3rd Symposium on Advances in Approximate Bayesian Inference

HWA: Hyperparameters Weight Averaging in Bayesian Neural Networks

Belhal Karimi and Ping Li
January 3rd 2021

Baidu Research, Cognitive Computing Lab

Agenda

1. Bayesian Deep Learning: Modeling and Training
2. HWA Method
3. Numerical Results

Intro: Bayesian Deep Learning

Modelling

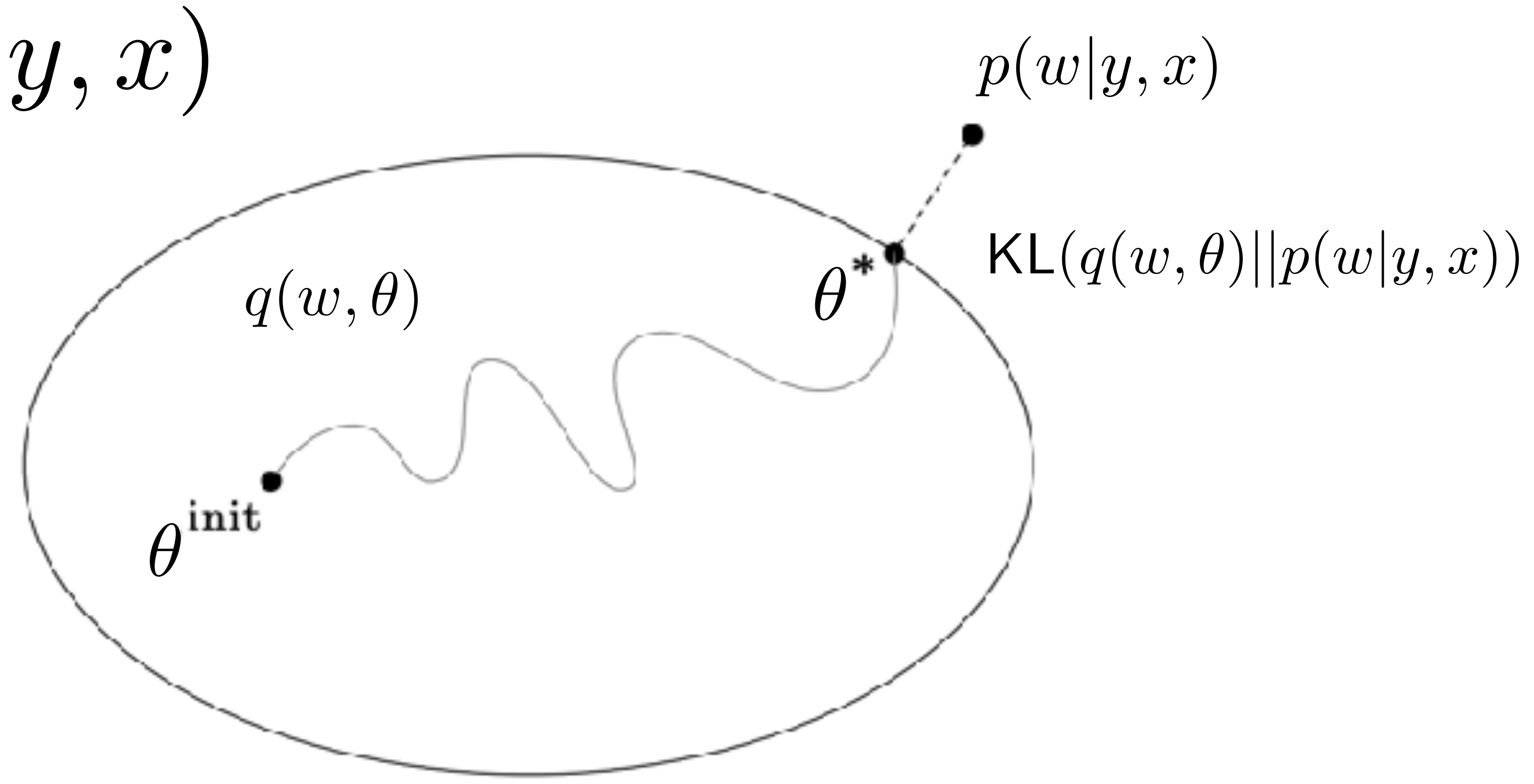
- **Can we combine the advantages of neural nets (Multi-layered yet fixed basis function) and Bayesian models (Posterior prediction, model averaging)?**
- **Bayesian Neural Networks (BNN):**
  - Input-output pairs $((x_i, y_i), 1 \leq i \leq n)$ and w the vector of weights on which we place a prior $\pi(w)$
  - Consider the following Classification problem $p(y_i|x_i, w) = \mathrm{Softmax}(f(x_i, w))$ where f is a Neural Network

