Targeted Maximum Likelihood Learning: An Optimization Perspective

Diyang Li
Cornell University
diyang01@cs.cornell.edu

Kyra Gan
Cornell University
kyragan@cornell.edu

Abstract

Targeted maximum likelihood estimation (TMLE) is a widely used debiasing algorithm for plug-in estimation. While its statistical guarantees, such as double robustness and asymptotic efficiency, are well-studied, the convergence properties of TMLE as an iterative optimization scheme have remained underexplored. To bridge this gap, we study TMLE's iterative updates through an *optimization-theoretic* lens, establishing *global convergence* under standard assumptions and regularity conditions. We begin by providing the first complete characterization of *different stopping criteria and their relationship to convergence* in TMLE. Next, we provide *geometric insights*. We show that each submodel induces a smooth, non-self-intersecting path (homotopy) through the probability simplex. We then analyze the solution space of the *estimating equation* and loss landscape. We show that all valid solutions form a submanifold of the statistical model, with the difference in dimension (i.e., codimension) exactly matching the dimension of the target parameter. Building on these geometric insights, we deliver the *first strict proof* of TMLE's convergence *from an optimization viewpoint*, as well as explicit sufficient criteria under which TMLE terminates in a single update. As a by-product, we discover an unidentified *overshooting* phenomenon wherein the algorithm can surpass feasible roots to the *estimating equation* along a homotopy path, highlighting a promising avenue for designing enhanced debias algorithms.

1 Introduction

Plug-in estimation, the approach of first estimating the data-generating distribution and then evaluating the target parameter on this estimate, is a natural strategy for estimating quantities such as quantiles, variance, *average treatment effects* (ATEs), and feature importance measures [1, 2]. Despite widespread use, these estimators often suffer from biases arising from the nonlinearity of parameter mappings, the finite-sample variability inherent to empirical distributions, or model misspecification, which can lead to unreliable inference in high-dimensional or complex settings [3, 4].

To address these limitations, Targeted Maximum Likelihood Estimation (TMLE) [5] offers a principled framework for constructing data-adaptive estimators by combining machine-learning (ML)-based initial estimates with targeted bias correction. At each step, TMLE updates the current distribution estimate along a fluctuation direction to reduce bias for the target parameter, achieving the desired double robustness and asymptotic efficiency under mild regularity conditions [6, 7]. These strengths have driven its adoption across areas including semi-supervised learning [8, 9], personalized medicine [10, 11], algorithmic fairness [12, 13], and off-policy evaluation [14], equipping practitioners with estimators that are both flexible and theoretically grounded.

While TMLE's asymptotic properties (e.g., double robustness and semi-parametric efficiency) are well-established [16], its *finite-sample behavior as an optimization procedure* remains underexplored. Current theory often treats iterative updates as implementation details for solving *estimating*

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

equations [16, 17]. While asymptotic guarantees depend on algorithmic convergence, this convergence is typically assumed, except in a few cases with known one- or two-step convergence (e.g., [42, 44]). However, most target parameters lack such one-step guarantees, making algorithmic convergence not just an implementation concern but a fundamental determinant of TMLE's debiasing performance. This gap is especially critical in finite-sample settings like healthcare [18, 19], where asymptotic guarantees offer no guidance for choosing convergence criteria, iteration limits, and tuning parameters. Without optimization-theoretic foundations, implementations remain ad-hoc [20, 21], risking reproducibility and silently degrading estimator quality.

We address this by analyzing TMLE's iterative trajectory and convergence through an *optimization-theoretic lens*. Unlike asymptotic efficiency results, our framework provides actionable insights for practical implementation. Conceptually, the convergence analysis of TMLE shares high-level similarities with that of *expectation-maximization* (EM) algorithms [22], both involving iterative schemes designed to optimize complex objective functions via tractable subproblems using alternating optimization frameworks. However, TMLE's *influence-function* (IF)-driven fluctuations within probability simplex-embedded homotopies pose unique challenges absent in classical EM literature [23, 24]. These technical differences necessitate tailored convergence analyses that differ significantly from those in EM. While recent works have explored optimization aspects for specific problems, e.g., causal effect estimation in exponential families [25] and off-policy evaluation with regularization [14], their scope remains limited to these special settings and does not apply to the general template.

Our contributions In this paper, we address a longstanding gap in the theoretical understanding of TMLE by recasting it as an explicit iterative optimization scheme. Our contributions are fourfold:

- **Geometric Insights.** We establish formal connections between different stopping criteria and the resulting algorithm convergence behavior (Theorem 1). Next, we show that each submodel induces a smooth homotopy mapping (or path), embedding each fluctuation into the probability simplex without self-intersection (Theorem 2), precluding cyclic or oscillatory behavior in TMLE. We demonstrate that the set of distributions satisfying the estimating-equation forms a smooth submanifold of the statistical model, whose codimension coincides exactly with the dimensionality of the target parameter (Theorem 3). This structural perspective reveals that TMLE iterates traverse a low-dimensional manifold within the ambient probability space, explaining their effectiveness in navigating a complex landscape with irregular likelihood surfaces (Theorem 3).
- **Convergence Guarantee.** We provide *the first rigorous proof* of TMLE's *global convergence* under mild regularity conditions (Theorem 4). Although this result is asymptotic, requiring an infinite number of iterations, it confirms long-standing empirical observations of TMLE's convergence behavior. Further, we derive explicit sufficient conditions under which TMLE terminates after a single update (Theorem 5). These conditions, based on the initial estimator and IF structure, offer verifiable criteria for simplified TMLE implementations.
- **Overshoot Behavior.** Our analysis reveals an unrecognized overshooting phenomenon, where TMLE may bypass feasible roots along the homotopy path (Theorem 6), potentially affecting finite-sample performance. This insight motivates safeguard strategies such as step-size control or root tracking that retain asymptotic guarantees while improving runtime and stability.
- **Interdisciplinary Impact.** By casting TMLE as an optimization procedure, we bridge statistical estimation with modern optimization theory. Conventional analyses of parametric optimizer using *convexity* typically require the submodel objective to be strictly (or strongly) convex. In contrast, our analyses based on non-intersection do not require such assumptions and thus hold under broader settings. This perspective enables reinterpretations of influence functions via tools like mirror descent, and suggests integrating adaptive acceleration or regularization paths into TMLE, potentially leading to new, theoretically grounded algorithms.

2 Preliminaries on Plug-in Estimation

This section introduces notation, reviews plug-in estimation, plug-in bias, and the influence function. Readers familiar with these concepts may skip ahead. Given the dataset $\{O_1, \ldots, O_n\}$ consists of n independent and identically distributed (i.i.d.) observations of a random variable o (e.g., an experimental unit) that fits an *unknown true* distribution P_0 with sample space $\mathcal{O}$. Our goal is to *efficiently* estimate a d-dimensional target parameter representing some statistical feature of interest (e.g., population mean and average treatment effect).

Notation Let P be a probability distribution with density p. Abusing the notation, *we use P and p interchangeably to denote the probability measure*. We let $\mathbb{P}_n$ denote the empirical measure (Definition A.5), and $\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(O_i)$. Let $L_0^2(P)$ denote the *space of mean-zero, finite-variance functions* with respect to the distribution P, i.e., $L_0^2(P) := \{h : \mathcal{O} \to \mathbb{R} : \mathbb{E}_P h(o)^2 < \infty, \mathbb{E}_P h(o) = 0\}$, where o is a generic random variable drawn from distribution P. We use $a \lesssim b$ to denote that there exists a constant C such that $a \leq Cb$. Let $\mathcal{D}_f$ be the Fréchet derivative, and we abbreviate $p_n(o)$ as p_n when no ambiguity arises. We use $\mathbb{C}^j$ to denote the class of mappings that are j-times continuously differentiable. A complete notation table is included in Appendix A.

Plug-in estimation We consider nonparametric estimation,[1] where the model class $\mathcal{M}$ contains all candidate distributions for P_0 on a σ-finite measurable space $(\Omega, \mathcal{F}, \nu)$, where each element $P \in \mathcal{M}$ admits a Radon-Nikodym density $p = dP/d\nu$ with respect to a dominating measure ν, satisfying $p \geq 0$ ν-a.e. and $\int_\Omega p \, d\nu = 1$. *To ease the derivation, we work with the density p of P with respect to a fixed dominating measure. The equivalent definitions in this section can be stated directly in terms of P, which is more common in the literature.* We assume p is uniformly bounded:[2]

Assumption 1 (Uniform Boundedness)**.** *There exists a $C_\infty < \infty$ s.t. $\|p\|_{L^\infty} \leq C_\infty$, $\forall p \in \mathcal{M}$.*

The target parameter functional $\Psi : \mathcal{M} \to \mathbb{R}^d$ then maps each candidate distribution to a corresponding feature of interest (e.g., for $d = 1, p \mapsto \mathbb{E}_P[o]$). The *plug-in estimator* $\hat{\psi}_n := \Psi(\widehat{p}_n)$ is obtained by applying Ψ to an empirical estimate $\hat{p}_n \in \mathcal{M}$ of the unknown p_0. Let $\psi_0 = \Psi(p_0)$ be the true value of our target parameter.

Plug-in Bias and Influence Function While plug-in estimators are often consistent under regularity conditions, they typically fail to achieve *asymptotic linearity* (Definition A.6)[3] due to bias inherited from the initial estimate $\hat{p}_n$[4] and the potential nonlinearity of Ψ. Even if $\hat{p}_n$ is consistent at the parametric rate, when Ψ is nonlinear in p, estimation errors in $\hat{p}_n$ can propagate nonlinearly through Ψ, introducing bias that invalidates a $\sqrt{n}$-linear expansion. As a result, plug-in estimators typically require bias correction to attain asymptotic linearity and efficiency.

A key tool for formalizing this limitation is the *influence function*, which quantifies how sensitive the target parameter Ψ is to small perturbations of the underlying distribution P. When $\mathcal{M}$ is nonparametric, the IF of the target parameter Ψ at a distribution P, $D^*_\Psi(p)(\cdot) : \mathcal{O} \to \mathbb{R} \in L_0^2(P)$ is unique. We formally introduce the influence function in Appendix A.4, Definition A.9.

To establish the asymptotic behavior of $\hat{\psi}_n$ can then be analyzed through a von Mises expansion. Let Ψ be *pathwise differentiable* (Definition A.8), and consider the perturbation path defined in Eq. (A.14) with $P = \widehat{P}_n$ and $Q = P_0$. Let P_0 be absolutely continuous with respect to $\widehat{P}_n$, and $dP_0/d\widehat{P}_n \in L_0^2(\widehat{P}_n)$. The estimation error in $\hat{\psi}_n$ can be decomposed into the following [26, 27]:

$$\hat{\psi}_n - \psi_0 = \mathbb{P}_n D^*_\Psi(p_0) - \underbrace{\mathbb{P}_n D^*_\Psi(\hat{p}_n)}_{\text{plug-in bias}} + \underbrace{(\mathbb{P}_n - P_0)[D^*_\Psi(\hat{p}_n) - D^*_\Psi(p_0)]}_{\text{empirical process term}} + \underbrace{R_2(\hat{p}_n, p_0)}_{\text{second-order remainder}}, \qquad (1)$$

where $R_2(\hat{p}_n, p_0)$ is the second-order remainder in the difference between $\hat{p}_n$ and p_0. The expansion in Eq. (1) resembles Definition A.6. While standard regularity conditions (e.g., Donsker class requirements and rate constraints on $\|\widehat{P}_n - P_0\|$) suffice to ensure the empirical process term and the second-order remainder are $o_{p_0}(1/\sqrt{n})$,[5] the first-order plug-in bias term typically remains non-negligible without correction.

[1] We analyze TMLE convergence in nonparametric models, revealing how target smoothness interacts with estimator complexity. Results extend to semi-parametric cases via subspace projections.

[2] This is commonly assumed in prior works to obtain statistical guarantees [5].

[3] Asymptotic linearity ensures $\sqrt{n}$-rate convergence to a normal distribution, with an asymptotic variance determined by the influence function, thereby facilitating efficient estimation and valid statistical inference.

[4] For example, $\hat{p}_n$ may not be consistent at the parametric rate, especially when estimated via neural networks.

[5] We refer readers to Van Der Laan and Rubin [5] and Cho et al. [27] for detailed assumptions.

3 Warm-up: TMLE Template

To correct the plug-in bias term, TMLE finds the solution p^*_n that solves the score equation

$$\mathbb{P}_n D^*_\Psi(p^*_n) = \frac{1}{n}\sum_{i=1}^{n} D^*_\Psi(p^*_n)(O_i) = \int_{\mathcal{O}} D^*_\Psi(p^*_n)(o) p_n d\nu(o) = \mathbf{0}, \tag{2}$$

by iterative refining the initial estimate $\hat{p}_n$.[6] We express this via sample space integration to connect with optimization-theoretic analysis.

Abusing the notation, let p^k_n be the k-th iteration of TMLE. At a high-level, after obtaining a "sufficiently good" initial $\hat{p}_n$ (e.g., via ML), TMLE selects a parametric submodel $\{p(\epsilon)\}_{\epsilon\in\mathbb{R}} \subset \mathcal{M}$ (Definition 2)[7] guided by the IF to maximize sensitivity to the target parameter Ψ at $\hat{p}_n$. It then updates $\hat{p}_n$ by fitting ϵ along this submodel (typically via minimizing a loss function) to obtain an updated estimate p^1_n. This procedure is repeated until no further improvement (i.e., no nonzero ϵ) can be found (if achieved). A pseudo-code of TMLE is provided in Algorithm 1. If TMLE terminates after k iterations, the final estimate P^k_n achieves asymptotic efficiency under standard regularity conditions. However, the actual convergence behavior of this iterative procedure remains unestablished.

Definition 1 (TMLE Loss)**.** *Define the TMLE loss function $\mathbf{L} : \mathcal{M} \to (\mathcal{O} \to \mathbb{R}_{\geq 0})$ such that*

$$p_0 \in \arg\min_{p} \int_{\Omega} \mathbf{L}(p) p_0(o) d\nu(o). \tag{3}$$

Similar to classical maximum likelihood estimation, we use a loss function $\mathbf{L}(\cdot)$ such that the mapping (3) is minimized at the true density p_0. E.g., one may use the negative log-likelihood, $\mathbf{L}(p) := -\log p$.

Definition 2 (Fluctuation Submodel)**.** *Let $\mathcal{R}$ denote an open subset of $\mathbb{R}^d$. We define a family of fluctuation submodel $\{p(\epsilon) : \epsilon \in \mathcal{R}\}$ (a.k.a. parametric working model) that follows (i) $\{p(\epsilon) : \epsilon\} \subset \mathcal{M}$; (ii) the submodel through p at $\epsilon = \mathbf{0}$; (iii) a linear combination of the components of "score" $d\mathbf{L}(p(\epsilon))/d\epsilon$ at $\epsilon = \mathbf{0}$ recovers the IF (cf. Definition 3).*

$\mathcal{R}$ denotes the set of ϵ values for which $p^k_n(\epsilon)$ is a proper density. Common parametric submodels include linear (Example 1) and exponential (Example 2).

Example 1 (Linear Reparameterization)**.** *For $\epsilon \in \mathcal{R}$, the instances of additive perturbation include*

$$p^k_n(\epsilon) \triangleq \left(1 + \epsilon^\top D^*_\Psi(p^k_n)\right) p^k_n. \tag{4}$$

Example 2 (Exponential Family)**.** *For $\epsilon \in \mathcal{R}$, the instances of exponential tilting include*

$$p^k_n(\epsilon) \triangleq C(\epsilon, p^k_n) \exp\left(\epsilon^\top D^*_\Psi(p^k_n)\right) p^k_n, \tag{5}$$

$$p^k_n(\epsilon) \triangleq C'(\epsilon, p^k_n) \left\{1 + \exp\left(-2\epsilon^\top D^*_\Psi(p^k_n)\right)\right\}^{-1} p^k_n, \tag{6}$$

for $C(\epsilon, p^k_n)$, $C'(\epsilon, p^k_n)$ be normalizing constants (defined in Definition B.10).

Definition 3 (Relaxed Score Condition)**.** *Let A be a constant matrix with $\|A\| < \infty$. TMLE requires that every parametric submodel has a sufficient statistic equal to the IF, i.e.,*

$$\left.\frac{d\mathbf{L}(p(\epsilon))(o)}{d\epsilon}\right|_{\epsilon=\mathbf{0}} \equiv A \cdot D^*_\Psi(p)(o) \quad \textit{for all possible values } o \in \mathcal{O}. \tag{7}$$

Definition 3 imposes the statistical connections between the loss and the IF in TMLE. Additionally, we make the following mild regularity assumption, ensuring that the solution of Algorithm 1 is achieved in the interior of $\mathcal{M}$:

Assumption 2 (van der Laan et al. [6])**.** *Let $\mathcal{M}$ be a sufficiently rich model class such that the iterated model is locally saturated. Under this condition, the minimization invoked in line 3 of Algorithm 1 admits its solution strictly within the interior of $\mathcal{M}$.*

[6] In Eq. (2), we adopt an equivalent integral representation to facilitate the convergence analysis.

[7] We follow the common TMLE convention of using p for both submodels and probabilities, where no ambiguity arises.

Algorithm 1 Targeted Maximum Likelihood Estimator (TMLE)

```
Input: Data $\{O_i\}_{i=1}^n$, canonical gradient $D_\Psi^*(p)$ of the interested functional $\Psi$, initial estimator $p_n^0$.
$k \leftarrow 0$, initialize $\epsilon_n^0 \neq \mathbf{0}$.
while $\epsilon_n^k \neq \mathbf{0}$ do
    $\epsilon_n^k \leftarrow \arg\min_{\{\epsilon: p_n^k(\epsilon) \in \mathcal{M}\}} \int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon))(o) p_n \, d\nu(o)$
    $p_n^{k+1} \leftarrow p_n^k\left(\epsilon_n^k\right)$, $k \leftarrow k+1$    \\ iterative debiasing
end while
$p_n^* \leftarrow p_n^{k-1}$, $\hat{\psi}_n \leftarrow \Psi(p_n^*)$    \\ plug-in estimation
Output: Targeted estimator $\hat{\psi}_n$.
```

4 Main Results

This section presents our main theoretical results, framing TMLE through an optimization lens for a pathwise differentiable target parameter (Definition A.8). Throughout, we assume that the standard regularity conditions (Assumptions 1, 2) as well as our mild assumptions (Assumptions 3, 4, 5) hold globally unless explicitly stated otherwise.

Assumption 3 (Smooth Link)**.** *Let the link function $\breve{f} \colon L_0(\nu) \mapsto L_0(\nu)$ be $\mathbb{C}^2$ smooth. We assume that $p_n^k(\epsilon)$ admits the representation $p_n^k(\epsilon) \triangleq \breve{f}(\epsilon^\top D_\Psi^*(p_n^k))$ where $\breve{f}(\cdot)$ is injective on its effective domain, namely the linear span of the components of D_Ψ^*.*

Assumption 3 ensures that the perturbation parameter ϵ and the IF $D_\Psi^*(\cdot)$ always appear jointly in the form $\epsilon^\top D_\Psi^*(\cdot)$. Meanwhile, the injectivity imposed on the mapping $\breve{f}$ is typically mild in practice and readily satisfied by virtually all mainstream submodels such as those defined in Examples 1 and 2. A formal justification of this injectivity condition is provided in Appendix D.1.

Assumption 4 (Differentiability and Lipschitz-in-path)**.** *The mappings $\epsilon \mapsto p(\epsilon)$ and $\epsilon \mapsto \mathbf{L}(\epsilon)$ are of class $\mathbb{C}^2$ and $\mathbb{C}^3$ in ϵ, respectively. The mappings $p \mapsto \mathbf{L}(p)$ and $p \mapsto D_\Psi^*(p)$ are twice continuously Fréchet-differentiable. And for $k \in \mathbb{Z}^+$, $\epsilon_1, \epsilon_2 \in \mathcal{R}$, and $p_1, p_2 \in \mathcal{M}$ we have*

$$\begin{aligned} \left\|\mathbf{L}(p_n^k(\epsilon_1)) - \mathbf{L}(p_n^k(\epsilon_2))\right\|_{L^2(\nu)} &\lesssim \|\epsilon_1 - \epsilon_2\|_2 \,, \\ \left\|\mathbf{L}\left(p_1\right) - \mathbf{L}\left(p_2\right)\right\|_{L^2(\nu)} &\lesssim \|p_1 - p_2\|_{L^2(\nu)} \,, \\ \left\|D_\Psi^*(p_n^k(\epsilon_1)) - D_\Psi^*(p_n^k(\epsilon_2))\right\|_{L^2(\nu)} &\lesssim \|\epsilon_1 - \epsilon_2\|_2 \,. \end{aligned} \tag{8}$$

Assumption 5 (Metric-subregularity of Gradient)**.** *There exists a neighborhood of origin s.t. for $k \in \mathbb{Z}^+$ and all ϵ_1, ϵ_2 in that neighborhood, with $\nabla_\epsilon \mathbf{L}(\cdot)$ evaluated at $o \leftarrow O_i$, we have*

$$\|\epsilon_1 - \epsilon_2\|_2 \lesssim \left\|\nabla_\epsilon \mathbf{L}(p_n^k(\epsilon_1)) - \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon_2))\right\|_2 \lesssim \|\epsilon_1 - \epsilon_2\|_2 \,. \tag{9}$$

Assumption 4 is a standard and widely adopted assumption within the optimization literature. Likewise, Assumption 5 is notably mild within the context of TMLE optimization. E.g., one can trivially verify that the log-likelihood under an exponential tilt naturally satisfies (9). Further, the second inequality in (9) is essentially inherited from the vanilla TMLE literature (albeit expressed differently).

4.1 What do we really mean when we talk about 'convergence'?

In TMLE practice, we observe that iterative stopping ("convergence") criteria are often applied inconsistently. To maintain theoretical rigor, we explicitly define TMLE convergence as the convergence in probability of the iterates to a deterministic limiting distribution in $\mathcal{M}$. While previous studies like [5, 6] have shown that the convergence implies solving (2), we generalize this result to Theorem 1.

Theorem 1 (Stopping Condition Equivalence)**.** *The condition $\lim_{k\to\infty} \epsilon_n^k = \mathbf{0}$ is both necessary and sufficient for the convergence of Algorithm 1. If Algorithm 1 does converge, then the iterates $\{p_n^k\}_{k\geq 0}$ admit a limit $\lim_{k\to\infty} p_n^k \rightsquigarrow p_n^*$ where p_n^* lies on the solution manifold of (2).*

Theorem 1 rigorously characterizes and clarifies the underlying relationships among several commonly employed stopping conditions, visually illustrating these connections in Figure 1. Theorem 1 serves as an essential foundation for our subsequent optimization studies. It is also worth noting that convergence of `TMLE` is a sufficient but *not* necessary condition for solving the Estimating Equation (2), which implies the potential occurrence of overshooting phenomena during algorithmic iterations (as discussed later in Section 4.4).

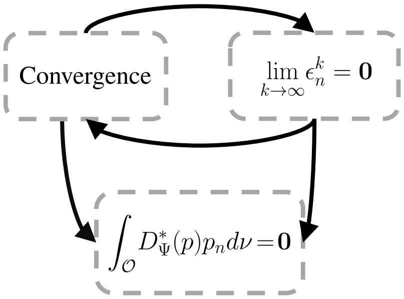

Convergence
$\lim_{k\to\infty} \epsilon_n^k = \mathbf{0}$
$\int_{\mathcal{O}} D^*_\Psi(p) p_n d\nu = \mathbf{0}$

Figure 1: Illustration of Thm. 1.

4.2 Global convergence

The state-of-the-art convergence result regarding Algorithm 1 is currently represented by Lemma 1.

Lemma 1. *In Algorithm 1, the sequence of empirical risks* $\left\{\int_{\mathcal{O}} \mathbf{L}(p_n^k) p_n d\nu\right\}_k$ *converges as* $k \rightsquigarrow \infty$.

Proof. See Section 3 in Van Der Laan and Rubin [5] for a detailed proof. □

However, it is a common consensus within the optimization community that convergence of the empirical loss alone does not necessarily imply convergence of the iterates themselves. This issue becomes even more challenging in a semi-parametric context.[8] We provide motivating examples of `TMLE` divergence without standard assumptions in Appendix I. Nonetheless, our analysis will build upon Lemma 1, and we first present several preparatory results essential to our analysis.

Theorem 2 (Non-self-intersection)**.** *The homotopy path* $\{p(\epsilon) : \epsilon\}$ *in the probability simplex never self-intersects for* $d = 1$. *Further assume the non-degeneracy condition*

$$\forall \beta \in \mathbb{R}^d \backslash \{\mathbf{0}\}, \;\; \mathbb{P}\left\{o \in \mathcal{O} : \beta^\top D^*_\Psi(p)(o) \neq 0\right\} > 0. \tag{10}$$

Then the $\epsilon \mapsto p(\epsilon)$ *defines a one-to-one* $\mathbb{C}^1$ *embedding of* $\mathbb{R}^d$ *into the probability simplex, and its image is free of self-intersections for every* $d \in \mathbb{Z}^+$.

Remark 1. *The assumption for* $d \geq 2$ *is to require that the* d *components of* $D^*_\Psi(p)(o)$ *are linearly independent in* $L^2(\nu)$. *Equivalently,* $\Sigma = \int D^*_\Psi(p)(o) D^*_\Psi(p)(o)^\top p(o) d\nu(o)$ *is a full-rank* $d \times d$ *matrix.*[9]

Theorem 3 (Solution Submanifold)**.** *Assume* $n < \infty$ *and the Fréchet derivative* $\mathcal{D}_f D^*_\Psi$ *is surjective for every* $p \in \mathcal{M}$, *then both the (i) IF Estimating Equation (2), and (ii) nonparametric loss landscape* $\min_{\{p \in \mathcal{M}\}} \int_{\mathcal{O}} \mathbf{L}(p) p_n d\nu$ *admit infinitely many solutions which form a smooth submanifold (or continuum) of a whole equivalence class in* $\mathcal{M}$.

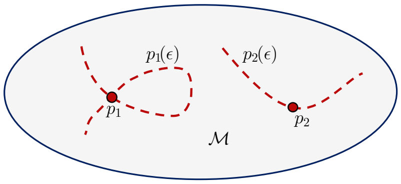
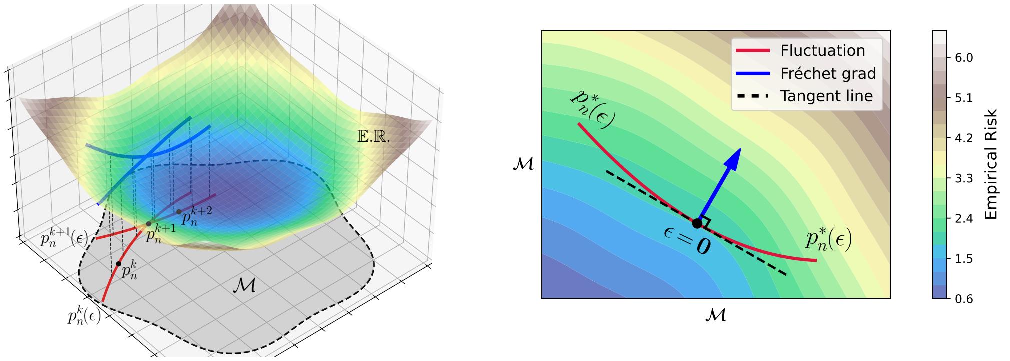


Figure 2: An illustration of self-intersection in $\mathcal{M}$. The path $p_1(\epsilon)$ (left) is self-intersect while the other $p_2(\epsilon)$ (right) is not.

