Plug-in Regularized Estimation of High-Dimensional Parameters in Nonlinear Semiparametric Models[*]

Victor Chernozhukov Denis Nekipelov Vira Semenova Vasilis Syrgkanis

Abstract

We develop a theory for estimation of a high-dimensional sparse parameter θ defined as a minimizer of a population loss function $L_D(\theta, g_0)$ which, in addition to θ, depends on a, potentially infinite dimensional, nuisance parameter g_0. Our approach is based on estimating θ via an ℓ_1-regularized minimization of a sample analog of $L_S(\theta, \hat{g})$, plugging in a first-stage estimate $\hat{g}$, computed on a hold-out sample. We define a population loss to be (Neyman) orthogonal if the gradient of the loss with respect to θ, has pathwise derivative with respect to g equal to zero, when evaluated at the true parameter and nuisance component. We show that orthogonality implies a second-order impact of the first stage nuisance error on the second stage target parameter estimate. Our approach applies to both convex and non-convex losses, albeit the latter case requires a small adaptation of our method with a preliminary estimation step of the target parameter. Our result enables oracle convergence rates for θ under assumptions on the first stage rates, typically of the order of $n^{-1/4}$.

We show how such an orthogonal loss can be constructed via a novel orthogonalization process for a general model defined by conditional moment restrictions. We apply our theory to high-dimensional versions of standard estimation problems in statistics and econometrics, such as: estimation of conditional moment models with missing data, estimation of structural utilities in games of incomplete information and estimation of treatment effects in regression models with non-linear link functions.

1 Introduction

Many questions in Economics and Statistics can be posed as an extremum estimation problem:

$$\theta_0 = \arg\min_{\theta\in\Theta} L_D(\theta, g_0), \tag{1}$$

where $L_D : \Theta \times \mathcal{G} \to \mathbb{R}$ is a population loss function induced by the data distribution D and dependent on a parameter of interest $\theta \in \Theta \subset \mathbb{R}^p$ and a potentially infinite dimensional nuisance

[*] Code is available at: `https://github.com/vsyrgkanis/plugin_regularized_estimation`

parameter $g \in \mathcal{G}$. The true value θ_0 is defined as the minimizer of the population loss $L_D(\cdot, g_0)$ evaluated at the true value of the nuisance component g_0. The nuisance component g_0 can itself be estimated based on some auxiliary estimation process, whose description depends on the application of interest.

We address the problem of estimating θ_0 based on a data set S of n i.i.d. samples, each drawn from distribution D and we consider a high-dimensional sparse regime, i.e., we allow the dimension p to exceed the sample size n: $p \gg n$, but require θ_0 to be sparse:

$$k = \|\theta_0\|_0 := |\{j : \theta_{0,j} \neq 0\}| \ll n$$

This framework extends standard semiparametric extremum estimation problems by allowing the finite dimensional parameter to be a high-dimensional sparse vector. Instances of this framework that we investigate in detail in this paper include estimating models defined via *conditional moment restrictions with missing data*, estimating the utility of agents in *games of incomplete information* and estimating *treatment effects* in a regression model with a nonlinear-link function. In all these settings, our work enables estimation in the high-dimensional regime, where among the p treatments/features only k of them have a non-zero effect on the outcome.

As is typical in semiparametric models, estimating g_0 is most times a much harder problem in terms of sample complexity than estimating θ_0, had we been given oracle access to the true g_0 (e.g. estimating g_0 requires a non-parametric regression or a high-dimensional regression with very dense parameters). This nature of semiparametric estimation extends even in the high dimensional θ_0 regime. Motivated by this observation, the main goal of our work is to develop an estimation algorithm for θ_0, whose performance is robust to errors in the estimation of g_0.

Two-stage Estimation. A natural way to estimate θ_0 is via a two-stage procedure, where a first-stage estimate $\hat{g}$ of the nuisance component is plugged into a ℓ_1-regularized sample analog of (1). Namely, we assume the existence a sample loss function $L_S(\cdot, g)$ that concentrates around $L_D(\cdot, g)$ conditional on any first-stage estimate g, as the sample size n becomes large. Given such an *empirical loss* function, we propose to estimate θ by the following two-stage algorithm:

Algorithm 1 Plug-in Regularized Extremum Estimator

```
Input: $\lambda \geqslant 0$, search set $\mathcal{T}$
1: Partition sample into a auxiliary sample $S'$ and a estimation sample $S$
2: Estimate the nuisance parameter $\hat{g} \in \mathcal{G}$ on the auxiliary sample $S'$.
3: Estimate $\theta_0$ via plug-in $\ell_1$-regularized local extremum estimation, i.e.:
       $\hat{\theta} = \text{local-}\min_{\theta \in \mathcal{T}} L_S(\theta, \hat{g}) + \lambda\|\theta\|_1,$    (2)
Return: $\hat{\theta}$
```

Our main result is to show that if the loss function L_S satisfies an orthogonality condition with respect to the nuisance component, as well as regularity conditions that are typical in high-dimensional estimation, then the convergence rate of the plug-in regularized extremum estimator presented in Algorithm 1, has a second order impact from the first-stage estimation error of g, i.e. it depends only on the squared error $\|g - g_0\|^2$.

Example 1 (High-Dimensional Heterogeneous Treatment Effects)**.** *To make matters more concrete, let us consider a stylized, albeit of practical importance, model of heterogeneous treatment effect estimation. In particular consider the following structural model, which corresponds to a high-dimensional extension of the classic Partially Linear Model (PLR) [19]:*

$$y = \tau \cdot u'\theta_0 + \underbrace{u'\alpha_0}_{f_0(u)} + \epsilon, \qquad \tau = \underbrace{u'\beta_0}_{h_0(u)} + \eta, \qquad \mathbb{E}[\epsilon|\tau, u] = 0, \quad \mathbb{E}[\eta|u] = 0, \quad \epsilon \perp\!\!\!\perp \eta \mid u$$

where $\tau \in \mathbb{R}$ is a base treatment variable, $u \in \mathbb{R}^p$ is a high-dimensional vector of features/control variables and $y \in \mathbb{R}$ is an outcome of interest. The target parameter θ_0 corresponds to a linear parametrization of the heterogeneous treatment effect of τ on y conditional on the features u. The features u also have a confounding effect, in the sense that they have a direct impact on the outcome apart from determining the treatment effect. This setting falls into our formulation, where $g_0 = \{f_0, h_0\}$ are the nuisance components in the estimation of the parameter of interest θ_0. Many times, the density of the coefficients α_0 and β_0 is much larger than the density of θ_0, i.e. many variables u have a direct effect on the outcome, but do not alter the effect of the treatment. Hence, our goal is to estimate θ_0 in a manner that does not depend on the support size of the coefficients α_0, β_0.

In this particular example, one could estimate θ via a direct approach, by regressing y on $x \cdot u$, u via the Lasso algorithm, i.e. minimizing the regularized loss:

$$\min_{\alpha,\theta} L_S(\theta, \alpha) := \frac{1}{2n} \sum_{i=1}^{n} \left(\tau_i \cdot u_i'\theta + u_i'\alpha\right)^2 + \lambda\|\alpha\|_1 + \lambda\|\theta\|_1$$

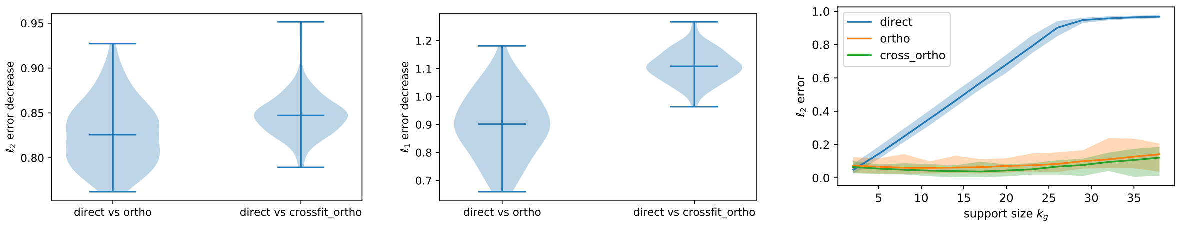

Figure 1: Direct vs orthogonal estimation for linear heterogeneous treatment effect model. We simulated 100 experiments, with data generating parameters $u \sim N(0, I_p)$, $\epsilon, \eta \sim N(0,1)$, $n = 5000$, $p = 200$, $k = 2$, $k_g = \|\alpha_0\|_0 = \|\beta_0\|_0 \in \{2, 5, \ldots, 38\}$. Supports of α_0 and β_0 are identical and all non-zero coefficients of $\alpha_0, \beta_0, \theta_0$ are equal to one. The left figure shows the distribution of the decrease in ℓ_2 and ℓ_1 estimation errors for θ, when going from the direct regression to the orthogonal plug-in estimation or to the orthogonal estimation with cross-fitting (see Section 2.3) for $k_g = 38$, across 100 experiments. Right figure shows distributions of ℓ_2 errors as k_g grows.

However, with such a direct approach, the convergence rate for the parameter θ_0 will depend on the support-size of both θ_0 and α_0. The framework that we will establish in this work, will show that if instead one invokes our two-stage Algorithm (1) *with a slightly modified loss function:*

$$L_S(\theta, g) = \frac{1}{2n} \sum_{i=1}^{n} ((\tau_i - \hat{h}(u_i)) \cdot u_i'\theta + \hat{q}(u_i))^2 + \lambda \|\theta\|_1 \tag{3}$$

where $\hat{h}$ is a first stage estimate of h and $\hat{q}$ is a first stage estimate of the function $q_0(u) = \mathbb{E}[y \mid u] = h_0(u) \cdot u'\theta_0 + f_0(u)$, then the convergence rate of the resulting estimate is asymptotically independent of the error in the estimation of h and q, i.e. the density of the coefficients of their linear representations enters only in a non-leading n^{-1} term. The crucial property that enables this result is that the modified loss satisfies an orthogonality condition, which we will define shortly and which renders it insensitive to local perturbations of the nuisance components, near the true values of both θ and g. This difference of the two estimation methods is not an artifact of the theoretical analysis, but exhibits itself clearly in their experimental performance as we show in Figure 1.

A detailed exposition of the latter result is given in Section 2.4. This approach extends beyond the linear setting to high-dimensional treatment effect estimation with non-linear link functions, i.e. $\mathbb{E}[y \mid x, u] = G(x'\theta_0 + f_0(z))$, where we present an orthogonal loss construction, which is novel even in the low-dimensional regime. This generalization is presented in Section 4.3 and a sample experimental performance of our approach for the logistic link, which arises in estimation of discrete choice models, is presented in Figure 2.

Outline of Main Result. The input to Algorithm 1 consists of the regularization parameter $\lambda \geqslant 0$ and a search set $\mathcal{T} \subset \mathbb{R}^p$. Depending on the convexity of the loss $L_S(\theta, g)$ in θ, the

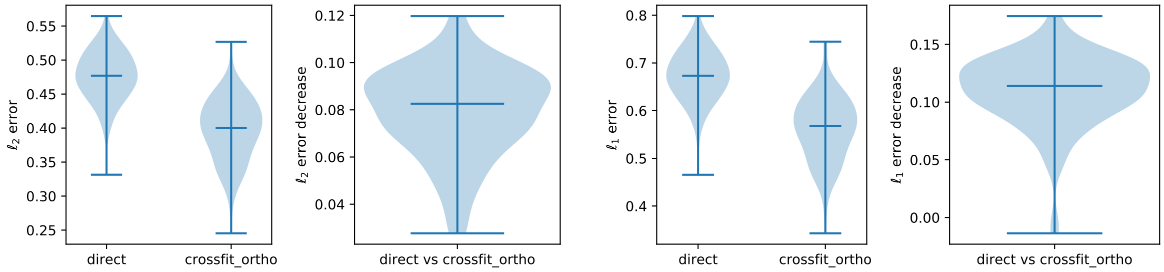

Figure 2: Direct vs orthogonal estimation for heterogeneous treatment effect model with a logistic link function. We simulated 100 experiments, with data generating parameters $u \sim U(-.5, .5)^p$, $\eta \sim N(0, 3)$, $n = 5000$, $p = 2000$, $k = 2$, $k_g = \|\alpha_0\|_0 = \|\beta_0\|_0 = 5$. Supports of α_0 and β_0 are identical and all non-zero coefficients of $\alpha_0, \beta_0, \theta_0$ are equal to one. The left figure shows the distribution of the ℓ_2 errors of the two methods as well as the distribution of the decrease in ℓ_2 error when going from from the direct regression to the orthogonal plug-in estimation (see Section 4.3) across the 100 experiments. The right figure shows the corresponding quantities for the ℓ_1 norm

algorithm defines either a global or local optimization problem, and we consider both cases in Section 2. In the convex case, we conduct a global search by setting $\mathcal{T} = \mathbb{R}^p$ and the regularization parameter λ to dominate the gradient of the loss with high probability

$$\lambda \gtrsim_P \|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty,$$

The rate at which the gradient of the empirical plugin loss evaluated at the true parameter goes to zero, is a proxy of how the *noise* of the problem decays to zero as the sample size grows. In the non-convex case, we conduct a local search determined by the properties of the loss $L_S(\theta, \hat{g})$, which will be discussed later, and set λ to dominate the gradient of the loss and the local violation of the convexity $L_S(\theta, \hat{g})$ around θ_0. In both cases, the error of the final estimator is proportional to the regularization parameter λ.

Hence, to understand the impact of the first stage estimation error $\|\hat{g} - g_0\|$ on the second stage estimate, one crucial aspect is characterizing how this error affects the *noise* of our second stage estimation problem, as captured by the empirical plugin gradient evaluated at the true value θ_0. We define a population loss to be (Neyman)-orthogonal to the nuisance parameter g if the pathwise derivative of the gradient of the loss $\nabla_\theta L_D(\theta, g)$ w.r.t g, evaluated at the true parameter and nuisance component value, is zero:

$$D_0[g - g_0, \nabla_\theta L_D(\theta_0, g_0)] := \nabla_r \nabla_\theta L_D(\theta_0, r(g - g_0) + g_0)|_{r=0} = 0.$$

In other words, at the true parameter value, local perturbations of the nuisance component around its true value, have a zero first-order effect on the gradient of the loss, i.e.:

$$\|\nabla_\theta L_D(\theta_0, \hat{g})\|_\infty = \|\nabla_\theta L_D(\theta_0, g_0)\|_\infty + O(\|\hat{g} - g_0\|^2)$$

As we will show later, in several estimation settings defined via conditional moment restrictions it is always possible to construct such an orthogonal loss.

Subsequently, we can use this property to show an analogue of it for the empirical loss. Crucially, this property allows us to set a regularization weight that only depends on $\|\hat{g} - g_0\|^2$, since that suffices for regularization to dominate the *noise* of the problem. Since the convergence rate of θ is determined by the required level of regularization, this leads to our desired second-order influence property. Moreover, if the quantity $g_n = \|\hat{g} - g_0\|^2$ is of lower order than the rate at which the oracle empirical gradient $\epsilon_n = \|\nabla_\theta L_S(\theta_0, g_0)\|_\infty$ converges to zero, then the estimation error of g can essentially be asymptotically ignored. In typically settings, ϵ_n will be of the order of $O_p\left(\sqrt{\frac{\log(p)}{n}}\right)$. Hence, the requirement for the oracle convergence property is essentially $g_n = o_p\left(\left(\frac{\log(p)}{n}\right)^{1/4}\right)$, which can be achieved by several non-parametric or high-dimensional parametric estimators. Even when g_n is not fast enough to ensure the oracle convergence property, orthogonality still benefits the estimation of θ_0, in that it renders it more robust to the nuisance component estimation.

The results of this paper accommodate estimation of g_0 by high-dimensional/highly complex modern machine learning (ML) methods, such as random forests, neural networks, and ℓ_1-shrinkage estimators, as well as earlier developed tools. The only requirement we impose on $\hat{g}$ is its uniform convergence to the g_0 at some slow rate g_n. Crucially, we do not impose further stringent complexity requirements on the function class in which the first stage estimates need to lie in. We achieve this by a sample splitting approach, already introduced in the low-dimensional regime. We also show that for a particular class of extremum estimators, namely M-estimators, the slight statistical inefficiency due to sample splitting can be alleviated via a cross-fitting scheme (see Algorithm 3).

Our formal proof requires further technical steps, primarily addressing the fact that the first-stage estimation error also has an effect on the second-order (strong convexity) properties of the loss function. In the convex setting this translates to a minimal requirement on the rate g_n, so that its effect on the second-order properties of the loss can be ignored after some constant number of samples n_0. This effect is much harder to handle in the non-convex setting, where the effect on the second order properties of the loss, need to also be dominated by the regularization strength λ, thereby entering the convergence rate. This seemingly leads to a first order effect of the nuisance estimation on the target parameter estimation. However, we show how to bypass this problem via a two-step estimation approach, where we first estimate a preliminary $\tilde{\theta}$ at a slow rate and then refine our search set around this preliminary estimate. With this addition, we arrive at an overall estimation algorithm (see Algorithm 2) that enjoys second-order dependence on the estimation of g. The sole drawback of this approach is that the requirement on g_n for oracle convergence is stricter than in the convex setting. In particular: $g_n = o_p\left(\frac{1}{k}\left(\frac{\log(p)}{n}\right)^{1/4}\right)$, which contains an extra $1/k$ in the right hand side, as compared to the convex case.

Constructing an Orthogonal Loss. Our main result is presented conditional on having access to an orthogonal loss. One might wonder how one arrives at such a loss from primitives of the model. In Section 3, we show how such an orthogonal loss can always be constructed, via a novel orthogonalization technique, when the model is defined via single-index conditional moment restrictions of the form:

$$\begin{aligned} \mathbb{E}[M(w, \Lambda(z, g_0(z))'\theta_0, g_0(z))|z] &= 0, \\ \mathbb{E}[v - g_0(z)|z] &= 0, \end{aligned}$$

for some real-valued moment function M, where the inner product $\Lambda(z, g_0(z))'\theta_0$ is referred to as the *index*. Crucially, the parameter θ_0 enters the moment function only through the index. The latter approach applies to our application on estimation with missing data and estimation in games of incomplete information. For our final application of non-linear treatment effect estimation, we develop a separate *partialing-out* approach to arrive at an orthogonal loss. This method is an extension and generalization of the loss function presented in Example 1.

Applications. We apply our general results to three classical problems in bio-statistics, structural econometrics and causal inference. Concretely we address estimation of conditional moment models with missing data (Section 4.1), estimation of agent utilities in games of incomplete information (Section 4.2) and estimation of high-dimensional treatment effects in regression problems with nonlinear link functions (Section 4.3). In all these settings, we extend prior work (e.g. [9], [7], [3]), from the low-dimensional target parameter setting, to its sparse high-dimensional counterpart. For each setting, we establish concrete conditions on the complexity of the nuisance components that lead to oracle convergence rates for our two-stage estimator.

Literature Review. This paper builds on the two theoretical bodies of research within the Statistics and Econometrics literature: (1) orthogonal/debiased machine learning and (2) sparse high-dimensional M-estimation and its extensions to non-convex settings, as well as uses the examples of the models described by conditional moment restrictions. The first literature ([7]) provides a $\sqrt{n}$-consistent and asymptotically normal estimates of low-dimensional target parameters θ_0 in the presence of high-dimensional/highly complex nonparametric nuisance functions.

The second literature establishes the convergence rates for ℓ_1-penalized M-estimation problems in convex ([18]) and non-convex ([17] and [15]) settings. As for the applications, we illustrate our results by applying them to Conditional Moment Models in presence of Missing Data ([9],[10]) applicable also to the models with measurement error and an error-free validation samples (as studied in [4], [5], [6], [14], [20]), Games of Incomplete Information (e.g. see [2] and [3] among others), and model of high-dimensional treatment effects with nonlinear link function, whose linear case was considered in [8].

2 Plug-in L_1 Regularized Extremum Estimation

In this section we derive the convergence rate for the Plug-in Regularized Extremum Estimator, outlined in Algorithm 1, which exhibits second-order impact from the first stage error in the estimation of g. We establish sufficient conditions under which this rate can be attained.

We assume that both the estimation sample $S = \{w_1, \dots, w_n\} \in \mathcal{W}^n$ and the auxiliary sample $S' = \{w_1', \dots, w_n'\} \in \mathcal{W}^n$ consist of n i.i.d. data points, each drawn from a data generating distribution D. We consider empirical loss function $L_S(\theta, g)$ and population loss $L_D(\theta, g)$, that depend on a target parameter $\Theta \in \mathbb{R}^p$ and a nuisance component g that can either be a finite-dimensional parameter or a function. We assume that g belongs to a convex set $\mathcal{G}$ equipped with some norm $\|\cdot\|$, whose choice will be specific to the type of the nuisance parameter and the application of interest.[1]

A leading example of our framework is an M-estimation problem, where sample and population losses are defined as the empirical and population expectation of an M-estimator loss function $\ell : \mathcal{W} \times \Theta \times \mathcal{G} \to \mathbb{R}$, i.e.:

$$L_S(\theta, g) = \frac{1}{n} \sum_{i \in S} \ell(w_i, \theta, g), \quad L_D(\theta, g) = \mathbb{E}\left[\ell(w, \theta, g)\right], \tag{4}$$

Our results are not specific to the M-estimation setting and also apply to loss functions that are not additively separable across samples.

