Supplement to "A Continuous-Time View of Early Stopping for Least Squares"

Alnur Ali Carnegie Mellon University
J. Zico Kolter Carnegie Mellon University
Ryan J. Tibshirani Carnegie Mellon University

This supplementary document contains additional details, proofs, and experiments for the paper "A Continuous-Time View of Early Stopping for Least Squares". All section, figure, and equation numbers in this document begin with the letter "S", to differentiate them from those appearing in the main paper (which appear without the prepended letter "S").

S.1 Proof of Lemma 3

Let $X^TX/n = VSV^T$ be an eigendecomposition of X^TX/n. Then we can rewrite the gradient descent iteration (2) as

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^T(y - X\beta^{(k-1)}) = (I - \epsilon VSV^T)\beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^T y.$$

Rotating by V^T, we get

$$\tilde{\beta}^{(k)} = (I - \epsilon S)\tilde{\beta}^{(k-1)} + \tilde{y},$$

where we let $\tilde{\beta}^{(j)} = V^T\beta^{(j)}$, $j = 1, 2, 3, \ldots$ and $\tilde{y} = (\epsilon/n)V^TX^Ty$. Unraveling the preceding display, we find that

$$\tilde{\beta}^{(k)} = (I - \epsilon S)^k\tilde{\beta}^{(0)} + \sum_{j=0}^{k-1}(I - \epsilon S)^j\tilde{y}.$$

Furthermore applying the assumption that the initial point $\beta^{(0)} = 0$ yields

$$\tilde{\beta}^{(k)} = \sum_{j=0}^{k-1}(I - \epsilon S)^j\tilde{y} = (\epsilon S)^{-1}(I - (I - \epsilon S)^k)\tilde{y},$$

with the second equality following after a short inductive argument.

Now notice that $\beta^{(k)} = V\tilde{\beta}^{(k)}$, since VV^T is the projection onto the row space of X, and $\beta^{(k)}$ lies in the row space. Rotating back to the original space then gives

$$\beta^{(k)} = V(\epsilon S)^{-1}(I - (I - \epsilon S)^k)\tilde{y} = \frac{1}{n}VS^{-1}(I - (I - \epsilon S)^k)V^TX^Ty.$$

Compare this to the solution of the optimization problem in Lemma 3, which is

$$(X^TX + nQ_k)^{-1}X^Ty = \frac{1}{n}(VSV^T + Q_k)^{-1}X^Ty.$$

Equating the last two displays, we see that we must have

$$VS^{-1}(I - (I - \epsilon S)^k)V^T = (VSV^T + Q_k)^{-1}.$$

Inverting both sides and rearranging, we get

$$Q_k = VS(I - (I - \epsilon S)^k)^{-1}V^T - VSV^T,$$

and an application of the matrix inversion lemma shows that $(I - (I - \epsilon S)^k)^{-1} = I + ((I - \epsilon S)^{-k} - I)^{-1}$, so

$$Q_k = VS((I - \epsilon S)^{-k} - I)^{-1}V^T,$$

as claimed in the lemma.

S.2 Proof of Lemma 4

Recall that Lemma 1 gives the gradient flow solution at time t, in (6). Compare this to the solution of the optimization problem in Lemma 4, which is

$$(X^TX + nQ_t)^{-1}X^Ty.$$

To equate these two, we see that we must have

$$(X^TX)^+(I - \exp(-tX^TX/n)) = (X^TX + nQ_t)^{-1},$$

i.e., writing $X^TX/n = VSV^T$ as an eigendecomposition of X^TX/n,

$$VS^+(I - \exp(-tS))V^T = (VSV^T + Q_t)^{-1}.$$

Inverting both sides and rearranging, we find that

$$Q_t = VS(I - \exp(-tS))^{-1}V^T - VSV^T,$$

which is as claimed in the lemma.

S.3 Proof of Lemma 5

For fixed β_0, and any estimator $\hat{\beta}$, recall the bias-variance decomposition

$$\text{Risk}(\hat{\beta}; \beta_0) = \|\mathbb{E}(\hat{\beta}) - \beta_0\|_2^2 + \text{tr}[\text{Cov}(\hat{\beta})].$$

For the gradient flow estimator in (6), we have

$$\begin{aligned}
\mathbb{E}[\hat{\beta}^{\text{gf}}(t)] &= (X^TX)^+(I - \exp(-tX^TX/n))X^TX\beta_0 \\
&= (X^TX)^+X^TX(I - \exp(-tX^TX/n))\beta_0 \\
&= (I - \exp(-tX^TX/n))\beta_0. \qquad \text{(S.1)}
\end{aligned}$$

In the second line, we used the fact that X^TX and $(I - \exp(-tX^TX/n))$ are simultaneously diagonalizable, and so they commute; in the third line, we used the fact that $(X^TX)^+X^TX = X^+X$ is the projection onto the row space of X, and the image of $I - \exp(-tX^TX/n)$ is already in the row space. Hence the bias is, abbreviating $\hat{\Sigma} = X^TX/n$,

$$\big\|\mathbb{E}[\hat{\beta}^{\text{gf}}(t)] - \beta_0\big\|_2^2 = \|\exp(-t\hat{\Sigma})\beta_0\|_2^2 = \sum_{i=1}^p |v_i^T\beta_0|^2 \exp(-2ts_i). \qquad \text{(S.2)}$$

As for the variance, we have

$$\begin{aligned}
\text{tr}\big(\text{Cov}[\hat{\beta}^{\text{gf}}(t)]\big) &= \sigma^2\text{tr}\big[(X^TX)^+(I - \exp(-t\hat{\Sigma}))(X^TX)(I - \exp(-t\hat{\Sigma}))(X^TX)^+\big] \\
&= \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}^+(I - \exp(-t\hat{\Sigma}))^2\big] \\
&= \frac{\sigma^2}{n}\sum_{i=1}^p \frac{(1 - \exp(-ts_i))^2}{s_i}, \qquad \text{(S.3)}
\end{aligned}$$

where in the second line we used the fact that $\hat{\Sigma}^+$ and $(I - \exp(-t\hat{\Sigma}))$ are simultaneously diagonalizable, and hence commute, and also the fact that $\hat{\Sigma}^+\hat{\Sigma}\hat{\Sigma}^+ = \hat{\Sigma}^+$. Putting together (S.2) and (S.3) proves the result in (11).

When β_0 follows the prior in (10), the variance (S.3) remains unchanged. The expectation of the bias (S.2) (over β_0) is

$$\mathbb{E}\big[\beta_0^T\exp(-2t\hat{\Sigma})\beta_0\big] = \text{tr}\big[\mathbb{E}(\beta_0\beta_0^T)\exp(-2t\hat{\Sigma})\big] = \frac{r^2}{p}\sum_{i=1}^p \exp(-2ts_i),$$

which leads to (12), after the appropriate definition of α.

