Proceedings of Machine Learning Research vol 291:1–56, 2025 38th Annual Conference on Learning Theory

Spike-and-Slab Posterior Sampling in High Dimensions

Syamantak Kumar SYAMANTAK@UTEXAS.EDU
University of Texas at Austin

Purnamrita Sarkar PURNA.SARKAR@UTEXAS.EDU
University of Texas at Austin

Kevin Tian KJTIAN@CS.UTEXAS.EDU
University of Texas at Austin

Yusong Zhu ZHUYS@UTEXAS.EDU
University of Texas at Austin

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Posterior sampling with the spike-and-slab prior (Mitchell and Beauchamp, 1988), a popular multimodal distribution used to model uncertainty in variable selection, is considered the theoretical gold standard method for Bayesian sparse linear regression (Carvalho et al., 2009; Rockova, 2018). However, designing provable algorithms for performing this sampling task is notoriously challenging. Existing posterior samplers for Bayesian sparse variable selection tasks either require strong assumptions about the signal-to-noise ratio (SNR) (Yang et al., 2016), only work when the measurement count grows at least linearly in the dimension (Montanari and Wu, 2024), or rely on heuristic approximations to the posterior.

We give the first provable algorithms for spike-and-slab posterior sampling that apply for any SNR, and use a measurement count sublinear in the problem dimension. Concretely, assume we are given a measurement matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and noisy observations $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\xi}$ of a signal $\boldsymbol{\theta}^\star$ drawn from a spike-and-slab prior π with a Gaussian diffuse density and expected sparsity k, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$. We give a polynomial-time high-accuracy sampler for the posterior $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$, for any SNR $\sigma^{-1} > 0$, as long as $n \geq k^3 \cdot \mathrm{polylog}(d)$ and $\mathbf{X}$ is drawn from a matrix ensemble satisfying the restricted isometry property. We further give a sampler that runs in near-linear time $\approx nd$ in the same setting, as long as $n \geq k^5 \cdot \mathrm{polylog}(d)$. To demonstrate the flexibility of our framework, we extend our result to spike-and-slab posterior sampling with Laplace diffuse densities, achieving similar guarantees when $\sigma = O(\frac{1}{k})$ is bounded.

Keywords: Posterior sampling, high-dimensional Bayesian statistics, sparse linear regression

1. Introduction

Sparse linear regression is a fundamental, well-studied problem in high-dimensional statistics. In this problem, there is an unknown sparse signal $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ that we are trying to estimate, and we are given a measurement matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, as well as noisy observations

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\xi}, \text{ where } \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n). \tag{1}$$

This problem has been studied for decades by statisticians from the perspective of *variable selection*, see e.g., Hocking (1976) and references therein, where $\mathrm{supp}(\boldsymbol{\theta}^\star)$ models the "important features" for a regression problem, and potentially $|\mathrm{supp}(\boldsymbol{\theta}^\star)| \ll d$. More recently, due to the increasing dimensionality of data in modern statistics and learning tasks, there has been significant interest in understanding the *computational complexity* of sparse linear regression.

One major algorithmic success story in sparse linear regression is a line of work building off breakthrough results of Candès and Romberg (2005); Candès and Tao (2005); Donoho (2006); Candès and Tao (2006); Candès et al. (2006) for compressed sensing, which we refer to collectively as "sparse recovery" methods (some of which are reviewed in Appendix A.3). These algorithms take a frequentist viewpoint of (1), and aim to return an estimate $\widehat{\boldsymbol{\theta}}$ such that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|$ for an appropriate norm $\|\cdot\|$, e.g., ℓ_2 or ℓ_∞, often via penalized maximum likelihood methods such as the Lasso (Tibshirani, 1996). Impressively, despite the nonconvexity of the problem (i.e., modeling the constraint that $\boldsymbol{\theta}^\star$ is sparse), the aforementioned line of work showed that as long as $\mathbf{X}$ satisfies the *restricted isometry property* (RIP, cf. Definition 7), a sparse vector $\boldsymbol{\theta}^\star$ can be recovered up to a function of the noise level σ in polynomial time. In fact, sparse recovery is tractable even in the *sublinear measurement* regime $n \ll d$, as long as n is sufficiently larger than the sparsity $|\mathrm{supp}(\boldsymbol{\theta}^\star)|$. This regime is natural for high-dimensional sparse linear regression, because its underconstrained nature necessitates making a structural assumption on the signal $\boldsymbol{\theta}^\star$, e.g., sparsity.

A potential shortcoming of sparse recovery methods is their inability to model uncertainty in variable selection. In particular, these frequentist algorithms commit to a single estimate $\widehat{\boldsymbol{\theta}}$, and do not return multiple plausible candidate supports. This issue is particularly manifest in the setting of moderate signal-to-noise ratios, where some coordinates of $\boldsymbol{\theta}^\star$ have magnitude at or around the noise level σ. In such settings, it is more appropriate to consider a distribution over estimates to reflect our uncertainty of $\boldsymbol{\theta}^\star$. This motivates the problem of *Bayesian sparse linear regression*, the main problem studied in this paper. In this problem, for some prior distribution π that is mostly supported on sparse vectors, the task is to return a sample from the posterior distribution,

$$\boldsymbol{\theta} \sim \pi\left(\cdot \mid \mathbf{X}, \mathbf{y}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right\|_2^2\right)\pi(\boldsymbol{\theta}). \tag{2}$$

Perhaps the most well-studied choice of prior in Bayesian sparse linear regression is the *spike-and-slab* prior, introduced by Mitchell and Beauchamp (1988) and subsequently studied in, e.g., George and McCulloch (1993); Chipman (1996); Geweke (1996). This prior is parameterized by $\mathbf{q} \in [0,1]^d$ and a *diffuse density* μ governing the non-negligible entries (the "slab," i.e., the selected variables). The prior then takes the product density form[1]

$$\pi := \bigotimes_{i \in [d]} \left((1 - \mathbf{q}_i)\delta_0 + \mathbf{q}_i \mu\right), \tag{3}$$

so that the i^{th} coordinate of $\boldsymbol{\theta}^\star \sim \pi$ is independently set to 0 except with probability $\mathbf{q}_i$. Throughout the paper, we define $k := \|\mathbf{q}\|_1$ to be the expected sparsity level of $\boldsymbol{\theta}^\star \sim \pi$; we are primarily interested in the regime $k \ll d$. Chernoff bounds show that $|\mathrm{supp}(\boldsymbol{\theta}^\star)| \lesssim k$ with high probability, so that at least the frequentist problem of sparse recovery is tractable with $n \ll d$ measurements.

1. We refer the reader to Section A.1 for notation used throughout the paper, and Section A.2 for the formal definition of our spike-and-slab posterior sampling problem.

Sampling from (2) when π has the form (3) (henceforth, *spike-and-slab posterior sampling*) is often referred to as the "theoretical ideal" or "gold standard" for modeling uncertainty in Bayesian variable selection (Johnstone and Silverman, 2004; Carvalho et al., 2009; Ishwaran and Rao, 2011; Castillo and Van Der Vaart, 2012; Rockova, 2018; Polson and Sun, 2019). While the favorable statistical properties of spike-and-slab posterior sampling are well-documented (see e.g. Castillo et al. (2015)), actually performing said sampling is notoriously challenging. This is in large part due to the combinatorial and nonconvex nature of (3), which causes the posterior (2) to be a mixture of potentially exponentially-many Gaussians. This has caused many researchers to consider approximations to (2), e.g., by relaxing the hard sparsity constraints, considering mean-field approximations, or using other heuristics (see Section 1.2 for further discussion, or Bai et al. (2021)).

We are aware of relatively few spike-and-slab posterior samplers that provably work without restrictive modeling assumptions. The two most relevant to us are Yang et al. (2016); Montanari and Wu (2024); we give an extended comparison to both in Section 1.2, but provide an overview here.

In the well-studied setting where $\mu = \mathcal{N}(0,1)$ in (3) is a Gaussian, Yang et al. (2016) (which formally studies a different family of sparse priors than (3)) gives a sampler that only works under very high or very low signal-to-noise ratios (SNR) σ^{-1}. In such regimes, the posterior distribution either collapses to a single candidate support (as all signal coordinates are above the noise level), or the posterior effectively reduces to the prior (as the noise overwhelms the signal). Both settings appear to be at odds with a major motivation behind Bayesian sparse linear regression, i.e., modeling partial uncertainty in variable selection. Conversely, while Montanari and Wu (2024) works for fairly general diffuse densities and SNRs (μ, σ^{-1}), it requires that the number of observations n grows at least linearly in the dimension d. This regime is again at odds with the high-dimensional, underparameterized setting $n \ll d$ where sparse modeling tasks are typically studied. To our knowledge, the following motivating question has remained largely unaddressed by the existing literature.

$$\begin{gathered}\textit{Is there a polynomial-time spike-and-slab posterior sampler} \\ \textit{that uses } n = o(d) \textit{ measurements and works for arbitrary SNRs } \sigma^{-1} > 0\textit{?}\end{gathered} \tag{4}$$

1.1. Our results

Gaussian prior. Our primary contribution is to resolve (4) when $\mu = \mathcal{N}(0,1)$ is a Gaussian, and the expected sparsity k is sufficiently small. For simplicity in this introduction, we only state our results in the case where the measurement matrix $\mathbf{X}$ has i.i.d. Gaussian entries. Our formal theorem statements extend the statements here broadly to any $\mathbf{X}$ drawn from any commonly-used RIP matrix ensemble (cf. Appendix A.3 for examples). This can yield improved runtimes when the ensemble supports fast matrix-vector multiplies, e.g., when $\mathbf{X}$ is a subsampled DFT matrix.

Theorem 1 (informal, see Theorem 26, Corollary 28) *Let $\mu = \mathcal{N}(0,1)$ in (3), and suppose $\mathbf{X} \in \mathbb{R}^{n\times d}$ has i.i.d. entries $\sim \mathcal{N}(0, \frac{1}{n})$. For any $\sigma > 0$, $\delta \in (0,1)$, and $\mathbf{q} \in [0,1]^d$, letting $k := \|\mathbf{q}\|_1$, the following hold for the problem of sampling from $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ (2).*

1. *If $n = \Omega(k^3 \operatorname{polylog}(\frac{d}{\delta}))$, there is an algorithm returning $\boldsymbol{\theta} \sim \pi'$ for $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{y}, \mathbf{X})) \le \delta$ with probability $\ge 1 - \delta$ over $\mathbf{X}$, $\boldsymbol{\theta}^\star$, and $\boldsymbol{\xi}$, in time $O(n^2 d^{1.5} \operatorname{polylog}(\frac{d}{\delta \min(1,\sigma)}))$.*
2. *If $n = \Omega(k^5 \operatorname{polylog}(\frac{d}{\delta}))$, there is an algorithm returning $\boldsymbol{\theta} \sim \pi'$ for $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{y}, \mathbf{X})) \le \delta$ with probability $\ge 1 - \delta$ over $\mathbf{X}$, $\boldsymbol{\theta}^\star$, and $\boldsymbol{\xi}$, which runs in time $O(nd \log(\frac{1}{\min(1,\sigma)}))$.*

We briefly remark on the form of Theorem 1. There are two sources of failure: the random model instance $(\mathbf{X}, \boldsymbol{\theta}^\star, \boldsymbol{\xi})$, and the inaccuracy of our sampler itself. With positive probability, the model will produce an instance with an intractable posterior sampling task (e.g., if $\mathbf{X}$ is not RIP or $\boldsymbol{\theta}^\star$ is not sparse). Assuming the success of the model, our sampler gives high-accuracy convergence guarantees to the true posterior (achieving total variation δ with a $\mathrm{polylog}(\frac{1}{\delta})$ dependence).

Our required sample complexity, or the number of random measurements n in Theorem 1, scales sublinearly in d for sufficiently small k. Qualitatively, this draws a parallel between Theorem 1 and results in the sparse recovery literature, which also estimate $\boldsymbol{\theta}^\star$ (via optimization instead of sampling) from $n = \mathrm{poly}(k, \log(\frac{d}{\delta}))$ measurements. It is plausible that (4) could even be answered affirmatively given the natural limit of $n \approx k$ observations. However, this would likely require fundamentally new techniques, which we leave as an exciting open question.

Laplace prior. Several qualitative properties of our Gaussian posterior sampler in Theorem 1 remain true for general diffuse densities μ in (3) (which we expand on in Section 2). For this reason, at least for a range of $\sigma > 0$, we believe our techniques extend to posterior sample from (2) for a fairly broad set of diffuse densities μ in (3). This is a desirable trait, shared e.g., by the sampler of Montanari and Wu (2024), which applies to the regime $n = \Omega(d)$. As a demonstration of our framework, in Appendix C, we show how to sample from (2) when $\mu = \mathrm{Lap}(0, 1)$ in (3).

Theorem 2 (informal, see Theorem 36, Corollary 37) *Let $\mu = \mathrm{Lap}(0, 1)$ in (3), and suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ has i.i.d. entries $\sim \mathcal{N}(0, \frac{1}{n})$. For any $\delta \in (0, 1)$, $\sigma \in (0, O((k + \log(\frac{1}{\delta}))^{-1})]$, and $\mathbf{q} \in [0, 1]^d$, letting $k := \|\mathbf{q}\|_1$, the following hold for the problem of sampling from $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ (2).*

1. *If $n = \Omega(k^3 \,\mathrm{polylog}(\frac{d}{\delta}))$, there is an algorithm returning $\boldsymbol{\theta} \sim \pi'$ for $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{y}, \mathbf{X})) \leq \delta$ with probability $\geq 1 - \delta$ over $\mathbf{X}, \boldsymbol{\theta}^\star$, and $\boldsymbol{\xi}$, in time $O(n^2 d^{1.5} \log(\frac{d}{\sigma\delta}) + \frac{k^4}{\delta^2} \,\mathrm{polylog}(\frac{d}{\sigma\delta}))$.*
2. *If $n = \Omega(k^5 \,\mathrm{polylog}(\frac{d}{\delta}))$, there is an algorithm returning $\boldsymbol{\theta} \sim \pi'$ for $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{y}, \mathbf{X})) \leq \delta$ with probability $\geq 1 - \delta$ over $\mathbf{X}, \boldsymbol{\theta}^\star$, and $\boldsymbol{\xi}$, in time $O(nd \log(\frac{d}{\sigma\delta}) + \frac{k^4}{\delta^2} \,\mathrm{polylog}(\frac{d}{\sigma\delta}))$.*

Theorems 1 and 2 have two main differences. First, the runtime of Theorem 2 scales polynomially with $\frac{1}{\delta}$ (in a low-order term), as opposed to Theorem 1's polylogarithmic scaling. This is because in the Gaussian setting, there is an explicit formula for the proportionality constant of each posterior component (corresponding to a possible $\mathrm{supp}(\boldsymbol{\theta}^\star)$). However, in the Laplace setting, we must rely on Monte Carlo estimation of normalizing constants, leading to a total variation error that scales with our estimate's approximation factor. Second, due to technical obstacles in the analysis of a rejection sampling step of our algorithm, Theorem 2 is only able to tolerate a bounded noise level $\sigma \lesssim \frac{1}{k}$. We think it is an interesting open direction to characterize the types of μ for which (4) can be answered affirmatively (potentially including $\mu = \mathrm{Lap}(0, 1)$).

1.2. Related work

Over the years, there has been a huge amount of work on high dimensional sparse regression in both the Bayesian and frequentist communities (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Efron et al., 2004; Wainwright, 2009; Bertsimas and Parys, 2020; Montanari and Wu, 2024). Under the Bayesian paradigm, some works focus on finding the mode of the posterior distribution in (2) (Ročková and George, 2014, 2018), but such an estimate ignores the uncertainty of the posterior. Directly sampling from the posterior, on the other hand, could address such drawbacks.

While there is an active line of research in this direction (Bhattacharya et al., 2016; Narisetty and He, 2019; Johndrow and Bhattacharya, 2020), computationally efficient algorithms with rigorous theoretical guarantees have been few, due to the key challenge that the posterior resulting from the spike-and-slab prior is high-dimensional and may be very far from unimodal.

MCMC-based methods. One line of work on posterior sampling exploits MCMC-based methods (Belloni and Chernozhukov, 2009; Richardson et al., 2010; Schreck et al., 2015; Yang et al., 2016). Motivated by the Bernstein-von Mises theorem, Belloni and Chernozhukov (2009) assume that the posterior is close to a normal density and design a sampler based on this assumption. Closer in spirit to our work, Yang et al. (2016) assumes their measurement matrix $\mathbf{X}$ satisfies a condition similar to RIP, and shows that the posterior on the set of "influential coordinates" (i.e., those larger than the noise level) is unimodal, if the SNR is very high or very low. However, their work does not capture a broad range of intermediate SNRs (see discussion after Eq. (10)), for which the posterior can display multimodal behavior. A similar result was given by Theorem 4.2 of Chen et al. (2024), again under a high SNR assumption. Moreover, Theorem 6 of Castillo et al. (2015) shows that under some mild conditions, the posterior converges to a *mixture of Gaussians*, and such a mixture will collapse into a single Gaussian only if the signal dominates the noise. Such a result uncovers a limitation of current MCMC-based methods: their conclusions are based on the premise that posterior is approximately a unimodal distribution, which reduces the difficulty of mixing time analyses.

Recent work by Bruna and Han (2024) studies the use of diffusion methods (based on stochastic localization) for posterior sampling in linear inverse problems. We think this is an interesting approach that merits further investigation, e.g., for our specific case of spike-and-slab posterior sampling. However, the aforementioned result is not analyzed in discrete time, and explicit bounds on various parameters used in its analysis (such as χ_t in Eq. (14)) are not readily available, so it is presently unclear what end-to-end algorithmic guarantees this framework yields for our setting. Moreover, Jiang (2024) provides mixing time guarantees for various posterior samplers, including stochastic localization methods and Gibbs sampling. However, the analyses in Jiang (2024) require strong assumptions about the SNR and/or side information about the true support (cf. discussion in Section 2.1.1), and thus appears to not address our main Question 4 in full generality.

Approximating the posterior. Another line of work on posterior sampling uses variational Bayesian methods, which obtain the closest approximation of a complex posterior distribution from a more tractable family (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017). The most popular approximating family consists of product distributions, and the corresponding approximation is known as a naïve mean-field approximation. For example, Ray and Szabó (2022) and Mukherjee and Sen (2022) show that the mean-field approximation has strong theoretical properties. However, the mean-field optimization problem has been shown to be highly nonconvex with spurious local optima in some simple learning settings (Mukherjee et al., 2018). Moreover, Ghorbani et al. (2018) show the failure of naïve mean-field for a simple high-dimensional problem. In some sense, our paper also obtains a variational approximation by sampling from an appropriate product distribution. However, instead of choosing the "best" candidate (requiring nonconvex optimization), we choose one guided by fine approximations of the true posterior and the RIP condition.

Moderate dimension. Recently, Montanari and Wu (2024) gave a computationally-efficient posterior sampler in a setting where the posterior can be highly multimodal. Their algorithm is motivated by decompositions in statistical physics, and introduces a latent variable that is easily sampleable

and induces the correct marginal distribution on $\boldsymbol{\theta}$. However, Montanari and Wu (2024) assume that their design matrix $\mathbf{X}$ has a number of linear measurements which grows at least linearly in the feature dimension d. This can be a costly assumption in practical high-dimensional settings.

In our paper, we remove the aforementioned constraints, i.e., we operate in the *high-dimensional regime* of $n = o(d)$ and under a range of SNRs σ^{-1} where the posterior is potentially *multimodal*.

2. Approach

In this section, we overview the pieces of our algorithm and its analysis. A straightforward calculation based on Bayes' theorem (see Lemmas 22 and 25) shows that when $\mu = \mathcal{N}(0, 1)$ in (3), the posterior distribution (2) is a mixture of Gaussians:

$$\begin{aligned} \pi\left(\cdot \mid \mathbf{X}, \mathbf{y}\right) &= \sum_{S \subseteq [d]} w(S)\, \mathcal{N}\left(\mathbf{A}_S^{-1}\mathbf{b}_S, \mathbf{A}_S^{-1}\right), \\ \text{where } w(S) &\propto \left(\prod_{i \in S} \frac{\mathbf{q}_i}{1 - \mathbf{q}_i}\right) \exp\left(\frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}}\right) \frac{1}{\sqrt{\det \mathbf{A}_S}}, \\ \mathbf{A}_S \in \mathbb{R}^{S \times S} &:= \frac{1}{\sigma^2}\left[\mathbf{X}^\top \mathbf{X}\right]_{S \times S} + \mathbf{I}_S,\ \mathbf{b}_S \in \mathbb{R}^S := \frac{1}{\sigma^2}\mathbf{X}_{S:}^\top \mathbf{y}, \end{aligned} \tag{5}$$

and w forms a distribution over $\{S \mid S \subseteq [d]\}$. The key challenge is that, even if one can establish that $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ is mostly supported on $\approx k$-sparse subsets S (where RIP holds and estimation is easy), there are still $\approx d^k$ candidate small subsets to enumerate over. Thus, a natural starting point is to correct an easier-to-sample proposal distribution that approximates (5).

Approximating with a product mixture. Our approach is inspired by first-order frequentist methods for sparse recovery (1), when $\mathbf{X}$ is RIP. This condition means that for bounded-size subsets $S \subseteq [d]$, we have $[\mathbf{X}^\top \mathbf{X}]_{S \times S} \approx \mathbf{I}_S$ is an approximate isometry, so $\mathbf{A}_S$ in (5) is close to a multiple of the identity (see Definition 7 for a formal statement). Many sparse recovery methods simulate well-conditioned gradient descent, a method for optimizing convex functions, on the naturally non-convex sparse linear regression objective by enforcing sparsity of iterates (Blumensath and Davies, 2009; Needell and Tropp, 2009) or projecting onto convex proxy sets for sparsity (Agarwal et al., 2010). Thus, a natural approach to sampling from (2), at least over components $S \subseteq [d]$ with small $|S|$, is to approximate each Gaussian in (5) with an isotropic Gaussian (i.e., a product density).

There is precedent: for the closely-related problem of sampling from the *sparse mean model*, where observations are of the form (1) but $\mathbf{X} = \mathbf{I}_d$, Castillo and Van Der Vaart (2012) gave a simple dynamic programming-based algorithm for drawing a candidate $S \sim w$. The key property enabling the Castillo and Van Der Vaart (2012) algorithm is that both the prior and mixture components are product distributions. Thus, our starting point is to approximate (1) with a sparse mean model.

Denoising the observations. One candidate for simulating the sparse mean model is

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}^\star + \mathbf{X}^\top \boldsymbol{\xi} \approx \boldsymbol{\theta}^\star + \mathbf{X}^\top \boldsymbol{\xi},$$

where the above approximation holds due to RIP, assuming the signal $\boldsymbol{\theta}^\star$ is sparse. At least for the natural setting of small σ, this calculation suggests $\mathbf{X}^\top \mathbf{y}$ (which we can recover from our observations) is a good estimate of $\boldsymbol{\theta}^\star$. Unfortunately, the coordinates of $\mathbf{X}^\top \mathbf{y}$ are highly non-uniform

in magnitude, due to randomness inherent in the diffuse density μ. This causes difficulties when setting up a rejection sampling scheme for sampling from w in (5), because the estimate

$$\exp\left(\frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}}\right) \approx \exp\left(\frac{\sigma^2}{2(1+\sigma^2)}\left\|\mathbf{b}_S\right\|_2^2\right), \text{ where } \mathbf{b} := \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

does not hold, even though $\mathbf{A}_S \approx (\frac{1+\sigma^2}{\sigma^2})\mathbf{I}_S$ is true for small S. This is simply because the scale of $\mathbf{b}_S$ is too large, due to the aforementioned nonuniformity in $\boldsymbol{\theta}^\star \sim \pi$.

We make the crucial observation that we can "denoise" the proportionality constants in (5) as follows. Let $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ be an estimate of our signal $\boldsymbol{\theta}^\star$ with $T := \operatorname{supp}(\widehat{\boldsymbol{\theta}})$, and let $T \subseteq S, S' \subseteq [d]$. Then, a straightforward calculation (Lemma 3) shows that

$$\frac{\exp\left(\frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}}\right)}{\exp\left(\frac{1}{2}\left\|\mathbf{b}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}}\right)} = \frac{\exp\left(\frac{1}{2}\left\|\mathbf{z}_S\right\|^2_{\mathbf{A}_S^{-1}}\right)}{\exp\left(\frac{1}{2}\left\|\mathbf{z}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}}\right)}, \text{ where } \mathbf{z} := \frac{1}{\sigma^2}\left(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\theta}}\right) + \widehat{\boldsymbol{\theta}}. \tag{6}$$

Thus, as long as almost all of the high-weight S in (5) are supersets of some $T \subseteq [d]$, we have freedom in choosing $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^T$ to shrink the scale of our proportionality constants using (6).

Constructing a posterior estimator. The next step in our plan is to construct an estimate $\widehat{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{y})$, for which we can show $\operatorname{supp}(\boldsymbol{\theta}) \supseteq \operatorname{supp}(\widehat{\boldsymbol{\theta}})$ with high probability over $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$.

We propose to use $\widehat{\boldsymbol{\theta}}$ obtained using sparse recovery with ℓ_∞ guarantees. To draw a connection between $\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^\star$ and $\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$, we extend a powerful result of Jalal et al. (2021) (Proposition 4) that characterizes posterior sampling as an estimation tool. Specifically, it shows that posterior sampling achieves estimation guarantees within a constant factor of *any other estimator* in any metric, that is learned from noisy measurements of a signal. Due to the existence of sparse recovery guarantees in ℓ_∞, we show that thresholding the coordinates of a frequentist estimate $\widehat{\boldsymbol{\theta}}$ yields a sufficient T for our framework. This can be intuitively viewed as learning the "obvious" coordinates of $\boldsymbol{\theta}^\star$ above the SNR σ^{-1}, which almost all posterior samples should also contain.

After applying this denoising, we are able to show that our new proportionality constants in (6) are much more closely approximated by a sparse mean model, for small $S \subseteq [d]$. However, we are left to prove that $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ is mainly supported on small $S \subseteq [d]$. While this is an intuitive property (as the prior is sparse), we were previously only aware of proofs in specific cases. For example, Castillo et al. (2015) showed sparsity of the spike-and-slab posterior when $\mu = \mathrm{Lap}(0, 1)$, but their proof technique required the negative log-density to be a norm, and hence did not apply to $\mu = \mathcal{N}(0, 1)$. By a very simple application of Proposition 4 (cf. Corollary 5), we prove such a sparsity concentration result for all densities μ, which is potentially of independent interest.

Our final sampler first learns a frequentist estimate $\widehat{\boldsymbol{\theta}}$ with large coordinates T, and draws $S \subseteq [d]$ from a product distribution restricted to $S \supseteq T$ and $|S| \lesssim k$. It then applies a denoised rejection sampling step to correct this proposal. Our proposal distribution is sampled via *conditional Poisson sampling* (Lemma 10), based on a dynamic programming strategy. Interestingly, the product distribution inducing our proposal is actually not concentrated on small subsets, so it only approximates our posterior (5) over a high-probability region of $\operatorname{supp}(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$.