Our goal in this section is to establish high probability bounds on the estimation error $\|\hat{\theta} - \theta_0\|$ of Algorithm 1, with respect to either the ℓ_2 or the ℓ_1 norm. To enable our results we first impose a set of sufficient conditions. Some of these conditions will be of a first order nature (e.g. orthogonality and strong convexity), while others will be easily satisfied under mild

[1] For instance, in the case of a finite dimensional parameter g, the norm $\|\cdot\|$ could be some L^p-norm of the finite dimensional vector space $\mathcal{G}$, and in the case of a vector-valued function $g(w)$, it could be the L^p norm with respect to the measure of w defined by the distribution D, i.e. $\|g\| = \|\left(\mathbb{E}_{w\sim D}[g(w)^p]\right)^{1/p}\|$, with the outer-norm being some finite dimensional L^p space norm.

smoothness and differentiability properties of the loss functions. We will typically refer to the latter as regularity assumptions.

First-Stage Rate. Our first regularity assumption requires that the first stage estimator achieves some non-trivial rate of convergence to the truth. In particular, Assumption 1 introduces a sequence of nuisance realization sets $\mathcal{G}_n \subset \mathcal{G}$ that contain the first-stage estimator $\hat{g}$ with high probability. As sample size n increases, set $\mathcal{G}_n$ shrinks around the true value g_0. The shrinkage speed is measured by the rate g_n and is referred to as the first-stage rate. At the end of the section we will characterize bounds on g_n under which the first stage error can be ignored and is not of the same order as the leading error term of the second stage estimation. However, our convergence rate for the second stage will be valid for any rate g_n and will still have a dampened impact from the first stage error, even if this impact is of a leading order. Only in our convex setting, we will impose the mild condition that $kg_n = o(1)$ for our convergence rate to be valid.

REGULARITY ASSUMPTION 1 (Nuisance Parameter Estimation Error)**.** *For any $\delta > 0$, w.p. at least $1 - \delta$, the first-stage estimate $\hat{g}$ belongs to a neighborhood $\mathcal{G}_n \subset \mathcal{G}$ of g_0, such that:*

$$\sup_{g \in \mathcal{G}_n} \|g - g_0\| \leqslant g_{n,\delta}.$$

Orthogonality of the population loss $L_D(\theta, g)$. To dampen the impact of the estimation error of the first-stage estimator $\hat{g}$ on the second-stage estimator $\hat{\theta}$, we require population loss $L_D(\theta_0, g)$ to be orthogonal with respect to g. We call a population loss (Neyman) orthogonal to the nuisance parameter g if the pathwise derivative of the loss gradient $\nabla_\theta L_D(\theta, g_0)$ w.r.t g is zero.

Definition 1 (Orthogonal Loss)**.** *Loss function $L : \Theta \times \mathcal{G} \to \mathbb{R}$ is orthogonal with respect to the nuisance function if the pathwise derivative map of its gradient at θ_0,*

$$D_r[g - g_0, \nabla_\theta L(\theta_0, g_0)] := \frac{\partial}{\partial r} \nabla_\theta L(\theta_0, r(g - g_0) + g_0) \tag{5}$$

exists $\forall r \in [0, 1)$ and $g \in \mathcal{G}$, and vanishes at $r = 0$:

$$\forall g \in \mathcal{G} : D_0[g - g_0, \nabla_\theta L(\theta_0, g_0)] = 0 \tag{6}$$

ASSUMPTION 2 (Orthogonality of Population Loss)**.** *The population loss function L_D is orthogonal.*

To guarantee that the impact of the first-stage estimator $\hat{g}$ on the second stage estimator $\hat{\theta}$ is second-order, we require an extra regularity assumption which is easily satisfied when $\nabla_\theta L_D(\theta_0, \cdot)$ is sufficiently smooth.

REGULARITY ASSUMPTION 3 (Bounded Hessian of the Gradient of Population Loss w.r.t. Nuisance)**.** *The second order path-wise derivative of the gradient* $\nabla_\theta L_D(\theta_0, \cdot)$ *w.r.t the nuisance parameter:*

$$D_r^2[g' - g, \nabla_\theta L_D(\theta_0, g)] := \frac{\partial^2}{\partial r^2} \nabla_\theta L_D(\theta_0, r(g' - g) + g) \tag{7}$$

exists $\forall r \in [0, 1)$ *and is bounded as:*

$$\exists B, \forall g \in \mathcal{G}_n, \forall r \in [0, 1) : \left\| D_r^2[g - g_0, \nabla_\theta L_D(\theta_0, g_0)] \right\|_\infty \leqslant B\|g - g_0\|^2.$$

When Assumption 2 and Regularity Assumption 3 are satisfied, the estimation error in $\hat{g}$ has a second-order impact on the gradient of the population loss, since by the second-order Taylor expansion:

$$\|\nabla_\theta L_D(\theta_0, \hat{g})\|_\infty \leqslant \|\nabla_\theta L_D(\theta_0, g_0)\|_\infty + B\|\hat{g} - g_0\|^2 = B\|\hat{g} - g_0\|^2,$$

where the last equality follows from the first order condition (FOC) for $L_D(\theta, g_0)$ being satisfied at θ_0.

Convergence of the gradient of the empirical loss $L_S(\theta_0, g)$ To ensure that the empirical oracle gradient $\nabla_\theta L_S(\theta_0, g)$ goes to zero in ℓ_∞ norm, we assume that the gradient of the empirical loss $\nabla_\theta L_S(\theta_0, g)$ concentrates well around its population analogue for each fixed instance of $g \in \mathcal{G}_n$. Crucially, by using different samples in the first and the second stages of Algorithm 1, we do not require the uniform convergence of $\nabla_\theta L_S(\theta_0, g)$ over the realization set $\mathcal{G}_n$ of the nuisance g, and therefore, we do not restrict the complexity of the function class $\mathcal{G}_n$. As a result, one can employ high-dimensional, highly complex methods to estimate g_0.

REGULARITY ASSUMPTION 4 (Convergence Rate of Empirical Gradient)**.** *We assume that for any fixed* $g \in \mathcal{G}_n$*, there exists a sequence* $\epsilon_{n,\delta}$ *such that* $\|\nabla_\theta L_S(\theta_0, g) - \nabla_\theta L_D(\theta_0, g)\|_\infty$ *converges at rate* $\epsilon_{n,\delta}$ *to zero w.p.* $1 - \delta$*. Formally, for any* $\delta > 0$*:*

$$\forall g \in \mathcal{G}_n : \Pr\left(\|\nabla_\theta L_S(\theta_0, g) - \nabla_\theta L_D(\theta_0, g)\|_\infty \leqslant \epsilon_{n,\delta}\right) \geqslant 1 - \delta.$$

This regularity assumption is a mild requirement. For example, in the case of the M-estimation problem with a bounded loss gradient $\|\nabla_\theta \ell(w, \theta_0, g)\|_\infty \leqslant B$, Assumption 4 follows from McDiarmid's inequality with $\epsilon_n = O\left(B\sqrt{\frac{\log(p/\delta)}{n}}\right)$.

Curvature of the loss. The mere fact that the estimator $\hat{\theta}$ is a local minimum of the empirical loss $L_S(\cdot, \hat{g})$ is not sufficient to guarantee that $\hat{\theta}$ is close to θ_0. Even if we knew that $\hat{\theta}$ was an approximate minimizer of the population oracle loss $L_D(\cdot, g_0)$, this would not imply that $\hat{\theta}$ is close to θ_0, unless the loss function $L_D(\cdot, g_0)$ has a large curvature within the search set $\mathcal{T}$. For a given direction $\nu \in \mathbb{R}^p$, we measure this curvature of loss function $L : \Theta \to \mathbb{R}$ by the symmetric Bregman distance, considered in [1, 16, 15] among others.[2]

Definition 2 (Symmetric Bregman distance). *For a differentiable function $L : \Theta \to \mathbb{R}$, define its symmetric Bregman distance as:*

$$H(\theta_0 + \nu, \theta_0) := \langle \nabla_\theta L(\theta_0 + \nu) - \nabla_\theta L(\theta_0), \nu \rangle \tag{8}$$

Given that assumptions presented below pertain to the second order properties of loss functions, they will depend on the overall convexity of the empirical loss $L_S(\theta, g)$ in θ. In the non-convex case, we will conduct a local optimization, where the search set $\mathcal{T}$ of Algorithm 1 depends on the problem features as discussed below. In addition, our further assumptions will be required to hold uniformly for *all directions ν in an ℓ_1 neighborhood* around θ_0. In a convex case, convexity of $L_S(\theta, g)$ ensures that the estimator $\hat{\theta}$ belongs to a *restricted cone*

$$\mathcal{C}(T; 3) = \{\nu \in \mathbb{R}^p : \|\nu_{T^c}\|_1 \leqslant 3\|\nu_T\|_1\}, \tag{9}$$

where T denotes the support of the true parameter θ_0, T^c its complement and by ν_T we denote p-dimensional vector such that $\nu_{i,T} = \nu_i$ on set of indices $i \in T \subset \{1, \ldots, p\}$ and $\nu_{i,T} = 0$ if $i \notin T$ (similarly for ν_{T^c}). Therefore, the uniformity requirement will apply to cone $\mathcal{C}(T; 3)$ only. For that reason, we formulate our assumptions with respect to a set $\mathcal{B} \subseteq \mathbb{R}^p$, subsuming the ℓ_1-neighborhood of θ_0 in the non-convex case and the restricted cone $\mathcal{C}(T; 3)$ in the convex case.

Definition 3 ((γ, κ_n, τ_n)-Generalized Restricted Strong Convexity on a set $\mathcal{B}$). *A differentiable function $L : \Theta \to \mathbb{R}$ satisfies the GRC property with curvature γ and tolerance parameters (κ_n, τ_n) on a set $\mathcal{B}$ if its symmetric Bregman distance satisfies:*

$$\forall \nu \in \mathcal{B} : H(\theta_0 + \nu, \theta_0) \geqslant \gamma\|\nu\|_2^2 - \kappa_n\|\nu\|_1 - \tau_n\|\nu\|_1^2.$$

If loss function L is twice differentiable, then a sufficient condition for the (γ, κ_n, τ_n)-GRC property is that for all $\nu \in \mathcal{B}$ and for all $\theta \in \Theta$:

$$\nu^T \nabla_{\theta\theta} L(\theta) \nu \geqslant \gamma\|\nu\|_2^2 - \kappa_n\|\nu\|_1 - \tau_n\|\nu\|_1^2. \tag{10}$$

Moreover, if the loss is also convex, then a sufficient condition for the GRC property is that the Bregman distance at θ_0, satisfies the same lower bound for all $\nu \in \mathcal{B}$, i.e.:

$$\forall \nu \in \mathcal{B} : D(\theta_0 + \nu \mid \theta_0) \geqslant \gamma\|\nu\|_2^2 - \kappa_n\|\nu\|_1 - \tau_n\|\nu\|_1^2. \tag{11}$$

[2] The latter quantity is referred to as the symmetric Bregman distance since it corresponds to a symmetrized version of the Bregman distance, defined as: $D(\theta' \mid \theta) = L(\theta') - L(\theta) - \langle \nabla_\theta L(\theta), \theta' - \theta \rangle$, which measures how far the value of L at θ' is from the value of a linear approximation of L when it is linearized around the point θ. Observe that $H(\theta, \theta') = D(\theta' \mid \theta) + D(\theta \mid \theta')$.

The latter is the condition that was employed in the analysis of [18], which studied regualarized convex loss based estimation.

ASSUMPTION 5 ($(\gamma,\, \kappa_{n,\delta}, \tau_{n,\delta})$-GRC Empirical Oracle Loss on set $\mathcal{B}$)**.** *There exist curvature $\gamma > 0$ and tolerance parameter sequences $(\kappa_{n,\delta}, \tau_{n,\delta})$ such that the empirical oracle loss $L_S(\cdot, g_0)$ satisfies the $(\gamma,\, \kappa_{n,\delta}, \tau_{n,\delta})$-GRC condition on the set $\mathcal{B}$. w.p. $1-\delta$.*

Assumption 5 states that the empirical oracle loss $L_S(\theta, g_0)$ has a positive curvature γ in all directions $\nu \in \mathcal{B}$, allowing for the violation described by the tolerance parameters τ_n, κ_n. In our further discussion we use notation $H_S(\theta', \theta, g)$ and $H_D(\theta', \theta, g)$ to denote, respectively, symmetric Bregman distances of the empirical and population losses evaluated at parameter values θ' and θ and nuisance g. In many biostatistic and econometric applications, it is plausible to assume that the population loss $L_D(\cdot, g_0)$ is strongly convex with no violations (satisfies $(\gamma, 0, 0)$-GRC). In this case, if the difference $H_S(\theta_0 + \nu, \theta_0, g_0) - H_D(\theta_0 + \nu, \theta_0, g_0)$ between the symmetric Bregman distances of the empirical and population loss converges to zero at rate κ_n uniformly over $\nu \in \mathcal{B}$, then $(\gamma, 0, 0)$-GRC of $L_D(\theta, g_0)$ implies $(\gamma, \kappa_n, 0)$-GRC of $L_S(\cdot, g_0)$. Also, if the empirical oracle loss $L_S(\cdot, g_0)$ is twice differentiable, and its Hessian converges at rate τ_n uniformly over the set $\{\theta_0 + r\nu : \nu \in \mathcal{B}, r \in [0,1]\}$ to its population counterpart, then $(\gamma, 0, 0)$-GRC of $L_D(\cdot, g_0)$ implies $(\gamma, 0, \tau_n)$-GRC of $L_S(\cdot, g_0)$.

Lemma 1 (From Population to Empirical GRC.)**.** *Suppose the difference between the sample and the population symmetric Bregman distances normalized by $\|\nu\|_1$ converges uniformly over $\nu \in \mathcal{B}$ to zero at rate $\kappa_{n,\delta}$, i.e., w.p. $1-\delta$:*

$$\sup_{\nu \in \mathcal{B}} \frac{|H_S(\theta_0 + \nu, \theta_0, g_0) - H_D(\theta_0 + \nu, \theta_0, g_0)|}{\|\nu\|_1} \leqslant \kappa_{n,\delta}$$

Then, $(\gamma, 0, 0)$-GRC of $L_D(\cdot, g_0)$ implies $(\gamma, \kappa_{n,\delta}, 0)$-GRC of $L_S(\cdot, g_0)$ on $\mathcal{B}$ w.p. $1-\delta$.

Lemma 2 (From Population to Empirical GRC with Twice Differentiability)**.** *Suppose that $L_S(\theta, g_0)$ is twice differentiable and its empirical Hessian concentrates uniformly over $\theta \in \{\theta_0 + r\nu : \nu \in \mathcal{B}, r \in [0,1]\}$ to its population counterpart at rate $\tau_{n,\delta}/2$, i.e., w.p. $1-\delta$:*

$$\sup_{\theta \in \{\theta_0 + r\nu : \nu \in \mathcal{B}, r \in [0,1]\}} \|\nabla_{\theta\theta} L_S(\theta, g_0) - \nabla_{\theta\theta} L_D(\theta, g_0)\|_\infty \leqslant \tau_{n,\delta}, \tag{12}$$

Then, $(\gamma, 0, 0)$-GRC of $L_D(\cdot, g_0)$ implies $(\gamma, 0, \tau_{n,\delta})$-GRC of $L_S(\cdot, g_0)$ on $\mathcal{B}$ w.p. $1-\delta$.

Lipschitz in nuisance symmetric Bregman distance. To control the impact of the first-stage estimation error of $\hat{g}$ on the second-stage estimate $\hat{\theta}$, we require a final regularity assumption that the symmetric Bregman distance $H_S(\theta_0 + \nu, \theta_0, g)$ is Lipschitz in g. If the loss

$L_S(\theta, g)$ is sufficiently smooth in g and additional mild requirements, Regularity Assumption 6 is satisfied on $\mathcal{B} = \mathbb{R}^p$.

REGULARITY ASSUMPTION 6 (Lipschitz symmetric Bregman distance on $\mathcal{B}$)**.** *The empirical symmetric Bregman distance*

$$H_S(\theta_0 + \nu, \theta_0, g) = \langle \nabla_\theta L_S(\theta_0 + \nu, g) - \nabla_\theta L_S(\theta_0, g), \nu \rangle,$$

satisfies the following Lipschitz condition in g uniformly over a set $\mathcal{B}$: $\forall g, g' \in \mathcal{G}_n$, w.p. $1 - \delta$:

$$\forall \nu \in \mathcal{B} : \left| H_S(\theta_0 + \nu, \theta_0, g) - H_S(\theta_0 + \nu, \theta_0, g') \right| \leqslant L(\|g - g'\| + \xi_{n,\delta}) \|\nu\|_1^2. \tag{13}$$

The fudge factor $\xi_{n,\delta}$ is used to account for the fact that the norm of the nuisance space might be defined with respect to the population measure, while the latter assumption is about Lipschitzness of the empirical loss. Hence, typically a slack variable will be required to account for this difference in measures. For the case of sup norms over the data, $\xi_{n,\delta}$ will be zero.

Lemma 3 (Sufficient Condition for Regularity Assumption 6)**.** *Suppose the loss $L_S(\theta, g)$ is twice differentiable and its Hessian $\nabla_{\theta\theta} L_S(\theta, g)$ is L-Lipshitz in $g \in \mathcal{G}_n$ uniformly over $\mathcal{B}$: $\forall g, g' \in \mathcal{G}_n$, w.p. $1 - \delta$*

$$\sup_{\theta \in \mathcal{B}} \|\nabla_{\theta\theta} L_S(\theta, g) - \nabla_{\theta\theta} L_S(\theta, g')\|_\infty \leqslant L(\|g - g'\| + \xi_{n,\delta}),$$

Then Regularity Assumption 6 holds on set $\mathcal{B}$.

2.1 Local Optimization of a Non-Convex Loss $L_S(\theta, g)$

To state our main theorem for non-convex losses, we will also make a benign assumption that a preliminary estimator that converges to θ_0 at some preliminary rate is available. After stating the main theorem, we will discuss how one can easily construct such an estimator by either employing a convex non-orthogonal loss that is readily available in some applications, or even the same orthogonal loss $L_S(\theta, g)$ with a sufficiently large search set $\mathcal{T}$ and a more aggressive regularization weight λ. The latter implies that one does not really need a separate estimator, but rather needs to repeat the orthogonal estimation process twice with different parameters.

ASSUMPTION 7 (Preliminary Estimator)**.** *We assume that there exists a preliminary estimator $\tilde{\theta}$ such that with probability $1 - \delta$:*

$$\|\tilde{\theta} - \theta_0\|_1 \leqslant R(n, \delta)$$

Main Theorem for Non-Convex Loss $L_S(\theta, g)$ We are now ready to state our main theorem for the non-convex case. It provides a bound on the ℓ_2 and ℓ_1 errors of the Plug-in Regularized Estimation procedure.

Theorem 4 (Convergence Rate of the Plug-in Regularized Estimator)**.** *Let Assumptions 1, 2, 3, 4, 7 hold, and $\tilde{\theta}$ be a preliminary estimator. Let Assumptions 5, 6 hold on a set $\mathcal{B} = \{\theta : \|\theta - \tilde{\theta}\|_1 \leqslant R(n, \delta)\}$. Then, the Plug-in Regularized Estimator of Algorithm 1 with regularization parameter λ satisfying $\frac{\lambda}{2} \geqslant \epsilon_{n,\delta} + Bg_{n,\delta}^2 + \kappa_{n,\delta} + (\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta}))R(n, \delta)$ and search set $\mathcal{T} = \mathcal{B}$ satisfies w.p. $1 - 5\delta$:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{3\sqrt{k}}{2\gamma}\lambda \qquad\qquad \|\hat{\theta} - \theta_0\|_1 \leqslant \frac{6k}{\gamma}\lambda \tag{14}$$

Construction of a preliminary estimator. Below we provide a three-step algorithm that is a generalization of Algorithm 1, augmented by the construction of a preliminary estimator. We show how such a preliminary estimator can be constructed using the same loss with a more aggressive regularization. In Appendix A.1 we also show provide concrete rates when a convex loss is used as a preliminary step and in Appendix A.2 we provide extra conditions under which a preliminary estimator might not even be required.

