

Maximum Likelihood is Likely Wrong

Paul Mayer, Lorenzo Luzi, Ali Siahkoohi, Richard Baraniuk

November 2024

1 Introduction

In this paper, it is argued that Maximum Likelihood Estimation (MLE) is wrong, both conceptually and in terms of results it produces (except in two very special cases, which are discussed). While the use of MLE can still be justified on the basis of its practical performance, it is argued that there are better estimation methods that overcome MLE’s empirical and philosophical shortcomings while retaining all of MLE’s benefits.

2 Background

Maximum Likelihood Estimation (MLE), in its modern form, was proposed and named by Sir Ronald Fisher (Hald, 1999). It is a form of parameter estimation where chosen parameters maximize the likelihood function $P(X|\theta)$, the probability of the observed data. MLE chooses parameters as follows:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta'} P(X|\theta') \quad (1)$$

MLE is used to train most generative model architectures, including variational autoencoders (VAEs) (Pu et al., 2016), normalizing flows (NFs) (Rezende and Mohamed, 2015), diffusion models (Ho et al., 2020), and generative adversarial networks (GANs) (Goodfellow et al., 2014)¹. Any deep learning model that uses the negative log likelihood as a loss function is performing maximum likelihood estimation (Vapnik, 1999, 1991). Despite MLE’s ubiquity, it often produces biased estimators of the underlying true parameters. The most famous example was pointed out by Neyman and Scott (1948), who showed that MLE can produce inconsistent results when the number of parameters is large relative to the amount of data (DasGupta, 2008). In the Neyman-Scott problem, there is not enough data relative to the number of parameters to mitigate the bias, leading to what Neyman called “false estimations of the parameters”, or statistics where the stochastic limits were unequal to the values of the parameters to be estimated (Stigler, 2007).

¹These observations apply to models trained on a lower bound of the likelihood function, such as the popular ELBO (Kingma and Welling, 2022).

3 Parameter Estimation

Suppose there is a value² θ one wants to obtain that characterizes a population. Unfortunately, due to the size of the population (which could be infinite), time, cost, or other considerations, neither θ nor the population can be observed directly. Instead, one is given access to a sample from the population \mathbf{X} . The goal of parameter estimation is to *guess* or *estimate* θ as accurately as possible from \mathbf{X} . Use $\hat{\theta}$ to denote an estimate of θ , sometimes explicitly writing $\hat{\theta} = H(\mathbf{X})$ to show this estimate is a statistic or function of the sample \mathbf{X} .

What makes the problem difficult is there can be different populations that conceivably generated \mathbf{X} , each corresponding to a different value of θ . Randomness is thus introduced due to the underdetermination of population parameters from the (usually finite) sample. This situation creates *epistemic* uncertainty regarding the true values of the parameters³. The sampling distribution of $\hat{\theta}$ thus represents the uncertainty in recovering θ from samples of a distribution parametrized by θ .

In many cases, this uncertainty is removed when sampling the entire population or letting the sample size grow to infinity. This idea is captured by *consistent* estimators, or estimators where $\lim_{n \rightarrow \infty} \mathbb{E} [|\hat{\theta} - \theta|_2^2] = 0$. When the sample size is finite and the problem is underdetermined, one often desires a way of choosing the “best” possible parameters over the set of parameters that *possibly* generated \mathbf{X} . For instance, the sample $\mathbf{X}_u = [0.2, 0.4, 0.6]$ *could* have been generated from a one-sided uniform distribution $U[0, a]$, for any value of $a \geq 0.6$. This means there are an uncountably infinite number of populations that could have given rise to the observed data. Reducing this set of possible a ’s is necessary to get a single estimate of the population parameters.

