A Variable Smoothing for Nonconvexly Constrained Nonsmooth Optimization with Application to Sparse Spectral Clustering

Keita Kume and Isao Yamada, Tokyo Institute of Technology

東京工業大学 Tokyo Institute of Technology

ICASSP 2024 KOREA

Summary

- **Target problem:**

$$\underset{\boldsymbol{x}\in C}{\text{Minimize}}\ f(\boldsymbol{x}) := \underbrace{h(\boldsymbol{x})}_{\text{smooth}} + \underbrace{g\circ G(\boldsymbol{x})}_{\text{nonsmooth}} \cdots (\bigstar)$$

($\boldsymbol{x}\in C$: constraint)

- $\mathcal{X}, \mathcal{Z}$ are Euclidean spaces
- $C(\neq\emptyset)\subset\mathcal{X}$ is a (possibly nonconvex) closed subset of $\mathcal{X}$
- $h:\mathcal{X}\to\mathbb{R}$: differentiable, ∇h is Lipschitz continuous over C
- $g:\mathcal{Z}\to\mathbb{R}$: weakly convex, nonsmooth Lipschitz continuous $\overset{\Leftrightarrow}{\text{def}}\ \exists\eta>0$ s.t. $g+\frac{\eta}{2}\|\cdot\|^2$ is convex
- $G:\mathcal{X}\to\mathcal{Z}$: smooth (possibly nonlinear) mapping

- **Typical applications:** sparsity-aware signal processing, e.g., sparse PCA, sparse spectral clustering, robust subspace recovery

- **Challenging issues:** { nonconvex constraint C; nonsmoothness and nonconvexity of g }

- **Our contributions:**
  - Proposal of an optimization algorithm of guaranteed global convergence to a stationary point (First available algorithm for ($\bigstar$), and generalization of [1]).
  - Application to sparse spectral clustering (SSC) based on nonconvex sparse regularizer (Inherently first nonconvex approach for SSC).

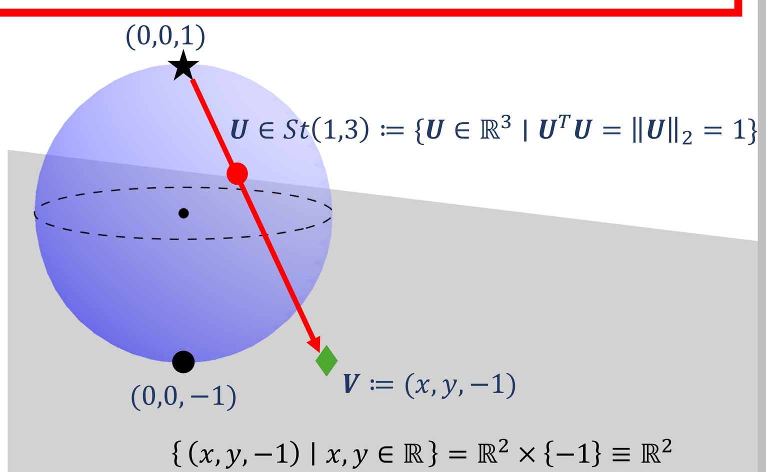
How to deal with constraint set C ? ⇨ parametrization

Key idea

Parameterize C in terms of the Euclidean space $\mathcal{Y}$ with a smooth mapping $F:\mathcal{Y}\to\mathcal{X}$ such that $C=\{F(\boldsymbol{y})\in\mathcal{X}\mid\boldsymbol{y}\in\mathcal{Y}\}$.

