Aust. N. Z. J. Stat. **46**(3), 2004, 505–510

CORRECTED PROOF OF THE RESULT OF 'A PREDICTION ERROR PROPERTY OF THE LASSO ESTIMATOR AND ITS GENERALIZATION' BY HUANG (2003)

SAHARON ROSSET[1]* AND JI ZHU[2]

IBM T.J. Watson Research Center and University of Michigan

Summary

The Lasso achieves variance reduction and variable selection by solving an ℓ_1-regularized least squares problem. Huang (2003) claims that 'there always exists an interval of regularization parameter values such that the corresponding mean squared prediction error for the Lasso estimator is smaller than for the ordinary least square estimator'. This result is correct. However, its proof in Huang (2003) is not. This paper presents a corrected proof of the claim, which exposes and uses some interesting fundamental properties of the Lasso.

Key words: lasso; least squares estimator; piecewise linear; prediction error.

1. Introduction

The Lasso (Tibshirani, 1996) achieves variance reduction and variable selection by solving an ℓ_1-penalized least squares problem:

$$\hat{\boldsymbol{\beta}}(\gamma) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma \sum_j |\beta_j| \,.$$

For 'large' values of γ (defined in a relative sense), $\hat{\boldsymbol{\beta}}(\gamma)$ has many zero components. This was a major motivating factor for the Lasso, as it implies that the Lasso is appropriate for 'sparse' models, where ridge regression is unlikely to succeed, since it forces all coefficients to be non-zero (Friedman *et al.*, 2004). In some cases, the Lasso or equivalent procedures have provable optimality properties, such as in the case of wavelet shrinkage (Donoho *et al.*, 1995).

Recently, it has been shown (Osborne, Presnell & Turlach, 2000; Efron *et al.*, 2004) that the path of optimal solutions for the Lasso, $\{\hat{\boldsymbol{\beta}}(\gamma), 0 \le \gamma \le \infty\}$ is piecewise linear, and thus the Lasso can be solved efficiently for *all* values of γ using an incremental algorithm. A simple example to illustrate the piecewise linear property can be seen in Figure 1, where we show the Lasso optimal solution paths for a four-variable synthetic dataset. The plot shows the optimal Lasso solutions $\hat{\boldsymbol{\beta}}(\gamma)$ as a function of the ℓ_1 norm $\|\hat{\boldsymbol{\beta}}(\gamma)\|_1$. Each line represents one coefficient and gives its values at the optimal solution for the range of $\|\hat{\boldsymbol{\beta}}(\gamma)\|_1$ values. We observe that between points marked '+' the lines are straight, i.e. the coefficient paths are piecewise-linear, and the one-dimensional curve $\hat{\boldsymbol{\beta}}(\gamma)$ is piecewise linear in $\mathbb{R}^4$.

Received July 2003; revised July 2003; accepted November 2003.

*Author to whom correspondence should be addressed.

[1] Data Analytics Research Group, IBM T.J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA. e-mail: srosset@us.ibm.com

[2] Dept of Statistics, 439 West Hall, 550 East University, University of Michigan, Ann Arbor, MI 48109, USA.

Acknowledgments. The authors thank F. Huang for helpful comments and R. Tibshirani for pointing out Huang (2003) to us.

