Dual Augmented Lagrangian Algorithm 法による スパース正則化

冨岡 亮太[1], **鈴木 大慈**[1], **杉山 将**[2]

[1] **東大 数理情報学専攻**
[2] **東工大 計算工学専攻**

tomioka@mist.i.u-tokyo.ac.jp

2010-4-13 @ **統計学輪講**

スパース回帰（組み合わせ的）

- **入出力の組み**

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m), \quad (\boldsymbol{x}_i \in \mathbb{R}^n).$$

- **仮定：**

$$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- **経験誤差：**

$$L(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)^2 = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{w}\|^2.$$

- **問題：**

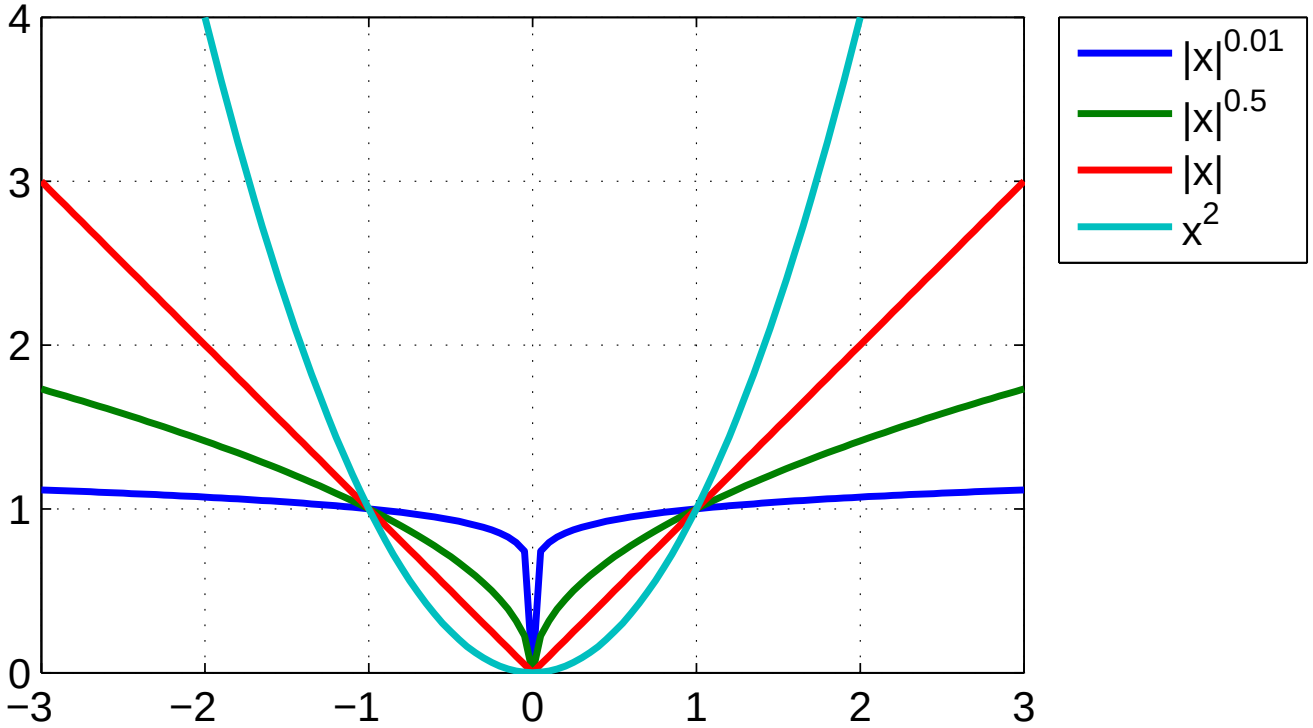
$$\text{minimize} \quad L(\boldsymbol{w}), \quad \text{s.t.} \ \|\boldsymbol{w}\|_0 \leq C \quad \text{(NP 困難!!)}$$

- **動機 1：現実には多くの場合 $m < n$.**
- **動機 2：なるべく少ない数の変数で説明したい！**
  （$\boldsymbol{w}$ の非ゼロ要素の数 $\|\boldsymbol{w}\|_0$ が少なければ少ないほどよい）

スパース回帰（連続）

- p-ノルムの p 乗（のようなもの）

$$\|\boldsymbol{w}\|_p^p = \sum_{j=1}^n |w_j|^p : \begin{cases} p \geq 1 \text{ ならば凸} \\ p < 1 \text{ ならば非凸} \end{cases}$$

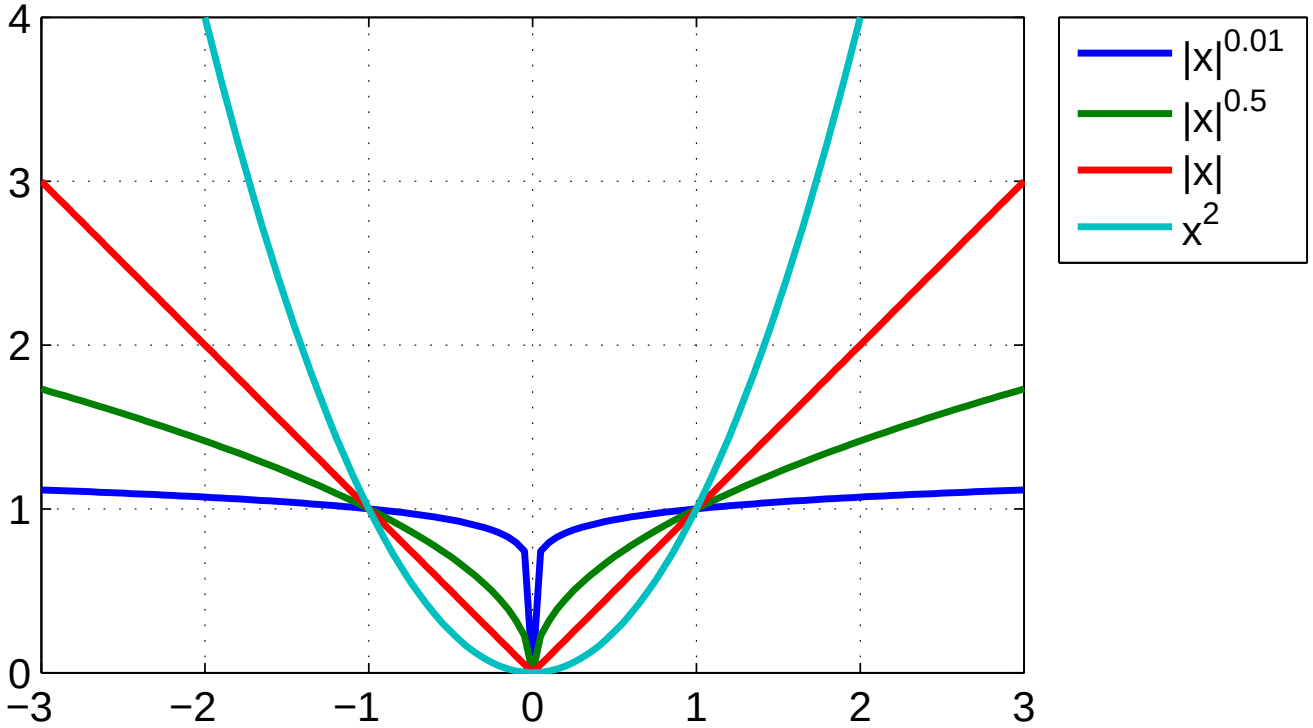


$\|w\|_1$ の正則化は凸の中ではもっとも $\|w\|_0$ の正則化に近い！

スパース回帰（連続）

- p-ノルムの p 乗（のようなもの）

$$\|\boldsymbol{w}\|_p^p = \sum_{j=1}^n |w_j|^p : \begin{cases} p \geq 1 \text{ ならば凸} \\ p < 1 \text{ ならば非凸} \end{cases}$$



$\|w\|_1$ の正則化は凸の中ではもっとも $\|w\|_0$ の正則化に近い！

Lasso 回帰（連続かつ凸）

- **問題 1:**

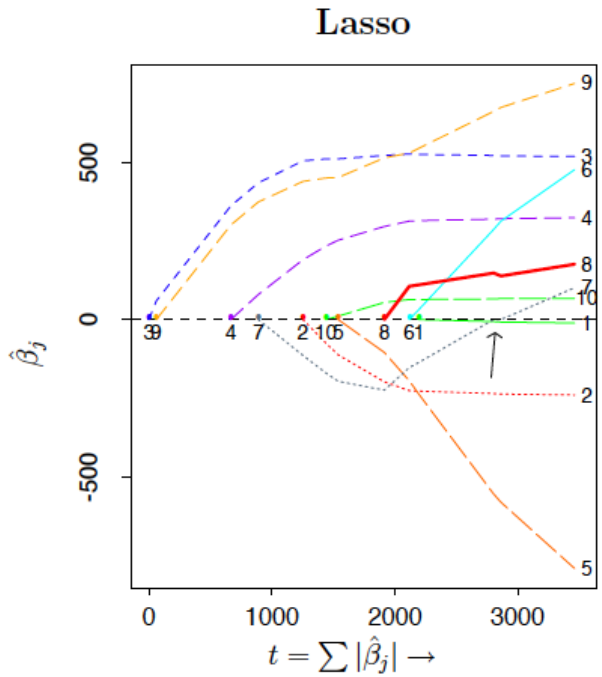
$$\text{minimize} \quad \|\boldsymbol{w}\|_1, \quad \text{s.t. } L(\boldsymbol{w}) \leq C.$$

