# Introduction to Debiased Machine Learning, Lecture 1: Double Lasso

Whitney K. Newey

2026 Nemmers Lectures
Northwestern University

# INTRODUCTION

Many economic and causal parameters depend on first steps that are high dimensional and/or nonparametric.

Example 1: A regression slope when there are many potential covariates to control for possible omitted variables bias.

Example 2: Average treatment effect (ATE) in causal analysis when many covariates should be conditioned on in order to have treatment behave as randomly assigned.

There are other important examples including policy effects, economic demand analysis, estimation of structural models where conditional choice probabilities are unknown functions of state variables, and many more.

Machine learning (ML) provides useful ways to allow high dimensional first steps, like covariate terms in regression with many covariates.

ML methods include Lasso, neural nets, random forests, and others.

ML provides excellent predictions for high dimensional regressions but care must be exercised when using ML to estimate economic and causal parameters of interest.

ML achieves excellent predictions by using various forms of regularization and/or model selection to control variance.

Regularization and model selection introduce bias that imparts large biases to estimators of parameters of interest.

Debiased machine learning (DML) reduces this bias to acceptable levels using Neyman orthogonal estimating equations, where ML has zero first-order effect on the parameter of interest.

Cross-fitting, a form of sample splitting, is also useful in reducing "overfitting" bias that ML can cause for parameters of interest.

These lectures introduce Neyman orthogonal estimating equations and how cross-fitting can the used to obtain DML estimators of parameters of interest.

Empirical and Monte Carlo examples are given.

# WARM UP: LINEAR REGRESSION WITH MANY CONTROLS

Consider the regression model,

$$Y = \theta_0 D + Z'\gamma_0 + \epsilon,\ E[D\epsilon] = 0,\ E[Z\epsilon] = 0, \tag{1}$$

where $D$ is the regressor interest, $Z$ consists of $p$ controls, and controls are high dimensional (not small compared to the available sample size $n$).

One way to make sure that ML has no first-order effect on estimation of $\theta_0$ is to "profile out" $Z$ from both $Y$ and $D$; let

$$\tilde{D} = D - Z'\gamma_{DZ}, \quad \gamma_{DZ} = \arg\min_{\gamma \in \mathbb{R}^p} E[(D - Z'\gamma)^2],$$

$$\tilde{Y} = Y - Z'\gamma_{YZ}, \quad \gamma_{YZ} = \arg\min_{\gamma \in \mathbb{R}^p} E[(Y - Z'\gamma)^2],$$

so $\tilde{D}$ and $\tilde{Y}$ are residuals from population least squares regression on $Z$.

Then

$$\tilde{Y} = \theta_0 \tilde{D} + \epsilon, \quad E[\tilde{D}\epsilon] = 0. \tag{2}$$

Then $\theta_0$ can be recovered from population linear regression of $\tilde{Y}$ on $\tilde{D}$:

$$\theta_0 = \arg\min_{a \in \mathbb{R}} E[(\tilde{Y} - \theta \tilde{D})^2] = (E[\tilde{D}^2])^{-1} E[\tilde{D}\tilde{Y}].$$

Note also that $\theta = \theta_0$ solves the moment equation:

$$E[(\tilde{Y} - \theta \tilde{D})\tilde{D}] = 0].$$

This is Frisch-Waugh-Lovell profiling-out.

Because of profiling out the last estimating equation will be Neyman orthogonal, where estimation of the $\gamma's$ has no first order effect on the estimating equation.

We now consider estimation of $\theta$ in high-dimensional setting; for estimation purposes we have a random sample $(Y_i, X_i)_{i=1}^n$, $X_i = (D_i, Z_i)$.

We will mimic in the sample the profiling-out procedure in the population.

If the number of regressors $p$ was small relative to $n$, we could employ ordinary least squares as the prediction method in the profiling-out steps.

To allow for $p$ not small, we can employ instead Lasso-based methods in the profiling-out steps.

