A Short Note on Finite Sample Analysis on Double/Debiased Machine Learning

Yonghan Jung

Purdue University
jung222@purdue.edu
http://yonghanjung.me/

October 14, 2023

This note provides learning guarantees for sample-splitting-based estimators, which include double/debiased machine learning (DML) (Chernozhukov et al., 2018) estimators. We prove consistency and Gaussian approximation of estimators using finite-sample arguments, extending the general asymptotic theory. Our work extends previous research (Chernozhukov et al., 2023; Quintas-Martinez, 2022) that studied learning guarantees for the expected linear functional in general sample-splitting-based estimators.

Contents

1. Introduction

This paper aims to explore and provide a thorough understanding of sample-splitting-based estimators. This class of estimators utilizes sample-splitting to improve the accuracy of estimation. Our research endeavors to establish learning guarantees for this class of estimators, with a specific focus on double/debiased machine learning (DML) estimators. Our results are based on finite-sample arguments and demonstrate consistency and Gaussian approximation.

1.1. Related Works

In the realm of related work, studies such as Chernozhukov et al. (2023) and Quintas-Martinez (2022) both focused on providing learning guarantees for the expected linear functional (also known as Riesz representer) as the main focus in (Chernozhukov et al., 2022b). While our research also delves into learning guarantees, we provide general results that apply to any sample-splitting-based estimators. By doing so, we hope to contribute to the current understanding of this class of estimators and provide researchers and practitioners with a solid foundation upon which to build future work.

2. Problem Setup

2.1. Notations

Each variable is represented with a capital letter (X), and its realized value with a small letter (x). We use P_0 as a probability distribution over a random vector $V \in \mathbb{R}^d$, and $\mathcal{D} \coloneqq \{V_{(i)}\}_{i=1}^n$ as a set of n identically and independently distributed samples drawn from $P_0(V)$. For a function f, we define the mean of $f(V)$ with respect to P as $\mathbb{E}_{P_0}[f(V)] \coloneqq \int f \; d[P_0]$. We use $\mathbb{E}_{\mathcal{D}}[f(V)] \coloneqq \frac{1}{n}\sum_{i=1}^n f(V_{(i)})$ to denote the sample average over $\mathcal{D}$. We will use $\Phi(x)$ as the cumulative distribution function (CDF) of the standard normal distribution at x and $\phi(x)$ to denote the corresponding density function.

2.2. Setup

An inference framework considered in this paper is to find a solution parameter $\psi_0 \in L_2(P)$ satisfying an equation $\mathbb{E}_{P_0}[\phi(V;\eta_0,\psi)] = 0$ where $\eta \in L_2(P)$ is a set of nuisance parameters for the true nuisances $\eta_0 \coloneqq \eta(P_0)$. Here, P_0 is a true distribution such that $\mathcal{D} \overset{iid}{\sim} P_0$ where $\mathcal{D} \coloneqq \{V_{(i)}\}_{i=1}^n$ are independently and identically distributed samples. We will use $\sigma_0^2 \coloneqq \mathbb{E}_{P_0}[\{\phi(V;\eta_0,\psi_0)\}^2]$. We assume that the score is differentiable and the derivative is bounded.

Assumption 1 (Bounded First Derivative). *For a set of nuisances $\eta = \{\eta^j\}_{j=1}^m$, the map $\eta^j \mapsto \varphi(V;\eta)$ is uniformly differentiable with respect to η^j. Additionally, $\left\|\phi'(V;\eta^j)\right\| < c_j$ for $\phi'(V;\eta^j) \coloneqq \frac{\partial}{\partial \eta^j}\phi(V;\eta)$.*

In this paper, we consider a special but practically important case where the score function is affine to the target parameter ψ. The linearity of the score is formally stated as follows:

Assumption 2 (Linear Score). *The score function ϕ is linear; i.e., it satisfies the following:*

$$\phi(V;\eta,\psi) = \varphi(V;\eta) - \psi. \tag{1}$$

We focus on analyzing the finite-sample properties of estimators that use the cross-fitting (i.e., sample-splitting) technique (Klaassen, 1987; Robins and Ritov, 1997; Zheng and van der Laan, 2011; Chernozhukov et al., 2018; Newey and Robins, 2018). We refer to these estimators as *cross-fit estimators*, following Newey and Robins (2018). The cross-fit estimator and its confidence interval is formally defined as follows:

Definition 1 (**Cross-Fit Estimator and its Confidence Interval**)**.** *An estimator is called a cross-fit estimator based on a score function* ϕ *if it is constructed using the following step: Given* n*-size samples* $\mathcal{D} := \{V_{(i)}\}_{i=1}^{n}$*, randomly partition the samples into* L *folds; i.e.,* $\mathcal{D} = \cup_{\ell=1}^{L}\mathcal{D}_\ell$ *where* $n_\ell := |\mathcal{D}_\ell|$*.*

1. *For each fold* ℓ*, estimate the nuisance* $\hat{\eta}_\ell$ *using* $\mathcal{D}_\ell^c := \mathcal{D}\backslash\mathcal{D}_\ell$*.*
2. *Estimate* ψ_0 *as* $\hat{\psi} := (1/L)\sum_{\ell=1}^{L}\hat{\psi}_\ell$ *where* $\hat{\psi}_\ell := \mathbb{E}_{\mathcal{D}_\ell}[\varphi(V;\hat{\eta}_\ell)]$*.*
3. *Estimate* $(1-\alpha)100\%$ *confidence interval of* $\hat{\psi}$ *as* $\hat{\psi} \pm \frac{1}{\sqrt{n}}q_\alpha\hat{\sigma}$ *where* q_α *is the* $1-\alpha/2$ *quantile of the standard Gaussian distribution and* $\hat{\sigma}^2 := (1/L)\sum_{\ell=1}^{L}\hat{\sigma}_\ell^2$ *where* $\hat{\sigma}_\ell^2 := \mathbb{E}_{\mathcal{D}_\ell}[\{\phi(V;\hat{\eta}_\ell,\hat{\psi}_\ell)\}^2]$*.*

We note that the double/debiased machine learning (DML) estimator (Chernozhukov et al., 2018) is a special case of the cross-fit estimator where the score function ϕ is orthogonal with respect to nuisances.

