ランダム特徴モデルのレプリカ計算のメモ

髙橋 昂

May 25, 2021

Abstract

機械学習の研究でこの数年よく出てくるランダム特徴モデル（このメモの設定では教師の構造にランダム特徴使ってないので、Hidden manifold modelと呼ばれることもある）のレプリカ計算のメモ。Gaussian equivalence theorem（GET）を使ったあとのレプリカ計算についてまとめている[1]。GETの導出自体はここでは扱っていない。元ネタは[GLK+20]。

1 setting

まず、分析者が与えられているデータの性質、および分析したい学習と予測の方法についてまとめる。

1.1 データ

この問題では、一次元の出力からなる回帰あるいは分類の問題を扱う。データ分析者が与えられているデータDは特徴量$\boldsymbol{x}_\mu \in \mathbb{R}^N$と、ターゲット$y_\mu \in \mathcal{Y} \subset \mathbb{R}$の組の集まりである:

$$D = \{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^M. \tag{1}$$

ここで、$\mathcal{Y}$は回帰なら$\mathbb{R}$であり、分類なら$\{1, -1\}$である。

$\boldsymbol{x}_\mu, y_\mu$の生成ルールは以下の通りである。まず、$\boldsymbol{c}_\mu, \mu = 1, 2, \ldots, M, \boldsymbol{\theta}_0 \in \mathbb{R}^D, F \in \mathbb{R}^{D\times N}$を以下のような量として定める:

$$\boldsymbol{c}_\mu \sim \mathcal{N}(\boldsymbol{0}_D, I_D), \quad \mu = 1, 2, \ldots, M \tag{2}$$

$$\boldsymbol{\theta}_0 \in \mathbb{R}^D, \theta_{0,k} \sim q_\theta, \quad k = 1, 2, \ldots, D \tag{3}$$

$$F \in \mathbb{R}^{D\times N}. \tag{4}$$

ここでは、Fの各行を$\boldsymbol{f}_i \in \mathbb{R}^D$としたとき、

$$\frac{1}{\sqrt{D}} \boldsymbol{f}_i^\top \boldsymbol{f}_j = \begin{cases} \sqrt{D}, & i = j \\ \mathcal{O}(1), & i \neq j \end{cases}, \tag{5}$$

を満たすような、後述のGaussian equivalence assumptionが成り立つ程度に性質のよい行列であるとしておく[2]。

このもとで、ターゲットは

$$y_\mu \sim q_y\left(\cdot \,\middle|\, \frac{1}{\sqrt{D}} \boldsymbol{c}_\mu^\top \boldsymbol{\theta}_0\right), \tag{6}$$

[1] 仮定した先は線形モデルなんで、目を瞑っててもできる、、、という人もいると思うけど、、、。

[2] これが成り立つ条件については、[GMKZ20]などを参照のこと。多分、回転不変なランダム行列とかなら大丈夫ではないかと思う。

と与えられるものとする。つまり、ターゲットy_μはD次元の回帰係数$\boldsymbol{\theta}_0$と特徴量$\boldsymbol{c}_\mu$から一般化線形観測によって得られている。いっぽう、特徴量$\boldsymbol{x}_\mu$については、ある関数$\sigma : \mathbb{R} \to \mathbb{R}$を用いて以下のようにして作られている

$$\boldsymbol{x}_\mu = \sigma\left(\frac{1}{\sqrt{D}} F^\top \boldsymbol{c}_\mu\right). \tag{7}$$

ただし、σは要素ごとに作用するものとする。この式をみると、分析者が持っている特徴量$\boldsymbol{x}_\mu$はD次元の真の特徴量[3]$\boldsymbol{c}_\mu$からFによってP次元に移した後、適当な非線形変換を加えることによって生成されているということになる。なお、解析を簡単にするために、$\mathbb{E}[\sigma(\xi)] = 0, \xi \sim \mathcal{N}(0,1)$となるもののみを扱うことにする。

1.1.1 熱力学極限

特に、ここでは$N, M, D \to \infty, M/N \to \alpha, D/N \to \gamma, \alpha, \gamma \in (0, \infty)$の大自由度極限に興味がある[4]。この大自由度極限を統計物理の監修に従って熱力学極限と呼ぶ。書くのが面倒なので、以下ではこの極限のことを単に$N \to \infty$と書くことにする。

1.1.2 モデルの解釈について

ランダム特徴モデルとして　分析者が$\boldsymbol{c}_\mu, F$を知っている場合、これはいわゆるランダム特徴と呼ばれるものになっていて、ある種のカーネル法だと思うことができる。こっちの設定がポピュラーだと思うが、その場合y_μの生成ルールについてもランダム特徴を使うようにする場合が多いかもしれない。

Hidden manifold modelとして　逆に、分析者が$\boldsymbol{c}_\mu, F$を知らない設定だと解釈すると、真の入出力関係はD次元の線形モデルなのだが、それとは何か全然違う次元Nの特徴量を与えられていてそこで戦わされているのだと思うこともできる。この場合には、$D < N$であるような場合に、本当は低次元の構造を持った入出力関係をうまいこと見つけられるのかというのが焦点になる[5]。

1.2 学習と予測

学習については、ランダム特徴$\boldsymbol{x}_\mu$を用いた線形モデルを、適当な損失関数$l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$と、正則化強度$\lambda$のRidge正則化を用いて以下のように学習する:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \mathbb{R}^N} \left[\sum_{\mu=1}^{M} l(y_\mu, \boldsymbol{w}^\top \boldsymbol{x}_\mu) + \frac{\lambda}{2} \sum_{i=1}^{N} w_i^2\right]. \tag{8}$$

さらに、予測については上で学習した$\hat{\boldsymbol{w}}$を用いてある予測関数fを用い、新たな入力点$\boldsymbol{x} \in \mathbb{R}^N$における出力の予測値$\hat{y}$を

$$\hat{y}(\boldsymbol{x}; \hat{\boldsymbol{w}}) = f(\hat{\boldsymbol{w}}^\top \boldsymbol{x}), \tag{9}$$