Training: Objective Loss Function

- **How to train BNNs? —> Variational Inference (VI)**
  - Minimize the KL between the candidate $q(w, \theta)$ and the true posterior $p(w|y, x)$
  - KL term is intracable: VI optimizes the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta) := -\mathbb{E}_{q(w;\theta)}\big[\log p(y|x, w)\big] + \mathbb{E}_{q(w;\theta)}\big[\log q(w;\theta)/\pi(w)\big]$$

  - ELBO is a lower bound of the incomplete log likelihood —> Maximizing it minimizes the KL



SGD based Training

Current SGD for ELBO Maximization (VI)

- SGD on the hyperparameters of the weights (and no longer on the weights themselves). Assume Normal candidate

$$\mu_\ell^{k+1} = \mu_\ell^k - \gamma_{k+1} \nabla \mathcal{L}(\mu_\ell^k) \longrightarrow \text{Variational Proposal:} \quad \mathcal{N}(\mu^k, \Sigma^k)$$

Existing Relevant Methods

- Stochastic Gradient Langevin Dynamics (SGLD) [Welling & Teh, ICML 2011]:

$$\mu^{k+1} = \mu^k - \alpha_k \nabla \tilde{U}\left(\mu^k\right) + \sqrt{2\alpha_k}\epsilon_k \longrightarrow \text{Variational Proposal:} \quad \mathcal{N}(\mu^k - \alpha_k \nabla \tilde{U}\left(\mu^k\right), \Sigma_{SGLD}^k)$$

$$\nabla \tilde{U}(\mu^k) = \nabla \log p(w|y, \theta^k)$$

- Cyclical Stochastic Gradient MCMC (CSGMCMC) [Zhang et. al., ICLR 2020]:

$$\alpha_k^{\text{new}} = \frac{\alpha_0}{2}\left[\cos\left(\frac{\pi \bmod (k-1, \lceil K/M \rceil)}{\lceil K/M \rceil}\right) + 1\right] \xrightarrow{\text{SGLD}} \text{Variational Proposal:} \quad \mathcal{N}(\mu^k - \alpha_k^{\text{new}} \nabla \tilde{U}\left(\mu^k\right), \Sigma_{new}^k)$$

SGD based Training

Current SGD for ELBO Maximization (VI)

- SGD on the hyperparameters of the weights (and no longer on the weights themselves). Assume Normal candidate

$$\mu_\ell^{k+1} = \mu_\ell^k - \gamma_{k+1} \nabla \mathcal{L}(\mu_\ell^k) \longrightarrow \text{Variational Proposal:} \quad \mathcal{N}(\mu^k, \Sigma^k)$$

What other choice of proposal can be derived?

- Averaging procedure applied to proposal parameters (Diagonal covariance)

$$\mu_\ell^{HWA} = \frac{n_{\mathrm{m}} \mu_\ell^{HWA} + \mu_\ell^{k+1}}{n_{\mathrm{m}} + 1} \quad \text{and} \quad \sigma^{HWA} = \frac{n_{\mathrm{m}} \sigma^{HWA} + (\mu_\ell^{k+1})^2}{n_{\mathrm{m}} + 1} - (\mu_\ell^{HWA})^2$$

- Low rank plus diagonal proposal covariance matrix as in [Maddox et. al., 2019] in SWAG

$$\Sigma = \frac{1}{2} \Sigma_{\mathrm{diag}} + \frac{\widehat{D}\widehat{D}^\top}{2(R-1)} \qquad \widehat{D}_r = \theta_r - \theta_r^{HWA}$$

quantifies how far the current estimate parameter deviate from the current averaged parameter.

