Is Memorization Helpful or Harmful? Prior Information Sets the Threshold

Chen Cheng CHENCHENG@UCHICAGO.EDU
Department of Statistics, University of Chicago

Rina Foygel Barber RINA@UCHICAGO.EDU
Department of Statistics, University of Chicago

Editors: Steve Hanneke and Tor Lattimore

Abstract

We examine the connection between training error and generalization error for arbitrary estimating procedures, working in an overparameterized linear model under general priors in a Bayesian setup. We find determining factors inherent to the prior distribution π, giving explicit conditions under which optimal generalization necessitates that the training error be (i) near interpolating relative to the noise size (i.e., memorization is necessary), or (ii) close to the noise level (i.e., overfitting is harmful). Remarkably, these phenomena occur when the noise reaches thresholds determined by the Fisher information and the variance parameters of the prior π.

Keywords: memorization, overparameterization, interpolation, overfitting, Fisher information

1. Introduction

In this work, we consider the question of *overfitting* in high-dimensional statistical models: should we allow our trained model to overfit to the training data, perhaps even interpolating the data perfectly, or should we instead constrain the training process to return low-complexity models that avoid overfitting? This core question lies at the heart of statistics and machine learning in practice, and has received substantial attention in the theoretical literature as well, over several decades. Despite its importance and ubiquity, however, this question is still not fully understood: there are a range of findings that are apparently at odds with each other. In classical statistical machine learning, it is standard to train a model that exhibits some intrinsic low-dimensional structure, such as sparsity or low-rank, to ensure that we avoid overfitting (Tibshirani, 1996; Candes et al., 2008; Candes and Recht, 2012). On the other hand, recent work by Cheng et al. (2022) suggests that, in an overparameterized regime, memorization is *necessary* for generalization—that is, it is necessary to choose a model that overfits, in order to achieve optimal prediction error. And, another line of literature on the "benign overfitting" phenomenon can be viewed as lying in between these two extremes, establishing that overfitting is not harmful (but, perhaps not necessary) for good predictive performance (Belkin et al., 2018, 2019; Bartlett et al., 2020; Hastie et al., 2022; Cheng and Montanari, 2024). How can we reconcile these different views?

We examine this question from the perspective of a Bayesian framework, in the context of a specific problem: high-dimensional linear regression. Concretely, we consider the model

$$y = X\theta + \mathsf{N}(0, \sigma^2 I_n),$$

where $\theta \in \mathbb{R}^d$ is drawn from a prior π, and where $d \geq n$ (the overparameterized regime). In this model, an overfitted estimator $\widehat{\theta}$ is one that has (mean square) training error substantially below the noise

level σ^2. At a high level, our findings reveal that the question of whether overfitting is necessary or is instead harmful for optimal predictive peformance depends entirely on the nature of the prior π. If π encodes some sort of latent low-dimensional structure (for instance, encouraging approximately sparse θ) then overfitting is harmful, while if π is uninformative then overfitting becomes necessary. Strikingly, our findings suggest the (asymptotically optimal) determining factors delineating these scenarios for the prior π with a differentiable density $p > 0$ are the Fisher information and variance parameters:[1]

$$\text{Fisher information:} \qquad \mathsf{J}_\pi = \mathbb{E}\left[\frac{1}{n}\|\nabla \log p(X\theta)\|^2 \mid X\right], \tag{1a}$$

$$\text{Variance parameter:} \qquad \mathsf{V}_\pi = \mathbb{E}\left[\frac{1}{n}\|X\theta\|^2 \mid X\right], \tag{1b}$$

where in order to define variance, from this point on we assume that π has mean zero, without loss of generality. Intuitively, J_π measures spikiness of the prior π—and thus J_π^{-1} acts as a local measure of effective dimension. In contrast, V_π measures a macroscopic notion of dimension of π by computing a variance. Our key message is that comparing the noise level σ^2 to J_π^{-1} determines when it is necessary to overfit, and comparing to V_π determines when it is harmful to overfit.

Indeed, the Fisher information and variance parameters depend directly on the push-forward distribution of $X\theta$. As we primarily treat the design matrix X as fixed, and therefore the prior π on θ directly specifies a distribution on $X\theta$, we refer to quantities depend on either the distribution of $X\theta$ or θ as information about the prior.

1.1. Related work

High-dimensional statistical models have played a central role in extending classical large-sample asymptotic theory to modern overparametrization settings. Pioneering work that exploits intrinsic low-rank and sparsity structures in data typically relies on explicit regularization mechanisms, such as ℓ_1-penalization (Tibshirani, 1996) in compressed sensing (Candes et al., 2008) and nuclear-norm regularization in matrix completion (Candes and Recht, 2012).

The remarkable success of deep learning over the past two decades questions the presumed necessity of explicit regularization for the accurate recovery of low-rank structures. In particular, the phenomenon of "implicit regularization"—whereby specific optimization procedures converge to good local minima exhibiting "benign overfitting"—has been the subject of extensive investigation, especially in the context of large-scale datasets and deep neural networks (Neyshabur, 2017; Gunasekar et al., 2017). Closely related to our work, Hastie et al. (2022) analyze the minimum-ℓ_2-norm interpolating solution in high-dimensional linear regression, with subsequent extensions to broader settings including kernel regression (Liang and Rakhlin, 2020), random feature models (Mei et al., 2022), and general Hilbert space frameworks (Bartlett et al., 2020; Cheng and Montanari, 2024).

Memorization, or the necessity of interpolating models, still remains largely underexplored. Feldman (2020) initiate the study of this question in the multi-class classification setting, where the data-generating distribution is modeled as a mixture of heavy-tailed subpopulations and, consequently, does not exhibit low-rank structure. Cheng et al. (2022) investigate the necessity of interpolation in

1. We note that J_π denotes the Fisher information for π alone (Dembo et al., 2002, Sec. C), while the classical definition is for a parametric family of distributions $\{p_\xi(\cdot)\}_{\xi\in\mathbb{R}^n}$. Our definition is compatible with the Fisher information evaluated at $\xi = 0$ for the location model by $p_\xi(\cdot) = p(\cdot + \xi)$.

overparameterized linear models, and here we extend their framework to (almost) arbitrary priors. Characterizing the fundamental limits of model capacity making interpolation necessary has also emerged in several recent works along this line of research (Shah et al., 2025; Müller et al., 2025).

We finally point out that our analyses, largely inspired by prior research on information-theoretic inequalities (Stam, 1959; Dembo et al., 2002), deeply connect to the well established results on the minimum mean square error (MMSE) in Gaussian channels (Guo et al., 2004, 2011; Ledoux, 2016).

2. Main results

2.1. Problem setup

We first specify our problem setting and introduce the corresponding quantities of interest.

Overparameterized linear model under a general prior π. Consider a design matrix $X \in \mathbb{R}^{n\times d}$, whose rows are vectors $x_1^\top, \cdots, x_n^\top \in \mathbb{R}^d$. The design matrix may be fixed or random, as we carry out all our calculations conditional on X. We assume that $d \geq n$ (the overparameterized regime), and that X has full row rank (almost surely).

Given a prior π on $\mathbb{R}^d$ and a noise level $\sigma^2 > 0$, we assume the following model for the observed response vector $y \in \mathbb{R}^n$:

$$\text{The Bayesian model } \mathscr{M}_X(\pi, \sigma^2)\text{: } y = X\theta + \sigma\tau \text{ where } (\theta, \tau) \sim \pi \times \mathsf{N}(0, I_n).$$

That is, this is a Gaussian linear model, with coefficients $\theta \in \mathbb{R}^d$ sampled from the prior π.

Training error and prediction error. We study estimating θ of the linear model given the design matrix X and the observation y. Given an estimator $\widehat{\theta} = \widehat{\theta}(X, y)$, we define its expected training error as

$$\mathsf{Train}(\widehat{\theta}) = \mathbb{E}\left[\frac{1}{n}\left\|X\widehat{\theta} - y\right\|^2 \mid X\right],$$

where implicitly, the expected value is computed with respect to the model $\mathscr{M}_X(\pi, \sigma^2)$. The training error will typically take values in the range $[0, \sigma^2]$. Indeed, any estimator $\widehat{\theta}$ that interpolates the data will have $\mathsf{Train}(\widehat{\theta}) = 0$—that is, perfect memorization. On the other hand, an estimator with zero overfitting will have $\mathsf{Train}(\widehat{\theta}) = \sigma^2$ (i.e., the true noise level).

We can also define its prediction error at a new test point x' as $\mathbb{E}\left[(x'^\top\widehat{\theta} - x'^\top\theta)^2 \mid X\right]$. In particular, if x' is random and mean-zero, and is independent from the training data y, we can equivalently define the prediction error as

$$\mathsf{Pred}_\Sigma(\widehat{\theta}) = \mathbb{E}\left[\left\|\widehat{\theta} - \theta\right\|_\Sigma^2 \mid X\right],$$

where $\Sigma = \mathrm{Cov}(x')$ is the covariance of the random test point, and where $\|v\|_\Sigma := \sqrt{v^\top\Sigma v}$. (If $\Sigma = I_d$, then this is simply the mean-squared error (MSE) of the estimator $\widehat{\theta}$.) We will also write

$$\mathsf{Cost}(\widehat{\theta}) = \mathsf{Pred}_\Sigma(\widehat{\theta}) - \mathsf{Pred}_\Sigma^* \text{ where } \mathsf{Pred}_\Sigma^* = \inf_{\widehat{\theta}} \mathsf{Pred}_\Sigma(\widehat{\theta}),$$

i.e., this measures the excess prediction error of the estimator $\widehat{\theta}$.[2]

2. Since X is treated as fixed (i.e., we condition on X throughout), we should interpret the definition of Pred_Σ^* as computing an infimum over all possible maps $y \mapsto \widehat{\theta}(X, y)$, i.e., functions $\mathbb{R}^n \to \mathbb{R}^d$. As a technical note, we will assume throughout that any estimator $\widehat{\theta}$ considered in this paper is square-integrable (with respect to the distribution of its input, the observed response y), so that all defined measures of training and prediction error are finite.

The key question of this paper is the following: what is the relationship between the training error and the prediction error of an estimator—and, to ensure optimal predictive performance (i.e., minimal prediction error), what training error should we aim for? In particular, we are interested in identifying settings where these interesting phenomena may occur:

- Settings where **memorization is necessary** for generalization: when is it true that any estimator $\widehat{\theta}$ that is (near) optimal for predictive error, must have training error close to zero, i.e.,

$$\mathsf{Cost}_\Sigma(\widehat{\theta}) \approx 0 \implies \mathsf{Train}(\widehat{\theta}) \leq o(\sigma^2) \ ?$$

- Settings where **overfitting is harmful** for generalization: when is it true that any estimator $\widehat{\theta}$ that is (near) optimal for predictive error, must have training error close to σ^2, i.e.,

$$\mathsf{Cost}_\Sigma(\widehat{\theta}) \approx 0 \implies \mathsf{Train}(\widehat{\theta}) \geq \sigma^2 - o(\sigma^2) \ ?$$

2.2. Comparing to the Bayes-optimal estimator

Under our assumption of the Bayesian model $\mathscr{M}_X(\pi, \sigma^2)$, the Bayes estimator is given by the posterior mean,

$$\widehat{\theta}_{\mathsf{B}}(X, y) = \mathbb{E}[\theta \mid X, y].$$

The following proposition demonstrates that $\widehat{\theta}_{\mathsf{B}}$ plays a crucial role for the motivating questions of this paper.

Proposition 1 *Under the setting and notation above, for any positive definite $\Sigma \in \mathbb{R}^{d\times d}$, $\widehat{\theta}_{\mathsf{B}}$ achieves the optimal prediction error, i.e.,*

$$\mathsf{Pred}_\Sigma(\widehat{\theta}_{\mathsf{B}}) = \inf_{\widehat{\theta}} \mathsf{Pred}_\Sigma(\widehat{\theta}) = \mathsf{Pred}^*_\Sigma.$$

Moreover, letting $\lambda_\Sigma = \frac{1}{n}\|X\Sigma^{-\frac{1}{2}}\|^2$, for any estimator $\widehat{\theta}$, its excess prediction error satisfies

$$\mathsf{Cost}_\Sigma(\widehat{\theta}) = \mathsf{Pred}_\Sigma(\widehat{\theta}) - \mathsf{Pred}^*_\Sigma = \mathbb{E}\left[\left\|\widehat{\theta} - \widehat{\theta}_{\mathsf{B}}\right\|^2_\Sigma \mid X\right] \geq \lambda_\Sigma^{-1}\left(\sqrt{\mathsf{Train}(\widehat{\theta})} - \sqrt{\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})}\right)^2.$$

In other words, this standard result tells us that the Bayes estimator is optimal for prediction error, regardless of Σ. Moreover, the cost of choosing a different estimator (in terms of excess prediction error) can be lower bounded via the difference in training errors, $\mathsf{Train}(\widehat{\theta})$ versus $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$. In particular, any estimator $\widehat{\theta}$ that is close to optimal in terms of prediction error, must have a training error that is similar to that of the Bayes estimator:

$$\mathsf{Cost}_\Sigma(\widehat{\theta}) \approx 0 \implies \mathsf{Train}(\widehat{\theta}) \approx \mathsf{Train}(\widehat{\theta}_{\mathsf{B}}).$$

2.3. Training error of the Bayes estimator

The result of Proposition 1 suggests that, to characterize the regimes in which memorization is necessary or in which overfitting is harmful, we need to examine the training error $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$ of the Bayes estimator. From this point on, we assume that π has a positive and differentiable density, and we write p as the density of $X\theta$ induced by $\theta \sim \pi$ (with X treated as fixed). Recall the Fisher

information (cf. Eq. (1a)) and variance (cf. Eq. (1b)) parameters J_π and V_π, which we assume are positive and finite in this section. In particular, we note an important and well-known property:[3]

For any π, it holds that $\mathsf{V}_\pi \geq \mathsf{J}_\pi^{-1}$, with equality if and only if $X\theta \sim \mathsf{N}(0, \nu^2 I_n)$ for some $\nu^2 > 0$.

The following result characterizes the training error of $\widehat{\theta}_{\mathsf{B}}$ in terms of these two parameters.

Theorem 2 *Fix any $d \geq n \geq 1$. Let π be any prior with positive and differentiable density, and let $X \in \mathbb{R}^{n\times d}$ have full row rank n. Then under the model $\mathscr{M}_X(\pi, \sigma^2)$, the training error of the Bayes estimator satisfies*

$$\frac{\sigma^4}{\mathsf{V}_\pi + \sigma^2} \leq \mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) \leq \frac{\sigma^4}{\mathsf{J}_\pi^{-1} + \sigma^2} \quad \text{for all } \sigma^2 > 0. \tag{2}$$

Moreover, the training error approaches the above upper bound for vanishing noise,

$$\mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) = \frac{\sigma^4}{\mathsf{J}_\pi^{-1} + \sigma^2} + o(\sigma^4) = \mathsf{J}_\pi \sigma^4 + o(\sigma^4) \quad \text{as} \quad \sigma^2 \to 0, \tag{3}$$

and approaches the above lower bound for increasing noise,

$$\mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) = \frac{\sigma^4}{\mathsf{V}_\pi + \sigma^2} + o(1) = \sigma^2 - \mathsf{V}_\pi + o(1) \quad \text{as} \quad \sigma^2 \to \infty. \tag{4}$$

The asymptotic statements (3) and (4) can be interpreted as follows. With vanishing noise $\sigma^2 \to 0$, the Bayes estimator exhibits memorization (since its training error is $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) = O(\sigma^4) = o(\sigma^2)$. In contrast, when $\sigma^2 \to \infty$, the Bayes estimator instead shows a limited amount of overfitting, since $\sigma^2 - \mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) = \mathsf{V}_\pi + o(1) = o(\sigma^2)$.[4]

2.4. The tradeoff between training error and prediction error

With the results above in place, we are now ready to examine the tradeoff between training error and prediction error, for any estimator $\widehat{\theta}$.

Theorem 3 *Fix any $d \geq n \geq 1$. Let π be any prior with positive and differentiable density, and let $X \in \mathbb{R}^{n\times d}$ have full row rank n. Fix any positive definite $\Sigma \in \mathbb{R}^{d\times d}$. Then under the model $\mathscr{M}_X(\pi, \sigma^2)$, for any $\sigma^2 > 0$ and any estimator $\widehat{\theta}$,*

$$\text{If } \mathsf{Train}(\widehat{\theta}) \geq \frac{\sigma^4}{\mathsf{J}_\pi^{-1} + \sigma^2} \text{ then } \mathsf{Cost}_\Sigma(\widehat{\theta}) \geq \lambda_\Sigma^{-1} \left(\sqrt{\mathsf{Train}(\widehat{\theta})} - \sqrt{\frac{\sigma^4}{\mathsf{J}_\pi^{-1} + \sigma^2}} \right)^2, \tag{5}$$

and

$$\text{If } \mathsf{Train}(\widehat{\theta}) \leq \frac{\sigma^4}{\mathsf{V}_\pi + \sigma^2} \text{ then } \mathsf{Cost}_\Sigma(\widehat{\theta}) \geq \lambda_\Sigma^{-1} \left(\sqrt{\frac{\sigma^4}{\mathsf{V}_\pi + \sigma^2}} - \sqrt{\mathsf{Train}(\widehat{\theta})} \right)^2. \tag{6}$$

3. Essentially this is the classical Cramér-Rao bound. See this result in (Stam, 1959, Eq. (2.1)). Indeed, the upper bound by V_π can be sharpened into the entropy power. We refer the reader to Lemma A.3 for details.
4. We emphasize that in the asymptotic claims (3) and (4), the statements hold for fixed d, n, fixed design matrix X, and a fixed choice of the prior π; the $o(\cdot)$ terms reveal the asymptotic dependence on the noise level σ^2 only.

