Bias of Estimators and Regularization Terms

Noboru Murata

July 21, 2001

Abstract

We deal with the role of regularization terms (penalty terms) from the view point of bias of the minimum training error estimation. In the field of neural networks, for instance, regularization terms are often utilized to avoid over-fitting, however most of the time cross-validation is chosen to determine the strength of the regularization.

First we will clarify the bias of minimum training error estimation, which is caused by the nonlinearity of the learning system and depends on the size of training samples. Then taking this bias into account, we consider an appropriate size of the regularization term which is minimizing the predictive errors. The optimal size of the regularization term in this sense is calculated from the second and third order information of the loss function. When the learning system has a large number of modifiable parameters, it is computationally expensive to calculate the higher order information, thus we propose a simple method of approximating the optimal size via a generalized AIC (NIC).

1 Introduction

In this manuscript, we will discuss a role of regularization terms (penalty terms) from the view point of optimizing the minimum training error estimation. In the field of neural networks, for instance, regularization terms are often utilized to avoid over-fitting mainly, and in many practical cases, it is a difficult problem how to choose the coefficient which determines the strength of the regularization.

One of the important problem that we consider here is to clarify the inherent bias and variance of minimum training error estimators. Most of these estimators are biased estimators and unbiasedness is only guaranteed asymptotically. Taking this into account, we discuss the optimality of the size of regularization terms. One idea to determine the size is minimizing generalization error (predictive loss) in a sense of ensemble average. This notion is closely related with generalized AIC (Murata et al. [1994]).

2 Bias and Variance of Estimators

First of all, we describe the mathematical model of the problem that we discuss here.

Let us define the pointwise loss for a learning system by

$$d(z,\theta) = d(x,y,\theta),$$

which measures the performance of the learning system with parameter $\theta = (\theta^1,\ldots,\theta^m)$ when a pair of input and output $z = (x,y)$ is given. We assume that function d satisfies some regularity conditions, such as d is at least three times differentiable with respect to parameter θ. We consider the problem of finding a good parameter by using a set of examples $\{z_1, z_2, \ldots, z_n\}$, which are i.i.d. random variables. This procedure is usually referred as learning from examples. The optimal parameter is defined by

$$\theta^* = \operatorname*{argmin}_{\theta} E^Z d(Z,\theta), \tag{1}$$

where E^Z denotes expectation with respect to the distribution of random variable Z. Here the optimality is discussed in terms of minimizing expected loss under a given environment, which is represented by the distribution of input and output $P(Z)$. A natural ways of estimating the optimal parameter is adopting a minimum training error estimator defined by

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{p=1}^{n} d(z_p,\theta) \left(= \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{p=1}^{n} d(z_p,\theta) \right), \tag{2}$$

which is optimal under the empirical distribution constructed from given examples.

Using the statistical asymptotic theory (see for example, Akahira and Takeuchi (1981)), we obtain ensemble characteristics of minimum training error estimators as follows. Hereafter we use tensor-like expressions and Einstein's notation without notice, such as

$$a^i b_i = \sum_i^n a^i b_i.$$

Also we use the differential operator

$$\partial_i : f \mapsto \frac{\partial f}{\partial \theta^i}$$

for derivatives with respect to parameter $\theta = (\theta^1, \ldots, \theta^m)$.

Lemma 1 *An estimator derived by minimizing the empirical loss*

$$\begin{aligned} E(\hat{\theta}) &= \theta^* + \frac{1}{n} b + o\left(\frac{1}{n}\right), & (3) \\ Cov(\hat{\theta}^i, \hat{\theta}^j) &= \frac{1}{n} q^{ik} q^{jl} g_{kl} + o\left(\frac{1}{n}\right) & (4) \end{aligned}$$

where

$$\begin{aligned} g_{ij} &= E^Z\left(\partial_i d(Z, \theta^*) \partial_j d(Z, \theta^*)\right) & (5) \\ q_{ij} &= E^Z\left(\partial_i \partial_j d(Z, \theta^*)\right) & (6) \\ t_{ijk} &= E^Z\left(\partial_i \partial_j \partial_k d(Z, \theta^*)\right) & (7) \\ s_{ijk} &= E^Z\left(\partial_i \partial_j d(Z, \theta^*) \partial_k d(Z, \theta^*)\right), & (8) \end{aligned}$$

matrices q^{ij} and g^{ij} are the inverse of matrices q_{ij} and g_{ij} respectively, and $b = (b^1, \ldots, b^m)$ is defined by

$$b^i = q^{ij} q^{kl} \left(s_{jkl} - \frac{1}{2} q^{k'l'} t_{jkk'} g_{ll'} \right). \quad (9)$$

Here we assume that the values $g_{ij}, q_{ij}, t_{ijk}, s_{ijk}$ exist. The proof is given in Appendix B.

Note that there exists bias of the estimator

$$E(\hat{\theta}) - \theta^* = \frac{1}{n} b + o\left(\frac{1}{n}\right),$$

which depends upon the number of examples. This means that in general minimum training error estimators of this kind are not unbiased and have only asymptotic consistency. Although the shape of the error surface $E^Z d(Z, \theta)$ can be well approximated by a quadratic function in a appropriate neighborhood of the optimal parameter θ^*, higher order terms should be considered outside of this neighborhood. The leading term of bias $b = (b^1, \ldots, b^m)$ consists of the third order derivatives t_{ijk}, which represents skewness of the error surface, and the covariance of the first and second order derivatives s_{ijk}, which roughly represents fluctuation of the second order derivative of pointwise loss depending on input and output pairs Z. Therefore, intuitively speaking, the bias comes from the distortion and fluttering of the error surface.

