Uncertainty Estimation and Generalization Bounds for Modern Deep Learning

Advances in Function-Space Variational Inference, Linearized Laplace Approximation, Deep Ensembles, and Chernoff-Based Generalization Bounds

Luis Antonio Ortega Andrés
Supervisor: Daniel Hernández Lobato

March 7, 2026

Department of Computer Science and Telecomunications
Escuela Politécnica Superior
Universidad Autónoma de Madrid

Table of contents

Motivation

Deep learning is accurate, but:

- **Uncertainty** is often miscalibrated
  - overconfidence on OOD inputs
  - unreliable risk estimates
- **Generalization** at scale is not fully understood
  - over-parameterization / interpolation
  - why SGD solutions generalize?

Thesis perspective

Unify three tools:

- Bayesian inference
- Function-space modeling
- Large-deviation theory

Thesis goal

Methodological goal

- Scalable Bayesian uncertainty for deep models.
- Function-space inference as the main tool, rather than parameter-space.

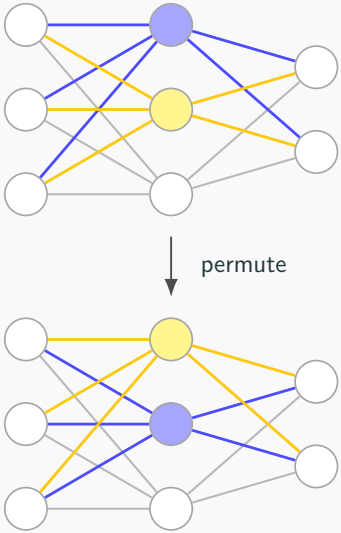
Theoretical goal

- Distribution-dependent view of generalization.
- Informative beyond capacity-only bounds (including interpolation).

Part I: Uncertainty estimation via function-space methods

Why function-space inference?

- Parameter-space posteriors are difficult in deep nets:
  - symmetries / non-identifiability
  - strong parameter dependencies
- Function-space inference is often better behaved:
  - directly targets predictions
  - avoids many parameterization pathologies
- Reference models in function space
  - Gaussian processes (GPs): canonical Bayesian prior on functions.
  - Implicit processes: sample-based priors that generalize GPs.



Implicit Processes $\rightarrow$ VIP $\rightarrow$ DVIP

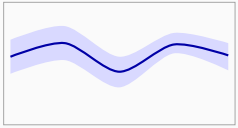
Implicit Process (IP)

- prior over functions
- easy to sample
- density is intractable

samples $f \sim p(f)$

Variational IP (VIP)

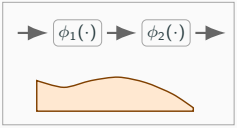
- GP surrogate for IP posterior
- matches mean & covariance
- linear-feature approximation



GP surrogate: mean $\pm$ uncertainty

Deep VIP (DVIP)

- composition of IP layers
- hidden-layer inputs random
- predictive outputs non-Gaussian



deep composition

DVIP definition

- **Implicit Process (IP).** An implicit process is a distribution over functions $f : \mathcal{X} \to \mathbb{R}$ specified by a *sampling procedure* rather than a tractable density:

$$f \sim \mathrm{IP}(g, P) \quad \Longleftrightarrow \quad \exists\, z \sim P \quad \text{s.t.} \quad f(\cdot) = g(\cdot, z).$$

  (Sampling is easy via $z \sim P$; the function-density is generally intractable)

- **Deep composition of IP layers.** For datapoint $\mathbf{x}_n$ and layer l:

$$\mathbf{f}^l_{h,n} = f^l_h(\mathbf{f}^{l-1}_{:,n}), \qquad \mathbf{f}^0_{:,n} = \mathbf{x}_n\,.$$

  where each hidden unit is an independent IP draw

$$f^l_h(\cdot) \sim \mathrm{IP}(g^l_h, P^l_h), \qquad h = 1, \ldots, H_l.$$

  Randomness propagates through layers because the inputs $\mathbf{f}^{l-1}_{:,n}$ are themselves random.

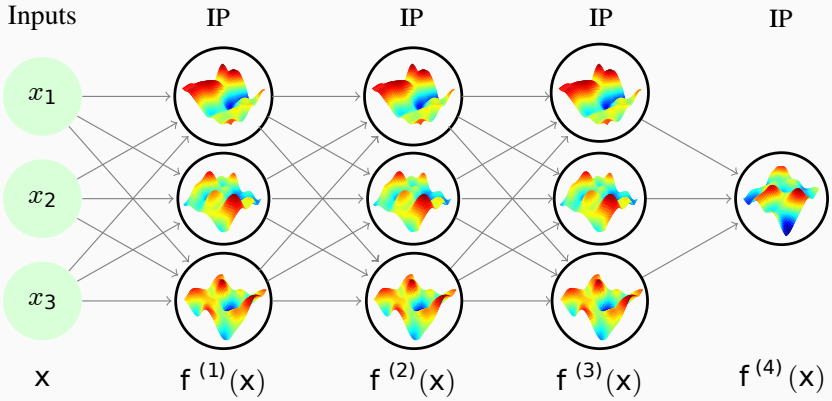
- **VIP/GP-style linear surrogate per unit (practical inference).** Approximate each unit-function by a linear model in learned features:

$$\mathbf{f}^l_{h,n} = \phi^l_h(\mathbf{f}^{l-1}_{:,n})^\top a^l_h + m^l_h(\mathbf{f}^{l-1}_{:,n}),$$

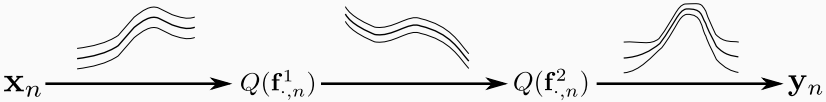
  where $\phi^l_h(\cdot)$ and $m^l_h(\cdot)$ are constructed from the sampled IP, and $\boldsymbol{a}$ are Gaussian weights.

DVIP architecture

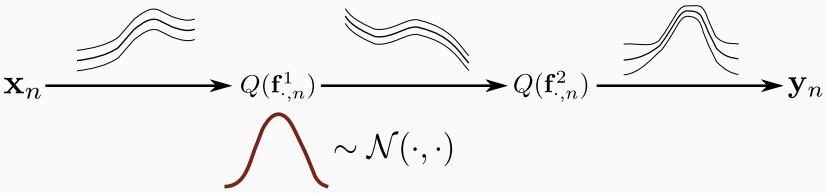
Fully connected DVIP: input propagated through L IP layers (each layer has H_l independent IP units).



