

Model Averaging and Double Machine Learning*

Achim Ahrens[†] Christian B. Hansen[‡] Mark E. Schaffer[§]

Thomas Wiemann[‡]

January 4, 2024

Abstract

This paper discusses pairing double/debiased machine learning (DDML) with *stacking*, a model averaging method for combining multiple candidate learners, to estimate structural parameters. We introduce two new stacking approaches for DDML: *short-stacking* exploits the cross-fitting step of DDML to substantially reduce the computational burden and *pooled stacking* enforces common stacking weights over cross-fitting folds. Using calibrated simulation studies and two applications estimating gender gaps in citations and wages, we show that DDML with stacking is more robust to partially unknown functional forms than common alternative approaches based on single pre-selected learners. We provide Stata and R software implementing our proposals.

Keywords: causal inference, partially linear model, high-dimensional models, super learners, nonparametric estimation

JEL: C21, C26, C52, C55, J01, J08

*Many thanks to Elliott Ash, Daniel Björkegren, David Cai, Ben Jann, Michael Knaus, Rafael Lalive, Moritz Marbach, Martin Huber, Blaise Melly, Gabriel Okasa, and Martin Spindler for helpful discussions and comments. We are also thankful for the helpful feedback we have received at the AI+Economics Workshop at the ETH Zürich in 2022, the Italian and Swiss Stata meetings in 2022, the 2022 Machine Learning in Economics Summer Institute in Chicago, the LISER workshop “Machine Learning in Program Evaluation, High-dimensionality and Visualization Techniques,” the IAAE Annual Conference in 2023, the London 2023 Stata meeting, the 2023 Stata Economics Virtual Symposium and the European Summer Meetings of the Econometric Society in 2023. All remaining errors are our own. An earlier version of the paper was presented under the title “A Practitioners’ Guide to Double Machine Learning.”

[†]ETH Zürich, Switzerland. *Email:* achim.ahrens@gess.ethz.ch

[‡]University of Chicago, United States. *Email:* Christian.Hansen@chicagobooth.edu (Hansen), wiemann@uchicago.edu (Wiemann).

[§]Heriot-Watt University, Edinburgh, United Kingdom and IZA Institute of Labor Economics. *Email:* M.E.Schaffer@hw.ac.uk.

1 Introduction

Motivated by their robustness to partially unknown functional forms, supervised machine learning estimators are increasingly leveraged for causal inference. For example, lasso-based approaches such as the post-double-selection lasso (PDS lasso) of Belloni, Chernozhukov, and Hansen (2014) have become popular estimators of causal effects under conditional unconfoundedness in applied economics (e.g. Gilchrist and Sands, 2016; Dhar, Jain, and Jayachandran, 2022). Yet, a recent literature also raises practical concerns about the use of machine learning for causal inference. Wüthrich and Zhu (2021) find that lasso often fails to select relevant confounds in small samples while inference based on linear regression performs relatively well. Giannone, Lenza, and Primiceri (2021) and Kolesár, Müller, and Roelsgaard (2023) argue that the sparsity assumption, on which the lasso fundamentally relies, is frequently not plausible in economic data sets. Angrist and Frandsen (2022) show that conditioning on confounders using random forests may yield spurious results in IV regressions.¹ In an application to the evaluation of active labor market programs, Goller et al. (2020) find that random forests are not suitable for the estimation of propensity scores. A key characteristic shared by many of these studies using machine learning for causal inference is the focus on a single pre-selected machine learner.

This paper revisits the application of machine learning for causal inference in light of this recent literature. In particular, we highlight the benefits of pairing double/debiased machine learning (DDML) estimators of Chernozhukov et al. (2018) with stacking (Wolpert, 1996; Breiman, 1996; Laan, Polley, and Hubbard, 2007). DDML can leverage generic machine learners meeting mild convergence rate requirements for the estimation of common (causal) parameters. Stacking improves the robustness to the underlying structure of the data by allowing the researcher to combine multiple candidate estimators with differing strengths rather than requiring an *ad-hoc* choice between them. Based on a diverse set of applications and calibrated simulation studies, we illustrate the finite sample perfor-

¹See also Angrist (2022) for additional discussion.

mance of stacking-based DDML estimators. The results suggest that stacking with a rich set of candidate estimators can address some of the shortcomings highlighted in the recent literature on causal inference with single pre-selected machine learners.

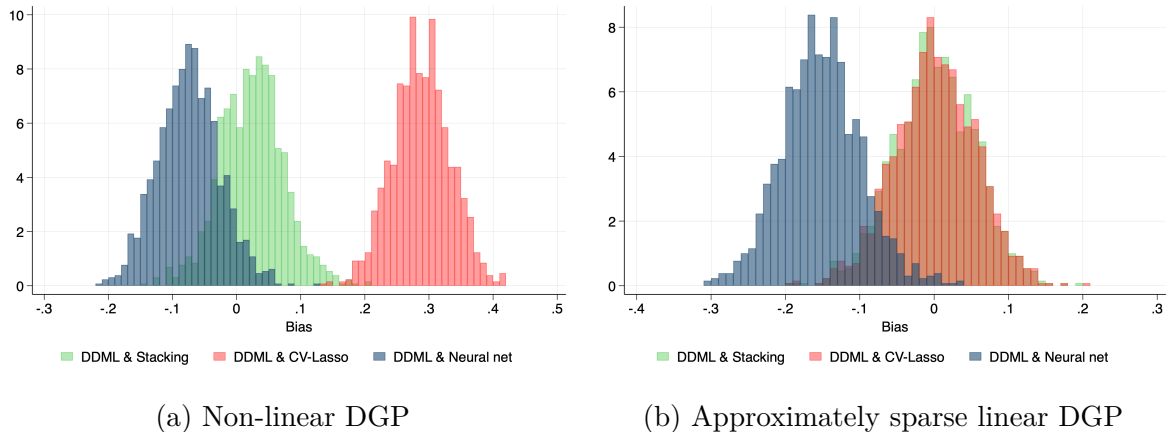
We further introduce two novel ways of combining stacking and DDML aimed at improving practical feasibility and stability in finite samples: *Short-stacking* leverages the cross-fitting step of DDML to reduce the computational burden of stacking substantially. *Pooled stacking* decreases the variance of stacking-based learners across the DDML cross-fitting folds. Both approaches facilitate interpretability compared to conventional stacking by enforcing common stacking weights. We complement the paper with software packages for Stata and R that implement the proposed approaches (Ahrens et al., 2023; Wiemann et al., 2023).

The benefits of combining multiple estimators into a ‘super learner’ via stacking to improve robustness to the structure of the underlying data-generating process are well-known in the statistics literature. Loss-minimizing combinations of a pre-specified set of estimators were introduced by Wolpert (1992) and Breiman (1996) and generalized by Laan, Polley, and Hubbard (2007). Under appropriate restrictions on the data generating process and loss-function, Laan and Dudoit (2003) show asymptotic equivalence between stacking and the best-performing candidate learner.²

Despite its theoretical appeal, stacking has hitherto been rarely used for the estimation of causal effects in economics or other social sciences. Instead, estimators are often based on parametric (frequently linear) specifications or single pre-selected machine learners. This can have severe consequences for the properties of causal effect estimators if the given choice is ill-suited for the application at hand. A simple example is shown in Figure 1 which compares the performance of DDML using either cross-validated (CV) lasso or a feed-forward neural network to estimate a partially linear model across two different data-generating processes. The results show that the bias associated with each learner strongly depends on the structure of the data. Since true functional forms are often unknown in

²See also Hansen and Racine (2012) for discussion of jackknife (leave-one-out) stacking. Hastie, Tibshirani, and Friedman (2009) and Laan and Rose (2011) provide textbook treatments of stacking and super learning.

Figure 1: Estimation bias of DDML with cross-validated lasso, feed-forward neural net and stacking



Notes: The figures compare the bias of DDML paired with either cross-validated lasso, a feed-forward neural net (with two hidden layers of size 20) or a stacking learner combining 13 candidate learners (including cross-validated lasso and ridge, random forests, gradient-boosted trees and feed-forward neural nets). See Ahrens et al. (2023), where this example is taken from, for details on the specification of each learner. With respect to the data-generating processes, we generate 1000 samples of size $n = 1000$ using the PLM $Y_i = \theta_0 D_i + c_Y g(X_i) + \varepsilon_i$, $D_i = c_D g(X_i) + u_i$ where X_i are drawn from $\mathcal{N}(0, \Sigma)$ with $\Sigma_{i,k} = 0.5^{|j-k|}$, ε_i and u_i are drawn from standard normal distributions. In Figure (a), the nuisance function is $g(X_i) = X_{i,1}X_{i,2} + X_{i,3}^2 + X_{i,5}X_{i,5} + X_{i,6}X_{i,7} + X_{i,8}X_{i,9} + X_{i,10} + xX_{i,11}^2 + X_{i,12}X_{i,13}$. In Figure (b), the nuisance function is $g(X_i) = \sum_j 0.9^j X_{ij}$. c_Y and c_D are two constants chosen to ensure that the R^2 of the regression of Y onto X is approximately 0.5.

the social sciences, indiscriminate choices of machine learners in practice can thus result in poor estimates. DDML with stacking is a practical solution to this problem. As the example showcases, DDML using stacking is associated with low bias when considering a rich set of candidate learners that are individually most suitable to different structures of the data.

We conduct simulation studies calibrated to real economic datasets to demonstrate that stacking approaches can safeguard against ill-chosen or poorly tuned estimators in practical settings. Throughout, stacking estimators are associated with relatively low bias regardless of the simulated data-generating process, strongly contrasting the data-dependent performance of the causal effect estimators based on single pre-selected learners. The proposed stacking approaches thus appear relevant in the ubiquitous scenario where there is uncertainty about the set of control variables, correct functional form or the appropriate regularization assumption.

By revisiting the simulation design of Wüthrich and Zhu (2021), we further show that stacking can outperform linear regression for even small sample sizes. We argue

that the poor small sample performance of lasso-based approaches is partially driven by the choice of covariate transformations and illustrate how stacking can accommodate a richer set of specifications, including competing parametric models. We also find that short-stacking and pooled stacking outperform DDML paired with conventional stacking in small to moderate sample sizes. Paired with its lower computational cost, this finding suggests that short-stacking may be an attractive baseline approach to select and combine competing reduced form specifications.

Finally, we demonstrate the value of pairing of DDML with stacking with two applications. First, we examine gender gaps in citations of articles published in top-30 economic journals from 1983 to 2020, and assess how the difference in citations change when conditioning on content and quality proxied by the abstract text. Estimating these conditional differences is a challenging statistical problem due to the non-standard nature of text data, which is increasingly encountered in economic applications (see also e.g., Ash and Hansen, 2023; Chen and Ornaghi, 2023; Eberhardt, Facchini, and Rueda, 2022). Second, we revisit a UK sample of the OECD Skill Survey to estimate semiparametric Kitagawa-Oaxaca-Binder estimates of the unexplained gender wage gap. Both applications highlight that estimators of structural parameters based on single learners can be highly sensitive to the underlying structure of the data and/or poor tuning. The applications further demonstrate that DDML with stacking is a simple and practical solution to resolve the difficult problem of choosing a particular candidate learner in practice. Further, we observe that the optimal stacking weights often vary across reduced-form equations – meaning that different conditional expectation functions in the same data set are best estimated using different learners. This behavior sharply contrasts with common estimation approaches, such as OLS and PDS lasso, that impose the same form for each conditional expectation function.

The remainder of the paper is organized as follows: Section 2 provides a brief review of DDML. Section 3 discusses DDML with stacking, short-stacking, and pooled stacking. Section 4 presents our calibrated simulation studies. Section 5 discusses the applications, and Section 6 concludes.

2 Double/Debiased Machine Learning

This section outlines double/debiased machine learning as discussed in Chernozhukov et al. (2018). Throughout, we focus on the partially linear model as a natural extension of commonly applied linear regression methods. Despite its simplicity, the partially linear model illustrates practical challenges in the application of DDML that can be addressed by stacking. We highlight, however, that our discussion also applies to the wide range of models outlined in Chernozhukov et al. (2018) and more generally to estimation of low-dimensional structural parameters in the presence of high-dimensional nuisance functions.³

The partially linear model is defined by a random vector (Y, D, X^\top, U) with joint distribution characterized by

$$Y = \theta_0 D + g_0(X) + U, \tag{1}$$

where Y is the outcome, D is the scalar variable of interest, and X is a vector of control variables. The parameter of interest θ_0 and the unknown nuisance function g_0 are such that the corresponding residual U satisfies the conditional orthogonality property $E[Cov(U, D|X)] = 0$. These properties are analogous to the orthogonality properties of residuals in multiple linear regression with the key difference here being that g_0 need not be linear in the controls.

Albeit a seemingly small change in specification, the partially linear model has several important advantages over linear regression. For discrete D , for example, results in Angrist and Krueger (1999) imply that θ_0 can be interpreted as a positively weighted average of incremental changes in the conditional expectation function $E[Y|D = d, X]$.

Under appropriate conditional unconfoundedness assumptions, θ_0 thus corresponds to a

³A key example not explicitly discussed in Chernozhukov et al. (2018) is doubly-robust estimation of difference-in-difference parameters with staggered treatment assignment as in Callaway and Sant'Anna (2021) and Chang (2020). In settings with conditional parallel trends assumptions, high-dimensional nuisance functions arise in the estimation of group-time specific average treatment effect on the treated. The pairing of DDML and stacking, as proposed in this paper, also directly applies to the estimator of Callaway and Sant'Anna (2021) under a conditional unconfoundedness assumption.

convex combination of conditional average treatment effects.⁴ Importantly, these interpretations remain valid even if the additive separability assumption of the partially linear model fails. Linear regression coefficients, in contrast, do not correspond to positively weighted averages of causal effects without imposing strong linearity assumptions that are questionable in real applications.⁵

The advantages of the partially linear model in the interpretation of its parameter of interest come at the cost of a more challenging estimation problem relative to estimating a model that is linear in a pre-specified set of variables. Estimators for θ_0 are based on the solution to the moment equation

$$E[(Y - \ell_0(X) - \theta_0(D - m_0(X)))(D - m_0(X))] = 0,$$

given by

$$\theta_0 = \frac{E[(Y - \ell_0(X))(D - m_0(X))]}{E[(D - m_0(X))^2]},$$

where $\ell_0(X) \equiv E[Y|X]$ and $m_0(X) \equiv E[D|X]$ are the conditional expectations of the outcome and variable of interest given the controls, respectively. Since conditional expectation functions are high-dimensional in the absence of strong functional form assumptions, a sample analogue estimator for θ_0 requires nonparametric first-step estimators for the nuisance parameters ℓ_0 and m_0 . While nonparametric estimation generally reduces bias compared to linear regression alternatives, the increased variance associated with more flexible functional form estimation introduces additional statistical challenges: To allow for statistical inference on θ_0 , the nonparametric estimators need to converge sufficiently quickly to the true conditional expectation functions as the sample size increases.

⁴Similarly, for continuous D , θ_0 corresponds to a positively weighted average of derivatives of the conditional expectation function $E[Y|D = d, X]$ with respect to d . Under a conditional unconfoundedness assumption, θ_0 is thus a convex combination of derivatives of the causal response function.

⁵In the context of IV estimation where instrument validity relies on observed confounders, Blandhol et al. (2022) emphasize that, in the absence of strong functional form assumptions, two stage least squares does not generally correspond to a convex combination of local average treatment effects (LATE). However, the IV analogue to the partially linear model discussed here does admit a causal interpretation under the LATE assumptions.

DDML defines a class of estimators that allows for statistical inference on the parameter of interest θ_0 while only imposing relatively mild convergence requirements on the nonparametric estimators. These mild requirements are central to the wide applicability of DDML as they permit the use of a large variety of machine learners.⁶

Two key devices permit the mild convergence requirements of DDML: Identification of the parameter of interest based on Neyman-orthogonal moment conditions and estimation using cross-fitting. Neyman-orthogonal moment conditions are insensitive to local perturbations around the true nuisance parameter.⁷ Cross-fitting is a sample-splitting approach that addresses the *own-observation bias* that arises when the nuisance parameter estimation and the estimation of θ_0 are applied to the same observation. In practice, cross-fitting is implemented by randomly splitting a sample $\{(Y_i, D_i, X_i^\top)\}_{i \in I}$ indexed by $I = \{1, \dots, n\}$ into K evenly-sized folds, denoted as I_1, \dots, I_K . For each fold k , the conditional expectations ℓ_0 and m_0 are estimated using only observations not in the k th fold — i.e., in $I_k^c \equiv I \setminus I_k$ — resulting in $\hat{\ell}_{I_k^c}$ and $\hat{m}_{I_k^c}$, respectively, where the subscript I_k^c indicates the subsample used for estimation. The out-of-sample predictions for an observation i in the k th fold are then computed via $\hat{\ell}_{I_k^c}(X_i)$ and $\hat{m}_{I_k^c}(X_i)$. Repeating this procedure for all K folds then allows for computation of the DDML estimator for θ_0 :

$$\hat{\theta}_n = \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\ell}_{I_{k_i}^c}(X_i)) (D_i - \hat{m}_{I_{k_i}^c}(X_i))}{\frac{1}{n} \sum_{i=1}^n (D_i - \hat{m}_{I_{k_i}^c}(X_i))^2},$$

where k_i denotes the fold of the i th observation.