We know that $\epsilon = \mathbf{0}$ implies $p(\epsilon) = p$, while Theorem 2 characterizes the converse direction, i.e., if $p(\epsilon) = p$ then $\epsilon = \mathbf{0}$. It also reveals a profound geometric structure underlying `TMLE`'s iterative process. Specifically, the homotopy mapping induced by each fluctuation submodel defines a smooth embedding into the probability simplex. This embedding represents a continuous deformation of the statistical model that preserves its topological structure without self-intersections (cf. Figure 2), ensuring that each `TMLE` fluctuation traverses a well-defined homotopy path that cannot revisit the same density twice, preventing potential cycling behavior. Theorem 3 implies that the solution submanifold is inherently infinite-dimensional, consisting of uncountably many solutions. However, it forms a well-defined, smooth "sheet" of equivalent solutions that simultaneously solve the estimating equation and the (approximate) empirical-risk minimization problem. Inspired by Theorem 3, we can geometrically interpret the iterative optimization procedure of `TMLE` on the

[8] In this context, “semi-parametric” refers to settings where the target parameter is defined parametrically, but the plug-in distribution is estimated using nonparametric methods.

[9] Since $\beta^\top \Sigma \beta = \int p(\beta^\top D^*_\Psi(p))^2 d\nu$, where by assumption the integrand is strictly positive in part of domain.



(a) Understanding of iterative process.

(b) Characterization of limiting distribution.

Figure 3: The conceptual diagram of `TMLE` procedure. In 3a, the gray region represents model $\mathcal{M}$ and $\mathbb{E}.\mathbb{R}.$ denotes the empirical risk. At k-th iteration, we span a submodel (red curve) around p_n^k in $\mathcal{M}$, then project the path onto the loss landscape (blue curve). We select the point corresponding with the minimal loss as p_n^{k+1}. 3b is the top-down perspective of 3a, which indicates the Fréchet gradient of loss evaluated at the p_n^* is orthogonal (in the L^2 sense) to the tangent line of the submodel at $\epsilon = \mathbf{0}$.

empirical loss landscape (cf. Figure 3a), as well as characterize the geometric structure of the limiting distribution (cf. Figure 3b).

Grounded in the insights from Theorems 2 and 3, our asymptotic convergence guarantees are now provided in Theorem 4, demonstrating the infinite-step convergence behavior of `TMLE` even when the initial estimator is heavily misspecified in $\mathcal{M}$.

Theorem 4 (Iterative Convergence)**.** *Let $\{p_n^k\}_{k\geq 0}$ be the sequence of density estimates produced by the `TMLE` Algorithm 1. There exists a limiting density $p_n^* \in \mathcal{M}$ such that $\lim_{k\to\infty} p_n^k \rightsquigarrow p_n^*$.*

Remark 2. *Theorem 4 covers many well-known `TMLE` instances like D´iaz and Rosenblum [25]. Note that even if Assumption 5 may not hold for some `TMLE` practices, we can still establish asymptotic regularity (or pseudo-convergence) of the iterates as $\lim_{k\to\infty} \|p_n^{k+1} - p_n^k\|_1 = 0$ (a.k.a., quasi-Cauchy sequence).*

We emphasize that the proof of Theorem 4 does not depend on the quality of the initial estimate p_n^0, meaning that Algorithm 1 guarantees asymptotic convergence to a solution of the *estimating equation* from *any* starting point (Theorem 1). However, a poor initial estimate can slow the decay of the empirical process term and the second-order remainder in Eq. (1), preventing the estimator from achieving parametric-rate efficiency asymptotically [6].

4.3 One-step property

Theorem 5 (One-step Property)**.** *The semi-parametric `TMLE` procedure performs exactly one update when one of the following conditions is met:*

(i) *initial density p_n^0 already on the solution manifold of (2) and $\int_{\mathcal{O}} \mathbf{L}$ is strictly convex in ϵ;*

(ii) *for $\forall o \in \mathcal{O}$, the mapping $\epsilon \mapsto D_\Psi^*(p(\epsilon))(o)$ is a conservative (i.e., curl-free) vector field in ϵ and the loss satisfies line integral $\mathbf{L}(p(\epsilon))(o) \triangleq \mathbf{L}(p(\mathbf{0}))(o) + A\int_0^\epsilon D_\Psi^*(p(u))(o)du$;*

(iii) *in some of the practical problems on outcome regression (e.g., ATE, propensity-score intervention) with the existence of "clever covariate" and a proper fluctuation submodel;*

(iv) *hit the user-set convergence criteria (e.g., machine precision, number of iterations).*

In the (*i*), (*ii*) of Theorem 5, we firstly establish a set of sufficient conditions under which the `TMLE` can be terminated after one single iteration. Putting together with existing conditions (*iii*) and (*iv*), the whole theorem significantly extends various scattered conditions previously dispersed throughout existing literature, e.g., [42, 43, 25, 44]. It is important to clarify that this one-step property does *not* imply convergence in the mathematical sense; instead, it indicates the algorithm has achieved a predefined stopping criterion (e.g., satisfying Estimating Equation (2)) and obtained desirable properties for the targeted estimator of interest.

4.4 Potential overshooting

As a by-product of our analysis into the optimization, we uncovered an interesting phenomenon wherein a `TMLE` update may overshoot a feasible solution to the *estimating equation* along the homotopy path induced by the fluctuation submodel (cf. Figure 4). We establish theoretical existence of this overshooting phenomenon through a concrete illustrative example presented in Example 3.

Example 3 (Degenerate Hyperplane)**.** *Consider Submodel (4) with log-likelihood loss. Solving*

$$\int_{\mathcal{O}} D^*_{\Psi}(p_n^k)(o)\left(1+\epsilon^\top D^*_{\Psi}(p_n^k)(o)\right)^{-1} p_n\, d\nu(o) = \mathbf{0} \tag{11}$$

with $\epsilon \neq \mathbf{0}$ gives the next movement of `TMLE`. If all of the $D^*_{\Psi}(p_n^k)(O_i)$ happen to lie in a single affine hyperplane $\{x : \epsilon^\top x = \chi\}$ where $\chi \in \mathbb{R}$ is a constant, the resulting distribution would exactly solve the Estimating Equation (2) while Algorithm 1 keeps looping.

Motivated by Example 3, we present a more formal definition of this phenomenon.

Definition 4 (Overshooting of `TMLE`)**.** *At the k-th iteration of `TMLE`, we say that the Algorithm 1 overshoots a feasible solution if there exists $\epsilon^\dagger \neq \epsilon_n^k$ such that $p_n^k(\epsilon^\dagger) \in \mathcal{M}$ and $\int_{\mathcal{O}} D^*_{\Psi}(p_n^k(\epsilon^\dagger))(o) p_n d\nu(o) = \mathbf{0}$.*

Theorem 6 (Overshoot Control)**.** *If we further assume that*

(i) the population Jacobian $\int_{\Omega} p_0 \nabla_\epsilon D^*_{\Psi}\left(p_n^k(\epsilon)\right)\big|_{\epsilon=\mathbf{0}} d\nu$ has positive minimal eigenvalue λ_o,

(ii) $\left\| \int_{\mathcal{O}} p_n\left[\nabla_\epsilon \mathbf{L}\left(p_n^k(\epsilon)\right) - \nabla_\epsilon \mathbf{L}\left(p_n^k(\mathbf{0})\right)\right] d\nu - \int_{\mathcal{O}} p_n \nabla^2_\epsilon \mathbf{L}(p_n^k(\epsilon))\big|_{\epsilon=\mathbf{0}} d\nu \cdot \epsilon \right\| \lesssim \|\epsilon\|^2 .$

Then, the probability that `TMLE` overshoots (o.s.) the nearest root of the estimating equation *is*

$$\mathbb{P}\left[o.s.\right] \lesssim 2d \exp\left(-\frac{n\lambda_o^2 C_o^2 \tilde{\mu}}{2B_o^2}\right) + 2d \exp\left(-\frac{n\tilde{\mu}}{2B_o^2}\right), \quad \tilde{\mu} = \left\| \int_{\Omega} p_0 D^*_{\Psi}(p_n^k) d\nu \right\|_\infty^2 , \tag{12}$$

where B_o, C_o are constants specified in Appendix H.

While a systematic characterization of the exact conditions leading to overshooting remains an open challenge, we derive preliminary probabilistic guarantees under stronger assumptions in Theorem 6, particularly the uniform boundedness of second-order remainders, as specified in Condition *(ii)* above. Our findings indicate that the exponent in (12) scales linearly with sample size n, leading to *an exponential reduction in the overshooting probability bound as n increases*. Additionally, *higher*-dimensional Ψ *increases* the likelihood of overshooting and complicates the convergence. Intuitively, when $d \geq 2$, the solution set forms a complex manifold, where the risk function may exhibit heterogeneous curvature, flat in some directions while sharply curved in others. Consequently, optimizers following steepest descent directions may escape the manifold before reaching a risk minimum, overshooting potential solutions. Furthermore, common subproblem solvers employing backtracking line searches or momentum acceleration are particularly prone to this phenomenon. The combination of adaptive step sizing and inertia can cause the algorithm to leap over viable roots before the termination criteria are triggered.

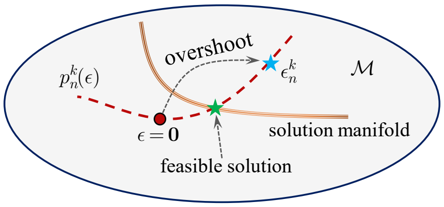


Figure 4: Illustration of an overshooting where the `TMLE` update passes through an feasible EIF-root ★ but continues to the next iterate ★.

5 Proof Overview

This section sketches the high-level ideas behind our proofs.

Proof Sketch of Theorem 1. This theorem naturally decomposes into 4 independent sub-results, which we denote by `E1`, `E2`, `E3`, and `E4` (see Figure 5). For `E1`, we begin by showing that the sequence $\{p_n^k\}_{k\geq 0}$ enjoys a tail-sum additivity property in the underlying metric space. By carefully bounding the increments and summing over the fluctuation steps, we conclude that $\{p_n^k\}$ is a Cauchy sequence. In proving `E2`, we derive both compact upper and lower bounds on the fluctuation parameters $\{\|\epsilon_n^k\|\}$. Applying the squeeze theorem we deduce that the norm of ϵ_n^k must converge to the origin. Sub-result `E3` follows directly from the classical analysis in Van Der Laan and Rubin [5], and it is worth noting that the proof remains straightforward under our framework. Lastly, sub-result `E4` emerges as a trivial generalization of sub-result `E3`. □

Proof Sketch of Theorem 2. Since the fluctuation submodel is defined via an injective link on its effective domain, in the scalar case ($d = 1$) we show that the Hellinger distance between $p(\epsilon_1)$ and $p(\epsilon_2)$ never vanishes. Hence $\epsilon \mapsto p(\epsilon)$ is strictly monotonic in the simplex. For general d, the non-degeneracy condition (10) guarantees that $\mathcal{D}_f p(\epsilon)[h] \neq 0$ whenever $h \neq 0$. One shows that $\|\epsilon_1 - \epsilon_2\| \lesssim \|p(\epsilon_1) - p(\epsilon_2)\| \lesssim \|\epsilon_1 - \epsilon_2\|$, thus p is a bi-Lipschitz embedding of $\mathbb{R}^d$ into the simplex and cannot self–intersect by Hadamard's global inverse function theorem. □

Proof Sketch of Theorem 3. We first show that the functional $h \mapsto \int_{\mathcal{O}} D_\Psi^*(p_0 + h)(o) p_n d\nu(o)$ is locally Lipschitz-type continuous in an $L^2(\nu)$-neighborhood of the reference solution p_0. Using a second-order Fréchet expansion we derive $\| \int_{\mathcal{O}} (D_\Psi^*(p_0 + h_1) - D_\Psi^*(p_0 + h_2)) p_n d\nu\|_{\mathbb{R}^d} \lesssim \|h_1 - h_2\|_{L^2(\nu)}$. Given that the Fréchet derivative $\mathcal{D}_f D_\Psi^*(p_0)$ is assumed surjective, the rank–nullity theorem for Banach spaces ensures that the null space must be infinite-dimensional. The infinite-dimensional implicit function theorem then applies due to continuity and surjectivity conditions, which guarantees that the solution locus is a $\mathbb{C}^1$ manifold. The proof of *(ii)* technically follows a similar process. □

Proof Sketch of Theorem 4. Along the real-analytic fluctuation path, we first show that $\|\epsilon_n^k\|^2 \lesssim \int_{\mathcal{O}} (\mathbf{L}(p_n^k)(o) - \mathbf{L}(p_n^{k+1})(o)) p_n d\nu(o)$ using Lipschitz-in-path. Based on Lemma 1, telescoping this inequality shows that the squared step sizes form a summable series $\sum_{k=0}^{\infty} \|\epsilon_n^k\|^2$. Therefore, we get $\epsilon_n^k \rightsquigarrow \mathbf{0}$, combining these facts with the equivalence statement in Theorem 1, the vanishing of ϵ_n^k is both necessary and sufficient for convergence of the `TMLE` iterates. □

Proof Sketch of Theorem 5. For case *(i)*, the result follows directly by substituting into the algorithmic framework and exploiting convexity arguments. In the analysis of *(ii)*, we leverage fundamental properties of line integrals along with optimality conditions to demonstrate that the updated density after one iteration lies within a component of the solution manifold. For case *(iv)*, we rigorously establish both the existence and appropriateness of stopping conditions introduced therein. We omit the proof for *(iii)*, as it is highly problem-specific and can be found in the corresponding paper. □

Proof Sketch of Theorem 6. We first linearize the empirical score map at the origin, writing it as a fixed Jacobian term plus a Lipschitz remainder. The minimal-norm root therefore lies within a factor of the score norm, and a single loss-based update stays in the same radius. Overshoot can happen if the score at the origin is already large compared with that radius. Each coordinate of that score is a bounded average, by applying Hoeffding's inequality twice we obtain the desired bound. □

Remark 3. *We also provide some toy numerical examples to validate partial results in Appendix B.1.*

6 Discussions

In this work, we have laid a rigorous optimization-theoretic foundation for `TMLE`. Our analysis assumes exact solution of each subproblem, whereas in numerical practice one always computes an approximate $\hat{\epsilon}_n^k$ satisfying $\|\hat{\epsilon}_n^k - \epsilon_n^k\|_2 \leq \sigma_\epsilon$ and setting $p_n^{k+1} \leftarrow p_n^k(\hat{\epsilon}_n^k)$. How such a gap affects convergence guarantees remains unclear. Meanwhile, the non-self-intersection property of submodel paths plays a subtle role and may carry implications for algorithmic stability. Further, our framework addresses only canonical first-order `TMLE` while extending to higher-order variants [28] is also an important directions for future research.

References

[1] Peter J Bickel and Ya'acov Ritov. Nonparametric estimators which can be" plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003.

[2] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[3] Simon J Sheather and James Stephen Marron. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416, 1990.

[4] Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, pages 2300–2312, 1997.

[5] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

[6] Mark J van der Laan, Sherri Rose, and Susan Gruber. Readings in targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series.*, 2009.

[7] Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

[8] Kristin E Porter, Susan Gruber, Mark J Van Der Laan, and Jasjeet S Sekhon. The relative performance of targeted maximum likelihood estimators. *The international journal of biostatistics*, 7(1):0000102202155746791308, 2011.

[9] S Ghazaleh Dashti, Katherine J Lee, Julie A Simpson, Ian R White, John B Carlin, and Margarita Moreno-Betancur. Handling missing data when estimating causal effects with targeted maximum likelihood estimation. *American Journal of Epidemiology*, 193(7):1019–1030, 2024.

[10] Michael R Kosorok and Erica EM Moodie. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, 2015.

[11] Susan Gruber, Rachael V Phillips, Hana Lee, Martin Ho, John Concato, and Mark J van der Laan. Targeted learning: toward a future informed by real-world evidence. *Statistics in Biopharmaceutical Research*, 16(1):11–25, 2024.

[12] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[13] Alexander Asemota and Giles Hooker. Targeted learning for data fairness. *arXiv preprint arXiv:2502.04309*, 2025.

[14] Aurelien Bibaut, Ivana Malenica, Nikos Vlassis, and Mark Van Der Laan. More efficient off-policy evaluation through regularized targeted learning. In *International Conference on Machine Learning*, pages 654–663. PMLR, 2019.

[15] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

[16] Susan Gruber and Mark Van Der Laan. tmle: an r package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51:1–35, 2012.

[17] Megan S Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185(1):65–73, 2017.

[18] Pranab K Sen, Julio M Singer, and Antonio C Pedroso de Lima. *From finite sample to asymptotic methods in statistics*. Cambridge University Press, 2010.

[19] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[20] Laura B Balzer and Ted Westling. Invited commentary: demystifying statistical inference when using machine learning in causal research. *American Journal of Epidemiology*, 192(9): 1545–1549, 2023.

[21] Helene CW Rytgaard and Mark J van der Laan. Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Analysis*, 30(1):4–33, 2024.

[22] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13 (6):47–60, 1996.

[23] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[24] Russell A Boyles. On the convergence of the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 45(1):47–50, 1983.

[25] Iván Díaz and Michael Rosenblum. Targeted maximum likelihood estimation using exponential families. *The international journal of biostatistics*, 11(2):233–251, 2015.

[26] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[27] Brian M Cho, Yaroslav Mukhin, Kyra Gan, and Ivana Malenica. Kernel debiased plug-in estimation: Simultaneous, automated debiasing without influence functions for many target parameters. In *International Conference on Machine Learning*, pages 8534–8555. PMLR, 2024.

[28] Mark van der Laan, Zeyi Wang, and Lars van der Laan. Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290*, 2021.

[29] A.W. van der Vaart. *Semiparametric statistics*, pages 331–457. Number 1781 in Lecture Notes in Math. Springer, 2002. MR1915446.

[30] Dennis K Burke. Cauchy sequences in semimetric spaces. *Proceedings of the American Mathematical Society*, 33(1):161–164, 1972.

[31] Houshang H Sohrab. *Basic real analysis*, volume 231. Springer, 2003.

[32] Raymond EAC Paley and Antoni Zygmund. On some series of functions,(3). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 190–205. Cambridge University Press, 1932.

[33] Richard Johnsonbaugh. Summing an alternating series. *The American Mathematical Monthly*, 86(8):637–648, 1979.

[34] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[35] Brian Davies. *Integral transforms and their applications*, volume 41. Springer Science & Business Media, 2002.

[36] D Martin and LV Ahlfors. *Complex analysis*. New York: McGraw-Hill, 1966.

[37] Reinhold Meise and Dietmar Vogt. *Introduction to functional analysis*. Clarendon press, 1997.

[38] Henk P Barendregt and Erik Barendsen. Introduction to lambda calculus. Unpublished manuscript, Radboud University Nijmegen, 1984.

[39] Jean-Pierre Aubin. *Applied functional analysis*. John Wiley & Sons, 2011.

[40] Antonio Ambrosetti and Giovanni Prodi. *A primer of nonlinear analysis*. Number 34 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[41] Stephen M Robinson. An implicit-function theorem for a class of nonsmooth functions. *Mathematics of operations research*, 16(2):292–309, 1991.

[42] Michael Rosenblum and Mark J Van Der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The international journal of biostatistics*, 6(1), 2010.

[43] Mireille E Schnitzer, Erica EM Moodie, and Robert W Platt. Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification. *Biostatistics*, 14(1):1–14, 2013.

[44] Helene CW Rytgaard and Mark J van der Laan. One-step tmle for targeting cause-specific absolute risks and survival curves. *arXiv preprint arXiv:2107.01537*, 2021.

[45] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[46] Jean-Philippe Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9(1-2):187–205, 1982.

Appendix

The appendix is organized as follows. Appendix A provides the Notation Table and Formal Definitions. Appendix B outlines the Proof Preliminaries. Appendix C presents the Proof of Theorem 1, including proofs of sufficiency (`E1`), necessity (`E2`), and conditions (`E3`) and (`E4`). Appendix D contains the Proof of Theorem 2, with concrete analysis on selected submodels. Appendices E through H present the Proofs of Theorem 3 to 6, respectively. Appendix I discusses the Non-convergence of `TMLE`.

A Notation Table and Formal Definitions

Table A.1: Table of main notations.

Notations	Descriptions
$a \lesssim b$	There exists a constant C such that $a \leq C\,b$.
$\mathcal{D}_f$	The Fréchet derivative operator.
$\mathbb{C}^j$	Class of mappings that are j-times continuously differentiable.
Ω	The domain on which random variables o are defined.
$\mathcal{F}$	A σ-algebra of subsets of Ω, defining which events are measurable.
$(\Omega, \mathcal{F}, \nu)$	A σ-finite measurable space with dominating measure ν.
P, p	Probability measure defined on $(\Omega, \mathcal{F})$ and its density $dP/d\nu$
P_0, p_0	Those associated with the true data-generating probability measure
$P \ll \nu$	P is dominated by measure ν.
$\mathcal{O}$	Sample space of the possible observations.
$\mathcal{O} \times \mathcal{M}$	The cartesian product of $\mathcal{O}$ and $\mathcal{M}$
$\mathbb{P}_n, p_n$	Empirical measure (Definition A.5), and its density $d\mathbb{P}_n/d\nu$.
$\mathbb{P}_n f$	$\frac{1}{n}\sum_{i=1}^n f(O_i)$
$L_0^2(P)$	Space of mean-zero, finite-variance functions with respect to P
$\widehat{P}_n, \hat{p}_n$	Estimated probability distribution using n samples and its density in $\mathcal{M}$
ψ_0	True parameter value $\Psi(p_0)$.
$D^*_\Psi(p)(o), D^*_{\Psi,j}$	Efficient influence function at p, and the j-th component of it
R_n	Remainder (substitution bias) in the first-order expansion of Ψ.
$\mathbf{1}\{\cdot\}$	Indicator function.
A	The constant matrix in $\mathbb{R}^{d\times d}$.
$\mathcal{R}$	An open subset of $\mathbb{R}^d$ where the submodel index is valid.
ϵ, ϵ_j	The fluctuation submodel parameter, and the j-th coordinate of it.
$\nabla f, \nabla^2 f$	The ordinary gradient operator on $f(\cdot)$, and the second-order gradient.
$:=$	Assignment of a term.
$\leftarrow$	Assignment of quantities (usually in algorithmic representations).
$\triangleq$	Defined by ... .
$\mathbb{Z}^+$	The set of positive integers.
$p_n^0;\ p_n^k$	Initial density estimate in $\mathcal{M}$; the k-th iteration of `TMLE`.
p_n^*	The output (limiting distribution) of `TMLE`.
$\rightsquigarrow$	Converge to or approach to ... .
Σ	Empirical covariance matrix of D^*_Ψ.
$L_0(\nu)$	Space of measurable functions on Ω.
$\Vert p\Vert_{L^\infty}$	Essential supremum norm $\Vert p\Vert_{L^\infty} = \inf\{C : \vert p(o)\vert \leq C\ \nu\text{-a.e.}\}$.
$\Vert f\Vert_{L^2(\nu)}$	L^2-norm with respect to ν, $\Vert f\Vert_{L^2(\nu)} = \left(\int \vert f(o)\vert^2\, d\nu(o)\right)^{1/2}$.
$\Vert f\Vert_{L^1(\nu)}$	L^1-norm w.r.t. ν, $\Vert f\Vert_{L^1(\nu)} = \int \vert f(o)\vert\, d\nu(o)$.
$\setminus$	Make a difference set.
int$(\cdot)$	The interior of a set.
$\Vert\cdot\Vert_\infty$	Infinity norm $\Vert x\Vert_\infty = \max_i \vert x_i\vert$.
$\Vert\cdot\Vert_2, \Vert\cdot\Vert_{\mathbb{R}^d}$	Euclidean norm $\Vert x\Vert_2 = (\sum_i x_i^2)^{1/2}$.
$\Vert\cdot\Vert_1$	ℓ_1-norm $\Vert x\Vert_1 = \sum_i \vert x_i\vert$.
$\Vert\cdot\Vert_{\text{op}}$	Operator (spectral) norm of matrix, i.e. largest singular value.

A.1 Empirical Measure

In the rest of this section, we use P and p interchangeably to align with the notation in the main body.

Definition A.5 (Empirical Measure)**.** *The empirical measure $\mathbb{P}_n$ of $O_1, \dots, O_n$ is*

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{O_i \in A\}, \quad A \in \mathcal{F}, \tag{A.13}$$

and its Radon–Nikodym derivative $p_n = d\mathbb{P}_n/d\nu$ placing mass $1/n$ at each observed O_i.

A.2 Asymptotic Linearity and Regular Estimator

Let $\hat{\psi}_n$ be an estimator of the target parameter under n samples.

Definition A.6 (Asymptotic linearity)**.** *An estimator $\hat{\psi}_n$ is asymptotically linear if $\hat{\psi}_n - \psi_0 = \mathbb{P}_n D^*_{\hat{\psi}_n}(p_0) + o_{p_0}(1/\sqrt{n})$, where $D^*_{\hat{\psi}_n}(p_0) : \mathcal{O} \to \mathbb{R}^d$ is the corresponding* influence function *of the estimator $\hat{\psi}_n$ at p_0.*

Since we are working with the nonparametric model class, the influence function of the estimator in this case coincides with the influence function of the target parameter, $D^*_\Psi(p_0)$, for all regular and asymptotically linear estimators. At a high level, an estimator $\hat{\psi}_n$ is said to be *regular* at $P \in \mathcal{M}$ if it converges to $\Psi(p)$ *locally uniformally* in a neighborhood of P within $\mathcal{M}$ [29]. In other words, regularity requires the estimator to be stable under small perturbations of the underlying distribution.

A.3 Gateaux Derivative and Pathwise Differentiability

To formalize the notion of sensitivity, we first introduce the concept of *Gateaux derivative*–a generalization of the directional derivative to allow differentiation along directions in infinite-dimensional vector spaces.

Definition A.7 (Gateaux Derivative)**.** *For any $P, Q \in \mathcal{M}$, consider the ϵ-perturbed distribution for some small $\epsilon \in \mathbb{R}$:*

$$P_\epsilon := (1-\epsilon)P + \epsilon Q = P + \epsilon(Q - P), \tag{A.14}$$

with $P_\epsilon \in \mathcal{M}$. We say that a target parameter Ψ is Gateaux differentiable at P *if, for all $Q \in \mathcal{M}$, the following limit exists:*

$$\frac{d}{d\epsilon}\Psi(p_\epsilon)\bigg|_{\epsilon=0} = \lim_{\epsilon \to 0} \frac{\Psi(p_\epsilon) - \Psi(p)}{\epsilon},$$

and the resulting map $Q \mapsto \frac{d}{d\epsilon}\Psi(p_\epsilon)\Big|_{\epsilon=0}$ is linear and continuous in Q.

Definition A.8 (Pathwise Differentiability)**.** *A target parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ is* pathwise differentiable *if, for every regular one-dimensional parametric submodel $\{P_\epsilon\}_{\epsilon \in (-\delta, \delta)}$ for some small $\delta \in \mathbb{R}$, with $P_0 = P$ and $P_\epsilon \in \mathcal{M}$, it is* Gateaux differentiable.

