Bayesian Online Natural Gradient (BONG)

Matt Jones[1], Peter Chang[2], Kevin Murphy[3]
(mcj@colorado.edu, gyuyoung@mit.edu, kpmurphy@google.com)

[1]U. Colorado, [2]MIT, [3]Google DeepMind

October 31, 2024

Matt Jones, Peter Chang, Kevin Murphy. Bayesian online natural gradient (BONG). NeurIPS24, arXiv:2405.19681, 2024.

Online Learning

- Data received sequentially
  - Input $\boldsymbol{x}_t$
  - Predict observation $\boldsymbol{y}_t$
- Often nonstationary
  - Covariate shift $p_t(\boldsymbol{x}_t)$
  - Concept shift $p_t(\boldsymbol{y}_t|\boldsymbol{x}_t)$
- Predictive model (e.g., NN)
  - Parameters $\boldsymbol{\theta}_t$ (e.g., weights)
  - $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t) = p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$ with $f_t(\boldsymbol{\theta}_t) = f(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$
- Prequential evaluation
  - Maximize log-likelihood of each upcoming observation
  - $\log p\left(\boldsymbol{y}_t|f_t\left(\hat{\boldsymbol{\theta}}_t\right)\right)$

Online Learning

- Data received sequentially
  - Input $\boldsymbol{x}_t$
  - Predict observation $\boldsymbol{y}_t$
- Often nonstationary
  - Covariate shift $p_t(\boldsymbol{x}_t)$
  - Concept shift $p_t(\boldsymbol{y}_t|\boldsymbol{x}_t)$
- Predictive model (e.g., NN)
  - Parameters $\boldsymbol{\theta}_t$ (e.g., weights)
  - $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t) = p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$ with $f_t(\boldsymbol{\theta}_t) = f(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$
- Prequential evaluation
  - Maximize log-likelihood of each upcoming observation
  - $\log p\left(\boldsymbol{y}_t|f_t\left(\hat{\boldsymbol{\theta}}_t\right)\right)$

Online Learning

- Data received sequentially
  - Input $\boldsymbol{x}_t$
  - Predict observation $\boldsymbol{y}_t$
- Often nonstationary
  - Covariate shift $p_t(\boldsymbol{x}_t)$
  - Concept shift $p_t(\boldsymbol{y}_t|\boldsymbol{x}_t)$
- Predictive model (e.g., NN)
  - Parameters $\boldsymbol{\theta}_t$ (e.g., weights)
  - $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t) = p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$ with $f_t(\boldsymbol{\theta}_t) = f(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$
- Prequential evaluation
  - Maximize log-likelihood of each upcoming observation
  - $\log p\left(y_t|f_t\left(\hat{\theta}_t\right)\right)$

Online Learning

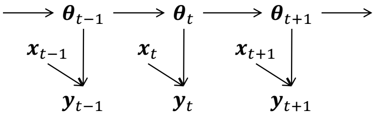
- Data received sequentially
  - Input $\boldsymbol{x}_t$
  - Predict observation $\boldsymbol{y}_t$
- Often nonstationary
  - Covariate shift $p_t(\boldsymbol{x}_t)$
  - Concept shift $p_t(\boldsymbol{y}_t|\boldsymbol{x}_t)$
- Predictive model (e.g., NN)
  - Parameters $\boldsymbol{\theta}_t$ (e.g., weights)
  - $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t) = p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$ with $f_t(\boldsymbol{\theta}_t) = f(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$
- Prequential evaluation
  - Maximize log-likelihood of each upcoming observation
  - $\log p\left(\boldsymbol{y}_t|f_t\left(\hat{\boldsymbol{\theta}}_t\right)\right)$

State Space Models (SSMs)

- Stochastic dynamics for latent parameter
  - $p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right)$
- Observation model
  - $p\left(\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right)$
- Bayesian filtering
  - Maintain posterior $p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t}\right)$
  - Predict step: $p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t-1}\right) = \int p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right) p\left(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{1:t-1}\right) \mathrm{d}\boldsymbol{\theta}_{t-1}$
  - Update step: $p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t}\right) \propto p\left(\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right) p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t-1}\right)$

State Space Models (SSMs)

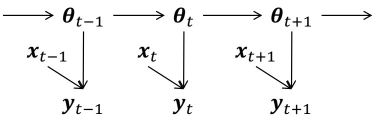
- Stochastic dynamics for latent parameter
  - $p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right)$
- Observation model
  - $p\left(\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right)$



- Bayesian filtering
  - Maintain posterior $p\left(\theta_t|\mathcal{D}_{1:t}\right)$
  - Predict step: $p\left(\theta_t|\mathcal{D}_{1:t-1}\right) = \int p\left(\theta_t|\theta_{t-1}\right) p\left(\theta_{t-1}|\mathcal{D}_{1:t-1}\right) \mathrm{d}\theta_{t-1}$
  - Update step: $p\left(\theta_t|\mathcal{D}_{1:t}\right) \propto p\left(y_t|f_t\left(\theta_t\right)\right) p\left(\theta_t|\mathcal{D}_{1:t-1}\right)$

State Space Models (SSMs)

- Stochastic dynamics for latent parameter
  - $p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$
- Observation model
  - $p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))$
- Bayesian filtering
  - Maintain posterior $p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t})$
  - Predict step: $p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t-1}) = \int p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) \, p(\boldsymbol{\theta}_{t-1} | \mathcal{D}_{1:t-1}) \, \mathrm{d}\boldsymbol{\theta}_{t-1}$
  - Update step: $p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t}) \propto p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \, p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t-1})$



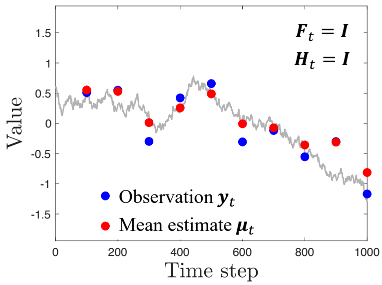
Kalman Filter

- Linear-Gaussian dynamics
  - $\boldsymbol{\theta}_t \sim \mathcal{N}\left(\boldsymbol{F}_t \boldsymbol{\theta}_{t-1}, \boldsymbol{Q}_t\right)$
- Linear-Gaussian observations
  - $\boldsymbol{y}_t \sim \mathcal{N}\left(\boldsymbol{H}_t \boldsymbol{\theta}_t, \boldsymbol{R}_t\right)$
- Predict step (exact)

$$p\left(\boldsymbol{\theta}_t | \mathcal{D}_{1:t-1}\right) = \mathcal{N}\left(\boldsymbol{\theta}_t | \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right)$$
$$\boldsymbol{\mu}_{t|t-1} = \boldsymbol{F}_t \boldsymbol{\mu}_{t-1}$$
$$\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{F}_t \boldsymbol{\Sigma}_{t-1} \boldsymbol{F}_t^{\intercal} + \boldsymbol{Q}_t$$

- Update step (exact, conjugate)

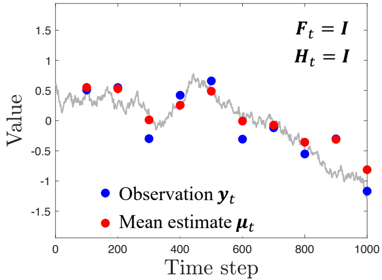
$$p\left(\boldsymbol{\theta}_t | \mathcal{D}_{1:t}\right) = \mathcal{N}\left(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right)$$
$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t \boldsymbol{H}_t^{\intercal} \boldsymbol{R}_t^{-1}\left(\boldsymbol{y}_t - \boldsymbol{H}_t \boldsymbol{\mu}_{t|t-1}\right)$$
$$\boldsymbol{\Sigma}_t = \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \boldsymbol{H}_t^{\intercal} \boldsymbol{R}_t^{-1} \boldsymbol{H}_t\right)^{-1}$$



Kalman Filter

- Linear-Gaussian dynamics
  - $\boldsymbol{\theta}_t \sim \mathcal{N}(\boldsymbol{F}_t \boldsymbol{\theta}_{t-1}, \boldsymbol{Q}_t)$
- Linear-Gaussian observations
  - $\boldsymbol{y}_t \sim \mathcal{N}(\boldsymbol{H}_t \boldsymbol{\theta}_t, \boldsymbol{R}_t)$
- Predict step (exact)

$$p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t-1}) = \mathcal{N}\left(\boldsymbol{\theta}_t | \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right)$$
$$\boldsymbol{\mu}_{t|t-1} = \boldsymbol{F}_t \boldsymbol{\mu}_{t-1}$$
$$\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{F}_t \boldsymbol{\Sigma}_{t-1} \boldsymbol{F}_t^{\mathsf{T}} + \boldsymbol{Q}_t$$



- Update step (exact, conjugate)

$$p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t}) = \mathcal{N}(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$
$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t \boldsymbol{H}_t^{\mathsf{T}} \boldsymbol{R}_t^{-1} \left(\boldsymbol{y}_t - \boldsymbol{H}_t \boldsymbol{\mu}_{t|t-1}\right)$$
$$\boldsymbol{\Sigma}_t = \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \boldsymbol{H}_t^{\mathsf{T}} \boldsymbol{R}_t^{-1} \boldsymbol{H}_t\right)^{-1}$$

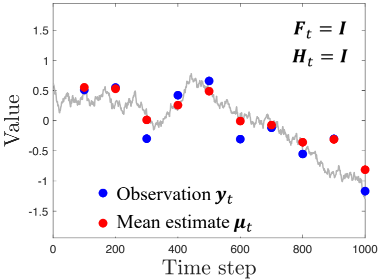
Kalman Filter

- Linear-Gaussian dynamics
  - $\boldsymbol{\theta}_t \sim \mathcal{N}(\boldsymbol{F}_t \boldsymbol{\theta}_{t-1}, \boldsymbol{Q}_t)$
- Linear-Gaussian observations
  - $\boldsymbol{y}_t \sim \mathcal{N}(\boldsymbol{H}_t \boldsymbol{\theta}_t, \boldsymbol{R}_t)$
- Predict step (exact)

$$p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t | \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})$$

$$\boldsymbol{\mu}_{t|t-1} = \boldsymbol{F}_t \boldsymbol{\mu}_{t-1}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{F}_t \boldsymbol{\Sigma}_{t-1} \boldsymbol{F}_t^{\intercal} + \boldsymbol{Q}_t$$



- Update step (exact, conjugate)

$$p(\boldsymbol{\theta}_t | \mathcal{D}_{1:t}) = \mathcal{N}(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t \boldsymbol{H}_t^{\intercal} \boldsymbol{R}_t^{-1} \left(\boldsymbol{y}_t - \boldsymbol{H}_t \boldsymbol{\mu}_{t|t-1}\right)$$

$$\boldsymbol{\Sigma}_t = \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \boldsymbol{H}_t^{\intercal} \boldsymbol{R}_t^{-1} \boldsymbol{H}_t\right)^{-1}$$

Variational Bayesian Inference (VI)

- Variational family

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx q_{\psi}(\boldsymbol{\theta})$$

- Minimize KL divergence from true posterior

$$\psi = \arg\min_{\psi} D_{\mathbb{KL}}\left(q_{\psi}(\theta) \,|\, \frac{1}{Z} p_0(\theta)\, p(\mathcal{D}|\theta)\right)$$
$$= \arg\min_{\psi} \mathcal{L}(\psi) + \text{const}$$

