

---

# EXACT CLOSED-FORM GAUSSIAN MOMENTS OF RESIDUAL LAYERS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

We study the problem of propagating the mean and covariance of a general multivariate Gaussian distribution through a deep (residual) neural network using layer-by-layer moment matching. We close a longstanding gap by deriving exact moment matching for the probit, GeLU, ReLU (as a limit of GeLU), Heaviside (as a limit of probit), and sine activation functions; for both feedforward and generalized residual layers. On random networks, we find orders-of-magnitude improvements in the KL divergence error metric, up to a millionfold, over popular alternatives. On real data, we find competitive statistical calibration for inference under epistemic uncertainty in the input. On a variational Bayes network, we show that our method attains hundredfold improvements in KL divergence from Monte Carlo ground truth over a state-of-the-art deterministic inference method. We also give an *a priori* error bound and a preliminary analysis of stochastic feedforward neurons, which have recently attracted general interest.

## 1 INTRODUCTION

We are interested in inference of the output distribution of a neural network when the input is Gaussian-distributed. There are at least four reasons to study this problem.

1. The pushforward distribution of a neural network can shed light on local robustness of the network to “typical” (as opposed to worst-case) input perturbations (Wright et al., 2024).
2. Uncertainty propagation allows a neural network trained on a population with certain inputs to make predictions on a shifted distribution of uncertain inputs (Bibi et al., 2018).
3. Deterministic distribution propagation can replace Monte Carlo in both training and inference of variational Bayes neural networks with random weights (Frey & Hinton, 1999; Wu et al., 2019; Petersen et al., 2024; Wright et al., 2024; Rui Li & Trapp, 2025).
4. Analytical moments can be used to understand how activation functions behave in deep networks, by applying the Central Limit Theorem in the wide limit (He et al., 2015).

This problem is of basic importance, easily stated in elementary terms, and widely studied. Yet even in the simplest case of a single hidden layer with two neurons, no existing method can compute the mean and covariance exactly, in closed form.

From the perspective of numerical analysis, a numerical approximation to a functional (such as moments with respect to a measure) should be judged by the largest class of functions on which it is exact (Press, 2007, Chapter 4). Jacobian linearization and the unscented transforms (Julier et al., 1995; Julier, 2002) are exact for linear functions. In one dimension, Gauss-Hermite quadrature using  $n$  points is exact for the first moment of polynomials of degree  $2n - 1$ , and for the second moment of polynomials of degree  $n - 1$ . But in dimension  $d$ , Gauss-Hermite quadrature needs  $n^d$  points to achieve the same accuracy. In search of efficiency, the activation function-based calculations in Abdelaziz et al. (2015); Huber (2020); Wagner et al. (2022); Akgül et al. (2025); Bibi et al. (2018); Wu et al. (2019) are exact only for one-layer networks *of a single neuron* (they approximate the covariance between neurons). A recent theoretical work proves optimal approximation for a single ReLU neuron (Petersen et al., 2024). The series expansion of Wright et al. (2024) is formally correct for a single hidden layer if taken to infinity, but must be computed by hand and truncated at finitely many terms.

---

054 In this work, we have managed to achieve, for the first time, exact moment propagation (to machine  
055 precision) for activation functions ReLU, used in the earliest deep networks; GeLU (Hendrycks &  
056 Gimpel, 2017), used in frontier LLMs; sine, conceived to overcome frequency bias (Parascandolo  
057 et al., 2017; Ziyin et al., 2020) and later appreciated for physics-informed learning (Sitzmann et al.,  
058 2020); probit, a close approximation to the logistic sigmoid (Huber, 2020); Heaviside, used in Bayes  
059 networks (Wright et al., 2024); as well as general residual connections. We derive closed-form  
060 means and full covariance matrices.

061 On multi-layer networks, our method applies a variational Gaussian approximation, for which we  
062 provide *a priori* error bounds. We argue that this approximation not only prudently trades off be-  
063 tween accuracy and scalability, but also essentially saturates a hard theoretical limit: exact integra-  
064 tion of a general deep neural network under high-dimensional uncertainty is known to be #P-hard in  
065 the number of input neurons (Feischl & Zehetgruber, 2025).<sup>1</sup>

066 **Contribution.** We derive exact first and second moment matching for propagating the mean and  
067 covariance matrix of a Gaussian distribution through a single layer of a (residual) neural network  
068 (Lemma 1). For deeper neural networks, the single-layer formula is chained over layers, with orders-  
069 of-magnitude accuracy improvement over other popular techniques (§5.1). We demonstrate appli-  
070 cations of uncertainty propagation to inference on real data: regression on a noisy input (§2) and  
071 classification with missing features (§5.3). We discuss hard instances on which methods fail (§4),  
072 and then prove a soft guarantee on the method’s accuracy in favorable conditions (§3). Finally, we  
073 briefly explore a possibility related to stochastic activations (§5.5).

074 **Related work: local approximation.** Existing results on distribution propagation through neural  
075 networks can be taxonomized by the assumptions they impose on the input distribution (Appendix A,  
076 Table 4). Some works assume a small covariance matrix. Then the mean and covariance can be prop-  
077 agated by Jacobian linearization (Titensky et al., 2018; Nagel & Huber, 2022; Petersen et al., 2024;  
078 Jungmann et al., 2025; Bergna et al., 2025; Rui Li & Trapp, 2025). This formula is justified by the  
079 delta method (van der Vaart, 1998, Chap. 3), dates back to Gauss’s study of heavenly bodies (Gauss,  
080 1857, Chapter 187), and is taught in textbooks (Taylor, 1997). The Unscented Transformation, used  
081 in Astudillo & Neto (2011); Abdelaziz et al. (2015), is also justified by a Taylor expansion (Julier,  
082 2002).

083 **Related work: analytical approximation.** For certain activation functions including ReLU and  
084 GeLU, the mean and the diagonal of the covariance matrix can be computed explicitly, but (prior  
085 to this paper) there is no closed-form known for off-diagonal covariances of a hidden layer. Some  
086 works set them to zero—the mean-field assumption (Huber, 2020; Goulet et al., 2021; Wagner et al.,  
087 2022; Rui Li & Trapp, 2025; Bergna et al., 2025; Akgül et al., 2025; Rui Li & Trapp, 2025). Bibi  
088 et al. (2018) uses an analytical approximation around zero mean, and Wu et al. (2019) uses an  
089 analytical approximation around infinite mean. For the logistic activation function  $\sigma(x) = (1 +$   
090  $e^{-x})^{-1}$ ; Astudillo & Neto (2011); Abdelaziz et al. (2015), and Huber (2020) approximate  $\sigma$  with  
091 another function having closed-form Gaussian moments, such as a piecewise exponential function or  
092 a rescaled Normal CDF  $\Phi$ . Appendix A, Table 5 catalogs the literature on moment approximations  
093 for activation functions. We exemplify the failure modes of the above methods by giving single-  
094 hidden-layer counterexamples in §4.

095 For a general activation function, Wright et al. (2024) uses a Fourier transform to derive the exact  
096 mean and covariance matrix of a general activation function as a formal power series in  $\rho$ , the inter-  
097 neuron correlation. The series must be truncated, as each coefficient needs to be derived by hand.

098 Our work supersedes the ReLU, GeLU, and Heaviside moment derivations of Frey & Hinton (1999);  
099 Bibi et al. (2018); Wu et al. (2019); Huber (2020); Akgül et al. (2025); Wright et al. (2024), as well  
100 as the sine moment derivations of Sitzmann et al. (2020) by virtue of being exact on single layers.  
101 While we do not have exact integrals for the logistic sigmoid function, we do for  $\Phi$ , which is a  
102 similarly shaped function.<sup>2</sup>

## 103 2 METHODOLOGY

---

106 <sup>1</sup>We discuss the runtime of our method in App. O.

107 <sup>2</sup>In fact, Huber (2020) proposes to train a network using logistic activation and compute its moments using  
a  $\Phi$  surrogate network. Our numerical experiments simply train a  $\Phi$  network directly.

The activation function of a neural network is denoted  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  and applies elementwise. Except for parameters  $A, b, C, d$ , capital letters refer to random variables. The layers of a neural network are indicated by superscripts, e.g.  $A^k$  is a matrix of parameters for the  $k$ th layer. If  $X$  is a square-integrable random vector, the notation  $\mathbb{N} X$  refers to a random variable distributed as  $\mathcal{N}(\mathbb{E} X, \text{Cov} X)$ .

A neural network is a composition of layer functions  $\mathbb{R}^n \rightarrow \mathbb{R}^m$

$$g(x; A, b, C, d) = \sigma(Ax + b) + Cx + d, \quad (1)$$

where  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ ,  $C \in \mathbb{R}^{m \times n}$ ,  $d \in \mathbb{R}^m$  are parameters.

**Definition 1.** A neural network with  $\ell$  layers is the function  $f : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_y}$  defined by

$$\begin{aligned} f(x) &= f^\ell(x) \\ f^k(x) &= g(f^{k-1}(x); A^k, b^k, C^k, d^k) \quad \forall k \in \{1 \dots \ell\} \\ f^0(x) &= x \end{aligned}$$

Stated formally, the problem of uncertainty propagation studied in this paper is:

**Problem 1.** Let  $f$  be a neural network with  $\ell$  layers. Given  $X \sim \mathcal{N}(\mu, \Sigma)$ , characterize the distribution of  $Y_0 = f(X)$ .

After layer-wise Gaussian approximation, this problem reduces to:

**Problem 2.** Given  $X \sim \mathcal{N}(\mu, \Sigma)$  and  $A, b, C, d$ ; find exact expressions for  $\mathbb{E} g(X; A, b, C, d)$  and  $\text{Cov} g(X; A, b, C, d)$ .

## 2.1 OUR ANALYTIC METHOD $Y_{\text{ana}}$

Our method, like Wright et al. (2024), re-approximates each layer by a Gaussian sharing its first two moments:

**Definition 2.** Let  $f$  be a neural network with  $\ell$  layers. Given  $X \sim \mathcal{N}(\mu, \Sigma)$ , the moment-matching Gaussian approximation of  $f(X)$ , is the random variable  $Y_{\text{ana}}$  defined by

$$\begin{aligned} Y_{\text{ana}} &= Y^\ell \\ Y^k &= \mathbb{N} g(Y^{k-1}; A^k, b^k, C^k, d^k) \quad \forall k \in \{1 \dots \ell\} \\ Y^0 &= X \end{aligned}$$

According to basic index manipulation, three transcendental functions are needed to compute the first two Gaussian moments of a layer defined by (1).

**Definition 3.** Given a nonlinear function  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ , the functions  $M_\sigma : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$  and  $K_\sigma, L_\sigma : \mathbb{R}^2 \times \mathbb{R}_{\geq 0}^{2 \times 2} \rightarrow \mathbb{R}$  are

$$\begin{aligned} M_\sigma(\mu; \nu) &= \mathbb{E} \sigma(X), & X &\sim \mathcal{N}(\mu, \nu) \\ K_\sigma(\mu_1, \mu_2; \nu_{11}, \nu_{22}, \nu_{12}) &= \text{Cov}(\sigma(X_1), \sigma(X_2)), & \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &\sim \mathcal{N} \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{12} & \nu_{22} \end{pmatrix} \right) \\ L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) &= \text{Cov}(\sigma(X_1), X_2), & \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &\sim \mathcal{N} \left( \begin{pmatrix} \mu_1 \\ \star \end{pmatrix}, \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{12} & \nu_{22} \end{pmatrix} \right) \end{aligned}$$

**Lemma 1.** For some activation function  $\sigma$ , let  $g$  be the function defined by  $g_\sigma(x; A, b, C, d) = \sigma(Ax + b) + Cx + d$ . Let  $X \sim \mathcal{N}(\mu, \Sigma)$ . Then

$$\left( \mathbb{E} g_\sigma(X; A, b, C, d) \right)_i = M_\sigma(\mu_i; \nu_{ii}) + (C\mu)_i + d_i$$

and

$$\begin{aligned} \left( \text{Cov} g_\sigma(X; A, b, C, d) \right)_{i,j} &= K_\sigma(\mu_i, \mu_j; \nu_{ii}, \nu_{jj}, \nu_{ij}) \\ &\quad + L_\sigma(\mu_i; \nu_{ii}, \tau_{jj}, \kappa_{ij}) + L_\sigma(\mu_j; \nu_{jj}, \tau_{ii}, \kappa_{ji}) \\ &\quad + \tau_{ij}. \end{aligned}$$

where for all valid indices  $(i, j)$ ,

$$\begin{aligned}\mu_i &= (A\mu + b)_i & \tau_{ij} &= (C\Sigma C^\top)_{i,j} \\ \nu_{ij} &= (A\Sigma A^\top)_{i,j} & \kappa_{ij} &= (A\Sigma C^\top)_{i,j}\end{aligned}$$

With a five-dimensional domain, these functions are too complex to be represented by a look-up table. We compute them analytically for:

**probit** in App. D by expressing  $\Phi(x) = \mathbb{E}[Z \leq x]$  in terms of an auxiliary standard Normal  $Z$ , generalizing (due to the inclusion of an affine term) and superseding the results of Huber (2020, § III.B) and Wright et al. (2024, App. B.4).

**GeLU** in App. E by repeated applications of the multivariate Stein’s lemma and the Gaussian ODE  $\phi'(x) + x\phi(x) = 0$ , generalizing and superseding the result of Wright et al. (2024, App. B.2).

**ReLU** in App. F by using the Dominated Convergence Theorem to take the pointwise limit

$$\text{ReLU}(x) = \lim_{\lambda \rightarrow \infty} \lambda^{-1} \text{GeLU}(\lambda x),$$

generalizing and superseding the results of Frey & Hinton (1999, App. C.3), Bibi et al. (2018, §3), Wu et al. (2019, App A.2.2), Huber (2020, § III.C), and Wright et al. (2024, App. B.2)

**Heaviside** in App. G by using the Dominated Convergence Theorem to take the pointwise limit

$$\text{Heaviside}(x) = \lim_{\lambda \rightarrow \infty} \Phi(\lambda x),$$

generalizing and superseding the results of Frey & Hinton (1999, App. C.2), Wu et al. (2019, App. A.2.1), and Wright et al. (2024, App. B.2).

**sine** in App. H by combining the characteristic function of the Normal distribution with the trigonometric identity  $\sin(x) = (e^{ix} - e^{-ix}) / (2i)$ , generalizing and superseding the results in Sitzmann et al. (2020, App. 1).

The calculations are interesting in themselves because they are probabilistic in nature: we never resort to Riemann integrals against the Gaussian density, instead working with higher-level properties of Gaussian variables derived in App. C.

The baseline methods  $Y_{\text{mfa}}$ , mean-field;  $Y_{\text{lin}}$ , linear;  $Y_{\text{u'95}}$ , unscented’95; and  $Y_{\text{u'02}}$ , unscented’02 are presented in Appendix B.

## 2.2 GROUND TRUTH(S) $Y_0$ AND $Y_1$

The true distribution is

$$Y_0 = f(X),$$

and the pseudo-true distribution is

$$Y_1 = \mathcal{N}(\mathbb{E} Y_0, \text{Cov} Y_0).$$

Whereas  $Y_0$  is the ideal answer to Problem 1,  $Y_1$  is the closest Gaussian approximation (by KL divergence) to  $Y_0$ . We obtain  $Y_0$  and  $Y_1$  in baselines by quasi-Monte Carlo simulation.

In §5.1, we evaluate each method by computing the KL divergence of its Normal approximation from  $Y_1$ . This measures how close the method’s Normal approximation is to the best Normal approximation. We also compare each Normal approximation to the true (non-Normal) distribution  $Y_0$  using Wasserstein distance.

## 3 THEORETICAL GUARANTEES

In Appendix I, we give a theoretical bound on the dissimilarity between the laws of  $Y_0$  and  $Y$ , as measured in Wasserstein distance:

**Definition 4.** Let  $X$  and  $Y$  be random variables taking values  $\mathbb{R}^n$ . The Wasserstein distance  $d_W(X, Y)$  is

$$d_W(X, Y) = \sup_{\| \nabla h \|_\infty \leq 1} \mathbb{E}(h(X) - h(Y)).$$

We prove that  $d_W(Y_0, Y)$  can be decomposed as the final step of a recursion

$$\begin{aligned} \text{error at layer } k \leq & (\text{Lipschitz constant of layer } k) (\text{error at layer } k - 1) \\ & + (\text{non-normality of layer } k) \end{aligned}$$

where all of the terms can be bounded in terms of the input distribution and the network weights. Even though this bound is too loose to be depended upon in practice, it lends attribution to the sources of error in our approach; in particular, that non-normality arises as a multilayer interaction between variance and nonlinearity.

## 4 ADVERSARIAL EXAMPLES

The Introduction claims that linear and unscented propagation are only exact for low-order (1 and 2, respectively) polynomials. We support this claim by applying them to general smooth functions.

**Example 1** (for  $Y_{\text{lin}}$ ). Consider the network  $Y = \sin(X)$ , where  $X \sim \mathcal{N}(0, \sigma^2)$ . Then  $Y_{\text{lin}} = \mathcal{N}(0, \sigma^2)$ . But in fact  $\text{Var } Y = (1 - e^{-\sigma^2})/2$  which tends to  $1/2$  for large  $\sigma^2$ . So by increasing  $\sigma^2$ , we can make  $Y_{\text{lin}}$  arbitrarily wrong.

**Example 2** (for  $Y_{\text{u95}}$  and  $Y_{\text{u02}}$ ). Suppose that  $X \sim \mathcal{N}(0, 1)$ , and the sigma points are  $X \in \{-\alpha, 0, \alpha\}$  for some  $\alpha > 0$ . Then on the neural network  $Y = \sin(\alpha^{-1}\pi X)$ ,  $Y_{\text{u95}}$  and  $Y_{\text{u02}}$  will be identically zero and arbitrarily wrong.

The Introduction claims that mean-field propagation is only exact for networks consisting of a single neuron. We support this claim by applying it to a network with multiple neurons.

**Example 3** (for  $Y_{\text{mfa}}$ ). Consider the following (linear) network, with scalar input  $X$ , hidden  $Y^1 \in \mathbb{R}^m$ , and scalar output  $Y$ .

$$\begin{aligned} Y &= \frac{1}{m} \sum_{i=1}^m Y_i^1 \\ Y_i^1 &= X \\ X &\sim \mathcal{N}(0, 1) \end{aligned}$$

The mean-field approximation treats each  $Y_i^1$  as independent  $\mathcal{N}(0, 1)$ , so it concludes  $Y \sim \mathcal{N}(0, m^{-1})$ . But  $Y$  is identical to  $X \sim \mathcal{N}(0, 1)$ . So by increasing  $m$ , we can make  $Y_{\text{mfa}}$  arbitrarily wrong.

However the mean and variance of our method  $Y_{\text{ana}}$  are exact on all single-hidden-layer networks, which includes all of the examples above. Unlike the works cited in Appendix A, we push our method past the breaking point, by an explicit multi-layer network combining strong non-normality with strong nonlinearity.

**Example 4** (for  $Y_{\text{ana}}$ ). Consider the following network, where input  $X$ , output  $Y$ , and hidden  $Y^1$  are scalars;  $u(x) = \mathbf{1}_{x \geq 0}$  is the Heaviside function; and  $\alpha$  is a weight.

$$\begin{aligned} Y &= \alpha u(Y^1 - 3) \\ Y^1 &= 2u(X) \\ X &\sim \mathcal{N}(0, 1) \end{aligned}$$

At the hidden layer,  $Y^1$  is approximated by  $\mathcal{N}(1, 1)$ . Therefore  $Y_{\text{ana}}$  is approximated by some nondegenerate Normal distribution, scaled by  $\alpha$ . By increasing  $\alpha$ , we can make  $Y_{\text{ana}}$  arbitrarily wrong:  $Y_0$  is identically zero because  $Y^1 - 3$  is always negative.

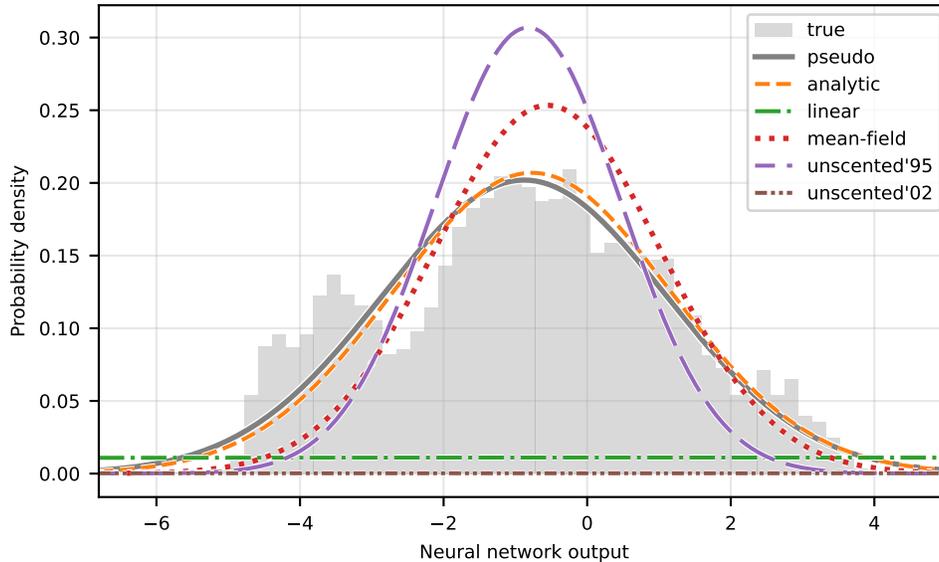
270  
271  
272  
273  
274  
275  
276  
277  
278

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(\cdot \  Y_1)$
pseudo-true ( $Y_1$ )	$9.531 \times 10^{-2} \pm 2.9 \times 10^{-5}$	0
analytic	$1.111 \times 10^{-1} \pm 3.5 \times 10^{-5}$	$2.086 \times 10^{-3} \pm 1.2 \times 10^{-6}$
mean-field	$3.101 \times 10^{-1} \pm 3.7 \times 10^{-5}$	$5.797 \times 10^{-2} \pm 7.0 \times 10^{-6}$
linear	$1.929 \times 10^{+1} \pm 2.5 \times 10^{-4}$	$1.627 \times 10^{+2} \pm 7.7 \times 10^{-3}$
unscented'95	$4.135 \times 10^{-1} \pm 3.2 \times 10^{-5}$	$1.353 \times 10^{-1} \pm 1.3 \times 10^{-5}$
unscented'02	$8.884 \times 10^{+2} \pm 1.0 \times 10^{-2}$	$3.067 \times 10^{+5} \pm 1.4 \times 10^{+1}$

279  
280  
281  
282  
283

Table 1: Summary statistics for Network (architecture=small, weights=trained, activation=probit residual), variance=large.  $d_W(\cdot, Y_0)$  is the Wasserstein distance to  $Y_0$ , the ground truth (lower is better), and  $d_{KL}(\cdot \| Y_1)$  is the KL divergence from  $Y_1$ , the pseudo-true Normal distribution having ground truth moments (lower is better).

284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302



303

Figure 1: Probability distributions for Network (architecture=small, weights=trained, activation=probit residual), variance=large.

304  
305

## 5 EXAMPLES, APPLICATIONS, AND EXTENSIONS

306

307

### 5.1 RANDOM NETWORKS

308

309

310

311

We apply our method and other benchmarks to 38 different ensembles of random neural networks with more than one hidden layer. They are designed to cover a range of architectures, weights, and activations, and to stress-test the assumptions of layer-by-layer moment matching. We sample one neural network from each ensemble and evaluate the goodness of approximation of the output distribution for three input distributions. In each case we evaluate  $Y_0, Y_1, Y_{ana}, Y_{mfa}, Y_{lin}, Y_{u95}$ , and  $Y_{u02}$ , and compare distance to the true distribution  $d_W(\cdot, Y_0)$  and KL divergence from the pseudo-true distribution  $d_{KL}(\cdot, Y_1)$ .

312

313

314

315

316

317

318

See Appendix N for full specifications of the 114 test cases, the Monte Carlo methodology, and the full results. For the case of large input variance, across all networks, our method is typically **one hundred** times better than Unscented'95, **ten thousand** times better than linear, and **one million** times better than Unscented'02 (measured by KL divergence from the pseudo-true distribution), as seen in Fig. 2. Visualizations for the medium- and small- input variance cases are available in Appendix N.1.

319

320

321

322

323

324  
 325  
 326  
 327  
 328  
 329  
 330  
 331  
 332  
 333  
 334  
 335  
 336  
 337  
 338  
 339  
 340  
 341  
 342  
 343  
 344  
 345  
 346  
 347  
 348  
 349  
 350  
 351  
 352  
 353  
 354  
 355  
 356  
 357  
 358  
 359  
 360  
 361  
 362  
 363  
 364  
 365  
 366  
 367  
 368  
 369  
 370  
 371  
 372  
 373  
 374  
 375  
 376  
 377

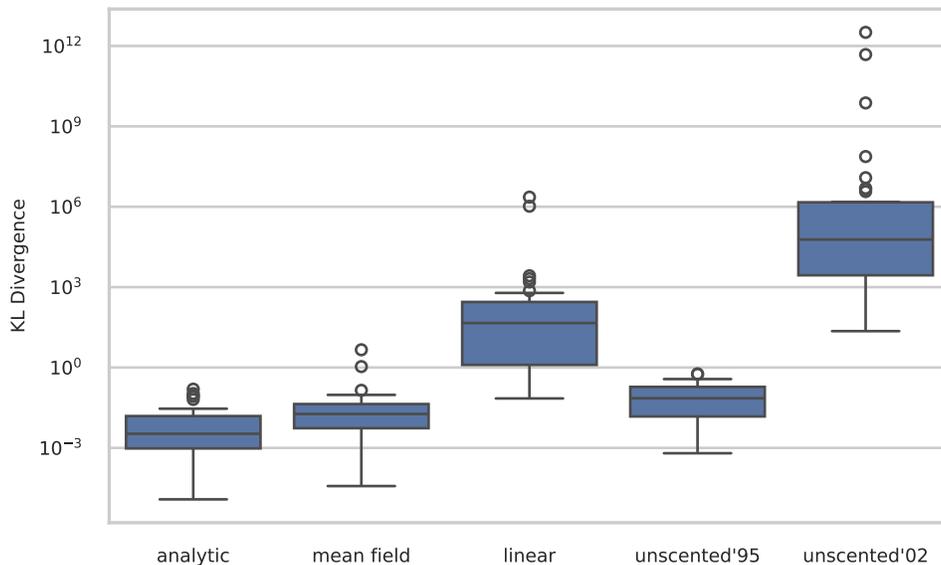


Figure 2: Comparison of goodness of approximation (lower KL divergence is better) for all random neural networks, grouped by approximation method, in the large input variance scenario.

One example, which we reproduce here in Table 1 and Fig. 1, is the large-variance case of a small network with trained weights and a probit-residual activation function, in which the output distribution is evidently non-normal, but our method is still closest to the true distribution (in Wasserstein distance) and to the pseudo-true distribution (in KL divergence).

Many examples; such as the small-variance test case with wide architecture, trained weights, and sine activation; exhibit the underdispersion of the mean-field approximation predicted in Example 3. It is remarkable that even in the small-variance regime when our method, linearization, and the unscented transformations are all justified by the delta method and asymptotically equivalent, our method is often still closest to the pseudo-true distribution.

## 5.2 INPUT UNCERTAINTY: REGRESSION

A generative neural network trained for regression on noiseless inputs is used to make predictions on noisy inputs. At inference, the network is provided with a perturbed input and the covariance matrix of the perturbation. The prediction is a distribution over the output.

We apply this procedure to the California Housing dataset (Kelley Pace & Barry, 1997). We report the average log probability density of the test  $y$  under the predictive distribution  $\hat{Y}$  in Table. 2, as well as coverage and interval width for nominal 95% confidence intervals. The analytic moment propagation method has the highest log probability density and the closest-to-nominal coverage. All uncertainty propagation methods outperform “certain,” which ignores the uncertainty in the input.

The model specification, training, inference, and Monte Carlo method are detailed in Appendix. J.

## 5.3 INPUT UNCERTAINTY: BINARY CLASSIFICATION

A neural network trained for binary classification (Bernoulli regression) is used to make predictions on a shifted distribution of missing features. The missing features are filled by linear regression, leading to input uncertainty that needs to be propagated to the output probability  $\hat{p}$ . Uncertainty propagation is needed because the calibration of  $\hat{p}$  matters in applications such as betting on the outcome of a binary event (Cover & Thomas, 2006, Example 6.1.1).

We apply this to the Taiwanese bankruptcy prediction dataset (Liang et al., 2016).

Method	log pdf	coverage (%)	interval width
certain	$-4.452 \pm 4.5 \times 10^{-2}$	$49.7 \pm 2.0 \times 10^{-1}$	1.67
analytic	$-1.420 \pm 7.6 \times 10^{-3}$	$96.0 \pm 1.5 \times 10^{-1}$	$4.06 \pm 3.5 \times 10^{-3}$
mean field	$-1.647 \pm 2.7 \times 10^{-3}$	$99.8 \pm 4.0 \times 10^{-2}$	$6.87 \pm 4.4 \times 10^{-4}$
linear	$-1.851 \pm 3.2 \times 10^{-3}$	$97.3 \pm 5.4 \times 10^{-2}$	$7.73 \pm 5.5 \times 10^{-3}$
unscented'95	$-1.457 \pm 8.8 \times 10^{-3}$	$93.6 \pm 1.8 \times 10^{-1}$	$3.89 \pm 4.3 \times 10^{-3}$
unscented'02	$-2.529 \pm 1.2 \times 10^{-3}$	$99.7 \pm 1.2 \times 10^{-2}$	$18.85 \pm 1.5 \times 10^{-2}$

Table 2: Performance of different uncertainty propagation methods on the California housing regression problem. Both coverage (closer to 95% is better) and interval width (smaller is better) are reported for nominal 95% confidence intervals.

Method for predicting $\hat{p}$	log probability of correct label
certain	$-0.202 \pm 2.9 \times 10^{-2}$
analytic	$-0.145 \pm 1.8 \times 10^{-2}$
mean field	$-0.163 \pm 2.2 \times 10^{-2}$
linear	$-0.202 \pm 2.9 \times 10^{-2}$
unscented'95	$-0.159 \pm 2.1 \times 10^{-2}$
unscented'02	$-0.144 \pm 1.8 \times 10^{-2}$

Table 3: Average (with standard error) log predicted probability (higher is better) of the correct label on test data in the Taiwanese bankruptcy dataset.

The missing data dramatically degrades the discrimination of  $\hat{p}$  across all uncertainty propagation methods, decreasing the area under the receiver operating characteristic by roughly 0.2 (App. K, Fig. 5). However, the analytic uncertainty-aware prediction has the best-calibrated probabilities up to sampling and numerical uncertainty in our results (Table. 3). All uncertainty propagation methods outperform “certain” prediction, which ignores the uncertainty in the imputation step.

We give the details of the model specification, training, and inference in Appendix. K.

#### 5.4 WEIGHT UNCERTAINTY: VARIATIONAL BAYES NETWORKS

We apply our distributional approximation to variational inference for Bayesian networks. In this case, the input  $x$  is fixed and the weights  $w$  are random with a prior distribution  $p(w)$ . The network predicts  $y \sim p(y | w, x)$ . The posterior distribution is approximated variationally as  $q(w; \theta)$  by optimizing the evidence lower bound (Blundell et al., 2015):

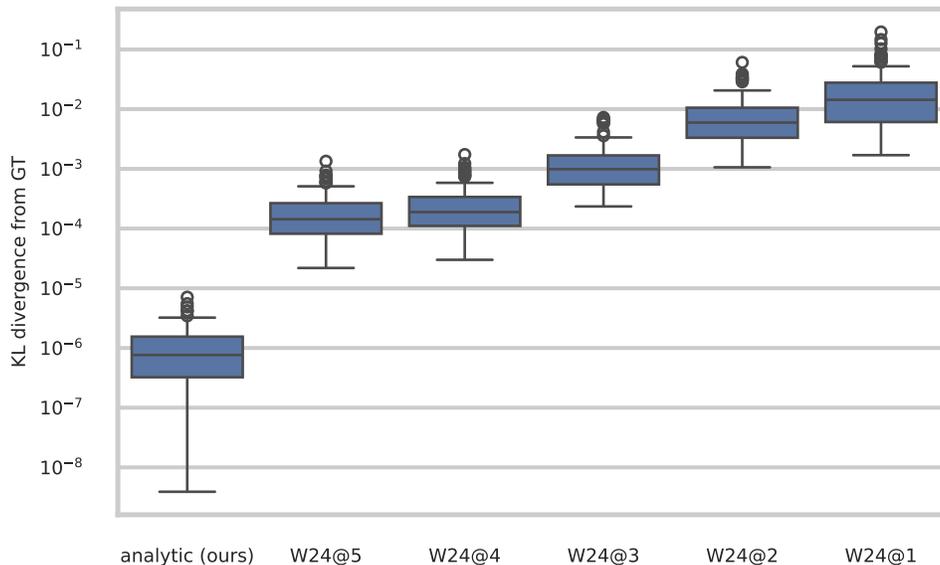
$$\theta^* = \arg \min_{\theta} \left\{ \mathbb{E}_{w \sim q(w; \theta)} \log p(y | w, x) + D_{\text{KL}, w}(q(w; \theta) | p(w)) \right\}. \quad (2)$$

Our example uses the GeLU activation function and, like Wright et al. (2024, §4.3), a single hidden layer; for other details, see App. L.

Monte Carlo Variational Inference (MCVI) is a stochastic gradient method that approximates  $\mathbb{E}_{w \sim q(w; \theta)}$  with Monte Carlo samples. In order to reduce the gradient variance and computational cost, Wu et al. (2019); Petersen et al. (2024); Wright et al. (2024); Rui Li & Trapp (2025) use moment-matching deterministic approximations to the evidence lower bound, and then apply the same deterministic approximation to evaluate the predictive distribution. While this benchmarking strategy assesses the end-to-end learning process, it does not necessarily reflect accurate distributional approximations: intuitively, we can expect a sufficiently expressive variational model to be *robust* to systematic errors in the distribution propagation. On one hand, this robustness favors end-to-end inference (training and testing with the same distributional approximation), but on the other hand, it means that end-to-end inference fails to interrogate whether the approximation itself is accurate.<sup>3</sup>

<sup>3</sup>“Learning can still improve a bound on the log likelihood of the data even when the posterior distribution over hidden states is computed incorrectly”; see also the surrounding discussion in Frey & Hinton (1999, §1.2).

432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450



451  
452  
453  
454

Figure 3: KL divergence (lower is better) between pseudo-true (ground truth moments) predictive distribution (by Monte Carlo) and approximations for the concrete compressive strength dataset.  $W24@k$  means the  $k$ th partial sum of the GeLU covariance series of Wright et al. (2024, Appendix B.3).

455  
456  
457  
458

Because our interest is in the correctness of distributional approximation, our experiments differ from the works cited in the following ways:

459  
460  
461  
462  
463  
464  
465  
466  
467  
468

1. We train the Bayesian network using MCVI with a large Monte Carlo batch size. After training has converged, the Monte Carlo predictions from this network are taken as ground truth mean and variance.
2. We test the Bayesian network using Monte Carlo on the variational posterior, as well as six different techniques to propagate distributions through activation functions: the power series of Wright et al. (2024), expanded to 1–5 terms, and our method.
3. The figure of merit is the KL divergence (lower is better) between the pseudo-true Gaussian distribution (via Monte Carlo) and each deterministic approximation.

469  
470  
471  
472  
473  
474  
475

We applied this method to four regression datasets from the UCI Machine Learning Repository. The data references and full results are in App. L; we highlight only the concrete compressive strength dataset here. As Fig. 3 shows, the power series expansion of Wright et al. (2024) shows a “dose response” i.e. becomes more accurate as more terms are added. But the most accurate approximation to the Monte Carlo predictive distribution is attained using exact moment matching, our method. It is not exact because, for a Bayes network with even a single hidden layer, we use a moment-matching Gaussian multiplication approximation (Goulet et al., 2021, eqq. 3–6).

476  
477

## 5.5 STOCHASTIC ACTIVATIONS

478  
479  
480  
481  
482

As publicly acclaimed last year, the earliest conceptions of an artificial neural networks were, like biological neural networks, stochastic (Davour). Later, a neuron’s stochastic activation was replaced by a deterministic sigmoid that represented its average behavior. Today, the biological similitude has ceased to be a driving motivation, yet as a curiosity, we analyze the output distribution of a neural

483  
484  
485

For a toy example, take  $Z \sim \mathcal{N}(0, 1)$ ; if the uncertainty propagation formula were  $aZ \sim \mathcal{N}(0, 2a)$ , which is patently incorrect, the variational model could learn to approximate the target distribution  $\mathcal{N}(0, \sigma^2)$  using the parameter  $a = \sigma^2/2$ . If the approximations were adequate, one would expect the different inference methods to be indistinguishable from each other and from MCVI, as they are in Wright et al. (2024, Table 3).

486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

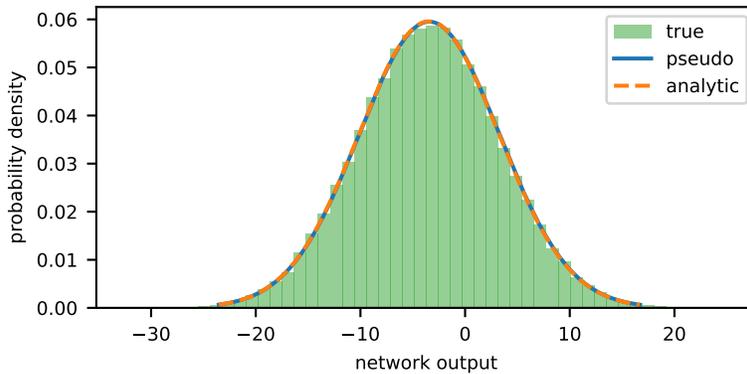


Figure 4: Output distribution of a stochastic neural network, pseudo-true Normal distribution (“pseudo”), and layer-by-layer moment-matched Normal distribution (“analytic”).

network whose activations are random processes modeled after a stochastic neuron:

$$\tilde{\sigma}(x, U) = 2\mathbf{1}_{U < \Phi(x)} - 1, \tag{3}$$

where at each artificial neuron,  $U \sim \text{Uniform}(0, 1)$  is an independent random variable.

In Appendix. M we derive a moment-matching approximation to the distribution of the output of a stochastic neural network. Applying this formula to 1 million samples from a stochastic version of the “deep” neural network of §5.1 with a constant zero input results in a normal distribution with a good subjective agreement (Fig. 4). Whether this line of inquiry deserves further methodological development we reserve for future work.

## 6 NOVELTY AND SIGNIFICANCE

Until now, the first and second moments of a neural network layer have been approximated in various ways (independence, approximate activation function, truncated power series, etc.). We compute them exactly for many popular activation functions. Our networks also generalize from previous work in uncertainty propagation by allowing for residual connections, which are common in modern neural networks.

This discovery enables layer-wise moment matching in deep networks, which we demonstrate to be orders of magnitude more accurate than other uncertainty propagation methods.

We finally demonstrate that this uncertainty propagation method is effective for uncertainty-aware inference on real data, as well as for deterministic inference on variational Bayes networks.

---

540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593

## 7 REPRODUCIBILITY STATEMENT

In the Supplementary Material, executable Python scripts are in `demo/`. They reference libraries in `src/` and generate output in `docs/`. `test_case.py` and `analyze_results.py` generate the results in §N. To reproduce §J, run `california_housing.py`. To reproduce §K, run `classification.py`. To reproduce §M, run `stochastic.py`. [Full reproduction code for the discussion version will be posted later in the ICLR review cycle.](#)

594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647

---

## REFERENCES

- Ahmed Hussen Abdelaziz, Shinji Watanabe, John R. Hershey, Emmanuel Vincent, and Dorothea Kolossa. Uncertainty propagation through deep neural networks. In *Interspeech 2015*, pp. 3561–3565. ISCA, September 2015. doi:10.21437/Interspeech.2015-706.
- Abdullah Akgül, Manuel Haußmann, and Melih Kandemir. Deterministic Uncertainty Propagation for Improved Model-Based Offline Reinforcement Learning, January 2025. arXiv:2406.04088 [cs].
- Ramón Fernandez Astudillo and João Paulo Da Silva Neto. Propagation of uncertainty through multilayer perceptrons for robust automatic speech recognition. In *Interspeech 2011*, pp. 461–464, ISCA, August 2011. ISCA. doi:10.21437/interspeech.2011-196.
- Angeliki Xifara Athanasios Tsanas. *Energy Efficiency*, 2012.
- Richard Bergna, Stefan Depeweg, Sergio Calvo Ordoñez, Jonathan Plenk, Alvaro Cartea, and José Miguel Hernández-Lobato. Post-Hoc Uncertainty Quantification in Pre-Trained Neural Networks via Activation-Level Gaussian Processes. In *7th Symposium on Advances in Approximate Bayesian Inference – Workshop Track*, 2025.
- Adel Bibi, Modar Alfadly, and Bernard Ghanem. Analytic Expressions for Probabilistic Moments of PL-DNN with Gaussian Input. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9099–9107, Salt Lake City, UT, June 2018. IEEE. doi:10.1109/cvpr.2018.00948.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622, Lille, France, July 2015. PMLR.
- Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, N.J, second edition edition, 2006. ISBN 978-0-471-74882-3 978-0-471-74881-6 978-1-118-58577-1 978-0-471-24195-9. doi:10.1002/047174882X.
- Anna Davour. The Nobel Prize in Physics 2024 - Popular science background.
- DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020.
- Zvi Drezner and G. O. Wesolowsky. On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation*, 35(1-2):101–107, March 1990. ISSN 0094-9655. doi:10.1080/00949659008811236. Publisher: Taylor & Francis eprint: <https://doi.org/10.1080/00949659008811236>.
- Michael Feischl and Fabian Zehetgruber. Computational Math with Neural Networks is Hard, May 2025. arXiv:2505.17751 [math] version: 1.
- Brendan J. Frey and Geoffrey E. Hinton. Variational Learning in Nonlinear Gaussian Belief Networks. *Neural Computation*, 11(1):193–213, January 1999. ISSN 0899-7667, 1530-888X. doi:10.1162/089976699300016872.
- Carl Friedrich Gauss. *Theory of the motion of the heavenly bodies moving about the sun in conic sections: a translation of Gauss’s” Theoria Motus.” with an appendix*. Little, Brown, 1857.
- James-A. Goulet, Luong Ha Nguyen, and Saeid Amiri. Tractable Approximate Gaussian Inference for Bayesian Neural Networks. *Journal of Machine Learning Research*, 22(251):1–23, 2021.

- 
- 648 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing  
649 Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, December 2015. doi:10.1109/ICCV.2015.123.  
650 ISSN: 2380-7504.  
651
- 652 Dan Hendrycks and Kevin Gimpel. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units, 2017.  
653  
654
- 655 Marco F. Huber. Bayesian Perceptron: Towards fully Bayesian Neural Networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 3179–3186, December 2020.  
656 doi:10.1109/CDC42340.2020.9303764. ISSN: 2576-2370.  
657
- 658 I-Cheng Yeh. Concrete Compressive Strength, 1998.  
659
- 660 Shida Jiang, Junzhe Shi, and Scott Moura. A New Framework for Nonlinear Kalman Filters, February 2025. arXiv:2407.05717 [eess].  
661  
662
- 663 S. Julier, J. Uhlmann, and H.F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, 45(3):  
664 477–482, March 2000. ISSN 0018-9286. doi:10.1109/9.847726. Publisher: Institute of Electrical  
665 and Electronics Engineers (IEEE).  
666
- 667 Simon J. Julier and Jeffrey K. Uhlmann. New extension of the Kalman filter to nonlinear systems. In *Defense, Security, and Sensing*, 1997.  
668  
669
- 670 S.J. Julier. The scaled unscented transformation. In *Proceedings of the 2002 American Control Conference*, volume 6, pp. 4555–4559 vol.6, May 2002. doi:10.1109/ACC.2002.1025369. ISSN:  
671 0743-1619.  
672
- 673 S.J. Julier and J.K. Uhlmann. Unscented Filtering and Nonlinear Estimation. *Proceedings of the IEEE*, 92(3):401–422, March 2004. ISSN 0018-9219. doi:10.1109/JPROC.2003.823141.  
674  
675
- 676 S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte. A new approach for filtering nonlinear systems. In *Proceedings of 1995 American Control Conference - ACC'95*, volume 3, pp. 1628–1632,  
677 Seattle, WA, USA, 1995. American Autom Control Council. doi:10.1109/acc.1995.529783.  
678
- 679 Paul Jungmann, Julia Poray, and Akash Kumar. Analytical Uncertainty Propagation in Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2495–2508, February 2025. ISSN 2162-2388. doi:10.1109/TNNLS.2023.3347156.  
680  
681  
682
- 683 Toni Karvonen and Simo Särkkä. Wasserstein bounds for non-linear Gaussian filters, March 2025. arXiv:2503.21643 [math].  
684
- 685 R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, May 1997. ISSN 01677152. doi:10.1016/S0167-7152(96)00140-X.  
686  
687
- 688 Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2):561–572, July 2016. ISSN 0377-2217. doi:10.1016/j.ejor.2016.01.012.  
689  
690
- 691 Lennart Ljung. unscentedKalmanFilter. In *System Identification Toolbox Reference*, pp. 1–2372. The MathWorks, Inc., 2025.  
692  
693
- 694 Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.  
695
- 696 Bengt Muthén. Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology*, 43(1):131–143, May 1990. ISSN 0007-1102, 2044-8317. doi:10.1111/j.2044-8317.1990.tb00930.x.  
697  
698  
699
- 700 Tobias Nagel and Marco F. Huber. Kalman-Bucy-Informed Neural Network for System Identification. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 1503–1508, December 2022. doi:10.1109/CDC51059.2022.9993245. ISSN: 2576-2370.  
701

---

702 Ivan Nourdin, Giovanni Peccati, and Gesine Reinert. Second order Poincaré inequalities and CLTs  
703 on Wiener space. *Journal of Functional Analysis*, 257(2):593–609, July 2009. ISSN 00221236.  
704 doi:10.1016/j.jfa.2008.12.017.  
705

706 Ivan Nourdin, Giovanni Peccati, and Anthony Réveillac. Multivariate normal approximation using  
707 Stein’s method and Malliavin calculus. *Annales de l’Institut Henri Poincaré, Probabilités et*  
708 *Statistiques*, 46(1):45–58, February 2010. ISSN 0246-0203. doi:10.1214/08-AIHP308. Publisher:  
709 Institut Henri Poincaré.  
710

711 David Nualart. *The Malliavin calculus and related topics*. Probability and its applications. Springer,  
712 Berlin ; New York, 2nd ed edition, 2006. ISBN 978-3-540-28328-7.  
713

714 D. B. Owen. A table of normal integrals: A table. *Communications in Statistics -*  
715 *Simulation and Computation*, 9(4):389–419, January 1980. ISSN 0361-0918, 1532-4141.  
716 doi:10.1080/03610918008812164.

717 Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Taming the waves: sine as  
718 activation function in deep neural networks, 2017.  
719

720 A. Cerdeira Paulo Cortez. Wine Quality, 2009.  
721

722 Felix Petersen, Aashwin Mishra, Hilde Kuehne, Christian Borgelt, Oliver Deussen, and Mikhail  
723 Yurochkin. Uncertainty Quantification via Stable Distribution Propagation, February 2024.  
724 arXiv:2402.08324 [cs].

725 Heysem Kaya Pnar Tfekci. Combined Cycle Power Plant, 2014.  
726

727 William H. Press (ed.). *Numerical recipes: the art of scientific computing*. Cambridge University  
728 Press, Cambridge, 3. ed edition, 2007. ISBN 978-0-521-88068-8.  
729

730 Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*.  
731 Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 3. print edition, 2008.  
732 ISBN 978-0-262-18253-9.  
733

734 Arno Solin Rui Li, Marcus Klasson and Martin Trapp. Streamlining Prediction in Bayesian Deep  
735 Learning. In *International Conference on Learning Representations (ICLR)*, 2025.  
736

737 Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Im-  
738 plicit Neural Representations with Periodic Activation Functions. In *Advances in Neural Infor-*  
739 *mation Processing Systems*, volume 33, pp. 7462–7473. Curran Associates, Inc., 2020.

740 Michael L. Stein. *Interpolation of Spatial Data*. Springer Series in Statistics. Springer, New York,  
741 NY, 1999. ISBN 978-1-4612-7166-6 978-1-4612-1494-6. doi:10.1007/978-1-4612-1494-6.  
742

743 John R. Taylor. *An introduction to error analysis: the study of uncertainties in physical measure-*  
744 *ments*. University Science Books, Sausalito, Calif, 2. ed edition, 1997. ISBN 978-0-935702-75-0  
745 978-0-935702-42-2.  
746

747 Jessica S. Titensky, Hayden Jananathan, and Jeremy Kepner. Uncertainty Propagation in Deep Neural  
748 Networks Using Extended Kalman Filtering. In *2018 IEEE MIT Undergraduate Research Tech-*  
749 *nology Conference (URTC)*, pp. 1–4, October 2018. doi:10.1109/URTC45901.2018.9244804.

750 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilis-  
751 tic Mathematics. Cambridge University Press, Cambridge, 1998. ISBN 978-0-521-78450-4.  
752 doi:10.1017/CBO9780511802256.  
753

754 Aad W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer  
755 Series in Statistics. Springer, New York, NY, 1996. ISBN 978-1-4757-2547-6 978-1-4757-2545-  
2. doi:10.1007/978-1-4757-2545-2.

---

756 Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau,  
757 Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der  
758 Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson,  
759 Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore,  
760 Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero,  
761 Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt,  
762 and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing  
763 in Python. *Nature Methods*, 17:261–272, 2020. doi:10.1038/s41592-019-0686-2.

764 Philipp Wagner, Xinyang Wu, and Marco F. Huber. Kalman Bayesian Neural Networks for Closed-  
765 form Online Learning, November 2022. arXiv:2110.00944 [cs].

766  
767 E.A. Wan and R. Van Der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pp. 153–158, Lake Louise, Alta., Canada, 2000. IEEE. ISBN 978-0-7803-5800-3. doi:10.1109/ASSPCC.2000.882463.

768  
769  
770  
771 Oren Wright, Yorie Nakahira, and José M. F. Moura. An Analytic Solution to Covariance Propagation in Neural Networks. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 4087–4095. PMLR, April 2024. ISSN: 2640-3498.

772  
773  
774  
775 Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E. Turner, José Miguel Hernández-Lobato,  
776 and Alexander L. Gaunt. Deterministic Variational Inference for Robust Bayesian Neural Networks, March 2019. arXiv:1810.03958 [cs].

777  
778 Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural Networks Fail to Learn Periodic Functions  
779 and How to Fix It. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1583–1594. Curran Associates, Inc., 2020.

780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

---

## SUPPLEMENTARY MATERIAL

Owing to the sheer length of the supplementary material, which includes a programmatically generated section of exhaustive random neural network test cases (§N), we begin with a table of contents to re-orient the reader.

### CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Methodology</b>	<b>2</b>
2.1	Our analytic method $Y_{\text{ana}}$ . . . . .	3
2.2	Ground truth(s) $Y_0$ and $Y_1$ . . . . .	4
<b>3</b>	<b>Theoretical Guarantees</b>	<b>4</b>
<b>4</b>	<b>Adversarial examples</b>	<b>5</b>
<b>5</b>	<b>Examples, applications, and extensions</b>	<b>6</b>
5.1	Random networks . . . . .	6
5.2	Input uncertainty: regression . . . . .	7
5.3	Input uncertainty: binary classification . . . . .	7
5.4	Weight uncertainty: variational Bayes networks . . . . .	8
5.5	Stochastic activations . . . . .	9
<b>6</b>	<b>Novelty and significance</b>	<b>10</b>
<b>7</b>	<b>Reproducibility statement</b>	<b>11</b>
<b>A</b>	<b>Supplement to literature review</b>	<b>24</b>
<b>B</b>	<b>Baselines</b>	<b>24</b>
B.1	The mean-field approximation $Y_{\text{mfa}}$ . . . . .	24
B.2	The linear approximation $Y_{\text{lin}}$ . . . . .	24
B.3	The unscented approximation(s) $Y_{\text{u95}}$ and $Y_{\text{u02}}$ . . . . .	25
<b>C</b>	<b>Preliminaries: Gaussian integrals</b>	<b>25</b>
<b>D</b>	<b>Derivation of uncertainty propagation formulas for probit activation</b>	<b>29</b>
<b>E</b>	<b>Derivation of uncertainty propagation formulas for GeLU activation</b>	<b>29</b>
<b>F</b>	<b>Derivation of uncertainty propagation formulas for ReLU activation</b>	<b>33</b>
<b>G</b>	<b>Derivation of uncertainty propagation formulas for Heaviside activation</b>	<b>35</b>

---

864	<b>H Derivation of uncertainty propagation formulas for sine activation</b>	<b>36</b>
865		
866	<b>I Theoretical Guarantees</b>	<b>38</b>
867		
868	I.1 Recursive triangle inequality . . . . .	38
869	I.2 Lipschitz constant . . . . .	38
870		
871	I.3 Non-normality . . . . .	38
872		
873	<b>J Supplement to §5.2</b>	<b>42</b>
874		
875	J.1 Data . . . . .	42
876	J.2 Network . . . . .	42
877	J.3 Training . . . . .	42
878		
879	J.4 Inference . . . . .	43
880		
881	<b>K Supplement to §5.3</b>	<b>44</b>
882		
883	K.1 Data . . . . .	44
884	K.2 Network . . . . .	44
885	K.3 Training . . . . .	44
886		
887	K.4 Reporting . . . . .	44
888	K.5 Additional figures . . . . .	44
889		
890	<b>L Supplement to §5.4</b>	<b>46</b>
891		
892	L.1 Results . . . . .	47
893		
894	<b>M Supplement to §5.5</b>	<b>50</b>
895		
896	<b>N Supplement to §5.1</b>	<b>51</b>
897		
898	N.1 Summaries . . . . .	53
899		
900	<b>O Comments on computational complexity</b>	<b>163</b>
901		
902	<b>P LLM usage statement</b>	<b>164</b>
903		
904		
905		
906		
907		
908		
909		
910		
911		
912		
913		
914		
915		
916		
917		

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

LIST OF FIGURES

1 Probability distributions for `Network(architecture=small, weights=trained, activation=probit residual), variance=large`. . . . . 6

2 Comparison of goodness of approximation (lower KL divergence is better) for all random neural networks, grouped by approximation method, in the large input variance scenario. . . . . 7

3 KL divergence (lower is better) between pseudo-true (ground truth moments) predictive distribution (by Monte Carlo) and approximations for the concrete compressive strength dataset.  $W24@k$  means the  $k$ th partial sum of the GeLU covariance series of Wright et al. (2024, Appendix B.3). . . . . 9

4 Output distribution of a stochastic neural network, pseudo-true Normal distribution (“pseudo”), and layer-by-layer moment-matched Normal distribution (“analytic”). . . . . 10

5 Receiver operator characteristic curve of test data in the Taiwanese bankruptcy dataset based on varying the threshold for  $\hat{p}$ . . . . . 45

6 KL divergence between ground truth predictive distribution (by Monte Carlo) and approximations for the combined cycle power plant dataset.  $W24@k$  means Wright et al. (2024) with  $k$  terms in the series expansion. . . . . 48

7 KL divergence between ground truth predictive distribution (by Monte Carlo) and approximations for the energy efficiency dataset.  $W24@k$  means Wright et al. (2024) with  $k$  terms in the series expansion. . . . . 48

8 KL divergence between ground truth predictive distribution (by Monte Carlo) and approximations for the wine quality dataset.  $W24@k$  means Wright et al. (2024) with  $k$  terms in the series expansion. . . . . 49

9 Comparison of goodness of approximation (lower KL divergence is better) for all random neural networks, grouped by approximation method, in the small input variance scenario. . . . . 53

10 Comparison of goodness of approximation (lower KL divergence is better) for all random neural networks, grouped by approximation method, in the medium input variance scenario. . . . . 53

11 Comparison of goodness of approximation (lower KL divergence is better) for all random neural networks, grouped by approximation method, in the large input variance scenario. . . . . 54

12 Probability distributions for `Network(architecture=wide, weights=initialized, activation=probit), variance=small` . . . . . 55

13 Probability distributions for `Network(architecture=wide, weights=initialized, activation=probit), variance=medium` . . . . . 56

14 Probability distributions for `Network(architecture=wide, weights=initialized, activation=probit), variance=large` . . . . . 57

15 Probability distributions for `Network(architecture=wide, weights=trained, activation=probit), variance=small` . . . . . 58

16 Probability distributions for `Network(architecture=wide, weights=trained, activation=probit), variance=medium` . . . . . 59

17 Probability distributions for `Network(architecture=wide, weights=trained, activation=probit), variance=large` . . . . . 60

18 Probability distributions for `Network(architecture=wide, weights=initialized, activation=probit residual), variance=small` . . . . . 61

972	19	Probability distributions for Network(architecture=wide, weights=initialized, activation=probit residual), variance=medium . . . . .	62
973			
974	20	Probability distributions for Network(architecture=wide, weights=initialized, activation=probit residual), variance=large . . . . .	63
975			
976	21	Probability distributions for Network(architecture=wide, weights=trained, activation=probit residual), variance=small . . . . .	64
977			
978	22	Probability distributions for Network(architecture=wide, weights=trained, activation=probit residual), variance=medium . . . . .	65
979			
980	23	Probability distributions for Network(architecture=wide, weights=trained, activation=probit residual), variance=large . . . . .	66
981			
982	24	Probability distributions for Network(architecture=wide, weights=initialized, activation=sine), variance=small . . . . .	67
983			
984	25	Probability distributions for Network(architecture=wide, weights=initialized, activation=sine), variance=medium . . . . .	68
985			
986	26	Probability distributions for Network(architecture=wide, weights=initialized, activation=sine), variance=large . . . . .	69
987			
988	27	Probability distributions for Network(architecture=wide, weights=trained, activation=sine), variance=small . . . . .	70
989			
990	28	Probability distributions for Network(architecture=wide, weights=trained, activation=sine), variance=medium . . . . .	71
991			
992	29	Probability distributions for Network(architecture=wide, weights=trained, activation=sine), variance=large . . . . .	72
993			
994	30	Probability distributions for Network(architecture=wide, weights=initialized, activation=sine residual), variance=small . . . . .	73
995			
996	31	Probability distributions for Network(architecture=wide, weights=initialized, activation=sine residual), variance=medium . . . . .	74
997			
998	32	Probability distributions for Network(architecture=wide, weights=initialized, activation=sine residual), variance=large . . . . .	75
999			
1000	33	Probability distributions for Network(architecture=wide, weights=trained, activation=sine residual), variance=small . . . . .	76
1001			
1002	34	Probability distributions for Network(architecture=wide, weights=trained, activation=sine residual), variance=medium . . . . .	77
1003			
1004	35	Probability distributions for Network(architecture=wide, weights=trained, activation=sine residual), variance=large . . . . .	78
1005			
1006	36	Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu), variance=small . . . . .	79
1007			
1008	37	Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu), variance=medium . . . . .	80
1009			
1010	38	Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu), variance=large . . . . .	81
1011			
1012	39	Probability distributions for Network(architecture=wide, weights=trained, activation=gelu), variance=small . . . . .	82
1013			
1014	40	Probability distributions for Network(architecture=wide, weights=trained, activation=gelu), variance=medium . . . . .	83
1015			
1016	41	Probability distributions for Network(architecture=wide, weights=trained, activation=gelu), variance=large . . . . .	84
1017			
1018			
1019			
1020			
1021			
1022			
1023			
1024			
1025			

---

1026	42	Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=small . . . . .	85
1027			
1028	43	Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=medium . . . . .	86
1029			
1030			
1031	44	Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=large . . . . .	87
1032			
1033	45	Probability distributions for Network(architecture=wide, weights=trained, activation=gelu residual), variance=small . . . . .	88
1034			
1035			
1036	46	Probability distributions for Network(architecture=wide, weights=trained, activation=gelu residual), variance=medium . . . . .	89
1037			
1038	47	Probability distributions for Network(architecture=wide, weights=trained, activation=gelu residual), variance=large . . . . .	90
1039			
1040			
1041	48	Probability distributions for Network(architecture=wide, weights=initialized, activation=relu), variance=small . . . . .	91
1042			
1043	49	Probability distributions for Network(architecture=wide, weights=initialized, activation=relu), variance=medium . . . . .	92
1044			
1045	50	Probability distributions for Network(architecture=wide, weights=initialized, activation=relu), variance=large . . . . .	93
1046			
1047			
1048	51	Probability distributions for Network(architecture=wide, weights=trained, activation=relu), variance=small . . . . .	94
1049			
1050	52	Probability distributions for Network(architecture=wide, weights=trained, activation=relu), variance=medium . . . . .	95
1051			
1052	53	Probability distributions for Network(architecture=wide, weights=trained, activation=relu), variance=large . . . . .	96
1053			
1054			
1055	54	Probability distributions for Network(architecture=wide, weights=initialized, activation=relu residual), variance=small . . . . .	97
1056			
1057	55	Probability distributions for Network(architecture=wide, weights=initialized, activation=relu residual), variance=medium . . . . .	98
1058			
1059			
1060	56	Probability distributions for Network(architecture=wide, weights=initialized, activation=relu residual), variance=large . . . . .	99
1061			
1062	57	Probability distributions for Network(architecture=wide, weights=trained, activation=relu residual), variance=small . . . . .	100
1063			
1064	58	Probability distributions for Network(architecture=wide, weights=trained, activation=relu residual), variance=medium . . . . .	101
1065			
1066			
1067	59	Probability distributions for Network(architecture=wide, weights=trained, activation=relu residual), variance=large . . . . .	102
1068			
1069	60	Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside), variance=small . . . . .	103
1070			
1071	61	Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside), variance=medium . . . . .	104
1072			
1073			
1074	62	Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside), variance=large . . . . .	105
1075			
1076	63	Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside residual), variance=small . . . . .	106
1077			
1078			
1079	64	Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside residual), variance=medium . . . . .	107

---

1080	65	Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside residual), variance=large . . . . .	108
1081			
1082			
1083	66	Probability distributions for Network(architecture=deep, weights=initialized, activation=probit), variance=small . . . . .	109
1084			
1085	67	Probability distributions for Network(architecture=deep, weights=initialized, activation=probit), variance=medium . . . . .	110
1086			
1087			
1088	68	Probability distributions for Network(architecture=deep, weights=initialized, activation=probit), variance=large . . . . .	111
1089			
1090	69	Probability distributions for Network(architecture=deep, weights=trained, activation=probit), variance=small . . . . .	112
1091			
1092	70	Probability distributions for Network(architecture=deep, weights=trained, activation=probit), variance=medium . . . . .	113
1093			
1094			
1095	71	Probability distributions for Network(architecture=deep, weights=trained, activation=probit), variance=large . . . . .	114
1096			
1097	72	Probability distributions for Network(architecture=deep, weights=initialized, activation=probit residual), variance=small . . . . .	115
1098			
1099			
1100	73	Probability distributions for Network(architecture=deep, weights=initialized, activation=probit residual), variance=medium . . . . .	116
1101			
1102	74	Probability distributions for Network(architecture=deep, weights=initialized, activation=probit residual), variance=large . . . . .	117
1103			
1104	75	Probability distributions for Network(architecture=deep, weights=trained, activation=probit residual), variance=small . . . . .	118
1105			
1106			
1107	76	Probability distributions for Network(architecture=deep, weights=trained, activation=probit residual), variance=medium . . . . .	119
1108			
1109	77	Probability distributions for Network(architecture=deep, weights=trained, activation=probit residual), variance=large . . . . .	120
1110			
1111	78	Probability distributions for Network(architecture=deep, weights=initialized, activation=sine), variance=small . . . . .	121
1112			
1113			
1114	79	Probability distributions for Network(architecture=deep, weights=initialized, activation=sine), variance=medium . . . . .	122
1115			
1116	80	Probability distributions for Network(architecture=deep, weights=initialized, activation=sine), variance=large . . . . .	123
1117			
1118			
1119	81	Probability distributions for Network(architecture=deep, weights=trained, activation=sine), variance=small . . . . .	124
1120			
1121	82	Probability distributions for Network(architecture=deep, weights=trained, activation=sine), variance=medium . . . . .	125
1122			
1123	83	Probability distributions for Network(architecture=deep, weights=trained, activation=sine), variance=large . . . . .	126
1124			
1125			
1126	84	Probability distributions for Network(architecture=deep, weights=initialized, activation=sine residual), variance=small . . . . .	127
1127			
1128	85	Probability distributions for Network(architecture=deep, weights=initialized, activation=sine residual), variance=medium . . . . .	128
1129			
1130			
1131	86	Probability distributions for Network(architecture=deep, weights=initialized, activation=sine residual), variance=large . . . . .	129
1132			
1133	87	Probability distributions for Network(architecture=deep, weights=trained, activation=sine residual), variance=small . . . . .	130

---

1134	88	Probability distributions for Network(architecture=deep, weights=trained, activation=sine residual), variance=medium . . . . .	131
1135			
1136			
1137	89	Probability distributions for Network(architecture=deep, weights=trained, activation=sine residual), variance=large . . . . .	132
1138			
1139	90	Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu), variance=small . . . . .	133
1140			
1141	91	Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu), variance=medium . . . . .	134
1142			
1143			
1144	92	Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu), variance=large . . . . .	135
1145			
1146	93	Probability distributions for Network(architecture=deep, weights=trained, activation=gelu), variance=small . . . . .	136
1147			
1148			
1149	94	Probability distributions for Network(architecture=deep, weights=trained, activation=gelu), variance=medium . . . . .	137
1150			
1151	95	Probability distributions for Network(architecture=deep, weights=trained, activation=gelu), variance=large . . . . .	138
1152			
1153	96	Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=small . . . . .	139
1154			
1155			
1156	97	Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=medium . . . . .	140
1157			
1158	98	Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=large . . . . .	141
1159			
1160			
1161	99	Probability distributions for Network(architecture=deep, weights=trained, activation=gelu residual), variance=small . . . . .	142
1162			
1163	100	Probability distributions for Network(architecture=deep, weights=trained, activation=gelu residual), variance=medium . . . . .	143
1164			
1165	101	Probability distributions for Network(architecture=deep, weights=trained, activation=gelu residual), variance=large . . . . .	144
1166			
1167			
1168	102	Probability distributions for Network(architecture=deep, weights=initialized, activation=relu), variance=small . . . . .	145
1169			
1170	103	Probability distributions for Network(architecture=deep, weights=initialized, activation=relu), variance=medium . . . . .	146
1171			
1172			
1173	104	Probability distributions for Network(architecture=deep, weights=initialized, activation=relu), variance=large . . . . .	147
1174			
1175	105	Probability distributions for Network(architecture=deep, weights=trained, activation=relu), variance=small . . . . .	148
1176			
1177	106	Probability distributions for Network(architecture=deep, weights=trained, activation=relu), variance=medium . . . . .	149
1178			
1179			
1180	107	Probability distributions for Network(architecture=deep, weights=trained, activation=relu), variance=large . . . . .	150
1181			
1182	108	Probability distributions for Network(architecture=deep, weights=initialized, activation=relu residual), variance=small . . . . .	151
1183			
1184	109	Probability distributions for Network(architecture=deep, weights=initialized, activation=relu residual), variance=medium . . . . .	152
1185			
1186			
1187	110	Probability distributions for Network(architecture=deep, weights=initialized, activation=relu residual), variance=large . . . . .	153

---

1188		
1189	111	Probability distributions for Network(architecture=deep, weights=trained, activation=relu residual), variance=small . . . . . 154
1190		
1191	112	Probability distributions for Network(architecture=deep, weights=trained, activation=relu residual), variance=medium . . . . . 155
1192		
1193	113	Probability distributions for Network(architecture=deep, weights=trained, activation=relu residual), variance=large . . . . . 156
1194		
1195	114	Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside), variance=small . . . . . 157
1196		
1197	115	Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside), variance=medium . . . . . 158
1198		
1199	116	Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside), variance=large . . . . . 159
1200		
1201	117	Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside residual), variance=small . . . . . 160
1202		
1203	118	Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside residual), variance=medium . . . . . 161
1204		
1205	119	Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside residual), variance=large . . . . . 162
1206		
1207		
1208		
1209		
1210		
1211		
1212		
1213		
1214		
1215		
1216		
1217		
1218		
1219		
1220		
1221		
1222		
1223		
1224		
1225		
1226		
1227		
1228		
1229		
1230		
1231		
1232		
1233		
1234		
1235		
1236		
1237		
1238		
1239		
1240		
1241		

1242 A SUPPLEMENT TO LITERATURE REVIEW

1243

1244

Assumption	References
$\Sigma$ is small (linearized)	Titensky et al. (2018); Nagel & Huber (2022) Petersen et al. (2024); Jungmann et al. (2025) Bergna et al. (2025); Rui Li & Trapp (2025)
$\Sigma$ is small (unscented)	Astudillo & Neto (2011); Abdelaziz et al. (2015)
$\Sigma$ is diagonal	Abdelaziz et al. (2015); Huber (2020) Goulet et al. (2021); Rui Li & Trapp (2025)
$\mu = 0$	Wagner et al. (2022); Akgül et al. (2025) Bibi et al. (2018)
$\mu \rightarrow \infty$	Wu et al. (2019)
no assumptions	Wright et al. (2024) this paper

1256

1257 Table 4: Comparison of assumptions imposed on Gaussian approximations of neural network layers  
1258 with input  $\mathcal{N}(\mu, \Sigma)$ .  
1259

1260

Activation function	References
piecewise linear	Bibi et al. (2018); Huber (2020) Wright et al. (2024); Akgül et al. (2025) Wu et al. (2019)
logistic ( $\approx$ piecewise exponential)	Astudillo & Neto (2011); Abdelaziz et al. (2015)
logistic ( $\approx \Phi$ )	Huber (2020)
Heaviside	Wu et al. (2019); Wright et al. (2024)
GeLU	Wright et al. (2024)
(sin, $\Phi$ , GeLU, ReLU, Heaviside) + affine	this paper ( <b>exact</b> )

1270

1271 Table 5: Activation functions for which moment propagation has been approximated. Note two  
1272 distinct approaches to approximating the logistic function.  
1273

1274

1275 B BASELINES

1276

1277 B.1 THE MEAN-FIELD APPROXIMATION  $Y_{\text{mfa}}$

1278

1279 Following Huber (2020); Wagner et al. (2022); Akgül et al. (2025), we define the mean-field analytic  
1280 approximation by assuming that neurons in the same hidden layer are independent:

1281

1282 
$$Y_{\text{mfa}} = Y^\ell$$

1283 
$$Y^k = \mathcal{N}(\mu^k, \Sigma^k) \quad \forall k \in \{1 \dots \ell\}$$

1284 
$$\mu^k = \mathbb{E} g(Y^{k-1}; A^k, b^k, C^k, d^k)$$

1285

1286 
$$\Sigma_{ij}^k = \begin{cases} [\text{Cov } g(Y^{k-1}; A^k, b^k, C^k, d^k)]_{ij}, & i = j \\ 0, & \text{else} \end{cases} \quad (\text{mean-field assumption})$$

1287

1288 
$$Y^0 = X$$

1289

1290

1291 B.2 THE LINEAR APPROXIMATION  $Y_{\text{lin}}$

1292 Following Titensky et al. (2018); Nagel & Huber (2022); Petersen et al. (2024); Jungmann et al.  
1293 (2025), we define the linear approximation as the asymptotically Normal output distribution (pur-  
1294 suant to the delta method) in the limit of small input variance:

1295 
$$Y_{\text{lin}} = \mathcal{N}(f(\mathbb{E} X), \nabla f(\mathbb{E} X) \text{Cov } X \nabla f(\mathbb{E} X)^\top)$$

1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349

### B.3 THE UNSCENTED APPROXIMATION(S) $Y_{u95}$ AND $Y_{u02}$

An unscented transformation is a quadrature rule for approximating a probability measure on  $\mathbb{R}^n$  by  $2n + 1$  point masses whose locations (called sigma points) and weights satisfy certain first- and second-order moment matching conditions. This technique was developed to improve upon linearization in nonlinear Kalman filtering (Julier et al., 1995; Julier & Uhlmann, 1997; Julier et al., 2000; Julier, 2002; Wan & Van Der Merwe, 2000; Julier & Uhlmann, 2004).

There are (at least) two unscented transformations: a one-parameter family, which we refer to as Unscented'95; and a three-parameter family, which we refer to as Unscented'02.

- Unscented'95 (Julier et al., 1995; Julier & Uhlmann, 1997; Julier et al., 2000), used in Astudillo & Neto (2011); Abdelaziz et al. (2015), is a one-parameter family of unscented transformations, parameterized by a single shape hyperparameter  $\kappa$ .
- Unscented'02 (Julier, 2002; Wan & Van Der Merwe, 2000) is a three-parameter family of unscented transformations, parameterized by three shape hyperparameters  $(\alpha, \beta, \kappa)$ . This version, with default values tuned to be drastically more localized in  $X$ -space than Unscented'95, is usually called “the” unscented transformation in current filtering research (Jiang et al., 2025) and tooling (Ljung, 2025).

## C PRELIMINARIES: GAUSSIAN INTEGRALS

**Definition 5.** *The bivariate normal CDF  $\Phi_2$  is defined by*

$$\Phi_2(h, k; \rho) = \mathbb{P}[Z_1 \leq h, Z_2 \leq k],$$

where

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

**Remark 1.** *While we consider it a valid “closed form” atom, the bivariate normal CDF can be a difficult transcendental function to evaluate. Many software packages such as SciPy (Wagner et al., 2022) implement the multivariate normal CDF by a (quasi-) Monte Carlo integration over  $\mathbb{R}^n$ , which is too expensive for our purposes. Furthermore, we frequently use the expression  $\Phi(h, k; \rho) - \Phi(h, k; 0)$ , which is vulnerable to cancellation error for extreme values of  $h, k$ , and  $\rho$ . To avoid this, we use 10-point Gaussian quadrature of the one-dimensional proper integral*

$$\Phi_2(h, k; \rho) - \Phi_2(h, k; 0) = \int_0^\rho \partial'_\rho \Phi_2(h, k; \rho') d\rho',$$

using

$$\partial_\rho \Phi_2(h, k; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(h^2+k^2-2\rho hk)},$$

a helpful identity found in Drezner & Wesolowsky (1990).