Pathwise differentiability guarantees the construction of estimators that (1) converge at the optimal $\sqrt{n}$-rate, and (2) exhibit predictable changes in their asymptotic distribution under small perturbations of the true distribution P. These properties form the foundation for valid confidence interval construction in semi-parametric models.

A.4 Influence Function

When $\mathcal{M}$ is nonparametric, the IF of the target parameter Ψ at a distribution P is the unique Riesz representer of the Gateaux derivative (Definition A.7, Appendix A.3) for the $L_0^2(P)$ inner product.[10] We formalize this argument in Definition A.9.

[10] When $\mathcal{M}$ is semi-parametric, a target parameter may admit multiple influence functions, but the efficient influence function is unique.

Definition A.9 (Influence Function)**.** *The influence function $D^*_\Psi(p)(\cdot) : \mathcal{O} \to \mathbb{R}, D^*_\Psi(p) \in L^2_0(P)$, satisfies the property that for any $Q \in \mathcal{M}$ with corresponding density q, the Gateaux derivative of Ψ at P (Definition A.7) along the perturbation direction $q - p \in L^2_0(P)$ is given by*

$$\frac{d}{d\epsilon}\Psi(p_\epsilon)\bigg|_{\epsilon=0} = \langle D^*_\Psi(p), q - p\rangle_{L^2_0(P)} := \mathbb{E}_P\left[D^*_\Psi(p)(O) \cdot (q(O) - p(O))\right],$$

with P_ϵ defined in Eq. (A.14).

B Proof Preliminaries

Without loss of generality, we assume that each density $p \in \mathcal{M}$ lies in the Hilbert space $L^2(\nu)$, so that the norm

$$\|h\|_{L^2(\nu)} \triangleq \left(\int_\Omega h(o)^2 d\nu(o)\right)^{1/2} \tag{B.15}$$

is well-defined for any measurable function h.

Definition B.10 (Normalizing Constant)**.** *We define the two normalizing constants appearing in Eq.* (5) *and Eq.* (6) *by*

$$\begin{aligned} C\left(\epsilon, p_n^k\right) :=& \frac{1}{\displaystyle\int_\Omega \exp\left(\epsilon^\top D^*_\Psi(p_n^k)\right) p_n^k \, d\nu(o)}, \\ C'\left(\epsilon, p_n^k\right) :=& \frac{1}{\displaystyle\int_\Omega \frac{p_n^k}{1 + \exp\left(-2\epsilon^\top D^*_\Psi(p_n^k)\right)} d\nu(o)}, \end{aligned} \tag{B.16}$$

which ensure that the fluctuation submodels remain valid densities.

Definition B.11 (Restricted Domain with Counting Measure)**.** *Define the finite subset $\mathcal{O}' = \{O_1, O_2, \ldots, O_n\} \subseteq \mathcal{O}$. We assume that the measure ν restricted to $\mathcal{O}'$ is the counting measure, i.e., $\nu\left(\{O_i\}\right) = 1$ for $i = 1, \ldots, n$.*

With these conventions in place, all subsequent integrals over densities and influence functions may be interpreted either as Lebesgue integrals with respect to ν, or on the restricted domain $\mathcal{O}'$ as empirical averages.

B.1 Simulations for Theorem 1 & 5

Before we diving into the proofs, we first show some numerical observations of the `E1`, `E2` of Theorem 1, and the (i), (ii) of Theorem 5. Specifically, we consider a target parameter defined as the square of a cumulative distribution function $\Psi(p) := \int_0^1 F_p(o)^2 \, \mathrm{d}o$, where F_p denotes the cumulative distribution function of distribution p. We adopt the negative log-likelihood as our loss function. For numerical experiments, we select sample size $n = 5000$ and dimension $d = 20$, with each sample generated from a uniform probability distribution. All numerical simulations were implemented in Python and executed on a CPU-based cluster. The convergence metrics are computed using the L^2 distance between probabilities. Different random seeds were used for each run of our numerical experiments.

We present the TMLE learning curves in Table B.2, B.3, B.4 below, where each table corresponds to one of the three distinct submodels of the paper.

Table B.2: Submodel 1 (corresponding to (4))

Iter #	1	3	6	10	14	18	22	26	30	31	32	33
$\left\|\epsilon_n^k\right\|$	0.6318	0.4211	0.2783	0.1962	0.1248	0.0863	0.0567	0.0364	0.0228	0.0125	0.0053	0.00
Distance	0.0178	0.0132	0.0095	0.0067	0.0046	0.0031	0.0020	0.0013	0.0008	0.0004	0.0001	0.00

Table B.3: Submodel 2 (corresponding to (5))

Iter #	0	5	10	15	20	25	30	35	40	45	50	55
$\left\|\epsilon_n^k\right\|$	1.7704	0.8260	0.2034	0.0230	0.00245	0.000260	0.0000276	0.00000292	0.00000031	1.8×10^{-7}	0.00	0.00
Distance	0.08916	0.04462	0.00957	0.00104	0.000110	0.0000117	0.000000117	0.000000013	0.0000000014	8×10^{-9}	0.00	0.00

Table B.4: Submodel 3 (corresponding to (6))

Iter #	0	5	10	15	20	25	30	35	40	45	50	55
$\left\|\epsilon_n^k\right\|$	0.712	0.524	0.379	0.284	0.213	0.157	0.119	0.084	0.056	0.031	0.009	0.00
Distance	0.0205	0.0172	0.0136	0.0098	0.0081	0.0062	0.0048	0.0031	0.0020	0.0011	0.0004	0.00

The above numerical experiments provide some empirical support for sub-conclusions `E1` and `E2` in our Theorem 1, specifically demonstrating the equivalence of the two convergence metrics. For Theorem 5, our experimental setup largely mirrors the previous one.

We first verify the condition (i) within Theorem 5. For simplicity, we choose p to be a degenerate distribution. We then follow the `TMLE`algorithm and execute one step of the pseudocode, obtaining the resulting ϵ from the solver (from `SciPy`) of the subproblem. Note that for each coordinate

$$D^*_\Psi(P)(O_i) = 2\left(\int_{O_i}^1 F(o)\mathrm{d}o - \Psi(p)\right) = 2(1-1) = 0.$$

The output result was (always) a zero vector $\epsilon_n^1 = [0., 0., \ldots, 0.]$ (length 20). Precisely the same result holds under the setting described in Example 3 of our manuscript.

Condition (ii) allows non-degenerate densities provided the perturbation direction forms a conservative vector field and the loss satisfies a path-independent line-integral representation. A simple way to enforce this (in a finite sample) is to construct the EIF values so that they sum to zero exactly, ensuring the empirical estimating equation holds at $\epsilon = 0$. Consider a concrete example, where we first draw $m = 50$ independent samples from a $\mathrm{Beta}(2, 5)$ distribution and compute their EIF values $(D_1, \ldots, D_{50})$. Next we form a symmetric sample of size $n = 2m$ by appending the negatives of those values $D^*_\Psi = (D_1, \ldots, D_{50}, -D_1, \ldots, -D_{50})$. This creates a perfectly antisymmetric EIF vector, and the output result still remains a zero vector

$$\epsilon_n^1 = [0., 0., \ldots, 0.] \quad \text{(length 20)}.$$

These simulations therefore demonstrate that when conditions (i) and (ii) of Theorem 5 are satisfied, the `TMLE` update parameter is exactly zero and the algorithm does converge in a single step.

C Proof of Theorem 1

For clarity, we decompose the conclusion of Theorem 1 into four subordinate results, as illustrated in Figure Figure 5. We now proceed to establish the four sub-results (`E1`)-(`E4`) in turn.

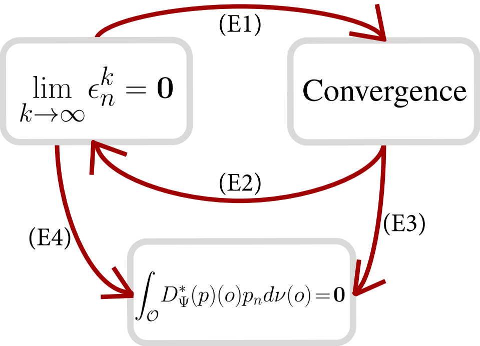

(E1)
$\lim_{k\to\infty} \epsilon_n^k = \mathbf{0}$
Convergence
(E2)
(E4)
(E3)
$\int_{\mathcal{O}} D^*_\Psi(p)(o)p_n d\nu(o) = \mathbf{0}$

Figure 5: A relationship diagram of the subresults in Theorem 1.

C.1 Proof of sufficiency (E1)

Proof. Consider an arbitrary $p \in \mathcal{M}$ and fluctuation parameter $\epsilon \in \mathcal{R}$, and expand the associated parametric submodel as

$$\begin{aligned} p(\epsilon) &= p(\mathbf{0}) + \frac{\partial}{\partial \epsilon} p(\epsilon)\Big|_{\epsilon=\mathbf{0}} \cdot (\epsilon - \mathbf{0}) + \int_0^1 (1-t)\,\mathcal{D}_f^2 p(t\epsilon)\Big[\frac{\partial}{\partial \epsilon} p(\epsilon)\Big|_{\epsilon=\mathbf{0}} \epsilon,\, \frac{\partial}{\partial \epsilon} p(\epsilon)\Big|_{\epsilon=\mathbf{0}} \epsilon\Big]\, dt \\ &= p + \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon + \int_0^1 (1-t)\,\mathcal{D}_f^2 p(t\epsilon)\Big[\frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon,\, \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon\Big]\, dt. \end{aligned} \tag{C.17}$$

Applying the TMLE update rule yields

$$\begin{aligned} &\left\|p_n^{k+1} - p_n^k\right\|_1 \\ &= \left\| \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k + \int_0^1 (1-t)\,\mathcal{D}_f^2 p_n^k(t\epsilon)\Big[\frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k,\, \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k\Big]\, dt \right\|_1 \\ &\le \left\| \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k \right\| + \left\| \int_0^1 (1-t)\,\mathcal{D}_f^2 p_n^k(t\epsilon)\Big[\frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k,\, \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k\Big]\, dt \right\| \\ &\le \sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \cdot \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} + \int_0^1 (1-t) \left\| \mathcal{D}_f^2 p_n^k(t\epsilon)\Big[\frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k,\, \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k\Big] \right\|_1 dt \\ &\le \sup_{p\in\mathcal{M}} \sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} + \int_0^1 (1-t) \cdot \left\| \mathcal{D}_f^2 p_n^k(t\epsilon) \right\|_{\text{op}} \cdot \left\| \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k \right\|_1^2 dt \\ &\le \sup_{p\in\mathcal{M}} \sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} + \left(\int_0^1 (1-t)^2 dt\right)^{\frac{1}{2}} \left(\int_0^1 \left(\left\| \mathcal{D}_f^2 p_n^k(t\epsilon) \right\|_{\text{op}} \cdot \left\| \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k \right\|_1^2 \right)^2 dt \right)^{\frac{1}{2}}. \end{aligned} \tag{C.18}$$

Introduce a closed Euclidean ball with positive radius $\bar{\alpha}$, $\mathbb{B}_{\bar{\alpha}}(0) := \left\{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| \le \bar{\alpha}\right\}$, and for every density $p \in \mathcal{M}$ we consider the parametric lift $\mathbf{v} \longmapsto p(\epsilon = \mathbf{v})$, $\mathbf{v} \in \mathbb{R}^d$. Construct the metric ball defined by $\mathcal{B}_{\bar{\alpha}}(p) := \{p(\mathbf{v}) : \mathbf{v} \in \mathbb{B}_{\bar{\alpha}}(0)\}$, where $\bar{\alpha}$ is chosen such that the local path $\mathbf{v} \mapsto p(\mathbf{v})$ does not leave the statistical model $\mathcal{M}$ and maintain uniform boundedness of all second-Fréchet derivatives $\mathcal{D}_f^2 p(\mathbf{v})$ over the ball $\mathcal{B}_{\bar{\alpha}}$. Using uniform boundedness of $\mathcal{D}_f^2 p$ within $\mathcal{B}_{\bar{\alpha}}$, it follows that

$$\begin{aligned} &\left\|p_n^{k+1} - p_n^k\right\|_1 \\ &\le \sup_{p\in\mathcal{M}} \sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} + \frac{\sqrt{3}}{3} \left(\int_0^1 \left(\sup_{\mathbf{v}\in\mathbb{B}_{\bar{\alpha}}} \left\| \mathcal{D}_f^2 p_n^k(\mathbf{v}) \right\|_{\text{op}} \cdot \left\| \frac{\partial p_n^k(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k \right\|_1^2 \right)^2 dt \right)^{\frac{1}{2}} \\ &\le \sup_{p\in\mathcal{M}} \sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \\ &\qquad + \frac{\sqrt{3}}{3} \left(\int_0^1 \left(\sup_{p\in\mathcal{M}} \sup_{\|\mathbf{v}\|\le\bar{\alpha}} \sup_{\substack{\mathbf{u},\mathbf{w}\in\mathbb{R}^d \\ \|\mathbf{u}\|=\|\mathbf{w}\|=1}} \left\| \mathcal{D}_f^2 p(\mathbf{v})[\mathbf{u},\mathbf{w}] \right\|_1 \cdot \left(\sup_{p\in\mathcal{M}} \sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \right)^2 \right)^2 dt \right)^{\frac{1}{2}} \\ &= \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \cdot \sup_{p\in\mathcal{M}} \Bigg[\sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \\ &\qquad + \frac{\sqrt{3}}{3} \left(\sup_{\mathbf{v}\in\mathcal{R}, \|\mathbf{v}\|\le 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| \right)^2 \left(\sup_{\|\mathbf{v}\|\le\bar{\alpha}} \sup_{\substack{\mathbf{u},\mathbf{w}\in\mathbb{R}^d \\ \|\mathbf{u}\|=\|\mathbf{w}\|=1}} \left\| \mathcal{D}_f^2 p(\mathbf{v})[\mathbf{u},\mathbf{w}] \right\|_1 \right) \cdot \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \Bigg] \\ &\triangleq \blacksquare. \end{aligned} \tag{C.19}$$

As $\epsilon_n^k \rightsquigarrow \mathbf{0}$, the subsequent asymptotic limit holds

$$\lim_{k\to\infty} \left\| p_n^{k+1} - p_n^k \right\|_1 \leq \lim_{k\to\infty} \blacksquare = 0 \cdot \sup_{p\in\mathcal{M}} \left[\sup_{\mathbf{v}\in\mathcal{R},\|\mathbf{v}\|\leq 1} \left\| \frac{\partial p(\epsilon)}{\partial \epsilon}\bigg|_{\epsilon=\mathbf{0}} \mathbf{v} \right\| + 0 \right] = 0. \tag{C.20}$$

Given summability $\sum_k \left\|\epsilon_n^k\right\|^2 < \infty$, invoking the Cauchy–Schwarz inequality provides

$$\begin{aligned} \sum_{k=0}^{\infty} \left\| p_n^{k+1} - p_n^k \right\|_1 &\lesssim \sum_{k=0}^{\infty} \|\epsilon_n^k\|_{\mathbb{R}^d} + \sum_{k=0}^{\infty} \|\epsilon_n^k\|_{\mathbb{R}^d}^2 \\ &\leq \left(\sum_{k=0}^{\infty} 1^2\right)^{1/2} \left(\sum_{k=0}^{\infty} \|\epsilon_n^k\|_{\mathbb{R}^d}^2\right)^{1/2} + \sum_{k=0}^{\infty} \|\epsilon_n^k\|_{\mathbb{R}^d}^2 < +\infty. \end{aligned} \tag{C.21}$$

Select an arbitrary positive constant $\delta > 0$. As the series $\sum_{k=0}^{\infty} \left(\|\epsilon_n^k\| + \|\epsilon_n^k\|^2 \right)$ converges, its tail must necessarily vanish. Hence there exists an index N such that for all $r \geq N$, the bound $\sum_{k=r}^{\infty} \left(\|\epsilon_n^k\| + \|\epsilon_n^k\|^2 \right) \lesssim \delta$ is satisfied. For any $m > n' \geq N$, apply the telescoping identity to deduce

$$\begin{aligned} \left\| p_n^m - p_n^{n'} \right\|_1 &= \left\| \left(p_n^m - p_n^{m-1}\right) + \left(p_n^{m-1} - p_n^{m-2}\right) + \cdots + \left(p_n^{n'+1} - p_n^{n'}\right) \right\|_1 \\ &\leq \sum_{k=n'}^{m-1} \left\| p_n^{k+1} - p_n^k \right\|_1 \\ &\lesssim \sum_{k=n'}^{m-1} \left(\|\epsilon_n^k\|_{\mathbb{R}^d} + \|\epsilon_n^k\|_{\mathbb{R}^d}^2 \right) \\ &\leq \sum_{k=n'}^{\infty} \left(\|\epsilon_n^k\|_{\mathbb{R}^d} + \|\epsilon_n^k\|_{\mathbb{R}^d}^2 \right) \lesssim \delta. \end{aligned} \tag{C.22}$$

Thus, for every $\delta > 0$, one identifies an index N ensuring $\left\| p_n^m - p_n^{n'} \right\|_1 \lesssim \delta$ whenever $m > n' \geq N$. Consequently, it holds that $\lim\limits_{n,m\to\infty} \sum\limits_{k=n}^{m-1} \|p_n^{k+1} - p_n^k\|_1 = 0$, which verifies that the sequence $\{p_n^k\}_k$ is Cauchy in $(\mathcal{M}, \|\cdot\|_1)$. By invoking Burke [30] the convergence of algorithmic sequence is thereby guaranteed. □

Remark 4. *While the result drawn in (C.20) may appear promising, it does not necessarily guarantee the desired convergence. We provide a simple counterexample as follows. Choose the flat baseline* $p_n^0(o) = 1$ *defined over* $0 \leq o \leq 1$*, thereby satisfying* $\int_0^1 p_n^0 = 1$ *and* $\left\|p_n^0\right\|_1 = 1$*. Define the perturbation* $h(o) := 4o - 2$ $(0 \leq o \leq 1)$*, resulting in* $\|h\|_1 = \int_0^1 |4o-2| do = 1$ *and* $\int_0^1 h(o) do = 0$*. Set the sequence* $\epsilon_n^k = \frac{1}{k+1}$ *for all* $k \geq 0$*, and define recursively*

$$p_n^{k+1} \leftarrow p_n^k + \frac{1}{k+1} h(o). \tag{C.23}$$

So that by induction

$$p_n^k = p_n^0 + \sum_{j=0}^{k-1} \frac{1}{j+1} h(o), \quad k \geq 1, \tag{C.24}$$

where each iterate p_n^k *retains total mass* 1 *since* $\int h = 0$*. Summing (C.23) leads to*

$$\sum_{k=0}^{\infty} \|p_n^{k+1} - p_n^k\|_1 = \sum_{k=0}^{\infty} \left\| \frac{1}{k+1} h \right\|_1 = \sum_{k=0}^{\infty} \frac{1}{k+1} \rightsquigarrow \infty. \tag{C.25}$$

As $k \rightsquigarrow \infty$ *we have both* $\epsilon_n^k \rightsquigarrow 0$ *and* $\|p_n^{k+1} - p_n^k\|_1 \rightsquigarrow 0$*, but* $\sum_{k=0}^{\infty} \epsilon_n^k = +\infty$ *(i.e., partial sum of the harmonic series). As a result the convergence of iterates never hold. Note that if* p_n^k *crosses* $[0,1]$ *at some* k *we may simply project back onto the set of densities, e.g.,* $p_n^k := \frac{\max\left(p_n^k, 0\right)}{\int_0^1 \max(p_n^k, 0)}$*. Such a projection modifies* p_n^k *by at most a factor proportional to* H_k^{-1} *where* H_k *is the* k*-th harmonic number. Therefore it does not affect the lower bound established above.*

C.2 Proof of necessity (E2)

Proof. Since ϵ_n^k is defined as the minimizer of empirical risk functional $\int_{\mathcal{O}} \mathbf{L}(p_n^k(\cdot))(o) p_n d\nu(o)$, it necessarily satisfies $\int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^k))(o) p_n d\nu(o) \leq \int_{\mathcal{O}} \mathbf{L}(p_n^k(\mathbf{0}))(o) p_n d\nu(o)$. For any fixed iteration index k and all admissible ϵ, there exists an intermediate $\bar{\epsilon}$ on the line segment joining $\mathbf{0}$ and ϵ_n^k such that

$$\begin{aligned}
&\int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^k))(o) p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}(p_n^k(\mathbf{0}))(o) p_n d\nu(o) \\
&= \int_{\mathcal{O}} \mathbf{L}(p_n^k(\mathbf{0}))(o) p_n d\nu(o) + \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}^{\top} \epsilon_n^k + \frac{1}{2}\left(\epsilon_n^k\right)^{\top} \int_{\mathcal{O}} \nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k \\
&\quad + \frac{1}{6}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{l=1}^{d} \frac{\partial^3 \int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} \epsilon_i \epsilon_j \epsilon_l \Big|_{\epsilon=\bar{\epsilon}} - \int_{\mathcal{O}} \mathbf{L}(p_n^k(\mathbf{0}))(o) p_n d\nu(o) \\
&= 0 + \frac{1}{2}\left(\epsilon_n^k\right)^{\top} \int_{\mathcal{O}} \nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k + \frac{1}{6}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{l=1}^{d} \int_{\mathcal{O}} \frac{\partial^3 \mathbf{L}(p_n^k(\epsilon))(o) p_n}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} d\nu(o)\Big|_{\epsilon=\bar{\epsilon}} \epsilon_i \epsilon_j \epsilon_l \\
&\quad + \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}^{\top} \epsilon_n^k \\
&\leq 0,
\end{aligned} \tag{C.26}$$

which subsequently leads to

$$\begin{aligned}
&\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}^{\top} \epsilon_n^k \\
&\leq -\frac{1}{2}\left(\epsilon_n^k\right)^{\top} \int_{\mathcal{O}} \nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}} \epsilon_n^k - \frac{1}{6}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{l=1}^{d} \int_{\mathcal{O}} \frac{\partial^3 \mathbf{L}(p_n^k(\epsilon))(o) p_n}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} d\nu(o)\Big|_{\epsilon=\bar{\epsilon}} \epsilon_i \epsilon_j \epsilon_l \\
&\leq \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \left(\frac{1}{2}\left\|\int_{\mathcal{O}} \nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\| + \frac{1}{6}\left|\left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{l=1}^{d} \int_{\mathcal{O}} \frac{\partial^3 \mathbf{L}(p_n^k(\epsilon))(o) p_n}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} d\nu(o)\Big|_{\epsilon=\bar{\epsilon}}\right|\right) \\
&\leq \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \left(\frac{1}{2\log(n+1)}\left\|\int_{\mathcal{O}'} \nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o)\big|_{\epsilon=\mathbf{0}} d\nu(o)\right\| + \frac{1}{6}\left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{l=1}^{d} \left|\int_{\mathcal{O}} \frac{\partial^3 \mathbf{L}(p_n^k(\epsilon))(o) p_n}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} d\nu(o)\Big|_{\epsilon=\bar{\epsilon}}\right|\right) \\
&\leq \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \left(\frac{1}{2\log(n+1)}\int_{\mathcal{O}'} \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o)\big|_{\epsilon=\mathbf{0}}\right\| d\nu(o) + \frac{d}{6}\left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \sum_{i=1}^{d}\sum_{j=1}^{d} \max_{l\in[1,d]^1\cap\mathbb{Z}} \int_{\mathcal{O}} p_n \left|\frac{\partial^3 \mathbf{L}(p_n^k(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l}\Big|_{\epsilon=\bar{\epsilon}}\right| d\nu(o)\right) \\
&\leq \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \left(\frac{1}{2\log(n+1)} \sup_{\left\{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \int_{\mathcal{O}'} \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\| d\nu(o)\right. \\
&\qquad\qquad \left. + \frac{d^2}{6}\left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \sum_{i=1}^{d} \max_{(j,l)\in[1,d]^2\cap\mathbb{Z}^2} \int_{\mathcal{O}} p_n \left|\frac{\partial^3 \mathbf{L}(p_n^k(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l}\Big|_{\epsilon=\bar{\epsilon}}\right| d\nu(o)\right) \\
&\leq \frac{\sup_{\left\{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \int_{\mathcal{O}'} \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\| d\nu(o)}{2\log(n+1)} \cdot \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \\
&\qquad + \frac{d^3}{6\log(n+1)} \sup_{\left\{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \left|\frac{\partial^3 \mathbf{L}(p_n^{k'}(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l}\right| d\nu(o) \cdot \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^3 .
\end{aligned}$$

$$
\begin{aligned}
\leq \sup_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+} \Bigg\{ & \frac{\int_{\mathcal{O}'} \sup_{\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\}} \left\| \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o) \right\| d\nu(o)}{2\log(n+1)} \cdot \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}^2 \\
& + \frac{d^3}{6\log(n+1)} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\{\epsilon : p_n^{k''}(\epsilon)\in\mathcal{M}\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l} \right| d\nu(o) \cdot \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}^3 \Bigg\}.
\end{aligned}
\tag{C.27}
$$

Observe that the direction of update ϵ_n^k satisfies the relation

$$
\begin{aligned}
& \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}}^\top \epsilon_n^k \\
= & \frac{\left\langle \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}},\ \epsilon_n^k \right\rangle}{\left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}} \cdot \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \cdot \left\| \epsilon_n^k \right\|_{\mathbb{R}^d} \\
\overset{(a)}{\geq} & \cos\left(\sup_{k'\geq 0} \arccos\left(\min\left\{ \max\left\{ -1, \left\langle \frac{\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}}}{\left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\| \left\| \epsilon_n^{k'} \right\|},\ \epsilon_n^{k'} \right\rangle \right\}, 1 \right\} \right) \right) \\
& \cdot \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\| \epsilon_n^k \right\|_{\mathbb{R}^d} \\
\triangleq & \cos\left(\sup_{k'\geq 0} \mathcal{C}\Big(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon)) \Big) \right) \cdot \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\| \epsilon_n^k \right\|_{\mathbb{R}^d},
\end{aligned}
\tag{C.28}
$$

where step (a) follows from the strict monotonicity of the cosine function on interval $[0, \pi]$. By synthesizing this observation with the result established in (C.27), it follows that

$$
\begin{aligned}
& \cos\left(\sup_{k'\geq 0} \mathcal{C}\Big(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon)) \Big) \right) \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\| \epsilon_n^k \right\|_{\mathbb{R}^d} \\
\leq & \sup_{k'\geq 0} \frac{\int_{\mathcal{O}'} \sup_{\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\}} \left\| \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o) \right\| d\nu(o)}{2\log(n+1)} \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}^2 \\
& + \frac{d^3}{6\log(n+1)} \sup_{k'\geq 0} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k'}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l} \right| d\nu(o) \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}^3.
\end{aligned}
\tag{C.29}
$$

