Calibrated Propensities for Causal Effect Estimation

Anonymous Authors[1]

Abstract

Propensity scores are commonly used to balance observed confounders while estimating treatment effects. When the confounders are high-dimensional or unstructured, the learned propensity scores can be miscalibrated and ineffective in the correction of confounding. We argue that the probabilistic output of a learned propensity score model should be calibrated, i.e. predictive treatment probability of 90% should correspond to 90% individuals being assigned the treatment group. We investigate the theoretical properties of a calibrated propensity score model and its role in unbiased treatment effect estimation. We demonstrate improved causal effect estimation with calibrated propensity scores in several tasks including high-dimensional genome-wide association studies.

1. Introduction

This paper studies the problem of inferring the causal effect of an intervention from observational data. For example, consider the problem of estimating the effect of a treatment on a medical outcome or the effect of a genetic mutation on a phenotype. A key challenge in this setting is confounding—e.g., if a treatment is only given to sick patients, it may appear to spuriously correlate with worse outcomes [8; 35]. Propensity score methods are a popular tool for correcting for confounding in observational data [31; 3; 35; 17; 36], and have been used to balance high-dimensional, unstructured covariates [29; 36; 37]. However, neural approximators of propensity score conditional on high-dimensional covariates may output volatile and miscalibrated probabilities close to 0 or 1 [11], thus making the propensity score methods unreliable [12; 18]. For example, when the propensity score model is overconfident (a known problem with neural network estimators [9]), predicted assignment probabilities can be too small [34], which yields a blow-up in the estimated causal effects. More generally, propensity score weighting stands to benefit from accurate uncertainty quantification [11].

This work argues that propensity score methods can be improved by leveraging calibrated uncertainty estimation in treatment assignment models. Intuitively, when a calibrated model outputs a treatment probability of 90%, then 90% of individuals with that prediction should be assigned to the treatment group [26; 14]. We argue that calibration is a necessary condition for propensity score models that also addresses the aforementioned problems of model overconfidence. This paper makes the following contributions: (1) we provide formal arguments that explain the benefits of uncertainty calibration in propensity score models; (2) we propose simple algorithms that enforce calibration; (3) we provide theoretical guarantees on the calibration and regret of these algorithms and we demonstrate their effectiveness in genome-wide association studies.

2. Background

Notation Formally, we are given an observational dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^{n}$ consisting of n units, each characterized by features $x^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, a binary treatment $t^{(i)} \in \{0, 1\}$, and a scalar outcome $y^{(i)} \in \mathcal{Y} \subseteq \mathbb{R}$. We assume $\mathcal{D}$ consists of i.i.d. realizations of random variables $X, Y, T \sim P$ from a data distribution P. Although we assume binary treatments and scalar outcomes, our approach naturally extends beyond this setting. The feature space $\mathcal{X}$ can be any continuous or discrete set.

2.1. Causal effect estimation using propensity scoring

We seek to estimate the true effect of $T = t$ in terms of its average treatment effect (ATE).

$$Y[x, t] = \mathbb{E}[Y|X = x, \text{do}(T = t)] \quad \text{ATE} = \mathbb{E}[Y[x, 1] - Y[x, 0]],$$

where $\text{do}(\cdot)$ denotes an intervention [25]. We assume strong ignorability, i.e., $(Y(0), Y(1)) \perp T|X$ and $0 < P(T|X) < 1$, for all $X \in \mathcal{X}, T \in \{0, 1\}$, where $Y(0)$ and $Y(1)$ denote potential outcomes. We also make the stable unit treatment value assumption (SUTVA) [31]. Under these assumptions, the propensity score defined as $e(X) = P(T = 1|X)$ satisfies the conditional independence $(Y(0), Y(1)) \perp T|e(X)$ [31]. Thus, ATE can be

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

expressed as $\tau = \mathbb{E}\left(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right)$. We define the Inverse Propensity Treatment Weight (IPTW) and Augmented Inverse Propensity Weight (AIPW) estimators for ATE in Appendix A.

2.2. Calibrated prediction for uncertainty estimation

This paper seeks to evaluate and improve the uncertainty of propensity scores. We say that a propensity score model Q is calibrated if the true probability of $T = 1$ conditioned on predicting a probability p matches the predicted probability, i.e., $P(T = 1 \mid Q(T = 1|X) = p) = p \ \ \forall p \in [0, 1]$.

3. Calibrated propensity scores

We start with the observation that a learned propensity scoring model $Q(T|X)$ must not only correctly output the treatment assignment, but also accurately estimate predictive uncertainty. Specifically, the *probability* of the treatment assignment must be correct, not just the class assignment. While a Bayes optimal Q will perfectly estimate uncertainty, suboptimal models will need to balance various aspects of predictive uncertainty, such as calibration and sharpness.

3.1. Calibration: A necessary condition for propensity scoring model

This paper argues that calibration improves propensity-scoring methods. Intuitively, if the model $Q(T = 1|X)$ predicts a treatment assignment probability of 90%, then 90% of these predictions should receive the treatment. If the prediction is larger or smaller, the downstream IPTW estimator will overcorrect or undercorrect for the biased treatment allocation. In other words, calibration is a *necessary condition* for a correct propensity scoring model. We formalize this intuition below.

Theorem 3.1. *When $Q(T|X)$ is not calibrated, there exists an outcome function such that an IPTW estimator based on Q yields an incorrect estimate of the true causal effect almost surely.*

Please refer to Appendix H.2 for a full proof.

3.2. Calibrated uncertainties improve propensity scoring models

We identify settings in which calibration is either sufficient or prevents common failure modes of IPTW estimators. Specifically, we identify and study two such regimes: (1) accurate but over-confident propensity scoring models (e.g., neural networks [9]); (2) high-variance IPTW estimators that take as input numerically small propensity scores.

3.2.1. Bounding the error of causal effect estimation using proper scores

Firstly, we relate the error of an IPTW estimator to the difference between a model $Q(T|X)$ and the true $P(T|X)$. We define $\pi_{t,y}(Q) = \sum_x P(y|x,t)\frac{P(t|x)}{Q(t|x)}P(x)$ as the estimated probability of y given t when using propensity score model Q. It is not hard to show that the true $Y[t] := \mathbb{E}_X Y[X,t] = \mathbb{E}_X \mathbb{E}[Y|X = x, \mathrm{do}(T = t)]$ can be written as $\sum_y y\pi_{y,t}(P)$ (see Appendix H.3). Similarly, the estimate of an IPTW estimator with propensity model Q in the limit of infinite data tends to $\hat{Y}_Q[1] - \hat{Y}_Q[0]$, where $\hat{Y}_Q[t] := \sum_y y\pi_{y,t}(Q)$. We may bound the expected L1 ATE error $|Y[1] - Y[0] - (\hat{Y}_Q[1] - \hat{Y}_Q[0])|$ by $\sum_t |Y[t] - \hat{Y}_Q[t]| \leq \sum_t \sum_y |y| \cdot |\pi_{y,t}(P) - \pi_{y,t}(Q)|$.

Our first lemma bounds the error $|\pi_{y,t}(P) - \pi_{y,t}(Q)|$ as a function of the difference between $Q(T|X)$ and the true $P(T|X)$. A bound on the ATE error follows from a simple corollary.

Lemma 3.2. *The expected error $|\pi_{y,t}(P) - \pi_{y,t}(Q)|$ induced by an IPTW estimator with propensity score model Q is bounded as*

$$|\pi_{y,t}(P) - \pi_{y,t}(Q)| \leq \mathbb{E}_{X\sim R_{y,t}}[\ell_\chi(P,Q)^{\frac{1}{2}}], \qquad (1)$$

where $R_{y,t} \propto P(Y = y|X, T = t)P(X)$ is a data distribution and $\ell_\chi(Q,P) = \left(1 - \frac{P(T=t|X)}{Q(T=t|X)}\right)^2$ is the chi-squared loss between the true propensity score and the model Q.

Proof (Sketch). Note that $|\pi_{y,t}(P) - \pi_{y,t}(Q)| \leq \mathbb{E}_{X\sim R_{y,t}}\left|1 - \frac{P(T=t|X)}{Q(T=t|X)}\right| \leq \mathbb{E}_{R_{y,t}}\ell_\chi(P,Q)^{\frac{1}{2}}$ □

See Appendix H.3.1 for the full proof.

Corollary 3.3. *Let $|y| \leq K$ for all $y \in \mathcal{Y}$. The error of an IPTW estimator with propensity score model Q is bounded by $2|\mathcal{Y}|K \max_{y,t} \mathbb{E}_{R_{y,t}}\ell_\chi(P,Q)^{\frac{1}{2}}$.*

Note that ℓ_χ is a type of proper loss or proper scoring rule: it is small only if Q correctly captures the probabilities in P. A model that is accurate, but that does not output correct probability will have a large ℓ_χ; conversely, when $Q = P$, the bound equals to zero and the IPTW estimator is perfectly accurate. To the best of our knowledge, this is the first bound that relates the accuracy of an IPTW estimator directly to the quality of uncertainties of the probabilistic model Q.

