Supplementary Notes: RRR (Ridge Risk Resolvents)*

Advanced Topics in Statistical Learning, Spring 2023

Pratik Patil

Note: The content of this supplement is pretty much taken shamelessly from Appendix C of Patil (2022) (continuing the trend from Ryan's supplementary notes on B-splines).

1 Calculus of asymptotic equivalents

We will use the language of asymptotic equivalents to express the limiting bias and variance resolvents of ridge regression. In this section, we provide a basic review of the definition of asymptotic equivalents and list several useful calculus rules that such equivalence obeys. For more details, see Dobriban and Sheng (2021); Patil (2022). The treatment below borrows heavily from the latter reference.

Definition 1 (Asymptotic equivalence)**.** Consider sequences $\{A_p\}_{p\geqslant 1}$ and $\{B_p\}_{p\geqslant 1}$ of (random or deterministic) matrices of growing dimension. We say that A_p and B_p are asymptotically equivalent and write $A_p \simeq B_p$ if $\lim_{p\to\infty} |\operatorname{tr}[C_p(A_p - B_p)]| = 0$ almost surely for any sequence C_p matrices with bounded trace norm such that $\limsup \|C_p\|_{\mathrm{tr}} < \infty$ as $p \to \infty$.

An observant reader will notice that Dobriban and Sheng (2021) use the notation $A_p \asymp B_p$ to denote asymptotic equivalence. We instead prefer to use the notation $A_p \simeq B_p$ for such equivalence to stress the fact that this equivalence is exact in the limit rather than up to constants as the "standard" use of the asymptotic notation $\asymp$ would hint at.

Lemma 1 (Calculus of asymptotic equivalents, Dobriban and Wager (2018), Dobriban and Sheng (2021))**.** *Let A_p, B_p, and C_p be sequences of (random or deterministic) matrices. The calculus of deterministic equivalents satisfy the following properties:*

1. *Equivalence: The relation $\simeq$ is an equivalence relation.*
2. *Sum: If $A_p \simeq B_p$ and $C_p \simeq D_p$, then $A_p + C_p \simeq B_p + D_p$.*
3. *Product: If A_p a sequence of matrices with bounded operator norms, i.e., $\|A_p\|_{\mathrm{op}} < \infty$, and $B_p \simeq C_p$, then $A_pB_p \simeq A_pC_p$.*
4. *Trace: If $A_p \simeq B_p$, then $\operatorname{tr}[A_p]/p - \operatorname{tr}[B_p]/p \to 0$ almost surely.*
5. *Differentiation: Suppose $f(z, A_p) \simeq g(z, B_p)$ where the entries of f and g are analytic functions in $z \in S$ and S is an open connected subset of $\mathbb{C}$. Suppose for any sequence C_p of deterministic matrices with bounded trace norm we have $|\operatorname{tr}[C_p(f(z, A_p) - g(z, B_p))]| \leqslant M$ for every p and $z \in S$. Then, we have $f'(z, A_p) \simeq g'(z, B_p)$ for every $z \in S$, where the derivatives are taken entry-wise with respect to z.*

The notion of *deterministic* equivalence is a special case of asymptotic equivalence where one of the sequences is a deterministic sequence. In the sequel, we will first record deterministic equivalent for the standard ridge resolvent in terms of the population covariance matrix Σ, and then derive deterministic equivalents for the bias and variance resolvents arising in the squared prediction risk of ridge regression.

A side comment: Some of you may be wondering what the term "resolvent" means. Resolvent formalism is a technique that uses complex-analytic machinery in the study of the spectrum of operators on Hilbert, Banach spaces, and more general spaces. You can quench your curiosity to know more about this topic at: https://en.wikipedia.org/wiki/Resolvent_formalism.

*Any connection to the eponymous 2022 movie (https://en.wikipedia.org/wiki/RRR_(film)) is purely coincidental! There is neither Rise nor Roar nor Revolt here!