Intuitively, our algorithm reflects uncertainty in the denoised sparse linear regression problem: we do not know the location of the remaining small signal coordinates, but we know they are few.

Extending to a Laplace prior. We now illustrate how to extend our findings to the Laplace prior setting (Theorem 2). Unlike the Gaussian case, the ℓ_1 norm arising from the Laplace prior precludes a closed-form expression for the density over the posterior support. Consequently, we rely on the triangle inequality to perform approximations, and subsequently establish in Lemma 33 that for noise level $\sigma = O\big(\frac{1}{k}\big)$, it is still possible to construct an approximate density that is both efficient to sample from and closely approximates the true posterior. A further challenge in designing an efficient algorithm is computing the rejection ratio, again due to the lack of a closed-form expression for the posterior density. To overcome this, we use an annealing-based algorithm (Proposition 29) combined with an approximate sampler for composite densities developed in Lee et al. (2021) to estimate normalizing constants in polynomial time, enabling us to bypass the need for an exact closed-form density while still ensuring efficient and accurate sampling from the posterior.

Due to space constraints, we present all results related to our extension to the Laplace prior setting (i.e., for proving Theorem 2) in Appendix C.

3. Overview of proof of Theorem 1

In this section, we provide a more formal overview of our proof of Theorem 1 (i.e., spike-and-slab posterior sampling with a Gaussian prior $\mu = \mathcal{N}(0,1)$); for a more complete version of this section with proofs and additional exposition, see Appendix B.

First, we give a formal definition of our sampling problem.

Model 1 (Spike-and-slab posterior sampling) *Let $\mathbf{q} \in [0,1]^d$, $n, d \in \mathbb{N}$, $k = \|\mathbf{q}\|_1$, and $\sigma > 0$ be known, and let $\mu = \mathcal{N}(0,1)$. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a known measurement matrix, and suppose that we observe $(\mathbf{X}, \mathbf{y})$ where, following the notation (14), $\mathbf{y}$ is generated via*

$$\boldsymbol{\theta}^\star \sim \pi, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 \mathbf{I}_d), \quad \mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\xi}, \tag{7}$$

and $\boldsymbol{\theta}^\star$ and $\boldsymbol{\xi}$ are independent. Our goal is to sample from the posterior $\pi(\cdot \mid \mathbf{X}, y)$.

Our first observation is the following fact about recentering proportionality constants (i.e., mixture weights) in (5), which helps us rewrite said weights in a way that is more closely approximated by a mixture of product distributions. We defer a proof to Appendix B.

Lemma 3 *In the setting of Model 1, let*

$$\mathbf{A} := \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{I}_d, \ \mathbf{b} := \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y},$$

and denote $\mathbf{A}_S := \mathbf{A}_{S\times S}$ for brevity. For some $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ with $T := \mathrm{supp}(\widehat{\boldsymbol{\theta}})$, let $\mathbf{z} := \mathbf{b} - \mathbf{A}\widehat{\boldsymbol{\theta}}$. Then,

$$\frac{\exp\left(\frac{1}{2}\|\mathbf{b}_S\|^2_{\mathbf{A}_S^{-1}}\right)}{\exp\left(\frac{1}{2}\|\mathbf{b}_{S'}\|^2_{\mathbf{A}_{S'}^{-1}}\right)} = \frac{\exp\left(\frac{1}{2}\|\mathbf{z}_S\|^2_{\mathbf{A}_S^{-1}}\right)}{\exp\left(\frac{1}{2}\|\mathbf{z}_{S'}\|^2_{\mathbf{A}_{S'}^{-1}}\right)}, \text{ for any } T \subseteq S, S' \subseteq [d].$$

Lemma 3 makes our goal clear; to use it, we must identify an estimate $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ such that $T := \mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq \mathrm{supp}(\boldsymbol{\theta}) =: S$, with high probability over $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$. This is because Lemma 3 only applies to mixture components that are supersets of the same T. To do so, we introduce the

following analysis tool, which also lets us prove a key structural property of $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$: that it is mostly supported on mixture components corresponding to small S.

Our main analysis tool shows that posterior sampling performs nearly as well at estimation as *any other procedure*, in a precise sense. To our knowledge, this fact was first stated in Jalal et al. (2021). We use this in several places to prove properties of our posterior. Here, we state a slight generalization of the Jalal et al. (2021) result, with a proof deferred to Appendix E.

Proposition 4 (Generalization of Theorem 3.4, Jalal et al. (2021)) *Let $\mathrm{m}(\cdot, \cdot)$ be an arbitrary metric over $\mathcal{K} \times \mathcal{K}$ and suppose μ is a distribution over $\mathcal{K}$. Let $\theta^\star \sim \mu$, let $\mathcal{F} : \mathcal{K} \to \Omega$ be an arbitrary (possibly randomized) forward operator, and let $\varphi = \mathcal{F}(\theta^\star)$. Suppose there is any (possibly randomized) algorithm $\mathcal{A} : \Omega \to \mathcal{K}$ such that for $\epsilon > 0$, $\delta \in (0, 1)$,*

$$\mathbb{P}\left[\mathrm{m}(\widehat{\theta}, \theta^\star) > \epsilon\right] \le \delta, \text{ for } \widehat{\theta} \sim \mathcal{A}(\varphi), \tag{8}$$

where probabilities are taken over the joint distribution of $(\theta^\star, \varphi, \widehat{\theta})$. Then, letting $\theta \sim \mu(\cdot \mid \varphi)$, i.e., the posterior distribution of $\theta^\star$ given observations φ, we have $\mathbb{P}[\mathrm{m}(\theta, \theta^\star) > 2\epsilon] \le 2\delta$.

To demonstrate the power of Proposition 4, we first prove a support size concentration result for spike-and-slab posteriors. A similar result was shown in Theorem 1, Castillo et al. (2015), without a quantitative convergence rate. Moreover, the strategy in Castillo et al. (2015) is somewhat more ad hoc, as it was specialized to Laplace priors (for example, it relied on the negative log-density obeying the triangle inequality, which does not hold for e.g., the Gaussian prior). In contrast, we prove a non-asymptotic result for arbitrary product priors, using an arguably simpler proof.

Corollary 5 (Sparsity of posterior) *In the setting of Model 1, let $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$. Then*

$$\mathbb{P}\left[|\mathrm{supp}(\boldsymbol{\theta})| \le 6\left(k + \log\left(\frac{3}{\delta}\right)\right)\right] \ge 1 - \delta, \text{ for all } \delta \in (0, 1), \tag{9}$$

Moreover, (9) holds for any choice of μ in Model 1.

Proof We apply Proposition 4 to the distribution $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$ defined in (15), (17). Let $\mathcal{K}$ be the set of subsets of $[d]$, and let $\varphi = (\mathbf{X}, \mathbf{y})$ (so $\Omega = \mathbb{R}^{n \times d} \times \mathbb{R}^n$). Moreover, let $\mathrm{m}(\cdot, \cdot)$ be the Hamming distance over $\mathcal{K} \times \mathcal{K}$, i.e., $\mathrm{m}(S_1, S_2) = |\{i \in [d] \mid (i \in S_1 \wedge i \notin S_2) \vee (i \in S_2 \wedge i \notin S_1)\}|$. It is well-known that $\mathrm{m}(\cdot, \cdot)$ is a metric. We let $\mathcal{A}(\varphi)$ always return the empty set $\emptyset$. By a Chernoff bound, letting $S^\star \sim \pi_{\mathrm{supp}}$, since $|S^\star|$ is the sum of variables $\sim \mathrm{Bern}(\mathbf{q}_i)$ for all $i \in [d]$,

$$\mathbb{P}\left[|S^\star| \ge 2\left(k + \log\frac{3}{\delta}\right)\right] = \mathbb{P}\left[\mathrm{m}(S^\star, \mathcal{A}(\varphi)) \ge 2\left(k + \log\frac{3}{\delta}\right)\right] \le \frac{\delta}{3}.$$

Thus, (8) holds with $\epsilon = 2(k + \log(\frac{3}{\delta}))$ so that, for $S \sim \pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$, $\mathbb{P}[\mathrm{m}(S, S^\star) \ge 4(k + \log\frac{3}{\delta})] \le \frac{2\delta}{3}$. Finally, by the triangle inequality on $\mathrm{m}(\cdot, \cdot)$, $|S| = \mathrm{m}(S, \mathcal{A}(\varphi)) \le \mathrm{m}(S, S^\star) + \mathrm{m}(S^\star, \mathcal{A}(\varphi)) = \mathrm{m}(S, S^\star) + |S^\star|$. Hence, we have

$$\begin{aligned}\mathbb{P}\left[|S| \le 6\left(k + \log\frac{3}{\delta}\right)\right] &\ge \mathbb{P}\left[|S^\star| + m(S, S^\star) \le 6\left(k + \log\frac{3}{\delta}\right)\right] \\ &\ge \mathbb{P}\left[|S^\star| \le 2\left(k + \log\frac{3}{\delta}\right) \wedge m(S, S^\star) \le 4\left(k + \log\frac{3}{\delta}\right)\right].\end{aligned}$$

The conclusion follows from a union bound over the above two events. ■

Proposition 4's conclusion holds over the randomness of both $(\theta^\star, \varphi)$ (collectively, defining the "model" given in the posterior sampling problem) and the sample $\theta \sim \mu(\cdot \mid \varphi)$. This guarantee is not quite compatible with our applications, as we would only like to reason about the failure probability of our sampler for a given posterior distribution, holding the model fixed. We give a simple reduction showing that an event which holds with very high probability over the randomness of $(\theta^\star, \varphi, \theta)$ also holds with high probability over $\theta \sim \mu(\cdot \mid \varphi)$, for "most" models $(\theta^\star, \varphi)$.

Lemma 6 *Let (α, β) be random variables, and let $\mathcal{E}_{\alpha,\beta}$ be an event that depends on their realizations. Suppose for some $\delta_1, \delta_2 \in (0,1)$ that $\mathbb{P}_{(\alpha,\beta)}[\mathcal{E}_{\alpha,\beta}] \geq 1 - \delta_1\delta_2$. Then, letting $g(\alpha) := \mathbb{P}_\beta[\mathcal{E}_{\alpha,\beta} \mid \alpha]$, $\mathbb{P}_\alpha[g(\alpha) \leq 1 - \delta_1] \leq \delta_2$.*

Our final theorem statements use Lemma 6 by applying it with $\alpha \leftarrow (\boldsymbol{\theta}^\star, \boldsymbol{\xi})$ (i.e., the randomness used to generate observations $\mathbf{y}$ in our model (1)), and $\beta \leftarrow \boldsymbol{\theta} \sim \mu(\cdot \mid \mathbf{X}, \mathbf{y})$. We refer to this usage by qualifying that statements hold "with probability $\geq 1 - \delta_1$ over the randomness of our model."

We next address the issue of coming up with an estimate $\widehat{\boldsymbol{\theta}}$ for use in Lemma 3. Our estimate $\widehat{\boldsymbol{\theta}}$ is learned via sparse recovery algorithms applied to the observations $(\mathbf{X}, \mathbf{y})$. We need the following standard definition from the sparse recovery literature to state our required estimation guarantees.

Definition 7 (Restricted isometry property) *We say $\mathbf{X} \in \mathbb{R}^{n\times d}$ satisfies the (ϵ, s)-restricted isometry property, or $\mathbf{X}$ is (ϵ, s)-*RIP*, if for all $\boldsymbol{\theta} \in \mathbb{R}^d$ with $\mathrm{nnz}(\boldsymbol{\theta}) \leq s$,*

$$(1-\epsilon)\left\|\boldsymbol{\theta}\right\|_2^2 \leq \left\|\mathbf{X}\boldsymbol{\theta}\right\|_2^2 \leq (1+\epsilon)\left\|\boldsymbol{\theta}\right\|_2^2.$$

An equivalent condition is that $\boldsymbol{\lambda}([\mathbf{X}^\top\mathbf{X}]_{S\times S}) \in [1-\epsilon, 1+\epsilon]^s$ for all $S \subseteq [d]$ with $|S| \leq s$.

Intuitively, RIP implies $\mathbf{X}$ acts as an approximate isometry when restricted to sparse vectors, so $[\mathbf{X}^\top\mathbf{X}]_{S\times S}$ is well-conditioned for any small S. Many random matrix ensembles, e.g., entrywise i.i.d. sub-Gaussian matrices, satisfy Definition 7 for $n \gtrsim \frac{s}{\epsilon^2}$; we give an extended discussion and more examples in Appendix A.3. We now state our main assumption on sparse recovery algorithms.

Model 2 (Sparse recovery) *Let $n, d \in \mathbb{N}$, $k \in [1, d]$, and $\sigma > 0$ be known. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a known measurement matrix, and let $\boldsymbol{\theta}^\star \in \mathbb{R}^d$, $\boldsymbol{\xi} \in \mathbb{R}^n$ satisfy $\mathrm{nnz}(\boldsymbol{\theta}^\star) \leq k$. Suppose we observe $(\mathbf{X}, \mathbf{y})$ where $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$. Our goal is to output an estimate of $\boldsymbol{\theta}^\star$.*

Assumption 1 *Let $\mathcal{A}(\mathbf{X}, \mathbf{y})$ be an algorithm with the following guarantee, parameterized by a universal constant $C_\mathcal{A}$ and $R_\mathcal{A} : \mathbb{N} \times (0,1) \times (0,1) \to \mathbb{R}_{\geq 0}$. In the setting of Model 2, suppose $\mathbf{X}$ is $(\epsilon, C_\mathcal{A}k)$-RIP for $\epsilon \in (0,1)$. Then $\mathcal{A}(\mathbf{X}, \mathbf{y})$ returns $\boldsymbol{\theta}'$ satisfying*

$$\left\|\boldsymbol{\theta}' - \boldsymbol{\theta}^\star\right\|_\infty \leq \sigma \cdot R_\mathcal{A}(k, \epsilon, \delta), \text{ with probability } \geq 1 - \delta,$$

for any $\delta \in (0,1)$, over the randomness of $\boldsymbol{\xi}$.[2]

In Appendix A.3 (specifically, Proposition 15), we state two sparse recovery algorithms $\mathcal{A}$ satisfying the premise of Assumption 1 with different tradeoffs. The first algorithm achieves $R_\mathcal{A}(k, \epsilon, \delta) \approx \sqrt{\log(d/\delta)}$, which is near-constant in problem parameters, and requires polynomial time. The

2. We state our result for deterministic algorithms $\mathcal{A}$ for simplicity, as Proposition 15 is deterministic.

second algorithm has the advantage of running in nearly-linear time, but at the cost of a weaker recovery guarantee of $R_{\mathcal{A}}(k,\epsilon,\delta) \approx \sqrt{k}\,\mathrm{polylog}(\frac{d}{\delta})$. We now show how to use a sparse recovery algorithm meeting Assumption 1, in conjunction with Proposition 4, to produce an estimator of the large coordinates of a sample from the posterior distribution $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ with high probability.

Lemma 8 *Let Assumption 1 hold. In the setting of Model 1, let μ be arbitrary, let $\epsilon, \delta \in (0,1)$, and suppose $\mathbf{X}$ is $(C_{\mathcal{A}}k^\star, \epsilon)$-RIP for some $k^\star \in \mathbb{N}$ with $k^\star \geq 2(k + \log(\frac{5}{\delta}))$. Then there is an estimation procedure $\mathcal{A}_{\text{est}}$ that takes as input $(\mathbf{X}, \mathbf{y})$, and produces $\widehat{\boldsymbol{\theta}}$ satisfying*

$$\mathbb{P}_{\boldsymbol{\theta}\sim\pi(\cdot|\mathbf{X},\mathbf{y})}\left[\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_\infty \leq 6\sigma \cdot R_{\mathcal{A}}\left(k^\star, \epsilon, \frac{\delta^2}{5}\right)\right) \wedge \left(\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq \mathrm{supp}(\boldsymbol{\theta})\right)\right] \geq 1 - \delta. \quad (10)$$

Proof By a Chernoff bound, with probability $\geq 1 - \frac{\delta^2}{5}$ over $\boldsymbol{\theta}^\star \sim \pi$ in Model 1, we have $|\mathrm{supp}(\boldsymbol{\theta}^\star)| \leq 4(k + \log(\frac{5}{\delta}))$. Under this condition, Assumption 1 states that $\mathcal{A}$ outputs $\boldsymbol{\theta}'$ with

$$\left\|\boldsymbol{\theta}' - \boldsymbol{\theta}^\star\right\|_\infty \leq \sigma \cdot R_{\mathcal{A}}\left(k^\star, \epsilon, \frac{\delta^2}{5}\right) \quad (11)$$

with probability $\geq 1 - \frac{\delta^2}{5}$ over the randomness of $\boldsymbol{\xi}$. Thus, after taking a union bound, applying Proposition 4 shows that with probability $\geq 1 - \frac{4\delta^2}{5}$ over the randomness of $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$ and Model 1, our estimate $\boldsymbol{\theta}'$ satisfies

$$\left\|\boldsymbol{\theta}' - \boldsymbol{\theta}\right\|_\infty \leq \left\|\boldsymbol{\theta}' - \boldsymbol{\theta}^\star\right\|_\infty + \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right\|_\infty \leq 3\sigma \cdot R_{\mathcal{A}}\left(k^\star, \epsilon, \frac{\delta^2}{5}\right). \quad (12)$$

Therefore, applying Lemma 6 with $\delta_1 \leftarrow \delta$ and $\delta_2 \leftarrow \frac{4\delta}{5}$, we have with probability $\geq 1 - \frac{4\delta}{5}$ over the randomness of Model 1 that (12) holds with probability $\geq 1 - \delta$ for $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$.

Define the operation $\mathrm{clip}(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ by $[\mathrm{clip}(\mathbf{x}, \alpha)]_i = \mathbf{x}_i$ if $|\mathbf{x}_i| > \alpha$ and 0 else, for all $i \in [d]$. Our estimation procedure $\mathcal{A}_{\text{est}}$ returns $\widehat{\boldsymbol{\theta}} \leftarrow \mathrm{clip}(\boldsymbol{\theta}', 3\sigma \cdot R_{\mathcal{A}}(k^\star, \epsilon, \frac{\delta^2}{5}))$. It is clear that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\|_\infty \leq 3\sigma \cdot R_{\mathcal{A}}(k^\star, \epsilon, \frac{\delta^2}{5})$, so applying the triangle inequality with (12) gives the first bound in (10). Moreover, to see that $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq \mathrm{supp}(\boldsymbol{\theta})$, suppose that some $i \in [d]$ has $i \notin \mathrm{supp}(\boldsymbol{\theta})$ but $i \in \mathrm{supp}(\widehat{\boldsymbol{\theta}})$. The latter condition and the definition of clip implies $|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i| = |\boldsymbol{\theta}'_i| > 3\sigma \cdot R_{\mathcal{A}}(k^\star, \epsilon, \frac{\delta^2}{5})$, which contradicts (12). The same logic with (11) shows $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq \mathrm{supp}(\boldsymbol{\theta}^\star)$. ■

The final tools we require are the following well-known results from the sampling literature: *rejection sampling* and *conditional Poisson sampling*. We defer proofs of these statements to Appendix E.

Lemma 9 *Let π, μ be distributions over the same domain Ω, and suppose $\pi \propto P$ and $\mu \propto Q$ for unnormalized densities P, Q. Moreover, suppose for all $\omega \in \Omega$, $\frac{P(\omega)}{Q(\omega)} \leq C$, $\frac{Q(\omega)}{P(\omega)} \leq C$. There is an algorithm $\mathsf{RejectionSample}(\mu, P, Q, \Omega, C, \delta)$ that, for any $\delta \in (0,1)$, outputs a sample within total variation δ from π, using $O(C^2 \log(\frac{1}{\delta}))$ samples from μ, and evaluating $\frac{P(\omega)}{Q(\omega)}$ $O(C^2 \log(\frac{1}{\delta}))$ times.*

Lemma 10 *Let $\mathbf{p} \in [0,1]^d$, and let $\pi := \bigotimes_{i\in[d]} \mathrm{Bern}(\mathbf{p}_i)$ be a product distribution over $\{0,1\}^d$, identified with sets $S \subseteq [d]$. Define $\Omega_k := \{S \subseteq [d] \mid |S| \leq k\}$. $\mathsf{ConditionalPoisson}$ (Algorithm 5) runs in time $O(dk)$ and returns $S \in \Omega_k$ such that $\mathbb{P}[S = T] = \pi(T \mid T \in \Omega_k)$.*

We now present the details of our posterior sampler, an instance of rejection sampling (Lemma 9). In Algorithm 1, we first define the proposal distribution $\tilde{\pi}_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$ that our rejection sampler is based on. This proposal is a product distribution, restricted to small subsets that we show approximates the true posterior density $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ over a high-probability region according to $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$. It takes an additional input $\widehat{\boldsymbol{\theta}}$ that is used to center its samples, as discussed following Lemma 3. Finally, it produces its output in Line 1 by calling our conditional Poisson sampler (Algorithm 5, cf. Lemma 10).

Algorithm 1 $\mathsf{ProductSampleGaussian}(\mathbf{X}, \mathbf{y}, \widehat{\boldsymbol{\theta}}, \sigma, \mathbf{q}, k^\star)$

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\sigma > 0$, $\mathbf{q} \in [0,1]^d$ produced by Model 1, $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$, $k^\star \in \mathbb{N}$.
Output: Sample S from a conditional Poisson distribution over

$$S \in \Omega_{\widehat{\boldsymbol{\theta}}, k^\star} := \left\{ S \subseteq [d] \mid S \supseteq \mathrm{supp}(\widehat{\boldsymbol{\theta}}), |S| \leq k^\star \right\} \tag{13}$$

$\mathbf{z} \leftarrow \frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}) - \widehat{\boldsymbol{\theta}}$
$T \leftarrow \mathrm{supp}(\widehat{\boldsymbol{\theta}})$
for $i \in T^c$ **do**
  $\mathbf{r}_i \leftarrow \mathbf{q}_i \sqrt{\frac{\sigma^2}{1+\sigma^2}} \exp\left(\frac{\sigma^2}{2(1+\sigma^2)} \mathbf{z}_i^2 \right)$
  $\mathbf{p}_i \leftarrow \frac{\mathbf{r}_i}{1 - \mathbf{q}_i + \mathbf{r}_i}$
end
return $T \cup \mathsf{ConditionalPoisson}(\mathbf{p}, k^\star - |T|)$

Let $k^\star = O(k \,\mathrm{polylog}(\frac{d}{\delta}))$ be a parameter such that Corollary 5 guarantees $|\mathrm{supp}(\boldsymbol{\theta})| \leq k^\star$ with high probability, for $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$. We next demonstrate that when conditioned on a support set $S \subseteq [d]$ with certain desirable properties — namely, that $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$ and $|S| \leq k^\star$ — the true (unnormalized) posterior density P and the approximate (unnormalized) density Q from Algorithm 1, defined in (22) and (21) respectively, are pointwise close. This proximity renders them well-suited for efficient rejection sampling (Lemma 9). Although we are only able to show that P and Q are close over this subset of the sample space, Lemma 8 implies that the true posterior distribution concentrates over this region of the sample space. Therefore, this conditional sampling is close in total variation to the true posterior.

Lemma 11 *Suppose that* $\mathbf{X}$ *is* $(\epsilon, k^\star)$*-RIP for* $\epsilon \in (0, \frac{1}{2})$. *Then for any* $S \subseteq [d]$ *such that* $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$ *and* $|S| \leq k^\star$, *we have* $\pi_{\mathrm{supp}}(S \mid \mathbf{y}, \mathbf{X}) \propto P(S)$, $\tilde{\pi}_{\mathrm{supp}}(S \mid \mathbf{y}, \mathbf{X}) \propto Q(S)$ *and*

$$\left(\frac{1}{1+\epsilon}\right)^{\frac{k^\star}{2}} \exp\left(-\frac{\sigma^2 \epsilon}{(1+\sigma^2)^2} \|\mathbf{z}_S\|_2^2 \right) \leq \frac{P(S)}{Q(S)} \leq \left(\frac{1}{1-\epsilon}\right)^{\frac{k^\star}{2}} \exp\left(\frac{\sigma^2 \epsilon}{(1+\sigma^2)^2} \|\mathbf{z}_S\|_2^2 \right),$$

where $P(S), Q(S)$, *and* $\mathbf{z}_S$ *are defined in* (22), (21), *and* (23).

We remark that in Lemmas 19 and 20, we use Proposition 4 once again to provide a high-probability bound on $\|\mathbf{z}_S\|_2$ (as required above) using our sparse recovery guarantees from Assumption 1.

We put all of these pieces together in Algorithm 2, to provide our high-accuracy posterior sampling algorithm over the support set, $S = \mathrm{supp}(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{x}, \mathbf{y})$. We use Algorithm 1 to generate

samples from the approximate density $\mu \propto Q$, conditioned on $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$ and $|S| \leq k^\star$. Given sample access to μ, we now use rejection sampling to obtain a sample from the true posterior, P. Finally, note that given a sample from the posterior over the support, $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$, the distribution of $\boldsymbol{\theta}$, is given as $\pi(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, S) = \mathcal{N}(\boldsymbol{\mu}_S, \mathbf{A}_S^{-1})$, for parameters $\boldsymbol{\mu}_S, \mathbf{A}_S$ stated in Lemma 25. Therefore, given the support, S, it is easy to efficiently sample the parameter, $\boldsymbol{\theta}$.

Algorithm 2 PosteriorSampleGaussian$(\mathbf{X}, \mathbf{y}, \widehat{\boldsymbol{\theta}}, \sigma, \mathbf{q}, k^\star, \delta)$

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\sigma > 0$, $\mathbf{q} \in [0,1]^d$ produced by Model 1, $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$, $k^\star \in \mathbb{N}$, $\delta \in (0,1)$
Output: Sample S approximately distributed as $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$ (cf. Proposition 24)
$\mu \leftarrow$ ProductSampleGaussian$(\mathbf{X}, \mathbf{y}, \widehat{\boldsymbol{\theta}}, \sigma, \mathbf{q}, \frac{3}{4}k^\star)$
$P, Q \leftarrow$ unnormalized distributions in (22), (21)
$\Omega \leftarrow \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star}$ defined in (13)
return RejectionSample$(\mu, P, Q, \Omega, 2, \frac{\delta}{2})$

Given Algorithm 2 along with the conditional distribution form $\pi(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, S) = \mathcal{N}(\boldsymbol{\mu}_S, \mathbf{A}_S^{-1})$, the proof of Theorem 1 follows from the guarantees of the particular sparse recovery procedure used in Assumption 1 (via Lemma 8) to obtain $\hat{\boldsymbol{\theta}}$. Given a $\hat{\boldsymbol{\theta}}$ satisfying $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$ with high probability, for S being sampled from the posterior restricted to sets with $|S| = O(k^\star)$, Lemma 11 shows that rejection sampling is efficiently possible given sample access to an approximate posterior Q and an estimator $\widehat{\boldsymbol{\theta}}$ causing the recentered vector $\mathbf{z}$ in Lemma 3 to be small, and Algorithm 1 provides an algorithm to efficiently sample from such a Q, concluding the proof.

