PAPER *Special Section on Information Theory and Its Applications*

Empirical Bayes Estimation for L_1 Regularization: A Detailed Analysis in the One-Parameter Lasso Model

Tsukasa YOSHIDA[†a)], *Nonmember and* **Kazuho WATANABE**[†b)], *Member*

SUMMARY Lasso regression based on the L_1 regularization is one of the most popular sparse estimation methods. It is often required to set appropriately in advance the regularization parameter that determines the degree of regularization. Although the empirical Bayes approach provides an effective method to estimate the regularization parameter, its solution has yet to be fully investigated in the lasso regression model. In this study, we analyze the empirical Bayes estimator of the one-parameter model of lasso regression and show its uniqueness and its properties. Furthermore, we compare this estimator with that of the variational approximation, and its accuracy is evaluated.

key words: *lasso regression, empirical Bayes, Laplace prior, local variational approximation*

1. Introduction

Regularization methods are often used in regression models for the purposes such as the suppression of overfitting and sparse modeling. Lasso (least absolute shrinkage and selection operator) regression is one of the most popular method for sparse modeling. It introduces the regularization term defined by L_1 norm of the regression coefficients, and has widely been used for variable selection and compressed sensing [1]–[3].

It is necessary for successful applications of regularization methods to determine the regularization parameter appropriately, which adjusts the degree of regularization. The cross-validation is the most common approach to the estimation of the regularization parameter. Since this approach generally requires huge computational costs, alternative approaches such as the empirical Bayes estimation has been applied. The empirical Bayes approach in the L_2 regularization, namely ridge regression, itself provides a sparsity-inducing mechanism called automatic relevance determination (ARD) [4], [5]. Furthermore, the empirical Bayes solution of the regularization parameter was fully analyzed in the ridge regression model with the identity design matrix [6], [7]. For the L_1 regularization, however, we no longer have the conjugacy between the Gauss model and the Laplace prior corresponding to the L_1 regularization. For non-conjugate models, the integral calculation of the marginal likelihood in the empirical Bayes estimation is generally intractable. Hence, an approximation method such as the local variational approximation (LVA) is required for the empirical Bayes approach in lasso regression [8], [9]. Because such an approximate solution of the empirical Bayes estimator is obtained by an iterative algorithm, the properties of the approximate or exact empirical Bayes estimators are yet to be fully investigated.

In this study, we analyze in detail the empirical Bayes estimator of the regularization parameter in the simplest one-parameter model of lasso regression. We prove the exact solution of this model and its upper bound. This gives a rare example where the empirical Bayes solution with a non-conjugate model is analytically obtained. The property of the marginal likelihood of this model such as unimodality is demonstrated in the proof of the main theorem. The asymptotic behavior of the solution is also analyzed. Furthermore, we compare the exact solution with the approximate solution given by the LVA to examine its accuracy analytically.

The rest of the paper is organized as follows. Section 2 introduces the one-parameter lasso model. The exact solution of the empirical Bayes estimator and its asymptotic expansions are given in Sect. 3 and Sect. 4, respectively. Section 5 derives the algorithm of the LVA and analyzes its solution. The exact and approximate solutions are numerically demonstrated in Sect. 6. The discussion and conclusion follow in Sect. 7 and Sect. 8.

2. One-Parameter Lasso Model

Given the data set $x^n = \{x_1, \dots, x_n\}$ $(x_i \in \mathbb{R}, i = 1, \dots, n)$, we consider the following one-parameter Gauss model with the parameter $w \in \mathbb{R}$ and the prior distribution of w with the hyperparameter $\lambda > 0$*,

$$\begin{aligned} p(x|w) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-w)^2}{2}\right\}, \\ p(w|\lambda) &= \frac{\lambda}{2} \exp\left(-\lambda |w|\right). \end{aligned} \tag{1}$$

This likelihood corresponds to the linear regression model with the intercept w and without any explanatory (input) variables. The maximum a posteriori (MAP) estimation of this model reduces to the (one-parameter) lasso regression

Manuscript received January 26, 2018.
Manuscript revised June 16, 2018.

[†]The authors are with the Dept. of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi-shi, 441-8580 Japan.
a) E-mail: tyoshida@lisl.cs.tut.ac.jp
b) E-mail: wkazuho@cs.tut.ac.jp

DOI: 10.1587/transfun.E101.A.2184

*This model turns out to be equivalent to the Gauss model $\mathcal{N}(w, \sigma^2)$ with the known variance σ^2 by the standardizing variable transformations, $\tilde{x} = x/\sigma$ and $\tilde{w} = w/\sigma$.

with the regularization parameter λ [1], which maximizes the following posterior distribution of w,

$$p(w|x^n, \lambda) = \frac{p(x^n|w)p(w|\lambda)}{Z(\lambda)},$$

where the likelihood is given by $p(x^n|w) = \prod_{i=1}^n p(x_i|w)$ under the i.i.d. assumption. Here,

$$Z(\lambda) = p(x^n|\lambda) = \int p(x^n|w)p(w|\lambda)dw \tag{2}$$

is the marginal likelihood also known as the evidence [10]. The empirical Bayes estimator of the regularization parameter λ is defined by the maximizer of $Z(\lambda)$,

$$\hat{\lambda} \equiv \operatorname*{argmax}_{\lambda} Z(\lambda).$$

We analyze the empirical Bayes estimator $\hat{\lambda}$ of the model (1).