Definition 2 (**Double/Debiased Machine Learning (DML) Estimator**(Chernozhukov et al., 2018, 2022a))**.** A cross-fit estimator based on a score function $\phi(V;\eta,\psi)$ is said to be a *double/debiased machine learning (DML) estimator* if the score function is orthogonal with respect to its nuisance; i.e.,

$$\frac{\partial}{\partial t}\mathbb{E}_P[\phi(V;\eta(P_t),\psi)]\bigg|_{t=0} = 0. \tag{2}$$

A general method for deriving a score function that satisfies the orthogonal property (*orthogonal score function*) in Eq. 2 was recently provided in (Chernozhukov et al., 2022a, Section 2.1). One implication of this is that an efficient influence function of a target parameter $\psi(P)$ is an orthogonal score function. Specifically, Jung et al. (2021) derived an influence function for any identifiable causal effects given a causal graph, and shown that such influence functions is an orthogonal score satisfying Assumption 2. Therefore, any one-step estimator based on an influence function equipped with the cross-fitting technique described in Def. 1 is a DML estimator. We collect an example of orthogonal score functions in Section xx.

To provide a clear example, we will describe a specific case in which we apply our general framework to the problem of treatment effect estimation.

Example 1. Covariate Adjustment. Let $V = (Z, X, Y)$, where X is the binary treatment variable, Z is a set of covariates, and Y is an outcome. The true parameter ψ_0 is an identification expression of the treatment effect $\mathbb{E}[Y(x)]$ under the ignorability assumption $Y(x) \perp\!\!\!\perp X|Z$ and a positivity assumption $\pi_0(X,Z) > 0$, where $\pi_0(X|Z) := \mathbb{1}_x(X)/P(X|Z)$. Under these assumptions, the true parameter is specified as $\psi_0 := \mathbb{E}_P[\mu_0(x,Z)]$, where $\mu_0(X,Z) := \mathbb{E}_P[Y|X,Z]$. An orthogonal score function is an influence function of ψ_0, which is given by: (Robins and Rotnitzky, 1995; Chernozhukov et al., 2017)

$$\phi(V;\eta=\{\mu,\pi\},\psi) := \pi(X|Z)\{Y-\mu(X,Z)\} + \mu(x,Z) - \psi, \tag{3}$$

where $\pi(X,Z)$ and $\mu(X,Z)$ are nuisance functions. Eq. (3) is a linear score satisfying Assumption 2.

Example 2. g-formula Let $V = \{Z_1, X_1, Z_2, X_2, \cdots, Z_m, X_m, Y\}$ be an ordered set, where $\forall i \in [m]$, X_i represents binary treatments, Z_i represents a set of covariates, and Y represents an outcome. Define $\overline{X}^i := \{X_1,\cdots,X_i\}$ and $\overline{Z}^i := \{Z_1,\cdots,Z_i\}$. The true parameter ψ_0 is an identification expression of the treatment effect $\mathbb{E}[Y(\mathbf{x})]$ for $\mathbf{x} := (x_1,\cdots,x_m)$ under ignorability assumptions $Y(\mathbf{x}) \perp\!\!\!\perp X_i|\overline{X}^{i-1},\overline{Z}^i$ and the positivity condition $\pi_0^i(X_i|\overline{X}^{i-1},\overline{Z}^i) > 0$ where $\pi_0^i(X_i|\overline{X}^{i-1},\overline{Z}^i) := \mathbb{1}_{x_i}(X_i)/P(X_i|\overline{X}^{i-1},\overline{Z}^i)$. Under these assumptions, the true parameter is defined as $\psi_0 := \mathbb{E}_P[\mu_0^1(x_1,Z_1)]$, where $\mu_0^k(\overline{X}^k,\overline{Z}^k) := \mathbb{E}_P[\mu_0^{k+1}(x_{k+1},\overline{X}^k,\overline{Z}^{k+1})|\overline{X}^k,\overline{Z}^k]$ for $k = 1,2,\cdots,m-1$, and $\mu_0^m(\overline{X}^m,\overline{Z}^m) := \mathbb{E}_P[Y|\overline{X}^m,\overline{Z}^m]$ (Robins, 1986; Bang and Robins, 2005; Rotnitzky et al., 2017). An orthogonal score function is an influence function of ψ_0, which is given by:

$$\phi(V;\eta=\{\pi^k,\mu^k : k\in[m]\},\psi) = \sum_{k=1}^{m}\overline{\pi}^k\{\mu^{k+1}(x_{k+1},\overline{X}^k,\overline{Z}^{k+1}) - \mu^k(\overline{X}^k,\overline{Z}^k)\}. \tag{4}$$

