A Bregman Extension of quasi-Newton Updates

Takafumi Kanamori[1] Atsumi Ohara[2]

[1]Nagoya university

[2]Osaka university

情報幾何関連分野研究会 2010

–情報工学への幾何学的アプローチ–

Plan of Presentation

1. quasi-Newton method
   - nonlinear optimization problem
   - Hessian update formula
   - variational view of quasi-Newton update
2. Bregman extension of quasi-Newton method
   - Bregman divergence with V-potential on $\mathbf{PD}(n)$
   - dual structure defined by Bregman divergence
3. Exploiting Sparsity of Hessian matrix
   - relation to U-boost and em-algorithm
4. Other topics
   - convergence property
   - invariance under group action
   - robustness against numerical errors
5. Concluding Remarks

— quasi-Newton method —

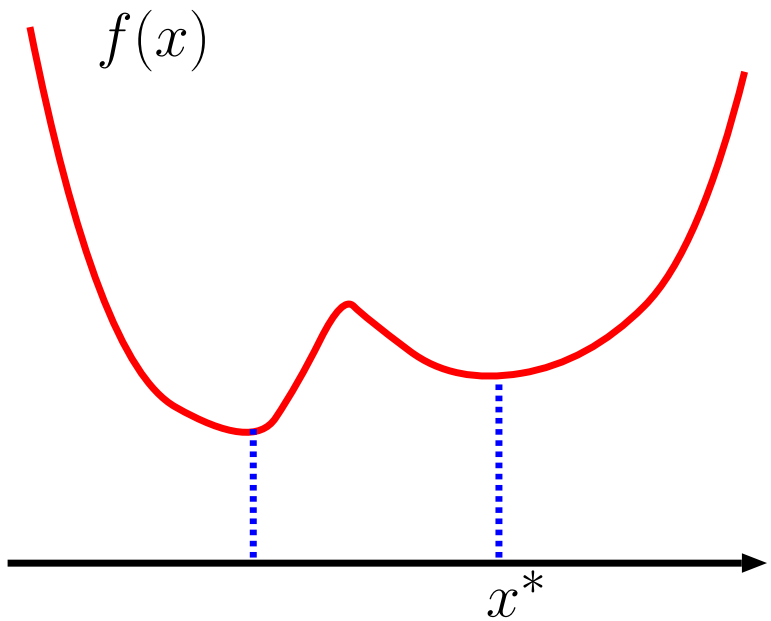
keywords: Hessian update, secant condition

Unconstrained nonlinear optimization problem

制約なし最適化問題： 目的関数 $f \in C^2(\mathbf{R}^n)$

$$\min_x \ f(x), \qquad x \in \mathbf{R}^n$$

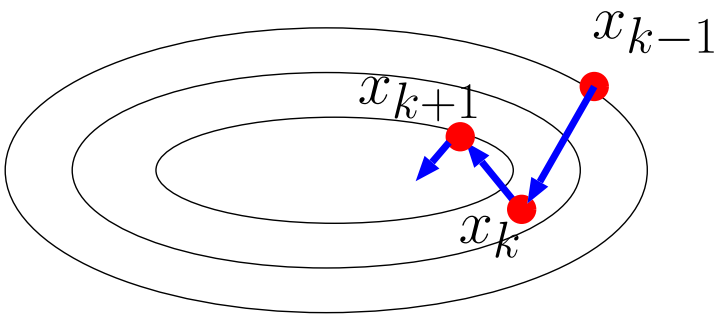
局所解 x^* を数値的に求める



Numerical algorithm

- 初期値 x_0, $B_0 : n \times n$ matrix, $\alpha_0 \geq 0$. 以下を繰り返す

$$x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k), \qquad k = 0, 1, 2, \ldots$$



- 各 x_k のまわりで 2 次近似

$$f(x) \cong f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2}(x - x_k)^\top \nabla^2 f(x_k)(x - x_k)$$

$\longrightarrow$ 2 次近似した関数を小さくする x を計算

  - 最急降下法：$B_k = I$, α_k: line search
  - ニュートン法：$B_k = \nabla^2 f(x_k)$, $\alpha_k = 1$
  - 準ニュートン法：$B_k = \nabla^2 f(x_k)$ の近似, α_k: line search

quasi-Newton method

ヘシアン行列 $\nabla^2 f(x_{k+1})$ の近似 B_{k+1} の構成

$$x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k), \qquad \alpha_k \geq 0$$

標準的な記法： $s = x_{k+1} - x_k \in \mathbf{R}^n, \quad y = \nabla f(x_{k+1}) - \nabla f(x_k) \in \mathbf{R}^n$

(正確には s_k, y_k と添字を付ける)

目標： B_k, s, y から B_{k+1} を作る.

BFGS update：

$$B_{k+1} = B^{BFGS}[B_k] := B_k - \frac{B_k s s^\top B_k}{s^\top B_k s} + \frac{y y^\top}{s^\top y}$$

DFP update：

$$B_{k+1} = B^{DFP}[B_k] := B_k - \frac{B_k s y^\top + y s^\top B_k}{s^\top y} + s^\top B_k s \frac{y y^\top}{(s^\top y)^2} + \frac{y y^\top}{s^\top y}$$

Some requirements for update formula

- セカント条件： $B_{k+1}s = y$

$$\nabla^2 f(x_{k+1})(x_{k+1} - x_k) \approx \nabla f(x_{k+1}) - \nabla f(x_k), \quad B_{k+1} \approx \nabla^2 f(x_{k+1})$$

- 正定値性の継承：

$$B_k \in \mathrm{PD}(n), \quad s^\mathsf{T} y > 0 \implies B_{k+1} \in \mathrm{PD}(n).$$

  - B_{k+1} が正定値 $\implies -B_{k+1}\nabla f(x_{k+1})$ が降下方向
  - 条件 $s^\mathsf{T} y > 0$ ： 係数 α_k を定める直線探索がある程度正確.

