DEBIASED PRIOR-DATA FITTED NETWORKS FOR AMORTIZED BAYESIAN SEMI-PARAMETRIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior-data fitted networks (PFNs) have recently gained popularity as an end-to-end approach to amortized Bayesian inference. Yet, their statistical behavior for semi-parametric targets, such as means under missingness or causal effects, remains poorly understood. In this work, we develop theoretical foundations for debiased PFNs to perform inference on finite-dimensional targets in general semi-parametric settings. First, we show that, despite being Bayes-optimal with respect to the target, an idealized PFN can exhibit a posterior concentration rate that implicitly depends on the concentration rates of (latent) nuisance functions and can become slower as the prior becomes more diverse. Then, to speed up the target posterior concentration rates, we introduce two generic debiasing strategies based on efficient influence functions: (i) a label correction and (ii) a prior correction. Finally, we show that the debiased PFNs (i) & (ii) possess concentration rates that asymptotically match a convergence rate given by traditional frequentist bias-corrected estimators. In this way, we establish a *novel connection between amortized Bayesian inference with PFNs and frequentist semi-parametric efficiency theory*. Under additional assumptions, this allows us to guarantee local and global optimality of PFN-based estimators (namely, the Bernstein–von Mises property and minimax optimality, respectively). Therefore, our work provides guidelines for designing debiased PFNs for semi-parametric inference with principled and provable asymptotic guarantees.

1 INTRODUCTION

Prior-data fitted networks (PFNs) have emerged as a powerful approach for building tabular foundation models (Müller et al., 2022; Hollmann et al., 2023; 2025). PFNs combine pretraining with synthetic data and in-context learning (Brown et al., 2020; Dong et al., 2024) to perform an approximate *amortized Bayesian inference* (ABI) (Gershman & Goodman, 2014; Le et al., 2017). Hence, a PFN (usually a transformer-based neural network) is trained once on datasets sampled from a prior over data-generating mechanisms, and then, at inference time, it maps a new dataset directly into a target posterior distribution.

PFNs have been applied widely across various machine learning (ML) tasks (Müller et al., 2025). For example, PFNs were adopted to tasks like time-series forecasting (Dooley et al., 2023; Ansari et al., 2024; Taga et al., 2025), survival analysis (Seletkov et al., 2026; Qi et al., 2026; Böhm et al., 2026), Bayesian optimization (Müller et al., 2023), and explainability (Fonseca & Stoyanovich, 2026; Brockschmidt et al., 2026), among others. Since recently, PFNs were also used for *semi-parametric inference tasks*[1] such as causal inference (Robertson et al., 2025; Balazadeh et al., 2025; Mahajan et al., 2025; Ma et al., 2026; Javurek et al., 2026; Reuter et al., 2026; Melnychuk et al., 2026). These developments point toward a broader future role of PFNs as general-purpose engines for automating any statistical inference task.

Surprisingly, the statistical properties of PFN-based inference remain only partially explored. Some theoretical aspects of the PFNs have been studied (Nagler, 2023; Ma et al., 2025; Nagler & Rügamer, 2025; Hülskamp et al., 2026; Ching et al., 2026). However, several core properties of PFN-based inference (and, especially, semi-parametric inference) remain *unclear*:

[1] In this work, we use the term "semi-parametric inference" to describe the estimation of a finite-dimensional target in the presence of an infinite-dimensional nuisance component (= nuisance function) (Bickel et al., 1993).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

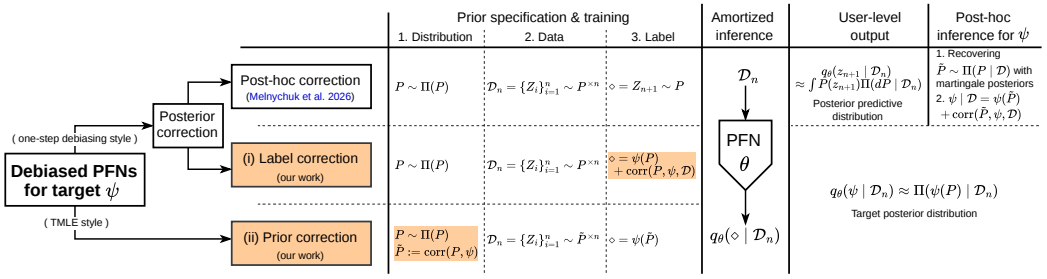

Figure 1: **Overview of our proposed debiasing strategies:** (i) label correction and (ii) prior correction (marked with orange). Compared to the existing approach of Melnychuk et al. (2026), our debiasing strategies (i) and (ii) are applied at the prior specification & training stages of PFNs and, therefore, do *not* require post-hoc inference to be performed by a downstream user.

(1) At what rates do naïve PFN-based posteriors concentrate around finite-dimensional targets?
(2) Can debiasing strategies improve these rates for a specific inference target?
(3) Do the proposed debiasing strategies yield posterior concentration rates that match the convergence rates of well-adopted debiased frequentist estimators, and are they asymptotically optimal?

Our work studies the above research questions, and thereby adds to the theoretical foundations for PFN-based inference through the lens of *asymptotic properties*. Our contributions[2] are three-fold:

(1) We start by restating that a PFN aims to exactly recover the Bayes-optimal posterior wrt. its prior (Nagler, 2023). Importantly, however, such Bayes optimality *does not by itself guarantee favorable concentration properties* for a particular finite-dimensional target. We show that the corresponding target posterior concentration rate is governed by two closely related features of the underlying semi-parametric inference task. First, even though a PFN can directly output a posterior for the finite-dimensional target, the concentration rate implicitly depends on the posterior concentration rates of (latent) nuisance functions. Second, informally, these rates can become slower as the prior becomes more diverse.

(2) Then, to speed up the target posterior concentration rate, we carefully tailor two generic debiasing strategies (Ray & van der Vaart, 2020; Yiu et al., 2025) specifically for PFNs (see Fig. 1): (i) a *one-step posterior label correction* that modifies the PFN's target label using an efficient influence function; and (ii) a *prior correction along a least-favorable direction* that alters the PFN's prior in a target-specific direction. Importantly, unlike the post-hoc approach of Melnychuk et al. (2026), our strategies (i)–(ii) are applied at the pretraining stage and thus do *not* introduce additional correction steps at test time.

(3) Finally, we show that the concentration rates of the resulting debiased PFNs asymptotically match the convergence rates achieved by well-established frequentist bias-corrected estimators. Hence, the corresponding target posteriors also concentrate at the same rates. This allows us to establish a **direct connection between PFN-based ABI and frequentist semi-parametric efficiency theory**. Furthermore, under additional assumptions, our debiasing strategies ensure local and global optimality guarantees in the form of semi-parametric Bernstein–von Mises property (Freedman, 1999) and minimax optimality (Balakrishnan et al., 2026), respectively.

To instantiate our theory, we later use three running examples of inferential targets: (A) *mean value* (classical statistics setting), (B) *mean value under missingness at random (MAR)* (prediction-powered inference setting (Angelopoulos et al., 2023; van der Laan & van der Laan, 2026)), and (C) *weighed average treatment effect* (WATE) (causal inference setting (Hirano et al., 2003)).

In sum, our main contribution is providing *guidelines for debiasing PFNs* for semi-parametric inference tasks with principled and provable asymptotic properties.

2 Related Work

In the following, we discuss existing works on (a) the statistical properties of PFNs and (b) semi-parametric inference with PFNs. In Appendix A, we provide further details on related works.

[2]Code is available at `https://anonymous.4open.science/r/Ortho-PFN-Anonymized-E4B7`.

Statistical properties of PFNs. Nagler (2023) has laid general statistical foundations of PFNs and argued whether transformers are suitable for consistent ABI. Multiple subsequent works then studied downstream aspects of the pretrained PFNs: e. g., the application of PFNs for downstream inference (Nagler & Rügamer, 2025; Ng et al., 2025), and the error decomposition of PFN predictions (Hülskamp et al., 2026; Fortini et al., 2026). Other works explored fundamental limits of PFNs pretraining itself: Kim et al. (2024); Ching et al. (2026) showed that transformer-based PFNs can in principle achieve minimax rates for Hölder smooth functions; Ma et al. (2025) demonstrated the adaptivity of PFN wrt. different smoothness levels; and Reuter et al. (2025) empirically showed the learnability of latent variables with PFNs. Our work falls into the latter category: We establish the asymptotic properties of *idealized PFNs used as semi-parametric inference engines*.

Semi-parametric inference with PFNs. Recent work has adapted PFNs to causal inference (Balazadeh et al., 2025; Robertson et al., 2025; Ma et al., 2026; Javurek et al., 2026); yet, these works primarily fit PFNs in a naïve way and, at best, only establish estimation consistency (without further discussing the asymptotic properties). The closest to our theoretical work is Melnychuk et al. (2026), who show that naïve PFN-based inference for the ATE can suffer from prior-induced confounding bias (Hahn et al., 2018; Linero & Antonelli, 2023) and who apply a *post-hoc* one-step posterior correction. Our work differs from the former: We characterize posterior concentration rates for general finite-dimensional semi-parametric targets, and we incorporate debiasing *directly* into PFN pretraining, thereby avoiding *post-hoc* correction at test time. **To the best of our knowledge, our work is the first to offer theoretical foundations for PFNs used in semi-parametric inference**.

3 Preliminaries

Notation. Let $Z \in \mathcal{Z} \subseteq \mathbb{R}^d$ denote a random variable with probability distribution $P = P(Z) \in \mathcal{P}(\mathcal{Z})$, where $\mathcal{P}(\mathcal{Z})$ is a class of probability distributions defined on $\mathcal{Z}$. With a slight abuse of notation, we also denote the density/probability mass function of observing $Z = z$ as $P(z)$. For $P \in \mathcal{P}(\mathcal{Z})$, we observe an i.i.d. dataset $\mathcal{D}_n = (Z_1, \ldots, Z_n) \sim P^{\otimes n}$. Similarly, we denote processes (distributions in the infinite-dimensional/functional spaces) with Π: e. g., $\Pi(P)$ is a prior process and $\Pi(P \mid \mathcal{D}_n)$ is a posterior process over the distributions of Z. Also, for any P-measurable function f, we write an $L_p(P)$-norm of f as $\|f\|_{L_p(P)} := (\mathbb{E}[|f(Z)|^p])^{1/p}$; a mean of f as $Pf := \mathbb{E}[f(Z)] = \int f(z)\,\mathrm{d}P(z)$; a sample mean of f as $P_n f := n^{-1}\sum_{i=1}^n f(Z_i)$; and the mean wrt. Bayesian bootstrap process as $\mathrm{BB}_n\{f(Z)\} = \sum_{i=1}^n W_i f(z_i)$ where $(W_1, \ldots, W_n)$ is a sample from a Dirichlet process $\mathrm{Dir}(n; 1, \ldots, 1)$. Similarly, we also define L_p-norm for a function f as $\|f\|_{L_p} := (\int |f(z)|^p\,\mathrm{d}z)^{1/p}$. For a sequence of random vectors X_n and a positive deterministic sequence a_n, we write $X_n = o_P(a_n)$ if $\|X_n\|_2/a_n \to 0$ in $P^{\otimes n}$-probability, and $X_n = O_P(a_n)$ if $\|X_n\|_2/a_n$ is bounded in $P^{\otimes n}$-probability. Finally, $a \lesssim b$ means that $a \leq C \cdot b$ for some $C \geq 0$.

Problem setup. In our work, we are interested in a statistical inference for a final-dimensional statistical functional (estimand) $\psi : \mathcal{P}(\mathcal{Z}) \to \Psi \subseteq \mathbb{R}$ (w. l. o. g. $\psi(P)$ is one-dimensional). Let $\eta(P) \in \mathcal{H}$ denote a nuisance functions space equipped with norm $\|\cdot\|_{\mathcal{H}}$ (namely, η are the attributes of P that have a variational influence on the target ψ). For a fixed distribution $P_0 \in \mathcal{P}(\mathcal{Z})$, we write $\psi_0 := \psi(P_0)$ as the target statistical estimand. When $\psi(P)$ is pathwise differentiable at P, we denote its *efficient influence function* (EIF) under the nonparametric model by $\phi_\psi(Z; P)$ (see Appendix B for details). Importantly, pathwise-differentiable estimands $\psi(P)$ admit a *von–Mises expansion*:

$$\psi(P) - \psi(P_0) = (P - P_0)\phi_\psi(\cdot; P) + R_{2,\psi}(P, P_0) = -P_0\phi_\psi(\cdot; P) + R_{2,\psi}(P, P_0), \tag{1}$$

where $R_{2,\psi}(P, P_0)$ is the second-order remainder term.

Inferential targets. In our work, we consider the following three running examples of estimands:

Example A (Mean value)**.** *We start with a trivial estimand,*[3] *a mean value of $Z \in [0, 1]$; $\psi(P) = \mathbb{E}[Z] = \int zP(z)\,\mathrm{d}z$. Then, the corresponding nuisance functions, EIF, and second-order remainder are*

$$\eta(P) = \{P(z)\};\ \phi_\psi(Z; P) = Z - \psi(P);\ R_{2,\psi}(P, P_0) = 0. \tag{2}$$

Example B (Mean value under missingness at random (MAR))**.** *Mean value under MAR is a classical semi-parametric target (Angelopoulos et al., 2023; van der Laan & van der Laan, 2026). Here,*

[3] While estimation of the mean does not involve nuisance functions, we consider it as the simplest running example that illustrates the main ideas before turning to genuinely semi-parametric settings.

$Z = (X, R, RY)$ with covariates $X \in [0,1]^{d_x}$, labeling indicator $R \in \{0,1\}$, and the (partially missing) outcome $RY \in [0,1]$. Here, the inference target is $\psi = \mathbb{E}[Y]$. Under standard MAR assumptions (i. e., conditional independence, $R \perp\!\!\!\perp Y \mid X$; and strong overlap, $\kappa < \pi(x) < 1 - \kappa$, with $\kappa \in (0, 1/2)$), $\psi(P) = \mathbb{E}[\mu(X)] = \int \mu(x)P(x)\,\mathrm{d}x$, where $\mu(x) = \mathbb{E}[RY \mid X = x, R = 1]$ is a conditional outcome mean, and $\pi(x) = \mathbb{P}(R = 1 \mid X = x)$ is a labeling score. Then, the corresponding nuisance functions, EIF, and second-order remainder are (Robins et al., 1995):

$$\eta(P) = \{\mu(x), P(x)\};\ \phi_\psi(Z; P) = \frac{R}{\pi(X)}(Y - \mu(X)) + \mu(X) - \psi(P);\ R_{2,\psi}(P, P_0) = P_0\left[\frac{\pi - \pi_0}{\pi}(\mu - \mu_0)\right]. \quad (3)$$

Example C (Weighed average treatment effect (WATE))**.** *Let $Z = (X, A, Y)$, with propensity score $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$ and outcome regressions $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$. For a user-specified weight $w(\cdot; \pi) \geq 0$, the WATE is $\psi_w(P) = \mathbb{E}[w(X; \pi)(\mu_1(X) - \mu_0(X))]/\mathbb{E}[w(X; \pi)]$, which includes the ATE ($w \equiv 1$) and overlap-weighted ATE ($w = \pi(1 - \pi)$) as special cases. We defer further details to Appendix C.*

Amortized Bayesian inference with PFNs. A PFN amortizes inference across synthetically generated datasets. Specifically, for a fixed dataset $\mathcal{D}_n$, let $q_\theta(z_{n+1} \mid \mathcal{D}_n)$ denote a conditional density on $\mathcal{Z}$ parameterized by a neural network with parameters $\theta \in \Theta$ (typically, a transformer (Vaswani et al., 2017)). Then, a **posterior-predictive PFN** (Müller et al., 2022) *with fixed n* aims to minimize a prior-predictive negative log-likelihood risk wrt. to a new observation $Z_{n+1} \sim P$:

$$\mathcal{R}^{\mathrm{PP}}_{\Pi,n}(\theta) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\mathbb{E}_{Z_{n+1}\sim P}\left[-\log q_\theta\big(Z_{n+1} \mid \mathcal{D}_n\big)\right], \quad (4)$$

where $\Pi = \Pi(P)$ is a prior process over the distributions P.

Bayesian self-consistency. The main reason why the prior-predictive risk in Eq. (4) performs ABI is the property of *Bayesian self-consistency* (Dawid, 1982): the conditional distribution of a held-out observation Z_{n+1} generated under the prior process Π, given the observed dataset $\mathcal{D}_n$, coincides with the Bayesian posterior-predictive distribution. More formally, a PFN $q_{\theta^*}(\cdot \mid \mathcal{D}_n)$ that minimizes the population risk in Eq. (4) also minimizes a Kullback-Leibler (KL) divergence wrt. a ground-truth posterior-predictive density $\pi(z_{n+1} \mid \mathcal{D}_n) := \int P(z_{n+1})\Pi(\mathrm{d}P \mid \mathcal{D}_n)$ (Nagler, 2023):

$$\theta^* = \arg\min_{\theta\in\Theta} \mathcal{R}^{\mathrm{PP}}_{\Pi,n}(\theta) = \arg\min_{\theta\in\Theta} \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\,\mathrm{KL}[\pi(\cdot \mid \mathcal{D}_n)\,||\,q_\theta(\cdot \mid \mathcal{D}_n)]. \quad (5)$$

Therefore, the statistical properties of PFN-based ABI are mainly guided by (a) the specification of the prior process Π (which, in turn, specifies the prior $\Pi(P)$, the implicit log-likelihood $\ell_n(P) := \log\prod_{i=1}^n P(Z_i)$, and, thus, the posterior $\Pi(\mathrm{d}P \mid \mathcal{D}_n) \propto \exp\{\ell_n(P)\}\Pi(\mathrm{d}P)$) and (b) the amortization errors due to limited expressivity of the PFN q_θ. In the following, we primarily focus on the aspect (a).

4 Semi-Parametric Inference with PFNs

In the following, we develop theoretical foundations for amortized Bayesian semi-parametric inference with PFNs along our three research questions: (1) In Sec. 4.1, we first show that, despite being Bayes-optimal, an idealized PFN can suffer from slow posterior concentration rates. (2) In Sec. 4.2, we propose two remedies against the slow concentration rates of PFNs: (i) label correction and (ii) prior correction. (3) In Sec. 4.3, we formalize the favorable asymptotic properties of the debiased PFNs. We provide detailed theoretical statements and all the proofs in Appendix D.

4.1 Bayes-Optimality and Slow Concentration Rates

The core idea of ABI with PFNs can be straightforwardly applied to semi-parametric inference. Specifically, a **PFN for target** ψ with fixed n is given by a conditional density $q_\theta(\psi \mid \mathcal{D}_n)$ in the parameter space Ψ that minimizes the following risk wrt. a label ψ:

$$\mathcal{R}^{\psi}_{\Pi,n}(\theta) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\left[-\log q_\theta\big(\psi(P) \mid \mathcal{D}_n\big)\right]. \quad (6)$$

Again, similarly, to the posterior-predictive PFNs (see Sec. 3), the minimization of the risk in Eq. (6) is equivalent to the KL-projection of $q_\theta(\psi \mid \mathcal{D}_n)$ on the ground-truth posterior distribution[4] $\Pi(\psi(P) \mid \mathcal{D}_n)$. This yields the following useful definition.

[4]More precisely, $\Pi(\psi(P) \mid \mathcal{D}_n)$ is the marginalization of the full posterior $\Pi(P \mid \mathcal{D}_n)$.

