Learning Hyperparameters via a Data-Emphasized Variational Objective


Ethan Harvey


Mikhail Petrov


Michael C. Hughes

Spotlight presentation

Tuning Hyperparameters is Expensive

Grid search or smarter search methods rely on

- Expensive separate runs
- Carving out a validation set reduces available training data

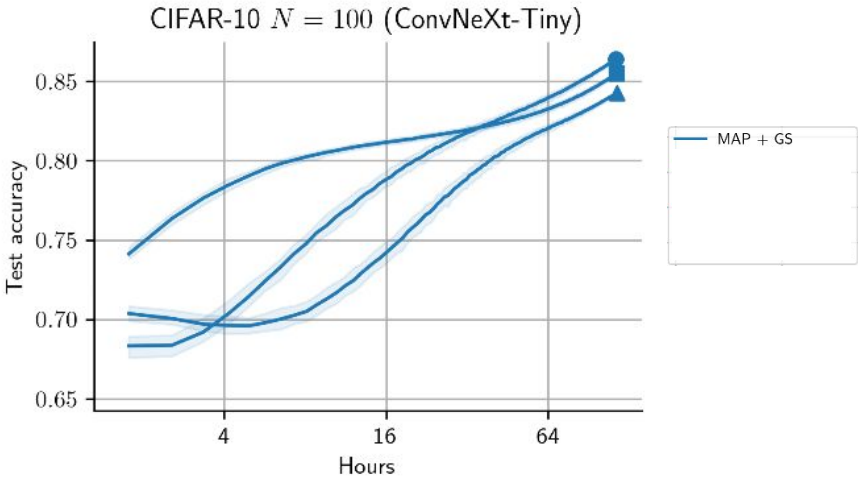


Figure: Test accuracy over time for ConvNeXt-Tiny (D > 1e7) on CIFAR-10 (N = 100).

Tuning Hyperparameters is Expensive

Grid search or smarter search methods rely on

- Expensive separate runs
- Carving out a validation set reduces available training data

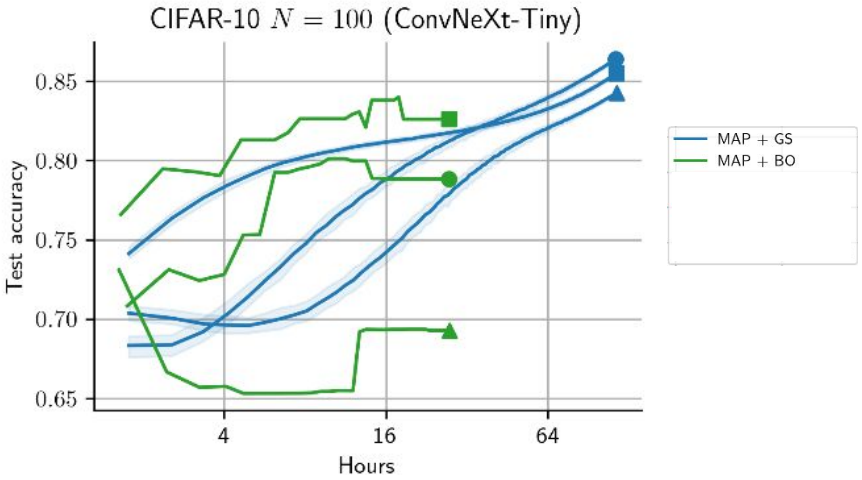


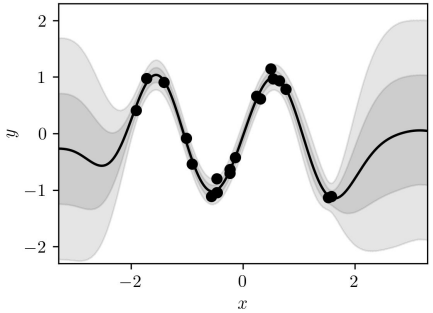
Figure: Test accuracy over time for ConvNeXt-Tiny (D > 1e7) on CIFAR-10 (N = 100).

