)} L_j}{2}\|x - y\|^2, \quad j = 1, \ldots, m$$

が成り立つ. この不等式を $j = 1, \ldots, m$ について足しあわせ, L, σ の定め方に注意すると, 任意の $u \in U$ に対して以下の不等式を得る.

$$\frac{\sigma}{2}\|x - y\|^2 \leq \sum_{j=1}^{m} u^{(j)} f_j(x) - \sum_{j=1}^{m} u^{(j)}[f_j(y) + \langle \nabla f_j(y), x - y \rangle] \leq \frac{L}{2}\|x - y\|^2$$

したがって, 任意の $u \in U$ に対して,

$$\begin{aligned}
&\sum_{j=1}^{m} u^{(j)}[f_j(y) + \langle \nabla f_j(y), x - y \rangle] - \phi(u) + \frac{\sigma}{2}\|x - y\|^2 \\
&\qquad \leq \sum_{j=1}^{m} u^{(j)} f_j(x) - \phi(u) \\
&\qquad \leq \sum_{j=1}^{m} u^{(j)}[f_j(y) + \langle \nabla f_j(y), x - y \rangle] - \phi(u) + \frac{L}{2}\|x - y\|^2
\end{aligned}$$

が成り立つ. この不等式で各辺の $u \in U$ に関する最大値をとれば, 求める不等式が得られる. □

[2] $f_j(x)$ が一次関数のときは $\sigma_j = 0$ であり, $L_j = 0$ としてよい.

第 3 章

劣勾配アルゴリズム

凸計画問題に対する反復的アルゴリズムの効率を議論する上で重要な要素として次の二点が挙げられる.

- 一反復に要する計算量
- 各反復で生成される近似解における目的関数値の最適値への収束率

ニュートン法やそれにもとづく内点法といったアルゴリズムでは, 最適値への収束が速いが, 一反復に多くの計算量が必要となる. この事実とは対照的に, 以下に述べる劣勾配アルゴリズムは最適値への収束が遅いものの, 一反復を少ない計算量で実行することができる[1] という特徴をもつ. したがって, 対象となる凸計画問題が非常に大規模で内点法などの適用が困難である場合, 劣勾配アルゴリズムは有効な手段となりうる. この章では, 既存の劣勾配アルゴリズムのいくつかを述べる.

3.1　古典的な劣勾配アルゴリズム

この節では, 古典的な劣勾配アルゴリズムである最急降下法と射影劣勾配法について述べる. これらのアルゴリズムは E の内積が定めるノルム $\|\cdot\|_2$ に関して記述, 解析される ($\|x\|_2 = \langle x, x\rangle^{1/2}$).

3.1.1　最急降下法

E 上で連続的微分可能な凸関数 $f : E \to \mathbb{R}$ に対して, 以下の凸計画問題を考える.

$$(P) \quad \min_{x\in E} f(x)$$

この問題 (P) に対して勾配 $\nabla f(x)$ を用いた反復的アルゴリズムを考える. 勾配がもつ性質として, 次の事実が基本的である.

補題 3.1.1 ([6], p. 17)**.** $\bar{x} \in E$ とする. 方向 $-\nabla f(\bar{x})$ は点 $\bar{x}$ において $f(x)$ の局所的な減少量を最大にする. すなわち, $s \in E,\ \|s\|_2 = 1$ に対して, $f(x)$ の $\bar{x}$ における方向 s に対する局所的変化量

$$\Delta(s) = \lim_{\alpha\downarrow 0} \frac{f(\bar{x}+\alpha s) - f(\bar{x})}{\alpha} = \langle \nabla f(\bar{x}), s\rangle$$

[1] 対象とする凸計画問題の複雑さによってはその限りではない.

は, $s = -\nabla f(\bar{x})/\|\nabla f(\bar{x})\|_2$ において最小となる.

この意味で, $-\nabla f(\bar{x})$ は $\bar{x} \in E$ における $f(x)$ の最急降下方向と呼ばれる. この事実から次の素朴なアルゴリズムが考えられる.

アルゴリズム 3.1.2 (最急降下法)**.** $x_0 \in E$ をえらび, $k = 0, 1, 2, \ldots$ に対して以下のようにして点列 $\{x_i\}$ を計算する.

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k), \quad \text{ただし}, \ \lambda_k > 0$$

このアルゴリズムにおけるパラメータ $\lambda_k > 0$ はステップサイズと呼ばれる. ステップサイズの選び方は, アルゴリズムの効率および解析に大きく関わるパラメータである. 最急降下法におけるステップサイズの選択に関しては, たとえば次の事実が知られている.

定理 3.1.3 ([6], p. 70)**.** 問題 (P) に対して, $\nabla f(x)$ が定数 $L > 0$ に対して E 上リプシッツ連続であるとする. このとき, ステップサイズ $\lambda_k \equiv 1/L$ に対する最急降下法は以下を満たす.

$$f(x_k) - f(x^*) \le \frac{2L\|x_0 - x^*\|_2^2}{k+4}$$

ただし, $x^* \in E$ は問題 (P) の最適解である. さらに, $f(x)$ が定数 $\sigma > 0$ に対して E 上強凸であるならば, $\lambda_k \equiv 2/(\sigma + L)$ に対する最急降下法は以下を満たす.

$$f(x_k) - f(x^*) \le \frac{L}{2}\left(\frac{L-\sigma}{L+\sigma}\right)^{2k} \|x_0 - x^*\|_2^2$$

この定理からわかる最急降下法の効率について述べよう. $D = \frac{1}{2}\|x_0 - x^*\|_2^2$ とおく.
(1) ステップサイズ $\lambda_k \equiv 1/L$ による最急降下法が与える目的関数値と最適値との残差は $4LD/(k+4)$ でおさえられる. 反復回数 k に関しては $O(1/k)$ の収束オーダーになっている. ここで, 精度 $\varepsilon > 0$ をもつ近似解 x_k を得るために十分な反復回数について考えよう.

$$\frac{4LD}{k+4} \le \varepsilon \iff k \ge \frac{4LD}{\varepsilon} - 4 \iff k \ge \left\lceil \frac{4LD}{\varepsilon} \right\rceil - 4$$

であるから, 精度 ε をもつ近似解 x_k を得るための反復回数 k の上界として $\lceil 4LD/\varepsilon \rceil - 4 = O(LD/\varepsilon)$ が得られる. また, $\nabla f(x)$ の一回の評価にかかる計算量の上界を G とおけば, 最急降下法における一反復の計算量の上界も G としてよい[2]. ゆえに, 精度 ε の近似解を得るために十分な計算量は以下のように与えられる.

$$G \times \left(\left\lceil \frac{4LD}{\varepsilon} \right\rceil - 4 \right)$$

(2) 目的関数 $f(x)$ が定数 σ に対して強凸であるとき, $r = \left(\dfrac{L-\sigma}{L+\sigma}\right)^2 \in (0,1)$ とおけば, $\lambda_k \equiv 2/(\sigma + L)$ による最急降下法が与える目的関数値と最適値との残差は LDr^k でおさえられる. したがって, 精度 $\varepsilon > 0$ をもつ近似解 x_k を得るための反復回数 k の上界は次のように与えられる.

$$LDr^k \le \varepsilon \iff k \ge \frac{1}{\log(1/r)}\left(\log\frac{1}{\varepsilon} + \log L + \log D\right) \Longleftarrow k \ge \left\lceil \frac{\log(1/\varepsilon) + \log L + \log D}{\log(1/r)} \right\rceil$$

[2] $E = \mathbb{R}^n$ のとき, $\nabla f(x)$ の評価には $\frac{\partial f}{\partial x_i}(x)$, $1 \le i \le n$ の計算が必要であり, 少なくとも $O(n)$ の計算量が必要であると一般的に仮定する.

ゆえに, 精度 ε を得るために十分な計算量として以下を得る.

$$G \times \left\lceil \frac{\log(1/\varepsilon) + \log L + \log D}{\log(1/r)} \right\rceil$$

3.1.2　射影劣勾配法

最急降下法はここに述べる射影劣勾配法として一般化することができる. ここでは, 閉凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ および閉凸集合 $Q \subset \operatorname{dom} f$ に対して以下の凸計画問題を考える.

$$\begin{array}{lll} (P) & \min & f(x) \\ & \text{s.t.} & x \in Q \end{array}$$

この問題に対して, 閉凸集合 Q への直交射影 $\pi_Q : z \mapsto \underset{x \in Q}{\operatorname{argmin}} \|x - z\|_2$ を用いて, 射影劣勾配法は次のように記述される.

アルゴリズム 3.1.4 (射影劣勾配法)**.** $x_0 \in Q$ をえらび, $k = 0, 1, 2, \ldots$ に対して以下のようにして点列 $\{x_i\}$ を計算する.

$$x_{k+1} = \pi_Q\left(x_k - \lambda_k g_k\right), \quad \text{ただし}, \lambda_k > 0,\ g_k \in \partial f(x_k)$$

$Q = E$ かつ目的関数 $f(x)$ が E 上連続的微分可能であるならば, $\partial f(x) = \{\nabla f(x)\}$ (定理 1.4.1) なので, 射影劣勾配法は最急降下法になる.

射影劣勾配法は次節に述べる mirror-descent アルゴリズムへとさらに一般化して議論されるので, ここでは詳しい解析については述べない.

3.2　Mirror-descent アルゴリズム

Mirror-Descent (MD) アルゴリズムは, Nemirovski と Yudin [4] によって提案された次の凸計画問題に対して提案されたアルゴリズムである.

$$\begin{array}{lll} (P) & \min & f(x) \\ & \text{s.t.} & x \in Q \end{array}$$

ただし, $f : E \to \mathbb{R} \cup \{+\infty\}$ は閉凸関数, $Q \subset \operatorname{dom} f$ は閉凸集合である. この節では, Beck と Teboulle [1] による MD の解析について述べる.

MD アルゴリズムは E 上の一般のノルム $\|\cdot\|$ に対して構築される. われわれは次を満たす凸関数 $d : E \to \mathbb{R} \cup \{+\infty\}$ の存在を仮定する.

- $Q \subset \operatorname{dom} d$
- $d(x)$ は Q 上で連続的微分可能かつ定数 $\sigma > 0$ に対して強凸な関数である.
- $s \in E$ に対して, 最適化問題 $\min_{x \in Q} \{\langle s, x \rangle + d(x)\}$ は (簡単に) 解くことができる.

この関数 $d(x)$ に対して以下の Bregman distance をつかう.

$$\xi(z, x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle, \quad x, z \in Q$$

このとき, MD アルゴリズムは次のように記述される.

アルゴリズム 3.2.1 (mirror-descent)**.** 上記の関数 $d(x)$ に対して以下をおこなう.

Step 0. $x_0 \in Q$ をえらぶ.
Step 1. $k = 0, 1, 2, \ldots$ に対して, 以下をおこなう.

$$x_{k+1} = \operatorname*{argmin}_{x \in Q} \left\{ \lambda_k [f(x_k) + \langle g_k, x - x_k \rangle] + \xi(x_k, x) \right\}, \quad \text{ただし}, \lambda_k > 0,\ g_k \in \partial f(x_k)$$

例 3.2.2. MD アルゴリズムが射影劣勾配法を含むことを確認しよう. $\|\cdot\| = \|\cdot\|_2$, $d(x) = \frac{1}{2}\|x\|_2^2$ とすると, 関数 $d(x)$ は定数 $\sigma = 1$ に対しての強凸関数であり, $\xi(z, x) = \frac{1}{2}\|x - z\|_2^2$ が成り立つ. このとき, MD アルゴリズムの反復は,

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in Q} \left\{ \lambda_k [f(x_k) + \langle g_k, x - x_k \rangle] + \frac{1}{2}\|x - x_k\|_2^2 \right\} \\ &= \operatorname*{argmin}_{x \in Q} \left\{ 2\lambda_k f(x_k) + 2\langle \lambda_k g_k, x - x_k \rangle + \|x - x_k\|_2^2 \right\} \quad (\because \text{ 正の定数倍では不変}) \\ &= \operatorname*{argmin}_{x \in Q} \left\{ \|\lambda_k g_k\|_2^2 + 2\langle \lambda_k g_k, x - x_k \rangle + \|x - x_k\|_2^2 \right\} \quad (\because \text{ 定数の差では不変}) \\ &= \operatorname*{argmin}_{x \in Q} \|(x - x_k) + \lambda_k g_k\|_2^2 \\ &= \pi_Q(x_k - \lambda_k g_k) \end{aligned}$$

となる. したがってこれは射影劣勾配法の反復でもある.

MD アルゴリズムの効率は以下のように解析される. 定理 3.2.3 の前半の主張は [1] の解析によるものであり, 後半の主張は Jensen の不等式 (定理 1.1.1) から得られる.

定理 3.2.3. $x^* \in Q$ を問題 (P) に対する最適解とする. Mirror-descent アルゴリズムにより得られる $\{x_k\} \subset Q$ は $k \geq 0$ に対して次の不等式を満たす.

$$\frac{\sum_{i=0}^k \lambda_i [f(x_i) - f(x^*)]}{\sum_{i=0}^k \lambda_i} \leq \frac{\xi(x_0, x^*) + \frac{1}{2\sigma}\sum_{i=0}^k \lambda_i^2 \|g_i\|_*^2}{\sum_{i=0}^k \lambda_i}$$

したがって, $\widehat{x}_k = \sum_{i=0}^k \lambda_i x_i / \sum_{i=0}^k \lambda_i$ とおくと $k \geq 0$ に対して以下が成り立つ.

$$\max\left\{ f(\widehat{x}_k) - f(x^*), \min_{0 \leq i \leq k} f(x_i) - f(x^*) \right\} \leq \frac{\xi(x_0, x^*) + \frac{1}{2\sigma}\sum_{i=0}^k \lambda_i^2 \|g_i\|_*^2}{\sum_{i=0}^k \lambda_i}$$

この定理から得られる目的関数値と最適値との残差に対する上界を

$$\delta_k = \frac{\xi(x_0, x^*) + \frac{1}{2\sigma}\sum_{i=0}^k \lambda_i^2 \|g_i\|_*^2}{\sum_{i=0}^k \lambda_i} \tag{3.2.1}$$

とおこう. $\sup_{k \geq 0} \|g_k\|_* \leq M$ であるときは, 以下のステップサイズ λ_k のとりかたは $\delta_k \to 0$ となるための十分条件である.

$$\lambda_k \to 0, \quad \sum_{k=0}^{\infty} \lambda_k = \infty$$

固定された N に対して, δ_N は $(\lambda_0,\dots,\lambda_N)\in\mathbb{R}^{N+1}$ に関する凸関数になっており, 最適なステップサイズの選び方を解析的に調べることができる. その結果として次の定理に述べるステップサイズが導かれる. 定理の証明自身は, 実際にステップサイズを代入して容易に確認することができる.

定理 3.2.4. 上記の δ_k に対して以下が成り立つ.

(1) 整数 $N\geq 0$ に対して次が成り立つ.

$$\lambda_k=\frac{\sqrt{2\sigma\xi(x_0,x^*)}}{\|g_k\|_*}\frac{1}{\sqrt{N+1}}\quad\Longrightarrow\quad\delta_N\leq\max_{0\leq k\leq N}\|g_k\|_*\sqrt{\frac{2\xi(x_0,x^*)}{\sigma}}\frac{1}{\sqrt{N+1}}$$

(2) $\sup_{k\geq 0}\|g_k\|_*\leq M$ とするとき, 整数 $N\geq 0$ に対して次が成り立つ.

$$\lambda_k=\frac{\sqrt{2\sigma\xi(x_0,x^*)}}{M}\frac{1}{\sqrt{N+1}}\quad\Longrightarrow\quad\delta_N\leq M\sqrt{\frac{2\xi(x_0,x^*)}{\sigma}}\frac{1}{\sqrt{N+1}}$$

上記の定理における δ_N の見積もりは固定された整数 N に対してのみ成り立つので, 非常に限定的である. この定理の (2) が与える δ_N の上界 $\sqrt{\frac{2\xi(x_0,x^*)}{\sigma}}\frac{M}{\sqrt{N+1}}$ に関して考えると, 精度 $\varepsilon>0$ の近似解を得るための反復回数 N の上界は以下のようになる.

$$\sqrt{\frac{2\xi(x_0,x^*)}{\sigma}}\frac{M}{\sqrt{N+1}}\leq\varepsilon\quad\Longleftrightarrow\quad N\geq\frac{2M^2\xi(x_0,x^*)}{\sigma\varepsilon^2}-1\quad\Longleftarrow\quad N\geq\left\lceil\frac{2M^2\xi(x_0,x^*)}{\sigma\varepsilon^2}\right\rceil-1$$

また, 劣勾配 $g_k\in\partial f(x_k)$ を得るための計算量の上界を G_f, 勾配 $\nabla d(x)$ を得るための計算量の上界を G_d とし, さらに補助問題 $\min\limits_{x\in Q}\{\langle s,x\rangle+d(x)\}$ を解くのに要する計算量の上界を A とすれば, 一反復の計算量の上界は G_f+G_d+A である. したがって, 精度 ε の近似解を得るために十分な計算量として次が得られる.

$$(G_f+G_d+A)\times\left(\left\lceil\frac{2M^2\xi(x_0,x^*)}{\sigma\varepsilon^2}\right\rceil-1\right)\tag{3.2.2}$$

この計算量で精度 ε の近似解を得るには, $N=\left\lceil\frac{2M^2\xi(x_0,x^*)}{\sigma\varepsilon^2}\right\rceil-1$ として, 定理 3.2.4 のステップサイズを用いた MD アルゴリズムを N 反復おこなえばよい. ただしこの方法では, あらかじめ N を決めておく必要があることや, N 反復目以外では効率的な δ_k の見積もりが得られないといった制限がある.

