機械学習における連続最適化の新しいトレンド

冨岡 亮太 [1]
共同研究者: 鹿島久嗣 [1]、杉山将 [2]、鈴木大慈 [1]、林浩平 [3]

[1] 東京大学 [2] 東京工業大学 [3] 奈良先端科学技術大学院大学

2011-10-25 @ RAMP 2011

最適化業界-機械学習業界の間の需給ミスマッチ

- 最適化業界
  - 最適化のことよく分からなくても使えるツールボックスが必要
  - ワンストップサービス — CVX (Grant & Boyd)
  - 連続最適化なら内点法 (80 年代〜)
- 機械学習業界
  - モデルが変わってもすぐ実装を変更できる方がよい．
  - なるべく簡単な手法が好ましい．
  - 並列化できるとなおよい．

⇒ 古い手法（60-70 年代）がどうやら熱い．

最適化業界-機械学習業界の間の需給ミスマッチ

- 最適化業界
  - 最適化のことよく分からなくても使えるツールボックスが必要
  - ワンストップサービス — CVX (Grant & Boyd)
  - 連続最適化なら内点法 (80 年代〜)
- 機械学習業界
  - モデルが変わってもすぐ実装を変更できる方がよい．
  - なるべく簡単な手法が好ましい．
  - 並列化できるとなおよい．

⇒ 古い手法（60-70 年代）がどうやら熱い．

  - (Accelerated) Proximal gradient methods

最適化業界-機械学習業界の間の需給ミスマッチ

- 最適化業界
  - 最適化のことよく分からなくても使えるツールボックスが必要
  - ワンストップサービス — CVX (Grant & Boyd)
  - 連続最適化なら内点法 (80 年代〜)
- 機械学習業界
  - モデルが変わってもすぐ実装を変更できる方がよい.
  - なるべく簡単な手法が好ましい.
  - 並列化できるとなおよい.

⇒ 古い手法（60-70 年代）がどうやら熱い.
  - (Accelerated) Proximal gradient methods
  - Dual decomposition (Uzawa's method)

最適化業界-機械学習業界の間の需給ミスマッチ

- 最適化業界
  - 最適化のことよく分からなくても使えるツールボックスが必要
  - ワンストップサービス — CVX (Grant & Boyd)
  - 連続最適化なら内点法 (80 年代〜)
- 機械学習業界
  - モデルが変わってもすぐ実装を変更できる方がよい．
  - なるべく簡単な手法が好ましい．
  - 並列化できるとなおよい．

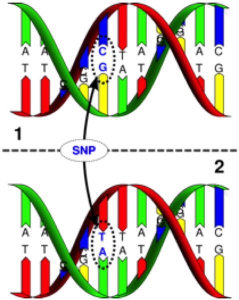
⇒ 古い手法（60-70 年代）がどうやら熱い．

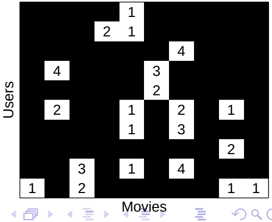
- (Accelerated) Proximal gradient methods
- Dual decomposition (Uzawa's method)
- Alternating Direction Method of Multipliers (ADMM)

機械学習における連続最適化の古いトレンド？

なぜこれらの手法がいま注目されるか — スパース推定

- 高次元データ (サンプル数 $\ll$ 次元)
  - バイオインフォマティクス（遺伝子発現，SNP 解析，etc）
  - テキストマイニング（系列ラベリング，係り受け解析）
  - イメージング（MRI） — 圧縮センシング
- 構造があるデータ
  - 協調フィルタリング— 低ランク構造
  - グラフィカルモデル推定 — グラフ構造





例 1: SNP（一塩基多型）解析

$\boldsymbol{x}_i$: **入力** (SNP)， $y_i = 1$: **病気**，$y_i = -1$: **健康**

目的: ゲノムの個人差 $\boldsymbol{x}_i$ と病気になるかならないか y_i の関係を知りたい．

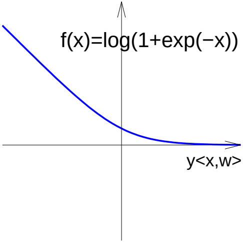
ロジスティック回帰：2 **値分類規則の学習法** $(y_i \in \{-1, +1\})$

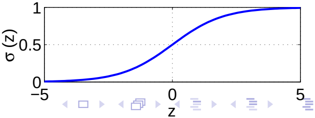
$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad \underbrace{\sum_{i=1}^{m} \log(1 + \exp(-y_i \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle))}_{\text{data-fit}} \quad + \quad \underbrace{\lambda \|\boldsymbol{w}\|_1}_{\text{Regularization}}$$

- **例えば，**SNP **の数** $n = 500,000$，**被験者の数** $m = 5,000$
- **事後確率最大化** (MAP) **法の一種．ロジスティック損失関数**:

$$\log(1 + e^{-yz}) = -\log P(Y = y|z)$$

where $P(Y = +1|z) = \frac{e^z}{1+e^z}$.





