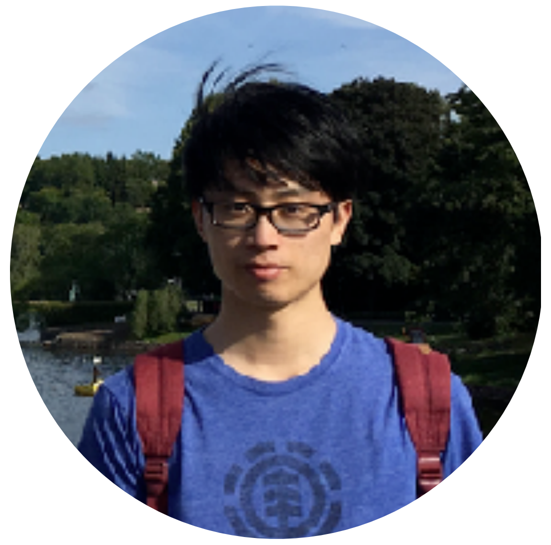
BOME! Bilevel Optimization Made Easy: A Simple First-Order Approach


***Bo Liu[1]** ***Mao Ye[1]** **Stephen Wright[2]** **Peter Stone[1,3]** **Qiang Liu[1]**

[1] The University of Texas at Austin, [2] University of Wisconsin, [3] Sony AI

* indicates equal contribution

Conference on Neural Information Processing Systems (NeurIPS), 2022

Problem

We consider the bilevel optimization (BO) problem:

$$\underbrace{\min_{v,\theta} f(v,\theta)}_{\text{outer problem}} \quad \text{s.t.} \quad \theta \in \underbrace{\arg\min_{\theta'} g(v,\theta')}_{\text{inner problem}}$$

Example (Hyper-parameter Tuning)

In machine learning, we often want to choose the right hyper-parameters v such that the model parameter θ achieves the best performance.

$$\min_{v,\theta} L_{\text{val}}(v,\theta) \quad \text{s.t.} \quad \theta \in \arg\min_{\theta'} L_{\text{train}}(v,\theta)$$

Problem

We consider the bilevel optimization (BO) problem:

$$\underbrace{\min_{v,\theta} f(v,\theta)}_{\text{outer problem}} \quad \text{s.t.} \quad \theta \in \underbrace{\arg\min_{\theta'} g(v,\theta')}_{\text{inner problem}}$$

Challenges in prior approaches:

- **Scalability**: often require computing **2nd order gradient** each iteration.

Problem

We consider the bilevel optimization (BO) problem:

$$\underbrace{\min_{v,\theta} f(v,\theta)}_{\text{outer problem}} \quad \text{s.t.} \quad \theta \in \underbrace{\arg\min_{\theta'} g(v,\theta')}_{\text{inner problem}}$$

Challenges in prior approaches:

- **Scalability**: often require computing **2nd order gradient** each iteration.
- **Theory**: lack convergence result when f, g are non-convex w.r.t. v, θ.

BOME! Method

BO objective: $$\min_{v,\theta} f(v,\theta) \quad s.t. \quad \theta \in \arg\min_{\theta'} g(v,\theta'),$$

General Idea Convert BO into a constrained optimization problem, in which g is required to be less than a certain threshold (ideally its optimal value for the given v). In other words,

*Optimize the outer problem s.t. the **optimality gap** for inner problem is 0*

BOME! Method

BO objective: $$\min_{v,\theta} f(v,\theta) \quad s.t. \quad \theta \in \arg\min_{\theta'} g(v,\theta'),$$

Step 1: Compute the **value function** (the optimality gap of the inner problem for g)

$$q(v,\theta) := g(v,\theta) - g^*(v)$$

$$g^*(v) := \min_\theta g(v,\theta)$$

Unknown

approximate value function

$$\hat{q}(v,\theta) = g(v,\theta) - g(v, \theta_k^{(T)}).$$

Obtained by T-step of gradient,
then **stop-gradient**

BOME! Method

The University of Texas at Austin

BO objective: $\min_{v,\theta} f(v,\theta) \quad s.t. \quad \theta \in \arg\min_{\theta'} g(v,\theta'),$

