Learning to Estimate Without Bias

Tzvi Diskin, Yonina C. Eldar, *Fellow, IEEE*, and Ami Wiesel, *Senior Member, IEEE*

***Abstract*—The Gauss–Markov theorem states that the weighted least squares estimator is a linear minimum variance unbiased estimation (MVUE) in linear models. In this paper, we take a first step towards extending this result to non-linear settings via deep learning with bias constraints. The classical approach to designing non-linear MVUEs is through maximum likelihood estimation (MLE) which often involves computationally challenging optimizations. On the other hand, deep learning methods allow for non-linear estimators with fixed computational complexity. Learning based estimators perform optimally on average with respect to their training set but may suffer from significant bias in other parameters. To avoid this, we propose to add a simple bias constraint to the loss function, resulting in an estimator we refer to as Bias Constrained Estimator (BCE). We prove that this yields asymptotic MVUEs that behave similarly to the classical MLEs and asymptotically attain the Cramer Rao bound. We demonstrate the advantages of our approach in the context of signal to noise ratio estimation as well as covariance estimation. A second motivation to BCE is in applications where multiple estimates of the same unknown are averaged for improved performance. Examples include distributed sensor networks and data augmentation in test-time. In such applications, we show that BCE leads to asymptotically consistent estimators.**

***Index Terms*—Machine learninig, estimation, Cramer Rao bounds.**

I. Introduction

PARAMETER estimation is a fundamental problem in many areas of science and engineering. The goal is to recover an unknown deterministic parameter $\boldsymbol{y}$ given realizations of random variables $\boldsymbol{x}$ whose distribution depends on $\boldsymbol{y}$. An estimator $\hat{\boldsymbol{y}}(\boldsymbol{x})$ is typically designed by minimizing some distance between the observed variables and their statistics, e.g., least squares, maximum likelihood estimation (MLE) or method of moments [1]. Performance is measured in terms of bias, variance and mean squared error (MSE), which depends on the unknowns. For example, MLE is known to be an asymptotic Minimum Variance Unbiased Estimator (MVUE) for any value of $\boldsymbol{y}$. In the last decade, there have been many works suggesting to design estimators using deep learning. Learning based estimators directly minimize the average MSE with respect to a given dataset. Their performance may deteriorate with respect to other values of $\boldsymbol{y}$. To close this gap, the goal of this article is to introduce a machine learning framework for Bias Constrained Estimators (BCE).

The starting point to our work are the classical performance measures in parameter estimation. Statistics define the MSE and the bias of an estimate $\hat{\boldsymbol{y}}(\boldsymbol{x})$ of $\boldsymbol{y}$ as (see for example 2.25 in [2], Chapter 2 in. [1] or [3]):

$$\begin{aligned}\mathrm{MSE}_{\hat{\boldsymbol{y}}}(\boldsymbol{y}) &= \mathrm{E}\left[\|\hat{\boldsymbol{y}}(\boldsymbol{x})-\boldsymbol{y}\|^2\right]\\ \mathrm{BIAS}_{\hat{\boldsymbol{y}}}(\boldsymbol{y}) &= \mathrm{E}\left[\hat{\boldsymbol{y}}(\boldsymbol{x})-\boldsymbol{y}\right].\end{aligned} \tag{1}$$

The "M" in the MSE and in all the expectations in (1) are with respect to to the distribution of $\boldsymbol{x}$ parameterized by $\boldsymbol{y}$. The unknowns are not integrated out and the metrics are functions of $\boldsymbol{y}$ [4]. The goal of parameter estimation is to minimize the MSE for any value of $\boldsymbol{y}$. This problem is ill-defined as different estimators are better for different values of $\boldsymbol{y}$. Therefore, it is standard to focus on minimizing the MSE only among unbiased estimators, that have zero bias for any value of $\boldsymbol{y}$. Fisher information provides a lower bound on this MSE known as the Cramer Rao bound (CRB). An unbiased estimator which achives the CRB is called an MVUE. Remarkably, the popular MLE is often asymptotically MVUE. See [1] for more details on this topic.

A competing framework is Bayesian statistics in which $\boldsymbol{y}$ is modeled as a random vector with a known prior. The goal of Bayesian estimation is also to minimize the MSE. But the definition of MSE is slightly different as the expectation is taken also with respect $\boldsymbol{y}$ (see for example Chapter 10 in [1] or Chapter 2.4 in [2]). To avoid confusion, we refer to this metric as BMSE:

$$\mathrm{BMSE}_{\hat{\boldsymbol{y}}} = \mathrm{E}\left[\mathrm{MSE}_{\hat{\boldsymbol{y}}}(\boldsymbol{y})\right]. \tag{2}$$

Unlike MSE, BMSE is not a function of $\boldsymbol{y}$. BMSE has a well defined minima but it is optimal on average with respect to $\boldsymbol{y}$ and depends on the chosen prior. In practice, this prior is often fictitious and does not really represent any prior knowledge on $\boldsymbol{y}$. Indeed, it is common to use simple uniform or wide Gaussian priors. Partial solutions to this issue include non-informative priors as Jeffrey's prior which require the Fisher information (or CRB) [5] and minimax approaches that are often too pessimistic [6].

In recent years, there is a growing trend of solving classical parameter estimation problems using deep learning. In brief, these methods are approximations to Bayesian methods where the underlying distribution is represented by a training dataset and the optimal solutions are implemented via expressive deep neural networks. This approach is advantageous when it is easy

Manuscript received 17 February 2022; revised 18 January 2023 and 15 May 2023; accepted 3 June 2023. Date of publication 8 June 2023; date of current version 20 June 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bo Chen. This work was supported in part by ISF under Grant 2672/21. *(Corresponding author: Tzvi Diskin.)*

Tzvi Diskin and Ami Wiesel are with The Hebrew University of Jerusalem, Jerusalem 94392, Israel (e-mail: zvidiskin@gmail.com; ami.wiesel@mail.huji.ac.il).

Yonina C. Eldar is with the Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

Digital Object Identifier 10.1109/TSP.2023.3284372

1053-587X © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

to collect or generate data but we do not have access to the exact probabilistic models. It also leads to networks that have fixed computational complexity which are often preferable. Examples include many fields including image reconstruction [7], [8], [9], phase retrieval [10], magnetic resonance imaging [11], frequency estimation [12], [13], Direction of arrival [14], [15], channel estimation [16] and robust regression [17]. Recent works also consider the use of neural networks for efficiently computing the theoretical Fisher information and CRBs [18], [19].

A main challenge in developing deep learning methods for parameter estimation is the choice of prior for $\boldsymbol{y}$. By construction, the learned networks are not optimal for specific values of $\boldsymbol{y}$ but only on average with respect to their training set. The goal of this article is to close this gap by introducing the learned Bias Constrained Estimator (BCE). It minimizes the BMSE while promoting unbiasedness for every value of $\boldsymbol{y}$. We prove that BCE converges to an MVUE for every value of $\boldsymbol{y}$ independently of the prior. Specifically, given a rich enough architecture and a sufficiently large number of samples, BCE is asymptotically unbiased and achieves the lowest possible MSE for any value of $\boldsymbol{y}$.

To gain more intuition into the BCE we consider as a special case the linear settings. Here, the classical Gauss Markov theorem states that the Weighted Least Squares (WLS) estimate is a linear MVUE for any unknown parameter. On the other hand, the linear minimal mean square error estimator (LMMSE) is optimal with respect to a specific Bayesian prior. Both of these can be interpreted as instances of the linear BCE (LBCE). The LBCE has a closed form solution which is a regularized LMMSE with a hyper-parameter λ that reduces the dependency on the prior. LBCE with $\lambda = 0$ reduces to the LMMSE. With a large λ, LBCE converges to the WLS. Generally, BCE provides a flexible bridge between these two extremes, and extends them to non-linear settings.

Numerical experiments support the theory and show that BCE leads to near-MVUEs for all $\boldsymbol{y}$. Estimators based on BMSE alone are better on average but can be worse for specific values of $\boldsymbol{y}$. We demonstrate this in two synthetic settings with a known probabilistic model: signal to noise ratio (SNR) estimation and structured covariance estimation. Next, we illustrate the advantages of BCE in a real world localization problem where we only have access to a training set with no model. The results clearly show that the predictions of BCE are less accurate than its competitors but are unbiased and centered around the ground truth as needed in many applications.

