機械学習における劣モジュラ性の利用と組合せ論的アルゴリズム

河原　吉伸

機械学習において，本質的には集合関数最適化である問題は数多く存在する．近年このような問題に対して，集合関数における凸性としてとらえられる離散構造である劣モジュラ性が，効率的な学習アルゴリズムの構築や理論的解析において有用であることが認知されてきている．本稿では，劣モジュラ性を用いた機械学習とそのアルゴリズムに関して，最近の研究動向の紹介を行うとともに基本的事項について解説する．

キーワード：機械学習，集合関数最適化，劣モジュラ性

1. はじめに

高度な知能情報処理を実現するための理論／アルゴリズム体系である機械学習は，近年ますます重要性を増す研究領域の一つである[1]．機械学習では，確率・統計や情報理論，最適化など種々の応用数理に基づき，データの背後にある規則性を推定・獲得するための方法論に関する議論が行われる．機械学習で扱われる問題の多くにおいて，組合せ的計算，特に集合関数の最適化は本質的役割を果たす．例えばデータのクラスタリングにおいては，すべてのサンプルにおのおのが属するクラスターのラベルを割り当てるという組合せ的なタスクを行う．また機械学習や統計において議論される特徴選択（詳細は後述）においては，すべての変数（特徴，次元）の中から，分類や回帰などに有用な一部の変数だけを選択する（つまり全変数の索引から成る集合から，その一部をある基準により選択する集合関数最適化問題としてとらえられる）．

機械学習分野では，このような本質的には組合せ的計算である問題に対して，一般に連続最適化への緩和を通した近似的なアプローチをとる．しかし一方で，組合せ的問題として直接扱うことが有効な手段となりうる場合も多い．最近ではそのような場面で，劣モジュラ性と呼ばれる，集合関数における凸性としてとらえられる離散構造が注目されている．劣モジュラ性は，組合せ最適化分野で 1970 年代に入ってその有用性が認識されるようになって以来，現在でも盛んに議論が行われている重要な概念である [10, 11]．最近では，劣モジュラ性を含む離散凸構造に関して，離散凸解析として統一的な体系化も進められている [32]．機械学習分野においても，多くの問題に対して，劣モジュラ性が効率的なアルゴリズムの構築や理論的解析などに有用であることが広く認識されつつある．

本稿ではこのような情勢を踏まえ，劣モジュラ性を用いた機械学習とそのアルゴリズムや応用に関して，最近の研究動向の紹介を含めて解説する．まず第 2 章では，機械学習における組合せ的問題の代表例を概観し，これらにおいて劣モジュラ性がどのような役割を持つかについて説明する．そして第 3 章では，劣モジュラ性を用いた学習アルゴリズムのいくつかの代表的な例について，基本的事項を説明する．

2. 機械学習と劣モジュラ性

機械学習において扱われる代表的問題には，特徴選択（変数選択）やクラスタリングなど，本質的に組合せ的計算を伴うものも多い．本節ではこのような問題の例を挙げ，これらにおいて劣モジュラ性がいかに有用であるかを概観する．具体的な学習アルゴリズムについては，次節においていくつか取り上げて解説する．

2.1 機械学習における組合せ的問題

上述のように，機械学習では組合せ的問題が数多く扱われる（表 1 を参照）．

例えば冒頭でもふれた（k クラス）クラスタリングでは，全サンプル（サンプルから成る集合を $\mathcal{V}$ とする）に対して，属するクラスタ 1〜k の割り当てを行う．これは明らかに一種の整数計画 (IP) 問題であり，実際

かわはら　よしのぶ
大阪大学 産業科学研究所
〒 567–0047 大阪府茨木市美穂ヶ丘 8–1
E-mail: kawahara@ar.sanken.osaka-u.ac.jp

[1] 良質な書籍としては [4] や [14] などがある．

表 1 機械学習における集合関数最適化の例

問題	有限集合	評価関数の例	文献（発表年順）
特徴選択	説明変数の索引	2 乗誤差	[9, 2]
能動学習	サンプル	Fisher 行列の比	[17]
クラスタリング	サンプル	カット関数，MDL	[37, 23]
画像セグメンテーション	画素	エネルギー（カット）	[6, 26]
グラフ構造学習	ノード	相互情報量	[35, 7]