3. Empirical Bayes Estimator

In this section, we prove the main theorem on the empirical Bayes estimator of the model (1). We use the following special function in the theorem,

$$\mathrm{erfcx}(x) \equiv e^{x^2}\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} e^{x^2} \int_x^\infty e^{-t^2} dt,$$

which is also known as the Mills ratio of the Gaussian random variable [11].

Theorem 1 (Main Result): Let $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ be the sample average. The empirical Bayes estimate $\hat{\lambda}$ for the model (1) is unique and is given by

$$\hat{\lambda} = \begin{cases} \infty & \left(|\overline{x}| \le \frac{1}{\sqrt{n}}\right), \\ \lambda^* & \left(|\overline{x}| > \frac{1}{\sqrt{n}}\right), \end{cases} \tag{3}$$

where λ^* is the unique λ satisfying

$$\left(\frac{\lambda^2}{n} - \overline{x}\lambda + 1\right) \mathrm{erfcx}\left\{\sqrt{\frac{n}{2}}\left(\frac{\lambda}{n} - \overline{x}\right)\right\} + \left(\frac{\lambda^2}{n} + \overline{x}\lambda + 1\right) \mathrm{erfcx}\left\{\sqrt{\frac{n}{2}}\left(\frac{\lambda}{n} + \overline{x}\right)\right\} = 2\sqrt{\frac{2}{n\pi}}\lambda \tag{4}$$

and it is evaluated as

$$0 < \lambda^* < \frac{2}{\sqrt{\overline{x}^2 - \frac{1}{n}}}. \tag{5}$$

If $\hat{\lambda} = \infty$, w is estimated to be 0. This is the effect of ARD. Detailed results on the MAP estimator of w are described in Sect. 7.2.

If the true data-generating distribution is the standard normal distribution $\mathcal{N}(0, 1)$, $\hat{\lambda} = \infty$ and $\hat{\lambda} = \lambda^*$ are selected with probabilities (approximately) 0.68% and 0.32, respectively since $\sqrt{n}\overline{x}$ follows $\mathcal{N}(0, 1)$.

(Proof of Theorem 1)

We have the following lemma.

Lemma 1: The marginal likelihood $Z(\lambda)$ of the model (1) is expressed as follows,

$$C\lambda \left[\mathrm{erfcx}\left\{\sqrt{\frac{n}{2}}\left(\frac{\lambda}{n} - \overline{x}\right)\right\} + \mathrm{erfcx}\left\{\sqrt{\frac{n}{2}}\left(\frac{\lambda}{n} + \overline{x}\right)\right\}\right], \tag{6}$$

where $C = \frac{1}{4\sqrt{n(2\pi)^{n-1}}} \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)$.

(Proof of Lemma 1)

By putting (1) into (2), dividing the integration into two parts, $w \in (-\infty,\ 0)$ and $[0,\ \infty)$, we have

$$Z(\lambda) = \frac{\lambda}{2(2\pi)^{\frac{n}{2}}}(I_+ + I_-),$$

where I_+ and I_- are given by the following integration with $+$ and $-$ chosen from $\pm$, respectively,

$$I_\pm = \int_0^\infty \exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i \pm w)^2 - \lambda w\right\} dw.$$

Completing the square with respect to w in the exponent of the integrand yields that

$$I_\pm = \exp\left\{\frac{n}{2}\left(\frac{\lambda}{n} \pm \overline{x}\right)^2 - \frac{1}{2}\sum_{i=1}^n x_i^2\right\} \cdot \int_0^\infty \exp\left[-\frac{n}{2}\left\{w + \left(\frac{\lambda}{n} \pm \overline{x}\right)\right\}^2\right] dw$$

Thus, we obtain (6) by applying the following formulas and definitions of the error functions,

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)dt,$$

$$\mathrm{erf}(-x) = -\mathrm{erf}(x), \tag{7}$$

$$\int_0^\infty \exp\left\{-a(t-b)^2\right\} dt = \frac{1}{2}\sqrt{\frac{\pi}{a}}\left\{1 + \mathrm{erf}(\sqrt{a}b)\right\},$$

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x), \tag{8}$$

$$\mathrm{erfcx}(x) = e^{x^2}\mathrm{erfc}(x). \tag{9}$$

□

Then, the theorem directly follows from the next lemma by the transformations,

$$\Lambda = \frac{\lambda}{\sqrt{2n}}, \text{ and } M = \sqrt{\frac{n}{2}}\overline{x}. \tag{10}$$

This lemma indicates that the marginal likelihood is always unimodal.

Lemma 2: The function

$$f(\Lambda) = \Lambda \{\mathrm{erfcx}(\Lambda - M) + \mathrm{erfcx}(\Lambda + M)\} \tag{11}$$

is strictly monotonically increasing for $\Lambda \geq 0$ if $|M| \leq 1/\sqrt{2}$, and strictly monotonically increasing for $\Lambda \in [0, \ \Lambda^*)$ and strictly monotonically decreasing for $\Lambda \in (\Lambda^*, \ \infty)$ if $|M| > 1/\sqrt{2}$. Here, Λ^* is the unique point satisfying

$$\begin{aligned}&(2\Lambda^2 - 2M\Lambda + 1)\mathrm{erfcx}(\Lambda - M)\\&+ (2\Lambda^2 + 2M\Lambda + 1)\mathrm{erfcx}(\Lambda + M) - \frac{4}{\sqrt{\pi}}\Lambda = 0,\end{aligned} \tag{12}$$

and is evaluated as

$$0 < \Lambda^* < \sqrt{\frac{2}{2M^2 - 1}}. \tag{13}$$

The proof of this lemma is given in Appendix A. □

4. Asymptotic Expansion

In this section, we analyze the asymptotic behavior of the empirical Bayes estimator obtained in Theorem 1. The next theorem is proved in Appendix B. The following asymptotic expansion of $\mathrm{erfcx}(x)$ for large x, which is obtained directly from that of $\mathrm{erfc}(x)$, is the key to the derivation of the asymptotic expansion in this theorem,

$$\mathrm{erfcx}(x) = \frac{1}{\sqrt{\pi}x} + O\left(\frac{1}{x^3}\right). \tag{14}$$