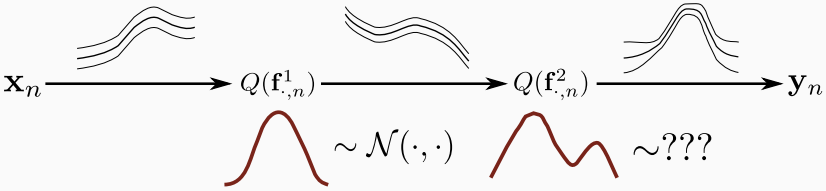
Stochasticity Propagation



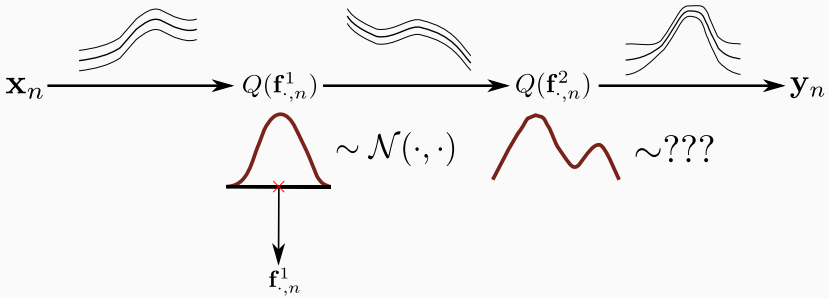
Stochasticity Propagation



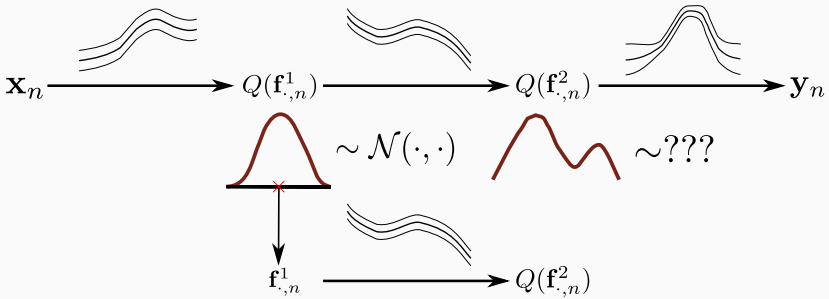
Stochasticity Propagation



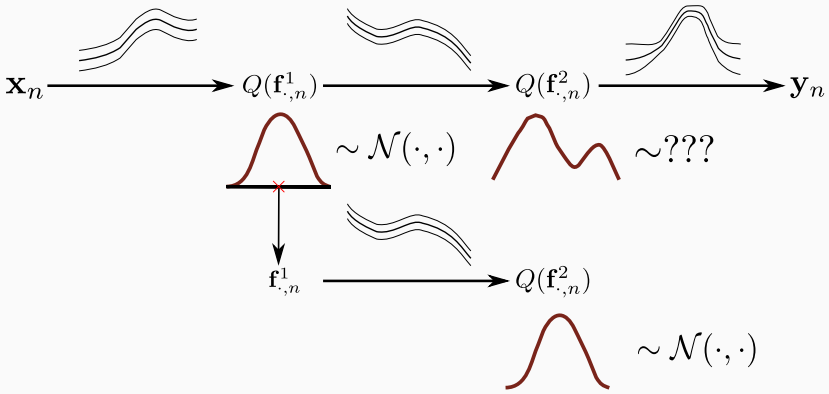
Stochasticity Propagation



Stochasticity Propagation



Stochasticity Propagation

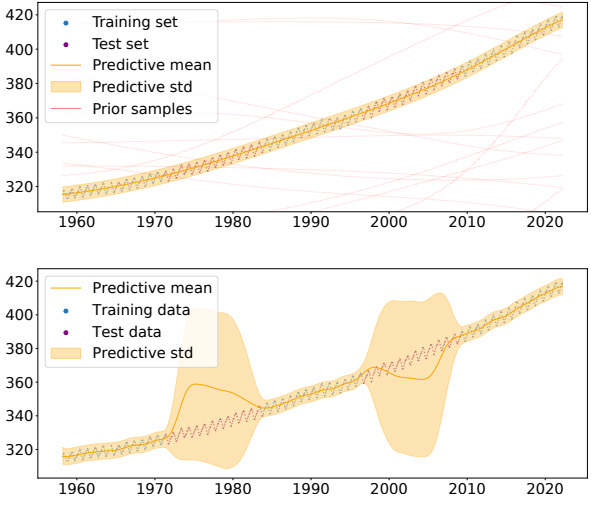


Inference and prediction in DVIP

- **Inference objective (function space)**
  - optimize a variational approximation to the posterior over functions/layer outputs
  - gradients estimated with Monte Carlo samples
- **Scalable optimization**
  - mini-batches over data points
  - stochastic estimates of the ELBO
- **Why predictions are non-Gaussian**
  - each layer receives *random* inputs (previous layer outputs)
  - composition of random mappings breaks Gaussianity
- **Prediction in practice**
  - sample functions per layer
  - forward-propagate samples through the network
  - use a Gaussian mixture as predictive distribution

DVIP Results Illustation

CO2 Prediction Dataset - DVIP Top and DGP Bottom



Post-hoc uncertainty for pretrained networks

- **Objective**
  - start from a strong pretrained deterministic DNN
  - add calibrated predictive uncertainty
- **Baseline: Linearized Laplace Approximation (LLA)**
  - strong performance in practice
  - scalability bottleneck: Jacobian / Hessian-related computations
- **This thesis: scalable function-space alternatives**
  - **VaLLA**: sparse variational GP surrogate of LLA (function space)
  - **FMGP**: fixed mean = pretrained predictor; learn covariance via VI

VaLLA: LLA as a Gaussian process

- **Network / MAP:** $f(\mathbf{x};\boldsymbol{\theta})$ is the network output at input $\mathbf{x}$, $\boldsymbol{\theta}_{\text{MAP}}$ are MAP parameters.
- **Local Gaussian posterior (Laplace):**

$$P(\boldsymbol{\theta}|D) \approx \mathcal{N}(\boldsymbol{\theta}_{\text{MAP}}, H^{-1}),$$

  where $H = \nabla^2_{\boldsymbol{\theta}}\big[-\log P(\boldsymbol{\theta}|D)\big]\big|_{\boldsymbol{\theta}_{\text{MAP}}}$ (often replaced by GGN Approximation).
- **Jacobian:**

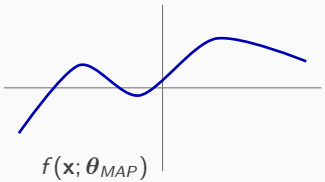
$$J_{\boldsymbol{\theta}_{\text{MAP}}}(\mathbf{x}) := \nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})\big|_{\boldsymbol{\theta}_{\text{MAP}}}.$$

- **Linearization induces function-space GP:**

$$f(\cdot) \sim \mathrm{GP}\big(m(x), k(x,x')\big), \quad m(\mathbf{x}) := f(\mathbf{x};\boldsymbol{\theta}_{\text{MAP}}), \quad k(\mathbf{x},\mathbf{x}') := J_{\boldsymbol{\theta}_{\text{MAP}}}(\mathbf{x}) H^{-1} J_{\boldsymbol{\theta}_{\text{MAP}}}(\mathbf{x}')^\top.$$

- **Dimensions (typical):** $J_{\boldsymbol{\theta}_{\text{MAP}}}(\mathbf{x}) \in \mathbb{R}^{d_{\text{out}} \times d_\theta}$, $H \in \mathbb{R}^{d_\theta \times d_\theta}$, so $k(\mathbf{x},\mathbf{x}') \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$.