3.1 Maximum Likelihood Estimation

MLE suggests reducing the set of candidate parameters by choosing the ones that makes the sample \mathbf{X} as likely as possible. On the surface, this is an intuitive solution to underdetermination, as it lets the data drive the selection process. However MLE ignores the critical interplay between the sampling distribution of $\hat{\theta}$ and the parametric form of the distribution under a given choice of parameters. In summary, MLE has a tendency to overfit the parameters to the sample, often at the expense of generalizing to the population.

Consider again estimating the value of a for a one-sided uniform distribution $U[0, a]$, given a sample $\mathbf{X}_u = [0.2, 0.4, 0.6]$. MLE chooses the smallest *possible* value of a . If a was any larger, the likelihood of \mathbf{X}_u would decrease since the density of the distribution is inversely related to its support. This choice corresponds to estimating a via the largest value in \mathbf{X}_u . This is shown in detail in Appendix A.

²Without losing generality, θ this can also be a vector.

³The connection between randomness and epistemology has been noted by many; see (Hoang, 2020; Cox, 1946; Jaynes, 2003; Hájek, 2023).

The problem is that \mathbf{X}_u is not a plausible sample from $U[0, 0.6]$. If $a = 0.6$ is really the true value of the parameter, one should not expect to *estimate* this value, since the only possible way this value can be estimated is by sampling 0.6 *exactly*. In other words, the density of samples that give rise to this estimate *under* the estimate itself have Lebesgue measure zero. While \mathbf{X}_u is rendered likely by choosing $a = 0.6$, estimating $\hat{a} = 0.6$ **itself** from a sample where the population parameter really is 0.6 is extremely unlikely, occurring almost never (a.n.).

3.2 Bias

The issue with MLE’s estimate of the one-sided uniform example described previously is not just that the estimate a is unlikely. Indeed any estimate of a is highly unlikely. The issue with MLE is that it will almost surely underestimate a . Consider the PDF of \hat{a}_{MLE} , and find its expected value conditioned on a variable representing the population value of the parameter, which is $\frac{n}{n+1}a$ (this is derived in Appendix A.1). The difference between this and the value of the parameter, a , is $-\frac{1}{n+1}a$. Because this difference is negative, it implies \hat{a}_{MLE} will, in expectation *underestimate* a .

This difference between the expected value of an estimate and the true value of the parameter is called the estimator’s *bias*, defined as $b(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}|\theta] - \theta$ (Johnson, 2013). In this example, we see that MLE produces biased results by essentially overfitting. Notice further that as $n \rightarrow \infty$ the bias disappears: this is due to the lack of uncertainty. In many cases, when there is no uncertainty, MLE is unbiased (hence why MLE is described as asymptotically unbiased in much of the literature (Johnson, 2013)). However for many distributions, MLE is biased when the sample size is finite.

4 What’s So Bad About Bias?

Parameters learned from biased estimators suffer from two serious issues: MADness and unfairness. MADness is a term in the machine learning literature referring to generative models that become progressively worse progressively in quality (precision) and diversity (recall) when trained on their own output (Alemohammad et al., 2023). The term MADness is a reference to bovine spongiform encephalopathy (BSE), the medical term for mad cow disease⁴. This phenomenon has become a growing concern for the machine learning community due to the availability and ubiquity of synthetic data (Nikolenko, 2021).

The “fully synthetic loop” described by Alemohammad et al. (2023) corresponds to a parameter estimation problem where the estimated parameters $\hat{\theta}$ are used to generate new samples $\hat{\mathbf{X}}$. From these new samples, the parameters can be estimated again, and so on, forming the basis of what is called a self-consumed or autophagous loop. This loop is illustrated in Figure 1. While

⁴Bovine spongiform encephalopathy (BSE) is a neurological disorder believed to be transmitted by cattle eating the remains of *other* (infected) cattle (Prusiner, 2001).

unbiased estimates center around the true value of the parameter, biased estimates diverge, as shown in Figure 1. This means that MLE estimates are far more susceptible to MADness than unbiased alternatives.

