5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–16

Online Laplace Model Selection Revisited

Jihao Andreas Lin[1,2] JAL232@CAM.AC.UK
Javier Antorán[1] JA666@CAM.AC.UK
José Miguel Hernández-Lobato[1] JMH233@CAM.AC.UK
[1] *University of Cambridge* [2] *Max Planck Institute for Intelligent Systems*

Abstract

The Laplace approximation provides a closed form model selection objective for neural networks (NN). *Online* variants, which optimise NN parameters jointly with hyperparameters, like weight decay strength, have seen renewed interest in the Bayesian deep learning community. However, these methods violate Laplace's method's critical assumption that the approximation is performed around a mode of the loss, calling into question their soundness. This work re-derives online Laplace methods, showing them to target a variational bound on a variant of the Laplace evidence which does not make stationarity assumptions, i.e, a mode-corrected method. Online Laplace and its mode-corrected counterpart share stationary points where 1. the NN parameters are a *maximum a posteriori*, satisfying Laplace's method's assumption, and 2. the hyperparameters maximise the Laplace evidence, motivating online methods. We demonstrate that these optima are roughly attained in practise by online algorithms using full-batch gradient descent on UCI regression datasets. Here, online model selection prevents overfitting and outperforms validation-based early stopping.

1. Introduction

Online model selection holds the promise of automatically tuning large numbers of neural network (NN) hyperparameters during a single training run, obsoleting cross validation (Clarke et al., 2022). Online Laplace methods (OL) interleave regular NN optimisation steps with steps of hyperparameter optimisation with respect to the Laplace-approximated model evidence (Foresee and Hagan, 1997; Friston et al., 2007; Immer et al., 2021a). The method has been used to learn layer-wise weight decay strengths, data noise scales, and even data augmentation hyperparameters (Immer et al., 2022).

Laplace's method constructs a second-order Taylor expansion around a mode of the posterior. Here, the first-order term vanishes, leaving a pure quadratic approximation to the log-density, that is, a Gaussian. The Laplace evidence is this distribution's normalising constant. The online setting implies non-convergence of the NN parameters, violating the stationarity assumption. Antorán et al. (2022b) show that applying Laplace's method to pre-trained NNs which do not attain zero loss gradient leads to severe deterioration in model selection performance. Immer et al. (2021a) experiment with inclusion of the first-order Taylor term in the OL procedure but find that it leads to training instability.

This work revisits online Laplace methods, taking steps towards reconciling their seeming unsoundness with their satisfactory empirical performance.

- We show that a variant of OL that includes the first-order Taylor term (and thus does not assume stationarity) corresponds to NN hyperparameter optimisation with the evidence of a tangent linear model. We re-derive the standard OL objective as a variational lower bound on the evidence of this tangent linear model, motivating its use.

© J.A. Lin, J. Antorán & J.M. Hernández-Lobato.

- We show OL shares fixed points, where the ELBO is tight, with the above first-order-corrected procedure. Here, the linear model's *maximum a posteriori* (MAP) parameters match the NN parameters, for which they are also a MAP, and the linear model's evidence matches the NN's Laplace evidence, further motivating its use.
- We show that these fixed points are roughly attained in practise by small NNs trained via full-batch gradient descent with OL hyperparameter tuning on UCI regression datasets, but they are not attained without online tuning. The hyperparameters found through online Laplace lead to NNs that outperform their non-online counterparts.

2. Preliminaries: Laplace approximation and online variants

We consider a supervised problem where n inputs, stacked as $X \in \mathbb{R}^{n \times d_x}$, are paired with scalar targets stacked as $y \in \mathbb{R}^n$. We introduce a NN regressor $f : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \to \mathbb{R}$ with $w \in \mathbb{R}^{d_w}$ as parameters. Omitting dependence on the inputs from our notation, as these are fixed, and working with arrays of stacked predictions $f(w) \in \mathbb{R}^n$, we model the targets as

$$y = f(w) + \epsilon \quad \text{with} \quad w \sim \mathcal{N}(0, \alpha^{-1} I_{d_w}) \quad \text{and} \quad \epsilon \sim \mathcal{N}(0, \beta^{-1} I_n),$$

where α is the scalar precision of our prior over NN parameters and β is the scalar homoscedastic noise precision. The set of both hyperparameters is $\mathcal{M} = \{\alpha, \beta\}$. Our results readily generalise to vector-valued hyperparameters and outputs (Antorán et al., 2023).

We train the NN to minimise the loss $\ell_f(w; \mathcal{M}) = \frac{\beta}{2}\|y - f(w)\|^2 + \frac{\alpha}{2}\|w\|^2$. This objective matches the unnormalised, negative log-joint density of our parameters and targets

$$e^{-\ell_f(w;\mathcal{M})} \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{d_w}{2}} = \mathcal{N}(y;\, f(w), \beta^{-1} I_n)\mathcal{N}(w;\, 0, \alpha^{-1} I_{d_w}) := p_f(y, w; \mathcal{M}).$$

Laplace's method (Mackay, 1992; Bishop, 2006) takes a second-order Taylor expansion of $\log p_f(y, w; \mathcal{M})$ around w_0, assumed to be a mode ($w_0 \in \arg\min_w \ell_f(w; \mathcal{M})$), and approximates the NN model evidence $p_f(y; \mathcal{M})$ by integrating the expanded expression as

$$\begin{aligned} &\int p_f(y, w; \mathcal{M})\, \mathrm{d}w \approx \int \exp\left(\log p_f(y, w_0; \mathcal{M}) - \frac{1}{2}(w - w_0)^\top H (w - w_0)\right) \mathrm{d}w \\ &= \exp\left(\frac{d_w}{2}\log\alpha + \frac{n}{2}\log\beta - \frac{\beta}{2}\|y - f(w_0)\|_2^2 - \frac{\alpha}{2}\|w_0\|_2^2 - \frac{1}{2}\log\det H - \frac{n}{2}\log 2\pi\right) \qquad (1) \\ &:= \exp \mathcal{L}_f(\mathcal{M}; w_0), \end{aligned}$$

where the first-order term vanishes since we assume $\partial_w \ell_f(w_0; \mathcal{M}) = 0$ and H is $\partial_w^2 \ell_f(w_0; \mathcal{M})$. In practice, the loss Hessian is replaced by the Generalised Gauss Newton matrix (GGN), that is, $H = \beta J^T J + \alpha I_{d_w} \in \mathbb{R}^{d_w \times d_w}$ for $J = \partial_w f(w_0) \in \mathbb{R}^{n \times d_w}$ the NN's Jacobian evaluated at the training inputs. This choice provides ease of computation and numerical stability (Martens, 2020; Immer et al., 2021b). We hereon assume H to be the GGN and suppress its dependence on the hyperparameters and the expansion point from our notation.