例 2: 圧縮センシング [Candes, Romberg, & Tao 06]

低次元（ノイズ入り）観測からの信号（MRI 画像）復元

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\Omega}\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{\Phi}\boldsymbol{w}\|_1$$

- $\boldsymbol{y}$: ノイズ入り観測信号
- $\boldsymbol{w}$: 原信号
- Ω: $\mathbb{R}^n \to \mathbb{R}^m$: 観測行列（ランダム，フーリエ変換）
- $\boldsymbol{\Phi}$: 原信号がスパースとなる基底への変換行列

※ $\boldsymbol{\Phi}^{-1}$ が存在すれば，より簡単な問題

$$\underset{\tilde{\boldsymbol{w}}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{w}}\|_2^2+\lambda\|\tilde{\boldsymbol{w}}\|_1,$$

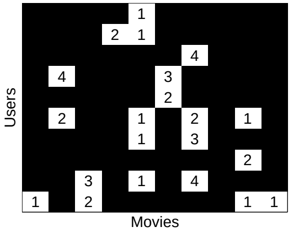
(ただし $\boldsymbol{A}=\Omega\boldsymbol{\Phi}^{-1}$) を解けばよい．

例 3: 低ランク行列の推定 [Fazel+ 01; Srebro+ 05]

行列 X を部分的な（ノイズ入り）観測 Y から復元したい：

$$\underset{\boldsymbol{X}}{\text{minimize}} \quad \frac{1}{2}\|\Omega(\boldsymbol{X}-\boldsymbol{Y})\|^2 + \lambda\|\boldsymbol{X}\|_{S_1}$$

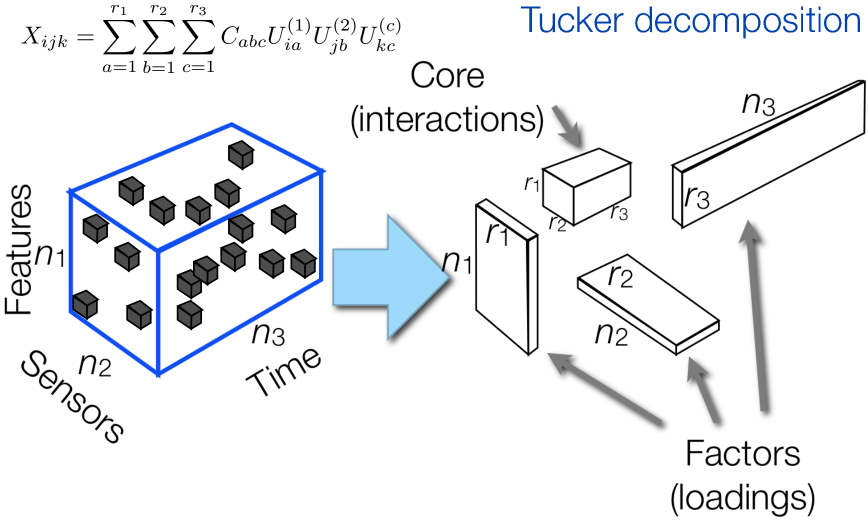
$$\text{where} \quad \|\boldsymbol{X}\|_{S_1} := \sum_{j=1}^{r}\sigma_j(\boldsymbol{X}) \quad \text{(Schatten 1-norm)}$$



特異値の線形和
⇒ **特異値の意味でスパース**
⇒ **低ランク**

例 4: 低ランクテンソルの補完 [Tucker 66]

$$X_{ijk} = \sum_{a=1}^{r_1}\sum_{b=1}^{r_2}\sum_{c=1}^{r_3} C_{abc} U_{ia}^{(1)} U_{jb}^{(2)} U_{kc}^{(c)}$$



単純スパース推定問題と構造付きスパース推定問題

- **単純スパース推定問題**

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

  - SNP 解析
  - 圧縮センシングで $\boldsymbol{\Phi}^{-1}$ が存在する場合（ウェーブレット）
  - 協調フィルタリング（行列穴埋め）

- **構造付きスパース推定問題**

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

  - 圧縮センシングで $\boldsymbol{\Phi}^{-1}$ が存在しない場合（Total variation)
  - テンソルの Tucker 分解

今日の内容

- **単純スパース推定問題のための最適化手法**
  - **（加速付き）近接勾配法** (proximal gradient method)
  - Dual Augmented Lagrangian (DAL)
- **構造付きスパース推定問題のための最適化手法**
  - Alternating Direction Method of Multipliers (ADMM)

単純スパース推定問題のための最適化手法

- **（加速付き）近接勾配法** (proximal gradient method)
- Dual Augmented Lagrangian (DAL)

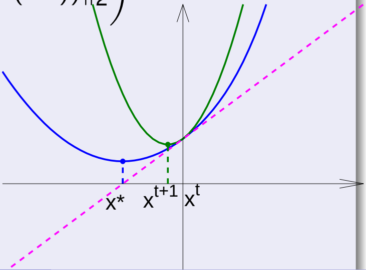
近接勾配法 (proximal gradient method)

最小化問題

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \underbrace{L(\boldsymbol{w})}_{\text{微分可能}} + \underbrace{\lambda\|\boldsymbol{w}\|_1}_{\text{微分不可能}}$$

