# Supplement to "Causal Strategic Linear Regression"

**Yonadav Shavit** [1]  **Benjamin L. Edelman** [1]  **Brian Axelrod** [2]

## 1. Appendix

### 1.1. Agent Outcomes

*Proof of Theorem 1.* Let's walk through the steps of the algorithm, bounding the error that accumulates along the way.

In the first round we set $\omega = 0$ in order to obtain an estimate for $E[\omega^{*T}x]$.

Since $\omega^*$ is a unit vector, the variance of $\omega^{*T}x$ is at most $\lambda_{max}$ plus a constant (from the $1-$subgaussian noise).

This means that $O(\epsilon^{-1}d\lambda_{max})$ samples suffice for the empirical estimator of $E[\omega^{*T}x]$ to have no more than $\frac{\epsilon}{4d}$ error with failure probability $\Omega(\frac{1}{2d})$. We call the output of this estimator $\hat{\mu}$ and let $\hat{\mu}_d$ be the r-dimensional vector with $\hat{\mu}$ in every coordinate.

Now we choose $\omega_1....\omega_d$ that form an orthonormal basis of the image of the diagonal matrix $V$. For each $\omega$ we observe the reward $\omega^{*T}(x + G\omega) + \eta$, subtract out $\hat{\mu}$, and plug it into the empirical mean estimator. For each $\omega_i$, let $\hat{\nu}_i$ be the resulting coefficient. After $O(\epsilon^{-1}d\lambda_{max})$ samples, each coefficient has at most $\frac{\epsilon}{4d}$ error with failure probability at most $\frac{1}{2d}$. Since we have computed $d+1$ estimators, each one with failure probability at most $\frac{1}{2d}$, a union bound gives us a total failure probability that is sub-constant.

We can now bound the total squared $\ell_2$ error between said coefficients and $G^T\omega^*$ in the $\omega_1...\omega_d$ basis (noting that the choice of basis does not affect the magnitude of the error). We can break up the error into two components using the triangle inequality: the error due to $\hat{\mu}_d$ and the error in the subsequent rounds. Each coordinate of $\hat{\mu}_d$ has error of magnitude at most $\frac{\epsilon}{4d}$, so the total magnitude of the error in $\hat{\mu}_d$ is at most $\frac{\epsilon}{4}$. The same argument applies for the error in the coordinate estimates, leading to a total $\ell_2$ error of at most $\epsilon/2$.

Recall that $\hat{\omega} = \hat{\nu}/\|\hat{\nu}\|$. Let $\nu := G^T\omega^*$. We can now bound the gap between the agent outcomes incentivized by $\hat{\omega}$ and by $\omega_{imp} = \nu/\nu$:

$$\text{AO}(\omega_{imp}) - \text{AO}(\hat{\omega}) = \nu^T \frac{\nu}{\|\nu\|} - \nu^T \frac{\hat{\nu}}{\|\hat{\nu}\|} \tag{1}$$

$$= \|\nu\| - \nu^T \frac{\hat{\nu}}{\|\hat{\nu}\|} \tag{2}$$

$$\leq \|\nu\| - \frac{\|\nu\|(\|\nu\| - \epsilon/2)}{\|\nu\| + \epsilon/2} \tag{3}$$

$$= \frac{\|\nu\|\epsilon}{\|\nu\| + \epsilon/2} \leq \epsilon \tag{4}$$

$\square$

[1]Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA [2]Stanford Computer Science Department, Palo Alto, CA, USA. Correspondence to: Yonadav Shavit <yonadav@g.harvard.edu>.

## 1.2. Prediction Risk

*Proof of Lemma 1.*

$$Risk(\omega) = \mathbb{E}_{x,a}\left[\left(\omega^T V(x + Ma) - \omega^{*T}(x + Ma)\right)^2\right]$$

$$= \mathbb{E}_{x,a}\left[\left(\left(\omega^T Vx - \omega^{*T}x\right) + \left(\omega^T VMa - \omega^{*T}Ma\right)\right)^2\right]$$

$$= \mathbb{E}_{x,a}\left[\left(\omega^T Vx - \omega^{*T}x\right)^2\right] + \mathbb{E}_{x,a}\left[(V\omega - \omega^*)^T x(Ma)^T(V\omega - \omega^*)\right] + \mathbb{E}_{x,a}\left[\left(\omega^T VMa - \omega^{*T}Ma\right)^2\right]$$

$$= \mathbb{E}_x\left[\left(\omega^T Vx - \omega^{*T}x\right)^2\right] + \mathbb{E}_a\left[\left(\omega^T VMa - \omega^{*T}Ma\right)^2\right]$$

