Meta-Learning with Highly Adaptive Lasso

Zeyi Wang
wangzeyi@berkeley.edu

Joint work with Wenxin Zhang, Brian Caffo, Martin Lindquist, and Mark van der Laan

2024 American Causal Inference Conference (ACIC), Seattle, WA
May 14, 2024

Outline

Introduction

Meta-Learning using the Highly Adaptive Lasso

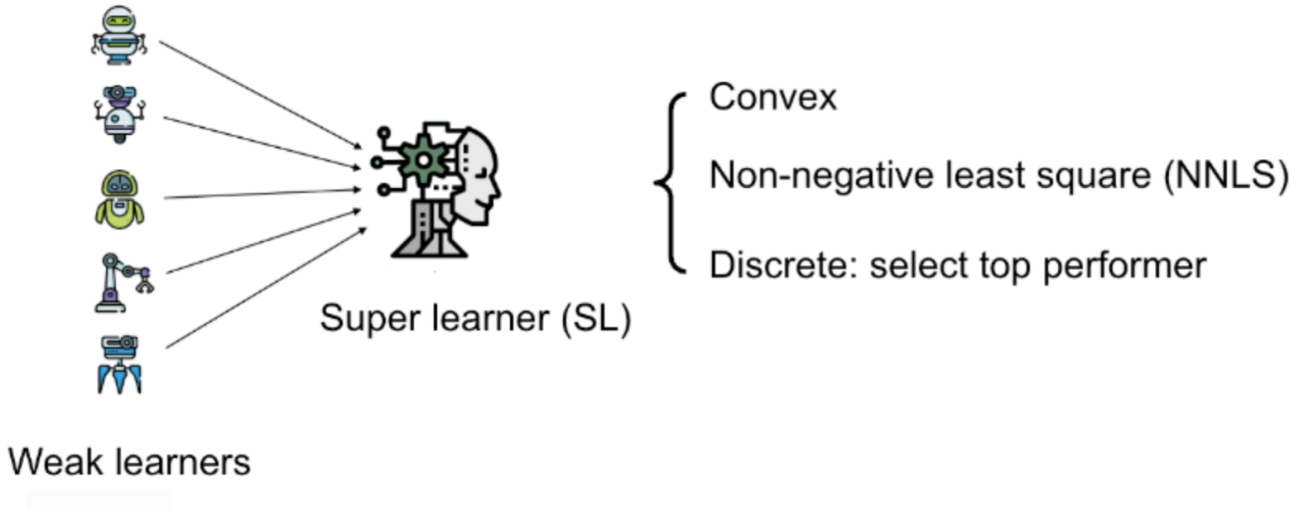


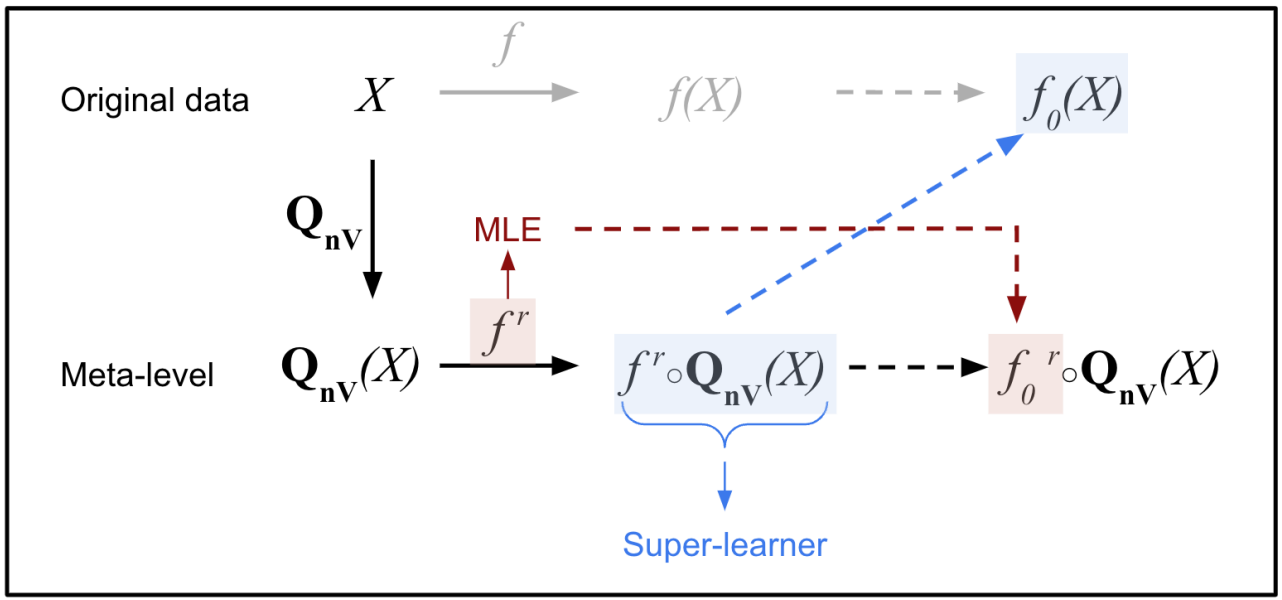
Figure: Super learning is a two-stage, meta-learning procedure.

By using HAL as the "ensembler", we can create

- A more flexible super learner,
- Consistent and asymptotically normal estimation (undersmoothing).

Tech report [Wang et al., 2023]: `https://arxiv.org/abs/2312.16953`.

Meta-HAL Algorithm



- f^r is an HAL-MLE with L1-norm constraint (the tuning parameter).
- $\boldsymbol{Q}_{nV}(X)$ is the cross-validated meta-level data (adaptiveness).

Different Estimation Goals

	Goal	Controlled term
Whole function $f \to f_0$	$\|f - f_0\|_{P,2}$	$\mathbb{E}L(f) - \mathbb{E}L(f_0)$
A smooth feature $\Psi(f) \to \Psi(f_0)$	$\Psi(f) \xrightarrow{P} \Psi(f_0)$ $\sqrt{n}(\Psi(f) - \Psi(f_0)) \rightsquigarrow N(0, \Sigma)$	$\frac{1}{n}\sum_{i=1}^{n} D(f, g)(O_i)$

When estimating the whole function:

- Double cross-validation is recommended for deciding hyperparameters of base learners and the "ensembler" algorithm;
- Only consider the approximation of f in the meta-level data, $\boldsymbol{Q}_{nV}(X)$.