δ_k の上界に関して, 任意の整数 k に対する主張としては, 以下の定理のような見積もりが得られる. しかし, オーダーは $O\left(\frac{\log k}{\sqrt{k}}\right)$ にまで増大してしまう.

定理 3.2.5. 上記の δ_k に対して以下が成り立つ.

(1) $$\lambda_k=\frac{\sqrt{2\sigma\xi(x_0,x^*)}}{\|g_k\|_*}\frac{1}{\sqrt{k+1}}\quad\Longrightarrow\quad\forall k\geq 0,\ \ \delta_k\leq\max_{0\leq i\leq k}\|g_i\|_*\sqrt{\frac{2\xi(x_0,x^*)}{\sigma}}\frac{\log(k+1)+2}{\sqrt{k+2}-1}$$

(2) $\sup_{k\geq 0}\|g_k\|_*\leq M$ とするとき, 次が成り立つ.

$$\lambda_k=\frac{\sqrt{2\sigma\xi(x_0,x^*)}}{M}\frac{1}{\sqrt{k+1}}\quad\Longrightarrow\quad\forall k\geq 0,\ \ \delta_k\leq M\sqrt{\frac{2\xi(x_0,x^*)}{\sigma}}\frac{\log(k+1)+2}{\sqrt{k+2}-1}$$

証明. 不等式

$$\sum_{i=0}^{k}\frac{1}{i+1} = 1 + \sum_{i=1}^{k}\frac{1}{i+1} \leq 1 + \sum_{i=1}^{k}\int_{i}^{i+1}\frac{dx}{x} = 1 + \int_{1}^{k+1}\frac{dx}{x} = 1 + \log(k+1)$$

および,

$$\sum_{i=0}^{k}\frac{1}{\sqrt{i+1}} \geq \sum_{i=0}^{k}\int_{i}^{i+1}\frac{dx}{\sqrt{x+1}} = \int_{0}^{k+1}\frac{dx}{\sqrt{x+1}} = 2(\sqrt{k+2}-1)$$

に注意して, ステップサイズ λ_k を δ_k の定義に代入すればよい. □

この定理のステップサイズを用いると, 各反復で残差の上界 δ_k を見積もることができ, あらかじめ実行する反復回数 N を決めておく必要がない. ただし, 上に述べたステップサイズの選び方は x^* が用いられているため現実的ではない。これを回避するには, $\xi(x_0, x^*)$ をその上界値 $\gamma > 0$ でおきかえれば良いが, この γ があらかじめ分かるかどうかは保証できない.

3.3　Dual-averaging アルゴリズム

この節では, Nesterov [9] によって提案 Dual-Averaging (DA) アルゴリズムについて述べる. これは mirror-descent アルゴリズムよりも解析的によい結果を導くことができるアルゴリズムである. われわれは前節と同じ以下の凸計画問題を対象にする.

$$(P) \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in Q \end{array}$$

ただし, $f : E \to \mathbb{R} \cup \{+\infty\}$ は閉凸関数, $Q \subset \operatorname{dom} f$ は閉凸集合である.

ここでは次のような凸関数 $d : E \to \mathbb{R} \cup \{+\infty\}$ の存在を仮定する.

- $Q \subset \operatorname{dom} d$
- $d(x)$ は Q 上で連続かつ定数 $\sigma > 0$ に対して強凸な関数である (微分可能性は要求しない).
- $x_0 = \operatorname*{argmin}_{x \in Q} d(x)$ とする. このとき $d(x_0) = 0$ が成り立つ (一般性を失うことなく仮定してよい[3]).
- $s \in E$ に対して, 最適化問題 $\min_{x \in Q}\{\langle s, x\rangle + d(x)\}$ は (簡単に) 解くことができる.

MD では各反復で正のパラメータとしてステップサイズ $\{\lambda_k\}_{k \geq 0}$ を用いていたが, DA ではさらに非減少な正のパラメータ $\{\beta_k\}_{k \geq -1}$ (スケーリングパラメータと呼ばれる) を導入し, より柔軟な解析を可能にしている. DA アルゴリズムは次のように記述される.

アルゴリズム **3.3.1** (dual-averaging)**.** 上記の関数 $d(x)$ に対して以下をおこなう.

Step 0. $x_0 = \operatorname*{argmin}_{x \in Q} d(x),\ \beta_{-1} > 0$ とする.

Step 1. $k = 0, 1, 2, \ldots$ に対して以下をおこなう.

$$x_{k+1} = \operatorname*{argmin}_{x \in Q}\left\{\sum_{i=0}^{k}\lambda_i[f(x_i) + \langle g_i, x - x_i\rangle] + \beta_k d(x)\right\}, \quad \text{ただし}, \lambda_k > 0,\ \beta_k \geq \beta_{k-1},\ g_k \in \partial f(x_k)$$

[3] さもなくば $d(x) - d(x_0)$ をあらたに $d(x)$ として置き換えればよい.

注意． このアルゴリズムは関数 $d(x)$ に対して初期点 $x_0 = \operatorname*{argmin}_{x \in Q} d(x)$ が固定されてしまうように見える．しかし, 任意の $\widehat{x}_0 \in Q$ に対して $\widehat{s}_0 \in \partial d(\widehat{x}_0)$ をとり,

$$\widehat{d}(x) = d(x) - d(\widehat{x}_0) - \langle \widehat{s}_0, x - \widehat{x}_0 \rangle$$

とおくと, $\widehat{d}(x) \geq \dfrac{\sigma}{2}\|x - \widehat{x}_0\|^2,\ x \in Q$ であるから (系 1.5.4), $\widehat{x}_0 = \operatorname*{argmin}_{x \in Q} \widehat{d}(x),\ \widehat{d}(\widehat{x}_0) = 0$ が成り立つ．すなわち, われわれがアルゴリズムで用いる関数 $d(x)$ をこの関数 $\widehat{d}(x)$ でおきかえることができる．

DA アルゴリズムは次のように解析される．結果は MD のものと似ており, スケーリングパラメータ β_k が新たに含まれている．

定理 3.3.2 ([9], Theorem 1)**.** $x^* \in Q$ を問題 (P) に対する最適解とする．Dual-Averaging アルゴリズムにより得られる $\{x_k\} \subset Q$ は $k \geq 0$ に対して,

$$\frac{\sum_{i=0}^k \lambda_i f(x_i)}{\sum_{i=0}^k \lambda_i} - f(x^*) \leq \frac{\beta_k d(x^*) + \frac{1}{2\sigma}\sum_{i=0}^k \frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2}{\sum_{i=0}^k \lambda_i}$$

を満たす．したがって, $\widehat{x}_k = \sum_{i=0}^k \lambda_i x_i / \sum_{i=0}^k \lambda_i$ とおくと $k \geq 0$ に対して以下が成り立つ．

$$\max\left\{ f(\widehat{x}_k) - f(x^*), \min_{0 \leq i \leq k} f(x_i) - f(x^*) \right\} \leq \frac{\beta_k d(x^*) + \frac{1}{2\sigma}\sum_{i=0}^k \frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2}{\sum_{i=0}^k \lambda_i}$$

この定理から得られる目的関数値と最適値との残差に対する上界を

$$\delta_k = \frac{\beta_k d(x^*) + \frac{1}{2\sigma}\sum_{i=0}^k \frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2}{\sum_{i=0}^k \lambda_i} \tag{3.3.1}$$

とおこう．とくに, $d(x)$ が連続的微分可能かつ $x_0 \in \operatorname{int} Q$ であるとき, $\xi(z, x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$ に対して $\xi(x_0, x^*) = d(x^*)$ となる (x_0 の定義と系 2.1.3 より)．このときに $\beta_k \equiv 1$ としたときの δ_k は MD に対するそれ (3.2.1) と一致する．

Nesterov [9] は, スケーリングパラメータ β_k の選び方に関して, 次の数列 $\{\widehat{\beta}_k\}_{k \geq -1}$ を導入している．

$$\widehat{\beta}_{-1} = \widehat{\beta}_0 = 1, \quad \widehat{\beta}_{k+1} = \widehat{\beta}_k + \widehat{\beta}_k^{-1}\ (k \geq 0) \tag{3.3.2}$$

この数列は以下のように定義しても同じである．

$$\widehat{\beta}_{-1} = 1, \quad \widehat{\beta}_k = \sum_{i=0}^k \frac{1}{\widehat{\beta}_{i-1}} \quad (k \geq 0) \tag{3.3.3}$$

数列 $\{\widehat{\beta}_k\}$ は以下の性質をみたす．

補題 3.3.3 ([9], Lemma 3)**.**

$$\sqrt{2k+1} \leq \widehat{\beta}_k \leq \frac{1}{1+\sqrt{3}} + \sqrt{2k+1}, \quad k \geq 0$$

この $\widehat{\beta}_k$ を用いたパラメータ $\{\lambda_k\}$ および $\{\beta_k\}$ の選び方に関して次の定理が得られる．

定理 3.3.4 ([9], Theorem 2,3)**.** 上記の δ_k に対して以下が成り立つ.

(1) $\sup_{k\geq 0}\|g_k\|_* \leq M$ とするとき, ステップサイズ λ_k とスケーリングパラメータ β_k を

$$\lambda_k \equiv 1, \quad \beta_k = \gamma\widehat{\beta}_k \quad \text{ただし } \gamma > 0$$

と選ぶ. このとき,

$$\forall k \geq 0, \quad \delta_k \leq \left(\gamma d(x^*) + \frac{M^2}{2\sigma\gamma}\right)\frac{0.5+\sqrt{2k+1}}{k+1}$$

が成り立つ. 括弧内の値は, $\gamma = M/\sqrt{2\sigma d(x^*)}$ に対して最小値 $M\sqrt{2d(x^*)/\sigma}$ をとる.
(2) ステップサイズ λ_k とスケーリングパラメータ β_k を

$$\lambda_k = \frac{1}{\|g_k\|_*}, \quad \beta_k = \frac{\widehat{\beta}_k}{\rho\sqrt{\sigma}} \quad \text{ただし } \rho > 0$$

と選ぶ. このとき,

$$\forall k \geq 0, \quad \delta_k \leq \max_{0\leq i\leq k}\|g_i\|_* \frac{1}{\sqrt{\sigma}}\left(\frac{d(x^*)}{\rho} + \frac{\rho}{2}\right)\frac{0.5+\sqrt{2k+1}}{k+1}$$

が成り立つ. 括弧内の値は, $\rho = \sqrt{2d(x^*)}$ のときに最小値 $\sqrt{2d(x^*)}$ をとる.

この定理は, 任意の $k \geq 0$ に対して $O(1/\sqrt{k})$ の上界を与えており, MD に対する以下の欠点を克服している.

- MD では, $O(1/\sqrt{k})$ のオーダーをもつ上界は固定された k に対してのみ得られる.
- MD では, 任意の $k \geq 0$ に対する上界は $O\left(\frac{\log k}{\sqrt{k}}\right)$ までしか示されていない.

この定理の (1), (2) における δ_k の最適な上界は以下のように与えられる.

$$\begin{aligned}(1)\ &:\ \delta_k \leq M\sqrt{\frac{2d(x^*)}{\sigma}}\frac{0.5+\sqrt{2k+1}}{k+1} \\ (2)\ &:\ \delta_k \leq \max_{0\leq i\leq k}\|g_i\|_*\sqrt{\frac{2d(x^*)}{\sigma}}\frac{0.5+\sqrt{2k+1}}{k+1}\end{aligned}$$

$\sqrt{\dfrac{2}{k+1}} \leq \dfrac{0.5+\sqrt{2k+1}}{k+1} \leq \sqrt{\dfrac{5/2}{k+1}}$ に注意すると, $\sup_{k\geq 0}\|g_k\|_* \leq M$ であるとき, δ_k はいずれも $\sqrt{\dfrac{5d(x^*)}{\sigma}}\dfrac{M}{\sqrt{k+1}}$ でおさえられることが分かる. したがって, 精度 $\varepsilon > 0$ の近似解は高々 $\left\lceil\dfrac{5M^2 d(x^*)}{\sigma\varepsilon^2}\right\rceil - 1$ の反復回数で得ることができる. また, 劣勾配 $g_k \in \partial f(x_k)$ を得るための計算量の上界を G, 補助問題 $\min_{x\in Q}\{\langle s, x\rangle + d(x)\}$ を解くのに要する計算量の上界を A とすると, 一反復の計算量の上界は $G + A$ である. したがって, 精度 ε の近似解を得るために十分な計算量として以下が得られる.

$$A + (G+A)\times\left(\left\lceil\frac{5M^2 d(x^*)}{\sigma\varepsilon^2}\right\rceil - 1\right)$$

MD では各反復で $\nabla d(x_k)$ を計算する必要があったが, DA では必要とならないため, 一反復の計算量はより小さくなっている.

DA における反復回数の上界 $\left\lceil \dfrac{5M^2 d(x^*)}{\sigma\varepsilon^2} \right\rceil - 1$ は, MD で示した上界と同じオーダーである．しかし, MD がこのオーダーの上界を実現するには以下の制限があった．

- ステップサイズの決定には最適解 x^* に関する情報が必要である．
- 反復回数をあらかじめ決める必要がある．

DA ではこの制限がない．実際, 定理 3.3.4 により, 問題 (P) の情報によらないパラメータの選び方でも残差の上界は $\forall k \geq 0,\ \delta_k \leq O(1/\sqrt{k})$ を保証する．したがって, 現実的なパラメータ選択のもとでも $O(1/\varepsilon^2)$ の反復回数で精度 ε の近似解を得ることが可能である．

3.4　微分可能な目的関数に対する Nesterov の勾配アルゴリズム

射影勾配法, mirror-descent アルゴリズムおよび dual-averaging アルゴリズムは目的関数に対して微分可能性を仮定していない．これらのアルゴリズムの収束のオーダーは $O(1/\sqrt{k})$ であった．目的関数が微分可能である場合, これらのアルゴリズムがより良い効率を証明できるかどうかはわからない．一方で最急降下法は微分可能な目的関数に対しての解析の結果として, これらのアルゴリズムよりも良い収束のオーダー $O(1/k)$ が得られている．以下では, Nesterov [8] によって提案された連続的微分可能な目的関数に対する勾配アルゴリズムについて述べる．このアルゴリズムは収束のオーダーが $O(1/k^2)$ であることが証明される．

$O(1/k^2)$ の収束オーダーをもつ勾配アルゴリズムが初めて与えられたのは, 1983 年の Nesterov [5] によるものである．その後, このアルゴリズムは彼の著書 [6] において estimate sequence という概念を用いて整理された．これらのアルゴリズムはいずれも E の内積が定めるノルム $\|\cdot\|_2$ にもとづいて構築されたが, 以下に述べる Nesterov [8] のアルゴリズムは, 一般のノルム $\|\cdot\|$ のもとで estimate sequence の概念の言い換えとも見なせる概念を用いて導かれ, $O(1/k^2)$ の収束オーダーをもつことが示された．

この節で用いられる概念は本論文の主結果を導くために重要である．

われわれは以下の凸計画問題を対象とする．

$$(P) \quad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in Q \end{array}$$

ただし, $f : E \to \mathbb{R} \cup \{+\infty\}$ は凸関数 $Q \subset \operatorname{dom} f$ は閉凸集合であり, $f(x)$ は Q 上で連続的微分可能とする．また, 目的関数の勾配 $\nabla f(x)$ は Q 上で定数 L に対してリプシッツ連続であるとする．ここでは, dual-averaging の場合と同じ次のような凸関数 $d : E \to \mathbb{R} \cup \{+\infty\}$ の存在を仮定する．

- $Q \subset \operatorname{dom} d$
- $d(x)$ は Q 上で連続かつ定数 $\sigma > 0$ に対して強凸な関数である (微分可能性は要求しない).
- $x_0 = \operatorname*{argmin}_{x \in Q} d(x)$ に対して, $d(x_0) = 0$ が成り立つ (一般性を失うことなく仮定してよい).
- $s \in E$ に対して, 最適化問題 $\min_{x \in Q} \{\langle s, x \rangle + d(x)\}$ は (簡単に) 解くことができる．

Nesterov の勾配アルゴリズムは, 以下の目標のもとに構築される：反復的にふたつの点列 $\{\hat{x}_k\}, \{x_k\} \subset$

Q を以下の関係式を満たすように生成する.

$$(R_k) \quad S_k f(\widehat{x}_k) \le \min_{x\in Q}\left\{\sum_{i=0}^{k}\lambda_i[f(x_i)+\langle\nabla f(x_i), x-x_i\rangle]+\frac{L}{\sigma}d(x)\right\} \tag{3.4.1}$$

ただし, λ_k は正のパラメータ (ステップサイズ) であり, $S_k=\sum_{i=0}^{k}\lambda_i$ とする. この関係式の重要性は次の事実による：問題 (P) の最適解 $x^*\in Q$ に対して,

$$(R_k) \implies f(\widehat{x}_k)-f(x^*) \le \frac{Ld(x^*)}{\sigma S_k}$$

が成り立つ. そこで, われわれは関係式 (R_k) を満たしつつ $1/S_k$ がなるべく速く 0 に収束するような $\{\widehat{x}_k\}$, $\{x_k\}, \{\lambda_k\}$ のとりかたが提案できないかを考える.

次のアルゴリズムはある条件下で上記の関係式 (R_k) を満たすようにつくられたものである.

