Adaptive Elastic Net GMM estimator: Model Selection and Estimation with Diverging Number of Parameters

Mehmet Caner
North Carolina State University *

Hao Helen Zhang
North Carolina State University

April 25, 2009

VERY PRELIMINARY AND PARTIAL, PLEASE DO NOT QUOTE

Abstract

This paper considers GMM with large number of parameters, and introduces a new model selection tool. This is extending Zhou and Zhang (2009, Annals of Statistics, Forthcoming) idea from LS. The proofs are entirely different than theirs. The adaptive elastic net basically is a data dependent model selection tool, with the oracle property. This also handles large systems with possibly correlated moments/regressors.

1 Model

Let \( \beta \) be a \( p \) dimensional parameter vector. The true value of \( \beta \) is \( \beta_0 \). \( \beta \in B_p \) which is a compact subset in \( \mathbb{R}^p \). We allow \( p \) to grow with the sample size, so when \( n \to \infty \), we have \( p \to \infty \). We do not provide a subscript of \( n \) for parameter space not to burden ourselves with the notation.

Obviously when \( n \) gets large and converges to infinity and then the parameter space is denoted as \( B_\infty \). \( \theta_0 \in B_\infty \). For example this may be infinite cube \([0, 1]_\infty \) in certain examples. The population orthogonality conditions are

\[
E[g(X, \beta_0) = 0],
\]

where the data are \( \{X_i : i = 1, 2 \cdots n\} \), \( g(.) \) is a known function, and the number of orthogonality restrictions are \( q \), \( q \geq p \). So we also allow \( q \) to grow with the sample size. From now on we denote \( g(X, \beta) \) as \( g_i(\beta) \). We first define various estimators, and eventually link them to adaptive elastic net estimators and use them in our proofs. The first one is the elastic net estimator:

\[
\hat{\beta}_{enet} = (1 + \lambda_2/n) \{ \arg\min_\beta S_n(\beta) \},
\]

where

\[
S_n(\beta) = \left[ \sum_{i=1}^{n} g_i(\beta) \right]' W_n [\sum_{i=1}^{n} g_i(\beta)] + \lambda_2 \| \beta \|_2^2 + \lambda_1 \| \beta \|_1,
\]

\( W_n \) is a \( q \times q \) weight matrix that will be defined in Assumptions below, and \( \lambda_1, \lambda_2 \) are positive constants that will be defined in theorems. Also \( \| \beta \|_1 = \sum_{j=1}^{p} |\beta_j| \), and \( \| \beta \|_2^2 = \sum_{j=1}^{p} |\beta_j|^2 \). We specifically analyze the following estimator, which is connected to elastic net estimator above

\[
\hat{\beta}(\lambda_2, \lambda_1) = \arg\min_\beta S_n(\beta).
\]

We will prove the consistency of \( \hat{\beta}(\lambda_2, \lambda_1) \) estimator which will show the consistency of the elastic net estimator. Then we define another estimator which is tied to adaptive elastic net estimator.

\[
\hat{\beta}_a = \arg\min_\beta \left[ \left( \sum_{i=1}^{n} g_i(\beta) \right)' W_n [\sum_{i=1}^{n} g_i(\beta)] + \lambda_2 \| \beta \|_2^2 + \lambda_1^* \sum_{j=1}^{p} |\hat{\beta}_j| \right],
\]

where \( \hat{\beta}_j = \frac{1}{|\hat{\beta}_{j, enet}|} \), \( \gamma \) is a positive constant. \( \lambda_1^* \) is a sequence of positive numbers, which is different than \( \lambda_1 \). Also if \( \hat{\beta}_{j, enet} \to 0 \), then \( \hat{\beta}_j \to \infty \). The other possibility is if \( \hat{\beta}_{j, enet} \to \beta_j \neq 0 \), then \( \hat{\beta}_j \to u_j = \frac{1}{|\beta_j|} \).

We now provide the main assumptions that are used to derive the consistency of elastic net estimator and \( \hat{\beta}_a \) estimates which is related to adaptive elastic net.

**Assumptions**

1. The following uniform law of large number holds

\[
\sup_{\beta \in B_p} \left[ \frac{1}{n} \sum_{i=1}^{n} |g_i(\beta) - Eg_i(\beta)| \right] \overset{p}{\to} 0.
\]
2. Define \( En^{-1} \sum_{i=1}^{n} g_i(\beta) = m_{1n}(\beta) \), then

(i). Assume uniformly over \( \beta \) that \( m_{1n}(\beta) \rightarrow m_1(\beta) \), \( m_{1n}(\beta_0) = 0 \), and \( m_1(\beta) \neq 0 \), \( \beta \neq \beta_0 \), \( m_1(\beta) \) is continuous in \( \beta \).