Step 2: Descent on the **outer** s.t. the **inner** also improves

$$(v_{k+1}, \theta_{k+1}) \leftarrow (v_k, \theta_k) - \xi \delta_k$$

where $\delta_k = \arg\min_{\delta} \underbrace{||\nabla f - \delta||^2}_{\text{descend } f} \quad \text{s.t.} \quad \underbrace{\langle \nabla \hat{q}, \delta \rangle \geq \phi \geq 0}_{\hat{q} \text{ does not ascend}}$

Find an update close to ∇_f

The update shares a positive angle with $\nabla \hat{q}$

BOME! Theory

General Idea Analyze BO from a constrained optimization perspective

Optimality Measure (KKT loss)

$$\mathcal{K}(v,\theta) = \min_{\lambda \geq 0} \underbrace{||\nabla f(v,\theta) + \lambda \nabla q(v,\theta)||^2}_{\text{local improvement}} + \underbrace{q(v,\theta)}_{\text{feasibility}}$$

Key Contribution: we analyze how KKT loss decreases w.r.t. # updates

BOME! Theory

For **smooth** and **non-convex** inner and outer objectives, we have:

Theorem 2. *Consider Algorithm 1 with* $\xi, \alpha \leq 1/L$, $\phi_k = \eta \left\|\nabla \hat{q}(v_k, \theta_k)\right\|^2$, *and* $\eta > 0$. *Suppose that Assumptions 2, 3, and 4 hold and that* $q^\diamond$ *is differentiable on* (v_k, θ_k) *at every iteration* $k \geq 0$. *Then there exists a constant* c *depending on* α, κ, η, L, *such that when* $T \geq c$, *we have*

$$\min_{k \leq K} \mathcal{K}^\diamond(v_k, \theta_k) = O\left(\sqrt{\xi} + \sqrt{\frac{1}{\xi K}} + \exp(-bT)\right),$$

where b *is a positive constant depending on* κ, L, *and* α.

Remark:

- As the inner objective is **non-convex**, the above achieves a rate of $O(K^{-1/4} + \exp(-bT))$
- When inner objective is **convex**, the rate can be improved to $O(K^{-1/3} + \exp(-bT))$

BOME! Summary

TEXAS
The University of Texas at Austin

Improved Scalability

- BOME! is a purely 1st-order method

Good Performance

- Better/comparable accuracy/speed compared with SOTA BO methods

Simplicity

- Easy to implement
- Fewer hyper parameters than prior methods, and is robust to them