⁶The exact convergence rate requirement for nonparametric estimators depends on the parameter of interest. Chernozhukov et al. (2018) name the crude rate requirement of $o(n^{-1/4})$, but provide examples where the rate requirement is considerably weaker. Recent contributions show that these requirements are satisfied by specific instances of machine learners; see, e.g., results for lasso (Bickel, Ritov, and Tsybakov, 2009; Belloni et al., 2012), random forests (Wager and Walther, 2016; Wager and Athey, 2018; Athey, Tibshirani, and Wager, 2019), neural networks (Schmidt-Hieber, 2020; Farrell, Liang, and Misra, 2021), and boosting (Luo, Spindler, and Kück, 2022). The exact asymptotic properties of many other machine learners remain an active research area.

⁷In the context of the partially linear model, the formal Neyman-orthogonality requirement is

$$0 = \frac{\partial}{\partial \lambda} E \left[(Y - \{\ell_0(X) + \lambda(\ell(X) - \ell_0(X))\}) - \tau_0(D - \{m_0(X) + \lambda(m(X) - m_0(X))\}) \right. \\ \left. \times (D - \{m_0(X) + \lambda(m(X) - m_0(X))\}) \right] \Big|_{\lambda=0}$$

for arbitrary measurable functions ℓ and m , which can easily be verified using properties of the residuals.

Since the cross-fitting algorithm depends on the randomized fold split, and since some machine learners rely on randomization too, DDML estimates vary with the underlying random-number generator and seed. To reduce dependence on randomization, it is thus worthwhile to repeat the cross-fitting procedure and apply mean or median aggregation over DDML estimates (see Remark 2 in Ahrens et al., 2023). We show in Section 5 that repeating the cross-fitting procedure is a useful diagnostic tool, allowing to gauge the stability of DDML estimators.

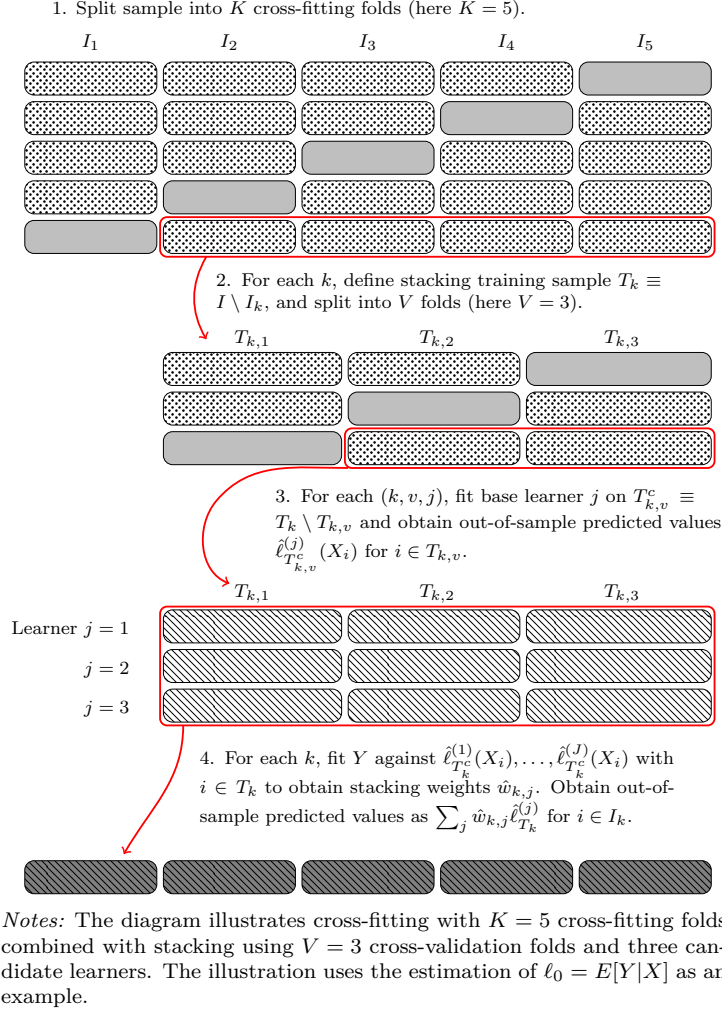
Under the conditions of Chernozhukov et al. (2018) — including, in particular, the convergence requirements on the nonparametric estimators — $\hat{\theta}_n$ is root- n asymptotically normal around θ_0 . As already highlighted by the example in Figure 1, however, a poorly chosen or poorly tuned machine learner for the estimation of nuisance parameters $\hat{\ell}$ and \hat{m} can have detrimental effects on the properties of $\hat{\theta}_n$. Since no machine learner can be best across all settings, this raises the difficult question of which learner to apply in a particular setting. In the next section, we discuss how DDML can be paired with stacking to provide a practical solution to the choice of learner. We also illustrate how the cross-fitting structure naturally arising in DDML estimators can be leveraged to substantially reduce the computational burden otherwise associated with stacking.

3 Pairing DDML with Stacking Approaches

This section discusses the estimation of structural parameters by pairing DDML with stacking approaches. After the discussion of DDML with conventional stacking, we introduce two stacking variants that leverage the cross-fitting structure of DDML estimators: short-stacking and pooled stacking. To fix ideas, we focus on the nuisance parameter $\ell_0(X) = E[Y|X]$ arising in the partially linear model where we consider an i.i.d. sample $\{(Y_i, X_i)\}_{i \in I}$. Further, we consider a rich set of J pre-selected base or candidate learners. The set of learners could include distinct parametric and nonparametric estimators — e.g., linear or logistic regression, regularized regression such as the lasso, or tree-based methods such as random forests — as well as the same algorithm with varying (hyper-

)tuning parameters or different (basis) expansions of the control variables. It is important to note that the set of candidate learners for stacking can readily incorporate commonly used unregularized learners such as linear or logistic regression; in practice, sometimes the best-performing candidate learner may be one such learner.

Figure 2: Cross-fitting with conventional stacking



DDML with conventional stacking. Combining DDML with conventional stacking involves two layers of re-sampling, as we illustrate in Figure 2. The *cross-fitting layer* divides the sample into K cross-fitting folds, denoted by I_1, \dots, I_K . In each cross-fitting step $k \in \{1, \dots, K\}$, the stacking learner is trained on the training sample which excludes fold I_k and which we label $T_k \equiv I \setminus I_k$. Fitting the stacking learner, in turn, requires subdividing the training sample T_k further into V cross-validation folds. This second sample split constitutes the *cross-validation layer*. We denote the cross-validation folds in cross-

fitting step k by $T_{k,1}, \dots, T_{k,V}$. Each candidate learner $j \in \{1, \dots, J\}$ is cross-validated on these folds, yielding cross-validated predicted values for each learner.

The final learner fits the outcome Y_i against the cross-validated predicted values of each candidate learner. The most common choice is to construct a convex combination via constrained least squares (CLS), with weights restricted to be non-negative and summing to one. Specifically, for each k , candidate learners are combined to solve

$$\min_{w_{k,1}, \dots, w_{k,J}} \sum_{i \in T_k} \left(Y_i - \sum_{j=1}^J w_{k,j} \hat{\ell}_{T_{k,v(i)}}^{(j)}(X_i) \right)^2 \quad \text{s.t. } w_{k,j} \geq 0, \sum_{j=1}^J |w_{k,j}| = 1.$$

Here, $\hat{\ell}_{T_{k,v(i)}}^{(j)}(X_i)$ denotes the out-of-sample predicted value for observation i , which is calculated from training candidate learner j on sub-sample $T_{k,v(i)}^c \equiv T_k \setminus T_{k,v(i)}$, i.e., all step- k cross-validation folds but fold $(k, v(i))$ which is the fold of the i th observation. We call the resulting $\hat{w}_{k,j}$ the *stacking weights*. The stacking predictions are obtained as $\sum_j \hat{w}_{k,j} \hat{\ell}_{T_k}^{(j)}(X_i)$ where each learner j is re-fit on T_k .

Although various options for combining candidate learners are available, CLS facilitates the interpretation of stacking as a weighted average of candidate learners (Hastie, Tibshirani, and Friedman, 2009). Due to this constraint, CLS tends to set some stacking weights to exactly zero. The constraint also regularizes the final estimator, which is important to mitigate issues arising from potential multicollinearity of the candidate learners. An alternative to CLS, which we refer to as *single-best learner*, is to impose the constraint that $w_{k,j} \in \{0, 1\}$ and $\sum_j w_{k,j} = 1$, implying that only the candidate learner with lowest cross-validated loss is used as the final estimator. Under appropriate restrictions on the data-generating process and loss function, Laan and Dudoit (2003) show asymptotic equivalence between stacking and the best-performing candidate learner.⁸

A drawback of DDML with stacking is its computational complexity. Considering the estimation of a single candidate learner as the unit of complexity (and ignoring the cost of fitting the final learner), DDML with stacking heuristically has a computational cost

⁸The *scikit-learn* (Buitinck et al., 2013) routines `StackingRegressor` and `StackingClassifier` implement stacking for Python. In Stata, stacking regression and classification are available via `pystacked`, which is a Stata front-end for these Python routines (Ahrens, Hansen, and Schaffer, 2022).

proportional to $K \times V \times J$. For example, when considering DDML with $K = 5$ cross-fitting folds and $J = 10$ candidate learners that are combined based on $V = 5$ fold cross-validation, more than 250 candidate learners need to be individually estimated. Although DDML with stacking is “embarrassingly parallel” and can thus be expected to decrease in computational time nearly linearly in the number of available computing processes, the increased complexity limits its application to moderately complex applications. Another potential concern (which we investigate in Section 4.2) is that DDML with stacking might not perform well in small samples, given that candidate learners are effectively trained on approximately $\frac{(K-1)(V-1)}{KV}\%$ of the full sample (see Figure 2). These two concerns motivate *short-stacking*.

DDML with short-stacking. In the context of DDML, we propose to take a shortcut: Instead of fitting the final learner on the cross-validated fitted values in each step k of the cross-fitting process, we can directly train the final learner on the cross-fitted values using the full sample; see Figure 3. Formally, candidate learners are then combined to solve

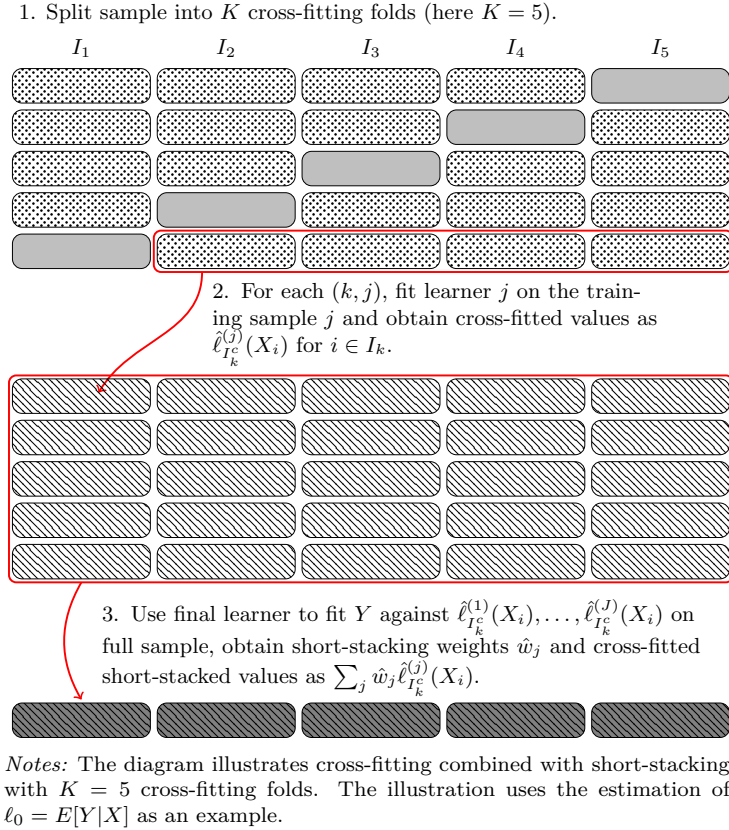
$$\min_{w_1, \dots, w_J} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^J w_j \hat{\ell}_{I_{k(i)}^c}^{(j)}(X_i) \right)^2 \quad \text{s.t. } w_j \geq 0, \sum_j |w_j| = 1$$

where w_j are the short-stacking weights. Cross-fitting thus serves a double purpose: First, it avoids the own-observation bias by avoiding overlap between the samples used for estimating high-dimensional nuisance functions and the samples used for estimating structural parameters. Second, it yields out-of-sample predicted values which we leverage for constructing the final stacking learner. As a consequence, the computational cost of DDML with short stacking is heuristically only proportional to $K \times J$ in units of estimated candidate learners. In the example from the previous paragraph, short-stacking thus requires estimating about 200 fewer candidate learners.

We recommend DDML with short-stacking in settings where the number of candidate learners is small relative to the sample size, i.e., $J \ll n$. If instead the number of considered learners is very large relative to the sample size — i.e., settings in which

inference for standard linear regression on J variables is invalid — pairing DDML with short-stacking may introduce bias.⁹ Since current applications of machine learning in economics and other social sciences rarely consider more than a few candidate learners, however, this is unlikely to be a strong concern in practice.

Figure 3: Cross-fitting with *short-stacking*



DDML with pooled stacking. While DDML with conventional stacking has one vector of weights per cross-fitting fold, short-stacking yields a single weight for each learner. A single weight for each learner decreases the variance of the final estimator and facilitates the interpretation of the stacking weights. Another way of achieving common stacking weights is DDML with pooled stacking. Pooled stacking relies on the two-layer

⁹Suppose, for simplicity, we consider ordinary (unconstrained) least squares as the final learner. Heuristically, the regression of Y_i against J sets of cross-fitted predicted values is akin to a conventional least squares regression of Y_i against J observed regressors where good performance would require $J/n \rightarrow 0$, ignoring that the cross-fitted predicted values are estimated. The additional regularization by constrained least squares should further weaken this rate requirement.

re-sampling strategy outlined above, but combines candidate learners to solve

$$\min_{w_1, \dots, w_J} \sum_{i \in I} \sum_{k \neq k(i)} \left(Y_i - \sum_{j=1}^J w_j \hat{\ell}_{T_{k,v(i)}^c}^{(j)}(X_i) \right)^2 \quad \text{s.t. } w_j \geq 0, \sum_{j=1}^J |w_j| = 1.$$

That is, pooled stacking collects the cross-validated predicted values that are calculated in each step k of the cross-fitting process for each learner j and estimates the stacking weights based on the pooled data set. We note that the computational costs are approximately the same as for DDML with conventional stacking.

4 The Practical Benefits of DDML with Stacking: Two Simulation Studies

In this section, we discuss two simulation studies illustrating the advantages of pairing DDML with stacking over alternative approaches based on single pre-selected learners. We begin with a simulation calibrated to household data on wealth and 401k eligibility from the 1991 wave of the Survey of Income and Program Participation (SIPP) in Section 4.1. In Section 4.2, we revisit the simulation of Wüthrich and Zhu (2021) to assess the robustness of DDML with stacking approaches in very small samples.

4.1 Simulation calibrated to the SIPP 1991 household data

To assess the performance of DDML with conventional stacking, short-stacking and pooled stacking in a realistic setting, we consider the analysis of 401(k) eligibility and total financial assets in Poterba, Venti, and Wise (1995) as the basis for an empirically calibrated Monte Carlo simulation. The application has recently been revisited by Belloni et al. (2017), Chernozhukov et al. (2018), and Wüthrich and Zhu (2021) to approximate high-dimensional confounding factors using machine learning. We focus on estimating the partially linear model discussed in the previous section. The outcome is measured as net financial assets, the treatment variable is an indicator for eligibility to the 401(k)

Let $\{(y_i, d_i, x_i)\}_{i=1, \dots, n}$ denote the observed sample, where i is a household in the 1991 SIPP and y_i , d_i , and x_i respectively denote net financial assets, an indicator for 401(k) eligibility, and the vector of control variables.