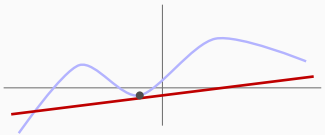
1) Pre-trained model



$f(\mathbf{x}; \boldsymbol{\theta}_{MAP})$

linearize

2) Linearization at $\boldsymbol{\theta}_{MAP}$



$f(\mathbf{x}; \boldsymbol{\theta}) \approx f(\mathbf{x}; \boldsymbol{\theta}_{MAP}) + J_{\boldsymbol{\theta}_{MAP}}(\mathbf{x})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$

decoupled sparse Gaussian process

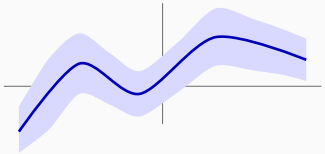
3) Variational inference

maximize (ELBO):

$$\mathcal{L} = \mathbb{E}_{Q(f)}[\log P(\mathcal{D}|f)] - \mathrm{KL}(Q(f)\|P(f))$$

use Q

4) Prediction (VaLLA)



$Q(f) = \mathcal{GP}(f(\mathbf{x}, \boldsymbol{\theta}_{MAP}), \boldsymbol{\Sigma}(\mathbf{x}))$

Function-space inference with Decoupled Sparse Gaussian Processes

- **Notation**
  - **Decoupled inducing sets:** $\mathbf{Z}_\alpha = \{\mathbf{z}_{\alpha,j}\}_{j=1}^{m_\alpha}$ (mean), $\quad \mathbf{Z}_\beta = \{\mathbf{z}_{\beta,j}\}_{j=1}^{m_\beta}$ (covariance)
  - GP prior: $P(f) = \mathcal{GP}(0, k)$
- **Decoupled sparse variational posterior in function space**

$$Q(f) \;=\; \mathcal{GP}\big(m_Q(\cdot),\, k_Q(\cdot,\cdot)\big),$$

with the standard decoupled parametrization

$$m_Q(\mathbf{x}) = K_{\mathbf{x},\mathbf{Z}_\alpha}\,\mathbf{a}, \qquad k_Q(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') + k_{\mathbf{x},\mathbf{Z}_\beta}\,K_{\mathbf{Z}_\beta,\mathbf{Z}_\beta}^{-1}\big(\mathbf{S} - K_{\mathbf{Z}_\beta,\mathbf{Z}_\beta}\big)K_{\mathbf{Z}_\beta,\mathbf{Z}_\beta}^{-1}\,k_{\mathbf{Z}_\beta,\mathbf{x}'},$$

where $\mathbf{a} \in \mathbb{R}^{m_\alpha \times d_{\text{out}}}$ controls the mean, and $\mathbf{S} \succeq 0$ controls the posterior covariance.

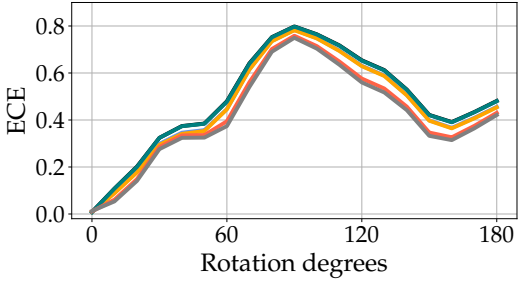
$$\mathbf{Z}_\alpha \text{ and } \mathbf{a} \text{ are fixed} \;\to\; m_Q(\mathbf{x}) \approx f(\mathbf{x}, \boldsymbol{\theta}_{MAP}).$$

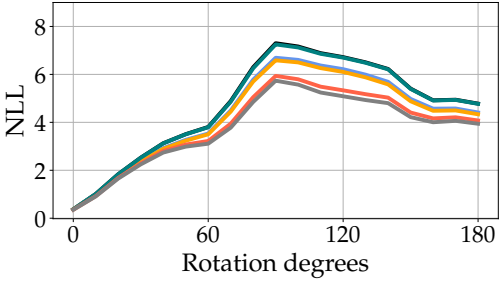
- **Function-space variational objective (ELBO)**

$$\log P(\mathcal{D}) \;\geq\; \mathcal{L}(Q) := \mathbb{E}_{Q(f)}\Big[\sum_{i=1}^{n} \log p\big(y_i \,|\, f(\mathbf{x}_i)\big)\Big] \;-\; \mathrm{KL}\big(Q(f)\,\|\,P(f)\big).$$

VaLLA Results on Fashion MNIST

Model	ACC	NLL	ECE	BRIER	OOD-AUC
MAP	86.6	0.373	**0.009**	0.193	0.874
LLA* KFAC	86.6	0.373	**0.008**	0.193	0.880
LLA*	86.6	0.373	**0.008**	0.193	0.882
ELLA	86.6	0.373	**0.008**	0.193	0.874
VaLLA 100	**87.4**	**0.335**	0.011	**0.182**	**0.923**
VaLLA 200	**87.6**	**0.332**	0.013	**0.181**	**0.933**





FMGP as Gaussian measures on a Hilbert space

- **Function prior as a Gaussian measure**
  - Work on a separable Hilbert space $\mathcal{H}$ of functions.
  - Gaussian measure on (a suitable completion of) $\mathcal{H}$:

$$P = \mathcal{N}(g, \mathcal{C}),$$

    where $g \in \mathcal{H}$ is a fixed mean and $\mathcal{C} : \mathcal{H} \to \mathcal{H}$ is a positive, self-adjoint covariance operator.
- **Infer only uncertainty (operator view)**
  - keep the mean fixed at g and update only the covariance operator
  - sparse/low-rank uncertainty updates via subspaces of $\mathcal{H}$
- **Practical benefit**
  - Bayesian uncertainty on top of a frozen network predictor (mean $f(\cdot, \boldsymbol{\theta}_{MAP})$)
  - avoids Jacobians / curvature of the DNN (inference happens in function space)

FMGP: Gaussian-measure posterior family and operator form

- **Posterior family as Gaussian measures**
  - Approximate posterior: $Q = \mathcal{N}(m^\star, \mathcal{C}^\star)$.
- **Mean constraint (Hilbert norm)**

$$\|m^\star - f(\cdot, \boldsymbol{\theta}_{MAP})\|_{\mathcal{H}} \leq \varepsilon \qquad \text{(FMGP enforces } m^\star \approx f(\cdot, \boldsymbol{\theta}_{MAP})\text{)}.$$

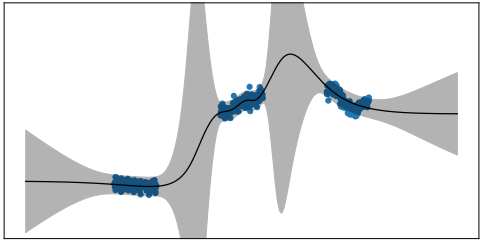
- **Sparse covariance update via a subspace**
  - Choose a finite-dimensional subspace $U = \operatorname{span}\{\phi_1, \ldots, \phi_{m_\beta}\} \subset \mathcal{H}$
  - Variational parameter: $\boldsymbol{A} \succeq 0$ acting on $\mathbb{R}^{m_\beta}$

$$K^\star_{\mathbf{Z}_\beta, \boldsymbol{A}}(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{Z}_\beta)\big(\boldsymbol{A} + K(\mathbf{Z}_\beta, \mathbf{Z}_\beta)\big)^{-1} K(\mathbf{Z}_\beta, \mathbf{x}').$$

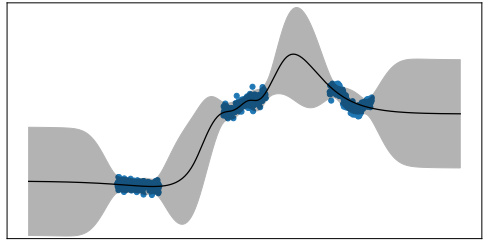
- **Learning:** optimize (VI) over $(\boldsymbol{A}, \mathbf{Z}_\beta)$ and kernel hyperparameters.