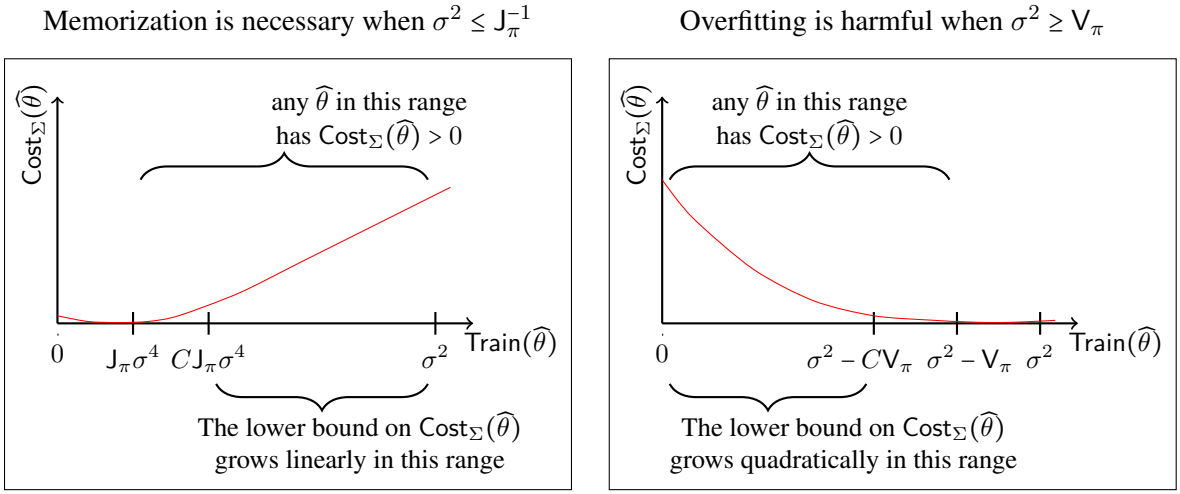

Figure 1: An illustration of the results of Corollary 4 (describing the regime where memorization is necessary), and Corollary 5 (describing the regime where overfitting is harmful). Here $C > 1$ is any constant.

Proof These results follow immediately from combining the bounds (2) in Theorem 2 (which calculates the training error of $\widehat{\theta}_{\mathsf{B}}$), with the result of Proposition 1 (which tells us that optimal prediction error can only be achieved by an estimator whose training error matches that of $\widehat{\theta}_{\mathsf{B}}$). ■

The two parts of this theorem immediately yield answers to our two key questions: under what regimes does it hold that **memorization is necessary**, or, that **overfitting is harmful**? The following two corollaries answer these questions.

First, we consider the question of when memorization is necessary: what is the cost of choosing an estimator whose training error is too high?

Corollary 4 *In the setting of Theorem 3, suppose also that* $\sigma^2 \le \mathsf{J}_\pi^{-1}$. *Then*

$$\text{If } \mathsf{Train}(\widehat{\theta}) > \mathsf{J}_\pi\sigma^4 \text{ then } \mathsf{Cost}_\Sigma(\widehat{\theta}) > 0.$$

Moreover, any estimator that fails to memorize sufficiently will incur excess prediction error that is linear in $\mathsf{Train}(\widehat{\theta})$*: for any* $C > 1$,

$$\text{If } \mathsf{Train}(\widehat{\theta}) \ge C \cdot \mathsf{J}_\pi\sigma^4 \text{ then } \mathsf{Cost}_\Sigma(\widehat{\theta}) \ge C' \cdot \mathsf{Train}(\widehat{\theta}),$$

where $C' = \lambda_\Sigma^{-1}(1 - C^{-1/2})^2$.

Next, we turn to the question of when overfitting is harmful: what is the cost of choosing an estimator whose training error is too low?

Corollary 5 *In the setting of Theorem 3, suppose also that* $\sigma^2 \ge \mathsf{V}_\pi$. *Then*

$$\text{If } \mathsf{Train}(\widehat{\theta}) < \sigma^2 - \mathsf{V}_\pi \text{ then } \mathsf{Cost}_\Sigma(\widehat{\theta}) > 0.$$

Moreover, any estimator that exhibits too much overfitting will incur excess prediction error that is quadratic in $\sigma^2 - \mathsf{Train}(\widehat{\theta})$*, i.e., in the amount of overfitting: for any* $C > 1$*,*

$$\text{If } \mathsf{Train}(\widehat{\theta}) \le \sigma^2 - C \cdot \mathsf{V}_\pi \text{ then } \mathsf{Cost}_\Sigma(\widehat{\theta}) \ge \frac{C'}{\sigma^2} \cdot \left(\sigma^2 - \mathsf{Train}(\widehat{\theta})\right)^2,$$

where $C' = (4\lambda_\Sigma)^{-1}(1 - C^{-1})^2$.

These results are illustrated in Figure 1.

Remark 6 *We point out that our characterizations are the cleanest at the extremes, when* $\sigma^2 \ge \mathsf{V}_\pi$ *or* $\sigma^2 \le \mathsf{J}_\pi^{-1}$*. As we will show in Fig. 3, under the moderate noise level* $\mathsf{J}_\pi^{-1} \le \sigma^2 \le \mathsf{V}_\pi$*, the behavior of the optimal training error relative to the noise size* $\mathsf{Train}(\widehat{\theta})/\sigma^2$ *can oscillate and be non-monotonic. It remains an interesting open question to investigate for a similar clean characterization for the necessity or harmfulness of overfitting in this regime.*

2.4.1. THE ROLE OF THE PRIOR: CONNECTIONS TO EFFECTIVE DIMENSION

In the results above, we have seen that if σ^2 is close to zero then memorization appears necessary, while if σ^2 is large then overfitting is harmful. How does this relate to the connection between overfitting and model complexity? In particular, we might expect the following:

- If θ is assumed to lie in a low-complexity model class (e.g., under sparsity), then any accurate estimator $\widehat{\theta}$ should have minimal overfitting.
- On the other hand, if we are in a truly high-dimensional setting (with no assumptions such as sparsity that would induce low-dimensional structure), then we may prefer to overfit, i.e., "benign overfitting".

In the Bayesian setting considered here, our results can be related to these principles by considering the parameters V_π and J_π^{-1}. Essentially, these parameters capture a notion of *effective dimension*.

For intuition, suppose $n = d$ and X is the identity. First, suppose our prior π is quite flat and uninformative. In this case, we would expect V_π and J_π^{-1} to both be $O(1)$ (for instance, if $\pi = \mathsf{N}(0, I_d)$ then $\mathsf{V}_\pi = \mathsf{J}_\pi^{-1} = 1$). In particular, if the noise level σ^2 is $o(1)$ then memorization becomes necessary: Corollary 4 tells us that cost is high for any estimator with training error $\gtrsim \sigma^4$.

At the other extreme, we can consider a setting where π encodes some knowledge of low-dimensional structure in θ—for instance, π may be concentrated near some low-dimensional subspace or manifold in $\mathbb{R}^d$. In such a setting, we would expect that overfitting would be harmful, since an optimal estimator should lie near this low-dimensional region and therefore would not have the capacity to overfit. For this type of prior π, J_π^{-1} would typically be quite low, and the regime in Corollary 4 (i.e., the regime in which our results show that memorization is necessary) becomes negligible. However, there is a subtlety here: if σ^2 is extremely close to zero (i.e., $\sigma^2 \le \mathsf{J}_\pi^{-1} = o(1)$), then we again find that memorization is necessary. This is an artifact of our Bayesian framework: even if π is strongly concentrated around some low-dimensional region within $\mathbb{R}^d$, since we require π to have a positive and differetiable density, if the noise level is extremely low then π is still effectively constant within any sufficiently small neighborhood. Thus if σ^2 is extremely small, π again acts (locally) as an uninformative prior (see Figure 2 for an illustration). On the other hand, if we choose π to be a distribution that places all its mass on a low-dimensional support (and therefore, π cannot have a density), this phenomenon would not occur—but this is beyond the scope of our framework.

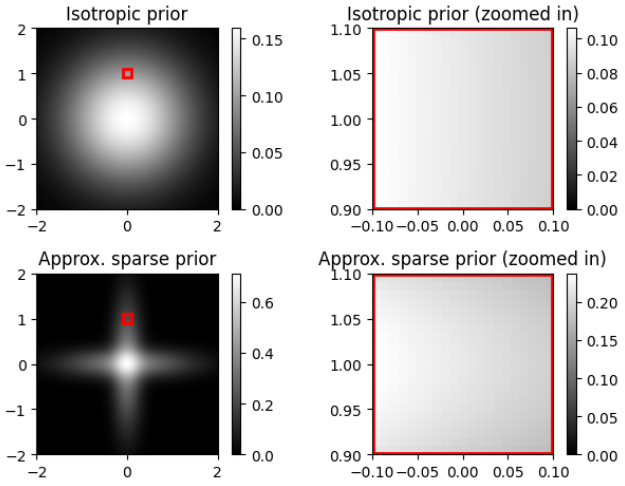

Figure 2: An illustration of the phenomenon discussed in Section 2.4.1. In the top row, we plot an isotropic prior, $\pi = \mathsf{N}(0, I_2)$, while the bottom row shows a prior that encourages approximate 1-sparsity, $\pi = 0.5\mathsf{N}(0, e_1e_1^\top + \eta e_2e_2^\top) + 0.5\mathsf{N}(0, \eta e_1e_1^\top + e_2e_2^\top)$, for $\eta = 0.05$. However, when we zoom in to the neighborhood of a single point $(0, 0.5)$, the two priors are both essentially constant.

In the next section, we explore these ideas through concrete examples to demonstrate the role of effective dimension in this problem.

3. Examples with differing levels of complexity

In this section, we apply our results to investigate concrete setups of $\mathscr{M}_X(\pi, \sigma^2)$ under different priors for the key determining parameters $\mathsf{V}_\pi, \mathsf{J}_\pi, \lambda_\Sigma$ in Theorem 3.

Throughout this section, we will work in a setting where the training data points x_i and the test data point x' all assumed to have mean zero and covariance $\Sigma = I_d$. To characterize high dimensional behaviors by exact limits, we leverage tools in high dimensional random matrix theory under proportional asymptotics (Bai and Silverstein, 2010), working under the following distributional assumptions on the training data features $x_1, \cdots, x_n \in \mathbb{R}^d$ and overparameterized proportional asymptotics.

Assumption 1 (Overparameterized proportional asymptotics) *For a given $\gamma \in (1, +\infty)$, consider a deterministic sequence of positive integers $d(n)$ indexed by n, such that $d/n \to \gamma$. For an infinite array of i.i.d. random variables $\{X_{ij}\}_{i,j=1}^{\infty}$ such that $\mathbb{E}[X_{ij}] = 0$, $\mathsf{Var}(X_{ij}) = 1$ with bounded fourth-moment $\mathbb{E}[X_{ij}^4] < +\infty$, define $X(n) = (X_{ij})_{i\in[n], j\in[d]} \in \mathbb{R}^{n\times d}$. We consider the sequence of linear models $\mathscr{M}_X(\pi, \sigma^2) = \mathscr{M}_{X(n)}(\pi(n), \sigma^2)$ indexed by n.*

3.1. Isotropic Gaussian prior

We start with the simplest prior of the isotropic Gaussian $\pi = \mathsf{N}(0, I_d/d)$. The following proposition shows the implications of our general results for this isotropic setting.[5]

Proposition 7 *Let Assumption 1 hold. For $\sigma^2 > 0$ and $\Sigma = I_d$, we have almost surely*

$$\mathsf{V}_\pi \to 1, \qquad \mathsf{J}_\pi \to \frac{\gamma}{\gamma - 1}, \qquad \lambda_\Sigma \to (1 + \sqrt{\gamma})^2,$$

Recalling the results of Section 2.4 (and Corollary 4 in particular), we see that memorization is necessary when the noise is not too high, since V_π and J_π are $O(1)$. We can interpret this finding as showing that, when π is essentially uninformative, we are in a genuinely overparameterized regime: in the absence of any implicit structural knowledge of θ, overfitting is unavoidably necessary, and strong predictive performance requires memorization.

3.2. Approximately-low-rank Gaussian prior

In the second example, we begin to investigate scenarios characterized by an underlying underparameterized structure, in which the model lacks sufficient capacity to overfit the noise. To capture this structure while still ensure $X\theta$ has a density p, we consider an approximately low-dimensional prior:

$$\pi = \pi(n) = \mathsf{N}(0, \Omega(n)), \qquad \Omega = \Omega(n) = \frac{1-\eta}{r}\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} + \frac{\eta}{d} I_d, \qquad 0 < \eta \le 1. \tag{7}$$

For small η, this prior is approximately supported on a low-rank subspace (spanned by the first r basis vectors). As $\eta \to 0+$, if $r < n$ then the model approaches an underdetermined linear model (where $\mathsf{Train}(\widehat{\theta})$ cannot be zero for any estimator), while if $r \ge n$ then the model essentially approaches the same setting as the isotropic Gaussian example of Section 3.1, except with r in place of d.

Proposition 8 *Let Assumption 1 hold, $r/n \to \rho \in (0, +\infty)$, and π defined as in (7) above. Define the constant $C = C(\gamma, \rho, \eta) = \eta^2(\gamma - \rho - 1) + \eta(\gamma\rho - \rho^2 - \rho) + \gamma(\rho - 1)$. For $\sigma^2 > 0$ and $\Sigma = I_d$,*

$$\mathsf{V}_\pi \to 1, \qquad \mathsf{J}_\pi \to \frac{-C + \sqrt{C^2 + 4\eta(\gamma-1)\gamma(\eta+\rho)}}{2\eta(\gamma-1)}, \qquad \lambda_\Sigma \to (1+\sqrt{\gamma})^2$$

almost surely. In particular, we have the following limits as $\eta \to 0+$:

$$\textit{Underparameterization limit: if } \rho < 1, \qquad \mathsf{J}_\pi = \frac{\gamma(1-\rho)}{\eta(\gamma-1)} + o(\eta^{-1});$$

$$\textit{Interpolation threshold limit: if } \rho = 1, \qquad \mathsf{J}_\pi = \sqrt{\frac{\gamma}{\eta(\gamma-1)}} + o(\eta^{-\frac{1}{2}});$$

$$\textit{Overparameterization limit: if } \rho > 1, \qquad \mathsf{J}_\pi = \frac{\rho}{\rho - 1} + o(1).$$

5. This is the same setup considered in Cheng et al. (2022), except that the authors originally examine $X(n)$ whose rows have a general covariance structure $\Sigma(n)$ with bounded condition numbers and converging empirical spectrum distributions (and measure test error against the same covariance $\Sigma(n)$); however, the qualitative conclusions remain the same by taking $\Sigma(n) = I_d$. See Appendix C.1 for extensions of our results to the case of a general $\Sigma(n)$.

In the regime $\rho > 1$ (i.e., $r > n$), even as $\eta \to 0+$ the model is still overparameterized, and essentially reduces to the same result as the isotropic Gaussian, in Proposition 7: we have $\mathsf{J}_\pi \to \frac{\rho}{\rho-1}$ in place of $\frac{\gamma}{\gamma-1}$, since the effective dimension is $r = \rho n$ instead of $d = \gamma n$. In particular, memorization is again necessary, since the results of Corollary 4 apply as soon as $\sigma^2 \le \mathsf{J}_\pi^{-1} \asymp 1$.

On the other hand, if $\rho < 1$ (i.e., $r < n$), we see different behavior. Indeed, if η is close to zero (i.e., the prior is strongly concentrated near the low-rank subspace), then if the noise σ^2 is not too small, intuitively we would expect overfitting to be harmful since we have knowledge of low-rank structure in θ (as expressed by π). This is confirmed by the theory: by Proposition 8 and Corollary 4, as $\eta \to 0+$ we see that memorization is necessary once $\sigma^2 \le \mathsf{J}_\pi^{-1} = O(\eta)$, but if $\eta \to 0+$ then any positive noise level will not fall into this regime.

For comparison, we can consider a prior with exact low-rank structure,

$$(\theta_1, \dots, \theta_r)^\top \sim \mathsf{N}(0, I_r/r), \qquad (\theta_{r+1}, \dots, \theta_d) = 0, \tag{8}$$

which we can view as a limit of (7) by setting $\eta = 0$. This prior does not have a density on $\mathbb{R}^d$ and thus lies outside the framework of our main results in Theorem 2, which bounds the training error of the Bayes estimator—but in this simple setting, we can instead compute $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$ directly.

Proposition 9 *Let Assumption 1 hold, let $r/n \to \rho \in (0,1)$, and let $\sigma^2 > 0$ be fixed. Then, for the low-rank prior π defined in (8), it holds almost surely that* $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) \to \sigma^2 \cdot (1-\rho) + \sigma^4 \cdot \frac{\rho^2}{1-\rho} + O(\sigma^6)$.

Consequently, any near-optimal estimator $\widehat{\theta}$ must have, at most, a bounded amount of overfitting: by Proposition 1, we will have $\mathsf{Cost}_\Sigma(\widehat{\theta}) \gtrsim \sigma^2$ if $\sigma^2 - \mathsf{Train}(\widehat{\theta}) \ge c\sigma^2$ for any constant $c > \rho$.