と予測するものとする。

汎化誤差ϵ_{g}は以下のように定める:

$$\epsilon_{\mathrm{g}} = \mathbb{E}_{\boldsymbol{x}, y}\left[(y - \hat{y}(\boldsymbol{x}; \hat{\boldsymbol{w}}))^2\right]. \tag{10}$$

[3] この辺の用語について正しい言い回しを髙橋はわかってない。

[4] 統計力学の解析全部がこういうスケーリング極限をとるわけではない。スパース推定とかだと、非ゼロ成分の個数を$\mathcal{O}(1)$にして計算したものもあるし…。ランダム系の平均場のなかでも比較的簡単になるところとして、こういう領域があるのだということだと思う。

[5] Hidden manifold modelと言う名前については、この設定を焦点にしたもののようだ。

ゴールは、上の汎化誤差の熱力学極限での振る舞いを調べることである。

2 method

2.1 Boltzmann分布と自由エネルギー

逆温度$\beta > 0$の$\mathbb{R}^N$上のBoltzmann分布[6]を以下の様に定める

$$p^{(\beta)}(\boldsymbol{w}|D) = \frac{1}{Z}\prod_{\mu=1}^{M} p_y(y_\mu|\boldsymbol{w}^\top \boldsymbol{x}_\mu)^\beta \prod_{i=1}^{N} p(w_i)^\beta, \tag{11}$$

$$Z = \int \prod_{\mu=1}^{M} p_y(y_\mu|\boldsymbol{w}^\top \boldsymbol{x}_\mu)^\beta \prod_{i=1}^{N} p(w_i)^\beta d\boldsymbol{w}, \tag{12}$$

$$p_y(y_\mu|\boldsymbol{w}^\top \boldsymbol{x}_\mu) = e^{-l(y_\mu, \boldsymbol{x}_\mu^\top \boldsymbol{w})}, \tag{13}$$

$$p(w_i) = e^{-\frac{\lambda}{2}w_i^2}. \tag{14}$$

尤度と事前分布のようなものを導入したのは、単なる便利のためである。また、以下では

$$\prod_{i=1}^{n} p(w_i) = p(\boldsymbol{w}), \tag{15}$$

といった記法も使うことにする。規格化定数Zは分配関数と呼ばれている。逆温度無限大$\beta \to \infty$の極限[7]では、このBoltzmann分布は$\sum_{\mu=1}^{M} \log p_y(y_\mu|\boldsymbol{w}^\top \boldsymbol{x}_\mu) - \sum_{i=1}^{N} \log p(w_i)$を最小にするような$\boldsymbol{w}$の上の一様分布に収束する。つまり、最適化問題(8)の解上の一様分布に収束する。というわけで、このBoltzmann分布の$\beta \to \infty$の$\beta \to \infty$での性質を調べることが学習の問題と等価である[8]。最適化問題の解析であったが、一応確率分布の解析の問題として扱うことができるようになる。分配関数を用いて、Helmholtzの自由エネルギー密度は以下のように定義される:

$$\tilde{f}(D) = \lim_{N\to\infty} -\frac{1}{\beta N}\log Z. \tag{16}$$

これはキュムラント生成関数になっているので、もしこの自由エネルギーの形がわかれば、大体欲しいものは何でも計算できることになる。といっても、これはほぼ方便で、実際にはレプリカ計算の過程を通じて、多体問題の効果がどのような有効的な場に置き換えられているのかを観察することになることが平均場スピングラス理論では多いと思う[9]。

自由エネルギーをあるDの実現値ごとに扱うのは難しいので、自己平均性が成立する(自由エネルギーの値に典型性が成立する、測度がある一点に集中する...)と期待し、期待値の評価$\mathbb{E}_D[\tilde{f}(D)]$に問題をすり替えることにする[10]。特に、我々は$\beta \to \infty$の問題に興味がある:

$$f = \lim_{N,\beta\to\infty} -\frac{1}{N\beta}\mathbb{E}_D[\log Z]. \tag{17}$$

[6] $d\boldsymbol{w}$も合わせてGibbs measureと呼ぶこともある。

[7] $1/\beta$が温度Tなので、ゼロ温度と呼ぶこともある。

[8] いまは凸問題なので、あんまりこのBoltzmann分布を選ぶというところに悩みはない。一般にはあるアルゴリズムによって得られる解の平衡分布があったとした場合に、それに対応するようなBoltzmann分布（かそれに類した何か）を構築する必要がある。そもそも平衡分布があるのかって言う問題はあるが。

[9] もちろん、熱力学関数そのものに興味があることだってあるが。

[10] これは別に自明なことではない。

2.2 レプリカ法

恒等式$\mathbb{E}_D[\log Z] = \lim_{n\to 0} n^{-1}\log \mathbb{E}[Z^n]$を用い、極限$N\to\infty$と$n\to 0$の順序を入れ替えられるとすれば、自由エネルギーの典型評価は以下のように書き直せる:

$$f = \lim_{n\to 0}\frac{1}{n}\phi_n,$$
$$\phi_n = \lim_{N,\beta\to\infty}\frac{-1}{\beta N}\log \mathbb{E}_D\left[Z^n\right].$$

これは（極限の順序の入れ替えを気にしなければ）ただの恒等式である。この表式の利点は$n=1,2,\dots$に対しては分配関数の定義式を用いてϕ_nが

$$\phi_n = \lim_{N,\beta\to\infty}\frac{-1}{\beta N}\log \mathbb{E}_D\left[\int\prod_{a=1}^{n}\left\{\prod_{\mu=1}^{M} p_y(y_\mu|\boldsymbol{w}_a^\top\boldsymbol{x}_\mu)^\beta\prod_{i=1}^{N}p(w_{a,i})^\beta\right\}d^n\boldsymbol{w}\right],$$