Predictive distibution of LLA and FMGP

LLA



FMGP



Part II: Generalization theory via diversity and large deviations

Ensemble diversity: core decomposition

For a distribution ρ over predictors $\boldsymbol{\Theta}$:

$$L(\rho) \leq \alpha\Big(\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \mathbb{D}(\rho)\Big),$$

where $\alpha = 1$ (squared or cross-entropy) and $\alpha = 4$ (0/1), and for squared loss the inequality is exact.

Diversity is captured by variance across ensemble members.

Diversity terms as expected variance

Examples (under data distribution ν):

$$\mathbb{D}_{\text{sq}}(\rho) := \mathbb{E}_{\nu}\left[\text{Var}_{\rho}(h_R(x;\boldsymbol{\theta}))\right],$$

$$\mathbb{D}_{ce}(\rho) := \mathbb{E}_{\nu}\left[\text{Var}_{\rho}\left(\frac{P(y|\mathbf{x},\boldsymbol{\theta})}{\sqrt{2}\max_{\boldsymbol{\theta}} P(y|\mathbf{x},\boldsymbol{\theta})}\right)\right],$$

$$\mathbb{D}_{0/1}(\rho) := \mathbb{E}_{\nu}\left[\text{Var}_{\rho}(\mathbf{1}(h_W(x;\boldsymbol{\theta}) \neq y))\right],$$

and, more generally,

$$\mathbb{D}(\rho) = \mathbb{E}_{\nu}\left[\text{Var}_{\rho}(f(y,x;\boldsymbol{\theta}))\right].$$

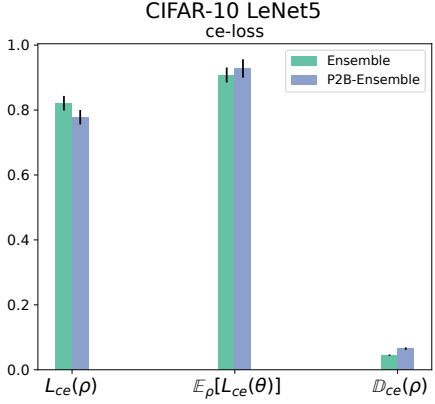
Key message: beneficial diversity reduces ensemble loss by reducing correlated errors.

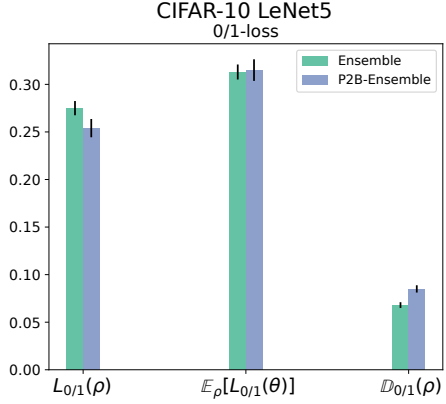
PAC–Bayes with diversity:

Theorem (PAC–Bayes bound). For any prior π over $\boldsymbol{\Theta}$ independent of D, any $\delta \in (0,1)$ and any $\lambda > 0$, with probability at least $1-\delta$ over draws $D \sim \nu^n$, *for all* posteriors ρ over $\boldsymbol{\Theta}$ (simultaneously),

$$L(\rho) \;\le\; \alpha\Big(\mathbb{E}_\rho\big[\hat{L}(\boldsymbol{\theta}, D)\big] \;-\; \hat{D}(\rho, D) \;+\; \frac{2\,\mathrm{KL}(\rho\,\|\,\pi) + \varepsilon}{\lambda n}\Big).$$

Interpretation: to generalize we want (i) low empirical risk, (ii) *high diversity* $\hat{D}$, and (iii) a posterior ρ *close to the prior* π (small KL).


CIFAR-10 LeNet5
ce-loss
Ensemble
P2B-Ensemble
1.0
0.8
0.6
0.4
0.2
0.0
$L_{ce}(\rho)$
$\mathbb{E}_\rho[L_{ce}(\theta)]$
$\mathbb{D}_{ce}(\rho)$


CIFAR-10 LeNet5
0/1-loss
Ensemble
P2B-Ensemble
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$L_{0/1}(\rho)$
$\mathbb{E}_\rho[L_{0/1}(\theta)]$
$\mathbb{D}_{0/1}(\rho)$

Rate function and inverse rate function

Define cumulant-generating function:

$$J_{\theta}(\lambda) = \log \mathbb{E}_{\nu} \left[\exp \left(\lambda (L(\theta) - \ell(y, \mathbf{x}, \theta)) \right) \right].$$

Rate function (Legendre transform):

$$\mathcal{I}_{\theta}(a) = \sup_{\lambda > 0} \left(\lambda a - J_{\theta}(\lambda) \right).$$

Inverse rate:

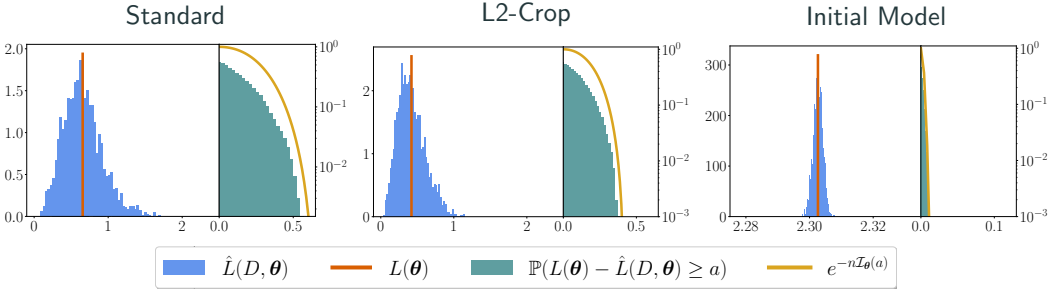
$$\mathcal{I}_{\theta}^{-1}(s) = \inf_{\lambda > 0} \frac{J_{\theta}(\lambda) + s}{\lambda}.$$

Chernoff bound

For fixed $\boldsymbol{\theta}$ and $a > 0$:

$$\mathbb{P}_{D\sim\nu^n}\Big(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta}) \geq a\Big) \leq \exp\big(-n\mathcal{I}_{\boldsymbol{\theta}}(a)\big).$$

Key message: if $\mathcal{I}_{\boldsymbol{\theta}}(\cdot)$ is large, large deviations are exponentially unlikely.