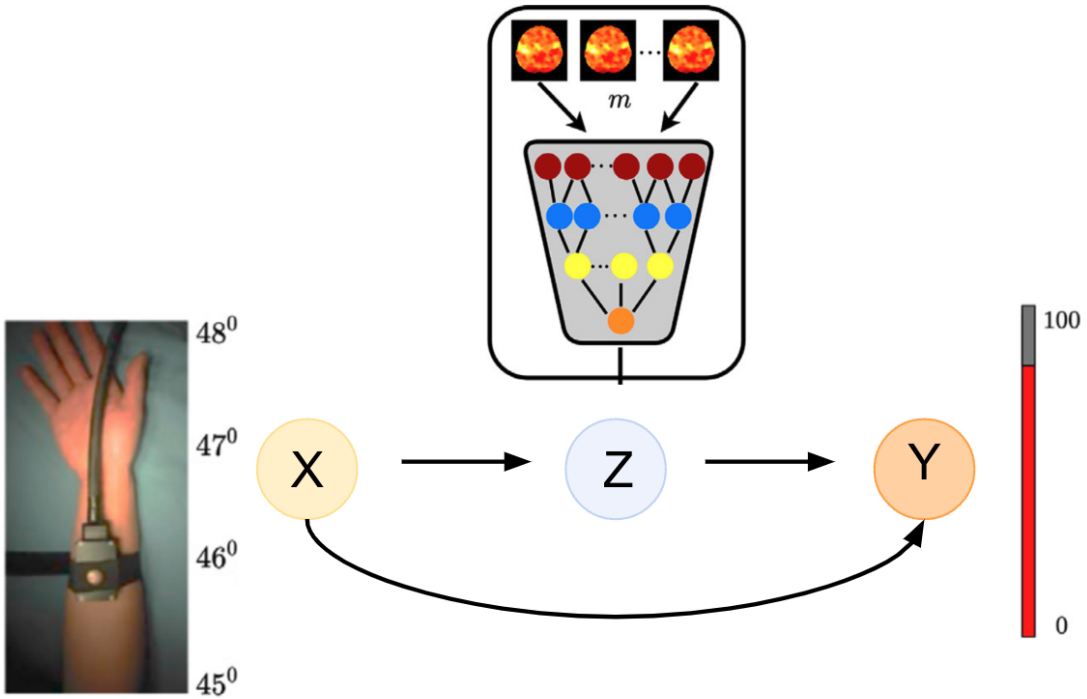
When estimating smooth features:

- Slight overfitting is acceptable and may be beneficial in undersmoothing;
- Consider approximating both f and g components in the score $D(f, g)$.

Code Examples

Application: Mediation with Neuroimaging Data

Pain and fMRI Data



Dimensions of Z:

- brain activation map: $91 \times 109 \times 91$,
- output of pretrained ResNet3D model `MedicalNet` [1]: 512 or 2048.
- meta-level mediator: 2.

[1] Chen S. et al. (2019). Med3D: Transfer Learning for 3D Medical Image Analysis.

Percentage Mediated

ATE: average treatment effect, contrast treated vs control.

NIE: natural indirect effect, achieved through impacting mediator.

$$\text{mediated\%} = \frac{\text{NIE}}{\text{ATE}} = \frac{\mathbb{E}\{Y(1) - Y(1, Z(0))\}}{\mathbb{E}\{Y(1) - Y(0)\}}$$

- True samples: bootstrap ($n = 10,000$) from observed data.
- Null samples: replace Z with random noise.

Meta-HAL Results

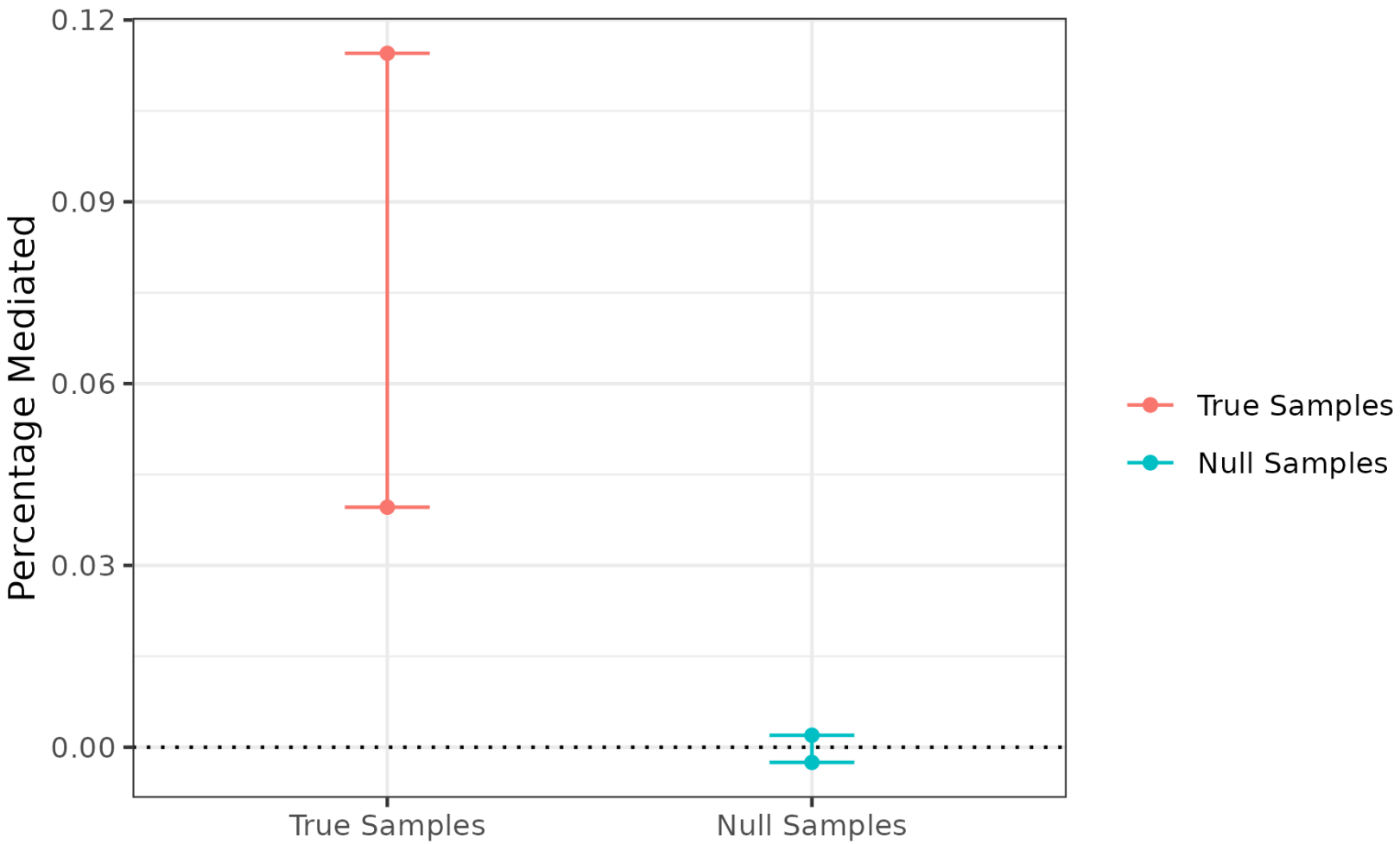


Figure: Undersmoothed meta-HAL plug-in.

Methods Comparison

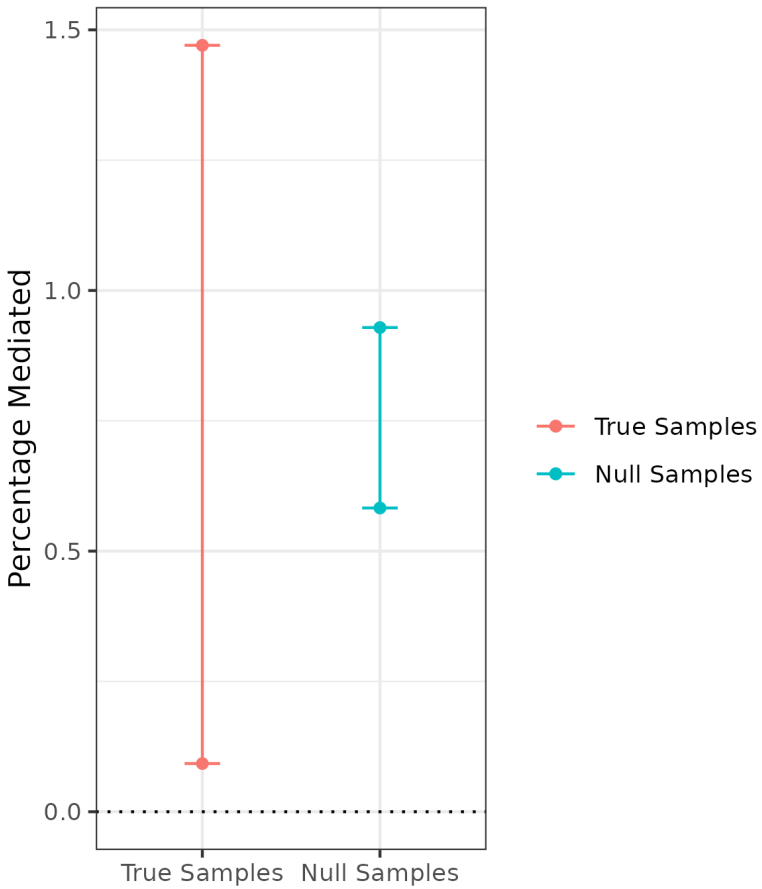


Figure: TMLE.