Lemma 2 (Deterministic equivalent for basic ridge resolvent, adapted from Theorem 1 of Rubio and Mestre (2011); see also Theorem 3.1 of Dobriban and Sheng (2021)). *Suppose $x_i \in \mathbb{R}^p$, $1 \leqslant i \leqslant n$, are i.i.d. random vectors where each $x_i = z_i \Sigma^{1/2}$, where z_i contains i.i.d. entries z_{ij}, $1 \leqslant j \leqslant p$, with $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[|z_{ij}|^{8+\alpha}] \leqslant M_\alpha$ for some $\alpha > 0$ and $M_\alpha < \infty$, and $\Sigma \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix such that $0 \preceq \Sigma \preceq r_{\max} I_p$ for some constant (independent of p) $r_{\max} < \infty$. Let $X \in \mathbb{R}^{n \times p}$ the matrix with x_i, $1 \leqslant i \leqslant n$ as rows and $\widehat{\Sigma} \in \mathbb{R}^{p \times p}$ denote the random matrix $X^\top X/n$. Define $\gamma_n = p/n$. Then, for $z \in \mathbb{C}^{>0}$, as $n, p \to \infty$ such that $0 < \liminf \gamma_n \leqslant \limsup \gamma_n < \infty$, we have*

$$(\widehat{\Sigma} - z I_p)^{-1} \simeq (c(e(z;\gamma_n))\Sigma - z I_p)^{-1}, \tag{1}$$

where $c(e(z;\gamma_n))$ is defined as

$$c(e(z;\gamma_n)) = \frac{1}{1 + \gamma_n e(z;\gamma_n)}, \tag{2}$$

and $e(z;\gamma_n)$ is the unique solution in $\mathbb{C}^{>0}$ to the fixed-point equation

$$e(z;\gamma_n) = \operatorname{tr}[\Sigma(c(e(z;\gamma_n))\Sigma - z I_p)^{-1}]/p. \tag{3}$$

We note that in defining $e(\lambda;\gamma_n)$, it is also implicitly a parameterized by Σ. We suppress this dependence for notational simplicity, and only explicitly indicate dependence on z and γ_n that will be useful for our purposes.

A helpful corollary of Lemma 2 is the following result that considers the "scaled" ridge resolvent. The reason why such scaling helps is because in the limit as $\lambda \to 0^+$, the ridge resolvent itself may blow up, but the scaled resolvent is well-behaved.

Corollary 1 (Deterministic equivalent for scaled ridge resolvent). *Assume the setting of Lemma 2. Then, for $\lambda > 0$, as $n, p \to \infty$ such that $0 < \liminf \gamma_n \leqslant \limsup \gamma_n < \infty$, we have*

$$\lambda(\widehat{\Sigma} + \lambda I_p)^{-1} \simeq (v(-\lambda;\gamma_n)\Sigma + I_p)^{-1},$$

where $v(-\lambda;\gamma_n)$ is the unique solution to the fixed-point equation

$$\frac{1}{v(-\lambda;\gamma_n)} = \lambda + \gamma_n \operatorname{tr}[\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}]/p.$$

We remark that in moving from Lemma 2 to Corollary 1, we have switched from a complex z to a real λ. For more details of how this is done, see the proof of Corollary 1 in Appendix C of Patil (2022) or the proof of Theorem 5 in Hastie et al. (2022) (that uses Lemma 2.2 of Knowles and Yin (2017)).

