Structured Dropout Variational Inference for Bayesian neural networks

Son Nguyen[1], Duong Nguyen[3], Khai Nguyen[2], Khoat Than[1,3], Hung Bui[*,1], Nhat Ho[*,2]

[1]VinAI Research, Vietnam

[2]University of Texas, Austin

[3]Hanoi University of Science and Technology

35th Conference on Neural Information Processing Systems (NeurIPS 2021)

Bayesian neural networks (BNNs)

Why BNNs?

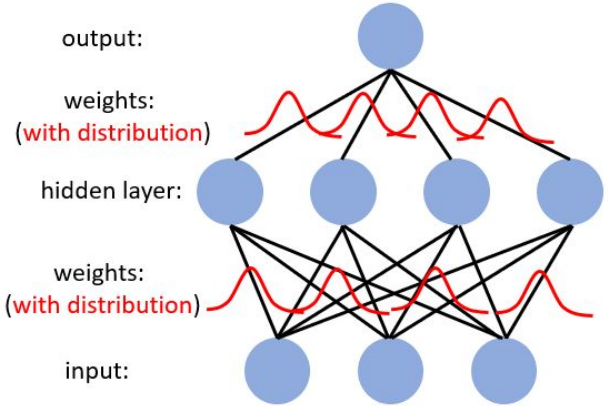
Limitation of deterministic neural nets:

- cannot properly represent uncertainty --> *miscalibrated prediction*
- not sufficiently robust: *overfit with small data, sensitive to ambiguous data*
- not sufficiently adaptive: *catastrophic forgetting*

Bayesian neural networks (BNNs)

What BNNs ?

- introduce random weights W with **prior distribution** $p(W)$
- infer a **posterior distribution** $p(W|\mathcal{D})$ instead of point estimates: $p(W|\mathcal{D}) \propto p(W)p(\mathcal{D}|W)$
- make predictions using the **posterior predictive distribution**: $p(y|x,\mathcal{D}) = \int p(W|\mathcal{D})p(y|W,x)dW$



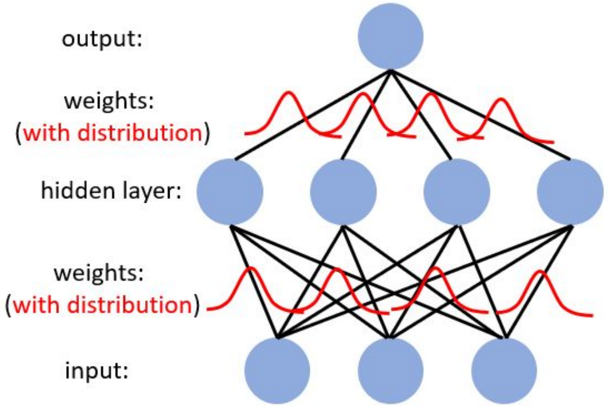
Bayesian neural networks (BNNs)

What BNNs ?

- introduce random weights W with **prior distribution** $p(W)$
- infer a **posterior distribution** $p(W|\mathcal{D})$ instead of point estimates: $p(W|\mathcal{D}) \propto p(W)p(\mathcal{D}|W)$
- make predictions using the **posterior predictive distribution**: $p(y|x,\mathcal{D}) = \int p(W|\mathcal{D})p(y|W,x)dW$

How BNNs ?

- **promising advantages:** *better generalization, robustness, uncertainty quantification, downstream tasks*



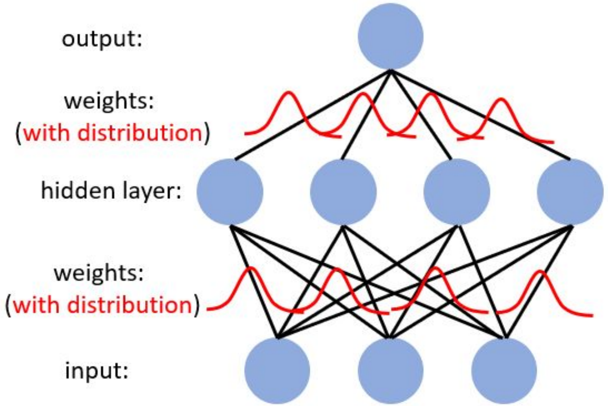
Bayesian neural networks (BNNs)

What BNNs ?

- introduce random weights W with **prior distribution** $p(W)$
- infer a **posterior distribution** $p(W|\mathcal{D})$ instead of point estimates: $p(W|\mathcal{D}) \propto p(W)p(\mathcal{D}|W)$
- make predictions using the **posterior predictive distribution**: $p(y|x,\mathcal{D}) = \int p(W|\mathcal{D})p(y|W,x)dW$

How BNNs ?

- **promising advantages:** *better generalization, robustness, uncertainty quantification, downstream tasks*
- but in practice, **exact inference is intractable**: *very high dimensionality, non-linearity*



Variational Inference for BNNs

Variational inference (VI) approximates the true posterior $p(\mathbf{W}|\mathcal{D})$ by a variational distribution $q_\phi(\mathbf{W})$ via optimizing ELBO:

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\mathbf{W})} \log p(\mathcal{D}|\mathbf{W}) - \mathbb{D}_{KL}(q_\phi(\mathbf{W})\|p(\mathbf{W}))$$

- **A central problem:** trade-off between **approximation expressiveness** and **computational efficiency**

Variational Inference for BNNs

Variational inference (VI) approximates the true posterior $p(\mathbf{W}|\mathcal{D})$ by a variational distribution $q_\phi(\mathbf{W})$ via optimizing ELBO:

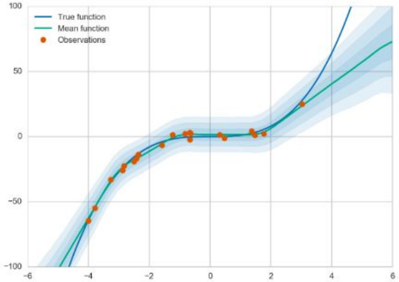
$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\mathbf{W})} \log p(\mathcal{D}|\mathbf{W}) - \mathbb{D}_{KL}(q_\phi(\mathbf{W}) \| p(\mathbf{W}))$$

❖ **A central problem:** trade-off between **approximation expressiveness** and **computational efficiency**

From the literature:

❖ **mean-field approximation:** ignores the strong statistical dependencies
- *underestimates posterior structure and model uncertainty*

mean-field



Variational Inference for BNNs

Variational inference (VI) approximates the true posterior $p(\mathbf{W}|\mathcal{D})$ by a variational distribution $q_\phi(\mathbf{W})$ via optimizing ELBO:

