# A Recovery Theory for Diffusion Priors: Deterministic Analysis of the Implicit Prior Algorithm

**Oscar Leong**
University of California, Los Angeles

**Yann Traonmilin**
University of Bordeaux

## Abstract

Recovering high-dimensional signals from corrupted measurements is a central challenge in inverse problems. Recent advances in generative diffusion models have shown remarkable empirical success in providing strong data-driven priors, but rigorous recovery guarantees remain limited. In this work, we develop a theoretical framework for analyzing deterministic diffusion-based algorithms for inverse problems, focusing on a deterministic version of the algorithm proposed by Kadkhodaie and Simoncelli [2021]. First, we show that when the underlying data distribution concentrates on a low-dimensional model set, the associated noise-convolved scores can be interpreted as time-varying projections onto such a set. This leads to interpreting previous algorithms using diffusion priors for inverse problems as generalized projected gradient descent methods with varying projections. When the sensing matrix satisfies a restricted isometry property over the model set, we can derive quantitative convergence rates that depend explicitly on the noise schedule. We apply our framework to two instructive data distributions: uniform distributions over low-dimensional compact, convex sets and low-rank Gaussian mixture models. In the latter setting, we can establish global convergence guarantees despite the nonconvexity of the underlying model set.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

## 1 INTRODUCTION

In this work, we consider solving linear inverse problems of the form

$$y = A\hat{x}$$

where $A \in \mathbb{R}^{m\times d}$ is a linear operator and $\hat{x} \in \mathbb{R}^d$ is the underlying signal of interest. In many settings arising in the physical sciences, the above problem is ill-posed due to the compressive nature of the linear operator $A$, meaning that there are infinitely many signals that fit the given measurements. Hence one requires additional prior knowledge or regularizers to overcome the inherent ill-posedness.

Traditionally, hand-crafted priors based on notions such as sparsity [Daubechies et al., 2004, Candès et al., 2006, Donoho, 2006, Tibshirani, 1994], low-rankness [Fazel, 2002, Candès and Recht, 2009, Recht et al., 2010], or smoothness [Rudin et al., 1992] have been widely integrated into variational approaches in solving inverse problems. With the development of modern machine learning techniques, we have seen a surge in more powerful, data-driven priors with strong empirical success. These include learning-based regularization penalties parameterized by deep neural networks [Lunz et al., 2018, Mukherjee et al., 2021, Shumaylov et al., Goujon et al., 2023] and methods based on plug-and-play denoisers [Venkatakrishnan et al., 2013, Romano et al., 2017].

There has been a surge of interest in the use of generative models as priors in inverse problems due to their ability to generate realistic samples from a variety of high-dimensional data distributions. Early works along these lines have exploited GANs/VAEs [Bora et al., 2017, Hand et al., 2018, 2024] and normalizing flows [Asim et al., 2020, Ardizzone et al., 2019]. More recently, we have seen significant progress in natural image synthesis through the advent of diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020, Song and Ermon, 2019, Song et al., 2021, Nichol and Dhariwal, 2021] and flow-matching models [Liu et al., 2023, Lipman et al., 2022]. Diffusion models, for example,

generate images by iteratively denoising noise samples with an iteration-dependent noise schedule. They are justified by an explicit relationship between the score of the noise-convolved data distribution and the minimum-mean-square-error (MMSE) denoising solution [Miyasawa et al., 1961, Vincent, 2011]. Once learned, these image priors have been successfully integrated into approaches for solving inverse problems [Chung et al., 2023, 2022, Feng et al., 2023, Kadkhodaie and Simoncelli, 2021, Wang et al., 2023, Zhang et al., 2024, Martin et al., 2025, Song et al., 2022, Sun et al., 2024, Xu and Chi, 2024, Song et al., 2023].

While the empirical performance of diffusion models for inverse problems is compelling, a rigorous recovery theory remains largely absent. Classical regularization theory provides quantitative guarantees on recovery error, stability, and robustness when using convex penalties or sparsity-inducing norms. These guarantees rely on an explicit low-dimensional model set $\Sigma \subset \mathbb{R}^d$ (e.g., a set of sparse vectors) that captures important structural properties of $\hat{x}$. By contrast, existing theoretical results for generative priors have primarily addressed representation error or sample complexity for GANs, VAEs, and normalizing flows, and most analyses of diffusion priors have focused on posterior sampling guarantees [Coeurdoux et al., 2024, Xu and Chi, 2024, Sun et al., 2024] or restricted data models such as linear subspaces [Rout et al., 2023b,a]. As a result, we lack a systematic understanding of how diffusion priors enable recovery of ground-truth signals from compressive measurements. In particular, it remains unclear which classes of measurement operators permit recovery, how many observations are needed, or how the choice of noise schedule impacts convergence. This work develops a framework that addresses these questions.

A central difficulty is that diffusion priors are defined implicitly through denoisers that approximate the score of noise-convolved distributions, rather than through explicit convex penalties or geometric constraints. We show in this work that *these denoisers can be viewed as approximate projection operators onto low-dimensional model sets*, which provides a direct connection between diffusion-based updates and generalized projected gradient methods. This perspective enables the analysis of recovery properties of diffusion priors, with guarantees that depend explicitly on the underlying data distribution and the noise schedule.

**Contributions.** This work develops a principled recovery theory for diffusion priors in linear inverse problems. Our main contributions are:

- We show that for natural data distributions, noise-convolved scores associated with diffusion models can be interpreted as *approximate projections* onto a low-dimensional model set $\Sigma$. Moreover, as the noise decreases, these approximate projections converge to a projection operator onto $\Sigma$.
- To exploit this perspective, we provide an in-depth analysis of a deterministic version of the sampling algorithm of Kadkhodaie and Simoncelli [2021] for signal recovery from linear compressive measurements. We show that this algorithm naturally connects diffusion priors directly to generalized projected gradient descent methods.
- For the case of low-rank Gaussian mixture models (LR-GMMs), we prove that the limiting score recovers the true projection operator. Under a restricted isometry condition on the operator $A$, we establish global convergence of the iterates to the true signal. We show that, after a *burn-in* period, the iterates converge at a rate that is a combination of a term exhibiting geometric decay and another term that depends on the noise schedule.
- We extend our analysis to other data distributions, including random sparse vectors and uniform distributions supported on convex bodies, where we prove that the limiting denoiser converges to the metric projection and quantify the approximation error.