**Definition 6.** *The bivariate Normal density is*

$$\phi_2(h, k; \rho) = \frac{\partial^2}{\partial h \partial k} \Phi_2(h, k; \rho).$$

**Definition 7.** *The partial derivative of the joint CDF is*

$$\Phi_{2;1}(h, k; \rho) = \frac{\partial}{\partial h} \Phi_2(h, k; \rho) \tag{4}$$

**Lemma 2.** *The function  $\Phi_{2;1}$  satisfies*

$$\Phi_{2;1}(h, k; \rho) = \phi(h) \Phi\left(\frac{k - \rho h}{\sqrt{1 - \rho^2}}\right).$$

1350 *Proof.* Letting  $Z_1, Z_2$  be standard normal with correlation  $\rho$ , there is a probabilistic interpretation  
 1351 of  $\Phi_{2;1}$ :

$$1352 \Phi_{2;1}(h, k; \rho) = \underbrace{f_{Z_1}(h)}_{\text{marginal density}} \underbrace{\mathbb{P}(Z_2 \leq k \mid Z_1 = h)}_{\text{conditional cdf}}$$

$$1353 = \phi(h) \Phi\left(\frac{k - \rho h}{\sqrt{1 - \rho^2}}\right).$$

1354  
1355  
1356  
1357  
1358 □

1359  
1360 **Lemma 3** (Multivariate Stein's lemma). *Let  $(X_1, \dots, X_n)$  be a multivariate normal random vector.*  
 1361 *Then*

$$1362 \mathbb{E}(X_1 - \mathbb{E} X_1) f(X_1, \dots, X_n) = \sum_{i=1}^n \text{Cov}(X_1, X_i) \mathbb{E} [\partial_{X_i} f(X_1, \dots, X_n)]$$

1363  
1364 **Lemma 4** (Stein simplification of  $L_\sigma$ ). *Let  $L_\sigma$  be defined as in Definition 3, and suppose that  $\sigma$  is*  
 1365 *differentiable. Then*

$$1366 L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} \mathbb{E} \sigma'(Z_1), \quad Z_1 \sim \mathcal{N}(\mu_1, \nu_{11}).$$

1367  
1368 *Proof.* Straightforward application of Lemma 3. □

1369  
1370 **Lemma 5** (Univariate integrals). *If  $Z \sim \mathcal{N}(\mu, \nu)$ , then*

$$1371 \mathbb{E} \Phi(Z) = \Phi\left(\frac{\mu}{\sqrt{1 + \nu}}\right) \tag{5a}$$

$$1372 \mathbb{E} \phi(Z) = \frac{1}{\sqrt{1 + \nu}} \phi\left(\frac{\mu}{\sqrt{1 + \nu}}\right) \tag{5b}$$

$$1373 \mathbb{E} Z \Phi(Z) = \frac{1}{(1 + \nu)^{3/2}} \phi\left(\frac{\mu}{\sqrt{1 + \nu}}\right) \tag{5c}$$

1374  
1375  
1376  
1377  
1378  
1379 *Proof of (5a).* Introducing an independent  $Z \sim \mathcal{N}(0, 1)$ , we have

$$1380 \mathbb{E} \Phi(Z) = \mathbb{E} \mathbb{P}[Z \leq X \mid X] \tag{6}$$

$$1381 = \mathbb{P}[Z \leq X] \tag{by the law of total probability}$$

$$1382 = \mathbb{P}[Z - X \leq 0] \tag{7}$$

1383  
1384 We conclude by noting that the random variable  $Z - X$  has a Normal distribution with mean  $-\mu$   
 1385 and variance  $1 + \nu$ . □

1386  
1387 *Proof of (5b).* We use  $\phi = \Phi'$ .

$$1388 \mathbb{E} \phi(Z) = \mathbb{E} \left. \frac{d}{dt} \right|_{t=0} \Phi(Z + t) \tag{8}$$

$$1389 = \left. \frac{d}{dt} \right|_{t=0} \mathbb{E} \Phi(Z + t) \tag{dominated convergence theorem}$$

$$1390 = \left. \frac{d}{dt} \right|_{t=0} \Phi\left(\frac{\mu + t}{\sqrt{1 + \nu}}\right) \tag{by (5a)}$$

1391  
1392 □

1393  
1394 *Proof of (5c).* We use the Gaussian ODE

$$1395 \phi'(x) + x\phi(x) = 0. \tag{9}$$

1404 Centering and applying Lemma 3,

$$\begin{aligned}
1406 \quad \mathbb{E} Z \phi(Z) &= \mathbb{E}(Z - \mu) \phi(Z) + \mu \mathbb{E} \phi(Z) && \text{(centering)} \\
1407 \quad &= \nu \mathbb{E} \phi'(Z) + \mu \mathbb{E} \phi(Z) && \text{(Stein's)} \\
1408 \quad &= -\nu \mathbb{E} Z \Phi(Z) + \mu \mathbb{E} \phi(Z) && \text{(Gaussian ODE)}
\end{aligned}$$

1410  
1411 Collecting like terms and solving for (I),

$$1412 \quad (1 + \nu) \mathbb{E} Z \Phi(Z) = \mu \mathbb{E} \phi(Z) \quad (10)$$

$$1413 \quad \mathbb{E} Z \Phi(Z) = \frac{\mu}{1 + \nu} \mathbb{E} \phi(Z) \quad (11)$$

$$1414 \quad = \frac{\mu_1}{(1 + \nu_{11})^{3/2}} \phi\left(\frac{\mu_1}{\sqrt{1 + \nu_{11}}}\right) \quad \text{(by (5b))}$$

1415  
1416  
1417  
1418  
1419 □

1420  
1421  
1422 **Lemma 6** (Bivariate integrals). *Let*

$$1423 \quad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{12} & \nu_{22} \end{pmatrix}\right).$$

1424  
1425  
1426  
1427  
1428 *Then*

$$1429 \quad \mathbb{E} \Phi(X_1) \Phi(X_2) = \Phi_2\left(\frac{\mu_1}{\sqrt{1 + \nu_{11}}}, \frac{\mu_2}{\sqrt{1 + \nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}}\right) \quad (12a)$$

$$1430 \quad \mathbb{E} \phi(X_1) \Phi(X_2) = \frac{1}{\sqrt{1 + \nu_{11}}} \Phi_{2;1}\left(\frac{\mu_1}{\sqrt{1 + \nu_{11}}}, \frac{\mu_2}{\sqrt{1 + \nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}}\right) \quad (12b)$$

$$1431 \quad \mathbb{E} \phi(X_1) \phi(X_2) = \frac{1}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}} \phi_2\left(\frac{\mu_1}{\sqrt{1 + \nu_{11}}}, \frac{\mu_2}{\sqrt{1 + \nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}}\right) \quad (12c)$$

$$1432 \quad \mathbb{E} \phi'(X_1) \Phi(X_2) = \frac{-\mu_1}{1 + \nu_{11}} \mathbb{E} \phi(X_1) \Phi(X_2) + \frac{-\nu_{12}}{1 + \nu_{11}} \mathbb{E} \phi(X_1) \phi(X_2) \quad (12d)$$

1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444 *Proof of (12a).* Introduce independent  $Z_1, Z_2 \sim \mathcal{N}(0, 1)$ . Then

$$1445 \quad \mathbb{E} \Phi(X_1) \Phi(X_2) = \mathbb{E} \mathbb{P}[Z_1 \leq X_1 \mid X_1] \mathbb{P}[Z_2 \leq X_2 \mid X_2] \quad (13)$$

$$1446 \quad = \mathbb{E} \mathbb{P}[Z_1 \leq X_1, Z_2 \leq X_2 \mid X_1, X_2] \quad \text{(independence)}$$

$$1447 \quad = \mathbb{P}[Z_1 \leq X_1, Z_2 \leq X_2] \quad \text{(by the law of total probability)}$$

$$1448 \quad = \mathbb{P}[Z_1 - X_1 \leq 0, Z_2 - X_2 \leq 0]. \quad (14)$$

1449  
1450  
1451  
1452 We conclude by using the fact that  $(Z_1 - X_1, Z_2 - X_2)$  is jointly Normal with distribution

$$1453 \quad \begin{pmatrix} Z_1 - X_1 \\ Z_2 - X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} -\mu_1 \\ -\mu_2 \end{pmatrix}, \begin{pmatrix} 1 + \nu_{11} & \nu_{12} \\ \nu_{12} & 1 + \nu_{22} \end{pmatrix}\right). \quad (15)$$

1454  
1455  
1456  
1457 □

1458 *Proof of (12b).* Using  $\phi = \Phi'$ , we have

$$1459 \mathbb{E} \phi(X_1)\Phi(X_2)$$

$$1460 = \mathbb{E} \frac{d}{dt} \Big|_{t=0} \Phi(X_1 + t)\Phi(X_2) \quad (16)$$

$$1461 = \frac{d}{dt} \Big|_{t=0} \mathbb{E} \Phi(X_1 + t)\Phi(X_2) \quad (\text{dominated convergence theorem})$$

$$1462 = \frac{d}{dt} \Big|_{t=0} \mathbb{P}(Z_1 - X_1 - t \leq 0, Z_2 - X_2 \leq 0) \quad (\text{introducing } Z_1, Z_2 \sim \mathcal{N}(0, 1))$$

$$1463 = \frac{d}{dt} \Big|_{t=0} \Phi_2 \left( \frac{\mu_1 + t}{\sqrt{1 + \nu_{11}}}, \frac{\mu_2}{\sqrt{1 + \nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}} \right) \quad (17)$$

$$1464 = \frac{1}{\sqrt{1 + \nu_{11}}} \Phi_{2;1} \left( \frac{\mu_1}{\sqrt{1 + \nu_{11}}}, \frac{\mu_2}{\sqrt{1 + \nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}} \right) \quad (18)$$

1474 □

1476 *Proof of (12c).* Using  $\phi = \Phi'$  in both terms,

$$1477 \mathbb{E} \phi(X_1)\phi(X_2) = \mathbb{E} \frac{\partial^2}{\partial t \partial s} \Big|_{t=0, s=0} \Phi(X_1 + t)\Phi(X_2 + s) \quad (19)$$

$$1480 = \frac{\partial^2}{\partial t \partial s} \Big|_{t=0, s=0} \mathbb{E} \Phi(X_1 + t)\Phi(X_2 + s) \quad (\text{dominated convergence theorem})$$

$$1481 = \frac{\partial^2}{\partial t \partial s} \Big|_{t=0, s=0} \mathbb{P}(Z_1 - X_1 - t \leq 0, Z_2 - X_2 - s \leq 0)$$

(introducing  $Z_1, Z_2 \sim \mathcal{N}(0, 1)$ )

$$1482 = \frac{\partial^2}{\partial t \partial s} \Big|_{t=0, s=0} \Phi_2 \left( \frac{\mu_1 + t}{\sqrt{1 + \nu_{11}}}, \frac{\mu_2 + s}{\sqrt{1 + \nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}} \right) \quad (20)$$

$$1483 = \frac{1}{\sqrt{1 + \nu_{11}} \sqrt{1 + \nu_{22}}} \phi_2 \left( \frac{\mu_1}{\sqrt{1 + \nu_{11}}}, \frac{\mu_2}{\sqrt{1 + \nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1 + \nu_{11})(1 + \nu_{22})}} \right)$$

(pdf is derivative of cdf)

1495 □

1497 *Proof of (12d).* We use the Gaussian ODE

$$1498 \phi'(x) + x\phi(x) = 0. \quad (21)$$

1500 Using this fact,

$$1501 \mathbb{E} \phi'(X_1)\Phi(X_2) = -\mathbb{E} X_1 \phi(X_1)\Phi(X_2) \quad (22)$$

$$1502 = -\mu_1 \mathbb{E} \phi(X_1)\Phi(X_2) - \mathbb{E}(X_1 - \mu_1)\phi(X_1)\Phi(X_2) \quad (\text{centering})$$

$$1503 = -\mu_1 \mathbb{E} \phi(X_1)\Phi(X_2) - \underbrace{\nu_{11} \mathbb{E} \phi'(X_1)\Phi(X_2)}_{\text{same as LHS}} - \nu_{12} \mathbb{E} \phi(X_1)\phi(X_2) \quad (\text{Lemma 3})$$

1504 Collecting like terms and solving for  $\mathbb{E} \phi'(X_1)\Phi(X_2)$ ,

$$1505 \mathbb{E} \phi'(X_1)\Phi(X_2) = \frac{-\mu_1}{1 + \nu_{11}} \mathbb{E} \phi(X_1)\Phi(X_2) + \frac{-\nu_{12}}{1 + \nu_{11}} \mathbb{E} \phi(X_1)\phi(X_2). \quad (23)$$

1511 □

---

1512 D DERIVATION OF UNCERTAINTY PROPAGATION FORMULAS FOR PROBIT  
1513 ACTIVATION  
1514

1515 In this appendix, we derive the  $M_\sigma$ ,  $K_\sigma$ , and  $L_\sigma$  functions (Def. 3) for the normal CDF activation  
1516 function

$$1517 \sigma(x) = \sqrt{\frac{2}{\pi}} \int_0^x e^{-\frac{1}{2}u^2} du = 2\Phi(x) - 1, \quad \Phi(x) = \mathbb{P}_{Z \sim \mathcal{N}(0,1)}(Z \leq x) \quad (24)$$

1520 **Lemma 7.** Let  $M_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (24).  
1521

$$1522 M_\sigma(\mu; \nu) = \sigma\left(\frac{\mu}{\sqrt{1+\nu}}\right)$$

1523  
1524  
1525 *Proof.* Let  $X \sim \mathcal{N}(\mu, \nu)$ .  
1526

$$1527 \mathbb{E} \sigma(X) = -1 + 2 \mathbb{E} \Phi(X) \quad (25)$$

1528 This is a direct application of Lemma 5, (5a).  $\square$   
1529

1530 **Lemma 8.** Let  $K_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (24).  
1531

$$1532 K(\mu_1, \mu_2; \nu_{11}, \nu_{22}, \nu_{12}) = 4 \Phi_2\left(\frac{\mu_1}{\sqrt{1+\nu_{11}}}, \frac{\mu_2}{\sqrt{1+\nu_{22}}}; \rho'\right) \Bigg|_{\rho'=0}^{\rho'=\frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}}},$$

1533  
1534 where  $\Phi_2$  is the bivariate normal CDF.  
1535

1536  
1537 *Proof.* The covariance may be expressed as

$$1538 \text{Cov}(\sigma(X_1), \sigma(X_2)) = 4 \text{Cov}(\Phi(X_1), \Phi(X_2)) \quad (26)$$

$$1539 = 4 \mathbb{E} \Phi(X_1)\Phi(X_2) - 4 \mathbb{E} \Phi(X_1) \mathbb{E} \Phi(X_2) \quad (27)$$

1540 Now apply Lemma 6 to the first term.  $\square$   
1541

1542 **Lemma 9.** Let  $L_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (24). Then  
1543

$$1544 L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = 2 \frac{\nu_{12}}{\sqrt{1+\nu_{11}}} \phi\left(\frac{\mu_1}{\sqrt{1+\nu_{11}}}\right).$$

1545  
1546 *Proof.* Using Lemma 4, we have  
1547

$$1548 L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} \mathbb{E} \sigma'(X_1) \quad (28)$$

$$1549 = 2\nu_{12} \mathbb{E} \phi(X_1) \quad (29)$$

1550 where  $\phi = \Phi'$ . By Lemma 5, applied to  $X_1 \sim \mathcal{N}(\mu_1, \nu_{11})$ ,  
1551

$$1552 \mathbb{E} \phi(X_1) = \frac{1}{\sqrt{1+\nu_{11}}} \phi\left(\frac{\mu_1}{\sqrt{1+\nu_{11}}}\right). \quad (30)$$

1553  
1554  
1555  $\square$   
1556

1557 E DERIVATION OF UNCERTAINTY PROPAGATION FORMULAS FOR GELU  
1558 ACTIVATION  
1559

1560 In this appendix, we derive the  $M_\sigma$ ,  $K_\sigma$  and  $L_\sigma$  functions (Def. 3) for the GeLU activation function  
1561

$$1562 \sigma(x) = x\Phi(x). \quad (31)$$

1563 **Lemma 10.** Let  $M_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (31). Then  
1564

$$1565 M_\sigma(\mu; \nu) = \frac{\nu}{\sqrt{1+\nu}} \phi\left(\frac{\mu}{\sqrt{1+\nu}}\right) + \mu \Phi\left(\frac{\mu}{\sqrt{1+\nu}}\right).$$



1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673

We apply Lemma 3 to (I.a) and (I.b) to get

$$(I.a) + (I.b) = \nu_{11} \mathbb{E} \phi(X_1)(X_2 - \mu_2)\Phi(X_2) + \nu_{12} \mathbb{E} \Phi(X_1)(X_2 - \mu_2)\phi(X_2) \quad (36)$$

$$= \nu_{11} \mathbb{E} [\nu_{12}\phi'(X_1)\Phi(X_2) + \nu_{22}\phi(X_1)\phi(X_2)] \quad (37)$$

$$+ \nu_{12} \mathbb{E} [\nu_{12}\phi(X_1)\phi(X_2) + \nu_{22}\Phi(X_1)\phi'(X_2)]$$

$$= \nu_{12}\nu_{11} \underbrace{\mathbb{E} \phi'(X_1)\Phi(X_2)}_{(I.c)} + \nu_{12}\nu_{22} \underbrace{\mathbb{E} \Phi(X_1)\phi'(X_2)}_{(I.d)} \quad (38)$$

$$+ (\nu_{11}\nu_{22} + \nu_{12}^2) \mathbb{E} \phi(X_1)\phi(X_2)$$

By Lemma 6, (12d), we have

$$(I.c) = \frac{-\mu_1}{1 + \nu_{11}} \mathbb{E} \phi(X_1)\Phi(X_2) + \frac{-\nu_{12}}{1 + \nu_{11}} \mathbb{E} \phi(X_1)\phi(X_2)$$

$$(I.d) = \frac{-\mu_2}{1 + \nu_{22}} \mathbb{E} \Phi(X_1)\phi(X_2) + \frac{-\nu_{12}}{1 + \nu_{22}} \mathbb{E} \phi(X_1)\phi(X_2)$$

Combining these last two equations with (35) and (38), (I) becomes

$$(I) = \nu_{12}\nu_{11} \left( \frac{-\mu_1}{1 + \nu_{11}} \mathbb{E} \phi(X_1)\Phi(X_2) + \frac{-\nu_{12}}{1 + \nu_{11}} \mathbb{E} \phi(X_1)\phi(X_2) \right)$$

$$+ \nu_{12}\nu_{22} \left( \frac{-\mu_2}{1 + \nu_{22}} \mathbb{E} \Phi(X_1)\phi(X_2) + \frac{-\nu_{12}}{1 + \nu_{22}} \mathbb{E} \phi(X_1)\phi(X_2) \right)$$

$$+ (\nu_{11}\nu_{22} + \nu_{12}^2) \mathbb{E} \phi(X_1)\phi(X_2)$$

$$+ \nu_{12} \mathbb{E} \Phi(X_1)\Phi(X_2) \quad (39)$$

which simplifies to

$$(I) = \frac{-\mu_1\nu_{12}\nu_{11}}{1 + \nu_{11}} \mathbb{E} \phi(X_1)\Phi(X_2) + \frac{-\mu_2\nu_{12}\nu_{22}}{1 + \nu_{22}} \mathbb{E} \Phi(X_1)\phi(X_2)$$

$$+ \left[ \nu_{11}\nu_{22} + \nu_{12}^2 \left( 1 - \frac{\nu_{11}}{1 + \nu_{11}} - \frac{\nu_{22}}{1 + \nu_{22}} \right) \right] \mathbb{E} \phi(X_1)\phi(X_2)$$

$$+ \nu_{12} \mathbb{E} \Phi(X_1)\Phi(X_2) \quad (40)$$

For terms (II-III) of (33), we apply Lemma 3 to get

$$(II) = \mathbb{E} \mu_1\Phi(X_1)(X_2 - \mu_2)\Phi(X_2) = \mu_1 (\nu_{12} \mathbb{E} \phi(X_1)\Phi(X_2) + \nu_{22} \mathbb{E} \Phi(X_1)\phi(X_2))$$

$$(III) = \mathbb{E}(X_1 - \mu_1)\Phi(X_1)\mu_2\Phi(X_2) = \mu_2 (\nu_{11} \mathbb{E} \phi(X_1)\Phi(X_2) + \nu_{12} \mathbb{E} \Phi(X_1)\phi(X_2))$$

Now (33) becomes:

$$\mathbb{E} X_1\Phi(X_1)X_2\Phi(X_2)$$

$$= \left( \mu_1\nu_{12} + \mu_2\nu_{11} - \frac{\mu_1\nu_{12}\nu_{11}}{1 + \nu_{11}} \right) \underbrace{\mathbb{E} \phi(X_1)\Phi(X_2)}_{(V.a)}$$

$$+ \left( \mu_2\nu_{12} + \mu_1\nu_{22} - \frac{\mu_2\nu_{12}\nu_{22}}{1 + \nu_{22}} \right) \underbrace{\mathbb{E} \Phi(X_1)\phi(X_2)}_{(V.b)}$$

$$+ \left[ \nu_{11}\nu_{22} + \nu_{12}^2 \left( 1 - \frac{\nu_{11}}{1 + \nu_{11}} - \frac{\nu_{22}}{1 + \nu_{22}} \right) \right] \underbrace{\mathbb{E} \phi(X_1)\phi(X_2)}_{(V.c)}$$

$$+ (\mu_1\mu_2 + \nu_{12}) \underbrace{\mathbb{E} \Phi(X_1)\Phi(X_2)}_{(V.d)} \quad (41)$$

1674 By Lemma 6, (12b),  
1675

$$1676 \quad (\text{V.a}) = \frac{1}{\sqrt{1+\nu_{11}}} \Phi_{2;1} \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}}, \frac{\mu_2}{\sqrt{1+\nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right) \quad (42)$$

1678 and  
1679

$$1680 \quad (\text{V.b}) = \frac{1}{\sqrt{1+\nu_{22}}} \Phi_{2;1} \left( \frac{\mu_2}{\sqrt{1+\nu_{22}}}, \frac{\mu_1}{\sqrt{1+\nu_{11}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right). \quad (43)$$

1682 By Lemma 6, (12c),  
1683

$$1684 \quad (\text{V.c}) = \frac{1}{\sqrt{1+\nu_{11}}\sqrt{1+\nu_{22}}} \phi_2 \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}}, \frac{\mu_2}{\sqrt{1+\nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right). \quad (44)$$

1686 By Lemma 6, (12a),  
1687

$$1688 \quad (\text{V.d}) = \Phi_2 \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}}, \frac{\mu_2}{\sqrt{1+\nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right) \quad (45)$$

1691 The final form of (41) is  
1692

$$1693 \quad \mathbb{E} X_1 \Phi(X_1) X_2 \Phi(X_2)$$

$$1694 = \left( \mu_1 \nu_{12} + \mu_2 \nu_{11} - \frac{\mu_1 \nu_{12} \nu_{11}}{1+\nu_{11}} \right) \frac{1}{\sqrt{1+\nu_{11}}} \Phi_{2;1} \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}}, \frac{\mu_2}{\sqrt{1+\nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right)$$

$$1695 + \left( \mu_2 \nu_{12} + \mu_1 \nu_{22} - \frac{\mu_2 \nu_{12} \nu_{22}}{1+\nu_{22}} \right) \frac{1}{\sqrt{1+\nu_{22}}} \Phi_{2;1} \left( \frac{\mu_2}{\sqrt{1+\nu_{22}}}, \frac{\mu_1}{\sqrt{1+\nu_{11}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right)$$

$$1696 + \frac{\nu_{11} \nu_{22} + \nu_{12}^2 \left( 1 - \frac{\nu_{11}}{1+\nu_{11}} - \frac{\nu_{22}}{1+\nu_{22}} \right)}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \phi_2 \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}}, \frac{\mu_2}{\sqrt{1+\nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right)$$

$$1697 + (\mu_1 \mu_2 + \nu_{12}) \Phi_2 \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}}, \frac{\mu_2}{\sqrt{1+\nu_{22}}}; \frac{\nu_{12}}{\sqrt{(1+\nu_{11})(1+\nu_{22})}} \right) \quad (46)$$

1700 The conclusion follows from subtracting  $M(\mu_1; \nu_{11})M(\mu_2; \nu_{22})$ .  $\square$   
1701

1702 **Lemma 12.** Let  $L_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (31). Then  
1703

$$1704 \quad L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} \frac{\mu_1}{(1+\nu_{11})^{3/2}} \phi \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}} \right) + \nu_{12} \Phi \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}} \right). \quad (47)$$

1707 *Proof.* By Lemma 4,  
1708

$$1709 \quad L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} \mathbb{E} \sigma'(X_1), \quad X_1 \sim \mathcal{N}(\mu_1, \nu_{11})$$

$$1710 \quad = \nu_{12} \underbrace{\mathbb{E} X_1 \phi(X_1)}_{\text{(I)}} + \nu_{12} \underbrace{\mathbb{E} \Phi(X_1)}_{\text{(II)}} \quad (\text{product rule})$$

1711 By Lemma 5, (5c)  
1712

$$1713 \quad \text{(I)} = \frac{\mu_1}{(1+\nu_{11})^{3/2}} \phi \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}} \right). \quad (48)$$

1714 Term (II) is given by  
1715

$$1716 \quad \text{(II)} = \mathbb{E} \Phi(X_1) = \Phi \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}} \right) \quad (49)$$

1717 Applying terms (I)–(II),  
1718

$$1719 \quad L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} \frac{\mu_1}{(1+\nu_{11})^{3/2}} \phi \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}} \right) + \nu_{12} \Phi \left( \frac{\mu_1}{\sqrt{1+\nu_{11}}} \right) \quad (50)$$

1720  $\square$

---

1728 F DERIVATION OF UNCERTAINTY PROPAGATION FORMULAS FOR ReLU  
1729 ACTIVATION  
1730

1731 In this appendix, we derive the uncertainty propagation formulas for ReLU activation  
1732

$$1733 \sigma(x) = \max(0, x) \quad (51)$$

1734 using the identity  
1735

$$1736 \text{ReLU}(x) = \lim_{\alpha \rightarrow \infty} \alpha^{-1} \text{GeLU}(\alpha x). \quad (52)$$

1737 See Muthén (1990) for another way to derive  $M$  and  $K$ .  
1738

1739 **Lemma 13.** *Let  $M_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (51). Then*  
1740

$$1741 M_\sigma(\mu; \nu) = \sqrt{\nu} \phi\left(\frac{\mu}{\sqrt{\nu}}\right) + \mu \Phi\left(\frac{\mu}{\sqrt{\nu}}\right).$$

1742 *Proof.* By the dominated convergence theorem,  
1743

$$1744 M_\sigma(\mu; \nu) = \lim_{\alpha \rightarrow \infty} \alpha^{-1} \mathbb{E} M_{\text{GeLU}}(\alpha\mu; \alpha^2\nu) \quad (53)$$

$$1745 = \lim_{\alpha \rightarrow \infty} \alpha^{-1} \left[ \frac{\alpha^2\nu}{\sqrt{1 + \alpha^2\nu}} \phi\left(\frac{\alpha\mu}{\sqrt{1 + \alpha^2\nu}}\right) + \mu \Phi\left(\frac{\alpha\mu}{\sqrt{1 + \alpha^2\nu}}\right) \right] \quad (\text{by Lemma 10})$$

$$1746 = \sqrt{\nu} \phi\left(\frac{\mu}{\sqrt{\nu}}\right) + \mu \Phi\left(\frac{\mu}{\sqrt{\nu}}\right) \quad (54)$$

1747  $\square$

1748 **Lemma 14.** *Let  $K_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (51). Then*  
1749

$$1750 K_\sigma(\mu_1, \mu_2; \nu_{11}, \nu_{22}, \nu_{12}) = \mu_2 \sqrt{\nu_{11}} \Phi_{2;1} \left( \frac{\mu_1}{\sqrt{\nu_{11}}}, \frac{\mu_2}{\sqrt{\nu_{22}}}; \frac{\nu_{12}}{\sqrt{\nu_{11}\nu_{22}}} \right)$$

$$1751 + \mu_1 \sqrt{\nu_{22}} \Phi_{2;1} \left( \frac{\mu_2}{\sqrt{\nu_{22}}}, \frac{\mu_1}{\sqrt{\nu_{11}}}; \frac{\nu_{12}}{\sqrt{\nu_{11}\nu_{22}}} \right)$$

$$1752 + \left( \sqrt{\nu_{11}\nu_{22}} - \frac{\nu_{12}^2}{\sqrt{\nu_{11}\nu_{22}}} \right) \phi_2 \left( \frac{\mu_1}{\sqrt{\nu_{11}}}, \frac{\mu_2}{\sqrt{\nu_{22}}}; \frac{\nu_{12}}{\sqrt{\nu_{11}\nu_{22}}} \right)$$

$$1753 + (\mu_1\mu_2 + \nu_{12}) \Phi_2 \left( \frac{\mu_1}{\sqrt{\nu_{11}}}, \frac{\mu_2}{\sqrt{\nu_{22}}}; \frac{\nu_{12}}{\sqrt{\nu_{11}\nu_{22}}} \right)$$

$$1754 - M_\sigma(\mu_1; \nu_{11}) M_\sigma(\mu_2; \nu_{22})$$

1755 *Proof.* By the dominated convergence theorem,  
1756

$$1757 \mathbb{E} \sigma(X_1) \sigma(X_2) = \mathbb{E} \lim_{\alpha \rightarrow \infty} \alpha^{-2} \text{GeLU}(\alpha X_1) \text{GeLU}(\alpha X_2) \quad (55)$$

$$1758 = \lim_{\alpha \rightarrow \infty} \alpha^{-2} \mathbb{E} \text{GeLU}(\alpha X_1) \text{GeLU}(\alpha X_2) \quad (56)$$

To compute  $\alpha^{-2} \mathbb{E} \text{GeLU}(\alpha X_1) \text{GeLU}(\alpha X_2)$ , we use (46) (Proof of Lemma 11) while scaling  $\mu$  by  $\alpha$  and  $\nu$  by  $\alpha^2$ .

$$\begin{aligned}
& \alpha^{-2} \mathbb{E} \text{GeLU}(\alpha X_1) \text{GeLU}(\alpha X_2) \\
&= \frac{\alpha \mu_1 \nu_{12} + \alpha \mu_2 \nu_{11} - \alpha^3 \frac{\mu_1 \nu_{12} \nu_{11}}{1 + \alpha^2 \nu_{11}}}{\sqrt{1 + \alpha^2 \nu_{11}}} \\
& \quad \Phi_{2;1} \left( \frac{\alpha \mu_1}{\sqrt{1 + \alpha^2 \nu_{11}}}, \frac{\alpha \mu_2}{\sqrt{1 + \alpha^2 \nu_{22}}}; \frac{\alpha^2 \nu_{12}}{\sqrt{(1 + \alpha^2 \nu_{11})(1 + \alpha^2 \nu_{22})}} \right) \\
& + \frac{\alpha \mu_2 \nu_{12} + \alpha \mu_1 \nu_{22} - \alpha^3 \frac{\mu_2 \nu_{12} \nu_{22}}{1 + \alpha^2 \nu_{22}}}{\sqrt{1 + \alpha^2 \nu_{22}}} \\
& \quad \Phi_{2;1} \left( \frac{\alpha \mu_2}{\sqrt{1 + \alpha^2 \nu_{22}}}, \frac{\alpha \mu_1}{\sqrt{1 + \alpha^2 \nu_{11}}}; \frac{\alpha^2 \nu_{12}}{\sqrt{(1 + \alpha^2 \nu_{11})(1 + \alpha^2 \nu_{22})}} \right) \\
& + \frac{\alpha^2 \nu_{11} \nu_{22} + \alpha^2 \nu_{12}^2 \left( 1 - \frac{\alpha^2 \nu_{11}}{1 + \alpha^2 \nu_{11}} - \frac{\alpha^2 \nu_{22}}{1 + \alpha^2 \nu_{22}} \right)}{\sqrt{(1 + \alpha^2 \nu_{11})(1 + \alpha^2 \nu_{22})}} \\
& \quad \phi_2 \left( \frac{\alpha \mu_1}{\sqrt{1 + \alpha^2 \nu_{11}}}, \frac{\alpha \mu_2}{\sqrt{1 + \alpha^2 \nu_{22}}}; \frac{\alpha^2 \nu_{12}}{\sqrt{(1 + \alpha^2 \nu_{11})(1 + \alpha^2 \nu_{22})}} \right) \\
& + (\mu_1 \mu_2 + \nu_{12}) \Phi_2 \left( \frac{\alpha \mu_1}{\sqrt{1 + \alpha^2 \nu_{11}}}, \frac{\alpha \mu_2}{\sqrt{1 + \alpha^2 \nu_{22}}}; \frac{\alpha^2 \nu_{12}}{\sqrt{(1 + \alpha^2 \nu_{11})(1 + \alpha^2 \nu_{22})}} \right)
\end{aligned} \tag{57}$$

Taking  $\alpha \rightarrow \infty$ ,

$$\begin{aligned}
\mathbb{E} \sigma(X_1) \sigma(X_2) &= \lim_{\alpha \rightarrow \infty} \alpha^{-2} \mathbb{E} \text{GeLU}(\alpha X_1) \text{GeLU}(\alpha X_2) \\
&= \mu_2 \sqrt{\nu_{11}} \Phi_{2;1} \left( \frac{\mu_1}{\sqrt{\nu_{11}}}, \frac{\mu_2}{\sqrt{\nu_{22}}}; \frac{\nu_{12}}{\sqrt{\nu_{11} \nu_{22}}} \right) \\
& + \mu_1 \sqrt{\nu_{22}} \Phi_{2;1} \left( \frac{\mu_2}{\sqrt{\nu_{22}}}, \frac{\mu_1}{\sqrt{\nu_{11}}}; \frac{\nu_{12}}{\sqrt{\nu_{11} \nu_{22}}} \right) \\
& + \left( \sqrt{\nu_{11} \nu_{22}} - \frac{\nu_{12}^2}{\sqrt{\nu_{11} \nu_{22}}} \right) \phi_2 \left( \frac{\mu_1}{\sqrt{\nu_{11}}}, \frac{\mu_2}{\sqrt{\nu_{22}}}; \frac{\nu_{12}}{\sqrt{\nu_{11} \nu_{22}}} \right) \\
& + (\mu_1 \mu_2 + \nu_{12}) \Phi_2 \left( \frac{\mu_1}{\sqrt{\nu_{11}}}, \frac{\mu_2}{\sqrt{\nu_{22}}}; \frac{\nu_{12}}{\sqrt{\nu_{11} \nu_{22}}} \right)
\end{aligned} \tag{58}$$

The conclusion follows after subtracting  $\mathbb{E} \sigma(X_1) \mathbb{E} \sigma(X_2)$ .  $\square$

**Lemma 15.** Let  $L_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (51). Then

$$L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} \Phi \left( \frac{\mu_1}{\sqrt{\nu_{11}}} \right).$$

*Proof.* By the dominated convergence theorem,

$$L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \lim_{\alpha \rightarrow \infty} \alpha^{-2} L_{\text{GeLU}}(\alpha \mu_1; \alpha^2 \nu_{11}, \alpha^2 \nu_{22}, \alpha^2 \nu_{12}) \tag{59}$$

$$= \lim_{\alpha \rightarrow \infty} \alpha^{-2} \left[ \alpha^2 \nu_{12} \frac{\alpha \mu_1}{(1 + \alpha^2 \nu_{11})^{3/2}} \phi \left( \frac{\alpha \mu_1}{\sqrt{1 + \alpha^2 \nu_{11}}} \right) + \alpha^2 \nu_{12} \Phi \left( \frac{\alpha \mu_1}{\sqrt{1 + \alpha^2 \nu_{11}}} \right) \right] \tag{by Lemma 12}$$

$$= \nu_{12} \Phi \left( \frac{\mu_1}{\sqrt{\nu_{11}}} \right) \tag{60}$$

1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889

□

## G DERIVATION OF UNCERTAINTY PROPAGATION FORMULAS FOR HEAVISIDE ACTIVATION

In this appendix, we state the  $M_\sigma$ ,  $K_\sigma$  and  $L_\sigma$  functions (Def. 3) for the Heaviside activation function

$$\sigma(x) = \mathbf{1}_{x>0} \tag{61}$$

$$= \lim_{\alpha \rightarrow \infty} (1/2 + 1/2\hat{\sigma}(\alpha x)) \tag{62}$$

where  $\hat{\sigma}$  is the odd probit sigmoid defined as in (51). The idea of the proofs is to take dominated limits of the results in Appendix. D, similar to Appendix F's treatment of Appendix E, so we omit them.

**Lemma 16.** *Let  $\sigma$  be the Heaviside function as in (61). Then the functions defined in Def. 3 are*

$$M_\sigma(\mu; \nu) = \Phi\left(\frac{\mu}{\sqrt{\nu}}\right),$$

$$K_\sigma(\mu_1, \mu_2; \nu_{11}, \nu_{22}, \nu_{12}) = \Phi_2\left(\frac{\mu_1}{\sqrt{\nu_{11}}}, \frac{\mu_2}{\sqrt{\nu_{22}}}; \rho'\right) \Bigg|_{\rho'=0}^{\rho' = \frac{\nu_{12}}{\sqrt{\nu_{11}\nu_{22}}}}, \quad \text{and}$$

$$L_\sigma(\mu_1; \nu_{11}, \nu_{22}, \nu_{12}) = \frac{\nu_{12}}{\sqrt{\nu_{11}}} \phi\left(\frac{\mu_1}{\sqrt{\nu_{11}}}\right).$$

---

1890 H DERIVATION OF UNCERTAINTY PROPAGATION FORMULAS FOR SINE  
1891 ACTIVATION  
1892

1893 In this appendix, we derive the  $M_\sigma$ ,  $K_\sigma$  and  $L_\sigma$  functions (Def. 3) for the sinusoidal activation  
1894 function

$$1895 \sigma(x) = \sin(x). \quad (63)$$

1896 We begin by recalling the identities that if  $Z \sim \mathcal{N}(\mu, \nu)$ ,

$$1897 \mathbb{E} \sin(Z) = e^{-\nu/2} \sin(\mu) \quad (64)$$

$$1898 \mathbb{E} \cos(Z) = e^{-\nu/2} \cos(\mu) \quad (65)$$

1901 These formulas, which can be derived from the characteristic function of the normal distribution,  
1902 allow moments of  $\sin(Z)$  and  $\cos(Z)$  to be computed exactly. There is no need, as in Sitzmann  
1903 et al. (2020, Lemma 1.6) to use analytic approximations in this step.

1904 From (64) immediately follows:

1905 **Lemma 17.** *Let  $M_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (63). Then*

$$1906 M_\sigma(\mu; \nu) = e^{-\nu/2} \sin(\mu).$$

1907 **Lemma 18.** *Let  $K_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (63). Then*

$$1908 K_\sigma(\mu_1, \mu_2; \nu_{11}, \nu_{22}, \nu_{12}) = \frac{1}{2} \left[ e^{\nu^* + \nu_{12}} - e^{\nu^*} \right] \cos(\mu_1 - \mu_2) \\ 1909 - \frac{1}{2} \left[ e^{\nu^* - \nu_{12}} - e^{\nu^*} \right] \cos(\mu_1 + \mu_2),$$

1910 where

$$1911 \nu^* = -\frac{\nu_{11} + \nu_{22}}{2}.$$

1912 *Proof.* Let

$$1913 \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{12} & \nu_{22} \end{pmatrix} \right). \quad (66)$$

1914 Then by inserting (63), we have

$$1915 \text{Cov}(\sigma(X_1), \sigma(X_2)) = \mathbb{E} \sin(X_1) \sin(X_2) - \mathbb{E} \sin(X_1) \mathbb{E} \sin(X_2). \quad (67)$$

1916 Using some trigonometric identities and (65), the first term of (67) becomes

$$1917 \mathbb{E} \sin(X_1) \sin(X_2) = \frac{1}{2} \mathbb{E} \cos \left( \underbrace{X_1 - X_2}_{\mathcal{N}(\mu_1 - \mu_2, \nu_{11} + \nu_{22} - 2\nu_{12})} \right) - \frac{1}{2} \mathbb{E} \cos \left( \underbrace{X_1 + X_2}_{\mathcal{N}(\mu_1 + \mu_2, \nu_{11} + \nu_{22} + 2\nu_{12})} \right) \quad (68) \\ 1918 = \frac{1}{2} \exp \left( -\frac{\nu_{11} + \nu_{22}}{2} + \nu_{12} \right) \cos(\mu_1 - \mu_2) \\ 1919 - \frac{1}{2} \exp \left( -\frac{\nu_{11} + \nu_{22}}{2} - \nu_{12} \right) \cos(\mu_1 + \mu_2) \quad (69)$$

1920 The second term of (67) becomes

$$1921 \mathbb{E} \sin(X_1) \mathbb{E} \sin(X_2) = \frac{1}{2} \exp \left( -\frac{\nu_{11} + \nu_{22}}{2} \right) \cos(\mu_1 - \mu_2) \\ 1922 - \frac{1}{2} \exp \left( -\frac{\nu_{11} + \nu_{22}}{2} \right) \cos(\mu_1 + \mu_2) \quad (70)$$

1923 The result follows from collecting like terms.  $\square$

1924 **Lemma 19.** *Let  $L_\sigma$  be defined as in Def. 3, with  $\sigma$  as in (63). Then*

$$1925 L_\sigma(\mu_1, \mu_2; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} e^{-\nu_{11}/2} \cos(\mu_1).$$

---

1944 *Proof.* Let

1945

1946

1947

1948

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{12} & \nu_{22} \end{pmatrix} \right). \quad (71)$$

1949

Using Lemma 4, we have

1950

$$L_\sigma(\mu_1, \mu_2; \nu_{11}, \nu_{22}, \nu_{12}) = \nu_{12} \mathbb{E} \sigma'(X_1) \quad (72)$$

1951

$$= \nu_{12} \mathbb{E} \cos(X_1) \quad (73)$$

1952

1953

$$= \nu_{12} e^{-\nu_{11}/2} \cos(\mu_1) \quad (74)$$

1954

by (65).

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

□

---

## I THEORETICAL GUARANTEES

The objective of this section is to provide a theoretical analysis of the dissimilarity between  $Y_0$  and  $Y$ . We recall the layer-by-layer Normal approximation resulting in  $Y$  alongside the exact neural network formula resulting in  $Y_0$ .

$$Y = Y^\ell \qquad Y_0 = Y_0^\ell \qquad (75a)$$

$$Y^k = N g^k(Y^{k-1}), \qquad Y_0^k = g^k(Y_0^{k-1}), \qquad k \in \{1 \dots \ell\}, \qquad (75b)$$

$$Y^0 = X \qquad Y_0^0 = X, \qquad (75c)$$

where  $g^k$  is the function

$$g^k(x) = g(x; A^k, b^k, C^k, d^k). \qquad (75d)$$

We will use the Wasserstein distance to measure the distance between  $Y$  and  $Y_0$ .

The ultimate goal is to show that  $d_W(Y_0, Y)$  is small. We will build up this quantity by recursion through the layers of the neural network. The key idea is that the error induced by Normal approximation gets worse with every subsequent layer of the network.

### I.1 RECURSIVE TRIANGLE INEQUALITY

Our basic triangle inequality is

$$\Delta^k := d_W(Y_0^k, Y^k) \qquad (76)$$

$$= d_W(g^k(Y_0^{k-1}), Y^k) \qquad (\text{definition of hidden layer (75b)})$$

$$\leq d_W(g^k(Y_0^{k-1}), g^k(Y^{k-1})) + d_W(g^k(Y^{k-1}), Y^k) \qquad (\text{triangle inequality})$$

$$\leq \left\| \nabla g^k \right\|_\infty d_W(Y_0^{k-1}, Y^{k-1}) + d_W(g^k(Y^{k-1}), Y^k) \qquad (\text{Lipschitz property of } d_W)$$

$$\leq \left\| \nabla g^k \right\|_\infty \Delta^{k-1} + d_W(g^k(Y^{k-1}), Y^k) \qquad (77)$$

If  $X$  is Gaussian, then  $\Delta^0 = 0$ . We need to fill two blanks to make this formula concrete.

### I.2 LIPSCHITZ CONSTANT

First, the Lipschitz constant  $\left\| \nabla g^k \right\|_\infty$ :

$$\nabla g(x; A, b, C, d) = \nabla [\sigma(Ax + b) + Cx + d] \qquad (78)$$

$$= A^\top \text{diag} \{ \sigma'(Ax + b) \} + C^\top \qquad (79)$$

To bound this quantity,

$$\left\| \nabla g(x; A, b, C, d) \right\| \leq \sup_x \left\| A^\top \text{diag} \{ \sigma'(Ax + b) \} + C^\top \right\| \qquad (80)$$

$$\leq \left\| \sigma' \right\|_\infty \|A\| + \|C\| \qquad (81)$$

### I.3 NON-NORMALITY

Second, we need to bound the distance between  $g^k(Y^{k-1})$  and  $Y^k$ . The importance of the recursion (77) is that at layer  $k$ , we assess the goodness of fit between  $g^k(Y^{k-1})$ , a nonlinear transformation of a Normal random vector, and  $Y^k$ , its approximant. As the input to  $\sigma$  is Normal, the non-normality of  $g^k(Y^{k-1})$  depends on the non-linearity of  $g^k$ , as measured by its second derivatives:

$$d_W \left( g^k(Y^{k-1}), Y^k \right) \leq \frac{3}{\sqrt{2}} \left\| \Sigma_k^{-1} \right\| \left\| \Sigma_k \right\|^{1/2} \left\| A^k \Sigma_{k-1} (A^k)^\top \right\|^{3/2} \cdot \left( \sum_{i=1}^{d^k} \mathbb{E} \left\| \sigma''(Y_i^{k-1}) \right\|^4 \right)^{1/4} \left( \sum_{i=1}^{d^k} \mathbb{E} \left\| \sigma'(Y_i^{k-1}) \right\|^4 \right)^{1/4} \qquad (82)$$

where  $d^k$  is the number of neurons in layer  $k$  and  $\Sigma^k = \text{Cov } Y^k$ . As a cruder approximation,

$$d_W \left( g^k(Y^{k-1}), Y^k \right) \leq \frac{3}{\sqrt{2}} \left( d^k \right)^{1/2} \|\sigma''\|_\infty \|\sigma'\|_\infty \left\| \Sigma_k^{-1} \right\| \|\Sigma_k\|^{1/2} \left\| A^k \Sigma_{k-1} \left( A^k \right)^\top \right\|^{3/2} \quad (83)$$

These inequalities are applications of Lem. 20, which follows from a result from the functional analysis of Gaussian spaces:

**Theorem 1** (Nourdin et al. (2009, Theorem 7.1)). *Fix  $d \geq 2$ , and let  $C = \{C(i, j); i, j = 1, \dots, d\}$  be a  $d \times d$  positive definite matrix. Suppose that  $F = (F_1, \dots, F_d)$  is a  $\mathbb{R}^d$ -valued random vector such that  $\mathbb{E} F_i = 0$  and  $F_i \in \mathbb{D}^{2,4}$  for every  $i = 1, \dots, d$ . Assume moreover that  $F$  has covariance matrix  $C$ . Then*

$$d_W(F, \mathcal{N}_d(0, C)) \leq \frac{3}{\sqrt{2}} \|C^{-1}\| \|C\|^{1/2} \left( \sum_{i=1}^d \mathbb{E} \|D^2 F_i\|_{\text{op}}^4 \right)^{1/4} \left( \sum_{j=1}^d \mathbb{E} \|DF_j\|_{\mathfrak{h}}^4 \right)^{1/4},$$

where  $\mathcal{N}_d(0, C)$  indicates a  $d$ -dimensional centered Gaussian vector, with covariance matrix equal to  $C$ , and  $D$  is the Malliavin derivative with respect to the underlying isonormal process.

This theorem is stated in terms of the deep generality of the Malliavin calculus for nonlinear functionals  $F$  of infinite-dimensional Gaussian processes. We adapt it to an inequality for finite-dimensional Gaussian  $X$ :

**Lemma 20** (Second-order Poincaré inequality). *Let  $X \sim \mathcal{N}(\mu_X, \Sigma_X)$  be a  $\mathbb{R}^n$ -valued random vector and  $Y = f(X)$  be a  $\mathbb{R}^d$ -valued random vector with mean  $\mu$  and covariance matrix  $\Sigma_Y$ . Then*

$$d_W(Y, \mathcal{N}(\mu, \Sigma_Y)) \leq \frac{3}{\sqrt{2}} \|\Sigma_Y^{-1}\| \|\Sigma_Y\|^{1/2} \|\Sigma_X\|^{3/2} \cdot \left[ \sum_{i=1}^d \left( \mathbb{E} \|\nabla^2 f_i(X)\|^4 \right)^{1/4} \right] \left[ \sum_{i=1}^d \left( \mathbb{E} \|\nabla f_i(X)\|^4 \right)^{1/4} \right],$$

where  $\nabla^2 f_i(X)$  is the Hessian of  $f_i$  evaluated at  $X$ , and  $\nabla f_i(X)$  is the gradient of  $f_i$  evaluated at  $X$ .

While this lemma is technically a basic application of Thm. 1, some nontrivial prerequisites are required to get it right.<sup>4</sup> Before proving this lemma, we first provide an exposition of isonormal Gaussian processes and the spaces  $\mathbb{D}^{m,p}$ . This material is derived from Nualart (2006); Nourdin et al. (2010).

Let  $\mathcal{I}$  be an abstract index set. A Gaussian process  $\{X(i)\}_{i \in \mathcal{I}}$  is a family of  $\mathbb{R}$ -valued random variables such that for every finite subset  $\mathcal{J} \subset \mathcal{I}$ , the random variables  $\{X(i)\}_{i \in \mathcal{J}}$  are jointly Normal. For example: let  $Z$  be a multivariate Normal random vector taking values in  $\mathbb{R}^n$ . Let  $\mathcal{I} = \{1, \dots, n\}$ . Define  $X(i) = Z_i$  for  $i \in \mathcal{I}$ . Then  $\{X(i)\}_{i \in \mathcal{I}}$  is a Gaussian process. Another example: in problems of learning a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  from data, the function of interest is modeled as a realization of a Gaussian process  $\{f(x)\}_{x \in \mathbb{R}^n}$  with a covariance kernel  $k(x, x')$  (Stein, 1999; Rasmussen & Williams, 2008).

The first example above had a finite index set. The second had a finite-dimensional index set. But the generality of Thm. 1 is amenable to an infinite-dimensional index set such as that obtained by a suitable restriction of  $L^2$ . For example, the modern theory of weak convergence is interested in identifying so-called Donsker classes of functions  $h : \mathbb{R}^n \rightarrow \mathbb{R}$  such that for a sequence of i.i.d.  $\mathbb{R}^n$ -valued random variables  $\{X_i\}_{i=1}^\infty$ ,

$$G(h) = \lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} \sum_{i=1}^n (h(X_i) - \mathbb{E} h(X))$$

<sup>4</sup>cf. Karvonen & Särkkä (2025) which does not correctly scale with the input covariance.

2106 converges in distribution to a Gaussian process indexed by  $h$  (Van Der Vaart & Wellner, 1996).  
 2107  
 2108 Prepared with the vocabulary of abstract Gaussian processes, we introduce isonormal Gaussian pro-  
 2109 cesses:

2110 **Definition 8** (Isonormal Gaussian process). *Let  $\mathfrak{h}$  be a separable Hilbert space. A stochastic process*  
 2111  *$W = \{W(h), h \in \mathfrak{h}\}$  is an isonormal Gaussian process if it satisfies the following two properties:*

- 2112 1. For every  $h \in \mathfrak{h}$ ,  $\mathbb{E} W(h) = 0$ .
- 2113 2. For every  $h_1, h_2 \in \mathfrak{h}$ ,  $\mathbb{E} W(h_1)W(h_2) = \langle h_1, h_2 \rangle_{\mathfrak{h}}$ .

2116 Let  $\{W(h)\}_{h \in \mathfrak{h}}$  be an isonormal Gaussian process. Let  $\mathcal{S}$  be the set of all random variables that  
 2117 depend “nicely” on a finite subset of  $W$ , i.e.

$$2118 \mathcal{S} = \{f(W(h_1), \dots, W(h_n)) \mid f \in C_c^\infty(\mathbb{R}^n), h_1, \dots, h_n \in \mathfrak{h}\}$$

2119 where  $C_c^\infty(\mathbb{R}^n)$  is the space of smooth functions with compact support. The Malliavin derivative is  
 2120 defined by

$$2122 Df(W(h_1), \dots, W(h_n)) = \sum_{i=1}^n \partial_i f(W(h_1), \dots, W(h_n)) h_i.$$

2124 The  $m$ th Malliavin derivative, which takes values in  $\mathfrak{h}^{\otimes m}$ , the  $m$ th tensor power of  $\mathfrak{h}$ , is defined  
 2125 iteratively. For  $m \geq 1$  and  $p \geq 1$ , define the space  $\mathbb{D}^{m,p}$  to be the closure of  $\mathcal{S}$  under the norm<sup>5</sup>

$$2127 \|F\|_{\mathbb{D}^{m,p}}^p = \mathbb{E} |F|^p + \sum_{i=1}^m \mathbb{E} \|D^i F\|_{\mathfrak{h}^{\otimes i}}^p.$$

2130 *Proof of Lemma 20.* In order to apply Thm. 1, we need to represent  $Y = f(X)$  as a member of  
 2131  $\mathbb{D}^{2,4}$ . Without loss of generality, we can assume that  $\mathbb{E} X = 0$ . Next, we represent  $X$  in terms of an  
 2132 isonormal Gaussian process. Let  $\mathfrak{h}$  be the Hilbert space  $\mathbb{R}^n$  equipped with the standard dot product.  
 2133 Let  $\xi$  be a standard Normal random vector taking values in  $\mathbb{R}^n$ . Define  $Z(h) = \langle h, \xi \rangle_{\mathfrak{h}}$  for  $h \in \mathfrak{h}$ .  
 2134 Then  $\{Z(h)\}_{h \in \mathfrak{h}}$  is an isonormal Gaussian process and  $\xi_i = Z(e_i)$  for  $i = 1, \dots, n$ , where  $e_i$  is the  
 2135  $i$ th standard basis vector. Next, we note that

$$2136 X \stackrel{\mathcal{D}}{=} \Sigma_X^{1/2} \xi$$

$$2137 \stackrel{\mathcal{D}}{=} \Sigma_X^{1/2} \sum_{i=1}^n \xi_i e_i.$$

2140 Therefore, by approximating  $f$  using smooth bounded functions, we can write

$$2141 Y = f(X)$$

$$2142 = f\left(\Sigma_X^{1/2} \sum_{j=1}^n \xi_j e_j\right)$$

2146 The Malliavin derivative of  $Y$  is given by

$$2147 DY_k = \sum_{i=1}^n \frac{\partial}{\partial \xi_i} f_k \left( \Sigma_X^{1/2} \sum_{j=1}^n \xi_j e_j \right) e_i$$

$$2148 = \sum_{i=1}^n \left\{ Df_k \left( \Sigma_X^{1/2} \sum_{j=1}^n \xi_j e_j \right) \left[ \frac{\partial}{\partial \xi_i} \left( \Sigma_X^{1/2} \sum_{j=1}^n \xi_j e_j \right) \right] \right\} e_i$$

$$2149 = \sum_{i=1}^n \left\{ Df_k \left( \Sigma_X^{1/2} \sum_{j=1}^n \xi_j e_j \right) \left[ \Sigma_X^{1/2} e_i \right] \right\} e_i$$

$$2150 = \Sigma_X^{1/2} \nabla f_k(X),$$

2158 <sup>5</sup>Compare to the Sobolev space  $W^{m,p}$ .

2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213

and likewise the second Malliavin derivative is (identifying  $\mathfrak{h} \otimes \mathfrak{h}$  with  $\mathbb{R}^{n \times n}$ )

$$D^2 Y_k = \Sigma_X^{1/2} \nabla^2 f_k(X) \Sigma_X^{1/2}.$$

Thus the conclusion of Thm. 1 becomes:

$$\begin{aligned} d_W(Y, \mathcal{N}_d(\mu_Y, \Sigma_T)) &\leq \frac{3}{\sqrt{2}} \|\Sigma_Y^{-1}\| \|\Sigma_Y\|^{1/2} \\ &\quad \cdot \underbrace{\left( \sum_{i=1}^d \mathbb{E} \|D^2 Y_i\|_{\text{op}}^4 \right)^{1/4}}_{\leq \|\Sigma_X\|^{2/2} \left( \sum_{i=1}^d \mathbb{E} \|\nabla^2 f_i(X)\|^4 \right)^{1/4}} \\ &\quad \cdot \underbrace{\left( \sum_{i=1}^d \mathbb{E} \|DY_i\|_{\mathfrak{h}}^4 \right)^{1/4}}_{\leq \|\Sigma_X\|^{1/2} \left( \sum_{i=1}^d \mathbb{E} \|\nabla f_i(X)\|^4 \right)^{1/4}}. \end{aligned}$$

□

---

## 2214 J SUPPLEMENT TO §5.2

2215

2216 A neural mapping  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and a variance  $\sigma^2$  are chosen to maximize the log-likelihood of the  
2217 generative model

$$2218 y | x \sim \mathcal{N}(f(x), \sigma^2) \quad (84)$$

2219

2220 for a set of inputs  $\{x_i\}_{i=1}^N$  and outputs  $\{y_i\}_{i=1}^N$ . At inference time, the input  $x$  is corrupted to  
2221  $\hat{X} = x + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$  is a perturbation with a known distribution. We make the prediction

$$2222 \hat{Y} | \hat{X} = f(\hat{X}) + \mathcal{N}(0, \sigma^2) \quad (85)$$

2223 which we approximate by

$$2224 \hat{Y} | \hat{X} \approx \mathcal{N}(\mathbb{E}(f(x) | \hat{X}), \sigma^2 + \text{Var}(f(x) | \hat{X})). \quad (86)$$

2225

2226 We apply this to the California housing dataset consisting of roughly 20,000  $(x, y)$  pairs, in which  
2227  $y$  is the log median house price of a block from the 1990 United States Census, and  $x$  is a vector of  
2228 eight features of that block consisting of latitude, longitude, and other demographic information. At  
2229 inference time, we corrupt the input  $x$  with Gaussian noise.

2230

### 2231 J.1 DATA

2232

2233 The California Housing dataset (Kelley Pace & Barry, 1997) consists of 20460 census blocks from  
2234 the 1990 United States Census. We split this dataset into roughly 70% training, 10% validation, and  
2235 20% test. We pre-processed this dataset by log-transforming some of the variables and standardizing  
2236 them using the means and standard deviations of the training set. The regressors are:

2237

- latitude of the block
- longitude of the block
- median income of the block
- (logarithm) average number of rooms per household
- (logarithm) average number of bedrooms per household
- (logarithm) population of the block
- (logarithm) number of households in the block

2245

2246 The predictor is the log median house price of the block.

2247

### 2248 J.2 NETWORK

2249

2250 We used a single-hidden-layer neural network to parameterize a model consisting of a superposition  
2251 of a linear term and a sinusoidal term:

$$2252 f(x) = C \sin(Ax + b) + C'x + d, \quad (87)$$

2253 implemented using two layers:

2254

$$2255 f(x) = \begin{pmatrix} C \\ C' \end{pmatrix} f^1(x) + d, \quad (88)$$

2256

$$2257 f^1(x) = \sin \left( \begin{pmatrix} A \\ 0 \end{pmatrix} x + \begin{pmatrix} b \\ 0 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (89)$$

2258

2259 The number of hidden neurons (the number of rows of  $A$ ) is 7. This number was the result of  
2260 hand-tuning to reduce the subjective appearance of overfitting in the validation set.

2261

### 2262 J.3 TRAINING

2263

2264 We initialized  $A$  with i.i.d. Gaussian entries having mean zero and variance equal to the reciprocal  
2265 of the number of columns. We initialized  $b$  with i.i.d. Uniform $(-\pi, \pi)$ . We initialized  $C$ ,  $C'$ , and  $d$   
2266 to zero.

2267

2267 We trained this network using the Adam optimizer on Optax (DeepMind et al., 2020) with learning  
rate  $10^{-1}$  for the first 2000 steps and  $10^{-2}$  for 6000 steps.

---

2268  
2269  
2270  
2271  
2272  
2273  
2274  
2275  
2276  
2277  
2278  
2279  
2280  
2281  
2282  
2283  
2284  
2285  
2286  
2287  
2288  
2289  
2290  
2291  
2292  
2293  
2294  
2295  
2296  
2297  
2298  
2299  
2300  
2301  
2302  
2303  
2304  
2305  
2306  
2307  
2308  
2309  
2310  
2311  
2312  
2313  
2314  
2315  
2316  
2317  
2318  
2319  
2320  
2321

#### J.4 INFERENCE

At inference time, we corrupt the input with Gaussian noise with zero variance in the latitude and longitude features, and 10X population covariance in the other six features. The two sources of uncertainty in the Monte Carlo simulation are the randomness of the train/val/test split (as judged relative to a hypothetical population) and the test data augmentation. In order to report a standard error that lumps both sources of uncertainty, we obtain 100 bootstrap samples of the test set. Within each bootstrap sample, we corrupt each input with 100 realizations of noise.

---

## 2322 K SUPPLEMENT TO §5.3

2323

2324 A neural mapping  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is trained to maximize the log-likelihood of the generative model

2325

$$2326 \quad y \mid x \sim \text{Bernoulli}(p) \quad (90)$$

2327

$$2327 \quad p := \Phi(f(x)) \quad (91)$$

2328

2329 for a set of inputs  $\{x_i\}_{i=1}^N$  and outputs  $\{y_i\}_{i=1}^N$ .

2330

2331 At inference time, the input  $x$  is corrupted to  $z = Px$ , where  $P$  is a wide measurement matrix. The  
2332 situation in which  $P$  is rank deficient corresponds to missing features. In order to use the model (90)  
2333 for inference, we impute the missing features using linear regression:

2333

$$2333 \quad \hat{X} \mid z = \mathcal{N}(\hat{\mu}, \hat{\Sigma}), \quad (92a)$$

2334

$$2334 \quad \hat{\mu} = \mu + \Sigma P^\top (P \Sigma P^\top)^{-1} (z - \mu), \quad (92b)$$

2335

$$2335 \quad \hat{\Sigma} = \Sigma - (\Sigma P^\top) (P \Sigma P^\top)^{-1} (P \Sigma). \quad (92c)$$

2336

2337 Here,  $\mu$  and  $\Sigma$  are the population mean and variance-covariance matrix of  $x$ , estimated from the  
2338 training data.

2339

2340 Afterwards we make the uncertainty-aware prediction

2341

$$2341 \quad \hat{Y} \mid \hat{X} \sim \text{Bernoulli}(\hat{p}) \quad (93)$$

2342

$$2342 \quad \hat{p} := \mathbb{E} \left[ \Phi(f(x)) \mid \hat{X} \right]. \quad (94)$$

2343

2344

2345 At inference time, we drop all features except ‘‘Operating Gross Margin’’ and impute the rest of the  
2346 balance sheet and cash flow features using linear regression following (92).

2347

### 2348 K.1 DATA

2349

2350 The Taiwanese bankruptcy dataset (Liang et al., 2016) consists of 6819 instances of Taiwanese com-  
2351 panies from 1999 to 2009. There are 95 features consisting of balance sheet (assets and liabilities)  
2352 and cash flow information. The target variable is a binary indicator of bankruptcy. We preprocessed  
2353 this dataset by standardizing the features using the means and standard deviations of the training set.  
2354 We split this dataset into roughly 70% training, 10% validation, and 20% test.

2355

### 2356 K.2 NETWORK

2357

2358 We used a single-hidden-layer neural network with a width of 200, depth of 3, and sine activation  
2359 function:

2360

$$2360 \quad f(x) = C \sin(Ax + b) + d \quad (95)$$

2361

2362 where  $A$  is a  $200 \times 95$  matrix,  $b$  is a  $200 \times 1$  vector,  $C$  is a  $1 \times 200$  vector, and  $d$  is a  $1 \times 1$  vector.  
2363 We initialized  $A$  with i.i.d. Gaussian entries having mean zero and variance equal to the reciprocal  
2364 of the number of columns. We initialized  $b$  with i.i.d. Uniform( $-\pi, \pi$ ). We initialized  $C$  to zero and  
2365  $d = \Phi^{-1}(\mathbb{E}_{\text{train}} y)$ .

2366

### 2367 K.3 TRAINING

2368

2369 We trained this network using the AdamW optimizer (Loshchilov & Hutter, 2019; DeepMind et al.,  
2370 2020) with learning rate  $10^{-5}$  for 5000 steps.

2371

### 2372 K.4 REPORTING

2373

2374 We report the average log predicted probability (higher is better) of the correct label on test data in  
2375 the Taiwanese bankruptcy dataset. We also report the standard error of the mean.

2376

### 2377 K.5 ADDITIONAL FIGURES

2376  
2377  
2378  
2379  
2380  
2381  
2382  
2383  
2384  
2385  
2386  
2387  
2388  
2389  
2390  
2391  
2392  
2393  
2394  
2395  
2396  
2397  
2398  
2399  
2400  
2401  
2402  
2403  
2404  
2405  
2406  
2407  
2408  
2409  
2410  
2411  
2412  
2413  
2414  
2415  
2416  
2417  
2418  
2419  
2420  
2421  
2422  
2423  
2424  
2425  
2426  
2427  
2428  
2429

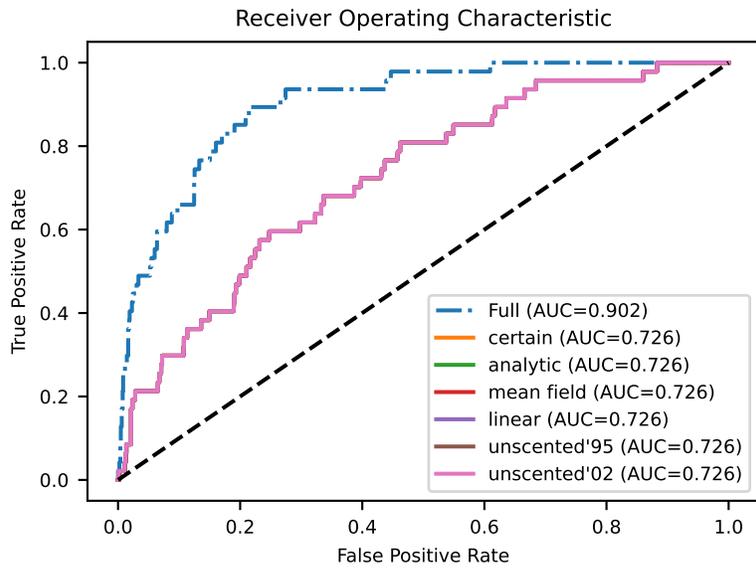


Figure 5: Receiver operator characteristic curve of test data in the Taiwanese bankruptcy dataset based on varying the threshold for  $\hat{p}$ .

2430  
2431  
2432  
2433  
2434  
2435  
2436  
2437  
2438  
2439  
2440  
2441  
2442  
2443  
2444  
2445  
2446  
2447  
2448  
2449  
2450  
2451  
2452  
2453  
2454  
2455  
2456  
2457  
2458  
2459  
2460  
2461  
2462  
2463  
2464  
2465  
2466  
2467  
2468  
2469  
2470  
2471  
2472  
2473  
2474  
2475  
2476  
2477  
2478  
2479  
2480  
2481  
2482  
2483

---

## L SUPPLEMENT TO §5.4

We expand on each of the terms in the evidence lower bound (2), repeated below.

$$\theta^* = \arg \min_{\theta} \left\{ \mathbb{E}_{w \sim q(w; \theta)} \log p(y | w, x) + D_{\text{KL}, w}(q(w; \theta) | p(w)) \right\}.$$

Our neural network  $f(x; w)$  consists of a single hidden layer. The weights  $w$  encompass:

- $A, b$ : pre-activation weights,
- $C, d$ : post-activation weights, as well as
- $\lambda$ : (homoscedastic) output log precision.

The log likelihood of  $w$  is Normal:

$$\log p(y | w, x) = \frac{1}{2} \left( -\lambda + e^{\lambda} [y - f(x; w)]^2 \right) + \frac{1}{2} \log [2\pi].$$

The prior distribution  $p(w)$  is the Kaiming initialization of independent Normal distributions, zero mean in each matrix entry and variance equal to  $2/\text{fan-in}$ . Like Wu et al. (2019); Petersen et al. (2024); Wright et al. (2024), the variational distribution  $q(w; \theta)$  is a Gaussian matrix with independent entries, parameterized by means and log precisions.

We train using the Adam optimizer with a learning rate of 0.1 for 5 000 epochs. Each epoch draws 10 random samples from the variational distribution  $q(w; \theta)$  per training instance and uses them to compute the gradient of the ELBO.

Each network takes about 2 minutes on a Nvidia T1200 GPU.

In reporting, we draw one million random variates from the variational posterior distribution for up to 100 instances from the test set.

The datasets are

- Combined cycle power plant (Pnar Tfekci, 2014)
- Concrete compressive strength (I-Cheng Yeh, 1998)
- Energy efficiency (we predict the heating load) (Athanasios Tsanas, 2012)
- Wine quality (Paulo Cortez, 2009)

2484 L.1 RESULTS

2485

2486

2487

2488

2489

Table 6: KL divergence between ground truth predictive distribution (by Monte Carlo) and approximations. W24@ $k$  means Wright et al. (2024) with  $k$  terms in the series expansion. Mean and standard error of the mean (over the test set).

2490

2491

2492

2493

2494

2495

2496

2497

2498

2499

2500

2501

2502

2503

2504

2505

2506

2507

2508

2509

2510

2511

2512

2513

2514

2515

2516

2517

2518

2519

2520

2521

2522

2523

2524

2525

2526

2527

2528

2529

2530

2531

2532

2533

2534

2535

2536

2537

Dataset	Method	KL divergence from GT
Combined Cycle Power Plant	analytic (ours)	$1.042 \times 10^{-6} \pm 1.0 \times 10^{-7}$
	W24@5	$2.460 \times 10^{-4} \pm 1.6 \times 10^{-5}$
	W24@4	$7.549 \times 10^{-4} \pm 6.0 \times 10^{-5}$
	W24@3	$1.336 \times 10^{-3} \pm 1.4 \times 10^{-4}$
	W24@2	$1.618 \times 10^{-3} \pm 2.0 \times 10^{-4}$
	W24@1	$1.794 \times 10^{-3} \pm 2.2 \times 10^{-4}$
Concrete Compressive Strength	analytic (ours)	$1.100 \times 10^{-6} \pm 1.2 \times 10^{-7}$
	W24@5	$2.230 \times 10^{-4} \pm 2.3 \times 10^{-5}$
	W24@4	$2.973 \times 10^{-4} \pm 3.0 \times 10^{-5}$
	W24@3	$1.601 \times 10^{-3} \pm 1.7 \times 10^{-4}$
	W24@2	$9.629 \times 10^{-3} \pm 1.0 \times 10^{-3}$
	W24@1	$2.520 \times 10^{-2} \pm 3.2 \times 10^{-3}$
Energy Efficiency	analytic (ours)	$1.069 \times 10^{-6} \pm 1.4 \times 10^{-7}$
	W24@5	$2.041 \times 10^{-4} \pm 1.2 \times 10^{-5}$
	W24@4	$2.459 \times 10^{-4} \pm 1.3 \times 10^{-5}$
	W24@3	$1.542 \times 10^{-3} \pm 7.8 \times 10^{-5}$
	W24@2	$7.640 \times 10^{-3} \pm 4.5 \times 10^{-4}$
	W24@1	$1.450 \times 10^{-2} \pm 9.5 \times 10^{-4}$
Wine Quality	analytic (ours)	$1.616 \times 10^{-6} \pm 1.9 \times 10^{-7}$
	W24@5	$5.292 \times 10^{-4} \pm 4.2 \times 10^{-5}$
	W24@4	$1.764 \times 10^{-3} \pm 1.2 \times 10^{-4}$
	W24@3	$7.337 \times 10^{-3} \pm 5.8 \times 10^{-4}$
	W24@2	$1.672 \times 10^{-2} \pm 1.7 \times 10^{-3}$
	W24@1	$2.876 \times 10^{-2} \pm 4.1 \times 10^{-3}$

Table 7: Train and test log-likelihoods of the Bayesian networks for all datasets. We do not report experimental uncertainty, as these networks as taken as fixed for the purpose of approximating the output distribution.

Dataset	Train log-likelihood	Test log-likelihood
Energy Efficiency	$-5.532 \times 10^{-1}$	$-6.029 \times 10^{-1}$
Concrete Compressive Strength	$-9.909 \times 10^{-1}$	$-9.253 \times 10^{-1}$
Wine Quality	$-1.292 \times 10^0$	$-1.242 \times 10^0$
Combined Cycle Power Plant	$-1.621 \times 10^{-1}$	$-9.716 \times 10^{-2}$

2538  
2539  
2540  
2541  
2542  
2543  
2544  
2545  
2546  
2547  
2548  
2549  
2550  
2551  
2552  
2553  
2554  
2555  
2556  
2557  
2558  
2559  
2560  
2561  
2562  
2563  
2564  
2565  
2566  
2567  
2568  
2569  
2570  
2571  
2572  
2573  
2574  
2575  
2576  
2577  
2578  
2579  
2580  
2581  
2582  
2583  
2584  
2585  
2586  
2587  
2588  
2589  
2590  
2591

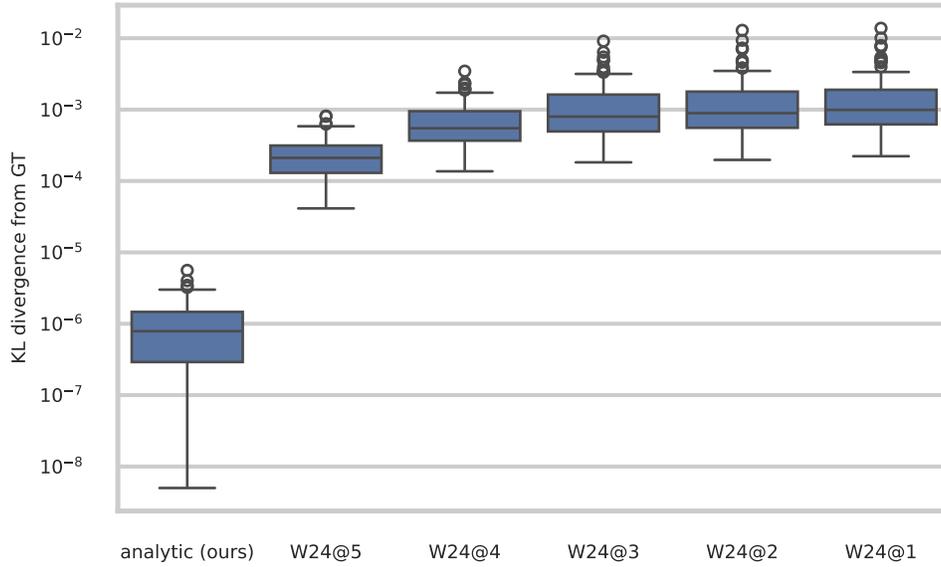


Figure 6: KL divergence between ground truth predictive distribution (by Monte Carlo) and approximations for the combined cycle power plant dataset. W24@ $k$  means Wright et al. (2024) with  $k$  terms in the series expansion.

