第18回情報論的学習理論ワークショップ

劣モジュラ関数による構造と学習の橋渡し：

構造正則化，確率的劣モジュラ

大阪大学 産業科学研究所　河原　吉伸

Email: ykawahara@sanken.osaka-u.ac.jp
Web:　http://www.ar.sanken.osaka-u.ac.jp/~kawahara/jp/

劣モジュラ関数

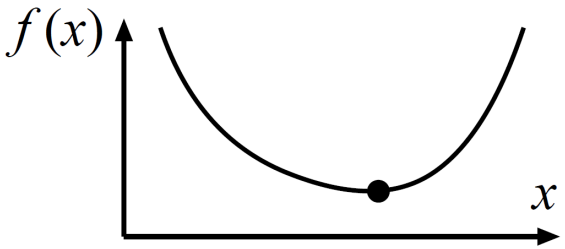
次式を満たす集合関数 $F : 2^V \to \mathbb{R}$ を劣モジュラ関数と呼ぶ:

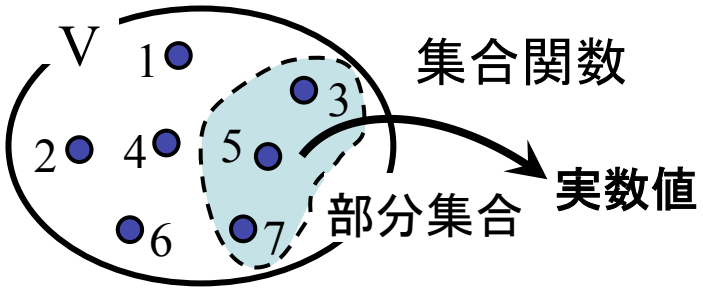
[定義] $$F(S) + F(T) \geq F(S \cup T) + F(S \cap T) \quad (\forall S, T \subseteq V)$$

⇒ 集合関数における凸関数として捉えられる（ただし凹的な性質も持つ）

凸性（連続関数） ⟷ 対応する概念 ⟷ 劣モジュラ性（集合関数）

自然に現れる構造

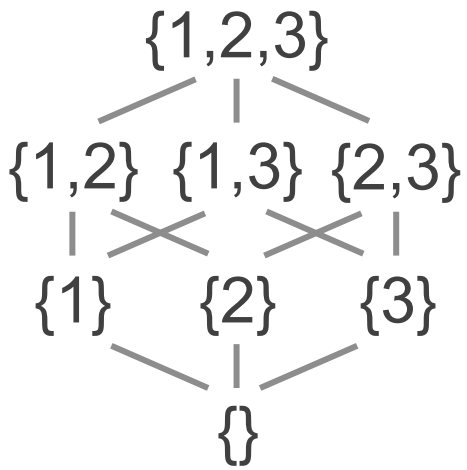




$$F(S) + F(T) \leq F(S \cap T) + F(S \cup T)$$
$$(S, T \subseteq V)$$

補足）べき集合と特性ベクトル

べき集合



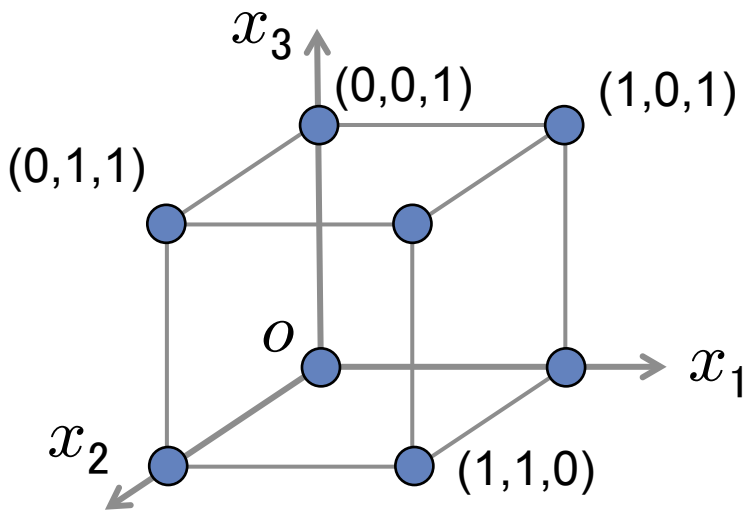
2^V

$\cup$

部分集合　S

$\approx$

（対応する $\{0,1\}^d$ の空間）



$\{0,1\}^d$

$\cup$

特性ベクトル　I_S

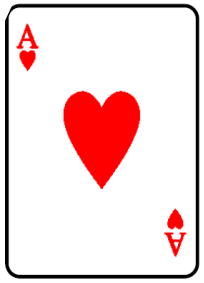
つまり，集合関数は $\{0,1\}^d$ 上の実数値関数とも見なせる

第 i 成分 $= \begin{cases} 1 & \text{if} \quad i \in S \\ 0 & \text{if} \quad i \notin S \end{cases}$

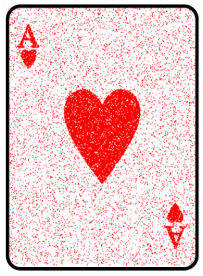
例）マルコフ確率場と劣モジュラ性

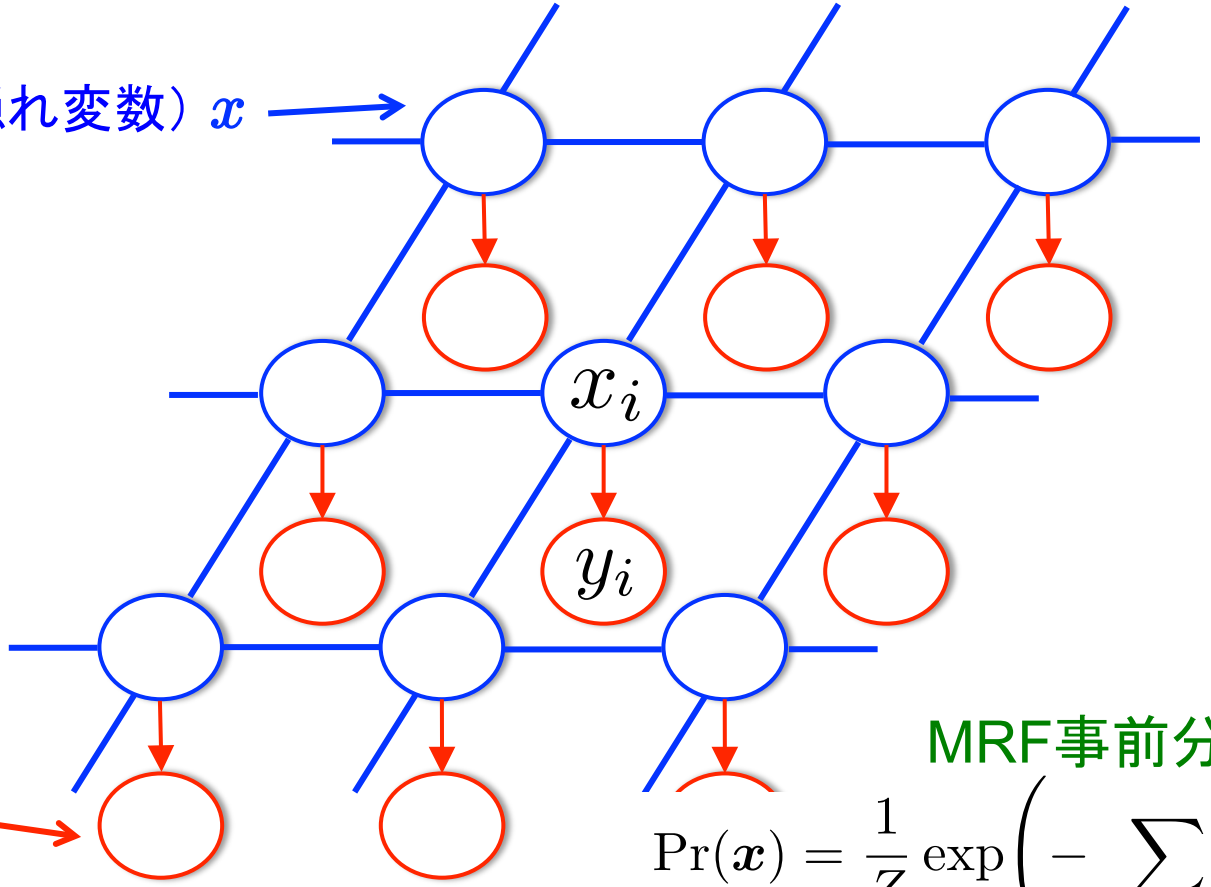
機械学習や統計分野などでよく用いられる構造的な確率モデル：

元の画像（隠れ変数） x



観測画像 y





MRF事前分布

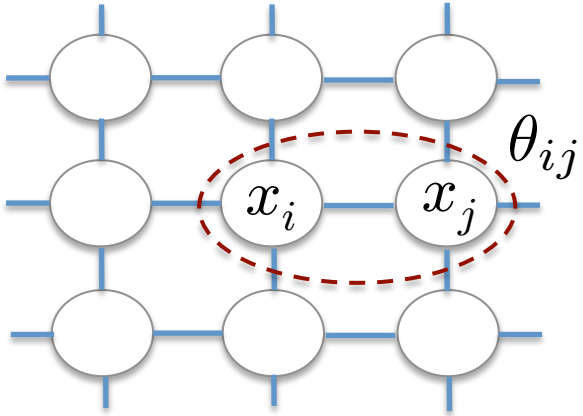
$$\Pr(\boldsymbol{x}) = \frac{1}{Z} \exp\left(- \sum_{(i,j)\in\mathcal{E}} \tilde{\theta}_{ij}(x_i, x_j) \right)$$

例）マルコフ確率場と劣モジュラ性

MRFの最大事後確率（MAP）推定：

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}\in\{0,1\}^V} \sum_{k=1}^{N} \Pr(\boldsymbol{x}|\boldsymbol{y}^{(k)})$$

k番目の観測



$$= \arg\max_{\boldsymbol{x}\in\{0,1\}^V} \sum_{k=1}^{N} \prod_{i\in V} \Pr(y_i^{(k)})|x_i) \Pr(\boldsymbol{x})$$

$$= \arg\max_{\boldsymbol{x}\in\{0,1\}^V} \left(N \log(\Pr(\boldsymbol{x})) + \sum_{k=1}^{N} \sum_{i\in V} \log(\Pr(y_i^{(k)}|x_i)) \right)$$

$$= \arg\min_{\boldsymbol{x}\in\{0,1\}^V} \left(N \sum_{(i,j)\in\mathcal{E}} \tilde{\theta}_{ij}(x_i, x_j) + N \log Z + \sum_{i\in V} \sum_{k=1}^{N} -\log(\Pr(y_i^{(k)}|x_i)) \right)$$

pairwise terms　　unary terms

等価

$$\min_{\boldsymbol{x}\in\{0,1\}^V} \sum_{i\in V} \theta_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \theta_{ij}(x_i, x_j)$$