S.4 Derivation of (13), (14)

As in the calculations in the last section, consider for the ridge estimator in (5),

$$\mathbb{E}[\hat{\beta}^{\text{ridge}}(\lambda)] = (X^TX + n\lambda I)^{-1}X^TX\beta_0 = (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}\beta_0, \tag{S.4}$$

where we have again abbreviated $\hat{\Sigma} = X^TX/n$. The bias is thus

$$\begin{aligned} \big\|\mathbb{E}[\hat{\beta}^{\text{ridge}}(\lambda)] - \beta_0\big\|_2^2 &= \big\|(\hat{\Sigma} + \lambda I)^{-1}(\hat{\Sigma} - I)\beta_0\big\|_2^2 \\ &= \big\|\lambda(\hat{\Sigma} + \lambda I)^{-1}\beta_0\big\|_2^2 \\ &= \sum_{i=1}^p |v_i^T\beta_0|^2 \frac{\lambda^2}{(s_i + \lambda)^2}, \end{aligned} \tag{S.5}$$

the second equality following after adding and subtracting λI to the second term in parentheses, and expanding. For the variance, we compute

$$\begin{aligned} \text{tr}\big(\text{Cov}[\beta^{\text{ridge}}(\lambda)]\big) &= \sigma^2\text{tr}\big[(X^TX + n\lambda I)^{-1}X^TX(X^TX + n\lambda I)^{-1}\big] \\ &= \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2}\big] \\ &= \frac{\sigma^2}{n}\sum_{i=1}^p \frac{s_i}{(s_i + \lambda)^2}, \end{aligned} \tag{S.6}$$

the second equality following by noting that $\hat{\Sigma}$ and $(\hat{\Sigma} + \lambda I)^{-1}$ are simultaneously diagonalizable, and therefore commute. Putting together (S.5) and (S.6) proves the result in (13). The Bayes result (14) follows by taking an expectation of the bias (S.5) (over β_0), just as in the last section for gradient flow.

S.5 Proof of Lemma 6

First, observe that for fixed β_0, and any estimator $\hat{\beta}$,

$$\text{Risk}^{\text{out}}(\hat{\beta}; \beta_0) = \mathbb{E}\|\hat{\beta} - \beta_0\|_\Sigma^2,$$

where $\|z\|_A^2 = z^TAz$. The bias-variance decomposition for out-of-sample prediction risk is hence

$$\text{Risk}^{\text{out}}(\hat{\beta}; \beta_0) = \|\mathbb{E}(\hat{\beta}) - \beta_0\|_\Sigma^2 + \text{tr}[\text{Cov}(\hat{\beta})\Sigma].$$

For gradient flow, we can compute the bias, from (S.1),

$$\big\|\mathbb{E}[\hat{\beta}^{\text{gf}}(t)] - \beta_0\big\|_\Sigma^2 = \|\exp(-t\hat{\Sigma})\beta_0\|_\Sigma^2 = \beta_0^T\exp(-t\hat{\Sigma})\Sigma\exp(-t\hat{\Sigma})\beta_0, \tag{S.7}$$

and likewise the variance,

$$\begin{aligned} \text{tr}\big(\text{Cov}[\beta^{\text{gf}}(t)]\big) &= \sigma^2\text{tr}\big[(X^TX)^+(I - \exp(-t\hat{\Sigma}))(X^TX)(I - \exp(-t\hat{\Sigma}))(X^TX)^+\Sigma\big] \\ &= \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}^+(I - \exp(-t\hat{\Sigma}))^2\Sigma\big]. \end{aligned} \tag{S.8}$$

Putting together (S.7) and (S.8) proves the result in (16). The Bayes result (17) follows by taking an expectation over the bias, as argued previously.

We note that the in-sample prediction risk is given by the same formulae except with Σ replaced by $\hat{\Sigma}$, which leads to

$$\begin{aligned} \text{Risk}^{\text{in}}(\hat{\beta}^{\text{gf}}(t); \beta_0) &= \beta_0^T\exp(-t\hat{\Sigma})\hat{\Sigma}\exp(-t\hat{\Sigma})\beta_0 + \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}^+(I - \exp(-t\hat{\Sigma}))^2\hat{\Sigma}\big] \\ &= \sum_{i=1}^p \bigg(|v_i^T\beta_0|^2 s_i\exp(-2ts_i) + \frac{\sigma^2}{n}(1 - \exp(-ts_i))^2\bigg), \end{aligned} \tag{S.9}$$

and

$$\begin{aligned}\text{Risk}^{\text{out}}(\hat{\beta}^{\text{gf}}(t)) &= \frac{\sigma^2}{n}\text{tr}\big[\alpha\exp(-2t\hat{\Sigma})\hat{\Sigma} + \hat{\Sigma}^{+}(I-\exp(-t\hat{\Sigma}))^2\hat{\Sigma}\big] \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{p}\big[\alpha s_i\exp(-2ts_i) + (1-\exp(-ts_i))^2\big]. \end{aligned} \tag{S.10}$$

S.6 Derivation of (18), (19)

For ridge, we can compute the bias, from (S.4),

$$\big\|\mathbb{E}[\hat{\beta}^{\text{ridge}}(\lambda)] - \beta_0\big\|_{\Sigma}^2 = \big\|\lambda(\hat{\Sigma}+\lambda I)^{-1}\beta_0\big\|_{\Sigma}^2 = \lambda^2\beta_0^T(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\beta_0, \tag{S.11}$$

and also the variance,

$$\begin{aligned}\text{tr}\big(\text{Cov}[\beta^{\text{ridge}}(\lambda)]\Sigma\big) &= \sigma^2\text{tr}\big[(X^TX+n\lambda I)^{-1}X^TX(X^TX+n\lambda I)^{-1}X^T\Sigma\big] \\ &= \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\Sigma\big]. \end{aligned} \tag{S.12}$$

Putting together (S.11) and (S.12) proves (18), and the Bayes result (19) follows by taking an expectation over the bias, as argued previously.

Again, we note that the in-sample prediction risk expressions is given by replacing Σ replaced by $\hat{\Sigma}$, yielding

$$\begin{aligned}\text{Risk}^{\text{in}}(\hat{\beta}^{\text{ridge}}(\lambda);\beta_0) &= \lambda^2\beta_0^T(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\beta_0 + \frac{\sigma^2}{n}\text{tr}\big[\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\big] \\ &= \sum_{i=1}^{p}\bigg(|v_i^T\beta_0|^2\frac{\lambda^2 s_i}{(s_i+\lambda)^2} + \frac{\sigma^2}{n}\frac{s_i^2}{(s_i+\lambda)^2}\bigg), \end{aligned} \tag{S.13}$$

and

$$\begin{aligned}\text{Risk}^{\text{in}}(\hat{\beta}^{\text{ridge}}(\lambda)) &= \frac{\sigma^2}{n}\text{tr}\big[\lambda^2\alpha(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma} + \hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\big] \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{\alpha\lambda^2 s_i + s_i^2}{(s_i+\lambda)^2}. \end{aligned} \tag{S.14}$$

S.7 Proof of Theorem 1, Part (c)

As we can see from comparing (11), (13) to (S.9), (S.13), the only difference in the latter in-sample prediction risk expressions is that each summand has been multiplied by s_i. Therefore the exact same relative bounds apply termwise, i.e., the arguments for part (a) apply here. The Bayes result again follows just by taking expectations.

S.8 Proof of Lemma 9

As in the proof of Lemma 8, because all matrices here are simultaneously diagonalizable, the claim reduces to one about eigenvalues, and it suffices to check that $e^{-2x} + (1-e^{-x})^2/x \leq 1.2147/(1+x)$ for all $x \geq 0$. Completing the square and simplifying,

$$\begin{aligned}e^{-2x} + \frac{(1-e^{-x})^2}{x} &= \frac{(1+x)e^{-2x} - 2e^{-x} + 1}{x} \\ &= \frac{(\sqrt{1+x}e^{-x} - \frac{1}{\sqrt{1+x}})^2}{x} + \frac{x}{1+x}. \end{aligned}$$

Now observe that, for any constant $C > 0$,

$$\begin{aligned}&\frac{(\sqrt{1+x}e^{-x} - \frac{1}{\sqrt{1+x}})^2}{x} + \frac{x}{1+x} \leq (1+C^2)\frac{1}{1+x} \\ &\iff |(1+x)e^{-x} - 1| \leq C\sqrt{x} \\ &\iff 1-(1+x)e^{-x} \leq C\sqrt{x}, \end{aligned} \tag{S.15}$$

the last line holding because the basic inequality $e^x \geq 1 + x$ implies that $e^{-x} \leq 1/(1+x)$, for $x > -1$. We see that for the above line to hold, we may take

$$C = \max_{x \geq 0} \big[1 - (1+x)e^{-x}\big]/\sqrt{x} = 0.4634,$$

which has been computed by numerical maximization, i.e., we find that the desired inequality (S.15) holds with $(1 + C^2) = 1.2147$.