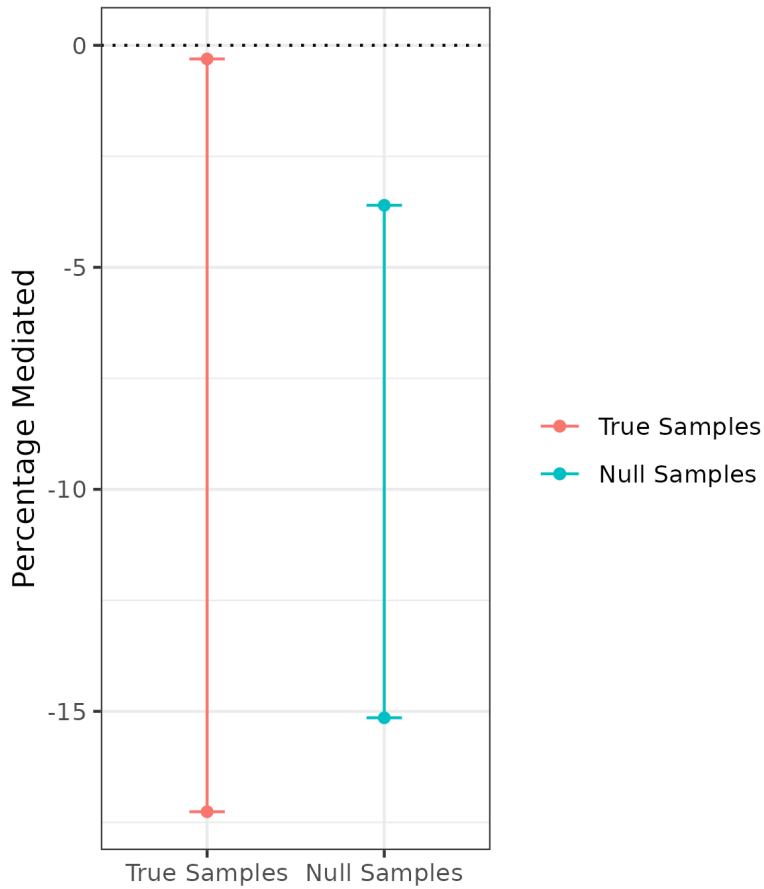


Figure: IPW.

Network Complexity Comparison

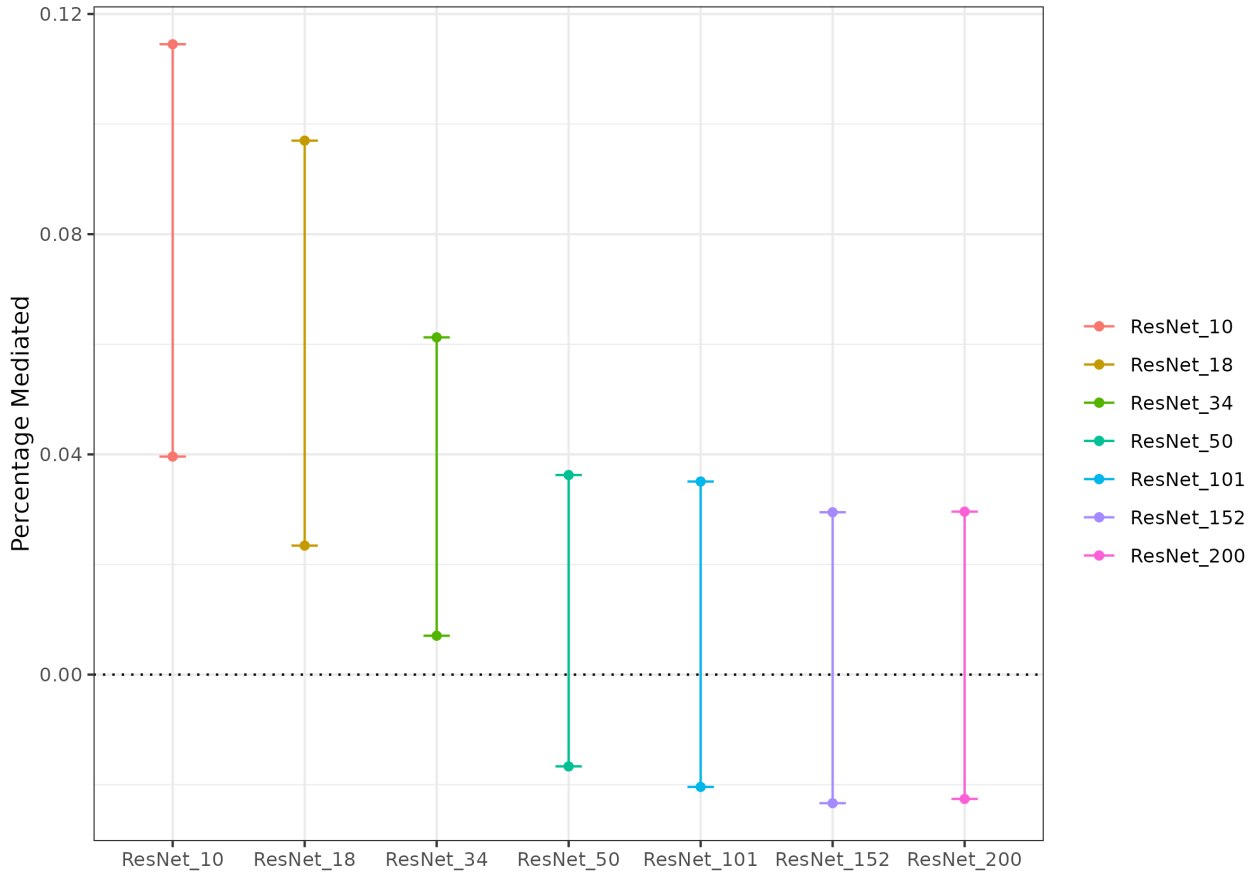


Figure: Different depths of pretrained ResNet networks.