2.1. Online hyperparameter tuning

We may tune the hyperparameters $\mathcal{M}$ for a NN by choosing them to maximise $\mathcal{L}_f(\mathcal{M}; w^\star)$, where $w^\star \in \arg\min_w \ell_f(w; \mathcal{M})$ is a true optima of the loss. This will not change our NN's predictions however, as its parameters are held fixed at $w_\star$. Mackay (1992) proposes to re-train the NN from scratch using the updated hyperparameters. This procedure terminates when a joint stationary point of the parameters and hyperparameters is found: $w_\star \in \arg\min_w \ell_f(w; \mathcal{M}^\star)$ and $\mathcal{M}^\star \in \arg\min_{\mathcal{M}} \mathcal{L}_f(\mathcal{M}; w^\star)$.

The size of NN parameter spaces and datasets has grown dramatically since 1992 and re-training the NN is now impractical. This motivates online Laplace (OL) approaches which, at timestep t with parameters w_t, perform a step of NN parameter optimisation to minimise $\ell_f(w_t; \mathcal{M})$, followed by a hyperparameter update to maximise $\mathcal{L}_f(\mathcal{M}; w_t)$ (Foresee and Hagan, 1997; Friston et al., 2007; Immer et al., 2021a)[1]. Importantly, the Taylor expansion point is chosen to match the current NN parameter setting w_t. Since optimisation has not converged, $w_t \notin \arg\min_w \ell_f(w; \mathcal{M})$, then $\mathcal{L}_f(\mathcal{M}; w_t)$, which discards the first-order expansion term, *does not provide a local approximation to the evidence.* Figure 1 illustrates that discarding this first-order term can result in a vastly different approximation.

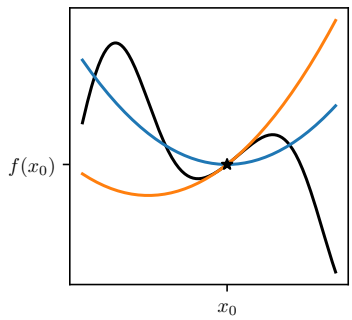

Figure 1: Second-order Taylor approximations of a function f (black) around x_0 (star), with (orange) and without (blue) the first-order term.

3. Understanding online Laplace through the tangent model

This section re-derives online Laplace methods for NNs as empirical Bayesian inference in a tangent linear model. This analysis provides a non-heuristic motivation for these methods.

Online hyperparameter optimisation including the first-order term To start, we integrate the Taylor expansion of $\log p_f(y, w; \mathcal{M})$ at w_t without discarding the first-order term

$$\int \exp\left(\log p_f(y, w_t; \mathcal{M}) - \partial_w \ell_f(w_t; \mathcal{M})(w - w_t) - \frac{1}{2}(w - w_t)^\top H(w - w_t)\right) \mathrm{d}w \tag{2}$$
$$= \exp\left(\frac{d_w}{2}\log\alpha + \frac{n}{2}\log\beta - \frac{\beta}{2}\|y - (f(w_t) + J(v^\star - w_t))\|_2^2 - \frac{\alpha}{2}\|v^\star\|_2^2 - \frac{1}{2}\log\det H - \frac{n}{2}\log 2\pi\right)$$
$$:= \exp \mathcal{L}_h(\mathcal{M}; w_t),$$

where the optimum of the quadratic expansion $v^\star$ is obtained by performing a Gauss-Newton update on the NN loss at w_t, that is, $v^\star = w_t - \beta H^{-1}\left(J^T(f(w_t) - y) + \alpha w_t\right)$. Gauss-Newton optimisation is common in the NN second order optimisation-literature; it makes more progress per step than gradient descent (Botev et al., 2017; Martens, 2020; Barbano et al., 2023). Since $\mathcal{L}_h(\mathcal{M}; w_t)$ represents a local evidence approximation, mode-corrected by a Gauss-Newton step, it may more closely match the Laplace evidence computed at an optimum $\mathcal{L}_f(\mathcal{M}; w^\star)$ than the Laplace evidence computed at the online expansion point $\mathcal{L}_f(\mathcal{M}; w_t)$. We

1. Friston et al. (2007) refers to the described online Laplace procedure as *Variational Laplace*.

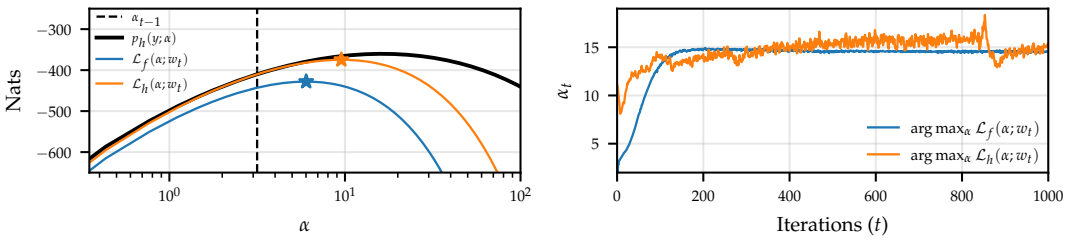

Figure 2: Illustration of exact linear model evidence, and LM and OL hyperparameter objectives at a single train step t (left). The latter two represent lower bounds as per (3). $\mathcal{L}_h$ is tighter which leads to larger updates, making hyperparameter trajectories unstable (right).

may use $\mathcal{L}_h(\mathcal{M}; w_t)$ to construct an algorithm where each step of NN parameter optimisation minimising $\ell_f(w_t; \mathcal{M})$ is interleaved with a hyperparameter update maximising $\mathcal{L}_h(\mathcal{M}; w_t)$.

Online hyperparameter optimisation with the tangent linear model The mode-corrected evidence approximation $\mathcal{L}_h(\mathcal{M}; w_t)$ matches the exact evidence of a *tangent model* with regressor $h : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \to \mathbb{R}$ given by a first-order Taylor approximation to f at w_t

$$y = h(v) + \epsilon \;\text{ with }\; h(v) \coloneqq f(w_t) + J(v - w_t) \;\text{ and }\; v \sim \mathcal{N}(0, \alpha^{-1} I_{d_w}), \;\; \epsilon \sim \mathcal{N}(0, \beta^{-1} I_n).$$

The tangent model's loss is $\ell_h(v; \mathcal{M}) = \frac{\beta}{2}\|y - h(v)\|^2 + \frac{\alpha}{2}\|v\|^2$. It is minimised by the MAP $v^\star \coloneqq \arg\min_v \ell_h(v; \mathcal{M})$, which matches the optimum of the quadratic expansion in (2), and its constant curvature matches the GGN $\partial_v^2 \ell_h(v^\star; \mathcal{M}) = H$. Thus, the above described mode-corrected algorithm selects hyperparameters that improve the tangent model's evidence $\mathcal{L}_h(\mathcal{M}; w_t)$ at each step. We hereon refer to this procedure as online Linear Model (LM).