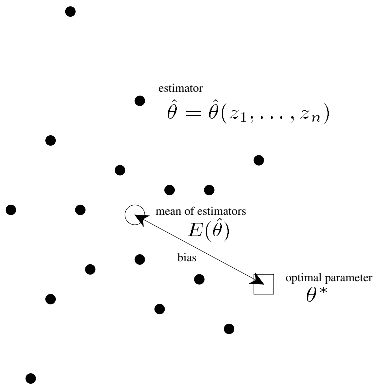

Figure 1: bias of batch learning

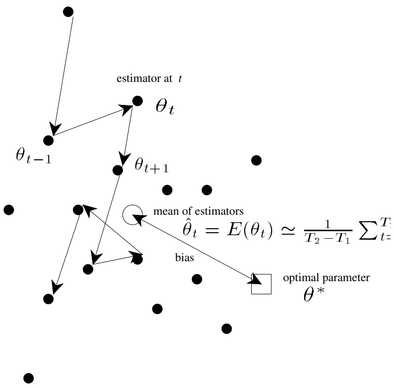

Figure 2: bias of on-line learning

3 Influence of Regularization on Estimators

Next, we consider another estimator with a regularization term. Regularization terms are introduced mainly to avoid so-called "over-fitting" problem, in which the learning system becomes too sensitive to specific examples and loses generalization capability. Hence, the regularization terms are often discussed in terms of smoothness constraints. Also other interpretations of the regularization terms are possible, such as Bayes prior, which determines confidence measure of parameter space, and penalty for model complexity reduction.

Let us consider an empirical loss with a certain regularization term $r(\theta)$

$$\sum_{p=1}^{n} d(z_p, \theta) + \alpha r(\theta), \tag{10}$$

where α determines the size of regularization. We can also adopt the definition

$$\frac{1}{n} \sum_{p=1}^{n} d(z_p, \theta) + \alpha r(\theta),$$

but for simplicity of the following discussion, we use definition (10). Note that in this way of definition, contribution of the regularization is scaled as $1/n$ automatically

$$\frac{1}{n} \sum_{p=1}^{n} d(z_p, \theta) + \frac{\alpha}{n} r(\theta).$$

The following lemma gives the discrepancy between an estimator with regularization

$$\bar{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{p=1}^{n} d(z_p, \theta) + \alpha r(\theta) \right\} \tag{11}$$

and an estimator without regularization

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{p=1}^{n} d(z_p, \theta).$$

Lemma 2 *The estimator is modified by regularization term as*

$$\bar{\theta}^i - \hat{\theta}^i = -\frac{\alpha}{n} \hat{q}^{ij} \partial_j r(\bar{\theta}) + O\left(\frac{1}{n^2}\right), \tag{12}$$

where

$$\hat{q}_{ij} = \frac{1}{n} \sum_{p=1}^{n} \partial_i \partial_j d(z_p, \hat{\theta}). \tag{13}$$

Typical examples of regularization terms are l_1-norm and l_2-norm of parameters:

$$\begin{aligned} r_{l_1}(\theta) &= \sum_i |\theta^i| \quad l_1\text{-norm}, \\ r_{l_2}(\theta) &= \sum_i |\theta^i|^2 \quad (l_2\text{-norm})^2 \end{aligned}$$

In these case, modifications for each element of the parameter are simply calculated as

$$\begin{gathered} -\frac{\alpha}{n} \sum_j \hat{q}^{ij} \operatorname{sign}(\bar{\theta}^j) + O\left(\frac{1}{n^2}\right), \\ -\frac{\alpha}{n} \sum_j \hat{q}^{ij} \bar{\theta}^j + O\left(\frac{1}{n^2}\right), \end{gathered}$$

respectively.

4 Optimizing Size of Regularization

From the above results, we know how large the bias and variance of estimators are in ensemble sense and how the modification by regularization terms behaves. One of the simplest application of this knowledge is minimizing bias of the estimator with appropriate choice of the size α, i.e.

$$\alpha_{opt} = \underset{\alpha}{\operatorname{argmin}} |E(\bar{\theta}) - \theta^*|$$

To avoid inherent bias of estimators, several methods of estimating bias are proposed, for example, Jackknife method (Tukey [1958]), Bootstrap method (Efron [1979]). We can also consider the problem of reducing the bias with regularization terms. Detailed discussion from this viewpoint is given in Appendix A.

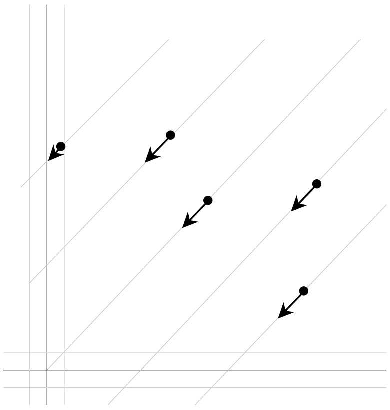

Figure 3: $l1$-norm regularization

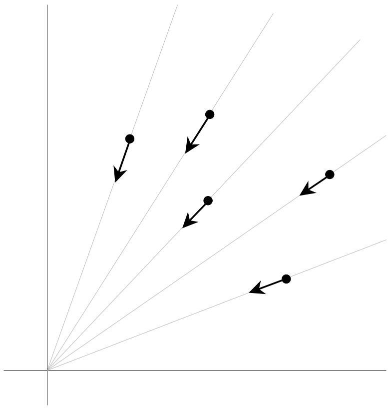

Figure 4: $l2$-norm regularization

In this paper, we consider the problem of balancing the bias and variance of the estimators in terms of minimizing generalization error.

