



On Implicit Bias in Overparameterized Bilevel Optimization

Paul Vicol^{1,2}, Jonathan Lorraine^{1,2}, Fabian Pedregosa³, David Duvenaud^{1,2}, Roger Grosse^{1,2}

¹University of Toronto, ²Vector Institute, ³Google Brain



VECTOR INSTITUTE

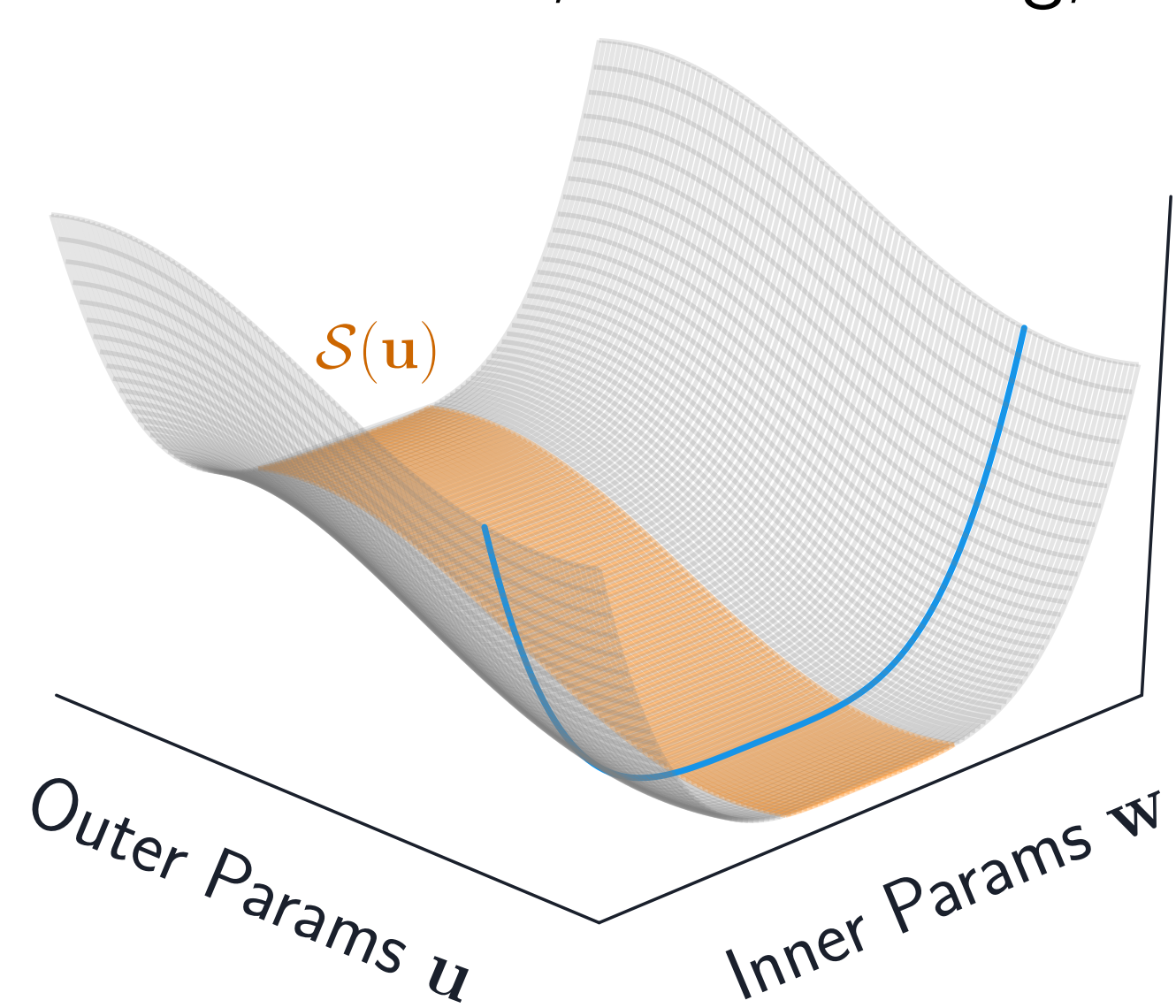
Motivation

- Bilevel problems** involve nested optimization, with an outer obj solved subject to the optimality of an inner obj:

$$\mathbf{u}^* \in \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} F(\mathbf{u}, \mathbf{w}^*)$$

$$\mathbf{w}^* \in \mathcal{S}(\mathbf{u}^*) = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmin}} f(\mathbf{u}^*, \mathbf{w})$$

- Examples:** hyperparameter optimization (HO), dataset distillation, meta-learning, NAS, and GANs.