と書き直せる点にある。ここで、$\boldsymbol{w}_a\in\mathbb{R}^N, a=1,2,\dots,n$であり、$d^n\boldsymbol{w} = d\boldsymbol{w}_1\dots d\boldsymbol{w}_n$である。もとの表式では、$\log Z$という、$\boldsymbol{w}$の積分をとって対数をとった後にはどのように$D$に依存しているのかよくわからないものの平均を取る必要があった。しかし、ここではD依存性が透明なものの平均をとるだけでよい。先にDについての平均をとってしまって、その後でN次元からNn次元に高次元化された問題を扱おうというわけである。こちらは、もし平均をとった後でも対称性が高ければ、統計力学でも統計学でも何でも普通の多体問題を扱う手段が使えるだろう。$n=1,2,\dots$に対して計算結果を得たとする。もしその表式がnの離散性を陽に含まなければ、$n\to 0$への外挿によって答を得ることができる。このようにしてべきの極限で難しいところを書き直し、整数べきからの外挿によって結果を得る方法を総称してレプリカ法と呼ぶ[11]。$\mathbb{E}[Z^n]$の整理の仕方は問題ごとの特性があるところである[12]。

2.3 Gaussian equivalence assumption

D平均をとるところで、$\boldsymbol{x}_\mu$での平均をとる必要がある。これは$\boldsymbol{c}_\mu$に非線形に依存していて、素朴にはこの平均をとるのは容易ではない。しかし、Fが前述の条件を満たしている場合に以下が成り立つと期待する[13]:

$$\boldsymbol{x}_\mu = \kappa_1\frac{1}{\sqrt{D}}F^\top\boldsymbol{c}_\mu + \kappa_*\boldsymbol{z}_\mu, \tag{18}$$
$$\boldsymbol{z}_\mu\sim\mathcal{N}(\boldsymbol{0}_N, I_N), \tag{19}$$
$$\kappa_1 = \mathbb{E}_\xi[\xi\sigma(\xi)], \kappa_* = \sqrt{\mathbb{E}_\xi[\sigma(\xi)^2]-\kappa_1^2}, \xi\sim\mathcal{N}(0,1). \tag{20}$$

つまり、非線形な関数を通して$\boldsymbol{c}_\mu$に依存しているが、実際にはデータのランダムネスの効果は線形にしか効かないとする。これはいくつかのFに対しては実際に成り立つことが証明されていて、Gaussian equivalence theoremと呼ばれている。ただ、ここではこの導出には主たる興味ではないので、これが成り立つとして計算を進める[14]。

---

[11] $\mathbb{E}[Z^n]$から$n\to 0$の極限をとるあたりまでを指してレプリカ計算と呼んでいる気がする。

[12] 今回の様に密に相互作用している系では、二次モーメントと分散で決まる秩序変数とその共役変数を適切に入れることで対処が済む。疎結合な場合には、それでは話が済まない。

[13] いま、$\mathbb{E}_\xi[\sigma(\xi)]=0$であるような$\sigma$に限定しているので、$\kappa_0$に相当する項はない。

[14] この瞬間にRS計算の結果が暗算できるよという人もいると思うが、そういう人はここで閉じていいです。

3 replica calculation

以下では$\mathbb{E}[Z^n]$の計算から、$n \to 0$の外挿がどう行われるか、そしてその様子から汎化誤差がどうなるかを考えていく。

3.1 $\mathbb{E}_D[Z^n]$の整理

まず、$\mathbb{E}_D[Z^n]$を書き下すと、

$$\mathbb{E}_D[Z^n] = \int \prod_{\mu=1}^{M} \int \mathbb{E}_{\boldsymbol{c}_\mu} \left[q_y \left(y_\mu | \frac{1}{\sqrt{D}} \boldsymbol{c}_\mu^\top \boldsymbol{\theta}_0 \right) \prod_{a=1}^{n} p_y \left(y_\mu | \boldsymbol{w}_a^\top \sigma \left(\frac{1}{\sqrt{D}} F^\top \boldsymbol{c}_\mu \right) \right)^\beta \right] dy_\mu \\ \times \prod_{a=1}^{n} p(\boldsymbol{w}_a)^\beta q_\theta(\boldsymbol{\theta}_0) d^n \boldsymbol{w} d\boldsymbol{\theta}_0, \tag{21}$$

である。

3.2 $\boldsymbol{c}_\mu$平均

(21)式の表現をみると、$\{\boldsymbol{c}_\mu\}$に関する平均は各μごとにとっていけばいいことがわかる。特に、上のGaussian equivalence仮定(18)を使うと、$\boldsymbol{c}_\mu$についての平均は以下のように書き直される:

$$\mathbb{E}_{\boldsymbol{c}_\mu} \left[q_y \left(y_\mu | \frac{1}{\sqrt{D}} \boldsymbol{c}_\mu^\top \boldsymbol{\theta}_0 \right) \prod_{a=1}^{n} p_y \left(y_\mu | \boldsymbol{w}_a^\top \sigma \left(\frac{1}{\sqrt{D}} F^\top \boldsymbol{c}_\mu \right) \right)^\beta \right] \\ = \mathbb{E}_{\boldsymbol{c}_\mu, \boldsymbol{z}_\mu} \left[q_y \left(y_\mu | \frac{1}{\sqrt{D}} \boldsymbol{c}_\mu^\top \boldsymbol{\theta}_0 \right) \prod_{a=1}^{n} p_y \left(y_\mu | \boldsymbol{w}_a^\top \left(\kappa_1 \frac{1}{\sqrt{D}} F^\top \boldsymbol{c}_\mu + \kappa_* \boldsymbol{z}_\mu \right) \right)^\beta \right], \tag{22}$$

$$\boldsymbol{c}_\mu \sim \mathcal{N}(\boldsymbol{0}_D, I_D), \boldsymbol{z}_\mu \sim \mathcal{N}(\boldsymbol{0}_N, I_N). \tag{23}$$

