Proceedings of the Thirtieth RAMP Symposium
Hiroshima, October 10–11, 2018

統計力学的アプローチによるリサンプリング手法の軽量化

Statistical mechanical approach to approximated resampling methods

小渕　智之[1*]

Tomoyuki Obuchi

概要　リサンプリングは広く使われている統計的手法であり，モデルや用途を問わず使える汎用性に強みがある．交差検証法・ブートストラップ法・ジャックナイフ法などがその伝統的利用例であるが，近年の機械学習の発展にともない，新たな利用例（例えばバギング・ブースティング等）が多く見出され，日々応用で活躍している．その有用性の一方で，リサンプリング手法は計算量的な負荷が高いことも知られている．最近我々は，いくつかの回帰・分類問題に対してこの計算量的な負荷を適切な解析法・近似法で軽減・回避する研究を行ったので，正則化付き線形回帰における交差検証法を例に，その一端を紹介する．これらの手法は統計力学，特にランダムスピン系の理論において揺籃され，過去 30 年の間に「情報統計力学」という分野を形成しながら手法自体が進化し今日に至っているものである．

キーワード　リサンプリング，交差検証法，線形応答理論，Cavity 法

1.　はじめに

（平衡）統計力学とは，物理学の一理論体系であり，ミクロな世界（cf. 運動方程式で記述される粒子系）とマクロな世界（cf. 熱力学に支配される気体）を結びつけるためのものである．具体的には系のエネルギー関数 $\mathcal{H}$（ハミルトニアン）から誘導される確率分布（ボルツマン分布）を導入し，それに関する平均操作を行うことで実現される．もっとも，身の回りにあるサイズの系を取り扱おうとすると，ボルツマン分布は極めて高次元[*1]になるため (いわゆる次元の呪い)，統計力学の処方箋はそのままは実行できない．しかし我々が知りたいマクロな世界とはそういうサイズの世界の話なので，そこをなんとか実行したいのである．

これを実行するための方法はいくつかあるが，本稿で紹介するのは「近似的に実行する」という，単純な発想のものである．これにより高次元の積分（和）を解析的に低次元に落としたり，計算したい量を他の（より計算しやすい）量に置き換えたりすることができる．上の出自

1　東京工業大学　情報理工学院　数理計算科学系，〒152-8552 東京都目黒区大岡山大岡山２丁目１２－１
Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552, Japan

∗　E-mail address: obuchi@c.titech.ac.jp

∗1　例えば，18 グラムの水がおよそ 1mol≈ 6.02×10^{23} の数の分子から成ることを思い出そう．ボルツマン分布の次元はその指数関数オーダー $O(e^{6.02\times 10^{23}})$ になる．

の話から想像がつくように，統計力学は高次元分布をハンドリングすることをその初期から宿命づけられてきており，結果として近似手法がいろいろと発達しているのである．

高次元確率分布の処理法に対する需要は，物理に限らず，統計や情報理論，信号処理といった様々な文脈で現れる．そのような情報処理に，統計力学的手法を積極的に活用しようという機運が過去 30 年の間に高まり，情報統計力学と呼ばれる分野が形成されることになった．同分野では，80 年代後半から 90 年代前半はニューラルネットワークや学習理論，90 年台後半から 00 年代前半は通信・符号化法や最適化問題に関する研究が精力的に行われていた．近年になって，データ科学・機械学習分野の台頭に伴い，実データから情報を抽出するため理論枠組みの研究へと分野全体がシフトしている*2．本稿で説明する内容も，実データを如何にハンドリングするか？という観点と密接に関わる理論研究である．

本稿では具体的問題として正則化付きの線形回帰を取り上げる．この問題において，正則化項の適切な強さをリサンプリング手法の一種である交差検証法で決めることが多い．リサンプリングは汎用的ではあるが，計算量が増大しがちであるというところに弱点がある．これを統計力学的手法で克服しよう，というのがここで考える話である．

以下では具体的内容を見ていく．2 章では，本稿で取り扱う問題を提示し，それを扱うための統計力学的定式化について見ていく．その枠組のもと，3 章では交差検証法とその近似について述べる．4 章は本稿のまとめにあてられる．

2. 問題設定と定式化

まずは，単純な線形回帰を通じて統計力学的定式化について説明しよう．計画行列を $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M)^\top \in \mathbb{R}^{M\times N}$，目的変数を $\boldsymbol{y} \in \mathbb{R}^M$ と置く．計画行列の各行ベクトル $\boldsymbol{x}_\mu^\top$ を説明変数と呼ぶ．さらに説明変数と目的変数のセットをまとめて $D = \{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^M$ とおきデータあるいはデータセットと呼ぶ．線形回帰の目的は

$$\boldsymbol{y} \approx X\boldsymbol{w}, \tag{1}$$

が良い近似と成るような係数 $\boldsymbol{w} \in \mathbb{R}^N$ を探すことである．通常は次のコスト関数

$$E(\boldsymbol{w}|D) = \frac{1}{2}\sum_{\mu=1}^{M}(y_\mu - \boldsymbol{x}_\mu^\top \boldsymbol{w})^2 = \frac{1}{2}||\boldsymbol{y} - X\boldsymbol{w}||_2^2 \tag{2}$$

を導入しこれを最小化することで, いわゆる最小二乗解 $\hat{\boldsymbol{w}}_{\rm LS} = (X^\top X)^{-1}X^\top \boldsymbol{y}$ を得る ($M > N$ かつ X のランクが十分大きいと仮定した)．ここで ℓ_p ノルム $||\boldsymbol{x}||_p \equiv (\sum_i |x_i|^p)^{1/p}$ を導入した．

さて線形回帰に対応する統計力学的定式化を導入する．与えられたコスト関数をハミルトニ

*2 といっても情報統計力学の研究グループは世界的に見ても手で数えられる程度しかいない．そういう意味では分野というよりコミュニティといったほうが良いかもしれない．

アンと見なし（$\mathcal{H} = E$），ボルツマン分布を次のように定義する：

$$P_\beta(\boldsymbol{w}|D) = \frac{1}{Z_\beta} e^{-\beta\mathcal{H}(\boldsymbol{w}|D)}, \tag{3}$$

ここで $Z_\beta \equiv \int d\boldsymbol{w}\ e^{-\beta\mathcal{H}}$ は規格化定数であるが，統計力学では分配関数と呼ばれる[*3]. また β は逆温度と呼ばれるパラメータで，確率分布の形をコスト関数の値に応じて制御する：$\beta = 0$ ではコスト関数の値に依らずフラットな定数関数になるし，$\beta = \infty$ では最小二乗解 $\hat{\boldsymbol{w}}_{\rm LS}$ のところにデルタ関数的に測度が集中する．そしてこの後者の性質からわかるように，$\boldsymbol{w}$ のボルツマン分布に関する平均を任意の β に対して計算できれば，最小二乗解を $\lim_{\beta\to\infty}\langle \boldsymbol{w}\rangle = \hat{\boldsymbol{w}}_{\rm LS}$ のように求めることができる．ここで $\langle\cdot\rangle$ をボルツマン分布に関する平均と定義した．実際 eq. (3) は指数関数の肩が $\boldsymbol{w}$ に関して二次形式なので，解析的に積分が実行出来て (ガウス積分)，任意の β に対する平均値を求めることができる．