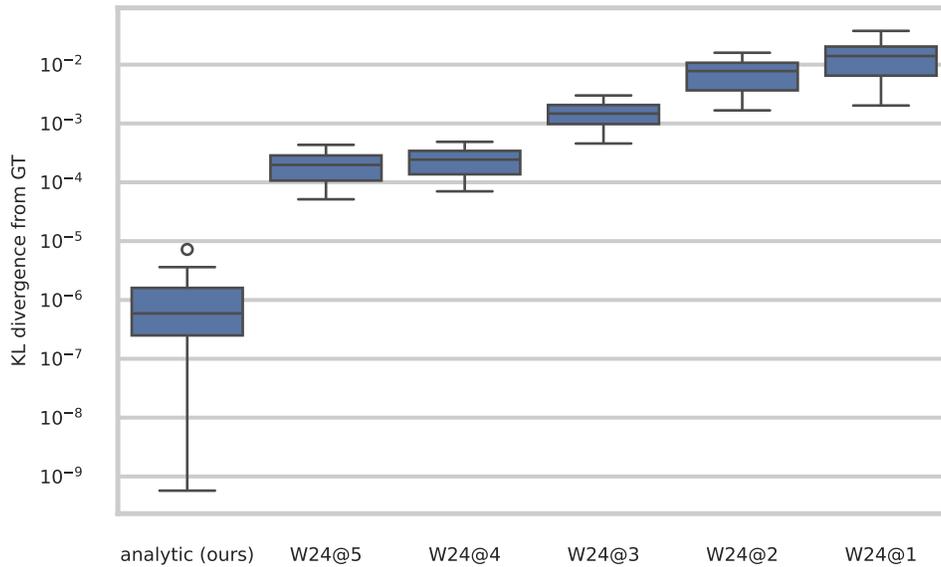


Figure 7: KL divergence between ground truth predictive distribution (by Monte Carlo) and approximations for the energy efficiency dataset. W24@ $k$  means Wright et al. (2024) with  $k$  terms in the series expansion.

2592  
2593  
2594  
2595  
2596  
2597  
2598  
2599  
2600  
2601  
2602  
2603  
2604  
2605  
2606  
2607  
2608  
2609  
2610  
2611  
2612  
2613  
2614  
2615  
2616  
2617  
2618  
2619  
2620  
2621  
2622  
2623  
2624  
2625  
2626  
2627  
2628  
2629  
2630  
2631  
2632  
2633  
2634  
2635  
2636  
2637  
2638  
2639  
2640  
2641  
2642  
2643  
2644  
2645

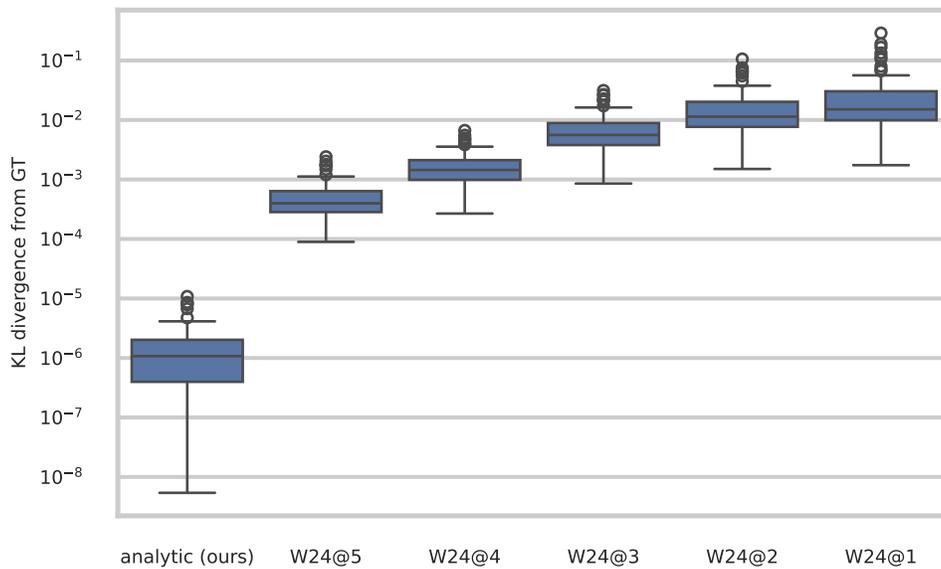


Figure 8: KL divergence between ground truth predictive distribution (by Monte Carlo) and approximations for the wine quality dataset. W24@ $k$  means Wright et al. (2024) with  $k$  terms in the series expansion.

2646  
2647  
2648  
2649  
2650  
2651  
2652  
2653  
2654  
2655  
2656  
2657  
2658  
2659  
2660  
2661  
2662  
2663  
2664  
2665  
2666  
2667  
2668  
2669  
2670  
2671  
2672  
2673  
2674  
2675  
2676  
2677  
2678  
2679  
2680  
2681  
2682  
2683  
2684  
2685  
2686  
2687  
2688  
2689  
2690  
2691  
2692  
2693  
2694  
2695  
2696  
2697  
2698  
2699

## M SUPPLEMENT TO §5.5

The random variable  $\mathbf{1}_{U < \Phi(x)}$  has a Bernoulli distribution with parameter  $\Phi(x)$ , which contributes an additional diagonal variance term to Lemma 1 as we see from applying the Law of Total Covariance:

$$\begin{aligned} (\text{Cov } g_{\tilde{\sigma}}(X; A, b, C, d))_{i,j} &= \left( \text{Cov } \mathbb{E} [g_{\tilde{\sigma}}(X; A, b, C, d) | X] \right)_{i,j} \\ &\quad + \left( \mathbb{E} \text{Cov} [g_{\tilde{\sigma}}(X; A, b, C, d) | X] \right)_{i,j} \end{aligned} \quad (96)$$

$$\left( \text{Cov } \mathbb{E} [g_{\tilde{\sigma}}(X; A, b, C, d) | X] \right)_{i,j} = (\text{Cov } g_{\sigma}(X; A, b, C, d))_{i,j} \quad (97)$$

since  $\mathbb{E}(\tilde{\sigma}(X, U) | X) = \sigma(X)$  and

$$\left( \mathbb{E} \text{Cov} [g_{\tilde{\sigma}}(X; A, b, C, d) | X] \right)_{i,j} = 4\delta_{ij} \mathbb{E} \text{Var} \left[ (g_{\tilde{\sigma}}(X; A, b, C, d))_i | X \right] \quad (98)$$

$$= 4\delta_{ij} \mathbb{E} \Phi(\xi_i) (1 - \Phi(\xi_i)) \quad (99)$$

where  $\xi_i \sim \mathcal{N}(\mu_i, \nu_{ii})$ . Using Owen's T function (Owen, 1980), we have

$$\mathbb{E} \Phi(\xi_i) (1 - \Phi(\xi_i)) = 2T \left( \frac{\mu_i}{\sqrt{1 + \nu_{ii}}}, \frac{\nu_{ii}}{1 + 2\nu_{ii}} \right). \quad (100)$$

---

## N SUPPLEMENT TO §5.1

We apply our method and other benchmarks to 38 different ensembles of random neural networks:

- network architecture  $\in$  {wide, deep}
- weights  $\in$  {initialized, trained}
- activation function  $\in$  {probit, probit residual, sine, sine residual}

(There are no trained networks containing Heaviside or Heaviside-residual layers.) From each ensemble, we sample one neural network and evaluate the goodness of approximation of the output distribution for input distributions:

- variance  $\in$  {small, medium, large}.

In each case we compare the distributions of:

- $Y_0$ , the true distribution
- $Y_1$ , the pseudo-true Gaussian distribution
- $Y_{\text{ana}}$ , the analytic layer-wise Gaussian approximation (our method)
- $Y_{\text{mfa}}$ , mean-field: applying our method for moment propagation and setting off-diagonal layer covariances to zero
- $Y_{\text{lin}}$ , linearization-based moment propagation
- $Y_{u'95}$ , unscented transform of the whole network using  $\kappa = 2$
- $Y_{u'02}$ , unscented transform of the whole network using  $\alpha = 0.001, \beta = 2, \kappa = 2$

### ENSEMBLES OF NEURAL NETWORKS

This section specifies the 38 ensembles of random neural networks and the three ensembles of inputs used to produce the 114 test cases that follow. The neural networks in this example are parameterized as in Def. 1. Let  $d_{\text{hidden}}$  and  $w_{\text{hidden}}$  be the depth and width of the hidden layers, respectively. A random neural network  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$  is specified with  $d_{\text{hidden}} + 1$  layers. The output layer is linear, with weights  $C^{d_{\text{hidden}}+1}$  and biases  $d^{d_{\text{hidden}}+1}$ . The first hidden layer has  $C^1 = 0_{w_{\text{hidden}} \times 1}$ ,  $d^1 = 0_{w_{\text{hidden}} \times 1}$ . The four degrees of freedom in the test cases are:

- architecture**
- if architecture = wide, then  $d_{\text{hidden}} = 5, w_{\text{hidden}} = 400$ ;
  - if architecture = deep, then  $d_{\text{hidden}} = 20, w_{\text{hidden}} = 100$ .
- weights**
- if weights = initialized:
    - $A$  matrices are initialized with i.i.d. Gaussian entries having mean zero and variance equal to the reciprocal of the number of columns times  $\sqrt{2}$
    - if activation  $\in$  {probit, probit residual}, then  $b$  vectors are initialized with independently sampled entries from  $\mathcal{N}(0, 1)$
    - if activation  $\in$  {sine, sine residual}, then  $b$  vectors are initialized with independently sampled entries from  $\mathcal{U}(-\pi, \pi)$
    - if activation  $\in$  {probit, sine}, then  $C$  matrices are initialized to the zero matrix.
    - if activation  $\in$  {probit residual, sine residual}, then square  $C$  matrices are initialized to the identity matrix, and all other  $C$  matrices are initialized to the zero matrix.
    - $d$  vectors are initialized to the zero vector.
  - if weights = trained and activation  $\notin$  {Heaviside, Heaviside residual}, the following initialization as above, the neural network is trained to minimize the mean squared error loss on a pseudorandomly generated dataset of ten  $(x, y)$  samples drawn from  $\mathcal{N}(0, 1)$ . Training consists of using the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of  $10^{-6}$  for 30,000 iterations and until the loss is less than  $10^{-8}$  (whichever is later). Implementation due to Optax, DeepMind et al. (2020)

- 
- 2754 **activation function** • if activation  $\in \{\text{probit}, \text{probit residual}\}$ , then  $\sigma(x) = 2\Phi(x) - 1$  where  $\Phi$   
2755 is the cumulative distribution function of the standard normal distribution  
2756 • if activation  $\in \{\text{sine}, \text{sine residual}\}$ , then  $\sigma(x) = \sin(x)$   
2757 • if activation  $\in \{\text{GeLU}, \text{GeLU residual}\}$ , then  $\sigma(x) = x\Phi(x)$ .  
2758 • if activation  $\in \{\text{ReLU}, \text{ReLU residual}\}$ , then  $\sigma(x) = \max(0, x)$ .  
2759 • if activation  $\in \{\text{Heaviside}, \text{Heaviside residual}\}$ , then  $\sigma(x) = \mathbf{1}_{\{x \geq 0\}}$   
2760  
2761 **variance** • if variance = small, then  $X \sim \mathcal{N}(0, 10^{-2}I)$   
2762 • if variance = medium, then  $X \sim \mathcal{N}(0, I)$   
2763 • if variance = large, then  $X \sim \mathcal{N}(0, 10^2I)$   
2764

## 2765 SIMULATION AND REPORTING

2767 The only source of uncertainty in our numerical results is the “true distribution” of  $Y_0 = f(X)$ . For  
2768 this we use twenty independent realizations of  $N = 2^{16}$  quasi-Monte Carlo samples (Virtanen et al.,  
2769 2020). We report statistical uncertainty in the form of mean  $\pm$  standard error within the independent  
2770 realizations. The distributional uncertainty is too small to visualize in figures, so we plot the pooled  
2771 data without an uncertainty indication.

2772 The Wasserstein distance between a Normal distribution and  $Y_0$  is computed by

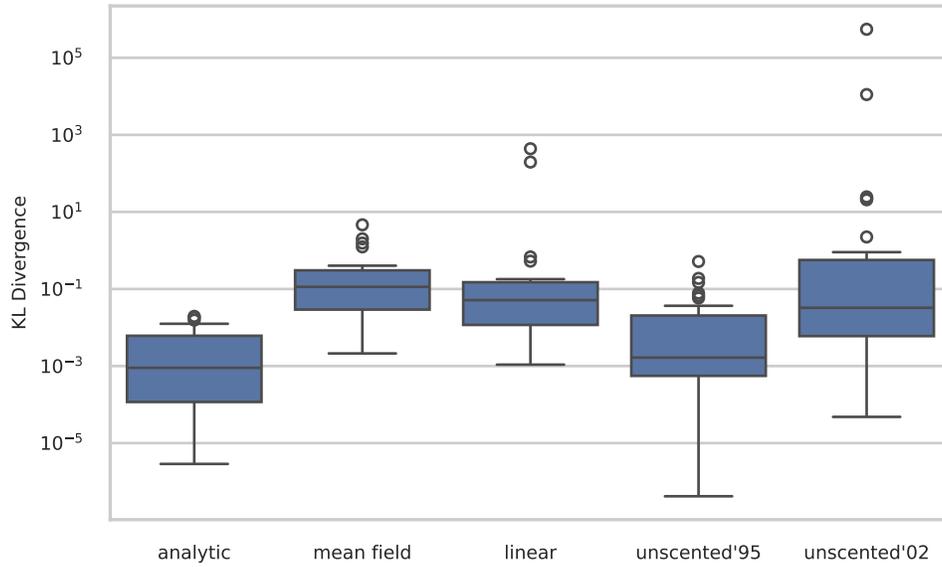
$$2773 d_W(\mathcal{N}(\mu, \sigma^2), Y_0) \approx \frac{1}{N} \sum_{i=1}^N \left| y_{(i)} - Q\left(\frac{i-1/2}{N}\right) \right|$$

2774 where  $\{y_{(i)}\}_{i=1}^N$  are the quasi-Monte Carlo samples of  $Y_0$  sorted in ascending order, and  $Q$  is the  
2775 quantile function of  $\mathcal{N}(\mu, \sigma^2)$ . When plotting distributions, we show  $Y_0$  using a histogram with 50  
2776 bins.

2780 The horizontal axes are scaled to include the 0.5th (99.5th) percentiles of the quasi-Monte Carlo sam-  
2781 ples of  $Y_0$ , or the mean minus (plus) three standard deviations of  $Y_1$ , whichever is smaller (greater).  
2782

2783 The entire suite of 72 test cases takes (including initialization, training, quasi-Monte Carlo, and  
2784 reporting) 18 minutes on an Ubuntu system with a 11th Gen Intel® Core™ i7-11850H CPU and  
2785 32GB of RAM.  
2786  
2787  
2788  
2789  
2790  
2791  
2792  
2793  
2794  
2795  
2796  
2797  
2798  
2799  
2800  
2801  
2802  
2803  
2804  
2805  
2806  
2807

2808 N.1 SUMMARIES  
 2809  
 2810  
 2811



2833 Figure 9: Comparison of goodness of approximation (lower KL divergence is better) for all random  
 2834 neural networks, grouped by approximation method, in the small input variance scenario.  
 2835  
 2836  
 2837  
 2838  
 2839  
 2840

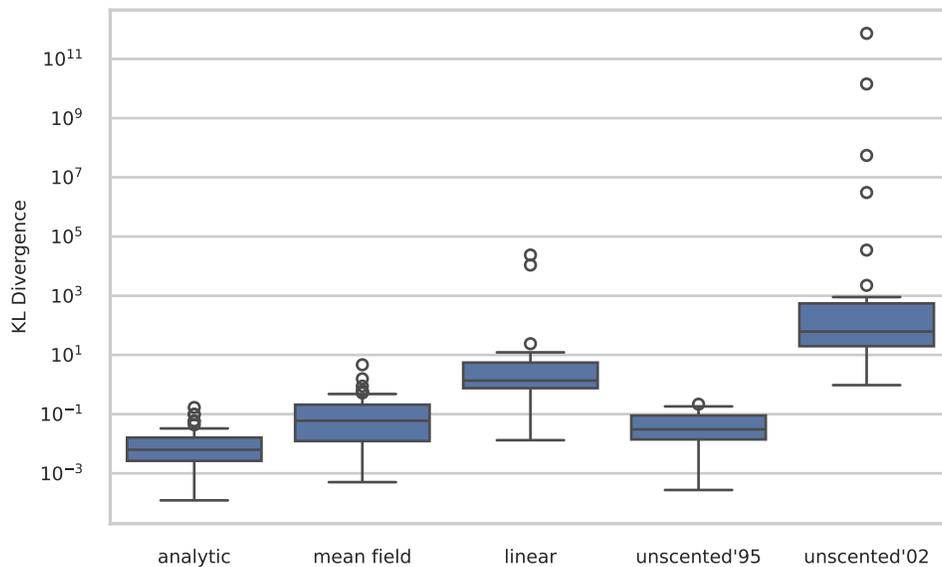


Figure 10: Comparison of goodness of approximation (lower KL divergence is better) for all random  
 neural networks, grouped by approximation method, in the medium input variance scenario.

2862  
2863  
2864  
2865  
2866  
2867  
2868  
2869  
2870  
2871  
2872  
2873  
2874  
2875  
2876  
2877  
2878  
2879  
2880  
2881  
2882  
2883  
2884  
2885  
2886  
2887  
2888  
2889  
2890  
2891  
2892  
2893  
2894  
2895  
2896  
2897  
2898  
2899  
2900  
2901  
2902  
2903  
2904  
2905  
2906  
2907  
2908  
2909  
2910  
2911  
2912  
2913  
2914  
2915

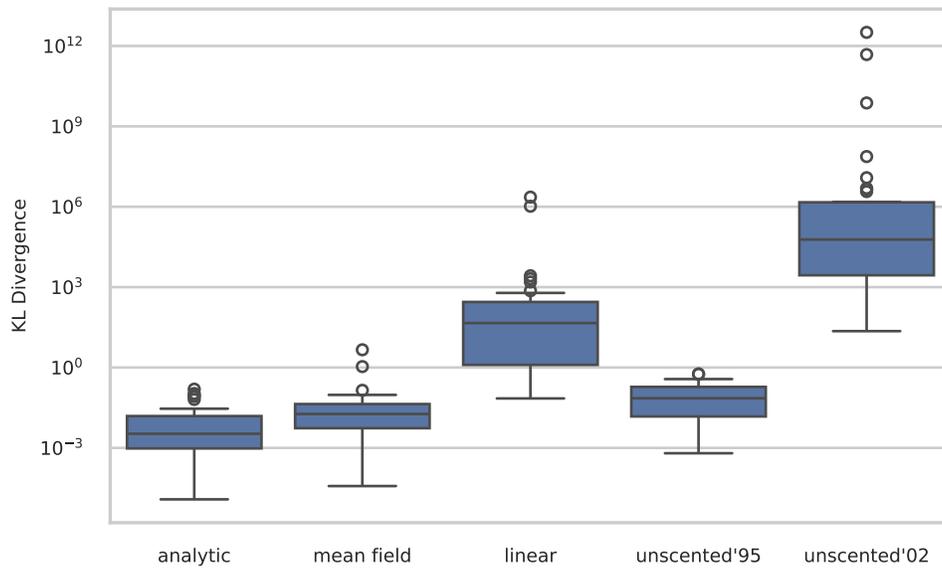


Figure 11: Comparison of goodness of approximation (lower KL divergence is better) for all random neural networks, grouped by approximation method, in the large input variance scenario.

2916  
 2917  
 2918  
 2919  
 2920  
 2921  
 2922  
 2923  
 2924  
 2925  
 2926  
 2927  
 2928  
 2929  
 2930  
 2931  
 2932  
 2933  
 2934  
 2935  
 2936  
 2937  
 2938  
 2939  
 2940  
 2941  
 2942  
 2943  
 2944  
 2945  
 2946  
 2947  
 2948  
 2949  
 2950  
 2951  
 2952  
 2953  
 2954  
 2955  
 2956  
 2957  
 2958  
 2959  
 2960  
 2961  
 2962  
 2963  
 2964  
 2965  
 2966  
 2967  
 2968  
 2969

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-3.856 \times 10^{-1} \pm 2.5 \times 10^{-8}$	$4.863 \times 10^{-4} \pm 7.6 \times 10^{-9}$
analytic	$-3.856 \times 10^{-1}$	$4.632 \times 10^{-4}$
mean-field	$-3.843 \times 10^{-1}$	$7.883 \times 10^{-4}$
linear	$-3.782 \times 10^{-1}$	$4.533 \times 10^{-4}$
unscented'95	$-3.856 \times 10^{-1}$	$4.485 \times 10^{-4}$
unscented'02	$-3.858 \times 10^{-1}$	$5.698 \times 10^{-4}$

Table 8: Comparison of moments for Network(architecture=wide, weights=initialized, activation=probit), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.745 \times 10^{-2} \pm 9.5 \times 10^{-7}$	0
analytic	$1.738 \times 10^{-2} \pm 1.2 \times 10^{-6}$	$5.828 \times 10^{-4} \pm 3.7 \times 10^{-7}$
mean-field	$3.560 \times 10^{-2} \pm 1.3 \times 10^{-6}$	$7.064 \times 10^{-2} \pm 4.9 \times 10^{-6}$
linear	$4.981 \times 10^{-2} \pm 1.9 \times 10^{-7}$	$5.747 \times 10^{-2} \pm 3.8 \times 10^{-7}$
unscented'95	$1.756 \times 10^{-2} \pm 9.0 \times 10^{-7}$	$1.598 \times 10^{-3} \pm 6.1 \times 10^{-7}$
unscented'02	$2.102 \times 10^{-2} \pm 7.5 \times 10^{-7}$	$6.681 \times 10^{-3} \pm 1.4 \times 10^{-6}$

Table 9: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=probit), variance=small

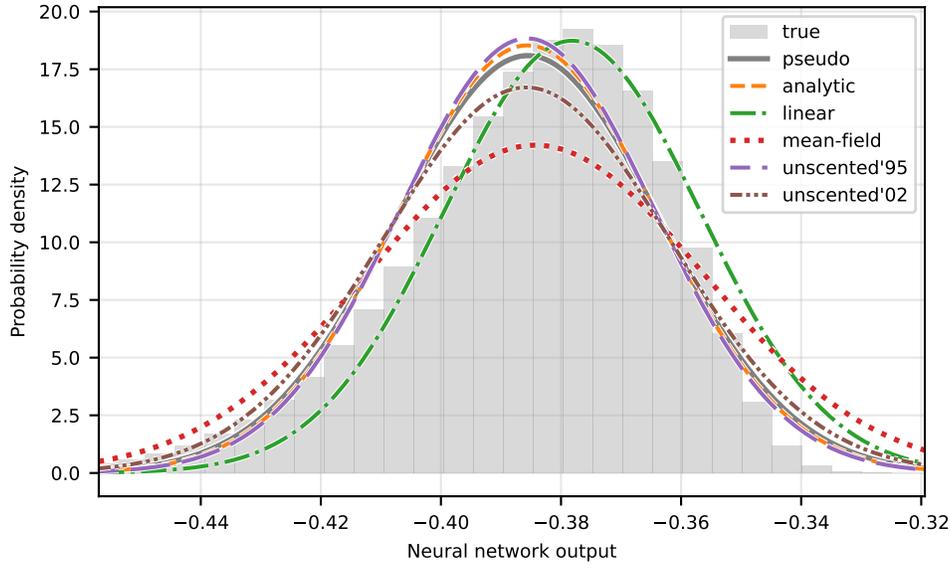


Figure 12: Probability distributions for Network(architecture=wide, weights=initialized, activation=probit), variance=small

2970  
 2971  
 2972  
 2973  
 2974  
 2975  
 2976  
 2977  
 2978  
 2979  
 2980  
 2981  
 2982  
 2983  
 2984  
 2985  
 2986  
 2987  
 2988  
 2989  
 2990  
 2991  
 2992  
 2993  
 2994  
 2995  
 2996  
 2997  
 2998  
 2999  
 3000  
 3001  
 3002  
 3003  
 3004  
 3005  
 3006  
 3007  
 3008  
 3009  
 3010  
 3011  
 3012  
 3013  
 3014  
 3015  
 3016  
 3017  
 3018  
 3019  
 3020  
 3021  
 3022  
 3023

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.609 \times 10^{-1} \pm 5.2 \times 10^{-7}$	$1.946 \times 10^{-2} \pm 2.2 \times 10^{-7}$
analytic	$-5.710 \times 10^{-1}$	$1.773 \times 10^{-2}$
mean-field	$-5.412 \times 10^{-1}$	$2.545 \times 10^{-2}$
linear	$-3.782 \times 10^{-1}$	$4.533 \times 10^{-2}$
unscented'95	$-5.943 \times 10^{-1}$	$8.656 \times 10^{-3}$
unscented'02	$-1.141 \times 10^0$	$1.210 \times 10^0$

Table 10: Comparison of moments for Network(architecture=wide, weights=initialized, activation=probit), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.924 \times 10^{-2} \pm 1.7 \times 10^{-6}$	0
analytic	$5.182 \times 10^{-2} \pm 3.1 \times 10^{-6}$	$4.751 \times 10^{-3} \pm 6.0 \times 10^{-7}$
mean-field	$5.895 \times 10^{-2} \pm 2.6 \times 10^{-6}$	$2.964 \times 10^{-2} \pm 2.1 \times 10^{-6}$
linear	$4.914 \times 10^{-1} \pm 2.3 \times 10^{-6}$	$1.099 \times 10^0 \pm 1.9 \times 10^{-5}$
unscented'95	$1.385 \times 10^{-1} \pm 3.3 \times 10^{-6}$	$1.562 \times 10^{-1} \pm 3.2 \times 10^{-6}$
unscented'02	$2.423 \times 10^0 \pm 8.2 \times 10^{-6}$	$3.719 \times 10^{+1} \pm 4.5 \times 10^{-4}$

Table 11: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=probit), variance=medium

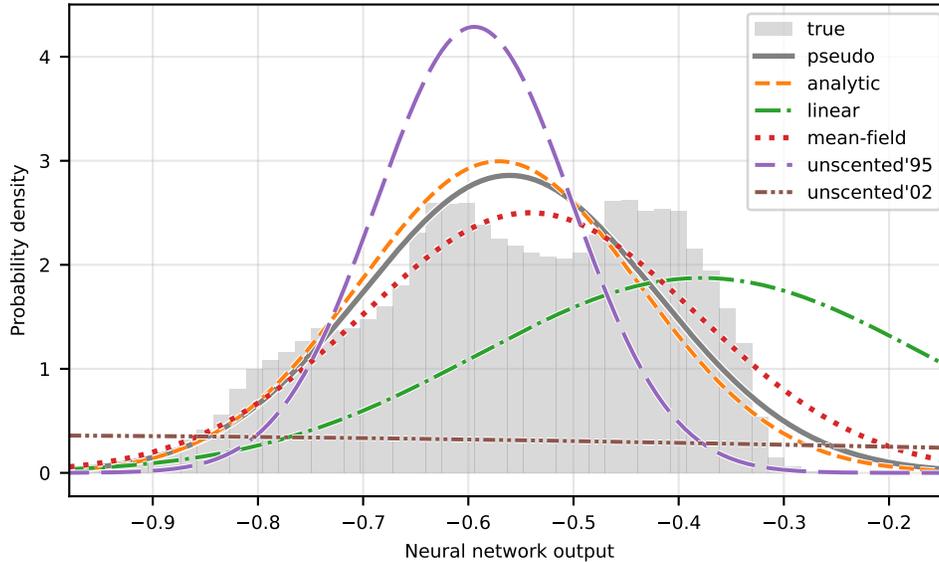


Figure 13: Probability distributions for Network(architecture=wide, weights=initialized, activation=probit), variance=medium

3024  
3025  
3026  
3027  
3028  
3029  
3030  
3031  
3032

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.141 \times 10^{-1} \pm 9.1 \times 10^{-6}$	$3.372 \times 10^{-2} \pm 2.8 \times 10^{-6}$
analytic	$-6.263 \times 10^{-1}$	$3.250 \times 10^{-2}$
mean-field	$-5.718 \times 10^{-1}$	$4.610 \times 10^{-2}$
linear	$-3.782 \times 10^{-1}$	$4.533 \times 10^0$
unscented'95	$-6.768 \times 10^{-1}$	$4.894 \times 10^{-2}$
unscented'02	$-7.669 \times 10^{-1}$	$1.165 \times 10^{+4}$

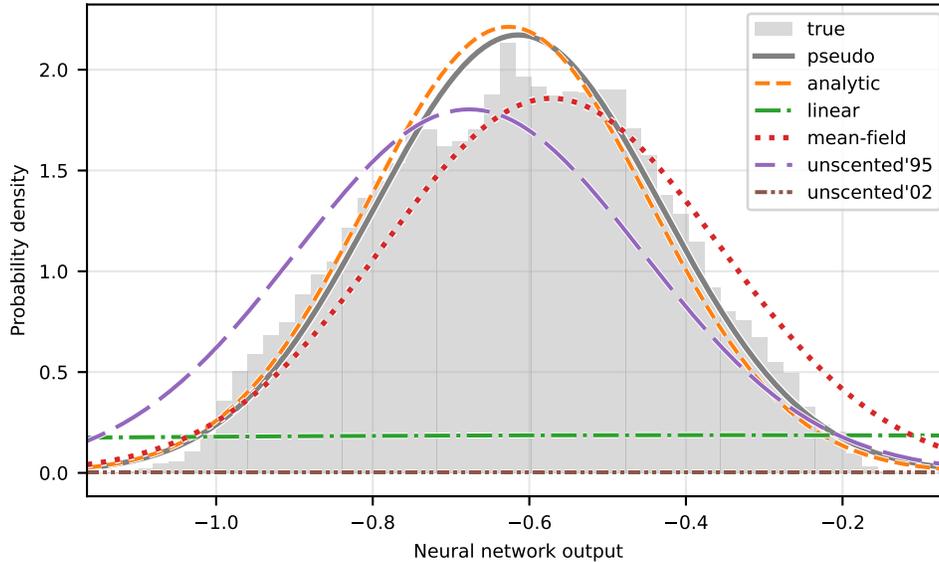
Table 12: Comparison of moments for Network(architecture=wide, weights=initialized, activation=probit), variance=large

3033  
3034  
3035  
3036  
3037  
3038  
3039  
3040  
3041  
3042  
3043  
3044  
3045

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.457 \times 10^{-2} \pm 1.6 \times 10^{-5}$	0
analytic	$3.666 \times 10^{-2} \pm 1.7 \times 10^{-5}$	$2.566 \times 10^{-3} \pm 3.5 \times 10^{-6}$
mean-field	$1.093 \times 10^{-1} \pm 2.0 \times 10^{-5}$	$5.363 \times 10^{-2} \pm 2.0 \times 10^{-5}$
linear	$3.637 \times 10^0 \pm 8.9 \times 10^{-5}$	$6.508 \times 10^{+1} \pm 5.6 \times 10^{-3}$
unscented'95	$1.561 \times 10^{-1} \pm 2.3 \times 10^{-5}$	$9.772 \times 10^{-2} \pm 3.0 \times 10^{-5}$
unscented'02	$2.486 \times 10^{+2} \pm 5.2 \times 10^{-3}$	$2.585 \times 10^{+5} \pm 2.1 \times 10^{+1}$

Table 13: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=probit), variance=large

3046  
3047  
3048  
3049  
3050  
3051  
3052  
3053  
3054  
3055  
3056  
3057  
3058  
3059  
3060  
3061  
3062  
3063  
3064  
3065  
3066  
3067



3068  
3069  
3070  
3071  
3072  
3073  
3074  
3075  
3076  
3077

Figure 14: Probability distributions for Network(architecture=wide, weights=initialized, activation=probit), variance=large

3078  
3079  
3080  
3081  
3082  
3083  
3084  
3085  
3086  
3087  
3088  
3089  
3090  
3091  
3092  
3093  
3094  
3095  
3096  
3097  
3098  
3099  
3100  
3101  
3102  
3103  
3104  
3105  
3106  
3107  
3108  
3109  
3110  
3111  
3112  
3113  
3114  
3115  
3116  
3117  
3118  
3119  
3120  
3121  
3122  
3123  
3124  
3125  
3126  
3127  
3128  
3129  
3130  
3131

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.698 \times 10^{-1} \pm 3.2 \times 10^{-7}$	$1.954 \times 10^{-1} \pm 1.2 \times 10^{-6}$
analytic	$+2.691 \times 10^{-1}$	$1.851 \times 10^{-1}$
mean-field	$+3.184 \times 10^{-1}$	$1.325 \times 10^{-3}$
linear	$+3.554 \times 10^{-1}$	$1.977 \times 10^{-1}$
unscented'95	$+2.673 \times 10^{-1}$	$1.934 \times 10^{-1}$
unscented'02	$+2.623 \times 10^{-1}$	$2.151 \times 10^{-1}$

Table 14: Comparison of moments for Network(architecture=wide, weights=trained, activation=probit), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$8.960 \times 10^{-2} \pm 2.4 \times 10^{-6}$	0
analytic	$9.035 \times 10^{-2} \pm 2.6 \times 10^{-6}$	$7.259 \times 10^{-4} \pm 1.6 \times 10^{-7}$
mean-field	$4.805 \times 10^{-1} \pm 2.4 \times 10^{-6}$	$2.006 \times 10^0 \pm 3.0 \times 10^{-6}$
linear	$1.288 \times 10^{-1} \pm 3.6 \times 10^{-7}$	$1.879 \times 10^{-2} \pm 1.0 \times 10^{-7}$
unscented'95	$9.089 \times 10^{-2} \pm 2.5 \times 10^{-6}$	$4.188 \times 10^{-5} \pm 2.7 \times 10^{-8}$
unscented'02	$9.728 \times 10^{-2} \pm 2.0 \times 10^{-6}$	$2.525 \times 10^{-3} \pm 3.1 \times 10^{-7}$

Table 15: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=probit), variance=small

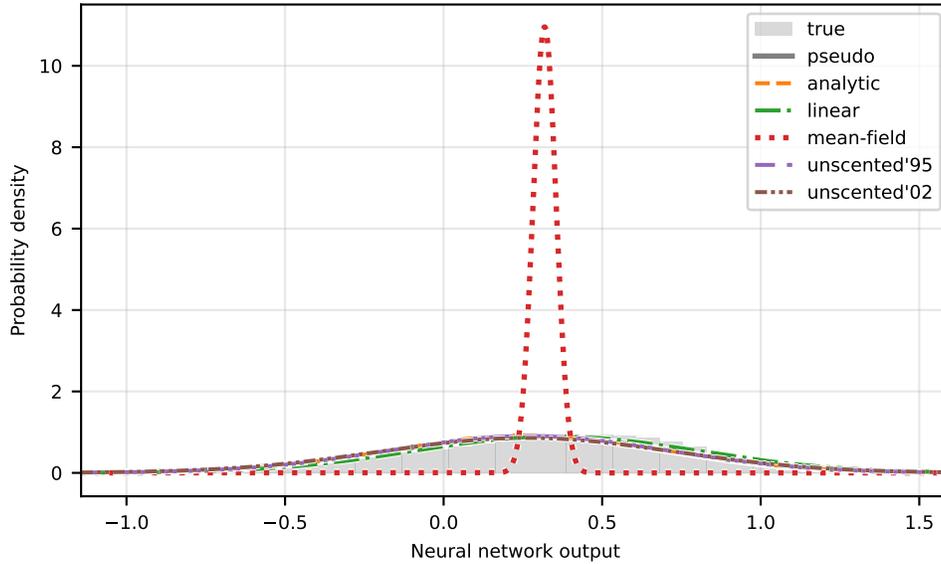


Figure 15: Probability distributions for Network(architecture=wide, weights=trained, activation=probit), variance=small

3132  
3133  
3134  
3135  
3136  
3137  
3138  
3139  
3140  
3141  
3142  
3143  
3144  
3145  
3146  
3147  
3148  
3149  
3150  
3151  
3152  
3153  
3154  
3155  
3156  
3157  
3158  
3159  
3160  
3161  
3162  
3163  
3164  
3165  
3166  
3167  
3168  
3169  
3170  
3171  
3172  
3173  
3174  
3175  
3176  
3177  
3178  
3179  
3180  
3181  
3182  
3183  
3184  
3185

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-8.887 \times 10^{-1} \pm 1.6 \times 10^{-6}$	$2.394 \times 10^0 \pm 6.7 \times 10^{-6}$
analytic	$-8.854 \times 10^{-1}$	$1.654 \times 10^0$
mean-field	$-6.635 \times 10^{-1}$	$4.026 \times 10^{-2}$
linear	$+3.554 \times 10^{-1}$	$1.977 \times 10^{+1}$
unscented'95	$-9.842 \times 10^{-1}$	$2.938 \times 10^0$
unscented'02	$-8.960 \times 10^0$	$1.933 \times 10^{+2}$

Table 16: Comparison of moments for Network(architecture=wide, weights=trained, activation=probit), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.316 \times 10^{-1} \pm 1.0 \times 10^{-5}$	0
analytic	$2.970 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$3.033 \times 10^{-2} \pm 4.3 \times 10^{-7}$
mean-field	$9.183 \times 10^{-1} \pm 7.8 \times 10^{-6}$	$1.562 \times 10^0 \pm 1.3 \times 10^{-6}$
linear	$1.915 \times 10^0 \pm 1.1 \times 10^{-5}$	$2.898 \times 10^0 \pm 1.1 \times 10^{-5}$
unscented'95	$2.453 \times 10^{-1} \pm 9.7 \times 10^{-6}$	$1.318 \times 10^{-2} \pm 3.2 \times 10^{-7}$
unscented'02	$9.627 \times 10^0 \pm 9.8 \times 10^{-6}$	$5.129 \times 10^{+1} \pm 1.5 \times 10^{-4}$

Table 17: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=probit), variance=medium

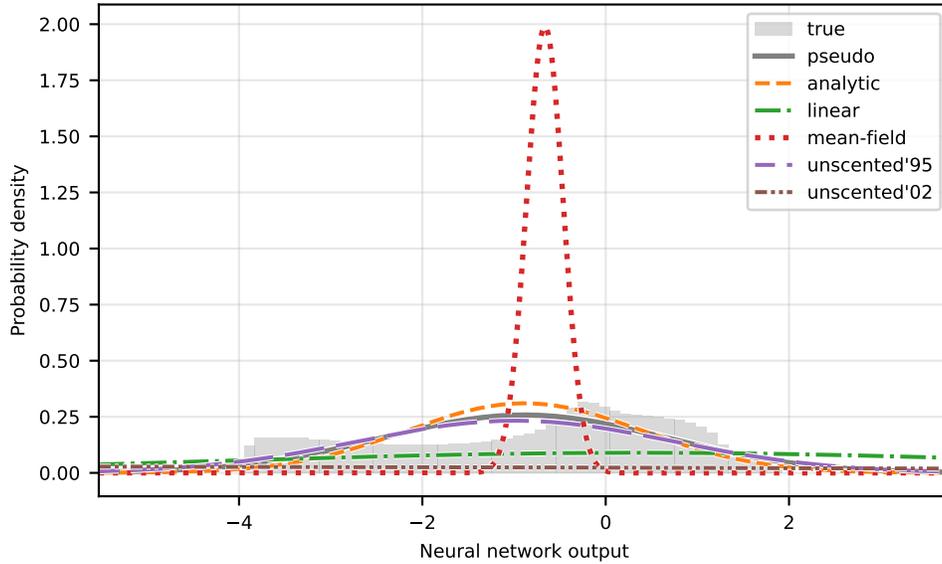


Figure 16: Probability distributions for Network(architecture=wide, weights=trained, activation=probit), variance=medium

3186  
3187  
3188  
3189  
3190  
3191  
3192  
3193  
3194

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-1.376 \times 10^0 \pm 1.6 \times 10^{-5}$	$1.342 \times 10^0 \pm 3.8 \times 10^{-5}$
analytic	$-1.316 \times 10^0$	$9.442 \times 10^{-1}$
mean-field	$-1.109 \times 10^0$	$6.238 \times 10^{-2}$
linear	$+3.554 \times 10^{-1}$	$1.977 \times 10^{+3}$
unscented'95	$-1.393 \times 10^0$	$1.282 \times 10^0$
unscented'02	$-9.306 \times 10^{+2}$	$1.735 \times 10^{+6}$

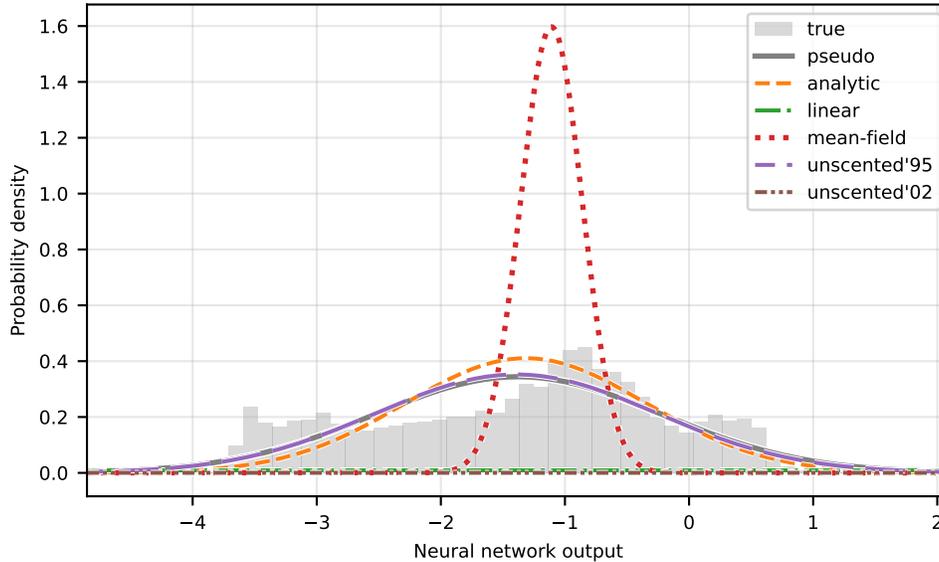
3195 Table 18: Comparison of moments for Network(architecture=wide, weights=trained, activa-  
3196 tion=probit), variance=large  
3197

3198  
3199  
3200  
3201  
3202  
3203  
3204  
3205  
3206  
3207

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.432 \times 10^{-1} \pm 1.5 \times 10^{-5}$	0
analytic	$2.051 \times 10^{-1} \pm 1.4 \times 10^{-5}$	$2.896 \times 10^{-2} \pm 4.2 \times 10^{-6}$
mean-field	$7.035 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$1.084 \times 10^0 \pm 1.3 \times 10^{-5}$
linear	$3.209 \times 10^{+1} \pm 2.4 \times 10^{-4}$	$7.334 \times 10^{+2} \pm 2.1 \times 10^{-2}$
unscented'95	$1.550 \times 10^{-1} \pm 1.4 \times 10^{-5}$	$6.325 \times 10^{-4} \pm 6.7 \times 10^{-7}$
unscented'02	$1.209 \times 10^{+3} \pm 8.5 \times 10^{-3}$	$9.681 \times 10^{+5} \pm 2.7 \times 10^{+1}$

3208 Table 19: Comparison of statistical distances for Network(architecture=wide, weights=trained, ac-  
3209 tivation=probit), variance=large  
3210

3211  
3212  
3213  
3214  
3215  
3216  
3217  
3218  
3219  
3220  
3221  
3222  
3223  
3224  
3225  
3226  
3227



3228  
3229  
3230  
3231  
3232 Figure 17: Probability distributions for Network(architecture=wide, weights=trained, activa-  
3233 tion=probit), variance=large  
3234

3235  
3236  
3237  
3238  
3239

3240  
3241  
3242  
3243  
3244  
3245  
3246  
3247  
3248  
3249  
3250  
3251  
3252  
3253  
3254  
3255  
3256  
3257  
3258  
3259  
3260  
3261  
3262  
3263  
3264  
3265  
3266  
3267  
3268  
3269  
3270  
3271  
3272  
3273  
3274  
3275  
3276  
3277  
3278  
3279  
3280  
3281  
3282  
3283  
3284  
3285  
3286  
3287  
3288  
3289  
3290  
3291  
3292  
3293

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.754 \times 10^{-1} \pm 1.8 \times 10^{-7}$	$4.923 \times 10^{-2} \pm 3.3 \times 10^{-7}$
analytic	$+1.754 \times 10^{-1}$	$4.870 \times 10^{-2}$
mean-field	$+1.775 \times 10^{-1}$	$3.412 \times 10^{-2}$
linear	$+1.858 \times 10^{-1}$	$5.340 \times 10^{-2}$
unscented'95	$+1.749 \times 10^{-1}$	$5.137 \times 10^{-2}$
unscented'02	$+1.746 \times 10^{-1}$	$5.365 \times 10^{-2}$

Table 20: Comparison of moments for Network(architecture=wide, weights=initialized, activation=probit residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.925 \times 10^{-2} \pm 3.7 \times 10^{-6}$	0
analytic	$2.948 \times 10^{-2} \pm 3.1 \times 10^{-6}$	$2.927 \times 10^{-5} \pm 3.6 \times 10^{-8}$
mean-field	$6.750 \times 10^{-2} \pm 3.0 \times 10^{-6}$	$2.988 \times 10^{-2} \pm 1.0 \times 10^{-6}$
linear	$2.734 \times 10^{-2} \pm 3.9 \times 10^{-6}$	$2.807 \times 10^{-3} \pm 2.8 \times 10^{-7}$
unscented'95	$2.957 \times 10^{-2} \pm 2.6 \times 10^{-6}$	$4.640 \times 10^{-4} \pm 1.5 \times 10^{-7}$
unscented'02	$3.159 \times 10^{-2} \pm 4.1 \times 10^{-6}$	$1.908 \times 10^{-3} \pm 3.0 \times 10^{-7}$

Table 21: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=probit residual), variance=small

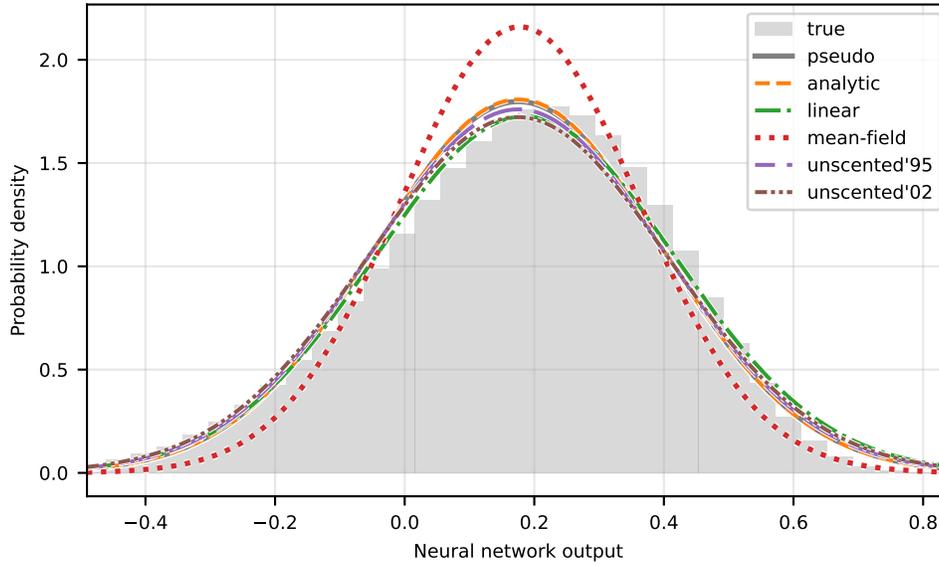


Figure 18: Probability distributions for Network(architecture=wide, weights=initialized, activation=probit residual), variance=small

3294  
3295  
3296  
3297  
3298  
3299  
3300  
3301  
3302  
3303  
3304  
3305  
3306  
3307  
3308  
3309  
3310  
3311  
3312  
3313  
3314  
3315  
3316  
3317  
3318  
3319  
3320  
3321  
3322  
3323  
3324  
3325  
3326  
3327  
3328  
3329  
3330  
3331  
3332  
3333  
3334  
3335  
3336  
3337  
3338  
3339  
3340  
3341  
3342  
3343  
3344  
3345  
3346  
3347

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.952 \times 10^{-2} \pm 3.6 \times 10^{-6}$	$1.045 \times 10^0 \pm 9.2 \times 10^{-6}$
analytic	$-2.266 \times 10^{-2}$	$9.126 \times 10^{-1}$
mean-field	$-6.728 \times 10^{-2}$	$1.105 \times 10^0$
linear	$+1.858 \times 10^{-1}$	$5.340 \times 10^0$
unscented'95	$-1.497 \times 10^{-1}$	$1.106 \times 10^0$
unscented'02	$-9.360 \times 10^{-1}$	$7.856 \times 10^0$

Table 22: Comparison of moments for Network(architecture=wide, weights=initialized, activation=probit residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.107 \times 10^{-1} \pm 1.3 \times 10^{-5}$	0
analytic	$1.225 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$5.062 \times 10^{-3} \pm 5.5 \times 10^{-7}$
mean-field	$1.094 \times 10^{-1} \pm 1.1 \times 10^{-5}$	$8.233 \times 10^{-4} \pm 2.5 \times 10^{-7}$
linear	$1.002 \times 10^0 \pm 9.5 \times 10^{-6}$	$1.267 \times 10^0 \pm 1.9 \times 10^{-5}$
unscented'95	$1.445 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$4.702 \times 10^{-3} \pm 3.9 \times 10^{-7}$
unscented'02	$1.600 \times 10^0 \pm 9.8 \times 10^{-6}$	$2.617 \times 10^0 \pm 3.2 \times 10^{-5}$

Table 23: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=probit residual), variance=medium

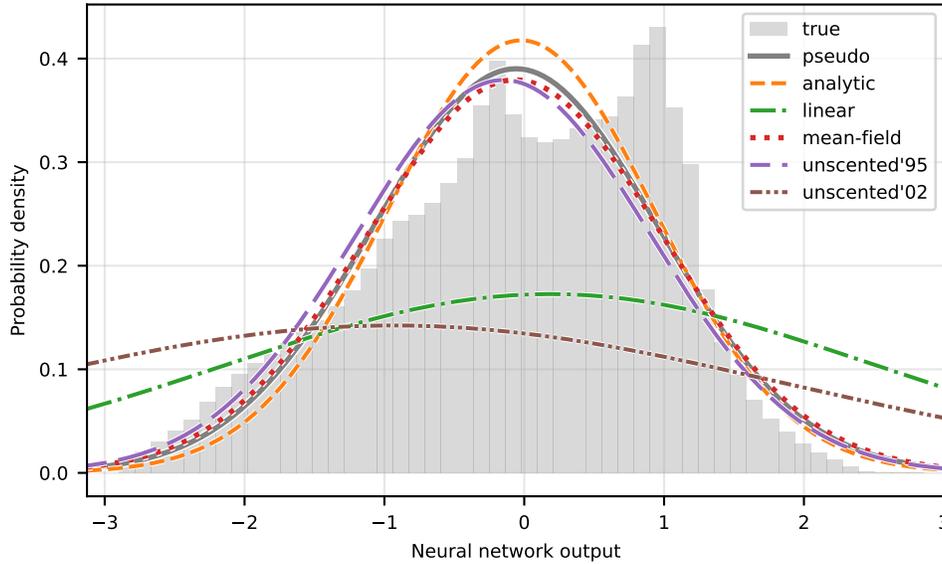


Figure 19: Probability distributions for Network(architecture=wide, weights=initialized, activation=probit residual), variance=medium

3348  
3349  
3350  
3351  
3352  
3353  
3354  
3355  
3356

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+6.429 \times 10^{-3} \pm 3.2 \times 10^{-5}$	$2.400 \times 10^0 \pm 1.8 \times 10^{-4}$
analytic	$+3.071 \times 10^{-2}$	$2.001 \times 10^0$
mean-field	$-1.149 \times 10^{-1}$	$2.268 \times 10^0$
linear	$+1.858 \times 10^{-1}$	$5.340 \times 10^{+2}$
unscented'95	$-2.363 \times 10^{-1}$	$1.624 \times 10^0$
unscented'02	$-1.120 \times 10^{+2}$	$2.569 \times 10^{+4}$

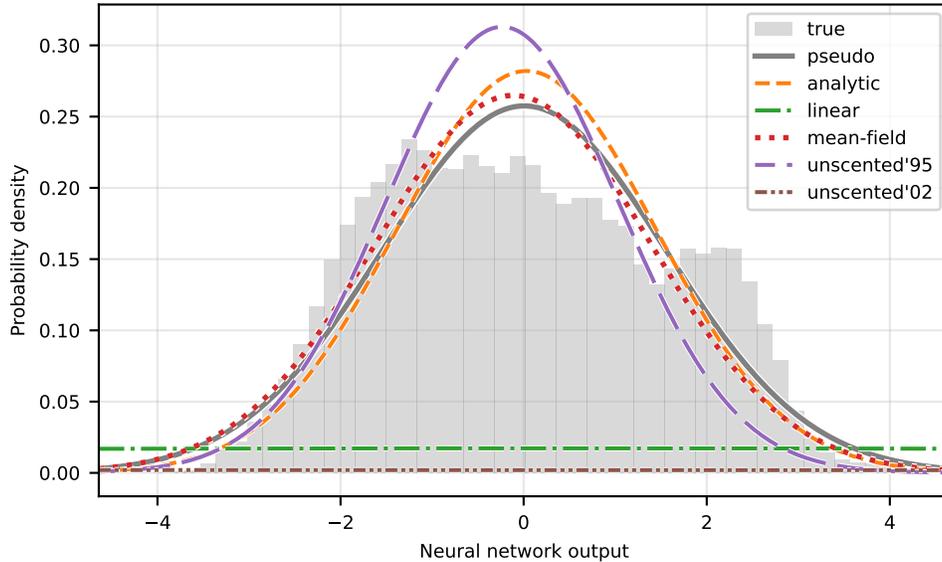
Table 24: Comparison of moments for Network(architecture=wide, weights=initialized, activation=probit residual), variance=large

3360  
3361  
3362  
3363  
3364  
3365  
3366  
3367  
3368  
3369

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.315 \times 10^{-1} \pm 4.2 \times 10^{-5}$	0
analytic	$1.714 \times 10^{-1} \pm 4.4 \times 10^{-5}$	$7.906 \times 10^{-3} \pm 6.3 \times 10^{-6}$
mean-field	$1.460 \times 10^{-1} \pm 4.3 \times 10^{-5}$	$3.846 \times 10^{-3} \pm 1.9 \times 10^{-6}$
linear	$1.377 \times 10^{+1} \pm 3.0 \times 10^{-4}$	$1.081 \times 10^{+2} \pm 8.3 \times 10^{-3}$
unscented'95	$2.530 \times 10^{-1} \pm 4.1 \times 10^{-5}$	$4.584 \times 10^{-2} \pm 1.0 \times 10^{-5}$
unscented'02	$1.260 \times 10^{+2} \pm 2.4 \times 10^{-3}$	$7.959 \times 10^{+3} \pm 6.0 \times 10^{-1}$

Table 25: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=probit residual), variance=large

3370  
3371  
3372  
3373  
3374  
3375  
3376  
3377  
3378  
3379  
3380  
3381  
3382  
3383  
3384  
3385  
3386  
3387  
3388  
3389  
3390  
3391  
3392



3393  
3394  
3395  
3396  
3397  
3398  
3399  
3400  
3401

Figure 20: Probability distributions for Network(architecture=wide, weights=initialized, activation=probit residual), variance=large

3402  
3403  
3404  
3405  
3406  
3407  
3408  
3409  
3410  
3411  
3412  
3413  
3414  
3415  
3416  
3417  
3418  
3419  
3420  
3421  
3422  
3423  
3424  
3425  
3426  
3427  
3428  
3429  
3430  
3431  
3432  
3433  
3434  
3435  
3436  
3437  
3438  
3439  
3440  
3441  
3442  
3443  
3444  
3445  
3446  
3447  
3448  
3449  
3450  
3451  
3452  
3453  
3454  
3455

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.878 \times 10^{-2} \pm 2.5 \times 10^{-7}$	$8.209 \times 10^{-2} \pm 4.9 \times 10^{-7}$
analytic	$-2.882 \times 10^{-2}$	$7.926 \times 10^{-2}$
mean-field	$-6.754 \times 10^{-3}$	$3.377 \times 10^{-2}$
linear	$+2.228 \times 10^{-2}$	$8.666 \times 10^{-2}$
unscented'95	$-3.091 \times 10^{-2}$	$8.257 \times 10^{-2}$
unscented'02	$-3.498 \times 10^{-2}$	$9.322 \times 10^{-2}$

Table 26: Comparison of moments for Network(architecture=wide, weights=trained, activation=probit residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.621 \times 10^{-2} \pm 4.1 \times 10^{-6}$	0
analytic	$3.678 \times 10^{-2} \pm 4.5 \times 10^{-6}$	$3.049 \times 10^{-4} \pm 1.0 \times 10^{-7}$
mean-field	$1.522 \times 10^{-1} \pm 2.4 \times 10^{-6}$	$1.528 \times 10^{-1} \pm 1.7 \times 10^{-6}$
linear	$9.538 \times 10^{-2} \pm 4.8 \times 10^{-7}$	$1.662 \times 10^{-2} \pm 2.9 \times 10^{-7}$
unscented'95	$3.773 \times 10^{-2} \pm 4.5 \times 10^{-6}$	$3.607 \times 10^{-5} \pm 1.9 \times 10^{-8}$
unscented'02	$4.863 \times 10^{-2} \pm 2.3 \times 10^{-6}$	$4.452 \times 10^{-3} \pm 4.1 \times 10^{-7}$

Table 27: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=probit residual), variance=small

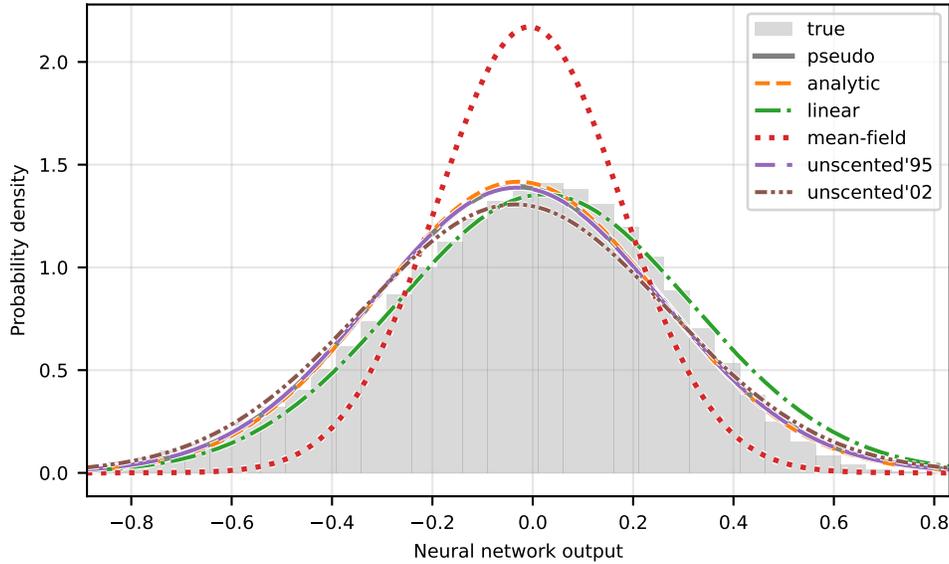


Figure 21: Probability distributions for Network(architecture=wide, weights=trained, activation=probit residual), variance=small

3456  
3457  
3458  
3459  
3460  
3461  
3462  
3463  
3464

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-4.463 \times 10^{-1} \pm 3.6 \times 10^{-6}$	$9.455 \times 10^{-1} \pm 7.6 \times 10^{-6}$
analytic	$-4.242 \times 10^{-1}$	$7.310 \times 10^{-1}$
mean-field	$-4.670 \times 10^{-1}$	$1.116 \times 10^0$
linear	$+2.228 \times 10^{-2}$	$8.666 \times 10^0$
unscented'95	$-4.659 \times 10^{-1}$	$9.297 \times 10^{-1}$
unscented'02	$-5.704 \times 10^0$	$7.424 \times 10^{+1}$

3465  
3466  
3467

Table 28: Comparison of moments for Network(architecture=wide, weights=trained, activation=probit residual), variance=medium

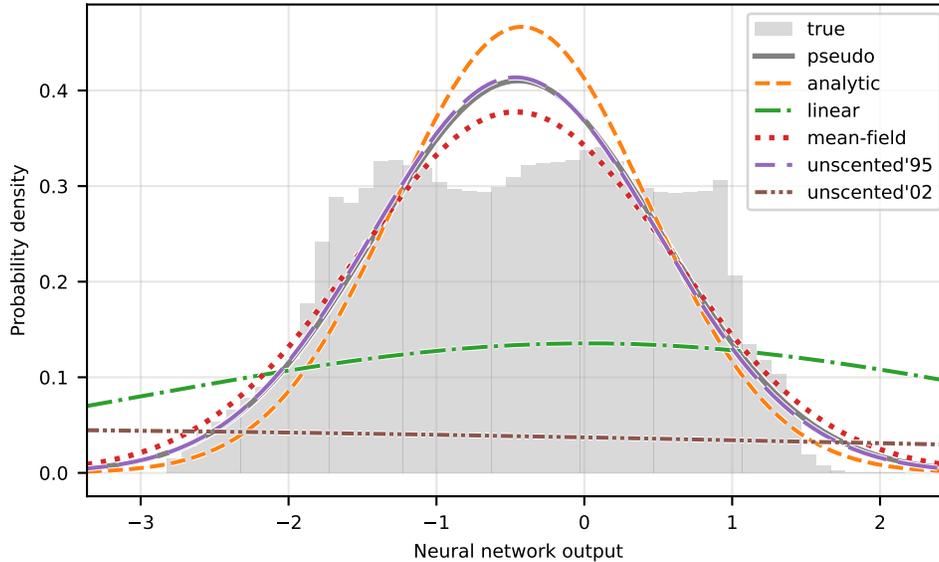
3468  
3469  
3470  
3471  
3472  
3473  
3474  
3475  
3476  
3477

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.057 \times 10^{-1} \pm 8.3 \times 10^{-6}$	0
analytic	$1.598 \times 10^{-1} \pm 1.1 \times 10^{-5}$	$1.548 \times 10^{-2} \pm 9.0 \times 10^{-7}$
mean-field	$1.033 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$7.492 \times 10^{-3} \pm 7.2 \times 10^{-7}$
linear	$1.591 \times 10^0 \pm 1.1 \times 10^{-5}$	$3.091 \times 10^0 \pm 3.4 \times 10^{-5}$
unscented'95	$1.096 \times 10^{-1} \pm 8.9 \times 10^{-6}$	$2.736 \times 10^{-4} \pm 1.1 \times 10^{-7}$
unscented'02	$7.592 \times 10^0 \pm 1.9 \times 10^{-5}$	$5.119 \times 10^{+1} \pm 4.3 \times 10^{-4}$

3478  
3479  
3480

Table 29: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=probit residual), variance=medium

3481  
3482  
3483  
3484  
3485  
3486  
3487  
3488  
3489  
3490  
3491  
3492  
3493  
3494  
3495  
3496  
3497  
3498  
3499



3500  
3501  
3502  
3503  
3504  
3505  
3506  
3507  
3508  
3509

Figure 22: Probability distributions for Network(architecture=wide, weights=trained, activation=probit residual), variance=medium

3510  
3511  
3512  
3513  
3514  
3515  
3516  
3517  
3518  
3519  
3520  
3521  
3522  
3523  
3524  
3525  
3526  
3527  
3528  
3529  
3530  
3531  
3532  
3533  
3534  
3535  
3536  
3537  
3538  
3539  
3540  
3541  
3542  
3543  
3544  
3545  
3546  
3547  
3548  
3549  
3550  
3551  
3552  
3553  
3554  
3555  
3556  
3557  
3558  
3559  
3560  
3561  
3562  
3563

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-3.398 \times 10^{-1} \pm 3.1 \times 10^{-5}$	$1.306 \times 10^0 \pm 1.4 \times 10^{-4}$
analytic	$-3.099 \times 10^{-1}$	$1.119 \times 10^0$
mean-field	$-3.785 \times 10^{-1}$	$2.281 \times 10^0$
linear	$+2.228 \times 10^{-2}$	$8.666 \times 10^{+2}$
unscented'95	$-4.517 \times 10^{-1}$	$1.085 \times 10^0$
unscented'02	$-5.721 \times 10^{+2}$	$6.556 \times 10^{+5}$

Table 30: Comparison of moments for Network(architecture=wide, weights=trained, activation=probit residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.819 \times 10^{-2} \pm 5.0 \times 10^{-5}$	0
analytic	$9.476 \times 10^{-2} \pm 6.3 \times 10^{-5}$	$6.010 \times 10^{-3} \pm 7.7 \times 10^{-6}$
mean-field	$2.552 \times 10^{-1} \pm 6.7 \times 10^{-5}$	$9.511 \times 10^{-2} \pm 4.0 \times 10^{-5}$
linear	$2.110 \times 10^{+1} \pm 6.1 \times 10^{-4}$	$3.282 \times 10^{+2} \pm 3.5 \times 10^{-2}$
unscented'95	$1.301 \times 10^{-1} \pm 4.1 \times 10^{-5}$	$1.283 \times 10^{-2} \pm 8.1 \times 10^{-6}$
unscented'02	$7.484 \times 10^{+2} \pm 2.0 \times 10^{-2}$	$3.763 \times 10^{+5} \pm 4.0 \times 10^{+1}$

Table 31: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=probit residual), variance=large

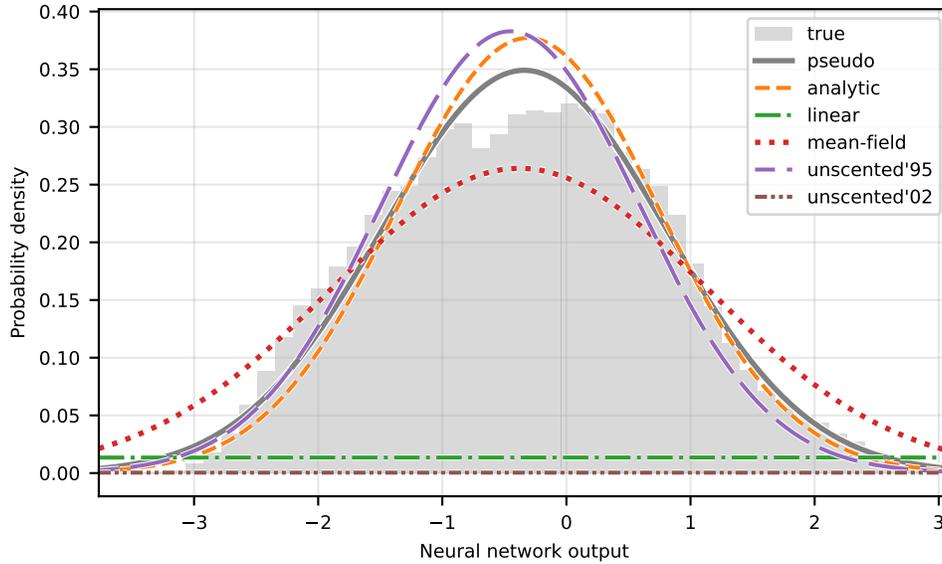


Figure 23: Probability distributions for Network(architecture=wide, weights=trained, activation=probit residual), variance=large

3564  
 3565  
 3566  
 3567  
 3568  
 3569  
 3570  
 3571  
 3572  
 3573  
 3574  
 3575  
 3576  
 3577  
 3578  
 3579  
 3580  
 3581  
 3582  
 3583  
 3584  
 3585  
 3586  
 3587  
 3588  
 3589  
 3590  
 3591  
 3592  
 3593  
 3594  
 3595  
 3596  
 3597  
 3598  
 3599  
 3600  
 3601  
 3602  
 3603  
 3604  
 3605  
 3606  
 3607  
 3608  
 3609  
 3610  
 3611  
 3612  
 3613  
 3614  
 3615  
 3616  
 3617

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-7.232 \times 10^{-1} \pm 1.3 \times 10^{-7}$	$3.095 \times 10^{-3} \pm 8.0 \times 10^{-8}$
analytic	$-7.224 \times 10^{-1}$	$2.803 \times 10^{-3}$
mean-field	$-7.099 \times 10^{-1}$	$8.446 \times 10^{-3}$
linear	$-7.459 \times 10^{-1}$	$2.995 \times 10^{-3}$
unscented'95	$-7.233 \times 10^{-1}$	$2.532 \times 10^{-3}$
unscented'02	$-7.222 \times 10^{-1}$	$4.117 \times 10^{-3}$

Table 32: Comparison of moments for Network(architecture=wide, weights=initialized, activation=sine), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.493 \times 10^{-2} \pm 2.5 \times 10^{-6}$	0
analytic	$2.565 \times 10^{-2} \pm 2.1 \times 10^{-6}$	$2.473 \times 10^{-3} \pm 1.2 \times 10^{-6}$
mean-field	$1.421 \times 10^{-1} \pm 1.9 \times 10^{-6}$	$3.909 \times 10^{-1} \pm 2.4 \times 10^{-5}$
linear	$9.653 \times 10^{-2} \pm 3.5 \times 10^{-7}$	$8.401 \times 10^{-2} \pm 1.1 \times 10^{-6}$
unscented'95	$2.625 \times 10^{-2} \pm 1.7 \times 10^{-6}$	$9.448 \times 10^{-3} \pm 2.4 \times 10^{-6}$
unscented'02	$4.092 \times 10^{-2} \pm 2.4 \times 10^{-6}$	$2.256 \times 10^{-2} \pm 4.3 \times 10^{-6}$

Table 33: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=sine), variance=small

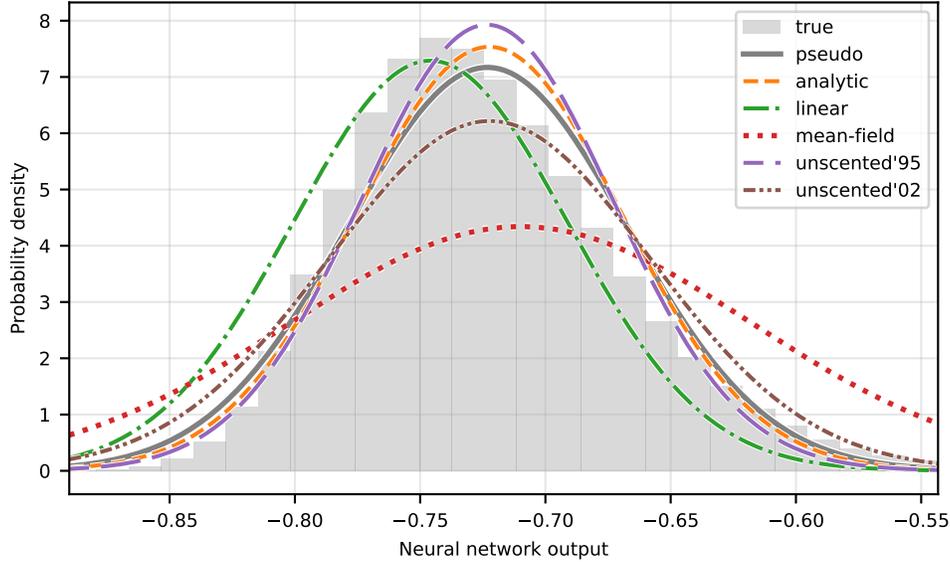


Figure 24: Probability distributions for Network(architecture=wide, weights=initialized, activation=sine), variance=small

3618  
3619  
3620  
3621  
3622  
3623  
3624  
3625  
3626  
3627  
3628  
3629  
3630  
3631  
3632  
3633  
3634  
3635  
3636  
3637  
3638  
3639  
3640  
3641  
3642  
3643  
3644  
3645  
3646  
3647  
3648  
3649  
3650  
3651  
3652  
3653  
3654  
3655  
3656  
3657  
3658  
3659  
3660  
3661  
3662  
3663  
3664  
3665  
3666  
3667  
3668  
3669  
3670  
3671

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-1.486 \times 10^{-1} \pm 7.9 \times 10^{-6}$	$1.199 \times 10^{-1} \pm 1.1 \times 10^{-5}$
analytic	$-9.940 \times 10^{-2}$	$1.031 \times 10^{-1}$
mean-field	$-2.147 \times 10^{-2}$	$1.125 \times 10^{-1}$
linear	$-7.459 \times 10^{-1}$	$2.995 \times 10^{-1}$
unscented'95	$-1.529 \times 10^{-1}$	$9.089 \times 10^{-2}$
unscented'02	$+1.623 \times 10^0$	$1.152 \times 10^{+1}$

Table 34: Comparison of moments for Network(architecture=wide, weights=initialized, activation=sine), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.159 \times 10^{-2} \pm 2.3 \times 10^{-5}$	0
analytic	$9.243 \times 10^{-2} \pm 1.9 \times 10^{-5}$	$1.553 \times 10^{-2} \pm 4.6 \times 10^{-6}$
mean-field	$2.164 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$6.839 \times 10^{-2} \pm 1.1 \times 10^{-5}$
linear	$1.015 \times 10^0 \pm 2.1 \times 10^{-5}$	$1.779 \times 10^0 \pm 1.8 \times 10^{-4}$
unscented'95	$5.197 \times 10^{-2} \pm 1.6 \times 10^{-5}$	$1.763 \times 10^{-2} \pm 1.1 \times 10^{-5}$
unscented'02	$4.814 \times 10^0 \pm 1.2 \times 10^{-4}$	$5.832 \times 10^{+1} \pm 5.6 \times 10^{-3}$

Table 35: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=sine), variance=medium