ここで、$\boldsymbol{c}_\mu, \boldsymbol{z}_\mu$はガウシアンなので、少なくともある固定された$F, \boldsymbol{w}_a, \boldsymbol{\theta}_0$に対しては、その線形結合である

$$u_0 = \frac{1}{\sqrt{D}} \boldsymbol{c}_\mu^\top \boldsymbol{\theta}_0, \tag{24}$$

$$u_a = \boldsymbol{w}_a^\top \left(\kappa_1 \frac{1}{\sqrt{D}} F^\top \boldsymbol{c}_\mu + \kappa_* \boldsymbol{z}_\mu, \right) a = 1, 2, \ldots, n, \tag{25}$$

もまた（相関した）ガウシアンとなる。したがって、上の$\boldsymbol{c}_\mu$平均は結局は

$$\int q_y(y_\mu | u_0) \prod_{a=1}^{n} p_y(y_\mu | u_a) \mathcal{N}(\boldsymbol{u}; \boldsymbol{0}_{n+1}, \Sigma) d\boldsymbol{u}, \tag{26}$$

と書き直される。ここで、Σは、

$$\boldsymbol{s}_a \equiv \frac{1}{\sqrt{N}} F \boldsymbol{w}_a \in \mathbb{R}^D, \quad a = 1, 2, \ldots, n, \tag{27}$$

として、$a \leq b$に対して以下のように定められる:

$$\Sigma_{ab} = \begin{cases} \Sigma_\rho = \frac{1}{D} \|\boldsymbol{\theta}_0\|_2^2, & a = b = 0 \\ \Sigma_m^{(a)} = \kappa_1 \frac{1}{D} \boldsymbol{\theta}_0^\top \boldsymbol{s}_a & a = 0, b = 1, 2, \ldots, n, \\ \Sigma_Q^{(ab)} = \kappa_1^2 \frac{1}{D} \boldsymbol{s}_a^\top \boldsymbol{s}_b + \kappa_*^2 \frac{1}{N} \boldsymbol{w}_a^\top \boldsymbol{w}_b, & 1 \leq a \leq < b \leq n \end{cases} \tag{28}$$

3.3 秩序変数の導入と鞍点法

$\{\boldsymbol{c}_\mu\}$平均の部分が上のような共分散のガウス積分によって書けることがわかったので、ここで

$$1 \doteq \int \delta(D\rho - \|\boldsymbol{\theta}_0\|_2^2) \prod_{a=1}^{n} \delta(Dm_a^{(s)} - \boldsymbol{\theta}_0^\top \boldsymbol{s}_a) \prod_{1\le a\le b\le n} \delta(NQ_{ab}^{(w)} - \boldsymbol{w}_a^\top \boldsymbol{w}_b)$$
$$\times \prod_{1\le a\le b\le n} \delta(DQ_{ab}^{(s)} - \boldsymbol{s}_a^\top \boldsymbol{s}_b) \prod_{a=1}^{n} \delta\left(\boldsymbol{s}_a - \frac{1}{\sqrt{N}} F\boldsymbol{w}_a\right) dQ^{(w)} dQ^{(s)} dm d\rho d^n \boldsymbol{s}_a, \quad (29)$$

によって秩序変数を導入する。ここで、$\doteq$は「自由エネルギー密度に対する寄与は等しい」を表す記号である。これを導入し、デルタ関数を適当にFourier変換すると、

$$\mathbb{E}_D[Z^n] \doteq \int \exp\left(D\rho\tilde{\rho} - D\sum_{a=1}^{n} m_a\tilde{m}_a + \frac{N}{2}\mathrm{Tr}Q^{(w)}\tilde{Q}^{(w)} + \frac{D}{2}\mathrm{Tr}Q^{(s)}\tilde{Q}^{(s)}\right)$$
$$\times \prod_{\mu=1}^{M}\left\{\int q_y(y_\mu|u_{0,\mu}) \prod_{a=1}^{n} p_y(y_\mu|u_{a,\mu}) \mathcal{N}(\boldsymbol{u}_\mu; \boldsymbol{0}_{n+1}, \Sigma) dy_\mu d\boldsymbol{u}_\mu\right\}$$
$$\times \left\{\int \prod_{a,b} \exp\left(-\frac{1}{2}\tilde{Q}_{ab}^{(w)} \boldsymbol{w}_a^\top \boldsymbol{w}_b - \frac{1}{2}\tilde{Q}_{ab}^{(s)} \boldsymbol{s}_a \boldsymbol{s}_b\right) \prod_{a=1}^{n} \exp\left(\tilde{m}_a^{(s)} \boldsymbol{\theta}_0^\top \boldsymbol{x}_a\right)\right.$$
$$\left.\times \prod_{a=1}^{n} p(\boldsymbol{w}_a)^\beta \delta\left(\boldsymbol{s}_a - \frac{1}{\sqrt{N}} F\boldsymbol{w}_a\right) q_\theta(\boldsymbol{\theta}_0) d^n\boldsymbol{w} d^n\boldsymbol{s}_a d\boldsymbol{\theta}_0\right\} dQ^{(w)} dQ^{(s)} dm d\rho d\tilde{Q}^{(w)} d\tilde{Q}^{(s)} d\tilde{m} d\tilde{\rho}, \quad (30)$$

