正則化学習法における最適化手法

鈴木 大慈
東京大学
情報理工学系研究科
数理情報学専攻

2013/2/18@九州大学伊都キャンパス
文部科学省委託事業数学協働プログラム
「最適化ワークショップ: 拡がっていく最適化」

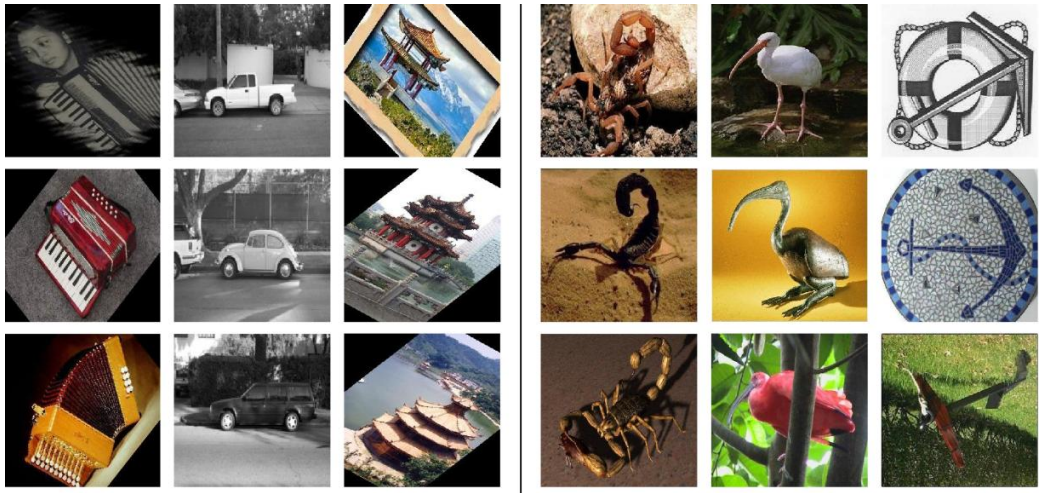
高次元データ
スパース正則化学習法

最適化手法
proximal point algorithm
確率最適化手法

問題設定
スパース正則化学習

高次元線形判別

物体認識

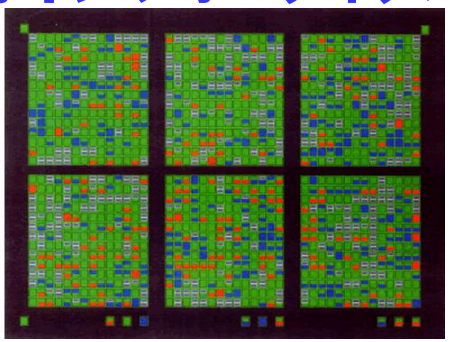


音声認識

自然言語処理

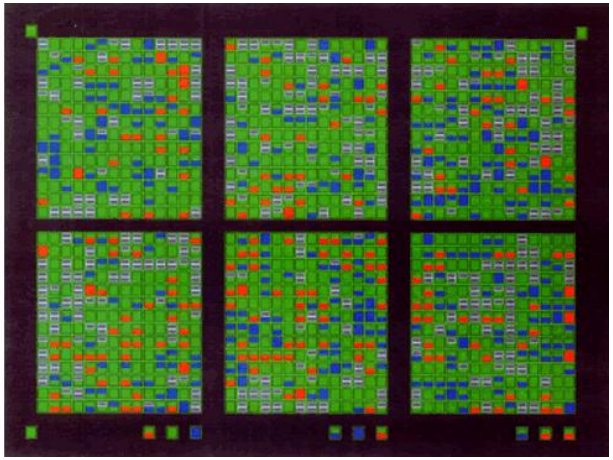
Google 日本

バイオインフォマティクス



特徴ベクトル

DNAデータ



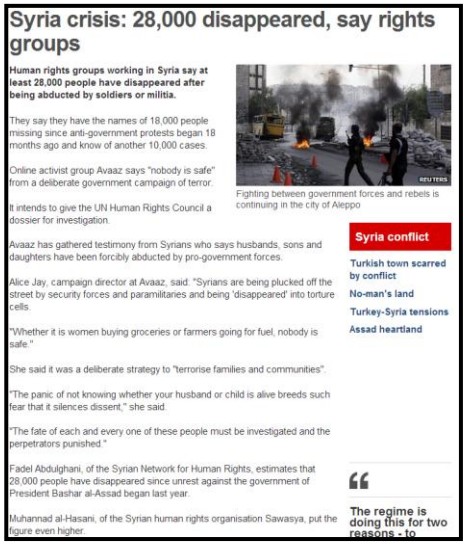
ベクトル化

判別

癌になりやすい
癌になりにくい

高次元

テキストデータ



Syria crisis: 28,000 disappeared, say rights groups

Human rights groups working in Syria say at least 28,000 people have disappeared after being abducted by soldiers or militia.

They say they have the names of 18,000 people missing since anti-government protests began 18 months ago and know of another 10,000 cases.

Online activist group Avaaz says "nobody is safe" from a deliberate government campaign of terror.

It intends to give the UN Human Rights Council a dossier for investigation.

Fighting between government forces and rebels is continuing in the city of Aleppo

Avaaz has gathered testimony from Syrians who says husbands, sons and daughters have been forcibly abducted by pro-government forces.

Alice Jay, campaign director at Avaaz, said: "Syrians are being plucked off the street by security forces and paramilitaries and being 'disappeared' into torture cells.

"Whether it is women buying groceries or farmers going for fuel, nobody is safe."

She said it was a deliberate strategy to "terrorise families and communities".

"The panic of not knowing whether your husband or child is alive breeds such fear that it silences dissent," she said.

"The fate of each and every one of these people must be investigated and the perpetrators punished."

Fadel Abdulghani, of the Syrian Network for Human Rights, estimates that 28,000 people have disappeared since unrest against the government of President Bashar al-Assad began last year.

Muhannad al-Hasani, of the Syrian human rights organisation Sawasya, put the figure even higher.