Theorem 2: λ^* in (3) has the asymptotic expansion as $\sqrt{n}|\overline{x}| \to \infty$,

$$\lambda^* \simeq \frac{1}{|\overline{x}|}\left\{1 + \frac{1}{n\overline{x}^2} + O\left(\frac{1}{(n\overline{x}^2)^2}\right)\right\}. \tag{15}$$

The limit $\sqrt{n}|\overline{x}| \to \infty$ includes the two cases, $n \to \infty$ for fixed $|\overline{x}|$ and $|\overline{x}| \to \infty$ for fixed n. In particular for the former case, if the true data-generating distribution is the normal distribution $\mathcal{N}(w_0, 1)$ with $w_0 \neq 0$, the empirical Bayes solution is given by this asymptotic expansion since λ^* is selected with probablity 1 as $n \to \infty$.

The upper bound (5) has the asymptotic expansion as $\sqrt{n}|\overline{x}| \to \infty$,

$$\frac{2}{\sqrt{\overline{x}^2 - \frac{1}{n}}} \simeq \frac{2}{|\overline{x}|}\left\{1 + \frac{1}{2n\overline{x}^2} + O\left(\frac{1}{(n\overline{x}^2)^2}\right)\right\}. \tag{16}$$

It follows from these asymptotic expansions that the upper bound (5) is loose by a factor of 2 asymptotically as $\sqrt{n}|\overline{x}| \to \infty$.

5. Local Variational Approximation

The LVA is commonly used for approximating the posterior distribution of the lasso regression model [8], [9]. The LVA for the model (1) forms the following upper bound to the term $|w| = \sqrt{w^2}$,

$$|w| = \sqrt{w^2} \leq \frac{1}{2\sqrt{\xi^2}}(w^2 - \xi^2) + \sqrt{\xi^2}, \tag{17}$$

which follows from the concavity of the square root. Here, ξ is a parameter, called variational parameter, and is used for optimizing the approximation.

Replacing $|w|$ with the upper bound (17), we put

$$\begin{aligned}\tilde{p}_\xi(w|\lambda) &\equiv \frac{\lambda}{2}\exp\left[-\lambda\left\{\frac{1}{2\sqrt{\xi^2}}(w^2 - \xi^2) + \sqrt{\xi^2}\right\}\right]\\&\leq p(w|\lambda).\end{aligned}$$

Thus, we obtain the following approximating posterior and the lower bound of the marginal likelihood $Z(\lambda)$,

$$\tilde{p}_\xi(w|x^n, \lambda) \equiv \frac{p(x^n|w)\tilde{p}_\xi(w|\lambda)}{\underline{Z}_\xi(\lambda)},$$

$$\underline{Z}_\xi(\lambda) \equiv \int p(x^n|w)\tilde{p}_\xi(w|\lambda)dw.$$

More specifically, the approximating posterior turns out to be the Gaussian distribution with the mean and variance,

$$\begin{aligned}&\mathbb{E}_{\tilde{p}_\xi(w|x^n,\lambda)}[W] = \frac{n\overline{x}}{\frac{\lambda}{\sqrt{\xi^2}} + n} \quad \text{and}\\&\mathbb{V}_{\tilde{p}_\xi(w|x^n,\lambda)}[W] = \frac{1}{\frac{\lambda}{\sqrt{\xi^2}} + n}.\end{aligned} \tag{18}$$

The lower bound of $Z(\lambda)$ is explicitly given by

$$\begin{aligned}\underline{Z}_\xi(\lambda) =& \frac{1}{2(2\pi)^{\frac{n-1}{2}}}\exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)\\&\cdot \frac{\lambda}{\sqrt{\frac{\lambda}{\sqrt{\xi^2}} + n}}\exp\left[-\frac{1}{2}\left\{\lambda\sqrt{\xi^2} - \frac{(n\overline{x})^2}{\frac{\lambda}{\sqrt{\xi^2}} + n}\right\}\right].\end{aligned}$$

To maximize $\underline{Z}_\xi(\lambda)$ with respect to the variational parameter and the regularization parameter, we can use the expectation-maximization (EM) algorithm, which updates ξ and λ so that

$$\mathbb{E}_{\tilde{p}_{\xi^{\mathrm{old}}}(w|x^n,\lambda^{\mathrm{old}})}[\log \tilde{p}_\xi(W|\lambda)]$$

is maximized for the fixed current estimates, ξ^{old} and λ^{old}, of ξ and λ [10], [12]. This update guarantees that $\underline{Z}_\xi(\lambda)$ is increased. Let ξ^{new} and λ^{new} be the updated parameters. Then, the update rules are explicitly given by