3.2.2. Calibration reduces variance of inverse probability estimators

A common failure mode of IPTW estimators arises when the probabilities from a propensity scoring model $Q(T|X)$ are small or even equal to zero—division by $Q(T|X)$ then causes the IPTW estimator to take on very large values or

be undefined. Furthermore, when $Q(T|X)$ is small, small changes in its value cause large changes in the IPTW estimator, which induces problematically high variance. Here, we show that calibration can help mitigate this failure mode. If Q is calibrated, then it cannot take on abnormally small values relative to P.

Theorem 3.4. *Let P be the data distribution, and suppose that $1-\delta > P(T|X) > \delta$ for all T, X and let Q be a calibrated model relative to P. Then $1-\delta > Q(T|X) > \delta$ for all T, X as well.*

Proof (Sketch). The proof is by contradiction. Suppose $Q(T=1|x) = q$ for some x and $q < \delta$. Then because Q is calibrated, of the times when we predict q, we have $P(T=1|Q(T=1|X)=q) = q < \delta$, which is impossible since $P(T=1|x) > \delta$ for every x.

See Appendix H.3.2 for the full proof. □

3.2.3. Calibration improves causal effect estimation with accurate propensity models

Unfortunately, calibration by itself is not sufficient to correctly estimate treatment effects. For example, consider defining $Q(T|X)$ as the marginal $P(T)$: this Q is calibrated, but cannot accurately estimate treatment effects. However, if the model Q is sufficiently accurate (as might be the case with a powerful neural network), calibration becomes the missing piece for an accurate IPTW estimator. Specifically, we define separability, a condition which states that when $P(T|X_1) \neq P(T|X_2)$ for $X_1, X_2 \in \mathcal{X}$, then the model Q satisfies $Q(T|X_1) \neq Q(T|X_2)$. Intuitively, the model Q is able to discriminate between various T—something that might be achievable with an expressive neural Q that has high classification accuracy. We show that a model that is separable and also calibrated achieves accurate causal effect estimation.

Theorem 3.5. *The error of an IPTW estimator with propensity model Q tends to zero as $n \to \infty$ if:*

1. *Separability holds, i.e., $\forall X_1, X_2 \in \mathcal{X}, P(T|X_1) \neq P(T|X_2) \implies Q(T|X_1) \neq Q(T|X_2])$*
2. *The model Q is calibrated, i.e., $\forall q \in (0,1), P(T=1|Q(T=1|X)=q) = q$*

See Appendix H.3.3 for the proof. In Appendix B, we also show that a post-hoc recalibrated model Q' has vanishing regret $\ell(Q', Q)$ with respect to a base model Q and a proper loss ℓ (including ℓ_χ used in our calibration bound).

4. Algorithms for calibrated propensity scoring

4.1. A framework for calibrated propensity scoring

Algorithm 1 Calibrated Propensity Scoring

1. Split $\mathcal{D}$ into training set $\mathcal{D}'$ and calibration set $\mathcal{C}$
2. Train a propensity score model $Q(T|X)$ on $\mathcal{D}'$
3. Train recalibrator R over output of Q on $\mathcal{C}$
4. Apply IPTW with $R \circ Q$ as prop. score model

We propose Algorithm 1 to produce calibrated propensity scoring models; it differs from standard propensity scoring methods by the addition of a post-hoc recalibration step (step #3) [26; 14] after training the model Q. The recalibration step is outlined in Algorithm 3 (Appendix B). The key idea is to learn an auxiliary model $R : [0,1] \to [0,1]$ such that the joint model $R \circ H$ is calibrated. In Appendix B, we discuss the choice of model R and prove that if R can approximate the density $P(T=1|Q(T|X)=p)$, $R \circ Q$ will be calibrated [14; 13].

5. Empirical evaluation

Genome-Wide Association Studies (GWASs) attempt to estimate the treatment effect of genetic mutations (called SNPs) on individual traits (called phenotypes) from observational datasets. Each SNP acts as a treatment. Confounding occurs because of hidden ancestry: individuals with shared ancestry have correlated genes and phenotypes.

The key takeaways can be summarized as follows. First, recalibration enables off-the-shelf IPTW estimators to match or outperform a state-of-the-art GWAS analysis system (LLM/LIMIX; see Tables 1 and 8). Second, our method enables the use of propensity score models that would otherwise be unusable due to the poor quality of their uncertainty estimates (e.g., Naive Bayes; see Table 7). Third, leveraging new types of propensity score models that are fast to train (such as Naive Bayes), improves the speed of GWAS analysis by more than two-fold (see Table 2).

Setup We simulate the genotypes and phenotypes of individuals following a range of standard models as described in Appendix F. The outcome is simulated as $Y = \beta^T G + \alpha^T Z + \epsilon$, where G is the vector of SNPs, Z contains the hidden confounding variables, ϵ is noise distributed as Gaussian, β is the vector of treatment effects corresponding to each SNP and α holds coefficients for the hidden confounding variables. We assume that the aspect of hidden population structure in Z that needs to be controlled for is fully contained in the observed genetic data to ensure ignorability [19]. To estimate the average marginal treatment effect corresponding to each SNP, we iterate suc-

Table 1. GWAS with calibrated propensities. We compare IPTW and AIPW estimates using calibrated propensity scores against standard baselines and a specialized GWAS analysis system (LMM/LIMIX). Results averaged over 10 reps and std error in braces.

Dataset	Spatial (α=0.1)	Spatial (α=0.3)	Spatial (α=0.5)	HGDP	TGP
Naive	16.23 (0.91)	11.76 (0.84)	9.81 (0.69)	11.82 (0.11)	12.24 (0.71)
PCA	9.60 (0.37)	9.54 (0.41)	9.38 (0.38)	11.69 (0.20)	10.73 (0.38)
FA	9.55 (0.34)	9.53 (0.44)	9.23 (0.30)	11.65 (0.16)	10.59 (0.32)
LMM	10.24 (0.41)	9.58 (0.45)	**8.15 (0.40)**	**10.09 (0.35)**	**9.44 (0.57)**
IPTW (Calib)	**8.13 (0.35)**	**8.69 (0.56)**	**8.32 (0.34)**	10.86 (0.13)	**9.57 (0.58)**
IPTW (Plain)	12.56 (1.25)	10.22 (0.81)	9.09 (0.48)	11.62 (0.12)	11.76 (0.86)
AIPW (Calib)	8.94 (0.29)	9.00 (0.58)	8.59 (0.39)	11.06 (0.12)	10.32 (0.43)
AIPW (Plain)	13.89 (0.76)	10.46 (0.72)	8.99 (0.51)	11.38 (0.11)	11.56 (0.65)
Δ_{ECE}	0.022 (0.001)	0.016 (0.007)	0.015 (0.001)	0.011 (0.001)	0.022 (0.001)

cessively over the vector of SNPs such that the selected SNP is treatment T and all the remaining SNPs are covariates X for predicting the phenotypic outcome Y. We use logistic regression as propensity model and isotonic regression as recalibrator. We measure ε_{ATE} as the l_2 norm of the difference between true and estimated marginal treatment effect vectors. We evaluate the calibration of the propensity score model using expected calibration error (ECE) (Appendix C). We compare the performance of these estimators with standard methods to perform GWAS, including Principal Components Analysis (PCA) [27; 28], Factor Analysis (FA), and Linear Mixed Models (LMMs) [42; 20], implemented in the popular LIMIX library [21]. 1% of total SNPs are causal and we have 4000 individuals in the dataset.

Results. In Table 1, we demonstrate the effectiveness of estimators using calibrated propensities on five different GWAS datasets (Appendix F). Here, we have a total of 100 SNPs. In Table 8 (Appendix G), we increase the proportion of causal SNPs for the Spatial simulation and continue to see improved performance under calibration. In Table 7 (Appendix G), we compare different base models to learn propensity scores and show that calibration improves the performance in each case. We also see that the performance of plain Naive Bayes as the base propensity score model is very poor owing to the simplistic conditional independence assumptions, but calibration improves its performance significantly. In Table 2, we compare the computational throughput of calibrated Naive Bayes as the propensity score model with logistic regression. Here, we have a total of 1000 SNPs. We see that using calibrated Naive Bayes obtains performance competitive with logistic regression at a significantly higher throughput.

In Appendix D, we demonstrate several additional experiments on the effectiveness of calibrated propensity scores under varying treatment assignment functions and base propensity models. In Appendix E we evaluate calibrated propensities for image as an unstructured confounder.

Table 2. Calibrated Naive Bayes yields lower ϵ_{ATE} (IPTW) and uses lower computational resources as compared to logistic regression.

Method	ϵ_{ATE}	Tput (SNPs/sec)
LMM	19.908 (3.592)	-
Calibrated NB	**18.210 (1.705)**	47.6
Plain NB	1455.992 (185.084)	68.6
Calibrated LR	23.618 (3.832)	19.5
Plain LR	27.921 (4.713)	20.1

6. Related work

Calibrated uncertainties have been used to improve deep reinforcement learning [23; 13], natural language processing [16], Bayesian optimization [4], etc. Lenis et al. [18], Kang and Schafer [12] demonstrate the degradation in treatment effect estimation in response to misspecified treatment and outcome models. Different notions of calibration have been proposed to reduce the bias in treatment effect estimation by optimizing the covariate balancing property [10; 43; 24] and by correcting measurement error [33]. Our notion of calibration is easier to implement and does not require modification to the training of propensity model.