4. Open problems and future work

Our work opens several promising directions for further investigation; we outline a few here.

Sample complexity at the information-theoretic limit. One natural question is whether, in the setting of Model 1, there exists a posterior sampler whose measurement counts can be reduced to the regime where $n = \widetilde{\Omega}(k)$, which matches the information-theoretic lower bound on n for the corresponding estimation task (i.e., Model 2) to be tractable. While our current analysis requires $n \gtrsim k^3$, a significant barrier arises from the determinant-based bound in Lemma 11. Specifically, the determinant ratio scales as $(1+\epsilon)^{k^\star}$, where $k^\star \approx k$ is a high-probability bound on the true signal's support size, and ϵ is the RIP parameter. This ratio becomes superconstant unless $\epsilon \ll \frac{1}{k}$. Addressing this would likely require fundamentally new tools to control estimation error without relying on such multiplicative bounds. Developing such techniques — perhaps via finer-grained probabilistic arguments or more robust surrogate objectives — remains an exciting open problem.

Randomized parameters in spike-and-slab prior. Our current model assumes a fixed and known prior over components. A more realistic and general setting is one where the mixing weights $\mathbf{q} \in [0,1]^d$ (cf. Model 1) are themselves drawn from a prior distribution, as is commonly considered in the sparse linear regression literature, see Section 2 of Montanari and Wu (2024) and references therein. Extending our method to accommodate such uncertainty is both practically relevant and theoretically nontrivial. Two potential approaches include: (i) marginalizing over $\mathbf{q}$ using Gibbs sampling as in Section 2.2 of Montanari and Wu (2024) to marginalize over the mixing probabilities,

or (ii) incorporating the prior on $\mathbf{q}$ directly into the proposal distribution. Whether our theoretical guarantees extend to these settings remains open and an important direction for future work.

Generalization to other priors. Our paper gives an absolute affirmative answer to Question (4) if $\mu = \mathcal{N}(0, 1)$ and a partial affirmative answer if $\mu = \mathrm{Lap}(0, 1)$ (i.e., for a limited range of σ). We think it is an interesting open direction to characterize the types of μ and the ranges of signal-to-noise ratios for which (4) can be answered affirmatively. In particular, a major analytical challenge in non-Gaussian settings is that we do not have an explicit formula for the proportionality constant of each posterior component, leading to potentially lossy bounds via approximation.

Acknowledgments

We would like to thank Eric Price for the pointer to Lemma 47, Arun Jambulapati for suggesting the approach in Appendix D.2, and Trung Dang for the pointer to Lemma 10. KT would like to thank Sameer Deshpande for helpful conversations on the spike-and-slab sampling literature. PS and SK gratefully acknowledge NSF grants 2217069, 2019844, and DMS 2109155.

References

Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010*, pages 37–45. Curran Associates, Inc., 2010.

Josh Alman, Ran Duan, Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, and Renfei Zhou. More asymmetry yields faster matrix multiplication. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2025*, pages 2005–2039. SIAM, 2025.

Ray Bai, Veronika Rockova, and Edward I. George. Spike-and-slab meets lasso: A review of the spike-and-slab lasso. *Handbook of Bayesian Variable Selection*, pages 81–108, 2021.

Alexandre Belloni and Victor Chernozhukov. On the computational complexity of mcmc-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055, 2009.

Dimitris Bertsimas and Bart Van Parys. Sparse high-dimensional regression. *The Annals of Statistics*, 48(1):300–323, 2020.

A. Bhattacharya, A. Chakraborty, and B. K. Mallick. Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991, 2016.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Nicolas Brosse, Alain Durmus, and Éric Moulines. Normalizing constants of log-concave densities. *Electronic Journal of Statistics*, 12:851–889, 2018.

Joan Bruna and Jiequn Han. Provable posterior sampling with denoising oracles via tilted transport. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.

Emmanuel J. Candès and Justin K. Romberg. Signal recovery from random projections. In *Computational Imaging III, 2005*, volume 5674 of *SPIE Proceedings*, pages 76–86. SPIE, 2005.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009*, volume 5 of *JMLR Proceedings*, pages 73–80. JMLR.org, 2009.

Ismael Castillo and Aad Van Der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 2012.

Ismael Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, pages 1986–2018, 2015.

Zizhong Chen and Jack J. Dongarra. Condition numbers of gaussian random matrices. *SIAM Journal on Matrix Analysis and Applications*, 27(3):603–620, 2005.

Zongchen Chen, Conor Sheehan, and Ilias Zadik. On the low-temperature MCMC threshold: the cases of sparse tensor pca, sparse regression, and a geometric rule. *CoRR*, abs/2408.00746, 2024.

H. Chipman. Bayesian variable selection with related predictors. *The Canadian Journal of Statistics*, 24:17–36, 1996.

Ben Cousins and Santosh S. Vempala. Gaussian cooling and $o^*(n^3)$ algorithms for volume and gaussian volume. *SIAM J. Comput.*, 47(3):1237–1273, 2018.

David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(1):407–499, 2004.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013. ISBN 9780817649487 0817649484. doi: 10.1007/978-0-8176-4948-7.

Rong Ge, Holden Lee, and Jianfeng Lu. Estimating normalizing constants for log-concave distributions: algorithms and lower bounds. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020*, pages 579–586. ACM, 2020.

Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

J. Geweke. Variable selection and model comparison in regression. *Bayesian Statistics*, 5:609–620, 1996.

B. Ghorbani, Hamid Haj Seyyed Javadi, and Andrea Montanari. An instability in variational inference for topic models. In *International Conference on Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:88522817.

Ishay Haviv and Oded Regev. The restricted isometry property of subsampled fourier matrices. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, pages 163–179. Springer, 2017.

R.R̃. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

Hemant Ishwaran and J. Sunil Rao. Consistency of spike and slab regression. *Statistics & Probability Letters*, 81(12):1920–1928, 2011.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G. Dimakis, and Jonathan I. Tamir. Robust compressed sensing MRI with deep generative priors. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 14938–14954, 2021.

Qijia Jiang. From estimation to sampling for bayesian linear regression with spike-and-slab prior. *CoRR*, abs/2405.13731, 2024.

P. Johndrow, J.and Orenstein and A. Bhattacharya. Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61, 2020.

Iain M. Johnstone and Bernard W. Silverman. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32(4):1594–1649, 2004.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Jonathan A. Kelner, Jerry Li, Allen Liu, Aaron Sidford, and Kevin Tian. Semi-random sparse recovery in nearly-linear time. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 2352–2398. PMLR, 2023.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 2993–3050. PMLR, 2021.

László Lovász and Santosh S. Vempala. Simulated annealing in convex bodies and an $o^*(n^4)$ volume algorithm. *J. Comput. Syst. Sci.*, 72(2):392–417, 2006.

Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Andrea Montanari and Yuchen Wu. Provably efficient posterior sampling for sparse linear regression via measure decomposition, 2024. URL https://arxiv.org/abs/2406.19550.

Soumendu Sunder Mukherjee, Purnamrita Sarkar, Y. X. Rachel Wang, and Bowei Yan. Mean field for the stochastic blockmodel: optimization landscape and convergence issues. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10717–10727, Red Hook, NY, USA, 2018. Curran Associates Inc.

Sumit Mukherjee and Subhabrata Sen. Variational inference in high-dimensional linear regression. *Journal of Machine Learning Research*, 23:1–56, 2022.

J. Narisetty, N. N.and Shen and X. He. Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*, 114(527):1205–1217, 2019.

Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321, 2009.

Yurii Nesterov and Arkadi Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

Victor Y. Pan and Zhao Q. Chen. The complexity of the matrix eigenproblem. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, pages 507–516. ACM, 1999.

Nicholas G. Polson and Lei Sun. Bayesian ℓ_0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, 2019.

Kolyan Ray and Botond Szabó. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281, 2022.

Sylvia Richardson, Leonardo Bottolo, and Jeffrey S. Rosenthal. Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics*, 9:539–569, 2010.

Veronika Rockova. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Annals of Statistics*, 46(1):401–437, 2018.

Veronika Ročková and Edward I. George. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

Amandine Schreck, Gersende Fort, Sylvain Le Corff, and Eric Moulines. A shrinkage-thresholding metropolis adjusted langevin algorithm for bayesian variable selection. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):366–375, 2015.

Maurice Sion. On general minimax theorems. *Pacific journal of mathematics*, 8(1):171–176, 1958.

Volker Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 13:354–356, 1969.

Kevin Tian. Cs 395t: Continuous algorithms, part ix (sparse recovery). https://kjtian.github.io/notes/CS

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

Ryan J Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1 constrained quadratic programming lasso. *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Yun Yang, Martin J. Wainwright, and Michael I. Jordan. On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics*, pages 2497–2532, 2016.

Appendix A. Preliminaries

A.1. Notation

General notation. We denote matrices in capital boldface and vectors in lowercase boldface. We define $[n] := \{i \in \mathbb{N} \mid i \leq n\}$. When S is a subset of a larger set clear from context (e.g., $[d]$), we let S^c denote its complement. We let $\mathbf{0}_d$ and $\mathbf{1}_d$ denote the all-zeroes and all-ones vectors in $\mathbb{R}^d$.

Matrices. We let $\mathbf{I}_d$ denote the $d \times d$ identity matrix, and $\mathbf{I}_S$ is the identity on $\mathbb{R}^S$ for an index set S. We use $\preceq$ to denote the Loewner partial order on the cone of positive semidefinite (PSD) matrices (denoted $\mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ when in d dimensions), a subset of the symmetric matrices (analogously, denoted $\mathbb{S}^{d\times d}$). When $\mathbf{M} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ is non-singular, we define its induced norm by $\|\mathbf{v}\|^2_{\mathbf{M}} := \mathbf{v}^\top \mathbf{M}\mathbf{v}$.

For $p \geq 1$ (including $p = \infty$), applied to a vector argument, $\|\cdot\|_p$ denotes the ℓ_p norm. For $p, q \geq 1$ and a matrix $\mathbf{M} \in \mathbb{R}^{n\times d}$, we also use the notation $\|\mathbf{M}\|_{p\to q} := \max_{\mathbf{v}\in\mathbb{R}^d \mid \|\mathbf{v}\|_p \leq 1} \|\mathbf{M}\mathbf{v}\|_q$.

We use $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{op}}$ to denote the Frobenius and $(2 \to 2)$ operator norms of a matrix argument, and for $\mathbf{M} \in \mathbb{S}^{d\times d}$, $\boldsymbol{\lambda}(\mathbf{M}) \in \mathbb{R}^d$ denotes its vector of eigenvalues sorted in nondecreasing order, i.e., $\boldsymbol{\lambda}_1(\mathbf{M}) \geq \boldsymbol{\lambda}_2(\mathbf{M}) \geq \ldots \geq \boldsymbol{\lambda}_d(\mathbf{M})$. We also use $\boldsymbol{\lambda}_{\max}(\mathbf{M})$ and $\boldsymbol{\lambda}_{\min}(\mathbf{M})$ to denote the maximal and minimal eigenvalues of $\mathbf{M} \in \mathbb{S}^{d\times d}$. For an asymmetric rank-r $\mathbf{M}$, we let $\boldsymbol{\sigma}(\mathbf{M}) \in \mathbb{R}^r_{>0}$ denote its singular values, and define $\boldsymbol{\sigma}_{\min}$, $\boldsymbol{\sigma}_{\max}$ similarly. For $\mathbf{M} \in \mathbb{S}^{d\times d}$, we let $\mathbf{M}^\dagger$ denote its pseudoinverse, i.e., the matrix in $\mathbb{S}^{d\times d}$ satisfying $\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger\mathbf{M}$ is the projection matrix onto the span of $\mathbf{M}$.

We let $\omega < 2.373$ (Alman et al., 2025) denote the matrix multiplication exponent, and assume multiplying, inverting (Strassen, 1969), and computing eigendecompositions (Pan and Chen, 1999) of $d \times d$ matrices takes $O(d^\omega)$ time.

Probability. For $p \in (0, 1)$, we let $\mathrm{Bern}(p)$ denote the Bernoulli distribution (over $\{0, 1\}$) with mean p. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the multivariate normal distribution with specified mean and covariance. For $t \in \mathbb{R}$ we let δ_t denote the Dirac (generalized) density at t, i.e., $x \sim \delta_t \implies x = t$ with probability 1. We let $\mathrm{Lap}(\mu, b)$ be the Laplace distribution with mean μ and scale b, so $\mathrm{Lap}(0, 1)$ is the density on $\mathbb{R}$ that is $\propto \exp(-|\cdot|)$. For an event $\mathcal{E}$, $\mathbb{I}_{\mathcal{E}}$ denotes its 0-1 indicator variable. We denote the total variation distance between two distributions μ, ν by $D_{\mathrm{TV}}(\mu, \nu)$.

For a density μ over $\mathbb{R}$, we denote $\mu^{\otimes S}$ to denote the product density over $\mathbb{R}^S$ whose coordinates are i.i.d. $\sim \mu$. We use $\mu^{\otimes d}$ to denote the case when $S = [d]$. More generally, we use $\bigotimes_{i\in[d]} \mu_i$ to mean the product density whose i^{th} coordinate is distributed $\sim \mu_i$.

For simplicity throughout the paper, we assume that all one-dimensional integration and sampling takes $O(1)$ time. All applications of this assumption involve reasonably well-behaved functions, and we expect this to be a fairly realistic assumption in practice.

Indexing. For a vector $\mathbf{v} \in \mathbb{R}^d$ and $S \subseteq [d]$, we use $\mathbf{v}_S \in \mathbb{R}^S$ to denote its restriction to its coordinates in S. We let $\mathrm{nnz}(\cdot)$ denote the number of nonzero entries in a matrix or vector argument. We let $\mathrm{supp}(\mathbf{v})$ denote the subset of nonzero entries in $\mathbf{v}$ (i.e., its support). For $\mathbf{M} \in \mathbb{R}^{n\times d}$, we use $\mathbf{M}_{i:}$ to denote its i^{th} row for $i \in [n]$, and $\mathbf{M}_{:j}$ to denote its j^{th} column for $j \in [d]$. When $S \subseteq [n]$ and $T \subseteq [d]$ are row and column indices, we let $\mathbf{M}_{S\times T}$ be the submatrix indexed by S and T; if $T = [d]$

we simply denote this submatrix as $\mathbf{M}_{S:}$ and similarly, we define $\mathbf{M}_{:T}$. We fix the convention that transposition is done prior to indexing, i.e., $\mathbf{M}^\top_{S\times T} := [\mathbf{M}^\top]_{S\times T}$.

A.2. Spike-and-slab posterior sampling

In this section, we formally define our problem. We begin by defining the *spike-and-slab* density, originally introduced by Mitchell and Beauchamp (1988). The spike-and-slab density is a simple product density that is parameterized by $\mathbf{q} \in [0,1]^d$ and a *diffuse density* μ over $\mathbb{R}$, given as follows:

$$\pi := \bigotimes_{i\in[d]} ((1-\mathbf{q}_i)\delta_0 + \mathbf{q}_i\mu)\,. \tag{14}$$

In other words, the i^{th} coordinate of $\boldsymbol{\theta} \sim \pi$ is nonzero with probability $\mathbf{q}_i$, for all $i \in [d]$. Finally, if μ is unspecified we assume $\mu = \mathcal{N}(0,1)$ is a standard Gaussian.

There is an alternative characterization of the distribution in (14) which follows by first describing the distribution π_{supp} of the support of $\boldsymbol{\theta} \sim \pi$, and then sampling $\boldsymbol{\theta}$ from the conditional distribution (i.e., via Bayes' theorem). We describe this alternative characterization as follows.

1. First, a support $S \subseteq [d]$ is sampled from π_{supp}, where

$$\pi_{\text{supp}}(S) = \left(\prod_{i\in S} \mathbf{q}_i\right)\left(\prod_{i\in S^c}(1-\mathbf{q}_i)\right) \text{ for all } S \subseteq [d]. \tag{15}$$

2. Second, $\boldsymbol{\theta}^\star_{S^c}$ is set to $\mathbf{0}_{S^c}$, and $\boldsymbol{\theta}^\star_S$ is sampled from

$$\boldsymbol{\theta}^\star_S \sim \mu^{\otimes S}. \tag{16}$$

We now define our posterior sampling problem.

Model 1 (Spike-and-slab posterior sampling) *Let $\mathbf{q} \in [0,1]^d$, $n, d \in \mathbb{N}$, $k = \|\mathbf{q}\|_1$, and $\sigma > 0$ be known, and let $\mu = \mathcal{N}(0,1)$. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a known measurement matrix, and suppose that we observe $(\mathbf{X}, \mathbf{y})$ where, following the notation (14), $\mathbf{y}$ is generated via*

$$\boldsymbol{\theta}^\star \sim \pi, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2\mathbf{I}_d), \quad \mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\xi}, \tag{7}$$

and $\boldsymbol{\theta}^\star$ and $\boldsymbol{\xi}$ are independent. Our goal is to sample from the posterior $\pi(\cdot \mid \mathbf{X}, y)$.

That is, $\mathbf{y}$ in (7) follows a standard noisy linear model, so our posterior task is a Bayesian linear regression problem with spike-and-slab prior (14). In light of our earlier viewpoint (15), (16) of sampling from π, there is similarly a two-stage process that characterizes sampling $\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$. We now provide an explicit description of this process in the case when $\mu = \mathcal{N}(0,1)$.

1. First, a support $S \subseteq [d]$ is sampled from $\pi_{\text{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$, where

$$\pi_{\text{supp}}(S \mid \mathbf{X}, \mathbf{y}) \propto \frac{(\prod_{i\in S}\mathbf{q}_i)(\prod_{i\in S^c}(1-\mathbf{q}_i))}{(2\pi)^{\frac{|S|}{2}}} \int_{\mathbb{R}^S} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_S\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta}_S\|_2^2\right) \mathrm{d}\boldsymbol{\theta}_S, \tag{17}$$

   and the integral extends each $\boldsymbol{\theta}_S \in \mathbb{R}^S$ to $\boldsymbol{\theta} \in \mathbb{R}^d$ via padding by zeroes.

2. Second, $\boldsymbol{\theta} \in \mathbb{R}^d$ is sampled from $\pi(\cdot \mid \mathbf{X}, \mathbf{y}, S)$, where

$$\pi(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, S) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{\theta}\right\|_2^2\right) \cdot \mathbb{I}_{\mathrm{supp}(\boldsymbol{\theta}) \subseteq S}. \tag{18}$$

In (17), we used the following standard calculation to compute the normalizing constants of the relevant multivariate Gaussian densities, which will be used several times throughout.

Fact 1 (Gaussian normalization) *Let $\boldsymbol{\Sigma} \in \mathbb{S}_{\succeq \mathbf{0}}^{d \times d}$ be non-singular and let $\boldsymbol{\mu} \in \mathbb{R}^d$. Then,*

$$\int \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{\mu}\right) \mathrm{d}\boldsymbol{\theta} = (2\pi)^{\frac{d}{2}} \exp\left(\frac{1}{2}\left\|\boldsymbol{\mu}\right\|_{\boldsymbol{\Sigma}}^2\right) \det(\boldsymbol{\Sigma})^{\frac{1}{2}}. \tag{19}$$

A.3. Sparse recovery

In this section, we state several preliminaries on noisy sparse recovery in the following model. We defer all proofs from this section to Appendix D, as most are well-known in the literature.

Model 2 (Sparse recovery) *Let $n, d \in \mathbb{N}$, $k \in [1, d]$, and $\sigma > 0$ be known. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a known measurement matrix, and let $\boldsymbol{\theta}^\star \in \mathbb{R}^d$, $\boldsymbol{\xi} \in \mathbb{R}^n$ satisfy $\mathrm{nnz}(\boldsymbol{\theta}^\star) \leq k$. Suppose we observe $(\mathbf{X}, \mathbf{y})$ where $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$. Our goal is to output an estimate of $\boldsymbol{\theta}^\star$.*

For a survey of techniques and known results in sparse recovery, we refer the reader to Chapter 7 of Wainwright (2019) and references therein. We provide two recovery results under Model 2 under assumptions on the pair $(\mathbf{X}, \boldsymbol{\xi})$. To state our assumptions, we require the following standard definition.

Definition 12 (Restricted isometry property) *We say $\mathbf{X} \in \mathbb{R}^{n \times d}$ satisfies the (ϵ, s)-restricted isometry property, or $\mathbf{X}$ is (ϵ, s)-*RIP*, if for all $\boldsymbol{\theta} \in \mathbb{R}^d$ with $\mathrm{nnz}(\boldsymbol{\theta}) \leq s$,*

$$(1 - \epsilon)\left\|\boldsymbol{\theta}\right\|_2^2 \leq \left\|\mathbf{X}\boldsymbol{\theta}\right\|_2^2 \leq (1 + \epsilon)\left\|\boldsymbol{\theta}\right\|_2^2.$$

An equivalent condition is that $\boldsymbol{\lambda}([\mathbf{X}^\top \mathbf{X}]_{S \times S}) \in [1 - \epsilon, 1 + \epsilon]^s$ for all $S \subseteq [d]$ with $|S| \leq s$.

Intuitively, RIP implies $\mathbf{X}$ acts as an approximate isometry when restricted to sparse vectors, so $[\mathbf{X}^\top \mathbf{X}]_{S \times S}$ is well-conditioned for any sparse support S. Various random matrix ensembles have been shown to satisfy Definition 7; we recall two such standard constructions below.

Proposition 13 (Theorem 9.2, Foucart and Rauhut (2013)) *Let $\mathbf{M}$ be an $n \times d$ random matrix, where entries of $\mathbf{M}$ are independent mean-zero sub-Gaussian random variables with variance 1. There exists a constant $C > 0$ such that for any $s \in [d]$, $\mathbf{X} = \frac{1}{\sqrt{n}}\mathbf{M}$ is (ϵ, s)-RIP with probability $\geq 1 - \delta$ if*

$$n \geq C \cdot \frac{s \ln \frac{d}{s} + \log \frac{1}{\delta}}{\epsilon^2}.$$

Proposition 14 (Theorem 4.5, Haviv and Regev (2017)) *Let $\mathbf{M} \in \mathbb{C}^{d \times d}$ be a discrete Fourier matrix whose elements are given by $\mathbf{M}_{jk} = d^{-1/2} \cdot \exp(2\pi\iota \cdot \frac{jk}{d})$ where $\iota := \sqrt{-1}$. Let $\mathbf{X} \in \mathbb{C}^{n \times d}$ be a matrix whose d rows are chosen uniformly and independently from the rows of $\mathbf{M}$ multiplied by $\sqrt{d/n}$. There exists a constant $C > 0$ such that for any $s \in [d]$, $\mathbf{X}$ is (ϵ, s)-RIP with probability $\geq 1 - 2^{-\Omega(\log d \log(s/\epsilon))}$ if*

$$n \geq \frac{s \cdot \log d \cdot \log^2(\frac{s}{\epsilon}) \cdot \log^2(\frac{1}{\epsilon})}{\epsilon^2}.$$

For more constructions of RIP matrix ensembles, including sampling bounded orthonormal systems and real trignometric polynomials, we refer the reader to Chapter 12 of Foucart and Rauhut (2013).

We next state the main sparse recovery results used in our samplers. These sparse recovery algorithms require some mild additional technical conditions for runtime bounds, which are satisfied by all of the constructions in Propositions 13 and 14 with high probability, after minor modifications.

Assumption 2 (Assumptions for ℓ_∞ sparse recovery) *In the setting of Model 2, suppose that $\mathbf{X} \in \mathbb{R}^{n \times d}$ is $(\frac{1}{4k}, k+1)$-RIP, that the columns of $\mathbf{X}$ are in general position (i.e., for any $r \in [n]$ and $S \subseteq [d]$ with $|S| = r$, $\mathbf{X}_{:S}$ is full-rank), and $\frac{\boldsymbol{\sigma}_1(\mathbf{X})}{\boldsymbol{\sigma}_n(\mathbf{X})} \leq \kappa$.*

In our applications of our framework (cf. Appendix B.3, Appendix C.3), we provide explicit estimates of κ which hold with high probability. The assumption that $\mathbf{X}$ has columns in general position holds with probability 1 if $\mathbf{X}$'s entries are drawn from continuous probability densities (measurable w.r.t. Lebesgue), as was observed in e.g., Tibshirani (2013). In other settings, we can enforce this general position assumption by adding infinitesimal noise to $\mathbf{X}$'s entries, and absorb the error under Model 2 into $\boldsymbol{\xi}$.

Assumption 3 (Assumptions for ℓ_2 sparse recovery) *In the setting of Model 2, suppose that for some $m \in [n]$, $\sqrt{m/n} \cdot \mathbf{X}_{[m]:}$ is $(\frac{1}{10}, C_2 k)$-RIP for a universal constant C_2.*

We include Assumption 3 for our sampling applications, which use $\mathbf{X}$ satisfying $(\epsilon, O(k))$-RIP for some $\epsilon \ll 1$. Since our ℓ_2 sparse recovery result only requires a constant ϵ parameter, but incurs error proportional to the noise level (scaling with the number of rows used), Assumption 3 (satisfied by all our RIP matrix ensembles) sharpens our measurement complexity's dependence on k.

We defer a proof of Proposition 15 to Appendices D.3 and D.4.

Proposition 15 *In the setting of Model 2, the following hold.*

1. *Under Assumption 2, there is an algorithm $\mathcal{A}_\infty$ that returns $\widehat{\boldsymbol{\theta}}$ satisfying*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_\infty \leq C_\infty r_\infty, \text{ for any } r_\infty \geq \left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty$$

   and for a universal constant C_∞. Moreover, $\mathcal{A}_\infty$ runs in time

$$O\left(d^{1.5} n^2 \log\left(d\kappa \cdot \left(1 + \frac{R_\infty}{r_\infty}\right)\right)\right), \text{ for any } R_\infty \geq \left\|\mathbf{X}^\top \mathbf{y}\right\|_\infty.$$

2. *Under Assumption 3, there is an algorithm $\mathcal{A}_2$ that returns $\widehat{\boldsymbol{\theta}}$ satisfying*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_2 \leq C_2 r_2, \text{ for any } r_2 \geq \left\|\boldsymbol{\xi}_{[m]}\right\|_2.$$

   Moreover, $\mathcal{A}_2$ runs in time

$$O\left(nd \log\left(\frac{R_2}{r_2}\right)\right), \text{ for any } R_2 \geq \left\|\boldsymbol{\theta}^\star\right\|_2.$$

Finally, we give a helper tool bounding quantities appearing in Proposition 15 for our model.

