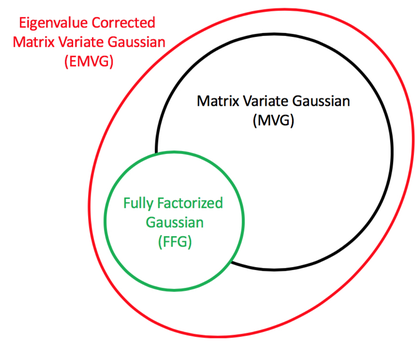
Eigenvalue Corrected Noisy Natural Gradient

Juhan Bae, Guodong Zhang, Roger Grosse
University of Toronto, Vector Institute

Summary

- We extend on a matrix-variate Gaussian posterior to compute a full diagonal variance.
- This leads to a more expressive posterior distribution in variational Bayesian neural networks.
- Better performance on BNN benchmarks.
- Scales up to large convolutional neural networks.



Natural Gradient

- Natural gradient descent is a second-order optimization technique which preconditions the gradient with the Fisher information matrix: $\tilde{\nabla}_{\mathbf{w}} h = \mathbf{F}^{-1} \nabla_{\mathbf{w}} h$, where the Fisher matrix $\mathbf{F}$ is given by:

$$\mathbf{F} = \mathbb{E}\left[\nabla_{\mathbf{w}} \log p(y|\mathbf{x}, \mathbf{w}) \nabla_{\mathbf{w}} \log p(y|\mathbf{x}, \mathbf{w})^\top\right]$$

- **Kronecker-Factored Approximate Curvature (K-FAC)**
  - K-FAC is a scalable approximation to natural gradient for neural networks.
  - It approximates the Fisher information matrix by assuming Kronecker structure:

$$\begin{aligned}\mathbf{F}_{\text{K-FAC}} &= \mathbb{E}[\text{vec}\{\nabla_{\mathbf{W}} h\}\text{vec}\{\nabla_{\mathbf{W}} h\}^\top] \\ &\approx \mathbb{E}[\nabla_{\mathbf{s}} h (\nabla_{\mathbf{s}} h)^\top] \otimes \mathbb{E}[\mathbf{a}\mathbf{a}^\top] \\ &= \mathbf{S} \otimes \mathbf{A}\end{aligned}$$

  - Using eigendecomposition, we can rewrite Kronecker-factored approximate Fisher matrix as:

$$\mathbf{S} \otimes \mathbf{A} = (\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})(\mathbf{\Lambda}_\mathbf{S} \otimes \mathbf{\Lambda}_\mathbf{A})(\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})^\top$$

- **Eigenvalue Corrected Kronecker-Factored Approximate Curvature (EK-FAC)**
  - EK-FAC is an extension of K-FAC which captures a more accurate diagonal re-scaling factor in K-FAC eigenbasis.
  - It computes the second moment of the gradient vector in K-FAC eigenbasis as follows:

$$\mathbf{R}_{ii} = \mathbb{E}[((\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})^\top \nabla_{\mathbf{w}})_i^2]$$

  - EK-FAC re-scaling factor minimizes the approximation error in K-FAC eigenbasis:

$$\mathbf{F}_{\text{EK-FAC}} \approx (\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})\mathbf{R}(\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})^\top$$

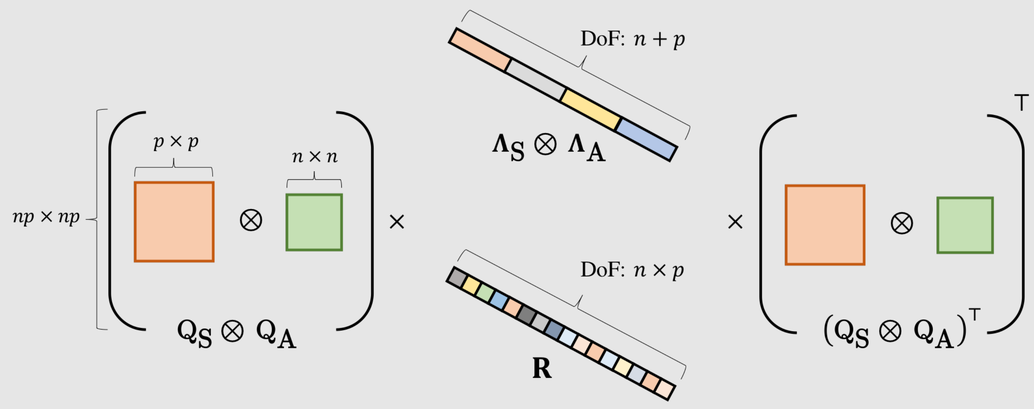


Figure: The diagonal re-scaling factor in K-FAC has Kronecker structure with $n + p$ degrees of freedom. The diagonal re-scaling matrix in EK-FAC is the second moment of the gradient vector with $n \times p$ degrees of freedom.