The estimation procedure can be summarized below:

# LASSO

Here the conditional mean is assumed to be linear in a known sequence of functions $(b_1(X), b_2(X), ...)$,

$$E[Y|X] = \gamma_0(X) = \sum_{j=1}^{\infty} \beta_{j0} b_j(X).$$

Have data $W_i = (Y_i, X_i)$, $(i = 1, ..., n)$; assume throughout that data are independently and identically distributed (i.i.d.).

For Lasso it is important to normalize so sample mean and variance of each $b_j(X)$ is 0 and 1 respectively, i.e. replace $b_j(X)$ with

$$\hat{b}_j(X) = \left(b(X) - \bar{b}_j\right)/\hat{s}_j,\ \bar{b}_j = \frac{1}{n}\sum_{i=1}^{n} b_j(X_i),\ \hat{s}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[b_j(X_i) - \bar{b}_j]^2}.$$

Normalizing $b_j(X)$ to have sample variance 1 is essential for good performance of Lasso in high dimensional settings.

Let $b(x) = (b_1(x), ..., b_p(x))'$ be "dictionary;" first $p$ functions; Lasso learner is

$$\hat{\gamma}(x) = \bar{Y} + \hat{\beta}' b(x), \quad \hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - \bar{Y} - b(X_i)'\beta]^2 + 2\lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\}.$$

Second term is absolute value penalty; objective function is convex in $\beta$.

Fast computation available in MATLAB and R; cross-validation to choose $\lambda$ is common in practice.

The absolute value term provides "regularization" that helps control the variance of the estimator when $p$ is much bigger than sample size.

Remarkably, convergence rate is only larger by $\ln(p)$ term than corresponding series estimator where we run OLS with best $p$ functions.

**Double Lasso:**

1. We run the Lasso regressions of $Y_i$ on $Z_i$ and $D_i$ on $Z_i$:

$$\hat{\gamma}_{YZ} = \text{arg} \min_{\gamma \in \mathbb{R}^p} \quad \sum_i (Y_i - Z_i'\gamma)^2 + \lambda_1 \sum_j |\gamma_j|,$$

$$\hat{\gamma}_{DZ} = \text{arg} \min_{\gamma \in \mathbb{R}^p} \quad \sum_i (D_i - Z_i'\gamma)^2 + \lambda_2 \sum_j |\gamma_j|,$$

   and obtain the resulting residuals, say using cross-validated choice of $\lambda's$;

$$\check{Y}_i = Y_i - Z_i'\hat{\gamma}_{YZ},$$

$$\check{D}_i = D_i - Z_i'\hat{\gamma}_{DZ}.$$

2. We run least squares of $\check{Y}_i$ on $\check{D}_i$ to obtain the estimator $\hat{\theta}$ as the root of:

$$\frac{1}{n}\sum_{i=1}^{n} (\check{Y}_i - \theta \check{D}_i)\check{D}_i = 0$$

   and use it for standard inference, proceeding as if the residuals were known.

It can be shown that, under regularity conditions, $\sqrt{n}(\hat{\theta}-\theta_0)$ converges in distribution $N(0,V)$ for $V>0$ where $V$ is asymptotic, heteroskedasticity robust variance of OLS.

Thus $\hat{\theta}$ concentrates in a $2\sqrt{V/n}$- neighborhood of $\theta$, with deviations controlled by the normal law.

Observe that the approximate behavior of the double lasso estimator is the same as the approximate behavior of the least squares estimator in low-dimensional models.

This result can be used for construction of confidence intervals.

Just like the low-dimensional case, we can use these results to construct a confidence interval for $\theta$.