この例からわかるように，ほとんど任意の最適化問題は，それに対応するボルツマン分布を導入することで統計力学的な定式化に移すことができる[*4]．線形回帰の例では eq. (2) の最小化が解析的に出来てしまうので，この定式化のありがたみが感じられないが，コスト関数自体のハンドリングが難しい場合には，ボルツマン分布を構成しておいて近似や確率的手法を導入していくことで，他の方法ではやりづらい計算を実行することができる．

本稿ではそのような例として正則化付きの線形回帰を取り上げよう．次の問題を考える：

$$\mathcal{H}(\boldsymbol{w}|D,\boldsymbol{\eta}) = E(\boldsymbol{w}|D) + J(\boldsymbol{w}|\boldsymbol{\eta}), \tag{4}$$

$$\hat{\boldsymbol{w}}(D,\boldsymbol{\eta}) = \arg\min_{\boldsymbol{w}}\{\mathcal{H}(\boldsymbol{w}|D,\boldsymbol{\eta})\}. \tag{5}$$

ここで E は eq. (2) で定義したように二乗誤差，$J(\boldsymbol{w}|\boldsymbol{\eta})$ は正則化項である．正則化項は，推定解 $\hat{\boldsymbol{w}}$ を何らかの意味で“良い性質”にするために導入される．“良い性質”は問題・目的に応じて変わるだろうが，例えば X によらず解が安定的に得られるようにしたい，というのもよくある要求だろう．そのような場合には ℓ_2 正則化 $J(\boldsymbol{w}|\boldsymbol{\eta} = (\lambda)) = \lambda||\boldsymbol{w}||_2^2$ を導入することで X の条件が悪くても (e.g., $M < N$), 解を求めることができる．色々な正則化を考えることが出来るが，近年の一つのトレンドは，推定解がスパースに成るような正則化を導入することである[*5]．この場合のスパース性とは，推定解 $\hat{\boldsymbol{w}}$ の要素にゼロが多いということである．これ

---

*3　確率論で出てくる積率母関数と本質的には同じ量である．

*4　これは多くの推定問題がベイズによる再定式化が可能なのと似ている．またベイズと統計力学の相性は良く [1]，問題をベイズで定式化して統計力学的に解くということも行われる．

*5　なぜそうしたいのか？ということについては言葉で説明するより，いくつかの美しい実験例を見ると説得力があるかもしれない．例えば [2] では，自然画像をフィルタリングする際，その表現がスパースであることを要求すると，表現基底が自動的にガボールフィルタのようになるということを示した；そこから一次視覚野がガボールフィルタのように成るのは，生物が情報処理にスパース性を利用している（リソースの制約上，表現をスパースにせざるを得ない）からではないかと論じている．似たような結果として，[3] では自然音をスパースに表現しようとすることで，自動的に和音を抽出するフィルターができあがることを示し，これが多くの生き物が和音に強く反応する理由であろうと論じている．

ところで，視覚・聴覚ときたら味覚で同じような結果が出せても良い気がする．なぜ味覚は甘味・酸味・塩味・苦味・旨

を実現する単純な方法の一つは，正則化を $J(\boldsymbol{w}|\boldsymbol{\eta}=(\lambda))=\lambda||\boldsymbol{w}||_p^p$ とし $p\leq 1$ を仮定することである．こうすると，かなり一般的な状況で推定解がスパースになることを証明することができる．特に $p=1$ ではスパース性に加え，コスト関数 $\mathcal{H}$ が凸になるというご利益があるため，解が比較的求めやすく幅広く使われている．

$p\leq 1$ では原点まわりの特異性が強く，最小化のためのアルゴリズムをどう組めばよいかということ自体が非自明となる．このアルゴリズムのデザインにも統計力学的手法は寄与できる [4, 5, 6]．本稿でも最後の方でこの点に少し触れる．ただ本稿で主に取り扱いたい話は，正則化を導入することによって新たに現れるハイパーパラメータ推定の問題 (正則化パラメータ $\boldsymbol{\eta}$ をどう決めるか？）である．以下ではこれを説明していく．

3. 交差検証法とその近似

交差検証法 (cross-validation, CV) とは，推定解に基づく統計的予測の良し悪しを，手元のデータのみから推定する方法である．線形回帰の場合，予測誤差 (または汎化誤差) は次のように定義するのが一般的である:

$$\epsilon_{\rm g}(D,\boldsymbol{\eta})=\int dy_{\rm new}d\boldsymbol{x}_{\rm new}P(\boldsymbol{x}_{\rm new},y_{\rm new})\ \frac{1}{2}||y_{\rm new}-\boldsymbol{x}_{\rm new}^\top\hat{\boldsymbol{w}}(D,\boldsymbol{\eta})||_2^2 \tag{6}$$

ここで $P(\boldsymbol{x}_{\rm new},y_{\rm new})$ は新しいデータ $(\boldsymbol{x}_{\rm new},y_{\rm new})$ の生成プロセスを表す確率分布である．もちろんこの確率分布は殆どの場合未知であるため，予測の良さを測るためには，eq. (6) に対する推定量を構成しなくてはならない．このために，CV では手元にあるデータ D を k 個に均等に分割する．これを $D=\{D_\nu\}_{\nu=1}^k$ と書こう．さらにこの分割の一つ $D_\mu=(\boldsymbol{X}_\mu,\boldsymbol{y}_\mu)$ を取っておき (テストデータ)，残りの $\{D_\nu\}_{\nu(\neq\mu)}\equiv D^{\backslash\mu}$(トレーニングデータ) を使って推定量 $\hat{\boldsymbol{w}}(D^{\backslash\mu},\boldsymbol{\eta})$ を構成する．これをすべての $q=1,\cdots,k$ に対して繰り返し，予測誤差の推定値を

$$\epsilon_{\rm CV}\left(\{D_\nu\}_{\nu=1}^k,\boldsymbol{\eta}\right)=\sum_{\mu=1}^k\frac{1}{2}||\boldsymbol{y}_\mu-\boldsymbol{X}_\mu\hat{\boldsymbol{w}}(D^{\backslash\mu},\boldsymbol{\eta})||_2^2, \tag{7}$$