1. Using the full sample, obtain the slope coefficient $\hat{\theta}_{OLS} \approx 5\,896$ from linear regression of d_i against d_i , and x_i in the original data. Construct the partial residuals $y_i^{(r)} = y_i - \hat{\theta}_{OLS}d_i$, $\forall i$.
2. Fit a supervised learning estimator (either linear regression or gradient boosting) to predict $y_i^{(r)}$ with the controls x_i . Denote the fitted estimator by \tilde{g} . Similarly, fit a supervised learning estimator to predict d_i with x_i and denote the fitted estimator by \tilde{h} .
3. Repeat to generate simulated samples of size n_b :
 - (a) Sample from the empirical distribution of x_i by bootstrapping n_b observations from the original data. Denote the bootstrapped sample by \mathcal{D}_b .
 - (b) Draw $\nu_i \stackrel{iid}{\sim} \mathcal{N}(0, \kappa_1)$ and $\varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \kappa_2)$, where κ_1 and κ_2 are simulation hyperparameters. Define

$$\begin{aligned}\tilde{d}_i^{(b)} &= \mathbb{1}\{\tilde{h}(x_i) + \nu_i \geq 0.5\} \\ \tilde{y}_i^{(b)} &= \theta_0 \tilde{d}_i^{(b)} + \tilde{g}(x_i) + \varepsilon_i \quad \forall i \in \mathcal{D}_b\end{aligned}$$

where we set $\theta_0 = 6\,000$ to roughly resemble the magnitude of the regression coefficient of 401(k) eligibility in the full data.

Notes: We set the hyper-parameter κ_1 and κ_2 to approximately match variance of 401(k) eligibility and log-wealth in the data. The values of the simulation hyperparameters (κ_1, κ_2) differ slightly depending on the supervised learning estimator used to fit the reduced form equations in the data. We take $\kappa_1 = 0.35$ in both scenarios but take $\kappa_2 = 55\,500$ when using linear regression and $\kappa_2 = 54\,000$ when using gradient boosting. Differences arise because gradient boosting reduces residual variance in the true data.

Algorithm 1: Algorithm for the calibrated Monte Carlo simulation

pension scheme, and the set of controls includes age, income, education in years, family size, as well as indicators for two-earner status, home ownership, and participation in two alternative pension schemes.

The simulation involves three steps. In the calibration step, we fit two generative models to the $n = 9\,915$ households from the 1991 wave of the Survey of Income and Program Participation. The first generative model is fully linear while the second is partially linear, allowing controls to enter non-linearly through gradient-boosted trees fitted to the real data. This approach is aimed at extracting and magnifying the linear or non-linear structures in the empirical conditional distributions, respectively, enabling us to compare the performance of estimators across favorable and unfavorable structures of the data. The generative step then simulates datasets of size $n_b = \{9\,915, 99\,150\}$ from the

respective fully linear model and the partially linear model. Throughout, we set the effect of 401(k) eligibility on total financial wealth to $\theta_0 = 6\,000$. Finally, in the estimation step, we fit various estimators to bootstrapped samples of the generated datasets and assess their statistical properties. We outline the steps used for constructing the two generative models in more detail in Algorithm 1.

For each bootstrap sample, we calculate estimates of the effect of 401(k) eligibility on simulated net financial assets. The estimators we consider are linear regression, the post-double selection (PDS) lasso estimator proposed by Belloni, Chernozhukov, and Hansen (2014), as well as DDML estimators with and without stacking. The candidate learners of the DDML estimators are linear regression, cross-validated lasso and ridge regression with interactions and second-order polynomial expansions of the controls, cross-validated lasso and ridge with no interactions but 10th-order polynomial expansions of the controls, two versions of random forests, two versions of gradient-boosted trees, and feed-forward neural nets with three hidden layers of size five (see Table 1 notes for details). We estimate DDML paired with conventional stacking, short-stacking and pooled stacking, and consider different methods to construct the final conditional expectation function estimator: CLS, unconstrained linear regression (OLS), selecting the single best estimator, and an unweighted average.

Table 1 presents the mean bias, median absolute bias (MAB) and coverage rates of a 95% confidence interval associated with estimates of the effect of 401(k) eligibility on net financial assets. The left and right panels correspond to results based on data simulated from the linear (Panel A) and non-linear (Panel B) generative models, respectively. The CLS weights associated with each candidate learner are shown in Table 2.¹⁰

Given the construction of the generative models, we would expect that linear regression performs best in the fully linear setting and that DDML with gradient boosting performs best in the nonlinear setting where the nuisance function is generated by gradient boosting. The simulation results confirm this intuition, showing that the two procedures achieve

¹⁰Further results are provided in the Appendix. Tables A.3 and A.2 show the stacking weights when using single-best and OLS as the final learner. Table A.1 in the Appendix gives the mean-squared prediction errors (MSPE) for each candidate learner for comparison.

Table 1: Bias and Coverage Rates in the Linear and Non-Linear DGP

	Panel (A): Linear DGP				Panel (B): Non-linear DGP							
	$n_b = 9,915$				$n_b = 9,915$							
	Bias	MAB	Rate		Bias	MAB	Rate					
$n_b = 99,150$												
	Bias	MAB	Rate		Bias	MAB	Rate					
Full sample:												
OLS	-24.8	820.2	0.95	-1.9	269.1	0.95	-2587.3	2642.4	0.95	-2644.9	2640.5	0.
PDS-Lasso	-25.4	821.2	0.95	0.6	269.4	0.95	-2598.4	2661.3	0.95	-2644.1	2638.4	0.
DDML methods:												
Candidate learners												
OLS	-30.5	826.8	0.95	-2.3	270.5	0.95	-2617.9	2622.6	0.95	-2647.2	2645.8	0.
Lasso with CV (2nd order poly)	-28.5	830.2	0.96	-1.5	272.6	0.95	746.2	1105.5	0.96	703.5	714.6	0.61
Ridge with CV (2nd order poly)	-25.5	821.1	0.95	-1.9	275.9	0.95	801.9	1143.6	0.95	714.3	725.3	0.60
Lasso with CV (10th order poly)	187.4	1047.6	0.95	61.4	281.7	0.94	-4684.7	2111.1	0.95	-2.1	284.5	0.94
Ridge with CV (10th order poly)	1069.3	1221.0	0.94	38.1	273.4	0.94	-3070.8	2499.6	0.94	-1.9	287.5	0.95
Random forest (low regularization)	-196.5	982.2	0.91	-33.3	356.7	0.87	-64.0	1065.0	0.91	-43.4	331.8	0.87
Random forest (high regularization)	-28.1	853.1	0.95	-22.5	288.7	0.94	-133.0	932.3	0.95	-18.5	272.8	0.94
Gradient boosting (low regularization)	-82.2	825.0	0.95	-19.4	270.9	0.95	52.4	924.3	0.95	13.6	267.7	0.95
Gradient boosting (high regularization)	28.6	819.6	0.96	71.4	279.8	0.94	199.8	895.4	0.96	182.3	319.3	0.93
Neural net	309.2	866.0	0.94	17.4	288.1	0.94	-620.3	1103.6	0.94	-142.9	292.0	0.92
Stacking approaches												
Stacking: CLS	8.8	829.3	0.95	-1.5	274.4	0.95	226.8	1129.3	0.95	24.9	262.9	0.95
Stacking: Average	-0.1	812.9	0.94	1.4	273.7	0.95	-101.7	1102.7	0.94	60.0	271.9	0.95
Stacking: OLS	-33.7	874.1	0.94	2.6	272.0	0.95	871.4	1264.7	0.94	26.6	265.7	0.94
Stacking: Single-best	-28.6	823.3	0.96	-3.5	272.1	0.95	179.2	1016.2	0.96	15.0	266.4	0.95
Short-stacking: CLS	-23.8	817.6	0.96	-1.5	274.3	0.95	221.9	897.3	0.96	21.9	261.7	0.95
Short-stacking: Average	-0.1	812.9	0.94	1.4	273.7	0.95	-101.7	1102.7	0.94	60.0	271.9	0.95
Short-stacking: OLS	-27.9	818.1	0.96	-2.4	274.4	0.95	169.0	888.2	0.96	15.8	262.5	0.94
Short-stacking: Single-best	-20.7	826.0	0.96	-3.4	272.4	0.95	116.1	900.1	0.96	15.0	266.4	0.95
Pooled stacking: CLS	-24.8	817.8	0.96	-1.5	273.4	0.95	251.5	960.1	0.96	24.8	264.8	0.95
Pooled stacking: Average	-0.1	812.9	0.94	1.4	273.7	0.95	-101.7	1102.7	0.94	60.0	271.9	0.95
Pooled stacking: OLS	-48.1	834.6	0.95	-2.2	272.1	0.95	376.8	1068.3	0.95	16.7	266.5	0.94
Pooled stacking: Single-best	-23.0	829.7	0.96	-3.3	270.9	0.95	131.0	952.0	0.96	15.0	266.4	0.95

Notes: The table reports mean bias, median absolute bias (MAB) and coverage rate of a 95% confidence interval for the listed estimators. We consider DDML with $K = 2$ cross-fit folds and the following individual learners: OLS with elementary covariates, CV lasso and CV ridge with second-order polynomials and interactions, CV lasso and CV ridge with 10th-order polynomials but no interactions, random forest with low regularization (8 predictors considered at each leaf split, no limit on the number of observations per node, bootstrap sample size of 70%), highly regularized random forest (5 predictors considered at each leaf split, at least 10 observation per node, bootstrap sample size of 70%), gradient-boosted trees with low regularization (500 trees and a learning rate of 0.01), gradient-boosted trees with high regularization (250 trees and a learning rate of 0.01), feed-forward neural nets with three hidden layers of size five. For reference, we report two estimators using the full sample: OLS and PDS lasso. Finally, we report results for DDML paired with conventional stacking, short-stacking and pooled stacking where the final estimator is either CLS, OLS, the unweighted average of candidate learners or the single-best candidate learner. Results are based on 1,000 replications.

among the lowest bias and median absolute bias in the data-generating processes that are based on them. Researchers are rarely certain of the functional structure in economic applications, however, so that it is more interesting to consider their respective performance in the non-favorable setting. In the non-linear data-generating process, linear regression is among the estimators with the worst performance across all three measures. Similarly, gradient boosting-based DDML is non-optimal in the linear data-generating process. It is outperformed by linear regression and CV lasso, both of which enforce a linear functional form on the control variables, in terms of MAB.

The simulation results are consequences of the “no free lunch” theorem in machine learning (Wolpert, 1996). Informally, the theorem states that there exists no estimator that performs best across all empirical settings. Researchers must, therefore, carefully match estimators to their application. However, with limited knowledge about underlying data-generating processes and few functional form restrictions implied by economic theory, the number of plausibly suitable estimators is typically large.

The bottom section of Table 1 reports results for DDML combined with the three stacking approaches outlined in Section 3. For each stacking approach, we consider stacking weights estimated by (CLS) as outlined in Section 3, set equal to $1/J$ (Average), estimated without constraint by OLS (OLS), and by selecting only the single best candidate learner (Single-best). We find that short-stacking performs similarly to, and sometimes better than, conventional and pooled stacking, while being computationally much cheaper (as shown in Table A.4). For example, at $K = 10$ and $V = 5$, DDML combined with short-stacking ran around 4.4 times faster on the full sample than DDML with conventional or pooled stacking, which is roughly in line with a speed improvement by a factor of $1/V$.¹¹

Even though the simulation set-up should favor single-best as the final learner, since there is one ‘true’ candidate learner, single-best does not clearly outperform CLS. The bias of the OLS final learner is overall similar to CLS, except when employing conventional

¹¹The computations were performed on the high-performance cluster of the ETH Zurich. Each instance used a single core of an AMD EPYC processor with 2.25-2.6GHz (nominal)/3.3-3.5 GHz (peak) and 4GB RAM. The run time of DDML with conventional stacking was 2393s on the full sample, while short-stacking ran in only 540s.

Table 2: Average stacking weights with CLS

	Stacking		Pooled stacking		Short-stacking	
	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$
<i>Panel (A): Linear DGP</i>						
OLS	0.768	0.329	0.820	0.349	0.768	0.287
Lasso with CV (2nd order poly)	0.069	0.021	0.057	0.011	0.069	0.015
Ridge with CV (2nd order poly)	0.072	0.041	0.057	0.028	0.093	0.051
Lasso with CV (10th order poly)	0.016	0.206	0.013	0.221	0.012	0.198
Ridge with CV (10th order poly)	0.023	0.112	0.017	0.098	0.020	0.128
Random forest (low regularization)	0.003	0.003	0.002	0.002	0.002	0.002
Random forest (high regularization)	0.007	0.009	0.005	0.006	0.005	0.006
Gradient boosting (low regularization)	0.018	0.162	0.014	0.178	0.013	0.190
Gradient boosting (high regularization)	0.004	0.007	0.003	0.004	0.002	0.002
Neural net	0.020	0.110	0.011	0.104	0.015	0.122
<i>Panel (B): Non-Linear DGP</i>						
OLS	0.	0.	0.	0.	0.	0.
Lasso with CV (2nd order poly)	0.	0.	0.	0.	0.	0.
Ridge with CV (2nd order poly)	0.	0.034	0.	0.035	0.	0.026
Lasso with CV (10th order poly)	0.	0.001	0.	0.001	0.	0.
Ridge with CV (10th order poly)	0.	0.038	0.	0.037	0.	0.027
Random forest (low regularization)	0.153	0.003	0.154	0.001	0.180	0.001
Random forest (high regularization)	0.	0.061	0.	0.065	0.	0.070
Gradient boosting (low regularization)	0.845	0.852	0.846	0.856	0.820	0.870
Gradient boosting (high regularization)	0.	0.	0.	0.	0.	0.
Neural net	0.002	0.011	0.	0.006	0.	0.006

Notes: The table shows the average stacking weights associated with the candidate learner for DDML with conventional stacking, pooled stacking and short-stacking. The final learner is CLS. The bootstrap sample size is $n_b = 9,915$ and the number of cross-fitting folds is $K = 2$. Results are based on 1,000 replications. See Table 1 for more information.

stacking under the non-linear DGP for $n_b = 9915$ where the average bias is almost four times as large. The unweighted average appears sub-optimal for $n_b = 99150$ under the non-linear DGP, and its performance likely deteriorates further if many poorly chosen candidate learners are included.

The CLS weights in Table 2 indicate that stacking approaches successfully assign the highest weights to the estimators aligning with the data-generating process (i.e., either OLS or gradient boosting) among the ten included candidate learners, illustrating the ability to adapt to different data structures. Specifically, the stacking methods applied to the linear data-generating process assign the largest weight to linear models while they assign the largest weights to the gradient-boosting estimators and the lowest weights to estimators that impose a linear functional form on the control variables in the non-linear data-generating process.¹² We conclude that DDML paired with stacking approaches reduces the burden of choice researchers face when selecting between candidate learners

¹²The rates at which each candidate learner is selected by the single-best final learner are shown in Table A.3 in the appendix and provide similar insights.

and specifications by allowing for the simultaneous consideration of multiple options, thus implying attractive robustness properties across a variety of data-generating processes.

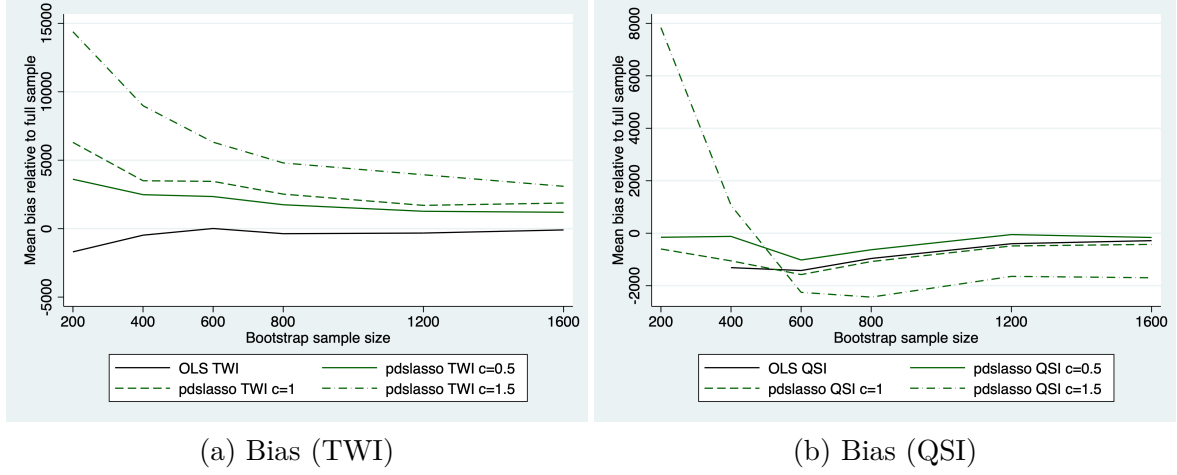
4.2 DDML and Stacking in Very Small Samples

A possible concern for estimators relying on machine learning is that they might not perform well for very small samples, given that their flexibility comes at the cost of increased variance compared to parametric estimators. Wüthrich and Zhu (2021, henceforth WZ) use two simulations to demonstrate that PDS lasso tends to underselect controls, which may result in a substantial small-sample bias. They also show that the bias heavily depends on the exact lasso penalty chosen (i.e., whether the plugin penalty of Belloni, Chernozhukov, and Hansen 2014 is scaled by 0.5 or 1.5), and argue in favor of OLS with appropriately chosen standard errors over PDS lasso in high-dimensional settings.