The main purpose of BCE is estimation of deterministic parameters, but we also present an additional use case. Here, the motivation is in the context of averaging estimators in test time. In this setting the goal is to learn a single network that will be applied to multiple inputs and then take their average as the final output. This is the case, for example, in a sensor network where multiple independent measurements of the same phenomena are available. Each sensor applies the network locally and sends its estimate to a fusion center. The global center then uses the average of the estimates as the final estimate [20]. Averaging in test-time has also become standard in image classification where the same network is applied to multiple crops at inference time [21]. In such settings, unbiasedness of the local estimates is a goal on its own, as it is a necessary condition for asymptotic consistency of the global estimate. BCE enforces this condition and improves accuracy of the global estimator even over the BMSE. We demonstrate this advantage over minimizing the MSE of each of the local estimators in the context of image classification on CIFAR10 dataset with test-time data augmentation.

For completeness, we note that BCE is closely related to the topics of "fairness" and "out of distribution (OOD) generalization" which have recently attracted considerable attention in the machine learning literature. The topics are related both in the terminology and in the solutions. Fair learning tries to eliminate biases in the training set and considers properties that need to be protected [22]. OOD works introduce an additional "environment" variable and the goal is to train a model that will generalize well on new unseen environments [23], [24]. Among the proposed solutions are distributionally robust optimization [25] which is a type of minmax estimator, as well as invariant risk minimization [26] and calibration constraints [27], both of which are reminiscent of BCE. A main difference is that in our work the protected properties are the labels themselves. Another core difference is that in parameter estimation we assume full knowledge of the generative model, whereas the above works are purely data-driven and discriminative.

The article is organized as follows. In Section II, we formalize the problem and define the bias constrained estimator. Next, in Section III, we prove that it asymptotically converges to the MVUE. In Section IV, we analyze the linear case to get intuition into the way that BCE works. An additional application of averaging in test time is discussed in Section V. We implement BCE using deep neural networks and demonstrate its performance on different settings in Section VI. Finally, we conclude and discuss some limitations of the work in Section VII.

II. Biased Constrained Estimation

A. Classical Parameter Estimation

Consider a random vector $\boldsymbol{x}$ whose probability distribution is parameterized by an unknown deterministic vector $\boldsymbol{y}$ in some region $S \subset \mathbb{R}^D$. We are interested in the case in which $\boldsymbol{y}$ is a deterministic variable without any prior distribution. We assume exact knowledge of the dependence of $\boldsymbol{x}$ on $\boldsymbol{y}$ via a likelihood function $p(\boldsymbol{x};\boldsymbol{y})$. Typically, this knowledge is based on well specified parametric models, e.g. physics based models. Our goal is to estimate $\boldsymbol{y}$ given $\boldsymbol{x}$.

The quality of an estimator $\hat{\boldsymbol{y}}(\boldsymbol{x})$ is usually measured by the MSE (1) which is a function of the unknown parameter $\boldsymbol{y}$. Minimizing the MSE for any value of $\boldsymbol{y}$ is an ill defined problem. A classical approach to bypass this issue is to consider only unbiased estimators as defined below.

Definition 1: An estimator $\hat{\boldsymbol{y}}(\boldsymbol{x})$ is called unbiased if it satisfies $\text{BIAS}_{\hat{\boldsymbol{y}}}(\boldsymbol{y}) = \boldsymbol{0}$ for all $\boldsymbol{y} \in S$.

Among the unbiased estimators, the MVUE, as defined below, is optimal:

Definition 2: An MVUE is an unbiased estimator $\tilde{\boldsymbol{y}}(\boldsymbol{x})$ that has a variance lower than or equal to that of any other unbiased estimator for all values of $\boldsymbol{y} \in S$.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

An MVUE does not always exists but can be guaranteed under favourable conditions, e.g., simple models in the exponential family [28, p. 88].

In practice, the most common approach to parameter estimation is MLE which is asymptotically near MVUE. It is the solution to the following optimization

$$\hat{\boldsymbol{y}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{x};\boldsymbol{y}). \tag{3}$$

Under favourable conditions, MLE is an MVUE. First, this holds when the observation vector $\boldsymbol{x} = [x_1, \ldots, x_Q]^T$ consists of independent and identically distributed (i.i.d.) elements all conditioned on the same unknown $\boldsymbol{y}$, and $q\mathbf{Q} \to \infty$ along with additional regularity conditions [1, p. 164]. Second, in specific signal in noise problems, MLE is an MVUE even for short data records if the signal to noise ratio (SNR) is high enough. For more details, see Example 7.6 and Problem 7.15 in [1]. Another advantage in these settings is that the performance of MLE attains the CRB: [1]:

$$Q \cdot \mathrm{cov}_{\hat{\boldsymbol{y}}_{\mathrm{MLE}}}(\boldsymbol{y}) \overset{Q\to\infty}{\to} \boldsymbol{F}^{-1}(\boldsymbol{y}) \tag{4}$$

where $\boldsymbol{F}(\boldsymbol{y})$ is the Fisher Information Matrix (FIM)

$$\boldsymbol{F} = \mathrm{E}\left[\left(\frac{\partial \log p(\boldsymbol{x};\boldsymbol{y})}{\partial \boldsymbol{y}}\right)^2\right], \tag{5}$$

where $p(\boldsymbol{x};\boldsymbol{y})$ is the likelihood of a single sample.

A main drawback is that it can be computationally expensive. In inference time, MLE requires the solution of a possibly non-convex and non-linear optimization for each observation vector $\boldsymbol{x}$. In many problems this optimization is intractable or impractical to implement.

B. Deep Learning for Estimation

The recent deep learning revolution has led many to apply machine learning methods for estimation [7], [8], [13], [29], [30], [31]. Deep learning relies on a computationally intensive fitting phase which is done offline, and yields a neural network with fixed complexity that can be easily applied in inference time. Therefore, it is a promising approach for deriving low complexity estimators when MLE is intractable. To avoid confusion, we note that the main motivation to machine learning is usually in data-driven settings, while our main focus is on model-driven settings where deep learning is used to provide low complexity approximations to MLEs and MVUEs.

Deep learning relies on a training dataset, and therefore the first step in using it for parameter estimation is synthetic generation of a dataset

$$\mathcal{D}_N = \{\boldsymbol{y}_i, \boldsymbol{x}_i\}_{i=1}^N. \tag{6}$$

For this purpose, the deterministic $\boldsymbol{y}$ is assumed random with some fictitious prior $p^{\mathrm{fake}}(\boldsymbol{y})$ such that $p^{\mathrm{fake}}(\boldsymbol{y}) \neq 0$ for all $\boldsymbol{y} \in S$. The dataset is then artificially generated according to $p^{\mathrm{fake}}(\boldsymbol{y})$ and the true $p(\boldsymbol{x};\boldsymbol{y})$. Next, a class of possible estimators $\mathcal{H}$ is chosen in order to tradeoff expressive power with computational complexity in test time. In the context of deep learning, the class $\mathcal{H}$ is usually a fixed differentiable neural network architecture.

Algorithm 1: BCE With Synthetic Data.

- Choose a fictitious prior $p^{\mathrm{fake}}(\boldsymbol{y})$.
- Generate N samples $\{\boldsymbol{y}_i\}_{i=1}^N \sim p^{\mathrm{fake}}(\boldsymbol{y})$.
- For each $\boldsymbol{y}_i,\ i = 1, \ldots, N$:
  Generate M samples $\{\boldsymbol{x}_j(\boldsymbol{y}_i)\}_{j=1}^M \sim p(\boldsymbol{x};\boldsymbol{y}_i)$.
- Solve the BCE optimization in (9).

Finally, the learned estimator is defined as the minimizer of the empirical MSE (EMMSE):

$$\mathrm{EMMSE}: \min_{\hat{\boldsymbol{y}}\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^N \left\|\hat{\boldsymbol{y}}(\boldsymbol{x}_i) - \boldsymbol{y}_i\right\|^2. \tag{7}$$

Assuming that $\boldsymbol{y}_i$ are i.i.d. and $N \to \infty$, the objective of EMMSE converges to the BMSE in (2), where the expectation over $\boldsymbol{y}$ is with a respect to the fictitious prior. If $\mathcal{H}$ is sufficiently expressive, EMMSE converges to the Bayesian MMSE estimator which can be expressed as $\hat{\boldsymbol{y}}(\boldsymbol{x}) = \mathrm{E}[\boldsymbol{y}|\boldsymbol{x}]$, [1, p. 346].