Example 3. Front-door Adjustment Let $V = (C, X, Z, Y)$, where X is the binary treatment variable, Z is a binary mediator variable, C is a set of covariates, and Y is an outcome. The true parameter ψ_0 is an identification expression of the treatment effect $\mathbb{E}[Y(x)]$ under the *front-door adjustment* identification assumption $Z(x) \perp\!\!\!\perp X|C$, $Y(x,z) \perp\!\!\!\perp Z(x)|X, W$ and a positivity assumption $\pi_0(X,C) > 0$ and $\xi_0(Z,X,C) > 0$ for $\pi_0(X,C) \coloneqq P(X|C)$ and $\xi_0(Z,X,C) \coloneqq P(Z|X,C)$ (Pearl, 2000; Fulcher et al., 2019). Assuming these conditions hold, the true parameter is defined as $\psi_0 \coloneqq \sum_{z,x' \in \mathcal{Z},\mathcal{X}} \mathbb{E}_P[\mu_0(z,x',C)\pi_0(x',C)\xi_0(z,x,C)]$, where $\mu_0(Z,X,C) \coloneqq \mathbb{E}_P[Y|Z,X,C]$. An orthogonal score function is an influence function of ψ_0, which is given as (Fulcher et al., 2019):

$$\begin{aligned}\phi(V; \eta = \{\mu, \xi, \pi\}, \psi) \coloneqq\ & \frac{\xi(Z|x,C)}{\xi(Z|X,C)}\{Y - \mu(Z,X,C)\} \\ & + \frac{\mathbb{1}_x(X)}{\pi(X|C)}\{\sum_{x'} \mu(Z,x',C)\pi(x';C) - \sum_{x',z'} \mu(z',x',C)\xi(z';x,W)\pi(x';C) \\ & + \sum_{z'} \mu(z',X,C)\xi(z';x,C) - \psi, \end{aligned} \tag{5}$$

where μ, π, ξ are nuisance functions. Eq. (5) is a linear score satisfying Assumption 2.

3. Finite Sample Learning Guarantees

In this section, we provide finite sample learning guarantees for the cross-fit estimator in Def. 1, which subsumes the DML estimators in Def. 2. We first provide an upper bound of the error of the estimator.

Theorem 1 (Finite-Sample Learning Guarantee). *Suppose Assumptions (1,2) hold. Let $\hat{\psi}$ be the cross-fit estimator of ψ_0, which is based on the score function ϕ. Let $\epsilon \in (0, \frac{1}{L+1})$ Then, with probability $1-(L+1)\epsilon$,*

$$|\hat{\psi} - \psi_0| \leq \frac{1}{\sqrt{n\epsilon}} \left(\sum_{\ell=1}^{L} \sum_{j=1}^{m} c_j \|\hat{\eta}^j_\ell - \eta^j_0\| + \sigma_0 \right) + \frac{1}{L} \sum_{\ell=1}^{L} |\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \,. \tag{6}$$

Theorem 1 implies that the estimator $\hat{\psi}$ is consistent and converges at a rate of $\sqrt{n}$, provided that each nuisance parameter $\hat{\eta}^j_\ell$ converges to η^j_0 in the $L_2(P)$ norm and the term $\mathbb{E}_P[\varphi(V;\hat{\eta}^\ell) - \varphi(V;\eta_0)]$ converges at a rate of $\sqrt{n}$. The analysis of the term $\mathbb{E}_P[\varphi(V;\hat{\eta}^\ell) - \varphi(V;\eta_0)]$, which is often called the *drift-term* in many literatures, including (Rotnitzky et al., 2017), is particularly crucial in studying the convergence behavior of the estimator. Since the behavior of the drift term depends on the specification of the function φ, we provide the analysis of the drift term for each working example in the next section.

We now provide a finite-sample learning guarantee for the confidence interval estimates. Note that the confidence interval estimation in Definition 1 is based on a Gaussian approximation of the estimator $\hat{\psi}$. This approximation is justified by the asymptotic behavior of the estimator, which exhibits asymptotic normality under certain conditions where each nuisance and the drift term converge quickly. The following result validates this approximation by providing an upper bound on the difference between the true distribution of the estimates and the standard Gaussian distribution.

Theorem 2 (Finite-Sample Gaussian Approximation). *Suppose Assumptions (1,2) hold. Let $\hat{\psi}$ be the cross-fit estimator of ψ_0, which is based on the score function ϕ. Let $\epsilon \in (0, 1/L)$. With probability of $1 - L\epsilon$,*

$$\left| P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x \right) - \Phi(x) \right| \leq \frac{1}{\sigma_0\sqrt{2\pi}}\Delta + \frac{0.4748\kappa_0^3}{\sigma_0^3\sqrt{n}}, \tag{7}$$

where $\sigma_0^2 := \mathbb{E}_P[\{\varphi(V;\eta_0) - \psi_0\}^2]$ *and* $\kappa_0^3 := \mathbb{E}_P[|\varphi(V;\eta_0) - \psi_0|^3]$, *and*

$$\Delta := \frac{1}{\sigma_0\sqrt{\epsilon}}\sum_{\ell=1}^{L}\sum_{j=1}^{m} c_j \left\|\hat{\eta}_\ell^j - \eta_0^j\right\| + \frac{\sqrt{n}}{\sigma_0 L}\sum_{\ell=1}^{L}\left|\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\right|. \tag{8}$$

Theorem 2 sheds light on how closely the Gaussian approximation of confidence intervals approximates the true confidence intervals of the cross-fit estimator. The drift term $\mathbb{E}_P[\varphi(V;\hat{\eta}) - \varphi(V;\eta_0)]$ is of particular importance as it determines the rate of convergence to limiting distribution. Specifically, the Gaussian approximation of the confidence interval converges to the true distribution only when each nuisance parameter $\hat{\eta}_\ell^j$ converges to η_0^j in the $L_2(P)$ norm and the term $\mathbb{E}_P[\varphi(V;\hat{\eta}^\ell) - \varphi(V;\eta_0)]$ converges at a rate of $\sqrt{n}$.

Note that finite-sample Gaussian approximation in Theorem 2 is expressed in terms of the true asymptotic variance σ_0^2. We will now provide an estimator for σ^2, denoted as $\hat{\sigma}^2$, to ensure the accuracy of the approximation.