BFGS, DFP は上の条件を満たす

A Variational View of quasi-Newton updates [Fletcher '91]

$\boldsymbol{B_{k+1}}$：セカント条件を満たす行列の中で $\boldsymbol{B_k}$ に最も「近い」

Matrix nearness: KL-divergence between $\boldsymbol{N(0, P)}$ and $\boldsymbol{N(0, Q)}$

$$\mathbf{KL}(\boldsymbol{P}, \boldsymbol{Q}) := \langle \boldsymbol{P}, \boldsymbol{Q}^{-1} \rangle - \log\det(\boldsymbol{P}\boldsymbol{Q}^{-1}) - \boldsymbol{n}, \quad \boldsymbol{P}, \boldsymbol{Q} \in \mathbf{PD}(\boldsymbol{n})$$

note: $\mathbf{KL}(\boldsymbol{P}, \boldsymbol{Q}) = \mathbf{KL}(\boldsymbol{Q}^{-1}, \boldsymbol{P}^{-1})$

(BFGS update) $\min_{\boldsymbol{B} \in \mathbf{PD}(\boldsymbol{n})} \mathbf{KL}(\boldsymbol{B}, \boldsymbol{B_k})$ subject to $\boldsymbol{B}\boldsymbol{s} = \boldsymbol{y}$

(DFP update) $\min_{\boldsymbol{B} \in \mathbf{PD}(\boldsymbol{n})} \mathbf{KL}(\boldsymbol{B_k}, \boldsymbol{B}),$ subject to $\boldsymbol{B}\boldsymbol{s} = \boldsymbol{y}$

— Bregman extension of quasi-Newton method —

keywords: Bregman divergence, projection

quasi-Newton updates with Bregman divergence

Bregman divergence: An extension of KL-divergence

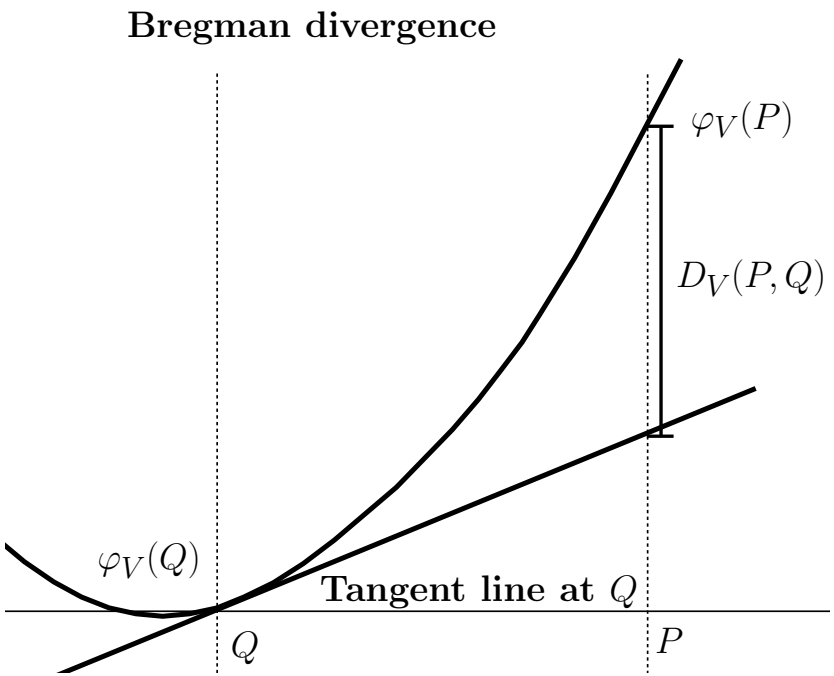
準ニュートン更新則の Bregman 拡張

$\mathbf{PD}(n)$ 上の Bregman divergence $\Longrightarrow$ { 幾何構造, 行列の更新則

Bregman divergence with V-potential [Ohara and Eguchi, '05]

V-potential : $\varphi_V(P) = V(\det P), \quad V : \mathbf{R}_+ \to \mathbf{R}$

Bregman div. : $D_V(P,Q) = \varphi_V(P) - \{\varphi_V(Q) + \langle \nabla\varphi_V(Q), P - Q\rangle\}$



$V(z) = -\log z \implies$ KL-divergence,
i.e. BFGS or DFP update.

$\varphi_V(P)$ is strictly convex $\iff$
$\nu(z) := -zV'(z) > 0$,
$\beta(z) := z\nu'(z)/\nu(z) < 1/n$

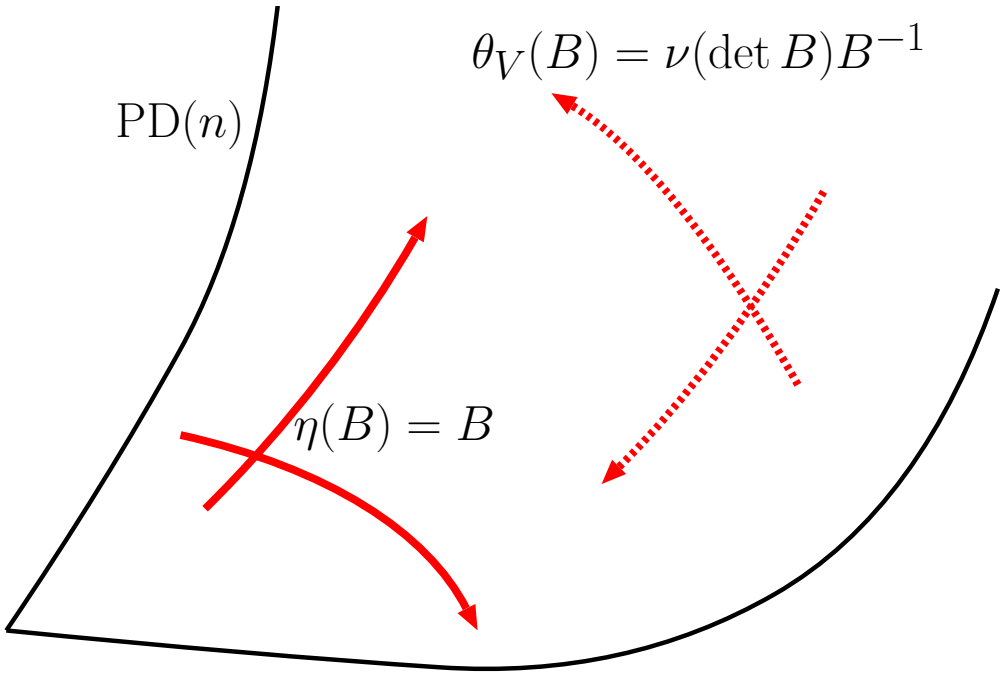
Information Geometry on $\mathrm{PD}(n)$ [Ohara and Eguchi, '05] -1/3-

two coordinate systems

PD(n) 上の ２つの座標系： for $B \in \mathbf{PD}(n)$

$$B \to \eta(B) = B, \qquad B \to \theta_V(B) = \nu(\det(B))B^{-1} \quad \text{(one-to-one)}$$