3.3. Mixture of approximately-sparse priors

We now move beyond the Gaussian prior setting, in which the absence of a closed-form expression for $\nabla \log p(y)$ makes it challenging for exact evaluation of the key Fisher information quantity J_π. Nevertheless, in certain cases we can still derive bounds that characterize its behavior. In particular, we consider the mixture of the collection of perturbed K-sparse priors:

$$\pi = \pi(n) = \frac{1}{\binom{d}{K}} \sum_{S \subset [d], |S|=K} \mathsf{N}(0, \Omega_S(n)), \qquad \Omega_S = \Omega_S(n) = \frac{1-\eta}{K} \sum_{i \in S} e_i e_i^\top + \frac{\eta}{d} I_d. \tag{9}$$

This prior is appropriate for a setting where we believe θ is K-sparse, but have no knowledge of its support. The distribution of $X\theta$ induced by $\theta \sim \pi$ then has the density

$$p(y) = \frac{1}{\binom{d}{K}} \sum_{S \subset [d], |S|=K} \phi(y; 0, X\Omega_S X^\top),$$

where $\phi(\cdot\,; \mu, \Sigma)$ denotes the density of the $\mathsf{N}(\mu, \Sigma)$ distribution.

Proposition 10 *Let Assumption 1 hold, and π defined as in (9) above. Assume also that $K/n \to 0$. For $\sigma^2 > 0$ and $\Sigma = I_d$, we have almost surely*

$$\frac{1}{\eta} \cdot e \left(1 - \frac{1}{\gamma}\right)^{\gamma-1} \le \liminf_{n\to\infty} \mathsf{J}_\pi \le \limsup_{n\to\infty} \mathsf{J}_\pi \le \frac{\gamma}{\eta(\gamma-1)}.$$

In addition, when $\gamma \to +\infty$, both sides converge to $1/\eta$.

Assuming $K/n \to 0$ ensures that, under the prior π, the model is essentially low-dimensional. Similarly to the results of Section 3.2 (for the underparameterized case $r < n$ considered there), here we see that, provided $\eta \approx 0$—i.e., the prior is strongly concentrated on approximately-sparse values of θ—the results only suggest that memorization is necessary if the noise level σ^2 is extremely small.

4. Examining the training error of the Bayes estimator

As implied by Theorem 3, quantifying the cost of overfitting (or not overfitting) hinges on characterization of the training error achieved by the optimal Bayes estimator under a given $\mathscr{M}_X$ specification. In this regard, Theorem 2 constitutes the principal driving mechanism for the central results presented in this work. In this section we examine this training error more closely:

1. In Sec. 4.1, we present Tweedie's formula and a Fisher information expression for $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$ valid for all $\sigma^2 > 0$, which serve as the foundation in the proof of Theorem 2 in Appendix B.2.
2. In Sec. 4.2, we analyze monotonicity properties of $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$ w.r.t. the noise level σ^2.

To simplify notation, throughout this section, we denote the density of $\mathsf{N}(0, \sigma^2 I_n)$ by ϕ_{σ^2}. Let π' be the probability distribution of $X\theta$ in $\mathbb{R}^n$, and let $\pi'_{\sigma^2} := \pi' * \mathsf{N}(0, \sigma^2 I_n)$, which is the marginal distribution of y. We also assume that the density $p > 0$ of π' always exists and is differentiable, and continue to assume that π has mean zero, as before. Write $p_{\sigma^2} = p * \phi_{\sigma^2}$ as the density of π'_{σ^2}.[6] Then

$$\mathsf{Train}\left(\widehat{\theta}_{\mathsf{B}}\right) = \frac{1}{n}\mathbb{E}_{y\sim\pi'_{\sigma^2}}\left[\left\|X\widehat{\theta}_{\mathsf{B}} - y\right\|^2 \mid X\right] = \frac{1}{n}\mathbb{E}_{\theta\sim\pi,\tau\sim\mathsf{N}(0,I_n)}\left[\left\|X\widehat{\theta}_{\mathsf{B}}(X, X\theta+\sigma\tau) - X\theta - \sigma\tau\right\|_2^2 \mid X\right].$$

The above manner in which $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$ depends on the specifications of the linear model $\mathscr{M}_X$ is not transparent. Thanks to the remarkable representation of the posterior mean through the score function, known as Tweedie's formula (Efron, 2011) in Bayesian estimation, we can concisely express $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$ in terms of the Fisher information of π'_{σ^2}.

4.1. Tweedie's formula and Fisher information

By Tweedie's formula,

$$\mathbb{E}[X\theta \mid X, y] = y + \sigma^2 \nabla \log p_{\sigma^2}(y),$$

and from this we can immediately see how to characterize the training error by the Fisher information of π'_{σ^2} by rearranging terms and taking expectation over y.

Lemma 4.1 *It holds for all noise level $\sigma^2 > 0$ that*

$$\mathsf{Train}\left(\widehat{\theta}_{\mathsf{B}}\right) = \frac{\sigma^2}{n} \cdot \mathbb{E}_{\theta\sim\pi,\tau\sim\mathsf{N}(0,I_n)}\left[\left\|\frac{\partial}{\partial\tau}\log\mathbb{E}_{\theta'\sim\pi}\left[\exp\left\{-\frac{\|X\theta' - X\theta - \sigma\tau\|^2}{2\sigma^2}\right\} \mid X, y\right]\right\|^2 \mid X\right]. \tag{10a}$$

Moreover, defining the Fisher information matrix of π'_{σ^2} as

$$\mathsf{I}(\sigma^2) := \mathbb{E}_{y\sim\pi'_{\sigma^2}}\left[\nabla \log p_{\sigma^2}(y) \nabla \log p_{\sigma^2}^{\mathsf{T}}(y)\right] = -\mathbb{E}_{y\sim\pi'_{\sigma^2}}\left[\nabla^2 \log p_{\sigma^2}(y)\right],$$

6. The differentiability assumption can be relaxed in certain cases, which we will discuss in the Appendix.

it holds that

$$\mathsf{Train}\left(\widehat{\theta}_{\mathsf{B}}\right) = \frac{\sigma^4}{n} \cdot \mathbb{E}_{y\sim\pi'_{\sigma^2}}\left[\|\nabla \log p_{\sigma^2}(y)\|^2\right] = \frac{\sigma^4}{n} \cdot \mathrm{trace}\left(\mathsf{I}(\sigma^2)\right). \tag{10b}$$

Remark 11 *Lemma 4.1 suggests that the trace of the Fisher information matrix* $\mathsf{I}(\sigma^2)$ *governs the training error of the Bayes estimator* $\widehat{\theta}_{\mathsf{B}}$*. In the low noise regime* $\sigma^2 \to 0^+$*, we anticipate*

$$\mathsf{Train}\left(\widehat{\theta}_{\mathsf{B}}\right) = (1+o(1))\frac{\sigma^4}{n} \lim_{\sigma^2\to 0^+} \mathrm{trace}\left(\mathsf{I}(\sigma^2)\right).$$

Higher-order expansions of $\mathsf{Train}\left(\widehat{\theta}_{\mathsf{B}}\right)$ *can be obtained from the asymptotics of* $\mathrm{trace}\left(\mathsf{I}(\sigma^2)\right)$ *as* $\sigma^2 \to 0+$*, which we defer to Appendix E. As* $\sigma^2 \to +\infty$*, the distribution* π'_{σ^2} *converges to* $\mathsf{N}(0, \sigma^2 I_n)$*, and the Fisher information satisfies* $\mathsf{I}(\sigma^2) = (1+o(1))I_n/\sigma^2$*. Hence, as* $\sigma^2 \to +\infty$*,*

$$\mathsf{Train}\left(\widehat{\theta}_{\mathsf{B}}\right) = (1+o(1))\frac{\sigma^4}{n} \cdot \frac{n}{\sigma^2} = (1+o(1))\sigma^2.$$

4.2. Monotonicity of the training error

Overloading notation, we next define $\mathsf{Train}(\sigma^2)$ as a function on $(0,+\infty)$ to represent the training error of $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})$ at noise level σ^2—that is, $\mathsf{Train}(\sigma^2)$ is defined as the right-hand side of (10a). We also define an auxiliary function J on $(0,+\infty)$ for the Fisher information as

$$\mathsf{J}(\sigma^2) = \frac{1}{n} \cdot \mathbb{E}_{y\sim\pi'_{\sigma^2}}\left[\|\nabla \log p_{\sigma^2}(y)\|^2\right] = \frac{1}{n}\mathrm{trace}(\mathsf{I}(\sigma^2)). \tag{11}$$

By the equivalence of Eqs. (10a) and (10b) from Lemma 4.1, we have $\mathsf{Train}(\sigma^2) = \sigma^4\mathsf{J}(\sigma^2)$. We also extend the definition to $\mathsf{J}(0) = \mathsf{J}_\pi$ and $\mathsf{Train}(0) = 0$ if π's push-forward distribution π' has finite Fisher information.

A natural direction for deepening our understanding of the behavior of the Bayes estimator is to investigate the evolution—in particular, monotonicity—of $\mathsf{Train}(\sigma^2)$ as a function of the noise variance parameter σ^2. Intuitively, $\mathsf{Train}(\sigma^2)$ should increase as the additional noise makes the optimal learning procedure less confident. Employing powerful tools for Gaussian channels (Guo et al., 2004, 2011) enables us to conclude the following.

Proposition 12 *If* J_π *is finite, then* $\mathsf{Train}(\sigma^2)$ *is monotonically non-decreasing on* $[0,+\infty)$*, and* $\mathsf{Train}(\sigma^2)/\sigma^4 = \mathsf{J}(\sigma^2)$ *is monotonically non-increasing on* $[0,+\infty)$*.*

However, the function $\mathsf{Train}(\sigma^2)/\sigma^2$ (i.e., the training error relative to the noise level) can in general be non-monotonic: see Fig. 3 for an example[7].

5. Discussion

Our results, extending the memorization phenomenon in Cheng et al. (2022) from Gaussian priors to general priors, provide summarizing parameters J_π^{-1} and V_π, which are intrinsic to the underlying parameter distribution π and capture a notion of effective dimension. These parameters yield asymptotically optimal guaranties for characterizing when interpolation is necessary or harmful.

7. See the code for the experiment in https://github.com/Moriartycc/is-memorization-helpful-or-harmful

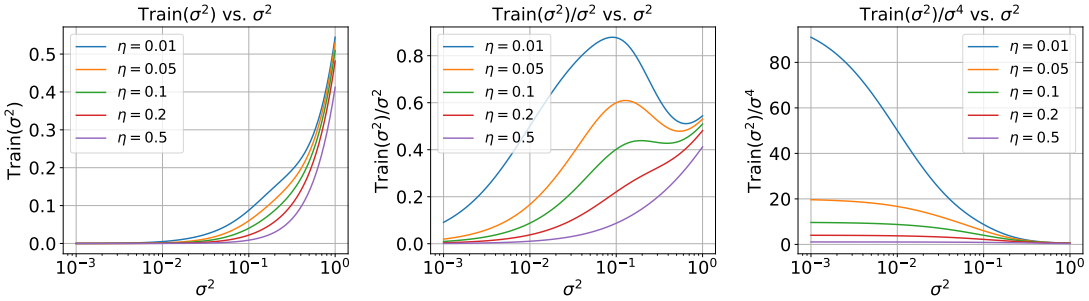

Figure 3: Numerical simulations for $\pi' = 0.5\mathsf{N}(-1, \eta) + 0.5\mathsf{N}(1, \eta)$. From left to right, we plot $\mathsf{Train}(\sigma^2)$, $\mathsf{Train}(\sigma^2)/\sigma^2$ and $\mathsf{Train}(\sigma^2)/\sigma^4$ vs. σ^2.

Furthermore, Theorem 2 does not rely on proportional asymptotics nor assume a high-dimensional ambient space, and thus holds in a fully general setting.

For linear models, the hypothesis class consists essentially of the parametric family $\mathcal{F}_\Theta = \{x^\top\theta \mid \theta \in \Theta = \mathbb{R}^d\}$. It is of considerable interest to extend the present analysis to more general model classes $\mathcal{F}_\Lambda = \{f_\lambda\}$, where Λ may be either a parametric or a nonparametric index set; it would also be valuable to investigate analogous phenomena outside the Bayesian framework without imposing a prior distribution on Θ.

Finally, the role of λ_Σ is inherently tied to the finite-dimensional ambient space $d < \infty$. In Hilbert spaces (Bartlett et al., 2020; Cheng and Montanari, 2024), one has $\lambda_\Sigma = \infty$, while the cited works show that benign overfitting occurs for arbitrary noise levels in such settings. A future theory of memorization in Hilbert spaces will therefore require a fundamentally different analytical approach.

Acknowledgments

R.F.B. was partially supported by the National Science Foundation via grant DMS-2023109. C.C. and R.F.B. were additionally supported by the Office of Naval Research via grant N00014-24-1-2544.

References

Shiri Artstein, Keith Ball, Franck Barthe, and Assaf Naor. Solution of shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 2006.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd international conference on artificial intelligence and statistics*, pages 1611–1619. PMLR, 2019.

Patrick Billingsley. *Probability and Measure*. Wiley, 3 edition, 1995.

Sergey G Bobkov, Gennadiy P Chistyakov, and Friedrich Götze. Fisher information and the central limit theorem. *Probability theory and related fields*, 159(1):1–59, 2014.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5):877–905, 2008.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52 (6):2879–2912, 2024.

Chen Cheng, John Duchi, and Rohith Kuditipudi. Memorize to generalize: on the necessity of interpolation in high dimensional linear regression. In *Conference on Learning Theory*, pages 5528–5560. PMLR, 2022.

Amir Dembo, Thomas M Cover, and Joy A Thomas. Information theoretic inequalities. *IEEE Transactions on Information theory*, 37(6):1501–1518, 2002.

Alex Dytso, H Vincent Poor, and Shlomo Shamai Shitz. A general derivative identity for the conditional mean estimator in gaussian noise and some applications. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1183–1188. IEEE, 2020.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, pages 954–959, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and mmse in gaussian channels. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, pages 349–349. IEEE, 2004.

Dongning Guo, Yihong Wu, Shlomo S Shitz, and Sergio Verdú. Estimation in gaussian noise: Properties of the minimum mean-square error. *IEEE Transactions on Information Theory*, 57(4): 2371–2385, 2011.

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

Peter J Huber and Elvezio M Ronchetti. *Asymptotic Minimax Theory for Estimating Location*, chapter 4, pages 71–103. John Wiley & Sons, Ltd, 2009. ISBN 9780470434697. doi: https://doi.org/10.1002/9780470434697.ch4. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470434697.ch4.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Michel Ledoux. Heat flow derivatives and minimum mean-square error in gaussian noise. *IEEE Transactions on Information Theory*, 62(6):3401–3409, 2016.

Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Manuel M Müller, Yuetian Luo, and Rina Foygel Barber. Are all models wrong? fundamental limits in distribution-free empirical model falsification. *arXiv preprint arXiv:2502.06765*, 2025.

Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Peter Petersen. *Riemannian geometry*. Springer, 2006.

Kulin Shah, Alkis Kalavasis, Adam R Klivans, and Giannis Daras. Does generation require memorization? creative diffusion models using ambient diffusion. *arXiv preprint arXiv:2502.21278*, 2025.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

Jack W Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

Aart J Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Appendix A. Technical lemmas

In this section, we gather some technical lemmas from existing work, that will be helpful in the main proofs.

The first result concerns the computation of moments of the multivariate normal distribution. Isserlis originally established the general formula for arbitrary moments, whereas we require only the following identity for the fourth-order moment.

Lemma A.1 (Isserlis's theorem (Isserlis, 1918)) *Let $X = (X_1, X_2, X_3, X_4)$ be a zero-mean Gaussian random vector. The following identity holds:*

$$\mathbb{E}[X_1X_2X_3X_4] = \mathbb{E}[X_1X_2]\mathbb{E}[X_3X_4] + \mathbb{E}[X_1X_3]\mathbb{E}[X_2X_4] + \mathbb{E}[X_1X_4]\mathbb{E}[X_2X_3].$$

The next result is Stam's Fisher information inequality (Stam, 1959) for sums of independent random variables, based on Shannon's entropy power inequality (Shannon, 1948). While Stam's original result is stated for scalar random variables, a vector-valued extension can be found in (Dembo et al., 2002, Thm. 13). We state an equivalent version given in (Dembo et al., 2002, Eq. (37)) as follows.

Lemma A.2 (Fisher information inequality) *For two independent vectors $X, Y \in \mathbb{R}^n$ with densities p, q, let $Z = X + Y$ with density $p * q$. The following inequality holds whenever all quantities are well-defined:*

$$\frac{1}{\mathbb{E}_Z\left[\|\nabla \log(p*q)(Z)\|^2\right]} \geq \frac{1}{\mathbb{E}_X\left[\|\nabla \log p(X)\|^2\right]} + \frac{1}{\mathbb{E}_Y\left[\|\nabla \log q(Y)\|^2\right]}.$$

In the same paper (Stam, 1959), Stam provides a lower bound for the Fisher information by the entropy power of a random variable, sharpening the Cramér-Rao bound by the variance. The result, often referred to as entropic isoperimetric inequality, is then later extended into the vector form. Here we present the version of the result stated in (Dembo et al., 2002, Thm. 16).

Lemma A.3 (Entropic isoperimetric inequality) *For a random vector $X \in \mathbb{R}^n$ with density p and for which the Fisher information exists,*

$$\frac{1}{n}\mathbb{E}_X\left[\|\nabla \log p(X)\|^2\right] \geq 2\pi e \cdot \exp\left\{-\frac{2}{n}\mathbb{E}_X\left[-\log p(X)\right]\right\}.$$

The equality holds iff $X \sim \mathsf{N}(0, \nu^2 I_n)$ for some $\nu^2 > 0$.