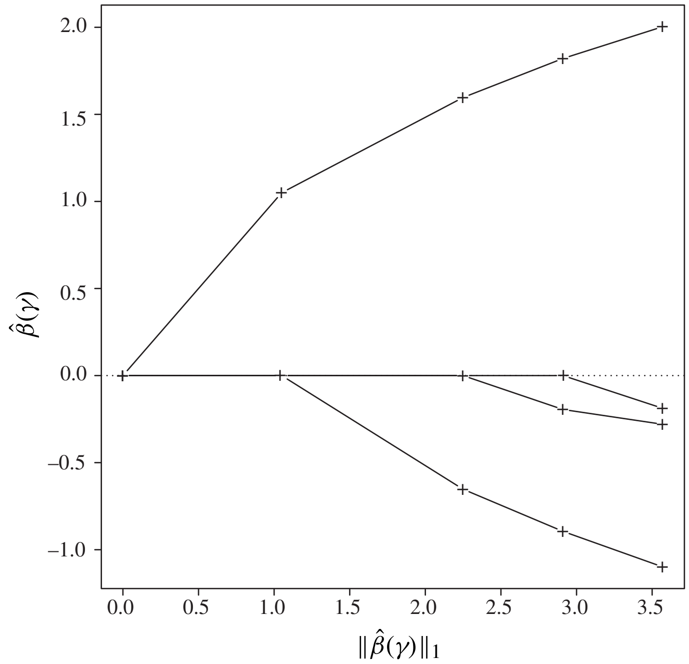

Figure 1. Piecewise linear solution paths for the Lasso on a simple four-variable example

Using this property, Efron *et al.* (2004) suggest the LAR-Lasso algorithm, which allows generation of the whole regularized solution path, $\{\hat{\boldsymbol{\beta}}(\gamma), 0 \le \gamma \le \infty\}$, for 'approximately' the computational cost of one least-squares calculation on the full dataset (the exact cost depends on some rather complicated properties of the regularized path, but the assumptions required to attain the above property are quite mild).

The paper by Huang (2003) describes another interesting and desirable property of the Lasso (Huang, 2003 p. 218, Theorem 2), shown here with modified notation:

There exists a value $\gamma_0 > 0$ *such that for* $\gamma \in (0, \gamma_0]$:

$$\mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\gamma)\|^2\big) < \mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\big)$$

That is, the mean squared prediction error of the Lasso estimator $\hat{\boldsymbol{\beta}}(\gamma)$ *is smaller than that of the least square estimator* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(0)$ *when the tuning parameter* γ *is small enough.*

This result is correct. However, its proof in that paper is not. Specifically, Theorem 3 therein is incorrect, and is fundamental in the proof of Theorem 2. In this note, we present a corrected version of Theorem 3, and a corrected proof of Theorem 2.

The notation throughout this paper is as in Huang (2003), except that we drop the (ℓ) and (0) superscripts for the Lasso solutions. As in Huang (2003) we assume throughout that the predictor matrix $\boldsymbol{X}$ is fixed. We denote by $\boldsymbol{y}$ the stochastic response vector used for fitting the model and by $\boldsymbol{y}_{\text{new}}$ a new, independent copy.

Unfortunately, our proof is not nearly as short and elegant as the proof using the incorrect Theorem 3. However, we believe it is of independent interest, as it exposes and uses some interesting fundamental properties of the Lasso — in particular Lemma 2 below and its proof which describes the 'piecewise linear' pieces of the Lasso path analytically.

2. Corrected results

2.1. Corrected Theorem 3 of Huang (2003 p. 219)

The original theorem in the paper reads, with modified notation:

There exists a value $\gamma_1 > 0$ *such that for* $\gamma \in [0, \gamma_1]$

$$\hat{\boldsymbol{\beta}}(\gamma) = \hat{\boldsymbol{\beta}} - \tfrac{1}{2}\gamma(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{s}(\hat{\boldsymbol{\beta}}) \qquad \textit{almost surely}.$$

Our corrected version is:

With probability 1 there exists a sample dependent value $\gamma_1(\boldsymbol{y}) > 0$ *such that for* $\gamma \in [0, \gamma_1(\boldsymbol{y})]$

$$\hat{\boldsymbol{\beta}}(\gamma) = \hat{\boldsymbol{\beta}} - \tfrac{1}{2}\gamma(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\mathbf{s}(\hat{\boldsymbol{\beta}}). \tag{1}$$

This corrected version does not assume there is a γ_1 which applies to (almost surely) all possible samples, but rather that it is sample dependent. The proof of Theorem 3 in Huang (2003 p. 226) actually proves this corrected version and requires no modification.