7. Conclusions

We proposed a simple technique to perform post-hoc calibration of the propensity score model. We show that calibration is a necessary condition to obtain accurate treatment effects and calibrated uncertainties improve propensity scoring models. We show improved treatment effect estimates for high-dimensional, unstructured covariates over a range of base models including the popular logistic regression. Calibration also allows us to utilize simpler models like Naive Bayes and obtain higher computational throughput while maintaining competitive performance for high-dimensional covariates.

References

[1] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[2] Anders Bergström, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, Hélène Blanché, Jean-François Deleuze, Howard Cann, Swapan Mallick, David Reich, Manjinder S. Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue, Richard Durbin, and Chris Tyler-Smith. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484): eaay5012, 2020. doi: 10.1126/science.aay5012. URL https://www.science.org/doi/abs/10.1126/science.aay5012.

[3] Ralph B D'Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.*, 17(19): 2265–2281, October 1998.

[4] Shachi Deshpande and Volodymyr Kuleshov. Calibrated uncertainty estimation improves bayesian optimization, 2023.

[5] Shachi Deshpande, Kaiwen Wang, Dhruv Sreenivas, Zheng Li, and Volodymyr Kuleshov. Deep multi-modal structural equations for causal effect estimation with unstructured proxies. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10931–10944. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/46e654963ca9f2b9ff05d1bbfce2420c-Paper-Conference.pdf.

[6] Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1): D941–D947, 10 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz836. URL https://doi.org/10.1093/nar/gkz836.

[7] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[8] Sander Greenland, Judea Pearl, and James M. Robins. Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29 – 46, 1999. doi: 10.1214/ss/1009211805. URL https://doi.org/10.1214/ss/1009211805.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[10] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *J. R. Stat. Soc. Series B Stat. Methodol.*, 76(1):243–263, January 2014.

[11] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020.

[12] Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), nov 2007. doi: 10.1214/07-sts227. URL https://doi.org/10.1214%2F07-sts227.

[13] Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11683–11693. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kuleshov22a.html.

[14] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression, 2018.

[15] Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, Carlos Soares, João Gama, and Alípio Jorge, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 68–85, Cham, 2015. Springer International Publishing. ISBN 978-3-319-23528-8.

[16] Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation, 2019.

[17] Stephanie T Lanza, Julia E Moore, and Nicole M Butera. Drawing causal inferences using propensity scores: a practical guide for community psychologists.

Am. J. Community Psychol., 52(3-4):380–392, December 2013.

[18] David Lenis, Benjamin Ackerman, and Elizabeth A Stuart. Measuring model misspecification: Application to propensity score methods with complex survey data. *Comput. Stat. Data Anal.*, 128:48–57, December 2018.

[19] D Y Lin and D Zeng. Correcting for population stratification in genomewide association studies. *J. Am. Stat. Assoc.*, 106(495):997–1008, September 2011.

[20] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011.

[21] Christoph Lippert, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle. Limix: genetic analysis of multiple traits. *BioRxiv*, 2014.

[22] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

[23] Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning, 2019.

[24] Yang Ning, Sida Peng, and Kosuke Imai. Robust estimation of causal effects via high-dimensional covariate balancing propensity score, 2018.

[25] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

[26] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[27] AL Price, NJ Patterson, RM Plenge, ME Weinblatt, Shadick NA, and Reich D. Principal components analysis corrects for stratification in genome-wide association studies., 2006.

[28] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics*, 11(7):459–463, 2010.

[29] Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. Predicting sales from the language of product descriptions. In *eCOM@ SIGIR*, 2017.

[30] James M Robins, Andrea Rotnitzky, and Mark van der Laan. On profile likelihood: Comment. *J. Am. Stat. Assoc.*, 95(450):477, June 2000.

[31] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41, April 1983.

[32] Matthew J. Smith, Camille Maringe, Bernard Rachet, Mohammad A. Mansournia, Paul N. Zivich, Stephen R. Cole, and Miguel Angel Luque-Fernandez. Tutorial: Introduction to computational causal inference using reproducible stata, r and python code, 2020.

[33] Til Stürmer, Sebastian Schneeweiss, Kenneth J Rothman, Jerry Avorn, and Robert J Glynn. Performance of propensity score calibration–a simulation study. *Am. J. Epidemiol.*, 165(10):1110–1118, May 2007.

[34] Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data, 2017.

[35] Tyler VanderWeele. The use of propensity score methods in psychiatric research. *Int. J. Methods Psychiatr. Res.*, 15(2):95–103, June 2006.

[36] Victor Veitch, Yixin Wang, and David M. Blei. Using embeddings to correct for unobserved confounding in networks, 2019.

[37] Victor Veitch, Dhanya Sridhar, and David M. Blei. Adapting text embeddings for causal inference, 2020.

[38] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[39] Larry Wasserman. *Nonparametric Curve Estimation*, pages 303–326. Springer New York, New York, NY, 2004. ISBN 978-0-387-21736-9. doi: 10.1007/978-0-387-21736-9_20. URL https://doi.org/10.1007/978-0-387-21736-9_20.

[40] Bruce S Weir and C Clark Cockerham. Estimating f-statistics for the analysis of population structure. *evolution*, pages 1358–1370, 1984.

[41] Florian Wilhelm. Causal inference and propensity score methods. https://florianwilhelm.info/2017/04/causal_inference_propensity_score/. URL https://florianwilhelm.info/2017/04/causal_inference_propensity_score/.

[42] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M

Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38 (2):203–208, 2006.

[43] Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions, 2017.

A. Estimators for Average Treatment Effects

We expressed ATE as $\tau = \mathbb{E}\left(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right)$. Following Smith et al. [32], we can simplify the following term

$$\begin{aligned}
\mathbb{E}\left[\frac{TY}{e(X)}\right] &= \mathbb{E}[\mathbb{E}\left(\frac{TY}{e(X)}|T,X\right)] \\
&= \mathbb{E}[\left(\frac{T\mathbb{E}(Y|T,X)}{e(X)}\right)] \\
&= \mathbb{E}[\left(\frac{T\mathbb{E}(Y|T=1,X)}{e(X)}\right)] \\
&= \mathbb{E}[\mathbb{E}\left(\frac{T\mathbb{E}(Y|T=1,X)}{e(X)}|X\right)] \\
&= \mathbb{E}[\left(\frac{\mathbb{E}(Y|T=1,X)P(T=1|X)}{e(X)}\right)] \\
&= \mathbb{E}[\mathbb{E}(Y|T=1,X)].
\end{aligned}$$

Similarly,

$$\mathbb{E}\left[\frac{(1-T)Y}{1-e(X)}\right] = \mathbb{E}[\mathbb{E}(Y|T=0,X)].$$

Thus, we can show that ATE is indeed equivalent to $\mathbb{E}\left(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right)$.

The Inverse Propensity of Treatment Weight (IPTW) estimator uses an approximate model $Q(T=1|X)$ of $P(T=1|X)$ to produce an estimate $\hat{\tau}$ of the ATE, which is computed as $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t^{(i)}y^{(i)}}{Q(T=1|x^{(i)})} - \frac{(1-t^{(i)})y^{(i)}}{1-Q(T=1|x^{(i)})}\right)$.

Due to sensitivity of the IPTW estimator toward misspecification of propensity score model, Robins et al. [30] propose doubly robust Augmented Inverse Propensity Weighted (AIPW) estimator for ATE. The AIPW estimate is asymptotically unbiased when either the treatment assignment (propensity) model or the outcome model is well-specified.

We define the outcome model as $f(X=x, T=t)$ to approximate the outcome $Y[X=x, T=t]$ as defined in Section 2.

With this, we define the AIPW estimator as

$$\begin{aligned}
\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}[f(X_i, T=1) - f(X_i, T=0) \quad + \\
\frac{T_i(Y_i - f(X_i, T=1))}{e(X_i)} - \frac{(1-T_i)(Y_i - f(X_i, T=0))}{1-e(X_i)}]
\end{aligned}$$

B. Algorithms for calibrated propensity scoring

B.1. A framework for calibrated propensity scoring

Algorithm 2 Calibrated Propensity Scoring

1. Split $\mathcal{D}$ into training set $\mathcal{D}'$ and calibration set $\mathcal{C}$
2. Train a propensity score model $Q(T|X)$ on $\mathcal{D}'$
3. Train recalibrator R over output of Q on $\mathcal{C}$
4. Apply IPW with $R \circ Q$ as prop. score model

Next, we propose algorithms that produce calibrated propensity scoring models. Our approach is outlined in Algorithm 2; it differs from standard propensity scoring methods by the addition of a post-hoc recalibration step (step #3) after training the model Q.

The recalibration step in Algorithm 2 implements a post-hoc recalibration procedure [26; 14] and is outlined in Algorithm 3. The key idea is to learn an auxiliary model $R : [0,1] \to [0,1]$ such that the joint model $R \circ H$ is calibrated. Below, we argue that if R can approximate the density $P(T = 1|Q(T|X) = p)$, $R \circ Q$ will be calibrated [14; 13].

Learning R that approximates $P(T = 1|Q(T|X) = p)$ requires specifying (1) a model class for R and (2) a learning objective ℓ. One possible model class for R are **non-parametric kernel density estimators** over $[0,1]$; their main advantage is that they can provably learn the one-dimensional conditional density $P(T = 1|Q(T|X) = p)$. Examples of such algorithms are RBF kernel density estimation or isotonic regression. Alternatively, one may use a family of **parametric models** for R: e.g., logistic regression, neural networks. Such parametric recalibrators can be implemented easily within deep learning frameworks and work well in practice, as we later demonstrate empirically.

