Overcoming Lower-Level Constraints in Bilevel Optimization: A Novel Approach with Regularized Gap Functions

Wei Yao, Haian Yin, Shangzhi Zeng, Jin Zhang

Problem Formulation

The **constrained** bilevel optimization (BiO) is formulated as:

$$\min_{x\in X, y\in Y} F(x,y) \quad \text{s.t.} \quad y \in S(x) := \operatorname*{argmin}_{y\in Y} \Big\{ f(x,y) \text{ s.t. } g(x,y) \leq 0 \Big\},$$

which tackles nested structures present in constrained learning tasks like constrained meta-learning, adversarial learning, and distributed bilevel optimization.

Assumptions:

- F is **Smooth** and **bounded below** over the feasible set.
- f is **Convex** in y, and **smooth** in both variables.
- g is **Convex** in y, **smooth**, and **Lipschitz continuous** (including its gradients).

Goal: develop a single-loop, first-order algorithm without projection onto the coupled lower-level constraint set.

Gap Function and Its Property

Here we introduce the doubly regularized gap function for the lower-level problem:

$$\mathcal{G}_\gamma(x,y,z) := \max_{\theta\in Y,\lambda\in\mathbb{R}^p_+} \left\{\mathcal{L}(x,y,\lambda) - \frac{1}{2\gamma_2}\|\lambda - z\|^2 - \mathcal{L}(x,\theta,z) - \frac{1}{2\gamma_1}\|\theta - y\|^2\right\},$$

where the Lagrangian function $\mathcal{L}(x,y,z) := f(x,y) + z^{\mathrm{T}} g(x,y)$.

Property I (Lemma 2.1.)

$$\mathcal{G}_\gamma(x,y,z) \le 0 \Leftrightarrow y \in S(x) \text{ and } z \in \mathcal{M}(x,y)$$

Property II (Lemma 2.2.)

$$\nabla\mathcal{G}_\gamma(x,y,z) = \begin{bmatrix} \nabla_x f(x,y) + (\lambda^*)^{\mathrm{T}}\nabla_x g(x,y) \\ \nabla_y f(x,y) + (\lambda^*)^{\mathrm{T}}\nabla_y g(x,y) \\ -(z-\lambda^*)/\gamma_2 \end{bmatrix} - \begin{bmatrix} \nabla_x f(x,\theta^*) + z^{\mathrm{T}}\nabla_x g(x,\theta^*) \\ (y-\theta^*)/\gamma_1 \\ g(x,\theta^*) \end{bmatrix}$$

$$\theta^* := \theta^*(x,y,z) := \operatorname*{argmin}_{\theta\in Y}\left\{\mathcal{L}(x,\theta,z) + \frac{1}{2\gamma_1}\|\theta - y\|^2\right\},$$

$$\lambda^* := \lambda^*(x,y,z) := \operatorname*{argmax}_{\lambda\in\mathbb{R}^p_+}\left\{\mathcal{L}(x,y,z) - \frac{1}{2\gamma_2}\|\lambda - z\|^2\right\} = \mathrm{Proj}_{\mathbb{R}^p_+}\left(z + \gamma_2 g(x,y)\right).$$

Reformulation

Based on the Property I of gap function, we can reformulate the BiO **equivalently** as

$$\min_{(x,y,z)\in X\times Y\times \mathbb{R}^p_+} F(x,y) \quad \text{s.t.} \quad \mathcal{G}_\gamma(x,y,z) \le 0.$$

To develop a gradient-based algorithm, we explore its **penalty** formulation:

$$\min_{(x,y,z)\in X\times Y\times Z} F(x,y) + c\,\mathcal{G}_\gamma(x,y,z),$$

Main Algorithm

- **Bi**level **C**onstrained **GA**p **F**unction-based **F**irst-order **A**lgorithm (**BiC-GAFFA**)
  1. Update Auxiliary Variables:

$$\theta^{k+1} = \mathrm{Proj}_Y\left(\theta^k - \eta_k \nabla_\theta\left(\mathcal{L}(x^k,\theta^k,z^k) + \frac{1}{2\gamma_1}\|\theta^k - y^k\|^2\right)\right)$$

$$\lambda^{k+1} = \mathrm{Proj}_{\mathbb{R}^p_+}\left(z^k + \gamma_2 g(x^k,y^k)\right)$$