- nELBO loss

$$\mathcal{L}(\psi) = \underbrace{\mathbb{E}_{\theta \sim q_{\psi}}\left[-\log p(\mathcal{D}|\theta)\right]}_{\text{fit: } \mathbb{E}[\text{NLL}]} + \underbrace{D_{\mathbb{KL}}(q_{\psi}|p_0)}_{\text{regularization to prior}}$$

Variational Bayesian Inference (VI)

- Variational family

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx q_{\psi}(\boldsymbol{\theta})$$

- Minimize KL divergence from true posterior

$$\begin{aligned}\psi &= \arg\min_{\psi} D_{\mathbb{KL}}\left(q_{\psi}(\boldsymbol{\theta}) \,|\, \frac{1}{Z} p_0(\boldsymbol{\theta}) \, p(\mathcal{D}|\boldsymbol{\theta})\right) \\ &= \arg\min_{\psi} \mathcal{L}(\psi) + \text{const}\end{aligned}$$

- nELBO loss

$$\mathcal{L}(\psi) = \underbrace{\mathbb{E}_{\boldsymbol{\theta} \sim q_{\psi}}\left[-\log p(\mathcal{D}|\boldsymbol{\theta})\right]}_{\text{fit: } \mathbb{E}[\text{NLL}]} + \underbrace{D_{\mathbb{KL}}(q_{\psi}|p_0)}_{\text{regularization to prior}}$$

Variational Bayesian Inference (VI)

- Variational family

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx q_{\boldsymbol{\psi}}(\boldsymbol{\theta})$$

- Minimize KL divergence from true posterior

$$\begin{aligned}\boldsymbol{\psi} &= \arg\min_{\boldsymbol{\psi}} D_{\mathbb{KL}}\left(q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) \,|\, \frac{1}{Z} p_0(\boldsymbol{\theta})\, p(\mathcal{D}|\boldsymbol{\theta})\right) \\ &= \arg\min_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\psi}) + \text{const}\end{aligned}$$

- nELBO loss

$$\mathcal{L}(\boldsymbol{\psi}) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim q_{\boldsymbol{\psi}}}\left[-\log p(\mathcal{D}|\boldsymbol{\theta})\right]}_{\text{fit: } \mathbb{E}[\text{NLL}]} + \underbrace{D_{\mathbb{KL}}(q_{\boldsymbol{\psi}}|p_0)}_{\text{regularization to prior}}$$

Online VI

- Approximate prior from previous timestep

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t-1}\right) \approx q_{\boldsymbol{\psi}_{t|t-1}}\left(\boldsymbol{\theta}_t\right) = \int q_{\boldsymbol{\psi}_{t-1}}\left(\boldsymbol{\theta}_{t-1}\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right) \mathrm{d}\boldsymbol{\theta}_{t-1}$$

- Variational filtering

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t}\right) \approx \frac{1}{Z} q_{\boldsymbol{\psi}_{t|t-1}}\left(\boldsymbol{\theta}_t\right) p\left(\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right) \approx q_{\boldsymbol{\psi}_t}\left(\boldsymbol{\theta}_t\right)$$

- Online VI loss

$$\boldsymbol{\psi}_t = \arg\min_{\boldsymbol{\psi}} \mathcal{L}_t\left(\boldsymbol{\psi}\right)$$

$$\mathcal{L}_t\left(\boldsymbol{\psi}\right) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\boldsymbol{\psi}}}\left[-\log p\left(\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right)\right]}_{\text{incremental fit, } L_t(\boldsymbol{\psi})} + \underbrace{D_{\mathbb{KL}}\left(q_{\boldsymbol{\psi}}|q_{\boldsymbol{\psi}_{t|t-1}}\right)}_{\text{recursive regularizer}}$$

Online VI

- Approximate prior from previous timestep

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t-1}\right) \approx q_{\psi_{t|t-1}}\left(\boldsymbol{\theta}_t\right) = \int q_{\psi_{t-1}}\left(\boldsymbol{\theta}_{t-1}\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right) \mathrm{d}\boldsymbol{\theta}_{t-1}$$

- Variational filtering

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_{1:t}\right) \approx \frac{1}{Z} q_{\psi_{t|t-1}}\left(\boldsymbol{\theta}_t\right) p\left(\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right) \approx q_{\psi_t}\left(\boldsymbol{\theta}_t\right)$$

- Online VI loss

$$\psi_t = \arg\min_{\psi} \mathcal{L}_t\left(\psi\right)$$

$$\mathcal{L}_t\left(\psi\right) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}_t \sim q_\psi}\left[-\log p\left(\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right)\right]}_{\text{incremental fit, } L_t(\psi)} + \underbrace{D_{\mathbb{KL}}\left(q_\psi|q_{\psi_{t|t-1}}\right)}_{\text{recursive regularizer}}$$

Online VI

- Approximate prior from previous timestep

$$p\left(\boldsymbol{\theta}_t | \mathcal{D}_{1:t-1}\right) \approx q_{\boldsymbol{\psi}_{t|t-1}}\left(\boldsymbol{\theta}_t\right) = \int q_{\boldsymbol{\psi}_{t-1}}\left(\boldsymbol{\theta}_{t-1}\right) p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}\right) \mathrm{d}\boldsymbol{\theta}_{t-1}$$

- Variational filtering

$$p\left(\boldsymbol{\theta}_t | \mathcal{D}_{1:t}\right) \approx \frac{1}{Z} q_{\boldsymbol{\psi}_{t|t-1}}\left(\boldsymbol{\theta}_t\right) p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right) \approx q_{\boldsymbol{\psi}_t}\left(\boldsymbol{\theta}_t\right)$$

- Online VI loss

$$\boldsymbol{\psi}_t = \arg\min_{\boldsymbol{\psi}} \mathcal{L}_t\left(\boldsymbol{\psi}\right)$$

$$\mathcal{L}_t\left(\boldsymbol{\psi}\right) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\boldsymbol{\psi}}}\left[-\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]}_{\text{incremental fit, } L_t(\boldsymbol{\psi})} + \underbrace{D_{\mathbb{KL}}\left(q_{\boldsymbol{\psi}} | q_{\boldsymbol{\psi}_{t|t-1}}\right)}_{\text{recursive regularizer}}$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p(\boldsymbol{y}_t | f_t(\theta_t))\right]$$

- Fixed point, implicit update
  - Exponential family: natural params ψ, dual (expectation) params ρ

$$\nabla_{\psi_t}\mathcal{L}_t(\psi_t) = 0 \quad \Longrightarrow \quad \psi_t = \psi_{t|t-1} + \nabla_{\rho_t}\mathbb{E}_{q_{\psi_t}}\left[\log p(\boldsymbol{y}_t | f_t(\theta_t))\right]$$

  - Gaussian: Recursive variational Gaussian approximation (RVGA) (Lambert et al. 2021)

$$q_{\psi_t}(\theta_t) = \mathcal{N}(\theta_t | \mu_t, \Sigma_t)$$
$$\mu_t = \mu_{t|t-1} + \Sigma_{t|t-1}\mathbb{E}_{q_{\psi_t}}\left[\nabla_{\theta_t}\log p(\boldsymbol{y}_t | f_t(\theta_t))\right]$$
$$\Sigma_t^{-1} = \Sigma_{t|t-1}^{-1} - \mathbb{E}_{q_{\psi_t}}\left[\nabla^2_{\theta_t}\log p(\boldsymbol{y}_t | f_t(\theta_t))\right]$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

- Fixed point, implicit update
  - Exponential family: natural params ψ, dual (expectation) params ρ

$$\nabla_{\psi_t}\mathcal{L}_t(\psi_t) = 0 \quad \Longrightarrow \quad \psi_t = \psi_{t|t-1} + \nabla_{\rho_t}\mathbb{E}_{q_{\psi_t}}\left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

  - Gaussian: Recursive variational Gaussian approximation (RVGA) (Lambert et al. 2021)

$$q_{\psi_t}(\boldsymbol{\theta}_t) = \mathcal{N}\left(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right)$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}\mathbb{E}_{q_{\psi_t}}\left[\nabla_{\boldsymbol{\theta}_t}\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} - \mathbb{E}_{q_{\psi_t}}\left[\nabla_{\boldsymbol{\theta}_t}^2\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p\left(\mathbf{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Fixed point, implicit update
  - Exponential family: natural params ψ, dual (expectation) params $\boldsymbol{\rho}$

$$\nabla_{\psi_t}\mathcal{L}_t(\psi_t) = 0 \quad \Longrightarrow \quad \psi_t = \psi_{t|t-1} + \nabla_{\boldsymbol{\rho}_t}\mathbb{E}_{q_{\psi_t}}\left[\log p\left(\mathbf{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

  - Gaussian: Recursive variational Gaussian approximation (RVGA) (Lambert et al. 2021)

$$\begin{aligned} q_{\psi_t}(\boldsymbol{\theta}_t) &= \mathcal{N}(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \\ \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}\mathbb{E}_{q_{\psi_t}}\left[\nabla_{\boldsymbol{\theta}_t} \log p\left(\mathbf{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right] \\ \boldsymbol{\Sigma}_t^{-1} &= \boldsymbol{\Sigma}_{t|t-1}^{-1} - \mathbb{E}_{q_{\psi_t}}\left[\nabla^2_{\boldsymbol{\theta}_t} \log p\left(\mathbf{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right] \end{aligned}$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Fixed point, implicit update (Lambert et al. 2021)

$$\boldsymbol{\psi}_t = \boldsymbol{\psi}_{t|t-1} + \nabla_{\boldsymbol{\rho}_t}\mathbb{E}_{q_{\psi_t}}\left[\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Iterative
  - Bayes by backprop (BBB) (Blundell et al. 2015): GD on nELBO

$$\boldsymbol{\psi}_i = \boldsymbol{\psi}_{i-1} - \alpha \nabla_{\boldsymbol{\psi}_{i-1}} \mathcal{L}(\boldsymbol{\psi}_{i-1})$$

  - Bayesian learning rule (BLR) (Khan and Rue 2023): NGD on nELBO

$$\begin{aligned}\psi_i &= \psi_{i-1} - \alpha \boldsymbol{F}_{\psi_{i-1}}^{-1} \nabla_{\psi_{i-1}} \mathcal{L}(\psi_{i-1}) \\ &= \psi_{i-1} - \alpha \nabla_{\rho_{i-1}} \mathcal{L}(\psi_{i-1})\end{aligned}$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_{\psi} | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_{\psi}}\left[-\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Fixed point, implicit update (Lambert et al. 2021)

$$\boldsymbol{\psi}_t = \boldsymbol{\psi}_{t|t-1} + \nabla_{\boldsymbol{\rho}_t} \mathbb{E}_{q_{\boldsymbol{\psi}_t}}\left[\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Iterative
    - Bayes by backprop (BBB) (Blundell et al. 2015): GD on nELBO

$$\boldsymbol{\psi}_i = \boldsymbol{\psi}_{i-1} - \alpha \nabla_{\boldsymbol{\psi}_{i-1}} \mathcal{L}(\boldsymbol{\psi}_{i-1})$$