Online Laplace as a variational bound on the tangent model's evidence Finally, we connect the LM procedure to the OL procedure described in Section 2.1. For this, we apply Antorán et al. (2023)'s lower bound on the linear model's evidence, i.e., the ELBO,

$$\begin{aligned}\mathcal{L}_h(\mathcal{M}; w_t) &\geq \mathbb{E}_{\mathcal{N}(v;\,\mu, H^{-1})}[\log p_h(y, v; \mathcal{M})] + \mathcal{H}(\mathcal{N}(\mu, H^{-1})) \coloneqq \mathcal{L}(\mathcal{M}; \mu, w_t) \qquad (3)\\ &= \exp\left(\frac{d_w}{2}\log\alpha + \frac{n}{2}\log\beta - \frac{\beta}{2}\|y - (f(w_t) + J(\mu - w_t))\|_2^2 - \frac{\alpha}{2}\|\mu\|_2^2 - \frac{1}{2}\log\det H - \frac{n}{2}\log 2\pi\right)\end{aligned}$$

where $\mathcal{H}$ refers to the differential entropy and $p_h(y, v; \mathcal{M}) \coloneqq e^{-\ell_h(v;\mathcal{M})}\left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{d_w}{2}}$.

The bound is tight when the variational posterior mean μ matches the linear model's MAP $v^\star$, that is, $\mathcal{L}(\mathcal{M}; v^\star, w_t) = \mathcal{L}_h(\mathcal{M}; w_t)$. This recovers the LM objective. On the other hand, by choosing the variational mean to match the linearisation point, $\mu = w_t$, we recover the OL evidence $\mathcal{L}(\mathcal{M}; w_t, w_t) = \mathcal{L}_f(\mathcal{M}; w_t)$. Thus, discarding the first-order term in the Taylor expansion of $\log p_f(y, w; \mathcal{M})$, which seemed unjustified, actually results in a lower bound on a variant of the Laplace evidence which corrects for the expansion point not matching the mode, that is, $\mathcal{L}_f(\mathcal{M}; w_t) \leq \mathcal{L}_h(\mathcal{M}; w_t)$. This is illustrated in Figure 2 (left).

Convergence behaviour Both the OL and LM procedures reach a shared fixed point when a simultaneous optimum of the NN weights and linear model hyperparameters is found, that is, $w^\star = \arg\min_w \ell_f(w; \mathcal{M}^\star)$ and $\mathcal{M}^\star = \arg\max_{\mathcal{M}} \mathcal{L}_h(\mathcal{M}; w^\star) = \arg\max_{\mathcal{M}} \mathcal{L}(\mathcal{M}; w^\star, w^\star)$.

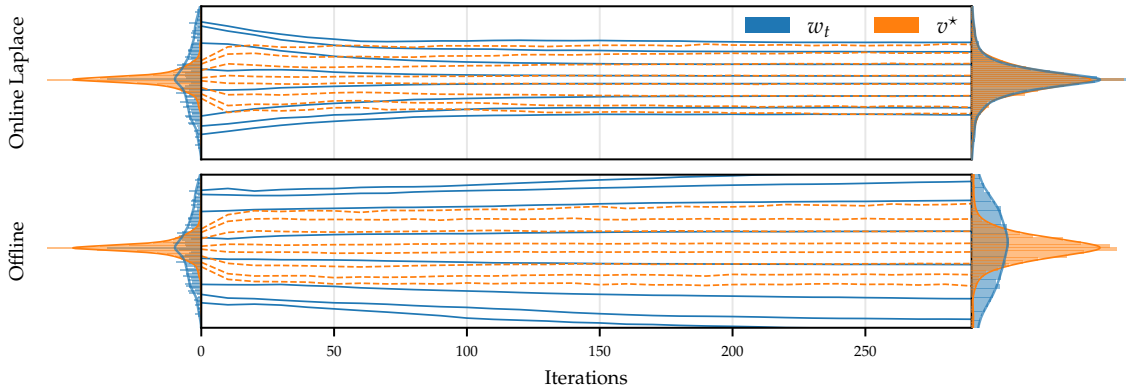

Figure 3: Evolution of neural network weights w_t and tangent linear model posterior mean $v^\star$ during training. With online Laplace (OL procedure), w_t and $v^\star$ converge to the same distribution. Offline training does not exhibit this behaviour.

Since the first-order term vanishes here, $\partial_w \ell_f(w^\star; \mathcal{M}) = 0$, the ELBO is tight and the linear model evidence $\mathcal{L}_h$ matches the regular Laplace evidence $\mathcal{L}_f$, satisfying the termination condition of Mackay (1992)'s NN re-training algorithm.

At convergence, we have $\partial_w \ell_f(w^\star; \mathcal{M}) = \partial_v \ell_h(w^\star; \mathcal{M}) = 0$. We may thus use similarity between $w^\star$ and $v^\star$ to check for convergence of the OL and LM procedures. Figure 3 shows the LM procedure results in the NN and tangent model converging to a shared MAP. This does not occur when performing standard NN training with fixed hyperparameters.

Since the NN and linear model share hyperparameters $\mathcal{M}$, their loss gradients match at the linearisation point $\partial_w \ell_f(w_t; \mathcal{M}) = \partial_v \ell_h(w_t; \mathcal{M})$. Thus, during training, each NN optimisation step brings its parameters closer to the MAP of the current tangent model while also changing the linearisation point, which in turn changes the tangent model. This is not guaranteed to result in a tangent linear model with increased evidence. Indeed, Figure 4 (right) shows both the OL and LM evidences exhibit non-monotonic training curves. Interestingly, the slack in the OL bound results in smaller hyperparameter steps, as shown in Figure 2. These may counteract training instability coming from a rapidly changing tangent model evidence. We find OL updates often lead to more stable convergence to a shared MAP for the NN and tangent model than LM updates. This can be seen in Figure 4 (left).

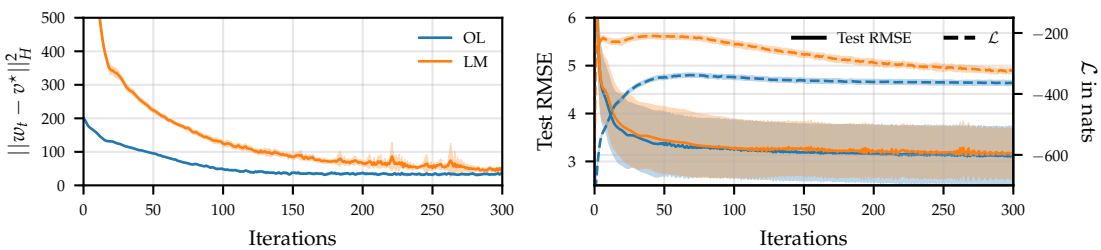

Figure 4: Difference between neural network weights w_t and linear model posterior mean $v^\star$ (left) and test RMSE versus ELBO $\mathcal{L}$ (right) throughout online training for both OL and LM procedures. A maximised ELBO does not imply optimal test RMSE.

