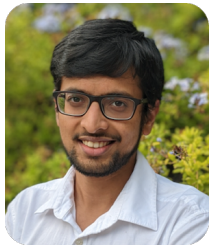
Regularization and Optimal Multiclass Learning

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

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?

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

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? $\mathrm{VC}(\mathcal{H}) < \infty$ [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? $\mathrm{DS}(\mathcal{H}) < \infty$ [BCDMY22]

Are there any 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)

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

Structural risk minimization (SRM)

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

Starting point: ERM & SRM

Empirical risk minimization (ERM)

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

Structural risk minimization (SRM)

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

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)

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

Structural risk minimization (SRM)

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

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 $\mathcal{H}$.

Starting point: ERM & SRM

Our motivating question:

What is the minimal augmentation of SRM that allows it to learn all (learnable) multiclass problems?

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 $\mathcal{H}$.

Relaxation 1: Local Regularization

Key obstruction: SRM is inherently proper

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

Relaxation 1: 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 $\mathcal{X}$

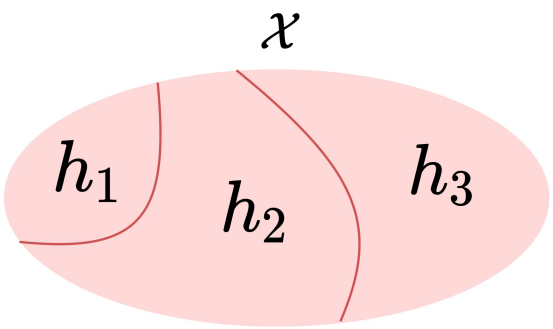
Relaxation 1: Local Regularization

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 $\mathcal{X}$



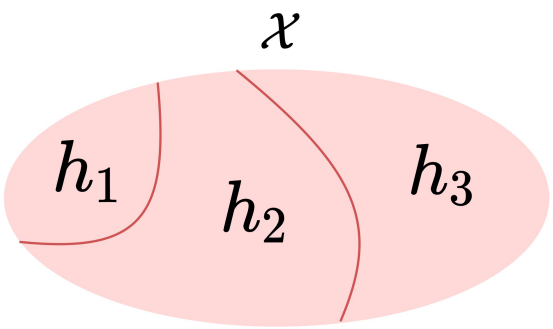
Relaxation 1: Local Regularization

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

$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{\mathcal{H}} L_S(h) + \psi(h, x)\}$$

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

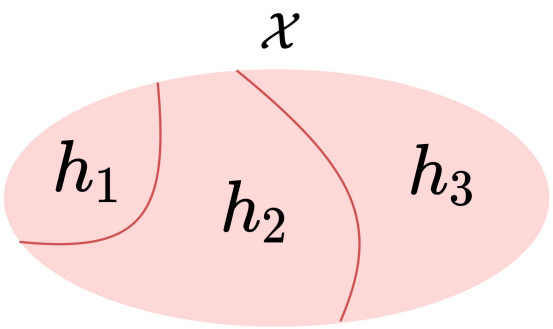
Relaxation 1: Local Regularization

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

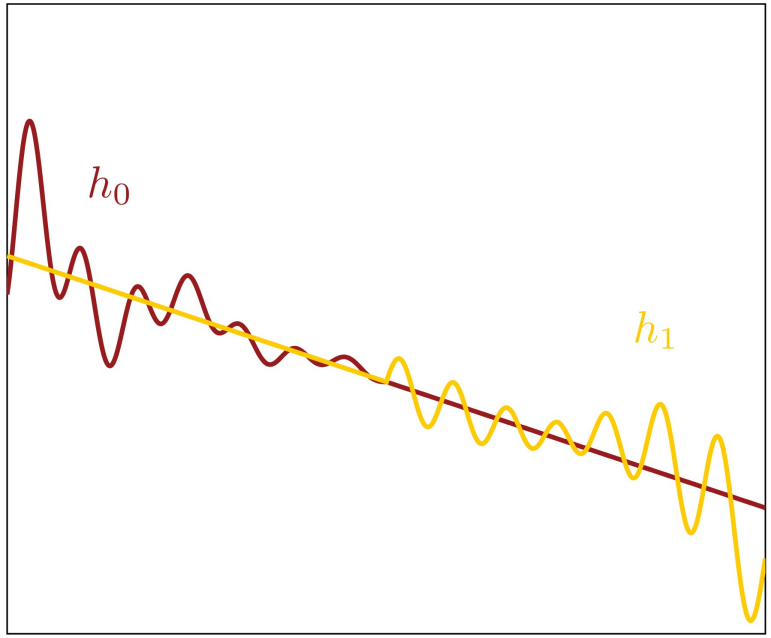
$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{\mathcal{H}} L_S(h) + \psi(h, x)\}$$

