A STATISTICAL THEORY FOR BLIND DENOISING: MINIMAX ESTIMATION OF THE NOISE LEVEL FROM A SINGLE SAMPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by blind denoising in diffusion models, we study estimation of an unknown Gaussian noise level from a single high-dimensional observation, assuming the signal law P is known. We characterize the minimax mean-squared error under two structural assumptions on P. For signals with covering complexity k, the minimax rate is $\widetilde{\Theta}_\Lambda(\min\{\Delta_\Lambda^2, d^{-1} + k^2 d^{-2}\})$, and the constrained MLE attains it up to logarithmic factors. For α-strongly log-concave signals, the rate is $\Theta_\Lambda(\min\{\Delta_\Lambda^2, (1+\alpha^{-1})^2 d^{-1}\})$, attained up to constants by a P-centered norm estimator. These results show that the structure of the signal law determines both the difficulty of blind noise estimation and the appropriate estimator.

1 INTRODUCTION

Diffusion models generate high-dimensional data by reversing a gradual corruption process through a denoiser or score field indexed by diffusion time (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). They have also proved effective for conditional generation (Dhariwal & Nichol, 2021). A closely related class of generative models is flow matching, which learns a time-dependent velocity field along a probability path connecting a simple base distribution to the data distribution (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023). In both cases, the model typically uses a network that takes time as an input, or the corresponding noise level along a Gaussian path (Song & Ermon, 2019; Ho et al., 2020; Lipman et al., 2023).

Recent work has asked whether the network needs this explicit time input. Noise-unconditioned denoising models can remain competitive with their conditioned counterparts (Sun et al., 2025), while blind denoising diffusion models omit the noise amplitude during both training and sampling (Kadkhodaie et al., 2026). Relatedly, in flow matching, Beckmann Transport Models show that, under suitable geometric conditions, an autonomous velocity field defines an exact transport and induces a one-step generative map (Lee et al., 2026). Taken together, these results suggest that explicit noise conditioning is not always necessary. What remains less understood is why a model can recover the missing noise information from the noisy state itself. This leads to a basic statistical question: if the noise level is not provided, what properties of the signal distribution make it recoverable from a single noisy observation?

Concretely, a blind denoiser must infer which noisy marginal produced its current input. This problem is closely related to classical noise-level estimation: estimating unknown noise parameters from a single corrupted observation is a classical difficulty in blind image denoising (Liu et al., 2013), and recent diffusion-based approaches likewise treat unknown noise parameters as an additional inference problem (Heurtel-Depeiges et al., 2024). More recently, Kadkhodaie et al. (2026) formalized this problem in their analysis of blind denoising diffusion models (BDDMs). Under a low-intrinsic-dimensionality assumption, quantified by the covering complexity of the signal support at the relevant noise scale, they showed that the posterior over the noise level concentrates near its true value, allowing the blind score to approximate the corresponding noise-conditioned score and the model to track an implicit noise schedule. Their analysis yields an estimation upper bound at the scale $d^{-1} + k^2 d^{-2}$, where d is the ambient dimension and k is the corresponding intrinsic dimension. This bound is sufficient to control the additional error introduced by blind sampling,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

but does not determine whether the estimation scale is statistically optimal or whether low intrinsic dimensionality is necessary for accurate estimation.

In this work, we study this estimation problem separately from the diffusion dynamics. Consider

$$Y = X + \lambda^{-1/2} Z, \qquad X \sim P, \quad Z \sim \mathcal{N}(0, I_d),$$

where the signal law P is known, the unknown noise level is parameterized by its precision $\lambda \in \Lambda$, and we observe a single d-dimensional vector Y. The estimator may use the full law P; this oracle formulation excludes distribution-learning, neural-network approximation, score-estimation, and discretization errors.

The ambient dimension might appear to provide d effective observations and hence a parametric rate. The signal coordinates, however, need not be independent, and variation in X can imitate a change in the noise level. Consequently, high dimension alone does not guarantee accurate estimation: it becomes useful only when the structure of P prevents signal variation from concealing noise variation. We therefore ask:

Which structural properties make blind noise estimation possible, and at what optimal rate?

1.1 Our Contributions

Our results show that accurate blind noise estimation is possible under more than one structural regime. We characterize the minimax difficulty in these regimes and give estimators attaining the corresponding rates.

Minimax rates. Let d denote the ambient dimension, k the intrinsic dimension in the low-dimensional regime, α the strong log-concavity parameter, and Δ_Λ the width of the fixed precision interval. Under the assumptions stated below, and up to constants depending on Λ, the minimax mean-squared error has order

$$\widetilde{\Theta}_\Lambda\left(\min\left\{\Delta_\Lambda^2, \frac{1}{d} + \frac{k^2}{d^2}\right\}\right)$$

under low intrinsic dimensionality, and

$$\Theta_\Lambda\left(\min\left\{\Delta_\Lambda^2, \frac{(1+\alpha^{-1})^2}{d}\right\}\right)$$

under strong log-concavity. The formal statements appear in Theorems 3.1 and 3.2.

Under low intrinsic dimensionality, we prove that the d^{-1} term is a parametric lower bound for every fixed signal law P, while the k^2/d^2 term arises from latent signal variation that can conceal a precision perturbation of order k/d. The constrained maximum-likelihood estimator (MLE) matches these lower bounds up to logarithmic factors, showing that the scale obtained in the blind-denoising analysis of Kadkhodaie et al. (2026) is essentially unimprovable in the fixed-P setting.

Under strong log-concavity, low-dimensional support is not necessary. For α-strongly log-concave signal laws, which may have full-dimensional support, we construct a P-centered norm estimator that attains the minimax rate up to constants. The MLE has the optimal dependence on α with an additional logarithmic factor in dimension. Thus, accurate noise estimation does not require low-dimensional support.

Norm-based and likelihood-based estimation. Strong log-concavity controls the radial fluctuations needed for the norm estimator through a Poincaré inequality. In contrast, low covering complexity alone does not control $\mathrm{Var}_P(\|X\|^2)$, so the standard energy-corrected norm statistic has no uniform variance bound in terms of k alone. This explains why the same simple statistic need not attain the low-dimensional rate, whereas the MLE adapts to the full signal law. We develop this comparison in Section 4.1 and Appendix B.

Proof ideas. We develop separate lower- and upper-bound arguments for the two structural regimes. For the k^2/d^2 lower bound, we construct a single variance-mixture signal law and approximate it by a finite point cloud, preserving overlap between noisy marginals while enforcing the

covering condition. For the MLE, we use two localization arguments. Under low intrinsic dimensionality, posterior-energy control and monotonicity yield a quadratic likelihood comparison. Under strong log-concavity, we instead establish Rényi separation and combine it with a one-dimensional covering argument. The separation-based argument isolates the structural input in a bound between the noisy marginals, suggesting a route to other signal classes for which comparable separation estimates can be established. Section 4 gives the proof overview, with complete arguments in Appendices B and C.

1.2 Prior Work

Blind denoising and noise-unconditioned generative models. Classical single-image noise estimators infer the noise level from low-texture or low-rank image patches (Liu et al., 2013; Chen et al., 2015). Learning-based denoisers have treated the role of the noise level in different ways. DnCNN demonstrated blind Gaussian denoising with an unknown noise level (Zhang et al., 2017), whereas FFDNet takes a noise-level map as an explicit input while using a single network across a range of noise levels (Zhang et al., 2018). Gnanasambandam & Chan (2020) instead study how to choose the training distribution over noise levels when one denoiser is used across the entire range. Blind denoisers have also been used as implicit priors for inverse problems (Kadkhodaie & Simoncelli, 2021). More recently, Sun et al. (2025) analyzed the error incurred by removing noise conditioning from denoising generative models, while Heurtel-Depeiges et al. (2024) proposed a Gibbs procedure that jointly samples the signal and unknown parametric noise variables using a diffusion prior. Kadkhodaie et al. (2026) is the closest work to ours: their BDDM analysis gives the low-dimensional noise-estimation upper bound, whereas we study the minimax limits of this estimation problem.

Diffusion models under low-dimensional structure. Empirical evidence suggests that the intrinsic dimension of natural images is much smaller than their ambient dimension (Pope et al., 2021). Theoretical work shows that score-based diffusion models can adapt their sampling or discretization complexity to such structure (Huang et al., 2026; Li & Yan, 2024; Potaptchik et al., 2025; Liang et al., 2025). These results concern sampling with supplied noise levels and accurate scores. In the blind setting, low-dimensional structure plays a different role: it can make the missing noise level recoverable from the current noisy state (Kadkhodaie et al., 2026). Our low-dimensional result characterizes the optimal estimation scale associated with this mechanism.

Theoretical analyses of diffusion models. Many analyses condition on an accurate score estimator to separate score-estimation error from initialization and discretization errors (Chen et al., 2023a;b; Lee et al., 2023). Complementary work studies the statistical problem of learning the score itself from finite data (Wibisono et al., 2024). Our formulation isolates noise-level estimation with the signal law given. Strong log-concavity is also a standard structural assumption in the theory of high-dimensional sampling and diffusion models, where it yields concentration and functional inequalities that support quantitative convergence guarantees (Saumard & Wellner, 2014).

Convolution models with unknown noise. Unknown Gaussian noise variance has also been studied in semiparametric convolution models (Matias, 2002; Butucea & Matias, 2005). In that literature, the signal density is unknown, information accumulates through multiple independent observations, and smoothness assumptions ensure identifiability. Here, one noisy vector is observed with P given, and the rates depend on its covering complexity k or strong log-concavity parameter α.

2 Background

2.1 From Blind Denoising to Estimation of the Noise Level

Additive Gaussian noise is a standard model of corruption in denoising and diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2021). For a clean signal $X \sim P$ and an independent $Z \sim \mathcal{N}(0, I_d)$, a noisy observation at level $\sigma > 0$ has the form

$$Y_\sigma = X + \sigma Z, \qquad P_\sigma := P * \mathcal{N}(0, \sigma^2 I_d).$$

Standard diffusion models typically use denoisers that take the noise level as an explicit input, leading to somewhat ad hoc choices of architecture and diffusion schedule. Recently, Kadkhodaie &

Simoncelli (2021) showed that one can instead use a *blind* denoiser, trained across noise levels $\sigma \sim \Pi_0$ without receiving σ itself as an input. Given a network $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$, the training objective is

$$\mathbb{E}_{X,\sigma,Z}\left[\|X - f_\theta(X + \sigma Z)\|^2\right], \qquad \sigma \sim \Pi_0, \quad Z \sim \mathcal{N}(0, I_d).$$

Let $f^\star_\sigma(y) := \mathbb{E}[X \mid Y_\sigma = y]$ denote the noise-conditioned Bayes denoiser. The population-optimal blind denoiser is

$$f^\star(y) = \int f^\star_\sigma(y)\, \Pi(d\sigma \mid y), \qquad \Pi(d\sigma \mid y) = \frac{p_\sigma(y)\Pi_0(d\sigma)}{\int p_\omega(y)\Pi_0(d\omega)}.$$

Surprisingly, blind denoising can even improve image generation quality despite the missing noise input (Kadkhodaie et al., 2026). This work gave a theoretical explanation under low intrinsic dimension: the posterior $\Pi(\cdot \mid Y)$ concentrates near the true noise level $\sigma_\star$, so the blind Bayes denoiser is close to the corresponding noise-conditioned denoiser $f^\star_{\sigma_\star}(Y)$. The same noisy marginal determines the latter through the Tweedie–Miyasawa identity

$$f^\star_\sigma(y) - y = \sigma^2 \nabla \log p_\sigma(y),$$

which links denoising to the score of P_σ (Miyasawa, 1961; Efron, 2011).

In this work, we study blind estimation of the noise level as a statistical problem in its own right. In particular, we ask whether the statistical rate obtained under low intrinsic dimension is tight, and under what other conditions blind noise estimation is possible.

The posterior $\Pi(\cdot \mid y)$ combines the prior weights in Π_0 with likelihood ratios between the marginals $\{P_\sigma\}_{\sigma>0}$. If two noise levels induce nearly indistinguishable laws for Y, then they cannot be reliably distinguished from the current state. We therefore study the statistical problem of estimating the noise level from Y for a fixed signal law P, independently of the choice of training distribution Π_0.

