Nonconvex Sparse Optimization and Algorithms

Xianchao Xiu

Department of Automation

上海大学
Shanghai University

Chinese Academy of Sciences, January 12, 2025

Joint work with Wanquan Liu (SYSU), Lingchen Kong (BJTU) and others

Outline

Sparse Optimization

- Sparse optimization considers

$$\min_{x\in\mathbb{R}^n} \; f(x) + \lambda\|x\|_0$$

$$\Updownarrow$$

$$\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad \|x\|_0 \leq s$$

  - x can be extended to matrices and tensors
  - $f(x)$ may be nonsmooth even nonconvex
  - $\|x\|_0$ counts the number of nonzeros
  - λ and s are parameters
- Also called compressed sensing and variable selection
- Broad applications in machine learning, pattern recognition and engineering
- `https://github.com/xianchaoxiu/Sparse-Optimization`

Algorithms

- Convex algorithms
  - Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B, 1996
  - Candès-Tao, Decoding by linear programming, IEEE TIT, 2005
  - Donoho, Compressed sensing, IEEE TIT, 2006
- Nonconvex algorithms
  - Fan-Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 2001
  - Chen-Xu-Ye, Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization, SIAM Journal on Scientific Computing, 2010
  - Xu-Chang-Xu-Zhang, $L_{1/2}$ regularization: A thresholding representation theory and a fast solver, IEEE TNNLS, 2012
- Direct algorithms
  - Blumensath-Davies, Iterative hard thresholding for compressed sensing, Applied and Computational Harmonic Analysis, 2009
  - Foucart, Hard thresholding pursuit: An algorithm for compressive sensing, SIAM Journal on Numerical Analysis, 2011
  - Yuan-Li-Zhang, Gradient hard thresholding pursuit, JMLR, 2018

More

- Bach-Jenatton-Mairal-Obozinski, Optimization with sparsity-inducing penalties, Foundations and Trends in Machine Learning, 2012
- Jain-Kar, Non-convex optimization for machine learning, Foundations and Trends in Machine Learning, 2017
- Hastie-Tibshirani-Wainwright, Statistical learning with sparsity: The Lasso and generalizations, CRC Press, 2015
- Zhao, Sparse optimization theory and methods, CRC Press, 2018
- Fan-Li-Zhang-Zou, Statistical foundations of data science, CRC Press, 2020
- Wright-Ma, High-dimensional data analysis with low-dimensional models: Principles, computation, and applications, Cambridge University Press, 2022
- Parhi-Nowak, Deep learning meets sparse regularization: A signal processing perspective, IEEE Signal Processing Magazine, 2023
- Tillmann-Bienstock-Lodi-Schwartz, Cardinality minimization, constraints, and regularization: A survey, SIAM Review, 2024

Outline

$\ell_1 - \ell_p$ Minimization

- Xiu-Kong-Li-Qi, Computational Optimization and Applications, 2018

$$\min_{x \in \mathbb{R}^n} \quad \|Ax - b\|_1 + \lambda \|x\|_p^p \ (0 < p < 1) \tag{1}$$

- Consider the following ϵ-approximations

$$\min_{x \in \mathbb{R}^n} \ F_{\alpha,\epsilon}(x) = \|Ax - b\|_1 + \lambda \sum_{i=1}^{n} (|x_i|^\alpha + \epsilon_i)^{\frac{p}{\alpha}}$$

$$\Downarrow$$

$$\min_{x \in \mathbb{R}^n} \ F_\epsilon(x) = \|Ax - b\|_1 + \lambda \sum_{i=1}^{n} h_{u_\epsilon}(x_i) \tag{2}$$

where

$$h_{u_\epsilon}(x_i) = \min_{0 \le s \le u_\epsilon} p \left(|x_i| s - \frac{p-1}{p} s^{\frac{p}{p-1}} \right), \quad u_\epsilon = \left(\frac{\epsilon}{\lambda n} \right)^{\frac{p-1}{p}}$$

$\ell_1 - \ell_p$ Minimization

- (Definition) We say that $x^* \in \mathbb{R}^n$ is a generalized first-order stationary point of (1) if

$$0 \in (A^\top \mathrm{sgn}(Ax^* - b))_i x_i^* + \lambda p |x_i^*|^p, \quad i = 1, 2, \cdots, n$$

  Furthermore, the following statement holds

$$|x_i^*| \ \geq \ \left(\frac{\lambda p}{\|A_i\|_1}\right)^{\frac{1}{1-p}}, \quad \forall i \in T \tag{3}$$

- (Lower Bound) Let ϵ be a constant such that

$$0 < \epsilon < \lambda n \left(\frac{\|A_i\|_1}{\lambda p}\right)^{\frac{p}{p-1}} \tag{4}$$

  Suppose that x^* is a generalized first-order stationary point of (2). Then, x^* is also a generalized first-order stationary point of (1). Moreover, the nonzero entries of x^* satisfy the lower bound property (3).