Ensemble of Pretrained Networks

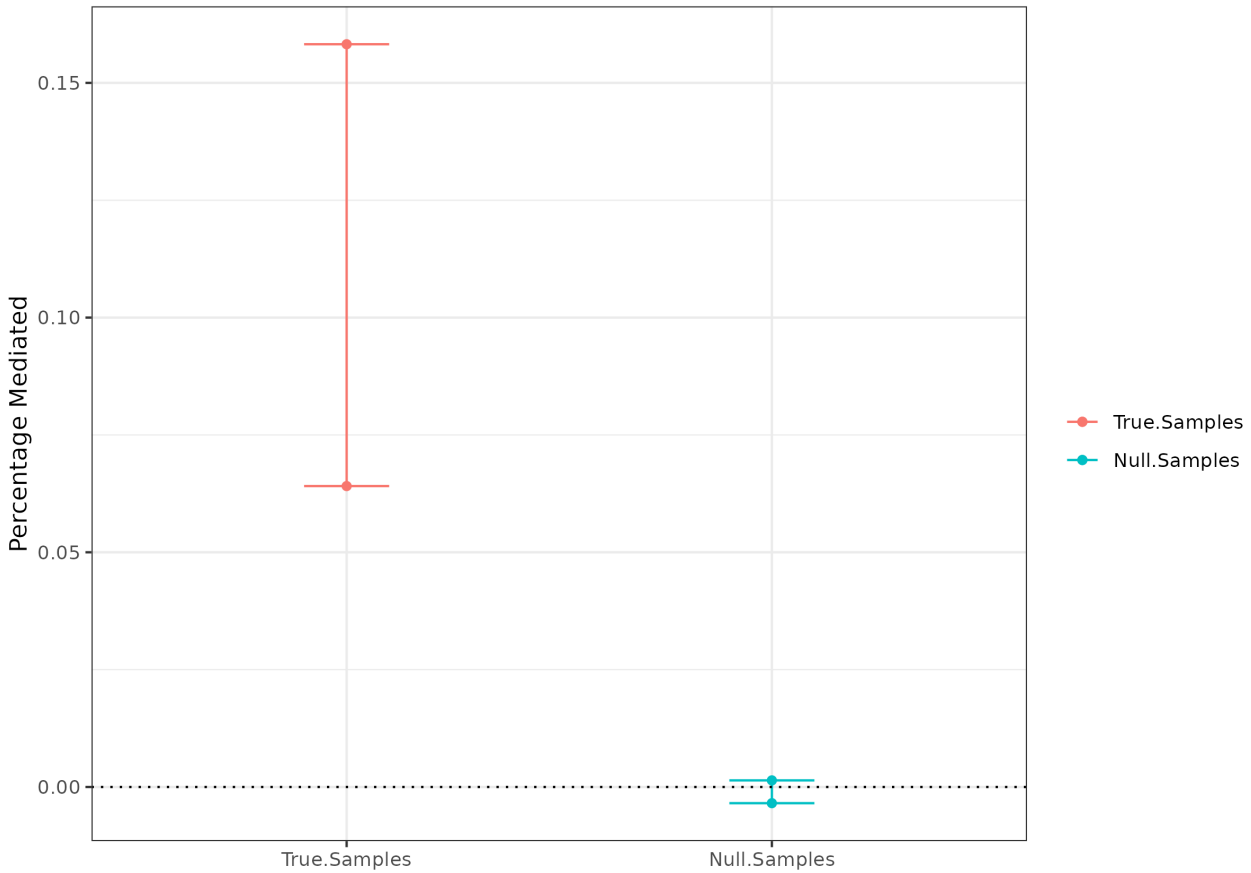


Figure: Ensemble of ResNet10, ResNet18, and ResNet34.

Ensemble of Pretrained Networks

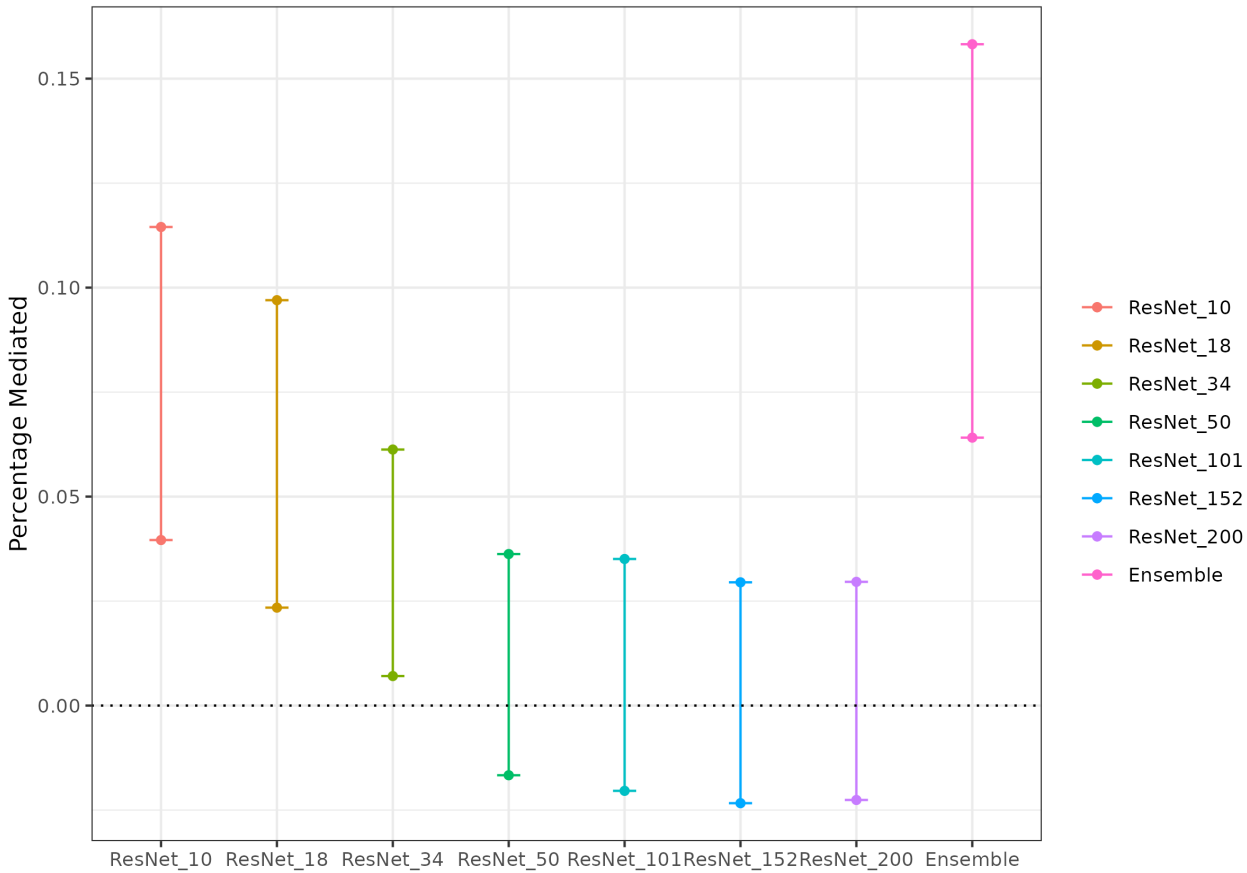


Figure: Ensemble vs single pretrained networks.

Summary

- Meta-HAL constructs more flexible super learners using HAL as the ensembler;
- L1-norm bound is the hyperparameter of meta-HAL, and can be decided with double cross-validation (for estimating the whole function) or other faster selectors (allowing slight overfitting when it is being undersmoothed for target parameters);
- Under different conditions on the meta-level data, a meta-HAL SL can achieve regular HAL rate in loss based dissimilarity or CAN plug-in estimate with undersmoothing.
- Potential advantage given high-dimensional, high-variation models.

Thank You

Thank you!

Questions?

Happy to discuss more!

Email: wangzeyi@berkeley.edu

References I

Zeyi Wang, Wenxin Zhang, and Mark van der Laan. Super ensemble learning using the highly-adaptive-lasso. *arXiv preprint arXiv:2312.16953*, 2023.

