On the Effect of Regularization on Nonparametric Mean-Variance Regression

Eliot Wong-Toi[1]
ewongtoi@uci.edu

Alex Boyd[2]
alex.boyd@gehealthcare.com

Vincent Fortuin[3,4]
vincent.fortuin@tum.de

Stephan Mandt[1]
mandt@uci.edu

[1]University of California, Irvine
[2]GE Healthcare
[3]Technical University of Munich
[4]Helmholtz AI

Abstract

Uncertainty quantification is vital for decision-making and risk assessment in machine learning. Mean-variance regression models, which predict both a mean and residual noise for each data point, provide a simple approach to uncertainty quantification. However, overparameterized mean-variance models struggle with signal-to-noise ambiguity, deciding whether prediction targets should be attributed to signal (mean) or noise (variance). At one extreme, models fit all training targets perfectly with zero residual noise, while at the other, they provide constant, uninformative predictions and explain the targets as noise. We observe a sharp phase transition between these extremes, driven by model regularization. Empirical studies with varying regularization levels illustrate this transition, revealing substantial variability across repeated runs. To explain this behavior, we develop a statistical field theory framework, which captures the observed phase transition in alignment with experimental results. This analysis reduces the regularization hyperparameter search space from two dimensions to one, significantly lowering computational costs. Experiments on UCI datasets and the large-scale ClimSim dataset demonstrate robust calibration performance, effectively quantifying predictive uncertainty.

Keywords: mean-variance model, heteroskedastic regression, phase transition, overfitting, regularization

1 Introduction

Deep learning models are now routinely trained in massively overparameterized regimes, where the number of parameters far exceeds the number of data points [Belkin et al., 2019, Zhang et al., 2021]. In many applications, such models perform reliably despite their overcapacity: they interpolate, generalize, and exhibit robust behavior under mild regularization [Krizhevsky et al., 2012, Mathew et al., 2021]. However, there exists a growing class of models and objectives for which overparameterization interacts in delicate ways with the underlying likelihood, exposing structural instabilities that persist even when standard architectural heuristics such as batch normalization, dropout, or classical weight decay are applied. In these settings, the model is not failing to optimize its objective; rather, it is faithfully optimizing an objective that is intrinsically ill-conditioned [Nix and Weigend, 1994a, Seitzer et al., 2022].

Such behavior is not well captured by conventional statistical learning theory. Classical analyses often rely on bounded-capacity assumptions, convexity, Lipschitz continuity, or uniform generalization bounds [Bartlett and Mendelson, 2001, Mohri et al., 2012, Vapnik and Izmailov, 2020], none of which adequately describe the collective behavior that emerges when many coupled degrees of freedom co-adapt during training. Instead, the phenomena resemble those studied in statistical physics, where abrupt qualitative changes in solution structure, sensitivity to small perturbations in hyperparameters, and competing "forces" between different components of the model are common [Landau and Lifshitz, 2013, Altland and Simons, 2010]. These parallels suggest that tools from the study of interacting systems, particularly phase transitions and variational principles, provide a more suitable lens [Ringel et al., 2025].

An illustrative example of these ideas is overparameterized mean–variance regression, in which a model learns both a predictive mean and an input-dependent noise level [Nix and Weigend, 1994a,b]. In principle, this formulation should enable calibrated uncertainty quantification [Detlefsen et al., 2019, Fortuin et al., 2022]. In practice, it often displays striking instabilities [Seitzer et al., 2022]. On one extreme, the mean network can nearly interpolate the data, pushing residuals and predicted variances toward zero. On the other, even modest regularization of the mean can cause the model to flatten its predictions and attribute nearly all structure to the variance [Stirn et al., 2023, Immer et al., 2023]. These behaviors arise across a range of architectures and optimization settings, and they appear as sharp transitions between qualitatively distinct regimes. As we show later, these extremes correspond closely to limits in which either the mean or variance channel is effectively unregularized. In such cases the objective becomes unbounded.

To move beyond empirical observations and toward an explanatory framework, we analyze mean–variance regression from a field-theoretic perspective. In this view, the learned mean and log-precision functions are treated as smooth fields governed by a free-energy-like functional [Altland and Simons, 2010, Landau and Lifshitz, 2013]. This continuum formulation abstracts away architectural details while preserving the essential interactions between data fit and regularization. Taking variational derivatives yields coupled Euler–Lagrange equations. These partial differential equations describe the stationary configurations of the model and explicitly reveal how the likelihood and regularization act as competing forces that redistribute prediction error across the input domain.

This field-theoretic analysis reveals several key properties that are not evident from neural experiments alone. Most notably, we show that the unregularized maximum likelihood regime admits no finite stationary solution, providing a mathematical explanation for the extreme overconfidence observed when the mean network interpolates the data [Zhang et al., 2021]. We further show that one-sided regularization, in which only the mean or only the variance is penalized, renders the objective unbounded. This implies that two-sided regularization is not merely empirically helpful but structurally necessary. More broadly, the PDE structure predicts that solutions organize into qualitatively distinct regimes separated by sharp or smooth transitions, forming a *phase diagram* over the space of regularization strengths.

We validate these predictions experimentally using synthetic datasets, standard UCI benchmarks [Kelly et al.], and a large-scale climate simulation dataset [Yu et al., 2023]. Across all settings, we observe a consistent qualitative picture: regimes of variance collapse, mean collapse, underfitting plateaus, and an intermediate region where both fields adapt stably to the data. Importantly, the solutions obtained from the neural models align closely with those predicted by the field theory, even though the latter is solved in a continuum limit.

Finally, our analysis suggests a natural reparameterization of the regularization strengths in which the effective balance between likelihood and smoothing is captured by a one-dimensional quantity. This substantially simplifies hyperparameter tuning and provides a principled method for avoiding pathological regimes. The next section reviews prior work on mean-variance regression, statistical-

physics perspectives on overparameterized models, and Bayesian formulations that motivate our field-theoretic approach.

A preliminary version of this work appeared as a conference paper [Wong-Toi et al., 2024]. The present manuscript substantially extends that version by developing a full field-theoretic formulation, establishing new well-posedness results, and introducing a Bayesian extension.

This work makes the following contributions:

1. We develop a field-theoretic formulation of verparameterized mean–variance regression, treating learned predictors as continuous fields governed by a variational principle.

2. We derive the coupled Euler–Lagrange equations that characterize stationary solutions and illustrate how likelihood and regularization act as competing forces on the mean and precision fields.

3. We prove that the unregularized variational MLE functional has no finite minimizer, clarifying the mathematical source of the collapse behavior seen in neural mean–variance regression.

4. We show that one-sided regularization renders the objective unbounded, demonstrating that both the mean and variance must be jointly regularized for the learning problem to be well-posed.

5. We introduce a reparameterized regularization scheme that reduces the effective hyperparameter search to one dimension, yielding practical advantages for tuning and stability.

6. We characterize the resulting phase transition structure of the solution space, identifying the main qualitative regimes and describing the boundaries that separate them.

7. We introduce a Bayesian Field Theory (BFT) perspective by placing Gaussian process–like priors on the predictor fields, recovering the deterministic field theory as the posterior mode and providing a principled pathway for capturing epistemic uncertainty.

8. We show that the Bayesian Field Theory connects naturally to the classical statistics literature on penalized likelihood, Gaussian process priors, and spline-based smoothing, placing overparameterized neural mean–variance regression within a unified framework that bridges modern deep learning and traditional nonparametric methods.

9. We validate the theoretical predictions across neural implementations and numerical solutions of the field theory, observing strong agreement in both topology and transition structure.

The code for all neural network and field-theoretic experiments is available at `https://github.com/ewongtoi/deep-heteroskedastic-regression`.

2 Related Work

At a high level, our analysis connects several strands of prior work. Classical and early neural formulations of mean–variance regression identified degeneracies such as variance collapse and proposed architectural or weight-based regularization schemes to mitigate them [Nix and Weigend, 1994a, Bishop, 1994, Bishop and Quazaz, 1996, Hjorth and Nabney, 1999, Li and Chan, 2000]. Modern deep-learning approaches revisit these models in highly overparameterized settings, highlighting practical instabilities and exploring regularization grids, architectural constraints, and optimization strategies to stabilize maximum-likelihood training [Seitzer et al., 2022, Detlefsen et al., 2019, Stirn

et al., 2023, Sluijterman et al., 2024, Immer et al., 2023]. Bayesian perspectives introduce smoothness priors or approximate posterior inference to improve uncertainty calibration [Yau and Kohn, 2003, Yuan and Wahba, 2004, Lakshminarayanan et al., 2017, Stirn and Knowles, 2020, Wenzel et al., 2020, Izmailov et al., 2021]. Our contribution differs in focus: we study the fully overparameterized regime at the level of function space, using tools from statistical physics and variational calculus to characterize when and why these instabilities arise and how regularization induces distinct solution phases.

Uncertainty and Mean-Variance Regression. Uncertainty is commonly divided into epistemic (model) and aleatoric (data) components [Hüllermeier and Waegeman, 2021], with the latter decomposed into homoskedastic and heteroskedastic noise. Handling input-dependent noise has long been an active area in statistics [Huber, 1967, Eubank and Thomas, 1993, Le et al., 2005, Uyanto, 2022] and machine learning [Abdar et al., 2021], but remains comparatively uncommon in modern deep learning [Kendall and Gal, 2017, Fortuin et al., 2022] due to the training instabilities studied in this work. Modeling heteroskedasticity can be interpreted as reweighting data points by their predictive uncertainty, a principle shown to improve robustness under label noise and imbalance [Mandt et al., 2016, Wang et al., 2017, Khosla et al., 2022].

Connections to Statistical Physics. There is increasing interest in applying tools from statistical physics to machine learning, particularly for understanding generalization, loss landscapes, and phase transitions [Ringel et al., 2025]. Analogies to spin glasses and jamming phenomena reveal transitions between underfitting and overfitting and the emergence of multimodal energy surfaces [Franz and Parisi, 2016, Geiger et al., 2019, Ros et al., 2019]. Related work on symmetry breaking and continuous symmetries in optimization [Bamler and Mandt, 2018] highlights how physical principles can influence the structure of learning dynamics and the geometry of model families. Sharp generalization transitions have been documented in regression and classification, including interpolation thresholds and double descent [Belkin et al., 2019, Wu and Sahai, 2023, George et al., 2023, Veiga et al., 2023]. Similar transition-like behavior appears in deep generative models [Wang et al., 2025]. Statistical mechanics approaches to inference [Zdeborová and Krzakala, 2016, Antenucci et al., 2019] further illuminate regimes where inference becomes computationally challenging. These perspectives motivate the field-theoretic description developed in our work.

Classical and Early Neural MVR. Neural mean–variance regression dates to the 1990s, when Nix and Weigend [1994a] and Bishop [1994] introduced Gaussian models that learn both mean and variance, noting degeneracies such as variance collapse. Subsequent work proposed unbiased variance estimators [Bishop and Quazaz, 1996], Bayesian regularization via Gaussian weight priors [Hjorth and Nabney, 1999], and analyses of heteroskedastic GLMs exhibiting similar instabilities [Li and Chan, 2000]. These studies show that ill-posedness is not unique to deep networks but arises whenever the likelihood can be increased by attributing residual structure to noise.

Modern Deep MVR and Regularization. Recent work has revisited these issues in overparameterized deep networks. Seitzer et al. [2022] analyze gradient blow-up as variances approach zero and propose reweighting schemes. Detlefsen et al. [2019], Stirn et al. [2023] decouple or constrain the variance pathway, while Sluijterman et al. [2024] map regularization grids and show that both channels must be regularized. New work on covariance and correlation learning [Shukla et al., 2024, 2025] further expands the space of heteroskedastic models.

Bayesian Perspectives. Bayesian neural networks [MacKay, 1992, Neal, 1995, Blundell et al., 2015, Gal et al., 2016] and approximate inference methods [Blundell et al., 2015, Welling and Teh, 2011, MacKay, 1992, Wenzel et al., 2020, Izmailov et al., 2021] provide mechanisms for epistemic uncertainty. Bayesian formulations of MVR place priors on mean and variance functions [Yau and Kohn, 2003, Yuan and Wahba, 2004] or infer precision parameters variationally [Stirn and Knowles, 2020]. These works motivate the Bayesian Field Theory developed later, which places priors directly on the predictor functions and recovers the deterministic field theory as its posterior mode.

3 Pitfalls of Overparameterized Mean–Variance Regression

This section introduces heteroskedastic mean–variance regression, explains why the maximum-likelihood objective becomes ill-posed in overparameterized settings, and motivates the regularization schemes used throughout. We conclude by describing the qualitative *phases* that arise across the regularization space.

3.1 Heteroskedastic Mean–Variance Regression

We consider independent data points $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with covariates $x_i \in \mathcal{X} \subset \mathbb{R}^d$ drawn from $p(x)$ and responses $y_i \in \mathbb{R}$. Throughout, we adopt the heteroskedastic Gaussian model

$$p(y \mid x; \mu, \Lambda) = \mathcal{N}\big(y \,|\, \mu(x),\, \Lambda(x)^{-1}\big), \tag{1}$$

with mean function $\mu : \mathcal{X} \to \mathbb{R}$ and precision function $\Lambda : \mathcal{X} \to \mathbb{R}_{>0}$.

To develop the continuum (field-theoretic) formulation, we interpret the dataset as a single realization of the process described by (1). No smoothness is assumed of the true (μ, Λ); the analysis depends only on the observed realization $y(\cdot)$. For analytic convenience we assume $y \in H^1(\mathcal{X})$. Later, after introducing the predictor functions $(\hat{\mu}, \hat{\Lambda})$, we impose $(\hat{\mu}, \hat{\Lambda}) \in H^1(\mathcal{X})^2$ and $\hat{\Lambda}(x) \geq c > 0$ a.e.

3.2 Overparameterized Neural Regression Models

A common approach is to parameterize μ and Λ using neural networks, whose universal approximation properties make them suitable for modeling flexible mean and precision functions [Hornik, 1991]. Let $\hat{\mu}_\theta : \mathcal{X} \to \mathbb{R}$ and $\hat{\Lambda}_\phi : \mathcal{X} \to \mathbb{R}_{>0}$ denote overparameterized feed-forward networks with parameters θ and ϕ. We assume they do not share parameters, although shared encoders have been explored in prior work [e.g., Detlefsen et al., 2019, Seitzer et al., 2022, Stirn et al., 2023]. For each x_i, these networks produce the values $\hat{\mu}_\theta(x_i)$ and $\hat{\Lambda}_\phi(x_i)$ that define the likelihood in (1).

The population cross-entropy between $p(x, y)$ and the predictive distribution $\hat{p}(y \,|\, x)p(x)$ is

$$\ell(\theta, \phi) = H(p, \hat{p}) = -\mathbb{E}_{p(x,y)}[\log \hat{p}(y \,|\, x)]\,. \tag{2}$$

Because the dataset $\mathcal{D}$ is a single Monte Carlo draw from $p(x, y)$, empirical training corresponds to replacing this expectation with its sample average. Replacing the expectation with this Monte Carlo estimate yields

$$\ell_{MLE}(\theta, \phi) = \frac{1}{2N} \sum_{i=1}^{N} \Big[\hat{\Lambda}_\phi(x_i)\hat{r}_\theta(x_i)^2 - \log \hat{\Lambda}_\phi(x_i)\Big]. \tag{3}$$

3.3 Why Maximum Likelihood Fails

Although (3) is well defined for finite models, it becomes fundamentally ill-posed in the overparameterized setting [Nix and Weigend, 1994a, Bishop, 1994, Bishop and Quazaz, 1996, Seitzer et al., 2022]. The two terms in (3) place contradictory pressures on $\hat{\Lambda}_\phi$: the residual term favors $\hat{\Lambda}_\phi \to 0$, while the $-\log \hat{\Lambda}_\phi$ term favors $\hat{\Lambda}_\phi \to \infty$. This competition alone does not stabilize $\hat{\Lambda}_\phi$.

At the same time, any sufficiently expressive mean network can interpolate at least one data point, driving $\hat{r}_\theta(x_i) \to 0$. Once a residual vanishes, the first term in (3) no longer restricts $\hat{\Lambda}_\phi(x_i)$, allowing it to diverge and driving the objective unbounded below. A related degeneracy arises in finite-dimensional heteroskedastic generalized linear models [Li and Chan, 2000] whenever exact interpolation is possible. Across both settings, the root cause is the same: the heteroskedastic Gaussian likelihood provides no mechanism to prevent collapse or explosion in Λ, and overparameterization exposes this structural weakness fully. This perspective explains why architectural heuristics such as weight decay or shared encoders may alleviate but do not eliminate the degeneracy. To better understand the interaction between likelihood and regularization, we next introduce explicit regularization schemes.

3.4 Regularization in Mean–Variance Regression

A natural remedy is to add L_2 penalties to both networks, as suggested by Hjorth and Nabney [1999], who also interpret these penalties as Gaussian weight priors:

$$\ell_{\alpha,\beta}(\theta, \phi) := \ell_{MLE}(\theta, \phi) + \alpha\|\theta\|_2^2 + \beta\|\phi\|_2^2, \tag{4}$$

where $\alpha, \beta \in \mathbb{R}_{\geq 0}$. Regularizing θ prevents mean overfitting, while regularizing ϕ prevents precision collapse and controls the complexity of the predicted uncertainty. As $\alpha \to \infty$, the mean becomes constant; as $\beta \to \infty$, the model becomes homoskedastic. [1]

Although intuitive, (α, β) spans an unbounded domain, so we adopt the bounded reparameterization

$$\ell_{\rho,\gamma}(\theta, \phi) := \rho\, \ell(\theta, \phi) + \bar{\rho}\left[\gamma\, \|\theta\|_2^2 + \bar{\gamma}\, \|\phi\|_2^2\right], \tag{5}$$

with $\rho, \gamma \in (0, 1)$ and $\bar{\rho} = 1 - \rho$, $\bar{\gamma} = 1 - \gamma$. This parameterization is one-to-one with (α, β), and both formulations yield proportional gradients. Here ρ determines the trade-off between data fit and total smoothness, whereas γ specifies whether smoothness is allocated primarily to the mean network or the precision network. The limits $\gamma = 1$ and $\gamma = 0$ correspond to unregularized precision and unregularized mean, respectively, and $\rho = 1$ recovers pure MLE.

3.5 Qualitative Phase Behavior

Across the (ρ, γ) space, the learned mean and precision functions exhibit a small number of recurring qualitative behaviors. These behaviors appear consistently across datasets, architectures, and both neural and field-theoretic solvers, suggesting an underlying structural organization. It is therefore useful to view the (ρ, γ) plane as a *phase diagram*, with representative solutions shown in Fig. 1.

Underfitting regimes. When ρ is small, the solution is dominated by the regularization terms. In Region U_Λ, both functions remain close to constants, effectively ignoring the data. As ρ increases and regularization is allocated primarily to the mean (Region U_μ), the precision becomes more responsive, and the model explains variation through noise rather than signal.

[1] Assuming an unpenalized bias term in the final layer or standardized data.