となる。ここで、Q, mなどを秩序変数、デルタ関数のフーリエ変換から出てきたチルダ付きの変数を共役変数と呼ぶことにする。

さて、

$$\psi_y = \log \int q_y(y_\mu|u_0) \prod_{a=1}^{n} p_y(y_\mu|u_a) \mathcal{N}(\boldsymbol{u}; \boldsymbol{0}_{n+1}, \Sigma) dy d\boldsymbol{u}, \quad (31)$$

$$\psi_w = \frac{1}{N} \log \int \prod_{a,b} \exp\left(-\frac{1}{2}\tilde{Q}_{ab}^{(w)} \boldsymbol{w}_a^\top \boldsymbol{w}_b - \frac{1}{2}\tilde{Q}_{ab}^{(s)} \boldsymbol{s}_a \boldsymbol{s}_b\right) \prod_{a=1}^{n} \exp\left(\tilde{m}_a^{(s)} \boldsymbol{\theta}_0^\top \boldsymbol{x}_a\right)$$
$$\times \prod_{a=1}^{n} p(\boldsymbol{w}_a)^\beta \delta\left(\boldsymbol{s}_a - \frac{1}{\sqrt{N}} F\boldsymbol{w}_a\right) q_\theta(\boldsymbol{\theta}_0) d^n\boldsymbol{w} d^n\boldsymbol{s}_a d\boldsymbol{\theta}_0, \quad (32)$$

によってψ_y, ψ_wを導入すると[15]、

$$\mathbb{E}_D[Z^n] = \int e^{N\tilde{g}(\rho, Q^{(w)}, Q^{(s)}, m^{(s)}, \tilde{\rho}, \tilde{Q}^{(w)}, \tilde{Q}^{(s)}, \tilde{m}^{(s)})} dQ^{(w)} dQ^{(s)} dm d\rho d\tilde{Q}^{(w)} d\tilde{Q}^{(s)} d\tilde{m} d\tilde{\rho}, \quad (33)$$

$$\tilde{g} = \gamma\rho\tilde{\rho} - \gamma\sum_{a=1}^{n} m_a\tilde{m}_a + \frac{1}{2}\mathrm{Tr}Q^{(w)}\tilde{Q}^{(w)} + \frac{\gamma}{2}\mathrm{Tr}Q^{(s)}\tilde{Q}^{(s)} + \alpha\psi_y + \psi_w \quad (34)$$

[15]ここではFによって相関が入るので、wについての有効問題が一体問題化できない。それでもRidge正則化ならば、後で見るように漸近固有値分布だけが効いて(determinantで済むから)Stieltjes変換で閉じた式が得られる。他の正則化だと、漸近的に満たすべき式は出てくるが、その方程式が低次元積分で書けない。

この事情は、特徴量が多変量正規分布で与えられている場合の線形回帰に似ている。[JM14]では熱力学極限での性能評価ではなくて、鞍点方程式を検定統計量の構成につかっている（有限のサイズなら直接逆行列計算すればよいので)。ちなみに、熱力学極限のマクロ構造を使って検定統計量作るというのは[TK18]でもやっている（ただし特徴量の構造が違う)。

となる。ここで、ψ_y, ψ_wはそれぞれ共役変数と$\boldsymbol{w}, \boldsymbol{s}$の有効的な生成関数である[16]。

$N \to \infty$で鞍点評価すると、

$$\lim_{N\to\infty} \log \mathbb{E}_D[Z^n] = \underset{\substack{\rho, Q^w, Q^s, m^{(s)}, \\ \tilde{\rho}, \tilde{Q}^{(w)}, \tilde{Q}^{(s)}, \tilde{m}^{(s)}}}{\mathrm{extr}} \left[\tilde{g}(\rho, Q^{(w)}, Q^{(s)}, m^{(s)}, \tilde{\rho}, \tilde{Q}^{(w)}, \tilde{Q}^{(s)}, \tilde{m}^{(s)}) \right], \tag{35}$$

である。

なお、$\lim_{n\to 0} \mathbb{E}[Z^n] = 1$の条件から、$n \to 0$では$\rho = \int \theta_0^2 q_\theta(\theta_0) d\theta_0, \tilde{\rho} = 1/\rho$が要請される。以降この値に固定して考える。

一応、鞍点条件を一般に書くと、以下のようになる:

$$Q_{ab}^{(w)} = -\frac{\partial}{\partial \tilde{Q}_{ab}^{(w)}} \psi_w, \tag{36}$$

$$Q_{ab}^{(s)} = -\frac{1}{\gamma} \frac{\partial}{\partial \tilde{Q}_{ab}^{(s)}} \psi_w, \tag{37}$$

$$m_a^{(s)} = \frac{1}{\gamma} \frac{\partial}{\partial \tilde{m}_a^{(s)}} \psi_w, \tag{38}$$

$$\tilde{Q}_{ab}^{(w)} = -\alpha \frac{\partial}{\partial Q_{ab}^{(w)}} \psi_y, \tag{39}$$

$$\tilde{Q}_{ab}^{(s)} = -\frac{\alpha}{\gamma} \frac{\partial}{\partial Q_{ab}^{(s)}} \psi_y, \tag{40}$$

$$\tilde{m}_a^{(s)} = \frac{\alpha}{\gamma} \frac{\partial}{\partial m_a^{(s)}} \psi_y, \tag{41}$$

右辺の微分を書き下すと、適当な有効問題の期待値になっている[17]。

[16]共役変数を直接出そうとすると面倒なので、q_y, p_yに直接秩序変数で書かれたキャビティ場を飛ばさないほうが綺麗な気はする。

[17]大抵の場合、これを一般のnに対して書いておいてから、そこにRS(B)の構造を突っ込むのが筋よしっぽい。RS(B)の自由エネルギーを書いてからだと、かえって計算が大変になることが多いような気がする。