The performance of EMMSE estimators is usually promising but depends on the fictitious prior chosen for $\boldsymbol{y}$. In some cases, simple fake priors such as uniform or wide Gaussians are sufficient. But, in some settings, the MSE for specific values of $\boldsymbol{y}$ can be high and unpredictable. See for example the discussion on data generation in [13] where a special purpose generation mechanism was developed to avoid such failures.

C. BCE

The contribution of this article is a competing BCE approach that tries to gain the benefits of both worlds, namely learn a low-complexity deep learning estimator which is near MVUE and is less sensitive to the fictitious prior. For this purpose, we minimize the empirical MSE along with an empirical squared bias regularization. We generate an enhanced dataset

$$\mathcal{D}_{NM} = \left\{\boldsymbol{y}_i, \{\boldsymbol{x}_{ij}\}_{j=1}^M\right\}_{i=1}^N \tag{8}$$

where the measurements $\boldsymbol{x}_{ij}$ are all generated with the same $\boldsymbol{y}_i$. That is, we first generate N samples $\{\boldsymbol{y}_i\}_{i=1}^N$ using the fictions prior and then we generate M samples $\{x_{ij}\}_{j=1}^M$ for each of y_i, using $p(\boldsymbol{x};\boldsymbol{y}_i\}$. BCE is then defined as the solution to:

$$\begin{aligned}\mathrm{BCE}: \min_{\hat{\boldsymbol{y}}\in\mathcal{H}} &\frac{1}{N}\sum_{i=1}^N \left\|\hat{\boldsymbol{y}}(\boldsymbol{x}_i) - \boldsymbol{y}_i\right\|^2 \\ &+ \lambda \frac{1}{N}\sum_{i=1}^N \left\|\frac{1}{M}\sum_{j=1}^M \hat{\boldsymbol{y}}(\boldsymbol{x}_{ij}) - \boldsymbol{y}_i\right\|^2.\end{aligned} \tag{9}$$

where the second term is an empirical squared bias regularization and $\lambda \geq 0$ is a hyperparameter. By choosing $\mathcal{H}$ to be a differentiable neural network architecture as described above, optimizing (9) can be done using standard deep learning tools such as stochastic gradient descent (SGD) [32] and its variants. The only difference from EMMSE is the enhanced dataset and the BCE regularization. The complete method is provided in Algorithm 1 below.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

III. Asymptotic MVUE Analysis

In this section, we show that, under asymptotic conditions and a sufficiently expressive architecture, BCE converges to an MVUE and behaves like MLE.

Theorem 1: Assume that

1) An MVUE denoted by $\tilde{\boldsymbol{y}}$ exists within $\mathcal{H}$.
2) The fictitious prior is non-singular, i.e., $p^{\text{fake}}(\boldsymbol{y}) \neq 0$ for all $\boldsymbol{y} \in S$.
3) The variance of the MVUE estimate under the fictitious prior is finite: $\int \text{VAR}_{\tilde{\boldsymbol{y}}}(\boldsymbol{y}) p^{\text{fake}}(\boldsymbol{y}) d\boldsymbol{y} < \infty$.

Then, BCE converges to MVUE for sufficiently large λ, M and N.

If an MVUE exists in the problem, the first assumption can be met by choosing a sufficiently expressive class of estimators. The second and third assumptions are technical and are less important in practice.

Proof: The main idea of the proof is that because the squared bias is non negative, it is equal to zero for all $\boldsymbol{y}$ if and only if its expectation over any non-singular prior is zero. Thus, taking λ to infinity enforces a solution that is unbiased for any value of $\boldsymbol{y}$. Among the unbiased solutions, only the MSE term in the BCE is left and thus the solution is the MVUE.

For sufficiently large M and N, the objective of (9) converges to its population form:

$$L_{\text{BCE}} = \text{E}\left[\|\hat{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}\|^2\right] + \lambda \text{E}\left[\|\text{E}\left[\hat{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}|\boldsymbol{y}\right]\|^2\right]. \tag{10}$$

The MVUE satisfies

$$\text{E}\left[\tilde{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}|\boldsymbol{y}\right] = 0, \qquad \forall\, \boldsymbol{y} \in S.$$

Thus

$$\text{E}\left[\|\text{E}\left[\tilde{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}|\boldsymbol{y}\right]\|^2\right] = 0.$$

Now we show that the BCE objective (10) of the MVUE $\tilde{\boldsymbol{y}}(\boldsymbol{x})$ is smaller both from any biased solution $\hat{\boldsymbol{y}}_1(\boldsymbol{x})$ and any unbiased solution $\hat{\boldsymbol{y}}_2(\boldsymbol{x})$. First, assume that the BCE $\hat{\boldsymbol{y}}_1(\boldsymbol{x})$ is biased for some $\boldsymbol{y} \in S$:

$$\text{E}\left[\hat{\boldsymbol{y}}_1(\boldsymbol{x}) - \boldsymbol{y}|\boldsymbol{y}\right] \neq 0.$$

Thus

$$\text{E}\left[\|\text{E}\left[\hat{\boldsymbol{y}}_1(\boldsymbol{x}) - \boldsymbol{y}|\boldsymbol{y}\right]\|^2\right] > 0.$$

Since $\text{E}[\|\tilde{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}\|^2]$ is finite, this means that for sufficiently large λ we have a contradiction

$$L_{\text{BCE}}(\tilde{\boldsymbol{y}}) < L_{\text{BCE}}(\hat{\boldsymbol{y}}_1).$$

Secondly, assume that the BCE $\hat{\boldsymbol{y}}_2$ is an unbiased estimator. By the definition of MVUE, for all $\boldsymbol{y} \in S$

$$\text{E}\left[\|\tilde{\boldsymbol{y}}(\boldsymbol{x}) - \text{E}\left[\tilde{\boldsymbol{y}}|\boldsymbol{y}\right]\|^2 |\boldsymbol{y}\right] \leq \text{E}\left[\|\hat{\boldsymbol{y}}_2(\boldsymbol{x}) - \text{E}\left[\hat{\boldsymbol{y}}_2|\boldsymbol{y}\right]\|^2 |\boldsymbol{y}\right]$$

and

$$\text{E}\left[\|\tilde{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}\|^2 |\boldsymbol{y}\right] \leq \text{E}\left[\|\hat{\boldsymbol{y}}_2(\boldsymbol{x}) - \boldsymbol{y}\|^2 |\boldsymbol{y}\right].$$

Thus

$$L_{\text{BCE}}(\tilde{\boldsymbol{y}}) = \text{E}\left[\|\tilde{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{y}\|^2\right] \leq \text{E}\left[\|\hat{\boldsymbol{y}}_2(\boldsymbol{x}) - \boldsymbol{y}\|^2\right] = L_{\text{BCE}}(\hat{\boldsymbol{y}}_2).$$

Together, the BCE loss of the MVUE is smaller than the BCE loss of any other estimator whether it is biased or not, completing the proof. □

To summarize, EMMSE (7) is optimal on average with respect to the fictitious prior. Otherwise, it can perform badly (see experiments in Section VI). On the other hand, if an MVUE for the problem exists, BCE is optimal for any value of $\boldsymbol{y}$ among the unbiased estimators, and is independent on the choice of the fictitious prior. In problems where a statistically efficient estimator exists, BCE achieves the Cramer Rao Bound (CRB), making its performance both optimal and predictable for any value of $\boldsymbol{y}$.

IV. Linear BCE

In this section we focus on the linear case in which the BCE has a closed form. The section is divided into two parts. In the first part we restrict the estimator to be linear where in the second part we assume that the likelihood model is also linear.

