On Implicit Bias in Overparameterized Bilevel Optimization

Paul Vicol, Jonathan Lorraine, Fabian Pedregosa, David Duvenaud, Roger Grosse





UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Google

Bilevel Optimization (BLO)

$$\underbrace{\mathbf{u}^*}_{\text{Outer parameters}} \in \text{“}\arg\min_{\mathbf{u}\in\mathcal{U}}\text{”} \underbrace{F(\mathbf{u}, \mathbf{w}^*)}_{\text{Outer objective}} \quad \textit{such that} \quad \underbrace{\mathbf{w}^*}_{\text{Inner parameters}} \in \mathcal{S}(\mathbf{u}^*) = \arg\min_{\mathbf{w}\in\mathcal{W}} \underbrace{f(\mathbf{u}^*, \mathbf{w})}_{\text{Inner objective}}$$

- **Examples:** *hyperparameter optimization, meta-learning, GANs, dataset distillation, etc.*

Bilevel Optimization (BLO)

$$\mathbf{u}^* \in \text{``}\arg\min_{\mathbf{u}\in\mathcal{U}}\text{''} F(\mathbf{u}, \mathbf{w}^*) \quad \textit{such that} \quad \mathbf{w}^* \in \mathcal{S}(\mathbf{u}^*) = \arg\min_{\mathbf{w}\in\mathcal{W}} f(\mathbf{u}^*, \mathbf{w})$$

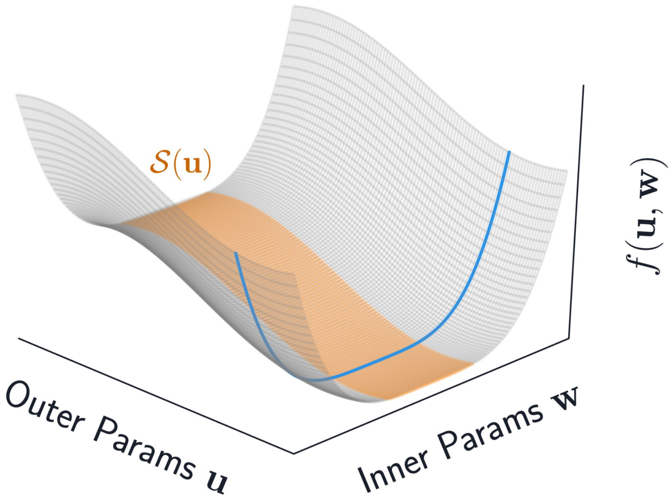
Outer parameters *Outer objective* *Inner parameters* *Inner objective*

- **Examples:** *hyperparameter optimization, meta-learning, GANs, dataset distillation, etc.*
- **Theory:** Typically *assumes that the solutions to the inner/outer objectives are unique*
- **Practice:** The inner and/or outer problems are often *underspecified*
  - There is a *manifold of optima*
  - The optimization dynamics can lead to implicit bias

➡ *Which of the many solutions do we obtain with common algorithms in practice?*

Inner and Outer Underspecification

Inner Underspecification

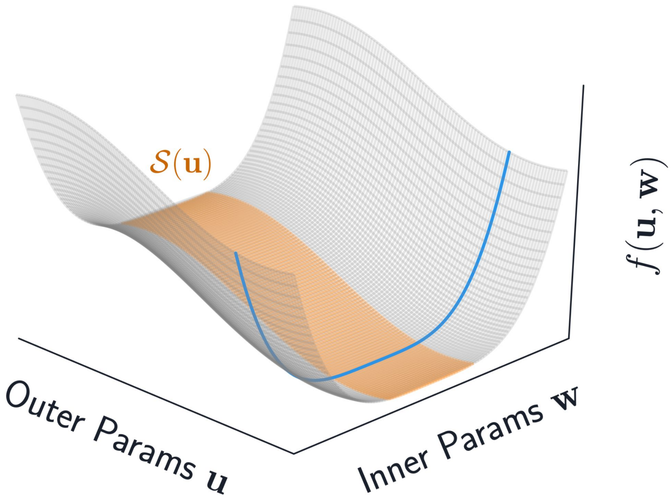


A manifold of optimal inner solutions for each outer parameter, $\mathcal{S}(\mathbf{u})$

Most BLO tasks in ML train a neural net in the inner level, which often yields an underspecified problem

Inner and Outer Underspecification

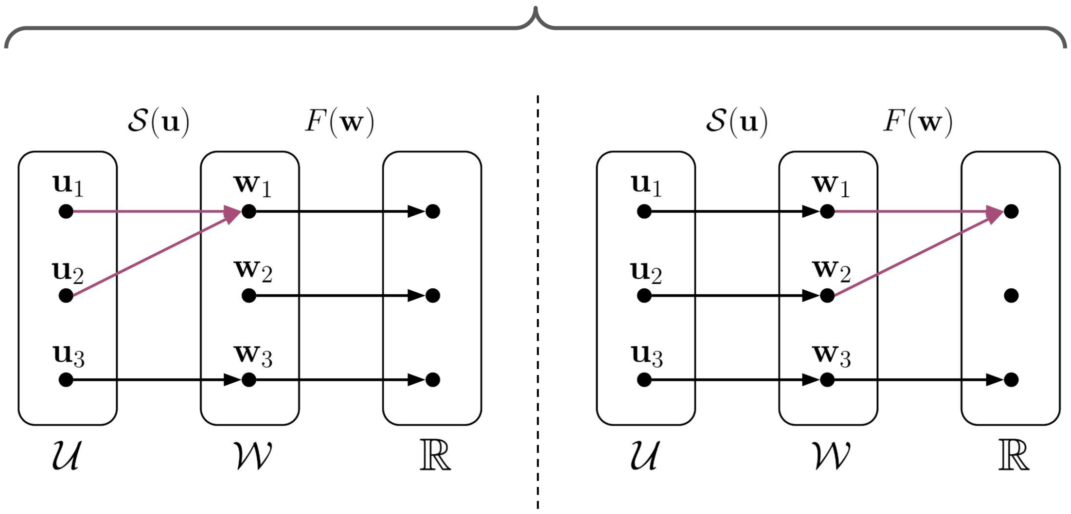
Inner Underspecification



A manifold of optimal inner solutions for each outer parameter, $\mathcal{S}(\mathbf{u})$