例：$V(z) = -\log z \Longrightarrow \theta(B) = B^{-1}$



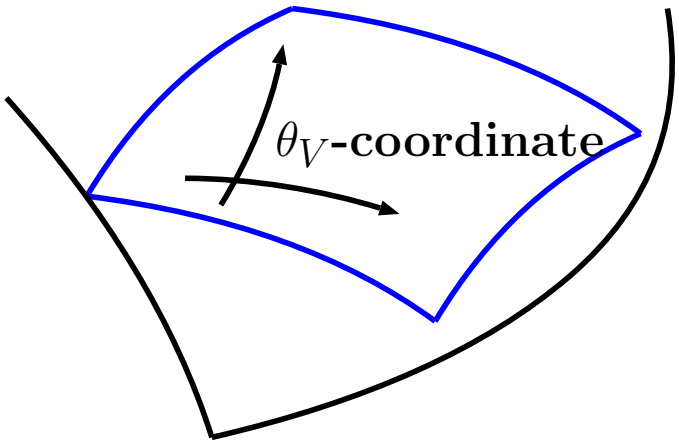
Autoparallel Submanifold

$\mathcal{M} \subset \mathbf{PD}(n)$: submanifold

- $\mathcal{M}$ が η-座標に関して affine 平面 $\Longrightarrow$ η-autoparallel
- $\mathcal{M}$ が θ_V-座標に関して affine 平面 $\Longrightarrow$ θ_V-autoparallel

  ($V(z) = -\log z$ のとき θ-autoparallel)

$$\mathcal{M} = \{B \in \mathrm{PD}(n) \mid \langle A, \theta_V(B) \rangle = c\} \quad (\theta_V\text{-autoparallel})$$



例：secant condition. η, θ-autoparallel (doubly autoparallel).

$$\mathcal{M} = \{B \in \mathbf{PD}(n) \mid \eta(B)s = y\} = \{B \in \mathbf{PD}(n) \mid s = \theta(B)y\}$$

$(\eta(B) = B,\ \theta(B) = B^{-1})$

Projection onto autoparallel submanifold

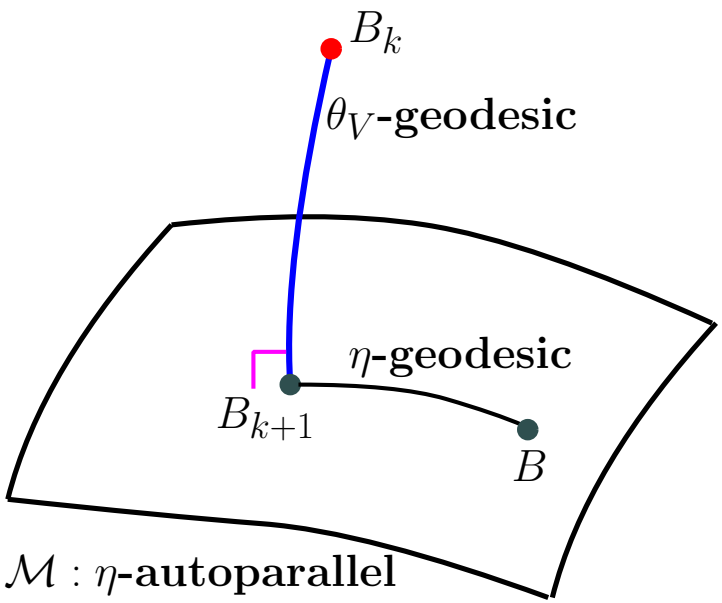
- $\mathcal{M}$: η-autoparallel submanifold in $\mathbf{PD}(n)$

B_k の $\mathcal{M}$ への θ_V-projection : $B_{k+1} = \underset{B}{\operatorname{argmin}}\, D_V(B, B_k),\ B \in \mathcal{M},$

B_k, B_{k+1} を結ぶ θ_V-geodesic と $\mathcal{M}$ は直交

$$\Longleftrightarrow \quad \langle \theta_V(B_k) - \theta_V(B_{k+1}),\ \eta(B) - \eta(B_{k+1}) \rangle = 0, \quad {}^{\forall} B \in \mathcal{M}$$

($\eta \longleftrightarrow \theta_V$ で同様の関係)

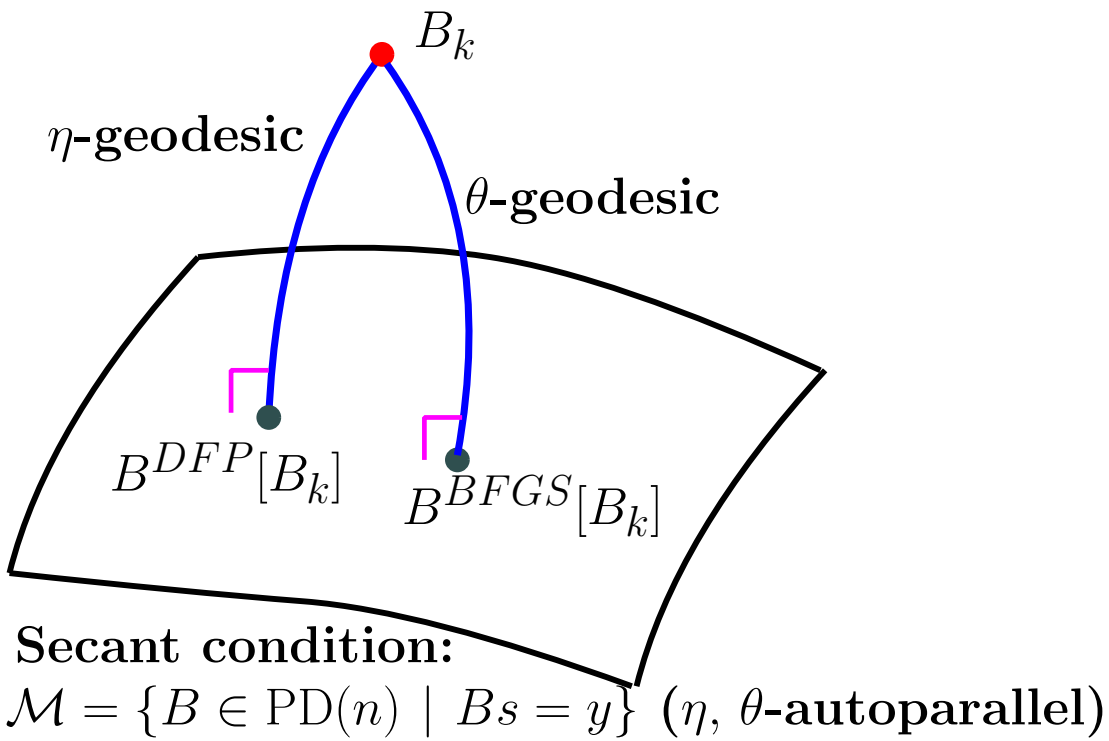


For all $B \in \mathcal{M}$,

$$D_V(B, B_k) = D_V(B, B_{k+1}) + D_V(B_{k+1}, B_k)$$

A Variational View of quasi-Newton updates [Fletcher '91] (revisited)

