Article

Information-Corrected Estimation: A Generalization Error Reducing Parameter Estimation Method

Matthew Dixon [1,†] and Tyler Ward [2,*,†]

[1] Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA; matthew.dixon@iit.edu

[2] Department of Financial Engineering, NYU Tandon School of Engineering, New York, NY 11201, USA

* Correspondence: tw623@nyu.edu

† These authors contributed equally to this work.

Abstract: Modern computational models in supervised machine learning are often highly parameterized universal approximators. As such, the value of the parameters is unimportant, and only the out of sample performance is considered. On the other hand much of the literature on model estimation assumes that the parameters themselves have intrinsic value, and thus is concerned with bias and variance of parameter estimates, which may not have any simple relationship to out of sample model performance. Therefore, within supervised machine learning, heavy use is made of ridge regression (i.e., L2 regularization), which requires the the estimation of hyperparameters and can be rendered ineffective by certain model parameterizations. We introduce an objective function which we refer to as Information-Corrected Estimation (ICE) that reduces KL divergence based generalization error for supervised machine learning. ICE attempts to directly maximize a corrected likelihood function as an estimator of the KL divergence. Such an approach is proven, theoretically, to be effective for a wide class of models, with only mild regularity restrictions. Under finite sample sizes, this corrected estimation procedure is shown experimentally to lead to significant reduction in generalization error compared to maximum likelihood estimation and L2 regularization.

Keywords: generalization error; overfitting; information criteria; entropy

1. Introduction

Kullback and Leibler [1] showed that minimizing a divergence $\rho_{KL}(f, g_\theta)$ between the truth, f, and a parametric model density, g_θ, is necessary and sufficient for making accurate predictions about data using the model defined by θ. Recent work [2] on Berk–Nash equilibria has shown the central role that KL divergence plays in game theoretic choice models such as multi-armed bandits and stochastic multi-party games. KL divergence thus plays a leading role in machine learning and neuroscience, with several inferential approaches developed in the information theory literature. Such approaches for minimizing KL divergence employ a range of methods, including data partitioning, Bayesian indirect inference and M-estimation [3–5]. These approaches are quite distinct from the standard penalized loss minimization framework and, as such, are non-trivial to combine with supervised learning methods such as neural networks.

It is well known that maximum likelihood estimation (MLE) introduces an asymptotic bias in the KL divergence minimizer which is problematic for both model estimation and model selection. For many models, where the parameters θ are themselves important, this may be investigated as parameter bias and parameter variance. However, for models common in modern machine learning, the parameters themselves do not have any easily interpreted meaning. For these models, the parameters themselves are irrelevant and only the accuracy (in terms of KL divergence) of the model predictions matter. Within the information theory literature, this has often been referred to simply as bias (e.g., $b(G)$ from [6]). To distinguish it from parameter bias, one might refer to it as "prediction bias"

or "generalization error". Generalization error is the more common terminology (see, for example, Equation 1.1.6 [7]) and will be used here.

Before the widespread use of machine learning, most models had interpretable parameters, and thus there is a large literature focused on reducing parameter bias. For instance, the jackknife [8] (leave-one-out cross-validation) estimator is an early example. More relevant to this paper is the approach of Firth [9] and later Kosmidis [10,11]. More recently, Pagui, Salvan, and Sartori [12] proposed a parameter bias reducing estimation methodology. An extensive review of the literature around this point can be found in [13]. Unfortunately, these approaches do not consider the impact on KL divergence-based generalization error and thus are not applicable to the field of machine learning where the parameters themselves are devoid of meaning. Heskes [14] shows that classifiers do have a notion of bias-variance decomposition for generalization error, but it is not computable from parameter bias and parameter variance. Therefore, parameter bias reducing formulations are not useful within machine learning unless it can be shown that they also reduce generalization error.

In fact, to seat the approach taken in this paper to generalization error, we recall much earlier and seminal work at the intersection of statistics and information theory. Akaike [15], and later Takeuchi [16], proposed information criteria (AIC and TIC, respectively) for model selection designed explicitly to reduce generalization error. Konishi and Kitagawa [6] extended the approach of Takeuchi to cases where MLE was not used to fit the underlying model, but still restricted themselves to the question of model selection. Stone [17] proved that Akaike's Information Criterion (AIC) is asymptotically equivalent to jackknifing when the estimator is finite. Takeuchi himself showed that TIC is an extension of AIC with fewer restrictions, and thus it too is equivalent to jackknifing whenever AIC would be valid.

For highly parameterized models, as are common in machine learning, model selection such as this is of limited utility. The parameter count may necessarily be very large, and thus none of the models fit using MLE may be acceptable. Then, merely choosing among them is unlikely to produce acceptable results. Within this field, typically L_2 or similar regularization is used to reduce generalization error. See Section 11.5.2 [18], for a typical example. For a more recent innovation, refer to [19]. Note that regularization schemes such as this often increase parameter bias while decreasing generalization error. Golub, Heath, and Wahba [20] showed that L_2 regularization is asymptotically equivalent to cross-validation for linear models, subject to certain assumptions. For nonlinear models, it has long been known that L_2 regularization is not always valid, and it is trivial to construct example models (See Section 4.1 for one such example) where this approach is always harmful in expectation.

Therefore, it is important to develop a method to reduce generalization error in model estimation analogous to the way that L_2 regularization would commonly be used for a highly parameterized model, but having applicability for a wider family of models, especially those for which L_2 regularization is not applicable. It is not the goal of this paper to perform a wide survey of generalization error reducing approaches, but we will rather propose an additional approach, investigate its properties, and show that it has superior performance when compared against L_2 regularization, which is currently the dominant generalization error reducing estimation procedure within the field of machine learning.

To this end, this paper introduces a generalization error reducing estimation approach referred to as Information Corrected Estimation (ICE). This estimator is proven to have a generalization error of only $O(n^{-\frac{3}{2}})$ instead of $O(n^{-1})$ as is the case for MLE, and is shown to be valid within a neighborhood around the MLE parameter estimate. Optimizing over this ICE objective function instead of the negative log likelihood thus produces parameters with superior out of sample performance.

Takeuchi's TIC and Firth's approach have never seen widespread use due to the computational and numerical issues that arise from the computation of this adjustment [21], and the ICE estimator in its raw form would have similar problems. Therefore, this paper also proposes an efficient approximation of this correction term, and shows through

numerical experiments that the approximation is effective at improving model performance across a range of models.

2. Preliminaries

Let us assume that we have data $x_n := \{x_1, \ldots, x_n\}$ generated from an unknown joint density function $f(x)$ of $X_n := \{X_1, \ldots, X_n\}$. Where necessary, we define Z_n to denote a second sample drawn from $f(x)$, independent of X_n, and x_n' is the observed realization of Z_n. We consider a model $\mathcal{M}_p$ given by a parametric family of densities $\mathcal{M}_p := \{g(\cdot|\theta) \mid \theta \in \Theta \subseteq \mathbb{R}^p\}$, for some compact Euclidean parameter space Θ, which is misspecified and hence excludes the truth f. Henceforth, the distribution over x identified by θ may be referred to as $g_\theta(x) := g(x|\theta)$ where it is notationally convenient to do so.

Suppose that θ_0 is the quasi-true parameter of model $\mathcal{M}$, and $\hat{\theta}(X_n)$ is the random variable representing the MLE of θ_0 fit on a dataset, x_n. The negative log-likelihood of X_n under the distribution g_θ is

$$-\ell(\theta, X_n) := -\frac{1}{n}\sum_{i=1}^{n} \log g_\theta(x_i), \qquad (1)$$

where $-\ell(\theta, X_n)$ is written including a $\frac{1}{n}$ to make the expectation of this quantity $O(1)$ and asymptotically independent of n. Similarly, the minus sign is incorporated because $-\ell(\theta, X_n)$ is a strictly non-negative quantity if $g_\theta(x_i)$ is a probability. The MLE, $\hat{\theta}(X_n)$, minimizes the negative log likelihood of the data set with respect to the model:

$$\hat{\theta}(x_n) := \underset{\theta}{\text{argmin}}[-\ell(\theta, x_n)]. \qquad (2)$$

The expectation of $-\ell(\theta, X_n)$ is the cross entropy between f and g_θ:

$$-\mathcal{L}(\theta) := \mathbb{E}_{X_n}[-\ell(\theta, X_n)]. \qquad (3)$$

Here, the expectation is a function only of θ and of the distribution f that generated the data X_n. As a function of the distribution f, this value is $O(1)$, but could be large for poorly conditioned f. The quasi-true parameter θ_0 is

$$\theta_0 := \underset{\theta}{\text{argmin}}[-\mathcal{L}(\theta)]. \qquad (4)$$

Generalization Error in KL Divergence Based Loss Functions

Kullback and Leibler [1] viewed "information" as discriminating the sample data drawn from one distribution against another, and defined the KL-divergence ρ_{KL} between distributions in terms of the ability to make predictions about one by knowing the other. Here,

$$\rho_{KL}(f, g_\theta) = \int \log\left[\frac{f(x)}{g_\theta(x)}\right] f(x) dx. \qquad (5)$$

This value is in general unknowable, but given a sample X_n from f, $-\ell(\theta, X_n)$ will converge asymptotically to $\rho_{KL}(f, g_\theta)$ plus an additive constant that depends only on f. The convergence relies on White's regularity conditions [22].

A well known result by Stone [17] shows that the MLE is a biased estimator of the minimum KL-divergence:

$$\mathbb{E}_{X_n}[-\ell(\hat{\theta}(X_n), X_n)] < \mathbb{E}_{X_n}[-\ell(\theta_0, X_n)], \qquad (6)$$

because it is evaluated on the data X_n which was used to fit $\hat{\theta}$. Cross-validation was developed as a model selection technique to select a model from a group that actually minimizes $\mathbb{E}_{X_n}[\rho_{KL}(g_{\theta_0}, g_{\hat{\theta}(X_n)})]$ and not merely $\mathbb{E}_{X_n}[-\ell(\hat{\theta}(X_n), X_n)]$ in the limit of large n. Takeuchi [16] and Akaike [15] explicitly modeled this bias (generalization error) of an estimation procedure $\theta(X_n)$ as

$$b := \mathbb{E}_{X_n}\Big[\ell(\boldsymbol{\theta}(X_n), X_n) - \mathbb{E}_{X'_n}[\ell(\boldsymbol{\theta}(X_n), X'_n)]\Big]. \tag{7}$$

Our goal is to obtain an estimate, b^*, of the generalization error b without using the MLE. We will then add this term to the objective function to develop the estimator $\boldsymbol{\theta}^*(X_n)$ so as to cancel the lower order terms of the generalization error. This estimator will then minimize $\mathbb{E}_{X_n}[\rho_{KL}(g_{\boldsymbol{\theta}_0}, g_{\boldsymbol{\theta}^*(X_n)})]$ more effectively than MLE, and potentially would in turn produce improved predictions from the model fitted over finite training sets.