$\ell_1 - \ell_p$ Minimization

- (Convergent Theorem) Assume that ϵ satisfies (4) and set q as $\frac{1}{p} + \frac{1}{q} = 1$. Suppose that x^* is an accumulation point of $\{x^k\}$. Then x^* is a generalized first-order stationary point of (1). Moreover, the nonzero entries of x^* satisfy the lower bound (3).

Choose an arbitrary $x^0 \in \mathbb{R}^n$ and ϵ such that (4) holds. Set $k = 0$

1) Solve the weighted ℓ_1 minimization problem

$$x^{k+1} \in \mathrm{argmin}_x \left\{ \|Ax - b\|_1 + \lambda p \sum_{i=1}^n s_i^k |x_i| \right\}$$

where $s_i^k = \min \left\{ \left(\frac{\epsilon}{\lambda n}\right)^{\frac{1}{q}}, |x_i^k|^{\frac{1}{q-1}} \right\}$ for all i

2) Set $k \leftarrow k + 1$ and go to step 1)

End

$\ell_1 - \ell_p$ Minimization

- Comparison with FISTA

m	n	FISTA	Alg. 2	Alg. 3	Alg. 4	FISTA	Alg. 2	Alg. 3	Alg. 4
100	500	11.1587	0.0455	0.0303	0.0223	0.0008	0.4183	0.3087	0.2453
200	1000	8.6122	0.2097	0.1518	0.1279	0.0020	0.1413	0.0564	0.0484
300	1500	2.0159	0.1498	0.1195	0.1079	0.0067	0.2095	0.1293	0.1265
400	2000	2.3528	0.1057	0.0877	0.0799	0.1093	0.3648	0.2905	0.2791
500	2500	1.1584	0.1672	0.1491	0.1091	0.0310	0.4761	0.4799	0.4583
600	3000	0.9855	0.0972	0.0972	0.0972	0.0386	0.9324	0.7700	0.7684
700	3500	1.1239	0.0947	0.0940	0.0872	0.0756	1.7057	1.6983	1.5231
800	4000	0.8065	0.0958	0.0924	0.0861	0.1598	2.5905	2.4271	2.3562
900	4500	0.8734	0.0982	0.0981	0.0823	0.1546	3.3103	3.2272	3.2263
1000	5000	1.1301	0.0942	0.0912	0.0851	0.1937	4.0071	3.9719	4.1359

Fused Regression

- Xiu-Liu-Li-Kong, Computational Statistics & Data Analysis, 2019

$$\min_{\beta} \quad \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\beta) + \sum_{i=1}^{p} \Phi_{\tau_2}(\beta_{i+1} - \beta_i)$$

  - Φ_{τ_1} and Φ_{τ_2} can be the same or different
  - Nonconvex penalty functions: ℓ_p, SCAD, MCP, capped ℓ_1
- For notational simplicity, define

$$\min_{\beta} \quad \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\beta) + \Phi_{\tau_2}(D\beta)$$

with

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}$$

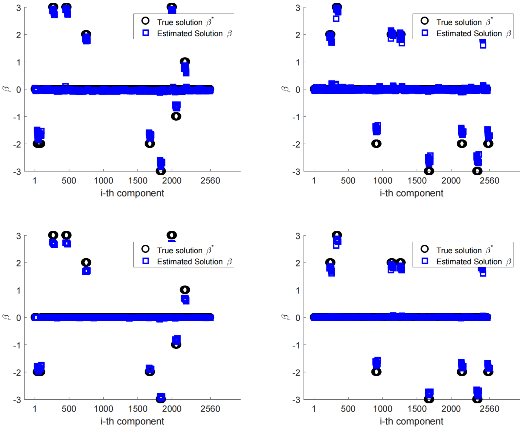
Fused Regression

- Alternating direction method of multipliers (ADMM)

$$
\begin{aligned}
\min_{\alpha,\gamma,\beta} \quad & \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\alpha) + \Phi_{\tau_2}(\gamma) \\
\text{s.t.} \quad & \alpha = \beta \\
& \gamma = D\beta
\end{aligned}
$$

- (Convergent Theorem) Suppose that $\{(\alpha^k, \gamma^k, \beta^k, w_1^k, w_2^k)\}$ is a generated sequence. Then the sequence converges to a stationary point.

