Tractable Structured Natural-Gradient Descent Using Local Parameterizations

Wu Lin (UBC), Frank Nielsen (Sony CSL), Mohammad Emtiyaz Khan (AIP, RIKEN), Mark Schmidt (UBC & AMII)

1 Introduction

Natural-gradient descent (NGD) on structured parameter spaces is computationally challenging.

We propose a flexible and efficient NGD method to incorporate structures. Our NGD method

- generalizes the exponential natural evolutionary strategy
- recovers existing Newton-like algorithms
- yields new structured second-order algorithms and structured adaptive-gradient methods
- gives new NGD to learn structured covariances of Gaussian, Wishart and their mixtures

Applications:
(1) deep learning (adaptive-gradient), (2) non-convex optimization (structured 2nd-order),
(3) evolution strategies (gradient-free), (4) variational inference (Monte Carlo (MC) gradient, multimodal density).

2 NGD for Optimization, Inference, and Search

A unified view: optimization over a parametric family $q(\mathbf{w}|\boldsymbol{\tau})$:

$$\min_{\boldsymbol{\tau}\in\Omega_\tau} \mathcal{L}(\boldsymbol{\tau}) := \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}[\ell(\mathbf{w})] + \gamma \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}[\log q(\mathbf{w}|\boldsymbol{\tau})]$$

where $\mathbf{w}$ is the decision variable, $\ell(\mathbf{w})$ is a loss function, Ω_τ is the parameter space of q, and $\gamma \geq 0$ is a constant.

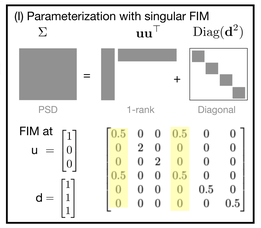
GD and Standard NGD: GD: $\boldsymbol{\tau}_{t+1} \leftarrow \boldsymbol{\tau}_t - \alpha \nabla_{\boldsymbol{\tau}_t}\mathcal{L}(\boldsymbol{\tau})$, Standard NGD: $\boldsymbol{\tau}_{t+1} \leftarrow \boldsymbol{\tau}_t - \beta\ [\mathbf{F}_\tau(\boldsymbol{\tau}_t)]^{-1} \nabla_{\boldsymbol{\tau}_t}\mathcal{L}(\boldsymbol{\tau})$

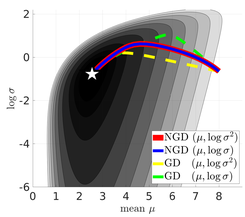
Advantages of NGD:
(1) recovers a Newton-like update for Gaussian q with mean $\boldsymbol{\mu}$ and precision $\mathbf{S}$

$$\text{Standard NGD: } \boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{\mu}_t - \beta \mathbf{S}_t^{-1}\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}[\nabla_w \ell(\mathbf{w})], \quad \mathbf{S}_{t+1} \leftarrow (1-\beta\gamma)\mathbf{S}_t + \beta \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\nabla^2_w \ell(\mathbf{w})\right]$$

(2) less sensitive to parameter transforms than GD
(3) converges faster than GD

Challenges of Standard NGD:
(1) NGD could violate parameterization constraint
(2) singular FIM $\mathbf{F}_\tau(\boldsymbol{\tau})$ of $q(\mathbf{w}|\boldsymbol{\tau})$
(3) limited covariance structures
(4) ad-hoc approximations for cost reductions
(5) complicated, inefficient NG computation





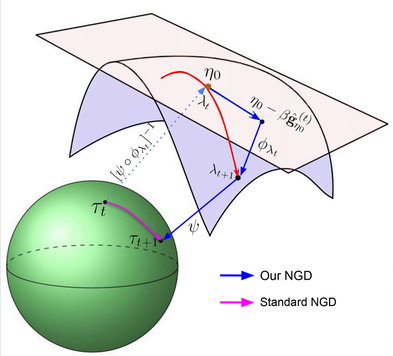
3 NGD using Local Parameterizations

Global parameterization: The original ('global') parameterization $\boldsymbol{\tau}$ for $q(\mathbf{w}|\boldsymbol{\tau})$
Auxiliary parameterization: A new 'auxiliary' parameterization $\boldsymbol{\lambda}$ with a surjective map: $\boldsymbol{\tau} = \psi(\boldsymbol{\lambda})$.
Local parameterization: A 'local' parameterization $\boldsymbol{\eta}$ for $\boldsymbol{\lambda}$ at a current value $\boldsymbol{\lambda}_t$ with a local map: $\boldsymbol{\lambda} = \phi_{\lambda_t}(\boldsymbol{\eta})$. ϕ_{λ_t} is tight at $\boldsymbol{\lambda}_t$: $\boldsymbol{\lambda}_t \equiv \phi_{\lambda_t}(\boldsymbol{\eta}_0)$ (we assume $\boldsymbol{\eta}_0 = \mathbf{0}$)

