MACHINE LEARNING

Prediction-powered inference

Anastasios N. Angelopoulos*†, Stephen Bates*†, Clara Fannjiang*†, Michael I. Jordan*†, Tijana Zrnic*†

Prediction-powered inference is a framework for performing valid statistical inference when an experimental dataset is supplemented with predictions from a machine-learning system. The framework yields simple algorithms for computing provably valid confidence intervals for quantities such as means, quantiles, and linear and logistic regression coefficients without making any assumptions about the machine-learning algorithm that supplies the predictions. Furthermore, more accurate predictions translate to smaller confidence intervals. Prediction-powered inference could enable researchers to draw valid and more data-efficient conclusions using machine learning. The benefits of prediction-powered inference were demonstrated with datasets from proteomics, astronomy, genomics, remote sensing, census analysis, and ecology.

Imagine a scientist has a machine-learning system that can supply accurate predictions about a phenomenon far more cheaply than any gold-standard experimental technique. The scientist may wish to use these predictions as evidence in drawing scientific conclusions. For example, accurate predictions of three-dimensional structures have been made for a vast catalog of known protein sequences (*1*, *2*) and are now being used in proteomics studies (*3*, *4*). Such machine-learning systems are increasingly common in modern scientific inquiry, in domains ranging from cancer prognosis to microclimate modeling. Predictions are not perfect, however, and this may lead to incorrect conclusions. Moreover, as predictions beget other predictions, the cumulative effect can amplify the imperfections. How can modern science leverage machine-learning predictions in a statistically principled way?

One way to use predictions is to follow the imputation approach: Proceed as if they are gold-standard measurements. Although this lets the scientist draw conclusions cheaply and quickly owing to the high-throughput nature of the machine-learning system, the conclusions may be invalid because the predictions may have biases.

Another possibility is to apply the classical approach: Ignore the machine-learning predictions and only use the available gold-standard measurements, which are typically far less abundant than predictions. The resulting discoveries will be statistically valid, but the smaller amount of data will limit the scope of possible discoveries.

This manuscript presents prediction-powered inference, a framework that achieves the best of both worlds: extracting information from the predictions of a high-throughput machine-learning system and guaranteeing statistical validity of the resulting conclusions. Prediction-powered inference provides a protocol for combining predictions, which are abundant but not always trustworthy, with gold-standard data, which are trusted but scarce, to compute confidence intervals and *P* values. The resulting confidence intervals and *P* values are statistically valid, as in the classical approach, but also leverage the information contained in the predictions, as in the imputation approach, to make the confidence intervals smaller and the *P* values more powerful.

Prediction-powered inference applies to any machine-learning system; as such, it absolves the need for case-by-case analyses dependent on the machine-learning algorithm on hand. The proposed protocol thereby could enable researchers to report on and assess the evidence for their conclusions in a fully standardized way.

Protocol for prediction-powered inference

The protocol for prediction-powered inference proceeds as follows. The scientist wishes to construct a confidence interval for a quantity θ^*, such as the mean outcome or a regression coefficient quantifying the statistical association between the outcome and a feature. Toward this goal, they have access to a small gold-standard dataset of features paired with outcomes, $(X, Y) = ((X_1, Y_1), \ldots, (X_n, Y_n))$, as well as the features of a large unlabeled dataset, $(X', Y') = \left((X'_1, Y'_1), \ldots, (X'_N, Y'_N)\right)$, where the true outcomes $Y'_1, \ldots, Y'_N$ are not observed. Typically, N is much larger than n . Both datasets are sampled at random from a larger population. Further, for both datasets the scientist has predictions of the outcomes made by a machine-learning algorithm based on the features, denoted $\left(\hat{Y}_1, \ldots, \hat{Y}_n\right)$ and $\left(\hat{Y}'_1, \ldots, \hat{Y}'_N\right)$, respectively. The following exposition focuses on confidence intervals; however, by the standard duality between confidence intervals and *P* values, the presented tools immediately carry over to valid *P*-value constructions and hypothesis tests; see supplementary materials (SM) for details.

Prediction-powered inference uses the gold-standard dataset to quantify and correct for the errors made by the machine-learning algorithm on the unlabeled dataset, thereby enabling researchers to reliably incorporate predictions when constructing confidence intervals. The three-step protocol is outlined below and visualized in Fig. 1.

1) Estimand. The first step is to select an estimand θ^*. The estimand is the quantity the scientist is interested in knowing—for example, the mean outcome $E[Y_i]$, median outcome $\text{median}(Y_i)$, a linear regression coefficient obtained by regressing Y onto X, etc.

2) Measure of fit and rectifier. The key step is to identify the right measure of fit m_θ and rectifier Δ_θ for the selected estimand. For every candidate value of the estimand θ, the measure of fit m_θ is computed on the unlabeled dataset imputed with predictions, $(X', \hat{Y}')$ and quantifies how likely θ^* is to be equal to θ on the basis of the imputed data. The closer m_θ is to zero, the more plausible it is for θ^* to be equal to θ.

The rectifier Δ_θ is a notion of prediction error that is relevant for the estimand of interest. It is defined as the difference of the measure of fit m_θ computed on the labeled data, (X, Y), and the labeled data when the true outcomes are replaced with predicted ones, $\left(X, \hat{Y}\right)$. If the predictions are perfect, the rectifier is equal to zero.

Table 1 states the appropriate measure of fit and rectifier for common estimands of interest: the mean outcome, median outcome, q-quantile of the outcome, and linear and logistic regression coefficients when regressing Y onto X. A general recipe for deriving the right measure of fit and corresponding rectifier for a broad class of other estimands is provided in the SM.

3) Prediction-powered confidence interval. Finally, the measure of fit and rectifier are carefully combined to form a prediction-powered confidence interval for θ^*. This process is called rectifying the confidence interval. The prediction-powered confidence interval is constructed as $C^{\text{PP}} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$ and is guaranteed to contain the estimand with probability at least $1 - \alpha$. Here, $w_\theta(\alpha)$ is a constant that depends on the confidence level; it is explicitly stated in Theorem S1 in the SM.

Properties of prediction-powered inference

We proved mathematically that prediction-powered inference yields a confidence interval that contains the true value of the estimand at the desired confidence level, such as 95%. Notably, this validity is guaranteed for any machine-learning algorithm and any underlying data distribution. Similarly, the corresponding *P* values are also valid for any machine-learning algorithm and data distribution. See SM for the details of the mathematical proof

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. Email: angelopoulos@berkeley.edu (A.N.A); stephenbates@berkeley.edu (S.B.); clarafy@berkeley.edu (C.F.); michael_jordan@berkeley.edu (M.I.J.); tijana.zrnic@berkeley.edu (T.Z.)
†These authors contributed equally to this work.

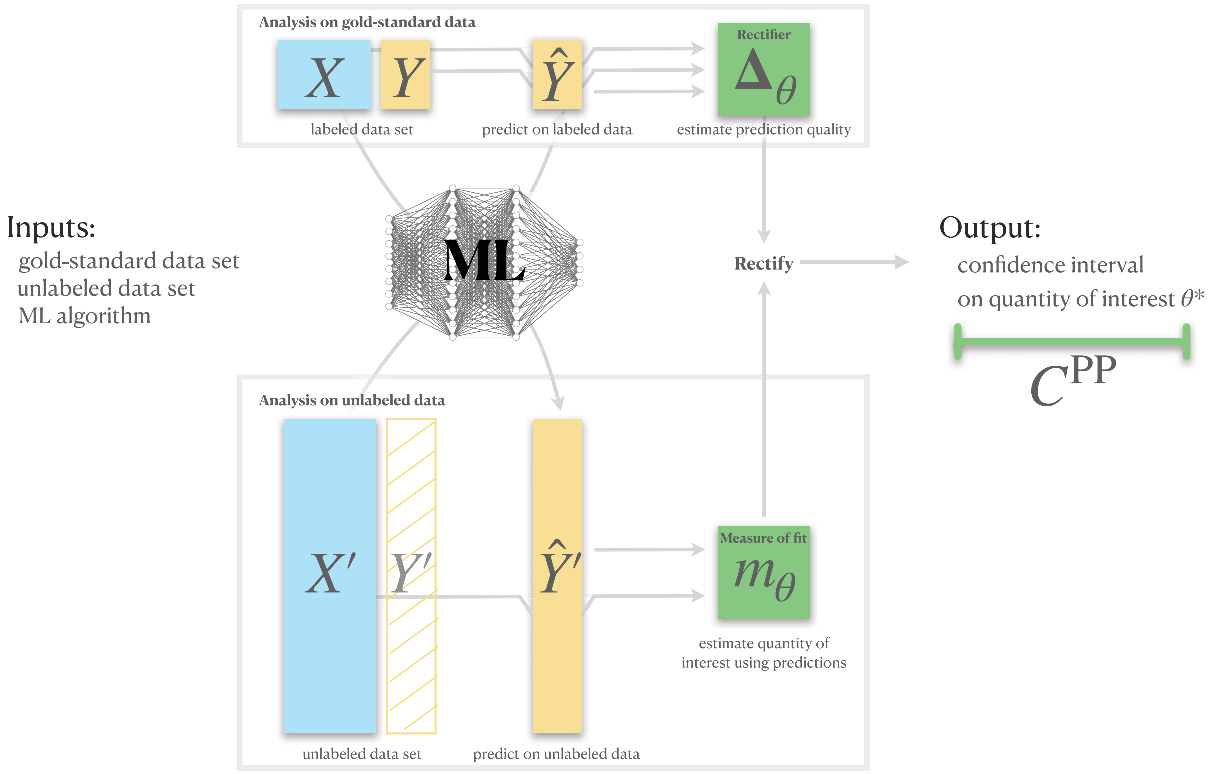

Fig. 1. Protocol for prediction-powered inference. The protocol is illustrated graphically as a block diagram. The inputs are the gold-standard dataset, the unlabeled dataset, and the machine-learning (ML) algorithm. The top block contains an analysis on gold-standard data, in which the rectifier, a measure of the prediction errors, is estimated using the labeled dataset. The bottom block contains an analysis on unlabeled data, wherein the quantity of interest is estimated using predictions. These analyses combine to form the prediction-powered confidence interval. For concrete examples of the rectifier and measure of fit, see Table 1. For a detailed theoretical exposition and more general definitions of these quantities, see SM.

of validity. A researcher relying on a deep neural network for predictions can therefore draw reliable conclusions, even though its predictions will inevitably be imperfect. Furthermore, prediction-powered inference enables more informative inferences than the classical approach, in which the researcher does not use machine-learning predictions: The confidence intervals are narrower, and the P values are more powerful. This is intuitive; prediction-powered inference carefully extracts information from the imputed data and thus has access to a larger sample size.

General applicability

Beyond quantities such as means, quantiles, and regression coefficients, the principle of prediction-powered inference can be used for constructing valid confidence intervals for any estimand that can be expressed as the minimizer of a convex objective function. This master protocol, which generalizes all the special cases instantiated in Table 1, is the core technical contribution of this work. We explained prediction-powered inference in greater generality and proved its validity in this general case in the SM. Because many important quantities can be expressed in terms of a convex-optimization problem, prediction-powered inference thus addresses many data-analysis goals beyond those explicitly demonstrated in this article.

Inference under distribution shift

Prediction-powered inference is also applicable to settings with distribution shift, i.e., the more challenging case where the unlabeled data are collected under different conditions than the gold-standard data. Two types of distribution shift are considered: label shift and covariate shift. The protocol retains the same properties as before: It is statistically valid for any machine-learning algorithm and boosts statistical power by making use of machine-learning predictions.

For covariate shift—the setting where only the feature distribution changes between the labeled and the unlabeled data—prediction-powered inference handles all estimation problems handled by the master protocol. This is done by appropriately reweighting the data; see Corollary S13 in the SM for details.

For label shift—the setting where only the label proportions change between the labeled and the unlabeled data—prediction-powered inference can be applied to estimands of the form $\theta^* = \mathrm{E}\left[\nu\left(Y_i'\right)\right]$, for a fixed function ν. For example, choosing $\nu(y) = 1\{y = k\}$ asks for inference on the proportion of instances that belong to class k. See Theorem S3 in the SM for a full description of the method.

Application of prediction-powered inference to real datasets

We demonstrated prediction-powered inference on several real tasks. In each, we computed a prediction-powered confidence interval for an estimand and compared it to intervals obtained through the classical approach and the imputation approach. In all cases, the imputation approach, which uses machine-learning predictions without accounting for prediction errors, did not contain the true value of the estimand. The widths of the two valid approaches, prediction-powered and classical, were compared as a function of the amount of labeled data used. In addition, we compared the number of labeled examples needed to reject a null hypothesis at level $1 - \alpha = 95\%$ with high probability. See (*5*) for a Python package implementing prediction-powered inference, which contains code for reproducing the experiments, and (*6*) for the data used in the experiments.

Relating protein structure and posttranslational modifications

The goal was to characterize whether various types of posttranslational modifications

Table 1. Prediction-powered inference for common statistical problems. Given a measure of fit m_θ and rectifier Δ_θ, prediction-powered inference computes a confidence interval as $C^{\mathrm{PP}} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$, where $w_\theta(\alpha)$ is a constant that depends on the error level α (see Theorem S1 in the SM). Algorithms S1 to S6 are stated in the SM. The last row ("convex minimizer") refers to a method that generalizes the methods in previous rows.

Estimand	Measure of fit m_θ	Rectifier Δ_θ	Procedure
Mean outcome	$\theta - \frac{1}{N}\sum_{i=1}^{N}\hat{Y}_i'$	$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)$	Alg. S1
Median outcome	$\frac{1}{2N}\sum_{i=1}^{N}\mathrm{sign}\left(\theta\text{-}\hat{Y}_i'\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \leq \theta\} - 1\left\{\hat{Y}_i \leq \theta\right\}\right)$	Alg. S2
q-quantile of outcome	$-q + \frac{1}{N}\sum_{i=1}^{N}1\left\{\hat{Y}_i' \leq \theta\right\}$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \leq \theta\} - 1\left\{\hat{Y}_i \leq \theta\right\}\right)$	Alg. S3
Linear regression	$\theta - (X')^{+}\hat{Y}'$	$X^{+}(\hat{Y} - Y)$	Alg. S4
Logistic regression	$\frac{1}{N}\sum_{i=1}^{N}X_i'\left(\frac{1}{1+e^{-\theta^{T}X_i'}} - \hat{Y}_i'\right)$	$\frac{1}{n}\sum_{i=1}^{n}X_i\left(\hat{Y}_i - Y_i\right)$	Alg. S5
Convex minimizer	$\frac{1}{N}\sum_{i=1}^{N}\nabla L_\theta\left(X_i', \hat{Y}_i'\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(\nabla L_\theta\left(X_i, \hat{Y}_i\right) - \nabla L_\theta(X_i, Y_i)\right)$	Alg. S6

(PTMs) occurred more frequently in intrinsically disordered regions (IDRs) of proteins (*7*). Recently, Bludau *et al.* (*3*) studied this relationship on an unprecedented proteome-wide scale by using structures predicted by AlphaFold (*1*) to predict IDRs, in contrast to previous work, which was limited to far fewer experimentally derived structures.

To quantify the association between PTMs and IDRs, the authors applied the imputation approach: They computed the odds ratio between AlphaFold-based IDR predictions and PTMs on a dataset of hundreds of thousands of protein sequence residues (*8*). Using prediction-powered inference, we could combine AlphaFold-based predictions together with gold-standard IDR labels to give a confidence interval for the true odds ratio that is statistically valid, in contrast with the interval constructed with the imputation approach, and smaller than the interval constructed using the classical approach. We used the fact that the odds ratio could be written in terms of two means and applied the recipe from the first row of Table 1; see SM for details.

We had 10,803 data points from Bludau *et al.* (*3*). For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and treated the remaining $N = 10,803 - n$ points as the unlabeled dataset for which we did not observe the IDR labels. For all values of n and all three different types of PTMs that we examined, the prediction-powered confidence intervals were smaller than classical intervals; see row A in Fig. 2. Often, the classical intervals were large enough that they contained the odds ratio value of one, which means the direction of the association could not be determined from the confidence interval. However, the imputed confidence interval was far too small and significantly overestimated the true odds ratio. To reject the null hypothesis that the odds ratio is no greater than one, prediction-powered inference required $n = 316$ labeled observations, and the classical approach required $n = 799$ labeled observations; see row A in Table 2.

Galaxy classification

The goal was to determine the demographics of galaxies with spiral arms, which are correlated with star formation in the disks of low-redshift galaxies, and therefore, contribute to the understanding of star formation in the Local Universe. A large citizen science initiative called Galaxy Zoo 2 (*9*) has collected human annotations of roughly 300,000 images of galaxies from the Sloan Digital Sky Survey (*10*) with the goal of measuring these demographics. We sought to explore the use of machine learning to improve the effective sample size and decrease the requisite number of human-annotated galaxies.

We focused on estimating the fraction of galaxies with spiral arms. We had 1,364,122 labeled galaxy images from Galaxy Zoo 2, from which we simulated labeled and unlabeled datasets as follows. For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and used the remaining $N = 1,364,122 - n$ points as the unlabeled dataset. We then used the first row of Table 1 to construct prediction-powered intervals. The prediction-powered confidence intervals for the mean were consistently much smaller than the classical intervals and they retained validity, and the imputation strategy failed to cover the ground truth; see Fig. 2, row B. To reject the null hypothesis that the fraction of galaxies with spiral arms is at most 0.2, prediction-powered inference required $n = 189$ labeled examples, and classical inference required $n = 449$ examples; see Table 2, row B.

Distribution of gene expression levels

Next, we constructed prediction-powered confidence intervals on quantiles that characterize how a population of promoter sequences affects gene expression. Recently, Vaishnav *et al.* (*11*) trained a state-of-the-art transformer model to predict the expression level of a particular gene induced by a promoter sequence. They used the model's predictions to study the effects of promoters—for example, by assessing how quantiles of predicted expression levels differ between different populations of promoters.

Here we focused on estimating different quantiles of gene expression levels induced by native yeast promoters. We had 61,150 labeled native yeast promoter sequences from Vaishnav *et al.* (*11*), from which we simulated labeled and unlabeled datasets as follows. For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and used the remaining $N = 61,150 - n$ points as the unlabeled dataset. We then used the second and third row of Table 1 to construct prediction-powered intervals for the median, as well as the 25% and 75% quantiles, of the expression levels. The prediction-powered confidence intervals for all three quantiles were much smaller than the classical intervals for all values of n. See row C in Fig. 2 for the results for the median and fig. S6 for the other two quantiles. We also evaluated the number of labeled examples required by prediction-powered inference and classical inference, respectively, to reject the null hypothesis that the median gene expression level is at most five. Prediction-powered inference required $n = 764$ examples and classical inference required $n = 900$ examples; see row C in Table 2.

Estimating deforestation in the Amazon

The goal was to estimate the fraction of the Amazon rainforest lost between 2000 and 2015. Gold-standard deforestation labels for parcels of land are scarce, having been collected in large part through field visits, an expensive process not suited for large areas (*12*). However, machine-learning predictions of forest cover based on satellite imagery are readily available for the entire Amazon (*13*). We began with 1596 gold-standard deforestation labels for parcels of land in the Amazon. For each of 100 trials, we randomly sampled n data points to serve as the labeled dataset and used the remaining data points as the unlabeled dataset. We used the first row of Table 1 to construct the prediction-powered intervals. The imputation approach yielded a small confidence interval that failed to cover the true deforestation fraction. The classical approach did cover the truth at the expense of a wider interval and, accordingly, diminished inferential power. The prediction-powered intervals were smaller than the classical intervals and retained validity; see row D in Fig. 2. We also compared the number of gold-standard deforestation labels required by prediction-powered inference and the classical approach to reject the null hypothesis that there is no deforestation. We obtain $n = 21$ labels for prediction-powered inference and $n = 35$ labels for the classical approach; see row D in Table 2.