- Recovery results



Sparse LDA

- Liu-Feng-Xiu-Liu, Pattern Recognition, 2024

$$
\begin{aligned}
\min_{Q} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda \|Q\|_{2,1} \\
\text{s.t.} \quad & Q^\top Q = I
\end{aligned}
$$

$$\Downarrow$$

$$
\begin{aligned}
\min_{P,Q,E} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda_1 \|Q\|_{2,1} + \lambda_2 \|E\|_1 \\
\text{s.t.} \quad & X = PQ^\top X + E, \; P^\top P = I
\end{aligned}
$$

$$\Downarrow$$

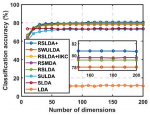
$$
\begin{aligned}
\min_{P,Q,E} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda_1 \|Q\|_{2,0} + \lambda_2 \|E\|_0 \\
\text{s.t.} \quad & X = PQ^\top X + E, \; P^\top P = I
\end{aligned}
$$

$$\Downarrow$$

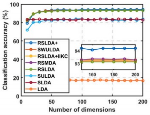
$$
\begin{aligned}
\min_{P,Q,E} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda_1 \|Q\|_{2,0} + \lambda_2 \|Q\|_0 + \lambda_3 \|E\|_0 \\
\text{s.t.} \quad & X = PQ^\top X + E, \; P^\top P = I
\end{aligned}
$$

Sparse LDA

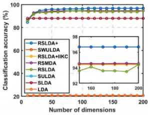
- Classification accuracy



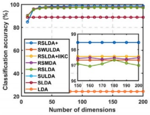
(a) AR(d=3)



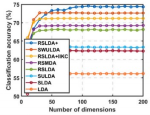
(b) AR (d=6)



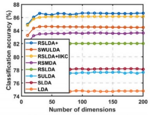
(c) AR (d=9)



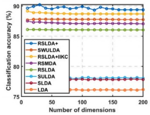
(d) AR (d=12)



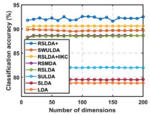
(e) MNIST (d=10)



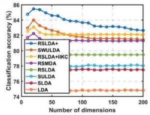
(f) MNIST (d=50)



(g) MNIST (d=100)

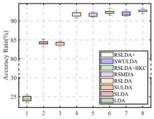


(h) MNIST (d=200)

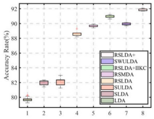


(i) HGR (d=4)

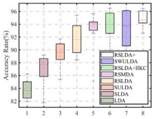
- Model stability



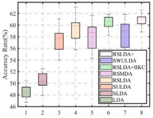
(a) AR



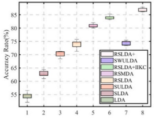
(b) MNIST



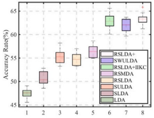
(c) HGR



(d) NEU



(e) Caltech



(f) CAR

Outline

Sparse CCA

- Xiu-Yang-Kong-Liu, Applied Mathematics and Computation, 2020

$$\min_{\beta,\theta} \quad -\beta^\top X^\top Y\theta + \lambda\|\beta\|_1 + \mu\|\theta\|_1$$
$$\text{s.t.} \quad \|X\beta\|^2 \leq 1, \ \|Y\theta\|^2 \leq 1$$

- Alternating minimization algorithm (AMA)
  - Update β by

$$\begin{aligned} \min_{\beta} \quad & -\beta^\top X^\top Y\theta + \lambda\|\beta\|_1 \\ \text{s.t.} \quad & \|X\beta\|^2 \leq 1 \end{aligned} \tag{5}$$

  - Update θ by

$$\min_{\theta} \quad -\beta^\top X^\top Y\theta + \mu\|\theta\|_1$$
$$\text{s.t.} \quad \|Y\theta\|^2 \leq 1$$