Most BLO tasks in ML train a neural net in the inner level, which often yields an underspecified problem

Outer Underspecification



The mapping $\mathcal{S}(\mathbf{u})$ is a function (e.g., not set-valued) and maps a range of outer parameters to the same inner parameter

F maps a range of inner parameters to the same objective value

Sources of Implicit Bias

- We consider *gradient-based BLO,* which requires the outer gradient $\frac{dF(\mathbf{u}, \mathbf{w}^\star(\mathbf{u}))}{d\mathbf{u}}$

1 The Bilevel Optimization Algorithm – Cold-Start vs Warm-Start

- **Cold-start:** re-initialize $\mathbf{w}$ and run inner optimization to convergence for each hypergradient computation

- **Warm-start:** jointly optimize $\mathbf{w}$ and $\mathbf{u}$ in an *online fashion*, e.g., alternating gradient steps with their respective objectives

Sources of Implicit Bias

- We consider *gradient-based BLO,* which requires the outer gradient $\frac{dF(\mathbf{u}, \mathbf{w}^\star(\mathbf{u}))}{d\mathbf{u}}$

1 The Bilevel Optimization Algorithm – Cold-Start vs Warm-Start

- **Cold-start:** re-initialize $\mathbf{w}$ and run inner optimization to convergence for each hypergradient computation
- **Warm-start:** jointly optimize $\mathbf{w}$ and $\mathbf{u}$ in an *online fashion*, e.g., alternating gradient steps with their respective objectives

2 The Hypergradient Approximation

- Computing the exact hypergradient is usually *intractable*
- Using *truncated unrolling* or *truncated implicit differentiation* is common

Warm-Start vs Cold-Start

$$\underbrace{\mathbf{u}^\star}_{\text{Learned datapoints}} \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \underbrace{\mathcal{D}_{\text{original}}}_{\text{Loss on the original dataset}})$$

$$\underbrace{\mathbf{w}^\star(\mathbf{u})}_{\text{Model params}} \in \overbrace{\arg\min_{\mathbf{w}} \underbrace{\mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}_{\text{"Training" loss on the synthetic data}}}^{\text{Training a model on the synthetic data}}$$

Warm-Start vs Cold-Start

$$\underbrace{\mathbf{u}^\star}_{\text{Learned datapoints}} \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \underbrace{\mathcal{D}_{\text{original}}}_{\text{Loss on the original dataset}})$$

Learned datapoints

Loss on the original dataset

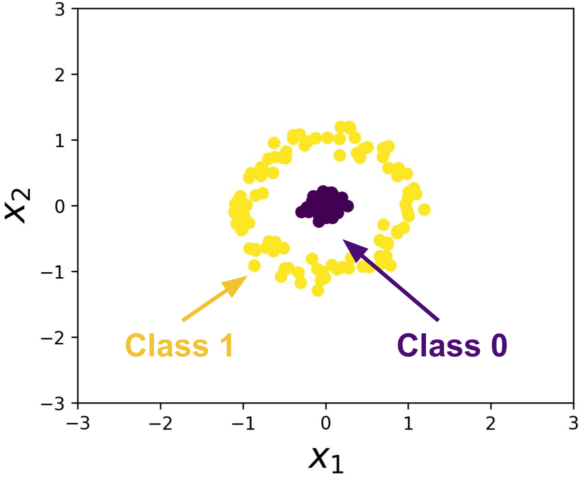
Training a model on the synthetic data

$$\mathbf{w}^\star(\mathbf{u}) \in \overbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}^{\text{Training a model on the synthetic data}}$$