We revisit the 401(k) simulation set-up in WZ to assess if DDML with stacking suffers from similar issues in small samples and to compare the performance of DDML paired with stacking with PDS lasso and OLS. Following WZ, we run simulations on bootstrap samples of the data for $n_b = \{200, 400, 800, 1600\}$ and approximate the bias as the mean difference relative to the full-sample estimates ($n = 9915$).¹³ WZ consider two sets of controls: two-way interactions (TWI), and quadratic splines with interactions (QSI) (as in Belloni et al., 2017). The number of predictors is 167 and 272, respectively. Figure 4 replicates the main results of WZ (Figure 8 in their paper). Panels (a) and (b) show the bias relative to the full sample estimate for the TWI and QSI specification based on OLS and PDS lasso with tuning parameter equal to the plugin penalty of Belloni, Chernozhukov, and Hansen 2014 scaled by c for $c \in \{0.5, 1, 1.5\}$. It is noteworthy that the speed at which the bootstrapped estimates converge to the full-sample estimate depends on the set of controls for the PDS lasso, but less so for OLS. While PDS lasso with $c = \{0.5, 1\}$ and OLS perform similarly if QSI controls are used, PDS lasso converges much more slowly to the full-sample estimate with TWI controls.

¹³The full-sample estimates are reported in Table B.1.

Figure 4: Replication of Figure 8 in Wüthrich and Zhu (2021).



Notes: The figures report the mean bias calculated as the mean difference to the full sample estimates. Full sample estimates reported in Table B.1. Following WZ, we draw 1000 bootstrap samples of size $n_b = \{200, 400, 600, 800, 1200, 1600\}$. ‘TWI’ indicates that the predictors have been expanded by two-way interactions. ‘QSI’ refers to the quadratic spline & interactions specification of Belloni et al. (2017).

The DDML-stacking framework allows us to choose between, and combine, OLS and lasso with both the TWI and QSI set of controls. Another advantage of DDML over PDS lasso is that we can leverage lasso with cross-validated penalization for a fully data-driven penalization approach. Table 3 compares the performance of the full-sample estimators OLS and PDS lasso (shown in Panel A) to DDML-stacking estimators only relying on OLS and CV lasso with TWI and QSI controls as candidate learners (Panel B). We again consider conventional stacking, short-stacking and pooled stacking together with either CLS or single-best as the final learner. We set the number of cross-fitting folds to $K = 10$ (but also consider $K = 2$ below for comparison in Panel E).

Across all sample sizes, the DDML-stacking estimators strictly outperform both OLS specifications, as well as PDS lasso with TWI, and exhibit overall similar performance to PDS lasso utilizing OLS controls and $c = \{0.5, 1\}$. The differences across DDML-stacking estimators are relatively minor. The CLS short-stacking weights reported in Table 4, Panel A-B, reveal that CV-lasso with QSI controls receives the largest weights, while both OLS specifications contribute jointly between nearly zero (at $n_b = 200$) and only up to 15% (for $n_b = 1600$ and the estimation of $E[D|X]$). When selecting only a single candidate learner, CV-lasso with QSI is chosen in more than 90% of bootstrap iterations for the estimation of $E[Y|X]$ and $E[D|X]$ (Panel C-D in Table 4), suggesting that CV-lasso with QSI controls

Table 3: Mean bias relative to full-sample estimates

	Bootstrap sample size n_b					
	200	400	600	800	1200	1600
<i>Panel A. Full-sample estimators</i>						
OLS QSI	-2083.5	-910.2	-806.4	-809.9	-677.2	-626.5
OLS TWI	-1694.5	-475.4	13.2	-366	-320.3	-91.3
Post double Lasso QSI $c=0.5$	409.2	-308.9	-204	-503.1	-571.6	-354.1
Post double Lasso QSI $c=1$	-179.1	-1113.5	-639.4	-1063.2	-1000.5	-523.5
Post double Lasso QSI $c=1.5$	8021.3	739.9	-1526.2	-2434.4	-2255.4	-1863.5
Post double Lasso TWI $c=0.5$	3611.2	2484.4	2347.2	1748.3	1270.4	1197.5
Post double Lasso TWI $c=1$	6303.3	3501.1	3453.1	2523.9	1702.4	1871.8
Post double Lasso TWI $c=1.5$	14386.1	8981.9	6317.9	4802.2	3939	3094.5
<i>Panel B. DDML-stacking with only OLS and CV lasso ($K = 10$)</i>						
Short-stacking: CLS	1020	-113.8	-181.1	-538.2	-575.6	-292.4
Short-stacking: Single-best	1002.3	-122.2	-270.1	-499.7	-550.3	-197.7
Pooled stacking: CLS	925.7	-237.3	-319.1	-628.1	-711	-370.5
Pooled stacking: Single-best	782.3	-200.5	-358.9	-541.2	-580.2	-237.5
Stacking: CLS	1155.8	-254.7	-266.9	-645	-633	-315.1
Stacking: Single-best	999.5	-23.6	-184.9	-503.9	-571.1	-248.2
<i>Panel C. DDML-stacking will all candidate learners ($K = 10$)</i>						
Short-stacking: CLS	1355.1	342.2	403.3	34.2	-103.9	43.8
Short-stacking: Single-best	669.2	113.5	144.6	-182.3	-272.6	48.9
Pooled stacking: CLS	2849.3	1345.7	1197	383.8	-102.3	-10.6
Pooled stacking: Single-best	724.1	-69.4	45	-250.7	-309	-19.4
Stacking: CLS	1394.1	296.9	344.5	2.8	-168.5	56.9
Stacking: Single-best	718.4	-47	104.3	-141.5	-318.6	42.5
<i>Panel D. DDML with candidate learners ($K = 10$)</i>						
OLS	963	-150.8	210	-161.7	-235.5	31.8
Lasso with CV (TWI)	5948.6	3223.1	2589.1	1706.2	872.2	734.1
Ridge with CV (TWI)	4137.3	1853.8	1617.5	951.8	657.5	879.2
Lasso with CV (QSI)	297.5	-343.9	-311.9	-551.8	-597.1	-239.8
Ridge with CV (QSI)	426.1	-111	85.3	-240.8	-294.4	-7.8
Random forest (low regularization)	1852.8	618.3	709.6	259.7	7.7	95.5
Random forest (high regularization)	9987.4	4270.1	2940.2	1919.5	1037.8	925
Gradient boosting (low regularization)	772.3	-25	306.3	70.7	-127.2	113
Gradient boosting (high regularization)	1060.8	94.3	564.6	292.5	44.2	228.6
Neural net	8892.3	7481.2	6915.4	5653.2	3716.5	2224.2
<i>Panel E. DDML-stacking will all candidate learners ($K = 2$)</i>						
Short-stacking: CLS	1842.3	1078.3	-144.4	61.2	446.7	282.9
Short-stacking: Single-best	1303.5	582.3	-436.4	-248.8	194	111.1
Pooled stacking: CLS	2799	1471.3	159.5	209.8	572.7	508.8
Pooled stacking: Single-best	1791.9	622.9	-542.3	-296.3	144.7	84.8
Stacking: CLS	1924.6	1196.1	-191.2	59.4	390.3	310.9
Stacking: Single-best	1173.4	549.6	-604.2	-285	181.8	138.3

Notes: The table reports the mean bias calculated as the mean difference to the full sample estimates. Following WZ, we draw 1000 bootstrap samples of size n_b . In Panel A, we show results for the full-sample estimators OLS and PDS lasso using either two-way interactions as controls (denoted TWI) or the quadratic spline & interactions specification of Belloni et al. (2017, denoted as QSI). We scale the PDS lasso penalty by $c = 0.5, 1$ or 1.5 . In Panel B, we report results for DDML with stacking approaches and only relying on OLS and CV lasso. In Panel C, we consider a larger set of candidate learners. These are: OLS, CV lasso and CV ridge with either TWI or QSI controls, random forest with low regularization (8 predictors considered at each leaf split, no limit on the number of observations per node, bootstrap sample size of 70%) or high regularization (5 splitting predictors, at least 10 observation per node, bootstrap sample size of 70%), gradient-boosted tree with either low (500 trees, learnings rate of 0.01) or high (250 trees, learning rate of 0.01) regularization, and a neural net with three hidden layers of size 5. Panel D shows results for these individual candidate learners. In Panels B–D, we use $K = 10$ cross-fitting folds and $R = 5$ cross-fitting repetitions. Panel D uses the same specifications as Panel C, but uses $K = 2$.

is strictly preferable over OLS and lasso with TWI controls in this application. This simulation exercise again highlights that relying on poorly chosen specifications that are not validated against other choices might be sub-optimal. In practice, the researcher does not know whether TWI or QSI controls perform better and whether to use OLS or lasso. Crucially, DDML paired with stacking allows for simultaneous consideration of OLS and lasso with both TWI and QSI controls and thus resolves the choice between learners and control specifications in a data-driven manner.

Table 4: Short-stacking weights

<i>Estimator</i>	<i>Observations</i>						
	200	400	600	800	1200	1600	9915
<i>Panel A. Constrained least squares. $E[Y X]$, $K = 10$</i>							
OLS (TWI)	.01	.042	.062	.078	.098	.113	.013
OLS (QSI)	0	0	.002	.008	.023	.032	.128
Lasso with CV (TWI)	.249	.2	.196	.171	.158	.14	.214
Lasso with CV (QSI)	.74	.758	.74	.742	.721	.716	.645
<i>Panel B. Constrained least squares. $E[D X]$, $K = 10$</i>							
OLS (TWI)	.005	.037	.055	.074	.1	.127	.13
OLS (QSI)	0	0	.001	.003	.011	.022	.134
Lasso with CV (TWI)	.264	.163	.137	.119	.114	.111	.232
Lasso with CV (QSI)	.731	.8	.807	.803	.775	.74	.504
<i>Panel C. Single-best. $E[Y X]$, $K = 10$</i>							
OLS (TWI)	0	0	0	0	.001	0	0
OLS (QSI)	0	0	0	0	0	0	0
Lasso with CV (TWI)	.186	.141	.128	.112	.09	.081	0
Lasso with CV (QSI)	.814	.859	.872	.888	.909	.919	1
<i>Panel D. Single-best. $E[D X]$, $K = 10$</i>							
OLS (TWI)	0	0	0	0	0	0	0
OLS (QSI)	0	0	0	0	0	0	0
Lasso with CV (TWI)	.239	.126	.098	.079	.06	.068	.003
Lasso with CV (QSI)	.761	.874	.902	.921	.94	.932	.997

Notes: The table reports the stacking weights corresponding to the DDML short-stacking estimators in Figure 3. Panel A-B use constrained least squares. Panel C-D rely on the single-best final learner. Panel A and C refer to the estimation of $E[Y|X]$; Panel C and D to the estimation of $E[D|X]$. See notes below Table 3 for more information.

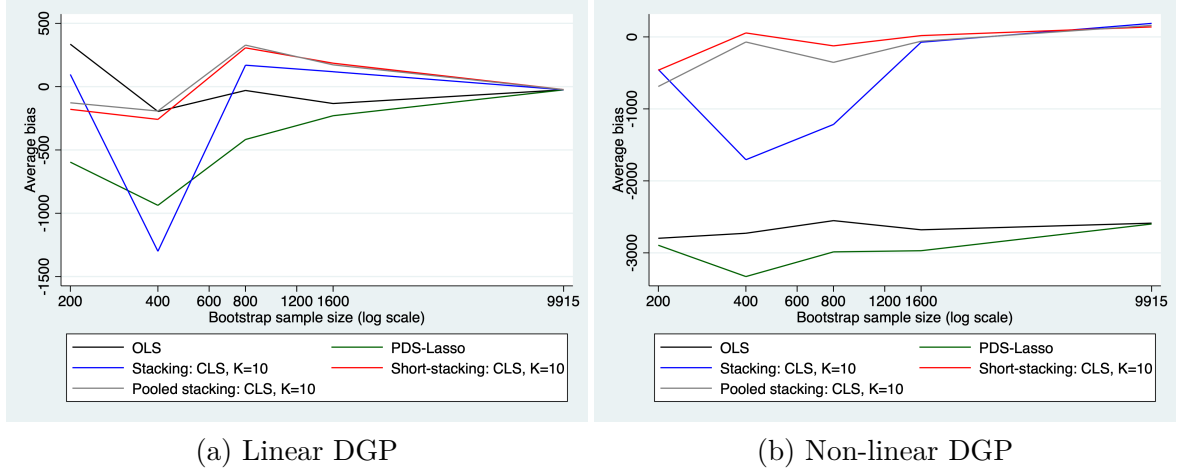
In the next step, we expand the set of candidate learners by two types of random forests, two types of gradient-boosted trees and a feed-forward neural net. In principle, widening the set of candidate learners increases robustness to a larger class of unknown confounding structures. We show the results in Panel C, Table 3. When measuring performance based on the difference to the full-sample estimates, we find there are benefits of extending the set of candidate learners for bootstrap sample sizes of $n_b = 800$ or larger. The results are generally comparable across conventional, short and pooled stacking. However, single-best exhibits a lower bias for small bootstrap sample sizes vs. CLS,

while pooled stacking with CLS appears to perform worse. The CLS weights reported in Appendix Table B.2 illustrate how DDML-stacking estimators adapt to the sample size. For example, for smaller sample sizes, a larger weight is put on OLS in the estimation of $E[Y|X]$. In Panel D, we report results for each candidate learner individually. DDML-stacking approaches perform better than most individual candidate learners and similar to the best-performing individual learner, which is DDML with CV lasso and QSI controls. In Panel E, we also show results if we reduce the number of folds to $K = 2$. The performance deteriorates drastically for smaller sample sizes, indicating that—while DDML stacking appears competitive for small sample sizes—it is important to increase the number of folds to ensure larger training samples for the CEF estimators.

A drawback of measuring the bias as the difference to the full-sample estimate is that we do not gain insights about convergence to the true parameter. We thus revisit the calibrated simulation exercise from Section 4.1, which allows us to measure the bias as the difference to the true parameter. When the DGP is linear (see Figure 5a), DDML with short-stacking or pooled stacking perform overall similarly to OLS. DDML with conventional stacking exhibits relatively large bias with $n_b = 400$. If the true DGP is non-linear, see Figure 5b, OLS and PDS-Lasso are unable to recover the true effect, while DDML with short and pooled stacking yield reasonably close approximations of the true parameter even for small sample sizes. DDML with conventional stacking is competitive only for larger samples. We provide extensive results for mean bias and coverage rates in Tables B.3 and B.4 in the Appendix.

To conclude, the results highlight the risks of relying on inappropriate functional form assumptions. DDML paired with stacking approaches—when combined with a diverse set of candidate learners—imposes weaker conditions on the underlying data-generating process compared to relying on a single pre-selected learner. Short-stacking and pooled stacking outperform conventional stacking in small samples. We conjecture the improvement is due to short and pooled stacking imposing common weights across cross-fitting folds.

Figure 5: Mean bias for very small sample sizes



Notes: The figure shows results from the calibrated simulation in Table 1, but with smaller bootstrap sample sizes. See table notes in Table 1 for more information. Full results for bias and coverage in small samples can be found in Table B.3 and B.4.

5 Applications

In this section, we use two applications to illustrate how pairing DDML and stacking can increase the robustness of structural parameter estimates to the underlying structure of the data. In the first application, we estimate gaps in citations of articles in top economics journals across different gender compositions among the authors. We condition on the abstract to proxy for the content and quality of the paper and demonstrate that stacking-based DDML is a practical solution to challenging estimation problems using text data. In the second application, we revisit the UK sample of the OECD Skills Survey for Kitagawa-Oaxaca-Binder estimates of the unexplained gender wage gap where we condition on a large set of individual characteristics. Both applications pertain to the literature on gender gaps in various domains, e.g., entry to STEM programs (Card and Payne, 2021), ICT literacy (Siddiq and Scherer, 2019) or wages (Strittmatter and Wunsch, 2021; Bonaccolto-Töpfer and Briel, 2022), and are methodologically also closely related to the broader literature on discriminatory attitudes towards minority groups (e.g., Hangartner, Kopp, and Siegenthaler, 2021).

5.1 Gender gap in citations

This section uses DDML with stacking to estimate a partially linear model applied to average differences in citations of articles published in top-30 economic journals from 1983 to 2020 by the gender composition of the authors. Following Card et al. (2020), we distinguish between papers with (imputed) all-male, all-female, and mixed-gender authorship.¹⁴ Instead of conditioning on hand-coded characteristics such as JEL codes, we leverage the abstract text as a proxy for the topic and quality of the article. Estimating these conditional differences is a challenging statistical problem due to the non-standard nature of text data, and researchers are faced with two key decisions when operationalizing an estimator using text data: how to encode the text data into numerical features, and how to select a suitable learner given the encoded data. Both decisions are ex-ante challenging, but also practically highly relevant as text data is becoming increasingly encountered in economic applications (e.g., Gentzkow and Shapiro, 2010; Chen and Ornaghi, 2023; Widmer, Galletta, and Ash, 2023). We show that these decisions can be consequential and that by simultaneously considering different encoding procedures and multiple learners, DDML with stacking provides a simple practical solution to both problems.