S.9 Proof of Theorem 3, Part (b)

The lower bounds for the in-sample and out-of-sample prediction risks follow by the same arguments as in the estimation risk case (the ridge estimator here is the Bayes estimator in the case of a normal-normal likelihood-prior pair, and the risks here do not depend on the specific form of the likelihood and prior).

For the upper bounds, for in-sample prediction risk, we can see from comparing (12), (14) to (S.10), (S.14), the only difference in the latter expressions is that each summand has been multiplied by s_i, and hence the same relative bounds apply termwise, i.e., the arguments for part (a) carry over directly here.

And for out-of-sample prediction risk, the matrix inside the trace in (17) when $t = \alpha$ is

$$\alpha \exp(-2\alpha\hat{\Sigma}) + \hat{\Sigma}^+(I - \exp(-\alpha\hat{\Sigma}))^2,$$

and the matrix inside the trace in (19) when $\lambda = 1/\alpha$ is

$$1/\alpha(\hat{\Sigma} + (1/\alpha)I)^{-2} + \hat{\Sigma}(\hat{\Sigma} + (1/\alpha)I)^{-2} = \alpha(\alpha\hat{\Sigma} + I)^{-1}.$$

By Lemma 9, we have

$$\alpha \exp(-2\alpha\hat{\Sigma}) + \hat{\Sigma}^+(I - \exp(-\alpha\hat{\Sigma}))^2 \preceq 1.2147\alpha(\alpha\hat{\Sigma} + I)^{-1}.$$

Letting A, B denote the matrices on the left- and right-hand sides above, since $A \preceq B$ and $\Sigma \succeq 0$, it holds that $\mathrm{tr}(A\Sigma) \leq \mathrm{tr}(B\Sigma)$, which gives the desired result.

S.10 Proof of Theorem 6

Denote $\mathbb{C}_- = \{z \in \mathbb{C} : \mathrm{Im}(z) < 0\}$. By Lemma 2 in Ledoit and Peche (2011), under the conditions stated in the theorem, for each $z \in \mathbb{C}_-$, we have

$$\lim_{n,p \to \infty} \frac{1}{p} \mathrm{tr}\big[(\hat{\Sigma} + zI)^{-1}\Sigma\big] \to \theta(z) := \frac{1}{\gamma}\left(\frac{1}{1 - \gamma + \gamma z m(F_{H,\gamma})(-z)} - 1\right), \tag{S.16}$$

almost surely, where $m(F_{H,\gamma})$ denotes the *Stieltjes transform* of the empirical spectral distribution $F_{H,\gamma}$,

$$m(F_{H,\gamma})(z) = \int \frac{1}{u - z}\, dF_{H,\gamma}(u). \tag{S.17}$$

It is evident that (S.16) is helpful for understanding the Bayes prediction risk of ridge regression (19), where the resolvent functional $\mathrm{tr}[(\hat{\Sigma} + zI)^{-1}\Sigma]$ plays a prominent role.

For the Bayes prediction risk of gradient flow (17), the connection is less clear. However, the Laplace transform is the key link between (17) and (S.16). In particular, defining $g(t) = \exp(tA)$, it is a standard fact that its Laplace transform $\mathcal{L}(g)(z) = \int e^{-tz} g(t)\, dt$ (meaning elementwise integration) is in fact

$$\mathcal{L}(\exp(tA))(z) = (A - zI)^{-1}. \tag{S.18}$$

Using linearity (and invertibility) of the Laplace transform, this means

$$\exp(-2t\hat{\Sigma})\Sigma = \mathcal{L}^{-1}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big)(2t), \tag{S.19}$$

Therefore, we have for the bias term in (17),

$$\begin{aligned}\frac{\sigma^2\alpha}{n}\mathrm{tr}\big[\exp(-2t\hat{\Sigma})\Sigma\big] &= \frac{\sigma^2\alpha}{n}\mathrm{tr}\big[\mathcal{L}^{-1}\big((\hat{\Sigma}+zI)^{-1}\Sigma\big)(2t)\big] \\ &= \frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma}+zI)^{-1}\Sigma\big]\Big)(2t), \end{aligned} \tag{S.20}$$

where in the second line we again used linearity of the (inverse) Laplace transform. In what follows, we will show that we can commute the limit as $n, p \to \infty$ with the inverse Laplace transform in (S.20), allowing us to apply the Ledoit-Peche result (S.16), to derive an explicit form for the limiting bias. We first give a more explicit representation for the inverse Laplace transform in terms of a line integral in the complex plane

$$\frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma}+zI)^{-1}\Sigma\big]\Big)(2t) = \frac{\sigma^2 p\alpha}{n}\frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty}\mathrm{tr}\big[p^{-1}(\hat{\Sigma}+zI)^{-1}\Sigma\big]\exp(2tz)\,dz,$$

where $i = \sqrt{-1}$, and $a \in \mathbb{R}$ is chosen so that the line $[a-i\infty, a+i\infty]$ lies to the right of all singularities of the map $z \mapsto \mathrm{tr}[p^{-1}(\hat{\Sigma}+zI)^{-1}\Sigma]$. Thus, we may fix any $a > 0$, and reparametrize the integral above as

$$\begin{aligned}\frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma}+zI)^{-1}\Sigma\big]\Big)(2t) &= \frac{\sigma^2 p\alpha}{n}\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{tr}\big[p^{-1}\big(\hat{\Sigma}+(a+ib)I\big)^{-1}\Sigma\big]\exp(2t(a+ib))\,db \\ &= \frac{\sigma^2 p\alpha}{n}\frac{1}{\pi}\int_{-\infty}^{0}\mathrm{Re}\Big(\mathrm{tr}\big[p^{-1}\big(\hat{\Sigma}+(a+ib)I\big)^{-1}\Sigma\big]\exp(2t(a+ib))\Big)\,db. \end{aligned} \tag{S.21}$$