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Aurélien F Bibaut and Mark J van der Laan. Fast rates for empirical risk minimization over c\adl\ag functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*, 2019.

M.J. van der Laan, D. Benkeser, and W. Cai. Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. Technical report, Division of Biostatistics, University of California, Berkeley, 2019.

References II

Mark van der Laan, Zeyi Wang, and Lars van der Laan. Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290*, 2021.

Appendix

Super Learner (SL) based on oracle inequality

For a loss function $L : \mathcal{Q} \to L^2(P_0)$ that satisfies $Q_0 = \arg\min_{Q \in \mathcal{Q}} P_0 L(Q)$, $\sup_{Q,o} \mid L(Q)(o) \mid < \infty$, and $\sup_{Q \in \mathcal{Q}} \frac{P_0(L(Q)-L(Q_0))^2}{P_0(L(Q)-L(Q_0))} < \infty$, the cross-validated selector is asymptotically equivalent with the oracle selector from the learner library [van der Laan et al., 2007].

- The size of the learner library grows at polynomial rates.
- E.g., discrete SL.
- Convex or non-negative combinations of $\log(n)$ learners.

SL based on Highly Adaptive Lasso (HAL)

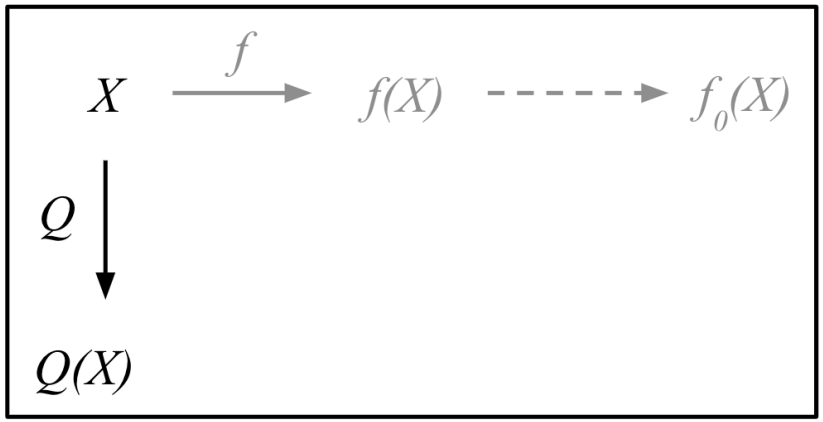
Consider d-variate functions with bounded sectional variation norms. The parametric MLE (HAL-MLE) converges to the true function at a rate of $n^{-2/3}(\log n)^d$ [Bibaut and van der Laan, 2019].

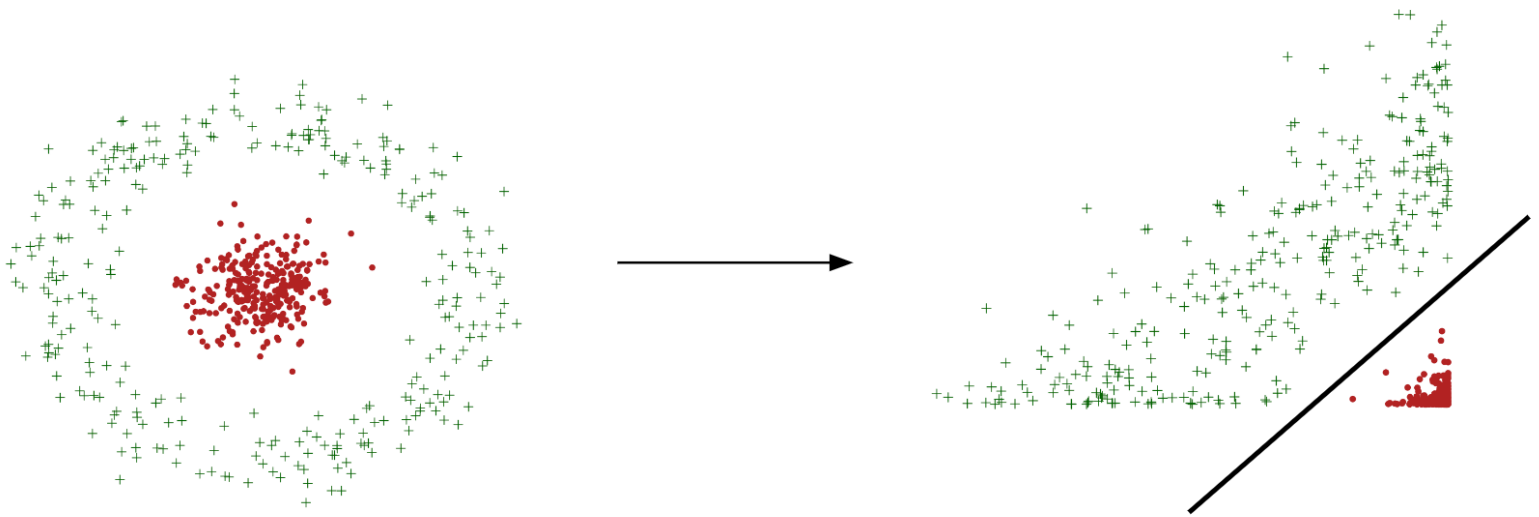
- Ensemble functions beyond convex/non-negative combinations.
- HAL-MLE can be undersmoothed to solve more score equations for semiparametric efficiency [van der Laan et al., 2019].

Meta-HAL SL vs HAL vs SL

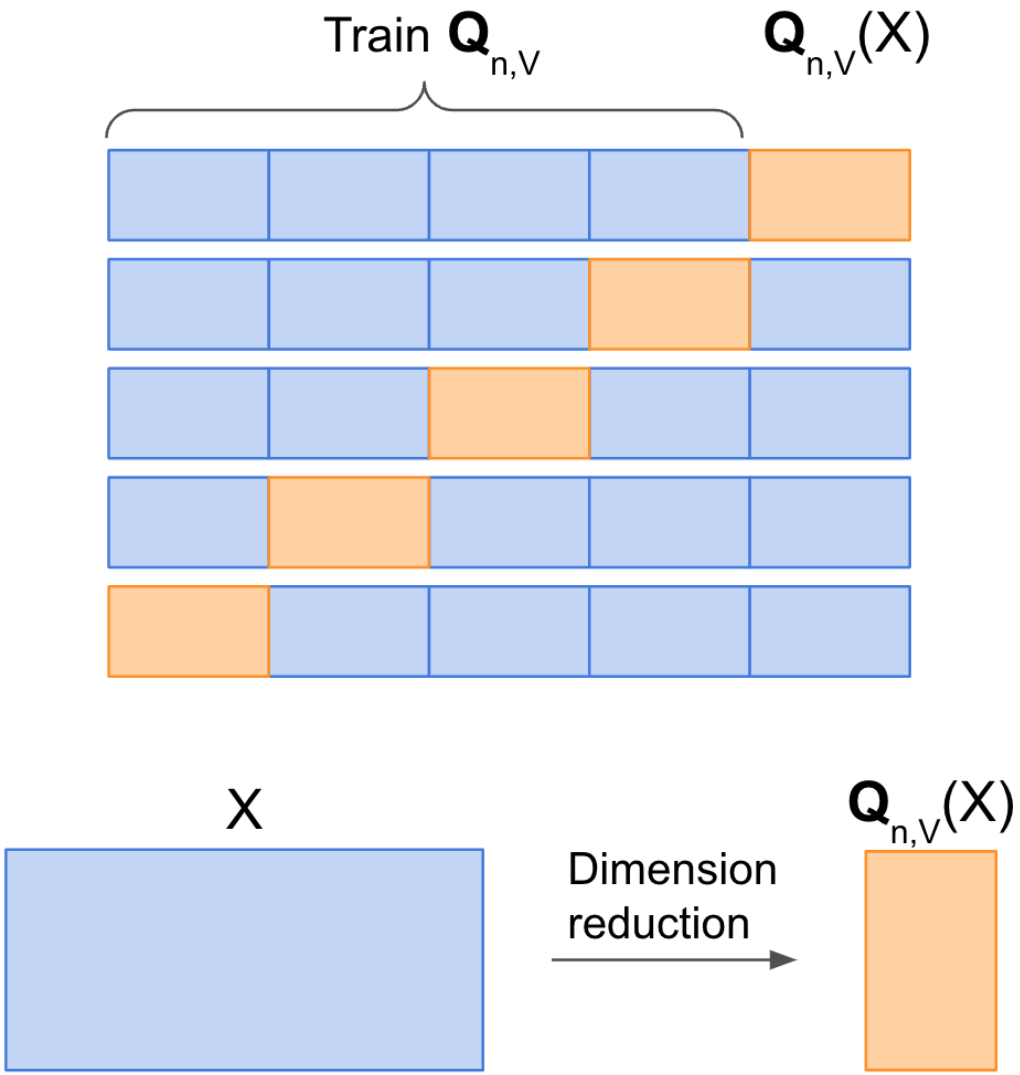
- Undersmoothed HAL does not utilize super learners. (Higher-order TMLE is another line of work; see van der Laan et al. [2021].)
- Computation costs non-linearly increase with **high-dimensional** data.
- Super learners have limitations.
  - Discrete SL can be more reliable.
  - Even the best candidate (the oracle learner) may be suboptimal; non-linear function of learners with interactions, learners with only partial information, especially common with **multi-modal** data.