のように構成する．この手順全体を k 分割交差検証法 (k-fold CV) と呼ぶ．データをすべて使わずに推定し，残しておいたデータに対して予測をして，その誤差を測るのがポイントである．Eq. (7) を以下では CV 誤差と呼ぶ．言うまでもなく，CV 誤差は小さい方が良い．CV をハイパーパラメータ推定に使う場合は，いくつもの異なる値の $\boldsymbol{\eta}$ を試しておいて，その中で最小の CV 誤差を与えるものを選ぶのが普通である．ちなみに k は大きいほうが良いとされる[*6]．これはトレーニングデータが大きいほうが推定精度が良くなるという自然な理由による．$k=M$ が可能な最大値であるが，この場合を特に leave-one-out (LOO) CV と呼ぶ．

CV のボトルネックは，CV 誤差の各項ごとに異なる推定解 $\hat{\boldsymbol{w}}(D^{\backslash\mu},\boldsymbol{\eta})$ を計算しなければな

味の５味なのかとか，旨味は摂取する食品群によっては表現基底に現れにくい不安定な要素であるとか，そういうことが示せたら面白そうである．

*6 もちろんアルゴリズムの安定性とか推定解のロバスト性などを議論しだすと話は変わる．

らないところにある．単純な線形回帰のように推定解がすぐ求まるならまだ良いが，スパース正則化を課すと推定解を求めること自体が大変になる．構造上，並列化が簡単に出来るとはいえ，アルゴリズムの安定性などを考えると，多重に最適化問題を解くこの状況はやはり回避したいところである．

そこで，この推定量 $\hat{\boldsymbol{w}}(D^{\setminus\mu}, \boldsymbol{\eta})$ を，データをフルに使った推定量 $\hat{\boldsymbol{w}}(D, \boldsymbol{\eta})$ から近似的に直接求めようというのが本章で考えるアプローチである．

3.1. 近似公式の導出

以下，簡単のため $k = M$ の LOOCV を考える．M が大きいとき，フルトレーニングデータ D と LOO トレーニングデータ $D^{\setminus\mu}$ の差は，M に対して極めて小さいと考えられる．であれば $\hat{\boldsymbol{w}}(D, \boldsymbol{\eta})$ と $\hat{\boldsymbol{w}}(D^{\setminus\mu}, \boldsymbol{\eta})$ の差 $\boldsymbol{d}_\mu \equiv \hat{\boldsymbol{w}}(D, \boldsymbol{\eta}) - \hat{\boldsymbol{w}}(D^{\setminus\mu}, \boldsymbol{\eta})$ も小さく，摂動的に取り扱えると考えるのが自然であろう．後は $\boldsymbol{d}_\mu$ を決める方程式を摂動で導けば良い．

以降簡単のため，$\hat{\boldsymbol{w}}^{\setminus\mu} \equiv \hat{\boldsymbol{w}}(D^{\setminus\mu}, \boldsymbol{\eta})$, $\mathcal{H}^{\setminus\mu}(\boldsymbol{w}) \equiv \mathcal{H}(\boldsymbol{w}|D^{\setminus\mu}, \boldsymbol{\eta})$ のようにデータへの依存性を引数としては書かず添字で表すこととする．他の変数についても同様である．

3.1.1. 特異性がない場合

スパース性のような特異性が全く無い場合，導出は簡単になる．正則化項 J が微分可能であるとする．コスト関数 $\mathcal{H}$ を微分してゼロとおけば，それが最小解を導く方程式となる．すなわち

$$\nabla\mathcal{H}(\boldsymbol{w}) = 0 \Rightarrow \hat{\boldsymbol{w}}, \tag{8}$$

$$\nabla\mathcal{H}^{\setminus\mu}(\boldsymbol{w}) = 0 \Rightarrow \hat{\boldsymbol{w}}^{\setminus\mu}, \tag{9}$$

である．フルと LOO データのコスト関数の差を $\epsilon_\mu(\boldsymbol{w}) = \mathcal{H}(\boldsymbol{w}) - \mathcal{H}^{\setminus\mu}(\boldsymbol{w})$ と置くと eq. (9) の左辺は次のように展開できる：

$$\begin{aligned} 0 &= \nabla\mathcal{H}^{\setminus\mu}(\hat{\boldsymbol{w}}^{\setminus\mu}) = \nabla\mathcal{H}^{\setminus\mu}(\hat{\boldsymbol{w}} - \boldsymbol{d}_\mu) \approx \nabla\mathcal{H}^{\setminus\mu}(\hat{\boldsymbol{w}}) - \partial^2\mathcal{H}^{\setminus\mu}(\hat{\boldsymbol{w}})\boldsymbol{d}_\mu \\ &= \nabla(\mathcal{H}(\hat{\boldsymbol{w}}) - \epsilon_\mu(\hat{\boldsymbol{w}})) - \partial^2\mathcal{H}^{\setminus\mu}(\hat{\boldsymbol{w}})\boldsymbol{d}_\mu = -\nabla\epsilon_\mu(\hat{\boldsymbol{w}}) - G^{\setminus\mu}(\hat{\boldsymbol{w}})\boldsymbol{d}_\mu, \end{aligned} \tag{10}$$

となる．最後の変形では eq. (8) から $\nabla\mathcal{H}(\hat{\boldsymbol{w}}) = 0$ であることを使い，かつ，コスト関数ヘシアンを $G^{\setminus\mu}(\hat{\boldsymbol{w}}) = \partial^2\mathcal{H}^{\setminus\mu}(\hat{\boldsymbol{w}})$ とおいた．故に

$$\boldsymbol{d}_\mu \approx -\left(G^{\setminus\mu}(\hat{\boldsymbol{w}})\right)^{-1}\nabla\epsilon_\mu(\hat{\boldsymbol{w}}). \tag{11}$$

ヘシアンの $^{\setminus\mu}$ が邪魔である．フルデータのヘシアンと比較すると

$$G^{\setminus\mu}(\hat{\boldsymbol{w}}) = G(\hat{\boldsymbol{w}}) - \partial^2\epsilon_\mu(\hat{\boldsymbol{w}}), \tag{12}$$

とかける．これを代入し，

$$\boldsymbol{d}_\mu \approx -\left(G(\hat{\boldsymbol{w}}) - \partial^2\epsilon_\mu(\hat{\boldsymbol{w}})\right)^{-1}\nabla\epsilon_\mu(\hat{\boldsymbol{w}}). \tag{13}$$

を得る．こうして

$$\hat{\boldsymbol{w}}^{\backslash\mu} = \hat{\boldsymbol{w}} - \boldsymbol{d}_\mu \approx \hat{\boldsymbol{w}} + \Big(G(\hat{\boldsymbol{w}}) - \partial^2\epsilon_\mu(\hat{\boldsymbol{w}})\Big)^{-1}\nabla\epsilon_\mu(\hat{\boldsymbol{w}}) \tag{14}$$