2.2 Statistical formulation and likelihood identities

We parameterize the noise by its precision $\lambda := \sigma^{-2}$, and fix $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ with $0 < \lambda_{\min} < \lambda_{\max} < \infty$. Since Λ is a fixed compact interval, estimating λ or $\sigma = \lambda^{-1/2}$ under squared-error loss gives equivalent rates, up to constants depending only on Λ. We work with λ because the likelihood takes a simpler form:

$$Y = X + \lambda^{-1/2} Z, \qquad P_\lambda := P * \mathcal{N}(0, \lambda^{-1} I_d), \qquad p_\lambda(y) = \left(\frac{\lambda}{2\pi}\right)^{d/2} \int e^{-\frac{\lambda}{2}\|x-y\|^2} P(dx).$$

Write $\Delta_\Lambda := \lambda_{\max} - \lambda_{\min}$.

For each fixed P, define

$$\mathcal{R}(P; \Lambda) := \inf_{\widehat{\lambda}_P} \sup_{\lambda \in \Lambda} \mathbb{E}_{\lambda,P}\left[\left(\widehat{\lambda}_P(Y) - \lambda\right)^2\right],$$

where the infimum is over estimators $\widehat{\lambda}_P : \mathbb{R}^d \to \Lambda$. Throughout, P is treated as known, and $\widehat{\lambda}_P$ may depend arbitrarily on P, including through its full likelihood. In Section 3, we take the supremum of $\mathcal{R}(P; \Lambda)$ over each structural class.

Let $\ell(\lambda \mid y) := -\log p_\lambda(y)$, and define the posterior quadratic error

$$D_\lambda(y) := \mathbb{E}_{\lambda,P}\left[\|X - y\|^2 \mid Y = y\right].$$

Direct differentiation gives

$$2\ell'(\lambda \mid y) = -\frac{d}{\lambda} + D_\lambda(y), \qquad \partial_\lambda D_\lambda(y) = -\frac{1}{2}\operatorname{Var}_{\lambda,P}\left(\|X - y\|^2 \mid Y = y\right) \le 0.$$

With $\lambda = \sigma^{-2}$, $D_\lambda(Y)$ is the conditional quadratic error appearing in the blind-denoising analysis. We consider the constrained MLE

$$\widehat{\lambda}_{\mathrm{MLE},P} \in \arg\min_{\lambda \in \Lambda} \ell(\lambda \mid Y).$$

Its localization is analyzed in Section 4.1; the likelihood derivations and auxiliary inequalities are collected in Appendix A.

3 RESULTS

We now formally state our assumptions and our main results.

3.1 RESULTS FOR LOW INTRINSIC DIMENSIONALITY

We first consider signal distributions with low geometric complexity. Define the resolution

$$r_0 := \frac{\sqrt{\lambda_{\min}}}{\lambda_{\max}\sqrt{d}}.$$

For $k \geq 1$ and $R > 0$, let $\mathcal{P}_{\mathrm{LD}}(k, R)$ be the class of probability measures P on $\mathbb{R}^d$ satisfying

$$\mathrm{supp}(P) \subseteq B(0, R), \qquad 1 + \log N(\mathrm{supp}(P), r_0) \leq k,$$

where $N(S, r_0)$ denotes the Euclidean covering number of S at scale r_0. The support may be discrete or highly nonconvex. Its covering complexity at this resolution is bounded by k, which need not scale with the ambient dimension d.

Using the risk $\mathcal{R}(P; \Lambda)$ defined in Section 2, set

$$\mathcal{R}^\star_{\mathrm{LD}}(d, k, R; \Lambda) := \sup_{P \in \mathcal{P}_{\mathrm{LD}}(k,R)} \mathcal{R}(P; \Lambda).$$

Our main result shows that, up to logarithmic factors, the minimax risk is governed by the two scales d^{-1} and $k^2 d^{-2}$.

Theorem 3.1 (Low intrinsic dimensionality)**.** *Assume $1 \leq k \lesssim_\Lambda d$ and $R^2 \gtrsim_\Lambda k$. Then*

$$\mathcal{R}^\star_{\mathrm{LD}}(d, k, R; \Lambda) \gtrsim \min\left\{\Delta_\Lambda^2,\ \lambda_{\min}^2\left(\frac{1}{d} + \frac{k^2}{d^2}\right)\right\}.$$

Conversely, for every fixed $P \in \mathcal{P}_{\mathrm{LD}}(k, R)$, the MLE defined in Section 2.2 satisfies

$$\sup_{\lambda^\star \in \Lambda} \mathbb{E}_{\lambda^\star, P}\left[(\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star)^2\right] \lesssim \min\left\{\Delta_\Lambda^2,\ \frac{\lambda_{\max}^4}{\lambda_{\min}^2}\left[\frac{\log(2d)}{d} + \frac{(k + \log(2d))^2}{d^2}\right]\right\}.$$

Hence, for a fixed compact precision interval Λ,

$$\mathcal{R}^\star_{\mathrm{LD}}(d, k, R; \Lambda) = \widetilde{\Theta}_\Lambda\left(\min\left\{\Delta_\Lambda^2,\ \frac{1}{d} + \frac{k^2}{d^2}\right\}\right).$$

The two terms dominate in different regimes. When $k \lesssim \sqrt{d}$, the parametric term d^{-1} determines the rate, whereas for $k \gtrsim \sqrt{d}$, the structural term k^2/d^2 dominates. The radius R enters through feasibility of the structural class and the lower-bound construction; once $R^2 \gtrsim_\Lambda k$, it does not determine the resulting minimax scale. This shows that the low-dimensional estimation rate obtained by Kadkhodaie et al. (2026) is tight up to logarithmic factors.

Proof sketch. The d^{-1} lower bound follows from a parametric two-point argument for fixed P. The k^2/d^2 term uses a variance-mixture construction with precision separation of order k/d; see Section 4.2 and Appendix B.

For the upper bound, control of $D_\lambda(Y)$ together with the likelihood identity in Section 2 compares the observed negative log-likelihood with a quadratic Gaussian reference. This yields a localization bound for the MLE; see Section 4.1 and Appendix B.

3.2 RESULTS FOR STRONG LOG-CONCAVITY

Next, we investigate whether we can go beyond low intrinsic dimensionality by identifying another class of structured signal priors for which blind noise estimation is possible. We consider strongly log-concave signal distributions. For $\alpha > 0$, let

$$\mathcal{P}_{\mathrm{SLC}}(\alpha) := \left\{P(dx) = Z_V^{-1} e^{-V(x)} dx : V \in C^2(\mathbb{R}^d),\ \nabla^2 V(x) \succeq \alpha I_d\right\}.$$

For $P \in \mathcal{P}_{\mathrm{SLC}}(\alpha)$, let $\mu_P := \mathbb{E}_P X$. Since P is known, so is μ_P. Translating (X, Y) to $(X - \mu_P, Y - \mu_P)$ preserves both the noise precision and strong log-concavity. Unlike the class in Section 3.1, this condition controls signal spread through α without restricting support dimension.

Using the risk $\mathcal{R}(P; \Lambda)$ defined in Section 2, set

$$\mathcal{R}^\star_{\mathrm{SLC}}(d, \alpha; \Lambda) := \sup_{P \in \mathcal{P}_{\mathrm{SLC}}(\alpha)} \mathcal{R}(P; \Lambda), \qquad \rho_{\alpha,d}(\Lambda) := \min\left\{\Delta_\Lambda^2,\ \frac{(1+\alpha^{-1})^2}{d}\right\}.$$

Our main result shows that, up to constants depending on Λ, the minimax risk is governed by the scale $(1+\alpha^{-1})^2/d$.

Theorem 3.2 (Strong log-concavity)**.** *For every $d \geq 1$ and $\alpha > 0$,*

$$\mathcal{R}^\star_{\mathrm{SLC}}(d, \alpha; \Lambda) \gtrsim_\Lambda \rho_{\alpha,d}(\Lambda).$$

Conversely, for every fixed $P \in \mathcal{P}_{\mathrm{SLC}}(\alpha)$, there exists a P-centered norm estimator $\widehat{\lambda}_{\mathrm{norm},P}$ such that

$$\sup_{\lambda \in \Lambda} \mathbb{E}_{\lambda,P}\left[\left(\widehat{\lambda}_{\mathrm{norm},P}(Y) - \lambda\right)^2\right] \leq \min\left\{\Delta_\Lambda^2,\ \frac{\lambda_{\max}^4}{d}\left(\frac{4}{\alpha^2} + \frac{4}{\alpha\lambda_{\min}} + \frac{2}{\lambda_{\min}^2}\right)\right\}.$$

The constrained fixed-P MLE defined in Section 2.2 additionally satisfies

$$\sup_{\lambda^\star \in \Lambda} \mathbb{E}_{\lambda^\star,P}\left[\left(\widehat{\lambda}_{\mathrm{MLE},P}(Y) - \lambda^\star\right)^2\right] \lesssim_\Lambda \min\left\{\Delta_\Lambda^2,\ \frac{(1+\alpha^{-1})^2 \log(2d)}{d}\right\}.$$

Hence,

$$\mathcal{R}^\star_{\mathrm{SLC}}(d, \alpha; \Lambda) \asymp_\Lambda \rho_{\alpha,d}(\Lambda).$$

In the factor $1 + \alpha^{-1}$, the constant represents the Gaussian noise scale and α^{-1} measures the permitted signal spread. The factor d^{-1} reflects averaging in high dimension. The risk saturates at the squared interval width Δ_Λ^2.

Proof sketch. For the lower bound, it suffices to restrict to the Gaussian prior $P_\alpha = \mathcal{N}(0, \alpha^{-1} I_d)$, for which the observation law remains Gaussian, and apply a two-point argument.

For the upper bound, writing $a = \lambda^{-1}$, consider

$$\widetilde{a}_P(Y) := \frac{\|Y - \mu_P\|^2 - \mathbb{E}_P\|X - \mu_P\|^2}{d}.$$

This statistic is unbiased for a. Strong log-concavity controls both $\mathbb{E}_P\|X - \mu_P\|^2$ and $\mathrm{Var}_P(\|X - \mu_P\|^2)$ through the Poincaré inequality, giving variance of order $(1+\alpha^{-1})^2/d$. Projection onto the admissible variance interval followed by inversion gives the norm estimator. For its analysis and the MLE argument, see Section 4.1 and Appendix C.

4 Techniques

We now give a technical overview of the main ideas behind the results in Section 3. Complete proofs of the following arguments appear in Appendices B and C.

4.1 Upper-bound techniques

MLE under low intrinsic dimensionality. We localize the MLE through a quadratic comparison of negative log-likelihoods. Recall that $2\ell'(\lambda \mid Y) = -d/\lambda + D_\lambda(Y)$, where $D_\lambda(Y)$ is the posterior quadratic error. The blind-denoising analysis gives high-probability control of $D_\lambda(Y)$ around the true Gaussian noise energy $d/\lambda^\star$ for each candidate $\lambda \in \Lambda$, with fluctuations of order $\sqrt{d} + k$, up to logarithmic factors and constants depending on Λ (Kadkhodaie et al., 2026). Integrating the

likelihood derivative therefore compares the observed negative log-likelihood with the Gaussian reference function

$$\Psi_{\lambda^\star}(\lambda) = \frac{d}{2}\left(\frac{\lambda}{\lambda^\star} - 1 - \log\frac{\lambda}{\lambda^\star}\right).$$

This function has curvature of order d, while the approximation error is linear in $|\lambda - \lambda^\star|$. The excess negative log-likelihood therefore has a lower bound of the form $d(\lambda-\lambda^\star)^2 - \text{error}\cdot|\lambda-\lambda^\star|$, and the defining inequality $\ell(\widehat{\lambda}_{\mathrm{MLE},P} \mid Y) \le \ell(\lambda^\star \mid Y)$ localizes the MLE.

The posterior-energy bound is taken from Kadkhodaie et al. (2026), where it is used to control posterior draws. For the fixed-P MLE, we combine this bound with the monotonicity of D_λ. It is enough to control $D_\lambda(Y)$ at the endpoints of Λ, which gives uniform control of the likelihood comparison over the entire interval.