The following lemmas cover several classical results concerning high-dimensional sample covariance matrices within the framework of random matrix theory that will be useful in our analyses for specific examples in Sec. 3. These include (i) the Bai–Yin law, which characterizes the asymptotic behavior of the largest and smallest eigenvalues, and (ii) the Marchenko–Pastur law, which describes the limiting empirical spectral distribution. We describe our results in a more general setting than Assumption 1, relaxed to also incorporate the underparameterized regime $\gamma \in (0, +\infty)$. We also assume the data features $X = Z\Sigma^{\frac{1}{2}}$ such that Z has i.i.d. entries with zero mean and unit variance and bounded fourth-moment, and we let $\Sigma = \Sigma(n)$ form a sequence of p.s.d. matrices. Let $\widehat{S} = XX^\top/d \in \mathbb{R}^{n\times n}$ and denote by the nonzero eigenvalues of $\widehat{S}$ by $\lambda_1 \geq \cdots \lambda_{n\vee d} > 0 = \lambda_{n\vee d+1} = \cdots = \lambda_n$ and the empirical spectrum distribution (e.s.d.) by $\mu_n(\mathrm{d}\lambda) = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}$, we have the standard Bai-Yin and Marchenko-Pastur theorems for $\Sigma = I$ (cf. (Bai and Silverstein, 2010, Thm. 3.6 and Thm. 5.10)).

Lemma A.4 (Marchenko-Pastur and Bai-Yin laws) *Let $d/n \to \gamma \in (0, +\infty)$. The largest and smallest nonzero eigenvalues of $\widehat{S}$ satisfies the Bai-Yin theorem:*

$$\lambda_1 \overset{a.s.}{\to} \left(1 + \gamma^{-\frac{1}{2}}\right)^2 =: \lambda_\gamma^+, \qquad \lambda_n \overset{a.s.}{\to} \left(1 - \gamma^{-\frac{1}{2}}\right)^2 =: \lambda_\gamma^-.$$

With probability one, the e.s.d. μ_n of $\widehat{S}$ converges weakly to the Marchenko-Pastur law:

$$\mu_n(\mathrm{d}\lambda) \overset{d}{\to} \mu_{\mathsf{MP},\gamma}(\mathrm{d}\lambda) = (1-\gamma)_+ \delta_0 + \frac{\gamma}{2\pi} \cdot \frac{\sqrt{(\lambda_\gamma^+ - \lambda)(\lambda - \lambda_\gamma^-)}}{\lambda} \mathbb{1}_{\lambda_\gamma^- \le \lambda \le \lambda_\gamma^+} \mathrm{d}\lambda$$

We then consider the case when Σ is a deterministic sequence of p.s.d. matrices whose e.s.d. converges. In particular, let $\sigma_1 \ge \cdots \ge \sigma_d \ge 0$ denote the eigenvalues of Σ and $\nu_n(\mathrm{d}\lambda) = \frac{1}{d}\sum_{i=1}^d \delta_{\sigma_i}$. To facilitate the statement, we introduce the Stieltjes transform $m_\mu(z) : \mathbb{C}_+ \to \mathbb{C}_+$ of a probability measure μ on $\mathbb{R}$, where $\mathbb{C}_+ = \{u + iv \mid v > 0\}$. $m_\mu(z)$ is then

$$m_\mu(z) = \int \frac{1}{\lambda - z} \mu(\mathrm{d}\lambda). \tag{12}$$

It is well-defined everywhere on $\mathbb{C}_+$ since $|(\lambda - z)^{-1}| = |(\lambda - u) + iv|/((\lambda - u)^2 + v^2) \le 1/|\Im z|$. The Stieltjes transform uniquely determines any finite signed measure on $\mathbb{R}$, see the inversion formula in (Bai and Silverstein, 2010, Thm. B.8) to recover μ for $m_\mu(z)$. The following generalized version of Lemma A.4 then holds (cf. (Bai and Silverstein, 2010, Thm. 4.3 and Thm. 6.3)).

Lemma A.5 (Generalized Marchenko-Pastur and Bai-Yin laws) *Let $d/n \to \gamma \in (0, +\infty)$, Σ_n have bounded spectral norm as $n \to +\infty$, and assume $\nu_n \overset{d}{\to} \nu$ with bounded support. With probability one, the e.s.d. μ_n of $\widehat{S}$ converges weakly to μ determined by $m_\mu(z)$ from the following fixed-point equation on $\mathbb{C}_+$:*

$$m_\mu(z) = -\left(z - \int \frac{\lambda}{1 + \lambda m_\mu(z)\gamma^{-1}} \nu(\mathrm{d}\lambda)\right)^{-1}.$$

In addition when $\gamma > 1$, if the smallest and largest eigenvalues of Σ_n converge to the smallest and largest numbers in the support of ν, then λ_1 and λ_n converge almost surely to the smallest and largest numbers in the support of μ.

We can also efficiently compute the limit of λ_1 and λ_n by the following lemma (Silverstein and Choi, 1995, Thm. 4.2).

Lemma A.6 *Under the same setup of Lemma A.5 for $\gamma > 1$. Let λ_μ^+ and λ_μ^- be the largest and smallest numbers to satisfy the following equation*

$$\frac{1}{m_\mu(\lambda)^2} = \frac{1}{\gamma} \int \frac{\lambda^2}{(1 + (\lambda/\gamma) m_\mu(\lambda))^2} \nu(d\lambda),$$

in addition to the fixed-point equation in Lemma A.5 and $m_\mu(\lambda) < 0$. Then $\lambda_1 \to \lambda_\mu^+$ and $\lambda_n \to \lambda_\mu^-$ with probability one.

Appendix B. Proofs for Section 2

B.1. Proof of Proposition 1

For any estimator $\widehat{\theta}$,

$$\begin{aligned}\mathsf{Pred}_\Sigma(\widehat{\theta}) &= \mathbb{E}\left[\left\|\widehat{\theta}-\theta\right\|_\Sigma^2 \mid X\right] = \mathbb{E}\left[\left\|\widehat{\theta}-\widehat{\theta}_{\mathsf{B}}+\widehat{\theta}_{\mathsf{B}}-\theta\right\|_\Sigma^2 \mid X\right] \\ &= \mathbb{E}\left[\left\|\widehat{\theta}-\widehat{\theta}_{\mathsf{B}}\right\|_\Sigma^2 + \left\|\widehat{\theta}_{\mathsf{B}}-\theta\right\|_\Sigma^2 \mid X\right] = \mathsf{Pred}_\Sigma(\widehat{\theta}_{\mathsf{B}}) + \mathbb{E}\left[\left\|\widehat{\theta}-\widehat{\theta}_{\mathsf{B}}\right\|_\Sigma^2 \mid X\right],\end{aligned}$$

where the next-to-last step holds since $\widehat{\theta}_{\mathsf{B}} = \mathbb{E}\left[\theta \mid X, y\right]$ and so

$$\mathbb{E}\left[(\widehat{\theta}-\widehat{\theta}_{\mathsf{B}})^\top \Sigma(\widehat{\theta}_{\mathsf{B}}-\theta) \mid X, y\right] = (\widehat{\theta}-\widehat{\theta}_{\mathsf{B}})^\top \Sigma \mathbb{E}\left[\widehat{\theta}_{\mathsf{B}}-\theta \mid X, y\right] = 0.$$

This verifies the equality $\mathsf{Pred}_\Sigma(\widehat{\theta}) - \mathsf{Pred}_\Sigma^* = \mathbb{E}\left[\left\|\widehat{\theta}-\widehat{\theta}_{\mathsf{B}}\right\|_\Sigma^2 \mid X\right]$. On the other hand, applying the Minkowski inequality in $\mathcal{L}^2$ yields

$$\begin{aligned}\left|\sqrt{\mathsf{Train}(\widehat{\theta})} - \sqrt{\mathsf{Train}(\widehat{\theta}_{\mathsf{B}})}\right| &= \left|\sqrt{\frac{1}{n}\mathbb{E}\left[\left\|X\widehat{\theta}-y\right\|^2 \mid X\right]} - \sqrt{\frac{1}{n}\mathbb{E}\left[\left\|X\widehat{\theta}_{\mathsf{B}}-y\right\|^2 \mid X\right]}\right| \\ &\le \sqrt{\frac{1}{n}\mathbb{E}\left[\left\|X(\widehat{\theta}-\widehat{\theta}_{\mathsf{B}})\right\|^2 \mid X\right]} \le \left\|\Sigma^{-\frac{1}{2}} \cdot \frac{X^\top X}{n} \cdot \Sigma^{-\frac{1}{2}}\right\|^{\frac{1}{2}} \cdot \sqrt{\mathbb{E}\left[\left\|\theta-\widehat{\theta}_{\mathsf{B}}\right\|_\Sigma^2 \mid X\right]}.\end{aligned}$$

B.2. Proof of Theorem 2

We divide our proofs into three parts, addressing the general bound in Eq. (2), the low noise asymptotic expansion in Eq. (3) and the high noise expansion in Eq. (4) respectively. The proofs, especially the first two parts, rely heavily on several remarkable results from the theory of information inequalities. We first refer the reader to the necessary backgrounds in Sec. 4.1 in which we define the important auxiliary functions $\mathsf{J}(\sigma^2)$ and $\mathsf{Train}(\sigma^2)$ such that $\mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) = \mathsf{J}(\sigma^2) = \sigma^4\mathsf{Train}(\sigma^2)$ and importantly $\mathsf{J}(0) = \mathsf{J}_\pi$, whose continuity $\mathsf{J}(0) = \mathsf{J}(0+)$ will be a main focus of our proof. In this proof, we will work under the notation that $t = \sigma^2$. We shall also use π' to denote the probability distribution of $X\theta$ on $\mathbb{R}^n$ induced by π.

Part I: The general lower and upper bounds. The Fisher information upper bound is a direct corollary of Stam's Fisher information inequality in Lemma A.2, applying to $\pi'_t = \pi' * \mathsf{N}(0, tI_n)$ and implying $\mathsf{J}(t)^{-1} \ge \mathsf{J}(0)^{-1} + t$. Therefore,

$$\mathsf{Train}(t) = t^2\mathsf{J}(t) \le \frac{t^2}{\mathsf{J}(0)^{-1}+t} = \frac{t^2}{\mathsf{J}_\pi^{-1}+t}.$$

The lower bound follows from the classical scalar version of Cramer-Rao bound (Lehmann and Casella, 1998, Chp. 2). We provide a simple two-line proof. Since by Cauchy-Schwarz and integration by parts,

$$\begin{aligned}(\mathsf{V}_{\pi'}+t)\mathsf{J}(t) &= \left(\frac{1}{n}\mathbb{E}_{y\sim\pi'_t}\left[\|y\|^2\right]\right)\cdot\left(\frac{1}{n}\mathbb{E}_{y\sim\pi'_t}\left[\|\nabla\log p_t(y)\|^2\right]\right) \ge \left(\frac{1}{n}\int y\cdot\nabla\log p_t(y)p_t(y)\mathrm{d}y\right)^2 \\ &= \left(-\frac{1}{n}\int \underbrace{\nabla\cdot y}_{=n}\, p_t(y)\mathrm{d}y\right)^2 = \left(\int p_t(y)\mathrm{d}y\right)^2 = 1.\end{aligned}$$

Thus we establish the lower bound:

$$\mathsf{Train}(t) = t^2\mathsf{J}(t) \geq \frac{t^2}{\mathsf{V}_\pi + t}.$$

Part II: Asymptotic expansion when $\sigma^2 \to 0+$. Since $\mathsf{Train}(t) = t^2\mathsf{J}(t)$, we only need to show the continuity of $\mathsf{J}(t)$ at $0+$, i.e.

$$\lim_{t\to 0+} \mathsf{J}(t) = \lim_{\sigma^2\to 0+} \frac{1}{n}\cdot \mathbb{E}_{y\sim\pi_t'}\left[\|\nabla \log p_t(y)\|^2\right] = \frac{1}{n}\cdot \mathbb{E}_{y\sim\pi'}\left[\|\nabla \log p(y)\|^2\right] = \mathsf{J}(0).$$

The right continuity of J at 0, seemingly trivial at first glance, indeed requires deep technical tools from information theory. Since $p_t = p * \phi_t$, we first use Stam's Fisher information inequality in Lemma A.2, which implies $\mathsf{J}(t)^{-1} \geq \mathsf{J}(0)^{-1} + t$ and

$$\limsup_{t\to 0+} \mathsf{J}(t) \leq \mathsf{J}(0). \tag{13}$$

It boils down to proving $\liminf_{t\to 0+} \mathsf{J}(t) \geq \mathsf{J}(0)$. Since π_t' converges weakly to π' trivially from $X\theta + \sqrt{t}\tau \xrightarrow{d} X\theta$, the desired inequality holds by the lower semi-continuity of Fisher information under weak convergence. For scalar random variables, this is known in (Bobkov et al., 2014, Prop. 3.1) and (Huber and Ronchetti, 2009, P. 78) by the convexity of Fisher information under vague topology. This is indeed true in general dimensions, and we provide a proof in what follows for completeness, using the similar variational principle in Huber and Ronchetti (2009) and (Artstein et al., 2004, Thm. 4). Let $p_0 = p$, we have the following lemma.

Lemma B.1 (Variational principle of the Fisher information) *Let $\mathcal{C}_c^\infty(\mathbb{R}^n;\mathbb{R}^n)$ be the family of all compactly supported and smooth vector fields from $\mathbb{R}^n$ to $\mathbb{R}^n$, then for all $t \geq 0$,*

$$\mathsf{J}(t) = \sup_{b\in\mathcal{C}_c^\infty(\mathbb{R}^n;\mathbb{R}^n)} \left\{-\frac{2}{n}\int p_t(y)\nabla\cdot b(y)\mathrm{d}y - \frac{1}{n}\int p_t(y)\,\|b(y)\|^2\,\mathrm{d}y\right\}.$$

Proof By nonnegativity

$$\frac{1}{n}\int p(y)\,\|b(y) - \nabla\log p_t(y)\|^2\,\mathrm{d}y = \epsilon_b \geq 0,$$

we immediately have $\mathsf{J}(t) \geq \epsilon_b + \frac{2}{n}\int p_t(y)b(y)\cdot\nabla\log p_t(y)\mathrm{d}y - \frac{1}{n}\int p_t(y)\,\|b(y)\|^2\,\mathrm{d}y$. While b is smooth and compactly supported, we can integrate by parts and obtain

$$\frac{2}{n}\int p_t(y)b(y)\cdot\nabla\log p_t(y)\mathrm{d}y = \frac{2}{n}\int b(y)\cdot\nabla p_t(y)\mathrm{d}y = -\frac{2}{n}\int p_t(y)\nabla\cdot b(y)\mathrm{d}y.$$

Here $\nabla\cdot b(y) := \sum_{i=1}^n \frac{\partial}{\partial y_i} b(y)$ denotes the divergence of $b(y)$. Thus taking supremum over all $b \in \mathcal{C}_c^\infty(\mathbb{R}^n;\mathbb{R}^n)$,

$$\begin{aligned}\mathsf{J}(t) &\geq \sup_{b\in\mathcal{C}_c^\infty(\mathbb{R}^n;\mathbb{R}^n)} \left\{\epsilon_b - \frac{2}{n}\int p_t(y)\nabla\cdot b(y)\mathrm{d}y - \frac{1}{n}\int p_t(y)\,\|b(y)\|^2\,\mathrm{d}y\right\}\\ &\geq \sup_{b\in\mathcal{C}_c^\infty(\mathbb{R}^n;\mathbb{R}^n)} \left\{-\frac{2}{n}\int p_t(y)\nabla\cdot b(y)\mathrm{d}y - \frac{1}{n}\int p_t(y)\,\|b(y)\|^2\,\mathrm{d}y\right\}.\end{aligned}$$

The equality holds since ϵ_b can be arbitrarily small. ■

We are now ready to show lower semi-continuity using Lemma B.1. Indeed, since $p_t(y)\mathrm{d}y \xrightarrow{d} p(y)\mathrm{d}y$, by weak convergence, for any compactly supported smooth b,

$$\begin{aligned}&\lim_{t\to 0+}\left(-\frac{2}{n}\int p_t(y)\nabla\cdot b(y)\mathrm{d}y-\frac{1}{n}\int p_t(y)\left\|b(y)\right\|^2\mathrm{d}y\right)\\&=-\frac{2}{n}\int p(y)\nabla\cdot b(y)\mathrm{d}y-\frac{1}{n}\int p(y)\left\|b(y)\right\|^2\mathrm{d}y.\end{aligned}$$

Thus

$$\liminf_{t\to 0+}\mathsf{J}(t)\geq-\frac{2}{n}\int p(y)\nabla\cdot b(y)\mathrm{d}y-\frac{1}{n}\int p(y)\left\|b(y)\right\|^2\mathrm{d}y.$$

Taking supremum on both sides for $b\in\mathcal{C}_c^\infty(\mathbb{R}^n;\mathbb{R}^n)$ yields

$$\liminf_{t\to 0+}\mathsf{J}(t)\geq\mathsf{J}(0). \tag{14}$$

We show $\lim_{t\to 0+}\mathsf{J}(t)=\mathsf{J}(0)$ combining Eqs. (13) and (14).