HWA in BNNs

HWA and its embedding in Variational Inference

- Periodic averaging of the hyperparameters

Algorithm 1 HWA: Hyperparameters Weight Averaging

Input: Iteration index k. Trained hyperparameters $\hat{\mu}_\ell$ and $\hat{\sigma}$. LR γ_k. Cycle length c. Gradient vector $\nabla \mathcal{L}_{i_k}(\theta^k)$

$\gamma \leftarrow \gamma(k)$ (Cyclical LR for the iteration)

SVI updates:

$\quad \mu_\ell^{k+1} \leftarrow \mu_\ell^k - \gamma_k \nabla_{\mu_\ell} \mathcal{L}_{i_k}(\mu_\ell^k)$

$\quad \sigma^{k+1} \leftarrow \sigma^k - \gamma_k \nabla_\sigma \mathcal{L}_{i_k}(\sigma^k)$

if $\mathrm{mod}(k, c) = 0$ **then**

$\quad n_\mathrm{m} \leftarrow k/c$ (Number of models to average over)

$$\mu_\ell^{HWA} \leftarrow \frac{n_\mathrm{m}\mu_\ell^{HWA} + \mu_\ell^{k+1}}{n_\mathrm{m}+1} \quad \text{and} \quad \sigma^{HWA} \leftarrow \frac{n_\mathrm{m}\sigma^{HWA} + (\mu_\ell^{k+1})^2}{n_\mathrm{m}+1} - (\mu_\ell^{HWA})^2$$

end if

Return: if mod$(k, c) = 0$, $(\{\mu_\ell^{HWA}\}_{l=1}^L, \sigma^{HWA})$ **else**, $(\{\mu_\ell^k\}_{l=1}^L, \sigma^k)$

HWA in BNNs

HWA and its embedding in Variational Inference

- Draw samples from candidate
- Compute MC integration of the expected gradient

- Plug HWA to obtain new mean and variance

Algorithm 2 Variational Inference with HWA for BNNs

Input: Trained hyperparameters $\hat{\mu}_\ell$ and $\hat{\sigma}$. Sequence of LR $\{\gamma_k\}_{k>0}$. Cycle length c. K iterations.
for $k = 0, 1, ...$ **do**
  Sample an index i_k uniformly on $[n]$
  Sample MC batch of weights $\{w_k^m\}_{m=1}^{M_k}$ from variational candidate $q(w, \theta^k)$ with $\theta^k = (\mu^k, \Sigma^k)$ and the covariance is either diagonal (4) or low rank (5).
  Compute MC approximation of the gradient vectors:

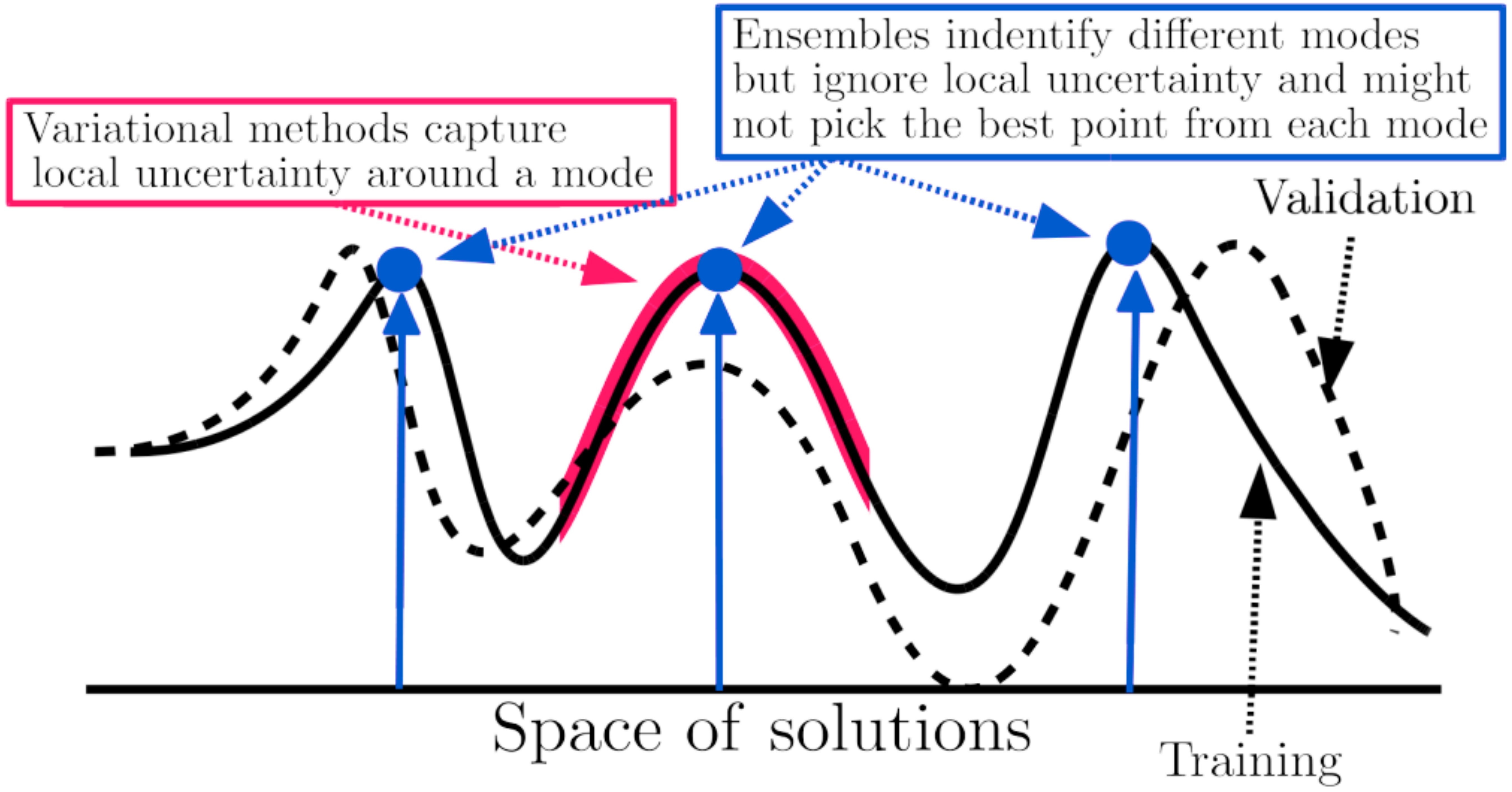
$$\nabla \mathcal{L}_{i_k}(\theta^k) \approx \frac{1}{M_k} \sum_{m=1}^{M_k} \log p(y_{i_k} | x_{i_k}, w_m^k) + \nabla KL(q(w, \theta^k) || \pi(w))$$

  Update the vector of parameter estimates calling Algorithm 1: $(\mu^K, \Sigma^K) = \text{HWA}(k, c, \gamma_k, \nabla \mathcal{L}_{i_k}(\theta^k))$
end for
Return Fitted parameters (μ^K, Σ^K).

Averaging in (Bayesian) NNs

Averaging Heuristics

- **Why Averaging makes sense?**
  - VI is unimodal (mode collapse)
  - Ensembles are great at training [Garipov et. al., NIPS 2018] (FGE) but bad at test time
    - Because of **Mode Connectivity (specific to NN)** and **Ensembling (Boosting)**
  - SWA averaging solution in [Izmailov et. al., UAI 2019]



From [Deep Ensembles: A Loss Landscape Perspective, Fort et. al., 2020]

- Ensemble of k models requires k times more computation.
- Averaging through the iterations can interpreted as an approximation to ensembles but with convenient test-time