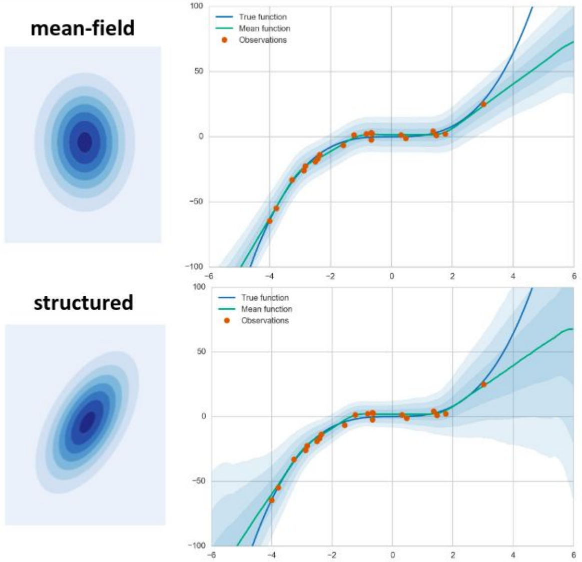
$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\mathbf{W})} \log p(\mathcal{D}|\mathbf{W}) - \mathbb{D}_{KL}(q_\phi(\mathbf{W}) \| p(\mathbf{W}))$$

❖ **A central problem:** trade-off between **approximation expressiveness** and **computational efficiency**

From the literature:

❖ **mean-field approximation:** ignores the strong statistical dependencies
- *underestimates posterior structure and model uncertainty*

❖ **richer or structured approximations:** Matrix Gaussian and variants, low-rank Gaussian, implicit distributions
- *improve both predictive accuracy and uncertainty calibration*
- *but some incur a large complexity & are difficult to integrate into CNNs*



Dropout Variational Inference for BNNs

What Dropout-VI ?

❖ interpret *Dropout regularization* in deterministic nns *as a form of approximate inference* in Bayesian deep models.

Dropout Variational Inference for BNNs

What Dropout-VI ?

- interpret *Dropout regularization* in deterministic nns *as a form of approximate inference* in Bayesian deep models.
- guaranteed via **KL-condition:** "*approximate Bayesian inference results in an identical objective to that of Dropout training*"

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(W)} \log p(\mathcal{D}|W) - \mathbb{D}_{KL}(q_\phi(W) \| p(W))$$

Dropout Variational Inference for BNNs

What Dropout-VI ?

- interpret *Dropout regularization* in deterministic nns *as a form of approximate inference* in Bayesian deep models.
- guaranteed via **KL-condition:** "*approximate Bayesian inference results in an identical objective to that of Dropout training*"

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(W)} \log p(\mathcal{D}|W) - \mathbb{D}_{KL}(q_\phi(W) \| p(W))$$

Dropout posterior

$$q_\phi(W) = \mathrm{Law}(\mathrm{diag}(\xi)\Theta)$$

ξ : Dropout noise

Θ : deterministic weight

$W := \mathrm{diag}(\xi)\Theta$: random weight

Dropout Variational Inference for BNNs

What Dropout-VI ?

- interpret ***Dropout regularization*** in deterministic nns ***as a form of approximate inference*** in Bayesian deep models.
- guaranteed via **KL-condition:** "*approximate Bayesian inference results in an identical objective to that of Dropout training*"

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(W)} \log p(\mathcal{D}|W) - \mathbb{D}_{KL}(q_\phi(W) \| p(W))$$

Dropout posterior

$q_\phi(W) = \text{Law}(\text{diag}(\xi)\Theta)$

ξ : Dropout noise

Θ : deterministic weight

$W := \text{diag}(\xi)\Theta$: random weight

ξ : Bernoulli noise $\mathcal{B}(p)$

$q_\phi(W) = \prod_{k=1}^{K} \left(p_k \mathcal{N}(\Theta_k, \sigma^2 \mathbf{I}_L) + (1 - p_k)\mathcal{N}(0, \sigma^2 \mathbf{I}_L) \right)$

$p(W) : \mathcal{N}(0, \lambda \mathbf{I})$

Dropout as a Bayesian Approximation:
Representing Model Uncertainty in Deep Learning

Yarin Gal YG279@CAM.AC.UK
Zoubin Ghahramani ZG201@CAM.AC.UK
University of Cambridge

ξ : Gaussian noise $\mathcal{N}(1, \text{diag}(\alpha))$

$q_\phi(W_{ij}) = \mathcal{N}(\Theta_{ij}, \alpha_i \Theta_{ij}^2)$

$p(|W_{ij}|) \propto 1/|w_{ij}|$

Variational Dropout and
the Local Reparameterization Trick

Diederik P. Kingma[*], Tim Salimans[×] and Max Welling[*†]

Dropout Variational Inference for BNNs

What Dropout-VI ?

- interpret *Dropout regularization* in deterministic nns *as a form of approximate inference* in Bayesian deep models.
- guaranteed via **KL-condition:** "*approximate Bayesian inference results in an identical objective to that of Dropout training*"

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(W)} \log p(\mathcal{D}|W) - \mathbb{D}_{KL}(q_\phi(W) \| p(W))$$

Dropout posterior

$q_\phi(W) = \mathrm{Law}(\mathrm{diag}(\xi)\Theta)$

ξ : Dropout noise

Θ : deterministic weight

$W := \mathrm{diag}(\xi)\Theta$: random weight

L2-Regularizer

$$\sum_{k=1}^{K} \frac{p_k l^2}{2N} \|\Theta_k\|_2^2$$

ξ : Bernoulli noise $\mathcal{B}(p)$

$q_\phi(W) = \prod_{k=1}^{K} \left(p_k \mathcal{N}(\Theta_k, \sigma^2 \mathbf{I}_L) + (1 - p_k)\mathcal{N}(0, \sigma^2 \mathbf{I}_L) \right)$

$p(W) : \mathcal{N}(0, \lambda \mathbf{I})$

**Dropout as a Bayesian Approximation:
Representing Model Uncertainty in Deep Learning**

Yarin Gal YG279@CAM.AC.UK
Zoubin Ghahramani ZG201@CAM.AC.UK
University of Cambridge

independent of Θ

$$\sum_{i=1}^{K} \left(0.5 \log(\alpha_i) + c_1 \alpha_i + c_2 \alpha_i^2 + c_3 \alpha_i^3 + C \right)$$

ξ : Gaussian noise $\mathcal{N}(1, \mathrm{diag}(\alpha))$

$q_\phi(W_{ij}) = \mathcal{N}(\Theta_{ij}, \alpha_i \Theta_{ij}^2)$

$p(|W_{ij}|) \propto 1/|w_{ij}|$

**Variational Dropout and
the Local Reparameterization Trick**

Diederik P. Kingma[*], Tim Salimans[×] and Max Welling[*†]

Dropout Variational Inference for BNNs

Why Dropout-VI ?