    - Bayesian learning rule (BLR) (Khan and Rue 2023): NGD on nELBO

$$\begin{aligned}\boldsymbol{\psi}_i &= \boldsymbol{\psi}_{i-1} - \alpha \boldsymbol{F}_{\boldsymbol{\psi}_{i-1}}^{-1} \nabla_{\boldsymbol{\psi}_{i-1}} \mathcal{L}(\boldsymbol{\psi}_{i-1}) \\ &= \boldsymbol{\psi}_{i-1} - \alpha \nabla_{\boldsymbol{\rho}_{i-1}} \mathcal{L}(\boldsymbol{\psi}_{i-1})\end{aligned}$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

- Fixed point, implicit update (Lambert et al. 2021)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_t} \mathbb{E}_{q_{\psi_t}}\left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

- Iterative
    - Bayes by backprop (BBB) (Blundell et al. 2015): GD on nELBO
    - Bayesian learning rule (BLR) (Khan and Rue 2023): NGD on nELBO
- Approximate likelihood
    - Linearized model: extended Kalman filter (EKF) (Singhal and Wu 1989; Puskorius and Feldkamp 1991)

$$\mathcal{N}\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right), \boldsymbol{R}_t\right) \approx \mathcal{N}\left(\boldsymbol{y}_t | \bar{f}_t\left(\boldsymbol{\theta}_t\right), \boldsymbol{R}_t\right)$$

$$\bar{f}_t\left(\boldsymbol{\theta}_t\right) = f_t\left(\boldsymbol{\mu}_{t|t-1}\right) + F_t\left(\boldsymbol{\theta}_t - \boldsymbol{\mu}_{t|t-1}\right)$$

$$F_t = \mathrm{jac}\left(f_t(\cdot)\right)\left(\boldsymbol{\mu}_{t|t-1}\right)$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Fixed point, implicit update (Lambert et al. 2021)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_t} \mathbb{E}_{q_{\psi_t}}\left[\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Iterative
  - Bayes by backprop (BBB) (Blundell et al. 2015): GD on nELBO
  - Bayesian learning rule (BLR) (Khan and Rue 2023): NGD on nELBO
- Approximate likelihood
  - Linearized model: extended Kalman filter (EKF) (Singhal and Wu 1989; Puskorius and Feldkamp 1991)
  - Linear-Gaussianized: Conditional moments EKF (CM-EKF) (Tronarp et al. 2018; Ollivier 2018)

$$p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right) \approx \mathcal{N}\left(\boldsymbol{y}_t | \bar{h}_t(\boldsymbol{\theta}_t), \hat{\boldsymbol{R}}_t\right)$$

$$\bar{h}_t(\boldsymbol{\theta}_t) = \mathbb{E}\left[\boldsymbol{y}_t | f_t\left(\boldsymbol{\mu}_{t|t-1}\right)\right] + \left.\frac{\partial \mathbb{E}\left[\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right]}{\partial \boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t = \boldsymbol{\mu}_{t|t-1}} \left(\boldsymbol{\theta}_t - \boldsymbol{\mu}_{t|t-1}\right)$$

$$\hat{\boldsymbol{R}}_t = \mathbb{V}\left[\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right]$$

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Fixed point, implicit update (Lambert et al. 2021)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_t} \mathbb{E}_{q_{\psi_t}}\left[\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Iterative
  - Bayes by backprop (BBB) (Blundell et al. 2015): GD on nELBO
  - Bayesian learning rule (BLR) (Khan and Rue 2023): NGD on nELBO
- Approximate likelihood
  - Linear-Gaussianized: EKF, CM-EKF (Singhal and Wu 1989; Puskorius and Feldkamp 1991; Ollivier 2018; Tronarp et al. 2018)
  - Plugin approximation
    - $q_{\psi_t}(\boldsymbol{\theta}_t) \to \delta_{\rho_t}(\boldsymbol{\theta}_t)$, gives implicit mirror decent

$$\psi_t = \psi_{t|t-1} + \nabla_{\boldsymbol{\theta}_t = \rho_t} \log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)$$

    - Gaussian family with fixed covariance Σ, yields online GD (Bencomo et al. 2023)

Approximate methods for VI

- Expected NLL intractable for NNs

$$\mathcal{L}_t(\psi) = L_t(\psi) + D_{\mathbb{KL}}\left(q_\psi | q_{\psi_{t|t-1}}\right), \quad L_t(\psi) = \mathbb{E}_{\theta_t \sim q_\psi}\left[-\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Fixed point, implicit update (Lambert et al. 2021)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_t} \mathbb{E}_{q_{\psi_t}}\left[\log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)\right]$$

- Iterative
  - Bayes by backprop (BBB) (Blundell et al. 2015): GD on nELBO
  - Bayesian learning rule (BLR) (Khan and Rue 2023): NGD on nELBO
- Approximate likelihood
  - Linear-Gaussianized: EKF, CM-EKF (Singhal and Wu 1989; Puskorius and Feldkamp 1991; Ollivier 2018; Tronarp et al. 2018)
  - Plugin approximation
    - $q_{\psi_t}(\boldsymbol{\theta}_t) \rightarrow \delta_{\rho_t}(\boldsymbol{\theta}_t)$, gives implicit mirror decent

$$\psi_t = \psi_{t|t-1} + \nabla_{\theta_t = \rho_t} \log p\left(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)\right)$$

    - Gaussian family with fixed covariance $\boldsymbol{\Sigma}$, yields online GD (Bencomo et al. 2023)

Bayesian online natural gradient (BONG)

- Online BLR (oBLR): regularize to iterative prior $q_{\psi_{t|t-1}}$ instead of p_0

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \underbrace{\left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p\left(y_t | f_t\left(\theta_t\right)\right)\right] - D_{\mathbb{KL}}\left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}}\right) \right)}_{\text{online VI loss } \mathcal{L}_t(\psi_{t,i-1})}$$

- BONG
  - Replace regularizer with implicit regularization from truncated update
  - Special case of oBLR with $I = 1$ iteration and $\alpha = 1$ (since $\nabla D_{\mathbb{KL}}\left(q_{\psi_{t|t-1}} | q_{\psi_{t|t-1}}\right) = 0$)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p\left(y_t | f_t\left(\theta_t\right)\right)\right]$$

Bayesian online natural gradient (BONG)

- Online BLR (oBLR): regularize to iterative prior $q_{\psi_{t|t-1}}$ instead of p_0

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \underbrace{\left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right] - D_{\mathbb{KL}}\left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}}\right) \right)}_{\text{online VI loss } \mathcal{L}_t(\psi_{t,i-1})}$$

- BONG
  - Replace regularizer with implicit regularization from truncated update
  - Special case of oBLR with $I = 1$ iteration and $\alpha = 1$ (since $\nabla D_{\mathbb{KL}}\left(q_{\psi_{t|t-1}} | q_{\psi_{t|t-1}}\right) = 0$)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

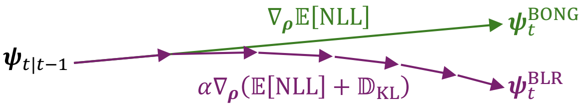
Bayesian online natural gradient (BONG)

- Online BLR (oBLR): regularize to iterative prior $q_{\psi_{t|t-1}}$ instead of p_0

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \Big(\underbrace{\mathbb{E}_{q_{\psi_{t,i-1}}} [\log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))] - D_{\mathbb{KL}} \left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}} \right)}_{\text{online VI loss } \mathcal{L}_t(\psi_{t,i-1})} \Big)$$

- BONG
  - Replace regularizer with implicit regularization from truncated update
  - Special case of oBLR with $I = 1$ iteration and $\alpha = 1$ (since $\nabla D_{\mathbb{KL}} \left(q_{\psi_{t|t-1}} | q_{\psi_{t|t-1}} \right) = 0$)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} [\log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))]$$



Bayesian online natural gradient (BONG)

- Online BLR (oBLR): regularize to iterative prior $q_{\psi_{t|t-1}}$ instead of p_0

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \underbrace{\left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p\left(\mathbf{y}_t | f_t\left(\boldsymbol{\theta}_t \right) \right) \right] - D_{\mathbb{KL}} \left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}} \right) \right)}_{\text{online VI loss } \mathcal{L}_t(\psi_{t,i-1})}$$

- BONG
  - Replace regularizer with implicit regularization from truncated update
  - Special case of oBLR with $I = 1$ iteration and $\alpha = 1$ (since $\nabla D_{\mathbb{KL}} \left(q_{\psi_{t|t-1}} | q_{\psi_{t|t-1}} \right) = 0$)

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p\left(\mathbf{y}_t | f_t\left(\boldsymbol{\theta}_t \right) \right) \right]$$

- Justifications
  - Explicit version of exact implicit update $\psi_t = \psi_{t|t-1} + \nabla_{\rho_t} \mathbb{E}_{q_{\psi_t}} \left[\log p\left(\mathbf{y}_t | f_t\left(\boldsymbol{\theta}_t \right) \right) \right]$
  - Bayes optimal when model is conjugate
  - Recovers several existing methods and new ones

Conjugate case

Theorem

Let the observation distribution be an exponential family with natural parameter $\boldsymbol{\theta}_t$ (with $\mathbf{y}_t$ encoded as sufficient statistics)

$$p_t(\mathbf{y}_t|\boldsymbol{\theta}_t) \propto \exp\left(\boldsymbol{\theta}_t^{\mathsf{T}}\mathbf{y}_t - A(\boldsymbol{\theta}_t) - b(\mathbf{y}_t)\right)$$

and let the prior be conjugate

$$q_{\boldsymbol{\psi}_{t|t-1}}(\boldsymbol{\theta}_t) = \exp\left(\boldsymbol{\psi}_{t|t-1}^{\mathsf{T}} T(\boldsymbol{\theta}_t) - \Phi(\boldsymbol{\psi}_{t|t-1})\right)$$

$$T(\boldsymbol{\theta}_t) = [\boldsymbol{\theta}_t; -A(\boldsymbol{\theta}_t)]$$

Then BONG agrees with the exact Bayesian update.

Proof (Sketch).