Remark 1. *We note that under MLE, $b = O(\frac{1}{n})$ [16]. Equivalently, one could say that a particular realization of the generalization error $\ell(\boldsymbol{\theta}(X_n), X_n) - \mathbb{E}_{X'_n}[\ell(\boldsymbol{\theta}(X_n), X'_n)]$ is itself $O_p(\frac{1}{n})$. Here, $O_p(\frac{1}{n})$ is used to indicate that the quantity is a random variable with finite variance, whose mean is $O(\frac{1}{n})$.*

3. Information Corrected Estimation (ICE)

We propose the following penalized likelihood function:

Definition 1 (ICE Objective).

$$-\ell^*(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \frac{1}{n}tr(I_{\boldsymbol{\theta}}J_{\boldsymbol{\theta}}^{-1}), \tag{8}$$

where $J_{\boldsymbol{\theta}}$ is the negative expected Hessian

$$J_{\boldsymbol{\theta}} := -\mathbb{E}_X[\partial_{\boldsymbol{\theta}}^2 \log g(X|\boldsymbol{\theta})] = -\int f(x)\partial_{\boldsymbol{\theta}}^2 \log g(x|\boldsymbol{\theta})dx, \tag{9}$$

and $I_{\boldsymbol{\theta}}$ is the Fisher Information matrix

$$I_{\boldsymbol{\theta}} := \mathbb{E}_X[\partial_{\boldsymbol{\theta}} \log g(X|\boldsymbol{\theta})\partial_{\boldsymbol{\theta}^T} \log g(X|\boldsymbol{\theta})]. \tag{10}$$

with $\hat{I}_{\boldsymbol{\theta}}$, $\hat{J}_{\boldsymbol{\theta}}$ being their estimates over the data.
Let $\boldsymbol{\theta}^*$ denote the minimizer of (8).

The trace term in Equation (8) will be familiar from Takeuchi [16]. However, Takeuchi showed only that this was the leading order of the bias for the MLE estimate $\hat{\boldsymbol{\theta}}$, and therefore the proof found there is not sufficient to justify a new estimator that will itself be the target of optimization, and is required to be valid away from $\hat{\boldsymbol{\theta}}$. As in Takeuchi, because I and J are unknowable, we will substitute their approximations computed from the training data, $\hat{I}_{\boldsymbol{\theta}}$ and $\hat{J}_{\boldsymbol{\theta}}$ during the actual computation of this objective. The numerical impact of this approximation will be examined in Section 4.2.1.

Remark 2. *Though AIC was developed before TIC, it is easily reproduced as a special case of TIC. Subject to certain conditions (guaranteed by the requirements of [15]), at least in expectation, $I_{\hat{\boldsymbol{\theta}}} = J_{\hat{\boldsymbol{\theta}}}$. Thus, the quantity within the TIC trace term, $I_{\hat{\boldsymbol{\theta}}}J_{\hat{\boldsymbol{\theta}}}^{-1}$, is the identity matrix. Therefore, its trace is equal to p, the parameter count of the model, recovering AIC. TIC itself can be derived using a proof that is similar to, though somewhat simpler than, the one we include in (A2), of which Takeuchi's proof is a special case that is valid only at the MLE estimate $\hat{\boldsymbol{\theta}}$.*

We also define $\hat{J}^*$ to be the negative hessian of $-\ell^*(\boldsymbol{\theta})$ rather than $-\ell(\boldsymbol{\theta})$, and similarly for $\hat{I}^*$, with expectations written as J^* and I^*. Analogously, $-\mathcal{L}^*(\boldsymbol{\theta})$ is the expectation of $-\ell^*(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^*$ is the minimizer of $-\ell^*(\boldsymbol{\theta})$, while $\boldsymbol{\theta}_0^*$ is the minimizer of $-\mathcal{L}^*(\boldsymbol{\theta})$.

We refer to the estimation of $\boldsymbol{\theta}^*$, by minimization of this corrected likelihood function as Information-Corrected Estimation (ICE). As the terminology suggests, we depart from the corrective approach used in Information Criterion, by directly minimizing the bias corrected likelihood function. Note that unlike L_2 regularization, the correction term is

parameter-free and thus would not require cross validation to estimate a hyperparameter such as the λ used by L_2.

General properties of this estimator are proved, and a set of regularity conditions are provided such that the estimator is asymptotically normal, and produces a bias that is $O_p(n^{-3/2})$ instead of the usual $O_p(n^{-1})$. Though this adds only a half-order to the bias correction, for most problems with reasonably large n, any increase in order is likely to greatly reduce bias. Experimental results demonstrate superior properties of ICE for linear models compared to MLE with and without L_2 regularization.

Remark 3. *For models satisfying White's regularity conditions (See [22]), it is known that J_{θ_0} is positive definite (thus non-singular) and continuous, and also that I_{θ_0} is continuous with respect to θ. Therefore, $\frac{1}{n}tr(I_{\theta_0}J_{\theta_0}^{-1})$ would always be well defined in an open region around θ_0. Similarly, the solution θ^* would be expected to have the same properties, and hence (for large enough n) the estimate $\frac{1}{n}tr(\hat{I}_{\theta^*}\hat{J}_{\theta^*}^{-1})$ would be well defined when computed using the estimates $\hat{I}_{\theta^*}$ and $\hat{J}_{\theta^*}$.*

Remark 4. *N.B: Though $-\ell^*(\theta)$ is an estimator of $\mathcal{L}(\theta)$ accurate to within $O(n^{-\frac{3}{2}})$, that does not mean that $\mathcal{L}(\theta^*)$ is reduced by any particular amount relative to $\mathcal{L}(\hat{\theta})$. We expect that using this corrected objective will always (if it can be calculated accurately) generate some improvement by virtue of more accurately representing the true performance of the model out of sample, but there is no proof that this level of improvement has any particular form or asymptotic behavior.*

Our approach preserves the linear complexity of training with respect to n. However, the computation of $\hat{J}_{\theta^*}^{-1}$ at each iteration of the numerical solver requires the inversion of a symmetric positive definite matrix with a complexity of $O(p^3)$. Hence the approach is not suitable for high dimensional datasets without adjustment. See Section 5 for optimized approximations that are viable for larger parameter counts. Further exploration of large models based on this approach are beyond the scope of the present work.

Remark 5. *It is clear from inspection that if $-\ell(\theta)$ is strictly convex, then so too is $-\ell^*(\theta)$ for large enough n.*

We first provide a proof of asymptotic convergence of θ^* under certain regularity conditions. With this convergence result in place, we then show that minimizing (8) leads to an $O(n^{-3/2})$ bias term, an improvement over the $O(\frac{1}{n})$ term produced by MLE.

Local Behavior of the ICE Objective

Suppose the following conditions hold:

1. $\mathcal{M}$ satisfies White's regularity conditions A1–A6 (see Appendix A.1 or [22]).
2. θ_0 is a global minimum of $-\mathcal{L}(\theta)$ in the compact space Θ defined in A2.
3. There exists a $\varepsilon > 0$ such that $-\mathcal{L}(\theta_0) < -\mathcal{L}(\theta_1) - \varepsilon$ for all other local minima θ_1.
4. For $k = 0, 1, 2, 3, 4, 5$ the derivative $\partial_{\theta}^k \mathcal{L}(\theta)$ exists, is continuous, and bounded on an open set around θ_0.
5. For $k = 0, 1, 2, 3, 4, 5$, the variance $\mathbb{V}[\partial_{\theta}^k \ell(\theta, X_n)] \to 0$ as $n \to \infty$ on an open set around θ_0.

Then for sufficiently large n there exists a compact subset $U \subset \Theta$ containing $\theta_0, \hat{\theta}$, such that:

1. For $k = 0, 1, 2, 3$ the derivative $\partial_{\theta}^k \ell^*(\theta, x_n)$ exists, is continuous, and bounded on U, almost surely.
2. For $k = 0, 1, 2, 3$, $\mathbb{V}[\partial_{\theta}^k \ell^*(\theta, X_n)] \to 0$ as $n \to \infty$ on U, almost surely.
3. $\theta^* \in U$ as $n \to \infty$ almost surely.
4. $\sqrt{n}(\theta^* - \theta_0^*) \to N(0, (\hat{J}_{\theta_0^*})^{-1} \hat{I}_{\theta_0^*} (\hat{J}_{\theta_0^*})^{-1})$ almost surely.
5. $-\mathcal{L}(\hat{\theta}^*) = -\ell^*(\theta^*(X_n), X_n) + O_p(n^{-3/2})$ almost surely.

Items (1–3) follow from Lemma A1 (see Appendix A.2). These are additional regularity conditions that are prerequisites for later theorems.

Item (4) follows from Theorem A1 in Appendix A.3. This states that the estimate θ^* is asymtotically normal in a way that is analogous to classical asymptotic normality results for MLE. It is only true almost surely because results (1–3) upon which it relies are only true almost surely.

Item (5) follows from Theorem A2 in Appendix A.4. This item establishes the superior accuracy of the ICE objective compared to the MLE objective function in predicting out of sample errors. Like item (4) this is only true almost surely because intermediate results on which it relies are only true almost surely.

The reduction in generalization error seen arises from the optimization over the superior ICE objective function, analogous to the way that L_2 regularization is used for this purpose.

