

A APPENDIX

A.1 CODE AVAILABILITY

Code is provided at the following link: <https://github.com/SimonSegert/specreg>.

A.2 FURTHER COMPARISON WITH HOCKING ET. AL.

The paper, similarly to us in spirit, introduces a class of biased linear estimators defined by a family of constrained optimization problems, and derive in this way the forms of Ridge, Stein Shrinkage, and PCR regressors. However, a closer inspection reveals that their setup is considerably different from ours.

For simplicity, we will make the assumption that $G = X^T X$ is diagonal, as this is sufficient to illustrate the differences. The form of estimators considered in Hocking et. al. is

$$\hat{\beta} = v \circ \Sigma \quad (14)$$

where as before, Σ is the vector of eigenvalues of G and \circ is Hadamard (element-wise) product. Immediately we see a difference that their setup has only d free variables, from the vector p , whereas we have $N \times p$ free variables in the matrix L .

By a bit of algebra,

$$\hat{\beta} = \Sigma \circ v = (\Sigma \circ v \oslash X^T Y) \circ X^T Y \quad (15)$$

$$= \text{diag}(\Sigma) \text{diag}(v) \text{diag}(X^T Y)^{-1} X^T Y \quad (16)$$

$$:= D X^T Y \quad (17)$$

where \oslash is element-wise division, and D is a diagonal matrix. Thus they effectively assume from the outset that $\hat{\beta} = D X^T Y$ for some diagonal matrix D , whereas we a priori allow $\hat{\beta} = L Y$ for $L \in \mathbb{R}^{N \times d}$ and derive that L is necessarily of this form.

They then show that each of the aforementioned estimators can be derived as the solution to an optimization of the form $\min_{v \in C} L(v)$, where L is a loss function and C is a constraint set, that depends on the estimator. The forms of the constraint sets are rather *ad-hoc* and the authors do not appear justify the form from more basic principles, or to relate them to each other in a meaningful way. By contrast, the forms of our constraints arise naturally from a standard family of matrix norms.

Finally, they do not precisely characterize the bias of each of the estimators, whereas our results characterize the estimators as being optimal for a fixed level of bias.

A.3 RULE OF THUMB FOR DEPTH VS. CURVATURE

In order to get insight into the tradeoff between curvature and depth, we consider a highly simplified setup which nonetheless captures some key features of selecting α through cross-validation. Let us assume that the function of average test error vs regularization strength is a parabola $f(x) = \kappa^2 x^2 / 2 + \mu$, for $-\delta \leq x \leq \delta$. Thus κ controls the curvature and μ controls the depth of the minimum. In real situations, f will likely have a more complex functional form, but we may imagine replacing it with its second-order Taylor expansion for the purpose of this analysis, which is likely to be a good approximation for sufficiently small δ .

We will model cross-validation as taking n iid samples $X_i \sim \text{Unif}(-\delta, \delta)$ and then estimating the minimum of f as $\hat{\mu} = \min_i (f(X_i))$. Note that in a real cross-validation setup there is the additional complication that we cannot perfectly observe the value $f(x)$ due to finite-sample variability; we do not model this effect here.

What is the distribution of $\hat{\mu}$? Well, for $\mu < y < \delta^2 \kappa^2 / 2 + \mu = f_{\max}$, we have

$$\mathbb{P}(f(X_i) > y) = \mathbb{P}(|X_i| > \sqrt{2(y - \mu)}/\kappa) = \frac{1}{2\delta} (2\delta - 2\sqrt{2(y - \mu)}/\kappa) \quad (18)$$

So,

$$\mathbb{P}(\hat{\mu} > y) = \mathbb{P}(f(X_i) > y)^n = (1 - \delta^{-1} \sqrt{2(y - \mu)}/\kappa)^n \quad (19)$$

The expected value is

$$\mathbb{E}\hat{\mu} = \int_0^\infty \mathbb{P}(\mu > y) dy = \mu + \int_\mu^{\delta^2 \kappa^2 / 2 + \mu} (1 - \sqrt{(2(y - \mu)\kappa^{-2}\delta^{-2})^n}) dy \quad (20)$$

Letting $u = \sqrt{2(y - \mu)/(\kappa^2\delta^2)}$, the integral becomes

$$\mu + \kappa^2\delta^2 \int_0^1 (1 - u)^n u du = \mu + \kappa^2\delta^2 \text{Beta}(2, n + 1) = \mu + \frac{(\kappa\delta)^2}{(n + 1)(n + 2)} \quad (21)$$

In the large sample limit, the contribution of the curvature vanishes, and we can exactly find the minimum. However, for finite samples this is not the case, and it could be that for two different parabolas the true minima satisfy $\mu_1 < \mu_2$ while the estimated minima satisfy $\mathbb{E}\hat{\mu}_1 > \mathbb{E}\hat{\mu}_2$.

A.4 PROOF OF MAIN THEOREM

Let us first make the simple but important observation that a minimizer in Equation 3 actually exists (since the constraint set is non-compact this is not completely immediate). One way to see this is to note that the problem is of the form $\min_{x \in D} |x|^2$ for some closed set D , and problems of this form always have a minimizer (even if D is non-compact). Combined with the observation that the objective function is strictly convex, we conclude that the problem in Equation 3 has *exactly one* minimizer.

The basic strategy of the proof is now to derive certain properties of the minimizer, and add these properties as further constraints until we get something tractable.