- **competitive accuracy** compared to structured VI, but with **much cheaper computational complexity**
- **complementary advantages** : *Bayesian inference* and *theoretical Dropout inductive biases*

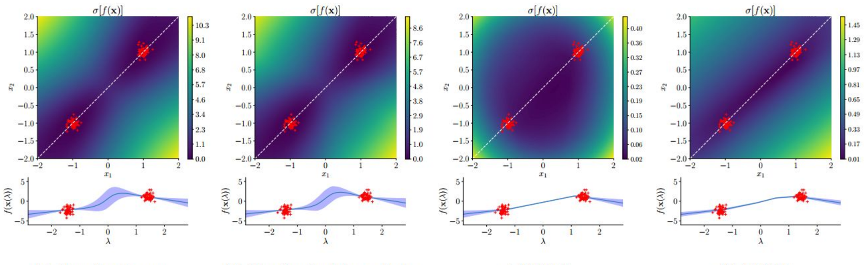
Dropout Variational Inference for BNNs

Why Dropout-VI ?

- **competitive accuracy** compared to structured VI, but with **much cheaper computational complexity**
- **complementary advantages**: *Bayesian inference* and *theoretical Dropout inductive biases*
- **research gap:** DVI also employed the simple structures of mean-field family for Dropout posterior

On the Expressiveness of Approximate Inference in Bayesian Neural Networks

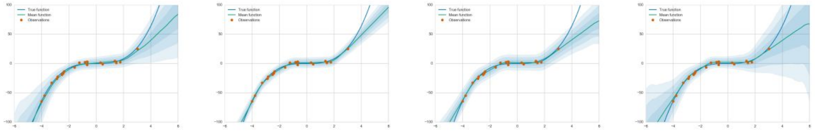
Andrew Y. K. Foong[*1] David R. Burt[*1] Yingzhen Li[2] Richard E. Turner[1 2]



(a) Gaussian Process (b) Hamiltonian Monte Carlo (c) MFVI (d) MCDO

Multiplicative Normalizing Flows for Variational Bayesian Neural Networks

Christos Louizos[1 2] Max Welling[1 3]



(a) Dropout $\pi = 0.5$ (b) Dropout learned π (c) FFLU (d) MNFG

ICML 2017*, empirical results*

NeurIPS 2020*, ...Theoretically, mean-field Gaussian and Dropout approximates cannot reasonably represent uncertainty"*

Dropout Variational Inference for BNNs

Intuition: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"

Dropout Variational Inference for BNNs

Intuition: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"

Challenges ?

1. maintain the backpropagation ***in parallel*** and optimize efficiently with gradient-based methods

Dropout Variational Inference for BNNs

Intuition: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"

Challenges ?

1. maintain the backpropagation *in parallel* and optimize efficiently with gradient-based methods
2. acquire *flexible* Bayesian inference in terms of both prior and approximate posterior, but guarantee **KL-condition**

Dropout Variational Inference for BNNs

Intuition: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"

Challenges ?

1. maintain the backpropagation *in parallel* and optimize efficiently with gradient-based methods
2. acquire *flexible* Bayesian inference in terms of both prior and approximate posterior , but guarantee **KL-condition**
3. address *theoretical pathologies* of previous Dropout-VI methods: non-Bayesian perspective

Variational Gaussian Dropout is not Bayesian

Jiri Hron
University of Cambridge
jh2084@cam.ac.uk

Alexander G. de G. Matthews
University of Cambridge
am554@cam.ac.uk

Zoubin Ghahramani
University of Cambridge, UBER AI Labs
zoubin@eng.cam.ac.uk

- *improper* prior --> ill-posed true posterior
- *singularity* in approximate posterior --> ELBO undefined

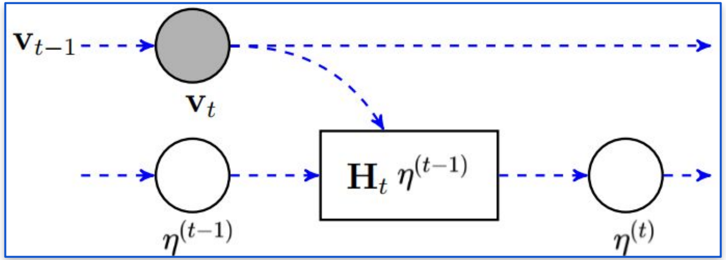
Variational Structured Dropout (VSD-our method)

- **Intuition**: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"
- **Approach:**
  - consider an original Dropout noise sampled from a Gaussian distribution: $\xi^{(0)} \sim \mathcal{N}(\mathbf{1}_K, \mathrm{diag}(\alpha))$

Variational Structured Dropout (VSD-our method)

- **Intuition**: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"
- **Approach:**
  - consider an original Dropout noise sampled from a Gaussian distribution: $\xi^{(0)} \sim \mathcal{N}(\mathbf{1}_K, \mathrm{diag}(\alpha))$
  - extract $\xi^{(0)} = 1 + \eta^{(0)}$ and successively transform $\eta^{(0)}$ through a chain of T Householder reflections

$$\xi^{(t)} := 1 + H_t H_{t-1} \dots H_1 \eta^{(0)} = 1 + U\eta^{(0)}$$



$\mathbf{v}_t = \text{FC-layer}(\mathbf{v}_{t-1})$

$\mathbf{H}_t = \mathbf{I} - 2\frac{\mathbf{v}_t \mathbf{v}_t^T}{\|\mathbf{v}_t\|_2^2}$ is called the Householder matrix

Variational Structured Dropout (VSD-our method)

- **Intuition**: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"
- **Approach:**
  - consider an original Dropout noise sampled from a Gaussian distribution: $\xi^{(0)} \sim \mathcal{N}(\mathbf{1}_K, \mathrm{diag}(\alpha))$
  - extract $\xi^{(0)} = 1 + \eta^{(0)}$ and successively transform $\eta^{(0)}$ through a chain of T Householder reflections

$$\xi^{(t)} := 1 + H_t H_{t-1} \dots H_1 \eta^{(0)} = 1 + U\eta^{(0)} \rightarrow \xi^{(t)} \sim \mathcal{N}(\mathbf{1}_K, U\mathrm{diag}(\alpha)U^T)$$

Variational Structured Dropout (VSD-our method)