Generalization error of the learning system with estimator $\bar{\theta}$ is defined as an averaged loss for possible inputs and outputs under a given environment $P(Z)$

$$E^Z d(Z, \bar{\theta}). \tag{14}$$

Taking the regularization term and the bias into account, the estimator is decomposed as

$$\bar{\theta}^i = \theta^{*i} + \frac{\omega^i}{\sqrt{n}} + \frac{b^i}{n} - \frac{\alpha}{n}\left(q^{ij}\partial_j r(\theta^*) + q^{ik}\partial_j\partial_k r(\theta^*)\frac{\omega^j}{\sqrt{n}} \right) + (\text{higher order term}), \tag{15}$$

where $\omega = (\omega^1, \ldots, \omega^m)$ is a random variable vector whose covariance is given by

$$Cov(\omega^i \omega^j) = q^{ik}q^{jl}g_{kl} + o\left(\frac{1}{\sqrt{n}}\right)$$

Minimizing ensemble average

$$E^{\bar{\theta}} E^Z d(Z, \bar{\theta})$$

with respect to the size of regularization α, we obtain the optimal size of α as follows. First we give the ensemble average of the generalization error.

Lemma 3

$$\begin{aligned} & E^{\bar{\theta}} E^Z d(Z, \bar{\theta}) \\ &= E^Z d(Z, \theta^*) \\ &\quad + \frac{1}{2n} q^{ij} g_{ij} \\ &\quad + \frac{1}{2n^2}\left\{ (b^i - \alpha \dot{r}^i)(b^j - \alpha \dot{r}^j) q_{ij} - 2\alpha(\ddot{r}^i_j q^{jk} g_{ki}) \right\} \\ &\quad + (\textit{higher order term}), \end{aligned} \tag{16}$$

where

$$\dot{r}^i = q^{ij}\partial_j r(\theta^*) \tag{17}$$

$$\ddot{r}^i_j = q^{ik}\partial_j\partial_k r(\theta^*). \tag{18}$$

The proof is given in Appendix D. From this relation, we can find the value of α which minimizes the term of order $O(1/n^2)$

$$(b^i - \alpha\dot{r}^i)(b^j - \alpha\dot{r}^j)q_{ij} - 2\alpha(\ddot{r}^i_j q^{jk} g_{ki}), \tag{19}$$

and obtain the following theorem.

Theorem 1 *The optimal size of the regularization in Eq.(10) which asymptotically minimizes expected generalization error is given by*

$$\alpha_{\mathrm{opt}} = \frac{b^i\partial_i r + q^{ik}q^{jl}g_{kl}\partial_i\partial_j r}{\partial_i r\partial_j r q^{ij}}, \tag{20}$$

where $\partial r = \partial r(\theta^*)$ and $\partial\partial r = \partial\partial r(\theta^*)$.

It is possible to calculate the approximation of α_{opt} by substituting the values at the optimal parameter θ^* with ones at the estimated parameter $\bar{\theta}$ and using the empirical distribution constructed from given examples. But practically it is difficult to calculate them in the case that learning systems have a large number of modifiable parameters, such as multi-layered Perceptrons, because these values include the second and third differentials of the loss function. The problem of calculating the optimal α is, however, rewritten as the problem of searching minimum of one parameter function by using notion of NIC (Murata et al. [1994]), which is a generalized AIC, as follows.

NIC is an estimator of the generalization error including regularization term

$$E^Z d(Z, \bar{\theta}) + \frac{\alpha}{n} r(\bar{\theta}), \tag{21}$$

and the original definition is given by

$$\frac{1}{n}\sum_{p=1}^{n} d(z_p, \bar{\theta}) + \frac{\alpha}{n}r(\bar{\theta}) + \frac{1}{2n}\bar{q}^{ij}\bar{g}_{ij} + \frac{1}{2n}q^{ij}g_{ij}, \tag{22}$$

where

$$\bar{g}_{ij} = \frac{1}{n}\sum_{p=1}^{n}\partial_i\left(d(z_p,\bar{\theta}) + \frac{\alpha}{n}r(\bar{\theta})\right)\partial_j\left(d(z_p,\bar{\theta}) + \frac{\alpha}{n}r(\bar{\theta})\right) \tag{23}$$

$$\bar{q}_{ij} = \frac{1}{n}\sum_{p=1}^{n}\partial_i\partial_j\left(d(z_p,\bar{\theta}) + \frac{\alpha}{n}r(\bar{\theta})\right), \tag{24}$$

and matrix $\bar{q}^{ij}$ is the inverse of matrix $\bar{q}_{ij}$. As the fourth term of Eq.(22) is not accessible because we don't know the true parameter θ^* and true distribution

$P(Z)$, matrices g and q are replaced with $\bar{g}$ and $\bar{q}$ respectively. So the practical version of the NIC is given by

$$\frac{1}{n}\sum_{p=1}^{n} d(z_p, \bar{\theta}) + \frac{\alpha}{n} r(\bar{\theta}) + \frac{1}{n}\bar{q}^{ij}\bar{g}_{ij}. \tag{25}$$

Roughly speaking, the third term of Eq.(21) is correction to obtain the realizable minimum error and the fourth term comes from the fluctuation caused by a finite number n of observations. Since we would like to find a good size α, we focus on the two terms which are directly related to the generalization error, i.e.

$$\frac{1}{n}\sum_{p=1}^{n} d(z_p, \bar{\theta}) + \frac{1}{2n}\bar{q}^{jk}\bar{g}_{jk}. \tag{26}$$

Expanding Eq.(26) around $\alpha = 0$, and minimizing it with respect to α in leading order, we obtain the following relationship.