アルゴリズム **3.4.1** (Nesterov [8])**.** 上記の関数 $d(x)$ に対して以下をおこなう.

Step 0. $x_0=\operatorname*{argmin}_{x\in Q} d(x)$, $\lambda_0>0$ とする.
Step 1. $k=0,1,2,\ldots$ に対して以下をおこなう.

$$\widehat{x}_k = \operatorname*{argmin}_{x\in Q}\left\{f(x_k)+\langle\nabla f(x_k), x-x_k\rangle+\frac{L}{2}\|x-x_k\|^2\right\}$$

$$z_k = \operatorname*{argmin}_{x\in Q}\left\{\sum_{i=0}^{k}\lambda_i[f(x_i)+\langle\nabla f(x_i), x-x_i\rangle]+\frac{L}{\sigma}d(x)\right\}$$

$$x_{k+1} = \frac{S_k\widehat{x}_k+\lambda_{k+1}z_k}{S_{k+1}} \quad \text{ただし } \lambda_{k+1}>0$$

このアルゴリズムに対して, 次の定理が成り立つ.

定理 3.4.2. $\lambda_k^2 \le S_k$, $k\ge 0$ を満たすようなステップサイズ $\{\lambda_k\}$ に対して, アルゴリズム 3.4.1 が生成する点列 $\{\widehat{x}_k\}, \{x_k\}\subset Q$ は, 任意の $k\ge 0$ に対して関係式 (R_k) を満たす. とくに, $\lambda_k=(k+1)/2$ ととれば以下が成り立つ.

$$\forall k\ge 0, \quad f(\widehat{x}_k)-f(x^*) \le \frac{4Ld(x^*)}{\sigma(k+1)(k+2)}$$

このアルゴリズムは各反復で二つの補助最適化問題を解く必要がある. とくに $\widehat{x}_k$ の計算はノルムによっては困難となることが予想される. この困難を修正するために Nesterov は以下の修正アルゴリズムの提案もおこなっている. ただし, このアルゴリズムでは関数 $d(x)$ が Q 上で連続的微分可能であることを仮定する必要がある. このとき, われわれは次の Bregman distance を利用する.

$$\xi(z,x)=d(x)-d(z)-\langle\nabla d(z), x-z\rangle, \quad x,z\in Q$$

修正アルゴリズムは以下のように記述される.

アルゴリズム **3.4.3** (Nesterov [8])**.** 上記の関数 $d(x)$ が Q 上で微分可能であるとする.

Step 0. $x_0 = \operatorname*{argmin}_{x\in Q} d(x)$, $\lambda_0 > 0$ とし, $\widehat{x}_0, \widehat{z}_0, z_0$ を以下のようにとる.

$$\widehat{x}_0 = \widehat{z}_0 = z_0 = \operatorname*{argmin}_{x\in Q} \left\{ \lambda_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \frac{L}{\sigma} d(x) \right\}$$

Step 1. $k = 0, 1, 2, \ldots$ に対して以下をおこなう.

$$\begin{aligned}
x_{k+1} &= \frac{\sum_{i=0}^{k} \lambda_i \widehat{z}_i + \lambda_{k+1} z_k}{S_{k+1}} \quad \text{ただし } \lambda_{k+1} > 0 \\
z_{k+1} &= \operatorname*{argmin}_{x\in Q} \left\{ \sum_{i=0}^{k+1} \lambda_i [f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] + \frac{L}{\sigma} d(x) \right\} \\
\widehat{z}_{k+1} &= \operatorname*{argmin}_{x\in Q} \left\{ \lambda_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] + \frac{L}{\sigma} \xi(z_k, x) \right\} \\
\widehat{x}_{k+1} &= \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i \widehat{z}_i
\end{aligned}$$

この修正アルゴリズムに関しても, 定理 3.4.2 と同じ主張が成り立つ ([8], Lemma 7). したがって, ステップサイズ $\lambda_k = (k+1)/2$ に対する修正アルゴリズムからは, 以下を満たす点列 $\{\widehat{x}_k\} \subset Q$ が得られる.

$$f(\widehat{x}_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma(k+1)(k+2)} \leq \frac{4Ld(x^*)}{\sigma(k+1)^2}$$

ゆえに, 精度 $\varepsilon > 0$ の近似解を得るのには高々 $\left\lceil \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil - 1$ の反復回数で十分である. とくに, 勾配 $\nabla f(x)$, $\nabla d(x)$ の一回の評価に必要な計算量の上界をそれぞれ G_f, G_d とし, 補助問題 $\min_{x\in Q}[\langle s, x\rangle + d(x)]$ を解くために要する計算量の上界を A とすると, 精度 ε の近似解を得るために十分な計算量は次のように与えられる.

$$(G_f + G_d + 2A) \times \left\lceil \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil$$

3.5　Extremal convex problem に対する Nesterov のアルゴリズム

Extremal convex problem に対するアルゴリズムは Nesterov [5] によって与えられた. このアルゴリズムは E の内積が定めるノルム $\|\cdot\|_2$ に関して提案されている. 一般のノルムに関しての議論は [8] によってなされているが, 問題のクラスは限定される. われわれは第 4 章で, 一般のノルムに関して extremal convex problem に対するアルゴリズムの提案をおこなう. この節では, Nesterov [5] のアルゴリズムを述べよう.

Nesterov [5] は, extremal convex problem のうち $\phi(u) \equiv 0$ であるものに対するアルゴリズムを提案する. すなわち, 次の凸計画問題 (P) を考える.

$$\begin{aligned}
(P) \qquad \min \quad & f(x) = \max_{u\in U} \sum_{j=1}^{m} u^{(j)} f_j(x) \\
\text{s.t.} \quad & x \in Q
\end{aligned}$$

ただし, Q は E 上の閉凸集合であり, $f_j(x): E \to \mathbb{R}$ は E 上連続的微分可能[4]な凸関数とする. また, 仮定 2.2.1 が成り立つとする.

さらに 各 $j = 1, \ldots, m$ に対して $\nabla f_j(x)$ が Q 上で定数 $L_j \geq 0$ に対してリプシッツ連続であると仮定する. このとき, $L = \max_{u \in U} \sum_{j=1}^{m} u^{(j)} L_j$ とおく. Nesterov [5] はリプシッツ定数 L_j が未知である場合にも対応できるアルゴリズムを提案している. とくに, リプシッツ定数が既知である場合は定数 L を利用して次のようなアルゴリズムが得られる.

アルゴリズム **3.5.1.** (Nesterov [5]) 上記の extremal convex problem に対して以下をおこなう.

Step 0. $\widehat{x}_0 = x_0 \in E, \quad a_0 = 1$ とする.
Step 1. $k = 0, 1, 2, \ldots$ に対して以下をおこなう.

$$\begin{aligned}
\widehat{x}_{k+1} &= \operatorname*{argmin}_{x \in Q} \left\{ f(x_k, x) + \frac{L}{2} \|x - x_k\|_2^2 \right\} \\
a_{k+1} &= \frac{1 + \sqrt{4a_k^2 + 1}}{2} \\
x_{k+1} &= \widehat{x}_k + \frac{a_k - 1}{a_k} (\widehat{x}_k - \widehat{x}_{k-1})
\end{aligned}$$

このアルゴリズムの収束率の見積もりを述べる.

定理 **3.5.2.** $x^* \in Q$ を問題 (P) の最適解とする. アルゴリズム 3.5.1 が生成する点列 $\{\widehat{x}_k\}_{k \geq 1} \subset Q$ は以下を満たす.

$$\forall k \geq 1, \quad f(\widehat{x}_k) - f(x^*) \leq \frac{2L\|x_0 - x^*\|_2^2}{(k+1)^2}$$

この定理から, アルゴリズム 3.5.1 の効率を調べることができる. 精度 $\varepsilon > 0$ に対して,

$$\frac{2L\|x_0 - x^*\|_2^2}{(k+1)^2} \leq \varepsilon \iff k \geq \sqrt{\frac{2L}{\varepsilon}} \|x_0 - x^*\|_2 - 1 \iff k \geq \left\lceil \sqrt{\frac{2L}{\varepsilon}} \|x_0 - x^*\|_2 \right\rceil - 1$$

であるから, 精度 ε の近似解 $\widehat{x}_k$ を得るのに必要な反復回数 k は高々 $\left\lceil \sqrt{\frac{2L}{\varepsilon}} \|x_0 - x^*\|_2 \right\rceil - 1$ である. 関数値 $(f_1(x), \ldots, f_m(x))$ を一回評価ために要する計算量の上界を F, 勾配 $(\nabla f_1(x), \ldots, \nabla f_m(x))$ を一回評価するために要する計算量の上界を G とする. さらに $a_1, \ldots, a_m \in E$, $b \in \mathbb{R}^m$, $y \in Q$ に対する補助問題

$$\min_{x \in Q} \left\{ \max_{u \in U} \sum_{j=1}^{m} u^{(j)} [\langle a_j, x \rangle + b^{(j)}] + \frac{1}{2} \|x - y\|_2^2 \right\}$$

の最適解を得るために必要な計算量の上界を A とすると, 一反復の計算量の上界は $F + G + A$ である. ゆえに, 精度 ε の近似解を得るためには次の計算量で十分である.

$$(F + G + A) \times \left(\left\lceil \sqrt{\frac{2L}{\varepsilon}} \|x_0 - x^*\|_2 \right\rceil - 1 \right) \tag{3.5.1}$$

[4] 本論文の定義 (2.2.1) では Q 上でしか微分可能性を仮定していないが, Nesterov のアルゴリズム [5] では一般に $\{x_k\} \subset Q$ とは限らないため, E 上での微分可能性を仮定している.

第 4 章

劣勾配アルゴリズムを統一的に扱うための新しい枠組み

第 3 章に述べた劣勾配アルゴリズムにおいて, mirror-descent や dual-averaging, さらには Nesterov [8] のアルゴリズムはいずれも, ある条件を満たす強凸関数 $d(x)$ を用いて各反復で補助問題を構築し, その補助最適解を利用した近似解の更新をおこなっていた. これらの共通点はこの章に述べる枠組みによって統一的に解析することができる. その結果として, 既存アルゴリズムに比べてより効率的なアルゴリズムを提案することができる.

4.1 Mirror-descent および dual-averaging の拡張概念

この節では, 次の凸計画問題を考える.

$$
\begin{array}{lll}
(P) & \min & f(x) \\
 & \text{s.t.} & x \in Q
\end{array}
$$

ただし, $f : E \to \mathbb{R} \cup \{+\infty\}$ は閉凸関数, $Q \subset \operatorname{dom} f$ は閉凸集合とする.

われわれは, 目的関数 $f(x)$ に関して以下の場合を考え, それぞれに対して効率的な反復的アルゴリズムを考察する.

(a) $f(x)$ が一般の閉凸関数のとき.

(b) $f(x)$ が Q 上で連続的微分可能な凸関数で, 勾配 $\nabla f(x)$ がリプシッツ連続のとき.

(c) 問題 (P) が extremal convex problem のとき.

とくに, (c) の場合を考えるとき, 目的関数は以下で与えられるとする.

$$
f(x) = \max_{u \in U} \left\{ \sum_{j=1}^{m} u^{(j)} f_j(x) - \phi(u) \right\}
$$

ただし, $\phi : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ は連続凸関数, $U \subset \operatorname{dom} \phi$ は有界閉集合, $f_j(x)$ は Q 上連続的微分可能な凸関数であり, 仮定 2.2.1 が成り立つとする. さらに $\nabla f_j(x)$ の Q 上でのリプシッツ連続性も仮定する.

(a) の場合に対しては, Mirror-Descent (MD) や Dual-Averaging (DA) のような $O(1/\sqrt{k})$ のオーダーで収束するアルゴリズムが存在することを 3 章で紹介した. (b) および (c) の場合に対しては, Nesterov の

アルゴリズム [5, 8] が $O(1/k^2)$ の収束率をもつことを示した．この章では, MD および DA を統一的に議論できるようなアルゴリズムの枠組みを示し, それによって導かれるアルゴリズムはより効率的な収束率の見積もりを与えることを証明する．さらには, この枠組みから (b) や (c) の場合に対しても効率的なアルゴリズムを提案することができることを示す．

まず, 共通の記号を導入する．整数 $k \geq 0$ に対して以下のものを考える．

$x_k \in Q$	:	目的関数 (またはその劣勾配など) の値を評価する点
$\widehat{x}_k \in Q$	:	問題 (P) に対する近似解
$g_k \in \partial f(x_k)$	:	x_k における $f(x)$ の劣勾配
$\psi_k(x)$	:	補助関数 (反復的に解く補助問題の目的関数)
$z_k \in Q$	:	補助問題 $\min\limits_{x \in Q} \psi_k(x)$ の最適解
$\widehat{\psi}_k(x)$	:	評価関数 (目的関数の上からの見積もりに関わる関数)

以降に考察する反復的アルゴリズムの流れに沿ってこれらの記号を概説すると次のように述べることができる：(1) 点 x_k において目的関数 $f(x)$ や劣勾配 g_k を評価する, (2) 補助関数 $\psi_k(x)$ を構成して補助問題 $\min_{x \in Q} \psi_k(x)$ の解 z_k を計算する, (3) 問題 (P) に対する近似解 $\widehat{x}_{k+1}$ (および x_{k+1}) を更新する, (4) このとき問題 (P) の最適値に対する近似値 $f(\widehat{x}_{k+1})$ は評価関数 $\widehat{\psi}_{k+1}(x)$ に関してのある基準を満たしている．

われわれは MD や DA などの場合と同様に, 次の条件を満たす凸関数 $d : E \to \mathbb{R} \cup \{+\infty\}$ の存在を仮定する．

- $Q \subset \operatorname{dom} d$
- $d(x)$ は Q 上で連続的微分可能かつ定数 $\sigma > 0$ に対して強凸である．
- $x_0 = \operatorname*{argmin}\limits_{x \in Q} d(x)$ とする．このとき $d(x_0) = 0$ が成り立つ (一般性を失うことなく仮定してよい)．
- $s \in E$ に対して, 補助最適化問題 $\min\limits_{x \in Q} \{\langle s, x \rangle + d(x)\}$ は (簡単に) 解くことができる．

ただし, (c) の場合, すなわち問題 (P) が extremal convex problem である場合には, 最後の条件を次のように変える必要がある：$a_1, \ldots, a_m,\ g \in E,\ b \in \mathbb{R}^m$ に対して, 補助最適化問題

$$\begin{array}{ll} \min & \max\limits_{u \in U} \left\{ \sum\limits_{j=1}^{m} u^{(j)} [\langle a_j, x \rangle + b^{(j)}] - \phi(u) \right\} + d(x) + \langle g, x \rangle \\ \text{s.t.} & x \in Q \end{array}$$

は (簡単に) 解くことができる．

われわれはこの関数 $d(x)$ に対して, 次の Bregman distance を定義する．

$$\xi(z, x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle, \quad x, z \in Q$$

この関数 $d(x)$ は補助問題の構築に用いられる, すなわち, 補助関数 $\psi_k(x)$ の定義に利用する．補助問題の構成には, このほかに次の二つのパラメータを用いる．

- ステップサイズ $\{\lambda_k\}_{k \geq 0},\ \lambda_k > 0$
- スケーリングパラメータ $\{\beta_k\}_{k \geq -1},\ \beta_{k+1} \geq \beta_k > 0$

ステップサイズは第 3 章で扱ったほとんどの劣勾配アルゴリズムで用いられたパラメータであり, スケーリングパラメータは DA において導入されたものである. たとえば, MD および DA の反復は以下のようにして点列 $\{x_k\}$ を更新していた.

$$\begin{aligned} \text{MD} \ : \ x_{k+1} &= \operatorname*{argmin}_{x \in Q} \{\lambda_k[f(x_k) + \langle g_k, x - x_k\rangle] + \xi(x_k, x)\} \\ \text{DA} \ : \ x_{k+1} &= \operatorname*{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i[f(x_i) + \langle g_i, x - x_i\rangle] + \beta_k d(x) \right\} \end{aligned}$$

すなわち, 以下のような補助関数 $\psi_k(x)$

$$\begin{aligned} \text{MD} \ : \ \psi_k(x) &= \lambda_k[f(x_k) + \langle g_k, x - x_k\rangle] + \xi(x_k, x) \\ \text{DA} \ : \ \psi_k(x) &= \sum_{i=0}^{k} \lambda_i[f(x_i) + \langle g_i, x - x_i\rangle] + \beta_k d(x) \end{aligned}$$

が定める補助問題の解 $z_k = \operatorname*{argmin}_{x \in Q} \psi_k(x)$ に対して, そのまま $x_{k+1} = z_k$ として更新している.

補助関数 $\psi_k(x)$ や評価関数 $\widehat{\psi}_k(x)$ を具体的に定めるために, 目的関数 $f(x)$ の近似モデルを定義すると便利である. 目的関数の構造に応じて, $f(x)$ の近似モデル $l_k(x)$ を次のように定める.

$$l_k(x) = \begin{cases} f(x_k) + \langle g_k, x - x_k\rangle & : \ \text{(a) の場合} \\ f(x_k) + \langle \nabla f(x_k), x - x_k\rangle & : \ \text{(b) の場合} \\ f(x_k, x) & : \ \text{(c) の場合} \end{cases} \tag{4.1.1}$$

(b) の場合は, 上の定義はすべて等しくなる. $l_k(x)$ は閉凸関数であり, $f(x)$ の下からの近似を与える, すなわち

$$f(x) \geq l_k(x), \quad x \in \operatorname{dom} f \tag{4.1.2}$$

が成り立つ ((a) は定義 1.2.4, (b) は定理 1.4.1 および (c) は命題 2.2.4 からわかる).