- $\mathcal{M}$: doubly autoparallel
  - BFGS: η-autoparallel $\mathcal{M}$ への $\boldsymbol{B_k}$ の θ-projection
  - DFP: θ-autoparallel $\mathcal{M}$ への $\boldsymbol{B_k}$ の η-projection

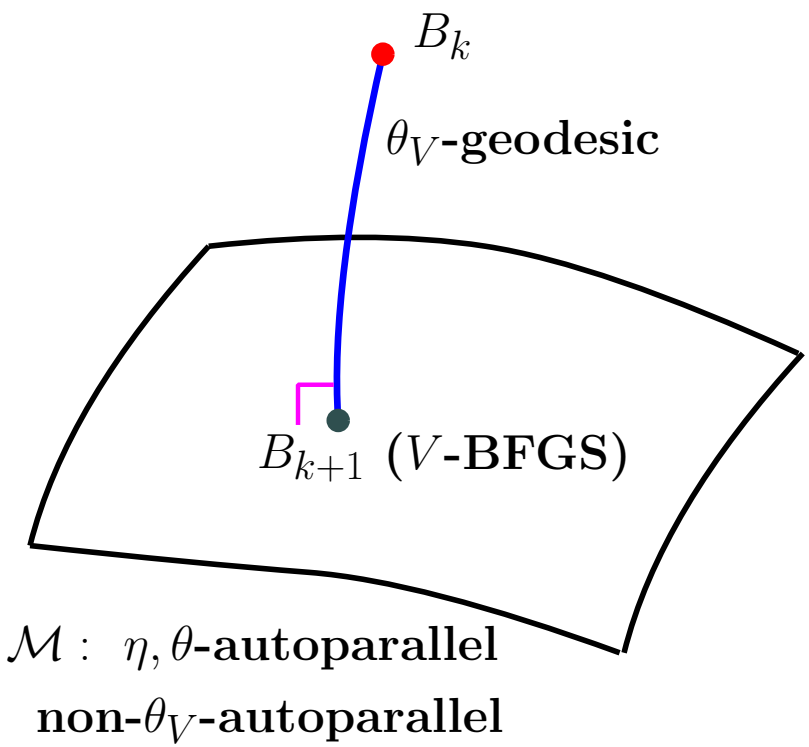


実用上は BFGS が数値的に良いとされている

V-extension of quasi-Newton update

V-BFGS update: $\min_{B \in \mathrm{PD}(n)} D_V(B, B_k)$ subject to $B \in \mathcal{M}$ (i.e. $Bs = y$)

V-DFP update: $\min_{B \in \mathrm{PD}(n)} D_V(B^{-1}, B_k^{-1})$ subject to $B \in \mathcal{M}$



note: $\min_B D_V(B_k, B)$ subject to $B \in \mathcal{M}$ may not be convex problem.

V-BFGS update formula

V-BFGS update

$$B_{k+1} = \frac{\nu(\det B_{k+1})}{\nu(\det B_k)} \cdot B^{BFGS}[B_k] + \left(1 - \frac{\nu(\det B_{k+1})}{\nu(\det B_k)}\right) \cdot \frac{yy^\top}{s^\top y},$$

where $\nu(z) = -zV'(z) > 0$.

- KL-div. $(V(z) = -\log(z)) \implies \nu(z) = 1,\ B_{k+1} = B^{BFGS}[B_k]$
- $B_k \in \mathrm{PD}(n),\ s^\top y > 0,\ \nu > 0 \implies B_{k+1} \in \mathrm{PD}(n)$
- BFGS update と (ほとんど) 同じ計算量
- Self-scaling quasi-Newton:

$$B_{k+1} = \phi_k \cdot B^{BFGS}[B_k] + (1 - \phi_k) \cdot \frac{yy^\top}{s^\top y}$$

V-BFGS では ϕ_k が V-potential から決まる.

V-BFGS update:

Initialization: Let $L_0 L_0^\top = B_0$ be the Cholesky decomposition of B_0, and $x_0 \in \mathbf{R}^n$ be an initial point. Set $k = 0$.

Repeat: If stopping criterion is satisfied, go to Output.

1. Let $x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k)$ with an appropriate $\alpha_k \geq 0$. The Cholesky decomposition $B_k = L_k L_k^\top$ is available to compute $B_k^{-1} \nabla f(x_k)$.
2. Update L_k to $\widetilde{L}$ which is the Cholesky decomposition of $B^{BFGS}[B_k; s_k, y_k]$. The Cholesky decomposition for rank-one update is available.
3. Compute $C = (\det \widetilde{L})^2 / \nu((\det L_k)^2)^{n-1}$ and find the root z^* of the equation $C \cdot \nu(z)^{n-1} = z,\ z > 0$.
4. Compute the Cholesky decomposition L_{k+1} such that

$$L_{k+1} L_{k+1}^\top = \frac{\nu(z^*)}{\nu((\det L_k)^2)} \widetilde{L}\widetilde{L}^\top + \left(1 - \frac{\nu(z^*)}{\nu((\det L_k)^2)}\right) \frac{y_k y_k^\top}{s_k^\top y_k}.$$

5. $k \leftarrow k + 1$.

Output: Local optimal solution x_k.