2 Deterministic equivalents for ridge

Lemma 3 (Deterministic equivalents for bias and variance ridge resolvents). *Assume the setting of Lemma 2. Then, for $\lambda > 0$, as $n, p \to \infty$ with $0 < \liminf \gamma_n \leqslant \limsup \gamma_n < \infty$, the following asymptotic deterministic equivalences hold:*

1. *Variance resolvent of ridge regression:*

$$(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma \simeq \widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma^2, \tag{4}$$

   where $v(-\lambda;\gamma_n) \geqslant 0$ is the unique solution to the fixed-point equation

$$v(-\lambda;\gamma_n)^{-1} = \lambda + \gamma_n \operatorname{tr}[\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}]/p, \tag{5}$$

   and $\widetilde{v}(-\lambda;\gamma_n)$ is defined via $v(-\lambda;\gamma_n)$ by the equation

$$\widetilde{v}(-\lambda;\gamma_n)^{-1} = v(-\lambda;\gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}]/p. \tag{6}$$

2. Bias resolvent of ridge regression:

$$\lambda^2(\widehat{\Sigma}+\lambda I_p)^{-1}\Sigma(\widehat{\Sigma}+\lambda I_p)^{-1} \simeq (1+\widetilde{v}_b(-\lambda;\gamma_n))(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}\Sigma(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}, \quad (7)$$

where $v(-\lambda;\gamma_n)$ as defined in (10), and $\widetilde{v}_b(-\lambda;\gamma_n)$ is defined via $v(-\lambda;\gamma_n)$ by the equation

$$\widetilde{v}_b(-\lambda;\gamma_n) = \frac{\gamma_n \operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma+I_p)^{-2}]/p}{v(-\lambda;\gamma_n)^{-2}-\gamma_n \operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma+I_p)^{-2}]/p}. \quad (8)$$

Proof. The plan of attack for both the first and second parts is to use Corollary 1 as the starting point, and apply the calculus rules for asymptotic deterministic equivalents listed in Section 1 to manipulate into the desired equivalents.

Part 1. For the first part, observe that we can express the resolvent of interest (associated with the variance of ridge regression) as a derivative (with respect to λ) of a certain resolvent:

$$(\widehat{\Sigma}+\lambda I_p)^{-2}\widehat{\Sigma}\Sigma = (\widehat{\Sigma}+\lambda I_p)^{-1}\Sigma - \lambda(\widehat{\Sigma}+\lambda I_p)^{-2}\Sigma = \frac{\partial}{\partial\lambda}[\lambda(\widehat{\Sigma}+\lambda I_p)^{-1}\Sigma)]. \quad (9)$$

To find a deterministic equivalent for $(\widehat{\Sigma}+\lambda I_p)^{-2}\widehat{\Sigma}\Sigma$, it thus suffices to obtain a deterministic equivalent for the resolvent $\lambda(\widehat{\Sigma}+\lambda I_p)^{-1}\Sigma$ and take its derivative, thanks to the differentiation rule from Lemma 1 (5).

Starting with Corollary 1, we have

$$\lambda(\widehat{\Sigma}+\lambda I_p)^{-1} \simeq (v(-\lambda;\gamma_n)\Sigma+I_p)^{-1},$$

where $v(-\lambda;\gamma_n)$ is the unique solution to the fixed point equation

$$v(-\lambda;\gamma_n)^{-1} = \lambda + \gamma_n \operatorname{tr}[\Sigma(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}]/p. \quad (10)$$

Since Σ has bounded operator norm (uniformly in p), from Lemma 1 (3), we have

$$\lambda(\widehat{\Sigma}+\lambda I_p)^{-1}\Sigma \simeq (v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}\Sigma, \quad (11)$$

where $v(-\lambda;\gamma_n)$ is as defined by (10). It now remains to take the derivative of the right hand side of (11) with respect to λ. One can verify that the differentiation rule indeed applies in this case. (You should check this!) Taking derivative, we have

$$\frac{\partial}{\partial\lambda}[(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}\Sigma] = -\frac{\partial}{\partial\lambda}[v(-\lambda;\gamma_n)](v(-\lambda;\gamma_n)\Sigma+I_p)^{-2}\Sigma^2. \quad (12)$$