線形化/最小化

$$\begin{aligned}
\boldsymbol{w}^{t+1} &= \underset{\boldsymbol{w}}{\text{argmin}} \left(\nabla L(\boldsymbol{w}^t)(\boldsymbol{w} - \boldsymbol{w}^t) + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \right) \\
&= \underset{\boldsymbol{w}}{\text{argmin}} \left(\lambda\|\boldsymbol{w}\|_1 + \frac{1}{2\eta_t}\|\boldsymbol{w} - (\boldsymbol{w}^t - \eta_t \nabla L(\boldsymbol{w}^t))\|_2^2 \right) \\
&= \text{prox}_{\lambda\eta_t}(\boldsymbol{w}^t - \eta_t \nabla L(\boldsymbol{w}^t)).
\end{aligned}$$

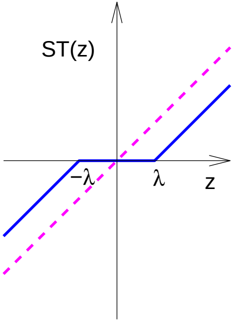


Proximal operator: 射影の一般化

$$\mathrm{prox}_{g}(\boldsymbol{z}) = \operatorname*{argmin}_{\boldsymbol{x}} \left(g(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 \right)$$

- **凸集合への射影**: $\mathrm{prox}_{\delta_C}(\boldsymbol{z}) = \mathrm{proj}_C(\boldsymbol{z})$.
- Soft-Threshold ($g(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$)

$$\begin{aligned}\mathrm{prox}_{\lambda}(\boldsymbol{z}) &= \operatorname*{argmin}_{\boldsymbol{x}} \left(\lambda\|\boldsymbol{x}\|_1 + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 \right) \\ &= \begin{cases} z_j + \lambda & (z_j < -\lambda), \\ 0 & (-\lambda \le z_j \le \lambda), \\ z_j - \lambda & (z_j > \lambda). \end{cases}\end{aligned}$$



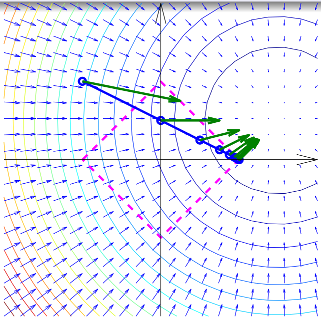
- **何らかの意味で分離可能な関数 r は Prox が簡単に計算できる.**
- **微分不可能でも解析的に計算できる.**

近接勾配法 (proximal gradient method)

近接勾配法 (Lions & Mercier 79; Figueiredo&Nowak 03; Daubechies 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：**

$$w^{t+1} \leftarrow \underbrace{\mathrm{prox}_{\eta_t\lambda}}_{\text{縮小}}\Bigl(\underbrace{w^t - \eta_t \nabla L(w^t)}_{\text{勾配ステップ}}\Bigr).$$



- **利点**: **実装が簡単.**
- **欠点**: **損失項 L のヘシアンの条件数が悪いと遅い.**
- **別名**: Forward-Backward splitting, Iterative Shrinkage/Thresholding

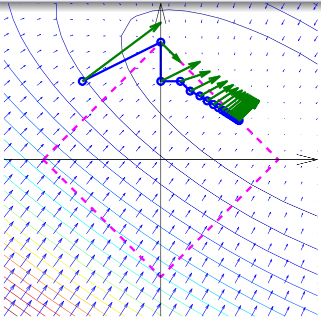
近接勾配法 (proximal gradient method)

近接勾配法 (Lions & Mercier 79; Figueiredo&Nowak 03; Daubechies 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：**

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

- **利点**: **実装が簡単.**
- **欠点**: **損失項 L のヘシアンの条件数が悪いと遅い.**
- **別名**: Forward-Backward splitting, Iterative Shrinkage/Thresholding



近接勾配法の収束レートと加速

- 損失項 L が強凸，かつリプシッツ定数

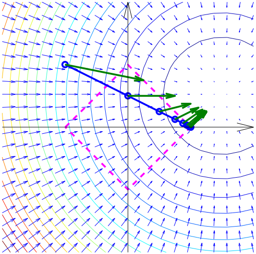
$$\|\nabla L(\boldsymbol{x}) - \nabla L(\boldsymbol{y})\| \leq H\|\boldsymbol{x} - \boldsymbol{y}\|$$

が存在すれば 1 次収束（勾配法と同じ）

- 強凸でない場合，ステップサイズ $\eta_t \leq 1/H$ とすることで，多項式レート

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq \frac{H\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2k}$$

- 下限 $O(1/k^2)$ を達成するための加速法も提案されている．(Nesterov 07; Beck & Teboulle 09)



Dual Augmented Lagrangian (DAL) [Tomioka & Sugiyama 09]

- 1 次ブラックボックスモデルでは下限が達成されている．
- もう少し**機械学習における**問題の構造を考慮したい．

1. 損失項は $L(\boldsymbol{w}) = f_\ell(\boldsymbol{A}\boldsymbol{w})$ と分解できる．f_ℓ: ロス関数，$\boldsymbol{A} \in \mathbb{R}^{m \times n}$: データ行列

   (例 1) $L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{w}\|_2^2$ (2 乗ロス回帰)

   (例 2) $L(\boldsymbol{w}) = \sum_{i=1}^{m} \log(1 + \exp(-y_i \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle))$ (ロジスティック回帰)

2. スパースな解に興味があるので，解がスパースであるほど効率的な解法が望ましい．(データ行列は必ずしもスパースではない)

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

主問題

$$\min_{\boldsymbol{w}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1}_{f(\boldsymbol{w})}$$

