

404	A Gaussian Integral Derivations	15
405	A.1 Gaussian Exp Integral	15
406	A.2 Gaussian NormCDF Integral	15
407	A.3 Probit Approximation for Gaussian Sigmoid Integral	15
408	A.4 Mean Field Approximation for Gaussian Softmax Integral	16
409	B Comparison with the Multinomial Probit Model	16
410	C Theoretical Analyses	17
411	C.1 Informal Analysis of Monte Carlo Approximations	17
412	C.2 Analysis of the Closed-Form Approximations (Theorem 2.1)	18
413	D Moment Matching Beta Distributions	21
414	D.1 Information Geometric Interpretation of the Pushforward through NormCDF	22
415	E List of Predictive and Dirichlet Parameters Formulas	23
416	E.1 Predictive Formulas	23
417	E.2 Dirichlet Parameters Formulas	24
418	F List of Uncertainty Estimators	25
419	F.1 Predictive	25
420	F.2 Monte Carlo	25
421	F.3 Dirichlet	25
422	G Experimental Setup	26
423	H Benchmark Metrics	27
424	H.1 Log Probability Proper Scoring Rule for the Predictive	27
425	H.2 Expected Calibration Error	27
426	H.3 Binary Log Probability Proper Scoring Rule for Correctness Prediction	28
427	H.4 Accuracy	28
428	H.5 Area Under the Receiver Operating Characteristic Curve for Out-of-Distribution	
429	Detection	28
430	I Benchmarked Methods	29
431	I.1 Spectral Normalized Gaussian Process	29
432	I.2 Heteroscedastic Classifier	29
433	I.3 Laplace Approximation	29
434	J CIFAR-10 Experiments	30
435	J.1 Quality of Sample-Free Predictives	30
436	J.2 Effects of Changing the Learning Objective	31
437	K Additional Results	32

438	K.1 Closed-form Softmax Predictives	32
439	K.2 Vision Transformer Results on ImageNet	33
440	K.3 CIFAR-100 Results	33
441	K.4 Alignment with the True Predictives	33
442	K.5 BCE Loss Performance Gains	34
443	K.6 Further Out-of-Distribution Detection Results	34

444 A Gaussian Integral Derivations

445 In this appendix, we derive the closed-form formula for the mean of Gaussian pushforwards through
 446 exp (Eq. (8)) and normCDF (Eq. (13)), as well as the approximations for pushforwards through
 447 sigmoid (Eq. (15)) and softmax.

448 A.1 Gaussian Exp Integral

449 By absorbing the exponential into the Gaussian probability density function, we get

$$\begin{aligned}
 \int_{\mathbb{R}} \exp(y) \mathcal{N}(\mu, \sigma^2)(dy) &= \int_{\mathbb{R}} \exp(y) \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy \\
 &= \int_{\mathbb{R}} \exp\left(\mu + \frac{\sigma^2}{2}\right) \exp\left(-\frac{1}{2\sigma^2}(y-\mu-\sigma^2)^2\right) dy \\
 &= \int_{\mathbb{R}} \exp\left(\mu + \frac{\sigma^2}{2}\right) \mathcal{N}(\mu + \sigma^2, \sigma^2)(dy) \\
 &= \exp\left(\mu + \frac{\sigma^2}{2}\right).
 \end{aligned} \tag{30}$$

450 A.2 Gaussian NormCDF Integral

451 Here, we derive the classical normCDF Gaussian integration formula [30, Eq. 10.010.8].

452 For $\lambda > 0$, $Z \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(\mu, \sigma^2)$ with Y and Z independent,

$$\begin{aligned}
 \int_{\mathbb{R}} \Phi(\lambda y) \mathcal{N}(\mu, \sigma^2)(dy) &= \int_{\mathbb{R}} p(Z \leq y/\lambda) \mathcal{N}(\mu, \sigma^2)(dy) \\
 &= p\left(Z \leq \frac{Y}{\lambda}\right) \\
 &= p\left(\frac{Z/\lambda - Y + \mu}{\sqrt{\lambda^{-2} + \sigma^2}} \leq \frac{\mu}{\sqrt{\lambda^{-2} + \sigma^2}}\right) \\
 &= \Phi\left(\frac{\mu}{\sqrt{\lambda^{-2} + \sigma^2}}\right)
 \end{aligned} \tag{31}$$

453 where we used $\frac{Z/\lambda - Y + \mu}{\sqrt{\lambda^{-2} + \sigma^2}} \sim \mathcal{N}(0, 1)$ for the last equality. Taking $\lambda = 1$, this gives the formula for
 454 the exact predictive with normCDF.

455 A.3 Probit Approximation for Gaussian Sigmoid Integral

456 Taylor expanding ρ to first order about 0,

$$\begin{aligned}
 \rho(y) &= \frac{1}{2} + \frac{1}{4}y + o(y), \\
 \Phi(y) &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}y + o(y).
 \end{aligned} \tag{32}$$

457 as $y \rightarrow 0$. Hence matching ρ and Φ to first order we get the approximation $\rho(y) \approx \Phi\left(\sqrt{\frac{\pi}{8}}y\right)$. So
 458 using Eq. (31) we derive the *probit approximation* [34, 23]

$$\int_{\mathbb{R}} \rho(y) \mathcal{N}(\mu, \sigma^2)(dy) \stackrel{(1)}{\approx} \int_{\mathbb{R}} \Phi\left(\sqrt{\frac{\pi}{8}}y\right) \mathcal{N}(\mu, \sigma^2)(dy) = \Phi\left(\frac{\mu}{\sqrt{\frac{8}{\pi} + \sigma^2}}\right) \stackrel{(2)}{\approx} \rho\left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}\right). \tag{33}$$

459 Note that the approximation (2) is not strictly needed, as Φ is computationally tractable. However,
 460 adding (2) empirically improves the quality of the overall approximation. This may be due to the
 461 fact the thicker tails of ρ in the integrand of the left-hand side are better captured by ρ than Φ on the
 462 right-hand side.

463 A.4 Mean Field Approximation for Gaussian Softmax Integral

464 For $\boldsymbol{\mu} \in \mathbb{R}^C$, $\boldsymbol{\sigma}^2 \in \mathbb{R}_{>0}^C$ and $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma}^2)$, the *mean field approximation* to the Gaussian softmax
 465 integral [21] is obtained as follows

$$\begin{aligned}
 \mathbb{E}_{\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\text{softmax}_c \mathbf{Y}] &= \mathbb{E} \left[\left(2 - C + \sum_{c' \neq c} \rho(Y_c - Y_{c'})^{-1} \right)^{-1} \right] \\
 &\stackrel{(1)}{\approx} \left(2 - C + \sum_{c' \neq c} \mathbb{E}[\rho(Y_c - Y_{c'})]^{-1} \right)^{-1} \\
 &\stackrel{(2)}{\approx} \left(2 - C + \sum_{c' \neq c} \mathbb{E}[\rho(Y_c - \mu_{c'})]^{-1} \right)^{-1} \\
 &\stackrel{(3)}{\approx} \left(2 - C + \sum_{c' \neq c} \rho \left(\frac{\mu_c - \mu_{c'}}{\sqrt{1 + \frac{\pi}{8} \sigma_c^2}} \right)^{-1} \right)^{-1} \\
 &= \text{softmax}_c \left(\frac{\boldsymbol{\mu}}{\sqrt{1 + \frac{\pi}{8} \boldsymbol{\sigma}^2}} \right)
 \end{aligned} \tag{34}$$

466 i.e.,

$$\mathbb{E}_{\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\text{softmax} \mathbf{Y}] \approx \text{softmax} \left(\frac{\boldsymbol{\mu}}{\sqrt{1 + \frac{\pi}{8} \boldsymbol{\sigma}^2}} \right). \tag{35}$$

467 (1) is the mean field approximation, and (3) uses the probit approximation Eq. (33). [21] provides
 468 two other variants of this approximation with other choices of approximation (2).

469 B Comparison with the Multinomial Probit Model

470 In this appendix, we show that a model whose output activation is a composition of an element-wise
 471 normCDF activation Φ and a normalisation \mathbf{n} is distinct from the classical multinomial probit model
 472 [6].

473 Given a logit $\mathbf{y} \in \mathbb{R}^C$, the multinomial probit model sets

$$Z(\mathbf{y}) = \arg \max_{1 \leq c \leq C} Y_c \tag{36}$$

474 where $Y_c = y_c + \epsilon_c$ and the ϵ_c are i.i.d. standard Normal. So

$$p(c \mid \mathbf{y}) = p(y_c + \epsilon_c > y_{c'} + \epsilon_{c'} \ \forall \ c' \neq c) \tag{37}$$

475 which is generally not analytically tractable. On the other hand, a model that uses normCDF and
 476 normalisation as output activation yields

$$p(c \mid \mathbf{y}) = \frac{p(y_c + \epsilon_c > 0)}{\sum_{c'=1}^C p(y_{c'} + \epsilon_{c'} > 0)} = \frac{\Phi(y_c)}{\sum_{c'=1}^C \Phi(y_{c'})}. \tag{38}$$

477 In the case $C = 2$, the multinomial probit model Eq. (37) outputs closed-form probabilities. This
 478 allows us to construct an explicit counterexample to the equivalence of the two models Eq. (37) and
 479 Eq. (38):

$$p(y_1 + \epsilon_1 > y_2 + \epsilon_2) = p\left(\frac{y_1 - y_2}{2} + \frac{\epsilon_1 - \epsilon_2}{2} > 0\right) = \Phi\left(\frac{y_1 - y_2}{2}\right) \neq \frac{\Phi(y_1)}{\Phi(y_1) + \Phi(y_2)}. \tag{39}$$

C Theoretical Analyses

In this appendix, we provide (formal and informal) theoretical analyses of the quality of various predictive approximations. This complements the empirical analyses, for instance in the synthetic experiment (Fig. 1) or in [7].