where the last line follows because $Ma$ and $x$ are uncorrelated. □

## 1.3. Parameter Estimation

In this section we describe how we recover $\hat{\omega}_{opt}$ in $L^2$-distance when there exists an $\omega$ such that $\Sigma + G\omega$ is full rank. Before we proceed we make a couple of observations. When there is no way to make the above matrix full rank, we cannot hope to recover the optimal $\hat{\omega}_{opt}$. If there is no natural variation in e.g. the last two features, and furthermore no agent can act along those features, it is not possible to disentangle their potential effects on the outcome. This also suggests that the parameter recovery is a more substantive demand for the decision maker than the standard linear regression setting. To discover this additional information, the decision maker can incentivize the agents to take actions that help the decision-maker recover the true outcome-governing parameters.

This motivates the algorithm we present in this section. It operates in two stages. First, it recovers the information necessary in order to to identify the decision rule which will provide the most informative agent samples after those agents have gamed. Second, it collects data while incentivizing this action. Finally, it computes an estimate of $\hat{\omega}_{opt}$ using the collected data. We present the complete procedure in Algorithm 1.

---

**Algorithm 1** Recovering the Causal Model

1: Let $k_1 = \lambda_{max}(G^T G)$ and $k_2 = ||\Sigma||^2$
2: Let $\kappa_{min} = \lambda_{min}(\Sigma)$
3: Choose an $\epsilon > 0$
4: Let $n_1 = O(\max(\frac{dk_1}{\kappa_{min}}, \frac{d^2 k_2}{\kappa_{min}}))$
5: Collect samples $x_1, \ldots, x_{n_1}$
6: Let $\hat{\mu} = \frac{1}{n_1}\sum x_i$
7: Let $\hat{\Sigma} = \frac{1}{n_1}\sum x_i x_i^T$
8: Let $n_2 = O(\max(d^2||\hat{\mu}||^2 \text{tr}(\Sigma), d^3||G||^2 \text{tr}(\Sigma)))$
9: **for** $i = 1...d$ **do**
10:   $\omega = e_i$
11:   Sample $x_1, \ldots, x_{n_2}^i$ and subtract $\hat{\mu}$ from each one.
12:   Let $\hat{G}_i = \frac{1}{n_2}\sum_{j=1}^{n_2} x_j$
13: **end for**
14: Let $\hat{\omega}_{opt} = \arg\min_{\omega} \hat{\Sigma} + 2\mu\omega^T \hat{G}^T + \hat{G}\omega\omega^T \hat{G}^T$
15: Let $n_3 = O(\frac{d}{\epsilon\kappa_{min}})$
16: Sample $x_1, \ldots, x_{n_3}$ with $\omega = \hat{\omega}_{opt}$.
17: Return the output of OLS on $x_1, \ldots, x_{n_3}$

---

The procedure in Algorithm 1 can be summarized as follows:

1. Estimate the first and second moments of the distribution of agents' features.

2. Estimate the Gramian of the action matrix $G$.

3. Compute the most informative choice of $\omega$.

4. Collect samples under the most informative $\omega$ and then return the output of OLS.

Before we proceed to the proof of correctness of Algorithm 1, let us build some intuition for why this procedure of choosing a single $\omega$ and collecting samples under said $\omega$ makes sense. As we show later, the convergence of OLS for linear regression can be controlled by the minimum eigenvalue of the second moment matrix of the samples. Our algorithm finds the value of $\omega$ that, after agents game, maximizes this minimum eigenvalue in expectation. It turns out the minimum eigenvalue of the expected second moment matrix of post-gaming samples is convex with respect to the choice of $\omega$. The convexity of the objective suggests that a priori, when choosing $\omega$s to obtain informative samples, the optimal strategy is choose a single specific $\omega$.

The main difficulty in the rest of the algorithm is achieving the necessary precision in the estimation to be able to set up the above optimization problem to identify such an $\omega$.

**Theorem 3.** *When $V = I$, the output of Algorithm 1 run with parameter $\epsilon$ satisfies $||\omega - \omega^*|| \leq \epsilon$ with probability greater than $\frac{2}{3}$.*

The proof of Theorem 3 relies on several lemmas. First we bound the $L_2$ error of OLS as a function of the empirical second moment matrix in Lemma 1. Note that the usual bound for the convergence of OLS is distribution dependent. That is, the expected error is small.