Relationship between income and private health insurance

The goal was to investigate the quantitative effect of income on the procurement of private health insurance using US census data. Concretely, we used the Folktables interface (*14*) to download census data from California in the year 2019 (378,817 individuals). As the labeled dataset with the health insurance indicator, n census entries were randomly sampled. The remaining data were used as the unlabeled dataset. We used a gradient-boosted tree (*15*) trained on the previous year's data to predict the health insurance indicator in 2019. We constructed a prediction-powered confidence interval on the logistic regression coefficient using the fifth row of Table 1. Results in row E in Fig. 2 show that prediction-powered inference covered the ground truth, the classical interval was wider, and the imputation strategy failed to cover the ground truth. We also compared the number of gold-standard labels required by prediction-powered inference and the classical approach to reject the null hypothesis that the logistic regression coefficient is no greater than 1.5×10^{-5}. We observed a significant sample size reduction with prediction-powered inference, which required $n = 5569$ labels, whereas classical inference required $n = 6653$ labels.

Relationship between age and income in a covariate-shifted population

The goal was to investigate the relationship between age and income using US census data. The same dataset was used as in the previous experiment, but the features were age and sex, and the target was yearly income in dollars. Furthermore, a shift in the distribution of the covariates was introduced between the gold-standard and unlabeled datasets by randomly sampling the unlabeled dataset with sampling weights of 0.8 for females and 0.2 for males. We used a gradient-boosted tree (*15*) trained on the previous year's raw data to predict the income in 2019. We constructed a prediction-powered confidence interval on the ordinary least squares (OLS) regression coefficient using a covariate-shift robust version of prediction-powered inference, stated in Corollary S13 in the SM. Results in row F in Fig. 2 show that

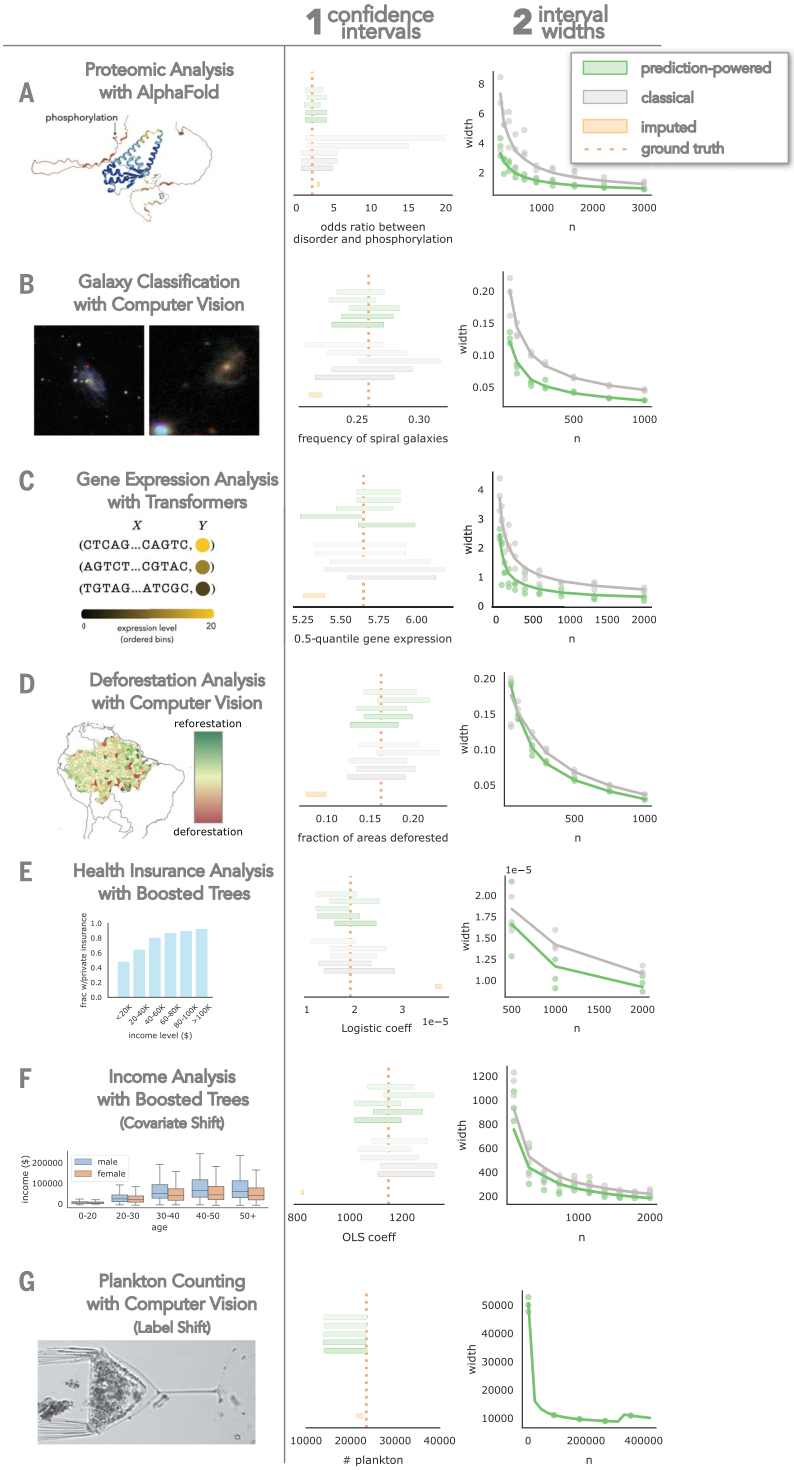

Fig. 2. Comparison of prediction-powered inference to the classical and imputation approaches on real tasks. Each row (**A** to **G**) is a different application domain. Panel 1 plots confidence intervals computed using the three approaches; for prediction-powered inference and the classical approach, intervals for five randomly chosen splits into labeled and unlabeled data are plotted. The value denoted as "ground truth" is the estimate computed on all $n + N$ data points (the true labels were available for all data points for the purpose of conducting the experiments). Panel 2 plots the average confidence interval width, as well as the width in five randomly chosen trials, for varying n, for prediction-powered inference and the classical approach; both are statistically valid solutions. The last problem setting (G) does not have a classical counterpart because the data are collected under distribution shift, hence the classical approach is not valid.

Table 2. Number of labeled examples needed to make a discovery with prediction-powered inference and classical inference. The rows (A to F) correspond to the application domains from Fig. 2. For each application, a null hypothesis about θ^* is tested at level 95%. For details, see the SM.

Problem	Prediction-powered inference	Classical inference
A Proteomic analysis with AlphaFold	$n = 316$	$n = 799$
B Galaxy classification with computer vision	$n = 189$	$n = 449$
C Gene expression analysis with transformers	$n = 764$	$n = 900$
D Deforestation analysis with computer vision	$n = 21$	$n = 35$
E Health insurance analysis with boosted trees	$n = 5569$	$n = 6653$
F Income analysis with boosted trees	$n = 177$	$n = 282$

prediction-powered inference covered the ground truth, the classical interval was wider, and the imputation strategy failed to cover the ground truth. We also compared the number of gold-standard labels required by prediction-powered inference and the classical approach to reject the null hypothesis that the OLS regression coefficient is no greater than 800 . We observed a significant sample size reduction with prediction-powered inference, which required $n = 177$labels, whereas classical inference required $n = 282$ labels.

Counting plankton

Assessment of the increases in phytoplankton growth during springtime warming is important for the study of global biogeochemical cycling in response to climate change. We counted the number of plankton observed by the Imaging FlowCytobot (*16*, *17*), an automated, submersible flow cytometry system, at Woods Hole Oceanographic Institution in the year 2014. We had access to data from 2013, which were labeled, and we imputed the 2014 data with machine-learning predictions from a state-of-the-art ResNet finetuned on all data up to and including 2012. The features, X_i, are images of organic matter taken by the FlowCytobot and the labels, Y_i, are one of {detritus, plankton}, where detritus represents unspecified organic matter.

The labeled dataset consisted of 421,238 image–label pairs from 2013, and we received 329,832 labeled images from 2014. We used the data from 2014 as our unlabeled data and confirmed our results against those that were hand-labeled. The years 2013 and 2014 had a distribution shift, primarily caused by the change in the base frequency of plankton observations with respect to detritus. To apply prediction-powered inference to count the number of plankton recorded in 2014, we used the label-shift-robust technique described in Theorem S3 in the SM. The results in row G in Fig. 2 show that prediction-powered inference covered the ground truth and the imputation strategy failed to cover the ground truth.

Related work

Thematically, prediction-powered inference is most similar to the work of Wang *et al.* (*18*), who introduced a method to correct machine-learning predictions for the purpose of subsequent inference. However, this procedure is not guaranteed to provide coverage in general and requires strong assumptions about the relationship between the prediction model and the true response, whereas prediction-powered inference provides provably valid conclusions under minimal assumptions about the data-generating distribution.

There has been an increasing body of work on estimation with many unlabeled data points and few labeled data points (*19*–*27*), focusing on efficiency in semiparametric or high-dimensional regimes. Prediction-powered inference continues in this vein but focuses on the setting where the scientist has access to a good predictive model fit on separate data. This allows tackling a much wider range of estimands (e.g., minimizers of any convex objective) and gives valid inferences without assumptions about the machine-learning model. Second, prediction-powered inference goes beyond random sampling and applies to certain forms of distribution shift.

Prediction-powered inference is conceptually related to conformal prediction (*28*). Both methodologies leverage a predictive model and a labeled dataset. From this point on, however, the two methods diverge: Prediction-powered inference has additional unlabeled data and gives a confidence set that contains a population-level quantity such as the mean outcome with high probability; conformal prediction gives a confidence set for a test instance that contains the true label with high probability. Thus, the goals of prediction-powered inference and conformal prediction differ greatly from the statistical perspective. Furthermore, the mathematical tools used in the frameworks are entirely different, and neither method can be applied nontrivially to solve the objective of the other.

See SM for a further discussion of related work and the relationship of prediction-powered inference to existing baselines, as well as for empirical comparisons.

Conclusions

The past decade has witnessed rapid development and deployment of large-scale machine-learning systems across science. This surge is proceeding, however, with little statistical justification to allow these black-box systems to be used to draw scientific conclusions responsibly. Prediction-powered inference is a standardized protocol for constructing provably valid confidence intervals and *P* values, allowing the scientist to use the power and scale of machine-learning systems. On an array of scientific problems, we demonstrated that prediction-powered inference achieved high statistical power owing to the use of machine-learning predictions and retained statistical validity.

One question that remains open is how to handle more general forms of distribution shift. In practice, distribution shifts are often a result of a joint influence of several different forms of shift, including covariate shift and label shift and possibly others. Understanding how to handle such settings remains an important avenue for future work.

A limitation of prediction-powered inference is that it does not improve upon the classical approach when the predictions are not accurate enough or when the unlabeled dataset is not large enough compared to the gold-standard dataset. These points are demonstrated, both theoretically and empirically, in SM section "Cases Where Prediction-Powered Inference Is Underpowered." Nevertheless, given the growing number of settings with excellent predictive models and abundant unlabeled data, there is increasing potential for prediction-powered inference to benefit scientific research.

REFERENCES AND NOTES

1. J. Jumper *et al.*, *Nature* **596**, 583–589 (2021).
2. K. Tunyasuvunakool *et al.*, *Nature* **596**, 590–596 (2021).
3. I. Bludau *et al.*, *PLOS Biol.* **20**, e3001636 (2022).
4. I. Barrio-Hernandez *et al.*, Clustering predicted structures at the scale of the known protein universe. bioRxiv **2023-03** (2023).
5. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, ppi-py: A Python package for scientific discovery using machine learning, Zenodo (2023); https://doi.org/10.5281/zenodo.8403931.
6. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, Prediction-Powered Inference: Data Sets, Zenodo (2023); https://doi.org/10.5281/zenodo.8397451.
7. L. M. Iakoucheva *et al.*, *Nucleic Acids Res.* **32**, 1037–1049 (2004).
8. UniProt Consortium, *Nucleic Acids Res.* **43**, D204–D212 (2015).
9. K. W. Willett *et al.*, *Mon. Not. R. Astron. Soc.* **435**, 2835–2860 (2013).
10. D. G. York *et al.*, *Astron. J.* **120**, 1579–1587 (2000).
11. E. D. Vaishnav *et al.*, *Nature* **603**, 455–463 (2022).
12. E. L. Bullock, C. E. Woodcock, C. Souza Jr., P. Olofsson, *Glob. Chang. Biol.* **26**, 2956–2969 (2020).
13. J. O. Sexton *et al.*, *Int. J. Digit. Earth* **6**, 427–448 (2013).
14. F. Ding, M. Hardt, J. Miller, L. Schmidt, in *Advances in Neural Information Processing Systems* **34** (2021), pp. 6478–6490.
15. T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

16. R. J. Olson, A. Shalapyonok, H. M. Sosik, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **50**, 301–315 (2003).
17. E. C. Orenstein, O. Beijbom, E. E. Peacock, H. M. Sosik, WHOI-Plankton- A large scale fine grained visual recognition benchmark dataset for plankton classification. arXiv:1510.00745 [cs.CV] (2015).
18. S. Wang, T. H. McCormick, J. T. Leek, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30266–30275 (2020).
19. M. S. Pepe, *Biometrika* **79**, 355–365 (1992).
20. J. Lafferty, L. Wasserman, in *Advances in Neural Information Processing Systems* **20** (2007), pp. 801–808.
21. A. Zhang, L. D. Brown, T. T. Cai, *Ann. Stat.* **47**, 2538–2566 (2019).
22. A. Chakrabortty, G. Dai, E. Tchetgen Tchetgen, A general framework for treatment effect estimation in semi-supervised and high dimensional settings. arXiv:2201.00468 [stat.ME] (2022).
23. A. Chakrabortty, T. Cai, *Ann. Stat.* **46**, 1541–1572 (2018).
24. Y. Zhang, J. Bradic, *Biometrika* **109**, 387–403 (2022).
25. S. Deng, Y. Ning, J. Zhao, H. Zhang, Optimal and safe estimation for high-dimensional semi-supervised learning. arXiv:2011.14185 [stat.ME] (2020).
26. D. Azriel *et al.*, *J. Am. Stat. Assoc.* **117**, 2238–2251 (2022).
27. A. Chakrabortty, G. Dai, R. J. Carroll, Semi-supervised quantile estimation: robust and efficient inference in high dimensional settings. arXiv:2201.10208 [stat.ME] (2022).
28. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World* (Springer, 2005), vol. 5.

ACKNOWLEDGMENTS

We thank A. Kohli for suggesting the term "rectifier," E. Orenstein for helpful discussions related to the WHO-I Plankton dataset, and S. Wang for help with the remote-sensing experiment. **Funding:** Office of Naval Research grant N00014-21-1-2840 (M.I.J.); National Science Foundation Graduate Research Fellowship (A.N.A., C.F.). **Author contributions:** Conceptualization: A.N.A., S.B., C.F., M.I.J., T.Z.; Methodology: A.N.A., S.B., C.F., M.I.J., T.Z.; Formal analysis: A.N.A., S.B., C.F., M.I.J., T.Z.; Funding acquisition: M.I.J.; Supervision: M.I.J.; Writing – original draft: A.N.A., S.B., C.F., M.I.J., T.Z.; Writing – review and editing: A.N.A., S.B., C.F., M.I.J., T.Z. **Competing interests:** None declared. **Data and materials availability:** The data and code are available in the accompanying Python package (*5*) and data repository (*6*). All other data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials.

SUPPLEMENTARY MATERIAL

science.org/doi/10.1126/science.adi6000
Supplementary Text
Figs. S1 to S6
References (*29*–*56*)

Submitted 7 May 2023; accepted 5 October 2023
10.1126/science.adi6000

Supplementary Materials for

Prediction-powered inference

Anastasios N. Angelopoulos *et al*.

Corresponding authors: Anastasios N. Angelopoulos, angelopoulos@berkeley.edu; Stephen Bates, stephenbates@berkeley.edu; Clara Fannjiang, clarafy@berkeley.edu; Michael I. Jordan, michael_jordan@berkeley.edu; Tijana Zrnic, tijana.zrnic@berkeley.edu

Science **382**, 669 (2023)
DOI: 10.1126/science.adi6000

The PDF file includes:

Supplementary Text
Figs. S1 to S6
References

Comparison to Baseline Procedures

The prediction-powered inference procedure was compared to three baseline procedures that also combine labeled and unlabeled data in performing statistical inference. The baselines were:

1. **Post-prediction inference.** The post-prediction inference procedure of Wang et al. (*18*) was used for estimating ordinary least-squares (OLS) coefficients. The procedure first fits a regression r to predict Y from $\hat{Y}$ on the gold-standard dataset. Subsequently, the regression function is used to correct the imputed labels on the unlabeled dataset. Confidence intervals are formed using the $r(\hat{Y}')$ as if they were gold-standard data. This procedure has no theoretical guarantees in general and requires strong distributional assumptions on the relationship between Y and $\hat{Y}$ to provide coverage. Our experiments indicated that this approach fails to cover in realistic conditions.

2. **Semi-supervised mean estimation.** The semi-supervised mean estimation procedure of Zhang and Bradic (*24*) involves cross-fitting a (possibly-regularized) linear model on K distinct folds of the gold-standard dataset. The average of the K model predictions on each unlabeled data point is taken as its corresponding $\hat{Y}'$, and the average bias $\hat{Y} - Y$ of the K models is also computed and used to debias the resulting mean estimate. The formal validity of this approach applies to mean estimation and requires the cross-fitting of linear models; it does not have formal guarantees for more flexible model classes. For this reason, it provided little improvement over the classical confidence interval in our experiments, since the variance reduction possible with linear models is typically limited.

3. **Conformal prediction for mean estimation.** A version of conformal prediction (*28*) was used to construct prediction intervals, which are then ensembled into a mean estimate. The procedure involves using the gold-standard data to construct conformal prediction sets with the residual score function at level α/N for each unlabeled example. The lower- and upper- endpoints of these sets were averaged to produce a confidence interval for the mean. This confidence interval is guaranteed validity for arbitrary models and distributions, unlike the other baseline approaches. However, it is extremely conservative: it output infinite intervals in the experiments. An ablation is performed without a Bonferroni correction (i.e., sets were constructed at level α instead of α/N), but this remained conservative and did not provide an improvement over the classical intervals.

Experimental Protocol

The methods were evaluated on an income prediction task on the same census dataset used for the logistic regression experiments in the main text. In the case of the semi-supervised and conformal baselines, the goal was to estimate the mean income in California in the year 2019 among employed individuals using a small amount of labeled data and a large amount of covariates. In the case of the post-prediction inference baseline, the target of inference was the OLS coefficient between age and income. The setup was the same as the logistic regression experiment described in the main text (including the use of the Folktables (*14*) interface and the gradient-boosted tree as the predictor).

Comparison to Post-Prediction Inference

Results of the post-prediction inference protocol as compared to the classical and prediction-powered approaches are shown in Fig. S1 for the previously-described OLS coefficient between age and income. The procedure did not cover at the proper rate and the intervals were biased.

Comparison to Semi-Supervised Mean Estimation

Results of the semi-supervised mean estimation protocol as compared to the classical and prediction-powered approaches are shown in Fig. S2 for the previously described mean income estimation task. The prediction-powered intervals dominated both the semi-supervised intervals and the classical ones in the experiment for all values of n.