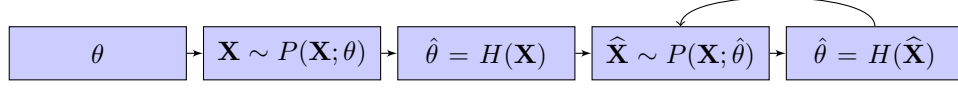


Figure 1: The self-consuming parameter estimation loop.

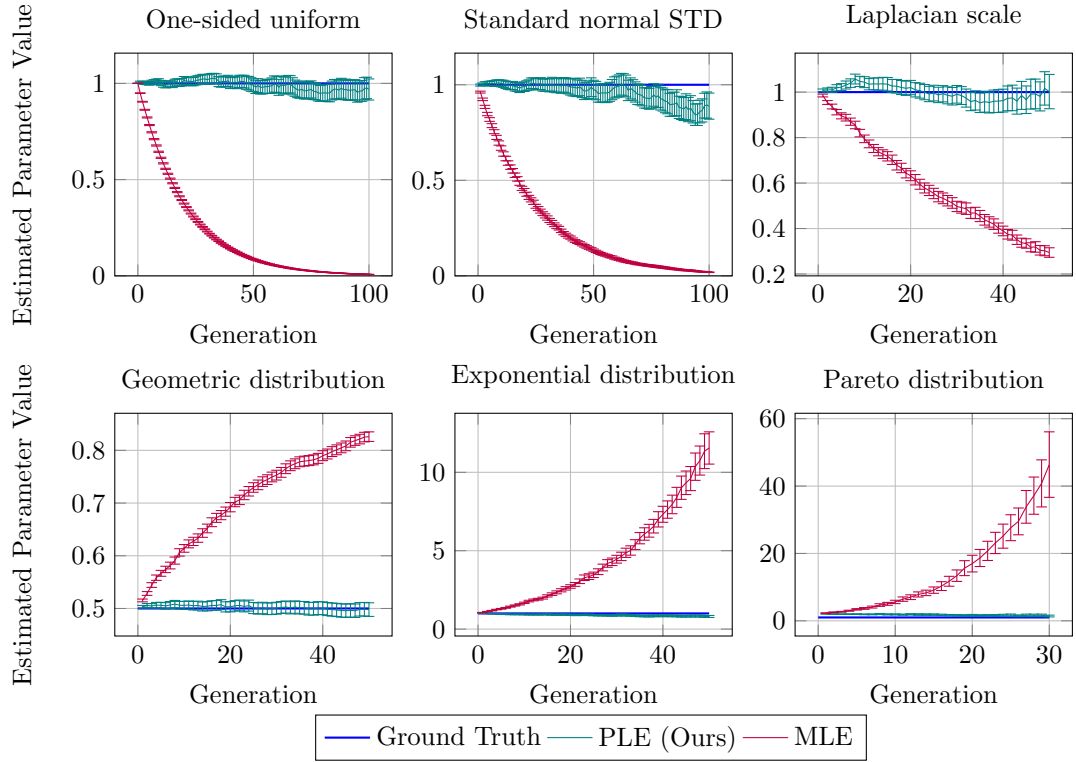


Figure 2: MLE vs PLE Estimates of the parameters of various distributions. Notice how MLE collapses into MADness much faster than PLE. More details on these experiments can be found in Appendix B.

4.1 Frequentism’s Failure

It is difficult to overstate the significance of the failure of MLE to produce unbiased estimates of population parameters. Although MLE can produce unbiased estimates when presented with an infinite amount of data, there is no practical setting where one has access to an infinite amount of data. Furthermore,

even if one could somehow process and access an infinite amount of data, there would be no reason to estimate, since one could simply exhaust the population and measure the parameter directly. Additionally, if the number of parameters scales with the amount of data, even an infinite amount of data will not solve the bias problem present in MLE, since Neyman and Scott (1948)’s example famously showed MLE’s bias can cause its results to be inconsistent (DasGupta, 2008).