Most work assumes that the inner/outer objectives have unique solutions, but in practice, at least one of them is underspecified \rightarrow non-unique solutions.

Gradient-Based Bilevel Optimization (BLO)

- Gradient-based BLO requires the total gradient of the outer objective w.r.t. the outer parameters, which we call the **hypergradient** (as in HO). For a given solution $\mathbf{w}^* \in \mathcal{S}(\mathbf{u})$, which is called a **best-response** to \mathbf{u} : $\frac{dF(\mathbf{u}, \mathbf{w}^*)}{d\mathbf{u}} = \frac{\partial F}{\partial \mathbf{u}} + \frac{\partial F}{\partial \mathbf{w}^*} \frac{\partial \mathbf{w}^*}{\partial \mathbf{u}}$

Warm-Start vs Cold-Start

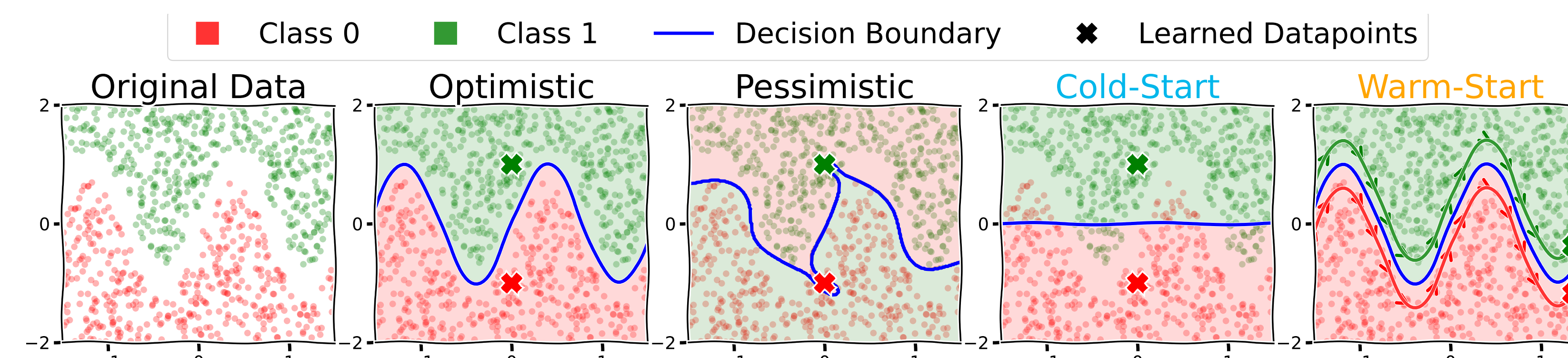
- Cold-start:** re-initialize \mathbf{w} and run inner optimization to convergence for each hypergradient computation
- Warm-start:** jointly optimize \mathbf{w} and \mathbf{u} online, alternating gradient steps with their respective objectives
- Let $\Xi^{(T)}(\mathbf{u}, \mathbf{w})$ denote T steps of inner optimization

Method	Inner Update
Cold-Start	$\mathbf{w}_{k+1}^* = \Xi^{(\infty)}(\mathbf{u}_{k+1}, \mathbf{w}_0)$
Full Warm-Start	$\mathbf{w}_{k+1}^* = \Xi^{(\infty)}(\mathbf{u}_{k+1}, \mathbf{w}_k^*)$
Partial Warm-Start	$\mathbf{w}_{k+1}^* = \Xi^{(T)}(\mathbf{u}_{k+1}, \mathbf{w}_k^*)$