Lemma 16 *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ satisfy $(\epsilon, 1)$-RIP and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$. Then*

$$\left\|\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \leq \sigma(1+\epsilon)\sqrt{2\log\left(\frac{d}{\delta}\right)}, \text{ with probability } \geq 1-\delta, \text{ for all } \delta \in (0,1).$$

Appendix B. Spike-and-slab posterior sampling with Gaussian prior

Here, we prove Theorem 1 by giving our posterior sampler under Model 1 when $\mu = \mathcal{N}(0,1)$, i.e., a Gaussian prior is used. To motivate our development, we make the following observation.

Lemma 17 *In the setting of Model 1, let*

$$\mathbf{A} := \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{I}_d, \ \mathbf{b} := \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y},$$

and denote $\mathbf{A}_S := \mathbf{A}_{S\times S}$ for brevity. For some $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ with $T := \mathrm{supp}(\widehat{\boldsymbol{\theta}})$, let $\mathbf{z} := \mathbf{b} - \mathbf{A}\widehat{\boldsymbol{\theta}}$. Then,

$$\frac{\exp\left(\frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}}\right)}{\exp\left(\frac{1}{2}\left\|\mathbf{b}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}}\right)} = \frac{\exp\left(\frac{1}{2}\left\|\mathbf{z}_S\right\|^2_{\mathbf{A}_S^{-1}}\right)}{\exp\left(\frac{1}{2}\left\|\mathbf{z}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}}\right)}, \text{ for any } T \subseteq S, S' \subseteq [d].$$

Proof First, observe that since $T \subseteq S, S'$,

$$\left[\mathbf{A}\widehat{\boldsymbol{\theta}}\right]_S = \left[\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \mathbf{I}_d\right)\widehat{\boldsymbol{\theta}}\right]_S = \mathbf{A}_S\widehat{\boldsymbol{\theta}}_S,$$

and similarly $[\mathbf{A}\widehat{\boldsymbol{\theta}}]_{S'} = \mathbf{A}_{S'}\widehat{\boldsymbol{\theta}}_{S'}$. The result follows by directly expanding:

$$\begin{aligned}
\frac{1}{2}\left\|\mathbf{z}_S\right\|^2_{\mathbf{A}_S^{-1}} - \frac{1}{2}\left\|\mathbf{z}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}} &= \frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}} - \widehat{\boldsymbol{\theta}}_S^\top\mathbf{A}_S\mathbf{A}_S^{-1}\mathbf{b}_S + \frac{1}{2}\left\|\mathbf{A}_S\widehat{\boldsymbol{\theta}}_S\right\|^2_{\mathbf{A}_S^{-1}} \\
&\quad - \frac{1}{2}\left\|\mathbf{b}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}} + \widehat{\boldsymbol{\theta}}_{S'}^\top\mathbf{A}_{S'}\mathbf{A}_{S'}^{-1}\mathbf{b}_{S'} - \frac{1}{2}\left\|\mathbf{A}_{S'}\widehat{\boldsymbol{\theta}}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}} \\
&= \frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}} - \frac{1}{2}\left\|\mathbf{b}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}} - \widehat{\boldsymbol{\theta}}_S^\top\mathbf{b}_S + \widehat{\boldsymbol{\theta}}_{S'}^\top\mathbf{b}_{S'} \\
&\quad + \frac{1}{2}\widehat{\boldsymbol{\theta}}_S^\top\mathbf{A}_S\widehat{\boldsymbol{\theta}}_S - \frac{1}{2}\widehat{\boldsymbol{\theta}}_{S'}^\top\mathbf{A}_{S'}\widehat{\boldsymbol{\theta}}_{S'} = \frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}} - \frac{1}{2}\left\|\mathbf{b}_{S'}\right\|^2_{\mathbf{A}_{S'}^{-1}},
\end{aligned}$$

where in the last equality, we used our support set assumption. ■

To explain how Lemma 3 is used, a direct calculation (combining (17) with Fact 1, see Lemma 22) shows that up to a term depending only on $\mathbf{q}_S$, we have

$$\pi_{\mathrm{supp}}(S \mid \mathbf{X}, \mathbf{y}) \propto \exp\left(\frac{1}{2}\left\|\mathbf{b}_S\right\|^2_{\mathbf{A}_S^{-1}}\right).$$

We would like to sample from this distribution using rejection sampling (Lemma 9) on top of a product distribution. However, some coordinates of $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\boldsymbol{\theta}^\star + \mathbf{X}^\top\boldsymbol{\xi} \approx \boldsymbol{\theta}^\star + \mathbf{X}^\top\boldsymbol{\xi}$ are potentially much larger than others, depending on the randomness of the model. This non-uniformity breaks straightforward applications of rejection sampling.

We instead use Lemma 3 with an estimated "hint" vector $\widehat{\boldsymbol{\theta}}$ that has two properties: $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})$ with high probability, and $\mathbf{z} := \mathbf{b} - \mathbf{A}\widehat{\boldsymbol{\theta}}$ is small in an appropriate norm. Lemma 3 then lets us use a product distribution that depends on the size of $\mathbf{z}$, rather than the less uniform $\mathbf{b}$. This hint vector estimation effectively denoises our observations by first crudely estimating of the large coordinates of $\boldsymbol{\theta}^\star$ up to the noise level, via sparse recovery.

In Appendix B.1, we give helper results leveraging Proposition 15 to produce an estimate $\widehat{\boldsymbol{\theta}}$ to guide our sampler. In Appendix B.2, we give a rejection sampling procedure based on a conditional Poisson distribution, building upon Lemma 3 to efficiently sample from a high-probability region over $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$. Finally, we combine these results to prove Theorem 26 in Appendix B.3.

B.1. Posterior estimation

We first give an estimator $\widehat{\boldsymbol{\theta}}$ whose residual $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}$ is bounded in an appropriate norm, with high probability over our model. Our estimator builds upon an arbitrary ℓ_∞ sparse recovery algorithm, e.g., those in Proposition 15. We state the guarantees we require of such an algorithm as follows.

Assumption 1 *Let $\mathcal{A}(\mathbf{X}, \mathbf{y})$ be an algorithm with the following guarantee, parameterized by a universal constant $C_{\mathcal{A}}$ and $R_{\mathcal{A}} : \mathbb{N} \times (0,1) \times (0,1) \to \mathbb{R}_{\geq 0}$. In the setting of Model 2, suppose $\mathbf{X}$ is $(\epsilon, C_{\mathcal{A}}k)$-RIP for $\epsilon \in (0,1)$. Then $\mathcal{A}(\mathbf{X}, \mathbf{y})$ returns $\boldsymbol{\theta}'$ satisfying*

$$\left\|\boldsymbol{\theta}' - \boldsymbol{\theta}^\star\right\|_\infty \leq \sigma \cdot R_{\mathcal{A}}(k, \epsilon, \delta), \text{ with probability } \geq 1 - \delta,$$

for any $\delta \in (0,1)$, over the randomness of $\boldsymbol{\xi}$.[3]

For example, $\mathcal{A} \leftarrow \mathcal{A}_\infty$ in Proposition 15 satisfies Assumption 1 with $C_{\mathcal{A}} = 2$ and $R_{\mathcal{A}}(k, \epsilon, \delta) \approx \sqrt{\log(d/\delta)}$ whenever $\epsilon \leq \frac{1}{4k}$, by using Lemma 16. Similarly, $\mathcal{A} \leftarrow \mathcal{A}_2$ in Proposition 15 satisfies Assumption 1 with $C_{\mathcal{A}} \leftarrow C_2$ and $R_{\mathcal{A}}(k, \epsilon, \delta) \approx \sqrt{k}$ (up to a $\mathrm{polylog}(\frac{1}{\delta})$ factor) whenever $\epsilon \leq \frac{1}{10}$.

We now show how to use Assumption 1 to produce an estimator of the large coordinates of a sample from the posterior distribution $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ with high probability, and which has a bounded residual. First we recall the statement of Lemma 8, proven in Section 3.

Lemma 18 *Let Assumption 1 hold. In the setting of Model 1, let μ be arbitrary, let $\epsilon, \delta \in (0,1)$, and suppose $\mathbf{X}$ is $(C_{\mathcal{A}}k^\star, \epsilon)$-RIP for some $k^\star \in \mathbb{N}$ with $k^\star \geq 2(k + \log(\frac{5}{\delta}))$. Then there is an estimation procedure $\mathcal{A}_{\mathrm{est}}$ that takes as input $(\mathbf{X}, \mathbf{y})$, and produces $\widehat{\boldsymbol{\theta}}$ satisfying*

$$\mathbb{P}_{\boldsymbol{\theta} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{y})}\left[\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_\infty \leq 6\sigma \cdot R_{\mathcal{A}}\left(k^\star, \epsilon, \frac{\delta^2}{5}\right)\right) \wedge \left(\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq \mathrm{supp}(\boldsymbol{\theta})\right)\right] \geq 1 - \delta. \quad (10)$$

We further show that Lemma 8 gives concentration bounds on both the residual and our estimator.

Lemma 19 *In the setting of Lemma 8, given (10), then*

$$\begin{gathered}\left\|\mathbf{X}_{S:}^\top\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right)\right\|_2^2 \leq 16\sigma^2 k^\star \cdot \left(R_{\mathcal{A}}\left(k^\star, \epsilon, \frac{\delta^2}{5}\right)^2 + \log\left(\frac{5d}{\delta}\right)\right), \\ \text{for all } S \subseteq [d] \text{ with } |S| + 4\left(k + \log\left(\frac{5}{\delta}\right)\right) \leq k^\star,\end{gathered} \quad (20)$$

3. We state our result for deterministic algorithms $\mathcal{A}$ for simplicity, as Proposition 15 is deterministic.

with probability $\geq 1-\delta$ *over the randomness of Model 1.*

Proof We prove that (20) holds. By Lemma 16, with probability $\geq 1-\frac{\delta}{5}$,

$$\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty \leq \sigma\sqrt{8\log\left(\frac{5d}{\delta}\right)}.$$

Condition on the above event, $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \supseteq \mathrm{supp}(\boldsymbol{\theta}^\star)$, and (11), which all hold with probability $\geq 1-\delta$ over Model 1 by a union bound. Then letting $T := S \cup \mathrm{supp}(\boldsymbol{\theta}^\star)$, which satisfies $|T| \leq k^\star$ by assumption, we have

$$\begin{aligned}
\left\|\mathbf{X}_{S:}^\top\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\theta}}\right)\right\|_2^2 &\leq 2\left\|\mathbf{X}_{S:}^\top\mathbf{X}\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}\right)\right\|_2^2 + 2\left\|\mathbf{X}_{S:}^\top\boldsymbol{\xi}\right\|_2^2 \\
&\leq 2\left\|\left[\mathbf{X}^\top\mathbf{X}\right]_{T\times T}\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}\right)\right\|_2^2 + 2|S|\left\|\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty^2 \\
&\leq 2|T|\left\|\left[\mathbf{X}^\top\mathbf{X}\right]_{T\times T}\right\|_{\mathrm{op}}^2\left\|\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}\right\|_\infty^2 + 16\sigma^2k^\star\log\left(\frac{5d}{\delta}\right) \\
&\leq 8k^\star\left\|\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}\right\|_\infty^2 + 16\sigma^2k^\star\log\left(\frac{5d}{\delta}\right)
\end{aligned}$$

where the first line used the inequality $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2+2\|\mathbf{b}\|_2^2$, the second used that $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq \mathrm{supp}(\boldsymbol{\theta}^\star) \subseteq T$ and $S \subseteq T$, the third used our bound on $\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty$, and the last used our assumption that $\mathbf{X}$ is RIP so $\|[\mathbf{X}^\top\mathbf{X}]_{T\times T}\|_{\mathrm{op}} \leq 2$. The conclusion follows from the above and (11). ■

Lemma 20 *In the setting of Lemma 8, let* $k_0 := 4(k+\log(\frac{10}{\delta}))$*, and suppose that for some* $N:(0,1)\to\mathbb{R}_{\geq 0}$ *the density* μ *in Model 1 satisfies*

$$\mathbb{P}_{\mathbf{v}\sim\mu^{\otimes k_0}}\left[\|\mathbf{v}\|_2^2 \geq N(\delta)\right] \leq \frac{\delta}{2}, \text{ for all } \delta\in(0,1).$$

Then for $\widehat{\boldsymbol{\theta}}$ *the estimator returned by Lemma 8, we have*

$$\left\|\widehat{\boldsymbol{\theta}}\right\|_2^2 \leq 8k_0\sigma^2R_{\mathcal{A}}\left(k^\star,\epsilon,\frac{\delta}{10}\right)^2 + 2N(\delta), \text{ with probability } \geq 1-\delta,$$

for any $\delta\in(0,1)$*, over the randomness of Model 1.*

Proof As argued in Lemma 8, with probability $\geq 1-\frac{\delta}{2}$ over Model 1, we have both

$$\mathrm{supp}(\boldsymbol{\theta}^\star) \leq k_0, \ \left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star\right\|_\infty \leq 2\sigma\cdot R_{\mathcal{A}}\left(k^\star,\epsilon,\frac{\delta}{10}\right).$$

Condition on the above events and that $\|\boldsymbol{\theta}^\star\|_2^2 \leq N(\delta)$, which gives the failure probability of δ by a union bound. Under these events we have the conclusion:

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}\right\|_2^2 &\leq 2\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star\right\|_2^2 + 2\|\boldsymbol{\theta}^\star\|_2^2 \leq 2k_0\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star\right\|_\infty^2 + 2N(\delta) \\
&\leq 8k_0\sigma^2R_{\mathcal{A}}\left(k^\star,\epsilon,\frac{\delta}{10}\right)^2 + 2N(\delta).
\end{aligned}$$

■

We will instantiate Lemma 20 with different choices of μ. For the rest of this section, we focus on the basic setting of $\mu = \mathcal{N}(0,1)$, for which we have the following corollary.

Corollary 21 *Let $\delta \in (0,1)$. In the setting of Lemma 8, let $k_0 := 4(k + \log(\frac{10}{\delta}))$, and suppose $\mu = \mathcal{N}(0,1)$. Then for $\widehat{\boldsymbol{\theta}}$ the estimator returned by Lemma 8, all of the following hold with probability $\geq 1 - \delta$ over the randomness of Model 1: (10) (with $\delta \leftarrow \frac{\delta}{2}$), (20) (with $\delta \leftarrow \frac{\delta}{2}$), and*

$$\left\|\widehat{\boldsymbol{\theta}}\right\|_2^2 \leq 8k_0 \left(\sigma^2 R_{\mathcal{A}}\left(k^\star, \epsilon, \frac{\delta}{10}\right)^2 + 1\right).$$

Proof We apply Lemma 20 with the following bound from Fact 3 with $t \leftarrow \log(\frac{2}{\delta})$:

$$N(\delta) \leq 2k_0 + 3\log\left(\frac{2}{\delta}\right) \leq 4k_0.$$

■

B.2. Centered rejection sampling

We now develop our posterior sampler, an instance of rejection sampling (Lemma 9). In Algorithm 1, we first define the proposal distribution our rejection sampler is based on. This proposal is a product distribution that we show approximates the true posterior density $\pi(\cdot \mid \mathbf{X}, \mathbf{y})$ over a high-probability region of candidate subsets according to $\pi_{\text{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$. It takes an additional input $\widehat{\boldsymbol{\theta}}$ that is used to center its samples, as discussed following Lemma 3. Finally, it produces its output in Line 1 by calling our conditional Poisson sampler (Algorithm 5, cf. Lemma 10).

We next derive unnormalized distributions according to $\pi_{\text{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$ (cf. (17)) and Algorithm 1.

Lemma 22 *In the setting of Model 1, let $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$, and let $\tilde{\pi}_{\text{supp}}(\cdot) := \text{Law}(S)$ where $S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star}$ (defined in (13)) is a sample from Algorithm 1. Then,*

$$\tilde{\pi}_{\text{supp}}(S) \propto \left(\prod_{i \in S} \frac{\mathbf{q}_i}{1 - \mathbf{q}_i}\right) \exp\left(\frac{\sigma^2}{2(1+\sigma^2)} \|\mathbf{z}_S\|_2^2\right) \left(\frac{\sigma^2}{1+\sigma^2}\right)^{\frac{|S|}{2}} =: Q(S), \tag{21}$$

and further, the conditional distribution of $\pi_{\text{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$ over $\Omega_{\widehat{\boldsymbol{\theta}},k^\star}$ satisfies

$$\pi_{\text{supp}}\left(S \mid \mathbf{X}, \mathbf{y}, S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star}\right) \propto \left(\prod_{i \in S} \frac{\mathbf{q}_i}{1 - \mathbf{q}_i}\right) \exp\left(\frac{1}{2} \|\mathbf{z}_S\|_{\mathbf{A}_S^{-1}}^2\right) \frac{1}{\sqrt{\det \mathbf{A}_S}} =: P(S), \tag{22}$$

where $\mathbf{z}$ and $\mathbf{A}_S$ are defined as

$$\mathbf{z} := \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}) - \widehat{\boldsymbol{\theta}}, \quad \mathbf{A}_S := \frac{1}{\sigma^2}[\mathbf{X}^\top\mathbf{X}]_{S \times S} + \mathbf{I}_S. \tag{23}$$

Proof First we show (21) is true. Observe that the output of Algorithm 1 is the conditional distribution over $\Omega_{\widehat{\boldsymbol{\theta}},k^\star}$, of a Bernoulli product density over subsets of $[d]$. Without loss of generality, we let $i \in T$ be included in the output with probability $\frac{\mathbf{r}_i}{1-\mathbf{q}_i+\mathbf{r}_i}$ defined consistently with Lines 1 and 1. Hence, for the unconditional product distribution, we have

$$\begin{aligned}\tilde{\pi}_{\text{supp}}(S) &= \left(\prod_{i\in S}\frac{\mathbf{r}_i}{1-\mathbf{q}_i+\mathbf{r}_i}\right)\cdot\left(\prod_{i\in S^c}\frac{1-\mathbf{q}_i}{1-\mathbf{q}_i+\mathbf{r}_i}\right)\\ &= \prod_{i\in S}\left(\frac{\mathbf{r}_i}{1-\mathbf{q}_i}\right)\cdot\prod_{i\in[d]}(1-\mathbf{q}_i)\cdot\prod_{i\in[d]}(1-\mathbf{q}_i+\mathbf{r}_i)\\ &\propto \prod_{i\in S}\left(\frac{\mathbf{r}_i}{1-\mathbf{q}_i}\right),\end{aligned}$$

where $\mathbf{p}_i$ is defined in Algorithm 1. This proportionality remains true after conditioning on $S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star}$. Finally, by the definition of $\mathbf{r}_i$, we have the claim:

$$\begin{aligned}\tilde{\pi}_{\text{supp}}(S) &\propto \prod_{i\in S}\left(\frac{\mathbf{q}_i}{1-\mathbf{q}_i}\sqrt{\frac{\sigma^2}{1+\sigma^2}}\exp\left(\frac{\sigma^2}{2(1+\sigma^2)}\mathbf{z}_i^2\right)\right)\\ &= \left(\prod_{i\in S}\frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right)\exp\left(\frac{\sigma^2}{2(1+\sigma^2)}\left\|\mathbf{z}_S\right\|_2^2\right)\left(\frac{\sigma^2}{1+\sigma^2}\right)^{\frac{|S|}{2}}.\end{aligned}$$

Next we prove (22). Recall from (17) that

$$\begin{aligned}\pi_{\text{supp}}(S\mid\mathbf{X},\mathbf{y}) &\propto \left(\prod_{i\in S}\frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right)\left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}}\int_{\mathbb{R}^S}\exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\theta}_S\right\|_2^2-\frac{1}{2}\left\|\boldsymbol{\theta}_S\right\|_2^2\right)\mathrm{d}\boldsymbol{\theta}_S\\ &\propto \left(\prod_{i\in S}\frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right)\left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}}\int_{\mathbb{R}^S}\exp\left(-\frac{1}{2}\boldsymbol{\theta}_S^\top\mathbf{A}_S\boldsymbol{\theta}_S+\left(\frac{1}{\sigma^2}[\mathbf{X}^\top\mathbf{y}]_S\right)^\top\boldsymbol{\theta}_S\right)\mathrm{d}\boldsymbol{\theta}_S\\ &\propto \left(\prod_{i\in S}\frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right)\exp\left(\frac{1}{2}\left\|\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}\right\|_{\mathbf{A}_S^{-1}}^2\right)\frac{1}{\sqrt{\det\mathbf{A}_S}},\end{aligned}\tag{24}$$

where in the last line, we used Fact 1. The conclusion follows from combining (24) with Lemma 3 because the resulting distribution is supported on sets that all contain $\text{supp}(\widehat{\boldsymbol{\theta}})$. ■

We next bound the ratio between the unnormalized distributions P, Q defined in (21), (22).

Lemma 23 *Suppose that* $\mathbf{X}$ *is* $(\epsilon, k^\star)$*-RIP for* $\epsilon \in (0, \frac{1}{2})$*. Then for any* $S \subseteq [d]$ *such that* $\text{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$ *and* $|S| \le k^\star$*, we have* $\pi_{\text{supp}}(S \mid \mathbf{y}, \mathbf{X}) \propto P(S), \tilde{\pi}_{\text{supp}}(S \mid \mathbf{y}, \mathbf{X}) \propto Q(S)$ *and*

$$\left(\frac{1}{1+\epsilon}\right)^{\frac{k^\star}{2}}\exp\left(-\frac{\sigma^2\epsilon}{(1+\sigma^2)^2}\left\|\mathbf{z}_S\right\|_2^2\right)\le\frac{P(S)}{Q(S)}\le\left(\frac{1}{1-\epsilon}\right)^{\frac{k^\star}{2}}\exp\left(\frac{\sigma^2\epsilon}{(1+\sigma^2)^2}\left\|\mathbf{z}_S\right\|_2^2\right),$$

where $P(S), Q(S)$*, and* $\mathbf{z}_S$ *are defined in* (22)*,* (21)*, and* (23)*.*

Proof First, we can directly compute that

$$\begin{aligned}\frac{P(S)}{Q(S)} &= \frac{\left(\prod_{i\in S}\frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right)\exp\left(\frac{1}{2}\left\|\mathbf{z}_S\right\|_{\mathbf{A}_S^{-1}}^2\right)\det(\mathbf{A}_S^{-1})^{\frac{1}{2}}}{\left(\prod_{i\in S}\frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right)\left(\frac{\sigma^2}{1+\sigma^2}\right)^{\frac{|S|}{2}}\exp\left(\frac{\sigma^2}{2(1+\sigma^2)}\left\|\mathbf{z}_S\right\|_2^2\right)} \\ &= \det(\mathbf{A}_S^{-1})^{\frac{1}{2}}\cdot\left(1+\frac{1}{\sigma^2}\right)^{\frac{|S|}{2}}\exp\left(\frac{1}{2}\left(\left\|\mathbf{z}_S\right\|_{\mathbf{A}_S^{-1}}^2-\frac{\sigma^2}{1+\sigma^2}\left\|\mathbf{z}_S\right\|_2^2\right)\right) \\ &= \det(\mathbf{A}_S^{-1})^{\frac{1}{2}}\cdot\left(1+\frac{1}{\sigma^2}\right)^{\frac{|S|}{2}}\exp\left(\frac{1}{2}\mathbf{z}_S^\top\left(\mathbf{A}_S^{-1}-\frac{\sigma^2}{1+\sigma^2}\mathbf{I}_S\right)\mathbf{z}_S\right). \end{aligned} \tag{25}$$

Since $\mathbf{X}$ is $(\epsilon, k^\star)$ RIP and $|S|\le k^\star$,

$$\left(1+\frac{1-\epsilon}{\sigma^2}\right)\mathbf{I}_S \preceq \mathbf{A}_S = \frac{1}{\sigma^2}\left[\mathbf{X}^\top\mathbf{X}\right]_{S\times S}+\mathbf{I}_S \preceq \left(1+\frac{1+\epsilon}{\sigma^2}\right)\mathbf{I}_S.$$

Thus, we have that

$$\begin{aligned}\left(\frac{\sigma^2}{1+\sigma^2+\epsilon}\right)^{k^\star} \le \det\left(\mathbf{A}_S^{-1}\right) \le \left(\frac{\sigma^2}{1+\sigma^2-\epsilon}\right)^{k^\star}, \\ \left\|\mathbf{A}_S^{-1}-\frac{\sigma^2}{1+\sigma^2}\mathbf{I}_S\right\|_{\text{op}} \le \frac{\sigma^2\epsilon}{(1+\sigma^2)(1+\sigma^2-\epsilon)}.\end{aligned}$$

Plugging these bounds into (25), we have the desired claims for $\epsilon\in(0,\frac{1}{2})$:

$$\begin{aligned}\frac{P(S)}{Q(S)} &\le \left(\frac{1+\sigma^2}{1+\sigma^2-\epsilon}\right)^{\frac{k^\star}{2}}\exp\left(\frac{\sigma^2\epsilon}{2(1+\sigma^2)(1+\sigma^2-\epsilon)}\left\|\mathbf{z}_S\right\|_2^2\right) \\ &\le \left(\frac{1}{1-\epsilon}\right)^{\frac{k^\star}{2}}\exp\left(\frac{\sigma^2\epsilon}{(1+\sigma^2)^2}\left\|\mathbf{z}_S\right\|_2^2\right), \\ \frac{P(S)}{Q(S)} &\ge \left(\frac{1+\sigma^2}{1+\sigma^2+\epsilon}\right)^{\frac{k^\star}{2}}\exp\left(-\frac{\sigma^2\epsilon}{2(1+\sigma^2)(1+\sigma^2+\epsilon)}\left\|\mathbf{z}_S\right\|_2^2\right) \\ &\ge \left(\frac{1}{1+\epsilon}\right)^{\frac{k^\star}{2}}\exp\left(-\frac{\sigma^2\epsilon}{(1+\sigma^2)^2}\left\|\mathbf{z}_S\right\|_2^2\right).\end{aligned}$$

■

Proposition 24 *In the setting of Model 1, let $\widehat{\boldsymbol{\theta}}$ be produced as in Corollary 21 with $\delta\leftarrow\frac{\delta}{4}$, and let*

$$k^\star \ge 16\left(k+\log\left(\frac{40}{\delta}\right)\right),\ \epsilon\le\frac{1}{32k^\star\cdot\left(R_{\mathcal{A}}\left(k^\star,\epsilon,\frac{\delta^2}{320}\right)^2+\log\left(\frac{40}{\delta}\right)\right)}. \tag{26}$$

Then if $\mathbf{X}$ is $(\epsilon, C_{\mathcal{A}}k^\star)$-RIP for the universal constant $C_{\mathcal{A}}$ from Assumption 1, with probability $\ge 1-\delta$ over the randomness of Model 1, Algorithm 2 returns $S\sim\nu$ satisfying $D_{\text{TV}}(\nu,\pi_{\text{supp}}(\cdot\mid\mathbf{X},\mathbf{y}))\le\delta$. If $\widehat{\boldsymbol{\theta}}$ is given, Algorithm 2 runs in time

$$O\left(nd+\left(dk^\star+n\left(k^\star\right)^{\omega-1}\right)\log\left(\frac{1}{\delta}\right)\right).$$

Proof First, applying Corollary 21 and Lemma 19, with $\delta \leftarrow \frac{\delta}{4}$ shows that with probability $\geq 1 - \frac{\delta}{4}$ over the randomness of Model 1, we have

$$\begin{aligned} \mathbb{P}_{S \sim \pi_{\mathrm{supp}}(\cdot | \mathbf{X}, \mathbf{y})} \left[\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq \mathrm{supp}(\boldsymbol{\theta}) \right] &\geq 1 - \frac{\delta}{4}, \\ \left\| \widehat{\boldsymbol{\theta}} \right\|_2^2 &\leq 32 \left(k + \log\left(\frac{40}{\delta}\right) \right) \left(\sigma^2 R_{\mathcal{A}} \left(k^\star, \epsilon, \frac{\delta}{40} \right)^2 + 1 \right), \\ \text{and } \left\| \mathbf{X}_{S:}^\top \left(\mathbf{y} - \mathbf{X} \widehat{\boldsymbol{\theta}} \right) \right\|_2^2 &\leq 16 \sigma^2 k^\star \cdot \left(R_{\mathcal{A}} \left(k^\star, \epsilon, \frac{\delta^2}{320} \right)^2 + \log\left(\frac{40d}{\delta}\right) \right), \\ \text{for all } S \subseteq [d] \text{ with } |S| &\leq k^\star - 4 \left(k + \log\left(\frac{40}{\delta}\right) \right) = \frac{3k^\star}{4}. \end{aligned} \tag{27}$$

Next, by applying Corollary 5 with $\delta \leftarrow \frac{\delta^2}{8}$, we have by Lemma 6 that with probability $\geq 1 - \frac{\delta}{2}$ over the randomness of Model 1, we have

$$\mathbb{P}_{S \sim \pi_{\mathrm{supp}}(\cdot | \mathbf{X}, \mathbf{y})} \left[|S| \leq \frac{3k^\star}{4} \right] \leq \frac{\delta}{4}, \tag{28}$$

for our choice of $k^\star$. In the remainder of the proof, we union bound over all conditions in (27) and (28) holding, which gives the failure probability of δ over Model 1.