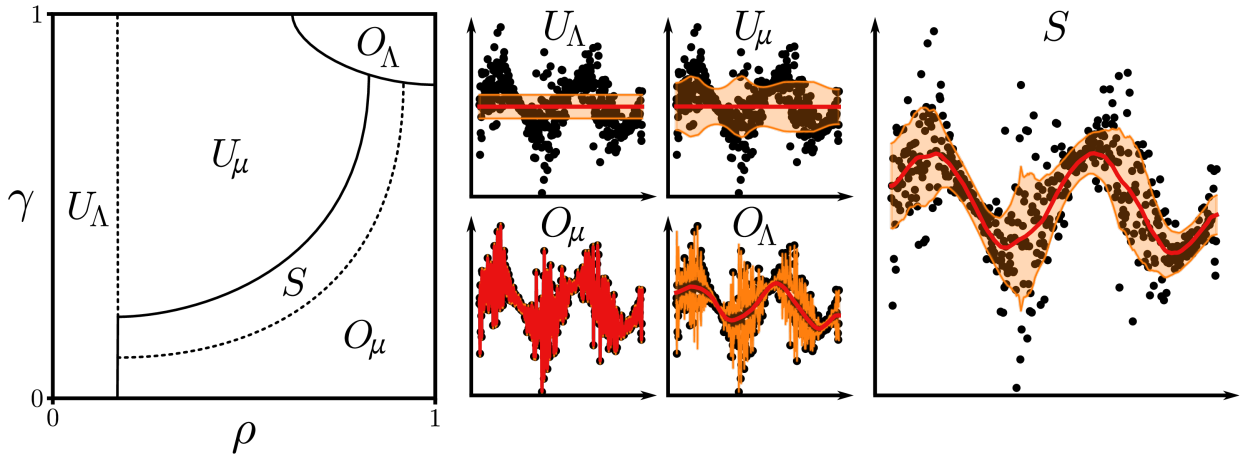


Figure 1: Phase-diagram sketch of mean–variance regression in the (ρ, γ) plane (*left*). Here ρ controls the data-fit vs. smoothness trade-off, and γ allocates smoothness between the mean and precision functions. Labeled regions indicate mean collapse (O_μ), variance collapse (O_Λ), underfitting (U_μ, U_Λ), and the stable regime S. Solid and dotted curves mark sharp and smooth transitions. Representative FT mean fits (red, with pointwise $\pm$ s.d. in orange) illustrate each regime (*middle*, *right*).

Overfitting regimes. Collapse occurs when one network is effectively unregularized. In Region O_μ, the mean nearly interpolates the data, leaving the precision largely irrelevant. In Region O_Λ, the precision adapts pointwise to fit residuals, producing spiky uncertainty estimates that mirror the data.

Stable regime. A comparatively narrow region S exists between these extremes, where neither learned network collapses. Here the mean captures the dominant structure and the precision reflects heteroskedastic variation rather than residual interpolation. This region typically yields the most stable and well-calibrated solutions.

As we show in Section 4, the field-theoretic formulation exhibits analogous qualitative phases and offers partial insight into the behavior of their boundaries in several limiting regimes.

4 Theoretical Considerations

We now develop a theoretical description of how regularization strengths shape the behavior of heteroskedastic regression models. This framework captures the limiting behavior of neural networks in the fully overparameterized regime and allows us to analyze edge cases of regularization settings. In particular, it yields necessary conditions that any optimal pair of mean and standard deviation functions must satisfy, independent of architectural details. Numerical solutions of the resulting *field theory*, introduced below, show strong qualitative agreement with practical neural network implementations.

Several of the structural transitions observed empirically in overparameterized networks resemble phase-transition phenomena analyzed in statistical physics. Prior work has documented analogous transitions in loss landscapes, jamming behavior, and generalization curves [Franz and Parisi, 2016, Geiger et al., 2019, Ros et al., 2019, Belkin et al., 2019]. Our field-theoretic formulation makes these parallels explicit by showing how competing likelihood and regularization terms give rise to sharp and smooth transitions in the learned predictors.

4.1 Field Theory

Having discussed the qualitative effects of regularization on a mean-variance model, we examine whether the extent to which these effects depend on a specific neural network architecture. Furthermore, we explore whether some of these effects can be characterized at the function level, independent of neural networks. To address these points, we develop *field theories* inspired by statistical mechanics.

Field theories are statistical descriptions of random functions, rather than discrete or continuous random variables [Altland and Simons, 2010]. A *field* is a function from spatial coordinates to scalar values (or vectors). Examples include electric charge density or a surface height. In field theory, we often seek the configuration (function) that minimizes an energy functional. Low-energy configurations of fields can display recurring patterns (e.g., waves) or undergo phase transitions (e.g., magnetism) upon varying model parameters. Since we can think of a function as an infinite-dimensional vector, field theory requires the usage of *functional analysis* over plain calculus. For example, we frequently ask for the field that minimizes a free energy functional that we obtain by calculating a functional derivative that we set to zero. The advantage of moving to a function-space description is that all details about neural architectures are abstracted away as long as the neural network is sufficiently over-parameterized.

Firstly, we abstract the neural networks $\hat{\mu}_\theta$ and $\hat{\Lambda}_\phi$ by smooth nonparametric functions $\hat{\mu}$ and $\hat{\Lambda}$ in the continuum limit; the exact Sobolev regularity conditions are specified in Appendix A. Though an "L_2" penalty does exist on a functional level (and could be applied to the FT), it penalizes the magnitude of the output of the function which is an inherently different aspect of the function. This would encourage predictions near zero rather than simple behaviors.

A comparable alternative is to directly penalize the output "complexity" of the models, measured via the *Dirichlet energies*, $\int p(x)\|\nabla\hat{\mu}(x)\|_2^2\,dx$ and $\int p(x)\|\nabla\hat{\Lambda}(x)\|_2^2\,dx$, corresponding to the respective mean and precision functions which we can weight via ρ, γ (or α, β as we did in the earlier neural network objective. A similar approach is taken by Yuan and Wahba [2004] where they apply separate roughness penalties to mean and noise functions. These quantities can be computed without any assumption on the particular parameterization of the functions. Note that these specific penalizations induce similar limiting behaviors for resulting solutions—$\rho \to 0$ implies functions that are quickly changing direction (overfitting) while $\rho \to \infty$ implies constant functions (underfitting). The discrete analogue to *Dirichlet energy* is the *geometric complexity (GC)*

$$GC(f, \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\|\nabla_x f(x)\|_F^2 \tag{6}$$

where $\|\nabla_x f(x)\|_F^2$ is the Frobenius norm of the network Jacobian. The geometric complexity of a neural network has been found to be lower in the presence of stronger L_2 penalties [Dherin et al., 2022] and in the case where $\hat{\mu}_\theta$ and $\hat{\Lambda}_\phi$ are linear models, this gradient penalty is equivalent to an L_2 penalty.

Using the assumptions outlined above and the same reparameterization of (α, β) to (ρ, γ) as in the neural network setting, the penalized cross-entropy can be interpreted as the action functional of a corresponding two-dimensional field theory (FT):

$$\mathcal{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}] = \int_{\mathcal{X}} p(x)\left\{\rho\int_{\mathcal{Y}} p(y\,|\,x)[-\log\hat{p}(y\,|\,x)]\,dy + \bar{\rho}\Big[\gamma\|\nabla\hat{\mu}(x)\|_2^2 + \bar{\gamma}\|\nabla\hat{\Lambda}(x)\|_2^2\Big]\right\}dx, \tag{7}$$

where $\hat{p}(y\,|\,x) = \mathcal{N}(y\,|\,\hat{\mu}(x), \hat{\Lambda}(x)^{-1})$. This formulation assumes continuous densities $p(x)$ and $p(y\,|\,x)$ and continuous predictor functions $\hat{\mu}(x)$ and $\hat{\Lambda}(x)$.

4.1.1 Empirical (sampled) field theory

Let $y(\cdot) = \{y(x)\}_{x\in\mathcal{X}}$ denote a realization of the stochastic process $y(x) \sim \mathcal{N}(\mu(x), \Lambda(x)^{-1})$, and assume that the realized data field satisfies $y \in H^1(\mathcal{X})$. In keeping with standard statistical and machine learning practice, all inference is performed *conditional on a single observed dataset*. Accordingly, we work with one realization $y(\cdot)$ rather than taking an expectation over multiple draws. This also avoids the computational burden of integrating over repeated noise realizations, yielding the *empirical* field-theoretic functional:

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}] = \int_{\mathcal{X}} p(x) \left\{ \rho\Big[\tfrac{1}{2}\hat{\Lambda}(x)\hat{r}^2(x) - \tfrac{1}{2}\log\hat{\Lambda}(x)\Big] + \bar{\rho}\Big[\gamma\|\nabla\hat{\mu}(x)\|_2^2 + \bar{\gamma}\|\nabla\hat{\Lambda}(x)\|_2^2\Big] \right\} dx, \tag{8}$$

where $\hat{r}(x) = \hat{\mu}(x) - y(x)$ is the pointwise residual field. By construction,

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}] \;\approx\; \mathcal{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}] + K, \qquad \text{where } K \text{ is independent of } (\hat{\mu}, \hat{\Lambda}). \tag{9}$$

The approximation arises solely from replacing the population expectation $\mathbb{E}_{p(y|x)}[\cdot]$ by its empirical counterpart under the observed realization $y(\cdot)$. This mirrors the finite-sample replacement used in the maximum-likelihood formulation: just as the empirical MLE objective approximates the population cross-entropy, the empirical functional $\mathcal{S}_{\rho,\gamma}$ approximates the population-level field-theoretic functional $\mathbf{S}_{\rho,\gamma}$.

Remark. The conditional Gaussian model $p(y \mid x) = \mathcal{N}(y \mid \mu(x), \Lambda(x)^{-1})$ is an *exact* specification of the data–generating process. Thus the field-theoretic functional itself remains exact under conditional normality; the symbol "≈" reflects only the finite-sample (Monte Carlo) replacement of the population expectation by the single observed dataset, which is both conceptually standard and computationally advantageous.

4.1.2 Function-space setting

We briefly record the regularity assumptions under which the functional $\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}]$ is well defined and its variational derivatives can be computed. Throughout, $\mathcal{X} \subset \mathbb{R}^d$ denotes a bounded Lipschitz domain and $p \in C^1(\overline{\mathcal{X}})$ is strictly positive on $\overline{\mathcal{X}}$, so that p is bounded above and below by positive constants.

We restrict the admissible fields to the usual Sobolev space $H^1(\mathcal{X})$, requiring square-integrability of both the functions and their weak gradients:

$$(\hat{\mu}, \hat{\Lambda}) \in H^1(\mathcal{X}) \times H^1(\mathcal{X}), \qquad \text{with} \qquad \hat{\Lambda}(x) \geq c > 0 \;\text{ a.e. in } \mathcal{X}.$$

The lower bound on the precision ensures non-degeneracy of the variance $\hat{\Lambda}^{-1}$ and, in particular, guarantees that the data-fidelity term $-\log\hat{\Lambda}$ remains finite. Under these assumptions, the weighted Dirichlet energies

$$\int_{\mathcal{X}} p(x)\,\|\nabla f(x)\|_2^2\,dx$$

are finite for all admissible $f \in H^1(\mathcal{X})$, since the positivity and essential boundedness of p imply the equivalence of the weighted and unweighted L^2 norms on $\mathcal{X}$.

From the perspective of ML and statistics, these conditions prevent degenerate behaviors such as vanishing or exploding variances and exclude functions with unbounded roughness. Consequently, the regularization term behaves as a well-posed smoothness penalty, while the likelihood contribution remains stable and well defined.

Taken together, these assumptions ensure that the functional $\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}]$ is finite for all admissible pairs $(\hat{\mu}, \hat{\Lambda})$ and that its associated Euler–Lagrange equations are meaningful in the usual weak

sense. Equivalently, the variational problem admits well-defined first-order optimality conditions that correspond to bona fide partial differential equations rather than ill-posed distributional identities.

4.2 Field-Theoretic Stationarity and Insights

Taking variational derivatives of the field-theory objective with respect to the mean and precision functions and imposing homogeneous Neumann boundary conditions yields the following *stationary conditions*, which hold almost everywhere with respect to the data density $p(x)$:

$$\rho\, p(x)\, \hat{\Lambda}^*(x)\, \hat{r}^*(x) = 2\bar{\rho}\, \gamma\, \mathcal{L}_p \hat{\mu}^*(x), \tag{10a}$$

$$\tfrac{\rho}{2}\, p(x) \Big[\hat{r}^*(x)^2 - \tfrac{1}{\hat{\Lambda}^*(x)} \Big] = 2\bar{\rho}\, \bar{\gamma}\, \mathcal{L}_p \hat{\Lambda}^*(x), \tag{10b}$$

where $\hat{r}^*(x) = \hat{\mu}^*(x) - y(x)$ and $\mathcal{L}_p f := -\nabla\cdot\big(p(x)\nabla f\big)$ denotes the (unnormalized) weighted Laplacian introduced in Lemma 1. When $p(x)$ is constant, $\mathcal{L}_p$ reduces to the standard Laplacian Δ [Engel and Dreizler, 2011], recovering the uniform-density setting studied in earlier work [Wong-Toi et al., 2024]. These equations coincide with the Euler–Lagrange conditions formalized in Prop. 2 of Appendix A. Below we summarize the theoretical implications obtained by taking limits over (ρ, γ).

Proposition 1 (Combined analytical structure of MVR)**.** *Assume that $\mathcal{X} \subset \mathbb{R}^d$ is a bounded, connected Lipschitz domain, that $p \in C^1(\overline{\mathcal{X}})$ is strictly positive on $\overline{\mathcal{X}}$, and that the data field satisfies $y \in H^1(\mathcal{X})$. Let $(\hat{\mu}, \hat{\Lambda}) \in H^1(\mathcal{X}) \times H^1(\mathcal{X})$ with $\hat{\Lambda}(x) \geq c > 0$ for a.e. $x \in \mathcal{X}$.*

Then any stationary point of the functional $\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}]$ satisfies the Euler–Lagrange system (10) under homogeneous Neumann (zero-flux) boundary conditions $p\nabla\hat{\mu}\cdot\mathbf{n} = 0$ and $p\nabla\hat{\Lambda}\cdot\mathbf{n} = 0$ on $\partial\mathcal{X}$. Furthermore:

- *(i) For $\rho = 1$ (no regularization), the EL equations admit no stationary solution.*
- *(ii) For $\rho = 0$ (no data term), the minimizer is non-unique: any constant pair $(\hat{\mu}, \hat{\Lambda})$ satisfies the Neumann equations on a connected domain.*
- *(iii) For $\gamma = 0$ or $\gamma = 1$ and $\rho > 0$, the functional $\mathbf{S}_{\rho,\gamma}$ is unbounded below; hence both regularization channels must be strictly positive for well-posedness.*
- *(iv) If in addition the precision is uniformly bounded $0 < c \leq \hat{\Lambda}(x) \leq \lambda_{\max} < \infty$ a.e. in $\mathcal{X}$, then for all $\rho, \gamma \in (0, 1)$ the variational problem admits at least one solution $(\hat{\mu}^*, \hat{\Lambda}^*) \in H^1(\mathcal{X})^2$; see Remark 3 for discussion of these additional assumptions.*

Hence, well-posed formulations require $\rho \in (0, 1)$ and $\gamma \in (0, 1)$, or equivalently two-sided positive penalties $\alpha, \beta > 0$.

The proofs, based on weighted Green's identities and weak Euler–Lagrange arguments, are given in Appendix A (see Props. 2 to 4 and Cor. 1).

4.2.1 Interpretation of Prop. 1

Under the assumptions of Prop. 1—namely, smooth positive $p(x)$ on a compact domain $\mathcal{X} \subset \mathbb{R}^d$ with homogeneous zero-flux boundaries—any stationary point $(\hat{\mu}^*, \hat{\Lambda}^*)$ of the field-theoretic objective satisfies the coupled PDEs in (10). Moreover, no solution with finite objective value exists in the limiting case $\rho \to 1$, consistent with the unregularized regime analyzed in Prop. 3.

Each equation in (10) expresses a balance between a data-fitting term (on the left) and a regularization-induced flux (on the right). The divergence operator plays a diffusion-like role,

redistributing prediction errors across the input domain rather than letting them concentrate at isolated points. The presence of $p(x)$ in the denominators modulates this diffusion according to the local data density: regions with higher $p(x)$ experience a weaker effective regularization, while low-density regions experience a stronger one. Equivalently, the local smoothing strength is proportional to

$$\text{effective regularization strength} \;\propto\; \frac{\bar{\rho}}{\rho}\,\gamma\,p(x)^{-1}.$$

Areas containing many data points therefore permit more functional complexity, whereas sparsely sampled regions are forced toward smoother, simpler predictions. This adaptive weighting makes the field theory sensitive to the empirical geometry of the data—a property absent from standard, unweighted L_2 regularization.

The weighted Laplacians $\mathcal{L}_p\hat{\mu}$ and $\mathcal{L}_p\hat{\Lambda}$ quantify the local curvature of the mean and precision functions under the data density $p(x)$. Accordingly, the hyperparameters ρ and γ directly determine how much curvature the model can sustain. Recall that $\rho \in (0,1)$ controls the overall balance between likelihood fitting and regularization (larger ρ emphasizes data fidelity, smaller ρ enforces stronger smoothing), while $\gamma \in (0,1)$ partitions the total regularization between the mean and precision fields (γ weighting the mean term and $\bar{\gamma} := 1-\gamma$ weighting the precision term). Large ρ (weak regularization) sharpens curvature and risks overconfident solutions, whereas small ρ (strong regularization) flattens both functions, leading to underfitting. Similarly, varying γ trades off smoothness between the mean and variance functions. These coupled PDEs therefore formalize the intuitive "phase diagram" of Figure 1, in which distinct regimes of (ρ,γ) correspond to qualitatively different equilibrium configurations of the fields.

The derivation of (10) and the associated boundary conditions follows from the weighted Green's identity in Lemma 1 and the Euler–Lagrange system proved in Prop. 2. Extreme and degenerate cases are analyzed in Prop. 3, which demonstrate the absence or unboundedness of solutions when either ρ or γ lies at the boundary of $[0,1]$. Finally, Cor. 2 shows that both mean and variance regularization terms must be strictly positive ($\gamma \in (0,1)$) to ensure well-posedness. Together, these results provide the analytic basis for the phase-transition structure schematized in Figure 1.

These limiting cases align with the intuition conveyed earlier and apply equally in the neural network setting. Assuming valid stationary solutions exist for $\rho,\gamma \in (0,1)$, one expects the system to exhibit either sharp transitions or smooth cross-overs between the behaviors described in the extremes as the regularization strengths vary. Empirically, Section 6 shows that the resulting phase diagrams resemble Figure 1. A complete analytical justification for the boundary types, their shapes, and precise placement within the (ρ,γ) plane is left for future work.

While the field theory is deterministic, its nonconvex objective can admit multiple distinct minimizers. This multiplicity reflects structural ambiguity in the learned predictor functions and provides a natural interpretation of *epistemic uncertainty*—not as stochasticity in the model itself, but as sensitivity to initialization and optimization. In particular, the existence results in Props. 2 and 4 do not guarantee uniqueness of minimizers, so different runs may converge to qualitatively distinct solutions. This perspective motivates the ensemble-based approximation introduced in Section 5.2, in which variability across local minimizers approximates the posterior spread of the Bayesian field theory (Section 5).

4.3 Numerically Solving the FT

We describe the numerical procedure used to approximate minimizers of the deterministic field theory (FT) objective introduced in Section 4.2. Because the Euler–Lagrange equations in (10) rarely admit

closed-form solutions, we work with a discrete approximation of the continuous functional $\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}]$. In practice we restrict attention to one-dimensional domains $\mathcal{X} \subset \mathbb{R}$.

4.3.1 Uniform lattice discretization

Let $\mathsf{L}^{(D)} = \{x_i^{(D)}\}_{i=1}^D$ denote a uniform grid on $\mathcal{X}$ with spacing $h \approx |\mathcal{X}|/D$. For boundary points, we pad one extra point on each side so that centered finite differences can be used at all interior indices; one-sided differences at the padded nodes enforce homogeneous Neumann boundary conditions. We define the discrete vectors

$$\vec{\mu}^{(D)} = (\hat{\mu}(x_i^{(D)}))_{i=1}^D, \qquad \vec{\Lambda}^{(D)} = (\hat{\Lambda}(x_i^{(D)}))_{i=1}^D, \qquad \vec{y}^{(D)} = (y(x_i^{(D)}))_{i=1}^D,$$

and let ∇_h be the standard centered finite-difference gradient. The discrete FT objective is

$$\mathsf{S}_{\rho,\gamma}^{(D)}(\vec{\mu}, \vec{\Lambda}) = \frac{1}{D} \sum_{i=1}^{D} \left\{ \rho \left[\frac{1}{2} \vec{\Lambda}_i \, (y_i - \vec{\mu}_i)^2 - \frac{1}{2} \log \vec{\Lambda}_i \right] + \bar{\rho} \left[\gamma \, \|\nabla_h \vec{\mu}\|_i^2 + \bar{\gamma} \, \|\nabla_h \vec{\Lambda}\|_i^2 \right] \right\}. \tag{11}$$

We minimize (11) via gradient descent to obtain lattice approximations of $\hat{\mu}$ and $\hat{\Lambda}$.