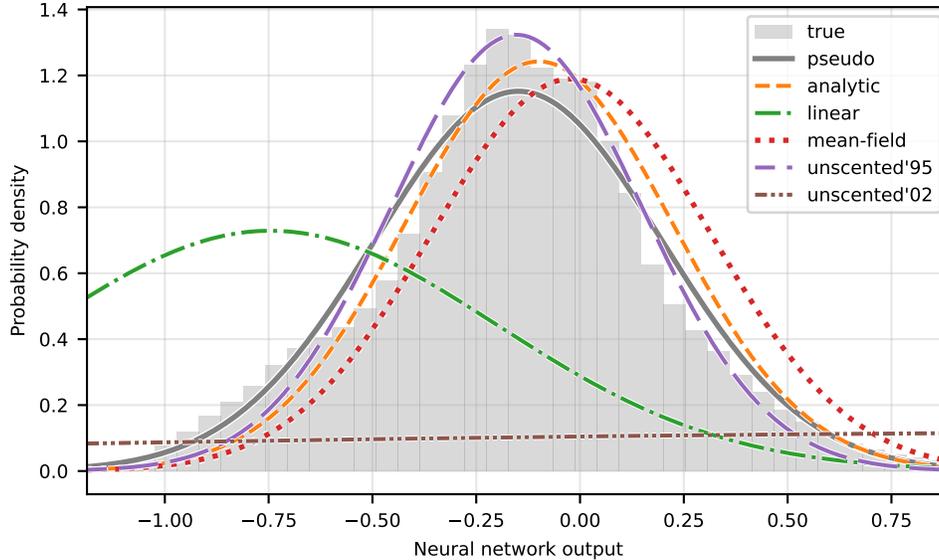


Figure 25: Probability distributions for Network(architecture=wide, weights=initialized, activation=sine), variance=medium

3672  
3673  
3674  
3675  
3676  
3677  
3678  
3679  
3680  
3681  
3682  
3683  
3684  
3685  
3686  
3687  
3688  
3689  
3690  
3691  
3692  
3693  
3694  
3695  
3696  
3697  
3698  
3699  
3700  
3701  
3702  
3703  
3704  
3705  
3706  
3707  
3708  
3709  
3710  
3711  
3712  
3713  
3714  
3715  
3716  
3717  
3718  
3719  
3720  
3721  
3722  
3723  
3724  
3725

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.423 \times 10^{-1} \pm 1.5 \times 10^{-4}$	$1.334 \times 10^{-1} \pm 1.1 \times 10^{-4}$
analytic	$+2.380 \times 10^{-1}$	$1.341 \times 10^{-1}$
mean-field	$+2.362 \times 10^{-1}$	$1.236 \times 10^{-1}$
linear	$-7.459 \times 10^{-1}$	$2.995 \times 10^{+1}$
unscented'95	$+2.682 \times 10^{-1}$	$9.459 \times 10^{-2}$
unscented'02	$+2.360 \times 10^{+2}$	$1.121 \times 10^{+5}$

Table 36: Comparison of moments for Network(architecture=wide, weights=initialized, activation=sine), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.009 \times 10^{-3} \pm 1.3 \times 10^{-4}$	0
analytic	$7.821 \times 10^{-3} \pm 2.0 \times 10^{-4}$	$8.083 \times 10^{-5} \pm 4.8 \times 10^{-6}$
mean-field	$1.992 \times 10^{-2} \pm 2.1 \times 10^{-4}$	$1.565 \times 10^{-3} \pm 3.1 \times 10^{-5}$
linear	$6.870 \times 10^0 \pm 1.5 \times 10^{-3}$	$1.127 \times 10^{+2} \pm 9.1 \times 10^{-2}$
unscented'95	$8.243 \times 10^{-2} \pm 1.7 \times 10^{-4}$	$2.898 \times 10^{-2} \pm 1.1 \times 10^{-4}$
unscented'02	$5.469 \times 10^{+2} \pm 1.1 \times 10^{-1}$	$6.284 \times 10^{+5} \pm 5.0 \times 10^{+2}$

Table 37: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=sine), variance=large

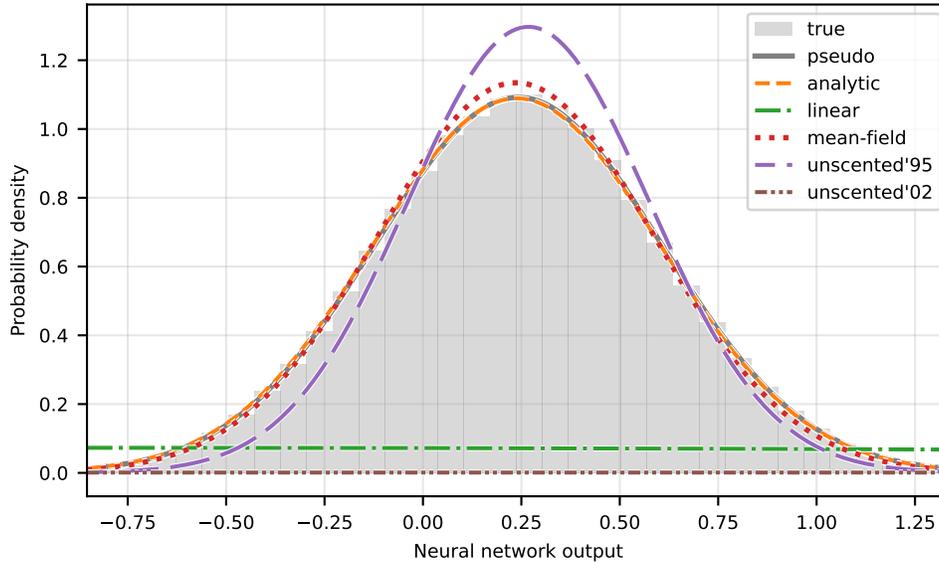


Figure 26: Probability distributions for Network(architecture=wide, weights=initialized, activation=sine), variance=large

3726  
3727  
3728  
3729  
3730  
3731  
3732  
3733  
3734  
3735  
3736  
3737  
3738  
3739  
3740  
3741  
3742  
3743  
3744  
3745  
3746  
3747  
3748  
3749  
3750  
3751  
3752  
3753  
3754  
3755  
3756  
3757  
3758  
3759  
3760  
3761  
3762  
3763  
3764  
3765  
3766  
3767  
3768  
3769  
3770  
3771  
3772  
3773  
3774  
3775  
3776  
3777  
3778  
3779

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-9.661 \times 10^{-2} \pm 2.2 \times 10^{-7}$	$4.201 \times 10^{-2} \pm 2.9 \times 10^{-7}$
analytic	$-9.662 \times 10^{-2}$	$3.998 \times 10^{-2}$
mean-field	$-5.941 \times 10^{-2}$	$8.820 \times 10^{-3}$
linear	$-3.517 \times 10^{-2}$	$4.359 \times 10^{-2}$
unscented'95	$-9.971 \times 10^{-2}$	$3.980 \times 10^{-2}$
unscented'02	$-1.027 \times 10^{-1}$	$5.272 \times 10^{-2}$

Table 38: Comparison of moments for Network(architecture=wide, weights=trained, activation=sine), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.363 \times 10^{-2} \pm 3.4 \times 10^{-6}$	0
analytic	$5.428 \times 10^{-2} \pm 3.3 \times 10^{-6}$	$5.980 \times 10^{-4} \pm 1.6 \times 10^{-7}$
mean-field	$1.947 \times 10^{-1} \pm 3.3 \times 10^{-6}$	$4.018 \times 10^{-1} \pm 2.7 \times 10^{-6}$
linear	$1.357 \times 10^{-1} \pm 4.7 \times 10^{-7}$	$4.527 \times 10^{-2} \pm 4.6 \times 10^{-7}$
unscented'95	$5.685 \times 10^{-2} \pm 3.2 \times 10^{-6}$	$8.258 \times 10^{-4} \pm 1.7 \times 10^{-7}$
unscented'02	$7.128 \times 10^{-2} \pm 3.3 \times 10^{-6}$	$1.438 \times 10^{-2} \pm 8.9 \times 10^{-7}$

Table 39: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=sine), variance=small

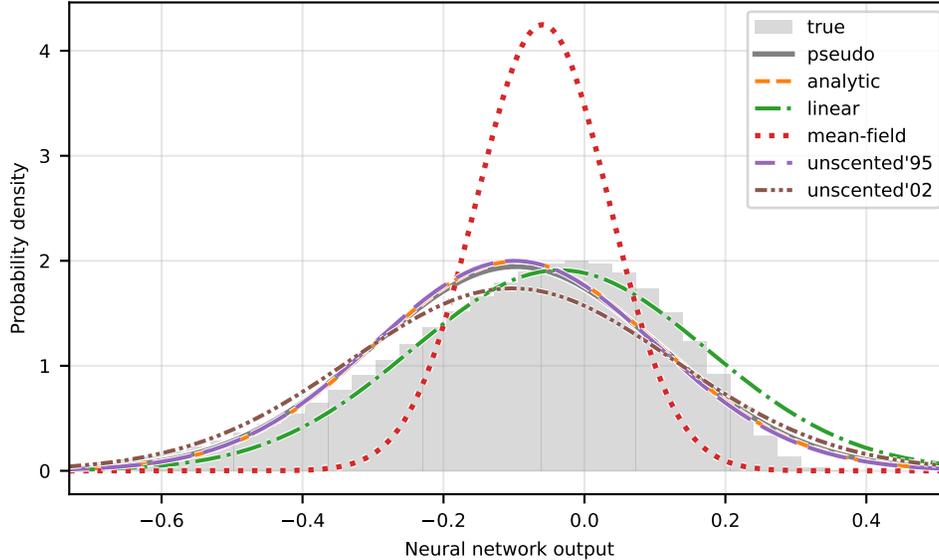


Figure 27: Probability distributions for Network(architecture=wide, weights=trained, activation=sine), variance=small

3780  
3781  
3782  
3783  
3784  
3785  
3786  
3787  
3788

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-1.067 \times 10^{-1} \pm 7.6 \times 10^{-6}$	$2.428 \times 10^{-1} \pm 1.0 \times 10^{-5}$
analytic	$-6.853 \times 10^{-2}$	$1.456 \times 10^{-1}$
mean-field	$+8.272 \times 10^{-3}$	$1.152 \times 10^{-1}$
linear	$-3.517 \times 10^{-2}$	$4.359 \times 10^0$
unscented'95	$-7.132 \times 10^{-2}$	$1.698 \times 10^{-1}$
unscented'02	$-6.790 \times 10^0$	$9.563 \times 10^{+1}$

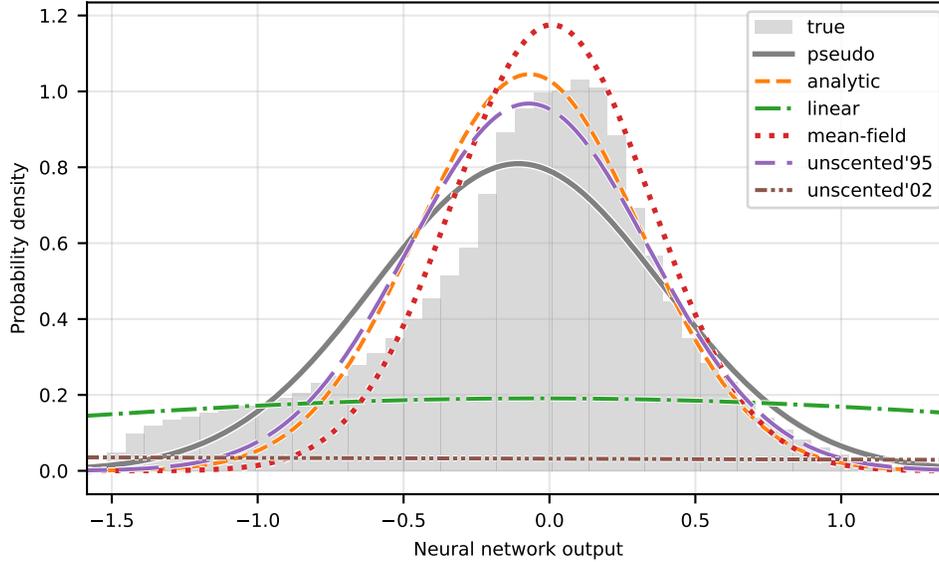
3789 Table 40: Comparison of moments for Network(architecture=wide, weights=trained, activa-  
3790 tion=sine), variance=medium  
3791

3792  
3793  
3794  
3795  
3796  
3797  
3798  
3799  
3800  
3801

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.000 \times 10^{-1} \pm 2.2 \times 10^{-5}$	0
analytic	$1.153 \times 10^{-1} \pm 1.2 \times 10^{-5}$	$5.852 \times 10^{-2} \pm 8.0 \times 10^{-6}$
mean-field	$1.785 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$1.374 \times 10^{-1} \pm 9.0 \times 10^{-6}$
linear	$1.837 \times 10^0 \pm 3.3 \times 10^{-5}$	$7.042 \times 10^0 \pm 3.7 \times 10^{-4}$
unscented'95	$9.049 \times 10^{-2} \pm 1.7 \times 10^{-5}$	$3.107 \times 10^{-2} \pm 5.9 \times 10^{-6}$
unscented'02	$1.322 \times 10^{+1} \pm 1.6 \times 10^{-4}$	$2.854 \times 10^{+2} \pm 1.2 \times 10^{-2}$

3802 Table 41: Comparison of statistical distances for Network(architecture=wide, weights=trained, ac-  
3803 tivation=sine), variance=medium  
3804

3805  
3806  
3807  
3808  
3809  
3810  
3811  
3812  
3813  
3814  
3815  
3816  
3817  
3818  
3819  
3820  
3821  
3822  
3823  
3824



3825  
3826 Figure 28: Probability distributions for Network(architecture=wide, weights=trained, activa-  
3827 tion=sine), variance=medium  
3828

3829  
3830  
3831  
3832  
3833

3834  
3835  
3836  
3837  
3838  
3839  
3840  
3841  
3842  
3843  
3844  
3845  
3846  
3847  
3848  
3849  
3850  
3851  
3852  
3853  
3854  
3855  
3856  
3857  
3858  
3859  
3860  
3861  
3862  
3863  
3864  
3865  
3866  
3867  
3868  
3869  
3870  
3871  
3872  
3873  
3874  
3875  
3876  
3877  
3878  
3879  
3880  
3881  
3882  
3883  
3884  
3885  
3886  
3887

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.743 \times 10^{-1} \pm 1.4 \times 10^{-4}$	$1.397 \times 10^{-1} \pm 1.1 \times 10^{-4}$
analytic	$+2.702 \times 10^{-1}$	$1.408 \times 10^{-1}$
mean-field	$+2.699 \times 10^{-1}$	$1.249 \times 10^{-1}$
linear	$-3.517 \times 10^{-2}$	$4.359 \times 10^{+2}$
unscented'95	$+3.025 \times 10^{-1}$	$1.039 \times 10^{-1}$
unscented'02	$-6.753 \times 10^{+2}$	$9.123 \times 10^{+5}$

Table 42: Comparison of moments for Network(architecture=wide, weights=trained, activation=sine), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.929 \times 10^{-3} \pm 1.1 \times 10^{-4}$	0
analytic	$7.570 \times 10^{-3} \pm 1.8 \times 10^{-4}$	$8.025 \times 10^{-5} \pm 4.3 \times 10^{-6}$
mean-field	$2.646 \times 10^{-2} \pm 2.2 \times 10^{-4}$	$3.105 \times 10^{-3} \pm 4.3 \times 10^{-5}$
linear	$2.676 \times 10^{+1} \pm 5.4 \times 10^{-3}$	$1.556 \times 10^{+3} \pm 1.2 \times 10^0$
unscented'95	$7.597 \times 10^{-2} \pm 1.6 \times 10^{-4}$	$2.281 \times 10^{-2} \pm 9.4 \times 10^{-5}$
unscented'02	$1.546 \times 10^{+3} \pm 3.0 \times 10^{-1}$	$4.898 \times 10^{+6} \pm 3.9 \times 10^{+3}$

Table 43: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=sine), variance=large

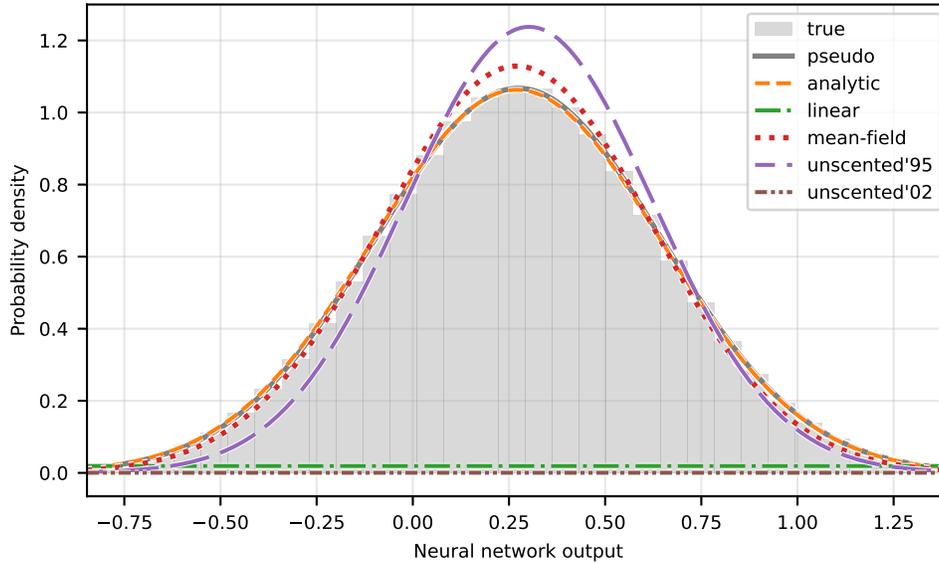


Figure 29: Probability distributions for Network(architecture=wide, weights=trained, activation=sine), variance=large

3888  
3889  
3890  
3891  
3892  
3893  
3894  
3895  
3896  
3897  
3898  
3899  
3900  
3901  
3902  
3903  
3904  
3905  
3906  
3907  
3908  
3909  
3910  
3911  
3912  
3913  
3914  
3915  
3916  
3917  
3918  
3919  
3920  
3921  
3922  
3923  
3924  
3925  
3926  
3927  
3928  
3929  
3930  
3931  
3932  
3933  
3934  
3935  
3936  
3937  
3938  
3939  
3940  
3941

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.778 \times 10^0 \pm 5.5 \times 10^{-7}$	$5.871 \times 10^{-2} \pm 6.9 \times 10^{-7}$
analytic	$+1.774 \times 10^0$	$7.277 \times 10^{-2}$
mean-field	$+1.646 \times 10^0$	$1.393 \times 10^{-1}$
linear	$+1.677 \times 10^0$	$7.931 \times 10^{-2}$
unscented'95	$+1.789 \times 10^0$	$5.573 \times 10^{-2}$
unscented'02	$+1.788 \times 10^0$	$1.040 \times 10^{-1}$

Table 44: Comparison of moments for Network(architecture=wide, weights=initialized, activation=sine residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$8.017 \times 10^{-2} \pm 3.5 \times 10^{-6}$	0
analytic	$7.802 \times 10^{-2} \pm 4.4 \times 10^{-6}$	$1.251 \times 10^{-2} \pm 1.4 \times 10^{-6}$
mean-field	$2.836 \times 10^{-1} \pm 4.0 \times 10^{-6}$	$4.026 \times 10^{-1} \pm 9.1 \times 10^{-6}$
linear	$2.043 \times 10^{-1} \pm 8.9 \times 10^{-7}$	$1.112 \times 10^{-1} \pm 2.6 \times 10^{-6}$
unscented'95	$8.962 \times 10^{-2} \pm 3.5 \times 10^{-6}$	$1.772 \times 10^{-3} \pm 2.4 \times 10^{-7}$
unscented'02	$1.538 \times 10^{-1} \pm 2.7 \times 10^{-6}$	$1.008 \times 10^{-1} \pm 4.6 \times 10^{-6}$

Table 45: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=sine residual), variance=small

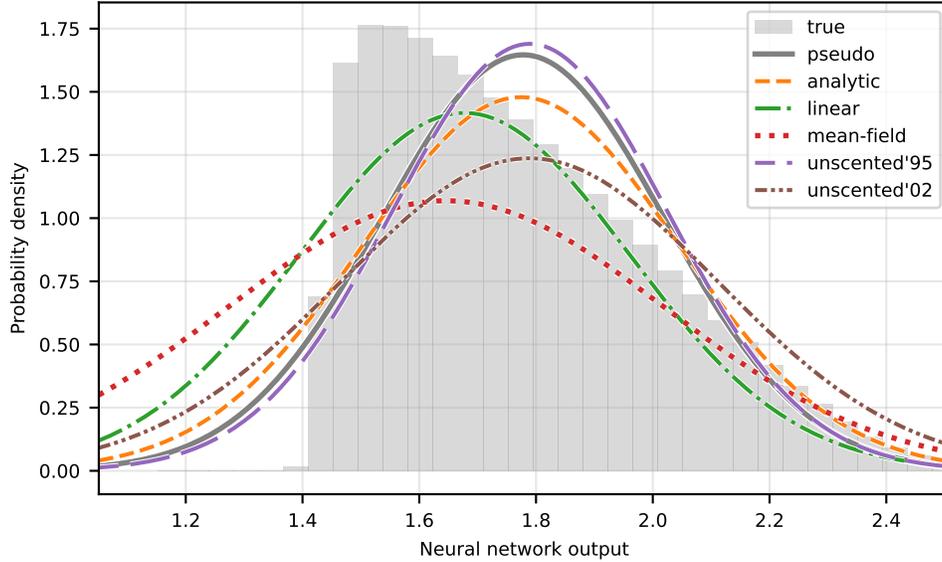


Figure 30: Probability distributions for Network(architecture=wide, weights=initialized, activation=sine residual), variance=small

3942  
3943  
3944  
3945  
3946  
3947  
3948  
3949  
3950

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+7.436 \times 10^{-1} \pm 4.8 \times 10^{-5}$	$2.285 \times 10^0 \pm 1.8 \times 10^{-4}$
analytic	$+5.891 \times 10^{-1}$	$1.899 \times 10^0$
mean-field	$+3.448 \times 10^{-1}$	$1.628 \times 10^0$
linear	$+1.677 \times 10^0$	$7.931 \times 10^0$
unscented'95	$+2.881 \times 10^{-1}$	$3.174 \times 10^0$
unscented'02	$+1.279 \times 10^{+1}$	$2.548 \times 10^{+2}$

3951  
3952  
3953

Table 46: Comparison of moments for Network(architecture=wide, weights=initialized, activation=sine residual), variance=medium

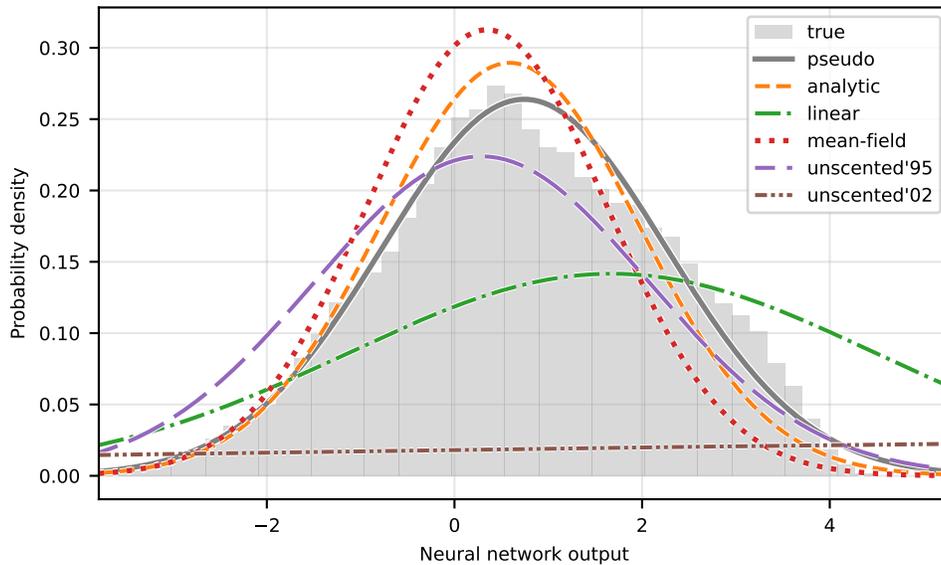
3954  
3955  
3956

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.599 \times 10^{-2} \pm 4.1 \times 10^{-5}$	0
analytic	$1.512 \times 10^{-1} \pm 3.7 \times 10^{-5}$	$1.326 \times 10^{-2} \pm 4.5 \times 10^{-6}$
mean-field	$3.332 \times 10^{-1} \pm 4.8 \times 10^{-5}$	$6.053 \times 10^{-2} \pm 6.2 \times 10^{-6}$
linear	$1.045 \times 10^0 \pm 6.2 \times 10^{-5}$	$8.041 \times 10^{-1} \pm 1.0 \times 10^{-4}$
unscented'95	$4.032 \times 10^{-1} \pm 4.7 \times 10^{-5}$	$7.563 \times 10^{-2} \pm 2.7 \times 10^{-5}$
unscented'02	$1.245 \times 10^{+1} \pm 2.8 \times 10^{-4}$	$8.465 \times 10^{+1} \pm 6.8 \times 10^{-3}$

3964  
3965  
3966

Table 47: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=sine residual), variance=medium

3967  
3968  
3969



3988  
3989

Figure 31: Probability distributions for Network(architecture=wide, weights=initialized, activation=sine residual), variance=medium

3990  
3991  
3992  
3993  
3994  
3995

3996  
3997  
3998  
3999  
4000  
4001  
4002  
4003  
4004  
4005  
4006  
4007  
4008  
4009  
4010  
4011  
4012  
4013  
4014  
4015  
4016  
4017  
4018  
4019  
4020  
4021  
4022  
4023  
4024  
4025  
4026  
4027  
4028  
4029  
4030  
4031  
4032  
4033  
4034  
4035  
4036  
4037  
4038  
4039  
4040  
4041  
4042  
4043  
4044  
4045  
4046  
4047  
4048  
4049

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.654 \times 10^{-2} \pm 5.7 \times 10^{-4}$	$1.868 \times 10^0 \pm 1.2 \times 10^{-3}$
analytic	$+5.135 \times 10^{-2}$	$1.862 \times 10^0$
mean-field	$+4.378 \times 10^{-2}$	$1.862 \times 10^0$
linear	$+1.677 \times 10^0$	$7.931 \times 10^{+2}$
unscented'95	$+1.063 \times 10^0$	$2.373 \times 10^0$
unscented'02	$+1.113 \times 10^{+3}$	$2.472 \times 10^{+6}$

Table 48: Comparison of moments for Network(architecture=wide, weights=initialized, activation=sine residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.228 \times 10^{-3} \pm 2.2 \times 10^{-4}$	0
analytic	$6.261 \times 10^{-3} \pm 3.1 \times 10^{-4}$	$1.346 \times 10^{-5} \pm 2.1 \times 10^{-6}$
mean-field	$1.123 \times 10^{-2} \pm 4.7 \times 10^{-4}$	$4.987 \times 10^{-5} \pm 4.2 \times 10^{-6}$
linear	$1.832 \times 10^{+1} \pm 3.2 \times 10^{-3}$	$2.095 \times 10^{+2} \pm 1.4 \times 10^{-1}$
unscented'95	$8.609 \times 10^{-1} \pm 5.4 \times 10^{-4}$	$2.867 \times 10^{-1} \pm 4.5 \times 10^{-4}$
unscented'02	$1.330 \times 10^{+3} \pm 2.1 \times 10^{-1}$	$9.932 \times 10^{+5} \pm 6.3 \times 10^{+2}$

Table 49: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=sine residual), variance=large

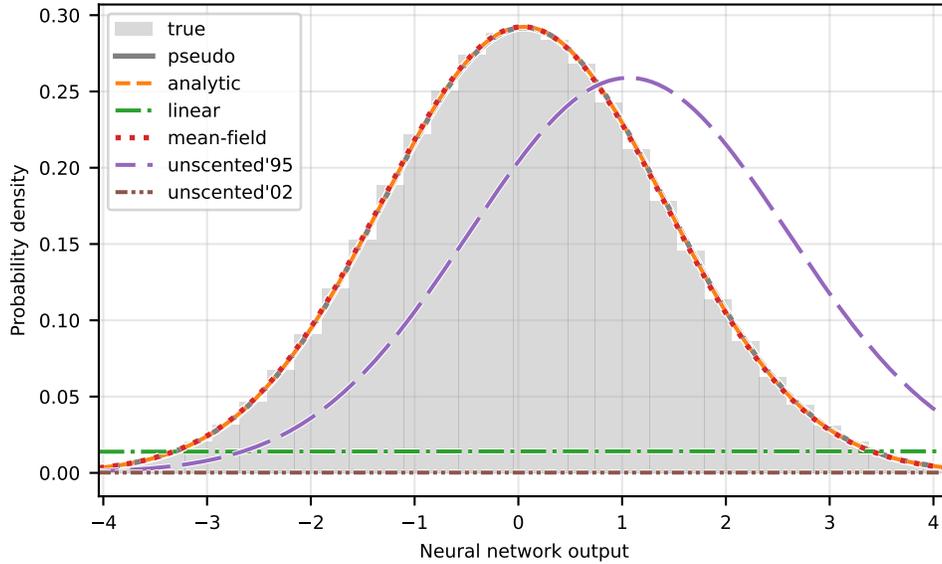


Figure 32: Probability distributions for Network(architecture=wide, weights=initialized, activation=sine residual), variance=large

4050  
4051  
4052  
4053  
4054  
4055  
4056  
4057  
4058

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+3.352 \times 10^{-1} \pm 5.0 \times 10^{-7}$	$1.858 \times 10^{-1} \pm 1.4 \times 10^{-6}$
analytic	$+3.330 \times 10^{-1}$	$1.887 \times 10^{-1}$
mean-field	$+3.550 \times 10^{-1}$	$1.432 \times 10^{-1}$
linear	$+3.526 \times 10^{-1}$	$2.274 \times 10^{-1}$
unscented'95	$+3.373 \times 10^{-1}$	$1.805 \times 10^{-1}$
unscented'02	$+3.176 \times 10^{-1}$	$2.299 \times 10^{-1}$

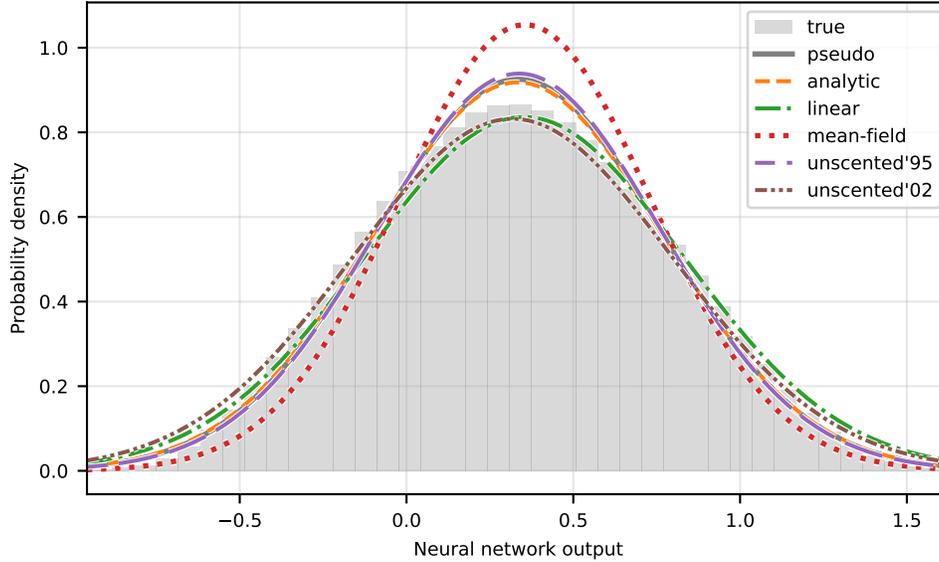
4059 Table 50: Comparison of moments for Network(architecture=wide, weights=trained, activation=sine  
4060 residual), variance=small  
4061

4062  
4063  
4064  
4065  
4066  
4067  
4068  
4069  
4070  
4071

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.299 \times 10^{-2} \pm 4.6 \times 10^{-6}$	0
analytic	$2.097 \times 10^{-2} \pm 4.2 \times 10^{-6}$	$7.531 \times 10^{-5} \pm 6.1 \times 10^{-8}$
mean-field	$7.953 \times 10^{-2} \pm 3.6 \times 10^{-6}$	$1.664 \times 10^{-2} \pm 8.6 \times 10^{-7}$
linear	$5.630 \times 10^{-2} \pm 3.5 \times 10^{-6}$	$1.175 \times 10^{-2} \pm 8.3 \times 10^{-7}$
unscented'95	$2.748 \times 10^{-2} \pm 4.4 \times 10^{-6}$	$2.177 \times 10^{-4} \pm 1.1 \times 10^{-7}$
unscented'02	$5.316 \times 10^{-2} \pm 2.7 \times 10^{-6}$	$1.302 \times 10^{-2} \pm 9.1 \times 10^{-7}$

4072 Table 51: Comparison of statistical distances for Network(architecture=wide, weights=trained, ac-  
4073 tivation=sine residual), variance=small  
4074

4075  
4076  
4077  
4078  
4079  
4080  
4081  
4082  
4083  
4084  
4085  
4086  
4087  
4088  
4089  
4090  
4091  
4092  
4093  
4094



4095  
4096 Figure 33: Probability distributions for Network(architecture=wide, weights=trained, activa-  
4097 tion=sine residual), variance=small  
4098

4099  
4100  
4101  
4102  
4103

4104  
4105  
4106  
4107  
4108  
4109  
4110  
4111  
4112

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.823 \times 10^{-1} \pm 4.7 \times 10^{-5}$	$1.672 \times 10^0 \pm 1.5 \times 10^{-4}$
analytic	$+2.142 \times 10^{-1}$	$1.758 \times 10^0$
mean-field	$+7.102 \times 10^{-2}$	$1.627 \times 10^0$
linear	$+3.526 \times 10^{-1}$	$2.274 \times 10^{+1}$
unscented'95	$+7.597 \times 10^{-2}$	$3.013 \times 10^0$
unscented'02	$-3.146 \times 10^0$	$4.722 \times 10^{+1}$

Table 52: Comparison of moments for Network(architecture=wide, weights=trained, activation=sine residual), variance=medium

4113  
4114  
4115  
4116  
4117  
4118  
4119  
4120  
4121  
4122  
4123  
4124  
4125

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.071 \times 10^{-2} \pm 6.4 \times 10^{-5}$	0
analytic	$5.506 \times 10^{-2} \pm 4.5 \times 10^{-5}$	$9.444 \times 10^{-4} \pm 1.9 \times 10^{-6}$
mean-field	$1.046 \times 10^{-1} \pm 4.4 \times 10^{-5}$	$3.885 \times 10^{-3} \pm 2.7 \times 10^{-6}$
linear	$2.445 \times 10^0 \pm 9.5 \times 10^{-5}$	$5.005 \times 10^0 \pm 5.5 \times 10^{-4}$
unscented'95	$3.110 \times 10^{-1} \pm 5.1 \times 10^{-5}$	$1.099 \times 10^{-1} \pm 3.8 \times 10^{-5}$
unscented'02	$4.571 \times 10^0 \pm 1.5 \times 10^{-4}$	$1.526 \times 10^{+1} \pm 1.6 \times 10^{-3}$

Table 53: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=sine residual), variance=medium

4126  
4127  
4128  
4129  
4130  
4131  
4132  
4133  
4134  
4135  
4136  
4137  
4138  
4139  
4140  
4141  
4142  
4143  
4144  
4145  
4146  
4147

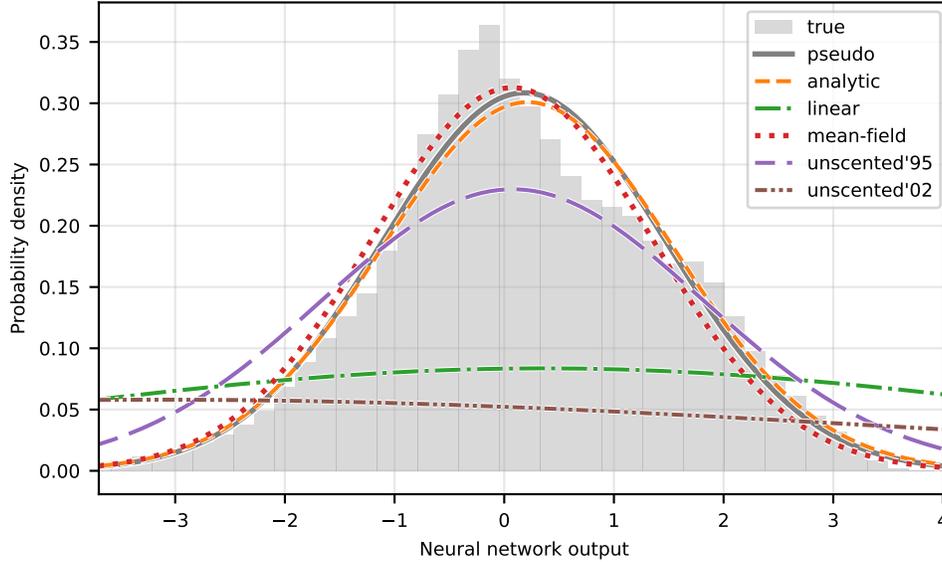


Figure 34: Probability distributions for Network(architecture=wide, weights=trained, activation=sine residual), variance=medium

4152  
4153  
4154  
4155  
4156  
4157

4158  
4159  
4160  
4161  
4162  
4163  
4164  
4165  
4166

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.018 \times 10^{-2} \pm 5.7 \times 10^{-4}$	$1.859 \times 10^0 \pm 1.2 \times 10^{-3}$
analytic	$+5.144 \times 10^{-3}$	$1.855 \times 10^0$
mean-field	$-9.526 \times 10^{-4}$	$1.862 \times 10^0$
linear	$+3.526 \times 10^{-1}$	$2.274 \times 10^{+3}$
unscented'95	$+1.029 \times 10^0$	$2.357 \times 10^0$
unscented'02	$-3.472 \times 10^{+2}$	$2.438 \times 10^{+5}$

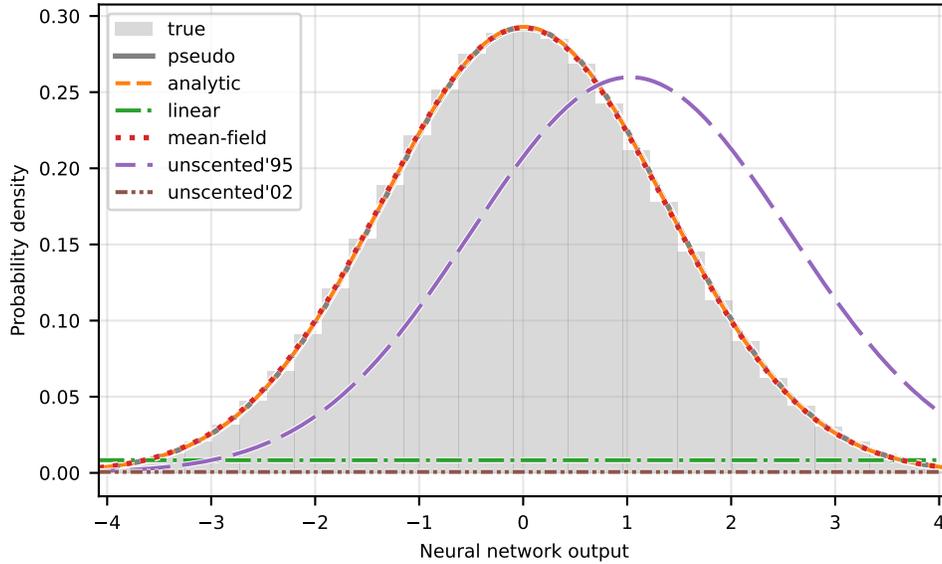
Table 54: Comparison of moments for Network(architecture=wide, weights=trained, activation=sine residual), variance=large

4170  
4171  
4172  
4173  
4174  
4175  
4176  
4177  
4178  
4179

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.921 \times 10^{-3} \pm 2.1 \times 10^{-4}$	0
analytic	$6.028 \times 10^{-3} \pm 3.1 \times 10^{-4}$	$1.203 \times 10^{-5} \pm 1.9 \times 10^{-6}$
mean-field	$9.905 \times 10^{-3} \pm 4.6 \times 10^{-4}$	$3.769 \times 10^{-5} \pm 3.4 \times 10^{-6}$
linear	$3.165 \times 10^{+1} \pm 5.4 \times 10^{-3}$	$6.075 \times 10^{+2} \pm 3.9 \times 10^{-1}$
unscented'95	$8.729 \times 10^{-1} \pm 5.4 \times 10^{-4}$	$2.946 \times 10^{-1} \pm 4.6 \times 10^{-4}$
unscented'02	$4.168 \times 10^{+2} \pm 6.7 \times 10^{-2}$	$9.798 \times 10^{+4} \pm 6.3 \times 10^{+1}$

Table 55: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=sine residual), variance=large

4180  
4181  
4182  
4183  
4184  
4185  
4186  
4187  
4188  
4189  
4190  
4191  
4192  
4193  
4194  
4195  
4196  
4197  
4198  
4199



4200  
4201  
4202  
4203  
4204  
4205  
4206  
4207  
4208  
4209  
4210  
4211

Figure 35: Probability distributions for Network(architecture=wide, weights=trained, activation=sine residual), variance=large

4212  
4213  
4214  
4215  
4216  
4217  
4218  
4219  
4220

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.777 \times 10^0 \pm 8.9 \times 10^{-8}$	$2.682 \times 10^{-2} \pm 1.0 \times 10^{-7}$
analytic	$-2.777 \times 10^0$	$2.743 \times 10^{-2}$
mean-field	$-2.783 \times 10^0$	$8.697 \times 10^{-3}$
linear	$-2.794 \times 10^0$	$2.999 \times 10^{-2}$
unscented'95	$-2.777 \times 10^0$	$2.814 \times 10^{-2}$
unscented'02	$-2.776 \times 10^0$	$3.064 \times 10^{-2}$

Table 56: Comparison of moments for Network(architecture=wide, weights=initialized, activation=gelu), variance=small

4221  
4222  
4223  
4224  
4225  
4226  
4227  
4228  
4229  
4230  
4231  
4232  
4233

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.772 \times 10^{-2} \pm 1.9 \times 10^{-6}$	0
analytic	$1.697 \times 10^{-2} \pm 2.5 \times 10^{-6}$	$1.302 \times 10^{-4} \pm 4.5 \times 10^{-8}$
mean-field	$1.423 \times 10^{-1} \pm 1.7 \times 10^{-6}$	$2.257 \times 10^{-1} \pm 1.3 \times 10^{-6}$
linear	$4.249 \times 10^{-2} \pm 1.1 \times 10^{-6}$	$8.757 \times 10^{-3} \pm 2.3 \times 10^{-7}$
unscented'95	$1.685 \times 10^{-2} \pm 2.6 \times 10^{-6}$	$5.865 \times 10^{-4} \pm 9.6 \times 10^{-8}$
unscented'02	$2.505 \times 10^{-2} \pm 1.2 \times 10^{-6}$	$4.667 \times 10^{-3} \pm 2.8 \times 10^{-7}$

Table 57: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=gelu), variance=small

4234  
4235  
4236  
4237  
4238  
4239  
4240  
4241  
4242  
4243  
4244  
4245  
4246  
4247  
4248  
4249  
4250  
4251  
4252  
4253  
4254  
4255  
4256

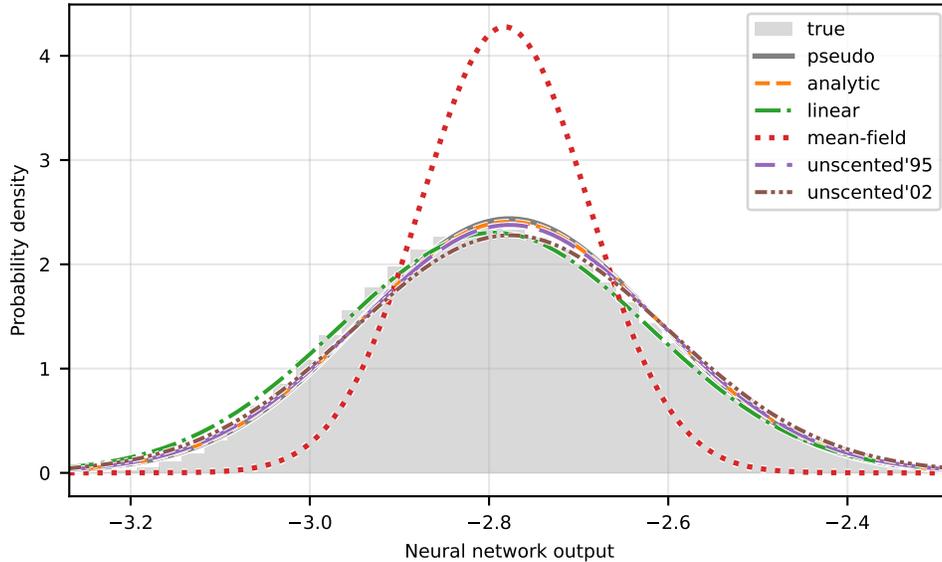


Figure 36: Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu), variance=small

4257  
4258  
4259  
4260  
4261  
4262  
4263  
4264  
4265

4266  
4267  
4268  
4269  
4270  
4271  
4272  
4273  
4274

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.176 \times 10^0 \pm 4.3 \times 10^{-6}$	$6.248 \times 10^{-1} \pm 1.4 \times 10^{-5}$
analytic	$-2.127 \times 10^0$	$5.620 \times 10^{-1}$
mean-field	$-2.228 \times 10^0$	$5.011 \times 10^{-1}$
linear	$-2.794 \times 10^0$	$2.999 \times 10^0$
unscented'95	$-2.106 \times 10^0$	$6.547 \times 10^{-1}$
unscented'02	$-9.887 \times 10^{-1}$	$9.520 \times 10^0$

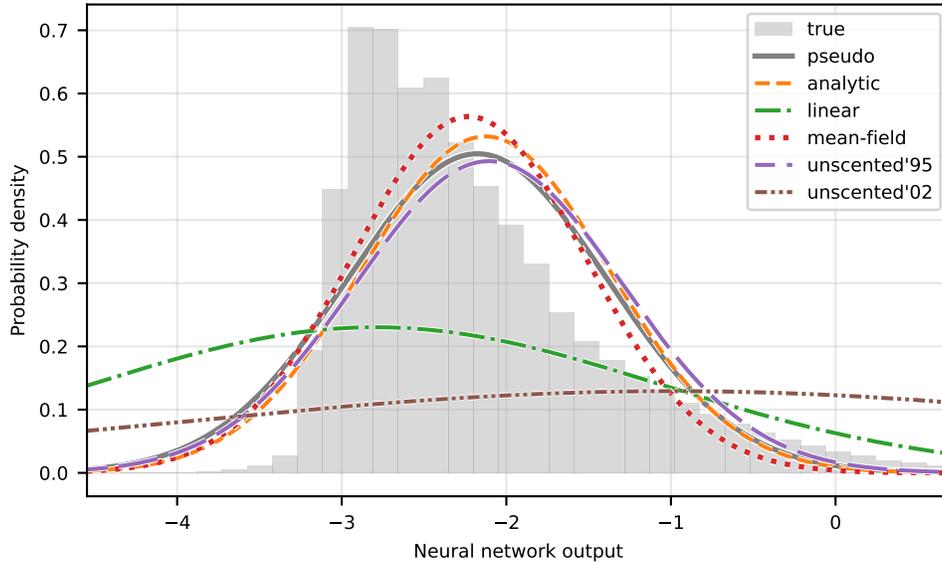
4275 Table 58: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
4276 tion=gelu), variance=medium  
4277

4278  
4279  
4280  
4281  
4282  
4283  
4284  
4285  
4286  
4287

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.103 \times 10^{-1} \pm 8.5 \times 10^{-6}$	0
analytic	$2.270 \times 10^{-1} \pm 9.0 \times 10^{-6}$	$4.684 \times 10^{-3} \pm 1.0 \times 10^{-6}$
mean-field	$1.903 \times 10^{-1} \pm 9.8 \times 10^{-6}$	$1.346 \times 10^{-2} \pm 2.4 \times 10^{-6}$
linear	$9.738 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$1.421 \times 10^0 \pm 4.8 \times 10^{-5}$
unscented'95	$2.438 \times 10^{-1} \pm 7.9 \times 10^{-6}$	$4.571 \times 10^{-3} \pm 9.5 \times 10^{-7}$
unscented'02	$2.444 \times 10^0 \pm 1.9 \times 10^{-5}$	$6.885 \times 10^0 \pm 1.9 \times 10^{-4}$

4288 Table 59: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
4289 activation=gelu), variance=medium  
4290

4291  
4292  
4293  
4294  
4295  
4296  
4297  
4298  
4299  
4300  
4301  
4302  
4303  
4304  
4305  
4306  
4307  
4308  
4309



4310  
4311  
4312 Figure 37: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
4313 tion=gelu), variance=medium  
4314  
4315  
4316  
4317  
4318  
4319

4320  
4321  
4322  
4323  
4324  
4325  
4326  
4327  
4328

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+9.844 \times 10^{-1} \pm 5.5 \times 10^{-5}$	$3.298 \times 10^{+1} \pm 1.3 \times 10^{-3}$
analytic	$+8.293 \times 10^{-1}$	$3.058 \times 10^{+1}$
mean-field	$+5.103 \times 10^{-1}$	$2.827 \times 10^{+1}$
linear	$-2.794 \times 10^0$	$2.999 \times 10^{+2}$
unscented'95	$+3.435 \times 10^0$	$3.525 \times 10^{+1}$
unscented'02	$+1.777 \times 10^{+2}$	$6.547 \times 10^{+4}$

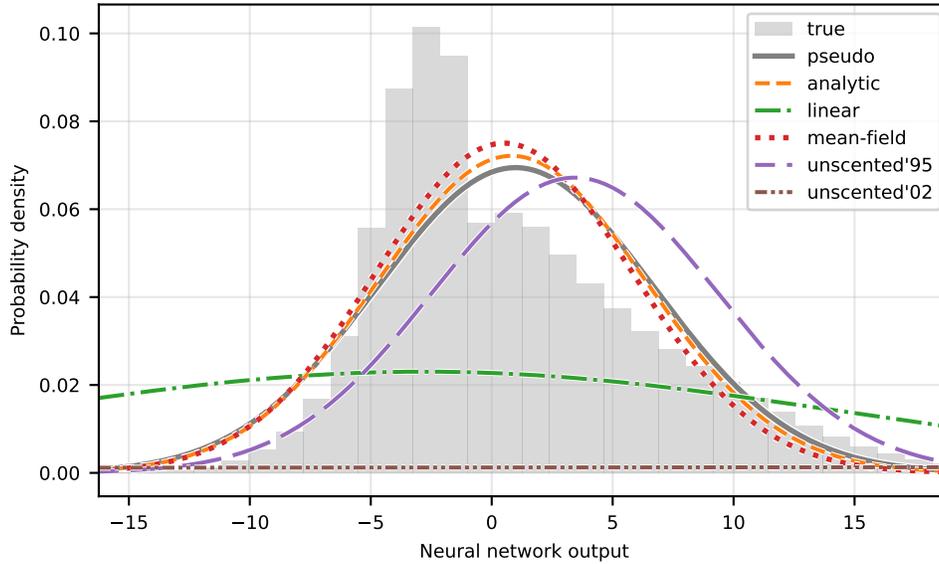
4329 Table 60: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
4330 tion=gelu), variance=large  
4331

4332  
4333  
4334  
4335  
4336  
4337  
4338  
4339  
4340  
4341

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.269 \times 10^{-1} \pm 4.4 \times 10^{-5}$	0
analytic	$4.155 \times 10^{-1} \pm 4.9 \times 10^{-5}$	$1.761 \times 10^{-3} \pm 1.4 \times 10^{-6}$
mean-field	$4.163 \times 10^{-1} \pm 5.2 \times 10^{-5}$	$9.056 \times 10^{-3} \pm 2.8 \times 10^{-6}$
linear	$3.984 \times 10^0 \pm 6.7 \times 10^{-5}$	$3.159 \times 10^0 \pm 1.7 \times 10^{-4}$
unscented'95	$1.057 \times 10^0 \pm 5.0 \times 10^{-5}$	$9.214 \times 10^{-2} \pm 6.4 \times 10^{-6}$
unscented'02	$1.035 \times 10^{+2} \pm 1.0 \times 10^{-3}$	$1.462 \times 10^{+3} \pm 5.7 \times 10^{-2}$

4342 Table 61: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
4343 activation=gelu), variance=large  
4344

4345  
4346  
4347  
4348  
4349  
4350  
4351  
4352  
4353  
4354  
4355  
4356  
4357  
4358  
4359  
4360  
4361  
4362  
4363  
4364



4365  
4366 Figure 38: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
4367 tion=gelu), variance=large  
4368

4369  
4370  
4371  
4372  
4373

4374  
4375  
4376  
4377  
4378  
4379  
4380  
4381  
4382  
4383  
4384  
4385  
4386  
4387  
4388  
4389  
4390  
4391  
4392  
4393  
4394  
4395  
4396  
4397  
4398  
4399  
4400  
4401  
4402  
4403  
4404  
4405  
4406  
4407  
4408  
4409  
4410  
4411  
4412  
4413  
4414  
4415  
4416  
4417  
4418  
4419  
4420  
4421  
4422  
4423  
4424  
4425  
4426  
4427

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.976 \times 10^{-2} \pm 2.3 \times 10^{-7}$	$4.419 \times 10^{-2} \pm 4.2 \times 10^{-7}$
analytic	$-7.000 \times 10^{-2}$	$4.266 \times 10^{-2}$
mean-field	$-4.962 \times 10^{-2}$	$9.842 \times 10^{-3}$
linear	$-2.054 \times 10^{-2}$	$4.617 \times 10^{-2}$
unscented'95	$-7.075 \times 10^{-2}$	$4.375 \times 10^{-2}$
unscented'02	$-7.263 \times 10^{-2}$	$5.159 \times 10^{-2}$

Table 62: Comparison of moments for Network(architecture=wide, weights=trained, activation=gelu), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.228 \times 10^{-2} \pm 5.2 \times 10^{-6}$	0
analytic	$4.302 \times 10^{-2} \pm 4.3 \times 10^{-6}$	$3.042 \times 10^{-4} \pm 1.6 \times 10^{-7}$
mean-field	$1.926 \times 10^{-1} \pm 3.5 \times 10^{-6}$	$3.668 \times 10^{-1} \pm 3.7 \times 10^{-6}$
linear	$1.073 \times 10^{-1} \pm 4.8 \times 10^{-7}$	$2.790 \times 10^{-2} \pm 4.5 \times 10^{-7}$
unscented'95	$4.315 \times 10^{-2} \pm 4.8 \times 10^{-6}$	$3.611 \times 10^{-5} \pm 4.5 \times 10^{-8}$
unscented'02	$5.126 \times 10^{-2} \pm 4.7 \times 10^{-6}$	$6.421 \times 10^{-3} \pm 8.0 \times 10^{-7}$

Table 63: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=gelu), variance=small

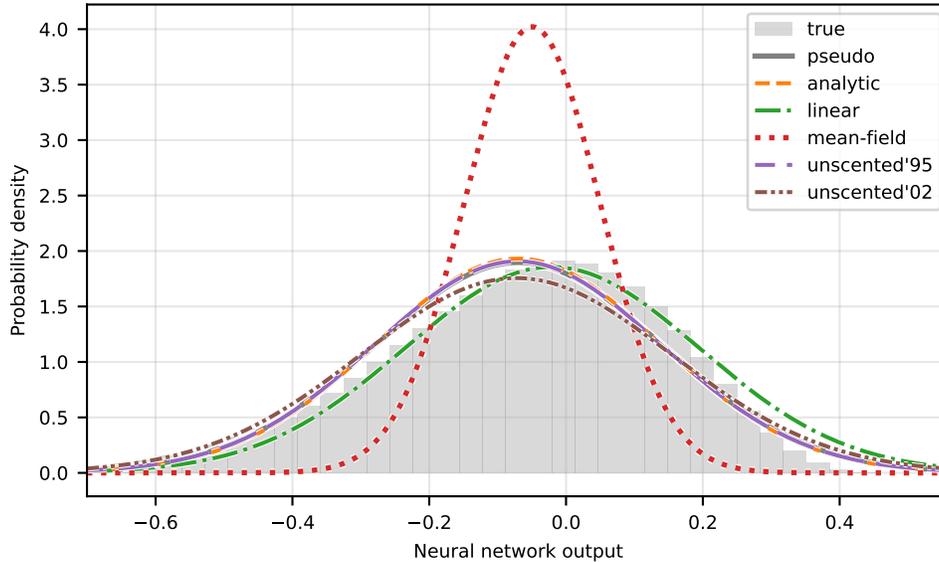


Figure 39: Probability distributions for Network(architecture=wide, weights=trained, activation=gelu), variance=small

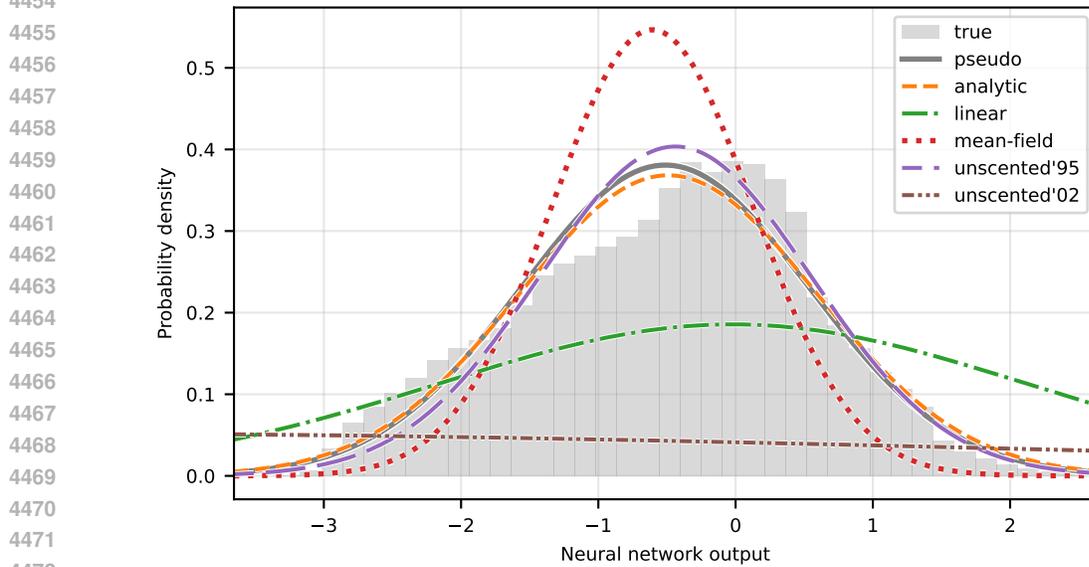
4428  
4429  
4430  
4431  
4432  
4433  
4434  
4435  
4436

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.119 \times 10^{-1} \pm 5.3 \times 10^{-6}$	$1.097 \times 10^0 \pm 1.3 \times 10^{-5}$
analytic	$-4.912 \times 10^{-1}$	$1.173 \times 10^0$
mean-field	$-6.076 \times 10^{-1}$	$5.327 \times 10^{-1}$
linear	$-2.054 \times 10^{-2}$	$4.617 \times 10^0$
unscented'95	$-4.392 \times 10^{-1}$	$9.772 \times 10^{-1}$
unscented'02	$-5.229 \times 10^0$	$5.888 \times 10^{+1}$

4437 Table 64: Comparison of moments for Network(architecture=wide, weights=trained, activa-  
4438 tion=gelu), variance=medium  
4439  
4440

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$7.352 \times 10^{-2} \pm 1.1 \times 10^{-5}$	0
analytic	$7.357 \times 10^{-2} \pm 1.3 \times 10^{-5}$	$1.333 \times 10^{-3} \pm 4.1 \times 10^{-7}$
mean-field	$2.944 \times 10^{-1} \pm 9.8 \times 10^{-6}$	$1.082 \times 10^{-1} \pm 3.0 \times 10^{-6}$
linear	$9.054 \times 10^{-1} \pm 7.6 \times 10^{-6}$	$9.951 \times 10^{-1} \pm 2.0 \times 10^{-5}$
unscented'95	$9.088 \times 10^{-2} \pm 1.1 \times 10^{-5}$	$5.652 \times 10^{-3} \pm 8.1 \times 10^{-7}$
unscented'02	$6.447 \times 10^0 \pm 2.5 \times 10^{-5}$	$3.447 \times 10^{+1} \pm 4.5 \times 10^{-4}$

4441  
4442  
4443  
4444  
4445  
4446  
4447  
4448  
4449  
4450 Table 65: Comparison of statistical distances for Network(architecture=wide, weights=trained, ac-  
4451 tivation=gelu), variance=medium  
4452  
4453  
4454



4455  
4456  
4457  
4458  
4459  
4460  
4461  
4462  
4463  
4464  
4465  
4466  
4467  
4468  
4469  
4470  
4471  
4472  
4473  
4474 Figure 40: Probability distributions for Network(architecture=wide, weights=trained, activa-  
4475 tion=gelu), variance=medium  
4476  
4477  
4478  
4479  
4480  
4481

4482  
4483  
4484  
4485  
4486  
4487  
4488  
4489  
4490

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.787 \times 10^0 \pm 7.5 \times 10^{-5}$	$3.069 \times 10^{+1} \pm 1.5 \times 10^{-3}$
analytic	$+2.645 \times 10^0$	$2.931 \times 10^{+1}$
mean-field	$+2.512 \times 10^0$	$2.951 \times 10^{+1}$
linear	$-2.054 \times 10^{-2}$	$4.617 \times 10^{+2}$
unscented'95	$+5.113 \times 10^0$	$3.012 \times 10^{+1}$
unscented'02	$-5.207 \times 10^{+2}$	$5.427 \times 10^{+5}$

Table 66: Comparison of moments for Network(architecture=wide, weights=trained, activation=gelu), variance=large

4491  
4492  
4493  
4494  
4495  
4496  
4497  
4498  
4499  
4500  
4501  
4502  
4503

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.469 \times 10^{-1} \pm 3.9 \times 10^{-5}$	0
analytic	$4.410 \times 10^{-1} \pm 3.8 \times 10^{-5}$	$8.506 \times 10^{-4} \pm 1.3 \times 10^{-6}$
mean-field	$4.407 \times 10^{-1} \pm 4.1 \times 10^{-5}$	$1.611 \times 10^{-3} \pm 1.3 \times 10^{-6}$
linear	$5.473 \times 10^0 \pm 1.1 \times 10^{-4}$	$5.794 \times 10^0 \pm 3.6 \times 10^{-4}$
unscented'95	$1.034 \times 10^0 \pm 4.3 \times 10^{-5}$	$8.826 \times 10^{-2} \pm 8.0 \times 10^{-6}$
unscented'02	$3.086 \times 10^{+2} \pm 3.9 \times 10^{-3}$	$1.330 \times 10^{+4} \pm 6.6 \times 10^{-1}$

Table 67: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=gelu), variance=large

4504  
4505  
4506  
4507  
4508  
4509  
4510  
4511  
4512  
4513  
4514  
4515  
4516  
4517  
4518  
4519  
4520  
4521  
4522  
4523  
4524  
4525  
4526  
4527

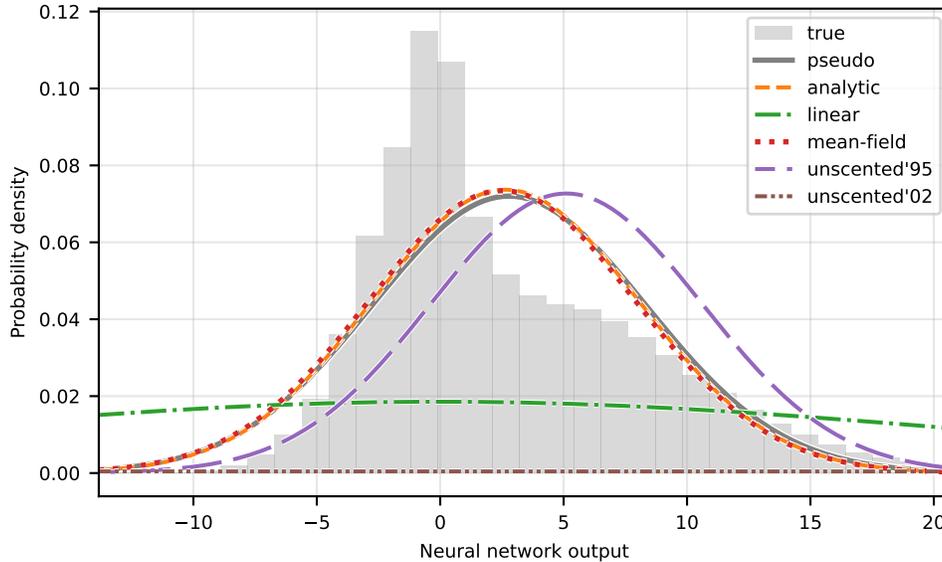


Figure 41: Probability distributions for Network(architecture=wide, weights=trained, activation=gelu), variance=large

4528  
4529  
4530  
4531  
4532  
4533  
4534  
4535

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.070 \times 10^0 \pm 3.8 \times 10^{-7}$	$2.447 \times 10^{-1} \pm 1.4 \times 10^{-6}$
analytic	$-5.069 \times 10^0$	$2.424 \times 10^{-1}$
mean-field	$-5.064 \times 10^0$	$1.377 \times 10^{-1}$
linear	$-5.098 \times 10^0$	$2.464 \times 10^{-1}$
unscented'95	$-5.070 \times 10^0$	$2.444 \times 10^{-1}$
unscented'02	$-5.069 \times 10^0$	$2.480 \times 10^{-1}$

Table 68: Comparison of moments for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.662 \times 10^{-2} \pm 3.2 \times 10^{-6}$	0
analytic	$2.694 \times 10^{-2} \pm 3.0 \times 10^{-6}$	$2.256 \times 10^{-5} \pm 2.5 \times 10^{-8}$
mean-field	$1.408 \times 10^{-1} \pm 2.7 \times 10^{-6}$	$6.890 \times 10^{-2} \pm 1.2 \times 10^{-6}$
linear	$3.954 \times 10^{-2} \pm 5.2 \times 10^{-7}$	$1.593 \times 10^{-3} \pm 4.2 \times 10^{-8}$
unscented'95	$2.646 \times 10^{-2} \pm 3.4 \times 10^{-6}$	$4.166 \times 10^{-7} \pm 3.2 \times 10^{-9}$
unscented'02	$2.755 \times 10^{-2} \pm 2.8 \times 10^{-6}$	$4.789 \times 10^{-5} \pm 3.9 \times 10^{-8}$

Table 69: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=small

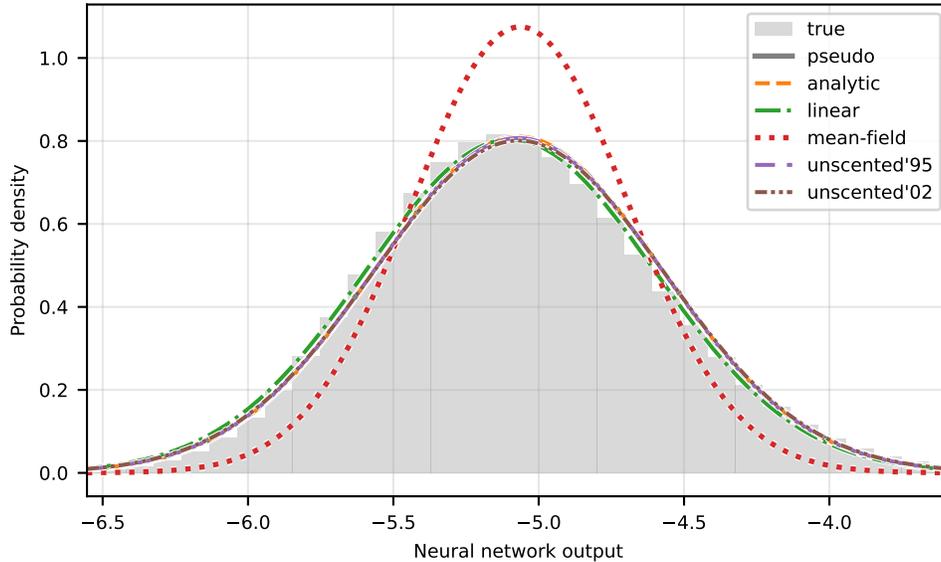


Figure 42: Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=small

4590  
4591  
4592  
4593  
4594  
4595  
4596  
4597  
4598  
4599  
4600  
4601  
4602  
4603  
4604  
4605  
4606  
4607  
4608  
4609  
4610  
4611  
4612  
4613  
4614  
4615  
4616  
4617  
4618  
4619  
4620  
4621  
4622  
4623  
4624  
4625  
4626  
4627  
4628  
4629  
4630  
4631  
4632  
4633  
4634  
4635  
4636  
4637  
4638  
4639  
4640  
4641  
4642  
4643

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.044 \times 10^0 \pm 1.3 \times 10^{-5}$	$8.996 \times 10^0 \pm 1.2 \times 10^{-4}$
analytic	$-6.034 \times 10^0$	$7.057 \times 10^0$
mean-field	$-6.139 \times 10^0$	$8.948 \times 10^0$
linear	$-5.098 \times 10^0$	$2.464 \times 10^{+1}$
unscented'95	$-6.119 \times 10^0$	$6.362 \times 10^0$
unscented'02	$-2.249 \times 10^0$	$4.087 \times 10^{+1}$

Table 70: Comparison of moments for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.041 \times 10^{-1} \pm 2.0 \times 10^{-5}$	0
analytic	$2.224 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$1.362 \times 10^{-2} \pm 1.4 \times 10^{-6}$
mean-field	$1.218 \times 10^{-1} \pm 1.7 \times 10^{-5}$	$5.037 \times 10^{-4} \pm 1.3 \times 10^{-7}$
linear	$9.831 \times 10^{-1} \pm 1.2 \times 10^{-5}$	$4.156 \times 10^{-1} \pm 1.2 \times 10^{-5}$
unscented'95	$2.747 \times 10^{-1} \pm 1.1 \times 10^{-5}$	$2.714 \times 10^{-2} \pm 1.9 \times 10^{-6}$
unscented'02	$2.438 \times 10^0 \pm 1.5 \times 10^{-5}$	$1.816 \times 10^0 \pm 3.2 \times 10^{-5}$

Table 71: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=medium

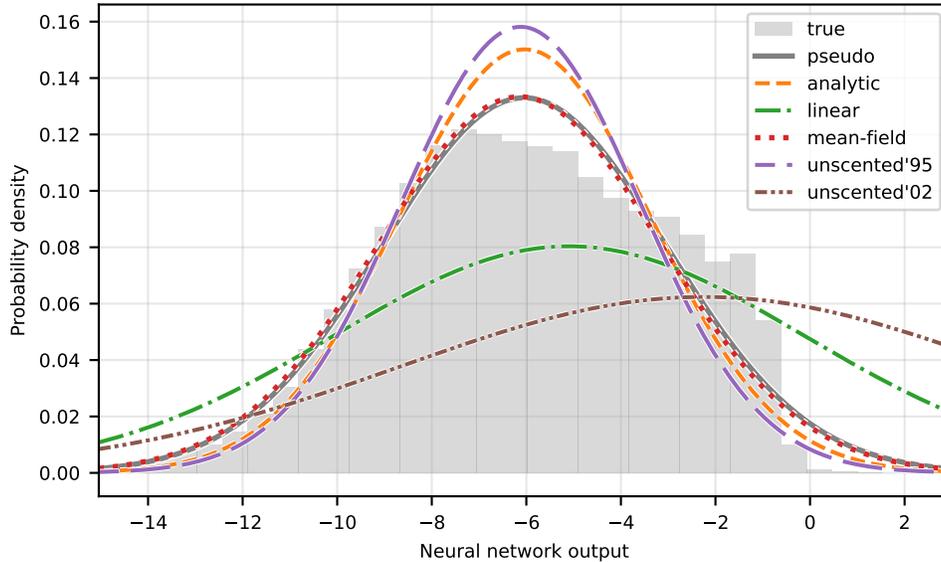


Figure 43: Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=medium

4644  
4645  
4646  
4647  
4648  
4649  
4650  
4651  
4652  
4653  
4654  
4655  
4656  
4657  
4658  
4659  
4660  
4661  
4662  
4663  
4664  
4665  
4666  
4667  
4668  
4669  
4670  
4671  
4672  
4673  
4674  
4675  
4676  
4677  
4678  
4679  
4680  
4681  
4682  
4683  
4684  
4685  
4686  
4687  
4688  
4689  
4690  
4691  
4692  
4693  
4694  
4695  
4696  
4697

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-3.746 \times 10^{+1} \pm 2.6 \times 10^{-4}$	$7.905 \times 10^{+2} \pm 1.9 \times 10^{-2}$
analytic	$-3.854 \times 10^{+1}$	$5.569 \times 10^{+2}$
mean-field	$-3.505 \times 10^{+1}$	$7.171 \times 10^{+2}$
linear	$-5.098 \times 10^0$	$2.464 \times 10^{+3}$
unscented'95	$-3.841 \times 10^{+1}$	$6.145 \times 10^{+2}$
unscented'02	$+2.797 \times 10^{+2}$	$1.647 \times 10^{+5}$

Table 72: Comparison of moments for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.608 \times 10^{-1} \pm 6.8 \times 10^{-5}$	0
analytic	$9.376 \times 10^{-1} \pm 4.0 \times 10^{-5}$	$2.814 \times 10^{-2} \pm 3.3 \times 10^{-6}$
mean-field	$6.046 \times 10^{-1} \pm 5.8 \times 10^{-5}$	$5.971 \times 10^{-3} \pm 1.5 \times 10^{-6}$
linear	$6.194 \times 10^0 \pm 6.0 \times 10^{-5}$	$1.153 \times 10^0 \pm 3.7 \times 10^{-5}$
unscented'95	$8.150 \times 10^{-1} \pm 2.7 \times 10^{-5}$	$1.519 \times 10^{-2} \pm 2.5 \times 10^{-6}$
unscented'02	$7.554 \times 10^{+1} \pm 4.6 \times 10^{-4}$	$1.646 \times 10^{+2} \pm 3.9 \times 10^{-3}$