Theory

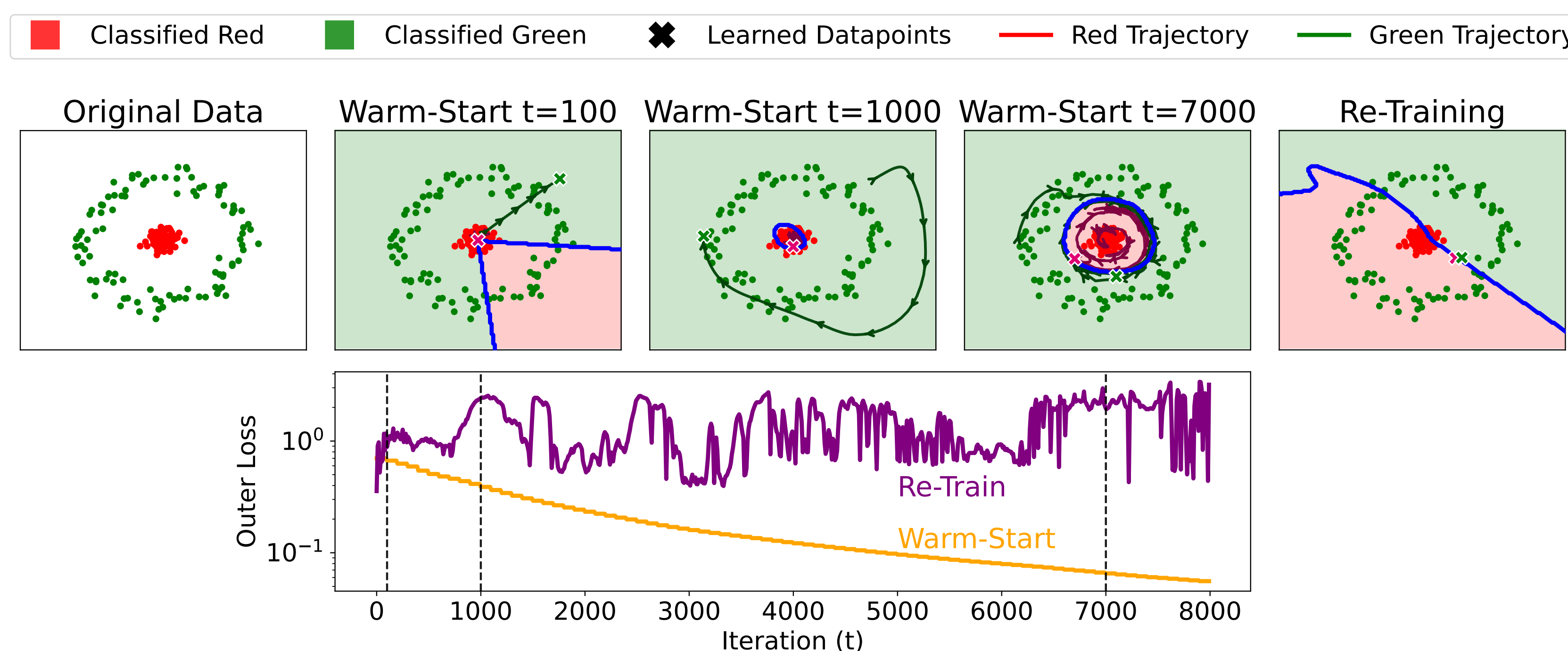
- Assuming F and f are quadratic and we use exact hypergradients, the converged \mathbf{u}^* minimizes distance to \mathbf{u}_0 : $\operatorname{argmin}_{\mathbf{u} \in \operatorname{argmin}_{\mathbf{w}} F(\mathbf{u}, \mathbf{w}^*)} \|\mathbf{u} - \mathbf{u}_0\|_2^2$. ($\mathbf{u}_0 = 0$ gives min-norm)
- For strongly-convex f , **full warm-start** \equiv **cold-start**
- Under conditions, **full warm-start** \equiv **partial warm-start**.