Comparison to Conformal Prediction for Mean Estimation

Results of the conformal mean estimation protocol, with and without a Bonferroni correction, were compared to the prediction-powered approach for the previously described mean income estimation task. The prediction-powered

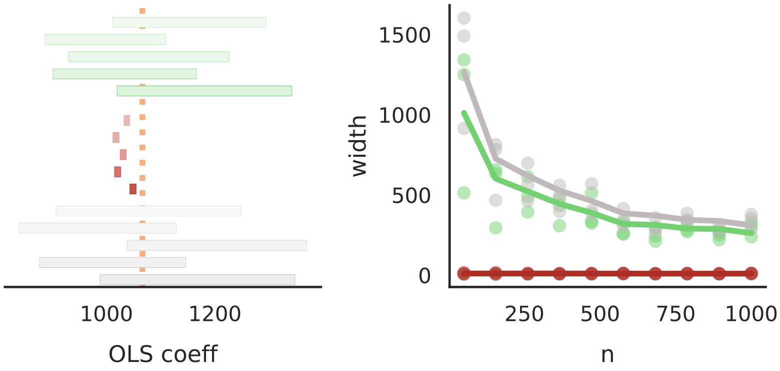

Figure S1: **Comparison to the post-prediction inference procedure.** On the left are five independent random draws of intervals with $n = 1000$. On the right is a line plot of interval width as a function of n, averaged over 100 independent trials. Five draws of interval widths are shown as a scatter plot at their respective n. The post-prediction inference approach is shown in red, the classical approach is in gray, and the prediction-powered approach is in green. The post-prediction inference approach had diminishing coverage in the experiment.

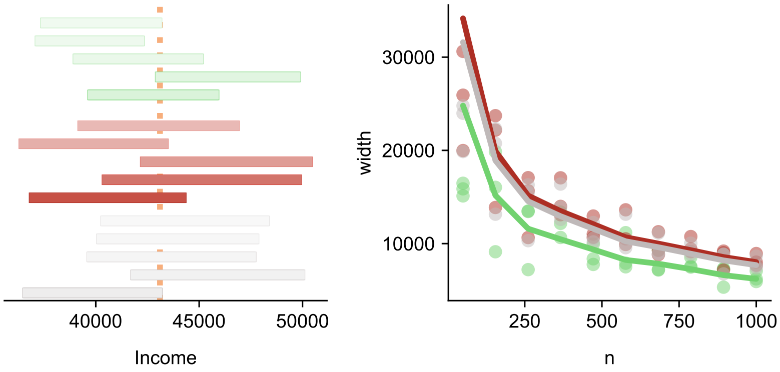

Figure S2: **Comparison to the semi-supervised mean estimation procedure.** The plot is the same as in Fig. S1, but with semi-supervised inference shown in red. The semi-supervised intervals had a similar width to the classical ones in this experiment, while the prediction-powered intervals dominated.

intervals dominated the conformal intervals for all values of n. The conformal intervals with Bonferroni correction were infinite in size and thus could not be plotted. Without the Bonferroni correction, though the method is statistically invalid, it remained quite conservative in the experiment, as can be seen by examining Fig. S3.

Cases Where Prediction-Powered Inference is Underpowered

Since standard confidence intervals scale with the standard error of the estimator, prediction-powered inference is powerful when a machine-learning model can provide a reduction in the estimator variance. At a high level, this happens when N is large enough relative to n and the model is accurate enough. This was the case in all the experiments shown in the main text. This section precisely quantifies what it means to have an accurate enough model and large enough N. Corroborating the theory, two cases where classical inference outperforms prediction-powered inference are presented: one where the model is not good enough and another where N is too small.

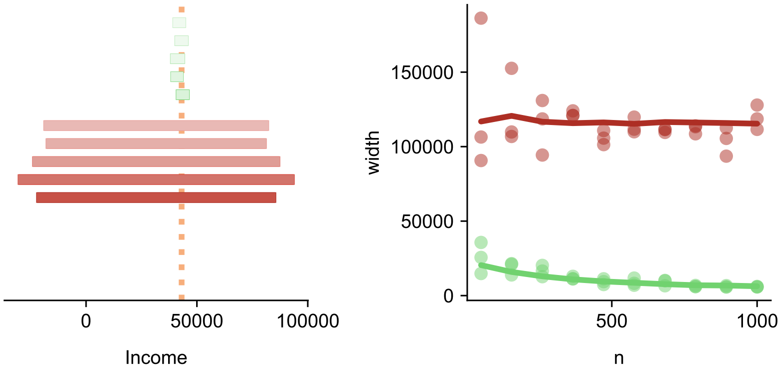

Figure S3: **Comparison to conformal prediction for mean estimation.** The plot is the same as in Fig. S1, but with conformal prediction (without a Bonferroni correction) shown in red. The intervals produced by conformal prediction with a Bonferroni correction were infinite and thus could not be plotted. Even the conformal intervals without a Bonferroni correction were wide compared to the prediction-powered intervals.

Mathematical Derivation

Consider the case of mean estimation, $\theta^* = \mathbb{E}[Y_i]$. The classical estimate of θ^* is the sample average of the outcomes on the labeled dataset,

$$\hat{\theta}^{\text{class}} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Prediction-powered inference, on the other hand, computes the estimate

$$\hat{\theta}^{\text{PP}} = \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i' - \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i).$$

Notice that both $\hat{\theta}^{\text{class}}$ and $\hat{\theta}^{\text{PP}}$ are unbiased, seeing that $\mathbb{E}[\hat{Y}_i'] = \mathbb{E}[\hat{Y}_i]$.

The widths of the classical confidence interval based on the central limit theorem and the prediction-powered confidence interval based on Theorem S1 scale with $\text{Var}(\hat{\theta}^{\text{class}})$ and $\text{Var}(\hat{\theta}^{\text{PP}})$, respectively. The classical estimator has variance equal to

$$\text{Var}(\hat{\theta}^{\text{class}}) = \frac{1}{n}\text{Var}(Y_i).$$

The variance of the prediction-powered estimator equals

$$\text{Var}(\hat{\theta}^{\text{PP}}) = \frac{1}{N}\text{Var}(\hat{Y}_i') + \frac{1}{n}\text{Var}(\hat{Y}_i - Y_i),$$

where independence of the two terms in the estimator is applied. Therefore, the prediction-powered confidence interval will be tighter when

$$\frac{1}{N}\text{Var}(\hat{Y}_i') + \frac{1}{n}\text{Var}(\hat{Y}_i - Y_i) < \frac{1}{n}\text{Var}(Y_i).$$

Since the predictions $\hat{Y}_i'$ will typically have a variance that is of the same order as the variance of Y_i, if $N \approx n$ one should not expect prediction-powered inference to help. Gains are expected when $N \gg n$. In that case, $\frac{1}{N}\text{Var}(\hat{Y}_i') \ll \frac{1}{n}\text{Var}(\hat{Y}_i - Y_i)$, and thus prediction-powered inference helps when

$$\text{Var}(\hat{Y}_i - Y_i) < \text{Var}(Y_i).$$

In other words, prediction-powered inference gives tighter confidence intervals when the predictions explain away some of the outcome variance.

To gain further intuition, suppose that the outcomes are binary, $Y_i \sim \text{Bern}(p)$, where $\text{Bern}(p)$ denotes the Bernoulli distribution with parameter p. In this case, $\theta^* = p$. For simplicity, suppose that $P(\hat{Y}_i = 0 | Y_i = 1) = P(\hat{Y}_i = 1 | Y_i = 0) = \eta$. Then, a direct variance calculation gives $\text{Var}(\hat{Y}_i - Y_i) = \eta - \eta^2(1 - 2p)^2$ and $\text{Var}(Y_i) = p(1 - p)$. This allows for a direct comparison of the variances in terms of the outcome bias p and model error η. For example, when $p = 0.5$, the model error η has to be smaller than 25% for prediction-powered inference to help; when $p = 0.1$, meaning the outcomes themselves have low variance, the model error η has to be smaller than about 9.5%. In general, the lower the variance of the outcome, the lower the model error has to be for prediction-powered inference to be helpful.

Putting everything together, the main takeaway is as follows: prediction-powered inference should only be applied when N is (preferably substantially) larger than n, and when the model has a high enough predictive accuracy to explain away some of the outcome variance. While this derivation focused on mean estimation, a similar intuition holds for other estimation problems.

Inaccurate Machine-Learning Model

The deforestation analysis experiment from the main text was repeated. However, instead of a gradient-boosted tree, a linear regression module was used as the machine-learning model. This led to a substantial enough reduction in performance that the classical baseline outperformed the prediction-powered approach. See Fig. S4 for the results. Due to the reduction of power, for the same null hypothesis as in the main text, the prediction-powered approach required $n = 40$ data points to reject, while the classical baseline required $n = 35$.

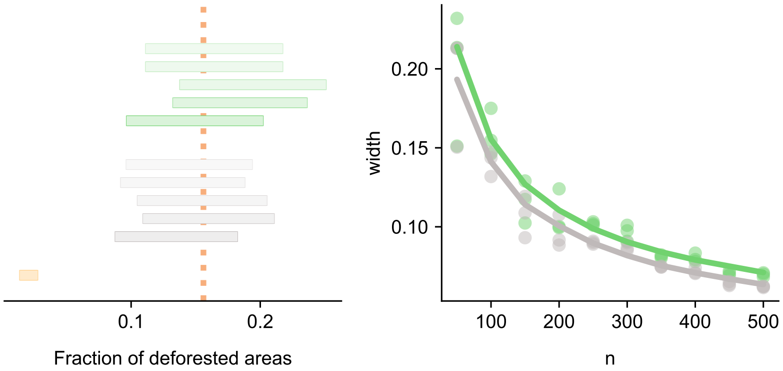

Figure S4: **Deforestation analysis with a linear model.** This is the same figure as Fig. 2D, with the same color coding; the prediction-powered approach is green, the classical approach is gray, and the imputation approach is gold. However, here the gradient-boosted tree was replaced with an ordinary linear regression. The drop in performance causes the classical intervals to outperform the prediction-powered intervals in terms of power.

Unlabeled Data Set is Too Small

The AlphaFold-based proteomic analysis from the main text was repeated. However, $N = 1000$ data points were randomly chosen as the unlabeled dataset. The rest of the procedure is performed exactly the same way as described in the main text. The decrease in sample size led to a reduction in power, and in the regime $n > N$, the classical baseline outperformed the prediction-powered approach. See Fig. S5 for the results. For the same null hypothesis as in the main text, the prediction-powered approach required $n = 869$ data points to reject, while the classical baseline required $n = 652$.

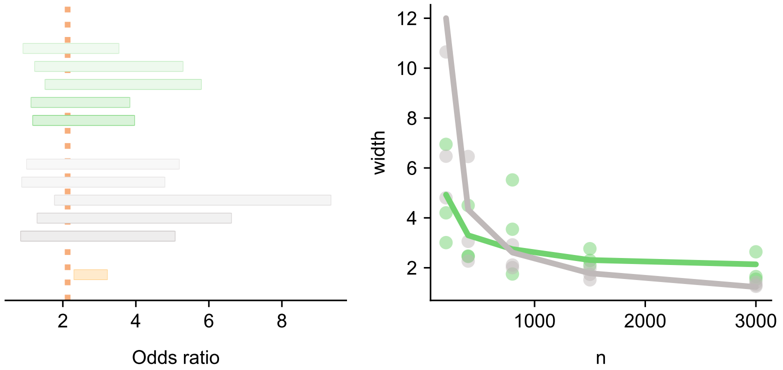

Figure S5: **AlphaFold analysis with a small unlabeled dataset.** This is the same figure as Fig. 2A, with the same color coding; the prediction-powered approach is green, the classical approach is gray, and the imputation approach is gold. However, here N was taken to be 1000. It can be seen that, when $n > N$, the classical baseline outperforms the prediction-powered one.

Validity of Prediction-Powered Inference

Our main contribution is a technique for inference on estimands that can be expressed as the solution to a *convex optimization problem*. Formally, estimands of the following form are considered

$$\theta^* = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} \ \mathbb{E}\left[L_\theta(X_i, Y_i)\right],$$

for a loss function $L_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that is convex in $\theta \in \mathbb{R}^p$, for some $p \in \mathbb{N}$. This paradigm captures problems such as mean estimation, median and general quantile estimation, estimation of linear and logistic regression coefficients, among others.

In the following, the term ∇L_θ denotes a subgradient of L_θ with respect to θ. Under mild conditions, convexity ensures that θ^* can also be expressed as the value solving

$$\mathbb{E}\left[\nabla L_{\theta^*}(X_i, Y_i)\right] = 0. \tag{S1}$$

Henceforth, convex estimation problems where θ^* satisfies (S1) will be called nondegenerate, and mild conditions that ensure this regularity will be later discussed.

The *measure of fit* on the imputed data will be defined as:

$$m_\theta = \frac{1}{N}\sum_{i=1}^{N} \nabla L_\theta(X_i', \hat{Y}_i').$$

Next, the *rectifier* is defined as the difference between the measure of fit computed on the labeled data, (X, Y), and the labeled data when the true outcomes are replaced with predicted ones, $(X, \hat{Y})$:

$$\boldsymbol{\Delta}_\theta = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla L_\theta(X_i, Y_i) - \nabla L_\theta(X_i, \hat{Y}_i)\right).$$

Further denoting by $\nabla L_{\theta,j}(x, y)$ the j-th coordinate of $\nabla L_\theta(x, y)$, the corresponding standard deviations are denoted as:

$$\hat{\sigma}^2_{\Delta_\theta, j} = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla L_{\theta,j}(X_i, Y_i) - \nabla L_{\theta,j}(X_i, \hat{Y}_i) - \boldsymbol{\Delta}_{\theta,j}\right)^2; \quad \hat{\sigma}^2_{m_\theta, j} = \frac{1}{N}\sum_{i=1}^{N}\left(\nabla L_{\theta,j}(X_i', \hat{Y}_i') - m_{\theta,j}\right)^2,$$

for $j \in [p]$.

Throughout, z_q denotes the q-quantile of the standard normal distribution.

The main mathematical result of this work is stated in Theorem S1.

Theorem S1 (Validity of prediction-powered inference)**.** *Let the labeled and unlabeled data be sampled i.i.d.. Suppose that the convex estimation problem is nondegenerate as in* (S1) *and that $\frac{n}{N} \to p$, for some $p \in (0,1)$. Fix $\alpha \in (0,1)$ and let*

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta : |m_\theta + \mathbf{\Delta}_\theta| \leq w_\theta(\alpha)\right\}, \tag{S2}$$

where $w_\theta(\alpha) \in \mathbb{R}^p$ has j-th coordinate equal to $w_{\theta,j}(\alpha) = z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_{\mathbf{\Delta}_\theta,j}}{n} + \frac{\hat{\sigma}^2_{m_\theta,j}}{N}}$, and the inequality in (S2) *is applied coordinatewise. Then,*

$$\liminf_{n,N\to\infty} P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1-\alpha.$$

Proof. The proof will show that $\theta^* \notin \mathcal{C}_\alpha^{\mathrm{PP}}$ with probability at most α in the limit; that is,

$$\limsup_{n,N\to\infty} P\left(|\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j}| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_{\mathbf{\Delta}_{\theta^*},j}}{n} + \frac{\hat{\sigma}^2_{m_{\theta^*},j}}{N}},\ \forall j \in [p]\right) \leq \alpha.$$

For each $j \in [p]$, the central limit theorem implies that

$$\sqrt{n}(\mathbf{\Delta}_{\theta^*,j} - \mathbb{E}[\mathbf{\Delta}_{\theta^*,j}]) \Rightarrow \mathcal{N}(0, \sigma^2_{\mathbf{\Delta}_{\theta^*},j}); \quad \sqrt{N}(m_{\theta^*,j} - \mathbb{E}[m_{\theta^*,j}]) \Rightarrow \mathcal{N}(0, \sigma^2_{m_{\theta^*},j}),$$

where $\sigma^2_{\mathbf{\Delta}_{\theta^*},j}$ is the variance of $\nabla L_{\theta^*,j}(X_i, Y_i) - \nabla L_{\theta^*,j}(X_i, \hat{Y}_i)$ and $\sigma^2_{m_{\theta^*},j}$ is the variance of $\nabla L_{\theta^*,j}(X_i, \hat{Y}_i)$. Therefore, by Slutsky's theorem,

$$\begin{aligned}\sqrt{N}\left(\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j} - \mathbb{E}[\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j}]\right) &= \sqrt{n}\left(\mathbf{\Delta}_{\theta^*,j} - \mathbb{E}[\mathbf{\Delta}_{\theta^*,j}]\right)\sqrt{\frac{N}{n}} + \sqrt{N}\left(m_{\theta^*,j} - \mathbb{E}[m_{\theta^*,j}]\right)\\ &\Rightarrow \mathcal{N}\left(0, \frac{1}{p}\sigma^2_{\mathbf{\Delta}_{\theta^*},j} + \sigma^2_{m_{\theta^*},j}\right).\end{aligned}$$

This in turn implies

$$\limsup_{n,N\to\infty} P\left(|\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j} - \mathbb{E}\left[\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j}\right]| > z_{1-\alpha/(2p)}\frac{\hat{\sigma}_j}{\sqrt{N}}\right) \leq \frac{\alpha}{p},$$

where $\hat{\sigma}_j^2$ is a consistent estimate of the variance $\frac{1}{p}\sigma^2_{\mathbf{\Delta}_{\theta^*},j} + \sigma^2_{m_{\theta^*},j}$. Define $\hat{\sigma}_j^2 = \hat{\sigma}^2_{\mathbf{\Delta}_{\theta^*},j}\frac{N}{n} + \hat{\sigma}^2_{m_{\theta^*},j}$; this estimate is consistent because the two terms are individually consistent estimates of the respective variances. Now notice that

$$\mathbb{E}\left[\mathbf{\Delta}_{\theta^*} + m_{\theta^*}\right] = \mathbb{E}\left[(\nabla L_{\theta^*}(X_i, Y_i) - \nabla L_{\theta^*}(X_i, \hat{Y}_i)) + \nabla L_{\theta^*}(X_i', \hat{Y}_i')\right] = \mathbb{E}[\nabla L_{\theta^*}(X_i, Y_i)] = 0,$$

where the last step follows by the nondegeneracy condition. Putting together the previous two displays and the choice of $\hat{\sigma}_j$ derived above, and applying a union bound yields

$$\begin{aligned}&\limsup_{n,N\to\infty} P\left(\exists j \in [p] : |\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j}| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_{\mathbf{\Delta}_{\theta^*},j}}{n} + \frac{\hat{\sigma}^2_{m_{\theta^*},j}}{N}}\right)\\ &\leq \sum_{j=1}^{p}\limsup_{n,N\to\infty} P\left(|\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j}| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_{\mathbf{\Delta}_{\theta^*},j}}{n} + \frac{\hat{\sigma}^2_{m_{\theta^*},j}}{N}}\right)\\ &= \sum_{j=1}^{p}\limsup_{n,N\to\infty} P\left(|\mathbf{\Delta}_{\theta^*,j} + m_{\theta^*,j}| > z_{1-\alpha/(2p)}\hat{\sigma}_j\right)\\ &\leq \sum_{j=1}^{p}\frac{\alpha}{p}\\ &= \alpha.\end{aligned}$$

Therefore, $\limsup_{n,N\to\infty} P(\theta^* \notin \mathcal{C}_\alpha^{\mathrm{PP}}) \leq \alpha$. □

Most practical problems are nondegenerate (S1). For example, if the loss is differentiable for all $\theta \in \mathbb{R}^p$, then the problem is immediately nondegenerate. Furthermore, if the data distribution does not have point masses and, for every θ, $L_\theta(x, y)$ is nondifferentiable only for a measure-zero set of (x, y) pairs, then the problem is again nondegenerate.

Algorithms

Prediction-Powered Confidence Intervals

The algorithms previously referenced in Table 1 are presented. The algorithms are derived by instantiating Theorem S1 with the appropriate loss function dependent on the estimand. In addition, guarantees of their formal validity are stated.