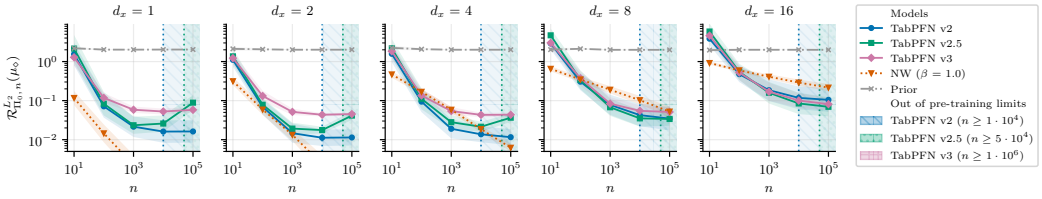

Figure 2: Slowing concentration rates for different versions of TabPFN ($\diamond \in \{\text{v2}, \text{v2.5}, \text{v3}\}$). Here, we show how Π_0-prior-averaged squared L_2-concentration errors of the conditional outcome mean $\mu(x)$ posteriors, $\mathcal{R}^{L_2}_{\Pi_0,n}(\mu_\diamond) := \mathbb{E}_{P_0\sim\Pi_0}\mathbb{E}_{\mathcal{D}_n\sim P_0^{\otimes n}}[M(P_0, \mathcal{D}_n, \Pi_\diamond)^2]$, decrease depending on dataset sizes n and covariate dimensionality d_x (prior-averaging is based on a sample of $D = 24$ datasets). For reference, we also show a Π_0-prior-averaged squared L_2-error of Nadaraya–Watson regression with a Matérn kernel (= NW ($\beta = 1.0$)). Note that the ground-truth Π_0 is based on the adapted prior of the TabPFN v1 (Hollmann et al., 2023) ($\mu(x) \sim \Pi_{\text{v1}}(\mu(x))$ and $P \sim N(\mu(x), \sigma^2 = 0.3^2)$) and is assumed to be in support of the priors of the higher versions.

Definition 1 (Idealized PFN)**.** *A PFN q_{θ^*} for target ψ with fixed n is called idealized if $\theta^* = \arg\min_{\theta\in\Theta} \mathcal{R}^{\psi}_{\Pi,n}(\theta)$ yields $Q_{\theta^*}(\psi \mid \mathcal{D}_n) = \Pi(\psi(P) \mid \mathcal{D}_n)$ for every $\mathcal{D}_n$ from a prior process Π (here: Q_{θ^*} is a distribution corresponding to the density q_{θ^*}). Hence, the PFN class is sufficiently expressive to capture the ground-truth posterior.*

Bayes-optimality. Equipped with the definition of idealized PFNs, we can use both the ground-truth posterior and its PFN approximation interchangeably. Also, it allows us to formalize a Bayes decision-theoretic property, i. e., Bayes-optimality, of PFNs for target ψ (see Appendix B for details).

Remark 1 (Bayes-optimality of the idealized PFN)**.** *The posterior mean of an idealized PFN, i.e., $\hat{\psi}_{\theta^*}(\mathcal{D}_n) := \int \psi\, q_{\theta^*}(\psi \mid \mathcal{D}_n)\, \mathrm{d}\psi$, is Bayes-optimal under the squared l_2 risk $\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\hat{\psi}(\mathcal{D}_n) - \psi(P)]^2$:*

$$\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}_{\theta^*}) = \inf_{\hat{\psi}:\mathcal{P}^{\otimes n}\to\Psi} \mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}) = \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\mathrm{Var}_\Pi[\psi(P) \mid \mathcal{D}_n]], \tag{7}$$

where the last term expresses the expected posterior uncertainty about the target.

Remark 1 justifies the optimality of q_{θ^*} from a Bayesian perspective. However, from a frequentist lens, we might be interested in how quickly $\Pi(\psi(P) \mid \mathcal{D})$ and, hence, $q_{\theta^*}(\psi \mid \mathcal{D}))$ *concentrate locally around* some **fixed** $\psi_0 = \psi(P_0)$ sampled from a prior. It turns out that the concentration rates of the target ψ implicitly depend on the concentration rate of the associated nuisance functions $\eta(P)$, even though those nuisance functions are not explicitly modeled by the idealized PFN q_{θ^*}.

Theorem 1 (Slow local L_2-concentration rates (short version))**.** *Assume that the target ψ admits a von–Mises expansion from Eq.* (1) *wrt. to any pair of P, P_0 independently drawn from $\Pi(P)$. Then, for a fixed target $\psi_0 = \psi(P_0)$ where $P_0 \in \operatorname{supp}\Pi$, and $\mathcal{D}_n \sim P_0^{\otimes n}$, the target L_2-concentration error, $\|\psi - \psi_0\|_{L_2(q_{\theta^*},\mathcal{D}_n)} := (\int(\psi - \psi_0)^2 q_{\theta^*}(\psi \mid \mathcal{D}_n)\, \mathrm{d}\psi)^{1/2}$, can be upper-bounded:*

$$\|\psi - \psi_0\|_{L_2(q_{\theta^*},\mathcal{D}_n)} \lesssim \underbrace{\left[\int \|\eta(P) - \eta(P_0)\|^2_{\mathcal{H}}\ \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}}_{F_\psi(P_0,\mathcal{D}_n,\Pi)} + \underbrace{\left[\int R^2_{2,\psi}(P, P_0)\, \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}}_{R_{2,\psi}(P_0,\mathcal{D}_n,\Pi)}, \tag{8}$$

where $F_\psi(P_0, \mathcal{D}_n, \Pi)$ and $R_{2,\psi}(P_0, \mathcal{D}_n, \Pi)$ are posterior-averaged first-order error and second-order remainder, respectively. Furthermore, assuming the L_2-concentration of the latent nuisance posteriors (i. e., $\int \|\eta(P) - \eta(P_0)\|^2_{\mathcal{H}}\, \Pi(\mathrm{d}P \mid \mathcal{D}_n) \to 0$ with $n \to \infty$), the upper-bound in Eq. (8) *is rate-sharp: Unless explicitly corrected, there exist a posterior $\Pi(P \mid \mathcal{D}_n)$ under which the target L_2-concentration error inherits the rate of the latent nuisance functions, namely, the rate of the first-order term $F_\psi(P_0, \mathcal{D}_n, \Pi)$.*

Theorem 1 shows that, despite the Bayesian optimality of the idealized PFNs, the target L_2-concentration error can linearly depend on **the concentration rate of the latent nuisance functions**. Hence, it can also be interpreted as a generalization of the previously discovered issue of *prior-induced bias* (Hahn et al., 2018). In our examples, Eq. (8) can be instantiated as follows.

Example A (Mean value)**.** *For a fixed pair of distributions P and P_0, we have the following:*

$$|\psi - \psi_0| = \left|\int_0^1 z\,(P(z) - P_0(z))\, \mathrm{d}z\right| \leq \|z - 1/2\|_{L_2} \cdot \|P - P_0\|_{L_2(z)} = 1/\sqrt{12} \cdot \|P - P_0\|_{L_2(z)}, \tag{9}$$

and, thus, $\|\eta(P) - \eta(P_0)\|_{\mathcal{H}} = \|P - P_0\|_{L_2(z)}$ *is the leading first-order error term. Furthermore, this bound is locally rate-sharp and is attained when the posterior* $\Pi(P \mid \mathcal{D}_n)$ *assigns asymptotically non-negligible mass to a region around the path* $P_t(z) = P_0(z) + t(z - 1/2)$ *for* $t > 0$ *(along such a path both nuisance errors and* $|\psi(P_t) - \psi_0|$ *are of order* t*).*

Example B (Mean value under MAR)**.** *Similarly, for a fixed pair of distributions* P *and* P_0*,*

$$|\psi - \psi_0| = \left|(P - P_0)\mu + P_0\Big[\frac{\pi_0}{\pi}(\mu - \mu_0)\Big] + R_{2,\psi}(P, P_0)\right| \tag{10}$$

$$\leq \|\mu\|_{L_2(x)} \cdot \|P - P_0\|_{L_2(x)} + \frac{1}{\kappa} \cdot \|\mu - \mu_0\|_{L_2(P_0)} + |R_{2,\psi}(P, P_0)|\,, \tag{11}$$

and, thus, $\|\eta(P) - \eta(P_0)\|_{\mathcal{H}} = \|P - P_0\|_{L_2(x)} + \|\mu - \mu_0\|_{L_2(P_0)}$ *is the leading first-order error term. Again, this bound is locally rate-sharp and is attained when the posterior* $\Pi(P \mid \mathcal{D}_n)$ *assigns asymptotically non-negligible mass to a region around a path* $P_t(x) = P_0(x) + th_0(x)$*,* $\mu_t(x) = \mu_0(x) + t$*, and* $\pi_t(x) = \pi_0(x)$*, where* h_0 *is any valid mean-zero direction satisfying* $\int \mu_0(x)h_0(x)\,\mathrm{d}x \neq 0$ *(along such a path both nuisance errors and* $|\psi(P_t) - \psi_0|$ *are of order* t*).*

The latter two examples are particularly relevant for idealized (yet uncorrected) PFNs: Here, the averaged errors of the conditional expectation ($P_0(\mu_t - \mu_0) = t$) implicitly propagate into the error of the target **in the same order** t. For a general posterior however, the squared error $\{P_0(\mu - \mu_0)\}^2$ is only upper-bounded by $\|\mu - \mu_0\|^2_{L_2(P_0)}$ (with equality achieved for specific, worst-case posteriors $\mu = \mu_t$). Thus, in general, the L_2-concentration rate of μ informs us of the *worst-case* L_2-concentration rate of target ψ.

Slow concentration rates. It turns out that, in fully non-parametric Bayesian models (and idealized PFNs can be viewed as such), the L_2-concentration rate of μ (i. e., the rate of $M(P_0, \mathcal{D}_n, \Pi)^2 := \int \|\mu - \mu_0\|^2_{L_2(P_0(x))}\,\Pi(\mathrm{d}P \mid \mathcal{D}_n)$) can be arbitrarily slow. For example, for many non-parametric Bayesian models with a more diverse prior $\Pi(P)$, the posterior concentration rate slows down[5] (e. g., for Gaussian processes with a rougher Matérn kernel (van der Vaart & van Zanten, 2011)).

Our findings have also implications for existing PFNs such as TabPFN (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025; 2026). In Fig. 2, we empirically show that more recent versions of TabPFN ($\diamond = \mathrm{v2} \to \mathrm{v2.5} \to \mathrm{v3}$) have slower rates of the L_2-concentration error, $M(P_0, \mathcal{D}_n, \Pi_\diamond)$. These insights motivate our debiasing strategies for PFNs, with the aim that *a more diverse prior does not result in much slower target concentration rate*.

4.2 Debiasing Strategies

In the following, we introduce two debiasing strategies for PFNs.

(i) Label correction. We start by tailoring (i) one-step posterior corrections (Yiu et al., 2025) specifically for PFNs. Unlike Melnychuk et al. (2026), we aim to perform bias correction not post-hoc but at the *training* stage of the PFN for target ψ. For this, we use a key property of Bayesian self-consistency (see Sec. 3): A joint sample from a prior process $(\mathcal{D}_n, \psi(P))$ can be interpreted as a joint sample of random $\mathcal{D}_n$ and a conditional (posterior) target $\psi(P) \mid \mathcal{D}_n$. Therefore, in the case of the PFNs, the one-step posterior correction takes the form of label ψ correction.

A **label correction** modifies the risk of a PFN with target ψ by performing a one-step posterior correction wrt. $\psi(P)$:

$$\tilde{\mathcal{R}}^{\psi}_{\Pi,n}(\theta) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\Big[-\log q_\theta\big(\tilde{\psi}(P) \mid \mathcal{D}_n\big)\Big], \quad \tilde{\psi}(P) := \psi(P) + \underbrace{\mathrm{BB}_n\{\phi_\psi(Z; P)\}}_{\mathrm{corr}(P,\psi,\mathcal{D}_n)}, \tag{12}$$

where the weights of $\mathrm{BB}_n\{\cdot\}$ are independent of P and $\mathcal{D}_n$.

Construction. Hence, once the EIF for a target ψ is derived, we automatically get the label correction (given that we can effectively infer the corresponding nuisance functions from the full prior $P \sim \Pi(P)$). Let us now revisit our running examples.

[5]Yet, this is not true in general: Classical non-parametric Bayesian theory (Ghosal & van der Vaart, 2017) characterizes posterior contraction through a trade-off between the amount of prior mass assigned to small neighborhoods of the data-generating distribution and the complexity of the effective support of the prior. Still, in our work, we consider the worst-case scenario, namely when the greater prior diversity leads to worse posterior concentration.

Example A (Mean value)**.** *For the mean value, the label correction effectively replaces the initial label, $\psi(P)$, with the Bayesian bootstrap, thereby removing the first-order error wrt. P:*

$$\tilde{\psi}(P) = \psi(P) + \mathrm{BB}_n\{Z - \psi(P)\} = \mathrm{BB}_n\{Z\}. \tag{13}$$

Example B (Mean value under MAR)**.** *Similarly, the label correction is:*

$$\tilde{\psi}(P) = \psi(P) + \mathrm{BB}_n\left\{\frac{R}{\pi(X)}(Y - \mu(X)) + \mu(X) - \psi(P)\right\} = \mathrm{BB}_n\left\{\frac{R}{\pi(X)}(Y - \mu(X)) + \mu(X)\right\}. \tag{14}$$

Interpretation. In both Eq. (13)-(14), the label correction effectively replaces the original target label with a Bayesian-bootstrap one-step estimator. This is done by design such that the first-order error term (i. e., $F_\psi(P_0, \mathcal{D}_n, \Pi)$ in Eq. (8)) gets replaced with *the empirical term that concentrates around zero with the parametric rate* $O_{P_0}(n^{-1/2})$.

(ii) Prior correction. An alternative approach to speed up the posterior concentration rates is an approach of prior corrections in the least-favorable direction (Ray & van der Vaart, 2020; Ray & Szabó, 2019; Yang et al., 2015). Existing constructions are typically expressed through target-specific parameterizations of nuisance functions (e. g., Gaussian process priors for $\mu(x)$ for the mean under MAR). Here, on the other hand, we formulate a general, distribution-level approach, suitable for PFNs. Specifically, we perturb each prior draw $P \sim \Pi(P)$ as follows.

A **prior correction** modifies the prior of a PFN with target ψ by augmenting prior draws $P \sim \Pi(P)$ with an independent perturbation along a least-favorable direction[6] $\mathrm{LF}_\psi(P; \lambda)$:

$$\tilde{P} := \underbrace{\mathrm{LF}_\psi(P; \lambda)}_{\mathrm{corr}(P,\psi)}, \quad \begin{cases} \lambda \sim N(0, \sigma_n^2) \text{ and } \lambda \perp\!\!\!\perp P, \\ \mathrm{LF}_\psi \text{ is such that } \mathrm{LF}_\psi(P; 0) = P \text{ and } \mathrm{S}[\mathrm{LF}_\psi(P; \lambda)]\big|_{\lambda=0} = \phi_\psi(\cdot; P), \end{cases} \tag{15}$$

where σ_n^2 is chosen in a way that $n\sigma_n^2 \to \infty$; $\mathrm{LF}_\psi(\cdot; \lambda) : \mathcal{P} \to \mathcal{P}$ is a distributional transformation; and $\mathrm{S}[\cdot] : \mathcal{P} \to L_2(\mathcal{P})$ is a score operator (i. e., for a distribution P_λ parametrized by λ with density $P(z; \lambda)$, $S[P_\lambda](z) = \frac{\partial}{\partial \lambda} \log P(z; \lambda)$). Furthermore, the draws of P and λ jointly induce a **corrected prior process** $\widetilde{\Pi}(\tilde{P})$, that is assumed to have the same support, $\operatorname{supp} \widetilde{\Pi} = \operatorname{supp} \Pi$.

Construction. We immediately see that the perturbation LF_ψ is given as a possibly non-unique solution to the operator equation $\mathrm{LF}_\psi(P; \cdot) = (S[\cdot]|_{\lambda=0})^{-1} \phi_\psi(\cdot; P)$. This non-uniqueness gives us a certain freedom in how to perform the correction; for example, we can either correct the full distribution or only the relevant nuisance function components. In the following running examples, we carefully adapt the existing approaches from the TMLE literature (van der Laan & Rubin, 2006), specifically for the debiased ABI with PFNs.

Example A (Mean value)**.** *Here, one of the alternatives for the prior correction is the exponential tilting (Díaz & Rosenblum, 2015). Then, the corrected prior density of P from $P \sim \Pi(P)$ is*

$$\tilde{P}(z) := \exp\{\lambda(z - \psi(P))\} / C_\psi(P) \cdot P(z), \quad C_\psi(P) = P[\exp\{\lambda(z - \psi(P))\}], \tag{16}$$

where $C_\psi(P)$ is a normalization constant, and it is easy to verify that $\tilde{P} = P$ when $\lambda = 0$ and $S[\tilde{P}](Z)|_{\lambda=0} = Z - \psi(P) = \phi_\psi(Z; P)$. Hence, the corrected prior gives each prior draw an additional degree of freedom, λ, exactly in the direction that moves its mean (we will explain soon why this helps to reduce the first-order error posterior).

Example B (Mean value under MAR)**.** *Assume, the density of a prior draw, $P(z)$, factorizes as $P(z) = P(x)\pi(x)^r(1 - \pi(x))^{1-r}P(y \mid x, R = 1)^r$, and the outcome distribution is heteroscedastic normal $P(y \mid x, R = 1) = N(\mu(x); \sigma^2(x))$. Now, the prior correction jointly (namely, with the same λ) perturbs $P(x)$ and $\mu(x)$ along their corresponding least-favorable directions:*

$$\begin{aligned} &\tilde{P}(x) := \exp\{\lambda(\mu(x) - \psi(P))\} / C_\psi(P) \cdot P(x), \quad C_\psi(P) = P[\exp\{\lambda(\mu(x) - \psi(P))\}]; \\ &\tilde{\mu}(x) := \mu(x) + \lambda \cdot \sigma^2(x)/\pi(x), \end{aligned} \tag{17}$$

where $C_\psi(P)$ is a normalization constant, and, again, it can be easily verified that the correction satisfies the definition in Eq. (15)*, in particular, because $S[\tilde{P}](X)|_{\lambda=0} = \mu(X) - \psi(P)$ and $S[\tilde{P}](Y \mid X, R = 1)|_{\lambda=0} = (Y - \mu(X))/\pi(X)$. Here, the prior correction acts differently wrt. different prior nuisance functions (i. e., it does the exponential tilting wrt. $P(x)$ and the location shift (Gruber & van der Laan, 2010) wrt. $\mu(x)$).*

[6] This direction is called *least favorable* because, among local perturbations of the data-generating distribution P, it represents the directions in which estimation of ψ is locally hardest.

Interpretation. Informally speaking, in all the examples, the prior correction alters the original prior such that it puts enough probability mass along directions that are *locally most relevant for changing the target* ψ. This allows to nudge the joint posterior $\widetilde{\Pi}(P, \lambda \mid \mathcal{D}_n)$ (and, thus, $\widetilde{\Pi}(\psi(\tilde{P}) \mid \mathcal{D}_n)$) toward what is supported by the data and, thereby, to correct for the first-order error term. The prior correction also has an important connection to the sharp counterexamples from Theorem 1: There, local perturbations of the nuisance functions induce a change in ψ of the same order, causing the target posterior to inherit the first-order nuisance error. The prior correction enriches the prior along such target-relevant directions, so that posterior updating can correct this first-order error directly.

4.3 Asymptotic Properties of Debiased PFNs

Our debiasing strategies allow us to significantly improve the previous theoretical results in Theorem 1.

Theorem 2 (L_2-concentration rates of debiased PFNs (short version))**.** *Assume the conditions of Theorem 1 and that either debiasing strategy (i)-(ii) from Sec. 4.2 is applied. Then, under mild regularity conditions, the L_2-concentration error of the debiased PFN, $\tilde{q}_{\theta^*}$, satisfies*

$$\left\|\tilde{\psi} - \psi_0\right\|_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)} \lesssim \underbrace{\tilde{F}_\psi(P_0, \mathcal{D}_n, \star)}_{=O_{P_0}(n^{-1/2})} + R_{2,\psi}(P_0, \mathcal{D}_n, \star), \quad \star = \begin{cases} \Pi, \textit{ (i) label correction,} \\ \widetilde{\Pi}, \textit{ (ii) prior correction,} \end{cases} \tag{18}$$

where $R_{2,\psi}(P_0, \mathcal{D}_n, \star)$ is the corresponding posterior-averaged second-order remainder.

