8th ICML Workshop on Automated Machine Learning (AutoML)

Leveraging Theoretical Tradeoffs in Hyperparameter Selection for Improved Empirical Performance

Parikshit Ram, Alexander G. Gray, Horst C. Samulowitz, **IBM Research AI**

Excess Risk in Hyperparameter Selection

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} E^{\nu}_{n_\nu}(\hat{g}_{n_t,\lambda}) \qquad \hat{g}_{n_t,\lambda} = \arg\min_{f \in \mathcal{F}_\lambda} E_{n_t}(f)$$

Outer DFO — Inner ERM

- *Hyperparameter selection*: Bilevel Optimization Problem
- *Inner level*:
  - Empirical risk minimization (ERM) problem
  - Objective depends on *training split* of the data set
- *Outer level*:
  - Derivative-free optimization (DFO) problem
  - Objective depends on *held-out validation split* of the data set

$$\mathcal{E} = E(\hat{g}_{n_t,\hat{\lambda}}) - E(f^\star)$$

Excess Risk

- *Excess risk* of model *learned via ERM* for hyperparameter *selected via DFO*
- *Both levels* of optimization use *empirical estimates* of true risk

$$E(f) = \int \ell(y, f(x))\,dP(x,y), \qquad E_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

True Risk — Empirical Risk

Additional Sources of Excess Risk

- *Inner ERM approximated*, implying following hyperparameter selection

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} E^{\nu}_{n_\nu}(\tilde{g}_{n_t,\lambda}), \quad \tilde{g}_{n_t,\lambda} \in \left\{ g \in \mathcal{F}_\lambda : E_{n_t}(g) \le \min_{f \in \mathcal{F}_\lambda} E_{n_t}(f) + \rho_{in} \right\}$$

Approximate Inner ERM

- After hyperparameter selection, *final model trained on data that combines training and validation splits*, leading to *model discrepancy*

$$\hat{f}_{n,\hat{\lambda}} = \arg\min_{f \in \mathcal{F}_{\hat{\lambda}}} E_n(f)$$

Final ERM with selected HP on full data

Contributions

- Provide novel excess risk bounds for above scenarios
- Propose novel data-driven practical heuristics for improved performance

Excess Risk with Exact ERM

No Model Discrepancy

Theorem 1 *Let $L = 2|\Lambda| + 2$. Then, with probability at least $1-\delta$ for any $\delta > 0$, the excess risk $\mathcal{E} = E(\hat{g}_{n_t,\hat{\lambda}}) - E(f^\star)$ is bounded from above as:*

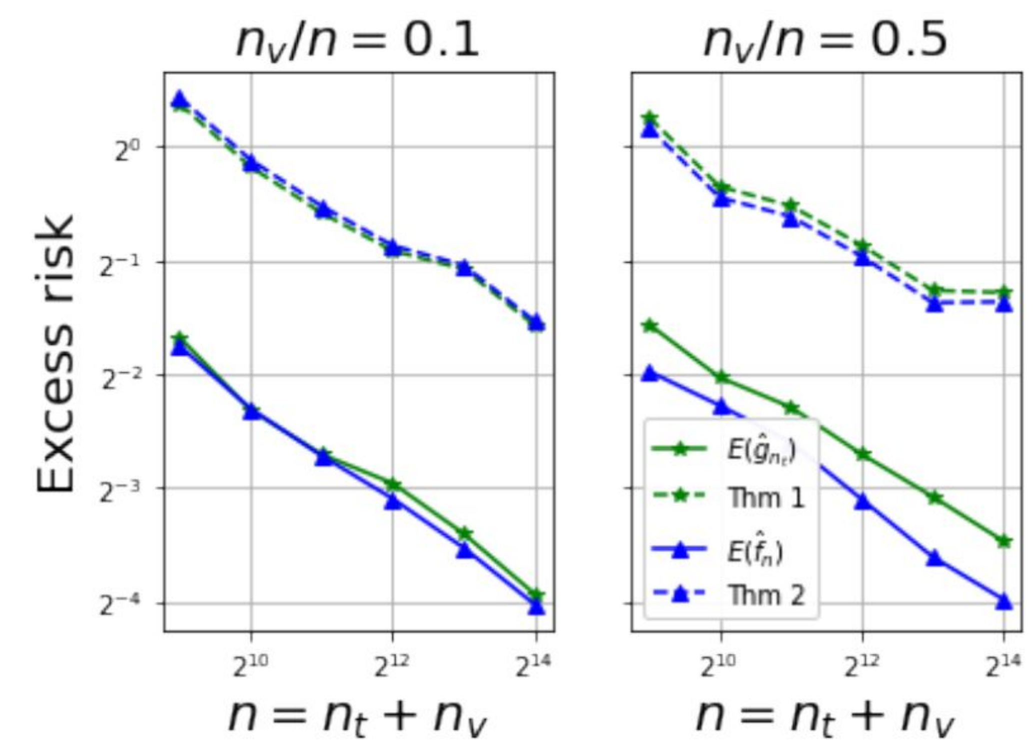
$$\mathcal{E} \le \min_{\lambda \in \Lambda} \left\{ 2\Delta(\mathcal{F}_\lambda, n_t, \delta/L_1) + \mathcal{E}_{app}(\lambda) \right\} + B\sqrt{2\log(L_1/\delta)/n_\nu}.$$