となる．LOO 解 $\boldsymbol{w}^{\backslash\mu}$ がフル解 $\hat{\boldsymbol{w}}$ で書けたので，目標は達成された．ここまで導出で，コスト関数が二乗誤差であることとか正則化項の性質などは一切使っていない．従って eq. (14) は特異性のない広い範囲の回帰や分類問題に対して適用可能な枠組みになっている．

正則化の無い通常の線形回帰の場合の式を最後にまとめておこう：

$$G(\hat{\boldsymbol{w}}) = X^\top X, \tag{15}$$
$$\epsilon_\mu(\hat{\boldsymbol{w}}) = \frac{1}{2}(y_\mu - \boldsymbol{x}_\mu^\top\hat{\boldsymbol{w}})^2, \tag{16}$$
$$\nabla\epsilon_\mu(\hat{\boldsymbol{w}}) = (y_\mu - \boldsymbol{x}_\mu^\top\hat{\boldsymbol{w}})\boldsymbol{x}_\mu, \tag{17}$$
$$\partial^2\epsilon_\mu(\hat{\boldsymbol{w}}) = \boldsymbol{x}_\mu\boldsymbol{x}_\mu^\top, \tag{18}$$

これらを全部 eq. (7) に代入し，シャーマン・モリソンの公式

$$\Big(A + \boldsymbol{b}\boldsymbol{c}^\top\Big)^{-1} = A^{-1} - \frac{A^{-1}\boldsymbol{b}\boldsymbol{c}^\top A^{-1}}{1 + \boldsymbol{c}^\top A^{-1}\boldsymbol{b}}, \tag{19}$$

を使うと LOOCV の場合の CV 誤差 (LOO 誤差)$\epsilon_{\rm LOO}$ はコンパクトな表現になる：

$$\begin{aligned}\epsilon_{\rm LOO} &= \sum_{\mu=1}^{M}\frac{1}{2}||y_\mu - \boldsymbol{x}_\mu^\top\hat{\boldsymbol{w}}^{\backslash\mu}||_2^2 \\ &= \sum_{\mu=1}^{M}\frac{1}{2}\Big(1 + \boldsymbol{x}_\mu^\top(X^\top X - \boldsymbol{x}_\mu\boldsymbol{x}_\mu^\top)^{-1}\boldsymbol{x}_\mu\Big)^2(y_\mu - \boldsymbol{x}_\mu^\top\hat{\boldsymbol{w}})^2, \qquad (20)\\ &= \sum_{\mu=1}^{M}\frac{1}{2}\Big(1 - \boldsymbol{x}_\mu^\top(X^\top X)^{-1}\boldsymbol{x}_\mu\Big)^{-2}(y_\mu - \boldsymbol{x}_\mu^\top\hat{\boldsymbol{w}})^2. \qquad (21)\end{aligned}$$

実は，これは近似ではなく厳密な結果である．というのはただの線形回帰の場合，eq. (13) の関係が厳密に成り立つからである．これに名前があるのかどうかわからないが (PRESS 統計量？)，一応標準的な結果のようである [7].

3.1.2. 正則化が ℓ_1 ノルムの場合

次に，正則化項を $J(\boldsymbol{w}|\boldsymbol{\eta} = (\lambda)) = \lambda||\boldsymbol{w}||_1$ とした場合について考える．この場合は特別な名前がついており LASSO と呼ばれる [8]．こうすると $\hat{\boldsymbol{w}}$ の各要素は，ゼロのものと非ゼロのものに分かれる．前者を非アクティブ，後者をアクティブ変数と呼び，アクティブ変数の添字集合を $\hat{A} = \{i|\hat{w}_i \neq 0\}$ と書き，サポートと呼ぶ．コスト関数の勾配に対して次の性質がある[*7]:

---

*7 ℓ_1 ノルムも原点を除いて区分的に微分可能であるので，勾配は定義域のほとんどで定義されている．原点は注意が必要だが，ここでは無視しても問題ない議論なので立ち入らない．

$$(\nabla\mathcal{H}(\hat{\boldsymbol{w}}))_i = \begin{cases} 0 & (i \in \hat{A}) \\ c_i(\neq 0) & (i \notin \hat{A}) \end{cases} . \tag{22}$$

これが意味することは，非アクティブ変数を最初からゼロとおいて，残った変数に関してだけの問題を考えれば，普通にゼロ勾配条件で解が得られるということである．そこで我々は次の仮定をおく：

仮定： フル解 $\hat{\boldsymbol{w}}$ と LOO 解 $\hat{\boldsymbol{w}}^{\setminus\mu}$ でアクティブセットは同じである．

こうすれば，上の特異性がない場合の議論をアクティブ変数に対してそのまま適用できる．ヘシアンも勾配も単にアクティブ変数に関してだけ計算すれば良い．結果は

$$\hat{\boldsymbol{w}}^{\setminus\mu}_{\hat{A}} \approx \hat{\boldsymbol{w}}_{\hat{A}} + \left(G_{\hat{A}\hat{A}} - (\partial^2\epsilon_\mu)_{\hat{A}\hat{A}}\right)^{-1} (\nabla\epsilon_\mu(\hat{\boldsymbol{w}}))_{\hat{A}}\,, \tag{23a}$$

$$\hat{\boldsymbol{w}}^{\setminus\mu}_{\hat{A}^c} \approx \boldsymbol{0} \tag{23b}$$

ここで $\hat{A}^c$ は $\hat{A}$ の補集合である．またベクトル $\boldsymbol{x}_{\hat{A}}$ は，$\boldsymbol{x}$ の要素のうち対応するアクティブな要素のみから成る小ベクトルを表す．行列の添字として使われた場合も意味は同じである．これにより LASSO における LOO 誤差は

$$\epsilon_{\mathrm{LOO}} \approx \sum_{\mu=1}^{M} \frac{1}{2}\left(1 + \left(\boldsymbol{x}_\mu^\top\right)_{\hat{A}} \left\{\left(X^\top X - \boldsymbol{x}_\mu\boldsymbol{x}_\mu^\top\right)_{\hat{A}\hat{A}}\right\}^{-1} (\boldsymbol{x}_\mu)_{\hat{A}}\right)^2 (y_\mu - \boldsymbol{x}_\mu^\top\hat{\boldsymbol{w}})^2, \tag{24}$$

$$= \sum_{\mu=1}^{M} \frac{1}{2}\left\{1 - \left(\boldsymbol{x}_\mu^\top\right)_{\hat{A}} \left\{(X^\top X)_{\hat{A}\hat{A}}\right\}^{-1} (\boldsymbol{x}_\mu)_{\hat{A}}\right\}^{-2} (y_\mu - \boldsymbol{x}_\mu^\top\hat{\boldsymbol{w}})^2, \tag{25}$$