  2. Update Main Variables:

$$(x^{k+1}, y^{k+1}, z^{k+1}) = \mathrm{Proj}_{X\times Y\times Z}\left((x^k,y^k,z^k) - \alpha_k \nabla\left(\frac{1}{c_k}F(x^k,y^k) + \mathcal{G}_\gamma(x^k,y^k,z^k)\right)\right)$$

Extension to Bilevel Optimization with Minimax Lower-level Problem

For bilevel optimization with minimax lower-level problem:

$$\min_{x\in X, y\in Y, z\in Z} F(x,y,z) \quad \text{s.t.} \quad (y,z)\in \mathcal{SP}(x) := \mathrm{Sol}\left[\min_{y\in Y}\max_{z\in Z} f(x,y,z)\right].$$

We use following gap function:

$$\mathcal{G}_\gamma^{\text{saddle}}(x,y,z) := \max_{\theta\in Y, \lambda\in\mathbb{R}^p_+} \left\{ f(x,y,z) - \frac{1}{2\gamma_2}\|\lambda - z\|^2 - f(x,\theta,z) - \frac{1}{2\gamma_1}\|\theta - y\|^2 \right\}.$$

And reformulate the BiO as:

$$\min_{(x,y,z)\in X\times Y\times\mathbb{R}^p_+} F(x,y,z) \quad \text{s.t.} \quad \mathcal{G}_\gamma^{\text{saddle}}(x,y,z)\le 0.$$

➢ Single-loop Hessian-free algorithm for BiO with minimax lower-level problem

1. Update Auxiliary Variables:

$$\theta^{k+1} = \mathrm{Proj}_Y\left(\theta^k - \eta_k \nabla_\theta\left(\mathcal{L}(x^k,\theta^k,z^k) + \frac{1}{2\gamma_1}\|\theta^k - y^k\|^2\right)\right)$$

$$\lambda^{k+1} = \mathrm{Proj}_{\mathbb{R}^p_+}\left(z^k + \gamma_2 g(x^k,y^k)\right)$$

2. Update Main Variables:

$$(x^{k+1}, y^{k+1}, z^{k+1})$$

$$= \mathrm{Proj}_{X\times Y\times Z}\left((x^k,y^k,z^k) - \alpha_k\nabla\left(\frac{1}{c_k}F(x^k,y^k) + \tilde{\mathcal{G}}_\gamma^{\text{saddle}}(x^k,y^k,z^k)\right)\right)$$

Synthetic Experiments

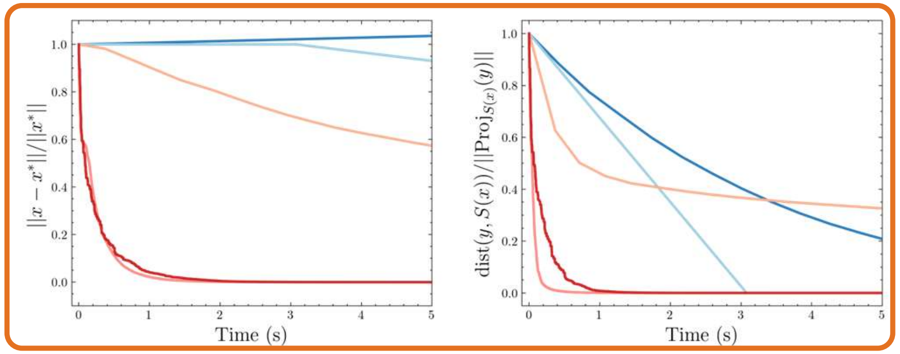
$$\begin{aligned}\min_{\substack{\mathbf{x}\in\mathcal{X},\\(\mathbf{y}_1,\mathbf{y}_2)\in\mathcal{Y}}}\quad & (\mathbf{y}_1 - 2\cdot\mathbb{1}_n)^{\mathrm{T}}(\mathbf{x}-\mathbb{1}_n) + \|\mathbf{y}_2+3\cdot\mathbb{1}_n\|^2\\ \text{s.t.}\quad & (\mathbf{y}_1,\mathbf{y}_2)\in\underset{(\mathbf{y}_1,\mathbf{y}_2)\in\mathcal{Y}}{\operatorname{argmin}}\left\{\frac{1}{2}\|\mathbf{y}_1\|^2-\mathbf{x}^{\mathrm{T}}\mathbf{y}_1+\mathbb{1}_n^{\mathrm{T}}\mathbf{y}_2\ \text{ s.t. }\ \sum_{i=1}^{n}\boxed{h(\mathbf{x}_i)}+\mathbb{1}_n^{\mathrm{T}}\mathbf{y}_1+\mathbb{1}_n^{\mathrm{T}}\mathbf{y}_2=0\right\}.\end{aligned}$$