We can write - $\partial/\partial\lambda[v(-\lambda;\gamma_n)]$ in terms of $v(-\lambda;\gamma_n)$ by taking derivative of (10) with respect to λ and solving for - $\partial/\partial\lambda[v(-\lambda;\gamma_n)]$. Taking the derivative of (10) yields the following equation:

$$-\frac{\partial}{\partial\lambda}[v(-\lambda;\gamma_n)]v(-\lambda;\gamma_n)^{-2} = 1+\gamma_n - \frac{\partial}{\partial\lambda}[v(-\lambda;\gamma_n)]\operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma+I_p)^{-2}]/p. \quad (13)$$

Denoting - $\partial/\partial\lambda[v(-\lambda;\gamma_n)]$ by $\widetilde{v}(-\lambda;\gamma_n)$ and solving for $\widetilde{v}(-\lambda;\gamma_n)$ in (13), we get

$$\widetilde{v}(-\lambda;\gamma_n)^{-1} = v(-\lambda;\gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma+I_p)^{-2}]/p. \quad (14)$$

Combining (9), (12), and (14), the statement follows. This completes the proof of the first part.

Part 2. For the second part, observe that we can express the resolvent of interest (appearing in the bias of ridge regression) as a derivative of a certain parameterized resolvent at a fixed value of the parameter:

$$\begin{aligned}\lambda^2(\widehat{\Sigma}+\lambda I_p)^{-1}\Sigma(\widehat{\Sigma}+\lambda I_p)^{-1} &= \lambda^2(\widehat{\Sigma}+\lambda I_p+\lambda\rho\Sigma)^{-1}\Sigma(\widehat{\Sigma}+\lambda I_p+\lambda\rho\Sigma)^{-1}|_{\rho=0} \\ &= -\frac{\partial}{\partial\rho}[\lambda(\widehat{\Sigma}+\lambda I_p+\lambda\rho\Sigma)^{-1}]\Big|_{\rho=0}.\end{aligned} \quad (15)$$

It is worth remarking that in contrast to Part 1, we needed to introduce another parameter ρ for this part to appropriately pull out the matrix Σ in the middle. This trick has been used in the proof of Theorem 5 in Hastie et al. (2022) in the context of bias calculation for ridge regression. Our strategy henceforth will be to obtain a deterministic equivalent for the resolvent $\lambda(\widehat{\Sigma} + \lambda I_p + \lambda\rho\Sigma)^{-1}$, take its derivative with respect to ρ, and set $\rho = 0$. Towards that end, we first massage it to make it amenable for application of Lemma 2 as follows:

$$\begin{aligned}
\lambda\big(\widehat{\Sigma} + \lambda I_p + \lambda\rho\Sigma\big)^{-1} &= \lambda\big(\widehat{\Sigma} + \lambda(I_p + \rho\Sigma)\big)^{-1} \\
&= (I_p + \rho\Sigma)^{-1/2}\lambda\big((I_p + \rho\Sigma)^{-1/2}\widehat{\Sigma}(I_p + \rho\Sigma)^{-1/2} + \lambda I_p\big)^{-1}(I_p + \rho\Sigma)^{-1/2} \\
&= (I_p + \rho\Sigma)^{-1/2}\lambda\big(\widehat{\Sigma}_{\rho,\Sigma} + \lambda I_p\big)^{-1}(I_p + \rho\Sigma)^{-1/2}, \qquad (16)
\end{aligned}$$

where $\widehat{\Sigma}_{\rho,\Sigma} := \Sigma_{\rho,\Sigma}^{1/2}(Z^\top Z/n)\Sigma_{\rho,\Sigma}^{1/2}$ and $\Sigma_{\rho,\Sigma} := (I_p + \rho\Sigma)^{-1/2}\Sigma(I_p + \rho\Sigma)^{-1/2}$. We will now obtain a deterministic equivalent for $\lambda(\widehat{\Sigma}_{\rho,\Sigma} + \lambda I_p)^{-1}$, and use the product rule to arrive at the deterministic equivalent for $\lambda(\widehat{\Sigma} + \lambda I_p + \lambda\rho\Sigma)^{-1}$.