Syria conflict

Turkish town scarred by conflict

No-man's land

Turkey-Syria tensions

Assad heartland

“

The regime is doing this for two reasons - to

特徴ベクトル

ベクトル化

判別

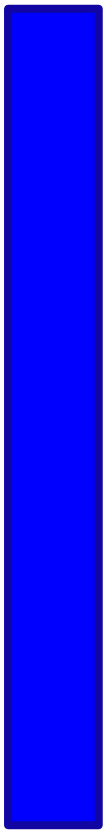
何の話題か？

高次元

特徴ベクトル

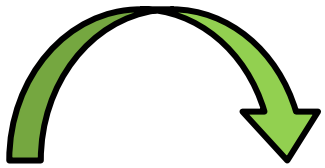
画像データ

ベクトル化

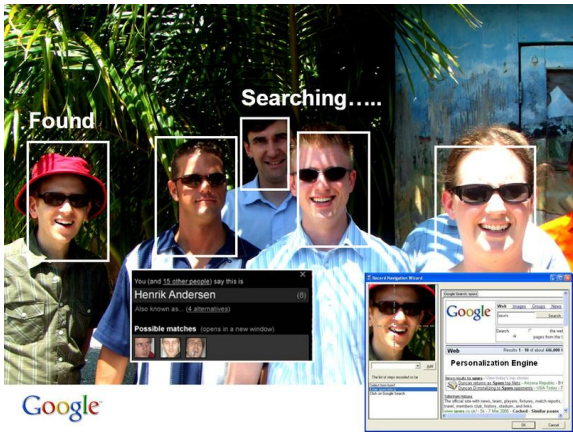


高次元

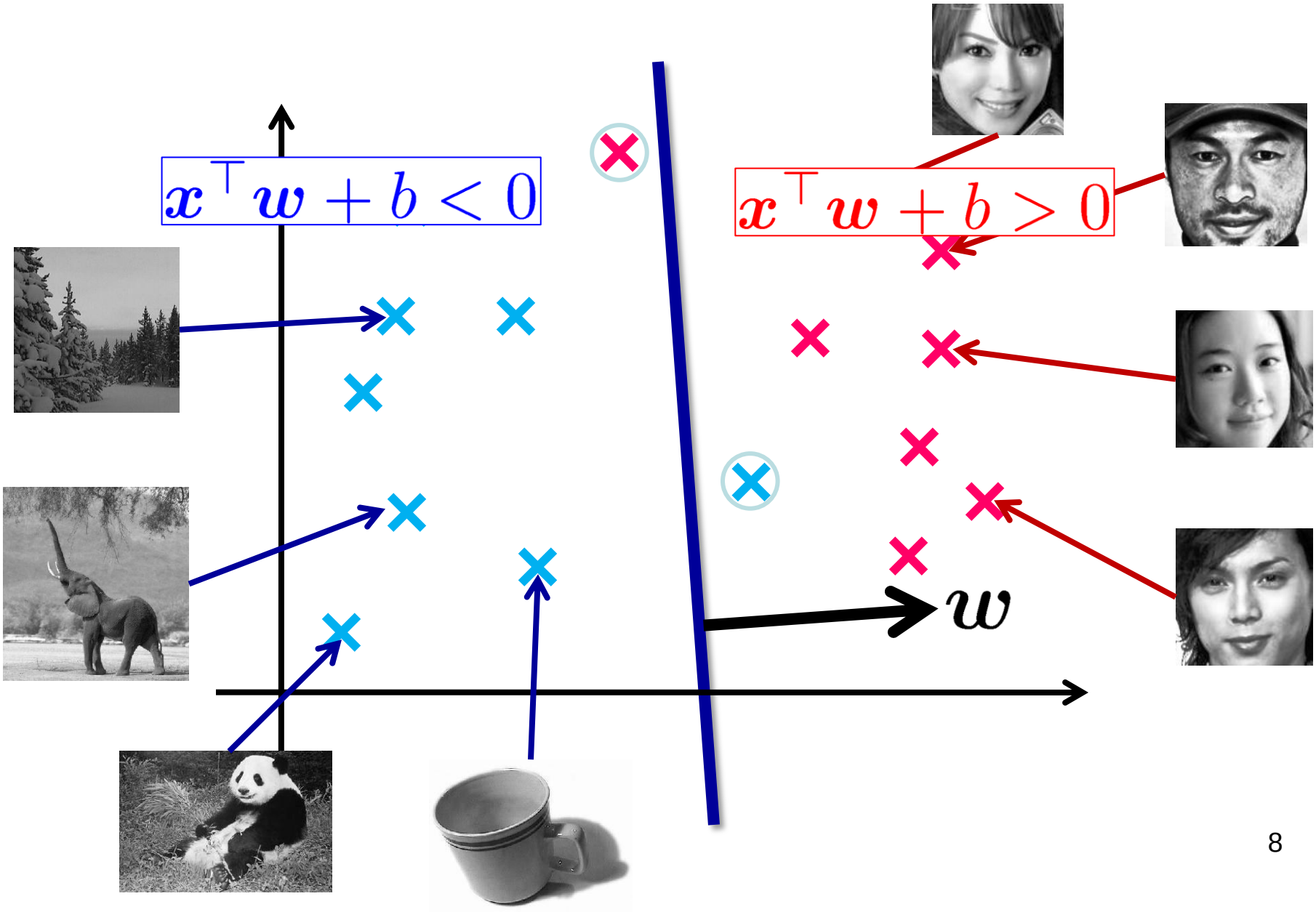
判別



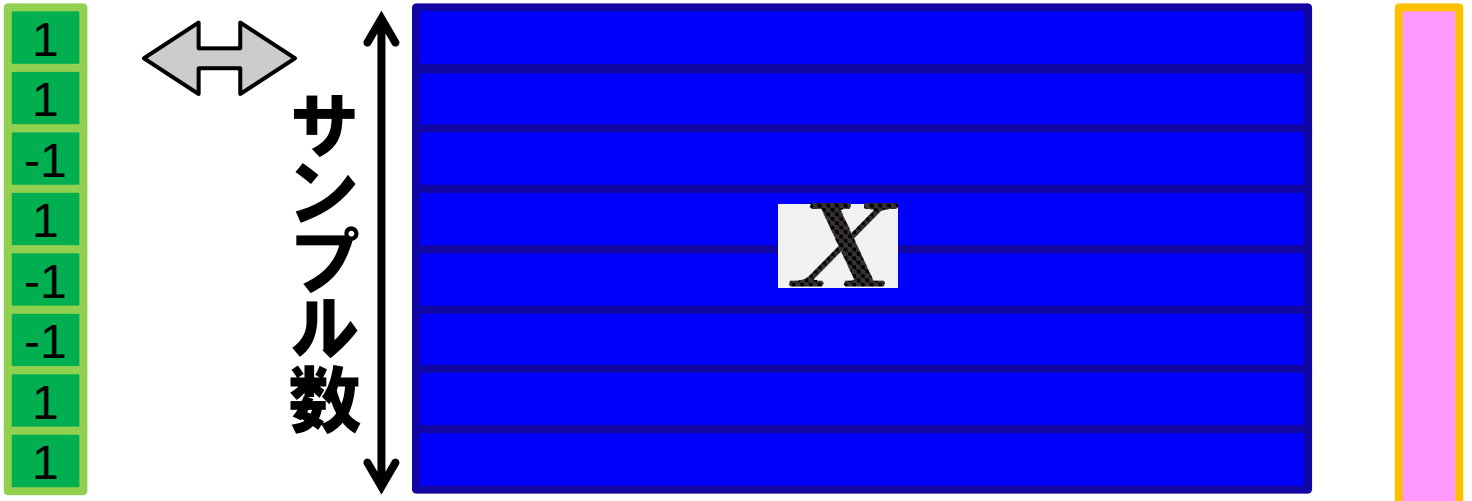
顔か？
顔でないか？



線形判別

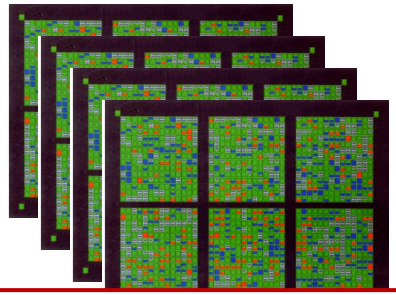


y $x^\top$ w

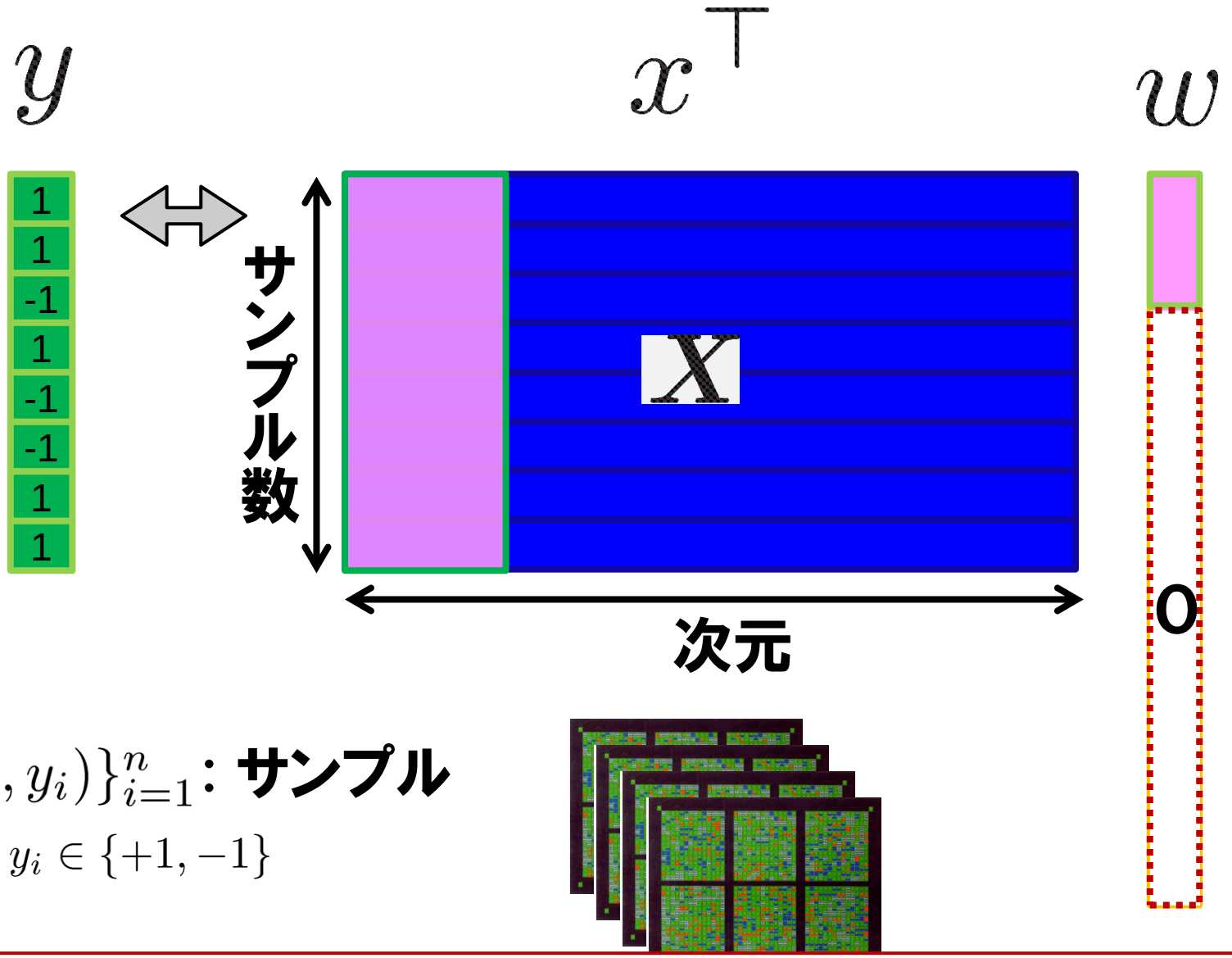


$\{(x_i, y_i)\}_{i=1}^n$: サンプル

$y_i \in \{+1, -1\}$



次元 > サンプル数 → 余分な情報を落としたい



$\{(x_i, y_i)\}_{i=1}^n$: **サンプル**

$y_i \in \{+1, -1\}$

次元 > サンプル数 → 余分な情報を落としたい

スパース推定：L1正則化

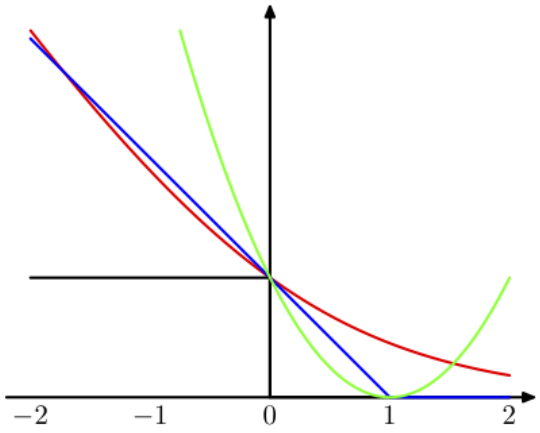
$$\min_{\boldsymbol{w}} \quad \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \boldsymbol{w}^\top \boldsymbol{x}_i) \quad + \quad C\|\boldsymbol{w}\|_1$$