Part III: Asymptotic expansion when $\sigma^2\to+\infty$. By the definition of $\widehat{\theta}_{\mathsf{B}}=\mathbb{E}[\theta\mid X,y]$, we can make use of the following Pythagorean theorems

$$\begin{aligned}\mathbb{E}_{y\sim\pi_t'}\left[\left\|X\widehat{\theta}_{\mathsf{B}}-y\right\|^2\mid X\right]&=\mathbb{E}_{y\sim\pi_t'}\left[\mathbb{E}_{\theta|X,y}\left[\left\|X\theta-y\right\|^2-\left\|X(\theta-\widehat{\theta}_{\mathsf{B}})\right\|^2\mid X,y\right]\right],\\\mathbb{E}_{\theta|X,y}\left[\left\|X(\theta-\widehat{\theta}_{\mathsf{B}})\right\|^2\mid X,y\right]&=\mathbb{E}_{\theta|X,y}\left[\left\|X\theta\right\|^2\mid X,y\right]-\left\|X\widehat{\theta}_{\mathsf{B}}\right\|^2,\end{aligned}$$

and obtain that

$$\begin{aligned}\mathsf{Train}(t)&=\frac{1}{n}\mathbb{E}_{y\sim\pi_t'}\left[\left\|X\widehat{\theta}_{\mathsf{B}}-y\right\|^2\mid X\right]\\&=\frac{1}{n}\mathbb{E}_{y\sim\pi_t'}\left[\mathbb{E}_{\theta|X,y}\left[\left\|X\theta-y\right\|^2-\left\|X\theta\right\|^2+\left\|X\widehat{\theta}_{\mathsf{B}}\right\|^2\mid X,y\right]\right]\\&=\frac{1}{n}\mathbb{E}_{\theta\sim\pi,\tau\sim\mathsf{N}(0,I_n),y=X\theta+\sigma\tau}\left[\left\|X\theta-y\right\|^2\right]-\frac{1}{n}\mathbb{E}_{\theta\sim\pi}\left[\left\|X\theta\right\|^2\mid X\right]+\frac{1}{n}\mathbb{E}_{y\sim\pi_t'}\left[\left\|X\widehat{\theta}_{\mathsf{B}}\right\|^2\mid X\right]\\&=t-\frac{1}{n}\mathbb{E}_{\theta\sim\pi,\tau\sim\mathsf{N}(0,I_n),y=X\theta+\sigma\tau}\left[\left\|X(\theta-\widehat{\theta}_{\mathsf{B}})\right\|^2\mid X\right]\\&=t-\frac{1}{n}\cdot\mathbb{E}\left[\left\|X\theta\right\|^2\mid X\right]+\frac{1}{n}\mathbb{E}\left[\left\|X\widehat{\theta}_{\mathsf{B}}\right\|^2\mid X\right].\end{aligned} \tag{15}$$

In the preceding displays, we marginalize over y and θ, using $y=X\theta+\sigma\tau$. The last two lines naturally give general lower and upper bounds as

$$t-\mathsf{V}_\pi=t-\frac{1}{n}\cdot\mathbb{E}\left[\left\|X\theta\right\|^2\mid X\right]\leq\mathsf{Train}(t)\leq t.$$

Next, we derive the high noise asymptotics when $t\to+\infty$ based on the preceding displays. Recall that

$$\widehat{\theta}_{\mathsf{B}}=\mathbb{E}\left[\theta\mid X,y\right]=\frac{\int\theta\exp\left\{-\frac{\left\|X\theta-y\right\|^2}{2t}\right\}\pi(\mathrm{d}\theta)}{\int\exp\left\{-\frac{\left\|X\theta-y\right\|^2}{2t}\right\}\pi(\mathrm{d}\theta)}.$$

We can then compute that

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\theta}_{\mathsf{B}}\widehat{\theta}_{\mathsf{B}}^{\top}\right] = \mathbb{E}_{y\sim\pi_t'}\mathbb{E}\left[\widehat{\theta}_{\mathsf{B}}\widehat{\theta}_{\mathsf{B}}^{\top} \mid X, y\right] \\
&= \mathbb{E}_{y'\sim\pi', \tau\sim\mathsf{N}(0,I_n)}\left[\frac{\iint \theta\tilde{\theta}^{\top} \exp\left\{-\frac{\|X\theta - y' - \sqrt{t}\tau\|^2}{2t} - \frac{\|X\tilde{\theta} - y' - \sqrt{t}\tau\|^2}{2t}\right\}\pi(\mathrm{d}\theta)\pi(\mathrm{d}\tilde{\theta})}{\left(\int \exp\left\{-\frac{\|X\theta - y' - \sqrt{t}\tau\|^2}{2t}\right\}\pi(\mathrm{d}\theta)\right)^2} \;\middle|\; X\right].
\end{aligned}$$

Since $\int \|\theta\|^2 \pi(d\theta) = \mathbb{E}[\|\theta\|^2] < +\infty$, we can apply dominated convergence theorem to take limit under the integral

$$\begin{aligned}
&\lim_{t\to+\infty} \mathbb{E}_{y\sim\pi_t'}\mathbb{E}\left[\widehat{\theta}_{\mathsf{B}}\widehat{\theta}_{\mathsf{B}}^{\top} \mid X, y\right] \\
&= \mathbb{E}_{y'\sim\pi', \tau\sim\mathsf{N}(0,I_n)}\left[\frac{\iint \theta\tilde{\theta}^{\top} \cdot \lim_{t\to+\infty} \exp\left\{-\frac{\|X\theta - y' - \sqrt{t}\tau\|^2}{2t} - \frac{\|X\tilde{\theta} - y' - \sqrt{t}\tau\|^2}{2t}\right\}\pi(\mathrm{d}\theta)\pi(\mathrm{d}\tilde{\theta})}{\left(\int \lim_{t\to+\infty} \exp\left\{-\frac{\|X\theta - y' - \sqrt{t}\tau\|^2}{2t}\right\}\pi(\mathrm{d}\theta)\right)^2} \;\middle|\; X\right] \\
&= \mathbb{E}_{\tau\sim\mathsf{N}(0,I_n)}\left[\frac{\iint \theta\tilde{\theta}^{\top} \exp\left\{-\|\tau\|^2\right\}\pi(\mathrm{d}\theta)\pi(\mathrm{d}\tilde{\theta})}{\exp\left\{-\|\tau\|^2\right\}} \;\middle|\; X\right] = \mathbb{E}\left[\theta\right]\mathbb{E}\left[\theta\right]^{\top} = 0.
\end{aligned}$$

Thus, $\widehat{\theta}_{\mathsf{B}} \xrightarrow{\mathcal{L}^2} 0$ by taking trace on the above limit, and

$$\lim_{t\to+\infty} \frac{1}{n}\mathbb{E}_{y\sim\pi_t'}\left[\left\|X\widehat{\theta}_{\mathsf{B}}\right\|^2 \mid X\right] = \frac{1}{n}\operatorname{trace}\left(X \cdot \lim_{t\to+\infty} \mathbb{E}_{y\sim\pi_t'}\left[\widehat{\theta}_{\mathsf{B}}\widehat{\theta}_{\mathsf{B}}^{\top}\right] \cdot X^{\top}\right) = 0.$$

The last part of the proof is complete combining with Eq. (15).

B.3. Proof of Corollary 4

First, since $\frac{\sigma^4}{\mathsf{J}_\pi^{-1}+\sigma^2} \le \mathsf{J}_\pi\sigma^4$, we have

$$\mathsf{Train}(\widehat{\theta}) \ge \mathsf{J}_\pi\sigma^4 \implies \mathsf{Train}(\widehat{\theta}) \ge \frac{\sigma^4}{\mathsf{J}_\pi^{-1} + \sigma^2}.$$

Therefore, we can apply the bound (5) from Theorem 3. If $\mathsf{Train}(\widehat{\theta}) > \mathsf{J}_\pi\sigma^4$, then (5) immediately yields $\mathsf{Cost}_\Sigma(\widehat{\theta}) > 0$. Moreover, if $\mathsf{Train}(\widehat{\theta}) \ge C \cdot \mathsf{J}_\pi\sigma^4$ for some $C > 1$, then by (5) we have

$$\begin{aligned}
\mathsf{Cost}_\Sigma(\widehat{\theta}) &\ge \lambda_\Sigma^{-1}\left(\sqrt{\mathsf{Train}(\widehat{\theta})} - \sqrt{\frac{\sigma^4}{\mathsf{J}_\pi^{-1} + \sigma^2}}\right)^2 \\
&\ge \lambda_\Sigma^{-1}\left(\sqrt{\mathsf{Train}(\widehat{\theta})} - \sqrt{C^{-1} \cdot \mathsf{Train}(\widehat{\theta})}\right)^2 \quad \text{since } \frac{\sigma^4}{\mathsf{J}_\pi^{-1} + \sigma^2} \le \mathsf{J}_\pi\sigma^4 \le C^{-1} \cdot \mathsf{Train}(\widehat{\theta}) \\
&= \lambda_\Sigma^{-1}(1 - C^{-1/2})^2 \cdot \mathsf{Train}(\widehat{\theta}).
\end{aligned}$$

B.4. Proof of Corollary 5

First, since $\frac{\sigma^4}{\mathsf{V}_\pi+\sigma^2} \geq \sigma^2 - \mathsf{V}_\pi$, we have

$$\mathsf{Train}(\widehat{\theta}) \leq \sigma^2 - \mathsf{V}_\pi \implies \mathsf{Train}(\widehat{\theta}) \leq \frac{\sigma^4}{\mathsf{V}_\pi + \sigma^2}.$$

Therefore, we can apply the bound (6) from Theorem 3. If $\mathsf{Train}(\widehat{\theta}) < \sigma^2 - \mathsf{V}_\pi$, then (6) immediately yields $\mathsf{Cost}_\Sigma(\widehat{\theta}) > 0$. Moreover, if $\mathsf{Train}(\widehat{\theta}) \leq \sigma^2 - C \cdot \mathsf{V}_\pi$ for some $C > 1$, then by (6) we have

$$\begin{aligned}
\mathsf{Cost}_\Sigma(\widehat{\theta}) &\geq \lambda_\Sigma^{-1}\left(\sqrt{\frac{\sigma^4}{\mathsf{V}_\pi+\sigma^2}} - \sqrt{\mathsf{Train}(\widehat{\theta})}\right)^2 \\
&\geq \lambda_\Sigma^{-1}\left(\sqrt{\sigma^2-\mathsf{V}_\pi} - \sqrt{\mathsf{Train}(\widehat{\theta})}\right)^2 \quad \text{since } \frac{\sigma^4}{\mathsf{V}_\pi+\sigma^2} \geq \sigma^2 - \mathsf{V}_\pi \geq \mathsf{Train}(\widehat{\theta}) \\
&\geq \lambda_\Sigma^{-1}\left(\sqrt{\sigma^2-\mathsf{V}_\pi} - \sqrt{\sigma^2-\Delta}\right)^2 \quad \text{defining } \Delta = \sigma^2 - \mathsf{Train}(\widehat{\theta}) \\
&\geq \lambda_\Sigma^{-1}\left(\sqrt{\sigma^2-C^{-1}\cdot\Delta} - \sqrt{\sigma^2-\Delta}\right)^2 \quad \text{since } \mathsf{Train}(\widehat{\theta}) \leq \sigma^2 - C\cdot\mathsf{V}_\pi \text{ implies } \Delta \geq C\cdot\mathsf{V}_\pi \\
&\geq \lambda_\Sigma^{-1}\left(\frac{\Delta - C^{-1}\cdot\Delta}{2\sigma}\right)^2 \quad \text{since } \sqrt{\sigma^2-a} - \sqrt{\sigma^2-b} \geq \frac{b-a}{2\sigma} \text{ for } 0 \leq a \leq b \leq \sigma^2 \\
&= \frac{(4\lambda_\Sigma)^{-1}(1-C^{-1})^2}{\sigma^2}\cdot\Delta^2.
\end{aligned}$$

Appendix C. Proofs for Section 3

C.1. Proof for the isotropic Gaussian prior

We give the proof for the isotropic results first. Since $X\theta \mid X \sim \mathsf{N}(0, XX^\top/d)$, we can denote by the n eigenvalues of $\widehat{S} := XX^\top/d$ by $\lambda_1 \geq \cdots \lambda_n > 0$. The classical results in random matrix theory enable the following almost sure limits (cf. Lemma A.4 in Appendix A),

$$\begin{aligned}
&\lambda_1 \to \left(1+\gamma^{-\frac{1}{2}}\right)^2 =: \lambda_\gamma^+, \qquad \lambda_n \to \left(1-\gamma^{-\frac{1}{2}}\right)^2 =: \lambda_\gamma^-, \\
&\mu_n(\mathrm{d}\lambda) := \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i} \xrightarrow{d} \frac{\gamma}{2\pi}\cdot\frac{\sqrt{(\lambda_\gamma^+-\lambda)(\lambda-\lambda_\gamma^-)}}{\lambda}\mathbb{1}_{\lambda_\gamma^-\leq\lambda\leq\lambda_\gamma^+}\mathrm{d}\lambda =: \mu_{\mathsf{MP},\gamma}(\mathrm{d}\lambda).
\end{aligned}$$

Since $\Sigma = I$, the parameter λ_Σ in Proposition 1 then has the exact limit $\lambda_\Sigma = \gamma\lambda_1 \to (1+\sqrt{\gamma})^2$. For $\mathsf{V}_\pi = \frac{1}{nd}\operatorname{trace}(XX^\top) = \frac{1}{nd}\sum_{i=1}^n\sum_{j=1}^d X_{ij}^2$, by law of large numbers we have $\mathsf{V}_\pi \to 1$ with probability one. Alternatively, we can also make use of the semicircle integral $\int_a^b \sqrt{(b-\lambda)(\lambda-a)}\mathrm{d}\lambda = \frac{\pi(b-a)^2}{8}$ to confirm that

$$\begin{aligned}
\mathsf{V}_\pi &= \frac{1}{n}\operatorname{trace}\left(\widehat{S}\right) = \int \lambda\mu_n(\mathrm{d}\lambda) \to \int \lambda\mu_{\mathsf{MP},\gamma}(\mathrm{d}\lambda) = \int \frac{\gamma}{2\pi}\cdot\sqrt{(\lambda_\gamma^+-\lambda)(\lambda-\lambda_\gamma^-)}\mathbb{1}_{\lambda_\gamma^-\leq\lambda\leq\lambda_\gamma^+}\mathrm{d}\lambda \\
&= \frac{\gamma}{2\pi}\cdot\frac{\pi\left((1+\gamma^{-\frac{1}{2}})^2-(1-\gamma^{-\frac{1}{2}})^2\right)}{8} = 1.
\end{aligned}$$

Finally, since $\pi' = \mathsf{N}(0, \widehat{S})$, we can compute

$$\begin{aligned}\mathsf{J}_\pi &= \mathbb{E}\left[\frac{1}{n}\left\|\nabla \log p(X\theta)\right\|^2 \mid X\right] = -\mathbb{E}\left[\frac{1}{n}\Delta \log p(X\theta) \mid X\right] = \frac{1}{n}\operatorname{trace}\left(\widehat{S}^{-1}\right) \\ &= \int \frac{1}{\lambda}\mu_n(\mathrm{d}\lambda) \to \int \frac{1}{\lambda}\mu_{\mathsf{MP},\gamma}(\mathrm{d}\lambda) = \frac{\gamma}{\gamma-1}.\end{aligned}$$

The last equality is from the resolvent identity at $z \to i \cdot 0+$ (Bai and Silverstein, 2010, Lemma 3.11).

For the general sequence of $\Sigma(n)$, we refer the reader for backgrounds in Stieltjes transform and random matrix theory to Lemmas A.5 and A.6 that are necessary for the general version of Proposition 7 below.

Proposition 13 *Let Assumption 1. Under the setups of Lemmas A.5 and A.6, for $\sigma^2 > 0$, we have almost surely*

$$\mathsf{V}_\pi \to \int \lambda\mu(\mathrm{d}\lambda), \qquad \mathsf{J}_\pi \to \int \frac{1}{\lambda}\mu(\mathrm{d}\lambda) \qquad \lambda_I \to \gamma\lambda_\mu^+,$$

where μ is the generalized Marchenko-Pastur limit for the sequences of covariance matrices $\Sigma(n)$ whose empirical spectrum distributions ν_n converge to some ν weakly.

The proof for Prop. 13, is identical to the preceding proof for $\Sigma = I$, and thus we omit the repetitive details. For λ_Σ, as it equals to $\frac{1}{n}\left\|\Sigma^{-\frac{1}{2}} \cdot X^\top X \Sigma^{-\frac{1}{2}}\right\| = \frac{1}{n}\left\|Z^\top Z\right\|$, the isotropic result applies and $\lambda_\Sigma \to (1+\sqrt{\gamma})^2$.

C.2. Proofs for the approximately-low-rank Gaussian prior

C.2.1. PROOF OF PROPOSITION 8

Part I: Limits of the parameters. Let $\widehat{S} = X\Omega X^\top$ and its e.s.d. be μ_n. Since $X\theta \mid X \sim \mathsf{N}(0, \widehat{S})$, the convergence of V_π follows from law of large numbers and

$$\mathbb{E}_X\left[\mathsf{V}_\pi\right] = \frac{1}{n}\operatorname{trace}(\mathbb{E}_X\left[\widehat{S}\right]) = \frac{1}{d}\operatorname{trace}(\Sigma) = 1.$$

As $\lambda_\Sigma = \frac{1}{n}\left\|X^\top X\right\|$, the same limit as in Proposition 7 hold. It remains only nontrivial to compute J_π. Denote by the e.s.d. of $d\Omega_n$ by ν_n, it is clear that

$$\nu_n \xrightarrow{d} \nu := \frac{\rho}{\gamma}\delta_{\gamma/\rho+\eta} + \left(1 - \frac{\rho}{\gamma}\right)\delta_\eta.$$

We can then apply Lemma A.5 since $\widehat{S} = X(d\Omega)X^\top/d$, which implies

$$\frac{1}{m_\mu} = -z + \int \frac{\lambda}{1+\lambda m_\mu \gamma^{-1}}\nu(\mathrm{d}\lambda) = -z + \frac{1+\eta\rho\gamma^{-1}}{1+m_\mu(\rho^{-1}+\eta\rho^{-1})} + \frac{\eta-\eta\rho\gamma^{-1}}{1+m_\mu\eta\gamma^{-1}}.$$

Set $z = 0$, this solves the quadratic equation

$$\eta(\gamma-1)m_\mu^2 + \underbrace{\left[\eta^2(\gamma-\rho-1) + \eta\left(\gamma\rho-\rho^2-\rho\right) + \gamma(\rho-1)\right]}_{=C} m_\mu - \gamma(\eta+\rho) = 0,$$

which yields for the positive branch solution:

$$m_\mu(0) = \frac{-C + \sqrt{C^2 + 4\eta(\gamma-1)\gamma(\eta+\rho)}}{2\eta(\gamma-1)}.$$

We thus have the explicit form for the limit of J_π by identifying

$$\mathsf{J}_\pi = -\mathbb{E}\left[\frac{1}{n}\Delta \log p(X\theta) \mid X\right] = \frac{1}{n}\operatorname{trace}\left(\widehat{S}^{-1}\right) = \int \frac{1}{\lambda}\mu_n(\mathrm{d}\lambda) \to \int \frac{1}{\lambda}\mu(\mathrm{d}\lambda) = m_\mu(0).$$

In this step we invoke Lemma A.5 implicitly, which allows us to restrict ourselves to the compact interval $[\lambda_\mu^- - \epsilon, \lambda_\mu^+ + \epsilon]$, since the extreme eigenvalues converge almost surely. Consequently, the function $\lambda \mapsto 1/\lambda$ is bounded on this interval for a suitable choice of ϵ, and thus weak convergence applies for the function $1/\lambda$.