エネルギー最小化

例）マルコフ確率場と劣モジュラ性

- 1階エネルギー最小化：

$$\min_{\boldsymbol{x}\in\{0,1\}^V} \sum_{i\in V} \theta_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \theta_{ij}(x_i, x_j)$$

⇒ 一般に「NP困難」

- 1階のエネルギー関数における劣モジュラ性：

$$\theta_{ij}(1,0) + \theta_{ij}(0,1) \geq \theta_{ij}(1,1) + \theta_{ij}(0,0)$$

一種のスムースネス
⇒ 物理的，応用的に妥当な仮定

（attractiveなイジング模型などはこの特殊ケース）

- 劣モジュラな1階エネルギー関数最小化は効率的に解ける

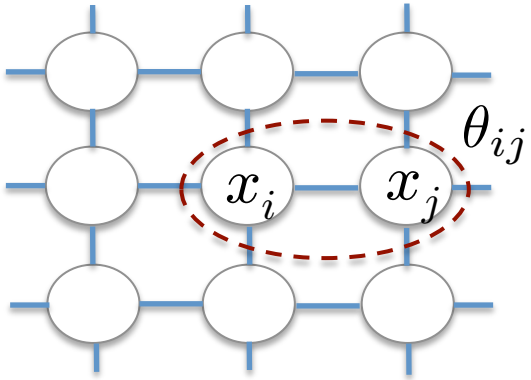
⇒ CVで用いられるグラフカット

例）マルコフ確率場と劣モジュラ性

劣モジュラ関数は，グラフ上の隣接構造よりもさらにリッチな表現が可能（ピクセル間の平滑性 ⇒ スーパーピクセル内の平滑性）．

ピクセル間の平滑性

$$\theta_{ij}(1,0)+\theta_{ij}(0,1) \geq \theta_{ij}(1,1)+\theta_{ij}(0,0)$$



$$\min_{\boldsymbol{x}\in\{0,1\}^V} \sum_{i\in V}\theta_i(x_i)+\sum_{(i,j)\in\mathcal{E}}\theta_{ij}(x_i,x_j)$$

ピクセル間の平滑性 ＋ 色の情報 ＋ ⋮ ＋

(Kohli+ 09) より

スーパーピクセルの平滑性

$$\min_{\boldsymbol{x}\in\{0,1\}^V} \sum_{i\in V}\theta_i(x_i)+\sum_{(i,j)\in\mathcal{E}}\theta_{ij}(x_i,x_j)+\sum_{c\in\mathcal{C}}\theta_c(\boldsymbol{x}_c)$$

この場合でも，（一般化された）グラフカットで高速に解ける
⇒ より一般にどういう分布を用いて学習に用いることができるのか？

本講演のトピック

今日の講演では，MRFとカット関数（劣モジュラ関数）との関係に代表される，劣モジュラ関数を用いた構造的学習について，特に

- *劣モジュラ関数と，構造的な学習モデルとの関係は？*
- *この関係を，どのようにして学習に用いることができるのか？*

という点に注目する．そして，その具体的なアプローチとしての

- 「劣モジュラ関数を用いた構造正則化学習」
- 「劣モジュラ関数から得られる確率分布を用いた学習」

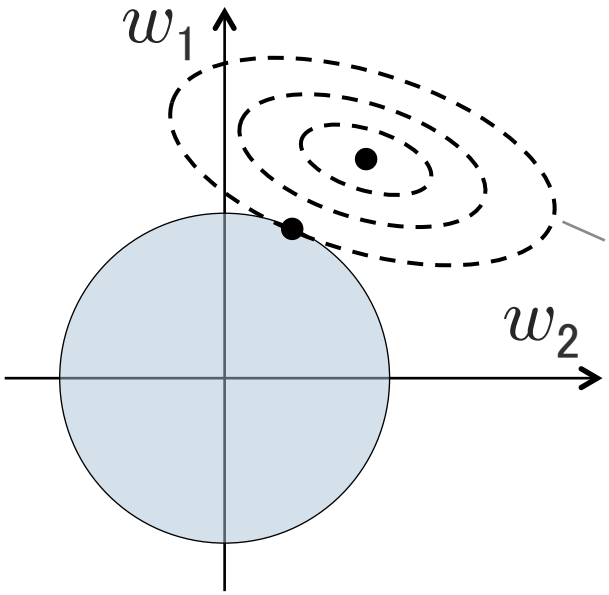
に焦点をあてて話を進める．

正則化とスパース性

l_pノルムによる正則化：$\min_{\boldsymbol{w}\in\mathbb{R}^d} l(\boldsymbol{w}) + \lambda \cdot \|\boldsymbol{w}\|_p^p$

l_2-正則化（$p=2$）：

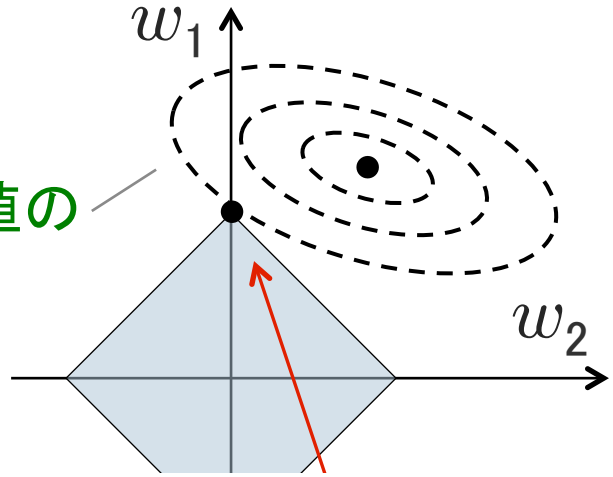
$$\|\boldsymbol{w}\|_2^2 = w_1^2 + \cdots + w_d^2$$



損失関数値の
等位集合

l_1-正則化（$p=1$）：

$$|\boldsymbol{w}| = |w_1| + \cdots + |w_d|$$



解が軸に乗りやすいため，
疎な解が得られやすい

構造正則化

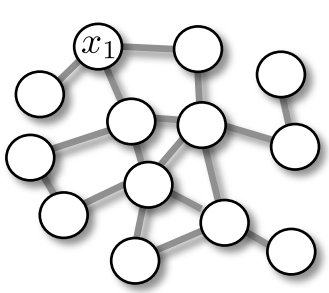
正則化の考え方を拡張して，既知の構造的関係に近づくように，
パラメータが推定されるような正則化項を設計して加える：

構造正則化項

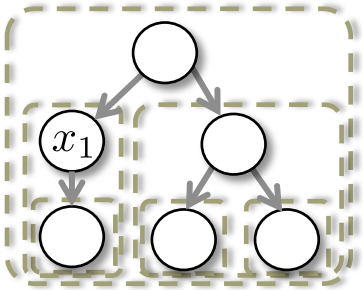
$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \; l(\boldsymbol{w}) + \lambda \cdot \Omega(\boldsymbol{w})$$

損失関数

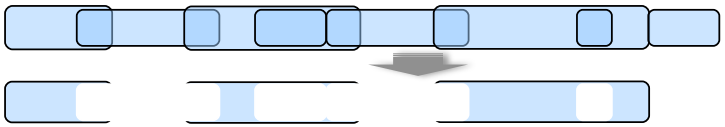
こういった変数間の組合せ的
構造を正則化項として組込む



グラフ構造



階層構造



グループ構造

その他，有向グラフ上のパスや，2次元グリッド上でのブロック構造など．

構造正則化の例

これまでに様々なタイプの構造正則化が提案され応用されてきた:

- グループ lasso (Yuan & Lin 07)
  - 与えたグループ単位でスパース性が得られる正則化. グループを階層的に構成した階層的 lasso (Jenatton+ 11) や, DAG上のパスを形成するよう構成したパス・コーディング (Mairal & Yu 13) などの拡張がある.
- (隠れ)グループ lasso (Jacob+ 09)
  - 与えたグループの補集合単位でのスパース性が得られる正則化. グラフ上で隣接する2変数をグループとする`グラフ lasso'も含まれる.
- 結合 lasso (Tibshirani+ 05), 一般化結合 lasso (Tibshirani & Taylor 11)
  - グラフ上で隣接するノードに対応する2変数の値が近くなるようなスパース性が得られる正則化. ハイパーグラフを用いた高階の場合への拡張 (Takeuchi+ 15) などがある.

– その他にも, 種々の応用中の構造に対する正則化が提案されている.

劣モジュラ関数を用いた正則化

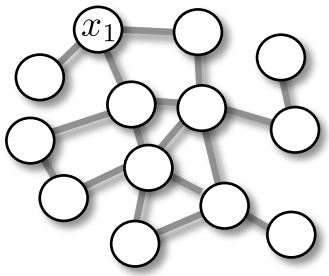
- 近年，多くの構造的スパース性が，劣モジュラ関数（の緩和）を正則化として用いて得られることが指摘される (Bach 10, 11).

構造正則化項

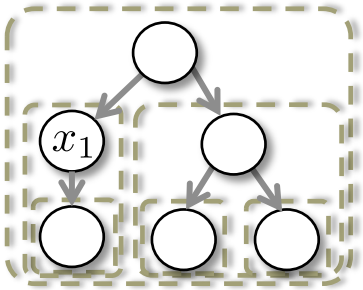
$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \; l(\boldsymbol{w}) + \lambda \cdot \Omega(\boldsymbol{w})$$

損失関数

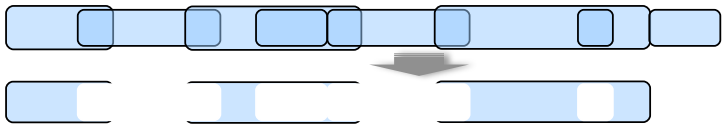
劣モジュラ関数の連続緩和



グラフ構造



階層構造



グループ構造

その他，有向グラフ上のパスや，2次元グリッド上でのブロック構造など.

劣モジュラ関数に基づく議論の利点

- 劣モジュラ関数という一般的な枠組みの中で議論が可能

⇒
  - 種々の構造を，統一的な枠組みの中で扱うことができる（最終的には，実装なども共通化できる）.
  - 組合せ最適化分野で議論されてきた，様々な理論やアルゴリズムが利用できる.

- 特に，劣モジュラ関数がグラフ表現可能な場合には，学習に伴う最適化が高速に計算可能となる.