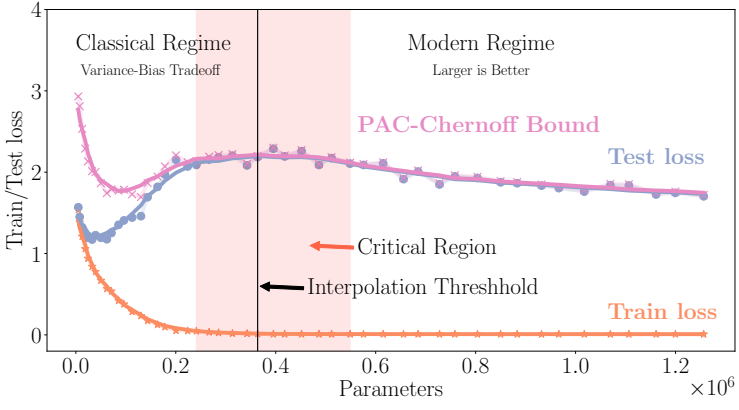


Distribution of empirical loss, tail probabilities and Chernoff bound for three Inception models on CIFAR-10 ($n = 50$).

PAC-Chernoff Bound

With h.p. $1-\delta$ over $D \sim \nu^n$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, simultaneously,

$$L(\boldsymbol{\theta}) \leq \hat{L}(D, \boldsymbol{\theta}) + \mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\tfrac{1}{n} \log \tfrac{k^p}{\delta}\right).$$



Optimal regularization at interpolation

Define (approximate) interpolators with $\hat{L}(D, \boldsymbol{\theta}) \leq \varepsilon$. Let:

$$\boldsymbol{\theta}_{\varepsilon}^{\star} = \underset{\hat{L}(D,\boldsymbol{\theta}) \leq \varepsilon}{\arg\min}\, L(\boldsymbol{\theta}), \qquad \boldsymbol{\theta}_{\varepsilon}^{\times} = \underset{\hat{L}(D,\boldsymbol{\theta}) \leq \varepsilon}{\arg\min} \left(\hat{L}(D, \boldsymbol{\theta}) + \mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\tfrac{1}{n} \log \tfrac{k^p}{\delta} \right) \right).$$

With high probability,

$$|L(\boldsymbol{\theta}_{\varepsilon}^{\star}) - L(\boldsymbol{\theta}_{\varepsilon}^{\times})| \leq \varepsilon.$$

Interpretation: minimizing the inverse rate function is an optimal regularization principle for interpolating regimes.

Understanding Existing Regularizers

Many common regularization techniques are **approximations** to the **optimal regularizer**:

- **Distance from initialization and ℓ_2-norm:**

$$\mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\tfrac{1}{n}\ln\tfrac{k^p}{\delta}\right) \leq \sqrt{2Ma}\,\|\boldsymbol{\theta}\|_2\,,$$

- **Input-gradient norm:**

$$\mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\tfrac{1}{n}\ln\tfrac{k^p}{\delta}\right) \leq \sqrt{\tfrac{1}{n}\ln\tfrac{k^p}{\delta}}\sqrt{M\mathbb{E}_{\nu}\left[\left\|\nabla_{\mathbf{x}}\ell(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})\right\|_2^2\right]}\,.$$

The distribution-dependent PAC-Chernoff Bound can be used to obtain **bounds over the number of parameters** of **interpolators**:

For any $\epsilon \in (0, L^\star)$ and any $\delta \in (0,1)$, with high probability $1-\delta$ over $D \sim \nu^n$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, simultaneously,

$$\text{if} \quad \hat{L}(D,\boldsymbol{\theta}) \leq \epsilon \quad \text{then} \quad p \geq \frac{n\mathcal{I}_{\boldsymbol{\theta}}(L^\star - \epsilon) + \ln\delta}{\ln k}\,.$$

where $L^\star = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$.

SGD implicit bias via large deviations

Defining the **abnormality** rate of a model as

$$\begin{aligned}\alpha(D,\boldsymbol{\theta}) &:= \mathcal{I}_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta}))\,,\\ &\asymp -\frac{1}{n}\ln \mathbb{P}_{S\sim\nu^n}\Big(L(\boldsymbol{\theta}) - \hat{L}(S,\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta})\Big)\end{aligned}$$

raises a decomposition on the empirical loss given by

$$\hat{L}(D,\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha(D,\boldsymbol{\theta}))\,.$$

The cumulative distribution of $\alpha(D,\boldsymbol{\theta})$ satisfies

$$\begin{aligned}\forall s>0 \quad & \mathbb{P}_{D\sim\nu^n}\big(\alpha(D,\boldsymbol{\theta}) \geq s\big) \leq e^{-n|s|}\,,\\ \forall s<0 \quad & \mathbb{P}_{D\sim\nu^n}\big(\alpha(D,\boldsymbol{\theta}) \leq s\big) \leq e^{-n|s|}\,,\end{aligned}$$

In conclusion, the distribution of empirical loss for *large* n values can be expressed as:

$$\hat{L}(D,\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(s), \quad s\sim \text{Laplace}(0,n)\,.$$

Implicit Bias of Gradient Descent (GD)

$$\hat{L}(D, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha(D, \boldsymbol{\theta}))$$

- **Two implicit incentives beyond lowering $L(\theta)$:**
  - Increase **abnormality** $\alpha(D, \boldsymbol{\theta})$ (move toward rare/left-tail realizations).
  - Increase $\mathcal{I}_{\boldsymbol{\theta}}^{-1}(\cdot)$ for a given α, i.e. implicitly favor **less concentrated** empirical-loss distributions (smaller $\mathcal{I}_{\boldsymbol{\theta}}$).
- **Consequence:** GD can drive $\hat{L}(D, \boldsymbol{\theta})$ down by exploiting unusually favorable deviations, often yielding **large generalization error**.

Implicit Bias of Stochastic Gradient Descent (SGD)

- **SGD objective (mini-batches B, size m):**

$$\min_{\boldsymbol{\theta}} \ \hat{L}(B, \boldsymbol{\theta}) \quad \text{and} \quad \mathbb{E}_{B \sim D}\big[\hat{L}(B, \boldsymbol{\theta})\big] = \hat{L}(D, \boldsymbol{\theta}).$$

Implicit Bias of Stochastic Gradient Descent (SGD)

- **SGD objective (mini-batches B, size m):**

$$\min_{\boldsymbol{\theta}} \ \hat{L}(B, \boldsymbol{\theta}) \quad \text{and} \quad \mathbb{E}_{B \sim D}\left[\hat{L}(B, \boldsymbol{\theta})\right] = \hat{L}(D, \boldsymbol{\theta}).$$

- **Batch-level decomposition:**

$$\hat{L}(B, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B}) \ \Rightarrow \ \hat{L}(D, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}_{B \sim D}\left[\mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B})\right],$$

so dynamics depend on the batch-induced distribution $Q(\alpha_{\boldsymbol{\theta}}^{B} | D, \boldsymbol{\theta})$.