3.4 レプリカ対称仮定

鞍点法の評価までは落としたが、現時点ではa, bの離散性が陽に残っていて、$n \to 0$の外挿ができない。そこで、鞍点の構造に、以下のレプリカ対称性を仮定することにする

$$m_a^{(s)} = m^{(s)}, \tag{42}$$
$$Q_{ab}^{(w)} = q^{(w)}, \quad a < b, \tag{43}$$
$$Q_{aa}^{(w)} = q^{(w)} + \frac{\chi^{(w)}}{\beta}, \tag{44}$$
$$Q_{ab}^{(s)} = q^{(s)}, \quad a < b, \tag{45}$$
$$Q_{aa}^{(s)} = q^{(s)} + \frac{\chi^{(s)}}{\beta}, \tag{46}$$
$$\tilde{m}_a^{(s)} = \beta \hat{m}^{(s)}, \tag{47}$$
$$\tilde{Q}_{ab}^{(w)} = -\beta^2 \hat{\chi}^{(w)}, \tag{48}$$
$$\tilde{Q}_{aa}^{(w)} = \beta \hat{Q}^{(w)} - \beta^2 \hat{\chi}^{(w)}, \tag{49}$$
$$\tilde{Q}_{ab}^{(s)} = -\beta^2 \hat{\chi}^{(s)}, \tag{50}$$
$$\tilde{Q}_{aa}^{(s)} = \beta \hat{Q}^{(s)} - \beta^2 \hat{\chi}^{(s)}. \tag{51}$$

この仮定は、nが1以上の整数であれば基本的には正しいと思うが、$n < 1$でも成り立っている保証はない。

上のようにしてから整理すると、nの離散性が陽に出ないようになり、$n \to 0$の外挿が行える形になる。以下、RS仮定下での鞍点条件の$n \to 0$極限を整理していくことにする。

3.4.1 共役変数側

上のようなRS構造を仮定した場合の鞍点方程式を考える。定数$A > 0$、およびp次元ベクトル$\boldsymbol{v}_a \in \mathbb{R}^p, a = 1, 2, \ldots, n$に対する恒等式

$$\exp\left(\frac{A}{2}\|\sum_{a=1}^n \boldsymbol{v}_a\|_2^2\right) = \mathbb{E}_{\boldsymbol{\xi}}\left[\prod_{a=1}^n \exp\left(\sqrt{A}\boldsymbol{\xi}^\top \boldsymbol{v}_a\right)\right], \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_p, I_p), \tag{52}$$

を用いると[18]、ψ_wの内側が以下のようになることに注目する:

$$\mathbb{E}_{\boldsymbol{\xi}_w, \boldsymbol{\xi}_s}\left[\prod_{a=1}^n \exp\left(-\frac{\beta(\hat{Q}^{(w)} + \lambda)}{2}\boldsymbol{w}_a^\top \boldsymbol{w}_a + \beta\sqrt{\hat{\chi}^{(w)}}\boldsymbol{\xi}_w^\top \boldsymbol{w}_a \right.\right.$$
$$\left.\left. -\frac{\beta\hat{Q}^{(s)}}{2}\boldsymbol{s}_a^\top \boldsymbol{s}_a + \beta\left(\sqrt{\hat{\chi}^{(s)}}\boldsymbol{\xi}_s + \hat{m}^{(s)}\boldsymbol{\theta}_0\right)^\top \boldsymbol{s}_a\right)\delta\left(\boldsymbol{s}_a - \frac{1}{\sqrt{N}}F\boldsymbol{w}_a\right)\right], \tag{53}$$
$$\boldsymbol{\xi}_w \sim \mathcal{N}(\mathbf{0}_N, I_N), \boldsymbol{\xi}_s \sim \mathcal{N}(\mathbf{0}_D, I_D). \tag{54}$$

RS仮定のもとでは、

$$\exp\left(-\frac{\beta(\hat{Q}^{(w)} + \lambda)}{2}\boldsymbol{w}_a^\top \boldsymbol{w}_a + \beta\sqrt{\hat{\chi}^{(w)}}\boldsymbol{\xi}_w^\top \boldsymbol{w}_a \right.$$
$$\left. -\frac{\beta\hat{Q}^{(s)}}{2}\boldsymbol{s}_a^\top \boldsymbol{s}_a + \beta\left(\sqrt{\hat{\chi}^{(s)}}\boldsymbol{\xi}_s + \hat{m}^{(s)}\boldsymbol{\theta}_0\right)^\top \boldsymbol{s}_a\right)\delta\left(\boldsymbol{s}_a - \frac{1}{\sqrt{N}}F\boldsymbol{w}_a\right) \tag{55}$$

[18] この恒等変形のことをHubbard-Stratonovich変換と呼ぶこともある。元ネタの論文だと、これに使う変数を 1 次元で入れてたけど、素直にp次元で入れたほうが見通しがよいように思う。

に比例した$\boldsymbol{w}, \boldsymbol{s}$の分布がパラメータ空間でのエフェクティブな分布で、$\boldsymbol{\xi}_w, \boldsymbol{\xi}_s$がその要素方向のばらつき、あるいは要素ごとの$D$に関する揺らぎであると解釈される。また、$\chi$は1次モーメントの局所場に対する応答の平均、$q$は1次モーメントの二乗の平均、$m$は$\boldsymbol{\theta}_0$と$\boldsymbol{s}$の一次モーメントの規格化された内積である[19]。これを踏まえると、秩序変数のマクロ表現を与える役割である共役変数の極値条件から出てくる鞍点方程式は下記のようになる:

$$\chi^{(w)} = \lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\left[\left((\hat{Q}^{(w)}+\lambda)I_N + \hat{Q}^{(s)}\frac{F^\top F}{N}\right)^{-1}\right], \tag{56}$$

$$q^{(w)} = \lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\left[\left(\hat{\chi}^{(w)}I_N + (\hat{\chi}^{(s)}+(\hat{m}^{(s)})^2)\frac{F^\top F}{N}\right)\left((\hat{Q}^{(w)}+\lambda)I_N + \hat{Q}^{(s)}\frac{F^\top F}{N}\right)^{-2}\right], \tag{57}$$