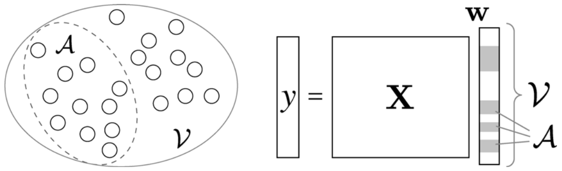

図 1 集合関数最適化としての，クラスタリング（左）と特徴選択（右）の概念図

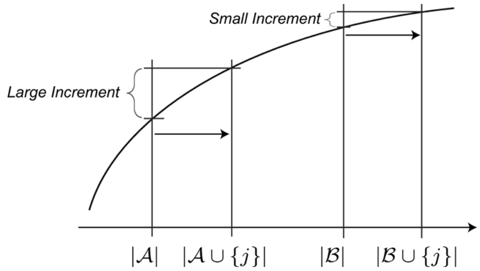

図 2 劣モジュラ関数の定義 (2.1) の概念図

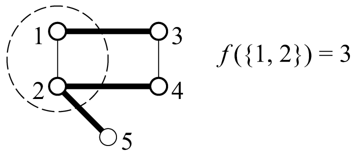

図 3 カット関数の例

IP によるアプローチに関する研究も古くから行われている [13]．なお $k = 2$ に限った場合は，クラスタリングは集合関数最適化問題としてもとらえられる（図 1 参照）．しかし一般に機械学習で議論されるクラスタリングへのアプローチは，直接組合せ的問題として扱うのではなく，緩和を通して連続最適化として定式化される．例えば機械学習における代表的アプローチであるスペクトラル・クラスタリングにおいては，一般化固有値問題として緩和解を計算し，それを整数解へ丸めることでクラスタリングを行う [31]．

また，先に例として挙げた特徴選択に関しても一般に同様のことが言える．教師あり学習[2]を例に考えると，出力 y の予測に有用となる特徴 $\mathbf{x}$（の次元）の部分集合を何らかの基準により選択する．例えば線形回帰モデル $y = \mathbf{w}^\top \mathbf{x}$（$\mathbf{w} \in \mathbb{R}^d$ は回帰パラメータ）の推定を考えた場合，手元にあるデータから自乗誤差などに基づき $\mathbf{w}$ を決定する．変数の索引から成る集合を $\mathcal{V}$ とした場合，特徴選択では各部分集合 $\mathcal{A} \subseteq \mathcal{V}$ に関して（サイズ制約などの制約の下で）基準を計算して最適な $\mathcal{A}$ を選ぶ（図 1 参照）．

上述のクラスタリングでは，その基準としてカット関数や最小記述長 (MDL) が用いられることが多い．一方特徴選択では，自乗誤差や相互情報量などが基準となることが多い．これらは，集合関数の凸関数としてとらえられる劣モジュラ関数であることが知られている．このように一般に機械学習で扱われる組合せ的問題は，劣モジュラ性との関連から議論できるものが多い．

2.2 劣モジュラ性

本節では，劣モジュラ性の基本的事項について紹介する．個々の具体的な学習アルゴリズムでの劣モジュラ性の利用の詳細に関しては，次節において述べる．

劣モジュラ性の定義

劣モジュラ性は，集合関数における凸性としてとらえることができる離散構造である [11, 32]．いくつかの等価な定義があるが，次の定義は直感的にもわかりやすくよく用いられる ($\forall\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}, \forall j \in \mathcal{V} \setminus \mathcal{B}$).

$$f(\mathcal{A}\cup\{j\}) - f(\mathcal{A}) \geq f(\mathcal{B}\cup\{j\}) - f(\mathcal{B}) \quad (2.1)$$

つまり包含関係にある 2 つの集合 $\mathcal{A}$ と $\mathcal{B}$ に関して，包含される集合 $\mathcal{A}$ へ新しい要素 j を加えた際の差分が，包含する集合 $\mathcal{B}$ の場合のそれより大きくなる（図 2 参照）．このように劣モジュラ関数は，サイズとともに増加が穏やかになる性質を持っており，経済における「規模の経済性」や「限界効用逓減の法則」をモデル化するものとしてもとらえられる．そのほかよく用いら

[2] 入力 $\mathbf{x} \in \mathbb{R}^d$ と出力 $y \in \mathbb{R}$ の組から成るデータ $\{\mathbf{x}_i, y_i\}_{i=1}^n$ が与えられたとき，教師あり学習では，入力 $\mathbf{x}$ から y までの予測器（分類器や回帰器など）を推定する．

れる劣モジュラ性の等価な定義としては，次のものがある．

$$f(\mathcal{A})+f(\mathcal{B})\geq f(\mathcal{A}\cup\mathcal{B})+f(\mathcal{A}\cap\mathcal{B})\ (\forall\mathcal{A},\mathcal{B}\subseteq\mathcal{V})$$