## 2 BACKGROUND AND RELATED WORK

**Diffusion priors in inverse problems.** Diffusion models have demonstrated remarkable empirical success in sampling from high-dimensional natural image distributions, which has led to their widespread use as unsupervised priors in inverse problems. Our focus here is not on the theory of generative modeling itself, but on theoretical results for their use as priors in inverse problems. A number of methods have been proposed; for a complete overview, see the survey [Daras et al., 2024]. Among the most influential is the Diffusion Posterior Sampling (DPS) algorithm [Chung et al., 2023], which generates reconstructions by exploiting a tractable approximation to the time-dependent log-likelihood $\log p_t(y|x_t)$. Extensions to latent diffusion models, such as Rout et al. [2023b], show that when the ground-truth data lies in a low-dimensional subspace, the latent DPS algorithm asymptotically samples from the true posterior and recovers the ground-truth. Related theory for inpainting (Repaint [Lugmayr et al., 2022]) has been established in Rout et al. [2023a]. Recent works [Wu et al., 2024,

Sun et al., 2024, Xu and Chi, 2024] have leveraged diffusion priors in plug-and-play and MCMC-style algorithms, avoiding approximations used in prior works and established non-asymptotic stationarity guarantees. Particle-based approaches have also been developed, yielding asymptotically exact posterior sampling as number of particles increases [Dou and Song, 2024, Wu et al., 2023, Cardoso et al., 2024]. Finally, Bruna and Han [2024] transforms the original posterior sampling problem to a simpler one when given a denoising oracle, connecting the geometry of the posterior to properties of the inverse problem (e.g., conditioning of the measurement operator). In contrast to these works, our analysis departs from posterior sampling guarantees and develops a deterministic recovery theory, showing how diffusion priors can act as approximate projections onto a low-dimensional model set and yield quantitative convergence guarantees for inverse problems. This perspective is useful as it directly relates the geometry learned by the diffusion model to their ability to solve inverse problems. To do so, we use injectivity properties of the measurement operator on the low-dimensional model set to develop sample complexity guarantees. To our knowledge, prior results have not connected these quantities to establish recovery guarantees for diffusion priors in inverse problems.

**Theory for learning diffusion models on data with low intrinsic dimension.** While our work is mainly focused on the use of diffusion models as priors for inverse problems, we would also like to highlight prior works that have analyzed learning diffusion models from data concentrated on low-dimensional sets and their sampling capabilities. For example, Liang et al. [2025] study the iteration complexity of sampling from a distribution using DDIM and DDPM-type samplers, showing that the induced distribution has TV distance to the true distribution on the order of $\epsilon$ when the number of steps scales like $O(k/\epsilon)$ where $k$ captures the log-covering number of the support of the distribution. Under a different manifold hypothesis on the support of the distribution, Pierret and Galerne [2025] exhibit convergence rates for the KL-divergence, showing that $O(k/\epsilon)$ iterations are needed to achieve KL-divergence on the order of $\epsilon$ where $k$ is the dimension of the manifold. On the sample complexity side, other works [Wang et al., 2024, Chen et al., 2023, Achilli et al., 2025, Merger and Goldt, 2025] have also aimed to characterize the sample complexity of learning distributions with intrinsic low-dimensional structure, such as LR-GMMs or data supported on a manifold, via diffusion models. Our work complements this line of research by showing that when diffusion models are trained on data with low intrinsic dimension, they provably exploit the underlying geometry when used as priors for inverse problems via approximate projections onto the data's support, allowing us to establish rigorous sample complexity and recovery guarantees in this setting.

**Solving ill-posed inverse problems with deep projective priors.** In the vast choice of methods for the recovery of low-dimensional signals from compressive measurements [Foucart and Rauhut, 2013], the most relevant to our work are methods based on the projected gradient descent algorithm (PGD). Called iterative hard thresholding for sparse recovery [Blumensath and Davies, 2010], generalizations of this algorithm yield stable recovery when the number of measurements is of the order of the dimensionality of the model set (up to logarithmic terms, typically through a restricted isometry condition) [Golbabaee and Davies, 2018, Bahmani et al., 2016]. The benefit of PGD is that only a projection on the model set is sufficient to perform iterations. This makes the PGD framework a natural tool to design algorithms using learned *deep projective priors*. In particular, plug-and-play methods rely on a general purpose denoiser (e.g. implemented with a deep neural network (DNN)) to perform a generalized projection algorithm. In particular, the "proximal gradient descent" variant of the algorithm linearly converges to a stable solution under a restricted isometry condition on measurements and a Lipschitz condition on the projection [Liu et al., 2021]. This was improved to a *restricted Lipschitz* condition leading to a common interpretation of sparse recovery and plug-and-play PGD-like methods [Traonmilin et al., 2024]. The restricted Lipschitz constant is a good indicator of the performance of PGD, as was shown by learning methods that improve this constant [Joundi et al., 2025]. Many other variations of plug-and-play methods exist [Romano et al., 2017, Reehorst and Schniter, 2018] with their dedicated study of convergence [Cohen et al., 2021, Hurault et al., 2022] but are out of the scope of this paper that makes the specific connection between the diffusion approach from Kadkhodaie and Simoncelli [2021] and a generalization of PGD.

### 2.1 Notation

Consider a set $\Sigma \subset \mathbb{R}^d$. We say that a (potentially set-valued) map $P$ is a (generalized) projection onto $\Sigma$ if for every $x \in \mathbb{R}^d$, $P(x) \subset \Sigma$. We reserve the notation $P_\Sigma^\perp$ for the (potentially set-valued) orthogonal projection (or metric projection) onto $\Sigma$:

$$P_\Sigma^\perp(x) := \mathrm{proj}_\Sigma(x) := \arg\min_{z \in \Sigma} \|x - z\|_2.$$

Given a function $f : \mathbb{R}^d \to \mathbb{R}^d$, we let $\|f\|_{\text{op}}$ denote the operator norm induced by the Euclidean norm

$$\|f\|_{\text{op}} := \sup_{x \neq 0} \frac{\|f(x)\|_2}{\|x\|_2}.$$

Note that when $f$ is positively homogeneous, we have $\|f\|_{\text{op}} = \sup_{\|x\|_2=1} \|f(x)\|_2$. We use the notation $a \lesssim b$ to refer to when there exists a positive absolute constant $C$ such that $a \leq Cb$. The normal distribution of mean $\xi$ and covariance matrix $\Theta$ is denoted $\mathcal{N}(\xi, \Theta)$ and its density evaluated at $x \in \mathbb{R}^d$ is denoted $\mathcal{N}(x; \xi, \Theta)$. For distributions $P$ and $Q$, let $P * Q$ denote their convolution, which corresponds to the distribution of $z = u + v$ where $u \sim P$ and $v \sim Q$.

## 3 PROBLEM SETUP

Suppose that we are given linear measurements $y = A\hat{x}$ of an underlying signal $\hat{x}$. In this work, we will assume that $\hat{x}$ lies in a model set $\Sigma \subset \mathbb{R}^d$ with low intrinsic dimension. We limit ourselves to the noiseless case to focus on the underlying low-dimensional geometry of diffusion methods for inverse problems.