例）結合正則化

一般化結合（generalized fused）**正則化** (Tibshirani & Taylor 11)：

各変数に対応するノードを持つ（有向）グラフ$G=(E,V)$上で，隣接する変数の値が近くなるような正則化．

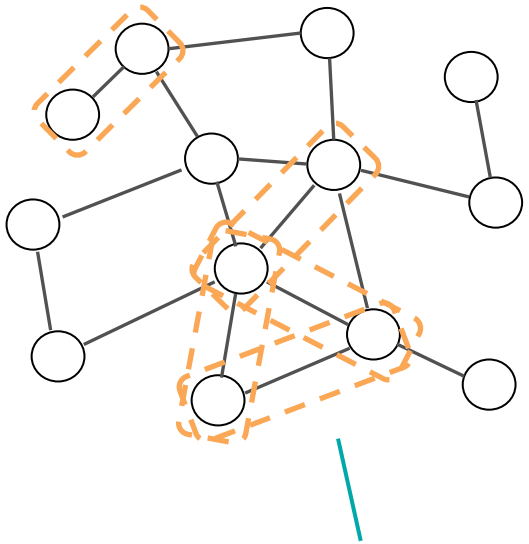
$$\Omega_{F,\infty}(\boldsymbol{w}) = \sum_{(i,j)\in E} a_{ij}|w_i - w_j|$$

隣接行列の要素

||　（等価）

カット関数のl_∞ノルムによる緩和（ロヴァース拡張）

$$F(S) = \sum_{(i,j)\in E} \{a_{ij} : i \in S, j \in V \setminus S\}$$



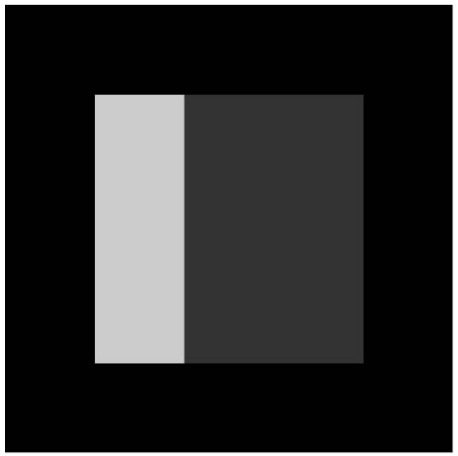
隣接する変数に関する係数が近い値になる

一階エネルギー関数におけるスムースネスを課すのと等価

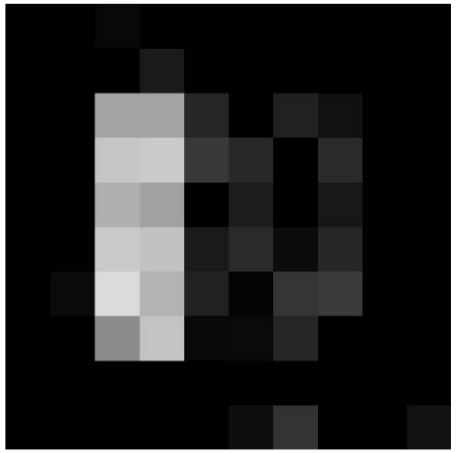
例）結合正則化

数値例：左図のような値のパラメータβを用いて，$y = \mathbf{X}\beta + \varepsilon$（ノイズ）のようにデータを生成.* 中図：ラッソによる推定．右図：2次元格子状グラフを用いた一般化結合正則化（$+l_1$項）による推定．

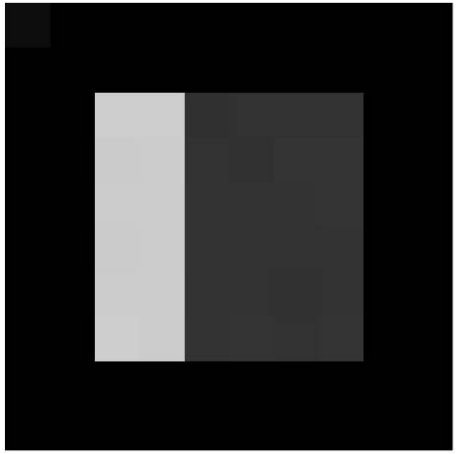
真のパラメータ

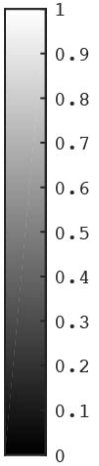


ラッソによる推定



一般化結合正則化による推定





*) 訓練データは80サンプル．ノイズは標準偏差0.05のガウス雑音．
正則化パラメータは5-fold CVにより選択．

例）グループ型の正則化

グループ型の正則化:

変数上に，グループ構造 $\mathcal{G}$（各要素が V の部分集合）が与えられたときに，
各グループ内の変数が同時にゼロになりやすくなるような正則化．

$$\Omega_{F,p}(\boldsymbol{w}) = \sum_{g \in \mathcal{G}} d_g \|\boldsymbol{w}_g\|_p$$

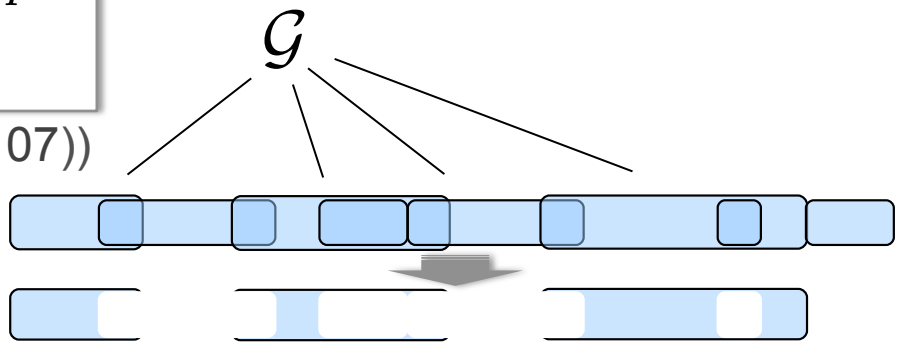
(グループ正則化 (Yuan & Lin 07))

≀≀ （ほぼ等価）

被覆関数のl_pノルムによる緩和:

$$F(S) = \sum \{d_g : g \in \mathcal{G}, g \cap S \neq \emptyset\}$$

グループへの重み

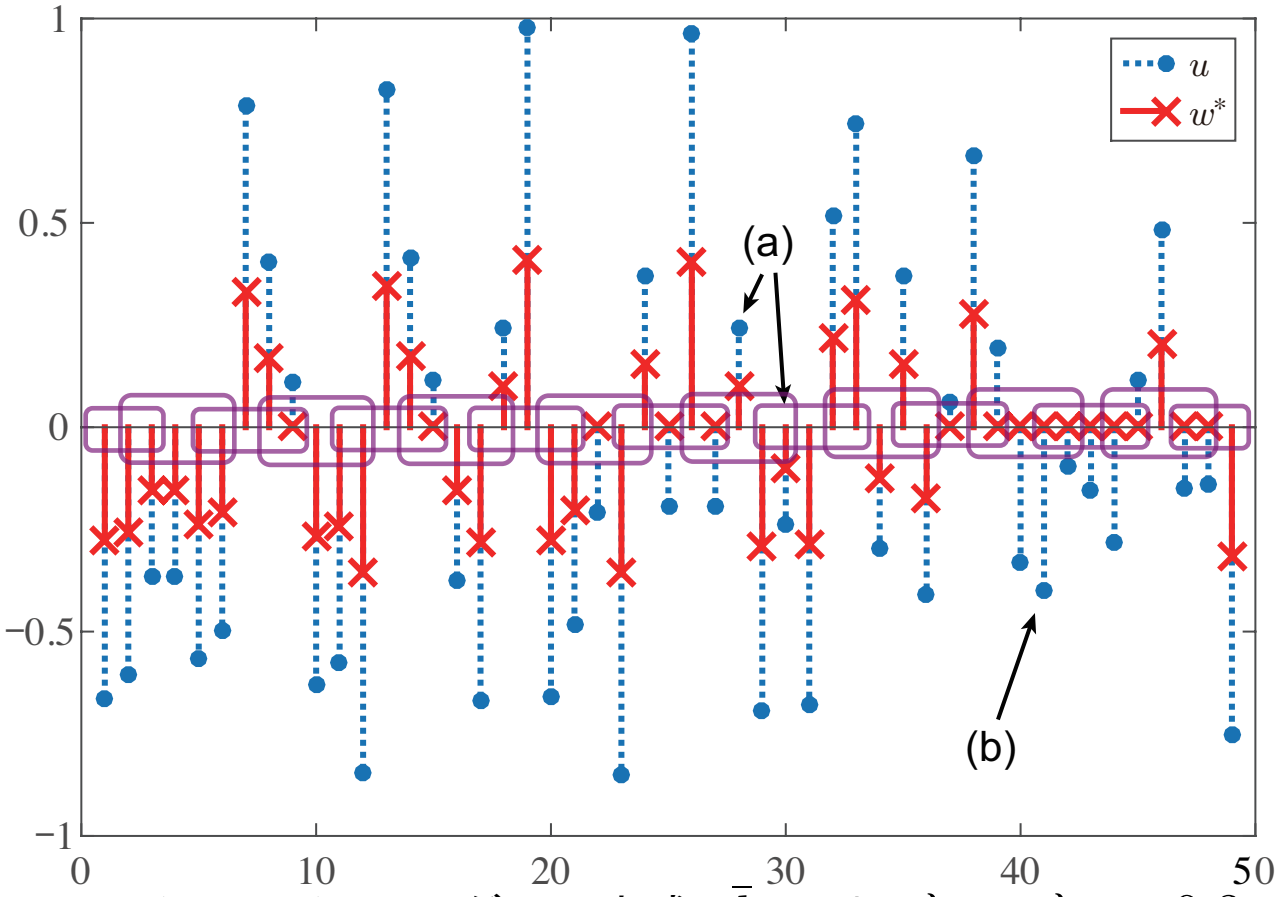


グループ内のものは同時に
ゼロになりやすい．

例）グループ型の正則化

数値例：グループは 1～3, 2～6, 5～9, 8～12, … のように設定.

$\min_{\boldsymbol{w}} \|\boldsymbol{u}-\boldsymbol{w}\| + \lambda_1|\boldsymbol{w}| + \lambda_2\Omega_{\text{group}}(\boldsymbol{w})$の計算例*



*) $\boldsymbol{u}$ は(−1,＋1)でランダムに生成 $\bar{d}_q = 2$, $\lambda_1 = \lambda_2 = 0.3$.