Warm-Start vs Cold-Start Solution Concepts

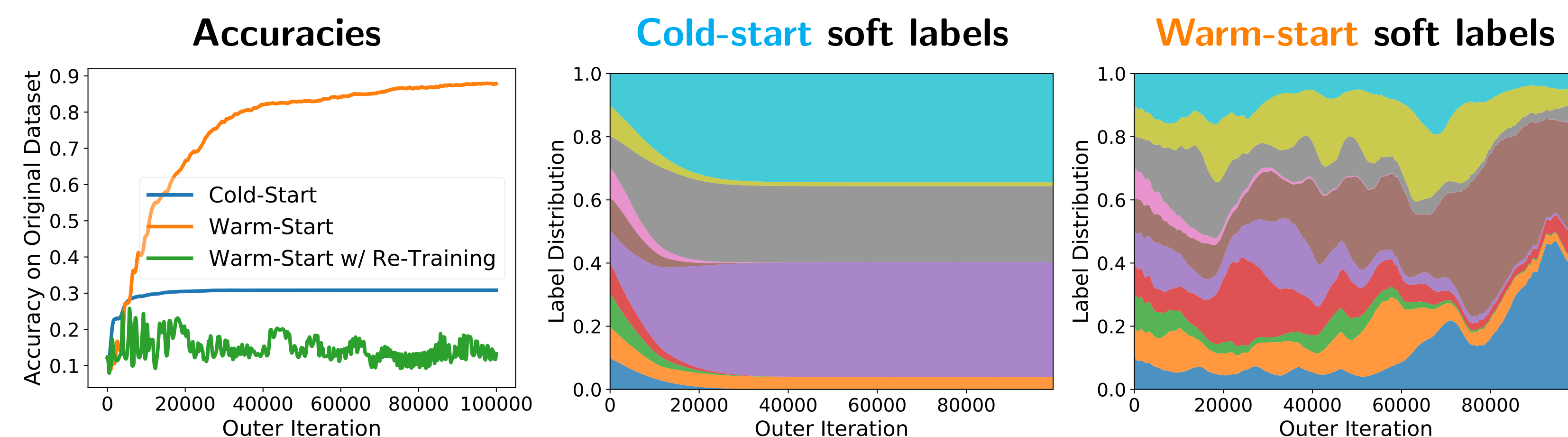


- Dataset distillation** for binary classification, with **two learned datapoints** (outer parameters) adapted jointly with the **model weights** (inner parameters).
- Optimistic:** \mathbf{w} achieves the **best outer-objective value**, $\operatorname{argmin}_{\mathbf{w}^* \in \mathcal{S}(\mathbf{u})} F(\mathbf{u}, \mathbf{w})$.
- Pessimistic:** \mathbf{w} achieves the **worst outer-objective value**, $\operatorname{argmax}_{\mathbf{w}^* \in \mathcal{S}(\mathbf{u})} F(\mathbf{u}, \mathbf{w})$.
- In practice:** Due to implicit bias of gradient descent, the solution $\mathbf{w}^* \in \mathcal{S}(\mathbf{u})$ we end up at depends on the initialization \mathbf{w}_0 : **cold-start** is biased towards simple solutions (e.g., min-norm solutions for quadratic f)
- With **warm-start**, the trajectory of outer parameters \mathbf{u} during joint optimization influences the inner parameters \mathbf{w} .