Theorem 2 *A size of regularization α which asymptotically minimizes Eq.(26) is given by*

$$\hat{\alpha}_{\mathrm{opt}} = \frac{\hat{b}^i \partial_i \hat{r} + \hat{q}^{ik}\hat{q}^{jl}\hat{g}_{kl}\partial_i\partial_j\hat{r}}{\partial_i \hat{r}\partial_j \hat{r}\hat{q}^{ij}}. \tag{27}$$

Therefore, we can substitute minimizing generalization error with minimizing NIC by using the following procedure.

Corollary 3 *Minimizing*

$$\sum_{i=1}^{n} d(z_i, \theta) + \alpha r(\theta)$$

with respect to parameter θ:

$$\bar{\theta}(\alpha) = \min_{\theta}\sum_{i=1}^{n} d(z_i, \theta) + \alpha r(\bar{\theta})$$

and minimizing

$$\sum_{i=1}^{n} d(z_i, \bar{\theta}) + \frac{1}{2}\bar{q}^{jk}\bar{g}_{jk}$$

with respect to (hyper)parameter α:

$$\bar{\alpha} = \min_{\alpha}\sum_{i=1}^{n} d(z_i, \bar{\theta}(\alpha)) + \frac{1}{2}\bar{q}^{jk}(\alpha)\bar{g}_{jk}(\alpha)$$

simultaneously, we can obtain an approximation $\bar{\alpha}$ of the optimal size α_{opt} in terms of minimum generalization error.

In this case, we need to calculate only the second differential, i.e. the Hessian matrix of the loss function. A good approximation of the Hessian can be accessed with quasi Newton methods like BFGS algorithm, for instance.

5 Experiments

6 Conclusion

We studied the asymptotic behavior of bias and variance of estimators and discussed the role of regularization terms balancing the two. Giving a concrete equation of the optimal regularization term in the asymptotic region, we clarify not only the effect of regularization upon a specific loss function, but also the relationship between the number of examples and the scaling of regularization. Moreover, based on the NIC method, we proposed a simple way of estimating the optimal scale of regularization. An intuitive explanation for this procedure is that the models with different regularization strengths can be regarded as continuously nested models (similar to the concept of structures in Vapnik [1995]) and therefore statistical fluctuations are canceled out, which would be a problem in comparison of different structured learning machines. That is, such fluctuations are almost common for all regularization strengths, and the estimation with NIC is not suffering from them, so the leading term that we focus on here is dominating.

References

[1] Akahira, M. & Takeuchi, K. (1981), *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency.* Lecture Notes in Statistics, vol. 7, Springer-Verlag.

[2] Bishop, C.M. (1995), *Neural Networks for Pattern Recognition.* Oxford University Press.

[3] Efron, B. (1979), "Bootstrap methods: Another look at the jackknife." *Annals of Statistics*, vol. 7, pp. 1–26.

[4] Efron, B. & Stein, C. (1981), "The jackknife estimate of variance." *Annals of Statistics*, vol. 9, pp. 586–596.

[5] Firth, D. (1993), "Bias reduction of maximum likelihood estimates." *Biometrika*, vol. 80, pp. 27–38.

[6] Geman, S., Bienenstock, E. & Doursat, R. (1992) "Neural networks and the bias/variance dilemma." *Neural Computation*, vol. 4, pp. 1–58.

[7] Moody, J.E. & Rögnvaldsson, T.S. (1996) "Smoothing regularizers for projective basis function networks." OGI CSE Technical Report 96-006.

[8] Murata, N., Yoshizawa, S. & Amari, S. (1994) "Network information criterion — determining the number of hidden units for an artificial neural network model." *IEEE Trans. NN*, vol. 5, no. 6, pp. 865–872.

[9] Poggio, T. & Girosi, F. (1990) "Networks for approximation and learning." *Proc. of the IEEE*, vol. 78, no. 9 pp. 1481–1497.

[10] Rissanen, J. (1978) "Modelling by shortest data description," *Automatica*, vol. 14, pp. 465-471.

[11] Stone, M. (1974) "Cross-validatory choice and assessment of statistical predictions." *J. Roy. Stat. Soc.*, vol. 36, pp. 111–133.

[12] Tukey, J.W. (1958) "Bias and confidence in not quite large samples, abstract." *Ann. Math. Statist.*, vol. 29, pp. 614.

[13] Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Springer-Verlag.