Algorithm S1 Prediction-powered mean estimation

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, error level $\alpha \in (0, 1)$
$\hat{\theta}^{\mathrm{PP}} \leftarrow \hat{\theta} - \boldsymbol{\Delta} := \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i' - \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)$
$\hat{\sigma}^2_{\hat{Y}'} \leftarrow \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i' - \hat{\theta})^2$
$\hat{\sigma}^2_{\hat{Y}-Y} \leftarrow \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i - \boldsymbol{\Delta})^2$
$w(\alpha) \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{\hat{Y}-Y}}{n} + \frac{\hat{\sigma}^2_{\hat{Y}'}}{N}}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\theta}^{\mathrm{PP}} \pm w(\alpha)\right)$
```

Algorithm S2 Prediction-powered median estimation

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, error level $\alpha \in (0, 1)$
Construct fine grid $\Theta_{\mathrm{grid}}$ between $\min_{i\in[N]} \hat{Y}_i'$ and $\max_{i\in[N]} \hat{Y}_i'$
for $\theta \in \Theta_{\mathrm{grid}}$ do
    $\boldsymbol{\Delta}_\theta \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\left\{Y_i \le \theta\right\} - \mathbb{1}\left\{\hat{Y}_i \le \theta\right\}\right)$
    $m_\theta \leftarrow \frac{1}{2N}\sum_{i=1}^{N} \mathrm{sign}\left(\theta - \hat{Y}_i'\right)$
    $\hat{\sigma}^2_{\boldsymbol{\Delta}_\theta} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\left\{Y_i \le \theta\right\} - \mathbb{1}\left\{\hat{Y}_i \le \theta\right\} - \boldsymbol{\Delta}_\theta\right)^2$
    $\hat{\sigma}^2_{m_\theta} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}\left\{\hat{Y}_i' \le \theta\right\} - m_\theta\right)^2$
    $w_\theta(\alpha) \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{\boldsymbol{\Delta}_\theta}}{n} + \frac{\hat{\sigma}^2_{m_\theta}}{N}}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta \in \Theta_{\mathrm{grid}} : |m_\theta + \boldsymbol{\Delta}_\theta| \le w_\theta(\alpha)\}$
```

Algorithm S3 Prediction-powered quantile estimation

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, quantile $q$, error level $\alpha \in (0, 1)$
Construct fine grid $\Theta_{\mathrm{grid}}$ between $\min_{i\in[N]} \hat{Y}_i'$ and $\max_{i\in[N]} \hat{Y}_i'$
for $\theta \in \Theta_{\mathrm{grid}}$ do
    $\boldsymbol{\Delta}_\theta \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\left\{Y_i \le \theta\right\} - \mathbb{1}\left\{\hat{Y}_i \le \theta\right\}\right)$
    $\hat{F}_\theta \leftarrow \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left\{\hat{Y}_i' \le \theta\right\}$
    $\hat{\sigma}^2_{\boldsymbol{\Delta}_\theta} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\left\{Y_i \le \theta\right\} - \mathbb{1}\left\{\hat{Y}_i \le \theta\right\} - \boldsymbol{\Delta}_\theta\right)^2$
    $\hat{\sigma}^2_{\hat{Y}',\theta} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}\left\{\hat{Y}_i' \le \theta\right\} - \hat{F}_\theta\right)^2$
    $w_\theta(\alpha) \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{\boldsymbol{\Delta}_\theta}}{n} + \frac{\hat{\sigma}^2_{\hat{Y}',\theta}}{N}}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta \in \Theta_{\mathrm{grid}} : |\hat{F}_\theta + \boldsymbol{\Delta}_\theta - q| \le w_\theta(\alpha)\right\}$
```

Algorithm S4 Prediction-powered linear regression

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, coefficient $j^*$, error level $\alpha \in (0,1)$
  $\hat{\theta}^{\mathrm{PP}} \leftarrow \hat{\theta} - \mathbf{\Delta} := X'^{\dagger}\hat{Y}' - X^{\dagger}(f - Y)$
  $\Sigma' \leftarrow \frac{1}{N}(X')^\top X'$, $M' \leftarrow \frac{1}{N}\sum_{i=1}^N (\hat{Y}_i' - (X_i')^\top \hat{\theta})^2 X_i'(X_i')^\top$
  $V' \leftarrow (\Sigma')^{-1} M' (\Sigma')^{-1}$
  $\Sigma \leftarrow \frac{1}{n} X^\top X$, $M \leftarrow \frac{1}{n}\sum_{i=1}^n (\hat{Y}_i - Y_i - X_i^\top \mathbf{\Delta})^2 X_i X_i^\top$
  $V \leftarrow \Sigma^{-1} M \Sigma^{-1}$
  $w(\alpha) \leftarrow z_{1-\alpha/2}\sqrt{\frac{V_{j^*j^*}}{n} + \frac{V'_{j^*j^*}}{N}}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\theta}_{j^*}^{\mathrm{PP}} \pm w(\alpha)\right)$
```

Algorithm S5 Prediction-powered logistic regression

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, error level $\alpha \in (0,1)$
  Construct fine grid $\Theta_{\mathrm{grid}} \subset \mathbb{R}^d$ of possible coefficients
  $\mathbf{\Delta}_j \leftarrow \frac{1}{n}\sum_{i=1}^n X_{i,j}(\hat{Y}_i - Y_i), \quad j \in [d]$
  $\hat{\sigma}^2_{\mathbf{\Delta},j} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(X_{i,j}(\hat{Y}_i - Y_i) - \mathbf{\Delta}_j\right)^2, \quad j \in [d]$
  for $\theta \in \Theta_{\mathrm{grid}}$ do
      $m_{\theta,j} \leftarrow \frac{1}{N}\sum_{i=1}^N X'_{i,j}\left(\mu_\theta(X_i') - \hat{Y}_i'\right), \quad j \in [d], \quad$ where $\mu_\theta(x) = \frac{1}{1+\exp(-x^\top\theta)}$
      $\hat{\sigma}^2_{m_\theta,j} \leftarrow \frac{1}{N}\sum_{i=1}^N \left(X'_{i,j}(\mu_\theta(X_i') - \hat{Y}_i') - m_{\theta,j}\right)^2, \quad j \in [d]$
      $w_{\theta,j}(\alpha) \leftarrow z_{1-\alpha/(2d)}\sqrt{\frac{\hat{\sigma}^2_{\mathbf{\Delta},j}}{n} + \frac{\hat{\sigma}^2_{m_\theta,j}}{N}}, \quad j \in [d]$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta \in \Theta_{\mathrm{grid}} : |m_{\theta,j} + \mathbf{\Delta}_j| \le w_{\theta,j}(\alpha), \forall j \in [d]\}$
```

Algorithm S6 Prediction-powered convex risk minimization

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, error level $\alpha \in (0,1)$ , gradient function
    $\nabla L_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^p$
  Construct fine grid $\Theta_{\mathrm{grid}} \subset \mathbb{R}^p$ of possible solutions
  for $\theta \in \Theta_{\mathrm{grid}}$ do
      $\mathbf{\Delta}_{\theta,j} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\nabla L_{\theta,j}(X_i, Y_i) - \nabla L_{\theta,j}(X_i, \hat{Y}_i)\right), \quad j \in [p]$
      $\hat{\sigma}^2_{\mathbf{\Delta}_\theta,j} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\nabla L_{\theta,j}(X_i, Y_i) - \nabla L_{\theta,j}(X_i, \hat{Y}_i) - \mathbf{\Delta}_{\theta,j}\right)^2, \quad j \in [p]$
      $m_{\theta,j} \leftarrow \frac{1}{N}\sum_{i=1}^N \nabla L_{\theta,j}(X_i', \hat{Y}_i'), \quad j \in [p]$
      $\hat{\sigma}^2_{m_\theta,j} \leftarrow \frac{1}{N}\sum_{i=1}^N \left(\nabla L_{\theta,j}(X_i', \hat{Y}_i') - m_{\theta,j}\right)^2, \quad j \in [p]$
      $w_{\theta,j}(\alpha) \leftarrow z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_{\mathbf{\Delta}_\theta,j}}{n} + \frac{\hat{\sigma}^2_{m_\theta,j}}{N}}, \quad j \in [p]$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta \in \Theta_{\mathrm{grid}} : |m_{\theta,j} + \mathbf{\Delta}_{\theta,j}| \le w_{\theta,j}(\alpha), \forall j \in [p]\}$
```

Regularity conditions. All algorithms stated in this section rely on confidence intervals derived from the central limit theorem. For such intervals to be asymptotically valid, it is required that the two quantities whose mean is being estimated, namely $\nabla L_\theta(X_i, Y_i) - \nabla L_\theta(X_i, \hat{Y}_i)$ and $\nabla L_\theta(X_i, \hat{Y}_i)$, have at least the first two moments (see Proposition S2).

For Corollary S3 to hold, the same conditions are required as those needed for classical linear regression intervals to cover the target. These conditions are very weak; in particular, it is *not* required that the true data-generating process be linear or the errors be homoskedastic. See Buja et al. (*29*) for a detailed discussion. The following are the required conditions, as stated in Theorem 3 of Halbert White's seminal paper (*30*). The data $(X_1, Y_1), \ldots, (X_n, Y_n)$ is generated as $X_i = h(Z_i)$, $Y_i = g(Z_i) + \epsilon_i$, where (Z_i, ϵ_i) are mean-zero i.i.d. random draws from some distribution such that $\mathbb{E}[Z_i Z_i^\top]$ and $\mathbb{E}[X_i X_i^\top]$ are finite and nonsingular, and $\mathbb{E}[\epsilon_i^2]$, $\mathbb{E}[Y_i^2 X_i X_i^\top]$, and $\mathbb{E}[X_{ij}^2 X_i X_i^\top]$ are all finite. In addition, it is assumed that h and g are measurable. Under these conditions,

$$\sqrt{n}(\hat{\theta}_{\mathrm{OLS}} - \theta^*) \Rightarrow \mathcal{N}(0, \Sigma^{-1} V \Sigma^{-1}),$$

where $\theta^* = \arg\min_\theta \mathbb{E}[(Y_1 - X_1^\top \theta)^2]$, $\hat{\theta}_{\mathrm{OLS}} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top \theta)^2$, $\Sigma = \mathbb{E}[X_1 X_1^\top]$, $V = \mathbb{E}[(Y_1 - X_1^\top \theta^*)^2 X_1 X_1^\top]$. Moreover, $\frac{1}{n} X^\top X \to \Sigma$ and $\frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top \hat{\theta}_{\mathrm{OLS}})^2 X_i X_i^\top \to V$ almost surely.

Corollary S1 (Mean estimation)**.** *Let θ^* be the mean outcome:*

$$\theta^* = \mathbb{E}[Y_i].$$

Then, the prediction-powered confidence interval in Algorithm S1 is valid: $\liminf_{n,N\to\infty} P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha$.

Proof. The prediction-powered confidence set constructed in Algorithm S1 is a special case of the prediction-powered confidence set constructed in Theorem S1. The proof then follows directly by the guarantee of Theorem S1.

Since $\nabla L_\theta(y) = \theta - y$, one has

$$\mathbf{\Delta}_\theta \equiv \mathbf{\Delta} = \frac{1}{n}\sum_{i=1}^n (\hat{Y}_i - Y_i); \quad m_\theta = \theta - \frac{1}{N}\sum_{i=1}^N \hat{Y}_i'.$$

Therefore, the set $\mathcal{C}_\alpha^{\mathrm{PP}}$ from Theorem S1 can be written as

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta : \left|\theta - \frac{1}{N}\sum_{i=1}^N \hat{Y}_i' + \frac{1}{n}\sum_{i=1}^n (\hat{Y}_i - Y_i)\right| \leq w_\theta(\alpha)\right\} = \left(\frac{1}{N}\sum_{i=1}^N \hat{Y}_i' - \frac{1}{n}\sum_{i=1}^n (\hat{Y}_i - Y_i) \pm w_\theta(\alpha)\right).$$

This is exactly the set constructed in Algorithm S1, which completes the proof. □

The median algorithm (Algorithm S2) is a special case of the general quantile algorithm (Algorithm S3), obtained by setting $q = 0.5$. Therefore, we only state the guarantees for Algorithm S3.

Corollary S2 (Quantile estimation)**.** *Let θ^* be the q-quantile:*

$$\theta^* = \min\{\theta : P(Y_i \leq \theta) \geq q\}.$$

Then, the prediction-powered confidence set in Algorithm S3 is valid: $\liminf_{n,N\to\infty} P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha$.

Proof. Like in the proof of Corollary S1, we proceed by showing that the prediction-powered confidence set constructed in Algorithm S3 is a special case of the prediction-powered confidence set constructed in Theorem S1. Then, we simply invoke Theorem S1.

Since $\nabla L_\theta(y) = -q + \mathbb{1}\{y \leq \theta\}$, we have

$$\mathbf{\Delta}_\theta = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{Y_i \leq \theta\} - \mathbb{1}\left\{\hat{Y}_i \leq \theta\right\}\right); \quad m_\theta = -q + \hat{F}_\theta,$$

where $\hat{F}_\theta = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\left\{\hat{Y}_i' \leq \theta\right\}$. Therefore, the set $\mathcal{C}_\alpha^{\mathrm{PP}}$ from Theorem S1 can be written as

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta : \left|\frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{Y_i \leq \theta\} - \mathbb{1}\left\{\hat{Y}_i \leq \theta\right\}\right) - q + \hat{F}_\theta\right| \leq w_\theta(\alpha)\right\} = \left\{\theta : \left|\hat{F}_\theta + \mathbf{\Delta}_\theta - q\right| \leq w_\theta(\alpha)\right\}.$$

This is exactly the set constructed in Algorithm S3. Therefore, the guarantee of Corollary S2 follows by the guarantee of Theorem S1. □

Corollary S3 (Linear regression)**.** *Fix $j^* \in [d]$. Let θ^* be the linear regression solution:*

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(Y_i - X_i^\top \theta)^2].$$

Then, the prediction-powered confidence interval in Algorithm S4 is valid: $\liminf_{n,N\to\infty} P\left(\theta^*_{j^*} \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha$.

Proof. For linear regression, we can derive more powerful prediction-powered confidence intervals than those implied by Theorem S1 by exploiting the linearity of the least-squares estimator.

As in Theorem S1, we assume that $\frac{n}{N} \to p$, for some fraction $p \in (0,1)$.

Theorem 3 of White (*31*) implies that

$$\sqrt{n}(\mathbf{\Delta} - \bar{\mathbf{\Delta}}) \Rightarrow \mathcal{N}(0, W); \quad \sqrt{N}(\hat{\theta} - \bar{\theta}) \Rightarrow \mathcal{N}(0, W'),$$

for appropriately defined coviariance matrices W and W', where $\bar{\theta} = (\mathbb{E}[X_i X_i^\top])^{-1}\mathbb{E}[X_i \hat{Y}_i]$ and $\bar{\mathbf{\Delta}} = (\mathbb{E}[X_i X_i^\top])^{-1}\mathbb{E}[X_i(\hat{Y}_i - Y_i)]$. With this, we can write the target estimand as $\theta^* = (\mathbb{E}[X_i X_i^\top])^{-1}\mathbb{E}[X_i Y_i] = \bar{\theta} - \bar{\mathbf{\Delta}}$.

Combining Theorem 3 of White with Slutsky's theorem, we get

$$\sqrt{N}(\hat{\theta}^{\mathrm{PP}} - \theta^*) = \sqrt{N}(\hat{\theta} - \bar{\theta}) - \sqrt{n}(\mathbf{\Delta} - \bar{\mathbf{\Delta}})\sqrt{\frac{N}{n}} \Rightarrow \mathcal{N}\left(0, W\frac{1}{p} + W'\right).$$

White also shows that V and V', as defined in Algorithm S4, are consistent estimates of W and W', respectively. Therefore, $\hat{\theta}^{\mathrm{PP}}$ is asymptotically normal and consistent, and we have a consistent estimate of its covariance. In particular,

$$V_{j^*j^*}\frac{N}{n} + V'_{j^*j^*} \to W_{j^*j^*}\frac{1}{p} + W'_{j^*j^*}.$$

This means that we can construct asymptotically valid confidence intervals via a normal approximation by choosing width $z_{1-\alpha/2}\sqrt{V_{j^*j^*}\frac{N}{n} + V'_{j^*j^*}}\sqrt{\frac{1}{N}} = z_{1-\alpha/2}\sqrt{\frac{V_{j^*j^*}}{n} + \frac{V'_{j^*j^*}}{N}}$, and this is precisely what Algorithm S4 accomplishes. □

Corollary S4 (Logistic regression)**.** *Let θ^* be the logistic regression solution:*

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[-Y_i \theta^\top X_i + \log(1 + \exp(\theta^\top X_i))\right].$$

Then, the prediction-powered confidence set in Algorithm S5 is valid: $\liminf_{n,N\to\infty} P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha$.

Proof. The proof follows a similar pattern as the first two corollaries, by arguing that the prediction-powered confidence set constructed in Algorithm S5 is a special case of the prediction-powered confidence set constructed in Theorem S1.

Since $\nabla L_\theta(x, y) = x(\mu_\theta(x) - y)$, we have

$$\mathbf{\Delta}_\theta \equiv \mathbf{\Delta} = \frac{1}{n}\sum_{i=1}^n X_i(\hat{Y}_i - Y_i); \quad m_\theta = \frac{1}{N}\sum_{i=1}^N X'_i(\mu_\theta(X'_i) - \hat{Y}'_i).$$

These quantities are explicitly computed in Algorithm S5. Moreover, the set $\mathcal{C}_\alpha^{\mathrm{PP}}$ constructed in Algorithm S5 exactly follows the recipe of Theorem S1, so the proof immediately follows. □

Corollary S5 (Convex risk minimization)**.** *Let θ^* be the convex risk minimizer:*

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} \mathbb{E}\left[L_\theta(X_i, Y_i)\right].$$

Then, the prediction-powered confidence set in Algorithm S6 is valid: $\liminf_{n,N\to\infty} P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha$.

The validity of Algorithm S6 follows directly by Theorem S1.

Prediction-Powered p-values

By relying on the standard duality between confidence intervals and p-values, we can immediately repurpose the presented theory to compute valid prediction-powered p-values.

To formalize this, suppose that we want to test the hull hypothesis $H_0 : \theta^* \in \Theta_0$, for some set $\Theta_0 \in \mathbb{R}^p$ (for example, a common choice when $p = 1$ is $\Theta_0 = \mathbb{R}_{\leq 0}$). Let $\mathcal{C}_\alpha$ be a valid confidence interval. Then, we can construct a valid p-value as

$$P = \inf\{\alpha : \theta_0 \notin \mathcal{C}_\alpha, \forall \theta_0 \in \Theta_0\}.$$

A p-value P is valid if it is super-uniform under the null, meaning $P(P \leq u) \leq u$ for all $u \in [0, 1]$. This is indeed the case for the p-value defined above, because when $\theta^* \in \Theta_0$, we have

$$P(P \leq u) \leq P(\theta^* \notin \mathcal{C}_u) \leq u.$$

The first inequality follows by the definition of P and the fact that $\theta^* \in \Theta_0$, and the second inequality follows by the validity of $\mathcal{C}_u$ at level $1 - u$. We are implicitly using the fact that $\mathcal{C}_u \subseteq \mathcal{C}_{u'}$ when $u \geq u'$.

The above derivation is a general recipe for deriving p-values from confidence intervals. For the prediction-powered confidence interval stated in Theorem S1, the corresponding prediction-powered p-value is given by:

$$P^{\mathrm{PP}} = \inf\left\{\alpha : |m_{\theta_0} + \mathbf{\Delta}_{\theta_0}| > w_{\theta_0}(\alpha),\ \forall \theta_0 \in \Theta_0\right\}.$$

Below we state analogues of Algorithms 1-5 when the goal is to compute a prediction-powered p-value.