Implicit Bias of Stochastic Gradient Descent (SGD)

- **SGD objective (mini-batches B, size m):**

$$\min_{\boldsymbol{\theta}} \ \hat{L}(B, \boldsymbol{\theta}) \quad \text{and} \quad \mathbb{E}_{B \sim D}\big[\hat{L}(B, \boldsymbol{\theta})\big] = \hat{L}(D, \boldsymbol{\theta}).$$

- **Batch-level decomposition:**

$$\hat{L}(B, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B}) \ \Rightarrow \ \hat{L}(D, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}_{B \sim D}\left[\mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B})\right],$$

  so dynamics depend on the batch-induced distribution $Q(\alpha_{\boldsymbol{\theta}}^{B} | D, \boldsymbol{\theta})$.
- **Why SGD differs from GD:** a Taylor expansion around $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ introduces an explicit **variance correction**:

Implicit Bias of Stochastic Gradient Descent (SGD)

- **SGD objective (mini-batches B, size m):**

$$\min_{\boldsymbol{\theta}} \ \hat{L}(B, \boldsymbol{\theta}) \quad \text{and} \quad \mathbb{E}_{B \sim D}\left[\hat{L}(B, \boldsymbol{\theta})\right] = \hat{L}(D, \boldsymbol{\theta}).$$

- **Batch-level decomposition:**

$$\hat{L}(B, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B}) \ \Rightarrow \ \hat{L}(D, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}_{B \sim D}\left[\mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B})\right],$$

  so dynamics depend on the batch-induced distribution $Q(\alpha_{\boldsymbol{\theta}}^{B} | D, \boldsymbol{\theta})$.

- **Why SGD differs from GD:** a Taylor expansion around $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ introduces an explicit **variance correction**:
  - larger $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ tends to **increase** the deviation effect,

Implicit Bias of Stochastic Gradient Descent (SGD)

- **SGD objective (mini-batches B, size m):**

$$\min_{\boldsymbol{\theta}} \ \hat{L}(B, \boldsymbol{\theta}) \quad \text{and} \quad \mathbb{E}_{B \sim D}\big[\hat{L}(B, \boldsymbol{\theta})\big] = \hat{L}(D, \boldsymbol{\theta}).$$

- **Batch-level decomposition:**

$$\hat{L}(B, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B}) \ \Rightarrow \ \hat{L}(D, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}_{B \sim D}\left[\mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B})\right],$$

  so dynamics depend on the batch-induced distribution $Q(\alpha_{\boldsymbol{\theta}}^{B} | D, \boldsymbol{\theta})$.

- **Why SGD differs from GD:** a Taylor expansion around $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ introduces an explicit **variance correction**:
  - larger $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ tends to **increase** the deviation effect,
  - larger $\mathrm{Var}(\alpha_{\boldsymbol{\theta}}^{B})$ tends to **decrease** it.

Implicit Bias of Stochastic Gradient Descent (SGD)

- **SGD objective (mini-batches B, size m):**

$$\min_{\boldsymbol{\theta}} \ \hat{L}(B, \boldsymbol{\theta}) \quad \text{and} \quad \mathbb{E}_{B \sim D}\big[\hat{L}(B, \boldsymbol{\theta})\big] = \hat{L}(D, \boldsymbol{\theta}).$$

- **Batch-level decomposition:**

$$\hat{L}(B, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B}) \ \Rightarrow \ \hat{L}(D, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}_{B \sim D}\left[\mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B})\right],$$

  so dynamics depend on the batch-induced distribution $Q(\alpha_{\boldsymbol{\theta}}^{B} | D, \boldsymbol{\theta})$.

- **Why SGD differs from GD:** a Taylor expansion around $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ introduces an explicit **variance correction**:
  - larger $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ tends to **increase** the deviation effect,
  - larger $\mathrm{Var}(\alpha_{\boldsymbol{\theta}}^{B})$ tends to **decrease** it.
- **Batch size:** when $m = n$, variance $\to 0$ and SGD reduces to GD; smaller m increases $\mathrm{Var}(\alpha_{\theta}^{B})$.

Implicit Bias of Stochastic Gradient Descent (SGD)

- **SGD objective (mini-batches B, size m):**

$$\min_{\boldsymbol{\theta}} \ \hat{L}(B, \boldsymbol{\theta}) \quad \text{and} \quad \mathbb{E}_{B \sim D}\big[\hat{L}(B, \boldsymbol{\theta})\big] = \hat{L}(D, \boldsymbol{\theta}).$$

- **Batch-level decomposition:**

$$\hat{L}(B, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B}) \ \Rightarrow \ \hat{L}(D, \boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}_{B \sim D}\left[\mathcal{I}_{\boldsymbol{\theta}}^{-1}(\alpha_{\boldsymbol{\theta}}^{B})\right],$$

  so dynamics depend on the batch-induced distribution $Q(\alpha_{\boldsymbol{\theta}}^{B} | D, \boldsymbol{\theta})$.
- **Why SGD differs from GD:** a Taylor expansion around $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ introduces an explicit **variance correction**:
  - larger $\mathbb{E}[\alpha_{\boldsymbol{\theta}}^{B}]$ tends to **increase** the deviation effect,
  - larger $\mathrm{Var}(\alpha_{\boldsymbol{\theta}}^{B})$ tends to **decrease** it.
- **Batch size:** when $m = n$, variance $\to 0$ and SGD reduces to GD; smaller m increases $\mathrm{Var}(\alpha_{\theta}^{B})$.
- **Resulting bias:** smaller batches weaken the push toward highly abnormal deviations and toward broad (high-variance) loss distributions, typically improving **generalization**.

Conclusions and Future Work

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.
- **Function-space Bayesian uncertainty (Part I):**

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.
- **Function-space Bayesian uncertainty (Part I):**
  - **DVIP:** deep implicit-process priors + variational inference in function space $\Rightarrow$ expressive *non-Gaussian* predictive uncertainty, competitive accuracy, and improved efficiency vs. deep GPs.

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.
- **Function-space Bayesian uncertainty (Part I):**
  - **DVIP:** deep implicit-process priors + variational inference in function space $\Rightarrow$ expressive *non-Gaussian* predictive uncertainty, competitive accuracy, and improved efficiency vs. deep GPs.
  - **Post-hoc Bayesianization: VaLLA** (function-space surrogate of linearized Laplace) and **FMGP** (freeze mean to pretrained network, learn covariance) $\Rightarrow$ calibrated uncertainty without performance loss.

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.
- **Function-space Bayesian uncertainty (Part I):**
  - **DVIP:** deep implicit-process priors + variational inference in function space $\Rightarrow$ expressive *non-Gaussian* predictive uncertainty, competitive accuracy, and improved efficiency vs. deep GPs.
  - **Post-hoc Bayesianization: VaLLA** (function-space surrogate of linearized Laplace) and **FMGP** (freeze mean to pretrained network, learn covariance) $\Rightarrow$ calibrated uncertainty without performance loss.
- **Distribution-dependent generalization theory (Part II):**

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.
- **Function-space Bayesian uncertainty (Part I):**
  - **DVIP:** deep implicit-process priors + variational inference in function space $\Rightarrow$ expressive *non-Gaussian* predictive uncertainty, competitive accuracy, and improved efficiency vs. deep GPs.
  - **Post-hoc Bayesianization: VaLLA** (function-space surrogate of linearized Laplace) and **FMGP** (freeze mean to pretrained network, learn covariance) $\Rightarrow$ calibrated uncertainty without performance loss.
- **Distribution-dependent generalization theory (Part II):**
  - **Diversity:** ensemble risk decomposes into average member error + a diversity (de-correlation) gain.