To address this problem, we take a Bayesian approach in which we assume access to a prior distribution $p$ that concentrates on the low-dimensional model set $\Sigma \subset \mathbb{R}^d$. Note that score-based diffusion priors obtain estimates of the score of the noise-convolved density $p_\sigma := p * \mathcal{N}(0, \sigma^2 I)$ for various values of $\sigma > 0$. Recall that via Tweedie's formula [Miyasawa et al., 1961], there is a connection between the score of the noise-convolved density $p_\sigma$ and the MMSE denoiser $D_\sigma(x) := \mathbb{E}[x_0 | x_0 + \sigma z = x]$ given by

$$\sigma^2 \nabla \log p_\sigma(x) = D_\sigma(x) - x.$$

Conditioned on linear measurements, one can compute the conditional score to estimate $\hat{x}$ from linear measurements [Kadkhodaie and Simoncelli, 2021]:

$$\sigma^2 \nabla \log p_\sigma(x|y) = A^T(y - Ax) + (I - A^T A)\sigma^2 \nabla \log p_\sigma(x).$$

By choosing a noise schedule $(\sigma_n)_{n \geq 0}$ that changes in a coarse-to-fine manner and adding Gaussian noise, we can use the above gradient steps to derive a stochastic estimate of $\hat{x}$.

In this work, we will consider a deterministic version of the algorithm presented in Kadkhodaie and Simoncelli [2021] and obtain recovery guarantees for $\hat{x}$. In particular, for a noise schedule $(\sigma_n)_{n \geq 0}$ that decays $\sigma_n \to 0$ as $n \to \infty$, we will consider iterates of the form

$$x_{n+1} = x_n - \sigma_n^2 \nabla \log p_{\sigma_n}(x|y) = x_n - (d_n + g_n) \quad (1)$$

where

$$\begin{aligned} d_n &:= A^T(Ax_n - y) \text{ and} \\ g_n &:= -\sigma_n^2 (I - A^T A) \nabla \log p_{\sigma_n}(x_n) \\ &= (I - A^T A)(x_n - D_{\sigma_n}(x_n)). \end{aligned}$$

Here $d_n$ is a direction that encourages data fit while $g_n$ is a direction that encourages high likelihood under the distribution $p_{\sigma_n}$. These are precisely the iterations considered in Kadkhodaie and Simoncelli [2021] with no noise and the step size set to the limiting choice of $\lim_{n \to \infty} h_n = 1$. Our goal is to establish conditions under which the sequence of iterates $x_n$ converges to $\hat{x}$ and to establish quantitative convergence rates.

## 4 THEORETICAL FRAMEWORK

The key insight of this work is that for many distributions concentrated on a low-dimensional model set $\Sigma$, the above iterations can be viewed as a variant of projected gradient methods where the projection operator is time-varying (i.e. iteration dependent). In particular, note that the prior direction is given by

$$-\sigma_n^2 \nabla \log p_{\sigma_n}(x) = x - D_{\sigma_n}(x) =: x - P^n(x)$$

where

$$P^n(x) := x + \sigma_n^2 \nabla \log p_{\sigma_n}(x) \quad (2)$$

is an iteration-dependent operator $P^n : \mathbb{R}^d \to \mathbb{R}^d$ that performs an approximate projection onto the model set $\Sigma$. The fundamental idea is that as $n \to \infty$, $P^n$ behaves like a true projection operator $P_\Sigma$ onto the model set $\Sigma$ in certain cases and the iterations (1) can be written as a generalized projected gradient descent with varying projections (GPGD-VP): algebraic manipulations yield that the iterates (1) can be written as

$$\begin{aligned} x_{n+1} &= x_n - A^T(Ax_n - y) - (I - A^T A)(x_n - P^n(x_n)) \\ &= P^n(x_n) - A^T(AP^n(x_n) - y). \end{aligned} \quad (3)$$

Note that we consider iterates before "projecting" with $P^n$ (PGD is often presented as iterates *after* projection). We will discuss several examples of data distributions and their corresponding projection operators in the following sections. In our analysis, we will consider the more general sequence of GPGD-VP iterates with step size $\mu > 0$:

$$x_{n+1} = P^n(x_n) - \mu A^T(AP^n(x_n) - y). \quad (4)$$

### 4.1 Preliminaries

For our theoretical results, we will assume certain properties of the underlying inverse problem, model

set, and projection operators to obtain convergence guarantees.

We will begin with the following definition of a restricted isometry constant, which is now classical in the context of signal recovery [Candès et al., 2006, Donoho, 2006].

**Definition 1.** *The operator $B$ has restricted isometry constant (RIC) $\delta < 1$ on the secant set $\Sigma - \Sigma = \{x_1 - x_2 : x_1, x_2 \in \Sigma\}$ if for all $x_1, x_2 \in \Sigma$,*

$$\|(B - I)(x_1 - x_2)\|_2 \leq \delta \|x_1 - x_2\|_2.$$

*We write $\delta_\Sigma(B)$ the smallest admissible restricted isometry constant (RIC). In the following we will only consider the RIC of $B = \mu A^T A$.*

Note that a lower restricted isometry property is necessary for the identification of a low-dimensional model set. With appropriate normalization $\mu$, this property is typically verified for a number of random Gaussian measurements $m \geq O(s \log(N/s))$ with high probability when $\Sigma = \Sigma_s$ the set of $s$-sparse vectors. This choice works more generally for random Gaussian matrices when $m \geq \Omega(w^2((\Sigma - \Sigma) \cap \mathbb{S}^{d-1})/\delta^2)$ where $w(T) := \mathbb{E}_{g \sim \mathcal{N}(0, I_d)}[\sup_{x \in T} \langle x, g \rangle]$ denotes the Gaussian width of a set $T$ [Liaw et al., 2017, Chandrasekaran et al., 2012]. It is also possible to construct compressive random measurements that satisfy the RIP for any low-dimensional model set, with a sample complexity on the order of the dimensionality of the model set (up to logarithmic factors) [Puy et al., 2017].

We now turn to properties of the projection operator. We will assume that our approximate projection operators $P^n$ converge (in a suitable sense) to a limiting projection operator $P_\Sigma$ that satisfies the following Lipschitz property with respect to the model set $\Sigma$. In Traonmilin et al. [2024], the restricted Lipschitz property was shown to guarantee the convergence of generalized projected gradient descent with fixed generalized projection.