Remark 6. *The regularity conditions described here are only slightly more strict than the conditions described by White [22]. In particular, models having three continuous derivatives as required by White, but not 5 as needed here are thought to be very rare. Requirement (2) is just the definition of θ_0, which White labels differently, and requirement (3) excludes a pathological corner case, the further study of which is beyond the scope of this paper.*

Remark 7. *Note that as $-\ell(\theta, x_n)$ is convex in the neighborhood of θ_0, so too is $-\ell^*(\theta)$ for large enough n because $-\ell^*(\theta) \to -\ell(\theta)$. Thus it can be concluded that the local behavior of $-\ell^*$ in the neighborhood of θ_0 is not appreciably worse than the behavior of $-\ell$ if the problem is not too ill conditioned.*

4. Direct Computation Results

The following experiments have been designed to compare MLE, MLE with L_2 regularization, and ICE for regression. Each experiment involves simulation of training and test sets and is implemented in R. See the attached code to run each experiment.

Each of these experiments has been performed using the raw formula for ICE provided in Equation (8) with minimal adjustments. All gradients are computed using R's default finite difference approach. This means that for a model with p parameters, the objective function is dominated by the inversion of J, which costs $O(p^3)$ time and $O(p^2)$ space. The use of finite difference gradients further increases the time complexity to $O(p^4)$, compounding the problem. This approach is therefore viable for small models with few parameters, but not realistic for larger models. Optimizations to overcome this limitation will be considered in upcoming Section 5. The use of finite difference derivatives was not found to produce appreciable numerical differences in the final output, so analytic derivatives were not used for this analysis.

The code and results for this section is provided in [23]. Throughout this section, the following estimators will be compared.

MLE	$\hat{\theta}(Z_n)$	$:= \mathrm{argmin}_{\theta}[-\ell(\theta)]$
L_2 regularization	$\theta^*_{L_2}(Z_n)$	$:= \mathrm{argmin}_{\theta,\lambda}[-\ell(\theta) + \lambda\|\theta\|^2_2]$
ICE	$\theta^*_{ICE}(Z_n)$	$:= \mathrm{argmin}_{\theta}[-\ell(\theta) + \frac{1}{n}tr(\hat{I}_{\theta}\hat{J}_{\theta}^{-1})]$

4.1. Gaussian Error Model

We begin by considering the simplest case of univariate linear regression with Gaussian residuals. The advantage of this simple model is that the exact form of the correction term can be derived analytically and aids therefore in building intuition on its behavior. For such a toy model, $y \sim N(\mu, \sigma^2)$ and, for simplicity, the following example will consider μ to be a constant, but it is equally applicable if $\mu = \mu(x)$. Consider the parameters of the model to therefore be $\theta := (\mu, \sigma)$ with their optimal values being $\theta_0 := (\mu_0, \sigma_0)$. The the probability density function is

$$g(y, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{11}$$

It is known a priori that L_2 regularization cannot improve this model, as if $\mu_0 \neq 0$, any decrease in the magnitude of μ is likely to be systematically harmful. Similarly, a decrease in σ below $\hat{\sigma}$ results in a decrease in model distribution entropy, and hence would be generally making overfitting worse, and would generate a correspondingly higher KL-divergence than the MLE estimate. Consequently, we would expect any λ computed through cross-validation to be statistically indistinguishable from zero, and L_2 regularization to be generally harmful whenever $\lambda \neq 0$.

Generalization Error Analysis

The Gaussian model described was generated with $\mu_0 = 0.2$, $\sigma_0 = 0.2$, and $dy = 0.001$. For each of $n \in \{16, 32, 64, 128, 256, 512, 1024\}$, 500 independent simulations of the data $y_1, \ldots, y_n$ were performed, and then the parameters were fit from that data. In each simulation, $\boldsymbol{\theta}$ was computed using MLE, MLE with L_2 regularization, and ICE. The λ parameter for L_2 regularization was computed using 2-way cross-validation on the available data, and as expected, none of the computed values of λ were statistically different from zero.

For each estimate of $\boldsymbol{\theta}$, the KL-divergence $\rho_{KL}(f, g_{\boldsymbol{\theta}})$ was computed (using the known value of $\boldsymbol{\theta}_0$), and the results were compared. The ICE parameter estimation method showed statistically significant improvement over MLE at the 5-sigma level out to $n = 64$, and was improved by just under 1-sigma at $n = 1024$.

The KL-divergence results graphed against n on a log-log scale are shown in Figure 1. Every value of n is normalized by the average KL divergence of the MLE methodology to improve legibility. The L_2 series is statistically indifferent from the MLE series at 2 standard deviations beyond $n = 32$, and the two are not materially different for any n. The ICE series is at least 4.5 standard deviations below the MLE series until $n = 1024$.
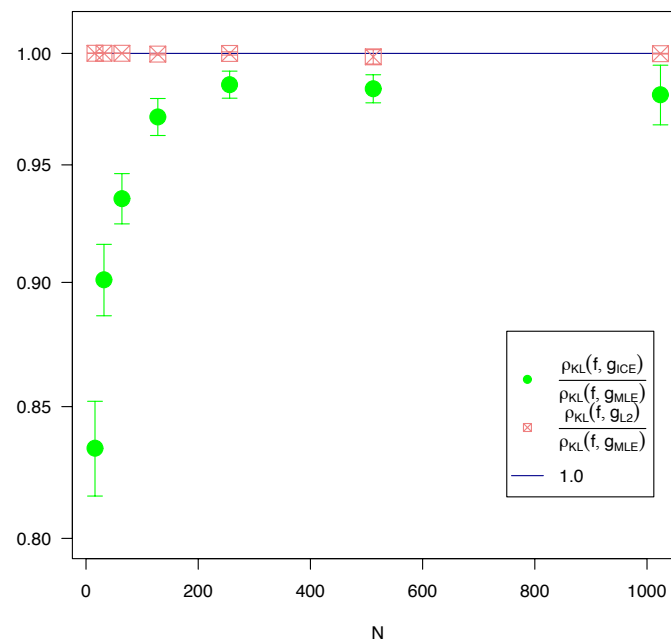
Figure 1. A comparison of the KL-divergence (y-axis) of various estimation methods against the number of training samples n. Each KL divergence value was divided by the average KL divergence of the MLE estimate for that value of n. The ICE and L_2 series are shown with 2 standard deviation error bars.

Remark 8. *In addition to the series shown in Figure 1, a series was computed using the true value of J, estimated from a much larger sample n = 1024 from the underlying distribution, and this series was indistinguishable from the series computed using $\hat{J}$ for every n, thus it was not graphed. This validates Takeuchi's approach of approximating J with $\hat{J}$ in this instance.*

As expected, the difference in μ between ICE and MLE is not statistically significant (at three standard deviations) for any n, but the ICE computed value of σ (shown in Figure 2) is considerably larger than the MLE estimate, especially for small values of n. This explains the greatly reduced KL-divergence noted in Figure 1.
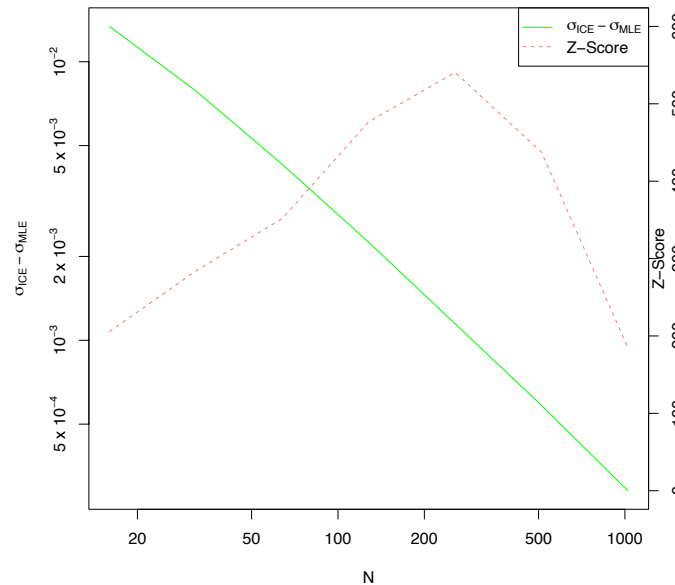
Figure 2. The error in the estimated $\hat{\sigma}_{ICE}$ and the Z-score of the estimate against the number of training samples n.

Note that the difference in estimated σ is always statistically significant when compared to the MLE value. This is because both MLE and ICE are fit on the same data, so ICE would always have a larger σ than MLE regardless of the actual data chosen from the distribution f. This is the cause of the large z-scores shown, always exceeding 200. We know from elementary statistics that correlation between the mean and std. deviation causes the MLE estimate of $\hat{\sigma}$ to be systematically low by a factor of $\frac{n-1}{n}$. Indeed, the ICE estimate of σ^* is closely tracking σ_0 whereas $\hat{\sigma}$ is closely tracking $\frac{\sigma_0(n-1)}{n}$ as expected. This is one example where reducing generalization error also reduces parameter bias as a side effect.

4.2. Friedman's Test Case

We now extend the example from Section 4.1 to the case where μ is no longer constants. For this example, we chose a standard regression test set, which is nonlinear in the features, based on Section 4.3 of [24]:

$$y_i = \mu_{\boldsymbol{\theta}}(x_i) + \varepsilon_i, \ \varepsilon \sim N(0, \sigma^2), \tag{12}$$

where the Friedman model is

$$\mu_{\boldsymbol{\theta}}(x_i) = \theta_0 sin(\pi x_{(i,0)} x_{(i,1)}) + \theta_1 (x_{(i,2)} - \theta_2)^2 + \theta_3 x_{(i,3)} + \theta_4 x_{(i,4)}. \tag{13}$$

The random features, X_j, are i.i.d. uniform random and the parameter values are fixed. The true parameter set, $\boldsymbol{\theta}_0 = (10.0, 20.0, 0.5, 10.0, 5.0, 1.0)$, reserves the last parameter (1.0) for the value of σ.

Note that here σ must be treated as an unknown parameter. To do otherwise implies that the modeler knows the amount of noise expected in the data. In the case of a known noise term, overfitting is impossible since overfitting arises when a model reduces the projected noise below its actual value, which can never arise when the noise level is known.