Flexible Data Transformation/Coordinate Transformation

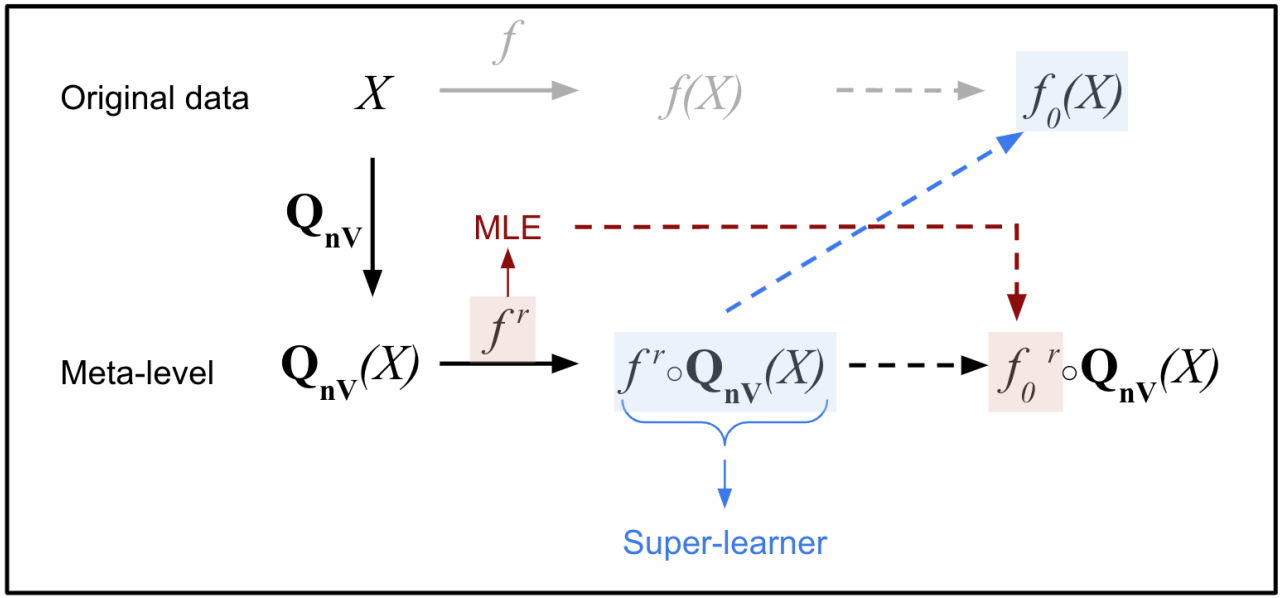




Data Transformation with Cross-validation



Meta-HAL Algorithm



Assumption: $L(f^r \circ \mathbf{Q}_{n,v})(O)$ depends on $\mathbf{Q}_{n,v}(X)$ only through $O^r(v, O)$.

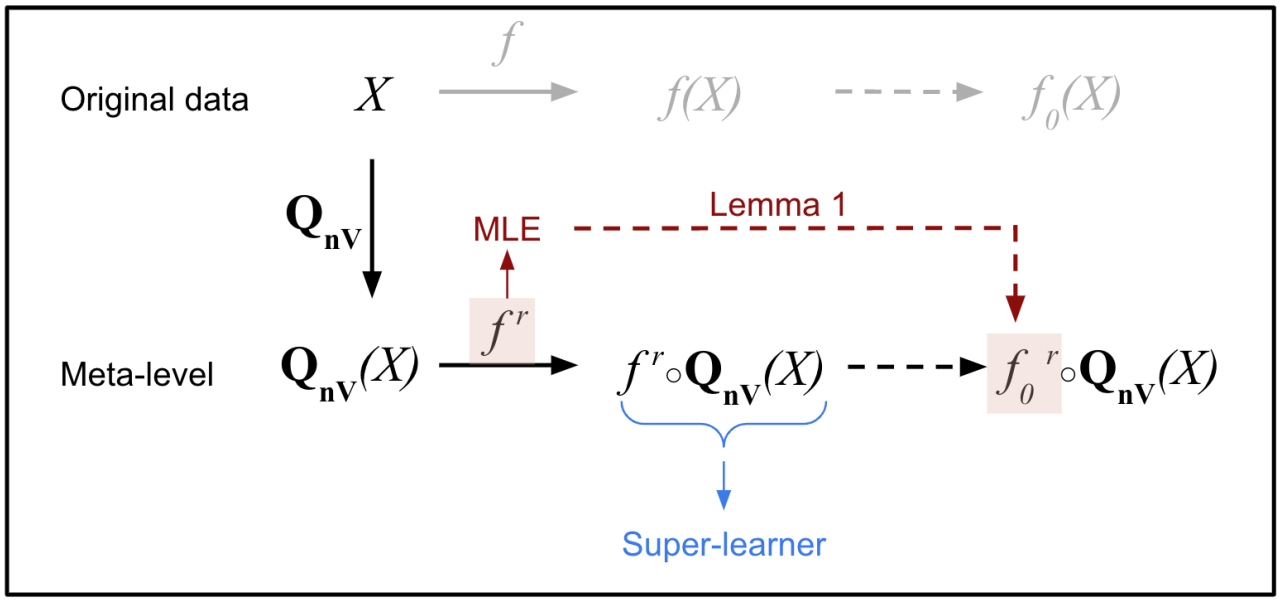
Example: $O = (X, Y)$, $O^r(v, O) = (\mathbf{Q}_{n,v}(X), Y)$, $L(f) = (Y - f(X))^2$, $L(f^r \circ \mathbf{Q}_{n,v}) = (Y - f^r \circ \mathbf{Q}_{n,v}(X))^2$.

$L^r(f^r)(O^r(v, O)) := L(f^r \circ \mathbf{Q}_{n,v})(O)$ defines a loss function.