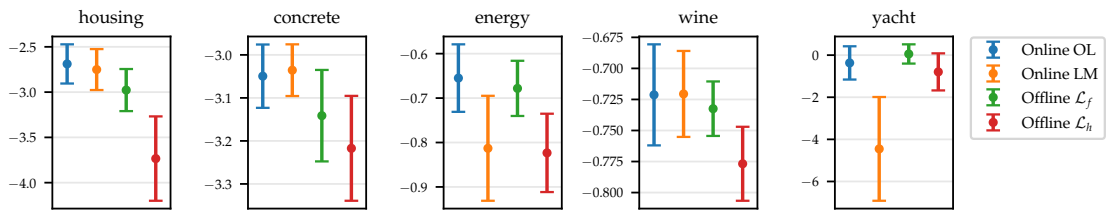

Figure 5: Test log-likelihood on UCI regression (mean ± standard error over 10 splits).

4. Experiments

This section demonstrates the OL and LM online hyperparameter tuning algorithms discussed in the previous section and contrasts them with standard NN training, that is, with fixed hyperparameters (labelled offline). To this end, we train single hidden layer, 50 hidden unit MLPs on 5 UCI datasets (Dua and Graff, 2017): housing, concrete, energy, wine, and yacht. We train our NNs with full-batch Adam and an exponentially decaying learning rate. For online methods, we train until convergence, updating the hyperparameters after every NN parameter step. The hyperparameter update step is detailed in Appendix D. For offline methods, we select the stopping point that minimises RMSE on a validation set consisting of 10% of training data points. We do not use normalisation layers, as these interfere with the Laplace approximation (Antorán et al., 2022a).

Convergence of NN and linear model posteriors Figure 4 (left) shows the Euclidean distance between the NN parameters and tangent model MAP decreasing throughout training for both the LM and OL procedures. Although the distance does not exactly reach 0, the NN and linear model parameter vectors present nearly identical distributions, as shown in Figure 3. This does not occur in the offline setting. The LM procedure takes larger hyperparameter steps than OL resulting in some training instability and overall slower convergence, as shown in Figure 2. Figure 4 (right) shows that both the LM and OL evidence estimates attain their maxima very early in training, before the lowest test RMSE is reached. As discussed in Section 3, these metrics target the evidence of a tangent model and may only transfer to the NN once online training has fully converged.

Predictive performance We obtain predictive distributions using the linearised Laplace method (Immer et al., 2021b). For NNs trained offline, linearised Laplace hyperparameters are chosen post-hoc my maximising either the $\mathcal{L}_f$ or $\mathcal{L}_h$ objectives. Figure 5 reports log-likelihood mean and standard errors obtained across 10 train-test splits where the test set size is 10% of the whole dataset size.

We find the OL procedure to consistently provide the best performance or be a close second best. The LM procedure performs very similarly to OL on all datasets except energy and yacht, where it performs the worse out of all methods. We attribute this to training instability. Offline methods only perform competitively on yacht, despite use of validation-based stopping. The online methods' improvement over their offline counterparts can mostly be explained by improved RMSE (see Figure 14), suggesting online hyperparameter optimisation helps find NN parameters that effectively navigate the bias-variance trade off.

The full set of experimental results, including hyperparameter trajectories and evidence estimates, is provided in Appendix A.

Acknowledgements

JAL was supported by the University of Cambridge Harding Distinguished Postgraduate Scholars Programme. JA acknowledges support from Microsoft Research, through its PhD Scholarship Programme, and from the EPSRC. JMHL acknowledges support from a Turing AI Fellowship under grant EP/V023756/1.

References

Javier Antorán, James Urquhart Allingham, David Janz, Erik Daxberger, Eric Nalisnick, and José Miguel Hernández-Lobato. Linearised Laplace Inference in Networks with Normalisation Layers and the Neural g-Prior. In *Advances in Approximate Bayesian Inference*, 2022a.

Javier Antorán, David Janz, James Urquhart Allingham, Erik A. Daxberger, Riccardo Barbano, Eric T. Nalisnick, and José Miguel Hernández-Lobato. Adapting the Linearised Laplace Model Evidence for Modern Deep Learning. In *International Conference on Machine Learning*, volume 162, pages 796–821, 2022b.

Javier Antorán, Shreyas Padhy, Riccardo Barbano, Eric Nalisnick, David Janz, and José Miguel Hernández-Lobato. Sampling-based inference for large linear models, with application to linearised Laplace. In *International Conference on Learning Representations*, 2023.

Riccardo Barbano, Javier Antorán, Johannes Leuschner, José Miguel Hernández-Lobato, Zeljko Kereta, and Bangti Jin. Fast and Painless Image Reconstruction in Deep Image Prior Subspaces. In *arXiv abs/2302.10279*, 2023.

Christopher M Bishop. *Pattern Recognition and Machine Learning.* springer, 2006.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical Gauss-Newton Optimisation for Deep Learning. In *International Conference on Machine Learning*, 2017.

Ross M Clarke, Elre Talea Oldewage, and José Miguel Hernández-Lobato. Scalable One-Pass Optimisation of High-Dimensional Weight-Update Hyperparameters by Implicit Differentiation. In *International Conference on Learning Representations*, 2022.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017.

F. Dan Foresee and Martin T. Hagan. Gauss-Newton approximation to Bayesian learning. In *International Conference on Neural Networks*, 1997.

Karl J. Friston, Jérémie Mattout, Nelson J. Trujillo-Barreto, John Ashburner, and William D. Penny. Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234, 2007.

Stephen F. Gull. *Bayesian Data Analysis: Straight-line fitting*, pages 511–518. Springer Netherlands, 1989.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Mohammad Emtiyaz Khan. Scalable Marginal Likelihood Estimation for Model Selection in Deep Learning. In *International Conference on Machine Learning*, 2021a.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of Bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics*, 2021b.

Alexander Immer, Tycho F.A. van der Ouderaa, Gunnar Ratsch, Vincent Fortuin, and Mark van der Wilk. Invariance Learning in Deep Neural Networks with Differentiable Laplace Approximations. In *Advances in Neural Information Processing Systems*, 2022.

Alexander Immer, Tycho F. A. van der Ouderaa, Mark van der Wilk, Gunnar Rätsch, and Bernhard Schölkopf. Stochastic Marginal Likelihood Gradients using Neural Tangent Kernels. In *arXiv abs/2306.03968*, 2023.

Mohammad Emtiyaz Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate Inference Turns Deep Networks into Gaussian Processes. In *Advances in Neural Information Processing Systems*, 2019.

Jihao Andreas Lin, Javier Antorán, Shreyas Padhy, David Janz, José Miguel Hernández-Lobato, and Alexander Terenin. Sampling from Gaussian Process Posteriors using Stochastic Gradient Descent, 2023.

David J. C. Mackay. *Bayesian Methods for Adaptive Models.* PhD thesis, 1992.