Algorithm 2 Plug-in Regularized Extremum Estimator with a Preliminary Step

```
Input: Preliminary loss $L_{pre} : \Theta \times \mathcal{M} \to \mathbb{R}$ and final loss $L : \Theta \times \mathcal{G} \to \mathbb{R}$.
Input: Radii $R_0, R_1$.
Input: Preliminary and final regularization weights $\lambda_{pre}$, $\lambda_{fin}$.
1: Partition sample into three samples $S_1, S_2, S_3$ each of size $n$
2: Estimate nuisance $\hat{m} \in \mathcal{M}$ needed to obtain preliminary estimator $\tilde{\theta}$ and nuisance $\hat{g} \in \mathcal{G}$ for
   final estimator $\hat{\theta}$ on the sample $S_1$.
3: Estimate $\theta_0$ via the plug-in $\ell_1$-regularized extremum estimation, i.e.:
       $\tilde{\theta} = \arg\min_{\theta \in \mathcal{T}_{pre}} L_{pre,S_2}(\theta, \hat{m}) + \lambda_{pre}\|\theta\|_1,$   (15)
   where the search set $\mathcal{T}_{pre} = \{\theta : \|\theta\|_1 \leqslant R_0\}$ is an $\ell_1$-ball of radius $R_0$ around 0 and an
   aggressive choice of the regularization parameter $\lambda_{pre}$.
4: Estimate $\theta_0$ via the plug-in $\ell_1$-regularized extremum estimation, i.e.:
       $\hat{\theta} = \arg\min_{\theta \in \mathcal{T}_{fin}} L_{S_3}(\theta, \hat{g}) + \lambda_{fin}\|\theta\|_1,$   (16)
   where the search set $\mathcal{T}_{fin} = \{\theta : \|\theta - \tilde{\theta}\|_1 \leqslant R_1\}$ is an $\ell_1$-ball of radius $R_1$ around $\tilde{\theta}$ and
   regularization parameter $\lambda_{fin}$.
Return: $\hat{\theta}$
```

Remark 1 (Achieving Second-Order Dependence on $g_{n,\delta}$)**.** *Observe that seemingly Theorem 4 declares first order dependence of the error in* $\widehat{\theta}$ *on the first stage error* $g_{n,\delta}$*, unless* $R(n,\delta)$ *is not decaying sufficiently fast (e.g. of order* $g_{n,\delta}$*). We will now argue that in fact Theorem 4 enables a second order three step estimation algorithm outlined in Algorithm 2.*

Suppose that we have an orthogonal loss $L_S(\theta, g)$ *that satisfies all the assumptions of Theorem 4 except the existence of a preliminary estimator. Then we can still apply the theorem with* $\tilde{\theta} = 0$ *and* $R(n,\delta) = R_0$*, for some upper bound* R_0 *on* $\|\theta_0\|_1$*. For instance, if we assume that the true coefficients are all bounded by* H*, then* $R_0 = kH$*. Then the Theorem states that the resulting estimator* $\tilde{\theta}$ *achieves a rate in terms of the* ℓ_1 *norm:*

$$R_1(n, 4\delta) = \frac{8k}{\gamma}\left(\epsilon_{n,\delta} + Bg_{n,\delta}^2 + \kappa_{n,\delta} + (\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta}))R_0\right) \tag{17}$$

Subsequently, we can use $\tilde{\theta}$ *as our preliminary estimator and invoke the theorem with the latter rate* $R(n,\delta) = R_1(n,\delta)$*, which will yield a new estimator* $\hat{\theta}$ *that achieves w.p.* $1 - 8\delta$*:*

$$\begin{aligned}\|\hat{\theta} - \theta_0\|_2 \leqslant\ & \frac{3\sqrt{k}}{\gamma}\left(\epsilon_{n,\delta} + Bg_{n,\delta}^2 + \kappa_{n,\delta} + (\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta}))R_1(n, 4\delta)\right)\\ =\ & \frac{\sqrt{k}}{\gamma} O\left(\max\left\{\epsilon_{n,\delta}, \kappa_{n,\delta}, Bg_{n,\delta}^2, \tau_{n,\delta}^2 \frac{kR_0}{\gamma}, (g_{n,\delta}^2 + \xi_{n,\delta}^2)\frac{L^2\, k\, R_0}{\gamma}\right\}\right)\end{aligned}$$

where in to simplify expressions we made the assumption that $k(\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta})) = o(1)$*. Thus we have recovered a rate that has only the second order dependence on the first stage error. This result leads to the following corollary.*

Corollary 5 (Convergence Rate of Plug-in Regularized Estimator with Preliminary Step)**.** *Let* $R = \max_{\theta\in\Theta} \|\theta\|_1$ *and suppose that Assumptions 1, 2, 3, 4, 5, 6 hold on set* $\mathcal{B} = \Theta$ *with* $k(\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta})) = o(1)$*. Then the estimator returned by Algorithm 2 with the orthogonal loss* L_S *as preliminary and final loss, search radii* $R_0 = R$*, and* $R_1 = R_1(n, 4\delta)$ *defined in Equation* (17) *and regularization weights:*

$$\lambda_{pre} = 2\left(\epsilon_{n,\delta} + Bg_{n,\delta}^2 + \kappa_{n,\delta} + (\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta}))R_0\right) \tag{18}$$

$$\lambda_{fin} = 2\left(\epsilon_{n,\delta} + Bg_{n,\delta}^2 + \kappa_{n,\delta} + (\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta}))R_1(n, 4\delta)\right) \tag{19}$$

satisfies w.p. $1 - 8\delta$*:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{\sqrt{k}}{\gamma} O\left(\max\left\{\epsilon_{n,\delta}, \kappa_{n,\delta}, Bg_{n,\delta}^2, \tau_{n,\delta}^2 \frac{k\, R_0}{\gamma}, (g_{n,\delta}^2 + \xi_{n,\delta}^2)\frac{L^2\, k\, R_0}{\gamma}\right\}\right) \tag{20}$$

If $k\, R_0\, (g_{n,\delta}^2 + \xi_{n,\delta}^2) = o(\max\{\epsilon_{n,\delta}, \kappa_{n,\delta}\})$*, then the estimation error of the nuisance component can asymptotically be ignored.*

2.2 Global Convergence Under Convexity

The statement below assumes the convexity of the empirical loss $L_S(\theta, g)$ with respect to the parameter of interest θ.

ASSUMPTION 8 (Convexity of Empirical Loss)**.** *The empirical loss $L_S(\theta, g)$ is a convex function of $\theta \in \Theta$ for any g in some neighborhood of $\mathcal{G}_n$ of the true g_0.*

The convexity assumption above allows to have a weaker $(\gamma, \kappa_n, \tau_n)$-GRC requirement on empirical oracle $L_S(\theta, g_0)$ than in a non-convex case. In particular, convexity ensures that the error vector $\nu = \hat{\theta} - \theta_0$ belongs to the restricted cone $\mathcal{C}(T; 3)$. Therefore, we require Assumptions 5 and 6 to hold on a set $\mathcal{B} = \mathcal{C}(T; 3)$, as opposed to a ℓ_1 p-dimensional ball of radius r around θ_0 as it was in a non-convex case.

Moreover, observe that in such a cone, $(\gamma, \kappa_{n,\delta}, \tau_{n,\delta})$-GRC of any loss, also implies $(\gamma - 16k\tau_{n,\delta}, \kappa_{n,\delta}, 0)$-GRC. Hence, assuming that $k\tau_{n,\delta} = o(1)$, for n large enough, the latter implies $(\gamma/2, \kappa_{n,\delta}, 0)$-GRC. Being able to incorporate the third component of the GRC condition inside the first one, at the expense of a constant factor, allows us to not require a preliminary estimator, since the rate of estimator $\widehat{\theta}$ will not even depend on $R(n, \delta)$.

Main Theorem for a Convex Loss $L_S(\theta, g)$ We are now ready to state our main theorem for the convex case.

Theorem 6 (Convergence Rate of Plug-in Regularized Estimator)**.** *Let Assumptions 1, 2, 3, 4, 5, 6, 8 are satisfied for $\mathcal{B} = \mathcal{C}(T; 3)$. Moreover, assume that $k(\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta})) = o(1)$ and n is large enough such that $16\, k\, (\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta})) \leqslant \gamma/2$. Then, the Plug-in Regularized Estimator of Algorithm 1 with regularization weight λ satisfying $\frac{\lambda}{2} \geqslant \epsilon_{n,\delta} + \kappa_{n,\delta} + B\, g_{n,\delta}^2$ and a search set $\mathcal{T} = \mathbb{R}^p$ satisfies w.p. $1 - 4\delta$:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{3\sqrt{k}}{\gamma}\lambda \qquad \|\hat{\theta} - \theta_0\|_1 \leqslant \frac{12k}{\gamma}\lambda \tag{21}$$

2.3 M-Estimator Losses and Cross-Fitting

One potential drawback of Algorithm 1 is that it does not use all of the available samples in the final estimation of the target parameter θ. The latter might lead to a worse strong-convexity property of the empirical loss, which could impact the empirical performance of the approach.

In this section we focus on M-estimator losses of the form described in Equation (4), based on some M-estimator loss $\ell : W \times \Theta \times \mathcal{G} \to \mathbb{R}$ and we show that for such loss functions a more statistically efficient version of Algorithm 1 enjoys the same properties. In particular, rather than only using half of the sample for the final stage estimation, we can use the entire dataset.

However, in order to avoid a requirement for the sample complexity of the function class $\mathcal{G}$, we will take a cross-fitting approach: i) estimate $\hat{g}_S$ on sample S, and use it as the nuisance function when evaluating the M-estimator loss ℓ on each sample $i \in S'$, ii) estimate $\hat{g}_{S'}$ on samples S' and use it for each sample $i \in S$. This leads to the following cross-fitting adaptation of Algorithm 1, that is specific to M-estimator losses:

Algorithm 3 Plug-in Regularized Extremum Estimator with Cross-Fitting

```
Input: $M$-estimator loss $\ell : W \times \Theta \times \mathcal{G} \to \mathbb{R}$, regularization weight $\lambda$, search set $\mathcal{T}$.
1: Partition sample into two samples $S, S'$ each of size $n$
2: Estimate nuisance parameter $\hat{g}_S \in \mathcal{G}$ using samples $S$ and nuisance parameter $\hat{g}_{S'}$ using sample $S'$.
3: Construct the cross-fitted empirical loss function:
     $$L_{S\cup S'}(\theta, \hat{g}_S, \hat{g}_{S'}) = \frac{1}{2n}\left(\sum_{i\in S} \ell(w_i, \theta, \hat{g}_{S'}) + \sum_{i\in S'} \ell(w_i, \theta, \hat{g}_S)\right) \tag{22}$$
4: Estimate $\theta_0$ as any local minimum $\hat{\theta} \in \mathcal{T}$ of the $\ell_1$-regularized cross-fitted loss function, i.e.:
     $$\hat{\theta} = \text{local-}\min_{\theta\in\mathcal{T}} L_{S\cup S'}(\theta, \hat{g}_S, \hat{g}'_S) + \lambda\|\theta\|_1, \tag{23}$$
Return: $\hat{\theta}$
```

We note that both our main Theorems 4 and 6 and their corollaries continue to hold for this modified empirical loss function with the appropriate modifications (of technical nature) provided thatnow this loss function takes as input two nuisance components, i.e. $L_{S\cup S'}(\theta, g, g')$, which we can just view as a single larger component $g_{\text{aug}} = \{g, g'\}$. Apart from the latter modification, the only other change in our proofs arises when arguing about convergence of the empirical gradient $\nabla_\theta L_{S\cup S'}(\theta, g_{\text{aug}})$ in Lemma 16. We need to alter this proof and, rather than conditioning on the augmented nuisance component g_{aug}, we need to partition the loss into the two parts computed from samples S and S' and analyze each part separately. When analyzing S, we can condition on nuisance $g_{S'}$ and when analyzing S' we can condition on nuisance g_S. The rest of the proofs of all theorems will be identical. Hence, for conciseness we omit this analysis. Most importantly, observe that when we invoke conditions on the empirical oracle loss $L_{S\cup S'}(\theta, \{g_0, g_0\})$, then both nuisances take the same value, and hence this loss function can be viewed as our original empirical loss function with a single nuisance function but with $2n$ samples. Hence, the GRC property will be invoked with $2n$ samples rather than n samples and thereby the constants $\kappa_{2n,\delta}, \tau_{2n,\delta}$ will be replacing $\kappa_{n,\delta}, \tau_{n,\delta}$ in the bounds of our theorems, and can become substantially smaller.

2.3.1 Sufficient Conditions for M-estimators

To further ease the exposition, we now consider a further simplification of the M-estimation setting, where the nuisance component is a vector valued function $g : W \to \mathbb{R}^d$, that maps data input w into a d-dimensional vector. In such a setting, the definition of pathwise derivatives further simplifies and several regularity conditions are implied by boundedness of standard gradients. We will further assume that the M-estimator loss function takes as input the output of the nuisance component and not the component itself, i.e. $\ell : W \times \Theta \times \mathbb{R}^d \to \mathbb{R}$ and is twice diffrerentiable.

For notational simplicity, for any function $f : W \times \Theta \times \mathbb{R}^d \to \mathbb{R}^p$, denote by $\nabla_\gamma f(w, \theta, \gamma)$ its Jacobian with respect to the final input γ. Observe that in such a setting the path-wise derivatives can be expressed as functions of conventional partial derivatives:

$$\begin{aligned} D_r[g - g_0, \nabla_\theta L_D(\theta, g_0)] &= \mathbb{E}[\nabla_{\gamma\theta}\ell(w, \theta, \bar{g}_r(w))\, (g(w) - g_0(w))] \\ \left(D_r^2[g - g_0, \nabla_\theta L_D(\theta, g_0)]\right)_i &= \mathbb{E}[(g(w) - g_0(w))' \nabla_{\gamma\gamma\theta_i}\ell(w, \theta, \bar{g}_r(w))\, (g(w) - g_0(w))], \end{aligned}$$

where $\bar{g}_r = r(g - g_0) + g_0$. In this simplified setting, the following corollary provides sufficient conditions for our regularity assumptions and for our main convergence theorems:

Corollary 7 (Convergence Rate for Regular M-Estimators)**.** *Suppose that nuisance space $\mathcal{G}$ is equipped with norm $\|g\|_{\infty,1} = \sup_{w \in W} \|g(w)\|_1$ and suppose that:*

1. *Each coefficient of the true parameter is bounded by H, i.e.: $\|\theta_0\|_\infty \leqslant H$*
2. *With probability $1 - \delta$, the first stage estimator $\widehat{g}$ is in $\mathcal{G}_n$, s.t.: $\forall g \in \mathcal{G}_n : \|g - g_0\|_{\infty,1} \leqslant g_{n,\delta}$, with $kg_{n,\delta} = o(1)$.*
3. *ℓ is three times differentiable such that $\forall w \in W$, $\theta \in \Theta$, $g \in \mathcal{G}_n$:*
$$\|\nabla_\theta \ell(w, \theta_0, g(w))\|_\infty,\ \|\nabla_{\gamma\gamma\theta}\ell(w, \theta_0, g(w))\|_\infty, \|\nabla_{\gamma\theta\theta}\ell(w, \theta, g(w))\|_\infty \leqslant\ H$$
4. *Loss ℓ satisfies the orthogonality condition:*
$$\forall g \in \mathcal{G}_n : \mathbb{E}[\nabla_{\gamma\theta}\ell(w, \theta_0, g_0(w))\, (g(w) - g_0(w))] = 0$$
5. *For all $\theta \in \Theta$, the population Hessian $\mathbb{E}[\nabla_{\theta\theta}\ell(w, \theta, g_0(w))]$ has minimum eigenvalue γ_D.*
6. *The empirical oracle Hessian converges uniformly to its population counterpart: w.p. $1 - \delta$*
$$\sup_{\theta \in \Theta} \|\mathbb{E}_S[\nabla_{\theta\theta}\ell(w, \theta, g_0(w))] - \mathbb{E}[\nabla_{\theta\theta}\ell(w, \theta, g_0(w))]\|_\infty = \hat{\tau}_{n,\delta}$$
with $k\tau_{n,\delta} = o(1)$.

Then Assumptions 1, 2, 3, 4, 5, 6 are satisfied with $B = H$*,* $L = H$*,* $\epsilon_{n,\delta} = H\sqrt{\frac{\log(2p/\delta)}{2n}}$*,* $\gamma = \gamma_D$*,* $\kappa_{n,\delta} = 0$*,* $\tau_{n,\delta} = \hat{\tau}_{n,\delta}$ *and* $\xi_{n,\delta} = 0$*. Thus Algorithm 2 with the parameters defined in Corollary 5 and for* $R_0 = k\,H$*, achieves w.p.* $1 - 8\delta$ *a rate of:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant H\frac{\sqrt{k}}{\gamma_D}O\left(\max\left\{\sqrt{\frac{\log(2p/\delta)}{2n}}, \hat{\tau}_{n,\delta}^2\frac{k^2}{\gamma_D}, g_{n,\delta}^2\left(1 + \frac{H^2\,k^2}{\gamma_D}\right)\right\}\right) \tag{24}$$

If further loss ℓ *is convex in* θ *then the estimator of Algorithm 1 with parameters outlined in Theorem 6, achieves w.p.* $1 - 4\delta$ *for* n *large enough:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant H\frac{\sqrt{k}}{\gamma_D}O\left(\max\left\{\sqrt{\frac{\log(2p/\delta)}{2n}}, g_{n,\delta}^2\right\}\right) \tag{25}$$

The same rates hold for the cross-fitting version of both Algorithms.

When the nuisance space is equipped with the norm $\|g\|_{2,1} = \sqrt{\mathbb{E}\left[\|g(w)\|_1^2\right]}$ *and, furthermore,* $\sup_{w,g\in\mathcal{G}_n}\|g(w)\|_\infty \leqslant H$*, then* $\xi_{n,\delta} = d\,H\sqrt{\frac{\log(2/\delta)}{2n}}$ *and* $g_{n,\delta}^2$ *in the latter bounds should be replaced by* $\max\{g_{n,\delta}^2, \xi_{n,\delta}^2\}$*.*

2.4 Illustrative Example: Linear Heterogeneous Treatment Effects

Before proceeding to a general theory of how to satisfy our assumptions starting from a set of conditional moment restrictions, let us portray that our assumptions are not vacuous by considering a simple example, which will be a special case of our more general development in the next sections. In particular, consider our running example of linear treatment effect estimation, defined by the structural equations:

$$y = q_0(u) + (x - h_0(u)) \cdot \phi(u)'\theta_0 + \epsilon, \quad \mathbb{E}[\epsilon|x, u] = 0 \tag{26}$$

$$x = h_0(u) + \eta, \quad \mathbb{E}[\eta|u] = 0, \tag{27}$$

with bounded high-dimensional controls u, with $\|u\|_\infty \leqslant H$ and heterogeneous treatment effects with respect to a base treatment x, with $|x| \leqslant H$, and heterogeneity captured by a known high-dimensional feature vector $\phi : \mathbb{R}^d \to \mathbb{R}^p$, with $\|\phi(u)\|_\infty \leqslant H$. Suppose that $q(u)$ and $h(u)$ are sparse linear functions of u with sparsity k_q and k_h correspondingly. Moreover, assume that the overall heterogeneous treatment effect $|\phi(u)'\theta_0|$ is also bounded by H. Finally, we assume that the errors ϵ and η in the two structural equations are bounded in absolute value by H.

We consider estimation based on a Robinson style M-estimator loss:

$$\ell(w, \theta, g) = \frac{1}{2}(y - q(u) - (x - h(u)) \cdot \phi(u)'\theta)^2 \tag{28}$$

and we show that our general theory in the previous sections implies an oracle convergence rate for this setting by verifying each of the required assumptions.

Assumption 1. We can estimate the nuisance functions $g = (q, h)'$ at a rate

$$g_{n,\delta} = O\left(H \frac{\max\{k_q, k_h\}}{\gamma_u} \sqrt{\frac{\log(d/\delta)}{n}}\right),$$

with respect to the norm $\|f\|_\infty = \max_{u:\|u\|_\infty \leqslant H} \|f(u)\|_\infty$ for a d_f-dimensional vector-valued function f, where γ_u is the minimum restricted eigenvalue of $\mathbb{E}[uu']$ in directions $\nu \in C(T; 3)$. This is achieved by running a lasso regressing y on u for estimating q and x on u for h.

Assumption 2. The loss function L_D defined by the M-estimator loss is indeed orthogonal to both q and h, since:

$$\begin{aligned}
D_0[q - q_0, \nabla_\theta L_D(\theta_0, q_0)] &= -\,\mathbb{E}[(x - h_0(u)) \cdot (q(u) - q_0(u)) \cdot \phi(u)] = 0 \\
D_0[h - h_0, \nabla_\theta L_D(\theta_0, h_0)] &= -\,\mathbb{E}\big[\big(y - q_0(u) - 2(x - h_0(u))\phi(u)'\theta_0\big) \cdot (h(u) - h_0(u)) \cdot \phi(u)\big] = 0
\end{aligned}$$

Assumption 3. The second-order pathwise derivative takes the simple form:

$$D_r^2[g - g_0, \nabla_\theta L_D(\theta_0, g_0)] = 2\mathbb{E}\big[\big((q(u) - q_0(u)) \cdot (h(u) - h_0(u)) + \phi(u)'\theta_0 \cdot (h(u) - h_0(u))^2\big) \cdot \phi(u)\big]$$

Using our boundedness assumptions we can upper bound the ℓ_∞ norm of the latter by $2(H + H^2)\|g - g_0\|_\infty^2$. Hence, our Regularity Assumption 3 is satisfied with $B = 2(H + H^2)$.

Assumption 4. Since we are in an M-estimator setting and since our loss ℓ is bounded by:

$$\|\nabla_\theta \ell(w, \theta_0, g)\|_\infty \leqslant \; H \left|(y - q(u) - (x - h(u)) \cdot \phi(u)'\theta_0)(x - h(u))\right| \leqslant O(H^3)$$

Where we used the boundedness of the residuals ϵ, η and the fact that for $(q, h)' \in \mathcal{G}_n$ the first stage errors are $o(1)$ and can be ignored for sufficiently large n. Thereby, by McDiarmid's inequality, Regularity Assumption 4 is satisfied with $\epsilon_{n,\delta} = O\left(H^3 \sqrt{\frac{\log(p/\delta)}{n}}\right)$.

Assumption 5. The Hessian of the loss with respect to θ takes the simple form:

$$\nabla_{\theta\theta} L_S(\theta, g) = \frac{1}{n} \sum_{i \in S} (x_i - h(u_i))^2 \phi(u_i)\phi(u_i)' \tag{29}$$

The latter is independent of θ and hence trivially concentrates uniformly over θ to its population counterpart $\mathbb{E}[(x - h(u))^2\phi(u)\phi(u)'] = \mathbb{E}[\eta^2\phi(u)\phi(u)']$ at a rate of $\tau_{n,\delta} = O\left(\sqrt{\frac{\log(p/\delta)}{n}}\right)$. Hence, by Lemma 2, to show that L_S has the $(\gamma, 0, \tau_n)$-GRC property, it suffices to show that the population Hessian has eigenvalues lower bounded by γ. The latter holds if the conditional variance of the residual in the second structural equation (i.e. the unconfounded randomness in the treatment) is lower bounded by σ^2, i.e. $\mathbb{E}[\eta^2|u] \geqslant \sigma^2$ and further the features $\phi(u)$ have non-trivial variance in all directions, i.e. $\mathbb{E}[\phi(u)\phi(u)'] \succeq \lambda^2 I_p$ (e.g. independent Gaussian features with variance λ^2). Then, the population loss satisfies the $(\sigma^2\lambda^2, 0, 0)$-GRC.