In documenting average differences in citations, the analysis presented also contributes to the broader literature on gender biases in academia (e.g., Lundberg and Stearns, 2019; Card et al., 2020; Hengel, 2022). It is well-documented that women are under-represented in academia, especially in senior positions (Ceci et al., 2014; Lundberg and Stearns, 2019). A possible reason for the persistent gap in representation include is that scholarly work produced by women faces more sceptical scrutiny compared to work produced by their male counterparts (Hengel, 2022; Krawczyk and Smyk, 2016). Higher scrutiny could be, for example, reflected at the refereeing stage when a publication decision is made and, as we examine here, after publication when scholarly work is attributed by other scholars through citations (Card et al., 2020; Roberts, Stewart, and Nielsen, 2020; Grossbard, Yilmazer, and Zhang, 2021).

¹⁴As we explain below, we impute the gender mix of authors from the authors' names.

Throughout our analysis, we focus on a descriptive characterization of the average gaps in citations across different gender compositions of the authors as given by θ_0 in the partially linear model of Equation (1) where Y denotes log-citations, D is a two-dimensional vector whose first component is an indicator for all-female authorship and whose second component is an indicator for mixed-gender authorship. The vector X collects the content of the abstract and a set of year-of-publication indicators. The two components of θ_0 may thus be interpreted as summarizing the average relative difference in total citations between all-male and all-female authorship, and all-male and mixed-gender authorship, respectively, conditional on the article’s year of publication and abstract. Throughout, we make no conditional unconfoundedness assumptions that would be necessary for causal interpretations.

We consider a sample of 29 185 articles that have been published between 1983–2020. The data was sourced from Scopus and is a sub-sample of the data analyzed in Advani et al. (2021), who kindly shared their data with us. For each article, we have a record of the citation count and the authors’ names, which we use to infer the authors’ gender.¹⁵ In the sample, 6.2% of articles are authored by only female authors and 23.5% have authors from both genders.

Before turning to estimation, the text of the abstract needs to be transformed into a numerical vector. To admit estimation conditional on the content of the abstract, it is necessary to find a representation (referred to as embedding) of the text that is lower-dimensional but captures its core meaning. An active literature in statistics and computer science provides solutions to this problem, suggesting a large variety of algorithms to construct text embeddings (see the overview in Ash and Hansen, 2023). Thus, in addition to the choice of candidate learner, researchers intent on using text data for their analysis are faced with the additional choice of embedding algorithm. To illustrate how stacking-based DDML can help support this choice, we consider two procedures for encoding the

¹⁵We use the software *Namsor* which ranks among the best-performing algorithms for gender classification using names (Sebo, 2021). Articles of authors whose gender could not be classified with a probability of less than 70% were excluded. Our sample includes 620 articles for which no citation is recorded. These were excluded from the analysis. We also provide results using the number of citations in Appendix Table C.1.

text of the abstract into numerical features: First, we consider a bag-of-word model summarizing the text as (stemmed) word counts (as used in, e.g., Enke, 2020; Esposito et al., 2023). In our data, this results in a 211-dimensional vector of word counts for each abstract. Second, since the bag-of-word approach disregards the word order and context, we construct word embeddings generated by a pre-trained BERT model, a transformer-based large-language model (Devlin et al., 2018). In particular, for each abstract, we extract the 768-dimensional vector of weights from the last hidden layer of the BERT model that was pre-trained on a large corpus of (uncased) English text data.¹⁶ Instead of embedding individual words, BERT attempts to reconstruct both whole sentences and the context of these sentences, making it particularly suitable to characterize the content of the abstracts. Recently, Bajari et al. (2023) use BERT to construct embeddings of product descriptions on Amazon.com.

The numerical abstract embeddings are then used in several base learners. We consider OLS, PDS lasso and DDML with CV lasso, CV ridge, **XGBoost** (Chen and Guestrin, 2016), random forests and a feed-forward neural net (see table notes for details).¹⁷ The base learners are aggregated by pairing DDML with either conventional stacking or short-stacking, and with either CLS or single-best.¹⁸ The final estimator thus simultaneously aggregates across both the text embedding algorithm and the base learner.

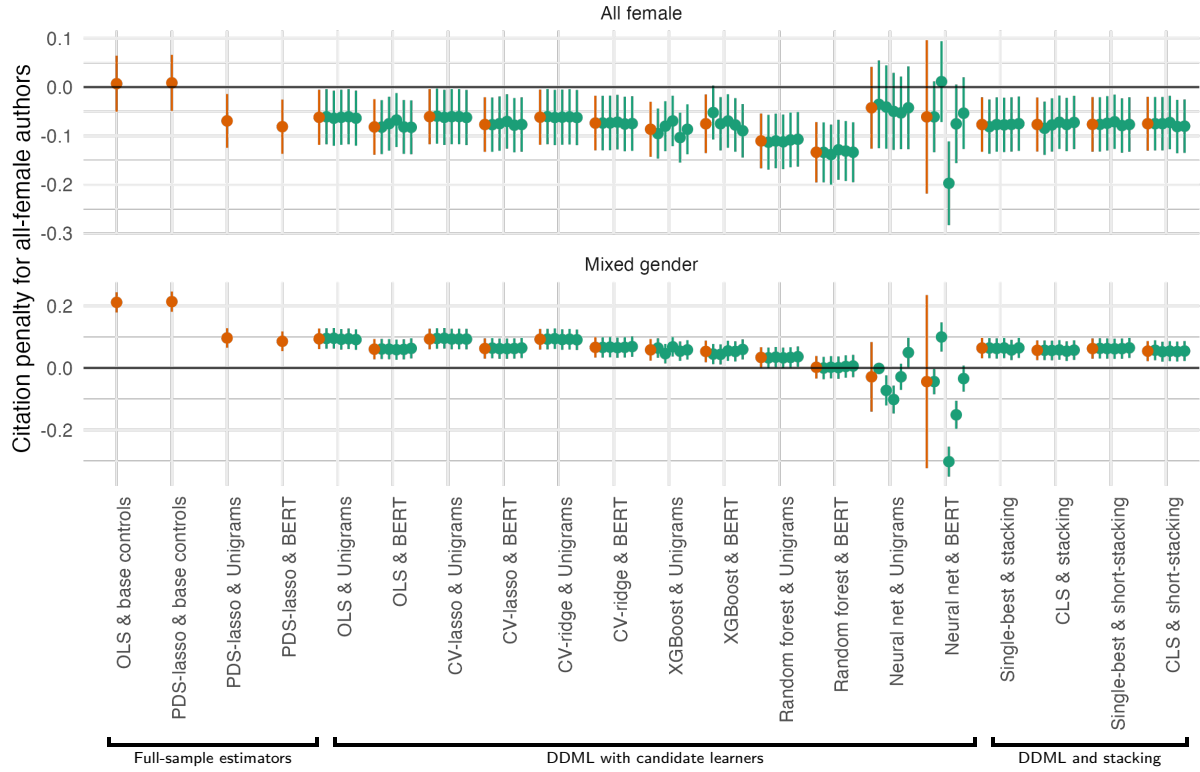
Figure 6 shows estimates of the average relative difference in total citations between all-male and all-female authorship (top-panel) and all-male and mixed-gender authorship (bottom-panel), respectively, for different control specifications and estimators. When we only condition on the publication year, the citation penalty for all-female authorship is close to zero, while there is a large positive effect of +21.2% ($s.e. = 1.7$) for mixed-gender authorship. We next employ PDS lasso to add the abstract text either in the form of word counts or as BERT features. Using the latter, the citation gap increases to -8.1% (2.8)

¹⁶The model **bert-base-uncased** is freely available from, among others, the Python library **huggingface**.

¹⁷To reduce the run time, we use regression approaches both for the estimation of $E[Y|D, X]$ and $E[D|X]$.

¹⁸We omit pooled stacking from this application since the R package **ddml**, which was used for this application, does currently not support pooled stacking.

Figure 6: The citation gap by authors' gender composition



Notes: The figure shows estimates of θ_0 summarizing average relative difference in total citations between all-male and all-female authorship, and all-male and mixed-gender authorship, respectively, conditional on the article's year of publication and abstract. Error-bars show heteroskedasticity-robust 95% confidence intervals. We consider the following estimators: OLS, PDS lasso and DDML with the following candidate learners: OLS, CV ridge, CV lasso, XGBoost (using 500 trees, learning rate of 0.3), random forest (using 500 trees) and feed-forward neural net (early stopping with 15 rounds, 0.5 dropout, 0.1 learning rate, 0.1 validation split, 50 epochs, 500 batch size and 3 hidden layers of size 10). Finally, we pair DDML with either conventional stacking or short-stacking, and with either CLS or single-best as the final learner based on the above candidate learners. Throughout, we use five cross-fitting repetitions, five cross-validation folds and five cross-fitting folds. Results from each cross-fitting replication are illustrated in green, and median aggregates across the cross-fitting replications are shown in orange. The sample includes 29 185 articles published between 1983–2020 in top-30 economics journals. A tabular version is provided in Table C.1.

for articles with all-female authorship, while the average relative difference of articles with mixed-gender authorship reduces to +8.6% (1.6). The estimates are qualitatively similar when using word counts instead of BERT features.

In the figure, we also show five cross-fitting repetitions of pairing DDML with each candidate learner (in green) and the median aggregates over these repetitions (in orange). There are considerable differences across DDML estimators, with the median estimates of the citation gap ranging between -4.2% (4.3) and -13.4 (3.2) for articles with all-female authorship and between -4.4 (14.3) and $+9.3$ (1.7) for articles with mixed-gender authorship, highlighting that different candidate learner specifications can yield vastly different effect sizes. These stark differences emphasize the need to choose and tune CEF estimators carefully. Without thoroughly validating each candidate learner, judging which results are more credible is difficult. Furthermore, it is noteworthy that some candidate learners exhibit substantial instability across cross-fitting repetitions, especially the neural net learners, which is also reflected in the large median-aggregate standard errors.

We show results from pairing DDML and stacking approaches on the right-hand side of the same figure. Relative to the DDML estimates based on the individual candidate learners, the stacking approaches yield lower variability over cross-fitting repetitions, suggesting higher stability. All four stacking-based approaches agree on an average relative difference in citations of -7.7% (2.8) for articles with all-female authorship and suggest a citation advantage of between $+5.4$ (1.6) and $+6.4$ (1.7) for articles with mixed-gender authorship.

Table 5 shows that stacking weights of conventional and short-stacking with constrained least squares as the final learner. The stacking estimators assign small weights to learners exhibiting a relatively large MSPE and large variability over cross-fitting repetitions. For example, in the CEF estimation of log citations, the neural nets have an MSPE that is over twice as large as that of other learners. Stacking assigns, as desired, zero weights to the neural nets, whereas OLS leveraging BERT as one of the best-performing learners receives the largest weights. It is noteworthy that the stacking weights often vary markedly across CEFs, highlighting that there is no reason to assume that the same

Table 5: Stacking weights in the gender citation gap application.

	<i>Citations</i>		<i>All female</i>		<i>Mixed gender</i>	
	<i>Conv.</i>	<i>Short</i>	<i>Conv.</i>	<i>Short</i>	<i>Conv.</i>	<i>Short</i>
<i>Panel A. Stacking and short-stacking weights</i>						
OLS & Unigrams	0.109	0.053	0.026	0.158	0.004	0.128
OLS & BERT	0.307	0.389	0.063	0.105	0.153	0.216
CV-lasso & Unigrams	0.	0.	0.	0.	0.	0.
CV-lasso & BERT	0.207	0.142	0.138	0.336	0.079	0.28
CV-ridge & Unigrams	0.	0.	0.	0.	0.	0.
CV-ridge & BERT	0.	0.	0.118	0.378	0.066	0.192
XGBoost & Unigrams	0.212	0.256	0.011	0.002	0.039	0.026
XGBoost & BERT	0.052	0.07	0.005	0.015	0.016	0.036
Random forest & Unigrams	0.037	0.093	0.	0.024	0.	0.139
Random forest & BERT	0.	0.	0.164	0.	0.022	0.
Neural net & Unigrams	0.	0.	0.	0.	0.081	0.
Neural net & BERT	0.	0.	0.	0.	0.124	0.
	<i>Citations</i>		<i>All female</i>		<i>Mixed gender</i>	
<i>Panel B. Mean-squared prediction error</i>						
OLS & Unigrams	1.335		0.058		0.169	
OLS & BERT	1.287		0.058		0.168	
CV-lasso & Unigrams	1.334		0.057		0.168	
CV-lasso & BERT	1.268		0.056		0.164	
CV-ridge & Unigrams	1.335		0.057		0.168	
CV-ridge & BERT	1.276		0.056		0.164	
XGBoost & Unigrams	1.428		0.074		0.198	
XGBoost & BERT	1.517		0.068		0.196	
Random forest & Unigrams	1.347		0.059		0.17	
Random forest & BERT	1.629		0.059		0.172	
Neural net & Unigrams	4.269		0.149		1.424	
Neural net & BERT	5.576		0.058		0.184	

Notes: Panel A shows stacking weights for conventional stacking (labelled ‘Conv.’) and short-stacking (labelled ‘Short’) by candidate learners and by variable. Panel B reports the mean-squared prediction error. The final learner is constrained least squares. The stacking weights are averaged over cross-fitting repetitions. Treatment variables are an indicator for all-female authors and mixed-gender authors.

learner is best suited for estimating both $E[Y|X]$ and $E[D|X]$. This insight is especially important since most estimation approaches (including OLS and PDS lasso) impose the same structure for each CEF.

The results on the citation gaps in top economic journals conditional on the content of the abstract are consistent with a citation penalty for all-female authored articles, possibly due to a higher degree of skepticism towards all-female author teams compared to all-male author teams. However, similar to Card et al. (2020) and Maddi and Gingras (2021), the estimates also suggest a conditional citation advantage of articles with mixed-gender authorship.

5.2 Gender gap in wages

The gap in wages between men and women is a central measure of economic gender equality and has been the focus of an extensive empirical literature (see, e.g., the review in Blau and Kahn, 2017). The classic approach to estimating the unexplained gender wage gap relies on a linear version of the Kitagawa-Oaxaca-Binder decomposition (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973; for an overview, see Fortin, Lemieux, and Firpo, 2011). Several recent articles by Bonaccolto-Töpfer and Briel (2022), Strittmatter and Wunsch (2021), Böheim and Stöllinger (2021) and Bach, Chernozhukov, and Spindler (2023), among others, focus instead on semi-parametric decompositions of the wage gap leveraging more flexible machine learning algorithms. Much of this literature focuses, however, on lasso-based approaches, even though there is no apparent reason to favor sparsity-based approaches over learners relying on other regularization assumptions. In contrast to the recent literature that primarily focuses on lasso-based approaches to estimate the high-dimensional nuisance functions, we consider a diverse set of candidate learners and aggregate them via stacking.

The parameter of interest in this application is the unexplained gender wage gap, which is the expected difference in wages after conditioning on observed characteristics. Formally,

$$\theta_0 \equiv E[E[Y|D = 1, X] - E[Y|D = 0, X] | D = 1],$$

where Y denotes the logarithm of wages, D is an indicator equal to one for women, and X is a vector of potentially many individual characteristics. The parameter is well-defined if $P(D = 1|X) > 0$ with probability 1.¹⁹

In the absence of functional form assumptions, estimation of θ_0 is a challenging statistical problem due to its dependence on unknown conditional expectation functions that need to be nonparametrically estimated. Analogous to the DDML estimator for the par-

¹⁹As in the previous section, we focus our analysis on a descriptive parameter of interest and do not make conditional unconfoundedness assumptions that would be necessary for causal interpretations.

tially linear model outlined in Section 2, we consider estimation of θ_0 via the split-sample analogue of the efficient score function for θ_0 – i.e.,

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n \left(\frac{D_i(Y_i - \hat{g}_{I_{k_i}^c}(0, X_i))}{\hat{p}_{I_{k_i}^c}} - \frac{\hat{m}_{I_{k_i}^c}(X_i)(1 - D_i)(Y_i - \hat{g}_{I_{k_i}^c}(0, X_i))}{\hat{p}_{I_{k_i}^c}(1 - \hat{m}_{I_{k_i}^c}(X_i))} \right),$$

where $\hat{g}_{I_k^c}$ and $\hat{m}_{I_k^c}$ are cross-fitted estimators for $g_0(D, X) \equiv E[Y|D, X]$ and $m_0(X) \equiv E[D|X]$, and $\hat{p}_{I_k^c}$ is a cross-fitted estimator of $P(D = 1)$.

