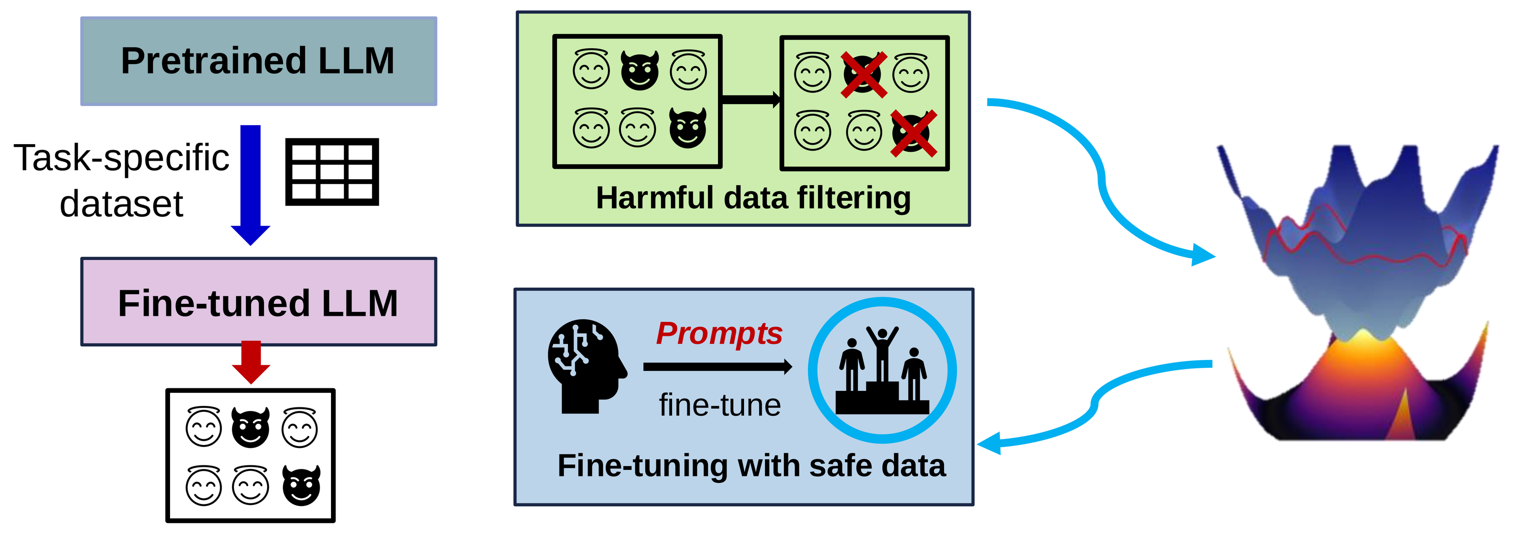
Unlocking Global Optimality in Bilevel Optimization: A Pilot Study

Quan Xiao, Tianyi Chen
xiaoq5@rpi.edu, chent18@rpi.edu
[1]Rensselaer Polytechnic Institute

Paper

1. Context & Motivation: Bilevel Optimization



Managing two hierarchic problems simultaneously

$$\min_{x,y\in\mathcal{S}(x)} \; f(x,y) \quad \text{(upper level)}$$
$$\text{s.t.} \quad \mathcal{S}(x)=\arg\min_{y} \; g(x,y) \quad \text{(lower level)}$$

x : safety score
y: model weight

Existing theoretical guarantees:

Stationary convergence.

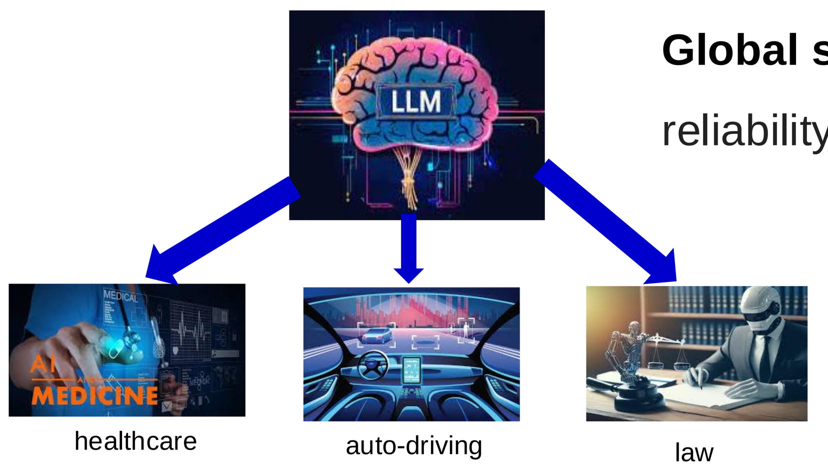
Second order information: SGD-based [Ghadimi & Wang, 18], [Hong et al, 23], [Ji et al, 21], [Chen et al, 21]; momentum-based [Khanduri et al, 21], [Yang et al, 21], [Dagreou et al, 22]; warm-started strategy [Li et al, 22], [Arbel et al, 21], etc