Theorem 2 shows that our debiasing strategies (i) and (ii) effectively replace a potentially slowly concentrating error term, $F_\psi(P_0, \mathcal{D}_n, \Pi)$, with the term that concentrates at the fast parametric speed. This finally allows us to establish a strong connection between the debiased ABI with PFNs and the classical frequentist bias-corrected estimators (e.g., A-IPTW / TMLE). As the concentration rate *matches up to the second-order remainder* with the convergence rate of the classical frequentist bias-corrected estimators, we consider two cases: (1) when the second-order remainder is negligible, and (2) when it is not.

(1) Second-order remainder is $o_{P_0}(n^{-1/2})$. In this case, we establish the Bernstein–von Mises property, namely, the asymptotic normality of the posterior given by the debiased PFN $\tilde{q}_{\theta^*}$.

Theorem 3 (Bernstein–von Mises property for debiased PFNs)**.** *Assume the conditions of Theorem 2, $R_{2,\psi}(P_0, \mathcal{D}_n, \Pi) = o_{P_0}(n^{-1/2})$, and standard regularity conditions for semi-parametric Bernstein–von Mises convergence. Then, for either debiasing strategy (i) or (ii), the debiased PFN posterior distribution $\widetilde{Q}_{\theta^*}$ is asymptotically Gaussian and efficient:*

$$d_{\mathrm{TV}}\left[\widetilde{Q}_{\theta^*}\left(\sqrt{n}(\psi - \psi_0) \mid \mathcal{D}_n\right); N\left(\Delta_n, V_{\mathit{eff}}(P_0) := \mathrm{Var}_{P_0}\left[\phi_\psi(\cdot; P_0)\right]\right)\right] \xrightarrow{P_0} 0, \tag{19}$$

where d_{TV} is total variation, and $\Delta_n := \sqrt{n} P_n[\phi_\psi(\cdot; P_0)]$ is a frequentist bias correction term.

Theorem 3 also establishes a **local optimality of the debiased PFNs** around P_0: Their posterior uncertainty asymptotically attains the semi-parametric efficiency bound, $\mathrm{Var}_{P_0}[\phi_\psi(\cdot; P_0)]$, and hence no other regular estimator can achieve a smaller asymptotic variance along local perturbations of P_0.

(2) Second-order remainder is non-negligible. When the second-order remainder concentrates slower than the parametric rate $n^{-1/2}$, its contribution to the posterior is non-negligible, and, in general, the asymptotic shape of the posterior cannot be established. Still, in this case, we can infer the *posterior concentration rates* for each term in Eq. (18) in Theorem 2. This, in turn, allows us to formulate two following global decision-theoretic guarantees for the debiased PFN, i.e., one for Π-prior-averaged and another for worst-case squared l_2 risks.

Theorem 4 (Rate optimality of debiased PFNs)**.** *Assume the conditions of Theorem 2, and suppose that, on an effective prior-supported class $\mathcal{P}_n^\Pi := \mathrm{supp}(\Pi) \cap \mathcal{P}_n^{SA}$ with $\Pi(\mathcal{P}_n^\Pi) \to 1$ (where $\mathcal{P}_n^{SA}$ is a structure-agnostic class of distributions that depends on the nuisance function estimators), the second-order contribution satisfies uniformly: $\sup_{P \in \mathcal{P}_n^\Pi} \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}[R_{2,\psi}^2(P, \mathcal{D}_n, \star)] \lesssim r_{2,n}^2$. Then, for either debiasing strategy $\star$ (i.e., label or prior correction), the following hold:*

$$\underbrace{\mathcal{R}_{\Pi,n}^{l_2}(\hat{\psi}_{\theta^*})}_{\textit{Bayes optimal risk, see Remark 1}} \leq \mathcal{R}_{\Pi,n}^{l_2}(\hat{\tilde{\psi}}_{\theta^*}) \leq \mathcal{R}_{\Pi,n}^{L_2}(\tilde{\psi}) \lesssim n^{-1} + r_{2,n}^2 \quad \textit{(Π-prior-averaged risks)}, \tag{20}$$

$$\underbrace{\mathfrak{M}_n(\mathcal{P}_n^\Pi)}_{\textit{minimax lower bound}} \leq \mathcal{W}_{\mathcal{P}_n^\Pi,n}^{l_2}(\hat{\tilde{\psi}}_{\theta^*}) \leq \mathcal{W}_{\mathcal{P}_n^\Pi,n}^{L_2}(\tilde{\psi}) \lesssim n^{-1} + r_{2,n}^2 \quad \textit{(worst-case risks)}, \tag{21}$$

where inequalities in red hold if Π *is fully supported on* $\mathcal{P}_n^{\Pi}$*;* $\mathcal{R}_{\Pi,n}^{l_2}(\hat{\psi}) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\hat{\psi}(\mathcal{D}_n) - \psi(P)]^2$*,* $\mathcal{R}_{\Pi,n}^{L_2}(\tilde{\psi}) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}||\tilde{\psi} - \psi||^2_{L_2(\tilde{q}_{\theta^*},\mathcal{D}_n)}$ *are* Π*-prior-averaged squared* l_2 *risks;* $\mathcal{W}^{l_2}_{\mathcal{P}_n^{\Pi},n}(\hat{\psi}) := \sup_{P\in\mathcal{P}_n^{\Pi}}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\hat{\psi}(\mathcal{D}_n) - \psi(P)]^2$ *and* $\mathcal{W}^{L_2}_{\mathcal{P}_n^{\Pi},n}(\tilde{\psi}) := \sup_{P\in\mathcal{P}_n^{\Pi}}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}||\tilde{\psi} - \psi||^2_{L_2(\tilde{q}_{\theta^*},\mathcal{D}_n)}$ *are worst-case squared* l_2 *risks; and* $\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) := \inf_{\hat{\psi}:\mathcal{P}^{\otimes n}\to\Psi}\mathcal{W}^{l_2}_{\mathcal{P}_n^{\Pi},n}(\hat{\psi})$ *denotes the minimax lower bound over* $\mathcal{P}_n^{\Pi}$*.*

Theorem 4 establishes a *rate*, rather than an exact notion of optimality. Notably, both strategies (i) and (ii) are generally *not Bayes-optimal under the original prior* Π (see Eq. (20)): as (i) modifies the training labels, and (ii) is only Bayes-optimal wrt. the corrected prior $\widetilde{\Pi}$. Nevertheless, whenever the matching minimax lower bound satisfies $\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \gtrsim n^{-1} + r_{2,n}^2$ (which, in particular, holds for Examples A–C, see Appendix E), the worst-case risk of the debiased PFN matches the minimax rate, while its original-prior-averaged risk remains bounded by the same rate (see Eq. (21)). Hence, the debiased PFNs are **globally minimax-rate optimal over the class** $\mathcal{P}_n^{\Pi}$.

5 Experiments

Setup. To validate our theory, we evaluate the debiased PFN with ***(i) label correction*** (Sec. 4.2) for the ATE (Example C with $w \equiv 1$). We use a synthetic prior Π over nonlinear structural causal models with observed confounding and heterogeneous treatment effects (Javurek et al., 2026). For each $n \in \{256, 512, 1024, 2048\}$, we train a *standard PFN* (Eq. (6)) and a *debiased PFN* (Eq. (12)) with the same architecture, training data, and optimization budget, so that the two differ only in the training label. Details are in Appendix F.2.

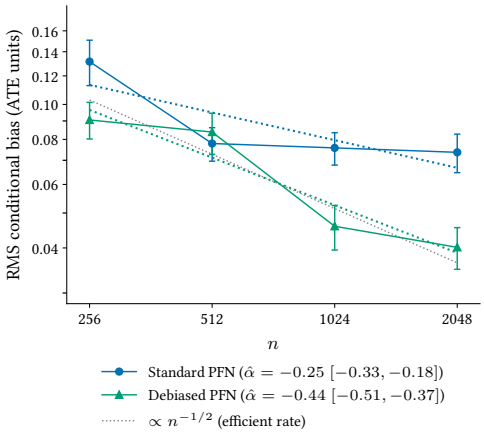

Figure 3: **RMS conditional bias of the posterior mean**, mean over $M = 200$ ground-truth P_0, each with $D = 200$ datasets. Displayed: power-law regression fits, RMS $\sim cn^{\alpha}$ (colored dotted lines); parametric rate, $\alpha = -0.5$ (grey dotted lines). Error bars: 95% bootstrap intervals. Lower is better; x- and y-axes are logarithmically scaled.

Evaluation. We want to evaluate the posteriors locally, at a *fixed* ground truth P_0 and, hence, to empirically validate the local optimality properties suggested by Theorems 1–3. For this, we draw $M = 200$ samples of $P_0 \sim \Pi$ with guaranteed strong overlap ($\kappa \geq 0.01$) and sample $D = 200$ datasets $\mathcal{D}_n \sim P_0^{\otimes n}$ from each P_0. Here, we report the empirical *root-mean-square (RMS) conditional bias of the posterior mean*, namely RMS $:= \left(\mathbb{E}_{P_0\sim\Pi}\left\{\mathbb{E}_{\mathcal{D}_n\sim P_0^{\otimes n}}\left[\hat{\psi}_{\theta^\star}(\mathcal{D}_n) - \psi(P_0)\right]\right\}^2\right)^{1/2}$. Importantly, this quantity *differs* from the Bayes risk (Remark 1) and allows us to directly evaluate *the squared bias* wrt. the P_0. In Appendix F.3, we further report other metrics: the posterior variance relative to the efficient variance $V_{\text{eff}}(P_0)/n$ and the conditional coverage of credible intervals.

Results. Results are shown in Fig. 3. Here, the conditional bias of the debiased PFN decays at a rate comparable to the parametric $n^{\alpha=-1/2}$ ($\hat{\alpha} = -0.44$, 95% CI $[-0.51, -0.37]$), whereas that of the standard PFN decays **markedly slower** ($\hat{\alpha} = -0.25$, $[-0.33, -0.18]$), as suggested by Theorems 1 and 2.

Other results. Additionally, we show that the Bayesian bootstrap in Eq. (12) (as a part of (i) label correction) is essential for calibrated uncertainty (see Appendix F.4), and that it adds the first-order sampling variance (see Appendix F.5). We refer to Appendix F for further experiments.

Conclusion. Our work provides two generic strategies for debiasing PFNs for semi-parametric inference with rigorous and provable asymptotic optimality guarantees.

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based coding assistants and agents) to implement methods, to provide feedback on the design of the experiments, and to assist in interpreting experimental results. We have also used generative AI tools to aid the formulation of mathematical claims, proofs, and hypotheses. All synthetic data were produced by our own simulation code, not generated by AI models; translation, dataset cleaning, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code and scientific figures and to draft and edit parts of the paper. We have reviewed all AI-assisted work. All theoretical results and proofs were checked and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is methodological and uses only simulated data; it involves no human subjects, personal data, or data releases. We do not foresee specific ethical concerns beyond the general caveat that uncertainty estimates from amortized models should be validated before use in high-stakes decisions.

REPRODUCIBILITY STATEMENT

The assumptions for all theoretical results are stated in the main text, and complete proofs are given in Appendix D. The data-generating prior, the training procedure of all PFN variants, and the evaluation protocol are described in Sec. 5 and Appendix F.3, with additional results in Appendix F. Anonymous source code covering data generation, training, inference, and metric computation, including a pipeline script with all random seeds fixed, is available at `https://anonymous.4open.science/r/Ortho-PFN-Anonymized-E4B7`.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *Advances in Neural Information Processing Systems*, 2017.

Ahmed M. Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, 2018.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Sivaraman Balakrishnan, Edward Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. *Statistical Science*, 41(3):659–670, 2026.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In *Advances in Neural Information Processing Systems*, 2025.

Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer New York, 1993.

PJ Bickel and BJK Kleijn. The semiparametric Bernstein-von Mises theorem. *The Annals of Statistics*, pp. 206–237, 2012.

Samuel Böhm, Lennart Purucker, Frank Hutter, and Pascal Schlosser. SurvPFN: Towards foundation models for survival predictions. *arXiv preprint arXiv:2606.04564*, 2026.

Marie Brockschmidt, Santo M. A. R. Thies, Maresa Schröder, Dennis Frauen, Valentyn Melnychuk, Maximilian Muschalik, Eyke Hüllermeier, and Stefan Feuerriegel. ConfoundingSHAP: Quantifying confounding strength in causal inference. *arXiv preprint arXiv:2605.10533*, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Victor Chernozhukov, Whitney Newey, Vıctor M. Quintas-Martınez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, 2022.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax optimal in-context nonparametric regression with transformers. In *International Conference on Machine Learning*, 2026.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 2010.

A. Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

Iván Díaz and Michael Rosenblum. Targeted maximum likelihood estimation using exponential families. *The International Journal of Biostatistics*, 11(2):233–251, 2015.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha V. Naidu, and Colin White. ForecastPFN: Synthetically-trained zero-shot forecasting. *Advances in Neural Information Processing Systems*, 2023.

Joao Fonseca and Julia Stoyanovich. ExplainerPFN: Towards tabular foundation models for model-free zero-shot feature importance estimations. *arXiv preprint arXiv:2601.23068*, 2026.

Sandra Fortini, Kenyon Ng, Sonia Petrone, Judith Rousseau, and Susan Wei. A principled framework for uncertainty decomposition in TabPFN. *arXiv preprint arXiv:2602.04596*, 2026.

David Freedman. Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141, 1999.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S.M. Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, 2018.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *The Annual Meeting of the Cognitive Science Society*, 2014.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. TabPFN-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. TabPFN-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Susan Gruber and Mark J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1):26, 2010.

P. Richard Hahn, Carlos M Carvalho, David Puelz, et al. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1), 2018.

J.A. Hartigan. Asymptotic normality of posterior distributions. In *Bayes theory*, pp. 107–118. Springer, 1983.

Tobias Hatt and Stefan Feuerriegel. Estimating average treatment effects via orthogonal regularization. In *International Conference on Information and Knowledge Management*, 2021.

Christian L. Hines and Oliver J. Hines. Automatic debiasing of neural networks via moment-constrained learning. In *Conference on Causal Learning and Reasoning*, 2025.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Manuel Hülskamp, Julius Kobialka, Emanuel Sommer, and David Rügamer. On the uncertainty in prior-data fitted network pretraining. In *ICML Workshop on Foundation Models for Structured Data*, 2026.

Emil Javurek, Dennis Frauen, Marie Brockschmidt, Jonas Schweisthal, and Stefan Feuerriegel. Amortizing causal sensitivity analysis via prior data-fitted networks. *arXiv preprint arXiv:2605.10590*, 2026.

Jikai Jin and Vasilis Syrgkanis. Sharp structure-agnostic lower bounds for general linear functional estimation. *arXiv preprint arXiv:2512.17341*, 2025a.

Jikai Jin and Vasilis Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. In *Annual Conference on Learning Theory*, 2025b.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pp. 207–236, 2024.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. *Advances in Neural Information Processing Systems*, 2024.

Tuan Anh Le, Atilim Gunes Baydin, and Frank Wood. Inference compilation and universal probabilistic programming. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Fan Li, Peng Ding, and Fabrizia Mealli. Bayesian causal inference: A critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153, 2023.

Antonio R. Linero and Joseph L. Antonelli. The how and why of Bayesian nonparametric causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1583, 2023.

Tianyi Ma, Tengyao Wang, and Richard J. Samworth. Optimal in-context adaptivity and distributional robustness of transformers. *arXiv preprint arXiv:2510.23254*, 2025.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *International Conference on Learning Representations*, 2026.

Divyat Mahajan, Jannes Gladrow, Agrin Hilmkil, Cheng Zhang, and Meyer Scetbon. Amortized inference of causal models via conditional fixed-point iterations. *Transactions on Machine Learning Research*, 2025.

Valentyn Melnychuk, Vahid Balazadeh, Stefan Feuerriegel, and Rahul G. Krishnan. Frequentist consistency of prior-data fitted networks for causal inference. In *International Conference on Machine Learning*, 2026.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Samuel Müller, Matthias Feurer, Noah Hollmann, and Frank Hutter. PFNs4BO: In-context learning for Bayesian optimization. In *International Conference on Machine Learning*, 2023.

Samuel Müller, Arik Reuter, Noah Hollmann, David Rügamer, and Frank Hutter. Position: The future of Bayesian prediction is prior-fitted. In *International Conference on Machine Learning Position Paper Track*, 2025.

Razieh Nabi. Semiparametric efficiency theory as differential calculus on a space of probability distributions. *arXiv preprint arXiv:2606.22784*, 2026.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *International Conference on Machine Learning*, 2023.

Thomas Nagler and David Rügamer. Uncertainty quantification for prior-data fitted networks using martingale posteriors. *arXiv preprint arXiv:2505.11325*, 2025.

Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6): 1349–1382, 1994.

Kenyon Ng, Edwin Fong, David T. Frazier, Jeremias Knoblauch, and Susan Wei. TabMGP: Martingale posterior with TabPFN. *arXiv preprint arXiv:2510.25154*, 2025.

Shi-ang Qi, Vahid Balazadeh, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. SurvivalPFN: Amortizing survival prediction via in-context Bayesian inference. *arXiv preprint arXiv:2605.15488*, 2026.

Kolyan Ray and Botond Szabó. Debiased Bayesian inference for average treatment effects. In *Advances in Neural Information Processing Systems*, 2019.

Kolyan Ray and Aad van der Vaart. Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.

Arik Reuter, Tim G.J. Rudner, Vincent Fortuin, and David Rügamer. Can transformers learn full Bayesian inference in context? In *International Conference on Machine Learning Position Paper Track*, 2025.

Arik Reuter, Anish Dhir, Cristiana Diaconu, Jake Robertson, Ole Ossen, Frank Hutter, Adrian Weller, Mark van der Wilk, and Bernhard Schölkopf. Use what you know: Causal foundation models with partial graphs. In *International Conference on Machine Learning*, 2026.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In *Advances in Neural Information Processing Systems*, 2025.

James M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

Dmitrii Seletkov, Paul Hager, Georgios Kaissis, Rickmer Braren, Daniel Rueckert, and Raphael Rehms. Survival in-context: Amortized Bayesian survival analysis via prior-fitted networks. *arXiv preprint arXiv:2603.29475*, 2026.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 2019.

Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. TimePFN: Effective multivariate time series forecasting with synthetic data. In *AAAI Conference on Artificial Intelligence*, 2025.

Anastasios A. Tsiatis. *Semiparametric theory and missing data*. Springer New York, 2006.

Alexandre B. Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Lars van der Laan and Mark van der Laan. Calibeating prediction-powered inference. *arXiv preprint arXiv:2604.21260*, 2026.

Mark J. van der Laan and Daniel B. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2, 2006.

Aad van der Vaart and Harry van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(6), 2011.

Aad W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, Cambridge, United Kingdom, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022.

Yun Yang, Guang Cheng, and David B. Dunson. Semiparametric Bernstein-von Mises theorem: Second order studies. *arXiv preprint arXiv:1503.04493*, 2015.

Andrew Yiu, Edwin Fong, Chris Holmes, and Judith Rousseau. Semiparametric posterior corrections. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1025–1054, 2025.

A Extended Related Work

A.1 Prior-data Fitted Networks

Amortized Bayesian inference & in-context learning. Amortized Bayesian inference (ABI) replaces repeated, dataset-specific posterior computation by a learned inference map that shares computation across related inference problems (Gershman & Goodman, 2014; Le et al., 2017). Closely related ideas appear in neural processes, which learn distributions over functions conditional on observed context sets (Garnelo et al., 2018; Kim et al., 2019). More recently, the emergence of *in-context learning* (ICL) in transformers has provided a particularly flexible mechanism for such amortization: A pretrained model adapts to a new task through its context alone, without updating its parameters (Brown et al., 2020; Dong et al., 2024). Several theoretical perspectives have been proposed to explain ICL, including implicit Bayesian inference (Xie et al., 2022), meta-learning, and the in-context implementation of statistical learning algorithms. The Bayesian and ICL viewpoints become especially closely connected when pretraining tasks are themselves sampled from a prior: conditioning on the in-context dataset then corresponds to learning a posterior predictive rule (Müller et al., 2022).

