Finding Global Minimizers for Optimization with Orthogonality Constraints

Honglin Yuan

Computational Mathematics, Peking University

yhlmath@pku.edu.cn

(Joint Work with Xiaoyi Gu[1], Rongjie Lai[2], Zaiwen Wen[3])

2016 International Workshop on Signal Processing, Optimization, and Compressed sensing

December 20, 2016

[1] Computational Mathematics, Peking University

[2] Department of Mathematics, Rensselaer Polytechnic Institute

[3] Beijing International Center for Mathematical Research, Peking University

Overview

Outline

Basic Problem

- Mathematically, the orthogonality constrained problem can be formulated as

$$\min \mathcal{F}(X), \quad \text{s.t.} \quad X^\top X = I, X \in \mathbb{R}^{n\times p}.$$

- Particularly, in the case of $p = 1$, the above orthogonality constraint is also known as *spherical constraint*.
- We also consider the similar problems with more than one constraints.

$$\min \mathcal{F}(X_1, \ldots, X_q), \quad \text{s.t. } X_i^\top X_i = I, X_i \in \mathbb{R}^{n\times p}, i = 1, \ldots, q.$$

Basic Problem (Examples)

- p-harmonic flow
- computation of the stability number
- Cryo-electron microscopy (Cryo-EM)
- community detection
- iterative closest point (ICP)

Challenges and Main Approach

- Non-convexity is the main challenges of orthogonality constrained problem since there may be many local minimizers and saddle points, from which finding global minimizers is generally NP-hard.
- Same issues when optimizing general non-convex unconstrained problems.
- A famous remedy in unconstrained case: adding white noise

$$\mathrm{d}x(t) = -\nabla f(x(t))\mathrm{d}t \quad \Rightarrow \quad \mathrm{d}x(t) = -\nabla f(x(t))\mathrm{d}t + \sigma(t)\mathrm{d}B(t),$$

where $B(t)$ is the standard Brownian motion (also known as Wiener Process).

Challenges and Main Approach (Cont'd)

- Our main ideas: generalize existing unconstrained methods to the orthogonality constrained case.

$$\mathrm{d}X(t) = -\nabla_{\mathcal{M}}\mathcal{F}(X(t))\mathrm{d}t + \sigma(t) \circ \mathrm{d}B_{\mathcal{M}}(t), \tag{1}$$

  where $\mathcal{M}$ is the Stiefel manifold, the differential geometric representation of orthogonality constraints.

- Our main contributions:
  - Derived and analyzed the extrinsic formulation of (1).
  - Designed a numerically efficient SDE solver with strong convergence.
  - Established overall global convergence.
  - Achieved promising numerical results in various problems.

Outline

Stiefel Manifold: A Riemmanian Geometric Overview

- Stiefel (Matrix) Manifold:

$$\mathcal{M}_{n,p} = \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$$

- Tangent Space:

$$\mathcal{T}_X\mathcal{M}_{n,p} = \{Z \in \mathbb{R}^{n\times p} : Z^\top X + X^\top Z = 0\}$$

- Embedded (Riemannian) Metric:

$$g_X^e(Z_1, Z_2) = \mathrm{tr}(Z_1^\top Z_2), \quad \text{where } Z_1, Z_2 \in \mathcal{T}_X\mathcal{M}_{n,p}$$

- Canonical (Riemannian) Metric (from quotient representation O_n/O_{n-p}):

$$g_X^c(Z_1, Z_2) = \mathrm{tr}\left(Z_1^\top \left(I - \frac{1}{2}XX^\top\right) Z_2\right), \quad \text{where } Z_1, Z_2 \in \mathcal{T}_X\mathcal{M}_{n,p}$$

Stiefel Manifold: A Riemmanian Geometric Overview (Cont'd)