Model params

"Training" loss on the synthetic data

Original dataset

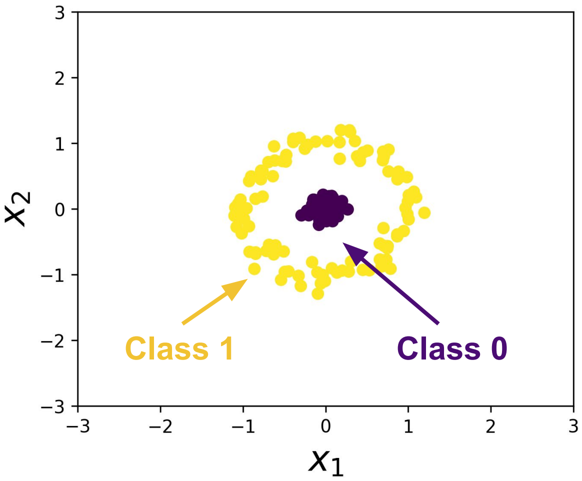


Warm-Start vs Cold-Start

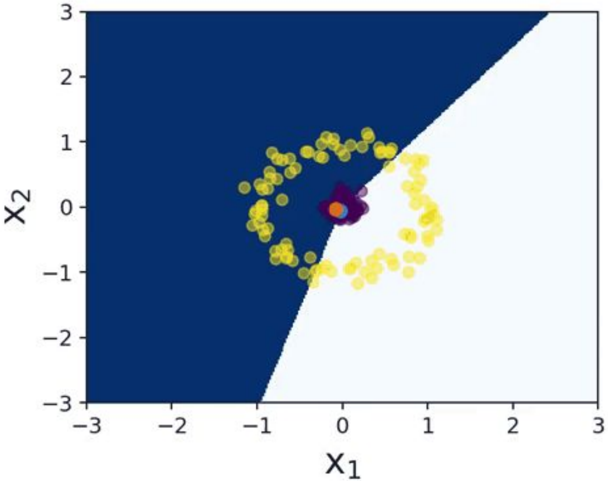
$$\underbrace{\mathbf{u}^\star}_{\text{Learned datapoints}} \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \underbrace{\mathcal{D}_{\text{original}})}_{\text{Loss on the original dataset}}$$

$$\underbrace{\mathbf{w}^\star(\mathbf{u})}_{\text{Model params}} \in \overbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \underbrace{\mathcal{D}(\mathbf{u}))}_{\text{“Training” loss on the synthetic data}}}^{\text{Training a model on the synthetic data}}$$

Original dataset



Warm-start joint optimization



Warm-Start vs Cold-Start

Training a model on the synthetic data

$$\mathbf{u}^\star \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \mathcal{D}_{\text{original}}) \qquad \mathbf{w}^\star(\mathbf{u}) \in \underbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}$$

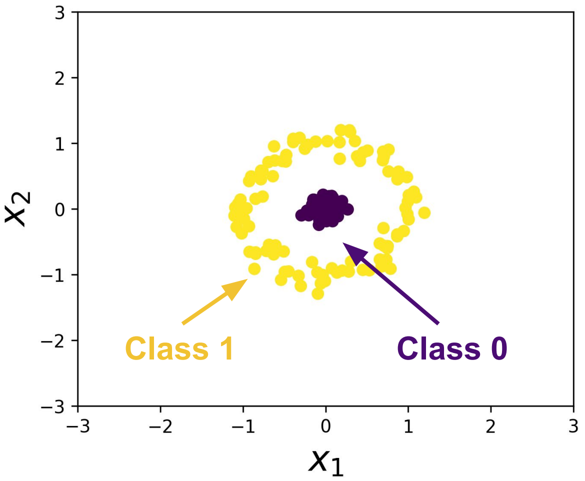
Learned datapoints

Loss on the original dataset

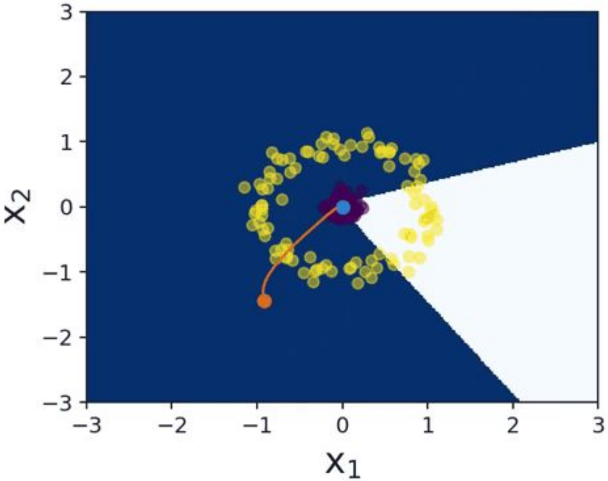
Model params

"Training" loss on the synthetic data

Original dataset



Warm-start joint optimization



Warm-Start vs Cold-Start

$$\mathbf{u}^\star \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \mathcal{D}_{\text{original}})$$

Learned datapoints

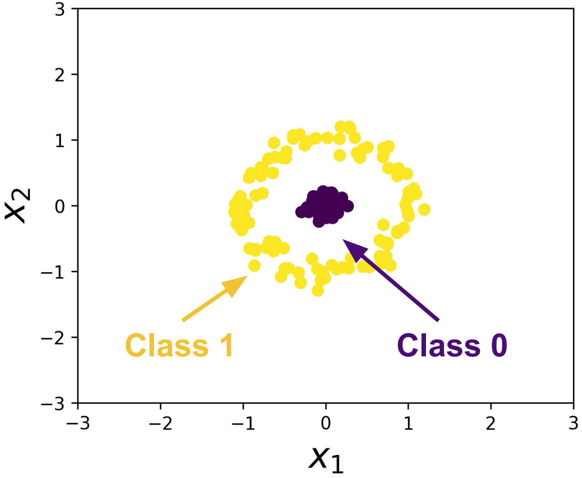
Loss on the original dataset

$$\mathbf{w}^\star(\mathbf{u}) \in \overbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}^{\text{Training a model on the synthetic data}}$$