PFNs. Prior-data fitted networks (PFNs) instantiate this connection between ABI and ICL explicitly (Müller et al., 2022). Rather than fitting a model separately to each observed dataset, a PFN is pretrained on synthetic datasets sampled from a user-specified prior over data-generating processes. By predicting held-out quantities conditional on the remaining observations, the network learns to approximate the corresponding Bayesian posterior-predictive distribution; after pretraining, inference on a new dataset reduces to a forward pass through the network. A prominent realization is TabPFN (Hollmann et al., 2023), which brought the PFN paradigm to tabular classification. Its substantially extended successor supports both classification and regression, categorical and missing features, and datasets with up to thousands of observations, while achieving strong predictive performance on small- and medium-sized tabular datasets (Hollmann et al., 2025). The PFN paradigm has since been extended beyond generic tabular prediction, including Bayesian optimization (Müller et al., 2023), time-series forecasting (Dooley et al., 2023; Ansari et al., 2024; Taga et al., 2025), and, more recently, survival analysis (Seletkov et al., 2026; Qi et al., 2026; Böhm et al., 2026); see Müller et al. (2025) for a broader perspective.

PFNs for semi-parametric inference. To our knowledge, explicit applications of PFNs to semi-parametric target estimation have so far concentrated on *causal inference*. Balazadeh et al. (2025) introduced CausalPFN, which pretrains on observational datasets generated under unconfoundedness and directly amortizes estimation of average and heterogeneous treatment effects. Furthermore, Robertson et al. (2025) proposed Do-PFN, which instead pretrains across heterogeneous structural causal models and interventions, thereby learning to predict interventional outcomes without requiring the causal graph at test time. Going beyond these specific constructions, Ma et al. (2026) developed CausalFM, a general framework for constructing PFN priors from structural causal models covering back-door, front-door, and instrumental-variable settings. Importantly, they formalized conditions under which such priors yield identifiable and consistent Bayesian causal estimators, highlighting that an inappropriate causal pretraining prior may remain biased even asymptotically. Subsequent work further incorporated partial causal-graph information (Reuter et al., 2026) or extended PFN-based amortization to causal sensitivity analysis (Javurek et al., 2026). Again, all the above-mentioned works (at best) established the consistency of estimation *without further studying the asymptotic properties of the PFNs*.

A.2 Semi-parametric Inference

Frequentist estimation. Classical semi-parametric inference constructs estimators of a low-dimensional target while treating other components of the data-generating process as infinite-dimensional nuisance functions (Kennedy, 2024). In causal inference, in particular, prominent examples include outcome-regression plug-in estimators, inverse-probability weighting, and estimators that combine the two. A central example is an *augmented inverse probability treatment weighted* (A-IPTW) estimator. Rather than relying solely on estimated nuisance functions, A-IPTW augments a plug-in estimate by an influence-function-based correction. Under standard regularity conditions and sufficiently accurate nuisance estimation, this construction yields $\sqrt{n}$-consistent, asymptotically normal, and semi-parametrically efficient estimation of finite-dimensional functionals

such as the ATE (Newey, 1994; Robins, 2000; Bickel et al., 1993; Tsiatis, 2006). On the other hand, *targeted maximum likelihood estimation* (TMLE) provides an alternative implementation of the same first-order debiasing principle and is asymptotically equivalent under suitable conditions (van der Laan & Rubin, 2006). Modern work increasingly combines such debiased (or orthogonal) estimators with flexible machine-learning estimates of the nuisance functions (e.g., Shi et al., 2019; Hatt & Feuerriegel, 2021; Chernozhukov et al., 2022; Hines & Hines, 2025). This separation between flexible nuisance learning and a target-specific correction is particularly important for our setting: We design our debiasing strategies such that the flexibility of PFNs does not come at the cost of slower target estimation rates.

Minimax optimality. Minimax theory characterizes the fundamental statistical difficulty of an estimation problem by asking for the smallest worst-case error achievable over a specified model class (Le Cam, 2012; van der Vaart, 2000; Tsybakov & Zaiats, 2009). A minimax lower bound thus establishes *a rate below which no estimator can uniformly improve*. At the same time, an upper bound can be obtained by analyzing existing estimators (e. g., debiased estimators from above). When the two bounds match, the corresponding estimator is *minimax optimal*. In semi-parametric problems, however, there is no "one-size-fits-all" optimal rate: The minimax rate generally depends on both the target functional and the restrictions imposed on the nuisance functions. Recent work has therefore derived lower bounds for specific classes of functionals: e. g., the integrated squared density and expected conditional covariance (Balakrishnan et al., 2026); weighted average treatment effects (Jin & Syrgkanis, 2025b); and general linear functionals with mixed-bias property (Jin & Syrgkanis, 2025a). These results motivate our target-specific analysis: In Appendix E, for each estimand, we compare the rate achieved by our debiased PFN with the corresponding minimax lower bound.

Bayesian estimation. Unlike the frequentist estimators, Bayesian semi-parametric inference places a prior on the unknown components of the data-generating process and propagates the resulting posterior uncertainty to the target estimand. Flexible nuisance models used for this purpose include Gaussian processes (Alaa & van der Schaar, 2017; 2018) and Bayesian additive regression trees (Chipman et al., 2010); see Li et al. (2023); Linero & Antonelli (2023) for broader discussions.

Bernstein–von Mises property of Bayesian estimators. For finite-dimensional functionals of an infinite-dimensional model, however, posterior consistency of the nuisance functions alone does not ensure valid uncertainty quantification for the target. The relevant asymptotic property is provided by *semi-parametric Bernstein–von Mises (BvM) theory* (Hartigan, 1983; Bickel & Kleijn, 2012). A BvM result establishes that, after appropriate centering and $\sqrt{n}$ scaling, the marginal posterior for the target approaches a normal distribution whose variance coincides with the semi-parametric efficiency bound. Consequently, Bayesian credible sets acquire an *asymptotically valid frequentist interpretation*. Establishing such a result is delicate because regularization or prior-induced bias (Hahn et al., 2018) in the nuisance components can propagate to the target and dominate its $\sqrt{n}$ fluctuations. This interaction between nuisance contraction, prior bias, and target-level posterior behavior is precisely the aspect of semi-parametric Bayesian theory that is most closely connected to our analysis.

B Background Materials

B.1 Efficient Influence Functions

We refer to Kennedy (2024) for a detailed introduction to semi-parametric efficiency theory and to Nabi (2026) for a geometric interpretation of efficient influence functions (EIFs). Below, we briefly introduce the concepts needed in our analysis; for more details, we refer to Bickel et al. (1993); Tsiatis (2006); van der Vaart (2000).

Parametric sub-models. Let $P_0 \in \mathcal{P}(\mathcal{Z})$ denote a fixed data-generating distribution. A one-dimensional regular parametric sub-model through P_0 is a collection $\{P_t : t \in (-\epsilon, \epsilon)\} \subseteq \mathcal{P}(\mathcal{Z})$ such that $P_{t=0} = P_0$ and the map $t \mapsto P_t$ is sufficiently smooth at $t = 0$. The corresponding *score* describes the infinitesimal direction in which the distribution is perturbed:

$$s(Z) := \frac{\partial}{\partial t} \log P_t(Z)\bigg|_{t=0}. \tag{22}$$

Any regular score satisfies $P_0 s = 0$ and $P_0 s^2 < \infty$, and therefore belongs to the Hilbert space $L_2^0(P_0) := \{f : P_0 f = 0, P_0 f^2 < \infty\}$, equipped with inner product $\langle f, g \rangle_{P_0} := P_0(fg)$. The collection of scores generated by all regular parametric sub-models determines the tangent space

$$\mathcal{T}(P_0) := \overline{\text{span}} \left\{ s : s = \frac{\partial}{\partial t} \log P_t\bigg|_{t=0} \text{ for some regular } P_t \right\} \subseteq L_2^0(P_0). \tag{23}$$

Thus, although $\mathcal{P}(\mathcal{Z})$ may be infinite-dimensional, a parametric sub-model probes it along a single local direction.

For a given target ψ, directions $s \in \mathcal{T}(P_0)$ along which $\psi(P_t)$ does not change to first order correspond to nuisance directions. Geometrically, the efficient influence function isolates the component of variation that is relevant for ψ from these target-irrelevant perturbations.

Pathwise derivatives. The functional $\psi : \mathcal{P}(\mathcal{Z}) \to \Psi$ is *pathwise differentiable* at P_0 if, along every regular parametric sub-model P_t through P_0, its directional derivative exists and defines a continuous linear functional of the corresponding score s:

$$\dot{\psi}_{P_0}(s) := \frac{\partial}{\partial t} \psi(P_t)\bigg|_{t=0}. \tag{24}$$

By the Riesz representation theorem, there then exists a unique element of the tangent space, denoted by $\phi_\psi(\cdot; P_0)$, such that

$$\dot{\psi}_{P_0}(s) = \langle \phi_\psi(\cdot; P_0), s \rangle_{P_0} = P_0 \left[\phi_\psi(Z; P_0)\, s(Z) \right], \qquad s \in \mathcal{T}(P_0). \tag{25}$$

This function is the *efficient influence function* (EIF), also called the canonical gradient, and satisfies $P_0 \phi_\psi(\cdot; P_0) = 0$. Equivalently, along a sufficiently smooth local path,

$$\psi(P_t) - \psi(P_0) = t \cdot P_0 \left[\phi_\psi(Z; P_0)\, s(Z) \right] + o(t). \tag{26}$$

The EIF therefore describes the first-order sensitivity of the target to perturbations of the underlying distribution. This infinitesimal representation is the local counterpart of the von–Mises expansion used throughout the main text.

The geometry also gives a useful interpretation of the EIF norm. By Cauchy–Schwarz,

$$\left| \dot{\psi}_{P_0}(s) \right| \le \| \phi_\psi(\cdot; P_0) \|_{L_2(P_0)} \cdot \| s \|_{L_2(P_0)}. \tag{27}$$

Hence, among unit-norm local perturbations, the EIF direction is the direction along which the target can change most rapidly, whenever such a score is attainable, or can be approximated, by regular parametric sub-models. This observation also motivates our prior correction, which introduces additional prior variation in a target-specific, EIF-aligned direction.

B.2 NOTIONS OF OPTIMALITY

We distinguish three notions of optimality that arise naturally in our analysis. *Local frequentist optimality* compares estimators in shrinking neighborhoods of a fixed data-generating distribution P_0; *global frequentist optimality* compares their worst-case behavior over an entire class of distributions; and *Bayes optimality* averages performance over distributions drawn from a prior Π. These notions are conceptually distinct and, in general, none implies the others.

Local frequentist optimality. Consider first a fixed P_0 and $\psi_0 = \psi(P_0)$. A regular asymptotically linear estimator $\hat{\psi}_n$ admits an expansion

$$\hat{\psi}_n - \psi_0 = P_n D(\cdot; P_0) + o_{P_0}(n^{-1/2}), \tag{28}$$

where $D(\cdot; P_0) \in L_2^0(P_0)$ is its influence function. Consequently,

$$\sqrt{n}\big(\hat{\psi}_n - \psi_0\big) \rightsquigarrow N\left(0, P_0 D^2(\cdot; P_0)\right). \tag{29}$$

Regularity requires this first-order behavior to remain stable under local perturbations of P_0 of order $n^{-1/2}$. Semi-parametric efficiency theory implies that the EIF has the smallest possible variance among influence functions of regular estimators:

$$P_0 D^2(\cdot; P_0) \geq P_0 \phi_\psi^2(\cdot; P_0) = \mathrm{Var}_{P_0}\left[\phi_\psi(Z; P_0)\right] =: V_{\mathrm{eff}}(P_0). \tag{30}$$

The quantity $V_{\mathrm{eff}}(P_0)$ is the *semi-parametric efficiency bound*. An estimator attaining this bound is locally efficient at P_0. The qualifier "local" is important: the comparison concerns infinitesimal, typically $n^{-1/2}$-scale, perturbations around a fixed P_0, rather than uniform performance over the entire model class.

For Bayesian procedures, a semi-parametric Bernstein–von Mises property provides the corresponding local frequentist guarantee. If a target posterior is asymptotically Gaussian with variance $V_{\mathrm{eff}}(P_0)/n$, its posterior uncertainty attains the same local efficiency bound and its credible sets have an asymptotically valid frequentist interpretation. This is the notion of local optimality later used for our debiased PFNs when the second-order remainder is $o_{P_0}(n^{-1/2})$.

Global frequentist optimality. Local efficiency does not characterize the behavior of an estimator uniformly over a potentially large class of data-generating distributions. For a class $\mathcal{P}' \subseteq \mathcal{P}(\mathcal{Z})$, we instead consider its worst-case squared l_2 risk

$$\mathcal{W}^{l_2}_{\mathcal{P}',n}(\hat{\psi}) := \sup_{P \in \mathcal{P}'} \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}\left[\hat{\psi}(\mathcal{D}_n) - \psi(P)\right]^2. \tag{31}$$

The corresponding minimax risk is

$$\mathfrak{M}_n(\mathcal{P}') := \inf_{\hat{\psi}: \mathcal{P}^{\otimes n} \to \Psi} \mathcal{W}^{l_2}_{\mathcal{P}',n}(\hat{\psi}). \tag{32}$$

The minimax risk is a property of the statistical problem itself: it gives the smallest worst-case error achievable by any estimator over $\mathcal{P}'$. If, for some sequence $r_n \to 0$,

$$\mathfrak{M}_n(\mathcal{P}') \gtrsim r_n^2 \qquad \text{and} \qquad \mathcal{W}^{l_2}_{\mathcal{P}',n}(\hat{\psi}) \lesssim r_n^2, \tag{33}$$

then $\hat{\psi}$ is *minimax-rate optimal* over $\mathcal{P}'$. Unlike local efficiency, global minimax optimality does not require a $\sqrt{n}$ rate or asymptotic normality: depending on the functional ψ and the complexity of the nuisance functions $\eta(P)$, the optimal rate r_n may be slower than $n^{-1/2}$.

In our analysis later, the relevant class is the effective prior-supported class $\mathcal{P}_n^\Pi$. Thus, the global question is whether the worst-case risk of a debiased PFN over $\mathcal{P}_n^\Pi$ matches a corresponding minimax lower bound. This notion is "global" relative to local efficiency because it controls the estimator uniformly across $\mathcal{P}_n^\Pi$, rather than only in shrinking neighborhoods of a single P_0.

Bayes optimality. Bayes optimality, on the other hand, evaluates an estimator by averaging its loss with respect to a prior Π over data-generating distributions. Under squared loss, the Bayes risk of an estimator $\hat{\psi}$ is

$$\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}) := \mathbb{E}_{P \sim \Pi} \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}\left[\hat{\psi}(\mathcal{D}_n) - \psi(P)\right]^2. \tag{34}$$

Conditioning on the observed dataset $\mathcal{D}_n$, the posterior mean

$$\hat{\psi}_n^{\Pi}(\mathcal{D}_n) := \mathbb{E}_{\Pi}\left[\psi(P) \mid \mathcal{D}_n\right] \tag{35}$$

minimizes the posterior expected squared loss and therefore also minimizes $\mathcal{R}_{\Pi,n}^{l_2}(\hat{\psi})$ over all estimators.

The Bayes risk is naturally related to the minimax risk introduced above. In particular, for any prior Π supported on a class $\mathcal{P}' \subseteq \mathcal{P}(\mathcal{Z})$,

$$\mathcal{R}_{\Pi,n}^{l_2}\left(\hat{\psi}_n^{\Pi}\right) = \inf_{\hat{\psi}} \mathcal{R}_{\Pi,n}^{l_2}(\hat{\psi}) \leq \inf_{\hat{\psi}} \sup_{P \in \mathcal{P}'} \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}\left[\hat{\psi}(\mathcal{D}_n) - \psi(P)\right]^2 = \mathfrak{M}_n(\mathcal{P}'). \tag{36}$$

Thus, the Bayes risk under any fixed prior provides a lower bound on the corresponding minimax risk. Equality may hold for a least-favorable prior, whereas a prior concentrating on statistically easier distributions can have a substantially smaller Bayes risk.

Bayes optimality is fundamentally an average-case criterion: it weights performance according to how likely different data-generating distributions are under Π. It therefore differs from local frequentist optimality, which concerns behavior in shrinking neighborhoods of a fixed P_0, and from global minimax optimality, which controls the worst-case risk over a class of distributions. In particular, an estimator can be Bayes optimal under Π while concentrating slowly for some fixed P_0 or performing poorly on distributions that receive little prior mass. Conversely, modifying the prior or the PFN training objective to improve frequentist concentration rates may sacrifice Bayes optimality under the original prior, while remaining Bayes optimal with respect to the modified prior.

C EXAMPLE C

Example C (Weighed average treatment effect (WATE))**.** *The WATE is one of the central estimands in causal inference (Hirano et al., 2003). Here, $Z = (X, A, Y = AY_1 + (1-A)Y_0)$ with pre-treatment covariates $X \in [0,1]^{d_x}$, a treatment $A \in \{0,1\}$, and the observed outcome $Y \in [0,1]$. Then, the WATE is defined as $\psi_w = \mathbb{E}[w(X;\pi)(Y_1 - Y_0)]/\mathbb{E}(w(X;\pi))$, where $w(\cdot;\pi) \geq 0$ is a user specified weighting function that depends on covariates x through a propensity score $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$. For example, $w(x;\pi) = 1$ yields an average treatment effect (ATE), and $w(x;\pi) = \pi(x)(1-\pi(x))$ yields the overlap-weighted ATE. Under standard causal assumptions (namely, unconfoundedness, $Y_0, Y_1 \perp\!\!\!\perp A \mid X$, and strong overlap, $\kappa < \pi(x) < 1-\kappa$, with $\kappa \in (0, 1/2)$), the WATE is identifiable as $\psi_w(P) = \mathbb{E}[w(X;\pi)(\mu_1(X) - \mu_0(X))]/\mathbb{E}[w(X;\pi)] = \int w(x;\pi)(\mu_1(x) - \mu_0(x))\,\mathrm{d}P(x) / \int w(x;\pi)\,\mathrm{d}P(x)$, where $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$ is a conditional expectation. Here, the corresponding nuisance functions, EIF, and second-order remainder are as follows:*

$$
\begin{aligned}
\eta(P) &= \{\mu_a(x), \mathbb{1}\{w' \not\equiv 1\}\pi(x), P(x)\},\\
\phi_{\psi,w}(Z;P) &= \frac{w(X;\pi)}{\bar{w}(P)}\left[\frac{A-\pi(X)}{\pi(X)(1-\pi(X))}(Y - \mu_A(X)) + \left((A-\pi(X))\cdot\frac{w'(X;\pi)}{w(X;\pi)} + 1\right)\big(\tau(X) - \psi_w(P)\big)\right],\\
R_{2,\psi,w}(P,P_0) &= \frac{1}{\bar{w}(P)}\Bigg[(\psi_w(P) - \psi_w(P_0))\cdot\big((P-P_0)w(\cdot;P) + P_0(\pi-\pi_0)w'(\cdot;\pi)\big)\\
&\qquad + P_0\Big\{\big(w(\cdot;\pi) - w(\cdot;\pi_0) - w'(\cdot;\pi)(\pi-\pi_0)\big)\cdot(\tau_0 - \psi_w(P_0))\Big\}\\
&\qquad + P_0\Bigg\{(\pi-\pi_0)\left(\left(\frac{w(\cdot;P)}{\pi} - w'(\cdot;P)\right)(\mu_1 - \mu_{1,0}) + \left(\frac{w(\cdot;P)}{1-\pi} + w'(\cdot;P)\right)(\mu_0 - \mu_{0,0})\right)\Bigg\},
\end{aligned}
\tag{37}
$$

where $\bar{w}(P) = \mathbb{E}[w(X;\pi)]$, $w'(\cdot;\pi)$ is a derivative of w wrt. to π, and $\tau(\cdot) = \mu_1(\cdot) - \mu_0(\cdot)$ is a conditional average treatment effect (CATE) (Hirano et al., 2003).