Assumption 6. From Equation (29), observe that the Hessian is Lipschitz in $g \in \mathcal{G}_n$, since:

$$\left\|\nabla_{\theta\theta} L_S(\theta, g) - \nabla_{\theta\theta} L_S(\theta, g')\right\|_\infty \leqslant H^2 \sup_{u:\|u\|_\infty \leqslant H} |(x - h(u))^2 - (x - h'(u))^2| \leqslant 6H^3 \|g - g'\|_\infty$$

Hence, by Lemma 3, we have that Regularity Assumpiton 6 is satisfied with $L = O(H^3)$.

Convexity Assumption 8. Finally, by Equation (29), the loss $L_S(\cdot, \hat{g})$ is convex in θ, since its Hessian is non-negative definite.

Concluding. We conclude that all our conditions are satisfied, and our main Theorem 6 for convex losses is valid for this example, leading to a convergence rate of:

$$\|\hat{\theta} - \theta_0\|_2 = O_p\left(H^3 \frac{\sqrt{k}}{\sigma^2 \lambda^2}\left(\sqrt{\frac{\log(p/\delta)}{n}} + \frac{H^2 \max\{k_q^2, k_h^2\} \log(d/\delta)}{\gamma_z^2\, n}\right)\right) \tag{30}$$

assuming that $k(\tau_n + g_n) = k\left(\sqrt{\frac{\log(p/\delta)}{n}} + H\frac{\max\{k_q, k_h\}}{\gamma_z}\sqrt{\frac{\log(d/\delta)}{n}}\right) = o(1)$. Crucially, observe that in Equation (30), the error term that comes from the first stage estimation of g, is of a lower order as it decays as $1/n$. Hence, assuming that k_q, k_h grow sufficiently slow (roughly slower than $n^{1/4}$), the second term can asymptotically be ignored.

3 Sparse High-Dimensional Conditional Moment Restrictions

In this section we show how our loss minimization framework of Section 2 can be applied to the estimation of models defined by moment restrictions, popular in statistical and econometric applications. Let $\rho : \mathcal{W} \times \Theta \times \mathcal{G} \to \mathbb{R}^p$ be a vector-valued function of the data vector $w \in \mathcal{W}$, the target parameter $\theta \in \Theta$, and a nuisance parameter $g \in \mathcal{G}$. The function ρ corresponds to a valid moment condition if the true parameter $\theta_0 \in \Theta$ satisfies:

$$\mathbb{E}\left[\rho(w, \theta_0, g_0)\right] = 0, \tag{31}$$

where g_0 is the true value of g. Moreover, we say that ρ is an *identifying* moment for θ_0, if θ_0 is the unique solution to the moment equation (31), $\mathbb{E}\left[\rho(w, \theta, g_0)\right] = 0$. Without loss of generality, we assume that the output dimension of ρ is the same as the dimension of θ. We will use the following definition of an orthogonal moment.

Definition 4 (Orthogonal Moment). *A moment vector-function $\rho : \mathcal{W} \times \Theta \times \mathcal{G} \to \mathbb{R}^p$ is orthogonal with respect to the nuisance realization $\mathcal{G}_n$ set if (31) holds, the pathwise derivative map $D_r[g - g_0, \mathbb{E}\left[\rho_j(w, \theta_0, g_0)\right]]$ of each function ρ_j exists $\forall r \in [0, 1)$, $g \in \mathcal{G}_n$, and $j \in \{1, 2, .., d_\rho\}$ and vanishes at $r = 0$:*

$$\forall g \in \mathcal{G}_n : D_0[g - g_0, \mathbb{E}\left[\rho_j(w, \theta_0, g_0)\right]] = 0 \quad \forall j \in \{1, 2, ..., d_\rho\}. \tag{32}$$

To employ the plug-in ℓ_1-regularized extremum estimation approach of Section 2, we need to ensure the existence of a loss function $\ell : \mathcal{W}, \times \Theta \times \mathcal{G} \to \mathbb{R}$, whose gradient with respect to θ is equal to the orthogonal moment ρ.

Definition 5 (Orthogonal M-estimator loss)**.** *A twice differentiable function $\ell : \mathcal{W} \times \Theta \times \mathcal{G} \to \mathbb{R}$ is an orthogonal M-estimator loss, corresponding to moment function ρ, if*

$$\nabla_\theta \ell(w, \theta, g) = \rho(w, \theta, g) \quad \forall w \in \mathcal{W}, \theta \in \Theta, g \in \mathcal{G}$$

The moment ρ itself is referred to as an orthogonal loss-generating moment.

3.1 Orthogonal Loss for Single-Index Conditionally Orthogonal Moments

We will now show that an orthogonal loss always exists for a broad class of orthogonal conditional moments. In the next section we will also analyze how one can arrive at an orthogonal loss even for non-orthogonal conditional moments via a novel *loss-admitting* orthogonalization procedure.

Consider the case where the moment restriction is a conditional one and the input into the nuisance function g is a subset of the conditioning set. Suppose that the moment depends only on the d-dimensional output of the nuisance function. The statistical model is defined by nuisance functions $g : \mathcal{Z} \to \mathbb{R}^d$, where $z \in \mathcal{Z}$ is a subvector of the data-point $w \in \mathcal{W}$ and a real-valued moment function $\phi : \mathcal{W} \times \Theta \times \mathbb{R}^d \to \mathbb{R}$. With this structure in place we analyze conditional moment restrictions of the form:

$$\mathbb{E}\left[\phi(w, \theta_0, g(z)) | z\right] = 0, \forall z \in \mathcal{Z} \tag{33}$$

For this model a simpler and widely applicable condition that implies orthogonality is the following notion of conditional orthogonality:

Definition 6 (Conditionally Orthogonal Moment)**.** *A moment function $\phi : \mathcal{W} \times \Theta \times \mathbb{R}^d \to \mathbb{R}$ is conditionally orthogonal if:*

$$\mathbb{E}\left[\nabla_\gamma \phi(w, \theta_0, g_0(z)) | z\right] = 0, \forall z \in \mathcal{Z}, \tag{34}$$

where $\nabla_\gamma \phi$, denotes the gradient of ϕ with respect to its third input, i.e. the output of the nuisance function.

Observe that conditional orthogonality of ϕ implies orthogonality of any vector valued moment function ρ of the form $\phi(w, \theta_0, g_0(z))\tau(z, g_0(z))$, for any $\tau : \mathcal{Z} \times \mathbb{R}^d \to \mathbb{R}^p$, since:

$$\begin{aligned} D_0[g - g_0, \mathbb{E}\left[\phi(w, \theta_0, g_0(z))\, \tau(z)\right] = \; & \mathbb{E}\left[\tau(z, g_0(z))\, \nabla_\gamma \phi(w, \theta_0, g_0(z))' \,(g(z) - g_0(z))\right] \\ & + \mathbb{E}\left[\phi(w, \theta_0, g_0(z))\, \nabla_\gamma \tau(z, g_0(z))\, (g(z) - g_0(z))\right] = 0, \end{aligned}$$

where the last inequality follows by invoking the tower property of expectations, validity of the conditional moment restriction and conditional orthogonality. Moreover, any such ρ is valid, since $\mathbb{E}[\rho(w, \theta_0, g_0(z))] = \mathbb{E}[\mathbb{E}[\phi(w, \theta_0, g_0(z)|z]\tau(z)] = 0$.

We conclude the section by showing that when a conditionally orthogonal moment function ρ has a single-index structure, i.e. it depends on the parameter θ only through a single-dimensional index $t = \Lambda(z, g(z))'\theta$, where $\Lambda(z, g(z))$ is an arbitrary known d-dimensional vector-valued function, then it always admits an orthogonal loss. We will refer to $\Lambda(z, g(z))$ as the vector of features.

Lemma 8. *Suppose that a real-valued moment function $\phi : \mathcal{W} \times \Theta \times \mathbb{R}^d \to \mathbb{R}$ is conditionally orthogonal and has a single-index structure, i.e.:*

$$\phi(w, \theta, g(z)) \equiv \Phi(w, \Lambda(z, g(z))'\theta, g(z)) \tag{35}$$

for some known function $\Lambda : \mathcal{Z} \times \mathbb{R}^d \to \mathbb{R}^p$. Then the vector valued moment

$$\rho(w, \theta, g(z)) = \Phi(w, \Lambda(z, g(z))'\theta, g(z))\, \Lambda(z, g(z)), \tag{36}$$

is a valid orthogonal loss-generating moment. The orthogonal M-estimator loss is given by:

$$\ell(w, \theta, g(z)) = K(w, \Lambda(z, g(z))'\theta, g(z)), \tag{37}$$

where K is any solution to the Ordinary Differential Equation (ODE):

$$\frac{d}{dt} K(w, t, g(z)) = \Phi(w, t, g(z)). \tag{38}$$

For the moment ρ to be identifying, it is sufficient that $\Phi(w, t, g_0(z))$ is strictly increasing in t with derivative $\frac{\partial \Phi(w,t,g_0(z))}{\partial t}$ bounded away from zero by some $\nu > 0$ and the covariance matrix $\Sigma = \mathbb{E}[\Lambda(z, g_0(z))\Lambda(z, g_0(z))']$, has minimum eigenvalue at least $\gamma_\Sigma > 0$. In the latter case, the population loss L_D satisfies the $(\nu\gamma_\Sigma, 0, 0)$-GRC condition.

Combining Lemma 8 with a set of benign regularity conditions on the smoothness and boundedness of the conditional moment ϕ and the vector of features Λ, and invoking Corollary 7, yields the following convergence rate result. The proof of this result also uses some auxiliary uniform convergence lemmas for function classes that comprise of Lipschitz functions of linear indices, where the parameter in the index is constrained in ℓ_1 norm so as to establish condition (6) in Corollary 7.

ASSUMPTION 9 (U-smooth)**.** *The conditional moment model defined by $\{\mathcal{W}, \mathcal{Z}, \Theta, \mathcal{G}, \phi, \Phi, \Lambda\}$ is U-smooth if ϕ and Λ are twice differentiable, with derivatives bounded by some constant U,* *i.e.* $\forall w \in W$, $\theta \in \Theta$, $g \in \mathcal{G}_n$:

$$\begin{aligned} |\phi(w,\theta,g(z))|, \|\nabla_\gamma \phi(w,\theta_0,g(z))\|_\infty, \|\nabla_{\gamma\gamma}\phi(w,\theta_0,g(z))\|_\infty &\leqslant U \\ \|\Lambda(z,g(z))\|_\infty, \|\nabla_\gamma \Lambda(z,g(z))\|_\infty, \|\nabla_{\gamma\gamma}\Lambda(z,g(z))\|_\infty, \|\nabla_\gamma \Lambda(z,g(z))'\theta\|_\infty &\leqslant U \\ |\nabla_t \Phi(w,\Lambda(z,g(z))'\theta,g(z))|, |\nabla_{tt}\Phi(w,\Lambda(z,g(z))'\theta,g(z))|, \|\nabla_{\gamma t}\Phi(w,\Lambda(z,g(z))'\theta,g(z))\| &\leqslant U \end{aligned}$$

where ∇_γ is the gradient w.r.t the output of g, ∇_t is the gradient w.r.t. the index.

Corollary 9 (Convergence Rate for Single-Index Orthogonal Conditional Moments)**.** *Suppose that the nuisance space $\mathcal{G}$ is equipped with the norm $\|g\|_{\infty,1} = \sup_{w\in W}\|g(w)\|_1$ and consider real-valued conditionally orthogonal moment ϕ with single-index form Φ. Suppose that the following regularity conditions are satisfied:*

1. *Each coefficient of the true parameter is bounded by H, $\|\theta_0\|_\infty \leqslant U$ and $k = o\left(\left(\frac{n}{\log(p/\delta)}\right)^{1/4}\right)$*
2. *With probability $1-\delta$, the first stage estimateor $\hat{g}$ is in $\mathcal{G}_n$, s.t.: $\forall g \in \mathcal{G}_n : \|g - g_0\|_{\infty,1} \leqslant g_{n,\delta}$, with $k\, g_{n,\delta} = o(1)$.*
3. *The conditional moment model $\{\mathcal{W},\mathcal{Z},\Theta,\mathcal{G},\phi,\Phi,\Lambda\}$ is U-smooth, as in Assumption 9.*
4. *$\Phi(w,t,g(z))$ is increasing in t and $\mathbb{E}[\Phi(w,t,g_0(z)) \,|\, z]$ is strictly increasing in t, with partial derivative bounded from below by some $\nu > 0$ and the covariance matrix $\Sigma = \mathbb{E}[\Lambda(z,g_0(z))\Lambda(z,g_0(z))']$, has minimum eigenvalue at least $\gamma_\Sigma > 0$.*

Then all conditions of Corollary 7 are satisfied for the orthogonal loss defined in Equation (37)*, with $H = \Theta(U^3)$ and $\gamma_D = \nu\,\gamma_\Sigma$ and $\hat{\tau}_{n,\delta} = O\left(kU^5\sqrt{\frac{\log(p)}{n}} + U^3\sqrt{\frac{\log(2p/\delta)}{n}}\right)$. Moreover, the loss ℓ is convex and $k(\tau_{n,\delta} + g_{n,\delta}) = o(1)$. Hence, the estimator produced by Algorithm 1 with parameters outlined in Theorem 6, achieves w.p. $1-4\delta$ for n large enough:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant U^3 \frac{\sqrt{k}}{\nu\gamma_\Sigma} O\left(\max\left\{\sqrt{\frac{\log(2p/\delta)}{2n}}, g_{n,\delta}^2\right\}\right) \tag{39}$$

The same rates hold for the cross-fitted version of Algorithm1, given in Algorithm 3. If the nuisance space is equipped with the $\|g\|_{2,1}$ norm, instead of $\|g\|_{\infty,1}$, then $\xi_{n,\delta}^2 = O\left(\frac{d^2\log(d/\delta)}{n}\right)$ needs to be appended to the latter maximum in the rate.

3.2 Orthogonalization: Orthogonal Loss from Non-Orthogonal Moments

In the previous section we assumed that the single index conditional moment ϕ was orthogonal. However, in many settings the model is represented by a non-orthogonal conditional moment restriction:

$$\mathbb{E}[m(w,\theta_0,g_0(z))|z] = 0, \tag{40}$$

where $\mathbb{E}[\nabla_\gamma m(w,\theta_0,g_0(z))|z] \neq 0$ and $m : \mathcal{W} \times \Theta \times \mathbb{R}^d \to \mathbb{R}$. For this model we can construct an orthogonal moment by utilizing d auxiliary moment conditions that identify the nuisance function g. For simplicity, we assume that these auxiliary moments have a simple linear form:

$$\mathbb{E}[v - g_0(z)|z] = 0, \tag{41}$$

i.e. g_0 is the conditional expectation of some observed d-dimensional random vector v.[3]

A recent orthogonalization technique [7], available for the estimation of small-dimensional parametric models, can be employed to arrive at an orthogonal moment starting from a non-orthogonal one, when the model is accompanied with these auxiliary moment conditions.

Lemma 10 (Moment Orthogonalization)**.** *Consider a model defined by the conditional moment constraints in Equations* (40) *and* (41)*. Then the following moment is valid and orthogonal conditional on* z*:*

$$\phi(w, \theta, \{g_0(z), h_0(z)\}) = m(w, \theta, g_0(z)) + \underbrace{\mathbb{E}[\nabla_\gamma m(\tilde{w}, \theta_0, g_0(z))|z]'}_{h_0(z)}(v - g_0(z)) \tag{42}$$

It is also orthogonal conditional on z *w.r.t. the* d*-dimensional valued nuisance function* h_0*.*

Denote by $\tilde{g} = \{g, h\}$ the augmented nuisance parameter and re-define:

$$\phi(w, \theta, \tilde{g}(z)) = m(w, \theta, g(z)) + h(z)'\,(v - g(z)) \tag{43}$$

Observe that if the moment m has a single-index structure:

$$m(w, \theta, g(z)) = M(w, \Lambda(z)'\theta, g(z)) \tag{44}$$

then the resulting orthogonal moment ϕ given by Lemma 10 also has a single index structure:

$$\Phi(w, t, \{\gamma, \chi\}) = M(w, t, \gamma) + \chi'(v - \gamma), \tag{45}$$

where γ is the output of g and χ the output of h and the corresponding orthogonal M-estimator loss takes the form:

$$\ell(w, \theta, \{\gamma, \chi\}) = K(w, \Lambda(z, \gamma)'\theta, \gamma) + \chi'(v - \gamma)\,\Lambda(z, \gamma)'\theta \tag{46}$$

where K is any solution of the ODE: $\frac{\partial}{\partial t}K(w, t, \{\gamma, \chi\}) = M(w, t, \{\gamma, \chi\})$.

Note that the correction term $\chi'(v - \gamma)$, has no effect on the Jacobian of the moment and hence on the second order properties of the resulting M-estimator loss, since it does not depend on θ. Hence, we can directly apply Corollary 9 to get the convergence rate result.

[3] Our approach extends even for non-linear auxiliary moments $\mathbb{E}[\eta(v, g)|z] = 0$, as long as η is Frechet differentiable, the conditional expectation of the inverse of the Frechet derivative conditional on z is invertible and can be estimated at a rate $g_{n,\delta}$ on the hold out sample. However, for simplicity of exposition we omit this extension.

Corollary 11 (Estimation of Non-Orthogonal Conditional Moment)**.** *Suppose that m defines a single-index non-orthogonal conditional moment model with index form M and let $\tilde{g} = \{g, h\}$, denote the augmented nuisance parameters. Further suppose that:*

1. *Each coefficient of the true parameter is bounded by H, $\|\theta_0\|_\infty \leqslant U$ and $k = o\left(\left(\frac{n}{\log(p/\delta)}\right)^{1/4}\right)$*
2. *With probability $1-\delta$, the first stage estimate $\tilde{g}$ is in $\tilde{\mathcal{G}}_n$, s.t.: $\forall \tilde{g} \in \tilde{\mathcal{G}}_n : \|\tilde{g} - \tilde{g}_0\|_{\infty,1} \leqslant \tilde{g}_{n,\delta}$, with $k\, g_{n,\delta} = o(1)$.*
3. *The model $\{\mathcal{W}, \mathcal{Z}, \Theta, \mathcal{G}, m, M, \Lambda\}$ is U-smooth and $\forall z \in \mathcal{Z}, \tilde{g} \in \tilde{\mathcal{G}}_n$: $|h(z)'g(z)| \leqslant U$.*
4. *$M(w, t, g_0(z))$ is increasing in t and $\mathbb{E}[M(w, t, g_0(z)) \,|\, z]$ is strictly increasing in t with partial derivative bounded from below by some $\nu > 0$ and the covariance matrix $\Sigma = \mathbb{E}[\Lambda(z, g_0(z))\Lambda(z, g_0(z))']$, satisfies $\Sigma \succeq \gamma_\Sigma I$ for $\gamma_\Sigma > 0$.*

Then all conditions of Corollary 7 are satisfied for the orthogonal loss defined in Equation (46)*, with $H = \Theta(U^3)$ and $\gamma_D = \nu\,\gamma_\Sigma$ and $\hat{\tau}_{n,\delta} = O\left(kU^5\sqrt{\frac{\log(p)}{n}} + U^3\sqrt{\frac{\log(2p/\delta)}{n}}\right)$. Moreover, the loss ℓ is convex and $k(\tau_{n,\delta} + g_{n,\delta}) = o(1)$. Hence, the estimator of Algorithm 1 with parameters outlined in Theorem 6, achieves w.p. $1-4\delta$ for n large enough:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant U^3 \frac{\sqrt{k}}{\nu\gamma_\Sigma} O\left(\max\left\{\sqrt{\frac{\log(2p/\delta)}{2n}}, g_{n,\delta}^2\right\}\right) \tag{47}$$

The same rates hold for the cross-fitting version of the Algorithm, given in Algorithm 3. If the nuisance space is equipped with the $\|g\|_{2,1}$ norm then $\xi_{n,\delta}^2 = O\left(\frac{d^2\log(d/\delta)}{n}\right)$ needs to be appended to the latter maximum in the rate.

Estimating h. Even though the moment ϕ is orthogonal, it depends on the estimation of an extra nuisance parameter h. One might wonder how hard it is to estimate h_0. Generically, h_0 can be estimated in a two-step manner as follows. Suppose that we have an estimator $\hat{g}$ of the original nuisance parameter g_0, as well as a preliminary estimator $\tilde{\theta}$ of our target parameter θ_0 (converging at a rate slower than $n^{-1/2}$). Then we can estimate $\hat{h}$ in a plug-in manner from moment equation:

$$\hat{h}(z) = \mathbb{E}[\nabla_\gamma m(\tilde{w}, \tilde{\theta}, \hat{g}(z))|z] \tag{48}$$

i.e. $\hat{h}(z)$ is a regression of the random variable $\nabla_\gamma m(w, \tilde{\theta}, \hat{g}(z))$ on a vector of covariates z.