First we restrict attention to the class of linear estimators which is easy to fit, store and implement. Note that using non linear features, this is applicable to any likelihood model and can be very expressive using random features [33] or deep features (by replacing the last layer of a DNN, see for example in [34], [35]). For simplicity, we also assume perfect training with $N, M \to \infty$ so that the empirical means converge to their true expectations. The following theorem gives a closed form solution to the linear BCE (LBCE):

Theorem 2: Consider the the BCE optimization problem (9) such that the estimator $\hat{\boldsymbol{y}}$ is restricted to the linear hypotheses class $\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{x}$ Then the solution is given by:

$$\boldsymbol{A} = \text{E}\left[\boldsymbol{y}\boldsymbol{x}^T\right]\left(\frac{1}{\lambda+1}\text{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] + \frac{\lambda}{\lambda+1}\mathbf{R}\right)^{-1} \tag{11}$$

where

$$\mathbf{R} = \text{E}\left[\text{E}\left[\boldsymbol{x}|\boldsymbol{y}\right]\text{E}\left[\boldsymbol{x}^T|\boldsymbol{y}\right]\right], \tag{12}$$

and we assume all are invertible. All the expectations over $\boldsymbol{y}$ are with respect to the fictitious prior.

The proof is straight forward by plugging the linear function into (9), and taking the derivative to zero. The full derivation is in Appendix A.

Plugging in $\lambda = 0$ in (11) yields the well known LMMSE [1, p. 380]:

$$\boldsymbol{A} \overset{\lambda=0}{\to} \text{E}\left[\boldsymbol{y}\boldsymbol{x}^T\right]\left(\text{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right]\right)^{-1}. \tag{13}$$

Just like LMMSE, the BCE in (11) holds for arbitrary distributions. It can account for highly non-linear relations between $\boldsymbol{y}$ and $\boldsymbol{x}$ by using pre-defined non-linear features of $\boldsymbol{x}$.

Next, we assume that the statistical relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is also linear plus additive zero mean noise. The solution to

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

this setting is well known as the Gauss Markov theorem. As expected, BCE agrees with this special case.

Theorem 3: Consider the linear case where

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{H}\boldsymbol{y} + \boldsymbol{n} \\ \hat{\boldsymbol{y}} &= \boldsymbol{A}\boldsymbol{x} \end{aligned} \tag{14}$$

where $\boldsymbol{n}$ is a zero mean random vector with covariance $\boldsymbol{\Sigma_n}$ and $\boldsymbol{y}$ is a parameter vector with a fictions prior with zero mean and covariance $\boldsymbol{\Sigma_y}$. For $N, M \to \infty$, LBCE reduces to

$$\boldsymbol{A} = \left(\boldsymbol{H}^T \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1} \boldsymbol{H} + \frac{1}{\lambda + 1} \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} \right)^{-1} \boldsymbol{H}^T \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1} \tag{15}$$

and we assume all are invertible. All the expectations over $\boldsymbol{y}$ are with respect to the fictitious prior.

The proof is simple and is based on plugging (14) into (11) and using the matrix inversion lemma.

Equivalently, (15) become:

$$\boldsymbol{A} = \left(\boldsymbol{H}^T \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1} \boldsymbol{H} + \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} \right)^{-1} \boldsymbol{H}^T \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1} \tag{16}$$

which is known as Bayesian linear regression [36, p. 153]. On the other hand, Plugging $\lambda \to \infty$ in (15) yields the seminal weighted least squares (WLS) estimator [1, p. 226]:

$$\boldsymbol{A} \overset{\lambda \to \infty}{\to} \left(\boldsymbol{H}^T \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1} \boldsymbol{H} \right)^{-1} \boldsymbol{H}^T \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1}. \tag{17}$$

For finite $\lambda > 0$, linear BCE provides a flexible bridge between the LMMSE and WLS estimators. The parameter λ attenuates the effect of the fictitious prior by dividing the prior covariance in (15) by $(\lambda + 1)$.

V. BCE for Averaging

The main motivation for BCE is approximating the MVUE using deep neural networks. In this section, we consider a different motivation to BCE where unbiasedness is a goal on its own. Specifically, BCEs are advantageous in scenarios where multiple estimators are combined together for improved performance. Typical examples include sensor networks where each sensor provides a local estimate of the unknown, or computer vision applications where the same network is applied on different crops of a given image to improve accuracy [21], [37].

The underlying assumption in BCE for averaging is that we have access, both in training and in test times, to multiple $\boldsymbol{x}_{ij}$ associate with the same $\boldsymbol{y}_i$. This assumption holds in learning with synthetic data (e.g., Algorithm 1), or real world data with multiple views or augmentations. In any case, the data structure allows us to learn a single network $\hat{\boldsymbol{y}}(\cdot)$ which will be applied to each of them and then output their average as summarized in Algorithm 2. The following theorem then proves that BCE results in asymptotic consistency.

Theorem 4: Assume that an unbiased estimator $\boldsymbol{y}^*(\boldsymbol{x})$ with finite variance exists within the hypothesis class, and that a BCE was learned with sufficiently large λ, M and N. Consider the case in which M_t independent and identically distributed (i.i.d.) measurements $\boldsymbol{x}_j$ of the same $\boldsymbol{y}$ are available in test time. Then,

Algorithm 2: BCE for Averaging.

- Let $\hat{\boldsymbol{y}}(\cdot)$ be the solution to BCE in (9).
- Define the global estimator as

$$\overline{\boldsymbol{y}}(\boldsymbol{x}) = \frac{1}{M_t} \sum_{j=1}^{M_t} \hat{\boldsymbol{y}}(\boldsymbol{x}_j), \tag{18}$$

where M_t is the number of local estimators at test time.

the average BCE in (18) is asymptotically consistent when M_t increases.

Proof: Following the proof of theorem 1, BCE with a sufficiently large λ, M and N results in a unbiased estimator, if one exists within the hypothesis class. The global metrics satisfy:

$$\begin{aligned} \mathrm{BIAS}_{\overline{\boldsymbol{y}}}(\boldsymbol{y}) &= \frac{1}{M} \sum_{j=1}^{M} \mathrm{BIAS}_{\hat{\boldsymbol{y}}}(\boldsymbol{y}) = \mathrm{BIAS}_{\hat{\boldsymbol{y}}}(\boldsymbol{y}) \\ \mathrm{VAR}_{\overline{\boldsymbol{y}}}(\boldsymbol{y}) &= \frac{1}{M^2} \sum_{j=1}^{M} \mathrm{VAR}_{\hat{\boldsymbol{y}}}(\boldsymbol{y}) = \frac{1}{M} \mathrm{VAR}_{\hat{\boldsymbol{y}}}(\boldsymbol{y}) \overset{M \to \infty}{\to} 0. \end{aligned} \tag{19}$$

Thus, the global variance decreases with M_t, whereas the global bias remains constant, and for an unbiased local estimator it is equal to zero. □

VI. Experiments

In this section, we present numerical experiments results. We focus on the main ideas and conclusions and leave the details to the appendix. Reproducible code with all the implementation details including architectures and hyperparameters will be provided at https://github.com/tzvid/BCE.

A. SNR Estimation

Our first experiment addresses a non-convex estimation problem of a single unknown. The unknown is scalar and therefore we can easily compute the MLE and visualize its performance. Specifically, we consider non-data-aided SNR estimation [38]. The received signal is

$$x_l = a_l h + w_l \tag{20}$$

where $a_l = \pm 1$ are equi-probable binary symbols, h is an unknown signal and w_l is a white Gaussian noise with unknown variance denoted by σ^2 and $l = 1, \ldots, Q$. The goal is to estimate the SNR defined as $y = \frac{h^2}{\sigma^2}$. Different estimators were proposed for this problem, including MLE [39] and method of moments [40]. For our purposes, we compare the MLE with two identical networks trained on synthetic data using EMMSE and BCE loss functions.