Due to its information-theoretic interpretation, a natural divergence to equip the probability simplex Δ^{C-1} with is the Kullback-Leibler (KL) divergence

$$D_{\text{KL}}(\mathbf{p}, \mathbf{q}) = \sum_{c=1}^C p_c (\log p_c - \log q_c) \quad (40)$$

which is well defined if \mathbf{p}, \mathbf{q} lie in the interior of the simplex ($p_i, q_i \neq 0, 1$ for all i). So, for a predictive approximation, $\hat{\mathbf{p}}$, we would like to analyse $D_{\text{KL}}(\mathbf{p}, \hat{\mathbf{p}})$, where $\mathbf{p} := \mathbb{E}_{\mathbf{P} \sim \mathcal{a}_* \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\mathbf{P}]$ is the true predictive, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are some logit space mean and covariance and \mathbf{a} is an output activation (e.g. $\mathbf{a} = \mathbf{n} \circ \boldsymbol{\varphi}$).

C.1 Informal Analysis of Monte Carlo Approximations

An N sample Monte Carlo estimate is defined as

$$\hat{\mathbf{P}}^S := \frac{1}{S} \sum_{s=1}^S \hat{\mathbf{P}}^{(s)} \quad (41)$$

where the $\hat{\mathbf{P}}^{(s)}$ are i.i.d. $\mathcal{a}_* \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The computational cost of MC integration is $\mathcal{O}(S \cdot C)$. This becomes prohibitive for large S and C . Thus, for a fair assessment of the quality of such an estimate in terms of the number of classes, one should consider MC estimates $\hat{\mathbf{P}}^{\lceil S/C \rceil}$.

We now give an informal theoretical argument for the linear growth of the KL divergence between \mathbf{p} and $\hat{\mathbf{P}}^{\lceil S/C \rceil}$ in terms of C , under the distributional conditions of the synthetic experiment (Fig. 1).

Taylor expanding Eq. (40) about \mathbf{p} to second order we obtain

$$\begin{aligned} D_{\text{KL}}(\mathbf{p}, \mathbf{q}) &= \sum_{c=1}^C p_c (\log p_c - \log q_c) \\ &\stackrel{(1)}{\approx} \sum_{c=1}^C p_c \left(1 - \frac{p_c}{q_c} + \frac{(p_c - q_c)^2}{2p_c^2} \right) \\ &= \underbrace{\sum_{c=1}^C p_c}_{=1} - \underbrace{\sum_{c=1}^C q_c}_{=1} + \sum_{c=1}^C \frac{(p_c - q_c)^2}{2p_c} \\ &= \sum_{c=1}^C \frac{(p_c - q_c)^2}{2p_c} \\ &\stackrel{(2)}{\approx} \frac{C}{2} \|\mathbf{p} - \mathbf{q}\|_2^2 \end{aligned} \quad (42)$$

where approximation (1) assumes $\|\mathbf{p} - \mathbf{q}\|_2$ is small, and (2) assumes $p_c \approx 1/C$.

In the synthetic experiment (Fig. 1), the logit class-wise means μ_c and variances σ_c^2 are sampled in an i.i.d. way. Let $\mathbf{Q} \sim \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2))$ the unnormalised ‘probabilities’ and $\mathbf{P} := \mathbf{Q} / \sum_{c=1}^C Q_c$ the probabilities. We have

$$\text{Var}[\mathbf{P}] = \mathbb{E} \left[\frac{\mathbf{Q}^2}{\left(\sum_{c=1}^C Q_c \right)^2} \right] - \mathbb{E} \left[\frac{\mathbf{Q}}{\sum_{c=1}^C Q_c} \right]^2 \approx \frac{\mathbb{E}[\mathbf{Q}^2] - \mathbb{E}[\mathbf{Q}]^2}{\left(\sum_{c=1}^C \mathbb{E}[Q_c] \right)^2} = \frac{\text{Var}[\mathbf{Q}]}{C^2 \mathbb{E}[Q_1]^2} \quad (43)$$

where all operations are taken element-wise. Thus

$$\mathbb{E} [\|\mathbf{p} - \mathbf{P}\|_2^2] = \sum_{c=1}^C \text{Var}[P_c] \approx \sum_{c=1}^C \frac{\text{Var}[Q_c]}{C^2 \mathbb{E}[Q_1]^2} = \frac{\text{Var}[Q_1]}{C \mathbb{E}[Q_1]^2}. \quad (44)$$

Now the MC samples $\hat{\mathbf{P}}^{(s)}$ are i.i.d. copies of \mathbf{P} . So we have

$$\mathbb{E}[\|\mathbf{p} - \hat{\mathbf{P}}^{\lceil S/C \rceil}\|_2^2] = \frac{1}{\lceil S/C \rceil} \mathbb{E}[\|\mathbf{p} - \mathbf{P}\|_2^2] \approx \frac{\text{Var}[Q_1]}{S\mathbb{E}[Q_1]^2}. \quad (45)$$

Plugging this into Eq. (42) we get

$$\mathbb{E}[\text{D}_{\text{KL}}(\mathbf{p}, \hat{\mathbf{P}}^{\lceil S/C \rceil})] \approx \frac{\text{Var}[Q_1]}{2\mathbb{E}[Q_1]^2} \cdot \frac{C}{S} \quad (46)$$

which grows linearly with the number of classes, as observed in Fig. 1.

C.2 Analysis of the Closed-Form Approximations (Theorem 2.1)

In this section we prove Theorem 2.1 and discuss whether its underlying assumptions are fulfilled in practice. We start with a lemma.

Lemma C.1. *Let $(X_n)_{n \geq 1}, (Y_n)_{n \geq 1}$ be sequences of $(0, M)$ -valued random variables for some $M > 0$. Suppose $\sup_n \mathbb{E}[Y_n^{-8}] < \infty$ and that $\text{Var}(Y_n) \rightarrow 0$ as $n \rightarrow \infty$. Then*

$$\left| \mathbb{E}\left[\frac{X_n}{Y_n}\right] - \frac{\mathbb{E}[X_n]}{\mathbb{E}[Y_n]} \right| = O(\text{Var}(Y_n)^{1/2}) \quad (47)$$

as $n \rightarrow \infty$.

Proof. Let X, Y be $(0, M)$ -valued random variables. We take a bivariate Taylor expansion of $\frac{X}{Y}$ for the random variables X and Y around $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ respectively:

$$\begin{aligned} \frac{X}{Y} &= \frac{\mathbb{E}[X]}{\mathbb{E}[Y]} + \frac{1}{\mathbb{E}[Y]}(X - \mathbb{E}[X]) - \frac{\mathbb{E}[X]}{\mathbb{E}[Y]^2}(Y - \mathbb{E}[Y]) \\ &\quad - \frac{1}{\xi(Y)^2}(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) + \frac{\eta(X)}{\xi(Y)^3}(Y - \mathbb{E}[Y])^2 \end{aligned} \quad (48)$$

where $\eta(X) \in [\mathbb{E}[X], X]$ or $[X, \mathbb{E}[X]]$ and $\xi(Y) \in [\mathbb{E}[Y], Y]$ or $[Y, \mathbb{E}[Y]]$, using the Lagrange form of the remainder. So by the Cauchy-Schwarz inequality,

$$\begin{aligned} \left| \mathbb{E}\left[\frac{X}{Y}\right] - \frac{\mathbb{E}[X]}{\mathbb{E}[Y]} \right| &= \left| -\mathbb{E}\left[\frac{1}{\xi(Y)^2}(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] + \mathbb{E}\left[\frac{\eta(X)}{\xi(Y)^3}(Y - \mathbb{E}[Y])^2\right] \right| \\ &\leq \mathbb{E}\left[\frac{1}{\xi(Y)^8}\right]^{1/4} \mathbb{E}[(X - \mathbb{E}[X])^4]^{1/4} \mathbb{E}[(Y - \mathbb{E}[Y])^2]^{1/2} \\ &\quad + \mathbb{E}\left[\frac{\eta(X)^2}{\xi(Y)^6}\right]^{1/2} \mathbb{E}[(Y - \mathbb{E}[Y])^4]^{1/2}. \end{aligned} \quad (49)$$

Note that

$$\begin{aligned} \eta(X_n) &\leq M, \\ \sup_n \mathbb{E}\left[\frac{1}{\xi(Y_n)^k}\right] &\leq \sup_n \mathbb{E}\left[\frac{1}{\min(Y_n^k, \mathbb{E}[Y_n]^k)}\right] \leq \sup_n \mathbb{E}\left[\frac{1}{Y_n^k} + \frac{1}{\mathbb{E}[Y_n]^k}\right] < \infty \text{ for } k \in \{6, 8\}, \\ \mathbb{E}[(X_n - \mathbb{E}[X_n])^4] &\leq M^4, \\ \mathbb{E}[(Y_n - \mathbb{E}[Y_n])^4] &\leq M^2 \mathbb{E}[(Y_n - \mathbb{E}[Y_n])^2]. \end{aligned} \quad (50)$$

So from Eq. (49) we obtain

$$\left| \mathbb{E}\left[\frac{X_n}{Y_n}\right] - \frac{\mathbb{E}[X_n]}{\mathbb{E}[Y_n]} \right| = O(\mathbb{E}[(Y_n - \mathbb{E}[Y_n])^2]^{1/2}) = O(\text{Var}(Y_n)^{1/2}) \quad (51)$$

as $n \rightarrow \infty$. \square

Now recall our notation

$$q = q(\mu, \sigma^2) := \mathbb{E}_{Q \sim \rho_* \mathcal{N}(\mu, \sigma^2)}[Q] \approx \rho\left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}\right) =: \hat{q}(\mu, \sigma^2) = \hat{q} \quad (52)$$

where \hat{q} is the probit approximation (Appendix A.3). Moreover we write $q_c := q(\mu_c, \sigma_c^2)$, $\hat{q}_c := \hat{q}(\mu_c, \sigma_c^2)$, \mathbf{p} for the true predictive, $\hat{\mathbf{p}}$ for our approximate predictive, and in addition $\tilde{\mathbf{p}}$ for the approximate predictive using the exact one dimensional integrals, i.e.

$$\mathbf{p} := \mathbb{E}_{\mathbf{P} \sim \mathcal{P}_{*}(\mu, \Sigma)}[\mathbf{P}], \quad \hat{\mathbf{p}} := \frac{\hat{q}_c}{\sum_{c'=1}^C \hat{q}_{c'}}, \quad \tilde{\mathbf{p}} := \frac{q_c}{\sum_{c'=1}^C q_{c'}}. \quad (53)$$

We now restate Theorem 2.1 with an explicit expression for $M(\mathcal{K})$:

Theorem C.2. Suppose the means and variances (μ_c, σ_c^2) lie in some compact set $\mathcal{K} \subset \mathbb{R} \times [0, \infty)$ for each class c . Using the compactness of \mathcal{K} , define

- $\delta(\mathcal{K}) := \sup_{(\mu, \sigma^2) \in \mathcal{K}} (\hat{q} - q)$,
- $u(\mathcal{K}) := \inf_{(\mu, \sigma^2) \in \mathcal{K}} q > 0$,
- $\Delta(\mathcal{K}) := \sup_{(\mu, \sigma^2) \in \mathcal{K}} \frac{q - \hat{q}}{q}$.