The model probability density $g(x, y|\boldsymbol{\theta})$ of y is given by

$$g(x_i, y_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mu_i - y_i)^2}{2\sigma^2}}. \tag{14}$$

Recall that in Section 4.1, the value of μ was considered to be a constant. This example is a natural extension of Section 4.1, and was chosen due to the well-explored difficulty of Friedman's problem.

We simulate 500 batches of equally sized training sets of length $n \in \{16, 32, 64, 128, 256, 512, 1024\}$. The test set is always of length 1024 to ensure accuracy for the smaller values of n. The starting point of the optimization is generated by adding a random perturbation, $\delta\boldsymbol{\theta} \sim N(0, 0.1)$, to each parameter. As before, the KL-divergence is computed between the distribution represented by the parameters and the true distribution, and these values are compared between estimation methods.

For each test sample, the KL divergence is computed using numerical integration with a dy increment of 0.01 over the interval containing $\mu \pm 10\sigma$ for both the true and model distributions. The computed probabilities are verified to numerically sum to unity within an error of $\pm 10^{-3}$.

In each simulation, $\boldsymbol{\theta}$ is computed using MLE, MLE with L_2 regularization, and ICE. The λ parameter for L_2 regularization is computed using 4-way cross validation on each batch of the training data.

As shown in Figure 3 and Table 1, L_2 is not effective for any value of n, and is is completely inactivated for $n > 32$. Where regularization is used (i.e., $\lambda \neq 0$), it generally underperforms MLE. ICE is effective across the entire data range, outperforming MLE for every n, and always by a statistically significant margin of at least 5 sigma.
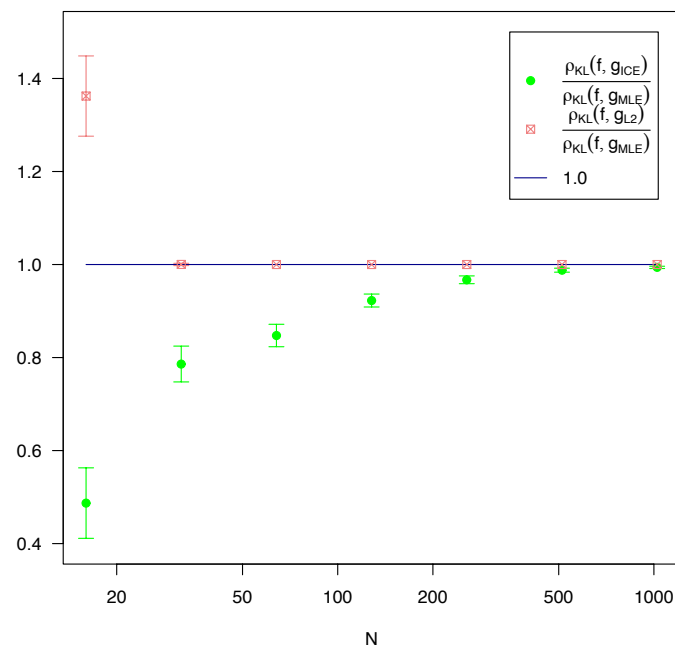
Figure 3. Comparison of the KL-divergence, averaged across 500 replications, of estimation methods against the number of training samples n. Each KL divergence value was divided by the average KL divergence of the MLE estimate for that value of n. The ICE and L_2 series are shown with 2 standard deviation error bars.

Table 1. Comparison of the average KL divergence across 500 replications for several model estimators given a fitting set size of n. For estimators other than $\hat{\theta}$, the values in parentheses denotes the t-statistic of the difference between this estimator and $\hat{\theta}$, with negative values indicating that the listed estimator has a lower KL divergence.

n	$\rho_{KL}(f, g_{\hat{\theta}})$	$\rho_{KL}(f, g_{\theta_{L2}})$	$\rho_{KL}(f, g_{\theta^*})$
16	6.19×10^{-1}	8.442×10^{-1} (8.40)	3.02×10^{-1} (−13.54)
32	1.74×10^{-1}	1.74×10^{-1} (0.16)	1.37×10^{-1} (−11.11)
64	6.85×10^{-2}	6.85×10^{-2} (0.0)	5.81×10^{-2} (−12.65)
128	3.82×10^{-2}	3.82×10^{-2} (0.0)	3.53×10^{-2} (−11.13)
256	2.19×10^{-2}	2.19×10^{-2} (0.0)	2.12×10^{-2} (−7.84)
512	1.53×10^{-2}	1.53×10^{-2} (0.0)	1.51×10^{-2} (−5.66)
1024	1.25×10^{-2}	1.25×10^{-2} (0.0)	1.24×10^{-2} (−5.03)

4.2.1. Impact of $\hat{J}$ Approximation

It was noted previously that Takeuchi used $\hat{J}$ (and likewise, $\hat{I}$) in place of the true value of J, and we do so here as well. Though there is no realistic way to avoid this approximation in the real world, and the optimized approach discussed in Section 5 has an entirely different set of approximations, the impact of this approximation will be briefly characterized here.

In Table 2, we revisit Table 1, but now drop the L_2 regualarization column, and add a new column where the ICE objective is allowed to use a much better approximated value of J, in this case approximated from 1024 independently drawn samples regardless of n.

Table 2. Comparison of the average KL divergence across 500 replications for ICE estimators with and without approximation of J given a fitting set size of n. For estimators other than $\hat{\theta}$, the values in parentheses denotes the t-statistic of the difference between this estimator and $\hat{\theta}$, with negative values indicating that the listed estimator has a lower KL divergence.

n	MLE	ICE ($\hat{J}$)	ICE (J)
16	6.19×10^{-1}	3.02×10^{-1} (−13.54)	6.21×10^{-1} (0.05)
32	1.74×10^{-1}	1.37×10^{-1} (−11.11)	1.51×10^{-1} (−3.74)
64	6.85×10^{-2}	5.81×10^{-2} (−12.65)	4.21×10^{-2} (−17.29)
128	3.82×10^{-2}	3.53×10^{-2} (−11.13)	2.99×10^{-2} (−11.82)
256	2.19×10^{-2}	2.12×10^{-2} (−7.84)	1.99×10^{-2} (−4.19)
512	1.53×10^{-2}	1.51×10^{-2} (−5.66)	1.48×10^{-2} (−1.62)
1024	1.25×10^{-2}	1.24×10^{-2} (−5.03)	1.24×10^{-2} (−0.72)

As can be seen from Table 2, using the true value of J is at most marginally helpful. In fact, for most values of n it displays slightly better average results, but slightly higher std. deviation of those results, and thus reduced T-statistics. Thus, we conclude that the Takeuchi's approximation, replacing J with $\hat{J}$ is reasonable. The same conclusion was reached in Section 4.1, see the remark there. We note also that the ICE estimator using $\hat{J}$ exhibits substantially better performance for very low sample sizes, but further investigation of this phenomenon is beyond the scope of the current paper.

In Table 3, we show the average matrix norms of J, $\hat{J}$, and also of the diagonal of $\hat{J}$, referred to as the matrix D. The matrix D will be examined further in Section 5, and is included here for completeness. We also show the norms of several matrix differences.

We note that the ICE objective values themselves exhibit much lower variation than the matrix norms show in Table 3. In particular though the matrix D is not actually converging to J as n increases, we see from the correction term it generates that this difference does not appear to have a material impact for larger n. We thus conclude that the major eigenvectors of $(\hat{J} - J)$ and $(D - J)$ are very nearly orthogonal to the gradient vectors used to construct $\hat{I}$ for large n.

Table 3. Mean matrix norms of J, its approximations, and differences from these approximations across 500 replications.

n	$\|J\|$	$\|\hat{J}\|$	$\|D\|$	$\|\hat{J}-J\|$	$\|D-J\|$	$\frac{1}{n}tr(IJ^{-1})$	$\frac{1}{n}tr(\hat{I}\hat{J}^{-1})$	$\frac{1}{n}tr(\hat{I}D^{-1})$
16	1423.29	925,489.26	1738.30	926,165.93	2905.64	0.1452	0.0027	6.9484
32	875.03	89,414.11	96.49	89,786.19	793.43	0.0521	0.0256	0.0291
64	214.05	4820.55	89.55	4646.51	125.02	0.0202	0.0160	0.0162
128	200.86	200.68	85.00	27.65	115.86	0.0097	0.0086	0.0086
256	194.36	191.81	82.70	19.12	111.67	0.0048	0.0045	0.0045
512	191.20	190.10	82.76	14.18	108.44	0.0023	0.0023	0.0023
1024	188.91	187.39	81.89	11.58	107.02	0.0012	0.0012	0.0012

It is not clear from examining the trace terms in Table 3 that D is a worse approximation of J than $\hat{J}$ is, even for small n where the impact of the ICE approach is most significant. A more complete investigation of the spectrum of these matrices is beyond the scope of the present work.

4.3. Multivariate Logistic Regression

The previous experiment is based on a well-known test case. In this second experiment, we assess the general performance of ICE under (i) varying dimensionality of the true data distribution, (ii) increasing misspecification, and (iii) increasing training set sizes. To achieve this goal, we generate a more exhaustive set of data from a more complex data generation process.

4.3.1. Data Generation Process

The synthetic data are designed to exhibit a number of characteristics needed to broadly evaluate the efficacy of ICE. First, the regressors should be sufficiently correlated so as to ensure that model selection is representative of typical datasets. However, we avoid multi-collinearity by ensuring the smallest eigenvalue is above a certain threshold. We additionally control the condition number of the covariance matrix Σ by randomly generating a symmetric positive definite covariance matrix $\Sigma \in \mathbb{R}^p$ using the eigen-decomposition

$$\Sigma = UDU^T, \tag{15}$$

where U is an orthogonal random matrix with elements $U_{ij} \sim N(0,1)$ and D is diagonal matrix of positive eigenvalues. The eigenvalues are uniformly distributed over the interval $[a, b]$ so that the condition number of Σ is b/a and the eigenvalues are kept distinct. Here, a is chosen to be 1×10^{-4} and b is chosen to be 0.1.