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

Local regularizer $\psi(h, x) =$ complexity of h at x

Relaxation 1: Local Regularization



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

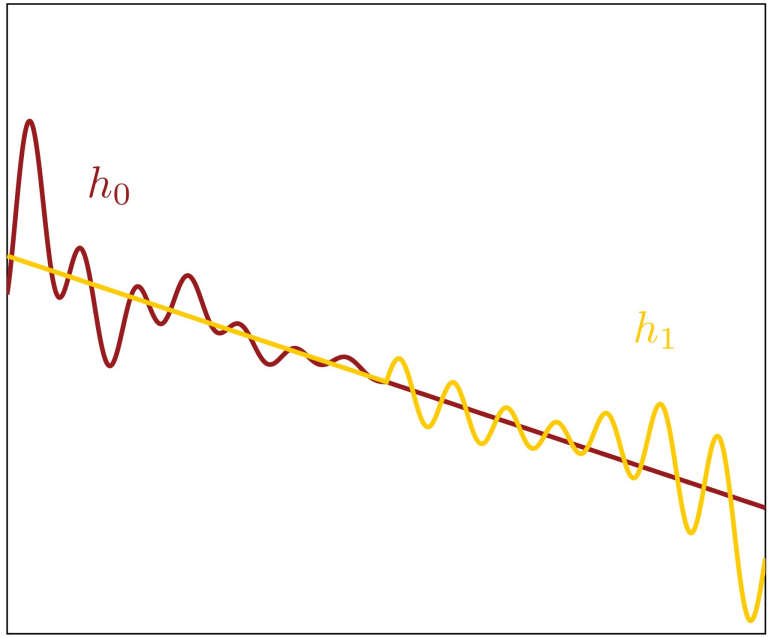
$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{\mathcal{H}} L_S(h) + \psi(h, x)\}$$

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

Local regularizer $\psi(h, x) =$ complexity of h at x

Relaxation 1: Local Regularization



Theorem: Even local regularization fails on learnable multiclass problems.

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

$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{\mathcal{H}} L_S(h) + \psi(h, x)\}$$

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

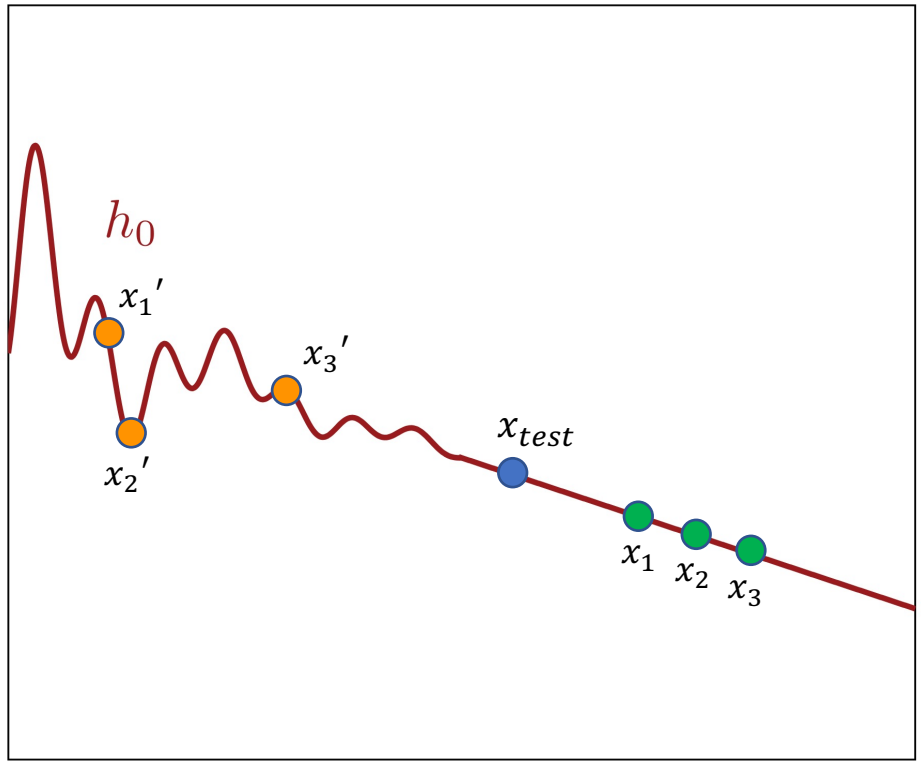
Local regularizer $\psi(h, x) =$ complexity of h at x