劣モジュラ関数のl_pノルムを用いた凸緩和

l_pノルムと，wのサポート上の単調劣モジュラ関数Fから成る関数

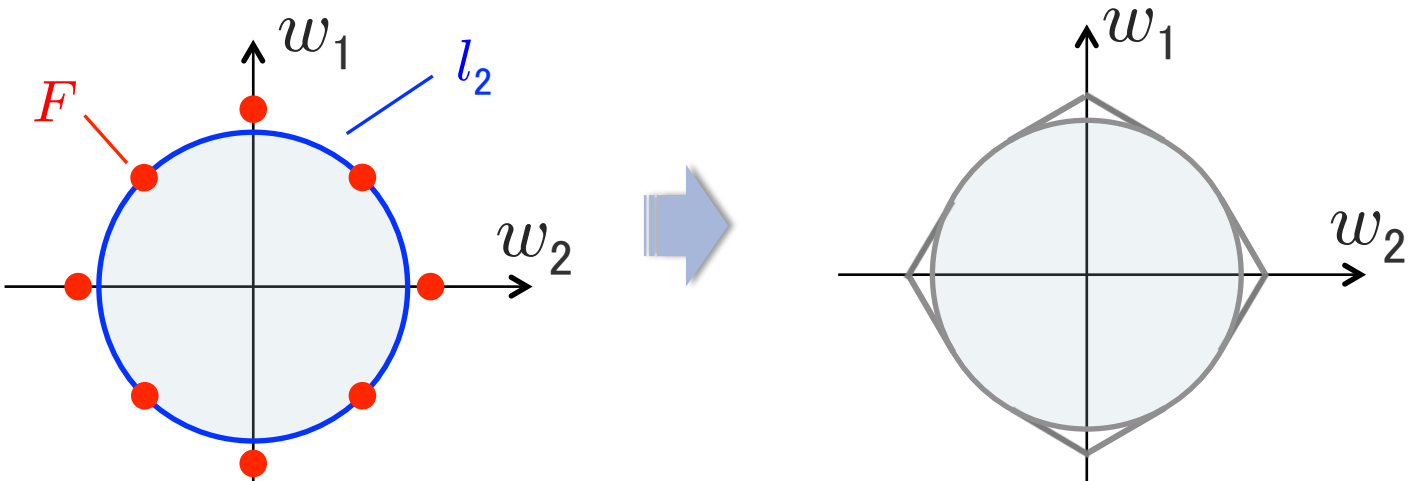
$$h(\boldsymbol{w}) = \frac{1}{p}\|\boldsymbol{w}\|_p^p + \frac{1}{r}F(\mathrm{supp}(\boldsymbol{w}))$$

のタイトな斉次凸下界（ただし，$1/p + 1/r = 1$）：

$$\tilde{\Omega}_{F,p}(\boldsymbol{w}) = \sup_{\boldsymbol{s}\in\mathbb{R}^V} \boldsymbol{w}^\top \boldsymbol{s} \quad \text{such that} \quad \|\boldsymbol{s}_S\|_r^r \le F(S)\ (\forall S \subseteq V)$$

(Obozinski & Bach 12)

例）$F(S) = |S|^{1/2}$の場合：



劣モジュラ関数から得られる構造正則化

劣モジュラ関数から得られる構造正則化の例:

- グループ型の正則化
  - グループlasso (Yuan & Lin 07) と同様のスパース性が，被覆関数の緩和により得られる (Obozinski & Bach 12). この拡張として，階層的構造 (Jenatton+ 11) や，DAG上のパスを形成するパス正則化 (Mairal & Yu 13) などの拡張が提案されている.
- 結合 lasso (Tibshirani+ 05), 一般化結合 lasso (Tibshirani & Taylor 11)
  - 無向グラフのカット関数のロヴァース拡張として得られる(Bach 11). ハイパーグラフを用いた高階への拡張 (Takeuchi+ 15) や，有向グラフを用いた拡張 (Bo+, in press) などがある.
- その他の正則化
  - スケールフリー・ネットワーク正則化 (Defazio & Caetano 12) など.

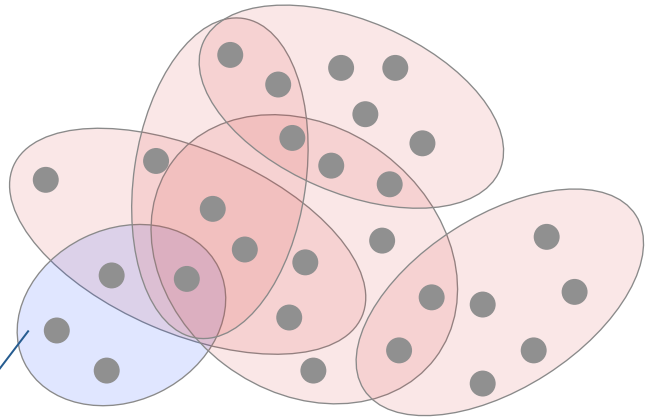
高階結合正則化

- 与えたグループ内（eg. 商品カテゴリ）のパラメータのとる値が，一致しやすくなるような正則化を用いる：

$$\Omega_{\mathrm{ho}}(\boldsymbol{\beta}) = \sum_{g \in G} \left(\sum_{i \in \{j_1, \ldots, j_{s-1}\}} (\beta_i - \beta_{j_s}) c^k_{1,i} + \beta_{j_s}(\theta^k_{\max} - \theta^k_1) + \beta_{j_t}(\theta^k_0 - \theta^k_{\max}) + \sum_{i \in \{j_{t+1}, \ldots, j_d\}} (\beta_{j_t} - \beta_i) c^k_{0,i} \right)$$

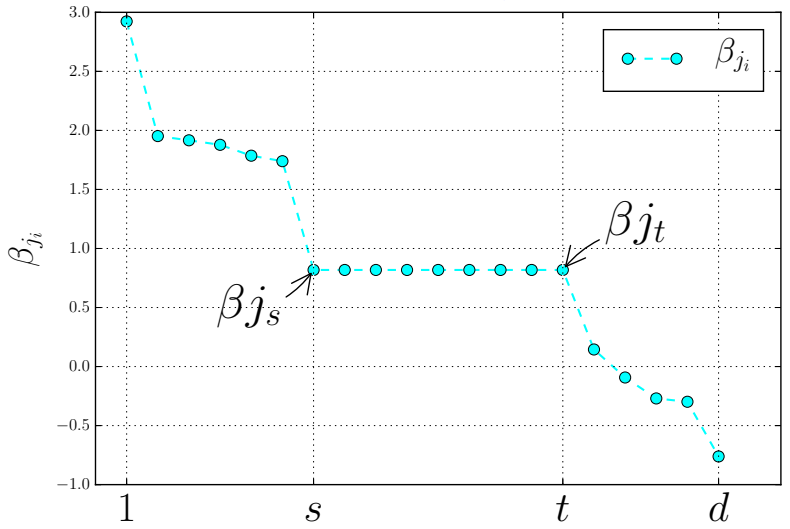
β をソートしたときの索引

このあたりは色々とパラメータ



こういったグループに含まれるパラメータの値が近くなりやすい

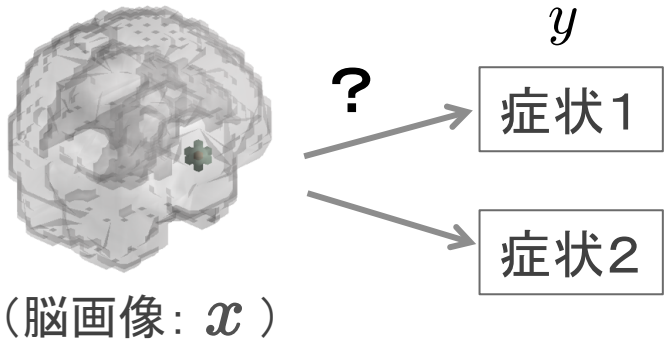
パラメータを大きい順に並べた図



K. Takeuchi, Y. Kawahara & T. Iwata, “Higher order fused regularization for supervised learning with grouped parameters,” in *Proc. of ECML-PKDD’15*, pp.577-593, 2015.

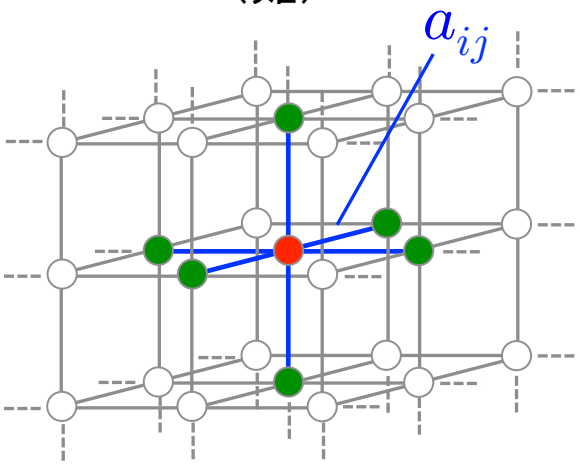
適用例: アルツハイマー病の分類

- 3D脳画像(MRI)を用いて，患者の症状の分類を行う:



(例: AD／NC, MCI_C／MCI_S)

(易) (難)



- 3次元格子状グラフの結合正則化項を加えたロジスティック回帰を用いる:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d, c\in\mathbb{R}} \sum_{i=1}^{n} \log(1+\exp(-y_i(\boldsymbol{w}^\top \boldsymbol{w}_i + c))) + \lambda_1|\boldsymbol{w}| + \lambda_2 \sum_{(i,j)\in\mathcal{E}} a_{ij}|w_i - w_j|$$

ロジスティック回帰の損失関数　　一般化Fused正則化項

B. Xin, Y. Kawahara, Y. Wang & W. Gao, "Efficient generalized fused Lasso with application to the diagnosis of Alzheimer's disease," in *Proc. of the 28th AAAI Conf. on Artificial Intelligence (AAAI'14)*, pp.2163-2169, 2014.

適用例: アルツハイマー病の分類

- データ: Alzheimer's Disease Neuroimaging Initiative (ANDI) database

  1.5T MRI (62 AD, 71 NC, 54 MCI_C, 87 MCI_S), および 3.0T MRI (66 AD, 104 ND, 8 MCI_C, 31 MCI_S) スキャンデータから構成.

- 前処理: DARTEL VBM pipeline*)を適用し, 0.2より大きいGMから成る2873個の8mm立方のボクセルを使用.