Theorem 3 (**Variance Estimation**). *Suppose Assumptions (1,2) hold. Let* $\xi^4 := \mathbb{E}_P[\{\phi(V;\eta_0,\psi_0)\}^4]$ *for* $\xi > 0$. *Let* $c^2 := \sum_{j=1}^m c_j^2$. *Let* $\epsilon \in (0, 1/4L)$. *With probability* $1 - 4L\epsilon$,

$$\hat{\sigma}^2 - \sigma_0^2 \leq \Delta_1 + 2\sqrt{\Delta_1}\left(\sqrt{\Delta_2} + \sigma_0\right) + \Delta_2, \tag{9}$$

where

$$\Delta_1 := \frac{3L\sigma_0^2}{n\epsilon^2} + \left(\frac{c^2}{L\epsilon} + \frac{3c^2}{n\epsilon^2}\right)\sum_{\ell=1}^{L}\sum_{j=1}^{m}\|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \frac{3}{L\epsilon}\sum_{\ell=1}^{L}\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2, \tag{10}$$

and $\Delta_2 := L\xi^2/\sqrt{n\epsilon}$.

4. Conclusion

This note is only a succinct and non-exhaustive discussion of finite sample guarantees of double/debiased machine (DML) learning-based inference framework. This work is a mild generalization of (Chernozhukov et al., 2023). We note that our work is actually not confined to the DML inference framework. As long as assumptions (1,2) are satisfied, our finite sample analysis is applicable. Finally, we note that statistically appealing properties such as doubly robustness can be claimed if the form $\varphi(V;\hat{\eta})$ is more specified. We hope that this work serves as a template for analyzing finite sample

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1).

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022b). Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25(3):576–601.

Chernozhukov, V., Newey, W. K., and Singh, R. (2023). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika*, 110(1):257–264.

Esseen, C.-G. (1942). On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28:1–19.

Fulcher, I. R., Shpitser, I., Marealle, S., and Tchetgen Tchetgen, E. J. (2019). Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*

Jung, Y., Tian, J., and Bareinboim, E. (2021). Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence.*

Kennedy, E. H., Balakrishnan, S., G'Sell, M., et al. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030.

Klaassen, C. A. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138.*

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York. 2nd edition, 2009.

Quintas-Martinez, V. (2022). Finite-sample guarantees for high-dimensional dml. *arXiv preprint arXiv:2206.07386.*

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Rotnitzky, A., Robins, J., and Babino, L. (2017). On the multiply robust estimation of the mean of the g-functional. *arXiv preprint arXiv:1705.08582.*

Shevtsova, I. (2014). On the absolute constants in the berry-esseen-type inequalities. In *Doklady Mathematics*, volume 89, pages 378–381. Springer.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer.

Supplement of A Short Note on Finite Sample Analysis on Double/Debiased Machine Learning

A. Proof

A.1. Proof of Theorem 1

For $\mathcal{D} := \{V_{(i)}\}_{i=1}^{n}$ and an sample index set $\{1, 2, \cdots, n\}$, let I_ℓ denote an index set for a set of samples in $\mathcal{D}_\ell$; i.e., $\mathcal{D}_\ell = \{V_{(i)}\}_{i \in I_\ell}$. We first note that

$$\begin{aligned}
\hat{\psi} - \psi_0 &= \frac{1}{L}\sum_{\ell=1}^{L} \hat{\psi}_\ell - \psi_0 \\
&= \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{D}_\ell}[\varphi(V;\hat{\eta}_\ell)] - \mathbb{E}_P[\varphi(V;\eta_0)] \\
&= \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\eta_0)] + \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] + \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \\
&= \mathbb{E}_{\mathcal{D} - P}[\varphi(V;\eta_0)] + \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] + \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \\
&= \mathbb{E}_{\mathcal{D} - P}[\phi(V;\eta_0,\psi_0)] + \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] + \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \\
&\leq \left|\mathbb{E}_{\mathcal{D} - P}[\phi(V;\eta_0,\psi_0)]\right| + \frac{1}{L}\sum_{\ell=1}^{L} \left|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\right| + \frac{1}{L}\sum_{\ell=1}^{L} \left|\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\right|.
\end{aligned}$$

We first study $|\mathbb{E}_{\mathcal{D}-P}[\phi(V;\eta_0,\psi_0)]|$. We first note that

$$\mathbb{V}_P[\mathbb{E}_{\mathcal{D}-P}[\phi(V;\eta_0,\psi_0)]] = \frac{1}{n}\mathbb{V}_P[\phi(V;\eta_0,\psi_0)] = \frac{\sigma_0^2}{n}.$$

By Chebyshev's inequality,

$$P\left(|\mathbb{E}_{\mathcal{D}-P}[\varphi(V;\eta_0)]| > t\frac{\sigma_0}{\sqrt{n}}\right) < \frac{1}{t^2}.$$

Choosing $t = 1/\sqrt{\epsilon}$, we have, with probability $1-\epsilon$,

$$|\mathbb{E}_{\mathcal{D}-P}[\phi(V;\eta_0,\psi_0)]| \overset{\text{wp } 1-\epsilon}{\leq} \frac{\sigma_0}{\sqrt{n\epsilon}}. \tag{A.1}$$

Now we will examine the second term $|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]|$. According to (Kennedy et al., 2020, Lemma 2), the following statement holds true: for any $t > 0$ and the usage of sample-splitting, and by Chebyshev's inequality, for each $\ell \in \{1, 2, \cdots, L\}$.

$$P\left(|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \geq t\sqrt{\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]]}\right) \leq \frac{1}{t^2}.$$

Note

$$\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]] = \mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]]$$

$$= \frac{1}{n_\ell}\mathbb{V}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)].$$

Therefore,

$$P\left(|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \geq t\frac{1}{\sqrt{n_\ell}}\sqrt{\mathbb{V}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]}\right) \leq \frac{1}{t^2}.$$

By choosing $t = 1/\sqrt{\epsilon}$, we have the following: With probability $1-\epsilon$,

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n_\ell \epsilon}}\sqrt{\mathbb{V}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]}.$$