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.
- **Function-space Bayesian uncertainty (Part I):**
  - **DVIP:** deep implicit-process priors + variational inference in function space $\Rightarrow$ expressive *non-Gaussian* predictive uncertainty, competitive accuracy, and improved efficiency vs. deep GPs.
  - **Post-hoc Bayesianization: VaLLA** (function-space surrogate of linearized Laplace) and **FMGP** (freeze mean to pretrained network, learn covariance) $\Rightarrow$ calibrated uncertainty without performance loss.
- **Distribution-dependent generalization theory (Part II):**
  - **Diversity:** ensemble risk decomposes into average member error + a diversity (de-correlation) gain.
  - **PAC–Chernoff / rate-function view:** bounds remain informative at interpolation.

Conclusion: A Unified Probabilistic View of Deep Learning

- **Thesis goal:** make uncertainty estimation scalable *and* make generalization analysis informative in modern (over-parameterized / interpolating) regimes.
- **Function-space Bayesian uncertainty (Part I):**
  - **DVIP:** deep implicit-process priors + variational inference in function space $\Rightarrow$ expressive *non-Gaussian* predictive uncertainty, competitive accuracy, and improved efficiency vs. deep GPs.
  - **Post-hoc Bayesianization: VaLLA** (function-space surrogate of linearized Laplace) and **FMGP** (freeze mean to pretrained network, learn covariance) $\Rightarrow$ calibrated uncertainty without performance loss.
- **Distribution-dependent generalization theory (Part II):**
  - **Diversity:** ensemble risk decomposes into average member error + a diversity (de-correlation) gain.
  - **PAC–Chernoff / rate-function view:** bounds remain informative at interpolation.
  - **SGD:** mini-batch noise induces an implicit regularization bias toward more stable (better-concentrated) solutions.

Future Work

- **DVIP extensions**

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.
- **Post-hoc uncertainty (VaLLA / FMGP)**

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.
- **Post-hoc uncertainty (VaLLA / FMGP)**
  - VaLLA: relax the Gaussian linearized-posterior assumption for better calibration in highly non-linear regions;

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.
- **Post-hoc uncertainty (VaLLA / FMGP)**
  - VaLLA: relax the Gaussian linearized-posterior assumption for better calibration in highly non-linear regions;
  - FMGP: design task-structured kernels (images, spatio-temporal) and pursue efficient approximations to Jacobian-induced kernels.

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.
- **Post-hoc uncertainty (VaLLA / FMGP)**
  - VaLLA: relax the Gaussian linearized-posterior assumption for better calibration in highly non-linear regions;
  - FMGP: design task-structured kernels (images, spatio-temporal) and pursue efficient approximations to Jacobian-induced kernels.
- **Theory directions**

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.
- **Post-hoc uncertainty (VaLLA / FMGP)**
  - VaLLA: relax the Gaussian linearized-posterior assumption for better calibration in highly non-linear regions;
  - FMGP: design task-structured kernels (images, spatio-temporal) and pursue efficient approximations to Jacobian-induced kernels.
- **Theory directions**
  - relax i.i.d. assumptions in large-deviation analyses;

Future Work

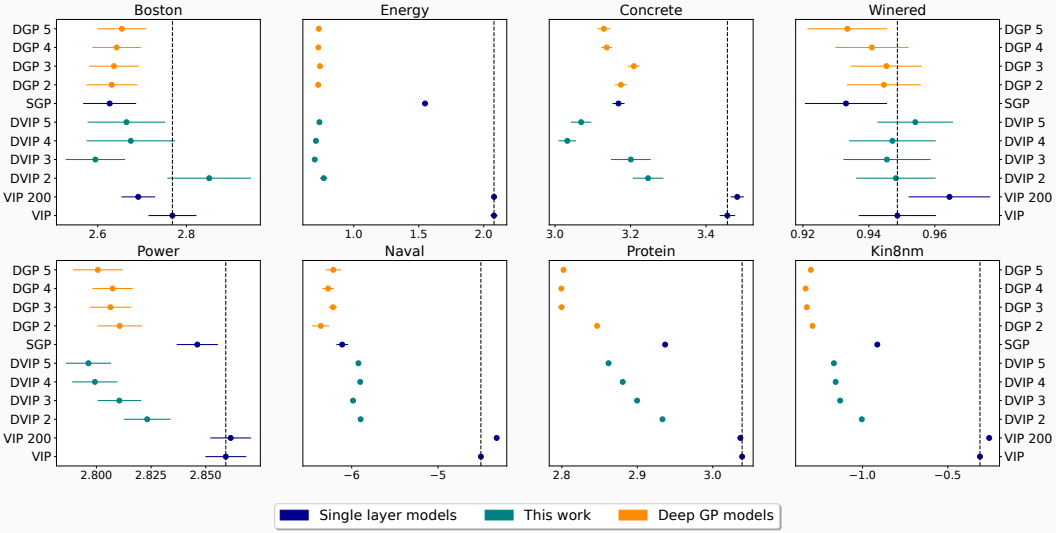
- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.
- **Post-hoc uncertainty (VaLLA / FMGP)**
  - VaLLA: relax the Gaussian linearized-posterior assumption for better calibration in highly non-linear regions;
  - FMGP: design task-structured kernels (images, spatio-temporal) and pursue efficient approximations to Jacobian-induced kernels.
- **Theory directions**
  - relax i.i.d. assumptions in large-deviation analyses;
  - connect PAC–Chernoff/rate-function analysis with information-theoretic bounds and algorithmic stability;

Future Work

- **DVIP extensions**
  - richer function-space posteriors beyond Gaussian/GP surrogates (e.g., flows or diffusion-based variational families);
  - stronger base approximations (e.g., sparse implicit processes) and *structured priors* (attention/recurrent for sequences, convolutional for spatial data);
  - integration into autoencoders / generative architectures to unify representation learning and uncertainty.
- **Post-hoc uncertainty (VaLLA / FMGP)**
  - VaLLA: relax the Gaussian linearized-posterior assumption for better calibration in highly non-linear regions;
  - FMGP: design task-structured kernels (images, spatio-temporal) and pursue efficient approximations to Jacobian-induced kernels.
- **Theory directions**
  - relax i.i.d. assumptions in large-deviation analyses;
  - connect PAC–Chernoff/rate-function analysis with information-theoretic bounds and algorithmic stability;
  - apply the rate-function formalism to transfer/continual learning and large language models.

Questions?