4.3.2 Consistency with the continuous FT

Because $\mathsf{L}^{(D)}$ is a uniform grid rather than a random sample from $p(x)$, convergence of the discrete objective to $\mathbf{S}_{\rho,\gamma}$ follows from deterministic quadrature rather than Monte–Carlo averaging. If $\hat{\mu}, \hat{\Lambda} \in H^1(\mathcal{X})$ and p is continuous and bounded above and below, then Riemann sum approximations yield

$$\lim_{D \to \infty} \mathsf{S}_{\rho,\gamma}^{(D)}\big(\Pi_D \hat{\mu}, \Pi_D \hat{\Lambda}\big) = \mathbf{S}_{\rho,\gamma}\big[\hat{\mu}, \hat{\Lambda}\big],$$

where Π_D denotes projection of the continuous fields onto the grid. Convergence of the discrete gradient terms follows from standard finite-difference consistency on uniform meshes [Fornberg, 1988, Brenner and Scott, 2008]. Violations of the regularity assumptions (e.g. if $\nabla \hat{\Lambda}$ is unbounded on a set of nonzero measure) may lead to instability or divergence of the discrete minimizer.

4.3.3 Relation to the general finite-element setting

The uniform grid considered here represents the one-dimensional analogue of the general mesh $\mathcal{G}_h$ in Appendix Section B.2. In higher dimensions, $\mathcal{G}_h$ is a shape-regular collection of elements supporting finite-element approximations of the weighted Laplacian $\mathcal{L}_p f = -\nabla \cdot (p \nabla f)$. Uniform grids in 1D can be viewed as a special case of such meshes with p incorporated solely through quadrature weights. Thus, standard FEM convergence theory applies directly [Brenner and Scott, 2008].

5 Bayesian Reformulation of the Field Theory

The deterministic field theory (FT) introduced in Section 4 can be extended to a fully probabilistic formulation by placing priors directly on the mean and log-precision fields $(\hat{\mu}, \hat{\Lambda})$. This yields a *Bayesian Field Theory* (BFT) in which the FT energy functional appears as a log-posterior, the deterministic solution arises as a maximum a posteriori (MAP) estimate, and posterior samples correspond to stochastic field realizations that quantify uncertainty.

This function-space viewpoint is closely related to classical Bayesian treatments of heteroskedastic regression. Prior statistical work places smoothness priors on spline-based mean and variance

functions [Yau and Kohn, 2003, Yuan and Wahba, 2004], while Lemm [2000, Section 3.7.1] analyzes homoskedastic Gaussian regression as a field theory with Gaussian priors defined by differential operators. The BFT developed here provides a continuous analogue adapted to the heteroskedastic setting and offers a principled route to uncertainty quantification that complements weight-space Bayesian neural network approaches.

We write the negative log-posterior (up to a constant) as

$$\Phi(\hat{\mu};\hat{\Lambda}) = \underbrace{-\log \pi(\hat{\mu},\hat{\Lambda})}_{\text{Prior}} \; \underbrace{- \; \rho \int_{\mathcal{X}} p(x) \log \hat{p}(y\,|\,x)\, dx}_{\text{Likelihood term}}, \qquad \hat{p}(y\,|\,x) = \mathcal{N}\big(y\,|\,\hat{\mu}(x), \hat{\Lambda}(x)^{-1}\big) \tag{12}$$

where $\pi(\hat{\mu},\hat{\Lambda})$ encodes smoothness priors over the predictor functions. Here ρ acts as a relative weighting (or inverse temperature) on the likelihood term, and the overall scaling between likelihood and prior is arbitrary up to a constant factor. All integrals are taken over the input domain $\mathcal{X}$ with respect to the normalized density $p(x)$ and can thus be interpreted as expectations under the input distribution.

Here $\hat{\mu}(x)$ denotes the predictive mean function and $\hat{\Lambda}(x) > 0$ the predictive *precision* (i.e., inverse-variance). We also define $\hat{\eta}(x) = \log \hat{\Lambda}(x)$ as the *log-precision* field, which is used in the subsequent parameterizations and enforces positivity after exponentiation. This notation ensures that the Gaussian likelihood term $\frac{1}{2}\hat{\Lambda}(y-\hat{\mu})^2 - \frac{1}{2}\log\hat{\Lambda}$ matches the standard negative log-likelihood of $\mathcal{N}(y\,|\,\hat{\mu},\hat{\Lambda}^{-1})$ up to an additive constant.

5.1 MAP-Equivalent Prior

As before, we write $\bar{\rho} = 1-\rho$ and $\bar{\gamma} = 1-\gamma$. To recover the deterministic FT exactly, we parameterize $\hat{\Lambda} = e^{\hat{\eta}}$, where $\hat{\eta}$ is the *model* log-precision function, and choose priors that yield the same gradient penalties as the FT functional.

The prior distributions are chosen to impose smoothness on the predictor functions, mirroring the geometry of the deterministic FT regularizers. We place a Gaussian field prior on the mean function to enforce smoothness and a log–convex prior on the log–precision to ensure positivity and stability of the noise function:

$$-\log \pi(\hat{\mu}) = \tfrac{\gamma}{2} \int_{\mathcal{X}} p(x)\, \|\nabla \hat{\mu}(x)\|_2^2\, dx, \tag{13a}$$

$$-\log \pi(\hat{\eta}) = \tfrac{\bar{\gamma}}{2} \int_{\mathcal{X}} p(x)\, e^{2\hat{\eta}(x)} \|\nabla \hat{\eta}(x)\|_2^2\, dx, \tag{13b}$$

where $\hat{\eta} := \log \hat{\Lambda}$ ensures $\hat{\Lambda} > 0$.

Since $\nabla\hat{\Lambda} = \nabla(e^{\hat{\eta}}) = e^{\hat{\eta}}\nabla\hat{\eta}$, we have $\|\nabla\hat{\Lambda}\|^2 = e^{2\hat{\eta}}\|\nabla\hat{\eta}\|^2$, so the penalty in (13b) is exactly the FT Dirichlet energy $\frac{\bar{\gamma}}{2}\int p\,\|\nabla\hat{\Lambda}\|_2^2\,dx$ expressed in the log–precision parameterization. (We adopt the same homogeneous weighted Neumann boundaries as in the FT derivation and omit them here for brevity; see Cor. 1 for details.)

Combining these priors with the likelihood yields the full posterior energy functional, whose stationary conditions coincide with those of the deterministic FT:

$$\Phi_{\text{MAP}}(\hat{\mu},\hat{\eta}) = \int_{\mathcal{X}} p \Big\{ \rho \big[\tfrac{1}{2} e^{\hat{\eta}}(y-\hat{\mu})^2 - \tfrac{1}{2}\hat{\eta}\big] + \tfrac{\bar{\rho}}{2}\big[\gamma \|\nabla\hat{\mu}\|^2 + \bar{\gamma}\, e^{2\hat{\eta}} \|\nabla\hat{\eta}\|^2\big] \Big\}. \tag{14}$$

For brevity, we omit explicit dependence on x where unambiguous. This functional coincides with $\mathbf{S}_{\rho,\gamma}$ up to constants, so minimizing Φ_{MAP} is precisely the FT optimization problem. The posterior

mode thus satisfies the same stationary equations as Prop. 2, confirming the equivalence MAP ≡ FT. (Up to the conventional $\frac{1}{2}$ scaling in Gaussian log–densities; see Appendix B for discussion.)

Our focus in this section is primarily analytical rather than computational: we study the structure of the Bayesian field theory (BFT) and its direct relationship to the deterministic FT, rather than performing full posterior sampling. Nevertheless, the Bayesian formulation clarifies how uncertainty arises in function space and provides a principled foundation for later ensemble-based approximations (see Section 5.2).

5.1.1 Connection to Gaussian processes

Before turning to computational approximations, it is helpful to relate the Bayesian field theory (BFT) to Gaussian processes (GPs), since both describe random functions and both encode smoothness through quadratic penalties.

A GP specifies its behavior directly through a covariance kernel $k(x, x')$. In contrast, a Gaussian field encodes smoothness implicitly through a linear differential operator $\mathcal{L}$ that penalizes roughness, much like a classical spline penalty in reproducing-kernel Hilbert spaces [Kimeldorf and Wahba, 1970, Wahba, 1990]. The covariance structure of the field is then determined by the inverse of this operator—specifically, by the Green's function solving $\mathcal{L}G = \delta$. This operator–kernel relationship, which plays a central role in modern SPDE-based Gaussian field models [Lindgren et al., 2011], provides the bridge between the field-theoretic and GP viewpoints. A formal derivation is given in Appendix B.3.

In our setting, the relevant operator is the weighted Laplacian $\mathcal{L}_p f = -\nabla \cdot (p \nabla f)$. With homogeneous Neumann boundary conditions, this operator leaves constant functions unchanged, and the corresponding prior is therefore "intrinsic" (improper) up to an additive constant. This is a standard phenomenon in intrinsic Gaussian Markov random fields [Rue and Held, 2005, Rue et al., 2009] and does not cause practical difficulty: the likelihood (or centering) fixes the overall offset of $\hat{\mu}$, yielding a proper posterior. Equivalently, the associated RKHS is the quotient space $H^1(\mathcal{X})/\{\text{constants}\}$.

For example, a prior of the form

$$p(\hat{\mu}) \propto \exp\left[-\tfrac{\gamma}{2} \int_{\mathcal{X}} p(x)\, \|\nabla \hat{\mu}(x)\|^2 \, dx\right]$$

corresponds to a Gaussian field whose precision is $\gamma \mathcal{L}_p$, and hence to a GP whose covariance is $(\gamma \mathcal{L}_p)^{-1}$. In this view, the Dirichlet energy simply plays the role of a smoothness penalty: in the GP formulation it arises from the covariance kernel, and in the field-theoretic formulation it arises from the corresponding precision operator. These two descriptions are mathematically equivalent [Kimeldorf and Wahba, 1970, Wahba, 1990, Lindgren et al., 2011].

5.2 Sampling and Approximation

In principle, posterior samples of the continuous functions can be generated using stochastic-gradient Langevin dynamics (SGLD), which treats the MAP functional as an energy landscape and simulates noisy gradient descent [Welling and Teh, 2011]:

$$z_{t+1} = z_t - \epsilon_t \, \widehat{\nabla} \Phi_{\text{MAP}}(z_t) + \sqrt{2\epsilon_t}\, \xi_t, \quad \xi_t \sim \mathcal{N}(0, I), \ \sum_t \epsilon_t = \infty, \ \sum_t \epsilon_t^2 < \infty.$$

Here z_t denotes the concatenated discretization of the functions $(\hat{\mu}, \hat{\eta})$ on a finite lattice or mesh (i.e., $\mathsf{L}^{(D)}$), so each SGLD iterate is a lattice-based approximation of the continuous predictor

functions. Sampling and optimization are therefore performed in a discretized representation of the BFT, whose precise construction is given in Section 5.2.1. Posterior samples $\{(\hat{\mu}^{(m)}, \hat{\eta}^{(m)})\}_{m=1}^{M}$ represent discrete function realizations drawn from this approximate posterior. Their dispersion across $\hat{\mu}^{(m)}$ captures epistemic uncertainty, while the Monte Carlo mean $\mathbb{E}[e^{-\hat{\eta}^{(m)}}]$ estimates the expected aleatoric variance function.

Rather than drawing full functional samples, we approximate the posterior by training an ensemble of independently initialized FT models, each converging to a distinct local MAP solution. The variability across the ensemble provides a Monte Carlo approximation to the Bayesian posterior $p(\hat{\mu}, \hat{\eta} \mid \mathcal{D})$. This ensemble view operationalizes the BFT: the deterministic FT gives the MAP equations, while multiple FT realizations collectively capture epistemic uncertainty through their dispersion in predictive means and noise functions.

The behavior of the BFT can be visualized by overlaying these independent FT realizations, which reveal the spread of predicted means and variances across ensemble members (Fig. 2). Rather than integrating out epistemic variation, we directly display the spread of predicted means $\hat{\mu}^{(m)}(x)$ and variances $e^{-\hat{\eta}^{(m)}(x)}$ across ensemble members. This representation highlights the posterior support and qualitative variability of solutions, which together approximate the epistemic uncertainty of the BFT.

5.2.1 Discretization and practical approximation of the continuum BFT

As in the deterministic FT (Section 4.3), the continuous predictor fields $(\hat{\mu}, \hat{\eta})$ are evaluated on a finite lattice $\mathsf{L}^{(D)} = \{x_i^{(D)}\}_{i=1}^{D} \subset \mathcal{X}$ to obtain discrete representations $(\hat{\mu}^{(D)}, \hat{\eta}^{(D)})$. We assume that the empirical measure of the lattice converges weakly to the data distribution,

$$\frac{1}{D}\sum_{i=1}^{D} \delta_{x_i^{(D)}} \;\Rightarrow\; p(x)\,dx,$$

so that weighted sums over $\mathsf{L}^{(D)}$ provide Monte–Carlo approximations to p–integrals.

In principle, one obtains finite-dimensional analogues of the continuous Gaussian field priors in Eqs. (13a) and (13b) by replacing spatial derivatives with centered finite-difference operators and integrals with weighted sums $\sum_i p(x_i^{(D)})\,(\cdot)$. As in the deterministic discretization, one ghost node is added on each side of the domain and reflected to enforce homogeneous Neumann boundary conditions, allowing centered differences to be used at every interior lattice point. The resulting discrete priors are Gaussian Markov random fields (GMRFs): multivariate Gaussian distributions with sparse precision matrices encoding local conditional independences on the lattice. These precision matrices provide consistent finite-difference approximations of the weighted elliptic operators associated with the continuous priors, including the weighted Laplacian $\mathcal{L}_p f = -\nabla \cdot (p \nabla f)$.

Under standard assumptions on regularity, boundary conditions, and mesh refinement, such GMRF priors converge weakly, in the sense of finite-dimensional distributions, to their continuous Gaussian field counterparts as $D \to \infty$ [Lindgren et al., 2011, 2022]. Further details on the discrete operators and their continuum limits are provided in Appendix Section B.2. Thus the lattice-based construction above yields a principled finite-dimensional approximation of the continuum Bayesian field theory.

In practice, however, the numerical implementation used in Section 6 adopts a simpler and more computationally tractable approximation. Rather than sampling from or explicitly forming the GMRF prior, we solve an ensemble of deterministic field theory optimizations, each initialized with a different random seed. Each run produces a solution of the deterministic FT on the lattice grid, and the resulting ensemble of solutions provides an empirical approximation to the posterior variability

that would be induced by the full BFT model. This ensemble-based procedure captures meaningful epistemic variability while avoiding the computational cost of full GMRF-based inference.

5.3 Predictive Decomposition

Because both latent functions $(\hat{\mu}, \hat{\eta})$ are random under the posterior, the BFT admits a hierarchical uncertainty decomposition that distinguishes variability in the predictive mean from variability in the noise function. This hierarchy allows the model to represent not only uncertainty about predictions, but also uncertainty about how predictable each region of the input space is. Formally, the Bayesian predictive distribution marginalizes over both functions,

$$p(y \,|\, x, \mathcal{D}) = \iint p(y \,|\, x, \hat{\mu}, \hat{\eta})\, p(\hat{\mu}, \hat{\eta} \,|\, \mathcal{D})\, d\hat{\mu}\, d\hat{\eta}, \tag{15}$$

with predictive variance

$$\mathrm{Var}[y \,|\, x, \mathcal{D}] = \mathbb{E}[e^{-\hat{\eta}(x)} \,|\, \mathcal{D}] + \mathrm{Var}[\hat{\mu}(x) \,|\, \mathcal{D}], \tag{16}$$

where the first term integrates over uncertainty in the noise function (aleatoric) and the second reflects uncertainty in the mean function (epistemic). This decomposition holds generally for any joint posterior $p(\hat{\mu}, \hat{\eta} \,|\, \mathcal{D})$ and does not depend on the specific choice of priors, so long as $\hat{\mu}$ and $\hat{\eta}$ are treated as random functions with finite second moments.

Although the predictive mean depends directly on $\hat{\mu}(x)$, we write the expectation $\mathbb{E}_{\hat{\mu},\hat{\eta}}[\hat{\mu}(x)]$ since the posterior $p(\hat{\mu}, \hat{\eta} \,|\, \mathcal{D})$ couples both functions through the likelihood. In practice, this reduces to a marginal expectation over $\hat{\mu}$ once $\hat{\eta}$ is integrated out.

These expectations can be estimated from a finite ensemble of MAP or SGLD function realizations. Let $\{(\hat{\mu}^{(m)}, \hat{\eta}^{(m)})\}_{m=1}^{M}$ denote M independent samples. Then the pointwise predictive mean and uncertainty components can be approximated as

$$\hat{\mu}^*(x) = \tfrac{1}{M} \sum_{m=1}^{M} \hat{\mu}^{(m)}(x), \qquad \hat{\sigma}^2_{\mathrm{epi}}(x) = \mathrm{Var}_m[\hat{\mu}^{(m)}(x)], \qquad \hat{\sigma}^2_{\mathrm{ale}}(x) = \tfrac{1}{M} \sum_{m=1}^{M} e^{-\hat{\eta}^{(m)}(x)}.$$

The total predictive uncertainty is given by $\hat{\sigma}_{\mathrm{tot}}(x) = (\hat{\sigma}^2_{\mathrm{epi}}(x) + \hat{\sigma}^2_{\mathrm{ale}}(x))^{1/2}$. These Monte Carlo estimators provide a practical implementation of the Bayesian predictive decomposition, linking the theoretical formulation directly to ensemble-based experiments.

5.4 Statistical Interpretation and Operator Geometry

The Bayesian FT induces Gaussian field priors whose precision operator is the weighted Laplacian $\mathcal{L}_p f = -\nabla \cdot (p \nabla f)$, as introduced in Lemma 1. This places the model squarely within the frameworks of operator–based Gaussian processes and reproducing–kernel Hilbert spaces (RKHS) [Kimeldorf and Wahba, 1970, Wahba, 1990, Lindgren et al., 2011]. In this view, the Dirichlet energy $\int p \, \|\nabla f\|^2$ plays the role of a smoothness penalty: in the GP formulation it arises from the covariance kernel $(\mathcal{L}_p)^{-1}$, while in the field–theoretic formulation it appears as the corresponding precision operator. These two descriptions are mathematically equivalent and provide a unified interpretation of the FT as a Gaussian prior over functions with weighted Sobolev structure.

This viewpoint clarifies the role of the additive FT geometry. Penalizing $\|\nabla \hat{\Lambda}\|^2$ enforces smoothness in the absolute noise level, while the log–precision parameterization $\hat{\eta} = \log \hat{\Lambda}$ ensures positivity and preserves the form of the Gaussian likelihood. When $\hat{\Lambda}$ varies smoothly and is bounded away from zero, these geometries differ only by a spatially varying rescaling through

$\nabla\hat{\eta} = \nabla\hat{\Lambda}/\hat{\Lambda}$, yielding similar stationary behavior within the stable region of the regularization space. The BFT thus recovers the deterministic FT at the posterior mode while also providing a principled probabilistic interpretation of the regularizers and the associated uncertainty decomposition.

This operator-based interpretation connects the FT to classical statistical treatments of heteroskedastic regression, in which smoothness penalties on mean and variance functions arise from Gaussian priors [Yuan and Wahba, 2004, Lemm, 2000]. The Bayesian formulation adopted here provides the continuum analogue: a Gaussian prior defined through a differential operator, yielding a flexible nonparametric model whose posterior mode corresponds exactly to the deterministic FT and whose ensemble-based approximations provide a practical route to epistemic uncertainty estimation.