- **問題 2:**

$$\text{minimize} \quad L(\boldsymbol{w}), \quad \text{s.t. } \|\boldsymbol{w}\|_1 \leq C'.$$

- **問題 3:**

$$\text{minimize} \quad L(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1.$$



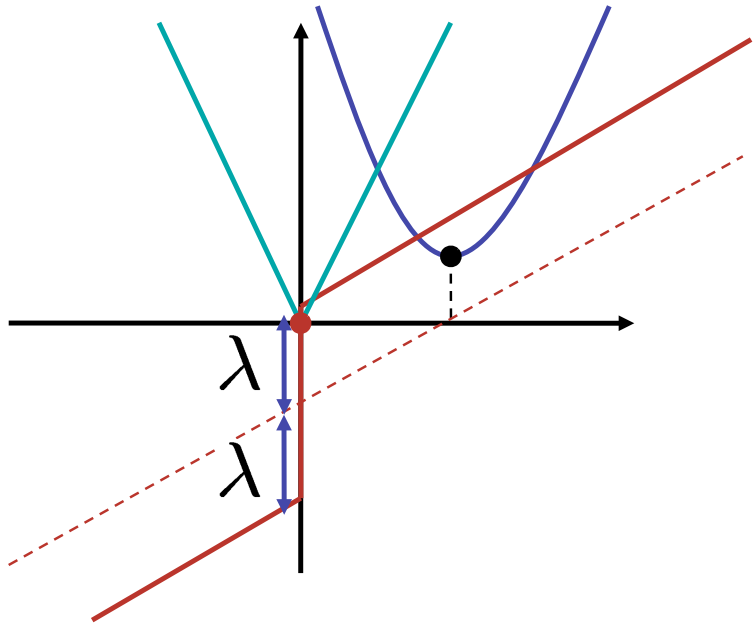
[From Efron et al. (2003)]

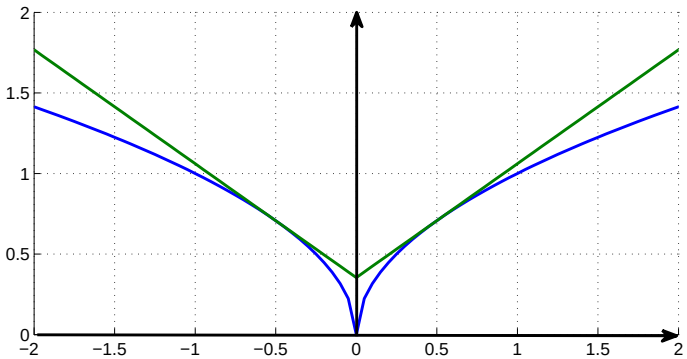
注意：

- **上の 3 つの問題はいずれも等価．**
- **正則化項やロス項に単調な非線形変換をしても等価．**
- **この発表では問題 3 を扱う．**

なぜ ℓ_1-正則化か？

- 凸の中で最も $\|\cdot\|_0$ に近い．
- 原点で微分不可能（有限の λ でゼロに打ち切ることができる．）
- 凸でない正則化
  $\longrightarrow$ 繰り返し（重み付き）ℓ_1-正則化問題を解けばよい．
- ベイズ周辺化尤度最大化
  $\longrightarrow$（特殊な場合に）繰り返し（重み付き）ℓ_1-正則化問題を解けばよい．
  (Wipf&Nagarajan, 08)





問題設定

以下の最適化問題を効率良く解くためのアルゴリズムが求められている．

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w}).$$

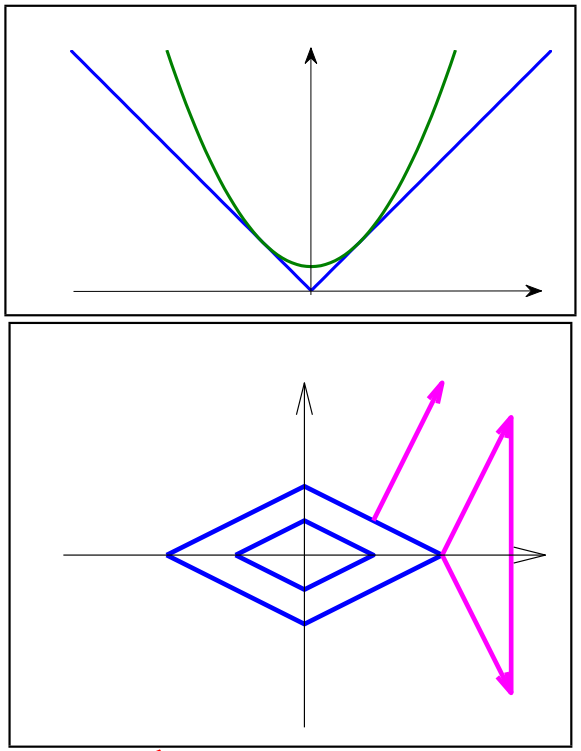
ただし，

- $\boldsymbol{A} \in \mathbb{R}^{m\times n}$（$m$: サンプル数，$n$: 未知変数の数）．
- f_ℓ は凸で 2 回微分可能．
- $\phi_\lambda(\boldsymbol{w})$ は例えば，$\phi_\lambda(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$ など，凸だが，微分不可能であってもよい．また，$\eta\phi_\lambda = \phi_{\eta\lambda}$ を仮定．
- 特定の f_ℓ に依存したアルゴリズム（LARS など）ではもの足りない．
- No Free Lunch – 観測の数が変数の数より少ない場合 ($m \ll n$) や $\boldsymbol{A}$ のコンディションが悪い場合が応用上重要．

どこが難しいか？

今までの見方: $\phi_\lambda(\boldsymbol{w})$ **の微分不可能性が原因．**

- **正則化項を微分可能な関数で上から押さえる．**
  - FOCUSS (Rao & Kreutz-Delgado, 99)
  - Majorization-Minimization (Figueiredo et al., 07)
- **微分不可能性を陽に考慮する．**
  - Sub-gradient L-BFGS (Andrew & Gao, 07; Yu et al., 08)



我々の見方: ***A* が変数の間にからみを導入するのが原因．**

どこが難しいか？

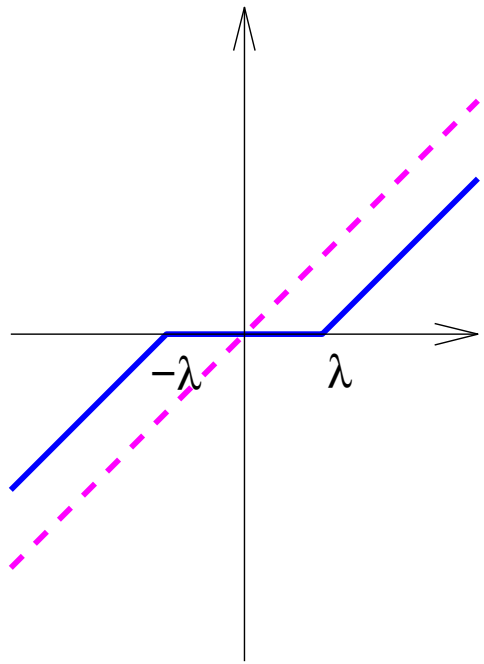
我々の見方:　**A が変数の間にからみを導入するのが原因．**

$A = I_n$ （単位行列の場合）

$$\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_1\right)=\sum_{j=1}^{n}\min_{w_j\in\mathbb{R}}\left(\frac{1}{2}(y_j-w_j)^2+\lambda|w_j|\right).$$

$$\Rightarrow\quad w_j^* = \mathrm{ST}_\lambda(y_j)$$
$$= \begin{cases} y_j-\lambda & (\lambda \le y_j), \\ 0 & (-\lambda \le y_j \le \lambda), \\ y_j+\lambda & (y_j \le -\lambda). \end{cases}$$

解析的に解ける！



本発表では ϕ_λ として，上の最小化が解析的に求められるもののみを扱う．

ϕ_λ に関する proximation は解析的に計算できる

仮定

ϕ_λ に関する proximation (soft-thresholding)

$$\mathrm{ST}_\lambda(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\phi_\lambda(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}\|_2^2 \right)$$

は解析的に計算できる.

- **グループラッソー**

$$\phi_\lambda(\boldsymbol{w}) = \sum_{\mathfrak{g}\in\mathfrak{G}} \|\boldsymbol{w}_\mathfrak{g}\|_2 : \textbf{グループ単位のノルムの和.}$$

- **トレースノルム**

$$\phi_\lambda(\boldsymbol{W}) = \sum_{j=1}^{r} \sigma_j(\boldsymbol{W}) : \quad \textbf{特異値の和（}\boldsymbol{W}\textbf{ は行列）.}$$

Proximation の解釈（脱線）

- Proximation は 集合に対する射影の一般化.
- 集合 C の定義関数 $\delta_C(\boldsymbol{x})$ を

$$\delta_C(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in C \text{ のとき}, \\ +\infty, & \text{その他}, \end{cases}$$

  と定義する.

  δ_C に関する proximation $\equiv$ 集合 C への２乗距離の意味での射影.

- Proximation による分解：

$$\mathrm{prox}_f(\boldsymbol{z}) + \mathrm{prox}_{f^*}(\boldsymbol{z}) = \boldsymbol{z}.$$

  ただし，f と f^* は凸共役な関数の組み.

- 参照： Moreau 1965, Rockafellar 1970.