これ以外の定義については [39] などを参照されたい．また $\forall\mathcal{A}\subseteq\mathcal{B}\subseteq\mathcal{V}$ に対し，集合関数 f が $f(\mathcal{A})\leq f(\mathcal{B})$ を満たすとき f は単調であるという．さらに $\forall\mathcal{A}\subseteq\mathcal{V}$ に対し，$f(\mathcal{A})=f(\mathcal{V}\setminus\mathcal{A})$ となるとき f は対称であるという（記号 $\setminus$ は差集合を表す）．劣モジュラ関数の最小化は，多項式時間で計算可能であることが知られている [11]．一方でその最大化や，制約付きの最小化 NP 困難であり，一般に効率的な計算方法が存在しない．

劣モジュラ関数の例

まず一つ目の例として，クラスタリングなどでも用いられるカット関数を紹介する．頂点集合を $\mathcal{V}=\{1,\ldots,d\}$，辺集合を $\mathcal{E}$ とする無向グラフが与えられ，さらに各辺 $e\in\mathcal{E}$ には正の重み c_e が定まっているとする．このとき，カット関数は次式で定義される．

$$f(\mathcal{A})=\sum\{c_e : e\in\mathcal{E}(\mathcal{A},\mathcal{V}\setminus\mathcal{A})\}\ \ (\mathcal{A}\subseteq\mathcal{V})$$

ここで $\mathcal{E}(\mathcal{A},\mathcal{V}\setminus\mathcal{A})$ は，端点の一方が $\mathcal{A}$ に，もう一方が $\mathcal{V}\setminus\mathcal{A}$ に含まれる辺の集合を表す．つまりカット関数は，頂点集合 $\mathcal{A}$ と $\mathcal{V}\setminus\mathcal{A}$ に両端点がまたがる辺上の重みの和になる．図 3 は，$d=5$ の場合の例を表す．

もう 1 つの例として（予測）自乗誤差について述べる．自乗誤差は，機械学習や統計の最も基本的な評価誤差の 1 つであると言える．Das–Kempe は，特徴集合の関数としての線形回帰における自乗誤差関数

$$f(\mathcal{A})=E[(y-\textstyle\sum_{i\in\mathcal{A}}\hat{w}_i x_i)^2] \tag{2.2}$$

が優モジュラ関数[3]であることを示した [9]．上式における $\hat{w}_i$ は，$\mathcal{A}$ に限定した場合の最適な回帰係数である．したがって，線形回帰における自乗誤差最小化に基づく特徴選択は，劣モジュラ最大化であることが言える．

2.3 緩和と劣モジュラ性に基づくアプローチの関連

上で述べたように，機械学習分野では，組合せ的な問題を緩和を通して連続最適化（凸最適化）として近似的な定式化を行う場合が多い．その大きな理由は，効率的な計算が可能になることが挙げられる．当然このアプローチでは，元の組合せ的問題の真の解とは異なる近似的な解しか得ることができない．しかし予測が目的である場合には，効率性と精度のトレード・オフの観点から，このアプローチで十分である場合が多い．データを生成している機構自体の解析が目的である場合には，（局所的ではなく）大域的な最適性を持った解の探索が必要となる．また後で見るように，組合せ的なアプローチ自体がそもそも効率的で，かつ精度の高い解法となる問題もある．このように問題や目的に応じて，緩和による定式化や組合せ的な定式化を考えることが重要であると言える．なお図 4 は，特徴選択における連続緩和・組合せ的方法の関係と研究例を示したものである．

3. 劣モジュラ性を用いた学習アルゴリズム

本節では，劣モジュラ性を用いた機械学習アルゴリズムに関して，代表的な具体的問題の基本事項の説明と最近の研究事例の紹介を行う．また図 5 は，劣モジュラ最適化の観点からの代表的な機械学習問題の分類を示したものである．

3.1 劣モジュラ最大化と貪欲法の適用

機械学習分野で劣モジュラ性が議論され始めたのは 2000 年代後半になってからのことであるが，当時最も多く行われたのは，扱う問題を劣モジュラ関数の最大化として定式化して貪欲法を適用する，という流れの研究である．先に紹介した特徴選択をはじめとして，多くの問題が劣モジュラ最大化として定式化できるため，関連した研究例が近年多数見られる [8, 17, 28, 41]．

$f:2^{\mathcal{V}}\rightarrow\mathbb{R}$ を劣モジュラ関数とした場合，サイズ制約下での劣モジュラ最大化は次式のように表される．

$$\max_{\mathcal{A}\subseteq\mathcal{V}}\ f(\mathcal{A})\quad \text{s.t. } |\mathcal{A}|\leq k$$