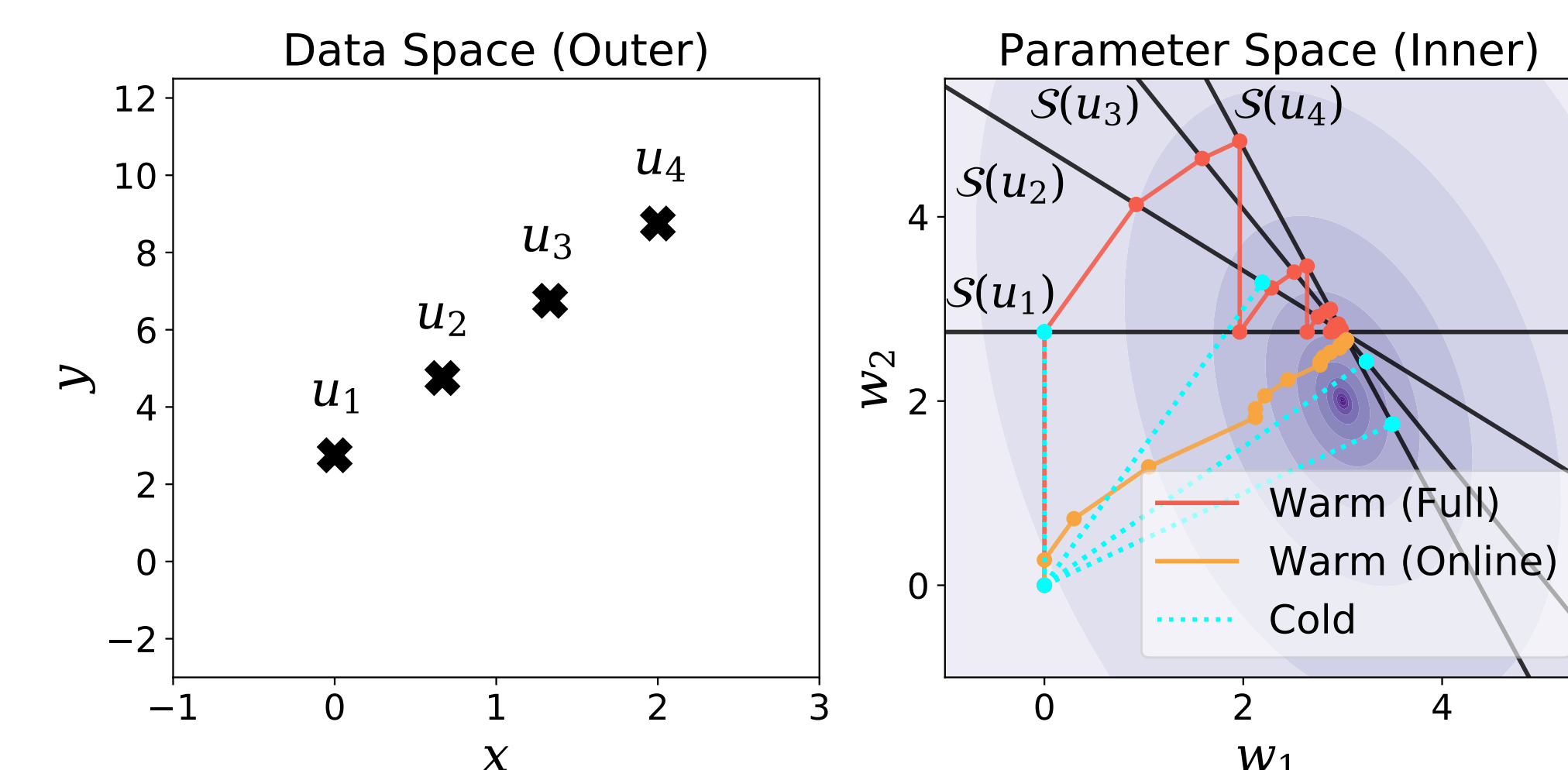
Inner Overparameterization: Dataset Distillation



- Because F is only used to update the outer parameters, one might think that all of the info about F is compressed into \mathbf{u} .
- Inner parameters can encode a surprising amount of information about the outer objective**, even when the outer parameters are low-dimensional.
- Warm-start BLO yields outer parameters that fail to generalize under re-initialization of the inner problem.