Write the natural parameters of the prior as $\boldsymbol{\psi}_{t|t-1} = [\boldsymbol{\chi}_{t|t-1}; \nu_{t|t-1}]$
The Bayesian update and BONG both yield

$$\boldsymbol{\chi}_t = \boldsymbol{\chi}_{t|t-1} + \mathbf{y}_t$$

$$\nu_t = \nu_{t|t-1} + 1$$

□

Variational case: Gaussian prior

- Prior $q_{\psi_{t|t-1}}(\boldsymbol{\theta}_t) = \mathcal{N}(\boldsymbol{\theta}_t | \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})$
- Natural parameters $\psi_t = (\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t, -\frac{1}{2}\boldsymbol{\Sigma}_t)$
- BONG update (matches explicit RVGA update; Lambert et al. 2021):

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} [\nabla_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))]$$

$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} - \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} [\nabla_{\boldsymbol{\theta}_t}^2 \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))]$$

- Derivation parallels BLR derivation of VON (Khan, Nielsen, et al. 2018)
  - Convert $\Delta\psi$ to $\Delta(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Bonnet (1964): $\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} [\ell] = \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} [\nabla_{\boldsymbol{\theta}} \ell]$
  - Price (1958): $\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} [\ell] = \frac{1}{2} \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} [\nabla_{\boldsymbol{\theta}}^2 \ell]$

Variational case: Gaussian prior

- Prior $q_{\psi_{t|t-1}}(\boldsymbol{\theta}_t) = \mathcal{N}\left(\boldsymbol{\theta}_t | \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right)$
- Natural parameters $\boldsymbol{\psi}_t = \left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t, -\frac{1}{2}\boldsymbol{\Sigma}_t\right)$
- BONG update (matches explicit RVGA update; Lambert et al. 2021):

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}}\left[\nabla_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))\right]$$

$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} - \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}}\left[\nabla^2_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))\right]$$

- Derivation parallels BLR derivation of VON (Khan, Nielsen, et al. 2018)
  - Convert $\Delta\psi$ to $\Delta(\mu, \Sigma)$
  - Bonnet (1964): $\nabla_\mu \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\Sigma)}[\ell] = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\Sigma)}[\nabla_\theta \ell]$
  - Price (1958): $\nabla_\Sigma \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\Sigma)}[\ell] = \frac{1}{2}\mathbb{E}_{\theta\sim\mathcal{N}(\mu,\Sigma)}[\nabla^2_\theta \ell]$

Variational case: Gaussian prior

- Prior $q_{\psi_{t|t-1}}(\boldsymbol{\theta}_t) = \mathcal{N}\left(\boldsymbol{\theta}_t | \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right)$
- Natural parameters $\psi_t = \left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t, -\frac{1}{2}\boldsymbol{\Sigma}_t\right)$
- BONG update (matches explicit RVGA update; Lambert et al. 2021):

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} \left[\nabla_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))\right]$$

$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} - \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} \left[\nabla^2_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))\right]$$

- Derivation parallels BLR derivation of VON (Khan, Nielsen, et al. 2018)
  - Convert $\Delta\psi$ to $\Delta(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Bonnet (1964): $\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\ell] = \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\nabla_{\boldsymbol{\theta}} \ell]$
  - Price (1958): $\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\ell] = \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}\left[\nabla^2_{\boldsymbol{\theta}} \ell\right]$

Space of methods

- Variational family
  - Tradeoff efficiency and expressiveness
  - Compare different parameterizations
- Update rule
  - Compare NGD to GD
  - Compare implicit regularization (1-step update) to explicit (iterated update)
- Approximating the expectations
  - $\mathbb{E}_{q_{\psi_{t|t-1}}}[\nabla_{\theta_t} \log p]$, $\mathbb{E}_{q_{\psi_{t|t-1}}}[\nabla^2_{\theta_t} \log p]$
  - Monte Carlo
  - Linearized methods
  - Empirical Fisher

Space of methods

- Variational family
  - Tradeoff efficiency and expressiveness
  - Compare different parameterizations
- Update rule
  - Compare NGD to GD
  - Compare implicit regularization (1-step update) to explicit (iterated update)
- Approximating the expectations
  - $\mathbb{E}_{q_{\psi_{t|t-1}}}[\nabla_{\theta_t} \log p]$, $\mathbb{E}_{q_{\psi_{t|t-1}}}[\nabla^2_{\theta_t} \log p]$
  - Monte Carlo
  - Linearized methods
  - Empirical Fisher

Space of methods

- Variational family
  - Tradeoff efficiency and expressiveness
  - Compare different parameterizations
- Update rule
  - Compare NGD to GD
  - Compare implicit regularization (1-step update) to explicit (iterated update)
- Approximating the expectations
  - $\mathbb{E}_{q_{\psi_{t|t-1}}}[\nabla_{\theta_t} \log p]$, $\mathbb{E}_{q_{\psi_{t|t-1}}}\left[\nabla^2_{\theta_t} \log p\right]$
  - Monte Carlo
  - Linearized methods
  - Empirical Fisher

Variational family

- Full-covariance (FC) Gaussian

$$\psi = \left(\mathbf{\Sigma}^{-1} \boldsymbol{\mu}, -\frac{1}{2} \mathbf{\Sigma}^{-1} \right)$$

- Diagonal Gaussian: linear scaling with model size

$$\mathbf{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}^2\right), \quad \psi = \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}^2}, -\frac{1}{2\boldsymbol{\sigma}^2} \right)$$

- FC Gaussian, moment parameterization: importance of natural parameters

$$\psi = (\boldsymbol{\mu}, \mathbf{\Sigma})$$

- Diagonal Gaussian, moment parameterization

$$\psi = \left(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\right)$$

- Diagonal + low-rank Gaussian (DLR; Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)

$$\mathcal{N}\left(\boldsymbol{\mu}, \left(\mathrm{Diag}\left(\boldsymbol{\Upsilon}\right) + \boldsymbol{W}\boldsymbol{W}^\intercal\right)^{-1}\right)$$

with $\boldsymbol{W} \in \mathbb{R}^{P \times R}$, $R \ll P$. Linear scaling but tracks correlations.
NGD methods: update with FC params, then SVD (Mishkin et al. 2018; Chang, Durán-Martín, et al. 2023)

Variational family

- Full-covariance (FC) Gaussian

$$\psi = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right)$$

- Diagonal Gaussian: linear scaling with model size

$$\boldsymbol{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}^2\right), \quad \psi = \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}^2}, -\frac{1}{2\boldsymbol{\sigma}^2}\right)$$

- FC Gaussian, moment parameterization: importance of natural parameters

$$\psi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Diagonal Gaussian, moment parameterization

$$\psi = \left(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\right)$$

- Diagonal + low-rank Gaussian (DLR; Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)

$$\mathcal{N}\left(\boldsymbol{\mu}, \left(\mathrm{Diag}\left(\boldsymbol{\Upsilon}\right) + \boldsymbol{W}\boldsymbol{W}^{\intercal}\right)^{-1}\right)$$

with $\boldsymbol{W} \in \mathbb{R}^{P \times R}$, $R \ll P$. Linear scaling but tracks correlations.
NGD methods: update with FC params, then SVD (Mishkin et al. 2018; Chang, Durán-Martín, et al. 2023)

Variational family

- Full-covariance (FC) Gaussian

$$\psi = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \right)$$

- Diagonal Gaussian: linear scaling with model size

$$\boldsymbol{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}^2\right), \quad \psi = \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}^2}, -\frac{1}{2\boldsymbol{\sigma}^2} \right)$$

- FC Gaussian, moment parameterization: importance of natural parameters

$$\psi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Diagonal Gaussian, moment parameterization

$$\psi = \left(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\right)$$

- Diagonal + low-rank Gaussian (DLR; Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)

$$\mathcal{N}\left(\boldsymbol{\mu}, \left(\mathrm{Diag}\left(\boldsymbol{\Upsilon}\right) + \boldsymbol{W}\boldsymbol{W}^{\intercal}\right)^{-1}\right)$$

with $\boldsymbol{W} \in \mathbb{R}^{P \times R}$, $R \ll P$. Linear scaling but tracks correlations.
NGD methods: update with FC params, then SVD (Mishkin et al. 2018; Chang, Durán-Martín, et al. 2023)

Variational family

- Full-covariance (FC) Gaussian

$$\psi = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right)$$

- Diagonal Gaussian: linear scaling with model size

$$\boldsymbol{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}^2\right), \quad \psi = \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}^2}, -\frac{1}{2\boldsymbol{\sigma}^2}\right)$$

- FC Gaussian, moment parameterization: importance of natural parameters

$$\psi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Diagonal Gaussian, moment parameterization

$$\psi = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$

- Diagonal + low-rank Gaussian (DLR; Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)

$$\mathcal{N}\left(\boldsymbol{\mu}, \left(\mathrm{Diag}\left(\boldsymbol{\Upsilon}\right) + \boldsymbol{W}\boldsymbol{W}^{\intercal}\right)^{-1}\right)$$

with $\boldsymbol{W} \in \mathbb{R}^{P \times R}$, $R \ll P$. Linear scaling but tracks correlations.
NGD methods: update with FC params, then SVD (Mishkin et al. 2018; Chang, Durán-Martín, et al. 2023)

Variational family

- Full-covariance (FC) Gaussian

$$\psi = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \right)$$

- Diagonal Gaussian: linear scaling with model size

$$\boldsymbol{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}^2\right), \quad \psi = \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}^2}, -\frac{1}{2\boldsymbol{\sigma}^2} \right)$$

- FC Gaussian, moment parameterization: importance of natural parameters

$$\psi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Diagonal Gaussian, moment parameterization

$$\psi = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$

- Diagonal + low-rank Gaussian (DLR; Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)

$$\mathcal{N}\left(\boldsymbol{\mu}, \left(\mathrm{Diag}\left(\boldsymbol{\Upsilon}\right) + \boldsymbol{W}\boldsymbol{W}^{\intercal}\right)^{-1}\right)$$

with $\boldsymbol{W} \in \mathbb{R}^{P \times R}$, $R \ll P$. Linear scaling but tracks correlations.
NGD methods: update with FC params, then SVD (Mishkin et al. 2018; Chang, Durán-Martín, et al. 2023)

Update rules

- BONG (Bayesian online natural gradient): 1 step NGD on NLL

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} [\log p(y_t|f_t(\theta_t))]$$

- oBLR (based on Khan and Rue 2023): iterate NGD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} [\log p(y_t|f_t(\theta_t))] - D_{\mathbb{KL}} \left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}} \right) \right)$$

- BOG (Bayesian online gradient): 1 step GD on NLL

$$\psi_t = \psi_{t|t-1} + \alpha \nabla_{\psi_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} [\log p(y_t|f_t(\theta_t))]$$

- oBBB (based on Blundell et al. 2015): iterate GD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\psi_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} [\log p(y_t|f_t(\theta_t))] - D_{\mathbb{KL}} \left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}} \right) \right)$$

Update rules

- BONG (Bayesian online natural gradient): 1 step NGD on NLL

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p\left(\mathbf{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

- oBLR (based on Khan and Rue 2023): iterate NGD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p\left(\mathbf{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right] - D_{\mathbb{KL}}\left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}}\right)\right)$$

- BOG (Bayesian online gradient): 1 step GD on NLL

$$\psi_t = \psi_{t|t-1} + \alpha \nabla_{\psi_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p\left(\mathbf{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

- oBBB (based on Blundell et al. 2015): iterate GD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\psi_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p\left(\mathbf{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right] - D_{\mathbb{KL}}\left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}}\right)\right)$$

Update rules

- BONG (Bayesian online natural gradient): 1 step NGD on NLL

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p \left(\boldsymbol{y}_t | f_t \left(\boldsymbol{\theta}_t \right) \right) \right]$$

- oBLR (based on Khan and Rue 2023): iterate NGD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p \left(\boldsymbol{y}_t | f_t \left(\boldsymbol{\theta}_t \right) \right) \right] - D_{\mathbb{KL}} \left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}} \right) \right)$$

- BOG (Bayesian online gradient): 1 step GD on NLL

$$\psi_t = \psi_{t|t-1} + \alpha \nabla_{\psi_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p \left(\boldsymbol{y}_t | f_t \left(\boldsymbol{\theta}_t \right) \right) \right]$$

- oBBB (based on Blundell et al. 2015): iterate GD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\psi_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p \left(\boldsymbol{y}_t | f_t \left(\boldsymbol{\theta}_t \right) \right) \right] - D_{\mathbb{KL}} \left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}} \right) \right)$$

Update rules

- BONG (Bayesian online natural gradient): 1 step NGD on NLL

$$\psi_t = \psi_{t|t-1} + \nabla_{\rho_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

- oBLR (based on Khan and Rue 2023): iterate NGD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\rho_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right] - D_{\mathbb{KL}}\left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}}\right)\right)$$