ただし $\mathcal{V}$ は有限集合，k は正の整数である．一般にこの問題は NP 困難であることが知られており，効率的な（厳密）アルゴリズムの存在は絶望的である [39]．しかし関数が単調である場合には，貪欲法により良い近似解が得られることが理論的にも保証される．サイズ制約下での劣モジュラ最大化への貪欲法では，サイズ制約を満たすまで 1 つずつ，解を最も改善する要素を加えていくという手順からなる：

```
A_0 ← ∅, i ← 1 に設定.
while i ≤ k do
    ρ^+_{j_i}(A_{i-1}) = arg max_{j∈V\A_{i-1}} ρ^+_j(A_{i-1}) となる
    要素 j_i ∈ V \ A_{i-1} を選択.
    A_i ← A_{i-1} ∪ {j_i}, i ← i+1 に設定.
```

[3] (-1) 倍が劣モジュラ関数であるような集合関数を優モジュラ関数と呼ぶ．

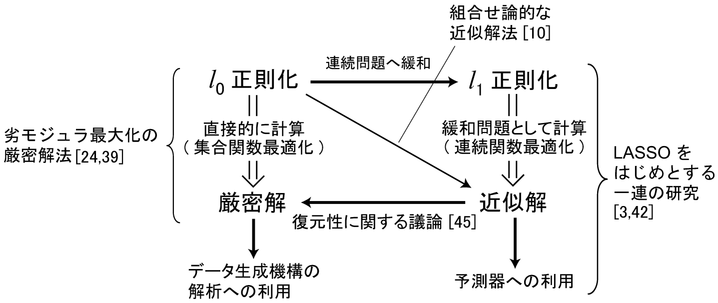

図 4　特徴選択の場合の緩和と組合せ的方法の関係

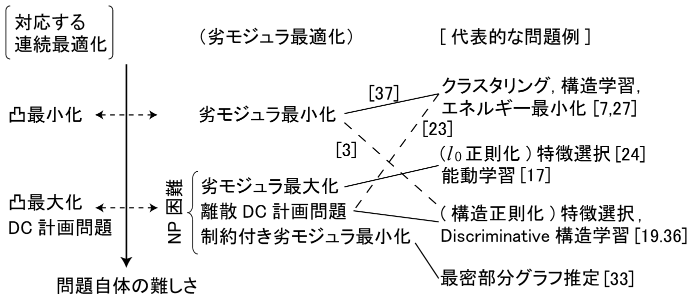

図 5　代表的な機械学習問題の分類

ただし $\rho_j^+(\mathcal{A}) := f(\mathcal{A} \cup \{j\}) - f(\mathcal{A})$ である．このように貪欲法は極めて単純な手順であるにもかかわらず，$(1 - 1/e) \approx 0.632$ の近似率を達成することが知られている（e は自然対数の底）[39]．経験的にも多くの問題で，問題を劣モジュラ最大化として定式化し貪欲法を適用することにより，極めて実用的な解が得られることが多数報告されており，その応用は，センサ配置 [28] やグラフマイニング [41]，能動学習 [17] など多岐にわたる．

例えば 2.1 節でも述べた自乗誤差に基づく特徴選択は，選択する特徴集合に関する自乗誤差 (2.2) を最小化する．この問題は選択する特徴の数の上限を k とした場合，上述のようにサイズ制約下の劣モジュラ最大化である [9]．したがって貪欲法により良い近似解が得られることが理論的にも保証される．貪欲法の適用は，統計的，あるいは経験的にも良い近似解が得られることがこれまでも報告されている [29, 44]．

3.2　グラフカットの利用

劣モジュラ性が機械学習において用いられるもう 1 つの代表例としては，画像処理などで議論されるエネルギー最小化へのグラフカット（最小カット）の適用が挙げられる [6, 26]．グラフカットは，いわゆる最小カット最大流定理に基づき，きわめて効率的な計算が可能な問題部類であることがよく知られている [12]．

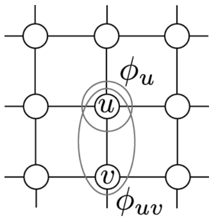

図 6　マルコフ確率場の概念図

コンピュータ・ビジョンのさまざまなタスクにおいては，画像中の各画素にラベルづけを行う問題として定式化されるものが多い．このような問題は一般に，格子状のグラフ構造 $(\mathcal{E}, \mathcal{V})$ を持った確率変数（マルコフ確率場：MRF）に関する推論問題として扱われることが多い（$\mathcal{E}$ と $\mathcal{V}$ はおのおのエッジとノードの集合，図 6 も参照）．エネルギー最小化は，MRF における最大事後確率推定を，事後確率の負の対数であるエネルギー関数の最小化問題として扱う枠組みである．最も簡単なケースである 2 値一階 MRF の場合には，エネルギー関数 $E : 2^{\mathcal{V}} \to \mathbb{R}$ は次のように表される．

$$E(\mathbf{x}) = \sum_{v \in \mathcal{V}} \phi_u(x_u) + \sum_{(u,v) \in \mathcal{E}} \phi_{uv}(x_u, x_v)$$