Relaxation 2: Unsupervised Learning of Regularizer

- Assigning hypothesis “complexity” may require more context than the test point
  - E.g., h is simple on $\{x_{test}, x_1, \dots, x_n\}$ but complex on $\{x_{test}, x_1', \dots, x_n'\}$

Relaxation 2: Unsupervised Learning of Regularizer

- Assigning hypothesis "complexity" may require more context than the test point
  - E.g., h is simple on $\{x_{test}, x_1, \dots, x_n\}$ but complex on $\{x_{test}, x_1', \dots, x_n'\}$



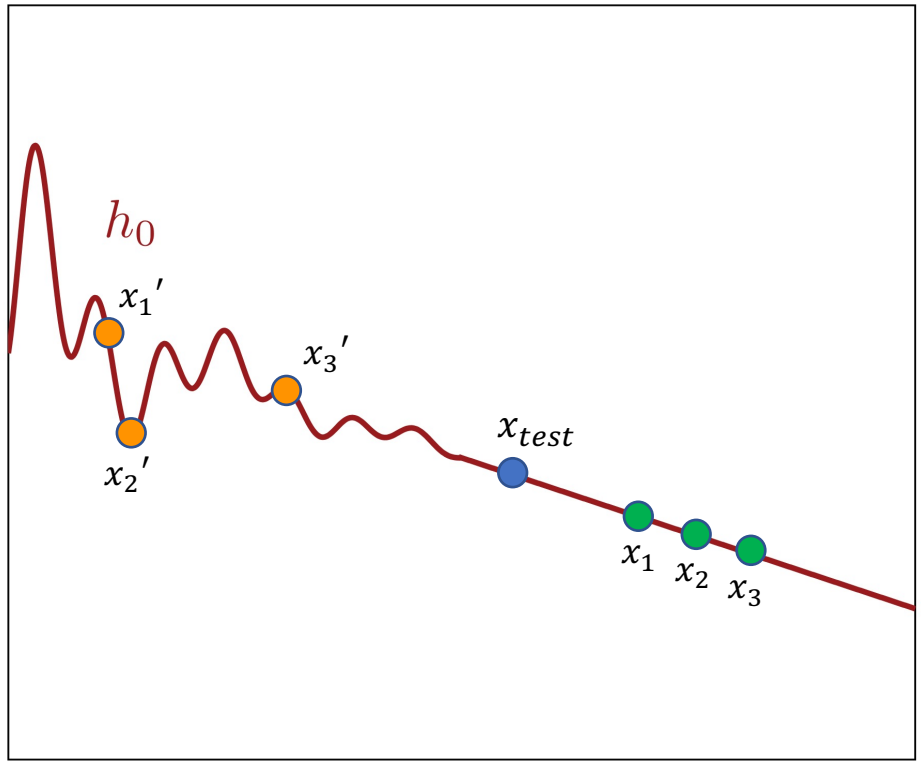
Relaxation 2: Unsupervised Learning of Regularizer

- Assigning hypothesis "complexity" may require more context than the test point
  - E.g., h is simple on $\{x_{test}, x_1, \ldots, x_n\}$ but complex on $\{x_{test}, x_1', \ldots, x_n'\}$

- Given unlabeled data S_X, learn local regularizer $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$

- Use local regularizer as before:

$$A(S)(x) \in \{h(x) : h \in \operatorname{argmin}_{\mathcal{H}} L_S(h) + \psi(h, x)\}$$



Our Results

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM.

Our Results

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM.

Good deterministic learner *but* loses factor 2, somewhat hard to interpret, fails in agnostic case.

Our Results

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM.

Good deterministic learner *but* loses factor 2, somewhat hard to interpret, fails in agnostic case.

Theorem: Every realizable or agnostic classification problem with finitely many labels admits an exactly optimal local unsupervised SRM (randomized).