発表の流れ

1. イントロ
   - スパース正則化
   - どこが難しいか： ~~微分不可能性~~ $\Rightarrow$ 変数の間のからみ
2. 最適化アルゴリズム
   - Iterative shrinkage-thresholding (IST)
   - Dual Augmented Lagrangian （提案手法）
3. 収束レート： 超１次収束
   - 厳密な内部最小化の場合
   - 近似的な 内部最小化の場合
4. 実験
   - OWLQN, SpaRSA, and FISTA　との比較．
5. まとめ

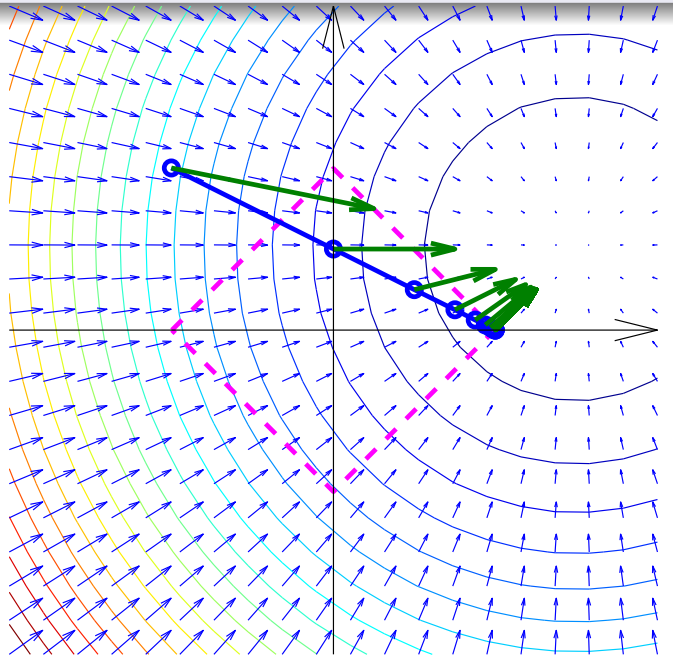
Iterative Shrinkage/Thresholding (IST)

IST 法 (Figueiredo&Nowak, 03; Daubechies et al., 04;...)

1. 適当に初期解 w^0 を決める.
2. 停止条件が満たされるまで反復：

$$\boldsymbol{w}^{t+1} \leftarrow \underbrace{\mathrm{ST}_{\eta_t \lambda}}_{\text{縮小}}\Big(\underbrace{\boldsymbol{w}^t - \eta_t \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)}_{\text{勾配ステップ}}\Big).$$

- 利点: 実装が簡単.
- 欠点: デザイン行列 A の条件数が悪いと遅い.



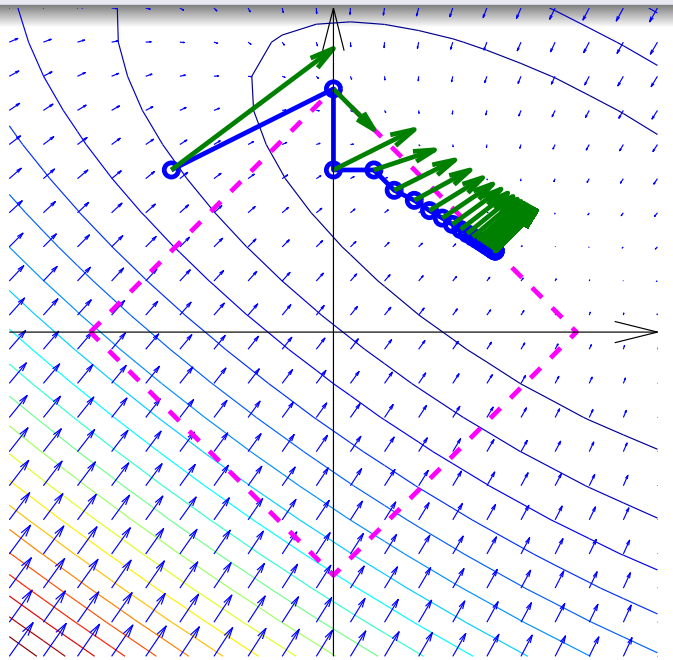
Iterative Shrinkage/Thresholding (IST)

IST 法 (Figueiredo&Nowak, 03; Daubechies et al., 04;...)

1. 適当に初期解 w^0 を決める.
2. 停止条件が満たされるまで反復：

$$w^{t+1} \leftarrow \underbrace{\mathrm{ST}_{\eta_t \lambda}}_{\text{縮小}}\Big(\underbrace{w^t - \eta_t A^\top \nabla f_\ell(Aw^t)}_{\text{勾配ステップ}}\Big).$$

- **利点**: 実装が簡単.
- **欠点**: デザイン行列 A の条件数が悪いと遅い.



Dual Augmented Lagrangian (DAL) 法（提案手法）

主問題

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

双対問題

$$\underset{\boldsymbol{\alpha},\boldsymbol{v}}{\text{maximize}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v})$$

$$\text{s.t.} \quad \boldsymbol{v} = \boldsymbol{A}^\top \boldsymbol{\alpha}$$

- f_ℓ^*，ϕ_λ^* は f_ℓ，ϕ_λ の凸共役：

$$f_\ell^*(\boldsymbol{\alpha}) = \sup_{\boldsymbol{z}\in\mathbb{R}^m} \left(\langle \boldsymbol{\alpha}, \boldsymbol{z}\rangle - f_\ell(\boldsymbol{z})\right)$$

$$\phi_\lambda^*(\boldsymbol{v}) = \sup_{\boldsymbol{w}\in\mathbb{R}^n} \left(\langle \boldsymbol{v}, \boldsymbol{w}\rangle - \phi_\lambda(\boldsymbol{w})\right)$$

- **主問題の最小値 = 双対問題の最大値（強双対性）**

Dual Augmented Lagrangian (DAL) 法（提案手法）

主問題

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

Proximal minimization:

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$$

$$(\eta_0 \leq \eta_1 \leq \cdots)$$

- 解析がしやすい．例えば
  $f(\boldsymbol{w}^{t+1}) + \frac{1}{2\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \leq f(\boldsymbol{w}^t)$.
- 実用的でない（もとの問題と同程度に難しい！）

双対問題

$$\underset{\boldsymbol{\alpha},\boldsymbol{v}}{\text{maximize}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v})$$

$$\text{s.t.} \quad \boldsymbol{v} = \boldsymbol{A}^\top \boldsymbol{\alpha}$$

⇔Augmented Lagrangian (Tomioka & Sugiyama, 09):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \alpha^t)$$

$$\alpha^t = \underset{\alpha}{\operatorname{argmin}}\, \varphi_t(\alpha)$$

- $\varphi_t(\alpha)$ の最小化は簡単（なめらか）.
- ステップサイズ η_t は増加.
- 同値性については Rockafellar 76 を参照.

Dual Augmented Lagrangian (DAL) 法（提案手法）

主問題

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

Proximal minimization:

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\text{argmin}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$$

$$(\eta_0 \leq \eta_1 \leq \cdots)$$

- **解析がしやすい．例えば**
  $f(\boldsymbol{w}^{t+1}) + \frac{1}{2\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \leq f(\boldsymbol{w}^t)$.
- **実用的でない（もとの問題と同程度に難しい！）**

双対問題

$$\underset{\boldsymbol{\alpha},\boldsymbol{v}}{\text{maximize}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v})$$

$$\text{s.t.} \quad \boldsymbol{v} = \boldsymbol{A}^\top \boldsymbol{\alpha}$$

⇔Augmented Lagrangian (Tomioka & Sugiyama, 09):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \alpha^t)$$

$$\alpha^t = \underset{\alpha}{\text{argmin}}\, \varphi_t(\alpha)$$

- $\varphi_t(\alpha)$ **の最小化は簡単（なめらか）．**
- **ステップサイズ** η_t **は増加．**
- **同値性については** Rockafellar 76 **を参照．**

Dual Augmented Lagrangian (DAL) 法（提案手法）

主問題

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \phi_\lambda(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

Proximal minimization:

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\text{argmin}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$$

$$(\eta_0 \leq \eta_1 \leq \cdots)$$

- **解析がしやすい．例えば**
  $f(\boldsymbol{w}^{t+1}) + \frac{1}{2\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \leq f(\boldsymbol{w}^t)$.
- **実用的でない（もとの問題と同程度に難しい！）**

双対問題

$$\underset{\boldsymbol{\alpha},\boldsymbol{v}}{\text{maximize}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - \phi_\lambda^*(\boldsymbol{v})$$

$$\text{s.t.} \quad \boldsymbol{v} = \boldsymbol{A}^\top \boldsymbol{\alpha}$$

⇔Augmented Lagrangian (Tomioka & Sugiyama, 09):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t)$$

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha}}{\text{argmin}}\, \varphi_t(\boldsymbol{\alpha})$$

- $\varphi_t(\boldsymbol{\alpha})$ **の最小化は簡単（なめらか）．**
- **ステップサイズ** η_t **は増加．**
- **同値性については** Rockafellar 76 **を参照．**

Dual Augmented Lagrangian 法 (ℓ_1-正則化)

1. 適当に初期解 $\boldsymbol{w}^1$ を決める．
2. 停止条件が満たされるまで反復：

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\eta_t \lambda}\left(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t\right)$$

ただし，

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\mathrm{argmin}} \Big(\underbrace{f_\ell^*(-\boldsymbol{\alpha})}_{\text{損失関数 } f_\ell \text{ の凸共役}} + \frac{1}{2\eta_t} \|\mathrm{ST}_{\eta_t \lambda}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha})\|_2^2 \Big)$$

Dual Augmented Lagrangian 法（ℓ_1-正則化の場合）

(1) **この計算は解析的にできると仮定．**

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\eta_t\lambda}\left(\boldsymbol{w}^t + \boldsymbol{A}^\top \alpha^t\right)$$