Our learning objective for R can be any proper scoring rule such as the L2 loss, the log-loss, or the Chi-squared loss. Optimizing it is a standard supervised learning problem.

Algorithm 3 Recalibration Step

Input: Pre-trained model $Q : \mathcal{X} \to [0,1]$, recalibrator $R : [0,1] \to [0,1]$, calibration set $\mathcal{C}$
Output: Recalibrated model $R \circ H : \mathcal{X} \to [0,1]$.

1. Create a recalibrator training set: $\mathcal{S} = \{(Q(x), y) \mid x, y \in \mathcal{C}\}$
2. Fit the recalibration model R on $\mathcal{S}$: $\min_R \sum_{(p,y)\in\mathcal{S}} L\left(R(p), y\right)$

B.2. Ensuring calibration in propensity scoring models

Next, we seek to show that Algorithms 2 and 3 provably yield a calibrated model $R \circ Q$. This shows that the desirable property of calibration can be maintained in practice.

Notation We have a calibration dataset $\mathcal{C}$ of size m sampled from P and we train a recalibrator $R : [0,1] \to [0,1]$ over the outputs of a base model Q to minimize a proper loss L. We denote the Bayes-optimal recalibrator by $B := P(T = 1 \mid Q(X))$; the probability of $T = 1$ conditioned on the forecast $(R \circ Q)(X)$ is $S := P(T = 1 \mid (R \circ Q)(X))$. To simplify notation, we omit the variable X, when taking expectations over X, T, e.g. $\mathbb{E}[L(R \circ Q, T)] = \mathbb{E}[L(R(Q(X)), T)]$.

Our first claim is that if we can perform density estimation, then we can ensure calibration. We first formally define the task of density estimation.

Task B.1 (Density Estimation)**.** *The model R approximates the density $B := P(T = t \mid Q(X))$. The expected proper loss of R tends to that of B as $m \to \infty$ such that w.h.p.:*

$$\mathbb{E}[L(B \circ Q, T)] \leq \mathbb{E}[L(R \circ Q, T)] < \mathbb{E}[L(B \circ Q, T)] + \delta$$

where $\delta > 0$, $\delta = o(m^{-k}), k > 0$ is a bound that decreases with m.

Note that non-parametric kernel density estimation is formally guaranteed to solve one-dimensional density estimation given enough data.

Fact B.2 (Wasserman [39])**.** *When R implements kernel density estimation and L is the log-loss, Task B.1 is solved with $\delta = o(1/m^{2/3})$.*

We now show that when we can solve Task B.1, our approach yields models that are asymptotically calibrated in the sense that their calibration error tends to zero as $m \to \infty$.

Theorem B.3. *The model $R \circ Q$ is asymptotically calibrated and the calibration error $\mathbb{E}[L_c(R \circ Q, S)] < \delta$ for $\delta = o(m^{-k}), k > 0$ w.h.p.*

See Appendix H.4.1 for the full proof.

B.3. No-regret calibration

Next, we show that Algorithms 2 and 3 produce a model $R \circ Q$ that is asymptotically just as good as the original Q as measured by the proper loss L.

Theorem B.4. *The recalibrated model has asymptotically vanishing regret relative to the base model: $\mathbb{E}[L(R \circ Q, T)] \leq \mathbb{E}[L(Q, T)] + \delta$, where $\delta > 0, \delta = o(m)$.*

Proof (Sketch). Solving Task B.1 implies $\mathbb{E}[L(R \circ Q, T)] \leq \mathbb{E}[L(B \circ Q, T)] + \delta \leq \mathbb{E}[L(Q, T)] + \delta$; the second inequality

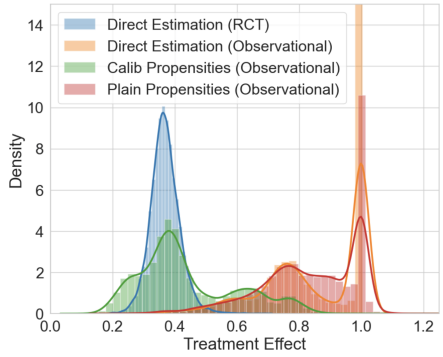

Figure 1. Recalibrating propensity score model reduces the bias in estimating treatment effect from observational data.

holds because a Bayes-optimal B has lower loss than an identity mapping. □

See Appendix H.4.2 for the full proof. Thus, given enough data, we are guaranteed to produce calibrated forecasts and preserve base model performance as measured by L (including L_χ used in our calibration bound).

C. Analysis of calibration

We evaluate the calibration of the propensity score model using expected calibration error (ECE) defined as $\mathbb{E}_{p\sim Q(T=1|X)}[|P(T=1|Q(T=1|X)=p)-p|]$, where $Q(T=1|X)$ models the treatment assignment mechanism $P(T=1|X)$. To compute ECE, we divide the probabilistic output range $[0,1]$ into equal-sized intervals $\{I_0, I_1, .., I_M\}$ such that we can generate buckets $\{B_i\}_{i=1}^M$, where $B_i = \{(X,T,Y)|Q(T=1|X) \in I_i\}$. The estimated ECE is then computed as $\sum_{i=1}^M \frac{|B_i|}{|\bigcup_{j=1}^M B_j|}|\text{avg}_i(B_i) - \text{pred}_i(B_i)|$, where $\text{avg}_i(B_i) = \sum_{j=1}^{|B_i|} T_j/|B_i|$ and $\text{pred}_i(B_i) = \sum_{j=1}^{|B_i|} Q(T=1|X_j)/|B_i|$.

D. Drug Effectiveness Study

We simulate an observational study of recovery time from disease in response to the administration of a drug [41]. The decision to treat an individual with the drug is dependent on the covariates specified as age, gender, and severity of disease. We use logistic regression as the propensity score model. In Figure 1, we see that weighing using recalibrated propensities allows us to approximate the distribution of individual treatment effect estimates better than uncalibrated propensities.

Experimental Setup. We model the outcome using random forests such that the covariates and treatment is taken as input. Logistic regression is used as the propensity score model and the inverse propensity scores are used to weigh each sample while training the outcome model. We use isotonic regression as the recalibrator. The treatment effect is expressed as the ratio $\mathbb{E}(Y(1))/\mathbb{E}(Y(0))$, where $Y(T)$ is the potential outcome Y obtained by setting treatment to T. The outcome is time taken by the patient to make full recovery from the disease. We use 10 cross-val splits to generate the recalibration dataset. Isotonic regression is used as the recalibrator. We use the Inverse-Propensity Treatment Weight (IPTW) and Augmented Inverse Propensity Weight (AIPW) estimators in our experiments. We compare the estimates obtained through calibrated propensities with baselines including estimators based on uncalibrated propensity scores. We measure the performance in terms of the absolute error in estimating ATE as $\epsilon_{ATE} = |\hat{\tau} - \tau|$, where τ is the true treatment effect and $\hat{\tau}$ is our estimated treatment effect.

D.1. Simulation

The covariates contain gender (x_1), age (x_2) and disease severity (x_3), while treatment (t) corresponds to administration of drug. Outcome (y) is the time taken for recovery of patient.

We simulate the covariates as

$$x_1 \sim \text{Bernoulli}(0.5)$$
$$x_2 \sim \text{Gamma}(\alpha = 8, \beta = 4)$$
$$x_3 \sim \text{Beta}(\alpha = 3, \beta = 1.5).$$

The outcome is simulated as

$$y \sim \text{Poisson}(2 + 0.5x_1 + 0.03x_2 + 2x_3 - t).$$

The treatment t is assigned on the basis of the covariates age, gender and severity of disease defined above. The simulations differ in their treatment assignment functions, which are described as follows

1. Simulation A: If $(x_1 = 1)$, set $t = (x_2 > 45)$ else set $t = (x_3 > 0.3)$.
2. Simulation B: If $(x_1 = 1)$, set $t = (x_3 > 0.3)$ else set $t = (x_2 > 40)$.
3. Simulation C: If $x_2 > 50$ AND $x_3 > 0.7$ then set $t = 1$ else $t = 0$.
4. Simulation D: If $x_2 > 50$ XOR $x_3 > 0.7$ then set $t = 1$ else $t = 0$.

For a linear model predicting treatment given covariates, Simulation C is easier to learn as compared to A, B and D.

Table 3. Recalibrating the output of the propensity score model results in a lower error in estimating causal effects. Reduction in ECE implies that the calibration of the model improves with our technique. Results are averaged over 10 experimental repetitions and braces contain the standard error.

Setting	ε_{ATE} with naive estimation	Plain Propensities		Recalibrated Propensities	
		ε_{ATE}	ECE	ε_{ATE}	ECE
Simulation A	0.495 (0.002)	0.477 (0.007)	0.033 (0.001)	0.156 (0.027)	0.027 (0.001)
Simulation B	0.222 (0.003)	0.210 (0.002)	0.040 (0.001)	0.193 (0.002)	0.016 (0.001)
Simulation C	0.273 (0.003)	0.153 (0.003)	0.053 (0.001)	0.147 (0.002)	0.025 (0.002)
Simulation D	0.290 (0.004)	0.066 (0.005)	0.118 (0.001)	0.026 (0.004)	0.026 (0.002)

D.2. Results

In Table 3, we employ different treatment assignment mechanisms in each simulated observational study, allowing us to compare mechanisms that may or may not be well-specified by a linear model. We see that calibrated propensities produce lower absolute error in estimating average treatment effect (ϵ_{ATE}) under varying mechanisms. Here, the naive estimation computes the outcomes without weighing the samples with propensities. In Table 5, we also compare a range of base propensity score models for Simulation A and see the benefits of calibration across these setups. In Figure 2, we see that the calibration curve of propensity score model gets closer to the diagonal after applying recalibration.