— Exploiting Sparsity of Hessian matrix —

keywords: iterative Bregman projection

Sparsity of Hessian matrix

適当な x について

$$F \supset \{(i,j) \mid (\nabla^2 f(x))_{ij} \neq 0\},$$
$$\mathcal{S} := \{B \in \mathrm{PD}(n) \mid B_{ij} = 0,\ (i,j) \in F^c\}$$

$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & & \\ * & * & * & & \\ * & & & * & \\ * & & & & * \end{pmatrix}$$

- 設定
  - x の次元 n が大きい
  - F の要素数は少ない
- 計算量を減らすために疎行列 $B_k \in \mathcal{S}$ を利用
- $\mathcal{S}$ は η-autoparallel
  $\Longrightarrow$ 情報幾何的にアルゴリズムを考察

Sparse quasi-Newton [Yamashita, '08]

$$\mathcal{M} = \{B \in \mathrm{PD}(n) \mid Bs = y\} \quad (\eta, \theta\text{-autoparallel})$$

$$\mathcal{S} = \{B \in \mathrm{PD}(n) \mid B_{ij} = 0,\ (i,j) \in F^c\} \quad (\eta\text{-autoparallel})$$

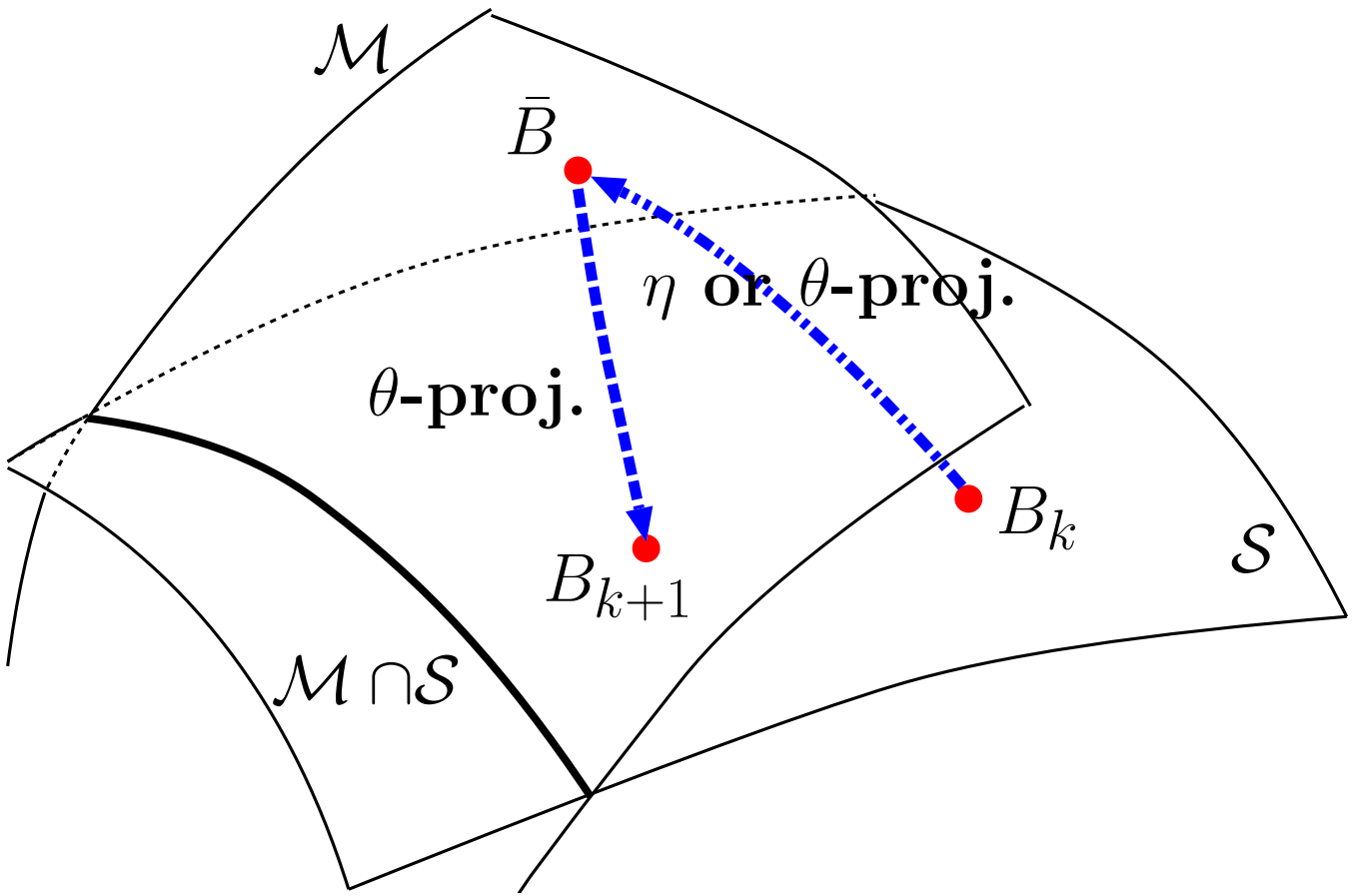
- Hessian update：$B_k \in \mathcal{S} \longrightarrow B_{k+1} \in \mathcal{S}$
- $B_{k+1} \in \mathcal{M} \cap \mathcal{S}$ (η-autoparallel) が理想的

計算が大変なので近似

1. B_k から $\bar{B} = B^{BFGS}[B_k]$ or $\bar{B} = B^{DFP}[B_k]$ を計算
   - $B^{BFGS}[B_k]$ は B_k の $\mathcal{M}$ への θ-projection
   - $B^{DFP}[B_k]$ は B_k の $\mathcal{M}$ への η-projection
2. $\bar{B}$ を $\mathcal{S}$ に θ-projection

$$B_{k+1} = \operatorname*{argmin}_{B \in \mathcal{S}} \mathrm{KL}(B, \bar{B})$$

Sparse quasi-Newton [Yamashita, '08]