2.2. Corrected proof of main result

We consider only the right derivative of the expected error at $\gamma = 0$ and prove it is negative. This concludes an existence proof for γ_0 in Theorem 2.

Define $\tilde{\boldsymbol{\beta}}(\gamma)$ as the solution which just extends the last piece of the Lasso path backwards,

$$\tilde{\boldsymbol{\beta}}(\gamma) = \hat{\boldsymbol{\beta}} - \tfrac{1}{2}\gamma(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{s}(\hat{\boldsymbol{\beta}}),$$

and compare this to (1); $\tilde{\boldsymbol{\beta}}(\gamma) = \hat{\boldsymbol{\beta}}(\gamma)$ if and only if $\gamma \le \gamma_1(\boldsymbol{y})$.

Proof of Theorem 2. Consider the right derivative of the error at $\gamma = 0$:

$$\begin{aligned}\frac{\partial}{\partial\gamma_+}&\big(\mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\gamma)\|^2\big) - \mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\big)\big)_{\gamma=0}\\ &= \lim_{\gamma\searrow 0}\frac{\mathrm{E}(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\gamma)\|^2) - \mathrm{E}(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2)}{\gamma}.\end{aligned}$$

We re-write the numerator as

$$\begin{aligned}\mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\gamma)\|^2\big) - \mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\big) &= \big(\mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\gamma)\|^2\big) - \mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\big)\big)\\ &\quad - \big(\mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\gamma)\|^2\big) - \mathrm{E}\big(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\gamma)\|^2\big)\big).\end{aligned}$$

The proof given for Theorem 2 in Huang (2003 p. 227) proves that

$$\lim_{\gamma\searrow 0}\frac{\mathrm{E}(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\gamma)\|^2) - \mathrm{E}(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2)}{\gamma} < 0.$$

And so all we have left to prove is

$$\lim_{\gamma\searrow 0}\frac{\mathrm{E}(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\gamma)\|^2) - \mathrm{E}(\|\boldsymbol{y}_{\text{new}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\gamma)\|^2)}{\gamma} = 0. \tag{2}$$

We start by proving a couple of useful lemmas.

Lemma 1. $\Pr(\hat{\boldsymbol\beta}(\gamma) \neq \tilde{\boldsymbol\beta}(\gamma)) \to 0$ *as* $\gamma \to 0$.

Proof. This follows from the corrected Theorem 3, because if $\gamma < \gamma_1(\boldsymbol y)$ then $\hat{\boldsymbol\beta}(\gamma) = \tilde{\boldsymbol\beta}(\gamma)$.

Lemma 2. *There exists* $M > 0$ *such that* $\|\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\| \le M\gamma$ *for all* $\gamma > 0$. *So we can say that* $\hat{\boldsymbol\beta}(\gamma)$ *is 'linearly' close to* $\tilde{\boldsymbol\beta}(\gamma)$.

Proof. We first observe that the Lasso optimal solution path $\hat{\boldsymbol\beta}(\gamma)$ is piecewise linear as a function of γ (see Efron *et al.*, 2004 for details). One result of Efron *et al.* (2004) indicates that within each linear piece of the solution path, the set of predictor variables with non-zero coefficients is constant, i.e. $\mathcal{A} = \{j\colon\ \hat{\boldsymbol\beta}_j(\gamma) \neq 0, j = 1, \ldots, p\}$ does not change within each linear piece, and the derivative of $\hat{\boldsymbol\beta}(\gamma)$ with respect to γ is equal to $-\frac{1}{2}(\boldsymbol X_{\mathcal A}^{\mathsf T}\boldsymbol X_{\mathcal A})^{-1}\boldsymbol s_{\mathcal A}$, where $\boldsymbol X_{\mathcal A}$ is the corresponding sub-matrix of $\boldsymbol X$, and $\boldsymbol s_{\mathcal A}$ is the vector containing the signs of $\hat{\boldsymbol\beta}_{\mathcal A}(\gamma)$. Hence it implies that there exist $0 = \gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_m = \infty$ and $\mathcal A_1, \ldots, \mathcal A_m \subset \{1, \ldots, p\}$ and $\boldsymbol s_0 \in \{-1, +1\}^p$, $\boldsymbol s_1 \in \{-1, +1\}^{|\mathcal A_1|}, \ldots, \boldsymbol s_m \in \{-1, +1\}^{|\mathcal A_m|}$ such that if $\gamma_k \le \gamma < \gamma_{k+1}$ then (with a slight abuse of notation)