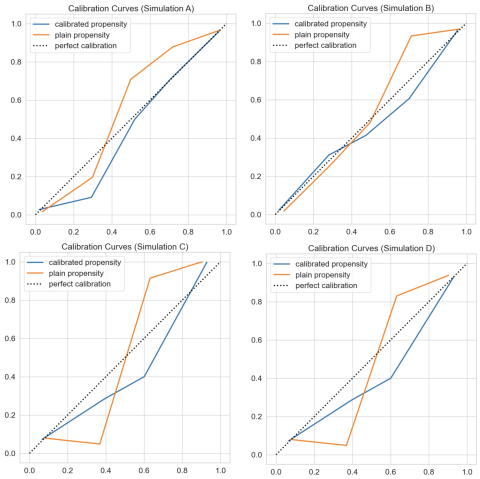

Figure 2. Calibration of propensity score model for Drug Effectiveness Study.

E. Unstructured Covariates Experiment

Setup. We use the Inverse-Propensity Treatment Weight (IPTW) and Augmented Inverse Propensity Weight (AIPW) estimators in our experiments. We compare the estimates obtained through calibrated propensities with several baselines including estimators based on uncalibrated propensity scores. We use sigmoid or isotonic regression as the recalibrator and utilize cross-validation splits to generate the calibration dataset. We measure the performance in terms of the absolute error in estimating ATE as $\epsilon_{ATE} = |\hat{\tau} - \tau|$, where τ is the true treatment effect and $\hat{\tau}$ is our estimated treatment effect.

We simulate a simple observational study following Louizos et al. [22] and Deshpande et al. [5] such that variables $X, T, Y \sim \mathbb{P}$ are binary and the true ATE is zero. Specifically, we generate a synthetic observational dataset consisting of binary variables $X, T, Y \sim \mathbb{P}$, such that

$$\mathbb{P}(Z=1) = \mathbb{P}(Z=0) = 0.5 \qquad \mathbb{P}(X=1|Z=1) = 0.3$$
$$\mathbb{P}(X=1|Z=0) = 0.1 \qquad \mathbb{P}(T=1|Z=1) = 0.4$$
$$\mathbb{P}(T=1|Z=0) = 0.2 \qquad Y = T \oplus Z.$$

Louizos et al. [22] show that the true ATE under this simulation is zero. We would like to note that the presence of hidden confounder Z implies that ignorability is not satisfied in this experiment. Following Deshpande et al. [5], we also introduce an unstructured image covariate $\mathbf{X}$ that represents X as a randomly chosen MNIST image of a zero or one, depending on whether $X = 0$ or $X = 1$. Specifically, $\mathbb{P}(\mathbf{X}|X = 1)$ is uniform over MNIST images of '1' and $\mathbb{P}(\mathbf{X}|X = 0)$ is uniform over MNIST images of '0'.

We use a multi-layer perceptron as the propensity score model and recalibrate its output. In Table 4, we compare the IPTW estimates for ATE using binary X and image $\mathbf{X}$ covariates. The ECE is higher for the plain propensity score model trained on image covariates, indicating higher miscalibration. We see that recalibration also improves ATE estimates with high-dimensional, unstructured covariates.

F. Simulated GWAS Datasets

We have N individuals and M number of total SNPs for each individual. Thus, we need to simulate a SNP matrix $G \in \{0, 1\}^{N \times M}$ and an outcome vector $Y \in \mathbb{R}^N$. We also have a matrix of confounding variables $Z \in \mathbb{R}^{N \times D}$ for these N individuals. We do not observe the confounding

Table 4. Comparison of structured and unstructured covariates.

Setting	ε_{ATE} with naive estimation	Plain Propensities		Recalibrated Propensities	
		ε_{ATE}	ECE	ε_{ATE}	ECE
Image Covariate	0.187 (0.010)	0.161 (0.046)	0.107 (0.029)	0.095 (0.005)	0.024 (0.003)
Binary Covariate	0.176 (0.019)	0.140 (0.029)	0.052 (0.011)	0.099 (0.008)	0.028 (0.004)

variables. Following Wang and Blei [38], we generate the following genotype simulations.

To generate the SNP matrix, we generate an allele frequency matrix $F \in \mathbb{R}^{N\times M}$ such that $F = S\Gamma^\top$, where $S \in \mathbb{R}^{N\times D}$ encodes genetic population structure and $\Gamma \in \mathbb{R}^{M\times D}$ maps how structure affects alleles.

Thus, $g_{ij} \sim \text{Binomial}(1, F_{ij})$.

The outcome is modeled as $Y = \beta^T G + \alpha^T Z + \epsilon$, where β is the vector of treatment effects for each SNP, α is the vector of coefficients corresponding to the hidden confounders in Z and ϵ is noise distributed independently as a Gaussian.

We simulate a high signal-to-noise ratio while simulating outcomes by replacing $\lambda_i = (\alpha^T Z)_i$ as

$$\lambda_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^m \beta_j g_{ij}\}_{i=1}^N}{\sqrt{\nu_{gene}}}\right]\left[\frac{\sqrt{\nu_{conf}}}{s.d.\{\lambda_i\}_{i=1}^N}\right]\lambda_i$$

$$\epsilon_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^m \beta_j g_{ij}\}_{i=1}^N}{\sqrt{\nu_{gene}}}\right]\left[\frac{\sqrt{\nu_{noise}}}{s.d.\{\epsilon_i\}_{i=1}^n}\right]\epsilon_i,$$

where $\nu_{gene} = 0.4, \nu_{conf} = 0.4$, and $\nu_{noise} = 0.2$.

Below, we reproduce the simulation details as described by Wang and Blei [38]. Γ and S are simulated in different ways to generate the following datasets.

1. **Spatial Dataset**: The matrix Γ was generated by sampling $\gamma_{ik} \sim 0.9 \times \text{Uniform}(0, 0.5)$, for $k = 1, 2$ and setting $\gamma_{ik} = 0.05$. The first two rows of S correspond to coordinates for each individual on the unit square and were set to be independent and identically distributed samples from $\text{Beta}(\alpha, \alpha), \alpha = 0.1, 0.3, 0.5$, while the third row of S was set to be 1, i.e. an intercept. As $\alpha \implies 0$, the individuals are placed closer to the corners of the unit square, while when $\alpha = 1$, the individuals are distributed uniformly.

2. **Balding-Nichols Model (BN)**: Each row i of Γ has three independent and identically distributed draws taken from the Balding- Nichols model: $\gamma_{ik} \sim BN(p_i, F_i)$, where $k \in 1, 2, 3$. The pairs (p_i, F_i) are computed by randomly selecting a SNP in the HapMap data set, calculating its observed allele frequency and estimating its FST value using the Weir & Cockerham estimator [40]. The columns of S were Multinomial(60/210,60/210,90/210), which reflect the subpopulation proportions in the HapMap dataset.

3. **1000 Genomes Project (TGP)** [1]: The matrix Γ was generated by sampling $\gamma_{ik} \sim 0.9\text{Uniform} \times (0, 0.5)$, for $k = 1, 2$ and setting $\gamma_{ik} = 0.05$. In order to generate S, we compute the first two principal components of the TGP genotype matrix after mean centering each SNP. We then transformed each principal com- ponent to be between (0,1) and set the first two rows of S to be the transformed principal components. The third row of S was set to 1, i.e. an intercept.

4. **Humane Genome Diversity Project (HGDP)** [6; 2]: Same as TGP but generating S with the HGDP genotype matrix.

These simulations and the ATE estimation experiments were all done on a laptop with 2.8GHz quad-core Intel i7 processor.

G. Additional Experimental Results

For the GWAS experiments, we provide a complete table of dataset simulations and acomparison against different base propensity models in Table 8 and Table 7 respectively.

H. Theoretical Analysis

H.1. Notation

As described in Section 2, we are given an observational dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^n$ consisting of n units, each characterized by features $x^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, a binary treatment $t^{(i)} \in \{0, 1\}$, and a scalar outcome $y^{(i)} \in \mathcal{Y} \subseteq \mathbb{R}$. We assume $\mathcal{D}$ consists of i.i.d. realizations of random variables $X, Y, T \sim P$ from a data distribution P. Although we assume binary treatments and scalar outcomes, our approach naturally extends beyond this setting. The feature space $\mathcal{X}$ can be any continuous or discrete set.

H.2. Calibration: a Necessary Condition for Propensity Scoring Models

Theorem H.1. *When $Q(T|X)$ is not calibrated, there exists an outcome function such that an IPTW estimator based on Q yields an incorrect estimate of the true causal effect almost surely.*

Table 5. Calibration reduces the bias in treatment effect estimation across different base models.