(ii). Define the following \( q \times p \) matrix \( \hat{G}_n(\beta) = \sum_{i=1}^{n} \frac{\partial g_i(\beta)}{\partial \beta} \), assume the following uniform law of large numbers in a neighborhood \( \mathcal{N} \) of \( \beta_0 \), \( \hat{G}_n(\beta)/n \rightarrow G(\beta) \), where \( G(\beta_0) \) is of full column rank \( p \), and \( G(\beta) \) is continuous in \( \beta \).

3. \( W_n \) is a positive definite matrix, and \( W_n \xrightarrow{p} W \), where \( W \) is symmetric and positive definite matrix.

4. (i). \( b \leq \text{Eig}_{\min}(G(\beta_0)'WG(\beta_0)), \quad \text{Eig}_{\max}(\Sigma) \leq B \), where \( \Sigma = G(\beta_0)'\Omega WG(\beta_0) \) and \( \Omega = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} Eg_i(\beta_0)g_i(\beta_0)' \). \( b \) and \( B \) are positive constants.

(ii). \( p/n \rightarrow 0 \) as \( n \to \infty \), \( p = n^\alpha, 0 < \alpha < 1 \), \( q/n \rightarrow 0 \) as \( n \to \infty \), and \( q \rightarrow \infty \).

Note that since \( B_p \) is compact \( B_\infty \) is compact through Theorem 6.17 (Tychonoff’s Theorem) in Davidson (1994). Assumption 1 simplifies the proofs since the main concern is the limit law and oracle property of the estimators. The uniform law of large number result can be obtained from a triangle array based Donsker result which is used in empirical process theory (Andrews, 1994). The primitives for that functional central limit theorem is the Lipschitz continuity and the uniform boundedness of the functions. Also we can benefit from another functional central limit theorem in van der Vaart (1998), Theorem 19.28 and the primitive of Lipshitz continuous functions are in Lemma 19.31 of van der Vaart (1998). Assumptions 2 and 3 are standard in the literature of GMM (Newey and Mc Fadden, 1994).

The following theorem provides consistency for both elastic net and an estimator closely connected to adaptive elastic net.

**Theorem 1.** Under Assumptions 1-3,

(i). additionally \( \lambda_2 = o(n^{2-\alpha}), \lambda_1 = o(n^{2-\alpha}) \), then we have

\[ \hat{\beta}_{\text{enet}} \xrightarrow{p} \beta_0. \]

(ii). additionally \( \lambda_2 = o(n^{2-\alpha}), \lambda_1^* = o_p(n^{(2-\alpha)+\frac{\gamma}{\alpha-1}}) \), then

\[ \hat{\beta}_w \xrightarrow{p} \beta_0. \]

Remarks. So it is clear that since \( \alpha > 0 \), the penalty is growing in a smaller amount compared with fixed number of parameters case (\( \alpha = 0 \)). For \( \lambda_1^* \) to diverge we need \( 0 < \gamma < \frac{2(2-\alpha)}{1-\alpha} \).

**Proof of Theorem 1(i).** Via Assumptions 1-3, uniformly over \( \beta \)

\[ \left[ n^{-1} \sum_{i=1}^{n} g_i(\beta) \right]'W_n \left[ n^{-1} \sum_{i=1}^{n} g_i(\beta) \right] \xrightarrow{p} m_1(\beta)'Wm_1(\beta). \] (1)
Now we show why we need the assumptions about the penalty rates in the proofs. First start with the following, if we had assumed the following instead

\[ \frac{\lambda_2}{n^2} \| \beta \|_2^2 \to m_2(\beta). \]  
(2)

\[ \frac{\lambda_1}{n^2} \| \beta \|_1 \to m_3(\beta). \]  
(3)

Note that the limit functions are positive. The uniformly over \( \beta \), the limit would have been

\[ \frac{S_n(\beta)}{n^2} \mathbf{p} \to m_1(\beta)'Wm_1(\beta) + m_2(\beta) + m_3(\beta). \]

So in that case,

\[ \hat{\beta}(\lambda_2, \lambda_1) \overset{p}{\to} \text{arginf}_\beta S_n(\beta), \]

due to the existence of \( m_2(\beta), m_3(\beta) \), under (2)(3), there is no consistency of the estimator. However, if we modify (2)(3) in the following way

\[ \frac{\lambda_2}{n^2} \| \beta \|_2^2 = \frac{\lambda_2}{n^2} \sum_{j=1}^p |\beta_j|^2 \to 0. \]  
(4)

\[ \frac{\lambda_1}{n^2} \| \beta \|_1 = \frac{\lambda_1}{n^2} \sum_{j=1}^p |\beta_j| \to 0. \]  
(5)

To get (4)(5) we need \( \frac{\lambda_2}{n^2} p \to 0 \) and \( \frac{\lambda_1}{n^2} p \to 0 \), where we need \( \lambda_2 = o(n^{2-\alpha}) \), \( \lambda_1 = o(n^{2-\alpha}) \) with the usage of \( p = n^\alpha \).