$$\hat{\boldsymbol\beta}(\gamma) = \hat{\boldsymbol\beta} - \tfrac{1}{2}\gamma_1(\boldsymbol X^{\mathsf T}\boldsymbol X)^{-1}\boldsymbol s_0 - \tfrac{1}{2}(\gamma_2 - \gamma_1)(\boldsymbol X_{\mathcal A_1}^{\mathsf T}\boldsymbol X_{\mathcal A_1})^{-1}\boldsymbol s_1 - \cdots - \tfrac{1}{2}(\gamma - \gamma_k)(\boldsymbol X_{\mathcal A_k}^{\mathsf T}\boldsymbol X_{\mathcal A_k})^{-1}\boldsymbol s_k\,.$$

Here $(\boldsymbol X_{\mathcal A}^{\mathsf T}\boldsymbol X_{\mathcal A})^{-1}\boldsymbol s$ is actually an $|\mathcal A| \times 1$ vector, rather than a $p \times 1$ vector. For notational simplicity, we have omitted the zero components, but this does not affect our claims below.

By the triangle inequality we then get

$$\|\hat{\boldsymbol\beta}(\gamma) - \hat{\boldsymbol\beta}\| \le \tfrac{1}{2}\gamma \max_j \|(\boldsymbol X_{\mathcal A_j}^{\mathsf T}\boldsymbol X_{\mathcal A_j})^{-1}\boldsymbol s_j\|\,,$$

which uses the specific data-dependent sequence of $\mathcal A_j, \boldsymbol s_j$, but we can easily generalize it to a data-independent result by observing that

$$\left|\left\{\boldsymbol s \in \{-1, +1\}^{|\mathcal A|}\colon\ \mathcal A \subset \{1, \ldots p\}\right\}\right| = 3^p - 1\,,$$

and thus we can find

$$M = \max_{(\mathcal A, \boldsymbol s)} \|(\boldsymbol X_{\mathcal A}^{\mathsf T}\boldsymbol X)^{-1}\boldsymbol s\|$$

and get the data-independent bound

$$\|\hat{\boldsymbol\beta}(\gamma) - \hat{\boldsymbol\beta}\| \le \tfrac{1}{2}\gamma M\,. \tag{3}$$

Next, we use the definition of $\tilde{\boldsymbol\beta}(\gamma)$ to bound

$$\|\tilde{\boldsymbol\beta}(\gamma) - \hat{\boldsymbol\beta}\| \le \tfrac{1}{2}\gamma \max_{\boldsymbol s} \|(\boldsymbol X^{\mathsf T}\boldsymbol X)^{-1}\mathbf{s}\| \le \tfrac{1}{2}\gamma M\,, \tag{4}$$

and combining (3) and (4) proves Lemma 2.

Now, consider the numerator of (2) again. We define

$$\Delta = \Delta(\gamma, \boldsymbol y, \boldsymbol y_{\text{new}}) = \|\boldsymbol y_{\text{new}} - \boldsymbol X\tilde{\boldsymbol\beta}(\gamma)\|^2 - \|\boldsymbol y_{\text{new}} - \boldsymbol X\hat{\boldsymbol\beta}(\gamma)\|^2\,,$$

and get

$$|\mathrm{E}(\Delta)| \le \left|\mathrm{E}\big(\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| < C)\big)\right| + \left|\mathrm{E}\big(\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| \ge C)\big)\right|.$$

We now analyse the two components of the right-hand side separately, via two additional lemmas.