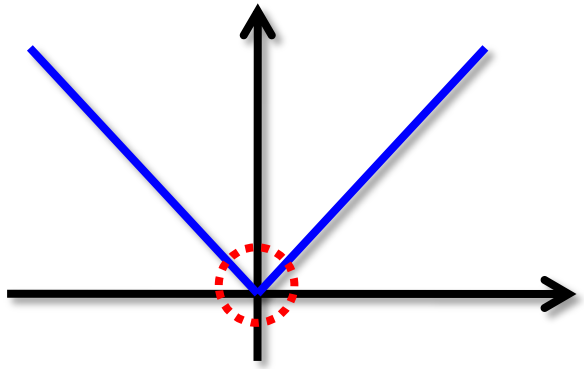
間違いへのペナルティ

平面の「複雑さ」
L1ノルム→スパース

ℓ ：ロス関数=間違いの度合いが大きいほど大きな値

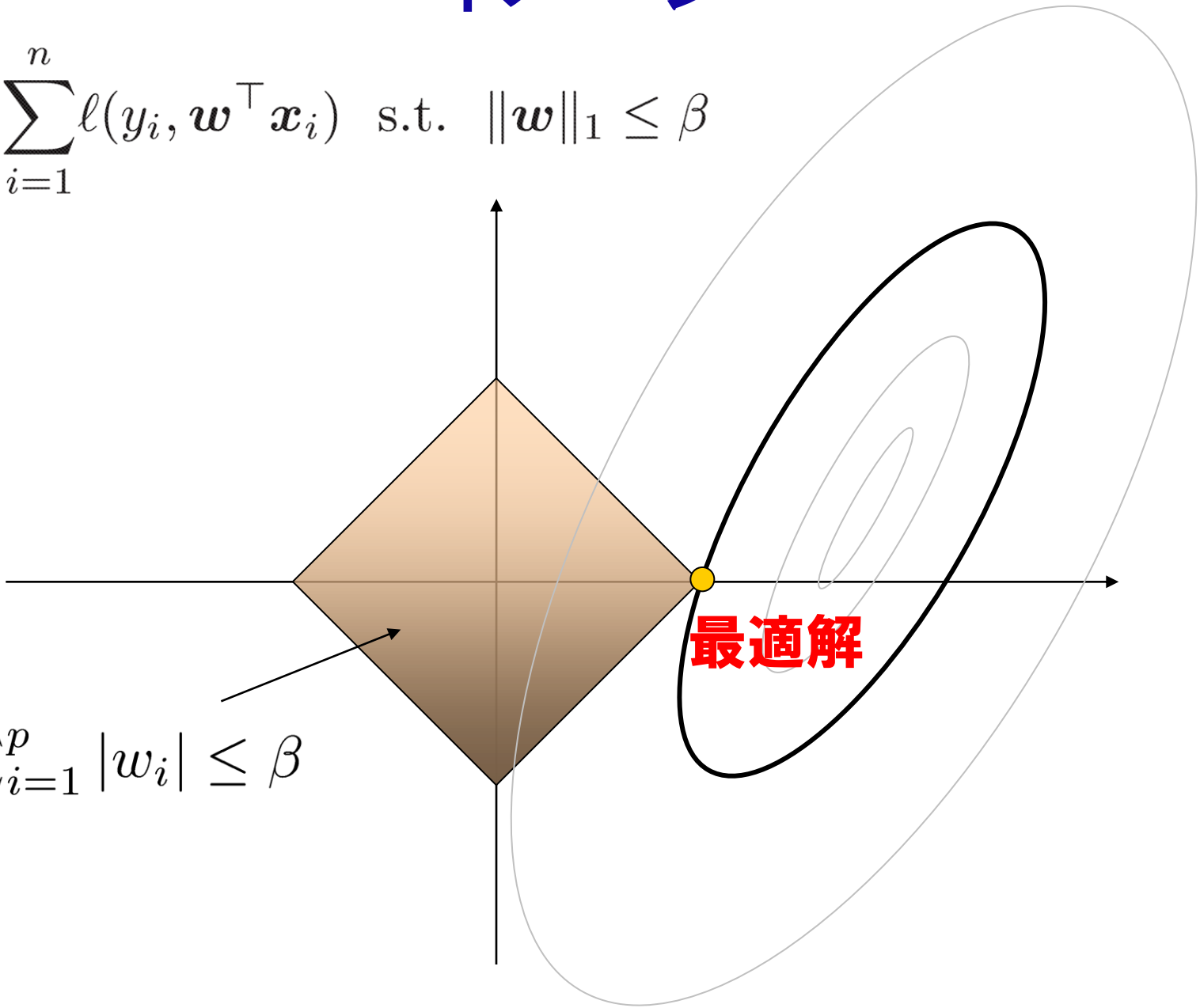
$$\|\boldsymbol{w}\|_1 := \sum_{j=1}^{p}|\boldsymbol{w}_j|$$





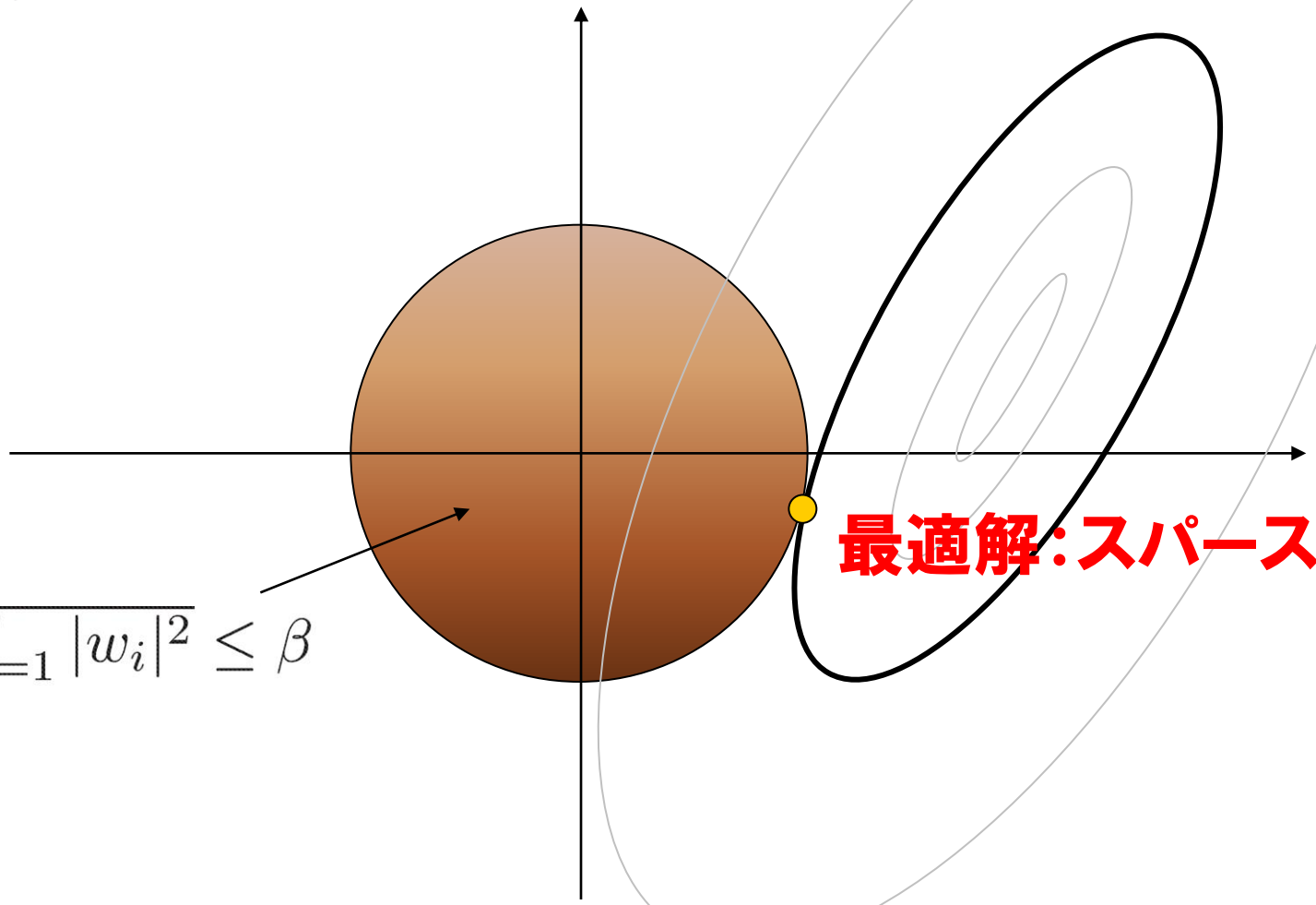
イメージ

$$\min_{\boldsymbol{w}} \quad \sum_{i=1}^{n} \ell(y_i, \boldsymbol{w}^\top \boldsymbol{x}_i) \quad \text{s.t.} \quad \|\boldsymbol{w}\|_1 \leq \beta$$