**Lemma 1.** *Assume $V = I$. Consider samples $x_1, \ldots, x_n$ and $y_i = \hat{\omega}_{opt}^T x_i + \eta_i$. Let $\omega$ be the output of OLS $(x_i, y_i)$. Then*

$$\mathbb{E}_\eta \left[ ||\omega - \hat{\omega}_{opt}||^2 \right] \leq \frac{d}{n\kappa_{min}}$$

The above proof is elementary and a slight modification of the standard textbook proof (see for example, (Liang, 2016)).

The proof also requires that the optimization to choose the optimal $\omega$ is convex.

**Lemma 2.** *The minimum eigenvalue of the following matrix is convex with respect to $\omega$ for any values of $x, G$.*

$$\sum_i (x_i + \hat{G}\omega)(x_i + \hat{G}\omega)^T$$

*Furthermore, when the following conditions are true, then the minimum eigenvalue of the above is within a constant factor of the optimal value.*

$\mathbb{E}[(x + G\omega)(x + \hat{G}\omega)^T]$.

- $||\hat{\Sigma} - \Sigma||^2 \leq \epsilon$

- $||\mu - \hat{\mu}||^2 \leq \frac{\lambda_{max}(G^T G)\epsilon}{d}$

- $||\hat{G} - G||^2 \leq \frac{\epsilon}{d||\mu||^2}$

- $||\hat{G} - G||^2 \leq \frac{\epsilon}{d^2||G||^2}$

*Finally, the above holds true even for an $\omega$ with distance at most $O(\frac{1}{poly(d)})$ from the optimum.*

Finally, we use a minor lemma for recover of a random vector via the empirical mean estimator. Note that we treat the matrix $G$ as a vector.

**Lemma 3.** *Assume $V = I$. Let $g_1^i, \ldots, g_n^i$ be drawn from the distribution $G_i + \xi$ and $\hat{G}$ be the empirical mean estimator computed from said $g_j^i$'s. Let $\Sigma$ be the expected second moment matrix of the $\xi$s. Then*

$$\mathbb{E}_\xi ||G - \hat{G}||^2 \leq \frac{d^2 \text{tr}(\Sigma)}{n}$$

We proceed with the proof of Theorem 3 below.

*Proof.* The first step of the algorithm is for recovering an estimate of $\Sigma$ and $\mu$. Note that $n_1$ samples suffice to recover $\hat{\Sigma}$ and $\hat{\mu}$ such that:

- $||\hat{\Sigma} - \Sigma||^2 \le \epsilon$

- $||\mu - \hat{\mu}||^2 \le \frac{\lambda_{max}(G^T G)\epsilon}{d}$

The for loop recovers an estimate of $G$. Via Lemma 3, the samples suffice to ensure that the following two conditions hold:

- $||\hat{G} - G||^2 \le \frac{\epsilon}{d||\mu||^2}$

- $||\hat{G} - G||^2 \le \frac{\epsilon}{d^2||G||^2}$

Then the algorithm computes an estimate of the optimal $\omega$. Via Lemma 2, we have that the optimum guarantees the minimum eigenvalue of an approximate solution will be within a constant factor of the optimum.

This $\omega$ guarantees that $n_3$ samples suffice to ensure the recover of $\omega^*$ within squared $L^2$-distance of $O(\epsilon)$ in expectation.

Finally the expectations can be used with a Markov inequality to ensure the algorithm succeeds with (arbitrarily high) constant probability. $\square$

Now we prove the lemmas. We begin with Lemma 1. This proof is a slight modification of the textbook proof for the convergence of OLS.

*Proof.* In this section we derive a bound on the convergence of the least squares estimator when a fixed design matrix $X$ is used. Note this is exactly the case we encounter, since the choice of $\omega$ lets us affect the entries of the design matrix. This is a standard, textbook result and not a main contribution of the paper.

In order to state the result more formally we have to introduce some notation. The goal of the procedure is to recover $\hat{\omega}_{opt}$, when given tuples $(x_i, \hat{\omega}_{opt} x_i + \eta)$ where $\eta$ is 1-subgaussian. We aim to characterize $||\omega - \hat{\omega}_{opt}||$ where $\omega$ is obtained from ordinary least squares. Let $X$ be the vector with the $x_i$'s in its columns. Let $\kappa_{min}$ be the minimum eigenvalue of $\frac{1}{n} X^T X$ (the second moment matrix).