James Martens. New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21:146:1–146:76, 2020.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A Scalable Laplace Approximation for Neural Networks. In *International Conference on Learning Representations*, 2018.

Michael E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Michael E. Tipping and Anita C. Faul. Fast Marginal Likelihood Maximisation for Sparse Bayesian Models. In *International Conference on Artificial Intelligence and Statistics*, 2003.

Joe Watson, Jihao Andreas Lin, Pascal Klink, and Jan Peters. Neural Linear Models with Functional Gaussian Process Priors. In *Advances in Approximate Bayesian Inference*, 2020.

Joe Watson, Jihao Andreas Lin, Pascal Klink, Joni Pajarinen, and Jan Peters. Latent Derivative Bayesian Last Layer Networks. In *International Conference on Artificial Intelligence and Statistics*, 2021.

David P. Wipf and Srikantan S. Nagarajan. A New View of Automatic Relevance Determination. In *Advances in Neural Information Processing Systems*, 2007.

Appendix A. Evolution of neural network training

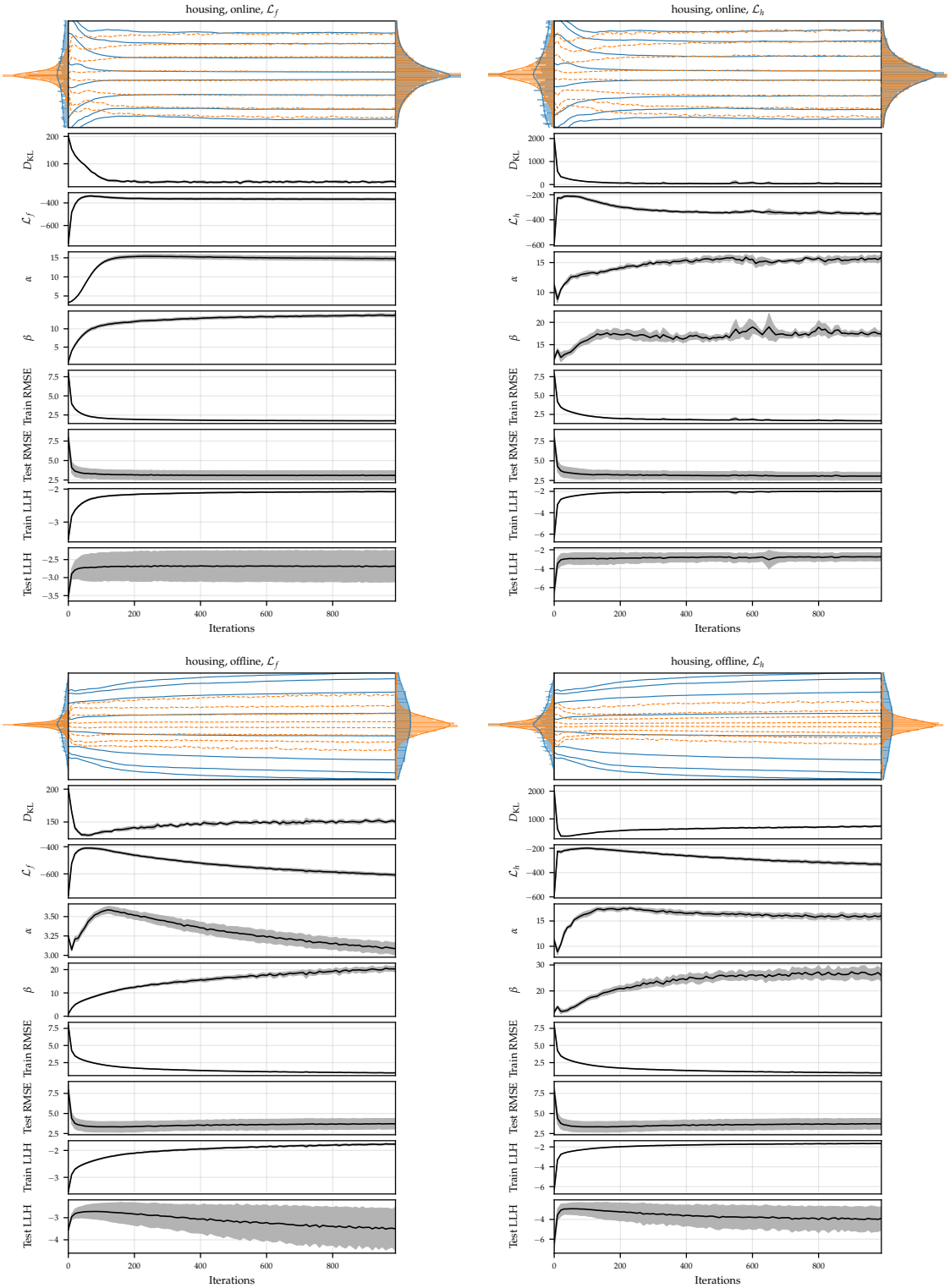

Figure 6: Evolution of neural network training on the housing dataset.