L2正則化の場合

$$\min_{\boldsymbol{w}} \quad \sum_{i=1}^{n} \ell(y_i, \boldsymbol{w}^\top \boldsymbol{x}_i) \quad \text{s.t.} \quad \|\boldsymbol{w}\|_2 \leq \beta$$



$$\sqrt{\sum_{i=1}^{p} |w_i|^2} \leq \beta$$

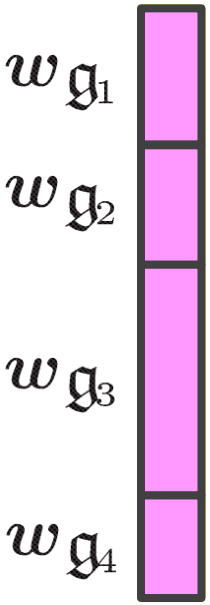
最適解：スパースでない

スパース正則化の例

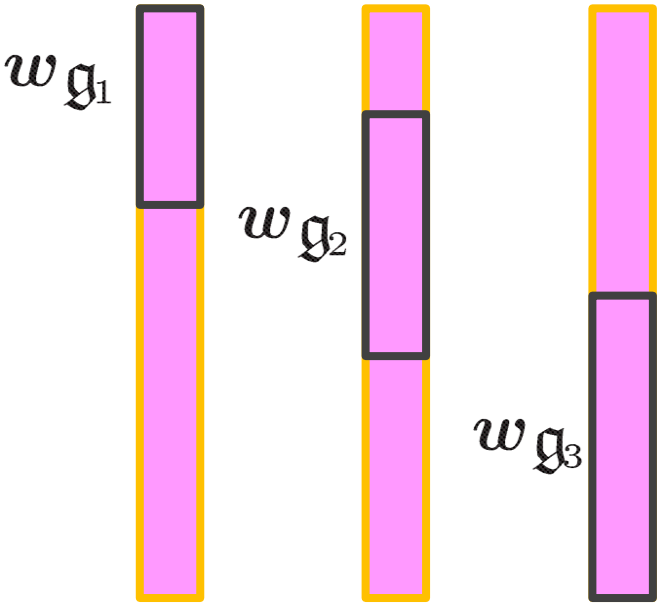
例1：Group Lasso

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, w^\top x_i) \;+\; C \sum_{\mathfrak{g}\in\mathfrak{G}} \|w_{\mathfrak{g}}\|_2$$

グループ構造



重複あり



例2：低ランク行列推定

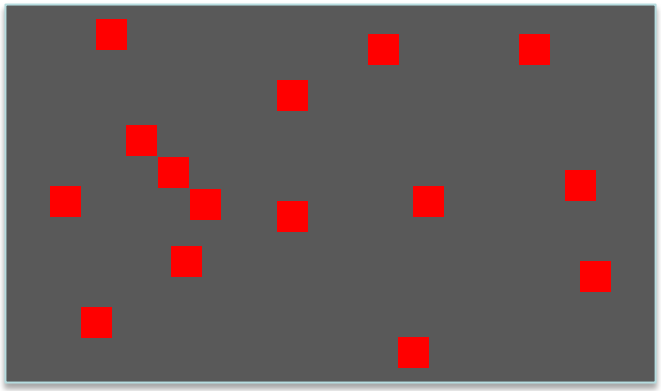
$$\min_{\boldsymbol{w}} \quad \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \mathrm{Tr}(\boldsymbol{X}_i\boldsymbol{W})) \;+\; C\,\|\boldsymbol{W}\|_{\mathrm{tr}}$$

$$\|\boldsymbol{W}\|_{\mathrm{tr}} = \mathrm{Tr}(\sqrt{W^\top W})$$

低ランク

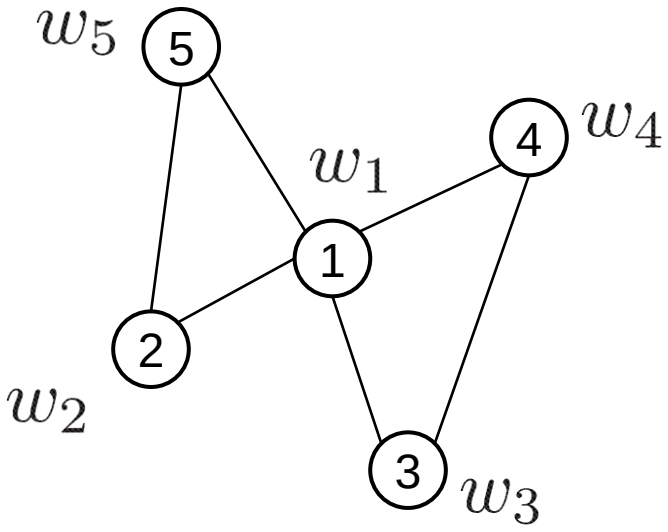
$\boldsymbol{W} =$

ユーザの趣向推定



例3：グラフ型正則化

$$\min_{\boldsymbol{w}} \quad \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \boldsymbol{w}^\top \boldsymbol{x}_i) \;+\; C \sum_{(i,j)\in\mathcal{E}} |w_i - w_j|$$



スパース正則化学習の最適化

$$\min_{\boldsymbol{w}} \quad L(X\boldsymbol{w}) + \psi(\boldsymbol{w})$$

proximal point algorithm [Rockafellar, 1976]

$$\boldsymbol{w}_{t+1} \leftarrow \min_{\boldsymbol{w}} \quad L(X\boldsymbol{w}) + \psi(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$$

双対

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{\rho},\boldsymbol{y}} \ \langle X\boldsymbol{w}, \boldsymbol{\rho}\rangle - L^*(\boldsymbol{\rho}) + \langle \boldsymbol{w}, \boldsymbol{y}\rangle - \psi^*(\boldsymbol{y}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$$

乗数法 [Hestenes, 1969; Powel 1969]

$$\min_{\boldsymbol{\rho},\boldsymbol{y}} \quad L^*(\boldsymbol{\rho}) + \psi^*(\boldsymbol{y}) - \boldsymbol{w}_t^\top(\boldsymbol{X}^\top\boldsymbol{\rho} - \boldsymbol{y}) + \frac{\eta_t}{2}\|\boldsymbol{X}^\top\boldsymbol{\rho} - \boldsymbol{y}\|^2$$

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta_t(\boldsymbol{X}^\top\boldsymbol{\rho}^* - \boldsymbol{y}^*)$$

proximal point alg.: $$w_{t+1} \leftarrow \min_{w} \quad L(Xw) + \psi(w) + \frac{1}{2\eta_t}\|w - w_t\|^2$$

乗数法: $$\min_{\rho,y} \quad L^*(\rho) + \psi^*(y) - w_t^\top(X^\top\rho - y) + \frac{\eta_t}{2}\|X^\top\rho - y\|^2$$
$$w_{t+1} \leftarrow w_t - \eta_t(X^\top\rho^* - y^*)$$

主問題	prox. point alg. FOBOS [Duchi&Singer,2009] FISTA [Beck&Teboulle,2009] (近似解法・近接勾配法)	乗数法 ADMM (Alternating Direction Multiplier Method) [Glowinski&Marrocco,75;Boyd et.al.,10] (近似解法)
双対問題	乗数法 DAL (Dual Augmented Lagrangian) [Tomioka&Sugiyama,2009; Tomioka,Suzuki&Sugiyama,2011]	prox. point alg. **SpicyMKL**

FOBOS (Forward Backward Splitting)

Nesterov (2007), Duchi&Singer (2009), FISTA:Beck&Teboulle (2009)

prox. point alg. $$\boldsymbol{w}_{t+1} \leftarrow \min_{\boldsymbol{w}} \; L(X\boldsymbol{w}) + \psi(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$$

線形近似

$$g_t \in \nabla_{\boldsymbol{w}} L(X\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{w}_t}$$

$$\boldsymbol{w}_{t+1} \leftarrow \min_{\boldsymbol{w}} \; \langle g_t, \boldsymbol{w}\rangle + \psi(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$$

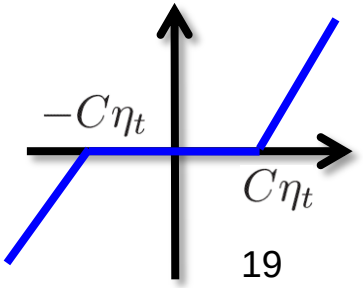
η_t : ステップ幅パラメータ

L1正則化での更新式 ($\psi(\boldsymbol{w}) = C\|\boldsymbol{w}\|_1$)

$$\boldsymbol{w}_{t+1}^{(i)} = \mathrm{sign}(\boldsymbol{w}_t^{(i)} - \eta_t g_t^{(i)}) \max[|\boldsymbol{w}_t^{(i)} - \eta_t g_t^{(i)}| - C\eta_t, 0]$$

最適化の途中でもスパースな解

Soft threshold



Proximal Operation

FOBOSは以下のproximal operationで更新

$$\mathrm{prox}(\boldsymbol{q}|\psi) := \arg\min_{\boldsymbol{w}} \left\{ \psi(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{q}\|^2 \right\}$$

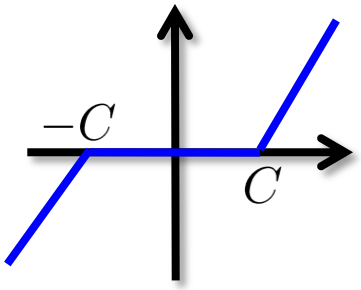
(射影の一般化)

FOBOS: $\boldsymbol{w}_{t+1} = \mathrm{prox}(\boldsymbol{w}_t - \eta_t g_t | \eta_t \psi)$

例：L1ノルムでのproximal operation ($\psi(\boldsymbol{w}) = C\|\boldsymbol{w}\|_1$)

$$\mathrm{prox}(\boldsymbol{q}|C\|\cdot\|_1) = (\mathrm{sign}(q_j)\max(|q_j|-C, 0))_j$$

Soft threshold



- **各変数ごとの最適化に分離．変数間の絡みがない．**
- **解析解の存在．**

収束レート

- 一般の凸ロス関数

$$L(X\boldsymbol{w}_T) + \psi(\boldsymbol{w}_T) - [L(X\boldsymbol{w}^*) + \psi(\boldsymbol{w}^*)] \leq C\frac{1}{\sqrt{T}}$$

- 滑らかな凸ロス関数

$$L(X\boldsymbol{w}_T) + \psi(\boldsymbol{w}_T) - [L(X\boldsymbol{w}^*) + \psi(\boldsymbol{w}^*)] \leq C\frac{1}{T}$$

- $X^\top X$ が正則で強凸かつ滑らかなロス関数

$$L(X\boldsymbol{w}_T) + \psi(\boldsymbol{w}_T) - [L(X\boldsymbol{w}^*) + \psi(\boldsymbol{w}^*)] \leq C\exp(-Tc)$$

proximal point alg.: $$w_{t+1} \leftarrow \min_{w} \quad L(Xw) + \psi(w) + \frac{1}{2\eta_t}\|w - w_t\|^2$$