5 Fixing MLE

The Bayesian perspective provides a path for solving the aforementioned issues with MLE. Under the Bayesian perspective, the sampling distribution of $\hat{\theta}$ represents the uncertainty choosing $\hat{\theta}$ from \mathbf{X} . However the uncertainty in estimating $\hat{\theta}$ is driven by two factors: the first is the underdetermination of $\hat{\theta}$ from a sample \mathbf{X} (which is usually finite). The second is more fundamental: one wants to capture the population parameters θ that not only generated \mathbf{X} , but other possible samples one *could* have observed if one had sampled differently. MLE successfully resolves the first kind of uncertainty, but ignores the second.

5.1 The True Goal of Parameter Estimation

The true goal of parameter estimation is *not* necessarily to maximize $P(\mathbf{X}|\theta)$. It is instead to find the θ that parametrizes the population. θ is a fixed (but unknown) value one wishes to find, so in Bayesian terms, $P(\theta)$ represents one’s confidence or degree of belief that the population parameter equals the argument value. The posterior distribution $P(\theta|\mathbf{X})$ represents one’s updated belief about θ after observing \mathbf{X} . On the surface, a reasonable way to estimate parameters given data \mathbf{X} is by maximizing this posterior distribution, an estimation method called Maximum A Posteriori (MAP):

$$\hat{\theta}_{\text{MAP}}^5 = \arg \max_{\theta'} P(\theta'|\mathbf{X}) = \arg \max_{\theta'} \frac{P(\mathbf{X}|\theta')P(\theta')}{P(\mathbf{X})} = \arg \max_{\theta'} P(\mathbf{X}|\theta')P(\theta') \quad (2)$$

MAP and MLE go wrong by failing to take into account that $P(\mathbf{X})$ itself depends on the fixed value of θ one wishes to estimate. We make the connection between the sample \mathbf{X} and the population parameters explicit by showing the conditionality of the posterior on θ :

$$\hat{\theta}_{\text{MAPFIX}} = \arg \max_{\theta'} P(\theta'|\mathbf{X}, \theta) = \arg \max_{\theta'} P(\mathbf{X}|\theta', \theta)P(\theta'|\theta) \quad (3)$$

Here, $P(\theta'|\theta)$ represents the probability one would estimate θ' if the true parameters are in fact θ . Crucially, $P(\theta'|\theta)$ does not only depend on the sample \mathbf{X} ; instead it represents the probability of estimating θ' from any counterfactual sample \mathbf{Y} that *could* have been sampled by θ . This new form needs to take into

⁵MLE is a case of MAP with a uniform prior $P(\theta)$.

account any possible sample that could have been generated by θ , since, as discussed at the beginning of the section, the goal is to choose θ that parametrizes the population, not just the sample \mathbf{X} . Let θ_Y be the parameters estimated from the counterfactual sample \mathbf{Y} . We can then write the corrected MAP as follows, letting \mathbb{I} be the indicator function:

$$\hat{\theta}_{\text{MAPFIX}} = \arg \max_{\theta'} P(\mathbf{X}|\theta', \theta) \int_{\mathbf{Y}} \mathbb{I}(\theta' = \theta_Y|\theta) P(\mathbf{Y}|\theta) d\mathbf{Y} \quad (4)$$

Notice here that $\int_{\mathbf{Y}} \mathbb{I}(\theta' = \theta_Y|\mathbf{Y}) P(\mathbf{Y}|\theta) d\mathbf{Y} = \mathbb{E}_{\mathbf{Y}|\theta}[\mathbb{I}(\theta' = \theta_Y)]$, the expected value of when the estimated parameters from \mathbf{X} equal the estimated parameters from the given counterfactual sample \mathbf{Y} . Since $\theta' = H(\mathbf{X})$ is a statistic of the sample, we can write the overall estimate in terms of the function H which estimates parameters from of a sample. Noting that our estimation method H could equally be applied to \mathbf{X} or any other sample \mathbf{Y} , we rewrite the optimization in terms of H .