Fully first order information: Strongly convex lower-level [Kwon et al, 23], [Chen et al, 23]; convex lower-level [Mehra et al, 21], [Lu et al, 23]; nonconvex lower-level [Shen et al, 23], [Kwon et al, 24], [Chen et al, 23], [Xiao et al, 23], etc

Local minimum convergence.

Adding random uniform noise helps escape saddle points [Huang et al, 23, Chen et al, 23]

Importance of global convergence:



Global solution in high-stakes field:
reliability, safety, and cost-effectiveness

NP hard in general. [Vicente et al, 94]

Computationally heavy or special problem setting. [Gümüş et al, 01], [Jeyakumar et al, 16], [Muu et al, 03], [Wang et al, 21, 22], …

Challenges of global convergence analysis:

$f(x,y)$ convex & non-decreasing

$g(x,y)$ convex → $\mathcal{S}(x)$ not necessarily convex

$F(x)=\min_{y\in\mathcal{S}(x)} f(x,y)$ → $F(x)$ not necessarily convex

- **Landscape of nested approach is complicated** → **constrained approach**

Approaches: Nested optimization & Constrained optimization

value function
$g^*(x) := \min_y g(x,y)$

$$\min_x F(x) \quad \text{hyper-function}$$
$$F(x)=\min_{y\in\mathcal{S}(x)} f(x,y)$$

$$\min_{x,y\in\mathcal{S}(x)} f(x,y) \quad \text{s.t.} \quad g(x,y)-g^*(x)\le 0$$

2. Penalty Based Global Analysis

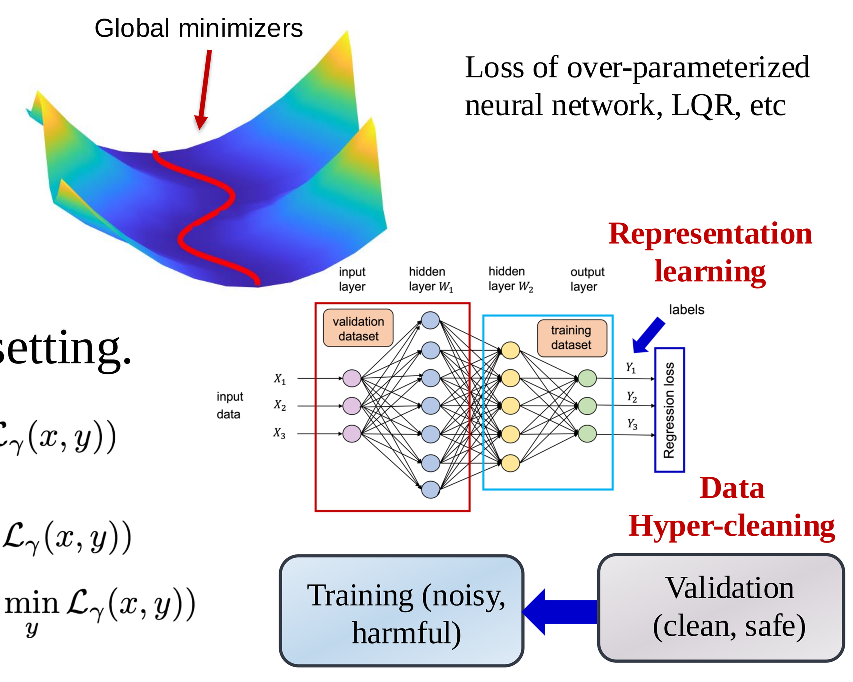
$$\min_{x,y\in\mathcal{S}(x)} f(x,y) \quad \text{(upper level)}$$
$$\text{s.t.} \quad g(x,y)-g^*(x)\le 0 \quad \text{(lower level)}$$

$\gamma \ge \epsilon^{-0.5}$ **Penalty problem**

$$\min_{x,y} \mathcal{L}_\gamma(x,y) := f(x,y)+\gamma(g(x,y)-g^*(x))$$

- **Polyak-Łojasiewicz (PL) condition**

$$\|\nabla F(x)\|^2 \ge F(x)-\min_x F(x)$$



❑ Generalize PL condition to two-variable setting.