The second line can be explained as follows. A straightforward calculation, given in Lemma S.1, shows that the function $h_{n,p}(z) = \mathrm{tr}[p^{-1}\big(\hat{\Sigma}+zI\big)^{-1}\Sigma]\exp(2tz)$ satisfies $h_{n,p}(\bar{z}) = \overline{h_{n,p}(z)}$; another short calculation, deferred to Lemma S.2, shows that for any function with such a property, its integral over a vertical line in the complex plane reduces to the integral of twice its real part, over the line segment below the real axis. Now, noting that the integrand above satisfies

$$|h_{n,p}(z)| \leq \|(\hat{\Sigma}+zI)^{-1}\|_2\|\Sigma\|_2 \leq C_2/a,$$

for all $z \in [a-i\infty, a+i\infty]$, we can take limits in (S.21) and apply the dominated convergence theorem, to yield that almost surely,

$$\begin{aligned}&\lim_{n,p\to\infty}\frac{\sigma^2 p\alpha}{n}\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma}+zI)^{-1}\Sigma\big]\Big)(2t) \\ &\qquad= \sigma^2\gamma\alpha_0\frac{1}{\pi}\int_{-\infty}^{0}\lim_{n,p\to\infty}\mathrm{Re}\Big(\mathrm{tr}\big[p^{-1}\big(\hat{\Sigma}+(a+ib)I\big)^{-1}\Sigma\big]\exp(2t(a+ib))\Big)\,db \\ &\qquad= \sigma^2\gamma\alpha_0\frac{1}{\pi}\int_{-\infty}^{0}\mathrm{Re}\big(\theta(a+ib)\exp(2t(a+ib))\big)\,db \\ &\qquad= \sigma^2\gamma\alpha_0\frac{1}{2\pi}\int_{-\infty}^{\infty}\theta(a+ib)\exp(2t(a+ib))\,db \\ &\qquad= \sigma^2\gamma\alpha_0\mathcal{L}^{-1}(\theta)(2t). \end{aligned} \tag{S.22}$$

In the second equality, we used the Ledoit-Peche result (S.16), which applies because $a+ib \in \mathbb{C}_-$ for b in the range of integration. In the third and fourth equalities, we essentially reversed the arguments leading to (S.20), but with $h(z) = \theta(z)\exp(2tz)$ in place of $h_{n,p}$ (note that h must also satisfy $h(\bar{z}) = \overline{h(z)}$, as it is the pointwise limit of $h_{n,p}$, which has this same property).

As for the variance term in (17), consider differentiating with respect to t, to yield

$$\begin{aligned}\frac{d}{dt}\frac{\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}^+(I-\exp(-t\hat{\Sigma}))^2\Sigma\big] &= \frac{2\sigma^2}{n}\mathrm{tr}\big[\hat{\Sigma}^+\hat{\Sigma}(I-\exp(-t\hat{\Sigma}))\exp(-t\hat{\Sigma})\Sigma\big] \\ &= \frac{2\sigma^2}{n}\mathrm{tr}\big[(I-\exp(-t\hat{\Sigma}))\exp(-t\hat{\Sigma})\Sigma\big], \end{aligned}$$

with the second line following because the column space of $I - \exp(-t\hat{\Sigma})$ matches that of $\hat{\Sigma}$. The fundamental theorem of calculus then implies that the variance equals

$$\frac{2\sigma^2}{n} \int_0^t \mathrm{tr}\big[(\exp(-u\hat{\Sigma}) - \exp(-2u\hat{\Sigma}))\Sigma\big]\, du = \frac{2\sigma^2 p}{n} \int_0^t \Big[\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(u) - \mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2u)\Big]\, du,$$

where the equality is due to inverting the Laplace transform fact (S.18), as done in (S.19) for the bias. The same arugments for the bias now carry over here, to imply

$$\lim_{n,p\to\infty} \frac{2\sigma^2}{n} \int_0^t \Big[\mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(u) - \mathcal{L}^{-1}\Big(\mathrm{tr}\big[p^{-1}(\hat{\Sigma} + zI)^{-1}\Sigma\big]\Big)(2u)\Big]\, du = 2\sigma^2\gamma \int_0^t \big(\mathcal{L}^{-1}(\theta)(u) - \mathcal{L}^{-1}(\theta)(2u)\big)\, du. \quad \text{(S.23)}$$

Putting together (S.22) and (S.23) completes the proof.

S.11 Supporting Lemmas

Lemma S.1. *For any real matrices $A, B \succeq 0$ and $t \geq 0$, define*

$$f(z) = \mathrm{tr}\big[(A + zI)^{-1}B\big] \exp(2tz),$$

over $z \in \mathbb{C}_+ = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$. Then $f(\overline{z}) = \overline{f(z)}$.

Proof. First note that $\exp(2t\overline{z}) = \overline{\exp(2tz)}$ by Euler's formula. As the conjugate of a product is the product of conjugates, it suffices to show that $\mathrm{tr}[(A + \overline{z}I)^{-1}B] = \overline{\mathrm{tr}[(A + zI)^{-1}B]}$. To this end, denote $C_z = (A + zI)^{-1}$, and denote by C_z^* its adjoint (conjugate transpose). Note that $\overline{\mathrm{tr}(C_z B)} = \mathrm{tr}(C_z^* B)$; we will show that $C_z^* = C_{\overline{z}}$, which would then imply the desired result. Equivalent to $C_z^* = C_{\overline{z}}$ is $\langle C_z x, y\rangle = \langle x, C_{\overline{z}} y\rangle$ for all complex vectors x, y (where $\langle \cdot, \cdot \rangle$ denotes the standard inner product). Observe

$$\begin{aligned}
\langle C_z x, y\rangle &= \langle C_z x,\, (A + \overline{z}I)C_{\overline{z}} y\rangle \\
&= \langle (A + \overline{z}I)^* C_z x,\, C_{\overline{z}} y\rangle \\
&= \langle (A + zI)C_z x,\, C_{\overline{z}} y\rangle \\
&= \langle x, C_{\overline{z}} y\rangle,
\end{aligned}$$

which completes the proof. □

Lemma S.2. *If $f : \mathbb{C} \to \mathbb{C}$ satisfies $f(\overline{z}) = \overline{f(z)}$, then for any $a \in \mathbb{R}$,*

$$\int_{-\infty}^{\infty} f(a + ib)\, db = 2\int_{-\infty}^{0} \mathrm{Re}(f(a + ib))\, db.$$

Proof. The property $f(\overline{z}) = \overline{f(z)}$ means that $\mathrm{Re}(f(a - ib)) = \mathrm{Re}(f(a + ib))$, and $\mathrm{Im}(f(a - ib)) = -\mathrm{Im}(f(a + ib))$. Thus

$$\begin{aligned}
\int_{-\infty}^{\infty} f(a + ib)\, db &= \int_{-\infty}^{\infty} \mathrm{Re}(f(a + ib))\, db + i\int_{-\infty}^{\infty} \mathrm{Im}(f(a + ib))\, db \\
&= 2\int_{-\infty}^{0} \mathrm{Re}(f(a + ib))\, db + 0,
\end{aligned}$$

which completes the proof. □

S.12 Asymptotics for Ridge Regression

Under the conditions of Theorem 5, for each $\lambda \geq 0$, the Bayes risk (14) of ridge regression converges almost surely to

$$\sigma^2 \gamma \int \frac{\alpha_0 \lambda^2 + s}{(s+\lambda)^2} \, dF_{H,\gamma}. \tag{S.24}$$

This is simply an application of weak convergence of $F_{\hat{\Sigma}}$ to $F_{H,\gamma}$ (as argued the proof of Theorem 5), and can also be found in, e.g., Chapter 3 of Tulino and Verdu (2004).

The limiting Bayes prediction risk is a more difficult calculation. It is shown in Dobriban and Wager (2018) that, under the conditions of Theorem 6, for each $\lambda \geq 0$, the Bayes prediction risk (19) of ridge regression converges almost surely to

$$\sigma^2 \gamma \big[\theta(\lambda) + \lambda(1 - \alpha_0 \lambda)\theta'(\lambda)\big], \tag{S.25}$$

where $\theta(\lambda)$ is as defined in (S.16). The calculation (19) makes use of the Ledoit-Peche result (S.16), and Vitali's theorem (to assure the convergence of the derivative of the resolvent functional in (S.16)).