Using a Cholesky decomposition $\Sigma = \Gamma\Gamma^T$ and the random mean vector $\mu \sim N(0,1)$, we generate correlated gaussian vectors of dimension p with the properties

$$X_i = \mu + \Gamma_{ij}Z_j, \; Z_j \sim N(0,1), \forall j \in 1, \ldots, p. \tag{16}$$

The data (x_n, y_n) are generated under a logistic regression

$$p(y = 1|x, \theta_0) = f(x|\theta_0) = \frac{1}{1 + e^{-x\theta_0}}. \tag{17}$$

A key challenge in assessing the efficacy of bias reduction is to avoid generating excessively low entropy distributions. In such cases, bias reduction will have marginal effect as the parameters are all nearly zero. To avoid such scenarios, the intercept parameter of the true model is adjusted a-posterior until the following conditions are met:

1.　$c < \mathbb{E}_Z[p(Y = 1|X, \theta_0)] < d$
2.　$-\mathcal{L}(\theta_0) > \epsilon$

where $c = 0.35$, $d = 0.65$, and $\epsilon = 0.2$. If these conditions can not be met, then the replication is discarded.

4.3.2. Model Performance Comparison

As in prior sections, KL divergence is computed between the estimated model and the true model for each of the estimation methods. The T-statistics of the difference with the corresponding MLE KL divergence are computed, with negative T-statistics showing that an approach is performing better than the MLE approach. For L_2 regularization in this section, the value of λ is computed via cross-validation, using two folds, on the provided fitting set.

Table 4 compares the KL divergences ρ_{KL} from the true distribution to the model distributions produced using various estimation approaches applied to misspecified data. Here m denotes the number of regressors that are not predictive, i.e., $\boldsymbol{\theta}_0$ contains m zeros. The experiment is replicated 300 times using the data generation process described above and the test set is fixed at 100,000 observations.

Table 4. Comparison of the KL divergence for the different estimation approaches applied to misspecified data. The values in parentheses denote the t-statistic relative to MLE. For $p = \{5, 10, 20\}$ there are $m = \{2, 4, 8\}$ non-explanatory variables added.

p	n	$\rho_{KL}(f, g_{\hat{\theta}})$	$\rho_{KL}(f, g_{\theta_{L_2}})$	$\rho_{KL}(f, g_{\theta^*})$
5	500	4.79×10^{-3}	3.01×10^{-3} (-13.43)	4.56×10^{-3} (-13.53)
5	1000	2.64×10^{-3}	1.76×10^{-3} (-10.94)	2.57×10^{-3} (-12.80)
5	2000	1.29×10^{-3}	1.09×10^{-3} (-6.15)	1.27×10^{-3} (-7.69)
5	5000	5.09×10^{-4}	4.60×10^{-4} (-4.69)	5.07×10^{-4} (-6.19)
10	500	9.79×10^{-3}	9.85×10^{-3} (0.16)	9.18×10^{-3} (-6.27)
10	1000	5.05×10^{-3}	5.13×10^{-3} (0.51)	4.83×10^{-3} (-4.90)
10	2000	2.50×10^{-3}	3.05×10^{-3} (5.99)	2.56×10^{-3} (1.70)
10	5000	1.06×10^{-3}	1.49×10^{-3} (7.72)	1.04×10^{-3} (-0.86)
20	500	2.18×10^{-2}	2.16×10^{-2} (-0.29)	1.95×10^{-2} (-8.71)
20	1000	1.13×10^{-2}	1.24×10^{-2} (3.79)	1.10×10^{-2} (-1.95)
20	2000	6.86×10^{-3}	7.52×10^{-3} (4.47)	6.72×10^{-3} (-1.67)
20	5000	3.57×10^{-3}	4.24×10^{-3} (6.56)	3.59×10^{-3} (0.45)

We observe that the t-statistic for $\boldsymbol{\theta}^*$ is most significant for relatively small sample sizes, particularly $n = 500$. For these small sizes, the improvement over MLE is greater, though noisier. There is uniform decay in improvement over $\hat{\boldsymbol{\theta}}$ as n grows, until for $p = 10$ and $p = 20$ the largest sizes are no longer statistically significant. This is expected, as both the MLE and ICE estimates are converging towards the true value of $\boldsymbol{\theta}_0$, and for large enough sample sizes the ICE correction would be dominated by numerical error, particularly the ill conditioning of J.

The L_2 estimate improves for small values of p, but then becomes progressively worse for large values of p. We observe that for dimensionality above $p = 5$, the L_2 regularization described here is no longer effective in reducing the KL-divergence. For low values of p the value of $\boldsymbol{\theta} x$ has comparatively low variance, and thus the logistic function is reasonably locally approximated as linear. For higher p this approximation is less realistic and the performance of L_2 regularization degrades.

For the ICE estimates, larger values of p show fluctuations that are often not statistically significant. It is apparent that larger p is increasing the variance of the ICE divergences, probably due to numerical errors and ill conditioning. Larger values of n reduce the absolute size of the divergence improvement whereas larger values of p seem to increase it.

Note that though the t-statistics are degrading for large n, the absolute magnitude of the differences is asymptotically small. For these sizes, the results are insignificant, but more importantly, immaterial.

4.3.3. Convergence Analysis for Large n

For 10 randomly chosen example problems, under which the model coefficients are now fixed, the convergence behavior for large n, the training set size, is explored. Note that the test set remains fixed at 100,000 observations for each problem. Table 5 compares the KL divergence (averaged over all 10 problems) under MLE ($\hat{\theta}$), L_2 regularization, and ICE for progressively larger sample sizes. The divergences $\rho_{KL}(f, g_{\hat{\theta}})$ and $\rho_{KL}(f, g_{\theta^*_{ICE}})$ converge to zero as $n \to \infty$, as does $\rho_{KL}(f, g_{\theta^*_{L_2}})$.

Table 5. Comparison of the KL divergence under the MLE $\hat{\theta}$, L_2 regularization and ICE regularization θ^*_{ICE} against a large sample size for the case when $p = 10$ and $m = 4$.

n	$\mathcal{L}(\theta_0)$	$\rho_{KL}(f, g_{\hat{\theta}})$	$\rho_{KL}(f, g_{\theta_{L_2}})$	$\rho_{KL}(f, g_{\theta^*})$
500	0.5439	9.28×10^{-3}	7.92×10^{-3}	7.74×10^{-3}
1000	0.5439	5.50×10^{-3}	5.86×10^{-3}	4.81×10^{-3}
2000	0.5439	2.65×10^{-3}	3.67×10^{-3}	2.65×10^{-3}
5000	0.5439	1.85×10^{-3}	2.72×10^{-3}	1.35×10^{-3}
10,000	0.5439	5.75×10^{-4}	1.57×10^{-3}	9.32×10^{-4}
20,000	0.5439	5.84×10^{-4}	8.10×10^{-4}	6.11×10^{-4}
50,000	0.5439	3.83×10^{-4}	4.09×10^{-4}	3.64×10^{-4}
100,000	0.5439	1.67×10^{-4}	1.15×10^{-3}	1.86×10^{-4}

Generally the θ^*_{ICE} estimates are seen to converge slightly faster than the $\hat{\theta}$ estimates. The regularization in $\theta^*_{L_2}$ is observed to be beneficial for very small sample sizes, but then becomes marginally detrimental for large n.

5. Optimized Computation Results

For any model satisfying White's Regularity Criteria, it is known that the matrix J is positive definite near the MLE optimum $\hat{\theta}$. This implies that J is diagonally dominated, and indeed considering just its diagonal elements D, it is known that $tr(ID^{-1}) > 0$. Indeed $tr(ID^{-1})$ differs strongly from $tr(IJ^{-1})$ most strongly for models with strong regressor interactions. Therefore, using finite difference gradients, consider the following approximations for the ICE objective function:

1. θ^*: J is computed directly, ICE is implemented as written.
2. θ^*_2: J is taken to be constant w.r.t. θ: ($J_\theta = J_{\hat{\theta}}$).
3. θ^*_3: J is taken to be diagonal: ($J = D$).
4. θ^*_4: J is taken to be the identity: ($J = \mathcal{I}$).

Clearly, we expect that θ^*_4 above is the least accurate approximation, and items θ^*_2 and θ^*_3 have varying levels of accuracy depending on the problem at hand. The cost comparison of these approaches is shown in Table 6.

Table 6. The asymptotic computational cost (per iteration) of various proposed approximations as a function of parameter count p. Cost is amortized when $J_\theta = J_{\hat{\theta}}$ assuming that $n \approx p$. Note that a typical model will cost $O(p)$ in time and space for both the objective function and its gradients.

Approximation	Objective Cost (Space)	Objective Cost (Time)	Gradient Cost (Space)	Gradient Cost (Time)
Direct Computation	$O(p^2)$	$O(p^3)$	$O(p^2)$	$O(p^4)$
$J_\theta = J_{\hat{\theta}}$	$O(p^2)$	$O(p^2)$	$O(p^2)$	$O(p^3)$
$J = D$	$O(p)$	$O(p)$	$O(p)$	$O(p^2)$
$J = \mathcal{I}$	$O(p)$	$O(p)$	$O(p)$	$O(p^2)$

Remark 9. *When computing gradients for use in a solver, often approximation error will have only a marginal impact on the final result, though it may increase the number of iterations needed for convergence. Broyden's method [25] is a typical example of this approach in action. Efficient approximations of $[\partial_\theta \hat{J}]$ might similarly have only a minor effect on accuracy and iteration count. The construction of approximate analytical derivatives is beyond the scope of the present work.*

These approximations were computed and compared for the Friedman (see Section 4.2) problem, and the results are shown in Table 7 below.