Example: smoothly parameterizable C

- Stiefel manifold
  $St(p,N) := \{\boldsymbol{U}\in\mathbb{R}^{N\times p}\mid \boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{I}_p\}$
  (with Cayley-type transforms [2,3])
- Bounded-rank matrices
  $\mathbb{R}^{M\times N}_{\leq r} := \{\boldsymbol{X}\in\mathbb{R}^{M\times N}\mid \operatorname{rank}(\boldsymbol{X})\leq r\}$
  (with the multiplication $\boldsymbol{X}=\boldsymbol{Y}\boldsymbol{Z}^T$ [$\boldsymbol{Y}\in\mathbb{R}^{M\times r}, \boldsymbol{Z}\in\mathbb{R}^{N\times r}$] [4])



Illustrative example of parametrization for $St(1,3)$ (stereographic projection [Riemann'1851])

We consider the following parameterized problem instead of ($\bigstar$):

$$\underset{\boldsymbol{y}\in\mathcal{Y}}{\text{Minimize}}\ f\circ F(\boldsymbol{y}) = (h+g\circ G)\circ F(\boldsymbol{y}) \cdots (\clubsuit)$$

($\boldsymbol{y}\in\mathcal{Y}$: Euclidean space)

We have the following relation of necessary conditions (optimality condition) of a local minimizer for ($\bigstar$) and ($\clubsuit$).

Theorem 4.1 [Relations of optimality conditions]

$$\underbrace{\boldsymbol{0}\in\partial f(F(\boldsymbol{y}^\star)) + N_C(F(\boldsymbol{y}^\star))}_{\text{Optimality condition for }(\bigstar)} \Leftrightarrow \underbrace{\boldsymbol{0}\in\partial(f\circ F)(\boldsymbol{y}^\star)}_{\text{Optimality condition for }(\clubsuit)}$$

under { the Clarke regularity on C (i.e,. C is sufficiently smooth); $N_C(F(\boldsymbol{y}^\star)) = \{\boldsymbol{x}\in\mathcal{X}\mid (\mathrm{D}F(\boldsymbol{y}^\star))^*(\boldsymbol{x})=\boldsymbol{0}\}$ at $\boldsymbol{y}^\star\in\mathcal{Y}$ }

※∂f denotes the general subdifferential.
$N_C(\boldsymbol{x})\subset\mathcal{X}$ denotes the general normal cone.
} Note: these are different senses from convex analysis (see [5]).
$(\mathrm{D}F(\boldsymbol{y}))^*$ denotes the adjoint of the Fréchet derivative (Jacobi matrix) at $\boldsymbol{y}\in\mathcal{Y}$.

How to deal with nonsmoothness of g ? ⇨ smoothing

Key idea (inspired by [1])

Use a smoothed surrogate function of η-weakly convex g.

➡ Moreau envelope ${}^{\mu}g$ of g with $\mu\in(0,\eta^{-1})$

$$(\bar{\boldsymbol{z}}\in\mathcal{Z})\ {}^{\mu}g(\bar{\boldsymbol{z}}) := \inf_{\boldsymbol{z}\in\mathcal{Z}}\left(g(\boldsymbol{z})+\frac{1}{2\mu}\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|^2\right)$$

- $\lim_{\mu\to0}{}^{\mu}g(\bar{\boldsymbol{z}}) = g(\bar{\boldsymbol{z}})$
- ${}^{\mu}g$ is differentiable, and $\nabla{}^{\mu}g$ is Lipschitz continuous

Theorem 3.1 [Characterization of optimality condition]

For $(\mu_n)_{n=1}^{\infty}(\subset(0,\eta^{-1}))\xrightarrow[n\to\infty]{}0$, and $(\boldsymbol{y}_n)_{n=1}^{\infty}\subset\mathcal{Y}\xrightarrow[n\to\infty]{}\exists\bar{\boldsymbol{y}}\in\mathcal{Y}$,

$$d(\boldsymbol{0},\partial(f\circ F)(\bar{\boldsymbol{y}}))\leq\liminf_{n\to\infty}\|\nabla((h+{}^{\mu_n}g\circ G)\circ F)(\boldsymbol{y}_n)\|$$

$\liminf_{n\to\infty}\|\nabla((h+{}^{\mu_n}g\circ G)\circ F)(\boldsymbol{y}_n)\|=0$ implies $d(\boldsymbol{0},\partial(f\circ F)(\bar{\boldsymbol{y}}))=0$, i.e., $\boldsymbol{0}\in\partial(f\circ F)(\bar{\boldsymbol{y}})$.