双対問題

$$\max_{\boldsymbol{\alpha},\boldsymbol{v}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - (\lambda\|\cdot\|_1)^*(\boldsymbol{v})$$

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

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

主問題

$$\min_{\boldsymbol{w}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1}_{f(\boldsymbol{w})}$$

Proximal minimization [Rockafellar 76]:

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

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

- 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

$$\max_{\boldsymbol{\alpha},\boldsymbol{v}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - (\lambda\|\cdot\|_1)^*(\boldsymbol{v})$$

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

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

主問題

$$\min_{\boldsymbol{w}} \quad \underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1}_{f(\boldsymbol{w})}$$

Proximal minimization [Rockafellar 76]:

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

$$(\eta_0 \le \eta_1 \le \cdots)$$

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

双対問題

$$\max_{\boldsymbol{\alpha}, \boldsymbol{v}} \quad -f_\ell^*(-\boldsymbol{\alpha}) - (\lambda\|\cdot\|_1)^*(\boldsymbol{v})$$

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

⇔Augmented Lagrangian [Powell 69; Hestenes 69]:

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

$$\boldsymbol{\alpha}^t = \operatorname*{argmin}_{\boldsymbol{\alpha}} \varphi_t(\boldsymbol{\alpha})$$

- $\varphi_t(\boldsymbol{\alpha})$ 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 $\boldsymbol{w}^0$ を決める．
2. 停止条件が満たされるまで反復：

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

ただし，

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

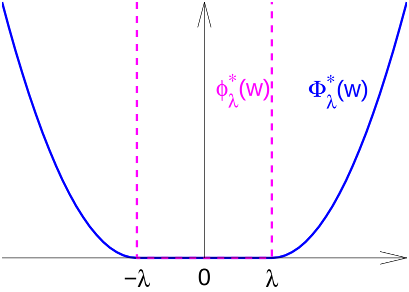
DAL の利点 (ℓ_1-正則化の場合)

(1) Prox **作用素は解析的に計算可能**

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

(2) **内部最適化は微分可能**

$$\boldsymbol{\alpha}^t = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}}\Bigg(\underbrace{f_\ell^*(-\boldsymbol{\alpha})}_{\text{微分可能. } \boldsymbol{A} \text{ のスケーリングの影響を受けない}} + \frac{1}{2\eta_t} \underbrace{\|\mathrm{prox}_{\lambda\eta_t}(\boldsymbol{w}^t + \eta_t \boldsymbol{A}^\top \boldsymbol{\alpha})\|^2}_{\text{非ゼロ成分の数に比例}} \Bigg)$$



近接勾配法と DAL の違い：いかに変数の間の絡みを除くか

目的関数 f に関する Proximation は難しい：

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\mathrm{argmin}} \Bigl(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{変数が絡みあっている}} + \lambda \|\boldsymbol{w}\|_1}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Bigr)$$

近接勾配法と DAL の違い：いかに変数の間の絡みを除くか

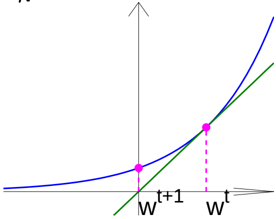
目的関数 f に関する Proximation は難しい：

$$\boldsymbol{w}^{t+1} = \mathop{\mathrm{argmin}}_{\boldsymbol{w}} \Big(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{変数が絡みあっている}} + \lambda \|\boldsymbol{w}\|_1}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Big)$$

- **近接勾配法（既存）**: **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**



近接勾配法と DAL の違い：いかに変数の間の絡みを除くか

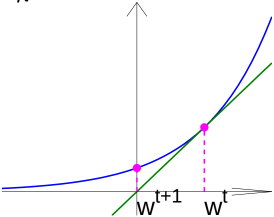
目的関数 f に関する Proximation は難しい：

$$\boldsymbol{w}^{t+1} = \underset{\boldsymbol{w}}{\mathrm{argmin}} \Bigg(\overbrace{\underbrace{f_\ell(\boldsymbol{A}\boldsymbol{w})}_{\text{変数が絡みあっている}} + \lambda \|\boldsymbol{w}\|_1}^{f(\boldsymbol{w})} + \frac{1}{2\eta_t} \|\boldsymbol{w} - \boldsymbol{w}^t\|^2 \Bigg)$$

- **近接勾配法（既存）：線形にロス項を 近似：**

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

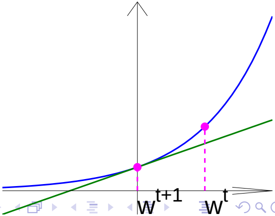
$\rightarrow$ **現在の点 $\boldsymbol{w}^t$ で最もタイト**



- **DAL（提案法）：線形なロス項の 下限**

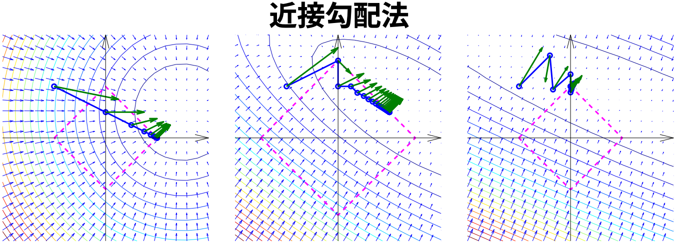
$$f_\ell(\boldsymbol{A}\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \left(-f_\ell^*(-\boldsymbol{\alpha}) - \boldsymbol{w}^\top \boldsymbol{A}^\top \boldsymbol{\alpha} \right)$$