$$\chi^{(s)} = \lim_{N\to\infty} \frac{1}{N\gamma}\mathrm{Tr}\left[F\left((\hat{Q}^{(w)}+\lambda)I_N + \hat{Q}^{(s)}\frac{F^\top F}{N}\right)^{-1}F^\top\right], \tag{58}$$

$$q^{(s)} = \lim_{N\to\infty} \frac{1}{N\gamma}\mathrm{Tr}\left[F\left((\hat{Q}^{(w)}+\lambda)I_N + \hat{Q}^{(s)}\frac{F^\top F}{N}\right)^{-1}\left(\hat{\chi}^{(w)}I_N + (\hat{\chi}^{(s)}+(\hat{m}^{(s)})^2)\frac{F^\top F}{N}\right)\right.$$
$$\left.\times\left((\hat{Q}^{(w)}+\lambda)I_N + \hat{Q}^{(s)}\frac{F^\top F}{N}\right)^{-1}F^\top\right], \tag{59}$$

$$m^{(s)} = \lim_{N\to\infty} \frac{\hat{m}^{(s)}}{N\gamma}\mathrm{Tr}\left[F\left((\hat{Q}^{(w)}+\lambda)I_N + \hat{Q}^{(s)}\frac{F^\top F}{N}\right)^{-1}F^\top\right]. \tag{60}$$

これは単にガウス積分の期待値、及び共分散を書き下して、$\boldsymbol{\xi}_w, \boldsymbol{\xi}_s$で期待値をとっただけである。この表式を得た上で、$F$を特異値分解して特異値で書き直し、$FF^\top/N$が漸近固有値分布$\tilde{\mu}$を持つと仮定して、$FF^\top/N$のStieltijes変換を$g_\mu(z) = \int \frac{d\tilde{\mu}(t)}{t-z}$、さらに$z = \frac{\hat{Q}^{(w)}+\lambda}{\hat{Q}^{(s)}}$として[20]、

$$\chi^{(w)} = \frac{\gamma}{\hat{Q}^{(w)}+\lambda}\left(\frac{1}{\gamma} - 1 + zg_\mu(-z)\right), \tag{61}$$

$$q^{(w)} = \frac{\hat{\chi}^{(w)}\gamma}{(\hat{Q}^{(w)}+\lambda)^2}\left(\frac{1}{\gamma} - 1 + z^2 g'_\mu(-z)\right) - \frac{\hat{\chi}^{(s)}+(\hat{m}^{(s)})^2}{\hat{Q}^{(s)}(\hat{Q}^{(w)}+\lambda)}\gamma\left(-zg_\mu(-z) + z^2 g'_\mu(-z)\right) \tag{62}$$

$$\chi^{(s)} = \frac{1}{\hat{Q}^{(s)}}(1 - zg_\mu(-z)), \tag{63}$$

$$q^{(s)} = \frac{\hat{\chi}^{(s)}+(\hat{m}^{(s)})^2}{(\hat{Q}^{(s)})^2}\left(1 - 2zg_\mu(-z) + z^2 g'_\mu(z)\right) - \frac{\hat{\chi}^{(w)}}{\hat{Q}^{(s)}(\hat{Q}^{(w)}+\lambda)}\left(zg_\mu(-z) - z^2 g'_\mu(-z)\right), \tag{64}$$

$$m^{(s)} = \frac{\hat{m}^{(s)}}{\hat{Q}^{(s)}}(1 - zg_\mu(-z)). \tag{65}$$

[19] というか、そうなるように定めた。このあたり、添字を抜いたりして大雑把に雰囲気だけ書いている。

[20] $F^\top F/N$のほうが綺麗になったのではないかという予感もないことはない。

3.4.2 秩序変数側

共役変数を定める秩序変数についての極値条件の側については、ϕ_yの内訳を考える必要がある。そこで、$\mathcal{N}(\boldsymbol{u};\mathbf{0}_{n+1},\Sigma)$を整理する。

$$q = \kappa_1^2 q^{(s)} + \kappa_*^2 q^{(w)}, \tag{66}$$

$$m = \kappa_1 m^{(s)}, \tag{67}$$

$$\chi = \kappa_1^2 \chi^{(s)} + \kappa_*^2 \chi^{(w)}, \tag{68}$$

と定めると、RS仮定のもとでは$\mathcal{N}(\boldsymbol{u};\mathbf{0}_{n+1},\Sigma)$に従う確率変数$\boldsymbol{u}$は以下のようにおけることに注目する:

$$u_0 = \sqrt{\rho - \frac{m^2}{q}}\tilde{u}_0 + \frac{m}{\sqrt{q}}\xi, \tag{69}$$

$$u_a = \sqrt{q}\xi + \sqrt{\frac{\chi}{\beta}}\tilde{u}_a, \quad a = 1,2,\ldots,n \tag{70}$$

$$\xi, \tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_n \sim \mathcal{N}(0,1). \tag{71}$$

すると、ξを固定すると、それぞれ$\mathcal{N}(u_0;\frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}), \mathcal{N}(u_a;\sqrt{q}\xi,\frac{\chi}{\beta}), a = 1,2,\ldots,n$に従っていると解釈できる。なので、$\psi_y$の内側は

$$\int q_y(y|u_0)\mathcal{N}\left(u_0;\frac{m}{\sqrt{q}}\xi,\rho-\frac{m^2}{q}\right)\prod_{a=1}^{n} p_y(y|u_a)^\beta \mathcal{N}\left(u_a;\sqrt{\xi},\frac{\chi}{\beta}\right)dy\frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}d\xi d^{n+1}u, \tag{72}$$