Fig. 1 shows the MSE and the bias of the different estimators as a function of the SNR. It is evident that BCE is a better approximation of MLE than EMMSE. EMMSE is very biased towards a narrow regime of the SNR. This is because the MSE scales as the square of the SNR and the average MSE loss is dominated by the large MSE examples. For completeness, we also

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

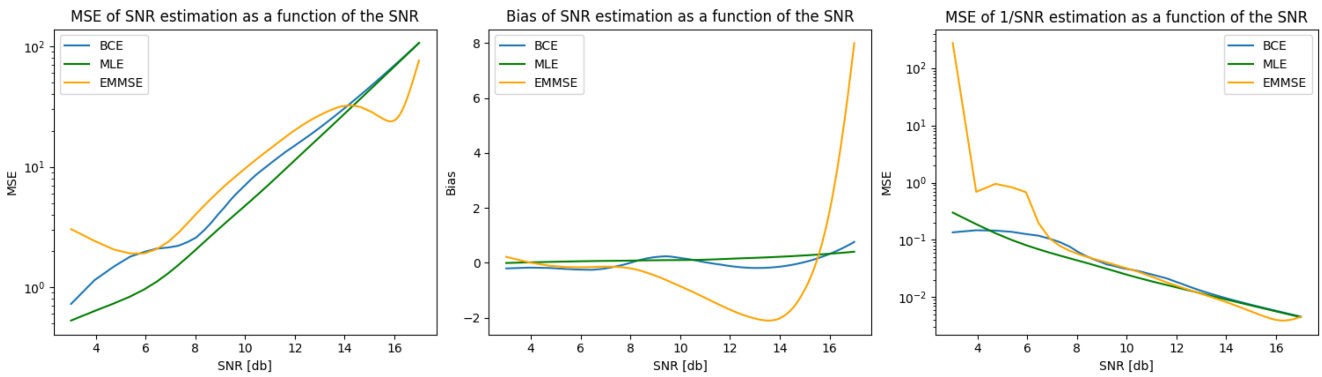

Fig. 1. Bias of SNR, MSE of SNR and MSE of inverse SNR.

plot the MSE in terms of inverse SNR defined as $\mathrm{E}[(\frac{1}{\hat{y}} - \frac{1}{y})^2]$. Functional invariance is a well known and attractive property of MLE. The figure shows that both MLE and BCE are robust to the inverse transformation, whereas EMMSE is unstable and performs poorly in low SNR.

B. Structured Covariance Estimation

Our second experiment considers a more interesting multivariate structured covariance estimation. Specifically, we consider the estimation of a sparse covariance matrix [41]. The measurement model is

$$p(\boldsymbol{x}; \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{y} = [y^{(1)}, \ldots, y^{(9)}]^T$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1+y^{(1)} & 0 & 0 & \frac{1}{2}y^{(6)} & 0 \\ 0 & 1+y^{(2)} & 0 & \frac{1}{2}y^{(7)} & 0 \\ 0 & 0 & 1+y^{(3)} & 0 & \frac{1}{2}y^{(8)} \\ \frac{1}{2}y^{(6)} & \frac{1}{2}y^{(7)} & 0 & 1+y^{(4)} & \frac{1}{2}y^{(9)} \\ 0 & 0 & \frac{1}{2}y^{(8)} & \frac{1}{2}y^{(9)} & 1+y^{(5)} \end{pmatrix} \tag{21}$$

and $0 \leq y^{(k)} \leq 1$ are unknown parameters. We train a neural network using $p^{\text{fake}}(y^{(k)}) \sim \mathcal{U}(0,1)$ using both EMMSE and BCE. Computing the MLE is non-trivial in these settings and we compare the performance to the theoretical asymptotic variance defined by the CRB [1]. The CRB depends on the specific values of $\boldsymbol{y}$. Therefore we take random realizations and provide scatter plots ordered according to the CRB value. Finally, as another baseline, we also report the result of an additional model named NORM. Following [42], NORM is trained with MSE normalized by the CRB. The idea is that the loss of each sample will be normalized by its "difficulty", and the performance will not be govern by the difficult examples.

Fig. 2 presents the results of two simulations. In the first, we generate data according to the training distribution $\boldsymbol{y} \sim \mathcal{U}(0,1)$. As expected, EMMSE which was optimized for this distribution provides the best MSEs. In the second, we generate test data according to a different distribution $\boldsymbol{y} \sim \mathcal{U}(0,0.1)$. Here the performance of EMMSE significantly deteriorates. NORM performs better in this out of distribution setting but is still far from the CRB. In contrast, in both settings BCE is near MVUE and provides MSEs close to the CRB while ignoring the prior distribution used in training.

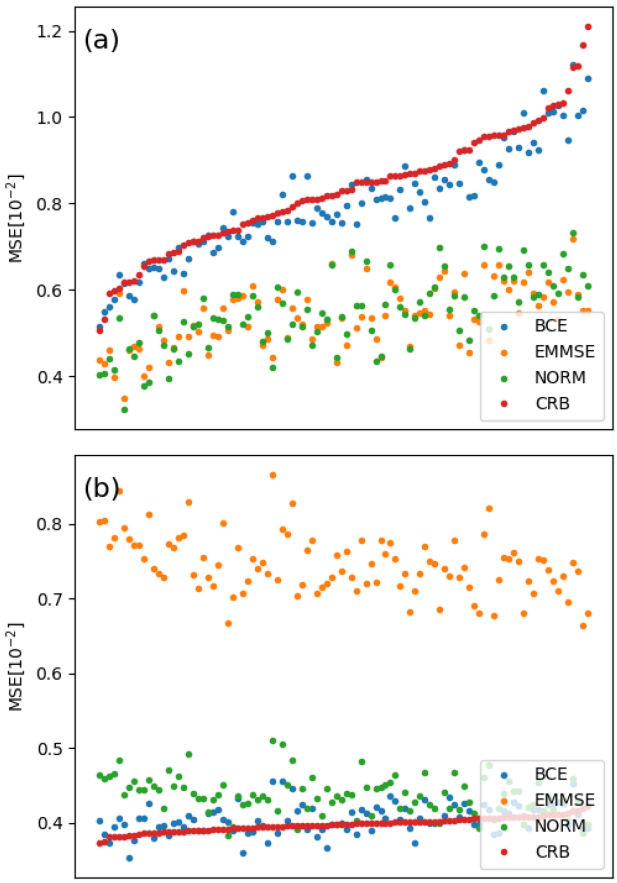

Fig. 2. Scatter plot of MSEs ordered by CRB values. In (a) the tested $\boldsymbol{y}$'s are generated from the training distribution. In (b) the test distribution is different. BCE is near-MVUE and close to the CRB for both distributions, whereas EMMSE is better in (a) and weaker in (b).

Note that BCE outperforms the NORM baseline and is also significantly easier to implement. NORM requires exact knowledge of the underlying model and the corresponding CRB. The main advantage of BCE is that it can be easily computed in a data-driven manner without access to the underlying model. The next example illustrates such real world settings.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

TABLE I
MSE AND BIAS OF EMSSE AND BCE FOR RSSI LOCALIZATION

METRIC/MODEL	MSE	BIAS^2
EMMSE	1.5	0.7
BCE	2.2	0.4

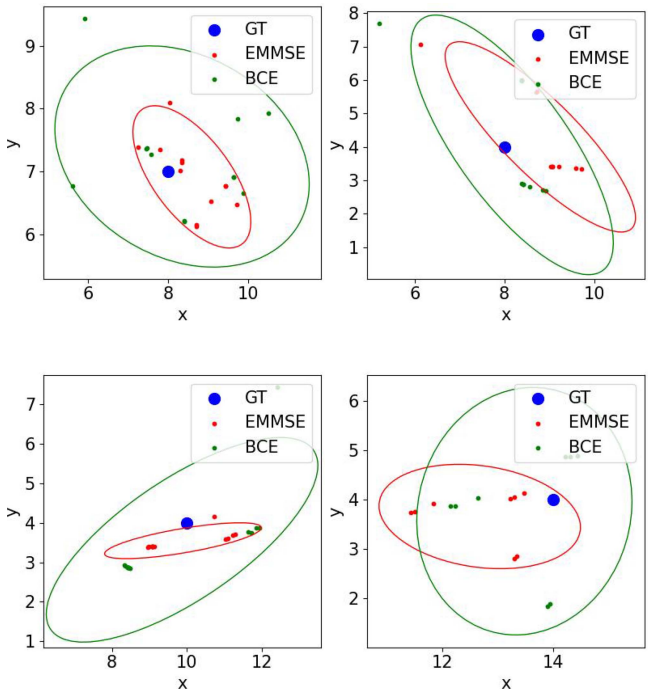

Fig. 3. RSSI localization: Different estimates of EMMSE and BCE for the same location, for 4 different locations. While the estimations of EMMSE has smaller variance, the center of BCE estimations is closer to the true location.