(a) の場合, MD アルゴリズム 3.2.1 および DA アルゴリズム 3.3.1 の反復は次の書きかえられる.

$$\text{MD} : \begin{cases} \psi_k(x) &= \ \lambda_k l_k(x) + \xi(x_k, x) \\ x_{k+1} &= \ z_k = \operatorname*{argmin}_{x \in Q} \psi_k(x) \\ \widehat{x}_{k+1} &= \ \dfrac{1}{S_{k+1}} \displaystyle\sum_{i=0}^{k+1} x_i \end{cases}$$

$$\text{DA} : \begin{cases} \psi_k(x) &= \ \displaystyle\sum_{i=0}^{k} \lambda_i l_i(x) + \beta_k d(x) \\ x_{k+1} &= \ z_k = \operatorname*{argmin}_{x \in Q} \psi_k(x) \\ \widehat{x}_{k+1} &= \ \dfrac{1}{S_{k+1}} \displaystyle\sum_{i=0}^{k+1} x_i \end{cases}$$

以下に MD および DA を統一的に扱うためのアルゴリズムの枠組みを提案しよう.

ステップサイズ $\lambda_k > 0$ に対して, $S_k = \sum_{i=0}^{k} \lambda_k$ とする．このとき, (P) に対する近似解 $\widehat{x}_k$ と評価関数 $\widehat{\psi}_k(x)$ に関して次の関係式 (R_k) を考える．

$$(R_k) \quad S_k f(\widehat{x}_k) \leq \min_{x \in Q} \widehat{\psi}_k(x) + C_k \tag{4.1.3}$$

ただし, C_k は目的関数の構造に応じて以下のように定める．

$$C_k = \begin{cases} \dfrac{1}{2\sigma} \displaystyle\sum_{i=0}^{k} \dfrac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2 & : \text{(a) の場合} \\ 0 & : \text{それ以外} \end{cases}$$

特別なアルゴリズムに対しては, 次の関係式についても考察する．

$$(\widehat{R}_k) \quad \sum_{i=0}^{k} \lambda_i f(x_i) \leq \min_{x \in Q} \widehat{\psi}_k(x) + C_k \tag{4.1.4}$$

関係式 (R_k) は, $f(\widehat{x}_k)$ の値に対してパラメータ $\{\lambda_i\}_{i=0}^{k}, \{\beta_i\}_{i=-1}^{k}$ および $\widehat{\psi}_k(x)$ による上からの見積もりを与えている．われわれは, $\{\lambda_k\}, \{\beta_k\}, \{\widehat{\psi}_k(x)\}$ に適当な条件を与え, そのもとでこの関係式 (R_k) を満たす点列 $\{\widehat{x}_k\}$ の更新式を具体的に見つけることによって, 反復的アルゴリズムを構築する．

Q 上で $d(x)$ を超えない関数 $d_k(x)$ に対して, 以下の枠組みを考える．ここで, $k = -1$ に対して $z_{-1} = x_0$ を便宜的に用いる．さらに, 評価関数については $\widehat{\psi}_{-1}(x)$ も考えることにする．

条件 4.1.1. 近似モデル $\{l_k(x)\}$, 補助問題の解 $\{z_k\}$, 正のパラメータ $\{\lambda_k\}$(ステップサイズ), 正の非減少パラメータ $\{\beta_k\}$(スケーリングパラメータ), 評価関数 $\{\widehat{\psi}_k(x)\}$ および Q 上で $d(x)$ を超えない関数 $d_k(x)$ は以下を満たす.

(i) $\min_{x \in Q} \widehat{\psi}_{-1}(x) = 0$

(ii) 任意の $k \geq -1$ に対して次の不等式が成り立つ.

$$\min_{x \in Q} \widehat{\psi}_{k+1}(x) \geq \min_{x \in Q} \widehat{\psi}_k(x) + \lambda_{k+1} l_{k+1}(z_{k+1}) + \beta_k \xi(z_k, z_{k+1})$$

(iii) 任意の $k \geq 0$ に対して次の不等式が成り立つ.

$$\min_{x \in Q} \widehat{\psi}_k(x) \leq \min_{x \in Q} \left[\sum_{i=0}^{k} \lambda_i l_i(x) + \beta_k d_k(x) \right]$$

条件 4.1.1 のもとでは, 関係式 (R_k) を満たす点列 $\{\widehat{x}_k\}$ を生成する更新式をいくつか提案できることを後に述べる．まずはこの仮定と関係式 (R_k) から得られる目的関数値の最適値との残差に関する不等式を示そう．

命題 4.1.2. $x^* \in Q$ を問題 (P) の最適解とし, 条件 4.1.1 が満たされるとする．

(1) 点列 $\{\widehat{x}_i\} \subset Q$ がある $k \geq 0$ に対して関係式 (R_k) を満たすならば, 以下が成り立つ．

$$f(\widehat{x}_k) - f(x^*) \leq \frac{\beta_k d_k(x^*) + C_k}{S_k}$$

(2) 点列 $\{x_i\} \subset Q$ がある $k \geq 0$ に対して関係式 $(\widehat{R}_k)$ を満たすならば, 以下が成り立つ.

$$\min_{0\leq i\leq k} f(x_i) - f(x^*) \leq \frac{\beta_k d_k(x^*) + C_k}{S_k}$$

証明. 条件 4.1.1 (iii) および (4.1.2) より,

$$\begin{aligned}
\min_{x\in Q} \widehat{\psi}_k(x) &\leq \min_{x\in Q} \left\{ \sum_{i=0}^{k} \lambda_i l_i(x) + \beta_k d_k(x) \right\} \\
&\leq \sum_{i=0}^{k} \lambda_i l_i(x^*) + \beta_k d_k(x^*) \\
&\leq \sum_{i=0}^{k} \lambda_i f(x^*) + \beta_k d_k(x^*) \\
&= S_k f(x^*) + \beta_k d_k(x^*)
\end{aligned}$$

となる. したがって, 関係式 (R_k) が成り立つとき,

$$S_k f(\widehat{x}_k) \leq \min_{x\in Q} \widehat{\psi}_k(x) + C_k \leq S_k f(x^*) + \beta_k d_k(x^*) + C_k$$

であるから, 両辺を S_k でわれば (1) の不等式が示される. また, (2) も $\displaystyle\min_{0\leq i\leq k} f(x_i) \leq \frac{1}{S_k}\sum_{i=0}^{k} \lambda_i f(x_i)$ に注意すれば, 関係式 $(\widehat{R}_k)$ が成り立つとき,

$$\begin{aligned}
S_k \min_{0\leq i\leq k} f(x_k) \leq S_k \times \frac{1}{S_k}\sum_{i=0}^{k} \lambda_i f(x_i) = \sum_{i=0}^{k} \lambda_i f(x_i) &\leq \min_{x\in Q} \widehat{\psi}_k(x) + C_k \\
&\leq S_k f(x^*) + \beta_k d_k(x^*) + C_k
\end{aligned}$$

であるから, 両辺を S_k でわることで, (2) の不等式が得られる. □

MD および DA がこの枠組みから導かれることを示すために, これらのアルゴリズムにおける補助関数 $\psi_k(x)$ の定義に対して条件 4.1.1 が成り立つ (ように評価関数を定義できる) ことを証明しよう. まず DA のほうから述べる.

命題 4.1.3. 補助関数 $\psi_k(x)$, 評価関数 $\widehat{\psi}_k(x)$ および $d_k(x)$ を以下のように定める.

$$\begin{aligned}
\psi_k(x) &= \widehat{\psi}_k(x) = \sum_{i=0}^{k} \lambda_i l_i(x) + \beta_k d(x) \quad (k \geq 0) \\
\widehat{\psi}_{-1}(x) &= \beta_{-1} d(x) \\
d_k(x) &= d(x) \quad (k \geq 0)
\end{aligned}$$

このとき, 条件 4.1.1 が成り立つ.

証明. $\beta_{-1} > 0$, $\min_{x\in Q} d(x) = 0$ であるから, (i) $\min_{x\in Q} \widehat{\psi}_{-1}(x) = 0$ が成り立つ. また, $\widehat{\psi}_k(x), d_k(x)$ の定め方より, (iii) において等号が成り立つ. 最後に (ii) を確認しよう. $k \geq -1$ とする. 補助関数と評価関

数の定め方および $z_{-1} = x_0 = \operatorname*{argmin}_{x \in Q} d(x)$ に注意すると, $z_k = \operatorname*{argmin}_{x \in Q} \widehat{\psi}_k(x)$ が成り立つ. ゆえに, $\widehat{\psi}_k(x)$ の作りかたに注意すれば, 定理 2.1.6 より,

$$\widehat{\psi}_k(x) \geq \widehat{\psi}_k(z_k) + \beta_k \xi(z_k, x), \quad x \in Q$$

が成り立つ. したがって, 任意の $x \in Q$ に対して, $d(x) \geq 0$ であることと $\beta_{k+1} \geq \beta_k > 0$ に注意すれば,

$$\begin{aligned}
\widehat{\psi}_{k+1}(x) &= \sum_{i=0}^{k} \lambda_i l_i(x) + \lambda_{k+1} l_{k+1}(x) + \beta_{k+1} d(x) \\
&\geq \sum_{i=0}^{k} \lambda_i l_i(x) + \lambda_{k+1} l_{k+1}(x) + \beta_k d(x) \\
&= \widehat{\psi}_k(x) + \lambda_{k+1} l_{k+1}(x) \\
&\geq \widehat{\psi}_k(z_k) + \lambda_{k+1} l_{k+1}(x) + \beta_k \xi(z_k, x)
\end{aligned}$$

となる. とくに $x = z_{k+1} \in Q$ とすれば (ii) が得られる. □

つぎに, MD にスケーリングパラメータ β_k を導入したモデルを考え, それが条件 4.1.1 を満たすことを示そう.

命題 4.1.4. 補助関数 $\psi_k(x)$, 評価関数 $\widehat{\psi}_k(x)$ および $d_k(x)$ を以下のように定める.

$$\begin{array}{lcl}
\psi_k(x) & = & \lambda_k l_k(x) + \beta_k d(x) - \beta_{k-1}[d(z_{k-1}) + \langle \nabla d(z_{k-1}), x - z_{k-1} \rangle] \quad (k \geq 0) \\
\widehat{\psi}_k(x) & = & \displaystyle\sum_{i=0}^{k-1} \psi_i(z_i) + \psi_k(x) \quad (k \geq 0) \\
\widehat{\psi}_{-1}(x) & = & \beta_{-1}[d(z_{-1}) + \langle \nabla d(z_{-1}), x - z_{-1} \rangle] \\
d_k(x) & = & d(z_k) + \langle \nabla d(z_k), x - z_k \rangle \quad (k \geq 0)
\end{array}$$

このとき, 条件 4.1.1 が成り立つ.

証明. まず, $d_{-1}(x) = d(z_{-1}) + \langle \nabla d(z_{-1}), x - z_{-1} \rangle$ とかくことにすれば,

$$\begin{array}{lcl}
\psi_k(x) & = & \lambda_k l_k(x) + \beta_k d(x) - \beta_{k-1} d_{k-1}(x) \quad (k \geq 0) \\
\widehat{\psi}_{-1}(x) & = & \beta_{-1} d_{-1}(x)
\end{array}$$

とあらわせる. $z_{-1} = x_0 = \operatorname*{argmin}_{x \in Q} d(x)$ であるから, 最適性条件 (系 2.1.3) より $\min_{x \in Q} d_{-1}(x) = d_{-1}(z_{-1}) = 0$ となる. したがって, $\min_{x \in Q} \widehat{\psi}_{-1}(x) = \widehat{\psi}_{-1}(z_{-1}) = 0$, すなわち (i) が成り立つ.

つぎに (ii) を示そう. $k \geq -1$ とする. $\widehat{\psi}_k(x)$ の定義のしかたより, $z_k = \operatorname*{argmin}_{x \in Q} \widehat{\psi}_k(x)$ が成り立つ. ゆえに,

$$\min_{x \in Q} \widehat{\psi}_k(x) = \widehat{\psi}_k(z_k) = \sum_{i=0}^{k} \psi_i(z_i)$$

である. したがって, $\beta_{k+1} \geq \beta_k$ に注意すると, 任意の $x \in Q$ に対して,

$$\begin{aligned}
\widehat{\psi}_{k+1}(x) &= \sum_{i=0}^{k} \psi_i(z_i) + \psi_{k+1}(x) \\
&= \min_{x \in Q} \widehat{\psi}_k(x) + \psi_{k+1}(x) \\
&= \min_{x \in Q} \widehat{\psi}_k(x) + \lambda_{k+1} l_{k+1}(x) + \beta_{k+1} d(x) - \beta_k d_k(x) \\
&\geq \min_{x \in Q} \widehat{\psi}_k(x) + \lambda_{k+1} l_{k+1}(x) + \beta_k \xi(z_k, x)
\end{aligned}$$

であるから, とくに $x = z_{k+1} \in Q$ とすれば, (ii) が得られる.

最後に (iii) を示そう.

$$\widehat{\Psi}_k(x) = \sum_{i=0}^{k} \lambda_i l_i(x) + \beta_k d_k(x) \quad (k \geq 0), \quad \widehat{\Psi}_{-1}(x) = \beta_{-1} d_{-1}(x)$$

とおくと, われわれの目標は $k \geq 0$ に対して $\min_{x \in Q} \widehat{\psi}_k(x) \leq \min_{x \in Q} \widehat{\Psi}_k(x)$ を示すことと等しい. 実際, $k \geq -1$ に対して,

$$\begin{aligned}
\widehat{\Psi}_{k+1}(x) &= \sum_{i=0}^{k} \lambda_i l_i(x) + \lambda_{k+1} l_{k+1}(x) + \beta_{k+1} d_{k+1}(x) \\
&= \sum_{i=0}^{k} \lambda_i l_i(x) + \beta_k d_k(x) \\
&\qquad - \beta_k d_k(x) + \lambda_{k+1} l_{k+1}(x) + \beta_{k+1} d(x) \\
&\qquad - \beta_{k+1} d(x) + \beta_{k+1} d_{k+1}(x) \\
&= \widehat{\Psi}_k(x) + \psi_{k+1}(x) - \beta_{k+1} \xi(z_{k+1}, x)
\end{aligned}$$

が成り立つので,

$$\widehat{\Psi}_k(x) = \widehat{\psi}_{-1}(x) + \sum_{i=0}^{k} [\psi_i(x) - \beta_i \xi(z_i, x)]$$

とあらわせる. $z_k = \operatorname*{argmin}_{x \in Q} \psi_k(x)$ に対して定理 2.1.6 より,

$$\psi_k(x) - \beta_k \xi(z_k, x) \geq \min_{x \in Q} \psi_k(x) = \psi_k(z_k), \quad \forall x \in Q$$

であることと (i) に注意すれば, 任意の $k \geq 0$ に対して,

$$\min_{x \in Q} \widehat{\Psi}_k(x) \geq \sum_{i=0}^{k} \psi_i(z_i) = \min_{x \in Q} \widehat{\psi}_k(x)$$

がわかる. □

4.2　微分不可能な目的関数に対するアルゴリズム

この節では, 4.1 節の問題 (P) において, (a) の場合を考える. すなわち, われわれは次の凸計画問題に対する効率的なアルゴリズムについて考察する.

$$(P) \qquad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in Q \end{array}$$

ただし, $f : E \to \mathbb{R} \cup \{+\infty\}$ は閉凸関数, $Q \subset \operatorname{dom} f$ は閉凸集合である.

このとき, 点列 $\{x_k\}$, $g_k \in \partial f(x_k)$ に対して, $f(x)$ の近似モデル $l_k(x)$ は式 (4.1.1) より

$$l_k(x) = f(x) + \langle g_k, x - x_k \rangle \tag{4.2.1}$$

で定義される. また (4.1.3) から, われわれは以下の関係式 (R_k) を満たす近似解 $\{\widehat{x}_k\}$ が得られるアルゴリズムを目指す：

$$(R_k) \qquad S_k f(\widehat{x}_k) \le \min_{x \in Q} \widehat{\psi}_k(x) + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2 \tag{4.2.2}$$

以降では, Q 上の最小点が計算可能な補助関数 $\psi_k(x)$ とその最小点 $z_k \in Q$ が, ステップサイズ $\lambda_k > 0$, スケーリングパラメータ $\beta_k > 0$, 評価関数 $\widehat{\psi}_k(x)$ および Q 上で $d(x)$ を超えない関数 $d_k(x)$ に対して, 条件 4.1.1 を満たすとする.

4.2.1　アルゴリズムの導出

以下に, 条件 4.1.1 と関係式 (R_k) に関する解析をおこなうのに必要な補題を与える.

補題 4.2.1. 任意の $a, b \in E$ に対して,

$$\frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|_*^2 \ge \langle a, b \rangle$$

が成り立つ.