Model Discrepancy

Theorem 2 *Let $L_2 = 2|\Lambda| + 3$. Let $\mathcal{I}_{n,n_t,\hat{\lambda}} = E_n(\hat{g}_{n_t,\hat{\lambda}}) - E_n(\hat{f}_{n,\hat{\lambda}})$ denote the "empirical risk improvement" obtained by refitting the model on the full training set. Then, with probability at least $1-\delta$ for any $\delta > 0$, the excess risk $\mathcal{E} = E(\hat{f}_{n,\hat{\lambda}}) - E(f^\star)$ is bounded from above by:*

$$\mathcal{E} \le \min_{\lambda \in \Lambda} \left\{ 2\Delta(\mathcal{F}_\lambda, n_t, \delta/L_2) + \mathcal{E}_{app}(\lambda) \right\} - \mathcal{I}_{n,n_t,\hat{\lambda}} + B\sqrt{2\log(L_2/\delta)}\left(1/\sqrt{n} + 1/\sqrt{n_\nu}\right).$$

Comparing Bounds to True Excess Risk

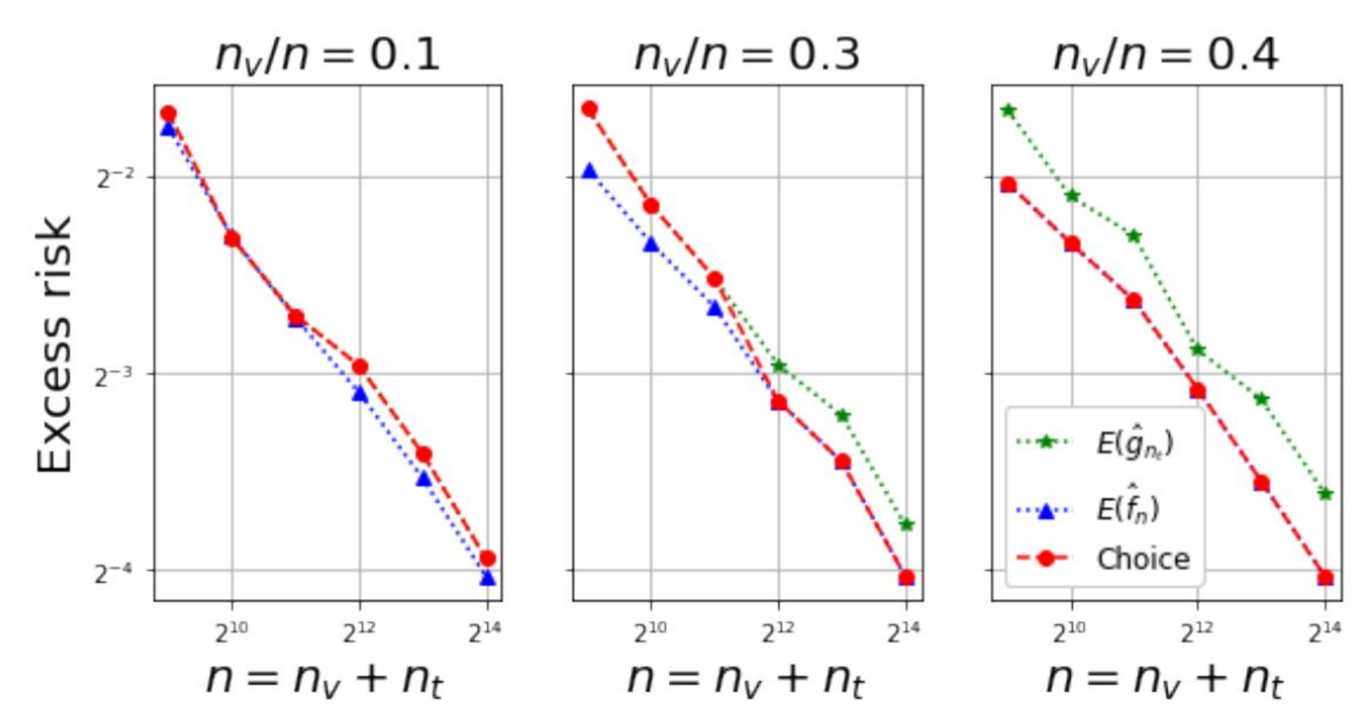


Risk bounds reflect relative performance

Data-driven Choice for Final Model

Heuristic 1 *Let us define the following data-dependent scalars α, β based on the quantities in Theorems 1 & 2, and we select $\hat{f}_{n,\hat{\lambda}}$ as the final model if $\alpha \ge \beta$, or select $\hat{g}_{n,\hat{\lambda}}$ otherwise:*

$$\alpha = B\sqrt{2\log(L_1/\delta)/n_\nu}, \quad \beta = -\mathcal{I}_{n,n_t,\hat{\lambda}} + B\sqrt{2\log(L_2/\delta)}\left(1/\sqrt{n} + 1/\sqrt{n_\nu}\right).$$