- Gradients can be derived once the Riemannian metric is introduced.
  - Let $G = \mathcal{DF}(X) = (\frac{\partial \mathcal{F}(X)}{\partial X_{i,j}})$ be the gradient in the embedded space
  - Embedded gradient: $\nabla^e_{\mathcal{M}} \mathcal{F} = X \mathrm{skew}(X^\top G) + (I - XX^\top)G$
  - Canonical gradient: $\nabla^c_{\mathcal{M}} \mathcal{F} = G - XG^\top X$
- First-order methods can be applied based on the above information, very similar to the unconstrained case.
  - Conjugate gradient (CG) on Stiefel Manifold (Edelman et al., 1998).
  - A feasible method based on Cayley transformation and nonmonotone curvilinear search with the Barzilai-Borwein step size (Wen et al., 2013).
  - A thorough introduction to optimization on matrix manifolds can be referred to (Absil et al., 2007).

Outline

Stochastic Integral and Stochastic Differential Equations

- Definition of Ito Integral

$$\int_{t_0}^{t} f(\tau)\mathrm{d}B_\tau := \lim_{\Delta\to\infty} \sum_{i=1}^{\Delta} f(t_{i-1})[B(t_i) - B(t_{i-1})]$$

  where $B(t)$ is a standard Brownian motion.

- Ito SDE: $\mathrm{d}x_t = b(x_t, t)\mathrm{d}t + \sigma(x_t, t)\mathrm{d}B_t$

$$\Leftrightarrow x_t = x_{t_0} + \int_{t_0}^{t} b(x_\tau, \tau)\mathrm{d}\tau + \int_{t_0}^{t} \sigma(x_\tau, \tau)\mathrm{d}B_\tau$$

- Definition of Stratonovich Integral (NOT equivalent!)

$$\int_{t_0}^{t} f(\tau) \circ \mathrm{d}B_\tau := \lim_{\Delta\to\infty} \sum_{i=1}^{\Delta} f(\frac{t_{i-1} + t_i}{2})[B(t_i) - B(t_{i-1})]$$

Diffusion Methods for Unconstrained Problem

- For unconstrained global optimization problem, consider

$$\mathrm{d}x_t = -\nabla f(x_t)\mathrm{d}t + \sigma(t)\mathrm{d}B_t \tag{2}$$

- Intuitively, the diffusion term $\sigma(t)\mathrm{d}B_t$ allows the trajectory to "climb over the mountains" and escape from the local stationary points.
- Kolmogorov Forward Equation (also known as Fokker-Planck Equation in physics) characterizes the density function $p(x,t)$ of (2):

$$\frac{\partial p}{\partial t} = \nabla \cdot (\nabla f(x)p) + \frac{1}{2}\sigma(t)^2 \Delta p$$

- If $\sigma(t)$ is constant, FPE has the equilibrium solution (Gibbs distribution):

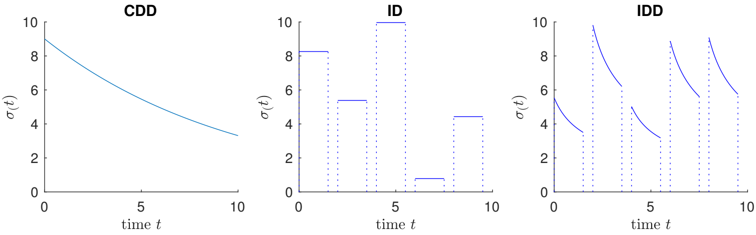
$$\bar{p}_\sigma(x) = \frac{1}{Z} e^{-\frac{2f(x)}{\sigma^2}}$$

Diffusion Methods for Unconstrained Problem (Cont'd)

- Classical work: Continuous Diminishing (Chiang et al., 1987).

$$\sigma(t) = \sqrt{c/\log(2+t)}, \quad \text{for large } c$$

- Intermittent Diffusion (ID) (Chow et al., 2009).
- Spatially inhomogeneous diffusion (Poliannikov et al., 2010).
- Interactive Diffusions: Parallelize the diffusion (Sun et al., 2013).
- We are interested in the Intermittent Diminishing Diffusion (IDD), shown in the figure below.



Diffusion Methods for Constrained Problem

Let $\sigma(t) = c/\log(2+t)$.