(10-CV平均の) 分類精度(Acc), 感度(Sens), 特異度(Spec)の比較:

	15ADNC			30ADNC			15MCI			30MCI		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	82.71%	80.65%	84.51%	**89.41%**	75.76%	**98.08%**	67.38%	40.74%	**83.91%**	82.05%	25.00%	**96.77%**
MLDA	**84.21%**	**84.51%**	83.87%	84.11%	**82.69%**	86.36%	63.83%	**65.52%**	61.11%	64.10%	58.06%	87.50%
LR	80.45%	74.19%	85.92%	87.06%	81.82%	90.38%	63.83%	50.00%	72.41%	79.49%	25.00%	93.55%
L1	81.20%	75.81%	85.92%	87.65%	78.79%	93.27%	68.79%	48.15%	81.61%	87.18%	50.00%	**96.77%**
GSR	**84.21%**	80.65%	**87.32%**	88.82%	77.27%	96.15%	**70.92%**	50.00%	**83.91%**	**89.74%**	**62.50%**	**96.77%**

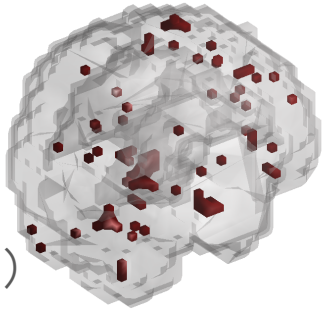
なお最近の医療画像解析分野で報告された方法のMCI分類性能は69.4% (B.Cheng+,12)

*) J. Ashburner et al., "A fast diffeomorphic image registration algorithm," *Neuroimage* **38**(1): 95-113, 2007.

適用例: アルツハイマー病の分類

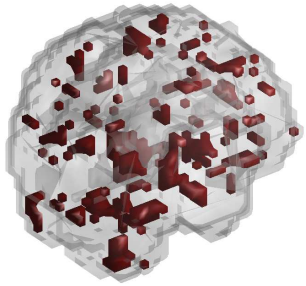
- 各スパース手法で（全データ10-CVにより）選択された脳部位（変数）:

構造なし（Lasso）

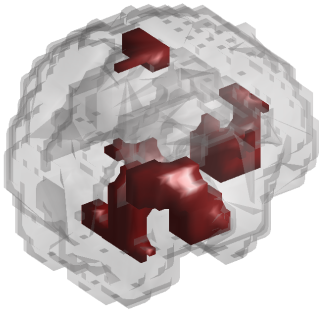


（最良）

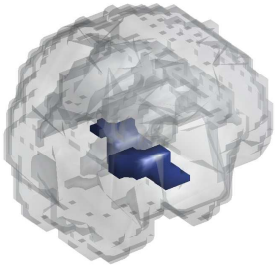
GFRと同数の
特徴数の場合



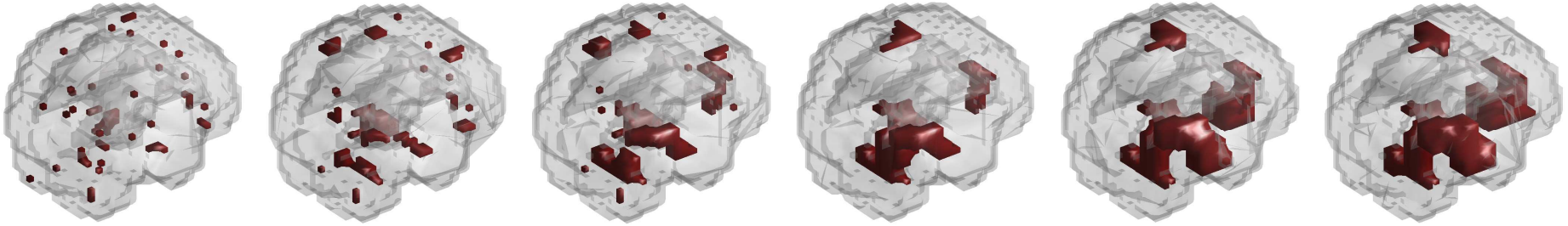
構造あり（GFL）



Top50の特徴



- 異なる一般化Fused項の程度と選択された脳部位:



(a) $\lambda_2 = 0$, 81.20%　(b) $\lambda_2 = 0.1$ 80.45%　(c) $\lambda_2 = 0.2$, 80.45%　(d) $\lambda_2 = 0.4$, 81.95%　(e) $\lambda_2 = 0.6$, 84.21%　(f) $\lambda_2 = 0.8$, 83.46%

λ_2 大

劣モジュラ関数に基づく議論の利点(再掲)

- 劣モジュラ関数という一般的な枠組みの中で議論が可能

⇒
  - 種々の構造を，統一的な枠組みの中で扱うことができる(最終的には，実装なども共通化できる)．
  - 組合せ最適化分野で議論されてきた，様々な理論やアルゴリズムが利用できる．

- 特に，劣モジュラ関数がグラフ表現可能な場合には，学習に伴う最適化が高速に計算可能となる．

最適化について

- 構造正則化では，微分可能な凸関数と，微分不可能な凸関数の和から成る評価関数を最小化する必要がある：

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \; l(\boldsymbol{w}) + \lambda \cdot \Omega_{F,p}(\boldsymbol{w})$$

微分可能な凸関数　　微分不可能な凸関数

- 一般的なアプローチ：

⇒ 近接勾配法，交互方向乗数法（ADMM）など

⇒ 近接演算子（proximity operator）の反復計算に帰着：

$$\mathrm{prox}_{\lambda\Omega_{F,p}}(\boldsymbol{u}) = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{w}\|_2^2 + \lambda\cdot\Omega_{F,p}(\boldsymbol{w}) \;\; (\boldsymbol{u}\in\mathbb{R}^d)$$

最適化について

近接演算子の計算:

- ロヴァース拡張で表される場合

  ⇒ 最小ノルム点アルゴリズムで計算可能 (Bach 10)

- 一般のlpノルムの場合

  ⇒ 分割アルゴリズム (Groenevelt 89) で計算可能
  (Obozinski & Bach 12)

- F がグラフ表現可能な場合(一般のlpノルム)

  ⇒ パラメトリック・フロー (Gallo+ 89) で計算可能
  (Mairal+11, Kawahara & Yamaguchi 15)

近接演算子の計算

$\Omega_{F,p}(\boldsymbol{w})$ に関する近接演算子の計算は，劣モジュラ多面体上の分離凸関数の最小化へ帰着できる (Kawahara & Yamaguchi 15)：

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{w}\|_2^2 + \lambda\cdot\Omega_{F,p}(\boldsymbol{w}) = \min_{\boldsymbol{w}\in\mathbb{R}^d}\max_{\boldsymbol{t}\in P_+(F)} \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{w}\|_2^2 + \lambda\sum_{i\in V} t_i^{1/r}|w_i|$$

$$= -\min_{\boldsymbol{t}\in P_+(F)}\sum_{i\in V} \underbrace{-\min_{w_i\in\mathbb{R}}\left\{\frac{1}{2}(w_i-u_i)^2 + \lambda t_i^{1/r}|w_i|\right\}}_{\text{凸関数}}$$

- 劣モジュラ多面体上での分離可能な凸関数の最小化
- F がグラフ表現可能な場合，パラメトリック最大流アルゴリズム（(Gallo+ 89)など）で高速に計算できる

グラフ表現可能な劣モジュラ関数

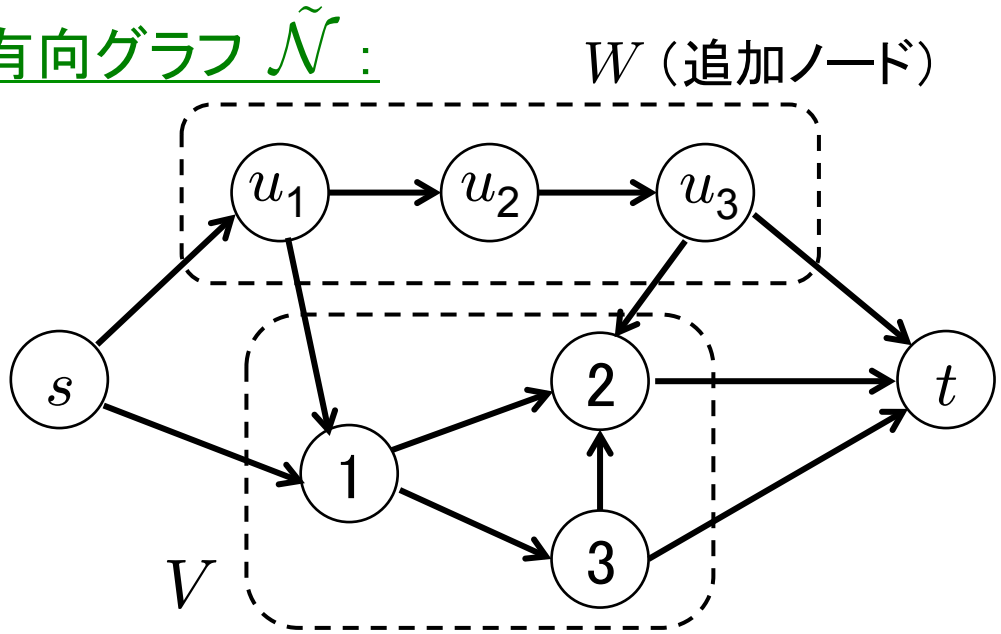
補助的な追加ノードを持つ s-t 有向グラフのカット関数として表現可能な劣モジュラ関数 (Jegelka+,11)

⇒ *フロー演算による高速な最小化が可能*

$$F(S) = \min_{Y \subseteq W} \kappa_{\tilde{\mathcal{N}}}(\{s\} \cup S \cup Y) + \text{const.}$$

有向グラフのカット関数

s-t 有向グラフ $\tilde{\mathcal{N}}$:

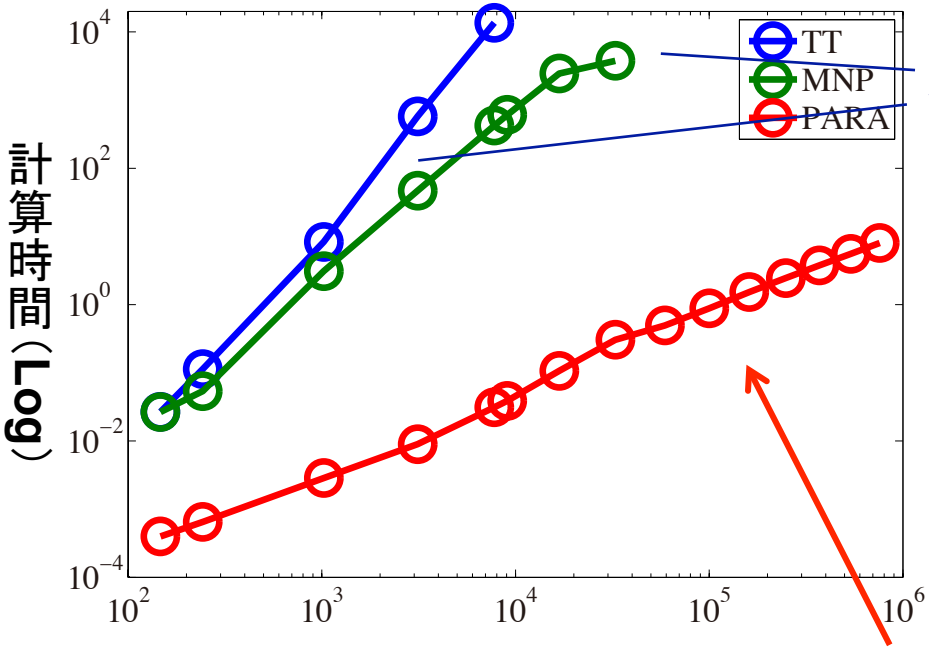


多くの構造正則化項を与える劣モジュラ関数はグラフ表現可能
(Kawahara & Yamaguchi 15)