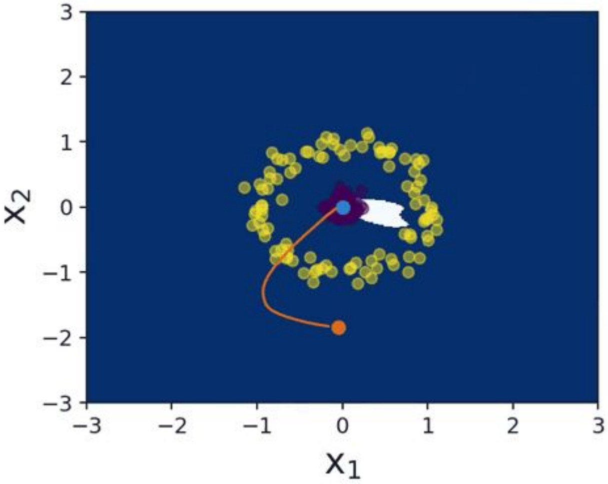
Model params

“Training” loss on the synthetic data

Original dataset



Warm-start joint optimization

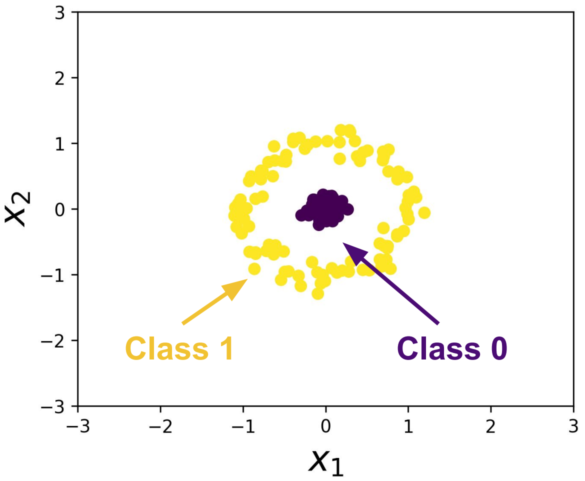


Warm-Start vs Cold-Start

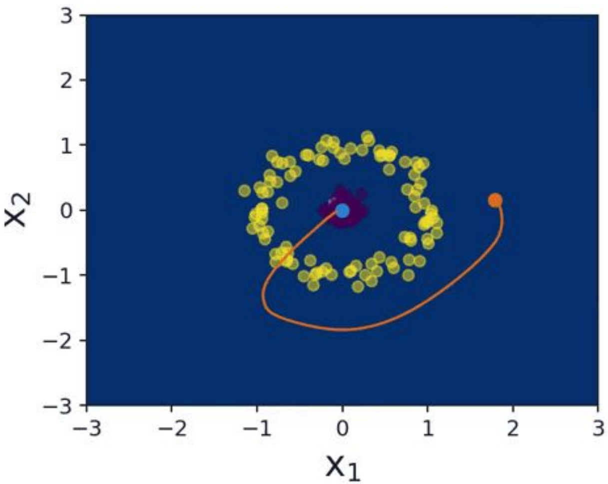
$$\underbrace{\mathbf{u}^\star}_{\text{Learned datapoints}} \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \underbrace{\mathcal{D}_{\text{original}}}_{\text{Loss on the original dataset}})$$

$$\underbrace{\mathbf{w}^\star(\mathbf{u})}_{\text{Model params}} \in \overbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \underbrace{\mathcal{D}(\mathbf{u})}_{\text{“Training” loss on the synthetic data}})}^{\text{Training a model on the synthetic data}}$$

Original dataset



Warm-start joint optimization



Warm-Start vs Cold-Start

$$\mathbf{u}^\star \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \mathcal{D}_{\text{original}}) \qquad \mathbf{w}^\star(\mathbf{u}) \in \underbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}^{\text{Training a model on the synthetic data}}$$

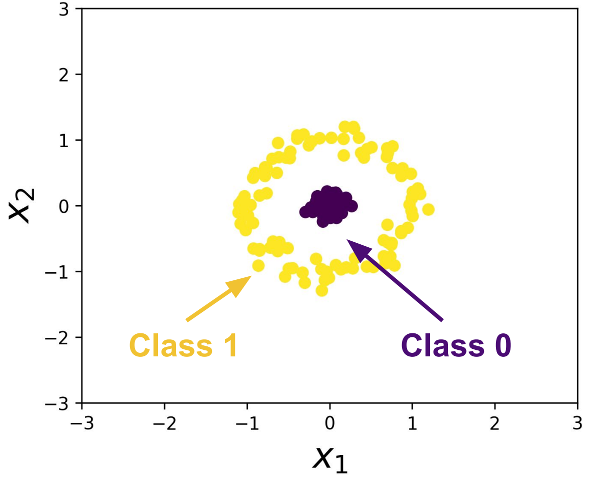
Learned datapoints

Loss on the original dataset

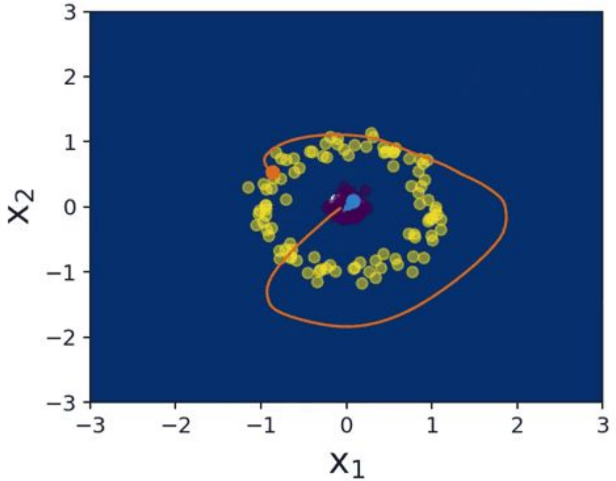
Model params

"Training" loss on the synthetic data

Original dataset



Warm-start joint optimization



Warm-Start vs Cold-Start

$$\mathbf{u}^\star \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \mathcal{D}_{\text{original}})$$