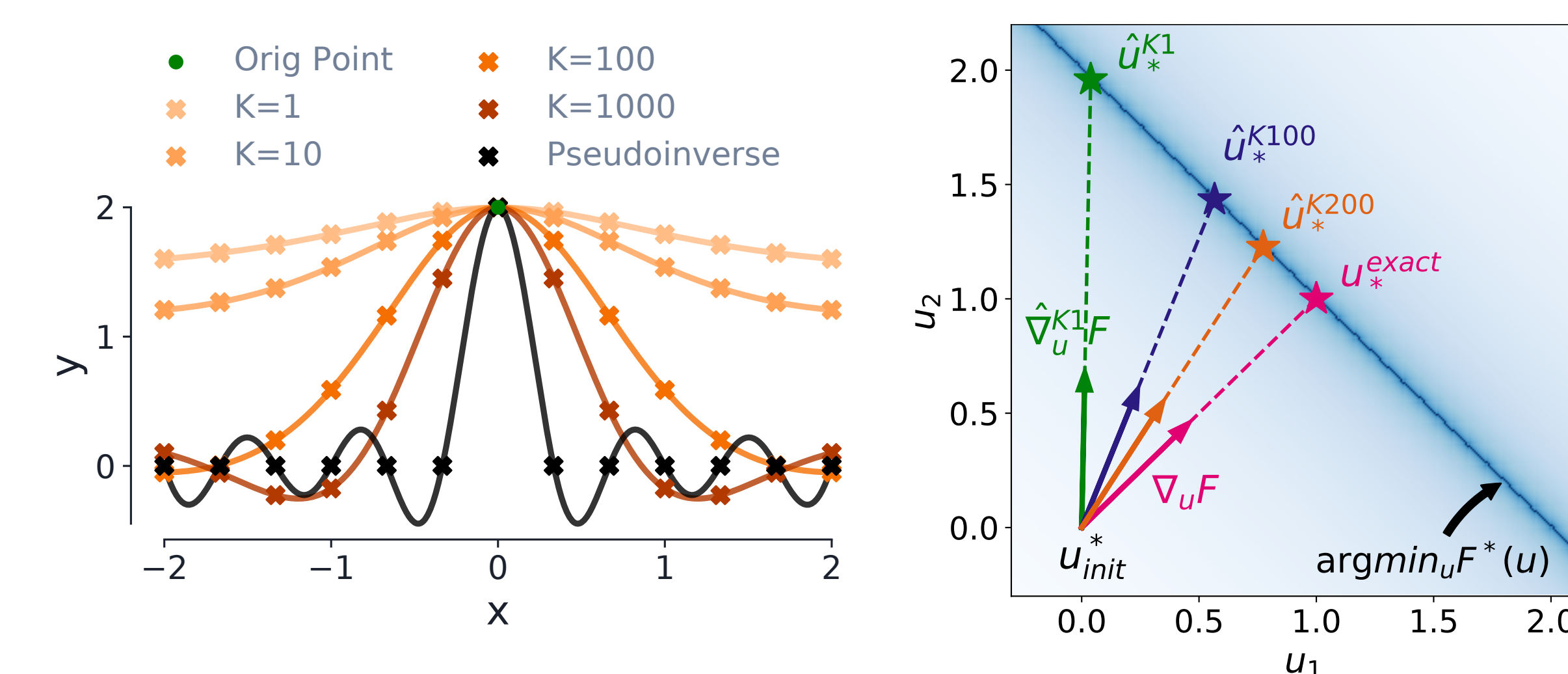


Explaining Warm-Start Memory



- Simplified parameter- and data-space view of **warm-start with full inner optimization**, **warm-start with partial inner optimization**, and **cold-start optimization**.
- Here, we cycle through outer param values $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$
- Cold-start** projects from the origin \mathbf{w}_0 onto the solution set for the current datapoint, $\mathcal{S}(\mathbf{u})$.
- Full warm-start** projects from the current weights \mathbf{w}_k .
- If we **successively project** between solution sets, \mathbf{w} can converge to the **intersection of solution sets**, so \mathbf{w} can perform well for multiple \mathbf{u} simultaneously.

Implicit Bias from Hypergradient Approximation



- The **truncated Neumann series approximates the damped Hessian inverse**: $\alpha \sum_{j=0}^K (\mathbf{I} - \alpha \mathbf{H})^j \approx (\mathbf{H} + \epsilon \mathbf{I})^{-1}$ where $\epsilon = \frac{1}{\alpha K}$.
- The damping prevents the inner optimization from moving far in low-curvature directions.
- Anti-distillation:** **more distilled datapoints than original datapoints**. We learn the y-coord of 13 synthetic datapoints such that a regressor trained on them will fit a single original datapoint, at the **green dot**.
- Left:** learned datapoints (outer parameters) from **different hypergradient approximations**: truncated Neumann/diff-through-unrolling with different # steps K
- Right:** The **exact hypergradient leads to the min-norm solution** $\|\mathbf{u} - \mathbf{u}_0\|_2^2$, while approximate Neumann hypergradients lead to different (valid) outer solutions.