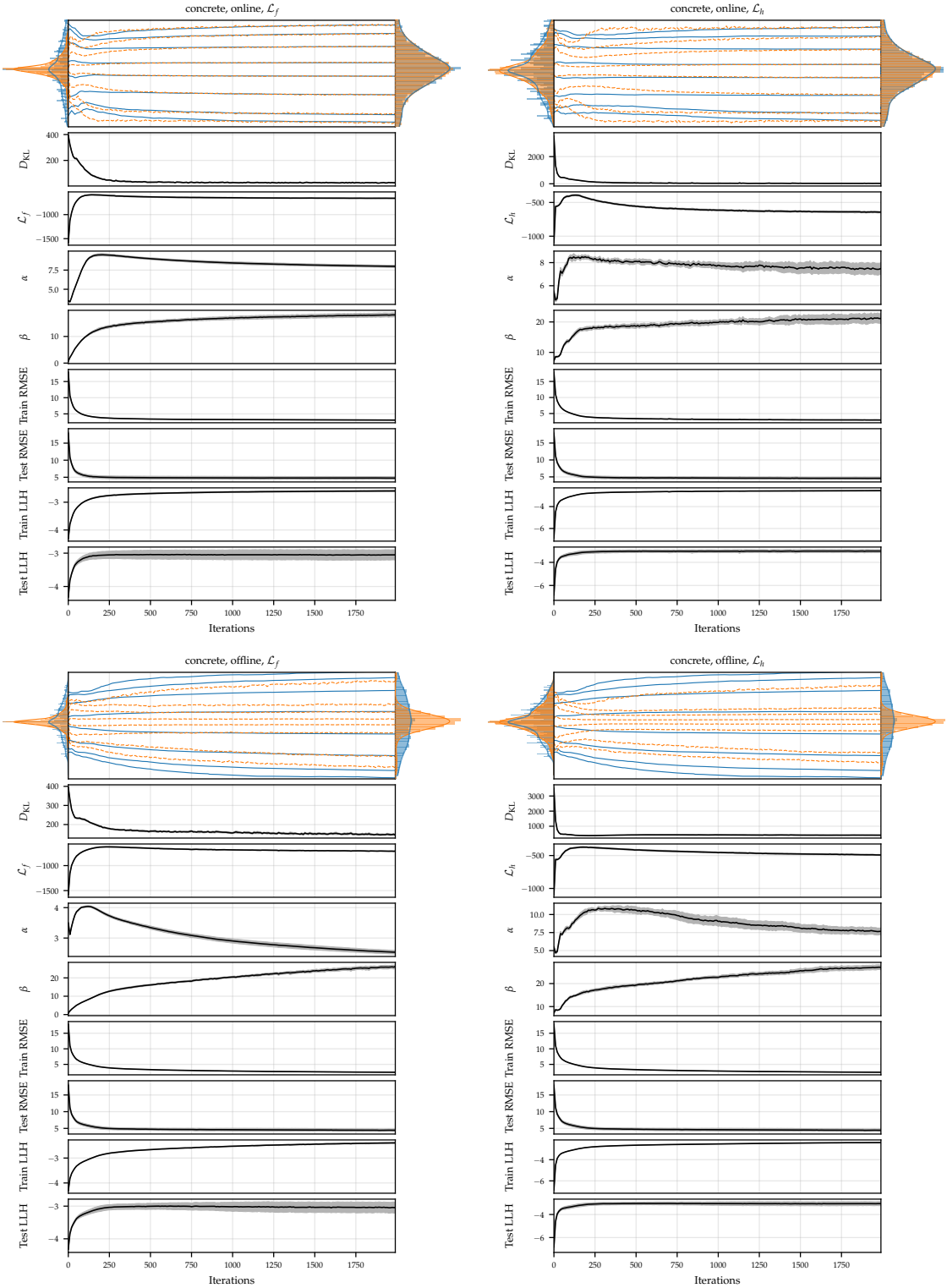

Figure 7: Evolution of neural network training on the concrete dataset.

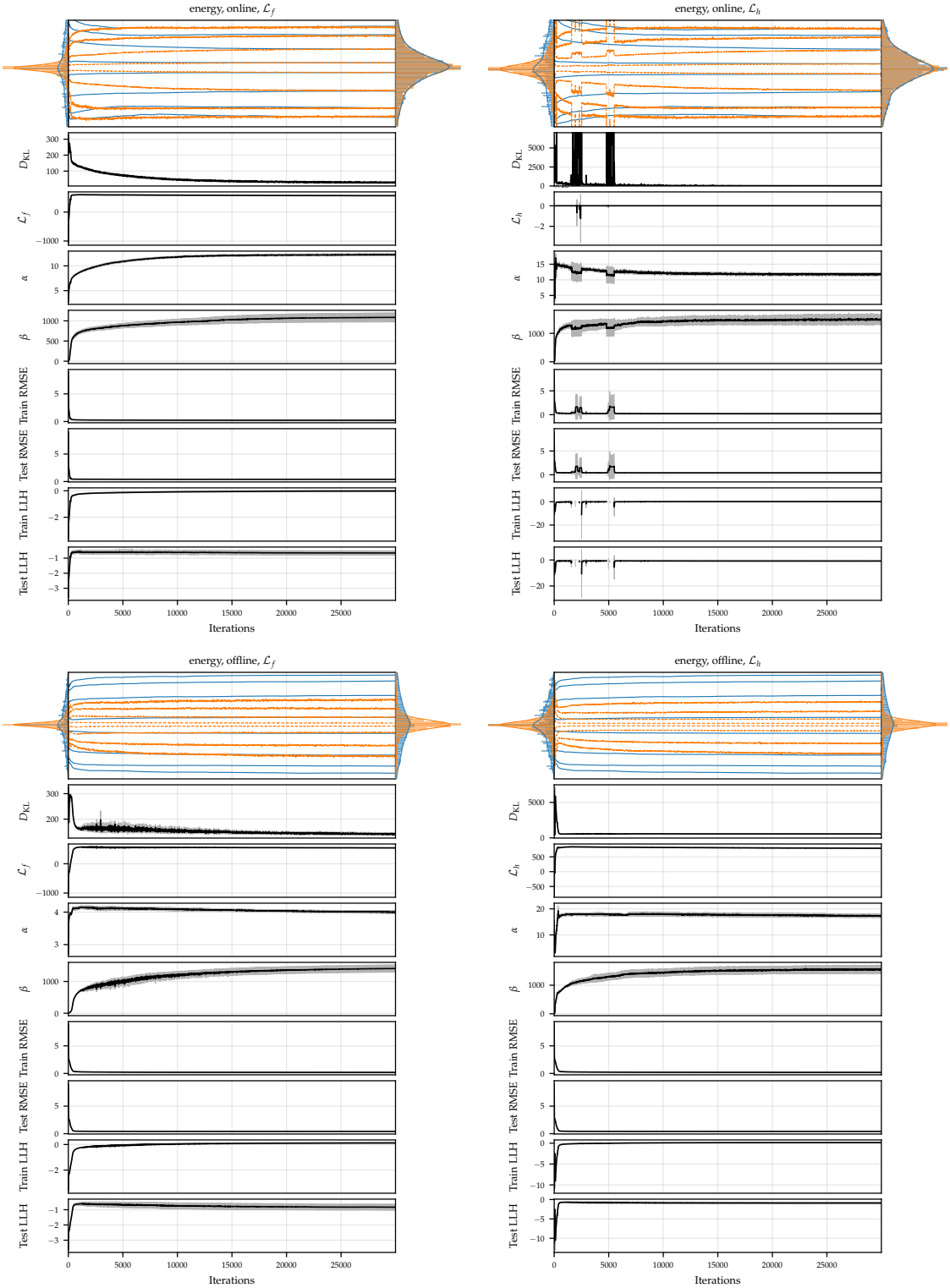

Figure 8: Evolution of neural network training on the energy dataset.