- (Parpas et al., 2006) Linear constraints $Ax = b$ and $x \geq 0$.

$$\mathrm{d}x(t) = P[-\nabla f(x(t)) + \mu x(t)^{-1}]\mathrm{d}t + \sqrt{\sigma(t)}P\mathrm{d}B(t)$$

  where $P = I - A^\top(AA^\top)^{-1}A$ and $\mu > 0$ is the barrier parameter.

- (Parpas et al., 2009) Nonlinear equality constraints $g(x) = 0$, $g : \mathbb{R}^n \to \mathbb{R}^m$.

$$\mathrm{d}x(t) = [-\nabla f(x(t)) - \nabla g(x(t))^\top \lambda(x(t), t)]\mathrm{d}t + \sqrt{\sigma(t)}\mathrm{d}B(t)$$

  where $\nabla f \in \mathbb{R}^{n\times 1}$, $\nabla g \in \mathbb{R}^{m\times n}$ and

$$\lambda(x, t) := [\nabla g(x)\nabla g(x)^\top]^{-1}[g(x) + \sigma(t)\Delta g(x) - \nabla g(x)\nabla f(x)].$$

- This nonlinear method is not feasible.

Outline

Extrinsic SDE

- Consider the intrinsic SDE on the Stiefel manifold

$$\mathrm{d}X(t) = -\nabla_{\mathcal{M}}\mathcal{F}(X(t))\mathrm{d}t + \sigma(t) \circ \mathrm{d}B_{\mathcal{M}}(t).$$

From now on we adopt the Canonical metric.

Theorem (extrinsic SDE)

The extrinsic SDE on the canonical Stiefel manifold is

$$\mathrm{d}X(t) = -\nabla_{\mathcal{M}}\mathcal{F}(X(t))\mathrm{d}t + \sigma(t) \sum_{u=1}^{n} \sum_{v=1}^{p} P_{uv}(X(t)) \circ \mathrm{d}B_{uv}(t)$$

where $\{B_{uv}(t)\}$ *is a series of (independent) 1D Brownian motion, and*

$$P_{uv}(X) = E_{uv} - \frac{\sqrt{2}}{2} X E_{uv}^{\top} X - \left(1 - \frac{\sqrt{2}}{2}\right) X X^{\top} E_{uv}, \quad X \in \mathcal{M}_{n,p},$$

$$u = 1, 2, \ldots, n, \quad v = 1, 2, \ldots, p.$$

Extrinsic SDE: well-posedness

Theorem (Well-posedness)

Let V be an arbitrary smooth vector field on the Stiefel manifold $\mathcal{M}_{n,p}$.

(a) *There exists some smooth extensions of $V(X)$ and $P_{uv}(X)$ in $\mathbb{R}^{n\times p}$ (denoted by $\tilde{V}(X)$ and $\tilde{P}_{uv}(X)$), which are globally Lipschitz. Once fixed, there exists a unique solution $X(t,w)$ for the extended SDE*

$$\mathrm{d}X(t) = \tilde{V}(X(t))\mathrm{d}t + \sigma(t)\sum_{u=1}^{n}\sum_{v=1}^{p}\tilde{P}_{uv}(X(t))\circ \mathrm{d}B_{uv}(t).$$

(b)

$$\mathbb{P}\{X(t)\in\mathcal{M}_{n,p}|X(0)\in\mathcal{M}_{n,p}\} = 1, \quad \forall\ t\geq 0.$$

$$\mathbb{P}\{\det(X(t)) = \det(X(0))|X(0)\in\mathcal{M}_{n,n}\} = 1, \quad \forall\ t\geq 0.$$

The solution is unique regardless of the extension of V and P_{uv}.