Let $\hat{f}^r \in \mathcal{F}^r$ denote the minimum loss estimator, **meta-HAL MLE**.

Ensemble estimator $\hat{f}^r \circ \mathbf{Q}_{n,V}$ is called **meta-HAL SL**.

Meta-HAL Rates

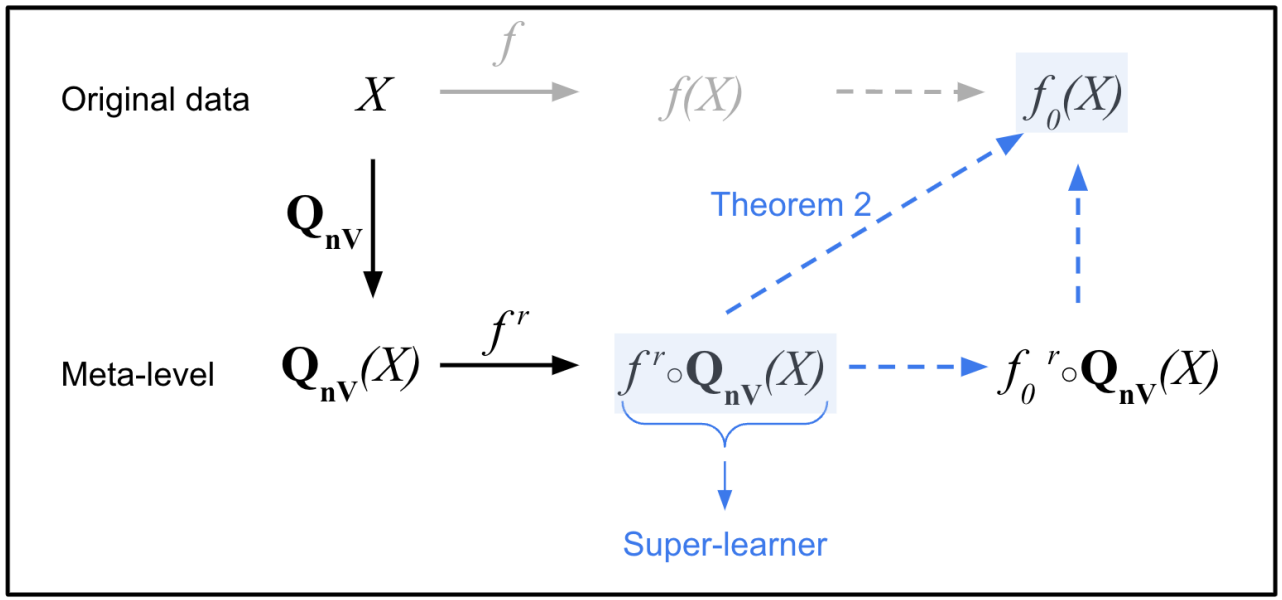


Lemma

Let d^r be the dimension of O_i^r. If $\{L^r(f^r) : f^r \in \mathcal{F}^r\} \subset \mathcal{D}_{d^r}[0, \tau^{d^r}]$, then,

$$d_0^r(\hat{f}^r, f_0^r) = O_P(n^{-2/3}(\log n)^{d^r}).$$

Meta-HAL Rates



Theorem

Suppose that $L(\cdot)$ is convex.

$$d_0\left(\frac{1}{V}\sum_{v=1}^{V}\hat{f}^r \circ \mathbf{Q}_{n,v}, f_0\right) = O_P\left(n^{-2/3}(\log n)^{d^r}\right) + \min_{f^r \in \mathcal{F}^r} \frac{1}{V}\sum_{v=1}^{V} d_0\left(f^r \circ \mathbf{Q}_{n,v}, f_0\right).$$

Undersmoothing

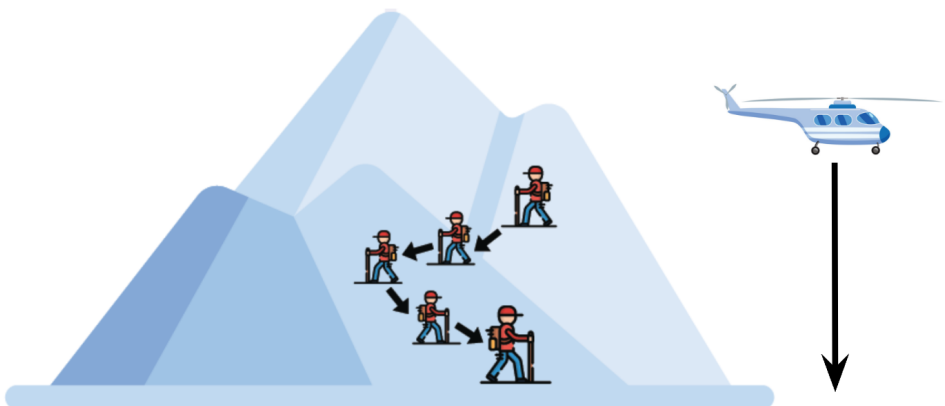
Recall $f^r \in \mathcal{F}^r = \{f^r \in \mathcal{D}_C[0,1]^J : \|f^r\|_v^* < C\}$.

Meta-HAL MLE, $\hat{f}^r = \hat{f}^r_{C_n}$, is indexed by a choice of C_n.

Undersmoothing

Under regularity conditions 1 and 2, we have

$$C_n \uparrow \quad \frac{1}{n}\sum_{i=1}^{n} D^*(\hat{f}^r_{C_n} \circ Q_{n,v_i}, g)(O_i) \downarrow$$



Adapted from Serrano L. (2021). Grokking Machine Learning.

Asymptotic Linearity Theorem

Theorem

Assume C_n is chosen large enough so that

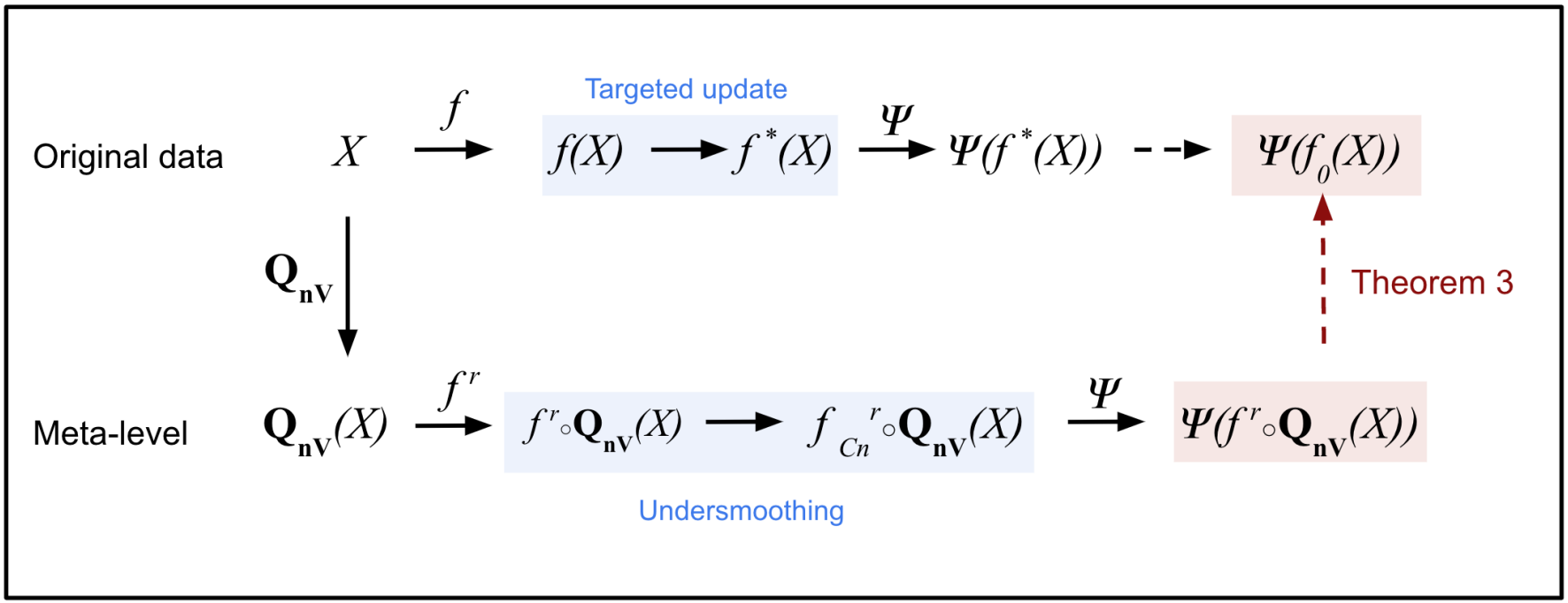
$$\frac{1}{n}\sum_{i=1}^{n} D^*(\hat{f}^r_{C_n} \circ \mathbf{Q}_{n,v_i}, \hat{g}) = o_P(n^{-1/2}),$$