Lemma 3. *The minimizer L_{opt} of the bias-constrained problem (Equation 3) takes the form QX^T for some $d \times d$ matrix Q*

Proof. Let X_\perp be any $N \times (N - d)$ matrix whose columns form a basis for the orthogonal complement of $\text{Colspace}(X)$. We may assume that the columns are orthonormal; $X_\perp^T X_\perp = I_{N-d}$. Note that the concatenated matrix $[X, X_\perp]$ is invertible. Therefore, we can express the optimum as $L = M \begin{pmatrix} X^T \\ X_\perp^T \end{pmatrix}$ for some $d \times N$ matrix M . Writing M in block form $M = [Q, Q_\perp]$ where $Q \in \mathbb{R}^{d \times d}$ and $Q_\perp \in \mathbb{R}^{d \times (N-d)}$, we have $L = QX^T + Q_\perp X_\perp^T$. Note that $X^T X_\perp = 0$; therefore

$$\|LL^T\|_F^2 = \|QX^T\|_F^2 + \|Q_\perp X_\perp^T\|_F^2 \quad (22)$$

$$= \|QX^T\|_F^2 + \text{Tr}(Q_\perp X_\perp^T X_\perp Q_\perp^T) \quad (23)$$

$$= \|QX^T\|_F^2 + \text{Tr}(Q_\perp Q_\perp^T) \quad (24)$$

$$= \|QX^T\|_F^2 + \|Q_\perp\|_F^2 \quad (25)$$

where $\|\cdot\|_F$ is the Frobenius norm. Similarly, the bias $LX - I = QX^T X - I$ does not depend on Q_\perp . Thus, any non-zero value of Q_\perp will strictly increase the value of the objective relative to setting $Q_\perp = 0$, without having any effect on the constraint. \square

Corollary 3.1. *The optimal solution to Equation 3 takes the form $L = (I + Q)G^{-1}X^T$ where*

$$Q = \text{argmin}_{Q' \in \mathbb{R}^{d \times d}} \text{Tr}(Q'G^{-1}Q'^T)/2 + \text{Tr}(Q'G^{-1}) + \alpha^{-1}\|Q'\|_p \quad (26)$$

and $\alpha \geq 0$ is determined by C .

Proof. By the Lemma, it is no loss of generality to assume that L has the indicated form. Now plug in to Equation 3 and simplify. The α term is just converting the constraint to a Lagrange multiplier. \square

By the discussion above, we conclude that there is *exactly one* minimum of Equation 26.

Proposition 3.1. *If Q is the solution to the equation 26 then Q is symmetric and commutes with G*

Before giving the proof, we first present a more technical matrix lemma.

Lemma 4. For any square matrix M and $p \geq 1$, $\|M_d\|_p \leq \|M\|_p$, where M_d is the diagonal part (i.e., the matrix with all non-diagonal entries set to zero).

Proof. Evidently the singular values of M_d coincide with the absolute values of the diagonal entries. By an inequality of Ky Fan (Fan, 1951),

$$\sum_{i \leq k} \sigma_i(M_d) \leq \sum_{i \leq k} \sigma_i(M) \quad (27)$$

for any $1 \leq k \leq d$, where σ_i denotes the singular values, ordered from largest to smallest. The lemma now follows from Schur convexity of the Euclidean p-norm. \square

Proof. (of proposition 3.1) Let $G^{-1} = UDU^T$ be the diagonalization, where U is an orthogonal matrix and D is diagonal. The objective is

$$Q = \operatorname{argmin}_Q \operatorname{Tr}(Q'UDU^T Q'^T)/2 + \operatorname{Tr}(Q'UDU^T) + \alpha^{-1} \|Q'\|_p \quad (28)$$

$$= \operatorname{argmin}_Q \operatorname{Tr}(U^T Q'UDU^T Q'^T U)/2 + \operatorname{Tr}(U^T Q'UD) + \alpha^{-1} \|U^T Q'U\|_p \quad (29)$$

$$UQU^T = \operatorname{argmin}_P \operatorname{Tr}(PDP^T)/2 + \operatorname{Tr}(PD) + \alpha^{-1} \|P\|_p \quad (30)$$

$$(31)$$

where we reparametrized as $P := U^T Q'U$. Evidently, the proposition will follow if we can show that the minimal P is diagonal. Let $P = P_d + P_{od}$ be the decomposition into diagonal and off-diagonal parts⁶. Plugging into the objective

$$\operatorname{Obj}(P) = \operatorname{Tr}((P_d + P_{od})D(P_d + P_{od}^T))/2 + \operatorname{Tr}(P_d D) + \operatorname{Tr}(P_{od} D) + \alpha^{-1} \|P\|_p \quad (32)$$

$$= \operatorname{Tr}(P_d D P_d)/2 + \operatorname{Tr}(P_d D) + \operatorname{Tr}(P_{od} D P_{od}^T)/2 + \operatorname{Tr}(P_{od} D) + \alpha^{-1} \|P\|_p \quad (33)$$

$$+ \operatorname{Tr}(P_d D P_{od}^T) \quad (34)$$

Now, we note that in general if A is diagonal, and B is off-diagonal, then $\operatorname{Tr}(AB) = 0$. So the expression simplifies to

$$\operatorname{Tr}(P_d D P_d)/2 + \operatorname{Tr}(P_d D) + \operatorname{Tr}(P_{od} D P_{od}^T)/2 + \alpha^{-1} \|P\|_p \quad (35)$$

Now, the third term is non-negative because it is the trace of a PSD matrix; thus $\operatorname{Obj}(P) \geq \operatorname{Tr}(P_d D P_d)/2 + \operatorname{Tr}(P_d D) + \alpha^{-1} \|P\|_p$. By Lemma 4, $\|P\|_p \geq \|P_d\|_p$, and therefore

$$\operatorname{Obj}(P) \geq \operatorname{Obj}(P_d) \quad (36)$$

However, P was assumed to be the minimum, which implies by strict convexity that $P = P_d$ \square

By the above, we now know that the optimal Q must satisfy $Q = U \operatorname{diag}(q) U^T$, where q denotes the vector of eigenvalues and U is the matrix of eigenvectors of G^{-1} . Since $\hat{G}^{-1} = (1 + Q)G^{-1}$ in the notation of the theorem statement, we have shown the first two claims, namely that \hat{G} is symmetric and commutes with G .

By plugging into the objective in 26 and simplifying, we see that

$$q = \operatorname{argmin}_{x \in \mathbb{R}^d} \frac{1}{2} \sum_i x_i^2 / \sigma_i^2 + \sum_i x_i / \sigma_i^2 + \alpha^{-1} |x|_p \quad (37)$$

where σ_i^2 are the eigenvalues of G , and $|\cdot|_p$ denotes the Euclidean p norm of a vector.

To show the last claim that $\hat{G} - G$ is non-negative definite, it is enough to show that $0 \geq q_i \geq -1$, since $\lambda_i(\hat{G}) = \lambda_i(G)/(1 + q_i)$. It is easy to see that the eigenvectors of Q must be non-positive (since if any eigenvector is positive, then switching the sign will leave the first and third terms alone, while strictly decreasing the second term). To see the second inequality, we first add a suitable constant to the objective and rewrite it as $\frac{1}{2} \sum_i (x_i + 1)^2 / \sigma_i^2 + \alpha^{-1} |x|_p$. Supposing that $q_i < -1$ for some i , let us replace it with $q'_i = -2 - q_i$. Doing so does not change the first term (i.e.