$$\begin{aligned} \hat{\theta}_{\text{MAPFIX}} &= H(\mathbf{X}) \\ H &= \arg \max_{H'} P(\mathbf{X}|H'(\mathbf{X}), \theta) \mathbb{E}_{\mathbf{Y}|\theta}[\mathbb{I}(H'(\mathbf{X}) = H'(\mathbf{Y}))]. \end{aligned} \quad (5)$$

6 Simplification

Needless to say, while Equation 5 is theoretically justified for estimating parameters, is difficult to solve for a myriad of reasons. The first is it requires knowledge of θ to generate the counterfactual samples \mathbf{Y} . If θ was known a priori, we would have no reason to estimate θ from samples in the first place. Furthermore, the indicator function inside the expectation, $\mathbb{I}(H'(\mathbf{X}) = H'(\mathbf{Y}))$, is not easy to compute: it requires computing the measure of when any estimate from a counterfactual sample equals the current estimator.

Computing this estimate can be made tractable with a few simplifying assumptions. The first is to remove the dependence of the estimate on knowledge of θ . The likelihood function is already implicitly conditioned on θ , since this value parametrizes the population \mathbf{X} is sampled from. For the expectation term, $\mathbb{E}_{\mathbf{Y}|\theta}[\mathbb{I}(H'(\mathbf{X}) = H'(\mathbf{Y}))]$, we can use $H'(\mathbf{X})$ as a proxy for the true parameters. With this change, \mathbf{Y} is a sample drawn from a distribution parametrized by $H'(\mathbf{X})$ rather than by θ itself: without knowledge of θ , we use our estimate of θ .

Next, we can simplify the computation of the expectation by removing the indicator function. For $P(\mathbf{X}|H'(\mathbf{X}), \theta) \mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[\mathbb{I}(H'(\mathbf{X}) = H'(\mathbf{Y}))]$ to take a nonzero value, it is necessary that $\mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[H'(\mathbf{X}) - H'(\mathbf{Y})] = 0$. We can turn this into an equality constraint on the optimization itself; we maximize the likelihood function such that there is no expected difference between our current estimate and an estimate from a counterfactual sample taken from distribution parametrized by the current estimate. This is captured via $\mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[H'(\mathbf{X}) - H'(\mathbf{Y})] = 0$. Using these assumptions, we arrive at the following form for parameter estimation.

$$\begin{aligned}\hat{\theta}_{\text{PLE}} &= H(\mathbf{X}), \\ H &= \arg \max_{H' \in \mathcal{H}} P(\mathbf{X}; H'(\mathbf{X})) \text{ s.t. } \mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[H'(\mathbf{Y}) - H'(\mathbf{X})] = 0.\end{aligned}\quad (6)$$

The equality constraint can be simplified via

$$\mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[H'(\mathbf{Y}) - H'(\mathbf{X})] = \int_{\mathbf{Y}} H'(\mathbf{Y})P(\mathbf{Y})d\mathbf{Y} - H'(\mathbf{X})$$

This assumption is the same as saying choose H' such that, over $\mathbf{Y} \sim H'(\mathbf{X})$,

$$H'(\mathbf{X}) = \mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[H'(\mathbf{Y})] \quad (7)$$

With this, we arrive at the final form for the PLE for a set of parameters:

$$\begin{aligned}\hat{\theta}_{\text{PLE}} &= H(\mathbf{X}), \\ H &= \arg \max_{H' \in \mathcal{H}} P(\mathbf{X}; H'(\mathbf{X})) \text{ s.t. } H'(\mathbf{X}) = \mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[H'(\mathbf{Y})].\end{aligned}\quad (8)$$

7 Implementing PLE Computationally

The constraint in Equation 8 requires taking an expectation over all counterfactual data \mathbf{Y} for each choice of H' . While this expectation is impossible to compute practically when dealing with infinite sets, it can be estimated via Monte Carlo approximation. Parametric bootstrapping for bias correction, such as that described by Hall (1992), is a special case of PLE where Monte Carlo Estimation is used to approximate $\mathbb{E}_{\mathbf{Y}|H'(\mathbf{X})}[H'(\mathbf{Y})]$.