$$\begin{aligned}\sqrt{(\xi^{\mathrm{new}})^2} &= \sqrt{\mathbb{E}_{\tilde{p}_{\xi^{\mathrm{old}}}(w|x^n,\lambda^{\mathrm{old}})}[W^2]}\\&= \frac{\sqrt{\frac{\lambda^{\mathrm{old}}}{\sqrt{(\xi^{\mathrm{old}})^2}} + n + (n\overline{x})^2}}{\frac{\lambda^{\mathrm{old}}}{\sqrt{(\xi^{\mathrm{old}})^2}} + n},\end{aligned}$$

$$\lambda^{\mathrm{new}} = \frac{\frac{\lambda^{\mathrm{old}}}{\sqrt{(\xi^{\mathrm{old}})^2}} + n}{\sqrt{\frac{\lambda^{\mathrm{old}}}{\sqrt{(\xi^{\mathrm{old}})^2}} + n + (n\overline{x})^2}},$$

which follows from $\frac{\partial}{\partial\lambda}\mathbb{E}_{\tilde{p}_{\xi^{\mathrm{old}}}(w|x^n,\lambda^{\mathrm{old}})}[\tilde{p}_{\xi}(W|\lambda)] = 0$, $\frac{\partial}{\partial\sqrt{\xi^2}}\mathbb{E}_{\tilde{p}_{\xi^{\mathrm{old}}}(w|x^n,\lambda^{\mathrm{old}})}[\tilde{p}_{\xi}(W|\lambda)] = 0$, and (18). Summarizing these rules, we have

$$\lambda^{\mathrm{new}} = \frac{(\lambda^{\mathrm{old}})^2 + n}{\sqrt{(\lambda^{\mathrm{old}})^2 + n + (n\overline{x})^2}}. \tag{19}$$

The solution to this update rule of the LVA for the model (1) is analyzed in the following theorem. Here, we put $\lambda^{(t)} = \lambda^{\mathrm{old}}$ and $\lambda^{(t+1)} = \lambda^{\mathrm{new}}$ for the estimates of λ at the tth update of (19), and let $\hat{\lambda}_{\mathrm{LVA}} = \lim_{t\to\infty}\lambda^{(t)}$.

Theorem 3: For any initial value $\lambda^{(0)} \geq 0$,

$$\hat{\lambda}_{\mathrm{LVA}} = \begin{cases} \infty & \left(|\overline{x}| \leq \frac{1}{\sqrt{n}}\right), \\ \frac{1}{\sqrt{\overline{x}^2 - \frac{1}{n}}} & \left(|\overline{x}| > \frac{1}{\sqrt{n}}\right). \end{cases} \tag{20}$$

That is, $\lambda^{(t)}$ grows unboundedly as $t \to \infty$ if $|\overline{x}| \leq \frac{1}{\sqrt{n}}$ and otherwise, $\lambda^{(t)}$ converges to $1/\sqrt{\overline{x}^2 - \frac{1}{n}}$.

(Proof of Theorem 3)

By the transformations, $\Lambda^{(t)} = \frac{\lambda^{(t)}}{\sqrt{2n}}$, $M = \sqrt{\frac{n}{2}}\overline{x}$, which were used also in the proof of Theorem 1, (19) is equivalent to

$$2(\Lambda^{(t+1)})^2 = \frac{\left\{2(\Lambda^{(t)})^2 + 1\right\}^2}{2(\Lambda^{(t)})^2 + 2M^2 + 1}$$

and further to

$$A^{(t+1)} - A^{(t)} = 1 - 2M^2 + \frac{4M^2}{A^{(t)} + 2M^2 + 1} \tag{21}$$

for $A^{(t)} \equiv 2(\Lambda^{(t)})^2$.

Assume that $A^{(t)}$ is bounded by a constant $U > 0$ for $0 < |M| \leq 1/\sqrt{2}$. Then, (21) implies that

$$\begin{aligned} A^{(t+1)} - A^{(t)} &\geq \frac{4M^2}{A^{(t)} + 2M^2 + 1} \\ &\geq \frac{4M^2}{U + 2M^2 + 1} > 0, \end{aligned}$$

which in turn implies that $A^{(t)}$ grows unboundedly. This contradicts our assumption that $A^{(t)}$ is bounded. Hence, $A^{(t)}$ diverges to infinity if $0 < |M| \leq 1/\sqrt{2}$.

It also follows from (21) that if $|M| \neq 1/\sqrt{2}$,

$$\begin{aligned} &\left|A^{(t+1)} - \frac{1}{2M^2 - 1}\right| \\ &= \frac{A^{(t)} + 2}{A^{(t)} + 2M^2 + 1}\left|A^{(t)} - \frac{1}{2M^2 - 1}\right|. \end{aligned}$$

Since $\frac{A^{(t)}+2}{A^{(t)}+2M^2+1} < 1$ if and only if $|M| > 1/\sqrt{2}$,

$$\lim_{t\to\infty} A^{(t)} = \begin{cases} \infty & \left(|M| \leq \frac{1}{\sqrt{2}}\right), \\ \frac{1}{2M^2 - 1} & \left(|M| > \frac{1}{\sqrt{2}}\right). \end{cases}$$

Expressing $A^{(t)}$ and M by $\lambda^{(t)}$ and $\overline{x}$, we obtain the theorem. □

Compared to the asymptotic expansion (15) of the exact empirical Bayes estimator, the approximate solution $\hat{\lambda}_{\mathrm{LVA}}$ provides a lower bound of $\hat{\lambda}$ asymptotically because (20) yields that

$$\hat{\lambda}_{\mathrm{LVA}} \simeq \frac{1}{|\overline{x}|}\left\{1 + \frac{1}{2n\overline{x}^2} + O\left(\frac{1}{(n\overline{x}^2)^2}\right)\right\}$$

as $\sqrt{n}|\overline{x}| \to \infty$.