Table 7. Comparison of the average KL divergence across 200 replications for MLE and several variants of ICE given a fitting set size of n. For estimators other than $\hat{\theta}$, the values in parentheses denotes the t-statistic of the difference between this estimator and $\hat{\theta}$, with negative values indicating that the listed estimator has a lower KL divergence.

n	$\rho_{KL}(f, g_{\hat{\theta}})$	$\rho_{KL}(f, g_{\theta^*})$	$\rho_{KL}(f, g_{\theta_2^*})$	$\rho_{KL}(f, g_{\theta_3^*})$	$\rho_{KL}(f, g_{\theta_4^*})$
8	$1.22 \times 10^{+1}$	$4.55 \times 10^{+0}$ (-4.89)	$5.70 \times 10^{+0}$ (-5.26)	$3.83 \times 10^{+0}$ (-5.22)	$1.28 \times 10^{+0}$ (-4.67)
16	6.68×10^{-1}	2.99×10^{-1} (-8.13)	3.53×10^{-1} (-10.56)	3.36×10^{-1} (-8.30)	5.47×10^{-1} (-2.11)
32	1.45×10^{-1}	1.14×10^{-1} (-6.90)	1.04×10^{-1} (-8.18)	1.08×10^{-1} (-10.16)	3.63×10^{-1} (19.60)
64	5.93×10^{-2}	4.80×10^{-2} (-10.42)	4.70×10^{-2} (-6.95)	4.81×10^{-2} (-9.81)	2.38×10^{-1} (39.08)
128	2.48×10^{-2}	2.24×10^{-2} (-6.38)	2.33×10^{-2} (-2.26)	2.26×10^{-2} (-6.00)	1.62×10^{-1} (61.85)
256	1.21×10^{-2}	1.16×10^{-2} (-4.28)	1.20×10^{-2} (-0.68)	1.16×10^{-2} (-4.11)	1.01×10^{-1} (68.82)
512	6.26×10^{-3}	6.10×10^{-3} (-2.41)	6.16×10^{-3} (-0.84)	6.10×10^{-3} (-2.39)	5.42×10^{-2} (63.84)
1024	3.05×10^{-3}	3.00×10^{-3} (-2.66)	3.04×10^{-3} (-0.37)	2.99×10^{-3} (-2.73)	2.61×10^{-2} (59.78)

From Table 7, it is apparent that approach θ_4^*, taking $J = \mathcal{I}$ is not effective. This is not surprising as the actual J matrix has dramatic differences in scale between regressors. Approximation θ_3^*, taking $J = D$ is accurate enough that it cannot be statistically distinguished from the direct computation of ICE by the test above. Approximation θ_2^* tends to underperform approximation (3).

Therefore, we propose taking $J = D$ as a more numerically stable approximation of the ICE objective.

6. Conclusions

Takeuchi [16] is believed to be the first to have proposed using an objective function similar to ICE in order to reduce generalization error, though it was applied via model selection. Firth [9] introduced a similar term to reduce parameter bias in model fitting, as opposed to model selection, though he derived it only for exponential model families and did not consider its effect on generalization error. It is not known why this approach did not find widespread use, but one may infer that the $O(p^4)$ computational cost and instability was enough to keep it from wider adoption.

In this paper, we reintroduce the objective function of [16] and provide a more general proof of its widespread applicability. We then show that efficient implementations costing only $O(p)$ are possible. Under finite sample sizes, this bias correction term is shown experimentally in several models to lead to significant reduction in bias compared to maximum likelihood estimation with and without L_2 regularization. ICE offers many advantages over L_2 penalized maximum likelihood estimation: (i) it's suitable for most nonlinear models, (ii) it's provably asymptotically convergent; and (iii) does not rely on any parameters which would need to be provided by the operator or deduced through cross-validation.

Author Contributions: Conceptualization, M.D. and T.W.; methodology, M.D. and T.W.; software, T.W.; validation, M.D. and T.W.; formal analysis, M.D. and T.W.; writing—original draft preparation, M.D. and T.W.; writing—review and editing, M.D. and T.W. All authors have read and agreed to the published version of the manuscript.

Appendix A. Proofs

Appendix A.1. White's Regularity Conditions

Definition A1 (White's regularity conditions). *White [22] provides the following regularity conditions:*

A1: *The independent random vectors, X_i, $i = 1, \ldots, n$, have common joint distribution function F on Ω, a measurable Euclidean space, with measurable Radon–Nikodym density $f = dF/dx$.*

A2: *The family of distribution functions $G(x|\boldsymbol{\theta})$ has Radon–Nikodym densities $g(x|\boldsymbol{\theta}) = dG(x|\boldsymbol{\theta})/dx$ which are measurable in x for every $\boldsymbol{\theta} \in \Theta$, a compact subset of p-dimensional Euclidean space, and continuous in $\boldsymbol{\theta}$ for every $x \in \Omega$.*

A3: *(a) $\mathbb{E}[\log f(X)]$ exists and $|\log g(x|\boldsymbol{\theta})| \leq m(x)$, $\forall \boldsymbol{\theta} \in \Theta$, where m is integrable with respect to F; (b) $\rho_{KL}(f, g_{\boldsymbol{\theta}})$ has a unique minimum at $\boldsymbol{\theta}_0 \in \Theta$.*

A4: *$\partial_{\boldsymbol{\theta}}(\log g(x|\boldsymbol{\theta}))$ are measurable functions of x for each $\boldsymbol{\theta} \in \Theta$ and continuously differentiable functions of $\boldsymbol{\theta}$ for each $x \in \Omega$.*

A5: *$|\partial_{\boldsymbol{\theta}}^2(\log g(x|\boldsymbol{\theta}))|$ and $|\partial_{\boldsymbol{\theta}}(\log g(x|\boldsymbol{\theta})) \cdot \partial_{\boldsymbol{\theta}}(\log g(x|\boldsymbol{\theta}))|$ are dominated by functions integrable in x with respect to F for all $x \in \Omega$ and $\boldsymbol{\theta} \in \Theta$.*

A6: *(a) $\boldsymbol{\theta}_0$ is an interior point of the parameter space; (b) $\mathbb{E}[\partial_{\boldsymbol{\theta}}(\log g(x|\boldsymbol{\theta})) \cdot \partial_{\boldsymbol{\theta}}(\log g(x|\boldsymbol{\theta}))]$ is non-singular; (c) $\boldsymbol{\theta}_0$ is a regular point of $\mathbb{E}[\partial_{\boldsymbol{\theta}}^2(\log g(x|\boldsymbol{\theta}))]$.*

Appendix A.2. Proof of Finite Variance

Lemma A1 (Finite variance). *Suppose the following conditions hold:*

1. *$\mathcal{M}$ satisfies White's regularity conditions A1–A6 (see Appendix A.1 or [22]).*
2. *$\boldsymbol{\theta}_0$ is a global minimum of $-\mathcal{L}(\boldsymbol{\theta})$ in the compact space Θ defined in A2.*
3. *There exists a $\varepsilon > 0$ such that $-\mathcal{L}(\boldsymbol{\theta}_0) < -\mathcal{L}(\boldsymbol{\theta}_1) - \varepsilon$ for all other local minima $\boldsymbol{\theta}_1$.*
4. *For $k = 0, 1, 2, 3, 4, 5$ the derivative $\partial_{\boldsymbol{\theta}}^k \mathcal{L}(\boldsymbol{\theta})$ exists, is continuous, and bounded on an open set around $\boldsymbol{\theta}_0$.*
5. *For $k = 0, 1, 2, 3, 4, 5$, the variance $\mathbb{V}[\partial_{\boldsymbol{\theta}}^k \ell(\boldsymbol{\theta}, X_n)] \to 0$ as $n \to \infty$ on an open set around $\boldsymbol{\theta}_0$.*

   Then, for sufficiently large n there exists a compact subset $U \subset \Theta$ containing $\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}$, such that

1. *For $k = 0, 1, 2, 3$ the derivative $\partial_{\boldsymbol{\theta}}^k \ell^*(\boldsymbol{\theta}, \boldsymbol{x}_n)$ exists, is continuous, and bounded on U, almost surely.*
2. *For $k = 0, 1, 2, 3$, $\mathbb{V}[\partial_{\boldsymbol{\theta}}^k \ell^*(\boldsymbol{\theta}, X_n)] \to 0$ as $n \to \infty$ on U, almost surely.*
3. *$\boldsymbol{\theta}^* \in U$ as $n \to \infty$ almost surely.*

Proof. Assumptions (4) and (5) establish existence of some set, S, containing $\boldsymbol{\theta}_0$ such that $\mathcal{L}(\boldsymbol{\theta})$ is bounded on S, and its estimate, $\ell(\boldsymbol{\theta}, X_n)$ has finite variance. Therefore, $-\ell(\boldsymbol{\theta}, \boldsymbol{x}_n)$ is also bounded on S almost surely. Similarly for the first 5 derivatives. White's criteria imply that $J_{\boldsymbol{\theta}_0}$ is positive definite on an open set around $\boldsymbol{\theta}_0$, and thus one can form a compact set $U \subset S$ containing an open set around $\boldsymbol{\theta}_0$ on which the minimum eigenvalue of $J_{\boldsymbol{\theta}}$ is bounded away from 0.

Note that $J_{\boldsymbol{\theta}}$ is three times differentiable on U by Assumption 4, as is $\hat{J}_{\boldsymbol{\theta}}$, as established above. Then $\hat{J}_{\boldsymbol{\theta}}^{-1}$ is also positive definite and bounded on U. It can be shown to also have three derivatives by using the well-known matrix relation

$$\partial_{\boldsymbol{\theta}} A^{-1} = -A^{-1}(\partial_{\boldsymbol{\theta}} A) A^{-1}. \tag{A1}$$

It follows that $\hat{J}_{\boldsymbol{\theta}}^{-1}$ is also positive definite, nonsingular, and bounded on U. Similarly for $\hat{I}_{\boldsymbol{\theta}}$, and thus $tr(\hat{I}_{\boldsymbol{\theta}}\hat{J}_{\boldsymbol{\theta}}^{-1})$ is bounded with finite variance on U. It also has three bounded derivatives with finite variance.