C. RSII Localization

Our third experiment considers BCE of unbiased localization using RSSI on real data. Specifically, we used the BLE RSSI Dataset for Indoor localization and Navigation Data Set from [43]. The dataset contains 1420 labeled examples of RSSI readings of an array of 13 beacons as input and the xy location as outputs. We used two thirds of the data for training and a third for test. We evaluate the MSE and the average squared bias of EMMSE and BCE. To estimate the average squared bias, we use examples with the same true location. We only evaluate locations with at least 8 different examples. Table I summarizes the results, showing that while EMMSE is better in terms of MSE, BCE results in a smaller bias. To gain more intuition, Fig 3 illustrates this difference on a few selected examples. For each example, we draw the ground truth location (GT) and the model predictions using the different inputs. We also plot uncertainty ellipses that correspond to two standard deviations. It is easy to see that while the EMMSE has smaller variance, the center of BCE estimations is closer to the true location.

D. Image Classification With Soft Labels

Our fourth experiment considers BCE for averaging in the context of image classification. We consider the popular CIFAR10 dataset. BCE is designed for regression rather than classification. Therefore we consider soft labels as proposed in the knowledge distillation technique [44]. The soft labels are obtained using a strong "teacher" network from [45]. To exploit the benefits of averaging we rely on data augmentation in the training and test phases [21]. For augmentation, we use random cropping and flipping. We train two small "student" convolution networks with identical architectures using the EMMSE loss and the BCE loss. More precisely, following other distillation works, we also add a hard valued cross entropy term to the losses.

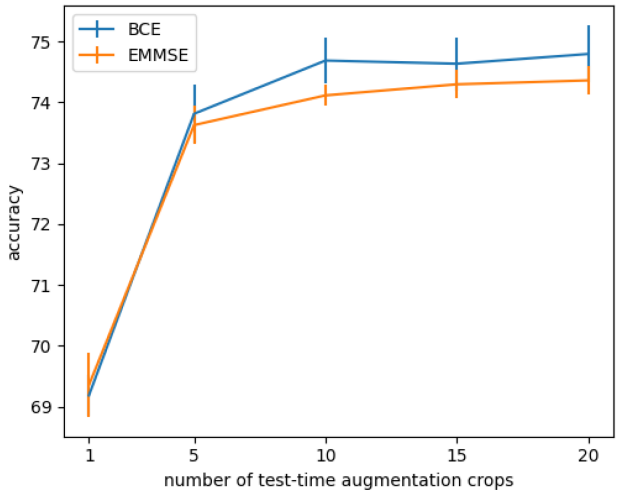

Fig. 4. BCE for averaging: accuracy as a function of the number of test-time augmentation crops.

Fig. 4 compares the accuracy of EMMSE vs BCE as a function of the number of test-time data augmentation crops. It can be seen that while on a single crop EMMSE achieves a slightly better accuracy, BCE achieves better results when averaged over many crops in the test phase.

VII. SUMMARY

In recent years, deep neural networks are replacing classical algorithms for estimation in many fields. While deep neural networks give remarkable improvement in performance "on average", in some situations one would prefer to use classical frequentist algorithms that have guarantees on their performance on any value of the unknown parameters. In this work we show that when a statistically efficient estimator exists, deep neural networks can be used to learn it using a bias constrained loss, provided that the architecture is expressive enough. BCE asymptotically converges to an MVUE but there are a few important caveats worth mentioning. First, in general problems, an MVUE does not always exist and the results only hold asymptotically. Just like MLE, BCE only attains the CRB when the number of samples is large and the errors are small. Otherwise, bias acts as a regularization mechanism which is typically preferable. Second, in many problems choosing a uniform or sufficiently wide prior leads to near MVUEs. BCE is mostly needed when the MSE changes significantly with the unknown parameters. In

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

particular, two of our experiments involved variance estimation where the MSE scales quadratically with the unknown variance.

APPENDIX A
PROOFS OF THEOREMS 2–3

A. Proof of Theorem 2

We insert $\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{x}$ in (9). Taking $M, N \to \infty$ gives the true expectations and thus:

$$\begin{aligned}\mathrm{BCE} &= \mathrm{MSE} + \lambda \mathrm{BIAS}^2 \\ &= \mathrm{E}\left[\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2\right] + \lambda \mathrm{E}\left[\|\mathrm{E}\left[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}|\boldsymbol{y}\right]\|^2\right]. \end{aligned} \tag{22}$$

Now the MSE term is equal to:

$$\begin{aligned}\mathrm{MSE} &= \mathrm{E}\left[\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2\right] \\ &= \mathrm{E}\left[\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{y}\right] \\ &= \mathrm{Tr}\left(\boldsymbol{A}\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right]\boldsymbol{A}^T - 2\boldsymbol{A}\mathrm{E}\left[\boldsymbol{x}\boldsymbol{y}^T\right] + \mathrm{E}\left[\boldsymbol{y}\boldsymbol{y}^T\right]\right), \end{aligned} \tag{23}$$

and the BIAS term is:

$$\begin{aligned}\mathrm{BIAS} &= \mathrm{E}\left[\|\mathrm{E}\left[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}|\boldsymbol{y}\right]\|^2\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[\boldsymbol{x}^T\boldsymbol{A}^T|\boldsymbol{y}\right]\mathrm{E}\left[\boldsymbol{A}\boldsymbol{x}|\boldsymbol{y}\right] - \mathrm{E}\left[\boldsymbol{x}^T\boldsymbol{A}^T|\boldsymbol{y}\right]\boldsymbol{y}\right] \\ &\quad - \mathrm{E}\left[\boldsymbol{y}^T\mathrm{E}\left[\boldsymbol{A}\boldsymbol{x}|\boldsymbol{y}\right] + \boldsymbol{y}^T\boldsymbol{y}\right] \\ &= \mathrm{Tr}\left(\boldsymbol{A}\mathbf{R}\boldsymbol{A}^T - 2\boldsymbol{A}\mathrm{E}\left[\mathrm{E}\left[\boldsymbol{x}|\boldsymbol{y}\right]\boldsymbol{y}^T\right]\right) \\ &= \mathrm{Tr}\left(\boldsymbol{A}\mathbf{R}\boldsymbol{A}^T - 2\boldsymbol{A}\mathrm{E}\left[\boldsymbol{x}\boldsymbol{y}^T\right]\right). \end{aligned} \tag{24}$$

Thus:

$$\begin{aligned}\mathrm{BCE} &= \mathrm{Tr}\left(\boldsymbol{A}\left(\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] + \lambda\mathbf{R}\right)\boldsymbol{A}^T\right) \\ &\quad - 2\left(\lambda+1\right)\mathrm{Tr}\left(\boldsymbol{A}\mathrm{E}\left[\boldsymbol{x}\boldsymbol{y}^T\right]\right) + \mathrm{E}\left[\boldsymbol{y}^T\boldsymbol{y}\right]. \end{aligned} \tag{25}$$

Taking the derivative with respect to $\boldsymbol{A}$ and equating to zero:

$$\frac{\partial \mathrm{BCE}}{\partial \boldsymbol{A}} = 2\left(\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] + \lambda\mathbf{R}\right)\boldsymbol{A}^T - 2\left(\lambda+1\right)\mathrm{E}\left[\boldsymbol{x}\boldsymbol{y}^T\right]. \tag{26}$$

Finally,

$$\boldsymbol{A} = \mathrm{E}\left[\boldsymbol{y}\boldsymbol{x}^T\right]\left(\frac{1}{\lambda+1}\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] + \frac{\lambda}{\lambda+1}\mathbf{R}\right)^{-1}. \tag{27}$$

B. Proof of Theorem 3

Using $\boldsymbol{x} = \boldsymbol{H}\boldsymbol{y} + \boldsymbol{n}$, we obtain:

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{y}\boldsymbol{x}^T\right] &= \mathrm{E}\left[\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{H}^T + \boldsymbol{y}\boldsymbol{n}^T\right] = \boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{H}^T \\ \mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] &= \mathrm{E}\left[\boldsymbol{H}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{H}^T + \boldsymbol{n}\boldsymbol{y}\boldsymbol{H}^T + \boldsymbol{H}\boldsymbol{y}\boldsymbol{n}^T + \boldsymbol{n}\boldsymbol{n}^T\right] \\ &= \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{n}}, \end{aligned} \tag{28}$$

and

$$\mathbf{R} = \mathrm{E}\left[\mathrm{E}\left[\boldsymbol{H}\boldsymbol{y} + \boldsymbol{n}|\boldsymbol{y}\right]\mathrm{E}\left[\boldsymbol{H}\boldsymbol{y} + \boldsymbol{n}|\boldsymbol{y}\right]^T\right] = \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{H}^T. \tag{29}$$

Plugging (28) and (29) in (11) gives:

$$\begin{aligned}\boldsymbol{A} &= \boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{H}^T + \frac{1}{\lambda+1}\boldsymbol{\Sigma}_{\boldsymbol{n}}\right)^{-1} \\ &= \left(\boldsymbol{H}^T\boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1}\boldsymbol{H} + \frac{1}{\lambda+1}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\right)^{-1}\boldsymbol{H}^T\boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1}, \end{aligned} \tag{30}$$

where the last step obtained using the matrix inversion lemma, completing the proof.

APPENDIX B
IMPLEMENTATION DETAILS

A. Implementation Details for Section VI-A

We train a simple fully connected model with one hidden layer. First the data is normalized by the second moment and then the input is augmented by hand crafted features: the fourth and sixth moments and different functions of them. We train the network using $Q = 50$ synthetic data in which the mean is sampled uniformly in [1, 10] and then SNR is sampled uniformly in [2, 50] (which corresponds to [3 dB, 16 dB]). The data is generated independently in each batch. We trained the model using the standard MSE loss, and using BCE with $\lambda = 1000$. We use ADAM solver with a multistep learning scheduler. We use batch sizes of $N = 10$ and $M = 100$ as defined (9).

B. Implementation Details for Section VI-B

We train a neural network for estimation the covariance matrix with the following architecture: First the sample covariance $\mathbf{C}_0$ is calculated from the input $\boldsymbol{x}$ as it is the sufficient statistic for the covariance in Gaussian distribution. Also a vector $\boldsymbol{\alpha}_{k=0}$ is initialized to $\boldsymbol{\alpha}_{k=0} = \frac{1}{2}\mathbf{1}_9$ and a vector $\boldsymbol{v}_{k=0}$ is initialized to zero. Next, a one hidden layer fully connected network with concatenated input of $\mathbf{C}_0$ and $\boldsymbol{v}_{k=0}$ is used to predict a modification $\Delta\boldsymbol{\alpha}$ and $\Delta\boldsymbol{v}$ for the vectors $\boldsymbol{\alpha}_k$ and $\boldsymbol{v}_k$ respectively, such that $\boldsymbol{\alpha}_{k+1} = \boldsymbol{\alpha}_k + 0.1\Delta\boldsymbol{\alpha}$ and similarly for $\boldsymbol{v}_{k+1}$. Then an updated covariance $\mathbf{C}_{k+1}$ is calculated using $\boldsymbol{\alpha}_{k+1}$ and (21). The process is repeated (with the updated $\mathbf{C}_k$ and $\boldsymbol{v}_k$ as an input to the fully connected network) for 50 iterations. The final covariance is the output of the network. The network in is tranied on synthetic data in which the covariance the paramaters of the covariance are generated uniformly in their valid region and then M different X's are generated from a normal distribution with a zero mean and the generated covariance. We use an ADAM solver with a "ReduceLROnPlateau" scheduler with the desired loss (BCE loss of BCE and MSE loss for EMMSE) on a synthetic validation set. We trained the model using the standard MSE loss, and using BCE with $\lambda = 1000$. We use batch sizes of $N = 1$ and $M = 20$ as defined in (9).

For the training of NORM, we use the loss:

$$\sum_{i,j}\left(\hat{\boldsymbol{y}}(\boldsymbol{x}_{ij}) - \boldsymbol{y}_i\right)^T\boldsymbol{F}(\boldsymbol{y}_i)\left(\hat{\boldsymbol{y}}(\boldsymbol{x}_{ij}) - \boldsymbol{y}_i\right) \tag{31}$$

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

C. Implementation Details for Section VI-C

We use a fully connected neural network with a single hidden layer of size 20. The data (both the inputs and outputs) was normalized to have zero mean and unit variance at each dimension. For the bias term of the BCE loss, at each batch we uniformly sample a location and take the average of the output of the model for all the examples of the same location.

D. Implementation details for Section VI-D

We generate soft labels using a "teacher" network. Specifically, we work on the CIFAR10 dataset, and use a DLA architecture [45] which achieves 0.95 accuracy as a teacher. Our student network is a very small convolutional neural network (CNN) with two convolutions layers with 16 and 32 channels respectively and a single fully connected layer. We now use the following notations: The original dataset is a set of N triplets of images $\boldsymbol{x}_i$, a one-hot vector of hard labels $\boldsymbol{y}_i^h$ and a vector of soft labels $\boldsymbol{y}_i^s$ $\{\boldsymbol{x}_i, \boldsymbol{y}_i^h, \boldsymbol{y}_i^s\}$. In the augmented data, M different images $\boldsymbol{x}_{ij}$ are generated randomly from each original image $\boldsymbol{x}_i$ using random cropping and flipping. The output of the network for the class l is denoted by z_{ij}^l and the vector of "probabilities" $\boldsymbol{q}_{ij}(T)$ is defined by:

$$q_{ij}^l(T) = \frac{\exp(q_{ij}^l/T)}{\sum_l \exp(z_{ij}^l/T)} \tag{32}$$

where T is a "temperature" that controls the softness of the probabilities. We define the following loss functions:

$$\begin{aligned} L_{hard} &= \sum_{ij}^{MN} CE(\boldsymbol{q}_{ij}(T=1), \boldsymbol{y}_i^h) \\ L_{MSE} &= \sum_{ij}^{MN} \left\|\boldsymbol{q}_{ij}(T=20) - \boldsymbol{y}_i^s\right\|^2 \\ L_{bias} &= \sum_{i}^{N} \left\| \sum_{j}^{M} \boldsymbol{q}_{ij}(T=20) - \boldsymbol{y}_i^s \right\|^2 \end{aligned} \tag{33}$$

where CE is the cross-entropy loss. The regular network and the BCE are trained with the following losses:

$$\begin{aligned} L_{EMMSE} &= L_{hard} + L_{MSE} \\ L_{BCE} &= L_{hard} + L_{bias}, \end{aligned} \tag{34}$$

using stochastic gradient decent (SGD). In training, we use 20 different random crops and flips. We test the trained network in five different checkpoints during training and calculate the average and the standard deviation. In test-time we use the same data augmentation as in the training, the scores of the different crops are averaged to get the final score as in Algorithm 2. Fig. 4 shows the average and the standard deviation of the accuracy as a function of the number of crops in the training set.

REFERENCES

[1] S. M. Kay and S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, vol. 1. Englewood Cliffs, NJ, USA:Prentice-Hall, 1993.

[2] J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Germany: Springer, 2017.

[3] E. Lehmann and H. Scheffe, "Completeness, similar regions, and unbiased estimation," *Bull. Amer. Math. Soc.*, vol. 54, no. 11, pp. 1080–1080, 1948.

[4] Y. C. Eldar, *Rethinking Biased Estimation: Improving Maximum Likelihood and the Cramér-Rao Bound*. Boston, MA, USA: Now Inc., 2008.

[5] R. E. Kass and L. Wasserman, "The selection of prior distributions by formal rules," *J. Amer. Stat. Assoc.*, vol. 91, no. 435, pp. 1343–1370, 1996.

[6] Y. C. Eldar and N. Merhav, "Minimax MSE-ratio estimation with signal covariance uncertainties," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1335–1347, Apr. 2005.

[7] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 39–56, May 2020.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[9] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proc. IEEE*, vol. 111, no. 5, pp. 465–499, May 2023.

[10] N. Naimipour, S. Khobahi, and M. Soltanalian, "Unfolded algorithms for deep phase retrieval," 2020, *arXiv:2012.11102*.