Sparse CCA

- The dual optimization problem of (5) is

$$\begin{aligned} \min_{\alpha,\gamma} \quad & \frac{1}{2}\|\alpha - Y\theta\|^2 + \delta_{\lambda B_\infty}(\gamma) \\ \text{s.t.} \quad & X^\top \alpha - \gamma = 0 \end{aligned}$$

$$\Downarrow$$

$$\mathcal{L}_\delta(\alpha, \gamma; \beta) = \frac{1}{2}\|\alpha - Y\theta\|^2 + \delta_{\lambda B_\infty}(\gamma) - \beta^\top (X^\top \alpha - \gamma) + \frac{\delta}{2}\|X^\top \alpha - \gamma\|^2$$

- Stage 1: Apply a semi-smooth Newton method for solving

$$(\alpha^{k+1}, \gamma^{k+1}) = \arg\min_{\alpha,\gamma} \{\mathcal{L}_\delta(\alpha, \gamma; \beta^k)\}$$

- Stage 2: Compute

$$\beta^{k+1} = \beta^k - \tau\delta_k(X^\top \alpha^{k+1} - \gamma^{k+1})$$

- (Convergent Theorem) The generated sequence $\{(\beta^k, \theta^k)\}$ converges to a stationary point.

Generalized CCA

- Li-Xiu-Liu-Miao, IEEE Signal Processing Letters, 2022

$$\min_{U, P_v} \sum_{v=1}^{M} \|U - X_v P_v\|_F^2$$
$$\text{s.t.} \quad U^\top U = I_d, \ \|P_v\|_{2,0} \leq s_v$$

- Alternating minimization algorithm (AMA)
  - Update U^{k+1} by
$$\min_{U} \sum_{v=1}^{M} \|U - X_v P_v^k\|_F^2$$
$$\text{s.t.} \quad U^\top U = I_d$$
  - Update $P_v^{k+1}(v = 1, \ldots, M)$ by
$$\min_{P_v} \sum_{v=1}^{M} \|U^{k+1} - X_v P_v\|_F^2 \qquad (6)$$
$$\text{s.t.} \quad \|P_v\|_{2,0} \leq s_v$$

Generalized CCA

- Denote $f(P_v) := \|U^{k+1} - X_v P_v\|_F^2$. Then

$$\nabla f(P_v) = 2X_v^\top (X_v P_v - U^{k+1}), \quad \nabla^2 f(P_v) = 2I_d \otimes X_v^\top X_v$$

- The α_v-stationary point of (6) can be given by

$$P_v = \Pi_{\mathcal{S}}(P_v - \alpha_v \nabla f(P_v))$$

$$\Downarrow$$

$$\begin{aligned} 0 &= P_v - \Pi_{\mathcal{S}}(P_v - \alpha_v \nabla f(P_v)) \\ &= \begin{pmatrix} (P_v)_{T_v} \\ (P_v)_{\overline{T}_v} \end{pmatrix} - \begin{pmatrix} (P_v)_{T_v} - \alpha_v \nabla_{T_v} f(P_v) \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \alpha_v \nabla_{T_v} f(P_v) \\ (P_v)_{\overline{T}_v} \end{pmatrix} \end{aligned}$$

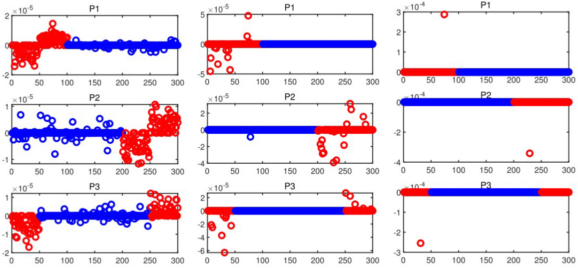
- Newton hard thresholding pursuit (NHTP)

Generalized CCA

- Runtime comparison

Problem Scale	GCCA	SGCCA	SCGCCA
(1,000;300;300;300)	0.04	0.04	**0.01**
(5,000;300;300;300)	0.23	0.28	**0.03**
(10,000;300;300;300)	0.40	0.41	**0.07**
(50,000;300;300;300)	2.32	2.27	**0.34**
(100,000;300;300;300)	4.58	4.35	**0.66**
(1,000;1,500;1,500;1,500)	0.42	0.40	**0.02**
(5,000;1,500;1,500;1,500)	1.35	1.16	**0.12**
(10,000;1,500;1,500;1,500)	2.63	2.24	**0.24**
(50,000;1,500;1,500;1,500)	13.21	10.56	**1.18**
(100,000;1,500;1,500;1,500)	26.60	22.53	**2.35**
(1,000;3,000;3,000;3,000)	1.53	1.58	**0.17**
(5,000;3,000;3,000;3,000)	3.92	3.49	**0.23**
(10,000;3,000;3,000;3,000)	6.87	5.65	**0.45**
(50,000;3,000;3,000;3,000)	32.02	23.18	**2.29**
(100,000;3,000;3,000;3,000)	667.69	629.54	**4.91**