References

- A. Hald, “On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares,” *Statistical Science*, vol. 14, no. 2, pp. 214–222, 1999, publisher: Institute of Mathematical Statistics. [Online]. Available: <https://www.jstor.org/stable/2676741>
- Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, “Variational Autoencoder for Deep Learning of Images, Labels and Captions,” in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/hash/eb86d510361fc23b59f18c1bc9802cc6-Abstract.html>
- D. Rezende and S. Mohamed, “Variational Inference with Normalizing Flows,” in *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, Jun. 2015, pp. 1530–1538, iSSN: 1938-7228. [Online]. Available: <https://proceedings.mlr.press/v37/rezende15.html>

- J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html>
- I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Networks,” Jun. 2014, arXiv:1406.2661. [Online]. Available: <http://arxiv.org/abs/1406.2661>
- D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” Dec. 2022, arXiv:1312.6114 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1312.6114>
- V. Vapnik, “An overview of statistical learning theory,” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, Sep. 1999, conference Name: IEEE Transactions on Neural Networks. [Online]. Available: <https://ieeexplore.ieee.org/document/788640>
- , “Principles of Risk Minimization for Learning Theory,” in *Advances in Neural Information Processing Systems*, vol. 4. Morgan-Kaufmann, 1991. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1991/hash/ff4d5fbbafdf976cfdc032e3bde78de5-Abstract.html
- J. Neyman and E. L. Scott, “Consistent Estimates Based on Partially Consistent Observations,” *Econometrica*, vol. 16, no. 1, pp. 1–32, 1948, publisher: [Wiley, Econometric Society]. [Online]. Available: <https://www.jstor.org/stable/1914288>
- A. DasGupta, “Maximum Likelihood Estimates,” in *Asymptotic Theory of Statistics and Probability*, A. DasGupta, Ed. New York, NY: Springer, 2008, pp. 235–258. [Online]. Available: https://doi.org/10.1007/978-0-387-75971-5_16
- S. M. Stigler, “The Epic Story of Maximum Likelihood,” *Statistical Science*, vol. 22, no. 4, pp. 598–620, 2007, publisher: Institute of Mathematical Statistics. [Online]. Available: <https://www.jstor.org/stable/27645865>
- L. N. Hoang, *The equation of knowledge: from Bayes’ rule to a unified philosophy of science*, first edition ed. Boca Raton, FL: CRC, 2020.
- R. T. Cox, “Probability, Frequency and Reasonable Expectation,” *American Journal of Physics*, vol. 14, no. 1, pp. 1–13, Jan. 1946, 781 citations (Crossref/DOI) [2024-11-15]. [Online]. Available: <https://doi.org/10.1119/1.1990764>
- E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, Apr. 2003, google-Books-ID: tTN4HuUNXjgC.

- A. Hájek, “Interpretations of Probability,” in *The Stanford Encyclopedia of Philosophy*, winter 2023 ed., E. N. Zalta and U. Nodelman, Eds. Metaphysics Research Lab, Stanford University, 2023. [Online]. Available: <https://plato.stanford.edu/archives/win2023/entries/probability-interpret/>
- D. H. Johnson, “Statistical signal processing,” URL <http://www.ece.rice.edu/~dhj/courses/elec531/notes.pdf>. Lecture notes, 2013.
- S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk, “Self-Consuming Generative Models Go MAD,” Jul. 2023. [Online]. Available: <http://arxiv.org/abs/2307.01850>
- S. B. Prusiner, “Neurodegenerative Diseases and Prions,” *New England Journal of Medicine*, vol. 344, no. 20, pp. 1516–1526, May 2001, publisher: Massachusetts Medical Society .eprint: <https://doi.org/10.1056/NEJM200105173442006>. [Online]. Available: <https://doi.org/10.1056/NEJM200105173442006>
- S. I. Nikolenko, *Synthetic Data for Deep Learning*, ser. Springer Optimization and Its Applications. Cham: Springer International Publishing, 2021, vol. 174. [Online]. Available: <https://link.springer.com/10.1007/978-3-030-75178-4>
- P. Hall, *The Bootstrap and Edgeworth Expansion*, ser. Springer Series in Statistics. New York, NY: Springer, 1992. [Online]. Available: <http://link.springer.com/10.1007/978-1-4612-4384-7>