[11] J. Schlemper et al., "Stochastic deep compressive sensing for the reconstruction of diffusion tensor cardiac MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2018, pp. 295–303.

[12] G. Izacard, S. Mohan, and C. Fernandez-Granda, "Data-driven estimation of sinusoid frequencies," 2019, *arXiv:1906.00823*.

[13] R. Dreifuerst and R. W. Heath Jr, "SignalNet: A low resolution sinusoid decomposition and estimation network," 2021, *arXiv:2106.05490*.

[14] J. P. Merkofer, G. Revach, N. Shlezinger, and R. J. van Sloun, "Deep augmented music algorithm for data-driven DoA estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 3598–3602.

[15] D. H. Shmuel, J. P. Merkofer, G. Revach, R. J. van Sloun, and N. Shlezinger, "Deep root music algorithm for data-driven DoA estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[16] B. Fesl, N. Turan, and W. Utschick, "Low-rank structured MMSE channel estimation with mixtures of factor analyzers," 2023, *arXiv:2304.14809*.

[17] T. Diskin, G. Draskovic, F. Pascal, and A. Wiesel, "Deep robust regression," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, 2017, pp. 1–5.

[18] H. V. Habi, H. Messer, and Y. Bresler, "Learning to bound: A generative Cramér-Rao bound," *IEEE Trans. Signal Process.*, vol. 71, pp. 1216–1231, 2023.

[19] T. T. Duy, L. V. Nguyen, V.-D. Nguyen, N. L. Trung, and K. Abed-Meraim, "Fisher information neural estimation," in *Proc. 30th Eur. Signal Process. Conf.*, 2022, pp. 2111–2115.

[20] J. Li and G. AlRegib, "Distributed estimation in energy-constrained wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3746–3758, Oct. 2009.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[22] A. Agarwal, M. Dudik, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 120–129.

[23] E. Creager, J.-H. Jacobsen, and R. Zemel, "Environment inference for invariant learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2189–2200.

[24] S. Maity, D. Mukherjee, M. Yurochkin, and Y. Sun, "There is no trade-off: Enforcing fairness can improve accuracy," 2020, *arXiv:2011.03173*.

[25] J. A. Bagnell, "Robust supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2005, pp. 714–719.

[26] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.

[27] Y. Wald, A. Feder, D. Greenfeld, and U. Shalit, "On calibration and out-of-domain generalization," 2021, arXiv:2102.10395.

[28] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Berlin, Germany: Springer, 2006.

[29] L. Gabrielli, S. Tomassetti, S. Squartini, and C. Zinato, "Introducing deep machine learning for parameter estimation in physical modelling," in *Proc. 20th Int. Conf. Digit. Audio Effects*, 2017.

[30] J. Rudi, J. Bessac, and A. Lenzi, "Parameter estimation with dense and convolutional neural networks applied to the FitzHugh-Nagumo ODE," 2020, *arXiv:2012.06691*.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.

[31] V. Dua, "An artificial neural network approximation based decomposition approach for parameter estimation of system of ordinary differential equations," *Comput. Chem. Eng.*, vol. 35, no. 3, pp. 545–553, 2011.
[32] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[33] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1177–1184.
[34] E. Rosenfeld, P. Ravikumar, and A. Risteski, "Domain-adjusted regression or: ERM may already learn features sufficient for out-of-distribution generalization," 2022, *arXiv:2202.06856*.
[35] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," 2022, *arXiv:2204.02937*.
[36] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. Berlin, Germany: Springer, 2006.
[37] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[38] N. S. Alagha, "Cramer-Rao bounds of SNR estimates for BPSK and QPSK modulated signals," *IEEE Commun. Lett.*, vol. 5, no. 1, pp. 10–12, Jan. 2001.
[39] A. Wiesel, J. Goldberg, and H. Messer, "Non-data-aided signal-to-noise-ratio estimation," in *Proc. IEEE Int. Conf. Commun.*, 2002, pp. 197–201.
[40] D. R. Pauluzzi and N. C. Beaulieu, "A comparison of SNR estimation techniques for the AWGN channel," *IEEE Trans. Commun.*, vol. 48, no. 10, pp. 1681–1691, Oct. 2000.
[41] S. Chaudhuri, M. Drton, and T. S. Richardson, "Estimation of a covariance matrix with zeros," *Biometrika*, vol. 94, no. 1, pp. 199–216, 2007.
[42] D. Deng, G. Chen, Y. Yu, F. Liu, and P.-A. Heng, "Uncertainty estimation by fisher information-based evidential deep learning," 2023, *arXiv:2303.02045*.
[43] M. Mohammadi, A. Al-Fuqaha, M. Guizani, and J. S. Oh, "Semi-supervised deep reinforcement learning in support of IoT and smart city services," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 624–635, Apr. 2018.
[44] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[45] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2403–2412.

Tzvi Diskin received the B.Sc. and M.Sc. degrees in physics from the Technion - Israel Institute of Technology, Haifa, Israel, in 2012 and 2015, respectively. He is currently working toward the Ph.D. degree with the Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel, under the supervision of Prof. A. Wiesel.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degree in physics and the second B.Sc. degree in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2002. She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel, where she holds the Dorothy and Patrick Gorman Professorial Chair and Heads the Center for Biomedical Engineering. She was previously a Professor with the Department of Electrical Engineering, Technion, Haifa, Israel, where she held the Edwards Chair of Engineering. She is also a Visiting Professor with MIT, a Visiting Scientist with Broad Institute, Cambridge, a Visiting Research Collaborator at Princeton, an Adjunct Professor with Duke University, Durham, NC, USA, an Advisory Professor with Fudan University, Shanghai, China, a Distinguished Visiting Professor with Tsinghua University, Beijing, China, and was a Visiting Professor with Stanford. She is a Member of the Israel Academy of Sciences and Humanities (elected 2017) and of the Academia Europaea (elected 2023), EURASIP Fellow, Fellow of the Asia-Pacific Artificial Intelligence Association, and Fellow of the 8400 Health Network. She is the author of the book *Sampling Theory: Beyond Bandlimited Systems* and co-author of seven other books. Her research interests include the broad areas of statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology, medical imaging and optics. Dr. Eldar was the recipient of the many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award in 2013, IEEE/AESS Fred Nathanson Memorial Radar Award in 2014, and IEEE Kiyo Tomiyasu Award in 2016. She was a Horev Fellow of the Leaders of Science and Technology Program, the Technion and an Alon Fellow. She was also the recipient of the Michael Bruno Memorial Award from the Rothschild Foundation, Weizmann Prize for Exact Sciences, Wolf Foundation Krill Prize for Excellence in Scientific Research, Henry Taub Prize for Excellence in Research (twice), Hershel Rich Innovation Award (three times), Award for Women with Distinguished Contributions, Andre and Bella Meyer Lectureship, Career Development Chair at the Technion, Muriel & David Jacknow Award for Excellence in Teaching, and Technion's Award for Excellence in Teaching (two times). She was also the recipient of the several best paper awards and best demo awards together with her research students and colleagues including the SIAM outstanding Paper Prize, UFFC Outstanding Paper Award, Signal Processing Society Best Paper Award and IET Circuits, Devices and Systems Premium Award. She was selected as one of the 50 most influential women in Israel and in Asia, and is a highly cited researcher. She was a Member of the Young Israel Academy of Science and Humanities and Israel Committee for Higher Education. She is the Editor in Chief of *Foundations and Trends in Signal Processing*, Member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. Earlier, she was a Signal Processing Society Distinguished Lecturer, Member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, *EURASIP Journal of Signal Processing*, *SIAM Journal on Matrix Analysis and Applications*, and *SIAM Journal on Imaging Sciences*. She was the Co-Chair and Technical Co-Chair of several international conferences and workshops.

Ami Wiesel (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering from the Technion - Israel Institute of Technology, Haifa, Israel, in 2007. During 2007–2009, he was a Postdoctoral Fellow with the University of Michigan, Ann Arbor, MI, USA. He is currently a Professor with the Rachel and Selim Benin School of Computer Science and Engineering, Hebrew University of Jerusalem, Jerusalem, Israel.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on June 23,2023 at 14:28:57 UTC from IEEE Xplore. Restrictions apply.