Therefore, $-\ell^*(\boldsymbol{\theta}, X_n) \to -\ell(\boldsymbol{\theta}, X_n)$, and $\boldsymbol{\theta}^* \to \boldsymbol{\theta}_0$ as $n \to \infty$, with the convergence being in probability. This means that U contains $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}$, and $\boldsymbol{\theta}^*$ almost surely for large enough n. Similarly, on U we have three continuous, bounded derivatives of $\partial_{\boldsymbol{\theta}}^k \ell^*(\boldsymbol{\theta}, x_n)$ almost surely. $\square$

Appendix A.3. Proof of Asymptotic Normality

Theorem A1 (Asymptotic Normality). *Provided the conditions hold in Lemma A1, namely, that*

1. *For $k = 0, 1, 2, 3$ the derivative $\partial_{\boldsymbol{\theta}}^k \ell^*(\boldsymbol{\theta}, x_n)$ exists, is continuous, and bounded on U.*
2. *For $k = 0, 1, 2, 3$, $\mathbb{V}[\partial_{\boldsymbol{\theta}}^k \ell^*(\boldsymbol{\theta}, x_n)] \to 0$ as $n \to \infty$ on U.*

Then

$$\sqrt{n}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*) \to N(0, (\hat{J}_{\boldsymbol{\theta}_0^*}^*)^{-1}\hat{I}_{\boldsymbol{\theta}_0^*}^*(\hat{J}_{\boldsymbol{\theta}_0^*}^*)^{-1}).$$

Proof. As the first derivatives of ℓ are continuous, the mean value theorem may be applied:

$$\partial_{\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}_0^*) = \partial_{\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)\hat{J}_{\bar{\boldsymbol{\theta}}}^* = (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)\hat{J}_{\bar{\boldsymbol{\theta}}}^*. \tag{A2}$$

$\bar{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_0^*$. Under the assumptions of Lemma A1, and given its finite variance, $\hat{J}_{\bar{\boldsymbol{\theta}}}$ is almost surely (in the large n limit) positive definite, and thus invertible as both $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_0^*$ are in U, and $\bar{\boldsymbol{\theta}}$ is between them. Therefore,

$$(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*) = (\hat{J}_{\bar{\boldsymbol{\theta}}}^*)^{-1}\partial_{\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}_0^*). \tag{A3}$$

Applying the mean value theorem a second time gives

$$\hat{J}_{\bar{\boldsymbol{\theta}}} = \hat{J}_{\boldsymbol{\theta}_0^*}^* + (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*)\hat{J}_{\boldsymbol{\theta}_1}^*, \tag{A4}$$

with $\boldsymbol{\theta}_1$ between $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0^*$. If the order of $(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*) = O_p(\delta)$, where $\delta := n^{-1/2}$, then

$$\hat{J}_{\bar{\boldsymbol{\theta}}}^* = \hat{J}_{\boldsymbol{\theta}_0^*}^* + O_p(\delta). \tag{A5}$$

As all of the $\hat{J}^*$ are bounded away from zero in probability, we have

$$(\hat{J}_{\bar{\boldsymbol{\theta}}}^*)^{-1} = (\hat{J}_{\boldsymbol{\theta}_0^*}^*)^{-1} + O_p(\delta), \tag{A6}$$

with the equality holding in probability. In the large n limit, $\delta \to 0$, and thus

$$(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*) = (\hat{J}_{\boldsymbol{\theta}_0^*}^*)^{-1}\partial_{\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}_0^*). \tag{A7}$$

As $\partial_{\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}_0^*)$ is the sum of n independent vectors, it is asymptotically normally distributed by the central limit theorem, and its mean is 0 by the definition of $\boldsymbol{\theta}_0^*$. Its variance is therefore $\mathbb{V}[\partial_{\boldsymbol{\theta}^*}\ell^*(\boldsymbol{\theta}_0^*)] = \mathbb{E}[\partial_{\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}_0^*)(\partial_{\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}_0^*))^T] = \frac{1}{n}\hat{I}_{\boldsymbol{\theta}_0^*}$

Substituting this into Equation (A7) yields

$$\sqrt{n}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*) = N(0, (\hat{J}_{\boldsymbol{\theta}_0^*}^*)^{-1}\hat{I}_{\boldsymbol{\theta}_0^*}^*(\hat{J}_{\boldsymbol{\theta}_0^*}^*)^{-1}), \tag{A8}$$

establishing the result. $\square$

Appendix A.4. Proof of Prediction Bias Order under ICE

Theorem A2 (Prediction Bias Estimation under ICE). *By minimizing $-\ell^*$ instead of $-\ell$, the first order terms of the prediction bias are cancelled leaving a $O_p(n^{-3/2})$ residual term*

$$-\mathcal{L}(\hat{\boldsymbol{\theta}}^*) = -\ell(\boldsymbol{\theta}^*(X_n), X_n) + O_p(n^{-3/2}). \tag{A9}$$

Proof. Note that in Takeuchi's proof [16], the use of $\hat{\boldsymbol{\theta}}$ was prescribed. Therefore, this approach could be used only for model selection, and not for model fitting. Now consider the bias under the ICE estimator $\boldsymbol{\theta}^*$.

$$
\begin{aligned}
b(\boldsymbol{\theta}^*(X_n), X_n) &= \mathbb{E}_{X_n}[\log g(X_n|\boldsymbol{\theta}^*(X_n)) - \log g(X_n|\boldsymbol{\theta}_0)] \\
&+ \mathbb{E}_{X_n}[\log g(X_n|\boldsymbol{\theta}_0) - n\mathbb{E}_{Z_n}[\log g(Z_n|\boldsymbol{\theta}_0)]] \\
&+ n\mathbb{E}_{X_n}[\mathbb{E}_{Z_n}[\log g(Z_n|\boldsymbol{\theta}_0)] - \mathbb{E}_{Z_n}[\log g(Z_n|\boldsymbol{\theta}^*(X_n))]].
\end{aligned}
$$

As the second term is zero, this can be simplified to

$$
\begin{aligned}
b(\boldsymbol{\theta}^*(X_n), X_n) &= -n\mathbb{E}_{X_n}[\ell(\boldsymbol{\theta}^*(X_n), X_n) - \ell(\boldsymbol{\theta}_0, X_n)] \\
&- n\mathbb{E}_{X_n}[\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}(\boldsymbol{\theta}^*(X_n))].
\end{aligned}
$$

Define $\delta = \frac{1}{\sqrt{n}}$, and then recall from White [22] that $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is $O_p(\delta)$. Similarly, recall from Theorem A1 that $(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)$ is also $O_p(\delta)$. As constructed, the error term $b(\boldsymbol{\theta}^*(X_n), X_n)$ is $O_p(1)$ (actually $O(1)$ as it is an expectation), and terms of order $O_p(\delta)$ and higher will be dropped. Therefore, as in Takeuchi's derivation [16], the Taylor expansions below will be truncated at second order in δ, dropping terms of order $O_p(\delta^3)$ or higher. Additionally, we will occasionally drop indications of X_n where the meaning is clear and it greatly simplifies the notation.

With that truncation, recalling that $\boldsymbol{\theta}_0$ is a minimum of $\mathcal{L}(\boldsymbol{\theta})$ and thus has zero gradient:

$$n\mathbb{E}_{X_n}[\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}(\boldsymbol{\theta}^*(X_n))] = \frac{n}{2}\mathbb{E}_{X_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)] + O(\delta). \tag{A10}$$

Recall Theorem A1, and recall the that for the quadratic form:

$$\mathbb{E}[\varepsilon^T A\varepsilon] = tr(A\Sigma) + \mu^T A\mu, \qquad \mathbb{E}[\varepsilon] = \mu, \mathbb{V}[\varepsilon] = \Sigma. \tag{A11}$$

Therefore,

$$
\begin{aligned}
\frac{n}{2}\mathbb{E}_{X_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)] &= \frac{1}{2}tr(J_{\boldsymbol{\theta}_0}[(J_{\boldsymbol{\theta}_0^*}^*)^{-1} I_{\boldsymbol{\theta}_0^*}^* (J_{\boldsymbol{\theta}_0^*}^*)^{-1})]) \\
&+ \frac{n}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\
&+ O(\delta).
\end{aligned}
$$

Now note that the second term on the right is a constant, and therefore would take no part in any optimization. Therefore, it can be safely ignored and

$$n\mathbb{E}_{X_n}[\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}(\boldsymbol{\theta}^*(X_n))] = \frac{1}{2}tr(J_{\boldsymbol{\theta}_0}[(J_{\boldsymbol{\theta}_0^*}^*)^{-1} I_{\boldsymbol{\theta}_0^*}^* (J_{\boldsymbol{\theta}_0^*}^*)^{-1})]) + O(\delta). \tag{A12}$$

Addressing the first term, again taking a Taylor expansion, we find that

$$\ell(\boldsymbol{\theta}_0) = \ell(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)^T \partial_{\boldsymbol{\theta}}^2\ell(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + O_p(\delta^3). \tag{A13}$$

Therefore, recalling that $nO_p(\delta^3) = O_p(\delta)$ gives

$$
\begin{aligned}
n\mathbb{E}_{\boldsymbol{X}_n}[\ell(\boldsymbol{\theta}^*(\boldsymbol{X}_n), X_n) - \ell(\boldsymbol{\theta}_0, X_n)] &= \frac{1}{2}tr(J_{\boldsymbol{\theta}^*}[(J_{\boldsymbol{\theta}_0}^*)^{-1}I_{\boldsymbol{\theta}_0}^*(J_{\boldsymbol{\theta}_0}^*)^{-1})]) \\
&+ \frac{n}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\
&+ n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
&+ O_p(\delta).
\end{aligned}
$$

Now examine the last term:

$$
\begin{aligned}
n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] &= n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
&+ n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)].
\end{aligned}
$$

However, $(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)$ does not depend on $\boldsymbol{X}_n$, so it can be pulled out of the expectation, then substitute in a first order Taylor expansion

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] &= (\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \mathbb{E}_{\boldsymbol{X}_n}[\partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
&= (\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \mathbb{E}_{\boldsymbol{X}_n}[\partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0^*) + (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)\partial_{\boldsymbol{\theta}}^2\ell(\boldsymbol{\theta}_0^*) + O_p(\delta^2)] \\
&= (\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_0^*) \\
&+ (\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)]\partial_{\boldsymbol{\theta}}^2\mathcal{L}(\boldsymbol{\theta}_0^*) + O(\delta^3) \\
&= (\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_0^*) + O(\delta^3),
\end{aligned}
$$

with the last equality following from the fact that $\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)] = 0$ and the substitution of $(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T O(\delta^2) = O(\delta^3)$. This term is therefore a constant, up to $O(\delta^3)$, and takes no part in optimization of $\boldsymbol{\theta}$, thus it can be dropped from further consideration. Therefore $n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] = O(\delta)$, and

$$
\begin{aligned}
n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] &= n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
&+ O(\delta).
\end{aligned}
$$

