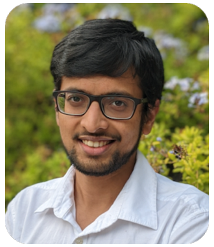
Open Problem: Can Local Regularization Learn All Multiclass Problems?

Julian Asilis

Siddartha Devic

Shaddin Dughmi

Vatsal Sharan



Shang-Hua Teng

Context on Classification

Binary classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y} = \{0, 1\}$
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

Multiclass classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y}$ (arbitrary)
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

Context on Classification

Binary classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y} = \{0, 1\}$
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

Multiclass classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y}$ (arbitrary)
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

Context on Classification

Binary classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y} = \{0, 1\}$
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn?

How to learn?

Multiclass classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y}$ (arbitrary)
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn?

How to learn?

Context on Classification

Binary classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y} = \{0, 1\}$
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn? VC($\mathcal{H}$) $< \infty$ [BEHW89]

How to learn? ERM

- Extremely simple
- Nearly optimal sample complexity

[BEHW89] – Blumer, Ehrenfeucht, Haussler, and Warmuth.
Learnability and the Vapnik-Chervonenkis Dimension

Multiclass classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y}$ (arbitrary)
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn?

How to learn?

Context on Classification

Binary classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y} = \{0, 1\}$
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn? VC($\mathcal{H}$) $< \infty$ [BEHW89]

How to learn? ERM

- Extremely simple
- Nearly optimal sample complexity

[BEHW89] – Blumer, Ehrenfeucht, Haussler, and Warmuth.
Learnability and the Vapnik-Chervonenkis Dimension

Multiclass classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y}$ (arbitrary)
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn? DS($\mathcal{H}$) $< \infty$ [BCDMY22]

How to learn? Not so clear…

- BCDMY learner is highly complex: subsampling, list PAC learning, sample compression, etc.

[BCDMY22] – Brukhim, Carmon, Dinur, Moran and Yehudayoff.
A Characterization of Multiclass Learnability

Context on Classification

Binary classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y} = \{0, 1\}$
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn? VC($\mathcal{H}$) $< \infty$ [BEHW89]

How to learn? ERM

- Extremely simple
- Nearly optimal sample complexity

Multiclass classification

Rules:

- Domain $\mathcal{X}$ (arbitrary)
- Label set $\mathcal{Y}$ (arbitrary)
- Loss function $\ell_{0-1}(y, y') = 1[y \neq y']$

When to learn? DS($\mathcal{H}$) $< \infty$ [BCDMY22]

Simple algorithmic templates for optimal multiclass learning?

[BEHW89] – Blumer, Ehrenfeucht, Haussler, and Warmuth.
Learnability and the Vapnik-Chervonenkis Dimension

[BCDMY22] – Brukhim, Carmon, Dinur, Moran and Yehudayoff.
A Characterization of Multiclass Learnability

Starting point: ERM & SRM

Empirical risk minimization (ERM)

$$A(S) = \mathrm{argmin}_{\mathcal{H}}\, L_S(h)$$

Structural risk minimization (SRM)

$$A(S) = \mathrm{argmin}_{\mathcal{H}}\, L_S(h) + \psi(h)$$

Starting point: ERM & SRM

Empirical risk minimization (ERM)

$$A(S) = \mathrm{argmin}_{\mathcal{H}}\ L_S(h)$$

Structural risk minimization (SRM)

$$A(S) = \mathrm{argmin}_{\mathcal{H}}\ L_S(h) + \psi(h)$$

Note ERM & SRM learners are proper,
always output functions in $\mathcal{H}$.

ERM characterizes learning for binary classification, but fails miserably for multiclass. *Why?*

Starting point: ERM & SRM

Empirical risk minimization (ERM)

$$A(S) = \operatorname{argmin}_{\mathcal{H}} L_S(h)$$

Structural risk minimization (SRM)

$$A(S) = \operatorname{argmin}_{\mathcal{H}} L_S(h) + \psi(h)$$

Note ERM & SRM learners are proper, always output functions in $\mathcal{H}$.

ERM characterizes learning for binary classification, but fails miserably for multiclass. *Why?*

Theorem [DS14]: In multiclass classification, there are learnable classes that cannot be learned by *any* proper learner.

Learning $\mathcal{H}$ can require emitting functions outside of $\mathcal{H}$.