Norm estimator under strong log-concavity. For a norm statistic, the relevant quantity is the variance of the signal energy. With $a = \lambda^{-1}$ and $Y = X + \sqrt{a}Z$, we have

$$\mathrm{Var}(\|Y\|^2) = \mathrm{Var}_P(\|X\|^2) + 4a\,\mathbb{E}_P\|X\|^2 + 2a^2 d.$$

Subtracting the mean signal energy under P leaves the variance term $\mathrm{Var}_P(\|X\|^2)$, which covering complexity k does not control: a distribution with small support complexity can assign mass to substantially different radii.

Strong log-concavity provides precisely this missing variance control. After centering at $\mu_P = \mathbb{E}_P X$, the Poincaré inequality (Bakry et al., 2014) gives

$$\mathbb{E}_P\|X-\mu_P\|^2 \le \frac{d}{\alpha}, \qquad \mathrm{Var}_P\big(\|X-\mu_P\|^2\big) \le \frac{4d}{\alpha^2}.$$

The P-centered norm statistic therefore estimates the noise variance with mean-squared error of order $(1+\alpha^{-1})^2/d$. Projection followed by inversion gives the estimator in Theorem 3.2.

MLE under strong log-concavity. The low-dimensional argument above relies on direct control of the posterior energy $D_\lambda(Y)$. Under strong log-concavity, the same approach does not preserve the desired dependence on α, so we instead compare the noisy marginals directly.

Strong log-concavity is preserved under Gaussian convolution, with inverse-curvature scale bounded by

$$s_{\alpha,\Lambda} := \alpha^{-1} + \lambda_{\min}^{-1}.$$

The change in noise level produces a corresponding separation of the covariances of $P_{\lambda^\star}$ and P_λ. Transportation and covariance inequalities (Saumard & Wellner, 2014; Bakry et al., 2014; Gelbrich, 1990) then give

$$D_{1/2}(P_{\lambda^\star}\|P_\lambda) \gtrsim_\Lambda \frac{d(\lambda-\lambda^\star)^2}{s_{\alpha,\Lambda}^2}.$$

This Rényi separation gives an exponential likelihood-ratio bound for each fixed candidate λ. A one-dimensional net over Λ, together with a derivative bound for $\lambda \mapsto \ell(\lambda \mid Y)$, extends the comparison to the whole interval and localizes the MLE.

4.2 Lower-bound techniques

Beyond the universal parametric d^{-1} lower bound, the k^2/d^2 term requires a fixed signal law that conceals a change in observation noise.

It is convenient to parameterize the noise by its variance $a = \lambda^{-1}$. In the non-saturated regime, take a variance increment $\delta \asymp_\Lambda k/d$ and consider the ideal signal mixture

$$P^\star = \frac{1}{M}\sum_{m=0}^{M-1} \mathcal{N}(0, m\delta I_d),$$

where M is a sufficiently large absolute constant. Under observation-noise variance a_0, the resulting marginal consists of Gaussian components with variance levels $a_0, a_0+\delta, \ldots, a_0+(M-1)\delta$.

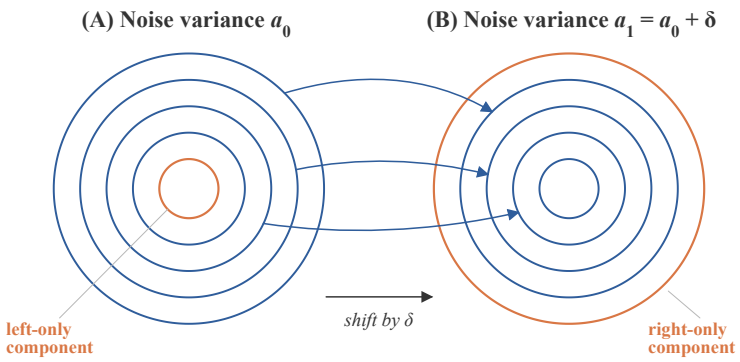

Figure 1: **Intrinsic-dimensional lower-bound construction.** Increasing the observation-noise variance from a_0 to $a_0 + \delta$ shifts each Gaussian component by one variance increment. The $M - 1$ interior components align across the two observation laws, leaving only the two endpoint components unmatched, each with weight $1/M$.

If the observation-noise variance is increased to $a_1 = a_0 + \delta$, the same components appear at $a_0 + \delta, \ldots, a_0 + M\delta$. Hence the $M - 1$ interior components coincide exactly, and

$$Q_0^\star - Q_1^\star = \frac{1}{M} \left\{ \mathcal{N}(0, a_0 I_d) - \mathcal{N}(0, (a_0 + M\delta) I_d) \right\}.$$

Therefore $\|Q_0^\star - Q_1^\star\|_{\mathrm{TV}} \le 1/M$. Figure 1 illustrates the overlapping components.

This mixture construction is related to lower bounds for noise-level estimation in semiparametric convolution models, where latent Gaussian variability can mask changes in the noise variance (Matias, 2002; Butucea & Matias, 2005). Those constructions, however, allow the signal law to vary between the hypotheses, whereas here the same fixed signal law must conceal both noise levels.

The ideal mixture has unbounded support and therefore does not satisfy the covering-number condition. We replace each Gaussian component by a finite empirical point cloud whose Gaussian smoothing is close to the corresponding population component in χ^2 divergence. A point-cloud size exponential in k gives the required approximation while keeping the total number of support points below e^k, and Gaussian concentration keeps the support radius at order $\sqrt{k}$. This approximation uses the same Gaussian-smoothed empirical-measure principle studied by Goldfeld et al. (2020), with the approximation measured here in χ^2 divergence. The resulting deterministic prior belongs to $\mathcal{P}_{\mathrm{LD}}(k, R)$ and preserves a constant overlap in total variation between the two observation laws. Since $\delta \asymp_\Lambda k/d$ corresponds to a precision separation of the same order on the compact interval Λ, the construction yields the k^2/d^2 term in Theorem 3.1 through the standard two-point method (Tsybakov, 2009).

For strong log-concavity, the class contains the Gaussian law $P_\alpha = \mathcal{N}(0, \alpha^{-1} I_d)$, under which $Y \sim \mathcal{N}(0, (\alpha^{-1} + \lambda^{-1}) I_d)$. A standard two-point comparison within this submodel gives the scale $(1 + \alpha^{-1})^2/d$ (Tsybakov, 2009).

5 Conclusion

In this work, we characterized the minimax risk of estimating a Gaussian noise level from one high-dimensional observation with a known, structurally constrained signal law. Under low intrinsic dimensionality, in the regime of Theorem 3.1, the minimax risk is $\widetilde{\Theta}_\Lambda(d^{-1} + k^2 d^{-2})$ up to saturation by Δ_Λ^2, and the fixed-P MLE attains this rate up to logarithmic factors. Under strong log-concavity, the minimax risk is $\Theta_\Lambda((1 + \alpha^{-1})^2/d)$ up to the same saturation, and a P-centered norm estimator is minimax optimal. Thus, low-dimensional support is sufficient but not necessary for accurate estimation of the noise level. More broadly, these results show that the difficulty of blind noise estimation is governed by the structure of the signal law rather than by ambient dimension alone, and that different structural assumptions naturally lead to different estimation procedures.

The analysis assumes an exactly known signal law and isotropic Gaussian noise with a single unknown scalar level. How the rates change when the signal law is learned or misspecified, as in practical diffusion models, remains open. Other directions include anisotropic or non-Gaussian corruption, as well as signal classes beyond low intrinsic dimensionality and strong log-concavity. A broader open question is to characterize which properties of the signal law make blind noise estimation possible. The separation-based argument for the strongly log-concave case suggests one possible route: establish direct quantitative separation between the corresponding noisy marginals.

AI USE STATEMENT

Generative AI tools were used in this work to assist with research ideation, literature retrieval and discovery, translation, drafting and polishing portions of the manuscript, and checking and developing mathematical arguments and proofs. AI-assisted suggestions were independently reviewed against the relevant literature and mathematical derivations. In particular, all mathematical claims and proofs were independently checked, and all cited references were independently verified. Full responsibility is taken for the final content of the paper, including all statements, proofs, and citations.

REPRODUCIBILITY STATEMENT

This work is theoretical and does not rely on empirical datasets or numerical experiments. The statistical formulation and assumptions are given in Section 2, and the main results are stated in Section 3. Section 4 provides an overview of the proof techniques. Complete proofs and supporting lemmas are provided in Appendices A, B, and C.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014.

Cristina Butucea and Catherine Matias. Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, 11(2):309–340, 2005.

Guangyong Chen, Fengyuan Zhu, and Pheng-Ann Heng. An efficient statistical method for image noise level estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 477–485, 2015.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4735–4763. PMLR, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

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

Matthias Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Abhiram Gnanasambandam and Stanley H. Chan. One size fits all: Can we train one denoiser for all noise levels? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3576–3586. PMLR, 2020.

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

David Heurtel-Depeiges, Charles Margossian, Ruben Ohana, and Bruno Régaldo-Saint Blancard. Listening to the noise: Blind denoising with Gibbs diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18284–18304. PMLR, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *Mathematics of Operations Research*, 2026.

Zahra Kadkhodaie and Eero P. Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13242–13254, 2021.

Zahra Kadkhodaie, Aram-Alexandre Pooladian, Sinho Chewi, and Eero P. Simoncelli. Blind denoising diffusion models and the blessings of dimensionality. *arXiv preprint arXiv:2602.09639*, 2026.

Cheuk-Kit Lee, Florentin Coeurdoux, Yuyuan Chen, Sophia Tang, Peter Potaptchik, Yilun Du, Michael Samuel Albergo, and Eric Vanden-Eijnden. Beckmann transport models: From autonomous flows to one-step maps. *arXiv preprint arXiv:2608.01692*, 2026.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 2023.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126297–126331, 2024.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation (extended abstract). In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3723–3729. PMLR, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Single-image noise level estimation for blind denoising. *IEEE Transactions on Image Processing*, 22(12):5226–5237, 2013.

Catherine Matias. Semiparametric deconvolution with unknown noise variance. *ESAIM: Probability and Statistics*, 6:271–292, 2002.

Koichi Miyasawa. An empirical bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38:181–188, 1961.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4668–4685. PMLR, 2025.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45–114, 2014.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57469–57502. PMLR, 2025.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4958–4991. PMLR, 2024.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

A Preliminaries

Throughout the appendix, let

$$\Lambda = [\lambda_{\min}, \lambda_{\max}], \qquad \Delta_\Lambda := \lambda_{\max} - \lambda_{\min},$$

and write $A \lesssim_\Lambda B$ when $A \le C_\Lambda B$ for a constant depending only on the fixed interval Λ. This appendix gives the likelihood identities, testing inequalities, and moment bounds used to prove Theorems 3.1 and 3.2.

A.1 Observation law and likelihood identities

Let P be a probability measure on $\mathbb{R}^d$, and consider

$$Y = X + \lambda^{-1/2} Z, \qquad X \sim P, \qquad Z \sim \mathcal{N}(0, I_d),$$

where X and Z are independent. The observation law is

$$P_\lambda := P * \mathcal{N}(0, \lambda^{-1} I_d),$$

with density

$$p_\lambda(y) = \left(\frac{\lambda}{2\pi}\right)^{d/2} \int \exp\left(-\frac{\lambda}{2}\|x - y\|^2\right) P(dx).$$

The posterior distribution of X given $Y = y$ is

$$q_\lambda(dx \mid y) = \frac{\exp\left(-\frac{\lambda}{2}\|x - y\|^2\right) P(dx)}{\int \exp\left(-\frac{\lambda}{2}\|u - y\|^2\right) P(du)}.$$

Define the negative log-likelihood and posterior quadratic error by

$$\ell(\lambda \mid y) := -\log p_\lambda(y), \qquad D_\lambda(y) := \mathbb{E}_{q_\lambda(\cdot|y)}\|X - y\|^2.$$

Lemma A.1 (Score and monotonicity identities). *For every $y \in \mathbb{R}^d$ and $\lambda > 0$,*

$$2\ell'(\lambda \mid y) = -\frac{d}{\lambda} + D_\lambda(y),$$

and

$$\partial_\lambda D_\lambda(y) = -\frac{1}{2}\operatorname{Var}_{q_\lambda(\cdot|y)}\big(\|X-y\|^2\big) \le 0.$$

In particular, $\lambda \mapsto D_\lambda(y)$ is nonincreasing.