where $\sup_{\{\mathbf{Q}_{n,v}:v\},v} P_0\{D^*(f_0^r \circ \mathbf{Q}_{n,v}, \hat{g}) - D^*(f_0, \tilde{g}_0)\}^2 \to_p 0$. *Then,*

$$\frac{1}{V}\sum_{v=1}^{V}\{\Psi(\hat{f}^r_{C_n} \circ \mathbf{Q}_{n,v})\} - \Psi(f_0) = \frac{1}{n}\sum_{i=1}^{n} D^*(f_0, \tilde{g}_0)(O_i) + o_P(n^{-1/2}).$$

Note: $\Psi(f_0^r \circ \mathbf{Q}_{n,v}) - \Psi(f_0) = o_P(n^{-1/2})$ under mild conditions.

Asymptotic Linearity Theorem



- Consistency: $\frac{1}{V}\sum_{v=1}^{V}\{\Psi(\hat{f}^r_{C_n} \circ \mathbf{Q}_{n,v})\} \xrightarrow{P} \Psi(f_0)$
- Asymptotic normality:
  $\sqrt{n}(\frac{1}{V}\sum_{v=1}^{V}\{\Psi(\hat{f}^r_{C_n} \circ \mathbf{Q}_{n,v})\} - \Psi(f_0)) \rightsquigarrow N(0, \Sigma)$

Full Data vs Meta-Level Data

With meta-level HAL, this suffices to have that for fixed $\mathbf{Q}_{n,v}$, $\{D^*(f^r \circ \mathbf{Q}, g^r_{n,v}) - D^*(f^r_0 \circ \mathbf{Q}_{n,v}, g^r_{0,n,v}) : f^r \in \mathcal{F}^r, g^r \in \mathcal{G}^r\}$ is a class of cadlag functions with a sectional variation norm bound not depending on $\mathbf{Q}_{n,v}$.

This is a much smaller class concerning lower-dimensional data, not over $\{D^*(f, g) : f \in \mathcal{F}, g \in \mathcal{G}\}$ for the original data.

Ensemble Learning/Meta-Learning

Consider the ensemble functions (of learners) with bounded sectional variation norms. This constructs a much larger class of ensembles than convex/non-negative combinations.

- Regular SL with constrained learner library sizes does NOT apply.
- Instead, rely on the convergence rate of HAL.

Cross-Validated Choice of L1 Norm

To fit the meta-level HAL-MLE (meta-HAL), we need

1. a cross-validation of the full data to construct coordinate transformation, and
2. a cross-validated choice of L_1-norm bounds (or equivalently a choice of λ for the L_1 regularization term).

Like any other super learner, an honest selection requires double cross-validation. But even without double cross-validation, the risk of overfitting is low with one continuous hyperparameter approximated by polynomial many grid points. Also, overfitting only leads to larger bounds on L_1-norms, which will be implemented in undersmoothing.

Comparison of Oracle Inequality and HAL Rates

Meta-HAL rates

Under mild conditions, the resulted Meta-HAL SL Q_n achieves the following decomposed risk.

$$d_0^{\bar{V}}(Q_n, Q_0) = \min_{Q^r \in \mathcal{Q}^r} \frac{1}{V} \sum_{v=1}^{V} P_0\{L(Q^r \circ \mathbf{Q}_{n,v}) - L(Q_0)\} + O_P(n^{-2/3}(\log n)^{d^r}).$$

For example, if the learner library includes an estimator $\hat{Q}_j$ such that $d_0(\hat{Q}_j(P_{n,v}), Q_0) = O_P(n^{-2/3}(\log n)^d)$, then it follows that

$$d_0^{\bar{V}}(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^{\max\{d, d^r\}}).$$

Not a necessary condition; consider an invertible transformation. In general, the risk of oracle ensemble is smaller than that of each learner.

Undersmoothing Conditions

Let D_n^r be the projection of D^r on to the linear space spanned by all the (unconstrained) score solved by meta-HAL-MLE. Assume

$$P_0^r\{D^r(\hat{f}_{C_n}^r, g_{0,n}^r) - D_n^r(\hat{f}_{C_n}^r, g_{0,n}^r)\} = o_P(n^{-1/2}), \tag{1}$$

$$C_n \min_{(s,j)\in\mathcal{J}_n(C_n)} |P_n^r \frac{d}{d\hat{f}_{C_n}^r} L^r(\hat{f}_{C_n}^r)(\phi_{s,j})| = o(n^{-1/2}). \tag{2}$$

See page 10.

Example: Treatment Specific Mean (TSM)

$O = (W, A, Y) \sim P_0 \in \mathcal{M}$.

- W: baseline,
- A: treatment (1: treated, 0: control),
- Y: outcome.

$f_0 = \mathbb{E}_{P_0}[Y|A = 1, W]$, $L(f) = A(Y - f(W))^2$.

$\Psi(f) = \mathbb{E}_{P_0} f(W)$.

$\Psi(f_0) = \mathbb{E}_{P_0}\mathbb{E}_{P_0}[Y|A = 1, W]$ is the target, TSM.

$D(f, g)(O) = \frac{A}{g(W)}(Y - f(W))$.

Double Robustness for TSM

Lemma

If $\mathbf{Q}_{n,v}(W)$ includes the correct propensity score model $g_0(W)$ (or if g_0 depends on W only through $\mathbf{Q}_{n,v}(W)$) for all $v = 1, \ldots, V$, or if $\mathbf{Q}_{n,v}(W)$ includes the correct outcome model $f_0(W)$ for all $v = 1, \ldots, V$, then we have $\frac{1}{V}\sum_{v=1}^{V} \Psi(f_0^r \circ \mathbf{Q}_{n,v}) - \Psi(f_0) = 0$. Moreover, if $\mathbf{Q}_{n,v}(W)$ includes a consistent estimator for either f_0 or g_0, then we have consistent estimation for $\Psi(f_0)$.

Key step.

$$|\Psi(f_0^r \circ \mathbf{Q}_{n,v}) - \Psi(Q_0)| \leq \|g_1 - g_0\|_{P_0} \|f_1 - f_0\|_{P_0}.$$