Using Corollary 1, we have

$$\lambda(\widehat{\Sigma}_{\rho,\Sigma} + \lambda I_p)^{-1} \simeq (v_b(-\lambda, \rho; \gamma_n)\Sigma_{\rho,\Sigma} + I_p)^{-1}, \qquad (17)$$

where $v_b(-\lambda, \rho; \gamma_n)$ is the unique solution to the fixed-point equation

$$v_b(-\lambda, \rho; \gamma_n)^{-1} = \lambda + \gamma_n \operatorname{tr}[\Sigma_{\rho,\Sigma}(v_b(-\lambda, \rho; \gamma_n)\Sigma_{\rho,\Sigma} + I_p)^{-1}]/p. \qquad (18)$$

Combining (16) with (17), and using the product rule from Lemma 1 (3) (which is applicable since $(I_p + \rho\Sigma)^{-1/2}$ is a deterministic matrix), we get

$$\begin{aligned}
\lambda(\widehat{\Sigma} + \lambda I_p + \lambda\rho\Sigma)^{-1} &= (I_p + \rho\Sigma)^{-1/2}\lambda(\widehat{\Sigma}_{\rho,\Sigma} + \lambda I_p)^{-1}(I_p + \rho\Sigma)^{-1/2} \\
&\simeq (I_p + \rho\Sigma)^{-1/2}(v_b(-\lambda, \rho; \gamma_n)\Sigma_{\rho,\Sigma} + I_p)^{-1}(I_p + \rho\Sigma)^{-1/2} \\
&= (I_p + \rho\Sigma)^{-1/2}(v_b(-\lambda, \rho; \gamma_n)(I_p + \rho\Sigma)^{-1/2}\Sigma(I_p + \rho\Sigma)^{-1/2} + I_p)^{-1}(I_p + \rho\Sigma)^{-1/2} \\
&= (v_b(-\lambda, \rho; \gamma_n)\Sigma + I_p + \rho\Sigma)^{-1}.
\end{aligned}$$

Similarly, the right hand side of the fixed-point equation (18) can be simplified by substituting back for $\Sigma_{\rho,\Sigma}$ to yield

$$\begin{aligned}
v_b(-\lambda, \rho; \gamma_n)^{-1} &= \lambda + \gamma_n \operatorname{tr}[(I_p + \rho\Sigma)^{-1/2}\Sigma(I_p + \rho\Sigma)^{-1/2}(v_b(-\lambda, \rho; \gamma_n)\Sigma_{\rho,\Sigma} + I_p)^{-1}]/p \\
&= \lambda + \gamma_n \operatorname{tr}[\Sigma(v_b(-\lambda, \rho; \gamma_n)(I_p + \rho\Sigma)^{1/2}\Sigma_{\rho,\Sigma}(I_p + \rho\Sigma)^{1/2} + (I_p + \rho\Sigma))^{-1}]/p \\
&= \lambda + \gamma_n \operatorname{tr}[\Sigma(v_b(-\lambda, \rho; \gamma_n)\Sigma + I_p + \rho\Sigma)^{-1}]/p. \qquad (19)
\end{aligned}$$

Finally, we will now use the differentiation rule from Lemma 1 (5) (with respect to ρ this time). It is easy verify that the differentiation rule applies in the neighborhood of $\rho = 0$. (Again, check this!) Taking derivative with respect to ρ, we get

$$\begin{aligned}
&-\frac{\partial}{\partial\rho}[(v_b(-\lambda, \rho; \gamma_n)\Sigma + I_p + \rho\Sigma)^{-1}] \\
&= (v_b(-\lambda, \rho; \gamma_n)\Sigma + I_p + \rho\Sigma)^{-1}\left(\frac{\partial}{\partial\rho}[v_b(-\lambda, \rho; \gamma_n)]\Sigma + \Sigma\right)(v_b(-\lambda, \rho; \gamma_n)\Sigma + I_p + \rho\Sigma)^{-1}.
\end{aligned} \qquad (20)$$