(2) **この計算はスパースであればあるほど効率的．** （**“アクティブな”未知変数の数に線形**）

$$\alpha^t = \underset{\alpha\in\mathbb{R}^m}{\mathrm{argmin}}\Big(\underbrace{f_\ell^*(-\alpha)}_{\text{A のスケーリングの影響を受けない}} + \underbrace{\frac{1}{2\eta_t}\|\mathrm{ST}_{\eta_t\lambda}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top\alpha)\|_2^2}_{\substack{\frac{\partial}{\partial\alpha}:\ \boldsymbol{A}\mathrm{ST}_{\eta_t\lambda}(\boldsymbol{w}^t+\eta_t\boldsymbol{A}^\top\alpha) \\ \frac{\partial}{\partial\alpha^2}:\ \eta_t\boldsymbol{A}_+{\boldsymbol{A}_+}^\top}} \Big)$$

（**$\boldsymbol{A}_+$ は $\boldsymbol{A}$ の“アクティブな”列からなる部分行列**; ℓ_1-**正則化の場合**）

(3) $\boldsymbol{A}$ **のスケーリングの悪さの影響を受けにくい．**

IST と DAL の違い：いかに変数の間の絡みを取り除くか

目的関数 f に関する Proximation **は難しい：**

$$\boldsymbol{w}^{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \Big(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{変数が絡みあっている}} + \phi_\lambda(\boldsymbol{w})}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Big)$$

- IST（既存手法）：線形にロス項を 近似：

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) \simeq f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)$$

$\rightarrow$ 現在の点 $\boldsymbol{w}^t$ で最もタイト

- DAL（提案法）：線形なロス項の 下限

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \left(-f_\ell^*(-\boldsymbol{\alpha}) - \boldsymbol{w}^\top \boldsymbol{A}^\top \boldsymbol{\alpha} \right)$$

$\rightarrow$ 次の点 $\boldsymbol{w}^{t+1}$ で最もタイト

IST と DAL の違い：いかに変数の間の絡みを取り除くか

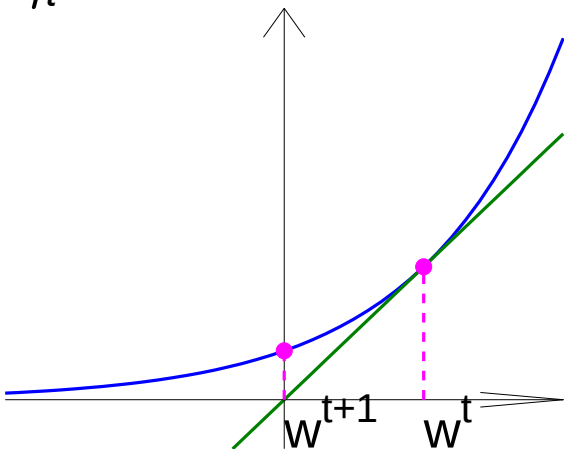
目的関数 f に関する Proximation は難しい：

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \Big(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{変数が絡みあっている}} + \phi_\lambda(\boldsymbol{w})}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Big)$$

- IST（既存手法）：線形にロス項を 近似：

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) \simeq f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)$$

$\rightarrow$ **現在の点 $\boldsymbol{w}^t$ で最もタイト**



- DAL（提案法）：線形なロス項の 下限

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\alpha \in \mathbb{R}^m} \left(-f_\ell^*(-\alpha) - \boldsymbol{w}^\top \boldsymbol{A}^\top \alpha \right)$$

$\rightarrow$ **次の点 $\boldsymbol{w}^{t+1}$ で最もタイト**

IST と DAL の違い：いかに変数の間の絡みを取り除くか

目的関数 f に関する Proximation **は難しい：**

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \Big(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{変数が絡みあっている}} + \phi_\lambda(\boldsymbol{w})}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Big)$$

- IST（既存手法）：線形にロス項を 近似：

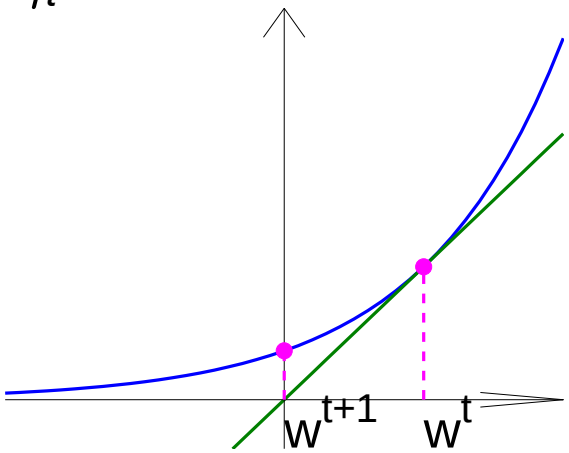
$$f_\ell(\boldsymbol{A}\boldsymbol{w}) \simeq f_\ell(\boldsymbol{A}\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \boldsymbol{A}^\top \nabla f_\ell(\boldsymbol{A}\boldsymbol{w}^t)$$

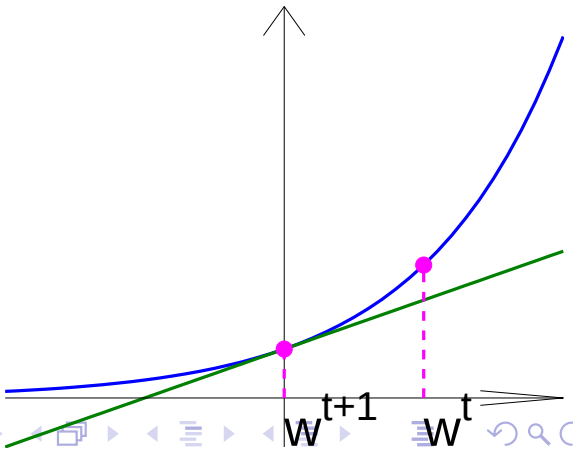
$\rightarrow$ **現在の点 $\boldsymbol{w}^t$ で最もタイト**

- DAL（提案法）：線形なロス項の 下限

$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left(-f_\ell^*(-\boldsymbol{\alpha}) - \boldsymbol{w}^\top \boldsymbol{A}^\top \boldsymbol{\alpha} \right)$$

$\rightarrow$ **次の点 $\boldsymbol{w}^{t+1}$ で最もタイト**

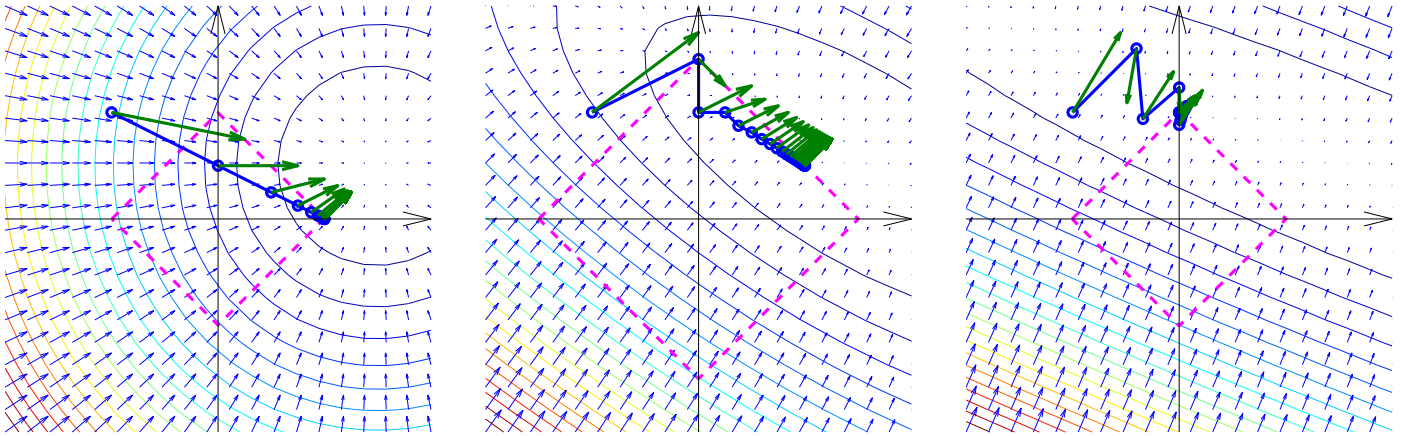




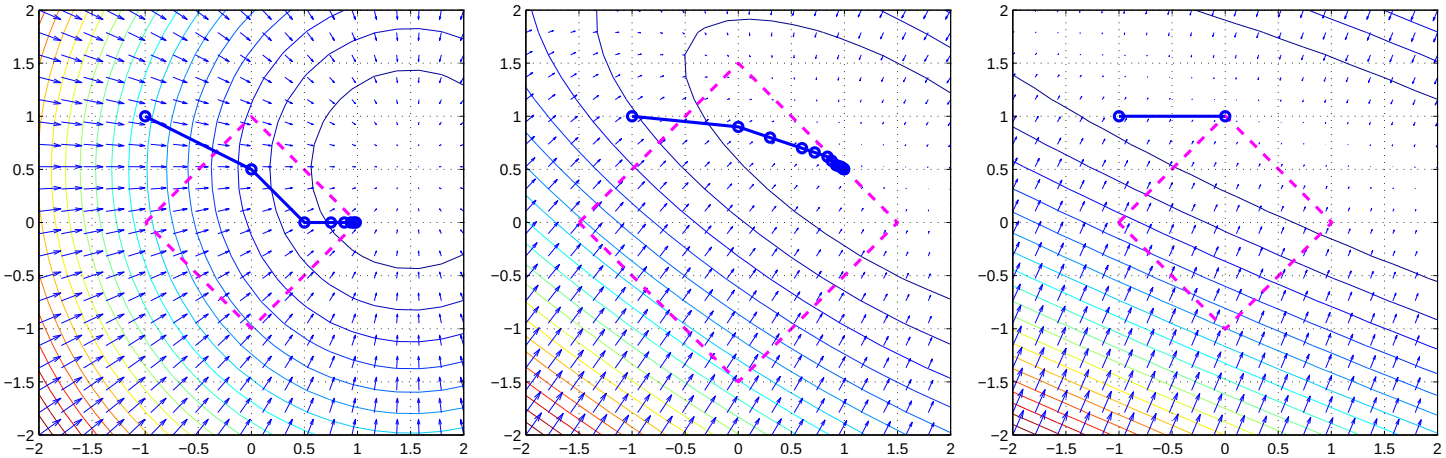
数値例

デザイン行列 A のコンデイションが悪くなるほど，DAL の方が有利．

IST



DAL



定理 1（厳密な最小化）

定義

- $\boldsymbol{w}^t$：厳密な DAL 法（$\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| = 0$）で得られる点列．
- $\boldsymbol{w}^*$： 目的関数 f を最小化する点．