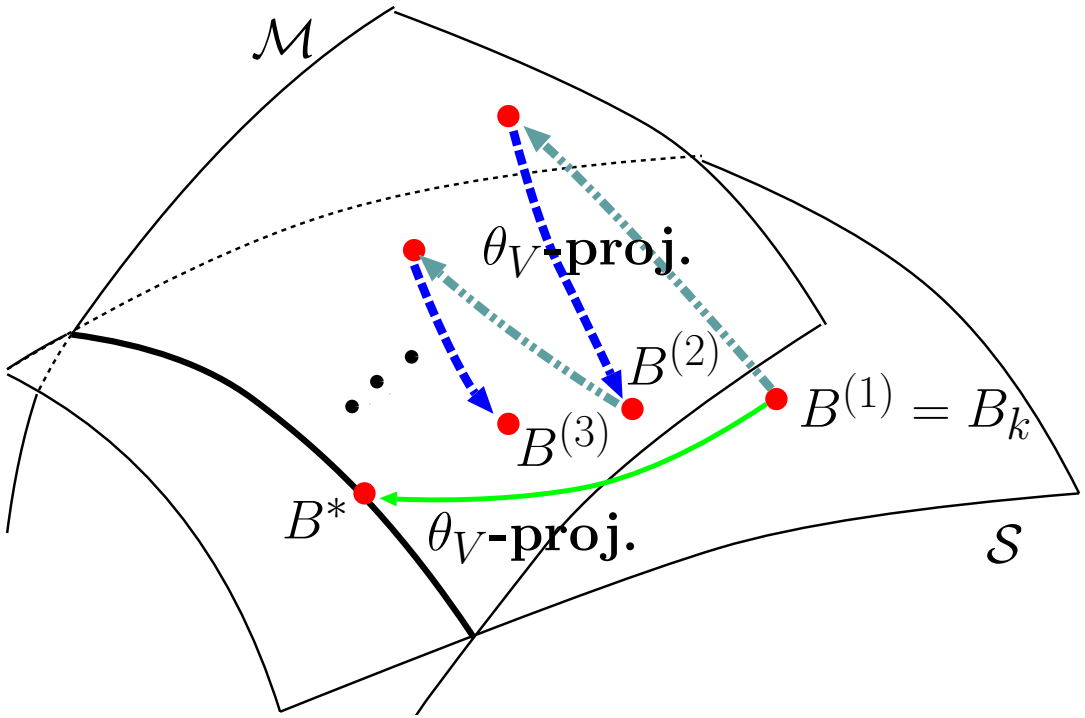


- $B_k \longrightarrow \bar{B} = B^{DFP}[B_k] \longrightarrow B_{k+1}$: em-algorithm
- $B_k \longrightarrow \bar{B} = B^{BFGS}[B_k] \longrightarrow B_{k+1}$: boosting

Sparse V-quasi-Newton – boosting-type extension –

1. $\bar{B}$ は B_k の $\mathcal{M}$ への θ_V-projection (V-BFGS)
2. B_{k+1} は B_k の $\mathcal{S}$ への θ_V-projection (計算可能)

θ_V-projection を繰り返すと $B^* = \mathrm{argmin}_{B \in \mathcal{M} \cap \mathcal{S}} D_V(B, B_k)$ に収束

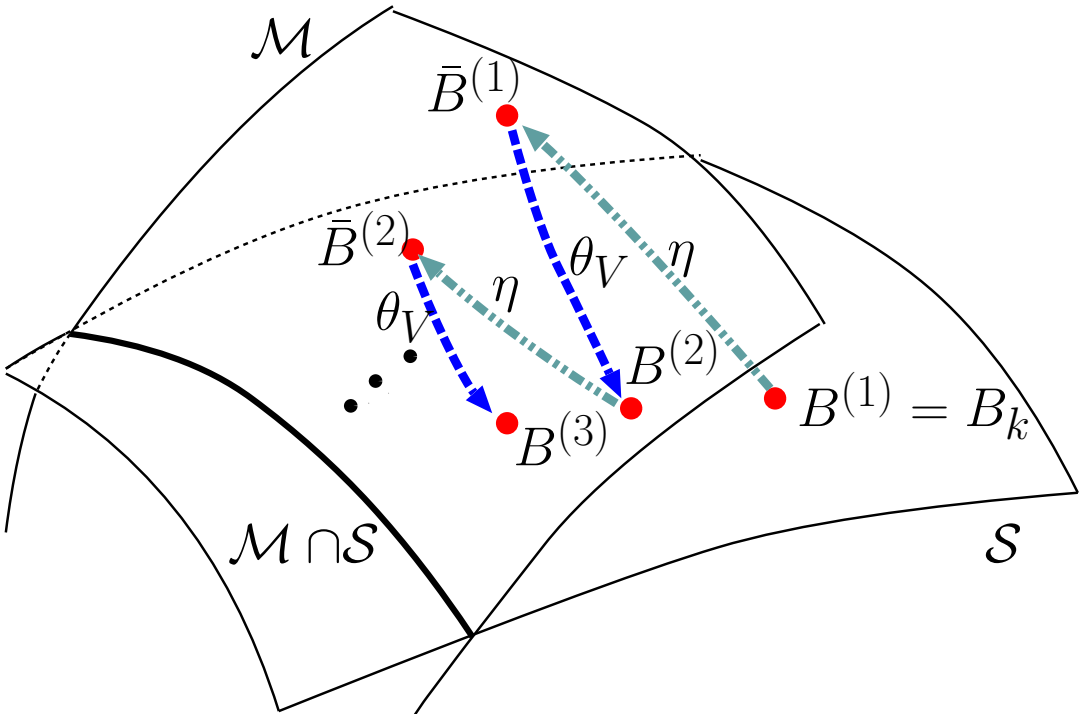


$$D_V(B^*, B^{(1)}) \geq D_V(B^*, B^{(2)}) \geq \cdots \geq D_V(B^*, B^{(T)})$$

Sparse V-quasi-Newton – em-type extension –

1. $\bar{\boldsymbol{B}}$ は $\boldsymbol{B_k}$ の $\boldsymbol{\mathcal{M}}$ への $\boldsymbol{\eta}$-projection (DFP)
2. $\boldsymbol{B_{k+1}}$ は $\boldsymbol{B_k}$ の $\boldsymbol{\mathcal{S}}$ への $\boldsymbol{\theta_V}$-projection

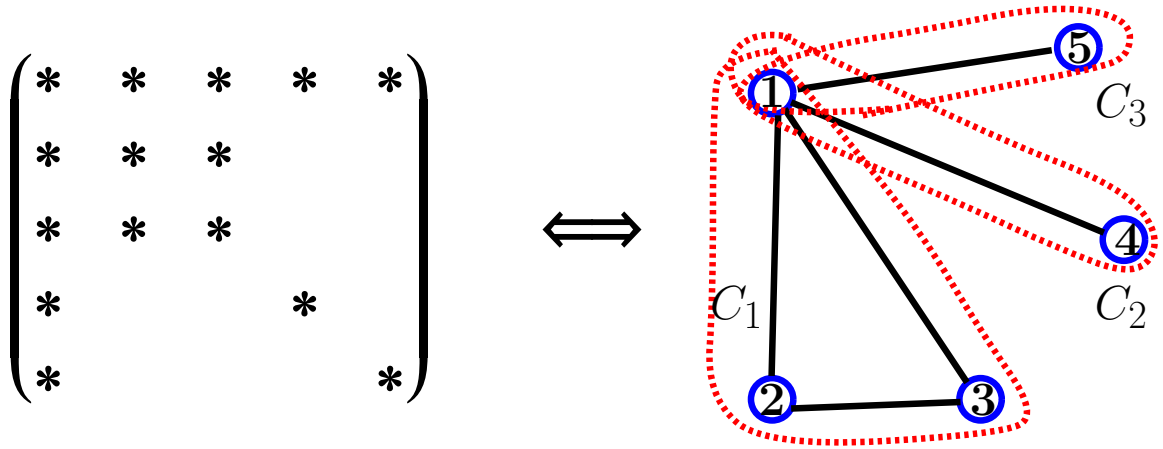
$\boldsymbol{\eta}$, $\boldsymbol{\theta_V}$-projection を繰り返したときの収束先？



$$\boldsymbol{D_V(B^{(1)}, \bar{B}^{(1)}) \geq D_V(B^{(2)}, \bar{B}^{(2)}) \geq \cdots \geq D_V(B^{(T)}, \bar{B}^{(T)})}$$