と近似される．ℓ_1 ノルムは二階微分すると消えてしまうので，コスト関数ヘシアン G に寄与が無いことに注意せよ．

さて上でおいた仮定は正しいのだろうか？結論からいうと正しくなく，任意のスパース性を持つアルゴリズムは LOO 操作によってサポートが変わることが証明されている [9]．にもかかわらず eq. (25) の近似は非常に良い．[10] ではこの理由を定量的に考察し，

1. サポートに出入りする変数の数がサポートの大きさに対して十分小さい，
2. サポートを出入りした変数の係数変化が十分小さい

の両方が成り立つ場合に，近似による誤差が無視できるくらい小さいことを示し，ℓ_1 正則化の場合にこれが実際に成り立つことを N, M に関するスケーリングの議論から示している．

3.1.3. 正則化が ℓ_1 ノルム以外の場合

上で述べた 2 つの条件のうち，前者は ℓ_1 ノルムを $p < 1$ の ℓ_p ノルムに置き換えても成り立つことが期待できるのに対し，後者は成立条件がもっと厳しい．小さい摂動が入って非アクティブ変数がアクティブになったとき，その係数が急激に大きくなるようだと困るわけであるが，$p < 1$ では実際そのようになることが知られている．従って eq. (25) のような便利な近似公式はこの場合存在しない．

しかしそもそも $p < 1$ の ℓ_p ノルムを正則化としてそのまま使うことは稀である．これは上

で述べた解の急激な変化が効率的に解を探索することを妨げてしまうためである．従って解の連続性を担保しつつ ℓ_1 ノルムの良くない点を克服する，というのがより現実的な方向である．

ℓ_1 ノルムの良くない点は推定解に強くバイアスがかかることにある．つまり真の解が本当に線形モデル $\boldsymbol{y} = X\boldsymbol{w}_0$ から生成していたとしても，推定解 $\hat{\boldsymbol{w}}$ は $\lambda||\boldsymbol{w}||_1$ のせいで，$\boldsymbol{w}_0$ より全体的に（ノルムの意味で）小さい推定量を返してしまうのである．

この原因は ℓ_1 ノルムの勾配が，w が大きい領域でも消失しない点にある．これを克服する一つの方法が，ℓ_1 ノルムと ℓ_0 ノルムを適切になめらかにつなぐというものである．一つの例として SCAD(smoothly clipped absolute deviation) 正則化というものが存在する [11]．一つの要素に対してその表式を顕に書くと

$$J(w|\boldsymbol{\eta} = (\lambda, a)) = \begin{cases} \lambda|w| & (|w| \leq \lambda) \\ -\frac{w^2 - 2a\lambda|w| + \lambda^2}{2(a-1)} & (\lambda < |w| \leq a\lambda) \\ \frac{(a+1)\lambda^2}{2} & (a\lambda < |w|) \end{cases}, \tag{26}$$

となる．ベクトルを引数とした場合は，各要素ごとにこの関数を適用し単純に和を取る．ℓ_1 と SCAD のプロットを図 1 に載せた．見てわかるように，SCAD では遠方での勾配がゼロにな

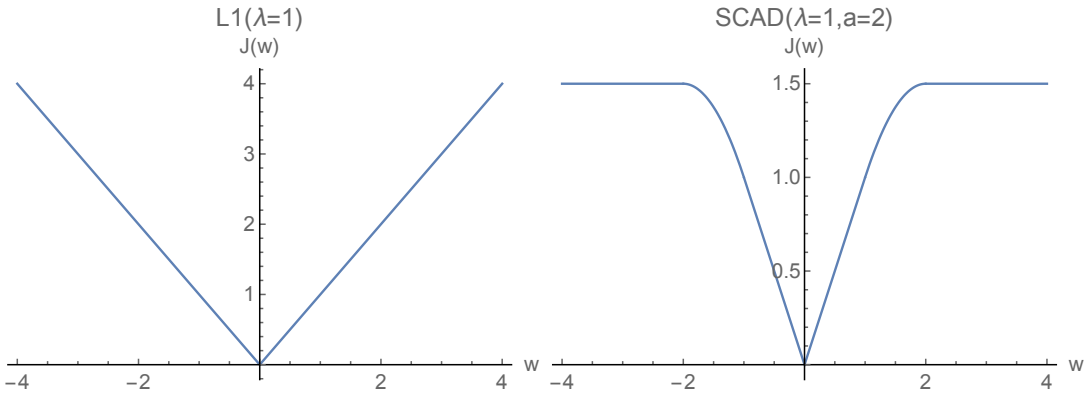

図 1　正則化項の振る舞い．(左) ℓ_1 の場合．(右) SCAD の場合．SCAD は遠方での勾配がゼロになっていることがわかる．

ることがわかる．これにより推定解へのバイアスが大幅に軽減される．

SCAD は $p < 1$ の ℓ_p ノルムとは異なり解の連続性が保たれるため，上で議論した条件が満たされる．従って eq. (23) と形式的に同じ近似公式が成り立つ．唯一の違いはコスト関数ヘシアンに SCAD 項からの寄与があるということである．すなわち

$$\left(\partial^2 J(\boldsymbol{w}|(\lambda, a))\right)_{ij} = \frac{1}{1-a}\delta_{ij} I\left(\lambda < |w_i| \leq a\lambda\right), \tag{27}$$

を G に足す必要がある．それ以外は ℓ_1 の場合と全く同様である．ちなみに I は指示関数で引数が真のとき 1，そうでないとき 0 を返す．

ここでは SCAD で説明したが，SCAD に限らず解の連続性が保たれる正則化であれば，eq.

(23) を適用することが出来て，CV の計算量の問題を回避することが出来る．これはかなり広い範囲の回帰・分類問題を含むので，色々な場面での応用が期待される．

3.2. さらなる近似

さて，2 章でボルツマン分布を導入しておきながら，ここまで全く使わずに単純な摂動論だけで話を進めてきた．実は以上の導出は，ボルツマン分布を経由して線形応答理論と呼ばれる枠組みに当てはめることで，統計力学的な定式化に移すことが出来る．実際 [10] ではその方法で導出している．しかし単純な方法で答えが出た以上，それを再度ここでやる意味は薄い．

代わりに再度 ℓ_1 の場合を考え，eq. (25) を更に近似して単純化することを考える．これを統計力学的定式化に基づいて行う[*8]．ちなみになぜ更なる近似を考えるかというと，eq. (25) は，一回で良いとはいえ，逆行列 $\left(X^\top X\right)^{-1}$ を取るゆえに $O(N^3)$ の計算量がかかるからである．２０１８年現在の現実的レベルの計算リソースを仮定した場合，これだと取り扱える問題は $N = O(10^4)$ 程度であろう．一方で，実応用の現場では線形回帰で $N = O(10^5), O(10^6)$ のサイズの問題を取り扱うこともしばしばある．そう考えると eq. (25) は計算コストをかけ過ぎなのである．