Rearranging the relevant expression yields an inequality involving $\left\| \epsilon_n^k \right\|$ as

$$
\begin{aligned}
& \sup_{k'\geq 0} \frac{\int_{\mathcal{O}'} \sup_{\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\}} \left\| \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o) \right\| d\nu(o)}{2\log(n+1)} \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}^2 \\
& + \frac{d^3}{6\log(n+1)} \sup_{k'\geq 0} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k'}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l} \right| d\nu(o) \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}^3 \\
& - \cos\left(\sup_{k'\geq 0} \mathcal{C}\Big(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon)) \Big) \right) \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\| \epsilon_n^k \right\|_{\mathbb{R}^d}
\end{aligned}
$$

$$
\begin{aligned}
&= \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \Bigg(\frac{d^3}{6\log(n+1)} \sup_{k'\geq 0} \int_{\mathcal{O}'} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \sup_{\left\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k'}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right| d\nu(o) \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \\
&\quad + \sup_{k'\geq 0} \frac{\int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\| d\nu(o)}{2\log(n+1)} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \\
&\quad - \cos\left(\sup_{k'\geq 0} \mathcal{C}\Big(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))\Big)\right) \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d} \Bigg) \\
&= \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \Bigg[\frac{d^3}{6\log(n+1)} \sup_{k''\geq 0} \int_{\mathcal{O}'} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \sup_{\left\{\epsilon : p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right| d\nu(o) \Bigg(\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \\
&\quad + \left(\frac{3 \sup\limits_{k'\geq 0} \frac{\int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\| d\nu(o)}{\log(n+1)}}{\frac{2d^3}{\log(n+1)} \sup\limits_{k''\geq 0} \max\limits_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right| d\nu(o)} \right)^2 \\
&\quad + \sup_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+} \int_{\mathcal{O}'} \frac{3\frac{\int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\| d\nu(o)}{\log(n+1)}}{\frac{d^3}{\log(n+1)} \max\limits_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|} d\nu(o) \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \\
&\quad - \frac{9}{4d^6} \sup_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+} \frac{\left(\int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\| d\nu(o)\right)^2}{\max\limits_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \left(\int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right| d\nu(o)\right)^2} \Bigg) \\
&\quad - \cos\left(\sup_{k'\geq 0} \mathcal{C}\Big(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))\Big)\right) \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d} \Bigg] \\
&\overset{(a)}{=} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \Bigg[\frac{d^3}{6\log(n+1)} \sup_{k''\geq 0} \int_{\mathcal{O}'} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \sup_{\left\{\epsilon : p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right| d\nu(o) \Bigg(\left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \\
&\quad + \sup_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+} \int_{\mathcal{O}'} \frac{3 \sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \frac{\left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|}{\log(n+1)}}{\frac{2d^3}{\log(n+1)} \int_{\mathcal{O}'} \max\limits_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right| d\nu(o)} d\nu(o) \Bigg)^2 \\
&\quad - \Bigg(\frac{3}{8d^3\log(n+1)} \sup_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+} \frac{\int_{\mathcal{O}'}\int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \left\{ \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\right\| \left\|\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o')\right\| \right\} d\nu(o)d\nu(o')}{\max\limits_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left|\frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right| d\nu(o)} \\
&\quad + \cos\left(\sup_{k'\geq 0} \mathcal{C}\Big(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))\Big)\right) \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d} \Bigg) \Bigg] \\
&\overset{(b)}{\geq} 0,
\end{aligned}
\tag{C.30}
$$

where step (a) utilizes properties of the empirical distribution, while step (b) follows directly from (C.29). Solving the resulting cubic inequality leads to the solution

$$\begin{aligned}
&\underbrace{\left(\frac{d^3}{3\log(n+1)}\sup_{k''\geq 0}\int_{\mathcal{O}'}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\right)}_{\clubsuit_L}\cdot\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}\\
&\geq\Bigg[\sup_{k'\geq 0}\int_{\mathcal{O}'}\int_{\mathcal{O}'}\frac{\sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}}\left\{\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\right\|\right\}}{4\log^2(n+1)}d\nu(o)d\nu(o')\\
&\quad+\frac{2d^3\cos\left(\sup_{k'\geq 0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)}{3\log(n+1)}\sup_{k''\geq 0}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\cdot\\
&\quad\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d}\Bigg]^{1/2}-\sup_{k'\geq 0}\frac{\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}}\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|d\nu(o)}{2\log(n+1)}\\
&\triangleq\clubsuit_R.
\end{aligned}\tag{C.31}$$

Now consider the gradient of empirical risk $\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)$ evaluated near ϵ_n^k. For some $\tilde{\epsilon}^k$ on the line segment connecting $\mathbf{0}$ and ϵ_n^k, we can have

$$\begin{aligned}
&\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\mathbf{0}}\\
&=\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\epsilon_n^k}-\int_{\mathcal{O}}\nabla_\epsilon^2\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\epsilon_n^k}\left(\epsilon_n^k-\mathbf{0}\right)\\
&\quad+\frac{1}{2}\int_{\mathcal{O}}\nabla_\epsilon^3\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\tilde{\epsilon}^k}\left(\epsilon_n^k-\mathbf{0}\right)^{\otimes 2}\\
&=-\int_{\mathcal{O}}\nabla_\epsilon^2\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\epsilon_n^k}\epsilon_n^k+\frac{1}{2}\int_{\mathcal{O}}\nabla_\epsilon^3\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\tilde{\epsilon}^k}\left(\epsilon_n^k\right)^{\otimes 2},
\end{aligned}\tag{C.32}$$

where $\otimes$ denotes the tensor product and in particular $(\epsilon)^{\otimes 2}:=\epsilon\otimes\epsilon$. Taking the inner product with $-\epsilon_n^k$ on both sides of the expression yields

$$\begin{aligned}
&\left\langle\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\mathbf{0}},\ -\epsilon_n^k\right\rangle\\
&=(\epsilon_n^k)^\top\int_{\mathcal{O}}\nabla_\epsilon^2\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\epsilon_n^k}\epsilon_n^k-\frac{1}{2}\left[\int_{\mathcal{O}}\nabla_\epsilon^3\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\tilde{\epsilon}^k}\left(\epsilon_n^k\right)^{\otimes 2}\right]^\top\epsilon_n^k\\
&\geq\left(\frac{\inf_{\{\epsilon:p_n^k(\epsilon)\in\mathcal{M}\}}(\epsilon_n^k)^\top\int_{\mathcal{O}}\nabla_\epsilon^2\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\epsilon_n^k}{\|\epsilon_n^k\|_{\mathbb{R}^d}^3}-\frac{1}{2}\left\|\int_{\mathcal{O}}\nabla_\epsilon^3\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\tilde{\epsilon}^k}\right\|\right)\|\epsilon_n^k\|_{\mathbb{R}^d}^3\\
&\geq\left(\frac{\inf\limits_{\{\epsilon:p_n^k(\epsilon)\in\mathcal{M}\setminus\{p_n^k\}\}}\int_{\mathcal{O}}\frac{\epsilon^\top\nabla_\epsilon^2\mathbf{L}(p_n^k(\epsilon))(o)p_n\epsilon}{\|\epsilon\|^2}d\nu(o)}{\|\epsilon_n^k\|_{\mathbb{R}^d}}-\frac{d^3}{2}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}}\int_{\mathcal{O}}\left|\frac{\partial^3\mathbf{L}(p_n^{k'}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}p_n\right|d\nu(o)\right)\|\epsilon_n^k\|_{\mathbb{R}^d}^3
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \left(\frac{\log(1+\frac{1}{n})}{\|\epsilon_n^k\|_{\mathbb{R}^d}} \inf_{\{\epsilon : p_n^k(\epsilon) \in \mathcal{M} \setminus \{p_n^k\}\}} \int_{\mathcal{O}'} \frac{\epsilon^\top \nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o)\epsilon}{\|\epsilon\|^2} d\nu(o) \right.\\
&\qquad \left. - \frac{d^3}{2\log(n+1)} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\left\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\right\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k'}(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} \right| d\nu(o) \right) \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^3 \\
&\geq \inf_{k'\geq 0} \sup_{k''\geq 0} \left\{ \log\left(1+\frac{1}{n}\right) \int_{\mathcal{O}'} \inf_{\left\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}} \epsilon^\top \frac{\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2} d\nu(o) \cdot \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \right.\\
&\qquad\qquad \left. - \frac{d^3}{2\log(n+1)} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\left\{\epsilon : p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} \right| d\nu(o) \cdot \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^3 \right\},
\end{aligned}
$$

with step (a) justified under the natural assumption $n \geq 1$. Analogous to the reasoning used in (C.28), the structure of inner product in (C.33) implies that

$$
\begin{aligned}
&\left\langle \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} ,\ \epsilon_n^k \right\rangle \\
&\geq \min_{k'\geq 0} \frac{ \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}}^\top \epsilon_n^{k'} }{ \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \|\epsilon_n^{k'}\|_{\mathbb{R}^d} } \cdot \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \\
&\geq \cos\left(\max_{k'\geq 0} \mathcal{C}\left(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))\right) \right) \cdot \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} .
\end{aligned}
\tag{C.34}
$$

Rewriting the expression leads to an inequality involving the $\left\|\epsilon_n^k\right\|$ as

$$
\begin{aligned}
&\cos\left(\pi - \max_{k'\geq 0} \mathcal{C}\left(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))\right) \right) \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \\
&\quad - \log\left(1+\frac{1}{n}\right) \inf_{k'\geq 0} \int_{\mathcal{O}'} \inf_{\left\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}} \epsilon^\top \frac{\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 d\nu(o) \\
&\quad + \frac{d^3}{2\log(n+1)} \sup_{k''\geq 0} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\left\{\epsilon : p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} \right| \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^3 d\nu(o) \\
= \ & \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \left(\frac{d^3}{2\log(n+1)} \sup_{k''\geq 0} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\left\{\epsilon : p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} \right| \left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 d\nu(o) \right.\\
&\quad - \log\left(1+\frac{1}{n}\right) \inf_{k'\geq 0} \int_{\mathcal{O}'} \inf_{\left\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}} \epsilon^\top \frac{\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} d\nu(o) \\
&\quad \left. + \cos\left(\pi - \max_{k'\geq 0} \mathcal{C}\left(\epsilon_n^{k'}, \nabla_\epsilon \mathbf{L}(p_n^{k'}(\epsilon))\right) \right) \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \right) \\
= \ & \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \left[\frac{3}{2} \clubsuit_L \left(\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 - \frac{\log^2(n+1) - \log(n)\log(n+1)}{d^3/2} \inf_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+} \int_{\mathcal{O}'} \inf_{\left\{\epsilon : p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}} \epsilon^\top \frac{\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2} \right.\right.
\end{aligned}
$$

$$\cdot \Bigg(\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\Bigg)^{-1}d\nu(o)\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}$$

$$+\left(\frac{\log\left(1+\frac{1}{n}\right)\inf_{k'\geq0}\int_{\mathcal{O}'}\inf_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}}\epsilon^\top\frac{\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)}{\|\epsilon\|^2}\epsilon d\nu(o)}{\frac{d^3}{\log(n+1)}\sup_{k''\geq0}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)}\right)^2\Bigg)$$

$$+\cos\left(\pi-\max_{k'\geq0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d}-\frac{d^3}{2\log(n+1)}$$

$$\cdot\sup_{k''\geq0}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\cdot\log^2\left(1+\frac{1}{n}\right)\cdot\Bigg(\inf_{k'\geq0}\int_{\mathcal{O}'}\inf_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}}\epsilon^\top$$

$$\frac{\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2}d\nu(o)\Bigg)^2\cdot\frac{\log^2(n+1)}{d^6}\left(\sup_{k''\geq0}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\right)^{-2}\Bigg]$$

$$\overset{(a)}{=}\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}\Bigg[\frac{3}{2}\clubsuit_L\left(\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}-\inf_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+}\int_{\mathcal{O}'}\frac{\frac{[\log(n+1)-\log(n)]\log(n+1)}{d^3}\inf_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}}\epsilon^\top\frac{\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)}{\|\epsilon\|^2}\epsilon}{\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)}d\nu(o)\right)^2$$

$$+\Bigg(\cos\left(\pi-\max_{k'\geq0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d}-\frac{\log(n+1)\left[\log(n+1)-\log(n)\right]^2}{2d^3}$$

$$\cdot\inf_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+}\int_{\mathcal{O}'}\int_{\mathcal{O}'}\frac{\inf_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}}\left\{\frac{\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon\epsilon^\top\right)}{\|\epsilon\|^2}\frac{\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\epsilon\epsilon^\top\right)}{\|\epsilon\|^2}\right\}}{\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)}d\nu(o)d\nu(o')\Bigg)\Bigg]$$

$$\geq 0, \tag{C.35}$$

where step (a) exploits the standard identity for the trace of a quadratic form. Given that the coefficient $\frac{3}{2}\clubsuit_L$ is non-negative, the resulting quadratic implicitly involved in (C.35) is convex and opens upward. Furthermore, since $\cos(\pi-\max_{k'\geq0}\mathcal{C}(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))))\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_nd\nu(o)\big|_{\epsilon=\mathbf{0}}\right\|>0$ for all $k\neq\infty$, the quadratic evaluates to a positive value at $\|\epsilon_n^k\|_{\mathbb{R}^d}=0$. This ensures that the minimizer ϵ_n^k must be sufficiently small under a valid fluctuation update. Consequently, $\left\|\epsilon_n^k\right\|$ is constrained to lie within the region where the quadratic remains nonnegative, i.e.,

$$3\clubsuit_L\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}$$

$$\leq\log\left(1+\frac{1}{n}\right)\inf_{k'\geq0}\int_{\mathcal{O}'}\inf_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}}\epsilon^\top\frac{\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2}d\nu(o)-\Bigg[\inf_{k'\geq0}\log^2\left(1+\frac{1}{n}\right)\int_{\mathcal{O}'}\int_{\mathcal{O}'}\inf_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}}$$

$$\left\{\frac{\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon\epsilon^\top\right)\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\epsilon\epsilon^\top\right)}{\|\epsilon\|^4}\right\}d\nu(o)d\nu(o')-\frac{2d^3}{\log(n+1)}\cos\left(\pi-\max_{k'\geq0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)$$

$$\left.\sup_{k''\geq 0}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\left\|\int_{\mathcal{O}}\nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\right]^{1/2}$$
$$\triangleq \clubsuit_{R'}. \tag{C.36}$$

Combining inequalities (C.31) and (C.36), it follows that

$$\max\left\{\frac{\clubsuit_R}{\clubsuit_L},\, 0\right\}\leq\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}\leq\frac{\clubsuit_{R'}}{3\clubsuit_L}. \tag{C.37}$$

We now establish that the limit of $\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}$ as $k\rightsquigarrow\infty$ must be $\mathbf{0}$. To this end, suppose for the sake of contradiction that the limiting gradient does not vanish, i.e., $\lim_{k\to\infty}\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}\neq\mathbf{0}$. By continuity of the gradient mapping with respect to k, there exists an index K such that for all $k\geq K$, the gradient $\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}$ remains bounded away from zero and uniformly close to its limiting value $\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^\infty(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}$. For each such k, define the descent direction $\eta:=-\frac{\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}}{\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}\right\|}$. Then by the directional derivative definition and smoothness of the loss, for sufficiently small $\delta>0$ we obtain

$$\begin{aligned}
&\int_{\mathcal{O}}\mathbf{L}(p_n^k(\delta\eta))(o)p_n d\nu(o)\\
=&\int_{\mathcal{O}}\mathbf{L}(p_n^k(\mathbf{0}))(o)p_n d\nu(o)+\delta\,\nabla_\epsilon\int_{\mathcal{O}}\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}^{\top}\eta+O(\delta)\\
=&\int_{\mathcal{O}}\mathbf{L}(p_n^k(\mathbf{0}))(o)p_n d\nu(o)-\delta\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|+O(\delta),
\end{aligned} \tag{C.38}$$

which is strictly less than $\int_{\mathcal{O}}\mathbf{L}(p_n^k(\mathbf{0}))(o)p_n d\nu(o)$ for small enough δ. Since ϵ_n^k is chosen to minimize $\int_{\mathcal{O}}\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)$ over all feasible ϵ, it follows that $\int_{\mathcal{O}}\mathbf{L}(p_n^k(\epsilon_n^k))(o)p_n d\nu(o)\leq\int_{\mathcal{O}}\mathbf{L}(p_n^k(\delta\eta))(o)p_n d\nu(o)<\int_{\mathcal{O}}\mathbf{L}(p_n^k(\mathbf{0}))(o)p_n d\nu(o)$, so the empirical risk at iteration k strictly decreases whenever $\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}\neq\mathbf{0}$. This implies that for infinitely many large k, the algorithm continues to reduce the risk, contradicting the hypothesis that $\{p_n^k\}$ converges to a stable limit p_n^∞. Therefore, the limit gradient $\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^\infty(\epsilon))(o)p_n d\nu(o)\big|_{\epsilon=\mathbf{0}}$ must be zero. Next, examining the left-hand side of (C.37), it can be observed that

$$\begin{aligned}
&\lim_{k\to\infty}\frac{\clubsuit_R}{\clubsuit_L}\\
=&\lim_{k\to\infty}\left\{\left[\sup_{k'\geq 0}\int_{\mathcal{O}'}\int_{\mathcal{O}'}\frac{\sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}}\left\{\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\right\|\right\}}{4\clubsuit_L^2\log^2(n+1)}d\nu(o)d\nu(o')\right.\right.\\
&+\frac{2d^3\cos\left(\sup_{k'\geq 0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)}{3\clubsuit_L^2\log(n+1)}\sup_{k''\geq 0}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\right\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\cdot\\
&\left.\left.\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d}\right]^{1/2}-\sup_{k'\geq 0}\frac{\int_{\mathcal{O}'}\sup_{\left\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\right\}}\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|d\nu(o)}{2\clubsuit_L\log(n+1)}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{|\clubsuit_L|}\Bigg[\sup_{k'\geq 0}\lim_{k\to\infty}\int_{\mathcal{O}'}\int_{\mathcal{O}'}\frac{\sup_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\}}\left\{\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\right\|\right\}}{4\log^2(n+1)}d\nu(o)d\nu(o') \\
&\quad + \frac{2d^3\cos\left(\sup_{k'\geq 0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)}{3\log(n+1)}\sup_{k''\geq 0}\lim_{k\to\infty}\left(\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\int_{\mathcal{O}'}\sup_{\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\right)\cdot \\
&\quad \lim_{k\to\infty}\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d}\Bigg]^{1/2} - \sup_{k'\geq 0}\frac{\int_{\mathcal{O}'}\sup_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\}}\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|d\nu(o)}{2\clubsuit_L\log(n+1)} \\
&= \frac{1}{\clubsuit_L}\Bigg[\sup_{k'\geq 0}\int_{\mathcal{O}'}\int_{\mathcal{O}'}\frac{\sup_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\}}\left\{\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\right\|\right\}}{4\log^2(n+1)}d\nu(o)d\nu(o') \\
&\quad + \frac{2d^3\cos\left(\sup_{k'\geq 0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)}{3\log(n+1)}\cdot 0\Bigg]^{1/2} - \sup_{k'\geq 0}\frac{\int_{\mathcal{O}'}\sup_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\}}\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|d\nu(o)}{2\clubsuit_L\log(n+1)} \\
&= \sup_{k'\geq 0}\frac{\sup\limits_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\}}\left(\sqrt{\int_{\mathcal{O}'}\int_{\mathcal{O}'}\left\{\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\right\|\right\}d\nu(o)d\nu(o')}-\int_{\mathcal{O}'}\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|d\nu(o)\right)}{2\clubsuit_L\log(n+1)} \\
&= \sup_{k'\geq 0}\frac{\sup\limits_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\}}\left(\int_{\mathcal{O}'}\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|d\nu(o)-\int_{\mathcal{O}'}\left\|\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\right\|d\nu(o)\right)}{2\clubsuit_L\log(n+1)} \\
&= \sup_{k'\geq 0}\frac{0}{2\clubsuit_L\log(n+1)} = 0.
\end{aligned}
\tag{C.39}
$$

Similarly, for the upper bound in (C.37), we obtain the following

$$
\begin{aligned}
&\lim_{k\to\infty}\frac{\clubsuit_{R'}}{\clubsuit_L} \\
&= \lim_{k\to\infty}\Bigg\{\frac{\log\left(1+\frac{1}{n}\right)}{\clubsuit_L}\inf_{k'\geq 0}\int_{\mathcal{O}'}\inf_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\}}\epsilon^\top\frac{\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2}d\nu(o) - \Bigg[\inf_{k'\geq 0}\log^2\left(1+\frac{1}{n}\right)\int_{\mathcal{O}'}\int_{\mathcal{O}'}\inf_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\}} \\
&\quad \frac{\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon\epsilon^\top\right)\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\epsilon\epsilon^\top\right)}{\|\epsilon\|^4}d\nu(o)d\nu(o') - \frac{2d^3}{\log(n+1)}\cos\left(\pi-\max_{k'\geq 0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right) \\
&\quad \sup_{k''\geq 0}\max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3}\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\mathbf{0}}\right\|\int_{\mathcal{O}'}\sup_{\{\epsilon:p_n^{k''}(\epsilon)\in\mathcal{M}\}}\left|\frac{\partial^3\mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i\partial\epsilon_j\partial\epsilon_l}\right|d\nu(o)\Bigg]^{1/2}\frac{1}{\clubsuit_L}\Bigg\} \\
&= \frac{\log\left(1+\frac{1}{n}\right)}{\clubsuit_L}\inf_{k'\geq 0}\int_{\mathcal{O}'}\inf_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\}}\epsilon^\top\frac{\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2}d\nu(o) - \Bigg[\inf_{k'\geq 0}\log^2\left(1+\frac{1}{n}\right)\int_{\mathcal{O}'}\int_{\mathcal{O}'}\inf_{\{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\}} \\
&\quad \frac{\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon\epsilon^\top\right)\operatorname{tr}\left(\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))(o')\epsilon\epsilon^\top\right)}{\|\epsilon\|^4}d\nu(o)d\nu(o') - \frac{2d^3}{\log(n+1)}\lim_{k\to\infty}\cos\left(\pi-\max_{k'\geq 0}\mathcal{C}\left(\epsilon_n^{k'},\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))\right)\right)
\end{aligned}
$$

$$\cdot \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o) p_n d\nu(o) \Big|_{\epsilon=\mathbf{0}} \right\|_{\mathbb{R}^d} \sup_{k''\geq 0} \max_{(i,j,l)\in[1,d]^3\cap\mathbb{Z}^3} \int_{\mathcal{O}'} \sup_{\left\{\epsilon: p_n^{k''}(\epsilon)\in\mathcal{M}\right\}} \left| \frac{\partial^3 \mathbf{L}(p_n^{k''}(\epsilon))(o)}{\partial\epsilon_i \partial\epsilon_j \partial\epsilon_l} \right| d\nu(o) \Bigg]^{1/2} \frac{1}{\clubsuit_L}$$

$$= \frac{\log\left(1+\frac{1}{n}\right)}{\clubsuit_L} \Bigg[\inf_{k'\geq 0} \int_{\mathcal{O}'} \inf_{\left\{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}} \epsilon^\top \frac{\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon}{\|\epsilon\|^2} d\nu(o) - \Bigg[\inf_{k'\geq 0} \int_{\mathcal{O}'}\int_{\mathcal{O}'} \inf_{\left\{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}\setminus\{p_n^{k'}\}\right\}}$$

$$\operatorname{tr}\left(\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o)\epsilon\epsilon^\top\right) \operatorname{tr}\left(\nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))(o')\epsilon\epsilon^\top\right) \|\epsilon\|^{-4} \; d\nu(o) d\nu(o') \Bigg]^{1/2} \Bigg] = 0. \tag{C.40}$$

It then follows from the bounds established in (C.37) that both the lower and upper estimates converge to zero. Applying the squeeze theorem [31], we conclude that $\lim_{k\to\infty} \left\|\epsilon_n^k\right\|_{\mathbb{R}^d} = 0 = \lim_{k\to\infty} \left\|\epsilon_n^k - \mathbf{0}\right\|_{\mathbb{R}^d}$. Invoking the sequential characterization of limits in normed spaces, this further implies $\lim_{k\to\infty} \epsilon_n^k = \mathbf{0}$. □

C.3 Proof of (E3) and (E4)

Proof. The validity of statement (E3) follows under the regularity conditions outlined in Result 1 of Van Der Laan and Rubin [5]. To establish (E4), we observe that it arises as a direct consequence of combining (E1) with (E3). □

D Proof of Theorem 2

Recall that the submodel is defined via $p_n^k(\epsilon) = \breve{f}(\epsilon^\top D_\Psi^*(p_n^k))$. To avoid ambiguity, we explicitly denote the map $\breve{f}$ as $\breve{f}(t)(o)$. Before proceeding with the proof, we first state the auxiliary Lemma 2.

Lemma 2 (Paley-Zygmund Inequality [32])**.** *Let $X \geq 0$ be a random variable with its mean denote by $\mathbb{E}[X]$. If $0 < \mathbb{E}\left[X^2\right] < \infty$, then for any $0 \leq \theta \leq 1$ we have*

$$\mathbb{P}(X > \theta\mathbb{E}[X]) \geq (1-\theta)^2 \frac{\mathbb{E}^2[X]}{\mathbb{E}\left[X^2\right]}. \tag{D.41}$$

With this preliminary, we now turn to the formal proof of Theorem 2.