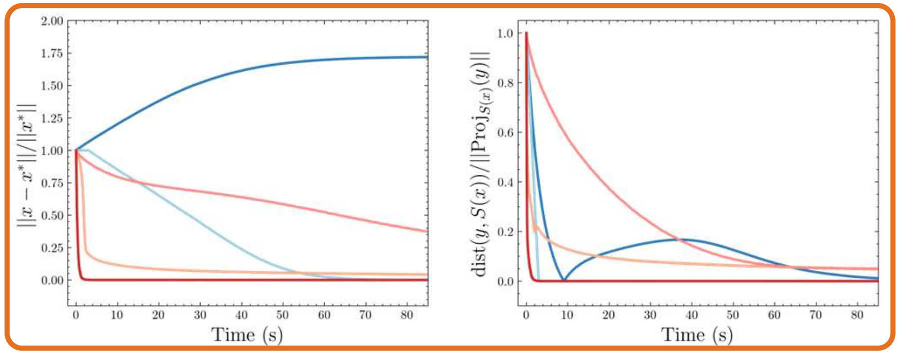
① $h(x) = x$

② $h(x) = x^3$

Performance (Compared with the SOTA Algorithms)

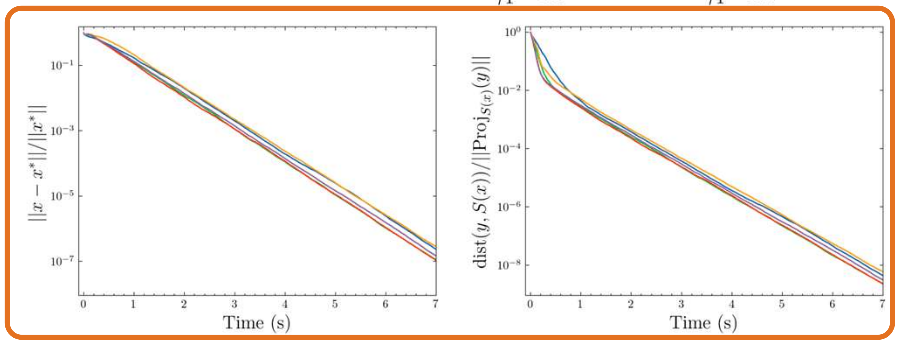
AIPOD GAM BVFSM LV-HBA BiC-GAFFA

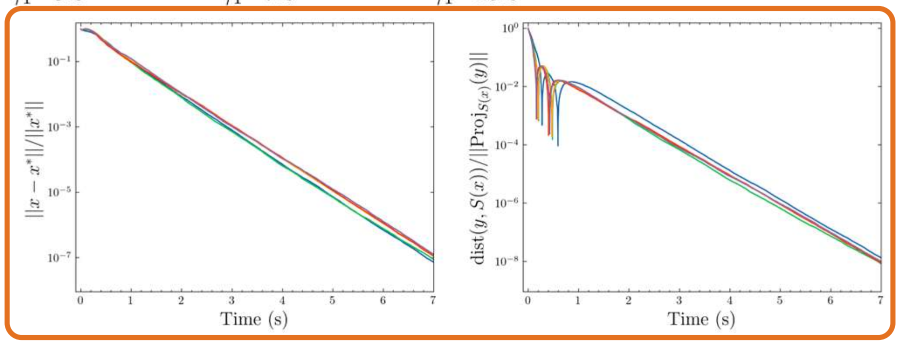




Sensitivity Analysis on Parameters

γ_1=1.0 γ_1=3.0 γ_1=5.0 γ_1=7.0 γ_1=10.0





Hyperparameter Optimization on Sparse Group Lasso Problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\lambda}\in\mathbb{R}_+^{M+1}} \frac{1}{2}\sum_{i\in I_{\mathrm{val}}} |b_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{a}_i|^2$$