- **Intuition**: "***a richer representation for variational noise could enrich Dropout posterior expressiveness***"
- **Approach:**
  - consider an original Dropout noise sampled from a Gaussian distribution: $\xi^{(0)} \sim \mathcal{N}(\mathbf{1}_K, \mathrm{diag}(\alpha))$
  - extract $\xi^{(0)} = 1 + \eta^{(0)}$ and successively transform $\eta^{(0)}$ through a chain of T Householder reflections

$$\xi^{(t)} := 1 + H_t H_{t-1} \ldots H_1 \eta^{(0)} = 1 + U\eta^{(0)} \rightarrow \xi^{(t)} \sim \mathcal{N}(\mathbf{1}_K, U\mathrm{diag}(\alpha)U^T)$$

  - inject structured noise $\xi^{(t)}$ into deterministic weight Θ:

$$\mathbf{W}^{(t)} := \mathrm{diag}(\xi^{(t)})\Theta$$

$$q_t(\mathbf{W}) = \mathrm{Law}(\mathrm{diag}(\xi^{(t)})\Theta)$$

Variational Structured Dropout (VSD-our method)

- **Intuition**: "*a richer representation for variational noise could enrich Dropout posterior expressiveness*"
- **Approach:**
  - consider an original Dropout noise sampled from a Gaussian distribution: $\xi^{(0)} \sim \mathcal{N}(\mathbf{1}_K, \mathrm{diag}(\alpha))$
  - extract $\xi^{(0)} = 1 + \eta^{(0)}$ and successively transform $\eta^{(0)}$ through a chain of T Householder reflections

$$\xi^{(t)} := 1 + H_t H_{t-1} \dots H_1 \eta^{(0)} = 1 + U\eta^{(0)} \rightarrow \xi^{(t)} \sim \mathcal{N}(\mathbf{1}_K, U\mathrm{diag}(\alpha)U^T)$$

  - inject structured noise $\xi^{(t)}$ into deterministic weight Θ:

$$\mathbf{W}^{(t)} := \mathrm{diag}(\xi^{(t)})\Theta$$

$$q_t(\mathbf{W}) = \mathrm{Law}(\mathrm{diag}(\xi^{(t)})\Theta)$$

ELBO → VSD

$$\mathcal{L}(\phi) := \mathbb{E}_{q_t(\mathbf{W})} \log p(\mathcal{D}|\mathbf{W}) - \mathbb{D}_{KL}(q_t(\mathbf{W})||p(\mathbf{W}))$$
$$= \mathbb{E}_{q_\alpha(\xi)} \log p(\mathcal{D}|\Theta, \xi^{(t)}) - \mathbb{D}_{KL}(q_t(\mathbf{W})||p(\mathbf{W}))$$

Variational Structured Dropout (VSD-our method)

❖ **Contribution 1:** VSD overcomes the **singularity issue** of approximate posterior in VD

$$\mathbf{W}^{(VD)} = \text{diag}(\xi^{(0)})\Theta = \Theta + \text{diag}(\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}(\text{diag}(\mathbf{e}_i)\Theta) = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}\Theta_{(i)}$$

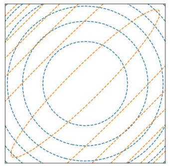
$$\mathbf{W}^{(VSD)} = \text{diag}(\xi^{(t)})\Theta = \Theta + \text{diag}(\mathbf{U}\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}(\text{diag}(\mathbf{U}_{i:})\Theta)$$

Variational Structured Dropout (VSD-our method)

❖ **Contribution 1:** VSD overcomes the **singularity issue** of approximate posterior in VD

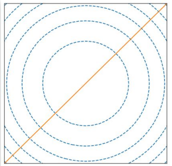
$$\mathbf{W}^{(VD)} = \text{diag}(\xi^{(0)})\Theta = \Theta + \text{diag}(\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)} (\text{diag}(\mathbf{e}_i)\Theta) = \Theta + \sum_{i=1}^{K} \eta_i^{(0)} \Theta_{(i)}$$

$$\mathbf{W}^{(VSD)} = \text{diag}(\xi^{(t)})\Theta = \Theta + \text{diag}(\mathbf{U}\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)} (\text{diag}(\mathbf{U}_{i:})\Theta)$$



No

Yes



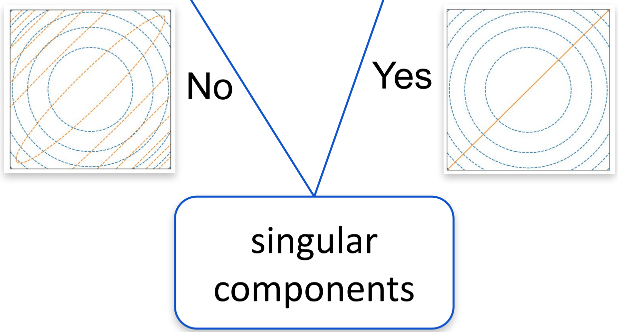
singular components

Variational Structured Dropout (VSD-our method)

❖ **Contribution 1:** VSD overcomes the **singularity issue** of approximate posterior in VD

$$\mathbf{W}^{(VD)} = \mathrm{diag}(\xi^{(0)})\Theta = \Theta + \mathrm{diag}(\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}(\mathrm{diag}(\mathbf{e}_i)\Theta) = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}\Theta_{(i)}$$

$$\mathbf{W}^{(VSD)} = \mathrm{diag}(\xi^{(t)})\Theta = \Theta + \mathrm{diag}(\mathbf{U}\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}(\mathrm{diag}(\mathbf{U}_{i:})\Theta)$$



$$\mathcal{KL}_B[\mu||\eta] = \sum_{i=1}^{|B|} \mu(B_i) \log \frac{\mu(B_i)}{\eta(B_i)}$$

infinite in VD, well-defined in VSD

Variational Structured Dropout (VSD-our method)

❖ **Contribution 1:** VSD overcomes the **singularity issue** of approximate posterior in VD

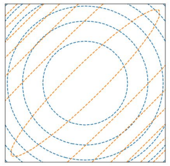
$$\mathbf{W}^{(VD)} = \text{diag}(\xi^{(0)})\Theta = \Theta + \text{diag}(\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}(\text{diag}(\mathbf{e}_i)\Theta) = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}\Theta_{(i)}$$

$$\mathbf{W}^{(VSD)} = \text{diag}(\xi^{(t)})\Theta = \Theta + \text{diag}(\mathbf{U}\eta^{(0)})\Theta = \Theta + \sum_{i=1}^{K} \eta_i^{(0)}(\text{diag}(\mathbf{U}_{i:})\Theta)$$

$$\mathbb{D}_{KL}(q_t(\mathbf{W})||p(\mathbf{W}))$$

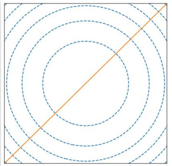
$$q_t(\mathbf{W}) = \text{Law}(\text{diag}(\xi^{(t)})\Theta)$$

is validly defined, but how to analyze ?