Part II: Asymptotics for $\eta \to 0+$. We compute the asymptotics for $m_\mu(0)$ as the limit of J_π.

Case I: $\rho > 1$**.** In this case $C(\gamma, \rho, 0+)$ has a positive limit $\gamma(\rho-1)$, and thus

$$\begin{aligned} m_\mu(0) &= \frac{-C + \sqrt{C^2 + 4\eta(\gamma-1)\gamma(\eta+\rho)}}{2\eta(\gamma-1)} = \frac{2\gamma(\eta+\rho)}{C + \sqrt{C^2 + 4\eta(\gamma-1)\gamma(\eta+\rho)}} \\ &\to \frac{\gamma\rho}{C(\gamma,\rho,0+)} = \frac{\rho}{\rho-1}. \end{aligned}$$

Case II: $\rho = 1$**.** When $\rho = 1$, taking $C(\gamma, 1, \eta) = \eta(\gamma-2) + o(\eta)$ into the above display gives

$$m_\mu(0) = \frac{2\gamma\rho + o(\eta)}{o(\eta^{\frac{1}{2}}) + \sqrt{4\eta(\gamma-1)\gamma\rho + o(\eta)}} = \sqrt{\frac{\gamma}{\eta(\gamma-1)}} + o(\eta^{-\frac{1}{2}}).$$

Case III: $0 < \rho < 1$**.** In this case $C(\gamma, \rho, 0+)$ has a negative limit $\gamma(\rho-1)$, and thus

$$m_\mu(0) = \frac{-C + \sqrt{C^2 + 4\eta(\gamma-1)\gamma(\eta+\rho)}}{2\eta(\gamma-1)} = \frac{2\gamma(1-\rho) + o(\eta)}{2\eta(\gamma-1)} = \frac{\gamma(1-\rho)}{\eta(\gamma-1)} + o(\eta^{-1}).$$

The proof is complete.

C.2.2. PROOF OF PROPOSITION 9

We divide this section into two parts. The main proof of the proposition is in only Part I. In Part II, we analyze the training error for $\eta > 0$ and derive asymptotic solutions under the scaling regime $\sigma = C\sqrt{\eta}$, for some constant C, in the limit as $\eta \to 0+$.

Part I: The unperturbed prior $\eta = 0$. We begin by examining the case $\eta = 0$, as it offers conceptual clarity while avoiding substantial technical complications. It is equivalent to having $X \in \mathbb{R}^{n\times r}$, $r/n \to \rho$, $\Sigma = I_r$ and $\theta \sim \mathsf{N}(0, I_r/r)$. The posterior mean (i.e. the Bayes estimator) for $y \sim \mathsf{N}(0, XX^\top/r + \sigma^2 I_r)$ is then $\widehat{\theta}_{\mathsf{B}} = (X^\top X + \sigma^2 r I_r)^{-1} X^\top y$, and therefore

$$\begin{aligned} \mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) &= \mathbb{E}\left[\frac{1}{n}\left\|X\widehat{\theta}_{\mathsf{B}} - y\right\|^2\right] \\ &= \frac{1}{n}\mathbb{E}\left[\left\|X(I - (X^\top X + \sigma^2 r I_r)^{-1} X^\top X)\theta\right\|^2\right] + \frac{1}{n}\mathbb{E}\left[\left\|\left(I - X(X^\top X + \sigma^2 r I_r)^{-1} X^\top\right)\epsilon\right\|^2\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{\sigma^4}{n}\operatorname{trace}\left((XX^\top/r+\sigma^2 I_n)^{-2}(XX^\top/r)\right) + \frac{\sigma^6}{n}\operatorname{trace}\left((XX^\top/r+\sigma^2 I_n)^{-2}\right)\\
&= \int \frac{\sigma^4\lambda}{(\lambda+\sigma^2)^2} + \frac{\sigma^6}{(\lambda+\sigma^2)^2}\mu_n(\mathrm{d}\lambda)\\
&\to \int \frac{\sigma^4}{\lambda+\sigma^2}\mu_{\mathsf{MP},\rho}(\mathrm{d}\lambda)\\
&= \sigma^4 m_{\mu_{\mathsf{MP},\rho}}(-\sigma^2) = \sigma^4\cdot\frac{-\left(\sigma^2\rho-\rho+1\right)+\sqrt{\left(\sigma^2\rho-\rho+1\right)^2-4\sigma^2\rho}}{2\sigma^2}.
\end{aligned}
$$

In the last line we use Lemma A.5 for $\nu = \delta_1$ and $z = -\sigma^2$. As it holds for $m = m_{\mu_{\mathsf{MP},\rho}}(-\sigma^2)$ that

$$
\frac{1}{m} = \sigma^2 + \frac{1}{1+m/\rho}.
$$

Using the asymptotics for $\rho > 1$ that

$$
\begin{aligned}
\frac{-\left(\sigma^2\rho+\rho-1\right)+\sqrt{\left(\sigma^2\rho+\rho-1\right)^2+4\sigma^2\rho}}{2\sigma^2} &= \frac{4\sigma^2\rho}{2\sigma^2\cdot\left(\left(\sigma^2\rho+\rho-1\right)+\sqrt{\left(\sigma^2\rho+\rho-1\right)^2+4\sigma^2\rho}\right)}\\
&= \frac{\rho}{\rho-1}+O(\sigma^2),
\end{aligned}
$$

and for $\rho < 1$ that

$$
\begin{aligned}
\frac{-\left(\sigma^2\rho+\rho-1\right)+\sqrt{\left(\sigma^2\rho+\rho-1\right)^2+4\sigma^2\rho}}{2\sigma^2} &= \frac{2(1-\rho)-\sigma^2\rho+\frac{\rho(1+\rho)}{1-\rho}\sigma^2+O(\sigma^4)}{2\sigma^2}\\
&= \frac{1-\rho}{\sigma^2}+\frac{\rho^2}{1-\rho}+O(\sigma^2).
\end{aligned}
$$

we complete the proof for $\eta = 0$.

Part II: The general training error when $\eta > 0$. Similar to the previous calculations by considering the exact Bayes estimator from the posterior mean of $y \sim \mathsf{N}(0, X\Omega X^\top + \sigma^2 I_d)$

$$
\widehat{\theta}_{\mathsf{B}} = \Omega X^\top (X\Omega X^\top + \sigma^2 I_d)^{-1} y.
$$

Let μ_n be the empirical spectrum distribution of $X\Omega X^\top$, we then have the exact formula for the Bayes error as (using the same calculations in the first part):

$$
\mathsf{Train}(\widehat{\theta}_{\mathsf{B}}) = \int \frac{\sigma^4}{\lambda+\sigma^2}\mu_n(\mathrm{d}\lambda) \to \int \frac{\sigma^4}{\lambda+\sigma^2}\mu(\mathrm{d}\lambda) = \sigma^4 m_\mu(-\sigma^2),
$$

where μ is the generalized Marchenko-Pastur limit by Lemma A.5 and $m_\mu(z)$ is its associated Stieltjes transform. Substituting $z = -\sigma^2$ into Lemma A.5 yields for $m_\eta = m_\mu(-\sigma^2)$,

$$
\frac{1}{m_\eta} = \sigma^2 + \frac{1+\eta\rho\gamma^{-1}}{1+m_\eta(\rho^{-1}+\eta\rho^{-1})} + \frac{\eta-\eta\rho\gamma^{-1}}{1+m_\eta\eta\gamma^{-1}}.
$$

When $\rho < 1$, we substitute in the ansatz $m_\eta = C'/\eta + o(\eta^{-1})$ under the noise asymptotics $\sigma = C\sqrt{\eta}$, the equation simplifies into

$$\frac{1-\rho}{C'} = C^2 + \frac{\gamma-\rho}{\gamma + C'},$$

which admits a positive solution $C' = C'(C)$ with $C'(0) = \frac{\gamma(1-\rho)}{\gamma-1}$.

C.3. Proof for the mixture of approximately-sparse priors

We provide the proofs in three parts. In the first part we upper bound the Fisher information by convexity. In the second part we utilize the isoperimetric lower bound from Lemma A.3.

Part I: Upper bound by convexity of Fisher information. For a mixture of density $p(y) = \sum_{i=1}^m \alpha_i p_i(y)$ where Fisher information exists for each individual p_i. We can upper bound pointwise by Cauchy-Schwarz inequality:

$$\|\nabla p(y)\|^2 = \left\|\sum_{i=1}^m \alpha_i \nabla p_i(y)\right\|^2 \le \left(\sum_{i=1}^m \alpha_i p_i(y)\right)\cdot\left(\sum_{i=1}^m \alpha_i \frac{\|\nabla p_i(y)\|^2}{p_i(y)}\right) = p(y)\cdot\left(\sum_{i=1}^m \alpha_i \frac{\|\nabla p_i(y)\|^2}{p_i(y)}\right).$$

Apply the pointwise bound to the mixture considered in the example with $m = \binom{d}{K}$, $\alpha_S = \binom{d}{K}^{-1}$ and $p_S = \phi(y; 0, X\Omega_S X^\top)$ and we have

$$\frac{\|\nabla p(y)\|^2}{p(y)} \le \sum_S \alpha_S \frac{\|\nabla p_S(y)\|^2}{p_S(y)} = \frac{1}{\binom{d}{K}}\sum_S \frac{\|\nabla \phi(y;0,X\Omega_S X^\top)\|^2}{\phi(y;0,X\Omega_S X^\top)}.$$

Integrating over y yields

$$\begin{aligned}
\mathsf{J}_\pi &= \int \frac{\|\nabla p(y)\|^2}{p(y)}\mathrm{d}y \le \int \frac{1}{\binom{d}{K}}\sum_S \frac{\|\nabla \phi(y;0,X\Omega_S X^\top)\|^2}{\phi(y;0,X\Omega_S X^\top)}\mathrm{d}y \\
&= \frac{1}{\binom{d}{K}}\sum_S \int \left\|\nabla \log \phi(y;0,X\Omega_S X^\top)\right\|^2 \phi(y;0,X\Omega_S X^\top)\mathrm{d}y \\
&= \frac{1}{\binom{d}{K}}\sum_S \mathbb{E}_{y\sim \mathsf{N}(0,X\Omega_S X^\top)}\left[\left\|\nabla \log \phi(y;0,X\Omega_S X^\top)\right\|^2\right] \\
&= \frac{1}{\binom{d}{K}}\sum_S \operatorname{trace}\left(\left(X\Omega_S X^\top\right)^{-1}\right) \le \frac{1}{\eta}\operatorname{trace}\left((XX^\top/d)^{-1}\right).
\end{aligned}$$

In the last inequality, we use the uniform lower bound $\Omega_S \succeq \eta I_d/d$. Let μ_n be the e.s.d. of $XX^\top/d$, $\mu_n \xrightarrow{d} \mu_{\mathsf{MP},\gamma}$ by Lemma A.4, and by the resolvent identity at $z \to i\cdot 0+$ in (Bai and Silverstein, 2010, Lemma 3.11) we conclude

$$\mathsf{J}_\pi \le \frac{1}{\eta}\operatorname{trace}\left((XX^\top/d)^{-1}\right) = \frac{1}{\eta}\int \frac{1}{\lambda}\mu_n(\mathrm{d}\lambda) \to \frac{1}{\eta}\int \frac{1}{\lambda}\mu_{\mathsf{MP},\gamma}(\mathrm{d}\lambda) \to \frac{\gamma}{\eta(\gamma-1)}.$$

We thus obtain the upper bound

$$\limsup_{n\to+\infty} \mathsf{J}_\pi \le \frac{\gamma}{\eta(\gamma-1)}.$$

Part II: Lower bound by entropic isoperimetric inequality. We begin by introducing the necessary notation that we will leverage to derive upper and lower bounds for J_π. Define the differential entropy for Ent_π for $y = X\theta$ by $\mathsf{Ent}_\pi = \mathbb{E}\left[-\log p(X\theta)\right] > 0$ (by Jensen's inequality). With a slight abuse of notations, we also denote S by the discrete random variable uniformly drawn from all $\binom{d}{k}$ subsets of size K of $[d]$. Then we can view the distribution of y as drawing S first and sampling $y \mid S \sim \mathsf{N}(0, X\Omega_S X^\top)$. We can then define the conditional entropy and mutual information as follows:

$$\mathsf{Ent}_{\pi|S} = \mathbb{E}_S\left[\mathbb{E}_{y\sim\pi|S}\left[-\log\phi(y \mid S; 0, X\Omega_S X^\top)\right]\right], \qquad \mathscr{I}(y;S) = \mathsf{Ent}_\pi - \mathsf{Ent}_{\pi|S}.$$

Combine the following explicit formula for the differential entropy for multivariate Gaussians, and the familiar entropy bound for the mutual information of a discrete random variable,

$$\mathsf{Ent}_{\pi|S} = \frac{1}{\binom{d}{K}}\sum_S\left(\frac{n}{2}\log 2\pi + \frac{n}{2} + \frac{1}{2}\det X\Omega_S X^\top\right) = \frac{n}{2} + \frac{n}{2}\log 2\pi + \frac{1}{2}\frac{1}{\binom{d}{K}}\sum_S \log\det\left(X\Omega_S X^\top\right),$$

$$\mathscr{I}(y;S) \le \mathscr{H}(S) = \log\binom{d}{K} \overset{\text{(i)}}{\le} K\log\frac{d}{K} + K,$$

where we use the Stirling's bound $K! \ge (K/e)^K$ in (i); we can thus upper bound

$$\begin{aligned}\mathsf{Ent}_\pi &= \mathsf{Ent}_{\pi|S} + \mathscr{I}(y;S)\\ &\le \frac{n}{2} + \frac{n}{2}\log 2\pi + K\log\frac{d}{K} + K + \frac{1}{2}\frac{1}{\binom{d}{K}}\sum_S\log\det\left(X\Omega_S X^\top\right).\end{aligned}$$

We are now ready to apply the entropic isoperimetric inequality in Lemma A.3 on J_π, which implies

$$\mathsf{J}_\pi = \frac{1}{n}\cdot\mathbb{E}\left[\|\nabla\log p(y)\|^2\right] \ge 2\pi e\cdot\exp\left\{-\frac{2}{n}\mathbb{E}\left[-\log p(y)\right]\right\} = \exp\left\{\log(2\pi) + 1 - \frac{2}{n}\mathsf{Ent}_\pi\right\}.$$

Comibining with the entropy bound yields

$$\mathsf{J}_\pi \ge \exp\left\{-\frac{2K}{n}\log\frac{d}{K} - \frac{2K}{n} - \frac{1}{\binom{d}{K}}\sum_S\frac{1}{n}\log\det(X\Omega_S X^\top)\right\}. \tag{16}$$

Utilizing the special structure of Ω_S, we can further upper bound the determinant by the following lemma.