Proof. We first prove that the one–dimensional fluctuation path (i.e. $d = 1$ case) cannot self-intersect within the model space. Formally, we aim to show that for any two distinct fluctuation parameters $\epsilon_1 \neq \epsilon_2$, the (squared) Hellinger distance

$$\begin{aligned} \mathbb{D}_{\mathrm{H}}\big(p\left(\epsilon_1\right), p\left(\epsilon_2\right)\big) &\triangleq \frac{1}{2} \int_\Omega \left(\sqrt{p(\epsilon_1)} - \sqrt{p(\epsilon_2)}\right)^2 d\nu(o) \\ &= \frac{1}{2} \int_\Omega \left(\sqrt{\breve{f}\left(\epsilon_1 D_\Psi^*\right)} - \sqrt{\breve{f}\left(\epsilon_2 D_\Psi^*\right)}\right)^2 d\nu(o) \end{aligned} \tag{D.42}$$

admits a strictly positive lower bound, so the homotopy path remains injective and thus cannot self-intersect. Concretely, let

$$\left[\underline{t}, \bar{t}\,\right] := \left[\min\left\{\left|\epsilon_1\right|, \left|\epsilon_2\right|\right\} \sup_{o\in\Omega} \left|D_\Psi^*(p)(o)\right|,\;\; \max\left\{\left|\epsilon_1\right|, \left|\epsilon_2\right|\right\} \sup_{o\in\Omega} \left|D_\Psi^*(p)(o)\right|\right]. \tag{D.43}$$

Since $D_\Psi^*(p)(o)$ is a given direction of fluctuation in a semi-parametric model, assume without loss of generality that $\|D_\Psi^*\|_\infty < \infty$. [11] Then we can derive a global lower bound on (D.42) as

$$\mathbb{D}_{\mathrm{H}}\big(p\left(\epsilon_1\right), p\left(\epsilon_2\right)\big) = \frac{1}{2} \int_\Omega \left(\frac{\breve{f}\left(\epsilon_1 D_\Psi^*(p)(o)\right) - \breve{f}\left(\epsilon_2 D_\Psi^*(p)(o)\right)}{\sqrt{\breve{f}\left(\epsilon_1 D_\Psi^*(p)(o)\right)} + \sqrt{\breve{f}\left(\epsilon_2 D_\Psi^*(p)(o)\right)}} \right)^2 d\nu(o)$$

[11] If D_Ψ^* were unbounded one can localize the argument on support where ϵD_Ψ^* must lie in the domain of $\breve{f}$.

$$
\begin{aligned}
&\geq \frac{1}{2}\int_{\Omega} \frac{\left(\breve{f}\left(\epsilon_1 D_{\Psi}^{*}(p)(o)\right) - \breve{f}\left(\epsilon_2 D_{\Psi}^{*}(p)(o)\right)\right)^2}{\left(\sqrt{\sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)} + \sqrt{\sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)}\right)^2}\, d\nu(o) \\
&\overset{(a)}{=} \frac{1}{2}\int_{\Omega} \frac{\left(\breve{f}'\big(\xi(o) D_{\Psi}^{*}(p)(o)\big)\left(\epsilon_2 D_{\Psi}^{*}(p)(o) - \epsilon_1 D_{\Psi}^{*}(p)(o)\right)\right)^2}{4 \sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)} d\nu(o) \\
&= \int_{\Omega} \frac{\left(\left|\breve{f}'\big(\xi(o) D_{\Psi}^{*}(p)(o)\big)\right| \cdot |\epsilon_2 - \epsilon_1| \cdot \left|D_{\Psi}^{*}(p)(o)\right|\right)^2}{8 \sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)} d\nu(o) \\
&\geq \frac{\left(\epsilon_2 - \epsilon_1\right)^2}{8 \sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)} \int_{\Omega} \left(\inf_{t\in[\underline{t},\bar{t}]} \left|\breve{f}'(t)\right|\right)^2 \left|D_{\Psi}^{*}(p)(o)\right|^2 d\nu(o) \qquad \text{(D.44)} \\
&\overset{(b)}{\geq} \frac{\left(\epsilon_2 - \epsilon_1\right)^2}{8 \sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)} \left(\inf_{t\in[\underline{t},\bar{t}]} \left|\breve{f}'(t)\right|\right)^2 \int_{\{o : |D_{\Psi}^{*}(p)(o)| \geq 1/m'\}} D_{\Psi}^{*}(p)(o)^2 d\nu(o) \\
&\geq \frac{\left(\epsilon_2 - \epsilon_1\right)^2}{8 \sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)} \left(\inf_{t\in[\underline{t},\bar{t}]} \left|\breve{f}'(t)\right|\right)^2 \int_{\{o : |D_{\Psi}^{*}(p)(o)| \geq 1/m'\}} \frac{1}{(m')^2} d\nu(o) \\
&= \frac{\left(\inf\limits_{t\in[\underline{t},\bar{t}]} |\breve{f}'(t)|\right)^2 \nu\left(\{o : |D_{\Psi}^{*}(p)(o)| \geq 1/m'\}\right)}{8(m')^2 \sup\limits_{t\in[\underline{t},\bar{t}]} \breve{f}(t)} \cdot \left(\epsilon_2 - \epsilon_1\right)^2 .
\end{aligned}
$$

In step (a), the pointwise mean-value theorem is applied for each $o \in \mathcal{O}$, which guarantees the existence of a point $\xi(o)$ lying on the segment connecting ϵ_1 and ϵ_2. Step (b) follows from the observation that the set $\{o : D_{\Psi}^{*}(p)(o) \neq 0\}$ can be expressed as the countable union $\bigcup_{m=1}^{\infty}\{o : |D_{\Psi}^{*}(p)(o)| \geq 1/m\}$. Hence, there exists some $m' \in \mathbb{Z}^{+}$ for which $\nu\left(\{o : |D_{\Psi}^{*}(p)(o)| \geq 1/m'\}\right) > 0$. On the other hand, if $D_{\Psi}^{*}(p)(o) \equiv 0$ the submodel becomes degenerate, and it is immediate that the path cannot self-intersect. By (D.44) we conclude that $\mathbb{D}_{\mathrm{H}}\big(p\left(\epsilon_1\right), p\left(\epsilon_2\right)\big) = 0$ if and only if $\epsilon_2 - \epsilon_1 = 0$.

Then we move to the more general fluctuations with dimension $d \in [2, \infty) \cap \mathbb{Z}^{+}$, adopting a different line of reasoning.[12] Concretely, take two arbitrary but distinct parameters ϵ_1, ϵ_2 and define the associated corresponding measurable set

$$
\mathbb{S}\left(\epsilon_1, \epsilon_2\right) := \left\{o \in \Omega : \left|\frac{\left(\epsilon_2 - \epsilon_1\right)^{\top} D_{\Psi}^{*}(p)(o)}{\left\|\epsilon_2 - \epsilon_1\right\|_2}\right| \geq \frac{1}{2}\int_{\Omega} p(o) \left|\frac{\left(\epsilon_2 - \epsilon_1\right)^{\top} D_{\Psi}^{*}(p)(o)}{\left\|\epsilon_2 - \epsilon_1\right\|_2}\right| d\nu(o)\right\}. \tag{D.45}
$$

Since $D_{\Psi}^{*}(p)(o)$ is a real vector-valued random variable and by condition (10)

$$
\mathbb{P}\left\{o \in \Omega : \left|\beta^{\top} D_{\Psi}^{*}(p)(o)\right| \neq 0\right\} > 0, \tag{D.46}
$$

its tail probabilities satisfy

$$
\lim_{\varpi \to \infty} \mathbb{P}\left(o \in \Omega : \|D_{\Psi}^{*}(p)(o)\| > \varpi\right) = 0. \tag{D.47}
$$

Hence we can pick ϖ such that

$$
\mathbb{P}\{\|D_{\Psi}^{*}(p)(o)\| \leq \varpi\} > 1 - \frac{1}{2}\mathbb{P}\left\{o \in \Omega : \left|\beta^{\top} D_{\Psi}^{*}(p)(o)\right| > 0\right\}. \tag{D.48}
$$

[12] Kindly note that the previous proof under the $d = 1$ dose not generalize well due to technical challenges. For instance, in a multi-dimensional setting the equality $|a \cdot b| = |a||b|$ no longer holds in general, and one can only assert the inequality like $|a \cdot b| \leq |a||b|$. Consequently, the direction of the bound is reversed compared with the scaling inequality presented in e.g., (D.44).

Then we assign the new $\underline{t}$ and $\bar{t}$ as follows

$$\left[\,\underline{t},\bar{t}\,\right] := \left[\min\left\{\,\|\epsilon_1\|\,,\|\epsilon_2\|\right\}\varpi,\ \max\left\{\,\|\epsilon_1\|\,,\|\epsilon_2\|\right\}\varpi\right]. \tag{D.49}$$

This construction yields the following pointwise lower bound

$$\begin{aligned}
\left\|p(\epsilon_2)-p(\epsilon_1)\right\|_1 &= \int_\Omega \left|\breve{f}\left(\epsilon_2^\top D_\Psi^*(p)(o)\right)-\breve{f}\left(\epsilon_1^\top D_\Psi^*(p)(o)\right)\right| d\nu(o)\\
&\geq \int_{\mathbb{S}(\epsilon_1,\epsilon_2)} \left|\breve{f}\left(\epsilon_2^\top D_\Psi^*(p)(o)\right)-\breve{f}\left(\epsilon_1^\top D_\Psi^*(p)(o)\right)\right| d\nu(o)\\
&\overset{(a)}{=} \int_{\mathbb{S}(\epsilon_1,\epsilon_2)} \left|\breve{f}'\left(\xi(o)^\top D_\Psi^*(p)(o)\right)\left(\epsilon_2^\top D_\Psi^*(p)(o)-\epsilon_1^\top D_\Psi^*(p)(o)\right)\right| d\nu(o)\\
&\geq \int_{\mathbb{S}(\epsilon_1,\epsilon_2)} \inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\left(\epsilon_2^\top D_\Psi^*(p)(o)-\epsilon_1^\top D_\Psi^*(p)(o)\right)\right| d\nu(o)\\
&= \inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \int_{\mathbb{S}(\epsilon_1,\epsilon_2)} \left|(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(o)\right| d\nu(o)\\
&\overset{(D.45)}{\geq} \frac{1}{2}\inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \left\|\epsilon_2-\epsilon_1\right\|_2 \int_{\mathbb{S}(\epsilon_1,\epsilon_2)}\int_\Omega p(u)\left|\frac{(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(u)}{\left\|\epsilon_2-\epsilon_1\right\|_2}\right| d\nu(u)d\nu(o)\\
&\geq \frac{1}{2}\inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \left\|\epsilon_2-\epsilon_1\right\|_2 \int_\Omega p(o)\mathbb{P}\left(\mathbb{S}\left(\epsilon_1,\epsilon_2\right)\right)\left|\frac{(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(o)}{\left\|\epsilon_2-\epsilon_1\right\|_2}\right| d\nu(o)\\
&\overset{(b)}{\geq} \frac{1}{2}\inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \left\|\epsilon_2-\epsilon_1\right\|_2 \cdot\\
&\qquad \int_\Omega p(o)\left(\int_\Omega p(u)\frac{\left(1-\frac{1}{2}\right)\left|\frac{(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(u)}{\left\|\epsilon_2-\epsilon_1\right\|_2}\right|}{\sqrt{\int_\Omega p(u')\left(\frac{(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(u')}{\left\|\epsilon_2-\epsilon_1\right\|_2}\right)^2 d\nu(u')}}d\nu(u)\right)^2\left|\frac{(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(o)}{\left\|\epsilon_2-\epsilon_1\right\|_2}\right| d\nu(o)\\
&\geq \frac{1}{8}\inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \frac{\left(\int_\Omega p(o)\left|\frac{(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(o)}{\left\|\epsilon_2-\epsilon_1\right\|_2}\right| d\nu(o)\right)^3}{\int_\Omega\left(\frac{(\epsilon_2-\epsilon_1)^\top D_\Psi^*(p)(u')}{\left\|\epsilon_2-\epsilon_1\right\|_2}\right)^2 p(u')d\nu(u')}\left\|\epsilon_2-\epsilon_1\right\|_2\\
&\geq \frac{1}{8}\inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \frac{\left(\inf_{\|\mathbf{u}\|_2=1}\int_\Omega p(u)\left|\mathbf{u}^\top D_\Psi^*(p)(u)\right| d\nu(u)\right)^3}{\sup_{\|\mathbf{u}\|_2=1}\int_\Omega\left(\mathbf{u}^\top D_\Psi^*(p)(u')\right)^2 p(u')d\nu(u')}\left\|\epsilon_2-\epsilon_1\right\|_2\\
&\geq \frac{1}{8\sqrt{d}}\inf_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \frac{\left(\inf_{\|\mathbf{u}\|_2=1}\int_\Omega p(u)\left|\mathbf{u}^\top D_\Psi^*(p)(u)\right| d\nu(u)\right)^3}{\int_\Omega \|D_\Psi^*(p)(u')\|_2^2\, p(u')d\nu(u')}\cdot\left\|\epsilon_2-\epsilon_1\right\|_1,
\end{aligned} \tag{D.50}$$

where step (a) invokes the pointwise mean-value theorem for each $o \in \mathcal{O}$, and step (b) applies the Paley-Zygmund inequality (cf. Lemma 2). Now from the established ℓ_1 separation we know $\left\|p(\epsilon_2)-p(\epsilon_1)\right\|_1$ is strictly positive if and only if $\epsilon_1 \neq \epsilon_2$. As a further remark, an argument parallel to that of (D.50) leads to

$$\left\|p(\epsilon_2)-p(\epsilon_1)\right\|_1 \lesssim \sup_{t\in[\underline{t},\bar{t}]}\left|\breve{f}'(t)\right| \sup_{\|\mathbf{u}\|_2=1}\sqrt{\int_\Omega\left(\mathbf{u}^\top D_\Psi^*(p)(u')\right)^2 p(u')d\nu(u')}\cdot\left\|\epsilon_2-\epsilon_1\right\|_1, \tag{D.51}$$

which demonstrates that the mapping $\epsilon \mapsto p(\epsilon)$ defines a bi-Lipschitz embedding. Consequently, the semi-parametric fluctuation path forms a one-to-one immersed submanifold of the model space, endowed with global linear separation. In particular, this excludes the possibility of self-intersection. □

D.1 Injectivity of fluctuations

We first present the formal statement in Lemma 3, followed by detailed proofs of each component.

Lemma 3 (Injectivity of $\breve{f}$)**.** *The function $\breve{f}(\cdot)$ appearing in the structure of Eq.* (4)*, Eq.* (5)*, and Eq.* (6) *is injective over its effective domain.*

Throughout, when we refer to the *effective domain* of $\breve{f}(\cdot)$, we mean the set of all values $t(o) = \epsilon^{\top} D^*_{\Psi}(p)(o)$ that arise from valid semi-parametric fluctuation directions. Equivalently, $t(o)$ must lie in the linear span of the components of the efficient influence function $D^*_{\Psi}(p)(o)$.

To formally prove Lemma 3, we begin by proving the claim in the case of Example 1.

Proof. We will make proof by contradiction. Assume there exist $t_1, t_2 \in L^0(\nu)$ with $\breve{f}(t_1) = \breve{f}(t_2)$ in $L^0(\nu)$, but the $\Delta t(o) \triangleq t_1(o) - t_2(o)$ does not vanish ν-a.e.. We set $E^\star = \{o \in \mathcal{O} : \Delta t(o) \neq 0\}$ so $\nu(E^\star) > 0$. Since $\breve{f}(t_1) = \breve{f}(t_2)$ a.e., it can be written as

$$p^k_n(o)\,(1 + t_1(o)) = p^k_n(o)\,(1 + t_2(o))\,, \tag{D.52}$$

which means $p^k_n(o)\Delta t(o) = 0$. So we have $\left|p^k_n(o)\Delta t(o)\right| = 0$, ν-a.e. Hence the integral of nonnegative function $o \mapsto \left|p^k_n(o)\Delta t(o)\right|$ is

$$0 = \int_{\Omega} \left|p^k_n(o)\Delta t(o)\right| d\nu(o). \tag{D.53}$$

Consider the set $\left\{o : p^k_n(o) \geq \frac{1}{n'}\right\}$ for each integer $n' \geq 1$, since $p^k_n > 0$ almost everywhere, we know $\bigcup_{n'=1}^{\infty} \left\{o : p^k(o) \geq \frac{1}{n'}\right\}$ covers the Ω up to a null set. Similarly, consider $\left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}$ for each integer $m \geq 1$. Since $\Delta t \neq 0$ on $E^\star$, $\bigcup_{m=1}^{\infty} \left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}$ covers $E^\star$ and thus covers a set of positive measure. On each cell $\left\{o : p^k_n(o) \geq \frac{1}{n'}\right\} \cap \left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}$, we know that $p^k_n(o) \geq \frac{1}{n'}$, $|\Delta t(o)| \geq \frac{1}{m}$, so

$$\left|p^k_n(o)\Delta t(o)\right| \geq \frac{1}{n'} \cdot \frac{1}{m} = \frac{1}{n'm}. \tag{D.54}$$

We may now express the quantity as

$$\begin{aligned}
0 &\overset{(D.53)}{=} \int_{\Omega} \left|p^k_n \Delta t\right| d\nu \\
&= \int_{\Omega} \sum_{n'=1}^{\infty} \sum_{m=1}^{\infty} \mathbf{1}_{\left\{o : p^k_n(o) \geq \frac{1}{n'}\right\} \cap \left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}}(o) \left|p^k_n(o)\Delta t(o)\right| d\nu(o) \\
&\overset{(a)}{=} \sum_{n'=1}^{\infty} \sum_{m=1}^{\infty} \int_{\left\{o : p^k_n(o) \geq \frac{1}{n'}\right\} \cap \left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}} \left|p^k_n(o)\Delta t(o)\right| d\nu(o) \\
&\overset{(D.54)}{\geq} \sum_{n'=1}^{\infty} \sum_{m=1}^{\infty} \int_{\left\{o : p^k_n(o) \geq \frac{1}{n'}\right\} \cap \left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}} \frac{1}{n'm} d\nu(o) \\
&= \sum_{n'=1}^{\infty} \sum_{m=1}^{\infty} \frac{1}{n'm} \nu\left(\left\{o : p^k_n(o) \geq \frac{1}{n'}\right\} \cap \left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}\right),
\end{aligned} \tag{D.55}$$

where step (a) applies the Fubini–Tonelli theorem to justify the interchange of integration and summation. Given that $E^\star = \bigcup_{m=1}^{\infty} \left\{o : |\Delta t(o)| \geq \frac{1}{m}\right\}$ and $\bigcup_{n'=1}^{\infty} \left\{o : p^k(o) \geq \frac{1}{n'}\right\}$ both have positive measure, there must exist at least one pair (n'_0, m_0) for which $\nu(\{o : p^k_n(o) \geq \frac{1}{n'_0}\} \cap \{o : |\Delta t(o)| \geq \frac{1}{m_0}\}) > 0$. As a result, the double summation series in (D.55) splits to the form

$$
\begin{aligned}
0 \;&\geq \sum_{n'=1}^{\infty}\sum_{m=1}^{\infty}\frac{1}{n'm}\nu\left(\left\{o:p_n^k(o)\geq\frac{1}{n'}\right\}\cap\left\{o:|\Delta t(o)|\geq\frac{1}{m}\right\}\right)\\
&\geq \underbrace{\frac{1}{n_0'm_0}\nu\left(\left\{o:p_n^k(o)\geq\frac{1}{n_0'}\right\}\cap\left\{o:|\Delta t(o)|\geq\frac{1}{m_0}\right\}\right)}_{(1)\ n'=n_0',m=m_0}+\underbrace{\sum_{\substack{m=1\\m\neq m_0}}^{\infty}\frac{1}{n_0'm}\nu\left(\left\{o:p_n^k(o)\geq\frac{1}{n_0'}\right\}\cap\left\{o:|\Delta t(o)|\geq\frac{1}{m}\right\}\right)}_{(2)\ n'=n_0',m\neq m_0}\\
&+\underbrace{\sum_{\substack{n'=1\\n'\neq n_0'}}^{\infty}\frac{1}{n'm_0}\nu\left(\left\{o:p_n^k(o)\geq\frac{1}{n'}\right\}\cap\left\{o:|\Delta t(o)|\geq\frac{1}{m_0}\right\}\right)}_{(3)\ n'\neq n_0',m=m_0}+\underbrace{\sum_{\substack{n'=1\\n'\neq n_0'}}^{\infty}\sum_{\substack{m=1\\m\neq m_0}}^{\infty}\frac{1}{n'm}\nu\left(\left\{o:p_n^k(o)\geq\frac{1}{n'}\right\}\cap\left\{o:|\Delta t(o)|\geq\frac{1}{m}\right\}\right)}_{(4)\ n'\neq n_0',m\neq m_0}\\
&\geq \frac{1}{n_0'm_0}\nu\left(\left\{o:p_n^k(o)\geq\frac{1}{n_0'}\right\}\cap\left\{o:|\Delta t(o)|\geq\frac{1}{m_0}\right\}\right)\;>\;0.
\end{aligned}
\tag{D.56}
$$

The only resolution to this contradiction is that the set $\{o:\Delta t(o)\neq 0\}$ has zero ν-measure. Equivalently, it follows that $t_1(o)=t_2(o)$ for ν-a.e. o, by definition we see that $\breve{f}$ is injective on its effective domain. □

We now proceed to the proof of Example 2, beginning with the first formulation given in Eq. (5).

Proof. We will make proof by contradiction. Recall that the Eq. (5) could be rewritten as

$$
p_n^k(\epsilon)=\left(\int_\Omega e^{t(u)}p_n^k(u)d\nu(u)\right)^{-1}e^{t(o)}p_n^k(o). \tag{D.57}
$$

Assume there exist $t_1,t_2\in L^0(\nu)$ with $\breve{f}\,(t_1)=\breve{f}\,(t_2)$ (ν-a.e.) in $L^0(\nu)$ but $t_1\not\equiv t_2$, on a set of positive measure. For convenience we set the null-set[13]

$$
N_\star=\left\{o:\left(\int_\Omega e^{t_1(u)}p_n^k(u)d\nu(u)\right)^{-1}e^{t_1(o)}p_n^k(o)\neq\left(\int_\Omega e^{t_2(u)}p_n^k(u)d\nu(u)\right)^{-1}e^{t_2(o)}p_n^k(o)\right\}, \tag{D.58}
$$

and denote its complement as $N_\star^\complement$. For $o\in N_\star^\complement$, we define $\tilde{c}\triangleq t_1(o)-t_2(o)$, ν-a.e. and we will show that this condition is equivalent to the identity $\breve{f}\,(t_1)=\breve{f}\,(t_2)$. To see this, we first observe that

$$
\int_{N_\star^\complement} e^{t_1(u)}p_n^{k'}(u)d\nu(u)=e^{\tilde{c}}\cdot\int_{N_\star^\complement} e^{t_2(u)}p_n^{k'}(u)d\nu(u), \tag{D.59}
$$

where step (a) follows by re-indexing the summation using the substitution $j=m-k$. This reveals that the ratio of the normalizing constants reduces algebraically to $e^{\tilde{c}}$. With this simplification, we are now in a position to compute the distance between the corresponding probability distributions

$$
\begin{aligned}
&\int_\Omega\left|\breve{f}\,(t_1)\,(o)-\breve{f}\,(t_2)\,(o)\right|d\nu(o)\\
\overset{(D.58)}{=}\;&\int_{N_\star}\left|\left(\int_\Omega e^{t_1(u)}p_n^k(u)d\nu(u)\right)^{-1}e^{t_1(o)}-\left(\int_\Omega e^{t_2(u)}p_n^k(u)d\nu(u)\right)^{-1}e^{t_2(o)}\right|p_n^k(o)d\nu(o)\\
&+\int_{N_\star^\complement}\left|\left(\int_\Omega e^{t_1(u)}p_n^k(u)d\nu(u)\right)^{-1}e^{t_1(o)}-\left(\int_\Omega e^{t_2(u)}p_n^k(u)d\nu(u)\right)^{-1}e^{t_2(o)}\right|p_n^k(o)d\nu(o)
\end{aligned}
$$

[13] In order to distinguish between two different local integrals, we temporarily replace the variable o with u.

$$\begin{aligned}
&= 0 + \int_{N_\star^\complement} \left| \left(\int_\Omega e^{t_1(u)} p_n^k(u) d\nu(u) \right)^{-1} e^{t_1(o)} - \left(\int_\Omega e^{t_2(u)} p_n^k(u) d\nu(u) \right)^{-1} e^{t_2(o)} \right| p_n^k(o) d\nu(o) \\
&= \int_{N_\star^\complement} \left(\int_\Omega e^{t_2(u)} p_n^k(u) d\nu(u) \right)^{-1} \left| \left(\int_\Omega e^{t_1(u)} p_n^k(u) d\nu(u) \right)^{-1} e^{t_1(o)} \int_\Omega e^{t_2(u)} p_n^k(u) d\nu(u) - e^{t_2(o)} \right| p_n^k(o) d\nu(o) \\
&= \left(\int_\Omega e^{t_2(u)} p_n^k(u) d\nu(u) \right)^{-1} \int_{N_\star^\complement} \left| e^{t_1(o)} \left(\int_{N_\star^\complement} e^{t_1(u)} p_n^k(u) d\nu(u) + \underbrace{\int_{N_\star} e^{t_1(u)} p_n^k(u) d\nu(u)}_{=0 \text{ as } \nu(N_\star)=0} \right)^{-1} \right. \\
&\qquad \left. \cdot \left(\int_{N_\star^\complement} e^{t_2(u)} p_n^k(u) d\nu(u) + \underbrace{\int_{N_\star} e^{t_2(u)} p_n^k(u) d\nu(u)}_{=0 \text{ as } \nu(N_\star)=0} \right) - e^{t_2(o)} \right| p_n^k(o) d\nu(o) \\
&\overset{(D.59)}{=} \left(\int_\Omega e^{t_2(u)} p_n^k(u) d\nu(u) \right)^{-1} \int_{N_\star^\complement} \left| \frac{e^{t_1(o)}}{e^{\tilde{c}}} \left(\int_{N_\star^\complement} e^{t_2(u)} p_n^k(u) d\nu(u) + 0 \right)^{-1} \int_{N_\star^\complement} e^{t_2(u)} p_n^k(u) d\nu(u) - e^{t_2(o)} \right| p_n^k(o) d\nu(o) \\
&= \left(\int_\Omega e^{t_2(u)} p_n^k(u) d\nu(u) \right)^{-1} \int_{N_\star^\complement} \left| e^{t_2(o)} \cdot 1 - e^{t_2(o)} \right| p_n^k(o) d\nu(o) \\
&= 0.
\end{aligned} \tag{D.60}$$

The Eq. (D.60) is valid from both two directions, it follows that the condition $t_1 - t_2 = \tilde{c}$ is both necessary and sufficient condition for the equality. We restrict attention to those $t(o)$ lying in the effective domain, so both t_1 and t_2 satisfy the centering condition $\int_\Omega t_i(o) p_n^k(o) d\nu(o) = 0$, for $i = 1, 2$. Therefore

$$\begin{aligned}
0 &= \int_\Omega t_1(o) p_n^k(o) d\nu(o) \\
&= \int_\Omega \left(t_2(o) + \tilde{c} \right) p_n^k(o) d\nu(o) \\
&= \underbrace{\int_\Omega t_2(o) p_n^k(o) d\nu(o)}_{=0} + \int_\Omega \tilde{c} p_n^k(o) d\nu(o) = \tilde{c} \int_\Omega p_n^k(o) d\nu(o) = \tilde{c},
\end{aligned} \tag{D.61}$$

which implies $t_1(o) - t_2(o) = 0$, contradicting with the assumption that $t_1 \not\equiv t_2$. Therefore, the mapping must be injective. □

We now proceed to the case of Eq. (6).