It is interesting to compare the limiting Bayes prediction risks (S.25) and (21). For concreteness, we can rewrite the latter as

$$\sigma^2 \gamma \bigg[\alpha_0 \mathcal{L}^{-1}(\theta)(2t) + 2 \int_0^t (\mathcal{L}^{-1}(\theta)(u) - \mathcal{L}^{-1}(\theta)(2u)) \, du\bigg]. \tag{S.26}$$

We see that (S.25) features θ and its derivative, while (S.26) features the inverse Laplace transform $\mathcal{L}^{-1}(\theta)$ and its antiderivative.

In fact, a similar structure can be observed by rewriting the limiting risks (S.24) and (20). By simply expanding $s = (s+\lambda) - \lambda$ in the numerator in (S.24), and using the definition of the Stieltjes transform (S.17), the limiting Bayes risk of ridge becomes

$$\sigma^2 \gamma \big[m(F_{H,\gamma})(-\lambda) - \lambda(1 - \alpha_0 \lambda) m(F_{H,\gamma})'(-\lambda)\big]. \tag{S.27}$$

By following arguments similar to the treatment of the variance term in the proof of Theorem 6, in Section S.10, the limiting Bayes risk of gradient flow becomes

$$\sigma^2 \gamma \bigg[\alpha_0 \mathcal{L}(f_{H,\gamma})(2t) + 2 \int_0^t (\mathcal{L}(f_{H,\gamma})(u) - \mathcal{L}(f_{H,\gamma})(2u)) \, du\bigg], \tag{S.28}$$

where $f_{H,\gamma} = dF_{H,\gamma}/ds$ denotes the density of the empirical spectral distribution $F_{H,\gamma}$, and $\mathcal{L}(f_{H,\gamma})$ its Laplace transform. We see (S.27) features $m(F_{H,\lambda})$ and its derivative, and (S.28) features $\mathcal{L}(f_{H,\gamma})$ and its antiderivative. But indeed $\mathcal{L}(\mathcal{L}(f_{H,\gamma}))(\lambda) = m(F_{H,\lambda})(-\lambda)$, since we can (in general) view the Stieltjes transform as an iterated Laplace transform. This creates a symmetric link between (S.27), (S.28) and (S.25), (S.26), where $m(F_{H,\gamma})(-\lambda)$ in the former plays the role of $\theta(\lambda)$ in the latter.

S.13 Additional Numerical Results

Here we show the complete set of numerical results comparing gradient flow and ridge regression. The setup is as described in Section 7. Figure S.1 shows the results for Gaussian features in the low-dimensional case ($n = 1000$, $p = 500$). The first row shows the estimation risk when $\Sigma = I$, with the left plot using $\lambda = 1/t$ calibration, and the right plot using ℓ_2 norm calibration (details on this calibration explained below). The second row shows the estimation risk when Σ has all off-diagonals equal to $\rho = 0.5$. The third row shows the prediction risk for the same Σ (n.b., the prediction risk when $\Sigma = I$ is the same as the estimation risk, so it is redundant to show both). The conclusions throughout are similar to that made in Section 7. Calibration by ℓ_2 norm gives extremely good agreement: the maximum ratio of gradient flow to ridge risk (over the entire path, in any of the three rows) is 1.0367. Calibration by $\lambda = 1/t$ is still quite good, but markedly worse: the maximum ratio of gradient flow to ridge risk (again over the entire path, in any of the three rows) is 1.4158.

Figures S.2 shows analogous results for Gaussian features in the high-dimensional case ($n = 500$, $p = 1000$). Figures S.3–S.6 show the results for Student t and Bernoulli features. The results are similar throughout: the maximum ratio of gradient flow to ridge risk, under ℓ_2 norm calibration (over the entire path, in any setting),

is 1.0371; the maximum ratio, under $\lambda = 1/t$ calibration (over the entire path, in any setting), is 1.4154. (One noticeable, but unremarkable difference between the settings is that the finite-sample risks seem to be converging slower to their asymptotic analogs in the case of t features. This is likely due to the fact that the tails here are very fat—they are as fat as possible for the t family, subject to the second moment being finite.)

It helps to give further details for a few of the calculations. For ℓ_2 norm calibration, note that we can compute the expected squared ℓ_2 norm of the ridge and gradient flow estimators under the data model (9) and prior (10):

$$\begin{aligned}
\mathbb{E}\|\hat{\beta}^{\text{ridge}}(\lambda)\|_2^2 &= \frac{1}{n}\Big(\text{tr}\big[\alpha(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}^2\big] + \text{tr}\big[(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{p}\frac{\alpha s_i^2 + s_i}{(s_i+\lambda)^2}, \\
\mathbb{E}\|\hat{\beta}^{\text{gf}}(t)\|_2^2 &= \frac{1}{n}\Big(\text{tr}\big[\alpha(I-\exp(-t\hat{\Sigma}))^2\big] + \text{tr}\big[(I-\exp(-t\hat{\Sigma}))^2\hat{\Sigma}^+\big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{p}\bigg(\alpha(1-\exp(-ts_i))^2 + \frac{(1-\exp(-ts_i))^2}{s_i}\bigg).
\end{aligned}$$

We thus calibrate according to the square root of the quantities above (this is what is plotted on the x-axis in the left columns of all the figures). The above expressions have the following limits under the asymptotic model studied in Theorem 5:

$$\begin{aligned}
\mathbb{E}\|\hat{\beta}^{\text{ridge}}(\lambda)\|_2^2 &\to \gamma\int\frac{\alpha_0 s^2+s}{(s+\lambda)^2}\,dF_{H,\gamma}(s), \\
\mathbb{E}\|\hat{\beta}^{\text{gf}}(t)\|_2^2 &\to \gamma\int\bigg(\alpha_0(1-\exp(-ts))^2 + \frac{(1-\exp(-ts))^2}{s}\bigg)\,dF_{H,\gamma}(s).
\end{aligned}$$

Furthermore, we note that when $\Sigma = I$, the empirical spectral distribution from Theorem 4 abbreviated as F_γ, sometimes called the *Marchenko-Pastur (MP) law* and has a closed form. For $\gamma \leq 1$, its density is

$$\frac{dF_\gamma(s)}{ds} = \frac{1}{2\pi\gamma s}\sqrt{(b-s)(s-a)},$$

and is supported on $[a,b]$, where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$. For $\gamma > 1$, the MP law F_γ has an additional point mass at zero of probability $1-1/\gamma$. This allows us to evaluate the integrals in (20), (S.24) via numerical integration, to compute limiting risks for gradient flow and ridge regression. (It also allows us to compute the integrals in the second to last display, to calibrate according to limiting ℓ_2 norms.)

References

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

Olivier Ledoit and Sandrine Peche. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1–2):233–264, 2011.

Antonia M. Tulino and Sergio Verdu. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1):1–182, 2004.