- BOG (Bayesian online gradient): 1 step GD on NLL

$$\psi_t = \psi_{t|t-1} + \alpha \nabla_{\psi_{t|t-1}} \mathbb{E}_{q_{\psi_{t|t-1}}} \left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right]$$

- oBBB (based on Blundell et al. 2015): iterate GD on online VI loss

$$\psi_{t,i} = \psi_{t,i-1} + \alpha \nabla_{\psi_{t,i-1}} \left(\mathbb{E}_{q_{\psi_{t,i-1}}} \left[\log p\left(\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right)\right] - D_{\mathbb{KL}}\left(q_{\psi_{t,i-1}} | q_{\psi_{t|t-1}}\right)\right)$$

Approximating the expectations

- Expected gradient

$$\boldsymbol{g}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} \left[\nabla_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Expected Hessian

$$\boldsymbol{G}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} \left[\nabla^2_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Monte Carlo: Draw M samples from $q_{\psi_{t|t-1}}$
- Linearized: Analytic expressions from approximate likelihoods
- Full Hessian: Second derivative at each observation
- Empirical Fisher (EF; e.g., Martens 2020): First-order approximation
- 4 combinations: MC-HESS, MC-EF, LIN-HESS, LIN-EF

Approximating the expectations

- Expected gradient

$$\boldsymbol{g}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} \left[\nabla_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Expected Hessian

$$\boldsymbol{G}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_{t|t-1}}} \left[\nabla^2_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Monte Carlo: Draw M samples from $q_{\psi_{t|t-1}}$
- Linearized: Analytic expressions from approximate likelihoods
- Full Hessian: Second derivative at each observation
- Empirical Fisher (EF; e.g., Martens 2020): First-order approximation
- 4 combinations: MC-HESS, MC-EF, LIN-HESS, LIN-EF

Approximating the expectations

- Expected gradient

$$\boldsymbol{g}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\boldsymbol{\psi}_{t|t-1}}} \left[\nabla_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Expected Hessian

$$\boldsymbol{G}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\boldsymbol{\psi}_{t|t-1}}} \left[\nabla^2_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Monte Carlo: Draw M samples from $q_{\boldsymbol{\psi}_{t|t-1}}$
- Linearized: Analytic expressions from approximate likelihoods
- Full Hessian: Second derivative at each observation
- Empirical Fisher (EF; e.g., Martens 2020): First-order approximation
- 4 combinations: MC-HESS, MC-EF, LIN-HESS, LIN-EF

Approximating the expectations

- Expected gradient

$$\boldsymbol{g}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\boldsymbol{\psi}_{t|t-1}}} \left[\nabla_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Expected Hessian

$$\boldsymbol{G}_t = \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\boldsymbol{\psi}_{t|t-1}}} \left[\nabla^2_{\boldsymbol{\theta}_t} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t)) \right]$$

- Monte Carlo: Draw M samples from $q_{\boldsymbol{\psi}_{t|t-1}}$
- Linearized: Analytic expressions from approximate likelihoods
- Full Hessian: Second derivative at each observation
- Empirical Fisher (EF; e.g., Martens 2020): First-order approximation
- 4 combinations: MC-HESS, MC-EF, LIN-HESS, LIN-EF

Monte Carlo methods

- Sample $\left\{\hat{\theta}_t^{(m)} : 1 \leq m \leq M\right\}$
- Approximate mean gradient

$$g_t^{\mathrm{MC}} = \frac{1}{M}\sum_{m=1}^{M}\hat{g}_t^{(m)}, \quad \hat{g}_t^{(m)} = \nabla_{\theta_t=\hat{\theta}_t^{(m)}} \log p(y_t|f_t(\theta_t))$$

- Approximate mean Hessian, 2nd-order method

$$G_t^{\mathrm{MC-HESS}} = \frac{1}{M}\sum_{m=1}^{M}\hat{G}_t^{(m)}, \quad \hat{G}_t^{(m)} = \nabla^2_{\theta_t=\hat{\theta}_t^{(m)}} \log p(y_t|f_t(\theta_t))$$

- Approximate mean Hessian, EF method

$$G_t^{\mathrm{MC-EF}} = -\frac{1}{M}\sum_{m=1}^{M}\hat{g}_t^{(m)}\hat{g}_t^{(m)\intercal}$$

Monte Carlo methods

- Sample $\left\{\hat{\boldsymbol{\theta}}_t^{(m)} : 1 \leq m \leq M\right\}$
- Approximate mean gradient

$$\boldsymbol{g}_t^{\mathrm{MC}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{g}}_t^{(m)}, \quad \hat{\boldsymbol{g}}_t^{(m)} = \nabla_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t^{(m)}} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))$$

- Approximate mean Hessian, 2nd-order method

$$\boldsymbol{G}_t^{\mathrm{MC-HESS}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{G}}_t^{(m)}, \quad \hat{\boldsymbol{G}}_t^{(m)} = \nabla^2_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t^{(m)}} \log p(\boldsymbol{y}_t | f_t(\boldsymbol{\theta}_t))$$

- Approximate mean Hessian, EF method

$$\boldsymbol{G}_t^{\mathrm{MC-EF}} = -\frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{g}}_t^{(m)} \hat{\boldsymbol{g}}_t^{(m)\intercal}$$

Monte Carlo methods

- Sample $\left\{\hat{\boldsymbol{\theta}}_t^{(m)} : 1 \leq m \leq M\right\}$
- Approximate mean gradient

$$\boldsymbol{g}_t^{\mathrm{MC}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{g}}_t^{(m)}, \quad \hat{\boldsymbol{g}}_t^{(m)} = \nabla_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t^{(m)}} \log p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$$

- Approximate mean Hessian, 2nd-order method

$$\boldsymbol{G}_t^{\mathrm{MC-HESS}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{G}}_t^{(m)}, \quad \hat{\boldsymbol{G}}_t^{(m)} = \nabla^2_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t^{(m)}} \log p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$$

- Approximate mean Hessian, EF method

$$\boldsymbol{G}_t^{\mathrm{MC-EF}} = -\frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{g}}_t^{(m)} \hat{\boldsymbol{g}}_t^{(m)\intercal}$$

Monte Carlo methods

- Sample $\left\{\hat{\boldsymbol{\theta}}_t^{(m)} : 1 \leq m \leq M\right\}$
- Approximate mean gradient

$$\boldsymbol{g}_t^{\mathrm{MC}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{g}}_t^{(m)}, \quad \hat{\boldsymbol{g}}_t^{(m)} = \nabla_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t^{(m)}} \log p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$$

- Approximate mean Hessian, 2nd-order method

$$\boldsymbol{G}_t^{\mathrm{MC-HESS}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{G}}_t^{(m)}, \quad \hat{\boldsymbol{G}}_t^{(m)} = \nabla^2_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t^{(m)}} \log p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t))$$

- Approximate mean Hessian, EF method

$$\boldsymbol{G}_t^{\mathrm{MC-EF}} = -\frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{g}}_t^{(m)}\hat{\boldsymbol{g}}_t^{(m)\intercal}$$

Linearized methods

- Assume exponential-family likelihood

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\theta}_t) = \exp\left(f_t\left(\boldsymbol{\theta}_t\right)^\intercal \boldsymbol{y}_t - A(f_t(\boldsymbol{\theta}_t)) - b(\boldsymbol{y}_t)\right)$$

- Define natural parameters $h_t\left(\boldsymbol{\theta}_t\right) = \mathbb{E}\left[\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right]$
  e.g., $f_t\left(\boldsymbol{\theta}_t\right)$ is logits and $h_t\left(\boldsymbol{\theta}_t\right) = \mathrm{softmax}\left(f_t\left(\boldsymbol{\theta}_t\right)\right)$ is probabilities
- **Linear(h)-Gaussian approximation** (Ollivier 2018; Tronarp et al. 2018): linearize $h_t\left(\boldsymbol{\theta}_t\right)$ and approximate likelihood as Gaussian

$$p\left(\boldsymbol{y}_t|h_t\left(\boldsymbol{\theta}_t\right)\right) \approx \mathcal{N}(\boldsymbol{y}_t| \underbrace{\bar{h}_t\left(\boldsymbol{\theta}_t\right)}_{\text{linearized about } \boldsymbol{\mu}_{t|t-1}}, \underbrace{\hat{R}_t}_{\text{conditional variance at } \boldsymbol{\mu}_{t|t-1}})$$

- **Linear(f)-delta approximation**: Linearize $f_t\left(\boldsymbol{\theta}_t\right)$ and use mean plug-in

$$p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t)) \approx \exp(\underbrace{\bar{f}_t\left(\boldsymbol{\theta}_t\right)^\intercal}_{\text{linearized about } \boldsymbol{\mu}_{t|t-1}} \boldsymbol{y}_t - A(\bar{f}_t(\boldsymbol{\theta}_t)) - b(\boldsymbol{y}_t))$$

$$q_{\psi_{t|t-1}}\left(\boldsymbol{\theta}_t\right) \approx \delta_{\boldsymbol{\mu}_{t|t-1}}\left(\boldsymbol{\theta}_t\right)$$

Linearized methods

- Assume exponential-family likelihood

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}_t) = \exp\left(f_t\left(\boldsymbol{\theta}_t\right)^\intercal \boldsymbol{y}_t - A(f_t(\boldsymbol{\theta}_t)) - b(\boldsymbol{y}_t)\right)$$

- Define natural parameters $h_t\left(\boldsymbol{\theta}_t\right) = \mathbb{E}\left[\boldsymbol{y}_t | f_t\left(\boldsymbol{\theta}_t\right)\right]$
  e.g., $f_t\left(\boldsymbol{\theta}_t\right)$ is logits and $h_t\left(\boldsymbol{\theta}_t\right) = \mathrm{softmax}\left(f_t\left(\boldsymbol{\theta}_t\right)\right)$ is probabilities
- **Linear(h)-Gaussian approximation** (Ollivier 2018; Tronarp et al. 2018): linearize $h_t\left(\boldsymbol{\theta}_t\right)$ and approximate likelihood as Gaussian

$$p\left(\boldsymbol{y}_t | h_t\left(\boldsymbol{\theta}_t\right)\right) \approx \mathcal{N}(\boldsymbol{y}_t | \underbrace{\bar{h}_t\left(\boldsymbol{\theta}_t\right)}_{\substack{\text{linearized} \\ \text{about } \boldsymbol{\mu}_{t|t-1}}}, \underbrace{\hat{\boldsymbol{R}}_t}_{\substack{\text{conditional} \\ \text{variance} \\ \text{at } \boldsymbol{\mu}_{t|t-1}}})$$

- **Linear(f)-delta approximation**: Linearize $f_t\left(\boldsymbol{\theta}_t\right)$ and use mean plug-in

$$p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t)) \approx \exp(\underbrace{\bar{f}_t\left(\boldsymbol{\theta}_t\right)^\intercal}_{\substack{\text{linearized} \\ \text{about } \boldsymbol{\mu}_{t|t-1}}} \boldsymbol{y}_t - A(\bar{f}_t(\boldsymbol{\theta}_t)) - b(\boldsymbol{y}_t))$$

$$q_{\psi_{t|t-1}}\left(\boldsymbol{\theta}_t\right) \approx \delta_{\boldsymbol{\mu}_{t|t-1}}\left(\boldsymbol{\theta}_t\right)$$