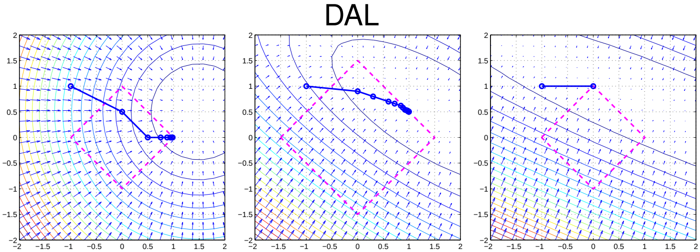
$\rightarrow$ **次の点 $\boldsymbol{w}^{t+1}$ で最もタイト**



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．





定理 1（厳密な最小化）

定義

- $\boldsymbol{w}^t$：厳密な DAL 法（$\|\nabla\varphi_t(\alpha^t)\| = 0$）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

- 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.

$$\|\nabla\varphi_t(\alpha^t)\| \leq \sqrt{\frac{\gamma}{\eta_t}}\|w^{t+1} - w^t\|$$

（$1/\gamma$: 損失関数の微分 ∇f_ℓ のリプシッツ定数.）

定理 2

定理 1 と同じ仮定のもとで

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

$\Rightarrow$ η_t が増加するなら，w^t は 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．

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

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 $\frac{\|\nabla\varphi_t(\boldsymbol{\alpha}^t)\|}{\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|} \leq O(1/\eta_t)$.

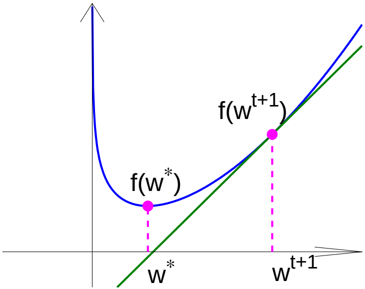
定理 1 の証明（エッセンス）

$\boldsymbol{w}^{t+1}$ 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)，

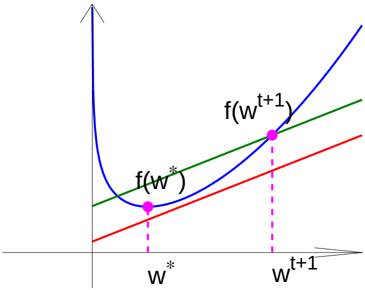
$$f(\boldsymbol{w}^*) - f(\boldsymbol{w}^{t+1}) \geq \left\langle (\boldsymbol{w}^t - \boldsymbol{w}^{t+1})/\eta_t, \boldsymbol{w}^* - \boldsymbol{w}^{t+1} \right\rangle.$$



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．**



構造付きスパース推定問題のための最適化手法

- Alternating Direction Method of Multipliers (ADMM)

拡張ラグランジュ法 [Powell 69; Hestenes 69]

最小化問題

$$\underset{\boldsymbol{x},\boldsymbol{z}}{\text{minimize}} \quad f(\boldsymbol{x}) + \lambda\|\boldsymbol{z}\|_1,$$
$$\text{s.t.} \quad \boldsymbol{z} = \boldsymbol{\Phi x}$$

拡張ラグランジアン

$$L_\eta(\boldsymbol{x}, \boldsymbol{z}, \alpha) = f(\boldsymbol{x}) + \lambda\|\boldsymbol{z}\|_1 + \alpha^\top(\boldsymbol{z} - \boldsymbol{\Phi x}) + \frac{\eta}{2}\|\boldsymbol{z} - \boldsymbol{\Phi x}\|^2.$$

拡張ラグランジュ法

拡張ラグランジアンを $\boldsymbol{x}, \boldsymbol{z}$ に関して最小化:

$$(\boldsymbol{x}^{t+1}, \boldsymbol{z}^{t+1}) = \underset{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^m}{\text{argmin}} \ L_{\eta_t}(\boldsymbol{x}, \boldsymbol{z}, \alpha^t).$$

ラグランジュ乗数を更新:

$$\alpha^{t+1} = \alpha^t + \eta_t(\boldsymbol{z}^{t+1} - \boldsymbol{\Phi x}^{t+1}).$$

$\boldsymbol{x}$ と $\boldsymbol{z}$ の間に絡みが発生!（別々に最小化できない）

Alternating Direction Method of Multipliers (ADMM; Gabay & Mercier 76)

拡張ラグランジアン

$$L_\eta(\boldsymbol{x}, \boldsymbol{z}, \alpha) = f(\boldsymbol{x}) + \lambda\|\boldsymbol{z}\|_1 + \alpha^\top(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{x}\|^2.$$

拡張ラグランジアンを $\boldsymbol{x}$ に関して最小化:

$$\boldsymbol{x}^{t+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} L_{\eta_t}(\boldsymbol{x}, \boldsymbol{z}^t, \alpha^t).$$

拡張ラグランジアンを $\boldsymbol{z}$ に関して最小化:

$$\boldsymbol{z}^{t+1} = \operatorname*{argmin}_{\boldsymbol{z}\in\mathbb{R}^m} L_{\eta_t}(\boldsymbol{x}^{t+1}, \boldsymbol{z}, \alpha^t).$$

