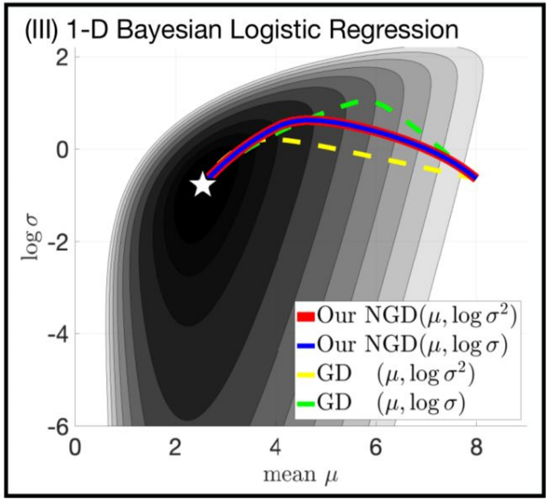
Tractable structured natural-gradient descent using local parameterizations

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

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

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

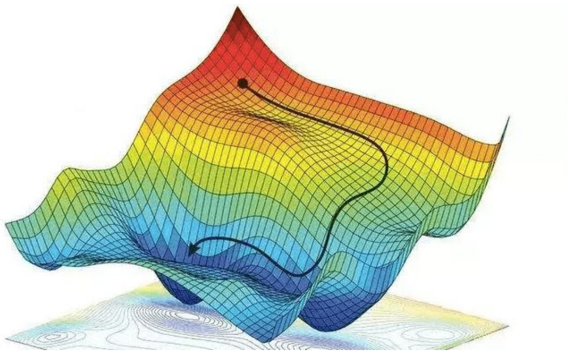
Challenges: how to incorporate structures (e.g. low-rank) ?

Tractable NGD for structured $\mathbf{F}(\tau)$

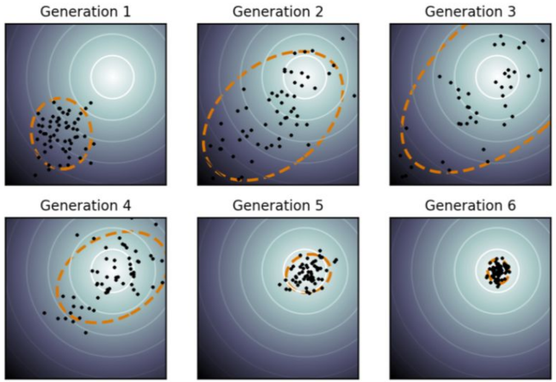
Contributions: by using local-parameter coordinates simple structure-preserving NG updates

- 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

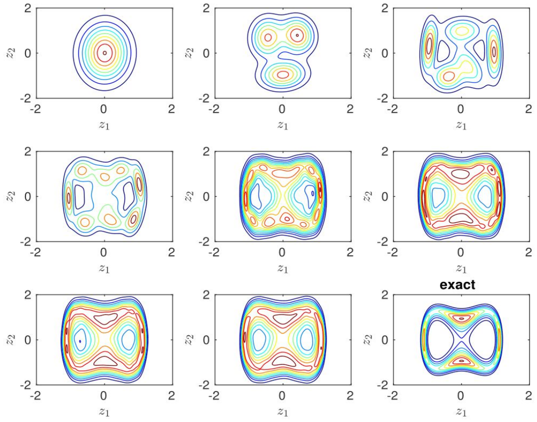
Many machine learning applications are optimization, search, inference problems.



Source: Google Images



Source: Wikipedia



Optimization (gradient/Hessian)

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

Search (gradient-free)

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

Inference (density)

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

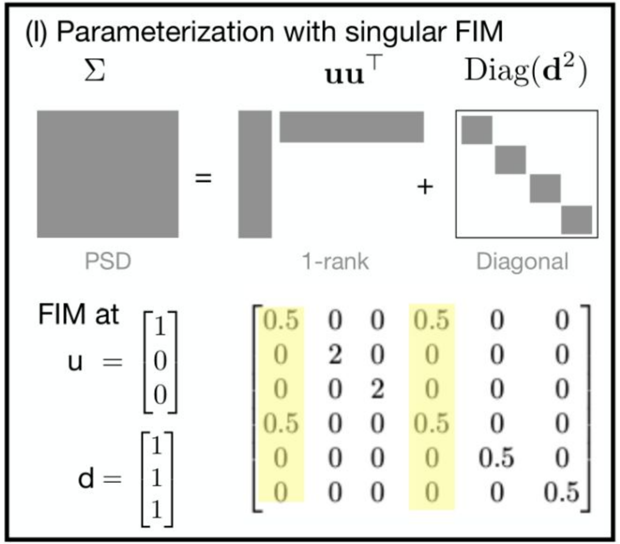
can be solved in one framework: minimization over parametric distribution q via **NGD**