**Definition 2** (Restricted Lipschitz property)**.** *Let $P$ be a generalized projection. Then $P$ has the restricted $\beta$-Lipschitz property with respect to $\Sigma$ if for all $z \in \mathbb{R}^d, x \in \Sigma, y \in P(z)$ we have*

$$\|y - x\|_2 \leq \beta \|z - x\|_2.$$

*We define $\beta_\Sigma(P)$ to be the smallest $\beta$ such that $P$ has the restricted $\beta$-Lipschitz property.*

This quantity characterizes how stable the projection operator $P_\Sigma$ is onto the model set $\Sigma$. Note that in general, we have that $\beta \geq 1$ with equality $\beta = 1$ in the case when $\Sigma$ is a closed, convex set and $P$ is the orthogonal projection $P_\Sigma^\perp$ onto $\Sigma$. For generic sets $\Sigma$, when the (set-valued) orthogonal projection defined as $P_\Sigma^\perp(x) = \arg\min_{u \in \Sigma} \|u - x\|_2$ is non-empty in $\mathbb{R}^d$, it has restricted Lipschitz property $\beta_\Sigma(P_\Sigma^\perp) \leq 2$. When $\Sigma = \{x \in \mathbb{R}^d : \|x\|_0 \leq s\}$ is the set of $s$-sparse vectors, it was shown (see [Traonmilin et al., 2024, Theorem 3.2]) that $\beta_\Sigma$ has a nearly optimal restricted Lipschitz constant $\beta_\Sigma(P_\Sigma^\perp) \leq \sqrt{\frac{3+\sqrt{5}}{2}} \approx 1.618$.

### 4.2 General results

We now turn towards our main results about iterations of the form (4) in recovering $\hat{x}$ from $y = A\hat{x}$. Our first main result is a general bound for the iterates. It shows that when the projection errors are bounded, we can obtain the following upper bound:

**Theorem 1** (Bound on iterates of GPGD-VP)**.** *Suppose $P_\Sigma$ is restricted $\beta$-Lipschitz. Let $\delta := \delta(\mu A^T A)$. Consider the iterations (4) with $\hat{x} \in \Sigma$ and assume the sequence $u_n := \|P^n(x_n) - P_\Sigma(x_n)\|_2$ is bounded. Then if $\delta\beta < 1$, there exists a constant $C$ depending on $\hat{x}, x_0, \mu, A, \beta, (u_n)_{n \geq 0}$ such that*

$$\|x_n - \hat{x}\|_2 \leq C((\delta\beta)^{n/2} + \max_{l = \lfloor n/2 \rfloor, n} \|P^l(x_l) - P_\Sigma(x_l)\|_2). \tag{5}$$

Depending on the behavior of $\|P^n - P_\Sigma\|_{\text{op}}$, we can guarantee fast convergence or only (potentially slow) convergence as described in the two following corollaries.

**Corollary 1** (Slow convergence)**.** *Under the hypotheses of Theorem 1, if $\lim_{n \to \infty} u_n = 0$, we have*

$$\lim_{n \to \infty} \|x_n - \hat{x}\|_2 = 0.$$

**Corollary 2** (Linear convergence)**.** *Suppose $P_\Sigma$ is restricted $\beta$-Lipschitz. Let $\delta := \delta(\mu A^T A)$. Consider the iterations (4) with $\hat{x} \in \Sigma$ and assume $\|P^n - P_\Sigma\|_{\text{op}} \leq O(e^{-cn})$ with $c > 0$. Then, there exist $C, D \geq 0$ such that*

$$\|x_n - \hat{x}\|_2 \leq C(\delta\beta)^{n/2} + De^{-\frac{c}{2}n}.$$

We will now analyze two examples of general distributions for which we can obtain recovery guarantees. The first considers data distributions concentrated on a low-dimensional convex set, while the latter focuses on the more challenging setting of data concentrated on a nonconvex union-of-subspaces.

### 4.3 Application to uniform measures on convex sets

Consider a measure $p$ that is uniform on a compact, convex set $\Sigma \subset \mathbb{R}^d$. In this case, one can show that for a given noise schedule $(\sigma_n)_{n \geq 0}$, the

iteration-dependent operator $P^n = D_{\sigma_n}$ in (2) converges uniformly to the metric projection $P_\Sigma^\perp(x) := \arg\min_{z\in\Sigma} \|x - z\|_2$ on bounded sets:

$$\lim_{n\to\infty} \sup_{\|x\|_2 \le R} \|P^n(x) - P_\Sigma^\perp(x)\|_2 = 0.$$

More specifically, one can obtain the following uniform error bound between the time-varying projection and the true projection operator onto $\Sigma$. Note that we allow for convex sets with low intrinsic dimension, which is captured by the dimensionality of its affine hull $\mathrm{aff}(\Sigma)$. Please see the Appendix (Section 8) for a more detailed discussion and proof of this result.

**Proposition 1.** *Suppose $\Sigma \subset \mathbb{R}^d$ is a compact, convex set containing the origin with intrinsic dimension $\dim(\mathrm{aff}(\Sigma)) = s \le d$ and let $p$ be the uniform measure on $\Sigma$. Then for any radius $R > 0$, there exists a constant $C_R$ such that the following holds:*

$$\sup_{\|x\|_2 \le R} \|P^n(x) - P_\Sigma^\perp(x)\|_2 \le C_R \sqrt{\left(\frac{s^2 R_\Sigma \lambda_s(\Sigma)}{\kappa_{\Sigma,s}}\right) \cdot \sigma_n^2 \log(1/\sigma_n)}$$

*where $R_\Sigma := \max_{z\in\Sigma} \|z\|_2$, $\lambda_s(\cdot)$ is the $s$-dimensional Hausdorff measure restricted to $\mathrm{aff}(\Sigma)$, and $\kappa_{\Sigma,s}$ depends on the volume of the $s$-dimensional Euclidean ball and inner and outer width of $\Sigma$ in $\mathrm{aff}(\Sigma)$ (see (12) in the Appendix).*

As a direct consequence, combining Proposition 1 with Theorem 1 and noting $P_\Sigma^\perp$ is $\beta = 1$-restricted Lipschitz in this setting, immediately yields the following convergence guarantee.

**Corollary 3.** *Suppose $\hat{x} \in \Sigma$ where $\Sigma$ is a compact, convex set containing the origin and let $p$ be the uniform measure on $\Sigma$. Consider the iterations (4). Then if $\delta := \delta(\mu A^T A) < 1$, there exists a constant $C$ such that*

$$\|x_n - \hat{x}\|_2 \le C\left(\delta^{n/2} + \sqrt{\sigma_n^2 \log(1/\sigma_n)}\right).$$

We observe that a geometric noise schedule is needed if we want to obtain a geometric convergence rate. Indeed, one can show that if $\sigma_n = \sigma_0 q^n$ for some $q \in (0,1)$, the rate is on the order of

$$\delta^{n/2} + \sqrt{n} q^n.$$

Choosing any $r \in (q, 1)$, one further has $\sqrt{n} \le (r/q)^n$ for $n$ sufficiently large so that $q^n\sqrt{n} \le r^n$. This gives a rate of the following form for all $n$ sufficiently large:

$$\|x_n - \hat{x}\|_2 \le C'\gamma^n, \text{ where } \gamma := \max\left\{\sqrt{\delta}, r\right\} < 1.$$

For nonconvex, low-dimensional sets associated with the LR-GMM in the next Section, we also observe the impact of the noise schedule on the convergence rate of GPGD-VP.