C.1 SLOW CONCENTRATION RATES

By extending Example B, we analogously show slow concentration rates for the WATE.

Example C (WATE)**.** *Assume that w is uniformly bounded, Lipschitz in π, and that $\mathbb{E}_P[w(X)]$ is bounded away from zero. Then, for a fixed pair of distributions P and P_0, it can be shown that*

$$
\begin{aligned}
|\psi_w - \psi_{w,0}| \lesssim\ & \|P - P_0\|_{L_2(x)} + \|\mu_1 - \mu_{1,0}\|_{L_2(P_0)} + \|\mu_0 - \mu_{0,0}\|_{L_2(P_0)}\\
&+ \mathbb{1}\{w \not\equiv \pi\}\,\|\pi - \pi_0\|_{L_2(P_0)}\,.
\end{aligned}
\tag{38}
$$

Hence, a corresponding nuisance-error norm is

$$
\begin{aligned}
\|\eta(P) - \eta(P_0)\|_{\mathcal{H}} :=\ & \|P - P_0\|_{L_2(x)} + \|\mu_1 - \mu_{1,0}\|_{L_2(P_0)} + \|\mu_0 - \mu_{0,0}\|_{L_2(P_0)}\\
&+ \mathbb{1}\{w \not\equiv \pi\}\,\|\pi - \pi_0\|_{L_2(P_0)}\,.
\end{aligned}
\tag{39}
$$

Again, this bound is locally rate-sharp. In particular, consider the path $\mu_{1,t}(x) = \mu_{1,0}(x) + t$, while keeping $\mu_{0,t} = \mu_{0,0}$, $\pi_t = \pi_0$, and $P_t(x) = P_0(x)$. Then $w_t = w_0$ and

$$
\psi_w(P_t) - \psi_w(P_0) = t, \qquad \|\mu_{1,t} - \mu_{1,0}\|_{L_2(P_0)} = |t|\,.
\tag{40}
$$

Similar paths can be analogously found wrt. other nuisance functions. Thus, whenever the posterior assigns asymptotically non-negligible mass around such paths, the WATE concentration error can inherit the concentration rate of the latent nuisance functions in the same order.

C.2 DEBIASING STRATEGIES

(i) Label correction. Compared to Example B, the label correction of the WATE additionally corrects for the propensity score errors.

Example C (WATE)**.** *Applying Eq.* (12) *with the EIF from Eq.* (37)*, the corrected label is*

$$
\tilde{\psi_w}_w(P) = \psi_w(P) + \mathrm{BB}_n\{\phi_{\psi,w}(Z;P)\} = \mathrm{BB}_n\{m_w(Z;P)\},
\tag{41}
$$

where the corresponding pseudo-outcome is

$$m_w(Z;P) = \psi_w(P) + \frac{1}{\bar{w}(P)}\Bigg[w(X;\pi)\Big\{\frac{A}{\pi(X)}\big(Y-\mu_1(X)\big) - \frac{1-A}{1-\pi(X)}\big(Y-\mu_0(X)\big)\Big\} + w(X;\pi)\big\{\tau(X)-\psi_w(P)\big\} + \big\{A-\pi(X)\big\}w'(X;\pi)\big\{\tau(X)-\psi_w(P)\big\}\Bigg]. \tag{42}$$

The first two terms inside the brackets correct the first-order errors arising from the outcome regressions and the covariate distribution, respectively, while the final term additionally corrects the first-order propensity-score error induced by the dependence of w on π. In particular, for the ATE, $w \equiv 1$ and hence $w' \equiv 0$, so the last term vanishes, and the label reduces to the usual A-IPTW pseudo-outcome. For the overlap-weighted ATE, $w(\pi) = \pi(1-\pi)$ and $w'(\pi) = 1-2\pi$, such that the additional propensity correction is generally non-zero.

(ii) Prior correction. Again, the prior correction of the WATE extends the previous Example B by additionally perturbing the prior for the propensity score.

Example C (WATE)**.** *Assume that a prior draw factorizes as $P(z) = P(x)\pi(x)^a(1-\pi(x))^{1-a}P(y \mid x,a)$ and that $P(y \mid x, A=a) = N(\mu_a(x);\sigma_a^2(x))$. Let $\bar{w}(P) = P[w(X;\pi)]$ and $\tau(x) = \mu_1(x) - \mu_0(x)$. One possible prior correction jointly perturbs $P(x)$, $\mu_1(x)$, $\mu_0(x)$, and $\pi(x)$ using the same λ:*

$$\tilde{P}(x) := \frac{\exp\Big\{\lambda\frac{w(x;\pi)}{\bar{w}(P)}\big(\tau(x)-\psi_w(P)\big)\Big\}}{C_{\psi,w}(P)}P(x), \quad \tilde{\mu}_1(x) := \mu_1(x) + \lambda\frac{\sigma_1^2(x)w(x;\pi)}{\bar{w}(P)\pi(x)},$$
$$\tilde{\mu}_0(x) := \mu_0(x) - \lambda\frac{\sigma_0^2(x)w(x;\pi)}{\bar{w}(P)(1-\pi(x))}, \quad \operatorname{logit}\tilde{\pi}(x) := \operatorname{logit}\pi(x) + \lambda\frac{w'(x;\pi)}{\bar{w}(P)}\big(\tau(x)-\psi_w(P)\big), \tag{43}$$

where

$$C_{\psi,w}(P) := P\left[\exp\left\{\lambda\frac{w(X;\pi)}{\bar{w}(P)}\big(\tau(X)-\psi_w(P)\big)\right\}\right]. \tag{44}$$

Indeed, at $\lambda = 0$, the four corresponding score contributions are

$$S[\tilde{P}] = \frac{w(X;\pi)}{\bar{w}(P)}\big(\tau(X)-\psi_w(P)\big),$$
$$S[\tilde{P}](Y \mid X, A) = \frac{w(X;\pi)}{\bar{w}(P)}\frac{A-\pi(X)}{\pi(X)(1-\pi(X))}\big(Y-\mu_A(X)\big),$$
$$S[\tilde{P}](A \mid X) = \big(A-\pi(X)\big)\frac{w'(X;\pi)}{\bar{w}(P)}\big(\tau(X)-\psi_w(P)\big), \tag{45}$$

whose sum equals $\phi_{\psi,w}(Z;P)$. Hence, the construction satisfies Eq. (15)*. Relative to Example B, the WATE correction perturbs both outcome regressions and, whenever w depends on π, additionally fluctuates the propensity score along its target-relevant direction. In particular, for the ATE with $w \equiv 1$, $w' \equiv 0$ and the propensity-score perturbation disappears.*

D THEORETICAL RESULTS

Here, we introduce additional notation. For a sequence of random vectors X_n and a positive deterministic sequence a_n, we write $X_n = \Omega_P(a_n)$ if $\|X_n\|_2/a_n$ is bounded away from zero in $P^{\otimes n}$-probability, i. e., for every $\epsilon > 0$, there exist constants $c_\epsilon > 0$ and $n_\epsilon < \infty$ such that $P^{\otimes n}(\|X_n\|_2/a_n \geq c_\epsilon) \geq 1 - \epsilon$ for all $n \geq n_\epsilon$. We write $X_n = \Theta_P(a_n)$ if both $X_n = O_P(a_n)$ and $X_n = \Omega_P(a_n)$. Finally, for two non-negative deterministic sequences a_n and b_n, we write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Remark 1 (Bayes-optimality of the idealized PFN)**.** *The posterior mean of an idealized PFN, i.e., $\hat{\psi}_{\theta^*}(\mathcal{D}_n) := \int \psi\, q_{\theta^*}(\psi \mid \mathcal{D}_n)\,\mathrm{d}\psi$, is Bayes-optimal under the squared l_2 risk $\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\hat{\psi}(\mathcal{D}_n) - \psi(P)]^2$:*

$$\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}_{\theta^*}) = \inf_{\hat{\psi}:\mathcal{P}^{\otimes n}\to\Psi} \mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}) = \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\mathrm{Var}_\Pi[\psi(P) \mid \mathcal{D}_n]]. \tag{46}$$

Proof. By definition of an idealized PFN, $q_{\theta^*}(\psi \mid \mathcal{D}_n)$ is the density of $\Pi\big(\psi(P) \mid \mathcal{D}_n\big)$, and hence its posterior mean satisfies

$$\hat{\psi}_{\theta^*}(\mathcal{D}_n) = \int \psi\, q_{\theta^*}(\psi \mid \mathcal{D}_n)\,\mathrm{d}\psi = \mathbb{E}_\Pi\big[\psi(P) \mid \mathcal{D}_n\big]. \tag{47}$$

Consider any estimator $\hat{\psi} : \mathcal{P}^{\otimes n} \to \Psi$. Conditional on $\mathcal{D}_n$, the squared error admits the usual bias–variance decomposition:

$$\mathbb{E}_\Pi\left[\big(\hat{\psi}(\mathcal{D}_n) - \psi(P)\big)^2 \Big| \mathcal{D}_n\right] = \Big(\hat{\psi}(\mathcal{D}_n) - \mathbb{E}_\Pi\big[\psi(P) \mid \mathcal{D}_n\big]\Big)^2 + \mathrm{Var}_\Pi\big[\psi(P) \mid \mathcal{D}_n\big]. \tag{48}$$

The second term does not depend on $\hat{\psi}$, while the first term is non-negative and is uniquely minimized by

$$\hat{\psi}(\mathcal{D}_n) = \mathbb{E}_\Pi\big[\psi(P) \mid \mathcal{D}_n\big] = \hat{\psi}_{\theta^*}(\mathcal{D}_n). \tag{49}$$

Taking expectations with respect to the joint distribution $P \sim \Pi$ and $\mathcal{D}_n \sim P^{\otimes n}$ therefore yields

$$\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}) = \mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}_{\theta^*}) + \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\left[\Big(\hat{\psi}(\mathcal{D}_n) - \hat{\psi}_{\theta^*}(\mathcal{D}_n)\Big)^2\right] \geq \mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}_{\theta^*}). \tag{50}$$

Consequently,

$$\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}_{\theta^*}) = \inf_{\hat{\psi}:\mathcal{P}^{\otimes n}\to\Psi} \mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}) = \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\left[\mathrm{Var}_\Pi\big[\psi(P) \mid \mathcal{D}_n\big]\right]. \tag{51}$$

This proves the claim. □

Theorem 1 (Slow local L_2-concentration rates (full version))**.** *Let q_{θ^*} be an idealized PFN, such that $q_{\theta^*}(\psi \mid \mathcal{D}_n)$ equals the push-forward of $\Pi(P \mid \mathcal{D}_n)$ through $\psi(\cdot) : P \to \psi(P)$. Fix $P_0 \in \operatorname{supp}\Pi$, let $\psi_0 := \psi(P_0)$, and suppose that $\mathcal{D}_n \sim P_0^{\otimes n}$. Assume that, for every relevant $P \in \operatorname{supp}\Pi$, the target admits the von–Mises expansion from Eq.* (1)*, and that the first-order term is locally Lipschitz in the nuisance functions: for some constant $C_\psi(P_0) < \infty$,*

$$|P_0\phi_\psi(\cdot; P)| \leq C_\psi(P_0) \cdot \|\eta(P) - \eta(P_0)\|_{\mathcal{H}}\,. \tag{52}$$

Then, whenever the corresponding posterior second moments are finite,

$$\|\psi - \psi_0\|_{L_2(q_{\theta^*},\mathcal{D}_n)} \lesssim \underbrace{\left[\int \|\eta(P) - \eta(P_0)\|^2_{\mathcal{H}}\ \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}}_{F_\psi(P_0,\mathcal{D}_n,\Pi)} + \underbrace{\left[\int R^2_{2,\psi}(P,P_0)\,\Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}}_{R_{2,\psi}(P_0,\mathcal{D}_n,\Pi)}. \tag{53}$$

In particular, if for deterministic sequences $\varepsilon_n \to 0$ and $r_{2,n} \to 0$,

$$F_\psi(P_0,\mathcal{D}_n,\Pi) = O_{P_0}(\varepsilon_n), \qquad R_{2,\psi}(P_0,\mathcal{D}_n,\Pi) = O_{P_0}(r_{2,n}), \tag{54}$$

then $\|\psi - \psi_0\|_{L_2(q_{\theta^*},\mathcal{D}_n)} = O_{P_0}(\varepsilon_n + r_{2,n})$. Hence, if $r_{2,n} = o(\varepsilon_n)$, the target posterior concentrates at most at the first-order nuisance rate $O_{P_0}(\varepsilon_n)$.

Moreover, this first-order bound is rate-sharp whenever the posterior places asymptotically non-negligible mass on target-relevant nuisance directions. In particular, suppose that there exist constants $0 < c_1 < c_2 < \infty$, $c_3 > 0$, and $\delta > 0$, together with measurable sets $\mathcal{A}_n \subseteq \mathcal{P}(\mathcal{Z})$, such that, with $P_0^{\otimes n}$-probability tending to one,

$$\Pi(\mathcal{A}_n \mid \mathcal{D}_n) \geq \delta, \tag{55}$$

and every $P \in \mathcal{A}_n$ satisfies

$$c_1\varepsilon_n \leq \|\eta(P) - \eta(P_0)\|_{\mathcal{H}} \leq c_2\varepsilon_n, \quad \text{and} \quad |\psi(P) - \psi_0| \geq c_3 \|\eta(P) - \eta(P_0)\|_{\mathcal{H}}. \tag{56}$$

If additionally $F_\psi(P_0, \mathcal{D}_n, \Pi) = O_{P_0}(\varepsilon_n)$ and $R_{2,\psi}(P_0, \mathcal{D}_n, \Pi) = o_{P_0}(\varepsilon_n)$, then

$$\begin{aligned} F_\psi(P_0, \mathcal{D}_n, \Pi) &= \Theta_{P_0}(\varepsilon_n), \\ \|\psi - \psi_0\|_{L_2(q_{\theta^*}, \mathcal{D}_n)} &= \Theta_{P_0}(\varepsilon_n). \end{aligned} \tag{57}$$

Thus, without an explicit first-order correction, the target posterior can inherit the concentration rate of the latent nuisance functions.

Proof. Since q_{θ^*} is an idealized PFN, its target posterior is exactly the push-forward of $\Pi(P \mid \mathcal{D}_n)$ under the map $\psi(\cdot) : P \to \psi(P)$. Hence,

$$\|\psi - \psi_0\|^2_{L_2(q_{\theta^*}, \mathcal{D}_n)} = \int \big(\psi - \psi_0\big)^2 q_{\theta^*}(\psi \mid \mathcal{D}_n)\, \mathrm{d}\psi = \int \big(\psi(P) - \psi_0\big)^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n). \tag{58}$$

Applying the von–Mises expansion from Eq. (1) gives

$$\psi(P) - \psi_0 = -P_0\phi_\psi(\cdot; P) + R_{2,\psi}(P, P_0). \tag{59}$$

Therefore, by the triangle inequality in $L_2(\Pi(\cdot \mid \mathcal{D}_n))$,

$$\|\psi - \psi_0\|_{L_2(q_{\theta^*}, \mathcal{D}_n)} \leq \left[\int \big\{P_0\phi_\psi(\cdot; P)\big\}^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} + \left[\int R^2_{2,\psi}(P, P_0)\Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}. \tag{60}$$

By Eq. (52),

$$\begin{aligned} \left[\int \big\{P_0\phi_\psi(\cdot; P)\big\}^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} &\leq C_\psi(P_0)\left[\int \|\eta(P) - \eta(P_0)\|^2_{\mathcal{H}}\, \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} \\ &= C_\psi(P_0) F_\psi(P_0, \mathcal{D}_n, \Pi). \end{aligned} \tag{61}$$

Combining this inequality with Eq. (60) proves Eq. (53). The rate statement in Eq. (54) follows immediately from $F_\psi(P_0, \mathcal{D}_n, \Pi) = O_{P_0}(\varepsilon_n)$ and $R_{2,\psi}(P_0, \mathcal{D}_n, \Pi) = O_{P_0}(r_{2,n})$.

It remains to establish rate-sharpness. On the event in Eq. (55), the definition of F_ψ, and the first inequality in Eq. (56) imply

$$\begin{aligned} F^2_\psi(P_0, \mathcal{D}_n, \Pi) &= \int \|\eta(P) - \eta(P_0)\|^2_{\mathcal{H}}\, \Pi(\mathrm{d}P \mid \mathcal{D}_n) \\ &\geq \int_{\mathcal{A}_n} \|\eta(P) - \eta(P_0)\|^2_{\mathcal{H}}\, \Pi(\mathrm{d}P \mid \mathcal{D}_n) \\ &\geq c_1^2\varepsilon_n^2 \Pi(\mathcal{A}_n \mid \mathcal{D}_n) \geq c_1^2\delta\varepsilon_n^2. \end{aligned} \tag{62}$$

Together with the assumed upper bound $F_\psi(P_0, \mathcal{D}_n, \Pi) = O_{P_0}(\varepsilon_n)$, this gives

$$F_\psi(P_0, \mathcal{D}_n, \Pi) = \Theta_{P_0}(\varepsilon_n). \tag{63}$$

Similarly, using the push-forward identity and the second inequality in Eq. (56),

$$\begin{aligned} \|\psi - \psi_0\|^2_{L_2(q_{\theta^*}, \mathcal{D}_n)} &= \int \big(\psi(P) - \psi_0\big)^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n) \\ &\geq \int_{\mathcal{A}_n} \big(\psi(P) - \psi_0\big)^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n) \\ &\geq c_3^2 \int_{\mathcal{A}_n} \|\eta(P) - \eta(P_0)\|^2_{\mathcal{H}}\, \Pi(\mathrm{d}P \mid \mathcal{D}_n) \\ &\geq c_1^2 c_3^2 \delta\varepsilon_n^2. \end{aligned} \tag{64}$$

Consequently,

$$\|\psi - \psi_0\|_{L_2(q_{\theta^*}, \mathcal{D}_n)} = \Omega_{P_0}(\varepsilon_n). \tag{65}$$

On the other hand, Eq. (53), together with $F_\psi(P_0, \mathcal{D}_n, \Pi) = O_{P_0}(\varepsilon_n)$ and $R_{2,\psi}(P_0, \mathcal{D}_n, \Pi) = o_{P_0}(\varepsilon_n)$, gives

$$\|\psi - \psi_0\|_{L_2(q_{\theta^*}, \mathcal{D}_n)} = O_{P_0}(\varepsilon_n). \tag{66}$$

Combining the lower and upper bounds proves Eq. (57). □

Theorem 2 (L_2-concentration rates of debiased PFNs (full version))**.** *Assume the conditions of Theorem 1. Fix $P_0 \in \operatorname{supp}\Pi$, let $\psi_0 := \psi(P_0)$, and suppose that $\mathcal{D}_n \sim P_0^{\otimes n}$. Write $P_n f := n^{-1}\sum_{i=1}^n f(Z_i)$ and $\phi_P := \phi_\psi(\cdot; P)$, with $\phi_0 := \phi_\psi(\cdot; P_0)$. Assume that $\phi_0 \in L_2(P_0)$ and that the posterior is stochastically equicontinuous in the sense that*

$$\left[\int \left\{(P_n - P_0)(\phi_P - \phi_0)\right\}^2 \Pi_\star(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} = o_{P_0}(n^{-1/2}), \tag{67}$$

where $\Pi_\star = \Pi$ for the label correction and $\Pi_\star$ denotes the P-marginal of the corrected posterior for the prior correction.