Learned datapoints

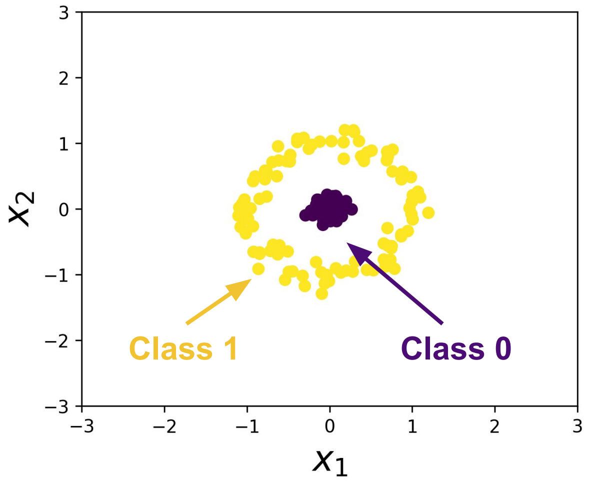
Loss on the original dataset

$$\mathbf{w}^\star(\mathbf{u}) \in \underbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}_{\text{Training a model on the synthetic data}}$$

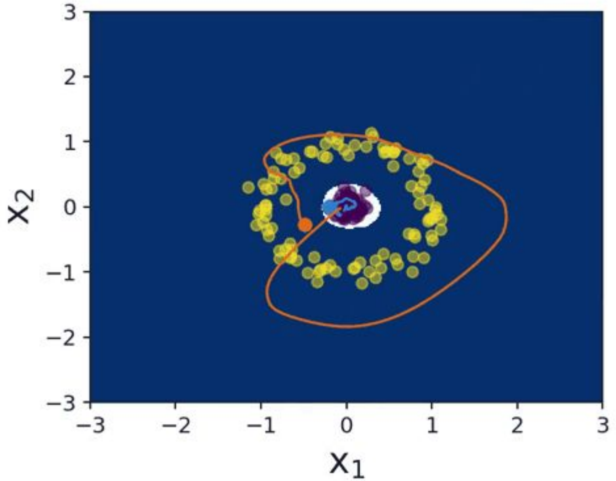
Model params

“Training” loss on the synthetic data

Original dataset



Warm-start joint optimization



Warm-Start vs Cold-Start

$$\mathbf{u}^\star \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \mathcal{D}_{\text{original}})$$

Learned datapoints

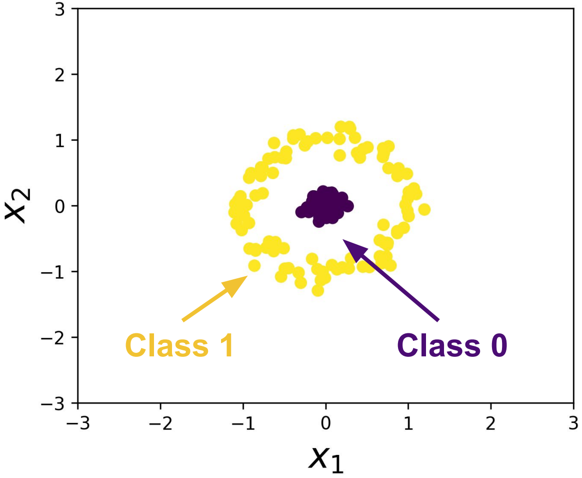
Loss on the original dataset

$$\mathbf{w}^\star(\mathbf{u}) \in \overbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}^{\text{Training a model on the synthetic data}}$$