Then with (4)(5) instead of (2)(3) we have by Assumptions 1-3

\[ \hat{\beta}(\lambda_2, \lambda_1) \overset{p}{\to} \text{arginf}_\beta [m_1(\beta)'Wm_1(\beta)] \]

Use Corollary 3.2.3 of van der Vaart and Wellner (1996) to have

\[ \beta_0 = \text{arginf}_\beta [m_1(\beta)'Wm_1(\beta)]. \]

This is also a unique minimum by Assumptions 2 and 3. So

\[ \hat{\beta}(\lambda_2, \lambda_1) \overset{p}{\to} \beta_0, \]

and by the definition of \( \hat{\beta}_{enet} \),

\[ \hat{\beta}_{enet} \overset{p}{\to} \beta_0. \]

Q.E.D.

Proof of Theorem 1(ii). This follows exactly in the case of (i), given \( \lambda_3 = o(n^{2-\alpha}) \), and

\[ \frac{\lambda_1^2}{n^2} \sum_{j=1}^p \hat{w}_j |\beta_j| \to 0. \]  
For the proof of the last point we need this to be true for the "zero" parameters as well as "nonzero" parameters.
For the zero parameter case, \( \hat{w}_j = \frac{1}{|\hat{\beta}_{j,enet}|} \) and \( \hat{\beta}_{j,enet} = O_p(\sqrt{p/n}) \) via Theorem 2i below (independent of the proof here in Theorem 1ii). So
\[
\hat{w}_j = O_p(n^{2(1-\alpha)}),
\]
where we use \( p = n^\alpha \). So for the zero parameters
\[
\frac{\lambda^*_1}{n^2} \sum_{j=1}^p \hat{w}_j |\beta_j| = O_p(\lambda^*_1 n^{\alpha-2}n^{2(1-\alpha)}) \to 0,
\]
where we use \( \sum_{j=1}^p |\beta_j| = O(n^{\alpha}) \). The last equation provides us with
\[
\lambda^*_1 = o_p(n^{(2-\alpha)+\gamma(\alpha-1)}). \tag{6}
\]
For the nonzero parameters, in elastic net estimator using Theorem 1i, \( \hat{\beta}_{j,enet} \xrightarrow{p} \beta_{j,0} \neq 0 \). So
\[
\frac{\lambda^*_1}{n^2} \sum_{j=1}^p \hat{w}_j |\beta_j| \xrightarrow{p} 0,
\]
if \( \frac{\lambda^*_1}{n^2} p \xrightarrow{p} 0 \), which is true if
\[
\lambda^*_1 = o_p(n^{2-\alpha}). \tag{7}
\]
To have the consistency we need the minimum of the rates in (6)(7) which is
\[
\min(2 - \alpha, (2 - \alpha) + \gamma\frac{2}{\alpha}(\alpha - 1)) = (2 - \alpha) + \frac{\gamma}{2}(\alpha - 1).
\]
This is true since \( \gamma > 0, \alpha < 1 \).

Q.E.D.

Proof of Theorem 2. In this proof we start by analyzing the GMM-Ridge estimator that is defined as follows:
\[
\hat{\beta}_R = \arg\min_{\beta} [\sum_{i=1}^n g_i(\beta)]'W_n[\sum_{i=1}^n g_i(\beta)] + \lambda_2 \|\beta\|^2.
\]
Note that this estimator is similar to the elastic net estimator, if we set \( \lambda_1 = 0 \), in elastic net estimator, then we obtain the GMM-Ridge estimator. So since elastic net estimator is consistent, GMM-Ridge will be consistent as well. Define also the following \( q \times p \) matrix \( \hat{G}_n(\hat{\beta}_R) = \sum_{i=1}^n \partial g_i(\beta_n) / \partial \beta \).
Then set \( \hat{\beta} \in (\beta_0, \hat{\beta}_R) \). A mean value theorem around \( \beta_0 \) applied to first order conditions provides, with \( g_n(\beta_0) = \sum_{i=1}^n g_i(\beta_0) \),
\[
\hat{\beta}_R = -(\hat{G}_n(\hat{\beta}_R)'W_n\hat{G}_n(\hat{\beta}_R) + \lambda_2 I_p)^{-1}[\hat{G}_n(\hat{\beta}_R)'W_n g_n(\beta_0) - \hat{G}_n(\hat{\beta}_R)'W_n \hat{G}_n(\hat{\beta}_R)\beta_0]. \tag{8}
\]
Also using the mean value theorem with first order conditions yields
\[
\hat{\beta}_R - \beta_0 = -(\hat{G}_n(\hat{\beta}_R)'W_n\hat{G}_n(\hat{\beta}_R) + \lambda_2 I_p)^{-1}[\hat{G}_n(\hat{\beta}_R)'W_n g_n(\beta_0) + \lambda_2 \beta_0]. \tag{9}
\]
We need the following expressions by using (8)