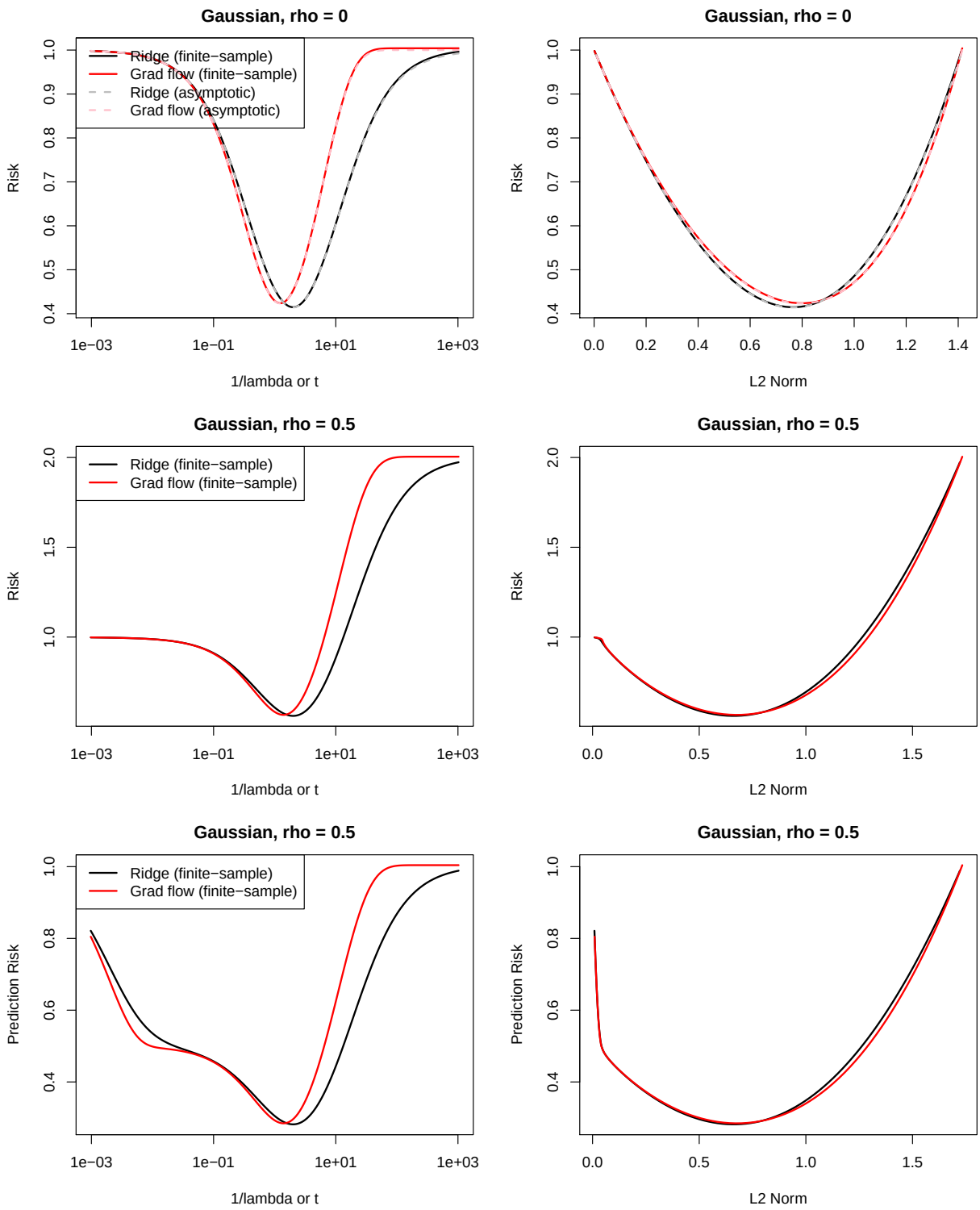

Figure S.1: *Gaussian features, with* $n = 1000$ *and* $p = 500$.

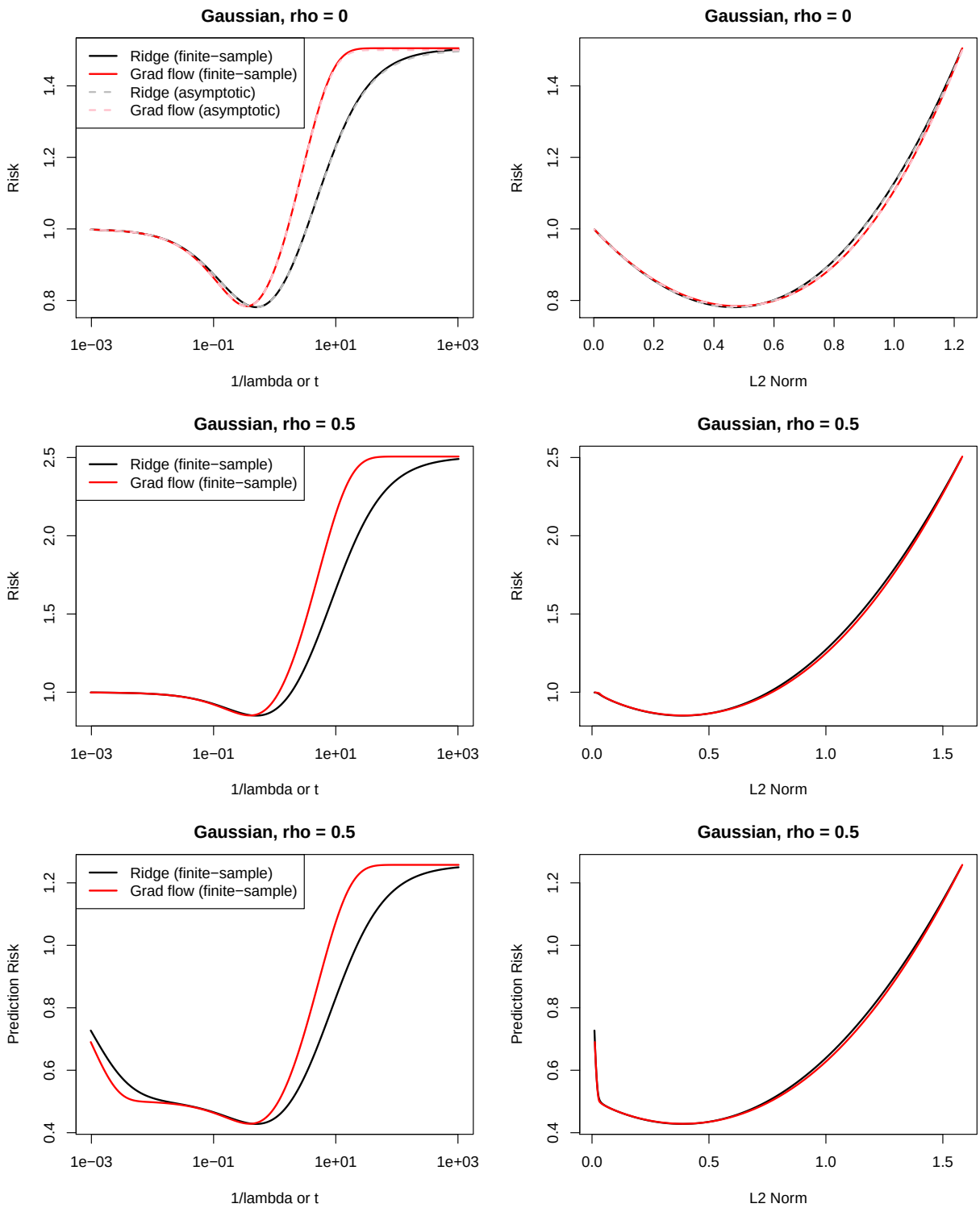

Figure S.2: *Gaussian features, with* $n = 500$ *and* $p = 1000$.