If $\Delta(\mathcal{K}) > 1$ then

$$\text{D}_{\text{KL}}(\mathbf{p}, \hat{\mathbf{p}}) \leq \log \left(\frac{1 + \delta/u}{1 - \Delta} \right) + O \left(\text{Var} \left(\sum_{c=1}^C Q_c \right) \right) \quad (54)$$

as $\text{Var} \left(\sum_{c=1}^C Q_c \right) \rightarrow 0$

Proof. We have

$$\text{D}_{\text{KL}}(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{c=1}^C p_c (\log p_c - \log \hat{p}_c) = \underbrace{\sum_{c=1}^C p_c (\log p_c - \log \tilde{p}_c)}_{(1)} + \underbrace{\sum_{c=1}^C p_c (\log \tilde{p}_c - \log \hat{p}_c)}_{(2)}. \quad (55)$$

Assuming $\Delta < 1$, we can bound (2):

$$\begin{aligned} (2) &= \sum_{c=1}^C \frac{q_c}{\sum_{c'=1}^C q_{c'}} \left(\log \left(\frac{q_c}{\sum_{c'=1}^C q_{c'}} \right) - \log \left(\frac{\hat{q}_c}{\sum_{c'=1}^C \hat{q}_{c'}} \right) \right) \\ &= \frac{1}{\sum_{c=1}^C q_c} \sum_{c=1}^C q_c \left(-\log \left(\frac{\hat{q}_c}{q_c} \right) + \log \left(\frac{\sum_{c'=1}^C \hat{q}_{c'}}{\sum_{c'=1}^C q_{c'}} \right) \right) \\ &= \frac{1}{\sum_{c=1}^C q_c} \sum_{c=1}^C q_c \left(-\log \left(1 - \frac{q_c - \hat{q}_c}{q_c} \right) + \log \left(1 + \frac{\sum_{c'=1}^C \hat{q}_{c'} - \sum_{c'=1}^C q_{c'}}{\sum_{c'=1}^C q_{c'}} \right) \right) \\ &\leq \frac{1}{\sum_{c=1}^C q_c} \sum_{c=1}^C q_c \left(-\log(1 - \Delta) + \log \left(1 + \frac{C\delta}{Cu} \right) \right) \\ &= \frac{1}{\sum_{c=1}^C q_c} \sum_{c=1}^C q_c \log \left(\frac{1 + \delta/u}{1 - \Delta} \right) \\ &= \log \left(\frac{1 + \delta/u}{1 - \Delta} \right). \end{aligned} \quad (56)$$

Now for (1), first note that $\mathbb{E} \left[\left(\sum_{c=1}^C Q_c \right)^{-8} \right] < \infty$ where² $Q_c \sim \rho_* \mathcal{N}(\mu_c, \sigma_c^2)$. By compactness of \mathcal{K} , we have in fact $\sup_{(\mu, \sigma) \in \mathcal{K}^C} \mathbb{E} \left[\left(\sum_{c=1}^C Q_c \right)^{-8} \right] < \infty$. Thus we can apply Theorem C.1 with

²In the case $\varphi = \Phi$ and $Q_c \sim \Phi_* \mathcal{N}(\mu_c, \sigma_c^2)$, due to the fast decay of the tail of Φ we may have $\mathbb{E} \left[\left(\sum_{c=1}^C Q_c \right)^{-8} \right] = \infty$, making the proof strategy fail in that case.

535 $(X_n)_{n \geq 1}$ a sequence of $Q_c \sim \rho_* \mathcal{N}(\mu_c, \sigma_c^2)$ and $(Y_n)_{n \geq 1}$ a sequence of $\sum_{c=1}^C Q_c$, to get

$$|p_c - \tilde{p}_c| = O \left(\text{Var} \left(\sum_{c=1}^C Q_c \right)^{1/2} \right) \quad (57)$$

536 as $\text{Var} \left(\sum_{c=1}^C Q_c \right) \rightarrow 0$. Thus Taylor expanding each term of (1) around p_c we get

$$\begin{aligned} (1) &= \sum_{c=1}^C p_c \left(\frac{p_c - \tilde{p}_c}{p_c} + \frac{(p_c - \tilde{p}_c)^2}{2\omega(\tilde{p}_c)^2} \right) \\ &= \underbrace{\sum_{c=1}^C p_c}_{=1} - \underbrace{\sum_{c=1}^C \tilde{p}_c}_{=1} + \frac{(p_c - \tilde{p}_c)^2}{2\omega(\tilde{p}_c)^2} \\ &= \frac{O \left(\text{Var} \left(\sum_{c=1}^C Q_c \right)^{1/2} \right)^2}{2\omega(\tilde{p}_c)^2} \\ &= O \left(\text{Var} \left(\sum_{c=1}^C Q_c \right) \right) \end{aligned} \quad (58)$$

537 as $\text{Var} \left(\sum_{c=1}^C Q_c \right) \rightarrow 0$, where $\omega_c(\tilde{p}_c) \in [p_c, \tilde{p}_c]$ or $[\tilde{p}_c, p_c]$.

538

□

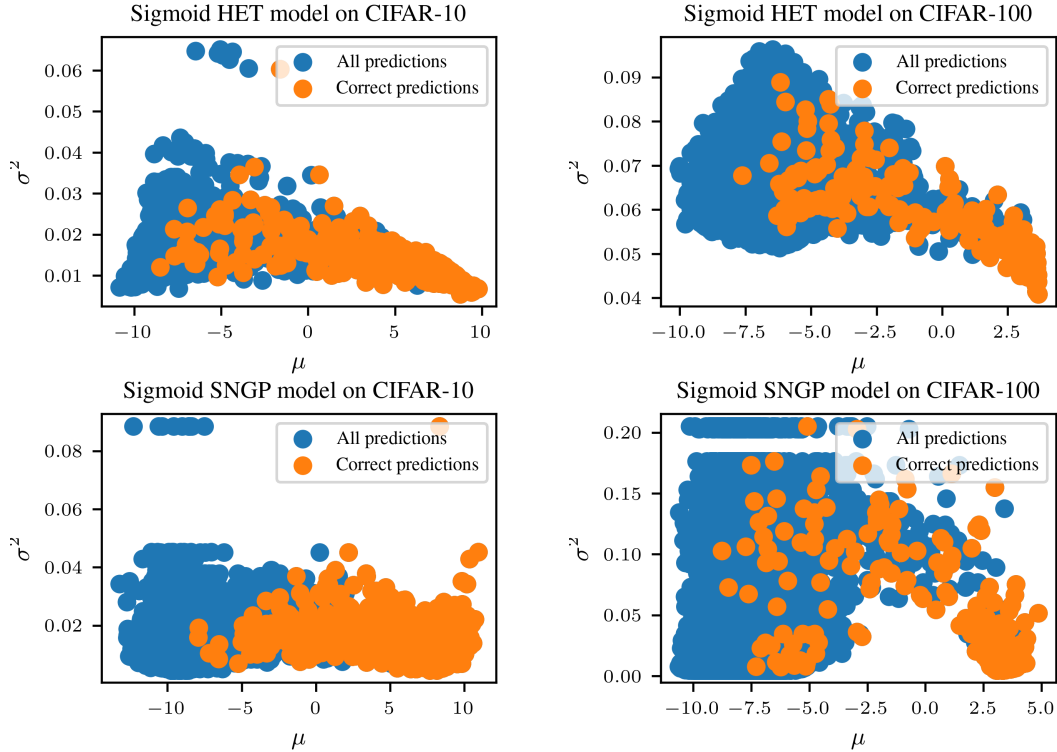


Figure C.1: Scatter plots of logit mean and variance pairs (μ_c, σ_c^2) on real models. We see that these seem to remain constrained to a compact set, with no much difference in the support as we increase the number of classes by an order of magnitude (from CIFAR-10 to CIFAR-100).

539 A key assumption in Theorem C.2 is the logit means and variances being restricted to a compact
 540 set. In Fig. C.1 we see empirically that this is approximately the case for HET and SNGP models,
 541 independently of the number of classes.

542 Theorem C.2 is of practical value as $M(\mathcal{K}) := \log \left(\frac{1+\delta/u}{1-\Delta} \right)$ is well-behaved:

- 543 1. $M(\mathcal{K})$ is independent of C ,
- 544 2. $M(\mathcal{K}) \rightarrow 0$ as $\delta \rightarrow 0$.

545 In other words, given knowledge of the worst case error in the approximation Eq. (15) on the compact
 546 set \mathcal{K} , we can bound the KL divergence in terms of that error independently of the number of classes.
 547 Due to the simplicity of our assumptions, the bound remains quite raw and could be strengthened
 548 with further distributional assumptions on the means and variances.

549 Finally, to obtain a meaningful bound in Theorem C.2, we needed to assume $\Delta(\mathcal{K}) :=$
 550 $\sup_{(\mu, \sigma^2) \in \mathcal{K}} \frac{q - \hat{q}}{q} < 1$. In Fig. C.2 we see that this can be assumed to hold on the compact sets
 551 on which logit means and variances tend to live (see also Fig. C.1).