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 $\mathbf{F}(\tau)$



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

Our approach

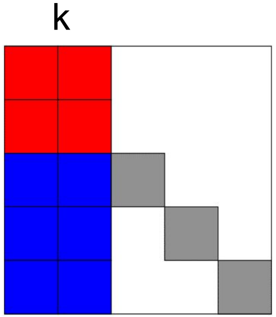
A class of structured matrix groups $\mathbf{A}$

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

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

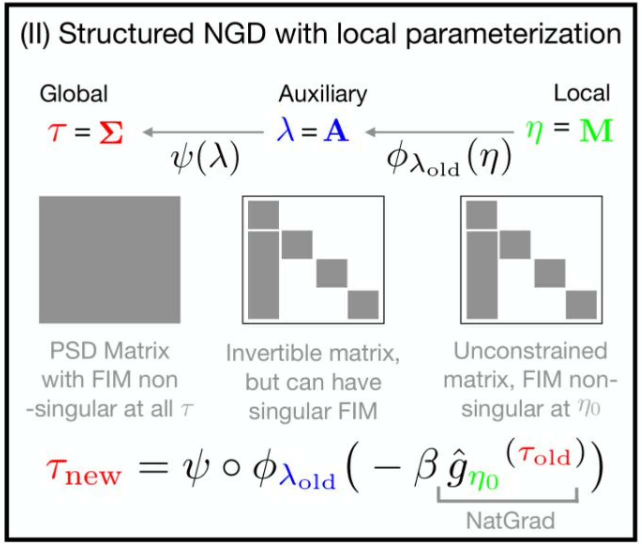
$$\mathbf{A}_{\text{old}} = \mathbf{A}_{\text{old}}\text{Exp}(\mathbf{0})$$

$\mathbf{A} =$



k=2, d = 5

Preserves group structures in A
Performs NGD in M
Recovers standard NGDs as special cases

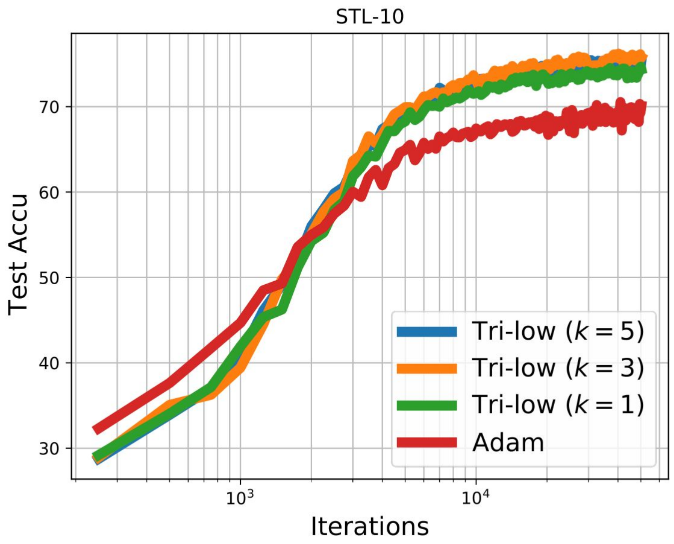


Structured second-order methods:

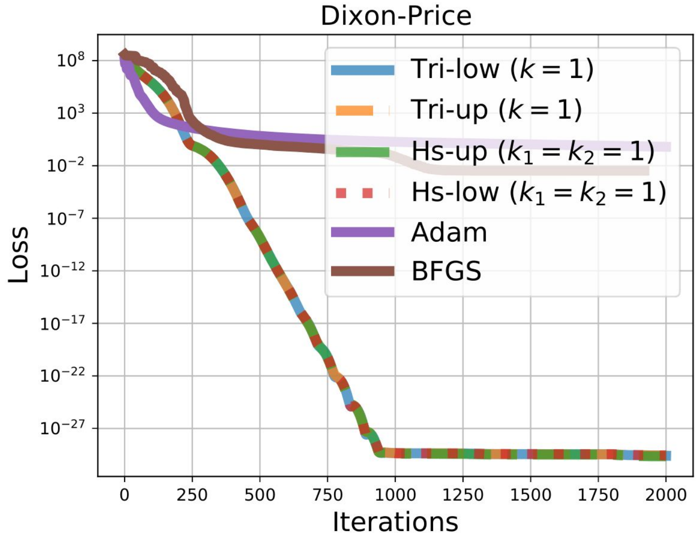
$$\text{Hessian } \mathbf{H} \approx \mathbf{A}\mathbf{A}^T$$

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