6 Experiments

The primary goal of our experiments is to visualize phase transitions in two-dimensional phase diagrams. We show that the qualitative structure of these phase diagrams is independent of any specific neural network architecture by demonstrating close agreement with the field-theoretic solutions. This analysis also yields a practical procedure for selecting well-suited (ρ, γ) regularization strengths, reducing a two-dimensional hyperparameter search to one dimension. Our main experiments use fully connected networks with (ρ, γ)–L_2 regularization, and we additionally assess the variability of model fits across multiple runs.

6.1 Field Theory and Neural Networks

The field-theoretic formulation developed in Section 4 describes the behavior of overparameterized mean–variance regression models in a function-space limit, abstracting away architectural details of specific predictors. In practice, however, we implement these ideas using fully connected neural networks. Although neural networks do not span the same function space as the nonparametric fields considered by the FT, modern overparameterized architectures are sufficiently expressive to approximate the relevant solution classes and to exhibit the same characteristic phase transitions predicted by the theory.

Our goal is therefore not to enforce an exact architectural correspondence, but to verify empirically that standard neural networks trained with simple (ρ, γ)–weighted L_2 regularization follow the qualitative regimes identified by the FT. We apply these penalties separately to the mean and precision networks, mirroring the allocation of smoothness in the field theory, and we observe sharp transitions between underfitting, stable, and overfitting regimes across datasets. The resulting neural-network phase diagrams closely match those obtained from numerically solving the FT, demonstrating that the field theory captures the coarse-grained behavior of practical heteroskedastic regressors without requiring architectural modifications.

6.2 Modeling Choices

We chose $\hat{\mu}_\theta, \hat{\Lambda}_\phi$ to be fully-connected networks with three hidden layers of 128 nodes and leaky ReLU activation functions. The first half of training was only spent on fitting $\hat{\mu}_\theta$, while in the second half of training, both $\hat{\mu}_\theta$ and $\hat{\Lambda}_\phi$ were jointly learned. This improves stability, since the precision is a dependent on the mean $\hat{\mu}_\theta$, and is similar in spirit to ideas presented in Detlefsen et al. [2019]. Complete training details can be found in Appendix C.2.

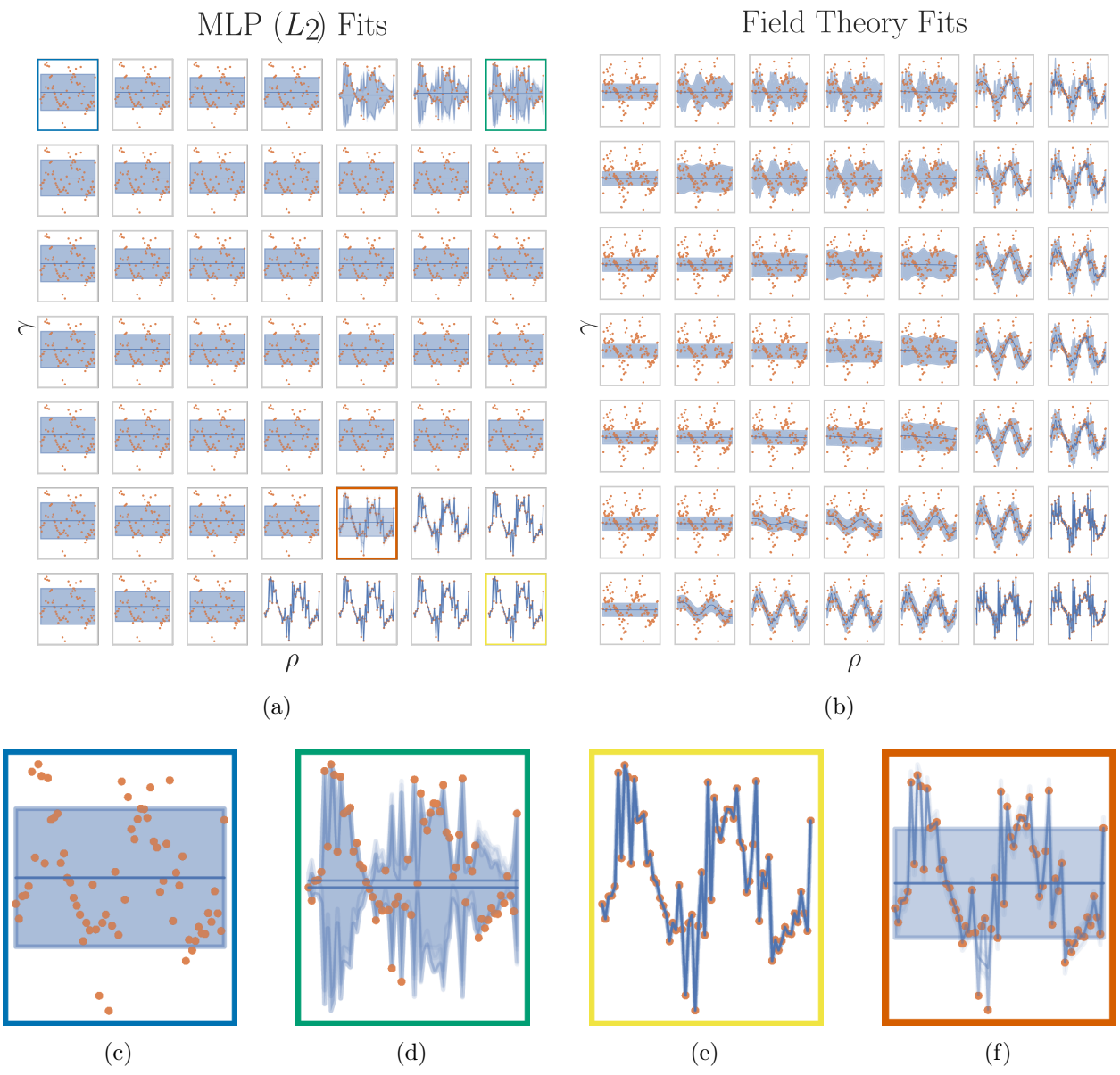


Figure 2: Ensemble fits from two modeling approaches. Training data are shown in orange; the ensemble mean (blue) and its pointwise ± 1 s.d. band (shaded) are overlaid for six independent runs. Panels (a) and (b) show a neural implementation and its FT counterpart, respectively. Panels (c)–(f) illustrate representative neural network fits in different overfitting and underfitting regimes, with panel (f) displaying phase coexistence in (ρ, γ) space.

6.3 Datasets

We study regularization effects on several one-dimensional simulated datasets and on standardized versions of the *Concrete* [Yeh, 2007], *Housing* [Harrison and Rubinfeld, 1978], *Power* [Tüfekci, 2014], and *Yacht* [Gerritsma, 1981] regression datasets from the UCI Repository [Kelly et al.], along with a scalar variable from the ClimSim dataset [Yu et al., 2023]. We fit neural networks to both simulated and real data, and additionally solve the FT on the simulated datasets. Dataset details appear in Appendix C.1. We show results for the *Sine* dataset and the four UCI datasets, with additional simulated results in Appendix C.4.

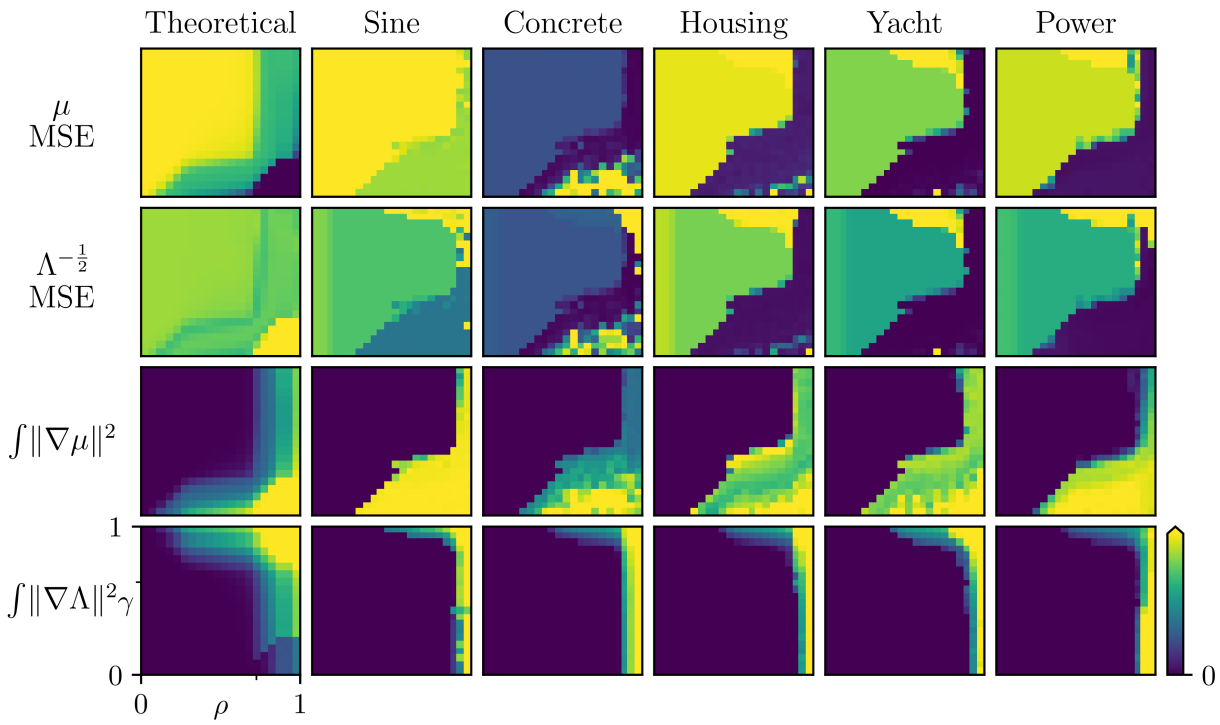


Figure 3: Array plot of evaluation metrics (*rows*) across datasets or fitting methods (*columns*) on the (ρ, γ) regularization grid. The *leftmost* column shows FT solutions; remaining columns show neural-network fits on held-out data. Each heatmap averages six runs. Ticks mark $\rho = 0.5$ and $\gamma = 0.5$ in the lower-left panel. Both axes use a logit parameterization of $\rho, \gamma \in (0, 1)$ to highlight limiting behaviors near 0 and 1. The FT captures the same transition structure observed in the empirical diagrams across datasets.

6.4 Qualitative Analysis

Our qualitative analysis aims at understanding architecture-independent aspects of mean-variance regression upon varying the regularization strength on the mean and variance functions, resulting in the observation of phase transitions.

6.4.1 Observables

We evaluate both the calibration and expressiveness of the learned models. For calibration, we compute mean squared error (MSE) for the predicted mean, $\hat{\mu}_\theta(x_i)$, and for the predicted standard

deviation, $\Lambda^{-1/2}(x_i)$, comparing the latter to the absolute residuals $|y_i - \hat{\mu}_\theta(x_i)|$. Well-fit models exhibit low errors for both measures, and we report $\Lambda^{-1/2}$-MSE due to its connection to variance-calibration metrics [Detlefsen et al., 2019, Levi et al., 2022].

To assess expressiveness, we compute the Dirichlet energy for the FT solutions and its discrete analogue, the geometric complexity [Dherin et al., 2022], for neural networks. The Dirichlet energy of a function f is $\int_{\mathcal{X}} p(x)|\nabla f(x)|_2^2 dx$, while geometric complexity is $N^{-1}\sum_{i=1}^{N}|\nabla f(x_i)|_2^2$. Both quantify the variability of the learned functions, with larger values indicating more expressive (less regularized) behavior and directly corresponding to the quantities penalized in the FT formulation.

6.4.2 Plot Interpretation

We present summaries of the fitted models in grids with ρ on the x-axis and γ on the y-axis in Figs. 3 and 4. The far right column ($\rho = 1$) corresponds to MLE solutions. The main focus is on qualitative traits of fits under different levels of regularization and how they behave in a relative sense, rather than a focus on absolute values. Fig. 5 show the summary statistics along the slice where $\rho = 1 - \gamma$. Zero on these plots corresponds to the upper left corner while one corresponds to the lower right corner. We provide model fits arranged in grids of the same orientation for the field theory and neural networks on the *Sine* dataset in Fig. 2.

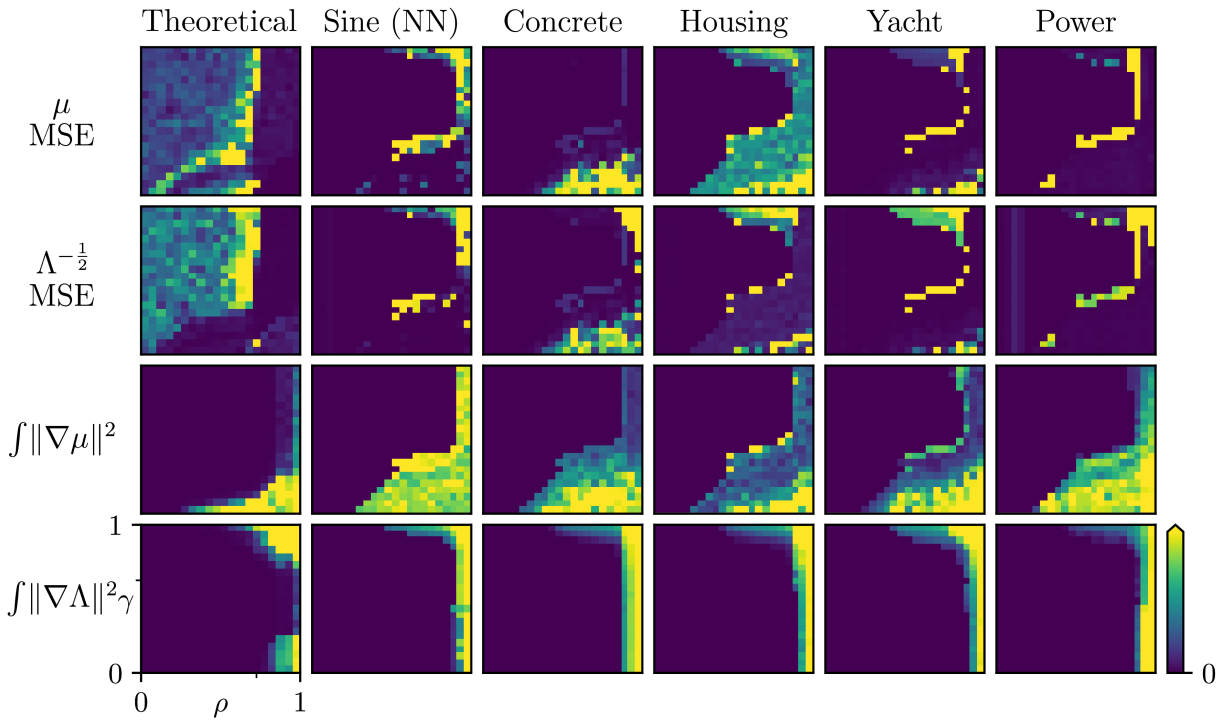


Figure 4: We compute the standard deviation over six runs for each metric in Fig. 3, illustrating how variability changes across the regularization space. The shapes of the instability regions remain consistent across datasets and between the neural networks and the FT, as reflected in the Dirichlet energies and geometric complexities. These quantities show the largest disagreement in overfitting regimes, though this does not always correspond to high variability in the MSEs.

6.4.3 Variability Over Runs

The preceding experiments focused on pointwise estimation of the mean and variance functions, capturing a single prediction and its associated *aleatoric* uncertainty. To assess *epistemic* uncertainty, we examine the variability of phase diagrams across multiple independent MVR fits. For the one-dimensional synthetic datasets, we visualize this variability directly by plotting individual model fits, illustrating how the consistency of learned functions changes across regions of the phase diagram.

Observation 1. *Our metrics show sharp phase transitions upon varying ρ, γ, as in a physical system.*

Fig. 3 and Fig. 5 show a sharp transition, both leading to worsening and improving performance when moving along the minor diagonal. In totality, across all metrics, the five regions are apparent. But not all of the regions in Fig. 1 appear in the heatmaps of each metric. For example, region O_Λ does not always appear in the metrics related to the mean. When using neural networks to approximate μ and Λ, there are sharper boundaries between phases than in the FT's numerical solutions. The boundary between U_μ and O_μ is sharply observed in the plots of $\int \|\nabla \mu(x)\|_2^2 \, dx$. However, in terms of μ-MSE, a smoother transition (i.e., region S) is visible.

Observation 2. *The FT insights and observed phases are consistent with both the numerically solved FT and the neural-network fits. Thus, our conclusions are not tied to a specific architecture or dataset.*

In line with the theoretical predictions, phases U_Λ and O_μ exhibit consistent behavior across γ-values (vertical slices in Fig. 3). Across all datasets we considered, the qualitative structure of the phase diagrams remains the same: the same phase types appear, the same ordering of regimes is observed, and the transitions occur along similarly shaped boundaries. Representative fitted models are shown in Fig. 2.

While the overall phase structure is shared across datasets, the precise locations of the transitions vary. Different datasets and input dimensionalities shift the (ρ, γ) values at which the boundaries occur, but the geometric shape and ordering of regions remain stable.

In the right-hand columns ($\rho \to 1$) the mean function nearly interpolates the data, and similar behavior is seen in the lower rows ($\gamma \to 0$). Across all metrics, the regions evolve with regularization strength in a comparable manner on all datasets. Notably, the region of small $\int \|\nabla \Lambda(x)\|_2^2 \, dx$ covers a larger portion of the diagram than the corresponding region for $\int \|\nabla \mu(x)\|_2^2 \, dx$, indicating that the precision function remains smoother than the mean under comparable levels of regularization.

Observation 3. *The neural network phase diagrams reveal different amounts of variability in model fits across the regularization space.*

This behavior hints at variability in the fitting procedure and can be considered a sign of the need to measure epistemic uncertainty. The standard deviations over the metrics displayed in Fig. 3 are shown in Fig. 4. The Dirichlet energies/geometric complexities show that there is the most variability in the overfitting regions O_μ and parts of O_Λ. This indicates that the functions themselves vary across runs. Actual fits of the *Sine* dataset are displayed in Fig. 2. However, when turning to quality of fits, the MSEs show a different pattern of regions of instability, and O_μ has low variability in terms of actual performance.

Note that in the region of high regularization (*far left column*) we see greater variability than in the moderately regularized regions in the upper middle. We posit that in the highly regularized regions we are essentially seeing the variability coming from the random initialization of the model weights. Meanwhile in the central region we see that the mean and variance functions are afforded

enough "flexibility" to adapt to the global mean and standard deviation, but not enough flexibility to fit to the data. Thus there is much less spread between the different ensemble members in this region.

Observation 4. *The neural network phase diagrams exhibit regions of instability: for certain* (ρ, γ) *values, independently trained MLPs produce qualitatively different fits, whereas the FT solutions are highly consistent across runs.*

Fig. 2 illustrates this phenomenon. Even with identical (ρ, γ) values and full-batch gradient descent, different random initializations of the MLP parameters can lead to distinct fitting behaviors—some runs overfit while others underfit the data (see the outlying curve in Fig. 2a). This variability is quantified in Fig. 4, which reports the standard deviation of MSEs and Dirichlet energies across runs and reveals pockets of substantial instability in the neural-network landscape.

In contrast, the field-theoretic fits (Fig. 2b) show almost no variation across runs. Although we also initialize the discretized fields μ and Λ randomly and optimize them by gradient descent, the FT energy appears to have a much smoother and more strongly regularized landscape: empirically, all runs converge to essentially the same solution for a fixed (ρ, γ). We do not claim uniqueness of the FT minimizer analytically, but in practice the FT optimization exhibits a single stable attractor, in sharp contrast to the multiple effective basins observed for the neural networks.

6.5 Quantitative Analysis

Our quantitative analysis aims to demonstrate the practical implications of our qualitative investigations that result in better calibration properties.

Observation 5. *We can search along* $\rho = 1 - \gamma$ *to find a well-calibrated* (ρ, γ)*-pair from region* S*.*

Table 1: Comparison of our mean-variance regression model (Ours) with diagonal regularization search and β-NLL [Seitzer et al., 2022]. Details on our MVR selection criteria can be found in Appendix E.2. We report the average and standard deviations of μ- and $\Lambda^{-\frac{1}{2}}$-MSE across six runs on test data.