Table 73: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=large

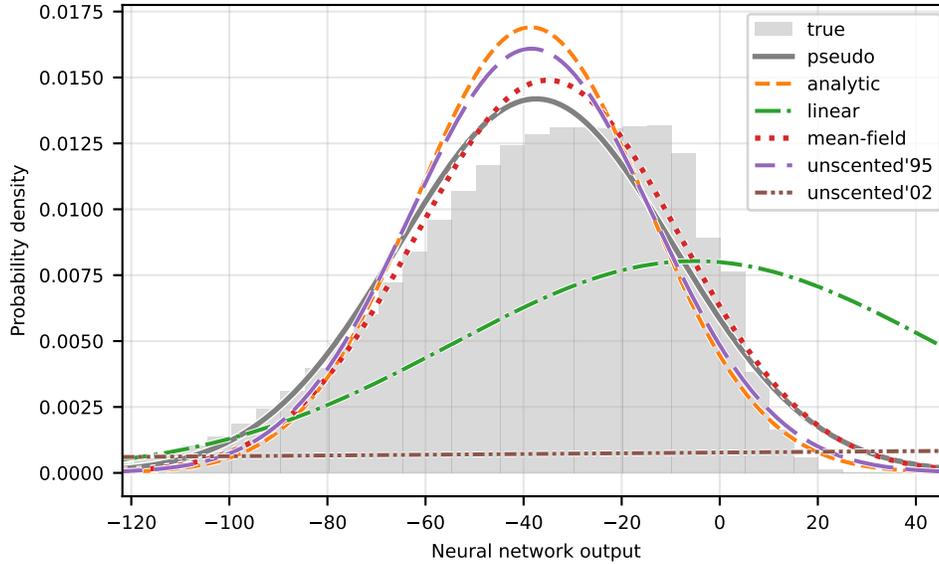


Figure 44: Probability distributions for Network(architecture=wide, weights=initialized, activation=gelu residual), variance=large

4698  
4699  
4700  
4701  
4702  
4703  
4704  
4705  
4706

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.997 \times 10^{-2} \pm 3.2 \times 10^{-7}$	$7.824 \times 10^{-2} \pm 7.3 \times 10^{-7}$
analytic	$+1.749 \times 10^{-2}$	$7.865 \times 10^{-2}$
mean-field	$+3.802 \times 10^{-2}$	$1.411 \times 10^{-1}$
linear	$+6.045 \times 10^{-2}$	$7.470 \times 10^{-2}$
unscented'95	$+1.659 \times 10^{-2}$	$7.546 \times 10^{-2}$
unscented'02	$+1.251 \times 10^{-2}$	$7.929 \times 10^{-2}$

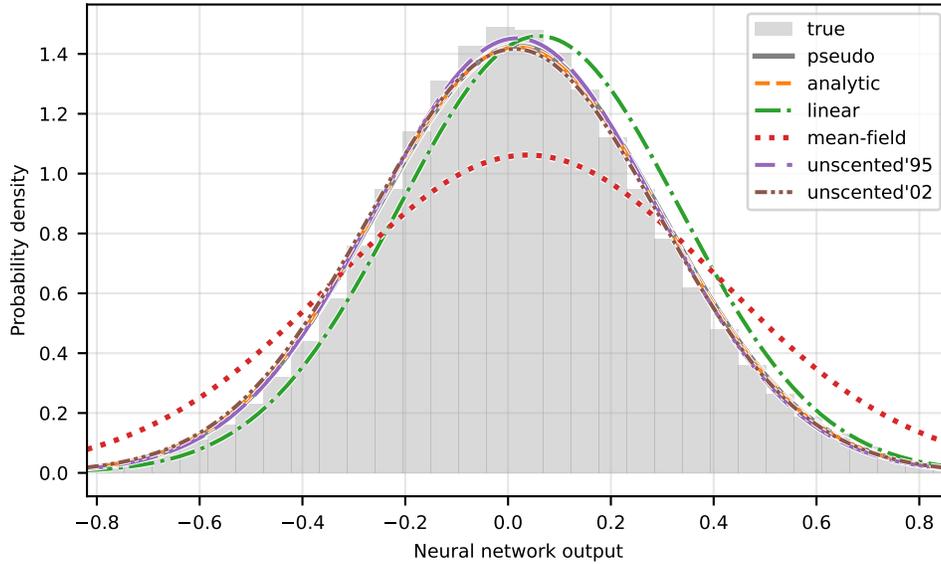
4707 Table 74: Comparison of moments for Network(architecture=wide, weights=trained, activa-  
4708 tion=gelu residual), variance=small  
4709

4711  
4712  
4713  
4714  
4715  
4716  
4717  
4718  
4719

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.176 \times 10^{-2} \pm 7.0 \times 10^{-6}$	0
analytic	$1.217 \times 10^{-2} \pm 6.1 \times 10^{-6}$	$4.612 \times 10^{-5} \pm 2.7 \times 10^{-8}$
mean-field	$1.536 \times 10^{-1} \pm 3.5 \times 10^{-6}$	$1.089 \times 10^{-1} \pm 3.8 \times 10^{-6}$
linear	$7.708 \times 10^{-2} \pm 3.3 \times 10^{-6}$	$1.100 \times 10^{-2} \pm 2.1 \times 10^{-7}$
unscented'95	$1.055 \times 10^{-2} \pm 6.5 \times 10^{-6}$	$3.966 \times 10^{-4} \pm 1.7 \times 10^{-7}$
unscented'02	$1.553 \times 10^{-2} \pm 7.9 \times 10^{-6}$	$4.003 \times 10^{-4} \pm 7.4 \times 10^{-8}$

4720 Table 75: Comparison of statistical distances for Network(architecture=wide, weights=trained, ac-  
4721 tivation=gelu residual), variance=small  
4722

4723  
4724  
4725  
4726  
4727  
4728  
4729  
4730  
4731  
4732  
4733  
4734  
4735  
4736  
4737  
4738  
4739



4742  
4743  
4744  
4745  
4746  
4747  
4748  
4749  
4750  
4751

4744 Figure 45: Probability distributions for Network(architecture=wide, weights=trained, activa-  
4745 tion=gelu residual), variance=small

4752  
4753  
4754  
4755  
4756  
4757  
4758  
4759  
4760  
4761  
4762  
4763  
4764  
4765  
4766  
4767  
4768  
4769  
4770  
4771  
4772  
4773  
4774  
4775  
4776  
4777  
4778  
4779  
4780  
4781  
4782  
4783  
4784  
4785  
4786  
4787  
4788  
4789  
4790  
4791  
4792  
4793  
4794  
4795  
4796  
4797  
4798  
4799  
4800  
4801  
4802  
4803  
4804  
4805

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-7.433 \times 10^{-1} \pm 1.2 \times 10^{-5}$	$2.825 \times 10^0 \pm 8.8 \times 10^{-5}$
analytic	$-7.555 \times 10^{-1}$	$3.514 \times 10^0$
mean-field	$-8.641 \times 10^{-1}$	$9.025 \times 10^0$
linear	$+6.045 \times 10^{-2}$	$7.470 \times 10^0$
unscented'95	$-7.988 \times 10^{-1}$	$1.499 \times 10^0$
unscented'02	$-4.733 \times 10^0$	$5.343 \times 10^{+1}$

Table 76: Comparison of moments for Network(architecture=wide, weights=trained, activation=gelu residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.348 \times 10^{-1} \pm 2.4 \times 10^{-5}$	0
analytic	$1.977 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$1.285 \times 10^{-2} \pm 3.8 \times 10^{-6}$
mean-field	$8.580 \times 10^{-1} \pm 1.9 \times 10^{-5}$	$5.193 \times 10^{-1} \pm 3.4 \times 10^{-5}$
linear	$8.027 \times 10^{-1} \pm 1.2 \times 10^{-5}$	$4.504 \times 10^{-1} \pm 2.9 \times 10^{-5}$
unscented'95	$2.733 \times 10^{-1} \pm 1.2 \times 10^{-5}$	$8.263 \times 10^{-2} \pm 7.3 \times 10^{-6}$
unscented'02	$4.372 \times 10^0 \pm 3.8 \times 10^{-5}$	$1.031 \times 10^{+1} \pm 3.7 \times 10^{-4}$

Table 77: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=gelu residual), variance=medium

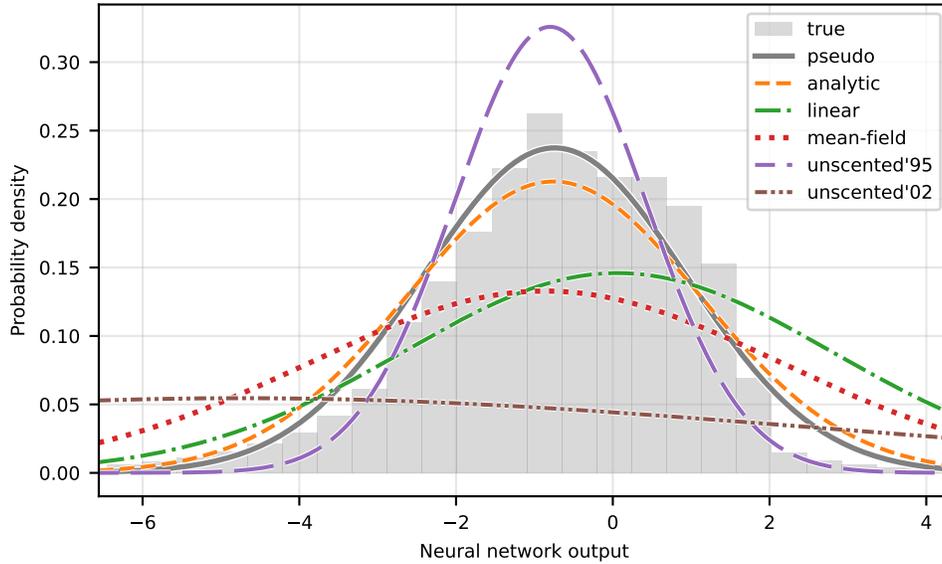


Figure 46: Probability distributions for Network(architecture=wide, weights=trained, activation=gelu residual), variance=medium

4806  
4807  
4808  
4809  
4810  
4811  
4812  
4813  
4814

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-9.417 \times 10^0 \pm 2.5 \times 10^{-4}$	$5.784 \times 10^{+2} \pm 1.3 \times 10^{-2}$
analytic	$-9.746 \times 10^0$	$4.943 \times 10^{+2}$
mean-field	$-6.172 \times 10^0$	$7.188 \times 10^{+2}$
linear	$+6.045 \times 10^{-2}$	$7.470 \times 10^{+2}$
unscented'95	$-8.083 \times 10^0$	$6.251 \times 10^{+2}$
unscented'02	$-4.789 \times 10^{+2}$	$4.595 \times 10^{+5}$

4815  
4816  
4817  
4818

Table 78: Comparison of moments for Network(architecture=wide, weights=trained, activation=gelu residual), variance=large

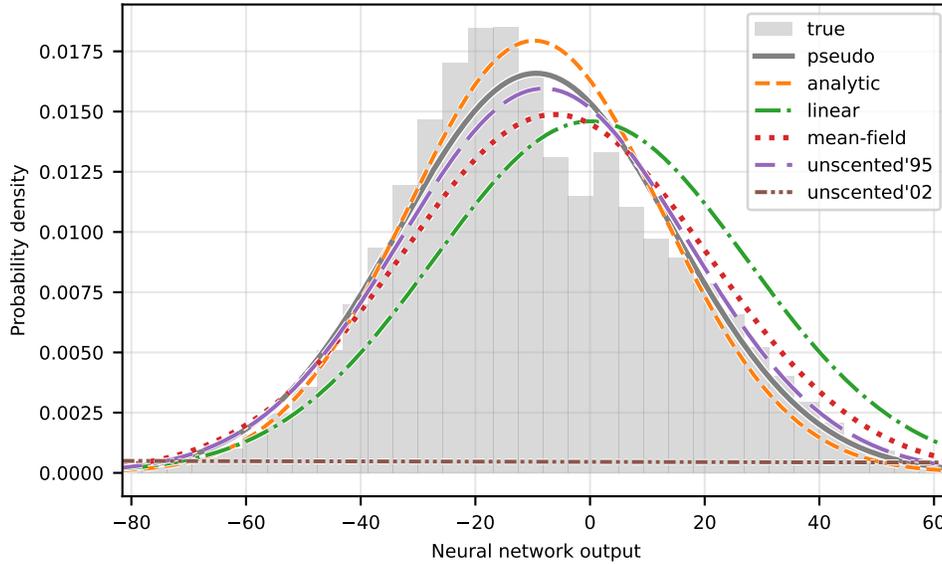
4819  
4820  
4821  
4822  
4823  
4824  
4825  
4826  
4827

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.188 \times 10^{-1} \pm 6.6 \times 10^{-5}$	0
analytic	$3.979 \times 10^{-1} \pm 6.8 \times 10^{-5}$	$5.957 \times 10^{-3} \pm 1.6 \times 10^{-6}$
mean-field	$8.734 \times 10^{-1} \pm 7.2 \times 10^{-5}$	$2.183 \times 10^{-2} \pm 3.0 \times 10^{-6}$
linear	$1.954 \times 10^0 \pm 5.5 \times 10^{-5}$	$9.551 \times 10^{-2} \pm 5.9 \times 10^{-6}$
unscented'95	$4.304 \times 10^{-1} \pm 6.8 \times 10^{-5}$	$3.085 \times 10^{-3} \pm 1.0 \times 10^{-6}$
unscented'02	$1.325 \times 10^{+2} \pm 7.4 \times 10^{-4}$	$5.839 \times 10^{+2} \pm 1.3 \times 10^{-2}$

4828  
4829  
4830

Table 79: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=gelu residual), variance=large

4831



4832  
4833  
4834  
4835  
4836  
4837  
4838  
4839  
4840  
4841  
4842  
4843  
4844  
4845  
4846  
4847  
4848  
4849  
4850  
4851

Figure 47: Probability distributions for Network(architecture=wide, weights=trained, activation=gelu residual), variance=large

4852  
4853  
4854  
4855  
4856  
4857  
4858  
4859

4860  
4861  
4862  
4863  
4864  
4865  
4866  
4867  
4868

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.548 \times 10^0 \pm 3.6 \times 10^{-7}$	$1.672 \times 10^{-2} \pm 1.7 \times 10^{-7}$
analytic	$-2.550 \times 10^0$	$1.674 \times 10^{-2}$
mean-field	$-2.551 \times 10^0$	$7.878 \times 10^{-3}$
linear	$-2.534 \times 10^0$	$1.529 \times 10^{-2}$
unscented'95	$-2.552 \times 10^0$	$1.753 \times 10^{-2}$
unscented'02	$-2.534 \times 10^0$	$1.529 \times 10^{-2}$

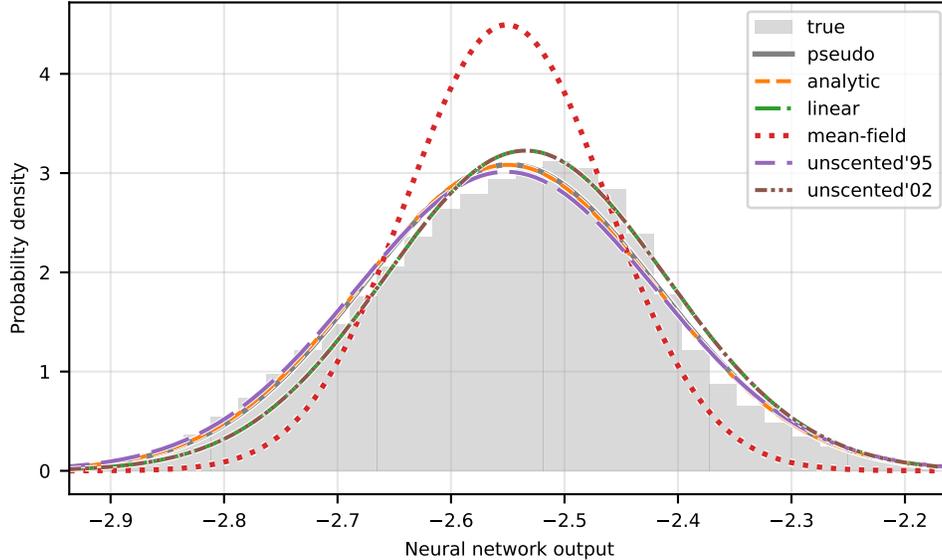
4869 Table 80: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
4870 tion=relu), variance=small  
4871

4872  
4873  
4874  
4875  
4876  
4877  
4878  
4879  
4880  
4881

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.237 \times 10^{-2} \pm 3.0 \times 10^{-6}$	0
analytic	$1.321 \times 10^{-2} \pm 2.2 \times 10^{-6}$	$7.681 \times 10^{-5} \pm 3.2 \times 10^{-8}$
mean-field	$9.102 \times 10^{-2} \pm 1.7 \times 10^{-6}$	$1.120 \times 10^{-1} \pm 2.8 \times 10^{-6}$
linear	$4.066 \times 10^{-2} \pm 1.5 \times 10^{-6}$	$8.034 \times 10^{-3} \pm 3.8 \times 10^{-7}$
unscented'95	$1.699 \times 10^{-2} \pm 2.1 \times 10^{-6}$	$8.407 \times 10^{-4} \pm 2.3 \times 10^{-7}$
unscented'02	$4.066 \times 10^{-2} \pm 1.5 \times 10^{-6}$	$8.034 \times 10^{-3} \pm 3.8 \times 10^{-7}$

4882 Table 81: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
4883 activation=relu), variance=small  
4884

4885  
4886  
4887  
4888  
4889  
4890  
4891  
4892  
4893  
4894  
4895  
4896  
4897  
4898  
4899  
4900  
4901  
4902  
4903



4904  
4905  
4906 Figure 48: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
4907 tion=relu), variance=small  
4908

4909  
4910  
4911  
4912  
4913

4914  
4915  
4916  
4917  
4918  
4919  
4920  
4921  
4922

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.086 \times 10^0 \pm 8.0 \times 10^{-6}$	$5.616 \times 10^{-1} \pm 1.5 \times 10^{-5}$
analytic	$-2.047 \times 10^0$	$5.068 \times 10^{-1}$
mean-field	$-2.129 \times 10^0$	$3.972 \times 10^{-1}$
linear	$-2.534 \times 10^0$	$1.529 \times 10^0$
unscented'95	$-1.992 \times 10^0$	$6.390 \times 10^{-1}$
unscented'02	$+1.315 \times 10^{+1}$	$4.933 \times 10^{+2}$

Table 82: Comparison of moments for Network(architecture=wide, weights=initialized, activation=relu), variance=medium

4923  
4924  
4925  
4926  
4927  
4928  
4929  
4930  
4931  
4932  
4933  
4934  
4935

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.010 \times 10^{-1} \pm 1.2 \times 10^{-5}$	0
analytic	$2.124 \times 10^{-1} \pm 9.6 \times 10^{-6}$	$3.897 \times 10^{-3} \pm 1.1 \times 10^{-6}$
mean-field	$1.878 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$2.849 \times 10^{-2} \pm 4.2 \times 10^{-6}$
linear	$5.572 \times 10^{-1} \pm 9.1 \times 10^{-6}$	$5.394 \times 10^{-1} \pm 2.5 \times 10^{-5}$
unscented'95	$2.555 \times 10^{-1} \pm 1.4 \times 10^{-5}$	$1.223 \times 10^{-2} \pm 3.0 \times 10^{-6}$
unscented'02	$2.468 \times 10^{+1} \pm 1.7 \times 10^{-4}$	$6.418 \times 10^{+2} \pm 1.7 \times 10^{-2}$

Table 83: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=relu), variance=medium

4936  
4937  
4938  
4939  
4940  
4941  
4942  
4943  
4944  
4945  
4946  
4947  
4948  
4949  
4950  
4951  
4952  
4953  
4954  
4955  
4956  
4957

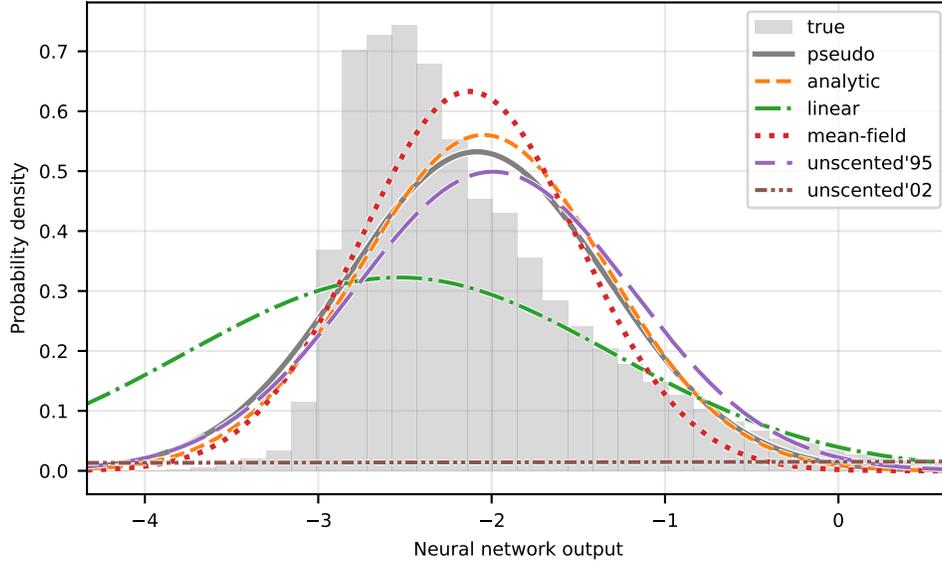


Figure 49: Probability distributions for Network(architecture=wide, weights=initialized, activation=relu), variance=medium

4962  
4963  
4964  
4965  
4966  
4967

4968  
4969  
4970  
4971  
4972  
4973  
4974  
4975  
4976  
4977  
4978  
4979  
4980  
4981  
4982  
4983  
4984  
4985  
4986  
4987  
4988  
4989  
4990  
4991  
4992  
4993  
4994  
4995  
4996  
4997  
4998  
4999  
5000  
5001  
5002  
5003  
5004  
5005  
5006  
5007  
5008  
5009  
5010  
5011  
5012  
5013  
5014  
5015  
5016  
5017  
5018  
5019  
5020  
5021

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+9.817 \times 10^{-1} \pm 5.8 \times 10^{-5}$	$3.271 \times 10^{+1} \pm 1.2 \times 10^{-3}$
analytic	$+8.530 \times 10^{-1}$	$2.964 \times 10^{+1}$
mean-field	$+5.436 \times 10^{-1}$	$2.726 \times 10^{+1}$
linear	$-2.534 \times 10^0$	$1.529 \times 10^{+2}$
unscented'95	$+3.402 \times 10^0$	$3.527 \times 10^{+1}$
unscented'02	$+7.403 \times 10^{+2}$	$1.104 \times 10^{+6}$

Table 84: Comparison of moments for Network(architecture=wide, weights=initialized, activation=relu), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.287 \times 10^{-1} \pm 4.3 \times 10^{-5}$	0
analytic	$4.185 \times 10^{-1} \pm 4.8 \times 10^{-5}$	$2.599 \times 10^{-3} \pm 1.6 \times 10^{-6}$
mean-field	$4.187 \times 10^{-1} \pm 4.9 \times 10^{-5}$	$1.075 \times 10^{-2} \pm 2.8 \times 10^{-6}$
linear	$2.396 \times 10^0 \pm 5.8 \times 10^{-5}$	$1.255 \times 10^0 \pm 7.3 \times 10^{-5}$
unscented'95	$1.049 \times 10^0 \pm 4.7 \times 10^{-5}$	$9.103 \times 10^{-2} \pm 5.8 \times 10^{-6}$
unscented'02	$4.327 \times 10^{+2} \pm 3.9 \times 10^{-3}$	$2.522 \times 10^{+4} \pm 8.9 \times 10^{-1}$

Table 85: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=relu), variance=large

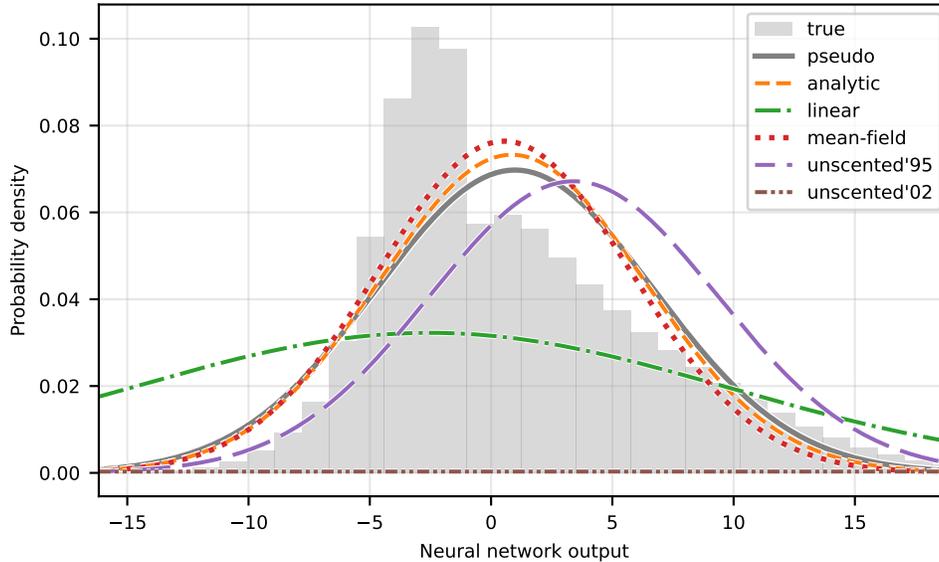


Figure 50: Probability distributions for Network(architecture=wide, weights=initialized, activation=relu), variance=large

5022  
5023  
5024  
5025  
5026  
5027  
5028  
5029

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.870 \times 10^{-1} \pm 4.3 \times 10^{-7}$	$1.633 \times 10^{-2} \pm 2.5 \times 10^{-7}$
analytic	$-2.877 \times 10^{-1}$	$1.361 \times 10^{-2}$
mean-field	$-2.822 \times 10^{-1}$	$8.050 \times 10^{-3}$
linear	$-2.263 \times 10^{-1}$	$2.322 \times 10^{-2}$
unscented'95	$-2.794 \times 10^{-1}$	$1.524 \times 10^{-2}$
unscented'02	$-2.263 \times 10^{-1}$	$2.322 \times 10^{-2}$

5030  
5031  
5032  
5033

Table 86: Comparison of moments for Network(architecture=wide, weights=trained, activation=relu), variance=small

5034

5035  
5036  
5037  
5038  
5039  
5040  
5041  
5042

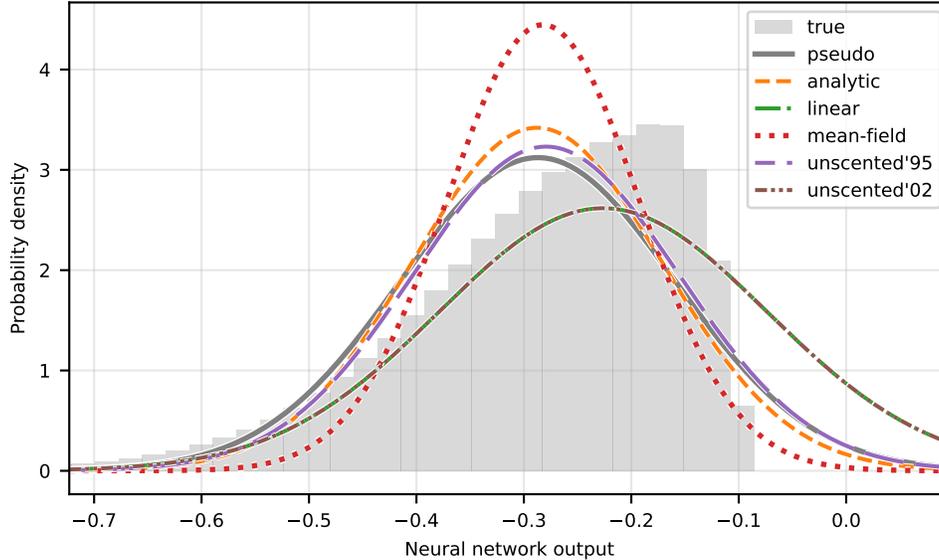
distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.259 \times 10^{-2} \pm 2.8 \times 10^{-6}$	0
analytic	$6.580 \times 10^{-2} \pm 3.4 \times 10^{-6}$	$7.816 \times 10^{-3} \pm 1.3 \times 10^{-6}$
mean-field	$9.224 \times 10^{-2} \pm 2.3 \times 10^{-6}$	$1.008 \times 10^{-1} \pm 3.9 \times 10^{-6}$
linear	$1.698 \times 10^{-1} \pm 9.1 \times 10^{-7}$	$1.479 \times 10^{-1} \pm 4.1 \times 10^{-6}$
unscented'95	$5.729 \times 10^{-2} \pm 2.5 \times 10^{-6}$	$2.914 \times 10^{-3} \pm 6.3 \times 10^{-7}$
unscented'02	$1.698 \times 10^{-1} \pm 9.1 \times 10^{-7}$	$1.479 \times 10^{-1} \pm 4.1 \times 10^{-6}$

5043  
5044  
5045  
5046

Table 87: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=relu), variance=small

5047

5048  
5049  
5050  
5051  
5052  
5053  
5054  
5055  
5056  
5057  
5058  
5059  
5060  
5061  
5062  
5063  
5064  
5065



5066  
5067  
5068  
5069

Figure 51: Probability distributions for Network(architecture=wide, weights=trained, activation=relu), variance=small

5070  
5071  
5072  
5073  
5074  
5075

5076  
5077  
5078  
5079  
5080  
5081  
5082  
5083  
5084

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-1.894 \times 10^{-1} \pm 8.2 \times 10^{-6}$	$5.868 \times 10^{-1} \pm 2.1 \times 10^{-5}$
analytic	$-1.771 \times 10^{-1}$	$6.492 \times 10^{-1}$
mean-field	$-3.079 \times 10^{-1}$	$4.012 \times 10^{-1}$
linear	$-2.263 \times 10^{-1}$	$2.322 \times 10^0$
unscented'95	$-1.181 \times 10^{-1}$	$4.681 \times 10^{-1}$
unscented'02	$-8.244 \times 10^0$	$1.309 \times 10^{+2}$

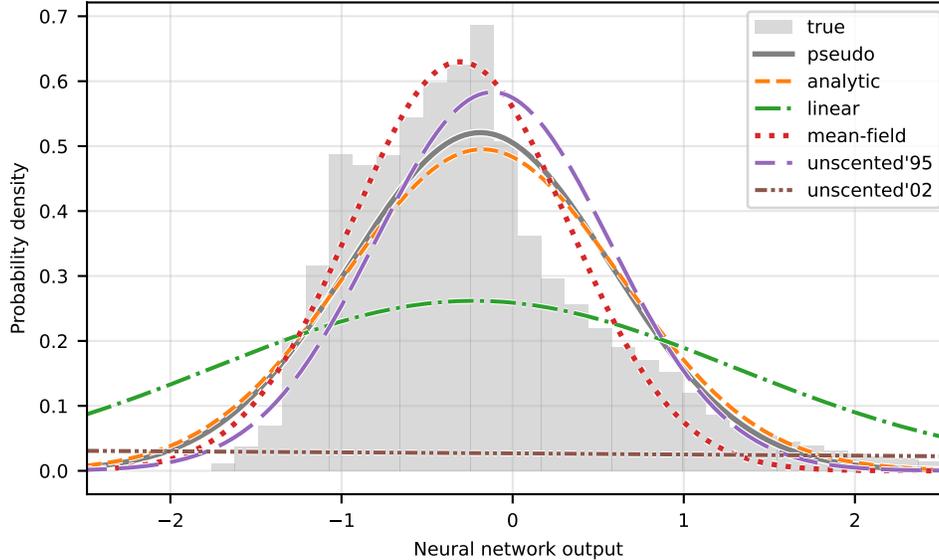
Table 88: Comparison of moments for Network(architecture=wide, weights=trained, activation=relu), variance=medium

5085  
5086  
5087  
5088  
5089  
5090  
5091  
5092  
5093  
5094  
5095  
5096  
5097

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.485 \times 10^{-1} \pm 1.3 \times 10^{-5}$	0
analytic	$1.619 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$2.772 \times 10^{-3} \pm 2.0 \times 10^{-6}$
mean-field	$1.360 \times 10^{-1} \pm 1.1 \times 10^{-5}$	$4.392 \times 10^{-2} \pm 6.1 \times 10^{-6}$
linear	$7.349 \times 10^{-1} \pm 1.4 \times 10^{-5}$	$7.923 \times 10^{-1} \pm 5.2 \times 10^{-5}$
unscented'95	$1.910 \times 10^{-1} \pm 1.2 \times 10^{-5}$	$1.618 \times 10^{-2} \pm 3.1 \times 10^{-6}$
unscented'02	$1.233 \times 10^{+1} \pm 1.2 \times 10^{-4}$	$1.636 \times 10^{+2} \pm 5.8 \times 10^{-3}$

Table 89: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=relu), variance=medium

5100  
5101  
5102  
5103  
5104  
5105  
5106  
5107  
5108  
5109  
5110  
5111  
5112  
5113  
5114  
5115  
5116  
5117



5118  
5119  
5120  
5121  
5122  
5123  
5124  
5125  
5126  
5127  
5128  
5129

Figure 52: Probability distributions for Network(architecture=wide, weights=trained, activation=relu), variance=medium

5130  
5131  
5132  
5133  
5134  
5135  
5136  
5137  
5138

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.938 \times 10^0 \pm 6.5 \times 10^{-5}$	$3.707 \times 10^{+1} \pm 1.3 \times 10^{-3}$
analytic	$+5.670 \times 10^0$	$2.890 \times 10^{+1}$
mean-field	$+5.268 \times 10^0$	$2.730 \times 10^{+1}$
linear	$-2.263 \times 10^{-1}$	$2.322 \times 10^{+2}$
unscented'95	$+8.636 \times 10^0$	$3.085 \times 10^{+1}$
unscented'02	$-5.028 \times 10^{+2}$	$5.054 \times 10^{+5}$

Table 90: Comparison of moments for Network(architecture=wide, weights=trained, activation=relu), variance=large

5143  
5144  
5145  
5146  
5147  
5148  
5149  
5150  
5151

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.875 \times 10^{-1} \pm 4.0 \times 10^{-5}$	0
analytic	$5.658 \times 10^{-1} \pm 3.0 \times 10^{-5}$	$1.523 \times 10^{-2} \pm 4.1 \times 10^{-6}$
mean-field	$5.534 \times 10^{-1} \pm 2.9 \times 10^{-5}$	$2.724 \times 10^{-2} \pm 5.0 \times 10^{-6}$
linear	$3.429 \times 10^0 \pm 5.0 \times 10^{-5}$	$2.228 \times 10^0 \pm 1.1 \times 10^{-4}$
unscented'95	$1.235 \times 10^0 \pm 3.1 \times 10^{-5}$	$1.061 \times 10^{-1} \pm 4.9 \times 10^{-6}$
unscented'02	$2.846 \times 10^{+2} \pm 2.6 \times 10^{-3}$	$1.030 \times 10^{+4} \pm 3.7 \times 10^{-1}$

Table 91: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=relu), variance=large

5152  
5153  
5154  
5155  
5156  
5157  
5158  
5159  
5160  
5161  
5162  
5163  
5164  
5165  
5166  
5167  
5168  
5169  
5170  
5171  
5172  
5173  
5174  
5175

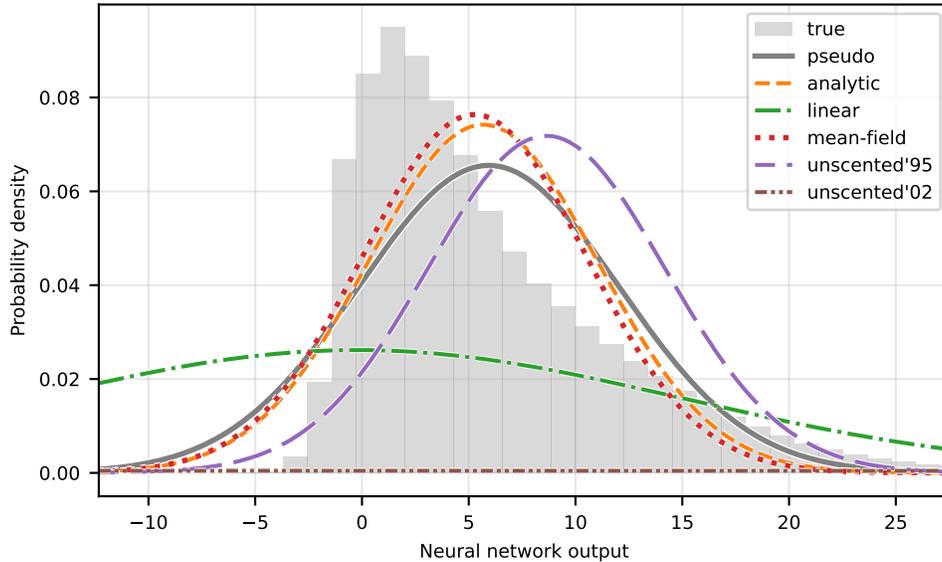


Figure 53: Probability distributions for Network(architecture=wide, weights=trained, activation=relu), variance=large

5176  
5177  
5178  
5179  
5180  
5181  
5182  
5183

5184  
5185  
5186  
5187  
5188  
5189  
5190  
5191  
5192

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.429 \times 10^0 \pm 8.7 \times 10^{-7}$	$2.035 \times 10^{-1} \pm 1.8 \times 10^{-6}$
analytic	$-5.427 \times 10^0$	$1.967 \times 10^{-1}$
mean-field	$-5.403 \times 10^0$	$1.344 \times 10^{-1}$
linear	$-5.335 \times 10^0$	$2.681 \times 10^{-1}$
unscented'95	$-5.439 \times 10^0$	$2.195 \times 10^{-1}$
unscented'02	$-5.335 \times 10^0$	$2.681 \times 10^{-1}$

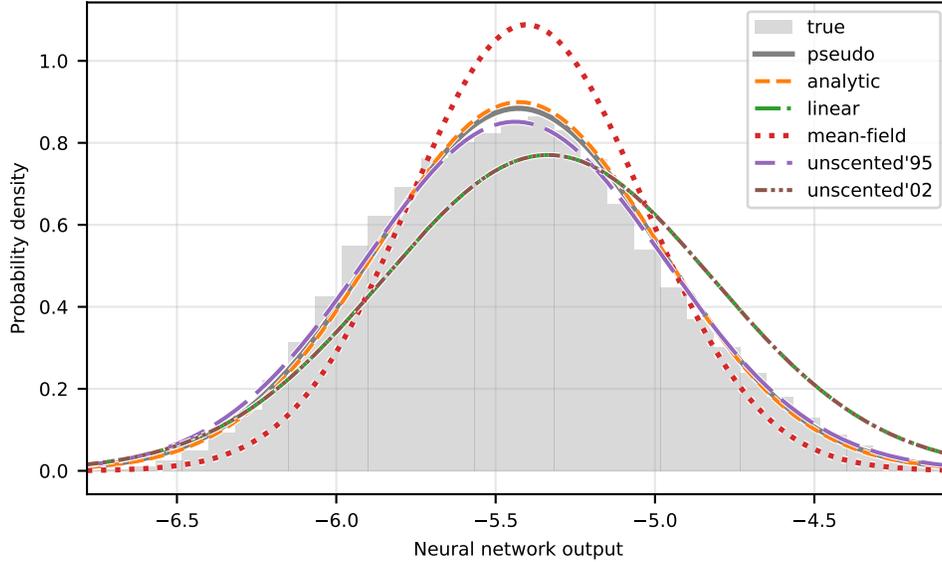
Table 92: Comparison of moments for Network(architecture=wide, weights=initialized, activation=relu residual), variance=small

5193  
5194  
5195  
5196  
5197  
5198  
5199  
5200  
5201  
5202  
5203  
5204  
5205

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.655 \times 10^{-2} \pm 3.6 \times 10^{-6}$	0
analytic	$2.999 \times 10^{-2} \pm 3.1 \times 10^{-6}$	$2.979 \times 10^{-4} \pm 1.5 \times 10^{-7}$
mean-field	$1.121 \times 10^{-1} \pm 2.0 \times 10^{-6}$	$3.923 \times 10^{-2} \pm 1.5 \times 10^{-6}$
linear	$1.608 \times 10^{-1} \pm 1.9 \times 10^{-6}$	$4.236 \times 10^{-2} \pm 1.6 \times 10^{-6}$
unscented'95	$2.274 \times 10^{-2} \pm 3.1 \times 10^{-6}$	$1.712 \times 10^{-3} \pm 3.6 \times 10^{-7}$
unscented'02	$1.608 \times 10^{-1} \pm 1.9 \times 10^{-6}$	$4.236 \times 10^{-2} \pm 1.6 \times 10^{-6}$

Table 93: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=relu residual), variance=small

5206  
5207  
5208  
5209  
5210  
5211  
5212  
5213  
5214  
5215  
5216  
5217  
5218  
5219  
5220  
5221  
5222  
5223  
5224  
5225  
5226  
5227



5228  
5229  
5230  
5231  
5232  
5233  
5234  
5235  
5236  
5237

Figure 54: Probability distributions for Network(architecture=wide, weights=initialized, activation=relu residual), variance=small

5238  
5239  
5240  
5241  
5242  
5243  
5244  
5245  
5246

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.570 \times 10^0 \pm 2.1 \times 10^{-5}$	$8.554 \times 10^0 \pm 1.4 \times 10^{-4}$
analytic	$-6.581 \times 10^0$	$6.679 \times 10^0$
mean-field	$-6.661 \times 10^0$	$8.103 \times 10^0$
linear	$-5.335 \times 10^0$	$2.681 \times 10^{+1}$
unscented'95	$-6.634 \times 10^0$	$5.961 \times 10^0$
unscented'02	$-6.031 \times 10^{+1}$	$6.072 \times 10^{+3}$

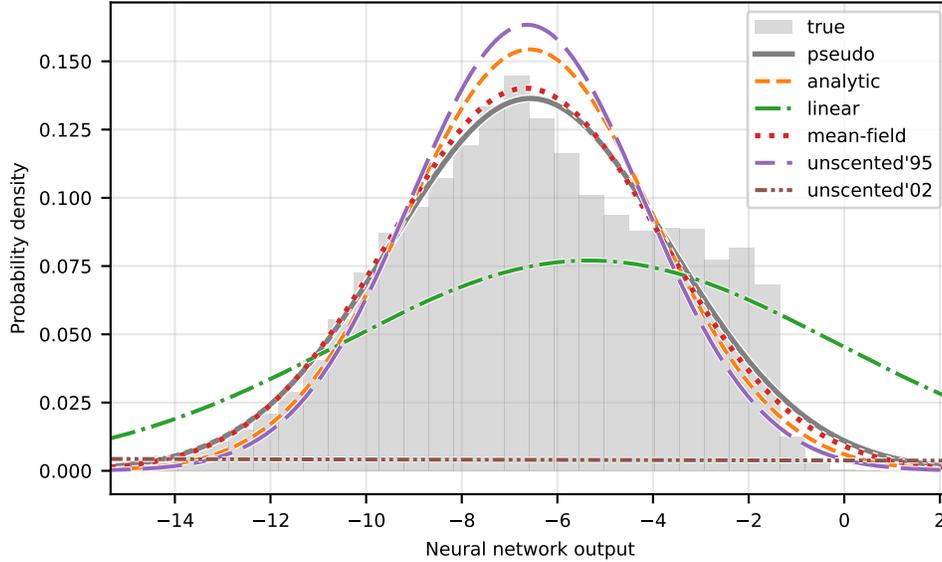
5247 Table 94: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
5248 tion=relu residual), variance=medium  
5249

5250  
5251  
5252  
5253  
5254  
5255  
5256  
5257  
5258  
5259

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$9.062 \times 10^{-2} \pm 2.6 \times 10^{-5}$	0
analytic	$2.083 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$1.413 \times 10^{-2} \pm 1.8 \times 10^{-6}$
mean-field	$1.237 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$1.206 \times 10^{-3} \pm 3.6 \times 10^{-7}$
linear	$1.183 \times 10^0 \pm 1.9 \times 10^{-5}$	$5.852 \times 10^{-1} \pm 1.8 \times 10^{-5}$
unscented'95	$2.634 \times 10^{-1} \pm 1.4 \times 10^{-5}$	$2.924 \times 10^{-2} \pm 2.4 \times 10^{-6}$
unscented'02	$4.361 \times 10^{+1} \pm 1.8 \times 10^{-4}$	$5.200 \times 10^{+2} \pm 8.7 \times 10^{-3}$

5260 Table 95: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
5261 activation=relu residual), variance=medium  
5262

5263  
5264  
5265  
5266  
5267  
5268  
5269  
5270  
5271  
5272  
5273  
5274  
5275  
5276  
5277  
5278  
5279



5280  
5281  
5282  
5283  
5284  
5285  
5286  
5287  
5288  
5289  
5290  
5291

5284 Figure 55: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
5285 tion=relu residual), variance=medium

5292  
5293  
5294  
5295  
5296  
5297  
5298  
5299  
5300  
5301  
5302  
5303  
5304  
5305  
5306  
5307  
5308  
5309  
5310  
5311  
5312  
5313  
5314  
5315  
5316  
5317  
5318  
5319  
5320  
5321  
5322  
5323  
5324  
5325  
5326  
5327  
5328  
5329  
5330  
5331  
5332  
5333  
5334  
5335  
5336  
5337  
5338  
5339  
5340  
5341  
5342  
5343  
5344  
5345

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-3.755 \times 10^{+1} \pm 2.5 \times 10^{-4}$	$7.896 \times 10^{+2} \pm 1.8 \times 10^{-2}$
analytic	$-3.858 \times 10^{+1}$	$5.527 \times 10^{+2}$
mean-field	$-3.508 \times 10^{+1}$	$7.094 \times 10^{+2}$
linear	$-5.335 \times 10^0$	$2.681 \times 10^{+3}$
unscented'95	$-3.855 \times 10^{+1}$	$6.209 \times 10^{+2}$
unscented'02	$-6.382 \times 10^{+3}$	$8.134 \times 10^{+7}$

Table 96: Comparison of moments for Network(architecture=wide, weights=initialized, activation=relu residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.563 \times 10^{-1} \pm 6.6 \times 10^{-5}$	0
analytic	$9.409 \times 10^{-1} \pm 4.3 \times 10^{-5}$	$2.904 \times 10^{-2} \pm 3.2 \times 10^{-6}$
mean-field	$6.105 \times 10^{-1} \pm 5.9 \times 10^{-5}$	$6.612 \times 10^{-3} \pm 1.6 \times 10^{-6}$
linear	$6.261 \times 10^0 \pm 5.5 \times 10^{-5}$	$1.243 \times 10^0 \pm 3.8 \times 10^{-5}$
unscented'95	$8.038 \times 10^{-1} \pm 3.5 \times 10^{-5}$	$1.400 \times 10^{-2} \pm 2.3 \times 10^{-6}$
unscented'02	$1.677 \times 10^{+3} \pm 9.5 \times 10^{-3}$	$7.699 \times 10^{+4} \pm 1.7 \times 10^0$

Table 97: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=relu residual), variance=large

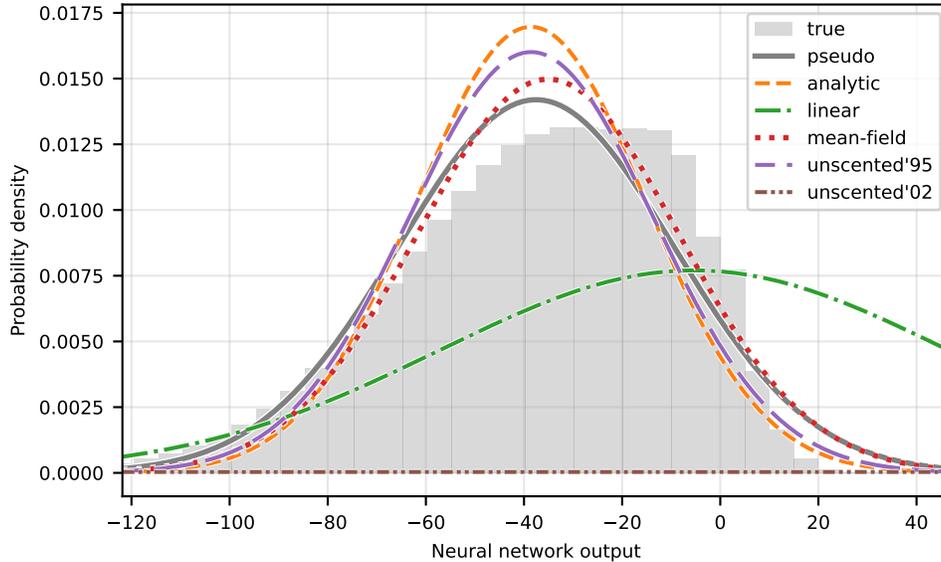


Figure 56: Probability distributions for Network(architecture=wide, weights=initialized, activation=relu residual), variance=large

5346  
5347  
5348  
5349  
5350  
5351  
5352  
5353  
5354  
5355  
5356  
5357  
5358  
5359  
5360  
5361  
5362  
5363  
5364  
5365  
5366  
5367  
5368  
5369  
5370  
5371  
5372  
5373  
5374  
5375  
5376  
5377  
5378  
5379  
5380  
5381  
5382  
5383  
5384  
5385  
5386  
5387  
5388  
5389  
5390  
5391  
5392  
5393  
5394  
5395  
5396  
5397  
5398  
5399

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+4.943 \times 10^{-2} \pm 8.5 \times 10^{-7}$	$5.939 \times 10^{-2} \pm 1.2 \times 10^{-6}$
analytic	$+5.056 \times 10^{-2}$	$5.561 \times 10^{-2}$
mean-field	$+6.542 \times 10^{-2}$	$1.351 \times 10^{-1}$
linear	$+1.866 \times 10^{-1}$	$7.861 \times 10^{-2}$
unscented'95	$+3.719 \times 10^{-2}$	$6.147 \times 10^{-2}$
unscented'02	$+1.866 \times 10^{-1}$	$7.861 \times 10^{-2}$

Table 98: Comparison of moments for Network(architecture=wide, weights=trained, activation=relu residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.563 \times 10^{-2} \pm 5.2 \times 10^{-6}$	0
analytic	$1.974 \times 10^{-2} \pm 3.4 \times 10^{-6}$	$1.072 \times 10^{-3} \pm 6.6 \times 10^{-7}$
mean-field	$1.966 \times 10^{-1} \pm 3.8 \times 10^{-6}$	$2.287 \times 10^{-1} \pm 1.3 \times 10^{-5}$
linear	$2.778 \times 10^{-1} \pm 2.2 \times 10^{-6}$	$1.800 \times 10^{-1} \pm 6.8 \times 10^{-6}$
unscented'95	$2.904 \times 10^{-2} \pm 3.2 \times 10^{-6}$	$1.560 \times 10^{-3} \pm 4.4 \times 10^{-7}$
unscented'02	$2.778 \times 10^{-1} \pm 2.2 \times 10^{-6}$	$1.800 \times 10^{-1} \pm 6.8 \times 10^{-6}$

Table 99: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=relu residual), variance=small

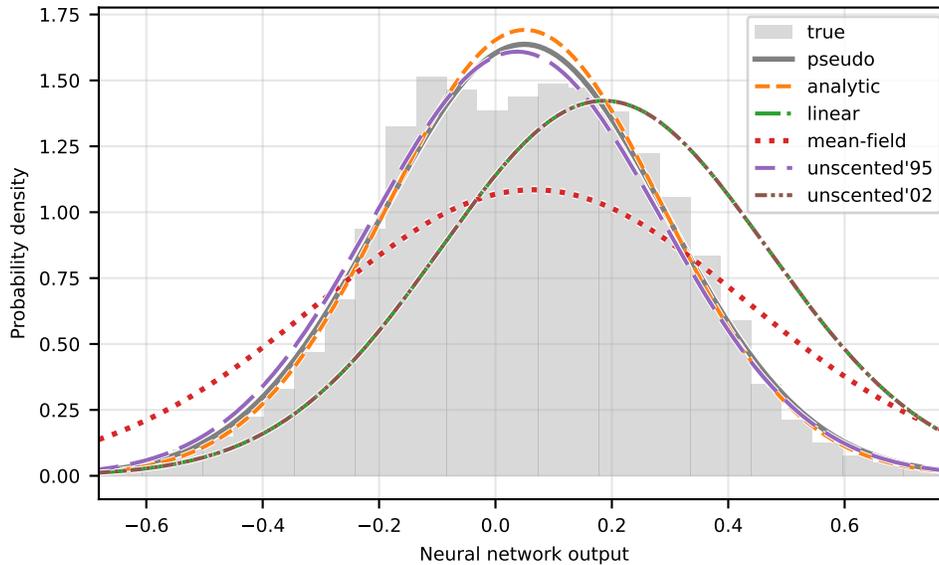


Figure 57: Probability distributions for Network(architecture=wide, weights=trained, activation=relu residual), variance=small

5400  
5401  
5402  
5403  
5404  
5405  
5406  
5407  
5408

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.646 \times 10^{-1} \pm 2.4 \times 10^{-5}$	$2.438 \times 10^0 \pm 1.0 \times 10^{-4}$
analytic	$-6.082 \times 10^{-1}$	$2.762 \times 10^0$
mean-field	$-6.252 \times 10^{-1}$	$8.171 \times 10^0$
linear	$+1.866 \times 10^{-1}$	$7.861 \times 10^0$
unscented'95	$-6.606 \times 10^{-1}$	$1.274 \times 10^0$
unscented'02	$+3.824 \times 10^{+1}$	$2.904 \times 10^{+3}$

Table 100: Comparison of moments for Network(architecture=wide, weights=trained, activation=relu residual), variance=medium

5409  
5410  
5411  
5412  
5413  
5414  
5415  
5416  
5417  
5418  
5419  
5420  
5421

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.621 \times 10^{-2} \pm 3.8 \times 10^{-5}$	0
analytic	$1.021 \times 10^{-1} \pm 3.2 \times 10^{-5}$	$4.452 \times 10^{-3} \pm 2.8 \times 10^{-6}$
mean-field	$8.479 \times 10^{-1} \pm 2.7 \times 10^{-5}$	$5.720 \times 10^{-1} \pm 5.0 \times 10^{-5}$
linear	$9.518 \times 10^{-1} \pm 2.9 \times 10^{-5}$	$6.428 \times 10^{-1} \pm 5.5 \times 10^{-5}$
unscented'95	$2.614 \times 10^{-1} \pm 1.7 \times 10^{-5}$	$8.772 \times 10^{-2} \pm 1.0 \times 10^{-5}$
unscented'02	$4.221 \times 10^{+1} \pm 4.7 \times 10^{-4}$	$9.004 \times 10^{+2} \pm 3.9 \times 10^{-2}$

Table 101: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=relu residual), variance=medium

5422  
5423  
5424  
5425  
5426  
5427  
5428  
5429  
5430  
5431  
5432  
5433  
5434  
5435  
5436  
5437  
5438  
5439  
5440  
5441  
5442  
5443  
5444  
5445

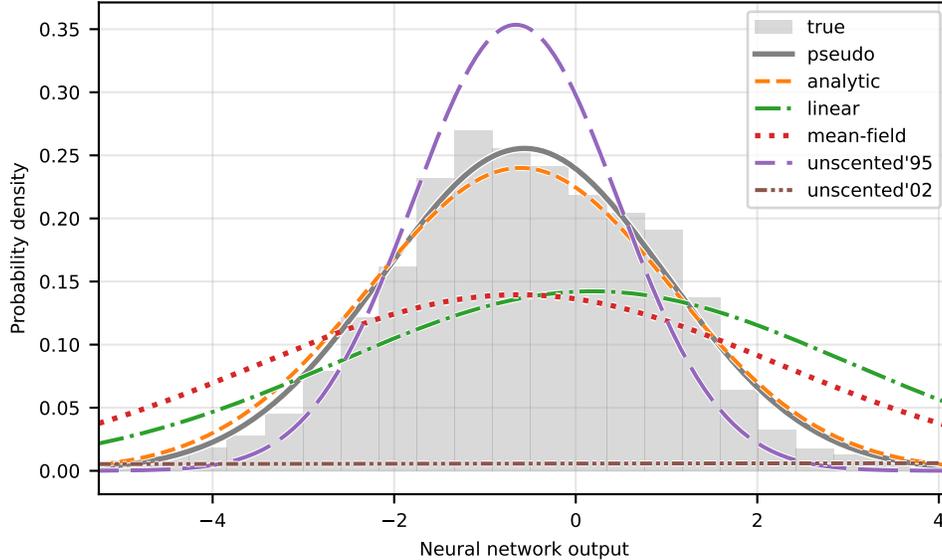


Figure 58: Probability distributions for Network(architecture=wide, weights=trained, activation=relu residual), variance=medium

5446  
5447  
5448  
5449  
5450  
5451  
5452  
5453

5454  
5455  
5456  
5457  
5458  
5459  
5460  
5461  
5462  
5463  
5464  
5465  
5466  
5467  
5468  
5469  
5470  
5471  
5472  
5473  
5474  
5475  
5476  
5477  
5478  
5479  
5480  
5481  
5482  
5483  
5484  
5485  
5486  
5487  
5488  
5489  
5490  
5491  
5492  
5493  
5494  
5495  
5496  
5497  
5498  
5499  
5500  
5501  
5502  
5503  
5504  
5505  
5506  
5507

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-4.090 \times 10^0 \pm 2.5 \times 10^{-4}$	$5.138 \times 10^{+2} \pm 1.2 \times 10^{-2}$
analytic	$-4.302 \times 10^0$	$4.533 \times 10^{+2}$
mean-field	$-5.117 \times 10^{-1}$	$7.093 \times 10^{+2}$
linear	$+1.866 \times 10^{-1}$	$7.861 \times 10^{+2}$
unscented'95	$-2.672 \times 10^0$	$5.364 \times 10^{+2}$
unscented'02	$-3.175 \times 10^{+3}$	$2.017 \times 10^{+7}$

Table 102: Comparison of moments for Network(architecture=wide, weights=trained, activation=relu residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.767 \times 10^{-1} \pm 7.1 \times 10^{-5}$	0
analytic	$5.756 \times 10^{-1} \pm 6.8 \times 10^{-5}$	$3.818 \times 10^{-3} \pm 1.4 \times 10^{-6}$
mean-field	$1.220 \times 10^0 \pm 7.1 \times 10^{-5}$	$4.147 \times 10^{-2} \pm 5.1 \times 10^{-6}$
linear	$1.477 \times 10^0 \pm 6.7 \times 10^{-5}$	$7.014 \times 10^{-2} \pm 7.0 \times 10^{-6}$
unscented'95	$6.572 \times 10^{-1} \pm 7.3 \times 10^{-5}$	$2.426 \times 10^{-3} \pm 9.3 \times 10^{-7}$
unscented'02	$9.296 \times 10^{+2} \pm 5.3 \times 10^{-3}$	$2.941 \times 10^{+4} \pm 6.7 \times 10^{-1}$

Table 103: Comparison of statistical distances for Network(architecture=wide, weights=trained, activation=relu residual), variance=large

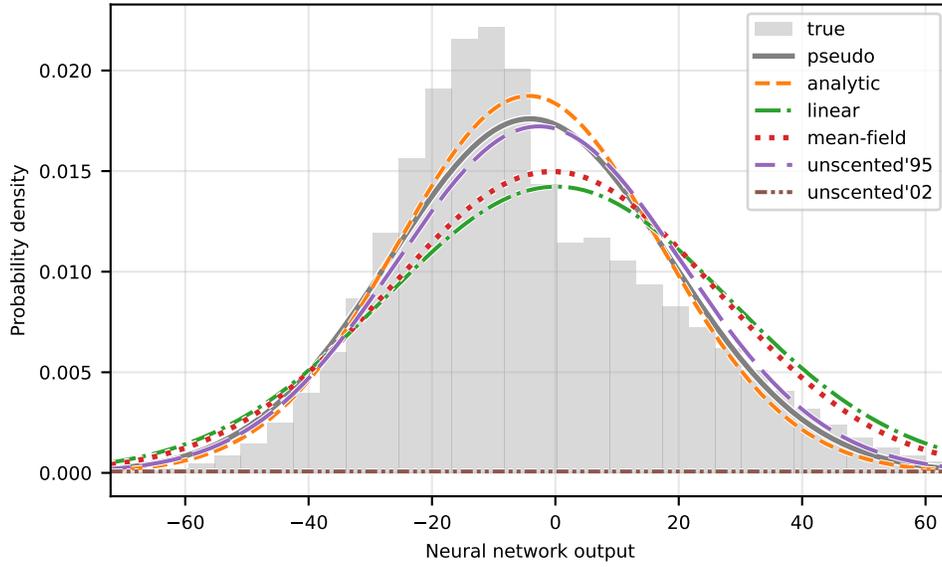


Figure 59: Probability distributions for Network(architecture=wide, weights=trained, activation=relu residual), variance=large

5508  
5509  
5510  
5511  
5512  
5513  
5514  
5515  
5516

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.998 \times 10^{-1} \pm 1.0 \times 10^{-4}$	$4.743 \times 10^{-2} \pm 2.3 \times 10^{-5}$
analytic	$-7.013 \times 10^{-1}$	$4.170 \times 10^{-2}$
mean-field	$-7.007 \times 10^{-1}$	$3.827 \times 10^{-2}$
linear	$-4.078 \times 10^{-1}$	0
unscented'95	$-9.009 \times 10^{-1}$	$8.230 \times 10^{-2}$
unscented'02	$-4.078 \times 10^{-1}$	$4.322 \times 10^{-21}$

5517  
5518  
5519

Table 104: Comparison of moments for Network(architecture=wide, weights=initialized, activation=heaviside), variance=small

5520

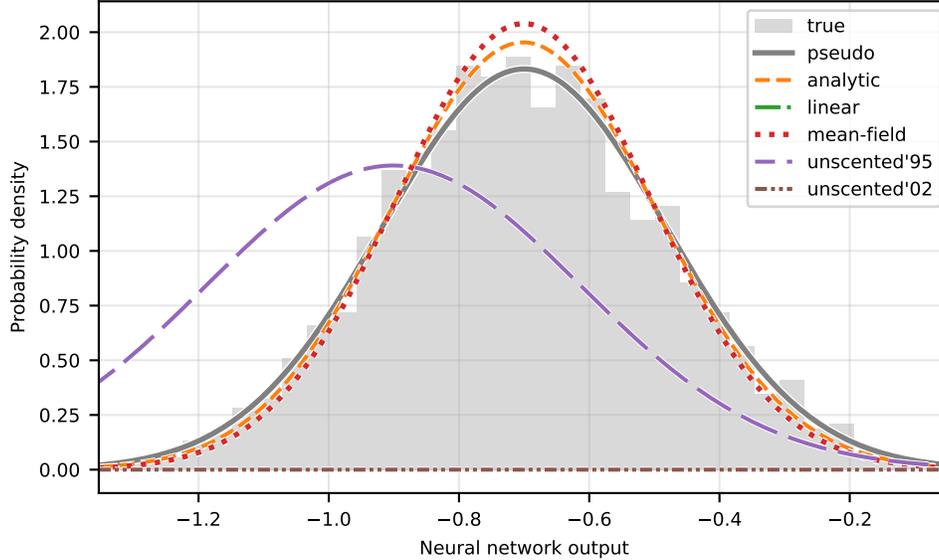
5521  
5522  
5523  
5524  
5525  
5526  
5527  
5528  
5529

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$8.746 \times 10^{-3} \pm 1.0 \times 10^{-4}$	0
analytic	$2.274 \times 10^{-2} \pm 1.1 \times 10^{-4}$	$3.998 \times 10^{-3} \pm 2.9 \times 10^{-5}$
mean-field	$3.737 \times 10^{-2} \pm 1.1 \times 10^{-4}$	$1.074 \times 10^{-2} \pm 4.6 \times 10^{-5}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$4.309 \times 10^{-1} \pm 2.4 \times 10^{-4}$	$5.181 \times 10^{-1} \pm 6.5 \times 10^{-4}$
unscented'02	$6.680 \times 10^{-1} \pm 1.8 \times 10^{-4}$	$2.232 \times 10^{+1} \pm 6.2 \times 10^{-4}$

5530  
5531  
5532

Table 105: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=heaviside), variance=small

5533  
5534  
5535  
5536  
5537  
5538  
5539  
5540  
5541  
5542  
5543  
5544  
5545  
5546  
5547  
5548  
5549



5550  
5551  
5552  
5553  
5554  
5555

Figure 60: Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside), variance=small

5556  
5557  
5558  
5559  
5560  
5561

5562  
5563  
5564  
5565  
5566  
5567  
5568  
5569  
5570

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.270 \times 10^{-1} \pm 1.9 \times 10^{-4}$	$5.427 \times 10^{-2} \pm 6.3 \times 10^{-5}$
analytic	$-6.293 \times 10^{-1}$	$5.005 \times 10^{-2}$
mean-field	$-6.308 \times 10^{-1}$	$4.576 \times 10^{-2}$
linear	$-4.078 \times 10^{-1}$	0
unscented'95	$-6.559 \times 10^{-1}$	$9.895 \times 10^{-2}$
unscented'02	$-1.626 \times 10^{+5}$	$5.287 \times 10^{+10}$

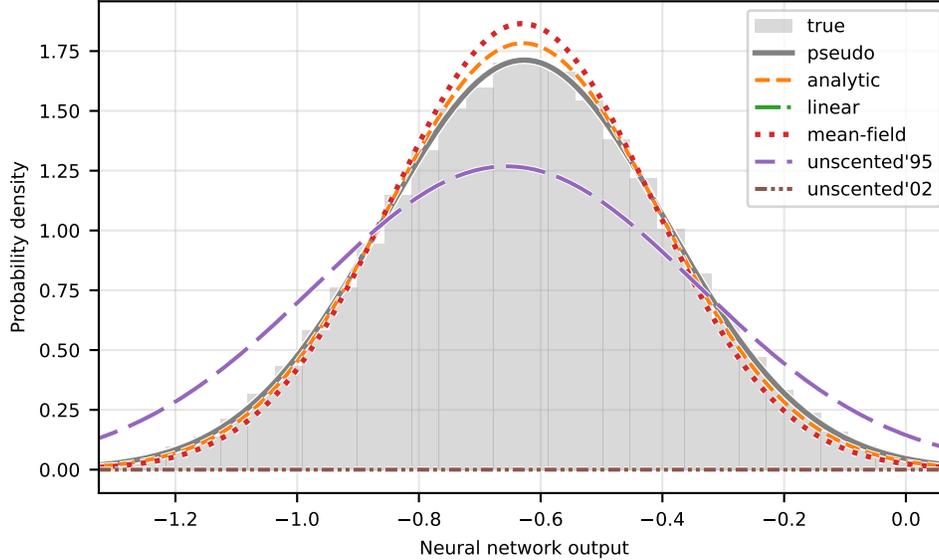
5571 Table 106: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
5572 tion=heaviside), variance=medium  
5573  
5574

5575  
5576  
5577  
5578  
5579  
5580  
5581  
5582  
5583

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.975 \times 10^{-3} \pm 1.1 \times 10^{-4}$	0
analytic	$1.647 \times 10^{-2} \pm 2.3 \times 10^{-4}$	$1.661 \times 10^{-3} \pm 4.5 \times 10^{-5}$
mean-field	$3.274 \times 10^{-2} \pm 2.4 \times 10^{-4}$	$7.035 \times 10^{-3} \pm 9.0 \times 10^{-5}$
linear	— ± —	$\infty \pm \text{—}$
unscented'95	$1.435 \times 10^{-1} \pm 3.4 \times 10^{-4}$	$1.190 \times 10^{-1} \pm 5.1 \times 10^{-4}$
unscented'02	$4.714 \times 10^{+5} \pm 1.4 \times 10^{+2}$	$7.306 \times 10^{+11} \pm 8.5 \times 10^{+8}$

5584 Table 107: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
5585 activation=heaviside), variance=medium  
5586  
5587

5588  
5589  
5590  
5591  
5592  
5593  
5594  
5595  
5596  
5597  
5598  
5599  
5600  
5601  
5602  
5603  
5604  
5605  
5606



5607  
5608  
5609  
5610  
5611  
5612  
5613  
5614  
5615

5608 Figure 61: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
5609 tion=heaviside), variance=medium

5616  
5617  
5618  
5619  
5620  
5621  
5622  
5623  
5624

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.306 \times 10^{-1} \pm 1.6 \times 10^{-4}$	$5.625 \times 10^{-2} \pm 7.4 \times 10^{-5}$
analytic	$-6.308 \times 10^{-1}$	$5.223 \times 10^{-2}$
mean-field	$-6.383 \times 10^{-1}$	$4.932 \times 10^{-2}$
linear	$-4.078 \times 10^{-1}$	0
unscented'95	$-7.735 \times 10^{-1}$	$5.789 \times 10^{-2}$
unscented'02	$-3.475 \times 10^{+5}$	$2.416 \times 10^{+11}$

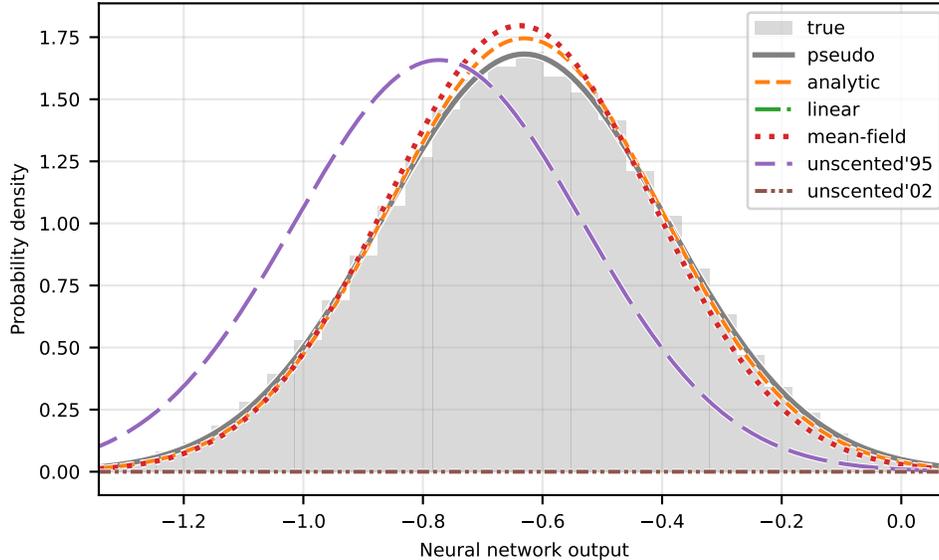
5625 Table 108: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
5626 tion=heaviside), variance=large  
5627

5628  
5629  
5630  
5631  
5632  
5633  
5634  
5635  
5636  
5637

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.546 \times 10^{-3} \pm 1.5 \times 10^{-4}$	0
analytic	$1.505 \times 10^{-2} \pm 2.3 \times 10^{-4}$	$1.355 \times 10^{-3} \pm 4.7 \times 10^{-5}$
mean-field	$2.879 \times 10^{-2} \pm 2.7 \times 10^{-4}$	$4.681 \times 10^{-3} \pm 8.4 \times 10^{-5}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$2.935 \times 10^{-1} \pm 3.4 \times 10^{-4}$	$1.818 \times 10^{-1} \pm 4.7 \times 10^{-4}$
unscented'02	$9.985 \times 10^{+5} \pm 3.3 \times 10^{+2}$	$3.221 \times 10^{+12} \pm 4.2 \times 10^{+9}$

5638 Table 109: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
5639 activation=heaviside), variance=large  
5640

5641  
5642  
5643  
5644  
5645  
5646  
5647  
5648  
5649  
5650  
5651  
5652  
5653  
5654  
5655  
5656  
5657



5662 Figure 62: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
5663 tion=heaviside), variance=large  
5664

5665  
5666  
5667  
5668  
5669

5670  
5671  
5672  
5673  
5674  
5675  
5676  
5677  
5678

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.766 \times 10^0 \pm 1.1 \times 10^{-4}$	$2.197 \times 10^{-1} \pm 1.0 \times 10^{-4}$
analytic	$-2.774 \times 10^0$	$1.893 \times 10^{-1}$
mean-field	$-2.734 \times 10^0$	$1.686 \times 10^{-1}$
linear	$-2.896 \times 10^0$	0
unscented'95	$-2.997 \times 10^0$	$1.551 \times 10^{-1}$
unscented'02	$-2.896 \times 10^0$	$8.872 \times 10^{-20}$

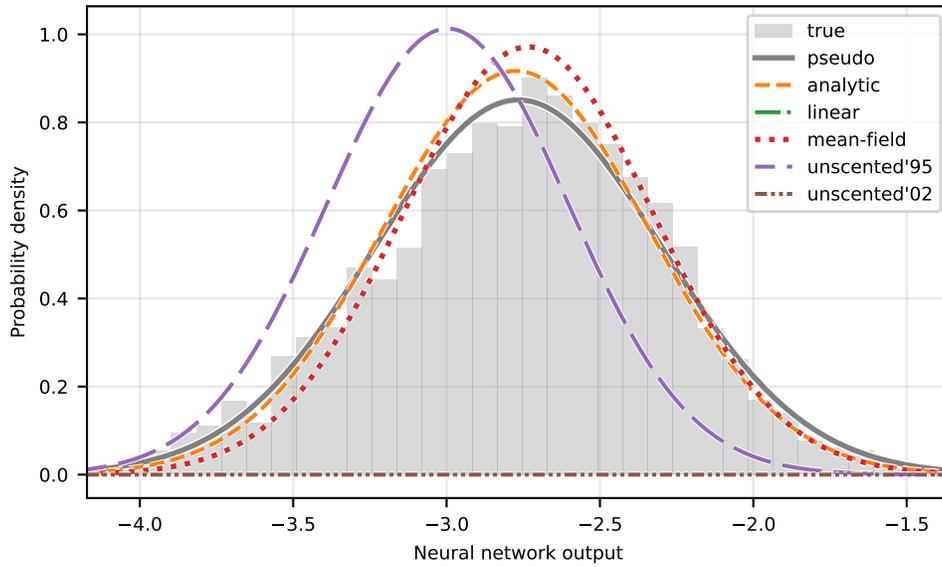
5679 Table 110: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
5680 tion=heaviside residual), variance=small  
5681

5682  
5683  
5684  
5685  
5686  
5687  
5688  
5689  
5690  
5691

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.325 \times 10^{-2} \pm 1.2 \times 10^{-4}$	0
analytic	$5.316 \times 10^{-2} \pm 1.6 \times 10^{-4}$	$5.447 \times 10^{-3} \pm 3.3 \times 10^{-5}$
mean-field	$6.836 \times 10^{-2} \pm 1.4 \times 10^{-4}$	$1.844 \times 10^{-2} \pm 5.0 \times 10^{-5}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$3.367 \times 10^{-1} \pm 1.6 \times 10^{-4}$	$1.481 \times 10^{-1} \pm 1.2 \times 10^{-4}$
unscented'02	$5.758 \times 10^{-1} \pm 9.8 \times 10^{-5}$	$2.072 \times 10^{+1} \pm 2.4 \times 10^{-4}$