Our Results

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM.

Good deterministic learner *but* loses factor 2, somewhat hard to interpret, fails in agnostic case.

Theorem: Every realizable or agnostic classification problem with finitely many labels admits an exactly optimal local unsupervised SRM (randomized).

Randomized learner attains exact optimality. Bayesian interpretation.

Our Results

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM.

Good deterministic learner *but* loses factor 2, somewhat hard to interpret, fails in agnostic case.

Theorem: Every realizable or agnostic classification problem with finitely many labels admits an exactly optimal local unsupervised SRM (randomized).

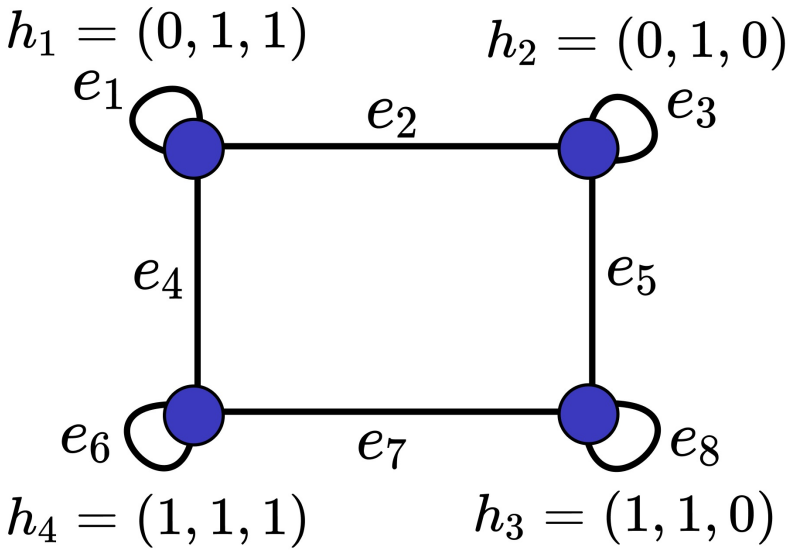
Randomized learner attains exact optimality. Bayesian interpretation.

(Results stated for transductive model. Transfer directly to PAC in black-box manner, up to lower-order factors.)

One-inclusion graphs

Throughout, we model learning using the *one-inclusion graph* (OIG) of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

- Vertex set $\mathcal{H}|_S$
- Hyperedges group hypotheses agreeing on $n-1$ points



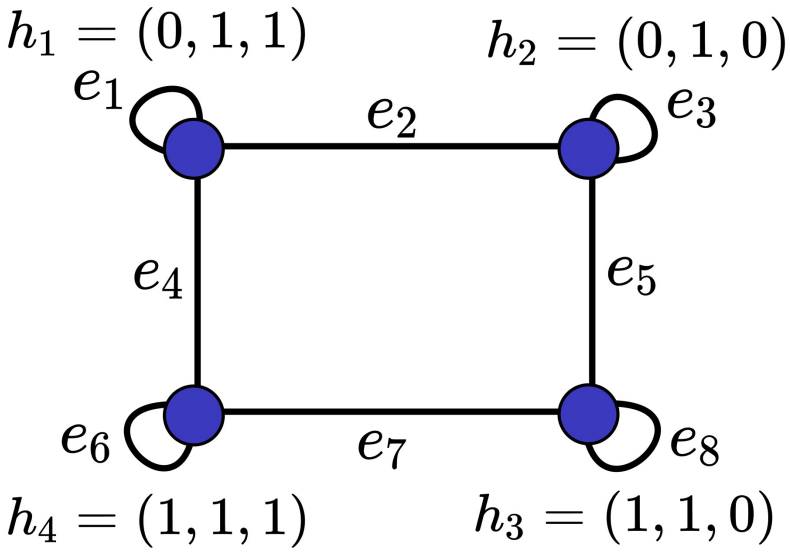
One-inclusion graphs

Throughout, we model learning using the *one-inclusion graph* (OIG) of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

- Vertex set $\mathcal{H}|_S$
- Hyperedges group hypotheses agreeing on $n-1$ points

Key facts:

1. Learner ≡ **orientation** of edges
2. Good learner ⇔ low outdegrees



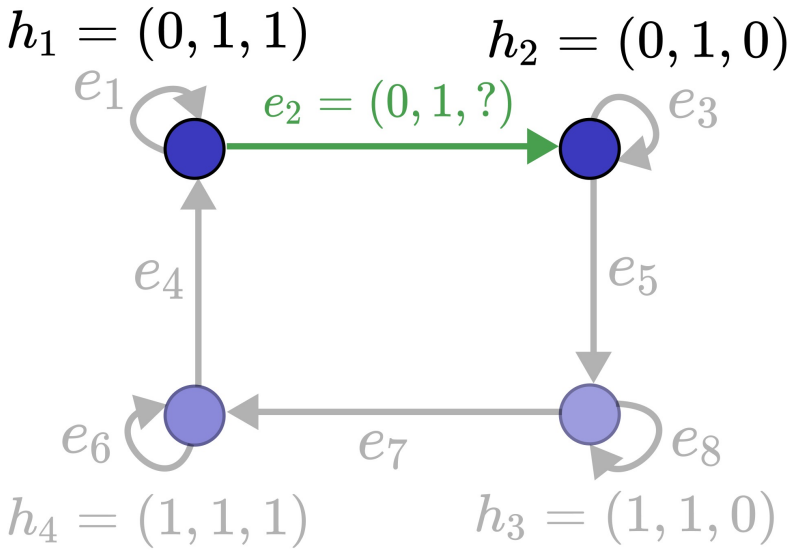
One-inclusion graphs

Throughout, we model learning using the *one-inclusion graph* (OIG) of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

- Vertex set $\mathcal{H}|_S$
- Hyperedges group hypotheses agreeing on $n-1$ points

Key facts:

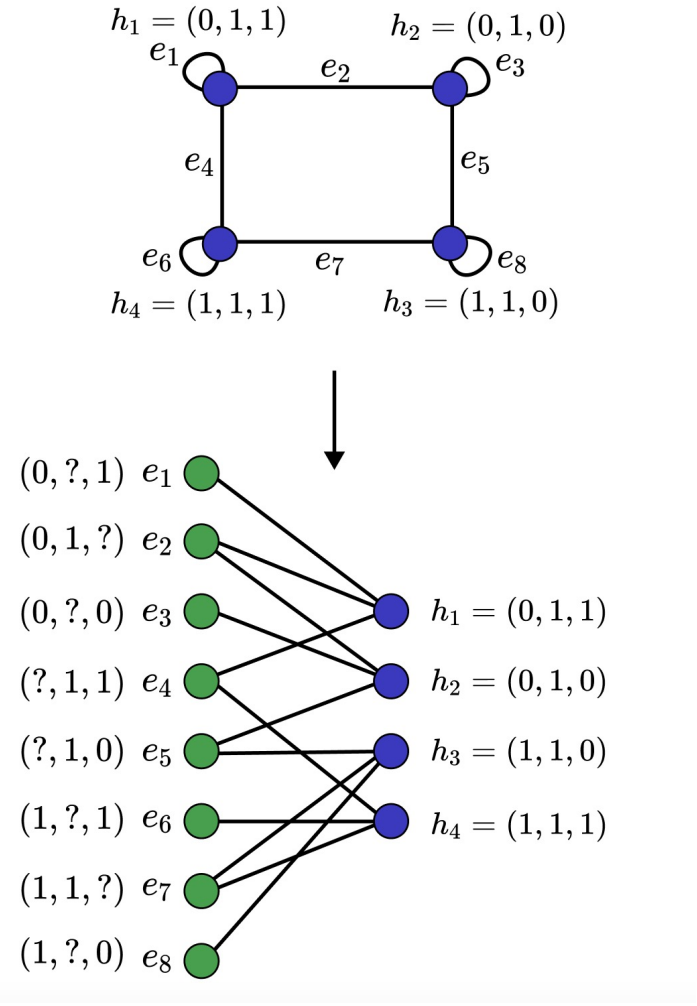
1. Learner ≡ **orientation** of edges
2. Good learner ⟺ low outdegrees



One-inclusion graphs

Throughout, we model learning using the *one-inclusion graph* (OIG) of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

- Vertex set $\mathcal{H}|_S$
- Hyperedges group hypotheses agreeing on $n-1$ points



Key facts:

1. Learner ≡ **orientation** of edges
2. Good learner ⟺ low outdegrees

Useful trick: pass to *bipartite version*

Good learner ≡ good matching/assignment in bipartite OIG

Glance at the Proofs

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM (deterministic).