以下では，統計力学で Cavity 法と呼ばれる方法に従ってその近似を導出する[*9]．

3.2.1. Cavity 法

ボルツマン分布を改めて書き下す：

$$P(\boldsymbol{w}|D,\boldsymbol{\eta}) = \frac{1}{Z}e^{-\beta E(\boldsymbol{w}|D)}e^{-\beta\lambda||\boldsymbol{w}||_1} = \frac{1}{Z}\prod_{\mu=1}^{M} e^{-\beta\frac{1}{2}(y_\mu - \boldsymbol{x}_\mu^\top \boldsymbol{w})^2}\prod_{i=1}^{N} e^{-\beta\lambda|w_i|}, \tag{28}$$

最後は構造がわかりやすいよう積的に分離した形で書いた．これを Cavity 法で次のように分解する：

$$\tilde{M}_{\mu\to i}(w_i) \propto \int \prod_{j(\neq i)} dw_j \; e^{-\frac{\beta}{2}\left(y_\mu - \boldsymbol{x}_\mu^\top \boldsymbol{w}\right)^2} \prod_{j(\neq i)} M_{j\to\mu}(w_j), \tag{29}$$

$$M_{i\to\mu}(w_i) \propto e^{-\beta\lambda|w_i|} \prod_{\nu(\neq\mu)} \tilde{M}_{\nu\to i}(w_i). \tag{30}$$

ここで $M, \tilde{M}$ をメッセージと呼ぶ．正確な解説は教科書に譲るとして [12, 13]，ここでは Cavity 法の気持ちだけ解説しよう．ボルツマン分布の周辺分布を２種類導入し次のように書く：

$$P_i(w_i) \equiv \int \prod_{j(\neq i)} dw_j P(\boldsymbol{w}), \tag{31}$$

$$P_i^{\backslash\mu}(w_i) \equiv \int \prod_{j(\neq i)} dw_j \frac{1}{Z^{\backslash\mu}} \prod_{\nu(\neq\mu)}^{M} e^{-\beta\frac{1}{2}(y_\nu - \boldsymbol{x}_\nu^\top \boldsymbol{w})^2} \prod_{j=1}^{N} e^{-\beta\lambda|w_j|}, \tag{32}$$

---

*8　それ以外の方法でやるのは恐らく可能だが，少なくとも筆者には統計力学的定式化のほうが簡単なので．

*9　コンピュータサイエンスでは類似の方法が存在し，それを信念伝搬法（Belief propagation, BP）と呼ぶ．

前者は w_i の通常の周辺分布である．後者はコスト関数のうち μ 番目のデータ点が無いとした LOO 系における周辺分布 (LOO 周辺分布) である．$M_{i\to\mu}(w_i)$ は LOO 周辺分布 $P_i^{\backslash\mu}(w_i)$ の “近似” 解と見なすことが出来る．$\tilde{M}_{\mu\to i}(w_i)$ はその近似解の “構成要素” の分布である．Eqs. (29,30) は LOO 周辺分布とその構成要素を Self-consistent に解くための関係式なのだが，その背後にあるのは “ツリー近似” である．言葉で説明するのは難しいので図 2 を見てもらいたい．左図はもともとのボルツマン分布 (28) をグラフィカルに表現したものであり，丸が変数

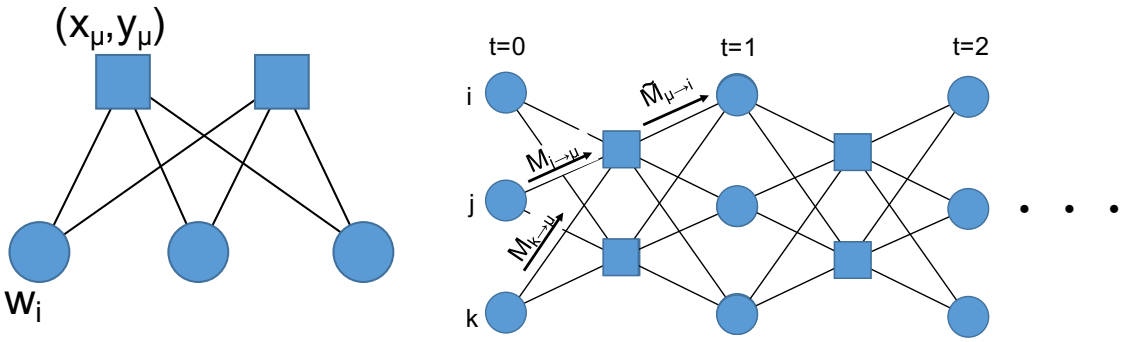

図 2 （左）Eq. (28) のグラフィカルモデル構造．（右）左図のグラフィカルモデルをツリー状，あるいはフィードフォワード型に展開したもの．左から右へとメッセージ $\tilde{M}_{\mu\to i}, M_{i\to\mu}$ が伝搬し更新されていく．その収束解が左図のグラフィカルモデルにおける周辺分布を近似したものとみなせる．

ノードと呼ばれるもので w_i とその事前重み (今の場合 $e^{-\beta\lambda|w_i|}$ と思ってほしい) を表す；四角が関数ノードと呼ばれるもので変数間の相互作用を考慮した重み $e^{-\frac{1}{2}\beta(y_\mu-\boldsymbol{x}_\mu^\top\boldsymbol{w})}$ を表現しているものと考える．右図はそれを９０度回して層状に重ねたもので，層から層へ，左から右へメッセージが伝搬していくと考える．$t=0$ と $t=1$ の間にある関数ノード μ は，$t=0$ の変数ノードからのメッセージ $\{M_{j\to\mu}\}_j$ を受け取り，自分の重み $e^{-\frac{1}{2}\beta(y_\mu-\boldsymbol{x}_\mu^\top\boldsymbol{w})}$ をかけて，$t=1$ の変数ノードへメッセージを伝搬する．この際，変数ノード i へ送る際には，$t=0$ の i からのメッセージ $M_{i\to\mu}$ はダイレクトには送らない．これを $t=1,\cdots$ と繰り返していくと，やがてメッセージ $M, \tilde{M}$ が収束する．その収束解を元々の問題の周辺分布の近似解とするのである．自分自身へのフィードバックが本質的に無いことからツリー近似と等価になっている．実際，元々のグラフがツリー構造なら，この処方箋は厳密な周辺分布を導く．

一方でグラフが極めて密に相互作用している場合も，広い範囲の X についてこのツリー近似が $N\to\infty$ の極限で厳密になると信じられている．特に $\boldsymbol{x}_\mu$ が $\forall\mu$ について独立同分布から引かれている場合は証明が存在する [14, 15]．密に相互作用していても正しいのは，他のノードからの影響による項が圧倒的に多いために，自分自身のフィードバックが相対的に無視でき