No

Yes



singular components

Variational Structured Dropout (VSD-our method)

❖ **Approach (cont'd):**

- consider isotropic Gaussian prior: $p(\mathbf{W}) = \prod_{j=1}^{Q} p(\mathbf{W}_{:j})$ with $p(\mathbf{W}_{:j}) = \mathcal{N}(0, \text{diag}(\beta_{:j}^{-1}))$

Variational Structured Dropout (VSD-our method)

❖ **Approach (cont'd):**

- consider isotropic Gaussian prior: $p(\mathbf{W}) = \prod_{j=1}^{Q} p(\mathbf{W}_{:j})$ with $p(\mathbf{W}_{:j}) = \mathcal{N}(0, \mathrm{diag}(\beta_{:j}^{-1}))$
- augment a **mutual information** term $\mathbf{I}(\mathbf{W}_{:1}, \mathbf{W}_{:2}, ..., \mathbf{W}_{:Q})$ to encourage correlations:

$$\begin{aligned}\mathcal{L}_{MI}(\phi) &:= \mathbb{E}_{q_\alpha(\xi)} \log p(\mathcal{D}|\Theta, \xi^{(t)}) - \mathbb{D}_{KL}(q_t(\mathbf{W})||p(\mathbf{W})) + \mathbf{I}(\mathbf{W}_{:1}, \mathbf{W}_{:2}, ..., \mathbf{W}_{:Q}) \\ &= \mathbb{E}_{q_\alpha(\xi)} \log p(\mathcal{D}|\Theta, \xi^{(t)}) - \mathbb{D}_{KL}(q_t^\star(\mathbf{W})||p(\mathbf{W})),\end{aligned}$$

with $q_t^\star(\mathbf{W}) := \prod_{j=1}^{Q} q_t(\mathbf{W}_{:j})$

Variational Structured Dropout (VSD-our method)

❖ **Approach (cont'd):**

- consider isotropic Gaussian prior: $p(\mathbf{W}) = \prod_{j=1}^{Q} p(\mathbf{W}_{:j})$ with $p(\mathbf{W}_{:j}) = \mathcal{N}(0, \text{diag}(\beta_{:j}^{-1}))$
- augment a **mutual information** term $\mathbf{I}(\mathbf{W}_{:1}, \mathbf{W}_{:2}, ..., \mathbf{W}_{:Q})$ to encourage correlations:

$$\begin{aligned}\mathcal{L}_{MI}(\phi) &:= \mathbb{E}_{q_\alpha(\xi)} \log p(\mathcal{D}|\Theta, \xi^{(t)}) - \mathbb{D}_{KL}(q_t(\mathbf{W})||p(\mathbf{W})) + \mathbf{I}(\mathbf{W}_{:1}, \mathbf{W}_{:2}, ..., \mathbf{W}_{:Q}) \\ &= \mathbb{E}_{q_\alpha(\xi)} \log p(\mathcal{D}|\Theta, \xi^{(t)}) - \mathbb{D}_{KL}(q_t^\star(\mathbf{W})||p(\mathbf{W})),\end{aligned}$$

with $q_t^\star(\mathbf{W}) := \prod_{j=1}^{Q} q_t(\mathbf{W}_{:j})$

- use Empirical Bayes (EB) to specify β:

$$\mathbb{D}_{KL}^{EB}(q_t^\star(\mathbf{W})||p(\mathbf{W})) = \frac{Q}{2}\sum_{i=1}^{K} \log \frac{1 + \sum_{j=1}^{K} \alpha_j U_{ij}^2}{\alpha_i}$$

→ **KL condition**

Variational Structured Dropout (VSD-our method)

❖ **Contribution 2:** VSD is flexible in terms of both approximate posterior and prior distribution.

- expand hierarchically prior distribution: $p(\mathbf{W}, \mathbf{z}) = p(\mathbf{W}|\mathbf{z}, \beta)p(\mathbf{z})$
- do joint inference with Dropout posterior: $q_t(\mathbf{W}, \mathbf{z}) = q_\psi(\mathbf{z})q_t(\mathbf{W}|\mathbf{z})$

Variational Structured Dropout (VSD-our method)

❖ **Contribution 2:** VSD is flexible in terms of both approximate posterior and prior distribution.

- expand hierarchically prior distribution: $p(\mathbf{W}, \mathbf{z}) = p(\mathbf{W}|\mathbf{z}, \beta)p(\mathbf{z})$
- do joint inference with Dropout posterior: $q_t(\mathbf{W}, \mathbf{z}) = q_\psi(\mathbf{z})q_t(\mathbf{W}|\mathbf{z})$
- drawing $\mathbf{W} \sim q_t(\mathbf{W}, \mathbf{z})$ follows a hierarchical Dropout procedure:

$$\mathbf{z} \sim q_\psi(\mathbf{z}), \quad \xi_n^{(t)} \sim \mathcal{N}(\mathbf{1}_K, \mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T)$$

Variational Structured Dropout (VSD-our method)

❖ **Contribution 2:** VSD is flexible in terms of both approximate posterior and prior distribution.

- expand hierarchically prior distribution: $p(\mathbf{W}, \mathbf{z}) = p(\mathbf{W}|\mathbf{z}, \beta)p(\mathbf{z})$
- do joint inference with Dropout posterior: $q_t(\mathbf{W}, \mathbf{z}) = q_\psi(\mathbf{z})q_t(\mathbf{W}|\mathbf{z})$
- drawing $\mathbf{W} \sim q_t(\mathbf{W}, \mathbf{z})$ follows a hierarchical Dropout procedure:

$$\mathbf{z} \sim q_\psi(\mathbf{z}), \quad \xi_n^{(t)} \sim \mathcal{N}(\mathbf{1}_K, \mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T)$$

↓

$$\widehat{\xi}_n^{(t)} = \mathbf{z} \odot \xi_n^{(t)}$$

Variational Structured Dropout (VSD-our method)

❖ **Contribution 2:** VSD is flexible in terms of both approximate posterior and prior distribution.

- expand hierarchically prior distribution: $p(\mathbf{W}, \mathbf{z}) = p(\mathbf{W}|\mathbf{z}, \beta)p(\mathbf{z})$
- do joint inference with Dropout posterior: $q_t(\mathbf{W}, \mathbf{z}) = q_\psi(\mathbf{z})q_t(\mathbf{W}|\mathbf{z})$
- drawing $\mathbf{W} \sim q_t(\mathbf{W}, \mathbf{z})$ follows a hierarchical Dropout procedure:

$$\mathbf{z} \sim q_\psi(\mathbf{z}), \quad \xi_n^{(t)} \sim \mathcal{N}(\mathbf{1}_K, \mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T)$$

$$\widehat{\xi}_n^{(t)} = \mathbf{z} \odot \xi_n^{(t)}$$

$$\mathbf{W} = \mathrm{diag}(\widehat{\xi}_n^{(t)})\Theta$$

Variational Structured Dropout (VSD-our method)

❖ **Contribution 2:** VSD is flexible in terms of both approximate posterior and prior distribution.