証明. 双対ノルムの性質より, $\langle a, b \rangle \le \|a\|\|b\|_*$ であるから,

$$0 \le \frac{1}{2}(\|a\| - \|b\|_*)^2 = \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b^2\| - \|a\|\|b\|_* \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b^2\| - \langle a, b \rangle$$

□

補題 4.2.2. $\lambda \in \mathbb{R}$, $\beta > 0$ および $x, z \in Q$ に対して,

$$\langle \lambda g_k, x - z \rangle + \beta \xi(z, x) + \frac{1}{2\sigma\beta}\|\lambda g_k\|_*^2 \ge 0, \quad \forall k \ge 0$$

が成り立つ. したがってとくに,

$$\lambda l_k(x) + \beta \xi(x_k, x) + \frac{\lambda^2}{2\sigma\beta}\|g_k\|_*^2 \ge \lambda f(x_k), \quad \forall k \ge 0$$

証明. Bregman distance の性質により,

$$\langle \lambda g_k, x-z\rangle + \beta\xi(z,x) + \frac{1}{2\sigma\beta}\|\lambda g_k\|_*^2 \geq \langle \lambda g_k, x-z\rangle + \frac{\sigma\beta}{2}\|x-z\|^2 + \frac{1}{2\sigma\beta}\|\lambda g_k\|_*^2$$

が成り立つ. この右辺が ≥ 0 であることは, 補題 4.2.1 において,

$$a = \sqrt{\sigma\beta}(x-z), \quad b = -\frac{\lambda g_k}{\sqrt{\sigma\beta}}$$

ととればわかるので, 前半の不等式が示された. また前半の不等式において $z = x_k$ とし, 両辺に $\lambda f(x_k)$ を加えれば, 後半の不等式がわかる. □

われわれの提案するアルゴリズムは, 次の定理によって得られる.

定理 4.2.3. 条件 4.1.1 が満たされているとする. このとき, 以下が成り立つ.
(1) $\widehat{x}_0 = x_0$ に対して, 関係式 (R_0) が成り立つ.
(2) $k \geq 0$ に対して (R_k) が成り立つとする. このとき, $x_{k+1}, \widehat{x}_{k+1} \in Q$ を

$$\begin{aligned} x_{k+1} &= z_k \\ \widehat{x}_{k+1} &= \frac{S_k\widehat{x}_k + \lambda_{k+1}x_{k+1}}{S_{k+1}} \end{aligned} \tag{4.2.3}$$

と定めると, (R_{k+1}) が成り立つ.
もしくは,
(3) $k \geq 0$ に対して (R_k) が成り立つとする. このとき, $x_{k+1}, \widehat{x}_{k+1} \in Q$ を

$$\widehat{x}_{k+1} = x_{k+1} = \frac{S_k\widehat{x}_k + \lambda_{k+1}z_k}{S_{k+1}} \tag{4.2.4}$$

と定めると, (R_{k+1}) が成り立つ.

証明. (1) 条件 4.1.1 の (ii) において $k = -1$ とすれば,

$$\begin{aligned} \min_{x\in Q}\widehat{\psi}_0(x) + \frac{\lambda_0^2}{2\sigma\beta_{-1}}\|g_0\|_*^2 &\geq \left[\min_{x\in Q}\widehat{\psi}_{-1}(x) + \lambda_0 l_0(z_0) + \beta_{-1}\xi(z_{-1},z_0)\right] + \frac{\lambda_0^2}{2\sigma\beta_{-1}}\|g_0\|_*^2 \\ &= \lambda_0 l_0(z_0) + \beta_{-1}\xi(z_{-1},z_0) + \frac{\lambda_0^2}{2\sigma\beta_{-1}}\|g_0\|_*^2 \qquad (\because \text{ 条件 4.1.1 の (i) より}) \\ &= \lambda_0 l_0(z_0) + \beta_{-1}\xi(x_0,z_0) + \frac{\lambda_0^2}{2\sigma\beta_{-1}}\|g_0\|_*^2 \\ &\geq \lambda_0 f(x_0) \qquad (\because \text{ 補題 4.2.2 より}) \\ &= S_0 f(\widehat{x}_0) \end{aligned}$$

すなわち, (R_0) が成り立つ.
(2) 条件 4.1.1 の (ii) と $x_{k+1}, \widehat{x}_{k+1}$ の定め方より,

$$\min_{x\in Q}\widehat{\psi}_{k+1}(x) + \frac{1}{2\sigma}\sum_{i=0}^{k+1}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2$$

$$
\begin{aligned}
&\geq \left[\min_{x\in Q}\widehat{\psi}_k(x)+\lambda_{k+1}l_{k+1}(z_{k+1})+\beta_k\xi(z_k,z_{k+1})\right]+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2+\frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2 \\
&=\min_{x\in Q}\widehat{\psi}_k(x)+\frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2+\left[\lambda_{k+1}l_{k+1}(z_{k+1})+\beta_k\xi(x_{k+1},z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2\right] \\
&\geq \left[\min_{x\in Q}\widehat{\psi}_k(x)+\frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2\right]+\lambda_{k+1}f(x_{k+1}) \qquad (\because \text{ 補題 4.2.2 より}) \qquad (4.2.5) \\
&\geq S_kf(\widehat{x}_k)+\lambda_{k+1}f(x_{k+1}) \qquad (\because \text{ 関係式 } (R_k) \text{ より}) \\
&\geq S_{k+1}f(\frac{S_k\widehat{x}_k+\lambda_{k+1}x_{k+1}}{S_{k+1}}) \quad (\because\ f \text{ の凸性より}) \\
&= S_{k+1}f(\widehat{x}_{k+1})
\end{aligned}
$$

したがって (R_{k+1}) が成り立つ.

(3) $x'_{k+1}=\dfrac{S_k\widehat{x}_k+\lambda_{k+1}z_{k+1}}{S_{k+1}}$ とおく. $x_{k+1}=\dfrac{S_k\widehat{x}_k+\lambda_{k+1}z_k}{S_{k+1}}$ なので,

$$
z_{k+1}-z_k=\frac{S_{k+1}}{\lambda_{k+1}}(x'_{k+1}-x_{k+1})
$$

であることに注意すると, 条件 4.1.1 の (ii) と関係式 (R_k) より,

$$
\begin{aligned}
&\min_{x\in Q}\widehat{\psi}_{k+1}(x)+\frac{1}{2\sigma}\sum_{i=0}^{k+1}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2 \\
&\qquad\geq \min_{x\in Q}\widehat{\psi}_k(x)+\lambda_{k+1}l_{k+1}(z_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2+\frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2 \\
&\qquad\geq S_kf(\widehat{x}_k)+\lambda_{k+1}l_{k+1}(z_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2 \\
&\qquad\geq S_kl_{k+1}(\widehat{x}_k)+\lambda_{k+1}l_{k+1}(z_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2 \\
&\qquad= S_{k+1}l_{k+1}(\frac{S_k\widehat{x}_k+\lambda_{k+1}z_{k+1}}{S_{k+1}})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2 \\
&\qquad= S_{k+1}l_{k+1}(x'_{k+1})+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2 \\
&\qquad= S_{k+1}f(x_{k+1})+\left\langle g_{k+1},S_{k+1}(x'_{k+1}-x_{k+1})\right\rangle \\
&\qquad\qquad\qquad +\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2 \\
&\qquad= S_{k+1}f(x_{k+1})+\langle\lambda_{k+1}g_{k+1},z_{k+1}-z_k\rangle+\beta_k\xi(z_k,z_{k+1})+\frac{\lambda_{k+1}^2}{2\sigma\beta_k}\|g_{k+1}\|_*^2 \\
&\qquad\geq S_{k+1}f(x_{k+1}) \qquad (\because \text{ 補題 4.2.2 より}) \\
&\qquad= S_{k+1}f(\widehat{x}_{k+1})
\end{aligned}
$$

ゆえに (R_{k+1}) が成り立つ. □

この定理から, 次の二通りのアルゴリズムが考えられる.

(1) $\widehat{x}_0 = x_0$ とし, $\{\widehat{x}_k\}, \{x_k\}$ を更新式 (4.2.3) により生成する. このとき, $\widehat{x}_k, x_k$ は以下のようにあらわすことができる.

$$x_k = z_{k-1}, \qquad \widehat{x}_k = \frac{1}{S_k}\sum_{i=0}^{k}\lambda_i x_i \tag{4.2.6}$$

(2) $\widehat{x}_0 = x_0$ とし, $\{\widehat{x}_k\}, \{x_k\}$ を更新式 (4.2.4) により生成する. このとき, $\widehat{x}_k, x_k$ は以下のようにあらわすことができる.

$$\widehat{x}_k = x_k = \frac{1}{S_k}\sum_{i=0}^{k}\lambda_i z_{i-1} \tag{4.2.7}$$

更新式 (4.2.3) に対しては, 次に述べるようなより強い性質を示すことができる. そのために, 以下のような関係式 $(\widehat{R}_k)$ を考える.

$$(\widehat{R}_k) \qquad \sum_{i=0}^{k}\lambda_i f(x_i) \le \min_{x\in Q}\widehat{\psi}_k(x) + \frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2 \tag{4.2.8}$$

$(\widehat{R}_0) \Longleftrightarrow (R_0)$ である. また, 点列 $\{\widehat{x}_k\}$ が $\widehat{x}_k = \dfrac{1}{S_k}\displaystyle\sum_{i=0}^{k}\lambda_i x_i$ でつくられるときは, Jensen の不等式より $S_k f(\widehat{x}_k) \le \displaystyle\sum_{i=0}^{k}\lambda_i f(x_i)$ であるから, $(\widehat{R}_k) \Longrightarrow (R_k)$ が成り立つ.

定理 4.2.4. 条件 4.1.1 が満たされているとする.
$k \ge 0$ に対して, 関係式 $(\widehat{R}_k)$ が成り立つとする. このとき, $x_{k+1}, \widehat{x}_{k+1} \in Q$ を

$$\begin{aligned} x_{k+1} &= z_k \\ \widehat{x}_{k+1} &= \frac{S_k\widehat{x}_k + \lambda_{k+1}x_{k+1}}{S_{k+1}} \end{aligned}$$

と定めると, $(\widehat{R}_{k+1})$ が成り立つ.

証明. 式 (4.2.5) のあとで (R_k) でなく $(\widehat{R}_k)$ を使えばよい. すなわち,

$$\begin{aligned} &\min_{x\in Q}\widehat{\psi}_{k+1}(x) + \frac{1}{2\sigma}\sum_{i=0}^{k+1}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2 \\ &\qquad \ge \left[\min_{x\in Q}\widehat{\psi}_k(x) + \frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_{i-1}}\|g_i\|_*^2\right] + \lambda_{k+1}f(x_{k+1}) \qquad (\because\ \text{式 (4.2.5) より}) \\ &\qquad \ge \sum_{i=0}^{k}\lambda_i f(x_i) + \lambda_{k+1}f(x_{k+1}) \qquad (\because\ \text{関係式 } (\widehat{R}_k) \text{ より}) \\ &\qquad = \sum_{i=0}^{k+1}\lambda_i f(x_i) \end{aligned}$$

□

(4.2.3) および (4.2.4) で与えられるアルゴリズムを以下にまとめる.

アルゴリズム 4.2.5. 条件 4.1.1 が満たされる補助関数 $\psi_k(x)$, ステップサイズ λ_k のもとで以下をおこなう.

Step 0. $\widehat{x}_0 = x_0 = z_{-1} = \operatorname*{argmin}_{x\in Q} d(x)$

Step 1. $k = 0, 1, 2, \ldots$ に対して, 以下をおこなう.

(1) 更新式 (4.2.3) にもとづくアルゴリズム	(2) 更新式 (4.2.4) にもとづくアルゴリズム
$x_{k+1} = z_k = \operatorname*{argmin}_{x\in Q} \psi_k(x)$ $\widehat{x}_{k+1} = \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i x_i$	$z_k = \operatorname*{argmin}_{x\in Q} \psi_k(x)$ $\widehat{x}_{k+1} = x_{k+1} = \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_{i-1}$

定理 4.2.3 より, 以下の定理が導かれる.

定理 4.2.6. $x^* \in Q$ を問題 (P) $\min_{x\in Q} f(x)$ に対する最適解とする. アルゴリズム 4.2.5 は以下を満たす.

(i) アルゴリズム (1) は任意の $k \geq 0$ に対して, (R_k) および $(\widehat{R}_k)$ を満たす. さらに次の不等式が成り立つ.

$$\forall k \geq 0, \quad \max\left\{ f(\widehat{x}_k) - f(x^*),\ \min_{0\leq i\leq k} f(x_i) - f(x^*) \right\} \leq \frac{\beta_k d_k(x^*) + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2}{S_k}$$

(ii) アルゴリズム (2) は任意の $k \geq 0$ に対して, (R_k) を満たす. さらに次の不等式が成り立つ.

$$\forall k \geq 0, \quad f(\widehat{x}_k) - f(x^*) \leq \frac{\beta_k d_k(x^*) + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2}{S_k}$$

証明. (i) アルゴリズム (1) の更新式は (4.2.3) を満たすので, 定理 4.2.3 および 定理 4.2.4 より, 関係式 (R_k) および $(\widehat{R}_k)$ が成り立つ. したがって, 命題 4.1.2 から, 求める不等式を得る. (ii) も (i) と同様. □

4.2.2 パラメータの選択

この定理 4.2.6 は, アルゴリズム 4.2.5 における目的関数値の最適値との残差の上からの見積もりを与えている. この上界を δ_k とおこう. すなわち,

$$\delta_k = \frac{\beta_k d_k(x^*) + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2}{S_k} \tag{4.2.9}$$

とおく. この上界に登場するパラメータ $\{\lambda_k\}$, $\{\beta_k\}$ の選び方については, Nesterov [9] の提案したもの (定理 3.3.4) が有効である. 式 (3.3.2) で定まる数列 $\{\widehat{\beta}_k\}$ を用いて次の定理を示そう.

定理 4.2.7. 上記の δ_k について以下が成り立つ.
(1) $\lambda_k \equiv 1,\ \ \beta_k = \gamma\widehat{\beta}_k\ \ (\gamma > 0)$ とするとき, $\sup_{k\geq 0}\|g_k\|_* \leq M$ であれば, δ_k は次を満たす.

$$\forall k \geq 0, \quad \delta_k \leq \left(\gamma d_k(x^*) + \frac{M^2}{2\sigma\gamma}\right)\frac{0.5+\sqrt{2k+1}}{k+1}$$

(2) $\lambda_k = \dfrac{1}{\|g_k\|_*},\ \ \beta_k = \dfrac{\widehat{\beta}_k}{\rho\sqrt{\sigma}}\ \ (\rho > 0)$ とするとき, δ_k は次を満たす.

$$\forall k \geq 0, \quad \delta_k \leq \max_{0\leq i\leq k}\|g_i\|_* \frac{1}{\sqrt{\sigma}}\left(\frac{d_k(x^*)}{\rho} + \frac{\rho}{2}\right)\frac{0.5+\sqrt{2k+1}}{k+1}$$

証明. (1) $\lambda_k \equiv 1,\ \ \beta_k = \gamma\widehat{\beta}_k$ を δ_k に代入して,

$$\begin{aligned}
\delta_k &= \frac{\gamma\widehat{\beta}_k d_k(x^*) + \dfrac{1}{2\sigma}\displaystyle\sum_{i=0}^{k}\frac{1}{\gamma\widehat{\beta}_{i-1}}\|g_i\|_*^2}{k+1} \\
&\leq \frac{\gamma\widehat{\beta}_k d_k(x^*) + \dfrac{M^2}{2\gamma\sigma}\displaystyle\sum_{i=0}^{k}\frac{1}{\widehat{\beta}_{i-1}}}{k+1} \\
&= \left(\gamma d_k(x^*) + \frac{M^2}{2\gamma\sigma}\right)\frac{\widehat{\beta}_k}{k+1} \quad (\because\ \text{式 (3.3.3) より})
\end{aligned}$$

が得られる. したがって, 補題 3.3.3 より求める不等式が示される.
(2) $\lambda_k = 1/\|g_k\|_*,\ \ \beta_k = \widehat{\beta}_k/(\rho\sqrt{\sigma})$ を δ_k に代入して,

$$\begin{aligned}
\delta_k &= \frac{\dfrac{\widehat{\beta}_k}{\rho\sqrt{\sigma}}d_k(x^*) + \dfrac{1}{2\sigma}\displaystyle\sum_{i=0}^{k}\frac{\rho\sqrt{\sigma}}{\widehat{\beta}_{i-1}}\frac{1}{\|g_i\|_*^2}\|g_i\|_*^2}{\displaystyle\sum_{i=0}^{k}1/\|g_i\|_*} \\
&\leq \frac{1}{\sqrt{\sigma}}\frac{\dfrac{d_k(x^*)}{\rho}\widehat{\beta}_k + \dfrac{\rho}{2}\displaystyle\sum_{i=0}^{k}\frac{1}{\widehat{\beta}_{i-1}}}{\displaystyle\min_{0\leq i\leq k}(1/\|g_i\|_*)\times(k+1)} \\
&= \max_{0\leq i\leq k}\|g_i\|_*\frac{1}{\sqrt{\sigma}}\left(\frac{d_k(x^*)}{\rho} + \frac{\rho}{2}\right)\frac{\widehat{\beta}_k}{k+1} \quad (\because\ \text{式 (3.3.3) より})
\end{aligned}$$

が得られる. したがって, 補題 3.3.3 より求める不等式が示される. □

注意 . われわれは, 上記の定理において $\sup_{k\geq 0}\|g_k\|_* \leq M$ であることを仮定した. この仮定が満たされる十分条件を二つ挙げよう.

- $Q \subset \operatorname{int}(\operatorname{dom} f)$ かつ $f(x)$ が $\operatorname{int}(\operatorname{dom} f)$ 上で定数 M に対してリプシッツ連続であるとき.
- $\{x_k\} \subset \operatorname{int} Q$ かつ $f(x)$ が $\operatorname{int} Q$ 上で定数 M に対してリプシッツ連続であるとき.

これらが十分条件となることは, 定理 1.3.2 からわかる.

