# LancBiO: Dynamic Lanczos-aided Bilevel Optimization via Krylov Subspace

Yan Yang [†] [§] Bin Gao [†] Ya-xiang Yuan [†]

[†] Academy of Mathematics and Systems Science, CAS
[§] University of Chinese Academy of Sciences

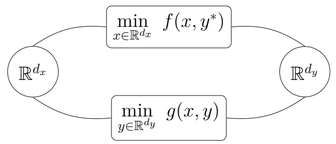
## Bilevel optimization problem (BiO)

**General form**

$$\begin{aligned} \min_{x\in\mathbb{R}^{d_x}} \quad & \varphi(x) := f\left(x, y^*(x)\right) \\ \text{s. t.} \quad & y^*(x) = \arg\min_{y\in\mathbb{R}^{d_y}} g(x,y) \end{aligned}$$

- $f, g: \mathbb{R}^{d_x}\times\mathbb{R}^{d_y} \longrightarrow \mathbb{R}$, differentiable, $g$ is **strongly convex**

The nested structure **couples** the **upper level** and **lower level**




## Approximate implicit differentiation (AID)

**Hyper-gradient computation**

$$\begin{aligned} \nabla\varphi(x) &= \nabla_x f\left(x, y^\star(x)\right) + \nabla_x y^\star(x)^\top \nabla_y f\left(x, y^\star(x)\right) \\ &= \nabla_x f\left(x, y^\star\right) - \nabla^2_{xy} g\left(x, y^\star\right)\left[\nabla^2_{yy} g\left(x, y^\star\right)\right]^{-1} \nabla_y f\left(x, y^\star\right) \end{aligned}$$

**Main difficulties**

- Solving the lower-level problem to obtain $y^*(x)$
- ★ Approximating the **Hessian inverse vector product**

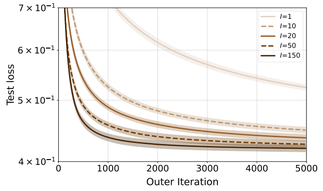
$$v^*(x) := \left[\nabla^2_{yy} g\left(x, y^*(x)\right)\right]^{-1} \nabla_y f\left(x, y^*(x)\right)$$

**Vanilla update rule**

$$\begin{aligned} \mathbf{1}\times: \; & x^+ = x - \beta\left(\nabla_x f\left(x,y\right) - \nabla^2_{xy} g\left(x,y\right)\left[\nabla^2_{yy} g\left(x,y\right)\right]^{-1}\nabla_y f\left(x,y\right)\right) \\ \mathbf{N}\times: \; & y^+ = y - \alpha\nabla_y g\left(x,y\right) \end{aligned}$$

## Motivation

**An observation**




**The more accurate $v^*$, the more enhanced descent!**

**How to tackle $v^*$ in BiO**

- **Approximation** principle
- **Amortization** principle

**How to adhere to these two principles**

- Subspace methods for **efficient Approximation** [Yuan, 2014]
- Subspace iteration for **reasonable Amortization** [Yuan, 1995]

## A subspace perspective

**Geometric interpretation**

$$\min_{v\in\mathcal{S}_k} \; m_k(v) := \frac{1}{2} v^\top \nabla^2_{yy} g\left(x_k, y_k\right) v - \nabla_y f\left(x_k, y_k\right)^\top v$$

**Low-dimensional subspace?**

$$\mathcal{S}_k = \mathbb{R}^{d_y} \Longrightarrow v_k = \nabla^2_{yy} g\left(x_k, y_k\right)^{-1} \nabla_y f\left(x_k, y_k\right) := A_k^{-1} b_k$$

**Krylov subspace** [Krylov, 1931] provides a good estimate for $A^{-1}b$ [Carmon and Duchi, 2018]

$$\mathcal{K}_N(A,b) := \operatorname{span}\left\{b, Ab, A^2b, \ldots A^{N-1}b\right\}$$

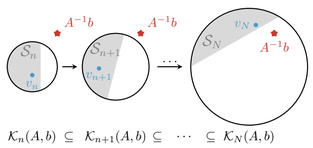
## Constructing 2-D subspace: SubBiO

Given $v_n \in \mathcal{K}_n(A,b)$, an initial approximation of $A^{-1}b$

$$v_n = \sum_{i=0}^{n-1} c_i \left(I - \eta A\right)^i b \approx A^{-1} b$$

Recursively, we can choose

$$v_{n+1} \in \mathcal{S}_{n+1} := \operatorname{span}\left\{b, (I-\eta A)v_n\right\} \subseteq \mathcal{K}_{n+1}(A,b)$$




**SubBiO**

$$\begin{aligned} \mathbf{1}\times: \; & \mathcal{S}_k := \operatorname{span}\left\{b_k, (I-\eta A_k)v_{k-1}\right\} \\ \mathbf{1}\times: \; & v_k := \arg\min_{v\in\mathcal{S}_k} \; m_k(v) := \frac{1}{2} v^\top A_k v - b_k v \\ \mathbf{1}\times: \; & x_{k+1} = x_k - \beta\left(\nabla_x f\left(x_k, y_k\right) - \nabla^2_{xy} g\left(x_k, y_y\right) v_k\right) \\ \mathbf{1}\times: \; & y_{k+1} = y - \alpha\nabla_y g\left(x_k, y_k\right) \end{aligned}$$