ただし ϕ_v および ϕ_{uv} は，各ノード／エッジ上に与えられた実数値関数（ポテンシャルと呼ばれる）である．Boykov らは，2 値ラベルの場合のエネルギー関数が劣

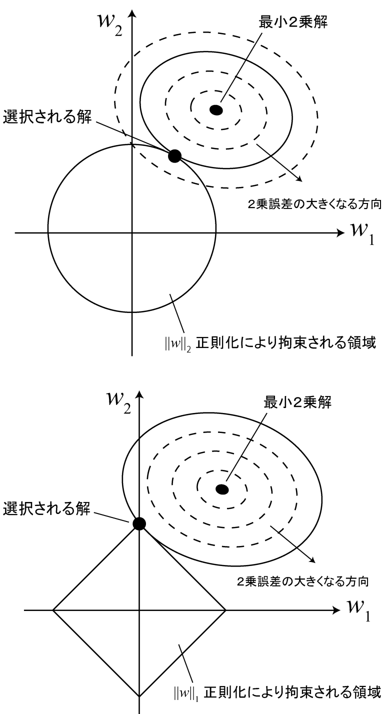

図 7 正則化による特徴選択の概念図

モジュラ性を満たす場合[4]，グラフカットにより厳密解を得ることができるのを示した [6]．実際の応用では，必ずしもエネルギー関数が劣モジュラ性を満たすわけではないが，多くのタスクにおいて近似的に満たすため，グラフカットにより有用な解が得られることが知られている [18, 27]．

エネルギー最小化以外の学習問題へのグラフカットの適用例もいくつか見られる．例えば Blum-Chawla は，一般的な分類学習への最小カットの適用について議論している [5]．また Azencott らは，グラフ構造情報を用いた特徴選択を最小カットにより定式化し，遺伝子データ解析へ適用している [1]．最近では，機械学習においてもさまざまな応用で重要となる劣モジュラ最小化に関して，最小カットにより効率的に計算するオンライン・アルゴリズムが提案されている [22]．

3.3 構造正則化と劣モジュラ最小化

近年注目される機械学習における枠組みの 1 つに，学習における事前情報としての構造（グループ構造，グラフ構造など）を利用した構造正則化学習がある [2, 43]．

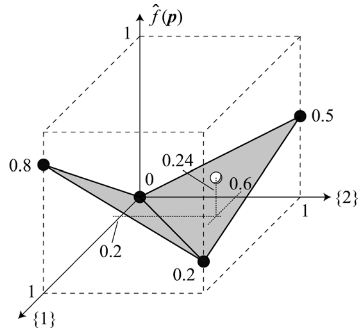

図 8 $d = 2$ の場合の Lovász 拡張の例

正則化に基づく学習は，学習タスクの評価関数（損失関数）$L : \mathbb{R}^d \rightarrow \mathbb{R}$ に対して，最適化されるパラメータへの一種のペナルティ項 $\Omega : \mathbb{R}^d \rightarrow \mathbb{R}$ （正則化項とも呼ばれる）を加えた定式化を行う（ [4] などを参照)．

$$\min_{\mathbf{w} \in \mathbb{R}^d} \; L(\mathbf{w}) + \lambda \cdot \Omega(\mathbf{w})$$

$\lambda \in \mathbb{R}$ は，損失関数と正則化項のバランスを調整するパラメータ（正則化パラメータ）である．正則化項の選択により解を疎（パラメータ $\mathbf{w}$ 中の多くの次元の値が 0）にできるため，機械学習分野では特徴選択の定式化としてよく用いられる．図 7 は $d = 2$ において，Ω として l_2 正則化 $(\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2)$ と l_1 正則化 $(\Omega(\mathbf{w}) = \|\mathbf{w}\|_1)$ をおのおの選択した場合の概念図である．図からわかるように，l_1 正則化を選択した場合のほうが，解が軸上に乗り 0 になる要素が多くなりやすい [42]．

構造正則化は，この正則化学習の枠組みにおいて，グループ構造やグラフ構造といった構造に関する正則化項 Ω を用いることで，問題に対する事前情報を反映した学習を行うための方法である．構造の例としては，次のようなものが挙げられる．

グラフ正則化 [16]　グラフ構造 $(\mathcal{E}, \mathcal{V})$ とエッジ上の重み c_{ij} $(i, j \in \mathcal{V})$ が与えられたとき，正則化項として正規化したグラフ・ラプラシアン[5]$\bar{\mathcal{L}}$ を用いて $\Omega(\mathbf{w}) = \sum_{ij} \bar{\mathcal{L}}_{ij} |w_i - w_j|$ とすると，グラフ上で連結性が高い変数が同時に選択されやすいように機能する．例えば画像のセグメンテーションを考えた際には，通常オブジェクトは連続した複数の画素にわたるため，隣接する画素は同一のセグメントへと含まれやすいと考えるのが自然である．このような場合は，格子状のグラフに関するラプラシアン正則化を用いることがで