Variational Bayesian Neural Networks

- Variational Bayesian neural networks attempt to fit an approximate posterior $q(\mathbf{w})$ to maximize the evidence lower bound (ELBO):

$$\mathcal{F} = \underbrace{\mathbb{E}[\log p(\mathcal{D} \mid \mathbf{w})]}_{\text{data fitting}} - \underbrace{\lambda \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}) \,\|\, p(\mathbf{w}))}_{\text{regularization}}$$

- **Noisy Natural Gradient (NNG)**
  - It is an efficient method to fit a multivariate Gaussian posterior by adding adaptive weight noise to ordinary natural gradient updates.
  - Assuming $q_\phi(\mathbf{w})$ is a multivariate Gaussian posterior parameterized by $\phi = (\boldsymbol{\mu}, \mathbf{\Sigma})$ and $p(\mathbf{w})$ is a spherical Gaussian, the update rules of NNG are:

$$\begin{aligned}\boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} + \tilde{\alpha}\left(\bar{\mathbf{F}} + \frac{\lambda}{N\eta}\mathbf{I}\right)^{-1}\left[\nabla_{\mathbf{w}} \log p(y \mid \mathbf{w}, \mathbf{x}) - \frac{\lambda}{N\eta}\mathbf{w}\right] \\ \bar{\mathbf{F}} &\leftarrow (1 - \tilde{\beta})\bar{\mathbf{F}} + \tilde{\beta}\nabla_{\mathbf{w}} \log p(y \mid \mathbf{w}, \mathbf{x}) \nabla_{\mathbf{w}} \log p(y \mid \mathbf{w}, \mathbf{x})^\top\end{aligned}$$

  - This resembles ordinary natural gradient updates for point estimation except that $\mathbf{w}$ is sampled from $q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \mathbf{\Lambda}^{-1} = \frac{\lambda}{N}\left(\mathbf{F} + \frac{\lambda}{N\eta}\mathbf{I}\right)^{-1}$$

Eigenvalue Corrected Noisy Natural Gradient

To capture an accurate diagonal variance in a matrix-variate Gaussian, we develop a new tractable instance of noisy natural gradient. EK-FAC approximation of the Fisher matrix yields an EMVG posterior in NNG.

$$\begin{aligned}\mathbf{\Sigma} &= \frac{\lambda}{N}(\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})(\mathbf{R}^\gamma)^{-1}(\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})^\top \\ &= \frac{\lambda}{N}(\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})\left(\mathbf{R} + \frac{\lambda}{N\eta}\mathbf{I}\right)^{-1}(\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})^\top\end{aligned}$$

The EMVG posterior is potentially powerful because:

- compactly represents covariances between weights.
- **efficiently computes a full diagonal variance in K-FAC eigenbasis.**

The inference is efficient because the covariance matrix is factorized with three small matrices $\mathbf{A}, \mathbf{S}$, and $\mathbf{R}$. The update rules for these factors are:

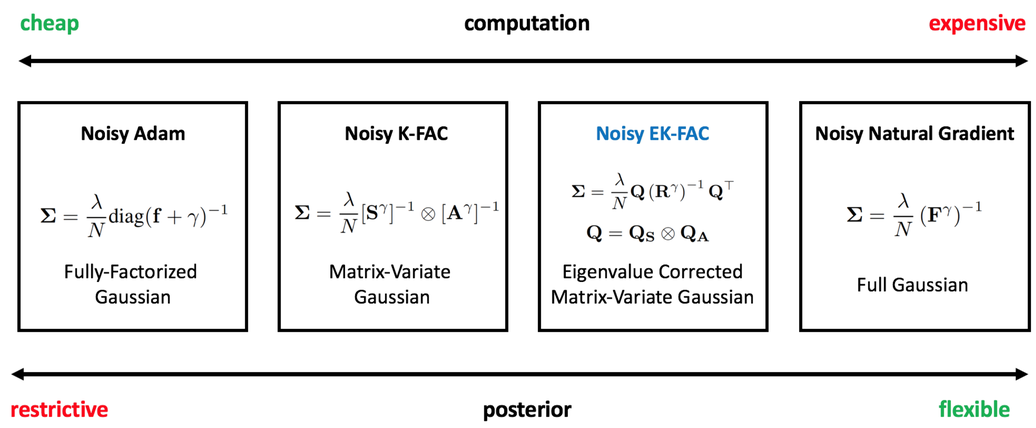
$$\begin{aligned}\mathbf{A} &\leftarrow (1 - \tilde{\beta})\mathbf{A} + \tilde{\beta}\mathbf{a}\mathbf{a}^\top \\ \mathbf{S} &\leftarrow (1 - \tilde{\beta})\mathbf{S} + \tilde{\beta}\nabla_{\mathbf{s}} \log p(y|\mathbf{x}, \mathbf{w}) \nabla_{\mathbf{s}} \log p(y|\mathbf{x}, \mathbf{w})^\top \\ \mathbf{R}_{ii} &\leftarrow (1 - \tilde{\omega})\mathbf{R}_{ii} + \tilde{\omega}\left[(\mathbf{Q}_\mathbf{S} \otimes \mathbf{Q}_\mathbf{A})^\top \nabla_{\mathbf{w}} \log p(y|\mathbf{x}, \mathbf{w})\right]_i^2\end{aligned}$$

Eigenvalue Corrected Matrix-Variate Gaussian

$$\text{vec}(\mathbf{W}) \sim \mathcal{N}(\text{vec}(\mathbf{M}), (\mathbf{Q}_\mathbf{V} \otimes \mathbf{Q}_\mathbf{U})\,\mathbf{R}\,(\mathbf{Q}_\mathbf{V} \otimes \mathbf{Q}_\mathbf{U})^\top)$$

where $\mathbf{M} \in \mathbb{R}^{n \times p}$ is the mean, $\mathbf{U} \in \mathbb{R}^{n \times n}$ is the covariance matrix among rows, $\mathbf{V} \in \mathbb{R}^{p \times p}$ is the covariance matrix among columns, and $\mathbf{R} \in \mathbb{R}^{np \times np}$ is the re-scaling matrix. $\mathbf{Q}_\mathbf{V}$ and $\mathbf{Q}_\mathbf{U}$ are eigenbasis of $\mathbf{V}$ and $\mathbf{U}$.

Family of Noisy Natural Gradient



Experiments

Performance on UCI Benchmarks:

Dataset	Test log-likelihood			
	BBB	Noisy Adam	Noisy K-FAC	Noisy EK-FAC
Boston	-2.602±0.031	-2.558 ±0.032	-2.409±0.047	**-2.378±0.044**
Concrete	-3.149±0.018	-3.145±0.023	-3.039±0.025	**-3.002±0.025**
Energy	-1.500±0.006	-1.629±0.020	**-1.421±0.004**	-1.448±0.004
Kin8nm	1.111±0.007	1.112±0.008	1.148±0.007	**1.149±0.012**
Naval	6.143±0.032	6.231±0.041	7.079±0.034	**7.287±0.002**
Pow. Plant	-2.807±0.010	-2.803±0.010	-2.776±0.012	**-2.774±0.012**
Protein	-2.882±0.004	-2.896±0.004	-2.836±0.002	**-2.819±0.007**
Wine	-0.977±0.017	-0.976±0.016	-0.969±0.014	**-0.964±0.002**
Yacht	-2.408±0.007	-2.412±0.006	-2.316±0.006	**-2.224±0.007**
Year	-3.614±NA	-3.620±NA	-3.595±NA	**-3.573±NA**

Generalization: VGG16 on CIFAR10.

Method	Test Accuracy			
		D	B	D + B
SGD	81.79	88.35	85.75	91.39
KFAC	82.39	88.89	86.86	92.13
Noisy-KFAC	85.52	89.35	88.22	92.01
Noisy-EKFAC	**87.07**	**89.86**	**88.45**	**92.22**

Convergence: curves of evidence lower bound.