Lemma 3.

$$\lim_{\gamma\searrow 0}\left|\mathrm{E}\Big(\frac{1}{\gamma}\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| < C\,)\Big)\right| = 0 \qquad \textit{for all } C\,.$$

Proof. We fix C. First we re-phrase the numerator:

$$\mathrm{E}\big(\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| < C\,)\big) = \mathrm{E}\Big(\big(2\boldsymbol{y}_{\mathrm{new}}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big) \\ - \big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big)^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) + \tilde{\boldsymbol\beta}(\gamma)\big)\big)\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| < C\,)\Big)\,. \quad (5)$$

The expectation is over the distribution of both $\boldsymbol y$ and $\boldsymbol{y}_{\mathrm{new}}$. All the $\boldsymbol\beta$ quantities we have depend on $\boldsymbol y$ only. So our next step is to remove dependence on $\boldsymbol y$, by bounding the expectation by its maximum

$$\big|\mathrm{E}\big(\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| < C\,)\big)\big| \le \Pr\big(\hat{\boldsymbol\beta}(\gamma) \ne \tilde{\boldsymbol\beta}(\gamma)\big) \max_{\boldsymbol y,\,\|\hat{\boldsymbol\beta}\|<C}\Big|2\mathrm{E}_{\boldsymbol{y}_{\mathrm{new}}}\big(\boldsymbol{y}_{\mathrm{new}}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big) \\ - \big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big)^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) + \tilde{\boldsymbol\beta}(\gamma)\big)\big)\Big|\,. \quad (6)$$

The next step is to bound the two expressions inside the maximum. Denote by λ the maximal absolute singular value of $\boldsymbol X$ (so λ^2 is the maximal eigenvalue of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$). Then Lemma 2 gives us

$$\big|\mathrm{E}_{\boldsymbol{y}_{\mathrm{new}}}\big(\boldsymbol{y}_{\mathrm{new}}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big)\big)\big| \le \|\mathrm{E}(\boldsymbol{y}_{\mathrm{new}})\|\lambda M\gamma\,, \quad (7)$$

$$\big|\big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big)^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) + \tilde{\boldsymbol\beta}(\gamma)\big)\big| \le \lambda^2 M\gamma\,(2C + \gamma M)\,, \quad (8)$$

and combining (6), (7) and (8) with Lemma 1 we get

$$\lim_{\gamma\searrow 0}\frac{|\mathrm{E}(\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\|<C\,))|}{\gamma} \le \lim_{\gamma\searrow 0}\Pr\big(\hat{\boldsymbol\beta}(\gamma) \ne \tilde{\boldsymbol\beta}(\gamma)\big)\big(2\|\mathrm{E}(\boldsymbol{y}_{\mathrm{new}})\|\lambda M + \lambda^2 M(2C + \gamma M)\big) = 0\,.$$

This proves Lemma 3.

Lemma 4.

$$\lim_{C\to\infty}\lim_{\gamma\searrow 0}\left|\mathrm{E}\Big(\frac{1}{\gamma}\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| \ge C\,)\Big)\right| = 0\,.$$

Proof. Using the same algebra as in (5) we can write

$$\left|\mathrm{E}\Big(\frac{1}{\gamma}\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| \ge C\,)\Big)\right| \le \big|\mathrm{E}\big(2\boldsymbol{y}_{\mathrm{new}}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big)\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| \ge C\,)\big)\big| \\ + \big|\mathrm{E}\big(\big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big)^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) + \tilde{\boldsymbol\beta}(\gamma)\big)\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| \ge C\,)\big)\big|\,. \quad (9)$$