⁶By definition an off-diagonal matrix is one with all zeros along the diagonal.

$(q_i+1)^2 = (q'_i+1)^2$, however it strictly decreases the second term, since $|q_i|' < |q_i|$ [7], contradicting the minimality of q [8].

Until now, we have not made any assumption about p except that $p \geq 1$. At this point, we separately analyze each of the three special cases $p = 1, 2, \infty$.

Before doing so, however, we first present the following well-known and elementary calculation, which we will employ several times in what follows:

Proposition 4.1. *Let $y, \tau > 0$ then*

$$\operatorname{argmin}_{x \in \mathbb{R}} \frac{1}{2}(x+y)^2 + \tau|x| = \min(\tau - y, 0) \quad (38)$$

Proof. It is clear that the optimal x cannot be positive, so we want to compute

$$\operatorname{argmin}_{x \leq 0} \frac{1}{2}(x+y)^2 - \tau x = \operatorname{argmin}_{x \leq 0} \frac{1}{2}(x+y-\tau)^2 + \frac{y^2 - (y-\tau)^2}{2} \quad (39)$$

where we completed the square. The second term does not depend on x and therefore has no effect on the argmin. Now the formula is immediate. If $\tau - y < 0$, then the minimum is plainly attained at $x = \tau - y < 0$. If $\tau - y > 0$, then the parabola is monotonically decreasing on the interval $(-\infty, 0)$, so the minimum is attained at $x = 0$. □

Nuclear case ($p=1$)

The objective [37] splits into a sum of separable one-dimensional problems:

$$\operatorname{argmin}_{x \in \mathbb{R}} \frac{1}{2}x^2/\sigma_i^2 + x/\sigma_i^2 + \alpha^{-1}|x| \quad (40)$$

$$\operatorname{argmin}_{x \in \mathbb{R}} \frac{1}{2}(x+1)^2 + \sigma_i^2\alpha^{-1}|x| \quad (41)$$

By Proposition 4.1

$$q_i = \min(\sigma_i^2/\alpha - 1, 0) \quad (42)$$

The eigenvalues of \hat{G} are thus given by

$$\lambda_i(\hat{G}) = \frac{\sigma_i^2}{1+q_i} = \frac{\sigma_i^2}{\min(\sigma_i^2/\alpha, 1)} = \max(\sigma_i^2, \alpha) \quad (43)$$

as claimed.

Frobenius case ($p=2$)

This is a simple calculus exercise and omitted.

Spectral case ($p=\infty$)

We know that there exists exactly one minimizer $q_{opt} \in \mathbb{R}^d$ of [37]. Fix some $C > \max(|q_{opt}|_\infty, \max_i(1 + \sigma_i^2/\alpha))$ and consider the problem

$$\min_{q: |q|_\infty \leq C} \frac{1}{2} \sum_i q_i^2/\sigma_i^2 + \sum_i q_i/\sigma_i^2 + \alpha^{-1}|q|_\infty \quad (44)$$

This clearly has the same minimum as the original unconstrained problem.

Note the following identity:

$$\alpha^{-1}|q|_\infty = \max_{y: |y|_1 \leq \alpha^{-1}} \langle q, y \rangle \quad (45)$$

⁷e.g., square both sides

⁸This argument doesn't quite go through if $p = \infty$, since in that case decreasing the magnitude of a component of q might not decrease the norm. But that's fine because we will derive the exact solution for $p = \infty$ shortly.

By Von Neumann’s minimax theorem, we can interchange the min and the max. So the optimal value of the objective is:

$$\max_{y:|y|_1 \leq \alpha^{-1}} \min_{q:|q|_\infty \leq C} \frac{1}{2} \sum_i q_i^2 / \sigma_i^2 + \sum_i q_i / \sigma_i^2 + \langle q, y \rangle \quad (46)$$

Now, the set $\{q : |q|_\infty \leq C\}$ is geometrically a product of intervals $[-C, C]^d$. Thus, we see that the inner objective splits into a sum of uncoupled 1-dimensional problems:

$$q_i = \operatorname{argmin}_{q_i \in [-C, C]} \frac{1}{2} q_i^2 / \sigma_i^2 + q_i / \sigma_i^2 + q_i y_i = \operatorname{argmin}_{q_i \in [-C, C]} \frac{1}{2\sigma_i^2} (q_i + 1 + \sigma_i^2 y_i)^2 - \frac{1}{2\sigma_i^2} (1 + \sigma_i^2 y_i)^2 \quad (47)$$

By assumption on C , $1 + \sigma_i^2 y_i \leq C$, and therefore for any fixed y_i , the minimum of the inner problem is attained at

$$q_i = -1 - \sigma_i^2 y_i \quad (48)$$

, with minimal value equal to $-\frac{1}{2\sigma_i^2} (1 + \sigma_i^2 y_i)^2 = -\frac{\sigma_i^2}{2} (\sigma_i^{-2} + y_i)^2$. Plugging back in to [46](#), we need to solve

$$\min_{y:|y|_1 \leq \alpha^{-1}} \frac{1}{2} \sum_i \sigma_i^2 (\sigma_i^{-2} + y_i)^2 \quad (49)$$

For this, we introduce a Lagrange multiplier λ , upon which the objective splits again into a sum of uncoupled 1d problems: $y_i = \operatorname{argmin}_{y \in \mathbb{R}} \frac{1}{2} \sigma_i^2 (\sigma_i^{-2} + y)^2 + \lambda |y|$.