A Correcting Bias by Projection

From the viewpoint of correcting the bias, the optimal regularization term is

$$\alpha r(\theta) = q^{ij}(\theta) g_{ij}(\theta),$$

where

$$\begin{aligned} g_{ij}(\theta) &= E^Z\left(\partial_i d(Z,\theta)\partial_j d(Z,\theta)\right) && (28)\\ q_{ij}(\theta) &= E^Z\left(\partial_i \partial_j d(Z,\theta)\right). && (29) \end{aligned}$$

As it includes averaging operation under unknown probability $P(Z)$, it is not computable. Even if we can substitute the empirical distribution for the true unknown distribution, it is quite tedious to calculate it at each learning step. A practical solution is to prepare a proper function

$$r(\theta)$$

and to choose an appropriate size α which can negate the bias as much as possible. If the modification by regularization is parallel to the bias, with appropriate α the bias can be completely canceled. Otherwise the optimal size α should be determined by projecting the bias vector to the direction caused by the regularization term. In this case, one of the reasonable metrics is q_{ij}.

Theorem 4 *The optimal size α which cancels the bias of the estimator is given by*

$$\begin{aligned} \alpha_{\mathrm{opt}} &= \frac{b^i \partial_i r}{\partial_i r \partial_j r q^{ij}} && (30)\\ &= \frac{\langle b^i, q^{jk}\partial_k r\rangle_{q_{ij}}}{\langle q^{ik}\partial_k r, q^{jl}\partial_l r\rangle_{q_{ij}}}. && (31) \end{aligned}$$

Especially in the case of maximum likelihood estimator with realizable model, this metric corresponds to Fisher information matrix.

As described in the previous section, calculation of the above optimal α is computationally expensive, we will try to resolve it into the one dimensional optimization problem like the previous case. This can be easily done by modifying a term of generalized AIC.

To put it concretely, first we define

$$\bar{g}_{ij} = \frac{1}{n}\sum_{p=1}^{n} \partial_i d(z_p,\bar{\theta})\partial_j d(z_p,\bar{\theta}) \tag{32}$$

$$\bar{q}_{ij} = \frac{1}{n}\sum_{p=1}^{n} \partial_i\partial_j d(z_p,\bar{\theta}). \tag{33}$$

Noting that $\bar{\theta}$ is a function of α, we try to minimize

$$\frac{1}{n}\sum_{p=1}^{n} d(z_p,\bar{\theta}) + \frac{1}{2n}\bar{q}^{ij}\bar{g}_{ij} \tag{34}$$

with respect to α. (There are several alternatives, but we have chosen one of the simplest.) This can be seen as minimizing generalized AIC without the regularization term at the estimated parameter with the regularization term.

Similar to the calculation of the previous section, the optimal α is determined by

$$\frac{1}{2}\left(\frac{\alpha}{n}\right)^2 \hat{q}^{ij}\partial_i\hat{r}\partial_j\hat{r} - \frac{1}{n}\cdot\frac{\alpha}{n}\hat{b}^i\partial_i\hat{r}.$$

Note that there is no term including the second differential of the regularization term. Using this α, the bias is asymptotically corrected by

$$\frac{1}{n}\cdot\frac{\langle \hat{b}^j, \hat{q}^{kl}\partial_l\hat{r}\rangle_{\hat{q}_{jk}}}{\langle \hat{q}^{jl}\partial_l\hat{r}, \hat{q}^{km}\partial_m\hat{r}\rangle_{\hat{q}_{jk}}}\cdot\hat{q}^{jk}\partial_k\hat{r}. \tag{35}$$

This correction is a projection of the bias on the direction of modification by the regularization term with metric $\hat{q}_{jk}$. In other word, a part of the bias which has the same direction of the regularization is corrected as much as possible.

The idea correcting the bias by using penalty terms is firstly proposed by Firth.

B Proof of Lemma 1

Let us define ω as

$$\hat{\theta} - \theta^* = \frac{\omega}{\sqrt{n}} \tag{36}$$

and expand the derivative of the empirical loss

$$\sum_{p=1}^{n} \partial_i d(z_p,\hat{\theta}) = 0$$

around the optimal parameter:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\partial_i d(z_p,\theta^*)\\
+&\left\{\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\partial_i\partial_j d(z_p,\theta^*)\right\}\frac{\omega^j}{\sqrt{n}}\\
+&\left\{\frac{1}{2\sqrt{n}}\sum_{p=1}^{n}\partial_i\partial_j\partial_k d(z_p,\theta^*)\right\}\frac{\omega^j}{\sqrt{n}}\frac{\omega^k}{\sqrt{n}}\\
+&\ \text{(higher order term)} = 0.
\end{aligned}$$

With values we introduced previously

$$\begin{aligned}
g_{ij} &= E^Z\left(\partial_i d(Z,\theta^*)\partial_j d(Z,\theta^*)\right)\\
q_{ij} &= E^Z\left(\partial_i\partial_j d(Z,\theta^*)\right)\\
t_{ijk} &= E^Z\left(\partial_i\partial_j\partial_k d(Z,\theta^*)\right)\\
s_{ijk} &= E^Z\left(\partial_i\partial_j d(Z,\theta^*)\partial_k d(Z,\theta^*)\right),
\end{aligned}$$

we can rewrite the above equation:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\partial_i d(z_p,\theta^*)\\
+&\left\{\frac{1}{n}\sum_{p=1}^{n}\partial_i\partial_j d(z_p,\theta^*) - q_{ij} + q_{ij}\right\}\omega^j\\
+&\frac{1}{2\sqrt{n}}\left\{\frac{1}{n}\sum_{p=1}^{n}\partial_i\partial_j\partial_k d(z_p,\theta^*) - t_{jik} + t_{ijk}\right\}\omega^j\omega^k\\
+&\ \text{(higher order term)} = 0,
\end{aligned}$$

and solve the equation for ω, then we obtain the expression

$$\begin{aligned}
\omega^i = &- q^{ij}\left\{\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\partial_j d(z_p,\theta^*)\right\}\\
&- q^{ij}\left\{\frac{1}{n}\sum_{p=1}^{n}\partial_j\partial_k d(z_p,\theta^*) - q_{jk}\right\}\omega^k\\
&- q^{ij}\frac{1}{2\sqrt{n}}\left\{\frac{1}{n}\sum_{p=1}^{n}\partial_j\partial_k\partial_l d(z_p,\theta^*) - t_{jkl} + t_{jkl}\right\}\omega^k\omega^l\\
&+ \text{(higher order term)}.
\end{aligned} \tag{37}$$