BOME! Experiment

Experiments

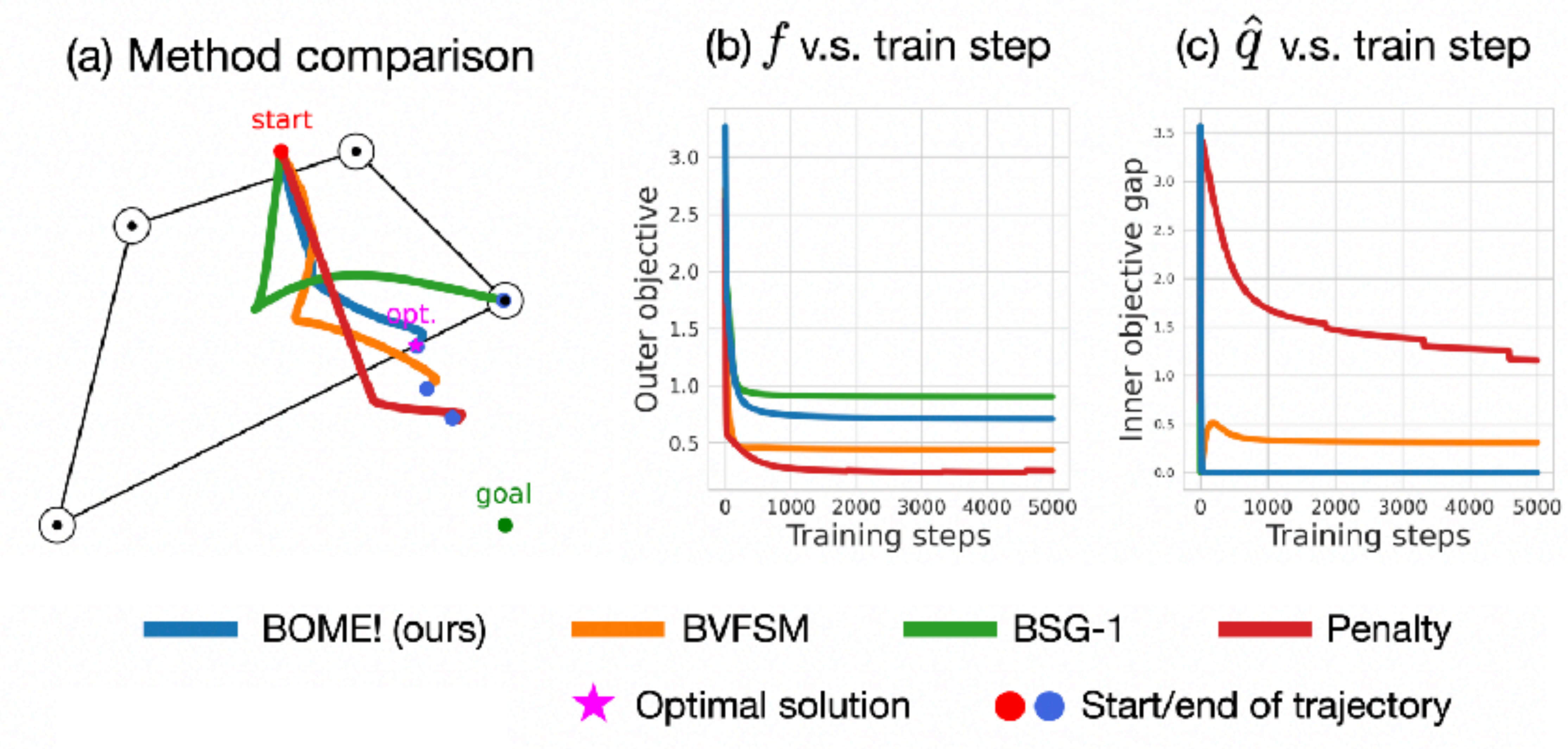
- We conduct experiments on 3 toy examples and 3 BO benchmarks.
- For simplicity, we show result on a toy example.

The Coreset Problem

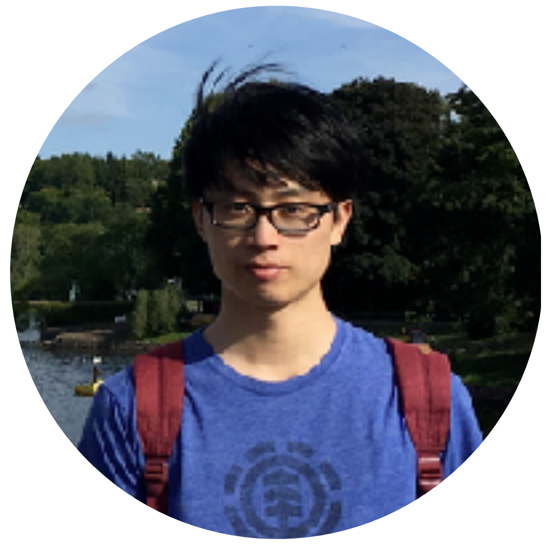
$$\min_{v,\theta} ||\theta - x_0||^2, \quad \text{s.t.} \quad \theta \in \arg\min_{\theta'} ||\theta' - X\sigma(v)||^2$$

$$\sigma(v) = \exp(v) / \sum_{i=1}^{4} \exp(v_i)$$

(i.e., find the closest point in the convex hull of X to x_0)



Thank you!


***Bo Liu[1]** ***Mao Ye[1]** **Stephen Wright[2]** **Peter Stone[1,3]** **Qiang Liu[1]**

[1] The University of Texas at Austin, [2] University of Wisconsin, [3] Sony AI

* indicates equal contribution

Paper Link:

Code Link:

https://github.com/Cranial-XIX/BOME

Conference on Neural Information Processing Systems (NeurIPS), 2022