Data-driven heuristic able to match best in most cases

Excess Risk with Approximate ERM

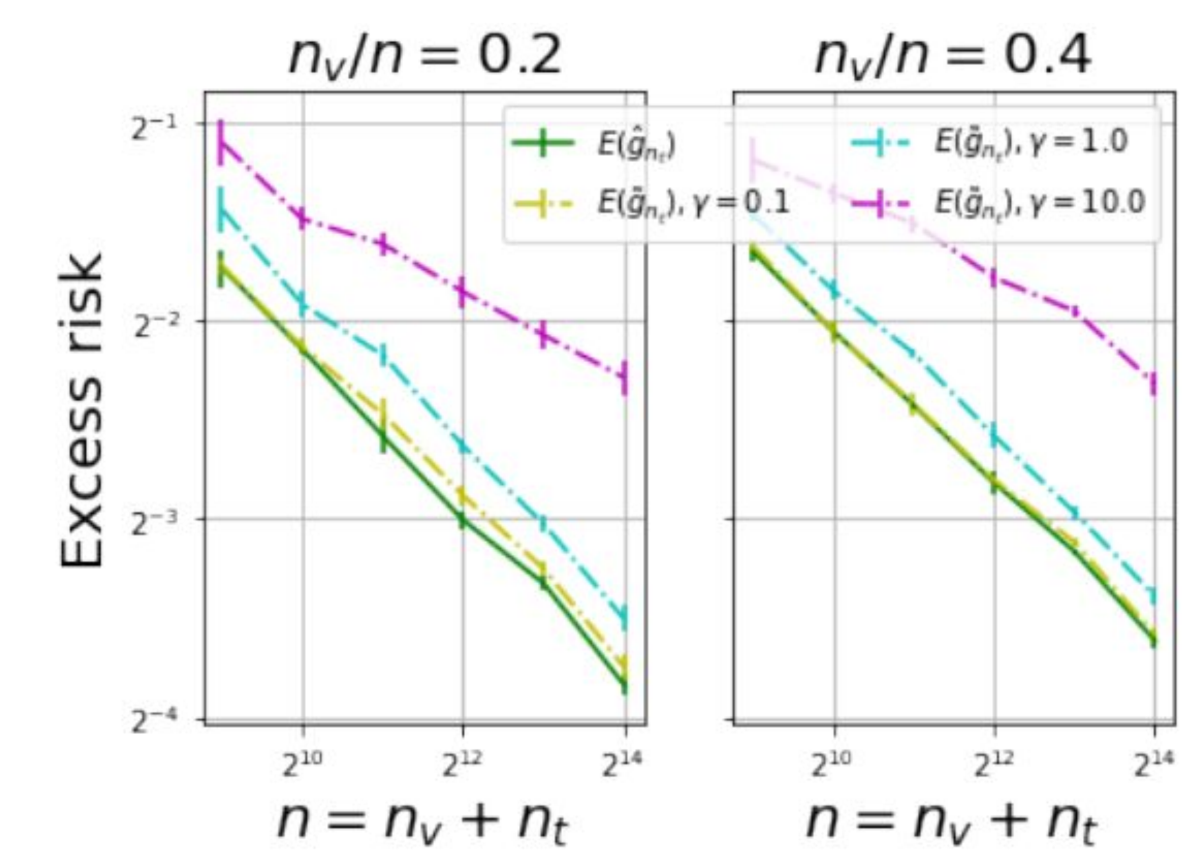
No Model Discrepancy

Theorem 3 *The excess risk $\mathcal{E} = E(\tilde{g}_{n_t,\hat{\lambda}}) - E(f^\star)$ can be bounded from above with probability at least $1-\delta$ for any $\delta > 0$ and $L_3 = (2 + 3|\Lambda|)$:*