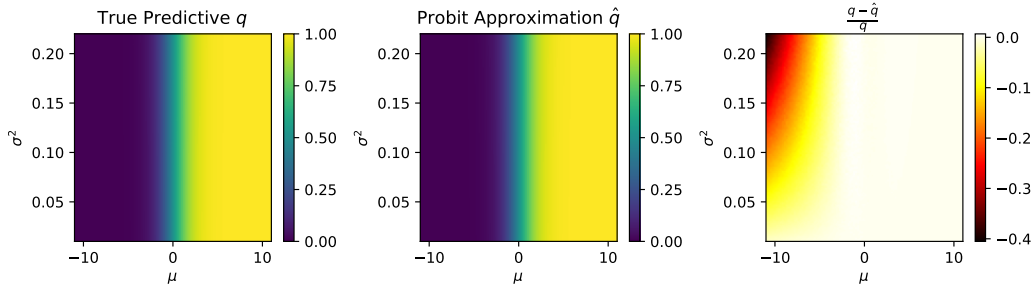


Figure C.2: Plot of the ‘true’ predictive q (approximated with a 50000 sample MC approximation) versus the probit approximation \hat{q} , as well as $\frac{q - \hat{q}}{q}$. We see that $\frac{q - \hat{q}}{q}$ tends to be negative, making the assumption $\sup_{(\mu, \sigma^2) \in \mathcal{K}} \frac{q - \hat{q}}{q} < 1$ from Theorem C.2 easily fulfilled.

552 D Moment Matching Beta Distributions

553 As noted in Section 3.2, when $\varphi = \Phi$ or ρ , we can construct a mapping

$$p: \mathcal{G}(\mathbb{R}^C) \rightarrow \mathcal{B}((0, 1))^C \quad (59)$$

554 by moment matching. Specifically, the parameters $\alpha, \beta \in (0, \infty)^C$ that match the moments of
 555 $Q \sim \varphi_* f$ for some $f \in \mathcal{G}(\mathbb{R}^C)$ must satisfy

$$\begin{aligned} \mathbb{E}[Q] &= \frac{\alpha}{\alpha + \beta}, \\ \mathbb{E}[Q^2] &= \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}, \end{aligned} \quad (60)$$

556 where all vector operations are element-wise. Multiplying out the denominators on the right-hand
 557 side of the equations of Eq. (60), we obtain a system of two linear equations with two unknowns (for
 558 each c), which can be solved uniquely, yielding

$$\begin{aligned} \alpha &:= \frac{\mathbb{E}[Q] - \mathbb{E}[Q^2]}{\mathbb{E}[Q^2] - \mathbb{E}[Q]^2} \mathbb{E}[Q], \\ \beta &:= \frac{\mathbb{E}[Q] - \mathbb{E}[Q^2]}{\mathbb{E}[Q^2] - \mathbb{E}[Q]^2} (1 - \mathbb{E}[Q]). \end{aligned} \quad (61)$$

559 which give us the parameters of the Beta distributions $p(f)$.

560 D.1 Information Geometric Interpretation of the Pushforward through NormCDF

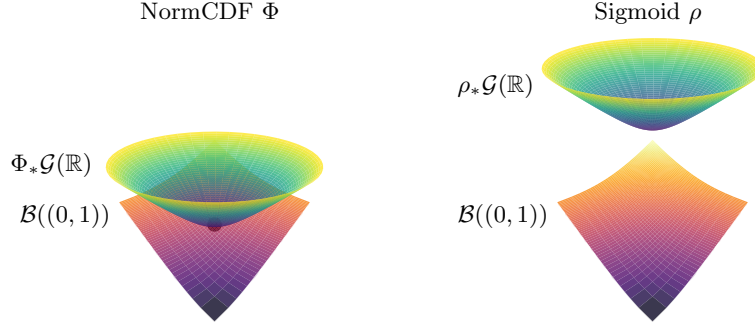


Figure D.1: Illustration of the statistical manifolds of the pushforward Gaussian distributions $\mathcal{G}(\mathbb{R})$ through the normCDF and the sigmoid respectively compared to the statistical manifold of Beta distributions $\mathcal{B}((0,1))$. NormCDF, unlike the sigmoid, makes the manifolds intersect at the point $\Phi_*\mathcal{N}(0,1) = \mathcal{B}(1,1)$.

561 Here, we argue that the normCDF activation is an ideal choice for approximating Gaussian push-
 562 forwards with Beta distributions by interpreting Fig. D.1. This extends the work from [22], as it
 563 shows how one can make sense of the ‘right’ basis for performing Laplace approximations in the
 564 classification setting.

565 $\Phi_*\mathcal{G}(\mathbb{R})$, $\rho_*\mathcal{G}(\mathbb{R})$ the space of pushforwards of Gaussian distributions by normCDF and sigmoid
 566 respectively, and $\mathcal{B}((0,1))$, the space of Beta distributions, are statistical manifolds naturally equipped
 567 with Riemannian metrics, that is their respective Fisher information metrics. We would like to
 568 visualise these manifolds. However, two difficulties arise.

569 The first difficulty is that, while these manifolds all lie in the infinite-dimensional vector space of
 570 signed measures on the open unit interval $\mathcal{M}((0,1))$, there is *no* subspace $\mathbb{V} \subset \mathcal{M}((0,1))$ which is
 571 3-dimensional ($\mathbb{V} \cong \mathbb{R}^3$) and contains any two of these statistical manifolds ($\Phi_*\mathcal{G}(\mathbb{R}), \mathcal{B}((0,1)) \subset \mathbb{V}$
 572 or $\rho_*\mathcal{G}(\mathbb{R}), \mathcal{B}((0,1)) \subset \mathbb{V}$). We will work around this by building distinct isometric embeddings
 573 $\Phi_*\mathcal{G}(\mathbb{R}) \hookrightarrow \mathbb{V}$, $\rho_*\mathcal{G}(\mathbb{R}) \hookrightarrow \mathbb{V}$ and $\mathcal{B}((0,1)) \hookrightarrow \mathbb{V}$ for some 3-dimensional vector space \mathbb{V} . This
 574 means that while the shape of the manifold illustrations is meaningful, the positioning of a manifold
 575 with respect to another is not, apart from some design choices that we describe below.

576 The second difficulty is that some—if not all—of these manifolds cannot be embedded isometrically
 577 into Euclidean space. As a workaround, we instead embed them into the 3-dimensional Minkowski
 578 space $\mathbb{R}^{2,1}$, that is \mathbb{R}^3 equipped with the pseudo-Riemannian metric $dx_1^2 + dx_2^2 - dx_3^2$.

579 The key observation is that $\Phi_*\mathcal{G}(\mathbb{R})$ and $\rho_*\mathcal{G}(\mathbb{R})$ are isometric to $\mathcal{G}(\mathbb{R})$. This is because Φ and ρ are
 580 diffeomorphisms, so in particular sufficient statistics, and the Fisher information metric is invariant
 581 under sufficient statistics [2, Section 5.1.4]. Visually, this means that $\varphi_*\mathcal{G}(\mathbb{R})$ has the same shape
 582 irrespectively of the diffeomorphism activation function φ . One can thus observe that, given that
 583 $\mathcal{B}((0,1))$ is not isometric to $\mathcal{G}(\mathbb{R})$, there exists no activation φ such that $\varphi_*\mathcal{G}(\mathbb{R}) = \mathcal{B}((0,1))$. To
 584 design an activation φ that maps Gaussians to Betas, the best one can hope to do is to map one
 585 specific Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to a specific Beta distribution $\mathcal{B}(\alpha, \beta)$. This is done with
 586 the map $F_{\alpha,\beta}^{-1} \circ \Phi_{\mu,\sigma^2}$ where Φ_{μ,σ^2} and $F_{\alpha,\beta}$ are the cumulative distribution functions of $\mathcal{N}(\mu, \sigma^2)$
 587 and $\mathcal{B}(\alpha, \beta)$ respectively. Taking $\mu = 0$, $\sigma^2 = 1$, $\alpha = \beta = 1$ we get $\Phi_{0,1} = \Phi$ and $F_{1,1} = \text{id}_{(0,1)}$,
 588 yielding $F_{\alpha,\beta}^{-1} \circ \Phi_{\mu,\sigma^2} = \Phi$.

589 Now $\mathcal{G}(\mathbb{R})$, and hence $\Phi_*\mathcal{G}(\mathbb{R})$ and $\rho_*\mathcal{G}(\mathbb{R})$, is isometric to the hyperbolic plane Ay et al. [2, Example
 590 3.1]. We can embed this isometrically into Minkowski space with the classical hyperboloid model of
 591 the hyperbolic plane [31],

$$\mathcal{G}(\mathbb{R}) \hookrightarrow \mathbb{R}^{2,1}. \quad (62)$$

592 For $\mathcal{B}((0,1))$, we use the isometric embedding from Le Brigant et al. [18, Proposition 2]:

$$\begin{aligned} \mathcal{B}((0,1)) &\hookrightarrow \mathbb{R}^{2,1} \\ (\alpha, \beta) &\mapsto (\eta(\alpha), \eta(\beta), \eta(\alpha + \beta)) \end{aligned} \quad (63)$$

where $\eta(a) := \int_1^a \sqrt{\psi'(r)} dr$ and ψ is the digamma function.

Finally, we choose our embedding Eq. (62) for $\Phi_*\mathcal{G}(\mathbb{R})$ such that it intersects the embedding Eq. (63) at a point, to highlight that the statistical manifolds $\Phi_*\mathcal{G}(\mathbb{R})$ and $\mathcal{B}((0, 1))$ intersect at a point in the infinite-dimensional ambient space $\mathcal{M}((0, 1))$, while $\rho_*\mathcal{G}(\mathbb{R})$ and $\mathcal{B}((0, 1))$ do not.

Moreover, since $\Phi_*\mathcal{N}(0, 1) = B(1, 1)$, our moment matching approximation Eq. (60) is exact when $f(x)$ is the standard Normal distribution.

E List of Predictive and Dirichlet Parameters Formulas

E.1 Predictive Formulas

We gather formulas for all predictive estimators \hat{p} of the true predictive $\mathbb{E}_{P \sim \alpha_*\mathcal{N}(\mu, \Sigma)}[P]$, $\Sigma = \text{diag}(\sigma^2)$, used in our experiments (Section 6).

E.1.1 Softmax

Monte Carlo One can Monte Carlo estimate the true predictive as follows:

$$\hat{p} := \frac{1}{S} \sum_{s=1}^S \hat{p}^{(s)} \quad (64)$$

where the $\hat{p}^{(s)}$ are sampled i.i.d. from $\mathcal{N}(\mu, \Sigma)$.