と書かれる。これはある入力点ごとに、q_y, p_yに入力される変数がそれぞれ$\mathcal{N}(u_0;\frac{m}{\sqrt{q}}\xi,\rho-\frac{m^2}{q}), \mathcal{N}(u;\sqrt{q}\xi,\frac{\chi}{\beta})$にしたがって分布しているということを表している[21]。さらに、$\hat{Q},\hat{\chi}$は、いまのRSの構成では基本的には

$$\sqrt{-1}\frac{u-\sqrt{q}\xi}{\chi/\beta} \tag{73}$$

の

$$p_y(y|u)^\beta \mathcal{N}\left(u;\sqrt{\xi},\frac{\chi}{\beta}\right) \tag{74}$$

に比例した分布での 1 次モーメントの二乗、および 1 次モーメント（の複素数倍）局所場に対する応答と書かれる[22]。また、$\hat{m}$は

$$\sqrt{-1}\frac{u-\sqrt{q}\xi}{\chi/\beta}, \tag{75}$$

の上の分布での平均と、

$$\sqrt{-1}\frac{u_0 - \frac{m}{\sqrt{q}}\xi}{\rho - \frac{m^2}{q}} \tag{76}$$

の積である。これらを外から、$q_{y|u0}, e^{-\frac{\xi^2}{2}}$をかけて$y,\xi$で平均をとる。このあたりで共役変数が標準化した変数の分散などで書かれるようになる（計算上の）起源がFourier変換から来ているので複素数倍がついていて符号で混乱するが、それはしょうがない。特に、$\beta\to\infty$の極限を考えて、

$$\hat{u} = \arg\min_{u\in\mathbb{R}}\left[-\log p(y|u) + \frac{\left(u-\sqrt{q}\xi\right)^2}{\chi}\right], \tag{77}$$

[21] 要素方向にもそうだろうし、入力点について平均をとることを考えてもそうなっているはず

[22] ただし、q,mの構成の問題で、κ由来の係数がつく。

と書くと、鞍点条件は以下のようになる:

$$\hat{Q}^{(w)} = -\alpha\kappa_*^2 \int \frac{\partial}{\partial(\sqrt{q}\xi)} \left(\frac{\hat{u} - \sqrt{q}\xi}{\chi} \right) q_y(y|u_0) \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) du_0 dy D\xi, \tag{78}$$

$$\hat{\chi}^{(w)} = \alpha\kappa_*^2 \int \left(\frac{\hat{u} - \sqrt{q}\xi}{\chi} \right)^2 q_y(y|u_0) \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) du_0 dy D\xi, \tag{79}$$

$$\hat{Q}^{(s)} = -\frac{\alpha\kappa_1^2}{\gamma} \int \frac{\partial}{\partial(\sqrt{q}\xi)} \left(\frac{\hat{u} - \sqrt{q}\xi}{\chi} \right) q_y(y|u_0) \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) du_0 dy D\xi, \tag{80}$$

$$\hat{\chi}^{(s)} = \frac{\alpha\kappa_1^2}{\gamma} \int \left(\frac{\hat{u} - \sqrt{q}\xi}{\chi} \right)^2 q_y(y|u_0) \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) du_0 dy D\xi, \tag{81}$$

$$\hat{m}^{(s)} = \frac{\alpha\kappa_1}{\gamma} \int \left(\frac{u_0 - \frac{m}{\sqrt{q}}\xi}{\rho - \frac{m^2}{q}} \right) \left(\frac{\hat{u} - \sqrt{q}\xi}{\chi} \right) q_y(y|u_0) \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) du_0 dy D\xi. \tag{82}$$

これは、[GLK+20]の鞍点方程式と、ノーテーションを合わせれば合っている（はず）[23]。

3.5 汎化誤差

上の、ある入力点でのq_y, p_yに入力される分布の解釈から、汎化誤差は以下のように書ける:

$$\epsilon_g = \int (y - f(u))^2 q_y(y|u_0) \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) \mathcal{N}(u; |\sqrt{q}\xi, \frac{\chi}{\beta}) dy du_0 du \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} d\xi. \tag{83}$$

ここで、$q = \kappa_1^2 q^{(s)} + \kappa_*^2 q^{(w)}, m = \kappa_1 m^{(s)}, \chi = \kappa_1^2 \chi^{(s)} + \kappa_*^2 \chi^{(w)}$は先の鞍点方程式の固定点である。

特に、q_yが単に$f_0(u_0)$を与えるだけのものである場合、

$$\epsilon_g = \int (f_0(u_0) - f(u))^2 \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) \mathcal{N}(u; \sqrt{q}\xi, \frac{\chi}{\beta}) du_0 du \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} d\xi, \tag{84}$$

である。

ややわかりにくいが、上の計算を逆にたどれば

$$\left(\int \mathcal{N}(u_0; \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}) \mathcal{N}(u; \sqrt{q}\xi, \frac{\chi}{\beta}) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} d\xi \right) du_0 du \tag{85}$$

の部分が平均ゼロ、共分散

$$\begin{pmatrix} \rho & m \\ m & q + \frac{\chi}{\beta} \end{pmatrix}, \tag{86}$$

の２変数ガウス測度を表していることがわかる。

特に、$\beta \to \infty$では$\chi/\beta \to 0$（これはゼロ温度で相転移したりしていなければ大丈夫だと思われる）であることに注意すると、論文の式に等しいことがわかる。

参考文献

[GLK+20] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová, *Generalisation error in learning with random features and the hidden manifold model*, International Conference on Machine Learning, PMLR, 2020, pp. 3452–3462.

[23] このパラメータの置き方が良いのかわからんなあ。

[GMKZ20] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová, *Modeling the influence of data structure on learning in neural networks: The hidden manifold model*, Physical Review X **10** (2020), no. 4, 041044.

[JM14] Adel Javanmard and Andrea Montanari, *Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory*, IEEE Transactions on Information Theory **60** (2014), no. 10, 6522–6554.

[TK18] Takashi Takahashi and Yoshiyuki Kabashima, *A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements*, Journal of Statistical Mechanics: Theory and Experiment **2018** (2018), no. 7, 073405.