るからである[*10]．以下この正しさを仮定し，さらなる近似公式を導こう．

3.2.2. さらなる近似公式の導出

Eqs. (29,30) をもとに近似公式を導く．ポイントは $\tilde{M}$ がガウシアン

$$\tilde{M}_{\mu\to i}(w_i) \propto e^{\beta\left(-\frac{1}{2}\tilde{\Gamma}_{\mu\to i}w_i^2+\tilde{h}_{\mu\to i}w_i\right)}, \tag{33}$$

と見なせるという点にある．これは eq. (29) において $\{w_j\}_{j(\neq i)}$ がそれぞれ独立な分布 $M_{j\to\mu}$ から生成され，かつ，それがコスト関数 E の中で単純な線形和 $\sum_j x_{\mu j}w_j$ としてのみ現れているため，中心極限定理が使えるからである．

一旦これを見切れば，あとはその係数 $\tilde{h},\tilde{\Gamma}$ が Eqs. (29,30) でどう Self-consistent に決まるかを考えれば良い．中心極限定理から $M_{j\to\mu}(w_j)$ 分布における w_j の分散 $V_j^{\backslash\mu}$ と平均 $m_j^{\backslash\mu}$ のみが $\tilde{h},\tilde{\Gamma}$ を決めるのに重要である．さらにガウシアン性から，分散と二次の係数 $\tilde{\Gamma}$ は，平均及び一次の係数 $\tilde{h}$ とは独立に取り扱うことが出来る．結局，この分散と二次係数の方程式は

$$\tilde{\Gamma}_{\mu\to i} = \frac{x_{\mu i}^2}{1+\sum_{j(\neq i)}x_{\mu j}^2\chi_j^{\backslash\mu}}, \tag{34a}$$

$$\chi_i^{\backslash\mu} = \begin{cases} \left(\sum_{\nu(\neq\mu)}\tilde{\Gamma}_{\nu\to i}\right)^{-1} & (i\in\hat{A}) \\ 0 & (i\notin\hat{A}) \end{cases}, \tag{34b}$$

と閉じさせることができる．ここで分散を β でリスケールして $\chi_j^{\backslash\mu}\equiv\beta V_j^{\backslash\mu}$ とおいた．

Eq. (34a) の分母において，$\sum_{j(\neq i)}x_{\mu j}^2\chi_j^{\backslash\mu}\approx\sum_j x_{\mu j}^2\chi_j^{\backslash\mu}$ が成り立つだろうことは，N という大きな数の項の和をとっていることから容易に想像できる．すなわち $\tilde{\Gamma}_{\mu\to i}$ の μ,i-依存性は分子の $x_{\mu i}^2$ に依っている．さらに eq. (34b) では，$\sum_{\nu(\neq\mu)}\tilde{\Gamma}_{\nu\to i}$ という $M-1$ 個の多数の項の和を取ることから，$\chi_i^{\backslash\mu}$ の μ,i-依存性もやはり弱いということが期待される．そこでこれを無視することを考え，

$$\chi_i^{\backslash\mu} \approx \begin{cases} \chi & (i\in\hat{A}) \\ 0 & (i\notin\hat{A}) \end{cases}, \tag{35}$$

と置く．こうすると eq. (34) は χ について解くことが出来て[*11]，

$$\chi \approx \frac{1}{(M-|\hat{A}|)\sigma_x^2}, \tag{36}$$

を得る．ここで $\sigma_x^2=\sum_\mu\sum_i x_{\mu i}^2/(MN)$ は計画行列要素の二乗平均である[*12]．さらに $\chi_i^{\backslash\mu}$ の

---

*10 ただし 2 ステップ前の自己からの寄与は有限に残ることが示せ，それはきちんと取り入れる．この項のことをオンサーガー反跳項 (Onsager reaction term) と呼ぶ．ちなみにこれを取り入れるか否かが，本稿の最後で述べる AMP アルゴリズムの収束性にクリティカルな影響を与える．

*11 サポートのサイズ $|\hat{A}|$ が十分大きいとすれば，eq. (34a) の分母は $1+\sum_{j(\neq i)}x_{\mu j}^2\chi\approx 1+\chi|\hat{A}|\sigma_x^2$ と近似できる．さらに $\sum_{\nu(\neq\mu)}x_{\nu i}^2\approx M\sigma_x^2$ である．

*12 いわゆる標準化をしている場合は $\sigma_x^2=1$ である．

元々の意味は LOO 系のボルツマン分布に関するリスケールした分散・共分散であった．アクティブ変数だけ見たとき, LOO ボルツマン分布はローカルには精度行列が $\left(X^\top X - \boldsymbol{x}_\mu \boldsymbol{x}_\mu^\top\right)_{\hat{A}\hat{A}}$ のガウス分布と見なせることから

$$\chi_{ij}^{\backslash\mu} \equiv \beta \mathrm{Cov}^{\backslash\mu}(w_i, w_j) = \left(\left(\left(X^\top X - \boldsymbol{x}_\mu \boldsymbol{x}_\mu^\top \right)_{\hat{A}\hat{A}} \right)^{-1} \right)_{ij}, \tag{37}$$

となる $(i, j \in \hat{A})$．Eqs. (34-36) は，$\chi_{ij}^{\backslash\mu}$ の対角要素の値が eq. (36) になり，かつ，非対角要素が無視できるということを示唆している，すなわち

$$\left(\left(\left(X^\top X - \boldsymbol{x}_\mu \boldsymbol{x}_\mu^\top \right)_{\hat{A}\hat{A}} \right)^{-1} \right)_{ij} \approx \frac{1}{(M - |\hat{A}|)\sigma_x^2} \delta_{ij}. \tag{38}$$

これを eq. (24) に代入すれば

$$\epsilon_{\mathrm{LOO}} \approx \left(\frac{M}{M - |\hat{A}|} \right)^2 \sum_{\mu=1}^{M} \frac{1}{2} (y_\mu - \boldsymbol{x}_\mu^\top \hat{\boldsymbol{w}})^2 = \left(\frac{M}{M - |\hat{A}|} \right)^2 \frac{1}{2} ||\boldsymbol{y} - X\hat{\boldsymbol{w}}||_2^2, \tag{39}$$

となり，目標のさらなる近似公式を得る．これはトレーニングエラー $\frac{1}{2}||\boldsymbol{y} - X\hat{\boldsymbol{w}}||_2^2$ に単純なファクター $\left(\frac{M}{M-|\hat{A}|}\right)^2$ を掛ければよいというだけの式であり，計算量がほとんどかからない．

実際の適用例として，UCI machine learning repository のワインクオリティデータ [16, 17] に対して LASSO を適用してみた結果を図 3 に乗せる．近似と実際に LOOCV を行った結果