**Computational complexity**

- $\min_{z\in\mathbb{R}^2} \frac{1}{2} z^\top (S_k^\top A_k S_k) z - b_k^\top S_k z$: $O(\mathbf{2}n^2)$
- $A_k v_{k-1}$: $O(n^2)$, $\nabla^2_{xy} g(x_k, y_k) v_k$: $O(n^2)$

**SubBiO costs $O(\mathbf{4}n^2)$ per iteration**

## Dynamic Lanczos process in BiO

**Core principles**

- Maintain an **orthogonal basis** $Q_j = [q_1, \ldots, q_j]$ approximating $\mathcal{K}_j(A_j, b_j)$, $\mathcal{S}_k = \operatorname{span}(Q_j)$
- Keep the (approximate) projection matrix $T_j$ **tridiagonal**
- Dynamically solve quadratic subproblems: $v_k := \arg\min_{v\in\mathcal{S}_k} \; m_k(v) := \frac{1}{2} v^\top A_k v - b_k^\top v$

**Adapt standard Lanczos process in BiO**

$$\begin{aligned} & u_j = A_j q_j - \beta_j q_{j-1}, \qquad \alpha_j = q_j^\top u_j \qquad \omega_j = u_j - \alpha_j q_j \\ & \beta_{j+1} = \|\omega_j\| \qquad\qquad q_{j+1} = \omega_j / \beta_{j+1} \end{aligned} \qquad T_j = \begin{pmatrix} T_{j-1} & \begin{matrix} \mathbf{0} \\ \beta_j \end{matrix} \\ \begin{matrix} \mathbf{0} & \beta_j \end{matrix} & \alpha_j \end{pmatrix}.$$

**Lanczos process is inherently unstable** [Paige, 1980; Meurant and Strakoš, 2006]

## Two "Res" strategies: LancBiO

**Restart mechanism**

- Restart subspaces each $m$ steps
- Mitigate the accumulation of difference among $\{A_1, \ldots, A_k\}$

**Residual minimization**

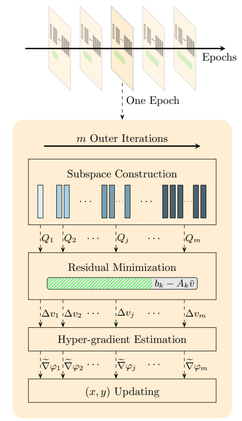
- Minimal residual subproblems

$$\min_{\Delta v\in\mathcal{S}_k} \left\|\left(b_k - A_k\bar{v}\right) - A_k \Delta v\right\|^2$$

- Correct current $\bar{v}$, $v_k = \bar{v} + \Delta v_k$
- Collect historical information

**Computational complexity**

- ☺ Low-dimensional subproblems
- ☺ No cost of Hessian projection
- ☺ LancBiO needs $O(\mathbf{2}n^2)$ instead of $O(\mathbf{4}n^2)$




## Theoretical analysis

**Proposition**

The dynamic Lanczos process in LancBiO with normalized $q_1$ and $\alpha_j, \beta_j, q_j$ satisfies

$$A_j^* Q_j = Q_j T_j + \beta_{j+1} q_{j+1} e_j^\top + \delta Q_j, \quad \text{for } j = 1, 2, \ldots, m, \quad \text{with} \quad \|\delta q_j\| \le L_{gyy}\varepsilon_j$$

**Theorem**

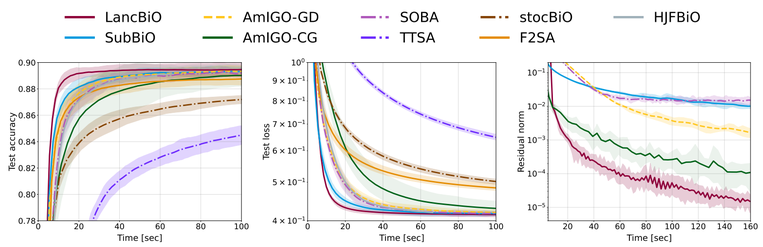
With appropriate step sizes $\theta$ for $y$ and $\lambda$ for $x$, the iterates $\{x_k\}$ generated by LancBiO satisfy

$$\frac{m}{K\left(m - m_0\right)} \sum_{\substack{k=0, \\ (k \bmod m) > m_0}}^{K} \left\|\nabla\varphi\left(x_k\right)\right\|^2 = \mathcal{O}\left(\frac{m\lambda^{-1}}{K\left(m - m_0\right)}\right)$$

where $m$ is the subspace dimension and $m_0 \sim \Omega(\log m)$

## Numerical Experiment

**Test on MNIST (pollution rate=0.8)**




More experiments are referred to our paper...