Proof idea:

local unsupervised SRM $\equiv$ assignment of number to each node in the *one-inclusion graph* G, by which you direct edges

Glance at the Proofs

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM (deterministic).

Proof idea:

local unsupervised SRM $\equiv$ assignment of number to each node in the *one-inclusion graph* G, by which you direct edges

$\equiv$ topological ordering of G

Glance at the Proofs

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM (deterministic).

Proof idea:

local unsupervised SRM $\equiv$ assignment of number to each node in the *one-inclusion graph* G, by which you direct edges

$\equiv$ topological ordering of G

$\equiv$ **acyclic** orientation of G

Glance at the Proofs

Theorem: Every realizable classification problem with at most countably many labels admits a near optimal (factor 2) local unsupervised SRM (deterministic).

Proof idea:

local unsupervised SRM $\equiv$ assignment of number to each node in the *one-inclusion graph* G, by which you direct edges

$\equiv$ topological ordering of G

$\equiv$ **acyclic** orientation of G

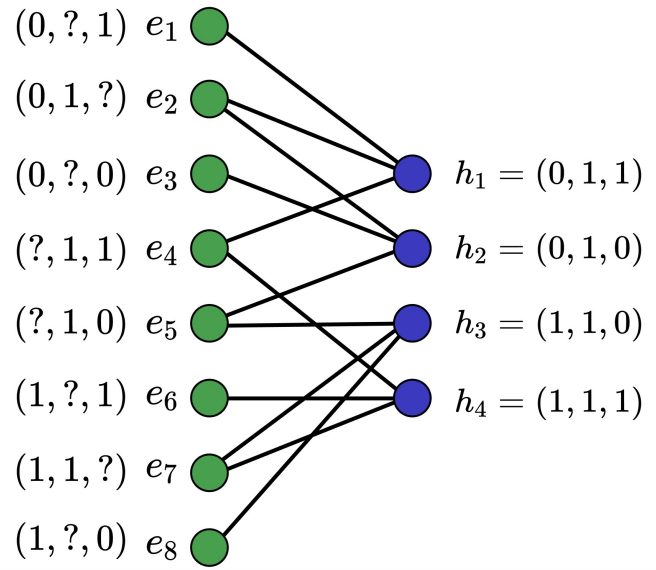
Key technical result: OIGs have optimal acyclic orientations (up to factor 2)

Glance at the Proofs

Theorem: Every realizable or agnostic classification problem with finitely many labels admits an exactly optimal local unsupervised SRM (randomized).

Proof idea:

Consider maximum entropy program over matchings in bipartite OIG



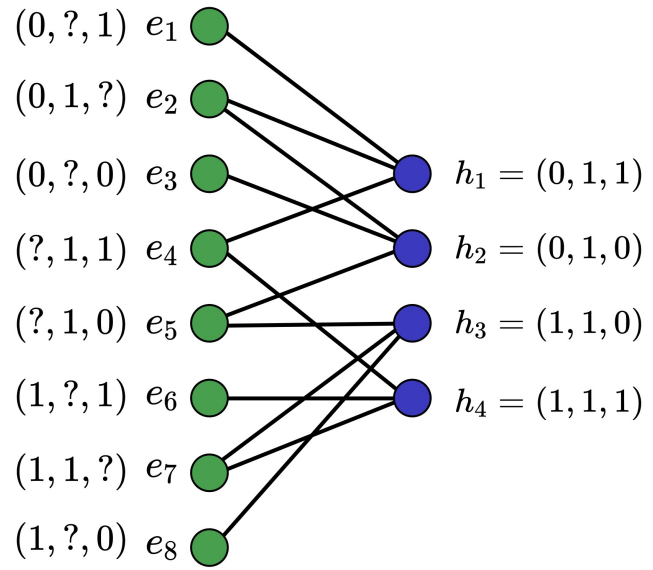
Glance at the Proofs

Theorem: Every realizable or agnostic classification problem with finitely many labels admits an exactly optimal local unsupervised SRM (randomized).

Proof idea:

Consider maximum entropy program over matchings in bipartite OIG

Dual optimum has one variable γ_v for each RHS node. Optimal randomized matching has *product structure* w.r.t. the γ_v's.