$$\text{s.t.} \quad \boldsymbol{\beta} \in \arg\min_{\hat{\boldsymbol{\beta}}\in\mathbb{R}^p} \left\{ \frac{1}{2}\sum_{i\in I_{\mathrm{tr}}} |b_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{a}_i|^2 + \boxed{\sum_{m=1}^{M} \lambda_m \|\hat{\boldsymbol{\beta}}^{(m)}\|_2 + \lambda_{M+1}\|\hat{\boldsymbol{\beta}}\|_1} \right\}.$$

$$\Big\downarrow \ \text{Hyperparameter Decoupling}$$

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}_+^{M+1}} \frac{1}{2}\sum_{i\in I_{\mathrm{val}}} |b_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{a}_i|^2$$

$$\text{s.t.} \quad \boldsymbol{\beta} \in \arg\min_{\hat{\boldsymbol{\beta}}\in\mathbb{R}^p} \left\{ \frac{1}{2}\sum_{i\in I_{\mathrm{tr}}} |b_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{a}_i|^2 \ \text{s.t.} \ \boxed{\|\hat{\boldsymbol{\beta}}^{(m)}\|_2^2 \le u_m, m\in[M], \ \|\hat{\boldsymbol{\beta}}\|_1 \le u_{M+1}} \right\}.$$

Method	nTr = 100, nVal = 100, nTest = 300			nTr = 300, nVal = 300, nTest = 300		
	Time (s)	Val Err	Test Err	Time (s)	Val Err	Test Err
Grid	17.3 ± 0.9	35.9 ± 7.2	37.7 ± 6.7	78.7 ± 1.9	18.9 ± 2.3	19.8 ± 1.8
Random	17.4 ± 0.7	33.6 ± 6.7	35.7 ± 6.2	78.6 ± 2.5	18.7 ± 2.4	19.5 ± 1.9
TPE	16.9 ± 0.7	33.9 ± 7.0	36.0 ± 5.6	74.7 ± 2.2	18.9 ± 2.3	19.8 ± 1.9
IGJO	21.2 ± 2.2	19.7 ± 2.8	25.6 ± 4.4	49.9 ± 2.6	16.5 ± 2.5	18.1 ± 1.4
VF-iDCA	12.4 ± 0.5	14.6 ± 2.6	25.4 ± 3.9	40.7 ± 1.7	14.9 ± 2.1	17.2 ± 1.3
BiC-GAFFA	21.4 ± 0.7	7.3 ± 1.3	22.3 ± 3.0	22.0 ± 1.0	12.8 ± 1.4	17.1 ± 1.3