\[
\beta_R^0[\bar{G}_n(\beta_R)'W_n\bar{g}_n(0) - \bar{G}_n(\beta_R)'W_n\bar{G}_n(\beta)\beta_0] = -[\bar{G}_n(\beta_R)'W_n\bar{g}_n(0) - \bar{G}_n(\beta_R)'W_n\bar{G}_n(\beta)\beta_0]' \\
\times [\bar{G}_n(\beta_R)'W_n\bar{G}_n(\beta) + \lambda_2 I]^{-1} \\
\times [\bar{G}_n(\beta_R)'W_n\bar{g}_n(0) - \bar{G}_n(\beta_R)'W_n\bar{G}_n(\beta)\beta_0](10)
\]

\[
\beta_R^0[\bar{G}_n(\bar{\beta}) + \lambda_2 I_p]\beta_R = [\bar{G}_n(\beta_R)'W_n\bar{g}_n(0) - \bar{G}_n(\beta_R)'W_n\bar{G}_n(\beta)\beta_0]' \\
\times [\bar{G}_n(\beta_R)'W_n\bar{G}_n(\bar{\beta}) + \lambda_2 I_p]^{-1}[\bar{G}_n(\beta_R)'W_n\bar{g}_n(0) - \bar{G}_n(\beta_R)'W_n\bar{G}_n(\bar{\beta})\beta_0]. (11)
\]

Note that the right hand side expression in (10) is just the negative of the right hand side of the expression in (11).

Next the aim is to rewrite the following GMM-Ridge objective function via a mean value expansion

\[
\left[\sum_{i=1}^n g_i(\beta_R)'W_n\left[\sum_{i=1}^n g_i(\beta_R)\right]\right] + \lambda_2\|\hat{\beta}_R\|^2_2 = g_n(\beta_0)'W_n\bar{g}_n(\beta_0) \\
\times g_n(\beta_0)'W_n\bar{G}_n(\bar{\beta})(\hat{\beta}_R - \bar{\beta}_0) + (\hat{\beta}_R - \beta_0)'\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) \\
\times (\hat{\beta}_R - \beta_0)'\bar{G}_n(\bar{\beta})'W_n\bar{G}_n(\bar{\beta})(\hat{\beta}_R - \bar{\beta}_0) + \lambda_2\|\hat{\beta}_R\|^2_2. (12)
\]

Furthermore we can rewrite the right hand side of (12) as

\[
g_n(\beta_0)'W_n\bar{g}_n(\beta_0) + \beta_R^0[\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) - \bar{G}_n(\bar{\beta})'W_n\bar{G}_n(\bar{\beta})\beta_0] \\
+ [\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) - \bar{G}_n(\bar{\beta})'W_n\bar{G}_n(\bar{\beta})\beta_0]'\hat{\beta}_R \\
+ \beta_R^0[\bar{G}_n(\bar{\beta})'W_n\bar{G}_n(\bar{\beta}) + \lambda_2 I_p]\hat{\beta}_R \\
- g_n(\beta_0)'W_n\bar{g}_n(\beta_0) - \beta_0^0\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) + \beta_0^0\bar{G}_n(\bar{\beta})'W_n\bar{G}_n(\bar{\beta})\beta_0 \\
= g_n(\beta_0)'W_n\bar{g}_n(\beta_0) - \beta_0^0\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) + \beta_0^0\bar{G}_n(\bar{\beta})'W_n\bar{G}_n(\bar{\beta})\beta_0. (13)
\]