Next, following notation of Algorithm 2, for $S \in \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star}$ and $\mathbf{z}$ defined in (23), we have

$$\begin{aligned} \|\mathbf{z}_S\|_2^2 &\leq \frac{2}{\sigma^4} \left\| \mathbf{X}_{S:}^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}) \right\|_2^2 + 2 \left\| \widehat{\boldsymbol{\theta}} \right\|_2^2 \\ &\leq 32k^\star \left(R_{\mathcal{A}} \left(k^\star, \epsilon, \frac{\delta^2}{320} \right)^2 \left(\frac{1}{\sigma^2} + \sigma^2 \right) + 1 + \frac{\log(\frac{40}{\delta})}{\sigma^2} \right) \\ &\leq 32k^\star \cdot \frac{(1+\sigma^2)^2}{\sigma^2} \cdot \left(R_{\mathcal{A}} \left(k^\star, \epsilon, \frac{\delta^2}{320} \right)^2 + \log\left(\frac{40}{\delta}\right) \right), \end{aligned}$$

where we used the second and third bounds in (27). Thus, by Lemma 11, for $S \in \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star}$,

$$\begin{aligned} \frac{P(S)}{Q(S)} &\leq \left(1 + \frac{1}{32k^\star} \right)^{\frac{k^\star}{2}} \exp\left(\frac{\sigma^2 \epsilon}{(1+\sigma^2)^2} \|\mathbf{z}_S\|_2^2 \right) \\ &\leq 1.1 \exp\left(32\epsilon k^\star \cdot \left(R_{\mathcal{A}} \left(k^\star, \epsilon, \frac{\delta^2}{320} \right)^2 + \log\left(\frac{40}{\delta}\right) \right) \right) \leq 2, \end{aligned}$$

and an analogous calculation shows $\frac{Q(S)}{P(S)} \leq 2$. Therefore, Lemma 9 shows that Algorithm 2 correctly returns a sample within total variation distance $\frac{\delta}{2}$ from

$$\pi_{\mathrm{supp}} \left(S \mid \mathbf{X}, \mathbf{y}, S \in \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star} \right).$$

This distribution is in turn within total variation distance $\frac{\delta}{2}$ from the unconditional distribution $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$, by the first bound in (27) and (28). This concludes the proof of correctness.

For the runtime bound, we first precompute $\mathbf{z}$ required by Algorithm 1 in $O(nd)$ time. Next, by Lemma 9, we need to produce a sample $S \sim \mu$ and calculate the likelihood ratio $\frac{P(S)}{Q(S)}$ at most $O(\log(\frac{1}{\delta}))$ times. By Lemma 10, each draw from μ takes time $O(dk^\star)$. Further, by the definition (22), each evaluation of $P(S)$ takes time

$$O\left(n|S|^{\omega-1}\right) \le O\left(n\left(k^\star\right)^{\omega-1}\right).$$

To see this, it is clear that we can compute $\mathbf{A}_S$ in this time, and the costs of inverting (Strassen, 1969) and computing the determinant (Pan and Chen, 1999) do not dominate. Similarly, it is clear that evaluating $Q(S)$ only takes $O(|S|)$ time once $\mathbf{z}$ is precomputed. Combining these costs yields the claim. ■

We conclude with a closed-form expression for $\pi(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, S)$ in (18).

Lemma 25 *In the setting of Model 1, if $\mu = \mathcal{N}(0,1)$, then $\pi(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, S) = \mathcal{N}(\boldsymbol{\mu}_S, \mathbf{A}_S^{-1})$, where $\mathbf{A}_S$ is defined in (23), and $\boldsymbol{\mu} := \frac{1}{\sigma^2}\mathbf{A}_S^{-1}\mathbf{X}^\top\mathbf{y}$.*

Proof This follows immediately from rewriting (18): for $\boldsymbol{\theta}$ with $\mathrm{supp}(\boldsymbol{\theta}) \subseteq S$,

$$\begin{aligned}
\pi(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, S) &\propto \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{\theta}\right\|_2^2\right) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\left(\frac{1}{\sigma^2}\left[\mathbf{X}^\top\mathbf{X}\right]_{S\times S} + \mathbf{I}_S\right)\boldsymbol{\theta} + \frac{1}{\sigma^2}(\mathbf{X}^\top\mathbf{y})^\top\boldsymbol{\theta}\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_S)^\top\mathbf{A}_S(\boldsymbol{\theta} - \boldsymbol{\mu}_S)\right).
\end{aligned}$$

■

B.3. Proof of Theorem 26

We now instantiate Proposition 24 with the solvers from Proposition 15 to prove our main theorem.

Theorem 26 *Let $\delta \in (0,1)$. In the setting of Model 1, suppose $\mathbf{X}$ is $(\epsilon, k^\star)$-RIP where*

$$k^\star = \Omega\left(k + \log\left(\frac{1}{\delta}\right)\right),\ \epsilon = O\left(\frac{1}{\left(k + \log\left(\frac{1}{\delta}\right)\right)\log\left(\frac{d}{\delta}\right)}\right),$$

for appropriate constants. There is an algorithm (Algorithm 2 using $\widehat{\boldsymbol{\theta}}$ produced by $\mathcal{A}_\infty$ in Proposition 15) that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \le \delta$, with probability $\ge 1 - \delta$ over the randomness of Model 1. The algorithm runs in time

$$O\left(d^{1.5}n^2\log\left(d\kappa\left(1 + \frac{1}{\sigma}\right)\right) + d\log^2\left(\frac{1}{\delta}\right) + n\log^\omega\left(\frac{1}{\delta}\right)\right), \text{ for any } \kappa \ge \frac{\boldsymbol{\sigma}_1(\mathbf{X})}{\boldsymbol{\sigma}_n(\mathbf{X})}.$$

Alternatively, suppose $\mathbf{X}$ is $(\epsilon, k^\star)$-RIP, and for some $m \in [n]$, $\sqrt{m/n}\cdot\mathbf{X}_{[m]:}$ is $(\frac{1}{10}, k^\star)$-RIP where

$$k^\star = \Omega\left(k + \log\left(\frac{1}{\delta}\right)\right),\ \epsilon = O\left(\frac{1}{(k + \log(\frac{1}{\delta}))(m + \log\left(\frac{1}{\delta}\right))}\right),$$

for appropriate constants. There is an algorithm (Algorithm 2 using $\widehat{\boldsymbol{\theta}}$ produced by $\mathcal{A}_2$ in Proposition 15) that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \leq \delta$, with probability $\geq 1 - \delta$ over the randomness of Model 1. The algorithm runs in time

$$O\left(nd \log\left(1 + \frac{1}{\sigma}\right) + \left(dk^\star + n\,(k^\star)^{\omega-1}\right) \log\left(\frac{1}{\delta}\right)\right).$$

Proof The first claim follows from the ℓ_∞ recovery result in Proposition 15, combined with Proposition 24. Note that Lemma 16 bounds $R_{\mathcal{A}}(k^\star, \epsilon, \frac{\delta^2}{320})^2 = O(\log(\frac{d}{\delta}))$ in this setting. Moreover, to obtain our runtime bound, we use the estimates

$$r_\infty = \Theta\left(\sigma\sqrt{\log\left(\frac{d}{\delta}\right)}\right), \; R_\infty = \Theta\left(r_\infty\left(1 + \frac{d}{\sigma}\right)\right).$$

Correctness of the former bound follows from Lemma 16, and the latter follows from

$$\left\|\mathbf{X}^\top \mathbf{y}\right\|_\infty \leq \left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty + \left\|\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}^\star\right\|_\infty = O\left(r_\infty + \|\boldsymbol{\theta}^\star\|_2\right) = O\left(r_\infty\left(1 + \frac{d}{\sigma}\right)\right),$$

where the first equality holds since $\boldsymbol{\theta}^\star$ is $k^\star$-sparse within the allotted failure probability, and we can bound $\|\boldsymbol{\theta}^\star\|_2^2 = O(k^\star + \log(\frac{1}{\delta}))$ using Fact 3 to obtain the second.

The second claim follows from the ℓ_2 recovery result in Proposition 15, combined with Proposition 24. Again by Fact 3, we have that $R_{\mathcal{A}}(k^\star, \epsilon, \frac{\delta^2}{320})^2 = O(m + \log(\frac{1}{\delta}))$ in this setting. For the runtime, we set $R_2^2 = O(k^\star + \log(\frac{1}{\delta}))$ as above, and $r_2^2 = O(\sigma^2 \cdot (m + \log(\frac{1}{\delta})))$ as before. The conclusion follows because $m \geq k^\star$: since $\mathbf{X}_{[m]:}$ is RIP, otherwise column subsets of size $k^\star$ would not be full rank.

Finally, we note that in both cases Proposition 24 only refers to the distribution of the support $S \subseteq [d]$. However, by applying Lemma 25 on the sampled S, we can exactly sample from the corresponding distribution over $\boldsymbol{\theta}$, which cannot increase the total variation distance by using the same coupling over S. The runtime of this last step does not dominate. ■

To give a concrete example of applying Theorem 26 to a well-studied random matrix ensemble, consider the case where $\mathbf{X} \in \mathbb{R}^{n \times d}$ has i.i.d. entries $\sim \mathcal{N}(0, \frac{1}{n})$ (to which Proposition 13 applies). We require the following concentration inequalities on $\kappa(\mathbf{X})$.

Lemma 27 (Chen and Dongarra (2005)) *Assume that d is sufficiently larger than n. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ have entries i.i.d. $\sim \mathcal{N}(0, \frac{1}{n})$, and let $\delta \in (0, 1)$. Then with probability $\geq 1 - \delta$,*

$$\frac{\boldsymbol{\sigma}_1(\mathbf{X})}{\boldsymbol{\sigma}_n(\mathbf{X})} = O\left(d + \left(\frac{1}{\delta}\right)^{O(\frac{1}{d})}\right). \tag{29}$$

We remark that similar high-probability condition number bounds exist for other matrix ensembles, see, e.g., Theorem 4.6.1 in Vershynin (2018) for the case of entrywise sub-Gaussian matrices.

By plugging in Proposition 13 and Lemma 27 into Theorem 26, we have the following specialization of our posterior sampler to Gaussian measurement matrices $\mathbf{X}$.

Corollary 28 *Let $\delta \in (0,1)$, and suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ has i.i.d. entries $\sim \mathcal{N}(0, \frac{1}{n})$. If*

$$n = \Theta\left(\left(k + \log\left(\frac{1}{\delta}\right)\right)^3 \log^3\left(\frac{d}{\delta}\right)\right)$$

for an appropriate constant, in the setting of Model 1, there is an algorithm that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \le \delta$, with probability $\ge 1 - \delta$ over the randomness of Model 1 and $\mathbf{X}$. The algorithm runs in time

$$O\left(n^2 d^{1.5} \log\left(d\left(1 + \frac{1}{\sigma}\right)\right) + n^2\sqrt{d}\log\left(\frac{d}{\delta}\left(1 + \frac{1}{\sigma}\right)\right)\right).$$

Alternatively, if

$$n = \Theta\left(\left(k + \log\left(\frac{1}{\delta}\right)\right)^5 \log^2\left(\frac{d}{\delta}\right)\log\left(d\right)\right).$$

for an appropriate constant, in the setting of Model 1, there is an algorithm that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \le \delta$, with probability $\ge 1 - \delta$ over the randomness of Model 1 and $\mathbf{X}$. The algorithm runs in time

$$O\left(nd\log\left(1 + \frac{1}{\sigma}\right)\right).$$

Proof The result using the ℓ_∞ estimator follows immediately from combining Theorem 26 with Proposition 13 and Lemma 27. To obtain the result using the ℓ_2 estimator, we first set

$$m = O(k^\star \log(d))$$

which is enough to guarantee the precondition of Theorem 26 on $\sqrt{m/n} \cdot \mathbf{X}_{[m]:}$ holds, by Proposition 13. We substitute this into our choice of ϵ and obtain our bound on n, again through Proposition 13. In both cases, we simplified by dropping all non-dominant runtime terms. ■

Appendix C. Spike-and-slab posterior sampling with Laplace prior

Here, we prove Theorem 2 by generalizing the results of Appendix B to the setting where $\mu(x) \propto \exp(-|x|)$ (i.e., a Laplace prior) is used in Model 1, for a range of signal-to-noise ratios σ. We begin by giving an annealing strategy in Section C.1, following Ge et al. (2020) (see also Lovász and Vempala (2006); Cousins and Vempala (2018); Brosse et al. (2018)), for estimating normalizing constants used in our approximate rejection sampler. Next, we provide our proposal distribution and bound normalizing constant ratios in Appendix C.2. Finally, we combine these results in Appendix C.3 to prove Theorem 36, our main result on posterior sampling with a Laplace prior.

C.1. Annealing

In this section, we consider the following computational task, parameterized by an approximation tolerance $\Delta \in (0,1)$, a failure probability $\delta \in (0,1)$, a matrix $\mathbf{A} \in \mathbb{S}^{k \times k}_{\succeq \mathbf{0}}$ satisfying $\mu \mathbf{I}_k \preceq \mathbf{A} \preceq$

$L\mathbf{I}_k$, and a vector $\mathbf{b} \in \mathbb{R}^k$. Our goal is to produce an estimate $\widetilde{Z}$, such that with probability $\geq 1-\delta$,

$$\int \exp\left(-f(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} \in \left[(1-\Delta)\widetilde{Z}, (1+\Delta)\widetilde{Z}\right], \\ \text{where } f\left(\boldsymbol{\theta}\right) := \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta} - \mathbf{b}^\top \boldsymbol{\theta} + \|\boldsymbol{\theta}\|_1 . \tag{30}$$

This general question was studied by Ge et al. (2020), who gave the following framework based on an annealing scheme, patterned in part off of prior work by Lovász and Vempala (2006); Cousins and Vempala (2018); Brosse et al. (2018).

Proposition 29 (Section 3, Appendix B, Ge et al. (2020)) *Let $f : \mathbb{R}^k \to \mathbb{R}$ be μ-strongly convex, let $Z := \int \exp(-f(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}$, and fix an approximation tolerance $\Delta \in (0,1)$ and a failure probability $\delta \in (0,1)$. Let $\sigma_1 > 0$ be a parameter satisfying*

$$\left(1-\frac{\Delta}{2}\right)\int_{\boldsymbol{\theta}\in\mathbb{R}^k} \exp\left(-\frac{\|\boldsymbol{\theta}\|_2^2}{2\sigma_1^2}\right)\mathrm{d}\boldsymbol{\theta} \leq \int_{\boldsymbol{\theta}\in\mathbb{R}^k} \exp\left(-f_1\left(\boldsymbol{\theta}\right)\right)\mathrm{d}\boldsymbol{\theta} \leq \int_{\boldsymbol{\theta}\in\mathbb{R}^k} \exp\left(-\frac{\|\boldsymbol{\theta}\|_2^2}{2\sigma_1^2}\right)\mathrm{d}\boldsymbol{\theta}, \tag{31}$$

where, for a sequence of increasing $\{\sigma_i\}_{i\in[M]}$ specified in Ge et al. (2020), we let

$$f_i(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \frac{\|\boldsymbol{\theta}\|_2^2}{2\sigma_i^2}, \text{ for all } i \in [M].$$

Further, let

$$M = \Theta\left(\sqrt{k}\log\left(\frac{k}{\mu\sigma_1^2}\right)\right), \; N = \Theta\left(\frac{M^2}{\Delta^2}\right),$$

for appropriate constants. Finally, let $\mathcal{A}$ be an algorithm that takes as input $i \in [M]$, and produces a sample within total variation distance $\frac{1}{8N}$ from the density on $\mathbb{R}^k$ that is $\propto \exp(-f_i)$. There is an algorithm that queries $\mathcal{A}$ $O(N\log(\frac{1}{\delta}))$ times, and produces an estimate $\widetilde{Z}$ satisfying

$$Z \in \left[(1-\Delta)\widetilde{Z}, (1+\Delta)\widetilde{Z}\right], \text{ with probability } \geq 1-\delta.$$

Proof This is almost the exact derivation carried out in Ge et al. (2020), except for two differences. First, they assume that the relevant negative log-density f is L-smooth in addition to being strongly convex. However, one can check that the only place this assumption is used in their analysis is in Lemma 3.1 to choose σ_1 satisfying (31), which we replace with by isolating the sufficient condition (31). Second, the Ge et al. (2020) result is stated for an algorithm which produces a correct estimate Z with probability $\geq \frac{3}{4}$, using exactly N calls to $\mathcal{A}$. However, by repeating $\log(\frac{1}{\delta})$ times and taking a median, Chernoff bounds show that we can boost the failure probability as stated. ■

We next provide a sufficient choice of σ_1 that satisfies (31) in the setting of (30).

Lemma 30 *Let $\mathbf{A} \in \mathbb{S}^{k\times k}_{\succeq \mathbf{0}}$, let $\mathbf{b} \in \mathbb{R}^k$ such that $\|\mathbf{b}\|_2 \leq R$, let $\Delta \in (0,1)$, and let*

$$\lambda \geq \frac{8R\sqrt{k}}{\Delta} + \frac{192k}{\Delta^2}\left(k + \log\left(\frac{4}{\Delta}\right)\right). \tag{32}$$

Denoting $\mathbf{M} := \mathbf{A} + \lambda \mathbf{I}_k$*, we have*

$$\left(1 - \frac{\Delta}{2}\right) Z_{\mathbf{M},\mathbf{b}} \leq \int \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta} - \|\boldsymbol{\theta}\|_1\right) \mathrm{d}\boldsymbol{\theta} \leq Z_{\mathbf{M},\mathbf{b}} \tag{33}$$

where $Z_{\mathbf{M},\mathbf{b}} := \int \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta}$.

Proof The right-hand side of (33) is immediate from $\|\boldsymbol{\theta}\|_1 \geq 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^k$. In the rest of the proof, we prove the left-hand side of (33). Define the set $\Omega := \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq \frac{\Delta}{4\sqrt{k}}\}$. Then, we have

$$\begin{aligned}
\int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta} - \|\boldsymbol{\theta}\|_1\right) \mathrm{d}\boldsymbol{\theta} &\geq \int_{\Omega} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta} - \|\boldsymbol{\theta}\|_1\right) \mathrm{d}\boldsymbol{\theta} \\
&= \exp\left(-\sup_{\Omega} \|\boldsymbol{\theta}\|_1\right) \int_{\Omega} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta} \\
&\geq \exp\left(-\frac{\Delta}{4}\right) \int_{\Omega} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta},
\end{aligned} \tag{34}$$

where the last line used $\|\boldsymbol{\theta}\|_1 \leq \sqrt{k}\,\|\boldsymbol{\theta}\|_2$. Moreover, we claim that

$$\mathbb{P}_{\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{b}, \mathbf{M}^{-1})}\left[\boldsymbol{\theta} \in \Omega\right] \geq 1 - \frac{\Delta}{4}, \tag{35}$$

which completes the proof when combined with (34), since

$$\begin{aligned}
\exp\left(-\frac{\Delta}{4}\right) \int_{\Omega} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta} &\geq \left(1 - \frac{\Delta}{4}\right) \exp\left(-\frac{\Delta}{4}\right) \\
&\quad \cdot \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta}\right) \\
&\geq \left(1 - \frac{\Delta}{2}\right) Z_{\mathbf{M},\mathbf{b}},
\end{aligned}$$

as $\exp(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta})$ is the unnormalized density corresponding to $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{b}, \mathbf{M}^{-1})$. We now prove (35). First, observe that for $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{b}, \mathbf{M}^{-1})$, we can write

$$\boldsymbol{\theta} = \mathbf{M}^{-1}\mathbf{b} + \mathbf{M}^{-\frac{1}{2}}\mathbf{z}, \text{ where } \mathbf{z} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{I}_k).$$

Since $\mathbf{M} \succeq \lambda \mathbf{I}_k$, we have

$$\left\|\mathbf{M}^{-1}\mathbf{b}\right\|_2 \leq \frac{R}{\lambda} \leq \frac{\Delta}{8\sqrt{k}}, \ \left\|\mathbf{M}^{-\frac{1}{2}}\mathbf{z}\right\|_2 \leq \frac{\|\mathbf{z}\|_2}{\sqrt{\lambda}},$$

for our range of λ. Returning to (35), this implies the desired

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{b}, \mathbf{M}^{-1})}\left[\boldsymbol{\theta} \notin \Omega\right] &\leq \mathbb{P}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{I}_k)}\left[\left\|\mathbf{M}^{-\frac{1}{2}}\mathbf{z}\right\|_2 \geq \frac{\Delta}{8\sqrt{k}}\right] \\
&\leq \mathbb{P}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{I}_k)}\left[\|\mathbf{z}\|_2 \geq \frac{\Delta\sqrt{\lambda}}{8\sqrt{k}}\right] \leq \frac{\Delta}{4},
\end{aligned}$$

where we used Fact 3 with our choice of λ. ■

To facilitate our annealing scheme, we use a sampler from Lee et al. (2021) for composite densities.

Proposition 31 (Corollary 2, Lee et al. (2021)) *Let $f : \mathbb{R}^k \to \mathbb{R}$ be L-smooth and μ-strongly convex, let $\kappa := \frac{L}{\mu}$, and let $g : \mathbb{R}^k \to \mathbb{R}$ have the form $g(\mathbf{x}) = \sum_{i\in[k]} g_i(\mathbf{x}_i)$ for scalar functions $\{g_i : \mathbb{R} \to \mathbb{R}\}_{i\in[k]}$. There is an algorithm which runs in $O(\kappa k \log^3(\frac{\kappa k}{\epsilon}))$ iterations, each querying ∇f and performing $O(k)$ additional work, and obtains ϵ total variation distance to the density over $\mathbb{R}^d$, $\pi \propto \exp(-f - g)$.*

Proof This is almost the statement of Corollary 2 in Lee et al. (2021); we briefly justify the differences here. First, if g is coordinatewise-separable, then the RGO access required by Lee et al. (2021) takes $O(k)$ time, under our assumption of $O(1)$-time integration and sampling over $\mathbb{R}$. Next, Corollary 2 in Lee et al. (2021) requires access to the minimizer of $f + g$, but the tolerance of the algorithm to inexactness is discussed in Appendix A of that paper, and it is justified why the cost of computing an approximate minimizer for composite functions does not dominate the sampler's gradient complexity. Finally, the runtime in Lee et al. (2021) is in expectation, but can be bounded with probability $1 - O(\delta)$ by using standard Chernoff bounds over the complexity of rejection sampling steps. The failure probability of the runtime being bounded can be charged to the total variation distance, via a union bound. ■

We put together the pieces to obtain our desired estimator for normalizing constants.

Corollary 32 *Suppose that $\mathbf{A} \in \mathbb{S}^{k\times k}_{\succeq \mathbf{0}}$ satisfies $\mu\mathbf{I}_k \preceq \mathbf{A} \preceq L\mathbf{I}_k$, and that $\mathbf{b} \in \mathbb{R}^k$ satisfies $\|\mathbf{b}\|_2 \leq R$. Fix an approximation tolerance $\Delta \in (0, 1)$ and a failure probability $\delta \in (0, 1)$. There is an algorithm for computing $\widetilde{Z}$ satisfying (30), with probability $\geq 1 - \delta$, using time*

$$O\left(\frac{Lk^4}{\mu\Delta^2}\log^4\left(\frac{LRk}{\mu\Delta}\right)\log\left(\frac{1}{\delta}\right)\right).$$

Proof We instantiate Proposition 29 with $\sigma_1^{-2} \leftarrow \lambda$ satisfying (32), and $\mathcal{A} \leftarrow$ the sampler from Proposition 31. In particular, for any $i \in [M]$, the density $\propto \exp(-f_i)$ has the structure required by Proposition 31, where in the statement of Proposition 31, we set

$$f(\boldsymbol{\theta}) \leftarrow \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{M}\boldsymbol{\theta} - \mathbf{b}^\top\boldsymbol{\theta} + \frac{1}{2\sigma_i^2}\|\boldsymbol{\theta}\|_2^2, \; g(\boldsymbol{\theta}) \leftarrow \|\boldsymbol{\theta}\|_1.$$

Observe that for any $i \in [M]$, the parameter κ in Proposition 31 satisfies

$$\kappa \leq \frac{L + \frac{1}{\sigma_i^2}}{\mu + \frac{1}{\sigma_i^2}} \leq \frac{L}{\mu}.$$

Thus, the time complexity of each call to $\mathcal{A}$ is at most

$$O\left(\frac{Lk^3}{\mu}\log^3\left(\frac{NLk}{\mu}\right)\right), \text{ for } N = \Theta\left(\frac{k}{\Delta^2}\log\left(\frac{k}{\mu\sigma_1^2}\right)\right) = \Theta\left(\frac{k}{\Delta^2}\log\left(\frac{kR}{\mu\Delta}\right)\right),$$

since the cost of computing $\nabla f(\boldsymbol{\theta})$ is $O(k^2)$ time. The conclusion follows from Proposition 29. ■

C.2. Approximate centered rejection sampling

In this section, we give the analog of Algorithm 1 in the Laplace prior setting.

