

Tractable structured natural-gradient descent using local parameterizations

Wu Lin (UBC)

Joint work with Frank Nielsen (Sony CSL), Emtiyaz Khan (AIP, RIKEN), Mark Schmidt (UBC & Amii)

Motivation

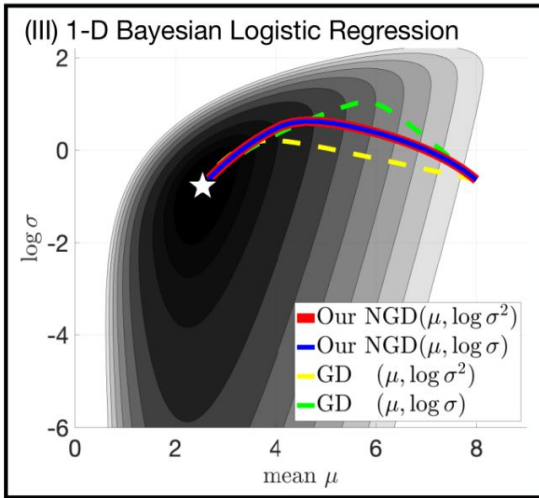
GD: dependent on parameterizations

NGD: less dependent on parameterizations

$$\text{NGD} : \tau \leftarrow \tau - (\mathbf{F}(\tau))^{-1} \mathbf{g}_\tau$$

Unstructured Fisher matrix $\mathbf{F}(\tau)$: high iteration cost for $\mathbf{F}^{-1}(\tau)$

Challenges: how to incorporate structures (e.g. low-rank) ?

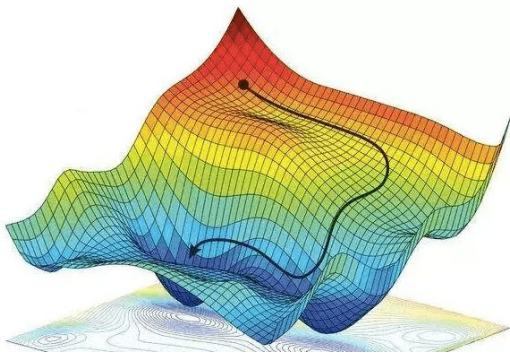


Tractable NGD for structured $F(\tau)$

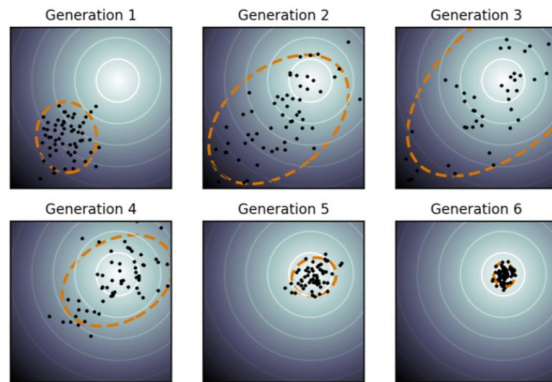
Contributions: by using local-parameter coordinates
simple structure-preserving NG updates

- Generalizes the exponential natural evolutionary-strategy
- Recovers existing Newton-like methods
- Gives new algorithms to learn structured covariances
- Obtains new efficient structured 2nd-order methods

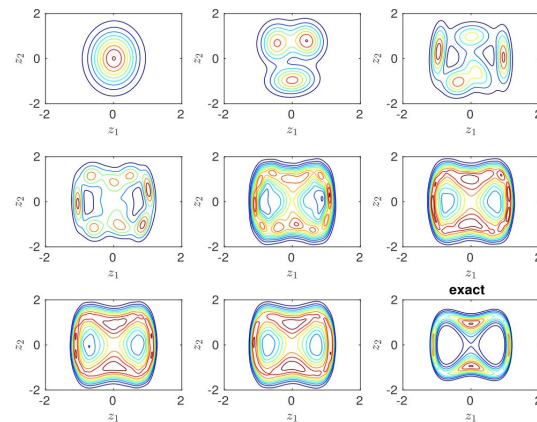
Many machine learning applications are optimization, search, inference problems.



Source: Google Images



Source: Wikipedia



Optimization
(gradient/Hessian)

$$\ell(\mathbf{w}^*) \leq \min_{q \in \Omega} E_{q(w)}[\ell(\mathbf{w})]$$

Search
(gradient-free)

$$\min_{q \in \Omega} E_{q(w)}[\ell(\mathbf{w})]$$

Inference
(density)

$$\min_{q \in \Omega} E_{q(w)}[\ell(\mathbf{w})] + E_{q(w)}[\log q(\mathbf{w})]$$

can be solved in one framework: minimization over parametric distribution q via **NGD**

e.g., minimization over a Gaussian family q

Existing works on structural Gaussian-covariances $q(\mathbf{w}|\tau)$

Issues:

- Complicated natural-gradient computations
- Singular Fisher matrices $\mathbf{F}(\tau)$ for arbitrary structures
- Case-by-case derivations
- Ad-hoc approximations for complexity reductions and singular $\mathbf{F}(\tau)$

(I) Parameterization with singular FIM

$\Sigma = \mathbf{u}\mathbf{u}^T + \text{Diag}(\mathbf{d}^2)$

The diagram illustrates the decomposition of a symmetric positive semi-definite (PSD) matrix Σ into a rank-1 matrix $\mathbf{u}\mathbf{u}^T$ and a diagonal matrix $\text{Diag}(\mathbf{d}^2)$. The PSD matrix is represented by a solid grey square. The rank-1 matrix is represented by a vertical grey bar and a horizontal grey bar. The diagonal matrix is represented by a square with grey squares along its main diagonal.