乗数法: $$\min_{\rho,y} \quad L^*(\rho) + \psi^*(y) - w_t^\top(X^\top\rho - y) + \frac{\eta_t}{2}\|X^\top\rho - y\|^2$$
$$w_{t+1} \leftarrow w_t - \eta_t(X^\top\rho^* - y^*)$$

	prox. point alg.	**乗数法**
主問題	FOBOS [Duchi&Singer,2009] FISTA [Beck&Teboulle,2009] (近似解法・近接勾配法)	ADMM (Alternating Direction Multiplier Method) [Glowinski&Marrocco,75;Boyd et.al.,10] (近似解法)
双対問題	**乗数法** DAL (Dual Augmented Lagrangian) [Tomioka&Sugiyama,2009; Tomioka,Suzuki&Sugiyama,2011]	**prox. point alg.** ×

Dual Augmented Lagrangian (DAL)

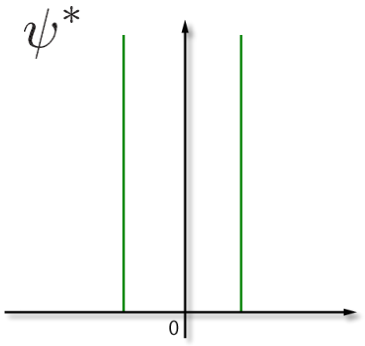
[Tomioka&Sugiyama,2009; Tomioka,Suzuki&Sugiyama,2011]

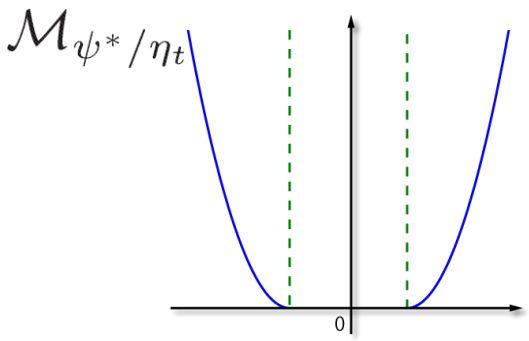
$$(\boldsymbol{\rho}_{t+1}, \boldsymbol{y}_{t+1}) \leftarrow \operatorname*{arg\,min}_{\boldsymbol{\rho},\boldsymbol{y}} \quad L^*(\boldsymbol{\rho}) + \psi^*(\boldsymbol{y}) - \boldsymbol{w}_t^\top (\boldsymbol{X}^\top \boldsymbol{\rho} - \boldsymbol{y}) + \frac{\eta_t}{2} \|\boldsymbol{X}^\top \boldsymbol{\rho} - \boldsymbol{y}\|^2$$

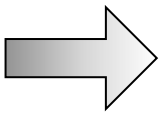
$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta_t (\boldsymbol{X}^\top \boldsymbol{\rho}_{t+1} - \boldsymbol{y}_{t+1})$$

$$\min_{\boldsymbol{\rho}} \quad L^*(\boldsymbol{\rho}) + \eta_t \underbrace{\min_{\boldsymbol{y}} \left\{ \frac{\psi^*(\boldsymbol{y})}{\eta_t} + \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\rho} + \frac{\boldsymbol{w}_t}{\eta_t} \right\|^2 \right\}}$$

$$\mathcal{M}_{\psi^*/\eta_t}(\boldsymbol{X}^\top \boldsymbol{\rho} - \boldsymbol{w}_t/\eta_t)$$: Moreau's envelope



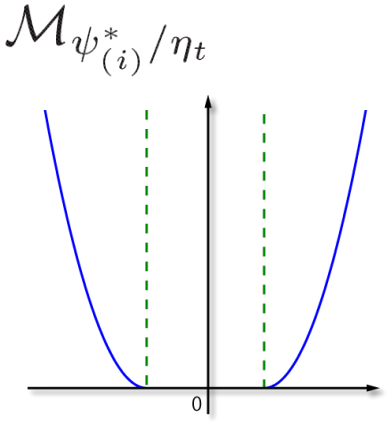




*L**が滑らかな場合，
ρに関するNewton法

DALの性質

$$\min_{\boldsymbol{\rho}} \quad L^*(\boldsymbol{\rho}) + \eta_t \underbrace{\mathcal{M}_{\frac{\psi^*}{\eta_t}}\left(\boldsymbol{X}^\top \boldsymbol{\rho} - \frac{\boldsymbol{w}_t}{\eta_t}\right)}_{\sum_{i=1}^{p} \mathcal{M}_{\frac{\psi^*_{(i)}}{\eta_t}}\left(\boldsymbol{X}_i^\top \boldsymbol{\rho} - \frac{w_t^{(i)}}{\eta_t}\right)}$$



- $\mathcal{M}_{\psi^*_{(i)}/\eta_t}$の微分をスキップできる
  → スパース性を利用した高速化
- （超）一次収束

$f(\boldsymbol{w}) - f(\boldsymbol{W}^*) \geq \sigma \|\boldsymbol{w} - \boldsymbol{W}^*\|^2$ なら

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{W}^*\| \leq \frac{1}{1+\sigma\eta_t} \|\boldsymbol{w}_t - \boldsymbol{W}^*\|$$

最適解が一意である必要はない

SpicyMKL [Suzuki&Tomioka, 2011]

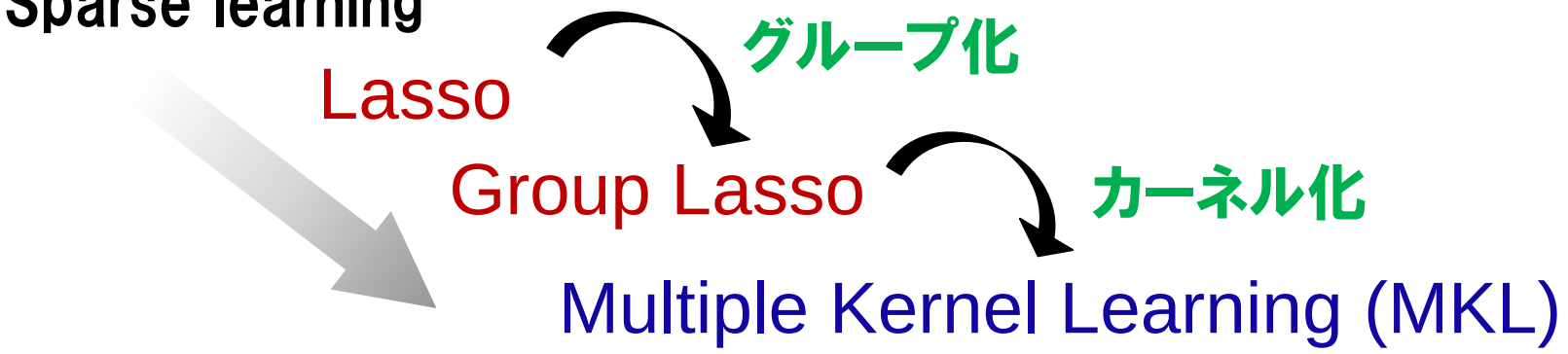
DALのMKLへの拡張

Multiple Kernel Learning (MKL) [Lanckriet et al. 2004]

グループ正則化の各グループを無限次元の<u>再生核ヒルベルト空間</u>とした方法.

沢山のカーネル関数とそれに付随した再生核ヒルベルト空間 $\{\mathcal{H}_m\}_{m=1}^M$

$$\min_{f_m \in \mathcal{H}_m} \quad \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, \sum_{m=1}^{M} f_m(\boldsymbol{x}_i)\right) + C \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}$$



ソフトウェアの公開

SpicyMKL

SpicyMKL is an optimization method of MKL (mutiple kernel learning) that scales well against the number of kernels. The software affords hinge, logistic and square losses, and L1-norm and elasticnet regularization.

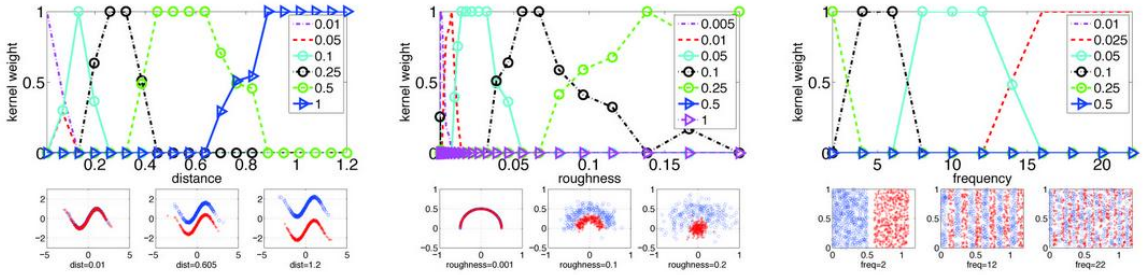
Matlab Implementation: SpicyMKL.zip (updated 1st Nov. 2010)

- SpicyMKL.m is the main function.
  - SpicyMKL.m calls normKj.mex and HessAugMexbias.mex. Please link blas library when compiling these files like

```
mex -lmwblas normKj.c
mex -lmwblas HessAugMexbias.c
```

  (startup.m contains this code).
- demo.m is a demo script.
  - The demo code requires SimpleMKL Toolbox in the path.

Examples



Matlabコード

http://www.simplex.t.u-tokyo.ac.jp/~s-taiji/software/SpicyMKL/

DAL

What is DAL?