Algorithm S7 Prediction-powered p-value for the mean

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, null set $\Theta_0$
$\hat{\theta}^{\mathrm{PP}} \leftarrow \hat{\theta} - \mathbf{\Delta} := \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i' - \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)$
$\hat{\sigma}^2_{\hat{Y}'} \leftarrow \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i' - \hat{\theta})^2$
$\hat{\sigma}^2_{\hat{Y}-Y} \leftarrow \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i - \mathbf{\Delta})^2$
Define $w(\alpha) := z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{\hat{Y}-Y}}{n} + \frac{\hat{\sigma}^2_{\hat{Y}'}}{N}}$
Output: prediction-powered p-value $P^{\mathrm{PP}} = \inf\{\alpha : \theta_0 \notin (\hat{\theta}^{\mathrm{PP}} \pm w(\alpha)), \forall \theta_0 \in \Theta_0\}$
```

Algorithm S8 Prediction-powered p-value for the median

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, null set $\Theta_0$
for $\theta \in \Theta_0$ do
    $\mathbf{\Delta}_\theta \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\{Y_i \leq \theta\} - \mathbb{1}\left\{\hat{Y}_i \leq \theta\right\}\right)$
    $m_\theta \leftarrow \frac{1}{2N}\sum_{i=1}^{N}\mathrm{sign}\left(\theta - \hat{Y}_i'\right)$
    $\hat{\sigma}^2_{\mathbf{\Delta}_\theta} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\{Y_i \leq \theta\} - \mathbb{1}\left\{\hat{Y}_i \leq \theta\right\} - \mathbf{\Delta}_\theta\right)^2$
    $\hat{\sigma}^2_{m_\theta} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}\left\{\hat{Y}_i' \leq \theta\right\} - m_\theta\right)^2$
    Define $w_\theta(\alpha) := z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{\mathbf{\Delta}_\theta}}{n} + \frac{\hat{\sigma}^2_{m_\theta}}{N}}$
Output: prediction-powered p-value $P^{\mathrm{PP}} = \inf\left\{\alpha : \theta \in \Theta_{\mathrm{grid}} : |m_{\theta_0} + \mathbf{\Delta}_{\theta_0}| > w_{\theta_0}(\alpha), \forall \theta_0 \in \Theta_0\right\}$
```

Algorithm S9 Prediction-powered p-value for the q-quantile

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, quantile $q$, null set $\Theta_0$
for $\theta \in \Theta_0$ do
    $\boldsymbol{\Delta}_\theta \leftarrow \frac{1}{n} \sum_{i=1}^n \left( \mathbb{1}\{Y_i \le \theta\} - \mathbb{1}\left\{\hat{Y}_i \le \theta\right\} \right)$
    $\hat{F}_\theta \leftarrow \frac{1}{N} \sum_{i=1}^N \mathbb{1}\left\{\hat{Y}'_i \le \theta\right\}$
    $\hat{\sigma}^2_{\boldsymbol{\Delta}_\theta} \leftarrow \frac{1}{n} \sum_{i=1}^n \left( \mathbb{1}\{Y_i \le \theta\} - \mathbb{1}\left\{\hat{Y}_i \le \theta\right\} - \boldsymbol{\Delta}_\theta \right)^2$
    $\hat{\sigma}^2_{\hat{Y}',\theta} \leftarrow \frac{1}{N} \sum_{i=1}^N \left( \mathbb{1}\left\{\hat{Y}'_i \le \theta\right\} - \hat{F}_\theta \right)^2$
    Define $w_\theta(\alpha) := z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}^2_{\boldsymbol{\Delta}_\theta}}{n} + \frac{\hat{\sigma}^2_{\hat{Y}',\theta}}{N}}$
Output: prediction-powered p-value $P^{\mathrm{PP}} = \inf\left\{\alpha : |\hat{F}_{\theta_0} + \boldsymbol{\Delta}_{\theta_0} - q| > w_{\theta_0}(\alpha), \forall \theta_0 \in \Theta_0\right\}$
```

Algorithm S10 Prediction-powered p-value for linear regression coefficients

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, coefficient $j^*$, null set $\Theta_0$
$\hat{\theta}^{\mathrm{PP}} \leftarrow \hat{\theta} - \boldsymbol{\Delta} := X'^{\dagger}\hat{Y}' - X^{\dagger}(\hat{Y} - Y)$
$\Sigma' \leftarrow \frac{1}{N}(X')^\top X'$, $M' \leftarrow \frac{1}{N} \sum_{i=1}^N (\hat{Y}'_i - (X'_i)^\top \hat{\theta})^2 X'_i (X'_i)^\top$
$V' \leftarrow (\Sigma')^{-1} M' (\Sigma')^{-1}$
$\Sigma \leftarrow \frac{1}{n} X^\top X$, $M \leftarrow \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i - X_i^\top \boldsymbol{\Delta})^2 X_i X_i^\top$
$V \leftarrow \Sigma^{-1} M \Sigma^{-1}$
Define $w(\alpha) := z_{1-\alpha/2} \sqrt{\frac{V_{j^*j^*}}{n} + \frac{V'_{j^*j^*}}{N}}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \inf\{\alpha : \theta_0 \notin (\hat{\theta}^{\mathrm{PP}}_{j^*} \pm w(\alpha)), \forall \theta_0 \in \Theta_0\}$
```

Algorithm S11 Prediction-powered p-value for logistic regression coefficients

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, null set $\Theta_0$
$\boldsymbol{\Delta}_j \leftarrow \frac{1}{n} \sum_{i=1}^n X_{i,j}(\hat{Y}_i - Y_i), \quad j \in [d]$
$\hat{\sigma}^2_{\boldsymbol{\Delta},j} \leftarrow \frac{1}{n} \sum_{i=1}^n \left( X_{i,j}(\hat{Y}_i - Y_i) - \boldsymbol{\Delta}_j \right)^2, \quad j \in [d]$
for $\theta \in \Theta_0$ do
    $m_{\theta,j} \leftarrow \frac{1}{N} \sum_{i=1}^N X'_{i,j} \left( \mu_\theta(X'_i) - \hat{Y}'_i \right), \quad j \in [d], \quad$ where $\mu_\theta(x) = \frac{1}{1+\exp(-x^\top \theta)}$
    $\hat{\sigma}^2_{m_\theta,j} \leftarrow \frac{1}{N} \sum_{i=1}^N \left( X'_{i,j}(\mu_\theta(X'_i) - \hat{Y}'_i) - m_{\theta,j} \right)^2, \quad j \in [d]$
    Define $w_{\theta,j}(\alpha) := z_{1-\alpha/(2d)} \sqrt{\frac{\hat{\sigma}^2_{\boldsymbol{\Delta},j}}{n} + \frac{\hat{\sigma}^2_{m_\theta,j}}{N}}, \quad j \in [d]$
Output: prediction-powered p-value $P^{\mathrm{PP}} = \inf\{\alpha : |m_{\theta_0,j} + \boldsymbol{\Delta}_j| > w_{\theta_0,j}(\alpha), \forall j \in [d], \theta_0 \in \Theta_0\}$
```

Corollary S6 (Mean p-value). *Let θ^* be the mean outcome:*

$$\theta^* = \mathbb{E}[Y_i].$$

Then, the prediction-powered p-value in Algorithm S7 is valid: under the null, $\liminf_{n,N\to\infty} P\left(P^{\mathrm{PP}} \le u\right) \le u, \forall u \in [0,1]$.

Corollary S7 (Quantile p-value). *Let θ^* be the q-quantile:*

$$\theta^* = \min\{\theta : P(Y_i \le \theta) \ge q\}.$$

Then, the prediction-powered p-value in Algorithm S9 is valid: under the null, $\liminf_{n,N\to\infty} P\left(P^{\mathrm{PP}} \le u\right) \le u, \forall u \in [0,1]$.

Corollary S8 (Linear regression p-value). *Fix $j^* \in [d]$. Let θ^* be the linear regression solution:*

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(Y_i - X_i^\top \theta)^2].$$

Then, the prediction-powered p-value in Algorithm S10 is valid: under the null, $\liminf_{n,N\to\infty} P\left(P^{\mathrm{PP}} \le u\right) \le u, \forall u \in [0,1]$.

Corollary S9 (Logistic regression p-value). *Let θ^* be the logistic regression solution:*

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[-Y_i \theta^\top X_i + \log(1 + \exp(\theta^\top X_i))\right].$$

Then, the prediction-powered p-value in Algorithm S11 is valid: under the null, $\liminf_{n,N\to\infty} P\left(P^{\mathrm{PP}} \leq u\right) \leq u, \forall u \in [0,1]$.

Nonasymptotic Analysis of Prediction-Powered Inference

Just like in most common confidence interval constructions, in Theorem S1 we used an argument based on the central limit theorem to construct the prediction-powered confidence set; as a result, the validity statement is asymptotic. We opted to present Theorem S1 as the main technical result due to its interpretability. However, the key underlying principles of prediction-powered inference can also be applied nonasymptotically. In this section we state the nonasymptotic counterparts of the previous results.

Validity

Below we state a nonasymptotic analogue of Theorem S1.

To state the result, let $\bar{m}_\theta$ denote the *population-level measure of fit* on the imputed data:

$$\bar{m}_\theta = \mathbb{E}[\nabla L_\theta(X_i, \hat{Y}_i)].$$

Similarly, we let $\bar{\mathbf{\Delta}}_\theta$ denote the *population-level rectifier*:

$$\bar{\mathbf{\Delta}}_\theta = \mathbb{E}\left[\nabla L_\theta(X_i, Y_i) - \nabla L_\theta(X_i, \hat{Y}_i)\right].$$

Theorem S2 (Validity of prediction-powered inference: nonasymptotic). *Let the labeled and unlabeled data be sampled i.i.d.. Suppose that the convex estimation problem is nondegenerate as in* (S1). *Fix $\alpha \in (0,1)$ and $\delta \in (0,\alpha)$. Suppose that, for any $\theta \in \mathbb{R}^p$, we can construct $\mathcal{R}_\theta(\delta)$ and $\mathcal{T}_\theta(\alpha-\delta)$ satisfying*

$$P\left(\bar{\mathbf{\Delta}}_\theta \in \mathcal{R}_\theta(\delta)\right) \geq 1-\delta; \quad P\left(\bar{m}_\theta \in \mathcal{T}_\theta(\alpha-\delta)\right) \geq 1-(\alpha-\delta).$$

Let $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta : 0 \in \mathcal{R}_\theta(\delta) + \mathcal{T}_\theta(\alpha-\delta)\}$, where $+$ denotes the Minkowski sum.[1] *Then,*

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1-\alpha.$$

Proof. We show that $\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}$ with probability at least $1-\alpha$; that is, with probability at least $1-\alpha$ it holds that

$$0 \in \mathcal{R}_{\theta^*}(\delta) + \mathcal{T}_{\theta^*}(\alpha-\delta).$$

Consider the event $E = \{\bar{\mathbf{\Delta}}_{\theta^*} \in \mathcal{R}_{\theta^*}(\delta)\} \cap \{\bar{m}_{\theta^*} \in \mathcal{T}_{\theta^*}(\alpha-\delta)\}$. By a union bound, $P(E) \geq 1-\alpha$. On the event E, we have that

$$\begin{aligned}\mathbb{E}[\nabla L_{\theta^*}(X_i, Y_i)] &= \mathbb{E}[\nabla L_{\theta^*}(X_i, Y_i)] - \mathbb{E}[\nabla L_{\theta^*}(X_i, \hat{Y}_i)] + \mathbb{E}[\nabla L_{\theta^*}(X_i, \hat{Y}_i)] \\ &= \bar{\mathbf{\Delta}}_{\theta^*} + \bar{m}_{\theta^*} \in \mathcal{R}_{\theta^*}(\delta) + \mathcal{T}_{\theta^*}(\alpha-\delta).\end{aligned}$$

The theorem finally follows by invoking the nondegeneracy condition, which ensures $\mathbb{E}[\nabla L_{\theta^*}(X_i, Y_i)] = 0$, so we have shown $0 \in \mathcal{R}_{\theta^*}(\delta) + \mathcal{T}_{\theta^*}(\alpha-\delta)$. □

We note that, because $\bar{m}_\theta$ and $\bar{\mathbf{\Delta}}_\theta$ are mean values of a well-specified quantity, the sets $\mathcal{R}_\theta(\delta)$ and $\mathcal{T}_\theta(\alpha-\delta)$ can be constructed using any off-she-shelf algorithm for computing a confidence intervals for the mean. In our explicit algorithm statements below, we choose a variance-adaptive confidence interval for the mean due to Waudby-Smith and Ramdas (*32*), which we state in Algorithm S15. We opt to present this construction as the default nonasymptotic confidence interval for the mean because of its strong practical performance. The only assumption required to apply Algorithm S15 is that the observations are almost surely bounded within a known interval.

[1] The Minkowski sum of two sets A and B is equal to $\{a+b : a \in A, b \in B\}$.

Algorithms

We state nonasymptotically-valid algorithms for prediction-powered mean estimation, quantile estimation, and logistic regression. These are nonasymptotic counterparts of Algorithms S1, S3, and S5, and they rely on the abstract recipe from Theorem 2.

Algorithm S12 Prediction-powered mean estimation (nonasymptotic)

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, error levels $\alpha, \delta \in (0,1)$, bound $B$
$(\hat{Y}^l(\alpha-\delta), \hat{Y}^u(\alpha-\delta)) \leftarrow \texttt{MeanCI}\Big(\{\hat{Y}_i'\}_{i=1}^N, \mathrm{err} = \alpha - \delta, \mathrm{range} = [0, B]\Big)$
$(\mathcal{R}^l(\delta), \mathcal{R}^u(\delta)) \leftarrow \texttt{MeanCI}\Big(\{\hat{Y}_i - Y_i\}_{i=1}^n, \mathrm{err} = \delta, \mathrm{range} = [-B, B]\Big)$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \Big(\hat{Y}^l(\alpha-\delta) - \mathcal{R}^u(\delta), \hat{Y}^u(\alpha-\delta) - \mathcal{R}^l(\delta)\Big)$
```

Algorithm S13 Prediction-powered quantile estimation (nonasymptotic)

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, quantile $q$, error levels $\alpha, \delta \in (0,1)$
Construct fine grid $\Theta_{\mathrm{grid}}$ between $\min_{i\in[N]} \hat{Y}_i'$ and $\max_{i\in[N]} \hat{Y}_i'$
for $\theta \in \Theta_{\mathrm{grid}}$ do
    $(\mathcal{R}_\theta^l(\delta), \mathcal{R}_\theta^u(\delta)) \leftarrow \texttt{MeanCI}\Big(\Big\{\mathbb{1}\{Y_i \le \theta\} - \mathbb{1}\Big\{\hat{Y}_i \le \theta\Big\}\Big\}_{i=1}^n, \mathrm{err} = \delta, \mathrm{range} = [-1, 1]\Big)$
    $(\hat{F}_\theta^l(\alpha-\delta), \hat{F}_\theta^u(\alpha-\delta)) \leftarrow \texttt{MeanCI}\Big(\Big\{\mathbb{1}\Big\{\hat{Y}_i' \le \theta\Big\}\Big\}_{i=1}^N, \mathrm{err} = \alpha - \delta, \mathrm{range} = [0, 1]\Big)$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \Big\{\theta \in \Theta_{\mathrm{grid}} : q \in \Big(\hat{F}_\theta^l(\alpha-\delta) + \mathcal{R}_\theta^l(\delta), \hat{F}_\theta^u(\alpha-\delta) + \mathcal{R}_\theta^u(\delta)\Big)\Big\}$
```

Algorithm S14 Prediction-powered logistic regression (nonasymptotic)

```
Input: labeled data $(X, Y)$, unlabeled features $X'$, predictions $(\hat{Y}, \hat{Y}')$, error levels $\alpha, \delta \in (0,1)$, bounds $B_j$
Construct fine grid $\Theta_{\mathrm{grid}} \subset \mathbb{R}^d$ of possible coefficients
$(\mathcal{R}_j^l(\delta), \mathcal{R}_j^u(\delta)) \leftarrow \texttt{MeanCI}\left(\{X_{i,j}(\hat{Y}_i - Y_i)\}_{i=1}^n, \mathrm{err} = \delta, \mathrm{range} = [-B_j, B_j]\right), j \in [d]$
for $\theta \in \Theta_{\mathrm{grid}}$ do
    $(m_{\theta,j}^l(\alpha-\delta), m_{\theta,j}^u(\alpha-\delta)) \leftarrow \texttt{MeanCI}\left(\{X_{i,j}'\left(\mu_\theta(X_i') - \hat{Y}_i'\right)\}_{i=1}^N, \mathrm{err} = \frac{\alpha-\delta}{d}, \mathrm{range} = [-B_j, B_j]\right), j \in [d]$,
    where $\mu_\theta(x) = \frac{1}{1+\exp(-x^\top\theta)}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \Big\{\theta \in \Theta_{\mathrm{grid}} : 0 \in \Big[m_{\theta,j}^l(\alpha-\delta) + \mathcal{R}_j^l(\delta), m_{\theta,j}^u(\alpha-\delta) + \mathcal{R}_j^u(\delta)\Big], \forall j \in [d]\Big\}$
```

Corollary S10 (Mean estimation: nonasymptotic)**.** *Let θ^* be the mean outcome:*

$$\theta^* = \mathbb{E}[Y_i].$$

Suppose that $Y_i, \hat{Y}_i \in [0, B]$ almost surely. Then, the prediction-powered confidence set in Algorithm S12 is valid: $P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \ge 1 - \alpha$.

Proof. The proof follows by instantiating the terms in Theorem 2. In particular, we have $\mathbb{E}[\nabla L_\theta(\hat{Y}_i)] = \theta - \mathbb{E}[\hat{Y}_i]$, hence it is valid to construct $\mathcal{T}_\theta(\alpha-\delta)$ as:

$$m_\theta \in \mathcal{T}_\theta(\alpha-\delta) = \theta - (\hat{Y}^l(\alpha-\delta), \hat{Y}^u(\alpha-\delta)).$$

Therefore, the condition $0 \in \mathcal{R}(\delta) + \mathcal{T}_\theta(\alpha-\delta)$ becomes

$$0 \in (\mathcal{R}^l(\delta), \mathcal{R}^u(\delta)) + \theta - (\hat{Y}^l(\alpha-\delta), \hat{Y}^u(\alpha-\delta)),$$

which after rearranging and simplifying is equivalent to

$$\theta \in \left(\hat{Y}^l(\alpha-\delta) - \mathcal{R}^u(\delta), \hat{Y}^u(\alpha-\delta) - \mathcal{R}^l(\delta)\right).$$

This set exactly matches the set $\mathcal{C}_\alpha^{\mathrm{PP}}$ constructed in Algorithm S12. □

Corollary S11 (Quantile estimation: nonasymptotic)**.** *Let θ^* be the q-quantile:*

$$\theta^* = \min\{\theta : P(Y_i \leq \theta) \geq q\}.$$

Then, the prediction-powered confidence set in Algorithm S13 is valid: $P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1 - \alpha$.

Proof. The proof follows by instantiating the terms in Theorem 2. First, we have $\mathbb{E}[\nabla L_\theta(\hat{Y}_i)] = -q + P(\hat{Y}_i \leq \theta)$; therefore, it is valid to construct $\mathcal{T}_\theta(\alpha - \delta)$ as:

$$\mathbb{E}[\nabla L_\theta(\hat{Y}_i)] \in \mathcal{T}_\theta(\alpha - \delta) = -q + \left(\hat{F}_\theta^l(\alpha - \delta), \hat{F}_\theta^u(\alpha - \delta)\right).$$

Therefore, the condition $0 \in \mathcal{R}_\theta(\delta) + \mathcal{T}_\theta(\alpha - \delta)$ becomes

$$q \in \left(\hat{F}_\theta^l(\alpha - \delta) + \mathcal{R}_\theta^l(\delta), \hat{F}_\theta^u(\alpha - \delta) + \mathcal{R}_\theta^u(\delta)\right),$$

which matches the condition used to form $\mathcal{C}_\alpha^{\mathrm{PP}}$ in Algorithm S13. □

Corollary S12 (Logistic regression: nonasymptotic)**.** *Let θ^* be the logistic regression solution:*

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}[-Y_i \theta^\top X_i + \log(1 + \exp(\theta^\top X_i))].$$

Suppose that $|X_{i,j}| \leq B_j$ and $Y_i, \hat{Y}_i \in [0,1]$ almost surely. Then, the prediction-powered confidence set in Algorithm S14 is valid: $P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1 - \alpha$.

Proof. We instantiate the relevant terms in Theorem 2. We have $\mathbb{E}[\nabla L_\theta(X_i, \hat{Y}_i)] = \mathbb{E}\left[-X_i \hat{Y}_i + X_i \frac{1}{1+\exp(-X_i^\top \theta)}\right]$. Note that, because X_i is coordinatewise bounded, and $Y_i, \frac{1}{1+\exp(-X_i^\top \theta)} \in [0,1]$, we have $|(\nabla L_\theta(X_i, \hat{Y}_i))_j| \leq B_j$ almost surely. Therefore, we can construct $\mathcal{T}_\theta(\alpha - \delta)$ as:

$$m_\theta \in \mathcal{T}_\theta(\alpha - \delta) = \left(m_\theta^l(\alpha - \delta), m_\theta^u(\alpha - \delta)\right) = \left(m_{\theta,1}^l(\alpha - \delta), m_{\theta,1}^u(\alpha - \delta)\right) \times \cdots \times \left(m_{\theta,d}^l(\alpha - \delta), m_{\theta,d}^u(\alpha - \delta)\right).$$

Since the rectifier has no dependence on θ, the condition $0 \in \mathcal{R}_\theta(\delta) + \mathcal{T}_\theta(\alpha - \delta)$ becomes

$$0 \in (\mathcal{R}_j^l(\delta), \mathcal{R}_j^u(\delta)) + \left(m_{\theta,j}^l(\alpha - \delta), m_{\theta,j}^u(\alpha - \delta)\right), \quad \forall j \in [d],$$

which matches the condition in $\mathcal{C}_\alpha^{\mathrm{PP}}$ in Algorithm S14. □

We note that there exists an analogous nonasymptotic algorithm for linear regression, however we do not recommend it in practice. The reason is that the refined (but asymptotic) analysis used to prove Corollary S3 shows that it is sufficient to analyze a one-dimensional rectifier, while directly invoking Theorem 2 would require analyzing a d-dimensional rectifier and thus yields more conservative intervals.

Algorithm S15 `MeanCI` (see Proposition S1)