Metric	Sine	Concrete	Housing	Power	Yacht	Solar Flux
μ-MSE						
Ours	0.80 ± 0.00	**0.11** ± 0.02	1.22 ± 0.00	**0.04** ± 0.01	**0.01** ± 0.01	**0.29** ± 0.00
β-NLL	**0.69** ± 0.05	0.55 ± 0.30	**0.32** ± 0.05	0.09 ± 0.01	**0.01** ± 0.01	0.38 ± 0.00
$\Lambda^{-\frac{1}{2}}$-MSE						
Ours	0.80 ± 0.00	**0.30** ± 0.51	**0.76** ± 0.00	**0.03** ± 0.01	**0.01** ± 0.01	**0.12** ± 0.00
β-NLL	**0.52** ± 0.07	1.09 ± 0.20	0.88 ± 0.03	0.31 ± 0.37	1.33 ± 0.02	0.32 ± 0.00

Our FT indicates that a slice across the minor diagonal of the phase diagram should always cross the S region (see Fig. 1). Fig. 5 shows that by searching along this diagonal, we indeed find a combination of regularization strengths where both $\hat{\mu}_\theta$ and $\hat{\Lambda}_\phi$ generalize well to held-out test data. This implies that there is no need to search all of the two-dimensional space, but only a single slice which reduces the number of models to fit from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, where N is the number of ρ and γ values that are tested. This finding is consistent with the suggestion from Sluijterman et al. [2024] to have stronger regularization on the variance than the mean. This corresponds to searching across a horizontal slice in the lower portion of our phase diagram and is generally consistent with where we posit the well-behaved S region tends to lie.

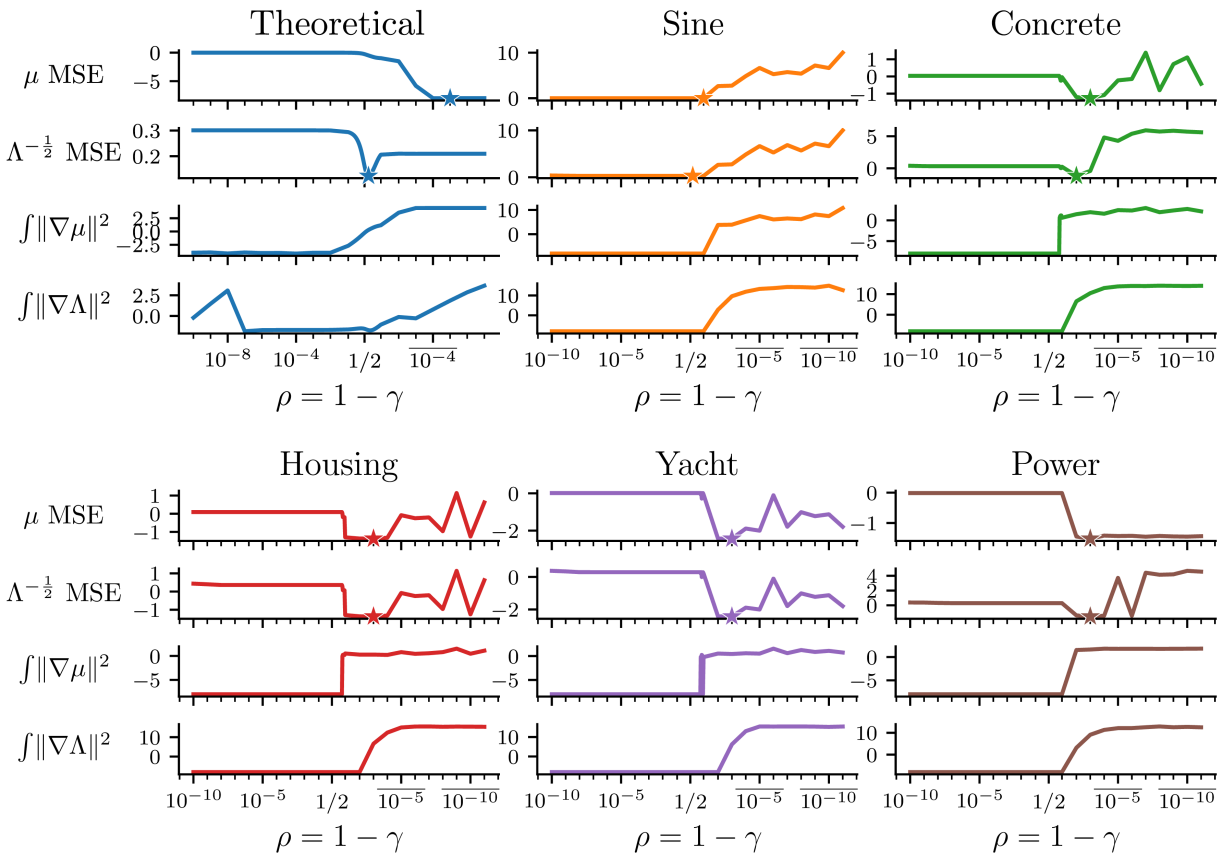


Figure 5: Test metrics across six runs along the $\rho = 1 - \gamma$ diagonal. Stars denote the minimum MSE for each dataset. All metrics are plotted on a $\log_{10}$ scale, and ρ is shown on a logit scale to highlight behavior near the boundaries. Errors drop sharply near the transition into the S phase and then increase again as ρ moves past this region, consistent with the qualitative structure in Fig. 1.

Fig. 5 shows that along the minor diagonal the performance is initially poor, improves, and then drops off again. These shifts from strong to weak performance are sharp. The regularization pairings that result in optimal performance with respect to μ- and $\Lambda^{-1/2}$-MSE are near each other along this diagonal for the real-world test data. As the theory predicts, the performance becomes highly variable as we approach the MLE solutions and the FT fails to converge in this region. In practice, we propose searching along this line to find the (ρ, γ)-combinations that minimize the μ- and $\Lambda^{-\frac{1}{2}}$-MSEs and averaging the regularization strengths to fit a model. We compare models chosen by our diagonal line search to two heteroskedastic modeling baselines in Appendix E on the synthetic and UCI datasets as well as a scalar quantity from the ClimSim dataset [Yu et al., 2023]. We present a subset of the results below in Table 1. In most cases the model chosen via the diagonal line search was competitive or better than the baselines.

7 Conclusion

Neural mean–variance models are known to exhibit significant training instabilities [Nix and Weigend, 1994a,b]. By developing a field-theoretic perspective grounded in statistical physics, we derived a continuum variational formulation of the learning problem that isolates structural causes of these pathologies [Lemm, 2000, Ringel et al., 2025]. We refer to this continuum formulation as the field theory, and it yields explicit, architecture-independent insights into the behavior of deep heteroskedastic regression. It also helps clarify why these models often require carefully tuned regularization and why they tend to exhibit transitions between qualitatively different regimes.

Building on this continuum formulation, we introduced a numerical discretization of the field theory and demonstrated close qualitative agreement with neural network solutions across both synthetic and real-world datasets. Across repeated fits, we observed two central challenges: inconsistent behavior across model runs and inconsistent behavior across regularization strengths. The field theory clarifies these effects and motivates a more principled strategy for tuning regularization, reducing a two-dimensional search to an effectively one-dimensional problem. We also found that independently trained neural networks often undergo their transitions at different points in the regularization space. This produces an effect reminiscent of phase coexistence, where ensemble variability captures structure that is not visible in any single fit. The Bayesian field-theoretic perspective clarifies part of this connection to epistemic uncertainty, although a full characterization within a complete Bayesian framework remains to be developed. The Bayesian formulation also links the continuum model to classical statistical frameworks, including Gaussian process priors, spline-based smoothing, and penalized likelihood methods [Wahba, 1990, Lindgren et al., 2011], which arise as special cases under particular choices of regularizers.

7.1 Limitations

The field-theoretic formulation captures several important aspects of neural behavior, but it also has limitations. By replacing the discrete neural objective with a population-level variational problem, it abstracts away optimization dynamics, architectural nonlinearities, and other discrete effects that are present in neural networks and may influence the sharpness of observed transitions. As a result, the continuum phase diagram should not be expected to match neural behavior in detail. Nevertheless, across datasets with different scales, dimensionalities, and smoothness levels, we observe similar qualitative transition structures, and the field theory reproduces this high-level organization. However, it does not fully capture quantitative differences such as the severity or abruptness of certain transitions seen in neural networks. Clarifying how the discrete neural parameterization relates to its continuum limit may help close this gap.

Finally, our approach illustrates a broader methodological trade-off. By passing to a continuum and imposing a structured variational framework, the field theory sacrifices some of the specificity of discrete neural architectures in exchange for analytic clarity. At the same time, it highlights structural mechanisms that are difficult to isolate directly, such as the competing effects of the likelihood and the regularizers, the influence of the data density, and the organization of solutions into distinct qualitative regimes. Related phenomena arise in other learning problems where objectives exhibit instabilities or undergo symmetry breaking, such as the continuous-symmetry scenarios studied by Bamler and Mandt [2018]. This reflects a theme familiar from the philosophy of mathematical physics, where idealized formulations sacrifice some realism but provide conceptual clarity about the forces that shape a system's behavior [Kronz and Lupher, 2025]. In this sense, the field-theoretic formulation should be viewed as a complementary perspective rather than a replacement for the neural model.

Overall, we hope that this work encourages further exploration of phase transitions, variational principles, and operator-based ideas as tools for understanding the collective and nonlinear behavior that arises in modern large-scale deep learning models.

Acknowledgments

Eliot Wong-Toi acknowledges support from the Hasso Plattner Research School at UC Irvine. Vincent Fortuin was supported by a Branco Weiss Fellowship. Stephan Mandt acknowledges support from the IARPA WRIVA program; the National Science Foundation (NSF) under the CAREER Award 2047418 and Grants 2003237 and 2007719; the Department of Energy, Office of Science under Grant DE-SC0022331; and gifts from Intel, Disney, and Qualcomm.

A Theoretical Details

The results summarized in Prop. 1 are derived here in full detail.

A.1 General Field Theory and Extreme Settings

Weighted Green's Identity and Natural Neumann Data

Lemma 1 (Weighted Green's identity)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a bounded Lipschitz domain with outward unit normal $\boldsymbol{n}$ on $\partial\mathcal{X}$, and let $p \in C^1(\overline{\mathcal{X}})$ satisfy $p > 0$. For any $f, g \in H^1(\mathcal{X})$,*

$$\int_{\mathcal{X}} p\,\nabla f \cdot \nabla g\,dx = -\int_{\mathcal{X}} g\,\nabla\cdot\big(p\nabla f\big)\,dx + \int_{\partial\mathcal{X}} g\,p\,\nabla f\cdot\boldsymbol{n}\,dS. \tag{17}$$

Equivalently, define the (negative) weighted Laplacian

$$\mathcal{L}_p f := -\nabla\cdot\big(p\,\nabla f\big),$$

so that $\mathcal{L}_p$ is a positive semidefinite, self-adjoint operator on $H^1(\mathcal{X})$ with respect to the weighted inner product $\langle f, g\rangle_p = \int_{\mathcal{X}} p\,f g\,dx$. Then (17) *becomes*

$$\int_{\mathcal{X}} p\,\nabla f\cdot\nabla g\,dx = \int_{\mathcal{X}} p\,(\mathcal{L}f)\,g\,dx + \int_{\partial\mathcal{X}} g\,p\,\nabla f\cdot\boldsymbol{n}\,dS. \tag{18}$$

Proof. Apply the divergence theorem to the vector field $g\,p\,\nabla f$:

$$\int_{\mathcal{X}} \nabla\cdot\big(g\,p\,\nabla f\big)\,dx = \int_{\partial\mathcal{X}} g\,p\,\nabla f\cdot\boldsymbol{n}\,dS.$$

Expanding the divergence gives

$$\nabla\cdot\big(g\,p\,\nabla f\big) = g\,\nabla\cdot\big(p\nabla f\big) + p\,\nabla f\cdot\nabla g.$$

Rearranging yields (17). Substituting $\mathcal{L}f = -\nabla\cdot(p\nabla f)$ gives (18). □

Natural Neumann boundary conditions. If the co-normal derivative $(p\nabla f)\cdot\boldsymbol{n}$ is required to vanish on $\partial\mathcal{X}$ (the homogeneous Neumann condition), then the boundary term in (18) disappears. This yields

$$\int_{\mathcal{X}} p\,\nabla f\cdot\nabla g\,dx = \int_{\mathcal{X}} p\,(\mathcal{L}_p f)\,g\,dx,$$

showing that homogeneous Neumann boundary conditions arise naturally as the *natural* boundary conditions of the weighted Dirichlet energy $\int p\,\|\nabla f\|^2$.

Remark 1 (On the weighted Laplacian)**.** *With the convention*

$$\mathcal{L}_p f := -\nabla\cdot(p\nabla f),$$

we may expand in Euclidean coordinates as

$$\mathcal{L}_p f = -p\,\Delta f - (\nabla p)\cdot\nabla f.$$

A normalized form,

$$\tilde{\mathcal{L}}_p f := \frac{1}{p}\mathcal{L}_p f = -\Delta f - (\nabla\log p)\cdot\nabla f,$$

makes explicit the drift term induced by the nonuniform weight $p(x)$. Both $\mathcal{L}_p$ and its normalized form encode the weighted Laplace operator that arises from integration by parts under the measure $p(x)\,dx$.

Corollary 1 (Natural (Neumann) boundary conditions)**.** *If $(p\nabla f)\cdot \boldsymbol{n} = 0$ on $\partial\mathcal{X}$, then for all $g \in H^1(\mathcal{X})$,*

$$\int_{\mathcal{X}} p\,\nabla f\cdot\nabla g\,dx = \int_{\mathcal{X}} p\,(\mathcal{L}_p f)\,g\,dx.$$

Thus quadratic Dirichlet energies $\int \frac{\kappa}{2} p\|\nabla f\|^2\,dx$ induce $\mathcal{L}_p$ as the Euler–Lagrange operator with homogeneous zero-flux as the natural boundary condition.

General Field Theory Formulation

Proposition 2 (General Field Theory)**.** *Assume that $\mathcal{X} \subset \mathbb{R}^d$ is a bounded Lipschitz domain, and that $p \in C^1(\overline{\mathcal{X}})$ is a strictly positive probability density on $\mathcal{X}$.*

Let $\hat{\mu}, \hat{\Lambda} \in H^1(\mathcal{X})$ satisfy

$$\hat{\Lambda}(x) \geq c > 0 \qquad \text{for a.e. } x \in \mathcal{X},$$

and assume their weighted Dirichlet energies are finite:

$$\int_{\mathcal{X}} p\,\|\nabla\hat{\mu}\|_2^2\,dx < \infty, \qquad \int_{\mathcal{X}} p\,\|\nabla\hat{\Lambda}\|_2^2\,dx < \infty.$$

Define the unnormalized weighted Laplacian

$$\mathcal{L}_p f := -\nabla\cdot\big(p\nabla f\big).$$

Let

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu},\hat{\Lambda}] = \int_{\mathcal{X}} p(x)\Big[-\rho\log\hat{p}(y\,|\,x) + \bar{\rho}\Big(\gamma\|\nabla\hat{\mu}(x)\|_2^2 + \bar{\gamma}\|\nabla\hat{\Lambda}(x)\|_2^2\Big)\Big]dx, \tag{19}$$

where $\hat{p}(y \mid x) = \mathcal{N}(y \mid \hat{\mu}(x), \hat{\Lambda}(x)^{-1})$, $\bar{\rho} = 1-\rho$, and $\bar{\gamma} = 1-\gamma$. Then stationary points satisfy the Euler–Lagrange equations

$$\rho\,p(x)\,\hat{\Lambda}(x)\big(\hat{\mu}(x) - y(x)\big) = 2\bar{\rho}\,\gamma\,\mathcal{L}_p\hat{\mu}(x), \tag{20a}$$

$$\frac{\rho}{2}\,p(x)\left[\big(\hat{\mu}(x)-y(x)\big)^2 - \frac{1}{\hat{\Lambda}(x)}\right] = 2\bar{\rho}\,\bar{\gamma}\,\mathcal{L}_p\hat{\Lambda}(x), \tag{20b}$$

with homogeneous Neumann (zero-flux) boundary conditions

$$p\,\nabla\hat{\mu}\cdot\boldsymbol{n} = p\,\nabla\hat{\Lambda}\cdot\boldsymbol{n} = 0 \qquad \text{on } \partial\mathcal{X}.$$

Proof. Rewrite the likelihood term using

$$-\log\hat{p}(y \mid x) = \tfrac{1}{2}\hat{\Lambda}(x)\,\hat{r}(x)^2 - \tfrac{1}{2}\log\hat{\Lambda}(x) + \text{const}, \qquad \hat{r}(x) = y(x) - \hat{\mu}(x).$$

Thus

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu},\hat{\Lambda}] = \int_{\mathcal{X}} p(x)\left\{\rho\Big[\tfrac{1}{2}\hat{\Lambda}\,\hat{r}^2 - \tfrac{1}{2}\log\hat{\Lambda}\Big] + \bar{\rho}\Big[\gamma\|\nabla\hat{\mu}\|_2^2 + \bar{\gamma}\|\nabla\hat{\Lambda}\|_2^2\Big]\right\}dx.$$

For test functions $\varphi, \psi \in H^1(\mathcal{X})$, consider perturbations $\hat{\mu}_\varepsilon = \hat{\mu} + \varepsilon\varphi$ and $\hat{\Lambda}_\varepsilon = \hat{\Lambda} + \varepsilon\psi$. We compute the first variations.

Variation with respect to $\hat{\mu}$. Only the terms $\frac{1}{2}\hat{\Lambda}\hat{r}^2$ and $\gamma\|\nabla\hat{\mu}\|^2$ depend on $\hat{\mu}$. Since $\hat{r} = y - \hat{\mu}$,

$$\frac{d}{d\varepsilon}\Big(\tfrac{1}{2}\hat{\Lambda}\,\hat{r}_\varepsilon^2\Big)\Big|_0 = \hat{\Lambda}(\hat{\mu} - y)\,\varphi.$$

The gradient term gives

$$\frac{d}{d\varepsilon}\|\nabla\hat{\mu}_\varepsilon\|^2\Big|_0 = 2\,\nabla\hat{\mu}\cdot\nabla\varphi.$$

Thus

$$\delta\mathbf{S}[\hat{\mu};\varphi] = \int_{\mathcal{X}} p\,\rho\,\hat{\Lambda}(\hat{\mu} - y)\varphi\,dx + 2\bar{\rho}\gamma\int_{\mathcal{X}} p\,\nabla\hat{\mu}\cdot\nabla\varphi\,dx.$$

Applying the weighted Green's identity

$$\int_{\mathcal{X}} p\,\nabla u\cdot\nabla v\,dx = \int_{\mathcal{X}} v\,\mathcal{L}_p u\,dx + \int_{\partial\mathcal{X}} p\,v\,\tfrac{\partial u}{\partial n}\,dS,$$

with $u = \hat{\mu}$, $v = \varphi$, yields

$$\delta\mathbf{S}[\hat{\mu};\varphi] = \int_{\mathcal{X}}\Big[\,p\rho\,\hat{\Lambda}(\hat{\mu} - y) + 2\bar{\rho}\gamma\,\mathcal{L}_p\hat{\mu}\,\Big]\varphi\,dx + 2\bar{\rho}\gamma\int_{\partial\mathcal{X}} p\,\varphi\,\tfrac{\partial\hat{\mu}}{\partial n}\,dS.$$

Stationarity for all φ on $\partial\mathcal{X}$ implies the natural boundary condition $p\,\frac{\partial\hat{\mu}}{\partial n} = 0$. Stationarity for all interior φ gives

$$p\,\rho\,\hat{\Lambda}(\hat{\mu} - y) = 2\bar{\rho}\gamma\,\mathcal{L}_p\hat{\mu},$$

which is (20a).