Linearized methods

- Assume exponential-family likelihood

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\theta}_t) = \exp\left(f_t\left(\boldsymbol{\theta}_t\right)^\intercal \boldsymbol{y}_t - A(f_t(\boldsymbol{\theta}_t)) - b(\boldsymbol{y}_t)\right)$$

- Define natural parameters $h_t\left(\boldsymbol{\theta}_t\right) = \mathbb{E}\left[\boldsymbol{y}_t|f_t\left(\boldsymbol{\theta}_t\right)\right]$
  e.g., $f_t\left(\boldsymbol{\theta}_t\right)$ is logits and $h_t\left(\boldsymbol{\theta}_t\right) = \mathrm{softmax}\left(f_t\left(\boldsymbol{\theta}_t\right)\right)$ is probabilities
- **Linear(h)-Gaussian approximation** (Ollivier 2018; Tronarp et al. 2018): linearize $h_t\left(\boldsymbol{\theta}_t\right)$ and approximate likelihood as Gaussian

$$p\left(\boldsymbol{y}_t|h_t\left(\boldsymbol{\theta}_t\right)\right) \approx \mathcal{N}(\boldsymbol{y}_t| \underbrace{\bar{h}_t\left(\boldsymbol{\theta}_t\right)}_{\substack{\text{linearized} \\ \text{about } \boldsymbol{\mu}_{t|t-1}}}, \underbrace{\hat{\boldsymbol{R}}_t}_{\substack{\text{conditional} \\ \text{variance} \\ \text{at } \boldsymbol{\mu}_{t|t-1}}})$$

- **Linear(f)-delta approximation**: Linearize $f_t\left(\boldsymbol{\theta}_t\right)$ and use mean plug-in

$$p(\boldsymbol{y}_t|f_t(\boldsymbol{\theta}_t)) \approx \exp(\underbrace{\bar{f}_t\left(\boldsymbol{\theta}_t\right)^\intercal}_{\substack{\text{linearized} \\ \text{about } \boldsymbol{\mu}_{t|t-1}}} \boldsymbol{y}_t - A(\bar{f}_t(\boldsymbol{\theta}_t)) - b(\boldsymbol{y}_t))$$

$$q_{\psi_{t|t-1}}\left(\boldsymbol{\theta}_t\right) \approx \delta_{\boldsymbol{\mu}_{t|t-1}}\left(\boldsymbol{\theta}_t\right)$$

Linearized methods

Theorem

Under a Gaussian variational distribution, the linear(h)-Gaussian and linear(f)-delta approximations yield the same values for the expected gradient and Hessian:

$$\boldsymbol{g}_t^{\mathrm{LIN}} = \boldsymbol{H}_t^{\intercal} \hat{\boldsymbol{R}}_t^{-1} (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)$$
$$\boldsymbol{G}_t^{\mathrm{LIN-HESS}} = -\boldsymbol{H}_t^{\intercal} \hat{\boldsymbol{R}}_t^{-1} \boldsymbol{H}_t$$

where $\boldsymbol{H}_t = \mathrm{jac}\left(h_t\left(\cdot\right)\right)\left(\boldsymbol{\mu}_{t|t-1}\right)$ *and* $\hat{\boldsymbol{y}}_t = h\left(\boldsymbol{\mu}_{t|t-1}\right)$.

Proof.

Direct calculation.
Intuition: Linear assumptions imply mean gradient equals gradient at mean.
For Hessian, the Gaussian and plug-in approximations require different linearizations to eliminate curvature of the NN, yielding the GGN approximation. □

Linearized methods

- Linear-EF method: Jacobian free
- Expected gradient

$$g_t^{\mathrm{LIN}} = H_t^\mathsf{T} \hat{R}_t^{-1} (y_t - \hat{y}_t)$$
$$= \nabla_{\theta_t = \mu_{t|t-1}} \left[-\frac{1}{2} (y_t - h_t(\theta_t))^\mathsf{T} \hat{R}_t^{-1} (y_t - h_t(\theta_t)) \right]$$

- Expected Hessian

$$G_t^{\mathrm{LIN-EF}} = -g_t^{\mathrm{LIN}} \left(g_t^{\mathrm{LIN}} \right)^\mathsf{T}$$

- Justification: If model were correct, meaning $\hat{y}_t = \mathbb{E}[y_t | x_t]$, then $\mathbb{E}[(y_t - \hat{y}_t)(y_t - \hat{y}_t)^\mathsf{T}] = \hat{R}_t$, implying $\mathbb{E}\left[G_t^{\mathrm{LIN-EF}}\right] = G_t^{\mathrm{LIN-HESS}}$

Linearized methods

- Linear-EF method: Jacobian free
- Expected gradient

$$\boldsymbol{g}_t^{\text{LIN}} = \boldsymbol{H}_t^{\mathsf{T}} \hat{\boldsymbol{R}}_t^{-1} (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)$$
$$= \nabla_{\boldsymbol{\theta}_t = \boldsymbol{\mu}_{t|t-1}} \left[-\frac{1}{2} (\boldsymbol{y}_t - h_t(\boldsymbol{\theta}_t))^{\mathsf{T}} \hat{\boldsymbol{R}}_t^{-1} (\boldsymbol{y}_t - h_t(\boldsymbol{\theta}_t)) \right]$$

- Expected Hessian

$$\boldsymbol{G}_t^{\text{LIN-EF}} = -\boldsymbol{g}_t^{\text{LIN}} \left(\boldsymbol{g}_t^{\text{LIN}}\right)^{\mathsf{T}}$$

- Justification: If model were correct, meaning $\hat{\boldsymbol{y}}_t = \mathbb{E}[\boldsymbol{y}_t | \boldsymbol{x}_t]$, then $\mathbb{E}\left[(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)^{\mathsf{T}}\right] = \hat{\boldsymbol{R}}_t$, implying $\mathbb{E}\left[\boldsymbol{G}_t^{\text{LIN-EF}}\right] = \boldsymbol{G}_t^{\text{LIN-HESS}}$

Linearized methods

- Linear-EF method: Jacobian free
- Expected gradient

$$
\begin{aligned}
\boldsymbol{g}_t^{\mathrm{LIN}} &= \boldsymbol{H}_t^{\intercal} \hat{\boldsymbol{R}}_t^{-1} (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t) \\
&= \nabla_{\boldsymbol{\theta}_t = \boldsymbol{\mu}_{t|t-1}} \left[-\frac{1}{2} (\boldsymbol{y}_t - h_t(\boldsymbol{\theta}_t))^{\intercal} \hat{\boldsymbol{R}}_t^{-1} (\boldsymbol{y}_t - h_t(\boldsymbol{\theta}_t)) \right]
\end{aligned}
$$

- Expected Hessian

$$
\boldsymbol{G}_t^{\mathrm{LIN-EF}} = -\boldsymbol{g}_t^{\mathrm{LIN}} \left(\boldsymbol{g}_t^{\mathrm{LIN}} \right)^{\intercal}
$$

- Justification: If model were correct, meaning $\hat{\boldsymbol{y}}_t = \mathbb{E}[\boldsymbol{y}_t | \boldsymbol{x}_t]$, then $\mathbb{E}[(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)^{\intercal}] = \hat{\boldsymbol{R}}_t$, implying $\mathbb{E}\left[\boldsymbol{G}_t^{\mathrm{LIN-EF}}\right] = \boldsymbol{G}_t^{\mathrm{LIN-HESS}}$

Linearized methods

- Linear-EF method: Jacobian free
- Expected gradient

$$\begin{aligned} \boldsymbol{g}_t^{\mathrm{LIN}} &= \boldsymbol{H}_t^\intercal \hat{\boldsymbol{R}}_t^{-1} (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t) \\ &= \nabla_{\boldsymbol{\theta}_t = \boldsymbol{\mu}_{t|t-1}} \left[-\frac{1}{2} (\boldsymbol{y}_t - h_t(\boldsymbol{\theta}_t))^\intercal \hat{\boldsymbol{R}}_t^{-1} (\boldsymbol{y}_t - h_t(\boldsymbol{\theta}_t)) \right] \end{aligned}$$

- Expected Hessian

$$\boldsymbol{G}_t^{\mathrm{LIN-EF}} = -\boldsymbol{g}_t^{\mathrm{LIN}} \left(\boldsymbol{g}_t^{\mathrm{LIN}} \right)^\intercal$$

- Justification: If model were correct, meaning $\hat{\boldsymbol{y}}_t = \mathbb{E}[\boldsymbol{y}_t | \boldsymbol{x}_t]$, then $\mathbb{E}[(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)^\intercal] = \hat{\boldsymbol{R}}_t$, implying $\mathbb{E}\left[\boldsymbol{G}_t^{\mathrm{LIN-EF}}\right] = \boldsymbol{G}_t^{\mathrm{LIN-HESS}}$

Space of methods

Time Complexity		Family and parameterization				
Method	Approx	FC, natural	FC, moment	Diag, natural	Diag, moment	DLR
BONG	MC-EF	$O(MP^2)^*$ [RVGA]	$O(MP^2)^*$	$O(MP)^*$	$O(MP)^*$	$O((R+M)^2P)^*$
oBLR	MC-EF	$O(IP^3)$	$O(IP^3)$	$O(IMP)^*$ [VON]	$O(IMP)^*$	$O(I(R+M)^2P)^*$ [SLANG]
BOG	MC-EF	$O(P^3)$	$O(MP^2)$	$O(MP)^*$	$O(MP)^*$	$O(RMP)^*$
oBBB	MC-EF	$O(IP^3)$	$O(IP^3)$	$O(IMP)^*$	$O(IMP)^*$ [BBB]	$O(IR(R+M)P)^*$
BONG	LIN-HESS	$O(CP^2)$ [CM-EKF]	$O(CP^2)$	$O(C^2P)$ [VD-EKF]	$O(C^2P)$	$O((R+C)^2P)$ [LO-FI]
oBLR	LIN-HESS	$O(IP^3)$	$O(IP^3)$	$O(IC^2P)$	$O(IC^2P)$	$O(I(2R+C)^2P)$
BOG	LIN-HESS	$O(P^3)$	$O(CP^2)$	$O(C^2P)$	$O(C^2P)$	$O(C(C+R)P)$
oBBB	LIN-HESS	$O(IP^3)$	$O(IP^3)$	$O(IC^2P)$	$O(IC^2P)$	$O(I(C+R)RP)$

- P: params, C: observation dim, M: MC samples, I: iterations, R: DLR rank
- $*$: MC-EF asymptotically faster than MC-HESS (otherwise equal)
- LIN-EF complexities: $C \to 1$
- RGVA: Lambert et al. (2021) (explicit update version)
- VON: Khan, Nielsen, et al. (2018) (modified for online)
- SLANG: Mishkin et al. (2018) (modified for online)
- BBB: Blundell et al. (2015) (modified for online)
- CM-EKF: Ollivier (2018) and Tronarp et al. (2018)
- VD-EKF: Chang, Murphy, et al. (2022)
- LO-FI: Chang, Durán-Martín, et al. (2023)