[4] 2 次のポテンシャル関数における劣モジュラ性は，集合関数の定義 (2.1) から次式のように導かれる [6].

$$\phi_{uv}(0, 0) + \phi_{uv}(1, 1) \leq \phi_{uv}(0, 1) + \phi_{uv}(1, 0)$$

[5] グラフの各辺上の重み行列（隣接行列）$\mathbf{C}$ が与えられたとき，グラフ・ラプラシアンは $\bar{\mathcal{L}} = \mathbf{D} - \mathbf{C}$ のように定義される．ただし $\mathbf{D} = \mathrm{diag}(\mathbf{C1})$ である．

きる．

グループ正則化 [21, 43]　重なりを許した集合の組 $\mathcal{G}$ が与えられているとする（つまり $\mathcal{G}$ の各要素 $g \in \mathcal{G}$ が $\mathcal{V}$ の部分集合 $(g \subset \mathcal{V})$）．このとき，次式で定義される l_1/l_q-正則化を用いることで，グループに含まれる変数は同時に選ばれやすく，かつグループに関しては疎になりやすい特徴選択が可能となる．

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_q$$

ここで $\mathbf{w}_g$ は，グループ g に含まれる次元のみで構成されたパラメータ・ベクトルである．この正則化は，階層的構造を持つ変数を持つデータや，時系列のようなシーケンス・データ，2 次元格子状の構造を持つデータなどへ適用可能である [3]．

近年 Bach は，機械学習分野で知られる構造正則化の多くは，劣モジュラ関数の Lovász 拡張で統一的に表されることを示した [2]．集合関数 $f : 2^d \to \mathbb{R}$ に対する Lovász 拡張 $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ の定義は次式のようになる．

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{m-1} (\hat{x}_i - \hat{x}_{i+1}) f(\mathcal{U}_i) + \hat{x}_m f(\mathcal{U}_m)$$

ここで $\hat{x}_1 \geq \cdots \geq \hat{x}_m (m \leq d)$ は $\mathbf{x}$ 中の異なる値を並べたものであり，$\mathcal{U}_i = \{j \in \mathcal{V} : x_j \geq \hat{x}_i\}$ である [30]．この Lovász 拡張が凸関数であるとき，またそのときに限り元の集合関数は劣モジュラ関数であることが知られている．図 8 は $d = 2$ の場合の例である[6]．Bach はこの関係に基づき，構造正則化への最も一般的な解法の 1 つである近接勾配法における各反復の勾配計算が，劣モジュラ最小化により計算可能であることを示した [2, 3]．

3.4　その他の学習アルゴリズム

上述のもの以外にも，機械学習で議論される劣モジュラ性を用いた学習アルゴリズムは数多く見られる．

まず，機械学習で扱う代表的問題の 1 つであるクラスタリングは，劣モジュラ関数であるカット関数や MDL などをその基準として扱う場合が多い [37]．Narasimhan–Bilmes は，クラスタ間のサンプル数に偏りの少ないクラスタリング（バランス・クラスタリング）に対して，劣モジュラ性に基づく局所探索アルゴリズムを提案している [38]．バランス・クラスタリングに対しては，著者らも離散ニュートン法（Dinkelbach 法）に基づく反復的なクラスタリング・アルゴリズムを提案している [23]．また永野らは，劣モジュラ性から導かれる離散構造（劣モジュラ多面体 [11] の性質）を利用して，クタスタ数を事前に指定する必要がないクラスタリング手法を提案している [34]．

また機械学習においても，データを生成する機構に関して解析を行うためには，大域的な解を探索するアルゴリズムは重要となる．これに関連して最近著者らは，機械学習においても重要となる代表的な NP 困難問題である，劣モジュラ最大化と離散 DC 計画（2 つの劣モジュラ関数の差の最小化）への大域最適化手法を提案している [24, 25]．また機械学習でも重要な役割を持つ劣モジュラ最小化に関しては，データが得られるたびに解を更新するオンライン・アルゴリズムもいくつか提案されている [15, 22]．最近では，統計的な独立性基準やダイバージェンス（分布に関する距離的なもの）の一種である Bregman ダイバージェンスと，劣モジュラ性との関連に関する研究も見られる [20, 40]．

これらに加えて，劣モジュラ最小化を用いたグラフ構造の学習 [7, 35] や，機械学習に有用な基礎的な劣モジュラ最適化のアルゴリズムに関する研究なども見られる [19, 33, 36]．