Using Proposition [4.1](#) we derive the solution

$$y_i = \sigma_i^{-2} \min(\lambda - 1, 0) = \sigma_i^{-2} (\min(\lambda, 1) - 1) \quad (50)$$

where λ is chosen to satisfy the original constraint $\sum_i |y_i| \leq \alpha^{-1}$. Using the relation [48](#) between q_i and y_i

$$q_i = -1 - (\min(\lambda, 1) - 1) = -\min(\lambda, 1) \quad (51)$$

for the solution to the problem [44](#). Since we took C to be large enough to contain the solution to the unconstrained problem, we conclude that this is also the solution to the unconstrained problem (i.e. $C = \infty$). Since x_i are the eigenvalues of Q , we conclude that the eigenvalues of \hat{G} are

$$\lambda_i(\hat{G}) = \frac{\sigma_i^2}{1 + q_i} = \frac{\sigma_i^2}{1 - \min(\lambda, 1)} \quad (52)$$

Clearly the denominator lies in $[0, 1]$, therefore we recover the claimed form in which all eigenvalues of \hat{G} are obtained by scalar multiplication with some factor > 1 . Note that the case $\lambda > 1$ corresponds to multiplication by infinity, i.e. setting \hat{G}^{-1} (and thus $\hat{\beta}$) to zero.

A.5 PROOF OF [2.3](#)

We first give the definition of the Appel hypergeometric function F_1 for reference.

The function is typically defined as

$$F_1(\alpha, \beta, \beta', \gamma, x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha)_{m+n} (\beta)_{m+n} (\beta')_{n}}{m! n! (\gamma)_{m+n}} x^m y^n \quad (53)$$

Here $(\cdot)_m$ is the Pochhammer symbol. The series is absolutely convergent for $|x|, |y| < 1$, and arbitrary $\alpha, \beta, \beta', \gamma$. In [2.3](#) we may possibly need to evaluate it at some x outside of this range; we do this by appropriate analytic continuation, discussed further below.

To prove the formula, we note that the extremal cases follow straightforwardly from the integral formula. So in what follows we will assume that $\alpha \in [\lambda_-, \lambda_+]$.

We rewrite the integral

$$\frac{Err}{\lambda} - \beta^2 = \int_{\lambda_-}^{\lambda_+} \beta^2 x^2 f_\alpha(x)^{-2} - 2\beta^2 x f_\alpha^{-1}(x) + \sigma^2 x f_\alpha^{-2}(x) d\mu \quad (54)$$

$$= \int_{\lambda_-}^{\alpha} \beta^2 \alpha^{-2} x^2 - 2\beta^2 \alpha^{-1} x + \sigma^2 \alpha^{-2} x d\mu \quad (55)$$

$$+ \int_{\alpha}^{\lambda_+} \beta^2 - 2\beta^2 + \sigma^2 x^{-1} d\mu \quad (56)$$

$$= \beta^2 \alpha^{-2} \int_{\lambda_-}^{\alpha} x^2 d\mu + (\sigma^2 \alpha^{-2} - 2\beta^2 \alpha^{-1}) \int_{\lambda_-}^{\alpha} x d\mu \quad (57)$$

$$- \beta^2 (1 - F_\lambda(\alpha)) + \sigma^2 \left(\frac{1}{1 - \lambda} - \int_{\lambda_-}^{\alpha} x^{-1} d\mu \right) \quad (58)$$

$$= c + \frac{\beta^2}{\alpha^2} I(2, \alpha) + (\sigma^2 \alpha^{-2} - 2\beta^2 \alpha^{-1}) I(1, \alpha) + \beta^2 F_\lambda(\alpha) - \sigma^2 I(-1, \alpha) \quad (59)$$

where we defined $I(r, \alpha) = \int_{\lambda_-}^{\alpha} x^r d\mu$.

So we have reduced the theorem to just evaluating $I(r, \alpha)$ for $r \in \{-1, 1, 2\}$. Now, we plug in the form of the MP density, and make the variable substitution $u = \frac{x - \lambda_-}{\alpha - \lambda_-}$. Clearly then $u(\alpha - \lambda_-) + \lambda_- = x$.

$$\begin{aligned} 2\pi I(r, \alpha) &= \int_{\lambda_-}^{\alpha} x^{r-1} \sqrt{(\lambda_+ - x)(x - \lambda_-)} dx \\ &= \int_0^1 (u(\alpha - \lambda_-) + \lambda_-)^{r-1} \sqrt{\lambda_+ - \lambda_- - u(\alpha - \lambda_-)} \sqrt{u(\alpha - \lambda_-)} (\alpha - \lambda_-) du \\ &= (\alpha - \lambda_-)^{3/2} \lambda_-^{r-1} \sqrt{\lambda_+ - \lambda_-} \int_0^1 \sqrt{u} (1 + u \frac{\alpha - \lambda_-}{\lambda_-})^{r-1} \sqrt{1 - u \frac{\alpha - \lambda_-}{\lambda_+ - \lambda_-}} du \end{aligned}$$

We can now express this in the form given in the proposition by means of the following formula (Bailey, 1934):

$$F_1(\alpha, \beta, \beta', \gamma, x, y) = \frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma - \alpha)} \int_0^1 u^{\alpha-1} (1-u)^{\gamma-\alpha-1} (1-ux)^{-\beta} (1-uy)^{-\beta'} du \quad (60)$$

if $\alpha, \gamma - \alpha > 0$. We use this formula to define the analytic continuation in case $|x| > 1$ as can happen in formula 2.3 \square

A.6 PROOF OF GENERALIZATION FORMULAS 2.1 AND 2.4

We first consider 2.4 Given the setup in the main text, consider computing the test error for a fixed training/testing pair:

$$|X_{test}(\hat{G}^{-1} X_{tr}^T (X_{tr} \beta_0 + \epsilon) - X_{test} \beta_0)|^2 = |X_{test}(\hat{G}^{-1} G - I) \beta_0 + X_{test} \hat{G}^{-1} X_{tr}^T \epsilon|^2 \quad (61)$$

$$= |X_2 D S B \beta_0 + X_2 D S \hat{G}^{-1} X_{tr}^T \epsilon|^2 \quad (62)$$

where S is the diagonal matrix containing the noise values $\sqrt{s_i}$, and D is the diagonal matrix containing $\sqrt{\lambda_i}$. By rotational symmetry of the Frobenius norm, when we marginalize X_2 we get

$$|D S B \beta_0 + D S \hat{G}^{-1} X_{tr}^T \epsilon|^2 \quad (63)$$

And marginalizing over ϵ further yields

$$|D S B \beta_0|^2 + \sigma^2 \text{Tr}(D S \hat{G}^{-1} G \hat{G}^{-1} S D) \quad (64)$$

where the cross term vanishes because $\mathbb{E}\epsilon = 0$. Remembering all matrices are actually diagonal, we obtain

$$n * err = \sum_{i=1}^d s_i \lambda_i (\beta_0)_i^2 (\lambda_i / f_\alpha(\lambda_i) - 1)^2 + \sigma^2 \sum_i s_i \lambda_i (\lambda_i / f_\alpha(\lambda_i))^2 \quad (65)$$