4.2.3 Dual-averaging の場合のアルゴリズム

アルゴリズム 4.2.5 において, DA の場合 (すなわち命題 4.1.3 の場合) を当てはめよう. すると, $\widehat{x}_0 = x_0 = z_{-1} = \operatorname*{argmin}_{x \in Q} d(x)$ を初期点とする次の更新法が得られる.

$$\text{Dual-averaging (1)} : \begin{cases} x_{k+1} &= z_k = \operatorname*{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle] + \beta_k d(x) \right\} \\ \widehat{x}_{k+1} &= \dfrac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i x_i \end{cases} \tag{4.2.10}$$

$$\text{Dual-averaging (2)} : \begin{cases} z_k &= \operatorname*{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle] + \beta_k d(x) \right\} \\ \widehat{x}_{k+1} &= x_{k+1} = \dfrac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_{i-1} \end{cases} \tag{4.2.11}$$

これらのアルゴリズムに対して定理 4.2.6 が成り立ち, 式 (4.2.9) で定義した上界 δ_k は,

$$\delta_k = \frac{\beta_k d(x^*) + \dfrac{1}{2\sigma} \sum_{i=0}^{k} \dfrac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2}{S_k} \tag{4.2.12}$$

とかける. アルゴリズム (1) はもともとの DA アルゴリズム 3.3.1 と一致しており, 上界 δ_k も同じである.

4.2.4 Mirror-descent の場合のアルゴリズム

アルゴリズム 4.2.5 において, MD の場合 (すなわち命題 4.1.4 の場合) を当てはめ, DA との比較をしてみよう. MD の場合, $\widehat{x}_0 = x_0 = z_{-1} = \operatorname*{argmin}_{x \in Q} d(x)$ を初期点として次の更新法が得られる.

$$\text{Mirror-descent (1)} : \begin{cases} x_{k+1} &= z_k = \operatorname*{argmin}_{x \in Q} \Big\{ \lambda_k [f(x_k) + \langle g_k, x - x_k \rangle] \\ & \qquad + \beta_k d(x) - \beta_{k-1} [d(z_{k-1}) + \langle \nabla d(z_{k-1}), x - z_{k-1} \rangle] \Big\} \\ \widehat{x}_{k+1} &= \dfrac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i x_i \end{cases} \tag{4.2.13}$$

$$\text{Mirror-descent (2)} : \begin{cases} z_k &= \operatorname*{argmin}_{x \in Q} \Big\{ \lambda_k [f(x_k) + \langle g_k, x - x_k \rangle] \\ & \qquad + \beta_k d(x) - \beta_{k-1} [d(z_{k-1}) + \langle \nabla d(z_{k-1}), x - z_{k-1} \rangle] \Big\} \\ \widehat{x}_{k+1} &= x_{k+1} = \dfrac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_{i-1} \end{cases} \tag{4.2.14}$$

とくに, アルゴリズム (1) は $x_k = z_{k-1}$, $k \geq 0$ を使うと, 点列 $\{z_k\}_{k\geq -1}$ を取り除いた次の更新法に書き直せる.

$$\text{Mirror-descent (1)} : \begin{cases} x_{k+1} = \underset{x\in Q}{\operatorname{argmin}} \Big\{ \lambda_k[f(x_k) + \langle g_k, x - x_k\rangle] \\ \qquad\qquad + \beta_k d(x) - \beta_{k-1}[d(x_k) + \langle \nabla d(x_k), x - x_k\rangle] \Big\} \\ \widehat{x}_{k+1} = \dfrac{1}{S_{k+1}} \displaystyle\sum_{i=0}^{k+1} \lambda_i x_i \end{cases} \tag{4.2.15}$$

これらのアルゴリズムに対して定理 4.2.6 が成り立ち, 式 (4.2.9) で定義した上界 δ_k は,

$$\delta_k = \frac{\beta_k[d(z_k) + \langle \nabla d(z_k), x^* - z_k\rangle] + \dfrac{1}{2\sigma}\displaystyle\sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_{i-1}} \|g_i\|_*^2}{S_k} \tag{4.2.16}$$

とかける. したがって, 定理 4.2.7 によるパラメータの選び方を用いて, 任意の $k \geq 0$ に対して $\delta_k \leq O(1/\sqrt{k})$ を保証することができる. もともとの MD アルゴリズム 3.2.1 に対する収束率の見積もりは $O(\log k/\sqrt{k})$ が限界であった.

この上界と DA の上界 (4.2.12) を比較すると, $d(x)$ の凸性から, 不等式

$$d(z_k) + \langle \nabla d(z_k), x^* - z_k\rangle \leq d(x^*)$$

が成り立つので,

$$\Big(\text{MD に対する } \delta_k\Big) \leq \Big(\text{DA に対する } \delta_k\Big)$$

がわかる.

アルゴリズム (1) において, とくに $\beta_k \equiv 1$ とすると,

$$\text{MD (1)}, \ \beta_k \equiv 1 : \begin{cases} x_{k+1} = \underset{x\in Q}{\operatorname{argmin}} \Big\{ \lambda_k[f(x_k) + \langle g_k, x - x_k\rangle] + \xi(x_k, x) \Big\} \\ \widehat{x}_{k+1} = \dfrac{1}{S_{k+1}} \displaystyle\sum_{i=0}^{k+1} \lambda_i x_i \end{cases} \tag{4.2.17}$$

とかける. これは, もともとの MD アルゴリズム 3.2.1 と同じ更新法である.

4.2.5　計算量

上記の MD や DA に対する計算量の見積もりを与えよう. $d_k(x)$ に対する仮定より, $d_k(x^*) \leq d(x^*)$ であるから, 定理 4.2.7 で得られた上界はさらに次のように上からおさえることができる.

$$(1) : \delta_k \leq \left(\gamma d(x^*) + \frac{M^2}{2\sigma\gamma}\right) \frac{0.5 + \sqrt{2k+1}}{k+1}$$

$$(2) : \delta_k \leq \max_{0\leq i\leq k} \|g_i\|_* \frac{1}{\sqrt{\sigma}} \left(\frac{d(x^*)}{\rho} + \frac{\rho}{2}\right) \frac{0.5 + \sqrt{2k+1}}{k+1}$$

この上界は (1) では $\gamma = M/\sqrt{2\sigma d(x^*)}$, (2) では $\rho = \sqrt{2d(x^*)}$ によって最適となる. $\sup_{k\geq 0} \|g_k\|_* \leq M$ ならば, 3.3 節の最後に述べたように

$$\forall k \geq 0, \quad \delta_k \leq \sqrt{\frac{5d(x^*)}{\sigma}} \frac{M}{\sqrt{k+1}}$$

が満たされる．したがって, 精度 $\varepsilon > 0$ の近似解 $\widehat{x}_k$ を高々 $\left\lceil \frac{5M^2 d(x^*)}{\sigma\varepsilon^2} \right\rceil - 1$ の反復回数で得ることができる．また, 劣勾配 $g_k \in \partial f(x_k)$ を得るための計算量の上界を G_f, 勾配 $\nabla d(x)$ を得るための計算量を G_d とし, さらに補助問題 $\min_{x\in Q}\{\langle s, x\rangle + d(x)\}$ に要する計算量の上界を A とすれば, 一反復の計算量の上界は次のように与えられる．

$$\mathrm{MD} : G_f + G_d + A, \qquad \mathrm{DA} : G_f + A$$

したがって, 精度 ε の近似解を得るために十分な計算量として以下が得られる．

$$\begin{aligned} \mathrm{MD} \quad &: \quad A + (G_f + G_d + A) \times \left(\left\lceil \frac{5M^2 d(x^*)}{\sigma\varepsilon^2} \right\rceil - 1 \right) \\ \mathrm{DA} \quad &: \quad A + (G_f + A) \times \left(\left\lceil \frac{5M^2 d(x^*)}{\sigma\varepsilon^2} \right\rceil - 1 \right) \end{aligned}$$

4.3 Extremal convex problem に対するアルゴリズム

この節では, 4.1 節の問題 (P) において, (c) の場合を考える．すなわち, 次の凸計画問題を対象とする．

$$\begin{array}{lll} (P) & \min & f(x) \\ & \text{s.t.} & x \in Q \end{array}$$

ここで $Q \subset \mathrm{dom}\, f$ は閉凸集合であり, 目的関数が以下のように与えられるとする．

$$f(x) = \max_{u\in U} \left\{ \sum_{j=1}^{m} u^{(j)} f_j(x) - \phi(u) \right\}$$

ただし, $\phi : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ は連続凸関数, $U \subset \mathrm{dom}\, \phi$ は有界閉集合, $f_j : E \to \mathbb{R} \cup \{+\infty\}$ は Q 上連続的微分可能な凸関数であり, 仮定 2.2.1 が成り立つとする．

このとき, 点列 $\{x_k\}$ に対して, 目的関数の近似モデル $l_k(x)$ は式 (4.1.1) より,

$$l_k(x) = f(x_k, x) = \max_{u\in U} \left\{ \sum_{j=1}^{m} u^{(j)} [f_j(x_k) + \langle \nabla f_j(x_k), x - x_k\rangle] - \phi(u) \right\} \tag{4.3.1}$$

である．われわれはアルゴリズムの解析をおこなうために, 各 $\nabla f_j(x)$ は Q 上で定数 L_j に対してリプシッツ連続であるとする．このとき, $L = \max_{u\in U} \sum_{j=1}^{m} u^{(j)} L_j$ とおくと, 命題 2.2.5 より以下が成り立つ．

$$f(x) \le l_k(x) + \frac{L}{2}\|x - x_k\|^2, \quad x \in Q \tag{4.3.2}$$

われわれは, (4.1.3) より次の関係式 (R_k) を満たす近似解 $\{\widehat{x}_k\}$ を生成するアルゴリズムを目指す：

$$(R_k) \quad S_k f(\widehat{x}_k) \le \min_{x\in Q} \widehat{\psi}_k(x)$$

ここでも, Q 上の最小点が計算可能な補助関数 $\psi_k(x)$ とその最小点 $z_k \in Q$ が, ステップサイズ $\lambda_k > 0$, スケーリングパラメータ $\beta_k > 0$, 評価関数 $\widehat{\psi}_k(x)$ および Q 上で $d(x)$ を超えない関数 $d_k(x)$ に対して, 条件 4.1.1 を満たすとする．

4.3.1　アルゴリズムの導出

定理 4.3.1. 条件 4.1.1 が満たされているとする. このとき, 以下が成り立つ.

(1) $\widehat{x}_0 = z_0$, $\dfrac{\sigma\beta_{-1}}{\lambda_0} \geq L$ ならば (R_0) が成り立つ.

(2) $k \geq 0$ に対して (R_k) が成り立つとする. このとき, $x_{k+1}, \widehat{x}_{k+1} \in Q$ を

$$\begin{aligned} x_{k+1} &= \frac{S_k \widehat{x}_k + \lambda_{k+1} z_k}{S_{k+1}} \\ \widehat{x}_{k+1} &= \frac{S_k \widehat{x}_k + \lambda_{k+1} z_{k+1}}{S_{k+1}} \end{aligned} \tag{4.3.3}$$

と定めるとき, $\dfrac{\sigma\beta_k S_{k+1}}{\lambda_{k+1}^2} \geq L$ ならば (R_{k+1}) が成り立つ.

もしくは,

(3) $k \geq 0$ に対して (R_k) が成り立つとする. このとき, $x_{k+1}, \widehat{x}_{k+1} \in Q$ を

$$\begin{aligned} x_{k+1} &= z_k \\ \widehat{x}_{k+1} &= \frac{S_k \widehat{x}_k + \lambda_{k+1} z_{k+1}}{S_{k+1}} \end{aligned} \tag{4.3.4}$$

と定めるとき, $\dfrac{\sigma\beta_k}{\lambda_{k+1}} \geq L$ ならば (R_{k+1}) が成り立つ.

証明. (1) 条件 4.1.1 の (ii) において $k = -1$ とすれば,

$$\begin{aligned} \min_{x \in Q} \widehat{\psi}_0(x) &\geq \min_{x \in Q} \widehat{\psi}_{-1}(x) + \lambda_0 l_0(z_0) + \beta_{-1} \xi(z_{-1}, z_0) \\ &= \lambda_0 \left(l_0(z_0) + \frac{\beta_{-1}}{\lambda_0} \xi(z_{-1}, z_0) \right) \quad (\because \text{条件 4.1.1 の (i) より}) \\ &= \lambda_0 \left(l_0(z_0) + \frac{\beta_{-1}}{\lambda_0} \xi(x_0, z_0) \right) \\ &\geq \lambda_0 \left(l_0(z_0) + \frac{\sigma\beta_{-1}}{\lambda_0} \frac{1}{2} \|z_0 - x_0\|^2 \right) \\ &\geq \lambda_0 f(z_0) \quad (\because \text{不等式 (4.3.2) より}) \\ &= S_0 f(\widehat{x}_0) \end{aligned}$$

すなわち (R_0) が成り立つ.

(2) x_{k+1}, $\widehat{x}_{k+1}$ の定め方より,

$$z_{k+1} - z_k = \frac{S_{k+1}}{\lambda_{k+1}} (\widehat{x}_{k+1} - x_{k+1})$$

であることに注意すると, 条件 4.1.1 の (ii) と関係式 (R_k) より,

$$
\begin{aligned}
\min_{x \in Q} \widehat{\psi}_{k+1}(x) &\geq \min_{x \in Q} \widehat{\psi}_k(x) + \lambda_{k+1} l_{k+1}(z_{k+1}) + \beta_k \xi(z_k, z_{k+1}) \\
&\geq S_k f(\widehat{x}_k) + \lambda_{k+1} l_{k+1}(z_{k+1}) + \beta_k \xi(z_k, z_{k+1}) \\
&\geq S_k l(\widehat{x}_k) + \lambda_{k+1} l_{k+1}(z_{k+1}) + \beta_k \xi(z_k, z_{k+1}) \\
&\geq S_{k+1} l_{k+1}(\frac{S_k \widehat{x}_k + \lambda_{k+1} z_{k+1}}{S_{k+1}}) + \beta_k \xi(z_k, z_{k+1}) \quad (\because\ l_k(x) \text{ の凸性}) \\
&\geq S_{k+1} l_{k+1}(\widehat{x}_{k+1}) + \frac{\sigma \beta_k}{2} \|z_{k+1} - z_k\|^2 \\
&= S_{k+1} \left(l_{k+1}(\widehat{x}_{k+1}) + \frac{\sigma \beta_k S_{k+1}}{\lambda_{k+1}^2} \frac{1}{2} \|\widehat{x}_{k+1} - x_{k+1}\|^2 \right) \\
&\geq S_{k+1} f(\widehat{x}_{k+1}) \quad (\because \text{不等式 (4.3.2) より})
\end{aligned}
$$

したがって (R_{k+1}) が成り立つ.

(3) 条件 4.1.1 の (ii) と $\widehat{x}_{k+1}, x_{k+1}$ の定め方より,

$$
\begin{aligned}
\min_{x \in Q} \widehat{\psi}_{k+1}(x) &\geq \min_{x \in Q} \widehat{\psi}_k(x) + \lambda_{k+1} l_{k+1}(z_{k+1}) + \beta_k \xi(z_k, z_{k+1}) \\
&= \min_{x \in Q} \widehat{\psi}_k(x) + +\lambda_{k+1} l_{k+1}(z_{k+1}) + \beta_k \xi(x_{k+1}, z_{k+1}) \\
&\geq \min_{x \in Q} \widehat{\psi}_k(x) + +\lambda_{k+1} \left(l_{k+1}(z_{k+1}) + \frac{\sigma \beta_k}{\lambda_{k+1}} \frac{1}{2} \|z_{k+1} - x_{k+1}\|^2 \right) \\
&\geq \min_{x \in Q} \widehat{\psi}_k(x) + +\lambda_{k+1} f(z_{k+1}) \quad (\because\ \text{不等式 (4.3.2) より}) \qquad (4.3.5) \\
&\geq S_k f(\widehat{x}_k) + \lambda_{k+1} f(z_{k+1}) \quad (\because\ \text{関係式 } (R_k) \text{ より}) \\
&\geq S_{k+1} f(\widehat{x}_{k+1}) \quad (\because\ f(x) \text{ の凸性と } \widehat{x}_{k+1} \text{ の定義より})
\end{aligned}
$$

ゆえに (R_{k+1}) が成り立つ. □

この定理により, 次のアルゴリズムを得る.

アルゴリズム **4.3.2.** 条件 4.1.1 が満たされる補助関数 $\psi_k(x)$, ステップサイズ λ_k のもとで以下をおこなう.

Step 0. $x_0 = z_{-1} = \operatorname*{argmin}_{x \in Q} d(x),\ \widehat{x}_0 = z_0 = \operatorname*{argmin}_{x \in Q} \psi_0(x)$

Step 1. $k = 0, 1, 2, \ldots$ に対して, 以下をおこなう.