Base classifier	Plain Propensities		Recalibrated Propensities	
	ε_{TE}	ECE	ε_{TE}	ECE
Logistic Regression	0.479 (0.005)	0.029 (0.001)	0.091 (0.022)	0.017 (0.001)
MLP	0.455 (0.042)	0.038 (0.001)	0.027 (0.031)	0.014 (0.001)
SVM	0.485 (0.004)	0.041 (0.001)	0.454 (0.013)	0.018 (0.000)
Naive Bayes	0.471 (0.003)	0.064 (0.000)	0.021 (0.018)	0.003 (0.000)

Table 7. Comparing propensity score models. We compare the AIPW estimate using calibrated propensities and observe reduction in error across a range of base propensity score models.

Dataset	Metrics	LR	MLP	Random Forest	Adaboost	NB
Spatial (α=0.1)	ε_{ATE} (plain)	13.886 (0.755)	17.403 (1.070)	12.911 (0.612)	16.234 (0.916)	582.731 (64.514)
	ε_{ATE} (calib)	8.942 (0.287)	14.661 (0.762)	8.706 (0.322)	8.524 (0.297)	8.526 (0.472)
	Δ_{ECE}	0.022 (0.001)	0.072 (0.003)	0.060 (0.001)	0.252 (0.006)	0.281 (0.002)
Spatial (α=0.3)	ε_{ATE} (plain)	10.460 (0.720)	12.636 (0.730)	10.578 (0.768)	11.764 (0.839)	400.643 (49.301)
	ε_{ATE} (calib)	9.000 (0.58)	11.550 (0.747)	9.277 (0.532)	8.909 (0.549)	9.121 (0.535)
	Δ_{ECE}	0.016 (0.007)	0.070 (0.003)	0.063 (0.001)	0.244 (0.006)	0.281 (0.002)
Spatial (α=0.5)	ε_{ATE} (plain)	8.990 (0.510)	10.408 (0.694)	9.277 (0.518)	9.814 (0.691)	276.017 (24.183)
	ε_{ATE} (calib)	8.590 (0.390)	9.728 (0.650)	8.687 (0.224)	8.520 (0.286)	8.592 (0.216)
	Δ_{ECE}	0.015 (0.001)	0.070 (0.002)	0.065 (0.001)	0.239 (0.007)	0.269 (0.003)
Balding Nichols	ε_{ATE} (plain)	17.660 (1.330)	18.282 (1.267)	18.419 (1.210)	19.248 (1.169)	95.892 (6.350)
	ε_{ATE} (calib)	16.810 (1.390)	17.033 (1.391)	16.611 (1.385)	16.938 (1.367)	16.833 (1.392)
	Δ_{ECE}	0.013 (0.002)	0.041 (0.002)	0.052 (0.002)	0.259 (0.010)	0.261 (0.009)
HGDP	ε_{ATE} (plain)	11.380 (0.110)	12.358 (0.197)	11.529 (0.107)	11.816 (0.108)	138.086 (5.086)
	ε_{ATE} (calib)	11.060 (0.120)	11.198 (0.106)	11.299 (0.143)	11.070 (0.123)	11.430 (0.133)
	Δ_{ECE}	0.011 (0.001)	0.069 (0.002)	0.053 (0.001)	0.275 (0.006)	0.206 (0.003)
TGP	ε_{ATE} (plain)	11.560 (0.650)	11.965 (0.754)	11.677 (0.614)	12.246 (0.713)	87.329 (5.716)
	ε_{ATE} (calib)	10.320 (0.430)	11.530 (0.633)	10.519 (0.402)	10.244 (0.398)	9.070 (0.316)
	Δ_{ECE}	0.022 (0.001)	0.061 (0.002)	0.070 (0.002)	0.204 (0.007)	0.267 (0.004)

Example. Consider a toy binary setting where $\mathcal{X} = \mathcal{T} = \{0,1\}, \mathcal{Y} = \{0,1\}^2$.

We set $Y = (X \oplus T, \bar{X} \oplus \bar{T})$, $P(T = 1|X = 0) = p_0, P(T = 1|X = 1) = p_1$ and $P(X = 1) = 0.5$ such that $\oplus$ is logical 'AND' and $\bar{V}$ denotes logical negation of binary variable V. We see that true ATE is $\tau = (0.5, -0.5)$. Let us assume that $Q(T = 1|X = 0) = q_0$ and $Q(T = 1|X = 1) = q_1$. Thus, with IPTW estimator based on Q, we estimate $\tau' = \mathbb{E}\left(\frac{TY}{Q(T=1|X)} - \frac{(1-T)Y}{1-Q(T=1|X)}\right) = (-\frac{0.5(1-p_0)}{1-q_0}, \frac{0.5.p_1}{q_1})$. The treatment effect $\tau' = \tau$ only when $q_0 = p_0$ and $q_1 = p_1$, which is not true if Q is not calibrated. □

Proof. Let $\mathcal{P}$ be a space of valid probability distributions on $\mathcal{Y}$. We would like to prove that $\exists P'(Y|X = x, T = t) \in \mathcal{P}$ such that

$$\lim_{n\to\infty} \text{Probability}(\hat{\tau}_n = \tau) = 0$$

where

- τ is the true ATE
- $\hat{\tau}_n$ is the ATE estimated using IPTW estimator such that we have n individuals and propensity score model is $Q(T = 1|X)$
- The probability is taken over all propensity models $Q(T = 1|X)$ such that $\exists q \in [0,1], P(T = 1|Q(T = 1|X) = q) \neq q$, and all data-generating distributions $P'(Y,T,X) = P'(Y|X,T).P(T,X)$.

Let $S_Q = \{q|\exists X \in \mathcal{X}, Q(T = 1|X) = q\}$. We partition $\mathcal{X}$ into buckets $\{B_q\}_{q\in S_q}$ such that $B_q = \{X|Q(T = 1|X) = q\}$.

Let $\hat{\tau}(Q) = \lim_{n\to\infty} \tau_n$. Thus, for discrete $\mathcal{X}$, we could

Table 8. Increasing proportion of causal SNPs. Calibrated propensities reduce the bias in treatment effect estimation across all setups and compare favorably against standard GWAS methods.

Method	1% Causal SNPs	2% Causal SNPs	5% Causal SNPs	10% Causal SNPs
Naive	22.408 (5.752)	15.150 (2.213)	23.388 (5.021)	14.846 (2.272)
PCA	18.104 (5.378)	13.699 (2.413)	15.837 (3.331)	11.683 (0.983)
FA	18.532 (3.641)	14.166 (2.259)	16.855 (2.764)	11.963 (0.958)
LMM	17.575 (3.408)	13.896 (2.152)	14.681 (3.366)	10.108 (0.827)
IPTW (Calib)	**17.237 (3.054)**	**13.113 (1.775)**	**14.587 (3.432)**	**8.625 (0.838)**
IPTW (Plain)	19.297 (3.425)	14.372 (1.482)	18.290 (3.788)	11.859 (0.95240)
AIPW (Calib)	17.647 (3.208)	13.382 (1.676)	15.166 (3.597)	9.078 (0.928)
AIPW (Plain)	20.652 (3.286)	13.720 (1.798)	21.321 (4.750)	12.904 (1.990)

write

$$\hat{\tau}(Q)$$
$$= \mathbb{E}_{Y\sim P'(.|T,X);T,X\sim P}\left[\left(\frac{TY}{Q(T=1|X)} - \frac{(1-T)Y}{1-Q(T=1|X)}\right)\right]$$

Computing expectation over X

$$= \sum_{X\in\mathcal{X}} \mathbb{E}_{Y\sim P'(.|T,X);T\sim P(.|X)}\left[\frac{TYP(X)}{Q(T=1|X)}\right] -$$
$$\sum_{X\in\mathcal{X}} \mathbb{E}_{Y\sim P'(.|T,X);T\sim P(.|X)}\left[\frac{(1-T)YP(X)}{1-Q(T=1|X)}\right]$$

Computing expectation over T

$$= \sum_{X\in\mathcal{X}} \mathbb{E}_{Y\sim P'(.|X,T=1)}\left[\left(\frac{P(T=1|X)Y}{Q(T=1|X)}\right)P(X)\right]$$
$$+ \sum_{X\in\mathcal{X}} \mathbb{E}_{Y\sim P'(.|X,T=0)}\left[\left(-\frac{(1-P(T=1|X))Y}{1-Q(T=1|X)}\right)P(X)\right]$$
$$= \sum_{X\in\mathcal{X}} (\mathbb{E}_{Y\sim P'(.|X,T=1)}\left[\left(\frac{P(T=1|X)Y}{Q(T=1|X)}\right)\right]$$
$$- \mathbb{E}_{Y\sim P'(.|X,T=0)}\left[\left(\frac{(1-P(T=1|X))Y}{1-Q(T=1|X)}\right)\right])P(X)$$

Expressing the summation over X differently

$$= \sum_{q\in S_Q}\sum_{X\in B_q} (\mathbb{E}_{Y\sim P'(.|X,T=1)}\left[\left(\frac{P(T=1|X)Y}{Q(T=1|X)}\right)\right]$$
$$- \mathbb{E}_{Y\sim P'(.|X,T=0)}\left[\left(\frac{(1-P(T=1|X))Y}{1-Q(T=1|X)}\right)\right])P(X)$$

Since $Q(T=1|X)$ is not calibrated, we know that $\exists q \in [0,1], P(T=1|Q(T=1|X)=q) \neq q$. Let us pick $q' \in S_Q$ such that $P(T=1|Q(T=1|X)=q') \neq q'$.