Proof. Differentiating the marginal log-density gives

$$\partial_\lambda \log p_\lambda(y) = \frac{d}{2\lambda} - \frac{1}{2}\mathbb{E}_{q_\lambda(\cdot|y)}\|X-y\|^2,$$

which proves the score identity. More generally, for any integrable function h,

$$\partial_\lambda \mathbb{E}_{q_\lambda(\cdot|y)} h(X) = -\frac{1}{2}\operatorname{Cov}_{q_\lambda(\cdot|y)}\big(h(X), \|X-y\|^2\big).$$

Taking $h(X) = \|X-y\|^2$ gives the second identity. □

A.2 Divergences and two-point lower bounds

For probability measures Q and Q', we use

$$\|Q-Q'\|_{\mathrm{TV}} := \sup_A |Q(A)-Q'(A)|, \qquad \mathrm{KL}(Q\|Q') := \int \log\left(\frac{dQ}{dQ'}\right) dQ,$$

and

$$\chi^2(Q\|Q') := \int \left(\frac{dQ}{dQ'}-1\right)^2 dQ'.$$

The order-$1/2$ Rényi divergence is

$$D_{1/2}(Q\|Q') := -2\log \int \sqrt{\frac{dQ}{d\nu}\frac{dQ'}{d\nu}}\, d\nu,$$

where ν is any common dominating measure.

We repeatedly use the standard comparisons

$$\|Q-Q'\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\mathrm{KL}(Q\|Q')}, \qquad \|Q-Q'\|_{\mathrm{TV}} \le \frac{1}{2}\sqrt{\chi^2(Q\|Q')},$$

as well as the data-processing inequality for relative entropy.

Lemma A.2 (Two-point lower bound). *Let Q_0, Q_1 be two observation laws associated with parameters $\theta_0, \theta_1 \in \mathbb{R}$. Then*

$$\inf_{\widehat{\theta}} \max_{i\in\{0,1\}} \mathbb{E}_{Q_i}\big(\widehat{\theta}-\theta_i\big)^2 \ge \frac{|\theta_1-\theta_0|^2}{8}\big(1-\|Q_0-Q_1\|_{\mathrm{TV}}\big).$$

Proof. This is the standard two-point reduction; see Tsybakov (2009, Chapter 2). □

For Gaussian laws with a common mean, if $v_0, v_1 > 0$, then

$$\mathrm{KL}\left(\mathcal{N}(m, v_0 I_d) \,\|\, \mathcal{N}(m, v_1 I_d)\right) = \frac{d}{2}\left(\frac{v_0}{v_1} - 1 - \log\frac{v_0}{v_1}\right).$$

The common mean does not affect the divergence.

A.3 Radial moments and strong log-concavity

The variance of the norm statistic depends on the signal's radial moments, as the following identity shows.

Lemma A.3 (Radial variance identity)**.** *Suppose that X has finite fourth moment, let $m \in \mathbb{R}^d$ be deterministic, and write $a = \lambda^{-1}$. Then*

$$\mathbb{E}_{\lambda,P}\|Y - m\|^2 = \mathbb{E}_P\|X - m\|^2 + ad,$$

and

$$\mathrm{Var}_{\lambda,P}\big(\|Y - m\|^2\big) = \mathrm{Var}_P\big(\|X - m\|^2\big) + 4a\,\mathbb{E}_P\|X - m\|^2 + 2a^2 d.$$

Taking $m = 0$ gives the identity used in the discussion of low intrinsic dimensionality; taking $m = \mu_P$ gives its centered form for the strongly log-concave case.

Proof. Set $U := X - m$. Since

$$Y - m = U + \sqrt{a}\,Z,$$

we have

$$\|Y - m\|^2 = \|U\|^2 + 2\sqrt{a}\langle U, Z\rangle + a\|Z\|^2.$$

The three terms on the right are pairwise uncorrelated. Moreover,

$$\mathbb{E}\langle U, Z\rangle^2 = \mathbb{E}_P\|U\|^2, \qquad \mathrm{Var}(\|Z\|^2) = 2d.$$

Taking expectations and variances proves the identities. □

Strong log-concavity gives the following bounds on these moments.

Lemma A.4 (Consequences of strong log-concavity)**.** *Let*

$$P(dx) = Z_V^{-1} e^{-V(x)}dx, \qquad \nabla^2 V(x) \succeq \alpha I_d.$$

Then

$$\mathrm{Cov}_P(X) \preceq \alpha^{-1} I_d, \qquad \mathbb{E}_P\|X - \mu_P\|^2 \le \frac{d}{\alpha},$$

and

$$\mathrm{Var}_P\big(\|X - \mu_P\|^2\big) \le \frac{4d}{\alpha^2}.$$

Proof. Strong log-concavity implies the Poincaré inequality

$$\mathrm{Var}_P(f) \le \frac{1}{\alpha}\mathbb{E}_P\|\nabla f(X)\|^2$$

for every sufficiently regular f; see Bakry et al. (2014). Applying it to $f_v(x) = \langle v, x\rangle$ yields

$$v^\top \mathrm{Cov}_P(X) v \le \frac{\|v\|^2}{\alpha},$$

and hence

$$\mathbb{E}_P\|X - \mu_P\|^2 = \mathrm{tr}\,\mathrm{Cov}_P(X) \le \frac{d}{\alpha}.$$

Finally, applying the Poincaré inequality to $f(x) = \|x - \mu_P\|^2$ gives

$$\mathrm{Var}_P\big(\|X - \mu_P\|^2\big) \le \frac{4}{\alpha}\mathbb{E}_P\|X - \mu_P\|^2 \le \frac{4d}{\alpha^2}.$$

□

B Proofs for Low Intrinsic Dimensionality

Throughout this section, let

$$\Lambda = [\lambda_{\min}, \lambda_{\max}], \qquad \Delta_\Lambda := \lambda_{\max} - \lambda_{\min}, \qquad r_0 := \frac{\sqrt{\lambda_{\min}}}{\lambda_{\max}\sqrt{d}},$$

and write

$$P_\lambda := P * \mathcal{N}(0, \lambda^{-1} I_d)$$

for the observation law under precision λ. To prove Theorem 3.1, we establish the universal parametric lower bound, construct the additional lower bound from covering complexity, and bound the MLE risk. The final subsection gives the norm-statistic calculation used in Section 4.1.

B.1 UNIVERSAL PARAMETRIC LOWER BOUND

The parametric lower bound holds for every fixed signal law P.

Proposition B.1 (Universal parametric lower bound). *For every probability measure P on $\mathbb{R}^d$,*

$$\mathcal{R}(P;\Lambda) \gtrsim \min\left\{\Delta_\Lambda^2,\ \frac{\lambda_{\min}^2}{d}\right\}.$$

Consequently,

$$\mathcal{R}^\star_{\mathrm{LD}}(d,k,R;\Lambda) \gtrsim \min\left\{\Delta_\Lambda^2,\ \frac{\lambda_{\min}^2}{d}\right\}.$$

Proof. Fix a signal law P and consider two nearby precision parameters. Let

$$\lambda_0 := \lambda_{\min}, \qquad \lambda_1 := \lambda_0 + 2\delta,$$

where we choose $\delta > 0$ below to keep the two observation laws close.

Let $\mu_i := P_{\lambda_i}$, $i \in \{0,1\}$. Augment the observation Y with the latent signal X. Under hypothesis i, the joint law of (X,Y) is

$$\widetilde{\mu}_i(dx,dy) = P(dx)\,\mathcal{N}(x,\lambda_i^{-1}I_d)(dy).$$

Since μ_i is the Y-marginal of $\widetilde{\mu}_i$, the data-processing inequality gives

$$\mathrm{KL}(\mu_0\|\mu_1) \le \mathrm{KL}(\widetilde{\mu}_0\|\widetilde{\mu}_1).$$

The distribution of X is the same under both hypotheses, so the chain rule for relative entropy yields

$$\begin{aligned}\mathrm{KL}(\widetilde{\mu}_0\|\widetilde{\mu}_1) &= \mathbb{E}_{X\sim P}\mathrm{KL}\left(\mathcal{N}(X,\lambda_0^{-1}I_d)\,\middle\|\,\mathcal{N}(X,\lambda_1^{-1}I_d)\right)\\ &= \mathrm{KL}\left(\mathcal{N}(0,\lambda_0^{-1}I_d)\,\middle\|\,\mathcal{N}(0,\lambda_1^{-1}I_d)\right)\\ &= \frac{d}{2}\left(\frac{\lambda_1}{\lambda_0} - 1 - \log\frac{\lambda_1}{\lambda_0}\right).\end{aligned}$$

Thus the divergence does not depend on the common mean.

Choose the separation by setting

$$\delta := c\min\left\{\Delta_\Lambda,\ \frac{\lambda_{\min}}{\sqrt{d}}\right\},$$

where $c > 0$ is a sufficiently small absolute constant. Then $\lambda_1 \in \Lambda$, and $2\delta/\lambda_0 \le 2c$. Using $u - \log(1+u) \le Cu^2$ for sufficiently small $u \ge 0$, we obtain

$$\mathrm{KL}(\mu_0\|\mu_1) \le Cd\frac{\delta^2}{\lambda_0^2} \le Cc^2.$$

For sufficiently small c, the KL divergence is bounded by a fixed numerical constant. Pinsker's inequality then gives $\|\mu_0 - \mu_1\|_{\mathrm{TV}} \le c_0$ for some absolute $c_0 < 1$.

The standard two-point reduction for squared-error estimation therefore implies

$$\inf_{\widehat{\lambda}_P}\sup_{i\in\{0,1\}}\mathbb{E}_{\mu_i}\left[(\widehat{\lambda}_P - \lambda_i)^2\right] \gtrsim |\lambda_1 - \lambda_0|^2\big(1 - \|\mu_0 - \mu_1\|_{\mathrm{TV}}\big),$$

see, e.g., Tsybakov (2009, Chapter 2). Since $|\lambda_1 - \lambda_0| = 2\delta$, this yields

$$\mathcal{R}(P;\Lambda) \gtrsim \delta^2 \asymp \min\left\{\Delta_\Lambda^2,\ \frac{\lambda_{\min}^2}{d}\right\}.$$

Since P was arbitrary, we can take the supremum over $P \in \mathcal{P}_{\mathrm{LD}}(k,R)$. □

B.2 LOWER BOUND FROM COVERING COMPLEXITY

For the second lower bound, we construct one signal law with low covering complexity for which the observation laws remain close in total variation at the required precision separation.

Lemma B.2 (Intrinsic-dimensional indistinguishable pair)**.** *There exists a sufficiently large absolute constant k_0 such that the following holds. Assume $k_0 \le k \lesssim_\Lambda d$ and $R^2 \gtrsim_\Lambda k$. Then there exist $P \in \mathcal{P}_{\mathrm{LD}}(k, R)$ and $\lambda_0, \lambda_1 \in \Lambda$ such that*

$$|\lambda_1 - \lambda_0| \gtrsim \min\left\{\Delta_\Lambda,\, \lambda_{\min}\frac{k}{d}\right\},$$

while

$$\|P_{\lambda_0} - P_{\lambda_1}\|_{\mathrm{TV}} \le c_0$$

for some absolute constant $c_0 < 1$.

The construction has an ideal Gaussian mixture and a finite-support approximation.

Ideal Gaussian-mixture construction. It is convenient to work with the noise variance $a = \lambda^{-1}$. Recall that the admissible variance interval is

$$\mathcal{A}_\Lambda := [\lambda_{\max}^{-1}, \lambda_{\min}^{-1}].$$

Set

$$a_0 := \lambda_{\max}^{-1}, \qquad \tau := \min\left\{\frac{k}{d},\, \frac{\Delta_\Lambda}{\lambda_{\min}},\, 1\right\},$$

and let

$$\delta := ca_0\tau, \qquad a_1 := a_0 + \delta, \qquad \lambda_i := a_i^{-1}, \quad i \in \{0, 1\},$$

where $c > 0$ is a sufficiently small absolute constant.