Computation of Projectoin onto $\mathcal{S}$

- $G = (\{1, \ldots, n\}, F)$: cordal graph とする (loop を除く)
- maximal clique of G: $C_1, \ldots, C_L$



- $\bar{B} = B^{BFGS}[B_k]$ or $B^{DFP}[B_k]$ の $\mathcal{S}$ への θ_V-projection

$$\min_{B} \; D_V(B, \bar{B}), \quad B \in \mathcal{S}$$

$\longrightarrow$ 解 $(B_{opt})^{-1}$: $\bar{H} = \bar{B}^{-1}$ の部分行列 $\bar{H}_{C_\ell C_\ell}$ から計算可能

  - Yamashita の方法：Sparse clique-factorization [Fukuda et al.,'00]
    $\longrightarrow$ V-quasi Newton にも適用可
- Computation cost of $B^{(t)} \to B^{(t+1)}$: $O(\sum_{\ell=1}^{L} |C_L|^2) \cong O(n)$

— Other Topics —

Convergence property of V-BFGS

標準的な仮定：

1. The objective function $f \in C^2(\mathbf{R}^n)$.
2. The level set $\mathcal{L} = \{x \in \mathbf{R}^n \mid f(x) \leq f(x_0)\}$ is convex, and there exist positive constants m and M such that

$$m\|z\|^2 \leq z^\top \nabla^2 f(x) z \leq M\|z\|^2$$

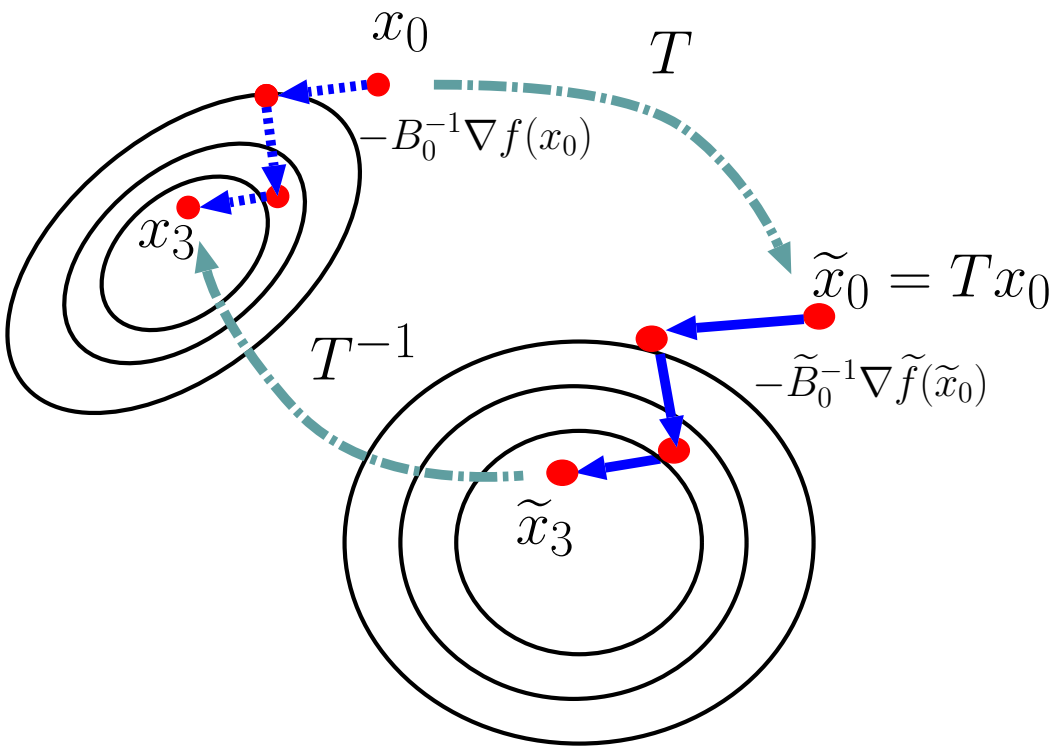
for all $z \in \mathbf{R}^n$ and $x \in \mathcal{L}$.

Theorem

$\exists L_1, L_2 > 0$ *such that* $L_1 \leq \nu \leq L_2 \implies \lim_{k\to\infty} x_k = x^*$ (*local opt.*)

Invariance under Group Action -1/2-

- 変数変換： $x \longrightarrow \widetilde{x} = Tx, \qquad \widetilde{f}(\widetilde{x}) := f(T^{-1}\widetilde{x})$
- 不変性：$\widetilde{x}_k = Tx_k,\ k = 0, 1, 2, \ldots,$ は成立するか？
  Newton method, BFGS or DFP with exact line search で成立



不変なアルゴリズム：変数変換のもとで収束性などの挙動は不変

Invariance under Group Action -2/2-

transformation: $x \longrightarrow \widetilde{x} = Tx$

- $T \in \mathbf{SL}(n) = \{T \in \mathbf{GL}(n) \mid \det T = 1\}$
  $\Longrightarrow$ 任意の potential V に対して V-BFGS は不変.
- $T \in \mathbf{GL}(n)$ に対して V-BFGS が不変
  $\Longleftrightarrow$ power potential $V(z) = (1 - z^{\gamma})/\gamma$
  - $\lim_{\gamma \to 0}(1 - z^{\gamma})/\gamma = -\log z$
  - $\mathbf{PD}(n)$ 上の射影：$\mathbf{GL}(n)$-group action に対して不変

V-BFGS update with power potential:

$$B_{k+1} = \left(\frac{s^{\top}y}{s^{\top}B_k s}\right)^{\alpha} B^{BFGS}[B_k] + \left(1 - \left(\frac{s^{\top}y}{s^{\top}B_k s}\right)^{\alpha}\right)\frac{yy^{\top}}{s^{\top}y},$$

$$\alpha = \frac{\gamma}{1 - (n-1)\gamma}, \quad \left(-\frac{1}{n-1} < \alpha < 1 \text{ for strict convexity}\right)$$

$\alpha = 1$: a popular self-scaling quasi-Newton.