Evidence or Marginal Likelihood

$$\log p_\eta(y) = \log \int p_\eta(y|\theta) p_\eta(\theta) d\theta$$

- Used in probabilistic models to learn hyperparameters
- Prefers the simplest model that fits the data best (MacKay, 1992)
- For fixed model and data, fits data well even when $D \gg N$

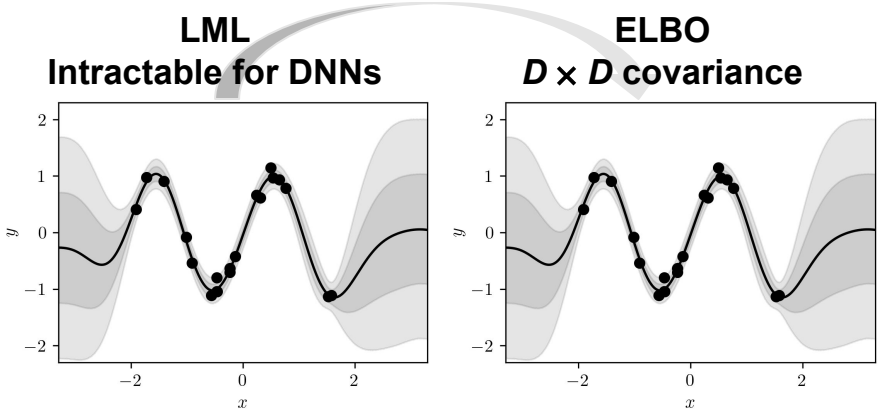
LML
Intractable for DNNs



Evidence Lower Bound (ELBO)

$$\log p_\eta(y) \geq \mathbb{E}_{q_\psi(\theta)}\left[\log p_\eta(y|\theta)\right] - D_{\mathrm{KL}}(q_\psi(\theta)\|p_\eta(\theta))$$

In variational inference, we assume an approximate posterior distribution $q_\psi(\theta)$ and maximize the ELBO