DAL is an efficient and flexibible MATLAB toolbox for solving the sparsity-regularized minimization problems, which arrises often in machine learning, of the following form:

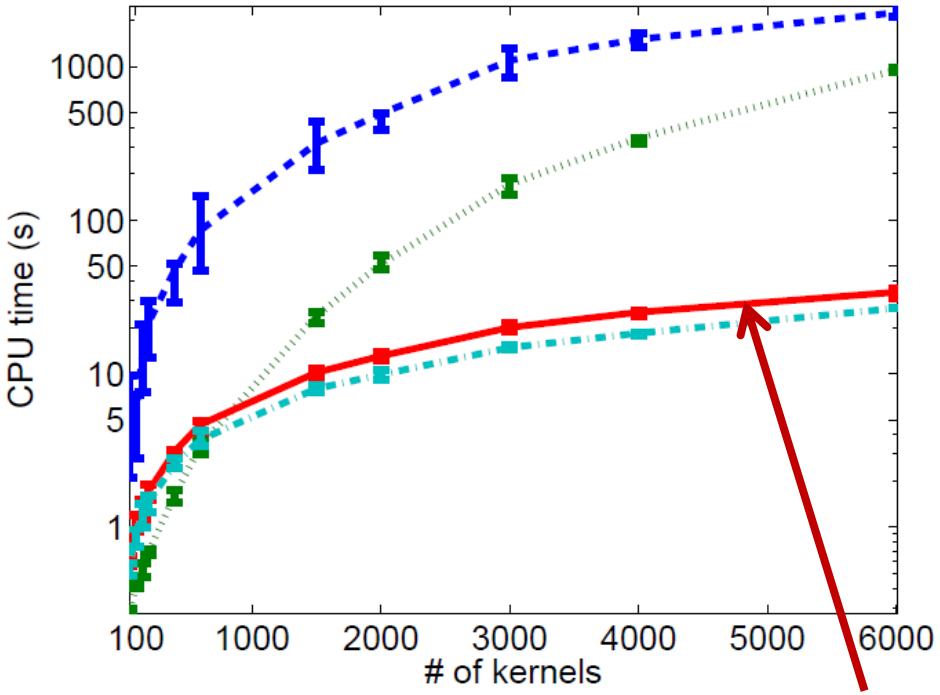
$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad f(\boldsymbol{A}\boldsymbol{w}) + \lambda g(\boldsymbol{w})$$

http://www.ibis.t.u-tokyo.ac.jp/ryotat/dal/

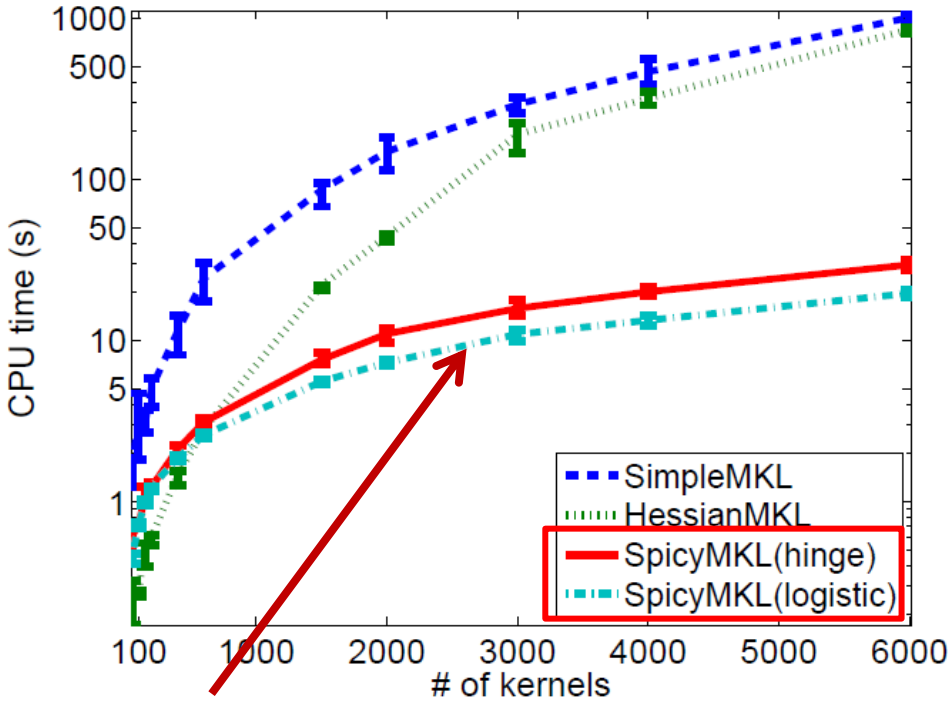
- DAL is efficient when m≪n (m: #samples, n: #unknowns) or the matrix A is poorly conditioned.

数値実験

CPU time v.s. # of kernels



Splice



Ringnorm

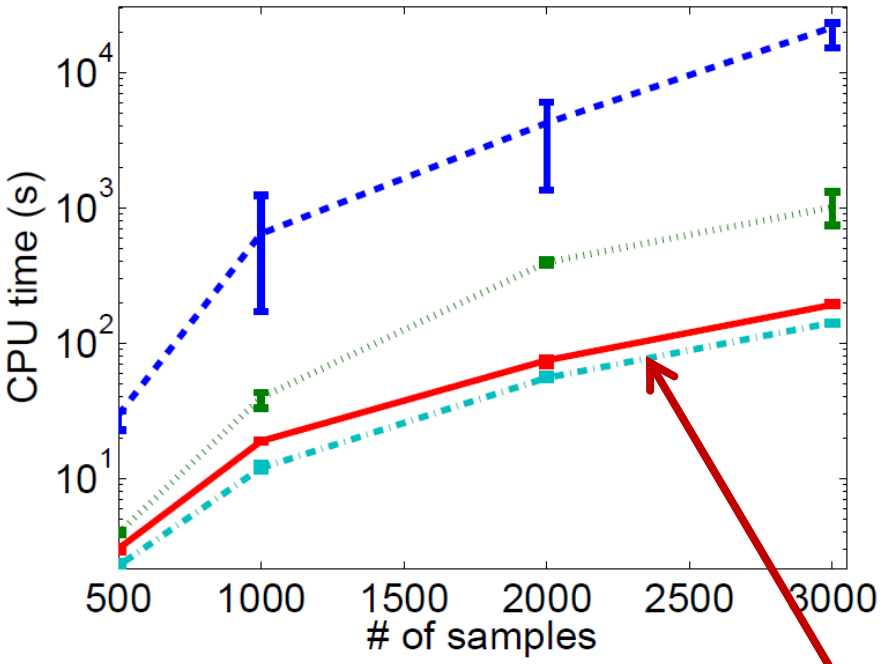
SpicyMKL

- カーネルの数に対し良くスケールする

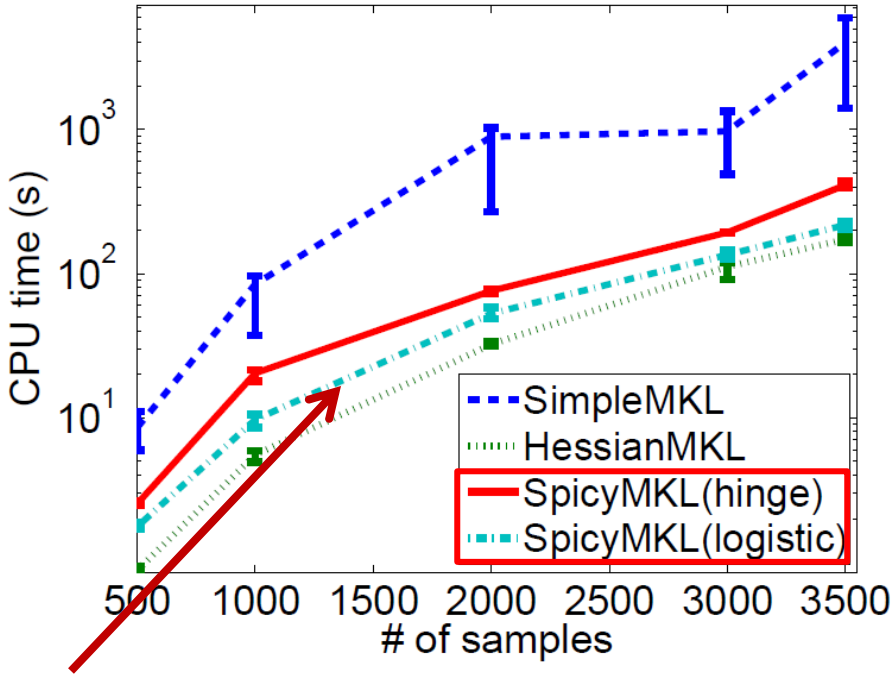
IDA data set, L1 regularization.

IDA data set, L1 regularization.