Since

$$\lambda_{\min}^{-1} - \lambda_{\max}^{-1} = \frac{\Delta_\Lambda}{\lambda_{\min}\lambda_{\max}} = a_0\frac{\Delta_\Lambda}{\lambda_{\min}},$$

the definition of τ implies

$$\delta \le c\left(\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right).$$

Thus, for $c \le 1$, we have $a_1 \in \mathcal{A}_\Lambda$. Moreover, $\delta \le ca_0$, and hence

$$|\lambda_1 - \lambda_0| = \frac{\delta}{a_0(a_0 + \delta)} = \frac{c\lambda_{\max}\tau}{1 + c\tau}.$$

Since $\tau \le 1$, this gives

$$|\lambda_1 - \lambda_0| \gtrsim \lambda_{\max}\tau \gtrsim \min\left\{\Delta_\Lambda,\, \lambda_{\min}\frac{k}{d}\right\}.$$

It remains to construct a prior with low covering complexity that keeps the two observation laws close in total variation.

Fix an integer $M \ge 10$, to be chosen as a sufficiently large absolute constant, and use the convention $\mathcal{N}(0, 0I_d) = \delta_0$. Consider the ideal signal distribution

$$P^\star := \frac{1}{M}\sum_{m=0}^{M-1}\mathcal{N}(0, m\delta I_d).$$

For $i \in \{0, 1\}$, let

$$Q_i^\star := P^\star * \mathcal{N}(0, a_i I_d).$$

By Gaussian convolution,

$$Q_0^\star = \frac{1}{M}\sum_{m=0}^{M-1}\mathcal{N}(0, (a_0 + m\delta)I_d),$$

whereas

$$Q_1^\star = \frac{1}{M}\sum_{m=0}^{M-1}\mathcal{N}(0,(a_0+(m+1)\delta)I_d).$$

Hence the two mixtures share $M-1$ of their M Gaussian components and differ only in their two endpoint components:

$$Q_0^\star - Q_1^\star = \frac{1}{M}\left\{\mathcal{N}(0,a_0 I_d) - \mathcal{N}(0,(a_0+M\delta)I_d)\right\}.$$

Therefore

$$\|Q_0^\star - Q_1^\star\|_{\mathrm{TV}} \le \frac{1}{M}.$$

The ideal signal law $P^\star$ has the required overlap but unbounded support. To obtain an admissible prior, we replace each Gaussian component by a finite point cloud.

Finite-support approximation. For each $m \in \{0,\dots,M-1\}$, choose points $x_{1,m},\dots,x_{J,m} \in \mathbb{R}^d$ and define

$$\widehat{P}_m := \frac{1}{J}\sum_{j=1}^{J}\delta_{x_{j,m}}.$$

Define the approximating prior by

$$P := \frac{1}{M}\sum_{m=0}^{M-1}\widehat{P}_m = \frac{1}{JM}\sum_{m=0}^{M-1}\sum_{j=1}^{J}\delta_{x_{j,m}}.$$

Since P has at most JM support points,

$$N(\mathrm{supp}(P), r_0) \le JM.$$

Thus the covering requirement

$$1+\log N(\mathrm{supp}(P), r_0) \le k$$

holds whenever

$$JM \le e^{k-1}, \qquad \text{equivalently} \qquad J \le \frac{e^{k-1}}{M}.$$

For $a>0$, define the smoothed empirical layer

$$\widehat{Q}_{m,a} := \widehat{P}_m * \mathcal{N}(0,aI_d) = \frac{1}{J}\sum_{j=1}^{J}\mathcal{N}(x_{j,m},aI_d),$$

and let

$$G_{m,a} := \mathcal{N}(0,(a+m\delta)I_d).$$

We choose the support points so that $\widehat{Q}_{m,a_i}$ approximates G_{m,a_i} for every layer m and both $i \in \{0,1\}$.

Single-layer χ^2 approximation. Fix $m \in \{0,\dots,M-1\}$ and $a>0$, and set $b := m\delta$. Draw

$$X_{1,m},\dots,X_{J,m} \overset{\text{iid}}{\sim} \mathcal{N}(0,bI_d), \qquad \widehat{P}_m := \frac{1}{J}\sum_{j=1}^{J}\delta_{X_{j,m}}.$$

Lemma B.3 (Single-layer approximation)**.** *With the notation above,*

$$\mathbb{E}\left[\chi^2(\widehat{Q}_{m,a}\|G_{m,a})\right] = \frac{1}{J}\left[\left(1+\frac{b}{a}\right)^d - 1\right] \le \frac{1}{J}\exp\left(\frac{bd}{a}\right).$$

Proof. Let

$$\varphi_a(y-x) := (2\pi a)^{-d/2} \exp\left(-\frac{\|y-x\|^2}{2a}\right)$$

be the density of $\mathcal{N}(x, aI_d)$. The density of $\widehat{Q}_{m,a}$ is

$$\widehat{q}(y) = \frac{1}{J}\sum_{j=1}^{J} \varphi_a(y - X_{j,m}),$$

while the density of $G_{m,a}$ is $g(y) := \varphi_{a+b}(y)$. Since $X_{j,m} \sim \mathcal{N}(0, bI_d)$, Gaussian convolution gives

$$\mathbb{E}\varphi_a(y - X_{j,m}) = \varphi_{a+b}(y) = g(y),$$

and hence $\mathbb{E}\widehat{q}(y) = g(y)$.

By definition,

$$\chi^2(\widehat{Q}_{m,a}\|G_{m,a}) = \int \frac{(\widehat{q}(y) - g(y))^2}{g(y)}\, dy = \int \frac{\widehat{q}(y)^2}{g(y)}\, dy - 1.$$

Expanding $\widehat{q}^2$ and separating diagonal and off-diagonal terms gives

$$\mathbb{E}\int \frac{\widehat{q}(y)^2}{g(y)}\, dy = \frac{J(J-1)}{J^2} + \frac{1}{J}\,\mathbb{E}I(X_{1,m}),$$

where

$$I(x) := \int \frac{\varphi_a(y-x)^2}{g(y)}\, dy.$$

Therefore

$$\mathbb{E}\chi^2(\widehat{Q}_{m,a}\|G_{m,a}) = \frac{1}{J}\left\{\mathbb{E}I(X_{1,m}) - 1\right\}.$$

A direct Gaussian calculation yields

$$I(x) = \left(\frac{(a+b)^2}{a(a+2b)}\right)^{d/2} \exp\left(\frac{\|x\|^2}{a+2b}\right).$$

Since $X_{1,m} \sim \mathcal{N}(0, bI_d)$,

$$\mathbb{E}\exp\left(\frac{\|X_{1,m}\|^2}{a+2b}\right) = \left(\frac{a+2b}{a}\right)^{d/2}.$$

Consequently,

$$\mathbb{E}I(X_{1,m}) = \left(1 + \frac{b}{a}\right)^d,$$

and hence

$$\mathbb{E}\left[\chi^2(\widehat{Q}_{m,a}\|G_{m,a})\right] = \frac{1}{J}\left[\left(1+\frac{b}{a}\right)^d - 1\right].$$

The final inequality follows from $(1+u)^d \le e^{du}$ for $u \ge 0$. □

Simultaneous approximation and support control. We now choose a deterministic realization of the point clouds that satisfies the approximation bounds for all layers and both noise variances. For each $m \in \{0, \dots, M-1\}$, draw

$$X_{1,m}, \dots, X_{J,m} \overset{\text{iid}}{\sim} \mathcal{N}(0, m\delta I_d).$$

Lemma B.3 gives, for $a \in \{a_0, a_1\}$,

$$\mathbb{E}\left[\chi^2(\widehat{Q}_{m,a}\|G_{m,a})\right] \le \frac{1}{J}\exp\left(\frac{M\delta d}{a_0}\right).$$

Fix a sufficiently small absolute constant $C_0 > 0$ and a sufficiently large absolute constant M. Then choose $c > 0$ sufficiently small, depending only on M, and choose J large enough that

$$J \geq C_1 \exp\left(C_2 \frac{M\delta d}{a_0}\right),$$

where $C_1, C_2 > 0$ are sufficiently large constants depending only on C_0 and M. With these choices,

$$\mathbb{E}\left[\chi^2(\widehat{Q}_{m,a} \| G_{m,a})\right] \leq \frac{C_0}{8M}$$

for every m and $a \in \{a_0, a_1\}$. Markov's inequality therefore gives

$$\mathbb{P}\left(\chi^2(\widehat{Q}_{m,a} \| G_{m,a}) > C_0\right) \leq \frac{1}{8M}.$$

A union bound over the $2M$ pairs (m, a) shows that, with probability at least $3/4$,

$$\chi^2(\widehat{Q}_{m,a} \| G_{m,a}) \leq C_0$$

simultaneously for every m and $a \in \{a_0, a_1\}$. On this event,

$$\|\widehat{Q}_{m,a} - G_{m,a}\|_{\mathrm{TV}} \leq \sqrt{C_0},$$

where we used the standard comparison between total variation and χ^2 divergence.

We must also ensure that the support lies inside $B(0, R)$. Standard Gaussian concentration and a union bound imply that, with probability at least $3/4$,

$$\max_{0 \leq m \leq M-1} \max_{1 \leq j \leq J} \|X_{j,m}\|^2 \leq CM\delta\big\{d + \log(JM)\big\}.$$

The chosen variance increment must satisfy both the approximation and covering-number requirements. Recall that

$$\delta = ca_0\tau, \qquad \tau = \min\left\{\frac{k}{d}, \frac{\Delta_\Lambda}{\lambda_{\min}}, 1\right\}.$$

Since $\tau \leq k/d$,

$$\frac{M\delta d}{a_0} = Mc\tau d \leq Mck.$$

Hence, if the absolute constant $c > 0$ is sufficiently small, the requirements

$$C_1 \exp\left(C_2 \frac{M\delta d}{a_0}\right) \leq J \leq \frac{e^{k-1}}{M}$$

are compatible for all sufficiently large k. We may therefore choose an integer J satisfying both inequalities.

The upper bound $JM \leq e^{k-1}$ gives

$$1 + \log N(\mathrm{supp}(P), r_0) \leq k.$$

Moreover, $\log(JM) \leq k$, so the support estimate becomes

$$\max_{m,j} \|X_{j,m}\|^2 \leq CM\delta(d + k) \leq CMca_0 \frac{k}{d}(d + k) \lesssim_\Lambda k,$$

where we used $\tau \leq k/d$ and $k \lesssim_\Lambda d$. Hence $\mathrm{supp}(P) \subseteq B(0, R)$ whenever $R^2 \gtrsim_\Lambda k$.

The simultaneous approximation event and the support-control event each have probability at least $3/4$, so their intersection has positive probability. We can therefore fix a deterministic collection of support points $\{x_{j,m}\}$ satisfying both bounds.

Total-variation comparison. Fix such a deterministic realization and define

$$P = \frac{1}{JM} \sum_{m=0}^{M-1} \sum_{j=1}^{J} \delta_{x_{j,m}}.$$

For $i \in \{0, 1\}$, let

$$\mu_i := P * \mathcal{N}(0, a_i I_d).$$

For $a = a_0$,

$$\mu_0 = \frac{1}{M} \sum_{m=0}^{M-1} \widehat{Q}_{m,a_0}, \qquad Q_0^\star = \frac{1}{M} \sum_{m=0}^{M-1} G_{m,a_0}.$$

By convexity of total variation,

$$\|\mu_0 - Q_0^\star\|_{\mathrm{TV}} \le \frac{1}{M} \sum_{m=0}^{M-1} \|\widehat{Q}_{m,a_0} - G_{m,a_0}\|_{\mathrm{TV}} \le \sqrt{C_0}.$$

Similarly,

$$\|\mu_1 - Q_1^\star\|_{\mathrm{TV}} \le \sqrt{C_0}.$$

Combining these estimates with $\|Q_0^\star - Q_1^\star\|_{\mathrm{TV}} \le 1/M$ gives

$$\|\mu_0 - \mu_1\|_{\mathrm{TV}} \le \|\mu_0 - Q_0^\star\|_{\mathrm{TV}} + \|Q_0^\star - Q_1^\star\|_{\mathrm{TV}} + \|Q_1^\star - \mu_1\|_{\mathrm{TV}} \le 2\sqrt{C_0} + \frac{1}{M}.$$

Choosing $C_0 > 0$ sufficiently small and then M sufficiently large yields

$$\|\mu_0 - \mu_1\|_{\mathrm{TV}} \le c_0$$

for some absolute constant $c_0 < 1$.

Completion of the construction. For bounded k, the universal d^{-1} lower bound from Proposition B.1 dominates the contribution from covering complexity. We can therefore restrict to $k \ge k_0$ for a sufficiently large numerical constant k_0.