Since s_i are independent of λ_i and have expectation 1, we can easily marginalize out, obtaining:

$$\sum_{i=1}^d \lambda_i (\beta_0)_i^2 (\lambda_i / f_\alpha(\lambda_i) - 1)^2 + \sigma^2 \sum_i \lambda_i (\lambda_i / f_\alpha(\lambda_i))^2 \quad (66)$$

At this point, the only variable that remains to marginalize over is λ . The result is:

$$d^{-1} |\beta_0|^2 \sum_i \mathbb{E}_{\lambda_1, \dots, \lambda_n} \lambda_i ((\lambda_i / f_\alpha(\lambda_i) - 1))^2 + \sigma^2 \mathbb{E}_{\lambda_1, \dots, \lambda_n} \sum_i \lambda_i^2 / f_\alpha(\lambda_i)^2 \quad (67)$$

where we used exchangeability of λ_i to bring out the factor of $|\beta_0|^2$. Now the proposition follows by taking the limit $n \rightarrow \infty$ and applying the law of large numbers. \square

The proof of [2.1](#) is very similar - as above, the idea is to use rotational symmetry to reduce the squared-error to a sum of the form $\mathbb{E}_{\lambda_1, \dots, \lambda_d} \sum_i g(\lambda_i)$. The main difference is that in this case the eigenvalues of G are no longer independent, but they are at least still exchangeable, so we can bring out the factor of $|\beta_0|^2$ in front of the sum like above. And since G now has a Wishart distribution, we can use the Marchenko-Pastur theorem ([Bai & Silverstein, 2010](#)) instead of the Law of large numbers to reduce the sum to the indicated integral form.

A.7 ESTIMATION OF MINIMA AND CURVATURES

To estimate the minimum and curvature of an error curve $Err(\alpha)$, we evaluate $Err(\alpha_i)$, $i = 1, \dots, n$, where where $n = 500$ and α_i are equally logarithmically spaced between 10^{-3} and 10^5 . The minimum is simply estimated as $\min_i Err(\alpha_i)$. To estimate the curvature at the minimum, we took the closest few points to the minimum and fit a quadratic function. That is, if $i_0 = \operatorname{argmin}_i Err(\alpha_i)$ then we fit a two-parameter linear model of the form

$$Err(\alpha_i) - Err(\alpha_{i_0}) \sim a(\alpha_i - \alpha_{i_0}) + b(\alpha_i - \alpha_{i_0})^2, i_0 - 5 \leq i \leq i_0 + 5 \quad (68)$$

with the estimated Hessian being $2b$.

A.8 RELATION BETWEEN α AND C IN THEOREM [2](#)

To express the relation between α and C , we consider two cases, corresponding to whether or not the constraint set contains the global minimizer of the objective $L \mapsto Tr(LL^T)$. The first case is when $C \geq \|I_d\|_p = d^{1/p}$. In this case, the optimum is evidently $L = 0_{d \times N}$, corresponding to $\alpha = \infty$.

In the case where $C < d^{1/p}$, $0_{d \times N}$ does not lie in the constraint set. Therefore, the optimum L lies on the boundary of the constraint set, i.e. $\|LX - I\|_p = C$. By plugging in the functional forms from Theorem [2](#) we obtain the following relations:

$$\sum_{\sigma_i < \alpha} 1 - \frac{\sigma_i}{\alpha} = C \quad (p = 1) \quad (69)$$

$$\sqrt{\sum_i \frac{\alpha^2}{(\sigma_i + \alpha)^2}} = C \quad (p = 2) \quad (70)$$

$$\alpha / (1 + \alpha) = C \quad (p = \infty) \quad (71)$$

$$(72)$$

where σ_i are the eigenvalues of $X^T X$.

A.9 GENERALIZATION ERROR FORMULAS FOR POWER-LAW SPECTRUM

Here we derive analytic expressions for the generalization error in the Diagonal matrix ensemble with power-law spectral density $d\nu(x)/dx = \gamma x^{\gamma-1}$, $0 < x < 1$.

We begin with the Nuclear case. In this case $f_\alpha(x) = \max(x, \alpha)$, therefore $\frac{x}{f_\alpha(x)}$ is either x/α or 1 depending on whether $x < \alpha$ or $x > \alpha$. Plugging in to Equation [12](#), we get

$$\gamma^{-1}\lambda^{-1}Err_1(\alpha) = \int_0^\alpha (\beta^2 x(1 - \alpha^{-1}x)^2 + \sigma^2 \alpha^{-2} x^2) x^{\gamma-1} dx \quad (73)$$

$$+ \int_\alpha^1 \sigma^2 x^{\gamma-1} dx \quad (74)$$

$$= \int_0^\alpha \beta^2 (x^\gamma - 2\alpha^{-1}x^{\gamma+1} + \alpha^{-2}x^{\gamma+2}) + \sigma^2 \alpha^{-2} x^{\gamma+1} dx \quad (75)$$

$$+ \int_\alpha^1 \sigma^2 x^{\gamma-1} dx \quad (76)$$

$$= \alpha^{\gamma+1} \frac{\beta^2}{\gamma+1} + \alpha^{\gamma+2} \frac{\sigma^2 \alpha^{-2} - 2\alpha^{-1}\beta^2}{\gamma+2} + \alpha^{\gamma+3} \frac{\alpha^{-2}\beta^2}{\gamma+3} \quad (77)$$

$$+ \frac{\sigma^2}{\gamma} (1 - \alpha^\gamma) \quad (78)$$

$$= \frac{\sigma^2}{\gamma} + \alpha^\gamma \sigma^2 \left(\frac{1}{\gamma+2} - \frac{1}{\gamma} \right) + \alpha^{\gamma+1} \beta^2 \left(\frac{1}{\gamma+1} - \frac{2}{\gamma+2} + \frac{1}{\gamma+3} \right) \quad (79)$$

$$= \frac{\sigma^2}{\gamma} - 2\alpha^\gamma \frac{\sigma^2}{\gamma(\gamma+2)} + 2\alpha^{\gamma+1} \frac{\beta^2}{(\gamma+1)(\gamma+2)(\gamma+3)} \quad (80)$$