Mean Field (Appendix A.4) The Mean Field predictive [21] uses the following approximation for the true predictive:

$$\hat{p} := \text{softmax} \left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8} \sigma^2}} \right). \quad (65)$$

Laplace Bridge The Laplace Bridge predictive [14] approximates the true predictive as follows:

$$\hat{p} := \frac{\frac{1}{\tilde{\sigma}^2} \left(1 - \frac{2}{C} + \frac{e^{\tilde{\mu}}}{C^2} \sum_{c=1}^C e^{-\tilde{\mu}_c} \right)}{\sum_{c=1}^C \frac{1}{\tilde{\sigma}_c^2} \left(1 - \frac{2}{C} + \frac{e^{\tilde{\mu}}}{C^2} \sum_{c'=1}^C e^{-\tilde{\mu}_{c'}} \right)} \quad (66)$$

where

$$\tilde{\mu}^2 := \sqrt{\frac{\sqrt{C/2}}{\sum_{c=1}^C \sigma_c^2}} \mu, \quad \tilde{\sigma}^2 := \frac{\sqrt{C/2}}{\sum_{c=1}^C \sigma_c^2} \sigma^2. \quad (67)$$

E.1.2 Exp

Our closed-form predictive for the exp activation function (Section 2.1) is given by

$$\hat{p} := \frac{\exp \left(\mu + \frac{\sigma^2}{2} \right)}{\sum_{c=1}^C \exp \left(\mu_c + \frac{\sigma_c^2}{2} \right)}. \quad (68)$$

E.1.3 NormCDF

The closed-form predictive for the normCDF activation function (Section 2.2) is computed as

$$\hat{p} := \frac{\Phi \left(\frac{\mu}{\sqrt{1 + \sigma^2}} \right)}{\sum_{c=1}^C \Phi \left(\frac{\mu_c}{\sqrt{1 + \sigma_c^2}} \right)}. \quad (69)$$

614 E.1.4 Sigmoid

615 For the sigmoid activation function (Section 2.2), the closed-form predictive can be computed as

$$\hat{\mathbf{p}} := \frac{\rho\left(\frac{\boldsymbol{\mu}}{\sqrt{1+\frac{\pi}{8}}\boldsymbol{\sigma}^2}\right)}{\sum_{c=1}^C \rho\left(\frac{\boldsymbol{\mu}_c}{\sqrt{1+\frac{\pi}{8}}\boldsymbol{\sigma}_c^2}\right)}. \quad (70)$$

616 E.2 Dirichlet Parameters Formulas

617 We now gather the formulas of the parameters γ for the Dirichlet approximations to the Gaussian
618 pushforwards.

619 E.2.1 Softmax

620 The Laplace Bridge method [14] uses the following Dirichlet parameters:

$$\boldsymbol{\gamma} := \frac{1}{\tilde{\boldsymbol{\sigma}}^2} \left(1 - \frac{2}{C} + \frac{e^{\tilde{\boldsymbol{\mu}}}}{C^2} \sum_{c=1}^C e^{-\tilde{\boldsymbol{\mu}}_c} \right) \quad (71)$$

621 where

$$\tilde{\boldsymbol{\mu}}^2 := \sqrt{\frac{\sqrt{C/2}}{\sum_{c=1}^C \sigma_c^2}} \boldsymbol{\mu}, \quad \tilde{\boldsymbol{\sigma}}^2 := \frac{\sqrt{C/2}}{\sum_{c=1}^C \sigma_c^2} \boldsymbol{\sigma}^2. \quad (72)$$

622 E.2.2 Exp

623 Exp uses the closed-form parameters derived from Moment Matching the Gaussian pushforwards
624 (Section 3.1):

$$\boldsymbol{\gamma} := \left(\prod_{c=1}^C \frac{a_c \cdot \max(\sum_{c'=1}^C a_{c'}, 1) - b_c}{b_c - a_c^2} \right)^{1/C} \frac{\mathbf{a}}{\sum_{c=1}^C a_c} \quad (73)$$

625 where

$$\mathbf{a} = \exp\left(\boldsymbol{\mu} + \frac{\boldsymbol{\sigma}^2}{2}\right), \quad \mathbf{b} = \exp(2\boldsymbol{\mu} + 2\boldsymbol{\sigma}^2). \quad (74)$$

626 E.2.3 NormCDF

627 NormCDF uses the closed-form parameters derived from Moment Matching the Gaussian pushfor-
628 wards (Section 3.1):

$$\boldsymbol{\gamma} := \left(\prod_{c=1}^C \frac{a_c \cdot \max(\sum_{c'=1}^C a_{c'}, 1) - b_c}{b_c - a_c^2} \right)^{1/C} \frac{\mathbf{a}}{\sum_{c=1}^C a_c} \quad (75)$$

629 where

$$\mathbf{a} = \Phi\left(\frac{\boldsymbol{\mu}}{\sqrt{1+\boldsymbol{\sigma}^2}}\right), \quad \mathbf{b} = \Phi\left(\frac{\boldsymbol{\mu}}{\sqrt{1+\boldsymbol{\sigma}^2}}\right) - 2\mathbf{T}\left(\frac{\boldsymbol{\mu}}{\sqrt{1+\boldsymbol{\sigma}^2}}, \frac{1}{\sqrt{1+2\boldsymbol{\sigma}^2}}\right). \quad (76)$$

630 E.2.4 Sigmoid

631 Sigmoid uses the closed-form parameters derived from Moment Matching the Gaussian pushforwards
632 (Section 3.1):

$$\boldsymbol{\gamma} := \left(\prod_{c=1}^C \frac{a_c \cdot \max(\sum_{c'=1}^C a_{c'}, 1) - b_c}{b_c - a_c^2} \right)^{1/C} \frac{\mathbf{a}}{\sum_{c=1}^C a_c} \quad (77)$$

633 where

$$\begin{aligned} a &= \rho \left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8} \sigma^2}} \right), \\ b &= \rho \left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8} \sigma^2}} \right) - \frac{1}{\sqrt{1 + \frac{\pi}{8} \sigma^2}} \rho \left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8} \sigma^2}} \right) \left(1 - \rho \left(\frac{\mu}{\sqrt{1 + \frac{\pi}{8} \sigma^2}} \right) \right). \end{aligned} \quad (78)$$

634 **F List of Uncertainty Estimators**

635 In this section, we list the uncertainty estimators used in our experiments (Section 6).

636 **F.1 Predictive**

637 Given a predictive p , we consider two uncertainty estimators.

638 **Maximum Probability**

$$\arg \max_{c \in \{1, \dots, C\}} p_c. \quad (79)$$

639 **Entropy**

$$-\sum_{c=1}^C p_c \log p_c. \quad (80)$$

640 **F.2 Monte Carlo**

641 Given S Monte Carlo samples $\hat{p}^{(1)}, \dots, \hat{p}^{(S)}$ with mean \hat{p} , one can calculate a predictive as their
642 average and derive the estimators in Appendix F.1. However, Monte Carlo samples allow one to
643 calculate two additional estimators detailed below.

644 **Expected Entropy**

$$-\frac{1}{S} \sum_{s=1}^S \sum_{c=1}^C \hat{p}_c^{(s)} \log \hat{p}_c^{(s)}. \quad (81)$$

645 **Mutual Information/Jensen-Shannon Divergence**

$$-\sum_{c=1}^C \hat{p}_c \log \hat{p}_c + \frac{1}{S} \sum_{s=1}^S \sum_{c=1}^C \hat{p}_c^{(s)} \log \hat{p}_c^{(s)}. \quad (82)$$

646 **F.3 Dirichlet**

647 Given a second-order Dirichlet distribution with parameters γ , one can obtain the expected entropy
648 and mutual information estimators without the need for Monte Carlo samples.

649 **Expected Entropy**

$$-\sum_{c=1}^C \frac{\gamma_c}{\sum_{c'=1}^C \gamma_{c'}} \left(\psi(\gamma_c + 1) - \psi \left(\sum_{c'=1}^C \gamma_{c'} + 1 \right) \right) \quad (83)$$

650 where ψ is the digamma function.

651 **Mutual Information**

$$\sum_{c=1}^C \frac{\gamma_c}{\sum_{c'=1}^C \gamma_{c'}} \left(-\log \gamma_c + \log \left(\sum_{c'=1}^C \gamma_{c'} \right) + \psi(\gamma_c + 1) - \psi \left(\sum_{c'=1}^C \gamma_{c'} + 1 \right) \right). \quad (84)$$

G Experimental Setup

This section describes our experimental setup in detail.

We have two main research questions:

- What are the effects of changing the learning objective?
- Do we have to sacrifice performance for sample-free predictives?

To answer the first question, we evaluate our closed-form predictives (Sigmoid, NormCDF, Exp) and moment-matched Dirichlet distributions against softmax models equipped with approximate inference tools (Laplace Bridge [14], Mean Field [21], Monte Carlo sampling). We consider **Heteroscedastic Classifiers (HET)** [5], **Spectral-Normalized Gaussian Processes (SNGP)** [20], and last-layer **Laplace approximation** methods [8] as backbones (see Appendix I for details).

The resulting 21 (method, activation, predictive) triplets are evaluated on ImageNet-1k [9] and CIFAR-10 [17] on five metrics aligning with practical needs from uncertainty estimates [26]:

1. Log probability proper scoring rule for the predictive,
2. Expected calibration error of the predictive’s maximum-probability confidence,
3. Binary log probability proper scoring rule for the correctness prediction task,
4. Accuracy of the predictive’s argmax,
5. AUROC for the out-of-distribution (OOD) detection task.

See Appendix H for details.

For ImageNet, we treat ImageNet-C [13] samples with 15 corruption types and 5 severity levels as OOD samples. For CIFAR-10, we use the CIFAR-10C corruptions.

For the second question, we consider fixed (method, activation) pairs and test whether our methods perform on par with the Monte Carlo sampled predictives.

To provide a fair comparison, we reimplement each method as simple-to-use wrappers around deterministic backbones.

For ImageNet evaluation, we use a ResNet-50 backbone pretrained with the softmax activation function, and train each (method, activation) pair for 50 ImageNet-1k epochs following Mucsányi et al. [27]. We train with the LAMB optimiser [40] using a batch size of 128 and gradient accumulation across 16 batches, resulting in an effective batch size of 2048, following Tran et al. [36]. We further use a cosine learning rate schedule with a single warmup epoch using a warmup learning rate of 0.0001. The learning rate is treated as a hyperparameter and selected from the interval $[0.0005, 0.05]$ based on the validation performance. The weight decay is selected from the set $\{0.01, 0.02\}$. During training, we keep track of the best-performing checkpoint on the validation set and load it before testing. We search for ideal hyperparameters with a ten-step Bayesian Optimization scheme [33] in Weights & Biases [3] based on the negative log-likelihood.