```
Input: data points $\{Z_1, \dots, Z_n\}$, error level $\alpha \in (0,1)$, range $[L, U]$ s.t. $Z_i \in [L, U]$
For all $i \in [n]$, let $Z_i \leftarrow (Z_i - L)/(U - L)$          ▷ normalize data to interval $[0,1]$
Construct fine grid $M_{\mathrm{grid}}$ of interval $[0,1]$
Initialize active set $\mathcal{A} = M_{\mathrm{grid}}$
for $t \in 1, \dots, n$ do
    Set $\hat{\mu}_t \leftarrow \frac{0.5 + \sum_{j=1}^t Z_j}{t+1}$, $\hat{\sigma}_t^2 \leftarrow \frac{0.25 + \sum_{j=1}^t (Z_j - \hat{\mu}_t)^2}{t+1}$, $\lambda_t \leftarrow \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{t-1}^2}}$
    for $m \in \mathcal{A}$ do
        $M_t^+(m) \leftarrow \left(1 + \min\left(\lambda_t, \frac{0.5}{m}\right)(Z_t - m)\right) M_{t-1}^+(m)$
        $M_t^-(m) \leftarrow \left(1 - \min\left(\lambda_t, \frac{0.5}{1-m}\right)(Z_t - m)\right) M_{t-1}^-(m)$
        $M_t(m) \leftarrow \frac{1}{2}\max\left\{M_t^+(m),\ M_t^-(m)\right\}$          ▷ construct test martingale for $m \in [0,1]$
        if $M_t(m) \geq 1/\alpha$ then
            $\mathcal{A} \leftarrow \mathcal{A} \setminus \{m\}$          ▷ Remove $m$ from active set
Output: Confidence set for the mean $\mathcal{C}_\alpha = \{m(U - L) + L : m \in \mathcal{A}\}$
```

Inference Under Distribution Shift

In the main text we focused on forming prediction-powered confidence intervals when the labeled and unlabeled data come from the same distribution. Herein, we extend our tools to the case where the labeled data (X, Y) comes from $\mathbb{P}$ and the unlabeled data (X', Y')—which defines the target of inference θ^*—comes from $\mathbb{Q}$, and these are related by either a label shift or a covariate shift. For covariate shift, we handle all estimation problems previously studied; for label shift, we handle certain types of linear problems.

We will write $\mathbb{E}_{\mathbb{Q}}, \mathbb{E}_{\mathbb{P}}$, etc to indicate which distribution the data inside the expectation is sampled from.

Covariate Shift

First, we assume that $\mathbb{Q}$ is a known *covariate shift* of $\mathbb{P}$. That is, if we denote by $\mathbb{Q} = \mathbb{Q}_X \cdot \mathbb{Q}_{Y|X}$ and $\mathbb{P} = \mathbb{P}_X \cdot \mathbb{P}_{Y|X}$ the relevant marginal and conditional distributions, we assume that $\mathbb{Q}_{Y|X} = \mathbb{P}_{Y|X}$. As in previous sections, we consider estimands of the form

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{Q}}[L_\theta(X_i, Y_i)]. \tag{S3}$$

Estimands of the form (S3) can be related to risk minimizers on $\mathbb{P}$ using the Radon-Nikodym derivative. In particular, suppose that $\mathbb{Q}_X$ is dominated by $\mathbb{P}_X$ and assume that the Radon-Nikodym derivative $w(x) = \frac{\mathbb{Q}_X}{\mathbb{P}_X}(x)$ is known. Then, we can rewrite (S3) as

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}}[L_\theta^w(X_i, Y_i)],$$

where $L_\theta^w(x, y) = w(x) L_\theta(x, y)$. In words, risk minimizers on $\mathbb{Q}$ can simply be written as risk minimizers on $\mathbb{P}$, but with a reweighted loss function. This permits inference on the rectifier to be based on data sampled from $\mathbb{P}$ as before. For concreteness, we explain the approach in detail for convex risk minimizers. Let

$$\bar{m}_\theta^w = \mathbb{E}_{\mathbb{P}}\left[\nabla L_\theta^w(X_i, \hat{Y}_i)\right]; \quad \bar{\boldsymbol{\Delta}}_\theta^w = \mathbb{E}_{\mathbb{P}}\left[\nabla L_\theta^w(X_i, Y_i) - \nabla L_\theta^w(X_i, \hat{Y}_i)\right],$$

where $\nabla L_\theta^w(x, y) = \nabla L_\theta(x, y) \cdot w(x)$ and ∇L_θ is a subgradient of L_θ as before. A confidence set for the above rectifier suffices for prediction-powered inference on θ^*.

Corollary S13 (Covariate shift)**.** *Let the unlabeled data distribution be a covariate shift of the labeled data distribution. Suppose that the problem* (S3) *is a nondegenerate convex estimation problem. Fix $\alpha \in (0, 1)$ and $\delta \in (0, \alpha)$. Suppose that, for any $\theta \in \mathbb{R}^p$, we can construct $\mathcal{R}_\theta(\delta)$ and $\mathcal{T}_\theta(\alpha - \delta)$ satisfying*

$$P\left(\bar{\boldsymbol{\Delta}}_\theta^w \in \mathcal{R}_\theta(\delta)\right) \geq 1 - \delta; \quad P\left(\bar{m}_\theta^w \in \mathcal{T}_\theta(\alpha - \delta)\right) \geq 1 - (\alpha - \delta).$$

Let $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta : 0 \in \mathcal{R}_\theta(\delta) + \mathcal{T}_\theta(\alpha - \delta)\}$, where $+$ denotes the Minkowski sum. Then,

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1 - \alpha.$$

Label Shift

Next, we analyze classification problems where the proportions of the classes in the labeled data is different from those in the unlabeled data. This problem has been studied before in the literature on domain adaptation, e.g. by Lipton et al. (*33*), but our treatment focuses on the formation of confidence intervals. Formally, let $\mathcal{Y} = \{1, ..., K\}$ be the label space and assume that $\mathbb{Q}_{X|Y} = \mathbb{P}_{X|Y}$. We consider estimands of the form

$$\theta^* = \mathbb{E}_{\mathbb{Q}_Y}[\nu(Y)],$$

where $\nu : \mathcal{Y} \to \mathbb{R}$ is a fixed function. For example, choosing $\nu(y) = \mathbb{1}\{y = k\}$ for some $k \in [K]$ asks for inference on the proportion of instances that belong to class k.

Using an analogous decomposition to the one for mean estimation, we can write

$$\theta^* = \mathbb{E}_{\mathbb{Q}_{\hat{Y}}}[\nu(\hat{Y})] + (\mathbb{E}_{\mathbb{Q}_Y}[\nu(Y)] - \mathbb{E}_{\mathbb{Q}_{\hat{Y}}}[\nu(\hat{Y})]) = \bar{\theta} + \bar{\boldsymbol{\Delta}},$$

where $\mathbb{Q}_{\hat{Y}}$ denotes the distribution of the prediction $\hat{Y}$ based on features $X \sim \mathbb{Q}_X$. The quantity $\bar{\theta}$ can be estimated using the unlabeled data from $\mathbb{Q}$ and the model. Estimating the quantity $\bar{\boldsymbol{\Delta}}$ using samples from $\mathbb{P}$ will require leveraging the structure of the distribution shift. Central to our analysis will be the confusion matrix

$$\mathcal{K}_{j,l} = \mathbb{Q}\left(\hat{Y} = j \;\middle|\; Y = l\right), \; j, l \in [K].$$

The label-shift assumption implies that $\mathcal{K}_{j,l} = \mathbb{P}\left(\hat{Y} = j \mid Y = l\right)$, which can be estimated from labeled data sampled from $\mathbb{P}$. In particular, we estimate $\mathcal{K}$ from the labeled data as

$$\widehat{\mathcal{K}}_{j,l} = \frac{1}{n(l)} \sum_{i=1}^{n} \mathbb{1}\left\{\hat{Y}_i = j, Y_i = l\right\}, \text{ where } n(l) = \sum_{i=1}^{n} \mathbb{1}\left\{Y_i = l\right\}.$$

Similarly, we can estimate $\mathbb{Q}_{\hat{Y}}(k), k \in [K]$ as

$$\widehat{\mathbb{Q}}_{\hat{Y}}(k) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{\hat{Y}'_i = k\right\}.$$

Treating $\mathbb{Q}_{\hat{Y}}$ and $\mathbb{Q}_Y$ as vectors, notice that we can write $\mathbb{Q}_{\hat{Y}} = \mathcal{K}\mathbb{Q}_Y$, and hence $\mathbb{Q}_Y = \mathcal{K}^{-1}\mathbb{Q}_{\hat{Y}}$. This leads to a natural estimate of $\mathbb{Q}_Y$, $\widehat{\mathbb{Q}}_Y = \widehat{\mathcal{K}}^{-1}\widehat{\mathbb{Q}}_{\hat{Y}}$. Below, we use these quantities to construct a prediction-powered confidence interval for $\theta^* = \mathbb{E}_{\mathbb{Q}_Y}[\nu(Y)]$.

Theorem S3 (Label shift)**.** *Let the unlabeled data distribution be a label shift of the labeled data distribution. Fix $\alpha \in (0,1)$ and $\delta \in (0, \alpha)$. Let*

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\mathbb{E}_{\widehat{\mathbb{Q}}_Y}[\nu(Y)] \pm \left(\max_{l,k \in [K]} \max_{p \in C_{l,k}} |\widehat{\mathcal{K}}_{l,k} - p| + \sqrt{\frac{1}{2N} \log \frac{2}{\alpha - \delta}}\right)\right),$$

where

$$C_{l,k} = \left\{p : n(k)\widehat{\mathcal{K}}_{l,k} \in \left[F^{-1}_{\mathrm{Binom}(n(k),p)}\left(\frac{\delta}{2K^2}\right), F^{-1}_{\mathrm{Binom}(n(k),p)}\left(1 - \frac{\delta}{2K^2}\right)\right]\right\}$$

and $F_{\mathrm{Binom}(n(k),p)}$ denotes the Binomial CDF. Then,

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1 - \alpha.$$

Naturally, the confidence interval becomes more conservative as the number of classes grows. Also, the power of the bound depends on the smallest number of instances observed for a particular class.

Proof. Notice that we can write $\mathbb{E}_{\mathbb{Q}_Y}[\nu(Y)] = \nu^\top \mathbb{Q}_Y$, where on the right-hand side we are treating $\nu = (\nu(1), \ldots, \nu(K))$ and $\mathbb{Q}_Y = (\mathbb{Q}_Y(1), \ldots, \mathbb{Q}_Y(K))$ as vectors of length K. We can write similar expressions for $\mathbb{Q}_{\hat{Y}}, \widehat{\mathbb{Q}}_Y$, etc. Using this notation, by triangle inequality we have

$$|\theta^* - \nu^\top \widehat{\mathbb{Q}}_Y| = |\nu^\top \mathbb{Q}_Y - \nu^\top \widehat{\mathbb{Q}}_Y| \leq \left|\nu^\top \widehat{\mathcal{K}}^{-1}(\mathbb{Q}_{\hat{Y}} - \widehat{\mathbb{Q}}_{\hat{Y}})\right| + \left|\nu^\top \mathcal{K}^{-1}\mathbb{Q}_{\hat{Y}} - \nu^\top \widehat{\mathcal{K}}^{-1}\mathbb{Q}_{\hat{Y}}\right|.$$

We bound the first term using Hölder's inequality,

$$\left|\nu^\top \widehat{\mathcal{K}}^{-1}(\mathbb{Q}_{\hat{Y}} - \widehat{\mathbb{Q}}_{\hat{Y}})\right| \leq \|\nu^\top \widehat{\mathcal{K}}^{-1}\|_1 \|\mathbb{Q}_{\hat{Y}} - \widehat{\mathbb{Q}}_{\hat{Y}}\|_\infty.$$

For the second term, we write

$$\left|\nu^\top \mathcal{K}^{-1}\mathbb{Q}_{\hat{Y}} - \nu^\top \widehat{\mathcal{K}}^{-1}\mathbb{Q}_{\hat{Y}}\right| = \left|\nu^\top \widehat{\mathcal{K}}^{-1}(\widehat{\mathcal{K}} - \mathcal{K})\mathcal{K}^{-1}\mathbb{Q}_{\hat{Y}}\right|.$$

In the above equation, the factor on the right, $\mathcal{K}^{-1}\mathbb{Q}_{\hat{Y}}$, is exactly equal to $\mathbb{Q}_Y$, and thus lives on the simplex, which we denote by Δ. Using this fact and Hölder's inequality,

$$\left|\nu^\top \widehat{\mathcal{K}}^{-1}(\widehat{\mathcal{K}} - \mathcal{K})\mathcal{K}^{-1}\mathbb{Q}_{\hat{Y}}\right| \leq \sup_{q \in \Delta} \left|\nu^\top \widehat{\mathcal{K}}^{-1}(\widehat{\mathcal{K}} - \mathcal{K})q\right| \leq \left\|\nu^\top \widehat{\mathcal{K}}^{-1}\right\|_1 \sup_{q \in \Delta} \left\|(\widehat{\mathcal{K}} - \mathcal{K})q\right\|_\infty.$$

Next, we have

$$\sup_{q\in\Delta} \|(\widehat{\mathcal{K}} - \mathcal{K})q\|_\infty = \max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty,$$

where $\mathcal{K}_k$ indexes the k-th column of $\mathcal{K}$. This yields the expression

$$\left\|\nu^\top \widehat{\mathcal{K}}^{-1}\right\|_1 \sup_{q\in\Delta} \left\|(\widehat{\mathcal{K}} - \mathcal{K})q\right\|_\infty = \left\|\nu^\top \widehat{\mathcal{K}}^{-1}\right\|_1 \max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty.$$

Putting everything together, we have

$$|\nu^\top \mathbb{Q}_Y - \nu^\top \widehat{\mathbb{Q}}_Y| \le \|\nu^\top \widehat{\mathcal{K}}^{-1}\|_1 \left(\|\mathbb{Q}_{\hat{Y}} - \widehat{\mathbb{Q}}_{\hat{Y}}\|_\infty + \max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty \right).$$

Since $\|\nu^\top \widehat{\mathcal{K}}^{-1}\|_1$ can be evaluated empirically, it remains to bound the distributional distances $\|\mathbb{Q}_{\hat{Y}} - \widehat{\mathbb{Q}}_{\hat{Y}}\|_\infty$ and $\max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty$.

For the first distance, $\|\mathbb{Q}_{\hat{Y}} - \widehat{\mathbb{Q}}_{\hat{Y}}\|_\infty$, we can simply apply the DKWM inequality (*34, 35*), which gives

$$\|\mathbb{Q}_{\hat{Y}} - \widehat{\mathbb{Q}}_{\hat{Y}}\|_\infty \le \sqrt{\frac{2}{N} \log \frac{2}{\alpha - \delta}}$$

with probability $1 - (\alpha - \delta)$. See (*36*) for details.

For the second term, $\max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty$, since we only have n observations for estimation, we use a more adaptive concentration result. In particular, for each $l, k \in [K]$, $n(k)\widehat{\mathcal{K}}_{l,k}$ (conditional on the k-th column) follows a binomial distribution with $n(k)$ draws and success probability $\mathcal{K}_{l,k}$. Therefore, if we let

$$C_{l,k} = \left\{ p : n(k)\widehat{\mathcal{K}}_{l,k} \in \left(F^{-1}_{\mathrm{Binom}(n(k),p)}\left(\frac{\delta}{2K^2}\right), F^{-1}_{\mathrm{Binom}(n(k),p)}\left(1 - \frac{\delta}{2K^2}\right) \right) \right\},$$

where $F_{\mathrm{Binom}(n(k),p)}$ denotes the Binomial CDF, then by a union bound:

$$P\left(\max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty \ge \max_{l,k\in[K]} \max_{p\in C_{l,k}} |\widehat{\mathcal{K}}_{l,k} - p| \right) \le \delta.$$

Combining the last three inequalities yields the final result. □

Extensions

Beyond Convex Estimation

The tools developed previously were tailored to unconstrained convex optimization problems. More specifically, they relied on the property given by equation (S1). In general, however, inferential targets can be defined in terms of nonconvex losses or they may have (possibly even nonconvex) constraints. For such general optimization problems, we cannot expect the condition (S1) to hold. In this section we generalize our approach to a broad class of risk minimizers:

$$\theta^* = \arg\min_{\theta\in\Theta} \mathbb{E}[L_\theta(X_i, Y_i)], \tag{S4}$$

where $L_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a possibly nonconvex loss function and Θ is an arbitrary set of admissible parameters. As before, if θ^* is not a unique minimizer, our method will return a set that contains all minimizers. We note that, technically, the approach in Algorithms S1-S6 is valid even for nonconvex problems as long as we know that condition (S1) holds; e.g., if the loss is differentiable and the optimization is unconstrained. However, even then the set of points satisfying the condition may be too large and thus the returned confidence sets could be large as well if the optimization problem does not have a unique minimizer.

The problem (S4) subsumes all previously studied settings. Indeed, when the loss L_θ is convex and subdifferentiable and $\Theta = \mathbb{R}^p$ for some p—which is the case for all problems previously studied—θ^* can be equivalently characterized via the condition (S1). In this section we provide a solution that can handle problems of the form (S4) in full generality. For concreteness, we provide an analogue of the previous nonasymptotic result, however one can analogously derive an analogue of the asymptotic statement. We note that the solution does not reduce to the one in Algorithms S1-S6 for convex estimation problems, and we expect the methods from Algorithms S1-S6 to be more powerful for convex estimation problems with low-dimensional rectifiers.

We rely on the following population-level measure of fit and rectifier:

$$\bar{m}_\theta = \mathbb{E}[L_\theta(X_i, \hat{Y}_i)]; \quad \bar{\boldsymbol{\Delta}}_\theta = \mathbb{E}\left[L_\theta(X_i, Y_i) - L_\theta(X_i, \hat{Y}_i)\right]. \tag{1}$$

Notice that the rectifier (1) is always one-dimensional, while previously the rectifier was p-dimensional.

One key difference relative to the approach designed for convex problems is that we have an additional step of data splitting. We need the additional step because, unlike in convex estimation where we know $\mathbb{E}[\nabla L_{\theta^*}(X_i, Y_i)] = 0$, for general problems we do not know the value of $\mathbb{E}[L_{\theta^*}(X_i, Y_i)]$. To circumvent this issue, we estimate $\mathbb{E}[L_{\theta^*}(X_i, Y_i)]$ by approximating θ^* with an imputed estimate on the first $N/2$ unlabeled data points (for simplicity, take N to be even). To state the main result, we define

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{2}{N} \sum_{i=1}^{N/2} L_\theta(X_i', \hat{Y}_i'), \quad L_\theta^{\hat{Y}} := \frac{2}{N} \sum_{i=N/2+1}^{N} L_\theta(X_i', \hat{Y}_i').$$

Theorem S4 (General risk minimization). *Let the labeled and unlabeled data be sampled i.i.d.. Fix $\alpha \in (0,1)$ and $\delta \in (0, \alpha)$. Suppose that, for any $\theta \in \Theta$, we can construct $(\mathcal{R}_\theta^l(\delta/2), \mathcal{R}_\theta^u(\delta/2))$ and $\left(\mathcal{T}_\theta^l\left(\frac{\alpha-\delta}{2}\right), \mathcal{T}_\theta^u\left(\frac{\alpha-\delta}{2}\right)\right)$ such that*

$$\begin{aligned}
&P\left(\bar{\boldsymbol{\Delta}}_\theta \le \mathcal{R}_\theta^u(\delta/2)\right) \ge 1 - \delta/2; \quad P\left(\bar{\boldsymbol{\Delta}}_\theta \ge \mathcal{R}_\theta^l(\delta/2)\right) \ge 1 - \delta/2; \\
&P\left(L_\theta^{\hat{Y}} - \bar{m}_\theta \le \mathcal{T}_\theta^u\left(\frac{\alpha-\delta}{2}\right)\right) \ge 1 - \frac{\alpha-\delta}{2}; \quad P\left(L_\theta^{\hat{Y}} - \bar{m}_\theta \ge \mathcal{T}_\theta^l\left(\frac{\alpha-\delta}{2}\right)\right) \ge 1 - \frac{\alpha-\delta}{2}.
\end{aligned}$$

Let

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta \in \Theta : L_\theta^{\hat{Y}} \le L_{\hat{\theta}}^{\hat{Y}} - \mathcal{R}_\theta^l\left(\frac{\delta}{2}\right) + \mathcal{R}_{\hat{\theta}}^u\left(\frac{\delta}{2}\right) + \mathcal{T}_\theta^u\left(\frac{\alpha-\delta}{2}\right) - \mathcal{T}_{\hat{\theta}}^l\left(\frac{\alpha-\delta}{2}\right)\right\}.$$