The next corollary evaluates the approximation accuracy of the LVA, which directly follows from (5) and (20).

Corollary 1: If $|\overline{x}| > 1/\sqrt{n}$, $\hat{\lambda}/\hat{\lambda}_{\mathrm{LVA}} < 2$.

6. Numerical Evaluation

We computed the exact solution of the empirical Bayes estimator in (3) by Newton's method for $n = 100$ (Fig. 1). We also compared it with its upper bound (5) and the solution of the LVA obtained by the update rule (19). Although we adopt $n = 100$ here, under the transformations (10), we can draw a similar figure for Λ and M with different scales of horizontal and vertical axes. This means that we can obtain the solutions for an arbitrary n only by changing the scales of axes in Fig. 1.

We see that all the solutions diverge to infinity for $|\overline{x}| \leq 1/\sqrt{n} = 10^{-1}$, and that the LVA solution provides a lower bound to the exact solution. We confirmed that the LVA solution is equal to the one proved in Theorem 3. We

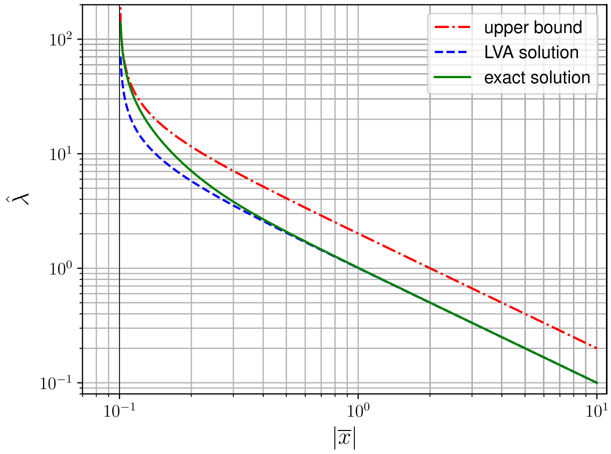

Fig. 1 The empirical Bayes estimates of λ against $|\overline{x}|$, the exact solution (solid line), its upper bound (dot-dashed line) and the solution of the LVA (dashed line).

also see that the upper bound is asymptotically tight in the limit $|\bar{x}| \to 10^{-1}+$ while the LVA solution is asymptotically tight as $|\bar{x}| \to \infty$. The latter asymptotic behavior is explained by Theorem 2 and the discussion below it. The former asymptotic behavior suggests that the accuracy of LVA becomes worse as $|\bar{x}|$ approaches $1/\sqrt{n}$ from above and we have $\hat{\lambda}_{\mathrm{LVA}} \approx 0.5\hat{\lambda}$, the worst case proved in Corollary 1, asymptotically.

7. Discussion

In this section, we compare the main result (Theorem 1) with the case of the L_2 regularization studied in a previous work [6], [7].

7.1 The Effect of ARD

If we replace the Laplace prior in (1) with the Gaussian prior,

$$p(w|\lambda) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\lambda \frac{w^2}{2}\right),$$

then, we obtain the following empirical Bayes estimator of λ,

$$\hat{\lambda}_{L_2} = \begin{cases} \infty & \left(|\bar{x}| \le \frac{1}{\sqrt{n}}\right), \\ \frac{1}{\bar{x}^2 - \frac{1}{n}} & \left(|\bar{x}| > \frac{1}{\sqrt{n}}\right), \end{cases} \tag{22}$$

and the posterior mean estimator of w,

$$\hat{w}_{L_2}(\hat{\lambda}_{L_2}) = \begin{cases} 0 & \left(|\bar{x}| \le \frac{1}{\sqrt{n}}\right), \\ \bar{x} - \frac{1}{n\bar{x}} & \left(|\bar{x}| > \frac{1}{\sqrt{n}}\right), \end{cases} \tag{23}$$

which is also the MAP estimator since the posterior distribution is Gaussian. Note that the condition, $|\bar{x}| \le 1/\sqrt{n}$, corresponds to the case where λ goes to infinity and hence w is estimated to be exactly zero by the mechanism of ARD. The main theorem shows that this condition is identical between the L_1 and L_2 regularization. This fact is suggested to be true as $n \to \infty$ by the following transformation of the model (1) to the hierarchical model discussed in [6], [7],

$$\begin{aligned} p(x|a,b) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-ab)^2}{2}\right\}, \\ p(a) &= \frac{1}{2}\exp(-|a|), \end{aligned} \tag{24}$$

where $a = \lambda w$ and $b = 1/\lambda$. This is because the effect of the prior tends to vanish as $n \to \infty$. The main theorem and (22) mean that the effect of ARD appears identically in the L_1 and L_2 regularization not only for $n \to \infty$ but also for all n.