The equality is obtained through (10)(11). Similar to the case above, for the estimator \(\hat{\beta}_w\) we have the following and the mean value theorem applied to that

\[
\left[\sum_{i=1}^n g_i(\hat{\beta}_w)'W_n\left[\sum_{i=1}^n g_i(\hat{\beta}_w)\right]\right] + \lambda_2\|\hat{\beta}_w\|^2_2 = g_n(\beta_0)'W_n\bar{g}_n(\beta_0) \\
+ \beta_w^0[\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) - \bar{G}_n(\bar{\beta})'W_n\bar{G}(\bar{\beta})\beta_0] \\
+ [\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) - \bar{G}_n(\bar{\beta})'W_n\bar{G}(\bar{\beta})\beta_0]'\hat{\beta}_w \\
+ \beta_w^0[\bar{G}_n(\bar{\beta})'W_n\bar{G}(\bar{\beta}) + \lambda_2 I_p]\hat{\beta}_w \\
- g_n(\beta_0)'W_n\bar{G}_n(\bar{\beta})\beta_0 - \beta_0^0\bar{G}_n(\bar{\beta})'W_n\bar{g}_n(\beta_0) \\
+ \beta_0^0\bar{G}_n(\bar{\beta})'W_n\bar{G}_n(\bar{\beta})\beta_0. (14)
\]
Then see that by assuming \( \hat{\beta} \) to be the same variable in the mean value theorem for both \( \beta_w, \beta_R \) analysis without losing any generality

\[
\hat{\beta}_w [\hat{G}_n(\beta)'W_ng_n(\beta_0) - \hat{G}_n(\beta)'W_n\hat{G}(\beta)\beta_0] = \hat{\beta}_w [\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p] [\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p]^{-1} \\
\times [\hat{G}_n(\beta)'W_ng_n(\beta_0) - \hat{G}_n(\beta)'W_n\hat{G}(\beta)\beta_0] \\
= -\hat{\beta}_w [\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p] \beta_R + o_p(1). \quad (15)
\]

The asymptotically negligible remainder in the last equality is due to (1) and Theorem 1 with \( \lambda_1 = 0 \). Next substitute (15) into (14) to have

\[
[\sum_{i=1}^{n} g_i(\hat{\beta}_w)]'W_n[\sum_{i=1}^{n} g_i(\hat{\beta}_w)] + \lambda_2 \| \hat{\beta}_w \|^2 = g_n(\beta_0)'W_n g_n(\beta_0) - \hat{\beta}_w [\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p] \hat{\beta}_R \\
+ \hat{\beta}_R [\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p] \hat{\beta}_w - g_n(\beta_0)'W_n\hat{G}(\beta)\beta_0 \\
- \beta_0 [\hat{G}_n(\beta)'W_n g_n(\beta_0) + \beta_0 [\hat{G}_n(\beta)'W_n\hat{G}(\beta)\beta_0 + o_p(1). \quad (16)
\]

Now subtract (13) from (16) to have

\[
[\sum_{i=1}^{n} g_i(\hat{\beta}_w)]'W_n[\sum_{i=1}^{n} g_i(\hat{\beta}_w)] + \lambda_2 \| \hat{\beta}_w \|^2 - (\sum_{i=1}^{n} g_i(\hat{\beta}_R)]'W_n[\sum_{i=1}^{n} g_i(\hat{\beta}_R)] + \lambda_2 \| \hat{\beta}_R \|^2 \\
= (\hat{\beta}_w - \hat{\beta}_R)'[\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p] (\hat{\beta}_w - \hat{\beta}_R) + o_p(1). \quad (17)
\]

After this important result see that by the definitions of \( \hat{\beta}_w, \hat{\beta}_R \)

\[
\lambda_1 \sum_{j=1}^{p} (\hat{\beta}_{j,R} - \hat{\beta}_{j,w}) \geq (\sum_{i=1}^{n} g_i(\hat{\beta}_w)]'W_n[\sum_{i=1}^{n} g_i(\hat{\beta}_w)] + \lambda_2 \| \hat{\beta}_w \|^2 \\
- (\sum_{i=1}^{n} g_i(\hat{\beta}_R)]'W_n[\sum_{i=1}^{n} g_i(\hat{\beta}_R)] + \lambda_2 \| \hat{\beta}_R \|^2. \quad (18)
\]

Then also see that

\[
\sum_{j=1}^{p} \hat{w}_j (|\hat{\beta}_{j,R}| - |\hat{\beta}_{j,w}|) \leq \sqrt{\sum_{j=1}^{p} (\hat{w}_j)^2 \| \beta_R - \beta_w \|_2}. \quad (19)
\]

Next benefit from (18), with (17)(19) to have

\[
(\hat{\beta}_w - \hat{\beta}_R)'[\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p] (\hat{\beta}_w - \hat{\beta}_R) + o_p(1) \leq \lambda_1 \sqrt{\sum_{j=1}^{p} (\hat{w}_j)^2 \| \beta_R - \beta_w \|_2}. \quad (20)
\]