仮定

正の定数 σ が存在して

$$f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2 \quad (t = 0, 1, 2, \ldots).$$

定理 1

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

$\Rightarrow \eta_t$ が増加するなら，$\boldsymbol{w}^t$ は $\boldsymbol{w}^*$ に超１次収束する．

定理 2（近似的最小化）

定義

- $\boldsymbol{w}^t$: 以下の停止基準による近似的な DAL 法で得られる点列．

$$\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| \leq \sqrt{\frac{\gamma}{\eta_t}}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\| \qquad \left(\begin{array}{l} 1/\gamma\text{: 損失関数の微分} \\ \nabla f_\ell \text{ のリプシッツ定数．} \end{array}\right)$$

定理 2

定理 1 と同じ仮定のもとで

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{1}{\sqrt{1 + 2\sigma\eta_t}}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|.$$

$\Rightarrow \eta_t$ が増加するなら，$\boldsymbol{w}^t$ は $\boldsymbol{w}^*$ に超１次収束する．

- 収束レートは厳密な場合 ($\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| = 0$) より少し悪い．
- 同程度の収束レートは内部最小化をもう少し厳しくすることで達成可能 $\frac{\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\|}{\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|} \leq O(1/\eta_t)$.

定理 2（近似的最小化）

定義

- $\boldsymbol{w}^t$: 以下の停止基準による近似的な DAL 法で得られる点列．

$$\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| \leq \sqrt{\frac{\gamma}{\eta_t}}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\| \qquad \left(\begin{array}{l} 1/\gamma\text{: 損失関数の微分} \\ \nabla f_\ell \text{ のリプシッツ定数.} \end{array}\right)$$

定理 2

定理 1 と同じ仮定のもとで

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{1}{\sqrt{1+2\sigma\eta_t}}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|.$$

$\Rightarrow \eta_t$ が増加するなら，$\boldsymbol{w}^t$ は $\boldsymbol{w}^*$ に超１次収束する．

- 収束レートは厳密な場合 ($\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\| = 0$) より少し悪い．
- 同程度の収束レートは内部最小化をもう少し厳しくすることで達成可能 $\frac{\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\|}{\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\|} \leq O(1/\eta_t)$.

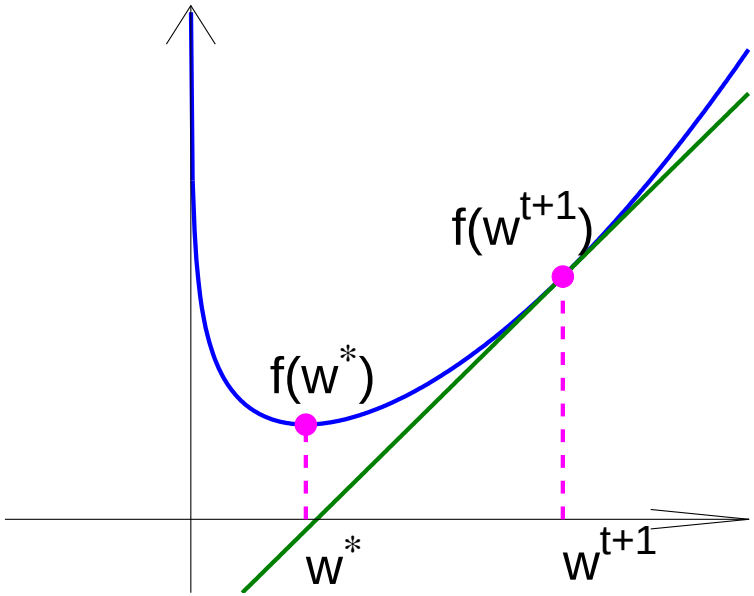
定理 1 の証明（エッセンス）

$\boldsymbol{w}^{t+1}$ は，$f(\boldsymbol{w}) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2$ を最小化するので，

$$(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t \in \partial f(\boldsymbol{w}^{t+1}) \quad \text{（劣微分に入る）}$$

従って (Beck & Teboulle 09)，

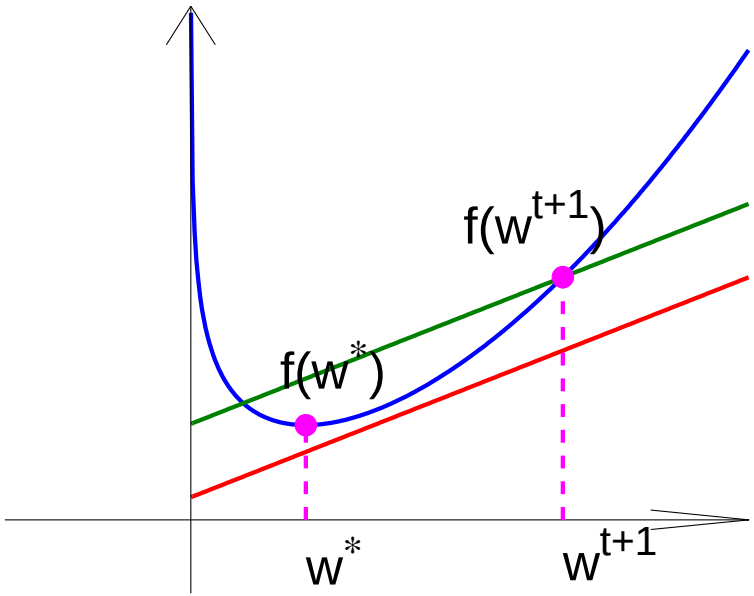
$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle .$$



定理 2 の証明（エッセンス）

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle \underbrace{-\frac{1}{2\gamma}\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\|^2}_{\text{近似最小化のコスト}} .$$

$1/\gamma$: **損失関数の微分 ∇f_ℓ のリプシッツ定数.**



他の手法との比較

	手法	説明	収束オーダー
1次	IST		$O(1/k)$
	FISTA	**IST の収束性を改善したもの**	$O(1/k^2)$
	SpaRSA	**曲率の情報を利用して IST のステップサイズを改善したもの**	?
	OWLQN	**劣微分を利用した擬似ニュートン法**	?
高次	**内点法**	Koh, Kim, & Boyd 2007	$O(e^{-k})$
	DAL	**提案法**	$o(e^{-k})$

数値実験: ℓ_1-正則化付きロジスティック回帰

#samples=1, 024, #unknowns=16, 384.

- FISTA

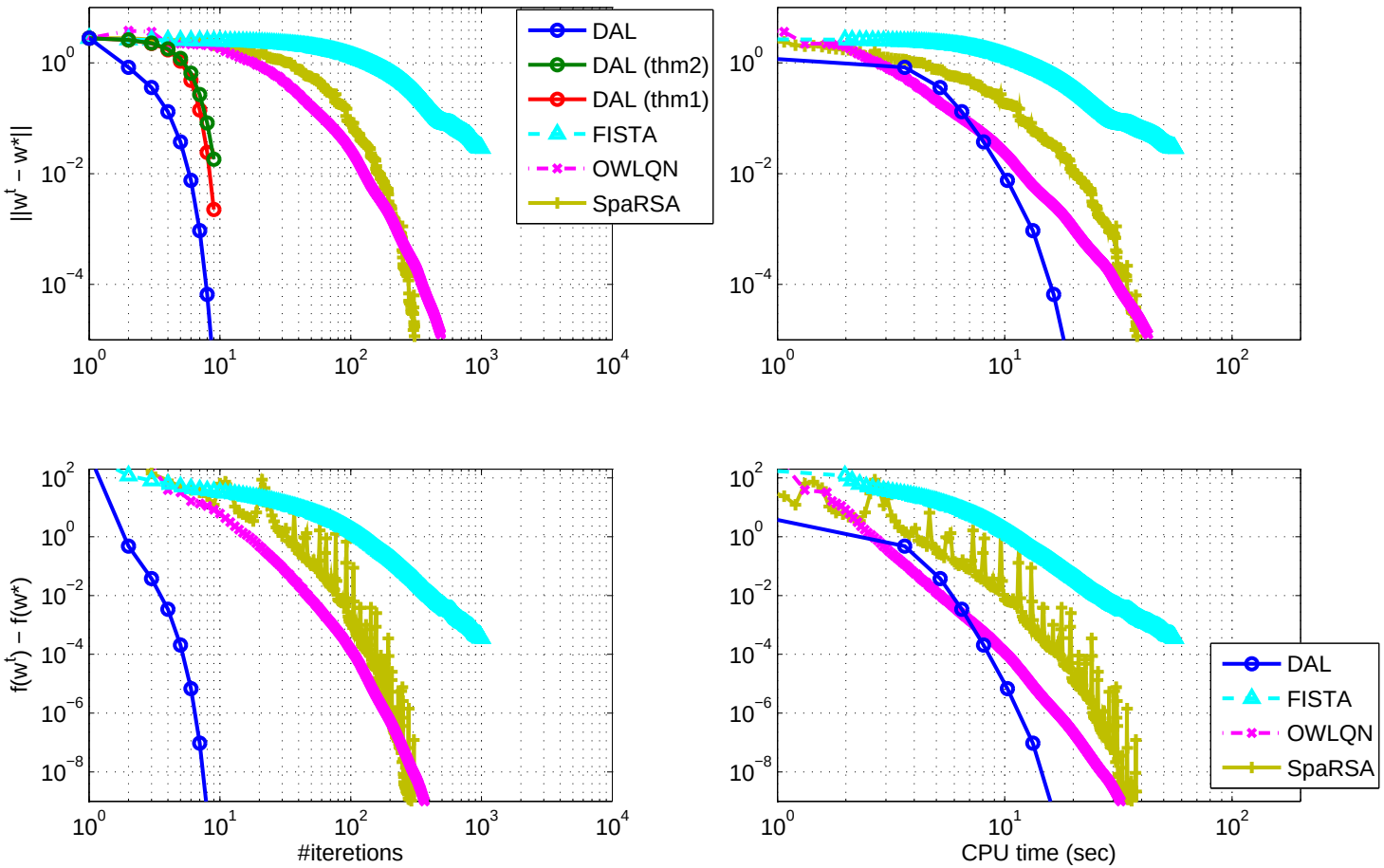
2段階 IST (Beck & Teboulle 09)