Algorithm 3 ProductSampleLaplace($\mathbf{X}, \mathbf{y}, \widehat{\boldsymbol{\theta}}, \sigma, \mathbf{q}, k^\star$)

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\sigma > 0$, $\mathbf{q} \in [0,1]^d$ produced by Model 1, $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$, $k^\star \in \mathbb{N}$
Output: Sample S from a conditional Poisson distribution over

$$S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star} := \left\{ S \subseteq [d] \mid S \supseteq \mathrm{supp}(\widehat{\boldsymbol{\theta}}), |S| \le k^\star \right\} \tag{36}$$

$\mathbf{u} \leftarrow \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}})$
$T \leftarrow \mathrm{supp}(\widehat{\boldsymbol{\theta}})$ **for** $i \in T^c$ **do**
  $\mathbf{v}_i^- \leftarrow \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1+\epsilon}{2\sigma^2}x^2 + \mathbf{u}_i x - |x|\right) \mathrm{d}x$
  $\mathbf{r}_i \leftarrow \mathbf{q}_i \cdot \mathbf{v}_i^-$
  $\mathbf{p}_i \leftarrow \frac{\mathbf{r}_i}{1-\mathbf{q}_i+\mathbf{r}_i}$
end
return $T \cup \mathsf{ConditionalPoisson}(\mathbf{p}, k^\star - |T|)$

Lemma 33 *In the setting of Model 1, take $\mu = \mathrm{Lap}(0,1)$. Let $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$, and let $\tilde{\pi}_{\mathrm{supp}}(\cdot) := \mathrm{Law}(S)$ where $S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star}$ (defined in (36)) is a sample from Algorithm 3. Then,*

$$\tilde{\pi}_{\mathrm{supp}}(S) \propto \exp\left(-\|\widehat{\boldsymbol{\theta}}\|_1\right) \prod_{i\in S} \left(\frac{\mathbf{q}_i \mathbf{v}_i^-}{1-\mathbf{q}_i}\right) =: Q(S), \tag{37}$$

and further, for $\mathbf{r} := \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}$, the conditional distribution of $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$ over $\Omega_{\widehat{\boldsymbol{\theta}},k^\star}$ satisfies

$$\begin{aligned} \pi_{\mathrm{supp}}\left(S \mid \mathbf{X}, \mathbf{y}, S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star}\right) &\propto \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \\ &\quad \cdot \int_{\mathbb{R}^S} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{X}_{:S}\boldsymbol{\Delta}_S - \mathbf{r}\right\|_2^2 - \left\|\boldsymbol{\Delta}_S + \widehat{\boldsymbol{\theta}}_S\right\|_1\right) \mathrm{d}\boldsymbol{\Delta}_S \\ &=: P(S). \end{aligned} \tag{38}$$

Furthermore, suppose that $\mathbf{X}$ is $(\epsilon, k^\star)$-RIP for $\epsilon \in (0, \frac{1}{2})$. Then if $\sigma \le \frac{1}{4}$, for any $S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star}$,

$$1 \le \frac{P(S)}{Q(S)} \le \exp\left(\frac{2\sigma^2\sqrt{k^\star\|\mathbf{u}_S\|_2^2} + 2\sigma^2\epsilon(\|\mathbf{u}_S\|_2^2 + k^\star)}{1-\epsilon^2} + \frac{6k^\star\sigma}{\sqrt{1-\epsilon}} + \frac{k^\star\epsilon}{1-\epsilon}\right). \tag{39}$$

Proof The proof that (37) holds proceeds analogously to the proof that (21) holds in Lemma 22. We next prove (38) holds. Following an analogous derivation to (17) but using $\mu = \mathrm{Lap}(0, 1)$,

$$\begin{aligned}\pi_{\mathrm{supp}}\left(S \mid \mathbf{X}, \mathbf{y}, S \in \Omega_{\widehat{\boldsymbol{\theta}},k^\star}\right) &\propto \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \\ &\quad \cdot \int_{\mathbb{R}^S} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{X}_{:S}\boldsymbol{\theta}_S\right\|_2^2 - \|\boldsymbol{\theta}_S\|_1\right) \mathrm{d}\boldsymbol{\theta}_S \\ &= \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \\ &\quad \cdot \int_{\mathbb{R}^S} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{X}_{:S}\boldsymbol{\Delta}_S - \mathbf{r}\right\|_2^2 - \left\|\boldsymbol{\Delta}_S + \widehat{\boldsymbol{\theta}}_S\right\|_1\right) \mathrm{d}\boldsymbol{\Delta}_S,\end{aligned} \tag{40}$$

where, performing the change of variables $\boldsymbol{\Delta} = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$, we have

$$\mathbf{X}_{:S}\boldsymbol{\theta}_S - \mathbf{y} = \mathbf{X}_{:S}(\boldsymbol{\Delta}_S + \widehat{\boldsymbol{\theta}}) - \mathbf{y} = \mathbf{X}_{:S}\boldsymbol{\Delta}_S - (\mathbf{y} - \mathbf{X}_{:S}\widehat{\boldsymbol{\theta}}_S) = \mathbf{X}_{:S}\boldsymbol{\Delta}_S - \mathbf{r},$$

and the last equality uses $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$. It remains to prove (39). First, define for all $i \in [d]$,

$$\begin{aligned}\mathbf{v}_i^- &:= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1+\epsilon}{2\sigma^2}x^2 + \mathbf{u}_i x - |x|\right) \mathrm{d}x, \\ \mathbf{v}_i^+ &:= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1-\epsilon}{2\sigma^2}x^2 + \mathbf{u}_i x + |x|\right) \mathrm{d}x.\end{aligned}$$

Since $[\mathbf{X}^\top\mathbf{X}]_{S\times S} \preceq (1+\epsilon)\mathbf{I}_S$, we have the lower bound

$$\begin{aligned}P(S) &= \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \int_{\mathbb{R}^S} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{X}_{:S}\boldsymbol{\Delta}_S - \mathbf{r}\right\|_2^2 - \left\|\boldsymbol{\Delta}_S + \widehat{\boldsymbol{\theta}}_S\right\|_1\right) \mathrm{d}\boldsymbol{\Delta}_S \\ &= \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \int_{\mathbb{R}^S} \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{\Delta}_S\right\|_{[\mathbf{X}^\top\mathbf{X}]_{S\times S}}^2 + \mathbf{u}_S^\top\boldsymbol{\Delta}_S - \left\|\widehat{\boldsymbol{\theta}}_S + \boldsymbol{\Delta}_S\right\|_1\right) \mathrm{d}\boldsymbol{\Delta}_S \\ &\overset{(a)}{\geq} \exp\left(-\|\widehat{\boldsymbol{\theta}}\|_1\right) \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \\ &\quad \cdot \int_{\mathbb{R}^S} \exp\left(-\frac{1+\epsilon}{2\sigma^2}\left\|\boldsymbol{\Delta}_S\right\|_2^2 + \mathbf{u}_S^\top\boldsymbol{\Delta}_S - \|\boldsymbol{\Delta}_S\|_1\right) \mathrm{d}\boldsymbol{\Delta}_S \\ &= \exp\left(-\|\widehat{\boldsymbol{\theta}}\|_1\right) \prod_{i\in S} \left(\frac{\mathbf{q}_i \mathbf{v}_i^-}{1-\mathbf{q}_i}\right),\end{aligned}$$

where (a) is by the triangle inequality on $\|\cdot\|_1$ and the fact that $\|\widehat{\boldsymbol{\theta}}\|_1 = \|\widehat{\boldsymbol{\theta}}_S\|_1$ as $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$.

Similarly, since $[\mathbf{X}^\top \mathbf{X}]_{S\times S} \succeq (1-\epsilon)\mathbf{I}_S$, we have the upper bound

$$\begin{aligned}
P(S) &= \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \int_{\mathbb{R}^S} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{X}_{:S}\Delta_S - \mathbf{r}\right\|_2^2 - \left\|\Delta_S + \widehat{\boldsymbol{\theta}}_S\right\|_1\right) \mathrm{d}\boldsymbol{\Delta}_S \\
&\le \exp\left(-\|\widehat{\boldsymbol{\theta}}\|_1\right) \left(\frac{1}{2\pi}\right)^{\frac{|S|}{2}} \left(\prod_{i\in S} \frac{\mathbf{q}_i}{1-\mathbf{q}_i}\right) \\
&\cdot \int_{\mathbb{R}^S} \exp\left(-\frac{1-\epsilon}{2\sigma^2}\left\|\boldsymbol{\Delta}_S\right\|_2^2 + \mathbf{u}_S^\top \boldsymbol{\Delta}_S + \|\boldsymbol{\Delta}_S\|_1\right) \mathrm{d}\boldsymbol{\Delta}_S \\
&= \exp\left(-\|\widehat{\boldsymbol{\theta}}\|_1\right) \prod_{i\in S} \left(\frac{\mathbf{q}_i \mathbf{v}_i^+}{1-\mathbf{q}_i}\right).
\end{aligned}$$

Now, by the definition of Q in (37), we have

$$1 \le \frac{P(S)}{Q(S)} \le \prod_{i\in S} \frac{\mathbf{v}_i^+}{\mathbf{v}_i^-}.$$

Next, we derive explicit formulas for $\mathbf{v}_i^+$ and $\mathbf{v}_i^-$. Let

$$\sigma_- := \frac{\sigma}{\sqrt{1-\epsilon}},\ \sigma_+ := \frac{\sigma}{\sqrt{1+\epsilon}},\ \mathbf{u}_i^+ := \mathbf{u}_i + 1,\ \mathbf{u}_i^- := \mathbf{u}_i - 1.$$

We have

$$\begin{aligned}
\mathbf{v}_i^- &= \frac{1}{\sqrt{2\pi}} \left(\int_{x\ge 0} \exp\left(-\frac{1}{2\sigma_+^2}x^2 + \mathbf{u}_i^- x\right) \mathrm{d}x + \int_{x\le 0} \exp\left(-\frac{1}{2\sigma_+^2}x^2 + \mathbf{u}_i^+ x\right) \mathrm{d}x\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\sigma_+^2(\mathbf{u}_i^-)^2}{2}\right) \int_{x\ge 0} \exp\left(-\frac{(x-\sigma_+^2\mathbf{u}_i^-)^2}{2\sigma_+^2}\right) \mathrm{d}x \\
&+ \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\sigma_+^2(\mathbf{u}_i^+)^2}{2}\right) \int_{x\le 0} \exp\left(-\frac{(x-\sigma_+^2\mathbf{u}_i^+)^2}{2\sigma_+^2}\right) \mathrm{d}x \\
&= \sigma_+ \left(\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^-)^2}{2}\right) \mathbb{P}\left(\mathcal{N}(\sigma_+^2\mathbf{u}_i^-, \sigma_+^2) \ge 0\right) + \exp\left(\frac{\sigma_+^2(\mathbf{u}_i^+)^2}{2}\right) \mathbb{P}\left(\mathcal{N}(\sigma_+^2\mathbf{u}_i^+, \sigma_+^2) \le 0\right)\right).
\end{aligned}$$

Similarly, we have

$$\mathbf{v}_i^+ = \sigma_- \left(\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^+)^2}{2}\right) \mathbb{P}\left(\mathcal{N}(\sigma_-^2\mathbf{u}_i^+, \sigma_-^2) \ge 0\right) + \exp\left(\frac{\sigma_-^2(\mathbf{u}_i^-)^2}{2}\right) \mathbb{P}\left(\mathcal{N}(\sigma_-^2\mathbf{u}_i^-, \sigma_-^2) \le 0\right)\right).$$

Using the fact that $\mathbb{P}[\mathcal{N}(\mu,\sigma^2) \le 0] = \mathbb{P}[Z \le -\frac{\mu}{\sigma}]$, where $Z \sim \mathcal{N}(0,1)$, we have:

$$\begin{aligned}
\mathbf{v}_i^+ &= \sigma_- \left(\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^+)^2}{2}\right) \mathbb{P}\left[Z \ge -\sigma_-\mathbf{u}_i^+\right] + \exp\left(\frac{\sigma_-^2(\mathbf{u}_i^-)^2}{2}\right) \mathbb{P}\left[Z \le -\sigma_-\mathbf{u}_i^-\right]\right), \\
\mathbf{v}_i^- &= \sigma_+ \left(\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^-)^2}{2}\right) \mathbb{P}\left[Z \ge -\sigma_+\mathbf{u}_i^-\right] + \exp\left(\frac{\sigma_+^2(\mathbf{u}_i^+)^2}{2}\right) \mathbb{P}\left[Z \le -\sigma_+\mathbf{u}_i^+\right]\right).
\end{aligned}$$

Thus, we have

$$\frac{\mathbf{v}_i^+}{\mathbf{v}_i^-} \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}} \cdot \frac{\max\left(\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^+)^2}{2}\right), \exp\left(\frac{\sigma_-^2(\mathbf{u}_i^-)^2}{2}\right)\right)\left(\mathbb{P}\left[Z \geq -\sigma_-\mathbf{u}_i^+\right] + \mathbb{P}\left[Z \leq -\sigma_-\mathbf{u}_i^-\right]\right)}{\min\left(\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^-)^2}{2}\right), \exp\left(\frac{\sigma_+^2(\mathbf{u}_i^+)^2}{2}\right)\right)\left(\mathbb{P}\left[Z \geq -\sigma_+\mathbf{u}_i^-\right] + \mathbb{P}\left[Z \leq -\sigma_+\mathbf{u}_i^+\right]\right)}.$$

We begin by bounding the exponential terms in the above display:

$$\begin{aligned}
\frac{\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^+)^2}{2}\right)}{\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^-)^2}{2}\right)} &= \exp\left(\frac{\sigma^2}{2}\left(\frac{(\mathbf{u}_i+1)^2}{1-\epsilon} - \frac{(\mathbf{u}_i-1)^2}{1+\epsilon}\right)\right) \\
&= \exp\left(\frac{\sigma^2}{1-\epsilon^2}\left(2\mathbf{u}_i - \epsilon(\mathbf{u}_i^2+1)\right)\right) \leq \exp\left(\frac{2\sigma^2|\mathbf{u}_i|}{1-\epsilon^2}\right), \\
\frac{\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^-)^2}{2}\right)}{\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^+)^2}{2}\right)} &= \exp\left(\frac{\sigma^2}{2}\left(\frac{(\mathbf{u}_i-1)^2}{1-\epsilon} - \frac{(\mathbf{u}_i+1)^2}{1+\epsilon}\right)\right) \\
&= \exp\left(\frac{\sigma^2}{(1-\epsilon^2)}\left(-2\mathbf{u}_i + \epsilon(\mathbf{u}_i^2+1)\right)\right) \leq \exp\left(\frac{2\sigma^2|\mathbf{u}_i| + \sigma^2\epsilon(\mathbf{u}_i^2+1)}{1-\epsilon^2}\right), \\
\frac{\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^+)^2}{2}\right)}{\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^+)^2}{2}\right)} &= \exp\left(\frac{\sigma^2(\mathbf{u}_i+1)^2}{2}\left(\frac{1}{1-\epsilon} - \frac{1}{1+\epsilon}\right)\right) \\
&\leq \exp\left(\frac{\sigma^2\epsilon(\mathbf{u}_i+1)^2}{1-\epsilon^2}\right) \leq \exp\left(\frac{2\sigma^2\epsilon(\mathbf{u}_i^2+1)}{1-\epsilon^2}\right), \\
\frac{\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^-)^2}{2}\right)}{\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^-)^2}{2}\right)} &= \exp\left(\frac{\sigma^2(\mathbf{u}_i-1)^2}{2}\left(\frac{1}{1-\epsilon} - \frac{1}{1+\epsilon}\right)\right) \\
&\leq \exp\left(\frac{\sigma^2\epsilon(\mathbf{u}_i-1)^2}{1-\epsilon^2}\right) \leq \exp\left(\frac{2\sigma^2\epsilon(\mathbf{u}_i^2+1)}{1-\epsilon^2}\right).
\end{aligned}$$

Therefore, combining these cases, we have

$$\frac{\max\left(\exp\left(\frac{\sigma_-^2(\mathbf{u}_i^+)^2}{2}\right), \exp\left(\frac{\sigma_-^2(\mathbf{u}_i^-)^2}{2}\right)\right)}{\min\left(\exp\left(\frac{\sigma_+^2(\mathbf{u}_i^-)^2}{2}\right), \exp\left(\frac{\sigma_+^2(\mathbf{u}_i^+)^2}{2}\right)\right)} \leq \exp\left(\frac{2\sigma^2|\mathbf{u}_i| + 2\sigma^2\epsilon(\mathbf{u}_i^2+1)}{1-\epsilon^2}\right).$$

Next we consider the ratio of probabilities. Suppose that $\epsilon, 4\sigma/\sqrt{2\pi} \le 1/2$.

$$\mathbb{P}\left[Z \ge -\sigma_- \mathbf{u}_i^+\right] + \mathbb{P}\left[Z \le -\sigma_- \mathbf{u}_i^-\right] = 1 + \mathbb{P}\left[Z \in \left[-\sigma_- \mathbf{u}_i^+, -\sigma_- \mathbf{u}_i^-\right]\right] \le 1 + 2\sigma_-,$$
$$\mathbb{P}\left[Z \ge -\sigma_+ \mathbf{u}_i^-\right] + \mathbb{P}\left[Z \le -\sigma_+ \mathbf{u}_i^+\right] = 1 - \mathbb{P}\left[Z \in \left[-\sigma_+ \mathbf{u}_i^-, -\sigma_+ \mathbf{u}_i^+\right]\right] \ge 1 - 2\sigma_+,$$

since the PDF of $\mathcal{N}(0,1)$ is pointwise bounded by 1. Combining, we have shown

$$\frac{\mathbf{v}_i^+}{\mathbf{v}_i^-} = \sqrt{\frac{1+\epsilon}{1-\epsilon}} \exp\left(\frac{2\sigma^2|\mathbf{u}_i| + 2\sigma^2\epsilon(\mathbf{u}_i^2+1)}{1-\epsilon^2}\right) \frac{1 + \frac{2\sigma}{\sqrt{1-\epsilon}}}{1 - \frac{2\sigma}{\sqrt{1+\epsilon}}}$$

Taking a product over the $|S| \le k^\star$ coordinates, and using $\frac{1}{1-x} \le 1 + \frac{x}{2}$ for $x \in (0, \frac{1}{2})$, as well as $\sqrt{\frac{1+\epsilon}{1-\epsilon}} = \sqrt{1 + \frac{2\epsilon}{1-\epsilon}} \le \exp(\frac{\epsilon}{1-\epsilon})$, we have the desired result:

$$\begin{aligned}\prod_{i\in S} \frac{\mathbf{v}_i^+}{\mathbf{v}_i^-} &\le \exp\left(\sum_{i\in S} \frac{2\sigma^2|\mathbf{u}_i| + 2\sigma^2\epsilon(\mathbf{u}_i^2+1)}{1-\epsilon^2}\right) \exp\left(\frac{2k^\star\sigma}{\sqrt{1-\epsilon}} + \frac{4k^\star\sigma}{\sqrt{1+\epsilon}}\right) \exp\left(\frac{k^*\epsilon}{1-\epsilon}\right) \\ &\le \exp\left(\frac{2\sigma^2\sqrt{k^\star\|\mathbf{u}_S\|_2^2} + 2\sigma^2\epsilon(\|\mathbf{u}_S\|_2^2 + k^\star)}{1-\epsilon^2} + \frac{6k^\star c\sigma}{\sqrt{1-\epsilon}} + \frac{k^\star\epsilon}{1-\epsilon}\right).\end{aligned}$$

■

Our posterior sampler uses rejection sampling with the unnormalized densities described in (38), (37). However, one difficulty is that unlike in the Gaussian case, there is no explicit formula (e.g., Fact 1) for $P(S)$. In Appendix C.1, we gave an algorithm for estimating $P(S)$ to multiplicative error, based on sampling access to the induced distribution over $\boldsymbol{\theta}_S \in \mathbb{R}^S$. To facilitate using these approximate evaluations of P, we provide the following helper tool.

Lemma 34 *Let $\pi, \tilde{\pi}$ be distributions over the some domain Ω, and suppose that $\pi \propto P$ and $\tilde{\pi} \propto \widetilde{P}$ for unnormalized densities $P, \widetilde{P}$. Moreover, suppose that for all $\omega \in \Omega$,*

$$\frac{P(\omega)}{\widetilde{P}(\omega)} \in [1-\Delta, 1+\Delta]$$

for some $\Delta \in (0, \frac{1}{2})$. Then, $D_{\mathrm{TV}}(\pi, \tilde{\pi}) \le \frac{3\Delta}{2}$.

Proof Throughout the proof, denote the relevant normalization constants by

$$Z := \int_\Omega P(\omega)\mathrm{d}\omega, \ \widetilde{Z} := \int_\Omega \widetilde{P}(\omega)\mathrm{d}\omega,$$

so $Z \in [(1-\Delta)\widetilde{Z}, (1+\Delta)\widetilde{Z}]$, and thus $\frac{1}{Z} \in [\frac{1}{(1+\Delta)\widetilde{Z}}, \frac{1}{(1-\Delta)\widetilde{Z}}] \subseteq [\frac{1-2\Delta}{\widetilde{Z}}, \frac{1+2\Delta}{\widetilde{Z}}]$. Then, we have the claim by applying the triangle inequality:

$$\begin{aligned}D_{\mathrm{TV}}(\pi, \tilde{\pi}) &= \frac{1}{2}\int_\Omega \left|\frac{P(\omega)}{Z} - \frac{\widetilde{P}(\omega)}{\widetilde{Z}}\right| \mathrm{d}\omega \\ &\le \frac{1}{2}\int_\Omega \left|\frac{P(\omega)}{Z} - \frac{\widetilde{P}(\omega)}{Z}\right| \mathrm{d}\omega + \frac{1}{2}\int_\Omega \left|\frac{\widetilde{P}(\omega)}{Z} - \frac{\widetilde{P}(\omega)}{\widetilde{Z}}\right| \mathrm{d}\omega \\ &\le \frac{\Delta}{2}\int_\Omega \frac{P(\omega)}{Z}\mathrm{d}\omega + \frac{2\Delta}{2}\int_\Omega \frac{\widetilde{P}(\omega)}{\widetilde{Z}}\mathrm{d}\omega = \frac{3\Delta}{2}.\end{aligned}$$

■

Finally, by combining these developments, we obtain the analog of Proposition 24.

Algorithm 4 PosteriorSampleLaplace$(\mathbf{X}, \mathbf{y}, \widehat{\boldsymbol{\theta}}, \sigma, \mathbf{q}, k^\star, \delta)$

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\sigma > 0$, $\mathbf{q} \in [0,1]^d$ produced by Model 1 with $\mu = \mathrm{Lap}(0,1)$, $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$, $k^\star \in \mathbb{N}$, $\delta \in (0,1)$

Output: Sample S approximately distributed as $\pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})$ (cf. Proposition 35)

$\mu \leftarrow$ ProductSampleLaplace$(\mathbf{X}, \mathbf{y}, \widehat{\boldsymbol{\theta}}, \sigma, \mathbf{q}, k^\star)$

$P, Q \leftarrow$ unnormalized distributions in (38), (37)

$\Omega \leftarrow \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star}$ defined in (36)

$\widetilde{P} \leftarrow$ unnormalized distribution satisfying $P \in [(1 - \frac{\delta}{12})\widetilde{P}, (1 + \frac{\delta}{12})\widetilde{P}]$ over Ω, using Corollary 32

return RejectionSample$(\mu, \widetilde{P}, Q, \Omega, 4, \frac{\delta}{8})$

Proposition 35 *In the setting of Model 1 where $\mu = \mathrm{Lap}(0,1)$, let $\widehat{\boldsymbol{\theta}}$ be produced as in Corollary 21 with $\delta \leftarrow \frac{\delta}{4}$, let $k^\star$, ϵ satisfy (26), and let $\sigma \le \frac{1}{6k^\star}$. Then if $\mathbf{X}$ is $(\epsilon, C_{\mathcal{A}}k^\star)$-RIP for the universal constant $C_{\mathcal{A}}$ from Assumption 1, with probabilty $\ge 1 - \delta$ over the randomness of Model 1, Algorithm 4 returns $S \sim \nu$ satisfying $D_{\mathrm{TV}}(\nu, \pi_{\mathrm{supp}}(\cdot \mid \mathbf{X}, \mathbf{y})) \le \delta$. If $\widehat{\boldsymbol{\theta}}$ is given, Algorithm 4 runs in time*

$$O\left(nd + n\,(k^\star)^{\omega-1} \log\left(\frac{1}{\delta}\right) + \frac{(k^\star)^4}{\delta^2} \log^4\left(\frac{1}{\sigma}\log\left(\frac{d}{\delta}\right)\right) \log^2\left(\frac{1}{\delta}\right)\right).$$

Proof As in the proof of Proposition 24, with probability $\ge 1 - \delta$ over the randomness of Model 1, all of the bounds in (27) and (28) hold, except for the second bound in (27) which is not used in this proof. In particular, this shows that with probability $\ge 1 - \delta$ over Model 1,

$$\|\mathbf{u}_S\|_2^2 \le \frac{16k^\star}{\sigma^2}\left(R_{\mathcal{A}}\left(k^\star, \epsilon, \frac{\delta^2}{320}\right) + \log\left(\frac{40d}{\delta}\right)\right),$$

for all $S \in \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star}$. Thus, by Lemma 33,

$$\frac{P(S)}{Q(S)} \le \exp\left(\frac{2\sigma^2\sqrt{k^\star\|\mathbf{u}_S\|_2^2} + 2\sigma^2\epsilon(\|\mathbf{u}_S\|_2^2 + k^\star)}{1 - \epsilon^2} + \frac{6k^\star\sigma}{\sqrt{1-\epsilon}} + \frac{k^\star\epsilon}{1-\epsilon}\right) \le 3,$$

for our choice of parameters in (26) and $\sigma \le \frac{1}{6k^\star}$. Thus, we can check that any $\widetilde{P}$ satisfying the criterion on Line 4 over $\Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star}$ has

$$\frac{1}{4} \le \frac{\widetilde{P}(S)}{Q(S)} \le 4, \text{ for all } S \in \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star}.$$

Therefore, Lemmas 9 and 34 show that Algorithm 4 samples within total variation $\frac{\delta}{4}$ from $\pi_{\mathrm{supp}}(S \mid \mathbf{X}, \mathbf{y}, S \in \Omega_{\widehat{\boldsymbol{\theta}}, \frac{3}{4}k^\star})$, which is in turn within total variation $\frac{\delta}{2}$ from $\pi_{\mathrm{supp}}(S \mid \mathbf{X}, \mathbf{y})$. Our argument

allows for a failure probability of $O(\frac{\delta}{\log(1/\delta)})$ in each computation of $\widetilde{P}$, so that all calls to $\widetilde{P}$ used by Lemma 9 succeed with probability $\geq 1 - \frac{\delta}{4}$.[4] This concludes the correctness proof.