Extrinsic SDE: Ito Representation

Theorem (Ito Version)

The corresponding Ito version of Stratonovich SDE on $\mathcal{M}_{n,p}$ (with feasible initial point $X(0) \in \mathcal{M}_{n,p}$) is given by

$$\mathrm{d}X(t) = \left(V(X(t)) - \frac{n-1}{2}\sigma^2(t)X(t) \right) \mathrm{d}t \\ + \sigma(t) \sum_{u=1}^{n} \sum_{v=1}^{p} \left(E_{uv} - \frac{\sqrt{2}}{2} X E_{uv}^\top X - (1 - \frac{\sqrt{2}}{2}) X X^\top E_{uv} \right) \mathrm{d}B_{uv}(t).$$

Extrinsic SDE: Generator of the Diffusion Term

Theorem

Suppose $W(t)$ is the solution of the following SDE

$$\mathrm{d}W(t) = \sum_{u=1}^{n}\sum_{v=1}^{p}\left(E_{uv} - \frac{\sqrt{2}}{2}WE_{uv}^{\top}W - (1-\frac{\sqrt{2}}{2})WW^{\top}E_{uv}\right)\circ \mathrm{d}B_{uv}(t),$$

and then $W(t)$ is driven by half of Laplacian-Beltrami operator $\Delta_{\mathcal{M}}$ on Stiefel Manifold, i.e.

$$\frac{1}{2}\Delta_{\mathcal{M}}\varphi(W(t)) = \mathcal{L}\varphi(W(t)) := \lim_{t\to 0+}\frac{\mathbb{E}[\varphi(W(t))|W(0)=w_0] - \varphi(w_0)}{t}.$$

Extrinsic SDE: Fokker-Planck Equation

Theorem (Fokker-Planck Equation of Extrinsic SDE)

The Fokker-Planck Equation of the following SDE

$$\mathrm{d}X(t) = -\nabla_{\mathcal{M}}\mathcal{F}(X(t))\mathrm{d}t + \sigma(t)\sum_{u=1}^{n}\sum_{v=1}^{p} P_{uv}(X(t)) \circ \mathrm{d}B_{uv}(t)$$

is given by

$$\frac{\partial p}{\partial t} = -\nabla_{\mathcal{M}} \cdot (p\nabla_{\mathcal{M}}\mathcal{F}) + \frac{1}{2}\sigma^2(t)\Delta_{\mathcal{M}}p.$$

Outline

Numerical SDE Scheme and Strong Convergence

Given time discretization $t_0 = \tau_0 < \tau_1 < \cdots < \tau_K = T$, we consider the following approximation based on Cayley transformation:

$$\begin{cases} Z_k = -\Delta_k G_k + \sigma_k (I_n - (1 - \frac{\sqrt{2}}{2}) Y_k Y_k^\top) \Delta B_k, \\ A_k = Z_k Y_k^\top - Y_k Z_k^\top, \\ Y_{k+1} = Y_k + A_k \left(\frac{Y_k + Y_{k+1}}{2} \right) = \left(I - \frac{A_k}{2} \right)^{-1} \left(I + \frac{A_k}{2} \right) Y_k, \end{cases}$$

where $k = 0, 1, \ldots, K-1$ and Y_0 is the initial point. Δ_k, $\tau_{k+1} - \tau_k$ and σ_k, G_k are the shorthands of $\tau_{k+1} - \tau_k$, $\sigma(t_k)$ and $\nabla \mathcal{F}(Y_k)$, respectively.

Theorem (Half Order Strong Convergence)

Let $\delta := \max_k(\tau_{k+1} - \tau_k)$, and then there exists a positive constant $C = C(T)$ independent of δ, as well as a constant $\delta_0 > 0$ such that

$$\mathbb{E}\|X(T) - Y_K\|_2^2 \le C\delta, \quad \forall \delta \in (0, \delta_0)$$

Acceleration of Numerical SDE Scheme

The following lemma can be applied to accelerate the numerical scheme.