Knowing that

$$E^{Z_1,\ldots,Z_n}\left(\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\partial_i d(Z_p,\theta^*)\right) = 0$$

$$E^{Z_1,\ldots,Z_n}\left(\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\partial_i d(Z_p,\theta^*)\frac{1}{\sqrt{n}}\sum_{p'=1}^{n}\partial_j d(Z_{p'},\theta^*)\right)=g_{ij}$$

$$E^{Z_1,\ldots,Z_n}\left(\frac{1}{\sqrt{n}}\sum_{p=1}^{n}\partial_i d(Z_p,\theta^*)\frac{1}{\sqrt{n}}\sum_{p'=1}^{n}\partial_j d(Z_{p'},\theta^*)\frac{1}{\sqrt{n}}\sum_{p''=1}^{n}\partial_k d(Z_{p''},\theta^*)\right)$$
$$=O\left(\frac{1}{\sqrt{n}}\right),$$

where $E^{Z_1,\ldots,Z_n}$ denotes expectation with respect to i.i.d. random variables $Z_1,\ldots,Z_n$, and that

$$\left\{\frac{1}{n}\sum_{p=1}^{n}\partial_j\partial_k d(z_p,\theta^*)-q_{jk}\right\},\quad \left\{\frac{1}{n}\sum_{p=1}^{n}\partial_j\partial_k\partial_l d(z_p,\theta^*)-t_{jkl}\right\}$$

are order $O_p(1/\sqrt{n})$, and using Eq.(37) repeatedly, the average of ω can be calculated as

$$\begin{aligned}
E(\omega^i) &= \frac{1}{\sqrt{n}}q^{ij}q^{kl}E\left(\frac{1}{n}\sum_{p=1}^{n}\partial_j\partial_k d(Z_p,\theta^*)\sum_{p'=1}^{n}\partial_l d(Z_{p'},\theta^*)\right)\\
&\quad -\frac{1}{2\sqrt{n}}q^{ij}q^{kl}q^{k'l'}t_{jkk'}E\left(\frac{1}{n}\sum_{p=1}^{n}\partial_l d(Z_p,\theta^*)\sum_{p'=1}^{n}\partial_{l'} d(Z_{p'},\theta^*)\right)\\
&\quad +o\left(\frac{1}{\sqrt{n}}\right)\\
&= \frac{1}{\sqrt{n}}\left(q^{ij}q^{kl}s_{jkl}-\frac{1}{2}q^{ij}q^{kl}q^{k'l'}t_{jkk'}g_{ll'}\right)+o\left(\frac{1}{\sqrt{n}}\right) \qquad (38)\\
&= \frac{1}{\sqrt{n}}b^i+o\left(\frac{1}{\sqrt{n}}\right). \qquad (39)
\end{aligned}$$

Therefore

$$E(\hat{\theta}^i-\theta^{*i})=E\left(\frac{\omega^i}{\sqrt{n}}\right)=\frac{1}{n}b^i+o\left(\frac{1}{\sqrt{n}}\right). \qquad (40)$$

The covariance of the estimator is given by

$$\begin{aligned}
Cov(\hat{\theta}^i,\hat{\theta}^j) &= E\left(\hat{\theta}^i-\theta^{*i}-\frac{1}{n}b^i+o\left(\frac{1}{n}\right)\right)\left(\hat{\theta}^j-\theta^{*j}-\frac{1}{n}b^j+o\left(\frac{1}{n}\right)\right)\\
&= \frac{1}{n}q^{ik}q^{jl}g_{kl}+o\left(\frac{1}{n}\right) \qquad (41)
\end{aligned}$$

from ordinal asymptotic theory. ∎

C Proof of Lemma 2

Let

$$\zeta = \bar{\theta} - \hat{\theta}.$$

be difference between two estimators. Taking account of

$$\frac{1}{n}\sum_{p=1}^{n} \partial_i d(z_p, \bar{\theta}) + \frac{\alpha}{n}\partial_i r(\bar{\theta}) = 0$$

and

$$\frac{1}{n}\sum_{p=1}^{n} \partial_i d(z_p, \hat{\theta}) = 0,$$

we can expand the empirical loss with the regularization term as follows:

$$\begin{aligned}
&\frac{1}{n}\sum_{p=1}^{n} \partial_i d(z_p, \hat{\theta}+\zeta) + \frac{\alpha}{n}\partial_i r(\bar{\theta}) \\
&= \frac{1}{n}\sum_{p=1}^{n} \partial_i d(z_p, \hat{\theta}) + \frac{1}{n}\sum_{p=1}^{n} \partial_i \partial_j d(z_p, \hat{\theta})\zeta^j + \frac{\alpha}{n}\partial_i r(\bar{\theta}) + \text{(higher order term)} \\
&= \frac{1}{n}\sum_{p=1}^{n} \partial_i \partial_j d(z_p, \hat{\theta})\zeta^j + \frac{\alpha}{n}\partial_i r(\bar{\theta}) + \text{(higher order term)} \\
&= 0.
\end{aligned}$$