Then, we have

$$P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \ge 1 - \alpha.$$

For example, if the loss $L_\theta(x, y)$ takes values in $[0, B]$ for all x, y, then we can set $\mathcal{T}_\theta(\alpha - \delta) = B\sqrt{\frac{\log(1/(\alpha-\delta))}{N}}$. The validity of this choice follows by Hoeffding's inequality.

Proof. Define

$$\bar{L}_\theta = \mathbb{E}[\ell_\theta(X_i, Y_i)], \quad \bar{L}_\theta^{\hat{Y}} = \mathbb{E}[L_\theta(X_i, \hat{Y}_i)].$$

By the definition of θ^*, we have

$$\begin{aligned}
L_{\theta^*}^{\hat{Y}} &= (L_{\theta^*}^{\hat{Y}} - \bar{L}_{\theta^*}) + (\bar{L}_{\theta^*} - \bar{L}_{\hat{\theta}}) + (\bar{L}_{\hat{\theta}} - L_{\hat{\theta}}^{\hat{Y}}) + L_{\hat{\theta}}^{\hat{Y}} \\
&\le (L_{\theta^*}^{\hat{Y}} - \bar{L}_{\theta^*}) + (\bar{L}_{\hat{\theta}} - L_{\hat{\theta}}^{\hat{Y}}) + L_{\hat{\theta}}^{\hat{Y}}.
\end{aligned}$$

By applying the validity of the confidence bounds, a union bound implies that with probability $1 - \alpha$ we have

$$\begin{aligned}
L_{\theta^*}^{\hat{Y}} &\le (\bar{L}_{\theta^*}^{\hat{Y}} - \bar{L}_{\theta^*}) + (\bar{L}_{\hat{\theta}} - \bar{L}_{\hat{\theta}}^{\hat{Y}}) + L_{\hat{\theta}}^{\hat{Y}} + \mathcal{T}_{\theta^*}^u\left(\frac{\alpha-\delta}{2}\right) - \mathcal{T}_{\hat{\theta}}^l\left(\frac{\alpha-\delta}{2}\right) \\
&= -\bar{\boldsymbol{\Delta}}_{\theta^*} + \bar{\boldsymbol{\Delta}}_{\hat{\theta}} + L_{\hat{\theta}}^{\hat{Y}} + \mathcal{T}_{\theta^*}^u\left(\frac{\alpha-\delta}{2}\right) - \mathcal{T}_{\hat{\theta}}^l\left(\frac{\alpha-\delta}{2}\right) \\
&\le -\mathcal{R}_{\theta^*}^l(\delta/2) + \mathcal{R}_{\hat{\theta}}^u(\delta/2) + L_{\hat{\theta}}^{\hat{Y}} + \mathcal{T}_{\theta^*}^u\left(\frac{\alpha-\delta}{2}\right) - \mathcal{T}_{\hat{\theta}}^l\left(\frac{\alpha-\delta}{2}\right).
\end{aligned}$$

Therefore, with probability $1 - \alpha$ we have that $\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}$, as desired. □

Mode estimation. A commonplace inference task that does not fall under convex estimation is the problem of estimating the mode of the outcome distribution. When the outcome takes values in a discrete set Θ, this can be done by using the loss function $L_\theta(y) = \mathbb{1}\{y \neq \theta\}, \theta \in \Theta$. A generalization of this approach to continuous outcome distributions is obtained by defining the loss $L_\theta(y) = \mathbb{1}\{|y - \theta| > \eta\}$, for some width parameter $\eta > 0$. The target of inference is thus the point $\theta \in \mathbb{R}$ that has the most probability mass in its η-neighborhood, $\theta^* = \arg\min_{\theta \in \mathbb{R}} P(|Y_i - \theta| > \eta)$. Theorem 4 applies directly in both the discrete and continuous cases.

Tukey's biweight robust mean. The Tukey biweight loss function is a commonly used loss in robust statistics that results in an outlier-robust mean estimate. It behaves approximately like a quadratic near the origin and is constant far away from the origin. Formally, Tukey's biweight loss function is given by

$$L_\theta(y) = \begin{cases} \frac{c^2}{6}\left(1 - \left(1 - \frac{(y-\theta)^2}{c^2}\right)^3\right), & |y - \theta| \leq c, \\ \frac{c^2}{6}, & \text{otherwise}, \end{cases}$$

where c is a user-specified tuning parameter. It is not hard to see that the function $L_\theta(y)$ is nonconvex and hence not amenable to the initial analysis; however, Theorem 4 applies.

Model selection. Nonconvex risk minimization problems are ubiquitous in model selection. For example, a common model selection strategy is best subset selection, which optimizes the squared loss, $L_\theta(x, y) = (y - x^\top \theta)^2$, subject to the constraint $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq k\}$. Here, Θ is the space of all k-sparse vectors for a user-chosen parameter k. Even though the loss function is convex, Θ is a nonconvex constraint set and hence we cannot rely on the condition (S1) to find the minimizer. However, Theorem 4 still applies.

Inference on a Finite Population

The techniques developed in this paper directly translate to the *finite-population* setting. Here, we treat (X', Y') as a fixed finite population consisting of N feature-outcome pairs, without imposing any distributional assumptions on the data points. Analogously to the i.i.d. setting, we observe all features X' and a small set of outcomes. Specifically, we assume that we observe $(Y'_i)_{i \in \mathcal{I}}$, where $\mathcal{I} = \{i_1, \ldots, i_n\}$ is a uniformly sampled subset of $[N]$ of size $n \ll N$. In this section we adapt all our main results to the finite-population context.

Given a loss function L_θ and parameter space Θ, the target estimand is the risk minimizer we would compute if we could observe the whole population:

$$\theta^* = \underset{\theta \in \Theta}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} L_\theta(X'_i, Y'_i). \tag{2}$$

The following results mirror the results for convex estimation. All results in this section are proved essentially identically as their i.i.d. counterparts.

In what follows, we construct prediction-powered confidence sets $\mathcal{C}_\alpha^{\mathrm{PP}}$ assuming a valid confidence set around the rectifier (defined below for the finite-population context). The confidence set for the rectifier can be constructed from $(X'_i, Y'_i)_{i \in \mathcal{I}}$ via a direct application of off-the-shelf results: in Proposition S4 we state an asymptotically valid interval for the mean based on a finite-population version of the central limit theorem, and in Proposition S3 we state a nonasymptotically valid interval for the mean for finite populations due to Waudby-Smith and Ramdas (*32*). The only assumption required to apply the latter is that $\nabla L_\theta(X'_i, Y'_i) - \nabla L_\theta(X'_i, \hat{Y}'_i)$ has a known bound valid for all $i \in [N]$.

In the finite-population setting, the mild nondegeneracy condition ensured by convexity takes the form

$$\frac{1}{N} \sum_{i=1}^{N} \nabla L_{\theta^*}(X'_i, Y'_i) = 0, \tag{3}$$

where ∇L_θ is a subgradient of L_θ. The population-level rectifier is thus:

$$\bar{\mathbf{\Delta}}_\theta = \frac{1}{N} \sum_{i=1}^{N} \left(\nabla L_\theta(X'_i, Y'_i) - \nabla L_\theta(X'_i, \hat{Y}'_i)\right).$$

Theorem S5 (Convex estimation, finite population)**.** *Let the labeled data be sampled uniformly at random from a finite population. Suppose that the convex estimation problem is nondegenerate* (3)*. Fix* $\alpha \in (0,1)$*. Suppose that, for any* $\theta \in \mathbb{R}^p$*, we can construct* $\mathcal{R}_\theta(\alpha)$ *satisfying*

$$P\left(\bar{\mathbf{\Delta}}_\theta \in \mathcal{R}_\theta(\alpha)\right) \geq 1-\alpha.$$

Let $\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta : -\frac{1}{N}\sum_{i=1}^N \nabla L_\theta(X_i', \hat{Y}_i') \in \mathcal{R}_\theta(\alpha)\right\}$*. Then,*

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1-\alpha.$$

We apply Theorem S5 in the context of mean estimation, quantile estimation, logistic regression, and linear regression. The target estimand θ^* is defined as in (2) with the loss function chosen appropriately. We remark that, just like in the i.i.d. case, the analysis for linear regression follows a more refined approach, as in the proof of Corollary S3.

Corollary S14 (Mean estimation, finite population)**.** *Let* θ^* *be the mean outcome. Fix* $\alpha \in (0,1)$*. Suppose that, for any* $\theta \in \mathbb{R}$*, we can construct an interval* $(\mathcal{R}^l(\alpha), \mathcal{R}^u(\alpha))$ *such that* $P\left(\bar{\mathbf{\Delta}} \in (\mathcal{R}^l(\alpha), \mathcal{R}^u(\alpha))\right) \geq 1-\alpha$*, where*

$$\bar{\mathbf{\Delta}} = \frac{1}{N}\sum_{i=1}^N \left(\hat{Y}_i' - Y_i'\right).$$

Let

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\frac{1}{N}\sum_{i=1}^N \hat{Y}_i' - \mathcal{R}^u(\alpha), \frac{1}{N}\sum_{i=1}^N \hat{Y}_i' - \mathcal{R}^l(\alpha)\right).$$

Then,

$$P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1-\alpha.$$

Corollary S15 (Quantile estimation, finite population)**.** *Let* θ^* *be the* q*-quantile. Fix* $\alpha \in (0,1)$*. Suppose that, for any* $\theta \in \mathbb{R}$*, we can construct an interval* $(\mathcal{R}_\theta^l(\alpha), \mathcal{R}_\theta^u(\alpha))$ *such that* $P\left(\bar{\mathbf{\Delta}}_\theta \in (\mathcal{R}_\theta^l(\alpha), \mathcal{R}_\theta^u(\alpha))\right) \geq 1-\alpha$*, where*

$$\bar{\mathbf{\Delta}}_\theta = \frac{1}{N}\sum_{i=1}^N \left(\mathbb{1}\left\{Y_i' \leq \theta\right\} - \mathbb{1}\left\{\hat{Y}_i' \leq \theta\right\}\right).$$

Let

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta \in \mathbb{R} : \frac{1}{N}\sum_{i=1}^N \mathbb{1}\left\{\hat{Y}_i' \leq \theta\right\} \in \left(q - \mathcal{R}_\theta^u(\alpha), q - \mathcal{R}_\theta^l(\alpha)\right)\right\}.$$

Then,

$$P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1-\alpha.$$

Corollary S16 (Linear regression, finite population)**.** *Let* θ^* *be the linear regression solution. Fix* $\alpha \in (0,1)$*. Suppose that we can construct* $\mathcal{R}^l(\alpha), \mathcal{R}^u(\alpha) \in \mathbb{R}^d$ *such that* $P(\bar{\mathbf{\Delta}}_j \in (\mathcal{R}_j^l(\alpha), \mathcal{R}_j^u(\alpha)), \forall j \in [d]) \geq 1-\alpha$*, where*

$$\bar{\mathbf{\Delta}} = X'^{\dagger}(\hat{Y}' - Y').$$

Let

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left(X'^{\dagger}\hat{Y}' - \mathcal{R}^u(\alpha), X'^{\dagger}\hat{Y}' - \mathcal{R}^l(\alpha)\right).$$

Then,

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1-\alpha.$$

Corollary S17 (Logistic regression, finite population)**.** *Let* θ^* *be the logistic regression solution. Fix* $\alpha \in (0,1)$*. Suppose that we can construct* $\mathcal{R}^l(\alpha), \mathcal{R}^u(\alpha) \in \mathbb{R}^d$ *such that* $P(\bar{\mathbf{\Delta}}_j \in (\mathcal{R}_j^l(\alpha), \mathcal{R}_j^u(\alpha)), \forall j \in [d]) \geq 1-\alpha$*, where*

$$\bar{\mathbf{\Delta}} = \frac{1}{N}\sum_{i=1}^N X_i'(\hat{Y}_i' - Y_i').$$

Let

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta \in \mathbb{R}^d : \frac{1}{N}\sum_{i=1}^{N} X'_{i,j}\left(\hat{Y}'_i - \frac{1}{1+\exp(-\theta^\top X'_i)}\right) \in \left(\mathcal{R}_j^l(\alpha), \mathcal{R}_j^u(\alpha)\right), \forall j \in [d]\right\}.$$

Then,

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geq 1-\alpha.$$

Confidence Intervals for the Mean

We give an overview of off-the-shelf confidence intervals for the mean. We state the results for two observation models: first for the i.i.d. sampling model considered in the main body and then for the finite-population setting discussed in the SM. In both cases, we provide a construction with nonasymptotic guarantees and one with asymptotic guarantees.

For the nonasymptotic confidence intervals, we rely on the results of Waudby-Smith and Ramdas (*32*), specifically their Theorem 3 and Theorem 4. We opt for these results because of their strong practical performance, which is primarily driven by variance adaptivity. These results assume that the observed random variables are bounded within a known interval. Without loss of generality we assume that the observations are bounded in $[0,1]$ (otherwise we can always normalize the observations to $[0,1]$).

For the asymptotic confidence intervals, we rely on the central limit theorem (CLT) and its variant for sampling without replacement; see (*37, 38*) for classical references.

Inference with i.i.d. Samples

In the following two results, assume that we observe $Z_1,\dots,Z_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and let $\mu = \mathbb{E}[Z_1]$.

Proposition S1 (Nonasymptotic CI: Theorem 3 in (*32*))**.** *Assume supp*$(\mathbb{P}) \subseteq [0,1]$. *Let*

$$\hat{\mu}_t = \frac{0.5+\sum_{j=1}^t Z_j}{t+1}, \quad \hat{\sigma}_t^2 = \frac{0.25+\sum_{j=1}^t (Z_j-\hat{\mu}_t)^2}{t+1}, \quad \lambda_t = \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{t-1}^2}}.$$

For every $m \in [0,1]$, *define the supermartingale:*

$$M_t(m) = \frac{1}{2}\max\left\{\prod_{j=1}^t\left(1+\min\left(\lambda_j, \frac{0.5}{m}\right)(Z_j-m)\right), \prod_{j=1}^t\left(1-\min\left(\lambda_j, \frac{0.5}{1-m}\right)(Z_j-m)\right)\right\}.$$

Let

$$\mathcal{C} = \bigcap_{t=1}^n \left\{m \in [0,1] : M_t(m) < 1/\alpha\right\}.$$

Then,

$$P\left(\mu \in \mathcal{C}\right) \geq 1-\alpha.$$

Intuitively, the supermartingale $M_t(m)$ should be thought of as the amount of evidence against m being the true mean. That is, $M_t(m)$ being big suggests that m is unlikely to be the true mean, so the final confidence set is the collection of all m for which the amount of such evidence is small.

For large n, computing the intersection in the definition of $\mathcal{C}$ can be intractable, so we conservatively choose a subsequence of $1,\dots,n$ for the computation.

Proposition S2 (Asymptotic CI: CLT interval)**.** *Assume* $\mathbb{P}$ *has a finite second moment. Let*

$$\mathcal{C} = \left(\frac{1}{n}\sum_{i=1}^n Z_i \pm z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right),$$

where $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (Z_i - \frac{1}{n}\sum_{j=1}^n Z_j)^2}$. *Then,*

$$\liminf_{n\to\infty} P\left(\mu \in \mathcal{C}\right) \geq 1-\alpha.$$

Inference on a Finite Population

In the following two results, we assume that there exists a *fixed* sequence $Z_1, \dots, Z_N$, and we observe $\{Z_i : i \in \mathcal{I}\}$, where $\mathcal{I} = \{i_1, \dots, i_n\}$ is a uniform random subset of $[N]$ with cardinality n. We let $\mu = \frac{1}{N}\sum_{i=1}^N Z_i$. For the asymptotic result, we assume that $Z_1, \dots, Z_N$ is the first N entries of an infinite underlying sequence $Z_1, Z_2, \dots$.

Proposition S3 (Nonasymptotic CI: Theorem 4 in (*32*))**.** *Assume $Z_i \in [0,1],\ i \in [N]$. Let*

$$\hat{\mu}_t = \frac{0.5 + \sum_{j=1}^t Z_{i_j}}{t+1}, \quad \hat{\sigma}_t^2 = \frac{0.25 + \sum_{j=1}^t (Z_{i_j} - \hat{\mu}_t)^2}{t+1}, \quad \lambda_t = \sqrt{\frac{2\log(2/\alpha)}{n\hat{\sigma}_{t-1}^2}}.$$

For every $m \in [0,1]$, define the supermartingale:

$$M_t(m) = \frac{1}{2}\max\left\{\prod_{j=1}^t\left(1 + \min\left(\lambda_j, \frac{0.5}{\mu_t(m)}\right)(Z_{i_j} - \mu_t(m))\right), \prod_{j=1}^t\left(1 - \min\left(\lambda_j, \frac{0.5}{1-\mu_t(m)}\right)(Z_{i_j} - \mu_t(m))\right)\right\},$$

where $\mu_t(m) = \frac{Nm - \sum_{j=1}^{t-1} Z_{i_j}}{N-t+1}$ is the putative mean. Let

$$\mathcal{C} = \bigcap_{t=1}^n \{m \in [0,1] : M_t(m) < 1/\alpha\}.$$

Then,

$$P(\mu \in \mathcal{C}) \geq 1 - \alpha.$$

Proposition S4 (Asymptotic CI: CLT for sampling without replacement)**.** *Let $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (Z_i - \mu)^2$, and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i \in \mathcal{I}}(Z_i - \hat{\mu})^2$. Assume that μ and σ have a limit and that $n/N \to p$ for some $p \in (0,1)$. Let*

$$\mathcal{C} = \left(\frac{1}{n}\sum_{i \in \mathcal{I}} Z_i \pm z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right).$$

Then,

$$\liminf_{n,N \to \infty} P(\mu \in \mathcal{C}) \geq 1 - \alpha.$$

Further Experimental Particulars

Relating Protein Structure and Post-Translational Modifications

The predictive model of whether a sequence position is in an IDR, f, is a logistic regression model that maps the relative solvent-accessible surface area (RSA) of each position, computed based on the AlphaFold-predicted structure using Bio.PDB (*39*), to a probability that the position is in an IDR. Following Bludau et al. (*3*), the RSA was locally smoothed with a window of $5, 10, 15, 20, 25, 30$, or 35 amino acids, and a sigmoid function was used to predict disorder from this smoothed RSA quantity. To fit the sigmoid, we used the data in (*3*) that had disorder labels but no PTM labels. The smoothing window size used for the final model was the value that resulted in the lowest variance of the bias, $Y - f$, on this data.

Galaxy Classification

We fine-tune a ResNet50 (*40*) on the training split of the Galaxy Zoo 2 data with a batch size of 32 and a learning rate of 0.0001 using Adam (*41*). We tune the entire backbone, not just the last layer. We use the remaining validation split as our labeled and unlabeled data, taking $n = [50, 100, 200, 300, 500, 750, 1000]$. We use Algorithm S1 for the naive prediction-powered approach, and Proposition S2 for the classical and imputation approaches.

Distribution of Gene Expression Levels