The first expression is bounded as follows:

$$\big|\mathrm{E}\big(\boldsymbol{y}_{\mathrm{new}}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol\beta}(\gamma) - \tilde{\boldsymbol\beta}(\gamma)\big)\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| \ge C\,)\big)\big| \le \|\mathrm{E}(\boldsymbol{y}_{\mathrm{new}})\|\lambda M\gamma\Pr\big(\|\hat{\boldsymbol\beta}\| \ge C\,\big)\,. \quad (10)$$

Next, (i) $\|\hat{\boldsymbol\beta}(\gamma) + \tilde{\boldsymbol\beta}(\gamma)\| \le 2\|\hat{\boldsymbol\beta}\| + 2M\gamma$, by Lemma 2, and (ii) $\mathrm{E}(\|\hat{\boldsymbol\beta}\|\,\mathrm{I}\,(\|\hat{\boldsymbol\beta}\| \ge C\,)) \to 0$ as $C \to \infty$ for the second term, by the fact that $\mathrm{E}(\|\hat{\boldsymbol\beta}\|) \le \sum_j \mathrm{E}(|\hat\beta_j|) < \infty$ as $\mathrm{E}(\hat\beta_j) = \beta_j^0$, the true parameter. This also implies $\Pr(\|\hat{\boldsymbol\beta}\| \ge C\,) \to 0$.

Thus we can now write

$$\begin{aligned}\Big|\mathrm{E}\Big(\big(\hat{\boldsymbol{\beta}}(\gamma)-\tilde{\boldsymbol{\beta}}(\gamma)\big)^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big(\hat{\boldsymbol{\beta}}(\gamma)+\tilde{\boldsymbol{\beta}}(\gamma)\big)\,\mathrm{I}\,(\|\hat{\boldsymbol{\beta}}\|\geq C)\Big)\Big| \\ \leq 2\lambda^2 M\gamma\Big(\mathrm{E}\big(\|\hat{\boldsymbol{\beta}}\|\,\mathrm{I}\,(\|\hat{\boldsymbol{\beta}}\|\geq C)\big)+2\gamma M\Big)\,, \qquad (11)\end{aligned}$$

and putting (9), (10) and (11) together we get

$$\begin{aligned}&\lim_{\gamma\searrow 0}\frac{|\mathrm{E}(\Delta\,\mathrm{I}\,(\|\hat{\boldsymbol{\beta}}\|\geq C))|}{\gamma} \\ &\quad\leq 2\|\mathrm{E}(\boldsymbol{y}_{\mathrm{new}})\|\lambda M\ \Pr(\|\hat{\boldsymbol{\beta}}\|\geq C)+2\lambda^2 M\,\mathrm{E}\big(\|\hat{\boldsymbol{\beta}}\|\,\mathrm{I}\,(\|\hat{\boldsymbol{\beta}}\|\geq C)\big)\to 0\ \text{ as }\ C\to\infty\,,\end{aligned}$$

which concludes the proof of Lemma 4.

Putting Lemma 3 and Lemma 4 together completes our proof, since for any C it gives us an upper bound on (2), and this bound converges to 0 as $C \to \infty$.

References

DONOHO, D., JOHNSTONE, I., KERKYACHAIRAN, G. & PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with Discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 301–369.

EFRON, B., HASTIE, T., JOHNSTONE, I.M. & TIBSHIRANI, R. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32**, 407–499.

FRIEDMAN, J., HASTIE, T., ROSSET, S., TIBSHIRANI, R. & ZHU, J. (2004). Discussion of ‘Consistency in boosting’ by W. Jiang, G. Lugosi, N. Vayatis & T. Zhang. *Ann. Statist.* **32**, 102–107.

HUANG, F. (2003). A prediction error property of the lasso and its generalization. *Aust. N. Z. J. Stat.* **45**, 217–228.

OSBORNE, M.R., PRESNELL, B. & TURLACH, B. (2000). On the lasso and its dual. *J. Comput. Graph. Statist.* **9**, 319–337.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.