See that \( Eig_{min}(\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) + \lambda_2 I_p) = Eig_{min}(\hat{G}_n(\beta)'W_n\hat{G}_n(\beta)) + \lambda_2 \). Use this in (20) to have

\[
\{Eig_{min}(\hat{G}_n(\beta)'W_n\hat{G}_n(\beta)) + \lambda_2\} \| \hat{\beta}_w - \hat{\beta}_R \|^2 \leq \lambda_1 \sum_{j=1}^{p} (\hat{w}_j)^2 \| \beta_R - \beta_w \|_2.
\]
This results in
\[
\|\hat{\beta}_w - \hat{\beta}_R\|_2 \leq \frac{\lambda_1 \sqrt{\sum_{j=1}^p \hat{\omega}_j^2} + o_p(1)}{E_{g_{\text{min}}}(\hat{G}_n(\beta)W_n\hat{G}_n(\beta)) + \lambda_2}. \tag{21}
\]
We also want to modify the last inequality. By the consistency of \(\hat{\beta}_w, \hat{\beta}_R, \hat{\beta} \overset{p}{\to} \beta_0\). Then since \(g(.)\) is continuously differentiable with the uniform law of large numbers on the partial derivative we have
\[
\begin{align*}
\left[ \frac{\hat{G}_n(\beta)}{n} \right]'_n W_n \left[ \frac{\hat{G}_n(\beta)}{n} \right] - G(\beta_0)'WG(\beta_0) & \overset{p}{\to} 0.
\end{align*}
\]
The last equation is true also for \(\hat{\beta}_w, \hat{\beta}_R\) replacing \(\hat{\beta}\). Then
\[
\hat{G}_n(\beta)'W_n\hat{G}_n(\beta) = n^2[G(\beta_0)'WG(\beta_0)] + o_p(n^2). \tag{22}
\]
Modify (21) in the following way given the last equality
\[
\|\hat{\beta}_w - \hat{\beta}_R\|_2 \leq \frac{\lambda_1 \sqrt{\sum_{j=1}^p \hat{\omega}_j^2} + o_p(1)}{n^2E_{g_{\text{min}}}(G(\beta_0)'WG(\beta_0)) + \lambda_2 + o_p(n^2)}. \tag{23}
\]
Now we consider the second part of the inequality \(|\|\hat{\beta}_R - \beta_0\||\). We need GMM ridge formula. Note that from (9)
\[
\begin{align*}
\hat{\beta}_R - \beta_0 &= -\lambda_2[\hat{G}_n(\beta_R)'W_n\hat{G}_n(\beta_R) + \lambda_2I_p]^{-1}\beta_0 \\
& \quad - [\hat{G}_n(\beta_R)'W_n\hat{G}_n(\beta) + \lambda_2I_p]^{-1}[\hat{G}_n(\beta_R)'W_ng_n(\beta_0)]. \tag{24}
\end{align*}
\]
We try to modify the equation above a little.
In the same way we obtain (22)
\[
\hat{G}_n(\beta_R)'W_ng_n(\beta_0) = n[G(\beta_0)'Wg_n(\beta_0)] + o_p(n). \tag{25}
\]
Second, see that \(Eg_n(\beta_0)g_n(\beta_0)' = nEg_i(\beta_0)g_i(\beta_0)' = n\Omega\). Next
\[
\begin{align*}
E[g_n(\beta_0)'WG(\beta_0)G(\beta_0)'Wg_n(\beta_0)] &= tr\{G(\beta_0)'WEG_n(\beta_0)'WG(\beta_0)\} \\
& = ntr\{G(\beta_0)'W\Omega WG(\beta_0)\} \\
& \leq npE_{\text{max}}(\Sigma), \tag{26}
\end{align*}
\]
where we use \(\Sigma = G(\beta_0)'W\Omega WG(\beta_0)\). Now we modify (24) using (25) (22)
\[
\begin{align*}
\hat{\beta}_R - \beta_0 &= -\lambda_2[n^2G(\beta_0)'WG(\beta_0) + \lambda_2I_p + o_p(n^2)]^{-1}\beta_0 \\
& \quad - [n^2G(\beta_0)'WG(\beta_0) + \lambda_2I_p + o_p(n^2)]^{-1}[nG(\beta_0)'Wg_n(\beta_0)]. \tag{27}
\end{align*}
\]
Then see that
\[
E(\|\hat{\beta}_R - \beta_0\|^2) \leq 2\lambda_0^2 \left[ n^2 E[g_{\min}(G(\beta_0)^\prime W G(\beta_0)) + \lambda_2 + o_p(n^2) \right]^{-2} \|\beta_0\|^2_2 \\
+ 2 \left[ n^2 E[g_{\min}(G(\beta_0)^\prime W G(\beta_0)) + \lambda_2 + o_p(n^2) \right]^{-2} \\
\times n^2 E[g_n(\beta_0)^\prime W G(\beta_0)^\prime W g_n(\beta_0)] + o_p(n^2) \\
\leq 2\left[ n^2 E[g_{\min}(G(\beta_0)^\prime W G(\beta_0)) + \lambda_2 + o_p(n^2) \right]^{-2} \\
\times [\lambda_0^2\|\beta_0\|^2 + n^3 p E[g_{\max}(\Sigma)] + o_p(n^2)],
\]
where the last inequality is by (26). Now use (23) and (28) to have
\[
E(\|\hat{\beta}_w - \beta_0\|^2) \leq 2E(\|\hat{\beta}_R - \beta_0\|^2) + 2E(\|\hat{\beta}_w - \hat{\beta}_R\|^2) \\
\leq \frac{4\lambda_0^2\|\beta_0\|^2_2 + 4n^3 p B + o_p(n^2) + 2\lambda_0^2 E\sum_{j=1}^p \hat{w}_j^2 + o_p(1)}{[n^2 b + \lambda_2 + o_p(n^2)]^2}.
\]
See that \( b = E[g_{\min}(G(\beta_0)^\prime W G(\beta_0)), B = E[g_{\max}(\Sigma)] \). Q.E.D