(1) 更新式 (4.3.3) にもとづくアルゴリズム

$$
\begin{aligned}
x_{k+1} &= \frac{\sum_{i=0}^{k} \lambda_i z_i + \lambda_{k+1} z_k}{S_{k+1}} \\
z_{k+1} &= \operatorname*{argmin}_{x \in Q} \psi_{k+1}(x) \\
\widehat{x}_{k+1} &= \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_i
\end{aligned}
$$

(2) 更新式 (4.3.4) にもとづくアルゴリズム

$$
\begin{aligned}
x_{k+1} &= z_k \\
z_{k+1} &= \operatorname*{argmin}_{x \in Q} \psi_{k+1}(x) \\
\widehat{x}_{k+1} &= \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_i
\end{aligned}
$$

定理 4.3.3. $x^* \in Q$ を extremal convex problem (P) に対する最適解とする．アルゴリズム 4.3.2 は以下を満たす．

(i) $\{\lambda_k\}$, $\{\beta_k\}$ が $\dfrac{\sigma\beta_{k-1}S_k}{\lambda_k^2} \geq L$ $(k \geq 0)$ を満たすならば, アルゴリズム (1) は任意の $k \geq 0$ に対して (R_k) を満たす．さらに次の不等式が成り立つ．

$$\forall k \geq 0, \quad f(\widehat{x}_k) - f(x^*) \leq \frac{\beta_k d_k(x^*)}{S_k}$$

(ii) $\{\lambda_k\}$, $\{\beta_k\}$ が $\dfrac{\sigma\beta_{k-1}}{\lambda_k} \geq L$ $(k \geq 0)$ を満たすならば, アルゴリズム (2) は任意の $k \geq 0$ に対して (R_k) を満たす．さらに次の不等式が成り立つ．

$$\forall k \geq 0, \quad f(\widehat{x}_k) - f(x^*) \leq \frac{\beta_k d_k(x^*)}{S_k}$$

証明. アルゴリズム (1), (2) はそれぞれ更新式 (4.3.3), (4.3.4) を満たすので, 定理 4.3.1 より, (R_k) が成り立つ．したがって, 命題 4.1.2 より, 求める不等式を得る．□

4.3.2　パラメータの選択

定理 4.3.3 から得られる目的関数値と最適値との残差に対する上界を δ_k とおこう．すなわち,

$$\delta_k = \frac{\beta_k d_k(x^*)}{S_k}$$

とおく．δ_k の 0 への収束をなるべく速くするようなパラメータ $\{\beta_k\}$, $\{\lambda_k\}$ の選び方を考察したい．ただし, アルゴリズム 4.3.2 の仮定より, L が次の上界をもつようにしなくてはならない．

$$\text{アルゴリズム (1)}: \frac{\sigma\beta_{k-1}S_k}{\lambda_k^2} \geq L, \quad k \geq 0 \tag{4.3.6}$$

$$\text{アルゴリズム (2)}: \frac{\sigma\beta_{k-1}}{\lambda_k} \geq L, \quad k \geq 0 \tag{4.3.7}$$

われわれは, 次のパラメータの選び方のなかで効率的なものをさがすことにする[1]．

$$\lambda_k = (k+1)^p \ (k \geq 0), \quad \beta_k = \beta_{-1}(k+2)^q \ (k \geq -1)$$

ただし, 下の補題を用いるために $p > -1$ とする．また, β_k は非減少であるから $q \geq 0$ とする．

補題 4.3.4. $p > -1$ に対して, 以下が成り立つ．

$$\min\left\{1, \frac{1}{1+p}\right\}(k+1)^{1+p} \leq \sum_{i=0}^{k}(i+1)^p \leq \max\left\{1, \frac{1}{1+p}\right\}(k+1)^{1+p}$$

[1] パラメータ (λ_k, β_{k-1}) は同時に正の定数倍をしても, 解析に影響がないので, $\lambda_0 = 1$ と固定して議論をおこなう．

証明. 場合分けをする. $-1 < p \leq 0$ のとき, 示すべき不等式は $(k+1)^{1+p} \leq \sum_{i=0}^{k}(i+1)^p \leq \frac{1}{1+p}(k+1)^{1+p}$ である. $\mathbb{R}_{>0} \ni x \mapsto x^p$ が単調減少することから,

$$\sum_{i=0}^{k}(i+1)^p \leq \sum_{i=0}^{k}\int_i^{i+1} x^p dx = \int_0^{k+1} x^p dx = \frac{1}{1+p}(k+1)^{1+p}$$

を得る. また, $a \in (0,1)$ に対して, $x \mapsto a^x$ が単調減少することから,

$$\sum_{i=0}^{k}(i+1)^p = (k+1)^p \sum_{i=0}^{k}\left(\frac{i+1}{k+1}\right)^p \geq (k+1)^p \sum_{i=0}^{k}\left(\frac{i+1}{k+1}\right)^0 = (k+1)^{1+p}$$

である. したがって求める不等式を得た. $p \geq 0$ のときは, 不等号を逆にして, $\frac{1}{1+p}(k+1)^{1+p} \leq \sum_{i=0}^{k}(i+1)^p \leq (k+1)^{1+p}$ が示される. ゆえに, 補題が証明された. □

この補題を用いると, $\lambda_k = (k+1)^p,\ p > -1$ と選ぶとき, $S_k = \sum_{i=0}^{k}\lambda_i$ は次の不等式を満たす.

$$\min\left\{1, \frac{1}{1+p}\right\}(k+1)^{1+p} \leq S_k \leq \max\left\{1, \frac{1}{1+p}\right\}(k+1)^{1+p}$$

ここで $\beta_k = \beta_{-1}(k+2)^q,\ q \geq 0$ と合わせると,

$$\delta_k = \frac{\beta_k d_k(x^*)}{S_k} \leq \beta_{-1} d_k(x^*) \max\{1, 1+p\} \frac{(k+2)^q}{(k+1)^{1+p}} = O(k^{q-p-1}) \tag{4.3.8}$$

まず, アルゴリズム (1) の場合について考える. 条件 (4.3.6) に与えられた L の上界は, 上の補題により次の不等式を満たす.

$$\sigma\beta_{-1}\min\left\{1, \frac{1}{1+p}\right\}(k+1)^{q-p+1} \leq \frac{\sigma\beta_{k-1}S_k}{\lambda_k^2} \leq \sigma\beta_{-1}\max\left\{1, \frac{1}{1+p}\right\}(k+1)^{q-p+1} = O(k^{q-p+1})$$

$L > 0$ であるとき, 条件 (4.3.6) が任意の $k \geq 0$ に対して成り立つためには, 上の不等式で右辺が 0 に収束しないこと, すなわち $q - p + 1 \geq 0$ が必要である. このもとで $\delta_k \leq O(k^{q-p-1})$ のオーダーを最小化すると, $q - p + 1 = 0$ ($\Leftrightarrow p = q + 1$) のときに最小オーダー $\delta_k \leq O(k^{-2})$ を達成する. このとき上の不等式は,

$$\frac{\sigma\beta_{-1}}{1+p} \leq \frac{\sigma\beta_{k-1}S_k}{\lambda_k^2} \leq \sigma\beta_{-1}$$

がとなる. したがって, 条件 (4.3.6) を満たす十分条件として $L = \sigma\beta_{-1}/(1+p) \Leftrightarrow \beta_{-1} = (1+p)L/\sigma$ を得る. このとき, δ_k の見積もり (4.3.8) は

$$\delta_k \leq \frac{(1+p)^2 L d_k(x^*)}{\sigma}\frac{(k+2)^{p-1}}{(k+1)^{p+1}}$$

となる. $p = q + 1, q \geq 0$ であったから, この見積もりは $p = 1,\ q = 0$ のときに最小となり,

$$\delta_k \leq \frac{4Ld_k(x^*)}{\sigma(k+1)^2}$$

を得る.

次にアルゴリズム (2) の場合について考える．条件 (4.3.7) は,

$$L \le \frac{\sigma\beta_{k-1}}{\lambda_k} = \sigma\beta_{-1}(k+1)^{q-p}$$

であり, これが任意の $k \ge 0$ に対して成り立つためには $q - p \ge 0$ が必要である． このもとで $\delta_k \le O(k^{q-p-1})$ のオーダーを最小化すると, $q - p = 0$ ($\Leftrightarrow p = q$) のときに最小オーダー $\delta_k \le O(k^{-1})$ を達成する．このとき, 条件 (4.3.7) は $\sigma\beta_{-1} \ge L$ となるから, $\beta_{-1} = L/\sigma$ と選ぶのがよい．すると, δ_k の見積もり (4.3.8) は,

$$\delta_k \le \frac{Ld_k(x^*)\max\{1, 1+p\}}{\sigma}\frac{(k+2)^p}{(k+1)^{1+p}}$$

となる．この見積もりは $p = q = 0$ のときに最小となり,

$$\delta_k \le \frac{Ld_k(x^*)}{\sigma(k+1)}$$

が得られる．

以上より, われわれは次の定理を得る．

定理 4.3.5. $x^* \in Q$ を extremal convex problem (P) に対する最適解とする．アルゴリズム 4.3.2 は以下を満たす．

(i) パラメータ $\lambda_k = (k+1)/2,\ \ \beta_k \equiv L/\sigma$ に対するアルゴリズム (1) が生成する近似解 $\{\widehat{x}_k\} \subset Q$ は次の不等式を満たす．

$$\forall k \ge 0,\quad f(\widehat{x}_k) - f(x^*) \le \frac{4Ld_k(x^*)}{\sigma(k+1)(k+2)}$$

(ii) パラメータ $\lambda_k \equiv 1,\ \ \beta_k \equiv L/\sigma$ に対するアルゴリズム (2) が生成する近似解 $\{\widehat{x}_k\} \subset Q$ は次の不等式を満たす．

$$\forall k \ge 0,\quad f(\widehat{x}_k) - f(x^*) \le \frac{Ld_k(x^*)}{\sigma(k+1)}$$

証明. (i) および (ii) のパラメータの取り方は, 定理 4.3.3 にパラメータ選択の条件を満たす．

(i) $S_k = (k+1)(k+2)/4$ なので,

$$f(\widehat{x}_k) - f(x^*) \le \frac{\beta_k d_k(x^*)}{S_k} = \frac{4Ld_k(x^*)}{\sigma(k+1)(k+2)}$$

(ii) $S_k = k+1$ なので,

$$f(\widehat{x}_k) - f(x^*) \le \frac{\beta_k d_k(x^*)}{S_k} = \frac{Ld_k(x^*)}{\sigma(k+1)}$$

□

4.3.3 Dual-averaging の場合のアルゴリズム

われわれはアルゴリズム 4.3.2 において二通りの更新法を得た．そのうち，アルゴリズム (1) は $f(\widehat{x}_k) - f(x^*) \le O(1/k^2)$ の収束オーダーを達成した．このアルゴリズム (1) について，DA の場合 (すなわち命題 4.1.3 の場合) を当てはめてみよう．近似モデル $l_k(x)$ の定義 (4.3.1) に注意すると，$x_0 = z_{-1} = \operatorname*{argmin}_{x \in Q} d(x)$ および

$$\widehat{x}_0 = z_0 = \operatorname*{argmin}_{x\in Q}\left\{\lambda_0 \max_{u\in U}\left\{\sum_{j=1}^m u^{(j)}[f_j(x_0) + \langle \nabla f_j(x_0), x - x_0\rangle] - \phi(u)\right\} + \beta_0 d(x)\right\}$$

を初期点として，$k \ge 0$ に対し以下のように更新をおこなう．

$$\begin{cases} x_{k+1} &= \dfrac{\sum_{i=0}^k \lambda_i z_i + \lambda_{k+1} z_k}{S_{k+1}} \\ z_{k+1} &= \operatorname*{argmin}_{x\in Q}\left\{\displaystyle\sum_{i=0}^{k+1} \lambda_i \max_{u\in U}\left\{\sum_{j=1}^m u^{(j)}[f_j(x_i) + \langle \nabla f_j(x_i), x - x_i\rangle] - \phi(u)\right\} + \beta_{k+1} d(x)\right\} \\ \widehat{x}_{k+1} &= \dfrac{1}{S_{k+1}}\displaystyle\sum_{i=0}^{k+1} \lambda_i z_i \end{cases} \tag{4.3.9}$$

ここでは，z_k を計算するのに ある $a_j^0, \dots, a_j^k \in E$ $(j = 1, \dots, m)$, $b_0, \dots, b_k \in \mathbb{R}^m$ に対する補助問題

$$\begin{array}{ll} \min & \displaystyle\sum_{i=0}^k \lambda_i \max_{u\in U}\left\{\sum_{j=1}^m u^{(j)}[\langle a_j^i, x\rangle + b_i^{(j)}] - \phi(u)\right\} + \beta_k d(x) \\ \text{s.t.} & x \in Q \end{array} \tag{4.3.10}$$

を解く必要がある．この補助問題の目的関数は，k に依存して複雑になっていくから，このアルゴリズムの実用は困難である．

パラメータ $\lambda_k = (k+1)/2$, $\beta_k \equiv L/\sigma$ によってこのアルゴリズムをおこなうとき，次の不等式が成り立つ．

$$f(\widehat{x}_k) - f(x^*) \le \frac{4Ld(x^*)}{\sigma(k+1)(k+2)} \tag{4.3.11}$$

4.3.4 Mirror-descent の場合のアルゴリズム

ここで，アルゴリズム 4.3.2 (1) について，MD の場合 (すなわち命題 4.1.4 の場合) を当てはめて，DA の場合と比べてみよう．MD の場合のアルゴリズム (1) は $x_0 = z_{-1} = \operatorname*{argmin}_{x\in Q} d(x)$ および

$$\begin{aligned} \widehat{x}_0 = z_0 = \operatorname*{argmin}_{x\in Q}\Bigg\{&\lambda_0 \max_{u\in U}\left\{\sum_{j=1}^m u^{(j)}[f_j(x_0) + \langle \nabla f_j(x_0), x - x_0\rangle] - \phi(u)\right\} \\ &+\beta_0 d(x) - \beta_{-1}[d(z_{-1}) + \langle \nabla d(z_{-1}), x - z_{-1}\rangle]\Bigg\} \end{aligned}$$

を初期点として, $k \geq 0$ に対し次のように更新をおこなう.

$$\begin{cases} x_{k+1} &= \dfrac{\sum_{i=0}^{k} \lambda_i z_i + \lambda_{k+1} z_k}{S_{k+1}} \\ z_{k+1} &= \underset{x \in Q}{\operatorname{argmin}} \left\{ \lambda_{k+1} \max_{u \in U} \left\{ \sum_{j=1}^{m} u^{(j)} [f_j(x_{k+1}) + \langle \nabla f_j(x_{k+1}), x - x_{k+1} \rangle] - \phi(u) \right\} \right. \\ & \qquad \left. + \beta_{k+1} d(x) - \beta_k [d(z_k) - \langle \nabla d(z_k), x - z_k \rangle] \right\} \\ \widehat{x}_{k+1} &= \dfrac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i z_i \end{cases} \tag{4.3.12}$$

このアルゴリズムにおける z_k は, ある $a_1, \ldots, a_m,\ g \in E,\ b \in \mathbb{R}^m$ に対しての補助問題

$$\begin{array}{ll} \min & \max_{u \in U} \left\{ \sum_{j=1}^{m} u^{(j)} [\langle a_j, x \rangle + b^{(j)}] - \phi(u) \right\} + \beta_k d(x) + \langle g, x \rangle \\ \text{s.t.} & x \in Q \end{array} \tag{4.3.13}$$

の最適解として計算される．この問題は DA に対する補助問題 (4.3.10) と違い, 複雑さは k に依存しない．また, 補助問題 (4.3.13) のような構造をもつ最適化問題に対しては Nesterov によって smoothing technique [8] や excessive gap technique [7] を用いたアルゴリズムが提案されている．したがって, このアルゴリズム (4.3.12) は extremal convex problem に対する有用な解法となりうる．このアルゴリズムにおいて, パラメータを $\lambda_k = (k+1)/2,\ \beta_k \equiv L/\sigma$ と選ぶとき, 次の不等式が成り立つ.

$$f(\widehat{x}_k) - f(x^*) \leq \frac{4L[d(z_k) + \langle \nabla d(z_k), x^* - z_k \rangle]}{\sigma(k+1)(k+2)} \tag{4.3.14}$$

$d(x)$ の凸性により

$$d(z_k) + \langle \nabla d(z_k), x^* - z_k \rangle \leq d(x^*)$$

であるから, この上界 (4.3.14) は DA に対する上界 (4.3.11) よりも小さいことがわかる.

4.3.5　計算量

われわれは, extremal convex problem に対して MD にもとづくアルゴリズム (4.3.12) が有効であることがわかった．このアルゴリズムの計算量を見積もり, Nesterov のアルゴリズム 3.5.1 と比較しよう．$d_k(x)$ に対する仮定より, $d_k(x^*) \leq d(x^*)$ であるから, 上の定理 4.3.5 (i) に与えられる収束率の見積もりは, さらに次のように上からおさえることができる.

$$f(\widehat{x}_k) - f(x^*) \leq \frac{4Ld_k(x^*)}{\sigma(k+1)(k+2)} \leq \frac{4Ld(x^*)}{\sigma(k+1)^2}$$

ここで, 精度 $\varepsilon > 0$ に対して,

$$\frac{4Ld(x^*)}{\sigma(k+1)^2} \leq \varepsilon \iff k \geq \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} - 1 \impliedby k \geq \left\lceil \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil - 1$$

であるから, 精度 ε の近似解 $\widehat{x}_k$ は高々 $\left\lceil \sqrt{\dfrac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil - 1$ の反復回数で得ることができる.