We could design $P'(Y|X,T) = \mathbb{I}(Y = T.\mathbb{I}(X \in B_{q'}))$.

Now, we can write

$$\hat{\tau}(Q) = \sum_{q\in S_Q}\sum_{X\in B_q} (\mathbb{E}_{Y\sim P'(.|X,T=1)}\left[\left(\frac{P(T=1|X)Y}{Q(T=1|X)}\right)\right]$$
$$- \mathbb{E}_{Y\sim P'(.|X,T=0)}\left[\left(\frac{(1-P(T=1|X))Y}{1-Q(T=1|X)}\right)\right])P(X)$$
(Since $Y=0$ when $T=0$ or $X \notin B_{q'}$)
$$= \sum_{X\in B_{q'}}\left(\left(\frac{P(T=1|X)P(X)}{Q(T=1|X)}\right)\right)$$
$$= \sum_{X\in B_{q'}}\left(\left(\frac{P(T=1|X)P(X)}{q'}\right)\right)$$
$$= \frac{P(T=1|X\in B_{q'})P(X\in B_{q'})}{q'}$$

Also, for the above data-generation process,

$$\tau = \hat{\tau}(P)$$
$$= \sum_{X\in\mathcal{X}} (\mathbb{E}_{Y\sim P'(Y|X,do(T=1))}[Y]$$
$$- \mathbb{E}_{Y\sim P'(Y|X,do(T=0))}[Y]).P(X)$$
$$= \sum_{q\in S_Q}\sum_{X\in B_q} (\mathbb{E}_{Y\sim P'(Y|X,do(T=1))}[Y]$$
$$- \mathbb{E}_{Y\sim P'(Y|X,do(T=0))}[Y]).P(X)$$
$$= \sum_{X\in B_{q'}} P(X)$$
$$= P(X\in B_{q'}))$$

Thus,

$$\lim_{n\to\infty} \text{Prob}(\tau_n = \tau)$$
$$= \text{Prob}(\hat{\tau}(Q) = \tau)$$
$$= \text{Prob}\left(\frac{P(T=1|X\in B_{q'})P(X\in B_{q'})}{q'} = P(X\in B_{q'})\right)$$
$$= \text{Prob}\left(P(T=1|X\in B_{q'}) = q'\right)$$
$$= \text{Prob}\left(P(T=1|Q(T=1|X)=q') = q'\right)$$
$$= 0,$$

since we began with the assumption that $P(T=1|Q(T=1|X)=q') \neq q'$.

Please note that we could have defined a set of outcome functions that produce $Y=0$ for $X \in B_{q'}$, thus, potentially letting us compute unbiased treatment effects despite working with a miscalibrated model. However, we want our IPTW estimator to provide unbiased ATE estimates over all possible outcome functions. Here, we can see that IPTW estimator for ATE that uses a miscalibrated propensity score model cannot obtain unbiased treatment effect estimates on all possible outcome functions.

□

H.3. Calibrated Uncertainties Improve Propensity Scoring Models

We define the true ATE as

$$\begin{aligned}\tau &= \mathbb{E}_{y\sim P(Y=y|do(T=1))}[y] - \mathbb{E}_{y\sim P(Y=y|do(T=0))}[y] \\ &= \sum_y y(\sum_X P(Y=y|X,do(T=1))P(X) - \\ &\sum_X P(Y=y|X,do(T=0))P(X)) \\ &= \sum_y y(\sum_X P(Y=y|X,T=1)P(X) - \\ &\sum_X P(Y=y|X,T=0)P(X))\end{aligned}$$

Next, recall that the finite-sample Inverse Propensity of Treatment Weight (IPTW) estimator with a model $Q(T=1|X)$ of $P(T=1|X)$ produces an estimate $\hat{\tau}_n(Q)$ of the ATE, which is computed as

$$\hat{\tau}_n(Q) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t^{(i)}y^{(i)}}{Q(T=1|x^{(i)})} - \frac{(1-t^{(i)})y^{(i)}}{1-Q(T=1|x^{(i)})}\right).$$

We define $\tau(Q)$ as the limit $lim_{n\to\infty}\hat{\tau}_n(Q)$ when the amount of data goes to infinity. Notice that we can write

$$\lim_{n\to\infty}(\hat{\tau}_n(Q)) = \hat{\tau}(Q) = \sum_y y[\pi_{y,1}(Q) - \pi_{y,0}(Q)],$$

where

$$\begin{aligned}\pi_{y,t}(Q) &= P(T=t)\sum_X P(Y=y|X,T=t)\frac{P(X|T=t)}{Q(T=t|X)} \\ &= \sum_X P(Y=y|X,T=t)\frac{P(T=t|X)}{Q(T=t|X)}P(X)\end{aligned}$$

We have a multiplicative term $P(T=t)$ in the above expression since we are dividing by n in the finite-sample formula as opposed to n_t (the number of samples with treatment t). In other words, in order for the finite-sample formula to be a valid Monte Carlo estimator with samples coming from $P(X|T=t)$, there needs to be an "effective adjustment factor" of n_t/n (such that $(n_t/n)\cdot(1/n_t) = (1/n)$), and this term is $P(T=t)$ in the limit of infinite data.

Clearly, if $Q=P$ we have $\hat{\tau}(Q) = \hat{\tau}(P) = \tau$. If not, we can consider the error

$$E = |(\hat{\tau}(P) - \hat{\tau}(Q))|.$$

H.3.1. Bounding the Error of Causal Effect Estimation Using Proper Losses

We can form a bound on E as

$$\begin{aligned}E &= |[\hat{\tau}(P) - \hat{\tau}(Q)]| \\ &= \left|\sum_y y[(\pi_{y,1}(P) - \pi_{y,0}(P)) - (\pi_{y,1}(Q) - \pi_{y,0}(Q))]\right| \\ &\leq \sum_t \left|\sum_y y[(\pi_{y,t}(P) - \pi_{y,t}(Q)]\right| \\ &\leq \sum_t \sum_y [|y||\pi_{y,t}(P) - \pi_{y,t}(Q)|] \\ &= \sum_t \sum_y |y|[\left|\sum_X P(Y=y|X,T=t)P(X)\left(1 - \frac{P(T=t|X)}{Q(T=t|X)}\right)\right|] \\ &\leq \sum_t \sum_y |y|[\sum_X P(Y=y|X,T=t)P(X)\left|1 - \frac{P(T=t|X)}{Q(T=t|X)}\right|] \\ &= \sum_t \sum_y |y|.[\sum_X P(Y=y|X,T=t)P(X)\ell_X(P,Q)^{1/2}] \\ &\text{where } \ell_X(P,Q) = \left(1 - \frac{P(T=t|X)}{Q(T=t|X)}\right)^2 \\ &= \sum_t \sum_y |y|.\mathbb{E}_{X\sim R_{y,t}}[\ell_X(P,Q)^{1/2}]\end{aligned}$$

where $R_{t,y} \propto P(Y=y|X,T=t)P(X)$ (i.e. $R_{t,y} \sim k.P(Y=y|X,T=t)P(X)$, k is constant) and $\ell_X(P,Q)$ is a type of expected Chi-Squared divergence between P,Q. It is a type of proper score. Thus when $P=Q$, we get zero error, and otherwise we get a bound.

In the above derivation, we see that the expected error $|\pi_{y,t}(P) - \pi_{y,t}(Q)|$ induced by an IPTW estimator with propensity score model Q is bounded as

$$|\pi_{y,t}(P) - \pi_{y,t}(Q)| \leq \mathbb{E}_{X\sim R_{y,t}}[\ell_\chi(P,Q)^{\frac{1}{2}}].$$

H.3.2. Calibration Reduces Variance of Inverse Probability Estimators

Theorem H.2. *Let P be the data distribution, and suppose that $1-\delta > P(T|X) > \delta$ for all T, X and let Q be a calibrated model relative to P. Then $1-\delta > Q(T|X) > \delta$ for all T, X as well.*

Proof. Suppose $Q(T=1|x) = q$ for some x and $q < \delta$. Since Q is calibrated, we have $P(T=1|Q(T=1|X) = q) = q < \delta$.

However $P(T = 1|x) > \delta$ for every x. Hence, $P(T = 1|X \in A) > \delta$, for all sets $A \subseteq \mathcal{X}$. This implies that $P(T = 1|Q(T = 1|X) = q) > \delta$ for all $q \in [0, 1]$.

Thus, we have a contradiction. □

H.3.3. Calibration Improves the Accuracy of Causal Effect Estimation

Theorem H.3. *The error of an IPTW estimator with propensity model Q tends to zero as $n \to \infty$ if:*

1. *Separability holds, i.e., $\forall X_1, X_2 \in \mathcal{X}, P(T|X_1) \neq P(T|X_2) \implies Q(T|X_1) \neq Q(T|X_2)$*
2. *The model Q is calibrated, i.e., $\forall q \in (0, 1), P(T = 1|Q(T = 1|X) = q) = q$*

Proof. We prove this for discrete inputs at first and then prove it for continuous inputs.

Discrete Input Space.