Proposed algorithm achieves $\liminf_{n\to\infty}\|\nabla((h+{}^{\mu_n}g\circ G)\circ F)(\boldsymbol{y}_n)\|=0$:

1. Set $\mu_n := \kappa n^{-\frac{1}{\alpha}}$ and $f_{[n]} := h+{}^{\mu_n}g\circ G$ $(\alpha>1, \exists\gamma_n>0, \exists\kappa>0)$
2. Update $\boldsymbol{y}_{n+1} := \boldsymbol{y}_n-\gamma_n\nabla(f_{[n]}\circ F)(\boldsymbol{y}_n)$

Increment n

Theorem 3.3 [Convergence analysis (informal)]

Assume that $\nabla(f_{[n]}\circ F)$ is Lipschitz continuous with a Lipschitz constant $\varpi\mu_n^{-1}$ with some $\varpi>0$, and $\gamma_n>0$ is computed by the so-called *backtracking algorithm*. Then, $(\boldsymbol{y}_n)_{n=1}^{\infty}$ generated by the proposed algorithm satisfies:

$$\liminf_{n\to\infty}\|\nabla((h+{}^{\mu_n}g\circ G)\circ F)(\boldsymbol{y}_n)\|=0.$$

Application to sparse Spectral Clustering (SC)

Goal: split given data $(\boldsymbol{\xi}_i)_{i=1}^N\subset\mathbb{R}^d$ into K groups without labeled data.

Outline of SC [6] ($\boldsymbol{D}\in\mathbb{R}^{N\times N}$: degree matrix; $\boldsymbol{W}\in\mathbb{R}^{N\times N}$: adjacency matrix)

1. Construct a similarity graph $\mathcal{G}$ of $(\boldsymbol{\xi}_i)_{i=1}^N$.
2. Compute K smallest eigenvectors $\boldsymbol{U}^\star\in St(K,N)$ of the graph Laplacian $\boldsymbol{L}:=\boldsymbol{I}-\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}\in\mathbb{R}^{N\times N}$.
3. Apply k-means algorithm to N row (normalized) vectors of $\boldsymbol{U}^\star$.

(Steps 2 and 3 correspond to splitting $\mathcal{G}$ into K connected subgraphs)

To improve SC, the Sparse SC (SSC) utilizes a prior knowledge that $\boldsymbol{U}^\star\boldsymbol{U}^{\star T}$ is sparse (block diagonal) in the ideal case [7].

Step 2 of SC can be refined along SSC as:

$$\text{Find }\boldsymbol{U}^\star\in\underset{\boldsymbol{U}\in St(K,N)}{\operatorname{argmin}}\ \underbrace{\operatorname{Tr}(\boldsymbol{U}^T\boldsymbol{L}\boldsymbol{U})}_{\text{Eigenvalue decomp.}}+\underbrace{\lambda\psi(\boldsymbol{U}\boldsymbol{U}^T)}_{\text{Promote sparsity of }\boldsymbol{U}^\star\boldsymbol{U}^{\star T}}\cdots(\spadesuit)$$

($\psi:\mathbb{R}^{N\times N}\to\mathbb{R}$: sparsity promoting function)