We now argue that in many cases, $h_0(z)$ is a known functional form of θ_0 and g_0. Hence, the final regression step is not required and estimating h_0 boils down to computing an estimator $\hat{g}$ of the original nuisance parameter g_0, as well as a preliminary estimator $\tilde{\theta}$ of θ_0 possibly converging at a slow rate. In the discussion below, let $\gamma = g(z)$ and $J(w, t, \gamma, \theta)$ denote the gradient of the non-orthogonal moment m w.r.t. γ, evaluated at data-point w, index t, nuisance output γ, and parameter θ, i.e.:

$$J(w, t, \gamma, \theta) \equiv \nabla_\gamma M(w, t, \gamma) + \nabla_t M(w, t, \gamma)\, \nabla_\gamma \Lambda(z, \gamma)'\theta, \tag{49}$$

where we make explicit the potential direct dependence of the gradient on the parameter θ, not only as part of the single index, due to the term $\nabla_\gamma \Lambda(z, \gamma)'\theta$. When the moment M decomposes linearly as $M(w, t, \gamma) = M_1(w) + M_2(t, \gamma)$, then h_0 takes a known functional form of θ_0, γ_0:

$$h_0(z) = \left(\nabla_\gamma M_2(t, \gamma) + \nabla_t M_2(w, t, \gamma)\, \nabla_\gamma \Lambda(z, \gamma)'\theta_0\right)\Big|_{t=\Lambda(z, g_0(z))'\theta_0,\, \gamma=g_0(z)} \tag{50}$$

Thus we can easily compute a plug-in estimator for h_0 as:

$$\hat{h} = \left(\nabla_\gamma M_2(t, \gamma) + \nabla_t M_2(w, t, \gamma)\, \nabla_\gamma \Lambda(z, \gamma)'\tilde{\theta}\right)\Big|_{t=\Lambda(z, \hat{g}(z))'\tilde{\theta},\, \gamma=\hat{g}(z)} \tag{51}$$

4 Applications

We apply the proposed M-estimation framework to three classes of problems: General Moment Problems with Missing Data, Games of Incomplete Information, and Treatment Effects in presence of a Nonlinear Link Function.

4.1 General Moment Problem with Missing Data

We consider a setting with missing data where we want to estimate a parameter θ, based on a conditional moment restriction:

$$\mathbb{E}[u(y, x'\theta_0)|x] = 0 \tag{52}$$

where $y \in \mathbb{R}^q$ and $x \in \mathbb{R}^p$. However, only some of the labels y are observed. In particular, if we denote with $d \in \{0, 1\}$ the indicator random variable that determines whether y is observed, then we assume that we only observe the quantities (x, dy). Hence, evaluating the conditional moment (52) directly is infeasible, since the variable y is not always observed.

Such models commonly occur in biostatistic and econometric applications, where the indicator d determines whether the data are corrupted due to some measurement error. A standard way to make progress in this problem is to assume that all variables $z \supseteq x$ that could have a direct effect on both the missing indicator d and the outcome y are also observable. Such variables are typically referred to as confounders or controls. This is formalized in the following assumption.

ASSUMPTION 10 (Observed Confounders (OC)). *The presence indicator d is independent from y conditional on an observed set of variables $z \subseteq x$, with $z \in \mathbb{R}^{d_z}$, i.e.: $d \perp y \,|\, z$.*

Under this restriction we can construct a feasible and valid conditional moment equation as follows: let $p_0(z) = \mathbb{E}[d|z]$ denote the propensity score of a data point missing, conditional on all the observables z. Then the single-index moment function $M(w, x'\theta, p(z)) = \frac{d\, u(y,x'\theta)}{p(z)}$ is a valid and feasible conditional moment, since due to the OC assumption and the tower law of expectations:

$$\begin{aligned}\mathbb{E}\left[M(w, x'\theta_0, p_0(z)) \,|\, x\right] =& \ \mathbb{E}\left[\mathbb{E}\left[\frac{d\, u(y, x'\theta_0)}{p_0(z)}\middle| z\right] \middle| x\right] = \mathbb{E}\left[\frac{\mathbb{E}[d \,|\, z]}{p_0(z)}\mathbb{E}[u(y, x'\theta_0) \,|\, z] \middle| x\right] \\ =& \ \mathbb{E}\left[\mathbb{E}[u(y, x'\theta_0) \,|\, z] \,|\, x\right] = \mathbb{E}[u(y, x'\theta_0) \,|\, x] = 0\end{aligned}$$

Moreover, the vector valued moment function $\rho(w, \theta, p(z)) = M(w, \theta, p(z))x$ is identifying for θ_0 if:

$$\Sigma_{\text{MD}}(\theta) = \mathbb{E}\left[\nabla_\theta \rho(w, \theta, p(z))\right] = \mathbb{E}\left[\nabla_t u(y, x'\theta) x\, x'\right] \succeq \gamma I \tag{53}$$

This is a moment equation that depends on the nuisance parameter p_0, which is typically unknown and also needs to be estimated. However, moment m is not conditionally orthogonal to p since:

$$\underbrace{\mathbb{E}\left[\nabla_\gamma M(w, x'\theta_0, p_0(z)) \,|\, z\right]}_{h_0(z)} = -\mathbb{E}\left[\frac{d\, u(y, x'\theta_0)}{p_0(z)^2} \,|\, z\right] = -\frac{1}{p_0(z)} \underbrace{\mathbb{E}\left[u(y, x'\theta_0) \,|\, z\right]}_{q_0(z)} \neq 0$$

Employing the techniques from Section 3 we will consider the orthogonal single-index conditional moment:

$$\Phi(w, x'\theta, \{p(z), h(z)\}) = M(w, x'\theta, p(z)) + h(z)\,(d - p(z)) \tag{54}$$

and its corresponding vector valued augmentation $\rho(w, \theta, \{p(z), h(z)\}) = \phi(w, \theta, \{p(z), h(z)\})x$, which is also orthogonal and loss admitting, with a convex M-estimator loss:

$$\ell(w, \theta, \{p(z), h(z)\}) = \underbrace{K(w, x'\theta, \{p(z), h(z)\})}_{\text{original loss}} + \underbrace{h(z)\,(d - p(z))\, x'\theta}_{\text{orthogonal correction}} \tag{55}$$

where K is any solution to the ODE: $\frac{\partial}{\partial t}K(w, t, \{p(z), h(z)\}) = M(w, t, \{p(z), h(z)\})$. We can then apply our M-estimation loss rates of Corollary 9 to get the following estimation result:

Corollary 12 (Orthogonal Estimation of General Moment Problem with Missing Data)**.** *Suppose that $\Sigma_{MD} \geqslant \gamma I$, $\|\theta\|_0 \leqslant k = o\left(\left(\frac{n}{\log(p/\delta)}\right)^{1/4}\right)$ and we can estimate each of the functions p and h on the hold-out sample at a rate of $g_{n,\delta}$ with respect to the root mean squared error norm $\|f\| = \sqrt{\mathbb{E}\left[f(w)^2\right]}$. Moreover, suppose that $p_0(z) \geqslant \underline{p} > 0$, all data are bounded and all functions h, p, u are bounded, twice differentiable with bounded derivatives. Then the Plug-in Regularized Extremum Estimator of Algorithm 1 with M-estimator loss defined in Equation* (55)*, regularization parameter $\lambda \sim \sqrt{\frac{\log(p/\delta)}{n}} + g_{n,\delta}^2$ and the search set $\mathcal{T} = \mathbb{R}^p$ obeys, w.p. $1-4\delta$:*

$$\|\hat{\theta} - \theta_0\|_2 = \frac{\sqrt{k}}{\gamma} O\left(\max\left\{\sqrt{\frac{\log(p/\delta)}{n}}, g_{n,\delta}^2\right\}\right) \tag{56}$$

Example estimation of p and h. We now give a concrete example where p and h can also be estimated at a rate that leads to oracle convergence for θ. Suppose that the data generating process is of the form:

$$\begin{aligned} y &= x'\theta + z'_{-x}\alpha + \zeta, & \mathbb{E}[z'_{-x}\alpha|x] = 0 \\ \Pr(d=1|z) &= \mathcal{L}(z'\beta), & \mathbb{E}[\zeta|d,z] = 0 \end{aligned}$$

where z_{-x} denotes the coordinates of z except x and $\mathcal{L}(\cdot)$ denotes the logistic function. Moreover, assume that all variables lie in a bounded range. Then $u(y, x'\theta) = x'\theta - y$ and h has a sparse linear form, while p a logistic sparse linear form. One particular example of this setting of interest is when $z_{-x} = x * v$ (where $*$ denotes pointwise multiplication) and $\mathbb{E}[v|x] = 0$. In this setting v corresponds to a treatment effect heterogeneity of the treatment x and the missing indicator is correlated with this heterogeneity variable.

If $\|\alpha\|_0, \|\beta\|_0 \leqslant k_z$, and $\mathbb{E}[zz^T|d=1]\Pr[d=1] \succeq \gamma_z I$, then function p and h can be estimated at a rate $g_{n,\delta} = O\left(\frac{k+k_z}{\gamma_z}\sqrt{\frac{\log(d_z/\delta)}{n}}\right)$. Function p can be estimated by running a logistic lasso regressing d on z to estimate α at an ℓ_1 rate $g_{n,\delta}$. Function h can be estimated by running a lasso between y and z on the non-missing data, i.e. for samples were $d=1$, to estimate both a preliminary estimate of θ and an estimate of α at an ℓ_1 rate $g_{n,\delta}$. This is a valid lasso regression, since $\mathbb{E}[\zeta|d,z] = 0$ (OC assumption). Subsequently $h(z) = \frac{z'_{-x}\alpha}{\mathcal{L}(z'\beta)}$. Thus if $(k+k_z)^2 = o\left(\frac{\sqrt{\log(p/\delta)\, n}}{\log(d_z/\delta)}\right)$, then the impact of the first stage estimation error of p and h on the error of the second stage estimate $\hat{\theta}$ can asymptotically be ignored. The main point of our orthogonal two-stage estimation result in this example is that even when k_z is growing asymptotically at a rate of $o(n^{1/4})$, we can achieve an estimate of θ that only depends on k, which could be a constant. The latter is a considerable advantage over the preliminary estimate produced by the first stage lasso described above.

4.2 Games of Incomplete Information

Consider a two-player discrete static game of incomplete information, where each player $i \in \{1,2\}$ can choose between two actions: 1 or 0. The payoffs for each player from action 1 are given by:

$$U_1 = x_1'\alpha_0^1 + y_2\Delta_0^1 + \epsilon_1, \quad \mathbb{E}[\epsilon_1|z] = 0 \tag{57}$$

$$U_2 = x_2'\alpha_0^2 + y_1\Delta_0^2 + \epsilon_1, \quad \mathbb{E}[\epsilon_2|z] = 0, \tag{58}$$

where $y_i \in \{1,0\}$ stands for an action of player i, $x_i, \alpha_0^i \in \mathcal{R}^{p-1}$ is a covariate vector that directly determines the utility of player i, and Δ_0^i is a strategic interaction parameter, showing the impact on the utility of player i of the action of his opponent for each player $i \in \{1,2\}$. In addition, each player privately observes a shock ϵ_i that is mean independent from the features $z = \{x_1, x_2\}$ that are commonly observed by both players and the researcher. In particular, we will focus on the case when ϵ_i are drawn from the Gumbel distribution, conditional on z.

Each player i makes a simultaneous choice y_i without observing her opponent's private shock ϵ_{-i}. We assume that players' choices correspond to *Bayes-Nash equilibrium* strategies in the game and, therefore, can be determines as

$$y_i = \arg\max_{y_i \in \{1,0\}} \{x_i'\alpha_0^i + \mathbb{E}[y_{-i}|z]\Delta_0^i + \epsilon_i, 0\},$$

where y_{-i} is the action of the opponent and the utility of an action 0 (outside option) is normalized to 0 for each player. Therefore, if a researcher observes samples from repeated plays of this game (where at each round each player gets a new draw of the additive logistic random variable), the estimated probabilities of choices by each player will also consistently estimate her opponent's beliefs regarding her choices.

Casting problem as estimation with nuisance component. Focusing on player $i \in \{1, 2\}$, the estimation of the model defined in (57)-(58) can be cast as the following special case of the single-index conditional moment restriction of Section 3.2:

$$\begin{aligned} \mathbb{E}[M(w, \Lambda(z, g(z))'\theta, g(z)) \,|\, z] := \mathbb{E}[\mathcal{L}(\Lambda(z, g(z))'\theta) - y \,|\, z] &= 0 \\ \mathbb{E}[v - g(z) \,|\, z] &= 0, \end{aligned}$$

where $\mathcal{L}(t) = 1/(1 + e^{-t})$ is the logistic function, $\Lambda(z, g(z)) = (x_i; g(z))$, $\theta = (\alpha^i; \Delta^i)$ is the combined vector of co-variate effects α^i and strategic effect Δ^i, the nuisance component $g : \mathbb{R}^{2p-2} \to \mathbb{R}$ is the conditional probability of opponent entry and $v \in \{0, 1\}$ is an action of the opponent. The corresponding vector valued moment $\rho(w, t, g(z)) = M(w, t, g(z))\, \Lambda(z, g(z))$ is identifying if

$$\Sigma_{\text{games}} := \mathbb{E}[\Lambda(z, g_0(z))\Lambda(z, g_0(z))'] \succeq \gamma I, \qquad \gamma > 0 \tag{59}$$

and if the value of the index $t \in \{\Lambda(z, g_0(z))'\theta : z \in \mathcal{Z}, \theta \in \Theta\}$ has a bounded range, with a constant bound, since then $\mathbb{E}[\nabla_t M(w, t, g(z)) \,|\, z] = \mathcal{L}(t)\,(1 - \mathcal{L}(t)) \geqslant \nu > 0$.

Moment M is not orthogonal with respect to g. However, we can use the technique developed in Section 3.2 to arrive at the orthogonal conditional moment:

$$\Phi(w, \Lambda(z, g(z))'\theta, \{g(z), h(z)\}) := \mathcal{L}(\Lambda(z, g(z))'\theta) - y + h(z)(v - g(z)), \tag{60}$$

where $h_0(z) = \Delta_0^i\, \mathcal{L}'(\Lambda(z, g_0(z))'\theta_0)$. Then we arrive at the corresponding orthogonal M-estimator loss which is also globally convex:

$$\ell(w, \theta, \{g(z), h(z)\}) := \underbrace{K(w, \Lambda(z, g(z))'\theta, g(z))}_{\text{non-orthogonal logistic loss}} + \underbrace{h(z)\,(v - g(z))\,\Lambda(z, g(z))'\theta}_{\text{orthogonal correction}}. \tag{61}$$

For this case, a solution to the ODE $\frac{\partial K(w,t,g(z))}{\partial t} = \mathcal{L}(t) - y$ is the logistic loss:

$$K(w,t,g(z)) := -y \log \mathcal{L}(t) - (1-y)\log(1-\mathcal{L}(t)) \tag{62}$$

We can then apply Corollary 11 to get the following estimation rate:

Corollary 13 (Orthogonal Estimation of Games of Incomplete Information)**.** *Suppose that* $\Sigma_{games} \geqslant \gamma I$*,* $\|\theta_0\|_0 \leqslant k = o\left(\left(\frac{n}{\log(p/\delta)}\right)^{1/4}\right)$*. Suppose that we can estimate the functions* g *and* h *at a rate* $\tilde{g}_{n,\delta}$ *with respect to the root mean squared error norm* $\|f\| = \sqrt{\mathbb{E}\left[f(w)^2\right]}$*, with* $k\tilde{g}_{n,\delta} = o(1)$*. Moreover, suppose that all data are bounded and* $\|\theta\|_\infty$ *is bounded for any* $\theta \in \Theta$*. Let* $\mathcal{L}(t)\left(1-\mathcal{L}(t)\right) \geqslant \nu$ *for any* $t \in \{\Lambda(z, g_0(z))'\theta : \theta \in \Theta, z \in \mathcal{Z}\}$*. Then the estimate of Algorithm 1 with M-estimator loss defined in Equation* (61)*, regularization parameter* $\lambda \sim \sqrt{\frac{\log(p/\delta)}{n}} + g_{n,\delta}^2$ *and the search set* $\mathcal{T} = \mathbb{R}^p$ *obeys, w.p.* $1-4\delta$*:*

$$\|\hat{\theta} - \theta_0\|_2 = \frac{\sqrt{k}}{\nu\gamma} O\left(\max\left\{\sqrt{\frac{\log(p/\delta)}{n}}, g_{n,\delta}^2\right\}\right) \tag{63}$$

Estimating g and h. If we assume that g follows a high-dimensional logistic parametric form: $g(z) = \mathcal{L}(\tau(z)'\psi)$ for some sparse parameter ψ with $\|\psi\|_0 \leqslant k_z$, then we can compute an estimate $\hat{g}$ of g_0 via a logistic lasso at a rate of $g_{n,\delta} = O\left(\frac{k_z}{\gamma_\tau}\sqrt{\frac{\log(p/\delta)}{n}}\right)$, where γ_τ is the minimum eigenvalue of $\mathbb{E}[\tau(z)\tau(z)']$. Moreover, an estimate $\hat{h}$ of h_0 can be estimated in a plugin manner using the estimate $\hat{g}$ of g_0 as well as a preliminary estimate $\tilde{\theta}$ of θ_0 and evaluating $\hat{h}(z) = \tilde{\Delta}^i \mathcal{L}'(\Lambda(z, \hat{g}(z))'\tilde{\theta})$. The preliminary estimate $\tilde{\theta}$ can be obtained via a non-orthogonal logistic lasso regressing y_i on $x_i, \hat{g}(z)$. The latter will lead to an ℓ_1 rate for $\tilde{\theta}$ of the order of $O\left(\frac{k}{\gamma}\left(\sqrt{\frac{\log(p/\delta)}{n}} + g_{n,\delta}\right)\right)$, where γ is the minimum eigenvalue of Σ_{games}. If k and k_z grow sufficiently slow, i.e. $k\, k_z = o\left(\left(\frac{n}{\log(p/\delta)}\right)^{-1/4}\right)$, then the estimation error of both $\hat{h}$ and $\hat{g}$ can asymptotically be ignored.

4.3 Nonlinear Treatment Effects

Consider the following model of Nonlinear Treatment Effects:

$$\mathbb{E}\left[G(x'\theta_0 + f_0(u)) - y|x,u\right] = 0, \tag{64}$$

where $x \in \mathbb{R}^p$ is a high-dimensional treatment vector, $u \in \mathcal{U}$ is a control vector that affects the outcome y through a composition of a partially linear index $x'\theta_0 + f_0(u)$ and a nonlinear known monotonically increasing link function $G : \mathbb{R} \to \mathbb{R}$. Define $f_0^\theta(u)$ as follows:

$$f_0^\theta(u) := \arg\min_{f \in \mathcal{M}} \mathbb{E}(G^{-1}(\mathbb{E}[y|x,u]) - x'\theta - f(u))^2, \tag{65}$$

where the set $\mathcal{M} = \{f(u) : \mathcal{U} \to \mathbb{R}\}$ consists of square integrable functions of u. Then:

$$f_0^\theta(u) := q_0(u) - h_0(u)'\theta, \tag{66}$$

where $q_0(u) := \mathbb{E}\big[G^{-1}(\mathbb{E}[y|x,u])|u\big]$ is the Conditional Expectation Function of $G^{-1}(\mathbb{E}[y|x,u])$ and $h_0(u) := \mathbb{E}[x|u]$ is the Conditional Expectation Function of the treatment. Observe that we can also re-write q_0 as: $q_0(u) = \mathbb{E}\left[x'\theta_0 + f_0(u)\right] = h_0(u)'\theta + f_0(u)$. This leads to the modified moment:

$$\mathbb{E}\left[G((x - h_0(u))'\theta_0 + q_0(u)) - y|x,u\right] = 0, \tag{67}$$

Unlike the linear case (i.e. $G(t) = t$), the latter moment is still not orthogonal with respect to h and q. The non-linearity requires us to make one extra modification to the moment: at a high level, optimally weight the moment by dividing by the conditional variance, to take away heteroskedasticiy. This modification turns out to lead to an orthogonal moment.

Casting problem as estimation with nuisance component. More formally, we will consider the following instance of the single-index model presented in Section 3: let $z = (x; u)$, $g(z) = \{h(u), q(u), V(z)\}$, with

$$V_0(z) = G'(x'\theta_0 + f_0(u)),$$

$\Lambda(z, g(z)) = x - h(u)$ and consider moment condition:

$$\Phi(w, \Lambda(z, g(z))'\theta, g(z)) := \frac{1}{V(z)}(G(\Lambda(z, g(z))'\theta + q(u)) - y) \tag{68}$$

Then we observe that the vector valued moment

$$\rho(w, \Lambda(z, g(z))'\theta, g(z)) := \Phi(w, \Lambda(z, g(z))'\theta, g(z))\, \Lambda(z, g(z)), \tag{69}$$

is orthogonal with respect to g (see proof in Appendix D). Moreover, it is identifying if for all $t \in \{(x - h_0(z))'\theta : \theta \in \Theta, w \in \mathcal{W}\}$:

$$\mathbb{E}\left[\nabla_t \Phi(w, t, g_0(z))|z\right] = \frac{G'(t + q_0(u))}{G'((x - h_0(u))'\theta_0 + f_0(u))} \geqslant \nu > 0 \tag{70}$$

and if:

$$\Sigma_{\mathrm{TE}} := \mathbb{E}[(x - h_0(z))\,(x - h_0(z))'] \geqslant \gamma I, \qquad \gamma > 0 \tag{71}$$

Moreover, since it has a single-index form, then ρ is also loss-generating with the orthogonal M-estimator loss $\ell(w, \theta, g(z))$ given by:

$$\ell(w, \theta, g(z)) := K(w, (x - h(u))'\theta + q(u), g(z)) \tag{72}$$

where K is a solution to the ODE

$$\frac{\partial}{\partial t}K(w,t,g(z)) = \frac{G(t)-y}{V(z)},\ t \in \mathbb{R}. \tag{73}$$

Now we can apply Corollary 9 to get a convergence rate result.