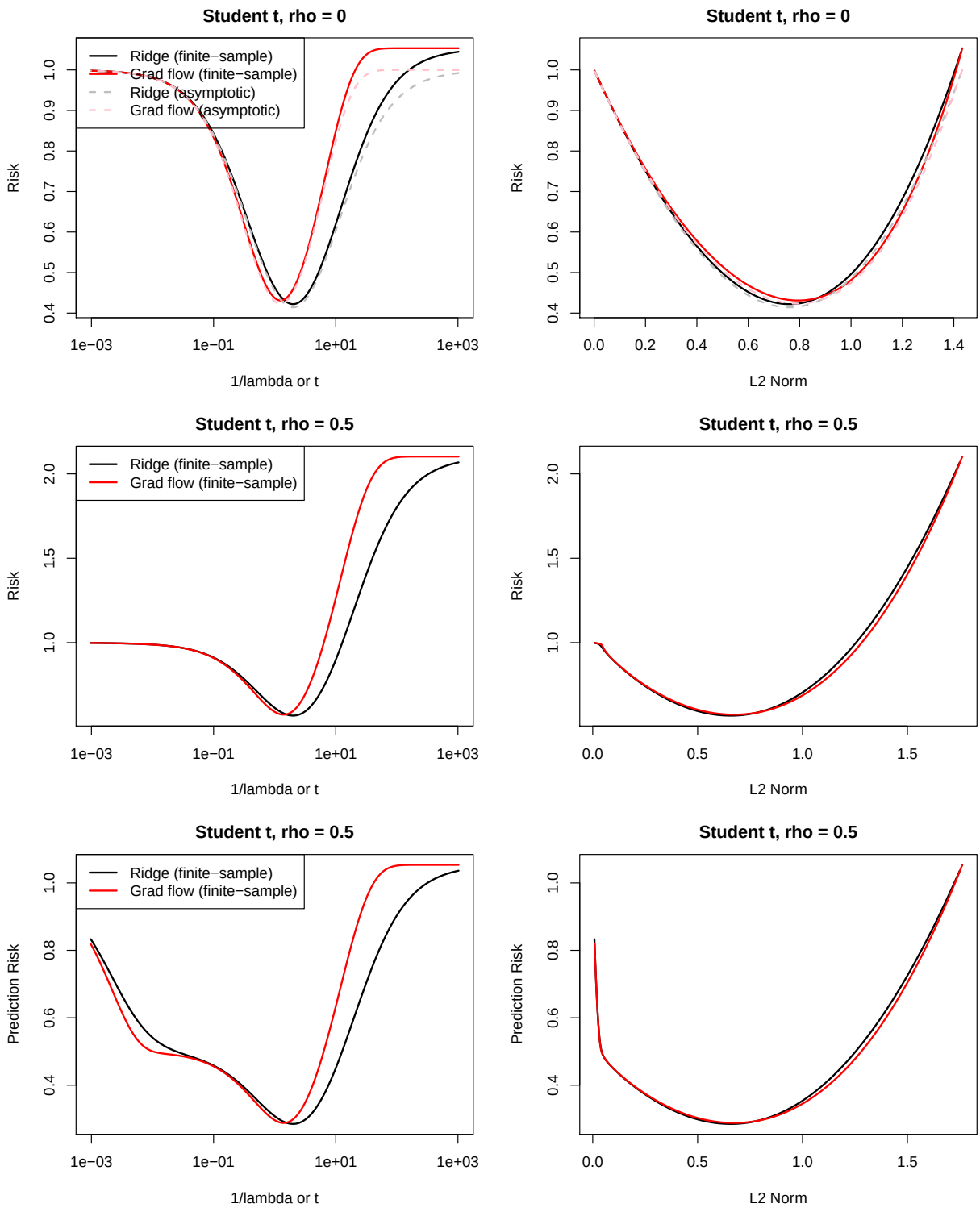

Figure S.3: *Student t features, with* $n = 1000$ *and* $p = 500$.

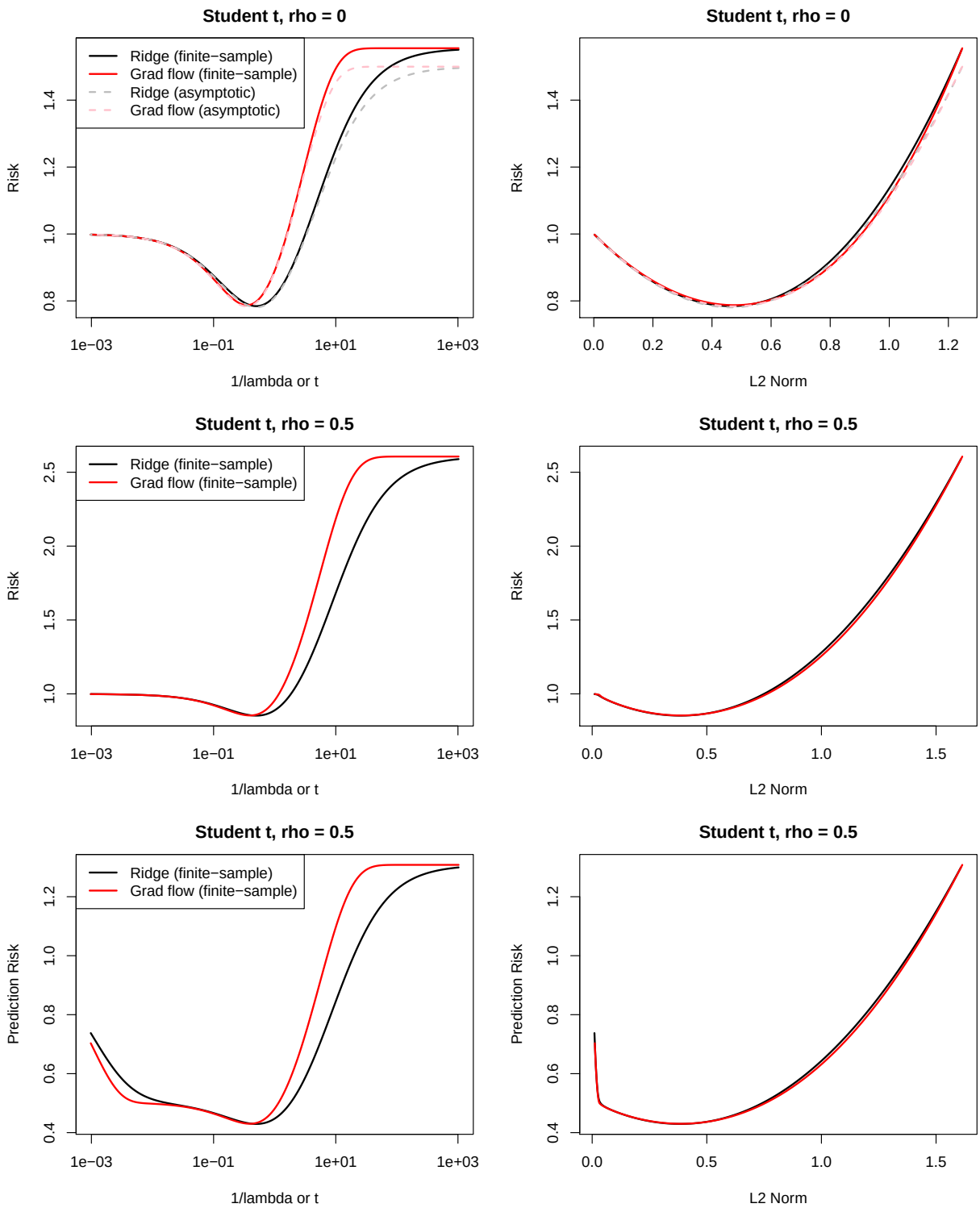

Figure S.4: *Student t features, with* $n = 500$ *and* $p = 1000$.