CPU time v.s. # of samples



Splice



Ringnorm

SpicyMKL

- サンプル数に対しては既存手法とほぼ同じスケール

proximal point alg.: $$w_{t+1} \leftarrow \min_{w} \quad L(Xw) + \psi(w) + \frac{1}{2\eta_t}\|w - w_t\|^2$$

乗数法: $$\min_{\rho,y} \quad L^*(\rho) + \psi^*(y) - w_t^\top(X^\top\rho - y) + \frac{\eta_t}{2}\|X^\top\rho - y\|^2$$
$$w_{t+1} \leftarrow w_t - \eta_t(X^\top\rho^* - y^*)$$

	主問題	**双対問題**
prox. point alg.	FOBOS [Duchi&Singer,2009] FISTA [Beck&Teboulle,2009] (近似解法・近接勾配法)	**乗数法** DAL (Dual Augmented Lagrangian) [Tomioka&Sugiyama,2009; Tomioka,Suzuki&Sugiyama,2011]
乗数法	ADMM (Alternating Direction Multiplier Method) [Glowinski&Marrocco,75;Boyd et.al.,10] (近似解法)	**prox. point alg.** ×

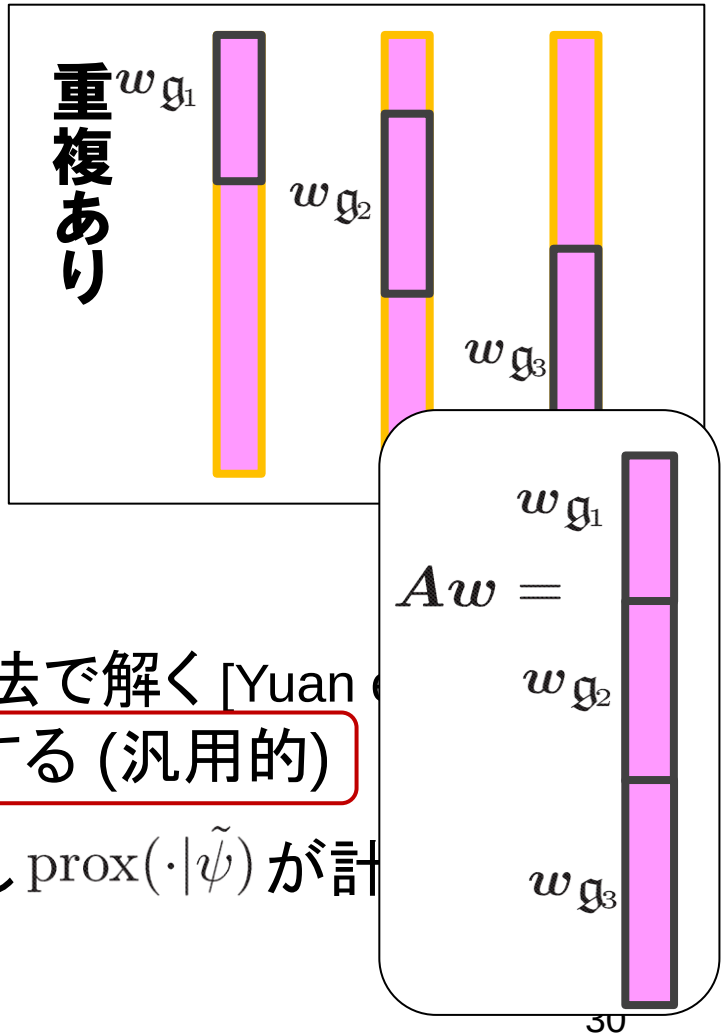
prox. op.が計算しにくい例

$$\mathrm{prox}(\boldsymbol{q}|\psi) := \arg\min_{\boldsymbol{w}} \left\{ \psi(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{q}\|^2 \right\}$$

- 重複ありグループ正則化

$$\psi(\boldsymbol{w}) = C \sum_{\mathfrak{g}\in\mathfrak{G}} \|\boldsymbol{w}_{\mathfrak{g}}\|_2$$

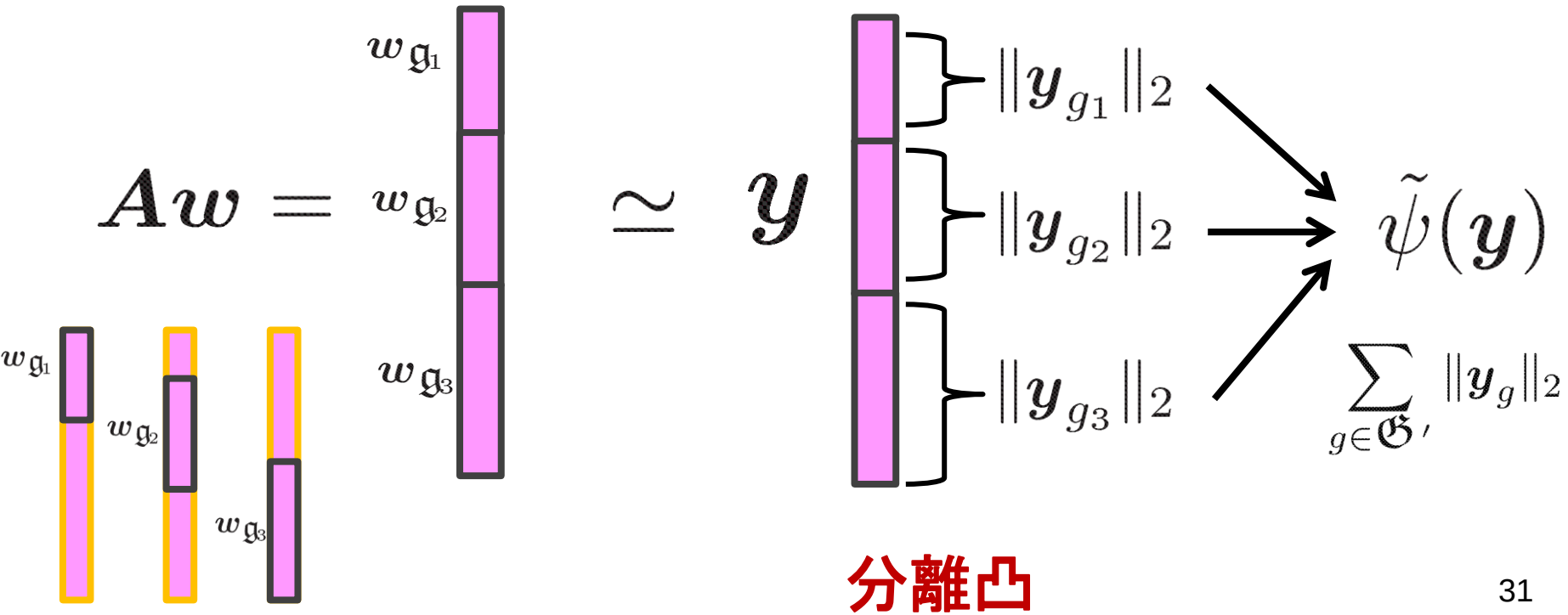
変数間に絡み



- 解決策
  - •各正則化関数に応じた賢い方法で解く [Yuan
  - •変数を増やして問題を簡単にする (汎用的)

idea: $\tilde{\psi}(\boldsymbol{A}\boldsymbol{w}) = \psi(\boldsymbol{w})$ を満たし $\mathrm{prox}(\cdot|\tilde{\psi})$ が計
$\tilde{\psi}$ を利用する.

$$\min_{\boldsymbol{w},\boldsymbol{y}} \quad L(\boldsymbol{X}\boldsymbol{w}) + \tilde{\psi}(\boldsymbol{y})$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} = \boldsymbol{y}$$



分離凸

ADMM
(Alternating Direction Multiplier Method)

- 乗数法：wとyを同時最適化

$$(\boldsymbol{w}_{t+1}, \boldsymbol{y}_{t+1}) \leftarrow \min_{\boldsymbol{w},\boldsymbol{y}} \; L(\boldsymbol{X}\boldsymbol{w}) + \tilde{\psi}(\boldsymbol{y}) - \boldsymbol{\lambda}_t^\top(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{y}) + \frac{\eta_t}{2}\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{y}\|^2$$

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \eta_t(\boldsymbol{A}\boldsymbol{w}_{t+1} - \boldsymbol{y}_{t+1})$$

- Splitting Technique

  wとyの最適化を分離

ADMM

$$\boldsymbol{w}_{t+1} \leftarrow \arg\min_{\boldsymbol{w}} \; L(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{\lambda}_t^\top(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{y}_t) + \frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{y}_t\|^2$$

$$\boldsymbol{y}_{t+1} \leftarrow \arg\min_{\boldsymbol{y}} \; \tilde{\psi}(\boldsymbol{y}) + \boldsymbol{\lambda}_t^\top \boldsymbol{y} + \frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{w}_{t+1} - \boldsymbol{y}\|^2 \qquad (\mathrm{prox}(\boldsymbol{A}\boldsymbol{w}_{t+1} - \boldsymbol{\lambda}_t/\rho|\tilde{\psi}/\rho))$$

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \rho(\boldsymbol{A}\boldsymbol{w}_{t+1} - \boldsymbol{y}_{t+1})$$

ADMMの収束レート

- リプシッツ連続凸ロス関数

$$L(X\boldsymbol{w}_T) + \psi(\boldsymbol{w}_T) - [L(X\boldsymbol{w}^*) + \psi(\boldsymbol{w}^*)] \leq C\frac{1}{T}$$

- $X^\top X$ が正則で強凸かつ滑らかなロス関数

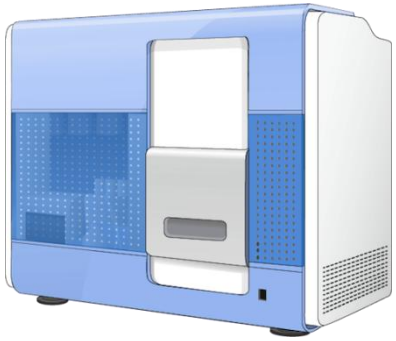
$$L(X\boldsymbol{w}_T) + \psi(\boldsymbol{w}_T) - [L(X\boldsymbol{w}^*) + \psi(\boldsymbol{w}^*)] \leq C\exp(-Tc)$$