(Even in realizable case!)

Dooms ERM & SRM – phrased as optimization problems over $\mathcal{H}$.

Proposed framework: local regularization

Key obstruction: SRM is inherently proper

- How to be improper while still optimizing over $\mathcal{H}$?

Proposed framework: local regularization

Key obstruction: SRM is inherently proper

- How to be improper while still optimizing over $\mathcal{H}$?

Solution: allow regularizer to depend on test point

- We call this a “local regularizer”
- A(S) can “glue” actions of different $h \in \mathcal{H}$ across $\mathcal{X}$

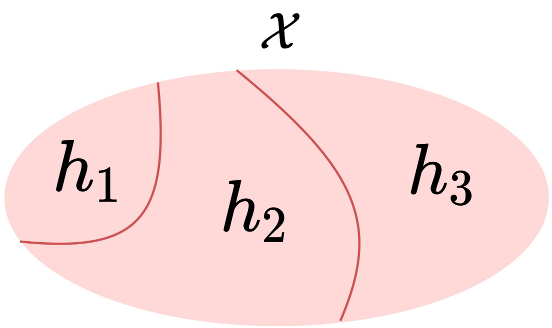
Proposed framework: local regularization

Key obstruction: SRM is inherently proper

- How to be improper while still optimizing over $\mathcal{H}$?

Solution: allow regularizer to depend on test point

- We call this a “local regularizer”
- A(S) can “glue” actions of different $h \in \mathcal{H}$ across $\mathcal{X}$



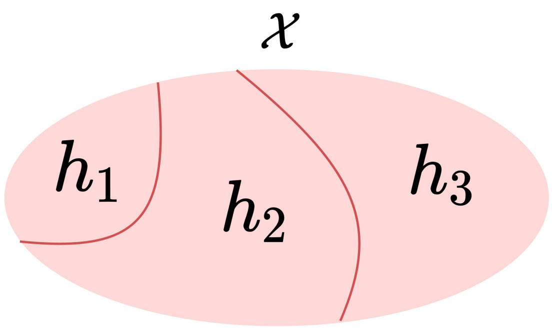
Proposed framework: local regularization

Key obstruction: SRM is inherently proper

- How to be improper while still optimizing over $\mathcal{H}$?

Solution: allow regularizer to depend on test point

- We call this a "local regularizer"
- A(S) can "glue" actions of different $h \in \mathcal{H}$ across $\mathcal{X}$



Formally, $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$,

$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{L_S^{-1}(0)} \psi(h, x)\}$$

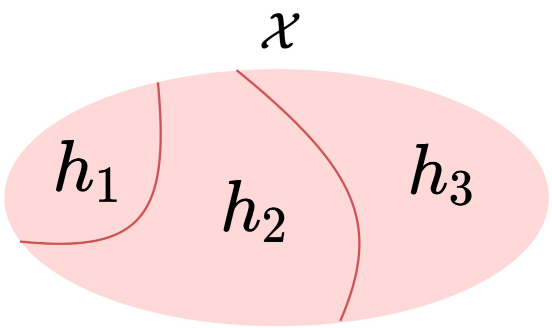
Proposed framework: local regularization

Key obstruction: SRM is inherently proper

- How to be improper while still optimizing over $\mathcal{H}$?

Solution: allow regularizer to depend on test point

- We call this a “local regularizer”
- A(S) can “glue” actions of different $h \in \mathcal{H}$ across $\mathcal{X}$



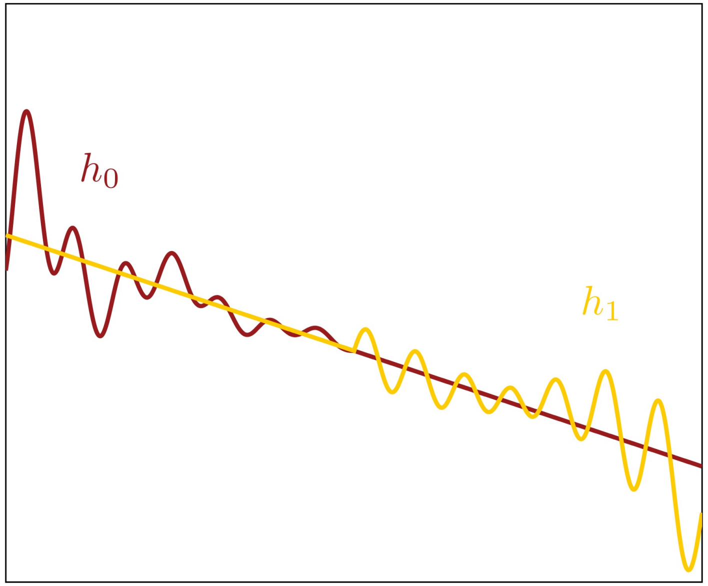
Formally, $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$,