Obtain the standard error of $\hat{\theta}$ as

$$\sqrt{\hat{V}/n},\ \hat{V}=n(\sum_{i=1}^{n}\check{D}_i^2)^{-1}\sum_{i=1}^{n}(\check{Y}_i-\hat{\theta}\check{D}_i)^2\check{D}_i^2(\sum_{i=1}^{n}\check{D}_i^2)^{-1};$$

this $\hat{V}$ is the Eicker-Huber-White estimator of $V$; the interval

$$[\hat{\theta}\pm 1.96\sqrt{\hat{V}/n}]$$

covers $\theta_0$ about 95% of the time.

# TESTING THE CONVERGENCE HYPOTHESIS

An empirical example of profiling-out with Lasso is to test the convergence hypothesis of growth theory.

Consider again the coefficient $\theta_0$ in the high-dimensional linear regression model:

$$Y = \theta_0 D + Z'\gamma_0 + \epsilon.$$

Specifically we are interested in how economic growth rates $(Y)$ are related to the initial wealth levels in each country $(D)$ controlling for country's institutional, educational, and other similar characteristics $(Z)$.

The relationship is captured by $\theta_0$, the "speed of convergence/divergence", which predicts the speed at which poor countries catch up ($\theta_0 < 0$) or fall behind ($\theta_0 > 0$) rich countries, after controlling for $Z$.

$$Y = \theta_0 D + Z'\gamma_0 + \epsilon.$$

Our inference question here is: do poor countries grow faster than rich countries, controlling for educational and other characteristics?

In other words, is the speed of convergence $\theta_0$ negative (this is the convergence hypothesis predicted by the several growth models, including the classical Solow model).

Under some strong assumptions that we will not state here, the regression exercise we are doing here can be given a causal interpretation.

The outcome ($Y$) is the realized annual growth rate of a country's wealth (Gross Domestic Product per capita).

The target regressor ($D$) is the initial level of the country's wealth.

The target parameter $\theta_0$ is the speed of convergence, which measures the speed at which poor countries catch up with rich countries.

The controls ($Z$) include measures of education levels, quality of institutions, trade openness, and political stability in the country.

The sample contains 90 countries and about 60 controls.

Thus $p \approx 60$, $n = 90$, and $p$ is not small relative to sample size.

We expect the least squares method to provide a poor/ noisy estimate of $\theta_0$; we expect the method based on profiling-out with Lasso to provide a high quality estimate of $\theta_0$.

| | Estimate | Std. Error | 95% CI |
|---|---|---|---|
| OLS | -.009 | .030 | [-.071, .052] |
| Double Lasso | -.050 | .014 | [-.078, -.022] |

As expected, least squares provides a rather noisy estimate of the speed of convergence, and does not allow us to answer the question about the convergence hypothesis.

In sharp contrast, double Lasso provides a more precise estimate.

The lasso based point estimate is $-5\%$ and the $95\%$ confidence interval for the (annual) rate of convergence is $-7.8\%$ to $-2.2\%$.

This empirical evidence does support the convergence hypothesis.

These inferences are justified by the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$ for double Lasso.

The regularization that come from the use of the penalty in the Lasso regression of $Y$ on $Z$ and of $D$ on $Z$ does not invalidate this because the partialled out regression is Neyman orthogonal with respect to the coefficients on $Z$.

The identifying moment equation that $\theta_0$ solves for the partialled out regression is

$$E[\tilde{D}(\tilde{Y} - \tilde{D}\theta)] = 0.$$

Because $Z$ is orthogonal to the least squares residuals for both $Y$ and $Z$, the derivative of the left-hand side of this equation with respect to any of the $\gamma$ terms is zero.

In particular

$$\partial_\gamma E[(D - Z'\gamma)\{\tilde{Y} - \tilde{D}\theta\}] = -E[Z(\tilde{Y} - \tilde{D}\theta)] = 0,$$

for any $\theta$, because orthogonality of population least squares residuals implies $E[Z\tilde{Y}] = E[Z\tilde{D}] = 0$.

Similarly, the derivatives with respect to the other $\gamma's$ are zero.