Experiments

- Comparisons
  - Implicit regularization: {BONG,BOG} vs {oBLR,oBBB}
  - NGD: {BONG,oBLR} vs {BOG,oBBB}
  - Linearization: LIN-HESS vs MC-EF
  - Parameterization: natural vs moment
- Datasets
  - Synthetic linear regression
  - MNIST: 10-way classification, $D = 784$, $N_{\text{train}} = 60k$, $N_{\text{test}} = 10k$, CNN with $P = 57{,}722$
  - SARCOS: 1d regression (robotic inverse dynamics, https://gaussianprocess.org/gpml/data/) $D = 22$, $N_{\text{train}} = 44{,}484$, $N_{\text{test}} = 4{,}449$, MLP (21-20-20-1) with $P = 881$
- Metrics
  - Speed
  - Misclassification (MNIST)
  - Negative log predictive density: $\text{NLPD}_t = -\frac{1}{N_{\text{test}}} \sum_{j \in \mathcal{D}_{\text{test}}} \log \mathbb{E}_{\theta_t \sim q_{\psi_t}} \left[p\left(y_j | f_j\left(\theta_t\right)\right)\right]$
  - Monte Carlo: sample $\theta_t^{1:S} \sim q_{\psi_t}$
  - Linear Monte Carlo: evaluate linear-Gaussianized model $\mathcal{N}\left(y_j | \bar{h}_j\left(\theta_t\right), \hat{R}_{tj}\right)$ (Immer et al. 2021)
  - Mean plug-in: $\log p\left(y_j | f_j\left(\mu_t\right)\right)$
- Hyperparams
  - Learning rate optimized on validation set (oBLR,BOG,oBBB)
  - Prior $q_{\psi_0}\left(\theta_0\right) = \mathcal{N}\left(\theta_0 | \mu_0, \sigma_0^2 I_P\right)$: optimize σ_0^2 and sample μ_0 from standard NN initializer

Experiments

- Comparisons
  - Implicit regularization: {BONG,BOG} vs {oBLR,oBBB}
  - NGD: {BONG,oBLR} vs {BOG,oBBB}
  - Linearization: LIN-HESS vs MC-EF
  - Parameterization: natural vs moment
- Datasets
  - Synthetic linear regression
  - MNIST: 10-way classification, $D = 784$, $N_{\text{train}} = 60k$, $N_{\text{test}} = 10k$, CNN with $P = 57{,}722$
  - SARCOS: 1d regression (robotic inverse dynamics, https://gaussianprocess.org/gpml/data/) $D = 22$, $N_{\text{train}} = 44{,}484$, $N_{\text{test}} = 4{,}449$, MLP (21-20-20-1) with $P = 881$
- Metrics
  - Speed
  - Misclassification (MNIST)
  - Negative log predictive density: $\text{NLPD}_t = -\frac{1}{N_{\text{test}}} \sum_{j \in \mathcal{D}_{\text{test}}} \log \mathbb{E}_{\theta_t \sim q_{\psi_t}} \left[p\left(\boldsymbol{y}_j | f_j\left(\boldsymbol{\theta}_t\right)\right)\right]$
  - Monte Carlo: sample $\boldsymbol{\theta}_t^{1:S} \sim q_{\psi_t}$
  - Linear Monte Carlo: evaluate linear-Gaussianized model $\mathcal{N}\left(\boldsymbol{y}_j | \bar{h}_j\left(\boldsymbol{\theta}_t\right), \hat{\boldsymbol{R}}_{tj}\right)$ (Immer et al. 2021)
  - Mean plug-in: $\log p\left(\boldsymbol{y}_j | f_j\left(\boldsymbol{\mu}_t\right)\right)$
- Hyperparams
  - Learning rate optimized on validation set (oBLR,BOG,oBBB)
  - Prior $q_{\psi_0}\left(\boldsymbol{\theta}_0\right) = \mathcal{N}\left(\boldsymbol{\theta}_0 | \boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{I}_P\right)$: optimize σ_0^2 and sample $\boldsymbol{\mu}_0$ from standard NN initializer

Experiments

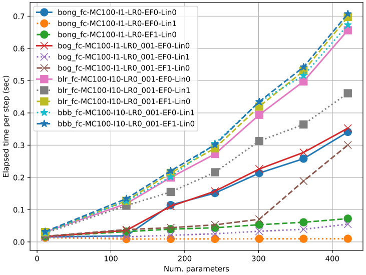
- Comparisons
  - Implicit regularization: {BONG,BOG} vs {oBLR,oBBB}
  - NGD: {BONG,oBLR} vs {BOG,oBBB}
  - Linearization: LIN-HESS vs MC-EF
  - Parameterization: natural vs moment
- Datasets
  - Synthetic linear regression
  - MNIST: 10-way classification, $D = 784$, $N_{\text{train}} = 60k$, $N_{\text{test}} = 10k$, CNN with $P = 57{,}722$
  - SARCOS: 1d regression (robotic inverse dynamics, https://gaussianprocess.org/gpml/data/) $D = 22$, $N_{\text{train}} = 44{,}484$, $N_{\text{test}} = 4{,}449$, MLP (21-20-20-1) with $P = 881$
- Metrics
  - Speed
  - Misclassification (MNIST)
  - Negative log predictive density: $\mathrm{NLPD}_t = -\frac{1}{N_{\text{test}}} \sum_{j \in \mathcal{D}_{\text{test}}} \log \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_t}} \left[p\left(\boldsymbol{y}_j | f_j\left(\boldsymbol{\theta}_t \right) \right) \right]$
  - Monte Carlo: sample $\boldsymbol{\theta}_t^{1:S} \sim q_{\psi_t}$
  - Linear Monte Carlo: evaluate linear-Gaussianized model $\mathcal{N}\left(\boldsymbol{y}_j | \bar{h}_j\left(\boldsymbol{\theta}_t \right), \hat{\boldsymbol{R}}_{tj} \right)$ (Immer et al. 2021)
  - Mean plug-in: $\log p\left(\boldsymbol{y}_j | f_j\left(\boldsymbol{\mu}_t \right) \right)$
- Hyperparams
  - Learning rate optimized on validation set (oBLR,BOG,oBBB)
  - Prior $q_{\psi_0}\left(\theta_0\right) = \mathcal{N}\left(\theta_0 | \mu_0, \sigma_0^2 \boldsymbol{I}_P\right)$: optimize σ_0^2 and sample μ_0 from standard NN initializer

Experiments

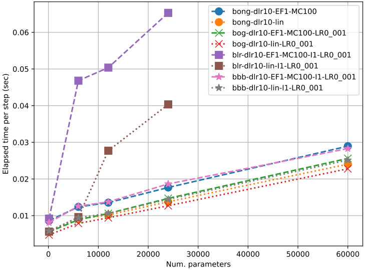
- Comparisons
  - Implicit regularization: {BONG,BOG} vs {oBLR,oBBB}
  - NGD: {BONG,oBLR} vs {BOG,oBBB}
  - Linearization: LIN-HESS vs MC-EF
  - Parameterization: natural vs moment
- Datasets
  - Synthetic linear regression
  - MNIST: 10-way classification, $D = 784$, $N_{\text{train}} = 60k$, $N_{\text{test}} = 10k$, CNN with $P = 57{,}722$
  - SARCOS: 1d regression (robotic inverse dynamics, https://gaussianprocess.org/gpml/data/) $D = 22$, $N_{\text{train}} = 44{,}484$, $N_{\text{test}} = 4{,}449$, MLP (21-20-20-1) with $P = 881$
- Metrics
  - Speed
  - Misclassification (MNIST)
  - Negative log predictive density: $\text{NLPD}_t = -\frac{1}{N_{\text{test}}} \sum_{j \in \mathcal{D}_{\text{test}}} \log \mathbb{E}_{\boldsymbol{\theta}_t \sim q_{\psi_t}} \left[p\left(\boldsymbol{y}_j | f_j\left(\boldsymbol{\theta}_t\right)\right)\right]$
  - Monte Carlo: sample $\boldsymbol{\theta}_t^{1:S} \sim q_{\psi_t}$
  - Linear Monte Carlo: evaluate linear-Gaussianized model $\mathcal{N}\left(\boldsymbol{y}_j | \bar{h}_j\left(\boldsymbol{\theta}_t\right), \hat{\boldsymbol{R}}_{tj}\right)$ (Immer et al. 2021)
  - Mean plug-in: $\log p\left(\boldsymbol{y}_j | f_j\left(\boldsymbol{\mu}_t\right)\right)$
- Hyperparams
  - Learning rate optimized on validation set (oBLR,BOG,oBBB)
  - Prior $q_{\psi_0}\left(\boldsymbol{\theta}_0\right) = \mathcal{N}\left(\boldsymbol{\theta}_0 | \boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{I}_P\right)$: optimize σ_0^2 and sample $\boldsymbol{\mu}_0$ from standard NN initializer

Timing

- Full covariance: {BONG,oBLR,BOG,oBBB} $\times$ {MC-HESS,LIN-HESS,MC-EF}
- DLR (rank $R = 10$): {BONG,oBLR,BOG,oBBB} $\times$ {LIN-HESS,MC-EF}
- Big speedups from linearization and implicit regularization ($I = 1$)
- NGD faster than GD for FC; slower for DLR because of SVD
- SVD dimension is larger for oBLR than BONG



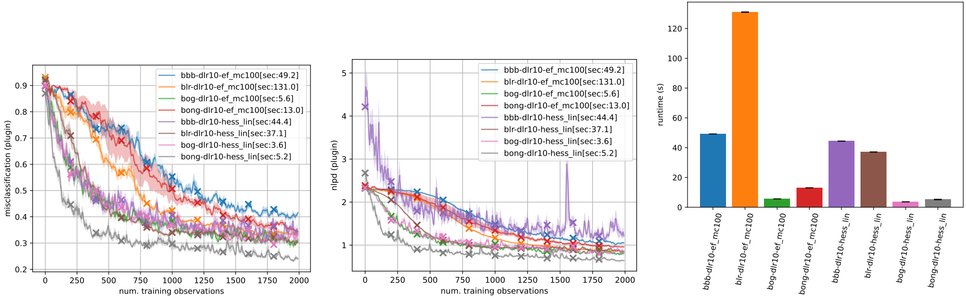
(a) Full covariance, 10 iterations for BLR,BBB.



(b) DLR10, 1 iteration.