$$A(S)(x) \in \{h(x) : h \in \text{argmin}_{L_S^{-1}(0)} \psi(h, x)\}$$

Intuition: ψ encodes *local* preferences on $\mathcal{H}$, rather than one *global* preference

Geometrically: $h \in \mathcal{H}$ can be “complex” in places, “simple” in others

Proposed framework: local regularization



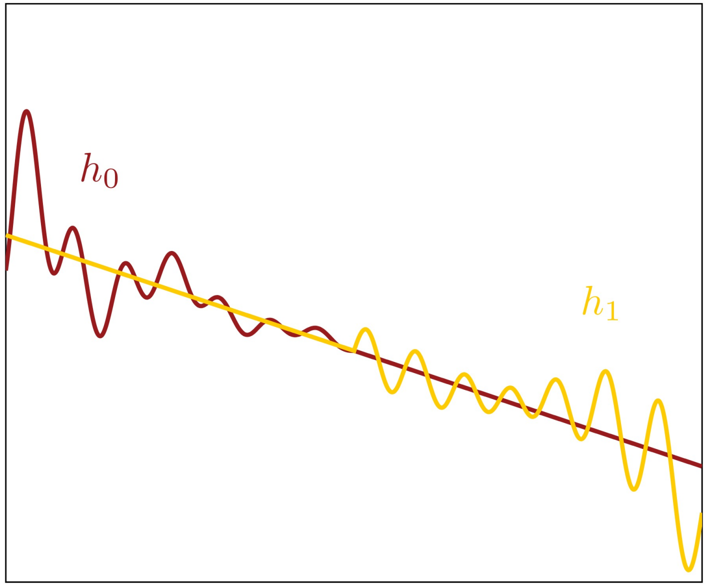
Formally, $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$,

$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{L_S^{-1}(0)} \psi(h, x)\}$$

Intuition: ψ encodes *local* preferences on $\mathcal{H}$, rather than one *global* preference

Geometrically: $h \in \mathcal{H}$ can be “complex” in places, “simple” in others

Proposed framework: local regularization



Formally, $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$,

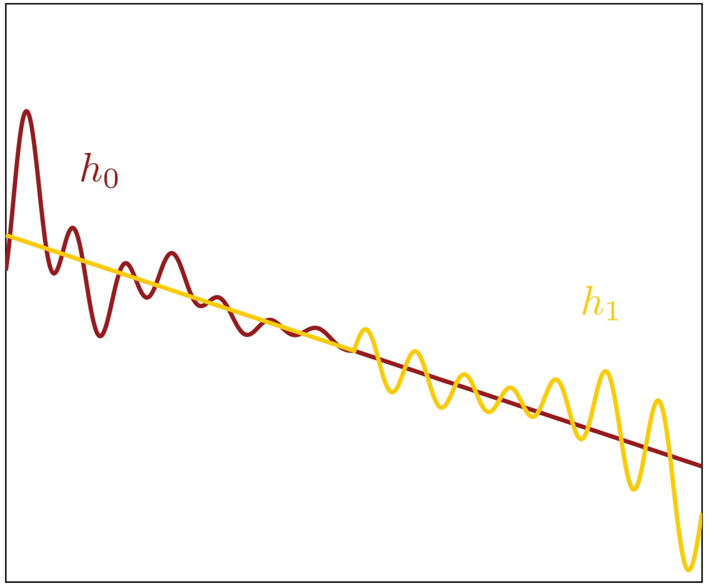
$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{L_S^{-1}(0)} \psi(h, x)\}$$

Intuition: ψ encodes *local* preferences on $\mathcal{H}$, rather than one *global* preference

Geometrically: $h \in \mathcal{H}$ can be "complex" in places, "simple" in others

Open problem: *In classification, can all learnable classes be learned by a local regularizer? If so, with (nearly) optimal sample complexity?*