Variation with respect to $\hat{\Lambda}$. The relevant terms are $\frac{1}{2}\hat{\Lambda}\hat{r}^2$, $-\frac{1}{2}\log\hat{\Lambda}$, and $\bar{\gamma}\|\nabla\hat{\Lambda}\|^2$. We have

$$\frac{d}{d\varepsilon}\Big(\tfrac{1}{2}\hat{\Lambda}_\varepsilon\hat{r}^2 - \tfrac{1}{2}\log\hat{\Lambda}_\varepsilon\Big)\Big|_0 = \tfrac{1}{2}\Big(\hat{r}^2 - \tfrac{1}{\hat{\Lambda}}\Big)\,\psi,$$

and

$$\frac{d}{d\varepsilon}\|\nabla\hat{\Lambda}_\varepsilon\|^2\Big|_0 = 2\,\nabla\hat{\Lambda}\cdot\nabla\psi.$$

Thus

$$\delta\mathbf{S}[\hat{\Lambda};\psi] = \int_{\mathcal{X}} p\,\tfrac{\rho}{2}\Big(\hat{r}^2 - \tfrac{1}{\hat{\Lambda}}\Big)\,\psi\,dx + 2\bar{\rho}\bar{\gamma}\int_{\mathcal{X}} p\,\nabla\hat{\Lambda}\cdot\nabla\psi\,dx.$$

Applying the weighted Green's identity,

$$\delta\mathbf{S}[\hat{\Lambda};\psi] = \int_{\mathcal{X}}\Big[\,p\,\tfrac{\rho}{2}\Big(\hat{r}^2 - \tfrac{1}{\hat{\Lambda}}\Big) + 2\bar{\rho}\bar{\gamma}\,\mathcal{L}_p\hat{\Lambda}\Big]\psi\,dx + 2\bar{\rho}\bar{\gamma}\int_{\partial\mathcal{X}} p\,\psi\,\tfrac{\partial\hat{\Lambda}}{\partial n}\,dS.$$

Stationarity for all ψ on $\partial\mathcal{X}$ gives $p\,\frac{\partial\hat{\Lambda}}{\partial n} = 0$. Stationarity for all interior ψ yields

$$p\,\frac{\rho}{2}\Big[(\hat{\mu} - y)^2 - \tfrac{1}{\hat{\Lambda}}\Big] = 2\bar{\rho}\bar{\gamma}\,\mathcal{L}_p\hat{\Lambda},$$

which is (20b).

Conclusion. We obtain the coupled PDE system

$$p\rho\,\hat{\Lambda}(\hat{\mu}-y) = 2\bar{\rho}\gamma\,\mathcal{L}_p\hat{\mu},$$
$$p\,\tfrac{\rho}{2}\Big[(\hat{\mu}-y)^2 - \tfrac{1}{\hat{\Lambda}}\Big] = 2\bar{\rho}\bar{\gamma}\,\mathcal{L}_p\hat{\Lambda},$$

with homogeneous Neumann data $p\,\nabla\hat{\mu}\cdot n = p\,\nabla\hat{\Lambda}\cdot n = 0$. □

Remark 2 (Uniform-density case)**.** *If $p(x) \propto 1$, then $\mathcal{L}_p = \Delta$ and the zero-flux conditions reduce to $\nabla\hat{\mu}\cdot\boldsymbol{n} = 0$ and $\nabla\hat{\Lambda}\cdot\boldsymbol{n} = 0$.*

A.1.1 Empirical vs. population objective

Unless otherwise stated, extremal arguments in this appendix (e.g., the unboundedness results when a regularizer is removed) are stated for the *empirical* FT in which $y(\cdot)$ is treated as a fixed field, equivalently the empirical/Monte Carlo objective induced by a finite dataset. In that regime, a sufficiently rich function class can interpolate the observations, i.e., there exist $\hat{\mu}$ with $\hat{r}(x) := y(x) - \hat{\mu}(x) \equiv 0$, and then removing a corresponding regularizer can drive the objective to $-\infty$ via the $-\frac{1}{2}\log\hat{\Lambda}$ term.

In contrast, for the *population* FT with expectation over $p(y\,|\,x)$,

$$\mathbb{E}_{y|x}\Big[\tfrac{1}{2}\,\hat{\Lambda}(x)(y-\hat{\mu}(x))^2 - \tfrac{1}{2}\log\hat{\Lambda}(x)\Big] = \tfrac{1}{2}\,\hat{\Lambda}(x)\Big(\operatorname{Var}[y\,|\,x] + \big(m(x)-\hat{\mu}(x)\big)^2\Big) - \tfrac{1}{2}\log\hat{\Lambda}(x),$$

where $m(x) := \mathbb{E}[y\,|\,x]$. If $\operatorname{Var}[y\,|\,x] > 0$ on a set of positive measure, the term growing linearly in $\hat{\Lambda}$ dominates $-\log\hat{\Lambda}$ as $\hat{\Lambda}\to\infty$, and the functional is *not* driven to $-\infty$ by such a blow-up. Thus, the unboundedness claims we make in the extreme settings (e.g., Prop. 3) pertain to the empirical/interpolating regime commonly used in practice, not to the noisily stochastic population risk.

Proposition 3 (Extreme Settings in the General FT)**.** *Assume $p \in C^1(\overline{\mathcal{X}})$ is strictly positive on a bounded, connected domain $\mathcal{X}\subset\mathbb{R}^d$, and that $y \in H^1(\mathcal{X})$. Impose homogeneous Neumann boundary conditions $p\nabla\hat{\mu}\cdot\boldsymbol{n} = 0$ and $p\nabla\hat{\Lambda}\cdot\boldsymbol{n} = 0$ on $\partial\mathcal{X}$. Then for the general field theory*

$$\mathbf{S}_{\rho,\gamma}\left[\hat{\mu},\hat{\Lambda}\right] = \int_{\mathcal{X}} p(x)\Big[-\rho\log\hat{p}(y\,|\,x) + \bar{\rho}\Big(\gamma\|\nabla\hat{\mu}(x)\|_2^2 + \bar{\gamma}\|\nabla\hat{\Lambda}(x)\|_2^2\Big)\Big]dx, \tag{21}$$

where $\hat{p}(y\mid x) = \mathcal{N}(y\mid\hat{\mu}(x),\hat{\Lambda}(x)^{-1})$, the following properties hold:

(i) If $\rho = 1$, no stationary solution exists.

(ii) If $\rho = 0$, the solution is non-unique (any constant pair minimizes the functional).

(iii) If $\gamma = 0$ and $\rho > 0$, the objective is unbounded below.

(iv) If $\gamma = 1$ and $\rho > 0$, the objective is likewise unbounded below.

Proof. The Euler–Lagrange equations corresponding to $\mathbf{S}_{\rho,\gamma}$, using the unnormalized weighted Laplacian $\mathcal{L}_p f := -\nabla\cdot(p\nabla f)$, are

$$\rho\,p\,\hat{\Lambda}(\hat{\mu}-y) = 2\bar{\rho}\,\gamma\,\mathcal{L}_p\hat{\mu}, \tag{22}$$

$$\frac{\rho}{2}\,p\Big[(\hat{\mu}-y)^2 - \frac{1}{\hat{\Lambda}}\Big] = 2\bar{\rho}\,\bar{\gamma}\,\mathcal{L}_p\hat{\Lambda}, \tag{23}$$

with Neumann boundary conditions

$$p\nabla\hat{\mu}\cdot\boldsymbol{n} = 0, \qquad p\nabla\hat{\Lambda}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\mathcal{X}.$$

(i) No regularization ($\rho = 1$)**.** Setting $\rho = 1$ forces $\bar{\rho} = 0$, so the right-hand sides of (22)–(23) vanish:

$$p\,\hat{\Lambda}(\hat{\mu} - y) = 0, \qquad \frac{p}{2}\Big[\hat{\Lambda}^{-1} - (\hat{\mu} - y)^2\Big] = 0.$$

Since $p > 0$, this implies

$$\hat{\Lambda}(\hat{\mu} - y) = 0, \qquad \hat{\Lambda}^{-1} = (\hat{\mu} - y)^2.$$

Multiplying the second equation by $\hat{\Lambda}$ gives $\hat{\Lambda}(\hat{\mu} - y)^2 = 1$, which contradicts $\hat{\Lambda}(\hat{\mu} - y) = 0$. Thus no stationary point exists.

(ii) No data term ($\rho = 0$)**.** Setting $\rho = 0$ eliminates the likelihood contribution:

$$\mathbf{S}_{0,\gamma}[\hat{\mu}, \hat{\Lambda}] = \int_{\mathcal{X}} p(x)\big[\gamma\|\nabla\hat{\mu}\|_2^2 + \bar{\gamma}\|\nabla\hat{\Lambda}\|_2^2\big]dx.$$

The Euler–Lagrange equations reduce to

$$\mathcal{L}_p\hat{\mu} = 0, \qquad \mathcal{L}_p\hat{\Lambda} = 0,$$

with Neumann boundary conditions. On a connected domain with $p > 0$, the only solutions are constants, so the minimizer is non-unique (any constant pair).

(iii) No mean regularization ($\gamma = 0$)**.** With $\gamma = 0$ (so $\bar{\gamma} = 1$), the functional becomes

$$\mathbf{S}_{\rho,0}[\hat{\mu}, \hat{\Lambda}] = \int_{\mathcal{X}} p\,\frac{\rho}{2}\big(\hat{\Lambda}(\hat{\mu} - y)^2 - \log\hat{\Lambda}\big)\,dx + \bar{\rho}\int_{\mathcal{X}} p\,\|\nabla\hat{\Lambda}\|_2^2\,dx.$$

Choose $\hat{\mu} \equiv y$ and $\hat{\Lambda} \equiv C > 0$ constant. Then $\nabla\hat{\Lambda} = 0$ and $\hat{r} = y - \hat{\mu} = 0$, giving

$$\mathbf{S}_{\rho,0}[y, C] = -\frac{\rho}{2}\Big(\int_{\mathcal{X}} p(x)\,dx\Big)\log C.$$

Since $\log C \to \infty$ as $C \to \infty$, the objective tends to $-\infty$. Thus the functional is unbounded below when $\gamma = 0$.

(iv) No variance regularization ($\gamma = 1$) **(so** $\bar{\gamma} = 0$)**.** Now

$$\mathbf{S}_{\rho,1}[\hat{\mu}, \hat{\Lambda}] = \int_{\mathcal{X}} p\,\frac{\rho}{2}\big(\hat{\Lambda}(\hat{\mu} - y)^2 - \log\hat{\Lambda}\big)\,dx + \bar{\rho}\int_{\mathcal{X}} p\,\|\nabla\hat{\mu}\|_2^2\,dx.$$

Again take $\hat{\mu} \equiv y$ and $\hat{\Lambda} \equiv C > 0$. Then $\nabla\hat{\mu} = \nabla y \in L^2$ and $\nabla\hat{\Lambda} = 0$, so

$$\mathbf{S}_{\rho,1}[y, C] = -\frac{\rho}{2}\Big(\int_{\mathcal{X}} p(x)\,dx\Big)\log C + \bar{\rho}\int_{\mathcal{X}} p\,\|\nabla y\|_2^2\,dx.$$

The second term is finite and independent of C, while the first tends to $-\infty$ as $C \to \infty$. Thus the objective is unbounded below when $\gamma = 1$. □

Corollary 2 (Necessity of two-sided regularization for $\rho > 0$ (general p))**.** *Under the assumptions of Prop. 3 with $\rho > 0$, any well-posed formulation requires*

$$\gamma \in (0, 1) \quad \Longleftrightarrow \quad \alpha = \tfrac{\bar{\rho}}{\rho}\gamma > 0 \;\; \textit{and} \;\; \beta = \tfrac{\bar{\rho}}{\rho}\bar{\gamma} > 0.$$

Equivalently, if either $\alpha = 0$ or $\beta = 0$, the objective is unbounded below.

Proof. If $\alpha = 0$ (i.e., $\gamma = 0$), part (iii) of Prop. 3 shows that the functional $\mathbf{S}_{\rho,0}$ is unbounded below: choosing $\hat{\mu} \equiv y$ and $\hat{\Lambda} \equiv C$ with $C \to \infty$ nulls the residual term while the $-\log \hat{\Lambda}$ contribution drives $\mathbf{S}_{\rho,0}[y, C] \to -\infty$. Similarly, if $\beta = 0$ (i.e., $\bar{\gamma} = 0$), part (iv) shows that the same construction yields $\mathbf{S}_{\rho,1}[y, C] \to -\infty$. Thus in either case ($\alpha = 0$ or $\beta = 0$) the objective is unbounded below, proving the necessity of $\alpha, \beta > 0$ for well-posedness when $\rho > 0$. □

Proposition 4 (Existence of a minimizer for interior regularization weights)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a bounded, connected Lipschitz domain, and let $p \in C^1(\overline{\mathcal{X}})$ be strictly positive on $\overline{\mathcal{X}}$. Fix $\rho, \gamma \in (0, 1)$ with $\bar{\rho} = 1 - \rho$ and $\bar{\gamma} = 1 - \gamma$. Assume that the observed data field satisfies $y \in H^1(\mathcal{X})$ (and hence $y \in L^2(\mathcal{X})$).*

Fix constants $0 < \lambda_{\min} < \lambda_{\max} < \infty$, and define the admissible set

$$\mathcal{A} := \Big\{ (\hat{\mu}, \hat{\Lambda}) \in H^1(\mathcal{X}) \times H^1(\mathcal{X}) \,:\, \lambda_{\min} \le \hat{\Lambda}(x) \le \lambda_{\max} \text{ a.e. in } \mathcal{X} \Big\}.$$

Then the variational objective

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}] = \int_{\mathcal{X}} p(x) \Big[-\rho \log \hat{p}(y \,|\, x) + \bar{\rho}\big(\gamma \|\nabla \hat{\mu}(x)\|_2^2 + \bar{\gamma} \|\nabla \hat{\Lambda}(x)\|_2^2\big) \Big] dx,$$

where $\hat{p}(y \,|\, x) = \mathcal{N}(y \,|\, \hat{\mu}(x), \hat{\Lambda}(x)^{-1})$, admits at least one minimizer $(\hat{\mu}^*, \hat{\Lambda}^*) \in \mathcal{A}$.

Proof. We apply the direct method of the calculus of variations (see Evans [2010, Sec. 8.2]). Because p is continuous and strictly positive on the bounded domain $\mathcal{X}$, it is bounded above and below by positive constants, so the weighted and unweighted L^2 and H^1 norms are equivalent.

1. Coercivity. Writing $(\hat{\mu}, \hat{\Lambda})$ for a generic admissible pair, the functional can be expressed as

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu}, \hat{\Lambda}] = \int_{\mathcal{X}} p(x) \Big[\frac{\rho}{2} \big(\hat{\Lambda}(x)(\hat{\mu}(x) - y(x))^2 - \log \hat{\Lambda}(x) \big) + \bar{\rho}\big(\gamma \|\nabla \hat{\mu}(x)\|_2^2 + \bar{\gamma} \|\nabla \hat{\Lambda}(x)\|_2^2\big) \Big] dx.$$

Since $\hat{\Lambda}(x) \in [\lambda_{\min}, \lambda_{\max}]$ a.e., we have, for all $x \in \mathcal{X}$,

$$\hat{\Lambda}(x)(\hat{\mu}(x) - y(x))^2 - \log \hat{\Lambda}(x) \;\ge\; \lambda_{\min}(\hat{\mu}(x) - y(x))^2 - \log \lambda_{\max}.$$

Multiplying by $p(x)\frac{\rho}{2}$ and integrating over $\mathcal{X}$ yields

$$\begin{aligned} \int_{\mathcal{X}} p(x)\, \frac{\rho}{2} \big(\hat{\Lambda}(x)(\hat{\mu}(x) - y(x))^2 - \log \hat{\Lambda}(x) \big)\, dx & \\ \ge \int_{\mathcal{X}} p(x)\, \frac{\rho}{2} \big(\lambda_{\min}(\hat{\mu}(x) - y(x))^2 - \log \lambda_{\max} \big)\, dx & \\ = \frac{\rho \lambda_{\min}}{2} \int_{\mathcal{X}} p(x)\, (\hat{\mu}(x) - y(x))^2\, dx \;-\; \frac{\rho \log \lambda_{\max}}{2} \int_{\mathcal{X}} p(x)\, dx. & \end{aligned}$$

Since p is continuous and strictly positive on the bounded domain $\overline{\mathcal{X}}$, there exist constants $0 < p_{\min} \le p(x) \le p_{\max} < \infty$ for all $x \in \overline{\mathcal{X}}$. Using $p(x) \ge p_{\min}$, we obtain

$$\int_{\mathcal{X}} p(x)\, (\hat{\mu}(x) - y(x))^2\, dx \;\ge\; p_{\min} \int_{\mathcal{X}} (\hat{\mu}(x) - y(x))^2\, dx = p_{\min}\, \|\hat{\mu} - y\|_{L^2(\mathcal{X})}^2.$$

Thus the data term is bounded below by

$$\int_{\mathcal{X}} p(x)\, \frac{\rho}{2} \big(\hat{\Lambda}(\hat{\mu} - y)^2 - \log \hat{\Lambda} \big)\, dx \;\ge\; c_1 \|\hat{\mu} - y\|_{L^2(\mathcal{X})}^2 - C_0,$$

for some constants $c_1 > 0$ and $C_0 > 0$ depending on $(\rho, \lambda_{\min}, \lambda_{\max}, p)$.

For the gradient terms, we similarly have

$$\int_{\mathcal{X}} p(x)\,\gamma\|\nabla\hat{\mu}(x)\|_2^2\,dx \;\geq\; \gamma p_{\min}\int_{\mathcal{X}}\|\nabla\hat{\mu}(x)\|_2^2\,dx = \gamma p_{\min}\,\|\nabla\hat{\mu}\|_{L^2(\mathcal{X})}^2 \quad \text{and}$$
$$\int_{\mathcal{X}} p(x)\,\bar{\gamma}\|\nabla\hat{\Lambda}(x)\|_2^2\,dx \;\geq\; \bar{\gamma} p_{\min}\int_{\mathcal{X}}\|\nabla\hat{\Lambda}(x)\|_2^2\,dx = \bar{\gamma} p_{\min}\,\|\nabla\hat{\Lambda}\|_{L^2(\mathcal{X})}^2.$$

Combining these estimates, we obtain

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu},\hat{\Lambda}] \;\geq\; c_1\|\hat{\mu}-y\|_{L^2(\mathcal{X})}^2 + c_2\Big(\|\nabla\hat{\mu}\|_{L^2(\mathcal{X})}^2 + \|\nabla\hat{\Lambda}\|_{L^2(\mathcal{X})}^2\Big) - C,$$

for some constants $c_1, c_2, C > 0$ depending on $(\rho, \gamma, \lambda_{\min}, \lambda_{\max}, p)$ and y.

Since

$$\|\hat{\mu}\|_{L^2(\mathcal{X})} \leq \|\hat{\mu}-y\|_{L^2(\mathcal{X})} + \|y\|_{L^2(\mathcal{X})}$$

and

$$\|\hat{\Lambda}\|_{L^2(\mathcal{X})}^2 = \int_{\mathcal{X}}|\hat{\Lambda}(x)|^2\,dx \leq \lambda_{\max}^2|\mathcal{X}| \quad \text{for all admissible } \hat{\Lambda},$$

we can absorb these constants to obtain

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu},\hat{\Lambda}] \;\geq\; C_1\Big(\|\hat{\mu}\|_{H^1(\mathcal{X})}^2 + \|\hat{\Lambda}\|_{H^1(\mathcal{X})}^2\Big) - C_2,$$

for some $C_1, C_2 > 0$, where

$$\|\hat{\mu}\|_{H^1(\mathcal{X})}^2 := \|\hat{\mu}\|_{L^2(\mathcal{X})}^2 + \|\nabla\hat{\mu}\|_{L^2(\mathcal{X})}^2, \quad \|\hat{\Lambda}\|_{H^1(\mathcal{X})}^2 := \|\hat{\Lambda}\|_{L^2(\mathcal{X})}^2 + \|\nabla\hat{\Lambda}\|_{L^2(\mathcal{X})}^2.$$

Thus $\mathbf{S}_{\rho,\gamma}$ is coercive on $\mathcal{A}$.