GAN with or with out Constraint

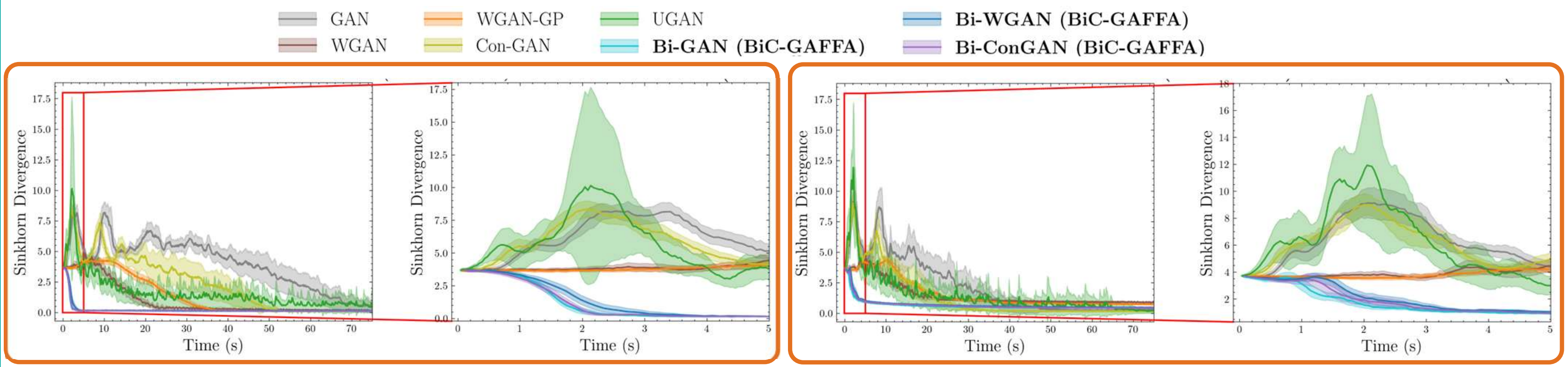
Model: $$\min_{G,D} \mathcal{L}_{\text{gen}}(G,D) \quad \text{s.t.} \quad D \in \operatorname*{argmin}_{D\in\mathcal{D}} \mathcal{L}_{\text{det}}(G,D).$$

WGAN:

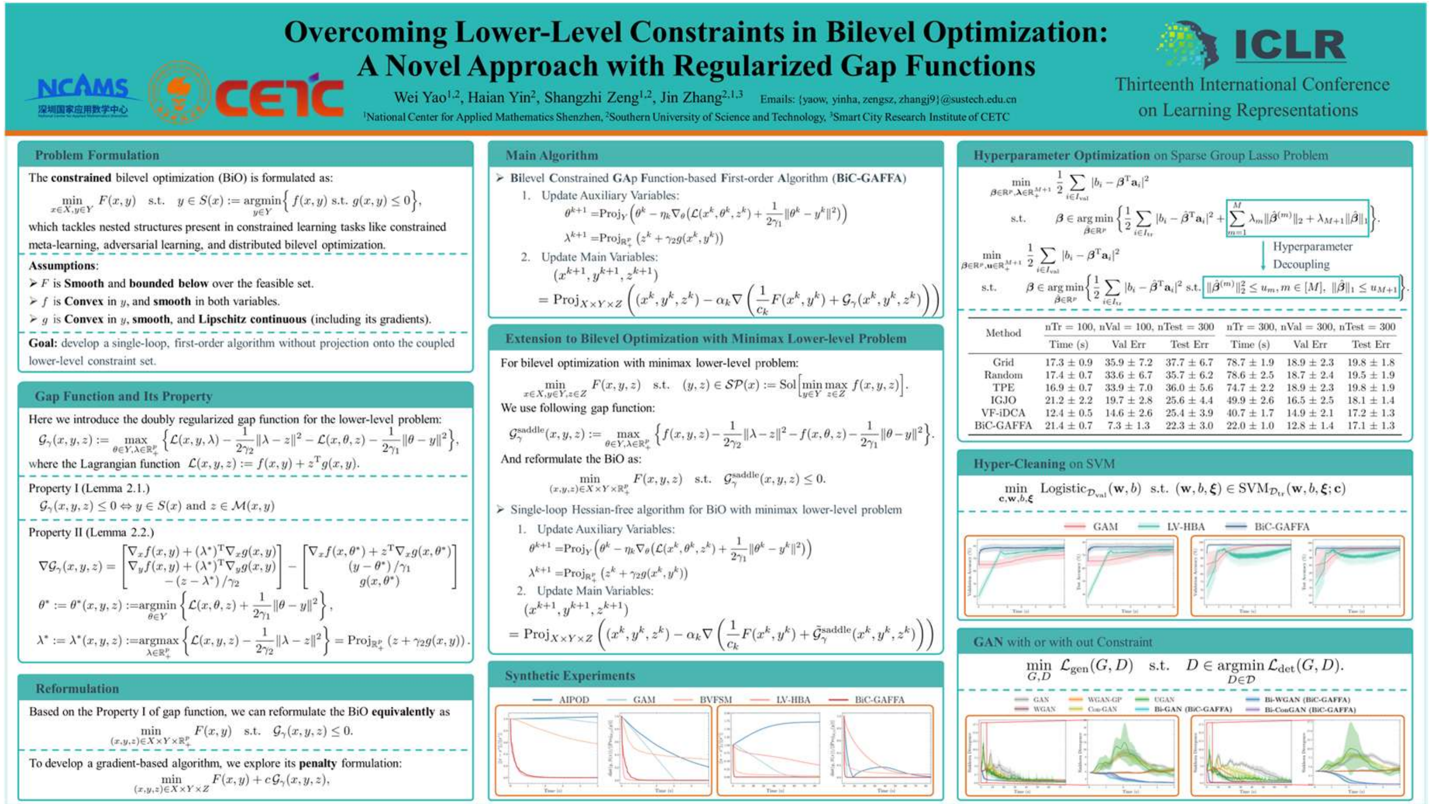
$$\mathcal{L}_{\text{gen}}(G,D) = -\,\mathbb{E}_{\mathbf{z}\sim\mathbb{N}}[D(G(\mathbf{z})],$$