Write \( \beta_0 = (\beta_{A,0}^\prime, 0)^\prime \) where \( \beta_{A,0} \) represents the vector of nonzero parameters (true values) that grow with the sample size. Then define
\[
\bar{\beta} = \arg\min_{\beta} [\sum_{i=1}^n g_i(\beta_A)^\prime W_n(\sum_{i=1}^n g_i(\beta_A) + \lambda_2 \sum_{j \in A} \beta_j^2 + \lambda_1^\prime \sum_{j \in A} \hat{w}_j|\beta_j|],
\]
where \( A = \{ j : \beta_{j0} \neq 0, j = 1, 2, \ldots, p \} \). The issue is to show that with probability one \( [(1 + \lambda_2/n)\bar{\beta}, 0] \) converging to the solution of the following
\[
\hat{\beta}_{aenet} = (1 + \lambda_2/n)\{\arg\min_{\beta} [\sum_{i=1}^n g_i(\beta)^\prime W_n(\sum_{i=1}^n g_i(\beta)) + \lambda_2 \|\beta\|^2_2 + \lambda_1^* \sum_{j=1}^p \hat{w}_j|\beta_j|],
\]
where this is the adaptive elastic net estimator.

**Proof of Theorem 3.** To prove the Theorem we need to show the following (Note that by Kuhn-Tucker conditions of (30)),
\[
P[\forall j \in A^c, |2\hat{G}_{n,j}(\bar{\beta})^\prime W_n(\sum_{i=1}^n g_i(\bar{\beta}))| \leq \lambda_1^* \hat{w}_j] \rightarrow 1,
\]
where \( \hat{G}_n(\bar{\beta}) = \sum_{i=1}^n \frac{\partial g_i(\bar{\beta})}{\partial \beta_j}, \) and \( A^c = \{ j : \beta_{j0} = 0, j = 1, 2, \ldots, p \} \). \( \hat{G}_{n,j}(\bar{\beta}) \) denotes the j th column of the partial derivative matrix which corresponds to zero parameters, evaluated at \( \bar{\beta} \). Or we need to show
\[
P[\exists j \in A^c, |2\hat{G}_{n,j}(\bar{\beta})^\prime W_n(\sum_{i=1}^n g_i(\bar{\beta}))| > \lambda_1^* \hat{w}_j] \rightarrow 0.
\]
Now set \( \eta = \min_{j \in A} |\beta_{j0}|, \hat{\eta} = \min_{j \in A} |\hat{\beta}_{j, aenet}| \). So
\[
P[\exists j \in A^c, |2\hat{G}_{n,j}(\bar{\beta})^\prime W_n(\sum_{i=1}^n g_i(\bar{\beta}))| > \lambda_1^* \hat{w}_j \] \leq \sum_{j \in A^c} P[|2\hat{G}_{n,j}(\bar{\beta})^\prime W_n(\sum_{i=1}^n g_i(\bar{\beta}))| > \lambda_1^* \hat{w}_j, \hat{\eta} > \eta/2] \\
+ P[\hat{\eta} \leq \eta/2].
\]
(31)
Then as in Zou and Zhang (2009), we can show that

\[ P[\hat{\eta} \leq \eta/2] \leq \frac{E\|\hat{\beta}_{enet} - \beta_0\|^2}{\eta^2/4} \]
\[ \leq 16 \frac{\lambda^2\||\beta_0\|^2 + n^3pB + \lambda^2_2p + o(n^2)}{[n^2b + \lambda_2 + o(n^2)]^2\eta^2}, \tag{32} \]