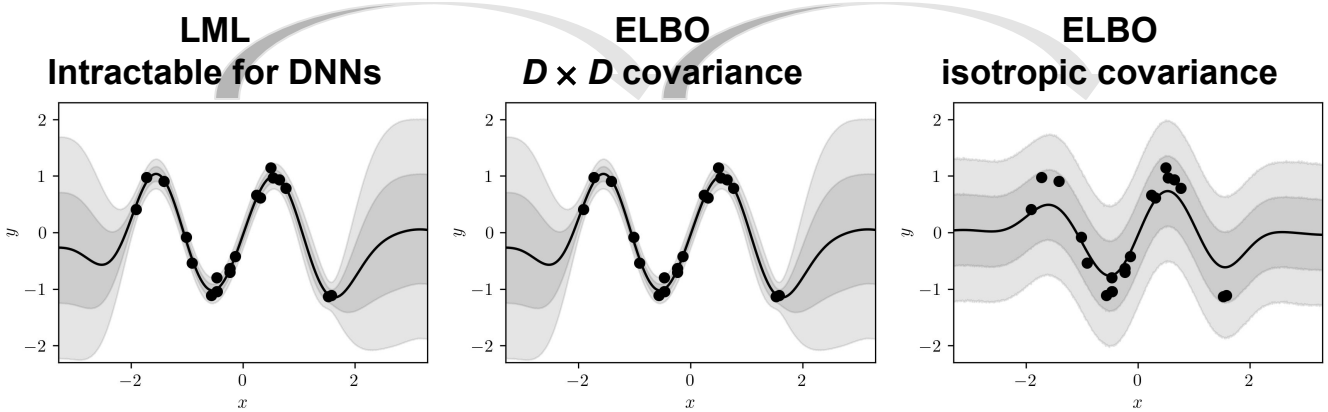


$D \gg N$ + Isotropic Covariance = ELBO Underfits

$$q_\psi(\theta) = \mathcal{N}(\theta|\bar{\theta}, \bar{\sigma}_q^2 I_D)$$

CONTRIBUTIONS

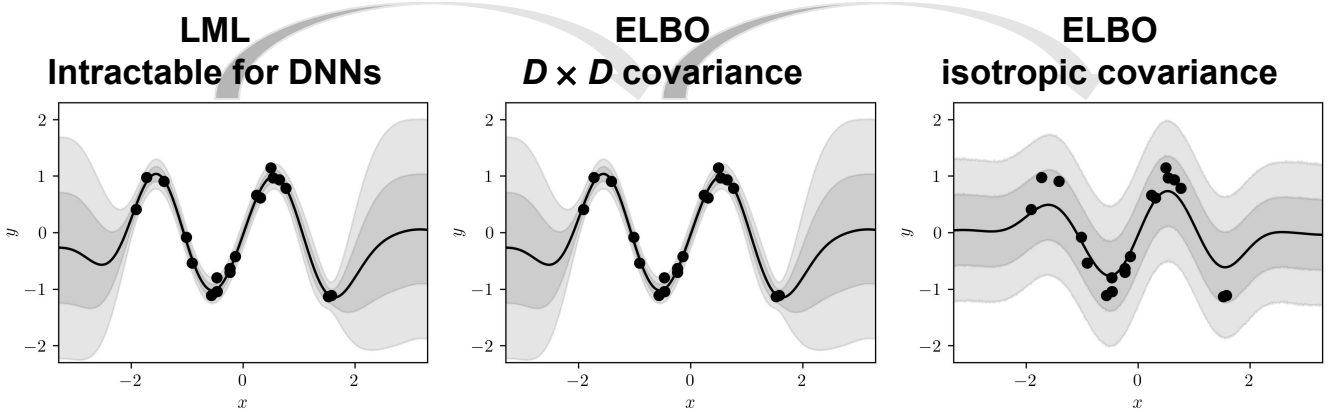
1. We show analytically that under these assumptions the ELBO underfits



Tempered Variational Inference

$$\kappa \cdot \mathbb{E}_{q_\psi(\theta)}\left[\log p_\eta(y|\theta)\right] - D_{\mathrm{KL}}(q_\psi(\theta)\|p_\eta(\theta))$$

Tempered variational inference (Mandt et al., 2016; Zhang et al., 2017; Aitchison, 2021), power posteriors (McLatchie et al., 2025), and β-VAE (Higgins et al., 2017)

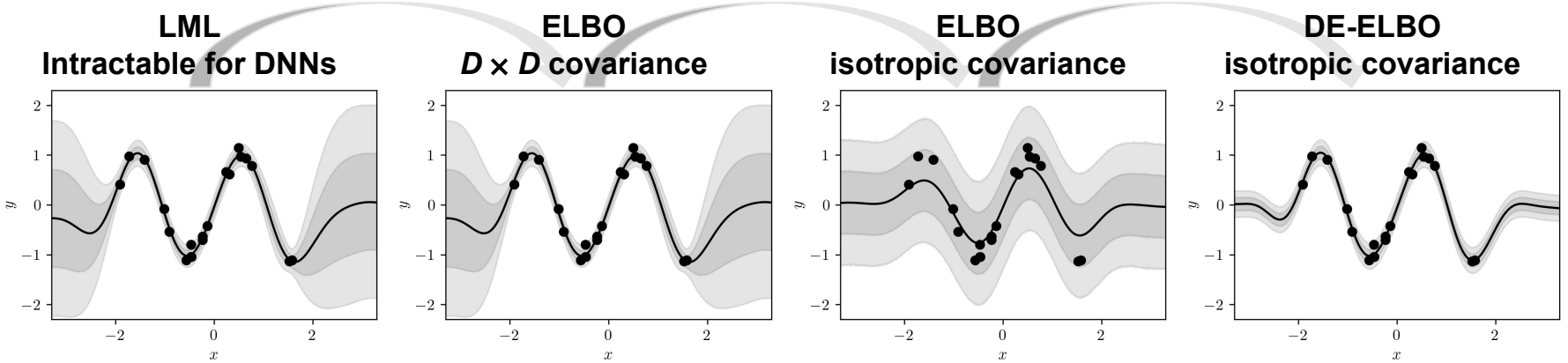


Data-Emphasized ELBO

$$\kappa \cdot \mathbb{E}_{q_\psi(\theta)}\left[\log p_\eta(y|\theta)\right] - D_{\mathrm{KL}}(q_\psi(\theta)\|p_\eta(\theta))$$

CONTRIBUTIONS

2. We propose the DE-ELBO (κ = *D*/*N*) for learning hyperparameters



Learning Hyperparameters is Affordable

- Gradient-based learning of hyperparameters via the DE-ELBO avoids the extreme time costs of grid search
- The accuracy of DE-ELBO is competitive with or better than recent Bayesian methods (Immer et al., 2021; Lotfi et al., 2022)