- OWLQN

劣微分準ニュートン (Andrew & Gao 07)

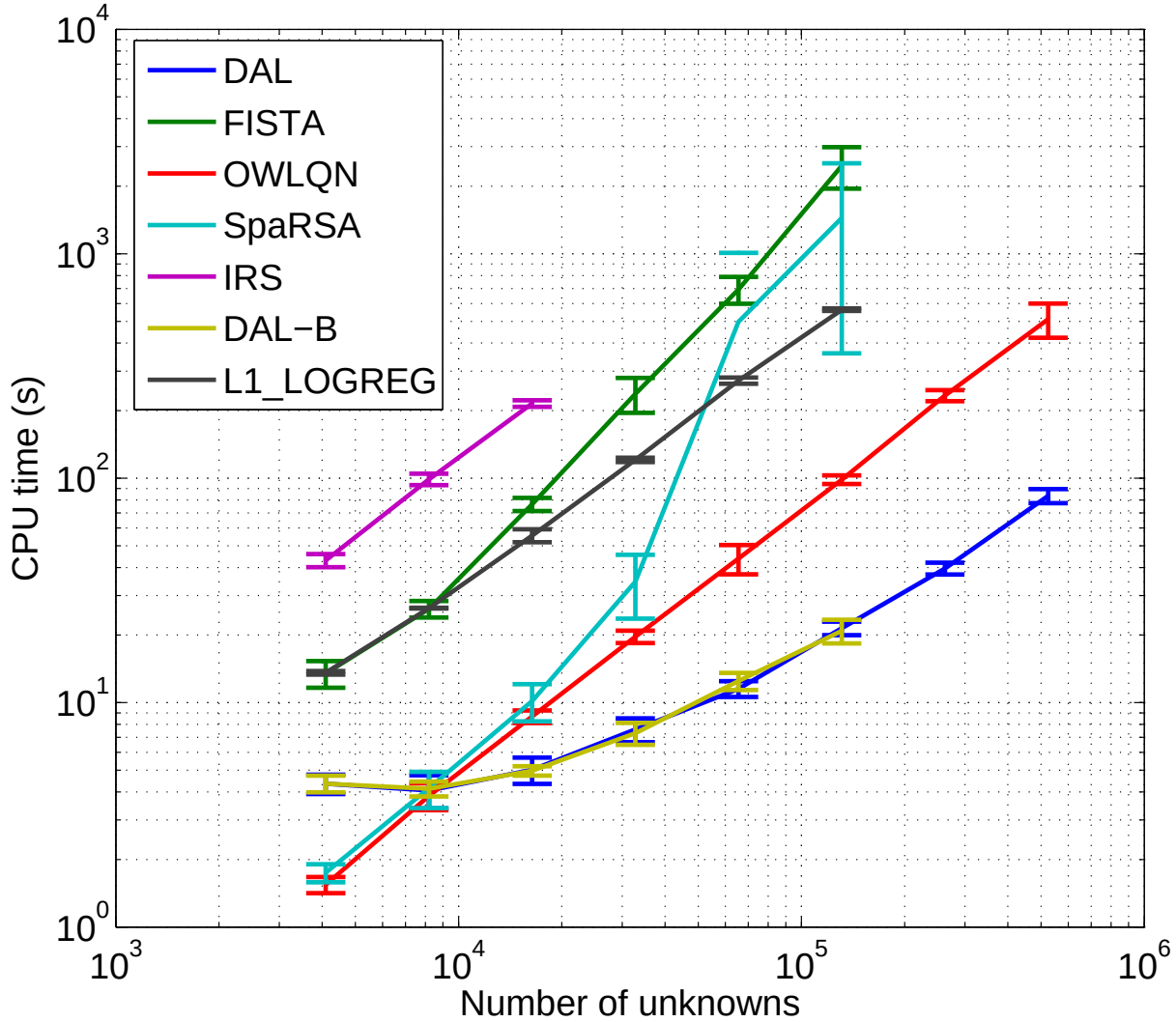
- SpaRSA

ステップサイズ改良 IST (Wright et al. 09)



未知変数の数に対する計算時間の増加

$m = 1{,}024$. $n = 4{,}096$–$524{,}288$



脳波解析における接続関係の推定

独立・
非ガウス信号

源信号

観測信号

ξ

$$z(t) = \sum_{p=1}^{P} \boldsymbol{H}^{(p)} \boldsymbol{z}(t-p) + \boldsymbol{\xi}(t)$$

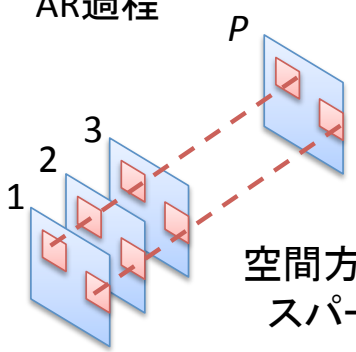
z

$$\boldsymbol{x}(t) = \boldsymbol{M}\boldsymbol{z}(t)$$

x

時空間
スパース
AR過程

空間方向
即時的
混合



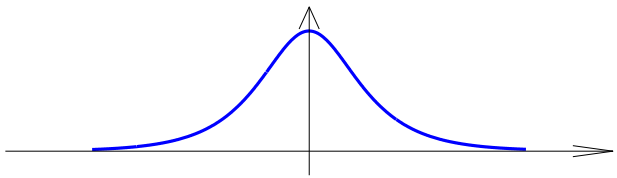
空間方向に
スパース

S. Haufe, R. Tomioka, G. Nolte,
K-R Müller and M. Kawanabe,
IEEE TBME 2010. accepted.

EM アルゴリズム

- E-ステップ：隠れ信号 $\boldsymbol{z}(t)$ の推定（$\boldsymbol{B}$ に関する最尤法）
  - 混合行列の逆行列 $\boldsymbol{B} := \boldsymbol{M}^{-1}$ とする．
  - $\boldsymbol{\xi}(t) = \boldsymbol{B}\boldsymbol{x}(t) - \sum_{p=1}^{P} \boldsymbol{H}^{(p)} \boldsymbol{B}\boldsymbol{x}(t-p)$ は *sech* 分布に従うと仮定．

$$p(\boldsymbol{\xi}) \propto \prod_{d=1}^{D} \frac{1}{e^{-\xi_d} + e^{\xi_d}}$$



- M-ステップ：AR 係数の推定（$\boldsymbol{H}^{(p)}$ に関するスパース推定）
  - 尤度：*sech* 分布
  - 正則化：（空間方向に）スパースな接続関係をみつけたい
    $\Rightarrow$ 時間方向にグループしたグループラッソー正則化．

結果（人工データ）

SCSA_EM（提案手法）が混合行列 M の推定および結合係数 $H^{(P)}$ の推定の両方の意味で優れている．

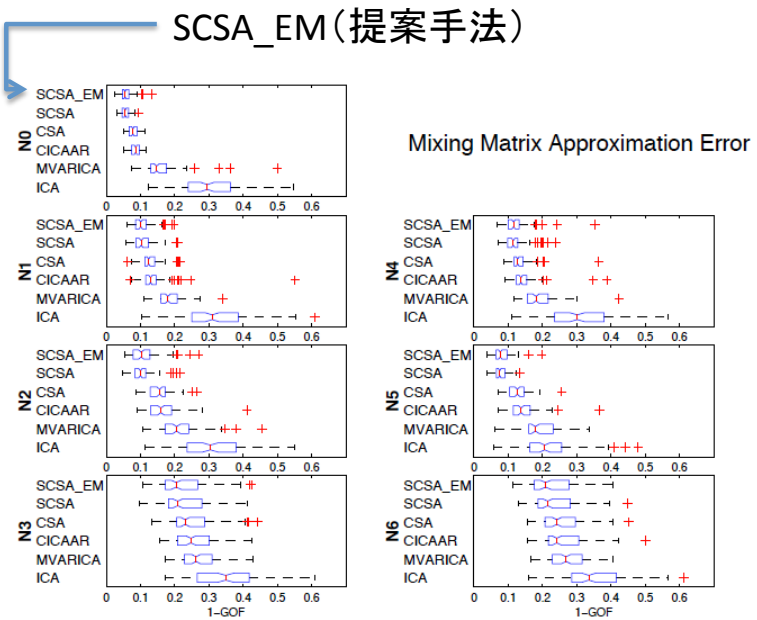


Fig. 3. Estimation errors of the mixing matrix according to the goodness-of-fit (GOF) criterion. Results are shown for the proposed (Sparsely-) Connected Sources Analysis variants (SCSA_EM, SCSA, CSA) and three alternative approaches (CICAAR, MVARICA, ICA). Different subfigures depict the methods' performance in the noiseless cass (N0), as well as in the presence of different types of noise (N1-N6, see TABLE I).