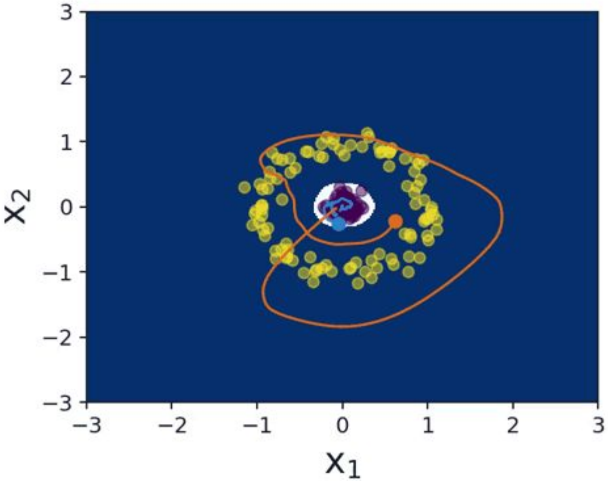
Model params

“Training” loss on the synthetic data

Original dataset



Warm-start joint optimization

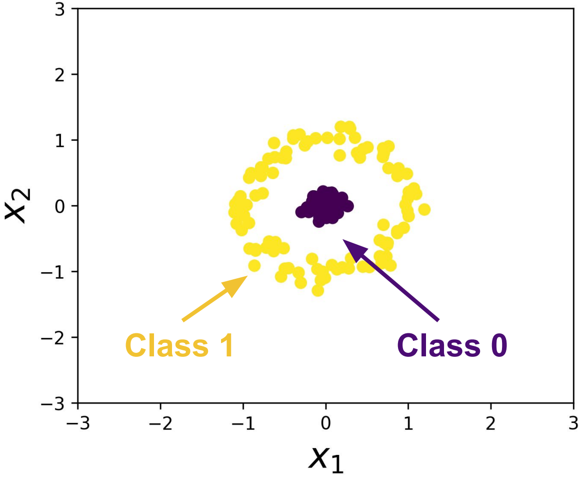


Warm-Start vs Cold-Start

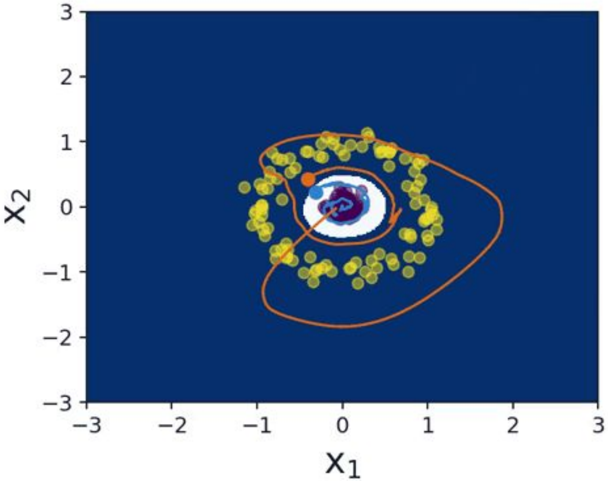
$$\underset{\text{Learned datapoints}}{\mathbf{u}^\star} \in \arg\min_{\mathbf{u}} \underbrace{\mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \mathcal{D}_{\text{original}})}_{\text{Loss on the original dataset}}$$

$$\underset{\text{Model params}}{\mathbf{w}^\star(\mathbf{u})} \in \overbrace{\arg\min_{\mathbf{w}} \underbrace{\mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}_{\text{"Training" loss on the synthetic data}}}^{\text{Training a model on the synthetic data}}$$

Original dataset



Warm-start joint optimization

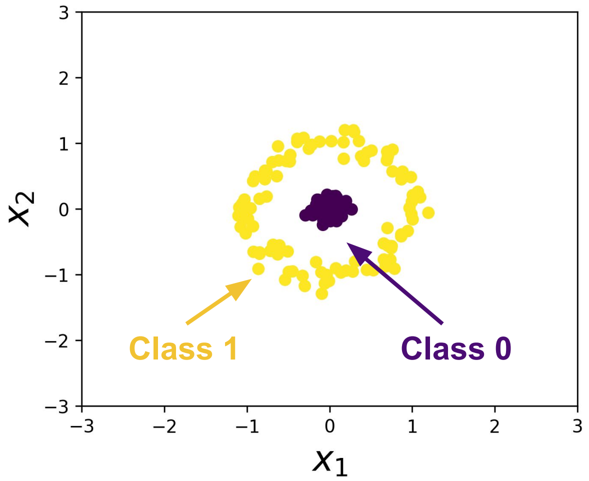


Warm-Start vs Cold-Start

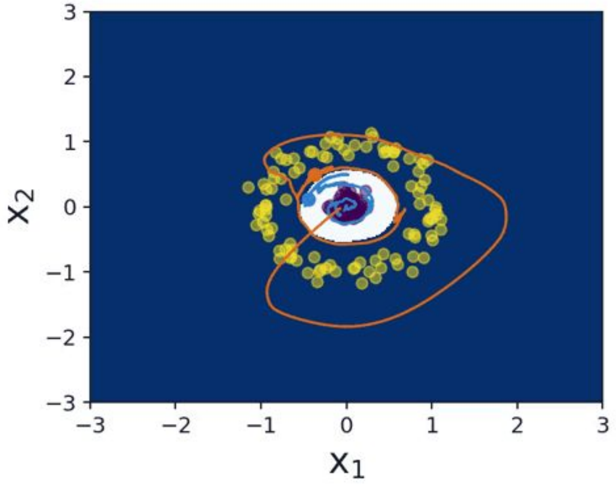
$$\underbrace{\mathbf{u}^\star}_{\text{Learned datapoints}} \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \underbrace{\mathcal{D}_{\text{original}}}_{\text{Loss on the original dataset}})$$

$$\underbrace{\mathbf{w}^\star}_{\text{Model params}}(\mathbf{u}) \in \overbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \underbrace{\mathcal{D}(\mathbf{u})}_{\text{“Training” loss on the synthetic data}})}^{\text{Training a model on the synthetic data}}$$

Original dataset



Warm-start joint optimization



Warm-Start vs Cold-Start

$$\mathbf{u}^\star \in \arg\min_{\mathbf{u}} \mathcal{L}(\mathbf{w}^\star(\mathbf{u}), \mathcal{D}_{\text{original}})$$

Learned datapoints

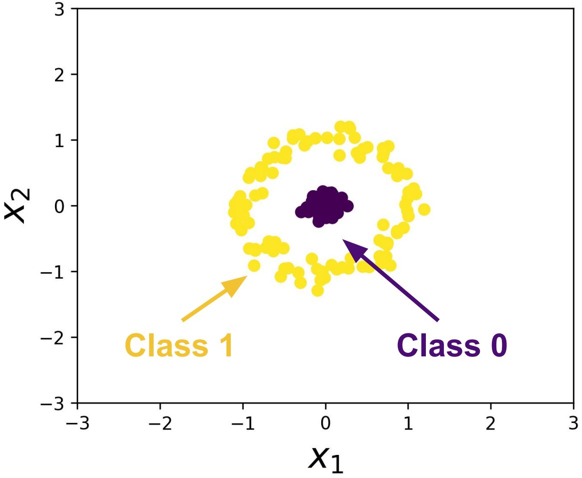
Loss on the original dataset

$$\mathbf{w}^\star(\mathbf{u}) \in \overbrace{\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}(\mathbf{u}))}^{\text{Training a model on the synthetic data}}$$