A Details of the One-Sided Uniform

Consider a n samples drawn from the following uniform distribution

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n], \mathbf{x}_i \sim U[0, a] \ i = 1, \dots, n.$$

We wish to estimate the parameter \hat{a} from \mathbf{X} so $\hat{a} = a$. First, we write out the likelihood function: $P_{\mathbf{X}|\hat{a}}(\mathbf{X}|\hat{a})$ as

$$P_{\mathbf{X}|\hat{a}}(\mathbf{X}|\hat{a}) = \begin{cases} 0 & \text{if } \hat{a} < \max(\mathbf{X}) \\ \frac{1}{\alpha \hat{a}^n} & \text{else} \end{cases}$$

Where α is a scaling factor that ensures the conditional PDF integrates to 1. Since this function is monotonic with respect to \hat{a} , the MLE is easily found as $\hat{a}_{\text{MLE}} = \arg \max_{\hat{a}} P_{\mathbf{X}|\hat{a}}(\mathbf{X}|\hat{a}) = \max(\mathbf{X})$. This also corresponds to the n -th order statistic.

A.1 Bias of the MLE of the One-Sided Uniform

Now that we have \hat{a}_{MLE} , we can calculate the bias as follows:

$$b(\hat{a}_{\text{MLE}}) = \mathbb{E}_{\mathbf{X}|a}[\hat{a}_{\text{MLE}}] - a = \mathbb{E}_{\mathbf{X}|a}[\max(\mathbf{X})] - a \quad (9)$$

The expected value of the maximum of \mathbf{X} (the n -th order statistic of \mathbf{X}) can be calculated by taking the derivative of the CDF of the maximum value with respect to the parameter in question:

$$F(\max(\mathbf{X})) = P(\max(\mathbf{X}) \leq \hat{a}) = \begin{cases} 0 & a < 0 \\ \left(\frac{\hat{a}}{a}\right)^n & \hat{a} \in [0, a] \\ 1 & \hat{a} > a \end{cases}$$

$$f(\max(\mathbf{X})) = P(\max(\mathbf{X}) = \hat{a}) = \begin{cases} 0 & a < 0 \\ \frac{n\hat{a}^{n-1}}{a^n} & \hat{a} \in [0, a] \\ 0 & \hat{a} > a \end{cases}$$

Now we can calculate $\mathbb{E}[\max(\mathbf{X})]$ as follows:

$$\mathbb{E}_{\mathbf{X}|a}[\max(\mathbf{X})] = \frac{n}{a^n} \int_0^a \hat{a}^n d\hat{a} = \frac{n}{n+1} a. \quad (10)$$

Therefore, the bias of the MLE is

$$b(\hat{a}_{\text{MLE}}) = \frac{n}{n+1} a - a = -\frac{1}{n+1} a.$$

Note that the bias here is negative, implying that the MLE \hat{a}_{MLE} will (in expectation) *underestimate* a .

B MLE vs PLE for Various Distributions

This section explains how the plots for Figure 2 were generated. Error bars show the standard error after either 100 or 1000 different initializations (some of the figures needed 1000 initializations for the error bars to decrease). Subfigure 1 shows the result from using the closed-form expression of PLE described in Section ??, Subfigures 2-6 use the data-driven form from Equation ??, with 100 synthetic samples ($m = 100$) used to estimate the expectation in Equation ??.