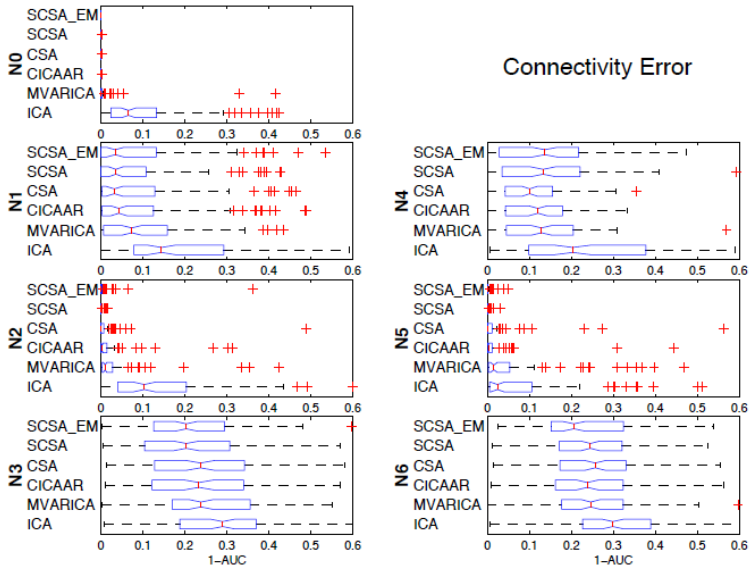


Fig. 5. Estimation errors regarding the source connectivity structure as measured by fitting an MVAR model subsequently to the demixed sources and testing the obtained coefficients for significant interaction. The performance measure reported is the area under the curve (AUC) score obtained by varying the significance level.

Summary

- なぜスパース正則化は難しい? – 変数の間の絡み
  - 微分不可能性は悪くない.
  - 内部最小化はスパースであればあるほど速い.
    微分不可能性は効率的な内部最小化のために使える
- いかに変数の間の絡みを取り除くか
  - 線形近似の代わりに パラメトリックな下限を使う.
- 厳密／近似的内部最小化のもとでの超１次収束.
  - スパース正則化の状況に特殊化することで，現実的・使いやすい条件のもとで，既存の収束レートを改善.
- 実験結果も理論をサポート
  - 既存手法 OWLQN, SpaRSA, and FISTA より速い．かつより複雑な正則化に対応可能.

Summary

- なぜスパース正則化は難しい? – 変数の間の絡み
  - 微分不可能性は悪くない.
  - 内部最小化はスパースであればあるほど速い.
    微分不可能性は効率的な内部最小化のために使える
- いかに変数の間の絡みを取り除くか
  - 線形近似の代わりに パラメトリックな下限を使う.
- 厳密／近似的内部最小化のもとでの超１次収束.
  - スパース正則化の状況に特殊化することで，現実的・使いやすい条件のもとで，既存の収束レートを改善.
- 実験結果も理論をサポート
  - 既存手法 OWLQN, SpaRSA, and FISTA より速い．かつより複雑な正則化に対応可能.

Summary

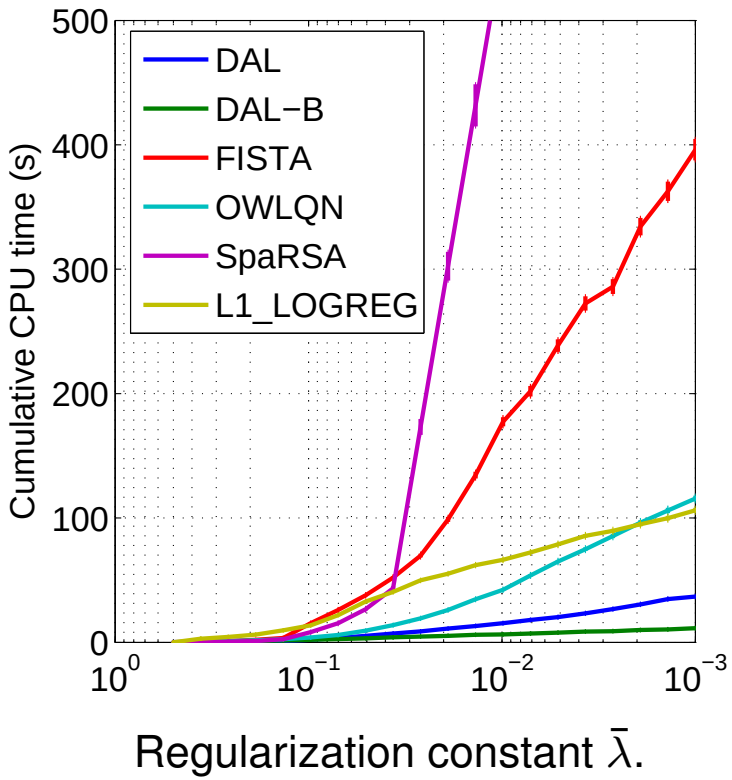
- なぜスパース正則化は難しい? – 変数の間の絡み
  - 微分不可能性は悪くない.
  - 内部最小化はスパースであればあるほど速い.
    微分不可能性は効率的な内部最小化のために使える
- いかに変数の間の絡みを取り除くか
  - 線形近似の代わりに パラメトリックな下限を使う.
- 厳密／近似的内部最小化のもとでの超１次収束.
  - スパース正則化の状況に特殊化することで，現実的・使いやすい条件のもとで，既存の収束レートを改善.
- 実験結果も理論をサポート
  - 既存手法 OWLQN, SpaRSA, and FISTA より速い．かつより複雑な正則化に対応可能.

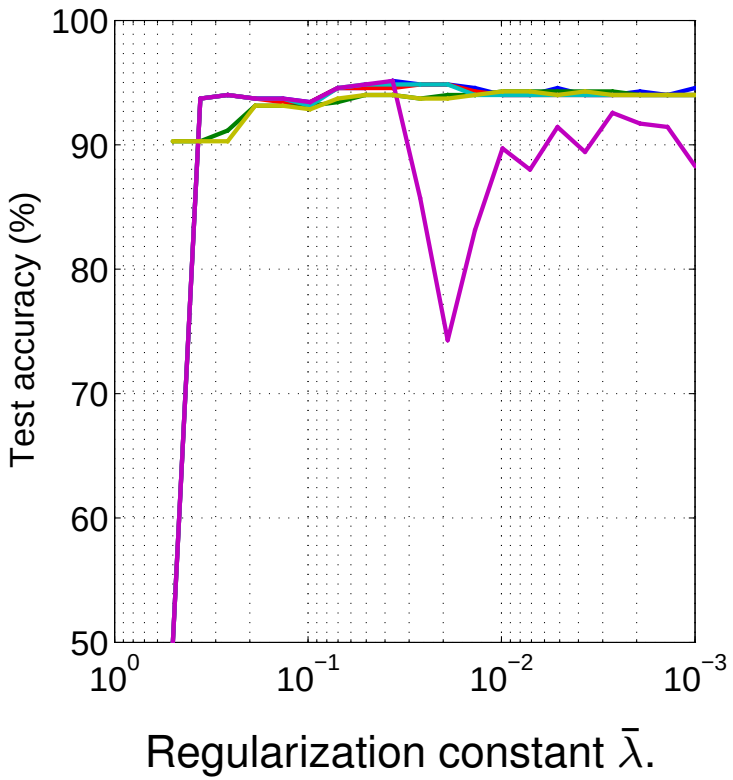
Summary

- なぜスパース正則化は難しい? – 変数の間の絡み
  - 微分不可能性は悪くない.
  - 内部最小化はスパースであればあるほど速い.
    微分不可能性は効率的な内部最小化のために使える
- いかに変数の間の絡みを取り除くか
  - 線形近似の代わりに パラメトリックな下限を使う.
- 厳密／近似的内部最小化のもとでの超１次収束.
  - スパース正則化の状況に特殊化することで，現実的・使いやすい条件のもとで，既存の収束レートを改善.
- 実験結果も理論をサポート
  - 既存手法 OWLQN, SpaRSA, and FISTA より速い．かつより複雑な正則化に対応可能.