where the second inequality is due to Theorem 2. Then we can also have

\[ \sum_{j \in A^c} P[|2\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^ng_i(\tilde{\beta}))| > \lambda^*_1\hat{w}_j, \hat{\eta} > \eta/2] \]

\[ \leq \sum_{j \in A^c} P[|2\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^ng_i(\tilde{\beta}))| > \lambda^*_1\hat{w}_j, \hat{\eta} > \eta/2, |\tilde{\beta}_{enet}| \leq M] \]
\[ + \sum_{j \in A^c} P[|\tilde{\beta}_{enet}| > M] \]
\[ \leq \sum_{j \in A^c} P[|2\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^ng_i(\tilde{\beta}))| > \lambda^*_1M^{-\gamma}, \hat{\eta} > \eta/2] \]
\[ + \sum_{j \in A^c} P[|\tilde{\beta}_{enet}| > M], \tag{33} \]

where \( M = \left( \frac{\lambda^*_1}{n\eta^2} \right)^{1/2}. \)

In (33) we consider the second term on the right hand side. Via Zou and Zhang (2009), and Theorem 2 here

\[ \sum_{j \in A^c} P[|\tilde{\beta}_{enet}| > M] \leq \frac{E\|\hat{\beta}_{enet} - \beta_0\|^2}{M^2} \]
\[ \leq 4 \frac{\lambda^2\||\beta_0\|^2 + 4n^3pB + \lambda^2_2p + o(n^2)}{[n^2b + \lambda_2 + o(n^2)]^2M^2}. \tag{34} \]

Next we can consider the first term on the right hand side of (33)

\[ \sum_{j \in A^c} P[|2\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^ng_i(\tilde{\beta}))| > \lambda^*_1M^{-\gamma}, \hat{\eta} > \eta/2] \]
\[ \leq \frac{4M^{2\gamma}}{\lambda^*_1} E\| \sum_{j \in A^c} [\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^ng_i(\tilde{\beta}))]^2I_{[\hat{\eta} \geq \eta/2]} \|. \tag{35} \]

So we try to simplify the term on the right hand side of (35). Now we evaluate

\[ \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^ng_i(\tilde{\beta}))|^2 \leq 2 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^ng_i(\beta_{A,0}))|^2 \]
\[ + 2 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_nG_n(\tilde{\beta} - \beta_{A,0})|^2, \tag{36} \]

where we have \( \tilde{\beta} \in (\beta_{A,0}, \tilde{\beta}), \) and

\[ g_i(\tilde{\beta}) = g_i(\beta_{A,0}) + \left[ \frac{\partial g_i(\tilde{\beta})}{\partial \beta} \right] (\tilde{\beta} - \beta_{A,0}). \]
Analyze each term in (36). Note that \( \tilde{\beta} \) is consistent if we go through the same steps as in Theorem 1. Then applying the Assumption 2ii (Uniform Law of Large Numbers)

\[
2 \sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^{n} g_i(\beta_{A,0}))|^2 \leq 2n^2\|G(\beta_{A,0})'W\sum_{i=1}^{n} g_i(\beta_{A,0})\|^2 + o_p(n^2). \tag{37}
\]

Then via (36)

\[
E[\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_n(\sum_{i=1}^{n} g_i(\beta_{A,0}))|^2] \leq n^3B + o_p(n^3), \tag{38}
\]

where we use \( \lim_{n \to \infty} (\sum_{i=1}^{n} g_i(\beta_{A,0}))'(\sum_{i=1}^{n} g_i(\beta_{A,0})) = n\Omega_\ast \), and

\[
\text{Eigmax}(\Sigma) \geq \text{Eigmax}(G(\beta_{A,0})'W\Omega_*WG(\beta_{A,0})),
\]

and \( \lim_{n \to \infty} \sum_{i=1}^{n} g_i(\beta_{A,0})g_i(\beta_{A,0})' = \Omega_\ast \). In the same manner we have

\[
\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_n\hat{G}_n(\tilde{\beta} - \beta_{A,0})|^2 \leq n^4|G(\beta_{A,0})'WG(\beta_{A,0})(\tilde{\beta} - \beta_{A,0})|^2 + o_p(n^4). \tag{39}
\]

Then designating \( \text{Eigmax}(G(\beta_{A,0})'WG(\beta_{A,0})) \leq B_1 \), where \( B_1 \) is a positive constant. So taking into account (39)

\[
\sum_{j \in A^c} |\hat{G}_{n,j}(\tilde{\beta})'W_n\hat{G}_n(\tilde{\beta} - \beta_{A,0})|^2 \leq n^4B_1^2\|\tilde{\beta} - \beta_{A,0}\|^2 + o_p(n^4). \tag{40}
\]