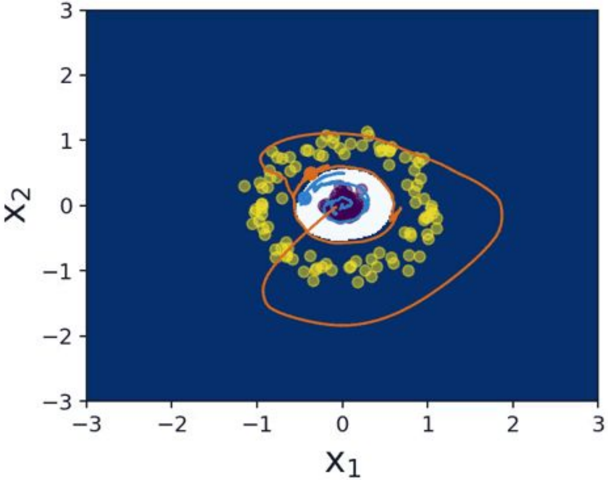
Model params

“Training” loss on the synthetic data

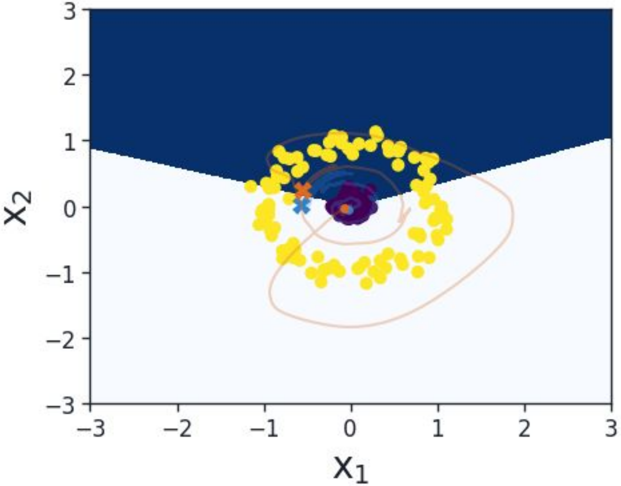
Original dataset



Warm-start joint optimization



Training from scratch on final points

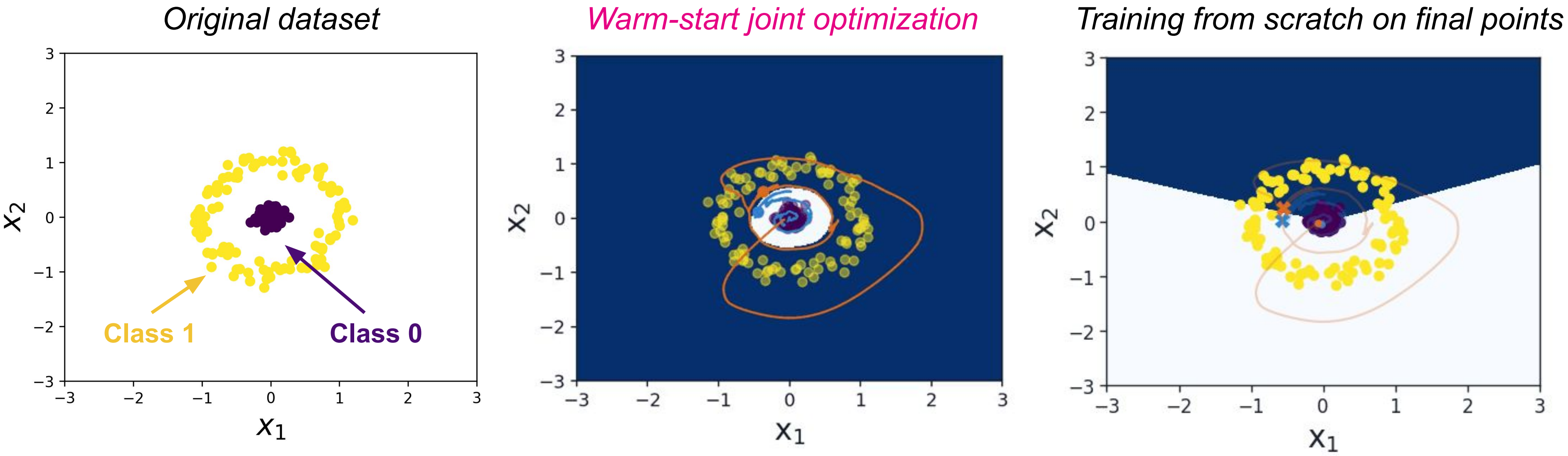


Warm-Start vs Cold-Start

Takeaways

1. A surprising amount of *information about the outer objective can leak to the inner parameters*, even when the outer parameters are low-dimensional

2. Warm-start bilevel optimization yields *outer parameters that fail to generalize* under re-initialization of the inner problem.