ラグランジュ乗数を更新:

$$\alpha^{t+1} = \alpha^t + \eta_t(\boldsymbol{z}^{t+1} - \boldsymbol{\Phi}\boldsymbol{x}^{t+1}).$$

- **今更新した $\boldsymbol{x}^{t+1}$ が $\boldsymbol{z}^{t+1}$ の計算に入っているところがポイント.**

ADMM (Gabay & Mercier 76)

拡張ラグランジアン

$$L_\eta(\boldsymbol{x},\boldsymbol{z},\alpha)=f(\boldsymbol{x})+\lambda\|\boldsymbol{z}\|_1+\alpha^\top(\boldsymbol{z}-\boldsymbol{\Phi}\boldsymbol{x})+\frac{\eta}{2}\|\boldsymbol{z}-\boldsymbol{\Phi}\boldsymbol{x}\|^2.$$

書き直すと

$$\begin{cases}\boldsymbol{x}^{t+1}=\underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\,L_{\eta_t}(\boldsymbol{x},\boldsymbol{z}^t,\alpha^t).\\ \boldsymbol{z}^{t+1}=\underset{\boldsymbol{z}\in\mathbb{R}^m}{\operatorname{argmin}}\,L_{\eta_t}(\boldsymbol{x}^{t+1},\boldsymbol{z},\alpha^t).\\ \alpha^{t+1}=\alpha^t+\eta_t(\boldsymbol{z}^{t+1}-\boldsymbol{\Phi}\boldsymbol{x}^{t+1}).\end{cases}$$

ADMM (Gabay & Mercier 76)

拡張ラグランジアン

$$L_\eta(\boldsymbol{x}, \boldsymbol{z}, \alpha) = f(\boldsymbol{x}) + \lambda\|\boldsymbol{z}\|_1 + \alpha^\top(\boldsymbol{z} - \boldsymbol{\Phi x}) + \frac{\eta}{2}\|\boldsymbol{z} - \boldsymbol{\Phi x}\|^2.$$

書き直すと

$$\begin{cases} \boldsymbol{x}^{t+1} = \underset{\boldsymbol{x} \in \mathbb{R}^n}{\operatorname{argmin}} \left(f(\boldsymbol{x}) + \frac{\eta_t}{2}\|\boldsymbol{z}^t - \boldsymbol{\Phi x} + \alpha^t/\eta_t\|^2 \right). \\ \boldsymbol{z}^{t+1} = \underset{\boldsymbol{z} \in \mathbb{R}^m}{\operatorname{argmin}} L_{\eta_t}(\boldsymbol{x}^{t+1}, \boldsymbol{z}, \alpha^t). \\ \alpha^{t+1} = \alpha^t + \eta_t(\boldsymbol{z}^{t+1} - \boldsymbol{\Phi x}^{t+1}). \end{cases}$$

ADMM (Gabay & Mercier 76)

拡張ラグランジアン

$$L_\eta(\boldsymbol{x}, \boldsymbol{z}, \alpha) = f(\boldsymbol{x}) + \lambda \|\boldsymbol{z}\|_1 + \alpha^\top(\boldsymbol{z} - \boldsymbol{\Phi x}) + \frac{\eta}{2}\|\boldsymbol{z} - \boldsymbol{\Phi x}\|^2.$$

書き直すと

$$\begin{cases} \boldsymbol{x}^{t+1} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} \left(f(\boldsymbol{x}) + \frac{\eta_t}{2}\|\boldsymbol{z}^t - \boldsymbol{\Phi x} + \alpha^t/\eta_t\|^2 \right). \\ \boldsymbol{z}^{t+1} = \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^m} \left(\lambda\|\boldsymbol{z}\|_1 + \frac{\eta_t}{2}\|\boldsymbol{z} - \boldsymbol{\Phi x}^{t+1} + \alpha^t/\eta_t\|^2 \right). \\ \alpha^{t+1} = \alpha^t + \eta_t(\boldsymbol{z}^{t+1} - \boldsymbol{\Phi x}^{t+1}). \end{cases}$$

ADMM (Gabay & Mercier 76)

拡張ラグランジアン

$$L_\eta(\boldsymbol{x},\boldsymbol{z},\alpha) = f(\boldsymbol{x}) + \lambda\|\boldsymbol{z}\|_1 + \alpha^\top(\boldsymbol{z}-\boldsymbol{\Phi x}) + \frac{\eta}{2}\|\boldsymbol{z}-\boldsymbol{\Phi x}\|^2.$$