Proof. Recall that the Eq. (6) could be rewritten as

$$p_n^k(\epsilon) = \left(\int_\Omega \frac{p_n^k(u)}{1 + e^{-2t(u)}} d\nu(u) \right)^{-1} \frac{p_n^k(o)}{1 + e^{-2t(o)}}. \tag{D.62}$$

Assume toward a contradiction that there exist functions $t_1, t_2 \in L^0(\nu)$ with $t_1 \not\equiv t_2$ but $\breve{f}(t_1) = \breve{f}(t_2)$. Hence we obtain

$$\left(1 + e^{-2t_2(o)} \right) \int_\Omega \frac{p_n^k(u')}{1 + e^{-2t_2(u')}} d\nu(u') = \left(1 + e^{-2t_1(o)} \right) \int_\Omega \frac{p_n^k(u)}{1 + e^{-2t_1(u)}} d\nu(u). \tag{D.63}$$

Then we can have

$$\begin{aligned}
0 &\equiv \int_\Omega t_2(o) p_n^k(o) d\nu(o) \\
&\overset{(D.63)}{=} -\frac{1}{2} \int_\Omega \ln \left[\left(\frac{1 + e^{-2t_2(o)}}{1 + e^{-2t_1(o)}} - 1 \right) + \frac{\displaystyle\int_\Omega \frac{p_n^k(u)}{1 + e^{-2t_1(u)}} d\nu(u)}{\displaystyle\int_\Omega \frac{p_n^k(u')}{1 + e^{-2t_2(u')}} d\nu(u')} e^{-2t_1(o)} \right] p_n^k(o) d\nu(o)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\int_\Omega \ln\left[\frac{\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)}{\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')}\left(e^{-2t_1(o)} + \left(\frac{\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)}{\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')} - 1\right)\frac{1+e^{-2t_1(o)}}{1+e^{-2t_2(o)}}\right)\right] p_n^k(o)d\nu(o) \\
&= -\frac{1}{2}\int_\Omega p_n^k(o)\left[\ln\frac{1+e^{-2t_2(o)}}{1+e^{-2t_1(o)}} + \ln e^{-2t_1(o)} + \ln\left(1 - \frac{1+e^{-2t_1(o)}}{1+e^{-2t_2(o)}}\left(1 - \frac{\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)}{\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')}\right)e^{2t_1(o)}\right)\right]d\nu(o) \\
&= -\frac{1}{2}\int_\Omega p_n^k(o)\left[\ln\frac{1+e^{-2t_2(o)}}{1+e^{-2t_1(o)}} - 2t_1 + \sum_{m=1}^{\infty}\left(\frac{\frac{\int_\Omega p_n^k(u)\left(1+e^{-2t_1(u)}\right)^{-1}d\nu(u)}{\int_\Omega p_n^k(u')\left(1+e^{-2t_2(u')}\right)^{-1}d\nu(u')} - 1}{\frac{\int_\Omega \left(1+e^{-2t_1(u)}\right)^{-1}p_n^k(u)d\nu(u)}{\int_\Omega \left(1+e^{-2t_2(u')}\right)^{-1}p_n^k(u')d\nu(u')}}\right)^m \frac{(-1)^{m+1}}{m}\, e^{2mt_1(o)}\right]d\nu(o) \\
&= \frac{1}{2}\ln\left(\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')\right)\int_\Omega p_n^k(o)d\nu(o) - \frac{1}{2}\ln\left(\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)\right) + \int_\Omega p_n^k(o)t_1(o)d\nu(o) \\
&\quad - \frac{1}{2}\sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}\left(\overbrace{\frac{\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u) - \int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')}{\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)}}^{\triangleq\blacktriangle(t_1(u),t_2(u'))}\right)^m \int_\Omega e^{2mt_1(o)}p_n^k(o)d\nu(o) \\
&= \frac{1}{2}\ln\left(\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')\right) - \frac{1}{2}\ln\left(\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)\right) \\
&\quad - \frac{1}{2}\sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}\cdot\Big(\blacktriangle\left(t_1(u),t_2(u')\right)\Big)^m\int_\Omega e^{2mt_1(o)}p_n^k(o)d\nu(o) \\
&\triangleq \diamondsuit.
\end{aligned}
\tag{D.64}
$$

Note that $\diamondsuit = 0$ if and only if

$$
\sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}\Big(\blacktriangle\left(t_1(u),t_2(u')\right)\Big)^m\cdot\int_\Omega e^{2m\cdot t_1(o)}p_n^k(o)d\nu(o) = \ln\left(\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')\right) - \ln\left(\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)\right),
\tag{D.65}
$$

where each moment term $\int e^{2mt_1}p_n^k d\nu > 0$. To proceed, we differentiate $\diamondsuit$ to calculate its total derivative with respect to the two normalizing constants. This yields

$$
\begin{aligned}
\diamondsuit' &= \frac{1}{\frac{2}{-\frac{}{1-\blacktriangle\left(t_1(u),t_2(u')\right)}}} - \frac{1}{2}\sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}\int_\Omega e^{2mt_1(o)}p_n^k(o)d\nu(o)\cdot\left(\Big(\blacktriangle\left(t_1(u),t_2(u')\right)\Big)^m\right)' \\
&= \frac{\blacktriangle\left(t_1(u),t_2(u')\right)-1}{2} - \frac{1}{2}\sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}\int_\Omega e^{2mt_1(o)}p_n^k(o)d\nu(o)\left[m\left(\frac{\blacktriangle\left(t_1(u),t_2(u')\right)\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)}{\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')}\right)^{m-1}\right. \\
&\quad \left.\cdot\left(\Big(1-\blacktriangle\left(t_1(u),t_2(u')\right)\Big)^{-m}\right)^{-1} - m\left(\frac{\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)}{\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')} - 1\right)^m\left(\Big(1-\blacktriangle\left(t_1(u),t_2(u')\right)\Big)^{-(m+1)}\right)^{-1}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\blacktriangle\left(t_1(u), t_2(u')\right) - 1}{2} - \frac{1}{2}\sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} m \int_{\Omega} e^{2mt_1(o)} p_n^k(o) d\nu(o) \left(\frac{\displaystyle\int_{\Omega} \frac{p_n^k(u)}{1+e^{-2t_1(u)}} d\nu(u)}{\displaystyle\int_{\Omega} \frac{p_n^k(u')}{1+e^{-2t_2(u')}} d\nu(u')} - 1 \right)^{m-1} \\
&\quad \cdot \Big(1 - \blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{m+1} \left[\frac{1}{1 - \blacktriangle\left(t_1(u), t_2(u')\right)} - \left(\frac{1}{1 - \blacktriangle\left(t_1(u), t_2(u')\right)} - 1 \right) \right] \\
&= \frac{\blacktriangle\left(t_1(u), t_2(u')\right) - 1}{2} - \frac{1}{2}\sum_{m=1}^{\infty} (-1)^{m+1} \int_{\Omega} e^{2mt_1(o)} p_n^k(o) d\nu(o) \left(\frac{1}{1 - \blacktriangle\left(t_1(u), t_2(u')\right)} - 1 \right)^{m-1} \\
&\quad \cdot \Big(1 - \blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{m+1} \cdot 1 \\
&= \frac{\blacktriangle\left(t_1(u), t_2(u')\right) - 1}{2} - \frac{1}{2}\sum_{m=1}^{\infty} (-1)^{m+1} \int_{\Omega} e^{2mt_1(o)} p_n^k(o) d\nu(o) \Big(\blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{m-1} \Big(1 - \blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{2}.
\end{aligned}
\tag{D.66}
$$

Observe that for every non-zero m, the function $f(x) = e^{2mx}$ is strictly convex since $\nabla^2 f = 4m^2 e^{2mx} > 0$, and Jensen's inequality implies

$$
\int_{\Omega} e^{2mt_1(o)} p_n^k(o) d\nu(o) \geq \exp\left(2m \int_{\Omega} t_1(o) p_n^k(o) d\nu(o) \right) = 1, \tag{D.67}
$$

and consequently we obtain the following upper bound

$$
\begin{aligned}
\Diamond' &\leq \frac{\blacktriangle\left(t_1(u), t_2(u')\right) - 1}{2} - \frac{1}{2}\sum_{m=1}^{\infty} (-1)^{m-1} \cdot 1 \cdot \Big(\blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{m-1} \Big(1 - \blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{2} \\
&\leq -\frac{1}{2} \frac{\displaystyle\int_{\Omega} \frac{p_n^k(u')}{1+e^{-2t_2(u')}} d\nu(u')}{\displaystyle\int_{\Omega} \frac{p_n^k(u)}{1+e^{-2t_1(u)}} d\nu(u)} - \frac{1}{2} \Big(1 - \blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{2} \sum_{m=1}^{\infty} (-1)^{m-1} \Big(\blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{m-1}.
\end{aligned}
\tag{D.68}
$$

To derive a sharp estimate on $\Diamond'$, it is necessary to analyze the range of the function $\blacktriangle\left(t_1(u), t_2(u')\right)$. Note that $\int_{\Omega} \frac{p_n^k(u)}{1+e^{-2t_i(u)}} d\nu(u) \in (0, \infty)$ for $i = 1, 2$, which ensures that

$$
\blacktriangle\left(t_1(u), t_2(u')\right) = \frac{\int_{\Omega} \frac{p_n^k(u)}{1+e^{-2t_1(u)}} d\nu(u) - \int_{\Omega} \frac{p_n^k(u')}{1+e^{-2t_2(u')}} d\nu(u')}{\int_{\Omega} \frac{p_n^k(u)}{1+e^{-2t_1(u)}} d\nu(u)} \in (-\infty, 1). \tag{D.69}
$$

We first consider the case where $\blacktriangle\left(t_1(u), t_2(u')\right) \in [0, 1)$. By the alternating-series lower bound [33] we obtain

$$
\begin{aligned}
\Diamond' &\leq -\frac{1}{2} \frac{\displaystyle\int_{\Omega} \frac{p_n^k(u')}{1+e^{-2t_2(u')}} d\nu(u')}{\displaystyle\int_{\Omega} \frac{p_n^k(u)}{1+e^{-2t_1(u)}} d\nu(u)} - \frac{\Big(1 - \blacktriangle\left(t_1(u), t_2(u')\right)\Big)^{2}}{2\left(\blacktriangle\left(t_1(u), t_2(u')\right) + 1\right)} \\
&< -\frac{1}{2} \frac{\displaystyle\int_{\Omega} \frac{p_n^k(u')}{1+e^{-2t_2(u')}} d\nu(u')}{\displaystyle\int_{\Omega} \frac{p_n^k(u)}{1+e^{-2t_1(u)}} d\nu(u)} - \frac{(1-1)^2}{2\,(1+1)} < 0.
\end{aligned}
\tag{D.70}
$$

For the case where $\blacktriangle\left(t_1(u), t_2(u')\right) \in (-\infty, 0)$, we rewrite $\blacktriangle := -\bar{\blacktriangle}$ with $\bar{\blacktriangle} > 0$. Then it follows that

$$
\begin{aligned}
\Diamond' &\leq -\frac{1}{2}\frac{\displaystyle\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')}{\displaystyle\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)} - \frac{1}{2}\Big(1+\bar{\blacktriangle}\left(t_1(u),t_2(u')\right)\Big)^2\sum_{m=1}^{\infty}(-1)^{m-1}\Big(-\bar{\blacktriangle}\left(t_1(u),t_2(u')\right)\Big)^{m-1} \\
&= -\frac{1}{2}\frac{\displaystyle\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')}{\displaystyle\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)} - \frac{1}{2}\Big(1+\bar{\blacktriangle}\left(t_1(u),t_2(u')\right)\Big)^2\sum_{m=1}^{\infty}(-1)^{m-1}\cdot(-1)^{m-1}\Big(\bar{\blacktriangle}\left(t_1(u),t_2(u')\right)\Big)^{m-1} \\
&= -\frac{1}{2}\left[\frac{\displaystyle\int_\Omega \frac{p_n^k(u')}{1+e^{-2t_2(u')}}d\nu(u')}{\displaystyle\int_\Omega \frac{p_n^k(u)}{1+e^{-2t_1(u)}}d\nu(u)} + \Big(1+\bar{\blacktriangle}\left(t_1(u),t_2(u')\right)\Big)^2\sum_{m=0}^{\infty}\Big(\bar{\blacktriangle}\left(t_1(u),t_2(u')\right)\Big)^{m}\right] \;<\; 0.
\end{aligned}
\tag{D.71}
$$

The reason behind last inequality is that if $\bar{\blacktriangle} \in (0,1)$, then $\sum_{m=0}^{\infty}\bar{\blacktriangle}^m = \frac{1}{1-\bar{\blacktriangle}} > 0$. If instead $\bar{\blacktriangle} \in [1,\infty]$, the series $\sum_{m=0}^{\infty}\bar{\blacktriangle}^m$ appearing in (D.71) diverges to $+\infty > 0$, and the inequality continues to hold trivially. Combine the (D.70) and (D.71) one obtains $\Diamond' < 0$. By observation we can learn that one root of the estimating equation (D.65) is $\blacktriangle = 0$ (i.e., $\int_\Omega \frac{p_n^k}{1+e^{-2t_1}}d\nu = \int_\Omega \frac{p_n^k}{1+e^{-2t_2}}d\nu$). Consequently,

$$
\begin{cases}
\Diamond > 0, & \text{when } 0 < \int_\Omega \frac{p_n^k}{1+e^{-2t_1}}d\nu < \int_\Omega \frac{p_n^k}{1+e^{-2t_2}}d\nu, \\
\Diamond = 0, & \text{when } \int_\Omega \frac{p_n^k}{1+e^{-2t_1}}d\nu = \int_\Omega \frac{p_n^k}{1+e^{-2t_2}}d\nu, \\
\Diamond < 0, & \text{when } \infty > \int_\Omega \frac{p_n^k}{1+e^{-2t_1}}d\nu > \int_\Omega \frac{p_n^k}{1+e^{-2t_2}}d\nu.
\end{cases}
\tag{D.72}
$$

As a result, the $\blacktriangle = 0$ is indeed the unique root of (D.65). With $\blacktriangle = 0$ in (D.64) we have

$$
\int \frac{p_n^k}{1+e^{-2t_1}}d\nu = \int \frac{p_n^k}{1+e^{-2t_2}}d\nu \;\Rightarrow\; 1+e^{-2t_2(o)} = 1+e^{-2t_1(o)},
\tag{D.73}
$$

so $t_1 = t_2$ a.e. on Ω. Thus our initial assumption $t_1 \not\equiv t_2$ leads to a contradiction and must be ruled out. Lastly, to ensure full rigor, we verify that the expansion used in (D.64) is valid within its region of application. Specifically, we observe that the relevant factor

$$
\left\| e^{2t_1(o)}\frac{\blacktriangle\left(t_1(u),t_2(u')\right)}{\int_\Omega p_n^k\left(1+e^{-2t_1(u)}\right)^{-1}d\nu(u)}\right\| \leq \left\|e^{2t_1(o)}\right\|_{L^2(\nu)}\cdot 0\cdot\left|\left(\int_\Omega p_n^k\left(1+e^{-2t_1(u)}\right)^{-1}d\nu(u)\right)^{-1}\right| < 1
\tag{D.74}
$$

lies within the standard convergence radius of Maclaurin series. Therefore, we conclude that the mapping $t \mapsto \breve{f}(t(\cdot))$ is injective on its domain. □

E Proof of Theorem 3

Proof. The existence of solutions to the empirical EIF estimating-equation is guaranteed under general conditions established by Bickel et al. [26]. Let $\hat{p}_0 \in \mathcal{M}$ denote such a feasible solution satisfying the

$$
\mathbf{0} = \int_{\mathcal{O}} D_\Psi^*(\hat{p}_0)(o)p_n d\nu(o).
$$

Now consider any perturbation $h \in \breve{\mathcal{T}}_p \triangleq \left\{h : \int_\Omega h(o)d\nu(o) = 0\right\}$. For each $o \in \mathcal{O}$, it follows that

$$
\begin{aligned}
D_\Psi^*(\hat{p}_0+h)(o) &= D_\Psi^*(\hat{p}_0)(o) + \int_0^d \mathcal{D}_f\frac{D_\Psi^*(\hat{p}_0+sd^{-1}h)}{d}[h](o)ds \\
&= D_\Psi^*(\hat{p}_0)(o) + \frac{1}{d}\int_0^d \mathcal{D}_f D_\Psi^*\left(\hat{p}_0+h\frac{s}{d}\right)[h](o)ds \\
&= D_\Psi^*(\hat{p}_0)(o) + \frac{1}{d}\int_0^d\left[\mathcal{D}_f D_\Psi^*(\hat{p}_0) + \frac{1}{d}\int_0^s \mathcal{D}_f^2 D_\Psi^*\left(\hat{p}_0+h\frac{v}{d}\right)[h]dv\right][h](o)ds
\end{aligned}
$$

$$
\begin{aligned}
&= D^*_\Psi(\hat{p}_0)(o) + \frac{1}{d}\mathcal{D}_f D^*_\Psi(\hat{p}_0)[h](o)\left(\int_0^d ds\right) + \frac{1}{d^2}\int_0^d\int_0^s \mathcal{D}^2_f D^*_\Psi\left(\hat{p}_0 + \frac{v}{d}h\right)[h,h](o)dvds \\
&\overset{(a)}{=} D^*_\Psi(\hat{p}_0)(o) + \frac{1}{d}\mathcal{D}_f D^*_\Psi(\hat{p}_0)[h](o)\cdot d + \frac{1}{d^2}\int_0^d\left(\int_v^d ds\right)\mathcal{D}^2_f D^*_\Psi\left(\hat{p}_0 + \frac{v}{d}h\right)[h,h](o)dv \\
&= D^*_\Psi(\hat{p}_0)(o) + \mathcal{D}_f D^*_\Psi(\hat{p}_0)[h](o) + \frac{1}{d^2}\int_0^d (d-v)\mathcal{D}^2_f D^*_\Psi\left(\hat{p}_0 + \frac{v}{d}h\right)[h,h](o)dv \\
&= D^*_\Psi(\hat{p}_0)(o) + \mathcal{D}_f D^*_\Psi(\hat{p}_0)[h](o) + \frac{1}{d^2}\left[\int_0^d (d-v)\left[\mathcal{D}^2_f D^*_\Psi\left(\hat{p}_0 + \frac{v}{d}h\right)[h,h](o)\right.\right. \\
&\qquad \left.\left. - \mathcal{D}^2_f D^*_\Psi(\hat{p}_0)[h,h](o)\right]dv + \mathcal{D}^2_f D^*_\Psi(\hat{p}_0)[h,h](o)\underbrace{\left[\int_0^d (d-v)dv\right]}_{=\frac{1}{2}d^2}\right] \\
&= D^*_\Psi(\hat{p}_0)(o) + \mathcal{D}_f D^*_\Psi(\hat{p}_0)[h](o) + \frac{1}{2}\mathcal{D}^2_f D^*_\Psi(\hat{p}_0)[h,h](o) \\
&\qquad + \frac{1}{d^2}\int_0^d (d-v)\left[\mathcal{D}^2_f D^*_\Psi\left(\hat{p}_0 + \frac{v}{d}h\right) - \mathcal{D}^2_f D^*_\Psi(\hat{p}_0)\right][h,h](o)dv,
\end{aligned}
\tag{E.75}
$$

where step (a) applies Fubini's Theorem to interchange the order of integration. The resulting integral term denoted by $\mathcal{R}(h, o; D^*_\Psi)$ satisfies the property

$$
\begin{aligned}
\mathcal{R}(h, o; D^*_\Psi) &\equiv \int_0^d (d-v)\left[\mathcal{D}^2_f D^*_\Psi\left(\hat{p}_0 + \frac{v}{d}h\right) - \mathcal{D}^2_f D^*_\Psi(\hat{p}_0)\right][h,h](o)dv, \\
&\lim_{\|h\|\to 0^+}\frac{\sup_{o\in\mathcal{O}}\left\|\mathcal{R}(h,o;D^*_\Psi)\right\|_{\mathbb{R}^d}}{\left\|h\right\|^2_{L^2(\nu)}} = 0.
\end{aligned}
\tag{E.76}
$$

For any two perturbations $h_1, h_2 \in \breve{\mathcal{T}}_{p_0}$, define $\Delta h = h_1 - h_2$. Then by applying the expansion (E.75) at $\hat{p}_0$, we deduce that

$$
\begin{aligned}
&\int_{\mathcal{O}} D^*_\Psi(\hat{p}_0 + h_1)(o)p_n d\nu(o) - \int_{\mathcal{O}} D^*_\Psi(\hat{p}_0 + h_2)(o)p_n d\nu(o) \\
&= \underbrace{\int_{\mathcal{O}} D^*_\Psi(\hat{p}_0)(o)p_n d\nu(o)}_{=0} + \mathcal{D}_f\int_{\mathcal{O}} D^*_\Psi(\hat{p}_0)[h_1](o)p_n d\nu(o) + \frac{1}{2}\mathcal{D}^2_f\int_{\mathcal{O}} D^*_\Psi(\hat{p}_0)[h_1,h_1](o)p_n d\nu(o) \\
&\quad + \frac{1}{d^2}\int_{\mathcal{O}}\mathcal{R}(h_1,o;D^*_\Psi)p_n d\nu(o) - \underbrace{\int_{\mathcal{O}} D^*_\Psi(\hat{p}_0)(o)p_n d\nu(o)}_{=0} - \frac{1}{d^2}\int_{\mathcal{O}}\mathcal{R}(h_2,o;D^*_\Psi)p_n d\nu(o) \\
&\quad - \mathcal{D}_f\int_{\mathcal{O}} D^*_\Psi(\hat{p}_0)[h_2](o)p_n d\nu(o) - \frac{1}{2}\mathcal{D}^2_f\int_{\mathcal{O}} D^*_\Psi(\hat{p}_0)[h_2,h_2](o)p_n d\nu(o) \\
&\overset{(a)}{=} \int_{\mathcal{O}}\mathcal{D}_f D^*_\Psi(\hat{p}_0)[\Delta h](o)p_n d\nu(o) + \frac{1}{2}\int_{\mathcal{O}}\left[\mathcal{D}^2_f D^*_\Psi(\hat{p}_0)[h_1,h_1](o)p_n - \mathcal{D}^2_f D^*_\Psi(\hat{p}_0)[h_2,h_2](o)p_n\right]d\nu(o) \\
&\quad + \int_{\mathcal{O}}\frac{\mathcal{R}(h_1,o;D^*_\Psi) - \mathcal{R}(h_2,o;D^*_\Psi)}{d^2}p_n d\nu(o) \\
&\overset{(b)}{\leq} \underbrace{\int_{\mathcal{O}}\mathcal{D}_f D^*_\Psi(\hat{p}_0)[\Delta h](o)p_n d\nu(o)}_{\frac{1}{2\pi\mathrm{i}}\oint_{|\zeta|=\rho}\int_{\mathcal{O}}\frac{D^*_\Psi(\hat{p}_0+\zeta\Delta h)(o)}{\zeta^2}p_n d\nu(o)d\zeta} + \frac{1}{2}\left[\underbrace{\int_{\mathcal{O}}\mathcal{D}^2_f D^*_\Psi(\hat{p}_0)[h_1,h_1](o)p_n d\nu(o)}_{\frac{1}{\pi\mathrm{i}}\oint_{|\zeta|=\rho}\int_{\mathcal{O}}\frac{D^*_\Psi(\hat{p}_0+\zeta h_1)(o)}{\zeta^3}p_n d\nu(o)d\zeta} - \int_{\mathcal{O}}\mathcal{D}^2_f D^*_\Psi(\hat{p}_0)[h_2,h_2](o)p_n d\nu(o)\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{d^2}\int_{\mathcal{O}} p_n \left(\oint_{|\zeta|=\rho} D_\Psi^*(\hat{p}_0+\zeta h_1)(o)\cdot \frac{\frac{d^2}{\zeta-1} - \left(\frac{d^2}{\zeta}+\frac{d^2}{\zeta^2}+\frac{d^2}{\zeta^3}\right)}{2\pi \mathrm{i}} d\zeta - \mathcal{R}(h_2,o;D_\Psi^*)\right) d\nu(o) \\
\overset{(c)}{=}\ & \frac{1}{2\pi \mathrm{i}}\left(\oint_{|\zeta|=\rho} \sum_{j=1}^{\infty} \frac{1}{\zeta^j} \int_{\mathcal{O}} \frac{D_\Psi^*(\hat{p}_0+\zeta h_1)(o)}{\zeta^3} d\nu(o) d\zeta - \oint_{|\zeta|=\rho} \sum_{k=1}^{\infty} \frac{1}{\zeta^k} \int_{\mathcal{O}} \frac{D_\Psi^*(\hat{p}_0+\zeta h_2)(o)}{\zeta^3} d\nu(o) d\zeta \right) \\
&+ \int_{\mathcal{O}} \mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o) p_n d\nu(o) + \frac{1}{2}\int_{\mathcal{O}} \left[\mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1,h_1](o)p_n - \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_2,h_2](o)p_n \right] d\nu(o) \\
=\ & \oint_{|\zeta|=\rho} \frac{\int_{\mathcal{O}} D_\Psi^*(\hat{p}_0+\zeta h_1)(o)p_n d\nu(o) - \int_{\mathcal{O}} D_\Psi^*(\hat{p}_0+\zeta h_2)(o) p_n d\nu(o)}{2\zeta^3(\zeta-1)\pi \mathrm{i}} d\zeta \\
&+ \int_{\mathcal{O}} \mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o) p_n d\nu(o) + \frac{1}{2}\int_{\mathcal{O}} \left[\mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1,h_1](o)p_n - \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_2,h_2](o)p_n \right] d\nu(o),
\end{aligned}
\tag{E.77}
$$

where step (a) follows from the application of the Lebesgue's dominated convergence theorem, step (b) invokes the use of complex contour-integrals [35, 36] in a Banach space and parameter $\rho > 1$ is chosen such that the integration contour lies entirely within the domain of analyticity [37]. The step (c) employs α-conversion for the functional term $\mathcal{R}(h_2, o; D_\Psi^*)$, following the formalism in Barendregt and Barendsen [38]. First we can bound the partial term $\oint_{|\zeta|=\rho} \frac{1}{|\zeta|^2|\zeta-1|} d\zeta$ by parameterizing the complex variable by $\zeta = \rho e^{\mathrm{i}\theta}, \theta \in [0, 2\pi]$. Then the differential is given by $d\zeta = \mathrm{i}\rho e^{\mathrm{i}\theta} d\theta$ with $|d\zeta| = \rho d\theta$. Hence,

$$
\begin{aligned}
\oint_{|\zeta|=\rho} \frac{1}{|\zeta|^2|\zeta-1|}|d\zeta| =& \frac{1}{\rho}\int_0^{2\pi} \frac{d\theta}{|\rho e^{\mathrm{i}\theta}-1|} \\
=& \frac{1}{\rho}\int_0^{2\pi} \frac{d\theta}{\sqrt{(\rho\cos\theta-1)^2+(\rho\sin\theta)^2}} \\
\leq& \frac{1}{\rho}\int_0^{2\pi} \frac{d\theta}{\min_{\theta\in[0,2\pi]}\sqrt{\rho^2-2\rho\cos\theta+1}} \\
=& \frac{1}{\rho}\int_0^{2\pi} \frac{d\theta}{\sqrt{\rho^2-2\rho\cdot 1+1}} \\
\leq& \frac{1}{\rho}\cdot\frac{2\pi}{\rho-1} = \frac{2\pi}{\rho(\rho-1)}.
\end{aligned}
\tag{E.78}
$$