### 4.4 Application to Low-Rank Gaussian Mixture Models (LR-GMMs)

We now provide an example of a distribution $\hat{x} \sim p$ that is expressive, concentrates on a nonconvex low-dimensional model set $\Sigma$ and whose noise-corrupted score can be interpreted in our theoretical framework. Consider a distribution $p$ given by a mixture of $K$ low-rank (degenerate) Gaussian distributions

$$p = \sum_{k=1}^K \pi_k \mathcal{N}(0, U_k U_k^T)$$

where each $U_k \in \mathbb{R}^{d\times r_k}$ has rank $r_k \le d$ and $\sum_{k=1}^K \pi_k = 1$ Here, we assume that each $U_k$ has orthonormal columns. In this example, the low-dimensional model set is given by the union of subspaces $E_k := \mathrm{Im}(U_k)$, $\Sigma := \bigcup_{k=1}^K E_k$.

Gaussian mixture models are an expressive class of distributions that are both theoretically interesting and practically useful. Indeed, it has been shown that Gaussian mixture models can approximate smooth densities [Dalal and Hall, 1983]. In particular, Low-Rank GMMs (LR-GMM), have been proposed to approximate low-dimensional data [Chen et al., 2021] and have proved their usefulness in the context of imaging inverse problems [Zoran and Weiss, 2011, Houdard et al., 2018].

Prior results have shown that one can analytically characterize the noise-convolved score for the LR-GMM. Using results in Wang et al. [2024] (see, e.g., Proposition 1), one can show that for $\sigma_n > 0$, the noise-convolved score is given by

$$\begin{aligned}
&\sigma_n^2 \nabla \log p_{\sigma_n}(x) \\
&= \frac{1}{1+\sigma_n^2} \frac{\sum_k \pi_k \mathcal{N}(x; 0, U_k U_k^T + \sigma_n^2 I) U_k U_k^T x}{\sum_l \pi_l \mathcal{N}(x; 0, U_l U_l^T + \sigma_n^2 I)} - x \\
&=: P^n(x) - x.
\end{aligned}$$

Letting $\omega_k(x,t) := \frac{\pi_k \mathcal{N}(x;0,U_k U_k^T + tI)}{\sum_l \pi_l \mathcal{N}(x;0,U_l U_l^T + tI)}$ and $P_{E_k}^\perp(x) := U_k U_k^T x$, one can see that the operator $P^n$ is the following weighted sum of orthogonal projections onto each linear subspace $E_k$:

$$P^n(x) = \frac{1}{1+\sigma_n^2} \sum_{k=1}^K \omega_k(x, \sigma_n^2) P_{E_k}^\perp(x). \tag{6}$$

**Computation of the limiting projection.** Note that in simple cases, it is clear that the approximate projections converge uniformly with a decreasing noise schedule $\sigma_n \to 0$ and overall convergence is easily obtained. For example, in the case of $k = 1$, we have $\Sigma = \mathrm{Im}(U_1)$. Moreover, $P^n(x) = \frac{1}{1+\sigma_n^2} U_1 U_1^T x =:$

$\frac{1}{1+\sigma_n^2}P_\Sigma^\perp(x)$. Finally, since $\sigma_n^2 \to 0$ as $n \to \infty$ we have

$$\begin{aligned}\|P^n - P_\Sigma^\perp\|_{\text{op}} &= \sup_{\|x\|_2=1} \|P^n(x) - P_\Sigma^\perp(x)\|_2 \\ &= \left|\frac{1}{1+\sigma_n^2} - 1\right| \\ &= O(\sigma_n^2) \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

For the more general case in (6), we first aim to characterize what the value of the limiting projection would be.

**Lemma 1.** *Let $\Sigma = \bigcup_{k=1}^K E_k$. Then for $x \in \mathbb{R}^d$, we have*

- *if $P_\Sigma^\perp(x)$ is single-valued,*

$$\lim_{n\to\infty} P^n(x) = P_\Sigma^\perp(x);$$

- *if there is $r$ such that $\dim(E_l) = r$, and $x$ is such that $P_\Sigma^\perp(x) = \{u_l : l \in L\}$ is multi-valued (where $L \subset \{1, \dots, K\}$), then*

$$\lim_{n\to\infty} P^n(x) = \frac{\sum_{l\in L} \pi_l u_l}{\sum_{l\in L} \pi_l}.$$

While given in the discrete $t_n$ case, this result is valid for continuous $t \to 0$. In the following, we consider the GMM case with $K$ subspaces $E_1, \dots, E_K$ with equal dimension $r$ with $t = t_n = \sigma_n^2 \to_{n\to\infty} 0$.

We note that the restricted Lipschitz constant of the orthogonal projection plays an important role in our convergence results. For a union-of-subspaces, the geometry and relative orientation of the subspaces affect the restricted Lipschitz constant of the orthogonal projection $P_\Sigma^\perp$. Specifically, the orthogonal projection onto a finite union-of-subspaces may admit a restricted Lipschitz constant arbitrarily close to the upper bound $\beta = 2$, as established for orthogonal projections onto nonconvex sets (see [Traonmilin et al., 2024, Remark 3.1]). Thus, the precise value of $\beta$ in our setting reflects the geometry of the underlying subspaces.

**Challenge.** As discussed in the previous section, if we have uniform convergence of the projection operators $P^n$ to $P_\Sigma$, then convergence of the entire scheme can be guaranteed. However, this may not always be the case for general distributions. In particular, for the LR-GMM setting, uniform control over $\|P^n(x) - P_\Sigma^\perp(x)\|_2$ cannot be expected, as the approximation error increases for points $x$ near the *frontier* $F$ between subspaces, i.e., the set of points for which $x$ is equidistant to multiple subspaces. Hence a more fine-grained analysis needs to be done for this special case. Nevertheless, one can show that the iterates are eventually repelled away from such frontiers, which enables global convergence guarantees. At a high-level, we guarantee that after a given *burn-in* time period, the iterates $x_n$ exhibit convergence to $\hat{x} \in \Sigma$ (with the right noise schedule) under the condition that $\delta\beta < 1$. The convergence rate is a combination of a geometric convergence rate and an additional term that depends on the choice of noise schedule.