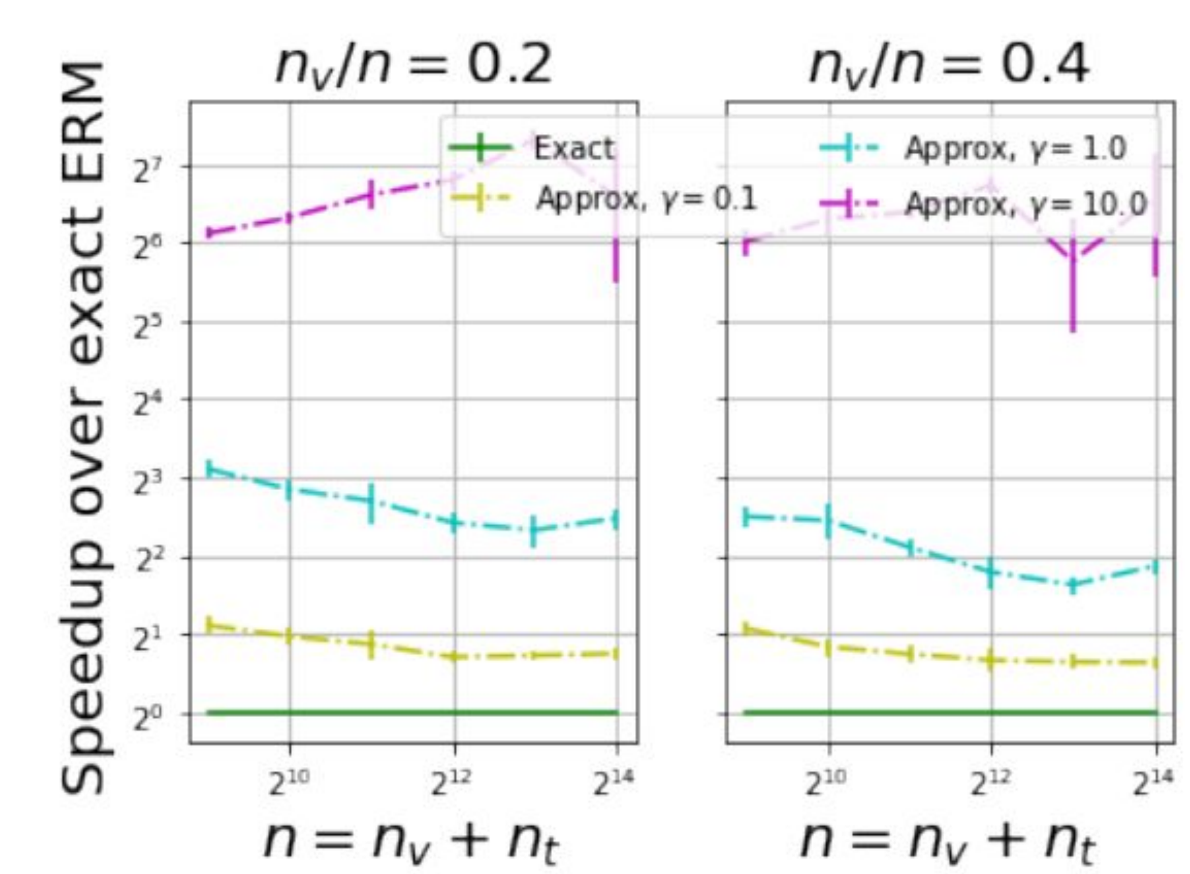
$$\mathcal{E} \le \min_{\lambda \in \Lambda} \left\{ 2\Delta(\mathcal{F}_\lambda, n_t, \delta/L_3) + \mathcal{E}_{app}(\lambda) \right\} + B\sqrt{2\log(L_3/\delta)/n_\nu} + \rho_{in}.$$

Data-driven Choice for ERM Approximation

Heuristic 2 *Based on the terms defined in Theorem 3, select a scaling parameter $\gamma > 0$ and set ρ_{in} as $\rho_{in} = \gamma B\sqrt{2\log(L_3/\delta)/n_\nu}$ such that $\rho_{in} \sim o(B\sqrt{2\log(L_3/\delta)/n_\nu})$. A value of $\gamma = 0.1$ suffices in our experience.*



Data-driven choice of approximation in inner ERM does not increase excess risk significantly over exact inner ERM



Data-driven choice of approximation in inner ERM provides 2X speedup over exact ERM with no additional excess risk, and can provide 4-6X speedup with slight increase in excess risk

Notations

$$f^\star : \mathcal{X} \to \mathcal{Y} \text{ such that } f^\star(x) = \arg\min_{\hat{y} \in \mathcal{Y}} \mathbb{E}\left[\ell(y, \hat{y}) | x\right]$$

Bayes optimal model

$$\bar{f}_\lambda = \arg\min_{f \in \mathcal{F}_\lambda} E(f)$$

True risk minimizer

$$\mathcal{E}_{app}(\lambda) = E(\bar{f}_\lambda) - E(f^\star)$$

Approximation risk

$$\sup_{f \in \mathcal{F}} |E_n(f) - E(f)| \le \Delta(\mathcal{F}, n, \delta)$$

Estimation risk bound