Lemma (Wen et al., 2013)

(1) *Rewrite* $A_k = U_k V_k^\top$ *for* $U_k = [Z_k, Y_k]$ *and* $V_k = [Y_k, -Z_k]$. *If* $I - \frac{1}{2} V_k^\top U_k$ *is invertible, then*

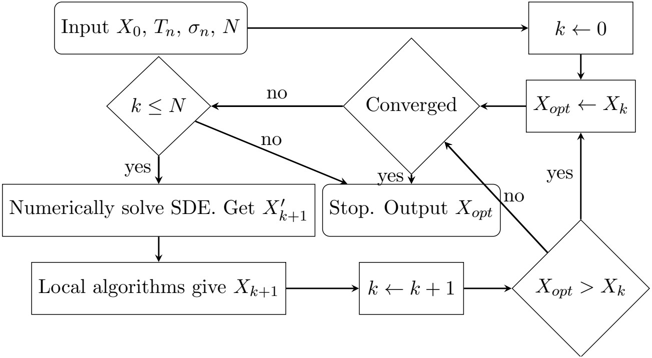
$$Y_{k+1} = Y_k + U_k \left(I - \frac{1}{2} V_k^\top U_k \right)^{-1} V_k^\top Y_k.$$

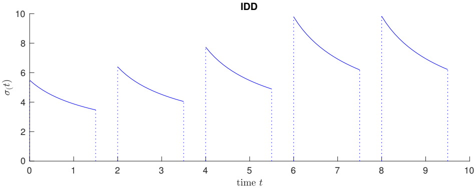
(2) *For the vector case,*

$$Y_{k+1} = Y_k + \frac{Z_k}{1 - (\frac{1}{2})^2 (Z_k^\top Y_k)^2 + (\frac{1}{2})^2 Z_k^\top Z_k} - \frac{Z_k^\top Y_k - \frac{1}{2}((Z_k^\top Y_k)^2) + Z_k^\top Z_k}{1 - (\frac{1}{2})^2 (Z_k^\top Y_k)^2 + (\frac{1}{2})^2 Z_k^\top Z_k} Y_k.$$

Outline

Intermittent Diminishing Diffusion on Stiefel Manifold (IDDM)





Outline

Gibbs Distribution

Theorem

Assuming $\sigma(t) = \sigma_0$ which is a constant, the distribution $p(x,t)$ (given by Fokker-Planck Equation) converges (ℓ_1) to the Gibbs distribution

$$\tilde{p}_{\sigma_0}(x) = \frac{1}{Z} e^{-2\mathcal{F}/\sigma_0^2},$$

where Z is the normalization constant

$$Z = \int_{\mathcal{M}} e^{2\mathcal{F}/\sigma_0^2} \mathrm{d}_{\mathcal{M}} X.$$

Global Convergence of ID

Theorem (Convergence Results of ID)

Assuming that the local algorithm satisfies $\mathcal{F}(X_k) \leq \mathcal{F}(X_k')$. Let the global minimum be $\mathcal{F}^*$, and suppose X_{opt} to be the optimal solution obtained by ID. For any given $\epsilon > 0$ and $\zeta > 0$, $\exists \sigma > 0$, $T(\sigma) > 0$ and $N_0 > 0$ such that if $\sigma_i \leq \sigma$, $T_i > T(\sigma_i)$ and $N > N_0$,

$$\mathbb{P}(\mathcal{F}(X_{opt}) < \mathcal{F}^* + \zeta) \geq 1 - \epsilon.$$

Outline

Test Problem 1: Homogeneous polynomial optimization

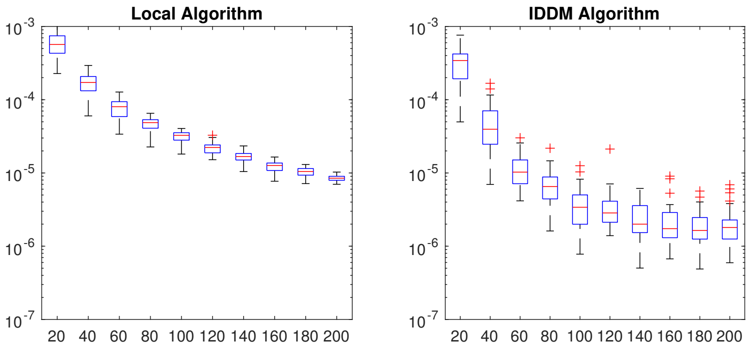
$$\min \mathcal{F}(x) = \sum_{1\leq i\leq n} x_i^6 + \sum_{1\leq i\leq n-1} x_i^3 x_{i+1}^3, \quad \text{s.t.} x^\top x = 1$$