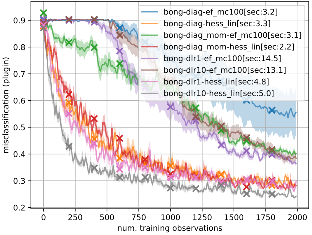
MNIST: main algorithms

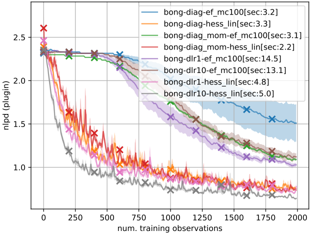
- {BONG,oBLR,BOG,oBBB} $\times$ {MC-EF,LIN-HESS}
- DLR (rank $R = 10$)
- Benefits of 3 main principles: implicit regularization, NGD, linearization
- Win for BONG-LIN (LO-FI, Chang, Durán-Martín, et al. 2023)

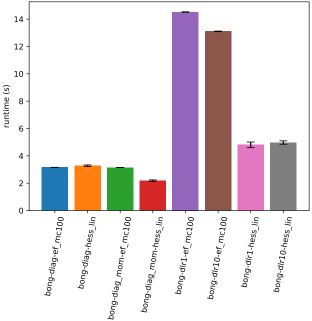


MNIST: BONG variants

- {Diag,Diag-Moment,DLR1,DLR10} × {MC-EF,LIN-HESS}
- FC ≈ DLR10 > DLR1 > Diag ≈ Diag-Moment
- DLR LIN-HESS (LO-FI) reasonably fast (LIN-EF not implemented)

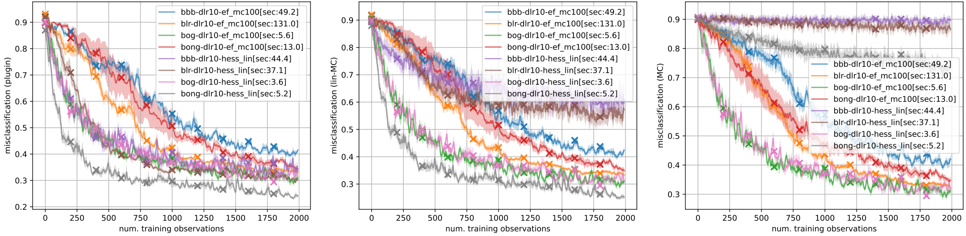






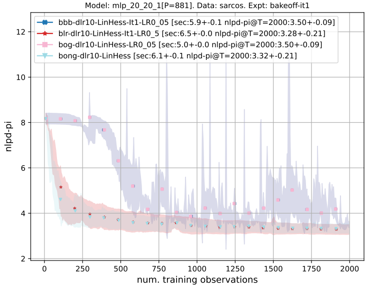
MNIST: predictive distributions

- Linearized methods do poorly with MC on nonlinear model (Immer et al. 2021)
- Predicting with linearized model matches mean plug-in
- Same pattern for NLPD (not shown)

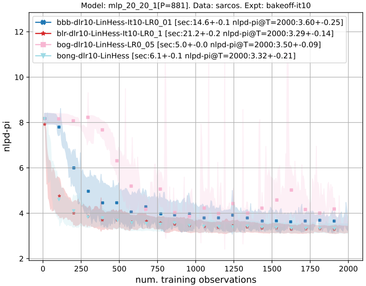


SARCOS: Linear methods

- {BONG,oBLR,BOG,oBBB} $\times$ {1 iteration, 10 iterations (oBLR,oBBB)}, all DLR10
- Advantage for NGD methods (BONG,oBLR)
- BLR with 10 iterations catches up to BONG
- Iterated methods are slower (oBLR,oBBB)



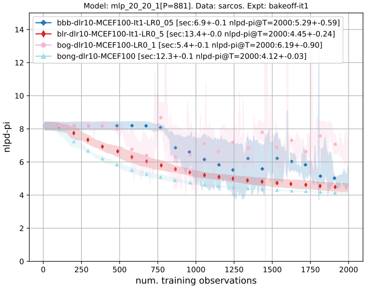
(a) 1 iteration.



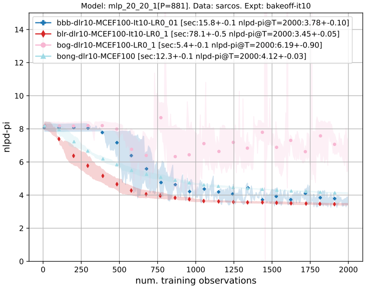
(b) 10 iterations.

SARCOS: MC methods

- {BONG,oBLR,BOG,oBBB} $\times$ {1 iteration, 10 iterations (oBLR,oBBB)}, all DLR10
- Advantage for NGD methods (BONG,oBLR)
- oBLR with 10 iterations outperforms BONG though 6x slower
- All methods learn slower than linear versions



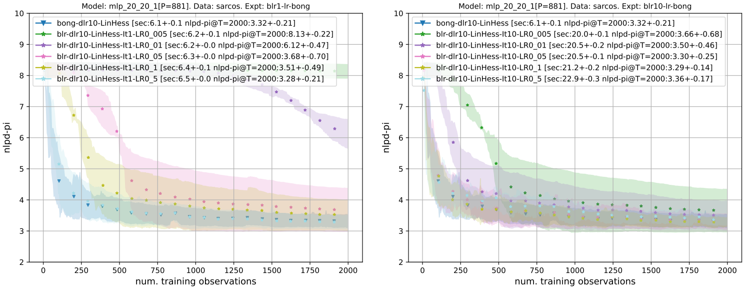
(a) 1 iteration.



(b) 10 iterations.

oBLR learning rates (SARCOS)

- $\alpha \in \{.005, .01, .05, .1, .5\}$, DLR10 with LIN-HESS
- Also compared to BONG ($\alpha \equiv 1$)
- oBLR sensitive to learning rate, though less so with more iterations

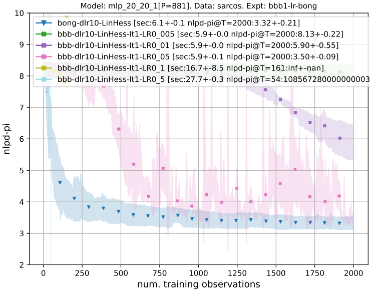


(a) 1 iteration.

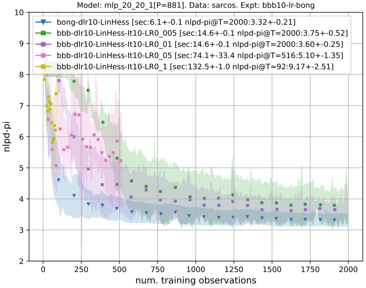
(b) 10 iterations.

oBBB learning rates (SARCOS)

- $\alpha \in \{.005, .01, .05, .1, .5\}$, DLR10 with LIN-HESS
- Also compared to BONG ($\alpha \equiv 1$)
- Sensitive to learning rate
- Performs better with more iterations, still $<$ BONG



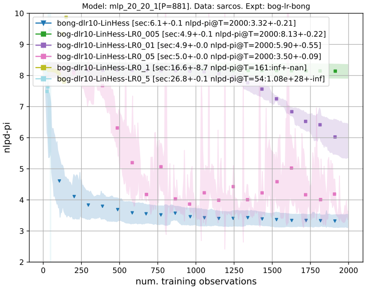
(a) 1 iteration.



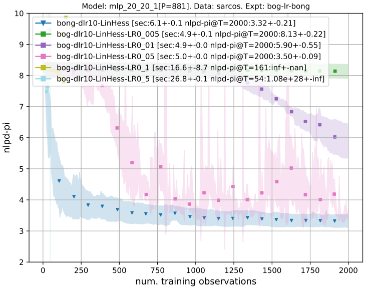
(b) 10 iterations.

BOG learning rates (SARCOS)

- $\alpha \in \{.005, .01, .05, .1, .5\}$, DLR10, using LIN-HESS and MC-EF
- Also compared to BONG ($\alpha \equiv 1$)
- Sensitive to learning rate and performs poorly



(a) LIN-HESS.



(b) MC-EF.

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Conclusions

- Novel approach to online VI
- Implicit regularization to prior using 1-step NGD (extends BLR; Khan and Rue 2023)
- Multiple routes to linearized model (extends CM-EKF; Tronarp et al. 2018; Ollivier 2018)
- Scale up using DLR Gaussian (Mishkin et al. 2018; Lambert et al. 2023; Chang, Durán-Martín, et al. 2023)
- Framework unifies many existing methods and new ones (e.g., SLANG-LIN; cf. Mishkin et al. 2018)
- Experiments support all three principles: implicit regularization, NGD, linearization
- Overall winner is LO-FI (Chang, Durán-Martín, et al. 2023): BONG-LIN on DLR prior
- Goals: scale to larger models and nonstationary settings

Bencomo, Gianluca M, Jake C Snell, and Thomas L Griffiths (2023). "Implicit Maximum a Posteriori Filtering via Adaptive Optimization". In: *arXiv preprint arXiv:2311.10580*.

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra (2015). "Weight Uncertainty in Neural Networks". In: *ICML*. URL: http://arxiv.org/abs/1505.05424.

Bonnet, Georges (1964). "Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire". In: *Annales des Télécommunications*. Vol. 19. Springer, pp. 203–220.

Chang, Peter G, Gerardo Durán-Martín, Alexander Y Shestopaloff, Matt Jones, and Kevin Murphy (May 2023). "Low-rank extended Kalman filtering for online learning of neural networks from streaming data". In: *COLLAS*. URL: http://arxiv.org/abs/2305.19535.

Chang, Peter G, Kevin Patrick Murphy, and Matt Jones (Dec. 2022). "On diagonal approximations to the extended Kalman filter for online training of Bayesian neural networks". In: *Continual Lifelong Learning Workshop at ACML 2022*. URL: https://openreview.net/forum?id=asgeEt25kk.

Immer, Alexander, Maciej Korzepa, and Matthias Bauer (2021). "Improving predictions of Bayesian neural nets via local linearization". In: *AISTATS*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 703–711. URL: https://proceedings.mlr.press/v130/immer21a.html.

Khan, Mohammad Emtiyaz, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava (2018). "Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam". In: *ICML*. URL: http://arxiv.org/abs/1806.04854.

Khan, Mohammad Emtiyaz and Håvard Rue (2023). "The Bayesian Learning Rule". In: *J. Mach. Learn. Res.* URL: http://arxiv.org/abs/2107.04562.

Lambert, Marc, Silvère Bonnabel, and Francis Bach (Dec. 2021). "The recursive variational Gaussian approximation (R-VGA)". In: *Stat. Comput.* 32.1, p. 10. URL: https://hal.inria.fr/hal-03086627/document.

– (2023). "The limited-memory recursive variational Gaussian approximation (L-RVGA)". In: *Statistics and Computing* 33.3, p. 70.

Martens, James (2020). "New insights and perspectives on the natural gradient method". In: *Journal of Machine Learning Research* 21.146, pp. 1–76.

Mishkin, Aaron, Frederik Kunstner, Didrik Nielsen, Mark Schmidt, and Mohammad Emtiyaz Khan (2018). "SLANG: Fast Structured Covariance

Approximations for Bayesian Deep Learning with Natural Gradient". In: *NIPS*. Curran Associates, Inc., pp. 6245–6255.

Ollivier, Yann (2018). "Online natural gradient as a Kalman filter". en. In: *Electron. J. Stat.* 12.2, pp. 2930–2961. URL: https://projecteuclid.org/euclid.ejs/1537257630.

Price, Robert (1958). "A useful theorem for nonlinear devices having Gaussian inputs". In: *IRE Transactions on Information Theory* 4.2, pp. 69–72.

Puskorius, G V and L A Feldkamp (1991). "Decoupled extended Kalman filter training of feedforward layered networks". In: *International Joint Conference on Neural Networks*. Vol. i, 771–777 vol.1. URL: http://dx.doi.org/10.1109/IJCNN.1991.155276.

Singhal, Sharad and Lance Wu (1989). "Training Multilayer Perceptrons with the Extended Kalman Algorithm". In: *NIPS*. Vol. 1.

Tronarp, Filip, Ángel F García-Fernández, and Simo Särkkä (2018). "Iterative Filtering and Smoothing in Nonlinear and Non-Gaussian Systems Using Conditional Moments". In: *IEEE Signal Process. Lett.* 25.3, pp. 408–412. URL: https://acris.aalto.fi/ws/portalfiles/portal/17669270/cm_parapub.pdf.