$$\text{Our NGD}: \quad \boldsymbol{\lambda}_{t+1} \leftarrow \phi_{\lambda_t}\left(\boldsymbol{\eta}_0 - \beta \hat{\mathbf{g}}^{(t)}_{\eta_0}\right), \quad \boldsymbol{\tau}_{t+1} \leftarrow \psi(\boldsymbol{\lambda}_{t+1})$$

where $\hat{\mathbf{g}}^{(t)}_{\eta_0}$ is the natural-gradient at $\boldsymbol{\eta}_0$ tied to $\boldsymbol{\lambda}_t$, which is computed by the chain rule,

$$\hat{\mathbf{g}}^{(t)}_{\eta_0} = \mathbf{F}_\eta(\boldsymbol{\eta}_0)^{-1}\left[\nabla_{\eta_0}\left[\psi \circ \phi_{\lambda_t}(\boldsymbol{\eta})\right]\ \nabla_{\tau_t}\mathcal{L}(\boldsymbol{\tau})\right], \quad \mathbf{F}_\eta(\boldsymbol{\eta}_0) \text{ is the (exact) FIM for } \boldsymbol{\eta}_0 \text{ tied to } \boldsymbol{\lambda}_t$$



Standard NGD is a special case of our NGD.
Gaussian case:

- $\mathbf{B}$: (closed) matrix Lie group
- $\mathbf{M}$: Lie sub-algebra

Full Gaussian q: with mean $\boldsymbol{\mu}$ and precision $\mathbf{S} = \boldsymbol{\Sigma}^{-1}$.
Global, Auxiliary, and Local parameterizations:

$$\boldsymbol{\tau} = \left\{\boldsymbol{\mu} \in \mathbb{R}^p, \mathbf{S} = \mathbf{S}\mathbf{S}^T \in \mathcal{S}^{p\times p}_{++}\right\}, \boldsymbol{\lambda} = \left\{\boldsymbol{\mu} \in \mathbb{R}^p, \mathbf{B} \in \mathrm{GL}^{p\times p}\right\}, \boldsymbol{\eta} = \left\{\boldsymbol{\delta} \in \mathbb{R}^p, \mathbf{M} \in \mathcal{S}^{p\times p}\right\}$$

Define $\mathbf{h}(\mathbf{M}) := \mathbf{I} + \mathbf{M} + \frac{1}{2}\mathbf{M}^2$. Maps ψ, ϕ_{λ_t}:

$$\begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{S}\end{Bmatrix} = \psi(\boldsymbol{\lambda}) := \begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\mathbf{B}^\top\end{Bmatrix}, \quad \begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\end{Bmatrix} = \phi_{\lambda_t}(\boldsymbol{\eta}) := \begin{Bmatrix}\boldsymbol{\mu} + \mathbf{B}_t^{-T}\boldsymbol{\delta}\\ \mathbf{B}_t\mathbf{h}(\mathbf{M})\end{Bmatrix}.$$

Tractable natural-gradient at $\boldsymbol{\eta}_0 = \{\boldsymbol{\delta}_0, \mathbf{M}_0\}$ tied to $\boldsymbol{\lambda}_t = \{\boldsymbol{\mu}_t, \mathbf{B}_t\}$:

$$\hat{\mathbf{g}}^{(t)}_{\eta_0} = \begin{pmatrix}\hat{\mathbf{g}}^{(t)}_{\delta_0}\\ \mathrm{vec}(\hat{\mathbf{g}}^{(t)}_{\mathbf{M}_0})\end{pmatrix} = \begin{pmatrix}\mathbf{I}_p & 0\\ 0 & 2\mathbf{I}_{p^2}\end{pmatrix}^{-1}\begin{bmatrix}\mathbf{B}_t^{-1}\mathbf{g}_{\mu_t}\\ \mathrm{vec}(-2\mathbf{B}_t^{-1}\mathbf{g}_{\boldsymbol{\Sigma}_t}\mathbf{B}_t^{-T})\end{bmatrix}$$