We used the transformer model developed and trained by Vaishnav et al. (*11*) to predict gene expression level, with the following modification that we found improved predictive performance. Given n labeled data points, five were randomly selected and used to train an affine (two-parameter) function mapping the scalar prediction of the transformer in (*11*) to a prediction of the conditional median of the label, using quantile regression. The predictions of this final model were used for the unlabeled dataset, and the remaining $n - 5$ data points that weren't used to fine-tune the transformer model were used as the labeled dataset. We use Algorithm S3 to form the prediction-powered confidence intervals and the standard quantile CLT confidence interval for the classical and imputation approaches. Results analogous to Fig. 2C in the main text for the 0.25- and 0.75-quantiles are plotted in Fig. S6.

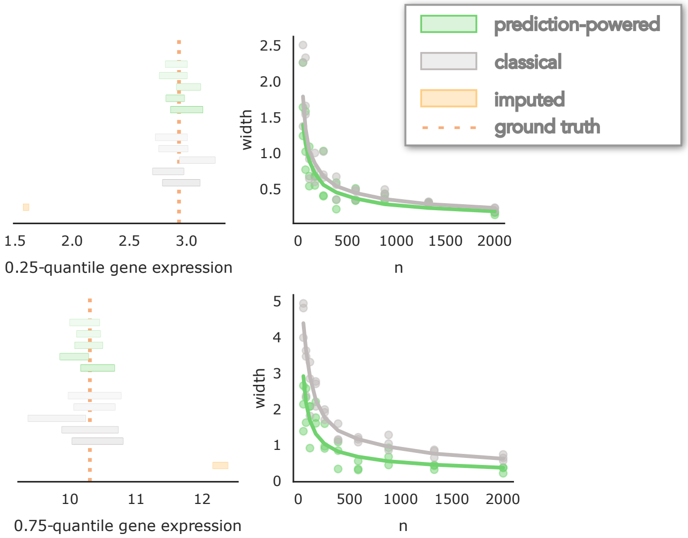

Figure S6: **Confidence intervals on gene expression quantiles** for $q = 0.25$ (top) and $q = 0.75$ (bottom). Following Fig. 2 in the main text, the left panel shows prediction-powered (green) and classical (gray) confidence intervals computed with five random splits of labeled and unlabeled data. The right panel shows the average interval width for varying values of n, the number of labeled data points, as well as the width for five randomly chosen trials.

Estimating Deforestation in the Amazon

The machine-learning model given by (*13*) outputs forest-cover predictions at 30m resolution for 3192 data points. We correspond these by latitude and longitude with gold-standard data points labeled as one of $\{$`deforestation, no deforestation`$\}$ from (*12*). In the first step, we split off half of the data to train a histogram-based gradient-boosted tree to predict deforestation labels from the forest-cover predictions. We take a random sample of $n = 100$ data points as the gold-standard data, and try to cover the true fraction of deforestation events on the $N = 1596$ remaining data points. We use Algorithm S1 to produce the prediction-powered confidence interval and Proposition S2 for the classical and imputation approaches.

Relationship Between Income and Private Health Insurance

We train a gradient-boosted tree (*15*) on the California Census data from 2018 acquired using the Folktables (*14*) interface. The boosted tree takes as input several covariates such as income, race, and sex, to predict whether an individual has private health insurance coverage. In the new year, 2019, we use $n = [200, 300, 500, 1000, 2000, 5000, 10000]$ labeled data points. We use Algorithm S5 to produce the prediction-powered confidence interval and the standard CLT confidence interval for the classical and imputation approaches.

Relationship Between Age and Income

The setting is the same as the above experiment on income and private health insurance, the main difference being that income is used as the target, and not as a covariate. We used Algorithm S4 to produce the prediction-powered confidence interval and the standard CLT confidence interval for the classical and imputation approaches.

Counting Plankton

We fine-tune a ResNet152 (*40*) on the WHOI-Plankton dataset (*17*) in the years 2006-2013 for two epochs with a batch size of 32 and a learning rate of 0.0001 using AdamW (*42*), with 5% of the data saved for validation. We tune the entire backbone, not just the last layer. Then we test in the year 2014, using all available data. We use Theorem S3 to produce the prediction-powered intervals and Proposition S2 for the imputation approach.

Related Work

This section expands the discussion of related work from the main body of the paper. Thematically, prediction-powered inference is most similar to the work of Wang et al. (*18*), who also introduce a method to correct machine-learning predictions for the purpose of subsequent inference. However, prediction-powered inference provides provably valid conclusions under minimal assumptions about the data-generating distribution, whereas the procedure of Wang et al. does not provide coverage in general and requires strong assumptions about the relationship between the prediction model and the true response. We compare against this baseline in "Comparison to Baseline Procedures" in the SM.

The technical results of this paper generalize tools from the model-assisted survey sampling literature (*43*), which provides methods to improve inference from surveys in the presence of auxiliary information. In particular, the prediction-powered mean estimator is the difference estimator, closely related to generalized regression estimators (*44*). It has long been recognized that model predictions can be leveraged as auxiliary data (*45*), and much work has gone into producing asymptotically valid confidence intervals when the predictive model is fit on the same data that is used for inference—see (*46*) for a recent overview. Prediction-powered inference is also related to the statistical literature on semiparametric inference, missing data, and multiple imputation (*47*). In particular, (*48-51*) study regression with missing data. The rectifier resembles debiasing strategies that are pervasive in this literature, an example being the AIPW estimator (*49*). Likewise, prediction-powered inference is related to measurement error (e.g., *52, 53*). Prediction-powered inference aims to provide simple, broadly applicable algorithms using similar debiasing tricks, while allowing the use of state-of-the-art black-box machine-learning systems.

There has been an increasing a body of work on estimation with many unlabeled data points and few labeled data points (*19-22*), focusing on efficiency in semiparametric or high-dimensional regimes. In particular, Chakrabortty and Cai (*23*), Deng et al. (*25*), and Azriel et al. (*26*) study efficient estimation of linear regression parameters, Chakrabortty et al. (*27*) study efficient quantile estimation, and Zhang and Bradic (*24*) study mean estimation in a high-dimensional setting. Finally, Song et al. (*55*) study M-estimation, using a projection-based correction to the classical M-estimator loss based on simple statistics (e.g. low-order polynomials) of the features. Prediction-powered inference continues in this vein but focuses on the setting where the scientist has access to a good predictive model fit on separate data and makes no assumptions about the model (such as consistency). The confidence intervals and resulting p-values from previous work rely on asymptotic approximations, while prediction-powered inference has both asymptotic and nonasymptotic variants. Furthermore, prediction-powered inference goes beyond random sampling and considers certain forms of distribution shift.

More distantly, the setting of prediction-powered inference, in which the scientist has access to some labeled data alongside unlabeled data, also appears in semi-supervised learning (*54*)—this literature studies the question of how to improve prediction accuracy with unlabeled data. Even further along is the literature on transfer learning, wherein estimation rates improve with access to out-of-distribution data; a similar debiasing technique also appears there (*56*).

Prediction-powered inference is conceptually related to conformal prediction (*28*). Both methodologies leverage a predictive model and a labeled dataset. From this point on, the two methods diverge: prediction-powered inference has additional unlabeled data and gives a confidence set that contains a population-level quantity such as the mean outcome; conformal prediction gives a confidence set for a test instance that contains the true label. These are two distinct goals, and neither method can be applied straightforwardly to solve the objective of the other.

References and Notes

1. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi:10.1038/s41586-021-03819-2 Medline

2. K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021). doi:10.1038/s41586-021-03828-1 Medline

3. I. Bludau, S. Willems, W.-F. Zeng, M. T. Strauss, F. M. Hansen, M. C. Tanzer, O. Karayel, B. A. Schulman, M. Mann, The structural context of posttranslational modifications at a proteome-wide scale. *PLOS Biol.* **20**, e3001636 (2022). doi:10.1371/journal.pbio.3001636 Medline

4. I. Barrio-Hernandez, J. Yeo, J. Jänes, T. Wein, M. Varadi, S. Velankar, P. Beltrao, M. Steinegger, Clustering predicted structures at the scale of the known protein universe. *bioRxiv* **2023-03** (2023).

5. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, ppi-py: A Python package for scientific discovery using machine learning, Zenodo (2023); https://doi.org/10.5281/zenodo.8403931.

6. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, Prediction-Powered Inference: Data Sets, Zenodo (2023); https://doi.org/10.5281/zenodo.8397451.

7. L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O’Connor, J. G. Sikes, Z. Obradovic, A. K. Dunker, The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037–1049 (2004). doi:10.1093/nar/gkh253 Medline

8. UniProt Consortium, UniProt: A hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015). doi:10.1093/nar/gku989 Medline

9. K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, D. Thomas, Galaxy Zoo 2: Detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* **435**, 2835–2860 (2013). doi:10.1093/mnras/stt1458

10. D. G. York, J. Adelman, J. E. Anderson Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. Fan,

G. R. Federwitz, L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C. Huang, C. Hull, D. Husby, S.-I. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Kent, R. S. J. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider, M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, S. Wang, M. Watanabe, D. H. Weinberg, B. Yanny, N. Yasuda, N. Yasuda, The Sloan digital sky survey: Technical summary. *Astron. J.* **120**, 1579–1587 (2000). doi:10.1086/301513

11. E. D. Vaishnav, C. G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis, D. A. Thompson, J. Z. Levin, F. A. Cubillos, A. Regev, The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455–463 (2022). doi:10.1038/s41586-022-04506-6 Medline

12. E. L. Bullock, C. E. Woodcock, C. Souza Jr., P. Olofsson, Satellite-based estimates reveal widespread forest degradation in the Amazon. *Glob. Chang. Biol.* **26**, 2956–2969 (2020). doi:10.1111/gcb.15029 Medline

13. J. O. Sexton, J. X.-P. Song, M. Feng, P. Noojipady, A. Anand, C. Huang, D.-H. Kim, K. M. Collins, S. Channan, C. DiMiceli, J. R. Townshend, Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error. *Int. J. Digit. Earth* **6**, 427–448 (2013). doi:10.1080/17538947.2013.786146

14. F. Ding, M. Hardt, J. Miller, L. Schmidt, “Retiring adult: New datasets for fair machine learning” in *Advances in Neural Information Processing Systems* **34** (2021), pp. 6478–6490.

15. T. Chen, C. Guestrin, “XGBoost: A scalable tree boosting system” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

16. R. J. Olson, A. Shalapyonok, H. M. Sosik, An automated submersible flow cytometer for analyzing pico-and nanophytoplankton: FlowCytobot. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **50**, 301–315 (2003). doi:10.1016/S0967-0637(03)00003-7

17. E. C. Orenstein, O. Beijbom, E. E. Peacock, H. M. Sosik, WHOI-Plankton- A large scale fine grained visual recognition benchmark dataset for plankton classification. arXiv:1510.00745 [cs.CV] (2015).

18. S. Wang, T. H. McCormick, J. T. Leek, Methods for correcting inference based on outcomes predicted by machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30266–30275 (2020). doi:10.1073/pnas.2001238117 Medline

19. M. S. Pepe, Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–365 (1992). doi:10.1093/biomet/79.2.355

20. J. Lafferty, L. Wasserman, "Statistical analysis of semi-supervised regression" in *Advances in Neural Information Processing Systems* **20** (2007), pp. 801–808.

21. A. Zhang, L. D. Brown, T. T. Cai, Semi-supervised inference: General theory and estimation of means. *Ann. Stat.* **47**, 2538–2566 (2019). doi:10.1214/18-AOS1756

22. A. Chakrabortty, G. Dai, E. Tchetgen Tchetgen, A general framework for treatment effect estimation in semi-supervised and high dimensional settings. arXiv:2201.00468 [stat.ME] (2022).

23. A. Chakrabortty, T. Cai, Efficient and adaptive linear regression in semi-supervised settings. *Ann. Stat.* **46**, 1541–1572 (2018). doi:10.1214/17-AOS1594

24. Y. Zhang, J. Bradic, High-dimensional semi-supervised learning: In search of optimal inference of the mean. *Biometrika* **109**, 387–403 (2022). doi:10.1093/biomet/asab042

25. S. Deng, Y. Ning, J. Zhao, H. Zhang, Optimal and safe estimation for high-dimensional semi-supervised learning. arXiv:2011.14185 [stat.ME] (2020).

26. D. Azriel, L. D. Brown, M. Sklar, R. Berk, A. Buja, L. Zhao, Semi-supervised linear regression. *J. Am. Stat. Assoc.* **117**, 2238–2251 (2022). doi:10.1080/01621459.2021.1915320

27. A. Chakrabortty, G. Dai, R. J. Carroll, Semi-supervised quantile estimation: robust and efficient inference in high dimensional settings. arXiv:2201.10208 [stat.ME] (2022).

28. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World* (Springer, 2005), vol. 5.

29. A. Buja, L. Brown, R. Berk, E. George, E. Pitkin, M. Traskin, K. Zhang, L. Zhao, Models as approximations I: Consequences illustrated with linear regression. *Stat. Sci.* **34**, 523–544 (2019). doi:10.1214/18-STS693

30. H. White, Using least squares to approximate unknown regression functions. *Int. Econ. Rev.* **21**, 149–170 (1980). doi:10.2307/2526245

31. H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838 (1980). doi:10.2307/1912934

32. I. Waudby-Smith, A. Ramdas, Estimating means of bounded random variables by betting. *J. R. Stat. Soc. Series B Stat. Methodol.* qkad009 (2023). doi:10.1093/jrsssb/qkad009

33. Z. Lipton, Y.-X. Wang, A. Smola, "Detecting and correcting for label shift with black box predictors" in *International Conference on Machine Learning* (2018), pp. 3122–3130.

34. A. Dvoretzky, J. Kiefer, J. Wolfowitz, Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27**, 642–669 (1956). doi:10.1214/aoms/1177728174

35. P. Massart, The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18**, 1269–1283 (1990). doi:10.1214/aop/1176990746

36. C. L. Canonne, A short note on learning discrete distributions. arXiv:2002.11457 [math.ST] (2020).

37. P. Erdos, On the central limit theorem for samples from a finite population. *Publ Math. Inst. Hung. Acad. Sci.* **4**, 49–61 (1959).

38. T. Höglund, Sampling from a finite population. A remainder term estimate. *Scand. J. Stat.* **5**, 69–71 (1978).

39. T. Hamelryck, B. Manderick, PDB file parser and structure class implemented in Python. *Bioinformatics* **19**, 2308–2310 (2003). doi:10.1093/bioinformatics/btg299 Medline

40. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

41. D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization" in International Conference on Learning Representations (2014).

42. I. Loshchilov, F. Hutter, "Decoupled Weight Decay Regularization" in *International Conference on Learning Representations* (2018).

43. C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling* (Springer Science & Business Media, 1992).

44. C. M. Cassel, C. E. Särndal, J. H. Wretman, Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620 (1976). doi:10.1093/biomet/63.3.615

45. C. Wu, R. R. Sitter, A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Stat. Assoc.* **96**, 185–193 (2001). doi:10.1198/016214501750333054

46. F. J. Breidt, J. D. Opsomer, Model-assisted survey estimation with modern prediction techniques. *Stat. Sci.* **32**, 190–205 (2017). doi:10.1214/16-STS589

47. R. J. Little, D. B. Rubin, *Statistical Analysis with Missing Data* (Wiley, 2019), vol. 793.

48. J. M. Robins, A. Rotnitzky, L. P. Zhao, Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **89**, 846–866 (1994). doi:10.1080/01621459.1994.10476818

49. J. M. Robins, A. Rotnitzky, Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* **90**, 122–129 (1995). doi:10.1080/01621459.1995.10476494

50. J. Chen, N. E. Breslow, Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. *Can. J. Stat.* **32**, 359–372 (2004). doi:10.2307/3316021

51. M. Yu, B. Nan, A revisit of semiparametric regression models with missing data. *Stat. Sin.* **2006**, 1193–1212 (2006).

52. R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective* (Chapman and Hall/CRC, 2006).

53. X. Chen, H. Hong, E. Tamer, Measurement error models with auxiliary data. *Rev. Econ. Stud.* **72**, 343–366 (2005). doi:10.1111/j.1467-937X.2005.00335.x

54. X. Zhu, A. B. Goldberg, *Introduction to Semi-Supervised Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3 (2009), pp. 1–130.

55. S. Song, Y. Lin, Y. Zhou, A general M-estimation theory in semi-supervised framework. *J. Am. Stat. Assoc.*, 1–11 (2023). doi:10.1080/01621459.2023.2169699

56. Y. Tian, Y. Feng, Transfer learning under high-dimensional generalized linear models. *J. Am. Stat. Assoc.*, 1–14 (2022). doi:10.1080/01621459.2022.2071278