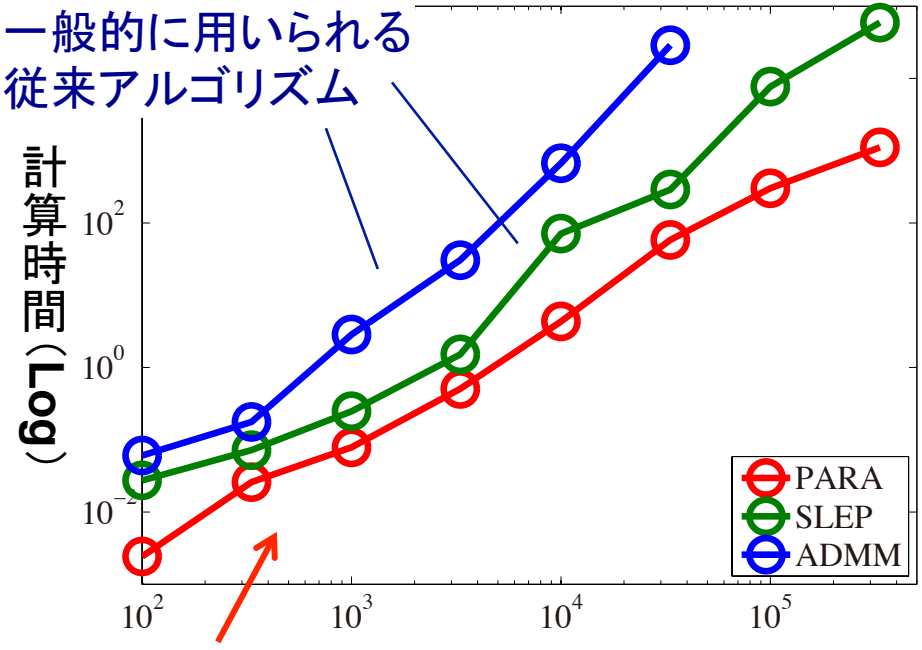
参考）近接演算子の計算時間の比較例

近接演算子の計算時間の比較例：

一般化結合正則化



グループ型正則化*



パラメトリック最大流による高速アルゴリズム

*) PARAと，SLEP・ADMMは，全く同じ疎性を与える正則化を計算している訳ではない

本講演のトピック

今日の講演では，MRFとカット関数（劣モジュラ関数）との関係に代表される，劣モジュラ関数を用いた構造的学習について，特に

- *劣モジュラ関数と，構造的な学習モデルとの関係は？*
- *この関係を，どのようにして学習に用いることができるのか？*

という点に注目する．そして，その具体的なアプローチとしての

- 「劣モジュラ関数を用いた構造正則化学習」
- 「劣モジュラ関数から得られる確率分布を用いた学習」

に焦点をあてて話を進める．

劣モジュラ関数から得られる確率分布

劣モジュラ関数が持つ離散構造を利用できる確率分布（対数劣モジュラ分布，対数優モジュラ分布）：

$$\Pr(S) = \frac{1}{\mathcal{Z}} \exp(\pm F(S))$$

劣モジュラ関数

正規化係数（分配関数）：
$\mathcal{Z} = \sum_{S \subseteq V} \exp(\pm F(S))$

- 多くの構造的な（2値ベクトル上の）確率分布を含む
- 劣モジュラ最適化による効率的な計算へ帰着できる場合がある

対数優モジュラ分布の例

対数優モジュラ分布: $\Pr(S) = \frac{1}{Z}\exp(-F(S))$

劣モジュラ関数

例）レギュラーなMRF

$$F(S) = \sum_{(i,j)\in E} \{a_{ij} : i \in S, j \in V \setminus S\}$$ の場合

隣接行列の要素

- 厳密なエネルギー最小化が可能なMRF
- より一般の劣モジュラ関数を用いることで，高階なMRFなどで同様の枠組みで扱うことができる.

対数劣モジュラ分布の例

対数劣モジュラ分布： $\Pr(S) = \frac{1}{Z}\exp(F(S))$

劣モジュラ関数

例）行列式点過程

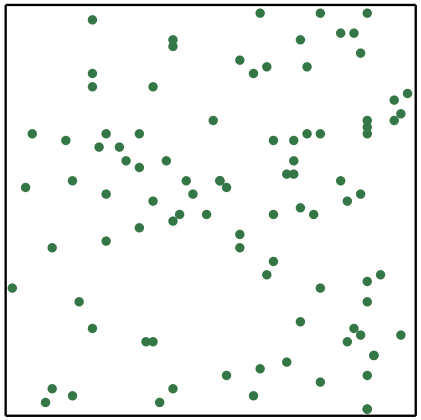
$$F(S) = \log|K_S| \quad \left[\begin{array}{l} K_S \text{は半正定値行列} K \text{の} \\ S \text{に対応する部分行列} \end{array}\right] \text{の場合}$$

- 行列式点過程自体は，古くから研究される離散分布 (Macchi 75) .
- （対数）劣モジュラ分布としての性質と，その一般化（劣モジュラ点過程）の議論はごく最近 (Iyer & Bilmes 15) .
- 行列式点過程は，機械学習へも利用される (Klesuza & Tasker 13).

対数劣モジュラ分布の例

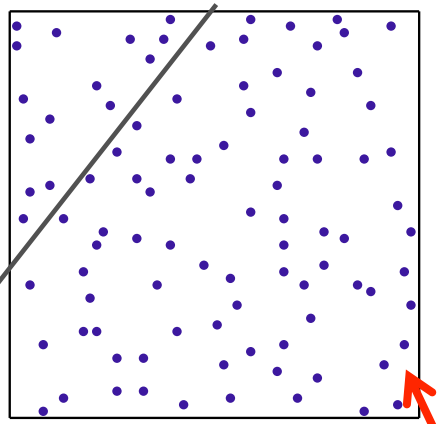
行列式点過程は，機械学習へも利用される (Klesuza & Tasker 13).

ランダムなサンプル: X



条件付き分布からのサンプル

$$\Pr(Y = S|X) \propto \det(L_S(X))$$



Diverse

$$L_{ij} = q_i(X)\phi_i(X)^\top \phi_j(X) q_j(X)$$

カーネル

$i \in V$ の価値 (given)

対数劣モジュラ分布の例

行列式点過程は，機械学習へも利用される (Klesuza & Tasker 13).

ランダムなサンプル：X

NASA and the Russian Space Agency have agreed to set aside a last-minute Russian request to launch an international space station into an orbit closer to Mir, officials announced Friday. . . .

A last-minute alarm forced NASA to halt Thursday's launching of the space shuttle Endeavour, on a mission to start assembling the international space station. This was the first time in three years . . .

The planet's most daring construction job began Friday as the shuttle Endeavour carried into orbit six astronauts and the first U.S.-built part of an international space station that is expected to cost more than $100 billion. . . .

Following a series of intricate maneuvers and the skillful use of the space shuttle Endeavour's robot arm, astronauts on Sunday joined the first two of many segments that will form the space station . . .

. . .

document cluster

多様性のある要約

On Friday the shuttle Endeavor carried six astronauts into orbit to start building an international space station. The launch occurred after Russia and U.S. officials agreed not to delay the flight in order to orbit closer to MIR, and after a last-minute alarm forced a postponement. On Sunday astronauts joining the Russian-made Zarya control module cylinder with the American-made module to form a 70,000 pounds mass 77 feet long. . . .

human summary

- NASA and the Russian Space Agency have agreed to set aside . . .
- A last-minute alarm forced NASA to halt Thursday's launching . . .
- This was the first time in three years, and 19 flights . . .
- After a last-minute alarm, the launch went off flawlessly Friday . . .
- Following a series of intricate maneuvers and the skillful . . .
- It looked to be a perfect and, hopefully, long-lasting fit. . . .

extractive summary

$$\Pr(Y = S|X) \propto \det(L_S(X))$$

推論と分配関数

周辺化計算:

$$\Pr(S \in [X,Y]) = \frac{1}{\mathcal{Z}} \sum_{X \subseteq A \subseteq Y} \exp(-F(A)) = e^{-F(X)} \frac{\mathcal{Z}_X^Y}{\mathcal{Z}}$$

$$\left[\text{ただし}, [X,Y] = \{A \mid X \subseteq A \subseteq Y\}, \mathcal{Z}_X^Y = \sum_{A \subseteq Y \setminus X} e^{-(F(X \cup A) - F(X))} \right]$$

条件付き確率:

$$\Pr(S = A \mid S \in [X,Y]) = \begin{cases} \exp(-F(A))/\mathcal{Z}_X^Y & \text{if } A \in [X,Y] \\ 0 & \text{otherwise} \end{cases}$$

⇒ 推論計算では，**分配関数**の計算が必要になる

⇒ 分配関数の計算は#P困難（近似は困難）

変分法によるアプローチ

- *モジュラ関数は，分配関数が厳密に計算可能*

分配関数に関する有用な性質：

モジュラ関数 l，u に対して，$\boldsymbol{l}(S) \leq F(S) \leq \boldsymbol{u}(S)\ (\forall S \subseteq V)$ なら

- $\sum_{S \subseteq V} \exp(+\boldsymbol{l}(S)) \leq \sum_{S \subseteq V} \exp(+F(S)) \leq \sum_{S \subseteq V} \exp(+\boldsymbol{u}(S))$
- $\sum_{S \subseteq V} \exp(-\boldsymbol{l}(S)) \geq \sum_{S \subseteq V} \exp(-F(S)) \geq \sum_{S \subseteq V} \exp(-\boldsymbol{u}(S))$