$$\mathbf{g}_\mu = \nabla_\mu \mathcal{L}(\boldsymbol{\tau}) = \mathbb{E}_q[\nabla_w \ell(\mathbf{w})], \quad \mathbf{g}_\Sigma = \nabla_{\mathcal{S}^{-1}}\mathcal{L}(\boldsymbol{\tau}) = \tfrac{1}{2}\mathbb{E}_q\left[\nabla^2_w \ell(\mathbf{w})\right] - \tfrac{\gamma}{2}\mathbf{S}$$

$$\text{Our NGD}: \quad \boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{\mu}_t - \beta \mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{g}_{\mu_t}, \quad \mathbf{B}_{t+1} \leftarrow \mathbf{B}_t\mathbf{h}\left(\beta \mathbf{B}_t^{-1}\mathbf{g}_{\Sigma_t}\mathbf{B}_t^{-T}\right)$$

Lemma 1: Update on $\mathbf{S}_{t+1} = \mathbf{B}_{t+1}\mathbf{B}_{t+1}^T$

$$\mathbf{S}_{t+1} = \underbrace{\overbrace{(1-\beta\gamma)\mathbf{S}_t + \beta\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\nabla^2_w \ell(\mathbf{w})\right]}^{\text{standard NGD on } \mathbf{S}} + \frac{\beta^2}{2}\mathbf{G}_t\mathbf{S}_t^{-1}\mathbf{G}_t}_{\text{RGD with retraction}} + O(\beta^3)$$

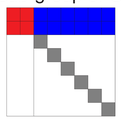
$$\mathbf{G}_t = \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\nabla^2_w \ell(\mathbf{w})\right] - \gamma\mathbf{S}_t$$

4 NGD for Gaussian with Flexible Covariance Structures

Notations:
$\mathrm{GL}^{p\times p}$: Invertible Matrices (General Linear Group), $\mathcal{D}^{p\times p}$: Diagonal Matrices, $\mathcal{D}^{p\times p}_{++}$: Diagonal Invertible Matrices (Diagonal Matrix Group).
$\mathcal{S}^{p\times p}_{++}$: Positive-definite Matrices, $\mathcal{S}^{p\times p}$: Symmetric Matrices.

Example : Auxiliary space: $\mathcal{B}_{\mathrm{up}}(k)$, a block upper-triangular sub-group of $\mathrm{GL}^{p\times p}$; $d_0 := p - k$.

$$\mathcal{B}_{\mathrm{up}}(k) := \left\{\begin{bmatrix}\mathbf{B}_A & \mathbf{B}_B\\ \mathbf{0} & \mathbf{B}_D\end{bmatrix} \middle| \mathbf{B}_A \in \mathrm{GL}^{k\times k}, \mathbf{B}_D \in \mathcal{D}^{d_0\times d_0}_{++}\right\},$$



When $k = 0$, $\mathcal{B}_{\mathrm{up}}(k) = \mathcal{D}^{p\times p}_{++}$. When $k = p$, $\mathcal{B}_{\mathrm{up}}(k) = \mathrm{GL}^{p\times p}$.
Local space: $\mathcal{M}_{\mathrm{up}}(k)$.

$$\mathcal{M}_{\mathrm{up}}(k) := \left\{\begin{bmatrix}\mathbf{M}_A & \mathbf{M}_B\\ \mathbf{0} & \mathbf{M}_D\end{bmatrix} \middle| \mathbf{M}_A \in \mathcal{S}^{k\times k}, \mathbf{M}_D \in \mathcal{D}^{d_0\times d_0}\right\}$$