- Extracted feature comparison



Distributed Optimization

- Qu-Chen-Xiu-Liu, Neurocomputing, 2024

$$\begin{aligned} \min_{Y \in \mathbb{R}^{n \times p}} \quad & \sum_{i=1}^{d} f_i(Y) \\ \text{s.t.} \quad & \|Y\|_{2,0} \leq s, \ \ Y^\top Y = I_p \end{aligned} \tag{7}$$

$$\Downarrow$$

$$\begin{aligned} \min_{Y \in \mathbb{R}^{n \times p}} \quad & \sum_{i=1}^{d} f_i(Y) + \frac{\mu}{4} \|Y^\top Y - I_p\|_F^2 \\ \text{s.t.} \quad & \|Y\|_{2,0} \leq s \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{Y, \{X_i\} \in \mathbb{R}^{n \times p}} \quad & \sum_{i=1}^{d} f_i(X_i) + \frac{\mu}{4} \|Y^\top Y - I_p\|_F^2 \\ \text{s.t.} \quad & X_i = Y, \ \forall i \in [d], \ \ \|Y\|_{2,0} \leq s \end{aligned} \tag{8}$$

Distributed Optimization

- (Lemma) Let $(\widetilde{Y}^*, \{\widetilde{X}_i^*\})$ be the (local) minimizer of (8). Then there exists $\mu_\epsilon > 0$ such that $\widetilde{Y}^*$ is an ϵ-(local) minimizer of (7) for any $\mu \geq \mu_\epsilon$.
- (Definition) We say $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a KKT point of (8) if it satisfies

$$\begin{cases} 0 \in \nabla g(Y^*) + \sum_{i=1}^d \Lambda_i^* + \mathcal{N}_{\mathcal{S}}(Y^*) \\ 0 = \nabla f_i(X_i^*) - \Lambda_i^*, \ \forall i \in [d] \\ 0 = X_i^* - Y^*, \ \forall i \in [d] \end{cases}$$

- (Definition) We say $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a stationary point of (8) if there exists $\alpha > 0$ such that

$$\begin{cases} Y^* = \mathcal{P}_{\mathcal{S}}(Y^* - \alpha(\nabla g(Y^*) + \sum_{i=1}^d \Lambda_i^*)) \\ 0 = \nabla f_i(X_i^*) - \Lambda_i^*, \ \forall i \in [d] \\ 0 = X_i^* - Y^*, \ \forall i \in [d] \end{cases}$$

- (Optimal Conditions) Suppose that $(Y^*, \{X_i^*\})$ is a local minimizer of (8). Then, there exists Λ_i^* $(i \in [d])$ such that $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a KKT point of (8).

Distributed Optimization

- (Nonincreasing Lemma) Let $\{(Y^k, \{X_i^k\}, \{\Lambda_i^k\})\}$ be the generated sequence and $\beta \geq \sqrt{2}r$. Then the generated augmented Lagrangian sequence is nonincreasing, i.e.,

$$\mathcal{L}_{\beta}(Y^{k+1}, \{X_i^{k+1}\}; \{\Lambda_i^{k+1}\}) \leq \mathcal{L}_{\beta}(Y^k, \{X_i^k\}; \{\Lambda_i^k\})$$

- (Bounded Lemma) Suppose that $\beta \geq 2r$ holds. Then the sequence $\{(Y^k, \{X_i^k\}, \{\Lambda_i^k\})\}$ is bounded. Moreover, it satisfies

$$\begin{cases} \lim\limits_{k\to\infty} \|Y^{k+1} - Y^k\|_F = 0 \\ \lim\limits_{k\to\infty} \|X_i^{k+1} - X_i^k\|_F = 0, \ \forall i \in [d] \\ \lim\limits_{k\to\infty} \|\Lambda_i^{k+1} - \Lambda_i^k\|_F = 0, \ \forall i \in [d] \end{cases}$$