Below all expectations are taken *only* over the random noise. We assume the second moment matrix is full rank.

$$
\begin{aligned}
\mathbb{E}[||\omega - \hat{\omega}_{opt}||^2] &\le \mathbb{E}[\frac{1}{n\kappa_{min}}(\omega - \hat{\omega}_{opt})X^T X(\omega - \hat{\omega}_{opt})] \\
&= \mathbb{E}[\frac{1}{n\kappa_{min}}||X(\omega - \hat{\omega}_{opt})||^2] \\
&= \frac{1}{n\kappa_{min}}\mathbb{E}[||X(X^T X)^{-1}X^T(X\hat{\omega}_{opt} + \eta) - X\hat{\omega}_{opt}||^2] \\
&= \frac{1}{n\kappa_{min}}\mathbb{E}[||X(X^T X)^{-1}X^T \eta||^2] \\
&\le \frac{d}{n\kappa_{min}}
\end{aligned}
$$

This motivates our procedure for parameter recovery. We do so in a fashion that attempts to maximize $\kappa_{min}$. Note that it is the minimum eigenvalue that determines the convergence rate. This is due to the fact that little variation along a dimension makes it hard to disentangle the features' effect on the outcome via $\hat{\omega}_{opt}$ from the constant-variance noise $\eta$. $\square$

Lemma 2 is somewhat more involved. It is proven in three parts. The first is that the optimization problem is convex. The second is that approximate recovery of $S, \mu$, and $G$ suffice for approximately minimizing the original expression. The third is that an approximate solution suffices.

*Proof.* In this section we describe how to choose the value of $\omega$ that maximizes the value of $\kappa_{min}$ for the samples we obtain.

To do so, we examine the expectation of the second moment matrix and make several observations. Let $\Sigma$ denote the expected second moment matrix of $x$ (i.e. $\mathbb{E}[xx^T]$. We have:

$$\mathbb{E}[(x + G\omega)(x + G\omega)^T] = \Sigma + 2\mu\omega^T G^T + G\omega\omega^T G^T$$

1. The minimum eigenvalue of the above expression is concave with respect to $\omega$. This follows due to the following: $x + G\omega$ is a linear operator, the minimum eigenvalue of a Gramian matrix $X^T X$ is concave with respect to $X$, and the expectation of a concave function is concave (Boyd & Vandenberghe, 2004).

2. Since the agent attempts to maximize their motion in the $\omega$ direction, we want to ensure that we move them toward toward the direction that maximizes the minimum eigenvalue of $\mathbb{E}[(x + G\omega)(x + G\omega)^T]$.

However, we do not operate with exact knowledge of $G$, etc. It turns out that even approximately solving this optimization problem with estimates for $G, \Sigma, \mu$ suffices for our purposes, as long as the $\omega$ we obtain from our optimization (using the estimates) results in a high value for the minimum eigenvalue of $\mathbb{E}[(x+G\omega)(x+G\omega)^T]$. Let $\hat{\omega}$ be the maximizing argument for the estimated optimization problem and let $\omega$ be the maximizing argument for the original optimization problem. Let $Q$ be the true maximized second moment matrix including gaming, and $\hat{Q}$ be the maximizing second moment matrix with gaming resulting from replacing the true $\Sigma, \mu, G$ with the estimates. In formal terms, we need to show the minimum eigenvalue of the following is large: $\mathbb{E}[(x + G\hat{\omega})(x + G\hat{\omega})^T]$. We note that when $y^T \hat{Q} y$ is within $\epsilon$ of $y^T Q y$ for all $y$ in the unit ball, the minimum eigenvalues may differ by at most $\epsilon$.

$$\begin{aligned} ||y^T \hat{Q} y - y^T Q y||^2 &= ||y^T (\hat{Q} - Q) y||^2 \\ &\leq \lambda_{max}^2 (\hat{Q} - Q)(y)||y||^2 \\ &\leq ||\hat{Q} - Q||^2 \end{aligned}$$

And now we bound the norm of $||\hat{\Sigma} - \Sigma||^2$ assuming the following:

1. $||\hat{\Sigma} - \Sigma||^2 \leq \epsilon$

2. $||\mu - \hat{\mu}||^2 \leq \frac{\lambda_{max}(G^T G)\epsilon}{d}$

3. $||\hat{G} - G||^2 \leq \frac{\epsilon}{d||\mu||^2}$

4. $||\hat{G} - G||^2 \leq \frac{\epsilon}{d^2||G||^2}$

We work out the bound below.