Lemma C.1 *For all $S \in [d], |S| = K$, we have*

$$\frac{1}{n}\log\det(X\Omega_S X^\top) \le \left(1 - \frac{K}{n}\right)\log\eta + \frac{K}{n}\log\frac{d}{K} + \frac{1}{n}\log\det\left(\frac{XX^\top}{d}\right).$$

Proof Denote by X_S the K columns of X whose indices belong to S. Then $X\Omega_S X^\top = \frac{1-\eta}{K}X_S X_S^\top + \frac{\eta}{d}XX^\top$. Making use of the determinant identity $\det(I_p + AB^\top) = \det(I_q + B^\top A)$ for any $A, B \in \mathbb{R}^{p\times q}$, it follows that

$$\det\left(X\Omega_S X^\top\right) = \det\left(\frac{1-\eta}{K}X_S X_S^\top + \frac{\eta}{d}XX^\top\right)$$

$$
\begin{aligned}
&= \eta^n \cdot \det\left(\frac{XX^\top}{d}\right) \cdot \det\left(I_d + \frac{(1-\eta)d}{\eta K}(XX^\top)^{-\frac{1}{2}} X_S X_S^\top (XX^\top)^{-\frac{1}{2}}\right) \\
&= \eta^n \cdot \det\left(\frac{XX^\top}{d}\right) \cdot \det\left(I_K + \frac{(1-\eta)d}{\eta K} X_S^\top (XX^\top)^{-1} X_S\right) \\
&\le \eta^n \cdot \det\left(\frac{XX^\top}{d}\right) \cdot \left(1 + \frac{(1-\eta)d}{\eta K}\right)^K,
\end{aligned}
$$

where in the last inequality, we use $\left\|X_S^\top (XX^\top)^{-1} X_S\right\| \le \left\|X_S^\top (X_S X_S^\top)^\dagger X_S\right\| = 1$. Taking logarithms on both sides, and applying the inequality $\eta K + (1-\eta)d \le d$, we complete the proof. ■

Applying Lemma C.1 to Eq. (16), we therefore obtain

$$
\mathsf{J}_\pi \ge \exp\left\{-\left(1 - \frac{K}{n}\right)\log\eta - \frac{3K}{n}\log\frac{d}{K} - \frac{2K}{n} - \frac{1}{n}\log\det\left(\frac{XX^\top}{d}\right)\right\}.
$$

We can once more exploit the convergence to the Marchenko–Pastur law as in the previous part, so that

$$
\frac{1}{n}\log\det\left(\frac{XX^\top}{d}\right) = \int \log\lambda \cdot \mu_n(\mathrm{d}\lambda) \to \int \log\lambda \cdot \mu_{\mathsf{MP},\gamma}(\mathrm{d}\lambda) = (1-\gamma)\cdot\log\left(1-\gamma^{-1}\right) - 1.
$$

The last equality is from (Bai and Silverstein, 2010, Example 1.1.1 & Sec. 9.12.3). We thus conclude the proof, additionally using $K = o(n)$ and $x\log x \to 0$ for $x \to 0+$,

$$
\liminf_{n\to+\infty} \mathsf{J}_\pi \ge \frac{1}{\eta} \cdot e\left(1 - \frac{1}{\gamma}\right)^{\gamma-1}.
$$

Indeed, the lower bound holds for general K by defining $\alpha = K/n \in [0,1]$ and

$$
\liminf_{n\to+\infty} \mathsf{J}_\pi \ge \left(\frac{1}{\eta}\right)^{1-\alpha} \cdot \left(\frac{\alpha}{\gamma}\right)^{3\alpha} \cdot e^{1-2\alpha}\left(1 - \frac{1}{\gamma}\right)^{\gamma-1}.
$$

Appendix D. Proofs for Section 4

D.1. Proof of Lemma 4.1

First, we remark that p does not need to be differentiable for this Lemma to hold, as Gaussian kernels smooth $\mathcal{L}^1$ functions into smooth functions for any $\sigma^2 > 0$—that is, p_{σ^2} is smooth for any $\sigma^2 > 0$, for any density p.

It follows from $\frac{\partial}{\partial y}\exp\left\{-\frac{\|X\theta'-y\|^2}{2\sigma^2}\right\} = \frac{X\theta'-y}{\sigma^2}\exp\left\{-\frac{\|X\theta'-y\|^2}{2\sigma^2}\right\}$ and we can interchange the order of differentiation and integration since the derivative is uniformly bounded, $\left\|\frac{X\theta'-y}{\sigma}\exp\left\{-\frac{\|X\theta'-y\|^2}{2\sigma^2}\right\}\right\| \le \sup_{t\in[0,+\infty)} te^{-t^2/2} \le e^{-1/2}$. To be specific,

$$
\frac{\partial}{\partial y}\log\mathbb{E}_{\theta'\sim\pi}\left[\exp\left\{-\frac{\|X\theta'-y\|^2}{2\sigma^2}\right\} \mid X, y\right] = \frac{\int \frac{\partial}{\partial y}\exp\left\{-\frac{\|X\theta'-y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta')}{\int \exp\left\{-\frac{\|X\theta'-y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta')} = \frac{X\widehat{\theta}_{\mathsf{B}} - y}{\sigma^2}.
$$

Eq. (10a) holds by substituting $y = X\theta + \sigma\tau$. Eq. (10b) holds given the existence of p since for $y' = X\theta'$,

$$\begin{aligned}
\exp\left\{-\frac{\|X\theta' - y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta) &= \exp\left\{-\frac{\|y' - y\|^2}{2\sigma^2}\right\}p(y')\mathrm{d}y' \\
&= C\phi_{\sigma^2}(y - y')p(y')\mathrm{d}y' \\
(X\theta' - y)\cdot\exp\left\{-\frac{\|X\theta' - y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta) &= (y' - y)\cdot\exp\left\{-\frac{\|y' - y\|^2}{2\sigma^2}\right\}p(y')\mathrm{d}y' \\
&= \sigma^2\frac{\partial}{\partial y}\exp\left\{-\frac{\|y - y'\|^2}{2\sigma^2}\right\}p(y')\mathrm{d}y' \\
&= C\sigma^2\nabla\phi_{\sigma^2}(y - y')p(y')\mathrm{d}y',
\end{aligned}$$

where $C = (2\pi\sigma^2)^{-\frac{n}{2}}$ is the integration constant. This confirms

$$\frac{\int \frac{\partial}{\partial y}\exp\left\{-\frac{\|X\theta' - y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta')}{\int \exp\left\{-\frac{\|X\theta' - y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta')} = \frac{\int \sigma^2\nabla\phi_{\sigma^2}(y - y')p(y')\mathrm{d}y'}{\int \phi_{\sigma^2}(y - y')p(y')\mathrm{d}y'} = \frac{\sigma^2\nabla p_{\sigma^2}(y)}{p_{\sigma^2}(y)} = \sigma^2\nabla\log p_{\sigma^2}(y).$$

D.2. Proof of Proposition 12

Let $t = \sigma^2$. The monotonicity of $\mathsf{Train}(t)/t^2 = \mathsf{J}(t)$ follows from Stam's Fisher information inequality in Lemma A.2, which implies $\mathsf{J}(t)^{-1} \geq \mathsf{J}(0)^{-1} + t$. The rest of the proof is for the monotonicity of $\mathsf{Train}(t)$ itself. This requires using the alternative form in Eq. (15),

$$\begin{aligned}
\mathsf{Train}(t) &= t - \frac{1}{n}\mathbb{E}_{y'\sim\pi',\tau\sim\mathsf{N}(0,I_n)}\left[\left\|y' - X\widehat{\theta}_{\mathsf{B}}\right\|^2 \mid X\right] \\
&= t - \frac{1}{n}\mathbb{E}_{y'\sim\pi',\tau\sim\mathsf{N}(0,I_n)}\mathbb{E}\left[\left\|y' - \mathbb{E}\left[y' \mid y\right]\right\|^2 \mid y = y' + \sqrt{t}\tau\right] \\
&=: t - \mathsf{mmse}(1/t),
\end{aligned}$$

identifying the formulation of MMSE in Gaussian channels. When $n = 1$, the explicit gradient $\mathsf{mmse}'(s)$ under the time inversion $s = 1/t$ is in (Guo et al., 2011, Cor. 2), given as $\mathsf{mmse}'(s) = -\mathbb{E}\left[\mathsf{Cov}(y' \mid y)^2\right]$. In the vector version, a similar formula indeed holds applying the general multi-dimensional derivative identity (Dytso et al., 2020, Prop. 1), which implies the following explicit equation.

Lemma D.1 *For all $s \in [0, +\infty)$ and $y = y' + \sqrt{1/s}\,\tau$,*

$$\mathsf{mmse}'(s) = -\frac{1}{n}\cdot\mathbb{E}\left[\left\|\mathsf{Cov}(y' \mid y)\right\|_F^2\right].$$

Proof Given the Jacobian identity from (Dytso et al., 2020, Prop. 1), we have

$$\frac{\partial\mathbb{E}\left[y' \mid y\right]}{\partial y} = s\mathsf{Cov}(y' \mid y),$$

and thus

$$\begin{aligned}\mathsf{mmse}'(s) &= \frac{1}{n}\cdot\mathbb{E}_y\left[\mathbb{E}\left[2\left(y'-\mathbb{E}\left[y'\mid y\right]\right)^\top\left(-\frac{\partial\mathbb{E}\left[y'\mid y\right]}{\partial y}\right)\mid y\right]\cdot\frac{\mathrm{d}y}{\mathrm{d}s}\right]\\ &= \frac{1}{n}\cdot\mathbb{E}_y\left[\mathbb{E}\left[2\left(y'-\mathbb{E}\left[y'\mid y\right]\right)^\top\left(-s\mathsf{Cov}(y'\mid y)\right)\mid y\right]\cdot\left(-\frac{1}{2}s^{-3/2}\right)\tau\right]\\ &= \frac{1}{n}\cdot\mathbb{E}\left[\mathrm{trace}\left(\mathbb{E}\left[\tau/\sqrt{s}\left(y'-\mathbb{E}\left[y'\mid y\right]\right)^\top\mid y\right]\cdot\mathsf{Cov}(y'\mid y)\right)\right].\end{aligned}$$

We then use that $\tau/\sqrt{s} = y - y'$ and conditional on y,

$$\mathbb{E}\left[\tau/\sqrt{s}\left(y'-\mathbb{E}\left[y'\mid y\right]\right)^\top\mid y\right] = -\mathbb{E}\left[\left(y'-\mathbb{E}\left[y'\mid y\right]\right)\left(y'-\mathbb{E}\left[y'\mid y\right]\right)^\top\mid y\right] = -\mathsf{Cov}(y'\mid y),$$

and the proof is done by

$$\mathsf{mmse}'(s) = -\frac{1}{n}\cdot\mathbb{E}\left[\mathrm{trace}\left(\mathsf{Cov}(y'\mid y)^2\right)\right] = -\frac{1}{n}\cdot\mathbb{E}\left[\left\|\mathsf{Cov}(y'\mid y)\right\|_F^2\right].$$

■

Given the derivative formula, by a variable transformation, we have

$$\begin{aligned}\mathsf{Train}'(t) &= 1-\mathsf{mmse}'(1/t)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{t}\right) = 1-\frac{1}{nt^2}\cdot\mathbb{E}\left[\left\|\mathsf{Cov}(y'\mid y)\right\|_F^2\right]\\ &= 1-\frac{1}{nt^2}\cdot\mathbb{E}\left[\left\|\mathbb{E}\left[\left(y'-\mathbb{E}\left[y'\mid y\right]\right)\left(y'-\mathbb{E}\left[y'\mid y\right]\right)^\top\right]\right\|_F^2\right]\\ &\overset{\text{(i)}}{\geq} 1-\frac{1}{nt^2}\cdot\mathbb{E}\left[\left\|\mathbb{E}\left[(y'-y)(y'-y)^\top\right]\right\|_F^2\right]\\ &= 1-\frac{1}{nt^2}\cdot\mathbb{E}\left[\left\|t\mathbb{E}\left[\tau\tau^\top\right]\right\|_F^2\right] = 1-\frac{1}{nt^2}\cdot nt^2 = 0,\end{aligned}$$

where in (i) we use the simple fact that conditional on y,

$$\begin{aligned}&\mathbb{E}\left[\left(y'-\mathbb{E}\left[y'\mid y\right]\right)\left(y'-\mathbb{E}\left[y'\mid y\right]\right)^\top\mid y\right]+\left(y-\mathbb{E}\left[y'\mid y\right]\right)\left(y-\mathbb{E}\left[y'\mid y\right]\right)^\top\\ &= \mathbb{E}\left[(y'-y)(y'-y)^\top\mid y\right].\end{aligned}$$

Then $\mathsf{Train}'(t)\geq 0$ and Train does not decrease in $[0,+\infty)$. The proof is complete.

D.3. Training error formula for the numerical simulation

In this section we give details for calculating the curves shown in Figure 3. Let $t = \sigma^2$. Since $\pi' = \frac{1}{2}\mathsf{N}(-1,\eta)+\frac{1}{2}\mathsf{N}(1,\eta)$, one has $\pi'_t = \frac{1}{2}\mathsf{N}(-1,\eta+t)+\frac{1}{2}\mathsf{N}(1,\eta+t)$. In this case, we can compute the explicit formula for its Fisher information $\mathsf{J}(t)$ from

$$\begin{aligned}\mathsf{J}(t) &= \mathbb{E}_{y\sim\pi'_t}\left[\left\{\frac{\mathrm{d}}{\mathrm{d}y}\log\left(\frac{1}{2}\phi_{\eta+t}(y-1)+\frac{1}{2}\phi_{\eta+t}(y+1)\right)\right\}^2\right]\\ &= \frac{1}{(\eta+t)^2}\cdot\mathbb{E}_{y\sim\pi'_t}\left[\left(\frac{(y-1)\phi_{\eta+t}(y-1)+(y+1)\phi_{\eta+t}(y+1)}{\phi_{\eta+t}(y-1)+\phi_{\eta+t}(y+1)}\right)^2\right]\end{aligned}$$

$$= \frac{1}{(\eta+t)^2}\cdot \mathbb{E}_{y\sim\pi'_t}\left[\left(y - \frac{\exp\left\{\frac{2y}{\eta+t}\right\}-1}{\exp\left\{\frac{2y}{\eta+t}\right\}+1}\right)^2\right] = \frac{1}{\eta+t} - \frac{1}{(\eta+t)^2}\mathbb{E}_{\tau\sim\mathsf{N}(0,1)}\left[\frac{4\exp\left(\frac{2(1+\sqrt{\eta+t}\tau)}{\eta+t}\right)}{\left(1+\exp\left(\frac{2(1+\sqrt{\eta+t}\tau)}{\eta+t}\right)\right)^2}\right].$$

At each value of (η, t), we can run Monte-Carlo to evaluate $\mathsf{J}(t)$.

Appendix E. Higher order asymptotics for the Bayes training error

In this section, we provide higher order asymptotics of $\mathsf{Train}(\sigma^2)$ when $\sigma^2 \to 0+$. We will use $t = \sigma^2$ in this section. As partially raised in Remark 11, the main technical hurdle that prevents us from directly taking limits such as $\lim_{t\to 0+}\mathsf{I}(t) = \mathsf{I}(0)$ hinges on the general insufficient regularities in the underlying noiseless data distribution π' along with its associated density p on $\mathbb{R}^n$.

We will need to introduce some necessary notations to facilitate delivering our results. On $\mathbb{R}^n$, we use the shorthand $\partial_i := \partial/\partial x_i$ when the context is clear, and denote by $\nabla := [\partial_1, \cdots, \partial_n]^\top \in \mathbb{R}^n$ the standard vector differential operator. Furthermore, we use the notation $u \cdot v$ to represent inner products, both between vectors and between differential operators and vector. We can then write the Hessian operator as $\nabla^2 = \nabla\nabla^\top$ and define the Laplace operator by $\Delta := \nabla\cdot\nabla = \mathrm{trace}(\nabla^2) = \sum_{i=1}^n \partial_i^2$.

To gain intuition and especially on the derivatives of $\mathsf{J}(t)$ at $t = 0+$, we first perform an informal analysis using formal power series in what follows, before studying properties of the Bayes training error function $\mathsf{Train}(t)$ at $t = 0+$ with full mathematical rigor. Let the formal power series ring $\mathbb{R}[[\sigma]]$ be the completion of the polynomial ring $\mathbb{R}[\sigma]$, we will write $\widetilde{o}(\sigma^k)$ to represent a power series with the leading order strictly larger than k, e.g. $\sum_{l=k+1}^{\infty}\sigma^l = \widetilde{o}(\sigma^k)$. We have:

Proposition 14 *Assume arbitrary differentiability of the density $p > 0$ of π', and we can always interchange differentiation with integration. Consider the following formal power series at any X, y and σ for $u_k \in \mathbb{R}^n$ and $\alpha_k \in \mathbb{R}$,*

$$\frac{1}{\sqrt{n}}\left(X\widehat{\theta}_{\mathsf{B}} - y\right) = \sum_{k=0}^{+\infty} u_k\sigma^k, \qquad \mathsf{Train}(\sigma^2) = \sum_{k=0}^{+\infty}\alpha_k\sigma^k,$$

whose coordinates and value belong to the polynomial ring $\mathbb{R}[\sigma]$. We have for the leading terms,

$$\frac{1}{\sqrt{n}}\left(X\widehat{\theta}_{\mathsf{B}} - y\right) = \frac{\sigma^2\nabla\log p(y)}{\sqrt{n}} + \frac{\sigma^4}{2\sqrt{n}}\left(\nabla(\Delta\log p(y)) + 2\nabla^2\log p(y)\nabla\log p(y)\right) + \widetilde{o}(\sigma^4),$$
$$\mathsf{Train}(\sigma^2) = \frac{\sigma^4}{n}\cdot\mathbb{E}_{y\sim\pi'}\left[\|\nabla\log p(y)\|^2\right] - \frac{\sigma^6}{n}\cdot\mathbb{E}_{y\sim\pi'}\left[\left\|\nabla^2\log p(y)\right\|_F^2\right] + \widetilde{o}(\sigma^6).$$

See the proof in Appendix E.1. We emphasize that the $\widetilde{o}(\cdot)$ terms in Proposition 14 do not yield explicit rates of convergence nor any finite radius of convergence in σ. But rather, they encapsulate only higher-order terms in a formal power series expansion. While this perspective is conceptually informative, it does not provide rigorous analytical guaranties. The central idea underlying the above informal results is to expand $p(y) = \exp\{\log p(y)\}$ into its Taylor series. However, deriving explicit convergence rates, justifying the interchange of integration and differentiation, and determining the radius of convergence all require substantially more sophisticated mathematical tools.

By employing powerful techniques such as the Bakry–Émery calculus, we rigorously establish the following characterizations of Train up to $o(\sigma^6)$ (distinct from the previous term $\widetilde{o}(\sigma^6)$)), when $\sigma^2 \to 0+$, validating the informal calculations in Proposition 14. The proof is provided in Appendix E.2.