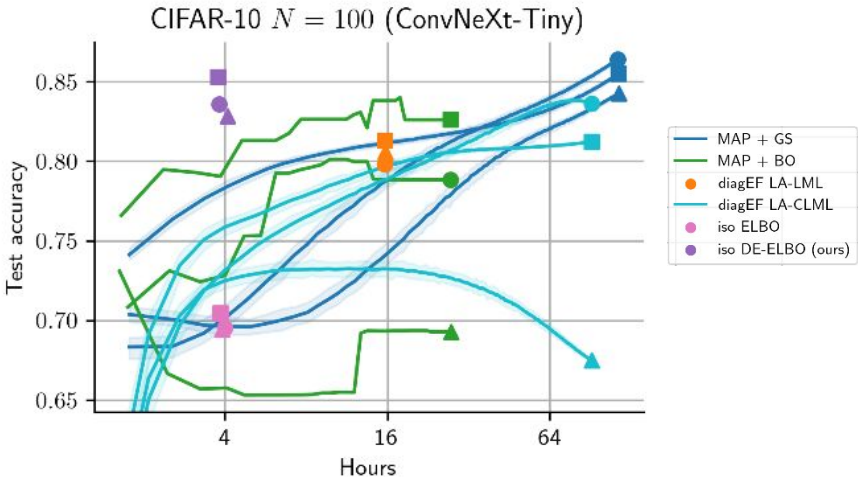
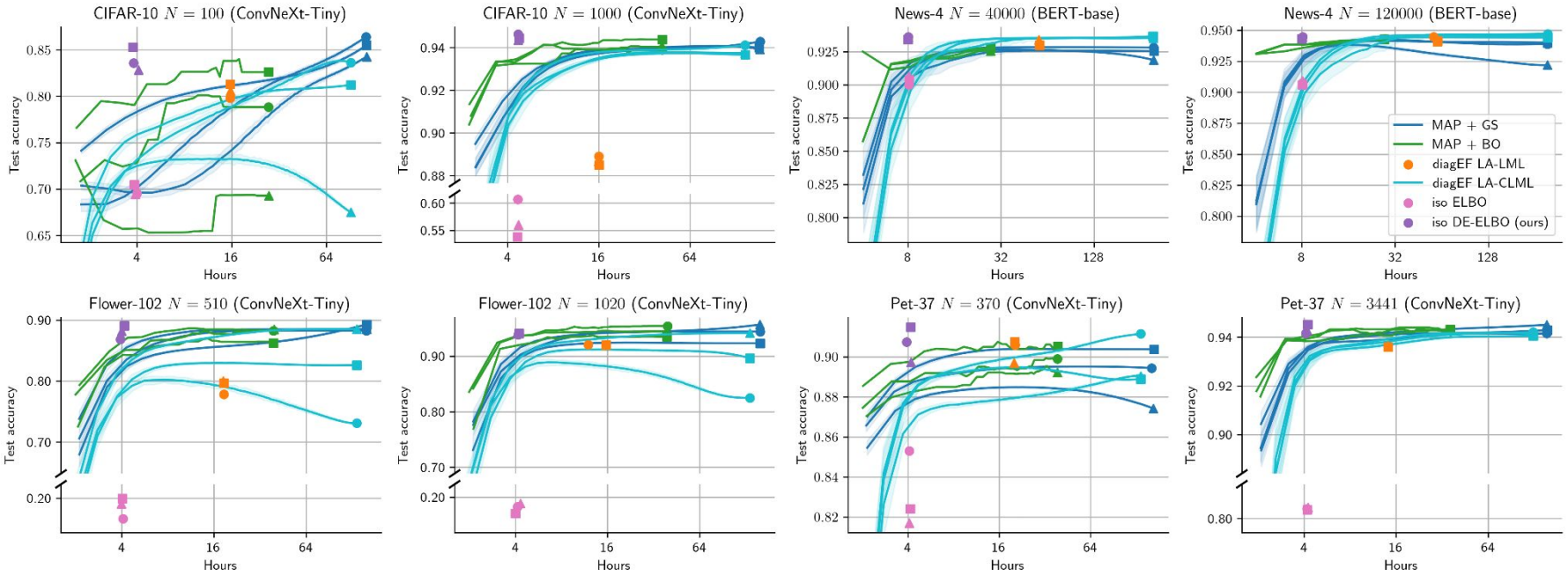


Figure: Test accuracy over time for ConvNeXt-Tiny (D > 1e7) on CIFAR-10 (N = 100).

Best Test Accuracy Given Limited Compute



Learning Hyperparameter via a Data-Emphasized Variational Objective

Poster Session: 1

Poster Number: 181

arXiv: arxiv.org/abs/2502.01861

GitHub: github.com/tufts-ml/data-emphasized-ELBO