- expand hierarchically prior distribution: $p(\mathbf{W}, \mathbf{z}) = p(\mathbf{W}|\mathbf{z}, \beta)p(\mathbf{z})$
- do joint inference with Dropout posterior: $q_t(\mathbf{W}, \mathbf{z}) = q_\psi(\mathbf{z})q_t(\mathbf{W}|\mathbf{z})$
- drawing $\mathbf{W} \sim q_t(\mathbf{W}, \mathbf{z})$ follows a hierarchical Dropout procedure:

$$\mathbf{z} \sim q_\psi(\mathbf{z}), \quad \xi_n^{(t)} \sim \mathcal{N}(\mathbf{1}_K, \mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T)$$

$$\widehat{\xi}_n^{(t)} = \mathbf{z} \odot \xi_n^{(t)}$$

$$\mathbf{W} = \mathrm{diag}(\widehat{\xi}_n^{(t)})\Theta$$

- satisfy the KL condition **w/o further simplifying assumption**

Variational Structured Dropout (VSD-our method)

❖ **Contribution 3:** VSD induces an adaptive regularization with several desirable inductive biases

$$R_{VSD} = \mathbb{E}_{(\mathbf{x} \sim \mathcal{B})} \sum_{i=1}^{L} \langle \mathbf{H}_i, \mathrm{diag}(\mathbf{h}_i) \mathbf{U} \mathrm{diag}(\alpha) \mathbf{U}^T \mathrm{diag}(\mathbf{h}_i) \rangle$$

Variational Structured Dropout (VSD-our method)

- **Contribution 3:** VSD induces an adaptive regularization with several desirable inductive biases

$$R_{VSD} = \mathbb{E}_{(\mathbf{x}\sim\mathcal{B})} \sum_{i=1}^{L} \langle \mathbf{H}_i, \mathrm{diag}(\mathbf{h}_i)\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i)\rangle$$

$$\begin{aligned} R_{VSD}^{(i)} &\approx \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \mathbf{Trace}\left(\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{H}_i(\mathbf{x})\right) \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \|\mathbf{H}_i^{1/2}(\mathbf{x})\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha^{1/2})\|_F^2. \end{aligned}$$

Variational Structured Dropout (VSD-our method)

❖ **Contribution 3:** VSD induces an adaptive regularization with several desirable inductive biases

$$R_{VSD} = \mathbb{E}_{(\mathbf{x}\sim\mathcal{B})} \sum_{i=1}^{L} \langle \mathbf{H}_i, \mathrm{diag}(\mathbf{h}_i)\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i)\rangle$$

$$R_{VSD}^{(i)} \approx \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \mathbf{Trace}\left(\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{H}_i(\mathbf{x})\right)$$
$$= \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \|\mathbf{H}_i^{1/2}(\mathbf{x})\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha^{1/2})\|_F^2.$$

$$\mathbf{Q} := \Gamma\Gamma^T = \mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))$$

$$= \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \left(\Theta^{[i:L]}\mathbf{Q}\Theta^{[i:L].T}\right)$$

VSD imposes a *Tikhonov-like regularization* and *reshapes the gradient*.

$$\Omega_i := \mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{J}_i^T(\mathbf{x})\mathbf{H}_{\mathrm{out}}(\mathbf{x})\mathbf{J}_i(\mathbf{x})\mathrm{diag}(\mathbf{h}_i\mathbf{x})$$

$$= \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \sum_{k=1}^{K} \alpha_k \mathbf{U}_{:k}^T \Omega_i \mathbf{U}_{:k}$$

VSD penalizes implicitly the *spectral norm* of weight matrices

Variational Structured Dropout (VSD-our method)

❖ **Contribution 3:** VSD induces an adaptive regularization with several desirable inductive biases

$$R_{VSD} = \mathbb{E}_{(\mathbf{x}\sim\mathcal{B})} \sum_{i=1}^{L} \langle \mathbf{H}_i, \mathrm{diag}(\mathbf{h}_i)\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i) \rangle$$

$$R^{(i)}_{VSD} \approx \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \mathbf{Trace}\left(\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{H}_i(\mathbf{x})\right)$$
$$= \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \|\mathbf{H}_i^{1/2}(\mathbf{x})\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha^{1/2})\|_F^2.$$

$$\mathbf{Q} := \Gamma\Gamma^T = \mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{U}\mathrm{diag}(\alpha)\mathbf{U}^T\mathrm{diag}(\mathbf{h}_i(\mathbf{x}))$$

$$\Omega_i := \mathrm{diag}(\mathbf{h}_i(\mathbf{x}))\mathbf{J}_i^T(\mathbf{x})\mathbf{H}_{\mathrm{out}}(\mathbf{x})\mathbf{J}_i(\mathbf{x})\mathrm{diag}(\mathbf{h}_i\mathbf{x})$$

$$= \mathbb{E}_{\mathbf{x}\sim\mathcal{B}}\left(\Theta^{[i:L]}\mathbf{Q}\Theta^{[i:L].T}\right)$$

$$= \mathbb{E}_{\mathbf{x}\sim\mathcal{B}} \sum_{k=1}^{K} \alpha_k \mathbf{U}_{:k}^T \Omega_i \mathbf{U}_{:k}$$

VSD imposes a ***Tikhonov-like regularization*** and ***reshapes the gradient***.

complementary advantages

VSD penalizes implicitly the ***spectral norm*** of weight matrices

Variational Structured Dropout (VSD-our method)

❖ **Contribution 4:** VSD gains noticeable empirical results compared to other variational methods.

- **Regression task:**

Table 10: Average test performance for UCI regression task. Results are reported with RMSE and Std. Errors.