Proposed framework: local regularization

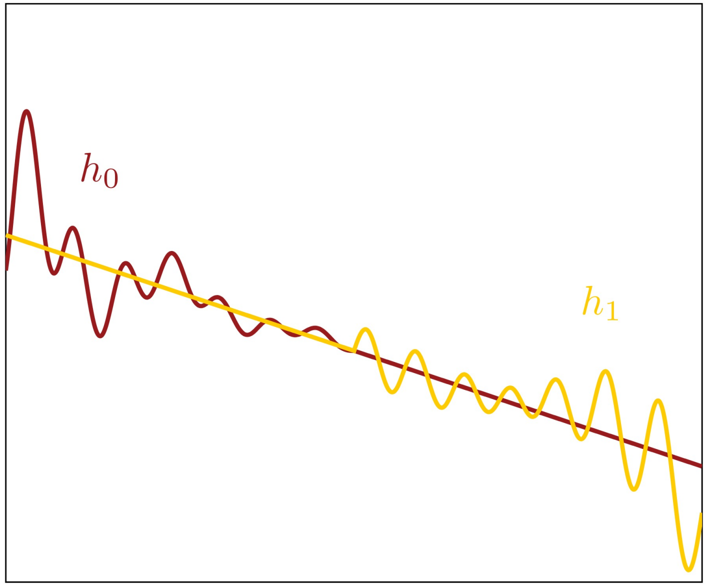


Open problem: *In classification, can all learnable classes be learned by a local regularizer? If so, with (nearly) optimal sample complexity?*

If **Yes**:

- Simpler algorithmic template for multiclass learning
  - Improves upon ***unsupervised local regularization*** [ADDST (COLT '24)]
- Reveals redundancy to *one-inclusion graph* learning algorithm (don't need unlabeled data)

Proposed framework: local regularization



Open problem: *In classification, can all learnable classes be learned by a local regularizer? If so, with (nearly) optimal sample complexity?*

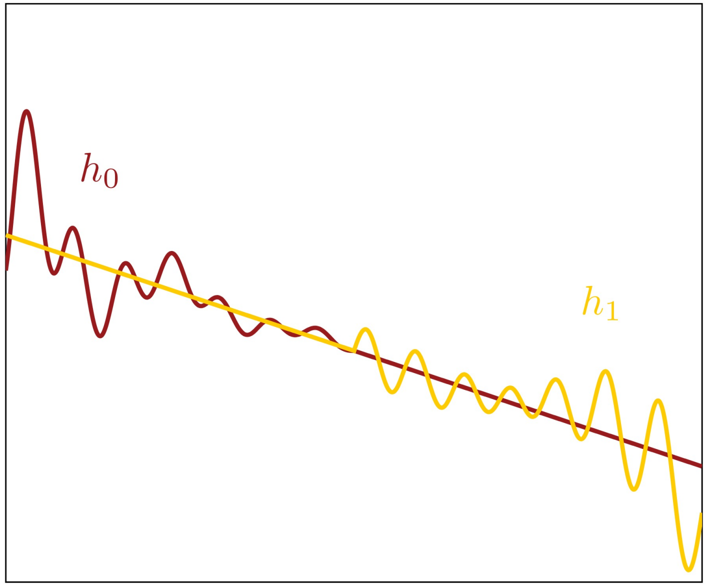
If **Yes**:

- Simpler algorithmic template for multiclass learning
  - Improves upon ***unsupervised local regularization*** [ADDST (COLT '24)]
- Reveals redundancy to *one-inclusion graph* learning algorithm (don't need unlabeled data)

If **No**:

- Impossibility result for understanding multiclass learners
- Suggests structure of OIGs is vital for learning (need unlabeled data)

Proposed framework: local regularization



Open problem: *In classification, can all learnable classes be learned by a local regularizer? If so, with (nearly) optimal sample complexity?*

If **Yes**:

- Simpler algorithmic template for multiclass learning
  - Improves upon ***unsupervised local regularization*** [ADDST (COLT '24)]
- Reveals redundancy to *one-inclusion graph* learning algorithm (don't need unlabeled data)

If **No**:

- Impossibility result for understanding multiclass learners
- Suggests structure of OIGs is vital for learning (need unlabeled data)

See our write-up for a possible counter-example!