($\spadesuit$) can be reformulated as ($\clubsuit$) with $h(\boldsymbol{U}):=\operatorname{Tr}(\boldsymbol{U}^T\boldsymbol{L}\boldsymbol{U})$, $g:=\lambda\psi$, $G(\boldsymbol{U}):=\boldsymbol{U}\boldsymbol{U}^T$ and the generalized Cayley transform [3]

$$F:=\Phi_{\boldsymbol{S}}^{-1}:\mathcal{Y}\to St(p,N):\boldsymbol{Y}\mapsto\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{Y})(\boldsymbol{I}+\boldsymbol{Y})^{-1}\boldsymbol{I}_{N\times p},$$

where $\mathcal{Y}:=\left\{\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^T\\ \boldsymbol{B} & \boldsymbol{0}\end{bmatrix}\in\mathbb{R}^{N\times N}\ \middle|\ \begin{matrix}\boldsymbol{A}^T=-\boldsymbol{A}\in\mathbb{R}^{p\times p}\\ \boldsymbol{B}\in\mathbb{R}^{(N-p)\times p}\end{matrix}\right\}$.

($St(p,N)$ and F above satisfy the assumption in Theorem 4.1)

We propose to solve ($\spadesuit$) with MCP (Minimax Concave Penalty) [8] as ψ.

Result: the proposed SSC with MCP achieves the best performance!

	iris		shuttle		segmentation		breast cancer		glass		wine		seeds	
	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
SC [6]	0.778 (0.000)	0.745 (0.000)	0.387 (0.043)	0.205 (0.059)	0.501 (0.043)	0.341 (0.056)	0.417 (0.000)	0.419 (0.000)	0.321 (0.026)	0.174 (0.019)	0.433 (0.000)	0.363 (0.000)	0.662 (0.007)	0.659 (0.010)
SSC(ℓ_1+relax) [7]	0.785 (0.000)	0.786 (0.000)	0.426 (0.047)	0.279 (0.063)	0.503 (0.036)	0.343 (0.052)	0.433 (0.000)	0.462 (0.000)	0.325 (0.021)	0.176 (0.016)	0.433 (0.000)	0.363 (0.000)	0.671 (0.032)	0.675 (0.034)
Proposed SSC(ℓ_1+Gr)	**0.823** (0.000)	**0.818** (0.000)	0.427 (0.047)	0.276 (0.069)	0.501 (0.037)	0.341 (0.052)	0.473 (0.000)	0.547 (0.000)	0.323 (0.025)	0.175 (0.020)	0.433 (0.000)	0.363 (0.000)	0.667 (0.000)	0.668 (0.000)
Proposed SSC(MCP+Gr)	**0.823** (0.000)	**0.818** (0.000)	**0.434** (0.049)	**0.294** (0.062)	**0.507** (0.042)	**0.351** (0.061)	**0.558** (0.000)	**0.664** (0.000)	**0.341** (0.030)	**0.180** (0.023)	**0.442** (0.000)	**0.376** (0.000)	**0.721** (0.037)	**0.708** (0.044)

[1] A. Böhm, S.J. Wright, "Variable smoothing for weakly convex composite functions," J. Optim. Theory Appl., 2021.
[2] K. Kume, I. Yamada, "A global Cayley parametrization of Stiefel manifold for direct utilization of optimization mechanisms over vector spaces," ICASSP, 2021.
[3] K. Kume , I. Yamada, "Generalized left-localized Cayley parametrization for optimization with orthogonality constraints," Optimization, 2022.
[4] E. Levin, J. Kileel, N. Boumal, "The effect of smooth parametrizations on nonconvex optimization landscapes," Math. Program., 2024.
[5] R. Rockafellar, R. J.-B. Wets, *Variational Analysis*, Springer Verlag, 3rd edition, 2010.
[6] A. Ng, M. Jordan, Y. Weiss, "On spectral clustering: Analysis and an algorithm," NeurIPS, 2001.
[7] C. Lu, S. Yan, Z. Lin, "Convex sparse spectral clustering: Single-view to multi-view," IEEE Trans. Image Process., 2016.
[8] C.-H. Zhang, "Nearly unbiased variable selection under mini- max concave penalty," The Annals of Statistics, 2010.

arXiv version