On CIFAR-10, we train ResNet-28 models from scratch for 100 epochs. The only exceptions are the SNGP models that are trained for 125 epochs [20]. We train with Momentum SGD using a batch size of 128 and no gradient accumulation. Similarly to ImageNet, we use a cosine learning rate schedule but with five warmup epochs and warmup learning rate $1e-5$. The learning rate is also treated as a hyperparameter on CIFAR-10. We use the interval $[0.05, 1]$ for Sigmoid and NormCDF, and $[0.01, 0.15]$ for Softmax and Exp. The optimal learning rates for Sigmoid and NormCDF are generally larger, as the class-wise binary cross-entropies are averaged instead of summed. The weight decay is selected from the interval $[1e-6, 1e-4]$. Similarly to ImageNet, we use the best-performing checkpoint in the tests and use a ten-step Bayesian Optimization scheme to select performant hyperparameters.

The hyperparameter optimization, training, and evaluation of the methods used in this paper took 0.8 GPU years on NVIDIA RTX 2080Ti GPUs in a university compute cluster. The individual runs required no more than 50 GB of RAM and 3 days of runtime.

699 H Benchmark Metrics

700 Our experiments use five tasks/metrics:

- 701 1. Log probability proper scoring rule for the predictive,
- 702 2. Expected calibration error of the predictive’s maximum-probability confidence,
- 703 3. Binary log probability proper scoring rule for the correctness prediction task,
- 704 4. Accuracy of the predictive’s argmax,
- 705 5. AUROC for the out-of-distribution detection task.

706 Below, we describe these metrics and their respective tasks.

707 H.1 Log Probability Proper Scoring Rule for the Predictive

708 First, we briefly discuss proper and strictly proper scoring rules over general probability measures based on [26].

710 Consider a function $S: \mathcal{Q} \times \mathcal{Y} \rightarrow \mathbb{R}$ where \mathcal{Q} is a family of probability distributions over the space \mathcal{Y} , called the label space.

712 S is called a proper scoring rule if and only if

$$\max_{q \in \mathcal{Q}} \mathbb{E}_{Y \sim p} S(q, Y) = \mathbb{E}_{Y \sim p} S(p, Y), \quad (85)$$

713 i.e., p is *one of* the maximisers of S in q in expectation. S is further *strictly* proper if $\arg \max_{q \in \mathcal{Q}} \mathbb{E}_{Y \sim p} S(q, Y) = p$ is the *unique* maximiser of S in q in expectation.

715 The log probability scoring rule for categorical distributions is defined as

$$S(q, c) = \sum_{c'=1}^C \delta_{c,c'} \log q_{c'}(x) = \log q_c(x), \quad (86)$$

716 where $c \in \{1, \dots, C\}$ is the true class and δ is the Kronecker delta. S defined this way is a strictly proper scoring rule, i.e. $\mathbb{E}_{Y \sim p} S(q, Y)$ is maximal if and only if

$$q(Y = c | x) = p(Y = c | x) \quad \forall c \in \{1, \dots, C\}. \quad (87)$$

718 The score above is equivalent to the negative cross-entropy loss.

719 H.2 Expected Calibration Error

720 To set up the required quantities for the Expected Calibration Error (ECE) metric [29], we follow the steps below, based on [26].

- 722 1. Train a neural network on the training dataset.
- 723 2. Create predictions and confidence estimates on the test data.
- 724 3. Group the predictions into M bins based on the confidences estimates. Define bin B_m to be the set of all indices n of predictions (\hat{y}_n, \tilde{c}_n) for which

$$\tilde{c}_n \in \left(\frac{m-1}{M}, \frac{m}{M} \right]. \quad (88)$$

726 The Expected Calibration Error (ECE) metric [29] is then defined as

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (89)$$

727 where

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{n \in B_m} 1(\hat{y}_n = c_n), \quad (90)$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{n \in B_m} \max_{1 \leq c \leq C} f_c(x_n). \quad (91)$$

Intuitively, the ECE is high when the model’s per-bin confidences match its accuracy on the bin. We use $M = 15$ bins in this paper.

H.3 Binary Log Probability Proper Scoring Rule for Correctness Prediction

The correctness prediction task measures the models’ ability to predict the correctness of their own predictions. We consider *correctness estimators* $\tilde{c}(x) \in [0, 1]$ for inputs $x \in \mathcal{X}$ derived from the predictives. Framed as a binary prediction task, the goal of these estimators is to predict the probability of the predicted class’ correctness. In particular, for an (input, target) pair (x, y) with $x \in \mathcal{X}, y \in \mathcal{Y}$, we set the correctness target to

$$\ell \equiv \ell(x, y) = 1 \left(\max_{1 \leq c \leq C} h_c(x) = y \right). \quad (92)$$

Dropping the dependency on $x \in \mathcal{X}$ for brevity, the log probability score for binary targets $\ell \in \{0, 1\}$ and estimators $\tilde{c} \in [0, 1]$ is defined as

$$S(\tilde{c}, \ell) = \begin{cases} \log c & \text{if } \ell = 1 \\ \log(1 - \tilde{c}) & \text{if } \ell = 0 \end{cases} = \ell \log \tilde{c} + (1 - \ell) \log(1 - \tilde{c}). \quad (93)$$

One can show that this is indeed a strictly proper scoring rule [26].

H.4 Accuracy

For completeness, the accuracy of a predictive h on a dataset $(x_n, c_n)_{n=1}^N$ is

$$\text{acc}(h; (x_n, c_n)_{n=1}^N) = \frac{1}{N} \sum_{n=1}^N 1 \left(\arg \max_{c \in \{1, \dots, C\}} h_c(x_n) = \tilde{c}_n \right). \quad (94)$$

H.5 Area Under the Receiver Operating Characteristic Curve for Out-of-Distribution Detection

Out-of-distribution detection is another binary prediction task where a general uncertainty estimator $u(x) \in \mathbb{R}$ (derived from predictives or second-order Dirichlet distributions) is tasked to separate ID and OOD samples from a balanced mixture. The target OOD indicator variable $o(x)$ is, therefore, binary. As the uncertainty estimator can take on any real value, we measure the Area Under the Receiver Operating Characteristic curve (AUROC), which quantifies the separability of ID and OOD samples w.r.t. the uncertainty estimator.

Given uncertainty estimates $u_n \equiv u(x_n)$ and target binary labels o_i on a balanced dataset $(x_n, o_n)_{n=1}^N$, as well as a threshold $t \in \mathbb{R}$, we predict 1 (out-of-distribution) when $u_n \geq t$ and 0 (in-distribution) when $u_n < t$. This lets us define the following index sets:

$$\begin{aligned} \text{True positives: } \text{TP}(t) &= \{n : o_n = 1 \wedge u_n \geq t\} \\ \text{False positives: } \text{FP}(t) &= \{n : o_n = 0 \wedge u_n \geq t\} \\ \text{False negatives: } \text{FN}(t) &= \{n : o_n = 1 \wedge u_n < t\} \\ \text{True negatives: } \text{TN}(t) &= \{n : o_n = 0 \wedge u_n < t\}. \end{aligned} \quad (95)$$

The Receiver Operating Characteristic (ROC) curve compares the following quantities:

$$\begin{aligned} \text{TPR}(t) &= \frac{|\text{TP}(t)|}{|\text{TP}(t)| + |\text{FN}(t)|} = \frac{|\text{TP}(t)|}{|P|} \\ \text{FPR}(t) &= \frac{|\text{FP}(t)|}{|\text{FP}(t)| + |\text{TN}(t)|} = \frac{|\text{FP}(t)|}{|N|}. \end{aligned} \quad (96)$$

Here, FPR tells us how many of the actual negative samples in the dataset are recalled (predicted positive) at threshold t .

One can draw a curve of $(\text{FPR}(t), \text{TPR}(t))$ for all t from $-\infty$ to $+\infty$. This is the *ROC curve*. The area under this curve quantifies how well the uncertainty estimator $u(x)$ can separate in-distribution and out-of-distribution inputs.

I Benchmarked Methods

This section describes our benchmarked methods and provides further implementation details.

I.1 Spectral Normalized Gaussian Process

Spectral normalized Gaussian processes (SNGP) [20] use spectral normalization of the parameter tensors for distance-awareness and a last-layer Gaussian process approximated by Fourier features to capture uncertainty. For an input $x \in \mathcal{X}$ and number of classes C , they predict a C -variate Gaussian distribution

$$\mathcal{N}\left(\mathbf{B}\phi(x), \phi(x)^\top (\Psi^\top \Psi + \mathbf{I})^{-1} \phi(x) \mathbf{I}_C\right) \quad (97)$$

in logit space.

- $\mathbf{B} \in \mathbb{R}^{C \times D}$ is a *learned* parameter matrix that maps pre-logits to logits.
- $\phi(x) = \cos(\mathbf{W} \mathbf{f}^{L-1}(x) + \mathbf{b}) \in \mathbb{R}^D$ is a random pre-logit embedding of the input $x \in \mathcal{X}$. $\mathbf{f}^{L-1}(x)$ denotes the pre-logit embedding. \mathbf{W} is a *fixed* semi-orthogonal random matrix, and \mathbf{b} is also a *fixed* random vector but sampled from $\text{Uniform}(0, 2\pi)$.
- $\Psi^\top \Psi$ is the (unnormalised) empirical covariance matrix of the pre-logits of the training set. This is calculated by accumulating the mini-batch estimates during the last epoch.³

The method applies spectral normalization to the hidden weights in each layer using a power iteration scheme with a single iteration per batch to obtain the largest singular value. Liu et al. [20] claim this helps with input distance awareness.

I.2 Heteroscedastic Classifier

Heteroscedastic classifiers (HET) [5] construct a Gaussian distribution in the logit space to model per-input uncertainties:

$$\mathcal{N}(\mathbf{f}(x), \Sigma(x)), \quad (98)$$

where $\mathbf{f}(x) \in \mathbb{R}^D$ is the logit mean for input $x \in \mathcal{X}$ and

$$\Sigma(x) = \mathbf{V}(x)^\top \mathbf{V}(x) + \text{diag}(\mathbf{d}(x)) \quad (99)$$

is a (positive definite) covariance matrix. Both the low-rank term $\mathbf{V}(x)$ and the diagonal term $\mathbf{d}(x)$ are calculated as a linear function of the pre-logit layer’s output.