**Theorem 2** (Global convergence)**.** *Let $\delta := \delta(\mu A^T A)$ and $\beta := \beta_\Sigma(P_\Sigma^\perp)$. Let $\hat{x} \in \Sigma = \bigcup_{k=1}^K E_k$ such that there exists a unique $k \in \{1, \dots, K\}$, such that $\hat{x} \in E_k$. Suppose $\delta\beta < 1$ and $\sigma_n \to 0$. Then, there are $c, C > 0$, $n_0$ such that for $n \geq n_0$*

$$\|x_n - \hat{x}\|_2 \leq C\left((\delta\beta)^{n/2} + \max_{l=\lfloor n/2\rfloor, n} \exp\left(-\frac{c}{\sigma_l^2}\right) + \sigma_l^2\right).$$

**Proof sketch.** The proof relies on establishing two properties. First, we show that away from the frontier between subspaces, the error between projection operators $\|P^n(x_n) - P_\Sigma(x_n)\|_2$ decays exponentially in $\eta/\sigma_n^2$, where $\eta > 0$ controls the distance to the frontier. Second, we establish that if iterates approach the frontier, the update rule introduces a repulsive effect that eventually drives the iterates into a stable region associated with the true subspace of $\hat{x}$. Within this neighborhood, iterates stay within a ball of the true signal $\hat{x}$ (see the definition of $\Omega_\eta$ and Theorem 3 in Section 9.2.1 of the Appendix). Establishing this result is the main technical difficulty of the proof. Combining these two ingredients yields the stated convergence rate. Similar results are typical in the literature of sparse recovery: once the subspace supporting the unknown signal is identified, linear convergence is observed (see e.g., Liang et al. [2018]). A detailed local convergence analysis, which quantifies the basin of attraction and proves stability away from the frontier, is provided in the Appendix.

Note that the requirement that $\hat{x}$ strictly belongs to a single subspace $E_k$ holds with probability 1 under $p$, since the intersection of two distinct subspaces has measure zero with respect to $p$. Similarly to convergence guarantees with convex model sets, Theorem 2 provides a bound on the iterates that depends on both a geometric term and the noise schedule. If the noise schedule follows a geometric rate, the same argument as in Corollary 2 yields geometric convergence.

**Remark 1** (Random sparse vectors)**.** *The LR-GMM framework can also provide an analysis for other classes of signals of interest, such as sparse vectors. In particular, consider the following generative model for random sparse vectors: we say that $x \sim p_s$ if and only if $x$ is generated by taking a subset $S \subset [d]$ of size $|S| = s$ uniformly and setting $x_i \sim \mathcal{N}(0,1)$ for*

*$i \in [S]$ and $x_i = 0$ otherwise. An equivalent way of describing this distribution is with a LR-GMM where the parameters are defined as follows: enumerate all possible subsets of $[d]$ of cardinality $s$ and denote this set by $\mathcal{S} := \{S \subset [d] : |S| = s\}$. Note that there are $K = \binom{d}{s}$ possible subsets. Then $p_s$ is a LR-GMM of the form*

$$p_s = \sum_{S \in \mathcal{S}} \frac{1}{\binom{d}{s}} \mathcal{N}(0, U_S U_S^T)$$

*where $U_S := [e_{i_1}, \dots, e_{i_s}] \in \mathbb{R}^{d \times s}$ and $S := \{i_1, \dots, i_s\} \subset [d]$. Here, $e_i$ is the $i$-th standard basis vector in $\mathbb{R}^d$. In this case, note that our model set is given by $\Sigma = \bigcup_{S \in \mathcal{S}} \mathrm{Im}(U_S) = \{x \in \mathbb{R}^d : \|x\|_0 \leq s\}$. Hence for $s$-sparse data $\hat{x}$, the previous convergence guarantees apply.*

**Remark 2** (Open question on global convergence rates)**.** *While we provide global convergence rates in the case when the data distribution is supported on a convex set in Corollary 3, it is more challenging to characterize the precise time $n_0$ where we enjoy the convergence rate in Theorem 2 for LR-GMMs. In particular, the precise dependence of $n_0$ on the data geometry is more subtle and highly non-trivial to characterize due to the nonconvexity of the underlying model set. In the proof in Section 9.2.1 of the Appendix, we used a continuity argument to link the convergence of the quantities $\eta_n^{L_n}$ (which characterize the distance of the iterates to certain subspaces with indices in $L_n$) with the convergence of the distance to the "frontier" (the set of points equidistant to all subspaces) $d_F(x_n)$. If we could obtain a stronger bound of the form $d_F(x_n) \leq O(\eta_n^{L_n})$, global convergence rates can be established. Preliminary analysis shows that it is not an easy task to obtain such bounds for a generic union of subspaces. In the Appendix (see Section 9.4), we show in the simpler case of when the model set is the union of two orthogonal subspaces that one has $d_F(x_n) \leq \frac{1}{2}\eta_n^{L_n}$. We leave characterizing $n_0$ in the general case of an arbitrary union of subspaces as an open question.*

## 5 EXPERIMENTS

We consider numerical experiments to support our theoretical findings. In particular, we will consider synthetic data from a LR-GMM and illustrate the convergence of the algorithm we analyze.

Given a LR-GMM distribution $p = \frac{1}{K}\sum_{k=1}^{K} \pi_K \mathcal{N}(0, U_k U_k^T)$, we aim to recover signals from a random Gaussian compressed sensing matrix $y = A\hat{x}$ where $A \in \mathbb{R}^{m \times d}$ with $A_{ij} \sim \mathcal{N}(0,1)$ and $\hat{x} \sim p$. Each $U_k \in \mathbb{R}^{d \times r}$ is a random matrix with orthonormal columns. We run our algorithm with various noise schedules $(\sigma_n)$ and a fixed step size of $\mu = 1.9/\|A\|_2^2$. For our different choices of schedules, setting $N$ equal to the maximum number of iterations of our algorithm, we consider the following set of noise schedules:

(geometric) $\sigma_n = \sigma_{\max}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^{n/N}$,

(linear) $\sigma_n^2 = \sigma_{\max}^2 + \frac{n}{N}(\sigma_{\min}^2 - \sigma_{\max}^2)$,

(cosine) $\sigma_n = \sigma_{\min} + \frac{1}{2}(\sigma_{\max} - \sigma_{\min})(1 + \cos(\pi n/N))$,

(infinite geometric) $\sigma_n = \sigma_{\max} a^n$ where $a \in (0,1)$.