Benchmark datasets (dorothea)

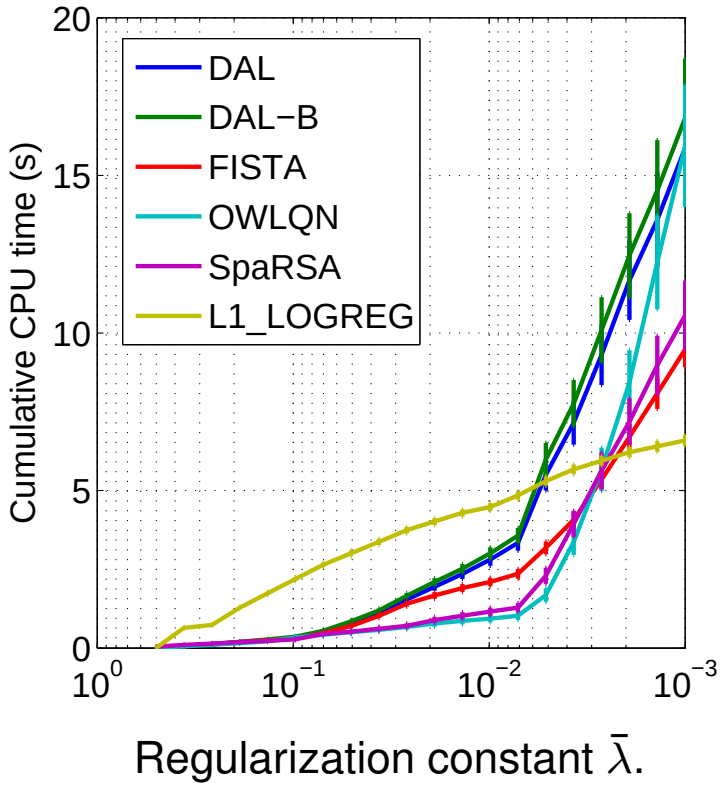
$m = 800$, $n = 100{,}000$.

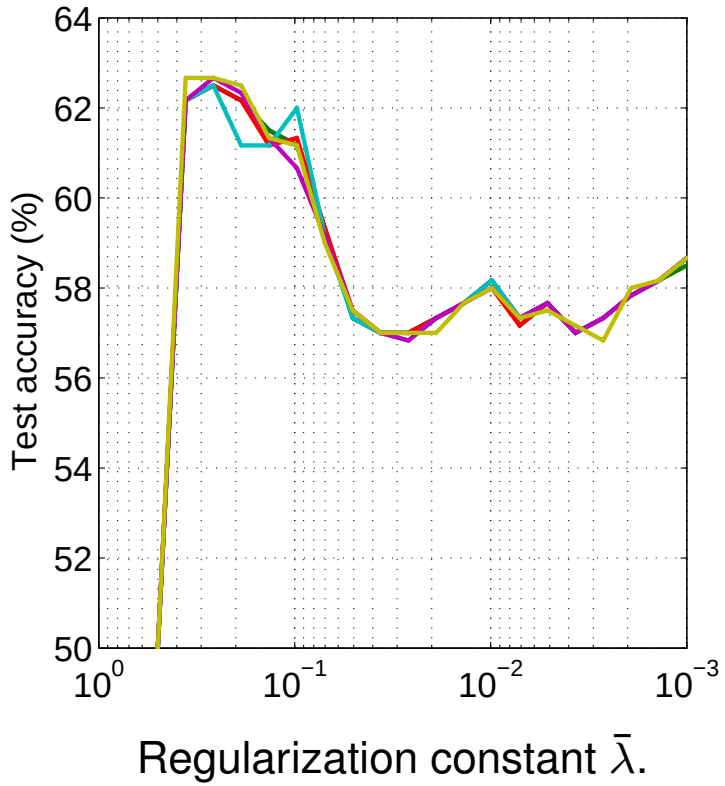




Benchmark datasets (madelon)

$m = 2{,}000$, $n = 500$.





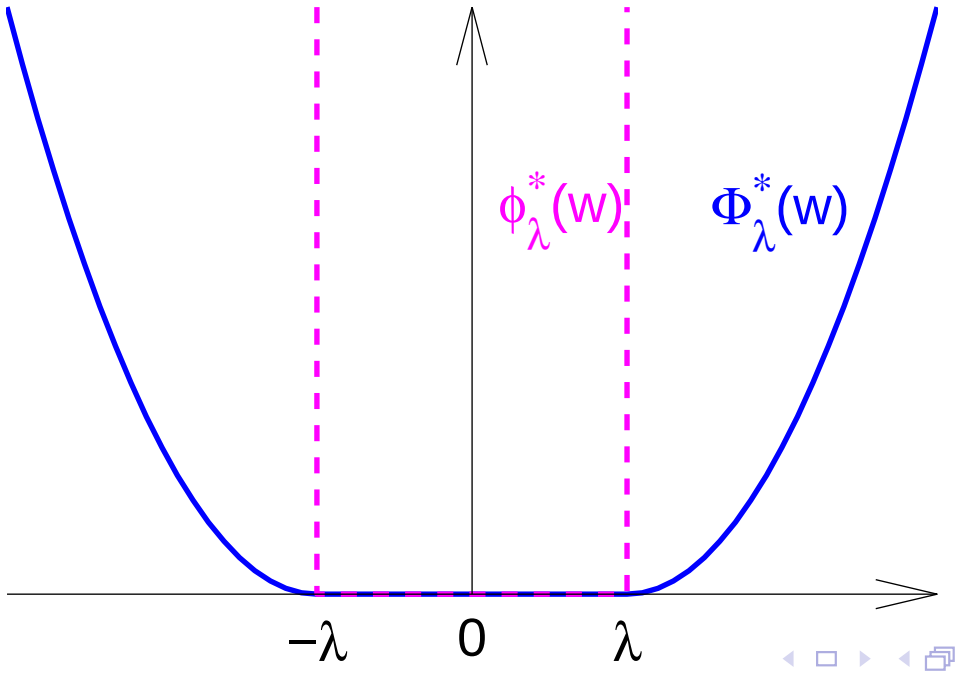
(1) Proximation wrt ϕ_λ is analytic (though non-smooth):

$$\boldsymbol{w}^{t+1} = \mathrm{ST}_{\eta_t \lambda}\left(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha}^t\right)$$

(2) Inner minimization is smooth:

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\mathrm{argmin}} \Big(\underbrace{f_\ell^*(-\boldsymbol{\alpha})}_{\text{independent of } \boldsymbol{A}.} + \frac{1}{2\eta_t} \underbrace{\|\mathrm{ST}_{\eta_t \lambda}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha})\|_2^2}_{= \Phi_\lambda^*(\cdot)} \Big)$$

(linear to the number of active variables)



Comparison to Rockafellar 76

Assumption

The multifunction ∇f^* is (locally) Lipschitz continuous at the origin:

$$\|\nabla f^*(\beta) - \nabla f^*(0)\| \leq L\|\beta\| \quad (\|\beta\| \leq \tau)$$

$\Rightarrow$ Implies our assumption with $\sigma = \frac{1}{2}\min(1/L, \tau/\|\boldsymbol{w}^0 - \boldsymbol{w}^*\|)$.

Convergence (exact minimization) – comparable to Thm 1

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{1}{\sqrt{1 + (\eta_t/L)^2}}\|\boldsymbol{w}^t - \boldsymbol{w}^*\|$$

Convergence (approximate minimization) – much worse than Thm 2

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \frac{\mu_t + \epsilon_t}{1 - \epsilon_t}\|\boldsymbol{w}^t - \boldsymbol{w}^*\| \quad \left(\mu_t = \frac{1}{\sqrt{1+(\eta_t/L)^2}}\right)$$

(assuming $\|\nabla\varphi_t\| \leq \epsilon_t\sqrt{\gamma/\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|$)

Outline of proof of Theorem 1

1. Since $\boldsymbol{w}^{t+1} = \mathrm{argmin}_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$,
   $(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t$ is a subgradient of f at $\boldsymbol{w}^{t+1}$. I.e.,
$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle.$$
2. For any $\mu > 0$,
$$\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\| \|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq \frac{\mu}{2} \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \frac{1}{2\mu} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$
3. Combining 1 & 2 and using $f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2$,
$$\frac{1}{2} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2 \geq ((1 + \sigma\eta_t)\mu - \frac{\mu^2}{2}) \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2.$$
4. Maximize RHS wrt μ.

Outline of proof of Theorem 1

1. Since $\boldsymbol{w}^{t+1} = \mathrm{argmin}_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$,
   $(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t$ is a subgradient of f at $\boldsymbol{w}^{t+1}$. I.e.,
   $$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle.$$
2. For any $\mu > 0$,
   $$\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\| \|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq \frac{\mu}{2} \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \frac{1}{2\mu} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$
3. Combining 1 & 2 and using $f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2$,
   $$\frac{1}{2} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2 \geq ((1 + \sigma\eta_t)\mu - \frac{\mu^2}{2}) \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2.$$
4. Maximize RHS wrt μ.

Outline of proof of Theorem 1

1. Since $\boldsymbol{w}^{t+1} = \mathrm{argmin}_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \right)$,
   $(\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t$ is a subgradient of f at $\boldsymbol{w}^{t+1}$. I.e.,
   $$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle.$$
2. For any $\mu > 0$,
   $$\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\| \|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq \frac{\mu}{2} \|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2 + \frac{1}{2\mu} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2.$$
3. Combining 1 & 2 and using $f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2$,
   $$\frac{1}{2} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2 \geq ((1 + \sigma\eta_t)\mu - \frac{\mu^2}{2}) \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2.$$
4. Maximize RHS wrt μ.

Outline of proof of Theorem 2

1. Let $\delta^t := \nabla\varphi_t(\alpha^t)$,

$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle - \frac{1}{2\gamma}\|\delta^t\|^2.$$

2. By assumption

$$f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*) \geq \sigma\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2,$$

$$\|\delta^t\|^2 \leq \frac{\gamma}{\eta_t}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2.$$

3. Combining 1 & 2,

$$\frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}^t\|^2 \geq (\sigma\eta_t + \frac{1}{2})\|\boldsymbol{w}^* - \boldsymbol{w}^{t+1}\|^2.$$