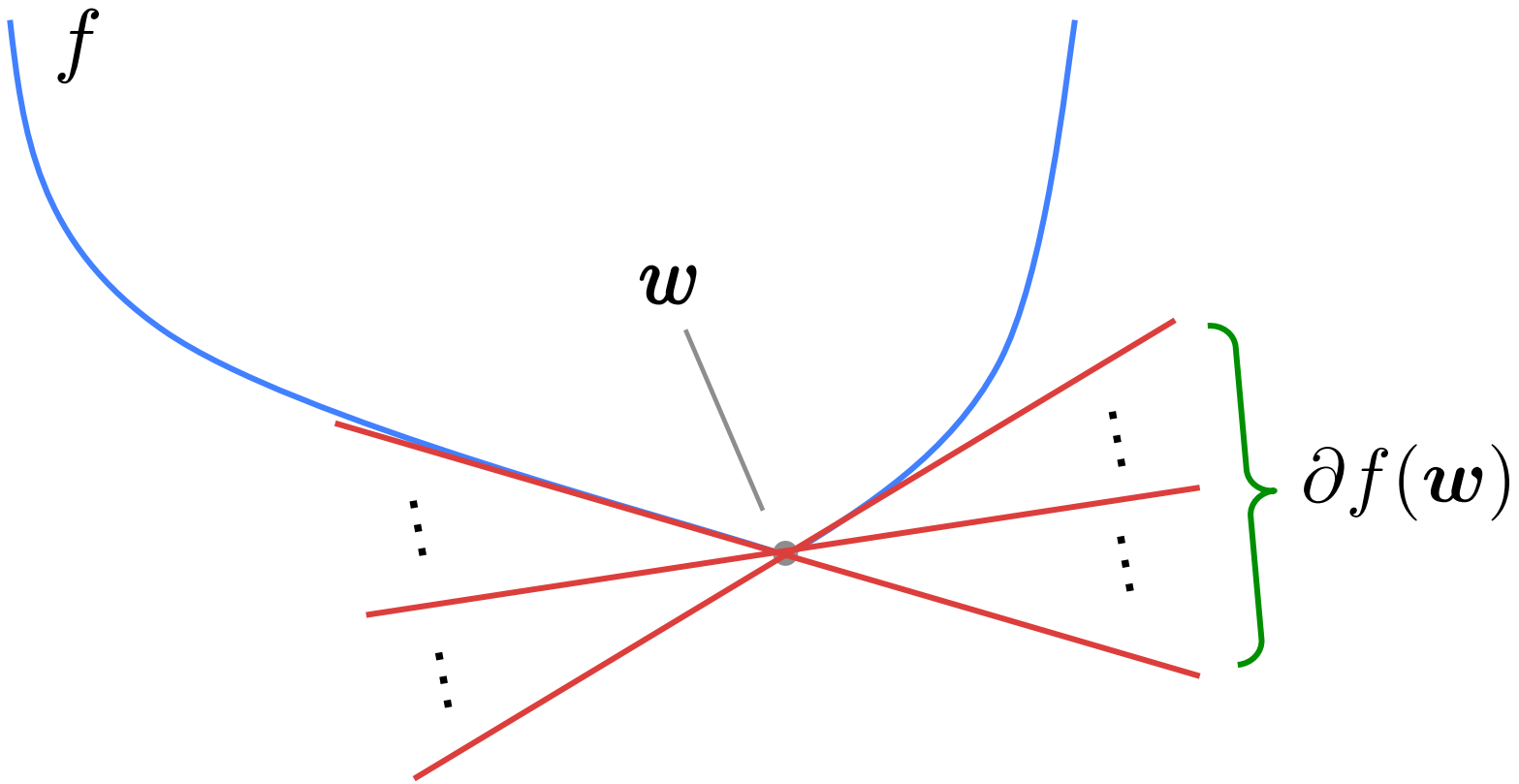
上と下から，モジュラ関数の分配関数でおさえられる

l，uとしては**劣勾配・優勾配**が利用可能

凸関数の劣勾配

凸関数 f の劣勾配:

$$\partial f(\boldsymbol{w}) = \{\boldsymbol{g} \mid f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + (\boldsymbol{u} - \boldsymbol{w})^\top \boldsymbol{g}\} \ (\forall \boldsymbol{u} \in \mathbb{R}^d)$$

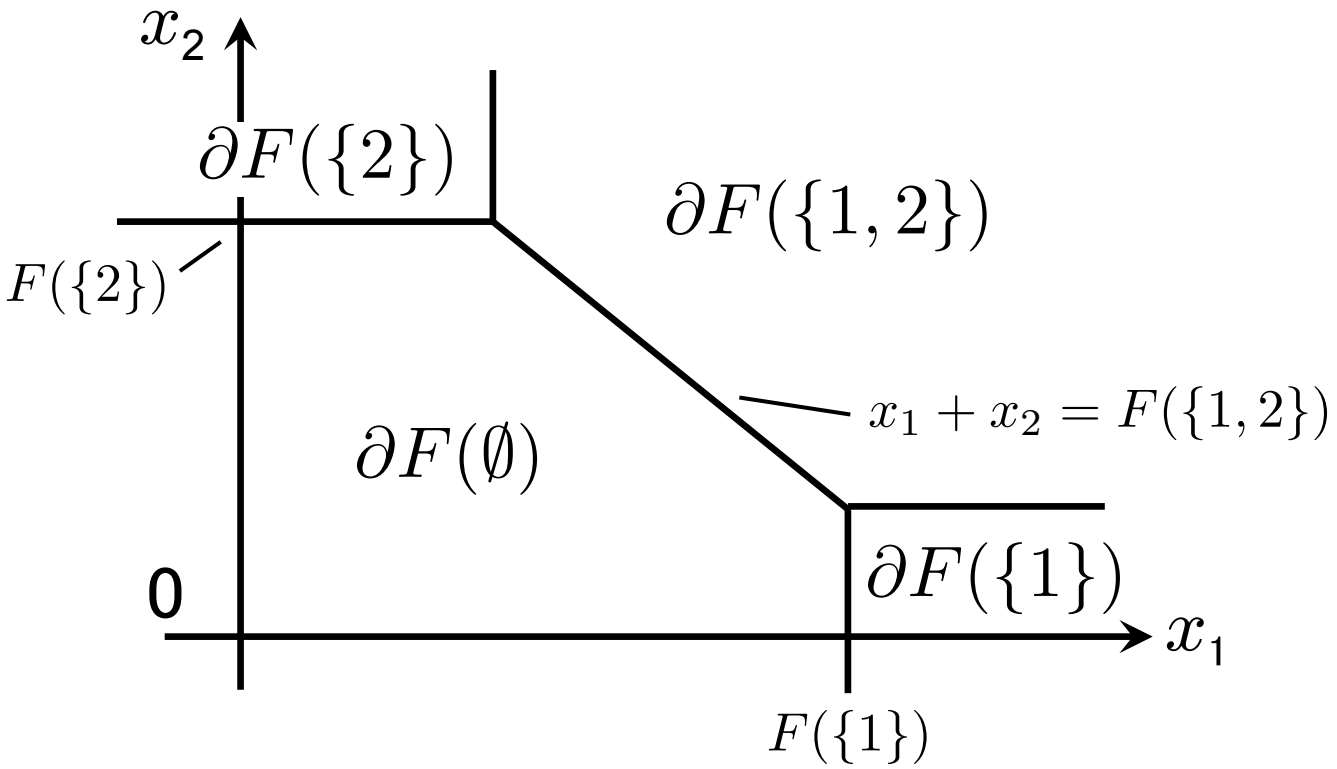


劣モジュラ関数の劣勾配

劣モジュラ関数に対しても劣勾配（モジュラ関数）が定義できる：

$$\partial F(X) = \{\boldsymbol{s} \in \mathbb{R}^d \mid F(Y) \geq F(X) + \boldsymbol{s}(Y) - \boldsymbol{s}(X), \forall Y \subseteq V\}$$

$|V|$=2の場合の概念図



変分法によるアプローチ（再掲）

- *モジュラ関数は，分配関数が厳密に計算可能*

分配関数に関する有用な性質：

モジュラ関数 l，u に対して，$\boldsymbol{l}(S) \leq F(S) \leq \boldsymbol{u}(S)\ (\forall S \subseteq V)$

- $\sum_{S \subseteq V} \exp(+\boldsymbol{l}(S)) \leq \sum_{S \subseteq V} \exp(+F(S)) \leq \sum_{S \subseteq V} \exp(+\boldsymbol{u}(S))$
- $\sum_{S \subseteq V} \exp(-\boldsymbol{l}(S)) \geq \sum_{S \subseteq V} \exp(-F(S)) \geq \sum_{S \subseteq V} \exp(-\boldsymbol{u}(S))$

上と下から，モジュラ関数の分配関数でおさえられる

l，uとしては優勾配・劣勾配が利用可能

変分法によるアプローチ（劣勾配）

劣勾配 $\boldsymbol{s} \in \partial_F(X)$ を用いると，任意の $X \subseteq V$ に対して

$$(\mathcal{Z} =) \sum_{S \subseteq V} \exp(-F(S)) \leq \underbrace{\sum_{S \subseteq V} \exp\left(-\boldsymbol{s}(S) + \boldsymbol{s}(X) - F(X)\right)}_{\mathcal{Z}_X^-(\boldsymbol{s})}$$

ただし，次の関係が知られている (Djolonga & Krause 14)：

$$\min_{\boldsymbol{s} \in \partial_F(\emptyset)} \mathcal{Z}_\emptyset^-(\boldsymbol{s}) \leq \min_{\boldsymbol{s} \in \partial_F(X)} \mathcal{Z}_X^-(\boldsymbol{s})$$

つまり，$X = \emptyset$ のときが最も良い上界を与える

最小ノルム点問題との関係

優モジュラ分布の
分配関数の上界

$$\min_{\boldsymbol{s}\in\partial_F(\emptyset)} \mathcal{Z}_{\emptyset}^{-}(\boldsymbol{s})$$

等価
(Djolonga & Krause 14)

基多面体上での分離強凸最小化

$$\min_{\boldsymbol{s}\in B(F)} \sum_{i\in V} \log(1+e^{-s_i})$$

等価
(Nagano & Aihara 13)

最小ノルム点アルゴリズム
(Fujishige+ 06) *が適用可能*
（現時点では，実用的に最も高速な
（一般の）劣モジュラ関数の最小化
アルゴリズム）

最小ノルム点問題

$$\min_{\boldsymbol{s}\in B(F)} \sum_{i\in V} s_i^2$$

平均場近似との関係

優モジュラ分布の
分配関数の上界

$$\min_{\boldsymbol{s}\in\partial_F(\emptyset)} \mathcal{Z}_{\emptyset}^{-}(\boldsymbol{s})$$

等価
(Djolonga & Krause 14)

基多面体上での分離強凸最小化

$$\min_{\boldsymbol{s}\in B(F)} \sum_{i\in V} \log(1+e^{-s_i})$$

Renyiダイバージェンス最小化

$$\min_{Q} D_{\infty}(P||Q)$$

$$\left(D_{\infty}(P||Q) = \log \sup_{S\subseteq V} \frac{P(S)}{Q(S)} \right)$$

等価

$$\hat{F}(\boldsymbol{q}) = \sup_{\boldsymbol{s}\in B(F)} \boldsymbol{s}^{\top}\boldsymbol{q}$$

（強）双対
(Djolonga & Krause 15)

L-fields

$$\max_{\boldsymbol{q}\in[0,1]^V} \mathbb{H}[\boldsymbol{q}] - \hat{F}(\boldsymbol{q})$$