Setting $\rho = 0$ and observing that $v_b(-\lambda, 0; \gamma_n) = v(-\lambda; \gamma_n)$, where $v(-\lambda; \gamma_n)$ is as defined in (10), we have

$$\begin{aligned}
&\frac{\partial}{\partial\rho}[(v_b(-\lambda, \rho; \gamma_n)\Sigma + I_p + \rho\Sigma)^{-1}]\,\Big|_{\rho=0} \\
&\quad = (v(-\lambda; \gamma_n)\Sigma + I_p)^{-1}\left(\frac{\partial}{\partial\rho}[v_b(-\lambda, \rho; \gamma_n)]\,\Big|_{\rho=0}\,\Sigma + \Sigma\right)(v(-\lambda; \gamma_n)\Sigma + I_p)^{-1}.
\end{aligned} \qquad (21)$$

To obtain an equation for $\partial/\partial\rho[v_b(-\lambda,\rho;\gamma_n)]|_{\rho=0}$, we can differentiate the fixed-point equation (19) with respect to ρ to yield

$$\begin{aligned}
&-\frac{\partial}{\partial\rho}[v_b(-\lambda,\rho;\gamma_n)]v_b(-\lambda,\rho;\gamma_n)^{-2} \\
&\quad= -\gamma_n\frac{\partial}{\partial\rho}[v_b(-\lambda,\rho;\gamma_n)]\,\mathrm{tr}[\Sigma^2(v_b(-\lambda,\rho;\gamma_n)\Sigma+I_p+\rho\Sigma)^{-2}]/p \\
&\qquad\quad -\gamma_n\,\mathrm{tr}[\Sigma^2(v_b(-\lambda,\rho;\gamma_n)\Sigma+I_p+\rho\Sigma)^{-2}]/p.
\end{aligned}$$

Setting $\rho = 0$ in the equation above, and using the fact that $v_b(-\lambda,0;\gamma_n) = v(-\lambda;\gamma_n)$, and denoting $\partial/\partial\rho[v_b(-\lambda,\rho;\gamma_n)]|_{\rho=0}$ by $\widetilde{v}_b(-\lambda;\gamma_n)$, we get that

$$\widetilde{v}_b(-\lambda;\gamma_n) = \frac{\gamma_n\,\mathrm{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma+I_p)^{-2}]/p}{v(-\lambda;\gamma_n)^{-2}-\gamma_n\,\mathrm{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma+I_p)^{-2}]/p}. \tag{22}$$

Therefore, from (15) and (21), we finally have

$$\begin{aligned}
\lambda^2(\widehat{\Sigma}+\lambda I_p)^{-1}\Sigma(\widehat{\Sigma}+\lambda I_p)^{-1} &\simeq (v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}(\widetilde{v}_b(-\lambda;\gamma_n)\Sigma+\Sigma)(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1} \\
&= (1+\widetilde{v}_b(-\lambda;\gamma_n))(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}\Sigma(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1},
\end{aligned}$$

where $v(-\lambda;\gamma_n)$ is as defined in (10), and $\widetilde{v}_b(-\lambda;\gamma_n)$ is as defined in (22). This completes the proof of the second part. □

3 Deterministic equivalents for ridgeless

Lemma 4 (Deterministic equivalents for bias and variance ridgeless resolvents)**.** *Assume the setting of Lemma 3 with $\gamma_n \in (1,\infty)$. Then, the following deterministic equivalences hold:*

1. *Variance resolvent of ridgeless regression:*

$$\widehat{\Sigma}^+\Sigma \simeq \widetilde{v}(0;\gamma_n)(v(0;\gamma_n)\Sigma+I_p)^{-2}\Sigma^2, \tag{23}$$

where $v(0;\gamma_n)$ is the unique solution to the fixed-point equation

$$\gamma_n^{-1} = \mathrm{tr}[v(0;\gamma_n)\Sigma(v(0;\gamma_n)\Sigma+I_p)^{-1}]/p, \tag{24}$$

and $\widetilde{v}(0;\gamma_n)$ is defined through $v(0;\gamma_n)$ via

$$\widetilde{v}(0;\gamma_n) = \big(v(0;\gamma_n)^{-2}-\gamma_n\,\mathrm{tr}[\Sigma^2(v(0;\gamma_n)\Sigma+I_p)^{-2}]/p\big)^{-1}. \tag{25}$$