Implicit Bias of the Hypergradient Approximation

- Another source of implicit bias is the *hypergradient approximation*
- Assuming uniqueness, using the implicit function theorem, the hypergradient is:

$$\frac{d}{d\mathbf{u}} F(\mathbf{u}, \mathbf{w}^\star(\mathbf{u})) = \frac{\partial F}{\partial \mathbf{u}} + \boxed{\left(\frac{\partial \mathbf{w}^\star(\mathbf{u})}{\partial \mathbf{u}}\right)}^\top \frac{\partial F}{\partial \mathbf{w}^\star(\mathbf{u})} \quad \textit{where} \quad \boxed{\frac{\partial \mathbf{w}^\star(\mathbf{u})}{\partial \mathbf{u}}} = -\underbrace{\left(\frac{\partial^2 f}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)^{-1}}_{\mathbf{H}^{-1}} \frac{\partial^2 f}{\partial \mathbf{w} \partial \mathbf{u}}$$

Intractable to store or invert

- We can compute $\mathbf{H}^{-1}$ using the *Neumann series*: $\mathbf{H}^{-1} = \alpha \sum_{k=0}^{\infty} (\mathbf{I} - \alpha \mathbf{H})^k$

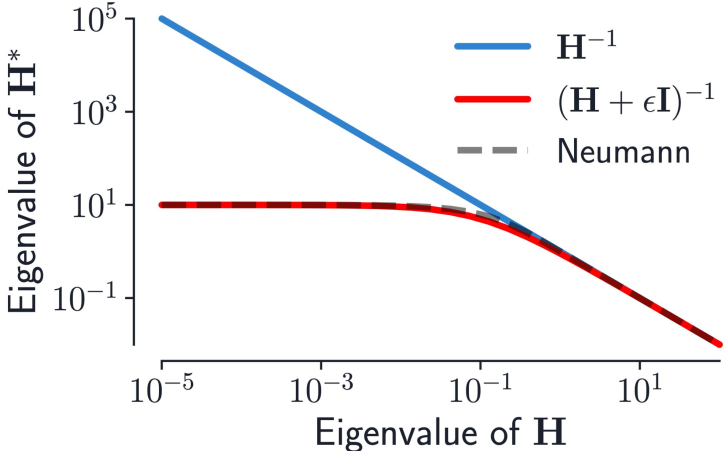
- In practice, we use a *truncation* of the infinite series

➡ What is the effect of using the truncated Neumann series on the outer optimization?

Implicit Bias of the Hypergradient Approximation

- Truncated Neumann approximates the inverse of the damped Hessian $(\mathbf{H} + \epsilon \mathbf{I})^{-1}$

$$\alpha \sum_{j=0}^{K} (\mathbf{I} - \alpha \mathbf{H})^j \approx (\mathbf{H} + \epsilon \mathbf{I})^{-1} \quad \text{where} \quad \epsilon = \frac{1}{\alpha K}$$

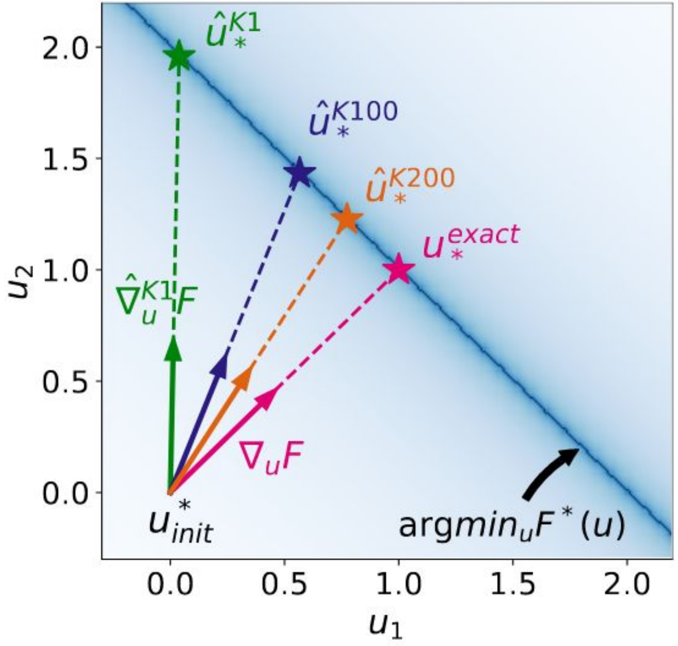


- The damped and un-damped inverse Hessians behave similarly in high curvature directions
- But the damped Hessian is *insensitive to low-curvature directions of the inner loss*

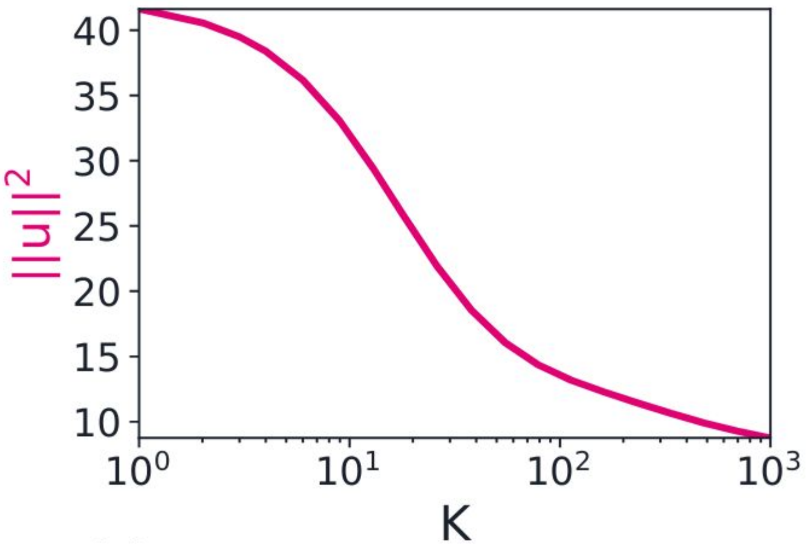
Implicit Bias of the Hypergradient Approximation

- Different *hypergradient approximations lead to different outer solutions*, with varying norms

Outer Optimization Trajectories



Converged Outer Parameter Norms



Thank you!