Global, Auxiliary, and Local parameterizations:

$$\boldsymbol{\tau} = \left\{\boldsymbol{\mu} \in \mathbb{R}^p, \mathbf{S} = \mathbf{B}\mathbf{B}^T \in \mathcal{S}^{p\times p}_{++} | \mathbf{B} \in \mathcal{B}_{\mathrm{up}}(k)\right\}, \boldsymbol{\lambda} = \left\{\boldsymbol{\mu} \in \mathbb{R}^p, \mathbf{B} \in \mathcal{B}_{\mathrm{up}}(k)\right\}, \boldsymbol{\eta} = \left\{\boldsymbol{\delta} \in \mathbb{R}^p, \mathbf{M} \in \mathcal{M}_{\mathrm{up}}(k)\right\}.$$

Maps ψ, ϕ_{λ_t}: see Block 3. $\kappa_{\mathrm{up}}(\mathbf{X})$ extracts non-zero entries of $\mathcal{M}_{\mathrm{up}}(k)$ from $\mathbf{X}$.

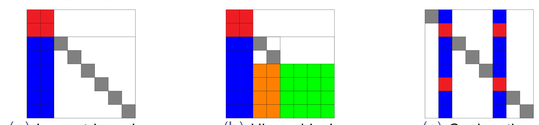
$$\text{Structured NGD}: \quad \boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{\mu}_t - \beta\mathbf{S}_t^{-1}\mathbf{g}_{\mu_t}, \quad \mathbf{B}_{t+1} \leftarrow \underbrace{\mathbf{B}_t\mathbf{h}\left(\beta\mathbf{C}_{\mathrm{up}} \odot \kappa_{\mathrm{up}}(2\mathbf{B}_t^{-1}\mathbf{g}_{\Sigma_t}\mathbf{B}_t^{-T})\right)}_{\text{structure-preserving}}$$

where $\odot$ is the elementwise product , $\kappa_{\mathrm{up}}(\mathbf{X}) \in \mathcal{M}_{\mathrm{up}}(k)$ and $\mathbf{C}_{\mathrm{up}} = \begin{bmatrix}\frac{1}{2}\mathbb{1} & \mathbb{1}\\ \mathbf{0} & \frac{1}{2}\mathbf{I}_D\end{bmatrix} \in \mathcal{M}_{\mathrm{up}}(k)$.

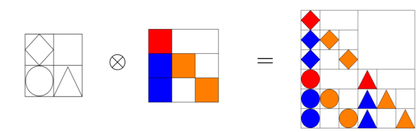
Lemma 2: Property of $\mathcal{M}_{\mathrm{up}}(k)$
$\mathbf{h}(\mathbf{M}) \in \mathcal{B}_{\mathrm{up}}(k)$ for any $\mathbf{M} \in \mathcal{M}_{\mathrm{up}}(k)$.

Our method: systemic approach to incorporate structures
obtain exact, non-singular FIMs

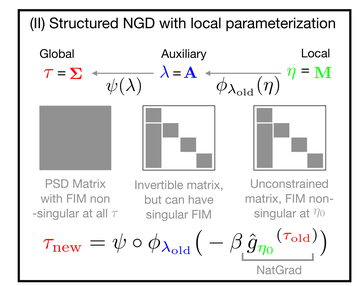
More flexible groups:



(a) Lower-triangular (b) Hierarchical (c) Conjugation



A Kronecker product group



5 Applications

Structured 2nd-order Methods for Non-convex Opt

Given an optimization problem

$$\min_{\mu\in\mathbb{R}^p} \ell(\boldsymbol{\mu}), \tag{1}$$

we formulate a new problem over Gaussian $q(\mathbf{w}|\boldsymbol{\tau})$.

$$\min_{\tau\in\Omega_\tau} \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}[\ell(\mathbf{w})] + \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}[\log q(\mathbf{w}|\boldsymbol{\tau})], \tag{2}$$

Use our structured NGD to solve (2) as

$$\boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{\mu}_t - \beta\mathbf{S}_t^{-1}\mathbf{g}_{\mu_t}, \quad \mathbf{B}_{t+1} \leftarrow \mathbf{B}_t\mathbf{h}\left(\beta\mathbf{C}_{\mathrm{up}} \odot \kappa_{\mathrm{up}}(2\mathbf{B}_t^{-1}\mathbf{g}_{\Sigma_t}\mathbf{B}_t^{-T})\right)$$