For strategy (i), additionally assume

$$\int P_n\left[\phi_P^2\right] \Pi(\mathrm{d}P \mid \mathcal{D}_n) = O_{P_0}(1). \tag{68}$$

For strategy (ii), let $\tilde{P} = \mathrm{LF}_\psi(P; \lambda)$, define

$$V_\psi(P) := P\left[\phi_P^2\right], \tag{69}$$

and suppose that the least-favorable path is locally regular so that

$$\psi\big(\mathrm{LF}_\psi(P; \lambda)\big) - \psi(P) = \lambda V_\psi(P) + r_{\mathrm{LF},\psi}(P, \lambda). \tag{70}$$

Furthermore, assume the standard local asymptotic normality and prior-flatness conditions along this path, including $n\sigma_n^2 \to \infty$, such that the joint corrected posterior satisfies

$$\left[\int \left\{\lambda V_\psi(P) - P_n\phi_P\right\}^2 \widetilde{\Pi}(\mathrm{d}P \times \mathrm{d}\lambda \mid \mathcal{D}_n)\right]^{1/2} = O_{P_0}(n^{-1/2}). \tag{71}$$

Then, for either debiasing strategy (i)–(ii), the L_2-concentration error of the corresponding idealized debiased PFN satisfies

$$\left\|\tilde{\psi} - \psi_0\right\|_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)} \le \underbrace{\widetilde{F}_\psi(P_0, \mathcal{D}_n, \star)}_{O_{P_0}(n^{-1/2})} + R_{2,\psi}(P_0, \mathcal{D}_n, \star), \qquad \star = \begin{cases} \Pi, & \textit{(i) label correction}, \\ \widetilde{\Pi}, & \textit{(ii) prior correction}, \end{cases} \tag{72}$$

where the corrected first-order terms are

$$\widetilde{F}_\psi(P_0, \mathcal{D}_n, \Pi) := \left[\int \mathbb{E}_W\left[\left\{(\mathrm{BB}_n - P_0)\phi_P\right\}^2 \,\middle|\, \mathcal{D}_n, P\right] \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}, \tag{73}$$

$$\widetilde{F}_\psi(P_0, \mathcal{D}_n, \widetilde{\Pi}) := \left[\int \left\{(P_n - P_0)\phi_P + \lambda V_\psi(P) - P_n\phi_P\right\}^2 \widetilde{\Pi}(\mathrm{d}P \times \mathrm{d}\lambda \mid \mathcal{D}_n)\right]^{1/2}. \tag{74}$$

The corresponding posterior-averaged second-order remainders are

$$R_{2,\psi}(P_0, \mathcal{D}_n, \Pi) := \left[\int R_{2,\psi}^2(P, P_0) \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}, \tag{75}$$

$$R_{2,\psi}(P_0, \mathcal{D}_n, \widetilde{\Pi}) := \left[\int \left\{R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda)\right\}^2 \widetilde{\Pi}(\mathrm{d}P \times \mathrm{d}\lambda \mid \mathcal{D}_n)\right]^{1/2}. \tag{76}$$

Proof. We consider the two correction strategies separately.

(i) Label correction. Recall from Eq. (12) that

$$\tilde{\psi}(P) = \psi(P) + \mathrm{BB}_n\{\phi_P\}. \tag{77}$$

Since the target functional satisfies the von–Mises expansion,

$$\psi(P) - \psi_0 = -P_0\phi_P + R_{2,\psi}(P, P_0), \tag{78}$$

we obtain the exact decomposition

$$\tilde{\psi}(P) - \psi_0 = (\mathrm{BB}_n - P_0)\phi_P + R_{2,\psi}(P, P_0). \tag{79}$$

For an idealized label-corrected PFN, $\tilde{q}_{\theta^*}$ is the conditional distribution induced by drawing $P \sim \Pi(P \mid \mathcal{D}_n)$ and independent Bayesian-bootstrap weights W. Hence, by Eq. (79) and the triangle inequality in the corresponding L_2 space,

$$\left\|\tilde{\psi} - \psi_0\right\|_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)} \le \widetilde{F}_\psi(P_0, \mathcal{D}_n, \Pi) + R_{2,\psi}(P_0, \mathcal{D}_n, \Pi). \tag{80}$$

It remains to show that the first term is of order $n^{-1/2}$. Decompose

$$(\mathrm{BB}_n - P_0)\phi_P = (\mathrm{BB}_n - P_n)\phi_P + (P_n - P_0)\phi_0 + (P_n - P_0)(\phi_P - \phi_0). \tag{81}$$

Applying the L_2 triangle inequality to Eq. (81) gives

$$\begin{aligned}\widetilde{F}_\psi(P_0, \mathcal{D}_n, \Pi) \le &\left[\int \mathbb{E}_W\left[\left\{(\mathrm{BB}_n - P_n)\phi_P\right\}^2 \,\middle|\, \mathcal{D}_n, P\right]\Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} \\ &+ |(P_n - P_0)\phi_0| \\ &+ \left[\int \left\{(P_n - P_0)(\phi_P - \phi_0)\right\}^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2}.\end{aligned} \tag{82}$$

For Bayesian-bootstrap weights $W \sim \mathrm{Dir}(n; 1, \ldots, 1)$,

$$\mathbb{E}_W\left[\left\{(\mathrm{BB}_n - P_n)\phi_P\right\}^2 \,\middle|\, \mathcal{D}_n, P\right] = \frac{P_n\left[\left(\phi_P - P_n\phi_P\right)^2\right]}{n+1} \le \frac{P_n[\phi_P^2]}{n+1}. \tag{83}$$

Therefore, Eq. (68) implies that the first term is $O_{P_0}(n^{-1/2})$. Since $\phi_0 \in L_2(P_0)$,

$$(P_n - P_0)\phi_0 = O_{P_0}(n^{-1/2}) \tag{84}$$

by the central limit theorem, while the final term is $o_{P_0}(n^{-1/2})$ by Eq. (67). Consequently,

$$\widetilde{F}_\psi(P_0, \mathcal{D}_n, \Pi) = O_{P_0}(n^{-1/2}). \tag{85}$$

(ii) Prior correction. Write $\tilde{P} = \mathrm{LF}_\psi(P; \lambda)$. By construction of the least-favorable path,

$$\frac{\partial}{\partial\lambda}\log \mathrm{LF}_\psi(P; \lambda)\Big|_{\lambda=0} = \phi_P. \tag{86}$$

Pathwise differentiability of ψ therefore gives

$$\frac{\partial}{\partial\lambda}\psi\big(\mathrm{LF}_\psi(P; \lambda)\big)\Big|_{\lambda=0} = P\left[\phi_P \,\frac{\partial}{\partial\lambda}\log \mathrm{LF}_\psi(P; \lambda)\Big|_{\lambda=0}\right] = P[\phi_P^2] = V_\psi(P). \tag{87}$$

Hence Eq. (70) is the first-order expansion of the target along the correction path. Combining this expansion with the von–Mises expansion around P_0 yields

$$\begin{aligned}\psi(\tilde{P}) - \psi_0 &= -P_0\phi_P + \lambda V_\psi(P) + R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda) \\ &= (P_n - P_0)\phi_P + \left\{\lambda V_\psi(P) - P_n\phi_P\right\} + R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda).\end{aligned} \tag{88}$$

Since an idealized prior-corrected PFN recovers the push-forward of the joint posterior $\widetilde{\Pi}(P, \lambda \mid \mathcal{D}_n)$ through $(P, \lambda) \mapsto \psi(\mathrm{LF}_\psi(P; \lambda))$, the L_2 triangle inequality implies

$$\left\| \tilde{\psi} - \psi_0 \right\|_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)} \leq \widetilde{F}_\psi(P_0, \mathcal{D}_n, \widetilde{\Pi}) + R_{2,\psi}(P_0, \mathcal{D}_n, \widetilde{\Pi}). \tag{89}$$

To bound the first-order contribution, write

$$(P_n - P_0)\phi_P = (P_n - P_0)\phi_0 + (P_n - P_0)(\phi_P - \phi_0). \tag{90}$$

Hence,

$$\widetilde{F}_\psi(P_0, \mathcal{D}_n, \widetilde{\Pi}) \leq |(P_n - P_0)\phi_0| + \left[\int \left\{ (P_n - P_0)(\phi_P - \phi_0) \right\}^2 \widetilde{\Pi}(\mathrm{d}P \ \mathrm{d}\lambda \mid \mathcal{D}_n) \right]^{1/2} + \left[\int \left\{ \lambda V_\psi(P) - P_n \phi_P \right\}^2 \widetilde{\Pi}(\mathrm{d}P \ \mathrm{d}\lambda \mid \mathcal{D}_n) \right]^{1/2}. \tag{91}$$

The first term is $O_{P_0}(n^{-1/2})$ by the central limit theorem, the second is $o_{P_0}(n^{-1/2})$ by Eq. (67), and the third is $O_{P_0}(n^{-1/2})$ by Eq. (71). Thus,

$$\widetilde{F}_\psi(P_0, \mathcal{D}_n, \widetilde{\Pi}) = O_{P_0}(n^{-1/2}). \tag{92}$$

This establishes Eq. (72) for both correction strategies. □

Theorem 3 (Bernstein–von Mises property for debiased PFNs)**.** *Assume the conditions of Theorem 2. Let $\phi_P := \phi_\psi(\cdot; P)$, $\phi_0 := \phi_\psi(\cdot; P_0)$, and*

$$V_0 := V_{\mathrm{eff}}(P_0) = P_0[\phi_0^2] = \mathrm{Var}_{P_0}[\phi_0] \in (0, \infty). \tag{93}$$

Suppose that the corresponding posterior-averaged second-order remainder is negligible at the parametric scale,

$$R_{2,\psi}(P_0, \mathcal{D}_n, \star) = o_{P_0}(n^{-1/2}), \qquad \star = \begin{cases} \Pi, & \textit{(i) label correction}, \\ \widetilde{\Pi}, & \textit{(ii) prior correction}. \end{cases} \tag{94}$$

Define the centered first-order fluctuation by

$$U_n^\star := \begin{cases} \sqrt{n}(\mathrm{BB}_n - P_n)\phi_P, & \textit{(i) label correction}, \\ \sqrt{n}\{\lambda V_\psi(P) - P_n \phi_P\}, & \textit{(ii) prior correction}, \end{cases} \tag{95}$$

where $V_\psi(P) := P[\phi_P^2]$. Assume the standard semi-parametric BvM regularity conditions such that, conditionally on $\mathcal{D}_n$,

$$d_{\mathrm{TV}}\left[\mathcal{L}_\star \left(U_n^\star \mid \mathcal{D}_n \right); N(0, V_0) \right] \xrightarrow{P_0} 0. \tag{96}$$

Then, for either debiasing strategy (i) or (ii),

$$d_{\mathrm{TV}}\left[\widetilde{Q}_{\theta^*} \left(\sqrt{n}(\psi - \psi_0) \mid \mathcal{D}_n \right); N\left(\Delta_n, V_{\mathrm{eff}}(P_0) \right) \right] \xrightarrow{P_0} 0, \tag{97}$$

where $\Delta_n := \sqrt{n} P_n[\phi_\psi(\cdot; P_0)]$. Moreover,

$$\Delta_n \rightsquigarrow N\left(0, V_{\mathrm{eff}}(P_0)\right), \tag{98}$$

so that the debiased PFN posterior asymptotically attains the semi-parametric efficiency bound.

Proof. Since ϕ_0 is an influence function, $P_0\phi_0 = 0$, and hence

$$\Delta_n = \sqrt{n}(P_n - P_0)\phi_0. \tag{99}$$

We show that, under either correction strategy,

$$\sqrt{n}(\tilde{\psi} - \psi_0) = \Delta_n + U_n^\star + A_n^\star, \tag{100}$$

where

$$A_n^\star := \sqrt{n}(P_n - P_0)(\phi_P - \phi_0) + \sqrt{n}\,\mathfrak{r}_\psi^\star,$$
$$\mathfrak{r}_\psi^\star := \begin{cases} R_{2,\psi}(P, P_0), & \text{(i) label correction,} \\ R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda), & \text{(ii) prior correction.} \end{cases} \tag{101}$$

Then, the assumptions of Theorem 2 together with Eq. (94) imply that $A_n^\star = o_{P_0}(1)$ in posterior L_2.

(i) Label correction. From Eq. (79),

$$\tilde{\psi}(P) - \psi_0 = (\mathrm{BB}_n - P_0)\phi_P + R_{2,\psi}(P, P_0). \tag{102}$$

Adding and subtracting $P_n\phi_P$ and decomposing ϕ_P around ϕ_0 gives

$$\begin{aligned}\sqrt{n}(\tilde{\psi}(P) - \psi_0) = \sqrt{n}(P_n - P_0)\phi_0 + \sqrt{n}(\mathrm{BB}_n - P_n)\phi_P \\ + \sqrt{n}(P_n - P_0)(\phi_P - \phi_0) + \sqrt{n}R_{2,\psi}(P, P_0).\end{aligned} \tag{103}$$

Thus, Eq. (100) holds with

$$A_n^\Pi := \sqrt{n}(P_n - P_0)(\phi_P - \phi_0) + \sqrt{n}R_{2,\psi}(P, P_0). \tag{104}$$

By the stochastic equicontinuity assumption in Eq. (67),

$$\left[\int n\big\{(P_n - P_0)(\phi_P - \phi_0)\big\}^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} = o_{P_0}(1). \tag{105}$$

Moreover, Eq. (94) gives

$$\left[\int nR_{2,\psi}^2(P, P_0)\Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} = o_{P_0}(1). \tag{106}$$

By the L_2 triangle inequality,

$$\left[\int (A_n^\Pi)^2 \Pi(\mathrm{d}P \mid \mathcal{D}_n)\right]^{1/2} = o_{P_0}(1). \tag{107}$$

(ii) Prior correction. From the decomposition established in the proof of Theorem 2,

$$\begin{aligned}\psi(\tilde{P}) - \psi_0 = (P_n - P_0)\phi_P + \big\{\lambda V_\psi(P) - P_n\phi_P\big\} \\ + R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda).\end{aligned} \tag{108}$$

Therefore,

$$\begin{aligned}\sqrt{n}(\psi(\tilde{P}) - \psi_0) = \sqrt{n}(P_n - P_0)\phi_0 + \sqrt{n}\big\{\lambda V_\psi(P) - P_n\phi_P\big\} \\ + \sqrt{n}(P_n - P_0)(\phi_P - \phi_0) \\ + \sqrt{n}\big\{R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda)\big\}.\end{aligned} \tag{109}$$

Hence, Eq. (100) holds with

$$\begin{aligned}A_n^{\widetilde{\Pi}} := \sqrt{n}(P_n - P_0)(\phi_P - \phi_0) \\ + \sqrt{n}\big\{R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda)\big\}.\end{aligned} \tag{110}$$

Again, Eq. (67) yields

$$\left[\int n\big\{(P_n - P_0)(\phi_P - \phi_0)\big\}^2 \widetilde{\Pi}(\mathrm{d}P \times \mathrm{d}\lambda \mid \mathcal{D}_n)\right]^{1/2} = o_{P_0}(1), \tag{111}$$

while Eq. (94) and the definition in Eq. (75) imply

$$\left[\int n\big\{R_{2,\psi}(P, P_0) + r_{\mathrm{LF},\psi}(P, \lambda)\big\}^2 \widetilde{\Pi}(\mathrm{d}P \times \mathrm{d}\lambda \mid \mathcal{D}_n)\right]^{1/2} = o_{P_0}(1). \tag{112}$$

Thus,

$$\left[\int (A_n^{\widetilde{\Pi}})^2 \widetilde{\Pi}(\mathrm{d}P \times \mathrm{d}\lambda \mid \mathcal{D}_n)\right]^{1/2} = o_{P_0}(1). \tag{113}$$

We have therefore shown for both correction strategies that

$$A_n^\star = o_{P_0}(1) \tag{114}$$

in posterior L_2. By the standard stability condition included in the semi-parametric BvM assumptions, this posterior-negligible perturbation does not affect the limiting total-variation law. Combining Eq. (100) with Eq. (96) and using that Δ_n is fixed conditionally on $\mathcal{D}_n$ therefore gives

$$d_{\mathrm{TV}}\left[\widetilde{Q}_{\theta^*}\left(\sqrt{n}(\psi - \psi_0) \mid \mathcal{D}_n\right); N(\Delta_n, V_0)\right] \xrightarrow{P_0} 0. \tag{115}$$

Finally, since $P_0\phi_0 = 0$ and $P_0[\phi_0^2] = V_0 < \infty$, the central limit theorem yields

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_0(Z_i) \rightsquigarrow N(0, V_0). \tag{116}$$

Since $V_0 = V_{\mathrm{eff}}(P_0)$, the limiting posterior variance equals the semi-parametric efficiency bound. □

Theorem 4 (Rate optimality of debiased PFNs)**.** *Assume the conditions of Theorem 2, and consider an effective prior-supported class $\mathcal{P}_n^\Pi := \mathrm{supp}(\Pi) \cap \mathcal{P}_n^{\mathrm{SA}}$ with $\Pi(\mathcal{P}_n^\Pi) \to 1$. Suppose that the first- and second-order contributions in Theorem 2 satisfy uniformly over $\mathcal{P}_n^\Pi$,*

$$\sup_{P \in \mathcal{P}_n^\Pi} \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}\left[\widetilde{F}_\psi^2(P, \mathcal{D}_n, \star)\right] \lesssim n^{-1}, \quad \sup_{P \in \mathcal{P}_n^\Pi} \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}\left[R_{2,\psi}^2(P, \mathcal{D}_n, \star)\right] \lesssim r_{2,n}^2, \tag{117}$$

where $\star = \Pi$ for (i) label correction, and $\star = \widetilde{\Pi}$ for (ii) prior correction. If Π is not fully supported on $\mathcal{P}_n^\Pi$, additionally assume that the contribution outside the effective class is negligible at the same rate:

$$\mathbb{E}_{P \sim \Pi}\left[\mathbb{1}\{P \notin \mathcal{P}_n^\Pi\}\, \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}\left\|\tilde{\psi} - \psi(P)\right\|^2_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)}\right] \lesssim n^{-1} + r_{2,n}^2. \tag{118}$$

Then, for either debiasing strategy,

$$\underbrace{\mathcal{R}_{\Pi,n}^{l_2}(\hat{\psi}_{\theta^*})}_{\text{Bayes optimal risk, see Remark 1}} \leq \mathcal{R}_{\Pi,n}^{l_2}(\hat{\tilde{\psi}}_{\theta^*}) \leq \mathcal{R}_{\Pi,n}^{L_2}(\tilde{\psi}) \lesssim n^{-1} + r_{2,n}^2 \quad (\Pi\text{-prior-averaged risks}), \tag{119}$$

$$\underbrace{\mathfrak{M}_n(\mathcal{P}_n^\Pi)}_{\text{minimax lower bound}} \leq \mathcal{W}_{\mathcal{P}_n^\Pi,n}^{l_2}(\hat{\tilde{\psi}}_{\theta^*}) \leq \mathcal{W}_{\mathcal{P}_n^\Pi,n}^{L_2}(\tilde{\psi}) \lesssim n^{-1} + r_{2,n}^2 \quad (\text{worst-case risks}), \tag{120}$$

where the inequalities in red hold whenever $\Pi(\mathcal{P}_n^\Pi) = 1$; $\mathcal{R}_{\Pi,n}^{l_2}(\hat{\psi}) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\hat{\psi}(\mathcal{D}_n) - \psi(P)]^2$, $\mathcal{R}_{\Pi,n}^{L_2}(\tilde{\psi}) := \mathbb{E}_{P\sim\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}||\tilde{\psi} - \psi||^2_{L_2(\tilde{q}_{\theta^*},\mathcal{D}_n)}$ are Π-prior-averaged squared l_2 risks; $\mathcal{W}_{\mathcal{P}_n^\Pi,n}^{l_2}(\hat{\psi}) := \sup_{P\in\mathcal{P}_n^\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}[\hat{\psi}(\mathcal{D}_n) - \psi(P)]^2$ and $\mathcal{W}_{\mathcal{P}_n^\Pi,n}^{L_2}(\tilde{\psi}) := \sup_{P\in\mathcal{P}_n^\Pi}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}||\tilde{\psi} - \psi||^2_{L_2(\tilde{q}_{\theta^*},\mathcal{D}_n)}$ are worst-case squared l_2 risks; and $\mathfrak{M}_n(\mathcal{P}_n^\Pi) := \inf_{\hat{\psi}:\mathcal{P}^{\otimes n}\to\Psi}\mathcal{W}_{\mathcal{P}_n^\Pi,n}^{l_2}(\hat{\psi})$ denotes the minimax lower bound over $\mathcal{P}_n^\Pi$.