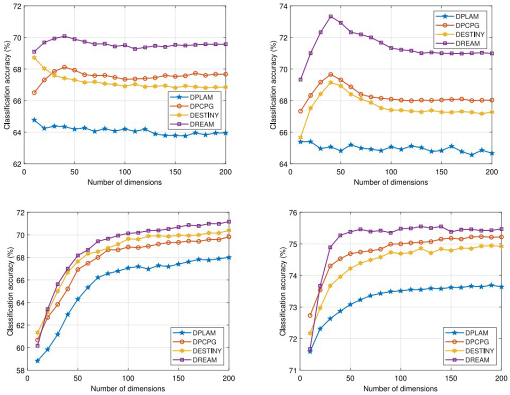
- (Convergent Theorem) Let $\{(Y^k, \{X_i^k\}, \{\Lambda_i^k\})\}$ be the generated sequence and $\beta \geq 2r$. Then, any accumulation point $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a stationary point of (8).

Distributed Optimization

- Runtime comparison

Dataset	DPLAM	DPCPG	DESTINY	DREAM
YALE	0.13	0.09	0.04	**0.02**
ORL	1.19	0.593	0.57	**0.22**
CAR	2.10	1.64	1.53	**0.82**
AR	2.83	2.19	2.01	**1.71**
Vegetable	3.60	2.95	2.50	**2.29**
CIFAR-10	4.79	3.64	3.74	**2.94**

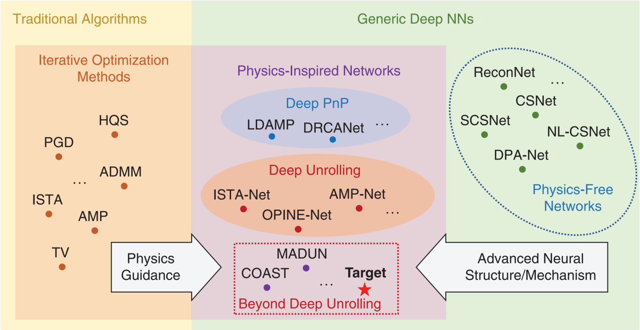
- Classification accuracy



Outline

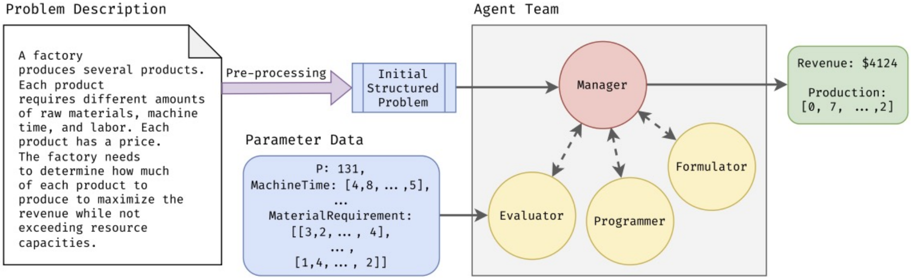
Deep Unfolding Networks

- Gregor-LeCun, Learning fast approximations of sparse coding, ICML, 2010
- Zhang-Ghanem, ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing, CVPR, 2018
- Chen-Liu-Yin, Learning to optimize: A tutorial for continuous and mixed-integer optimization, Science China Mathematics, 2024



Large Language Models

- Yang-Wang-Lu et al., Large language models as optimizers, ICLR, 2024
- AhmadiTeshnizi-Gao-Udell, OptiMUS: Scalable optimization modeling with (MI) LP solvers and large language models, ICML, 2024
- Romera-Barekatain-Novikov et al., Mathematical discoveries from program search with large language models, Nature, 2024



References

- Liu-Feng-Xiu-Liu, Towards robust and sparse linear discriminant analysis for image classification, Pattern Recognition, 2024
- Qu-Chen-Xiu-Liu, Distributed sparsity constrained optimization over the Stiefel manifold, Neurocomputing, 2024
- Li-Xiu-Liu-Miao, An efficient Newton-based method for sparse generalized canonical correlation analysis, IEEE Signal Processing Letters, 2022
- Xiu-Yang-Kong-Liu, tSSNALM: A fast two-stage semi-smooth Newton augmented Lagrangian method for sparse CCA, Applied Mathematics and Computation, 2020
- Xiu-Liu-Li-Kong, Alternating direction method of multipliers for nonconvex fused regression problems, Computational Statistics & Data Analysis, 2019
- Xiu-Kong-Li-Qi, Iterative reweighted methods for $\ell_1 - \ell_p$ minimization, Computational Optimization and Applications, 2018