Numerical Results

Bayesian LeNet and Bayesian VGG

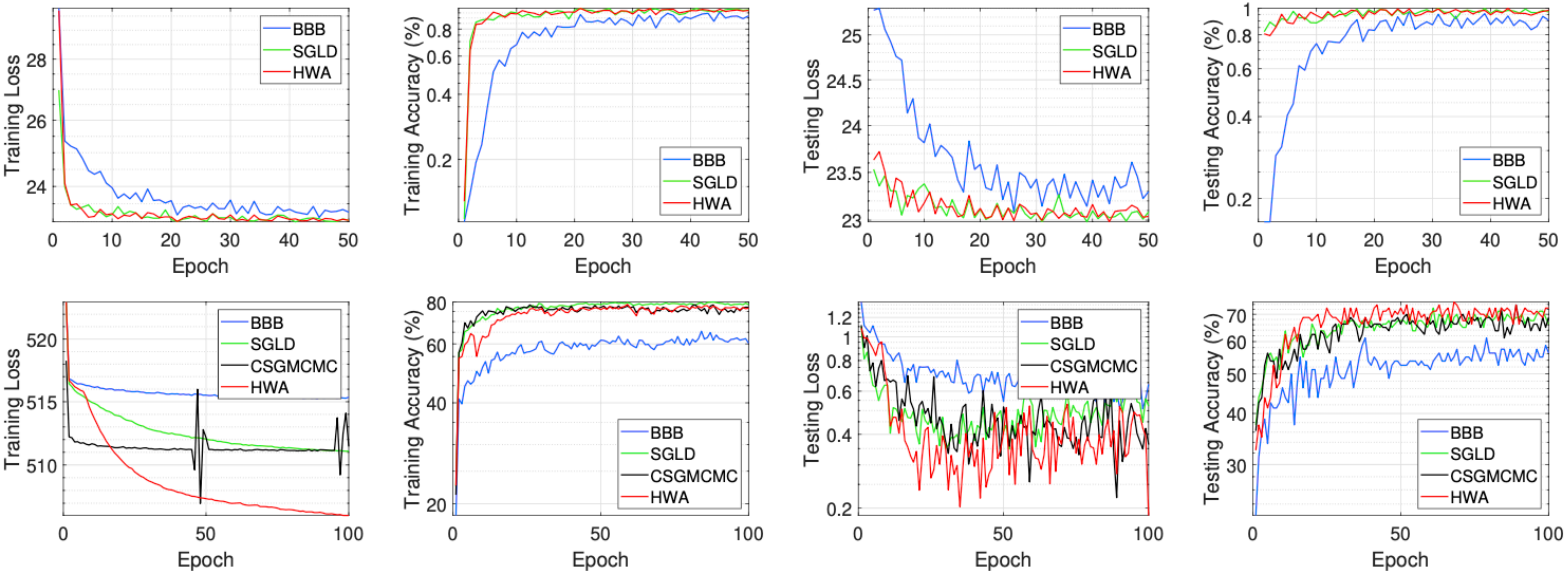


Figure 1: Comparison for Bayesian LeNet CNN architecture on MNIST dataset (top) and Bayesian VGG architecture on CIFAR-10 dataset (bottom). The plots are averaged over 5 repetitions.

Perspectives

Landscapes and Variance Reduction

- **Focus on the initial algorithm HWA: Hyperparameters Weight Averaging**
  - Plot Loss Landscape in 2D (PCA components)
    - Observe Mode Connectivity?
    - Observe Better Generalization?

- Find a better multiplier constant per « weak learner » (snapshots of the BNN in HWA)

Algorithm 1 HWA: Hyperparameters Weight Averaging

1: **Input:** Iteration index k. Trained hyperparameters $\hat{\mu}_\ell$ and $\hat{\sigma}$. LR γ_k. Cycle length c. Gradient vector $\nabla \mathcal{L}_{i_k}(\theta^k)$
2: $\gamma \leftarrow \gamma(k)$ (Cyclical LR for the iteration)
3: **SVI updates:**
4: $\quad \mu_\ell^{k+1} \leftarrow \mu_\ell^k - \gamma_k \nabla_{\mu_\ell} \mathcal{L}_{i_k}(\mu_\ell^k)$
5: $\quad \sigma^{k+1} \leftarrow \sigma^k - \gamma_k \nabla_\sigma \mathcal{L}_{i_k}(\sigma^k)$
6: **if** mod$(k, c) = 0$ **then**
7: $\quad n_{\text{m}} \leftarrow k/c \quad$ (Number of models to average over)

$$\mu_\ell^{HWA} \leftarrow \frac{n_{\text{m}}\mu_\ell^{HWA} + \mu_\ell^{k+1}}{n_{\text{m}} + 1} \quad \text{and} \quad \sigma^{HWA} \leftarrow \frac{n_{\text{m}}\sigma^{HWA} + (\mu_\ell^{k+1})^2}{n_{\text{m}} + 1} - (\mu_\ell^{HWA})^2$$

8: **end if**
9: **Return** hyperparameters $(\{\mu_\ell^{HWA}\}_{l=1}^L, \sigma^{HWA})$.

Change the multiplier (here it is 1/n_{models}) (cf. variance reduction method)

Thank You!

belhal.karimi@gmail.com