To learn the per-input covariance matrices from the training set, one has to construct a predictive estimate from $\mathcal{N}(\mathbf{f}(x), \Sigma(x))$ using any of the methods in Appendix E. This predictive estimate is then trained using a standard cross-entropy (NLL) loss.

HET uses a temperature parameter to scale the logits before calculating the BMA. This is chosen using a validation set.

The off-diagonal terms of the covariance matrix do not affect the approximate predictive (Eq. (12)). This means that, in our framework, one can discard the low-rank term $\mathbf{V}(x)$ and only model the diagonal term $\mathbf{d}(x)$ without a decrease in expressivity. To keep comparisons fair and use the same backbone with the same number of parameters, we also only model $\mathbf{d}(x)$ for softmax-based predictives.

I.3 Laplace Approximation

The Laplace approximation [8] approximates the posterior $p(\theta \mid \mathcal{D})$ over the network parameters θ for a Gaussian prior $p(\theta)$ and likelihood defined by the network architecture by a Gaussian. In its simplest form, it uses the maximum a posteriori (MAP) weights $\theta_{\text{MAP}} \in \mathbb{R}^P$ as the mean and the

³As we use a cosine learning rate decay in all experiments, the model makes negligible changes in its pre-logit feature space in the last epoch. Thus, the empirical covariance matrix is approximately consistent.

Table J.1: Comparison of ECE results for different predictives using a fixed Laplace backbone.

Method	Mean	Std
Softmax Laplace		
MC 100	0.0096	0.0013
MC 1000	0.0102	0.0016
MC 10	0.0120	0.0013
Mean Field	0.0121	0.0029
Laplace Bridge Predictive	0.5933	0.0105
NormCDF Laplace		
Closed-form	0.0074	0.0012
MC 1000	0.0092	0.0022
MC 100	0.0095	0.0015
MC 10	0.0100	0.0020

inverse Hessian of the *regularised* loss over the training set $\tilde{\mathcal{L}}(\theta; \mathcal{D}) = \mathcal{L}(\theta; \mathcal{D}) + \lambda \|\theta\|_2^2$ evaluated at the MAP as the covariance matrix:

$$\mathcal{N} \left(\theta_{\text{MAP}}, \left(\frac{\partial^2 \tilde{\mathcal{L}}(\theta; \mathcal{D})}{\partial \theta_i \partial \theta_j} \bigg|_{\theta_{\text{MAP}}} \right)^{-1} \right) = \mathcal{N} \left(\theta_{\text{MAP}}, \left(\frac{\partial^2 \mathcal{L}(\theta; \mathcal{D})}{\partial \theta_i \partial \theta_j} \bigg|_{\theta_{\text{MAP}}} + \lambda \mathbf{I}_P \right)^{-1} \right). \quad (100)$$

This is a *locally optimal* post-hoc Gaussian approximation of the true posterior $p(\theta \mid \mathcal{D})$ based on a second-order Taylor approximation. For details, see [35].

For deep neural networks, the Hessian matrix is often replaced with the Generalized Gauss-Newton (GGN) matrix. The GGN is guaranteed to be positive semidefinite even for suboptimal weights and has efficient approximation schemes, such as Kronecker-Factored Approximate Curvature [24] or low-rank approximations.

Denoting our curvature estimate of choice by G , the logit-space Gaussian is obtained by pushing forward the weight-space Gaussian measure through the *linearised* model around θ_{MAP} . For an input $x \in \mathcal{X}$, this results in

$$\mathcal{N} \left(\mathbf{f}(x, \theta_{\text{MAP}}), (\mathbf{J}_{\theta_{\text{MAP}}} \mathbf{f}(x)) (G + \lambda \mathbf{I}_P)^{-1} (\mathbf{J}_{\theta_{\text{MAP}}} \mathbf{f}(x))^\top \right), \quad (101)$$

where $\mathbf{J}_{\theta_{\text{MAP}}} \mathbf{f}(x) \in \mathbb{R}^{C \times P}$ is the model Jacobian matrix.

We use a last-layer KFAC Laplace variant in our experiments and use the *full* training set for calculating the GGN instead of a mini-batch based on recent works on the bias in mini-batch estimates [35].

J CIFAR-10 Experiments

This appendix section repeats the experiments presented in the main paper on the CIFAR-10 dataset. For a detailed description of the experimental setup, refer to Appendix G. Appendix H describes the used tasks and metrics.

As stated in the main paper, our two research questions are:

- What are the effects of changing the learning objective? (Appendix J.2)
- Do we have to sacrifice performance for sample-free predictives? (Appendix J.1)

J.1 Quality of Sample-Free Predictives

Similarly to the main paper, in this section, we investigate our first research question: whether there is a price to pay for sample-free predictives. Table J.1 showcases the two best-performing (activation, method) pairs on the ECE metric and the CIFAR-10 dataset: Softmax and NormCDF Laplace. Mean Field (MF) is a strong alternative for sample-free predictives, but it has no guarantees and can fall behind MC sampling (see also Fig. 1). Empirically, our closed-form predictives always perform on par with MC sampling.

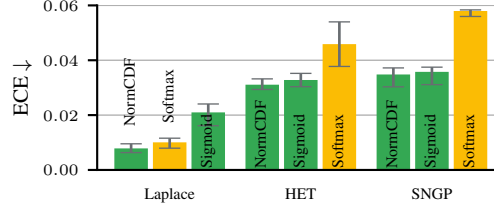


Figure J.1: CIFAR-10 ECE results. Our closed-form predictives (■) outperform Softmax (■) on HET and SNGP. Laplace tunes its hyperparameters based on the ECE metric – NormCDF Laplace is the overall best method. Note the restricted y -limits for readability.

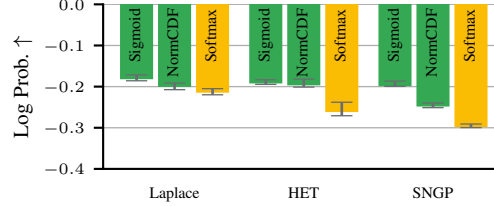


Figure J.2: CIFAR-10 log probability proper scoring results for the binary correctness prediction task. Our closed-form predictives (■) consistently outperform Softmax (■) on all methods.

J.2 Effects of Changing the Learning Objective

As in the main paper, in this section, we use the best-performing predictive and estimator (see Appendix E) for softmax models and employ our methods with the closed-form predictives.

J.2.1 Calibration and Proper Scoring

We first evaluate calibration using the log probability scoring rule [12] and the Expected Calibration Error (ECE) metric [29]. Fig. 3b shows that on CIFAR-10, the score of our closed-form predictives (Sigmoid, NormCDF) are consistently better than the corresponding softmax results for all methods.

Fig. J.1 shows that on CIFAR-10, our closed-form predictives have a clear advantage on HET and SNGP. Laplace is a post-hoc method that tunes its hyperparameters on the ECE metric, hence its enhanced performance. Our NormCDF predictive is on par with Softmax.

J.2.2 Correctness Prediction

Fig. J.2 shows that on the correctness prediction task, our closed-form predictives outperform all Softmax predictives across all methods, as measured by the log probability proper scoring rule.

J.2.3 Accuracy

Closed-form predictives do not sacrifice accuracy. Fig. J.3 evidences this claim on CIFAR-10: our closed-form predictives either outperform or are on par with Softmax predictives. The most accurate method is Sigmoid HET. These results support the findings of Wightman et al. [38] that showcase desirable training dynamics of the class-wise cross-entropy loss.

J.2.4 Out-of-Distribution Detection

Finally, we consider the OOD detection task on a balanced mixture of ID (CIFAR-10) and OOD inputs. As OOD inputs, we consider corrupted CIFAR-10C samples. We use the AUROC metric to evaluate the methods' performance. As shown in Fig. J.4, the best-performing method is Sigmoid SNGP, a closed-form method. Generally, Softmax performs on par with our closed-form predictives. Intuitively, separating ID and OOD samples does not require a fine-grained representation of uncertainty, unlike the ECE or proper scoring rules. Nevertheless, the closed-form predictives

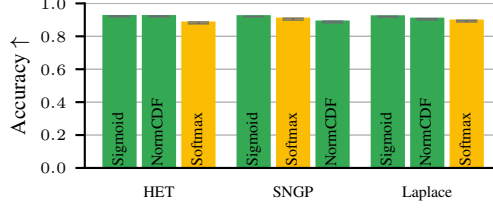


Figure J.3: **Closed-form predictives do not sacrifice accuracy.** CIFAR-10 accuracies. Our closed-form predictives (■) either outperform or are on par with Softmax (■) across all methods.

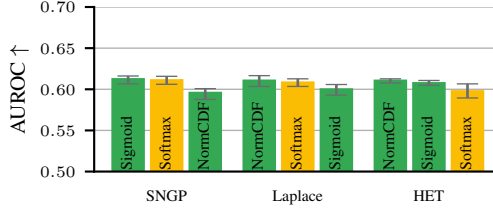


Figure J.4: CIFAR-10C OOD detection AUROC results for severity level one. Across all methods, the best-performing predictive is closed-form (■).

849 and second-order Dirichlet distributions are considerably cheaper to calculate than Softmax MC
 850 predictions (see Section 5).

851 K Additional Results

852 K.1 Closed-form Softmax Predictives

853 One can directly apply Eq. (10) to a neural network trained with the softmax cross-entropy loss
 854 in Eq. (4) to obtain closed-form predictives. However, we empirically found this approach to decrease
 855 performance, as the denominator of Eq. (10) increased to the order of billions during training.
 856 Considering a multivariate normal random variable $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ representing a logit distribution,

$$\text{Var} \left(\sum_{i=1}^C e^{x_i} \right) \geq \text{Var}(e^{x_j}) = e^{2\mu_j + \Sigma_{jj}} \underbrace{(e^{\Sigma_{jj}} - 1)}_{>0}, \quad (102)$$

857 i.e., when e^{μ_j} is large for some $j \in \{1, \dots, C\}$, the variance of the denominator necessarily explodes,
 858 violating our assumption.