関数値 $(f_1(x), \ldots, f_m(x))$ の一回の評価に必要な計算量の上界を F, 勾配 $(\nabla f_1(x), \ldots, \nabla f_m(x))$ を一回評価するのに要する計算量の上界を G_f とする. さらに勾配 $\nabla d(x)$ を得るための計算量の上界を G_d, 補助問題 (4.3.13) を解くのにかかる計算量の上界を A とすれば, アルゴリズム (4.3.12) の一反復の計算量の上界は $F + G_f + G_d + A$ である. したがって, 精度 ε の近似解を得るために十分な計算量として次が得られる.

$$A + (F + G_f + G_d + A) \times \left\lceil \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil \tag{4.3.15}$$

この計算量の上界を, Nesterov のアルゴリズム 3.5.1 と比較するために, $\phi(u) \equiv 0$ の場合を考えよう. $\|\cdot\| = \|\cdot\|_2$, $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $x_0 \in Q$ とすると $\sigma = 1$ であり,

$$\xi(z, x) = \frac{1}{2}\|(x - x_0) - (z - x_0)\|_2^2 = \frac{1}{2}\|x - z\|_2^2$$

が成り立つ. このとき, $\lambda_k = (k+1)/2$, $\beta_k \equiv L$ としたときの MD にもとづくアルゴリズム (4.3.12) は, $x_0 = z_{-1} \in Q$ および $\widehat{x}_0 = z_0 = \operatorname*{argmin}_{x \in Q} \left\{ \dfrac{1}{2} f(x_0, x) + \dfrac{L}{2}\|x - z_{-1}\|_2^2 \right\}$ を初期点として, $k \geq 0$ に対し,

$$\left\{ \begin{array}{rcl} x_{k+1} & = & \dfrac{\sum_{i=0}^{k} \lambda_i z_i + \lambda_{k+1} z_k}{S_{k+1}} \\ z_{k+1} & = & \operatorname*{argmin}_{x \in Q} \left\{ \dfrac{k+3}{2} f(x_{k+1}, x) + \dfrac{L}{2}\|x - z_k\|_2^2 \right\} \\ \widehat{x}_{k+1} & = & \dfrac{1}{S_{k+1}} \displaystyle\sum_{i=0}^{k+1} \lambda_i z_i \end{array} \right.$$

によっておこなわれる. $\nabla d(x) = x$ であるから, 一反復の計算量は Nesterov のアルゴリズム 3.5.1 と同じ上界 $F + G_f + A$ をもつとしてよい. さらにこのとき, 精度 ε の近似解を得るための計算量の上界は,

$$(F + G_f + A) \times \left\lceil \sqrt{\frac{2L}{\varepsilon}} \|x_0 - x^*\|_2 \right\rceil$$

となり, Nesterov のアルゴリズムにおける上界 (3.5.1) とほとんど同じである. したがって, われわれの提案するアルゴリズムは任意のノルムに関する強凸関数 $d(x)$ に関して記述できるという点において, Nesterov のアルゴリズムよりも柔軟なものになっているといえる.

4.4 微分可能な目的関数に対するアルゴリズム

この節では, 4.1 節の問題 (P) において, (b) の場合を考える. すなわち, 閉凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ と閉凸集合 $Q \subset \operatorname{dom} f$ に対して, $f(x)$ が Q 上で連続的微分可能であり, $\nabla f(x)$ が Q 上で定数 L に対してリプシッツ連続であるとして, 次の凸計画問題を考える.

$$\begin{array}{lll} (P) & \min & f(x) \\ & \text{s.t.} & x \in Q \end{array}$$

この問題 (P) が extremal convex problem に帰着されることを示し, 前節で得られたアルゴリズムがそのまま適用できることを示そう.

$$m=1,\quad f_1(x)=f(x),\quad ,U=\{1\}\subset\mathbb{R}^1,\quad \phi(u)\equiv 0\ \ (u\in\mathbb{R}^1) \tag{4.4.1}$$

とすれば,

$$f(x)=\max_{u\in U}\left\{\sum_{j=1}^{m}u^{(j)}f_j(x)-\phi(u)\right\}$$

が成り立つ. したがって, 問題 (P) は extremal convex problem に帰着される.

また, $y\in Q$ における $f(x)$ の線形化は,

$$f(y,x)=\max_{u\in U}\left\{\sum_{j=1}^{m}u^{(j)}[f_j(y)+\langle\nabla f_j(y),x-y\rangle]-\phi(u)\right\}=f(y)+\langle\nabla f(y),x-y\rangle$$

とかける. また, 点列 $\{x_k\}$ に対する近似モデル (4.1.1) は $l_k(x)=f(x_k)+\langle\nabla f(x_k),x-x_k\rangle$ で与えられるから, 定理 1.4.2 より

$$f(x)\le l_k(x)+\frac{L}{2}\|x-x_k\|^2$$

がなりたつ. これは不等式 (4.3.2) と一致する. したがって, この extremal convex problem に対して 4.3 節に述べた議論をそのまま適用することができる. すなわち, 問題 (P) に対してアルゴリズム 4.3.2 が有効であり, その効率は定理 4.3.3 で与えられる. とくに, 定理 4.3.5 によるパラメータの選び方により $f(\widehat{x}_k)-f(x^*)\le O(1/k^2)$ となる近似解 $\{\widehat{x}_k\}\subset Q$ を得ることができる.

以下ではとくに DA および MD の場合について考えよう.

4.4.1 Dual-averaging の場合のアルゴリズム

アルゴリズム 4.3.2 に対する DA の場合 (4.3.9) において上の (4.4.1) を代入すると, 次のアルゴリズムが得られる. $x_0=z_{-1}=\operatorname*{argmin}_{x\in Q}d(x)$ および

$$\widehat{x}_0=z_0=\operatorname*{argmin}_{x\in Q}\{\lambda_0[f(x_0)+\langle\nabla f(x_0),x-x_0\rangle]+\beta_0 d(x)\}$$

を初期点として, $k\ge 0$ に対して以下をおこなう.

$$\left\{\begin{array}{lcl} x_{k+1} & = & \dfrac{\sum_{i=0}^{k}\lambda_i z_i+\lambda_{k+1}z_k}{S_{k+1}} \\[2ex] z_{k+1} & = & \operatorname*{argmin}_{x\in Q}\left\{\displaystyle\sum_{i=0}^{k+1}\lambda_i[f(x_i)+\langle\nabla f(x_i),x-x_i\rangle]+\beta_{k+1}d(x)\right\} \\[2ex] \widehat{x}_{k+1} & = & \dfrac{1}{S_{k+1}}\displaystyle\sum_{i=0}^{k+1}\lambda_i z_i \end{array}\right. \tag{4.4.2}$$

実際には, z_k の計算するのに目的関数値 $\{f(x_i)\}$ は計算しなくてもよいことに注意する. このアルゴリズムを $\lambda_k=(k+1)/2,\ \beta_k\equiv L/\sigma$ に対しておこなえば, 定理 4.3.5 より,

$$f(\widehat{x}_k)-f(x^*)\le\frac{4Ld(x^*)}{\sigma(k+1)(k+2)} \tag{4.4.3}$$

が成り立つ.

4.4.2　Mirror-descent の場合のアルゴリズム

アルゴリズム 4.3.2 に対する MD の場合 (4.3.12) において上の (4.4.1) を代入すると, 次のアルゴリズムが得られる. $x_0 = z_{-1} = \operatorname*{argmin}_{x\in Q} d(x)$ および

$$\widehat{x}_0 = z_0 = \operatorname*{argmin}_{x\in Q} \Big\{ \lambda_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \beta_0 d(x) - \beta_{-1}[d(z_{-1}) - \langle \nabla d(z_{-1}), x - z_{-1}\rangle] \Big\}$$

を初期点として, $k \geq 0$ に対して以下をおこなう.

$$\left\{ \begin{array}{rcl} x_{k+1} & = & \dfrac{\sum_{i=0}^{k} \lambda_i z_i + \lambda_{k+1} z_k}{S_{k+1}} \\ z_{k+1} & = & \operatorname*{argmin}_{x\in Q} \Big\{ \lambda_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\ & & \qquad + \beta_{k+1} d(x) - \beta_k [d(z_k) - \langle \nabla d(z_k), x - z_k\rangle] \Big\} \\ \widehat{x}_{k+1} & = & \dfrac{1}{S_{k+1}} \displaystyle\sum_{i=0}^{k+1} \lambda_i z_i \end{array} \right. \tag{4.4.4}$$

この場合も, z_k の計算に目的関数値 $\{f(x_i)\}$ の計算は必要ないことに注意する. このアルゴリズムを $\lambda_k = (k+1)/2$, $\beta_k \equiv L/\sigma$ に対しておこなえば, 定理 4.3.5 より,

$$f(\widehat{x}_k) - f(x^*) \leq \frac{4L[d(z_k) + \langle \nabla d(z_k), x^* - z_k\rangle]}{\sigma(k+1)(k+2)} \tag{4.4.5}$$

が成り立つ. $d(x)$ の凸性により

$$d(z_k) + \langle \nabla d(z_k), x^* - z_k\rangle \leq d(x^*)$$

であるから, この上界 (4.4.5) は DA に対する上界 (4.4.3) よりも小さいことがわかる.

4.4.3　Nesterov のアルゴリズムとの比較

上に挙げた二つのアルゴリズムにおいて, $\beta_k \equiv L/\sigma$ としたものを考えるとき, 補助問題の解 z_k はそれぞれ次のように計算される.

$$\begin{aligned} \text{DA (4.4.2)} \ &: \ z_k = \operatorname*{argmin}_{x\in Q} \left\{ \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] + \frac{L}{\sigma} d(x) \right\} \\ \text{MD (4.4.4)} \ &: \ z_k = \operatorname*{argmin}_{x\in Q} \left\{ \lambda_k [f(x_k) + \langle \nabla f(x_k), x - x_k\rangle] + \frac{L}{\sigma} \xi(z_{k-1}, x) \right\} \end{aligned}$$

ここで, Nesterov のアルゴリズム 3.4.3 を思い出そう: $x_0 = z_{-1} = \operatorname*{argmin}_{x\in Q} d(x)$ および

$$\widehat{x}_0 = \widehat{z}_0 = z_0 = \operatorname*{argmin}_{x\in Q} \left\{ \lambda_0 [f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \frac{L}{\sigma} d(x) \right\}$$

を初期点として, $k \geq 0$ に対して以下をおこなう.

$$\left\{\begin{array}{llcl} \text{(a)} & x_{k+1} & = & \dfrac{\sum_{i=0}^{k} \lambda_i \widehat{z}_i + \lambda_{k+1} z_{k+1}}{S_{k+1}} \\ \text{(b)} & z_{k+1} & = & \operatorname*{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k+1} \lambda_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + \dfrac{L}{\sigma} d(x) \right\} \\ \text{(c)} & \widehat{z}_{k+1} & = & \operatorname*{argmin}_{x \in Q} \left\{ \lambda_{k+1} [f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle] + \dfrac{L}{\sigma} \xi(z_k, x) \right\} \\ \text{(d)} & \widehat{x}_{k+1} & = & \dfrac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i \widehat{z}_i \end{array}\right. \tag{4.4.6}$$

このことから, 次の事実がわかる.

- アルゴリズム (4.4.6) において (c) を除いて, $\widehat{z}_k$ を z_k におきかえると, DA アルゴリズム (4.4.2) において $\beta_k \equiv L/\sigma$ としたものになる.
- アルゴリズム (4.4.6) において (b) を除いて, z_k を $\widehat{z}_k$ におきかえると, MD アルゴリズム (4.4.4) において $\beta_k \equiv L/\sigma$ と同じ更新式が得られる (初期点 $\widehat{x}_0$ のとりかたは異なる).

また, Nesterov のアルゴリズム (4.4.6) に対する収束率の見積もりは (4.4.3) と同じであった. このことは, Nesterov のアルゴリズムの各反復でおこなわれる二つの補助問題を一つに減らしてもなお同じ収束率の見積もりが保証できることを意味する. とくに, MD アルゴリズム (4.4.4) の場合はさらに良い見積もり (4.4.5) が保証できる.

4.4.4　計算量

上記に得られた二つのアルゴリズム (4.4.2), (4.4.4) に対する計算量の見積もりを与えよう. 目的関数値と最適値との残差に対する上界 (4.4.3) および (4.4.5) はいずれも次を満たす.

$$f(\widehat{x}_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma(k+1)^2}$$

したがって, 精度 ε の近似解 $\widehat{x}_k$ を高々 $\left\lceil \sqrt{\dfrac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil - 1$ の反復回数で得ることができる.

勾配 $\nabla f(x)$, $\nabla d(x)$ の評価に要する計算量の上界をそれぞれ G_f, G_d とし, $s \in E$ に対して補助最適化問題 $\min_{x \in Q} \{\langle s, x \rangle + d(x)\}$ を解くために必要な計算量の上界を A とする. このとき, アルゴリズム (4.4.2) および (4.4.4) では一反復の計算量の上界は次のように与えられる.

$$\text{DA (4.4.2)} \ : \ G_f + A, \qquad \text{MD (4.4.4)} \ : \ G_f + G_d + A$$

したがって, 精度 ε の近似解を得るための計算量の上界は,

$$\begin{aligned} \text{DA (4.4.2)} &: A + (G_f + A) \times \left\lceil \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil \\ \text{MD (4.4.4)} &: A + (G_f + G_d + A) \times \left\lceil \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil \end{aligned}$$

となる. Nesterov のアルゴリズム 3.4.3 における上界は,

$$(G + 2A) \times \left\lceil \sqrt{\frac{4Ld(x^*)}{\sigma\varepsilon}} \right\rceil$$

であったから, とくに DA (4.4.2) はより小さな上界を与えている.

4.5 まとめ

われわれは閉凸関数 $f : E \to \mathbb{R} \cup \{+\infty\}$ と閉凸集合 $Q \subset \mathrm{dom}\, f$ に対する凸計画問題

$$\begin{array}{lll} (P) & \min & f(x) \\ & \text{s.t.} & x \in Q \end{array}$$

に対して, 劣勾配アルゴリズムに関する統一的な枠組みを提案した. この枠組みに対して, Mirror-Descent (MD) および Dual-Averaging (DA) の場合を当てはめることにより具体的なアルゴリズムを提案した. $x^* \in Q$ を問題 (P) の最適解とするとき, 提案アルゴリズムが生成する $\{\widehat{x}_k\} \subset Q$ に対する $f(\widehat{x}_k) - f(x^*)$ は MD の方が DA よりも小さな上界を与えた. さらに, 既存アルゴリズムとの比較をおこなった結果, 以下に述べる有意義な結論が得られた.

4.5.1 目的関数に微分可能性を仮定しない場合

目的関数に微分可能性を仮定しない場合には二つの効率的なアルゴリズムが提案できた. いずれのアルゴリズムも生成される近似解 $\{\widehat{x}_k\} \subset Q$ に対して $f(\widehat{x}_k) - f(x^*) \le O(1/\sqrt{k})$ が達成された. これに DA の場合を当てはめると, 一方は Nesterov [9] による本来の DA アルゴリズムが得られたが, 他方からも同じ効率をもつアルゴリズムが得られた. MD の場合を当てはめて得られるアルゴリズムはもともとの MD アルゴリズム [1, 4] にスケーリングパラメータ $\{\beta_k\}$ を導入した拡張アルゴリズムと見なすことができる. この拡張された MD は DA よりも $f(\widehat{x}_k) - f(x^*)$ を小さな上界でおさえることができる. もともとの MD の効率が $f(\widehat{x}_k) - f(x^*) \le O(\log k/\sqrt{k})$ でったのに対し, 今回の拡張では $O(1/\sqrt{k})$ の上界が達成できている.

4.5.2 目的関数が微分可能な場合

微分可能な目的関数に対しては $f(\widehat{x}_k) - f(x^*) \le O(1/k^2)$ を達成するアルゴリズムが提案できた. MD および DA の場合を当てはめて得られるアルゴリズムは, Nesterov [8] によるアルゴリズムと同じ効率を保つにもかかわらず, 一反復の補助問題の回数を半分に減らすことができた.

4.5.3 Extremal convex problem の場合

Extremal convex problem に対しては MD にもとづくアルゴリズム (4.3.12) を提案し, $f(\widehat{x}_k) - f(x^*) \leq O(1/k^2)$ が達成されることを示した. Nesterov [5] によるアルゴリズムはノルムが $\|\cdot\|_2$ に限定されていたのに対し, われわれの提案するアルゴリズムは一般のノルム上の強凸関数 $d(x)$ に関して記述することができる.

謝辞

本論文を書くにあたり, 指導教員の福田光浩准教授には初歩から丁寧にご指導いただき, 原稿を精読し数々の助言を与えてくださいました. また, 副指導教員の山下真准教授には研究における有益な示唆や原稿に対する多くのご指摘をいただき, 大変参考になりました. ここに深く感謝の意を申し上げます.

参考文献

[1] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Operations Research Letters*, **31**, pp. 167–175, 2003.

[2] J. Borwin and A. Lewis, *Convex analysis and nonlinear optimization*, Springer, 2006.

[3] L. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, *USSR Computational Mathematics and Methematical Physics*, **7**, pp. 200–217, 1967.

[4] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Nauka Publishers, 1978 (in Russian); English translation: John Wiley & Sons, 1983.

[5] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, *Soviet Mathematics Doklady*, **27**, pp. 372–376, 1983.

[6] Y. Nesterov, *Introductory lectures on convex optimization : A basic course*, Kluwer Academic Publishers, 2004.

[7] Y. Nesterov, Excessive gap technique in nonsmooth convex minimization, *SIAM Journal on Optimization*, **16**, pp. 235–249, 2005.

[8] Y. Nesterov, Smooth minimization of non-smooth functions, *Mathematical Programming*, Ser. A, **103**, pp. 127–152, 2005.

[9] Y. Nesterov, Primal-dual subgradient methods for convex problems, *Mathematical Programming*, Ser. B, **120**, pp. 221–259, 2009.

[10] R. T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.

[11] 高橋 渉, 『非線形関数解析学』, 近代科学社, 1988.