Thus, the partialled out moment equation is Neyman orthogonal with respect to each $Z$ coefficient, so Lasso estimation has no first order effect on the estimator of $\theta$.

# WHAT HAPPENS WITHOUT NEYMAN ORTHOGONALITY

If we don't have Neyman orthogonality, we should not expect to get high-quality estimates of the target parameters.

For example, a standard approach that is still used in some statistical applications applies Lasso regression of $Y$ on $D$ and $Z$ to select relevant covariates $\tilde{Z}$, as those covariates that have positive Lasso coefficients, in addition to the covariate of interest, then refits the model by least squares of $Y$ on $D$ and $\tilde{Z}$, and carries out conventional inference.

This is a post Lasso estimation where we always include $D$ in the second regression.

Despite its use in some applications, this is not a valid approach to perform inference on $\theta_0$, and it can result in very misleading conclusions, as we demonstrate below.

(It is a fine approach to use if our goal is solely a prediction problem but not for inference for $\theta_0$).

This approach relies on the (classical) moment condition

$$E[D(Y - D\theta - Z'\gamma_0)] = 0.$$

which is satisfied at the true value $\theta = \theta_0$.

$$E[D(Y - D\theta - Z'\gamma_0)] = 0.$$

We don't have Neyman Orthogonality here, since

$$\partial_\gamma E[D(Y - D\theta - Z'\gamma)] = -E[DZ] \neq 0,$$

unless $D$ is orthogonal to $Z$.

Consequently, this Lasso based selection estimator has a well known selection bias; the usual confidence intervals ignoring selection are not asymtotically correct.

Consequently, while the procedure provides an estimator of $\theta_0$ that is consistent, i.e. will approach the true value in large samples, the bias precludes the use of standard inference.

We can set up a simulation experiment to verify that this approach gives a low quality estimator of $\theta_0$.

We compare the performance of the naive and orthogonal methods in a computational experiment, where

$$p = n = 100,$$

where $\gamma_j = 1/j^2$ and $(\gamma_{DZ})_j = 1/j^2$, and where

$$Y = 1 \cdot D + Z'\gamma + \varepsilon_Y, \quad Z \sim N(0, I), \quad \varepsilon_Y \sim N(0, 1)$$

$$D = Z'\gamma_{DZ} + \tilde{D}, \quad \tilde{D} \sim N(0, 1)/4.$$

for this estimation strategy, and the orthogonal estimator, based on partialling out, is approximately unbiased and approximately Gaussian.