If, additionally, a matching minimax lower bound is available as

$$\mathfrak{M}_n(\mathcal{P}_n^\Pi) \gtrsim n^{-1} + r_{2,n}^2, \tag{121}$$

then the debiased PFN posterior mean is minimax-rate optimal:

$$\mathcal{W}_{\mathcal{P}_n^\Pi,n}^{l_2}(\hat{\tilde{\psi}}_{\theta^*}) \asymp \mathfrak{M}_n(\mathcal{P}_n^\Pi) \asymp n^{-1} + r_{2,n}^2. \tag{122}$$

Proof. We first relate the squared error of the debiased PFN posterior mean to its posterior L_2-concentration error. For every fixed P and $\mathcal{D}_n$, Jensen's inequality gives

$$\Big[\hat{\tilde{\psi}}_{\theta^*}(\mathcal{D}_n) - \psi(P)\Big]^2 = \left[\int \big(\tilde{\psi} - \psi(P)\big)\tilde{q}_{\theta^*}(\tilde{\psi} \mid \mathcal{D}_n)\,\mathrm{d}\tilde{\psi}\right]^2 \tag{123}$$

$$\leq \int \big(\tilde{\psi} - \psi(P)\big)^2 \tilde{q}_{\theta^*}(\tilde{\psi} \mid \mathcal{D}_n)\,\mathrm{d}\tilde{\psi} = \left\|\tilde{\psi} - \psi(P)\right\|^2_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)}. \tag{124}$$

Taking expectations and either averaging over $P \sim \Pi$ or taking the supremum over $P \in \mathcal{P}_n^{\Pi}$ therefore yields

$$\mathcal{R}^{l_2}_{\Pi,n}(\hat{\tilde{\psi}}_{\theta^*}) \leq \mathcal{R}^{L_2}_{\Pi,n}(\tilde{\psi}), \quad \mathcal{W}^{l_2}_{\mathcal{P}_n^{\Pi},n}(\hat{\tilde{\psi}}\theta^*) \leq \mathcal{W}^{L_2}_{\mathcal{P}_n^{\Pi},n}(\tilde{\psi}). \tag{125}$$

Next, Theorem 2 gives, for every $P \in \mathcal{P}_n^{\Pi}$,

$$\left\|\tilde{\psi} - \psi(P)\right\|_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)} \leq \widetilde{F}_{\psi}(P, \mathcal{D}_n, \star) + R_{2,\psi}(P, \mathcal{D}_n, \star). \tag{126}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\left\|\tilde{\psi} - \psi(P)\right\|^2_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)} \leq 2\widetilde{F}^2_{\psi}(P, \mathcal{D}_n, \star) + 2R^2_{2,\psi}(P, \mathcal{D}_n, \star). \tag{127}$$

Taking expectation over $\mathcal{D}_n \sim P^{\otimes n}$ and then the supremum over $P \in \mathcal{P}_n^{\Pi}$, Eq. (117) implies

$$\mathcal{W}^{L_2}_{\mathcal{P}_n^{\Pi},n}(\tilde{\psi}) \lesssim n^{-1} + r^2_{2,n}. \tag{128}$$

Together with Eq. (123), this proves the two upper bounds in Eq. (21).

For the prior-averaged risk, split the expectation according to whether P lies in the effective class:

$$\mathcal{R}^{L_2}_{\Pi,n}(\tilde{\psi}) = \mathbb{E}_{P\sim\Pi}\left[\mathbb{1}\{P \in \mathcal{P}_n^{\Pi}\}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\left\|\tilde{\psi} - \psi(P)\right\|^2_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)}\right]$$

$$+ \mathbb{E}_{P\sim\Pi}\left[\mathbb{1}\{P \notin \mathcal{P}_n^{\Pi}\}\mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\left\|\tilde{\psi} - \psi(P)\right\|^2_{L_2(\tilde{q}_{\theta^*}, \mathcal{D}_n)}\right]. \tag{129}$$

The first term is bounded by $\mathcal{W}^{L_2}_{\mathcal{P}_n^{\Pi},n}(\tilde{\psi})$ and hence by Eq. (128); the second is controlled by Eq. (118). Therefore,

$$\mathcal{R}^{L_2}_{\Pi,n}(\tilde{\psi}) \lesssim n^{-1} + r^2_{2,n}. \tag{130}$$

Together with Eq. (123), this establishes the upper bounds in Eq. (20).

By Remark 1, $\hat{\psi}_{\theta^*}$ is the Bayes estimator under the original prior Π and squared l_2 loss. Since $\hat{\tilde{\psi}}_{\theta^*}$ is another measurable estimator based on $\mathcal{D}_n$,

$$\mathcal{R}^{l_2}_{\Pi,n}(\hat{\psi}_{\theta^*}) \leq \mathcal{R}^{l_2}_{\Pi,n}(\hat{\tilde{\psi}}_{\theta^*}), \tag{131}$$

which proves the first inequality in Eq. (119).

Likewise, by definition of the minimax risk,

$$\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) = \inf_{\hat{\psi}} \mathcal{W}^{l_2}_{\mathcal{P}_n^{\Pi},n}(\hat{\psi}) \leq \mathcal{W}^{l_2}_{\mathcal{P}_n^{\Pi},n}(\hat{\tilde{\psi}}_{\theta^*}), \tag{132}$$

establishing the first inequality in Eq. (120).

If $\Pi(\mathcal{P}_n^{\Pi}) = 1$, then for any non-negative risk function $g(P)$,

$$\mathbb{E}_{P\sim\Pi}[g(P)] \leq \sup_{P\in\mathcal{P}_n^{\Pi}} g(P). \tag{133}$$

Applying this identity to the point-estimation and posterior L_2 risks gives the two red inequalities.

Finally, if the matching minimax lower bound in Eq. (121) holds, then

$$n^{-1} + r^2_{2,n} \lesssim \mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \leq \mathcal{W}^{l_2}_{\mathcal{P}_n^{\Pi},n}(\hat{\tilde{\psi}}_{\theta^*}) \lesssim n^{-1} + r^2_{2,n}, \tag{134}$$

which proves Eq. (122).

□

E MATCHING MINIMAX LOWER BOUNDS FOR EXAMPLES A–C

We briefly verify when the matching lower-bound condition in Theorem 4 holds for our running examples. Structure-agnostic lower bounds are typically formulated over local classes defined through the accuracy of black-box nuisance estimators (Balakrishnan et al., 2026; Jin & Syrgkanis, 2025b;a). Since our effective class additionally intersects such a class with the support of the PFN prior, these results apply provided that $\mathcal{P}_n^{\Pi}$ contains the corresponding hard structure-agnostic sub-model. More precisely, if $\mathcal{H}_n^{\psi}$ denotes a hard sub-model from one of the lower-bound constructions below and

$$\mathcal{H}_n^{\psi} \subseteq \mathcal{P}_n^{\Pi}, \tag{135}$$

then, by monotonicity of the minimax risk,

$$\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \geq \mathfrak{M}_n(\mathcal{H}_n^{\psi}). \tag{136}$$

Thus, the results below should be understood subject to this mild prior-support requirement.

Example A (Mean value)**.** *For $\psi(P) = \mathbb{E}[Z]$, there is no non-parametric nuisance estimation and $R_{2,\psi} \equiv 0$, such that $r_{2,n} = 0$. The structure-agnostic restriction is therefore vacuous. If $\mathcal{P}_n^{\Pi}$ contains a regular one-dimensional sub-model with non-zero variance (for example, the Bernoulli family with success probabilities in a neighborhood of $1/2$) the classical two-point or local asymptotic minimax lower bound gives (van der Vaart, 2000):*

$$\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \gtrsim n^{-1} = n^{-1} + r_{2,n}^2. \tag{137}$$

Hence, the n^{-1} squared-risk rate obtained by the debiased PFN is minimax-rate optimal.

Example B (Mean value under MAR)**.** *Let $e_{\mu,n}$ and $e_{\pi,n}$ denote squared $L_2(P_X)$ radii for black-box estimators of the outcome regression and labeling score, respectively:*

$$\|\mu - \hat{\mu}_n\|_{L_2(P_X)}^2 \leq e_{\mu,n}, \qquad \|\pi - \hat{\pi}_n\|_{L_2(P_X)}^2 \leq e_{\pi,n}. \tag{138}$$

Under strong overlap, the corresponding second-order remainder satisfies

$$|R_{2,\psi}(P, P_0)| = \left| P_0 \left[\frac{\pi - \pi_0}{\pi} (\mu - \mu_0) \right] \right| \lesssim \|\pi - \pi_0\|_{L_2(P_0)} \|\mu - \mu_0\|_{L_2(P_0)}. \tag{139}$$

Consequently, over a structure-agnostic class with the above nuisance radii, the squared second-order contribution is of order

$$r_{2,n}^2 \asymp e_{\mu,n} e_{\pi,n}. \tag{140}$$

The MAR mean is a linear functional of μ with an affine regression score and Riesz representer proportional to $1/\pi$. Under strong overlap, the L_2 error of this Riesz representer is equivalent up to constants to the L_2 error of π. Hence, the affine-score structure-agnostic lower bound of Jin & Syrgkanis (2025a) yields

$$\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \gtrsim n^{-1} + e_{\mu,n} e_{\pi,n} \asymp n^{-1} + r_{2,n}^2. \tag{141}$$

This is the same characteristic "parametric variance plus product of nuisance errors" lower bound established for canonical mixed-bias functionals by Balakrishnan et al. (2026). Thus, whenever the corresponding hard local class is contained in $\mathcal{P}_n^{\Pi}$, the rate of our debiased PFN is structure-agnostic minimax optimal.

Example C (WATE)**.** *We first consider a fixed weight $w = w(x)$ and let*

$$\|\mu_a - \hat{\mu}_{a,n}\|_{L_2(P_X)}^2 \leq e_{a,n},\ a \in \{0,1\}, \qquad \|\pi - \hat{\pi}_n\|_{L_2(P_X)}^2 \leq f_n. \tag{142}$$

For the unnormalized WATE, Jin & Syrgkanis (2025b) establish a structure-agnostic lower bound of the form

$$\mathfrak{M}_n^{\mathrm{WATE}} \gtrsim \max\{e_{0,n}, e_{1,n}\} f_n \|w\|_{L_\infty(P)}^2 + \frac{\|w\|_{L_2(P)}^2}{n}. \tag{143}$$

Their result is stated as a quantile lower bound and therefore also implies the corresponding expected squared-risk lower bound. If w is uniformly bounded and non-degenerate and $\bar{w}(P) = P[w(X)]$ is bounded away from zero, the same rate applies to our normalized WATE.

In particular, one may restrict the lower-bound construction to a sub-model with fixed P_X, for which $\bar{w}(P)$ is constant. Absorbing weight-dependent constants gives

$$\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \gtrsim n^{-1} + \max\{e_{0,n}, e_{1,n}\} f_n. \tag{144}$$

At the same time, the second-order remainder of the first-order corrected WATE contains products between propensity-score and outcome-regression errors, such that, up to constants,

$$r_{2,n}^2 \asymp f_n \left(e_{0,n} + e_{1,n}\right) \asymp f_n \max\{e_{0,n}, e_{1,n}\}. \tag{145}$$

Thus,

$$\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \gtrsim n^{-1} + r_{2,n}^2. \tag{146}$$

This result directly covers the ATE, $w \equiv 1$, as a special case. Jin & Syrgkanis (2025b) additionally establish the analogous structure-agnostic lower bound for the ATT, which corresponds to a normalized propensity-dependent weighting of the treatment effect.

For the overlap-weighted ATE, $w(x;\pi) = \pi(x)(1-\pi(x))$, a complementary lower bound follows from Balakrishnan et al. (2026). Indeed, its numerator can be written exactly as the expected conditional covariance,

$$P\left[\pi(X)\big(1-\pi(X)\big)\tau(X)\right] = P\left[\mathrm{Cov}(A, Y \mid X)\right]. \tag{147}$$

For squared L_2 nuisance radii r_n and s_n of the corresponding outcome regression and propensity score, respectively, Balakrishnan et al. (2026) show that the structure-agnostic minimax risk for this functional satisfies

$$\mathfrak{M}_n^{\mathrm{cov}} \gtrsim r_n s_n + n^{-1}. \tag{148}$$

Restricting to a hard sub-model on which $P[\pi(X)(1-\pi(X))]$ is fixed, or varies only at a smaller order, transfers the same lower bound to the normalized overlap-weighted ATE. Since its second-order remainder contains the same product-order nuisance contribution, $r_{2,n}^2 \asymp r_n s_n$ on this sub-model, yielding again

$$\mathfrak{M}_n(\mathcal{P}_n^{\Pi}) \gtrsim n^{-1} + r_{2,n}^2. \tag{149}$$

Taken together, these lower bounds verify the matching-rate condition from Theorem 4 for Example A, Example B, and the canonical fixed-weight, ATE, and overlap-weighted instances of Example C, whenever the effective prior-supported class contains the corresponding structure-agnostic hard sub-models.

F Additional Results and Details

This appendix provides the details and additional experiments for testing predictions derived from theory. Appendices F.1–F.3 describe the implementation, the prior, and the evaluation metrics. Appendices F.4 and F.5 test the uncertainty part of Theorem 3, namely that the Bayesian bootstrap supplies the efficient variance. Appendix F.6 tests the prior-averaged prediction. Appendix F.7 justifies an implementation choice. Appendix F.8 evaluates the second debiasing strategy, prior correction. Appendix F.9 separates the statistical effect from amortization errors. Appendix F.10 discusses the limitations of our experiments.

Code is available at `https://anonymous.4open.science/r/Ortho-PFN-Anonymized-E4B7`.

F.1 Implementation Details

Architecture. Both PFNs share the same architecture: a transformer that embeds every scalar cell of $\mathcal{D}_n$ as a token and alternates attention across features and across rows, similar to TabPFN v2 (Hollmann et al., 2025). A learned query row, which the observed rows do not attend to, is pooled across features and mapped to a 5-component Gaussian mixture density $q_\theta(\psi \mid \mathcal{D}_n)$. We use 3 layers, embedding dimension 64, 4 attention heads, and feed-forward width 128. Inputs and labels are normalized per dataset.

Training. We train a separate network for every n and method with AdamW (learning rate $3 \cdot 10^{-4}$, weight decay 10^{-5}), batch size 31 datasets, gradient-norm clipping at 1, and mixed precision for 126,900 updates. The learning rate is halved when the validation loss has not improved for 5 validation cycles (every 2,115 updates), and we keep the checkpoint with the lowest validation loss. Both methods use the same initialization, data order, and seed.

Oracle one-step estimator. With the ground-truth nuisance functions of P_0, let $m_i = \psi(P_0) + \phi_\psi(Z_i; P_0)$ be the pseudo-outcomes. The oracle posterior is the Bayesian-bootstrap distribution of $\mathrm{BB}_n\{m\}$, with mean $\bar{m} = \mathbb{P}_n m$, variance $\sum_i (m_i - \bar{m})^2/[n(n+1)]$, and interval endpoints from 1,000 Bayesian-bootstrap draws. It is thus the target of the label-corrected risk in Eq. (12) with P known, and it serves as a privileged reference rather than as a baseline.

Compute. All networks were trained on NVIDIA H200 GPUs. One fit takes approximately 4–5.5 hours for $n \leq 1024$ and approximately 10 hours for $n = 2048$, with up to two fits sharing one GPU.

F.2 Experimental Setup

Prior. Each prior draw $P \sim \Pi$ is a structural causal model over $Z = (X, A, Y)$. Covariates $X \in \mathbb{R}^5$ are selected nodes of a random layered nonlinear directed acyclic graph, whose node noises mix normal, uniform, Laplace, and logistic families. The propensity score $\pi(x)$ is a random $\tanh$ multilayer perceptron with a sigmoid output, and the outcome regressions $\mu_a(x)$ are given by a separate random $\tanh$ multilayer perceptron on (X, A). Then, $A \mid X \sim \mathrm{Bernoulli}(\pi(X))$ and $Y = \mu_A(X) + \varepsilon$ with $\varepsilon \sim N(0, 1)$ independent of (X, A). Hence, the prior contains nonlinear observed confounding and heterogeneous treatment effects, while both outcome regressions remain exactly inferable (e. g., by simply setting $\varepsilon = 0$ to obtain oracle outcome regression instead of a Y sample). Architectures and parameter laws follow Javurek et al. (2026), with the additive-noise outcome model above. We do not apply propensity clipping, overlap floor, or trimming in the prior.

Ground truth. For every P, the ATE $\psi(P) = \mathbb{E}[\mu_1(X) - \mu_0(X)]$ and the efficient variance $V_{\text{eff}}(P) = \mathrm{Var}(\mu_1(X) - \mu_0(X)) + \mathbb{E}[1/\pi(X) + 1/(1-\pi(X))]$ (with the known unit noise variance) are computed on an independent reference sample of covariates of size $L = 2^{18}$ for the training labels and $L = 2^{23}$ for the evaluation populations.

Training data. We draw $M = 69{,}632$ populations from Π with one dataset of 2,048 rows each, and use 65,536 populations for training and 4,096 for validation. For $n < 2048$, the datasets are the first n rows of the stored datasets, so the training data are nested across n. The debiased PFN is trained with $K = 4$ independent Bayesian-bootstrap draws $(W_1, \ldots, W_n) \sim \mathrm{Dir}(n; 1, \ldots, 1)$ per dataset, i. e., four labels $\tilde{\psi}(P)$ in Eq. (12) per dataset, whose losses are averaged. The ATE pseudo-outcomes entering the label are $\psi(P) + \phi_\psi(Z; P) = \mu_1(X) - \mu_0(X) + \frac{A}{\pi(X)}(Y - \mu_1(X)) - \frac{1-A}{1-\pi(X)}(Y - \mu_0(X))$ (see Appendix C). The standard PFN is trained on $\psi(P)$ for the same datasets.

Evaluation populations. The fixed-population evaluation (main text) uses $M = 200$ populations $P_0 \sim \Pi$ with $\varepsilon_{P_0} = \min\{\min_x \pi(x), 1 - \max_x \pi(x)\} \geq 0.01$ on a screening sample of 2^{20} covariate draws, i. e., the strong-overlap regime of our theory. For every population, we draw $D = 200$ datasets $\mathcal{D}_n^{(r)} \sim P_0^{\otimes n}$, nested across n. The prior-averaged evaluation uses $M = 3{,}000$ fresh prior draws with one dataset each. All evaluation populations are disjoint from training and validation.

F.3 Evaluation Metrics

The metrics translate Theorem 3 into checkable quantities at a fixed population P_0. If the posterior is asymptotically centered at an efficient estimator with efficient variance, then its conditional bias vanishes faster than $n^{-1/2}$, its variance relative to $V_{\text{eff}}(P_0)/n$ tends to one, and its credible intervals attain nominal coverage. Theorem 1 predicts that the standard PFN can fail on the first of these.