4.　まとめ

劣モジュラ性は，集合関数における凸性としてとらえられる離散構造であり，近年では効率的な学習アルゴリズムの構築や理論的解析において有用であることが認知されるようになった．本稿では，最近の関連する研究動向の紹介を行うとともに，劣モジュラ性を用いた代表的な学習アルゴリズムの基本的事項に関して解説を行った．特に，学習問題を劣モジュラ最大化として定式化後に貪欲法を適用する一連の研究や，コンピュータ・ビジョンなどへのグラフカットの適用，また構造情報を利用した学習（構造正則化）と劣モジュラ性との関連に関し詳細に説明した．劣モジュラ性を利用した機械学習の研究は最近になり議論が行われるようになった領域であり，今後のより盛んな研究が期待される．

[6] 関数値が $f(\emptyset) = 0$，$f(\{1\}) = 0.8$，$f(\{2\}) = 0.5$，$f(\{1,2\}) = 0.2$ と与えられた場合（図中黒丸），Lovász 拡張は図中の灰色の平面で表される．例えば $x = (0.2, 0.6)$ の場合，定義に従い $x^1 = 0.6$，$x^2 = 0.2$，および，$\mathcal{U}_1 = \{2\}$，$\mathcal{U}_2 = \{1, 2\}$ となり，$\hat{f} = (0.6 - 0.2) \times f(\{2\}) + 0.2 \times f(\{1, 2\}) = 0.24$ のように計算される．

参考文献

[1] C. A. Azencott, D. Grimm, Y. Kawahara and K. M. Borgwardt, A min-cut solution to mapping phenotypes to networks of genetic markers, *arXiv:1211.2315*, 2012.

[2] F. Bach, Structured sparsity-inducing norms through submodular functions, In *Advances in Neural Information Processing Systems*, **23**, 118–126. 2010.

[3] F. Bach, R. Jenatton, J. Mairal and G. Obozinski, Optimization with sparsity-inducing penalties, *Foundations and Trends in Machine Learning*, **4**(1), 1–106, 2012.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[5] A. Blum and S. Chawla, Learning from labeled and unlabeled data using graph mincuts, In *Proc. of the 18th Int'l Conf. on Machine Learning (ICML'01)*, 19–26, 2001.

[6] Y. Boykov, O. Veksler and R. Zabih, Fast approximate energy minimization via graph cuts, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **23**(9), 1222–1239, 2001.

[7] A. Chechetka and C. Guestrin, Efficient principled learning of thin junction trees, In *Advances in Neural Information Processing Systems*, **20**, 273–280. 2008.

[8] A. Das, A. Dasgupta and R. Kumar, Selecting diverse features via spectral regularization, In *Advances in Neural Information Processing Systems*, **25**, 1592–1600. 2012.

[9] A. Das and D. Kempe, Algorithms for subset selection in linear regression, In *Proc. of the 40th Ann. ACM Symp. on Theory of Computing (STOC'08)*, 45–54, 2008.

[10] J. Edmonds, Submodular functions, matroids and certain polyhedra, In Richard K. Guy, H. Hanani, N. Sauer, and J. Schönheim, eds., *Combinatorial Structures and Their Applications*, 69–87. Gordon and Breach, 1970.

[11] S. Fujishige, *Submodular Functions and Optimization*, Elsevier, 2nd edition, 2005.

[12] A. V. Goldberg and R. E. Tarjan, A new approach to the maximum-flow problem, *Journal of the ACM*, **35**(4), 921–940, 1988.

[13] P. Hansen and B. Jaumard, Cluster analysis and mathematical programming, *Mathematical Programming*, **79**(1-3):191–215, 1997.

[14] T. Hastie, R. Tibshirani and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Verlag, 2009.

[15] E. Hazan and S. Kale, Beyond convexity: Online submodular minimization, In *Advances in Neural Information Processing Systems*, **22**, 700–708. 2009.

[16] H. Hoefling, A path algorithm for the fused lasso signal approximator, *Journal of Computational and Graphical Statistics*, **19**(4), 984–1006, 2010.

[17] S. C. H. Hoi, R. Jin, J. Zhu and M. R. Lyu, Batch mode active learning and its application to medical image classification, In *Proc. of the 23rd Int'l Conf. on Machine Learning (ICML'06)*, 417–424, 2006.

[18] H. Ishikawa, Transformation of general binary MRF minimization to the first-order case, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **33**(6), 1234–1249, 2011.

[19] R. Iyer and J. Bilmes, Algorithms for approximate minimization of the difference between submodular functions, with applications, In *Proc. of the 28th Ann. Conf. on Uncertainty in Artificial Intelligence (UAI'12)*, 407–417, 2012.

[20] R. Iyer and J. Bilmes, Submodular-Bregman and the Lovasz-Bregman divergences with applications, In *Advances in Neural Information Processing Systems*, **25**, 2942–2950. 2012.