Since

$$\sqrt{\mathbb{V}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]} \leq \|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\|,$$

we have the following: With probability $1-\epsilon$,

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n_\ell \epsilon}}\|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\|.$$

Finally, by applying the Mean-value theorem with Assumption 1, we have,

$$\|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\| = \sum_{j=1}^{m}\|\varphi'(V;\hat{\eta}_\ell)\|\left\|\hat{\eta}^j_\ell - \eta^j_0\right\| \leq \sum_{j=1}^{m} c_j \left\|\hat{\eta}^j_\ell - \eta^j_0\right\|.$$

Combining, with probability $1-\epsilon$, for each $\ell \in \{1, 2, \cdots, L\}$, we have

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n_\ell \epsilon}}\sum_{j=1}^{m} c_j\|\hat{\eta}^j_\ell - \eta^j_0\|. \tag{A.2}$$

By applying Boole's inequality, with probability $1 - L\epsilon$, we have

$$\frac{1}{L}\sum_{\ell=1}^{L}|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n_\ell \epsilon}}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{m} c_j\|\hat{\eta}^j_\ell - \eta^j_0\|. \tag{A.3}$$

Since $1/\sqrt{n_\ell}L < 1/\sqrt{n_\ell L} = 1/\sqrt{n}$, we have the following: with probability $1 - L\epsilon$,

$$\frac{1}{L}\sum_{\ell=1}^{L}|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n\epsilon}}\sum_{\ell=1}^{L}\sum_{j=1}^{m} c_j\|\hat{\eta}^j_\ell - \eta^j_0\|. \tag{A.4}$$

Finally, we note that Eqs. (A.1 and A.4) hold in probability $1-(L+1)\epsilon$. This completes the proof.

A.2. Proof of Theorem 2

Lemma A.1. *Suppose Assumptions (1,2) hold. Let* $\bar{\psi} \coloneqq \mathbb{E}_{\mathcal{D}}[\varphi(V;\eta_0)]$*. At least with probability* $1 - L\epsilon$*,*

$$\frac{\sqrt{n}}{\sigma_0}\left|\bar{\psi} - \hat{\psi}\right| \leq \frac{1}{\sigma_0\sqrt{\epsilon}}\sum_{\ell=1}^{L}\sum_{j=1}^{m} c_j\left\|\hat{\eta}^j_\ell - \eta^j_0\right\| + \frac{\sqrt{n}}{\sigma_0 L}\sum_{\ell=1}^{L}|\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]|.$$

Proof for Lemma A.1. By Assumption 2, we note that

$$\left|\bar{\psi} - \hat{\psi}\right| = \left|\frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{\mathcal{D}_\ell}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\right|$$

$$= \left| \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] + \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}_{P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \right|$$

$$\leq \frac{1}{L} \sum_{\ell=1}^{L} \left| \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \right| + \frac{1}{L} \sum_{\ell=1}^{L} \left| \mathbb{E}_{P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \right|.$$

By (Kennedy et al., 2020, Lemma 2), the following holds: for any $t > 0$ and each $\ell \in \{1, 2, \cdots, L\}$,

$$P\left(\frac{|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]|}{\|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\| / \sqrt{n_\ell}} \geq t \right) \leq \frac{1}{t^2}.$$

By choosing $t = 1/\sqrt{\epsilon}$, we have

$$P\left(|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \geq \frac{1}{\sqrt{n_\ell \epsilon}} \|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\| \right) \leq \epsilon.$$

That is,

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n_\ell \epsilon}} \|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\| \qquad \text{with probability (w.p) } 1 - \epsilon$$

Finally, by applying the Mean-value theorem with Assumption 1, we have

$$\|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\| = \sum_{j=1}^{m} \|\varphi'(V;\hat{\eta}_\ell)\| \left\| \hat{\eta}_\ell^j - \eta_0^j \right\| \leq \sum_{j=1}^{m} c_j \left\| \hat{\eta}_\ell^j - \eta_0^j \right\|.$$

That is,

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n_\ell \epsilon}} \sum_{j=1}^{m} c_j \left\| \hat{\eta}_\ell^j - \eta_0^j \right\| \qquad \text{w.p } 1 - \epsilon.$$

By applying Boole's inequality and the fact that $L\sqrt{n_\ell} > \sqrt{L n_\ell} = \sqrt{n}$, we have

$$\frac{1}{L} \sum_{\ell=1}^{L} |\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \leq \frac{1}{\sqrt{n\epsilon}} \sum_{\ell=1}^{L} \sum_{j=1}^{m} c_j \left\| \hat{\eta}_\ell^j - \eta_0^j \right\| \qquad \text{w.p } 1 - L\epsilon.$$

In conclusion, with probability $1 - L\epsilon$,

$$\frac{\sqrt{n}}{\sigma_0} \left| \bar{\psi} - \hat{\psi} \right| \leq \frac{1}{\sigma_0 \sqrt{\epsilon}} \sum_{\ell=1}^{L} \sum_{j=1}^{m} c_j \left\| \hat{\eta}_\ell^j - \eta_0^j \right\| + \frac{\sqrt{n}}{\sigma_0 L} \sum_{\ell=1}^{L} |\mathbb{E}_{P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]|.$$

□

Proposition A.1 (**Berry–Esseen Inequality** (Berry, 1941; Esseen, 1942; Shevtsova, 2014))**.** *Suppose $\mathcal{D} = \{X_1, \cdots, X_n\}$ are independent and identically distributed random variables with $\mathbb{E}_P[X_i] = 0$, $\mathbb{E}_P[X_i^2] = \sigma^2$ and $\mathbb{E}_P[|X_i|^3] = \kappa^3$. Then, for all x and n,*

$$\left| P\left(\frac{\sqrt{n}}{\sigma_0} \mathbb{E}_{\mathcal{D}}[X] < x \right) - \Phi(x) \right| \leq \frac{0.4748\kappa^3}{\sigma^3 \sqrt{n}}.$$