For the runtime, we proceed similarly to Proposition 24, except we substitute the cost of computing $[\mathbf{X}^\top \mathbf{X}]_{S\times S}$ and $\widetilde{P}$ to the required accuracy by using Corollary 32 in each iteration, with $\delta \leftarrow O(\frac{\delta}{\log(1/\delta)})$, $\frac{L}{\mu} = O(1)$ (due to RIP), and for $\mathbf{b} := \frac{1}{\sigma^2}\mathbf{X}_{S:}^\top \mathbf{y}$,

$$\|\mathbf{b}\|_2 \leq \frac{1}{\sigma^2}\left\|\mathbf{X}_{S:}^\top \mathbf{y}\right\| = O\left(\text{poly}\left(k^\star, \log\left(\frac{d}{\delta}\right), \frac{1}{\sigma}\right)\right).$$

The above holds by using our bound on $\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty$ in the proof of Lemma 8, and because

$$\|\boldsymbol{\theta}^\star\|_2^2 = O\left(k^\star \log^2\left(\frac{d}{\delta}\right)\right) \tag{41}$$

within the failure probability on Model 1, by standard tail bounds on Laplace random variables. ■

C.3. Proof of Theorem 36

We now combine Propositions 15 and 35 to provide our main sampling result under a Laplace prior.

Theorem 36 *Let $\delta \in (0,1)$. In the setting of Model 1, suppose $\mu = \text{Lap}(0,1)$, $\sigma = O(\frac{1}{k^\star})$, and that $\mathbf{X}$ is $(\epsilon, k^\star)$-RIP, where*

$$k^\star = \Omega\left(k + \log\left(\frac{1}{\delta}\right)\right),\ \epsilon = O\left(\frac{1}{\left(k + \log\left(\frac{1}{\delta}\right)\right)\log\left(\frac{d}{\delta}\right)}\right),$$

for appropriate constants. There is an algorithm (Algorithm 4 using $\widehat{\boldsymbol{\theta}}$ produced by $\mathcal{A}_\infty$ in Proposition 15) that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\text{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \leq \delta$, with probability $\geq 1 - \delta$ over the randomness of Model 1. The algorithm runs in time

$$O\left(d^{1.5} n^2 \log\left(\frac{d\kappa \log\left(\frac{1}{\delta}\right)}{\sigma}\right) + \frac{(k^\star)^4}{\delta^2}\log^4\left(\frac{\log\left(\frac{d}{\delta}\right)}{\sigma}\right)\log^2\left(\frac{1}{\delta}\right)\right).$$

Alternatively, suppose $\mu = \text{Lap}(0,1)$, $\sigma = O(\frac{1}{k^\star})$, $\mathbf{X}$ is $(\epsilon, k^\star)$-RIP, and for some $m \in [n]$, $\sqrt{m/n} \cdot \mathbf{X}_{[m]:}$ is $(\frac{1}{10}, k^\star)$-RIP where

$$k^\star = \Omega\left(k + \log\left(\frac{1}{\delta}\right)\right),\ \epsilon = O\left(\frac{1}{\left(k + \log\left(\frac{1}{\delta}\right)\right)\left(m + \log\left(\frac{1}{\delta}\right)\right)}\right),$$

for appropriate constants. There is an algorithm (Algorithm 4 using $\widehat{\boldsymbol{\theta}}$ produced by $\mathcal{A}_2$ in Proposition 15) that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\text{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \leq \delta$, with probability $\geq 1 - \delta$ over the randomness of Model 1. The algorithm runs in time

$$O\left(nd\log\left(\frac{\log\left(\frac{d}{\delta}\right)}{\sigma}\right) + n\,(k^\star)^{\omega-1}\log\left(\frac{1}{\delta}\right) + \frac{(k^\star)^4}{\delta^2}\log^4\left(\frac{\log\left(\frac{d}{\delta}\right)}{\sigma}\right)\log^2\left(\frac{1}{\delta}\right)\right).$$

4. To make sure our algorithm is consistent in its choice of $\widetilde{P}$, we store all S sampled by Algorithm 3, along with the estimates $\widetilde{P}(S)$ computed. This does not dominate the asymptotic space or time complexity.

Proof The proof is almost identical to the proof of Theorem 26, where we substitute Proposition 35 in place of Proposition 24, which also gives the upper bound on σ. We summarize the remaining differences here. First, when using $\mathcal{A}_\infty$, we use the same choice of r_∞ as in Theorem 26, and

$$R_\infty = \Theta\left(r_\infty + k^\star \log^2\left(\frac{d}{\delta}\right)\right)$$

by substituting the bound (41). Second, when using $\mathcal{A}_2$, we similarly use the bound $R_2^2 = O(k^\star \log^2(\frac{d}{\delta}))$ as before. Finally, in the Laplace prior setting, we no longer have an exact characterization for sampling $\boldsymbol{\theta} \sim \pi(\mathbf{X}, \mathbf{y}, S)$, as in Lemma 25. However, using Proposition 31 to perform this sampling to total variation error $\frac{\delta}{2}$ does not dominate the runtime, and it suffices to adjust the failure probability of other steps of the algorithm by a constant factor. ■

For convenience, we provide the following specialization of Theorem 36 in the specific case of Gaussian measurements $\mathbf{X} \in \mathbb{R}^{n\times d}$ with i.i.d. entries $\sim \mathcal{N}(0, \frac{1}{n})$, analogously to Corollary 28.

Corollary 37 *Let $\delta \in (0,1)$, and suppose $\mathbf{X} \in \mathbb{R}^{n\times d}$ has i.i.d. entries $\sim \mathcal{N}(0, \frac{1}{n})$. If*

$$n = \Theta\left(\left(k + \log\left(\frac{1}{\delta}\right)\right)^3 \log^3\left(\frac{d}{\delta}\right)\right),$$

for an appropriate constant, in the setting of Model 1 where $\mu = \mathrm{Lap}(0,1)$ and $\sigma = O(\frac{1}{k+\log(1/\delta)})$, there is an algorithm that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \le \delta$, with probability $\ge 1-\delta$ over the randomness of Model 1 and $\mathbf{X}$. The algorithm runs in time

$$O\left(n^2 d^{1.5} \log\left(\frac{d\log\left(\frac{1}{\delta}\right)}{\sigma}\right) + n^2\sqrt{d}\log\left(\frac{d}{\sigma\delta}\right) + \frac{\left(k+\log(\frac{1}{\delta})\right)^4}{\delta^2}\log^4\left(\frac{\log\left(\frac{d}{\delta}\right)}{\sigma}\right)\log^2\left(\frac{1}{\delta}\right)\right).$$

Alternatively, if

$$n = \Theta\left(\left(k + \log\left(\frac{1}{\delta}\right)\right)^5 \log^2\left(\frac{d}{\delta}\right)\log(d)\right)$$

for an appropriate constant, in the setting of Model 1 where $\mu = \mathrm{Lap}(0,1)$ and $\sigma = O(\frac{1}{k+\log(1/\delta)})$, there is an algorithm that returns $\boldsymbol{\theta} \sim \pi'$ for a distribution π' satisfying $D_{\mathrm{TV}}(\pi', \pi(\cdot \mid \mathbf{X}, \mathbf{y})) \le \delta$, with probability $\ge 1-\delta$ over the randomness of Model 1 and $\mathbf{X}$. The algorithm runs in time

$$O\left(nd\log\left(\frac{\log\left(\frac{d}{\delta}\right)}{\sigma}\right) + \frac{\left(k+\log(\frac{1}{\delta})\right)^4}{\delta^2}\log^4\left(\frac{\log\left(\frac{d}{\delta}\right)}{\sigma}\right)\log^2\left(\frac{1}{\delta}\right)\right).$$

Appendix D. Helper results on sparse recovery

D.1. Sparse recovery preliminaries

Concentration. Here we state standard bounds on Gaussian random variables.

Fact 2 (Mill's inequality) *If $Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable and $t > 0$,*

$$\mathbb{P}\left[|Z| > t\right] \le \sqrt{\frac{2}{\pi}} \cdot \frac{1}{t}\exp\left(-\frac{t^2}{2}\right).$$

Fact 3 (χ^2 tail bounds, Lemma 1, Laurent and Massart (2000)) *Let $\{Z_i\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and $\mathbf{a} \in \mathbb{R}^n_{\geq 0}$. Then,*

$$\mathbb{P}\left[\sum_{i\in[n]} \mathbf{a}_i Z_i^2 - \|\mathbf{a}\|_2^2 \geq 2\|\mathbf{a}\|_2 \sqrt{t} + 2\|\mathbf{a}\|_\infty t\right] \leq \exp(-t),$$

$$\mathbb{P}\left[\sum_{i\in[n]} \mathbf{a}_i Z_i^2 - \|\mathbf{a}\|_2^2 \leq -2\|\mathbf{a}\|_2 \sqrt{t}\right] \leq \exp(-t).$$

As a consequence of Facts 2 and 3 we have the following.

Lemma 38 *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ satisfy $(\epsilon, 1)$-RIP and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$. Then*

$$\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty \leq \sigma(1+\epsilon)\sqrt{2\log\left(\frac{d}{\delta}\right)}, \text{ with probability } \geq 1-\delta, \text{ for all } \delta \in (0,1).$$

Proof We will show that for all $j \in [d]$, $|\mathbf{X}_{:j}^\top \boldsymbol{\xi}| \leq$ with probability $\geq 1 - \frac{\delta}{d}$. Since $\mathbf{X}$ satisfies $(\epsilon, 1)$-RIP, for any $j \in [d]$, we have $\|\mathbf{X}\mathbf{e}_j\|_2^2 = \|\mathbf{X}_{:j}\|_2 \leq 1+\epsilon$. Moreover, $\mathbf{X}_{:j}^\top \boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \|\mathbf{X}_{:j}\|_2^2)$, so

$$\mathbb{P}\left[|\mathbf{X}_{:j}^\top \boldsymbol{\xi}| > \sigma(1+\epsilon)\sqrt{2\log\left(\frac{d}{\delta}\right)}\right] \leq \frac{\delta}{d},$$

by Fact 1. The conclusion follows by taking a union bound over all columns $j \in [d]$. ■

Isometric conditions. We require another notion of isometry from the sparse recovery literature, used in proving Proposition 15.

Definition 39 (Mutual incoherence) *We say $\mathbf{X} \in \mathbb{R}^{n\times d}$ satisfies the mutual incoherence (MI) condition over $S \subseteq [d]$ with parameter $\alpha \in [0,1)$, or $\mathbf{X}$ is α-MI over S, if*

$$\max_{j\in S^c} \left\|[\mathbf{X}^\top \mathbf{X}]_{S\times S}^{-1} \mathbf{X}_{S:}^\top \mathbf{X}_{:j}\right\|_1 \leq \alpha.$$

We next show how Definitions 7 and 39 relate.

Lemma 40 (RIP implies MI) *For any $\mathbf{X} \in \mathbb{R}^{n\times d}$, if $\mathbf{X}$ is $(\epsilon, s+1)$-RIP, then $\mathbf{X}$ is α-MI over S for any $S \subseteq [d]$ with $|S| \leq s$, and $\alpha = \sqrt{2s\epsilon/(1-\epsilon)}$.*

Proof Fix $S \subseteq [d]$ such that $|S| \leq k$, and fix $j \in S^c$. We define a projection operator:

$$\mathbf{P}_S := \mathbf{X}_{:S}[\mathbf{X}^\top \mathbf{X}]_{S\times S}^{-1} \mathbf{X}_{S:}^\top.$$

It is well known that $\mathbf{P}_S$ is the projection matrix of the linear subspace: $\text{span}\{\mathbf{X}_{:j} \in \mathbb{R}^n \mid j \in S\}$. Let $T := S \cup j$. Since $\mathbf{X}$ satisfies $(\epsilon, s+1)$-RIP, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, we have

$$\|\mathbf{X}_{:T}\boldsymbol{\theta}_T\|_2^2 = \|\mathbf{X}_{:S}\boldsymbol{\theta}_S + \mathbf{X}_{:j}\boldsymbol{\theta}_j\|_2^2 \geq (1-\epsilon)\|\boldsymbol{\theta}_T\|_2^2 = (1-\epsilon)(\|\boldsymbol{\theta}_S\|_2^2 + \boldsymbol{\theta}_j^2). \tag{42}$$

By the properties of projection operators, we have

$$\|\mathbf{X}_{:j}\|_2^2 = \|\mathbf{P}_S\mathbf{X}_{:j}\|_2^2 + \|(\mathbf{I}-\mathbf{P}_S)\mathbf{X}_{:j}\|_2^2\,, \tag{43}$$

$$\|(\mathbf{I}-\mathbf{P}_S)\mathbf{X}_{:j}\|_2^2 = \min_{\boldsymbol{\theta}_S\in\mathbb{R}^S} \|\mathbf{X}_{:S}\boldsymbol{\theta}_S + \mathbf{X}_{:j}\|_2^2 \geq 1-\epsilon, \tag{44}$$

where the inequality in (44) follows from (42) with $\boldsymbol{\theta}_j = 1$ and $\|\boldsymbol{\theta}_S\|_2 \geq 0$. Inserting (44) into (43),

$$\|\mathbf{P}_S\mathbf{X}_{:j}\|_2^2 \leq \|\mathbf{X}_{:j}\|_2^2 - (1-\epsilon).$$

Next, let $\mathbf{x}_j := [\mathbf{X}^\top\mathbf{X}]^{-1}_{S\times S}\mathbf{X}^\top_{S:}\mathbf{X}_{:j} \in \mathbb{R}^S$. Then applying RIP again, we have

$$\|\mathbf{P}_S\mathbf{X}_{:j}\|_2^2 = \|\mathbf{X}_{:S}\mathbf{x}_j\|_2^2 \geq (1-\epsilon)\,\|\mathbf{x}_j\|_2^2\,.$$

Thus, we have by combining the above two displays that

$$\|\mathbf{x}_j\|_2^2 \leq \frac{\|\mathbf{X}_{:j}\|_2^2}{1-\epsilon} - 1.$$

Moreover, notice that $\mathbf{x}_j \in \mathbb{R}^S$, so the above display implies

$$\|\mathbf{x}_j\|_1 \leq \sqrt{s}\,\|\mathbf{x}_j\|_2 \leq \sqrt{s}\cdot\sqrt{\frac{\|\mathbf{X}_{:j}\|_2^2}{1-\epsilon} - 1}.$$

Finally, by RIP applied with the 1-sparse test vector $\mathbf{e}_j$,

$$\|\mathbf{X}\mathbf{e}_j\|_2^2 = \|\mathbf{X}_{:j}\|_2^2 \in [1-\epsilon, 1+\epsilon] \implies \|\mathbf{x}_j\|_1 \leq \sqrt{\frac{2s\epsilon}{1-\epsilon}}.$$

Because our choices of S and $j \in S^c$ were arbitrary, we have the desired MI parameter bound. ■

D.2. Solving the Lasso to high precision

Here, we consider the (Lagrangian) Lasso problem, a.k.a. ℓ_1-penalized linear regression, defined as:

$$\hat{\boldsymbol{\theta}} := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2}\,\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda\,\|\boldsymbol{\theta}\|_1\,, \tag{45}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^n$ and $\lambda > 0$. We provide a polynomial-time algorithm achieving ℓ_∞-norm convergence guarantees to the optimal solution. The reason this is nontrivial using off-the-shelf convex optimization tools, e.g., cutting-plane methods, is because (45) is not strongly convex, so function approximation guarantees do not transfer to distance bounds.

Throughout the section, we make the following assumption.

Assumption 4 *The columns of* $\mathbf{X}$ *are in general position.*

We now restate a result from Tibshirani (2013) that shows Assumption 4 implies uniqueness of the Lasso solution (45), and characterizes the solution. The former fact is attributed to Osborne et al. (2000).

Lemma 41 (Lemma 4, Tibshirani (2013)) *Under Assumption 4, (45) has a unique solution, $\hat{\boldsymbol{\theta}}$, satisfying*

$$\hat{\boldsymbol{\theta}}_S = [\mathbf{X}^\top \mathbf{X}]^{-1}_{S\times S}(\mathbf{X}^\top_{S:}\mathbf{y} - \lambda \mathbf{s}), \quad \hat{\boldsymbol{\theta}}_{S^C} = \mathbf{0}, \quad \mathbf{s} := \operatorname{sign}(\mathbf{X}^\top_{S:}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})),$$

where $S := \{i \in [d] \mid |\mathbf{X}^\top_{i:}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}})| = \lambda\}$.

We next derive a constrained dual problem for (45).

Lemma 42 (Lasso dual) *The dual problem for (45) is given as*

$$\hat{\mathbf{z}} = \underset{\|\mathbf{z}\|_\infty \le 1}{\arg\min} \ (\mathbf{z} - \mathbf{a})^\top (\mathbf{X}^\top \mathbf{X})^\dagger (\mathbf{z} - \mathbf{a}), \quad \mathbf{a} := \frac{1}{\lambda}\mathbf{X}^\top \mathbf{y}. \tag{46}$$

Furthermore, if Assumption 4 holds, then (46) has a unique minimizer satisfying

$$\hat{\boldsymbol{\theta}}_S = [\mathbf{X}^\top \mathbf{X}]^{-1}_{S\times S}(\mathbf{X}^\top_{S:}\mathbf{y} - \lambda \hat{\mathbf{z}}_S), \quad \hat{\boldsymbol{\theta}}_{S^c} = \mathbf{0}, \quad \hat{\mathbf{z}} = \mathbf{a} - \frac{1}{\lambda}\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\theta}}$$

where $\hat{\boldsymbol{\theta}}$ is the unique minimizer of (45) and $S = \{i \in [d] \mid |\hat{\mathbf{z}}_i| = 1\}$.

Proof First, by the standard identity $\|\boldsymbol{\theta}\|_1 = \max_{\|\mathbf{z}\|_\infty \le 1} \mathbf{z}^\top \boldsymbol{\theta}$, (45) is equivalently

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \max_{\|\mathbf{z}\|_\infty \le 1} \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \mathbf{z}^\top \boldsymbol{\theta}. \tag{47}$$

Note that the set $\|\mathbf{z}\|_\infty \le 1$ is a compact set, and $f(\boldsymbol{\theta}, z) := \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \mathbf{z}^\top \boldsymbol{\theta}$ is convex-concave. Therefore, using Sion's minimax theorem, Sion (1958), strong duality holds so (47) equals

$$\max_{\|\mathbf{z}\|_\infty \le 1} \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \mathbf{z}^\top \boldsymbol{\theta}. \tag{48}$$

Solving the inner minimization shows that for fixed $\mathbf{z}$, we should choose

$$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^\dagger (\mathbf{X}^\top \mathbf{y} - \lambda \mathbf{z}).$$

Substituting this back into (48) we have for the optimal solution, $\hat{\mathbf{z}}$, of the dual (48), that

$$\begin{aligned}
\hat{\mathbf{z}} &= \underset{\|\mathbf{z}\|_\infty \le 1}{\arg\max} \ -\frac{\lambda^2}{2}(\mathbf{z} - \mathbf{a})^\top (\mathbf{X}^\top \mathbf{X})^\dagger (\mathbf{z} - \mathbf{a}) + \frac{\lambda^2}{2}\mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^\dagger \mathbf{a} \\
&\quad + \frac{1}{2}\left\| \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\dagger \mathbf{X}^\top\right)\mathbf{y} \right\|_2^2 \\
&= \underset{\|\mathbf{z}\|_\infty \le 1}{\arg\max} \ -\frac{\lambda^2}{2}(\mathbf{z} - \mathbf{a})^\top (\mathbf{X}^\top \mathbf{X})^\dagger (\mathbf{z} - \mathbf{a}) = \underset{\|\mathbf{z}\|_\infty \le 1}{\arg\min} \ (\mathbf{z} - \mathbf{a})^\top (\mathbf{X}^\top \mathbf{X})^\dagger (\mathbf{z} - \mathbf{a}).
\end{aligned}$$

This proves the first conclusion. For the second conclusion, since strong duality holds, by using KKT conditions, we have for primal and dual optimal solutions $\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}$ respectively,

$$\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}^\top \mathbf{y} - \lambda \hat{\mathbf{z}}, \quad \|\hat{\mathbf{z}}\|_\infty \le 1.$$

Note that $\hat{\mathbf{z}}$ is fixed given $\hat{\boldsymbol{\theta}}$. Under Assumption 4, we conclude uniqueness of $\hat{\mathbf{z}}$. Next, we have

$$\hat{\mathbf{z}}_i = \frac{1}{\lambda}\mathbf{X}_i^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) \implies \max_{i\in[d]} |\mathbf{X}_i^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})| \leq \lambda.$$

Let $S := \{i \in [d] : |\hat{\mathbf{z}}_i| = 1\}$, which by the above is consistent with the definition in Lemma 41. We have the remaining conclusions due to Lemma 41. ■

Our strategy is now to show that restricted to $\mathrm{range}(\mathbf{X}^\top)$, the dual problem (46) is strongly convex, and consequently, we can suboptimality guarantees to distance bounds on the parameter $\mathbf{z}$.

Lemma 43 *Let $\mathbf{X}$ satisfy Assumption 4 with $n \leq d$, with singular value decomposition*

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \mathbf{U} \in \mathbb{R}^{n\times n}, \boldsymbol{\Sigma} = \mathbf{diag}\,(\boldsymbol{\sigma}) \text{ for } \boldsymbol{\sigma} \in \mathbb{R}^n_{>0},\ \mathbf{V} \in \mathbb{R}^{d\times n}.$$

Consider the following optimization problem:

$$\hat{\mathbf{w}} = \operatorname*{arg\,min}_{\|\mathbf{V}\mathbf{w}\|_\infty \leq 1} (\mathbf{w} - \mathbf{b})^\top \boldsymbol{\Sigma}^{-2}(\mathbf{w} - \mathbf{b}),\ \ \mathbf{V}\mathbf{b} := \frac{1}{\lambda}\mathbf{X}^\top\mathbf{y}. \tag{49}$$

Then, $\hat{\mathbf{z}} = \mathbf{V}\hat{\mathbf{w}}$, where $\hat{\mathbf{z}}$ is the unique optimal dual solution as defined in Lemma 42.

Proof By Lemma 42, the optimal solution to (46) is of the form

$$\hat{\mathbf{z}} = \mathbf{a} - \frac{1}{\lambda}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\theta}} = \frac{1}{\lambda}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})$$

where $\mathbf{a} := \frac{1}{\lambda}\mathbf{X}^\top\mathbf{y}$. Therefore, $\hat{\mathbf{z}} \in \mathrm{range}\left(\mathbf{X}^\top\right)$ and $\exists \mathbf{w} \in \mathbb{R}^n$ such that $\hat{\mathbf{z}} = \mathbf{V}\hat{\mathbf{w}}$. Similarly, $\mathbf{a} \in \mathrm{range}\left(\mathbf{X}^\top\right)$. Therefore, $\exists \mathbf{b} \in \mathbb{R}^n$ such that $\mathbf{V}\mathbf{b} = \frac{1}{\lambda}\mathbf{X}^\top\mathbf{y}$.

Thus, if we were to restrict the dual problem to solutions of the form

$$\mathbf{z} = \mathbf{V}\mathbf{w} \tag{50}$$

then we would restrict our constraint set to a subset of the original set in (46), $\{\mathbf{z} \mid \|\mathbf{z}\|_\infty \leq 1\}$. However, since $\hat{\mathbf{z}} = \mathbf{V}\hat{\mathbf{w}}$ holds, this restriction does not affect the problem. The uniqueness of $\hat{\mathbf{w}}$ implies the uniqueness of $\hat{\mathbf{z}}$. The conclusion of the lemma follows by using (50). ■

Lemma 43 shows that (49), our reformulation of (46), is strongly convex, with a convex constraint set. Consequently, a standard convex optimization algorithm can now be used to achieve function approximation guarantees, which we formalize, using a result from Nesterov and Nemirovski (1994), restated below.

Lemma 44 (Eq. (8.1.5), Nesterov and Nemirovski (1994)) *Consider the following optimization problem:*

$$\min \psi(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{c}^\top\mathbf{x},\ \textit{subject to}\ \mathbf{x} \in \mathbb{R}^n, -\mathbf{c}_i^\top\mathbf{x} + \mathbf{r}_i \geq 0\ \textit{for all}\ i \in [d],$$

where $\mathbf{A} \in \mathbb{S}^{n\times n}_{\succeq \mathbf{0}}$ *and* $\{\mathbf{c}_i\}_{i\in[d]} \cup \{\mathbf{c}\} \subset \mathbb{R}^n$*, and* $\mathbf{r} \in \mathbb{R}^d$*. Define the sets*

$$\mathcal{G} := \left\{\mathbf{x} \mid \mathbf{c}_i^\top \mathbf{x} + \mathbf{r}_i \geq 0 \text{ for all } i \in [d]\right\},\ \mathcal{G}' := \left\{\mathbf{x} \mid \mathbf{c}_i^\top \mathbf{x} + \mathbf{r}_i > 0 \text{ for all } i \in [d]\right\}.$$

If $\mathcal{G}'$ *is non-empty, then for any* $\epsilon > 0$*, there is an algorithm which returns* $\mathbf{x}_\epsilon \in \mathcal{G}'$ *satisfying*

$$\psi(\mathbf{x}_\alpha) - \min_{\mathbf{x}\in\mathcal{G}} \psi(\mathbf{x}) \leq \epsilon \left(\max_{\mathbf{x}\in\mathcal{G}} \psi(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{G}} \psi(\mathbf{x})\right),$$

initialized at $\mathbf{x}_0 \in \mathcal{G}$*, and runs in time*

$$O\left(d^{1.5}n^2 \log\left(\frac{d}{\alpha\left(G, \mathbf{x}_0\right)\epsilon}\right)\right)$$

where $\alpha\left(\mathcal{G}, \mathbf{x}_0\right) := \max\left\{\alpha \mid \mathbf{x}_0 + \alpha(\mathbf{x}_0 - \mathcal{G}) \subset \mathcal{G}\right\}$.

We next bound the asymmetry coefficient $\alpha(\mathcal{G}, \mathbf{x}_0)$ for our quadratic program (49).

Lemma 45 *Let* $\mathcal{G} := \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{V}\mathbf{w}\|_\infty \leq 1\}$ *where* $\mathbf{V} \in \mathbb{R}^{d\times n}$ *satisfies* $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$*. Then*

$$\alpha(\mathcal{G}, \mathbf{0}_n) := \max\left\{\alpha \mid \mathbf{0}_n + \alpha(\mathbf{0}_n - \mathcal{G}) \subset \mathcal{G}\right\} \geq \frac{1}{2}.$$

Proof Since $\mathcal{G}$ is symmetric about the origin and nonempty, $-\frac{1}{2}\mathcal{G} = \frac{1}{2}\mathcal{G} \subset \mathcal{G}$ as claimed. ■

Finally, we combine these results to provide a fast algorithm for approximating (45).