2. *Bias resolvent of ridgeless regression:*

$$(I_p-\widehat{\Sigma}^+\widehat{\Sigma})\Sigma(I_p-\widehat{\Sigma}^+\widehat{\Sigma}) \simeq (1+\widetilde{v}_b(0;\gamma_n))(v(0;\gamma_n)\Sigma+I_p)^{-1}\Sigma(v(0;\gamma_n)\Sigma+I_p)^{-1}, \tag{26}$$

where $v(0;\gamma_n)$ is as defined in (24), and $\widetilde{v}_b(0;\gamma_n)$ is defined via $v(0;\gamma_n)$ by

$$\widetilde{v}_b(0;\gamma_n) = \gamma_n\,\mathrm{tr}[\Sigma^2(v(0;\gamma_n)\Sigma+I_p)^{-2}]/p\cdot\big(v(0;\gamma_n)^{-2}-\gamma_n\,\mathrm{tr}[\Sigma^2(v(0;\gamma_n)\Sigma+I_p)^{-2}]/p\big)^{-1}. \tag{27}$$

Proof. The plan of attack for both the parts is to use the results of Lemma 3 and limiting arguments as $\lambda\to 0^+$. The results of Lemma 3 are pointwise in λ, but can be strengthened to be uniform in λ over a range that includes $\lambda = 0$ allowing one to take the limits of the deterministic equivalents obtained in Lemma 3 as $\lambda\to 0^+$.

Part 1. We will use the result in Part 1 of Lemma 3 as our starting point. Let $\Lambda := [0, \lambda_{\max}]$ where $\lambda_{\max} < \infty$, and let T be a matrix with bounded trace norm. Note that

$$|\operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma T]| \leqslant \|(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma\|_{\mathrm{op}} \operatorname{tr}[T] \leqslant C\|(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}} \leqslant C \tag{28}$$

for some constant $C < \infty$. Here, the last inequality follows because $s_i^2/(s_i^2 + \lambda)^2 \leqslant 1$ where s_i^2, $1 \leqslant i \leqslant p$, are the eigenvalues of $\widehat{\Sigma}$, and the operator norm Σ is assumed to be bounded. Consider the magnitude of the derivative (in λ) of the map $\lambda \mapsto \operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma T]$ given by

$$\left|\frac{\partial}{\partial\lambda} \operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma T]\right| = 2|\operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-3}\widehat{\Sigma}\Sigma T]|.$$

Following the argument in (28), for $\lambda \in \Lambda$, observe that

$$|\operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-3}\widehat{\Sigma}\Sigma T]| \leqslant \|(\widehat{\Sigma} + \lambda I_p)^{-3}\widehat{\Sigma}\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}} \operatorname{tr}[T] \leqslant C$$

for some constant $C < \infty$. Similarly, in the same interval $\operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma^2 T] \leqslant C$. In addition, it is easy to show that the map $\lambda \mapsto \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma T]$ is differentiable in λ and that the derivative for $\lambda \in \Lambda$ is bounded. Therefore, the family of functions $\operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma T] - \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma^2 T]$ forms an equicontinuous family in λ over $\lambda \in \Lambda$. Thus, the convergence in Part 1 of Lemma 3 is uniform in λ. We can now use the Moore-Osgood theorem to interchange the limits to obtain

$$\begin{aligned}
&\lim_{p\to\infty} \Big\{ \operatorname{tr}[\widehat{\Sigma}^+\Sigma T] - \operatorname{tr}[\widetilde{v}(0;\gamma_n)(v(0;\gamma_n)\Sigma + I_p)^{-2}\Sigma^2 T] \Big\} \\
&= \lim_{p\to\infty} \lim_{\lambda\to 0^+} \Big\{ \operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma T] - \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma^2 T)] \Big\} \\
&= \lim_{\lambda\to 0^+} \lim_{p\to\infty} \Big\{ \operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^{-2}\widehat{\Sigma}\Sigma T] - \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma^2 T)] \Big\} \\
&= 0.
\end{aligned}$$