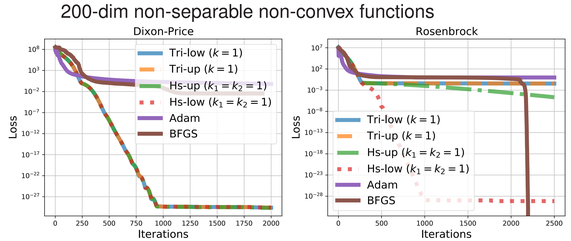
Obtain an update to solve (1) with a structural invariance:

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t - \beta\mathbf{S}_t^{-1}\nabla_{\mu_t}\ell(\boldsymbol{\mu}),\\ \mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t\mathbf{h}\left[\beta\mathbf{C}_{\mathrm{up}} \odot \kappa_{\mathrm{up}}\left(\mathbf{B}_t^{-1}\nabla^2_{\mu_t}\ell(\boldsymbol{\mu})\mathbf{B}_t^{-T} - \mathbf{I}\right)\right]\end{aligned} \tag{3}$$

by using $\mathbf{g}_{\mu_t} \approx \nabla_{\mu_t}\ell(\boldsymbol{\mu})$ and $\mathbf{g}_{\Sigma_t} \approx \frac{1}{2}\left[\nabla^2_{\mu_t}\ell(\boldsymbol{\mu}) - \mathbf{S}_t\right]$. (4)

Time complexity:
$O(k^2p)$ for triangular, $O((k_1^2 + k_2^2)p)$ for Heisenberg.
Implementation: Hessian-vector products (HVPs);
Off-diagonal: k HVPs (triangular), $k_1 + k_2$ HVPs (Heisenberg);
Diagonal: compute/approximate diagonal entries of $\nabla^2_{\mu_t}\ell(\boldsymbol{\mu})$

200-dim non-separable non-convex functions



Structured Adaptive-gradient Methods for DL

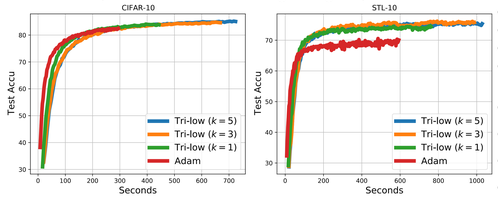
A block-diagonal Gaussian $q(\mathbf{w}|\boldsymbol{\tau}) = \prod_i q(\mathbf{w}^i|\boldsymbol{\mu}^i, \mathbf{S}^i)$.
Kronecker product group: i-th layer, $q(\mathbf{w}^i|\boldsymbol{\mu}^i, \mathbf{S}^i)$

Adaptive-gradient updates with a structural invariance: approximating $\nabla^2_{\mu_t}\ell(\boldsymbol{\mu})$ by the Gauss-Newton in (4).

Time complexity:
$O(kp)$ for triangular, $O((k_1 + k_2)p)$ for Heisenberg.

Easy to parallelize, no sequential conjugate-grad steps

Kronecker product of lower-triangular groups for a CNN



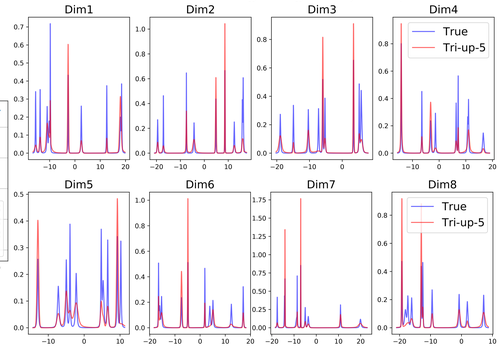
Variational Inference (VI) with Gaussian Mixtures

Structured multimodal mixture: $q(\mathbf{w}|\boldsymbol{\tau}) = \frac{1}{K}\sum_{k=1}^K q(\mathbf{w}|\boldsymbol{\mu}_k, \mathbf{S}_k)$

NGVI for structured Gaussian mixtures with MC gradients

First 8 marginal distributions of a mixture approximation to a 80-dim mixture of Student's t distributions:

Block upper-triangular group ($k = 5$)



Group Structural Invariance

Lemma 3: Group Structural Invariance
If $\mathbf{B} \in \mathcal{B}_{\mathrm{up}}(k)$, the update in (3) is invariant under any (group) transform $\mathbf{R}^T \in \mathcal{B}_{\mathrm{up}}(k)$ of (1) such as $\min_{y\in\mathbb{R}^p}\ell(\mathbf{R}\,\mathbf{y})$.