Dataset	BBB	VMG	MNF	SLANG	MCD	VD	D.E	VSD
Boston	3.43 ± 0.20	2.70 ± 0.13	2.98 ± 0.06	3.21 ± 0.19	2.83 ± 0.17	2.98 ± 0.18	3.28 ± 0.22	**2.64 ± 0.17**
Concrete	6.16 ± 0.13	4.89 ± 0.12	6.57 ± 0.04	5.58 ± 0.19	4.93 ± 0.14	5.16 ± 0.13	6.03 ± 0.13	**4.72 ± 0.11**
Energy	0.97 ± 0.09	0.54 ± 0.02	2.38 ± 0.07	0.64 ± 0.03	1.08 ± 0.03	0.64 ± 0.02	2.09 ± 0.06	**0.47 ± 0.01**
Kin8nm	0.08 ± 0.00	0.08 ± 0.00	0.09 ± 0.00	0.08 ± 0.00	0.09 ± 0.00	0.08 ± 0.00	0.09 ± 0.00	**0.08 ± 0.00**
Naval	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.00 ± 0.00**
Power Plant	4.21 ± 0.03	4.04 ± 0.04	4.19 ± 0.01	4.16 ± 0.04	4.00 ± 0.04	3.99 ± 0.03	4.11 ± 0.04	**3.92 ± 0.04**
Wine	0.64 ± 0.01	0.63 ± 0.01	**0.61 ± 0.00**	0.65 ± 0.01	0.61 ± 0.01	0.62 ± 0.01	0.64 ± 0.00	0.63 ± 0.01
Yacht	1.13 ± 0.06	0.71 ± 0.05	2.13 ± 0.05	1.08 ± 0.06	0.72 ± 0.05	1.09 ± 0.09	1.58 ± 0.11	**0.69 ± 0.06**

Table 11: Average test performance for UCI regression task. Results are reported with test LL and Std. Errors.

Dataset	BBB	VMG	MNF	SLANG	MCD	VD	D.E	VSD
Boston	-2.66 ± 0.06	-2.46 ± 0.09	-2.51 ± 0.06	-2.58 ± 0.05	-2.40 ± 0.04	-2.39 ± 0.04	-2.41 ± 0.06	**-2.35 ± 0.05**
Concrete	-3.25 ± 0.02	-3.01 ± 0.03	-3.35 ± 0.04	-3.13 ± 0.03	**-2.97 ± 0.02**	-3.07 ± 0.03	-3.06 ± 0.04	**-2.97 ± 0.02**
Energy	-1.45 ± 0.10	-1.06 ± 0.03	-3.18 ± 0.07	-1.12 ± 0.01	-1.72 ± 0.01	-1.30 ± 0.01	-1.38 ± 0.05	**-1.06 ± 0.01**
Kin8nm	1.07 ± 0.00	1.10 ± 0.01	1.04 ± 0.00	1.06 ± 0.00	0.97 ± 0.00	1.14 ± 0.01	1.20 ± 0.00	**1.17 ± 0.01**
Naval	4.61 ± 0.01	2.46 ± 0.00	3.96 ± 0.01	4.76 ± 0.00	4.76 ± 0.01	4.81 ± 0.00	5.63 ± 0.00	**4.83 ± 0.01**
Power Plant	-2.86 ± 0.01	-2.82 ± 0.01	-2.86 ± 0.01	-2.84 ± 0.01	**-2.79 ± 0.01**	-2.82 ± 0.01	-2.79 ± 0.01	**-2.79 ± 0.01**
Wine	-0.97 ± 0.01	-0.95 ± 0.01	-0.93 ± 0.00	-0.97 ± 0.01	**-0.92 ± 0.01**	-0.94 ± 0.01	-0.94 ± 0.03	-0.95 ± 0.01
Yacht	-1.56 ± 0.02	-1.30 ± 0.02	-1.96 ± 0.05	-1.88 ± 0.01	-1.38 ± 0.01	-1.42 ± 0.02	-1.18 ± 0.05	**-1.14 ± 0.02**

Variational Structured Dropout (VSD-our method)

❖ **Contribution 4:** VSD gains noticeable empirical results compared to other variational methods.

- **Image classification task:**

Table 4: Image classification using AlexNet architecture. Results are averaged over 5 random seeds.

AlexNet	CIFAR10 NLL	CIFAR10 ACC	CIFAR10 ECE	CIFAR100 NLL	CIFAR100 ACC	CIFAR100 ECE	SVHN NLL	SVHN ACC	SVHN ECE	STL10 NLL	STL10 ACC	STL10 ECE
MAP	1.038	69.58	0.121	4.705	40.23	0.393	0.418	87.56	0.033	2.532	65.70	0.267
BBB	0.994	65.38	0.062	2.659	32.41	**0.049**	0.476	87.30	0.094	1.707	65.46	0.222
MCD	0.717	75.22	0.023	2.503	42.91	0.151	0.401	88.03	0.023	1.059	63.65	**0.052**
VD	0.702	77.28	**0.028**	2.582	43.10	0.106	0.327	90.76	0.010	2.130	65.48	0.195
ELRG	0.723	76.87	0.065	2.368	42.90	0.099	0.312	90.66	**0.006**	1.088	59.99	**0.018**
VSD	**0.656**	**78.21**	**0.046**	**2.241**	**46.85**	0.112	**0.290**	**91.62**	**0.008**	**1.019**	**67.98**	0.079
D.E	0.872	77.56	0.115	3.402	46.42	0.314	0.319	90.30	**0.008**	2.229	**68.51**	0.241
SWAG	**0.651**	**78.14**	0.059	**1.958**	**49.81**	**0.028**	0.331	90.04	0.031	1.522	**68.41**	0.161

Table 5: Image classification using ResNet18 architecture. Results are averaged over 5 random seeds.

ResNet18	CIFAR10 NLL	CIFAR10 ACC	CIFAR10 ECE	CIFAR100 NLL	CIFAR100 ACC	CIFAR100 ECE	SVHN NLL	SVHN ACC	SVHN ECE	STL10 NLL	STL10 ACC	STL10 ECE
MAP	0.644	86.34	0.093	2.410	55.38	0.243	0.232	95.32	0.028	1.401	71.26	0.199
BBB	0.697	76.63	0.071	2.239	41.07	0.100	0.218	94.53	0.047	1.290	71.55	0.179
MCD	0.534	87.47	0.084	2.121	59.28	0.227	0.207	95.78	0.026	1.333	72.28	0.188
VD	0.451	87.68	**0.024**	2.888	56.80	0.284	0.164	96.11	0.017	1.084	73.29	0.084
ELRG	**0.382**	87.24	**0.018**	1.634	58.14	**0.096**	0.145	96.03	**0.003**	0.811	73.66	**0.080**
VSD	0.464	87.44	0.061	**1.504**	**60.15**	0.116	**0.140**	**96.41**	**0.003**	**0.769**	**74.50**	**0.083**
D.E	0.488	**88.91**	0.069	1.913	**60.16**	0.203	0.171	**96.36**	0.020	1.197	73.16	0.177
SWAG	**0.330**	**88.77**	0.026	**1.417**	**62.45**	**0.028**	**0.130**	**96.72**	0.016	0.843	73.15	**0.069**

Variational Structured Dropout (VSD-our method)

❖ **Contribution 4:** VSD gains noticeable empirical results compared to other variational methods.