[21] L. Jacob, G. Obozinski and J. P. Vert, Group Lasso with overlaps and graph Lasso, In *Proc. of the 26th Int'l Conf. on Machine Learning (ICML'09)*, 433–440, 2009.

[22] S. Jegelka, H. Liu and J. Bilmes, On fast approximate submodular minimization, In *Advances in Neural Information Processing Systems*, **24**, 460–468. 2011.

[23] Y. Kawahara, K. Nagano and Y. Okamoto, Submodular fractional programming for balanced clustering, *Pattern Recognition Letters*, **32**(2), 235–243, 2011.

[24] Y. Kawahara, K. Nagano, K. Tsuda and J. A. Bilmes, Submodularity cuts and applications, In *Advances in Neural Information Processing Systems*, **22**, 916–924. 2009.

[25] Y. Kawahara and T. Washio, Prismatic algorithm for discrete D. C. programming problems, In *Advances in Neural Information Processing Systems*, **24**, 2106–2114. 2011.

[26] V. Kolmogorov, Y. Boykov and C. Rother, Applications of parametric maxflow in computer vision, In *Proc. of the 11th Int'l Conf. on Computer Vision (ICCV'07)*, 2007.

[27] V. Kolmogorov and R. Zabih, What energy functions can be minimized via graph cuts?, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **26**(2), 147–159, 2004.

[28] A. Krause, H. McMahan, C. Guestrin and A. Gupta, Robust submodular observation selection, *Journal of Machine Learning Research*, **9**, 2761–2801, 2008.

[29] H. Lin and J. Bilmes, Multi-document summarization via budgeted maximization of submodular functions, In *Proc. of Human Language Technologies (HLT'10): The 2010 Ann. Conf. of the North American Chapter of the ACL*, 912–920, 2010.

[30] L. Lovász, Submodular functions and convexity, In A. Bachem, M. Grötschel and B. Korte, eds., *Mathematical Programming – The State of the Art*, 235–257. 1983.

[31] U. V. Luxburg, A tutorial on spectral clustering, *Statistics and Computing*, **17**(4), 395–416, 2007.

[32] K. Murota, *Discrete Convex Analysis*, Monographs on Discrete Math and Applications. SIAM, 2003.

[33] K. Nagano, Y. Kawahara and K. Aihara, Size-constrained submodular minimization through minimum norm base, In *Proc. of the 28th Int'l Conf. on Machine Learning (ICML'11)*, 977–984, 2011.

[34] K. Nagano, Y. Kawahara and S. Iwata, Minimum average cost clustering, In *Advances in Neural Information Processing Systems*, **23**, 1759–1767. 2010.

[35] M. Narasimhan and J. A. Bilmes, PAC-learning bounded tree-width graphical models, In *Proc. of the 20th Ann. Conf. on Uncertainty in Artificial Intelligence (UAI'04)*, 410–417, 2004.

[36] M. Narasimhan and J. A. Bilmes, A submodular-supermodular procedure with applications to discriminative structure learning, In *Proc.of the 21st*

Ann. Conf. on Uncertainty in Artificial Intelligence (UAI'05), 404–412, 2005.

[37] M. Narasimhan, N. Jojic and J. A. Bilmes, Q-clustering, In *Advances in Neural Information Processing Systems*, **18**, 979–986. 2006.

[38] N. Narasimhan and J. A. Bilmes, Local search for balanced submodular clustering, In *Proc. of the 12th Int'l Joint Conference on Artificial Intelligence (IJCAI'07)*, 981–986, 2007.

[39] G. L. Nemhauser and L. A. Wolsey, *Maximizing Submodular Set Functions: Formulation and Analysis of Algorithms*, Annals of Discrete Mathematics. 1981.

[40] B. Steudel, D. Janzing and B. Scholkopf, Causal markov condition for submodular information measures, In *Proc. of the 23rd Ann. Conf. on Learning Theory (COLT'10)*, 464–476. Omnipress, 2010.

[41] M. Thoma, H. Cheng, A. Gretton, J. Han, H. Kriegel, A. Smola, S. Le Song Philip, X. Yan and K. Borgwardt, Near-optimal supervised feature selection among frequent subgraphs, In *Proc. of the 2009 SIAM Conf. on Data Mining (SDM'09)*, 1076–1087, 2009.

[42] R. Tibshirani, Regression shrinkage and selection via the Lasso, *Journal of Royal Statistical Society B*, **58**(1), 267–288, 1996.

[43] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B*, **68**, 49–67, 2006.

[44] T. Zhang, Adaptive forward-backward greedy algorithm for sparse learning with linear models, In *Advances in Neural Information Processing Systems*, **21**, 1921–1928. 2008.