Subfigure 1 (top-left) was generated from a one-sided Uniform distribution $\mathbf{X} \sim U[0, a]$ with true parameter $a = 1$, using $n = 20$ datapoints. The MLE is $a_{\text{MLE}} = \max \mathbf{X}$, which is derived in Appendix A, while $a_{\text{PLE}} = \frac{n+1}{n} \max(\mathbf{X})$, which is derived in Section ???. The parameter \hat{a} is estimated each iteration. Error bars show the standard error after 100 different initializations.

Subfigure 2 (top-middle) shows samples generated from a standard Gaussian (normal) distribution $\mathbf{X} \sim N[\mu, \sigma]$, with true parameters $\mu = 0, \sigma = 1$. The mean $\hat{\mu}$ and the standard deviation $\hat{\sigma}$ are estimated each iteration. We use

$\mu_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\sigma_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$, which is derived in Section ???. The PLE estimates use the data-driven form from Equation ??, with 100 synthetic samples ($m = 100$). The estimates are generated with $n = 20$ points, and the results are averaged from 1000 initializations.

Subfigure 3 (top-right) shows samples generated from a Laplacian distribution $\mathbf{X} \sim \text{Laplace}[\mu, b]$ with true parameters $\mu = 0, b = 1$. The mean μ and the scale parameter b are estimated each iteration. The MLE of the parameters are $\mu_{\text{MLE}} = \text{median}(\mathbf{X})$ and $b_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|$, and PLE estimates use the data-driven form from Equation ??, with 100 synthetic samples ($m = 100$). The estimates are generated with $n = 25$ points, and the results are averaged from 1000 initializations.

Subfigure 4 (bottom-left) shows samples generated from a Geometric distribution $\mathbf{X} \sim \text{Geometric}[p]$, where the true parameter $p = 0.5$. The parameter \hat{p} is estimated each iteration. The MLE of p is $p_{\text{MLE}} = \frac{n}{\sum_{i=1}^n x_i}$, and PLE estimates use the data-driven form from Equation ??, with 100 synthetic samples ($m = 100$). The estimates are generated with $n = 25$ points, and the results are averaged over 1000 initializations.

Subfigure 5 (bottom-middle) shows samples generated from an Exponential distribution $\mathbf{X} \sim \text{Exponential}[\lambda]$, where the true parameter $\lambda = 0.5$. The parameter $\hat{\lambda}$ is estimated each iteration. The MLE of λ is $\lambda_{\text{MLE}} = \frac{n}{\sum_{i=1}^n x_i}$, and PLE estimates use the data-driven form from Equation ??, with 100 synthetic samples ($m = 100$). The estimates are generated with $n = 25$ points, and the results are averaged over 1000 initializations.

Subfigure 6 (bottom-right) shows samples generated from a Type-I Pareto distribution $\mathbf{X} \sim \text{Pareto}[b]$, where the true parameter $b = 1.0$. The PDF of this distribution is $f(x, b) = \frac{b}{x^{b+1}}$, and \hat{b} is estimated each iteration. The MLE of b is $b_{\text{MLE}} = \frac{n}{\sum_{i=1}^n (\log(x_i)) - n \log(\min(\mathbf{X}))}$, and PLE estimates use the data-driven form from Equation ??, with 100 synthetic samples ($m = 100$). The estimates are generated with $n = 25$ points, and the results are averaged over 100 initializations.

The upper-left and upper-middle sub-figures show PLE estimated parameters slope down slightly. This is due to the fact that for a few runs, the variance goes to zero and cannot “recover” via a multiplicative constant. These degenerate runs bring the overall average down slightly, as there is no analogous degeneracy for large values. In essence, for the few estimates of the variance that are near zero, the result becomes clipped. This is sometimes described as variance collapse or model collapse in the literature (Alemohammad et al., 2023).