$$\begin{aligned}
||\hat{Q} - Q||^2 &= ||\Sigma + 2\mu\hat{\omega}_{opt}^T G^T + G\hat{\omega}_{opt}\hat{\omega}_{opt}^T G^T - (\hat{\Sigma} + 2\hat{\mu}\omega^T \hat{G}^T + \hat{G}\hat{\omega}_{opt}\omega_{opt}^T \hat{G}^T)||^2 \\
&\leq ||\Sigma - \hat{\Sigma}||^2 + 2||\mu\hat{\omega}_{opt}^T G^T - \hat{\mu}\hat{\omega}_{opt}^T \hat{G}^T||^2 + ||\dots||^2 \\
&\leq \epsilon + 2||\mu\hat{\omega}_{opt}^T G^T + \hat{\mu}\hat{\omega}_{opt}^T G^T - \hat{\mu}\hat{\omega}_{opt}^T G^T - \hat{\mu}\hat{\omega}_{opt}^T \hat{G}^T||^2 + ||\dots||^2 \\
&\leq \epsilon + d||\mu - \hat{\mu}||^2||\hat{\omega}_{opt}^T G||^2 + ||\hat{G} - G||^2||\hat{\mu}\omega^T||^2 + \dots \\
&\leq \epsilon + \epsilon + ||\hat{G} - G||^2||\hat{\mu}\omega^T + \mu\omega^T - \mu\omega^T||^2 + \dots \\
&\leq \epsilon + \epsilon + ||\hat{G} - G||^2(||\hat{\mu} - \mu||^2 + ||\mu\hat{\omega}_{opt}^T||) + \dots \\
&\leq \epsilon + \epsilon + ||\hat{G} - G||^2 d||\mu||^2 + \dots \\
&\leq 3\epsilon + ||\hat{G}\hat{\omega}_{opt}\hat{\omega}_{opt}^T \hat{G}^T - G\hat{\omega}_{opt}\hat{\omega}_{opt}^T G^T||^2 \\
&\leq 3\epsilon + ||\hat{G}\hat{\omega}_{opt}\hat{\omega}_{opt}^T \hat{G}^T - \hat{G}\hat{\omega}_{opt}\hat{\omega}_{opt}^T G + \hat{G}\hat{\omega}_{opt}\hat{\omega}_{opt}^T G - G\hat{\omega}_{opt}\hat{\omega}_{opt}^T G^T||^2 \\
&\leq 3\epsilon + ||(\hat{G} - G)\hat{\omega}_{opt}\hat{\omega}_{opt}^T \hat{G}^T||^2 + ||(\hat{G} - G)\hat{\omega}_{opt}\hat{\omega}_{opt}^T G^T||^2 \\
&\leq 3\epsilon + ||(\hat{G} - G)\hat{\omega}_{opt}\hat{\omega}_{opt}^T \hat{G}^T - (\hat{G} - G)\hat{\omega}_{opt}\hat{\omega}_{opt}^T G^T + (\hat{G} - G)\hat{\omega}_{opt}\hat{\omega}_{opt}^T G^T||^2 + d^2||\hat{G} - G||^2||G||^2 \\
&\leq 4\epsilon + ||(\hat{G} - G)\hat{\omega}_{opt}\hat{\omega}_{opt}^T(\hat{G} - G)||^2 + ||(\hat{G} - G)\hat{\omega}_{opt}\hat{\omega}_{opt}^T G^T||^2 \\
&\leq 5\epsilon + d^2||\hat{G} - G||^4 \\
&\leq 6\epsilon
\end{aligned}$$

This means if we find an $\epsilon-$approximate solution to the system with the estimated values, we obtain a $\kappa_{min}$ within $6\epsilon$ of the optimal. $\qquad\square$

Finally, we present the proof of Lemma 3:

*Proof.* Recall that when the decision-maker fixes $\omega$, it receives samples of the form $x + G\omega$. We note this can be used to recover the matrix $G$. In particular, we show how $d$ rounds, each with $O(\frac{d\operatorname{tr}(\Sigma)}{\epsilon})$ samples, suffices to recover the matrix to squared Frobenius norm $\epsilon$. Recall the procedure we propose simply chooses $\omega = e_1, \dots e_d$, one-hot coordinate vectors in each round. We first bound the error in $\hat{G}$. coordinate-wise: $\mathbb{E}[||\hat{G}_{i,j} - G_{i,j}||^2] \leq \frac{\mathbb{E}[x_i^2]}{n}$. A union bound across coordinates shows that $O(\frac{d^2\operatorname{tr}(\Sigma)}{\epsilon})$ samples suffice to recover $G$ within squared Frobenius norm $\epsilon$. $\qquad\square$

# References

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge university press, 2004.

Liang, P. Cs229t/stat231: Statistical learning theory (lecture notes). https://web.stanford.edu/class/cs229t/notes.pdf, April 2016.