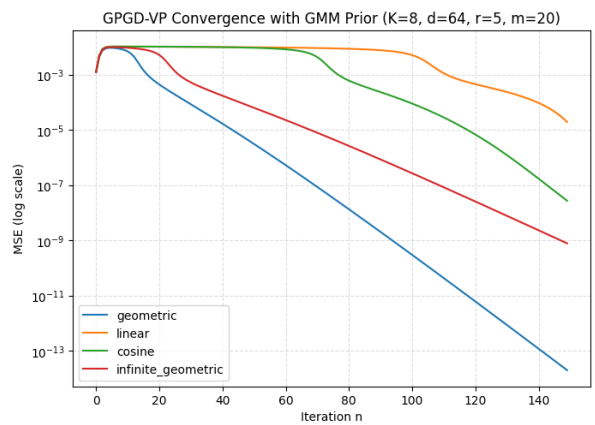


Figure 1: We take $\hat{x} \sim p = \frac{1}{8}\sum_{k=1}^{8} \mathcal{N}(0, U_k U_k^T)$ and recover $\hat{x}$ from $y = A\hat{x}$ where $A \in \mathbb{R}^{m \times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Here, $d = 64, r = 5$, and $m = 20$. We fix the step size $\mu = 1.9/\|A\|_2^2$ and consider 4 different noise schedules with $\sigma_{\max} = 0.5$ and $\sigma_{\min} = 10^{-4}$. We show the mean squared error over all iterations with $N = 150$.

Recovery results for our algorithm in the case when $K = 8$, $d = 64$, $r = 5$, and $m = 20$ are shown in Figure 1. We note that for a geometrically decaying noise schedule, the iterates exhibit linear convergence after an initial *burn-in* period, which is shorter for the geometric schedule as compared to each of the other noise schedules. In Figure 2, we further analyze this by zooming into the first part of the optimization and also display the distance of the iterates to each individual subspace $\mathrm{dist}(x_n, \mathrm{Im}(U_i)) := \|(I - U_i U_i^T)x_n\|_2$. We see that during this burn-in period, the iterates begin to reach the frontier, as their distance to each of the subspaces becomes close. After the correct subspace is identified, linear convergence is observed.

## 6 DISCUSSION

In this work, we developed a principled recovery theory for diffusion priors in linear inverse problems. Our

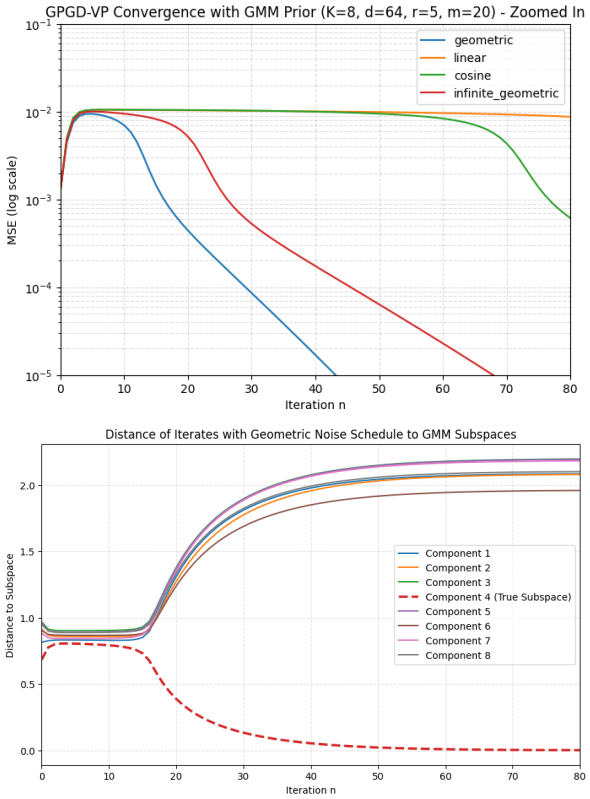


Figure 2: We consider the same experiment as in Figure 1. Top: a closer look at the first 80 iterations of the algorithm. Bottom: for the geometric noise schedule, we calculate the distance of the iterates to each subspace, with the correct underlying subspace error in the dashed line. We see that during this burn-in period, the iterates move closer towards the frontier, but are propelled away from it after a short time.

analysis exploited the fact that noise-convolved scores can be interpreted as approximate projections onto low-dimensional model sets. This perspective allows us to interpret certain diffusion-based algorithms as generalized projected gradient descent methods with time-varying projections, enabling recovery guarantees across a range of measurement operators and data distributions. Under a restricted isometry condition, we showed that (i) for uniform measures on convex sets, the limiting denoiser converges to the metric projection with an explicit error bound, giving global convergence for recovering signals from this model set. Also, (ii) for low-rank Gaussian mixture models (LR-GMMs), we showed that the limiting operator recovers a projection operator onto a union-of-subspaces and we established global convergence to the true signal. These results naturally cover (iii) recovery of random sparse signals as well. In all cases, we demonstrated that the convergence rate depends explicitly on the choice of noise schedule. These results bridge modern generative priors with the well-studied theory of regularization in inverse problems, leaving several open directions for future work.

First, while our analysis treated specific data distributions, it would be natural to extend the theory to other rich classes of distributions, such as those supported on nonlinear manifolds. Second, integrating our framework into sampling approaches such as DPS [Chung et al., 2023] or diffusion-based plug-and-play methods [Feng et al., 2023, Xu and Chi, 2024] could yield novel sample complexity guarantees for posterior sampling algorithms. Third, many applications involve nonlinear forward operators (e.g., phase retrieval), and extending our analysis to such settings would broaden the impact of this theory. Finally, it would be important to develop robustness guarantees with respect to measurement noise or model mismatch between the learned diffusion prior and the true data distribution.

## References

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and generalization in generative diffusion under the manifold hypothesis. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073401, 2025.

Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. In *International Conference on Learning Representations*, 2019.

Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Sohail Bahmani, Petros T Boufounos, and Bhiksha Raj. Learning model-based sparsity via projected gradient descent. *IEEE Transactions on Information Theory*, 62(4):2092–2099, 2016.

Thomas Blumensath and Mike E Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of selected topics in signal processing*, 4(2):298–309, 2010.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International conference on machine learning*, pages 537–546. PMLR, 2017.

Joan Bruna and Jiequn Han. Provable posterior sampling with denoising oracles via tilted transport. *Advances in Neural Information Processing Systems*, 37:82863–82894, 2024.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Founda-*

*tions of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

Gabriel Victorino Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided diffusion for bayesian linear inverse problems. In *ICLR International Conference on Learning Representations*, 2024.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023.

Yanxi Chen, Cong Ma, H Vincent Poor, and Yuxin Chen. Learning mixtures of low-rank models. *IEEE Transactions on Information Theory*, 67(7):4613–4636, 2021.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *International Conference on Learning Representations (ICLR)*, 2023.

Florentin Coeurdoux, Nicolas Dobigeon, and Pierre Chainais. Plug-and-play split gibbs sampler: embedding deep generative priors in bayesian inference. *IEEE Transactions on Image Processing*, 33:3496–3507, 2024.

Regev Cohen, Michael Elad, and Peyman Milanfar. Regularization by denoising via fixed-point projection (red-pro). *SIAM Journal on Imaging Sciences*, 14(3):1374–1406, 2021.