Taking the Euclidean norm in $\mathbb{R}^d$) and applying the triangle inequality for both sides of (E.77), we obtain

$$
\begin{aligned}
&\left\| \int_{\mathcal{O}} D_\Psi^*(\hat{p}_0+h_1)(o)p_n d\nu(o) - \int_{\mathcal{O}} D_\Psi^*(\hat{p}_0+h_2)(o)p_n d\nu(o) \right\|_{\mathbb{R}^d} \\
\leq& \left\| \int_{\mathcal{O}} \mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o) p_n d\nu(o) + \frac{1}{2}\int_{\mathcal{O}} \left[\mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1,h_1](o)p_n - \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_2,h_2](o)p_n \right] d\nu(o) \right. \\
&\left. + \oint_{|\zeta|=\rho} \frac{\int_{\mathcal{O}} D_\Psi^*(\hat{p}_0+\zeta h_1)(o)p_n d\nu(o) - \int_{\mathcal{O}} D_\Psi^*(\hat{p}_0+\zeta h_2)(o) p_n d\nu(o)}{2\zeta^3(\zeta-1)\pi \mathrm{i}} d\zeta \right\|_{\mathbb{R}^d} \\
\leq& \left\| \int_{\mathcal{O}} \mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o) p_n d\nu(o) \right\|_{\mathbb{R}^d} + \frac{1}{2} \left\| \left(\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1,h_1](o)p_n d\nu(o) - \int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1,h_2](o)p_n d\nu(o) \right) \right. \\
&+ \left(\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1,h_2](o)p_n d\nu(o) - \int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_2,h_2](o)p_n d\nu(o) \right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{\pi \mathrm{i}} \oint_{|\zeta|=\rho} \int_{\mathcal{O}} p_n \frac{D_\Psi^*(\hat{p}_0+\zeta h_1)(o)-D_\Psi^*(\hat{p}_0+\zeta h_2)(o)}{\zeta^3(\zeta-1)} d\nu(o) d\zeta \Bigg\|_{\mathbb{R}^d} \\
\overset{(a)}{\leq}\ & \int_{\mathcal{O}} p_n \left\|\mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o)\right\|_{\mathbb{R}^d} d\nu(o)+\frac{1}{2}\Bigg\|\left(\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_2, h_1](o) p_n d\nu(o)-\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_2, h_2](o) p_n d\nu(o)\right) \\
&+\left(\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1, h_1](o) p_n d\nu(o)-\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1, h_2](o) p_n d\nu(o)\right)\Bigg\|_{\mathbb{R}^d} \\
&+\frac{1}{2}\left\|\frac{1}{\pi \mathrm{i}} \oint_{|\zeta|=\rho} \int_{\mathcal{O}} p_n \frac{\left[D_\Psi^*(\hat{p}_0+\zeta h_1)-D_\Psi^*(\hat{p}_0+\zeta h_2)\right](o)}{\zeta^3(\zeta-1)} d\nu(o) d\zeta\right\|_{\mathbb{R}^d} \\
\overset{(b)}{\leq}\ & \int_{\mathcal{O}} p_n \left\|\mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o)\right\|_{\mathbb{R}^d} d\nu(o)+\frac{1}{2\pi} \oint_{|\zeta|=\rho} \int_{\mathcal{O}} p_n \frac{\left\|D_\Psi^*(\hat{p}_0+\zeta h_1)(o)-D_\Psi^*(\hat{p}_0+\zeta h_2)(o)\right\|_{\mathbb{R}^d}}{|\zeta|^3 \cdot|\zeta-1|} d\nu(o)\,|d\zeta| \\
&+\frac{1}{2}\left\|\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1, h_1-h_2](o) p_n d\nu(o)+\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_2, h_1-h_2](o) p_n d\nu(o)\right\|_{\mathbb{R}^d} \\
\overset{(c)}{\leq}\ & \int_{\mathcal{O}'} \frac{2\pi\left\|\mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o)\right\|_{\mathbb{R}^d}+\oint_{|\zeta|=\rho} \frac{\left\|D_\Psi^*(\hat{p}_0+\zeta h_1)(o)-D_\Psi^*(\hat{p}_0+\zeta h_2)(o)\right\|_{\mathbb{R}^d}}{|\zeta|^3 \cdot|\zeta-1|}|d\zeta|}{2 \log(n+1) \pi} d\nu(o) \\
&+\frac{1}{2}\left\|\int_{\mathcal{O}} \mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1+h_2, h_1-h_2](o) p_n d\nu(o)\right\|_{\mathbb{R}^d} \\
\overset{(d)}{\leq}\ & \int_{\mathcal{O}'} \frac{2\pi\left\|\mathcal{D}_f D_\Psi^*(\hat{p}_0)[\Delta h](o)\right\|_{\mathbb{R}^d}+\oint_{|\zeta|=\rho} \frac{|\zeta| \sup\limits_{t \in[0,1]}\left\|\mathcal{D}_f D_\Psi^*\left(\hat{p}_0+\zeta\left(h_2+t(h_1-h_2)\right)\right)(o)\right\|_{\mathbb{R}^d}\left\|h_1-h_2\right\|_{L^2(\nu)}}{|\zeta|^3|\zeta-1|}|d\zeta|}{2 \log(n+1) \pi} d\nu(o) \\
&+\frac{1}{2} \int_{\mathcal{O}}\left\|\mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1+h_2, h_1-h_2](o)\right\|_{\mathbb{R}^d} p_n d\nu(o) \\
\leq\ & \frac{16n}{\log(n+1)} \sup_{h \neq 0} \frac{\left\|\mathcal{D}_f D_\Psi^*\left(\hat{p}_0\right)[h]\right\|_{L^2(\nu) \rightarrow L^2(\nu ; \mathbb{R}^d)}}{\|h\|_{L^2(\nu)}}\|\Delta h\|_{L^2(\nu)}+\frac{1}{2} \int_{\mathcal{O}}\left\|\mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[h_1+h_2, \Delta h](o)\right\|_{\mathbb{R}^d} p_n d\nu(o) \\
&+8n \frac{\sup\left\{\left\|\mathcal{D}_f D_\Psi^*(\hat{p}_0+\tilde{h})\right\|:\|\tilde{h}\|_{L^2(\nu)} \leq \rho \max\{\|h_1\|_{L^2(\nu)},\|h_2\|_{L^2(\nu)}\}\right\}}{\log(n+1) \pi}\left\|h_1-h_2\right\|_{L^2(\nu)} \oint_{|\zeta|=\rho} \frac{1}{|\zeta|^2|\zeta-1|}|d\zeta| \\
\overset{(E.78)}{\leq}\ & \frac{16n}{\log(n+1)} \sup_{h \neq 0} \frac{\left\|\mathcal{D}_f D_\Psi^*\left(\hat{p}_0\right)[h]\right\|_{L^2(\nu) \rightarrow L^2(\nu ; \mathbb{R}^d)}}{\|h\|_{L^2(\nu)}}\|\Delta h\|_{L^2(\nu)}+\int_{\mathcal{O}'}\left\|\mathcal{D}_f^2 D_\Psi^*(\hat{p}_0)[\Delta h](o)\right\|_{\mathbb{R}^d} \frac{\left\|h_1+h_2\right\|_{L^2(\nu)}}{2 \log(n+1)} d\nu(o) \\
&+8n \frac{\sup\left\{\left\|\mathcal{D}_f D_\Psi^*(\hat{p}_0+\tilde{h})\right\|:\|\tilde{h}\|_{L^2(\nu)} \leq \rho \max\{\|h_1\|_{L^2(\nu)},\|h_2\|_{L^2(\nu)}\}\right\}}{\log(n+1) \pi}\left\|h_1-h_2\right\|_{L^2(\nu)} \cdot \frac{2\pi}{\rho(\rho-1)} \\
\leq\ & \frac{16n}{\log(n+1)} \sup_{h \neq 0} \frac{\left\|\mathcal{D}_f D_\Psi^*\left(\hat{p}_0\right)[h]\right\|_{L^2(\nu) \rightarrow L^2(\nu ; \mathbb{R}^d)}}{\|h\|_{L^2(\nu)}}\|\Delta h\|_{L^2(\nu)}+\frac{8n\left\|h_1+h_2\right\|_{L^2(\nu)}}{\log(n+1)} \sup_{h, h' \neq 0} \frac{\left\|\mathcal{D}_f^2 D_\Psi^*\left(\hat{p}_0\right)[h, h']\right\|}{\|h\|_{L^2(\nu)}\|h'\|_{L^2(\nu)}}\|\Delta h\|_{L^2(\nu)} \\
&+16n \frac{\sup_{\|\tilde{h}\| \leq \rho \max\{\|h_1\|,\|h_2\|\}}\left\|\mathcal{D}_f D_\Psi^*(\hat{p}_0+\tilde{h})\right\|_{L^2(\nu) \rightarrow L^2(\nu ; \mathbb{R}^d)}}{\rho(\rho-1) \log(n+1)}\|\Delta h\|_{L^2(\nu)},
\end{aligned}
\tag{E.79}
$$

where step (a) follows from the triangle inequality for Bochner integrals [39], step (b) uses symmetric bilinearity as established by Clairaut's theorem [39], step (c) holds when we naturally constrain $n \geq 1$, and step (d) applies the mean-value theorem for the Fréchet derivative [40]. Noting that

$\|h_1\|_{L^2(\nu)} \leq \|h_2\|_{L^2(\nu)} + \|\Delta h\|_{L^2(\nu)}$, finally we deduce

$$\begin{aligned}
&\left\|\int_{\mathcal{O}} \left(D^*_{\Psi}(\hat{p}_0 + h_1)(o) - D^*_{\Psi}(\hat{p}_0 + h_2)(o)\right) p_n d\nu(o)\right\|_{\mathbb{R}^d} \\
&\quad \leq \frac{16n}{\log(n+1)} \left(C_1 + \frac{C_2}{2}\left(\|h_1\|_{L^2(\nu)} + \|h_2\|_{L^2(\nu)}\right) + \frac{C_3}{\rho(\rho-1)}\right) \left\|(\hat{p}_0 + h_1) - (\hat{p}_0 + h_2)\right\|_{L^2(\nu)},
\end{aligned} \tag{E.80}$$

in which we set the $C_1 = \sup_{h\neq 0} \frac{\|\mathcal{D}_f D^*_{\Psi}(\hat{p}_0)[h]\|}{\|h\|_{L^2(\nu)}}$, $C_2 = \sup_{h,h'\neq 0} \frac{\left\|\mathcal{D}^2_f D^*_{\Psi}(\hat{p}_0)[h,h']\right\|}{\|h\|_{L^2(\nu)}\|h'\|_{L^2(\nu)}}$, and $C_3 = \sup_{\|h\|\leq\rho\max\{\|h_1\|,\|h_2\|\}} \|\mathcal{D}_f D^*_{\Psi}\left(\hat{p}_0 + h\right)\|$. Since $\breve{\mathcal{T}}_{p_0}$ is infinite-dimensional and the range of $\mathcal{D}_f \int_{\mathcal{O}} D^*_{\Psi}\left(\hat{p}_0\right)(o) p_n d\nu(o)$ is finite-dimensional, the rank–nullity theorem of Banach spaces implies that the nullity $\mathcal{K}_0(\hat{p}_0) = \{\hat{h} \in \breve{\mathcal{T}}_{p_0} : \int_{\mathcal{O}} \mathcal{D}_f D^*_{\Psi}(\hat{p}_0)[\hat{h}] p_n d\nu(o) = \mathbf{0}\}$ must be infinite. Since $\|\Delta h\|_{L^2(\nu)} > 0$ can be arbitrarily small, the inequality (E.80) confirms that $\int_{\mathcal{O}} D^*_{\Psi}(\hat{p}_0+h)(o)p_n d\nu(o)$ is locally Lipschitz continuous w.r.t. h near the $\hat{p}_0$, and the conditions for applying the infinite-dimensional implicit function theorem [41] are met. Now it guarantees that the set $\{\hat{h} \in \breve{\mathcal{T}}_{p_0} : \int_{\mathcal{O}} D^*_{\Psi}(\hat{p}_0+\hat{h})(o)p_n d\nu(o) = \mathbf{0}\}$ is a $\mathbb{C}^1$-manifold whose dimension equals that of $\mathcal{K}_0(\hat{p}_0)$. Therefore, in any sufficiently small $L^2(\nu)$-neighborhood of $\hat{p}_0$ there exists an infinite-dimensional manifold (of codimension at most d) w.r.t. perturbations $\hat{h} \in \breve{\mathcal{T}}_{p_0}$ such that $\int_{\mathcal{O}} D^*_{\Psi}(\hat{p}_0 + \hat{h})(o)p_n d\nu(o) = \mathbf{0}$. Thus, the finite d-dimensional condition Eq. (2) defines infinite number of solutions.

The fact that $\min\limits_{\{p\in\mathcal{M}\}} \int_{\mathcal{O}} \mathbf{L}(p)(o)p_n d\nu(o)$ also admits infinitely many solutions follows by a nearly identical argument to the one for the EIF estimating-equation. Under our standing differentiability and surjectivity assumptions, the first-order Fréchet derivative of the loss functional at any minimizer p^* is a surjective linear map onto the cotangent space of $\mathcal{M}$. By the Banach-manifold implicit-function theorem, the vanishing set of that derivative in a neighborhood of p^* is therefore a smooth submanifold of $\mathcal{M}$. Crucially, the kernel of the derivative is infinite-dimensional, so one can flow along any direction in this null-space without changing the loss to first order-producing a local continuum of distinct minimizers. Since every step of this Banach-manifold construction mirrors the classical proof for the EIF estimating-equation above, we omit the redundant details here. □

F Proof of Theorem 4

Proof. Recall from Assumption 5 that the empirical line-search loss admits a Lipschitz continuous gradient at every iterate k. Fix an arbitrary reference point $\varepsilon \in \mathbb{R}^d$, and let $\epsilon \in \mathrm{int}(\mathcal{R})$ be any interior point. Consider the line segment connecting ε and ϵ, along which the following holds

$$\begin{aligned}
\int_{\mathcal{O}} \mathbf{L}\big(p^k_n(\epsilon)\big)(o)p_n d\nu(o) &= \int_{\mathcal{O}} \mathbf{L}\big(p^k_n(\varepsilon + (\epsilon - \varepsilon))\big)(o)p_n d\nu(o) \\
&= \int_{\mathcal{O}} \mathbf{L}\big(p^k_n(\varepsilon)\big)(o)p_n d\nu(o) + \int_0^1 \int_{\mathcal{O}} \nabla_{\epsilon}\mathbf{L}\big(p^k_n(\epsilon)\big)(o)\Big|^{\top}_{\epsilon=\varepsilon+t(\epsilon-\varepsilon)} (\epsilon - \varepsilon)\, p_n d\nu(o)dt.
\end{aligned} \tag{F.81}$$

This expression could be rearranged into the form

$$\begin{aligned}
&\int_{\mathcal{O}} \mathbf{L}\big(p^k_n(\varepsilon)\big)(o)p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}\big(p^k_n(\epsilon)\big)(o)p_n d\nu(o) \\
&= -\int_0^1 \int_{\mathcal{O}} \nabla_{\epsilon}\mathbf{L}\big(p^k_n(\epsilon)\big)(o)\Big|^{\top}_{\epsilon=\varepsilon+t(\epsilon-\varepsilon)} (\epsilon - \varepsilon)\, p_n d\nu(o)dt \\
&= \int_{\mathcal{O}} \nabla_{\epsilon}\mathbf{L}\big(p^k_n(\epsilon)\big)(o)\Big|^{\top}_{\epsilon=\varepsilon} (\varepsilon - \epsilon)\, p_n d\nu(o) \\
&\quad - \int_0^1 \int_{\mathcal{O}} \left(\nabla_{\epsilon}\mathbf{L}\big(p^k_n(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon+t(\epsilon-\varepsilon)} - \nabla_{\epsilon}\mathbf{L}\big(p^k_n(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon}\right)^{\top} (\epsilon - \varepsilon)\, p_n d\nu(o)dt.
\end{aligned} \tag{F.82}$$

Invoking the Cauchy–Schwarz and utilizing the assumed Lipschitz property,

$$\begin{aligned}
&\int_0^1 \int_{\mathcal{O}} \left(\nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon+t(\epsilon-\varepsilon)} - \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon} \right)^\top (\epsilon-\varepsilon)\, p_n d\nu(o) dt \\
&\leq \int_0^1 \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon+t(\epsilon-\varepsilon)} p_n d\nu(o) - \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon} p_n d\nu(o) \right\|_{\mathbb{R}^d} \cdot \left\|\epsilon-\varepsilon\right\|_{\mathbb{R}^d} dt \\
&\leq \int_0^1 \int_{\mathcal{O}} \operatorname*{ess\,sup}_{\epsilon: p_n^k(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^k(\epsilon))(o)\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu(o) \cdot t \left\|\epsilon-\varepsilon\right\|_{\mathbb{R}^d}^2 dt \\
&\leq \frac{1}{2} \left\|\epsilon-\varepsilon\right\|_{\mathbb{R}^d}^2 \sup_{k'\in\mathbb{Z}^+} \operatorname*{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu.
\end{aligned} \tag{F.83}$$

Therefore, for every ϵ, it follows that

$$\begin{aligned}
&\int_{\mathcal{O}} \mathbf{L}\big(p_n^k(\varepsilon)\big)(o) p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}\big(p_n^k(\epsilon)\big)(o) p_n d\nu(o) \\
&\geq \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon}^\top (\varepsilon-\epsilon)\, p_n d\nu(o) - \frac{1}{2}\left\|\epsilon-\varepsilon\right\|_{\mathbb{R}^d}^2 \sup_{k'\in\mathbb{Z}^+} \operatorname*{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu.
\end{aligned} \tag{F.84}$$

Followed by the `TMLE` updates, taking the minimum over ϵ on the both sides of (F.84) yields

$$\begin{aligned}
&\min_\epsilon \left[\int_{\mathcal{O}} \mathbf{L}\big(p_n^k(\varepsilon)\big)(o) p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}\big(p_n^k(\epsilon)\big)(o) p_n d\nu(o) \right] \\
&= \sup_{\mathbf{u}\in\mathbb{R}^d} \left[\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon}^\top \mathbf{u}\, p_n d\nu(o) - \frac{1}{2} \sup_{k'\in\mathbb{Z}^+} \operatorname*{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu \cdot \left\|\mathbf{u}\right\|_{\mathbb{R}^d}^2 \right]_{\mathbf{u}:=\varepsilon-\epsilon}.
\end{aligned} \tag{F.85}$$

On the right-hand side of (F.85), the coefficient in front of $\|\mathbf{u}\|^2$ is strictly negative. This implies that the expression represents a concave quadratic form, which attains its supremum at a unique stationary point $\mathbf{u}^*$, explicitly characterized by

$$\begin{aligned}
&\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon} p_n d\nu(o) - \sup_{k'\in\mathbb{Z}^+} \operatorname*{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu \cdot \mathbf{u}^* = \mathbf{0}, \\
&\qquad \implies \quad \mathbf{u}^* = \frac{\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\big|_{\epsilon=\varepsilon} p_n d\nu(o)}{\sup_{k'\in\mathbb{Z}^+} \operatorname{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu}.
\end{aligned} \tag{F.86}$$

Consequently,

$$\begin{aligned}
&\min_\epsilon \left[\int_{\mathcal{O}} \mathbf{L}\big(p_n^k(\varepsilon)\big)(o) p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}\big(p_n^k(\epsilon)\big)(o) p_n d\nu(o) \right] \\
&\geq \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon}^\top \mathbf{u}^*\, p_n d\nu(o) - \frac{1}{2} \sup_{k'\in\mathbb{Z}^+} \operatorname*{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu \cdot \left\|\mathbf{u}^*\right\|_{\mathbb{R}^d}^2 \\
&\geq \frac{\left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\big|_{\epsilon=\varepsilon} p_n d\nu(o) \right\|^2}{\sup_{k'\in\mathbb{Z}^+} \operatorname{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu} - \frac{1}{2} \sup_{k'\in\mathbb{Z}^+} \operatorname*{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu \\
&\qquad \cdot \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon} p_n d\nu(o) \right\|_{\mathbb{R}^d}^2 \left(\sup_{k'\in\mathbb{Z}^+} \operatorname*{ess\,sup}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu \right)^{-2} \\
&= \frac{1}{2} \inf_{k'\in\mathbb{Z}^+} \operatorname*{ess\,inf}_{\epsilon: p_n^{k'}(\epsilon)\in\mathcal{M}} \left(\sup_{\mathbf{v}\neq\mathbf{0}} \int_{\mathcal{O}} \frac{\mathbf{v}^\top \nabla_\epsilon^2 \mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2} p_n d\nu \right)^{-1} \left\| \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\varepsilon} p_n d\nu(o) \right\|_{\mathbb{R}^d}^2.
\end{aligned} \tag{F.87}$$

Similarly, invoking the gradient metric-subregularity at $\epsilon = \mathbf{0}$ leads to

$$\begin{aligned} \left\|\epsilon_n^k\right\| &\leq \sup_{\substack{\epsilon\in\mathcal{R}\\ \epsilon\neq\mathbf{0}}} \frac{\|\epsilon\|}{\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^k(\epsilon))p_n d\nu\right\|} \left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\epsilon_n^k} p_n d\nu(o) - \int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\mathbf{0}} p_n d\nu(o)\right\|_{\mathbb{R}^d} \\ &\leq \sup_{k'\in\mathbb{Z}^+}\sup_{\substack{\epsilon\in\mathcal{R}\\ \epsilon\neq\mathbf{0}}} \frac{\|\epsilon\|}{\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^{k'}(\epsilon))p_n d\nu\right\|} \left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}\big(p_n^k(\epsilon)\big)(o)\Big|_{\epsilon=\mathbf{0}} p_n d\nu(o)\right\|_{\mathbb{R}^d}. \end{aligned} \tag{F.88}$$

Combining the results of (F.87) and (F.88), we arrive at

$$\begin{aligned} &\int_{\mathcal{O}}\mathbf{L}(p_n^k)(o)p_n d\nu(o) - \int_{\mathcal{O}}\mathbf{L}(p_n^{k+1})(o)p_n d\nu(o) \\ &\geq \frac{1}{2}\inf_{(k',k'')\in\mathbb{Z}^+\times\mathbb{Z}^+}\left(\sup_{\substack{\epsilon\in\mathcal{R}\\ \epsilon\neq\mathbf{0}}}\frac{\|\epsilon\|}{\left\|\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^{k''}(\epsilon))p_n d\nu\right\|}\sqrt{\operatorname*{ess\,sup}_{\epsilon:p_n^{k'}(\epsilon)\in\mathcal{M}}\sup_{\mathbf{v}\neq\mathbf{0}}\int_{\mathcal{O}}\frac{\mathbf{v}^\top\nabla_\epsilon^2\mathbf{L}(p_n^{k'}(\epsilon))\mathbf{v}}{\|\mathbf{v}\|^2}p_n d\nu}\right)^{-2}\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \\ &\triangleq \spadesuit\cdot\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2. \end{aligned} \tag{F.89}$$

Summing (F.89) over $k = 0$ to $K - 1$ and taking the limit results in

$$\begin{aligned} &\sum_{k=0}^{\infty}\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 = \lim_{K\to\infty}\sum_{k=0}^{K-1}\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}^2 \\ &\leq \lim_{K\to\infty}\spadesuit^{-1}\cdot\left(\int_{\mathcal{O}}\mathbf{L}(p_n^0)(o)p_n d\nu(o) - \int_{\mathcal{O}}\mathbf{L}(p_n^1)(o)p_n d\nu(o) + \ldots + \int_{\mathcal{O}}\mathbf{L}(p_n^{k-1})(o)p_n d\nu(o) - \int_{\mathcal{O}}\mathbf{L}(p_n^K)(o)p_n d\nu(o)\right) \\ &= \spadesuit^{-1}\cdot\left(\int_{\mathcal{O}}\mathbf{L}(p_n^0)(o)p_n d\nu(o) - \underbrace{\lim_{K\to\infty}\int_{\mathcal{O}}\mathbf{L}(p_n^K)(o)p_n d\nu(o)\Big)}_{\text{bounded by Lemma 1}} < +\infty. \end{aligned} \tag{F.90}$$

Define the partial sum sequence $s_K := \sum_{k=0}^{K}\left\|\epsilon_n^k\right\|^2$. From (F.90), it follows that $\{s_K\}$ converges in $\mathbb{R}^1$ to some limit $S_\infty < \infty$. Fix an arbitrary $\delta^\dagger > 0$. Since $s_K \rightsquigarrow S_\infty$, there exists a threshold index $K' \in \mathbb{Z}^+$ such that for all $K \geq K'$, the following bounds hold

$$|s_K - S_\infty| < \frac{1}{2}\delta^\dagger \quad \text{and} \quad |s_{K-1} - S_\infty| < \frac{1}{2}\delta^\dagger. \tag{F.91}$$

For any $K \geq K'$, observe that $\left\|\epsilon_n^K\right\|^2 = s_K - s_{K-1} = (s_K - S_\infty) - (s_{K-1} - S_\infty)$. Applying the triangle inequality yields

$$\left\|\epsilon_n^K\right\|_{\mathbb{R}^d}^2 \leq |s_K - S_\infty| + |s_{K-1} - S_\infty| < \frac{1}{2}\delta^\dagger + \frac{1}{2}\delta^\dagger = \delta^\dagger. \tag{F.92}$$

Since $\delta^\dagger > 0$ was arbitrary, it follows that $\lim_{k\to\infty}\left\|\epsilon_n^k\right\|^2 = 0$, and hence

$$\left\|\lim_{k\to\infty}\epsilon_n^k\right\| = \sqrt{\lim_{k\to\infty}\|\epsilon_n^k\|_{\mathbb{R}^d}^2} = 0. \tag{F.93}$$

Invoking the complete mathematical equivalence established in the Theorem 1 finishes the proof. □

G Proof of Theorem 5

Proof of Case (i). Consider an arbitrary $\epsilon \in \mathbb{R}^d$. The corresponding risk difference can be expressed as

$$\int_{\mathcal{O}}\mathbf{L}(p_n^0(\epsilon))(o)p_n d\nu(o) - \int_{\mathcal{O}}\mathbf{L}(p_n^0(\mathbf{0}))(o)p_n d\nu(o)$$

$$
\begin{aligned}
&= \int_0^1 \int_{\mathcal{O}} \left[\nabla_\epsilon \mathbf{L}(p_n^0(s\epsilon))(o)\right]^\top p_n d\nu(o)\, \epsilon\, ds \\
&= \int_0^1 \int_{\mathcal{O}} p_n \left[\nabla_\epsilon \mathbf{L}(p_n^0(s\epsilon))(o) - \nabla_\epsilon \mathbf{L}(p_n^0(\epsilon))(o)\big|_{\epsilon=\mathbf{0}}\right]^\top \epsilon\, d\nu(o)\, ds \\
&= \int_0^1 \int_{\mathcal{O}} p_n \int_0^s \frac{d}{dt} \left[\nabla_\epsilon \mathbf{L}(p_n^0(t\epsilon))(o)\right]^\top \epsilon\, dt\, d\nu(o)\, ds \qquad \text{(G.94)} \\
&= \int_{\mathcal{O}} p_n \int_0^1 \int_0^s \epsilon^\top \nabla_\epsilon^2 \mathbf{L}(p_n^0(t\epsilon))(o) \epsilon\, dt\, ds\, d\nu(o) \\
&= \int_{\mathcal{O}} p_n \int_0^1 \int_t^1 \epsilon^\top \nabla_\epsilon^2 \mathbf{L}(p_n^0(t\epsilon))(o) \epsilon\, ds\, dt\, d\nu(o),
\end{aligned}
$$

where the final equality follows by an application of Fubini's Theorem. Since the integrand in (G.94) is continuous and non-negative, the following chain of inequalities holds

$$
\begin{aligned}
\int_{\mathcal{O}} \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}(p_n^0(\mathbf{0}))(o) p_n d\nu(o) &= \int_{\mathcal{O}} p_n \int_0^1 (1-t)\, \epsilon^\top \nabla_\epsilon^2 \mathbf{L}(p_n^0(t\epsilon))(o) \epsilon\, dt\, d\nu(o) \\
&\geq \int_{\mathcal{O}} p_n \int_0^1 (1-t) \left[\inf_{0\leq s\leq 1} \epsilon^\top \nabla_\epsilon^2 \mathbf{L}(p_n^0(s\epsilon))(o)\epsilon\right] dt\, d\nu(o) \\
&= \frac{1}{2} \inf_{0\leq s\leq 1} \int_{\mathcal{O}} p_n \epsilon^\top \nabla_\epsilon^2 \mathbf{L}(p_n^0(s\epsilon))(o) \epsilon\, d\nu(o) \; > 0,
\end{aligned}
\tag{G.95}
$$

where the last inequality holds when $\epsilon \neq \mathbf{0}$. It exactly states that $\epsilon \mapsto \int_{\mathcal{O}} \mathbf{L}(p_n^0(\epsilon)) p_n d\nu$ achieves its global minimum at a single point, namely $\epsilon = \mathbf{0}$. Therefore the TMLE update must satisfy

$$
\{\mathbf{0}\} = \underset{\{\epsilon : p_n^0(\epsilon) \in \mathcal{M}\}}{\arg\min} \int_{\mathcal{O}} \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o). \tag{G.96}
$$