If our input space $\mathcal{X}$ is discrete, then the number of distinct values that $Q(T = 1|X)$ can take is countable. Let us assume that $Q(T = 1|X)$ takes values $\{q_i\}_{i=1}^{M}$. Thus, we can partition $\mathcal{X}$ into buckets $\{B_i\}_{i=1}^{M}$ such that $B_i = \{X|Q(T = 1|X) = q_i\}$. Due to separability, we have $\forall X_1, X_2 \in \mathcal{X}, Q(T|X_1) = Q(T|X_2) \implies P(T|X_1) = P(T|X_2)$. Thus, we have $\forall i, \forall X_1, X_2 \in B_i, Q(T = 1|X_1) = Q(T = 1|X_2)$, and $P(T = 1|X_1) = P(T = 1|X_2)$.

Let us assume that for each bucket B_i, our true propensity $P(T = 1|X)$ is p_i, i.e, if $X \in B_i$ then $Q(T = 1|X) = q_i$ and $P(T = 1|X) = p_i$.

Assuming positivity, $0 < p_i < 1$.

Now, for all i, we can write

$$P(T = 1|Q(T = 1|X) = q_i) = P(T = 1|X \in B_i) = p_i.$$

If Q is calibrated, then by definition $p_i = q_i$.

Now, we can write the expression for ATE τ as

$$\tau = \hat{\tau}(P) = \mathbb{E}_{Y,T,X}[\frac{TY}{P(T = 1|X)} - \frac{(1 - T)Y}{(1 - P(T = 1|X))}] = \sum_{i=1}^{N} P(X \in B_i)\mathbb{E}_{Y,T}\left(\frac{TY}{p_i} - \frac{(1 - T)Y}{(1 - p_i)}\right)$$

Using our propensity score model $Q(T = 1|X)$, we estimate $\hat{\tau}$ as

$$\hat{\tau}(Q) = \mathbb{E}_{Y,T,X}[\frac{TY}{Q(T = 1|X)} - \frac{(1 - T)Y}{(1 - Q(T = 1|X))}] = \sum_{i=1}^{N} P(X \in B_i)\mathbb{E}_{Y,T}\left(\frac{TY}{q_i} - \frac{(1 - T)Y}{(1 - q_i)}\right)$$

If our model Q is calibrated, then $p_i = q_i$. Hence, $0 < q_i < 1$ and $\hat{\tau}$ is well-defined. Also, $\tau = \hat{\tau}(P) = \hat{\tau}(Q)$.

When our observational data contains n units, the IPTW estimator based on model $Q(T = 1|X)$ is $\hat{\tau}_n = \frac{1}{n}\sum_{i=0}^{n}\left(\frac{T^{(i)}Y^{(i)}}{Q(T=1|X^{(i)})} - \frac{(1-T^{(i)})Y^{(i)}}{1-Q(T=1|X^{(i)})}\right)$.

Hence, $\lim_{n\to\infty} \hat{\tau}_n = \hat{\tau}(Q) = \hat{\tau}(P) = \tau$.

Continuous Input Space.

When X is continuous, the number of buckets can be uncountable. The buckets can now be formed as $B_q = \{X|Q(T = 1|X) = q\}, \forall q \in [0, 1]$. It is easy to see that $\{B_q\}_{q\in[0,1]}$ partitions $\mathcal{X}$. Note that B_q can be empty if there exists no X such that $Q(T = 1|X) = q$.

Due to separability, $\forall X_1, X_2 \in \mathcal{X}, Q(T|X_1) = Q(T|X_2) \implies P(T|X_1) = P(T|X_2)$. Thus, for all q, $P(T = 1|X)$ takes on a unique value for all $X \in B_q$, i.e., $\forall q \in [0, 1], P(T = 1|X \in B_q) = f(q)$, where function $f : [0, 1] \to [0, 1]$.

Hence, we can write

$$\forall q \in [0, 1], P(T = 1|Q(T = 1|X) = q) = P(T = 1|X \in B_q) = f(q).$$

When model $Q(T = 1|X)$ is calibrated by our definition, then $\forall q \in [0, 1], q = f(q)$.

Therefore, $\forall q \in [0, 1], Q(T = 1|X \in B_q) = q = f(q) = P(T = 1|X \in B_q)$.

Since $\{B_q\}_{q\in[0,1]}$ partitions $\mathcal{X}$, we have $\forall X \in \mathcal{X}, P(T = 1|X) = Q(T = 1|X)$. Thus, $\hat{\tau}(P) = \hat{\tau}(Q)$.

□

H.4. Algorithms for Calibrated Propensity Scoring

H.4.1. Asymptotic Calibration Guarantee

Theorem H.4. *The model $R \circ Q$ is asymptotically calibrated and the calibration error $\mathbb{E}[L_c(R \circ Q, S)] < \delta(m)$ for $\delta(m) = o(m^{-k}), k > 0$ w.h.p.*

Proof. Any proper loss can be decomposed as: proper loss = calibration - sharpness + irreducible term [9]. The calibration term consists of the error $\mathbb{E}[L_c(R \circ Q, S)]$. The

sharpness and irreducible term can be represented as the refinement term $\mathbb{E}(L_r(S))$. Table 9 provides examples of some proper loss functions and the respective decompositions. The rest of our proof uses the techniques of Kuleshov and Deshpande [13] in the context of propensity scores.

Kull and Flach [15] show that the refinement term can be further divided as $\mathbb{E}(L_r(S)) = \mathbb{E}(L_g(S, B\circ Q))+\mathbb{E}(L(B\circ Q, T))$. Here, B is the Bayes optimal recalibrator $P(T = 1|Q(T = 1|X))$ and S is $P(T = 1|R \circ Q)$.

Recall that if we solve the Task B.1, we have for $\delta(m) = o(1)$

$$\begin{aligned}\mathbb{E}(L(B \circ Q, T)) &\leq \mathbb{E}(L(R \circ Q, T)) \\ &\leq \mathbb{E}(L(B \circ Q, T)) + \delta(m)\end{aligned}$$

Using Gneiting et al. [7], Kull and Flach [15] we decompose $\mathbb{E}(L(R \circ Q, T))$

$$\begin{aligned}&\implies \mathbb{E}(L(B \circ Q, T)) \leq \\ &\quad (\mathbb{E}(L_c(R \circ Q, S))+\mathbb{E}(L_g(S, B \circ Q))+ \\ &\qquad \mathbb{E}(L(B \circ Q, T))) \leq \mathbb{E}(L(B \circ Q, T)) + \delta(m) \\ &\implies \mathbb{E}(L_c(R \circ Q, S)) + \mathbb{E}(L_g(S, B \circ Q)) \leq \delta(m) \\ &\implies \mathbb{E}(L_c(R \circ Q, S)) \leq \delta(m)\end{aligned}$$

Thus, solving Task B.1 allows us to obtain asymptotically calibrated $R \circ Q$ such that the calibration error is bounded as $\mathbb{E}[L_c(R \circ Q, S)] < \delta(m)$.

□

H.4.2. No-Regret Calibration

Theorem H.5. *The recalibrated model has asymptotically vanishing regret relative to the base model:* $\mathbb{E}[L(R \circ Q, T)] \leq \mathbb{E}[L(Q, T)] + \delta$, *where* $\delta > 0, \delta = o(m^{-k}), k > 0$.

Proof. Solving Task B.1 implies $\mathbb{E}[L(R\circ Q, T)] \leq \mathbb{E}[L(B\circ Q, T)] + \delta \leq \mathbb{E}[L(Q, T)] + \delta$. The first inequality comes from definition of Task B.1 and the second inequality holds because a Bayes-optimal B has lower loss than an identity mapping [13].

□

Proper Score	**Loss** $L(F,G)$	**Calibration** $L_c(F,S)$	**Refinement** $L_r(S)$
Logarithmic	$\mathbb{E}_{y\sim G}\log f(y)$	$\mathrm{KL}(s\|\|f)$	$H(s)$
CRPS	$\mathbb{E}_{y\sim G}\,(F(y)-G(y))^2$	$\int_{-\infty}^{\infty}(F(y)-S(y))^2\mathrm{dy}$	$\int_{-\infty}^{\infty}S(y)(1-S(y))dy$
Quantile	$\mathbb{E}_{y\sim G}^{\tau\in U[0,1]}\rho_\tau(y-F^{-1}(\tau))$	$\int_0^1\int_{S^{-1}(\tau)}^{F^{-1}(\tau)}(S(y)-\tau)dyd\tau$	$\mathbb{E}_{y\sim S}^{\tau\in U[0,1]}\rho_\tau(y-S^{-1}(\tau))$

Table 9. Proper loss functions. A proper loss is a function $L(F,G)$ over a forecast F targeting a variable $y\in\mathcal{Y}$ whose true distribution is G and for which $S(F,G)\geq S(G,G)$ for all F. Each $L(F,G)$ decomposes into the sum of a calibration loss term $L_c(F,S)$ (also known as reliability) and a refinement loss term $L_r(S)$ (which itself decomposes into sharpness and an uncertainty term). Here, $S(y)$ denotes the cumulative distribution function of the conditional distribution $\mathbb{P}(Y=y\mid F_X=F)$ of Y given a forecast F, and $s(y), f(y)$ are the probability density functions of S and F, respectively. We give three examples of proper losses: the log-loss, the continuous ranked probability score (CRPS), and the quantile loss.