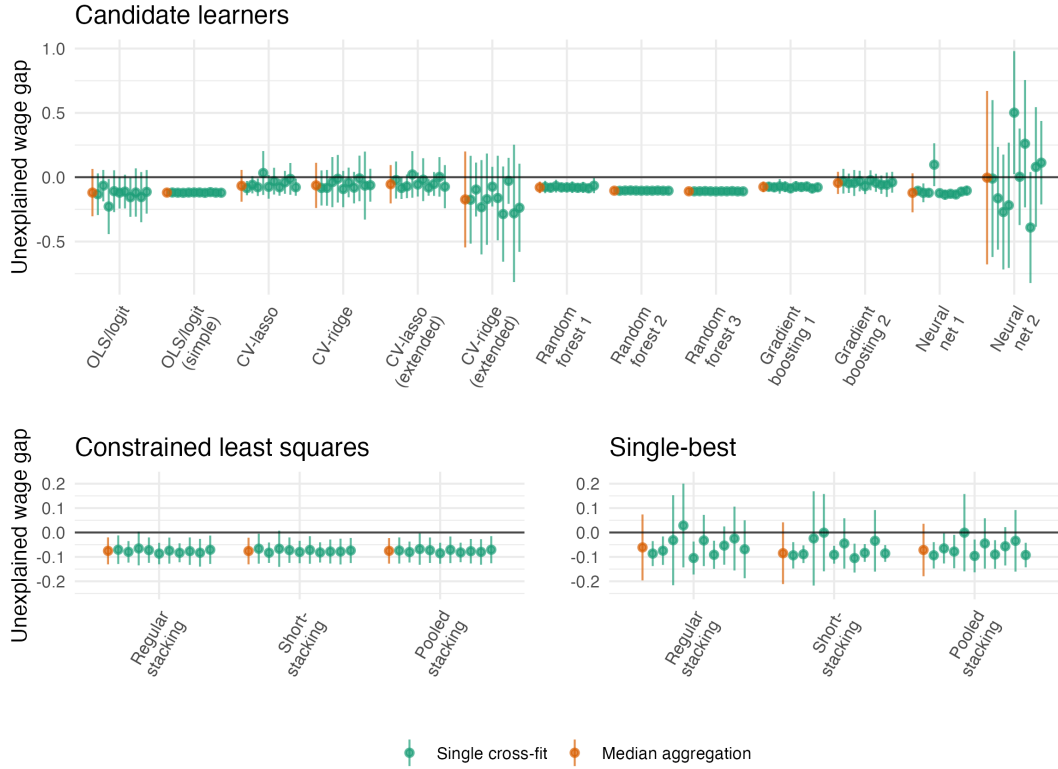
Following Forshaw et al. (2023), we take the data for this application from the UK sample of the OECD Skills Survey, which was collected in 2011-12 and comes with a rich set of covariates, including age, experience, education, occupation, and industry. The final data includes 4 836 British respondents, allowing us to test the performance of DDML with stacking on a relatively small sample. We specify three sets of control variables: The reduced set of controls only includes a selection of essential covariates: age (in levels and squared), years of education, a literacy and numeracy test score, years of tenure in the current job (in levels and squared), education level, hours worked per week, and number of children. The base set of controls adds, among others, management level, age of children, and parents' education level. Furthermore, we interact age and tenure with all categorical covariates. The expanded set comprises all variables and interacts each continuous covariate with each categorical covariate.²⁰

We include a diverse set of candidate learners to allow for a high level of flexibility. We employ regression approaches for the estimation of the CEF of log wages (i.e., $E[Y|D, X]$) and classification approaches for the CEF estimation of gender (i.e., $E[D|X]$). Our candidate learners are linear (or logistic) regression with the reduced and base set of controls; linear (or logistic) CV-lasso and CV-ridge with the base and extended set of controls; three random forests with 500 regression (or classification) trees and minimum leaf sizes of 1, 50 and 100; two types of gradient-boosted regression (or classification) trees

²⁰The base set adds the following variables to the reduced set: area of study, part of larger organization, management position, type of contract, job satisfaction, health status, living with a partner, age of youngest child, immigration age, mother's and father's highest level of education, immigration status of parents, informal job-related education in last 12 months, informal non-job-related education in last 12 months.

with and without early stopping; two feed-forward neural nets with hidden layer sizes of (40, 20, 1, 20, 50) and (30, 30, 30), and early stopping. Finally, we aggregate the candidate learners via conventional, short and pooled stacking, and using either CLS or single-best as the final learner.

Figure 7: Unexplained gender wage gap



Notes: The figure reports DDML estimates of the unexplained gender wage gap based. 95% heteroskedasticity-robust confidence intervals are shown. The candidate learners (shown on the left-hand side) are as follows: OLS (for the outcome equation) and logit (for the propensity scores) with the reduced and base set of controls; CV-lasso and CV-ridge with the base and extended set of controls; three random forests with 500 trees and minimum leaf sizes of 1, 50 and 100; two gradient-boosted trees with and without early stopping; two feed-forward neural nets with hidden layer sizes of (40, 20, 1, 20, 50) and (30, 30, 30), and early stopping. On the right-hand side, we show DDML paired with conventional, short and pooled stacking based on the above candidate learners, and with either CLS or single-best as the final learner. We use 10 cross-fitting folds and 10 cross-fitting repetitions. Results from each cross-fitting replication are illustrated in green, and median aggregates across the cross-fitting replications are shown in orange. A tabular version is provided in Table D.4-D.5.

Figure 7 reports results for individual candidate learners (on the top) and stacking approaches (on the bottom). We show results from 10 cross-fitting repetitions (in green) and the median aggregates (in orange). We again find that some candidate learners exhibit substantial variability over cross-fitting repetitions, which is also reflected in the large median-aggregate standard errors. The variability is especially large for CV-ridge with the extended set of controls and the neural nets, which are the candidate learners exhibiting the largest MSE (see Appendix Table D.1). The stacking results are, in contrast, relatively

stable over cross-fit repetitions when using CLS as the final learner and a little less stable when relying on single-best as the final learner, indicating that a combination of candidate learners seems to better fit the data than a single learner. The stacking weights and MSE in Appendix Table D.1 confirm that there is no single candidate learner dominating the others. The instability of the single-best final learner is reflected in the stacking standard errors. Given this potential for instability of choosing a single candidate learner, we recommend favoring constrained least squares over single-best if one is not confident that one of the chosen learners will be significantly better than the rest, and thus stably selected, which seems likely to be the most common setting in practice.

6 Conclusion

This article assesses the performance of DDML estimators in realistic settings using applications and simulation studies calibrated to real economic data. We highlight that estimators of structural parameters based on single pre-selected (machine) learners can be highly sensitive to the underlying structure of the data and/or poor tuning, and we show that pairing DDML with stacking can help alleviate these concerns, provided that a sufficiently diverse set of candidate learners is considered.

We discuss pairing DDML with conventional stacking but also suggest two novel stacking approaches: Short-stacking, which substantially reduces the computational burden by leveraging the cross-fitting naturally arising in the computation of DDML estimates, and pooled stacking, which decreases the variance of the stacking estimator by imposing common stacking weights over cross-fitting folds. In our simulations, both strategies are competitive with conventional stacking in settings with large and moderate sample sizes and are better in small samples. The advantages of short-stacking are particularly worth highlighting, given its substantially lower computational cost.

A key advantage of the DDML-stacking approach is that it accommodates both traditional parametric and nonparametric specifications by allowing simultaneous consideration of, for example, OLS with several sets of controls, sparsity-based learners, tree-based

ensembles and neural networks. In this sense, researchers are not forcibly deviating from standard (often linear) specifications unless the data suggests there is reason to. While machine-learning-based causal methods may yield fundamentally different results from linear regression only in specific examples, the additional robustness to unexpected structures in the data thus seems to come at relatively little cost.

References

- Advani, Arun, Elliott Ash, David Cai, and Imran Rasul (2021). *Race-related research in economics and other social sciences*. Discussion Paper 16115. CEPR.
- Ahrens, Achim, Christian B Hansen, and Mark E Schaffer (2018). *PDSLASSO: Stata module for post-selection and post-regularization OLS or IV estimation and inference*. URL: <https://ideas.repec.org/c/boc/bocode/s458459.html>.
- Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer (2022). *pystacked: Stacking generalization and machine learning in Stata*. URL: <https://arxiv.org/abs/2208.10896>.
- Ahrens, Achim, Christian B. Hansen, Mark E. Schaffer, and Thomas Wiemann (2023). *ddml: Double/debiased machine learning in Stata*. URL: <https://arxiv.org/abs/2301.09397>.
- Angrist, Joshua D. (2022). “Empirical Strategies in Economics: Illuminating the Path From Cause to Effect”. *Econometrica* 90.6, pp. 2509–2539.
- Angrist, Joshua D and Brigham Frandsen (2022). “Machine labor”. *Journal of Labor Economics* 40.S1, S97–S140.
- Angrist, Joshua D and Alan B Krueger (1999). “Empirical strategies in labor economics”. In: *Handbook of Labor Economics*. Vol. 3. Elsevier, pp. 1277–1366.
- Ash, Elliott and Stephen Hansen (2023). “Text Algorithms in Economics”. *Annual Review of Economics* 15.1, annurev-economics-082222-074352.
- Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). “Generalized random forests”. *Annals of Statistics* 47.2, pp. 1148–1178.
- Bach, Philipp, Victor Chernozhukov, and Martin Spindler (2023). “Heterogeneity in the US gender wage gap”. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnad091.
- Bajari, Patrick et al. (2023). “Hedonic prices and quality adjusted price indices powered by AI”. *arXiv preprint arXiv:2305.00044*.

- Belloni, A, V Chernozhukov, I Fernández-Val, and C Hansen (2017). “Program Evaluation and Causal Inference With High-Dimensional Data”. *Econometrica* 85.1, pp. 233–298.
- Belloni, Alexandre, D Chen, Victor Chernozhukov, and Christian Hansen (2012). “Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain”. *Econometrica* 80.6, pp. 2369–2429.
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). “Inference on treatment effects after selection among high-dimensional controls”. *Review of Economic Studies* 81, pp. 608–650.
- Bickel, Peter J, Ya’acov Ritov, and Alexandre B Tsybakov (2009). “Simultaneous analysis of Lasso and Dantzig selector”. *Annals of statistics* 37.4, pp. 1705–1732.
- Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky (2022). “When is TSLS Actually LATE?” *BFI Working Paper* 2022-16.
- Blau, Francine D. and Lawrence M. Kahn (2017). “The Gender Wage Gap: Extent, Trends, and Explanations”. *Journal of Economic Literature* 55.3, pp. 789–865.
- Blinder, Alan (1973). “Wage Discrimination: Reduced Form and Structural Estimates”. *Journal of Human Resources* 8, pp. 436–455.
- Böheim, René and Philipp Stöllinger (2021). “Decomposition of the gender wage gap using the LASSO estimator”. *Applied Economics Letters* 28.10, pp. 817–828.
- Bonaccolto-Töpfer, Marina and Stephanie Briel (2022). “The gender pay gap revisited: Does machine learning offer new insights?” *Labour Economics* 78, p. 102223.
- Breiman, Leo (1996). “Stacked regressions”. *Machine Learning* 24.1, pp. 49–64.
- Buitinck, Lars et al. (2013). “API design for machine learning software: experiences from the scikit-learn project”. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.
- Callaway, Brantly and Pedro HC Sant’Anna (2021). “Difference-in-differences with multiple time periods”. *Journal of Econometrics* 225.2, pp. 200–230.
- Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberry (2020). “Are Referees and Editors in Economics Gender Neutral?*”. *The Quarterly Journal of Economics* 135.1, pp. 269–327.

- Card, David and A. Abigail Payne (2021). “High School Choices and the Gender Gap in Stem”. *Economic Inquiry* 59.1, pp. 9–28.
- Ceci, Stephen J., Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams (2014). “Women in Academic Science: A Changing Landscape”. *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 15.3, pp. 75–141.
- Chang, Neng-Chieh (2020). “Double/debiased machine learning for difference-in-differences models”. *The Econometrics Journal* 23.2, pp. 177–191.
- Chen, Daniel L and Arianna Ornaghi (2023). “Gender Attitudes in the Judiciary: Evidence from US Circuit Courts”. *American Economic Journal: Applied Economics*.
- Chen, Tianqi and Carlos Guestrin (2016). “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). “Double/debiased machine learning for treatment and structural parameters”. *The Econometrics Journal* 21.1, pp. C1–C68.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. *arXiv preprint arXiv:1810.04805*.
- Dhar, Diva, Tarun Jain, and Seema Jayachandran (2022). “Reshaping adolescents’ gender attitudes: Evidence from a school-based experiment in India”. *American Economic Review* 112.3, pp. 899–927.
- Eberhardt, Markus, Giovanni Facchini, and Valeria Rueda (2022). “Gender Differences in Reference Letters: Evidence from the Economics Job Market”. *SSRN Electronic Journal*.
- Enke, Benjamin (2020). “Moral Values and Voting”. *Journal of Political Economy* 128.10, pp. 3679–3729.
- Esposito, Elena, Tiziano Rotesi, Alessandro Saia, and Mathias Thoenig (2023). “Reconciliation narratives: The birth of a nation after the US civil war”. *American Economic Review* 113.6, pp. 1461–1504.

- Farrell, Max H, Tengyuan Liang, and Sanjog Misra (2021). “Deep neural networks for estimation and inference”. *Econometrica* 89.1, pp. 181–213.
- Forshaw, Rachel, Vsevolod Iakovlev, Mark E. Schaffer, and Cristina Tealdi (2023). *Using machine learning methods to estimate the gender wage gap*.
- Fortin, Nicole, Thomas Lemieux, and Sergio Firpo (2011). “Chapter 1 - Decomposition Methods in Economics”. In: *Handbook of Labor Economics*. Ed. by Orley Ashenfelter and David Card. Vol. 4. Elsevier, pp. 1–102. URL: <https://www.sciencedirect.com/science/article/pii/S0169721811004072>.
- Gentzkow, Matthew and Jesse M. Shapiro (2010). “What drives media slant? Evidence from U.S. daily newspapers”. *Econometrica : journal of the Econometric Society* 78.1, pp. 35–71.
- Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri (2021). “Economic predictions with big data: The illusion of sparsity”. *Econometrica* 89.5, pp. 2409–2437.
- Gilchrist, Duncan Sheppard and Emily Glassberg Sands (2016). “Something to talk about: Social spillovers in movie consumption”. *Journal of Political Economy* 124.5, pp. 1339–1382.
- Goller, Daniel, Michael Lechner, Andreas Moczall, and Joachim Wolff (2020). “Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany’s programmes for long term unemployed”. *Labour Economics* 65, p. 101855.
- Grossbard, Shoshana, Tansel Yilmazer, and Lingrui Zhang (2021). “The gender gap in citations of articles published in two demographic economics journals”. *Review of Economics of the Household* 19.3, pp. 677–697.
- Hangartner, Dominik, Daniel Kopp, and Michael Siegenthaler (2021). “Monitoring hiring discrimination through online recruitment platforms”. *Nature* 589.7843, pp. 572–576.
- Hansen, Bruce E. and Jeffrey S. Racine (2012). “Jackknife model averaging”. *Journal of Econometrics* 167.1, pp. 38–46.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. 2nd ed. New York: Springer-Verlag.

- Hengel, Erin (2022). “Publishing While Female: are Women Held to Higher Standards? Evidence from Peer Review”. *The Economic Journal* 132.648, pp. 2951–2991.
- Kitagawa, Evelyn M. (1955). “Components of a Difference Between Two Rates”. *Journal of the American Statistical Association* 50.272, pp. 1168–1194.
- Kolesár, Michal, Ulrich K. Müller, and Sebastian T. Roelsgaard (2023). *The Fragility of Sparsity*. URL: <http://arxiv.org/abs/2311.02299>.
- Krawczyk, Michał and Magdalena Smyk (2016). “Author’s gender affects rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment”. *European Economic Review*. Social identity and discrimination 90, pp. 326–335.
- Laan, Mark J. Van der, Eric C Polley, and Alan E. Hubbard (2007). “Super Learner”. *Statistical Applications in Genetics and Molecular Biology* 6.1.
- Laan, Mark J. van der and Sandrine Dudoit (2003). “Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples”. In: vol. 130. U.C. Berkeley Division of Biostatistics Working Paper Series.
- Laan, Mark J Van der and Sherri Rose (2011). *Targeted learning: causal inference for observational and experimental data*. Vol. 4. New York: Springer.
- Lundberg, Shelly and Jenna Stearns (2019). “Women in Economics: Stalled Progress”. *Journal of Economic Perspectives* 33.1, pp. 3–22.
- Luo, Ye, Martin Spindler, and Jannis Kück (2022). “High-Dimensional L_2 Boosting: Rate of Convergence”. *arXiv preprint arXiv:1602.08927*.
- Maddi, Abdelghani and Yves Gingras (2021). “Gender diversity in research teams and citation impact in economics and management”. *Journal of Economic Surveys* 35.5, pp. 1381–1404.
- Oaxaca, Ronald (1973). “Male-Female Wage Differentials in Urban Labor Markets”. *International Economic Review* 14, pp. 693–709.
- Poterba, James M, Steven F Venti, and David A Wise (1995). “Do 401 (k) contributions crowd out other personal saving?” *Journal of Public Economics* 58.1, pp. 1–32.

- Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen (2020). “Adjusting for Confounding with Text Matching”. *American Journal of Political Science* 64.4, pp. 887–903.
- Schmidt-Hieber, Johannes (2020). “Nonparametric regression using deep neural networks with ReLU activation function”. *Annals of Statistics* 48.4, pp. 1875–1897.
- Sebo, Paul (2021). “Performance of gender detection tools: a comparative study of name-to-gender inference services”. *Journal of the Medical Library Association : JMLA* 109.3, pp. 414–421.
- Siddiq, Fazilat and Ronny Scherer (2019). “Is there a gender gap? A meta-analysis of the gender differences in students’ ICT literacy”. *Educational Research Review* 27, pp. 205–217.
- Strittmatter, Anthony and Conny Wunsch (2021). *The Gender Pay Gap Revisited with Big Data: Do Methodological Choices Matter?* URL: <https://arxiv.org/abs/2102.09207>.
- Wager, Stefan and Susan Athey (2018). “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. *Journal of the American Statistical Association* 113.523, pp. 1228–1242.
- Wager, Stefan and Guenther Walther (2016). “Adaptive concentration of regression trees, with application to random forests”. *arXiv preprint arXiv:1503.06388*.
- Widmer, Philine, Sergio Galletta, and Elliott Ash (2023). *Media Slant is Contagious*.
- Wiemann, Thomas, Achim Ahrens, Christian B Hansen, and Mark E Schaffer (2023). *ddml: Double/Debiased Machine Learning in R*.
- Wolpert, David H. (1992). “Stacked generalization”. *Neural Networks* 5.2, pp. 241–259.
- Wolpert, David H (1996). “The lack of a priori distinctions between learning algorithms”. *Neural computation* 8.7, pp. 1341–1390.
- Wüthrich, Kaspar and Ying Zhu (2021). “Omitted variable bias of Lasso-based inference methods: A finite sample analysis”. *Review of Economics and Statistics* 0.(0), pp. 1–47.

Supplementary material

A The benefits of pairing DDML and stacking

Table A.1: Mean-squared prediction error

<i>Panel (A): Linear DGP</i>	$n_b = 9,915$		$n_b = 99,150$	
	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$
<i>Candidate learners</i>				
OLS	3.095	0.200	3.089	0.200
Lasso with CV (2nd order poly)	3.097	0.200	3.089	0.200
Ridge with CV (2nd order poly)	3.101	0.200	3.089	0.200
Lasso with CV (10th order poly)	3.213	0.202	3.095	0.200
Ridge with CV (10th order poly)	3.347	0.205	3.094	0.200
Random forest (low regularization)	3.613	0.233	3.699	0.239
Random forest (high regularization)	3.183	0.205	3.197	0.207
Gradient boosting (low regularization)	3.131	0.201	3.102	0.200
Gradient boosting (high regularization)	3.152	0.202	3.138	0.200
Neural net	1.238	0.167	1.248	0.168
<i>Panel (B): Non-Linear DGP</i>				
	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$
<i>Candidate learners</i>				
OLS	3.685	0.203	3.673	0.203
Lasso with CV (2nd order poly)	3.482	0.201	3.451	0.200
Ridge with CV (2nd order poly)	3.480	0.201	3.450	0.200
Lasso with CV (10th order poly)	5.955	0.225	3.423	0.200
Ridge with CV (10th order poly)	7.062	0.235	3.425	0.200
Random forest (low regularization)	3.795	0.231	3.515	0.236
Random forest (high regularization)	3.591	0.204	3.252	0.205
Gradient boosting (low regularization)	3.353	0.200	3.095	0.198
Gradient boosting (high regularization)	3.405	0.200	3.216	0.199
Neural net	1.433	0.174	1.618	0.175

Notes: The table shows the mean-squared prediction error of each candidate learner. The bootstrap sample size is $n_b = 9\,915$ or $99\,150$. Results are based on 1 000 replications. See Table 1 for more information.

Table A.2: Average stacking weights using OLS as the final learner

<i>Panel (A): Linear DGP</i>	Stacking		Pooled stacking		Short-stacking	
	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$
OLS	1.180	0.937	1.042	0.820	0.910	0.692
Lasso with CV (2nd order poly)	-0.025	0.004	0.035	0.049	0.144	0.074
Ridge with CV (2nd order poly)	-0.145	-0.186	-0.093	-0.128	-0.062	-0.076
Lasso with CV (10th order poly)	-0.018	0.107	-0.017	0.089	-0.042	0.096
Ridge with CV (10th order poly)	-0.025	-0.021	-0.007	-0.005	0.023	0.036
Random forest (low regularization)	0.002	-0.006	0.002	-0.007	0.003	-0.007
Random forest (high regularization)	-0.011	0.003	-0.008	0.007	-0.016	0.011
Gradient boosting (low regularization)	-0.065	-0.132	-0.052	-0.120	-0.032	-0.081
Gradient boosting (high regularization)	0.101	0.292	0.097	0.292	0.079	0.249
Neural net	-248.176	0.015	0.008	0.031	0.	0.033
<i>Panel (B): Non-Linear DGP</i>						
<i>Panel (B): Non-Linear DGP</i>	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$
OLS	0.012	0.049	0.	0.040	-0.044	0.028
Lasso with CV (2nd order poly)	-0.125	-0.319	-0.136	-0.202	0.016	-0.375
Ridge with CV (2nd order poly)	0.369	0.534	0.395	0.453	0.166	0.564
Lasso with CV (10th order poly)	-0.008	0.122	0.019	0.106	0.095	0.109
Ridge with CV (10th order poly)	0.066	0.027	0.047	0.001	-0.017	0.034
Random forest (low regularization)	0.048	-0.015	0.052	-0.016	0.052	-0.017
Random forest (high regularization)	-0.097	0.071	-0.102	0.077	-0.096	0.065
Gradient boosting (low regularization)	1.114	0.024	1.167	0.054	1.345	0.164
Gradient boosting (high regularization)	-0.551	0.525	-0.618	0.506	-0.715	0.429
Neural net	-11.432	0.028	0.156	0.040	0.173	0.052

Notes: The table shows the (average) stacking weights of each candidate learner for conventional stacking, pooled stacking and short-stacking using OLS as the final learner. The bootstrap sample size is $n_b = 9915$ or 99150 . Results are based on 1 000 replications. See Table 1 for more information.

Table A.3: Average stacking weights using single-best

<i>Panel (A): Linear DGP</i>	Stacking		Pooled stacking		Single-Best	
	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$
OLS	0.825	0.649	0.901	0.758	0.793	0.665
Lasso with CV (2nd order poly)	0.155	0.271	0.097	0.228	0.174	0.251
Ridge with CV (2nd order poly)	0.016	0.019	0.002	0.003	0.031	0.032
Lasso with CV (10th order poly)	0.002	0.040	0.	0.009	0.001	0.023
Ridge with CV (10th order poly)	0.002	0.014	0.	0.001	0.001	0.028
Random forest (low regularization)	0.	0.	0.	0.	0.	0.
Random forest (high regularization)	0.	0.	0.	0.	0.	0.
Gradient boosting (low regularization)	0.	0.004	0.	0.001	0.	0.001
Gradient boosting (high regularization)	0.	0.004	0.	0.	0.	0.
Neural net	0.	0.	0.	0.	0.	0.
<i>Panel (B): Non-Linear DGP</i>						
<i>Panel (B): Non-Linear DGP</i>	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$	$E[Y X]$	$E[D X]$
OLS	0.	0.	0.	0.	0.	0.
Lasso with CV (2nd order poly)	0.097	0.152	0.072	0.142	0.051	0.079
Ridge with CV (2nd order poly)	0.104	0.116	0.100	0.095	0.059	0.080
Lasso with CV (10th order poly)	0.089	0.055	0.052	0.031	0.044	0.033
Ridge with CV (10th order poly)	0.022	0.045	0.007	0.012	0.004	0.035
Random forest (low regularization)	0.	0.	0.	0.	0.	0.
Random forest (high regularization)	0.002	0.001	0.001	0.	0.001	0.
Gradient boosting (low regularization)	0.672	0.364	0.766	0.409	0.831	0.637
Gradient boosting (high regularization)	0.014	0.268	0.002	0.311	0.008	0.136
Neural net	0.001	0.001	0.	0.	0.002	0.

Notes: The table shows the (average) rates at which each candidate learner is selected by the single-best final learner when using conventional stacking, pooled stacking and short-stacking. The bootstrap sample size is $n_b = 9915$ or 99150 . Results are based on 1 000 replications. See Table 1 for more information.

Table A.4: Computational time of DDML with conventional and short-stacking

Folds K	Obs.	DDML		OLS	PDS lasso	Ratio
		Stacking Conv.	Short			
2	200	24.69	6.34	0.0072	0.0617	0.2567
	400	26.02	6.76	0.0073	0.0640	0.2597
	800	29.51	7.67	0.0074	0.0665	0.2598
	1600	41.23	10.53	0.0082	0.0780	0.2554
	9915	210.78	53.01	0.0170	0.2131	0.2515
	99150	3434.07	778.17	0.1094	1.6571	0.2266
5	200	59.41	13.46	0.0069	0.0588	0.2266
	400	69.18	15.76	0.0070	0.0617	0.2278
	800	88.57	20.77	0.0074	0.0662	0.2345
	1600	137.77	31.92	0.0082	0.0781	0.2317
	9915	848.27	196.97	0.0148	0.1841	0.2322
10	200	120.47	26.01	0.0068	0.0583	0.2159
	400	141.29	30.95	0.0070	0.0608	0.2191
	800	189.87	42.98	0.0075	0.0677	0.2264
	1600	295.87	68.22	0.0082	0.0778	0.2306
	9915	1962.00	453.13	0.0159	0.1998	0.2310

Notes: The table reports the computational time in seconds of DDML paired with conventional stacking ('Conv.') or short-stacking ('Short') as implemented in Ahrens et al. (2023), OLS as implemented in Stata's **regress**, post-double-selection lasso as implemented in **pdslasso** (Ahrens, Hansen, and Schaffer, 2018). DDML uses $V = 5$ cross-validation folds and K cross-fitting folds as indicated. Times reported are in seconds (average over 1 000 replications). The computations were performed on the high-performance cluster of the ETH Zurich. Each instance used a single core of an AMD EPYC processor with 2.25-2.6GHz (nominal)/3.3-3.5 GHz (peak) and 4GB RAM.

B DDML and stacking in very small samples

Table B.1: Estimates based on the full sample ($N = 9,915$).

<i>Estimator</i>	<i>Estimate</i>
<i>Panel A. No sample splitting</i>	
OLS TWI	6751.907
OLS QSI	5988.413
Post double Lasso TWI $c=0.5$	6562.923
Post double Lasso QSI $c=0.5$	5648.14
Post double Lasso TWI $c=1$	6630.751
Post double Lasso QSI $c=1$	4646.575
Post double Lasso TWI $c=1.5$	7474.508
Post double Lasso QSI $c=1.5$	4472.324
<i>Panel B. DDML with candidate learners</i>	
Neural net	6433.092
OLS	6463.73
Lasso with CV (TWI)	6780.161
Ridge with CV (TWI)	6760.134
Lasso with CV (QSI)	5722.624
Ridge with CV (QSI)	5995.346
Random forest (low regularization)	6089.389
Random forest (high regularization)	6552.221
Gradient boosting (low regularization)	7003.373
Gradient boosting (high regularization)	7992.538
<i>Panel C. DDML with stacking approaches</i>	
Neural net	6433.092
OLS	6463.73
Lasso with CV (TWI)	6780.161
Ridge with CV (TWI)	6760.134
Lasso with CV (QSI)	5722.624
Ridge with CV (QSI)	5995.346
Random forest (low regularization)	6089.389
Random forest (high regularization)	6552.221
Gradient boosting (low regularization)	7003.373
Gradient boosting (high regularization)	7992.538

Notes: In the case of DDML estimators, the average estimates and standard errors are based on 50 replications. Panel A is reproduced from Table 1 in WZ.

Table B.2: Short-stacking weights using CLS

<i>Estimator</i>	<i>Observations</i>						
	200	400	600	800	1200	1600	9915
<i>Panel A. $E[Y X]$, $K = 10$</i>							
OLS	.164	.152	.115	.079	.037	.019	0
Neural net	.047	.045	.048	.067	.098	.05	.076
Lasso with CV (TWI)	.043	.034	.034	.035	.03	.033	.091
Ridge with CV (TWI)	.056	.048	.041	.025	.011	.006	.032
Lasso with CV (QSI)	.252	.274	.266	.264	.271	.297	.639
Ridge with CV (QSI)	.194	.252	.297	.328	.341	.357	.153
Random forest (low regularization)	.095	.097	.113	.131	.161	.2	.01
Random forest (high regularization)	.081	.04	.025	.021	.018	.016	0
Gradient boosting (low regularization)	.041	.04	.049	.041	.03	.021	0
Gradient boosting (high regularization)	.028	.019	.013	.009	.002	.001	0
<i>Panel B. $E[D X]$, $K = 10$</i>							
OLS	.132	.196	.234	.252	.245	.257	.163
Neural net	.04	.041	.038	.036	.031	.029	.038
Lasso with CV (TWI)	.053	.031	.025	.02	.016	.012	.106
Ridge with CV (TWI)	.038	.018	.013	.015	.008	.005	.029
Lasso with CV (QSI)	.173	.225	.25	.248	.25	.228	.413
Ridge with CV (QSI)	.202	.124	.072	.06	.068	.064	0
Random forest (low regularization)	.103	.123	.144	.187	.249	.307	.006
Random forest (high regularization)	.159	.129	.107	.09	.051	.031	.102
Gradient boosting (low regularization)	.043	.046	.054	.047	.045	.041	.144
Gradient boosting (high regularization)	.059	.065	.064	.046	.038	.025	0
<i>Panel C. $E[Y X]$, $K = 10$</i>							
OLS	.122	.098	.066	.026	.003	.001	0
Neural net	0	0	0	0	0	0	0
Lasso with CV (TWI)	.03	.022	.01	.013	.014	.023	0
Ridge with CV (TWI)	.074	.077	.079	.052	.03	.013	0
Lasso with CV (QSI)	.323	.376	.361	.381	.393	.405	.995
Ridge with CV (QSI)	.239	.314	.379	.428	.478	.479	.005
Random forest (low regularization)	.129	.058	.05	.049	.049	.044	0
Random forest (high regularization)	.022	.005	.001	.001	0	.001	0
Gradient boosting (low regularization)	.025	.033	.046	.046	.032	.034	0
Gradient boosting (high regularization)	.035	.016	.009	.004	0	0	0
<i>Panel D. $E[D X]$, $K = 10$</i>							
OLS	.038	.108	.17	.189	.173	.132	.005
Neural net	0	0	0	0	0	0	0
Lasso with CV (TWI)	.058	.032	.017	.011	.005	.003	.002
Ridge with CV (TWI)	.06	.013	.009	.01	.002	.001	0
Lasso with CV (QSI)	.232	.309	.313	.287	.261	.168	.754
Ridge with CV (QSI)	.242	.105	.032	.034	.05	.032	0
Random forest (low regularization)	.079	.028	.011	.008	.004	.003	0
Random forest (high regularization)	.185	.249	.256	.304	.344	.507	0
Gradient boosting (low regularization)	.009	.022	.048	.064	.115	.141	.24
Gradient boosting (high regularization)	.098	.135	.143	.092	.046	.013	0

Notes: The table reports the stacking weights corresponding to the DDML stacking estimator in Figure 3. The stacking weights are averaged over folds, based on 10-fold cross-fitting and shows for the estimation of $E[Y|X]$ and $E[D|X]$ in Panel A and B, respectively. See notes below Table 3 for more information.