The above holds for $\alpha < 1$. If $\alpha > 1$, then the formula becomes

$$\gamma^{-1}\lambda^{-1}Err_1(\alpha) = \int_0^1 \beta^2 (x^\gamma - 2\alpha^{-1}x^{\gamma+1} + \alpha^{-2}x^{\gamma+2}) + \sigma^2 \alpha^{-2} x^{\gamma+1} dx \quad (81)$$

$$= \frac{\beta^2}{\gamma+1} - 2\alpha^{-1} \frac{\beta^2}{\gamma+2} + \alpha^{-2} \left(\frac{\beta^2}{\gamma+3} + \frac{\sigma^2}{\gamma+2} \right) \quad (82)$$

$$(83)$$

Now consider the Ridge case. Here equation [12](#) becomes

$$\gamma^{-1}\lambda^{-1}Err_2(\alpha) = \int_0^1 \beta^2 x^\gamma \left(1 - \frac{x}{x+\alpha} \right)^2 + \sigma^2 \frac{x^2}{(x+\alpha)^2} dx \quad (84)$$

$$= \int_0^1 \beta^2 \frac{x^\gamma}{(\alpha^{-1}x+1)^2} + \sigma^2 \alpha^{-2} \frac{x^2}{(\alpha^{-1}x+1)^2} dx \quad (85)$$

$$(86)$$

Now, we make use of the standard formula for the Gauss hypergeometric function F :

$$Beta(b, c-b)F(a, b, c, z) = \int_0^1 x^{b-1} (1-x)^{c-b-1} (1-zx)^{-a} dx \quad (87)$$

which holds assuming that $c > b$ and the integral converges. In particular,

$$\int_0^1 \frac{x^\gamma}{(\alpha^{-1}x+1)^2} dx = Beta(\gamma+1, 1)F(2, \gamma+1, \gamma+2, -\alpha^{-1}) = F(2, \gamma+1, \gamma+2, -\alpha^{-1})/(\gamma+1) \quad (88)$$

and

$$\int_0^1 \frac{x^2}{(\alpha^{-1}x+1)^2} dx = Beta(3, 1)F(2, 3, 4, -\alpha^{-1}) = F(2, 3, 4, -\alpha^{-1})/3 \quad (89)$$

Therefore we get the formula

$$\gamma^{-1}\lambda^{-1}Err_2(\alpha) = \frac{\beta^2}{\gamma+1}F(2, \gamma+1, \gamma+2, -\alpha^{-1}) + \frac{\sigma^2}{3}\alpha^{-2}F(2, 3, 4, -\alpha^{-1}) \quad (90)$$

Finally, in the Spectral case, we have $x/f_\alpha(x) = 1/(1+\alpha)$, so the expression becomes

$$\begin{aligned} \gamma^{-1}\lambda^{-1}Err_\infty(\alpha) &= \int_0^1 \beta^2 x^\gamma (1 - \frac{1}{1+\alpha})^2 + \sigma^2 x^{\gamma-1} \frac{1}{(1+\alpha)^2} dx \\ &= \int_0^1 \beta^2 x^\gamma \frac{\alpha^2}{(1+\alpha)^2} + \sigma^2 x^{\gamma-1} \frac{1}{(1+\alpha)^2} dx \\ &= \frac{\beta^2 \alpha^2}{(\gamma+1)(1+\alpha)^2} + \frac{\sigma^2}{\gamma(1+\alpha)^2} \end{aligned}$$

A.10 EFFECT OF NUMBER OF α VALUES

As pointed out in the main text, the performance of the cross-validated models can depend on the number n of α values used for cross-validation. To do so, we follow the methodology used in [3.2.1](#), with the only difference being we use a smaller hyperparameter range $\sigma \in [.5, 2, 3.5]$, $\lambda = .5$, $\rho \in [0, .5, .9]$, and also vary the total number n of α values used in the cross validation $n \in [9, 15, 20, 30, 50]$. We keep the limits of the range of α values as before, and also maintain the equal logarithmic spacing.

We show the average error and win probability in Figure [7](#) and Figure [8](#) respectively. As per the discussion in Sections [2.2.3](#) and [5](#), we see that the Ridge often benefits drastically from increased number of α s, and can overtake the Nuclear for large number of α in cases when the Nuclear performs better for small number of α .

A.11 SPARSELY STRUCTURED DATA AND COMPARISON TO LASSO

We consider a variant of the setup in Section [3.2.1](#), in which the data is constructed to have sparse structure. In this case, when generating the ground truth coefficient vector β_0 , we select a set of indices $I \subset \{1, \dots, 10\}$, $|I| = 3$ at random, and generate β_0 as

$$(\beta_0)_i = N(0, 1), i \in I \quad (91)$$

$$(\beta_0)_i = N(0, 1)/10, i \notin I; \quad (92)$$

We also use the smaller hyperparameter ranges $\rho = 0$, $\lambda = .5$, $\sigma \in [.5, 1, 1.5, 2, 2.5, 3]$. Otherwise we follow the methodology of Section [3.2.1](#).

We also include Lasso in the set of considered models, since it is designed to deal with sparse coefficient vectors. Note, however, that Lasso is not a Linear Estimator in the sense defined in Section [2.1](#).

We show the average error and win probability in Figure [9](#) and Figure [10](#) respectively.

A.12 REAL DATA EXPERIMENTS

We evaluate the models on real (i.e., non-synthetic) data. We consider the well-known Diabetes dataset (N=442, d=10) and California housing dataset (N=20640, d=8), both available from the `sklearn.datasets` library.