- **Predictive entropy:**

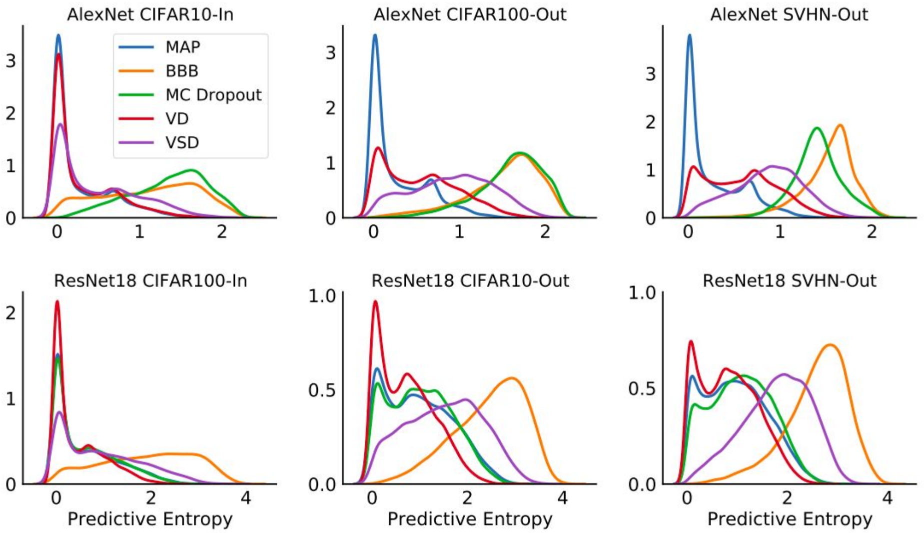


Figure 4: Histograms of predictive entropy for AlexNet (top) and ResNet18 (bottom) trained on CIFAR10 and CIFAR100 respectively.

Variational Structured Dropout (VSD-our method)

❖ **Contribution 4:** VSD gains noticeable empirical results compared to other variational methods.

- **OOD metrics:**

LeNet-5 (SVHN)	CIFAR10					CIFAR100				
	FPR	Det. err.	AUROC	AUPR IN	AUPR OUT	FPR	Det. err.	AUROC	AUPR IN	AUPR OUT
MAP	0.78	0.23	0.83	0.93	0.58	0.76	0.22	0.84	0.93	0.60
BBB	0.56	0.17	0.90	0.96	0.73	0.54	0.17	0.90	0.96	0.75
MCD	0.50	0.15	0.92	**0.97**	0.78	0.49	0.15	**0.92**	**0.97**	0.78
VD	0.62	0.17	0.89	0.96	0.71	0.64	0.17	0.89	0.96	0.71
VSD	**0.45**	**0.14**	**0.93**	**0.97**	**0.81**	**0.47**	**0.14**	**0.92**	**0.97**	**0.79**

AlexNet (CIFAR10)	CIFAR100					SVHN				
	FPR	Det. err.	AUROC	AUPR IN	AUPR OUT	FPR	Det. err.	AUROC	AUPR IN	AUPR OUT
MAP	0.88	0.35	0.70	0.73	0.65	**0.89**	0.33	0.71	0.59	**0.83**
BBB	0.93	0.46	0.55	0.54	0.54	0.99	0.45	0.53	0.33	0.70
MCD	0.91	0.41	0.63	0.63	0.60	0.97	0.39	0.59	0.47	0.74
VD	0.87	0.35	0.69	0.72	0.64	0.89	0.32	0.72	0.60	**0.83**
VSD	**0.85**	**0.33**	**0.72**	**0.76**	**0.68**	0.91	**0.30**	**0.73**	**0.65**	**0.83**

ResNet-18 (CIFAR100)	CIFAR10					SVHN				
	FPR	Det. err.	AUROC	AUPR IN	AUPR OUT	FPR	Det. err.	AUROC	AUPR IN	AUPR OUT
MAP	0.89	**0.37**	0.67	0.70	0.63	0.91	0.36	0.68	0.56	0.81
BBB	0.93	0.41	0.62	0.66	0.58	0.89	0.37	0.68	0.51	0.82
MCD	0.89	**0.37**	0.68	0.71	0.63	0.89	0.34	0.71	0.58	0.83
VD	0.90	0.38	0.66	0.70	0.62	0.87	0.34	0.70	0.58	0.83
VSD	**0.87**	**0.37**	**0.69**	**0.72**	**0.65**	**0.83**	**0.31**	**0.76**	**0.65**	**0.86**

Variational Structured Dropout (VSD-our method)

❖ **Contribution 5:** VSD exhibits significant computational efficiency

Table 6: Computational complexity per layer of MAP and different variational methods.

Method	Time	Memory
MAP	$\mathcal{O}(KL\|\mathcal{B}\|)$	$\mathcal{O}(L\|\mathcal{B}\|)$
BBB	$\mathcal{O}(sKL\|\mathcal{B}\|)$	$\mathcal{O}(sKL + L\|\mathcal{B}\|)$
BBB-LTR	$\mathcal{O}(2KL\|\mathcal{B}\|)$	$\mathcal{O}(2L\|\mathcal{B}\|)$
VMG	$\mathcal{O}(m^3 + 2KL\|\mathcal{B}\|)$	$\mathcal{O}(KL\|\mathcal{B}\|)$
SLANG	$\mathcal{O}(r^2KL + rsKL\|\mathcal{B}\|)$	$\mathcal{O}(rKL + sKL\|\mathcal{B}\|)$
ELRG	$\mathcal{O}(r^3 + (r+2)KL\|\mathcal{B}\|)$	$\mathcal{O}((r+2)L\|\mathcal{B}\|)$
VSD	$\mathcal{O}(K^2 + KL\|\mathcal{B}\|)$	$\mathcal{O}(K^2 + K\|\mathcal{B}\|)$
VSD-low rank	$\mathcal{O}(rK + KL\|\mathcal{B}\|)$	$\mathcal{O}(K^2 + K\|\mathcal{B}\|)$

Table 7: Computation time of variational methods compared to standard MAP (1x).

Methods	**Time/epoch** (s)		
	LeNet5	AlexNet	ResNet18
BBB-LTR	1.53x	1.75x	3.28x
MNF	2.86x	3.40x	4.88x
VD	1.18x	1.15x	1.32x
VSD $T = 1$	1.25x	1.32x	1.86x
VSD $T = 2$	1.35x	1.49x	2.90x
time-scaling	1.08	1.13	1.56

Conclusion

- Introducing a novel Dropout Variational Inference framework for BNNs

Email: v.sonnv27@vinai.io