Let $\bar{\psi} \coloneqq \mathbb{E}_{\mathcal{D}}[\varphi(V;\eta_0)]$. By Berry-Esseen Inequality in Prop. A.1,

$$P\left(\frac{\sqrt{n}}{\sigma_0} (\bar{\psi} - \psi_0) < x \right) - \Phi(x) \leq \frac{0.4748\kappa^3}{\sigma^3 \sqrt{n}}.$$

We recap that, by Lemma A.1, with a probability $1 - L\epsilon$,

$$\frac{\sqrt{n}}{\sigma_0}\left|\bar{\psi} - \hat{\psi}\right| < \Delta.$$

Then, with a probability $1 - L\epsilon$,

$$\begin{aligned}
P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right) - \Phi(x) &= P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \bar{\psi} + \bar{\psi} - \psi_0) < x\right) - \Phi(x) \\
&= P\left(\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \psi_0) < x + \frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \hat{\psi})\right) - \Phi(x) \\
&\leq P\left(\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \psi_0) < x + \Delta\right) - \Phi(x).
\end{aligned}$$

Also,

$$\begin{aligned}
P\left(\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \psi_0) < x + \Delta\right) - \Phi(x) &= P\left(\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \psi_0) < x + \Delta\right) - \Phi(x + \Delta) + \Phi(x + \Delta) - \Phi(x) \\
&= \frac{0.4748\kappa^3}{\sigma^3\sqrt{n}} + \Phi(x + \Delta) - \Phi(x) \\
&= \frac{0.4748\kappa^3}{\sigma^3\sqrt{n}} + \phi(x')\Delta \\
&\leq \frac{0.4748\kappa^3}{\sigma^3\sqrt{n}} + \frac{1}{\sqrt{2\pi}}\Delta,
\end{aligned}$$

since $\Phi(x + \Delta) - \Phi(x) = \phi(x')\Delta$ by Mean value theorem, where $\phi(x)$ is a standard Gaussian density.

Now, we are proving for $\Phi(x) - P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right)$. Note

$$\begin{aligned}
P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right) &= P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right) \\
&= P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \bar{\psi} + \bar{\psi} - \psi_0) < x\right) \\
&= P\left(\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \psi_0) < x - \frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \hat{\psi})\right) \\
&\geq P\left(\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \psi_0) < x - \Delta\right),
\end{aligned}$$

where the last inequality holds since $\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \hat{\psi}) \leq \Delta$. Then,

$$\begin{aligned}
\Phi(x) - P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right) &= \Phi(x) - \Phi(x - \Delta) + \Phi(x - \Delta) - P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right) \\
&\leq \frac{1}{\sqrt{2\pi}}\Delta + \Phi(x - \Delta) - P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right) \\
&\leq \frac{1}{\sqrt{2\pi}}\Delta + \Phi(x - \Delta) - P\left(\frac{\sqrt{n}}{\sigma_0}(\bar{\psi} - \psi_0) < x - \Delta\right) \\
&\leq \frac{1}{\sqrt{2\pi}}\Delta + \frac{0.4748\kappa^3}{\sigma^3\sqrt{n}}.
\end{aligned}$$

Therefore,

$$\left|P\left(\frac{\sqrt{n}}{\sigma_0}(\hat{\psi} - \psi_0) < x\right) - \Phi(x)\right| \leq \frac{1}{\sqrt{2\pi}}\Delta + \frac{0.4748\kappa^3}{\sigma^3\sqrt{n}}.$$

A.3. Proof of Theorem 3

Lemma S.2. *Suppose Assumptions (1,2,) hold. With a probability* $1-2\epsilon$*,*

$$\hat{\psi}_\ell - \psi_0 \le \frac{1}{\sqrt{n_\ell \epsilon}} \left(\sum_{j=1}^{m} c_j \|\hat{\eta}_\ell^j - \eta_0^j\| + \sigma_0 \right) + |\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \, .$$

Proof for Lemma S.2. By Assumption 2,

$$\begin{aligned} \hat{\psi}_\ell - \psi_0 &= \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\eta_0)] + \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] + \mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \\ &= \mathbb{E}_{\mathcal{D}_\ell - P}[\phi(V;\eta_0,\psi_0)] + \mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] + \mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)] \\ &\le |\mathbb{E}_{\mathcal{D}_\ell - P}[\phi(V;\eta_0,\psi_0)]| + |\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| + |\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \, . \end{aligned}$$

By Chevyshev's inequality,

$$P\left(|\mathbb{E}_{\mathcal{D}_\ell - P}[\phi(V;\eta_0,\psi_0)]| > \frac{1}{\sqrt{\epsilon}} \sqrt{\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\phi(V;\eta_0,\psi_0)]]} \right) < \epsilon,$$

which implies that

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\eta_0)]| \le \frac{1}{\sqrt{\epsilon}} \sqrt{\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\phi(V;\eta_0,\psi_0)]]} \qquad \text{w.p } 1-\epsilon.$$

Since

$$\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\phi(V;\eta_0,\psi_0)]] = \mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell}[\phi(V;\eta_0,\psi_0)]] = \frac{1}{n_\ell}\mathbb{V}_P[\phi(V;\eta_0,\psi_0)] = \frac{1}{n_\ell}\sigma_0^2,$$

we note that

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\eta_0)]| \le \frac{1}{\sqrt{n_\ell \epsilon}}\sigma_0 \qquad \text{w.p } 1-\epsilon. \tag{A.5}$$

By (Kennedy et al., 2020, Lemma 2), we note that

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \le \frac{1}{\sqrt{n_\ell \epsilon}} \|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\| \qquad \text{w.p } 1-\epsilon.$$