$$\mathcal{L}_{\text{det}}(G,D) = -\,\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_r}[D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim\mathbb{N}}[D(G(\mathbf{z})],$$

$$\mathcal{D} := \left\{ D(\mathbf{x}) \quad \text{s.t.} \quad \max_{\hat{\mathbf{x}}\in\mathbb{P}_{\hat{\mathbf{x}}}} \|\nabla D(\hat{\mathbf{x}})\|_2 \le 1 \right\}.$$



Thanks


Overcoming Lower-Level Constraints in Bilevel Optimization:
A Novel Approach with Regularized Gap Functions
Wei Yao1,2, Haian Yin2, Shangzhi Zeng1,2, Jin Zhang2,1,3
Emails: {yaow, yinha, zengsz, zhangj9}@sustech.edu.cn
1National Center for Applied Mathematics Shenzhen, 2Southern University of Science and Technology, 3Smart City Research Institute of CETC
NCAMS
深圳国家应用数学中心
CETC
ICLR
Thirteenth International Conference on Learning Representations
Problem Formulation
The constrained bilevel optimization (BiO) is formulated as:
which tackles nested structures present in constrained learning tasks like constrained meta-learning, adversarial learning, and distributed bilevel optimization.
Assumptions:
F is Smooth and bounded below over the feasible set.
f is Convex in y, and smooth in both variables.
g is Convex in y, smooth, and Lipschitz continuous (including its gradients).
Goal: develop a single-loop, first-order algorithm without projection onto the coupled lower-level constraint set.
Gap Function and Its Property
Here we introduce the doubly regularized gap function for the lower-level problem:
where the Lagrangian function
Property I (Lemma 2.1.)
Property II (Lemma 2.2.)
Reformulation
Based on the Property I of gap function, we can reformulate the BiO equivalently as
To develop a gradient-based algorithm, we explore its penalty formulation:
Main Algorithm
Bilevel Constrained GAp Function-based First-order Algorithm (BiC-GAFFA)
1. Update Auxiliary Variables:
2. Update Main Variables:
Extension to Bilevel Optimization with Minimax Lower-level Problem
For bilevel optimization with minimax lower-level problem:
We use following gap function:
And reformulate the BiO as:
Single-loop Hessian-free algorithm for BiO with minimax lower-level problem
1. Update Auxiliary Variables:
2. Update Main Variables:
Synthetic Experiments
AIPOD
GAM
BVFSM
LV-HBA
BiC-GAFFA
Hyperparameter Optimization on Sparse Group Lasso Problem
Hyperparameter Decoupling
Method
nTr = 100, nVal = 100, nTest = 300
nTr = 300, nVal = 300, nTest = 300
Time (s)
Val Err
Test Err
Grid
Random
TPE
IGJO
VF-iDCA
BiC-GAFFA
Hyper-Cleaning on SVM
GAM
LV-HBA
BiC-GAFFA
GAN with or with out Constraint