Glance at the Proofs

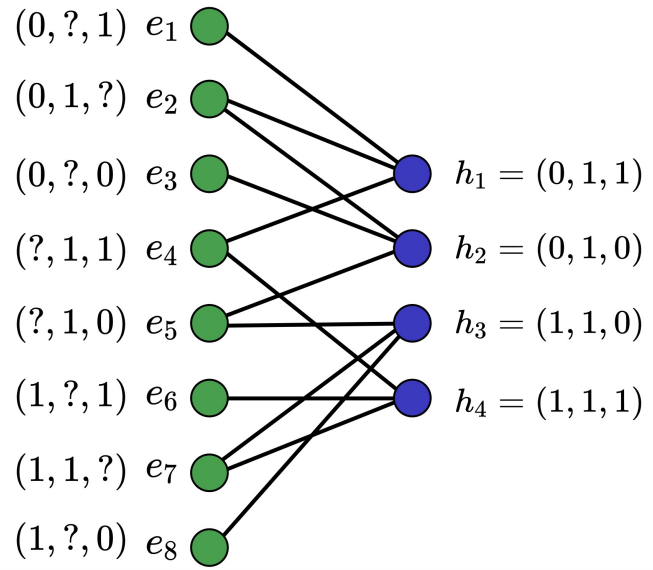
Theorem: Every realizable or agnostic classification problem with finitely many labels admits an exactly optimal local unsupervised SRM (randomized).

Proof idea:

Consider maximum entropy program over matchings in bipartite OIG

Dual optimum has one variable γ_v for each RHS node. Optimal randomized matching has *product structure* w.r.t. the γ_v's.

Fractions of dual variables yield *prior probabilities* for RHS nodes. Optimal learner has **Bayesian** structure.

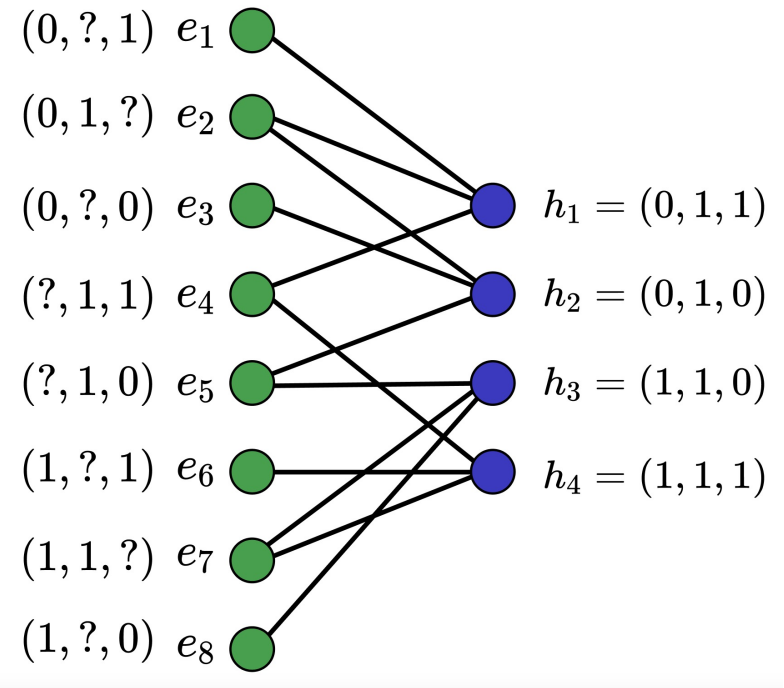


Companion Result: Hall Complexity

Theorem: The Hall complexity of a class $\mathcal{H}$ *exactly* characterizes its transductive error rate.

Hall complexity defined/analyzed by applying Hall's theorem to the bipartite OIGs.

Improves upon maximum subgraph density of [DS14], off by factor 2.



Conclusion

Small step toward algorithmic templates for multiclass learning

- Learning through *unsupervised local regularization*
- Concrete connections to Bayesian learning, maximum entropy principle

Conclusion

Small step toward algorithmic templates for multiclass learning

– Learning through *unsupervised local regularization*

– Concrete connections to Bayesian learning, maximum entropy principle

Future work: See our Open Problem talk tomorrow!

Open Problem: Can Local Regularization Learn All Multiclass Problems?

Asilis, Julian; Devic, Siddartha; Dughm, Shaddin

Session: Open Problems 2, Tuesday, Jul 02, 10:00-10:30

Thank you!