By Assumption 1, we have

$$\|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\| \le \sum_{j=1}^{m} \|\hat{\eta}_\ell^j - \eta_0^j\|.$$

Therefore,

$$|\mathbb{E}_{\mathcal{D}_\ell - P}[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]| \le \sum_{j=1}^{m} \|\hat{\eta}_\ell^j - \eta_0^j\| \qquad \text{w.p } 1-\epsilon. \tag{A.6}$$

By applying Boole's inequality, we note that Eqs. (A.5,A.6) hold with probability $1-2\epsilon$. This completes the proof. □

Lemma S.3. *Suppose Assumptions (1,2,) hold. Let* $c^2 := \sum_{j=1}^{m} c_j^2$*. With a probability* $1-2\epsilon$*,*

$$(\hat{\psi}_\ell - \psi_0)^2 \le \frac{3\sigma_0^2}{n_\ell \epsilon} + \frac{3c^2}{n_\ell \epsilon} \sum_{j=1}^{m} \|\hat{\eta}_\ell^j - \eta_0^j\|^2 + 3\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2.$$

Proof of Lemma S.3. By Lemma S.2, it suffices to show

$$\left\{\frac{1}{\sqrt{n_\ell \epsilon}}\left(\sum_{j=1}^{m} c_j\|\hat{\eta}_\ell^j - \eta_0^j\| + \sigma_0\right) + |\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]|\right\}^2 \leq \frac{\sigma_0^2}{n_\ell \epsilon} + \frac{c^2}{n_\ell \epsilon}\sum_{j=1}^{m}\|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2.$$

Using the inequality $(a+b+c)^2 < 3a^2 + 3b^2 + 3c^2$ by Cauchy-Schwartz inequality, we can see that the l.h.s of the above is upper bounded by

$$\frac{3}{n_\ell \epsilon}\left(\sum_{j=1}^{m} c_j\|\hat{\eta}_\ell^j - \eta_0^j\|\right)^2 + \frac{3\sigma_0^2}{n_\ell \epsilon} + 3\left\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\right\}^2.$$

By Cauchy-Schwartz inequality,

$$\frac{3}{n_\ell \epsilon}\left(\sum_{j=1}^{m} c_j\|\hat{\eta}_\ell^j - \eta_0^j\|\right)^2 \leq \frac{3c^2}{n_\ell \epsilon}\sum_{j=1}^{m}\|\hat{\eta}_\ell^j - \eta_0^j\|^2.$$

This completes the proof. □

Theorem 3 (Variance Estimation). *Suppose Assumptions (1,2) hold. Let $\xi^4 := \mathbb{E}_P[\{\phi(V;\eta_0,\psi_0)\}^4]$ for $\xi > 0$. Let $c^2 := \sum_{j=1}^m c_j^2$. Let $\epsilon \in (0, 1/4L)$. With probability $1 - 4L\epsilon$,*

$$\hat{\sigma}^2 - \sigma_0^2 \leq \Delta_1 + 2\sqrt{\Delta_1}\left(\sqrt{\Delta_2} + \sigma_0\right) + \Delta_2, \tag{9}$$

where

$$\Delta_1 := \frac{3L\sigma_0^2}{n\epsilon^2} + \left(\frac{c^2}{L\epsilon} + \frac{3c^2}{n\epsilon^2}\right)\sum_{\ell=1}^{L}\sum_{j=1}^{m}\|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \frac{3}{L\epsilon}\sum_{\ell=1}^{L}\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2, \tag{10}$$

and $\Delta_2 := L\xi^2/\sqrt{n\epsilon}$.

Proof of Theorem 3. To simplify notations, we will use $\hat{\phi}_\ell := \phi(V;\hat{\eta}_\ell,\hat{\psi})$ and $\phi_0 := \phi(V;\eta_0,\psi_0)$. Note

$$\begin{aligned}\hat{\sigma}_\ell^2 &:= \mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell\}^2] \\ &= \mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0 + \phi_0\}^2] \\ &= \mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2] + \mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] + 2\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}\phi_0],\end{aligned}$$

which implies that

$$\hat{\sigma}_\ell^2 - \mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] = \mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2] + 2\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}\phi_0].$$

Then,

$$\begin{aligned}\hat{\sigma}_\ell^2 - \sigma_0^2 &= \hat{\sigma}_\ell^2 - \mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] + \mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2 \\ &= \mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2] + 2\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}\phi_0] + \mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2.\end{aligned}$$

Note

$$\begin{aligned}2\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}\phi_0] &\leq 2\sqrt{\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2]}\sqrt{\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2]} \\ &= 2\sqrt{\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2]}\sqrt{\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2 + \sigma_0^2} \\ &\leq 2\sqrt{\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2]}\left(\sqrt{|\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2|} + \sigma_0\right).\end{aligned}$$

Therefore,

$$\hat{\sigma}_\ell^2 - \sigma_0^2 \leq \mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2] + 2\sqrt{\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2]}\left(\sqrt{|\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2|} + \sigma_0\right) + \left|\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2\right|.$$

The r.h.s. composed of two terms: $\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2]$ and $\left|\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2\right|$. The first term can be bounded as follow: By Markov inequality,

$$P\left(\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2] > t\right) \leq \frac{\mathbb{E}_P[\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2]]}{t} = \frac{\mathbb{E}_P[\{\hat{\phi}_\ell - \phi_0\}^2]}{t} = \frac{1}{t}\left\|\hat{\phi}_\ell - \phi_0\right\|^2,$$

and by choosing $t = \frac{\|\hat{\phi}_\ell - \phi_0\|^2}{\epsilon}$, the following holds:

$$\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2] \leq \frac{\|\hat{\phi}_\ell - \phi_0\|^2}{\epsilon} \qquad \text{w.p } 1 - \epsilon. \tag{A.7}$$