Recombining these terms yields

$$
\begin{aligned}
n\mathbb{E}_{\boldsymbol{X}_n}[\ell(\boldsymbol{\theta}^*(\boldsymbol{X}_n), X_n) - \ell(\boldsymbol{\theta}_0, X_n)] &= \frac{1}{2}tr(J_{\boldsymbol{\theta}^*}[(J_{\boldsymbol{\theta}^*}^*)^{-1}I_{\boldsymbol{\theta}^*}^*(J_{\boldsymbol{\theta}^*}^*)^{-1})]) \\
&+ \frac{n}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\
&+ n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
&+ O(\delta).
\end{aligned}
$$

Thus the bias (neglecting the constant terms) is then

$$
\begin{aligned}
b(\boldsymbol{\theta}^*(\boldsymbol{X}_n), \boldsymbol{X}_n) &= -\frac{1}{2}tr(J_{\boldsymbol{\theta}^*}[(J_{\boldsymbol{\theta}^*}^*)^{-1}I_{\boldsymbol{\theta}^*}^*(J_{\boldsymbol{\theta}^*}^*)^{-1})]) \\
&- \frac{1}{2}tr(J_{\boldsymbol{\theta}_0}[(J_{\boldsymbol{\theta}_0}^*)^{-1}I_{\boldsymbol{\theta}_0}^*(J_{\boldsymbol{\theta}_0}^*)^{-1})]) \\
&- \frac{n}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\
&- n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
&+ O(\delta).
\end{aligned}
$$

Because $\ell^*(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + O_p(\delta^2)$, it follows that $J_{\boldsymbol{\theta}}^* = J_{\boldsymbol{\theta}} + O_p(\delta^2)$ and thus $J_{\boldsymbol{\theta}^*}(J_{\boldsymbol{\theta}^*}^*)^{-1} = I + O(\delta^2)$. Similarly for $J_{\boldsymbol{\theta}_0}$. In addition $I_{\boldsymbol{\theta}^*} = I_{\boldsymbol{\theta}_0^*} + O(\delta)$, thus the two trace terms can be simplified and combined:

$$
\begin{aligned}
b(\boldsymbol{\theta}^*(\boldsymbol{X}_n), \boldsymbol{X}_n) \;=\;& -tr(I_{\boldsymbol{\theta}^*} J_{\boldsymbol{\theta}^*}^{-1}) \\
& -\; \frac{n}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\
& -\; n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
& +\; O(\delta).
\end{aligned}
$$

As $\boldsymbol{\theta}_0^*$ is a minimum of $\mathcal{L}^*$, begin by Taylor expanding the derivatives

$$
\partial_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_0^*) = \partial_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_0) - (\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)J_{\boldsymbol{\theta}_0} = -(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)J_{\boldsymbol{\theta}_0}. \tag{A14}
$$

Moreover, show from the definition of $\mathcal{L}^*$ that

$$
\partial_{\boldsymbol{\theta}}\mathcal{L}^*(\boldsymbol{\theta}_0^*) = 0 = \partial_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_0^*) + \frac{1}{n}\partial_{\boldsymbol{\theta}}tr(IJ^{-1}) \tag{A15}
$$

Then, after Taylor expanding $\mathcal{L}(\boldsymbol{\theta}_0^*)$ around $\boldsymbol{\theta}_0$, it is seen that

$$
(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) = \frac{1}{n}J_{\boldsymbol{\theta}_0}^{-1}\partial_{\boldsymbol{\theta}}tr(IJ^{-1}). \tag{A16}
$$

Noting that $J_{\boldsymbol{\theta}}^{-1} = O(1)$ and $tr(IJ^{-1}) = O(1)$, it holds that $(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) = O(\delta^2)$.
Therefore, $\frac{n}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) = O(\delta^2)$, and can be neglected.
Then, the bias becomes

$$
\begin{aligned}
b(\boldsymbol{\theta}^*(\boldsymbol{X}_n), \boldsymbol{X}_n) \;=\;& -tr(I_{\boldsymbol{\theta}^*} J_{\boldsymbol{\theta}^*}^{-1}) \\
& -\; n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)^T \partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] \\
& +\; O(\delta).
\end{aligned}
$$

However, for the same reason, $\partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*) = O(\delta^2)$, so the last term $n\mathbb{E}_{\boldsymbol{X}_n}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*)^T$ $\partial_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*)] = O(\delta)$, and it too can be absorbed into the residual.

Therefore,

$$
\begin{aligned}
b(\boldsymbol{\theta}^*(\boldsymbol{X}_n), \boldsymbol{X}_n) \;=\;& -tr(I_{\boldsymbol{\theta}^*} J_{\boldsymbol{\theta}^*}^{-1}) \\
& +\; O(\delta),
\end{aligned}
$$

and

$$
-\mathcal{L}(\hat{\boldsymbol{\theta}}) = -\ell(\boldsymbol{\theta}^*(Z_n), Z_n) + \frac{1}{n}tr(I_{\hat{\boldsymbol{\theta}}} J_{\hat{\boldsymbol{\theta}}}^{-1}) + O_p(\delta^3) + C, \tag{A17}
$$

where the constant C is composed of the neglected constant terms from earlier stages

$$
\begin{aligned}
C \;=\;& -\frac{1}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\
& +\; -(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T \partial_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_0^*).
\end{aligned}
$$

However, the last term is $O(\delta^3)$, and the first is $O(\delta^2)$, so this may be approximated as

$$
C = -\frac{1}{2}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0)^T J_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0). \tag{A18}
$$

Recalling again that $(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) = O(\delta^2)$, it is clear that $C = O(\delta^4)$, and may thus be absorbed into the $O(\delta^3)$ residual term. Therefore,

$$
-\mathcal{L}(\hat{\boldsymbol{\theta}}) = -\ell(\boldsymbol{\theta}^*(Z_n), Z_n) + \frac{1}{n}tr(I_{\hat{\boldsymbol{\theta}}} J_{\hat{\boldsymbol{\theta}}}^{-1}) + O(\delta^3). \tag{A19}
$$

Comparing this to the form of $\ell^*(\boldsymbol{\theta})$:

$$- \mathcal{L}(\hat{\boldsymbol{\theta}}) = -\ell^*(\boldsymbol{\theta}^*(Z_n), Z_n) + O_p(\delta^3). \tag{A20}$$

Thus, by minimizing $-\ell^*$ instead of $-\ell$, the first order terms of the prediction bias are canceled and, in expectation, a more accurate model is produced. □

References

1. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
2. Esponda, I.; Pouzo, D. Berk–Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models. *Econometrica* **2016**, *84*, 1093–1130. [CrossRef]
3. Jiang, B. Approximate Bayesian Computation with Kullback-Leibler Divergence as Data Discrepancy. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, Playa Blanca, Spain, 9–11 April 2018*; Proceedings of Machine Learning Research; Storkey, A., Perez-Cruz, F., Eds.; PMLR: Playa Blanca, Lanzarote, 2018; Volume 84, pp. 1711–1721.
4. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [CrossRef]
5. Pérez-Cruz, F. Kullback-Leibler divergence estimation of continuous distributions. In Proceedings of the 2008 IEEE International Symposium on Information Theory, Toronto, ON, Canada, 6–11 July 2008; pp. 1666–1670.
6. Konishi, S.; Kitagawa, G. Generalised Information Criteria in Model Selection. *Biometrika* **1996**, *83*, 875–890. [CrossRef]
7. Roelofs, R. Measuring Generalization and Overfitting in Machine Learning. Ph.D. Thesis, University of California, Berkeley, CA, USA, 2019.
8. Miller, R.G. The Jackknife—A Review. *Biometrika* **1974**, *61*, 1–15.
9. Firth, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **1993**, *80*, 27–38. [CrossRef]
10. Kosmidis, I. Bias Reduction in Exponential Family Nonlinear Models. Ph.D. Thesis, University of Warwick, Coventry, UK, 2007.
11. Kosmidis, I.; Firth, D. A generic algorithm for reducing bias in parametric estimation. *Electron. J. Stat.* **2010**, *4*, 1097–1112. [CrossRef]
12. Kenne Pagui, E.C.; Salvan, A.; Sartori, N. Median bias reduction of maximum likelihood estimates. *Biometrika* **2017**, *104*, 923–938. [CrossRef]
13. Kosmidis, I. Bias in parametric estimation: Reduction and useful side-effects. *Wiley Interdiscip. Rev. Comput. Stat.* **2014**, *6*, 185–196. [CrossRef]
14. Heskes, T. Bias/Variance Decompositions for Likelihood-Based Estimators. *Neural Comput.* **1998**, *10*, 1425–1433. [CrossRef] [PubMed]
15. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, 2–8 September 1971; Petrov, B.N., Csaki, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267–281.
16. Takeuchi, K. Distribution of information statistics and validity criteria of models. *Math. Sci.* **1976**, *153*, 12–18.
17. Stone, M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 44–47. [CrossRef]
18. Hastie, T.; Tibshirani R.; Friedman J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2009.
19. Ross, A.; Doshi-Velez, F. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
20. Golub, G.H.; Heath, M.; Wahba, G. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* **1979**, *21*, 215–223. [CrossRef]
21. Yanagihara, H.; Tonda, T.; Matsumoto, C. Bias correction of cross-validation criterion based on Kullback–Leibler information under a general condition. *J. Multivar. Anal.* **2006**, *97*, 1965–1975. [CrossRef]
22. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
23. Ward, T. Information Corrected Estimation: A Bias Correcting Parameter Estimation Method, Supplementary Materials, 2021. Available online: https://figshare.com/articles/software/Information_Corrected_Estimation_A_Bias_Correcting_Parameter_Estimation_Method_Supplementary_Materials/14312852/1 (accessed on 2 October 2021).
24. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
25. Broyden, C.G. A Class of Methods for Solving Nonlinear Simultaneous Equations. *Math. Comput.* **1965**, *19*, 577–593. [CrossRef]