2. Minimizing sequence and compactness. Let $(\hat{\mu}_n, \hat{\Lambda}_n) \in \mathcal{A}$ be a minimizing sequence, i.e.

$$\lim_{n\to\infty}\mathbf{S}_{\rho,\gamma}[\hat{\mu}_n,\hat{\Lambda}_n] = \inf_{(\hat{\mu},\hat{\Lambda})\in\mathcal{A}}\mathbf{S}_{\rho,\gamma}[\hat{\mu},\hat{\Lambda}].$$

By coercivity (Step 1), the sequence $(\hat{\mu}_n, \hat{\Lambda}_n)$ is bounded in $H^1(\mathcal{X}) \times H^1(\mathcal{X})$. Since $H^1(\mathcal{X})$ is a Hilbert (hence reflexive) space, every bounded sequence has a weakly convergent subsequence. Passing to such a subsequence (not relabeled), there exists $(\hat{\mu}^*, \hat{\Lambda}^*) \in H^1(\mathcal{X})^2$ such that

$$\hat{\mu}_n \rightharpoonup \hat{\mu}^* \quad \text{and} \quad \hat{\Lambda}_n \rightharpoonup \hat{\Lambda}^* \quad \text{weakly in } H^1(\mathcal{X}).$$

Weak convergence controls functions and their gradients only in an averaged sense, which is insufficient to pass to the limit in the nonlinear term $\hat{\Lambda}(\hat{\mu}-y)^2$. To obtain pointwise and L^2 convergence, we use a standard compactness result: on any bounded Lipschitz domain, the Sobolev embedding $H^1(\mathcal{X}) \hookrightarrow L^2(\mathcal{X})$ is *compact* (Rellich–Kondrachov). Thus, up to a further subsequence,

$$\hat{\mu}_n \to \hat{\mu}^*, \qquad \hat{\Lambda}_n \to \hat{\Lambda}^* \quad \text{strongly in } L^2(\mathcal{X}),$$

and in particular almost everywhere on $\mathcal{X}$.

Because each $(\hat{\mu}_n, \hat{\Lambda}_n)$ lies in the admissible set $\mathcal{A}$, we have the pointwise bounds $\lambda_{\min} \leq \hat{\Lambda}_n(x) \leq \lambda_{\max}$ a.e. Strong L^2 convergence implies almost-everywhere convergence along the subsequence, so these bounds pass to the limit:

$$\lambda_{\min} \leq \hat{\Lambda}^*(x) \leq \lambda_{\max} \quad \text{for a.e. } x.$$

Hence $(\hat{\mu}^*, \hat{\Lambda}^*) \in \mathcal{A}$.

3. Lower semicontinuity and passing to the limit. The gradient regularizers

$$\int_{\mathcal{X}} p(x)\,\gamma\|\nabla\hat{\mu}(x)\|_2^2\,dx, \qquad \int_{\mathcal{X}} p(x)\,\bar{\gamma}\|\nabla\hat{\Lambda}(x)\|_2^2\,dx$$

are weakly lower semicontinuous in $H^1(\mathcal{X})$, as they are convex quadratic forms in the gradients.

For the data-fitting term, the strong L^2 convergence of $\hat{\mu}_n$ and $\hat{\Lambda}_n$, together with the uniform bounds $\lambda_{\min} \le \hat{\Lambda}_n \le \lambda_{\max}$ and continuity of the map

$$(u,\lambda) \mapsto \lambda(u-y)^2 - \log\lambda \quad \text{on } \mathbb{R}\times[\lambda_{\min},\lambda_{\max}],$$

implies

$$\int_{\mathcal{X}} p(x)\left[\hat{\Lambda}_n(x)(\hat{\mu}_n(x)-y(x))^2 - \log\hat{\Lambda}_n(x)\right]dx \tag{24}$$

$$\longrightarrow \int_{\mathcal{X}} p(x)\left[\hat{\Lambda}^*(x)(\hat{\mu}^*(x)-y(x))^2 - \log\hat{\Lambda}^*(x)\right]dx. \tag{25}$$

Thus the data term is continuous along the minimizing subsequence.

Combining the weak lower semicontinuity of the gradient terms with the continuity of the data term yields

$$\mathbf{S}_{\rho,\gamma}[\hat{\mu}^*,\hat{\Lambda}^*] \;\le\; \liminf_{n\to\infty} \mathbf{S}_{\rho,\gamma}[\hat{\mu}_n,\hat{\Lambda}_n] = \inf_{(\hat{\mu},\hat{\Lambda})\in\mathcal{A}} \mathbf{S}_{\rho,\gamma}[\hat{\mu},\hat{\Lambda}].$$

Thus $(\hat{\mu}^*,\hat{\Lambda}^*)$ attains the infimum of $\mathbf{S}_{\rho,\gamma}$ over $\mathcal{A}$. This completes the existence proof via the direct method of the calculus of variations; see Evans [2010, Sec. 8.2, Thm. 2]. □

Remark 3 (Interpretation of Bounded Precision Assumption)**.** *The restriction $\lambda_{\min} \le \hat{\Lambda}(x) \le \lambda_{\max}$ ensures that the predicted precision (inverse variance) remains within a physically and statistically meaningful range. From a modeling standpoint, this prevents pathological behavior:*

- *Allowing $\hat{\Lambda}(x) \to 0$ corresponds to arbitrarily large predictive variance, i.e., extreme uncertainty, which is typically uninformative and numerically unstable.*
- *Allowing $\hat{\Lambda}(x) \to \infty$ implies vanishing predictive variance, i.e., extreme overconfidence, even in regions with limited or noisy data—this can lead to poor generalization.*

It is important to note that this assumption is made purely for mathematical tractability: the existence of such bounds is sufficient for the argument, and they need not be tight. For instance, one may take $\lambda_{\min} = 2^{-100}$, $\lambda_{\max} = 2^{100}$ and the proof still holds.

B Bayesian Field Theory (Supplementary Derivations)

This section provides the explicit mathematical formulation and discretization details for the Bayesian reformulation of the field theory (BFT) introduced in Section 5. It formalizes the two parameterizations of the variance field, derives the MAP and functional gradients, and outlines the weak convergence of the corresponding discretized priors.

B.1 MAP-equivalent parameterization

Let $\bar{\rho} := 1-\rho$ and $\bar{\gamma} := 1-\gamma$. We parameterize $\hat{\Lambda} = e^{\hat{\eta}}$ to enforce $\hat{\Lambda} > 0$ and choose priors so that the deterministic FT regularizers are recovered exactly at the posterior mode, making the MAP equations identical to those of the FT.

B.1.1 Priors

$$-\log \pi(\hat{\mu}) = \tfrac{\gamma}{2} \int_{\mathcal{X}} p(x)\, \|\nabla \hat{\mu}(x)\|^2\, dx, \tag{26}$$

$$-\log \pi(\hat{\eta}) = \tfrac{\bar{\gamma}}{2} \int_{\mathcal{X}} p(x)\, e^{2\hat{\eta}(x)}\, \|\nabla \hat{\eta}(x)\|^2\, dx. \tag{27}$$

Homogeneous Neumann boundary conditions $p\nabla\hat{\mu}\cdot\mathbf{n} = p\, e^{2\hat{\eta}}\nabla\hat{\eta}\cdot\mathbf{n} = 0$ ensure vanishing boundary terms. These priors act as Gaussian Markov random field (GMRF) and log-convex field priors, respectively, enforcing smoothness and positivity.

B.1.2 Scaling conventions between FT and BFT

The field–theoretic functional in Eq. (7) omits the global $\frac{1}{2}$ factor in its Dirichlet energies for notational simplicity and consistency with our simulations. In the Bayesian Field Theory (BFT) formulation, an equivalent $\frac{1}{2}$ appears in the Gaussian prior and likelihood log–densities (see Eq. (13)). This factor can be absorbed into the definition of the prior precision or covariance, so it merely rescales the regularization weights without changing the stationary equations. Consequently, the posterior mode and the field–theoretic stationary conditions remain identical, and the equivalence MAP $\equiv$ FT holds exactly.

B.1.3 Posterior and functional gradients

Combining likelihood and priors gives

$$\Phi(\hat{\mu},\hat{\eta}) = \int p \left\{ \rho\big[\tfrac{1}{2} e^{\hat{\eta}}(y-\hat{\mu})^2 - \tfrac{1}{2}\hat{\eta}\big] + \tfrac{\bar{\rho}}{2}\big[\gamma\|\nabla\hat{\mu}\|^2 + \bar{\gamma}\, e^{2\hat{\eta}}\|\nabla\hat{\eta}\|^2\big] \right\}. \tag{28}$$

Its gradients (before discretization) are:

$$\nabla_{\hat{\mu}}\Phi = \rho\, p\, e^{\hat{\eta}}(\hat{\mu}-y) - \bar{\rho}\,\gamma\,\nabla\cdot(p\nabla\hat{\mu}), \tag{29}$$

$$\nabla_{\hat{\eta}}\Phi = \tfrac{\rho p}{2}\Big(e^{\hat{\eta}}(y-\hat{\mu})^2 - 1\Big) - \bar{\rho}\,\bar{\gamma}\,\nabla\cdot(p\, e^{2\hat{\eta}}\nabla\hat{\eta}) + \bar{\rho}\,\bar{\gamma}\, p\, e^{2\hat{\eta}}\|\nabla\hat{\eta}\|^2. \tag{30}$$

At $\nabla\Phi = 0$, these coincide with the FT Euler–Lagrange equations.

Note on normalization The prior on $\hat{\eta}$ guarantees $\hat{\Lambda} > 0$ but leaves constant (mean) modes unpenalized. To ensure a proper posterior, one may add a small stabilizer term $\frac{\varepsilon}{2}\int p(\hat{\eta}-\hat{\eta}_0)^2\, dx$, which normalizes the prior without affecting the stationary equations or MAP solution.

B.2 Discretization and weak convergence

Let $\mathcal{G}_h$ denote a shape–regular mesh of $\mathcal{X}$ with mesh size $h = \max_{K\in\mathcal{G}_h} \mathrm{diam}(K) \to 0$. Shape–regularity means that all elements (triangles, tetrahedra, or grid cells) have uniformly bounded aspect ratio, i.e., no element becomes arbitrarily thin or degenerate as the mesh is refined. This ensures numerical stability of the discrete gradient and Laplacian operators.

The mesh need not be uniform. In our setting, the node density of $\mathcal{G}_h$ is proportional to the sampling density $p(x)$, corresponding to the empirical discretization used throughout the field theory. Regions of higher $p(x)$ therefore receive finer resolution, while maintaining shape–regularity in each local neighborhood.

Denote by $\mathbf{L}_{p,h}$ the standard finite–element (or weighted graph) discretization of the weighted Laplacian

$$\mathcal{L}_p f = -\nabla\cdot(p\nabla f)$$

with homogeneous Neumann boundary conditions (zero normal flux on boundary facets). Let ∇_h be the discrete gradient operator and $\mathbf{M}_{p,h}$ the (possibly lumped) p–weighted mass matrix. The discrete priors then read

$$-\log\pi(\hat{\boldsymbol{\mu}}) = \tfrac{\gamma}{2}\,\hat{\boldsymbol{\mu}}^\top \mathsf{L}_{p,h}\,\hat{\boldsymbol{\mu}}, \qquad -\log\pi(\hat{\boldsymbol{\eta}}) = \tfrac{\tilde{\gamma}}{2}\left(\mathrm{e}^{\hat{\boldsymbol{\eta}}}\odot\nabla_h\hat{\boldsymbol{\eta}}\right)^\top \mathbf{M}_{p,h}\left(\mathrm{e}^{\hat{\boldsymbol{\eta}}}\odot\nabla_h\hat{\boldsymbol{\eta}}\right),$$

where $\odot$ denotes elementwise multiplication.

Under standard finite–element assumptions (shape–regularity, bounded domain, and $h\to 0$), these discrete priors converge weakly to their continuous counterparts. If the node density of $\mathcal{G}_h$ tracks $p(x)$, then $\mathbf{L}_{p,h}$ and $\mathbf{M}_{p,h}$ provide consistent approximations of the weighted operators associated with $\mathcal{L}_p$. This includes the graph–based discretizations used in practice, which can be viewed as stochastic FEM quadratures under the empirical measure $p(x)\,dx$. For rigorous convergence results in this setting, see Lindgren et al. [2011], who prove weak convergence of discrete GMRFs and SPDE operators to their continuous Matérn and elliptic limits.

B.3 Gaussian Field–Process Equivalence

A Gaussian random field (GRF) $f(x)$ is a Gaussian measure over functions $f:\mathcal{X}\to\mathbb{R}$ specified by a precision (inverse-covariance) operator $\mathcal{L}$. Formally, the prior density (up to normalization) is

$$p(f) \;\propto\; \exp\!\left[-\tfrac{1}{2}\int_{\mathcal{X}} f(x)\,(\mathcal{L}f)(x)\,dx\right], \tag{31}$$

where $\mathcal{L}$ is typically a positive-definite elliptic operator such as the (weighted) Laplacian $\mathcal{L}_p f = -\nabla\cdot(p(x)\nabla f(x))$.

The Green's function $G(x,x')$ of $\mathcal{L}$ satisfies $\mathcal{L}G(x,\cdot\,) = \delta(x-\cdot\,)$, and defines the covariance kernel

$$k(x,x') \;=\; G(x,x') \;=\; \big(\mathcal{L}^{-1}\big)(x,x'). \tag{32}$$

Hence a GRF with precision $\mathcal{L}$ is equivalent to a Gaussian process $\mathcal{GP}(0,k)$ whose kernel is the inverse of $\mathcal{L}$:

$$f\sim\mathcal{N}(0,\mathcal{L}^{-1}) \quad\Longleftrightarrow\quad f(x)\sim\mathcal{GP}(0,k(x,x')).$$

For example, the Dirichlet-energy prior used throughout the field theory,

$$p(\hat{\mu})\propto\exp\Big[-\tfrac{\gamma}{2}\int_{\mathcal{X}} p(x)\,\|\nabla\hat{\mu}(x)\|^2 dx\Big], \tag{33}$$

is a zero-mean GRF with precision operator $\gamma\,\mathcal{L}_p$ and thus corresponds to a GP with covariance operator $(\gamma\,\mathcal{L}_p)^{-1}$. Consequently, the Dirichlet energy acts as the negative log-density of a Gaussian process whose kernel is the Green's function of the weighted Laplacian $\mathcal{L}_p$:

$$k_p(x,x') = \gamma^{-1}G_p(x,x'), \qquad \mathcal{L}_p G_p(x,\cdot\,) = \delta(x-\cdot\,).$$

This operator-based Gaussian prior is consistent with the classical RKHS–Bayesian correspondence of Kimeldorf and Wahba [1970], in which quadratic smoothness penalties correspond to Gaussian priors whose covariance operators are the inverses of the associated differential operators.

This is a concrete instance of the general operator–kernel correspondence used to construct Gaussian fields from elliptic SPDEs, including the Matérn family [Lindgren et al., 2011], and it links our field-theoretic prior directly to GP and RKHS smoothness priors.

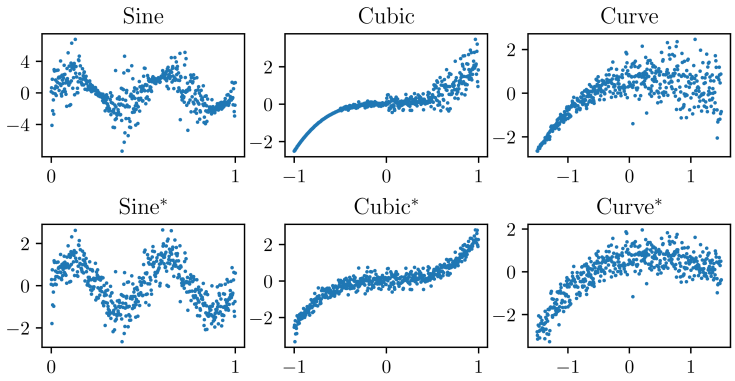


Figure 6: Visualization of heteroskedastic and homoskedastic versions of simulated datasets. Specific details for the functional form of these can be found in Table 2.

Table 2: Simulated datasets. Each dataset is defined by a true μ function and then a noise function f. All data is generated as $\mu(x) + \epsilon(x)$ where $\epsilon(x) \sim \mathcal{N}(0, f(x)^2)$. After the datasets were generated they were scaled to have mean zero and standard deviation one. The homoskedastic versions of each dataset fix $f(x) = 1$. The datasets are shown in Fig. 6.

Dataset	Mean (μ)	Noise Pattern (f)	Domain
Sine	$\mu(x) = 2\sin(4\pi x)$	$f(x) = \sin(6\pi x) + 1.25$	$x \in [0, 1]$
Cubic	$\mu(x) = x^3$	$f(x) = \begin{cases} 0.1 & \text{for } x < -0.5 \\ 1 & \text{for } x \in [-0.5, 0.0) \\ 3 & \text{for } x \in [0.0, 0.5) \\ 10 & \text{for } x \geq .5 \end{cases}$	$x \in [-1, 1]$
Curve	$\mu(x) = x - 2x^2 + 0.5x^3$	$f(x) = x + 1.5$	$x \in [-1.5, 1.5]$

C Experimental Details

C.1 Datasets

We chose 64 datapoints in each of the simulated datasets. The generating processes for each simulated dataset is included in Table 2 and can be seen in Fig. 6. The homoskedastic data is simulated in the same way, but with $f(x) = 1$. For testing, we simulate a new dataset of 64 datapoints with the same process. Table 3 summarizes the UCI datasets. We provide a description of the ClimSim climate data in Appendix E.4.

Table 3: UCI dataset details.

Dataset	Train Size	Test Size	Input Dimension
Concrete	687	343	8
Housing	337	168	13
Power	6379	3189	4
Yacht	204	102	6

C.2 Training Details

For the neural-network experiments, we evaluated 22 values of (ρ, γ) spaced log-uniformly on the logit scale between 10^{-10} and $1 - 10^{-5}$. For the field-theoretic models we used 20 values between 10^{-6} and $1 - 10^{-7}$, again logit-spaced. The ranges differ slightly due to numerical stability during fitting. Along the line $\rho = 1 - \gamma$, we sampled 100 values between 10^{-11} and 1, using a dense log-uniform grid near the extremes and a uniform grid on $[0.1, 0.9]$. The limiting cases $\rho, \gamma \in \{0, 1\}$ were excluded for stability. The exact grids used for all sweeps are provided in the public repository. All experiments were run on Nvidia Quadro RTX 8000 GPUs, totaling approximately 500 GPU hours.

C.3 Field Theory

For the discretized field theory we take $n_{ft} = 4096$ evenly spaced points on the interval $[-1, 1]$. There are two datapoints placed beyond $[-1, 1]$ because the method we use to estimate the gradients requires the datapoints to have left and right neighbors. These datapoints were not included when computing our metrics. Of these 4096 datapoints 64 were randomly selected to be used for training neural networks $\hat{\mu}_\theta, \hat{\Lambda}_\phi$. The field theory results were consistent across choices of $n_{ft} \in \{256, 512, 1024, 2048, 4096\}$. We present results for $n_{ft} = 4096$ in the main paper. We train for 100000 epochs and use the Adam optimizer with a basic triangular cycle that scales initial amplitude by half each cycle on the learning rate. The minimum and maximum learning rates were 0.0005 and 0.01. The cycles were 5000 epochs long. We clip the gradients at 1000.