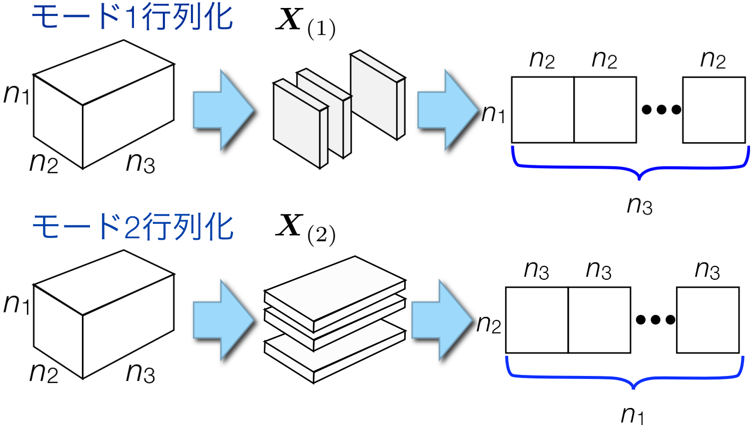
書き直すと

$$\begin{cases} \boldsymbol{x}^{t+1} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\left(f(\boldsymbol{x}) + \frac{\eta_t}{2}\|\boldsymbol{z}^t - \boldsymbol{\Phi x} + \alpha^t/\eta_t\|^2\right). \\ \boldsymbol{z}^{t+1} = \underset{\boldsymbol{z}\in\mathbb{R}^m}{\operatorname{argmin}}\left(\lambda\|\boldsymbol{z}\|_1 + \frac{\eta_t}{2}\|\boldsymbol{z} - \boldsymbol{\Phi x}^{t+1} + \alpha^t/\eta_t\|^2\right). \\ \alpha^{t+1} = \alpha^t + \eta_t(\boldsymbol{z}^{t+1} - \boldsymbol{\Phi x}^{t+1}). \end{cases}$$

- $\boldsymbol{z}$ **に関する最小化は** Prox **作用素** $\text{prox}_{\lambda/\eta_t}$ **(簡単).**
- $\boldsymbol{x}$ **に関する最小化は行列** $\boldsymbol{\Phi}$ **が変数を絡ませるのでちょっと難しい.**
- 1 **反復あたりのコストが同じなら近接勾配法より経験的に速い(理論的には不明)**
- **双対側での** Douglas Rachford Splitting **と等価** $\Rightarrow$ **ステップサイズ** η **によらず** ADMM **は安定.** (Lions & Mercier 76; Eckstein & Bertsekas 92)

テンソルの穴埋め問題への凸最適化の適用 [Liu+09, Signoretto +10, Tomioka+10, Gandy+11]

- **凸最適化の適用のポイント: テンソルの行列化** (Matricization)
- **テンソルが** Tucker **分解の意味で低ランク**
  ⇔ **そのテンソルの行列化は（行列の意味で）低ランク**



テンソルの穴埋め問題への ADMM の適用

数学的な定式化:

$$\begin{aligned} \underset{\boldsymbol{x},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_K\in\mathbb{R}^N}{\text{minimize}} \quad & \frac{1}{2\lambda}\|\boldsymbol{\Omega x}-\boldsymbol{y}\|^2+\sum_{k=1}^{K}\gamma_k \underbrace{\|\boldsymbol{Z}_k\|_{S_1}}_{\text{低ランク化}}, \\ \text{s.t.} \quad & \boldsymbol{P}_k\boldsymbol{x}=\boldsymbol{z}_k \quad (k=1,\ldots,K), \end{aligned}$$

- $\boldsymbol{x}$ は推定すべきテンソルをベクトルとして書いたもの．
- $\boldsymbol{y}\in\mathbb{R}^M$ は観測．($M \ll N = n_1 n_2 \cdots n_K$)
- $\boldsymbol{P}_k$ はモード k 行列化の操作を行列で表現したもの．
- $\boldsymbol{P}_k{}^\top \boldsymbol{P}_k=\boldsymbol{I}$（行列化は直交変換）．
- すべてのモードが同時に低ランクになるように正則化．

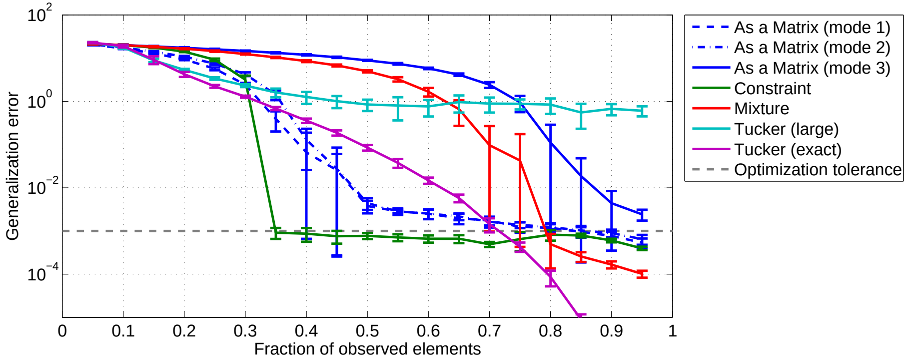
テンソルの穴埋め問題への ADMM の適用

拡張ラグランジアン

$$L_\eta(\boldsymbol{x}, \{\boldsymbol{Z}_k\}_{k=1}^K, \{\alpha_k\}_{k=1}^K) = \frac{1}{2\lambda}\|\boldsymbol{\Omega x} - \boldsymbol{y}\|^2 + \sum_{k=1}^K \gamma_k \|\boldsymbol{Z}_k\|_{S_1}$$
$$+ \sum_{k=1}^K \left(\alpha_k{}^\top(\boldsymbol{P}_k\boldsymbol{x} - \boldsymbol{z}_k) + \frac{\eta}{2}\|\boldsymbol{P}_k\boldsymbol{x} - \boldsymbol{z}_k\|^2\right).$$