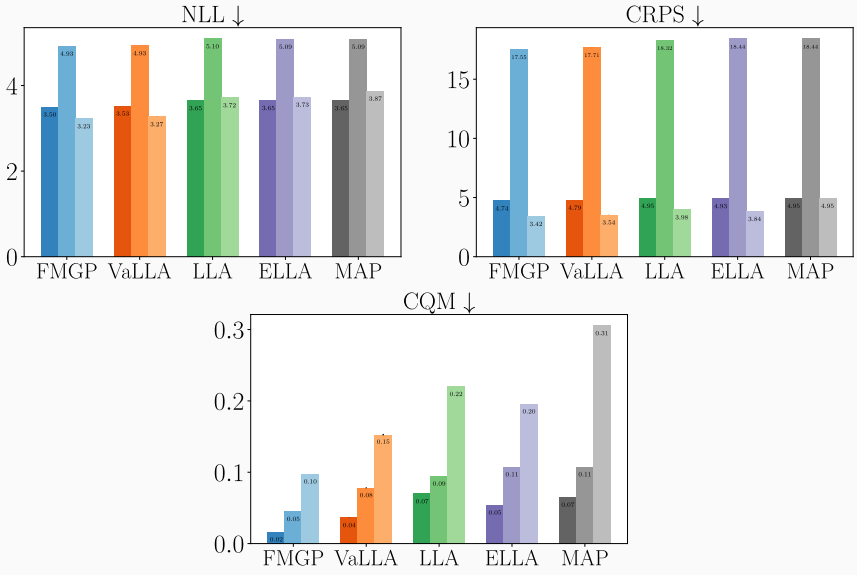
DVIP Regression Results



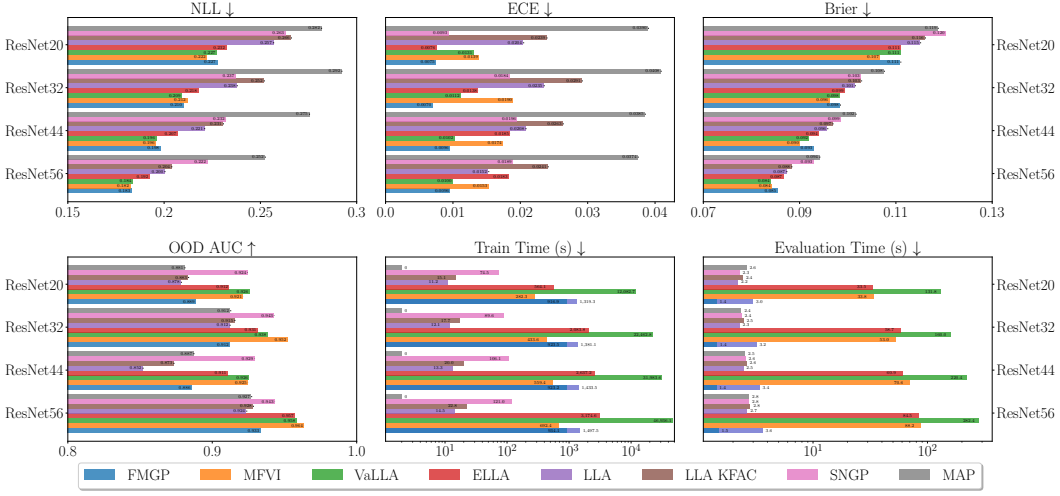
VaLLA Results

Method	ResNet-20			ResNet-32			ResNet-44			ResNet-56			Rank
	ACC	NLL	ECE	ACC	NLL	ECE	ACC	NLL	ECE	ACC	NLL	ECE	
MAP	**92.6**	0.282	0.039	**93.5**	0.292	0.041	**94.0**	0.275	0.039	**94.4**	0.252	0.037	–
MF-VI	**92.7**	**0.231**	0.016	**93.5**	0.222	0.020	**93.9**	0.206	0.018	**94.4**	0.188	0.016	–
SNGP	92.4	0.266	0.024	93.2	0.256	0.025	93.8	0.242	0.028	93.8	0.229	0.022	–
GP - Subset	**92.6**	0.555	0.299	**93.4**	0.462	0.247	93.6	0.424	0.225	**94.4**	0.403	0.221	–
LLA Diag	92.2	0.728	0.404	92.7	0.755	0.430	92.8	0.778	0.445	**92.9**	0.843	0.480	–
LLA KFAC	92.0	0.852	0.467	91.8	1.027	0.547	91.4	1.091	0.566	89.8	1.174	0.579	–
LLA*	**92.6**	0.269	0.034	**93.5**	0.259	0.033	**94.0**	0.237	0.028	**94.4**	0.213	0.022	–
LLA* KFAC	**92.6**	0.271	0.035	**93.5**	0.260	0.033	**94.0**	0.232	0.028	**94.4**	0.202	0.024	–
ELLA	92.5	0.233	0.009	**93.5**	**0.215**	**0.008**	**93.9**	0.204	**0.007**	**94.4**	0.187	**0.007**	2.37
Sampled LLA	92.5	**0.231**	**0.006**	**93.5**	0.217	**0.008**	**94.0**	**0.200**	**0.007**	**94.4**	**0.185**	0.015	**2.00**
VaLLA	**92.6**	**0.228**	**0.007**	**93.5**	**0.211**	**0.007**	**94.0**	**0.198**	**0.008**	**94.4**	**0.183**	**0.009**	**1.37**

FMGP Results



FMGP Results



Model	Method	NLL	ECE	Train Time	Test Time
ResNet18	MAP	**1.247±0.000**	0.026±0.000	**0.0±0.0**	**5.058±0.029**$\times 10^2$
	ELLA	1.248±0.000	**0.025**±0.000	**7.890±0.275**$\times 10^3$	8.060±0.010$\times 10^2$
	FMGP	1.248±0.001	**0.015±0.001**	1.835±0.099$\times 10^4$	**7.324±0.001**$\times 10^2$
	MFVI	**1.242±0.001**	0.040±0.000	7.602±0.032$\times 10^4$	3.773±0.308$\times 10^4$
ResNet34	MAP	**1.081±0.000**	0.035±0.000	**0.0±0.0**	**5.088±0.004**$\times 10^2$
	ELLA	1.082±0.000	**0.034±0.000**	**1.201±0.373**$\times 10^4$	1.087±0.018$\times 10^3$
	FMGP	**1.077±0.000**	**0.016±0.000**	1.942±0.103$\times 10^4$	**8.563± 0.011**$\times 10^2$
ResNet50	MAP	**0.962±0.000**	0.037±0.000	**0.0±0.0**	**4.954±0.010**$\times 10^2$
	ELLA	**0.962±0.000**	**0.036±0.000**	2.997±1.215$\times 10^4$	1.954±0.018$\times 10^3$
	FMGP	**0.958±0.001**	**0.018±0.001**	**2.543±0.046**$\times 10^4$	**1.100±0.010**$\times 10^3$
ResNet101	MAP	**0.912±0.000**	0.049±0.000	**0.0±0.0**	**5.059±0.001**$\times 10^2$
	ELLA	0.913±0.000	**0.048±0.000**	4.464±1.649$\times 10^4$	2.808±0.001$\times 10^3$
	FMGP	**0.900±0.000**	**0.030±0.001**	**2.654±0.064**$\times 10^4$	**1.134±0.001**$\times 10^3$
ResNet152	MAP	**0.876±0.000**	0.050±0.000	**0.0±0.0**	**6.324±0.004**$\times 10^2$
	ELLA	0.877±0.000	**0.048±0.000**	6.820±0.526$\times 10^4$	3.877±0.007$\times 10^3$
	FMGP	**0.865±0.001**	**0.024±0.001**	**2.973±0.069**$\times 10^4$	**1.267±0.002**$\times 10^3$