For a fixed population P_0 with $\psi_0 = \psi(P_0)$, let $s = (V_{\text{eff}}(P_0)/n)^{1/2}$ be the efficient standard deviation, and let $\hat{\psi}_r$, v_r, and $[l_r, u_r]$ be the posterior mean, the posterior variance, and the equal-tailed 95% credible interval of a PFN given $\mathcal{D}_n^{(r)}$. With $z_r = (\hat{\psi}_r - \psi_0)/s$, we compute per population

$$\begin{aligned}
\beta_{P_0} &= \tfrac{1}{R}\textstyle\sum_r z_r, \\
\tilde{\beta}^2_{P_0} &= \beta^2_{P_0} - \tfrac{1}{R}\widehat{\mathrm{Var}}_r(z_r), \\
\nu_{P_0} &= \tfrac{1}{R}\textstyle\sum_r v_r/s^2, \\
c_{P_0} &= \tfrac{1}{R}\textstyle\sum_r \mathbb{1}\{\psi_0 \in [l_r, u_r]\}, \\
\rho_{P_0} &= \tfrac{1}{R}\textstyle\sum_r \left(z_r^2 + v_r/s^2\right),
\end{aligned}$$

i. e., the standardized conditional bias, its Monte-Carlo-corrected square, the variance ratio (ideal: 1), the conditional coverage (ideal: 0.95), and the standardized posterior L_2 error of Theorem 2 (ideal: ≈ 2). Over populations, we report the root-mean-square conditional bias in ATE units, $\left(\mathrm{mean}_{P_0}\, s^2\tilde{\beta}^2_{P_0}\right)^{1/2}$; its standardized version $B = \left(\mathrm{mean}_{P_0}\, \tilde{\beta}^2_{P_0}\right)^{1/2}$; the median of ν_{P_0}; the lower decile $q_{0.10}(c_{P_0})$; the median of c_{P_0}; and the upper decile $q_{0.90}(\rho_{P_0})$. With $D = 200$, even a perfectly calibrated method has $q_{0.10}(c_{P_0}) \approx 0.93$ (tenth percentile of $\mathrm{Bin}(200, 0.95)/200$). Unlike prior averages, these quantities are not pinned by the Bayes-optimality of an idealized standard PFN (Remark 1): its bias vanishes on average over Π, but not at a given P_0.

For the prior-averaged evaluation, we report the normalized RMSE $\left(\mathrm{mean}_P\, n(\hat{\psi} - \psi(P))^2/V_{\text{eff}}(P)\right)^{1/2}$ (efficient: 1) and the negative log posterior density at the truth, $-\log q_\theta(\psi(P) \mid \mathcal{D}_n)$, averaged over prior draws.

Rates are least-squares fits of $\log(\cdot)$ (or of the negative log density) on $\log n$ over the four sample sizes. All intervals are 95% percentile intervals from 2,000 bootstrap resamples of populations (or prior draws), shared across methods and n, so that differences are paired. They are conditional on the trained networks, since every combination of method and n is a single training run (see Appendix F.10).

F.4 Ablation: Label Correction without the Bayesian Bootstrap

In Theorem 3, the first-order correction removes the bias, while the randomness of the Bayesian-bootstrap label provides the first-order sampling variance $\mathrm{Var}_{P_0}[\phi_\psi]/n$. Removing the Bayesian bootstrap should therefore leave the bias reduction intact but yield posteriors that are too narrow.

We replace the Bayesian-bootstrap label in Eq. (12) by its expectation over the weights, $\psi(P) + \mathbb{P}_n\phi_\psi(Z; P)$. By the law of total variance over W, an idealized PFN trained on this label has the same posterior mean as the idealized debiased PFN, and its posterior variance lacks the first-order sampling variance $\mathbb{E}\left[\sum_i (m_i - \bar{m})^2/[n(n+1)] \mid \mathcal{D}_n\right]$. Fig. 4 is consistent with both predictions: the conditional bias is of the same magnitude, but the posterior variance falls below efficient (median $\nu = 0.55$ at $n = 2048$), and the median coverage drops to 0.72, compared with 0.95 for the debiased PFN. Hence, the first-order correction reduces the bias, and the Bayesian bootstrap supplies the calibrated uncertainty.

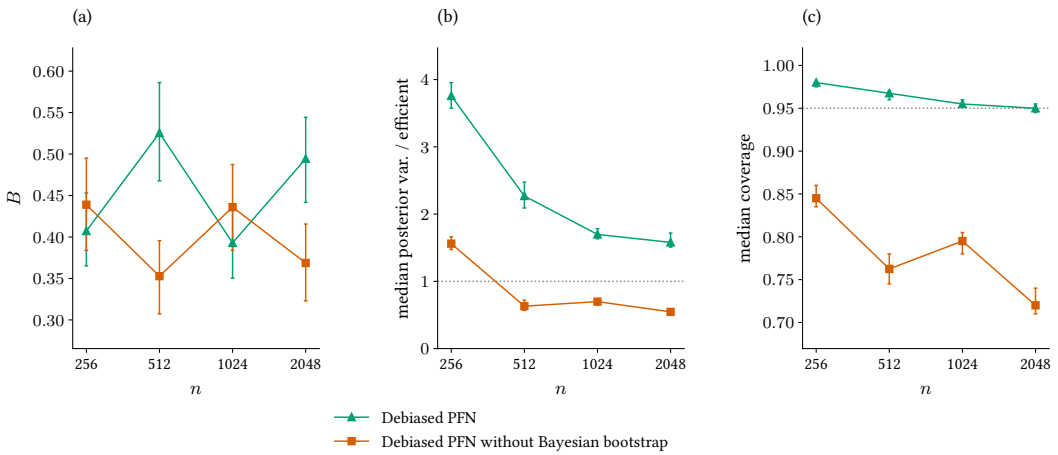

Figure 4: **Ablation: debiased PFN with and without the Bayesian bootstrap.** Same populations as Fig. 3. **(a)** Standardized conditional bias B. **(b)** Median posterior variance relative to efficient. **(c)** Median conditional coverage of 95% credible intervals. Error bars: 95% population-bootstrap intervals.

F.5 Variance Decomposition

This section makes the mechanism behind Appendix F.4 quantitative: the Bayesian bootstrap should add exactly the first-order sampling variance to the posterior, which is what yields the efficient variance in Theorem 3.

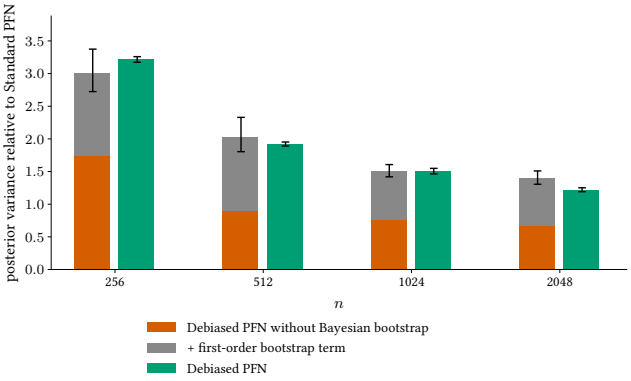

Figure 5: **The Bayesian bootstrap adds the first-order sampling variance.** For each n, posterior variances averaged over 3,000 prior draws (one dataset each), each relative to the standard PFN's posterior variance on the same dataset. Left bars: the debiased PFN without the Bayesian bootstrap, plus the first-order bootstrap term $\sum_i (m_i - \bar{m})^2/[n(n+1)]$ at the true nuisance functions. Right bars: the debiased PFN. Error bars: 95% bootstrap over prior draws.

By the law of total variance, the posterior variance of the debiased PFN equals that of the debiased PFN without the Bayesian bootstrap plus the posterior expectation of the first-order term $\sum_i (m_i - \bar{m})^2/[n(n+1)]$ (Appendix F.4). This posterior expectation is not observable, but averaged over the prior it equals the prior average of the first-order term at the ground-truth nuisance functions (tower property), also after normalizing by any data-dependent quantity; we use the standard PFN's posterior variance. Fig. 5 confirms the decomposition for $n \leq 1024$ (ratios of the variance increment to the first-order term: 1.17, 0.91, and 1.00).

F.6 Prior-Averaged Performance

Averaged over the prior, the idealized standard PFN is Bayes-optimal (Remark 1), so debiasing cannot improve prior-averaged proper scores in the idealized limit. Theorem 4 guarantees only that the prior-averaged risk of the debiased PFN retains the rate $n^{-1} + r_{2,n}^2$. This section checks that the local gains at fixed P_0 do not come at a cost when averaged over the prior.

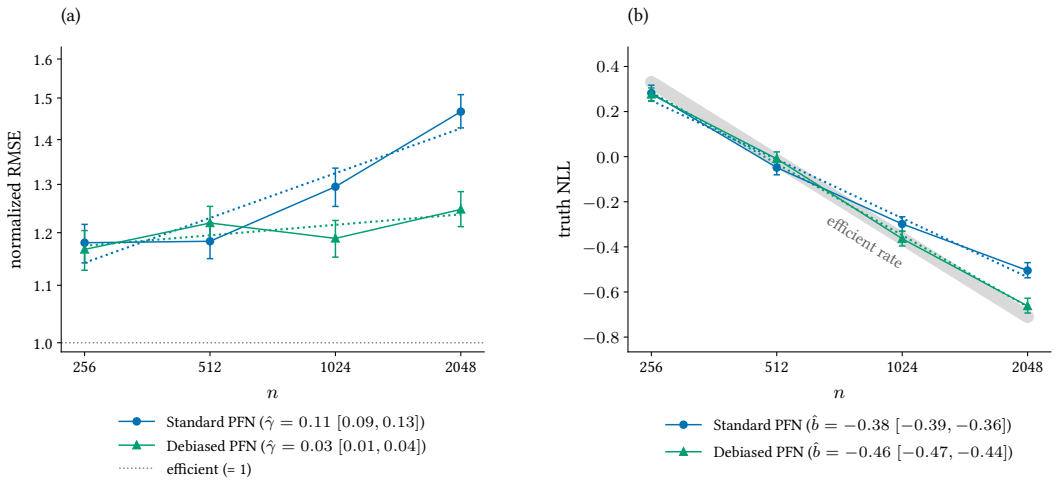

Figure 6: **Performance averaged over the prior**, mean over $M = 3{,}000$ prior draws, each with $D = 1$ dataset. **(a)** Normalized RMSE of the posterior mean (log scale; 1 is efficient). **(b)** Negative log posterior density at the true ATE; the grey band has the efficient slope $-\frac{1}{2}$ in $\ln n$, with its level fitted to the debiased PFN. Solid: estimates; dotted: least-squares fits, with 95% bootstrap intervals for the exponent (a) or slope (b) in the legend. Error bars: 95% bootstrap over prior draws, paired across methods and n.

Fig. 6 reports performance averaged over the prior Π rather than at fixed populations. The normalized RMSE of the debiased PFN is nearly constant in n (fitted exponent 0.03, $[0.01, 0.04]$), whereas that of the standard PFN increases (exponent 0.11, $[0.09, 0.13]$). Likewise, the negative log posterior density at the truth of the debiased PFN decreases at a slope closer to the efficient $-\frac{1}{2}$ in $\ln n$ (-0.46, $[-0.47, -0.44]$) than that of the standard PFN (-0.38, $[-0.39, -0.36]$). By Bayes-optimality (Remark 1), an *idealized* standard PFN cannot be outperformed on a prior-averaged proper score; hence, differences in favor of the debiased PFN reflect amortization errors of the trained standard PFN (see also Appendix F.9).

F.7 Number of Bayesian-Bootstrap Labels

The number K of Bayesian-bootstrap labels per training dataset is an implementation choice that our theory does not constrain: training with sampled labels targets the Bayesian-bootstrap distribution for any K, and K only controls the Monte Carlo noise of the training signal.

Table 1: **Bayesian-bootstrap labels per training dataset** ($K = 1$ vs. $K = 4$) for the debiased PFN at $n = 2048$ (overlap-restricted panel; last two columns: prior-averaged). Best value per column in bold (closest to target for coverage and ν). Parentheses: bootstrap standard errors.

Method	Labels/context	B	Median ν	$q_{.10}(c_p)$	$q_{.90}(\rho_p)$	Truth NLL	Mean z^2
Debiased PFN	1	0.715 (0.043)	2.359 (0.078)	**0.920 (0.009)**	6.622 (0.335)	-0.578	1.851
Debiased PFN	4	**0.494 (0.025)**	**1.579 (0.052)**	0.905 (0.008)	**4.659 (0.207)**	**-0.662**	**1.555**

Using $K = 4$ instead of a single Bayesian-bootstrap draw per training dataset reduces the Monte Carlo noise of the training label at no additional cost per update. Tab. 1 shows that this lowers the conditional bias ($B = 0.49$ vs. 0.71) and the posterior variance (1.58 vs. 2.36 times efficient) at $n = 2048$, with comparable coverage. We therefore use $K = 4$ throughout.

F.8 Prior Correction

Theorems 2 and 3 cover both debiasing strategies, but the main text evaluates only the label correction (i). Here, we evaluate the prior correction (ii) of Eq. (15) under the same protocol, implemented for the ATE following Example B.

Each prior draw P is perturbed jointly by an exponential tilt of the covariate distribution, $\tilde{P}(x) \propto \exp\{\lambda(\tau(x) - \psi(P))\}\, P(x)$ with $\tau = \mu_1 - \mu_0$, and by a location shift of both outcome regressions, $\tilde{\mu}_1(x) = \mu_1(x) + \lambda/\pi(x)$ and $\tilde{\mu}_0(x) = \mu_0(x) - \lambda/(1 - \pi(x))$. Here, the outcome noise has unit variance in our prior. For numerical stability, propensities are clipped at 0.01 in the shift, and the tilt weights are capped; the cap affects 23 of the 69,632 training populations. We set $\sigma_n^2 = c^2/(\sqrt{n}\, V_{\text{eff}}(P))$, so that $n\sigma_n^2 \to \infty$ as required. Since $\partial_\lambda \psi(\tilde{P})|_{\lambda=0} = V_{\text{eff}}(P)$, the induced perturbation of the target has prior standard deviation $c\, n^{1/4} \sqrt{V_{\text{eff}}(P)/n}$, i. e., $c\, n^{1/4}$ efficient posterior standard deviations. The PFN is trained with the standard label $\psi(\tilde{P})$ under the same protocol as the other variants; we report $c = 1$, as a smaller correction ($c = 0.25$) reduced bias less. Prior correction matches the debiased PFN's conditional bias only at the largest context size, $n = 2048$, where its bias also decays at close to the efficient rate. For smaller n, its bias is larger than that of the debiased PFN, and its prior-averaged NLL is worse than that of the standard PFN (Fig. 7). The debiased PFN thus matches or improves on prior correction at every n, without changing the prior.

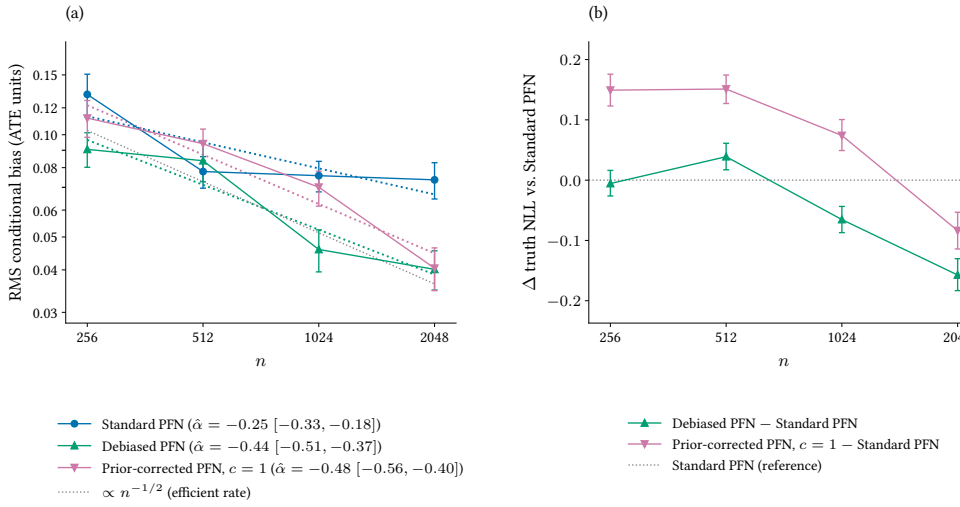

Figure 7: Prior correction ($c = 1$) vs. standard and debiased PFN. (a) RMS conditional bias with fitted exponents $\hat{\alpha}$. (b) Prior-averaged truth NLL relative to the standard PFN. 95% bootstrap intervals.

F.9 Amortization Budget

Our theory concerns the idealized PFN q_{θ^*}; a trained network additionally incurs an amortization error. The slower rate of the standard PFN in Sec. 5 may therefore partly reflect amortization error rather than the statistical mechanism of Theorem 1. To separate the two, we retrain both PFNs with four independent datasets per training population ($4\times$ training data) and the same number of updates. With $4\times$ data, the conditional bias of the standard PFN decreases to roughly the level of the debiased PFN trained on $1\times$ data (RMS bias 0.108 vs. 0.091 at $n = 256$ and 0.045 vs. 0.046 at $n = 1024$), and its fitted exponent becomes -0.49 ($[-0.56, -0.41]$). Hence, in this range of n, a part of the slower rate of the standard PFN is likely attributable to amortization error. At $4\times$, the debiased PFN retains a lower conditional bias for $n \leq 1024$; we do not interpret the $4\times$ exponents, since they hinge on the $n = 2048$ fits.

F.10 Limitations of the Experiments

Our experiments are designed to test the predictions of our theory in a controlled setting where the ground truth is known, rather than to benchmark against existing estimators. Hence, we compare

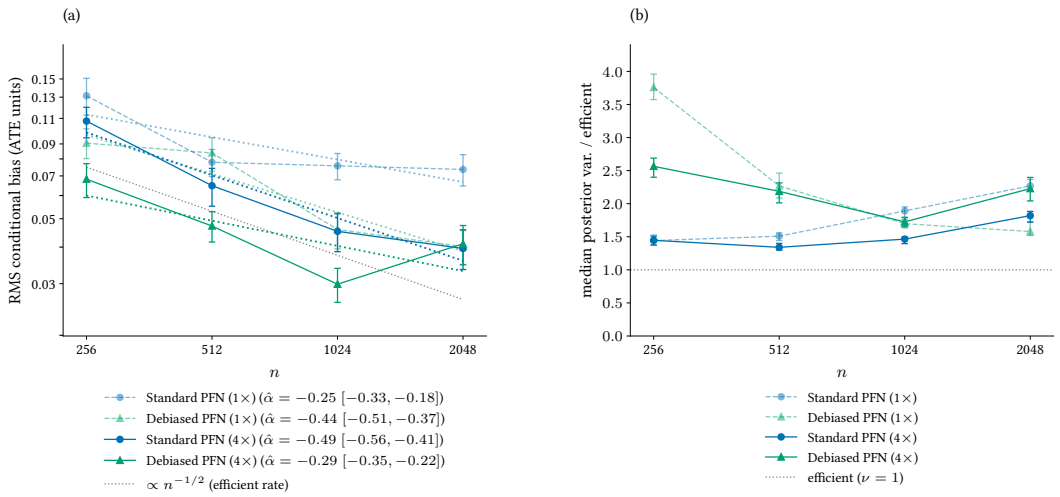

Figure 8: **Amortization budget: $1\times$ vs. $4\times$ training data.** Panels as in Fig. 3; solid: four datasets per training population, dashed: one. All fits use the same number of updates, so the larger training sets are traversed fewer times. The $4\times$ fits at $n = 2048$ did not finish annealing under the plateau schedule.

PFNs that differ only in the training label and do not include frequentist estimators with estimated nuisance functions (e. g., A-IPTW or TMLE) or post-hoc corrections of PFN posteriors (Melnychuk et al., 2026). For computational reasons, every combination of method and n is a single training run, so our intervals are conditional on the trained networks, and rates are fitted over $n = 256$ to 2048. Finally, we only focus on the ATE under a single synthetic prior with strong overlap, yet it is sufficiently complex to empirically validate our theoretical results.

G EXTENSIONS

Combined strategies. Interestingly, both correction strategies aim at the same goal (= reduction of the first-order error), yet they operate in different spaces: While (i) the label correction works in the target parameter space Ψ, (ii) the prior correction modifies the prior in the distributional space $\mathcal{P}$. Notably, both strategies can be combined: E. g., in Example B, the label correction can be used to correct the marginal distribution $P(x)$, and the prior correction is then applied only wrt. to the conditional outcome mean $\mu(x)$. In this way, both our strategies provide a very flexible toolkit to debias PFNs (depending on the ease of implementation, either (i) or (ii) can be chosen for each underlying nuisance function).