5692 Table 111: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
5693 activation=heaviside residual), variance=small  
5694

5695  
5696  
5697  
5698  
5699  
5700  
5701  
5702  
5703  
5704  
5705  
5706  
5707  
5708  
5709  
5710  
5711



5712 Figure 63: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
5713 tion=heaviside residual), variance=small  
5714

5715  
5716  
5717  
5718  
5719  
5720  
5721  
5722  
5723

5724  
5725  
5726  
5727  
5728  
5729  
5730  
5731  
5732  
5733  
5734  
5735  
5736  
5737  
5738  
5739  
5740  
5741  
5742  
5743  
5744  
5745  
5746  
5747  
5748  
5749  
5750  
5751  
5752  
5753  
5754  
5755  
5756  
5757  
5758  
5759  
5760  
5761  
5762  
5763  
5764  
5765  
5766  
5767  
5768  
5769  
5770  
5771  
5772  
5773  
5774  
5775  
5776  
5777

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.330 \times 10^0 \pm 2.3 \times 10^{-4}$	$3.146 \times 10^{-1} \pm 2.3 \times 10^{-4}$
analytic	$-2.338 \times 10^0$	$3.152 \times 10^{-1}$
mean-field	$-2.383 \times 10^0$	$4.215 \times 10^{-1}$
linear	$-2.896 \times 10^0$	0
unscented'95	$-2.077 \times 10^0$	$1.681 \times 10^{-1}$
unscented'02	$+5.462 \times 10^{+4}$	$5.967 \times 10^{+9}$

Table 112: Comparison of moments for Network(architecture=wide, weights=initialized, activation=heaviside residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.107 \times 10^{-3} \pm 1.3 \times 10^{-4}$	0
analytic	$1.199 \times 10^{-2} \pm 3.0 \times 10^{-4}$	$1.223 \times 10^{-4} \pm 6.6 \times 10^{-6}$
mean-field	$1.095 \times 10^{-1} \pm 2.9 \times 10^{-4}$	$2.808 \times 10^{-2} \pm 1.4 \times 10^{-4}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$3.453 \times 10^{-1} \pm 3.1 \times 10^{-4}$	$1.823 \times 10^{-1} \pm 2.1 \times 10^{-4}$
unscented'02	$1.020 \times 10^{+5} \pm 1.9 \times 10^{+1}$	$1.422 \times 10^{+10} \pm 1.0 \times 10^{+7}$

Table 113: Comparison of statistical distances for Network(architecture=wide, weights=initialized, activation=heaviside residual), variance=medium

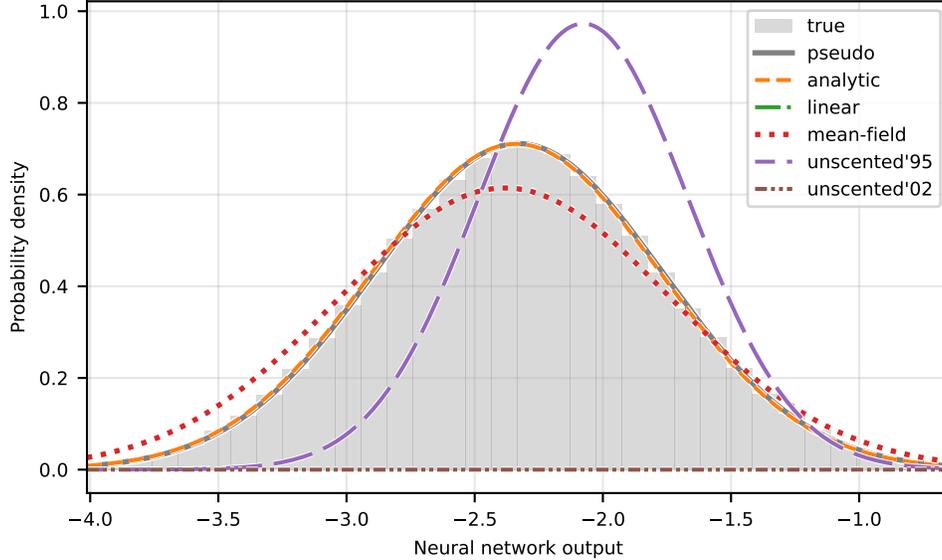


Figure 64: Probability distributions for Network(architecture=wide, weights=initialized, activation=heaviside residual), variance=medium

5778  
5779  
5780  
5781  
5782  
5783  
5784  
5785  
5786

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.018 \times 10^0 \pm 3.0 \times 10^{-4}$	$3.376 \times 10^{-1} \pm 1.9 \times 10^{-4}$
analytic	$-2.035 \times 10^0$	$3.613 \times 10^{-1}$
mean-field	$-2.143 \times 10^0$	$5.100 \times 10^{-1}$
linear	$-2.896 \times 10^0$	0
unscented'95	$-2.078 \times 10^0$	$4.143 \times 10^{-1}$
unscented'02	$-4.102 \times 10^{+4}$	$3.364 \times 10^{+9}$

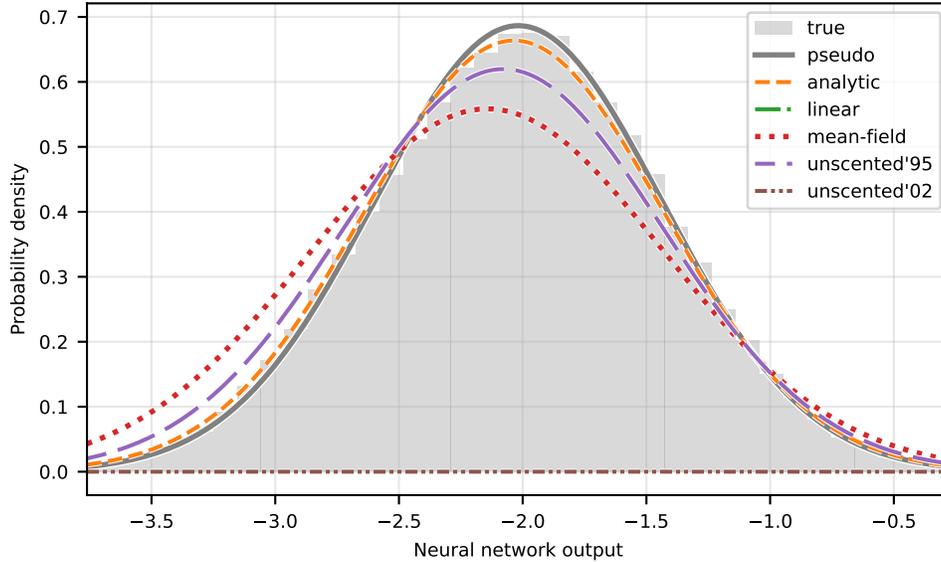
5787 Table 114: Comparison of moments for Network(architecture=wide, weights=initialized, activa-  
5788 tion=heaviside residual), variance=large  
5789

5791  
5792  
5793  
5794  
5795  
5796  
5797  
5798  
5799

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.042 \times 10^{-2} \pm 1.8 \times 10^{-4}$	0
analytic	$3.051 \times 10^{-2} \pm 3.2 \times 10^{-4}$	$1.643 \times 10^{-3} \pm 2.9 \times 10^{-5}$
mean-field	$1.986 \times 10^{-1} \pm 3.1 \times 10^{-4}$	$7.239 \times 10^{-2} \pm 2.2 \times 10^{-4}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$9.586 \times 10^{-2} \pm 3.1 \times 10^{-4}$	$1.669 \times 10^{-2} \pm 9.7 \times 10^{-5}$
unscented'02	$7.529 \times 10^{+4} \pm 1.1 \times 10^{+1}$	$7.473 \times 10^{+9} \pm 4.3 \times 10^{+6}$

5800 Table 115: Comparison of statistical distances for Network(architecture=wide, weights=initialized,  
5801 activation=heaviside residual), variance=large  
5802

5803  
5804  
5805  
5806  
5807  
5808  
5809  
5810  
5811  
5812  
5813  
5814  
5815  
5816  
5817  
5818  
5819  
5820  
5821



5822 Figure 65: Probability distributions for Network(architecture=wide, weights=initialized, activa-  
5823 tion=heaviside residual), variance=large  
5824  
5825  
5826  
5827  
5828  
5829  
5830  
5831

5832  
5833  
5834  
5835  
5836  
5837  
5838  
5839  
5840  
5841  
5842  
5843  
5844  
5845  
5846  
5847  
5848  
5849  
5850  
5851  
5852  
5853  
5854  
5855  
5856  
5857  
5858  
5859  
5860  
5861  
5862  
5863  
5864  
5865  
5866  
5867  
5868  
5869  
5870  
5871  
5872  
5873  
5874  
5875  
5876  
5877  
5878  
5879  
5880  
5881  
5882  
5883  
5884  
5885

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+9.823 \times 10^{-1} \pm 2.3 \times 10^{-10}$	$1.938 \times 10^{-7} \pm 9.1 \times 10^{-13}$
analytic	$+9.823 \times 10^{-1}$	$1.944 \times 10^{-7}$
mean-field	$+9.823 \times 10^{-1}$	$3.562 \times 10^{-7}$
linear	$+9.823 \times 10^{-1}$	$2.069 \times 10^{-7}$
unscented'95	$+9.823 \times 10^{-1}$	$1.956 \times 10^{-7}$
unscented'02	$+9.823 \times 10^{-1}$	$2.069 \times 10^{-7}$

Table 116: Comparison of moments for Network(architecture=deep, weights=initialized, activation=probit), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.046 \times 10^{-4} \pm 1.2 \times 10^{-7}$	0
analytic	$3.010 \times 10^{-4} \pm 1.2 \times 10^{-7}$	$2.877 \times 10^{-6} \pm 6.9 \times 10^{-9}$
mean-field	$5.874 \times 10^{-3} \pm 8.7 \times 10^{-8}$	$1.155 \times 10^{-1} \pm 2.0 \times 10^{-6}$
linear	$5.152 \times 10^{-4} \pm 7.2 \times 10^{-8}$	$1.088 \times 10^{-3} \pm 1.6 \times 10^{-7}$
unscented'95	$2.856 \times 10^{-4} \pm 1.2 \times 10^{-7}$	$2.159 \times 10^{-5} \pm 2.2 \times 10^{-8}$
unscented'02	$5.579 \times 10^{-4} \pm 7.3 \times 10^{-8}$	$1.099 \times 10^{-3} \pm 1.6 \times 10^{-7}$

Table 117: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=probit), variance=small

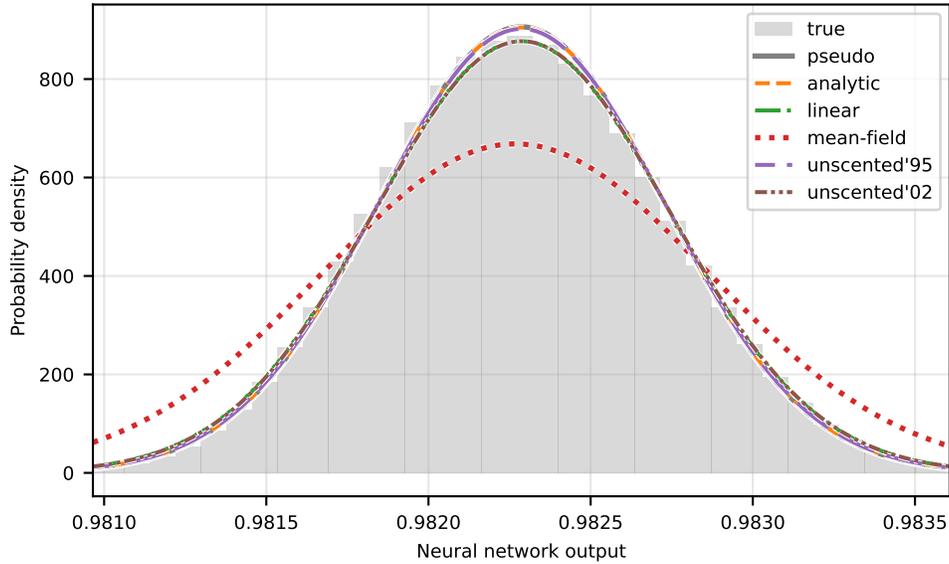


Figure 66: Probability distributions for Network(architecture=deep, weights=initialized, activation=probit), variance=small

5886  
5887  
5888  
5889  
5890  
5891  
5892  
5893  
5894

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+9.817 \times 10^{-1} \pm 8.1 \times 10^{-9}$	$4.914 \times 10^{-6} \pm 5.4 \times 10^{-11}$
analytic	$+9.817 \times 10^{-1}$	$4.609 \times 10^{-6}$
mean-field	$+9.814 \times 10^{-1}$	$1.063 \times 10^{-5}$
linear	$+9.823 \times 10^{-1}$	$2.069 \times 10^{-5}$
unscented'95	$+9.816 \times 10^{-1}$	$4.722 \times 10^{-6}$
unscented'02	$+9.825 \times 10^{-1}$	$2.077 \times 10^{-5}$

5895  
5896  
5897

Table 118: Comparison of moments for Network(architecture=deep, weights=initialized, activation=probit), variance=medium

5898  
5899  
5900

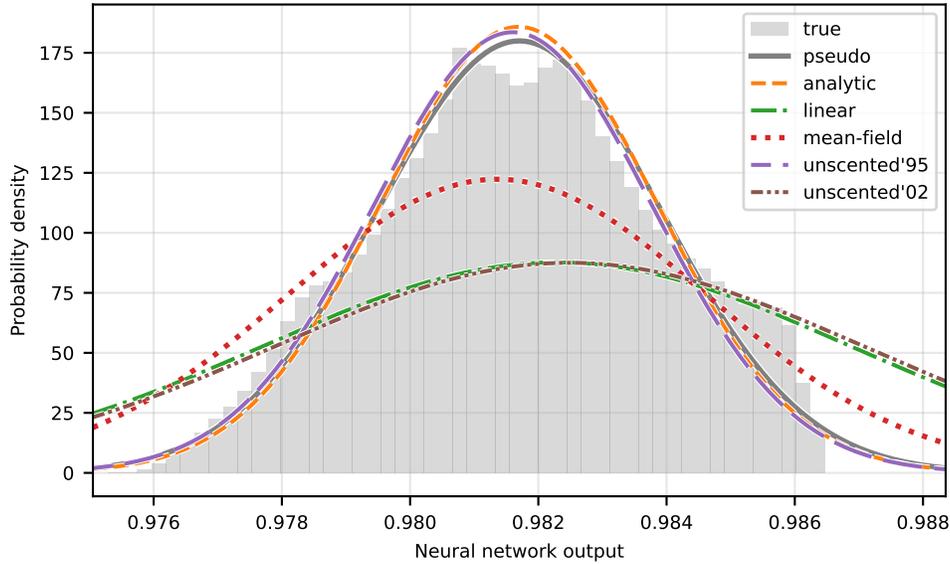
distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.313 \times 10^{-3} \pm 3.1 \times 10^{-7}$	0
analytic	$3.086 \times 10^{-3} \pm 2.6 \times 10^{-7}$	$1.014 \times 10^{-3} \pm 3.3 \times 10^{-7}$
mean-field	$1.787 \times 10^{-2} \pm 3.2 \times 10^{-7}$	$2.081 \times 10^{-1} \pm 6.9 \times 10^{-6}$
linear	$3.995 \times 10^{-2} \pm 5.1 \times 10^{-7}$	$9.210 \times 10^{-1} \pm 1.8 \times 10^{-5}$
unscented'95	$3.355 \times 10^{-3} \pm 3.5 \times 10^{-7}$	$1.433 \times 10^{-3} \pm 1.7 \times 10^{-7}$
unscented'02	$4.112 \times 10^{-2} \pm 5.9 \times 10^{-7}$	$9.546 \times 10^{-1} \pm 1.8 \times 10^{-5}$

5901

5902  
5903  
5904  
5905  
5906  
5907

Table 119: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=probit), variance=medium

5908  
5909  
5910  
5911



5912  
5913  
5914  
5915  
5916  
5917  
5918  
5919  
5920  
5921  
5922  
5923  
5924  
5925  
5926  
5927  
5928  
5929

5930  
5931  
5932  
5933

Figure 67: Probability distributions for Network(architecture=deep, weights=initialized, activation=probit), variance=medium

5934  
5935  
5936  
5937  
5938  
5939

5940  
5941  
5942  
5943  
5944  
5945  
5946  
5947  
5948  
5949  
5950  
5951  
5952  
5953  
5954  
5955  
5956  
5957  
5958  
5959  
5960  
5961  
5962  
5963  
5964  
5965  
5966  
5967  
5968  
5969  
5970  
5971  
5972  
5973  
5974  
5975  
5976  
5977  
5978  
5979  
5980  
5981  
5982  
5983  
5984  
5985  
5986  
5987  
5988  
5989  
5990  
5991  
5992  
5993

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+9.802 \times 10^{-1} \pm 1.1 \times 10^{-7}$	$1.055 \times 10^{-5} \pm 6.2 \times 10^{-10}$
analytic	$+9.802 \times 10^{-1}$	$9.472 \times 10^{-6}$
mean-field	$+9.794 \times 10^{-1}$	$1.949 \times 10^{-5}$
linear	$+9.823 \times 10^{-1}$	$2.069 \times 10^{-3}$
unscented'95	$+9.801 \times 10^{-1}$	$1.202 \times 10^{-5}$
unscented'02	$+1.002 \times 10^0$	$2.841 \times 10^{-3}$

Table 120: Comparison of moments for Network(architecture=deep, weights=initialized, activation=probit), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.341 \times 10^{-3} \pm 1.2 \times 10^{-6}$	0
analytic	$7.496 \times 10^{-3} \pm 1.6 \times 10^{-6}$	$2.916 \times 10^{-3} \pm 3.1 \times 10^{-6}$
mean-field	$2.005 \times 10^{-2} \pm 2.4 \times 10^{-6}$	$1.412 \times 10^{-1} \pm 2.8 \times 10^{-5}$
linear	$5.893 \times 10^{-1} \pm 1.1 \times 10^{-5}$	$9.509 \times 10^{+1} \pm 5.7 \times 10^{-3}$
unscented'95	$5.693 \times 10^{-3} \pm 1.9 \times 10^{-6}$	$4.550 \times 10^{-3} \pm 4.2 \times 10^{-6}$
unscented'02	$7.639 \times 10^{-1} \pm 1.2 \times 10^{-5}$	$1.538 \times 10^{+2} \pm 9.2 \times 10^{-3}$

Table 121: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=probit), variance=large

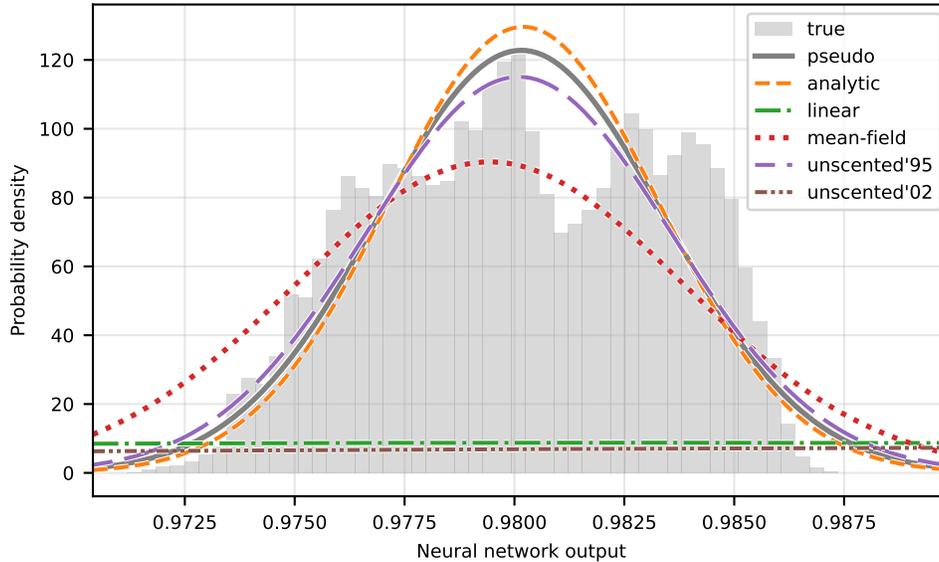


Figure 68: Probability distributions for Network(architecture=deep, weights=initialized, activation=probit), variance=large

5994  
5995  
5996  
5997  
5998  
5999  
6000  
6001  
6002

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.932 \times 10^{-3} \pm 2.9 \times 10^{-7}$	$5.431 \times 10^{-2} \pm 5.2 \times 10^{-7}$
analytic	$-4.371 \times 10^{-3}$	$4.186 \times 10^{-2}$
mean-field	$+3.967 \times 10^{-2}$	$1.949 \times 10^{-6}$
linear	$+8.458 \times 10^{-2}$	$3.872 \times 10^{-2}$
unscented'95	$-5.370 \times 10^{-3}$	$4.734 \times 10^{-2}$
unscented'02	$-7.562 \times 10^{-3}$	$5.570 \times 10^{-2}$

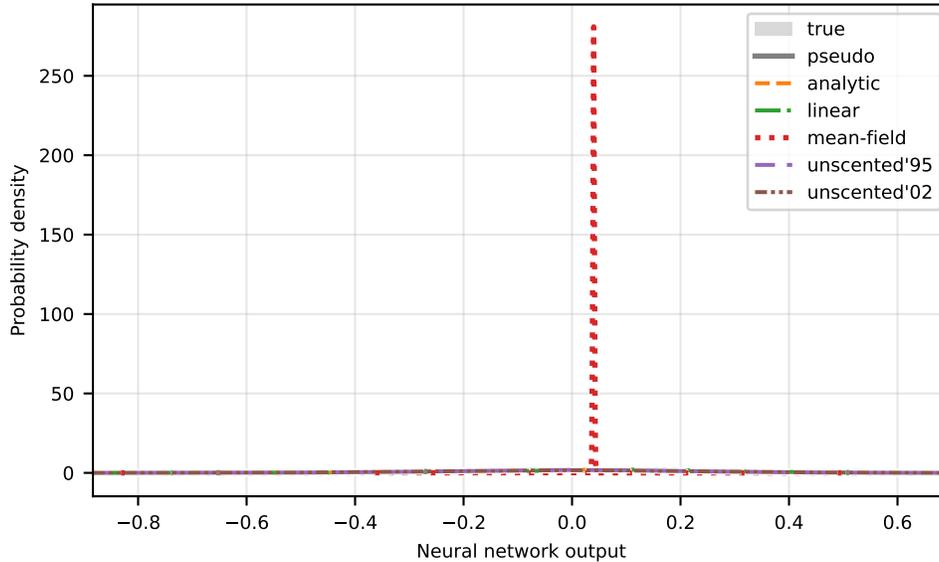
Table 122: Comparison of moments for Network(architecture=deep, weights=trained, activation=probit), variance=small

6003  
6004  
6005  
6006  
6007  
6008  
6009  
6010  
6011  
6012  
6013  
6014  
6015

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.147 \times 10^{-1} \pm 2.3 \times 10^{-6}$	0
analytic	$1.129 \times 10^{-1} \pm 1.9 \times 10^{-6}$	$1.559 \times 10^{-2} \pm 1.1 \times 10^{-6}$
mean-field	$3.569 \times 10^{-1} \pm 1.2 \times 10^{-6}$	$4.634 \times 10^0 \pm 4.8 \times 10^{-6}$
linear	$1.813 \times 10^{-1} \pm 4.8 \times 10^{-7}$	$9.615 \times 10^{-2} \pm 1.1 \times 10^{-6}$
unscented'95	$1.132 \times 10^{-1} \pm 2.2 \times 10^{-6}$	$4.562 \times 10^{-3} \pm 6.1 \times 10^{-7}$
unscented'02	$1.191 \times 10^{-1} \pm 2.6 \times 10^{-6}$	$3.580 \times 10^{-4} \pm 1.4 \times 10^{-7}$

Table 123: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=probit), variance=small

6016  
6017  
6018  
6019  
6020  
6021  
6022  
6023  
6024  
6025  
6026  
6027  
6028  
6029  
6030  
6031  
6032  
6033  
6034  
6035  
6036  
6037



6038  
6039  
6040  
6041  
6042  
6043  
6044  
6045  
6046  
6047

Figure 69: Probability distributions for Network(architecture=deep, weights=trained, activation=probit), variance=small

6048  
6049  
6050  
6051  
6052  
6053  
6054  
6055  
6056  
6057  
6058  
6059  
6060  
6061  
6062  
6063  
6064  
6065  
6066  
6067  
6068  
6069  
6070  
6071  
6072  
6073  
6074  
6075  
6076  
6077  
6078  
6079  
6080  
6081  
6082  
6083  
6084  
6085  
6086  
6087  
6088  
6089  
6090  
6091  
6092  
6093  
6094  
6095  
6096  
6097  
6098  
6099  
6100  
6101

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.472 \times 10^{-1} \pm 1.3 \times 10^{-5}$	$8.328 \times 10^{-1} \pm 2.3 \times 10^{-5}$
analytic	$-8.363 \times 10^{-1}$	$3.840 \times 10^{-1}$
mean-field	$-1.277 \times 10^0$	$5.188 \times 10^{-5}$
linear	$+8.458 \times 10^{-2}$	$3.872 \times 10^0$
unscented'95	$-5.163 \times 10^{-1}$	$1.106 \times 10^0$
unscented'02	$-9.129 \times 10^0$	$1.737 \times 10^{+2}$

Table 124: Comparison of moments for Network(architecture=deep, weights=trained, activation=probit), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.094 \times 10^{-1} \pm 1.2 \times 10^{-5}$	0
analytic	$4.192 \times 10^{-1} \pm 2.0 \times 10^{-5}$	$1.678 \times 10^{-1} \pm 9.0 \times 10^{-6}$
mean-field	$9.681 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$4.661 \times 10^0 \pm 1.4 \times 10^{-5}$
linear	$1.094 \times 10^0 \pm 1.8 \times 10^{-5}$	$1.296 \times 10^0 \pm 6.1 \times 10^{-5}$
unscented'95	$2.019 \times 10^{-1} \pm 2.0 \times 10^{-5}$	$2.269 \times 10^{-2} \pm 4.7 \times 10^{-6}$
unscented'02	$1.262 \times 10^{+1} \pm 9.6 \times 10^{-5}$	$1.453 \times 10^{+2} \pm 4.0 \times 10^{-3}$

Table 125: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=probit), variance=medium

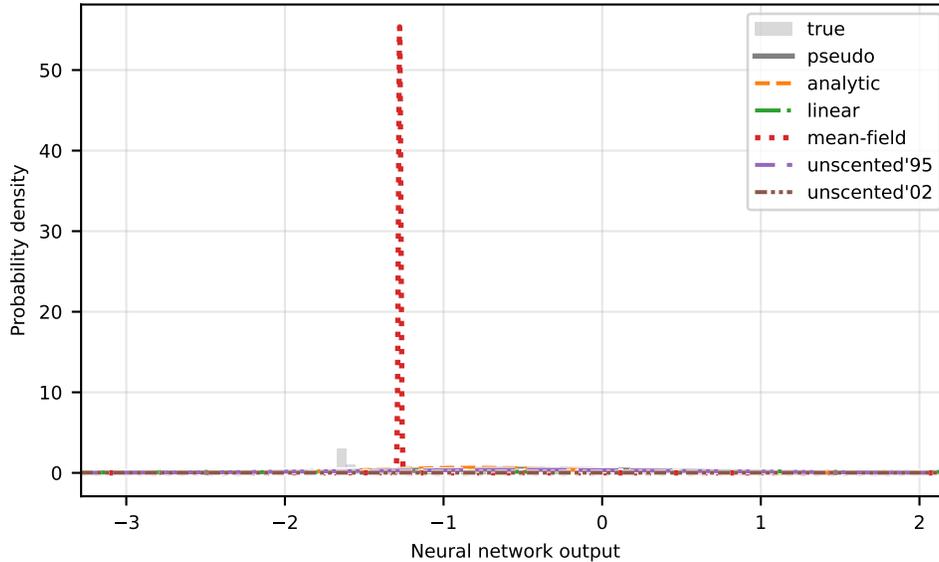


Figure 70: Probability distributions for Network(architecture=deep, weights=trained, activation=probit), variance=medium

6102  
6103  
6104  
6105  
6106  
6107  
6108  
6109

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-8.683 \times 10^{-1} \pm 3.2 \times 10^{-5}$	$8.802 \times 10^{-1} \pm 5.5 \times 10^{-5}$
analytic	$-9.489 \times 10^{-1}$	$3.595 \times 10^{-1}$
mean-field	$-1.559 \times 10^0$	$5.722 \times 10^{-5}$
linear	$+8.458 \times 10^{-2}$	$3.872 \times 10^{+2}$
unscented'95	$-8.587 \times 10^{-1}$	$9.796 \times 10^{-1}$
unscented'02	$-9.212 \times 10^{+2}$	$1.698 \times 10^{+6}$

6110  
6111  
6112  
6113

Table 126: Comparison of moments for Network(architecture=deep, weights=trained, activation=probit), variance=large

6114

6115  
6116  
6117  
6118  
6119  
6120  
6121  
6122

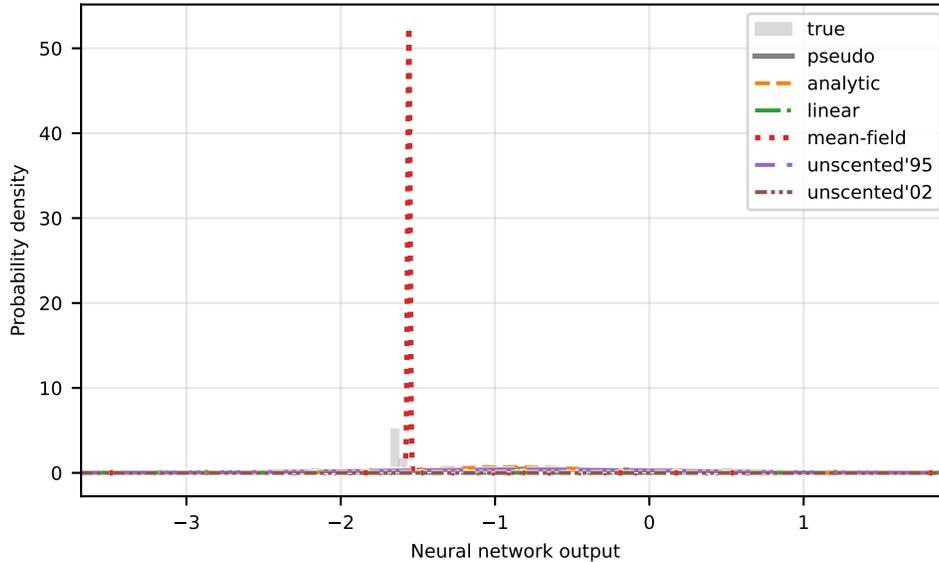
distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.908 \times 10^{-1} \pm 3.4 \times 10^{-5}$	0
analytic	$4.425 \times 10^{-1} \pm 3.4 \times 10^{-5}$	$1.557 \times 10^{-1} \pm 2.0 \times 10^{-5}$
mean-field	$7.749 \times 10^{-1} \pm 2.8 \times 10^{-5}$	$4.591 \times 10^0 \pm 3.5 \times 10^{-5}$
linear	$1.549 \times 10^{+1} \pm 2.6 \times 10^{-4}$	$2.169 \times 10^{+2} \pm 1.4 \times 10^{-2}$
unscented'95	$3.925 \times 10^{-1} \pm 3.4 \times 10^{-5}$	$3.018 \times 10^{-3} \pm 3.7 \times 10^{-6}$
unscented'02	$1.330 \times 10^{+3} \pm 2.1 \times 10^{-2}$	$1.446 \times 10^{+6} \pm 9.0 \times 10^{+1}$

6123  
6124  
6125  
6126

Table 127: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=probit), variance=large

6127

6128  
6129  
6130  
6131  
6132  
6133  
6134  
6135  
6136  
6137  
6138  
6139  
6140  
6141  
6142  
6143  
6144  
6145



6146  
6147  
6148  
6149

Figure 71: Probability distributions for Network(architecture=deep, weights=trained, activation=probit), variance=large

6150  
6151  
6152  
6153  
6154  
6155

6156  
6157  
6158  
6159  
6160  
6161  
6162  
6163  
6164

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+7.399 \times 10^0 \pm 3.2 \times 10^{-6}$	$4.314 \times 10^{-1} \pm 9.8 \times 10^{-6}$
analytic	$+7.409 \times 10^0$	$3.469 \times 10^{-1}$
mean-field	$+7.582 \times 10^0$	$7.771 \times 10^{-1}$
linear	$+7.650 \times 10^0$	$7.550 \times 10^{-1}$
unscented'95	$+7.308 \times 10^0$	$1.613 \times 10^{-1}$
unscented'02	$+6.993 \times 10^0$	$1.616 \times 10^0$

6165  
6166  
6167

Table 128: Comparison of moments for Network(architecture=deep, weights=initialized, activation=probit residual), variance=small

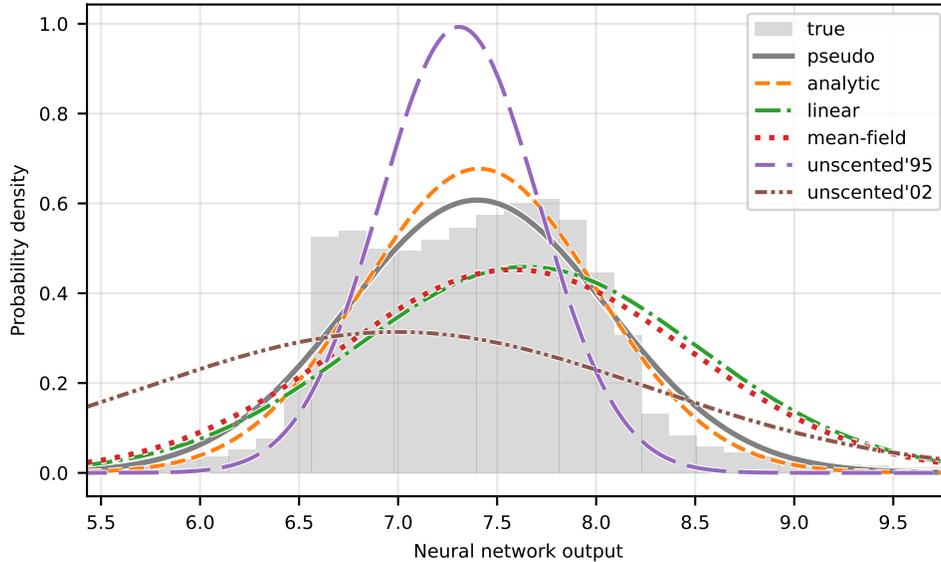
6168  
6169  
6170  
6171  
6172  
6173  
6174  
6175  
6176  
6177

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.736 \times 10^{-2} \pm 1.3 \times 10^{-5}$	0
analytic	$7.097 \times 10^{-2} \pm 1.1 \times 10^{-5}$	$1.115 \times 10^{-2} \pm 2.2 \times 10^{-6}$
mean-field	$3.262 \times 10^{-1} \pm 9.6 \times 10^{-6}$	$1.452 \times 10^{-1} \pm 1.0 \times 10^{-5}$
linear	$3.753 \times 10^{-1} \pm 1.1 \times 10^{-5}$	$1.678 \times 10^{-1} \pm 1.0 \times 10^{-5}$
unscented'95	$2.501 \times 10^{-1} \pm 4.8 \times 10^{-6}$	$1.885 \times 10^{-1} \pm 6.9 \times 10^{-6}$
unscented'02	$7.446 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$9.041 \times 10^{-1} \pm 3.6 \times 10^{-5}$

6178  
6179  
6180

Table 129: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=probit residual), variance=small

6181  
6182  
6183  
6184  
6185  
6186  
6187  
6188  
6189  
6190  
6191  
6192  
6193  
6194  
6195  
6196  
6197  
6198  
6199  
6200



6201  
6202  
6203  
6204  
6205  
6206  
6207  
6208  
6209

Figure 72: Probability distributions for Network(architecture=deep, weights=initialized, activation=probit residual), variance=small

6210  
6211  
6212  
6213  
6214  
6215  
6216  
6217  
6218  
6219  
6220  
6221  
6222  
6223  
6224  
6225  
6226  
6227  
6228  
6229  
6230  
6231  
6232  
6233  
6234  
6235  
6236  
6237  
6238  
6239  
6240  
6241  
6242  
6243  
6244  
6245  
6246  
6247  
6248  
6249  
6250  
6251  
6252  
6253  
6254  
6255  
6256  
6257  
6258  
6259  
6260  
6261  
6262  
6263

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+6.488 \times 10^0 \pm 6.5 \times 10^{-5}$	$8.614 \times 10^0 \pm 4.0 \times 10^{-4}$
analytic	$+6.419 \times 10^0$	$7.410 \times 10^0$
mean-field	$+6.860 \times 10^0$	$8.171 \times 10^0$
linear	$+7.650 \times 10^0$	$7.550 \times 10^{+1}$
unscented'95	$+6.199 \times 10^0$	$4.770 \times 10^0$
unscented'02	$-5.797 \times 10^{+1}$	$8.688 \times 10^{+3}$

Table 130: Comparison of moments for Network(architecture=deep, weights=initialized, activation=probit residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.098 \times 10^{-1} \pm 4.8 \times 10^{-5}$	0
analytic	$2.348 \times 10^{-1} \pm 4.2 \times 10^{-5}$	$5.680 \times 10^{-3} \pm 3.0 \times 10^{-6}$
mean-field	$2.172 \times 10^{-1} \pm 3.8 \times 10^{-5}$	$8.719 \times 10^{-3} \pm 3.3 \times 10^{-6}$
linear	$2.719 \times 10^0 \pm 7.2 \times 10^{-5}$	$2.875 \times 10^0 \pm 1.8 \times 10^{-4}$
unscented'95	$4.377 \times 10^{-1} \pm 3.4 \times 10^{-5}$	$7.725 \times 10^{-2} \pm 9.3 \times 10^{-6}$
unscented'02	$5.243 \times 10^{+1} \pm 6.3 \times 10^{-4}$	$7.415 \times 10^{+2} \pm 3.5 \times 10^{-2}$

Table 131: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=probit residual), variance=medium

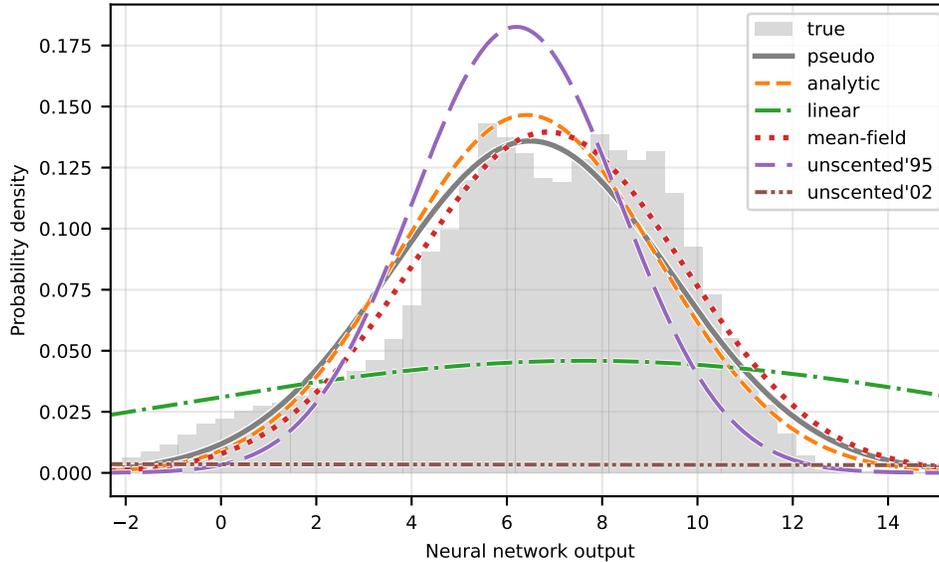


Figure 73: Probability distributions for Network(architecture=deep, weights=initialized, activation=probit residual), variance=medium

6264  
6265  
6266  
6267  
6268  
6269  
6270  
6271  
6272  
6273  
6274  
6275  
6276  
6277  
6278  
6279  
6280  
6281  
6282  
6283  
6284  
6285  
6286  
6287  
6288  
6289  
6290  
6291  
6292  
6293  
6294  
6295  
6296  
6297  
6298  
6299  
6300  
6301  
6302  
6303  
6304  
6305  
6306  
6307  
6308  
6309  
6310  
6311  
6312  
6313  
6314  
6315  
6316  
6317

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+3.429 \times 10^0 \pm 3.0 \times 10^{-4}$	$1.417 \times 10^{+1} \pm 1.8 \times 10^{-3}$
analytic	$+3.589 \times 10^0$	$1.220 \times 10^{+1}$
mean-field	$+4.069 \times 10^0$	$1.242 \times 10^{+1}$
linear	$+7.650 \times 10^0$	$7.550 \times 10^{+3}$
unscented'95	$+3.591 \times 10^0$	$6.302 \times 10^0$
unscented'02	$-6.533 \times 10^{+3}$	$8.557 \times 10^{+7}$

Table 132: Comparison of moments for Network(architecture=deep, weights=initialized, activation=probit residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.358 \times 10^{-2} \pm 1.4 \times 10^{-4}$	0
analytic	$1.549 \times 10^{-1} \pm 1.2 \times 10^{-4}$	$6.207 \times 10^{-3} \pm 8.1 \times 10^{-6}$
mean-field	$3.305 \times 10^{-1} \pm 1.6 \times 10^{-4}$	$1.857 \times 10^{-2} \pm 1.3 \times 10^{-5}$
linear	$3.423 \times 10^{+1} \pm 1.2 \times 10^{-3}$	$2.635 \times 10^{+2} \pm 3.4 \times 10^{-2}$
unscented'95	$5.266 \times 10^{-1} \pm 1.2 \times 10^{-4}$	$1.284 \times 10^{-1} \pm 3.4 \times 10^{-5}$
unscented'02	$4.715 \times 10^{+3} \pm 1.5 \times 10^{-1}$	$4.528 \times 10^{+6} \pm 5.7 \times 10^{+2}$

Table 133: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=probit residual), variance=large

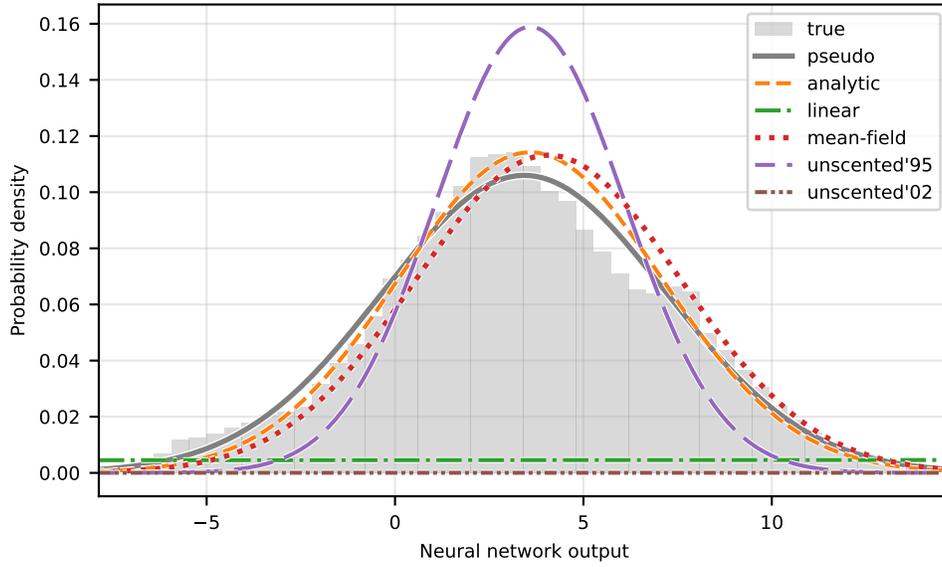


Figure 74: Probability distributions for Network(architecture=deep, weights=initialized, activation=probit residual), variance=large

6318  
6319  
6320  
6321  
6322  
6323  
6324  
6325  
6326  
6327  
6328  
6329  
6330  
6331  
6332  
6333  
6334  
6335  
6336  
6337  
6338  
6339  
6340  
6341  
6342  
6343  
6344  
6345  
6346  
6347  
6348  
6349  
6350  
6351  
6352  
6353  
6354  
6355  
6356  
6357  
6358  
6359  
6360  
6361  
6362  
6363  
6364  
6365  
6366  
6367  
6368  
6369  
6370  
6371

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+4.098 \times 10^{-1} \pm 3.1 \times 10^{-6}$	$3.397 \times 10^{-1} \pm 1.1 \times 10^{-5}$
analytic	$+3.878 \times 10^{-1}$	$3.891 \times 10^{-1}$
mean-field	$+5.497 \times 10^{-1}$	$8.286 \times 10^{-1}$
linear	$+6.993 \times 10^{-1}$	$4.761 \times 10^{-1}$
unscented'95	$+2.598 \times 10^{-1}$	$3.086 \times 10^{-1}$
unscented'02	$-2.036 \times 10^{-1}$	$2.107 \times 10^0$

Table 134: Comparison of moments for Network(architecture=deep, weights=trained, activation=probit residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.032 \times 10^{-2} \pm 1.2 \times 10^{-5}$	0
analytic	$7.162 \times 10^{-2} \pm 9.2 \times 10^{-6}$	$5.537 \times 10^{-3} \pm 2.3 \times 10^{-6}$
mean-field	$3.717 \times 10^{-1} \pm 8.8 \times 10^{-6}$	$3.026 \times 10^{-1} \pm 2.5 \times 10^{-5}$
linear	$3.796 \times 10^{-1} \pm 5.8 \times 10^{-6}$	$1.554 \times 10^{-1} \pm 1.2 \times 10^{-5}$
unscented'95	$2.047 \times 10^{-1} \pm 9.2 \times 10^{-6}$	$3.535 \times 10^{-2} \pm 1.6 \times 10^{-6}$
unscented'02	$1.141 \times 10^0 \pm 1.6 \times 10^{-5}$	$2.242 \times 10^0 \pm 1.0 \times 10^{-4}$

Table 135: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=probit residual), variance=small

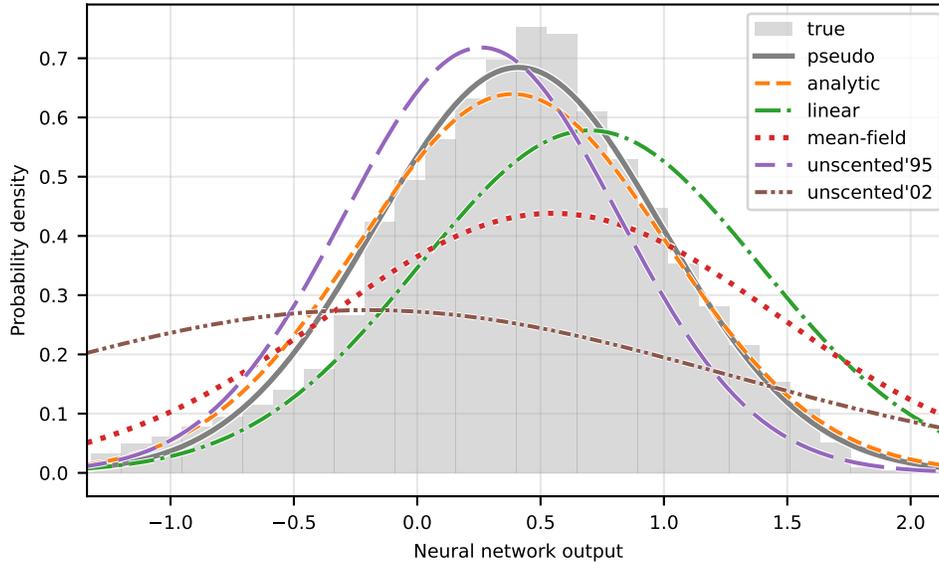


Figure 75: Probability distributions for Network(architecture=deep, weights=trained, activation=probit residual), variance=small

6372  
6373  
6374  
6375  
6376  
6377  
6378  
6379  
6380  
6381  
6382  
6383  
6384  
6385  
6386  
6387  
6388  
6389  
6390  
6391  
6392  
6393  
6394  
6395  
6396  
6397  
6398  
6399  
6400  
6401  
6402  
6403  
6404  
6405  
6406  
6407  
6408  
6409  
6410  
6411  
6412  
6413  
6414  
6415  
6416  
6417  
6418  
6419  
6420  
6421  
6422  
6423  
6424  
6425

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.459 \times 10^0 \pm 5.9 \times 10^{-5}$	$5.535 \times 10^0 \pm 2.8 \times 10^{-4}$
analytic	$+2.739 \times 10^0$	$6.167 \times 10^0$
mean-field	$+3.668 \times 10^0$	$8.286 \times 10^0$
linear	$+6.993 \times 10^{-1}$	$4.761 \times 10^{+1}$
unscented'95	$+2.535 \times 10^0$	$4.672 \times 10^0$
unscented'02	$-8.959 \times 10^{+1}$	$1.635 \times 10^{+4}$

Table 136: Comparison of moments for Network(architecture=deep, weights=trained, activation=probit residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.405 \times 10^{-1} \pm 5.3 \times 10^{-5}$	0
analytic	$2.169 \times 10^{-1} \pm 3.4 \times 10^{-5}$	$1.010 \times 10^{-2} \pm 4.0 \times 10^{-6}$
mean-field	$8.063 \times 10^{-1} \pm 3.5 \times 10^{-5}$	$1.788 \times 10^{-1} \pm 2.1 \times 10^{-5}$
linear	$2.477 \times 10^0 \pm 7.8 \times 10^{-5}$	$3.005 \times 10^0 \pm 2.1 \times 10^{-4}$
unscented'95	$1.861 \times 10^{-1} \pm 4.6 \times 10^{-5}$	$7.311 \times 10^{-3} \pm 4.1 \times 10^{-6}$
unscented'02	$8.207 \times 10^{+1} \pm 1.1 \times 10^{-3}$	$2.238 \times 10^{+3} \pm 1.1 \times 10^{-1}$

Table 137: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=probit residual), variance=medium

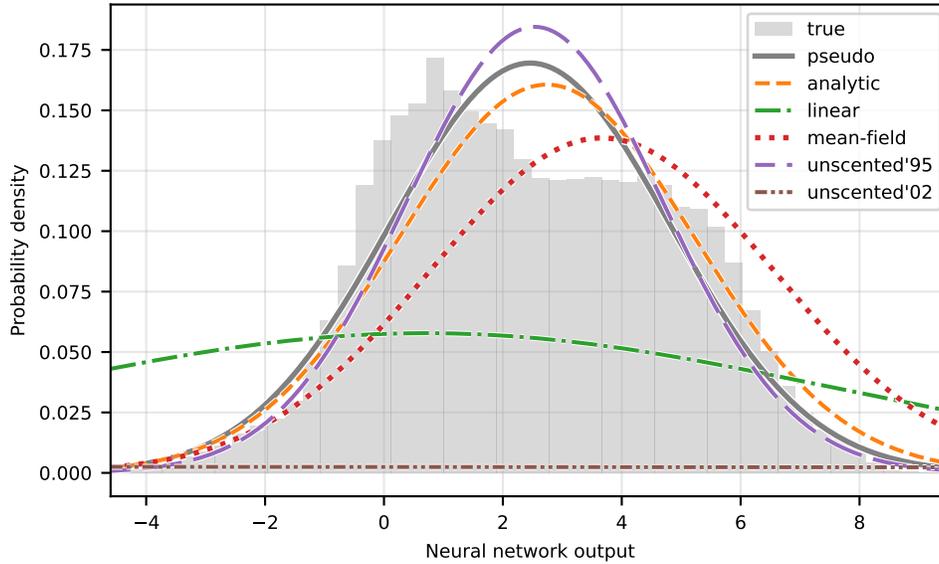


Figure 76: Probability distributions for Network(architecture=deep, weights=trained, activation=probit residual), variance=medium

6426  
6427  
6428  
6429  
6430  
6431  
6432  
6433  
6434  
6435  
6436  
6437  
6438  
6439  
6440  
6441  
6442  
6443  
6444  
6445  
6446  
6447  
6448  
6449  
6450  
6451  
6452  
6453  
6454  
6455  
6456  
6457  
6458  
6459  
6460  
6461  
6462  
6463  
6464  
6465  
6466  
6467  
6468  
6469  
6470  
6471  
6472  
6473  
6474  
6475  
6476  
6477  
6478  
6479

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.793 \times 10^0 \pm 2.6 \times 10^{-4}$	$9.967 \times 10^0 \pm 1.7 \times 10^{-3}$
analytic	$+1.831 \times 10^0$	$1.005 \times 10^{+1}$
mean-field	$+2.630 \times 10^0$	$1.248 \times 10^{+1}$
linear	$+6.993 \times 10^{-1}$	$4.761 \times 10^{+3}$
unscented'95	$+2.269 \times 10^0$	$2.237 \times 10^0$
unscented'02	$-8.971 \times 10^{+3}$	$1.610 \times 10^{+8}$

Table 138: Comparison of moments for Network(architecture=deep, weights=trained, activation=probit residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.560 \times 10^{-2} \pm 1.5 \times 10^{-4}$	0
analytic	$4.399 \times 10^{-2} \pm 1.5 \times 10^{-4}$	$9.368 \times 10^{-5} \pm 1.4 \times 10^{-6}$
mean-field	$4.740 \times 10^{-1} \pm 1.5 \times 10^{-4}$	$4.881 \times 10^{-2} \pm 3.9 \times 10^{-5}$
linear	$2.957 \times 10^{+1} \pm 1.3 \times 10^{-3}$	$2.353 \times 10^{+2} \pm 4.1 \times 10^{-2}$
unscented'95	$7.773 \times 10^{-1} \pm 1.2 \times 10^{-4}$	$3.707 \times 10^{-1} \pm 6.2 \times 10^{-5}$
unscented'02	$7.065 \times 10^{+3} \pm 3.0 \times 10^{-1}$	$1.212 \times 10^{+7} \pm 2.1 \times 10^{+3}$

Table 139: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=probit residual), variance=large

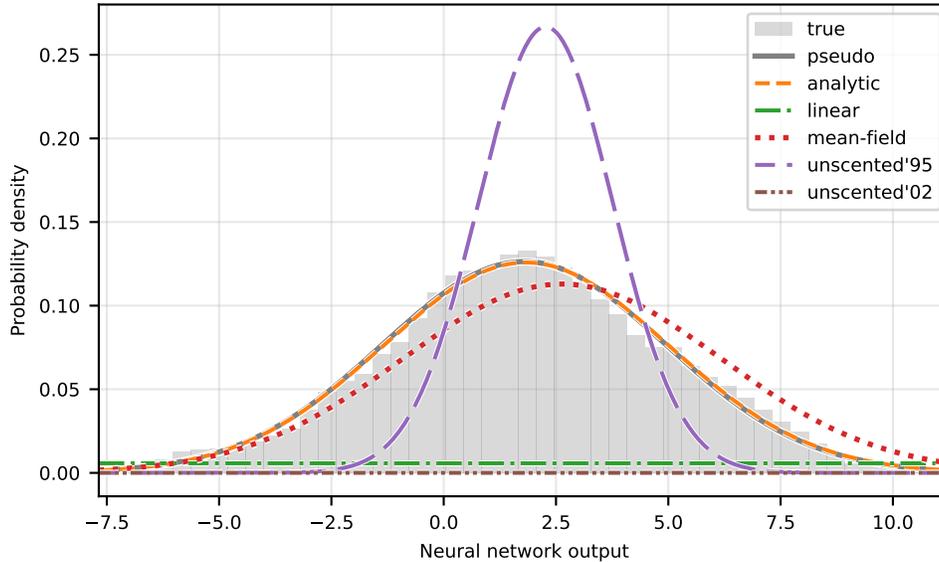


Figure 77: Probability distributions for Network(architecture=deep, weights=trained, activation=probit residual), variance=large

6480  
6481  
6482  
6483  
6484  
6485  
6486  
6487  
6488  
6489  
6490  
6491  
6492  
6493  
6494  
6495  
6496  
6497  
6498  
6499  
6500  
6501  
6502  
6503  
6504  
6505  
6506  
6507  
6508  
6509  
6510  
6511  
6512  
6513  
6514  
6515  
6516  
6517  
6518  
6519  
6520  
6521  
6522  
6523  
6524  
6525  
6526  
6527  
6528  
6529  
6530  
6531  
6532  
6533

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+8.681 \times 10^{-2} \pm 1.7 \times 10^{-7}$	$1.227 \times 10^{-2} \pm 9.2 \times 10^{-8}$
analytic	$+8.729 \times 10^{-2}$	$1.224 \times 10^{-2}$
mean-field	$+7.827 \times 10^{-2}$	$7.985 \times 10^{-3}$
linear	$+1.006 \times 10^{-1}$	$1.994 \times 10^{-2}$
unscented'95	$+8.267 \times 10^{-2}$	$1.311 \times 10^{-2}$
unscented'02	$+8.112 \times 10^{-2}$	$2.070 \times 10^{-2}$

Table 140: Comparison of moments for Network(architecture=deep, weights=initialized, activation=sine), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$7.001 \times 10^{-2} \pm 3.1 \times 10^{-6}$	0
analytic	$6.975 \times 10^{-2} \pm 3.3 \times 10^{-6}$	$1.113 \times 10^{-5} \pm 1.1 \times 10^{-8}$
mean-field	$9.025 \times 10^{-2} \pm 3.3 \times 10^{-6}$	$4.323 \times 10^{-2} \pm 1.3 \times 10^{-6}$
linear	$8.098 \times 10^{-2} \pm 3.8 \times 10^{-6}$	$7.746 \times 10^{-2} \pm 2.4 \times 10^{-6}$
unscented'95	$7.278 \times 10^{-2} \pm 3.1 \times 10^{-6}$	$1.795 \times 10^{-3} \pm 2.5 \times 10^{-7}$
unscented'02	$1.092 \times 10^{-1} \pm 1.6 \times 10^{-6}$	$8.327 \times 10^{-2} \pm 2.6 \times 10^{-6}$

Table 141: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=sine), variance=small

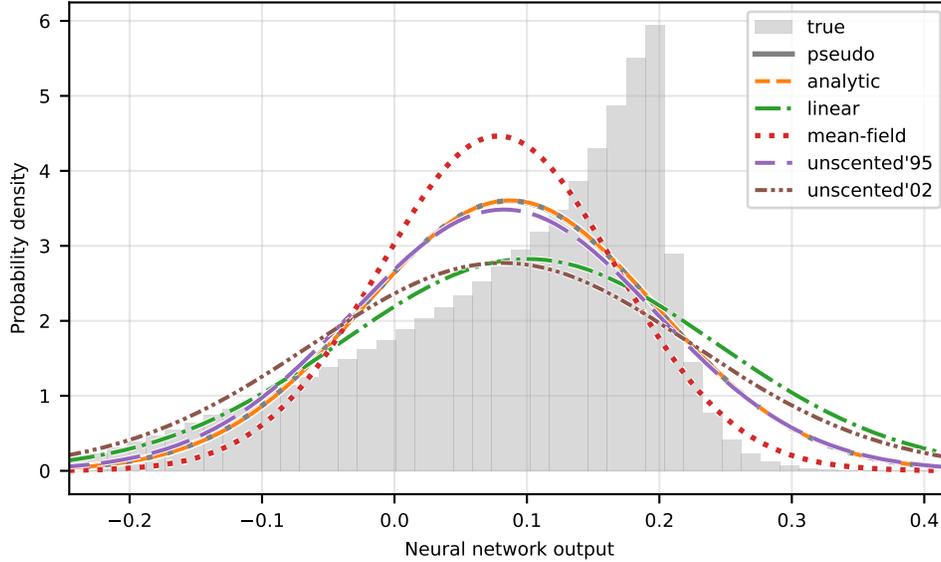


Figure 78: Probability distributions for Network(architecture=deep, weights=initialized, activation=sine), variance=small

6534  
6535  
6536  
6537  
6538  
6539  
6540  
6541  
6542  
6543  
6544  
6545  
6546  
6547  
6548  
6549  
6550  
6551  
6552  
6553  
6554  
6555  
6556  
6557  
6558  
6559  
6560  
6561  
6562  
6563  
6564  
6565  
6566  
6567  
6568  
6569  
6570  
6571  
6572  
6573  
6574  
6575  
6576  
6577  
6578  
6579  
6580  
6581  
6582  
6583  
6584  
6585  
6586  
6587

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-7.468 \times 10^{-2} \pm 1.8 \times 10^{-5}$	$3.835 \times 10^{-2} \pm 7.9 \times 10^{-6}$
analytic	$-8.706 \times 10^{-2}$	$3.731 \times 10^{-2}$
mean-field	$-1.600 \times 10^{-1}$	$3.821 \times 10^{-2}$
linear	$+1.006 \times 10^{-1}$	$1.994 \times 10^0$
unscented'95	$-1.997 \times 10^{-1}$	$4.768 \times 10^{-2}$
unscented'02	$-1.848 \times 10^0$	$9.588 \times 10^0$

Table 142: Comparison of moments for Network(architecture=deep, weights=initialized, activation=sine), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.136 \times 10^{-2} \pm 4.1 \times 10^{-5}$	0
analytic	$4.568 \times 10^{-2} \pm 3.7 \times 10^{-5}$	$2.187 \times 10^{-3} \pm 5.0 \times 10^{-6}$
mean-field	$1.934 \times 10^{-1} \pm 4.6 \times 10^{-5}$	$9.502 \times 10^{-2} \pm 5.3 \times 10^{-5}$
linear	$2.210 \times 10^0 \pm 1.5 \times 10^{-4}$	$2.393 \times 10^{+1} \pm 5.3 \times 10^{-3}$
unscented'95	$2.834 \times 10^{-1} \pm 4.9 \times 10^{-5}$	$2.164 \times 10^{-1} \pm 1.1 \times 10^{-4}$
unscented'02	$6.191 \times 10^0 \pm 3.4 \times 10^{-4}$	$1.627 \times 10^{+2} \pm 3.4 \times 10^{-2}$

Table 143: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=sine), variance=medium

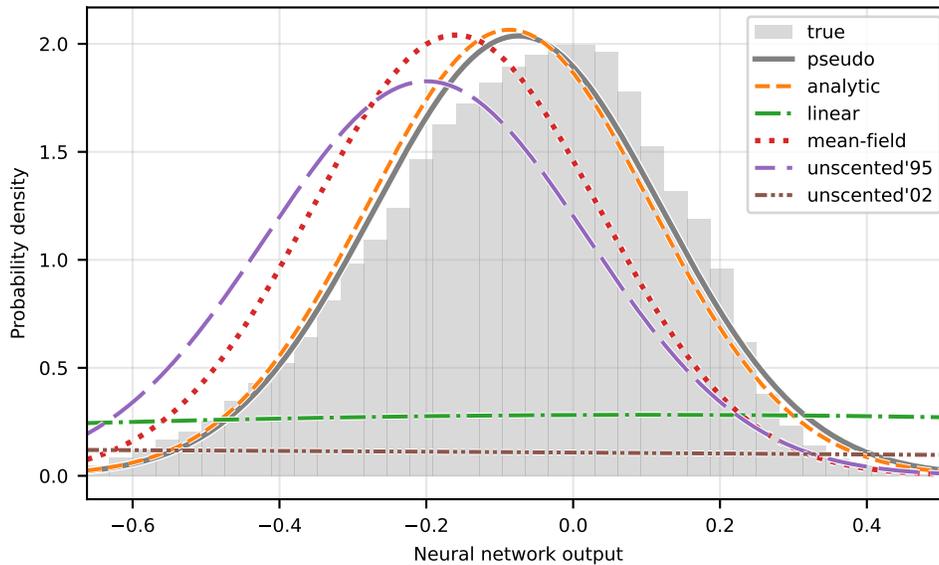


Figure 79: Probability distributions for Network(architecture=deep, weights=initialized, activation=sine), variance=medium

6588  
6589  
6590  
6591  
6592  
6593  
6594  
6595  
6596

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-1.076 \times 10^{-1} \pm 1.4 \times 10^{-4}$	$3.749 \times 10^{-2} \pm 5.1 \times 10^{-5}$
analytic	$-1.069 \times 10^{-1}$	$3.680 \times 10^{-2}$
mean-field	$-1.584 \times 10^{-1}$	$3.866 \times 10^{-2}$
linear	$+1.006 \times 10^{-1}$	$1.994 \times 10^{+2}$
unscented'95	$-7.389 \times 10^{-2}$	$3.900 \times 10^{-2}$
unscented'02	$-1.946 \times 10^{+2}$	$7.604 \times 10^{+4}$

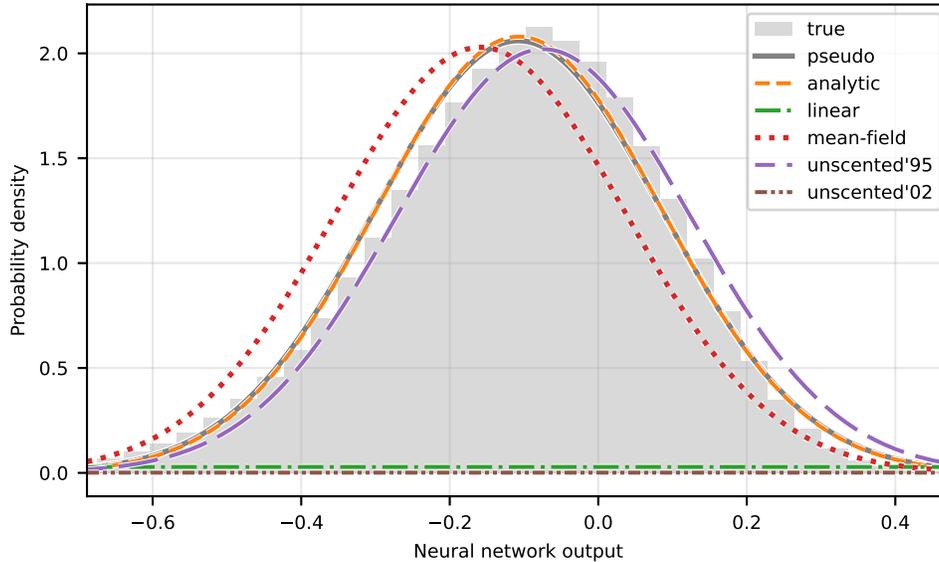
6597 Table 144: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
6598 tion=sine), variance=large  
6599

6600  
6601  
6602  
6603  
6604  
6605  
6606  
6607  
6608  
6609

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.691 \times 10^{-2} \pm 2.1 \times 10^{-4}$	0
analytic	$2.623 \times 10^{-2} \pm 2.3 \times 10^{-4}$	$1.076 \times 10^{-4} \pm 1.3 \times 10^{-5}$
mean-field	$1.175 \times 10^{-1} \pm 2.9 \times 10^{-4}$	$3.469 \times 10^{-2} \pm 2.2 \times 10^{-4}$
linear	$2.526 \times 10^{+1} \pm 8.9 \times 10^{-3}$	$2.656 \times 10^{+3} \pm 3.7 \times 10^0$
unscented'95	$7.668 \times 10^{-2} \pm 3.0 \times 10^{-4}$	$1.559 \times 10^{-2} \pm 1.2 \times 10^{-4}$
unscented'02	$6.193 \times 10^{+2} \pm 2.1 \times 10^{-1}$	$1.519 \times 10^{+6} \pm 2.1 \times 10^{+3}$

6610 Table 145: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
6611 activation=sine), variance=large  
6612

6613  
6614  
6615  
6616  
6617  
6618  
6619  
6620  
6621  
6622  
6623  
6624  
6625  
6626  
6627  
6628  
6629  
6630  
6631  
6632



6633  
6634  
6635  
6636  
6637  
6638  
6639  
6640  
6641

6634 Figure 80: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
6635 tion=sine), variance=large

6642  
6643  
6644  
6645  
6646  
6647  
6648  
6649  
6650

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-3.309 \times 10^{-1} \pm 3.4 \times 10^{-7}$	$2.760 \times 10^{-2} \pm 4.2 \times 10^{-7}$
analytic	$-3.310 \times 10^{-1}$	$2.474 \times 10^{-2}$
mean-field	$-2.899 \times 10^{-1}$	$9.013 \times 10^{-3}$
linear	$-2.892 \times 10^{-1}$	$2.971 \times 10^{-2}$
unscented'95	$-3.366 \times 10^{-1}$	$2.708 \times 10^{-2}$
unscented'02	$-3.311 \times 10^{-1}$	$3.322 \times 10^{-2}$

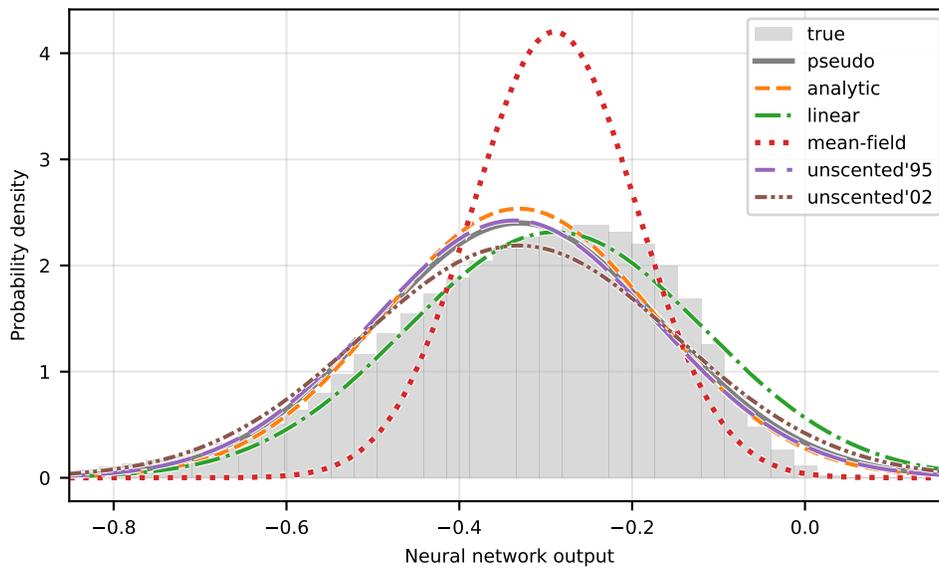
6651 Table 146: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
6652 tion=sine), variance=small  
6653  
6654

6655  
6656  
6657  
6658  
6659  
6660  
6661  
6662  
6663

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.504 \times 10^{-2} \pm 3.2 \times 10^{-6}$	0
analytic	$4.779 \times 10^{-2} \pm 3.8 \times 10^{-6}$	$2.875 \times 10^{-3} \pm 7.9 \times 10^{-7}$
mean-field	$1.453 \times 10^{-1} \pm 2.4 \times 10^{-6}$	$2.534 \times 10^{-1} \pm 4.7 \times 10^{-6}$
linear	$1.023 \times 10^{-1} \pm 9.3 \times 10^{-7}$	$3.288 \times 10^{-2} \pm 1.2 \times 10^{-6}$
unscented'95	$5.108 \times 10^{-2} \pm 4.4 \times 10^{-6}$	$6.692 \times 10^{-4} \pm 1.5 \times 10^{-7}$
unscented'02	$5.361 \times 10^{-2} \pm 4.3 \times 10^{-6}$	$9.147 \times 10^{-3} \pm 1.6 \times 10^{-6}$

6664 Table 147: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
6665 tivation=sine), variance=small  
6666  
6667

6668  
6669  
6670  
6671  
6672  
6673  
6674  
6675  
6676  
6677  
6678  
6679  
6680  
6681  
6682  
6683  
6684  
6685  
6686



6687 Figure 81: Probability distributions for Network(architecture=deep, weights=trained, activa-  
6688 tion=sine), variance=small  
6689  
6690  
6691  
6692  
6693  
6694  
6695

6696  
6697  
6698  
6699  
6700  
6701  
6702  
6703  
6704  
6705  
6706  
6707  
6708  
6709  
6710  
6711  
6712  
6713  
6714  
6715  
6716  
6717  
6718  
6719  
6720  
6721  
6722  
6723  
6724  
6725  
6726  
6727  
6728  
6729  
6730  
6731  
6732  
6733  
6734  
6735  
6736  
6737  
6738  
6739  
6740  
6741  
6742  
6743  
6744  
6745  
6746  
6747  
6748  
6749

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.705 \times 10^{-2} \pm 2.4 \times 10^{-5}$	$1.056 \times 10^{-1} \pm 1.7 \times 10^{-5}$
analytic	$+3.115 \times 10^{-2}$	$5.628 \times 10^{-2}$
mean-field	$-2.939 \times 10^{-2}$	$4.324 \times 10^{-2}$
linear	$-2.892 \times 10^{-1}$	$2.971 \times 10^0$
unscented'95	$-1.594 \times 10^{-1}$	$5.871 \times 10^{-2}$
unscented'02	$-4.479 \times 10^0$	$3.808 \times 10^{+1}$

Table 148: Comparison of moments for Network(architecture=deep, weights=trained, activation=sine), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.870 \times 10^{-2} \pm 3.2 \times 10^{-5}$	0
analytic	$1.373 \times 10^{-1} \pm 4.2 \times 10^{-5}$	$9.720 \times 10^{-2} \pm 3.5 \times 10^{-5}$
mean-field	$1.568 \times 10^{-1} \pm 3.6 \times 10^{-5}$	$1.513 \times 10^{-1} \pm 4.9 \times 10^{-5}$
linear	$2.004 \times 10^0 \pm 1.2 \times 10^{-4}$	$1.222 \times 10^{+1} \pm 2.3 \times 10^{-3}$
unscented'95	$2.508 \times 10^{-1} \pm 3.8 \times 10^{-5}$	$1.545 \times 10^{-1} \pm 4.1 \times 10^{-5}$
unscented'02	$1.046 \times 10^{+1} \pm 4.7 \times 10^{-4}$	$2.706 \times 10^{+2} \pm 4.5 \times 10^{-2}$

Table 149: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=sine), variance=medium