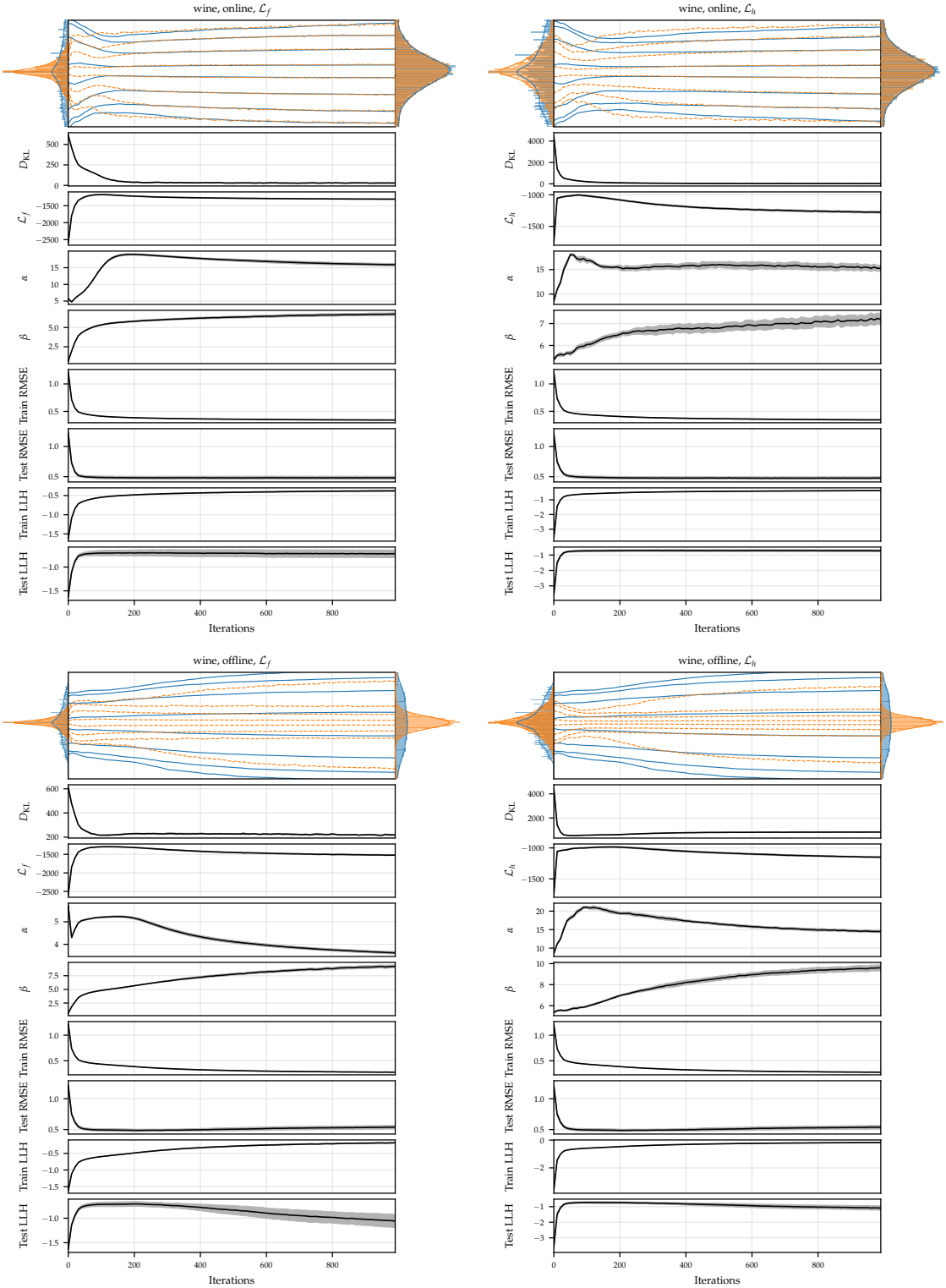

Figure 9: Evolution of neural network training on the wine dataset.

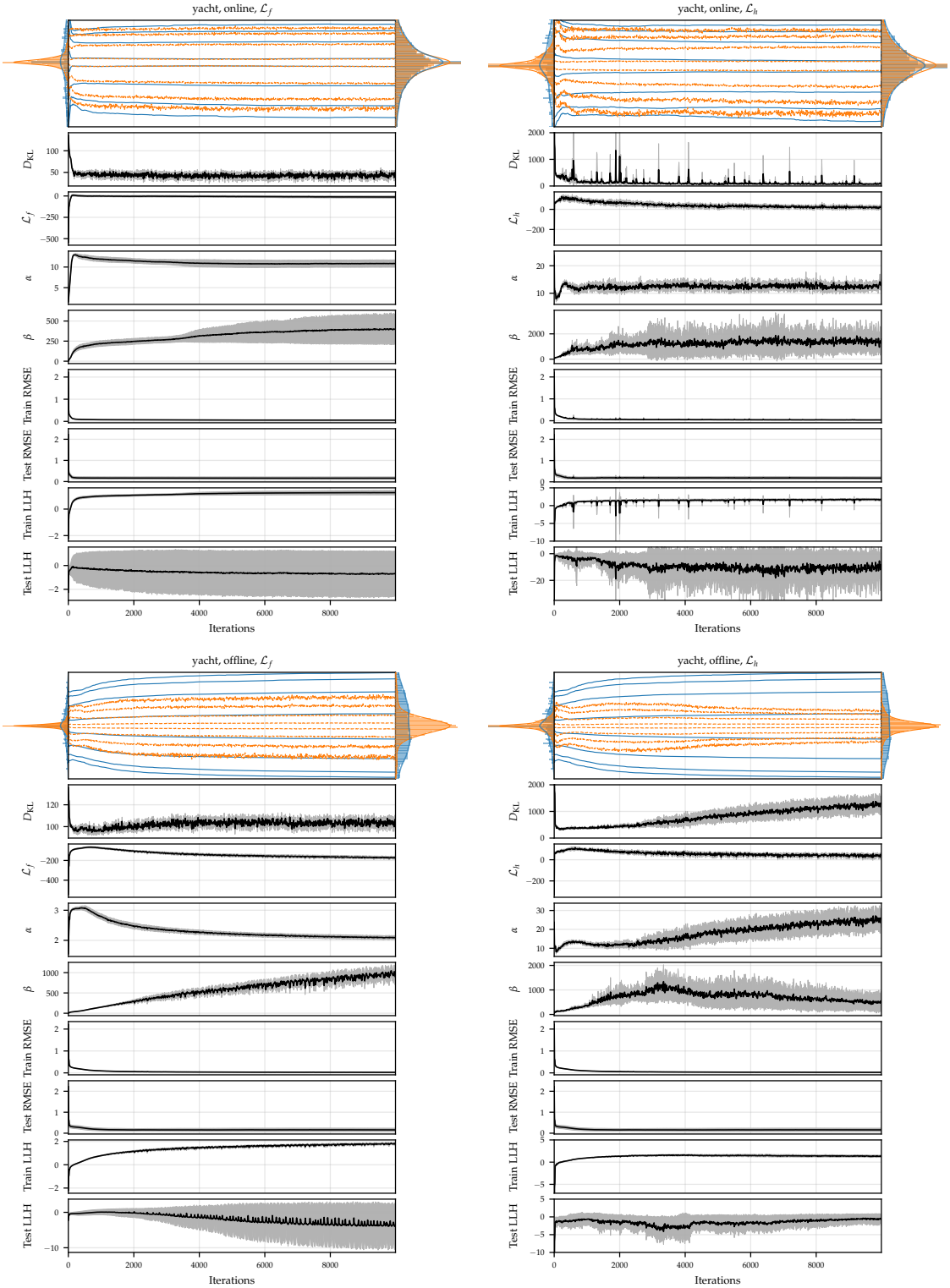

Figure 10: Evolution of neural network training on the yacht dataset.