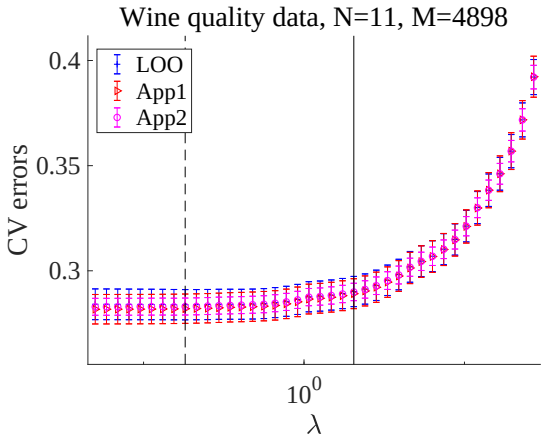

図 3　ワインクオリティデータに対する LASSO の CV 誤差の λ に対するプロット．App1 が eq. (25), App2 がさらなる近似 eq. (39) による結果．いずれも実際に LOOCV を行った結果と極めて良い一致を見せていることがわかる．点線は CV 誤差の最小値をとる λ の位置, 実線はいわゆる 1 シグマルールによって選ばれる最適な λ の位置を表している．この数値実験では Glmnet[18] を使って Solution path を求めた．

はほとんど重なっていて，良い精度の近似ができていることがわかる．近似公式におけるエラーバーは，eqs. (25,39) のいずれの場合もすべての項の標準偏差を $\sqrt{M}$ で割ることにより付けている．

4. まとめ

本稿では，交差検証法を近似的に行うことでその計算量を低減し，ハイパーパラメータ推定を効率的に行う方法について紹介した．本稿では主に ℓ_1 正則化付き線形回帰を例に説明してきたが，近似の基本発想である線形近似と Cavity 法は汎用的であり，コスト関数や正則化を別のものに変えることで様々な問題に適用できる．実際筆者らはこれまでに以下の問題に対して近似公式を開発・実装してきた：

- LASSO[10].
- SCAD 正則化付き線形回帰 (準備中).
- ２次元 Total variation 正則化付き線形回帰 [19]．(近似公式の MATLAB パッケージを [20] より配布中).
- Elastic net 正則化付き多項ロジスティック回帰による多クラス分類 [21]．(近似公式の MATLAB パッケージを [22] より，python パッケージを [23] より配布中).

多項ロジスティック回帰と Total variation 正則化付き線形回帰では，近似公式自体が込み入っているため，数値計算用のパッケージを公開している．興味があればぜひ使ってみてほしい．

また，本稿は交差検証法に限定した話であったが，別のリサンプリング手法としてブートストラップ法による信頼区間の推定を，やはり統計力学的手法で近似することで，高速に行うことが出来る．LASSO の場合にこれを実装したのが [24]．こちらはコードは公開していないが，MATLAB 版は既に手元にあるので，興味があればこれも筆者に連絡をとってほしい．

最後に，Cavity 法を解説した都合上，最適解 $\hat{\boldsymbol{w}}$ を求める統計力学的アルゴリズムについて少し触れておこう．本稿では近似公式に必要な分散と二次の係数についてのみ方程式を導いたが，平均 $m_j^{\backslash\mu}$ と一次の係数 $\tilde{h}_{\mu\to j}$ の関係式も同様に導くことができる．それを解く（大体の場合は反復代入をする）と最適解 $\hat{\boldsymbol{w}}$ 自体を計算するアルゴリズムとなる (平均値 $\boldsymbol{m}$ を $\hat{\boldsymbol{w}}$ と同一視することが出来る)．この方法は Approximate message passing(AMP) と呼ばれ [4, 5, 6]，線形回帰に限らず様々な問題に適用され新規なアルゴリズムが開発されている．LASSO の場合，AMP の計算量は $O(N^2)$ となり，汎用的な凸関数最小化法（内点法など）の計算量 $O(N^3)$ に比べると少なくてすむ．筆者の知る限り，現時点で AMP を計算量のオーダーで下回るアルゴリズムは存在せず，State of the art の結果となっている．

参考文献

[1] 幸人, 伊庭. ベイズ統計と統計物理. 岩波書店, 2003.

[2] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, Vol. 381, No. 6583, p. 607, 1996.

[3] Hiroki Terashima and Haruo Hosoya. Sparse codes of harmonic natural sounds and their modulatory interactions. *Network: Computation in Neural Systems*, Vol. 20, No. 4, pp. 253–267, 2009.

[4] Yoshiyuki Kabashima. A cdma multiuser detection algorithm on the basis of belief propagation.

Journal of Physics A: Mathematical and General, Vol. 36, No. 43, p. 11111, 2003.

[5] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, Vol. 106, No. 45, pp. 18914–18919, 2009.

[6] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pp. 2168–2172. IEEE, 2011.

[7] Hastie Trevor, Tibshirani Robert, and Friedman Jerome. *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. Springer-Verlag New York, 2009.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[9] Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 34, No. 1, pp. 187–193, 2012.

[10] Tomoyuki Obuchi and Yoshiyuki Kabashima. Cross validation in lasso and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2016, No. 5, pp. 53304–53339, 2016.

[11] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, Vol. 96, No. 456, pp. 1348–1360, 2001.

[12] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[13] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[14] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, Vol. 57, No. 2, pp. 764–785, 2011.

[15] Jean Barbier, Nicolas Macris, Mohamad Dia, and Florent Krzakala. Mutual information and optimality of approximate message-passing in random linear estimation. *arXiv preprint arXiv:1701.05823*, 2017.

[16] M. Lichman. UCI machine learning repository, 2013.

[17] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, Vol. 47, No. 4, pp. 547–553, 2009.

[18] Jerome Friedman, Trevor Hastie, Noah Simon, Junyang Qian, and Rob Tibshirani. glmnet. `https://CRAN.R-project.org/package=glmnet`.

[19] Tomoyuki Obuchi, Shiro Ikeda, Kazunori Akiyama, and Yoshiyuki Kabashima. Accelerating cross-validation with total variation and its application to super-resolution imaging. *PloS one*, Vol. 12, No. 12, p. e0188012, 2017.

[20] Tomoyuki Obuchi. Matlab package of approximate CV on linear regression penalized by l1 and two-dimensional total variation. `https://github.com/T-Obuchi/AcceleratedCVon2DTVLR`, 2017.

[21] Tomoyuki Obuchi and Yoshiyuki Kabashima. Accelerating cross-validation in multinomial logistic regression with l1-regularization. *arXiv preprint arXiv:1711.05420*, 2017.

[22] Tomoyuki Obuchi. Matlab package of ACV on MLR. `https://github.com/T-Obuchi/AcceleratedCVonMLR_matlab`, 2017.

[23] Takashi Takahashi and Tomoyuki Obuchi. Python package of ACV on MLR. `https://github.com/T-Obuchi/AcceleratedCVonMLR_python`, 2017.

[24] Tomoyuki Obuchi and Yoshiyuki Kabashima. Semi-analytic resampling in lasso. *arXiv preprint arXiv:1802.10254*, 2018.