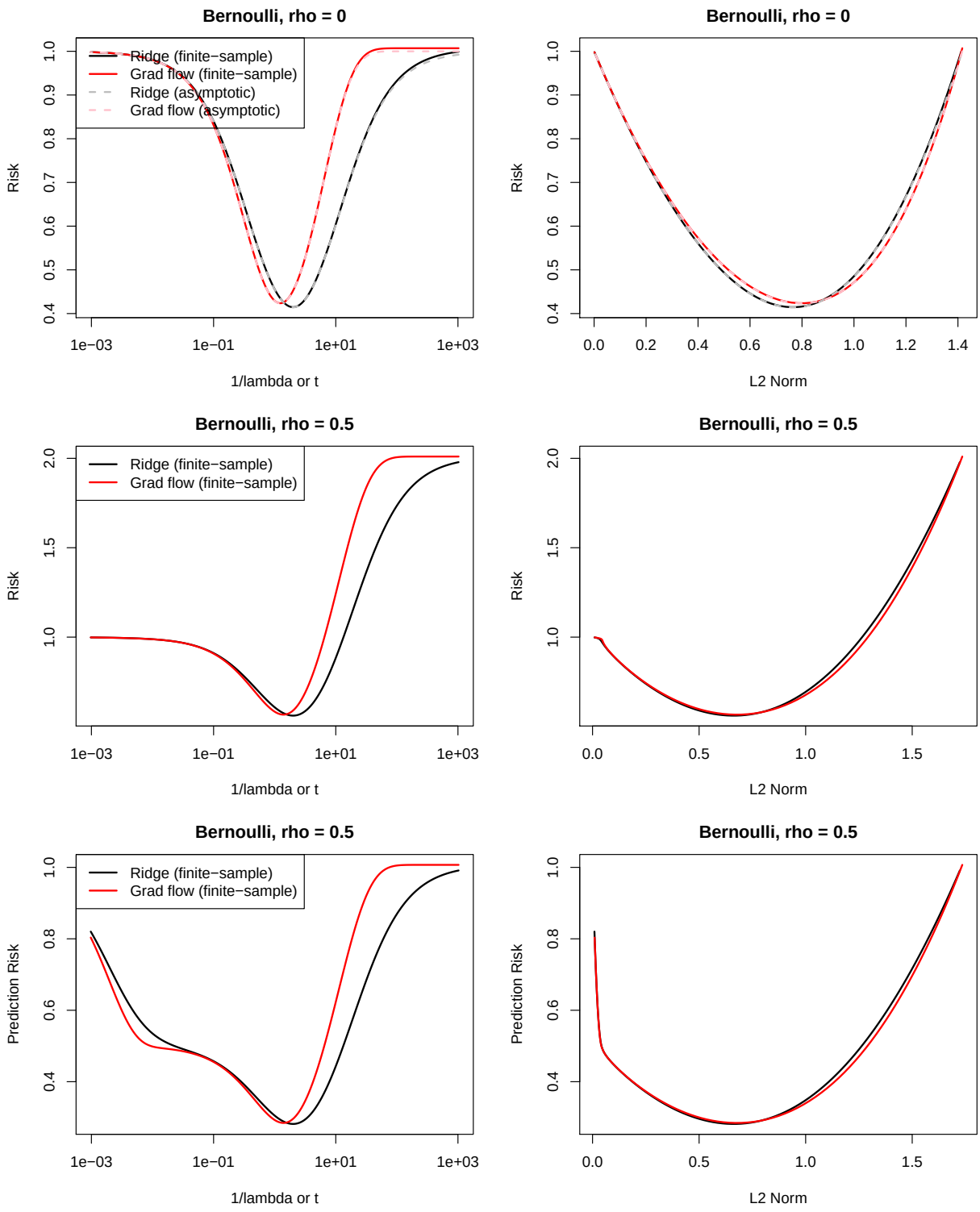

Figure S.5: *Bernoulli features, with* $n = 1000$ *and* $p = 500$.

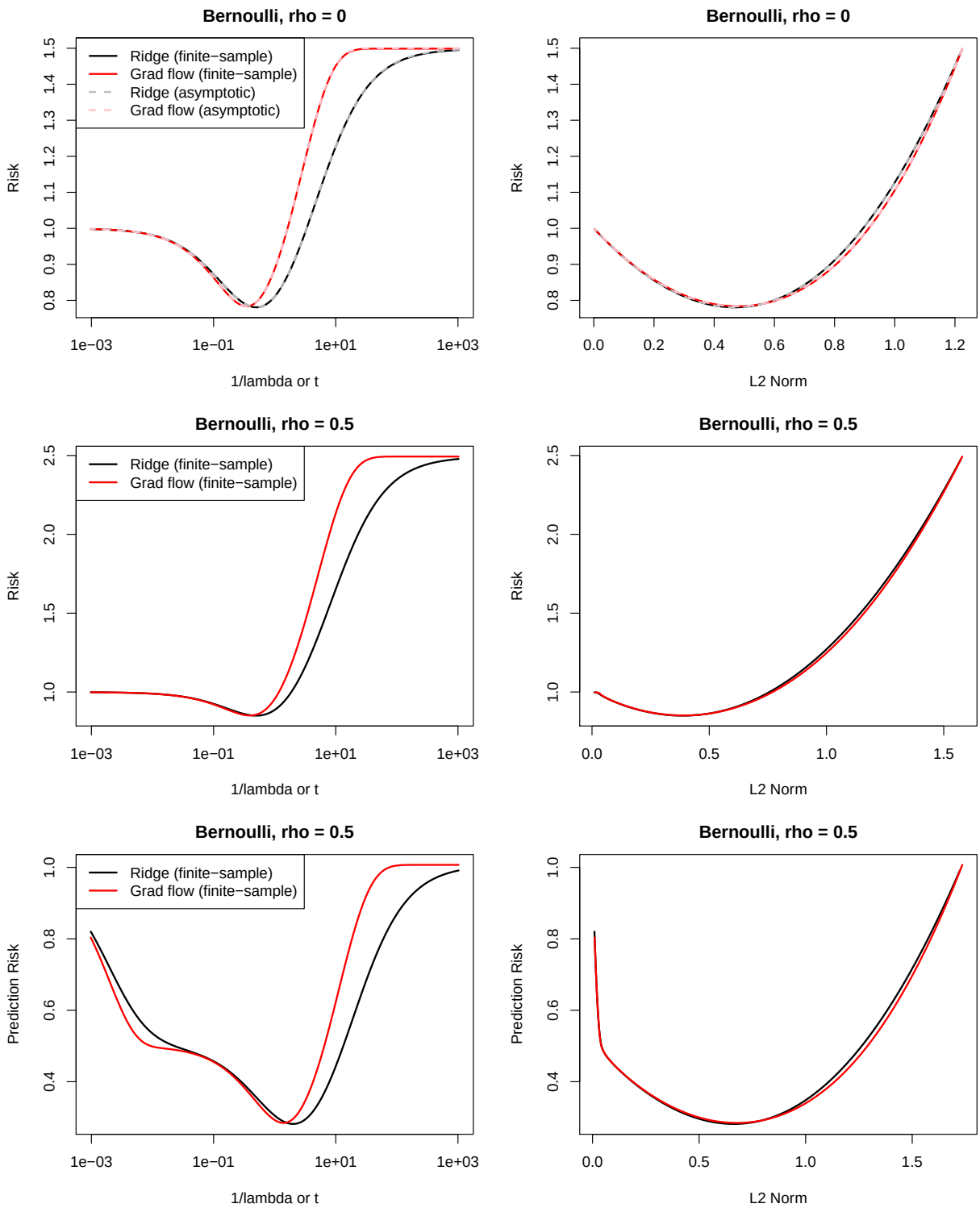

Figure S.6: *Bernoulli features, with* $n = 500$ *and* $p = 1000$.