- **Joint PL.** $\|\nabla \mathcal{L}_\gamma(x,y)\|^2 \ge 2\mu_l(\mathcal{L}_\gamma(x,y)-\min_{x,y}\mathcal{L}_\gamma(x,y))$
- **Blockwise PL.** $\|\nabla_x \mathcal{L}_\gamma(x,y)\|^2 \ge 2\mu_x(\mathcal{L}_\gamma(x,y)-\min_x\mathcal{L}_\gamma(x,y))$
  and $\|\nabla_y \mathcal{L}_\gamma(x,y)\|^2 \ge 2\mu_y(\mathcal{L}_\gamma(x,y)-\min_y\mathcal{L}_\gamma(x,y))$

❑ Additivity of PL functions.

$$\min_{x,y} \mathcal{L}_\gamma(x,y) := \underbrace{\boxed{f(x,y)}}_{\text{joint PL or blockwise PL}} + \gamma\underbrace{\boxed{g(x,y)-g^*(x)}}_{\text{joint PL \& blockwise PL}}$$

$\mathcal{L}_\gamma(x,y)$ is joint PL or blockwise PL if

- the PL condition of $f(x,y)$ and $g(x,y)-g^*(x)$ has additivity
- Convex function has additivity, but not all PL functions have additivity

Toy example

$$\min_{x,y} \; f(x,y)=\frac{1}{2}(x-\sin(y))^2,$$
$$\text{s.t.} \; y\in\arg\min_y g(x,y)=\frac{1}{2}(x-y)^2$$

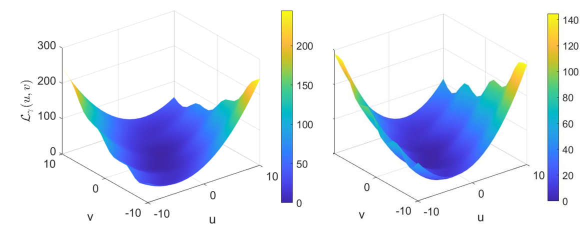


Figure 1: Landscape of penalty objective with $\gamma=0.5, \gamma=1$ from left to right.

4. Theoretical Analysis & Case Study

❑ Additivity of PL functions: A special case

Theorem (Additivity of PL functions)

Linear composited with strongly convex functions $h_1(Az), h_2(Bz)$ are PL functions. Moreover, the addition of them $h_1(Az)+h_2(Bz)$ is PL function.

PL functions – linear composited with strong convex function has additivity

❑ Data hyper-cleaning: Blockwise PL condition

Linear model

$$\min_{u\in\mathcal{U},\, W\in\mathcal{S}(u)} \frac{1}{2}\|Y_{\text{val}}-X_{\text{val}}W\|^2 \quad \text{s.t.} \quad \mathcal{S}(u)=\arg\min_W \frac{1}{2}\left\|\sqrt{\psi_N(u)}\,(Y_{\text{trn}}-X_{\text{trn}}W)\right\|^2.$$

Lemma (informal)

If training data is linearly independent, $\mathcal{L}_\gamma(u,W)$ is blockwise PL over the PBGD-B trajectory.

PBGD-B converges almost linearly to the global optimal solution of the data hyper-cleaning

❑ Representation learning: Joint PL condition

Linear model

$$\min_{W_1,W_2\in\mathcal{S}(W_1)} \frac{1}{2}\|Y_{\text{val}}-X_{\text{val}}W_1W_2\|^2, \quad \text{s.t.} \quad \mathcal{S}(W_1)=\arg\min_{W_2}\frac{1}{2}\|Y_{\text{trn}}-X_{\text{trn}}W_1W_2\|^2$$

Lemma (informal)

For overparameterized setting, $\mathcal{L}_\gamma(u,v)$ is joint PL over the PBGD-J trajectory.