859 To mitigate this issue, one may optimize the regularised cross-entropy loss

$$\mathcal{L}((x_n, c_n)_{n=1}^N) = \tilde{\mathcal{L}}((x_n, c_n)_{n=1}^N) + \lambda \sum_{n=1}^N \left(\sum_{c=1}^C \exp(f_c(x_n)) - 1 \right)^2 \quad (103)$$

860 or the more numerically stable version

$$\mathcal{L}((x_n, c_n)_{n=1}^N) = \tilde{\mathcal{L}}((x_n, c_n)_{n=1}^N) + \lambda \sum_{n=1}^N \left(\log \sum_{c=1}^C \exp(f_c(x_n)) \right)^2 \quad (104)$$

861 for an appropriately tuned λ hyperparameter. The latter formulation can be stably trained even at
 862 an ImageNet scale. However, there are fundamental limitations to the predictive given by Eq. (10).
 863 Namely, for *isotropic* logit-space Gaussian distributions, $\sigma^2(x)/2$ introduces a constant shift in the log-
 864 its, under which the softmax activation function is invariant. Therefore, the predictive approximation
 865 collapses into the MAP prediction. The SNGP method predicts such isotropic logit-space Gaussians.
 866 Further, we empirically found that the Laplace method’s predictives for models trained with the
 867 regularised cross-entropy loss were also approximately isotropic; thus, this integral approximation
 868 yielded diminishing returns.

869 For completeness, we share the ImageNet and CIFAR-10 results of $\varphi = \exp$ in Tables K.1 and K.2.

Table K.1: Performance metrics of $\varphi = \text{exp}$ on the CIFAR-10 validation dataset. We report the mean and two standard deviations.

Metric	Exp SNGP	Exp HET	Exp Laplace
NLL	0.376 ± 0.007	0.353 ± 0.059	0.370 ± 0.024
ECE	0.041 ± 0.006	0.029 ± 0.007	0.045 ± 0.007
Log Prob.	-0.255 ± 0.005	-0.238 ± 0.038	-0.255 ± 0.016
Accuracy	0.884 ± 0.013	0.888 ± 0.023	0.893 ± 0.007
AUROC	0.592 ± 0.006	0.603 ± 0.009	0.596 ± 0.011

Table K.2: Performance metrics of $\varphi = \text{exp}$ on the ImageNet validation dataset. We report the mean and two standard deviations.

Metric	Exp SNGP	Exp HET	Exp Laplace
NLL	0.866 ± 0.031	0.929 ± 0.042	0.986 ± 0.149
ECE	0.058 ± 0.004	0.022 ± 0.007	0.058 ± 0.013
Log Prob.	-0.380 ± 0.009	-0.403 ± 0.011	-0.402 ± 0.014
Accuracy	0.784 ± 0.001	0.779 ± 0.001	0.785 ± 0.002
AUROC	0.606 ± 0.001	0.611 ± 0.001	0.615 ± 0.003

870 K.2 Vision Transformer Results on ImageNet

871 Table K.3 shows results on ViT Little [11] backbones from the `timm` [37] library on the ImageNet
 872 validation set. As in the main paper, the hyperparameters are optimized using a ten-step Bayesian
 873 hyperparameter optimization scheme of Weights and Biases.

Table K.3: NLL and accuracy metrics of ViT Little models on the ImageNet validation set. Error bars represent two standard deviations.

Method	NLL	Accuracy (%)
Laplace		
Softmax	0.81988 ± 0.0020	79.598 ± 0.145
NormCDF	0.79603 ± 0.0087	80.678 ± 0.042
Sigmoid	0.79479 ± 0.0002	80.104 ± 0.206
HET		
Softmax	0.87096 ± 0.0242	78.604 ± 0.378
NormCDF	0.79106 ± 0.0075	80.514 ± 0.425
Sigmoid	0.88175 ± 0.0356	78.738 ± 0.324
GP		
Softmax	0.86724 ± 0.0229	78.898 ± 0.187
NormCDF	0.76974 ± 0.0088	80.694 ± 0.575
Sigmoid	0.81178 ± 0.0115	79.980 ± 0.412

874 K.3 CIFAR-100 Results

875 We repeat the NLL and accuracy evaluation on CIFAR-100 using WideResNet-28-5 models following
 876 a ten-step Bayesian hyperparameter sweep on Weights and Biases. Results are shown in ??.

877 K.4 Alignment with the True Predictives

878 Table K.5 shows the divergence of predictives using various approximation techniques from the true
 879 predictive estimated using 10,000 Monte Carlo samples. Our closed-form predictives, NormCDF
 880 and Sigmoid, are favorable to both the mean field and Laplace bridge approximations.

Table K.4: NLL and accuracy metrics of WideResNet-28-5 models on the CIFAR-100 test set. Error bars represent two standard deviations.

Method	NLL	Accuracy (%)
Laplace		
Softmax	0.92176 ± 0.0128	78.225 ± 0.412
NormCDF	0.88823 ± 0.0094	78.900 ± 0.325
Sigmoid	0.94209 ± 0.0136	78.050 ± 0.436
HET		
Softmax HET	0.98513 ± 0.0254	77.050 ± 0.228
NormCDF HET	0.94643 ± 0.0022	78.117 ± 0.286
Sigmoid HET	0.95251 ± 0.0142	77.583 ± 0.175
SNGP		
Softmax SNGP	0.97156 ± 0.0348	77.750 ± 0.462
NormCDF SNGP	0.90011 ± 0.0106	79.010 ± 0.298
Sigmoid SNGP	0.99996 ± 0.0265	78.350 ± 0.215

Table K.5: Kullback-Leibler (KL) divergence to the true predictive distributions (estimated using 10,000 Monte Carlo samples) for different approximation methods on the ImageNet validation set using ViT Little backbones. The mean KL divergence is calculated over the validation set. Error bars represent two standard deviations over independently trained models using different seeds.

Method	KL Divergence to True Predictive
NormCDF	0.0057 ± 0.0008
Sigmoid	0.0064 ± 0.0009
Softmax mean field	0.0330 ± 0.0042
Softmax Laplace bridge	1.7900 ± 0.1254

881 K.5 BCE Loss Performance Gains

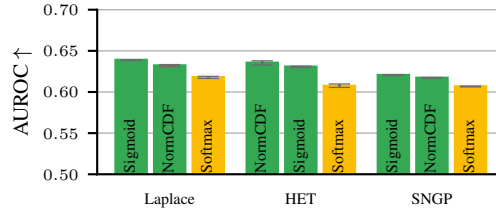
882 Table K.6 shows that the Softmax and Sigmoid activation functions (and corresponding losses)
883 already show improved performance on the *vanilla models* compared to softmax, highlighting that the
884 performance gains we observe cannot only be attributed to our closed-form predictive approximations
885 but also the favorable training dynamics of the class-wise BCE loss.

Table K.6: NLL results on ImageNet using ResNet-50 backbones and different activation functions. Error bars represent two standard deviations.

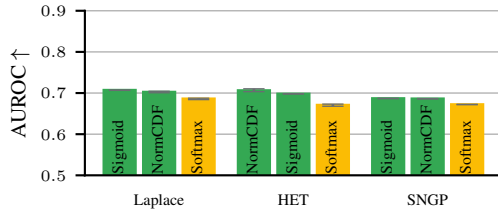
Activation Function	NLL
NormCDF	0.9345 ± 0.0072
Softmax	0.9369 ± 0.0025
Sigmoid	0.9146 ± 0.0046

886 K.6 Further Out-of-Distribution Detection Results

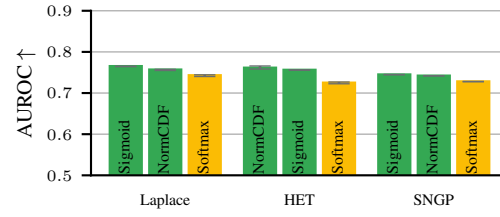
887 Fig. K.1 shows OOD detection results across all ImageNet-C severity levels. Our closed-form
888 predictives consistently outperform Softmax.



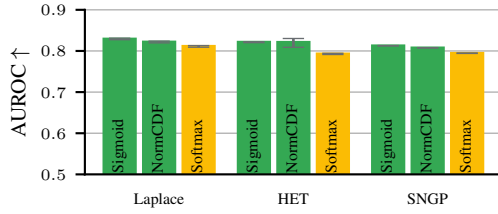
(a) OOD detection AUROC with severity level one.



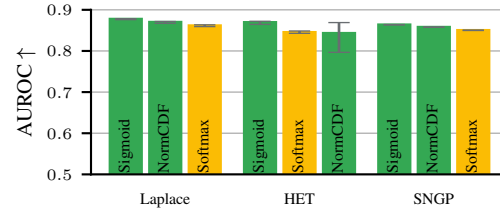
(b) OOD detection AUROC with severity level two.



(c) OOD detection AUROC with severity level three.



(d) OOD detection AUROC with severity level four.



(e) OOD detection AUROC with severity level five.

Figure K.1: The OOD detection performance of all methods increases steadily as we increase the severity of the perturbed half of the mixed dataset on the ImageNet validation dataset. Our closed-form predictives consistently outperform Softmax.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims of improved uncertainty quantification capabilities are thoroughly verified in Section 6 and Appendices J, K.3 and K.6.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper's limitations are discussed in Section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All theoretical claims include the full set of assumptions and are proved in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper accurately describes the experimental setup in Section 6 and Appendix G and the code is provided as supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See above.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The main training, implementation, and hyperparameter optimization details are discussed in Section 6 and Appendices G and I.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We run each experiment on five different seeds and report the min, mean, and max metric values.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed-form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The compute resources and time of execution are discussed at the end of Appendix G.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We fulfill all requirements.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our paper conducts no societally harmful research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: There is no risk in abusing ImageNet, CIFAR-100, or CIFAR-10 classifiers.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the ImageNet, ImageNet-Real, CIFAR-10, CIFAR-100, and CIFAR-10H datasets and the `timm` library for the code and the pretrained ResNet-50s.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets; we only record metrics.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not use crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not conduct experiments with humans.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- 1197
- 1198
- 1199
- 1200
- 1201
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
 - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

1202

16. Declaration of LLM usage

1203

1204

1205

1206

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

1207

Answer: [NA]

1208

Justification: The paper involves no use of LLMs.

1209

Guidelines:

- 1210
- 1211
- 1212
- 1213
- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
 - Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.