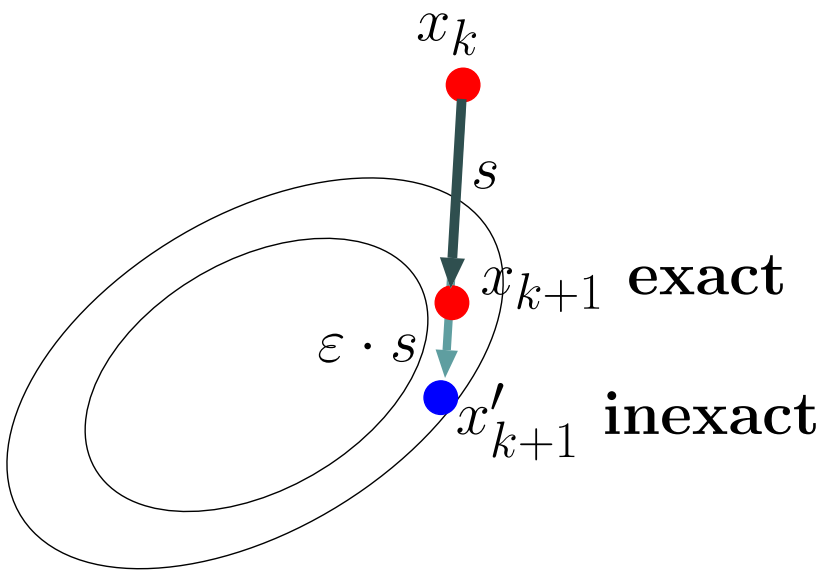
Robustness against inexact line search -1/3-

line search の誤差 $\Longrightarrow$ quasi-Newton update に影響

$$x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k), \quad s = x_{k+1} - x_k, \quad y = \nabla f(x_{k+1}) - \nabla f(x_k)$$

(Inexact line search) $\min_{B \in \mathrm{PD}(n)} D_V(B, B_k)$ subject to $B(1+\varepsilon)s = y + \varepsilon \tilde{y}$

$\Longrightarrow$ opt. sol. $B_{k+1}^{(\varepsilon)}$



$$s \longrightarrow s + \varepsilon \cdot s$$

$$y \longrightarrow y + \varepsilon \cdot \nabla^2 f(x_{k+1})\, s + O(\varepsilon^2)$$

Robustness against inexact line search -2/3-

A measure of sensitivity against numerical error:

$$\text{Influence function: } dB_V(B_k, \tilde{y}) := \lim_{\varepsilon \to 0} \frac{B_{k+1}^{(\varepsilon)} - B_{k+1}^{(0)}}{\varepsilon}$$

$$\text{Gross error sensitivity } : \max_{B_k, \tilde{y}} \|dB_V(B_k, \tilde{y})\|$$

（B_k, $\tilde{y}$ の範囲を適当に取る）

- Gross error sensitivity が小さい $\Longleftrightarrow$ 誤差の影響が小さい
- robust statistics で使われる評価法

Gross error sensitivity を最小にする potential V を求める

For given $s, y \in \mathbf{R}^n$ s.t. $s^\top y > 0$,

$$\min_{V} \max_{B_k, \tilde{y}} \|dB_V(B_k, \tilde{y})\| \quad \text{subject to } B_k \in \mathrm{PD}(n),\ \tilde{y} \in \tilde{\mathcal{Y}}$$

$\tilde{\mathcal{Y}} \subset \mathbf{R}^n$: a bounded subset ($\|\nabla^2 f\| \leq M < \infty$ に相当)

Robustness against inexact line search -3/3-

- V-BFGS, V-DFP
- $B \cong \nabla^2 f$ update, $H = B^{-1} \cong (\nabla^2 f)^{-1}$ update

4 通りについて Gross error sensitivity の値 を計算

	V-BFGS	V-DFP
B update	BFGS のみ有界	∞
$H = B^{-1}$ update	∞	∞

(for all s, y s.t. $s^\top y > 0$)

- $\nabla^2 f$ を近似する BFGS が (V-拡張のなかで) 最適
- sensitivity for line search: duality is violated.

$$(B, s, y) \leftrightarrow (H, y, s)$$

$$(B, (1+\varepsilon)s, y + \varepsilon\tilde{y}) \nleftrightarrow (H, y + \varepsilon\tilde{y}, (1+\varepsilon)s)$$

Concluding Remarks

Future works

- Superlinear convergence of V-quasi-Newton updates
- Which V is preferable?
  - Robustness against numerical error:
    BFGS for B-update, i.e. $V(z) = -\log(z)$
  - Optimally conditioned update formula

    [Dennis and Wolkovicz, sizing and least-change secant methods, '93]
- (intensive) Numerical experiments
  - rate of convergence
  - computation cost
  - robustness
- Link between computation and geometry

  [Ohara and Tsuchiya, An information geometric approach to polynomial-time interior-point algorithms, '07]

References

- Optimization
  - R. Fletcher. A new result for quasi-Newton formulae. *SIAM J. Optim.*, 1:18–21, 1991.
  - H. Yamashita, Sparse quasi-Newton updates with positive definite matrix completion, Mathematical programming, 2008, vol. 115, no1, pp. 1-30.
  - I. S. Dhillon and J. A. Tropp, Matrix nearness problems with Bregman divergences. *SIAM J. Matrix Anal. Appl.*, 29(4):1120–1146, 2007.
  - O. Güler, F. Gurtuna, and O. Shevchenko. Duality in quasi-newton methods and new variational characterizations of the DFP and BFGS updates. *Optimization Methods and Software*, 24(1):45–62, 2009.
  - J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.
- Information Geometry
  - S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, 2000.
  - A. Ohara and S. Eguchi, Geometry on positive definite matrices and v-potential function, Technical report, ISM Research Memo, 2005.
  - N. Murata, T. Takenouchi, T. Kanamori and S. Eguchi, Information geometry of U-Boost and Bregman divergence, Neural Computation, 16, 1437-1481, 2004.
- Robust Statistics
  - F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust Statistics. The Approach based on Influence Functions*. John Wiley and Sons, Inc., 1986