PBGD-J converges almost linearly to the global optimal solution of the representation learning

3. Penalty Based Algorithms

PBGD-J: Penalty Bilevel Gradient Descent (Joint PL)

For k = 0, 1, 2, …, K do
- **S1)** $z_k^T = T$ step GD updates on $g(x_k,y)$
- **S2)** $y_{k+1}=y_k-\alpha(\nabla_y f(x_k,y_k)+\gamma\nabla_y g(x_k,y_k))$
- **S3)** $x_{k+1}=x_k-\alpha(\nabla_x f(x_k,y_k)+\gamma(\nabla_x g(x_k,y_k)-\nabla_x g(x_k,z_k^T)))$

PBGD-B: Penalty Bilevel Gradient Descent (Blockwise PL)

For k = 0, 1, 2, …, K do
- **S1)** $z_k^T = T$ step GD updates on $g(x_k,y)$
- **S2)** $y_k^T = T$ step GD updates on $f(x_k,y)+\gamma g(x_k,y)$
- **S3)** $x_{k+1}=x_k-\alpha(\nabla_x f(x_k,y_k)+\gamma(\nabla_x g(x_k,y_k^T)-\nabla_x g(x_k,z_k^T)))$

Theorem (almost linear convergence to global optimum)

Under either joint PL or blockwise PL conditions and lower-level PL condition, consider running PBGD-J or PBGD-B for $k=1,2,\ldots,K$. With small enough step sizes and $T_k \gtrsim \log\epsilon^{-1}$, it holds

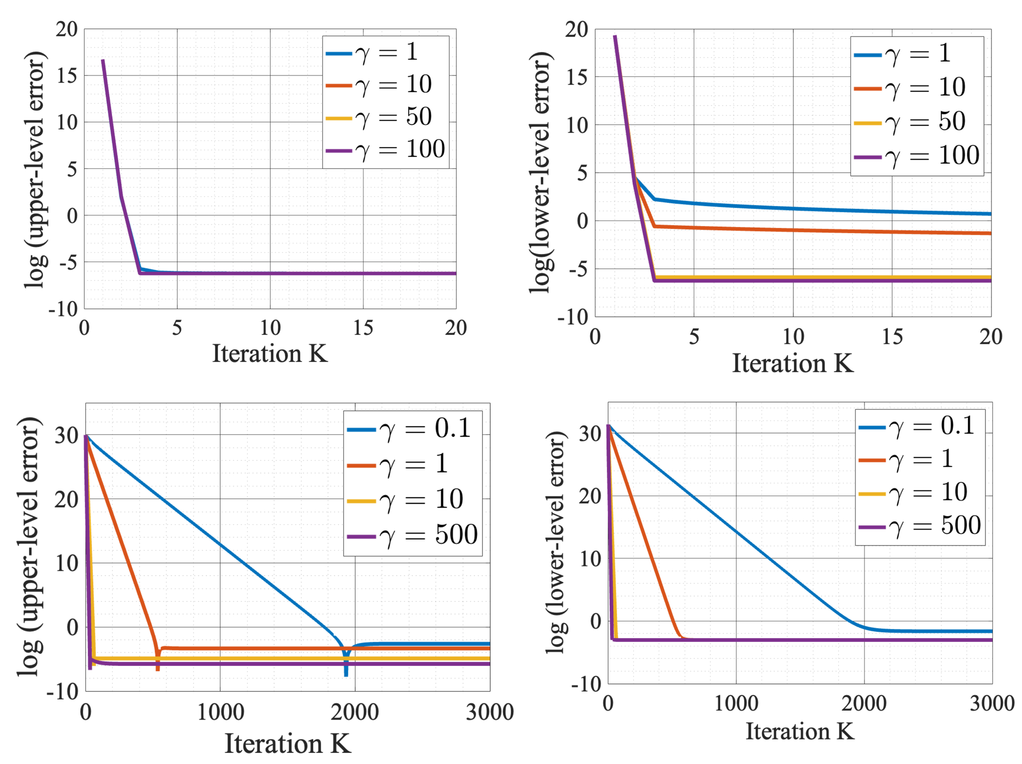
$$\mathcal{L}_\gamma(x_K,y_K)-\min_{x,y}\mathcal{L}_\gamma(x,y) \le (1-\alpha\mu_l)^K\left(\mathcal{L}_\gamma(x_0,y_0)-\min_{x,y}\mathcal{L}_\gamma(x,y)\right)+\mathcal{O}(\epsilon)$$

- With $\gamma \gtrsim \epsilon^{-0.5}$, it implies the $\mathcal{O}((\log\epsilon^{-1})^2)$ iterations to global optimum

5. Empirical Validation and Future Direction

Data hyper-cleaning

Representation learning



PBGD-B/J converges almost linearly to the global optimal solution!

Future Direction:

❑ Extend the global analysis to neural network model

❑ Investigate more bilevel coupling structures