Corollary 14 (Orthogonal Estimation of Non-linear Treatment Effects)**.** *Suppose that $\Sigma_{TE} \geqslant \gamma I$, $\|\theta_0\|_0 \leqslant k = o\left(\left(\frac{n}{\log(p/\delta)}\right)^{1/4}\right)$. Suppose that we can estimate the functions q, h and V at a rate $\tilde{g}_{n,\delta}$ with respect to the $\ell_{\infty,1}$ norm $\|f\|_{\infty,1} = \sup_w \|f(w)\|_1$, with $k\tilde{g}_{n,\delta} = o(1)$. Moreover, suppose that all data are bounded, $\|\theta\|_\infty$ is bounded for any $\theta \in \Theta$ is bounded by a constant. Let $\inf_{t\in\mathcal{I}} G'(t) \geqslant \nu > 0$ where $\mathcal{I}$ is the index space $\mathcal{I} := \{(x-h(z))'\theta + q(u) : \theta \in \Theta, \{h,q,V\} \in \mathcal{G}_n, w \in \mathcal{W}\}$. Then the estimate of Algorithm 1 with M-estimator loss defined in Equation* (72)*, regularization parameter $\lambda \sim \sqrt{\frac{\log(p/\delta)}{n}} + g_{n,\delta}^2$ and the search set $\mathcal{T} = \mathbb{R}^p$ obeys, w.p. $1-4\delta$:*

$$\|\hat{\theta} - \theta_0\|_2 = \frac{\sqrt{k}}{\nu\gamma} O\left(\max\left\{\sqrt{\frac{\log(p/\delta)}{n}}, g_{n,\delta}^2\right\}\right) \tag{74}$$

Example 2 (Discrete Choice)**.** *When the link function G is the logistic function $\mathcal{L}$, then the resulting orthogonal loss in Equation* (72) *corresponds to the weighted logistic maximum likelihood estimator with an offset $q(u)$ and sample weight $V(z)^{-1}$:*

$$\ell(w,\theta,g(z)) = \frac{y\cdot\log\left(\mathcal{L}\left((x-h(u))'\theta + q(u)\right)\right) + (1-y)\cdot\log\left(1-\mathcal{L}((x-h(u))'\theta + q(u))\right)}{V(z)}$$

The latter link function would arise when $x'\theta_0 + f_0(u)$ corresponds to the utility of an agent in a binary discrete choice problem and the unobserved heterogeneity follows a Gumbel distribution. In this, setting θ_0 corresponds to the effect of co-variate vector x on the utility of the agent. Our approach also extends to more than two choices, albeit with some extra technical arguments, which we omit for simplicity of exposition.

Example 3 (Partially Linear Regression Model)**.** *In the case of the Partially Linear Regression Model (PLR) of [19], the link function G is the identity and the parameter θ_0 is interpreted as the causal effect of co-variate x on an outcome y. In this setting, the resulting orthogonal loss in Equation* (72) *is simply the squared loss:*

$$\ell(w,\theta,g(z)) = \frac{1}{2}((x-h(u))'\theta + q(u) - y)^2$$

An illustrative exposition of this setting was given in Section 2.4.

Estimating q, h and V. First, observe that h has a high dimensional output of size $d = p$. Hence, in the worst-case computing an estimate $\hat{h}$ of h with respect to an $\ell_{\infty,1}$ norm would scale linearly with p if there is no relation across the coordinates of h. Hence, we will need to make an assumption that h is really determined by a lower dimensional nuisance function $\pi(u)$, which is then projected back to a high dimensional space via a known linear transform: $B(u)\pi(u)$. If we further assume that $\|B(u)\|_{\infty,1} = \max_i \|B_i(u)\|_1 \leqslant H$, then it suffices to estimate the lower dimensional nuisance π in $\ell_{\infty,\infty}$ norm. For instance, if the high-dimensional treatment x is really determined by lower dimensional treatment $\tau \in \mathbb{R}^d_\tau$, with d_τ being a constant, via a known linear transform: $B(u)'\tau$, then observe that $\mathbb{E}[x|u] = B(u)\mathbb{E}[\tau|u] = B(u)\pi_0(u)$, with $\pi_0(u) = \mathbb{E}[\tau|u]$. If further the base treatments have a sparse linear relation with u, i.e. $\pi_{0,j}(u) = u'\gamma_{0.j}$ with $\|\gamma_{0,j}\|_0 \leqslant k_\gamma$, then we can compute an estimate $\hat{\pi}_j$ of each $\pi_{0,j}$ via a lasso, regressing τ_j on u. Further, if we assume that $f_0(u) = u'\alpha_0$ for some sparse coefficient, $\|\alpha_0\|_0 \leqslant k_\alpha$, then we can compute a preliminary estimate of $\tilde{\theta}$ and $\hat{\alpha}$ by running a non-orthogonal non-linear lasso based on the loss $\int_0^{x'\theta+u'\alpha}(G(t)-y)dt$. With this estimates at hand, we can estimate q_0 and V_0 in a plugin manner:

$$\begin{aligned}\hat{q}(u) &= \hat{h}(u)'\tilde{\theta} + u'\hat{\alpha} = \hat{\pi}(u)'B(u)'\tilde{\theta} + u'\hat{\alpha}\\ \hat{V}(z) &= G'(x'\tilde{\theta} + u'\hat{\alpha})\end{aligned}$$

Assuming that k_γ and k_α grow at a sufficiently some slow rate of $n/\log(p)$, then the estimation error of the nuisance component can asymptotically be ignored.

An example of the latter setting is the heterogeneous treatment effect motivation presented as an illustrative application in Section 2. We note that the analysis in that section is slightly more refined as it does not put an $\ell_{\infty,1}$ constraint on the matrix B but rather an ℓ_∞ constraint on the low-dimensional vector $B(u)'\theta_0$. The latter can also be done at the level of generality of this section if one makes the low dimensional assumption on the treatment from the beginning and treats π rather than h as the nuisance component. We omit this refinement for simplicity.

Remark 2. *We note that an alternative orthogonal moment can be constructed as*

$$\tilde{\rho}(w, \Lambda(z, g(z))'\theta, g(z)) := \frac{1}{\tilde{V}(z)}(G(\Lambda(z, g(z))'\theta + q(u)) - y)G'(\Lambda(z, g(z))'\theta + q(u)), \tag{75}$$

with $\tilde{V}_0(z) = (G'(x'\theta_0 + f_0(u)))^2$. The orthogonal M-estimator loss that generates this moment takes the form

$$\tilde{\ell}(w, \theta, g(z)) = \frac{1}{2}\left(\frac{G(\Lambda(z, g(z))'\theta + q(u)) - y}{\sqrt{\tilde{V}(z)}}\right)^2.$$

This loss corresponds to the classic nonlinear least squares estimator. These observations recover the relationship between the losses produced by our approach and the standard losses used for regression models. We do not further analyze the nonlinear least squares loss, however, given that it is not guaranteed to be globally convex.

4.4 Single Index Unknown Monotone Link Models

Our framework also applies to the setting of the Single Index Model:

$$\mathbb{E}\left[y - G(x'\theta_0)|x\right] = 0, \tag{76}$$

where $x \in \mathbb{R}^p$ is a vector of regressors that affects the outcome $y \in \mathbb{R}$ through an unknown monotonically increasing link function $G : \mathbb{R} \to \mathbb{R}$ with the normalization $G(0) = 0$. Note that the difference between this model and the treatment effect model that we considered before is that the link function $G(\cdot)$ now needs to be estimated and starts playing the role of a nuisance function.

Models of this type generalize parametric nonlinear regression problems and have been extensively studied in Econometrics and Statistics literature in cases where parameter vector θ is finite-dimensional, e.g. see [13], [11] and [12].

The orthogonal moment for (76) is obtained by defining a vector $h_0 = E[x]$ and a scalar $q_0 = h_0'\theta_0$ with the moment taking form

$$\rho(w, \theta, g) = (y - G(\langle x - h, \theta\rangle + q))[x - h] + G(\langle x - h, \tilde{\theta}\rangle + q)[x - h],$$

where nuisance g includes G, h and preliminary estimates $\tilde{\theta}$ and $\tilde{q}$. In this objective nuisance h is allowed to have the first-order effect on the second stage given that it is a sample mean of x. This loss, however, is orthogonal with respect to slower converging nuisances G and q. The corresponding orthogonal M-estimator loss can be obtained from

$$\frac{d}{dt}K(w, t, q) = y - G(t + q) + G(\langle x - h(z), \tilde{\theta}\rangle + q(z))[x - h(z)]$$

by setting $\ell(w, \theta, g) = K(w, \langle\theta, x - h\rangle, q)$. The preliminary estimates can be obtained by splitting the initial sample S_1 into subsamples S_1' and S_1'' where the subsample S_1' is used to optimize convex but non-orthogonal loss

$$\ell(w, \theta, g) = \int_0^{\langle x - h, \theta\rangle + q} (y - G(t))\, dt$$

for each G and subsample S_1'' is used to run standard nonparametric regression of y on $G(\langle x - \tilde{h}, \tilde{\theta}\rangle + \tilde{q})$ where $\tilde{h}$, $\tilde{\theta}$ and $\tilde{q}$ are obtained from the first subsample.

References

[1] Sreangsu Acharyya, Arindam Banerjee, and Daniel Boley. Bregman divergences and triangle inequality. *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013.

[2] Patrick Bajari, Han Hong, John Krainer, and Denis Nekipelov. Estimating static models of strategic interactions. *Journal of Business & Economic Statistics*, 28(4):469–482, 2010.

[3] Patrick Bajari, Han Hong, and Denis Nekipelov. Game theory and econometrics: A survey of some recent research. In *Advances in economics and econometrics, 10th world congress*, volume 3, pages 3–52, 2013.

[4] Raymond J Carroll, David Ruppert, and Leonard A Stefanski. Measurement error in nonlinear models. number 63 in monographs on statistics and applied probability, 1995.

[5] Raymond J Carroll and MP Wand. Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 573–585, 1991.

[6] Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, pages 808–843, 2008.

[7] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/Debiased Machine Learning for Treatment and Causal Parameters. *ArXiv e-prints*, July 2016.

[8] Victor Chernozhukov, Matt Goldman, Vira Semenova, and Matt Taddy. Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*, 2017.

[9] Bryan Graham. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 2011.

[10] Bryan Graham. Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 2012.

[11] Wolfgang Hardle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The annals of Statistics*, pages 157–178, 1993.

[12] Joel L Horowitz and Wolfgang Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.

[13] Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

[14] Lung-fei Lee and Jungsywan H Sepanski. Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, 90(429):130–140, 1995.

[15] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *Ann. Statist.*, 45(2):866–896, 04 2017.

[16] Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 06 2012.

[17] Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[18] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 11 2012.

[19] Peter Mark Robinson. Root- n-consistent semiparametric regression. *Econometrica*, 1988.

[20] Jungsywan H Sepanski and Raymond J Carroll. Semiparametric quasilikelihood and variance function estimation in measurement error models. *Journal of Econometrics*, 58(1-2):223–256, 1993.

[21] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

A Omitted Remarks from Section 2

A.1 Preliminary Estimator from a Convex Non-Orthogonal Loss

In some cases a preliminary rate can be achieved in a computationally more desirable manner by using a convex non-orthogonal loss $L_{pre,S} : \Theta \times \mathcal{M} \to \mathbb{R}$. That loss could itself be dependent on a separate nuisance component $m \in \mathcal{M}$, which is estimable at a rate $\mu_{n,\delta}$ on a hold out set of size n (and let $\mathcal{M}_n$ denote the shrinking sets). If this loss also satisfies the $(\gamma^{pre}, \kappa_{n,\delta}^{pre}, \tau_{n,\delta}^{pre})$-GRC condition, defined in Assumption 5, convergence of the gradient defined in Assumption 4 at a rate $\epsilon_{n,\delta}^{pre}$ and the Lipschitz symmetric Bregman distance condition defined in Assumption 6 with constant $B^{pre}, \xi_{n,\delta}^{pre}$ uniformly on a set $\mathcal{B} = \mathcal{C}(T; 3)$, then a slight modification of our proof of the main Theorem 6 for convex losses in the next section, shows that such an estimator achieves an

ℓ_1 rate of the order of

$$R(n, 3\delta) = \frac{16k}{\gamma^{pre}}(\epsilon^{pre}_{n,\delta} + \kappa^{pre}_{n,\delta} + K^{pre}\mu_{n,\delta}) \tag{77}$$

with probability $1 - 3\delta$ and for n large enough such that $k(\tau^{pre}_{n,\delta} + B^{pre}(\mu_{n,\delta} + \xi^{pre}_{n,\delta})) \leqslant \frac{\gamma^{pre}}{2}$, where K^{pre} is an upper bound on the pathwise derivative of $L_{pre,S}$:

$$\forall \hat{m} \in \mathcal{M}_n : D_r[\hat{m} - m_0, \nabla_\theta L_{pre,D}(\theta_0, m_0] \leqslant K^{pre}\|\hat{m} - m_0\| \tag{78}$$

Then, $\hat{\theta}$ chosen as output of Algorithm 2 with respective losses $L_{pre,S}(\theta, m)$ and $L_S(\theta, g)$ achieves the following rate:

Corollary 15 (Convergence Rate of Plug-in Regularized Estimator with Convex Non-Orthogonal Preliminary Step)**.** *Let $L_{pre,S} : \Theta \times \mathcal{M} \to \mathbb{R}$ be a convex non-orthogonal loss that satisfies the Assumptions defined in Remark A.1 and L_S a non-convex orthogonal loss that satisfies Assumptions 1, 2, 3, 4, 5, 6 on a set $\mathcal{B} = \Theta$. Moreover, assume that $\epsilon^{pre}_{n,\delta} = \epsilon_{n,\delta}$, $\kappa^{pre}_{n,\delta} = \kappa_{n,\delta}$, $k(\tau^{pre}_{n,\delta} + B^{pre}(\mu_{n,\delta} + \xi^{pre}_{n,\delta})) = o(1)$ and $k(\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta})) = o(1)$. Then the estimate returned by Algorithm 2 with the convex non-orthogonal loss $L_{pre,S}$ as a preliminary loss, the non-convex orthogonal loss L_S as final loss, search radii $R_0 = \infty$, and $R_1 = R(n, 3\delta)$ defined in Equation* (77) *and regularization weights:*

$$\lambda_{pre} = 2\left(\epsilon^{pre}_{n,\delta} + \kappa^{pre}_{n,\delta} + K^{pre}\mu_{n,\delta}\right) \tag{79}$$

$$\lambda_{fin} = 2\left(\epsilon_{n,\delta} + Bg^2_{n,\delta} + \kappa_{n,\delta} + (\tau_{n,\delta} + L(g_{n,\delta} + \xi_{n,\delta}))R(n, 3\delta)\right) \tag{80}$$

satisfies for n large enough such that $k(\tau^{pre}_{n,\delta} + B^{pre}(\mu_{n,\delta} + \xi^{pre}_{n,\delta})) \leqslant \frac{\gamma^{pre}}{2}$ w.p. $1 - 7\delta$:

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{\sqrt{k}}{\gamma} O\left(\max\left\{\epsilon_{n,\delta}, \kappa_{n,\delta}, Bg^2_{n,\delta}, \frac{kK^{pre}\tau_{n,\delta}\ (\mu_{n,\delta} + \xi^{pre}_{n,\delta})}{\gamma^{pre}}, \frac{K^{pre}Lk}{\gamma^{pre}}(g_{n,\delta} + \xi_{n,\delta})(\mu_{n,\delta} + \xi^{pre}_{n,\delta})\right\}\right)$$

If $g^2_{n,\delta}$, $k(\mu_{n,\delta} + \xi^{pre}_{n,\delta})\tau_{n,\delta}$ and $k(g_{n,\delta} + \xi_{n,\delta})(\mu_{n,\delta} + \xi^{pre}_{n,\delta})$ is of lower order than $\epsilon_{n,\delta}, \kappa_{n,\delta}$, then the estimation error of the nuisance component can asymptotically be ignored.

A.2 No Need in Preliminary Estimator for Non-Convex Losses

Remark 3 (No Need in Preliminary Estimator in case of Uniform Orthogonality and Uniform Gradient Convergence)**.** *Let us conclude the section with an additional assumption, under which the preliminary estimator is not required.*

Definition 7 (Uniform Orthogonality)**.** *A loss function $L_D : \Theta \times \mathcal{G} \to \mathbb{R}$ is called uniformly orthogonal over Θ w.r.t a nuisance parameter $g \in \mathcal{G}$ if its pathwise derivative w.r.t. g is zero for all $\theta \in \Theta$:*

$$D_0[g - g_0, L_D(\theta, g_0)] = 0 \quad \forall \theta \in \Theta.$$

Let Assumptions 1, 3, and 5, 6 hold on a set $\mathcal{B} = \{\theta : \|\theta\|_1 \leqslant R_0\}$ for some constant R_0. In addition, suppose that the loss function $L_D(\theta, g)$ is uniformly orthogonal in $\theta \in \mathcal{B}$ in the sense of Definition 7 and that the gradient of $L_S(\theta, g)$ converges at a rate $\epsilon_{n,\delta}$ uniformly over the set $\mathcal{B}$, i.e. for any $g \in \mathcal{G}_n$ w.p. $1 - \delta$:

$$\sup_{\theta \in \mathcal{B}} \|\nabla_\theta L_S(\theta, g) - \nabla_\theta L_D(\theta, g)\| \leqslant \epsilon_{n,\delta} \tag{81}$$

Then, the Plug-in Regularized Estimator of Algorithm 1 with the regularization parameter $\lambda \geqslant 2(\epsilon_{n,\delta} + B\, g^2_{n,\delta} + \kappa_{n,\delta} + 2\, B\, g^2_{n,\delta} + 2\epsilon_{n,\delta} + \tau_{n,\delta}\, R_0)$ and the search set $\mathcal{T} = \mathcal{B}$ converges to θ_0 at rate (14) with probability $1 - 4\delta$. Observe that the latter rate has a second order impact from the first stage errors, even when the preliminary rate R_0, albeit bounded, is not shrinking to zero at any rate, i.e. w.p. $1 - 4\delta$:

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{\sqrt{k}}{\gamma} O\left(\max\{\epsilon_{n,\delta}, \kappa_{n,\delta}, \tau_{n,\delta}\, R_0, B g^2_{n,\delta}\}\right) \tag{82}$$

A rough proof sketch of the latter remark is that under the latter uniform assumptions it is easy to show that the $(\gamma, \kappa_{n,\delta}, \tau_{n,\delta})$-GRC condition of the empirical oracle loss implies the $(\gamma, \kappa_{n,\delta} + 2B\, g^2_{n,\delta} + 2\epsilon_{n,\delta}, \tau_{n,\delta})$-GRC condition of the empirical plugin loss. This follows by applying uniform convergence followed by uniform orthogonality to the definition of the symmetric Bregman distance. The latter alters the step in Equation (90) *in our proof of Theorem 4. The rest of the proof remains almost identical.*

B Omitted Proofs for Section 2

Proof of Lemma 1. With probability $1 - \delta$:

$$\begin{aligned} H_S(\theta_0 + \nu, \theta_0, g_0) &\geqslant H_D(\theta_0 + \nu, \theta_0, g_0) - \sup_{\nu \in \mathcal{B}} \frac{|H_D(\theta_0 + \nu, \theta_0, g_0) - H_S(\theta_0 + \nu, \theta_0, g_0)|}{\|\nu\|_1} \|\nu\|_1 \\ &\geqslant \gamma \|\nu\|_2^2 - \kappa_{n,\delta} \|\nu\|_1 \end{aligned}$$

□

Proof of Lemma 2. On the event $\mathcal{E}_2 := \{\sup_{\theta \in \Theta} \|\nabla_{\theta\theta} L_S(\theta, g_0) - \nabla_{\theta\theta} L_D(\theta, g_0)\|_\infty < \tau_{n,\delta}\}$, which occurs with probability $1-\delta$:

$$\begin{aligned} H_S(\theta_0+\nu, \theta_0, g_0) =& \int_0^1 \frac{\partial}{\partial r} \nu^T \nabla_\theta L_S(\theta_0 + r\nu) dr \\ =& \int_0^1 \nu^T \nabla_{\theta\theta} L_S(\theta_0 + r\nu)\nu dr \\ \geqslant& \int_0^1 \nu^T \nabla_{\theta\theta} L_D(\theta_0 + r\nu)\nu dr - \tau_{n,\delta}\|\nu\|_1^2 \\ =& H_D(\theta_0+\nu, \theta_0) - \tau_{n,\delta}\|\nu\|_1^2 \\ \geqslant& \gamma\|\nu\|_2^2 - \tau_{n,\delta}\|\nu\|_1^2 \end{aligned}$$

□

B.1 Proof of Main Theorem for Non-Convex Losses

First, let us prove that orthogonality of the loss (Assumptions 2 and 3) implies that the noise of the problem, summarized by $\|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty$, decays at least as fast as $O_p\left(g_n^2 + \epsilon_n\right)$.