The results shown below are based on 50 independent run. In each run, we call Random-Start for 10 times and compare it with one call of IDDM with 10 cycles.

n	Random-Start			IDDM		
	min	mean	max	min	mean	max
20	2.28e-04	6.03e-04	1.41e-03	4.98e-05	3.30e-04	7.60e-04
40	6.02e-05	1.70e-04	2.93e-04	6.98e-06	5.03e-05	1.68e-04
60	3.38e-05	7.89e-05	1.27e-04	4.16e-06	1.16e-05	3.02e-05
80	2.27e-05	4.68e-05	6.53e-05	1.61e-06	7.17e-06	2.18e-05
100	1.81e-05	3.18e-05	4.05e-05	7.78e-07	3.76e-06	1.26e-05
120	1.52e-05	2.20e-05	3.27e-05	1.39e-06	3.52e-06	2.12e-05
140	1.04e-05	1.67e-05	2.34e-05	5.03e-07	2.47e-06	6.16e-06
160	7.70e-06	1.24e-05	1.65e-05	6.72e-07	2.24e-06	9.04e-06
180	7.15e-06	1.04e-05	1.31e-05	4.88e-07	1.99e-06	5.66e-06
200	7.00e-06	8.53e-06	1.03e-05	5.95e-07	2.09e-06	6.90e-06

Test Problem 1: Cont'd

We plot the above data in the following boxplot. [4]



[4] On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol.

Outline

Test Problem 2: Stability Number

- Given an undirected graph $G = (V, E)$, the stability number $\alpha(G)$ is the cardinality of the LARGEST stable subset of V (whose vertices are not connected to each other).

- Motzkin and Straus showed that the stability number can be computed by spherically constrained problem:

$$\alpha(G)^{-1} = \min_{\|x\|_2=1} \sum_{i=1}^{n} x_i^4 + 2 \sum_{(i,j)\in E} x_i^2 x_j^2 .$$

- The results shown below are based on 50 independent run. In each run, we call Random-Start for 10 times and compare it with one call of IDDM with 10 cycles. Please take note that $\alpha(G)$ is expected to be maximized.

Problem		Random-Start (Motzkin)			IDDM(Motzkin)		
name	n	median $\alpha(G)$	max $\alpha(G)$	cpu	median $\alpha(G)$	max $\alpha(G)$	cpu
hamming-9-8	512	169	179	0.293	172	186	0.093
hamming-10-2	1024	65	67	0.907	66	73	0.860
keller4	171	9	11	0.257	11	11	0.197
G43	1000	181	189	0.654	188	195	0.502
G44	1000	182	187	0.678	190	198	0.497
G45	1000	180	187	0.660	187.5	196	0.502
G46	1000	180	182	0.675	188.5	200	0.502
G47	1000	184	191	0.673	192	202	0.509
G51	1000	332	337	0.818	343	347	0.604
G52	1000	330	336	0.842	341	345	0.615
G53	1000	330	334	0.761	340	344	0.563
G54	1000	323	329	0.708	334	337	0.540
1et.512	512	91	95	0.204	92	98	0.178
1tc.512	512	101	106	0.169	102.5	108	0.153
1zc.512	512	52	53	0.305	52	55	0.289
1dc.1024	1024	69	72	1.055	70	73	1.022
1et.1024	1024	154	158	0.416	158.5	164	0.361
1tc.1024	1024	175	180	0.359	182	188	0.302
1zc.1024	1024	91	94	0.737	93	102	0.651
1dc.2048	2048	119	123	2.689	124	129	2.256
1et.2048	2048	271	276	0.981	289.5	294	0.902
1tc.2048	2048	306	312	0.803	324	331	0.743
1zc.2048	2048	160.5	164	1.770	170	174	1.383