By the definition of update rule, the first TMLE iterate is given by $p_n^1 = p_n^0\left(\epsilon_n^0\right) = p_n^0(\mathbf{0}) = p_n^0$. Consequently, the TMLE terminates after its first fluctuation step, having already satisfied the EIF estimating-equation. □

Proof of Case (ii). Let $D^*_{\Psi,j}(p_n^0(u))(o)$ denote the j-th component of the vector $D^*_\Psi(p_n^0(u))(o)$. By our curl–freeness assumption we have $\partial_{\epsilon_i} D^*_{\Psi,j}(p_n^0(\epsilon)) = \partial_{\epsilon_j} D^*_{\Psi,i}(p_n^0(\epsilon))$. We consider the 1-form $\sum_{j=1}^d D^*_{\Psi,j}(p_n^0(\epsilon))(o) \mathrm{d}\epsilon_j$, which is closed since its exterior derivative vanishes as

$$
\mathrm{d}\left(\sum_{j=1}^d D^*_{\Psi,j}(p_n^0(\epsilon))(o) \mathrm{d}\epsilon_j\right) = \sum_{i<j} \Big(\partial_{\epsilon_i} D^*_{\Psi,j}(p_n^0(\epsilon))(o) - \partial_{\epsilon_j} D^*_{\Psi,i}(p_n^0(\epsilon))(o)\Big)\, \mathrm{d}\epsilon_i \wedge \mathrm{d}\epsilon_j = 0. \tag{G.97}
$$

The Poincaré's lemma guarantees that on any simply connected region $\mathcal{R} \subset \mathbb{R}^d$, every closed differential form (G.97) is exact. In the 1-form case, this means the field admits a global scalar potential $\Phi(\epsilon, o)$ such that

$$
\mathrm{d}\Phi(\epsilon, o) = \sum_{j=1}^d D^*_{\Psi,j}(p_n^0(\epsilon))(o) \mathrm{d}\epsilon_j, \tag{G.98}
$$

which indicates $\nabla_\epsilon \Phi(\epsilon, o) = D^*_\Psi(p_n^0(\epsilon))(o)$ and $\Phi(\mathbf{0}, o) = 0$. Now the condition provided in Case (ii) of Theorem 5 is

$$
\begin{aligned}
\mathbf{L}(p_n^0(\epsilon))(o) &= \mathbf{L}(p_n^0(\mathbf{0}))(o) + A \cdot \Phi(\epsilon, o) \\
&= \mathbf{L}(p_n^0(\mathbf{0}))(o) + A \int_0^1 \left[D^*_\Psi(p_n^0(t\epsilon))(o)\right]^\top \epsilon\, dt \\
&= \mathbf{L}(p_n^0(\mathbf{0}))(o) + A \cdot \sum_{j=1}^d \epsilon_j \int_0^1 D^*_{\Psi,j}(p_n^0(t\epsilon))(o) dt.
\end{aligned}
\tag{G.99}
$$

We differentiate (G.99) with respect to ϵ_k.[14] By the dominated convergence theorem, it follows that

$$\begin{aligned}
\partial_{\epsilon_k}\mathbf{L}(p_n^0(\epsilon))(o) &= 0 + A\partial_{\epsilon_k}\left[\sum_{j=1}^{d}\epsilon_j\int_0^1 D^*_{\Psi,j}(p_n^0(t\epsilon))(o)dt\right] \\
&= A\left[\int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt + \sum_{j=1}^{d}\epsilon_j\int_0^1 \partial_{\epsilon_k}D^*_{\Psi,j}(p_n^0(t\epsilon))(o)dt\right] \\
&\overset{(a)}{=} A\left[\int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt + \sum_{j=1}^{d}\epsilon_j\int_0^1 \sum_{\ell=1}^{d}\partial_{u_\ell}D^*_{\Psi,j}(p_n^0(u))(o)\Big|_{u=t\epsilon}\cdot t\delta_{\ell k}\,dt\right] \\
&= A\left[\int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt + \int_0^1 t\sum_{j=1}^{d}\epsilon_j\partial_{u_k}D^*_{\Psi,j}(p_n^0(t\epsilon))(o)dt\right] \\
&\overset{(b)}{=} A\left[\int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt + \int_0^1 t\sum_{j=1}^{d}\epsilon_j\partial_{u_j}D^*_{\Psi,k}(p_n^0(u))(o)dt\right] \\
&= A\left[\int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt + \int_0^1 t\frac{\mathrm{d}}{\mathrm{d}t}D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt\right],
\end{aligned} \tag{G.100}$$

where step (a) involves the Kronecker delta $\delta_{\ell k}$ defined as $\delta_{\ell k} := \begin{cases}1, & \ell = k \\ 0, & \ell \neq k\end{cases}$, and step (b) utilizes the curl–freeness assumption. Applying integration by parts with respect to t, we obtain

$$\begin{aligned}
\int_0^1 t\frac{\mathrm{d}}{\mathrm{d}t}D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt &= \Big[tD^*_{\Psi,k}(p_n^0(t\epsilon)(o)\Big]_{t=0}^{t=1} - \int_0^1 1\cdot D^*_{\Psi,k}(p_n^0(t\epsilon)(o)dt \\
&= D^*_{\Psi,k}(p_n^0(\epsilon)(o) - \int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon)(o)dt.
\end{aligned} \tag{G.101}$$

Substituting (G.101) into (G.100) yields

$$\begin{aligned}
\partial_{\epsilon_k}\mathbf{L}(p_n^0(\epsilon))(o) &= A\left[\int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon))(o)dt + D^*_{\Psi,k}(p_n^0(\epsilon)(o) - \int_0^1 D^*_{\Psi,k}(p_n^0(t\epsilon)(o)dt\right] \\
&= A\cdot D^*_{\Psi,k}(p_n^0(\epsilon)(o) \equiv A\partial_{\epsilon_k}\Phi(\epsilon, o).
\end{aligned} \tag{G.102}$$

Consequently, it follows that $AD^*_\Psi(p)(o) \equiv \dfrac{\mathrm{d}\mathbf{L}(p(\epsilon))(o)}{\mathrm{d}\epsilon}$ for each o and all ϵ. Adopting the same techniques utilized in the derivation of (D.50), we obtain the bound as

$$\left\|\nabla_\epsilon\mathbf{L}(p_n^0(\epsilon_1)) - \nabla_\epsilon\mathbf{L}(p_n^0(\epsilon_2))\right\|_2 \leq \|A\|_{\mathrm{op}}\sup_{\{\epsilon: p_n^0(\epsilon)\in\mathcal{M}\}}\left\|\int_{\mathcal{O}}\nabla_\epsilon D^*_\Psi(p(\epsilon))(o)p_n d\nu(o)\right\|_{\mathrm{op}}\cdot\|\epsilon_1 - \epsilon_2\|_2\,. \tag{G.103}$$

The continuity implied by (G.103), together with the assumption that $\mathcal{R}$ is simply connected, ensures the empirical loss attains its minimum at the first iteration. Let $\epsilon_n^0 \in \mathrm{int}\,(\mathcal{R})$ denote any minimiser returned by the solver. Then the optimality condition tells

$$\int_{\mathcal{O}}\nabla_\epsilon\mathbf{L}(p_n^0(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\epsilon_n^0} = \int_{\mathcal{O}}D^*_\Psi(p_n^*)(o)p_n d\nu(o) = \mathbf{0}. \tag{G.104}$$

Thus, the first `TMLE` fluctuation already solves the empirical EIF estimating-equation (2). □

Remark on Proof of Case (iii). We do not provide an explicit proof of this claim here. For treatments of particular TMLE formulations, the reader is referred to their original papers, which typically include detailed, problem-specific statistical analyses and numerical validations. Examples include

[14] The k is used as a generic index rather than iteration numbers, since we focus on the first few updates only.

Rosenblum and Van Der Laan [42], Schnitzer et al. [43], D´iaz and Rosenblum [25], Rytgaard and van der Laan [44], etc. A comprehensive and systematic characterization of these conditions remains highly challenging and is therefore left to future work. □

Proof of Case (iv). Without loss of generality, we discuss several commonly employed stopping rules of `TMLE`. Fix a tolerance $\delta_n > 0$, in practice we declare one-step convergence the moment we find

$$\left\|\epsilon_n^k\right\|_{\mathbb{R}^d} \leq \delta_n \quad \text{or} \quad \left\|\int_{\mathcal{O}} D_\Psi^*(p_n^k) p_n d\nu\right\|_{\mathbb{R}^d} \leq \delta_n, \tag{G.105}$$

whichever we prefer to monitor. By Definition 2 we know that the ϵ is a valid parametric index of probability submodel. Hence $\epsilon \in \mathcal{R} \subsetneq \mathbb{R}^d$ and the stepsize $\left\|\epsilon_n^k\right\|_{\mathbb{R}^d}$ is always bounded above. In the first stopping rule if one set $\delta_n \in [\sup_{\epsilon\in\mathcal{R}} \|\epsilon\|, \infty)$, the `TMLE` is guaranteed to converge in the first iteration. And if $\delta_n \in (0, \sup_{\epsilon\in\mathcal{R}} \|\epsilon\|)$ we have the probability to make `TMLE` converge in one step. On the another side, by design we know that the targeting step minimizes the empirical loss at $k = 1$. First-order expansion gives

$$\begin{aligned}
&\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o)\bigg|_{\epsilon=\epsilon_n^0} - \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o)\bigg|_{\epsilon=\mathbf{0}} \\
&= \left\langle \int_{\mathcal{O}} \nabla_\epsilon^2 \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o)\bigg|_{\epsilon=\xi_0}, \epsilon_n^0 - \mathbf{0} \right\rangle,
\end{aligned} \tag{G.106}$$

where $\mathbf{0} \preceq \xi_0 \preceq \epsilon_n^0$. We can estimate the magnitude of $\int_{\mathcal{O}} D_\Psi^*(p_n^0(\epsilon_n^0)) p_n d\nu$ as

$$\begin{aligned}
\left\|\int_{\mathcal{O}} D_\Psi^*(p_n^0(\epsilon_n^0))(o) p_n d\nu(o)\right\|_{\mathbb{R}^d} &= \left\|\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o)\bigg|_{\epsilon=\mathbf{0}}\right\|_{\mathbb{R}^d} \\
&\leq \left\|\int_{\mathcal{O}} \nabla_\epsilon^2 \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o)\bigg|_{\epsilon=\xi_0}^\top \epsilon_n^0\right\|_{\mathbb{R}^d} \\
&\leq \left\|\epsilon_n^0\right\|_{\mathbb{R}^d} \max_{(j,l)\in[1,d]^2\cap\mathbb{Z}^2} \sup_{t\in[0,1]} \left|\frac{\partial^2}{\partial\epsilon_j\partial\epsilon_l} \int_{\mathcal{O}} \mathbf{L}(p_n^0(\epsilon))(o) p_n d\nu(o)\bigg|_{\epsilon=t\epsilon_n^0}\right| \\
&\triangleq \check{\delta_n} < +\infty.
\end{aligned} \tag{G.107}$$

Similarly, if one set $\delta_n \in [\check{\delta_n}, \infty)$, the `TMLE` is guaranteed to converge in the first iteration. And if $\delta_n \in (0, \check{\delta_n})$ we have the chances to make `TMLE` converge in one step. In summary, it is theoretically possible to identify such a threshold $\delta_n > 0$ for which the stopping conditions ensuring convergence of `TMLE` in a single step are satisfied.

Practitioners may also explicitly set $k_{\max} := 1$ for various practical reasons, such as licensing restrictions of commercial solvers, limited computational time or memory constraints. Under such setting the `TMLE` algorithm will evidently terminate after a single iteration step. □

H Proof of Theorem 6

Lemma 4 (Hoeffding's Inequality [45])**.** *Let $X_1, \ldots, X_n$ be independent random variables with mean μ_n and almost sure bounds $a_i \leq X_i \leq b_i$, $i = 1, \ldots, n$. Write $S_n := \sum_{i=1}^n X_i$. Then for any $t > 0$,*

$$\mathbb{P}\left(\left|S_n - \mu_n\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

We now proceed to establish Theorem 6.

Proof. Under the Assumption (*ii*), for all $\|\epsilon\|_2$ in a fixed neighborhood of 0 we can have

$$\int_{\mathcal{O}} D_\Psi^*(p_n^k(\epsilon))(o) p_n d\nu(o) = \int_{\mathcal{O}} D_\Psi^*(p_n^k(\mathbf{0}))(o) p_n d\nu(o) + \int_{\mathcal{O}} \nabla_\epsilon D_\Psi^*(p_n^k(\epsilon))(o)\big|_{\epsilon=\mathbf{0}} \epsilon p_n d\nu(o) + R_o(\epsilon). \tag{H.108}$$

Define $r_o = \dfrac{\lambda_o}{4L_o}$, and for $\|\epsilon\|_2 \leq r_o$ there exists a constant L_o such that the remainder term satisfies $\|R_o(\epsilon)\|_2 \leq L_o \|\epsilon\|_2^2$. Within the ball $\{\|\epsilon\|_2 \leq r_o\}$, one verifies that the

$$\begin{aligned}\epsilon^\dagger = &- \left[\int_{\mathcal{O}} \nabla_\epsilon D^*_\Psi(p_n^k(\epsilon))(o)\big|_{\epsilon=\mathbf{0}}\, p_n d\nu(o)\right]^{-1} \int_{\mathcal{O}} D^*_\Psi(p_n^k)(o) p_n d\nu(o) \\ &- \left[\int_{\mathcal{O}} \nabla_\epsilon D^*_\Psi(p_n^k(\epsilon))(o)\big|_{\epsilon=\mathbf{0}}\, p_n d\nu(o)\right]^{-1} R_o\left(\epsilon^\dagger\right) \end{aligned} \tag{H.109}$$

is a contraction (with finite Lipschitz constant) and maps the ball into itself. By the Banach fixed-point theorem, this ensures the existence and uniqueness of a fixed point $\epsilon^\dagger$ within the ball that solves (2). We obtain

$$\begin{aligned}\left\|\epsilon^\dagger\right\|_2 &\lesssim \frac{1}{\lambda_o} \cdot \left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o) p_n d\nu(o)\right\|_2 + \frac{1}{\lambda_o} \cdot \left\|R_o\left(\epsilon^\dagger\right)\right\|_2 \\ &\lesssim \frac{1}{\lambda_o}\left[\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o) p_n d\nu(o)\right\|_2 + L_o \left\|\epsilon^\dagger\right\|_2^2\right] \\ &\overset{(a)}{\lesssim} \frac{1}{\lambda_o}\left[\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o) p_n d\nu(o)\right\|_2 + \frac{\lambda_o}{4}\left\|\epsilon^\dagger\right\|_2\right], \end{aligned} \tag{H.110}$$

where step (a) makes use of the condition $\|\epsilon\|_2 \leq \dfrac{\lambda_o}{4L_o}$. Solving the resulting inequality (H.110) gives

$$\left\|\epsilon^\dagger\right\|_2 \lesssim \frac{4}{3\lambda_o}\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o) p_n d\nu(o)\right\|_2. \tag{H.111}$$

Without loss of generality, we assume that the solution to subproblem is analytically subject to one-step Newton–Kantorovich iterate

$$\epsilon_n^k \mathrel{\hat{=}} - \left[\int_{\mathcal{O}} \nabla_\epsilon D^*_\Psi(p_n^k(\epsilon))(o)\big|_{\epsilon=\mathbf{0}}\, p_n d\nu(o)\right]^{-1} \int_{\mathcal{O}} D^*_\Psi(p_n^k(\mathbf{0}))(o) p_n d\nu(o). \tag{H.112}$$

A parallel second-order expansion of the empirical loss reveals that the `TMLE` update satisfies

$$\epsilon_n^k - \epsilon^\dagger \mathrel{\hat{=}} \left[\int_{\mathcal{O}} \nabla_\epsilon D^*_\Psi(p_n^k(\epsilon))(o)\big|_{\epsilon=\mathbf{0}}\, p_n d\nu(o)\right]^{-1} R_o\left(\epsilon^\dagger\right). \tag{H.113}$$

On the overshoot event $o.s. \triangleq \left\{\left\|\epsilon_n^k\right\|_2 > \left\|\epsilon^\dagger\right\|_2\right\}$, combine (H.113) with the reverse triangle inequality to get

$$\begin{aligned}\left\|\epsilon_n^k\right\|_2 - \left\|\epsilon^\dagger\right\|_2 \leq \left\|\epsilon_n^k - \epsilon^\dagger\right\|_2 &\lesssim \frac{1}{\lambda_o} \cdot \left\|R_o\left(\epsilon^\dagger\right)\right\|_2 \\ &\lesssim \frac{1}{\lambda_o}\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o) p_n d\nu(o)\right\|_2. \end{aligned} \tag{H.114}$$

Thus, we have

$$o.s. \subseteq \left\{\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o) p_n d\nu(o)\right\|_\infty \geq \frac{\lambda}{\sqrt{d}} \underbrace{\left(\left\|\epsilon_n^k\right\|_2 - \left\|\epsilon^\dagger\right\|_2\right)}_{>0 \textit{ and data-dependent}}\right\}. \tag{H.115}$$

Fix any non-random threshold $\tau > 0$. We split the probability as

$$\mathbb{P}\left[o.s.\right] \leq \underbrace{\mathbb{P}\left(\{o.s.\} \cap \left\{\|\epsilon_n^k\|_2 - \|\epsilon^\dagger\|_2 \leq \tau\right\}\right)}_{\text{small jump}} + \underbrace{\mathbb{P}\left(\{o.s.\} \cap \left\{\|\epsilon_n^k\|_2 - \left\|\epsilon^\dagger\right\|_2 > \tau\right\}\right)}_{\text{large jump}}. \tag{H.116}$$

On the small-jump term, we observe that

$$\{o.s.\} \cap \left\{\left\|\epsilon_n^k\right\|_2 - \left\|\epsilon^\dagger\right\|_2 \leq \tau\right\} \subseteq \left\{\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k) p_n d\nu\right\|_\infty \geq \frac{\lambda_o \tau}{\sqrt{d}}\right\}. \tag{H.117}$$

Note the $\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)p_n d\nu\right\|_\infty \ge a$ implies $\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)p_n d\nu - \tilde{\mu}\right\|_\infty \ge a/2$ or $\|\tilde{\mu}\|_\infty \ge a/2$. Hence with $a := \frac{\lambda_o \tau}{\sqrt{d}}$ we have

$$\mathbb{P}\left(\{o.s.\} \cap \left\{\left\|\epsilon_n^k\right\|_2 - \left\|\epsilon^\dagger\right\|_2 \le \tau\right\}\right) \lesssim \mathbb{P}\left(\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)p_n d\nu - \tilde{\mu}\right\|_\infty \ge \frac{\lambda_o \tau}{2\sqrt{d}}\right) + \mathbf{1}\left\{\|\tilde{\mu}\|_\infty \ge \frac{\lambda_o \tau}{2\sqrt{d}}\right\}. \tag{H.118}$$

Define $B_o := \sup_{o \in \mathcal{O}} \left\|D^*_\Psi(p_n^k, o)\right\|_\infty$. Given that each O_i $(i = 1, \dots, n)$ is an independent and identically distributed realization of the random variable o, and all mappings involved are deterministic functions of oc, we apply Hoeffding's inequality (cf. Lemma 4) and the union bound to yield

$$\mathbb{P}\left(\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)p_n d\nu - \tilde{\mu}\right\|_\infty \ge t\right) \lesssim 2d \exp\left(-\frac{nt^2}{2B_o^2}\right), \tag{H.119}$$

where we substitute $t = \lambda_o \tau / (2\sqrt{d})$ to bound the first term. Turning to the large-jump term, we know that if $\|\epsilon_n^k\|_2 - \|\epsilon^\dagger\|_2 > \tau$, then $\|\epsilon_n^k\|_2 > \tau$. Invoking (H.111) and (H.114), one gets

$$\begin{aligned} \left\|\epsilon_n^k\right\|_2 &\le \left\|\epsilon^\dagger\right\|_2 + \left\|\epsilon_n^k - \epsilon^\dagger\right\|_2 \\ &\lesssim \frac{4}{3\lambda_o}\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o)p_n d\nu(o)\right\|_2 + \frac{L_o}{\lambda_o}\left(\frac{4}{3\lambda_o}\right)^2 \left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o)p_n d\nu(o)\right\|_2^2 \\ &\lesssim \underbrace{\left(\frac{4}{3\lambda_o} + \frac{16 B_o L_o}{9\lambda_o^3}\right)}_{\triangleq C_o} \left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)(o)p_n d\nu(o)\right\|_2. \end{aligned} \tag{H.120}$$

We derive that

$$\left\{\left\|\epsilon_n^k\right\|_2 - \left\|\epsilon^\dagger\right\|_2 > \tau\right\} \subseteq \left\{\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)p_n d\nu\right\|_\infty \ge \frac{\tau}{C_o\sqrt{d}}\right\}. \tag{H.121}$$

Arguing as before, applying a second Hoeffding with $t = \tau/(2C_o\sqrt{d})$ gives

$$\begin{aligned} \mathbb{P}\left(\{o.s.\} \cap \left\{\left\|\epsilon_n^k\right\|_2 - \left\|\epsilon^\dagger\right\|_2 > \tau\right\}\right) &\lesssim \mathbb{P}\left(\left\|\int_{\mathcal{O}} D^*_\Psi(p_n^k)p_n d\nu - \tilde{\mu}\right\|_\infty \ge \frac{\tau}{2C_o\sqrt{d}}\right) + \mathbf{1}\left\{\|\tilde{\mu}\|_\infty \ge \frac{\tau}{2C_o\sqrt{d}}\right\} \\ &\lesssim 2d \exp\left(-\frac{n\tau^2}{8C_o^2 d B_o^2}\right) + \mathbf{1}\left\{\|\tilde{\mu}\|_\infty \ge \frac{\tau}{2C_o\sqrt{d}}\right\}. \end{aligned} \tag{H.122}$$

Taking the magnitude as $\tau = 2C_o\sqrt{d}\|\tilde{\mu}\|_\infty$, then we see that both deterministic indicators $\mathbf{1}\left\{\|\tilde{\mu}\|_\infty \ge \frac{\lambda_o\tau}{2\sqrt{d}}\right\}$ and $\mathbf{1}\left\{\|\tilde{\mu}\|_\infty \ge \frac{\tau}{2C_o\sqrt{d}}\right\}$ vanish. Combining the exponential bounds in (H.119) and (H.122) yields

$$\mathbb{P}\left[\text{o.s.}\right] \lesssim 2d \exp\left(-\frac{n\lambda_o^2 C_o^2 \tilde{\mu}}{2B_o^2}\right) + 2d \exp\left(-\frac{n\tilde{\mu}}{2B_o^2}\right). \tag{H.123}$$

□

Remark 5. *The concentration bound derived above extends naturally to all iterations. In particular, by applying a union bound across the sequence of* `TMLE` *updates, one obtains a uniform probabilistic guarantee that holds simultaneously over the entire run.*

I Non-convergence of `TMLE`

We have already provided a counterexample demonstrating the non-convergence of `TMLE` in Remark 4. Here, we present an additional example illustrating failure of convergence. Note that this scenario lies outside the scope of Assumption 3, 4, and 5.

Definition I.12 (κ-strong Quasi-Convexity [46])**.** *A differentiable function $f(x)$ is strongly quasi-convex about $x = x^*$ if the inequality*

$$f\left(x^*\right) \geq f\left(x\right) + \nabla_x f\left(x\right)^\top \left(x^* - x\right) + \frac{\kappa}{2}\left\|x^* - x\right\|^2 \tag{I.124}$$

holds for some $\kappa > 0$ and all $x \in dom(f)$.

Example 4 (Failure of Convergence)**.** *Suppose the empirical risk $\int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)$ is κ-strongly quasi-convex about $\epsilon = \mathbf{0}$, and has Lipschitz continuous gradient in ϵ. If the curvature constant κ is on the order of $\kappa \gtrsim \frac{d}{n\spadesuit}$, where $\spadesuit$ is defined in (F.89). Then the sequence of fluctuation updates $\left\{\epsilon_n^k\right\}$ produced by Algorithm 1 does not converge to any limit in $\mathcal{R}$.*

Proof Sketch. By the algorithmic construction of the TMLE procedure, it follows that

$$\sum_{k=2}^{K}\left(\int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k-1}))(o)p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k}))(o)p_n d\nu(o)\right) \geq 0. \tag{I.125}$$

Utilizing the Lipschitz continuity of the gradient, one obtains the estimate

$$\begin{aligned}\left\|\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\epsilon_n^{k-1}}\right\| &= \left\|\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\epsilon_n^{k}} - \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\epsilon_n^{k-1}}\right\| \\ &\lesssim \spadesuit^{-1}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|.\end{aligned} \tag{I.126}$$

Returning to (I.125), this yields

$$\begin{aligned}&\sum_{k=2}^{K}\left(\int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k}))(o)p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k-1}))(o)p_n d\nu(o)\right) \\ \overset{(I.124)}{\geq} &\sum_{k=2}^{K}\left\langle \int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\epsilon_n^{k-1}}, \epsilon_n^k - \epsilon_n^{k-1}\right\rangle + \frac{\kappa}{2}\sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2 \\ \gtrsim &-\sqrt{2}d\sum_{k=2}^{K}\left\|\int_{\mathcal{O}} \nabla_\epsilon \mathbf{L}(p_n^k(\epsilon))(o)p_n d\nu(o)\Big|_{\epsilon=\epsilon_n^{k-1}}\right\|\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\| + \frac{\kappa}{2}\sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2 \\ \overset{(I.126)}{\gtrsim} &-\frac{1+2\sqrt{2}d}{n\spadesuit}\sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2 + \frac{\kappa}{2}\sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2 = \left(\frac{\kappa}{2} - \frac{1+2\sqrt{2}d}{n\spadesuit}\right)\sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2.\end{aligned} \tag{I.127}$$

Since $\kappa > 0$ and $\spadesuit \geq 0$, it is evident that the expression $\frac{\kappa}{2} - \frac{1+2\sqrt{2}d}{n\spadesuit}$ is strictly positive whenever the curvature constant satisfies $\kappa \gtrsim 2(n\spadesuit)^{-1}(1+2\sqrt{2}d)$. Under this condition, the term contributes a net descent in the empirical risk

$$\sum_{k=2}^{K}\left(\int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k}))(o)p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k-1}))(o)p_n d\nu(o)\right) \gtrsim \sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2 \geq 0. \tag{I.128}$$

However, in light of the inequality given in (I.125), we deduce that

$$0 \geq \sum_{k=2}^{K}\left(\int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k}))(o)p_n d\nu(o) - \int_{\mathcal{O}} \mathbf{L}(p_n^k(\epsilon_n^{k-1}))(o)p_n d\nu(o)\right) \gtrsim \sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2. \tag{I.129}$$

Therefore, it follows that

$$\sum_{k=2}^{K}\left\|\epsilon_n^k - \epsilon_n^{k-1}\right\|^2 \leq 0 \quad \Rightarrow \quad \left\|\epsilon_n^k - \epsilon_n^{k-1}\right\| \equiv 0 \text{ with } k = 2, \ldots, K. \tag{I.130}$$

□