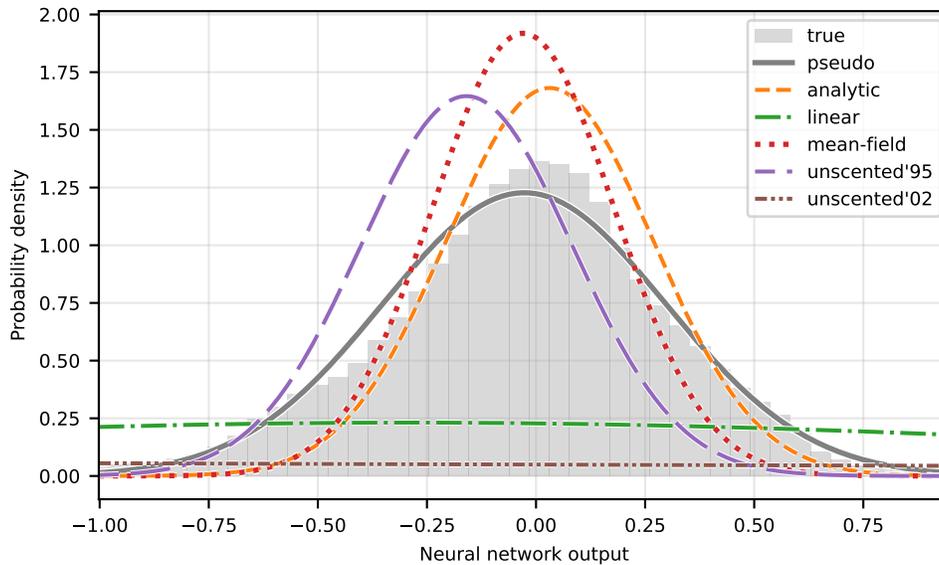


Figure 82: Probability distributions for Network(architecture=deep, weights=trained, activation=sine), variance=medium

6750  
6751  
6752  
6753  
6754  
6755  
6756  
6757  
6758  
6759  
6760  
6761  
6762  
6763  
6764  
6765  
6766  
6767  
6768  
6769  
6770  
6771  
6772  
6773  
6774  
6775  
6776  
6777  
6778  
6779  
6780  
6781  
6782  
6783  
6784  
6785  
6786  
6787  
6788  
6789  
6790  
6791  
6792  
6793  
6794  
6795  
6796  
6797  
6798  
6799  
6800  
6801  
6802  
6803

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.810 \times 10^{-2} \pm 1.6 \times 10^{-4}$	$7.180 \times 10^{-2} \pm 5.9 \times 10^{-5}$
analytic	$+3.082 \times 10^{-2}$	$5.672 \times 10^{-2}$
mean-field	$-4.625 \times 10^{-3}$	$4.329 \times 10^{-2}$
linear	$-2.892 \times 10^{-1}$	$2.971 \times 10^{+2}$
unscented'95	$+5.393 \times 10^{-2}$	$2.043 \times 10^{-2}$
unscented'02	$-4.205 \times 10^{+2}$	$3.534 \times 10^{+5}$

Table 150: Comparison of moments for Network(architecture=deep, weights=trained, activation=sine), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.247 \times 10^{-3} \pm 1.3 \times 10^{-4}$	0
analytic	$4.589 \times 10^{-2} \pm 1.8 \times 10^{-4}$	$1.292 \times 10^{-2} \pm 8.6 \times 10^{-5}$
mean-field	$1.065 \times 10^{-1} \pm 2.0 \times 10^{-4}$	$6.192 \times 10^{-2} \pm 1.8 \times 10^{-4}$
linear	$2.616 \times 10^{+1} \pm 5.5 \times 10^{-3}$	$2.065 \times 10^{+3} \pm 1.7 \times 10^0$
unscented'95	$1.966 \times 10^{-1} \pm 1.5 \times 10^{-4}$	$2.754 \times 10^{-1} \pm 2.9 \times 10^{-4}$
unscented'02	$1.136 \times 10^{+3} \pm 2.3 \times 10^{-1}$	$3.692 \times 10^{+6} \pm 3.0 \times 10^{+3}$

Table 151: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=sine), variance=large

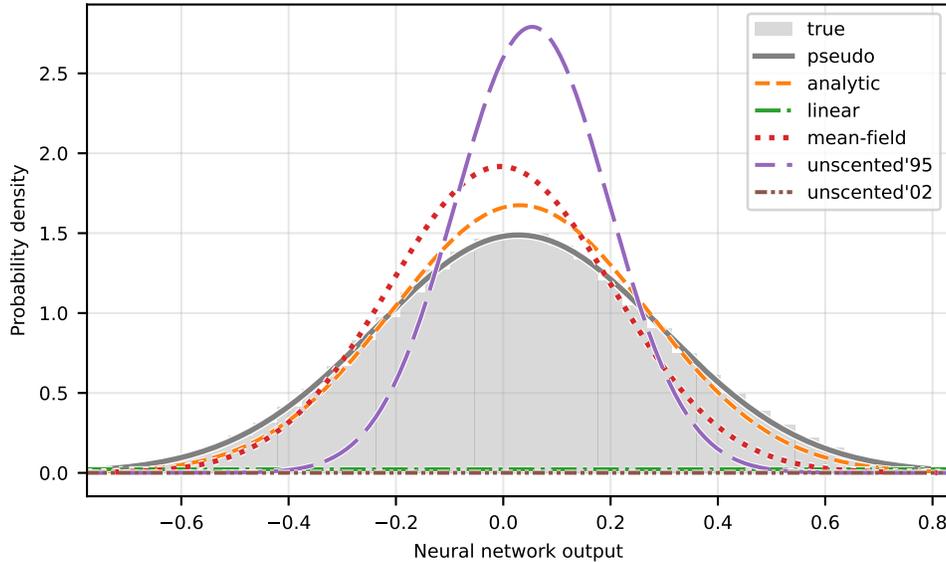


Figure 83: Probability distributions for Network(architecture=deep, weights=trained, activation=sine), variance=large

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.833 \times 10^0 \pm 9.3 \times 10^{-4}$	$5.519 \times 10^0 \pm 3.5 \times 10^{-3}$
analytic	$+1.634 \times 10^0$	$6.420 \times 10^0$
mean-field	$+1.506 \times 10^0$	$7.240 \times 10^0$
linear	$+5.635 \times 10^0$	$4.854 \times 10^{+3}$
unscented'95	$+1.465 \times 10^0$	$3.321 \times 10^0$
unscented'02	$+2.030 \times 10^{+2}$	$8.267 \times 10^{+4}$

Table 152: Comparison of moments for Network(architecture=deep, weights=initialized, activation=sine residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.564 \times 10^{-2} \pm 4.7 \times 10^{-4}$	0
analytic	$1.521 \times 10^{-1} \pm 4.8 \times 10^{-4}$	$9.624 \times 10^{-3} \pm 6.7 \times 10^{-5}$
mean-field	$2.590 \times 10^{-1} \pm 4.8 \times 10^{-4}$	$2.989 \times 10^{-2} \pm 1.3 \times 10^{-4}$
linear	$3.510 \times 10^{+1} \pm 6.0 \times 10^{-3}$	$4.371 \times 10^{+2} \pm 2.8 \times 10^{-1}$
unscented'95	$3.414 \times 10^{-1} \pm 3.5 \times 10^{-4}$	$6.712 \times 10^{-2} \pm 1.3 \times 10^{-4}$
unscented'02	$1.839 \times 10^{+2} \pm 3.0 \times 10^{-2}$	$1.115 \times 10^{+4} \pm 7.1 \times 10^0$

Table 153: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=sine residual), variance=small

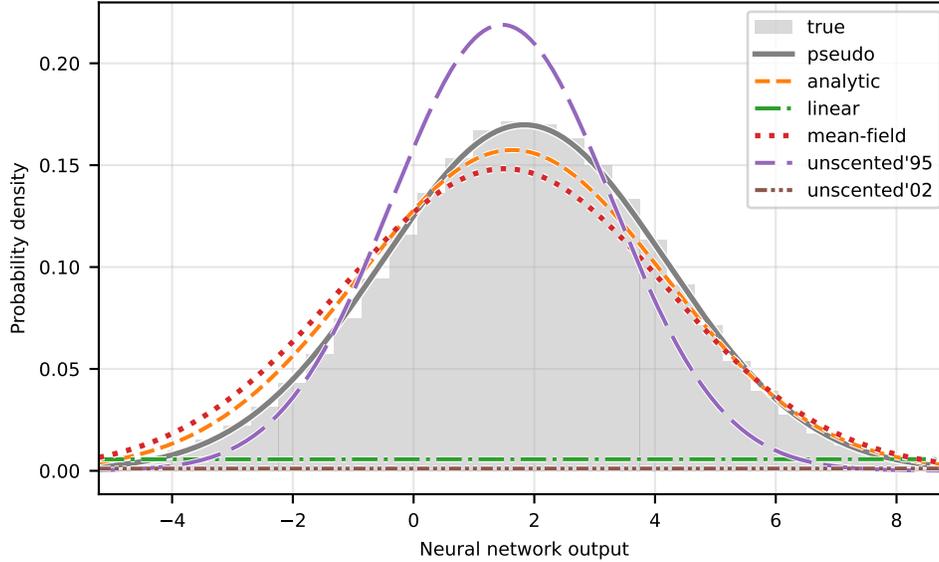


Figure 84: Probability distributions for Network(architecture=deep, weights=initialized, activation=sine residual), variance=small

6858  
6859  
6860  
6861  
6862  
6863  
6864  
6865  
6866  
6867  
6868  
6869  
6870  
6871  
6872  
6873  
6874  
6875  
6876  
6877  
6878  
6879  
6880  
6881  
6882  
6883  
6884  
6885  
6886  
6887  
6888  
6889  
6890  
6891  
6892  
6893  
6894  
6895  
6896  
6897  
6898  
6899  
6900  
6901  
6902  
6903  
6904  
6905  
6906  
6907  
6908  
6909  
6910  
6911

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.948 \times 10^{-1} \pm 1.9 \times 10^{-3}$	$1.022 \times 10^{+1} \pm 1.0 \times 10^{-2}$
analytic	$+6.953 \times 10^{-1}$	$1.046 \times 10^{+1}$
mean-field	$+8.278 \times 10^{-1}$	$1.010 \times 10^{+1}$
linear	$+5.635 \times 10^0$	$4.854 \times 10^{+5}$
unscented'95	$-2.237 \times 10^{-2}$	$8.705 \times 10^0$
unscented'02	$+1.935 \times 10^{+4}$	$7.487 \times 10^{+8}$

Table 154: Comparison of moments for Network(architecture=deep, weights=initialized, activation=sine residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.137 \times 10^{-2} \pm 5.1 \times 10^{-4}$	0
analytic	$5.630 \times 10^{-2} \pm 1.0 \times 10^{-3}$	$6.376 \times 10^{-4} \pm 2.5 \times 10^{-5}$
mean-field	$1.303 \times 10^{-1} \pm 1.0 \times 10^{-3}$	$2.700 \times 10^{-3} \pm 4.1 \times 10^{-5}$
linear	$3.095 \times 10^{+2} \pm 8.0 \times 10^{-2}$	$2.374 \times 10^{+4} \pm 2.4 \times 10^{+1}$
unscented'95	$3.467 \times 10^{-1} \pm 1.0 \times 10^{-3}$	$2.476 \times 10^{-2} \pm 1.4 \times 10^{-4}$
unscented'02	$1.514 \times 10^{+4} \pm 3.9 \times 10^0$	$5.494 \times 10^{+7} \pm 5.6 \times 10^{+4}$

Table 155: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=sine residual), variance=medium

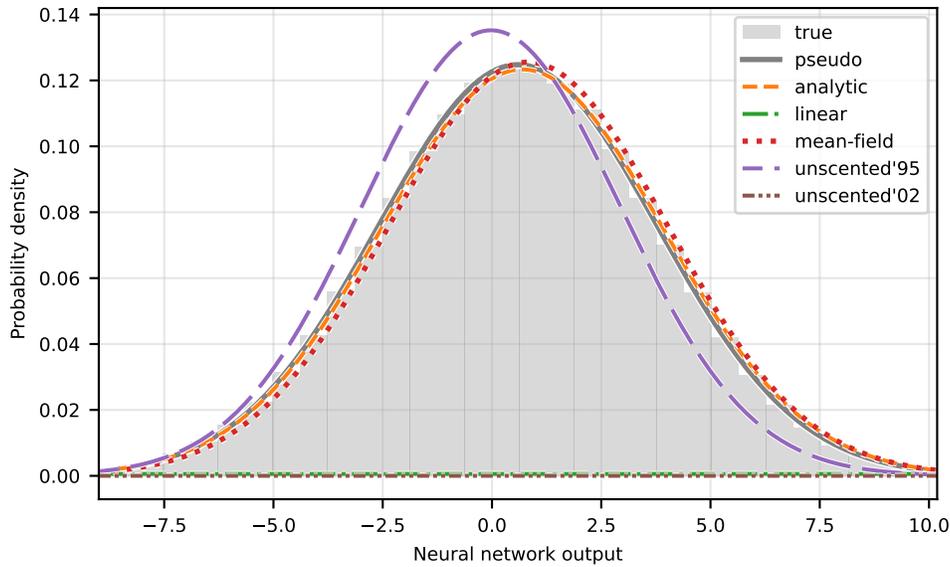


Figure 85: Probability distributions for Network(architecture=deep, weights=initialized, activation=sine residual), variance=medium

6912  
6913  
6914  
6915  
6916  
6917  
6918  
6919  
6920

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+7.725 \times 10^{-1} \pm 2.2 \times 10^{-3}$	$1.062 \times 10^{+1} \pm 1.2 \times 10^{-2}$
analytic	$+7.669 \times 10^{-1}$	$1.060 \times 10^{+1}$
mean-field	$+6.359 \times 10^{-1}$	$1.038 \times 10^{+1}$
linear	$+5.635 \times 10^0$	$4.854 \times 10^{+7}$
unscented'95	$-2.682 \times 10^0$	$9.806 \times 10^0$
unscented'02	$-1.840 \times 10^{+6}$	$6.771 \times 10^{+12}$

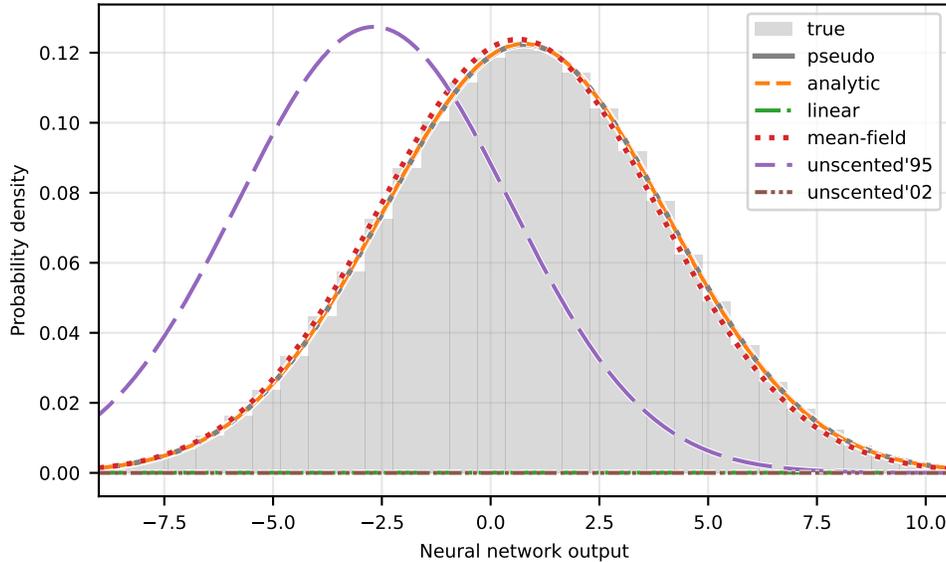
6921 Table 156: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
6922 tion=sine residual), variance=large  
6923  
6924

6925  
6926  
6927  
6928  
6929  
6930  
6931  
6932  
6933

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.743 \times 10^{-3} \pm 2.9 \times 10^{-4}$	0
analytic	$8.862 \times 10^{-3} \pm 6.3 \times 10^{-4}$	$1.424 \times 10^{-5} \pm 3.0 \times 10^{-6}$
mean-field	$7.593 \times 10^{-2} \pm 1.2 \times 10^{-3}$	$1.020 \times 10^{-3} \pm 2.7 \times 10^{-5}$
linear	$3.077 \times 10^{+3} \pm 8.8 \times 10^{-1}$	$2.284 \times 10^{+6} \pm 2.6 \times 10^{+3}$
unscented'95	$1.913 \times 10^0 \pm 1.5 \times 10^{-3}$	$5.632 \times 10^{-1} \pm 1.1 \times 10^{-3}$
unscented'02	$1.426 \times 10^{+6} \pm 4.1 \times 10^{+2}$	$4.780 \times 10^{+11} \pm 5.4 \times 10^{+8}$

6934 Table 157: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
6935 activation=sine residual), variance=large  
6936  
6937

6938  
6939  
6940  
6941  
6942  
6943  
6944  
6945  
6946  
6947  
6948  
6949  
6950  
6951  
6952  
6953  
6954  
6955  
6956



6957  
6958 Figure 86: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
6959 tion=sine residual), variance=large  
6960  
6961  
6962  
6963  
6964  
6965

6966  
6967  
6968  
6969  
6970  
6971  
6972  
6973  
6974

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.816 \times 10^0 \pm 1.1 \times 10^{-3}$	$5.545 \times 10^0 \pm 4.2 \times 10^{-3}$
analytic	$+1.657 \times 10^0$	$6.417 \times 10^0$
mean-field	$+1.526 \times 10^0$	$7.238 \times 10^0$
linear	$+2.756 \times 10^0$	$2.222 \times 10^{+3}$
unscented'95	$+1.666 \times 10^0$	$3.020 \times 10^0$
unscented'02	$-1.425 \times 10^{+3}$	$4.076 \times 10^{+6}$

6975  
6976  
6977

Table 158: Comparison of moments for Network(architecture=deep, weights=trained, activation=sine residual), variance=small

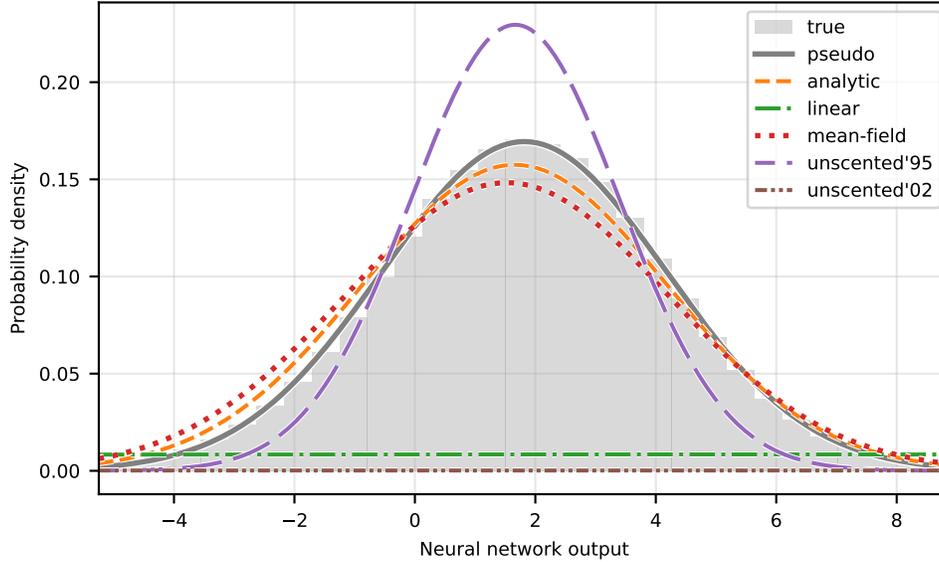
6978  
6979  
6980

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.305 \times 10^{-2} \pm 3.9 \times 10^{-4}$	0
analytic	$1.324 \times 10^{-1} \pm 4.7 \times 10^{-4}$	$7.895 \times 10^{-3} \pm 6.0 \times 10^{-5}$
mean-field	$2.408 \times 10^{-1} \pm 4.8 \times 10^{-4}$	$2.701 \times 10^{-2} \pm 1.2 \times 10^{-4}$
linear	$2.330 \times 10^{+1} \pm 4.9 \times 10^{-3}$	$1.970 \times 10^{+2} \pm 1.5 \times 10^{-1}$
unscented'95	$3.278 \times 10^{-1} \pm 5.5 \times 10^{-4}$	$7.811 \times 10^{-2} \pm 1.8 \times 10^{-4}$
unscented'02	$1.300 \times 10^{+3} \pm 2.5 \times 10^{-1}$	$5.511 \times 10^{+5} \pm 4.2 \times 10^{+2}$

6987  
6988  
6989  
6990

Table 159: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=sine residual), variance=small

6991  
6992



7003  
7004  
7005  
7006  
7007  
7008  
7009  
7010  
7011

Figure 87: Probability distributions for Network(architecture=deep, weights=trained, activation=sine residual), variance=small

7012  
7013  
7014  
7015  
7016  
7017  
7018  
7019

7020  
7021  
7022  
7023  
7024  
7025  
7026  
7027  
7028

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.940 \times 10^{-1} \pm 2.0 \times 10^{-3}$	$1.022 \times 10^{+1} \pm 1.1 \times 10^{-2}$
analytic	$+6.930 \times 10^{-1}$	$1.046 \times 10^{+1}$
mean-field	$+8.258 \times 10^{-1}$	$1.010 \times 10^{+1}$
linear	$+2.756 \times 10^0$	$2.222 \times 10^{+5}$
unscented'95	$-1.853 \times 10^{-1}$	$9.735 \times 10^0$
unscented'02	$+4.562 \times 10^{+3}$	$4.183 \times 10^{+7}$

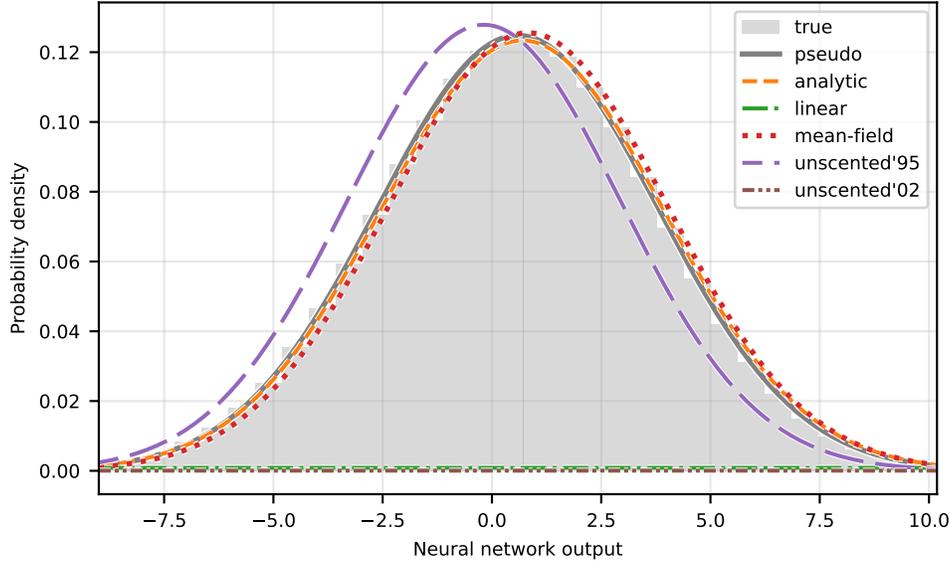
7029 Table 160: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
7030 tion=sine residual), variance=medium  
7031

7032  
7033  
7034  
7035  
7036  
7037  
7038  
7039  
7040  
7041

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.207 \times 10^{-2} \pm 5.1 \times 10^{-4}$	0
analytic	$5.551 \times 10^{-2} \pm 1.1 \times 10^{-3}$	$6.270 \times 10^{-4} \pm 2.7 \times 10^{-5}$
mean-field	$1.297 \times 10^{-1} \pm 1.1 \times 10^{-3}$	$2.673 \times 10^{-3} \pm 4.4 \times 10^{-5}$
linear	$2.090 \times 10^{+2} \pm 5.8 \times 10^{-2}$	$1.087 \times 10^{+4} \pm 1.2 \times 10^{+1}$
unscented'95	$4.359 \times 10^{-1} \pm 1.1 \times 10^{-3}$	$3.031 \times 10^{-2} \pm 1.5 \times 10^{-4}$
unscented'02	$3.575 \times 10^{+3} \pm 9.8 \times 10^{-1}$	$3.065 \times 10^{+6} \pm 3.4 \times 10^{+3}$

7042 Table 161: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
7043 tivation=sine residual), variance=medium  
7044

7045  
7046  
7047  
7048  
7049  
7050  
7051  
7052  
7053  
7054  
7055  
7056  
7057  
7058  
7059  
7060  
7061  
7062  
7063  
7064



7066 Figure 88: Probability distributions for Network(architecture=deep, weights=trained, activa-  
7067 tion=sine residual), variance=medium  
7068  
7069  
7070  
7071  
7072  
7073

7074  
7075  
7076  
7077  
7078  
7079  
7080  
7081  
7082

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+7.718 \times 10^{-1} \pm 2.1 \times 10^{-3}$	$1.063 \times 10^{+1} \pm 1.2 \times 10^{-2}$
analytic	$+7.654 \times 10^{-1}$	$1.059 \times 10^{+1}$
mean-field	$+6.345 \times 10^{-1}$	$1.038 \times 10^{+1}$
linear	$+2.756 \times 10^0$	$2.222 \times 10^{+7}$
unscented'95	$-2.734 \times 10^0$	$9.988 \times 10^0$
unscented'02	$-2.299 \times 10^{+4}$	$1.059 \times 10^{+9}$

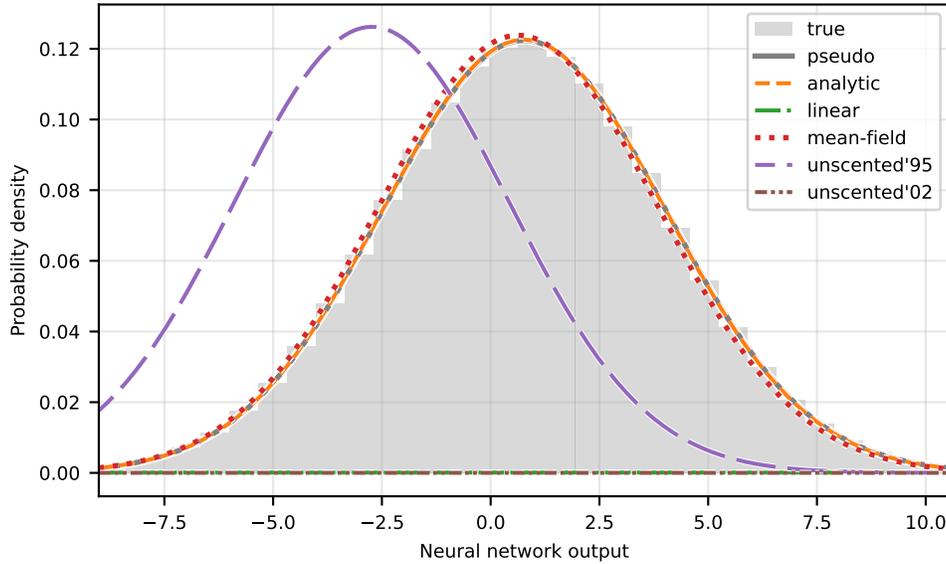
7083 Table 162: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
7084 tion=sine residual), variance=large  
7085

7086  
7087  
7088  
7089  
7090  
7091  
7092  
7093  
7094  
7095

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.821 \times 10^{-3} \pm 3.1 \times 10^{-4}$	0
analytic	$8.940 \times 10^{-3} \pm 6.7 \times 10^{-4}$	$1.504 \times 10^{-5} \pm 3.2 \times 10^{-6}$
mean-field	$7.640 \times 10^{-2} \pm 1.2 \times 10^{-3}$	$1.033 \times 10^{-3} \pm 2.5 \times 10^{-5}$
linear	$2.082 \times 10^{+3} \pm 5.9 \times 10^{-1}$	$1.046 \times 10^{+6} \pm 1.2 \times 10^{+3}$
unscented'95	$1.942 \times 10^0 \pm 1.5 \times 10^{-3}$	$5.794 \times 10^{-1} \pm 1.1 \times 10^{-3}$
unscented'02	$1.782 \times 10^{+4} \pm 5.1 \times 10^0$	$7.468 \times 10^{+7} \pm 8.5 \times 10^{+4}$

7096 Table 163: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
7097 tivation=sine residual), variance=large  
7098

7099  
7100  
7101  
7102  
7103  
7104  
7105  
7106  
7107  
7108  
7109  
7110  
7111  
7112  
7113  
7114  
7115  
7116  
7117



7118  
7119  
7120  
7121  
7122  
7123  
7124  
7125  
7126  
7127

7120 Figure 89: Probability distributions for Network(architecture=deep, weights=trained, activa-  
7121 tion=sine residual), variance=large

7128  
7129  
7130  
7131  
7132  
7133  
7134  
7135  
7136

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.886 \times 10^0 \pm 1.8 \times 10^{-7}$	$1.239 \times 10^{-2} \pm 1.5 \times 10^{-7}$
analytic	$+2.886 \times 10^0$	$1.254 \times 10^{-2}$
mean-field	$+2.894 \times 10^0$	$1.642 \times 10^{-2}$
linear	$+2.870 \times 10^0$	$1.262 \times 10^{-2}$
unscented'95	$+2.886 \times 10^0$	$1.281 \times 10^{-2}$
unscented'02	$+2.886 \times 10^0$	$1.316 \times 10^{-2}$

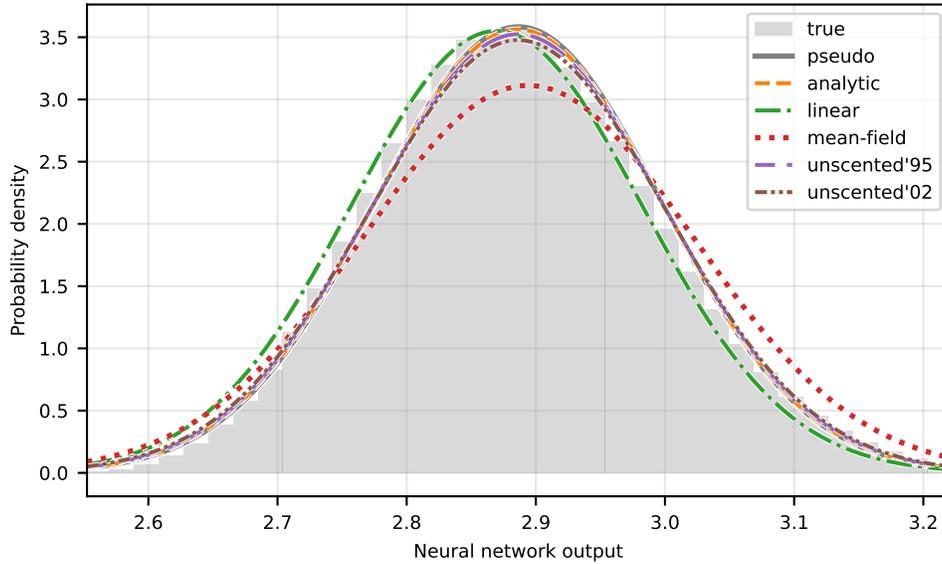
7137 Table 164: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
7138 tion=gelu), variance=small  
7139  
7140

7141  
7142  
7143  
7144  
7145  
7146  
7147  
7148  
7149

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.205 \times 10^{-2} \pm 2.2 \times 10^{-6}$	0
analytic	$1.165 \times 10^{-2} \pm 2.8 \times 10^{-6}$	$4.131 \times 10^{-5} \pm 7.0 \times 10^{-8}$
mean-field	$4.870 \times 10^{-2} \pm 2.7 \times 10^{-6}$	$2.410 \times 10^{-2} \pm 2.1 \times 10^{-6}$
linear	$5.025 \times 10^{-2} \pm 4.9 \times 10^{-7}$	$1.143 \times 10^{-2} \pm 2.5 \times 10^{-7}$
unscented'95	$1.236 \times 10^{-2} \pm 3.5 \times 10^{-6}$	$2.818 \times 10^{-4} \pm 2.1 \times 10^{-7}$
unscented'02	$1.358 \times 10^{-2} \pm 4.0 \times 10^{-6}$	$9.330 \times 10^{-4} \pm 3.8 \times 10^{-7}$

7150 Table 165: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
7151 activation=gelu), variance=small  
7152  
7153

7154  
7155  
7156  
7157  
7158  
7159  
7160  
7161  
7162  
7163  
7164  
7165  
7166  
7167  
7168  
7169  
7170  
7171



7172  
7173  
7174  
7175  
7176  
7177  
7178  
7179  
7180  
7181

7174 Figure 90: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
7175 tion=gelu), variance=small

7182  
7183  
7184  
7185  
7186  
7187  
7188  
7189  
7190

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.815 \times 10^0 \pm 7.4 \times 10^{-6}$	$2.040 \times 10^{-1} \pm 1.0 \times 10^{-5}$
analytic	$+2.834 \times 10^0$	$2.371 \times 10^{-1}$
mean-field	$+2.635 \times 10^0$	$5.798 \times 10^{-1}$
linear	$+2.870 \times 10^0$	$1.262 \times 10^0$
unscented'95	$+2.732 \times 10^0$	$1.270 \times 10^{-1}$
unscented'02	$+4.517 \times 10^0$	$6.690 \times 10^0$

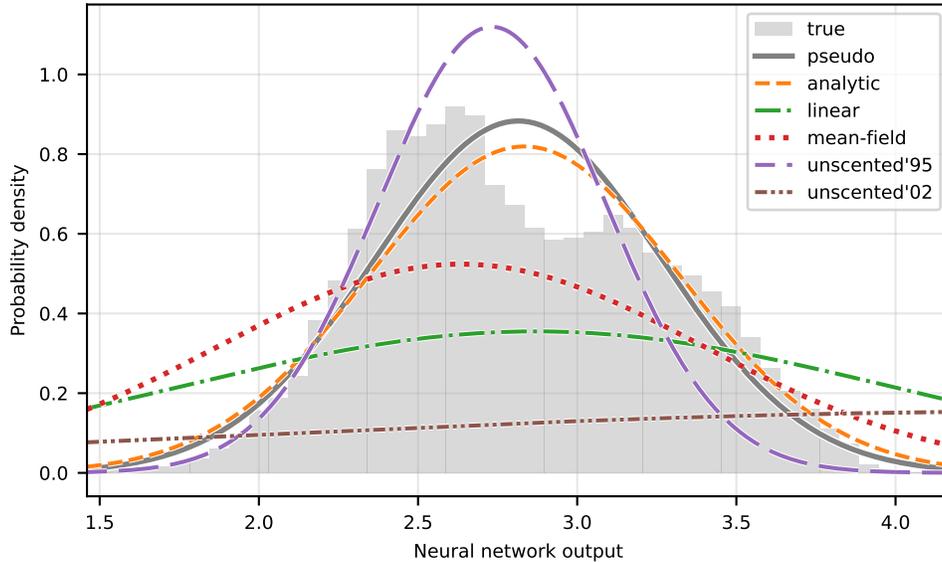
Table 166: Comparison of moments for Network(architecture=deep, weights=initialized, activation=gelu), variance=medium

7191  
7192  
7193  
7194  
7195  
7196  
7197  
7198  
7199  
7200  
7201  
7202  
7203

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.632 \times 10^{-2} \pm 1.8 \times 10^{-5}$	0
analytic	$7.115 \times 10^{-2} \pm 2.2 \times 10^{-5}$	$6.835 \times 10^{-3} \pm 3.8 \times 10^{-6}$
mean-field	$4.028 \times 10^{-1} \pm 1.4 \times 10^{-5}$	$4.783 \times 10^{-1} \pm 5.3 \times 10^{-5}$
linear	$7.860 \times 10^{-1} \pm 2.6 \times 10^{-5}$	$1.689 \times 10^0 \pm 1.3 \times 10^{-4}$
unscented'95	$1.395 \times 10^{-1} \pm 1.8 \times 10^{-5}$	$6.516 \times 10^{-2} \pm 7.7 \times 10^{-6}$
unscented'02	$3.322 \times 10^0 \pm 5.4 \times 10^{-5}$	$2.125 \times 10^{+1} \pm 1.1 \times 10^{-3}$

Table 167: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=gelu), variance=medium

7204  
7205  
7206  
7207  
7208  
7209  
7210  
7211  
7212  
7213  
7214  
7215  
7216  
7217  
7218  
7219  
7220  
7221  
7222  
7223  
7224  
7225  
7226



7227  
7228  
7229  
7230  
7231  
7232  
7233  
7234  
7235

Figure 91: Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu), variance=medium

7236  
7237  
7238  
7239  
7240  
7241  
7242  
7243  
7244

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+7.526 \times 10^0 \pm 1.1 \times 10^{-4}$	$1.930 \times 10^{+1} \pm 2.9 \times 10^{-3}$
analytic	$+8.032 \times 10^0$	$1.048 \times 10^{+1}$
mean-field	$+7.986 \times 10^0$	$1.344 \times 10^{+1}$
linear	$+2.870 \times 10^0$	$1.262 \times 10^{+2}$
unscented'95	$+8.265 \times 10^0$	$1.355 \times 10^{+1}$
unscented'02	$+1.676 \times 10^{+2}$	$5.441 \times 10^{+4}$

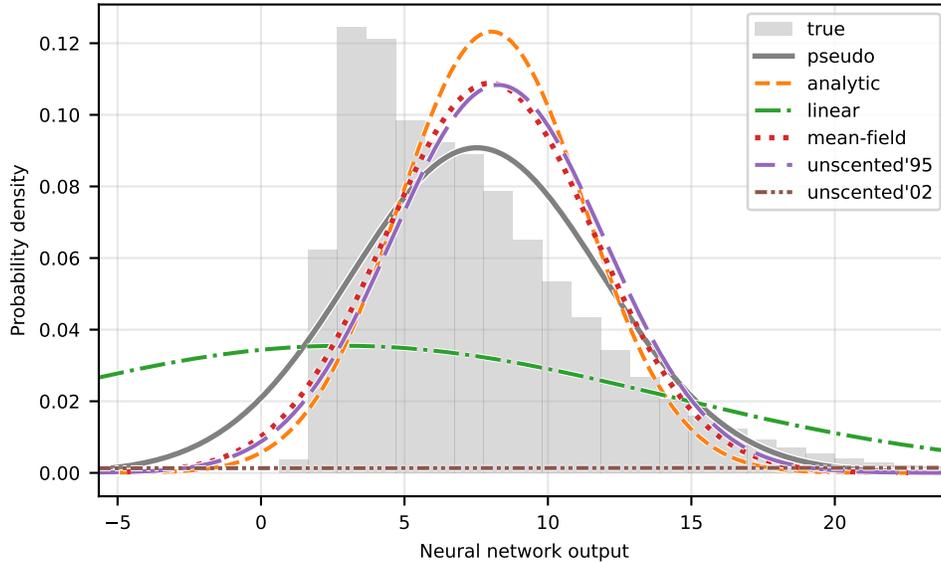
7245 Table 168: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
7246 tion=gelu), variance=large  
7247

7248  
7249  
7250  
7251  
7252  
7253  
7254  
7255  
7256  
7257

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.622 \times 10^{-1} \pm 8.9 \times 10^{-5}$	0
analytic	$6.447 \times 10^{-1} \pm 7.4 \times 10^{-5}$	$8.356 \times 10^{-2} \pm 3.2 \times 10^{-5}$
mean-field	$5.737 \times 10^{-1} \pm 7.7 \times 10^{-5}$	$3.468 \times 10^{-2} \pm 2.1 \times 10^{-5}$
linear	$3.070 \times 10^0 \pm 1.7 \times 10^{-4}$	$2.392 \times 10^0 \pm 4.9 \times 10^{-4}$
unscented'95	$6.487 \times 10^{-1} \pm 6.9 \times 10^{-5}$	$4.204 \times 10^{-2} \pm 1.9 \times 10^{-5}$
unscented'02	$1.079 \times 10^{+2} \pm 4.1 \times 10^{-3}$	$2.069 \times 10^{+3} \pm 3.1 \times 10^{-1}$

7258 Table 169: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
7259 activation=gelu), variance=large  
7260

7261  
7262  
7263  
7264  
7265  
7266  
7267  
7268  
7269  
7270  
7271  
7272  
7273  
7274  
7275  
7276  
7277



7278  
7279  
7280  
7281  
7282  
7283  
7284  
7285  
7286  
7287  
7288  
7289

7282 Figure 92: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
7283 tion=gelu), variance=large

7290  
7291  
7292  
7293  
7294  
7295  
7296  
7297  
7298

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-5.675 \times 10^{-1} \pm 3.1 \times 10^{-7}$	$6.305 \times 10^{-2} \pm 5.0 \times 10^{-7}$
analytic	$-5.702 \times 10^{-1}$	$6.299 \times 10^{-2}$
mean-field	$-5.180 \times 10^{-1}$	$2.304 \times 10^{-2}$
linear	$-5.470 \times 10^{-1}$	$8.199 \times 10^{-2}$
unscented'95	$-5.727 \times 10^{-1}$	$6.783 \times 10^{-2}$
unscented'02	$-5.755 \times 10^{-1}$	$8.361 \times 10^{-2}$

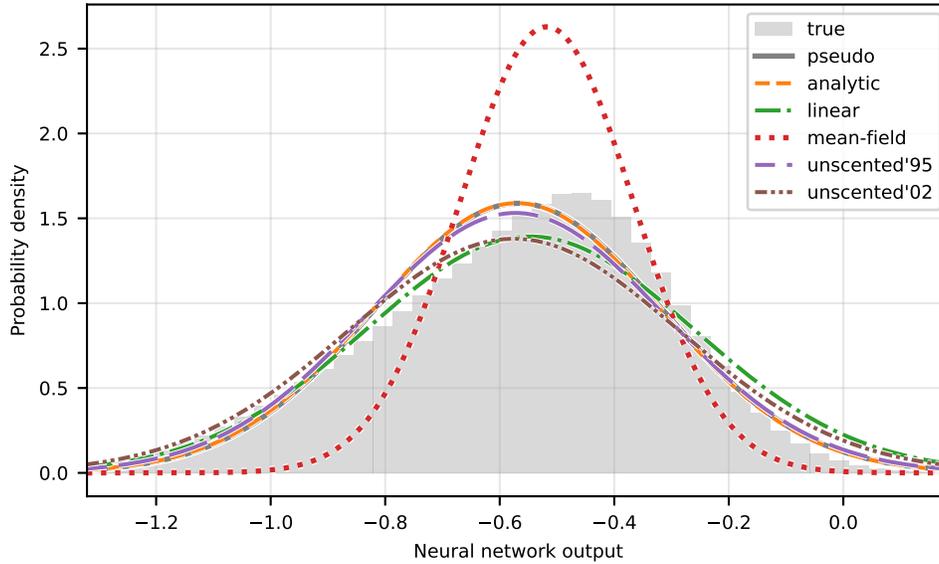
7299 Table 170: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
7300 tion=gelu), variance=small  
7301

7302  
7303  
7304  
7305  
7306  
7307  
7308  
7309  
7310  
7311

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.070 \times 10^{-2} \pm 4.1 \times 10^{-6}$	0
analytic	$4.193 \times 10^{-2} \pm 4.1 \times 10^{-6}$	$5.721 \times 10^{-5} \pm 1.4 \times 10^{-8}$
mean-field	$1.656 \times 10^{-1} \pm 2.1 \times 10^{-6}$	$2.056 \times 10^{-1} \pm 2.3 \times 10^{-6}$
linear	$5.535 \times 10^{-2} \pm 7.2 \times 10^{-6}$	$2.221 \times 10^{-2} \pm 1.2 \times 10^{-6}$
unscented'95	$4.372 \times 10^{-2} \pm 5.3 \times 10^{-6}$	$1.577 \times 10^{-3} \pm 3.0 \times 10^{-7}$
unscented'02	$7.483 \times 10^{-2} \pm 2.8 \times 10^{-6}$	$2.243 \times 10^{-2} \pm 1.3 \times 10^{-6}$

7312 Table 171: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
7313 tivation=gelu), variance=small  
7314

7315  
7316  
7317  
7318  
7319  
7320  
7321  
7322  
7323  
7324  
7325  
7326  
7327  
7328  
7329  
7330  
7331  
7332  
7333



7334  
7335  
7336 Figure 93: Probability distributions for Network(architecture=deep, weights=trained, activa-  
7337 tion=gelu), variance=small  
7338

7339  
7340  
7341  
7342  
7343

7344  
7345  
7346  
7347  
7348  
7349  
7350  
7351  
7352

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.262 \times 10^{-1} \pm 1.0 \times 10^{-5}$	$3.663 \times 10^{-1} \pm 1.7 \times 10^{-5}$
analytic	$-7.138 \times 10^{-2}$	$3.740 \times 10^{-1}$
mean-field	$+7.814 \times 10^{-3}$	$7.079 \times 10^{-1}$
linear	$-5.470 \times 10^{-1}$	$8.199 \times 10^0$
unscented'95	$-3.967 \times 10^{-1}$	$5.968 \times 10^{-1}$
unscented'02	$-3.399 \times 10^0$	$2.446 \times 10^{+1}$

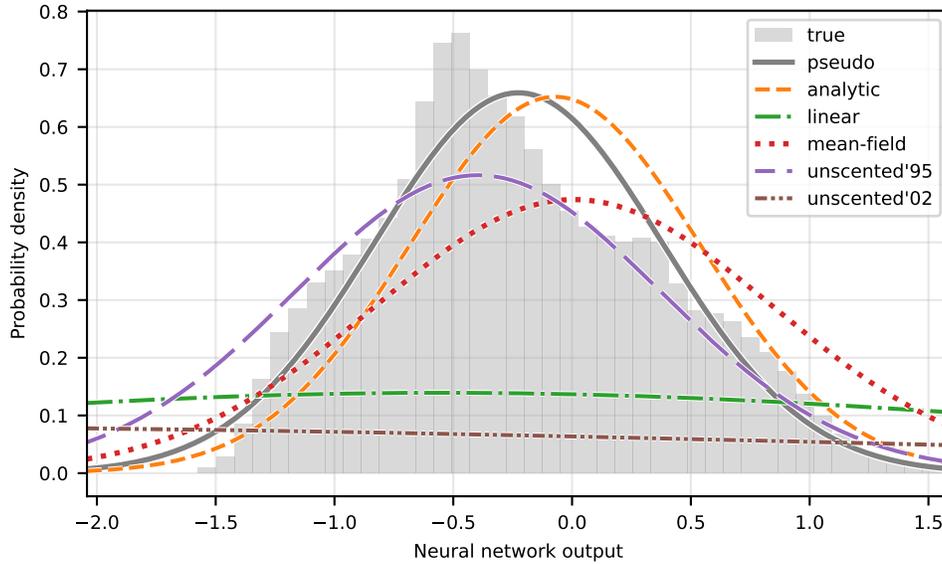
Table 172: Comparison of moments for Network(architecture=deep, weights=trained, activation=gelu), variance=medium

7353  
7354  
7355  
7356  
7357  
7358  
7359  
7360  
7361  
7362  
7363  
7364  
7365

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.726 \times 10^{-2} \pm 2.0 \times 10^{-5}$	0
analytic	$2.096 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$3.284 \times 10^{-2} \pm 6.0 \times 10^{-6}$
mean-field	$3.801 \times 10^{-1} \pm 2.1 \times 10^{-5}$	$2.117 \times 10^{-1} \pm 3.1 \times 10^{-5}$
linear	$2.320 \times 10^0 \pm 4.2 \times 10^{-5}$	$9.278 \times 10^0 \pm 5.0 \times 10^{-4}$
unscented'95	$2.379 \times 10^{-1} \pm 1.5 \times 10^{-5}$	$1.103 \times 10^{-1} \pm 1.4 \times 10^{-5}$
unscented'02	$5.556 \times 10^0 \pm 7.7 \times 10^{-5}$	$4.454 \times 10^{+1} \pm 2.1 \times 10^{-3}$

Table 173: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=gelu), variance=medium

7366  
7367  
7368  
7369  
7370  
7371  
7372  
7373  
7374  
7375  
7376  
7377  
7378  
7379  
7380  
7381  
7382  
7383  
7384  
7385  
7386  
7387



7388  
7389  
7390  
7391  
7392  
7393  
7394  
7395  
7396  
7397

Figure 94: Probability distributions for Network(architecture=deep, weights=trained, activation=gelu), variance=medium

7438  
7439  
7400  
7401  
7402  
7403  
7404  
7405  
7406

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+3.776 \times 10^0 \pm 1.7 \times 10^{-4}$	$1.469 \times 10^{+1} \pm 2.2 \times 10^{-3}$
analytic	$+4.338 \times 10^0$	$8.886 \times 10^0$
mean-field	$+4.591 \times 10^0$	$1.377 \times 10^{+1}$
linear	$-5.470 \times 10^{-1}$	$8.199 \times 10^{+2}$
unscented'95	$+3.805 \times 10^0$	$1.557 \times 10^{+1}$
unscented'02	$-2.854 \times 10^{+2}$	$1.632 \times 10^{+5}$

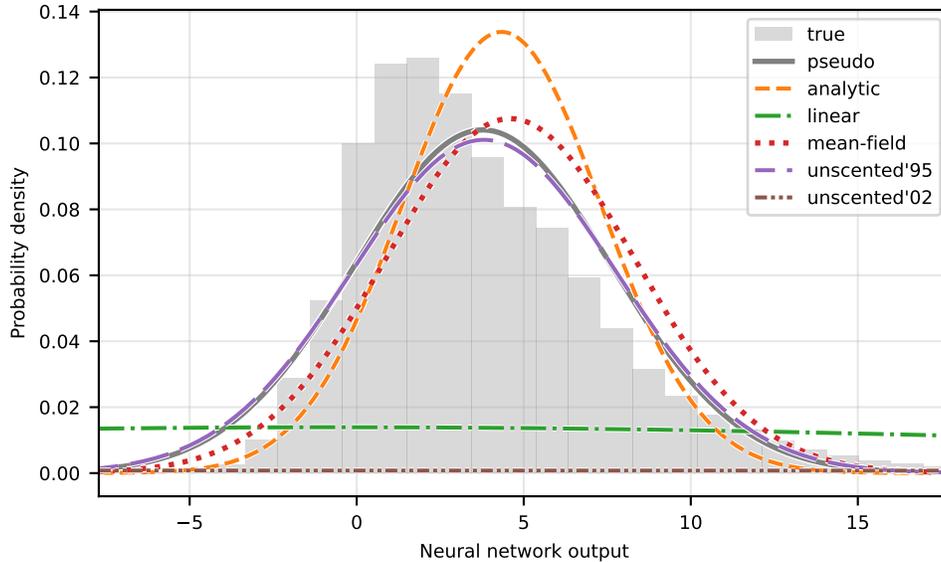
7407 Table 174: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
7408 tion=gelu), variance=large  
7409

7411  
7412  
7413  
7414  
7415  
7416  
7417  
7418  
7419

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.440 \times 10^{-1} \pm 9.7 \times 10^{-5}$	0
analytic	$5.409 \times 10^{-1} \pm 8.8 \times 10^{-5}$	$6.445 \times 10^{-2} \pm 2.5 \times 10^{-5}$
mean-field	$5.635 \times 10^{-1} \pm 4.6 \times 10^{-5}$	$2.363 \times 10^{-2} \pm 8.6 \times 10^{-6}$
linear	$1.029 \times 10^{+1} \pm 4.6 \times 10^{-4}$	$2.604 \times 10^{+1} \pm 4.2 \times 10^{-3}$
unscented'95	$3.634 \times 10^{-1} \pm 7.7 \times 10^{-5}$	$9.051 \times 10^{-4} \pm 4.8 \times 10^{-6}$
unscented'02	$2.038 \times 10^{+2} \pm 7.8 \times 10^{-3}$	$8.398 \times 10^{+3} \pm 1.3 \times 10^0$

7420 Table 175: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
7421 tivation=gelu), variance=large  
7422

7423  
7424  
7425  
7426  
7427  
7428  
7429  
7430  
7431  
7432  
7433  
7434  
7435  
7436  
7437  
7438  
7439



7440  
7441  
7442  
7443  
7444  
7445  
7446  
7447  
7448  
7449  
7450  
7451

7444 Figure 95: Probability distributions for Network(architecture=deep, weights=trained, activa-  
7445 tion=gelu), variance=large

7452  
7453  
7454  
7455  
7456  
7457  
7458  
7459  
7460

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-7.048 \times 10^{+1} \pm 2.5 \times 10^{-4}$	$7.272 \times 10^{+3} \pm 9.5 \times 10^{-2}$
analytic	$-7.235 \times 10^{+1}$	$7.475 \times 10^{+3}$
mean-field	$-6.834 \times 10^{+1}$	$9.767 \times 10^{+3}$
linear	$-9.331 \times 10^{+1}$	$1.011 \times 10^{+4}$
unscented'95	$-6.667 \times 10^{+1}$	$6.940 \times 10^{+3}$
unscented'02	$-4.291 \times 10^{+1}$	$1.518 \times 10^{+4}$

7461  
7462  
7463

Table 176: Comparison of moments for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=small

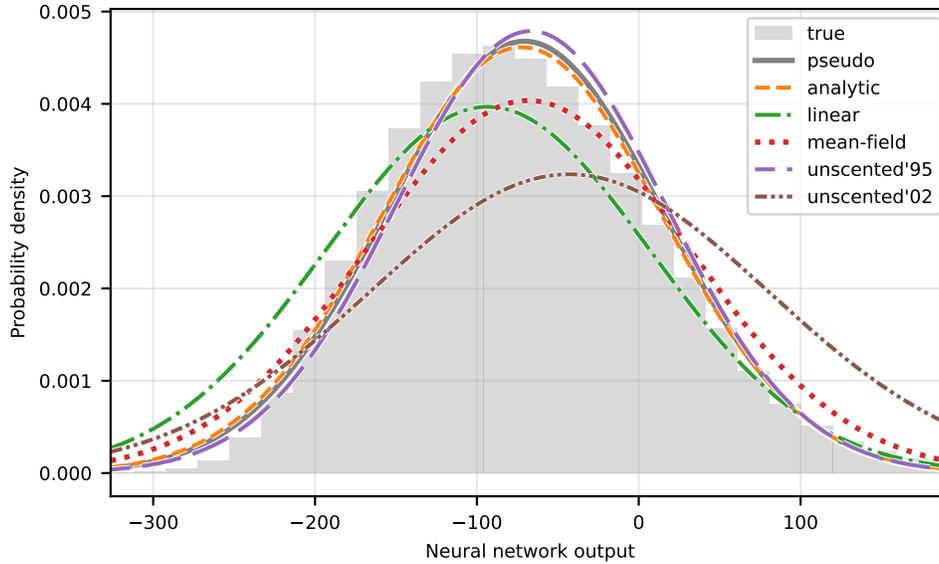
7464  
7465  
7466  
7467  
7468  
7469  
7470  
7471  
7472  
7473

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.311 \times 10^{-1} \pm 7.9 \times 10^{-5}$	0
analytic	$5.682 \times 10^{-1} \pm 5.8 \times 10^{-5}$	$4.297 \times 10^{-4} \pm 1.9 \times 10^{-7}$
mean-field	$1.403 \times 10^0 \pm 6.2 \times 10^{-5}$	$2.437 \times 10^{-2} \pm 2.3 \times 10^{-6}$
linear	$2.471 \times 10^0 \pm 2.7 \times 10^{-5}$	$6.610 \times 10^{-2} \pm 3.0 \times 10^{-6}$
unscented'95	$8.604 \times 10^{-1} \pm 7.9 \times 10^{-5}$	$1.541 \times 10^{-3} \pm 3.0 \times 10^{-7}$
unscented'02	$4.445 \times 10^0 \pm 5.8 \times 10^{-5}$	$2.282 \times 10^{-1} \pm 8.0 \times 10^{-6}$

7474  
7475  
7476

Table 177: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=small

7477  
7478  
7479  
7480  
7481  
7482  
7483  
7484  
7485  
7486  
7487  
7488  
7489  
7490  
7491  
7492  
7493  
7494  
7495  
7496  
7497



7498  
7499

Figure 96: Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=small

7500  
7501  
7502  
7503  
7504  
7505

7506  
7507  
7508  
7509  
7510  
7511  
7512  
7513  
7514

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+3.447 \times 10^{+1} \pm 8.2 \times 10^{-3}$	$3.583 \times 10^{+5} \pm 1.4 \times 10^{+1}$
analytic	$+8.599 \times 10^{+1}$	$4.080 \times 10^{+5}$
mean-field	$+1.463 \times 10^{+2}$	$5.186 \times 10^{+5}$
linear	$-9.331 \times 10^{+1}$	$1.011 \times 10^{+6}$
unscented'95	$+1.674 \times 10^{+2}$	$3.425 \times 10^{+5}$
unscented'02	$+4.944 \times 10^{+3}$	$5.175 \times 10^{+7}$

7515  
7516  
7517

Table 178: Comparison of moments for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=medium

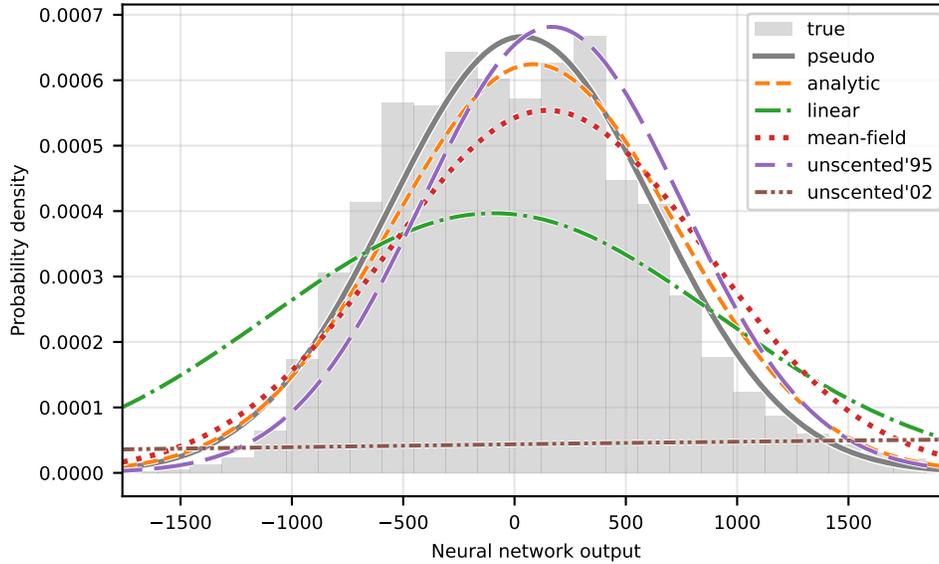
7518  
7519  
7520  
7521  
7522  
7523  
7524  
7525  
7526  
7527

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.827 \times 10^0 \pm 3.2 \times 10^{-4}$	0
analytic	$3.593 \times 10^0 \pm 3.9 \times 10^{-4}$	$8.122 \times 10^{-3} \pm 3.0 \times 10^{-6}$
mean-field	$6.621 \times 10^0 \pm 4.0 \times 10^{-4}$	$5.629 \times 10^{-2} \pm 9.6 \times 10^{-6}$
linear	$1.377 \times 10^{+1} \pm 5.0 \times 10^{-4}$	$4.146 \times 10^{-1} \pm 3.6 \times 10^{-5}$
unscented'95	$5.953 \times 10^0 \pm 4.4 \times 10^{-4}$	$2.517 \times 10^{-2} \pm 3.1 \times 10^{-6}$
unscented'02	$2.730 \times 10^{+2} \pm 2.7 \times 10^{-3}$	$1.029 \times 10^{+2} \pm 4.1 \times 10^{-3}$

7528  
7529  
7530

Table 179: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=medium

7531  
7532  
7533  
7534  
7535  
7536  
7537  
7538  
7539  
7540  
7541  
7542  
7543  
7544  
7545  
7546  
7547  
7548  
7549



7550  
7551  
7552  
7553  
7554  
7555  
7556  
7557  
7558  
7559

Figure 97: Probability distributions for Network(architecture=deep, weights=initialized, activation=gelu residual), variance=medium

7560  
7561  
7562  
7563  
7564  
7565  
7566  
7567  
7568

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.290 \times 10^{+3} \pm 8.3 \times 10^{-2}$	$4.372 \times 10^{+7} \pm 2.1 \times 10^{+3}$
analytic	$+5.838 \times 10^{+3}$	$3.727 \times 10^{+7}$
mean-field	$+4.691 \times 10^{+3}$	$3.766 \times 10^{+7}$
linear	$-9.331 \times 10^{+1}$	$1.011 \times 10^{+8}$
unscented'95	$+7.423 \times 10^{+3}$	$1.861 \times 10^{+7}$
unscented'02	$+4.778 \times 10^{+5}$	$4.569 \times 10^{+11}$

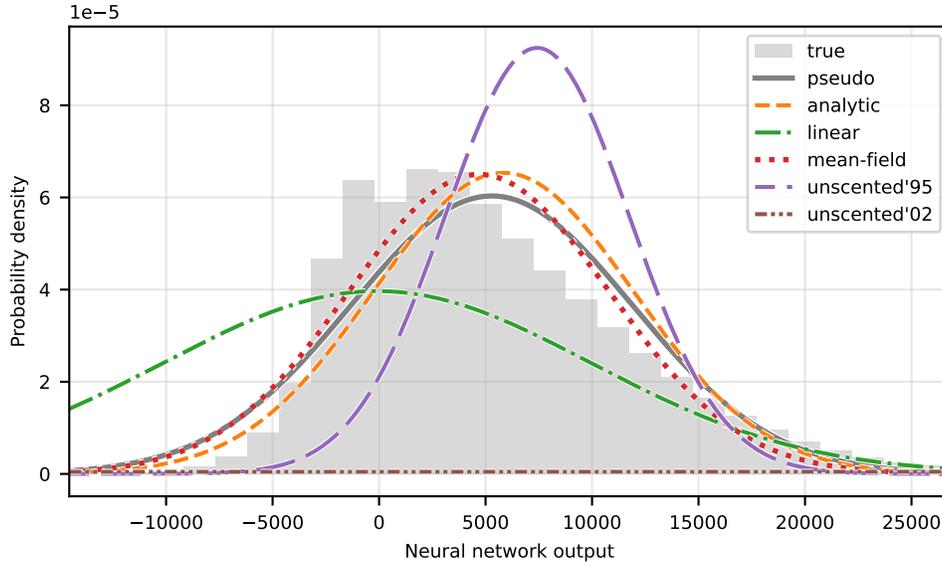
7569 Table 180: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
7570 tion=gelu residual), variance=large  
7571

7572  
7573  
7574  
7575  
7576  
7577  
7578  
7579  
7580  
7581

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.207 \times 10^{+1} \pm 1.7 \times 10^{-3}$	0
analytic	$1.529 \times 10^{+1} \pm 1.5 \times 10^{-3}$	$9.480 \times 10^{-3} \pm 3.3 \times 10^{-6}$
mean-field	$1.129 \times 10^{+1} \pm 1.6 \times 10^{-3}$	$9.404 \times 10^{-3} \pm 3.5 \times 10^{-6}$
linear	$6.621 \times 10^{+1} \pm 1.2 \times 10^{-3}$	$5.681 \times 10^{-1} \pm 4.6 \times 10^{-5}$
unscented'95	$3.666 \times 10^{+1} \pm 1.1 \times 10^{-3}$	$1.919 \times 10^{-1} \pm 1.1 \times 10^{-5}$
unscented'02	$8.146 \times 10^{+3} \pm 9.8 \times 10^{-2}$	$7.773 \times 10^{+3} \pm 3.7 \times 10^{-1}$

7582 Table 181: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
7583 activation=gelu residual), variance=large  
7584

7585  
7586  
7587  
7588  
7589  
7590  
7591  
7592  
7593  
7594  
7595  
7596  
7597  
7598  
7599  
7600  
7601  
7602  
7603  
7604



7606 Figure 98: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
7607 tion=gelu residual), variance=large  
7608

7609  
7610  
7611  
7612  
7613

7614  
7615  
7616  
7617  
7618  
7619  
7620  
7621  
7622

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.117 \times 10^{+1} \pm 1.9 \times 10^{-4}$	$1.663 \times 10^{+3} \pm 6.1 \times 10^{-2}$
analytic	$+4.852 \times 10^{+1}$	$1.777 \times 10^{+3}$
mean-field	$+3.612 \times 10^{+1}$	$9.507 \times 10^{+3}$
linear	$+4.609 \times 10^{+1}$	$6.393 \times 10^{+2}$
unscented'95	$+5.265 \times 10^{+1}$	$1.452 \times 10^{+3}$
unscented'02	$+7.208 \times 10^{+1}$	$1.990 \times 10^{+3}$

7623  
7624  
7625

Table 182: Comparison of moments for Network(architecture=deep, weights=trained, activation=gelu residual), variance=small

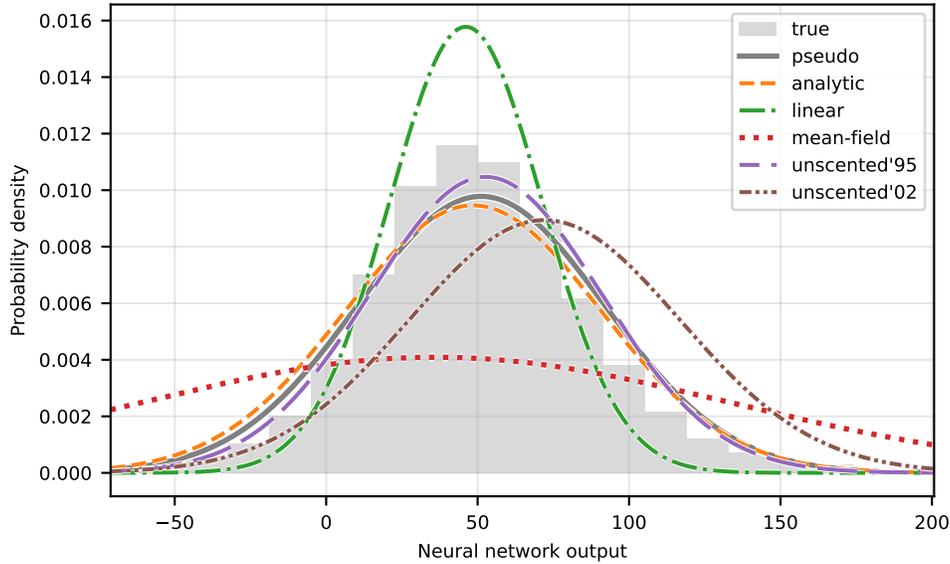
7626  
7627  
7628  
7629  
7630  
7631  
7632  
7633  
7634  
7635

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$7.140 \times 10^{-1} \pm 1.0 \times 10^{-4}$	0
analytic	$8.277 \times 10^{-1} \pm 1.2 \times 10^{-4}$	$3.241 \times 10^{-3} \pm 1.3 \times 10^{-6}$
mean-field	$7.635 \times 10^0 \pm 1.3 \times 10^{-4}$	$1.554 \times 10^0 \pm 8.9 \times 10^{-5}$
linear	$1.621 \times 10^0 \pm 6.2 \times 10^{-5}$	$1.780 \times 10^{-1} \pm 1.1 \times 10^{-5}$
unscented'95	$6.240 \times 10^{-1} \pm 8.5 \times 10^{-5}$	$5.081 \times 10^{-3} \pm 2.3 \times 10^{-6}$
unscented'02	$3.337 \times 10^0 \pm 7.9 \times 10^{-5}$	$1.400 \times 10^{-1} \pm 9.3 \times 10^{-6}$

7636  
7637  
7638

Table 183: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=gelu residual), variance=small

7639  
7640  
7641  
7642  
7643  
7644  
7645  
7646  
7647  
7648  
7649  
7650  
7651  
7652  
7653  
7654  
7655  
7656  
7657



7658  
7659  
7660  
7661  
7662  
7663  
7664  
7665  
7666  
7667

Figure 99: Probability distributions for Network(architecture=deep, weights=trained, activation=gelu residual), variance=small

7668  
7669  
7670  
7671  
7672  
7673  
7674  
7675  
7676

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-1.195 \times 10^{+2} \pm 5.4 \times 10^{-3}$	$1.574 \times 10^{+5} \pm 1.4 \times 10^{+1}$
analytic	$-1.198 \times 10^{+2}$	$2.787 \times 10^{+5}$
mean-field	$-7.058 \times 10^{+1}$	$5.193 \times 10^{+5}$
linear	$+4.609 \times 10^{+1}$	$6.393 \times 10^{+4}$
unscented'95	$-3.022 \times 10^{+1}$	$1.027 \times 10^{+5}$
unscented'02	$+2.644 \times 10^{+3}$	$1.357 \times 10^{+7}$

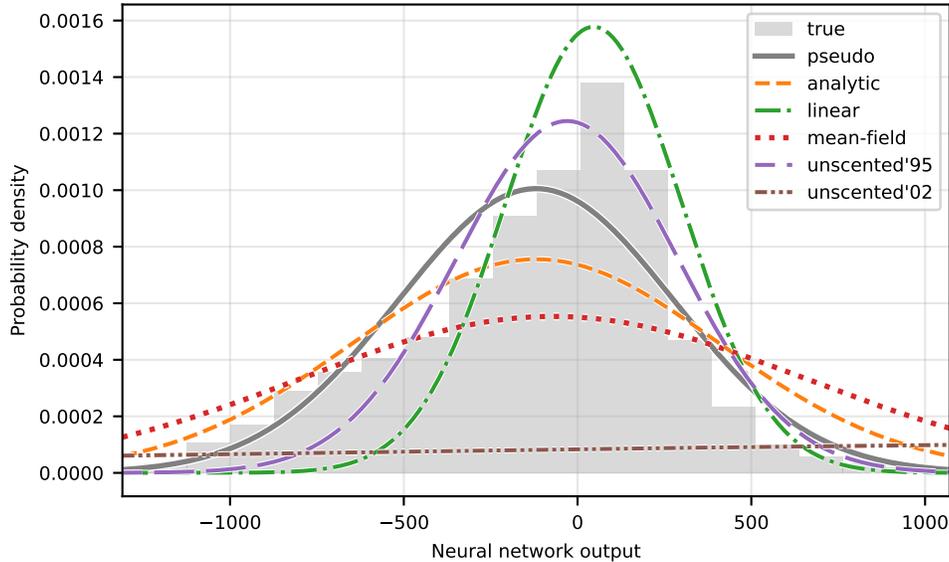
7677 Table 184: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
7678 tion=gelu residual), variance=medium  
7679

7680  
7681  
7682  
7683  
7684  
7685  
7686  
7687  
7688  
7689

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.730 \times 10^0 \pm 5.6 \times 10^{-4}$	0
analytic	$6.343 \times 10^0 \pm 4.1 \times 10^{-4}$	$9.965 \times 10^{-2} \pm 3.3 \times 10^{-5}$
mean-field	$1.362 \times 10^{+1} \pm 6.8 \times 10^{-4}$	$5.604 \times 10^{-1} \pm 9.9 \times 10^{-5}$
linear	$8.801 \times 10^0 \pm 4.3 \times 10^{-4}$	$2.407 \times 10^{-1} \pm 2.0 \times 10^{-5}$
unscented'95	$4.771 \times 10^0 \pm 3.7 \times 10^{-4}$	$6.496 \times 10^{-2} \pm 1.4 \times 10^{-5}$
unscented'02	$1.746 \times 10^{+2} \pm 3.9 \times 10^{-3}$	$6.463 \times 10^{+1} \pm 5.7 \times 10^{-3}$

7690 Table 185: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
7691 tivation=gelu residual), variance=medium  
7692

7693  
7694  
7695  
7696  
7697  
7698  
7699  
7700  
7701  
7702  
7703  
7704  
7705  
7706  
7707  
7708  
7709



7710  
7711  
7712  
7713  
7714  
7715  
7716  
7717  
7718  
7719  
7720  
7721

7714 Figure 100: Probability distributions for Network(architecture=deep, weights=trained, activa-  
7715 tion=gelu residual), variance=medium

7722  
7723  
7724  
7725  
7726  
7727  
7728  
7729  
7730

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.667 \times 10^{+3} \pm 7.4 \times 10^{-2}$	$3.189 \times 10^{+7} \pm 1.7 \times 10^{+3}$
analytic	$+2.912 \times 10^{+3}$	$3.073 \times 10^{+7}$
mean-field	$+1.762 \times 10^{+3}$	$3.778 \times 10^{+7}$
linear	$+4.609 \times 10^{+1}$	$6.393 \times 10^{+6}$
unscented'95	$+4.419 \times 10^{+3}$	$1.409 \times 10^{+7}$
unscented'02	$+2.537 \times 10^{+5}$	$1.287 \times 10^{+11}$

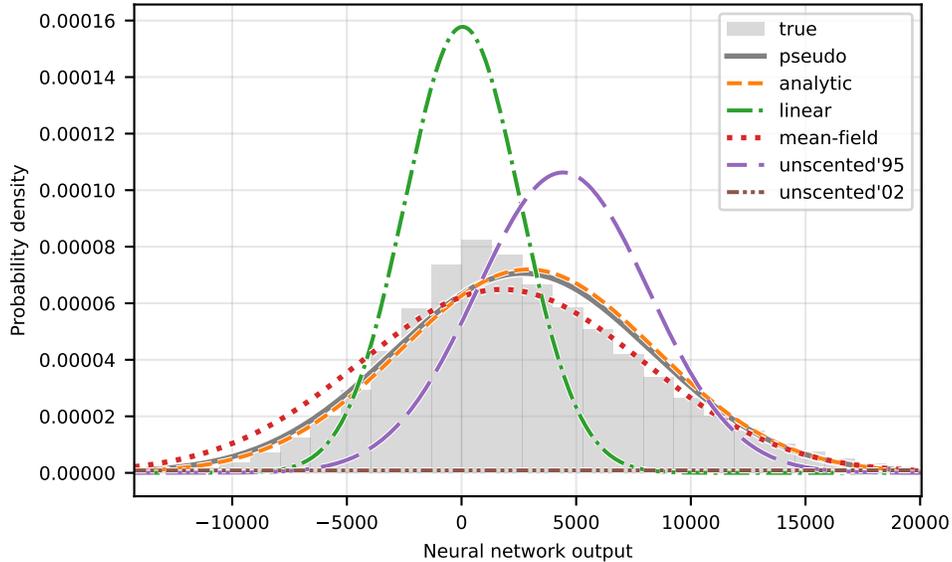
7731 Table 186: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
7732 tion=gelu residual), variance=large  
7733  
7734

7735  
7736  
7737  
7738  
7739  
7740  
7741  
7742  
7743

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.428 \times 10^0 \pm 1.3 \times 10^{-3}$	0
analytic	$7.260 \times 10^0 \pm 1.4 \times 10^{-3}$	$1.279 \times 10^{-3} \pm 1.0 \times 10^{-6}$
mean-field	$1.205 \times 10^{+1} \pm 9.7 \times 10^{-4}$	$2.044 \times 10^{-2} \pm 5.8 \times 10^{-6}$
linear	$4.083 \times 10^{+1} \pm 8.4 \times 10^{-4}$	$5.115 \times 10^{-1} \pm 1.7 \times 10^{-5}$
unscented'95	$3.094 \times 10^{+1} \pm 8.9 \times 10^{-4}$	$1.774 \times 10^{-1} \pm 1.3 \times 10^{-5}$
unscented'02	$4.662 \times 10^{+3} \pm 6.4 \times 10^{-2}$	$3.001 \times 10^{+3} \pm 1.6 \times 10^{-1}$