To analyze the performance of the models on each dataset, we create random train-test splits in which the size of the training set is always set to 300, and the test set comprises the remaining observations. Each model is fit on the training set (including the regularization strength α , using the same cross-validation procedure as in Section [3.2.1](#)), and the mean-square-error is evaluated on the

⁹To see this, one can note that in the case of $d = 1$, the Lasso estimator has the well-known closed form $\beta_{Lasso} = \text{Sign}(\langle X, Y \rangle) \max(\frac{\langle X, Y \rangle}{|X|^2} - \alpha, 0)$, which is clearly not a linear function of Y

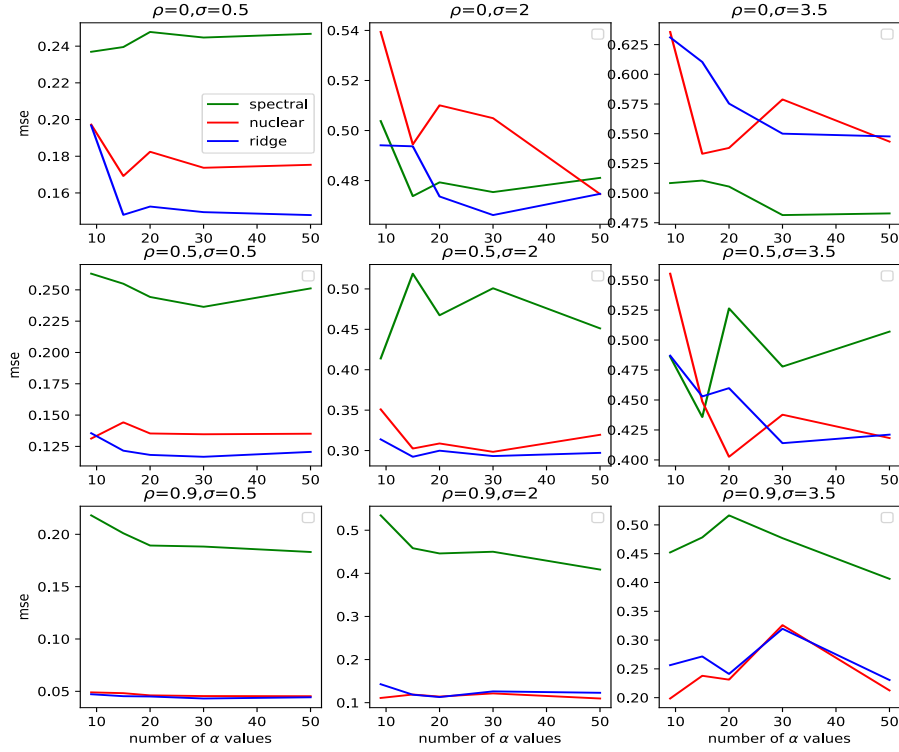


Figure 7: Average test error for varied number of regularization strength values α used in cross-validation. Each point represents an aggregate over 100 datasets.

test set. We use the same set of 9 α values as in Section 3.2.1 for all models. We constructed a total of 200 splits in this way, and evaluated each model on each split as before (so that we can evaluate each model’s probability of winning on a given split, as well as the average error over splits).

We show the average error and win probability in Figure 11 and Figure 12 respectively. We see a similar pattern as in Figure 5, with the Nuclear having a clear advantage over the other models in terms of win probability. The Nuclear also attains the lowest average MSE in the diabetes data, and is essentially tied with Ridge for lowest average MSE in the housing data.

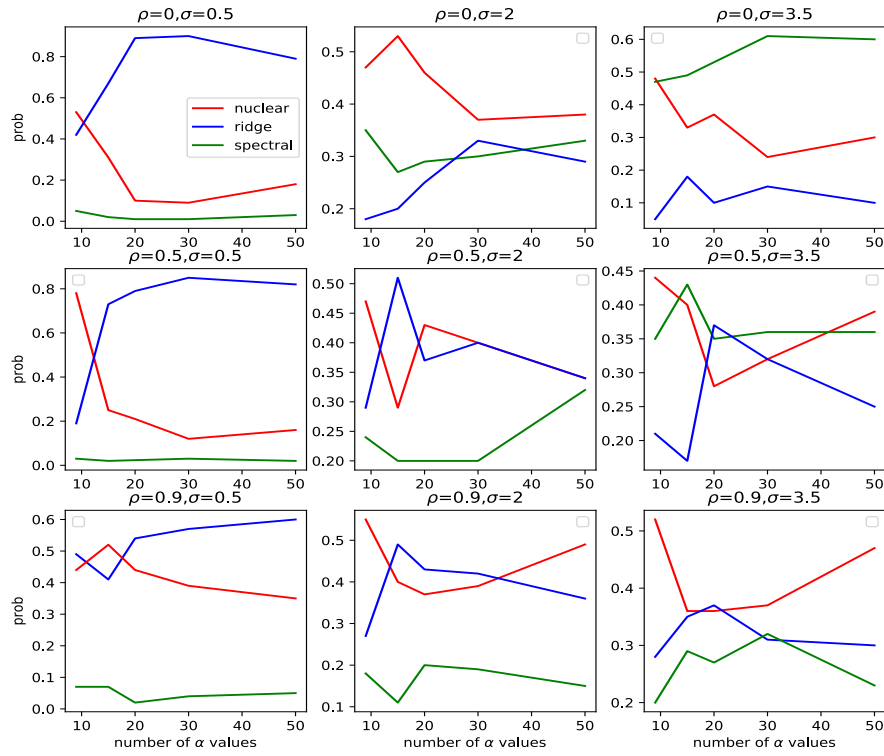


Figure 8: Same as Figure 7 except showing the probability that each model attains the lowest test error on a given dataset. Each point represents an aggregate over 100 datasets.

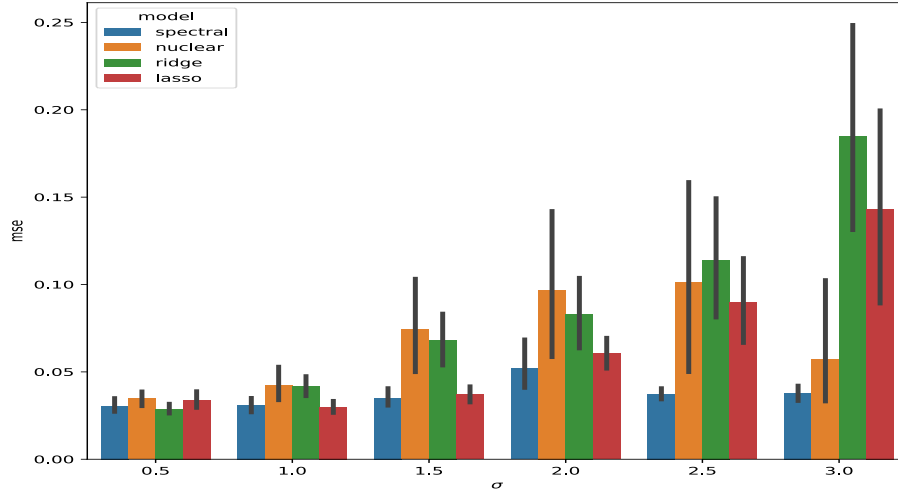


Figure 9: Average test error, with 95 percent confidence intervals, on Gaussian data with sparse ground truth coefficient vector. Each bar is an aggregate over 100 datasets.