Here we define the following values

$$\hat{g}_{ij} = \frac{1}{n}\sum_{p=1}^{n} \partial_i d(z_p, \hat{\theta})\partial_j d(z_p, \hat{\theta}) \tag{42}$$

$$\hat{q}_{ij} = \frac{1}{n}\sum_{p=1}^{n} \partial_i \partial_j d(z_p, \hat{\theta}) \tag{43}$$

$$\hat{t}_{ijk} = \frac{1}{n}\sum_{p=1}^{n} \partial_i \partial_j \partial_k d(z_p, \hat{\theta}) \tag{44}$$

$$\hat{s}_{ijk} = \frac{1}{n}\sum_{p=1}^{n} \partial_i \partial_j d(z_p, \hat{\theta})\partial_k d(z_p, \hat{\theta}), \tag{45}$$

which are similar to $g_{ij}, q_{ij}, t_{ijk}, s_{ijk}$, but are defined at parameter $\hat{\theta}$ and averaged with respect to the empirical distribution based on given examples. Then we obtain

$$\begin{aligned}
\zeta^i &= -\frac{\alpha}{n}\hat{q}^{ij}\partial_j r(\bar{\theta}) + O\left(\frac{1}{n^2}\right) \\
&= -\frac{\alpha}{n}\hat{q}^{ij}\partial_j \hat{r} + O\left(\frac{1}{n^2}\right),
\end{aligned} \tag{46}$$

where $\hat{r} = r(\hat{\theta})$. ∎

D Proof of Lemma 3

the estimator is decomposed as

$$\bar{\theta}^i = \theta^{*i} + \frac{\omega^i}{\sqrt{n}} + \frac{b^i}{n} - \frac{\alpha}{n}\left(\dot{r}^i + \ddot{r}^i_j \frac{\omega^j}{\sqrt{n}}\right) + (\text{higher order term}), \tag{47}$$

where $\omega = (\omega^1, \ldots, \omega^m)$ is a random variable vector which is subject to normal distribution

$$N(0, \Sigma), \quad \Sigma = (q^{ik} q^{jl} g_{kl})$$

and

$$\begin{aligned}
\dot{r}^i &= q^{ij} \partial_j r(\theta^*) \\
\ddot{r}^i_j &= q^{ik} \partial_j \partial_k r(\theta^*).
\end{aligned}$$

Then the ensemble average is given by

$$\begin{aligned}
&E^{\bar{\theta}} E^Z d(Z, \bar{\theta}) \\
&\quad = E^Z d(Z, \theta^*) \\
&\qquad + E^Z \partial_i \partial_j d(Z, \theta^*) \\
&\qquad \quad \times E^{\bar{\theta}} \left(\frac{\omega^k}{\sqrt{n}} \left(\delta^i_k - \frac{\alpha}{n} \ddot{r}^i_k \right) + \frac{1}{n}(b^i - \alpha \dot{r}^i) \right) \left(\frac{\omega^l}{\sqrt{n}} \left(\delta^j_l - \frac{\alpha}{n} \ddot{r}^j_l \right) + \frac{1}{n}(b^j - \alpha \dot{r}^j) \right) \\
&\qquad + (\text{higher order term}) \\
&\quad = E^Z d(Z, \theta^*) \\
&\qquad + \frac{1}{n} \left(\delta^i_k - \frac{\alpha}{n} \ddot{r}^i_k \right) \left(\delta^j_l - \frac{\alpha}{n} \ddot{r}^j_l \right) q^{kk'} q^{ll'} g_{k'l'} q_{ij} \\
&\qquad + \frac{1}{n^2}(b^i - \alpha \dot{r}^i)(b^j - \alpha \dot{r}^j) q_{ij} \\
&\qquad + (\text{higher order term}),
\end{aligned} \tag{48}$$

where the equality

$$E^Z \partial_i d(Z, \theta^*) = 0$$

is used.

E Proof of Theorem 2

The first term becomes

$$\frac{1}{n} \sum_{p=1}^{n} d(z_p, \hat{\theta} + \zeta)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{p=1}^{n} d(z_p,\hat{\theta}) + \frac{1}{2}\left(\frac{\alpha}{n}\right)^2 \hat{q}_{ij}\hat{q}^{ik}\hat{q}^{jl}\partial_k\hat{r}\partial_l\hat{r} + O\left(\frac{1}{n^3}\right) \\
&= \frac{1}{n}\sum_{p=1}^{n} d(z_p,\hat{\theta}) + \frac{1}{2}\left(\frac{\alpha}{n}\right)^2 \hat{q}^{ij}\partial_i\hat{r}\partial_j\hat{r} + O\left(\frac{1}{n^3}\right). \qquad (49)
\end{aligned}
$$