Lemma 16 (Second Order Influence on Oracle Gradient)**.** *Assumptions 1, 2, 3 and 4 imply that w.p.* $1-2\delta$*:*

$$\|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty \leqslant \epsilon_{n,\delta} + Bg_{n,\delta}^2 \tag{83}$$

Proof of Lemma 16. We consider the event that the probabilistic statements in Assumptions 1 and 4 occur. The latter happens with probability at least $1-2\delta$, by the union bound. We further condition on the value of the first stage estimate to be $\hat{g} \in \mathcal{G}_n$.

By the first-order optimality condition, $\nabla_\theta L_D(\theta_0, g_0) = 0$ and by the triangle inequality:

$$\|\nabla L_S(\theta_0, \hat{g})\|_\infty \leqslant \|\nabla_\theta[L_D(\theta_0, \hat{g}) - L_D(\theta_0, g_0)]\|_\infty + \|\nabla_\theta[L_S(\theta_0, \hat{g}) - L_D(\theta_0, \hat{g})]\|_\infty \tag{84}$$

Observe that because we estimated $\hat{g}$ on a separate sample S', the fact that we have conditioned on $G = \hat{g}$ does not affect the distribution of the samples in S. Hence, if we let G denote the random variable corresponding to the first stage estimate, then:

$$\Pr\left(\|\nabla_\theta[L_S(\theta_0, \hat{g}) - L_D(\theta_0, \hat{g})]\|_\infty \geqslant \epsilon_{n,\delta} | G = \hat{g}\right) = \Pr\left(\|\nabla_\theta[L_S(\theta_0, \hat{g}) - L_D(\theta_0, \hat{g})]\|_\infty \geqslant \epsilon_{n,\delta}\right)$$

By Assumption 4 the term on the right hand side is at most δ. Hence, it suffices to bound the first term on the right hand side of Equation (84) by $Bg_{n,\delta}^2$ w.p. $1-\delta$.

Consider the vector valued function $f(r) = \nabla_\theta L_D(\theta_0, g_0 + r(\hat{g} - g_0))$. By a first order Taylor expansion of $f^i(1)$ around $f^i(0)$ for each $i \in [d]$ and the definition of the pathwise derivative, we have:

$$\nabla_{\theta_i} L_D(\theta_0, \hat{g}) = \nabla_{\theta_i} L_D(\theta_0, g_0) + \underbrace{D_0[\hat{g} - g_0, \nabla_{\theta_i} L_D(\theta_0, g_0)]}_{\Gamma} + \underbrace{D_{\bar{r}_i}^2[\hat{g} - g_0, \nabla_{\theta_i} L_D(\theta_0, g_0)]}_{\Delta} \tag{85}$$

for some $\bar{r}_i \in [0,1]$. By Assumption 2, we have that $\Gamma = 0$ and by Regularity Assumption 3, we have that $|\Delta| \leqslant B\|\hat{g} - g_0\|^2$. By Assumption 1, we also have $\|\hat{g} - g_0\| \leqslant g_{n,\delta}$. Combining all the above we have:

$$\|\nabla_\theta [L_D(\theta_0, \hat{g}) - L_D(\theta_0, g_0)]\|_\infty \leqslant B g_{n,\delta}^2$$

which completes the proof of the Lemma. □

Proof of Theorem 4. Consider the event $\mathcal{E}_n$ that all probabilistic statements in Assumptions 1, 4, 5, 6 and 7 occur for some confidence parameter δ. If each individual occurs with probability $1-\delta$, then by a union bound their intersection occurs w.p. $1-5\delta$. On this event $\mathcal{E}_n$ the following proof holds. For simplicity of notation we drop the confidence parameter δ from all subscripts as it is fixed.

Since $\hat{\theta} \in \mathcal{T}$ is a local minimum of the empirical plug-in loss function $L_S(\cdot, \hat{g})$, we can write from the first-order condition that for any $\theta \in \mathcal{T}$:

$$\langle \nabla_\theta L_S(\hat{\theta}, \hat{g}) + \lambda\, \mathtt{sign}(\hat{\theta}), \hat{\theta} - \theta \rangle \leqslant 0. \tag{86}$$

By Assumption 7, we have that with probability approach 1, that $\theta_0 \in \mathcal{T}$. Hence, we can apply the latter inequality for $\theta = \theta_0$. Letting $\nu = \hat{\theta} - \theta_0$ we can write:

$$\langle \nabla_\theta L_S(\hat{\theta}, \hat{g}) + \lambda\, \mathtt{sign}(\hat{\theta}), \nu \rangle \leqslant 0. \tag{87}$$

By Cauchy-Schwarz and the definition of norms:

$$\langle \mathtt{sign}(\hat{\theta}), \theta_0 - \hat{\theta} \rangle \leqslant \|\mathtt{sign}(\hat{\theta})\|_\infty \|\theta_0\|_1 - \langle \mathtt{sign}(\hat{\theta}), \hat{\theta} \rangle = \|\theta_0\|_1 - \|\hat{\theta}\|_1.$$

Combining the two inequalities we get:

$$\langle \nabla_\theta L_S(\hat{\theta}, \hat{g}), \nu \rangle \leqslant \lambda \left(\|\theta_0\|_1 - \|\hat{\theta}\|_1 \right). \tag{88}$$

By the $(\gamma, \kappa_n, \tau_n)$-GRC condition on the empirical oracle loss

$$H_S(\hat{\theta}, \theta_0, g_0) \geqslant \gamma \|\nu\|_2^2 - \kappa_n \|\nu\|_1 - \tau_n \|\nu\|_1^2 \tag{89}$$

Combining the inequality above with Assumption 6, we derive a GRC condition for the empirical plugin loss:

$$H_S(\hat{\theta}, \theta_0, \hat{g}) \geqslant\ H_S(\hat{\theta}, \theta_0, g_0) - L(g_n + \xi_n)\|\nu\|_1^2 \geqslant\ \gamma\|\nu\|_2^2 - \kappa_n\|\nu\|_1 - \underbrace{(\tau_n + L(g_n + \xi_n))}_{\zeta_n} \cdot \|\nu\|_1^2 \tag{90}$$

Combining Equation (88) and (90):

$$\begin{aligned}\langle \nabla_\theta L_S(\theta_0, \hat{g}), \nu \rangle = &\ \langle \nabla_\theta L_S(\hat{\theta}, \hat{g}), \nu \rangle - H_S(\hat{\theta}, \theta_0, \hat{g}) \\ \leqslant &\ \lambda \left(\|\theta_0\|_1 - \|\hat{\theta}\|_1 \right) - \gamma \|\nu\|_2^2 + \kappa_n \|\nu\|_1 + \zeta_n \|\nu\|_1^2 \end{aligned} \tag{91}$$

By Cauchy-Schwarz, we also have:

$$\langle \nabla_\theta L_S(\theta_0, \hat{g}), \nu \rangle \geqslant -\|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty \|\nu\|_1 \tag{92}$$

Combining Equations (91) and (92):

$$\gamma \|\nu\|_2^2 \leqslant \lambda \left(\|\theta_0\|_1 - \|\hat{\theta}\|_1 \right) + \kappa_n \|\nu\|_1 + \zeta_n \|\nu\|_1^2 + \|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty \|\nu\|$$

By Lemma 16, $\frac{\lambda}{2} \geqslant \zeta_n \|\nu\|_1 + \kappa_n + \epsilon_n + B g_n^2$ (which is assumed by the Theorem) implies that $\frac{\lambda}{2} \geqslant \zeta_n \|\nu\|_1 + \kappa_n + \|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty$. Hence, we have:

$$\gamma \|\nu\|_2^2 \leqslant \lambda \left(\|\theta_0\|_1 - \|\hat{\theta}\|_1 \right) + \frac{\lambda}{2} \|\nu\|_1 \tag{93}$$

By ν_T we denote p-dimensional vector such that $\nu_{i,T} = \nu_i$ on set of indices $i \in T \subset \{1, \ldots, p\}$ and $\nu_{i,T} = 0$ if $i \notin T$. Also, let T^c be the complement of T Observe that:

$$\|\theta_0\|_1 - \|\hat{\theta}\|_1 \leqslant \|\nu_T\|_1 - \|\nu_{T^c}\|_1 \tag{94}$$

Combining Equations (93) and (94), we get:

$$\gamma \|\nu\|_2^2 \leqslant \frac{3\lambda}{2} \|\nu_T\|_1 - \frac{\lambda}{2} \|\nu_{T^c}\|_1 \leqslant \frac{3\lambda}{2} \|\nu_T\|_1 \leqslant \frac{3\lambda\sqrt{k}}{2} \|\nu_T\|_2 \leqslant \frac{3\lambda\sqrt{k}}{2} \|\nu\|_2 \tag{95}$$

Dividing both sides by $\|\nu\|_2$ yields the rate:

$$\|\nu\|_2 \leqslant \frac{3\sqrt{k}}{2\gamma} \cdot \lambda \tag{96}$$

For the ℓ_1 convergence rate, observe that from Equation (95), we can also derive that:

$$0 \leqslant \gamma \|\nu\|_2^2 \leqslant \frac{3\lambda}{2} \|\nu_T\|_1 - \frac{\lambda}{2} \|\nu_{T^c}\|_1 \tag{97}$$

Thus we get that: $\|\nu_{T^c}\|_1 \leqslant 3\|\nu_T\|_1$, which implies the ℓ_1 convergence theorem, since:

$$\|\nu\|_1 \leqslant 4\|\nu_T\|_1 \leqslant 4\sqrt{k}\|\nu_T\|_2 \leqslant 4\sqrt{k}\|\nu\|_2$$

□

B.2 Proof of Main Theorem for Convex Losses

Throughout the proof we assume that the probabilistic Assumptions 1, 4, 5 and 6 hold deteriministically. By a union bound, the probability of that event is at least $1 - 4\delta$. Hence, the proof below holds with probability $1 - 3\delta$. Moreover, given that we fix the confidence level δ we will drop it from the subscripts throughout the proof.

Before moving into the technical details we begin by providing an outline of the proof. We first show (Lemma 17) that if the empirical oracle loss $L_S(\cdot, g_0)$ satisfies the $(\gamma, \kappa_n, \tau_n)$-GRC condition, then the empirical plug-in loss $L_S(\cdot, \hat{g})$ also satisfies a $(\gamma/2, \kappa_n, 0)$-GRC condition for sufficiently large sample size n, assuming that the first stage estimation is consistent at a reasonable rate. Subsequently (Lemma 18), we show that if the regularization weight is at least $2\|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty$, then the estimation error $\nu = \hat{\theta} - \theta_0$ must lie in a small cone of the high dimensional space, were most of the mass is placed on the coordinates of the true support. Restricted strong convexity of the plug-in empirical loss and the fact that the estimation error lies in the restricted cone, together with a slightly stronger condition on the weight $\lambda \geqslant 2\left(\|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty + \kappa_n\right)$, then imply (Lemma 19) that the estimation error must be of the order $\frac{\lambda\sqrt{k}}{\gamma}$. We then conclude the proof of the main theorem by invoking orthogonality of the loss function to show that the ℓ_∞ norm of the gradient of the empirical plug-in loss evaluated at the true parameter, decays at least as fast as $\epsilon_n + Bg_n^2$ (Lemma 16). Hence, it suffices to set a regularization weight that decays at a rate $\epsilon_n + Bg_n^2 + \kappa_n$, which leads to the final convergence rate claimed in Theorem 6.

Lemma 17 (No First Stage Influence on Restricted Strong Convexity)**.** *Let Assumptions 1, 5, 6 with $\mathcal{B} = \mathcal{C}(T; 3)$ hold deterministically and $k(\tau_n + L\, g_n) = o(1)$. Then, the empirical plug-in loss $L_S(\cdot, \hat{g})$ satisfies the $(\gamma/2, \kappa_n, 0)$-GRC condition with $\mathcal{B} = \mathcal{C}(T; 3)$, for n sufficiently large, such that $16k(\tau_n + L\,(g_n + \xi_n)) \leqslant \frac{\gamma}{2}$.*

Proof. Assumption 6 implies:

$$\sup_{\nu \in \mathcal{C}(T;3)} \frac{|H_S(\theta_0 + \nu, \theta_0, \hat{g}) - H_S(\theta_0 + \nu, \theta_0, g_0)|}{\|\nu\|_1^2} \leqslant Lg_n.$$

Subsequently, invoking the Cauchy-Schwarz inequality and the fact that $\nu = \theta - \theta_0 \in C(T, 3)$:

$$\begin{aligned}\max_{\nu \in \mathcal{C}(T;3)} \frac{|H_S(\theta_0 + \nu, \theta_0, \hat{g}) - H_S(\theta_0 + \nu, \theta_0, g_0)|}{\|\nu\|_2^2} &\leqslant \max_{\nu \in \mathcal{C}(T;3)} \frac{\|\nu\|_1^2 Lg_n}{\|\nu\|_2^2} \leqslant \frac{((1+3)\|\nu_T\|_1)^2 L(g_n + \xi_n)}{\|\nu\|_2^2} \\ &\leqslant 16 \frac{\|\nu_T\|_1^2 L(g_n + \xi_n)}{\|\nu_T\|_2^2} \\ &\leqslant 16kL(g_n + \xi_n) \leqslant 16kL(g_n + \xi_n)\end{aligned}$$

By Assumption 5, the empirical oracle loss $L_S(\cdot, g_0)$ satisfies the $(\gamma, \kappa_n, \tau_n)$-GRC condition: $\forall \nu \in C(T; 3)$

$$\begin{aligned} H_S(\theta_0 + \nu, \theta_0, \hat{g}) &\geqslant H_S(\theta_0 + \nu, \theta_0, g_0) - 16kL(g_n + \xi_n)\|\nu\|_2^2 \\ &\geqslant \gamma\|\nu\|_2^2 - \kappa_n\|\nu\|_1 - \tau_n\|\nu\|_1^2 - 16kL(g_n + \xi_n)\|\nu\|_2^2 \\ &\geqslant \gamma\|\nu\|_2^2 - \kappa_n\|\nu\|_1 - 16k(\tau_n + L\left(g_n + \xi_n\right))\|\nu\|_2^2 \geqslant \frac{\gamma}{2}\|\nu\|_2^2 - \kappa_n\|\nu\|_1 \end{aligned}$$

□

Lemma 18. *If $\frac{\lambda}{2} \geqslant \|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty$ and Assumption 8 holds, then $\nu \in \mathcal{C}(T; 3)$, where $\nu = \hat{\theta} - \theta_0$.*

Proof. Since $\hat{\theta}$ minimizes the penalized loss, we have:

$$L_S(\hat{\theta}, \hat{g}) - L_S(\theta_0, \hat{g}) \leqslant \lambda\left(\|\theta_0\|_1 - \|\hat{\theta}\|_1\right) \leqslant \lambda\left(\|\nu_T\|_1 - \|\nu_{T^c}\|_1\right), \tag{98}$$

where the second inequality follows from the observation that $\|\hat{\theta}\|_1 = \|\theta_0 + \nu_T\|_1 + \|\nu_{T^c}\|_1 \geqslant \|\theta_0\|_1 - \|\nu_T\|_1 + \|\nu_{T^c}\|_1$. By convexity of $L_S(\theta, \hat{g})$, Cauchy-Schwarz inequality and the lower bound assumption on λ:

$$L_S(\hat{\theta}, \hat{g}) - L_S(\theta_0, \hat{g}) \geqslant \nabla_\theta L_S(\theta_0, \hat{g}) \cdot (\hat{\theta} - \theta_0) \geqslant -\|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty\|\nu\|_1 \geqslant -\frac{\lambda}{2}\|\nu\|_1 \tag{99}$$

Combining Equations (98) and 99:

$$\lambda\left(\|\nu_T\|_1 - \|\nu_{T^c}\|_1\right) \geqslant L_S(\hat{\theta}, \hat{g}) - L_S(\theta_0, \hat{g}) \geqslant -\frac{\lambda}{2}\|\nu\|_1$$

Dividing by λ and re-arranging we get $3\|\nu_T\|_1 \geqslant \|\nu_{T^c}\|_1$. □

Lemma 19 (Oracle Inequality for θ)**.** *If Assumptions 1, 2, 3, 4, 5, 6, 8 with $\mathcal{B} = \mathcal{C}(T; 3)$ hold deterministically and $\frac{\lambda}{2} \geqslant \epsilon_n + B\, g_n^2 + \kappa_n$ then the second stage estimate $\hat{\theta}$, for n sufficiently large, such that $16k(\tau_n + L\left(g_n + \xi_n\right)) \leqslant \frac{\gamma}{2}$, satisfies:*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{3\sqrt{k}}{\gamma}\lambda \qquad\qquad \|\hat{\theta} - \theta_0\|_1 \leqslant \frac{12k}{\gamma}\lambda \tag{100}$$

Proof. Let $\nu = \hat{\theta} - \theta_0$. Similar to Equation (98), since $\hat{\theta}$ minimizes the penalized objective:

$$L_S(\hat{\theta}, \hat{g}) - L_S(\theta_0, \hat{g}) \leqslant \lambda\left(\|\theta_0\|_1 - \|\hat{\theta}\|_1\right) \leqslant \lambda\left(\|\nu_T\|_1 - \|\nu_{T^c}\|_1\right), \tag{101}$$

By Lemma 16, 17 and 18 we have:

$$
\begin{aligned}
\lambda\left(\|\theta_0\|_1 - \|\hat{\theta}\|_1\right) &\geqslant \nabla_\theta L_S(\hat{\theta}, \hat{g}) \cdot (\hat{\theta} - \theta_0) && \text{(Optimality of } \hat{\theta}\text{)}\\
&\geqslant \nabla_\theta L_S(\theta_0, \hat{g}) \cdot (\hat{\theta} - \theta_0) + H_S(\hat{\theta}, \theta_0, \hat{g}) && \text{(symmetric Bregman distance)}\\
&\geqslant \nabla_\theta L_S(\theta_0, \hat{g}) \cdot (\hat{\theta} - \theta_0) + \frac{\gamma}{2}\|\nu\|_2^2 - \kappa_n \|\nu\|_1 && \text{(Lemmas 17, 18)}\\
&\geqslant -\|\nabla_\theta L_S(\theta_0, \hat{g})\|_\infty \cdot \|\nu\|_1 + \frac{\gamma}{2}\|\nu\|_2^2 - \kappa_n \|\nu\|_1 && \text{(Cauchy-Schwarz)}\\
&\geqslant -(\epsilon_n + B\, g_n^2) \cdot \|\nu\|_1 + \frac{\gamma}{2}\|\nu\|_2^2 - \kappa_n \|\nu\|_1 && \text{(Lemma 16)}\\
&\geqslant -\frac{\lambda}{2} \cdot \|\nu\|_1 + \frac{\gamma}{2}\|\nu\|_2^2 && \text{(Assumption on } \lambda\text{)}
\end{aligned}
$$

Combining Equation (101) with the latter inequality we get:

$$
\frac{\gamma}{2}\|\nu\|_2^2 \leqslant \frac{3\lambda}{2}\|\nu_T\|_1 - \frac{\lambda}{2}\|\nu_{T^c}\|_1 \leqslant \frac{3\lambda}{2}\|\nu_T\|_1 \leqslant \frac{3\lambda\sqrt{k}}{2}\|\nu\|_2
$$

Dividing over by $\|\nu\|_2$, yields the theorem. The ℓ_1 rate is derived by the fact that in the cone $\mathcal{C}(T;3)$: $\|\nu\|_1 \leqslant 4\sqrt{k}\|\nu\|_2$. □

B.3 Proof of Corollary 7

Part I. We first prove the statement for the case of the $\ell_{\infty,1}$ norm in the nuisance space. We verify that each of the Assumptions 1, 2, 3, 4, 5, 6 holds for the specified parameter instantiation. The Corollary then follows.

Assumption 1. Follows directly from Condition 2 of the corollary.

Assumption 2. Follows from Condition 4 of the corollary since:

$$
D_0[g - g_0, \nabla_\theta \mathbb{E}[\ell(w, \theta, g_0)]] = \mathbb{E}[\nabla_{\gamma\theta}\ell(w, \theta_0, g_0(w))\,(g(w) - g_0(w))] = 0
$$

Assumption 3. Let $\bar{g}_r = r(g - g_0) + g$. By Cauchy-Schwarz inequality, and Condition 3 of the corollary, the i-th coordinate of the second-order pathwise derivative is bounded as:

$$
\begin{aligned}
\left(D_r^2[g - g_0, \nabla_\theta L_D(\theta_0, g_0)]\right)_i &= \mathbb{E}[(g(w) - g_0(w))' \nabla_{\gamma\gamma\theta_i}\ell(w, \theta_0, \bar{g}_r(w))\,(g(w) - g_0(w))]\\
&\leqslant \mathbb{E}[\|\nabla_{\gamma\gamma\theta_i}\ell(w, \theta_0, \bar{g}_r(w))\|_\infty\, \|g(w) - g_0(w)\|_1^2]\\
&\leqslant H\,\mathbb{E}[\|g(w) - g_0(w)\|_1^2]\\
&\leqslant H \sup_{w \in \mathcal{W}} \|g(w) - g_0(w)\|_1^2\\
&\leqslant H\,\|g - g_0\|_{\infty,1}^2
\end{aligned}
$$

Assumption 4. Since by Condition 3, $\|\nabla_\theta \ell(w, \theta_0, g)\|_\infty \leqslant H$, we have that each coordinate of the empirical gradient is a sum of n independent random variables with mean equal to the population gradient and bounded a.s. in absolute value by H. Hence, by McDiarmid's inequality, each term is within $\epsilon_{n,\delta} = H\sqrt{\frac{\log(2p/\delta)}{2n}}$ of each mean with probability at least $1 - \frac{\delta}{p}$. By a union bound over the p coordinates we get that all coordinates are within $\epsilon_{n,\delta}$ from their mean with probability at least $1 - \delta$.