Theorem 15 *Let π' be the probability distribution of $X\theta$ induced by $\theta \sim \pi$. Suppose that π' has a density $p > 0$ on $\mathbb{R}^n$ and p is twice differentiable. If π' has finite Fisher information $\mathsf{J}_\pi = \mathsf{J}(0) < +\infty$, the following quantities are well-defined for sufficiently small $t > 0$ and the limit holds*

$$\lim_{t\to 0+} \frac{1}{n}\mathbb{E}_{y\sim\pi'_t}\left[\left\|\nabla^2 \log p_t(y)\right\|_F^2\right] = \frac{1}{n}\mathbb{E}_{y\sim\pi'}\left[\left\|\nabla^2 \log p(y)\right\|_F^2\right] =: \left(-\mathsf{J}'_\pi\right) < +\infty.$$

then for sufficiently small $t > 0$,

$$\mathsf{J}'(t) = -\frac{1}{n}\mathbb{E}_{y\sim\pi'_t}\left[\left\|\nabla^2 \log p_t(y)\right\|_F^2\right],$$

and $\mathsf{J}'(0)$ exists and is equal to J'_π above. When $\sigma^2 \to 0+$, $|\mathsf{Train}(\sigma^2) - \sigma^4\mathsf{J}_\pi - \sigma^6\mathsf{J}'_\pi| = o(\sigma^6)$, i.e.

$$\left|\mathsf{Train}(\sigma^2) - \frac{\sigma^4}{n}\cdot\mathbb{E}_{y\sim\pi'}\left[\|\nabla \log p(y)\|^2\right] + \frac{\sigma^6}{n}\cdot\mathbb{E}_{y\sim\pi'}\left[\left\|\nabla^2 \log p(y)\right\|_F^2\right]\right| = o(\sigma^6).$$

As a final remark concluding Theorem 15, we provide in the following lemma a sufficient condition under which the convergence $\mathsf{J}'(t) \to \mathsf{J}'_\pi$ holds. See its proof in Appendix E.3.

Lemma E.1 *Let $p > 0$ be twice differentiable. If $\nabla p, \nabla^2 p$ are integrable, and*

$$\Phi(y) := \frac{\left\|\nabla^2 p(y)\right\|_F^2}{p(y)}, \qquad \text{and} \qquad \Psi(y) := \frac{\|\nabla p(y)\|^4}{p(y)^3}$$

belong to $\mathcal{L}^{1+\delta}(\mathbb{R}^n)$ for some $\delta > 0$. Then $\mathsf{J}'(t)$ exists for all $t > 0$. In addition, J'_π exists and is finite and $\mathsf{J}'(t) \to \mathsf{J}'_\pi$.

E.1. Proof of Proposition 14

To avoid heavy notations, we write $o_{\mathbb{R}[\sigma]}(\cdot)$ as $o(\cdot)$ in this proof, which subsumes higher-order terms in the polynomial ring $\mathbb{R}[\sigma]$.

Step I: Formal power series conditional on y. We first consider expanding the normalized residual term for fixed X, y in $\mathbb{R}^n$. By a variable transformation $z = (X\theta - y)/\sigma$,

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(X\widehat{\theta}_{\mathsf{B}} - y\right) &= \frac{1}{\sqrt{n}}\mathbb{E}\left[X\theta - y \mid X, y\right] = \frac{1}{\sqrt{n}}\frac{\int (X\theta - y)\exp\left\{-\frac{\|X\theta - y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta)}{\int \exp\left\{-\frac{\|X\theta - y\|^2}{2\sigma^2}\right\}\pi(\mathrm{d}\theta)} \\ &= \frac{\sigma}{\sqrt{n}}\frac{\int z\exp\left\{-\frac{\|z\|^2}{2}\right\}p(y + \sigma z)\mathrm{d}z}{\int \exp\left\{-\frac{\|z\|^2}{2}\right\}p(y + \sigma z)\mathrm{d}z} = \frac{\sigma}{\sqrt{n}}\frac{\int z\phi_1(z)p(y + \sigma z)\mathrm{d}z}{\int \phi_1(z)p(y + \sigma z)\mathrm{d}z}. \end{aligned} \tag{17}$$

Let $\ell(y) = \log p(y)$, and define

$$g = \nabla\ell(y) \in \mathbb{R}^n, \qquad H = \nabla^2\ell(y) \in \mathbb{R}^2, \qquad T = \nabla^3\ell(y) \in \mathbb{R}^3,$$

where T is the symmetric tensor for the third-order derivatives $T_{ijk} = \partial_i\partial_j\partial_k\ell(y)$. With the tensor-vector product notation $T[z,z,z] := \sum_{i,j,k} T_{ijk}z_iz_jz_k$, we can then explicitly write out the expansion of p in its leading terms

$$p(y+\sigma z) = \exp\{\log p(y+\sigma z)\} = \exp\left\{\ell(y) + \sigma g\cdot z + \frac{\sigma^2}{2}z^\top Hz + \frac{\sigma^3}{6}T[z,z,z] + o(\sigma^3)\right\}$$
$$= p(y)\cdot\left\{1 + \sigma g\cdot z + \frac{\sigma^2}{2}\left(z^\top Hz + (g\cdot z)^2\right) + \frac{\sigma^3}{6}\left(T[z,z,z] + 3(g\cdot z)\cdot z^\top Hz + (g\cdot z)^3\right) + o(\sigma^3)\right\}.$$

Since $\phi_1(z)\mathrm{d}z$ is the density for $\mathsf{N}(0, I_n)$, to substitute the above display into Eq. (17) requires computation of moments the standard Gaussian vector $Z \sim \mathsf{N}(0, I_n)$. Indeed, by Isserlis's formula in Lemma A.1,

$$\mathbb{E}[Z_iZ_jZ_kZ_l] = \mathbb{1}_{i=j}\mathbb{1}_{k=l} + \mathbb{1}_{i=k}\mathbb{1}_{j=l} + \mathbb{1}_{i=l}\mathbb{1}_{j=k},$$

we can compute the following non-zero expectations (we omit the zero terms of odd-moments by symmetry),

$$\mathbb{E}\left[Z(g\cdot Z)\right] = \nabla\ell(y), \qquad \mathbb{E}\left[Z^\top HZ + (g\cdot Z)^2\right] = \mathrm{trace}(H) + \|g\|^2 = \Delta\ell(y) + \|\nabla\ell(y)\|^2,$$
$$\mathbb{E}\left[ZT[Z,Z,Z]\right] = 3[\partial_1\Delta\ell(y),\cdots,\partial_n\Delta\ell(y)]^\top = 3\nabla(\Delta\ell(y)),$$
$$\mathbb{E}\left[(g\cdot Z)\cdot Z^\top HZ\right] = \mathrm{trace}(H)g + 2Hg = \Delta\ell(y)\nabla\ell(y) + 2\nabla^2\ell(y)\nabla\ell(y),$$
$$\mathbb{E}\left[(g\cdot Z)^3\right] = 3\|g\|^2 g = 3\|\nabla\ell(y)\|^2\nabla\ell(y).$$

Combining the above displays with the expansion for $p(y+\sigma z)$ yields

$$\frac{\mathbb{E}[p(y+\sigma Z)]}{p(y)} = 1 + \frac{\sigma^2}{2}\left(\Delta\ell(y) + \|\nabla\ell(y)\|^2\right) + o(\sigma^3),$$
$$\frac{\mathbb{E}[Zp(y+\sigma Z)]}{p(y)} = \sigma\nabla\ell(y) + \frac{\sigma^3}{2}\left(\nabla(\Delta\ell(y)) + \Delta\ell(y)\nabla\ell(y) + 2\nabla^2\ell(y)\nabla\ell(y) + \|\nabla\ell(y)\|^2\nabla\ell(y)\right) + o(\sigma^3).$$

Finally, identifying the preceding equations as the denominator and numerator for Eq. (17) we have

$$\frac{1}{\sqrt{n}}\left(X\widehat{\theta}_{\mathsf{B}} - y\right)$$
$$= \frac{\sigma}{\sqrt{n}}\cdot\frac{\sigma\nabla\ell(y) + \frac{\sigma^3}{2}\left(\nabla(\Delta\ell(y)) + \Delta\ell(y)\nabla\ell(y) + 2\nabla^2\ell(y)\nabla\ell(y) + \|\nabla\ell(y)\|^2\nabla\ell(y)\right) + o(\sigma^3)}{1 + \frac{\sigma^2}{2}\left(\Delta\ell(y) + \|\nabla\ell(y)\|^2\right) + o(\sigma^3)}$$
$$= \frac{\sigma^2}{\sqrt{n}}\cdot\left\{\nabla\ell(y) + \frac{\sigma^2}{2}\left(\nabla(\Delta\ell(y)) + 2\nabla^2\ell(y)\nabla\ell(y)\right) + o(\sigma^2)\right\},$$

where in the last equality we make use of $(1+cx)^{-1} = 1 - cx + o(x)$.

Step II: Computing $\mathsf{Train}_X(\widehat{\theta}_{\mathsf{B}})$ marginalizing over y. To compute the training error, it requires further using the expansion $y = y' + \sigma\tau$. Similarly we define the gradient, Hessian and third-order derivative tensor as g', H' and T' at y'. We denote by $T'[\tau,\tau] \in \mathbb{R}^n$ whose i-th entry is $\sum_{j,k} T'_{ijk}\tau_j\tau_k$. Using those notations, we obtain the following.

$\mathsf{Train}(\sigma^2)$

$$= \frac{\sigma^4}{n} \cdot \mathbb{E}_{y' \sim \pi', \tau \sim \mathsf{N}(0, I_n)} \left[\left\| g' + \sigma \cdot H'\tau + \frac{\sigma^2}{2} \left(T'[\tau, \tau] + \nabla(\Delta \ell(y')) + 2H'g' \right) + o(\sigma^2) \right\|^2 \right]$$

$$= \frac{\sigma^4}{n} \cdot \mathbb{E}_{y' \sim \pi', \tau \sim \mathsf{N}(0, I_n)} \left[\left\| \nabla \ell(y') \right\|^2 + \sigma^2 \left(\left\| H' \right\|_F^2 + g' \cdot T'[\tau, \tau] + g' \cdot \nabla(\Delta \ell(y')) + 2g'H'g' \right) + o(\sigma^2) \right],$$

where we use $\mathbb{E}[\tau] = 0$ and $\mathbb{E}[\tau\tau^\top] = I_n$. To fully remove τ, it still remains to calculate the expectation of $g' \cdot T'[\tau, \tau]$ over τ in what follows.

$$\mathbb{E}_{\tau \sim \mathsf{N}(0, I_n)} \left[g' \cdot T'[\tau, \tau] \right] = \sum_{i,j} g_i' T_{ijj} = g' \cdot \nabla(\Delta \ell(y')).$$

Combined with integration by parts for $g'H'g'$,

$$\mathbb{E}\left[g'H'g' \right] = \mathbb{E}\left[\nabla \left\| \nabla \ell(y') \right\|^2 \cdot \nabla \ell(y') \right] = -\mathbb{E}\left[\Delta \left\| \nabla \ell(y') \right\|^2 \right] = -\mathbb{E}\left[\Delta \left\| g' \right\|^2 \right],$$

we have

$$\mathsf{Train}(\sigma^2) = \frac{\sigma^4}{n} \cdot \mathbb{E}\left[\left\| \nabla \ell(y') \right\|^2 + \sigma^2 \left(\left\| H' \right\|_F^2 + 2g' \cdot \nabla(\Delta \ell(y')) - 2\Delta \left\| g' \right\|^2 \right) \right] + o(\sigma^6).$$

In the last step, we use Bochner's identity (Petersen, 2006, Prop. 9.2.2) for Euclidean space

$$\frac{1}{2} \Delta \left\| g' \right\|^2 = g' \cdot \nabla(\Delta \ell(y')) + \left\| H' \right\|_F^2,$$

we conclude the proof with

$$\mathsf{Train}(\sigma^2) = \frac{\sigma^4}{n} \cdot \mathbb{E}\left[\left\| \nabla \ell(y') \right\|^2 - \sigma^2 \left\| H' \right\|_F^2 \right] + o(\sigma^6).$$

E.2. Proof of Theorem 15

By Theorem 2, we have $\mathsf{J}(t) = \mathsf{J}(0) + o(t)$. It remains to be shown that $\mathsf{J}'(0) = \lim_{t \to 0+} \mathsf{J}'(t)$—due to the smoothing by the Gaussian kernels, the differentiability of $\mathsf{J}(t)$ when $t > 0$ is free, since we can always interchange integration and differentiation. However, the exact formula for $\mathsf{J}'(t)$, is much more involved, requiring Bakry and Émery's calculus (Bakry and Émery, 2006; Bakry et al., 2013). We refer to the following explicit formula[8] from (Ledoux, 2016, Eq. (3.2)) as

$$\mathsf{J}'(t) = -\frac{1}{n} \mathbb{E}_{y \sim \pi_t'} \left[\left\| \nabla^2 \log p_t(y) \right\|_F^2 \right].$$

We then use Theorem 2 to show the differentiability of $\mathsf{J}(t)$ at $t = 0+$. Indeed,

$$\frac{\mathsf{J}(t) - \mathsf{J}(0)}{t} = \lim_{\epsilon \to 0+} \frac{\mathsf{J}(t) - \mathsf{J}(\epsilon)}{t} = \frac{1}{t} \int_{0+}^{t} \mathsf{J}'(t) \mathrm{d}t.$$

Taking $t \to 0$, and by the assumption that $\lim_{t \to 0+} \mathsf{J}'(t)$ exists, we conclude

$$\mathsf{J}'(0) = \lim_{t \to 0} \frac{1}{t} \int_{0+}^{t} \mathsf{J}'(t) \mathrm{d}t = \lim_{t \to 0+} \left(-\frac{1}{n} \mathbb{E}_{y \sim \pi_t'} \left[\left\| \nabla^2 \log p_t(y) \right\|_F^2 \right] \right) = -\frac{1}{n} \mathbb{E}_{y \sim \pi'} \left[\left\| \nabla^2 \log p(y) \right\|_F^2 \right].$$

The proof is complete.

8. We adapt to our scaling convention $t = \sigma^2$. The original form in the referenced paper is for the scaling $t = \sigma$.

E.3. Proof of Lemma E.1

Since $p, \nabla p, \nabla^2 p$ are integrable, by Lebesgue differentiation theorem, one has almost surely

$$p_t = p * \varphi_t \to p, \qquad \nabla p_t = \nabla p * \varphi_t \to \nabla p, \qquad \nabla^2 p_t = \nabla^2 p * \varphi_t \to \nabla^2 p,$$

where we use for any two differentiable functions f and g on the same space, $\nabla(f * g) = \nabla f * g = f * \nabla g$. Therefore, $p_t \nabla^2 \log p_t \to p \nabla^2 \log p$. To upgrade a.s. convergence to convergence in the mean, we want to invoke Vitali's convergence theorem, which requires uniform integrability certifications. We construct auxiliary functions

$$\Phi_t(y) := \frac{\left\|\nabla^2 p_t(y)\right\|_F^2}{p_t(y)}, \qquad \text{and} \qquad \Psi_t(y) := \frac{\|\nabla p_t(y)\|^4}{p_t(y)^3},$$

and we can write the upper bound

$$\left\|\nabla^2 \log p_t\right\|_F^2 p_t = \left\| \frac{\nabla^2 p_t}{p_t} - \frac{\nabla p_t \left(\nabla p_t\right)^\top}{p_t^2} \right\|_F^2 p_t \le 2(\Phi_t + \Psi_t). \tag{18}$$

By Hölder's inequality, for any number $q > 0$ and functions f and $g > 0$,

$$\|f(\cdot)\varphi_t(y - \cdot)\|_{\mathcal{L}^1} \le \left\| \frac{f(\cdot)}{g^{\frac{q}{q+1}}(\cdot)} \varphi_t^{\frac{1}{q+1}}(y - \cdot) \right\|_{\mathcal{L}^{q+1}} \left\| g^{\frac{q}{q+1}}(\cdot) \varphi_t^{\frac{q}{q+1}}(y - \cdot) \right\|_{\mathcal{L}^{q^{-1}+1}}.$$

Rearranging terms gives

$$\frac{(f * \varphi_t)^{q+1}}{(g * \varphi_t)^q} \le \left(\frac{f^{q+1}}{g^q} \right) * \varphi_t.$$

Applying the above inequality pointwise to Φ_t and Ψ_t gives

$$\Phi_t = \frac{\left\|\nabla^2 p_t\right\|_F^2}{p_t} = \frac{\left\|\nabla^2 p * \varphi_t\right\|_F^2}{p * \varphi_t} \le \Phi * \varphi_t,$$
$$\Psi_t = \frac{\|\nabla p_t\|^4}{p_t^3} = \frac{\|\nabla p * \varphi_t\|^4}{(p * \varphi_t)^3} \le \Psi * \varphi_t.$$

Return to Eq. (18), the above display then implies

$$\left\| \left\|\nabla^2 \log p_t\right\|_F^2 p_t \right\|_{\mathcal{L}^{1+\delta}} \le 2 \|\Phi + \Psi\|_{1+\delta} \|\varphi_t\|_1 = 2 \|\Phi + \Psi\|_{1+\delta}.$$

Hence, the functions $\{\left\|\nabla^2 \log p_t\right\|_F^2 p_t\}$ indexed by t are uniformly integrable in $\mathcal{L}^1$. By Vitali's convergence theorem (Billingsley, 1995, Thm. 16.14), we conclude that

$$\mathsf{J}'(t) = -\int \left\|\nabla^2 \log p_t(y)\right\|_F^2 p_t(y) \mathrm{d}y \to -\int \left\|\nabla^2 \log p(y)\right\|_F^2 p(y) \mathrm{d}y = \mathsf{J}'_\pi.$$

The proof is complete.