Recall that

$$a_0 = \lambda_{\max}^{-1}, \qquad \tau = \min\left\{\frac{k}{d}, \frac{\Delta_\Lambda}{\lambda_{\min}}, 1\right\}, \qquad \delta = c a_0 \tau,$$

and

$$a_1 = a_0 + \delta, \qquad \lambda_i = a_i^{-1}.$$

By construction, $a_0, a_1 \in \mathcal{A}_\Lambda$, and hence $\lambda_0, \lambda_1 \in \Lambda$. Furthermore,

$$|\lambda_1 - \lambda_0| = \frac{\delta}{a_0(a_0 + \delta)} = \frac{c\lambda_{\max}\tau}{1 + c\tau} \gtrsim \min\left\{\Delta_\Lambda, \lambda_{\min}\frac{k}{d}\right\}.$$

The deterministic point cloud constructed above satisfies both

$$1 + \log N(\mathrm{supp}(P), r_0) \le k$$

and $\mathrm{supp}(P) \subseteq B(0, R)$, so $P \in \mathcal{P}_{\mathrm{LD}}(k, R)$. The corresponding observation laws satisfy

$$\|P_{\lambda_0} - P_{\lambda_1}\|_{\mathrm{TV}} \le c_0 < 1.$$

This proves Lemma B.2.

Applying the standard two-point lower bound to the pair supplied by Lemma B.2 yields

$$\mathcal{R}^\star_{\mathrm{LD}}(d, k, R; \Lambda) \gtrsim \min\left\{\Delta_\Lambda^2, \lambda_{\min}^2 \frac{k^2}{d^2}\right\}.$$

Together with Proposition B.1, which gives

$$\mathcal{R}^\star_{\mathrm{LD}}(d, k, R; \Lambda) \gtrsim \min\left\{\Delta_\Lambda^2, \frac{\lambda_{\min}^2}{d}\right\},$$

we obtain

$$\mathcal{R}^\star_{\mathrm{LD}}(d, k, R; \Lambda) \gtrsim \min\left\{\Delta_\Lambda^2, \lambda_{\min}^2\left(\frac{1}{d} + \frac{k^2}{d^2}\right)\right\}.$$

Here we used the elementary equivalence

$$\max\{\min(A, u), \min(A, v)\} \asymp \min\{A, u + v\}$$

for nonnegative A, u, v.

B.3 MLE UPPER BOUND

We now prove the MLE upper bound. Fix $P \in \mathcal{P}_{\mathrm{LD}}(k, R)$ and let $\lambda^\star \in \Lambda$ denote the true precision, so that $Y \sim P_{\lambda^\star}$. Recall

$$\ell(\lambda \mid y) := -\log p_\lambda(y), \qquad D_\lambda(y) := \mathbb{E}_{q_\lambda(\cdot|y)}\|X - y\|^2,$$

and the score identity

$$2\ell'(\lambda \mid y) = -\frac{d}{\lambda} + D_\lambda(y).$$

Low intrinsic dimensionality bounds the posterior quadratic error uniformly over candidate precisions.

Lemma B.4 (Uniform posterior-energy approximation)**.** *Let $P \in \mathcal{P}_{\mathrm{LD}}(k, R)$, $\lambda^\star \in \Lambda$, and $Y \sim P_{\lambda^\star}$. For every $\eta \in (0, 1/2)$, with probability at least $1 - \eta$,*

$$\sup_{\lambda \in \Lambda} \left| D_\lambda(Y) - \frac{d}{\lambda^\star} \right| \le B_\eta,$$

where

$$B_\eta := \frac{C}{\lambda_{\min}} \left(\sqrt{d \log \frac{2}{\eta}} + k + \log \frac{2}{\eta} \right)$$

for a numerical constant $C > 0$.

Proof. Fix a candidate precision $\lambda \in \Lambda$ and write

$$\widetilde{\sigma} := (\lambda^\star)^{-1/2}, \qquad \sigma := \lambda^{-1/2}.$$

Proposition F.2 of Kadkhodaie et al. (2026), applied with $\sigma_T = \lambda_{\max}^{-1/2}$, $\sigma_0 = \lambda_{\min}^{-1/2}$, and $\ell = 2$, gives exactly the required bound for $D_\lambda(Y)$. In the present notation, with probability at least $1 - \delta$,

$$\left| D_\lambda(Y) - \frac{d}{\lambda^\star} \right| \le C \left[\frac{1}{\lambda^\star} \sqrt{d \log \frac{1}{\delta}} + \left(\frac{1}{\lambda^\star} + \frac{1}{\lambda} \right) \left(k + \log \frac{1}{\delta} \right) \right].$$

Since $\lambda, \lambda^\star \ge \lambda_{\min}$, the right-hand side is at most

$$\frac{C}{\lambda_{\min}} \left(\sqrt{d \log \frac{1}{\delta}} + k + \log \frac{1}{\delta} \right).$$

We apply this estimate only at the two endpoints $\lambda_{\min}$ and $\lambda_{\max}$, each with $\delta = \eta/2$. A union bound shows that, with probability at least $1 - \eta$,

$$\left| D_{\lambda_{\min}}(Y) - \frac{d}{\lambda^\star} \right| \le B_\eta, \qquad \left| D_{\lambda_{\max}}(Y) - \frac{d}{\lambda^\star} \right| \le B_\eta.$$

Monotonicity extends the endpoint bounds to the entire interval. Differentiating the posterior expectation with respect to λ gives

$$\partial_\lambda D_\lambda(y) = -\frac{1}{2} \operatorname{Var}_{q_\lambda(\cdot|y)} \left(\|X - y\|^2 \right) \le 0.$$

Hence $D_\lambda(y)$ is nonincreasing in λ, and therefore

$$D_{\lambda_{\max}}(Y) \le D_\lambda(Y) \le D_{\lambda_{\min}}(Y) \qquad \text{for every } \lambda \in \Lambda.$$

Since both endpoint values lie within B_η of $d/\lambda^\star$, the same holds uniformly over $\lambda \in \Lambda$. □

Proposition B.5 (MLE upper bound)**.** *Let $P \in \mathcal{P}_{\mathrm{LD}}(k, R)$ and*

$$\widehat{\lambda}_{\mathrm{MLE},P}(Y) \in \arg\min_{\lambda \in \Lambda} \ell(\lambda \mid Y).$$

For every $\lambda^\star \in \Lambda$ and $\eta \in (0, 1/2)$, with probability at least $1 - \eta$ under $Y \sim P_{\lambda^\star}$,

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right| \le C\frac{\lambda_{\max}^2}{\lambda_{\min}}\left(\sqrt{\frac{\log(2/\eta)}{d}} + \frac{k + \log(2/\eta)}{d}\right).$$

Consequently,

$$\sup_{\lambda^\star \in \Lambda} \mathbb{E}_{\lambda^\star, P}\left[\left(\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right)^2\right] \lesssim \min\left\{\Delta_\Lambda^2,\ \frac{\lambda_{\max}^4}{\lambda_{\min}^2}\left[\frac{\log(2d)}{d} + \frac{(k + \log(2d))^2}{d^2}\right]\right\}.$$

Proof. Define the Gaussian reference excess likelihood

$$\Psi_{\lambda^\star}(\lambda) := \frac{d}{2}\left[\frac{\lambda}{\lambda^\star} - 1 - \log\left(\frac{\lambda}{\lambda^\star}\right)\right].$$

It satisfies

$$\Psi_{\lambda^\star}(\lambda^\star) = \Psi'_{\lambda^\star}(\lambda^\star) = 0,$$

and

$$\Psi'_{\lambda^\star}(\lambda) = \frac{1}{2}\left(-\frac{d}{\lambda} + \frac{d}{\lambda^\star}\right), \qquad \Psi''_{\lambda^\star}(\lambda) = \frac{d}{2\lambda^2}.$$

Since $\lambda \le \lambda_{\max}$, Taylor's theorem gives

$$\Psi_{\lambda^\star}(\lambda) \ge \frac{d}{4\lambda_{\max}^2}(\lambda - \lambda^\star)^2.$$

Let B_η be as in Lemma B.4, and work on the event supplied by that lemma. By the score identity and the fundamental theorem of calculus,

$$\begin{aligned}\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y) - \Psi_{\lambda^\star}(\lambda) \\ = \frac{1}{2}\int_{\lambda^\star}^{\lambda}\left(D_u(Y) - \frac{d}{\lambda^\star}\right) du.\end{aligned}$$

Therefore

$$|\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y) - \Psi_{\lambda^\star}(\lambda)| \le \frac{B_\eta}{2}|\lambda - \lambda^\star|.$$

Combining this with the quadratic lower bound yields

$$\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y) \ge \frac{d}{4\lambda_{\max}^2}(\lambda - \lambda^\star)^2 - \frac{B_\eta}{2}|\lambda - \lambda^\star|.$$

By optimality of the constrained MLE,

$$\ell(\widehat{\lambda}_{\mathrm{MLE},P} \mid Y) \le \ell(\lambda^\star \mid Y).$$

Substituting $\lambda = \widehat{\lambda}_{\mathrm{MLE},P}$ into the preceding inequality gives

$$0 \ge \frac{d}{4\lambda_{\max}^2}\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right|^2 - \frac{B_\eta}{2}\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right|.$$

Hence

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right| \le C\frac{\lambda_{\max}^2}{d}B_\eta,$$

which, after substituting the expression for B_η, gives

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right| \le C\frac{\lambda_{\max}^2}{\lambda_{\min}}\left(\sqrt{\frac{\log(2/\eta)}{d}} + \frac{k + \log(2/\eta)}{d}\right).$$

To bound the mean-squared error using this probability estimate, assume first that $d \ge 2$ and take $\eta = d^{-3}$. On the event above,

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right|^2 \le C\frac{\lambda_{\max}^4}{\lambda_{\min}^2}\left[\frac{\log(2d)}{d} + \frac{(k + \log(2d))^2}{d^2}\right].$$

On the complementary event, both $\widehat{\lambda}_{\mathrm{MLE},P}$ and $\lambda^\star$ lie in Λ, so deterministically

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right|^2 \le \Delta_\Lambda^2.$$

Taking expectations gives

$$\mathbb{E}_{\lambda^\star,P}\left[\left(\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right)^2\right] \le C\frac{\lambda_{\max}^4}{\lambda_{\min}^2}\left[\frac{\log(2d)}{d} + \frac{(k+\log(2d))^2}{d^2}\right] + \Delta_\Lambda^2 d^{-3}.$$

Combining this estimate with the deterministic diameter bound gives the claimed minimum with Δ_Λ^2. The bounded-dimensional case follows after enlarging the numerical constant. The result is uniform over $\lambda^\star \in \Lambda$ and over fixed $P \in \mathcal{P}_{\mathrm{LD}}(k, R)$. □

B.4 Norm-based estimators

We give the norm-statistic calculation used in Section 4.1. This is not an impossibility result for all estimators depending on $\|Y\|^2$. The calculation concerns the standard energy-corrected norm statistic: its variance cannot be bounded uniformly in terms of the covering-complexity parameter k alone.

Write

$$a := \lambda^{-1}, \qquad Y = X + \sqrt{a}Z, \qquad Z \sim \mathcal{N}(0, I_d),$$

with Z independent of X. The raw norm statistic

$$\widetilde{a}_{\mathrm{raw}} := \frac{\|Y\|^2}{d}$$

satisfies

$$\mathbb{E}_{\lambda,P}\widetilde{a}_{\mathrm{raw}} = a + \frac{\mathbb{E}_P\|X\|^2}{d}.$$

Thus the raw norm confounds signal energy with noise energy. For example, if $P = \delta_\theta$ with $\|\theta\| = r \le R$, then P has covering complexity one but

$$\mathbb{E}_{\lambda,P}\widetilde{a}_{\mathrm{raw}} - a = \frac{r^2}{d}.$$

Since P is known, this bias can be removed by the energy-corrected statistic

$$\widetilde{a}_P(Y) := \frac{\|Y\|^2 - \mathbb{E}_P\|X\|^2}{d}, \qquad \mathbb{E}_{\lambda,P}\widetilde{a}_P(Y) = a.$$

By Lemma A.3 with $m = 0$,

$$\mathrm{Var}_{\lambda,P}(\widetilde{a}_P(Y)) = \frac{\mathrm{Var}_P(\|X\|^2) + 4a\,\mathbb{E}_P\|X\|^2 + 2a^2 d}{d^2}.$$

The covering-number condition gives no bound on the first term using k alone. For example, take

$$P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_u, \qquad \|u\| = r \le R.$$

This distribution has constant covering complexity, whereas

$$\mathbb{E}_P\|X\|^2 = \frac{r^2}{2}, \qquad \mathrm{Var}_P(\|X\|^2) = \frac{r^4}{4}.$$

Therefore

$$\mathrm{Var}_{\lambda,P}(\widetilde{a}_P(Y)) = \frac{r^4}{4d^2} + \frac{2ar^2}{d^2} + \frac{2a^2}{d}.$$

The covering assumption allows r to be as large as the support radius R. Taking $r \asymp R$, the first term is of order

$$\frac{R^4}{d^2}.$$

Since the class allows $R^2 \gg k$, this can be much larger than the intrinsic-dimensional scale k^2/d^2. Thus the covering condition alone does not provide a uniform bound on radial fluctuations in terms of k.