Assumption 5. By Condition 5, the population loss L_D satisfies the $(\gamma_D, 0, 0)$-GRC condition. Moreover, by Condition 6, the Hessian of the empirical oracle loss concentrates uniformly over $\theta \in \Theta$ to its population counterpart at rate $\hat{\tau}_{n,\delta}$. Hence, we can apply Lemma 2, to show that L_S has the $(\gamma, 0, \tau_{n,\delta})$-GRC property.

Assumption 6. From Condition 3, we have:

$$\begin{aligned}
|\nabla_{\theta_i\theta_j} L_S(\theta, g) - \nabla_{\theta_i\theta_j} L_S(\theta, g')| =&\ |\mathbb{E}_S[\nabla_{\theta_i\theta_j}\ell(w, \theta, g) - \nabla_{\theta_i\theta_j}\ell(w, \theta, g')]| \\
\leqslant&\ \mathbb{E}_S[|\nabla_{\theta_i\theta_j}\ell(w, \theta, g) - \nabla_{\theta_i\theta_j}\ell(w, \theta, g')|] \\
\leqslant&\ \sup_{w \in \mathcal{W}} |\nabla_{\theta_i\theta_j}\ell(w, \theta, g) - \nabla_{\theta_i\theta_j}\ell(w, \theta, g')| \\
=&\ \sup_{w \in \mathcal{W}} |\nabla_{\gamma\theta_i\theta_j}\ell(w, \theta, \bar{g})'(g(w) - g'(w))| \\
\leqslant&\ \sup_{w \in \mathcal{W}} H\|g(w) - g'(w)\|_1 \\
=&\ H\,\|g - g'\|_{\infty,1}
\end{aligned}$$

Part II. We now verify the conditions for the case of the $\ell_{2,1}$ norm in the nuisance space. The proofs of all assumptions except Assumptions 3 and 6 remains unchanged. For Assumption 3 we observe that the quantity in the third step of the proof is exactly equal to $H\,\|g - g_0\|_{2,1}^2$, which proves the claim. Thus it remains to verify Assumption 6.

From Part I we have shown that:

$$\begin{aligned}
|\nabla_{\theta_i\theta_j} L_S(\theta, g) - \nabla_{\theta_i\theta_j} L_S(\theta, g')| \leqslant&\ \mathbb{E}_S[|\nabla_{\theta_i\theta_j}\ell(w, \theta, g) - \nabla_{\theta_i\theta_j}\ell(w, \theta, g')|] \\
=&\ \mathbb{E}_S[|\nabla_{\gamma\theta_i\theta_j}\ell(w, \theta, \bar{g})'(g(w) - g'(w))] \\
\leqslant&\ H\,\mathbb{E}_S[\|g(w) - g'(w)\|_1]
\end{aligned}$$

Now observe that since $\|g(w) - g'(w)\|_\infty \leqslant H$, we have that $\|g(w) - g'(w)\|_1 \leqslant d\,H$. By McDiarmid's inequality we then have that for any fixed $g, g' \in \mathcal{G}_n$, with probability $1 - \delta$:

$$\left|\mathbb{E}_S[\|g(w) - g'(w)\|_1] - \mathbb{E}[\|g(w) - g'(w)\|_1]\right| \leqslant d\,H\sqrt{\frac{\log(2/\delta)}{2n}} := \xi_{n,\delta}$$

Combining with the previous inequality we have:

$$|\nabla_{\theta_i\theta_j} L_S(\theta, g) - \nabla_{\theta_i\theta_j} L_S(\theta, g')| \leqslant\ H\left(\mathbb{E}[\|g(w) - g'(w)\|_1] + \xi_{n,\delta}\right)$$

By Jensens' inequality:

$$\mathbb{E}[\|g(w)-g'(w)\|_1] \leqslant \sqrt{\mathbb{E}[\|g(w)-g'(w)\|_1^2]} = \|g-g'\|_{2,1}$$

Combining completes the proof.

C Omitted Proofs from Section 3

C.1 Proof of Lemma 8

Proof of Lemma 8. First we verify that the gradient of the loss ℓ is equal to the moment ℓ, by a simple application of the chain rule:

$$\begin{aligned}\nabla_\theta \ell(w,\theta,g(z)) &= \nabla_\theta K(w,\Lambda(z,g(z))'\theta,g(z)) = \frac{\partial}{\partial t}K(w,\Lambda(z,g(z))'\theta,g(z))\,\Lambda(z,g(z))\\ &= \Phi(w,\Lambda(z,g(z))'\theta,g(z))\,\Lambda(z,g(z)) = \rho(w,\theta,g(z))\end{aligned}$$

Next, we verify orthogonality of the loss:

$$\begin{aligned}D_0[g-g_0,\nabla_\theta\mathbb{E}[\ell(w,\theta_0,g_0(z))]] &= D_0[g-g_0,\mathbb{E}[\rho(w,\theta_0,g_0(z))]]\\ &= \mathbb{E}\left[\nabla_\gamma\rho(w,\theta_0,g_0(z))\,(g(z)-g_0(z))\right]\\ &= \mathbb{E}\left[\Lambda(z,g_0(z))\,\nabla_\gamma\phi(w,\theta_0,g_0(z))'\,(g(z)-g_0(z))\right]\\ &\quad+\mathbb{E}\left[\phi(w,\theta_0,g_0(z))\,\nabla_\gamma\Lambda(z,g_0(z))\,(g(z)-g_0(z))\right]\end{aligned}$$

Both terms on the right-hand-side are equal to zero, by applying the tower property and noting that $\mathbb{E}[\nabla_\gamma\phi(w,\theta_0,g_0(z))|z]=0$ by conditional orthogonality and $\mathbb{E}[\phi(w,\theta_0,g_0(z)|z]=0$ by conditional moment constraint.

For the final part of the Lemma, observe that the Hessian of the loss ℓ, equivalently the Jacobian of the moment ρ, takes the form:

$$\begin{aligned}\mathbb{E}[\nabla_{\theta\theta}\ell(w,\theta,g_0(z))\,|\,z] &= \mathbb{E}\left[\nabla_\theta\rho(w,\theta,g_0(z))\,|\,z\right]\\ &= \underbrace{\mathbb{E}\left[\frac{\partial}{\partial t}\Phi(w,\Lambda(z,g_0(z))'\theta,g_0(z))\,|\,z\right]}_{\geqslant\nu}\underbrace{\Lambda(z,g_0(z))\,\Lambda(z,g_0(z))'}_{\Sigma(z)\succeq 0}\end{aligned}$$

Since $\Sigma(w)$ is non-negative definite and since the scalar multiplier of the matrix is lower bounded by ν from the increasing rate with respect to the index assumption, we have that:

$$\mathbb{E}\left[\nabla_{\theta\theta}\ell(w,\theta,g_0(z))\,|\,z\right]\succeq\nu\Sigma(z)$$

Taking expectation on both sides we then have:

$$\nabla_{\theta\theta}L_D(\theta,g_0)=\mathbb{E}\left[\nabla_{\theta\theta}\ell(w,\theta,g_0)\right]\succeq\nu\,\mathbb{E}\left[\Sigma(z)\right]=\nu\,\Sigma\succeq\nu\,\gamma_\Sigma\,I$$

Thus the population loss is strongly convex and satisfies the $(\nu\,\gamma_\Sigma, 0, 0)$-GRC. Thus it has a unique local minimum with respect to θ and hence a unique solution to the first order condition $\nabla_\theta L_D(\theta, g_0) = 0$. Therefore, the moment equation has a unique solution, since $\nabla_\theta L_D(\theta, g_0) = \mathbb{E}[\rho(w, \theta, g_0(z))]$. Therefore ρ is also an identifying moment. □

C.2 Proof of Corollary 9

Proof of Corollary 9. We verify each of the conditions of Corollary 7. The corollary then follows.

Condition 1. Follows directly from Condition 1 of the corollary.

Condition 2. Follows directly from Condition 2 of the corollary.

Condition 3. We verify the boundedness of each of the derivative terms in Corollary 7, using the upper bounds in Condition 3 of the corollary.

$$\|\nabla_\theta \ell(w, \theta_0, g(z)))\|_\infty = \|\phi(w, \theta_0, g(z))\,\Lambda(z, g(z))\|_\infty \leqslant |\phi(w, \theta_0, g(z))|\,\|\Lambda(z, g(z))\|_\infty \leqslant U^2$$

For any $k \in \{1, \dots, p\}$:

$$\begin{aligned}
\|\nabla_{\gamma\gamma\theta_k} \ell(w, \theta_0, g(z)))\|_\infty =&\ \|\nabla_{\gamma\gamma}\rho_k(w, \theta_0, g(z))\|_\infty \\
=&\ \|\Lambda_k(z, g(z))\nabla_{\gamma\gamma}\phi(w, \theta_0, g(z)) + \phi(w, \theta_0, g(z))\nabla_{\gamma\gamma}\Lambda_k(z, g(z))\|_\infty \\
&+ \|\nabla_\gamma\phi(w, \theta_0, g(z))\nabla_\gamma\Lambda_k(z, g(z))' + \nabla_\gamma\Lambda_k(z, g(z))\nabla_\gamma\phi(w, \theta_0, g(z))'\|_\infty \\
\leqslant&\ 4U^2
\end{aligned}$$

For any $j, k \in \{1, \dots, p\}$:

$$\begin{aligned}
\|\nabla_{\gamma\theta_j\theta_k} \ell(w, \theta, g(z)))\|_\infty =&\ \|\nabla_{tt}\Phi(w, \Lambda(z, g(z))'\theta, g(z))\Lambda_j(z, g(z))\Lambda_k(z, g(z))\nabla_\gamma\Lambda(z, g(z))'\theta\|_\infty \\
&+ \|\nabla_t\Phi(w, \Lambda(z, g(z))'\theta, g(z))\nabla_\gamma\Lambda_j(z, g(z))\Lambda_k(z, g(z))\|_\infty \\
&+ \|\nabla_t\Phi(w, \Lambda(z, g(z))'\theta, g(z))\Lambda_j(z, g(z))\nabla_\gamma\Lambda_k(z, g(z)))\|_\infty \\
\leqslant&\ U^4 + 2U^3
\end{aligned}$$

Condition 4. Orthogonality of the loss defined in Equation (37) was established in Lemma 8.

Condition 5. As was established in the proof of Lemma 8, since by Condition 4 of the corollary, $\mathbb{E}[\frac{\partial}{\partial t}\Phi(w, \Lambda(z, g_0(z))'\theta, g_0(z))\,|\,z] \geqslant \nu$ and $\Sigma \succeq \gamma_\Sigma I$, we conclude that $\mathbb{E}[\nabla_{\theta\theta}\ell(w, \theta, g_0(z))] \succeq \nu\,\gamma_\Sigma\,I$.

Condition 6. We argue about the uniform convergence of each entry of the Hessian. The result would then follow via a union bound over the p^2 entries of the Hessian. For $i, j \in \{1, \dots, p\}$, the (i, j) entry takes the form:

$$\nabla_{\theta_i\theta_j}\ell(w, \theta, g_0(z)) = \nabla_t\Phi(w, \Lambda(z, g_0(z))'\theta, g_0(z))\Lambda_i(z, g_0(z))\Lambda_j(z, g_0(z))$$

Since by Condition 3 of the corollary, we have that $|\nabla_{tt}\Phi(w, \Lambda(z, g_0(z))'\theta, g_0(z))| \leqslant U$ and $\|\Lambda(z, g_0(z))\|_\infty \leqslant U$, we have that the latter function depends on the parameter θ only through a U^3-Lipschitz function of a linear index of θ. By invoking Lemma 26.9 of [21], we know that the Rademacher complexity of any such class is upper bounded by U^3 times the Rademacher complexity of the function class $\mathcal{F} = \{\Lambda(\cdot, g_0(\cdot))'\theta : \theta \in \Theta\}$. Since $\|\theta\|_1 \leqslant kU$ for any $\theta \in \Theta$ and $\|\Lambda(z, g_0(z))\|_\infty \leqslant U$, we can invoke Lemma 26.11 of [21] to bound the Rademacher complexity of $\mathcal{F}$ by $U^2 k\sqrt{\frac{2\log(p)}{n}}$. Thus the overall rademacher complexity of the function class defined by the entry of the Hessian is at most $\mathcal{R} = U^5 k\sqrt{\frac{2\log(p)}{n}}$. Subsequently, since these functions are also absolutely bounded by U^3, via standard results in statistical learning theory (see e.g. Lemma 26.2 and 26.5 of [21], we have that with probability at least $1 - \delta/p^2$:

$$\sup_{\theta\in\Theta} |\mathbb{E}_S[\nabla_{\theta_i\theta_j}\ell(w, \theta, g_0(z))] - \mathbb{E}[\nabla_{\theta_i\theta_j}\ell(w, \theta, g_0(z))]| \leqslant 4\mathcal{R} + U^3\sqrt{\frac{4\log(2\,p/\delta)}{n}}$$

Taking a union bound over the p^2 entries yields the condition for the stated $\hat{\tau}_{n,\delta}$.

Convexity. Finally, we note that the loss ℓ is convex, since:

$$\nabla_{\theta\theta}\ell(w, \theta, g(z)) = \nabla_t\Phi(w, \Lambda(z, g(z))'\theta, g(z))\,\Lambda(z, g(z))\Lambda(z, g(z))' \succeq 0.$$

Moreover, from Condition 2, we have that $k\,d\,g_{n,\delta} = o(1)$ and from our assumption on the rate of growth of k and the proven bound on $\hat{\tau}_{n,\delta}$, we also have that:

$$k\tau_{n,\delta} = O\left(k^2\sqrt{\frac{\log(p/\delta)}{n}}\right) = o\left(1\right)$$

Thus we conclude that $k(\tau_{n,\delta} + d\,g_{n,\delta}) = o(1)$. □

C.3 Proof of Lemma 10

Proof of Lemma 10. We first verify conditional orthogonality of ϕ with respect to the original nuisance function g: let ∇_γ be the gradient with respect to the output of g

$$\mathbb{E}\left[\nabla_\gamma\phi(w, \theta_0, \{g_0(z), h_0(z)\})|z\right] = \mathbb{E}[\nabla_\gamma m(w, \theta_0, g_0(z))|z] - h_0(z) = 0$$

by the definition of h_0. Finally, we verify orthogonality with respect to h_0: let ∇_χ be the gradient with respect to the output of h

$$\mathbb{E}\left[\nabla_\chi\phi(w, \theta_0, \{g_0(z), h_0(z)\})|z\right] = \mathbb{E}\left[v - g_0(z)|z\right] = 0$$

where we invoked the auxiliary conditional moment restriction. □

C.4 Proof of Corollary 11

Proof of Corollary 11. The corollary follows by simply noting that under the conditions of the corollary for model $\mathcal{I} = \{\mathcal{W}, \mathcal{Z}, \Theta, \mathcal{G}, m, M, \Lambda\}$ imply that all conditions of Corollary 9 are satisfied for the orthogonal version of the model $\mathcal{I}' = \{\mathcal{W}, \mathcal{Z}, \Theta, \tilde{\mathcal{G}}, \phi, \Phi, \Lambda\}$, defined by Equation (43) and the augmented nuisance space. In particular, U-smoothness of $\mathcal{I}$ imply $2U$-smoothness of $\mathcal{I}'$. Hence, we can directly apply the conclusion of Corollary 9 for model $\mathcal{I}'$ to get the result. □

D Omitted Proofs from Section 4

D.1 Omitted Proofs from Section 4.3

Lemma 20. *The vector-valued moment ρ defined in Equation* (69) *is valid and orthogonal with respect to $g = \{h, q, V\}$.*

Proof. We first observe that validity of ρ follows from the conditional validity of Φ:

$$\mathbb{E}[\Phi(w, \Lambda(z, g_0(z))'\theta, g_0(z)) \mid z] = \frac{1}{V_0(z)}\mathbb{E}[G(x'\theta_0 + f_0(u)) - y \mid z] = 0$$

For notational simplicity, let $\rho(w, \theta, g(z)) := \rho(w, \Lambda(z, g(z))'\theta, g(z))$. We now verify conditional orthogonality for each nuisance component conditional on the corresponding input of that component, were we invoke the original moment condition in Equation (64) as well as the definition of $h_0(u) = \mathbb{E}[x|u]$:

$$\begin{aligned}
\mathbb{E}[\nabla_h \rho(w, \theta_0, g_0(z)) \mid u] =& -\mathbb{E}\left[\frac{G'(x'\theta_0 + f_0(u))}{V_0(z)}(x - h_0(u))\theta_0' + \frac{G(x'\theta_0 + f_0(u)) - y}{V_0(z)} I_p \,\middle|\, u\right] \\
=& -\mathbb{E}\left[(x - h_0(u))\theta_0' + \frac{G(x'\theta_0 + f_0(u)) - y}{V_0(z)} I_p \,\middle|\, u\right] \\
=& -\mathbb{E}\left[x - h_0(u) \mid u\right]\theta_0' - \mathbb{E}\left[\frac{\mathbb{E}\left[G(x'\theta_0 + f_0(u)) - y|x, u\right]}{V_0(z)} I_p \,\middle|\, u\right] = 0 \\
\mathbb{E}[\nabla_q \rho(w, \theta_0, g_0(z)) \mid u] =& \mathbb{E}\left[\frac{G'(x'\theta_0 + f_0(u))}{V_0(z)}(x - h_0(u)) \,\middle|\, u\right] = \mathbb{E}\left[x - h_0(u) \,\middle|\, u\right] = 0 \\
\mathbb{E}[\nabla_V \rho(w, \theta_0, g_0(z)) \mid z] =& -\mathbb{E}\left[\frac{G(x'\theta_0 + f_0(u)) - y}{V_0(z)^2} \,\middle|\, z\right] = 0
\end{aligned}$$

where I_p denotes the p-dimensional identity matrix. □

Lemma 21. *The vector-valued moment $\tilde{\rho}$ defined in Equation* (75) *is valid and orthogonal with respect to $g = \{h, q, \tilde{V}\}$.*

Proof. We first observe that validity of $\tilde{\rho}$ follows from the validity of conditional expectation $\mathbb{E}[G(x'\theta_0 + f_0(u)) - y \mid x, u] = 0$:

$$\mathbb{E}[\tilde{\rho}(w, \Lambda(z, g_0(z))'\theta, g_0(z)) \mid z] = \frac{G'(x'\theta_0 + f_0(u))}{\tilde{V}_0(z)}\mathbb{E}[G(x'\theta + f_0(u)) - y \mid z] = 0$$

As before, for notational simplicity, let $\tilde{\rho}(w,\theta,g(z)) := \tilde{\rho}(w,\Lambda(z,g(z))'\theta,g(z))$. We now verify conditional orthogonality for each nuisance component conditional on the corresponding input of that component, were we invoke the original moment condition in Equation (64) as well as the definition of $h_0(u) = \mathbb{E}[x|u]$:

$$\begin{aligned}
\mathbb{E}[\nabla_h \tilde{\rho}(w,\theta_0,g_0(z))\,|\,u] =& -\mathbb{E}\Bigg[\frac{\left(G'(x'\theta_0+f_0(u))\right)^2}{\tilde{V}_0(z)}(x-h_0(u))\theta_0' \\
&+\frac{G(x'\theta_0+f_0(u))-y}{\tilde{V}_0(z)}G''(x'\theta_0+f_0(u))[x-h_0(u)] \\
&+\frac{G(x'\theta_0+f_0(u))-y}{\tilde{V}_0(z)}G'(x'\theta_0+f_0(u))I_p\,\Bigg|\,u\Bigg] \\
=& -\mathbb{E}\left[(x-h_0(u))\theta_0' + \frac{G(x'\theta_0+f_0(u))-y}{\tilde{V}_0(z)}I_p\,\middle|\,u\right] \\
=& -\mathbb{E}\Bigg[(x-h_0(u))\theta_0' \\
&+\frac{\mathbb{E}[G(x'\theta_0+f_0(u))-y\,|\,x,u]}{\tilde{V}_0(z)}G''(x'\theta_0+f_0(u))[x-h_0(u)] \\
&+\frac{\mathbb{E}[G(x'\theta_0+f_0(u))-y\,|\,x,u]}{\tilde{V}_0(z)}G'(x'\theta_0+f_0(u))I_p\,\Bigg|\,u\Bigg] = 0 \\
\mathbb{E}[\nabla_q \rho(w,\theta_0,g_0(z))\,|\,u] =& \mathbb{E}\Bigg[\frac{\left(G'(x'\theta_0+f_0(u))\right)^2}{\tilde{V}_0(z)}(x-h_0(u)) \\
&+\frac{G(x'\theta_0+f_0(u))-y}{\tilde{V}_0(z)}G''(x'\theta_0+f_0(u))(x-h(u))\,\Bigg|\,u\Bigg] \\
=& \mathbb{E}\Bigg[x-h_0(u) \\
&+\frac{\mathbb{E}[G(x'\theta_0+f_0(u))-y\,|\,x,u]}{\tilde{V}_0(z)}G''(x'\theta_0+f_0(u))(x-h(u))\,\Bigg|\,u\Bigg] = 0 \\
\mathbb{E}[\nabla_V \rho(w,\theta_0,g_0(z))\,|\,z] =& -\mathbb{E}\left[\frac{G(x'\theta_0+f_0(u))-y}{\tilde{V}_0(z)^2}G'(x'\theta_0+f_0(u))[x-h(u)]\,\middle|\,u\right] = 0
\end{aligned}$$

where I_p denotes the p-dimensional identity matrix. □