Outline

Test Problem 3: Iterative Closest Point

- Consider two data sets $P = \{p_1, p_2, ..., p_n\} \in \mathbb{R}^{n\times d}$ and $Q = \{q_1, q_2, ..., q_n\} \in \mathbb{R}^{n\times d}$. Assume P and Q are the same up to a rigid motion in $\mathbb{R}^d$. The iterative closest point (ICP) considers the following problem

$$\min_{R,\sigma} \|PR - \sigma Q\|_F^2, \quad \text{s.t.} \quad R^\top R = I_d, \sigma^\top \sigma = I_n, \sigma \geq 0.$$

- Given a parameter $\mu > 0$, a simple relaxation is given by

$$\min_{R,\sigma} \|PR - \sigma Q\|_F^2 + \mu \langle I_{(\sigma<0)}, \sigma\rangle^2, \quad \text{s.t.} \quad R^\top R = I_d, \sigma^\top \sigma = I_n.$$

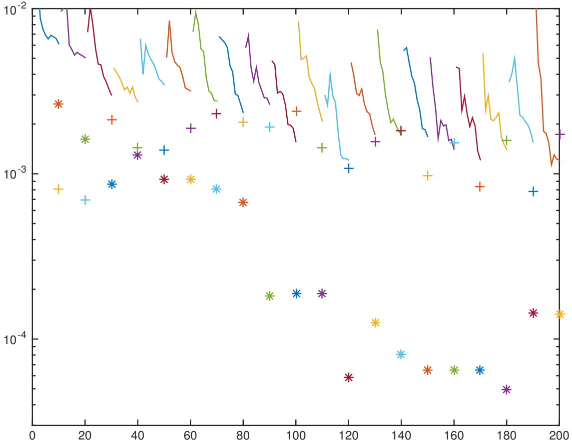
Test Problem 3: Numerical Results

- The results shown below are based on 50 independent run. In each run, we call Random-Start for 20 times and compare it with 1 call of IDDM with 20 cycles. We recognize the recovery successful if the exact rotation is found for at least once in this run.

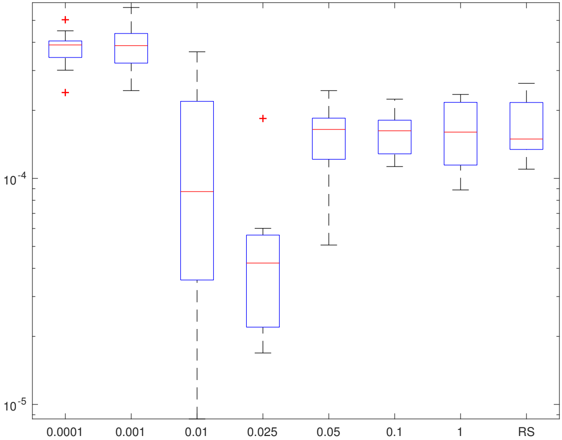
Problem		Random-Start		IDDM	
n	d	recovery rate	cpu	recovery rate	cpu
6	3	4 %	1.966	84%	1.039
6	4	6 %	3.097	82%	1.504
6	5	6 %	3.877	74%	2.034
8	3	0 %	2.174	66%	3.044
8	4	0 %	2.816	64%	3.075
8	5	0 %	3.234	74%	3.418
10	3	0 %	2.857	22%	2.212
10	4	0 %	3.686	20%	4.574
10	5	0 %	3.977	54%	4.734
12	3	0 %	2.761	6%	6.395
12	4	0 %	3.449	16%	5.585
12	5	0 %	4.390	4%	8.189
14	3	0 %	3.738	2%	7.066
14	4	0 %	4.829	10%	9.101
14	5	0 %	5.356	6%	10.381

Outline

Non-Independence over Cycles



Sensitivity to Diffusion Strength



Thanks For Your Attention!

Main References

Wen, Z. & Yin, W. (2012)
A feasible method for optimization with orthogonality constraints
Mathematical programming 142, 397-434

Edelman, A.; Arias, T. A. & Smith, S. T. (1998)
The Geometry of Algorithms with Orthogonality Constraint
SIAM Journal on Matrix Analysis and Applications 20, 303-353

Hsu, E. (2001)
Stochastic Analysis on Manifolds
American Mathematical Society 38

Villani, C. (2012)
Optimal Transport