Appendix B. Additional predictive metrics

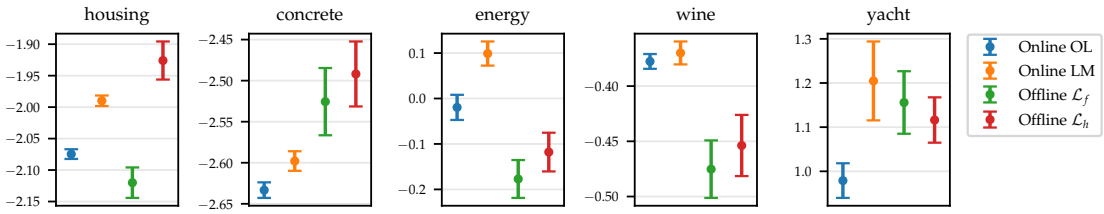

Figure 11: Train log-likelihood on UCI regression (mean ± standard error over 10 splits).

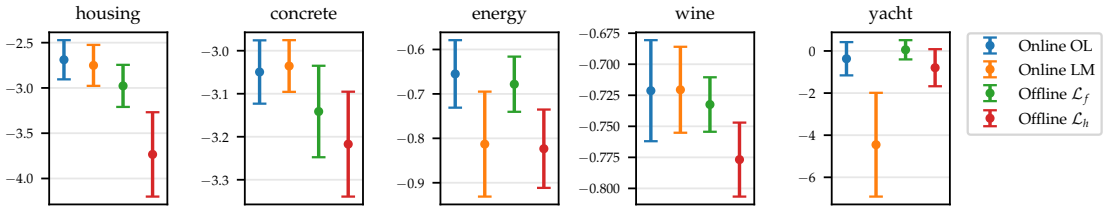

Figure 12: Test log-likelihood on UCI regression (mean ± standard error over 10 splits).

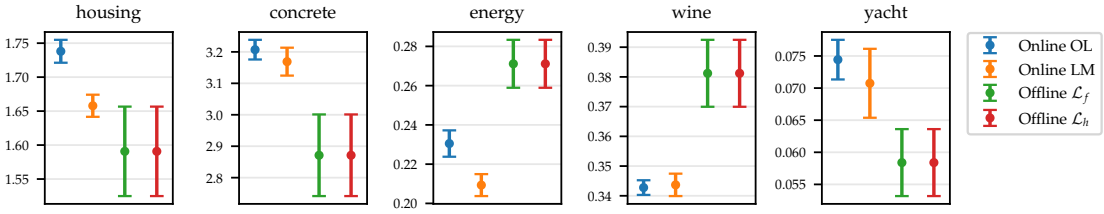

Figure 13: Train RMSE on UCI regression (mean ± standard error over 10 splits).

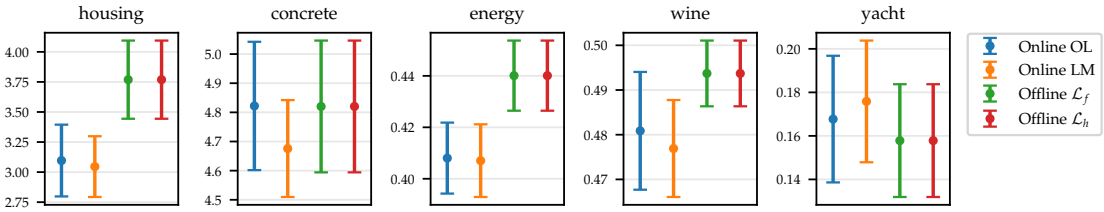

Figure 14: Test RMSE on UCI regression (mean ± standard error over 10 splits).

Appendix C. Related work

This work builds on the rich literature of Bayesian linear regression, which was first introduced by Gull (1989). We recommend chapter 3 of Bishop (2006) and chapter 2 of Rasmussen and Williams (2006) for an introduction. The use of both the linear model evidence and Laplace model evidence for hyperparameter selection are introduced in Mackay (1992). These techniques are analysed and extended by Tipping (2001); Tipping and Faul (2003); Wipf and Nagarajan (2007) for linear model and by Immer et al. (2021a); Antorán et al. (2022b) for NNs.

Despite their analytic tractability, linear models are held back by a cost of inference cubic in the number of parameters when expressed in primal form, or cubic in the number of observations for the dual (i.e. kernelised or Gaussian Process) form. Although we do not deal with this computationally intractability in this work, there exist a number of approximations which aim to make inference in linearised NNs scalable (Ritter et al., 2018; Khan et al., 2019; Watson et al., 2020, 2021; Antorán et al., 2023; Immer et al., 2023; Lin et al., 2023).

Appendix D. MacKay's hyperparameter update

Given mean μ and predictions $\hat{y}$, the hyperparameters α and β can be updated by

$$\alpha = \frac{\gamma}{\|\mu\|_2^2}, \qquad \beta = \frac{n-\gamma}{\|y-\hat{y}\|_2^2}, \qquad \gamma = d_w - \alpha \mathrm{Tr}(H^{-1}), \tag{4}$$

to maximise the linear model's evidence (Mackay, 1992), where γ is the effective number of dimensions. For OL, $\mu = w$ and $\hat{y} = f(w)$, and in the case of LM, $\mu = v^\star$ and $\hat{y} = h(v^\star)$.

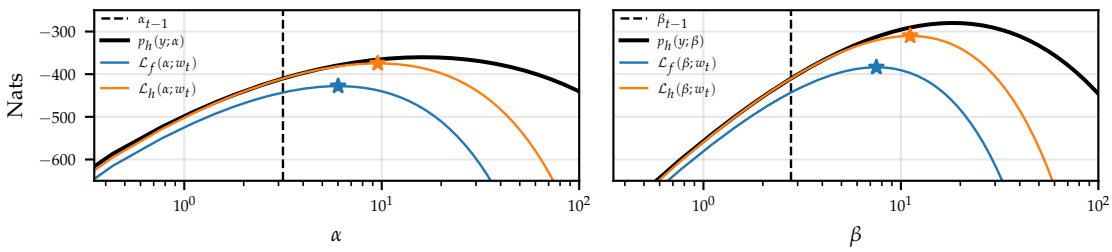

Figure 15: Illustration of a hyperparameter update at a single timestep t. Starting at the current value (dashed black), the updated value (star) is obtained by maximising the ELBO.

Appendix E. Implementation details

The initial learning rate used for Adam was set to 0.01 and the exponential learning rate decay factor was set to 0.9999. In the online setting, housing and wine were trained for 1000, concrete for 2000, energy for 30000 and yacht for 10000 steps until convergence. In the offline setting, the validation RMSE was tracked for a maximum of 30000 epochs.