In the first equality above, we used the fact that $\widehat{\Sigma}^+ = \widehat{\Sigma}^+\widehat{\Sigma}\widehat{\Sigma}^+ = \lim_{\lambda\to 0^+}(\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{-1}$, and that the functions $v(\cdot;\gamma_n)$ and $\widetilde{v}(\cdot;\gamma_n)$ are continuous (which is easy to verify). This provides the right hand side of (23). Similarly, the fixed-point equation (10) as $\lambda \to 0^+$ becomes

$$v(0;\gamma_n)^{-1} = \gamma_n \operatorname{tr}[\Sigma(v(0;\gamma_n)\Sigma + I_p)^{-1}]/p.$$

Moving $v(0;\gamma_n)$ to the other side (it follows easily that $v(0;\gamma_n) > 0$ for $\gamma_n \in (1,\infty)$, so we are safe in doing this), we arrive at the desired result.

Part 2. As done in Part 1, it is not difficult to show that over $\lambda \in \Lambda$ the family of functions $\operatorname{tr}[\lambda^2(\widehat{\Sigma} + \lambda I_p)^{-1}\Sigma(\widehat{\Sigma}+\lambda I_p)^{-1}T] - \operatorname{tr}[(1+\widetilde{v}_b(-\lambda;\gamma_n))(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}\Sigma(v(-\lambda;\gamma_n)\Sigma+I_p)^{-1}T]$ form an equicontinuous family. Therefore, the convergence in Part 2 of Lemma 3 is uniform in λ over Λ (that includes 0). Using the Moore-Osgood theorem to the interchange the limits, one has

$$\begin{aligned}
&\lim_{p\to\infty} \Big\{ \operatorname{tr}[(I_p - \widehat{\Sigma}^+\widehat{\Sigma})\Sigma(I_p - \widehat{\Sigma}^+\widehat{\Sigma})T] \\
&\qquad - \operatorname{tr}[(1 + \widetilde{v}_b(0;\gamma_n))(v(0;\gamma_n)\Sigma + I_p)^{-1}\Sigma(v(0;\gamma_n)\Sigma + I_p)^{-1}T] \Big\} \\
&= \lim_{p\to\infty} \lim_{\lambda\to 0^+} \Big\{ \operatorname{tr}[\lambda^2(\widehat{\Sigma} + \lambda I_p)^{-1}\Sigma(\widehat{\Sigma} + \lambda I_p)^{-1}T] \\
&\qquad\qquad - \operatorname{tr}[(1 + \widetilde{v}_b(-\lambda;\gamma_n))(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}T] \Big\} \\
&= \lim_{\lambda\to 0^+} \lim_{p\to\infty} \Big\{ \operatorname{tr}[\lambda^2(\widehat{\Sigma} + \lambda I_p)^{-1}\Sigma(\widehat{\Sigma} + \lambda I_p)^{-1}T] \\
&\qquad\qquad - \operatorname{tr}[(1 + \widetilde{v}_b(-\lambda;\gamma_n))(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}T] \Big\} \\
&= 0.
\end{aligned}$$

Now, both (25) and (27) follow by taking $\lambda \to 0^+$ in (7) and (8), respectively. □

References

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2): 918–943, 2021.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352, 2017.

Pratik Patil. *Facets of regularization in high-dimensional learning: Cross-validation, risk monotonization, and model complexity*. PhD thesis, Carnegie Mellon University, 2022.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & probability letters*, 81(5):592–602, 2011.