Proposition 46 *Let Assumption 4 hold. Following the notation in* (45)*, there is an algorithm which returns* $\boldsymbol{\theta}$ *satisfying* $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \leq \epsilon$*, in time*

$$O\left(d^{1.5}n^2 \log\left(\frac{(\lambda + \left\|\mathbf{X}^\top\mathbf{y}\right\|_2)d}{\epsilon} \cdot \frac{\boldsymbol{\sigma}_1(\mathbf{X})}{\boldsymbol{\sigma}_n(\mathbf{X})}\right)\right).$$

Proof We first claim it suffices to solve (49) to function error

$$\Delta := \frac{\epsilon^2}{\lambda^2} \cdot \frac{\boldsymbol{\sigma}_n(\mathbf{X})^4}{\boldsymbol{\sigma}_1(\mathbf{X})^2}.$$

To see this, let $\mathbf{w}$ satisfy $f(\mathbf{w}) - f(\hat{\mathbf{w}}) \leq \Delta$, where $\hat{\mathbf{w}}$ minimizes (49). By strong convexity of (49),

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_2^2 \leq \boldsymbol{\sigma}_1(\boldsymbol{\Sigma})^2\Delta = \boldsymbol{\sigma}_1(\mathbf{M})^2\Delta.$$

Moreover, letting $\mathbf{z} := \mathbf{V}\mathbf{w}$ and $\hat{\mathbf{z}} := \mathbf{V}\hat{\mathbf{w}}$, so that $\hat{\mathbf{z}}$ minimizes (46) by Lemma 43, we have $\|\mathbf{w} - \hat{\mathbf{w}}\|_2 = \|\mathbf{z} - \hat{\mathbf{z}}\|_2$ since $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$. Finally, by Lemma 42, letting $\widehat{\boldsymbol{\theta}} := (\mathbf{X}^\top\mathbf{X})^\dagger(\mathbf{X}^\top\mathbf{y} - \lambda\hat{\mathbf{z}})$ optimize (45) and $\boldsymbol{\theta} := (\mathbf{X}^\top\mathbf{X})^\dagger(\mathbf{X}^\top\mathbf{y} - \lambda\mathbf{z})$, we have

$$\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|_2 = \lambda\left\|(\mathbf{X}^\top\mathbf{X})^\dagger(\mathbf{z} - \hat{\mathbf{z}})\right\|_2 \leq \frac{\lambda}{\boldsymbol{\sigma}_n(\mathbf{X})^2}\|\mathbf{z} - \hat{\mathbf{z}}\|_2 = \frac{\lambda}{\boldsymbol{\sigma}_n(\mathbf{X})^2}\|\mathbf{w} - \hat{\mathbf{w}}\|_2.$$

It remains to bound the complexity of achieving Δ function error. Letting $\mathbf{A} := 2\boldsymbol{\Sigma}^{-2}$ and $\mathbf{c} := 2\boldsymbol{\Sigma}^{-2}\mathbf{b}$ in the setting of Lemma 44,

$$\begin{aligned}\max_{\mathbf{w}\in\mathcal{G}} \psi(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{G}} \psi(\mathbf{w}) &\leq \max_{\mathbf{w}\in\mathcal{G}} (\mathbf{w}-\mathbf{b})^\top \boldsymbol{\Sigma}^{-2}(\mathbf{w}-\mathbf{b}) \\ &\leq \frac{1}{\boldsymbol{\sigma}_n(\mathbf{X})^2} \max_{\|\mathbf{V}\mathbf{w}\|_\infty \leq 1} \|\mathbf{w}-\mathbf{b}\|_2^2 \\ &\leq \frac{1}{\boldsymbol{\sigma}_n(\mathbf{X})^2} \cdot \left(\max_{\|\mathbf{V}\mathbf{w}\|_2 \leq \sqrt{d}} 2\|\mathbf{w}\|_2^2 + 2\|\mathbf{b}\|_2^2 \right) \leq \frac{2(d+\|\mathbf{b}\|_2^2)}{\boldsymbol{\sigma}_n(\mathbf{X})^2}.\end{aligned}$$

The conclusion follows from Lemma 44 using $\|\mathbf{b}\|_2 \leq \frac{1}{\lambda}\left\|\mathbf{X}^\top\mathbf{y}\right\|_2$ by our choice of $\mathbf{b} = \frac{1}{\lambda}\mathbf{V}^\top\mathbf{X}^\top\mathbf{y}$. ■

D.3. Proof of Proposition 15: ℓ_∞ recovery

In this section, we prove the first statement in Proposition 15. Our main tool is the following known result on the performance of the Lasso for sparse recovery under MI (Definition 39).

Lemma 47 (Theorem 7.21, Wainwright (2019)) *In the setting of Model 2, let* $\mathrm{supp}(\boldsymbol{\theta}^\star) = S$ *and* $|S| = k$. *If* $\boldsymbol{\lambda}_{\min}([\mathbf{X}^\top\mathbf{X}]_{S\times S}) > 0$ *and* $\mathbf{X}$ *is* α*-MI over* S, *then for any regularization parameter* λ *such that*

$$\lambda \geq \frac{2}{1-\alpha}\left\|\mathbf{X}_{S^c:}^\top \boldsymbol{\Pi}_{S^\perp}(\mathbf{X})\boldsymbol{\xi}\right\|_\infty, \tag{51}$$

where $\boldsymbol{\Pi}_{S^\perp}(\mathbf{X}) := \mathbf{I}_n - \mathbf{X}_{:S}[\mathbf{X}^\top\mathbf{X}]_{S\times S}^{-1}\mathbf{X}_{S:}^\top$, *the Lasso solution*

$$\widehat{\boldsymbol{\theta}} := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{ \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\} \tag{52}$$

satisfies

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_\infty \leq \left\| \left[\mathbf{X}^\top\mathbf{X}\right]_{S\times S}^{-1} \mathbf{X}_{S:}^\top\boldsymbol{\xi} \right\|_\infty + \lambda \max_{i\in S} \sum_{j\in S} \left| \left[\mathbf{X}^\top\mathbf{X}\right]_{ij}^{-1} \right|. \tag{53}$$

We next give a bound on the parameters in (51), (53) under RIP.

Lemma 48 *In the setting of Model 2, let* $\mathrm{supp}(\boldsymbol{\theta}^\star) = S$ *and* $|S| = k$. *If* $\mathbf{X}$ *satisfies* $(\epsilon, k+1)$*-RIP, then for any regularization parameter* λ *such that*

$$\lambda \geq \frac{2(1+\alpha)}{1-\alpha}\left\|\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty, \tag{54}$$

the Lasso solution (52) *satisfies*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_\infty \leq \left(\frac{1}{1-\epsilon} + \frac{2\sqrt{k\epsilon}}{1-\epsilon^2}\right)\left(\left\|\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty + \lambda\right). \tag{55}$$

Proof Since $\mathbf{X}$ is $(\epsilon, k+1)$ RIP, we have by Lemma 40 that $\mathbf{X}$ is α-MI over any $S \subseteq [d]$ with $|S| \leq k$, for $\alpha = \sqrt{2k\epsilon/(1-\epsilon)}$. We begin by bounding the right-hand side of (51). By the definition of $\mathbf{\Pi}_{S^\perp}(\mathbf{X})$ and the triangle inequality, we have that

$$\left\|\mathbf{X}_{S^c:}^\top \mathbf{\Pi}_{S^\perp}(\mathbf{X})\boldsymbol{\xi}\right\|_\infty \leq \left\|\mathbf{X}_{S^c:}^\top \boldsymbol{\xi}\right\|_\infty + \left\|\left(\left[\mathbf{X}^\top\mathbf{X}\right]_{S\times S}^{-1}\mathbf{X}_{S:}^\top\mathbf{X}_{:S^c}\right)^\top \mathbf{X}_{S:}^\top\boldsymbol{\xi}\right\|_\infty. \tag{56}$$

Also by the definition of $\|\cdot\|_\infty$, we have

$$\left\|\left(\left[\mathbf{X}^\top\mathbf{X}\right]_{S\times S}^{-1}\mathbf{X}_{S:}^\top\mathbf{X}_{:S^c}\right)^\top \mathbf{X}_{S:}^\top\boldsymbol{\xi}\right\|_\infty \leq \max_{j\in S^c}\left\|\left[\mathbf{X}^\top\mathbf{X}\right]_{S\times S}^{-1}\mathbf{X}_{S:}^\top\mathbf{X}_{:j}\right\|_1 \cdot \left\|\mathbf{X}_{S:}^\top\boldsymbol{\xi}\right\|_\infty \leq \alpha\left\|\mathbf{X}_{S:}^\top\boldsymbol{\xi}\right\|_\infty, \tag{57}$$

where the last inequality used MI. Combining (56) and (57), we have shown

$$\left\|\mathbf{X}_{S^c\times S^c}^\top \mathbf{\Pi}_{S^\perp}(\mathbf{X})\boldsymbol{\xi}\right\|_\infty \leq (1+\alpha)\left\|\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty,$$

proving that (54) is a sufficient condition for (51) to apply. Next we bound the right-hand side of (53). Let $\mathbf{M} := [\mathbf{X}^\top\mathbf{X}]_{S\times S}^{-1}$. Because $\|\mathbf{M}\|_{\infty\to\infty} = \max_{i\in S}\|\mathbf{M}_{i:}\|_1$, we have from (53) that

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_\infty \leq \|\mathbf{M}\|_{\infty\to\infty}\left(\left\|\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty + \lambda\right), \tag{58}$$

so it is enough to provide a bound on $\|\mathbf{M}\|_{\infty\to\infty}$. Because $\mathbf{X}$ satisfies $(\epsilon, k+1)$-RIP, we have

$$\boldsymbol{\lambda}(\mathbf{M}^{-1}) \in [1-\epsilon, 1+\epsilon]^k \implies \boldsymbol{\lambda}(\mathbf{M}) \in \left[\frac{1}{1+\epsilon}, \frac{1}{1-\epsilon}\right]^k. \tag{59}$$

Thus, because $\|\mathbf{e}_i\|_2 = 1$,

$$\mathbf{M}_{ii} = \mathbf{e}_i^\top\mathbf{M}\mathbf{e}_i \in [\boldsymbol{\lambda}_{\min}(\mathbf{M}), \boldsymbol{\lambda}_{\max}(\mathbf{M})] = \left[\frac{1}{1+\epsilon}, \frac{1}{1-\epsilon}\right]. \tag{60}$$

Moreover, we have that

$$\sum_{\substack{j\in S\\ j\neq i}}\mathbf{M}_{ij}^2 = \|\mathbf{M}\mathbf{e}_i\|_2^2 - \mathbf{M}_{ii}^2 \leq \frac{1}{(1-\epsilon)^2} - \frac{1}{(1+\epsilon)^2} = \frac{4\epsilon}{(1-\epsilon^2)^2},$$

where the inequality used the upper bound in (59) and the lower bound in (60). Finally, since $\mathbf{v} \in \mathbb{R}^S$ has $\|\mathbf{v}\|_1 \leq \sqrt{k}\,\|\mathbf{v}\|_2$, by combining the above display with the upper bound in (60),

$$\|\mathbf{M}_{i:}\|_1 = \mathbf{M}_{ii} + \sum_{\substack{j\in S\\ j\neq i}}|\mathbf{M}_{ij}| \leq \frac{1}{1-\epsilon} + \frac{2\sqrt{k\epsilon}}{1-\epsilon^2}.$$

Finally, plugging the above into (58) yields the desired claim (55). ■

We are now ready to prove the first part of Proposition 15.

Proof [Proof of Proposition 15: ℓ_∞ recovery] Under Assumption 2, we have by Lemma 40 that for

$$\alpha = \sqrt{\frac{2k\epsilon}{1-\epsilon}} \le \sqrt{\frac{2}{3}},\ \lambda = \frac{2(1+\alpha)}{1-\alpha}\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty = \Theta\left(\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty\right),$$

that $\mathbf{X}$ is α-MI, and hence Lemma 48 shows that the solution to (53) satisfies

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_\infty = O\left(\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty\right).$$

Now it suffices to return an estimate of $\widehat{\boldsymbol{\theta}}$ within ℓ_∞ distance $O(\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty)$. To do so, we use Proposition 46, giving the conclusion. ■

D.4. Proof of Proposition 15: ℓ_2 recovery

In this section, we prove the second statement in Proposition 15. Our proof is a simple modification of a fairly standard approach based on projected gradient descent over ℓ_1 balls. Our result is implicit in Kelner et al. (2023), but we give a brief explanation of how to derive it.

Proof [Proof of Proposition 15: ℓ_2 recovery] This result is implicit in the proof of Theorem 4, Kelner et al. (2023), which handles a much more general setting of a *semi-random* RIP observation matrix. The exposition in Kelner et al. (2023) loses several additional logarithmic factors in the runtime due to the need to construct a "noisy step oracle," but in our simpler standard RIP setting, we can use uniform weights over the rows of $\mathbf{X}_{[m]:}$ to skip this step. There is a presentation of this simpler setting in Tian (2024), Theorem 3, in the noiseless case, and it extends to the noisy case in the same way as done in Kelner et al. (2023). In particular, in the noisy case it suffices to apply Theorem 3 in Tian (2024) with $R \leftarrow R_2$ and $r \leftarrow r_2$, i.e., the noise level at which point we can no longer guarantee progress. ■

Appendix E. Deferred proofs from Section 3

Proposition 49 (Generalization of Theorem 3.4, Jalal et al. (2021)) *Let $\mathrm{m}(\cdot,\cdot)$ be an arbitrary metric over $\mathcal{K} \times \mathcal{K}$ and suppose μ is a distribution over $\mathcal{K}$. Let $\theta^\star \sim \mu$, let $\mathcal{F} : \mathcal{K} \to \Omega$ be an arbitrary (possibly randomized) forward operator, and let $\varphi = \mathcal{F}(\theta^\star)$. Suppose there is any (possibly randomized) algorithm $\mathcal{A} : \Omega \to \mathcal{K}$ such that for $\epsilon > 0$, $\delta \in (0,1)$,*

$$\mathbb{P}\left[\mathrm{m}(\widehat{\theta}, \theta^\star) > \epsilon\right] \le \delta, \text{ for } \widehat{\theta} \sim \mathcal{A}(\varphi), \tag{8}$$

where probabilities are taken over the joint distribution of $(\theta^\star, \varphi, \widehat{\theta})$. Then, letting $\theta \sim \mu(\cdot \mid \varphi)$, i.e., the posterior distribution of $\theta^\star$ given observations φ, we have $\mathbb{P}[\mathrm{m}(\theta, \theta^\star) > 2\epsilon] \le 2\delta$.

Proof Let $\mathcal{E}_{\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}}$ denote the event that $\mathrm{m}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}) \le \epsilon$, and similarly define $\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}}$. We have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}^\star,\varphi,\boldsymbol{\theta}}\left[\mathrm{m}(\boldsymbol{\theta},\boldsymbol{\theta}^\star) \le 2\epsilon\right] &\ge \mathbb{P}_{\boldsymbol{\theta}^\star,\varphi,\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}}\left[\mathrm{m}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}) \le \epsilon \wedge \mathrm{m}(\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}) \le \epsilon\right] \\
&= \mathbb{P}_{\boldsymbol{\theta}^\star,\varphi,\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}}\left[\mathcal{E}_{\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}} \wedge \mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}}\right] = \mathbb{E}_{\boldsymbol{\theta}^\star,\varphi,\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}}\left[\mathbb{I}_{\mathcal{E}_{\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}}} \cdot \mathbb{I}_{\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}}}\right] \\
&= \mathbb{E}_{\varphi,\widehat{\boldsymbol{\theta}}}\left[\mathbb{E}_{\boldsymbol{\theta}^\star,\boldsymbol{\theta}}\left[\mathbb{I}_{\mathcal{E}_{\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}}} \cdot \mathbb{I}_{\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right]\right] \\
&= \mathbb{E}_{\varphi,\widehat{\boldsymbol{\theta}}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{I}_{\mathcal{E}_{\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right] \cdot \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\mathbb{I}_{\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right]\right] \\
&= \mathbb{E}_{\varphi,\widehat{\boldsymbol{\theta}}}\left[\mathbb{P}_{\boldsymbol{\theta}}\left[\mathcal{E}_{\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right] \cdot \mathbb{P}_{\boldsymbol{\theta}^\star}\left[\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right]\right] \\
&= \mathbb{E}_{\varphi,\widehat{\boldsymbol{\theta}}}\left[\mathbb{P}_{\boldsymbol{\theta}^\star}\left[\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right]^2\right].
\end{aligned}$$

The first line used that m is a metric and applied the triangle inequality, the second used the definitions of $\mathcal{E}_{\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}}, \mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}}$, the third iterated expectations, the fourth used that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\star$ are independent conditioned on $\varphi, \widehat{\boldsymbol{\theta}}$, and the last used that $(\boldsymbol{\theta}^\star, \varphi, \widehat{\boldsymbol{\theta}})$ and $(\boldsymbol{\theta}, \varphi, \widehat{\boldsymbol{\theta}})$ have the same joint distribution. Finally, the conclusion follows by Jensen's inequality and (8):

$$\begin{aligned}
\mathbb{E}_{\varphi,\widehat{\boldsymbol{\theta}}}\left[\mathbb{P}_{\boldsymbol{\theta}^\star}\left[\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right]^2\right] &\ge \mathbb{E}_{\varphi,\widehat{\boldsymbol{\theta}}}\left[\mathbb{P}_{\boldsymbol{\theta}^\star}\left[\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}} \mid \varphi,\widehat{\boldsymbol{\theta}}\right]\right]^2 \\
&= \mathbb{P}_{\boldsymbol{\theta}^\star,\varphi,\widehat{\boldsymbol{\theta}}}\left[\mathcal{E}_{\boldsymbol{\theta}^\star,\widehat{\boldsymbol{\theta}}}\right] = (1-\delta)^2 \ge 1 - 2\delta.
\end{aligned}$$

■

Lemma 50 *Let (α, β) be random variables, and let $\mathcal{E}_{\alpha,\beta}$ be an event that depends on their realizations. Suppose for some $\delta_1, \delta_2 \in (0,1)$ that $\mathbb{P}_{(\alpha,\beta)}[\mathcal{E}_{\alpha,\beta}] \ge 1 - \delta_1\delta_2$. Then, letting $g(\alpha) := \mathbb{P}_\beta\left[\mathcal{E}_{\alpha,\beta} \mid \alpha\right]$, $\mathbb{P}_\alpha[g(\alpha) \le 1 - \delta_1] \le \delta_2$.*

Proof Let $p := \mathbb{P}_\alpha[g(\alpha) \le 1 - \delta_1]$. By expanding, and using $g(\alpha) \in [0,1]$ for all α,

$$\begin{aligned}
1 - \delta_1\delta_2 &\le \mathbb{P}_{(\alpha,\beta)}\left[\mathcal{E}_{\alpha,\beta}\right] = \mathbb{E}_\alpha\left[g(\alpha)\right] \\
&= p\mathbb{E}_\alpha\left[g(\alpha) \mid g(\alpha) \le 1 - \delta_1\right] + (1-p)\mathbb{E}_\alpha\left[g(\alpha) \mid g(\alpha) > 1 - \delta_1\right] \\
&\le p(1-\delta_1) + 1 - p = 1 - p\delta_1.
\end{aligned}$$

Rearranging shows that $p \le \delta_2$ as claimed. ■

We next restate (a special case of) Lemma 12 from Lee et al. (2021), which we use to prove Lemma 9.

Lemma 51 (Lemma 12, Lee et al. (2021)) *Let π, μ be distributions over the same domain, and suppose that $\pi \propto P$ and $\mu \propto Q$ for unnormalized densities P, Q. Moreover, suppose for $\epsilon \in (0,1)$ and $C \ge 1$, there is a set Ω such that $\mathbb{P}_{\omega\sim\pi}[\omega \in \Omega] \ge 1 - \epsilon$, and*

$$\frac{P(\omega)}{Q(\omega)} \le C \text{ for all } \omega \in \Omega, \ \frac{\int Q(\omega)\mathrm{d}\omega}{\int P(\omega)\mathrm{d}\omega} \le 1.$$

There is an algorithm that, for any $\delta \in (0,1)$ outputs a sample within total variation distance $\epsilon + \delta$ from π. The algorithm uses $O(\frac{C}{1-\epsilon}\log(\frac{1}{\delta}))$ samples from μ, and evaluates $\frac{P(\omega)}{Q(\omega)}$ $O(\frac{C}{1-\epsilon}\log(\frac{1}{\delta}))$ times.

Now we present the proof of Lemma 9.

Lemma 52 *Let π, μ be distributions over the same domain Ω, and suppose $\pi \propto P$ and $\mu \propto Q$ for unnormalized densities P, Q. Moreover, suppose for all $\omega \in \Omega$, $\frac{P(\omega)}{Q(\omega)} \le C$, $\frac{Q(\omega)}{P(\omega)} \le C$. There is an algorithm* RejectionSample$(\mu, P, Q, \Omega, C, \delta)$ *that, for any $\delta \in (0,1)$, outputs a sample within total variation δ from π, using $O(C^2\log(\frac{1}{\delta}))$ samples from μ, and evaluating $\frac{P(\omega)}{Q(\omega)}$ $O(C^2\log(\frac{1}{\delta}))$ times.*

Proof We apply Lemma 51 with $\epsilon = 0$, i.e., with Ω as the whole domain. For all ω, let $Q'(\omega) := \frac{Q(\omega)}{C}$. Then, we have the result by applying Lemma 51 with $P \leftarrow P$ and $Q \leftarrow Q'$, since

$$\frac{P(\omega)}{Q'(\omega)} = \frac{CP(\omega)}{Q(\omega)} \le C^2, \ \int Q'(\omega)\mathrm{d}\omega = \frac{1}{C}\int Q(\omega)\mathrm{d}\omega \le \int P(\omega)\mathrm{d}\omega.$$

■

Lemma 53 *Let $\mathbf{p} \in [0,1]^d$, and let $\pi := \bigotimes_{i\in[d]} \mathrm{Bern}(\mathbf{p}_i)$ be a product distribution over $\{0,1\}^d$, identified with sets $S \subseteq [d]$. Define $\Omega_k := \{S \subseteq [d] \mid |S| \le k\}$.* ConditionalPoisson *(Algorithm 5) runs in time $O(dk)$ and returns $S \in \Omega_k$ such that $\mathbb{P}[S = T] = \pi(T \mid T \in \Omega_k)$.*

Proof The algorithm is displayed in Algorithm 5, and its runtime is clearly $O(dk)$. We now prove correctness. First, observe that for all $S \subseteq [d]$,

$$\pi(S) \propto \prod_{i\in S} \frac{\mathbf{p}_i}{1-\mathbf{p}_i} =: P(S).$$

We claim that after Lines 5 to 5 have finished running, we have for all $i \in [d]$, $j \in [k]$, that

$$F(i,j) = \sum_{\substack{S\subseteq\{i,i+1,\ldots,d\}\\|S|=j}} P(S). \tag{61}$$

In (61), we define empty sums to be 0 and $P(\emptyset) := 1$. Note that $F(d+1,j)$ for $j \in [0,k]$ satisfies (61). Now, supposing all entries $F(i+1,j)$ for $j \in [0,k]$ satisfy (61), we have for any $j \in [0,k]$ that

$$\begin{aligned}\sum_{\substack{S\subseteq\{i,i+1,\ldots,d\}\\|S|=j}} P(S) &= \sum_{\substack{S\subseteq\{i+1,i+2,\ldots,d\}\\|S|=j}} P(S) + \frac{\mathbf{p}_i}{1-\mathbf{p}_i}\sum_{\substack{S\subseteq\{i+1,i+2,\ldots,d\}\\|S|=j-1}} P(S)\\ &= F(i+1,j) + \frac{\mathbf{p}_i}{1-\mathbf{p}_i}F(i+1,j-1) = F(i,j),\end{aligned}$$

as computed by Lines 5 to 5. Thus, for any $S \in \Omega_k$, we have

$$\pi(S \mid |S| \le k) = \frac{P(S)}{\sum_{S||S|\le k} P(S)} = \frac{P(S)}{\sum_{j\in[0,k]} F(1,j)}.$$

In Algorithm 5, we first sample ℓ with probability $\frac{F(1,\ell)}{\sum_{j\in[k]} F(1,\ell)}$. Let $S = \{i_1, i_2, \ldots, i_\ell\}$ have $|S| = \ell$. Algorithm 5 outputs S with probability:

$$\frac{F(2,\ell)}{F(1,\ell)}\frac{F(3,\ell)}{F(2,\ell)}\cdots\frac{\mathbf{p}_{i_1}}{1-\mathbf{p}_{i_1}}\frac{F(i_1+1,\ell-1)}{F(i_1,\ell)}\cdots\frac{\mathbf{p}_{i_l}}{1-\mathbf{p}_{i_l}}\frac{F(i_l+1,0)}{F(i_l,1)}\frac{F(d+1,0)}{F(d,0)} = \frac{\prod_{i\in S}\frac{\mathbf{p}_i}{1-\mathbf{p}_i}}{F(1,\ell)}$$
$$= \frac{P(S)}{F(1,\ell)}.$$

Thus, we have

$$\mathbb{P}(\text{output } S) = \frac{P(S)}{F(1,\ell)}\cdot\frac{F(1,\ell)}{\sum_{j\in[0,k]}F(1,j)} = \pi(S \mid |S|\le k).$$

■

Algorithm 5 ConditionalPoisson$(\mathbf{p}, k)$

Input: $\mathbf{p} \in [0,1]^d$ inducing $\pi := \bigotimes_{i \in [d]} \mathrm{Bern}(\mathbf{p}_i)$, maximum allowed successes $k \in [d]$
Output: $S \in \Omega_k := \{S \subseteq [d] \mid |S| \leq k\}$ distributed as $\pi(S \mid S \in \Omega_k)$
for $j \in [0, k]$ **do**
  // Precompute dynamic programming table.
  $F(d+1, j) \leftarrow \mathbb{I}_{j=0}$
end
for $i = d$ **to** 1 **do**
  for $j \in [0, k]$ **do**
    if $j = 0$ **then**
      $F(i, j) \leftarrow 1$
    else
      $F(i, j) \leftarrow F(i+1, j) + \frac{\mathbf{p}_i}{1-\mathbf{p}_i} F(i+1, j-1)$
    end
  end
end
Sample $\ell \in [0, k]$ proportionally to $F(1, \ell)$ // Sample target number of successes.
$S \leftarrow \emptyset, r \leftarrow \ell$; // r is the number of successes remaining to be allocated.

for $i = 1$ **to** d **do**
  // Sequentially select indices.
  if $r = 0$ **then**
    break; // No more successes to allocate.
  end
  Let $\alpha \sim \mathrm{Bern}(\frac{\mathbf{p}_i F(i+1, r-1)}{(1-\mathbf{p}_i) F(i, r)})$
  if $\alpha = 1$ **then**
    $S \leftarrow S \cup \{i\}$
    $r \leftarrow r - 1$
  end
end
return S