Finally, the map $a \mapsto a^{-1}$ has derivative bounded above and below on the compact interval $[\lambda_{\max}^{-1}, \lambda_{\min}^{-1}]$. Thus fluctuations in a variance estimator translate, up to factors determined by $\lambda_{\min}$ and $\lambda_{\max}$, into fluctuations of the same local order for the corresponding precision estimator.

The centered norm statistic therefore gives no uniform upper bound based only on covering complexity. The fixed-P MLE uses the full signal law and satisfies the bound proved above.

C Proofs for Strong Log-Concavity

For a fixed $P \in \mathcal{P}_{\mathrm{SLC}}(\alpha)$, let $\mu_P := \mathbb{E}_P X$. Replacing (X, Y) by $(X - \mu_P, Y - \mu_P)$ preserves the precision parameter and the strong log-concavity constant. We can therefore center the signal throughout this section, taking $\mathbb{E}_P X = 0$.

For $P \in \mathcal{P}_{\mathrm{SLC}}(\alpha)$, write

$$\Lambda = [\lambda_{\min}, \lambda_{\max}], \qquad \Delta_\Lambda := \lambda_{\max} - \lambda_{\min}.$$

With P centered so that $\mathbb{E}_P X = 0$, the observation law at precision $\lambda \in \Lambda$ is

$$P_\lambda = P * N(0, \lambda^{-1} I_d).$$

We also use

$$\rho_{\alpha,d}(\Lambda) := \min\left\{\Delta_\Lambda^2, \frac{(1+\alpha^{-1})^2}{d}\right\}.$$

C.1 Gaussian lower bound

For the lower bound in Theorem 3.2, consider the centered Gaussian signal law

$$P_\alpha := N(0, \alpha^{-1} I_d).$$

Its potential is $V(x) = \alpha\|x\|^2/2$, so $\nabla^2 V(x) = \alpha I_d$ and hence $P_\alpha \in \mathcal{P}_{\mathrm{SLC}}(\alpha)$. Under this fixed prior,

$$Y \sim N\big(0,\, (\alpha^{-1} + \lambda^{-1}) I_d\big)\,.$$

It suffices to bound from below the risk $\mathcal{R}(P_\alpha; \Lambda)$.

Let

$$\lambda_{\mathrm{mid}} := \frac{\lambda_{\min} + \lambda_{\max}}{2},$$

and let $h > 0$ be chosen below. Set

$$\lambda_0 := \lambda_{\mathrm{mid}} - h, \qquad \lambda_1 := \lambda_{\mathrm{mid}} + h,$$

and write

$$\mu_i := N(0, v_i I_d), \qquad v_i := \alpha^{-1} + \lambda_i^{-1}, \qquad i \in \{0, 1\}.$$

We will choose h small enough that $\lambda_0, \lambda_1 \in \Lambda$.

By Lemma A.2,

$$R(P_\alpha; \Lambda) \geq \frac{|\lambda_1 - \lambda_0|^2}{8}\left(1 - \|\mu_0 - \mu_1\|_{\mathrm{TV}}\right).$$

We choose h to keep the two Gaussian laws at total variation distance bounded away from one.

For centered Gaussians with scalar covariance matrices,

$$\mathrm{KL}(\mu_0 \| \mu_1) = \frac{d}{2}\left(\frac{v_0}{v_1} - 1 - \log\frac{v_0}{v_1}\right).$$

Since $\lambda_0 < \lambda_1$,

$$0 \leq \frac{v_0}{v_1} - 1 = \frac{\lambda_0^{-1} - \lambda_1^{-1}}{\alpha^{-1} + \lambda_1^{-1}} = \frac{2h}{\lambda_0(1 + \lambda_1/\alpha)} \leq C_\Lambda \frac{h}{1 + \alpha^{-1}}.$$

Hence, provided $h \le c_\Lambda(1+\alpha^{-1})$ for a sufficiently small $c_\Lambda > 0$, we have $1 \le v_0/v_1 \le 3/2$. Using $r - 1 - \log r \le C(r-1)^2$ on this interval gives

$$\mathrm{KL}(\mu_0 \| \mu_1) \le C_\Lambda \frac{dh^2}{(1+\alpha^{-1})^2}.$$

We now choose

$$h := c_\Lambda \min\left\{\Delta_\Lambda, \frac{1+\alpha^{-1}}{\sqrt{d}}\right\},$$

where $c_\Lambda > 0$ is sufficiently small. This choice ensures both $\lambda_0, \lambda_1 \in \Lambda$ and

$$\mathrm{KL}(\mu_0 \| \mu_1) \le \frac{1}{8}.$$

Pinsker's inequality then gives

$$\|\mu_0 - \mu_1\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\mathrm{KL}(\mu_0 \| \mu_1)} \le \frac{1}{4}.$$

The two-point lower bound therefore yields

$$R(P_\alpha; \Lambda) \gtrsim h^2 \asymp_\Lambda \min\left\{\Delta_\Lambda^2, \frac{(1+\alpha^{-1})^2}{d}\right\} = \rho_{\alpha,d}(\Lambda).$$

Since $P_\alpha \in \mathcal{P}_{\mathrm{SLC}}(\alpha)$,

$$\mathcal{R}^\star_{\mathrm{SLC}}(d, \alpha; \Lambda) \ge R(P_\alpha; \Lambda) \gtrsim_\Lambda \rho_{\alpha,d}(\Lambda),$$

which proves the lower bound in Theorem 3.2.

C.2 Norm-estimator upper bound

The matching upper bound follows from the P-centered norm estimator. By Lemma A.4,

$$\mathbb{E}_P\|X - \mu_P\|^2 \le \frac{d}{\alpha}, \qquad \mathrm{Var}_P\big(\|X - \mu_P\|^2\big) \le \frac{4d}{\alpha^2}.$$

Write $a = \lambda^{-1}$ and let

$$A_\Lambda := [\lambda_{\max}^{-1}, \lambda_{\min}^{-1}]$$

denote the admissible noise-variance interval. Define the P-centered statistic

$$\widetilde{a}_P(Y) := \frac{\|Y - \mu_P\|^2 - \mathbb{E}_P\|X - \mu_P\|^2}{d}.$$

Define its projection by

$$\widehat{a}_P(Y) := \Pi_{A_\Lambda}\big(\widetilde{a}_P(Y)\big),$$

and define the corresponding precision estimator by

$$\widehat{\lambda}_{\mathrm{norm},P}(Y) := \widehat{a}_P(Y)^{-1}.$$

Fix $\lambda \in \Lambda$ and write $a = \lambda^{-1}$. By Lemma A.3, applied with $m = \mu_P$,

$$\mathbb{E}_{\lambda,P}\widetilde{a}_P(Y) = a$$

and

$$\mathrm{Var}_{\lambda,P}\big(\|Y - \mu_P\|^2\big) = \mathrm{Var}_P\big(\|X - \mu_P\|^2\big) + 4a\,\mathbb{E}_P\|X - \mu_P\|^2 + 2a^2 d.$$

Using Lemma A.4 and $a \le \lambda_{\min}^{-1}$,

$$\begin{aligned}\mathbb{E}_{\lambda,P}\left[\big(\widetilde{a}_P(Y) - a\big)^2\right] = \frac{\mathrm{Var}_{\lambda,P}\big(\|Y - \mu_P\|^2\big)}{d^2} &\le \frac{1}{d}\left(\frac{4}{\alpha^2} + \frac{4}{\alpha\lambda_{\min}} + \frac{2}{\lambda_{\min}^2}\right) \\ &\lesssim_\Lambda \frac{(1+\alpha^{-1})^2}{d}.\end{aligned}$$

Since $a \in A_\Lambda$, Euclidean projection cannot increase the estimation error:

$$|\widehat{a}_P(Y) - a| \le |\widetilde{a}_P(Y) - a|.$$

Furthermore, both a and $\widehat{a}_P(Y)$ lie in A_Λ, and hence

$$\left|\widehat{\lambda}_{\mathrm{norm},P}(Y) - \lambda\right| = \left|\frac{1}{\widehat{a}_P(Y)} - \frac{1}{a}\right| = \frac{|\widehat{a}_P(Y) - a|}{a\widehat{a}_P(Y)} \le \lambda_{\max}^2 |\widehat{a}_P(Y) - a|.$$

Therefore

$$\sup_{\lambda\in\Lambda} \mathbb{E}_{\lambda,P}\left[\left(\widehat{\lambda}_{\mathrm{norm},P}(Y) - \lambda\right)^2\right] \lesssim_\Lambda \frac{(1+\alpha^{-1})^2}{d}.$$

Because both $\widehat{\lambda}_{\mathrm{norm},P}(Y)$ and λ belong to Λ, their squared distance is bounded by Δ_Λ^2. Combining the two bounds gives

$$\sup_{\lambda\in\Lambda} \mathbb{E}_{\lambda,P}\left[\left(\widehat{\lambda}_{\mathrm{norm},P}(Y) - \lambda\right)^2\right] \lesssim_\Lambda \min\left\{\Delta_\Lambda^2, \frac{(1+\alpha^{-1})^2}{d}\right\} = \rho_{\alpha,d}(\Lambda).$$

Since this holds for every fixed $P \in \mathcal{P}_{\mathrm{SLC}}(\alpha)$,

$$\mathcal{R}^\star_{\mathrm{SLC}}(d,\alpha;\Lambda) \lesssim_\Lambda \rho_{\alpha,d}(\Lambda).$$

Together with Appendix C.1, this proves the minimax-rate statement in Theorem 3.2.

C.3 Rényi separation for noisy marginals

To analyze the MLE, we first bound the separation of the noisy marginals. Fix a true precision $\lambda^\star \in \Lambda$ and suppose

$$Y \sim P_{\lambda^\star}.$$

Define

$$s_{\alpha,\Lambda} := \alpha^{-1} + \lambda_{\min}^{-1}.$$

Since Λ is fixed, $s_{\alpha,\Lambda} \asymp_\Lambda 1 + \alpha^{-1}$.

For $a > 0$, write

$$\mu_a := P * N(0, aI_d),$$

and denote its density by p_a. Gaussian convolution preserves strong log-concavity: if P is α-strongly log-concave, then μ_a is $(\alpha^{-1}+a)^{-1}$-strongly log-concave; see Saumard & Wellner (2014). Hence, for every $a \in [\lambda_{\max}^{-1}, \lambda_{\min}^{-1}]$,

$$-\nabla^2 \log p_a(y) \succeq \frac{1}{\alpha^{-1}+a} I_d \succeq \frac{1}{s_{\alpha,\Lambda}} I_d.$$

Thus every noisy marginal in the parameter interval is $s_{\alpha,\Lambda}^{-1}$-strongly log-concave.