独立なベルヌーイ確率変数のエントロピー

$$\left[P(S) = \frac{1}{\mathcal{Z}_p}\exp(-F(S)),\ Q(S) = \frac{1}{\mathcal{Z}_q}\exp(-\boldsymbol{q}(S)) \right]$$

平均場近似との関係

$$E_{\boldsymbol{q}}[F] = \sum_{S \subseteq V} \prod_i q_i^{[i \in S]} (1 - q_i)^{[i \notin S]} F(S)$$

優モジュラ分布の
分配関数の上界

$$\min_{\boldsymbol{s} \in \partial_F(\emptyset)} \mathcal{Z}_{\emptyset}^{-}(\boldsymbol{s})$$

等価
(Djolonga & Krause 14)

基多面体上での分離強凸最小化

$$\min_{\boldsymbol{s} \in B(F)} \sum_{i \in V} \log(1 + e^{-s_i})$$

Renyiダイバージェンス最小化

$$\min_Q D_\infty(P||Q)$$

$$\left(D_\infty(P||Q) = \log \sup_{S \subseteq V} \frac{P(S)}{Q(S)} \right)$$

等価

$$\hat{F}(\boldsymbol{q}) = \sup_{\boldsymbol{s} \in B(F)} \boldsymbol{s}^\top \boldsymbol{q}$$

（強）双対
(Djolonga & Krause 15)

L-fields

$$\max_{\boldsymbol{q} \in [0,1]^V} \mathbb{H}[\boldsymbol{q}] - \hat{F}(\boldsymbol{q})$$

独立なベルヌーイ確率変数のエントロピー

KLの最小化

$$\min_Q D_{KL}(P||Q)$$

等価

平均場近似

$$\max_{\boldsymbol{q} \in [0,1]^V} \mathbb{H}[\boldsymbol{q}] - E_{\boldsymbol{q}}[F]$$

平均場近似との関係

優モジュラ分布の
分配関数の上界

$$\min_{\boldsymbol{s}\in\partial_F(\emptyset)} \mathcal{Z}_{\emptyset}^{-}(\boldsymbol{s})$$

等価
(Djolonga & Krause 14)

基多面体上での分離強凸最小化

$$\min_{\boldsymbol{s}\in B(F)} \sum_{i\in V} \log(1+e^{-s_i})$$

Renyiダイバージェンス最小化

$$\min_{Q} D_\infty(P||Q)$$

$$\left(D_\infty(P||Q) = \log \sup_{S\subseteq V} \frac{P(S)}{Q(S)}\right)$$

等価

$$\hat{F}(\boldsymbol{q}) = \sup_{\boldsymbol{s}\in B(F)} \boldsymbol{s}^\top \boldsymbol{q}$$

（強）双対
(Djolonga & Krause 15)

L-fields

$$\max_{\boldsymbol{q}\in[0,1]^V} \mathbb{H}[\boldsymbol{q}] - \hat{F}(\boldsymbol{q})$$

独立なベルヌーイ確率変数のエントロピー

⇒ $D_\infty(P||Q)$ に基づいた期待値伝搬法などの設計も可能となる

変分法によるアプローチ（再掲）

- *モジュラ関数は，分配関数が厳密に計算可能*

分配関数に関する有用な性質：

モジュラ関数 l，u に対して，$\boldsymbol{l}(S) \leq F(S) \leq \boldsymbol{u}(S)\ (\forall S \subseteq V)$

- $\sum_{S \subseteq V} \exp(+\boldsymbol{l}(S)) \leq \sum_{S \subseteq V} \exp(+F(S)) \leq \sum_{S \subseteq V} \exp(+\boldsymbol{u}(S))$
- $\sum_{S \subseteq V} \exp(-\boldsymbol{l}(S)) \geq \sum_{S \subseteq V} \exp(-F(S)) \geq \sum_{S \subseteq V} \exp(-\boldsymbol{u}(S))$

上と下から，モジュラ関数の分配関数でおさえられる

l，uとしては優勾配・劣勾配が利用可能

変分法によるアプローチ（再掲）

- *モジュラ関数は，分配関数が厳密に計算可能*

分配関数に関する有用な性質：

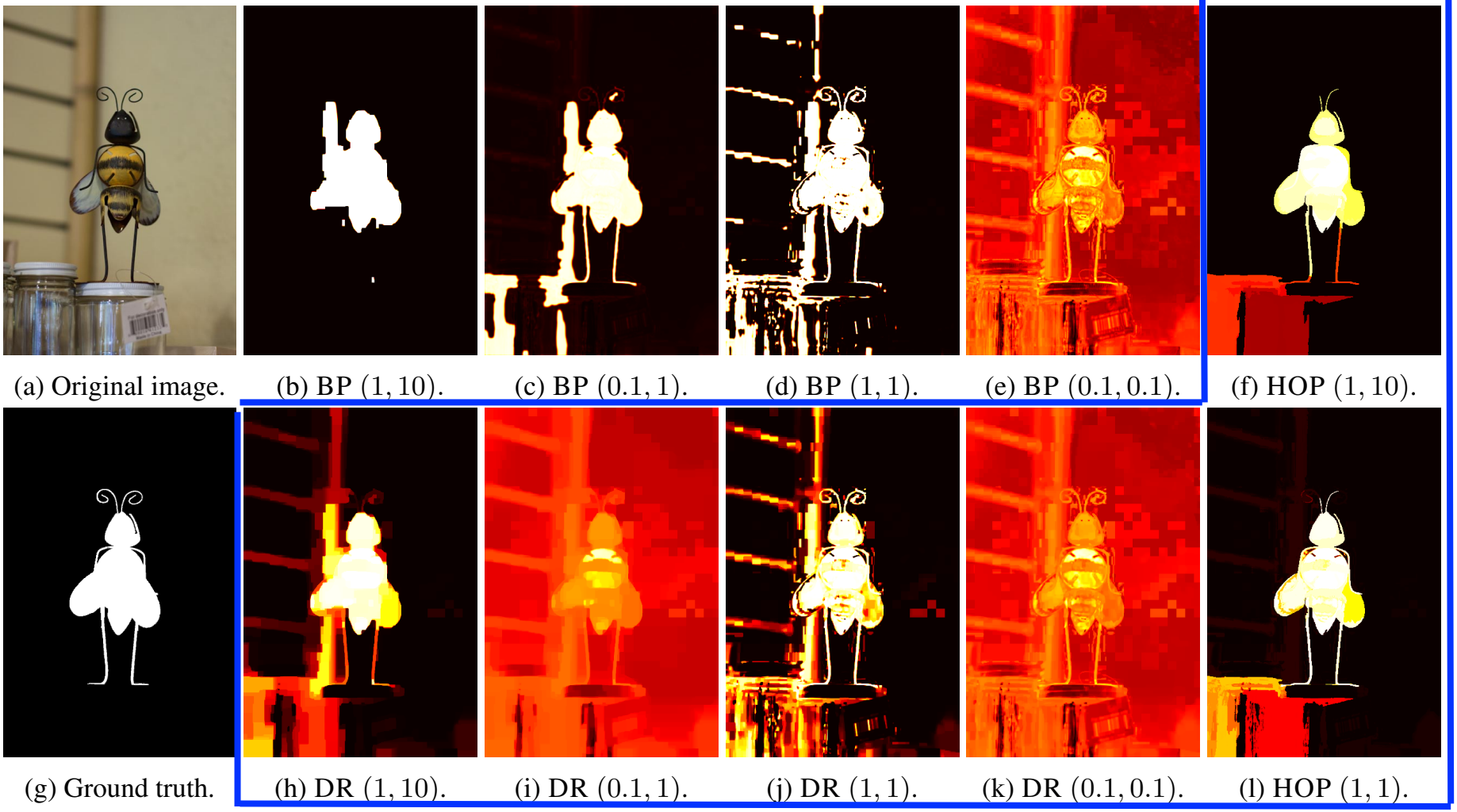
モジュラ関数 l，u に対して，$\boldsymbol{l}(S) \leq F(S) \leq \boldsymbol{u}(S)\ (\forall S \subseteq V)$

- $\sum_{S\subseteq V} \exp(+\boldsymbol{l}(S)) \leq \sum_{S\subseteq V} \exp(+F(S)) \leq \sum_{S\subseteq V} \exp(+\boldsymbol{u}(S))$
- $\sum_{S\subseteq V} \exp(-\boldsymbol{l}(S)) \geq \sum_{S\subseteq V} \exp(-F(S)) \geq \sum_{S\subseteq V} \exp(-\boldsymbol{u}(S))$

上と下から，モジュラ関数の分配関数でおさえられる

⇒ 劣モジュラ関数の最小化を何回か計算すれば最適化できる

L-fieldの適用例 (Kjolonga & Krause 15)



(a) Original image. (b) BP $(1, 10)$. (c) BP $(0.1, 1)$. (d) BP $(1, 1)$. (e) BP $(0.1, 0.1)$. (f) HOP $(1, 10)$.

(g) Ground truth. (h) DR $(1, 10)$. (i) DR $(0.1, 1)$. (j) DR $(1, 1)$. (k) DR $(0.1, 0.1)$. (l) HOP $(1, 1)$.

関連する論文

- 対数優モジュラ分布のMAP推定とその変分法による最適化 (Djolonga & Krause 14).
- 対数優モジュラ分布，対数劣モジュラ分布上での近似推論のためのMCMC（ギブス・サンプリングなど） (Gotovos+ 15)
- 上記のダイバージェンス的な解釈，期待値伝搬法や分散化などによる大規模化 (Djolonga & Krause 15).

- 行列式点過程の対数優モジュラ分布としての定式化と推論 (Iyer & Bilmes 14).
- 上記の劣モジュラ点過程としての一般化や，分配関数の近似率 (Iyer & Bilmes 15).

まとめ

- 劣モジュラ関数と，構造的な学習モデルとの関係に着目し，
  - 劣モジュラ関数の緩和を用いた構造正則化とその最適化
  - 劣モジュラ関数から得られる確率分布を用いた推論と，従来からある概念との関係

  を中心に紹介した．
- 劣モジュラ関数を用いた構造正則化については下記も参考になる．
  - F. Bach, “Learning with submodular function: A convex optimization perspective,” Foundations and Trends in ML, **6**(2-3): 145-373, 2013.
  - 河原吉伸，永野清仁, “劣モジュラ最適化と機械学習,” 講談社, 2015.

謝辞：紹介した自身の研究の一部は次の方々との成果を含みます．

B. Xin氏，Y. Wang氏（北京大），山口勇太郎氏（東大），永野清仁氏（未来大），竹内孝氏，岩田具治氏（CS研），永田啓介氏（アステラス），鷲尾隆氏（阪大）．