By Assumption 2 and the triangle inequality,

$$\left\|\hat{\phi}_\ell - \phi_0\right\|^2 \leq \|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\|^2 + \left\|\hat{\psi}_\ell - \psi_0\right\|^2.$$

By Assumption 1 and Cauchy-Schwartz inequality,

$$\|\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)\|^2 \leq \left\{\sum_{j=1}^m c_j^2\right\}\sum_{j=1}^m \|\hat{\eta}_\ell^j - \eta_0^j\|^2 = c^2\sum_{j=1}^m \|\hat{\eta}_\ell^j - \eta_0^j\|^2. \tag{A.8}$$

By Lemma S.3, we have

$$\left\|\hat{\psi}_\ell - \psi_0\right\|^2 \leq \frac{3\sigma_0^2}{n_\ell \epsilon} + \frac{3c^2}{n_\ell \epsilon}\sum_{j=1}^m \|\hat{\eta}_\ell^j - \eta_0^j\|^2 + 3\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2 \qquad \text{w.p } 1 - 2\epsilon. \tag{A.9}$$

Combining Eqs. (A.7, A.8, A.9) using Boole's inequality, we have the following with probability $1 - 3\epsilon$:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_\ell}[\{\hat{\phi}_\ell - \phi_0\}^2] &\leq \frac{c^2}{\epsilon}\sum_{j=1}^m \|\hat{\eta}_\ell^j \ - \eta_0^j|^2 + \frac{3\sigma_0^2}{n_\ell \epsilon^2} + \frac{3c^2}{n_\ell \epsilon^2}\sum_{j=1}^m \|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \frac{3}{\epsilon}\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2 \\ &= \frac{3\sigma_0^2}{n_\ell \epsilon^2} + \left(\frac{c^2}{\epsilon} + \frac{3c^2}{n_\ell \epsilon^2}\right)\sum_{j=1}^m \|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \frac{3}{\epsilon}\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2 \qquad (\text{A.10})\end{aligned}$$

Let

$$\Delta_{1,\ell} \coloneqq \frac{3\sigma_0^2}{n_\ell \epsilon^2} + \left(\frac{c^2}{\epsilon} + \frac{3c^2}{n_\ell \epsilon^2}\right)\sum_{j=1}^m \|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \frac{3}{\epsilon}\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2.$$

The second term $\left|\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2] - \sigma_0^2\right| = \left|\mathbb{E}_{\mathcal{D}_\ell - P}[\{\phi_0\}^2]\right|$ can be bounded as follow: By Chevyshev's inequality,

$$P\left(\left|\mathbb{E}_{\mathcal{D}_\ell - P}[\{\phi_0\}^2]\right| \geq \frac{1}{\sqrt{\epsilon}}\sqrt{\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\{\phi_0\}^2]]}\right) < \epsilon,$$

where

$$\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\{\phi_0\}^2]] = \mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell}[\{\phi_0\}^2]] = \frac{1}{n_\ell}\mathbb{V}_P[\{\phi_0\}^2] \leq \frac{\xi^4}{n_\ell}$$

Then,

$$\left|\mathbb{E}_{\mathcal{D}_\ell - P}[\{\phi_0\}^2]\right| \overset{w.p\ 1-\epsilon}{\leq} \frac{1}{\sqrt{\epsilon}}\sqrt{\mathbb{V}_P[\mathbb{E}_{\mathcal{D}_\ell - P}[\{\phi_0\}^2]]} \leq \frac{\xi^2}{\sqrt{n_\ell \epsilon}}. \tag{A.11}$$

Let

$$\Delta_{2,\ell} \coloneqq \frac{\xi^2}{\sqrt{n_\ell \epsilon}}.$$

Combining Eqs. (A.10, A.11), we have, with probability $1 - 4\epsilon$

$$\hat{\sigma}_\ell^2 - \sigma_0^2 \leq \Delta_{1,\ell} + 2\sqrt{\Delta_{1,\ell}}\left(\sqrt{\Delta_{2,\ell}} + \sigma_0\right) + \Delta_{2,\ell}.$$

Finally, let

$$\begin{aligned}
\Delta_1 &\coloneqq \frac{1}{L}\sum_{\ell=1}^{L}\Delta_{1,\ell} \\
&= \frac{1}{L}\sum_{\ell=1}^{L}\frac{3\sigma_0^2}{n_\ell \epsilon^2} + \frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{c^2}{\epsilon} + \frac{3c^2}{n_\ell \epsilon^2}\right)\sum_{j=1}^{m}\|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \frac{1}{L}\sum_{\ell=1}^{L}\frac{3}{\epsilon}\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2 \\
&= \frac{3L\sigma_0^2}{n\epsilon^2} + \left(\frac{c^2}{L\epsilon} + \frac{3c^2}{n\epsilon^2}\right)\sum_{\ell=1}^{L}\sum_{j=1}^{m}\|\hat{\eta}_\ell^j - \eta_0^j\|^2 + \frac{3}{L\epsilon}\sum_{\ell=1}^{L}\{\mathbb{E}_P[\varphi(V;\hat{\eta}_\ell) - \varphi(V;\eta_0)]\}^2.
\end{aligned}$$

Also, because

$$\frac{1}{L}\sum_{\ell=1}^{L}\Delta_{2,\ell} = \frac{L\xi^2}{L\sqrt{n_\ell \epsilon}} \leq \frac{L\xi^2}{\sqrt{n\epsilon}},$$

let

$$\Delta_2 \coloneqq \frac{L\xi^2}{\sqrt{n\epsilon}}.$$

Then,

$$\hat{\sigma}^2 - \sigma_0^2 \leq \Delta_1 + 2\sqrt{\Delta_1}\left(\sqrt{\Delta_2} + \sigma_0\right) + \Delta_2 \qquad \text{w.p } 1 - 4L\epsilon.$$

□