7744 Table 187: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
7745 tivation=gelu residual), variance=large  
7746  
7747

7748  
7749  
7750  
7751  
7752  
7753  
7754  
7755  
7756  
7757  
7758  
7759  
7760  
7761  
7762  
7763  
7764  
7765  
7766



7767  
7768  
7769  
7770  
7771  
7772  
7773  
7774  
7775

7768 Figure 101: Probability distributions for Network(architecture=deep, weights=trained, activa-  
7769 tion=gelu residual), variance=large

7776  
7777  
7778  
7779  
7780  
7781  
7782  
7783  
7784

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.616 \times 10^0 \pm 5.0 \times 10^{-7}$	$5.245 \times 10^{-3} \pm 1.4 \times 10^{-7}$
analytic	$+2.615 \times 10^0$	$5.611 \times 10^{-3}$
mean-field	$+2.634 \times 10^0$	$1.284 \times 10^{-2}$
linear	$+2.611 \times 10^0$	$1.690 \times 10^{-2}$
unscented'95	$+2.603 \times 10^0$	$5.793 \times 10^{-3}$
unscented'02	$+2.611 \times 10^0$	$1.690 \times 10^{-2}$

7785  
7786  
7787  
7788

Table 188: Comparison of moments for Network(architecture=deep, weights=initialized, activation=relu), variance=small

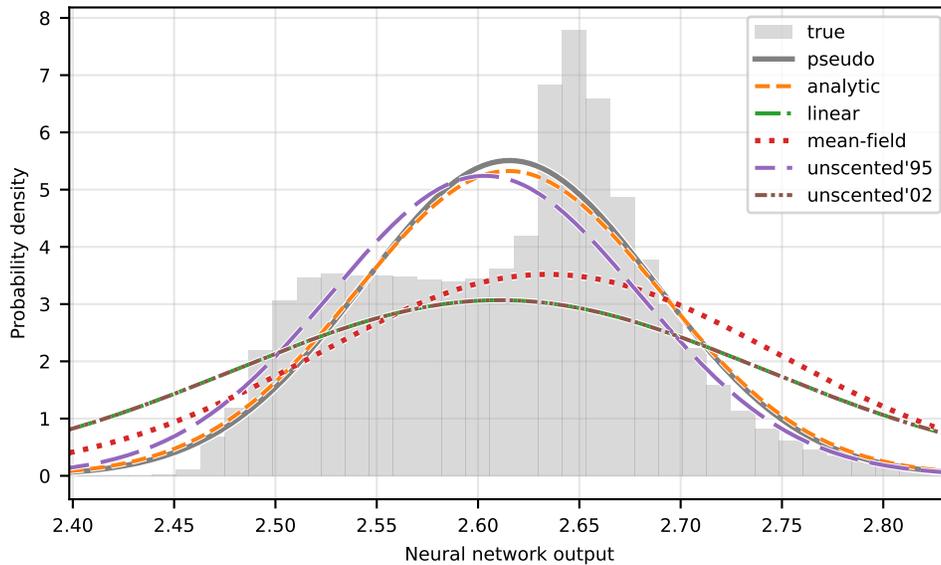
7789  
7790  
7791  
7792  
7793  
7794  
7795  
7796  
7797

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.790 \times 10^{-2} \pm 3.4 \times 10^{-6}$	0
analytic	$2.852 \times 10^{-2} \pm 4.4 \times 10^{-6}$	$1.186 \times 10^{-3} \pm 8.9 \times 10^{-7}$
mean-field	$1.245 \times 10^{-1} \pm 2.6 \times 10^{-6}$	$3.103 \times 10^{-1} \pm 2.0 \times 10^{-5}$
linear	$1.724 \times 10^{-1} \pm 3.3 \times 10^{-6}$	$5.280 \times 10^{-1} \pm 2.9 \times 10^{-5}$
unscented'95	$4.575 \times 10^{-2} \pm 3.0 \times 10^{-6}$	$1.660 \times 10^{-2} \pm 1.8 \times 10^{-6}$
unscented'02	$1.724 \times 10^{-1} \pm 3.3 \times 10^{-6}$	$5.280 \times 10^{-1} \pm 2.9 \times 10^{-5}$

7798  
7799  
7800

Table 189: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=relu), variance=small

7801  
7802  
7803  
7804  
7805  
7806  
7807  
7808  
7809  
7810  
7811  
7812  
7813  
7814  
7815  
7816  
7817



7821  
7822  
7823  
7824  
7825  
7826  
7827  
7828  
7829

Figure 102: Probability distributions for Network(architecture=deep, weights=initialized, activation=relu), variance=small

7830  
7831  
7832  
7833  
7834  
7835  
7836  
7837  
7838

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+2.429 \times 10^0 \pm 6.6 \times 10^{-6}$	$7.734 \times 10^{-2} \pm 6.4 \times 10^{-6}$
analytic	$+2.411 \times 10^0$	$7.888 \times 10^{-2}$
mean-field	$+2.392 \times 10^0$	$3.227 \times 10^{-1}$
linear	$+2.611 \times 10^0$	$1.690 \times 10^0$
unscented'95	$+2.355 \times 10^0$	$3.979 \times 10^{-2}$
unscented'02	$+2.611 \times 10^0$	$1.690 \times 10^0$

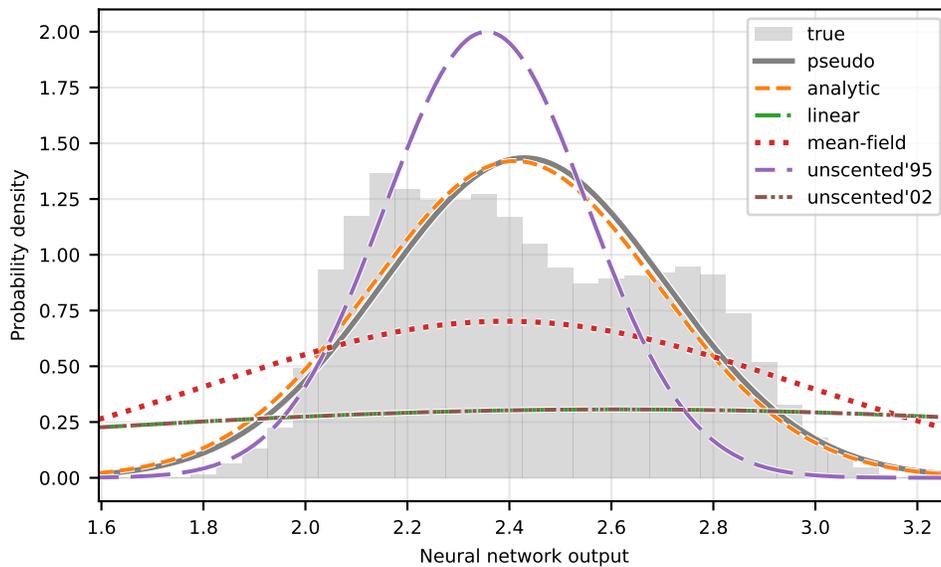
7839 Table 190: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
7840 tion=relu), variance=medium  
7841  
7842

7843  
7844  
7845  
7846  
7847  
7848  
7849  
7850  
7851

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$6.887 \times 10^{-2} \pm 1.6 \times 10^{-5}$	0
analytic	$6.803 \times 10^{-2} \pm 1.4 \times 10^{-5}$	$2.270 \times 10^{-3} \pm 1.9 \times 10^{-6}$
mean-field	$4.132 \times 10^{-1} \pm 2.6 \times 10^{-5}$	$8.809 \times 10^{-1} \pm 1.3 \times 10^{-4}$
linear	$1.555 \times 10^0 \pm 4.6 \times 10^{-5}$	$9.094 \times 10^0 \pm 8.8 \times 10^{-4}$
unscented'95	$1.607 \times 10^{-1} \pm 1.9 \times 10^{-5}$	$1.250 \times 10^{-1} \pm 1.8 \times 10^{-5}$
unscented'02	$1.555 \times 10^0 \pm 4.6 \times 10^{-5}$	$9.094 \times 10^0 \pm 8.8 \times 10^{-4}$

7852 Table 191: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
7853 activation=relu), variance=medium  
7854  
7855

7856  
7857  
7858  
7859  
7860  
7861  
7862  
7863  
7864  
7865  
7866  
7867  
7868  
7869  
7870  
7871  
7872  
7873  
7874



7875  
7876 Figure 103: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
7877 tion=relu), variance=medium  
7878  
7879  
7880  
7881  
7882  
7883

7884  
7885  
7886  
7887  
7888  
7889  
7890  
7891  
7892

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+7.572 \times 10^0 \pm 1.0 \times 10^{-4}$	$1.890 \times 10^{+1} \pm 2.5 \times 10^{-3}$
analytic	$+8.042 \times 10^0$	$9.281 \times 10^0$
mean-field	$+7.961 \times 10^0$	$1.191 \times 10^{+1}$
linear	$+2.611 \times 10^0$	$1.690 \times 10^{+2}$
unscented'95	$+8.271 \times 10^0$	$1.248 \times 10^{+1}$
unscented'02	$-3.548 \times 10^{+1}$	$3.074 \times 10^{+3}$

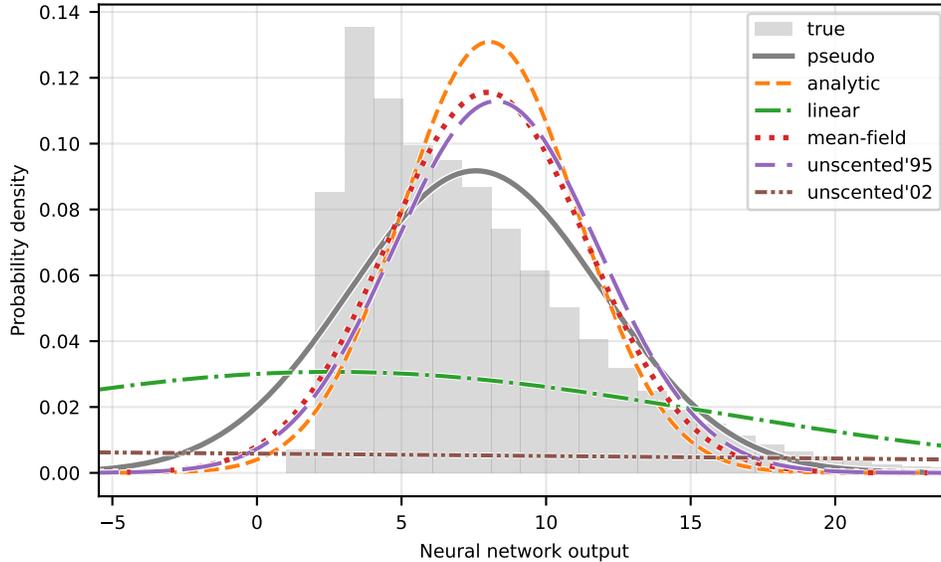
Table 192: Comparison of moments for Network(architecture=deep, weights=initialized, activation=relu), variance=large

7893  
7894  
7895  
7896  
7897  
7898  
7899  
7900  
7901  
7902  
7903  
7904  
7905

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.696 \times 10^{-1} \pm 8.3 \times 10^{-5}$	0
analytic	$6.699 \times 10^{-1} \pm 7.2 \times 10^{-5}$	$1.070 \times 10^{-1} \pm 3.2 \times 10^{-5}$
mean-field	$5.803 \times 10^{-1} \pm 6.7 \times 10^{-5}$	$4.992 \times 10^{-2} \pm 2.3 \times 10^{-5}$
linear	$3.750 \times 10^0 \pm 1.6 \times 10^{-4}$	$3.525 \times 10^0 \pm 6.0 \times 10^{-4}$
unscented'95	$6.525 \times 10^{-1} \pm 6.9 \times 10^{-5}$	$5.053 \times 10^{-2} \pm 1.9 \times 10^{-5}$
unscented'02	$2.595 \times 10^{+1} \pm 9.0 \times 10^{-4}$	$1.273 \times 10^{+2} \pm 1.7 \times 10^{-2}$

Table 193: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=relu), variance=large

7906  
7907  
7908  
7909  
7910  
7911  
7912  
7913  
7914  
7915  
7916  
7917  
7918  
7919  
7920  
7921  
7922  
7923  
7924  
7925  
7926  
7927



7928  
7929  
7930  
7931  
7932  
7933  
7934  
7935  
7936  
7937

Figure 104: Probability distributions for Network(architecture=deep, weights=initialized, activation=relu), variance=large

7938  
7939  
7940  
7941  
7942  
7943  
7944  
7945  
7946

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-6.083 \times 10^{-1} \pm 8.5 \times 10^{-7}$	$4.457 \times 10^{-2} \pm 6.2 \times 10^{-7}$
analytic	$-6.074 \times 10^{-1}$	$4.083 \times 10^{-2}$
mean-field	$-6.079 \times 10^{-1}$	$1.516 \times 10^{-2}$
linear	$-6.030 \times 10^{-1}$	$5.371 \times 10^{-2}$
unscented'95	$-5.946 \times 10^{-1}$	$6.185 \times 10^{-2}$
unscented'02	$-6.030 \times 10^{-1}$	$5.371 \times 10^{-2}$

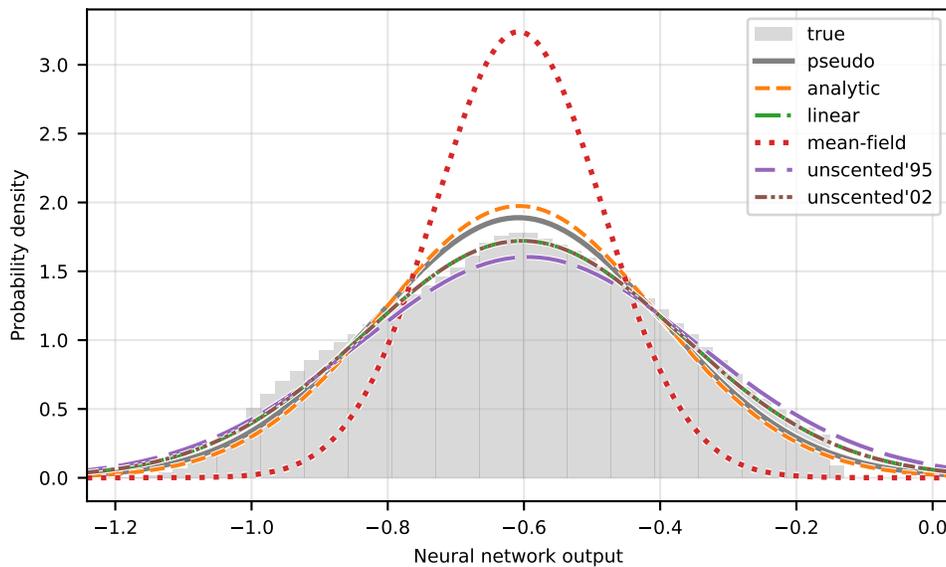
7947 Table 194: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
7948 tion=relu), variance=small  
7949

7950  
7951  
7952  
7953  
7954  
7955  
7956  
7957  
7958  
7959

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.680 \times 10^{-2} \pm 3.6 \times 10^{-6}$	0
analytic	$3.606 \times 10^{-2} \pm 4.0 \times 10^{-6}$	$1.875 \times 10^{-3} \pm 5.8 \times 10^{-7}$
mean-field	$1.644 \times 10^{-1} \pm 3.0 \times 10^{-6}$	$2.092 \times 10^{-1} \pm 4.6 \times 10^{-6}$
linear	$2.810 \times 10^{-2} \pm 2.6 \times 10^{-6}$	$9.589 \times 10^{-3} \pm 1.5 \times 10^{-6}$
unscented'95	$5.845 \times 10^{-2} \pm 2.8 \times 10^{-6}$	$3.215 \times 10^{-2} \pm 2.8 \times 10^{-6}$
unscented'02	$2.810 \times 10^{-2} \pm 2.6 \times 10^{-6}$	$9.589 \times 10^{-3} \pm 1.5 \times 10^{-6}$

7960 Table 195: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
7961 tivation=relu), variance=small  
7962

7963  
7964  
7965  
7966  
7967  
7968  
7969  
7970  
7971  
7972  
7973  
7974  
7975  
7976  
7977  
7978  
7979



7980  
7981  
7982  
7983  
7984  
7985  
7986  
7987  
7988  
7989  
7990  
7991

7984 Figure 105: Probability distributions for Network(architecture=deep, weights=trained, activa-  
7985 tion=relu), variance=small

7992  
7993  
7994  
7995  
7996  
7997  
7998  
7999  
8000  
8001  
8002  
8003  
8004  
8005  
8006  
8007  
8008  
8009  
8010  
8011  
8012  
8013  
8014  
8015  
8016  
8017  
8018  
8019  
8020  
8021  
8022  
8023  
8024  
8025  
8026  
8027  
8028  
8029  
8030  
8031  
8032  
8033  
8034  
8035  
8036  
8037  
8038  
8039  
8040  
8041  
8042  
8043  
8044  
8045

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.734 \times 10^{-1} \pm 1.3 \times 10^{-5}$	$2.430 \times 10^{-1} \pm 1.9 \times 10^{-5}$
analytic	$-1.693 \times 10^{-1}$	$1.785 \times 10^{-1}$
mean-field	$-1.942 \times 10^{-1}$	$3.867 \times 10^{-1}$
linear	$-6.030 \times 10^{-1}$	$5.371 \times 10^0$
unscented'95	$-2.971 \times 10^{-1}$	$3.028 \times 10^{-1}$
unscented'02	$-6.030 \times 10^{-1}$	$5.371 \times 10^0$

Table 196: Comparison of moments for Network(architecture=deep, weights=trained, activation=relu), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.927 \times 10^{-2} \pm 2.3 \times 10^{-5}$	0
analytic	$1.696 \times 10^{-1} \pm 1.8 \times 10^{-5}$	$4.375 \times 10^{-2} \pm 8.1 \times 10^{-6}$
mean-field	$1.932 \times 10^{-1} \pm 2.4 \times 10^{-5}$	$7.632 \times 10^{-2} \pm 2.6 \times 10^{-5}$
linear	$2.091 \times 10^0 \pm 6.3 \times 10^{-5}$	$9.229 \times 10^0 \pm 8.3 \times 10^{-4}$
unscented'95	$6.066 \times 10^{-2} \pm 2.1 \times 10^{-5}$	$1.422 \times 10^{-2} \pm 9.2 \times 10^{-6}$
unscented'02	$2.091 \times 10^0 \pm 6.3 \times 10^{-5}$	$9.229 \times 10^0 \pm 8.3 \times 10^{-4}$

Table 197: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=relu), variance=medium

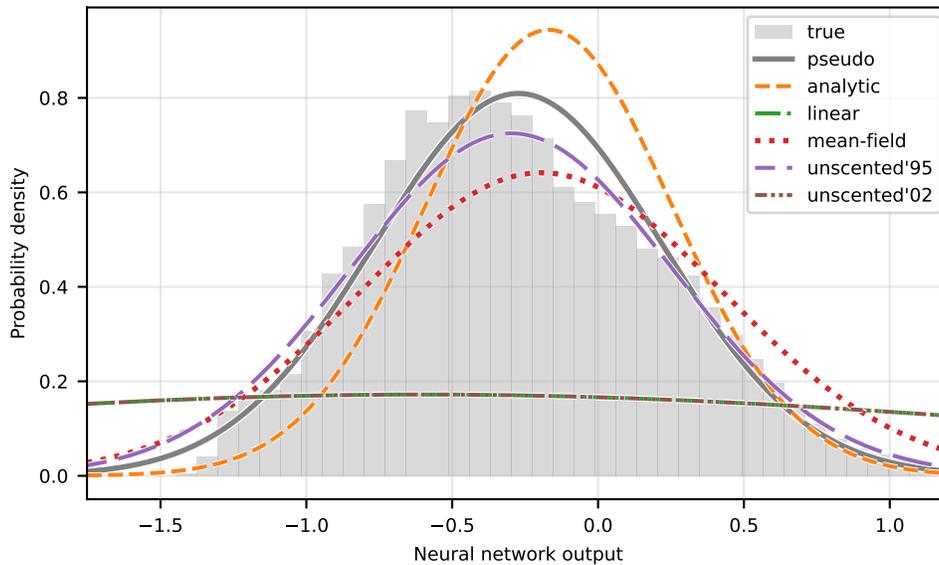


Figure 106: Probability distributions for Network(architecture=deep, weights=trained, activation=relu), variance=medium

8046  
8047  
8048  
8049  
8050  
8051  
8052  
8053  
8054

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+3.361 \times 10^0 \pm 1.5 \times 10^{-4}$	$1.637 \times 10^{+1} \pm 2.3 \times 10^{-3}$
analytic	$+3.930 \times 10^0$	$8.832 \times 10^0$
mean-field	$+4.131 \times 10^0$	$1.254 \times 10^{+1}$
linear	$-6.030 \times 10^{-1}$	$5.371 \times 10^{+2}$
unscented'95	$+3.242 \times 10^0$	$1.518 \times 10^{+1}$
unscented'02	$+1.249 \times 10^{+3}$	$3.124 \times 10^{+6}$

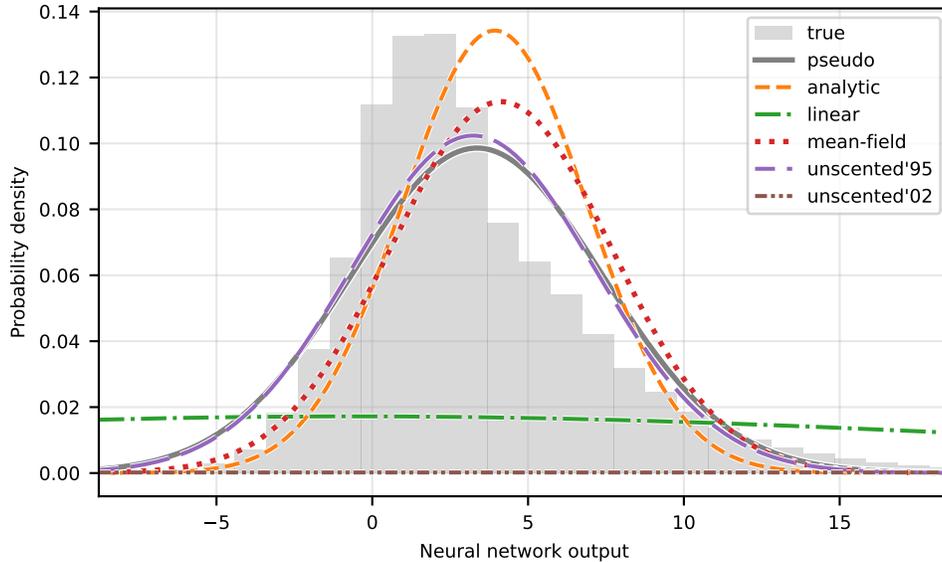
Table 198: Comparison of moments for Network(architecture=deep, weights=trained, activation=relu), variance=large

8055  
8056  
8057  
8058  
8059  
8060  
8061  
8062  
8063  
8064  
8065  
8066  
8067

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.964 \times 10^{-1} \pm 1.0 \times 10^{-4}$	0
analytic	$6.038 \times 10^{-1} \pm 1.1 \times 10^{-4}$	$8.819 \times 10^{-2} \pm 3.3 \times 10^{-5}$
mean-field	$5.667 \times 10^{-1} \pm 8.2 \times 10^{-5}$	$3.443 \times 10^{-2} \pm 1.7 \times 10^{-5}$
linear	$7.814 \times 10^0 \pm 3.5 \times 10^{-4}$	$1.464 \times 10^{+1} \pm 2.3 \times 10^{-3}$
unscented'95	$3.663 \times 10^{-1} \pm 9.9 \times 10^{-5}$	$1.817 \times 10^{-3} \pm 5.0 \times 10^{-6}$
unscented'02	$8.674 \times 10^{+2} \pm 3.1 \times 10^{-2}$	$1.428 \times 10^{+5} \pm 2.0 \times 10^{+1}$

Table 199: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=relu), variance=large

8068  
8069  
8070  
8071  
8072  
8073  
8074  
8075  
8076  
8077  
8078  
8079  
8080  
8081  
8082  
8083  
8084  
8085  
8086  
8087



8088  
8089  
8090  
8091  
8092  
8093  
8094  
8095  
8096  
8097  
8098  
8099

Figure 107: Probability distributions for Network(architecture=deep, weights=trained, activation=relu), variance=large

8100  
8101  
8102  
8103  
8104  
8105  
8106  
8107  
8108  
8109  
8110  
8111  
8112  
8113  
8114  
8115  
8116  
8117  
8118  
8119  
8120  
8121  
8122  
8123  
8124  
8125  
8126  
8127  
8128  
8129  
8130  
8131  
8132  
8133  
8134  
8135  
8136  
8137  
8138  
8139  
8140  
8141  
8142  
8143  
8144  
8145  
8146  
8147  
8148  
8149  
8150  
8151  
8152  
8153

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+6.455 \times 10^{+1} \pm 5.0 \times 10^{-4}$	$6.699 \times 10^{+3} \pm 1.8 \times 10^{-1}$
analytic	$+6.526 \times 10^{+1}$	$6.392 \times 10^{+3}$
mean-field	$+6.421 \times 10^{+1}$	$8.737 \times 10^{+3}$
linear	$+4.588 \times 10^{+1}$	$3.313 \times 10^{+3}$
unscented'95	$+6.665 \times 10^{+1}$	$5.455 \times 10^{+3}$
unscented'02	$+4.588 \times 10^{+1}$	$3.313 \times 10^{+3}$

Table 200: Comparison of moments for Network(architecture=deep, weights=initialized, activation=relu residual), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.234 \times 10^0 \pm 8.6 \times 10^{-5}$	0
analytic	$1.204 \times 10^0 \pm 6.8 \times 10^{-5}$	$5.794 \times 10^{-4} \pm 5.9 \times 10^{-7}$
mean-field	$1.820 \times 10^0 \pm 7.3 \times 10^{-5}$	$1.932 \times 10^{-2} \pm 4.1 \times 10^{-6}$
linear	$2.105 \times 10^0 \pm 5.6 \times 10^{-5}$	$1.254 \times 10^{-1} \pm 7.0 \times 10^{-6}$
unscented'95	$1.224 \times 10^0 \pm 8.9 \times 10^{-5}$	$1.020 \times 10^{-2} \pm 2.4 \times 10^{-6}$
unscented'02	$2.105 \times 10^0 \pm 5.6 \times 10^{-5}$	$1.254 \times 10^{-1} \pm 7.0 \times 10^{-6}$

Table 201: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=relu residual), variance=small

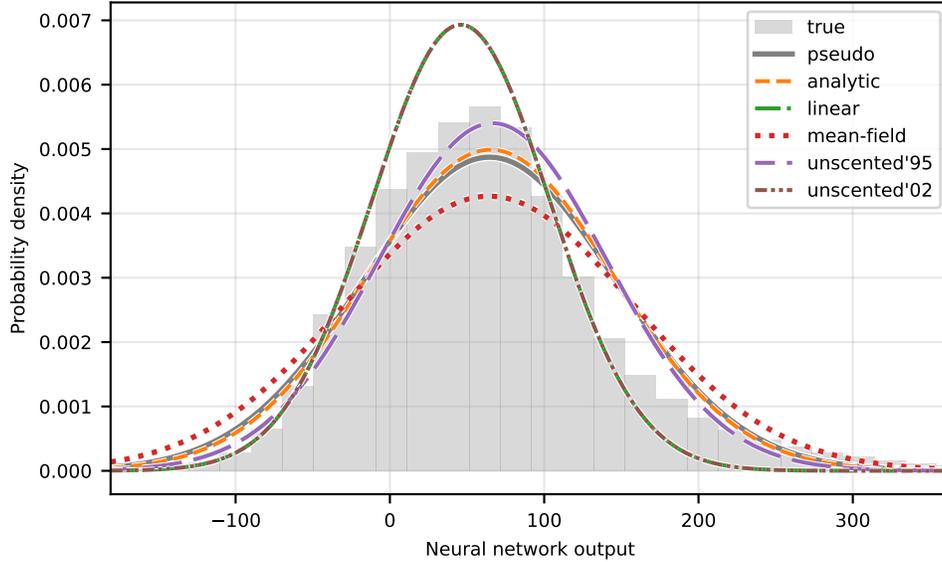


Figure 108: Probability distributions for Network(architecture=deep, weights=initialized, activation=relu residual), variance=small

8154  
8155  
8156  
8157  
8158  
8159  
8160  
8161  
8162  
8163  
8164  
8165  
8166  
8167  
8168  
8169  
8170  
8171  
8172  
8173  
8174  
8175  
8176  
8177  
8178  
8179  
8180  
8181  
8182  
8183  
8184  
8185  
8186  
8187  
8188  
8189  
8190  
8191  
8192  
8193  
8194  
8195  
8196  
8197  
8198  
8199  
8200  
8201  
8202  
8203  
8204  
8205  
8206  
8207

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+1.366 \times 10^{+2} \pm 7.3 \times 10^{-3}$	$3.198 \times 10^{+5} \pm 1.8 \times 10^{+1}$
analytic	$+1.641 \times 10^{+2}$	$3.460 \times 10^{+5}$
mean-field	$+2.213 \times 10^{+2}$	$4.806 \times 10^{+5}$
linear	$+4.588 \times 10^{+1}$	$3.313 \times 10^{+5}$
unscented'95	$+2.183 \times 10^{+2}$	$3.264 \times 10^{+5}$
unscented'02	$+8.620 \times 10^{+4}$	$1.484 \times 10^{+10}$

Table 202: Comparison of moments for Network(architecture=deep, weights=initialized, activation=relu residual), variance=medium

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.050 \times 10^0 \pm 4.1 \times 10^{-4}$	0
analytic	$2.867 \times 10^0 \pm 4.1 \times 10^{-4}$	$2.761 \times 10^{-3} \pm 2.3 \times 10^{-6}$
mean-field	$6.383 \times 10^0 \pm 3.3 \times 10^{-4}$	$5.890 \times 10^{-2} \pm 1.5 \times 10^{-5}$
linear	$3.815 \times 10^0 \pm 3.2 \times 10^{-4}$	$1.318 \times 10^{-2} \pm 2.9 \times 10^{-6}$
unscented'95	$4.264 \times 10^0 \pm 4.0 \times 10^{-4}$	$1.054 \times 10^{-2} \pm 2.0 \times 10^{-6}$
unscented'02	$5.053 \times 10^{+3} \pm 7.2 \times 10^{-2}$	$3.478 \times 10^{+4} \pm 2.0 \times 10^0$

Table 203: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=relu residual), variance=medium

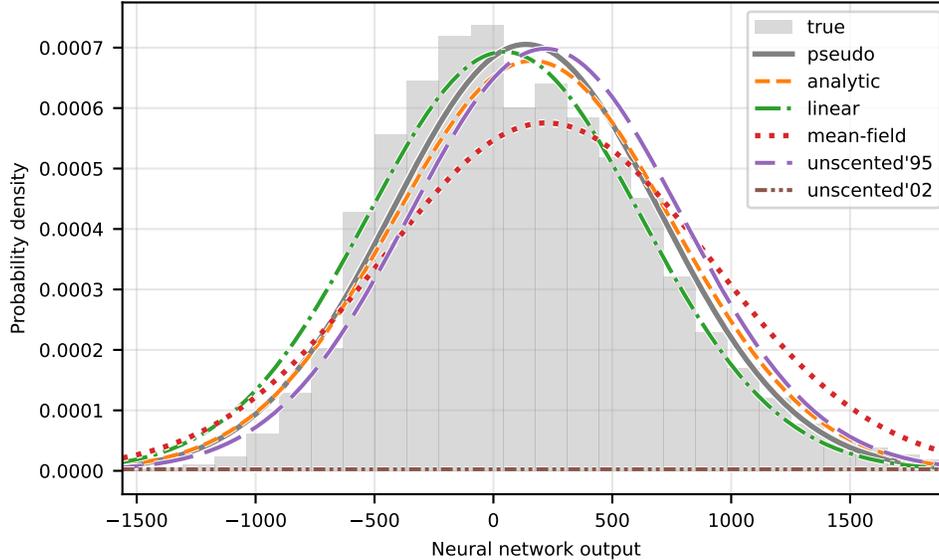


Figure 109: Probability distributions for Network(architecture=deep, weights=initialized, activation=relu residual), variance=medium

8208  
8209  
8210  
8211  
8212  
8213  
8214  
8215  
8216  
8217  
8218  
8219  
8220  
8221  
8222  
8223  
8224  
8225  
8226  
8227  
8228  
8229  
8230  
8231  
8232  
8233  
8234  
8235  
8236  
8237  
8238  
8239  
8240  
8241  
8242  
8243  
8244  
8245  
8246  
8247  
8248  
8249  
8250  
8251  
8252  
8253  
8254  
8255  
8256  
8257  
8258  
8259  
8260  
8261

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.307 \times 10^{+3} \pm 7.5 \times 10^{-2}$	$4.356 \times 10^{+7} \pm 2.1 \times 10^{+3}$
analytic	$+5.810 \times 10^{+3}$	$3.855 \times 10^{+7}$
mean-field	$+4.672 \times 10^{+3}$	$3.783 \times 10^{+7}$
linear	$+4.588 \times 10^{+1}$	$3.313 \times 10^{+7}$
unscented'95	$+7.418 \times 10^{+3}$	$1.873 \times 10^{+7}$
unscented'02	$+1.113 \times 10^{+6}$	$2.477 \times 10^{+12}$

Table 204: Comparison of moments for Network(architecture=deep, weights=initialized, activation=relu residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$1.207 \times 10^{+1} \pm 1.6 \times 10^{-3}$	0
analytic	$1.484 \times 10^{+1} \pm 1.4 \times 10^{-3}$	$6.496 \times 10^{-3} \pm 2.6 \times 10^{-6}$
mean-field	$1.125 \times 10^{+1} \pm 1.6 \times 10^{-3}$	$9.378 \times 10^{-3} \pm 3.2 \times 10^{-6}$
linear	$6.476 \times 10^{+1} \pm 1.1 \times 10^{-3}$	$3.349 \times 10^{-1} \pm 1.2 \times 10^{-5}$
unscented'95	$3.635 \times 10^{+1} \pm 9.3 \times 10^{-4}$	$1.881 \times 10^{-1} \pm 1.1 \times 10^{-5}$
unscented'02	$1.909 \times 10^{+4} \pm 2.3 \times 10^{-1}$	$4.251 \times 10^{+4} \pm 2.0 \times 10^0$

Table 205: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=relu residual), variance=large

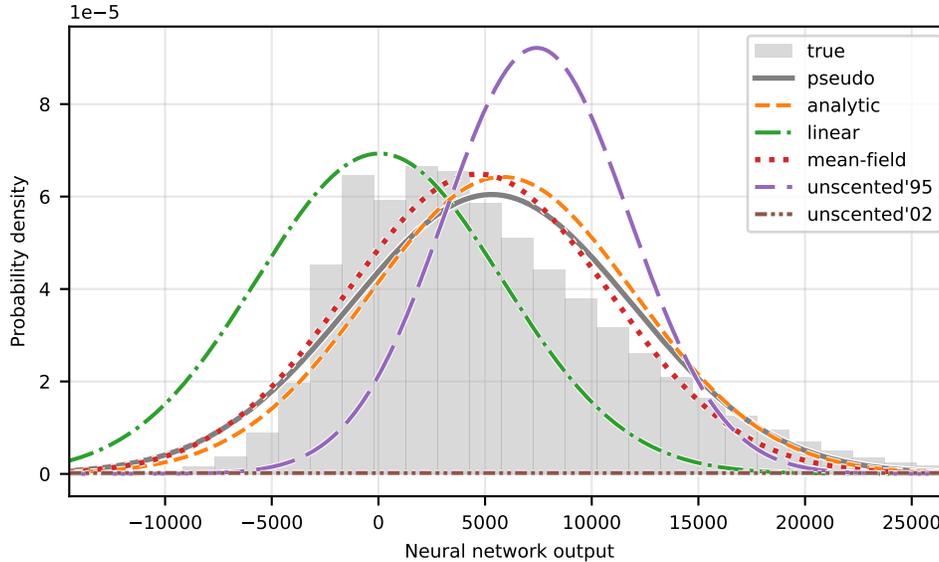


Figure 110: Probability distributions for Network(architecture=deep, weights=initialized, activation=relu residual), variance=large

8262  
8263  
8264  
8265  
8266  
8267  
8268  
8269  
8270

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-2.520 \times 10^{+1} \pm 4.7 \times 10^{-4}$	$1.799 \times 10^{+3} \pm 1.2 \times 10^{-1}$
analytic	$-2.716 \times 10^{+1}$	$1.354 \times 10^{+3}$
mean-field	$-3.235 \times 10^{+1}$	$9.135 \times 10^{+3}$
linear	$-4.096 \times 10^{+1}$	$6.231 \times 10^{+3}$
unscented'95	$-1.945 \times 10^{+1}$	$1.267 \times 10^{+3}$
unscented'02	$-4.096 \times 10^{+1}$	$6.231 \times 10^{+3}$

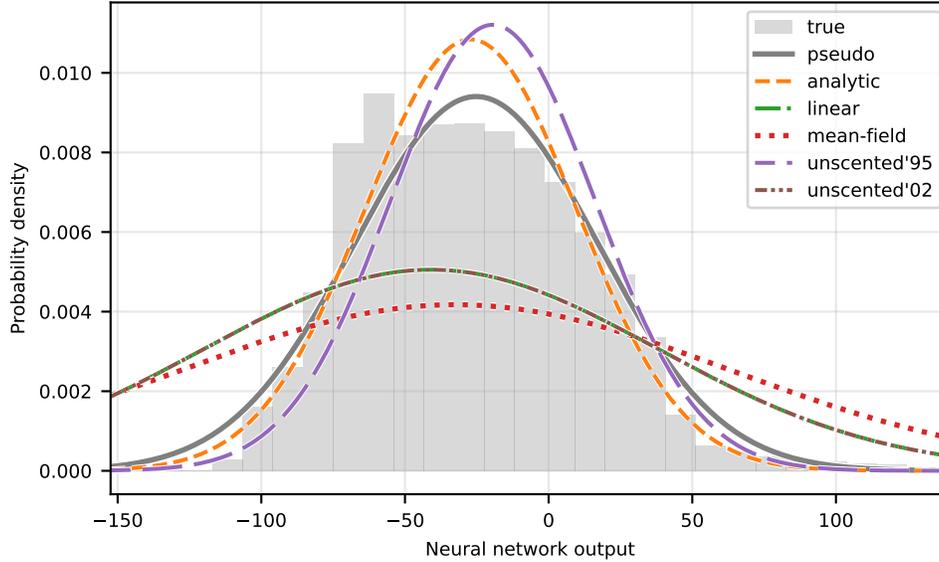
8271 Table 206: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
8272 tion=relu residual), variance=small  
8273

8274  
8275  
8276  
8277  
8278  
8279  
8280  
8281  
8282  
8283

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$7.204 \times 10^{-1} \pm 1.7 \times 10^{-4}$	0
analytic	$7.507 \times 10^{-1} \pm 1.2 \times 10^{-4}$	$1.946 \times 10^{-2} \pm 8.3 \times 10^{-6}$
mean-field	$6.635 \times 10^0 \pm 2.0 \times 10^{-4}$	$1.241 \times 10^0 \pm 1.3 \times 10^{-4}$
linear	$4.870 \times 10^0 \pm 1.6 \times 10^{-4}$	$6.797 \times 10^{-1} \pm 8.4 \times 10^{-5}$
unscented'95	$1.278 \times 10^0 \pm 6.3 \times 10^{-5}$	$3.666 \times 10^{-2} \pm 8.5 \times 10^{-6}$
unscented'02	$4.870 \times 10^0 \pm 1.6 \times 10^{-4}$	$6.797 \times 10^{-1} \pm 8.4 \times 10^{-5}$

8284 Table 207: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
8285 tivation=relu residual), variance=small  
8286

8287  
8288  
8289  
8290  
8291  
8292  
8293  
8294  
8295  
8296  
8297  
8298  
8299  
8300  
8301  
8302  
8303  
8304  
8305



8306  
8307  
8308  
8309  
8310  
8311  
8312  
8313  
8314  
8315

8308 Figure 111: Probability distributions for Network(architecture=deep, weights=trained, activa-  
8309 tion=relu residual), variance=small

8316  
8317  
8318  
8319  
8320  
8321  
8322  
8323  
8324

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$-1.372 \times 10^{+1} \pm 7.8 \times 10^{-3}$	$1.394 \times 10^{+5} \pm 1.6 \times 10^{+1}$
analytic	$-2.415 \times 10^{+1}$	$2.175 \times 10^{+5}$
mean-field	$+2.939 \times 10^{+1}$	$4.827 \times 10^{+5}$
linear	$-4.096 \times 10^{+1}$	$6.231 \times 10^{+5}$
unscented'95	$+6.028 \times 10^{+1}$	$9.951 \times 10^{+4}$
unscented'02	$-4.096 \times 10^{+1}$	$6.231 \times 10^{+5}$

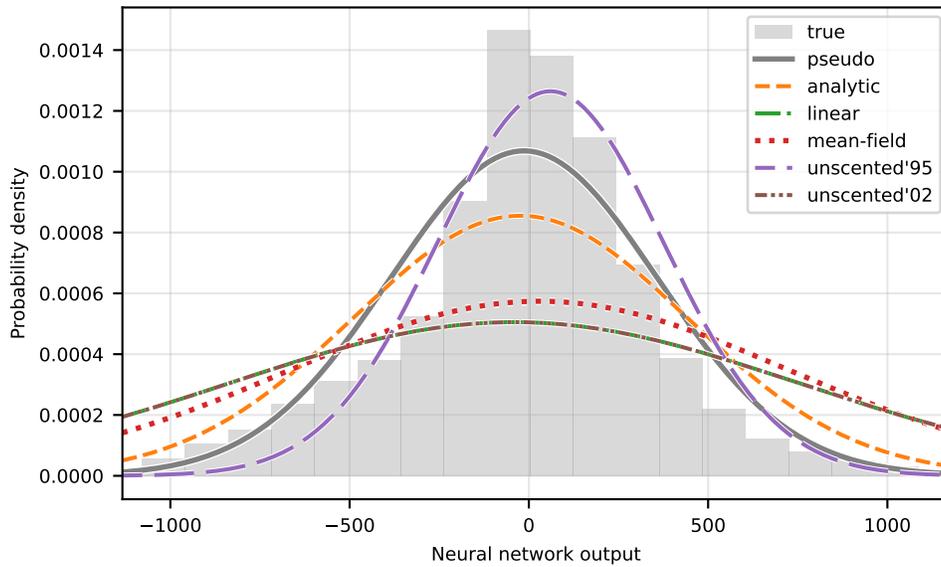
8325 Table 208: Comparison of moments for Network(architecture=deep, weights=trained, activa-  
8326 tion=relu residual), variance=medium  
8327

8328  
8329  
8330  
8331  
8332  
8333  
8334  
8335  
8336  
8337

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$2.432 \times 10^0 \pm 8.2 \times 10^{-4}$	0
analytic	$5.344 \times 10^0 \pm 7.5 \times 10^{-4}$	$5.820 \times 10^{-2} \pm 3.1 \times 10^{-5}$
mean-field	$1.467 \times 10^{+1} \pm 8.9 \times 10^{-4}$	$6.173 \times 10^{-1} \pm 1.4 \times 10^{-4}$
linear	$1.861 \times 10^{+1} \pm 1.0 \times 10^{-3}$	$9.894 \times 10^{-1} \pm 1.9 \times 10^{-4}$
unscented'95	$4.410 \times 10^0 \pm 3.7 \times 10^{-4}$	$4.508 \times 10^{-2} \pm 1.5 \times 10^{-5}$
unscented'02	$1.861 \times 10^{+1} \pm 1.0 \times 10^{-3}$	$9.894 \times 10^{-1} \pm 1.9 \times 10^{-4}$

8338 Table 209: Comparison of statistical distances for Network(architecture=deep, weights=trained, ac-  
8339 tivation=relu residual), variance=medium  
8340

8341  
8342  
8343  
8344  
8345  
8346  
8347  
8348  
8349  
8350  
8351  
8352  
8353  
8354  
8355  
8356  
8357



8358  
8359  
8360  
8361  
8362  
8363  
8364  
8365  
8366  
8367  
8368  
8369

8362 Figure 112: Probability distributions for Network(architecture=deep, weights=trained, activa-  
8363 tion=relu residual), variance=medium

8370  
8371  
8372  
8373  
8374  
8375  
8376  
8377  
8378  
8379  
8380  
8381  
8382  
8383  
8384  
8385  
8386  
8387  
8388  
8389  
8390  
8391  
8392  
8393  
8394  
8395  
8396  
8397  
8398  
8399  
8400  
8401  
8402  
8403  
8404  
8405  
8406  
8407  
8408  
8409  
8410  
8411  
8412  
8413  
8414  
8415  
8416  
8417  
8418  
8419  
8420  
8421  
8422  
8423

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+3.974 \times 10^{+3} \pm 6.7 \times 10^{-2}$	$3.397 \times 10^{+7} \pm 1.7 \times 10^{+3}$
analytic	$+4.166 \times 10^{+3}$	$3.255 \times 10^{+7}$
mean-field	$+2.983 \times 10^{+3}$	$3.785 \times 10^{+7}$
linear	$-4.096 \times 10^{+1}$	$6.231 \times 10^{+7}$
unscented'95	$+5.757 \times 10^{+3}$	$1.384 \times 10^{+7}$
unscented'02	$-3.249 \times 10^{+4}$	$2.168 \times 10^{+9}$

Table 210: Comparison of moments for Network(architecture=deep, weights=trained, activation=relu residual), variance=large

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$9.128 \times 10^0 \pm 1.1 \times 10^{-3}$	0
analytic	$9.622 \times 10^0 \pm 1.2 \times 10^{-3}$	$9.921 \times 10^{-4} \pm 1.0 \times 10^{-6}$
mean-field	$1.297 \times 10^{+1} \pm 8.6 \times 10^{-4}$	$1.747 \times 10^{-2} \pm 3.6 \times 10^{-6}$
linear	$5.259 \times 10^{+1} \pm 9.7 \times 10^{-4}$	$3.510 \times 10^{-1} \pm 3.2 \times 10^{-5}$
unscented'95	$3.335 \times 10^{+1} \pm 9.4 \times 10^{-4}$	$1.995 \times 10^{-1} \pm 1.2 \times 10^{-5}$
unscented'02	$5.818 \times 10^{+2} \pm 7.9 \times 10^{-3}$	$4.890 \times 10^{+1} \pm 2.5 \times 10^{-3}$

Table 211: Comparison of statistical distances for Network(architecture=deep, weights=trained, activation=relu residual), variance=large

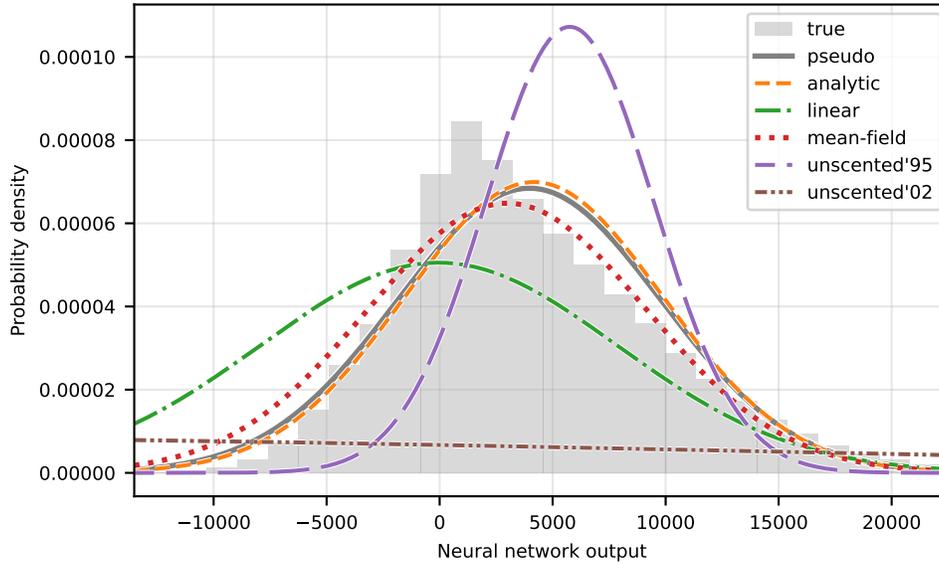


Figure 113: Probability distributions for Network(architecture=deep, weights=trained, activation=relu residual), variance=large

8424  
8425  
8426  
8427  
8428  
8429  
8430  
8431  
8432  
8433  
8434  
8435  
8436  
8437  
8438  
8439  
8440  
8441  
8442  
8443  
8444  
8445  
8446  
8447  
8448  
8449  
8450  
8451  
8452  
8453  
8454  
8455  
8456  
8457  
8458  
8459  
8460  
8461  
8462  
8463  
8464  
8465  
8466  
8467  
8468  
8469  
8470  
8471  
8472  
8473  
8474  
8475  
8476  
8477

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.368 \times 10^{-1} \pm 1.1 \times 10^{-4}$	$9.674 \times 10^{-2} \pm 3.7 \times 10^{-5}$
analytic	$+5.948 \times 10^{-1}$	$9.177 \times 10^{-2}$
mean-field	$+5.946 \times 10^{-1}$	$8.172 \times 10^{-2}$
linear	$+8.223 \times 10^{-2}$	0
unscented'95	$+5.531 \times 10^{-1}$	$9.697 \times 10^{-2}$
unscented'02	$+8.223 \times 10^{-2}$	$1.303 \times 10^{-22}$

Table 212: Comparison of moments for Network(architecture=deep, weights=initialized, activation=heaviside), variance=small

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.623 \times 10^{-2} \pm 4.9 \times 10^{-5}$	0
analytic	$1.172 \times 10^{-1} \pm 1.6 \times 10^{-4}$	$1.811 \times 10^{-2} \pm 6.4 \times 10^{-5}$
mean-field	$1.170 \times 10^{-1} \pm 1.5 \times 10^{-4}$	$2.400 \times 10^{-2} \pm 6.4 \times 10^{-5}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$6.479 \times 10^{-2} \pm 1.1 \times 10^{-4}$	$1.379 \times 10^{-3} \pm 1.8 \times 10^{-5}$
unscented'02	$8.272 \times 10^{-1} \pm 1.8 \times 10^{-4}$	$2.460 \times 10^{+1} \pm 5.1 \times 10^{-4}$

Table 213: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=heaviside), variance=small

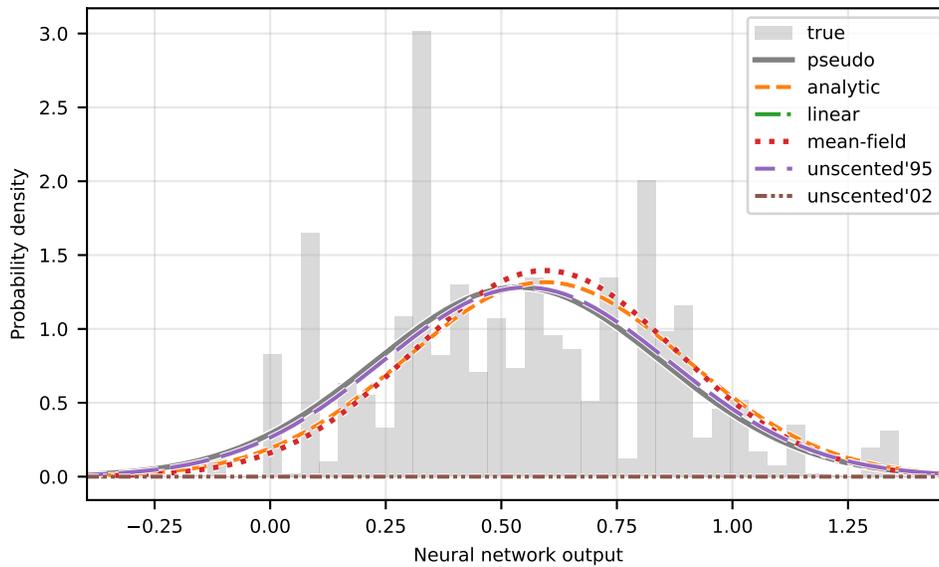


Figure 114: Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside), variance=small

8478  
8479  
8480  
8481  
8482  
8483  
8484  
8485  
8486

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.383 \times 10^{-1} \pm 1.2 \times 10^{-4}$	$9.035 \times 10^{-2} \pm 7.9 \times 10^{-5}$
analytic	$+5.949 \times 10^{-1}$	$9.201 \times 10^{-2}$
mean-field	$+5.957 \times 10^{-1}$	$8.197 \times 10^{-2}$
linear	$+8.223 \times 10^{-2}$	0
unscented'95	$+4.318 \times 10^{-1}$	$6.143 \times 10^{-2}$
unscented'02	$+8.223 \times 10^{-2}$	$1.303 \times 10^{-22}$

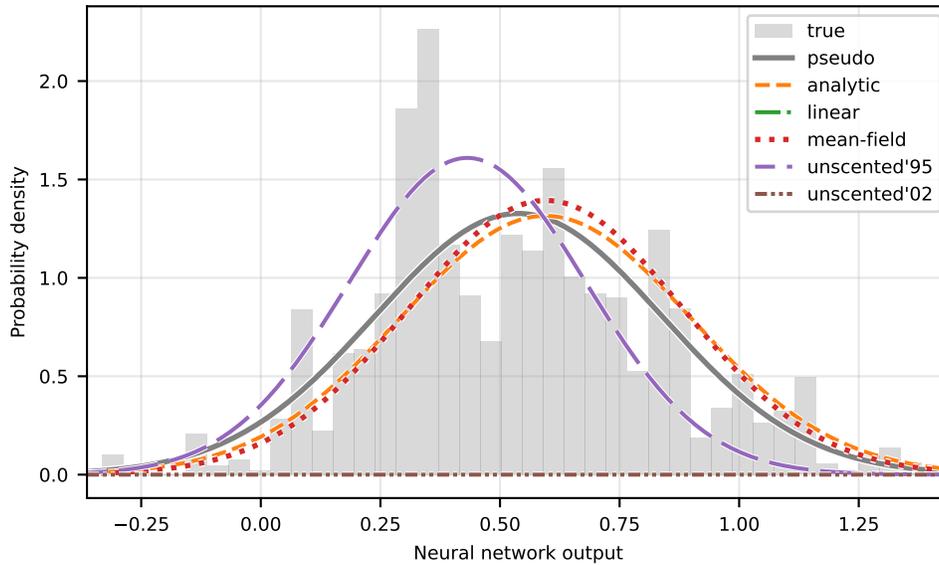
8487 Table 214: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
8488 tion=heaviside), variance=medium  
8489  
8490

8491  
8492  
8493  
8494  
8495  
8496  
8497  
8498  
8499

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.662 \times 10^{-2} \pm 1.3 \times 10^{-4}$	0
analytic	$1.058 \times 10^{-1} \pm 2.1 \times 10^{-4}$	$1.777 \times 10^{-2} \pm 7.7 \times 10^{-5}$
mean-field	$1.085 \times 10^{-1} \pm 2.0 \times 10^{-4}$	$2.049 \times 10^{-2} \pm 8.4 \times 10^{-5}$
linear	— ± —	$\infty \pm \text{—}$
unscented'95	$1.951 \times 10^{-1} \pm 2.2 \times 10^{-4}$	$9.570 \times 10^{-2} \pm 1.6 \times 10^{-4}$
unscented'02	$8.522 \times 10^{-1} \pm 2.5 \times 10^{-4}$	$2.465 \times 10^{+1} \pm 8.9 \times 10^{-4}$

8500 Table 215: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
8501 activation=heaviside), variance=medium  
8502  
8503

8504  
8505  
8506  
8507  
8508  
8509  
8510  
8511  
8512  
8513  
8514  
8515  
8516  
8517  
8518  
8519  
8520  
8521



8522  
8523  
8524 Figure 115: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
8525 tion=heaviside), variance=medium  
8526  
8527  
8528  
8529  
8530  
8531

8532  
8533  
8534  
8535  
8536  
8537  
8538  
8539  
8540

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.414 \times 10^{-1} \pm 2.1 \times 10^{-4}$	$8.995 \times 10^{-2} \pm 8.5 \times 10^{-5}$
analytic	$+5.949 \times 10^{-1}$	$9.189 \times 10^{-2}$
mean-field	$+5.952 \times 10^{-1}$	$8.185 \times 10^{-2}$
linear	$+8.223 \times 10^{-2}$	0
unscented'95	$+6.727 \times 10^{-1}$	$1.985 \times 10^{-1}$
unscented'02	$+8.223 \times 10^{-2}$	$1.303 \times 10^{-22}$

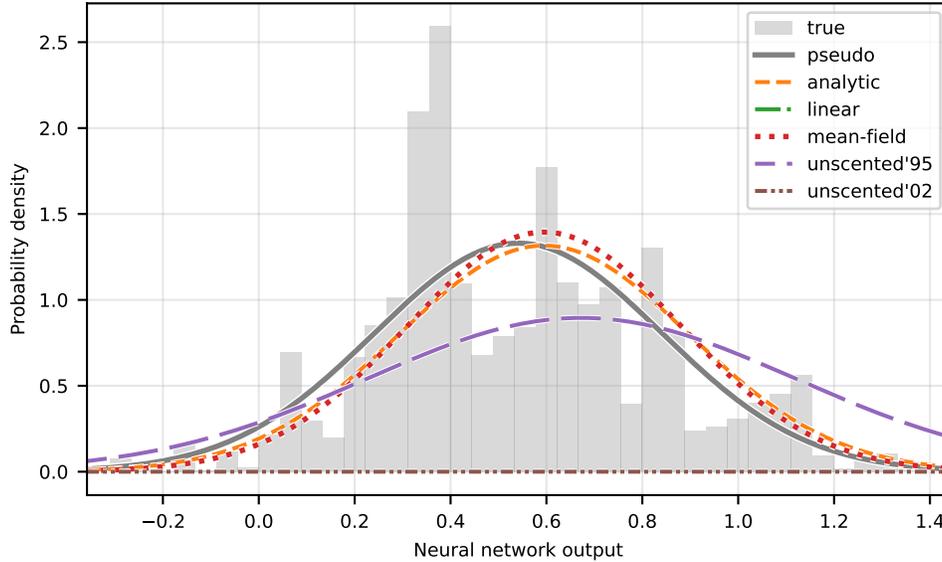
8541 Table 216: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
8542 tion=heaviside), variance=large  
8543  
8544

8545  
8546  
8547  
8548  
8549  
8550  
8551  
8552  
8553

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$5.311 \times 10^{-2} \pm 1.5 \times 10^{-4}$	0
analytic	$1.025 \times 10^{-1} \pm 3.5 \times 10^{-4}$	$1.604 \times 10^{-2} \pm 1.3 \times 10^{-4}$
mean-field	$1.057 \times 10^{-1} \pm 3.4 \times 10^{-4}$	$1.827 \times 10^{-2} \pm 1.3 \times 10^{-4}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$3.236 \times 10^{-1} \pm 3.6 \times 10^{-4}$	$3.036 \times 10^{-1} \pm 7.7 \times 10^{-4}$
unscented'02	$8.558 \times 10^{-1} \pm 3.6 \times 10^{-4}$	$2.466 \times 10^{+1} \pm 1.2 \times 10^{-3}$

8554 Table 217: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
8555 activation=heaviside), variance=large  
8556

8557  
8558  
8559  
8560  
8561  
8562  
8563  
8564  
8565  
8566  
8567  
8568  
8569  
8570  
8571  
8572  
8573  
8574  
8575  
8576



8577 Figure 116: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
8578 tion=heaviside), variance=large  
8579  
8580  
8581  
8582  
8583  
8584  
8585

8586  
8587  
8588  
8589  
8590  
8591  
8592  
8593  
8594

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+8.067 \times 10^{-1} \pm 1.6 \times 10^{-4}$	$4.988 \times 10^{-1} \pm 2.0 \times 10^{-4}$
analytic	$+7.753 \times 10^{-1}$	$4.018 \times 10^{-1}$
mean-field	$+8.268 \times 10^{-1}$	$5.413 \times 10^{-1}$
linear	$+7.963 \times 10^{-1}$	0
unscented'95	$+5.844 \times 10^{-1}$	$5.985 \times 10^{-1}$
unscented'02	$+7.963 \times 10^{-1}$	$6.996 \times 10^{-21}$

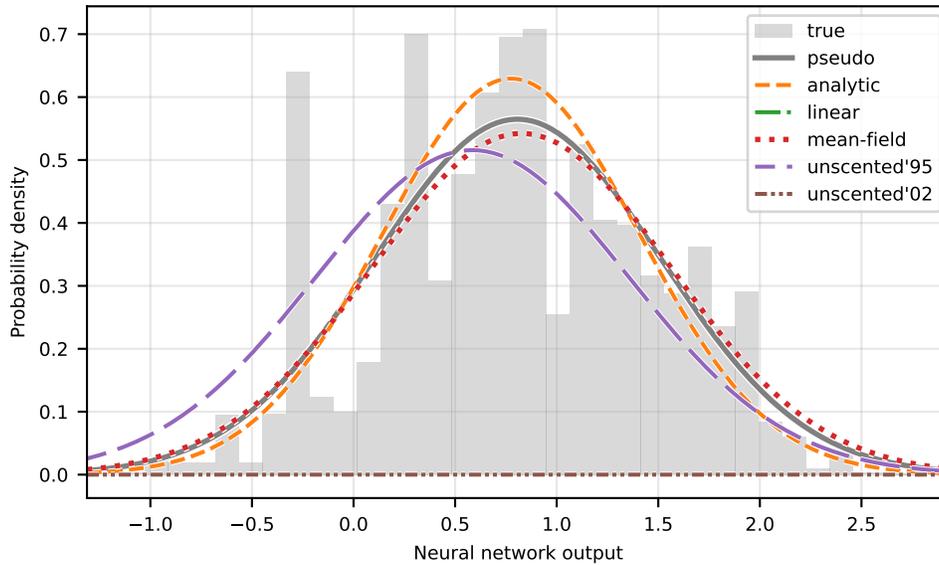
8595 Table 218: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
8596 tion=heaviside residual), variance=small  
8597

8598  
8599  
8600  
8601  
8602  
8603  
8604  
8605  
8606  
8607

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.940 \times 10^{-2} \pm 5.1 \times 10^{-5}$	0
analytic	$8.399 \times 10^{-2} \pm 1.0 \times 10^{-4}$	$1.190 \times 10^{-2} \pm 3.3 \times 10^{-5}$
mean-field	$5.953 \times 10^{-2} \pm 6.8 \times 10^{-5}$	$2.123 \times 10^{-3} \pm 1.4 \times 10^{-5}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$2.677 \times 10^{-1} \pm 2.1 \times 10^{-4}$	$5.836 \times 10^{-2} \pm 1.2 \times 10^{-4}$
unscented'02	$6.676 \times 10^{-1} \pm 9.1 \times 10^{-5}$	$2.236 \times 10^{+1} \pm 2.0 \times 10^{-4}$

8608 Table 219: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
8609 activation=heaviside residual), variance=small  
8610

8611  
8612  
8613  
8614  
8615  
8616  
8617  
8618  
8619  
8620  
8621  
8622  
8623  
8624  
8625  
8626  
8627



8632 Figure 117: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
8633 tion=heaviside residual), variance=small  
8634

8635  
8636  
8637  
8638  
8639

8640  
8641  
8642  
8643  
8644  
8645  
8646  
8647

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+7.707 \times 10^{-1} \pm 3.1 \times 10^{-4}$	$1.283 \times 10^0 \pm 5.4 \times 10^{-4}$
analytic	$+6.602 \times 10^{-1}$	$1.241 \times 10^0$
mean-field	$+5.408 \times 10^{-1}$	$1.006 \times 10^0$
linear	$+7.963 \times 10^{-1}$	0
unscented'95	$+1.122 \times 10^0$	$8.345 \times 10^{-1}$
unscented'02	$+7.963 \times 10^{-1}$	$6.996 \times 10^{-21}$

8648  
8649  
8650  
8651

Table 220: Comparison of moments for Network(architecture=deep, weights=initialized, activation=heaviside residual), variance=medium

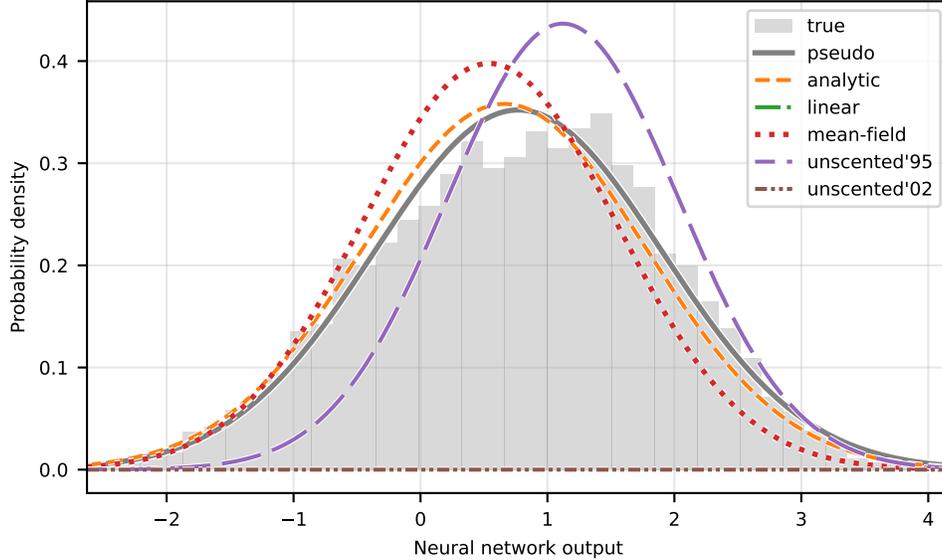
8652

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$4.528 \times 10^{-2} \pm 2.2 \times 10^{-4}$	0
analytic	$1.078 \times 10^{-1} \pm 2.9 \times 10^{-4}$	$5.025 \times 10^{-3} \pm 2.7 \times 10^{-5}$
mean-field	$2.201 \times 10^{-1} \pm 2.3 \times 10^{-4}$	$3.429 \times 10^{-2} \pm 6.8 \times 10^{-5}$
linear	— ± —	$\infty \pm \text{—}$
unscented'95	$3.297 \times 10^{-1} \pm 3.0 \times 10^{-4}$	$8.825 \times 10^{-2} \pm 9.9 \times 10^{-5}$
unscented'02	$8.681 \times 10^{-1} \pm 1.3 \times 10^{-4}$	$2.283 \times 10^{+1} \pm 2.1 \times 10^{-4}$

8662  
8663  
8664

Table 221: Comparison of statistical distances for Network(architecture=deep, weights=initialized, activation=heaviside residual), variance=medium

8665  
8666  
8667  
8668  
8669  
8670  
8671  
8672  
8673  
8674  
8675  
8676  
8677  
8678  
8679  
8680  
8681



8682  
8683  
8684  
8685  
8686  
8687

Figure 118: Probability distributions for Network(architecture=deep, weights=initialized, activation=heaviside residual), variance=medium

8688  
8689  
8690  
8691  
8692  
8693

8694  
8695  
8696  
8697  
8698  
8699  
8700  
8701  
8702

distribution	$\mu$	$\sigma^2$
pseudo-true ( $Y_1$ )	$+5.804 \times 10^{-1} \pm 4.0 \times 10^{-4}$	$1.098 \times 10^0 \pm 7.2 \times 10^{-4}$
analytic	$+5.070 \times 10^{-1}$	$1.071 \times 10^0$
mean-field	$+4.866 \times 10^{-1}$	$1.190 \times 10^0$
linear	$+7.963 \times 10^{-1}$	0
unscented'95	$+7.638 \times 10^{-1}$	$1.589 \times 10^0$
unscented'02	$+7.963 \times 10^{-1}$	$6.996 \times 10^{-21}$

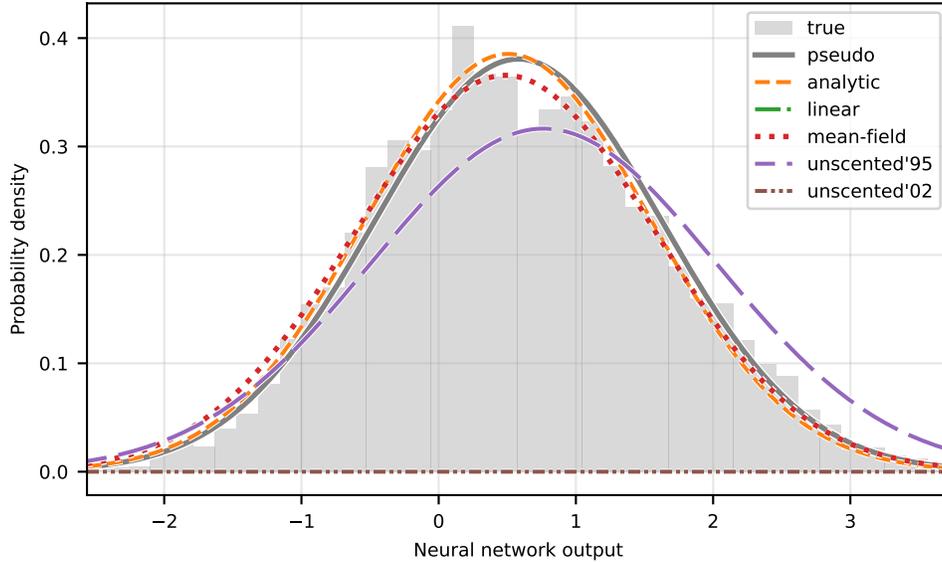
8703 Table 222: Comparison of moments for Network(architecture=deep, weights=initialized, activa-  
8704 tion=heaviside residual), variance=large  
8705  
8706

8707  
8708  
8709  
8710  
8711  
8712  
8713  
8714  
8715

distribution	$d_W(\cdot, Y_0)$	$D_{KL}(Y_1 \parallel \cdot)$
pseudo-true ( $Y_1$ )	$3.778 \times 10^{-2} \pm 2.7 \times 10^{-4}$	0
analytic	$7.241 \times 10^{-2} \pm 3.8 \times 10^{-4}$	$2.611 \times 10^{-3} \pm 2.8 \times 10^{-5}$
mean-field	$9.352 \times 10^{-2} \pm 3.7 \times 10^{-4}$	$5.700 \times 10^{-3} \pm 4.3 \times 10^{-5}$
linear	$— \pm —$	$\infty \pm —$
unscented'95	$2.401 \times 10^{-1} \pm 3.8 \times 10^{-4}$	$5.412 \times 10^{-2} \pm 1.8 \times 10^{-4}$
unscented'02	$8.522 \times 10^{-1} \pm 2.3 \times 10^{-4}$	$2.277 \times 10^{+1} \pm 3.1 \times 10^{-4}$

8716 Table 223: Comparison of statistical distances for Network(architecture=deep, weights=initialized,  
8717 activation=heaviside residual), variance=large  
8718  
8719

8720  
8721  
8722  
8723  
8724  
8725  
8726  
8727  
8728  
8729  
8730  
8731  
8732  
8733  
8734  
8735  
8736  
8737



8738  
8739  
8740  
8741  
8742  
8743  
8744  
8745  
8746  
8747

8740 Figure 119: Probability distributions for Network(architecture=deep, weights=initialized, activa-  
8741 tion=heaviside residual), variance=large

8748  
8749  
8750  
8751  
8752  
8753  
8754  
8755  
8756  
8757  
8758  
8759  
8760  
8761  
8762  
8763  
8764  
8765  
8766  
8767  
8768  
8769  
8770  
8771  
8772  
8773  
8774  
8775  
8776  
8777  
8778  
8779  
8780  
8781  
8782  
8783  
8784  
8785  
8786  
8787  
8788  
8789  
8790  
8791  
8792  
8793  
8794  
8795  
8796  
8797  
8798  
8799  
8800  
8801

---

## O COMMENTS ON COMPUTATIONAL COMPLEXITY

The cost of moment matching, like that of linearized covariance propagation, is cubic in the number of hidden neurons due to the need to evaluate “sandwich” covariances such as  $A\Sigma A^\top$ .<sup>6</sup> By contrast, the computational complexity of evaluating a neural network is quadratic in the number of hidden neurons. Thus, theoretically, moment matching is one polynomial order of magnitude slower than point prediction, on the same order as linearization. For neural networks with high input dimension, moment matching is much faster than Monte Carlo, which scales exponentially with dimension.

We performed a simple experiment using Jupyter Notebook’s `%%timeit` to compare the computational cost of moment matching and Monte Carlo. We compare our method with the ground truth method used in §1, which runs quasi-Monte Carlo (QMC) on  $2^{16}$  samples. We use GeLU because it has the most complicated covariance expression (see Appendix E).

On “wide residual” networks (defined in Appendix N):

- Our method takes  $88.4 \text{ ms} \pm 5.96 \text{ ms}$ .
- QMC takes  $1.79 \text{ s} \pm 83 \text{ ms}$ .

On “deep residual” networks (defined in Appendix N):

- Our method takes  $23.7 \text{ ms} \pm 957 \mu\text{s}$ .
- QMC takes  $878 \text{ ms} \pm 61.8 \text{ ms}$ .

---

<sup>6</sup>Contrary to occasional claims that the cost of propagating a covariance matrix scales *quadratically* with the number of hidden neurons (Akgül et al., 2025, §5).

---

8802 P LLM USAGE STATEMENT  
8803

8804 We consider that LLMs were not involved “at the level of a contributing author.” We nevertheless  
8805 include a usage statement for full transparency. LLM usage consisted of early ideation involved  
8806 in checking whether  $\Phi$  had exact integrals, as well as autocomplete (executable code and LaTeX)  
8807 and limited code generation (executable code). We take full responsibility for the content of this  
8808 manuscript.  
8809

8810  
8811  
8812  
8813  
8814  
8815  
8816  
8817  
8818  
8819  
8820  
8821  
8822  
8823  
8824  
8825  
8826  
8827  
8828  
8829  
8830  
8831  
8832  
8833  
8834  
8835  
8836  
8837  
8838  
8839  
8840  
8841  
8842  
8843  
8844  
8845  
8846  
8847  
8848  
8849  
8850  
8851  
8852  
8853  
8854  
8855