確率的最適化
（オンライン学習）

近年のデータ

twitter
2億ツイート/日 (2011/6/30)
トルストイ「戦争と平和」8163冊分

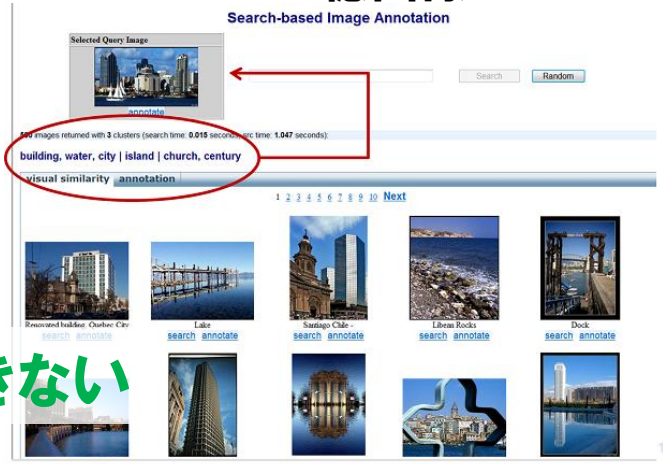
次世代シーケンサ
60億本 x100塩基



多量

Flickr
100万枚/日

ARISTA
20億画像



全データをメモリに保持できない

確率的最適化

prox. point alg. $$\boldsymbol{w}_{t+1} \leftarrow \min_{\boldsymbol{w}} \ \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{x}_i^\top \boldsymbol{w}) + \psi(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$$

全データを保持するのではなく,
一つサンプルを見たらwを更新してサンプルを捨てる方法

- FOBOS (Forward Backward Splitting) [Duchi&Singer,2009]

$$g_t \in \nabla \ell_t(\boldsymbol{w}_t)$$

$$\ell_t(\boldsymbol{w}) = \ell(y_t, \boldsymbol{x}_t^\top \boldsymbol{w})$$

$$\boldsymbol{w}_{t+1} \leftarrow \min_{\boldsymbol{w}} \ \langle g_t, \boldsymbol{w}\rangle + \psi(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$$

一つのサンプル・線形近似

- RDA (Regularized Dual Averaging) [Xiao,09; Nesterov,09]

$$\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$$

$$\boldsymbol{w}_{t+1} \leftarrow \min_{\boldsymbol{w}} \ \langle \bar{g}_t, \boldsymbol{w}\rangle + \psi(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w}\|^2$$

双対変数（勾配）の平均で過去の情報を保持

確率的ADMM

[Suzuki, ICML2013]

ADMM + 確率的最適化

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{y}} \quad & L(\boldsymbol{X}\boldsymbol{w}) + \tilde{\psi}(\boldsymbol{y}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{w} = \boldsymbol{y} \end{aligned}$$

- FOBOS型ADMM

$$\boldsymbol{w}_{t+1} \leftarrow \Pi_{\mathcal{W}}\left[-\frac{\eta_t}{\gamma}\left\{g_t - \boldsymbol{A}^\top(\boldsymbol{\lambda}_t - \rho(\boldsymbol{A}\boldsymbol{w}_t - \boldsymbol{y}_t)\right\} + \boldsymbol{w}_t\right]$$

- RDA型ADMM

$$\boldsymbol{w}_{t+1} \leftarrow \Pi_{\mathcal{W}}\left[-\frac{\eta_t}{\gamma}\left\{\bar{g}_t - \boldsymbol{A}^\top(\bar{\boldsymbol{\lambda}}_t - \rho(\boldsymbol{A}\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{y}}_t)\right\}\right]$$

$$\boldsymbol{y}_{t+1} \leftarrow \mathrm{prox}(\boldsymbol{A}\boldsymbol{w}_{t+1} - \boldsymbol{\lambda}_t/\rho|\tilde{\psi}/\rho)$$

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \rho(\boldsymbol{A}\boldsymbol{w}_{t+1} - \boldsymbol{y}_{t+1})$$

実装が簡単！

収束レート

データ: $D_{1:T-1} = (\boldsymbol{x}_t, y_t)_{t=1}^{T-1}$

- 一般の凸ロス関数

$$\mathrm{E}_{D_{1:T-1}}[\mathrm{E}_{(\boldsymbol{x},y)}[\ell(y, \boldsymbol{x}^\top \bar{\boldsymbol{w}}_T) + \psi(\bar{\boldsymbol{w}}_T) - \ell(y, \boldsymbol{x}^\top \boldsymbol{w}^*) - \psi(\boldsymbol{w}^*)]] \leq \frac{C}{\sqrt{T}}$$

- 強凸正則化関数

$$\mathrm{E}_{D_{1:T-1}}[\mathrm{E}_{(\boldsymbol{x},y)}[\ell(y, \boldsymbol{x}^\top \bar{\boldsymbol{w}}_T) + \psi(\bar{\boldsymbol{w}}_T) - \ell(y, \boldsymbol{x}^\top \boldsymbol{w}^*) - \psi(\boldsymbol{w}^*)]] \leq C\frac{\log(T)}{T}$$

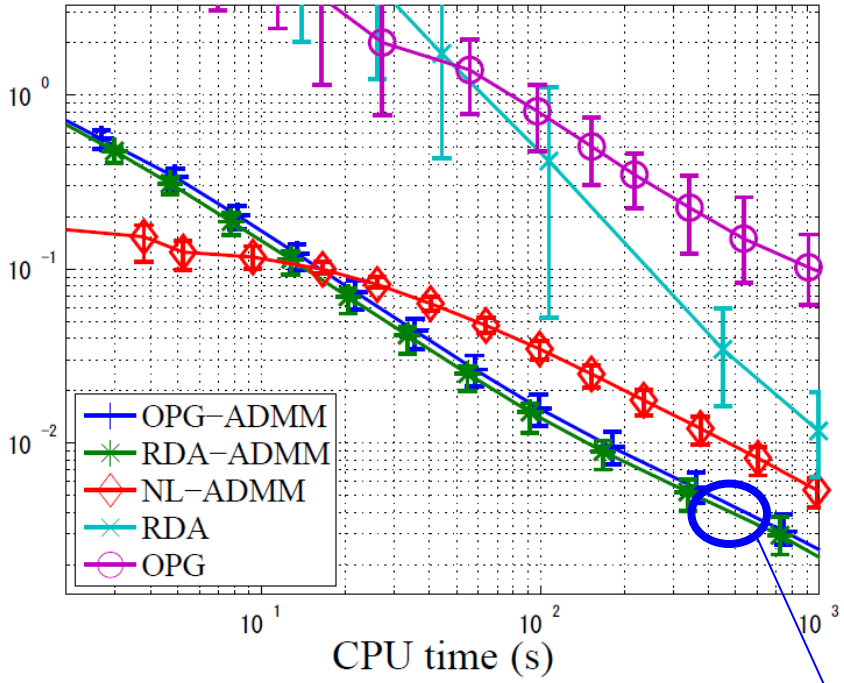
条件

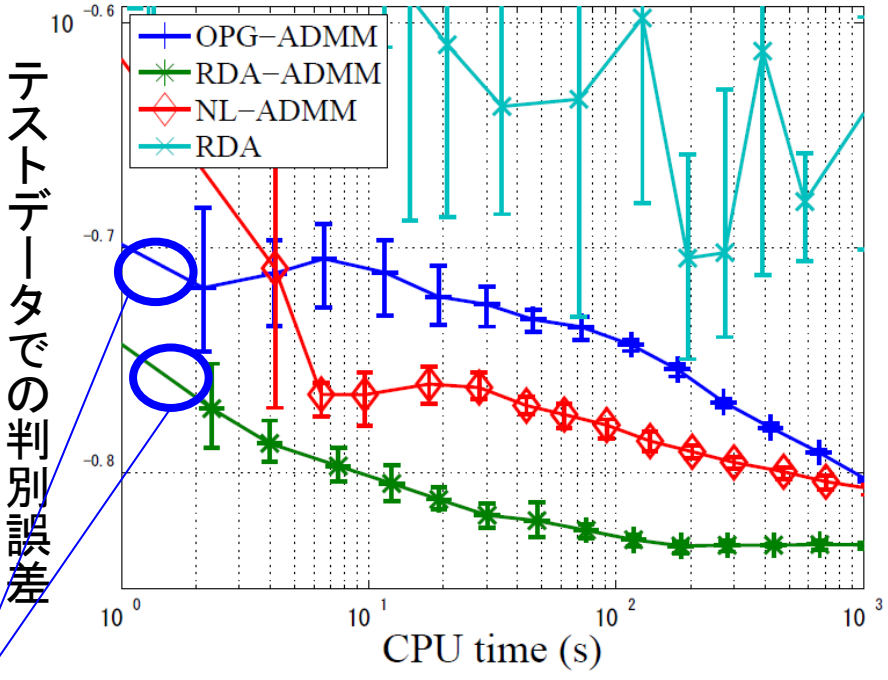
- データはi.i.d.系列
- ロスと正則化項はLipschitz連続
- wのドメインは有界

数値実験：確率的ADMM

人工データ

実データ（Adult, a9a
@LIVSVM data sets)





提案手法

1,024次元
512サンプル
重複ありグループ正則化

15,252次元
32,561サンプル
重複ありグループ正則化+ L1正則化

バッチデータに対する確率的最適化

Stochastic Dual Coordinate Ascent

[Shalev-Shwartz&Zhang,2012]

$$\min_{\boldsymbol{\rho}} \frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\rho_i) + \psi^*\left(-\frac{1}{n}X\boldsymbol{\rho}\right)$$

1. iをランダムに選択$(1 \leq i \leq n)$

2.次元i方向に最適化

$$\min_{\Delta\rho_i} \frac{1}{n}\ell_i^*(\Delta\rho_i + \rho_i) + \psi^*\left(-\frac{1}{n}X\boldsymbol{\rho} - X_i\frac{\Delta\rho_i}{n}\right)$$

3. 上の1,2を繰り返す.

ℓ_i^*が強凸でψ^*が滑らかな時,

双対ギャップの期待値 $\leq C\left(1-\frac{s}{n}\right)^T$

まとめ

- 正則化学習の最適化法

  proximal point algorithm ⇔ 乗数法

双対の関係

prox. point alg. FOBOS (近似解法・近接勾配法)	**乗数法** ADMM (近似解法)
乗数法 DAL	**prox. point alg.**

- 確率的最適化法
  - FOBOS・RDA
  - 確率的ADMM

←最近のトレンド