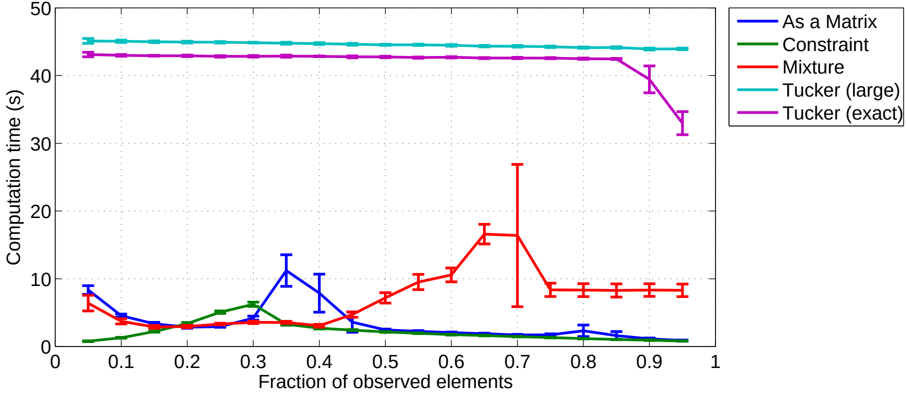
- **$\boldsymbol{x}$ に関する最小化 $\boldsymbol{P}_k$ が直交行列なので解析的に $O(N)$ で計算可能．**
- **$\boldsymbol{Z}_k$（$\boldsymbol{z}_k$ を行列として並べたもの）に関する最小化は Schatten 1-ノルムに関する Prox 作用素．**
- **ラグランジュ乗数ベクトルは制約の数（モードの数）だけ必要．**

テンソル結果 1: 予測精度



- **提案手法 Constraint は 35% くらい見えればほぼ完璧に予測可能．ランクを前もって決める必要なし．**
- **既存手法 Tucker (EM アルゴリズム) はランクが合っていれば OK．ランクが間違っていると汎化誤差が収束しない．**

テンソル結果 2: 計算速度



- **しかも凸最適化は速い！**

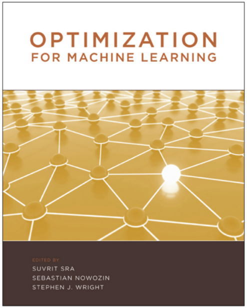
結論

- 最適化（凸最適化）：機械学習研究者に取って欠かせないツール
- ブラックボックス最適化から中身を考慮した最適化へ
  - 単純スパース推定
  - 構造付きスパース推定
- 理論解析の中でも最適化を含めた話が重要
  - Stochastic Optimization in Machine Learning (Nathan Srebro, tutorial at ICML 2010)
- 並列化，オンライン化などがホットな話題
  - LCCC : NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds

ご清聴ありがとうございました！

宣伝

Optimization for Machine Learning (MIT Press, 2011)



謝辞

これらの研究の様々な段階でコメントを頂いた土谷隆先生，小島政和先生，福島雅夫先生に感謝します．この研究は科研費 22700138 **および** NTT **コミュニケーション科学基礎研究所の支援をうけています．**

References

Recent surveys

- Tomioka, Suzuki, & Sugiyama (2011) Augmented Lagrangian Methods for Learning, Selecting, and Combining Features. In Sra, Nowozin, Wright., editors, *Optimization for Machine Learning*, MIT Press.
- Combettes & Pesquet (2010) Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer-Verlag.
- Boyd, Parikh, Peleato, & Eckstein (2010) Distributed optimization and statistical learning via the alternating direction method of multipliers.

IST/FISTA

- Moreau (1965) Proximité et dualité dans un espace Hilbertien. Bul letin de la S. M. F.
- Nesterov (2007) Gradient Methods for Minimizing Composite Objective Function.
- Beck & Teboulle (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM J Imag Sci 2, 183–202.

Augmented Lagrangian

- Rockafellar (1976) Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. of Oper. Res. 1.
- Bertsekas (1982) Constrained Optimization and Lagrange Multiplier Methods. Academic Press.
- Tomioka, Suzuki, & Sugiyama (2011) Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparse Learning. JMLR 12.

References

ADMM

- Gabay & Mercier (1976) A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput Math Appl 2, 17–40.
- Lions & Mercier (1979) Splitting Algorithms for the Sum of Two Nonlinear Operators. SIAM J Numer Anal 16, 964–979.
- Eckstein & Bertsekas (1992) On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators.

Matrices/Tensor

- Fazal, Hindi, & Boyd (2001) A Rank Minimization Heuristic with Application to Minimum Order System Approximation. Proc. of the American Control Conference.
- Srebro, Rennie, & Jaakkola (2005) Maximum-Margin Matrix Factorization. Advances in NIPS 17, 1329–1336.
- Cai, Candès, & Shen (2008) A singular value thresholding algorithm for matrix completion.
- Mazumder, Hastie, & Tibshirani (2010) Spectral Regularization Algorithms for Learning Large Incomplete Matrices. JMLR 11, 2287–2322.
- Tomioka, Hayashi, & Kashima (2011) Estimation of low-rank tensors via convex optimization. arXiv:1010.0789.

Total variation

- Rudin, Osher, Fetemi. (1992) Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60.
- Goldstein & Osher (2009) Split Bregman method for L1 regularization problems. SIAM J. Imag. Sci. 2.