FIM at $\mathbf{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

$\mathbf{d} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

0.5	0	0	0.5	0	0
0	2	0	0	0	0
0	0	2	0	0	0
0.5	0	0	0.5	0	0
0	0	0	0	0.5	0
0	0	0	0	0	0.5

$$\Sigma = \mathbf{u}\mathbf{u}^T + \text{Diag}(\mathbf{d}^2)$$

Our approach

A class of structured matrix groups \mathbf{A}

$$\Sigma = \mathbf{A}\mathbf{A}^T$$

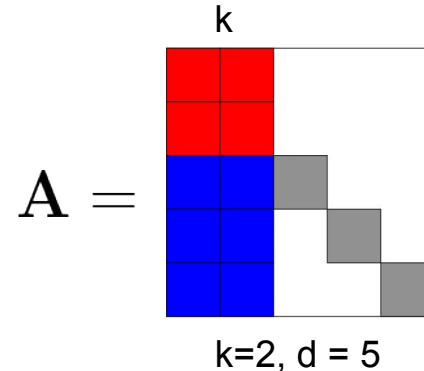
$$\mathbf{A} = \mathbf{A}_{\text{old}} \text{Exp}(\mathbf{M})$$

$$\mathbf{A}_{\text{old}} = \mathbf{A}_{\text{old}} \text{Exp}(\mathbf{0})$$

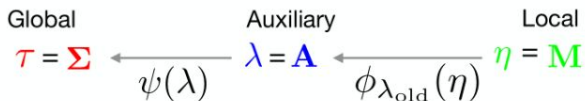
Preserves group structures in \mathbf{A}

Performs NGD in \mathbf{M}

Recovers standard NGDs as special cases



(II) Structured NGD with local parameterization



PSD Matrix
with FIM non
-singular at all τ



Invertible matrix,
but can have
singular FIM



Unconstrained
matrix, FIM non-
singular at η_0

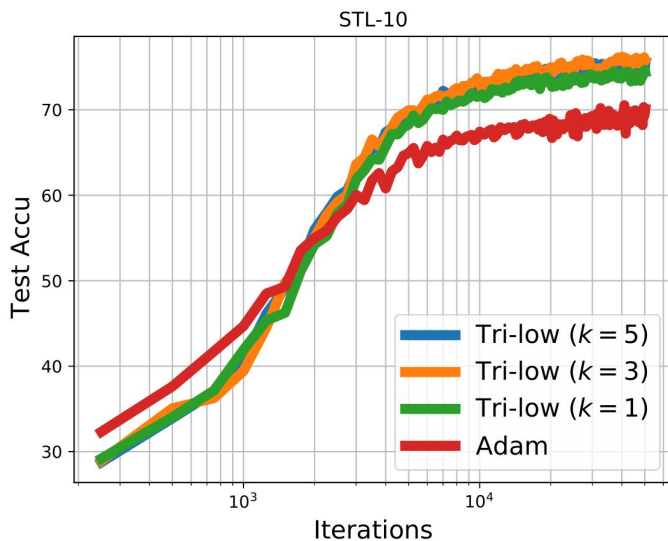
$$\tau_{\text{new}} = \psi \circ \phi_{\lambda_{\text{old}}} \left(- \underbrace{\beta \hat{g}_{\eta_0}(\tau_{\text{old}})}_{\text{NatGrad}} \right)$$

Structured second-order methods:

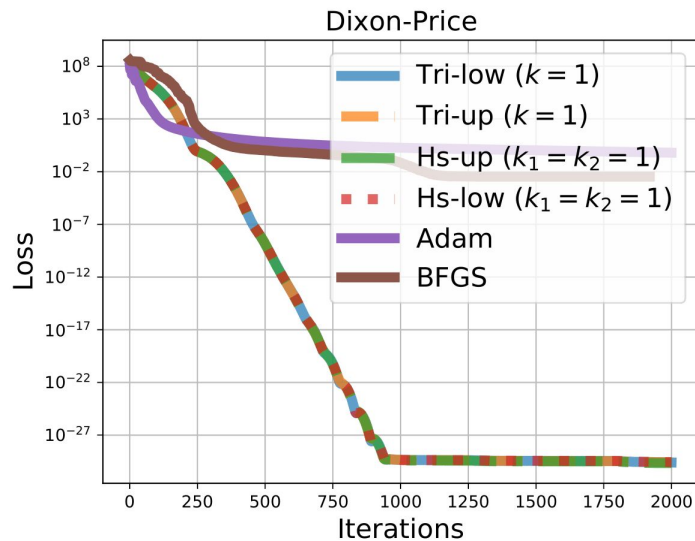
$$\text{Hessian } \mathbf{H} \approx \mathbf{A}\mathbf{A}^T$$

Our contributions

- A systematic approach to incorporate group structures
- Non-singular and closed-form FIMs for Gaussians and Wisharts
- Efficient and simple NG updates for many groups
- Structured 2nd-order methods for unconstrained optimization
- Structured adaptive algorithms for NN with a linear iteration cost



NN training



Structured 2nd-order optimization