7.2 The MAP Estimator

Under the L_1 regularization for a fixed regularization parameter λ, the MAP estimator of the parameter w in (1) is given by

$$\hat{w}_{\mathrm{MAP}}(\lambda) = \begin{cases} \bar{x} - \frac{\lambda}{n} & \left(\bar{x} > \frac{\lambda}{n}\right), \\ 0 & \left(|\bar{x}| \le \frac{\lambda}{n}\right), \\ \bar{x} + \frac{\lambda}{n} & \left(\bar{x} < -\frac{\lambda}{n}\right). \end{cases}$$

Theorem 1 shows that w is estimated to be 0 by the empirical Bayes approch (ARD) if $|\bar{x}| \le 1/\sqrt{n}$. As demonstrated in Fig. 1, the empirical Bayes solution of λ monotonically decreases from ∞ to 0 as $|\bar{x}|$ grows from $1/\sqrt{n}$. This means that there exists a unique $\bar{x} > 1/\sqrt{n}$ such that $\lambda^*/n = \bar{x}$ holds. Let this unique $\bar{x}$ be x_c. For $1/\sqrt{n} < |\bar{x}| < x_c$, ARD does not imply $w = 0$ while the MAP estimate does imply $w = 0$ if λ is set by ARD to a finite value. Therefore, the MAP estimator with its regularization parameter estimated by the empirical Bayes approach is

$$\hat{w}_{\mathrm{MAP}}(\hat{\lambda}) = \begin{cases} \bar{x} - \frac{\lambda^*}{n} & (\bar{x} > x_c), \\ 0 & (|\bar{x}| \le x_c), \\ \bar{x} + \frac{\lambda^*}{n} & (\bar{x} < -x_c). \end{cases} \tag{25}$$

In other words, the range of $|\bar{x}|$ for which $\hat{w}_{\mathrm{MAP}}(\hat{\lambda}) = 0$ is slightly extended compared to the case of the L_2 regularization in (23) because of the L_1 regularization.

Considering the intersection of $\lambda = n\bar{x}$ and the upper bound (5), the upper bound of x_c is given. Equating the upper bound to $n\bar{x}$ yields the following range of x_c,

$$\frac{1}{\sqrt{n}} < x_c < \sqrt{\frac{1+\sqrt{17}}{2n}}. \tag{26}$$

7.3 More General Case

As discussed above, the empirical Bayes solution of L_1 regularization is close to the solution of L_2 regularization if n is sufficiently large. Therefore, the behavior of the solution of L_2 regularization provides useful insight into the extension of the main result of this paper. Regarding L_2 regularization, a detailed analysis has been given in the case of the linear regression whose design matrix is identity [6], [7]. The detailed behavior of the solution in the case of a general design matrix has yet to be clarified although it can be reduced to the case of a hierarchical model as in (24) by using the pseudoinverse of the design matrix [13]. Our next task is to analyze the empirical Bayes solution of L_1 regularization whose design matrix has a special structure such as identity.

As we have proved in Lemma 2, the marginal likelihood is unimodal in the one-parameter case. The unimodality implies the convergence of iterative algorithms optimizing the hyperparameter to the global maximum. It is also an important undertaking to investigate such a property of the marginal likelihood in more complex practical cases.

8. Conclusion

We analyzed the empirical Bayes estimator of the regularization parameter in the one-parameter model of the L_1 regularization. It was shown that the condition that the empirical Bayes method yields the sparse solution is identical to the case of the L_2 regularization. We also compared the exact solution to the approximate solution given by the LVA and its accuracy was analytically evaluated.

Acknowledgments

This research was supported in part by JSPS grants 25120014, 15K16050 and 16H02825.

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," J. Royal Statistical Society, Series B, vol.58, pp.267–288, 1994.

[2] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, 1st ed., Springer, 2010.

[3] T. Hastie, R. Tibshirani, and M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, Chapman & Hall/CRC, 2015.

[4] D.J. MacKay, "Bayesian methods for backpropagation networks," in Models of Neural Networks III, E. Domany, van Hemmen, and K. Schulten, eds., pp.211–254, Physics of Neural Networks, Springer New York, 1996.

[5] M.E. Tipping, "Sparse Bayesian learning and the relevance vector machine," J. Machine Learn. Res., vol.1, pp.211–244, 2001.

[6] S. Nakajima and S. Watanabe, "Generalization error of an empirical Bayes approach," Proc. Workshop on Information-Based Induction Sciences (IBIS2004), pp.28–33, 2004 (in Japanese).

[7] S. Nakajima and S. Watanabe, "Generalization performance of subspace Bayes approach in linear neural networks," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.1128–1138, 2006.

[8] M. Girolami, "A variational method for learning sparse and overcomplete representations," Neural Comput., vol.13, no.11, pp.2517–2532, 2001.

[9] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," J. Machine Learn. Res., vol.9, pp.759–813, 2008.

[10] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.

[11] G. Grimmett and S. Stirzaker, Probability Theory and Random Processes, 3rd ed., Oxford University Press, 2001.

[12] K. Watanabe, M. Okada, and K. Ikeda, "Divergence measures and a general framework for local variational approximation," Neural Networks, vol.24, no.10, pp.1102–1109, 2011.

[13] S. Nakajima, M. Sugiyama, S.D. Babacan, and R. Tomioka, "Global analytic solution of fully-observed variational Bayesian matrix factorization," J. Machine Learn. Res., vol.14, pp.1–37, 2013.