C.4 Neural Network Training for Synthetic and UCI Data

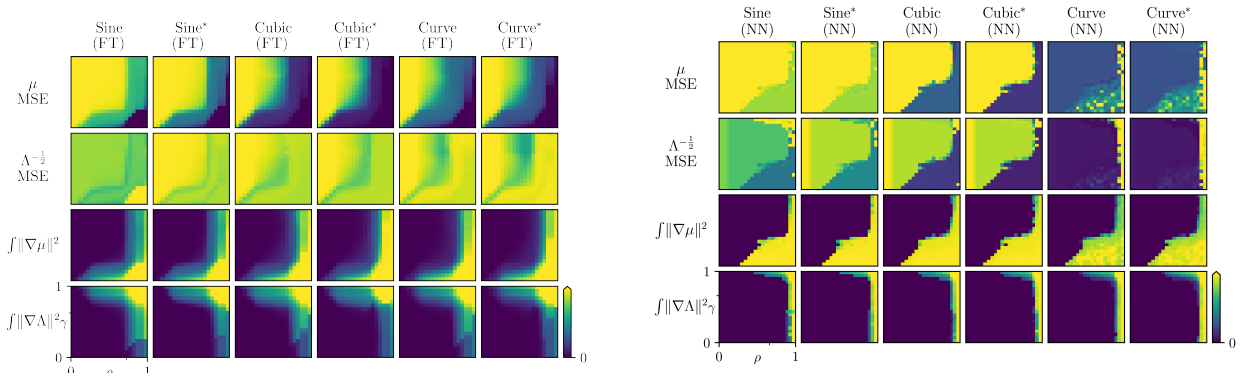
Across all experiments we train the networks using Adam with a triangular cyclic learning rate schedule (min. 0.0001, max. 0.01), where each cycle reduces its amplitude by half. Unless otherwise noted, the cycle length is 50,000 epochs and gradients are clipped at 1000. Training proceeds in two phases: we first train only the mean network $\hat{\mu}_\theta$ for the initial portion of training, and then train both $\hat{\mu}_\theta$ and $\hat{\Lambda}_\phi$ for the remainder.

Synthetic datasets. For all synthetic datasets except *Sine*, we train for 600,000 epochs: the first 250,000 epochs train only $\hat{\mu}_\theta$, followed by 350,000 epochs training both networks. For the *Sine* dataset, we use the same setup but train for 2,500,000 epochs total.

UCI datasets. For *Concrete*, *Housing*, and *Yacht*, we train for 500,000 epochs: 250,000 epochs on $\hat{\mu}_\theta$ alone and 250,000 epochs jointly on $\hat{\mu}_\theta$ and $\hat{\Lambda}_\phi$, using the same learning rate schedule. For the *Power* dataset, minibatching is required due to its size. We use a batch size of 1000 and train for 50,000 epochs in total, with 25,000 epochs devoted to $\hat{\mu}_\theta$ alone and the remaining 25,000 epochs training both networks. The same cyclic learning-rate schedule is used, but with a reduced cycle length of 5000.

D Additional Results

Both FT and neural networks were fit to the heteroskedastic and homoskedastic synthetic datasets described in Table 2. The main results for these displayed as phase diagrams of various metrics can be seen in Fig. 7a and Fig. 7b respectively. We largely see the same trends as were exhibited by the real-world datasets seen in Fig. 3.



(a) Field theory.

(b) Neural network.

Figure 7: Phase diagrams for the field theory (*left*) and neural networks (*right*) on six synthetic datasets, in the same configuration as Fig. 3. Dataset names with an $*$ denote homoskedastic counterparts.

E Comparison to Baselines

We compare the performance of our diagonal $\rho = 1 - \gamma$ search against two baselines, β-NLL [Seitzer et al., 2022] and an ensemble of six MLE-fit heteroskedastic regression models [Lakshminarayanan et al., 2017]. We use μ MSE, $\Lambda^{-\frac{1}{2}}$ MSE, and expected calibration error (ECE) to evaluate the models. In all cases lower values are better. Note that the method of ensembling multiple individual MLE-fit models from Lakshminarayanan et al. [2017] could be implemented on our method or β-NLL as well.

E.1 Model Architecture

All (individual) models have the same architecture: fully connected neural networks with three hidden layers of 128 nodes and leaky ReLU activations for the synthetic and UCI datasets and fully connected neural networks with three hidden layers of 256 nodes for the ClimSim data [Yu et al., 2023]. Note that both baselines model the variance while our approach models the precision (inverse-variance). In all cases we use a softplus on the final layer of the variance/precision networks to ensure the output is positive.

For the $\beta-$NLL implementation we take $\beta = 0.5$ as suggested in Seitzer et al. [2022]. The ensemble method we use fits 6 individual heteroskedastic neural networks and combines their outputs into a mixture distribution that is approximated with a normal distribution. We do not add in adversarial noise as the authors state it does not make a significant difference. We fit six $\beta-$NLL models and six MLE-ensembles.

E.2 Diagonal selection criteria

After conducting our diagonal search we found the model that minimized μ MSE and the model that minimized $\Lambda^{-\frac{1}{2}}$ MSE on the *training* data. In some cases these models coincided. We then used the model that was on the midpoint (on a logit scale) of the $\rho = 1 - \gamma$ line between these two models to compare. The results are reported in Table 4. In all cases our method is competitive with or exceeds the performance of these two baselines–particularly on real-world data. Note that our goal is to show that we are able to find models that model the mean and standard deviation of the data well, that is, lie in our proposed region S of the phase diagram. We do not claim that this

method will provide the globally optimal model.

E.3 Training Details

Training for our method follows Appendices C.2 and C.4. For the baselines, we use the same optimizer, gradient clipping, and cyclic learning-rate schedule. On synthetic datasets we train for 600,000 epochs, and on the UCI datasets we train for 500,000 epochs (or 50,000 epochs with a batch size of 1000 for *Power* due to its size).

E.4 ClimSim Dataset

The ClimSim dataset [Yu et al., 2023] is a largescale climate dataset. Its input dimension is 124 and output dimension is 128. We use all 124 inputs to model a single output, *Visible direct solar flux, SOLS [W/m^2]*. We train on 10,091,520 of the approximately 100 million points for training and we use a randomly selected 7,209 points to evaluate our models.

Table 4: Comparison of our deep heteroskedastic regression model (with diagonal regularization search; see Appendix E.2) against two baselines [Seitzer et al., 2022, Lakshminarayanan et al., 2017]. We report the mean ± standard deviation of ECE, μ-MSE, and $\Lambda^{-1/2}$-MSE on test data, with the best mean in bold. MLE ensembles often fail to converge—typically through divergence of thepredicted standard deviation—producing inf or `nan` values and illustrating the instability of MLE-based heteroskedastic training.

Dataset Metric	Mean-Variance Ours	β-NLL Seitzer et al. [2022]	MLE Ensemble Lakshminarayanan et al. [2017]
Cubic			
ECE	**0.2380** ± 0.03	0.2385 ± 0.02	0.2411 ± 0.02
μ MSE	0.2339 ± 0.01	**0.1500** ± 0.01	1.1809 ± 1.88
$\Lambda^{-\frac{1}{2}}$MSE	0.2397 ± 0.02	**0.1397** ± 0.01	inf ± nan
Curve			
ECE	0.1804 ± 0.02	**0.1754** ± 0.02	0.2432 ± 0.00
μ MSE	**0.4318** ± 0.12	0.4877 ± 0.16	1.0067 ± 0.19
$\Lambda^{-\frac{1}{2}}$MSE	0.4655 ± 0.09	**0.4187** ± 0.20	inf ± nan
Sine			
ECE	0.2499 ± 0.00	**0.2082** ± 0.03	0.2313 ± 0.05
μ MSE	**0.7968** ± 0.00	4.4107 ± 6.90	0.9716 ± 0.06
$\Lambda^{-\frac{1}{2}}$MSE	**0.7968** ± 0.00	4.3524 ± 6.89	inf ± nan
Concrete			
ECE	0.2471 ± 0.01	0.2552 ± 0.00	**0.0655** ± 0.01
μ MSE	**0.1055** ± 0.02	0.5461 ± 0.30	2.2454 ± 1.74
$\Lambda^{-\frac{1}{2}}$MSE	**0.3028** ± 0.51	1.0867 ± 0.20	$1.3 \times 10^5 \pm 1.2 \times 10^5$
Housing			
ECE	0.0653 ± 0.00	**0.2631** ± 0.01	0.1332 ± 0.02
μ MSE	**1.2236** ± 0.00	0.3175 ± 0.06	155.4494 ± 128.27
$\Lambda^{-\frac{1}{2}}$MSE	**0.7610** ± 0.00	0.8820 ± 0.03	218.8269 ± 195.38
Power			
ECE	0.2233 ± 0.01	0.2370 ± 0.00	**0.0285** ± 0.01
μ MSE	0.0350 ± 0.01	0.1013 ± 0.01	**0.0177** ± 0.00
$\Lambda^{-\frac{1}{2}}$MSE	0.0343 ± 0.01	0.3081 ± 0.37	**0.0091** ± 0.00
Yacht			
ECE	0.3038 ± 0.04	0.2882 ± 0.02	**0.0463** ± 0.02
μ MSE	**0.0077** ± 0.01	0.0137 ± 0.01	6.2670 ± 13.96
$\Lambda^{-\frac{1}{2}}$MSE	**0.0076** ± 0.01	1.3275 ± 0.02	8.0599 ± 19.18
Solar Flux			
ECE	**0.1503** ± 0.00	0.3007 ± 0.00	0.1924 ± 0.04
μ MSE	**0.2887** ± 0.00	0.3771 ± 0.00	1.0067 ± 0.19
$\Lambda^{-\frac{1}{2}}$MSE	**0.1175** ± 0.00	0.3217 ± 0.00	$4.6 \times 10^9 \pm 9.9 \times 10^9$

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, December 2021. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008.

Alexander Altland and Ben D. Simons. *Condensed Matter Field Theory*. Cambridge University Press, Cambridge, 2 edition, 2010. ISBN 978-0-521-76975-4. doi: 10.1017/CBO9780511789984.

Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová. Glassy nature of the hard phase in inference problems. *Phys. Rev. X*, 9:011020, Jan 2019. doi: 10.1103/PhysRevX.9.011020.

Robert Bamler and Stephan Mandt. Improving optimization for models with continuous symmetry breaking. In *International Conference on Machine Learning*, pages 423–432. PMLR, 2018.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. In G. Goos, J. Hartmanis, J. Van Leeuwen, David Helmbold, and Bob Williamson, editors, *Computational Learning Theory*, volume 2111, pages 224–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-42343-0 978-3-540-44581-4. doi: 10.1007/3-540-44581-1_15. Series Title: Lecture Notes in Computer Science.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Christopher Bishop and Cazhaow Quazaz. Regression with Input-Dependent Noise: A Bayesian Treatment. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

Christopher M. Bishop. Mixture density networks. *Mixture density networks*, 1994. ISSN NCRG/94/004. Place: Birmingham Publisher: Aston University.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR.

Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 2008. ISBN 978-0-387-75933-3 978-0-387-75934-0. doi: 10.1007/978-0-387-75934-0.

Nicki Skafte Detlefsen, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Benoit Dherin, Michael Munn, Mihaela Rosca, and David GT Barrett. Why neural networks find simple solutions: The many regularizers of geometric complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Eberhard Engel and Reiner M. Dreizler. *Density Functional Theory: An Advanced Course*. Theoretical and Mathematical Physics. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-14089-1 978-3-642-14090-7. doi: 10.1007/978-3-642-14090-7.

R. L. Eubank and Will Thomas. Detecting Heteroscedasticity in Nonparametric Regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):145–155, 1993. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1993.tb01474.x.

Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2nd edition, 2010.

Bengt Fornberg. Generation of Finite Difference Formulas on Arbitrarily Spaced Grids. *Mathematics of Computation*, 51:699–706, October 1988.

Vincent Fortuin, Mark Collier, Florian Wenzel, James Allingham, Jeremiah Liu, Dustin Tran, Balaji Lakshminarayanan, Jesse Berent, Rodolphe Jenatton, and Effrosyni Kokiopoulou. Deep Classifiers with Label Noise Modeling and Distance Awareness. *Transactions on Machine Learning Research*, 2022.

Silvio Franz and Giorgio Parisi. The simplest model of jamming. *Journal of Physics A: Mathematical and Theoretical*, 49(14):145001, February 2016. doi: 10.1088/1751-8113/49/14/145001. Publisher: IOP Publishing.

Yarin Gal et al. Uncertainty in deep learning. 2016.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Phys. Rev. E*, 100:012115, Jul 2019. doi: 10.1103/PhysRevE.100.012115.

Erin George, Michael Murray, William Joseph Swartworth, and Deanna Needell. Training shallow reLU networks on noisy data using hinge loss: when do we overfit and is it benign? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

J. Gerritsma. Geometry, resistance and stability of the Delft Systematic Yacht hull series. *TU Delft, Faculty of Marine Technology, Ship Hydromechanics Laboratory, Report No. 520-P, Published in: International Shipbuilding Progress, ISP, Delft, The Netherlands, Volume 28, No. 328, also 7th HISWA Symposium, Amsterdam, The Netherlands*, 1981.

David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, March 1978. ISSN 0095-0696. doi: 10.1016/0095-0696(78)90006-2.

L.U. Hjorth and I.T. Nabney. Regularisation of mixture density networks. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, volume 2, pages 521–526 vol.2, September 1999. doi: 10.1049/cp:19991162. ISSN: 0537-9989.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, January 1991. ISSN 0893-6080. doi: 10.1016/0893-6080(91)90009-T.

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 5.1:221–234, January 1967.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Alexander Immer, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Effective bayesian heteroscedastic regression with deep neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI Machine Learning Repository. URL `https://archive.ics.uci.edu`.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Savya Khosla, Chew Kin Whye, Jordan T. Ash, Cyril Zhang, Kenji Kawaguchi, and Alex Lamb. Neural Active Learning on Heteroskedastic Distributions, November 2022.

George S. Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970. ISSN 00034851, 21688990.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Fred Kronz and Tracy Lupher. Quantum Theory and Mathematical Rigor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2025 edition, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Neural Information Processing Systems*, 2017.

L. D. Landau and E. M. Lifshitz. *Statistical Physics: Volume 5*. Elsevier, October 2013. ISBN 978-0-08-057046-4.

Quoc V. Le, Alex J. Smola, and Stéphane Canu. Heteroscedastic Gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 489–496, Bonn, Germany, 2005. ACM Press. ISBN 978-1-59593-180-1. doi: 10.1145/1102351.1102413.

J. C. Lemm. Bayesian Field Theory: Nonparametric Approaches to Density Estimation, Regression, Classification, and Inverse Quantum Problems, March 2000. arXiv:physics/9912005.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and Calibrating Uncertainty Prediction in Regression Tasks. *Sensors (Basel, Switzerland)*, 22(15):5540, July 2022. ISSN 1424-8220. doi: 10.3390/s22155540.

Kim-Hung Li and Nai Ng Chan. Degeneracy in Heteroscedastic Regression Models. *Journal of Multivariate Analysis*, 74(2):282–295, 2000. ISSN 0047-259X. doi: https://doi.org/10.1006/jmva.1999.1887.

Finn Lindgren, Håvard Rue, and Johan Lindström. An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4):423–498, September 2011. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2011.00777.x.

Finn Lindgren, David Bolin, and Håvard Rue. The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics*, 50:100599, 2022. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2022.100599.

David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, May 1992. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1992.4.3.448.

Stephan Mandt, James McInerney, Farhan Abrol, Rajesh Ranganath, and David Blei. Variational tempering. In *Artificial intelligence and statistics*, pages 704–712. PMLR, 2016.

Amitha Mathew, P. Amudha, and S. Sivakumari. Deep Learning Techniques: An Overview. In Aboul Ella Hassanien, Roheet Bhatnagar, and Ashraf Darwish, editors, *Advanced Machine Learning Technologies and Applications*, Advances in Intelligent Systems and Computing, pages 599–608, Singapore, 2021. Springer. ISBN 9789811533839. doi: 10.1007/978-981-15-3383-9_54.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Mass. London, 2012. ISBN 978-0-262-01825-8.

Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 1995.

D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, June 1994a. doi: 10.1109/ICNN.1994.374138.

David Nix and Andreas Weigend. Learning Local Error Bars for Nonlinear Regression. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994b.

Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, and Inbar Seroussi. Applications of Statistical Field Theory in Deep Learning, April 2025. arXiv:2502.18553 [stat].

Valentina Ros, Gerard Ben Arous, Giulio Biroli, and Chiara Cammarota. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Phys. Rev. X*, 9:011003, Jan 2019. doi: 10.1103/PhysRevX.9.011003.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, April 2009. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2008.00700.x.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022.

Megh Shukla, Mathieu Salzmann, and Alexandre Alahi. TIC-TAC: A framework for improved covariance estimation in deep heteroscedastic regression. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45244–45257. PMLR, 21–27 Jul 2024.

Megh Shukla, Aziz Shameem, Mathieu Salzmann, and Alexandre Alahi. Towards self-supervised covariance estimation in deep heteroscedastic regression. In *The Thirteenth International Conference on Learning Representations*, 2025.

Laurens Sluijterman, Eric Cator, and Tom Heskes. Optimal training of Mean Variance Estimation neural networks. *Neurocomputing*, 597:127929, September 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.127929.

Andrew Stirn and David A. Knowles. Variational Variance: Simple, Reliable, Calibrated Heteroscedastic Noise Variance Parameterization, October 2020.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5593–5613. PMLR, 25–27 Apr 2023.

Pınar Tüfekci. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140, September 2014. ISSN 0142-0615. doi: 10.1016/j.ijepes.2014.02.027.

Stanislaus S. Uyanto. Monte Carlo power comparison of seven most commonly used heteroscedasticity tests. *Communications in Statistics - Simulation and Computation*, 51(4):2065–2082, April 2022. ISSN 0361-0918. doi: 10.1080/03610918.2019.1692031.

Vladimir Vapnik and Rauf Izmailov. Complete statistical theory of learning: learning using statistical invariants. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin, editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 4–40. PMLR, 09–11 Sep 2020. URL `https://proceedings.mlr.press/v128/vapnik20a.html`.

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks*. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114008, November 2023. ISSN 1742-5468. doi: 10.1088/1742-5468/ad01b1.

G. Wahba. Spline models for observational data. *Regional Conference Series in Applied Mathematics*, 59, 1990.

Kevin Wang, Hongqian Niu, Yixin Wang, and Didong Li. Deep Generative Models: Complexity, Dimensionality, and Approximation. *Journal of Machine Learning Research*, 26(143):1–37, 2025.

Yixin Wang, Alp Kucukelbir, and David M. Blei. Robust probabilistic modeling with Bayesian data reweighting. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3646–3655, Sydney, NSW, Australia, August 2017. JMLR.org.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, 2011.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020.

Eliot Wong-Toi, Alex James Boyd, Vincent Fortuin, and Stephan Mandt. Understanding pathologies of deep heteroskedastic regression. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

David Xing Wu and Anant Sahai. Precise asymptotic generalization for multiclass classification with overparameterized linear models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Paul Yau and Robert Kohn. Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, 13(3):191–208, August 2003. ISSN 1573-1375. doi: 10.1023/A:1024293931757.

I-Cheng Yeh. Concrete Compressive Strength, 2007.

Sungduk Yu, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius Busecke, Nora Loose, Charles Stern, Tom Beucler, Bryce Harrop, Benjamin Hillman, Andrea Jenney, Savannah L. Ferretti, Nana Liu, Animashree Anandkumar, Noah Brenowitz, Veronika Eyring, Nicholas Geneva, Pierre Gentine, Stephan Mandt, Jaideep Pathak, Akshay Subramaniam, Carl Vondrick, Rose Yu, Laure Zanna, Tian Zheng, Ryan Abernathey, Fiaz Ahmed, David Bader, Pierre Baldi, Elizabeth Barnes, Christopher Bretherton, Peter Caldwell, Wayne Chuang, Yilun Han, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Karthik Kashinath, Marat Khairoutdinov, Thorsten Kurth, Nicholas Lutsko, Po-Lun Ma, Griffin Mooers, J. David Neelin, David Randall, Sara Shamekh, Mark Taylor, Nathan Urban, Janni Yuval, Guang Zhang, and Mike Pritchard. ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation. *Advances in Neural Information Processing Systems*, 36:22070–22084, December 2023.

Ming Yuan and Grace Wahba. Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & Probability Letters*, 69(1):11–20, 2004. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2004.03.009.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. doi: 10.1080/00018732.2016.1211393.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, February 2021. ISSN 0001-0782. doi: 10.1145/3446776.