Noting that

$$
\begin{aligned}
\bar{q}_{ij} &= \frac{1}{n}\sum_{p=1}^{n}\partial_i\partial_j d(z_p,\hat{\theta}+\zeta) + \frac{\alpha}{n}\partial_i\partial_j r(\hat{\theta}+\zeta) \\
&= \frac{1}{n}\sum_{p=1}^{n}\partial_i\partial_j d(z_p,\hat{\theta}) + \frac{1}{n}\sum_{p=1}^{n}\partial_i\partial_j\partial_k d(z_p,\hat{\theta})\zeta^k + \frac{\alpha}{n}\partial_i\partial_j r(\hat{\theta}) + \frac{\alpha}{n}\partial_i\partial_j\partial_k r(\hat{\theta})\zeta^k \\
&\quad +(\text{higher order term}) \\
&= \hat{q}_{ij} - \frac{\alpha}{n}\left(\hat{t}_{ijk}\hat{q}^{kl}\partial_l r(\hat{\theta}) - \partial_i\partial_j\hat{r}\right) + O\left(\frac{1}{n^2}\right) \\
\bar{g}_{ij} &= \frac{1}{n}\sum_{p=1}^{n}\left(\partial_i d(z_p,\hat{\theta}+\zeta) + \frac{\alpha}{n}\partial_i r(\hat{\theta}+\zeta)\right)\left(\partial_j d(z_p,\hat{\theta}+\zeta) + \frac{\alpha}{n}\partial_j r(\hat{\theta}+\zeta)\right) \\
&= \frac{1}{n}\sum_{p=1}^{n}\partial_i d(z_p,\hat{\theta})\partial_j d(z_p,\hat{\theta}) \\
&\quad +\frac{1}{n}\sum_{p=1}^{n}\partial_i\partial_k d(z_p,\hat{\theta})\zeta^k\partial_j d(z_p,\hat{\theta}) + \frac{1}{n}\sum_{p=1}^{n}\partial_i d(z_p,\hat{\theta})\partial_j\partial_k d(z_p,\hat{\theta})\zeta^k \\
&\quad +\frac{1}{n}\sum_{p=1}^{n}\partial_i d(z_p,\hat{\theta})\frac{\alpha}{n}\partial_j r(\hat{\theta}) + \frac{1}{n}\sum_{p=1}^{n}\partial_j d(z_p,\hat{\theta})\frac{\alpha}{n}\partial_i r(\hat{\theta}) \\
&\quad +(\text{higher order term}) \\
&= \hat{g}_{ij} - \frac{\alpha}{n}\left(\hat{s}_{kij} + \hat{s}_{kji}\right)\hat{q}^{kl}\partial_l\hat{r} + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

and that

$$
\partial q^{ij} = -q^{ik}q^{lj}\partial q_{kl}
$$

holds for the derivative of the inverse matrix, the second term becomes

$$
\begin{aligned}
&\frac{1}{2n}\hat{q}^{ij}\hat{g}_{ij} \\
&\quad -\frac{1}{2n}\cdot\frac{\alpha}{n}\hat{q}^{ij}\partial_i\hat{r}\left(\left(\hat{s}_{jkl}+\hat{s}_{jlk}\right)\hat{q}^{kl} - \hat{t}_{jkl}\hat{q}^{kk'}\hat{q}^{ll'}\hat{g}_{k'l'}\right) \\
&\quad -\frac{1}{2n}\cdot\frac{\alpha}{n}\hat{q}^{ik}\hat{q}^{jl}\hat{g}_{kl}\partial_i\partial_j\hat{r} \\
&\quad +O\left(\frac{1}{n^3}\right) \\
&= \frac{1}{2n}\hat{q}^{ij}\hat{g}_{ij} - \frac{1}{n}\cdot\frac{\alpha}{n}\left(\hat{b}^i\partial_i\hat{r} + \hat{q}^{ik}\hat{q}^{jl}\hat{g}_{kl}\partial_i\partial_j\hat{r}\right) + O\left(\frac{1}{n^3}\right), \qquad (50)
\end{aligned}
$$

where $\hat{b}$ is defined similar to the definition of bias b by substituting $\hat{g}_{jk}, \hat{q}_{jk}, \hat{t}_{jkl}, \hat{s}_{jkl}$ for $g_{jk}, q_{jk}, t_{jkl}, s_{jkl}$ respectively. Neglecting higher order terms and finding the value of α that minimizes the quadratic form

$$\frac{1}{2}\left(\frac{\alpha}{n}\right)^2 \hat{q}^{ij}\partial_i\hat{r}\partial_j\hat{r} - \frac{1}{n}\cdot\frac{\alpha}{n}\left(\hat{b}^i\partial_i\hat{r} + \hat{q}^{ik}\hat{q}^{jl}\hat{g}_{kl}\partial_i\partial_i\hat{r}\right),$$

we obtain

$$\hat{\alpha}_{\mathrm{opt}} = \frac{\hat{b}^i\partial_i\hat{r} + \hat{q}^{ik}\hat{q}^{jl}\hat{g}_{kl}\partial_i\partial_j\hat{r}}{\partial_i\hat{r}\partial_j\hat{r}\hat{q}^{ij}} \tag{51}$$

$$= \frac{\langle\hat{b}^i, \hat{q}^{jk}\partial_k\hat{r}\rangle_{\hat{q}_{ij}} + \hat{q}^{ik}\hat{q}^{jl}\hat{g}_{kl}\partial_i\partial_j\hat{r}}{\langle\hat{q}^{ik}\partial_k\hat{r}, \hat{q}^{jl}\partial_l\hat{r}\rangle_{\hat{q}_{ij}}}, \tag{52}$$

where $\langle\cdot\rangle_{\hat{q}_{ij}}$ is inner product with metric $\hat{q}_{ij}$.