Set

$$a := \lambda^{-1}, \qquad a^\star := (\lambda^\star)^{-1}.$$

Define the Hellinger midpoint distribution

$$\nu(dy) := \frac{\sqrt{p_a(y)p_{a^\star}(y)}}{\int \sqrt{p_a(z)p_{a^\star}(z)}\,dz}\,dy.$$

By direct calculation,

$$D_{1/2}(\mu_{a^\star}\|\mu_a) = \mathrm{KL}(\nu\|\mu_{a^\star}) + \mathrm{KL}(\nu\|\mu_a).$$

Since both $\mu_{a^\star}$ and μ_a are $s_{\alpha,\Lambda}^{-1}$-strongly log-concave, Talagrand's transportation inequality gives

$$\begin{aligned} D_{1/2}(\mu_{a^\star}\|\mu_a) &\ge \frac{1}{2s_{\alpha,\Lambda}}\left\{W_2^2(\nu,\mu_{a^\star}) + W_2^2(\nu,\mu_a)\right\} \\ &\ge \frac{1}{4s_{\alpha,\Lambda}} W_2^2(\mu_{a^\star},\mu_a), \end{aligned}$$

where the second inequality follows from the triangle inequality for W_2 and $u^2+v^2 \ge (u+v)^2/2$.

Let

$$\Sigma := \mathrm{Cov}_P(X),$$

and denote its eigenvalues by $\tau_1, \dots, \tau_d$. The Poincaré inequality gives

$$0 \le \tau_i \le \alpha^{-1}, \qquad i = 1, \dots, d.$$

Moreover,

$$\mathrm{Cov}(\mu_a) = \Sigma + aI_d, \qquad \mathrm{Cov}(\mu_{a^\star}) = \Sigma + a^\star I_d.$$

The covariance lower bound for the 2-Wasserstein distance (Gelbrich, 1990) therefore yields

$$\begin{aligned} W_2^2(\mu_{a^\star}, \mu_a) &\ge \sum_{i=1}^{d} \left(\sqrt{\tau_i + a^\star} - \sqrt{\tau_i + a}\right)^2 \\ &= (a - a^\star)^2 \sum_{i=1}^{d} \frac{1}{\left(\sqrt{\tau_i + a^\star} + \sqrt{\tau_i + a}\right)^2}. \end{aligned}$$

Since $\tau_i + a, \tau_i + a^\star \le s_{\alpha,\Lambda}$,

$$W_2^2(\mu_{a^\star}, \mu_a) \ge \frac{d(a - a^\star)^2}{4 s_{\alpha,\Lambda}}.$$

Combining the preceding bounds gives

$$D_{1/2}(\mu_{a^\star} \| \mu_a) \ge \frac{d(a - a^\star)^2}{16 s_{\alpha,\Lambda}^2}.$$

Finally,

$$|a - a^\star| = \frac{|\lambda - \lambda^\star|}{\lambda\lambda^\star} \ge \frac{|\lambda - \lambda^\star|}{\lambda_{\max}^2},$$

and hence

$$D_{1/2}\big(P_{\lambda^\star} \| P_\lambda\big) \ge c_\Lambda \frac{d(\lambda - \lambda^\star)^2}{s_{\alpha,\Lambda}^2}. \tag{C.1}$$

This bound quantifies the Rényi separation needed to localize the MLE.

C.4 Uniform likelihood-ratio control

We use the separation bound equation C.1 to control likelihood ratios uniformly over candidate precisions.

Recall that

$$\ell(\lambda \mid y) := -\log p_\lambda(y).$$

For a fixed candidate $\lambda \in \Lambda$ and any $u \ge 0$,

$$\begin{aligned} &\mathbb{P}_{\lambda^\star}\left(\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y) \le u\right) \\ &\quad = \mathbb{P}_{\lambda^\star}\left(\frac{p_\lambda(Y)}{p_{\lambda^\star}(Y)} \ge e^{-u}\right). \end{aligned}$$

Applying Markov's inequality to the square root of the likelihood ratio gives

$$\begin{aligned} \mathbb{P}_{\lambda^\star}\left(\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y) \le u\right) &\le e^{u/2} \mathbb{E}_{\lambda^\star}\left[\left(\frac{p_\lambda(Y)}{p_{\lambda^\star}(Y)}\right)^{1/2}\right] \\ &= \exp\left\{\frac{u}{2} - \frac{1}{2} D_{1/2}\big(P_{\lambda^\star} \| P_\lambda\big)\right\}. \end{aligned}$$

Using equation C.1 and adjusting c_Λ,

$$\mathbb{P}_{\lambda^\star}\left(\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y) \le u\right) \le \exp\left\{\frac{u}{2} - c_\Lambda \frac{d(\lambda - \lambda^\star)^2}{s_{\alpha,\Lambda}^2}\right\}. \tag{C.2}$$

It remains to make this bound uniform over $\lambda \in \Lambda$. By the score identity from Section 2,

$$\partial_\lambda \ell(\lambda \mid y) = -\frac{d}{2\lambda} + \frac{1}{2} D_\lambda(y),$$

where

$$D_\lambda(y) := \mathbb{E}_{q_\lambda(\cdot|y)} \|X - y\|^2.$$

We first show that

$$D_\lambda(y) \le \mathbb{E}_P \|X - y\|^2. \tag{C.3}$$

Let

$$U := \|X - y\|^2, \qquad w(U) := e^{-\lambda U/2}.$$

Then

$$D_\lambda(y) = \frac{\mathbb{E}_P[U w(U)]}{\mathbb{E}_P[w(U)]}.$$

If U' is an independent copy of U, then

$$2\,\mathrm{Cov}(U, w(U)) = \mathbb{E}\left[(U - U')\big(w(U) - w(U')\big)\right] \le 0,$$

because w is decreasing. Hence

$$\mathbb{E}_P[U w(U)] \le \mathbb{E}_P[U]\,\mathbb{E}_P[w(U)],$$

which proves equation C.3.

Since $\mathbb{E}_P X = 0$ and $\mathbb{E}_P \|X\|^2 \le d/\alpha$,

$$D_\lambda(y) \le \mathbb{E}_P \|X - y\|^2 = \mathbb{E}_P \|X\|^2 + \|y\|^2 \le \frac{d}{\alpha} + \|y\|^2.$$

Consequently,

$$\sup_{\lambda \in \Lambda} |\partial_\lambda \ell(\lambda \mid y)| \le C_\Lambda \left\{ s_{\alpha,\Lambda} d + \|y\|^2 \right\}. \tag{C.4}$$

Under the true precision $\lambda^\star$, the noisy marginal $P_{\lambda^\star}$ is $s_{\alpha,\Lambda}^{-1}$-strongly log-concave. Moreover,

$$\mathbb{E}_{\lambda^\star} \|Y\|^2 = \mathbb{E}_P \|X\|^2 + \frac{d}{\lambda^\star} \le d s_{\alpha,\Lambda}.$$

Standard concentration for strongly log-concave measures, applied to the 1-Lipschitz map $y \mapsto \|y\|$, therefore gives, for every $t \ge 1$,

$$\mathbb{P}_{\lambda^\star}\left(\|Y\|^2 > C_\Lambda s_{\alpha,\Lambda}(d + t) \right) \le e^{-t}. \tag{C.5}$$

On the complementary event, equation C.4 implies

$$\sup_{\lambda \in \Lambda} |\partial_\lambda \ell(\lambda \mid Y)| \le L_t, \qquad L_t := C_\Lambda s_{\alpha,\Lambda}(d + t).$$

Set

$$\varepsilon_t := L_t^{-1}$$

and choose an ε_t-net $\{\lambda_1, \dots, \lambda_{N_t}\}$ of Λ. Since Λ is a fixed compact interval,

$$N_t \le 1 + \frac{\Delta_\Lambda}{\varepsilon_t} \le C_\Lambda s_{\alpha,\Lambda}(d + t). \tag{C.6}$$

On the event in equation C.5, for every $\lambda \in \Lambda$ and a nearest net point λ_j,

$$|\ell(\lambda \mid Y) - \ell(\lambda_j \mid Y)| \le L_t \varepsilon_t = 1.$$

Suppose there exists $\lambda \in \Lambda$ such that

$$|\lambda - \lambda^\star| \ge r + \varepsilon_t$$

and

$$\ell(\lambda \mid Y) \le \ell(\lambda^\star \mid Y).$$

Its nearest net point then satisfies

$$|\lambda_j - \lambda^\star| \geq r$$

and

$$\ell(\lambda_j \mid Y) - \ell(\lambda^\star \mid Y) \leq 1.$$

Applying equation C.2 with $u = 1$ and taking a union bound over the net points yields

$$\mathbb{P}_{\lambda^\star}\left(\inf_{\substack{\lambda\in\Lambda:\\ |\lambda-\lambda^\star|\geq r+\varepsilon_t}} \{\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y)\} \leq 0\right) \leq e^{-t} + N_t \exp\left\{\frac{1}{2} - c_\Lambda \frac{dr^2}{s^2_{\alpha,\Lambda}}\right\}. \tag{C.7}$$

This controls the likelihood ratio uniformly over the interval.

C.5 MLE LOCALIZATION

The uniform bound equation C.7 gives the MLE risk bound in Theorem 3.2. Recall that

$$\widehat{\lambda}_{\mathrm{MLE},P}(Y) \in \arg\min_{\lambda\in\Lambda} \ell(\lambda \mid Y).$$

Because both $\widehat{\lambda}_{\mathrm{MLE},P}$ and $\lambda^\star$ belong to Λ,

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right|^2 \leq \Delta^2_\Lambda.$$

It remains to consider the regime in which the risk bound is nontrivial:

$$\frac{s^2_{\alpha,\Lambda}\log(2d)}{d} \leq \Delta^2_\Lambda. \tag{C.8}$$

Take

$$t := 3\log(2d).$$

By equation C.6,

$$N_t \leq C_\Lambda s_{\alpha,\Lambda}(d+t).$$

Under equation C.8,

$$\log N_t \lesssim_\Lambda \log(2d).$$

Choose

$$r_t := C_\Lambda s_{\alpha,\Lambda}\sqrt{\frac{\log(2d)}{d}},$$

with C_Λ sufficiently large. Then equation C.7 gives

$$\mathbb{P}_{\lambda^\star}\left(\inf_{\substack{\lambda\in\Lambda:\\ |\lambda-\lambda^\star|\geq r_t+\varepsilon_t}} \{\ell(\lambda \mid Y) - \ell(\lambda^\star \mid Y)\} \leq 0\right) \leq 2(2d)^{-3}. \tag{C.9}$$

By definition of the constrained MLE,

$$\ell(\widehat{\lambda}_{\mathrm{MLE},P} \mid Y) \leq \ell(\lambda^\star \mid Y).$$

Therefore, on the complement of the event in equation C.9,

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right| \leq r_t + \varepsilon_t.$$

Since

$$\varepsilon_t = \frac{1}{C_\Lambda s_{\alpha,\Lambda}(d+t)}$$

and $s_{\alpha,\Lambda} \gtrsim_\Lambda 1$, we have

$$\varepsilon_t \lesssim_\Lambda s_{\alpha,\Lambda}\sqrt{\frac{\log(2d)}{d}}.$$

Hence, with probability at least $1 - 2(2d)^{-3}$,

$$\left|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right| \lesssim_\Lambda s_{\alpha,\Lambda}\sqrt{\frac{\log(2d)}{d}}. \tag{C.10}$$

Using the deterministic bound $|\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star| \le \Delta_\Lambda$ on the complementary event, we obtain

$$\begin{aligned}\mathbb{E}_{\lambda^\star,P}\left[\left(\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right)^2\right] &\lesssim_\Lambda \frac{s_{\alpha,\Lambda}^2 \log(2d)}{d} + \Delta_\Lambda^2 (2d)^{-3} \\ &\lesssim_\Lambda \frac{s_{\alpha,\Lambda}^2 \log(2d)}{d}.\end{aligned}$$

The estimate is uniform over $\lambda^\star \in \Lambda$. Since $s_{\alpha,\Lambda} \asymp_\Lambda 1 + \alpha^{-1}$,

$$\sup_{\lambda^\star\in\Lambda} \mathbb{E}_{\lambda^\star,P}\left[\left(\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right)^2\right] \lesssim_\Lambda \frac{(1+\alpha^{-1})^2 \log(2d)}{d}$$

in the nontrivial regime. Combining this with the deterministic diameter bound gives, for every $d \ge 1$ and $\alpha > 0$,

$$\sup_{\lambda^\star\in\Lambda} \mathbb{E}_{\lambda^\star,P}\left[\left(\widehat{\lambda}_{\mathrm{MLE},P} - \lambda^\star\right)^2\right] \lesssim_\Lambda \min\left\{\Delta_\Lambda^2,\ \frac{(1+\alpha^{-1})^2 \log(2d)}{d}\right\}.$$

This proves the MLE statement in Theorem 3.2.