Appendix A: Proof of Lemma 2

We can restrict ourselves to the domain $\Lambda \geq 0$ and $M \geq 0$ without loss of generality. From

$$\frac{d}{dx}\mathrm{erfcx}(x) = 2x\mathrm{erfcx}(x) - \frac{2}{\sqrt{\pi}},$$

we have

$$\begin{aligned} f'(\Lambda) &= (2\Lambda^2 - 2M\Lambda + 1)\mathrm{erfcx}(\Lambda - M) \\ &\quad + (2\Lambda^2 + 2M\Lambda + 1)\mathrm{erfcx}(\Lambda + M) - \frac{4}{\sqrt{\pi}}\Lambda. \end{aligned}$$

By using the five polynomials of Λ,

$$\begin{aligned} P_1(\Lambda) &\equiv 2\Lambda^2 + 2M\Lambda + 1, \\ P_2(\Lambda) &\equiv 2\Lambda^2 - 2M\Lambda + 1, \\ P_3(\Lambda) &\equiv 2\Lambda^4 + (1 - 2M^2)\Lambda^2 + 1, \\ P_4(\Lambda) &\equiv 2M\Lambda^3 + 2M^2\Lambda^2 - 1, \\ P_5(\Lambda) &\equiv (2M^2 - 1)\Lambda^2 - 2, \end{aligned}$$

we define the following functions and calculate their derivatives to analyze $f(\Lambda)$,

$$\begin{aligned} f'(\Lambda) &= P_1\mathrm{erfcx}(\Lambda + M) + P_2\mathrm{erfcx}(\Lambda - M) - \frac{4}{\sqrt{\pi}}\Lambda \\ &= g_1(\Lambda) \cdot g_2(\Lambda), \\ g_1(\Lambda) &\equiv \frac{2}{\sqrt{\pi}}P_1 e^{(\Lambda+M)^2}, \\ g_2(\Lambda) &\equiv \int_{\Lambda+M}^{\infty} e^{-t^2}dt + \frac{P_2}{P_1}e^{-4M\Lambda}\int_{\Lambda-M}^{\infty} e^{-t^2}dt \\ &\quad - \frac{2\Lambda}{P_1}e^{-(\Lambda+M)^2}, \\ g_2'(\Lambda) &= -\frac{8MP_3}{(P_1)^2}e^{-4M\Lambda}\int_{\Lambda-M}^{\infty} e^{-t^2}dt + \frac{4P_4}{(P_1)^2}e^{-(\Lambda+M)^2} \\ &= h_1(\Lambda) \cdot h_2(\Lambda), \\ h_1(\Lambda) &\equiv -\frac{8MP_3}{(P_1)^2}e^{-4M\Lambda}, \\ h_2(\Lambda) &\equiv \int_{\Lambda-M}^{\infty} e^{-t^2}dt - \frac{P_4}{2MP_3}e^{-(\Lambda-M)^2}, \\ h_2'(\Lambda) &= \frac{\Lambda P_1 P_5}{M(P_3)^2}e^{-(\Lambda-M)^2}, \end{aligned}$$

where we have omitted the dependency of the polynomials on Λ for notational simplicity. Since $P_1(\Lambda) > 0$, all the functions are continuous for all $\Lambda \geq 0$ and $M \geq 0$ except for h_2 and h_2', which have singularities at the zeros of $P_3(\Lambda)$.

If $0 \leq M \leq \frac{1}{\sqrt{2}}$, then we have $P_1(\Lambda) > 0$, $P_3(\Lambda) > 0$, and $P_5(\Lambda) < 0$. This means that $h_2'(\Lambda) < 0$, and thus $h_2(\Lambda)$ is monotonically decreasing. Furthermore, it follows from the continuity of $h_2(\Lambda)$ and $\lim_{\Lambda\to\infty} h_2(\Lambda) = 0$ that $h_2(\Lambda) > 0$, and also from $P_1(\Lambda) > 0$, and $P_3(\Lambda) > 0$ that $h_1(\Lambda) < 0$. These facts imply $g_2'(\Lambda) = h_1(\Lambda) \cdot h_2(\Lambda) < 0$, and hence $g_2(\Lambda)$ is monotonically decreasing. Combined with the facts that $g_2(\Lambda)$ is continuous and $\lim_{\Lambda\to\infty} g_2(\Lambda) = 0$, this proves that $g_2(\Lambda) > 0$. Since $P_1(\Lambda) > 0$, $g_1(\Lambda) > 0$. It finally follows that $f'(\Lambda) = g_1(\Lambda) \cdot g_2(\Lambda) > 0$ meaning that $f(\Lambda)$ is strictly monotonically increasing if $0 \leq M \leq \frac{1}{\sqrt{2}}$.

Next, we turn to the case of $M > \frac{1}{\sqrt{2}}$. In this case, $P_5(\Lambda)$ has a unique zero at

$$s_5 \equiv \sqrt{\frac{2}{2M^2 - 1}}.$$

We further divide the discussion into the three cases, (i) $\frac{1}{\sqrt{2}} < M < \sqrt{\frac{1}{2}+\sqrt{2}}$, (ii) $M = \sqrt{\frac{1}{2}+\sqrt{2}}$, and (iii) $M > \sqrt{\frac{1}{2}+\sqrt{2}}$.

(i) $\frac{1}{\sqrt{2}} < M < \sqrt{\frac{1}{2}+\sqrt{2}}$

In this case, it holds that $P_3(\Lambda) > 0$. Hence, the sign chart of P_5 yields that of h_2', which combined with the continuity, $\lim_{\Lambda\to\infty} h_2(\Lambda) = 0$ and the intermediate value theorem, shows that h_2 has a unique zero. Let this zero be denoted by s_h. We have $s_h < s_5$. Similarly, we can prove that g_2 has a unique zero by the fact $\lim_{\Lambda\to\infty} g_2(\Lambda) = 0$. Letting this zero of g_2 be s_g, we have $s_g < s_h < s_5$. Because f' changes its sign from plus to minus around s_g, f has the maximum at $\Lambda = s_g$ satisfying $f'(s_g) = 0$ and $s_g < s_5$, which yields the upper bound (13).