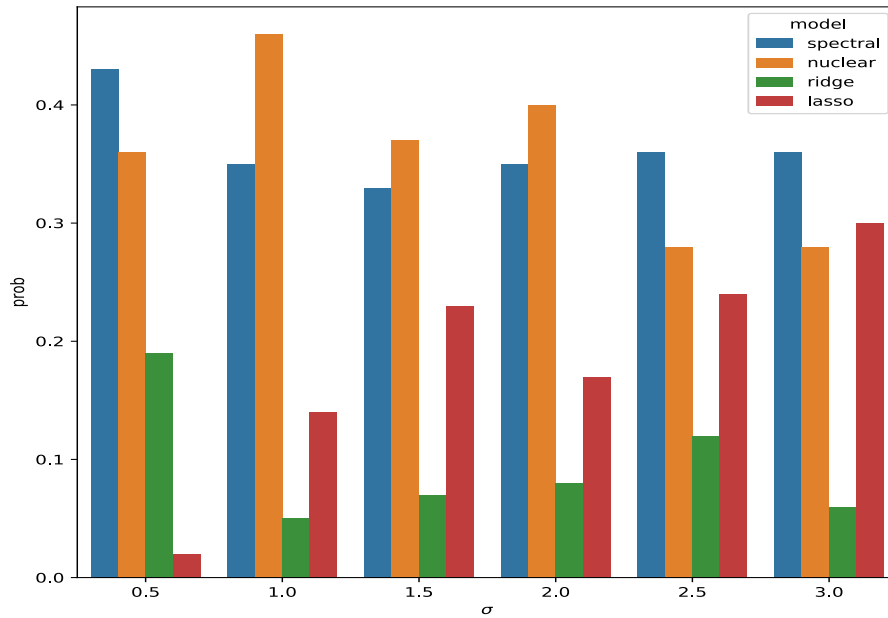


Figure 10: Same as Figure 9 except showing the probability that each model attains the lowest test error on a given dataset. Each bar is an aggregate over 100 datasets.

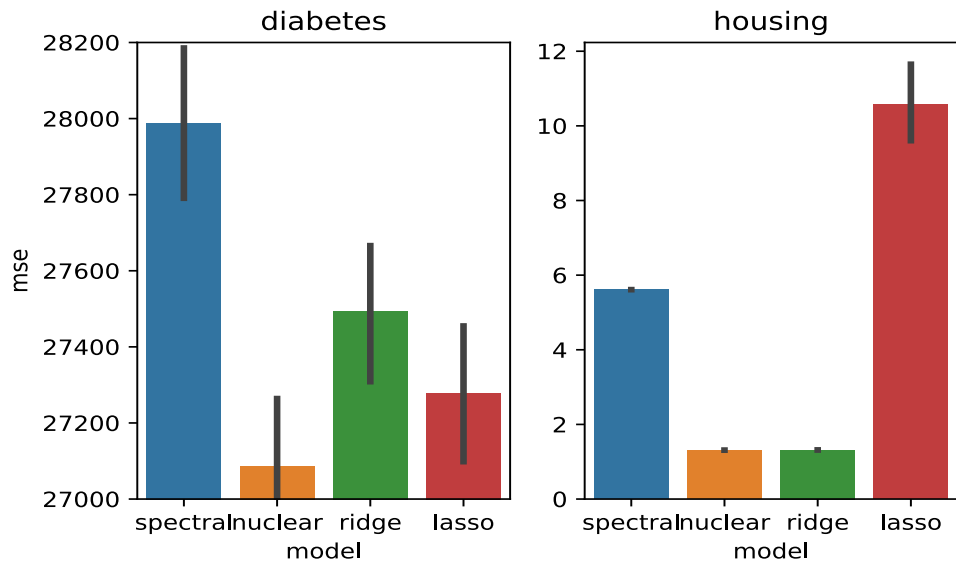


Figure 11: Average test error, with 95 percent confidence intervals, on a random train-test split, for the diabetes and California housing datasets. Each bar is an aggregate over 200 splits.

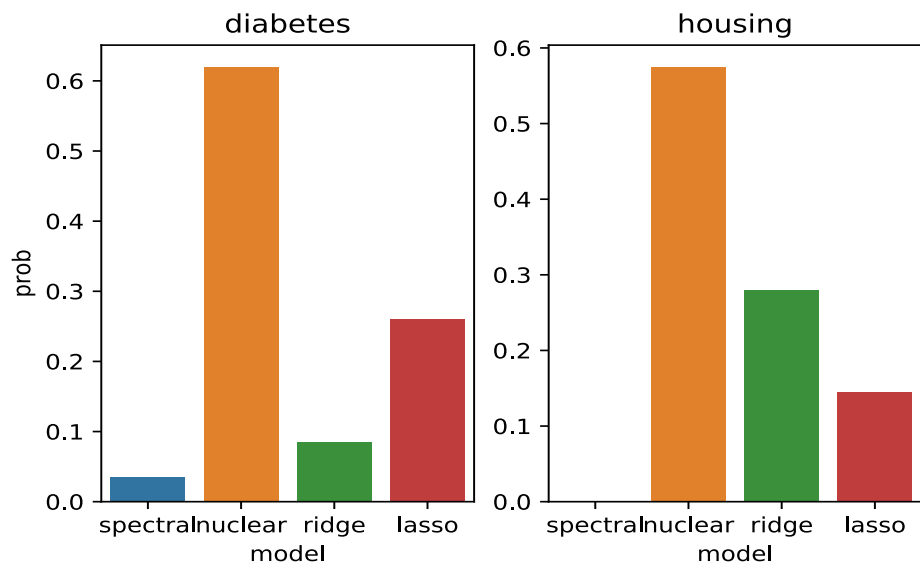


Figure 12: Same as Figure 11, except showing the probability that each model attains the lowest test error on a given split. Each bar is an aggregate over 200 splits.