Table B.3: Mean bias in small samples based on the calibrated Monte Carlo in Section 4.1

	Panel A. Linear DGP					Panel B. Non-linear DGP				
	200	400	800	1600	9915	200	400	800	1600	9915
Full sample estimators:										
OLS	-246.2	-297.0	248.3	161.8	-20.2	-2291.4	-2023.2	-2631.2	-2857.2	-2582.1
PDS-Lasso	-1247.5	-1001.1	-159.3	72.0	-18.7	-2143.2	-2524.0	-3105.3	-3120.1	-2591.5
DDML methods:										
<i>Candidate learners (K = 10)</i>										
OLS	-282.8	-308.6	255.4	164.3	-19.9	-2833.5	-2364.7	-2861.6	-2952.6	-2600.8
Lasso with CV (2nd order poly)	-236.3	-206.1	277.8	174.3	-19.6	-912.8	-163.8	-82.8	264.4	754.2
Ridge with CV (2nd order poly)	-144.7	-187.2	303.5	151.2	-22.0	-1803.8	-728.0	122.1	697.7	778.0
Lasso with CV (10th order poly)	6339.9	180.0	1258.7	565.9	2.4	-7471.8	-3708.5	182.8	4054.5	-951.6
Ridge with CV (10th order poly)	5321.1	4716.5	6706.9	89.8	290.3	691.7	3137.6	-6506.9	-8867.9	515.8
Random forest (low regularization)	-149.1	-337.2	162.3	78.1	-110.7	-274.3	233.1	-448.8	-293.4	-15.0
Random forest (high regularization)	633.6	132.6	454.7	288.4	-15.3	-381.2	185.4	-361.3	-313.4	-74.3
Gradient boosting (low regularization)	-489.1	-498.4	125.6	113.6	-53.5	-539.0	182.4	-348.5	-209.5	59.7
Gradient boosting (high regularization)	-240.3	-284.2	291.9	257.5	40.8	-316.7	329.6	-159.4	-32.6	213.5
Neural net	3554.5	3955.7	3570.1	2153.1	129.7	1277.7	2089.6	1354.8	-132.9	-465.6
<i>Meta learners (K = 10)</i>										
Stacking: CLS	96.2	-1300.3	169.5	118.2	-24.9	-452.0	-1706.0	-1216.1	-73.8	188.5
Stacking: Single-best	-366.2	-1414.4	185.2	124.6	-25.2	-1774.8	-313.1	-622.0	-708.9	112.6
Short-stacking: CLS	-179.0	-258.7	306.6	185.3	-23.0	-462.7	55.0	-124.6	18.4	138.8
Short-stacking: Single-best	-308.2	-321.7	246.8	148.5	-22.6	-681.2	-149.9	-217.6	36.4	55.3
Pooled stacking: CLS	-127.3	-192.6	328.1	172.0	-23.3	-688.8	-70.8	-353.6	-60.2	160.4
Pooled stacking: Single-best	-236.5	-341.1	242.8	156.2	-22.5	-797.4	-215.0	-329.1	13.0	53.6
<i>Meta learners (K = 2)</i>										
Stacking: CLS	3032.5	1043.6	-618.3	-104.4	8.8	109.3	-638.7	-275.4	-1462.4	226.8
Stacking: Single-best	1496.3	-348.9	-111.6	-206.4	-28.6	-2024.5	-746.4	-38.9	-95.9	179.2
Short-stacking: CLS	730.8	-58.7	25.6	-117.5	-23.8	-518.0	-382.4	-302.4	-54.6	221.9
Short-stacking: Single-best	454.5	-186.4	-33.7	-172.1	-20.7	-966.8	-665.8	-460.4	-72.0	116.1
Pooled stacking: CLS	827.4	160.8	317.8	-37.1	-24.8	-719.8	-5280.9	488.6	207.5	251.5
Pooled stacking: Single-best	953.6	-334.9	-17.0	-170.8	-23.0	-1323.5	-2172.3	-1216.1	109.9	131.0

Notes: The table reports mean bias, median absolute bias (MAB) and coverage rate of a 95% confidence interval for the listed estimators. We consider DDML with the following individual learners: OLS with elementary covariates, CV lasso and CV ridge with second-order polynomials and interactions, CV lasso and CV ridge with 10th-order polynomials but no interactions, random forest with low regularization (8 predictors considered at each leaf split, no limit on the number of observations per node, bootstrap sample size of 70%), highly regularized random forest (5 predictors considered at each leaf split, at least 10 observation per node, bootstrap sample size of 70%), gradient-boosted trees with low regularization (500 trees and a learning rate of 0.01), gradient-boosted trees with high regularization: 250 trees and a learning rate of 0.01, feed-forward neural nets with three hidden layers of size five. For reference, we report two estimators using the full sample: OLS and PDS lasso. We report results for four meta learners: Stacking with CLS, short-stacking with CLS, single best overall and single best by fold. Results are based on 1,000 replications.

Table B.4: Coverage in small samples based on the calibrated Monte Carlo in Section 4.1

	Panel A. Linear DGP					Panel B. Non-linear DGP				
	200	400	800	1600	9915	200	400	800	1600	9915
Full sample estimators:										
OLS	0.95	0.95	0.96	0.96	0.95	0.94	0.95	0.92	0.91	0.59
PDS-Lasso	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.91	0.89	0.59
DDML methods:										
<i>Candidate learners</i> ($K = 10$)										
OLS	0.94	0.95	0.95	0.95	0.95	0.93	0.94	0.93	0.91	0.59
Lasso with CV (2nd order poly)	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.90
Ridge with CV (2nd order poly)	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.89
Lasso with CV (10th order poly)	0.90	0.92	0.93	0.93	0.95	0.87	0.85	0.85	0.88	0.95
Ridge with CV (10th order poly)	0.81	0.89	0.87	0.89	0.95	0.82	0.82	0.83	0.87	0.94
Random forest (low regularization)	0.92	0.92	0.93	0.92	0.91	0.93	0.92	0.93	0.93	0.91
Random forest (high regularization)	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.93
Gradient boosting (low regularization)	0.92	0.93	0.94	0.94	0.95	0.92	0.93	0.95	0.96	0.94
Gradient boosting (high regularization)	0.93	0.94	0.95	0.95	0.95	0.94	0.94	0.96	0.96	0.94
Neural net	0.94	0.92	0.90	0.91	0.95	0.94	0.95	0.95	0.97	0.94
<i>Meta learners</i> ($K = 10$)										
Stacking: CLS	0.93	0.94	0.95	0.95	0.95	0.94	0.93	0.94	0.94	0.94
Stacking: Single-best	0.93	0.94	0.95	0.95	0.95	0.94	0.95	0.95	0.94	0.94
Short-stacking: CLS	0.95	0.95	0.96	0.95	0.95	0.96	0.96	0.95	0.95	0.94
Short-stacking: Single-best	0.95	0.95	0.96	0.95	0.95	0.95	0.96	0.95	0.94	0.94
Pooled stacking: CLS	0.95	0.95	0.96	0.95	0.95	0.96	0.95	0.96	0.96	0.94
Pooled stacking: Single-best	0.95	0.95	0.96	0.95	0.95	0.95	0.96	0.95	0.95	0.94
<i>Meta learners</i> ($K = 2$)										
Stacking: CLS	0.90	0.94	0.91	0.94	0.95	0.92	0.93	0.93	0.94	0.93
Stacking: Single-best	0.92	0.95	0.91	0.94	0.96	0.92	0.94	0.94	0.95	0.92
Short-stacking: CLS	0.93	0.95	0.92	0.95	0.96	0.93	0.95	0.94	0.96	0.93
Short-stacking: Single-best	0.93	0.95	0.92	0.94	0.96	0.93	0.96	0.94	0.96	0.94
Pooled stacking: CLS	0.93	0.95	0.92	0.95	0.96	0.93	0.94	0.94	0.95	0.93
Pooled stacking: Single-best	0.93	0.95	0.91	0.94	0.96	0.93	0.95	0.95	0.95	0.93

Notes: The table reports mean bias, median absolute bias (MAB) and coverage rate of a 95% confidence interval for the listed estimators. We consider DDML with the following individual learners: OLS with elementary covariates, CV lasso and CV ridge with second-order polynomials and interactions, CV lasso and CV ridge with 10th-order polynomials but no interactions, random forest with low regularization (8 predictors considered at each leaf split, no limit on the number of observations per node, bootstrap sample size of 70%), highly regularized random forest (5 predictors considered at each leaf split, at least 10 observation per node, bootstrap sample size of 70%), gradient-boosted trees with low regularization (500 trees and a learning rate of 0.01), gradient-boosted trees with high regularization: 250 trees and a learning rate of 0.01, feed-forward neural nets with three hidden layers of size five. For reference, we report two estimators using the full sample: OLS and PDS lasso. We report results for four meta learners: Stacking with CLS, short-stacking with CLS, single best overall and single best by fold. Results are based on 1,000 replications.

C Gender citation gap

Table C.1: Estimates for the citation penalty of all-female and mixed-gender authored articles

	Log citations		Citation counts	
	<i>All female</i>	<i>Mixed gender</i>	<i>All female</i>	<i>Mixed gender</i>
OLS & base controls	0.007 (0.029)	0.212* (0.017)	-9.839** (4.299)	11.368* (2.889)
PDS-lasso & base controls	0.009 (0.029)	0.214* (0.017)	-9.608** (4.293)	11.558* (2.886)
PDS-lasso & Unigrams	-0.069** (0.028)	0.097* (0.016)	-15.77* (4.141)	2.809 (2.896)
PDS-lasso & BERT	-0.081* (0.028)	0.086* (0.016)	-14.634* (4.236)	2.459 (2.89)
OLS & Unigrams	-0.062** (0.029)	0.094* (0.017)	-12.227** (5.957)	2.464 (3.48)
OLS & BERT	-0.082* (0.029)	0.061* (0.017)	-10.43** (5.968)	1.58 (3.608)
CV-lasso & Unigrams	-0.061** (0.029)	0.093* (0.017)	-12.512** (5.948)	2.359 (3.477)
CV-lasso & BERT	-0.077* (0.029)	0.063* (0.017)	-10.055** (5.967)	2.054 (3.495)
CV-ridge & Unigrams	-0.062** (0.029)	0.093* (0.017)	-12.398** (5.947)	1.859 (3.479)
CV-ridge & BERT	-0.074* (0.029)	0.066* (0.017)	-10.108** (5.961)	1.893 (3.499)
XGBoost & Unigrams	-0.087* (0.029)	0.058* (0.018)	32.376* (6.153)	19.504* (3.622)
XGBoost & BERT	-0.075** (0.031)	0.053* (0.018)	7.336 (7.049)	7.883** (3.894)
Random forest & Unigrams	-0.11* (0.029)	0.034** (0.017)	0.037 (5.912)	4.672 (3.469)
Random forest & BERT	-0.134* (0.032)	0.002 (0.018)	6.671 (6.025)	6.319** (3.53)
Neural net & Unigrams	-0.042 (0.043)	-0.029 (0.057)	-15.544 (10.062)	-11.383** (5.62)
Neural net & BERT	-0.061 (0.08)	-0.044 (0.143)	-17.923* (6.154)	-11.332* (3.682)
Single-best & stacking	-0.077* (0.028)	0.064* (0.017)	-10.177** (5.969)	1.981 (3.512)
CLS & stacking	-0.077* (0.028)	0.057* (0.016)	-13.085** (5.978)	-2.618 (3.454)
Single-best & short-stacking	-0.076* (0.029)	0.063* (0.017)	-10.108** (5.961)	1.893 (3.499)
CLS & short-stacking	-0.075* (0.028)	0.054* (0.016)	-9.556 (5.933)	1.507 (3.492)

Notes: The table shows median-aggregated estimates of the gender citation gap for all-female and mixed-gender authored articles. We show results using both log citations and citation counts as the outcome variable. Standard errors are robust to heteroskedasticity. See Table C.1 for information on the candidate learners and stacking approaches.

D Gender wage gap

Table D.1: Stacking weights in the gender wage gap application.

	<i>Conventional stacking</i>			<i>Short-stacking</i>			<i>Mean-squared error</i>		
	$g_0(0, X)$	$g_0(1, X)$	$m_0(X)$	$g_0(0, X)$	$g_0(1, X)$	$m_0(X)$	$g_0(0, X)$	$g_0(1, X)$	$m_0(X)$
OLS/logit	0.023	0.012	0.242	0.027	0.013	0.211	0.369	0.347	0.161
OLS/logit (simple)	0.004	0.	0.	0.	0.	0.	0.267	0.204	0.223
CV-lasso	0.103	0.136	0.109	0.03	0.076	0.047	0.236	0.178	0.16
CV-ridge	0.189	0.04	0.064	0.225	0.024	0.108	0.237	0.18	0.161
CV-lasso (extended)	0.041	0.157	0.016	0.035	0.266	0.002	0.238	0.18	0.161
CV-ridge (extended)	0.011	0.04	0.011	0.003	0.024	0.022	0.336	0.194	0.161
Random forest 1	0.435	0.506	0.275	0.483	0.507	0.28	0.23	0.176	0.161
Random forest 2	0.	0.	0.	0.	0.	0.	0.258	0.19	0.171
Random forest 3	0.	0.	0.	0.	0.	0.	0.274	0.199	0.179
Gradient boosting 1	0.025	0.008	0.039	0.011	0.003	0.022	0.239	0.183	0.16
Gradient boosting 2	0.15	0.059	0.216	0.175	0.063	0.285	0.254	0.196	0.161
Neural net 1	0.013	0.022	0.	0.	0.	0.	0.349	0.263	0.241
Neural net 2	0.008	0.02	0.027	0.01	0.023	0.023	0.643	0.357	0.176

Notes: The table shows weights of conventional and short-stacking along with the mean-squared prediction error by candidate learners and by variable. The final learner is constrained least squares. The stacking weights are averaged over cross-fitting repetitions. Pooled stacking weights are shown in Appendix Table D.2.

Table D.2: Stacking weights of pooled stacking using constrained least squares.

	<i>Pooled stacking</i>		
	$g_0(0, X)$	$g_0(1, X)$	$m_0(X)$
OLS/logit	0.014	0.001	0.257
OLS/logit (simple)	0.	0.	0.
CV-lasso	0.15	0.23	0.136
CV-ridge	0.205	0.064	0.063
CV-lasso (extended)	0.	0.078	0.
CV-ridge (extended)	0.	0.019	0.
Random forest 1	0.462	0.521	0.288
Random forest 2	0.	0.	0.
Random forest 3	0.	0.	0.
Gradient boosting 1	0.	0.	0.008
Gradient boosting 2	0.165	0.071	0.23
Neural net 1	0.	0.	0.
Neural net 2	0.004	0.016	0.018

Notes: The table shows pooled stacking weights for each of the considered candidate learners. The final learner is constrained least squares. The stacking weights are averaged over cross-fitting repetitions.

Table D.4: Median aggregate estimates by stacking approach and by final learner

	<i>Final learner</i>			
	<i>Unweighted average</i>	<i>CLS</i>	<i>OLS</i>	<i>Single-best</i>
Regular stacking	-0.101 (0.017)*	-0.075 (0.028)*	-0.197 (11.894)	-0.061 (0.069)
Short- stacking	-0.101 (0.017)*	-0.076 (0.028)*	-0.001 (0.184)	-0.085 (0.065)

Notes: The table reports median aggregate estimates by stacking type and final learner. See Figure 7 for more information.

Table D.3: Stacking weights using single-best final learner.

	<i>Conventional stacking</i>			<i>Short-stacking</i>			<i>Pooled stacking</i>		
	$g_0(0, X)$	$g_0(1, X)$	$m_0(X)$	$g_0(0, X)$	$g_0(1, X)$	$m_0(X)$	$g_0(0, X)$	$g_0(1, X)$	$m_0(X)$
OLS/logit	0.	0.	0.	0.	0.	0.	0.	0.	0.
OLS/logit (simple)	0.	0.	0.	0.	0.	0.	0.	0.	0.
CV-lasso	0.06	0.03	0.79	0.	0.	1.	0.	0.	0.4
CV-ridge	0.07	0.	0.04	0.	0.	0.	0.	0.	0.1
CV-lasso (extended)	0.02	0.06	0.03	0.	0.	0.	0.	0.	0.1
CV-ridge (extended)	0.	0.	0.01	0.	0.	0.	0.	0.	0.
Random forest 1	0.85	0.91	0.02	1.	1.	0.	1.	1.	0.
Random forest 2	0.	0.	0.	0.	0.	0.	0.	0.	0.
Random forest 3	0.	0.	0.	0.	0.	0.	0.	0.	0.
Gradient boosting 1	0.	0.	0.11	0.	0.	0.	0.	0.	0.4
Gradient boosting 2	0.	0.	0.	0.	0.	0.	0.	0.	0.
Neural net 1	0.	0.	0.	0.	0.	0.	0.	0.	0.
Neural net 2	0.	0.	0.	0.	0.	0.	0.	0.	0.

Notes: The table shows weights of conventional stacking, short-stacking and pooled stacking by candidate learners and by conditional expectation function. The stacking weights are averaged over cross-fitting repetitions.

Table D.5: Median aggregate estimates for each candidate learner

	<i>Gender wage gap</i>
OLS/logit	-0.12 (0.094)
CV-lasso	-0.067 (0.063)
CV-ridge	-0.064 (0.09)
OLS/logit (simple)	-0.12 (0.016)*
CV-lasso (extended)	-0.055 (0.076)
CV-ridge (extended)	-0.173 (0.19)
Random forest 1	-0.079 (0.023)*
Random forest 2	-0.105 (0.016)*
Random forest 3	-0.11 (0.015)*
Gradient boosting 1	-0.075
Observations	4836

Notes: The table reports median aggregate estimates by candidate learner. See Figure 7 for more information.