(ii) $M = \sqrt{\frac{1}{2}+\sqrt{2}}$

In this case, also P_3 has a unique zero at $s_5 = 1/\sqrt[4]{2}$. Similarly to the case (i), the sign charts of P_3 and P_5 combined with the intermediate value theorem show that h_2 has a unique zero at $\Lambda = s_h < s_5$. Then, we know that $\Lambda = s_h$ is the unique zero of g_2' with the special treatment of $\Lambda = s_5$ to prove $g_2'(s_5) > 0$ due to the singularity of h_2'. Similarly, it is proved that g_2 has a unique zero at $\Lambda = s_g$ satisfying $s_g < s_h < s_5$, which corresponds to the zero of f' and hence the maximum of f.

(iii) $M > \sqrt{\frac{1}{2}+\sqrt{2}}$

In this case, P_3 has two zeros,

$$s_3^{(1)} = \frac{1}{2}\sqrt{(2M^2-1) - \sqrt{(2M^2-1)^2-8}}, \quad \text{and}$$
$$s_3^{(2)} = \frac{1}{2}\sqrt{(2M^2-1) + \sqrt{(2M^2-1)^2-8}}.$$

For $s_3^{(1)} < \Lambda < s_3^{(2)}$, $P_3(\Lambda) < 0$. Then, we can prove that $s_3^{(1)} < s_5 < s_3^{(2)}$, and h_2 has a local minimum at $\Lambda = s_5$. By specifically proving that $h_2(s_5) > 0$, we know that h_2 has a unique zero, s_h, between 0 and $s_3^{(1)}$. Similarly to (i) and (ii), with the facts that $g_2'(s_3^{(1)}) > 0$ and $g_2'(s_3^{(2)}) > 0$, s_h turns out to be the unique zero of g_2'. The rest of the proof is same as those of the cases (i) and (ii).[†] □

Appendix B: Proof of Theorem 2

We can assume without loss of generality that $\overline{x} > 0$ by the symmetry. Under the transformations (10), we consider the limit of large M. By the formulas of the error functions (7), (8), and (9), we have

$$\operatorname{erfcx}(\Lambda - M) = 2\exp\left\{(M-\Lambda)^2\right\} + \operatorname{erfcx}(M-\Lambda).$$

Substituting this expression into (12) yields

$$\begin{aligned}&2\left(2\Lambda^2 - 2M\Lambda + 1\right)\exp\left\{(M-\Lambda)^2\right\}\\&+\left(2\Lambda^2 - 2M\Lambda + 1\right)\operatorname{erfcx}(M-\Lambda)\\&+\left(2\Lambda^2 + 2M\Lambda + 1\right)\operatorname{erfcx}(M+\Lambda)\\&-\frac{4}{\sqrt{\pi}}\Lambda = 0.\end{aligned} \qquad \text{(A· 1)}$$

It follows from (13) that

$$\Lambda^* = O\left(\frac{1}{M}\right). \qquad \text{(A· 2)}$$

This implies that $M - \Lambda$ and $M + \Lambda$ are $O(M)$ for $\Lambda = \Lambda^*$. Hence, by the asymptotic expansion of $\operatorname{erfcx}(x)$ in (14), the left hand side of (A· 1) is expressed as

$$2\left(2\Lambda^2 - 2M\Lambda + 1\right)\exp\left\{(M-\Lambda)^2\right\} + O\left(\frac{1}{M}\right).$$

This means that

$$2\Lambda^2 - 2M\Lambda + 1 + o\left(e^{-M^2}\right) = 0 \qquad \text{(A· 3)}$$

holds for $\Lambda = \Lambda^*$ since otherwise the left hand side of (A· 1) is away from zero. Let the $o\left(e^{-M^2}\right)$ term in this equation be c. Then, the solution to (A· 3) satisfying (A· 2) is

$$\begin{aligned}\Lambda &= \frac{M}{2}\left(1 - \sqrt{1 - \frac{2(1+c)}{M^2}}\right)\\&\simeq \frac{1}{2M} + \frac{1}{4M^3} + O\left(\frac{1}{M^5}\right),\end{aligned}$$

where we have used $\sqrt{1+x} \simeq 1 + \frac{x}{2} - \frac{x^2}{8} + O(x^3)$ for small x. Expressing Λ and M by λ and $\overline{x}$ yields (15). □

Tsukasa Yoshida received the B.E. degree from Toyohashi University of Technology, Japan, in 2018. He is currently a master course student at the Department of Computer Science and Engineering, Toyohashi University of Technology. His research interests include statistical machine learning theory and algorithms.

[†] From the proofs of the cases (ii) and (iii), we obtain a slightly tighter upper bound of Λ^* than (13),

$$\Lambda^* < s_3^{(1)} = \frac{1}{2}\sqrt{(2M^2-1) - \sqrt{(2M^2-1)^2-8}},$$

for $|M| \geq \sqrt{\frac{1}{2}+\sqrt{2}}$.

Kazuho Watanabe received the B.E., M.E. and Ph.D. degrees from Tokyo Institute of Technology, Japan, in 2002, 2004 and 2006, respectively. From 2007 to 2008, he was a postdoctoral fellow and a research associate at the Department of Complexity Science and Engineering, University of Tokyo. From 2009 to 2013, he was an assistant professor at the Graduate School of Information Science, Nara Institute of Science and Technology. He is currently a lecturer at the Department of Computer Science and Engineering, Toyohashi University of Technology. His research interests include statistical machine learning theory and algorithms.