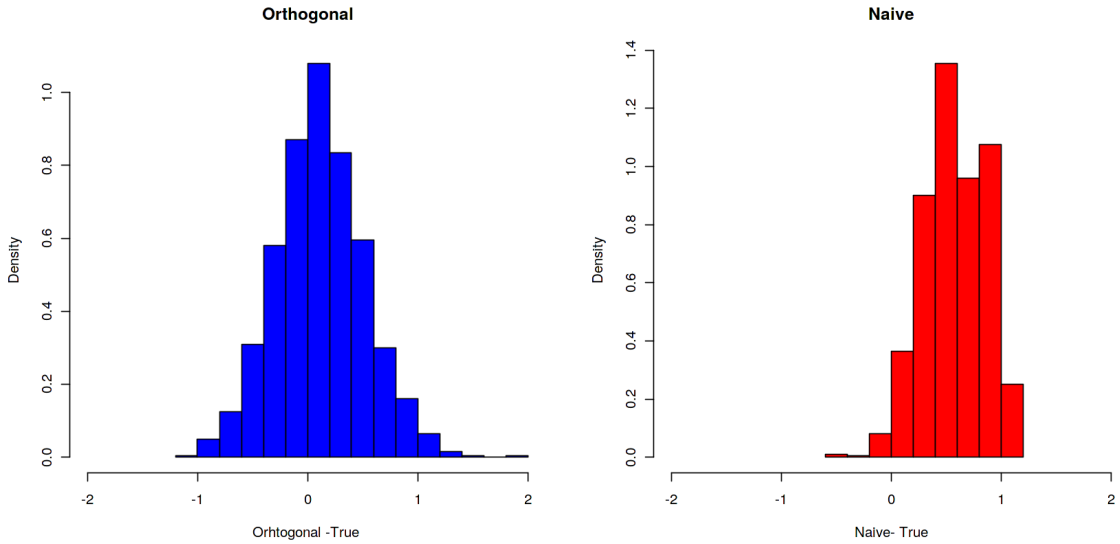



**Figure 12.1: Left Panel:** Simulated distribution of the centered orthogonal estimator minus the true value. **Right Panel:** Simulated distribution of the naive (single-selection) non-orthogonal estimator minus the true value

The reason, that the naive estimator does not perform well, is that it only selects controls $W_j$' s that are strong predictors of outcome, thereby omitting weak predictors, but the very same predictors could be strong predictors of $D$, and dropping these controls results in a strong omitted variable bias. In contrast, the orthogonal approach solves two prediction problems: one to predict $Y$ and another predict $D$ – and finds controls that are relevant for either. The resulting residuals are therefore approximately "de-counfounded".

## 12.3 General Debiased GMM

The general construction relies on a method-of-moments estimator for some low-dimensional target parameter $\theta_0$ based upon the empirical analog of the moment condition

$$\mathrm{E}g(W;\theta_0,\eta_0) = 0, \tag{12.4}$$

where we call $g$ the score function, the $W$ denotes a data vector, $\theta_0$ denotes the true value of a low-dimensional parameter of interest, and $\eta$ denotes parameters or (nuisance functions, more generally) with true value $\eta_0$, estimated by penalized regression such as lasso.

From the histograms the naive estimator is heavily biased, as expected from the failure of the Neyman orthogonality for this estimation strategy.

In contrast, the orthogonal estimator, based on partialling out, is approximately unbiased and approximately Gaussian.

The reason, that the naive estimator does not perform well, is that it only selects controls $Z_j$' s that are strong predictors of outcome, thereby omitting weak predictors, but the very same predictors could be strong predictors of $D$, and dropping these controls results in a strong omitted variable bias.

In contrast, the partialling out approach solves two prediction problems: one to predict $Y$ and another predict $D$ – and finds controls that are relevant for either.

The resulting residuals are therefore approximately “de-confounded".

# CROSS-FITTING

Cross-fitting is a form of sample splitting, where we average over different observations in forming moment conditions for parameters of interest than we use to estimate unknown parts of moment functions.

Cross-fitting eliminates one source of bias and is needed when we partial out using ML other than Lasso, such as neural nets.

Partition the observation indices $(i = 1, ..., n)$ into $L$ groups $I_\ell$, $(\ell = 1, ..., L)$.

Let $\hat{\gamma}_{YZ\ell}$ and $\hat{\gamma}_{DZ\ell}$ be Lasso regression estimates using observations *not* in $I_\ell$ for Lasso regression of $Y$ on $Z$ and $D$ on $Z$ respectively and cross-fit residuals be

$$\hat{U}_{i\ell Y} = Y_i - Z_i'\hat{\gamma}_{YZ\ell}, \text{ and } \hat{U}_{i\ell D} = D_i - Z_i'\hat{\gamma}_{DZ\ell}, \ i \in I_\ell.$$

Form a cross-fit sample moment function as

$$\hat{\psi}(\theta) = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{U}_{i\ell D}(\hat{U}_{i\ell Y} - \hat{U}_{i\ell D}\theta).$$

We choose $\hat{\theta}$ so that $\hat{\psi}(\hat{\theta}) = 0$ and solve for $\hat{\theta}$ to obtain

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}.$$

This is a cross-fit debiased machine learner of a function $\theta_0$ of a conditional mean.