SR Dalal and WJ Hall. Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):278–286, 1983.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Ingrid Daubechies, Michel Defrise, and Christine de Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

David Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Maryam Fazel. Matrix rank minimization with applications. *Ph.D. Thesis, Department of Electrical Engineering, Stanford University*, 2002.

Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10520–10531, 2023.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013. ISBN 0817649476.

Mohammad Golbabaee and Mike E Davies. Inexact gradient projection and fast data driven compressed sensing. *IEEE Transactions on Information Theory*, 64(10):6707–6721, 2018.

Alexis Goujon, Sebastian Neumayer, Pakshal Bohra, Stanislas Ducotterd, and Michael Unser. A neural-network-based convex regularizer for inverse problems. *IEEE Transactions on Computational Imaging*, 9:781 – 795, 2023.

Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Paul Hand, Oscar Leong, and Vladislav Voroninski. Compressive phase retrieval: Optimal sample complexity with deep generative priors. *Communications on Pure and Applied Mathematics*, 77(2):1147–1223, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Antoine Houdard, Charles Bouveyron, and Julie Delon. High-dimensional mixture models for unsupervised image denoising (hdmi). *SIAM Journal on Imaging Sciences*, 11(4):2815–2846, 2018.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022.

Ali Joundi, Yann Traonmilin, and Alasdair Newson. Stochastic orthogonal regularization for deep projective priors. *arXiv preprint arXiv:2505.13078*, 2025.

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

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation (extended abstract). In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3723–3729. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/liang25a.html.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local linear convergence analysis of primal–dual splitting methods. *Optimization*, 67(6):821–853, 2018.

Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, pages 277–299. Springer, 2017.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

Jiaming Liu, Salman Asif, Brendt Wohlberg, and Ulugbek Kamilov. Recovery analysis for plug-and-play priors using the restricted eigenvalue condition. *Advances in Neural Information Processing Systems*, 34:5921–5933, 2021.

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

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Adversarial regularizers in inverse problems. *Advances in neural information processing systems*, 31, 2018.

Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. Pnp-flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations*, 2025.

Claudia Merger and Sebastian Goldt. Generalization dynamics of linear diffusion models. *arXiv preprint arXiv:2505.24769*, 2025.

Koichi Miyasawa et al. An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, 38(181-188):1–2, 1961.

Subhadip Mukherjee, Sören Dittmer, Zakhar Shumaylov, Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Learned convex regularizers for inverse problems. *arXiv preprint arXiv:2008.02839*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Emile Pierret and Bruno Galerne. Diffusion models for gaussian distributions: Exact solutions and wasserstein errors. In *International Conference on Machine Learning*, pages 49355–49381. PMLR, 2025.

Gilles Puy, Mike E Davies, and Rémi Gribonval. Recipes for stable linear embeddings from hilbert spaces to $\mathbb{R}^m$. *IEEE Transactions on Information Theory*, 63(4):2171–2187, 2017.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Edward T Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE transactions on computational imaging*, 5(1):52–67, 2018.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM journal on imaging sciences*, 10 (4):1804–1844, 2017.

Litu Rout, Advait Parulekar, Constantine Caramanis, and Sanjay Shakkottai. A theoretical justification for image inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2302.01217*, 2023a.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:49960–49990, 2023b.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algo-

rithms. *Physica D: nonlinear phenomena*, 60(1-4): 259–268, 1992.

Zakhar Shumaylov, Jeremy Budd, Subhadip Mukherjee, and Carola-Bibiane Schönlieb. Weakly convex regularisers for inverse problems: Convergence of critical points and primal-dual optimisation. In *Forty-first International Conference on Machine Learning*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *The Twelfth International Conference on Learning Representations*, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022.

Yu Sun, Zihui Wu, Yifan Chen, Berthy T Feng, and Katherine L Bouman. Provable probabilistic imaging using score-based generative priors. *IEEE Transactions on Computational Imaging*, 2024.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Yann Traonmilin, Jean François Aujol, and Antoine Guennec. Towards optimal algorithms for the recovery of low-dimensional models with linear rates. *arXiv preprint arXiv:2410.06607*, 2024.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pages 945–948. IEEE, 2013.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23 (7):1661–1674, 2011.

Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*, 2024.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 37:118389–118427, 2024.

Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *Advances in Neural Information Processing Systems*, 37:36148–36184, 2024.

Yasi Zhang, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, and Oscar Leong. Flow priors for linear inverse problems via iterative corrupted trajectory matching. *Advances in Neural Information Processing Systems*, 37:57389–57417, 2024.

Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *2011 international conference on computer vision*, pages 479–486. IEEE, 2011.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/**Not Applicable**]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]

(c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   Justification: all instructions needed to reproduce the synthetic experiments are provided in Section 5.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]

   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]

   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

## 7 Proofs for Section 4.2

We bound the recovery error $\|x_n - \hat{x}\|_2$ with the following Lemma.

**Lemma 2.** *Suppose $P_\Sigma$ is restricted $\beta$-Lipschitz with respect to $\Sigma$. Let $\delta := \delta(\mu A^T A)$. Consider $x_n$ resulting from iterations (4). We have, for $n \geq 1$,*

$$\|x_n - \hat{x}\|_2 \leq (\delta\beta)^n \|x_0 - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \sum_{l=0}^{n-1} (\delta\beta)^{n-l-1} \|P^l(x_l) - P_\Sigma(x_l)\|_2. \tag{7}$$

*Proof.* With the definition of the RIC $\delta := \delta(\mu A^T A)$ and the restricted Lipschitz constant $\beta$ we have

$$\begin{aligned} \|x_{n+1} - \hat{x}\|_2 &= \|(I - \mu A^T A)(P^n(x_n) - \hat{x})\|_2 \\ &= \|(I - \mu A^T A)(P_\Sigma(x_n) - \hat{x}) + (I - \mu A^T A)(P^n(x_n) - P_\Sigma(x_n))\|_2 \\ &\leq \delta\beta \|x_n - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \|P^n(x_n) - P_\Sigma(x_n))\|_2 \end{aligned} \tag{8}$$

We verify the result by induction. The bound (8) for $n = 0$ is exactly the case $n = 1$ of (7). Suppose (7) is verified at step $n$, then

$$\begin{aligned} \|x_{n+1} - \hat{x}\|_2 \leq & \delta\beta \left( (\delta\beta)^n \|x_0 - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \sum_{l=0}^{n-1} (\delta\beta)^{n-l-1} \|P^l(x_l) - P_\Sigma(x_l)\|_2 \right) \\ & + \|(I - \mu A^T A)\|_{\text{op}} \|P^n(x_n) - P_\Sigma(x_n)\|_2 \\ = & (\delta\beta)^{n+1} \|x_0 - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \sum_{l=0}^{n-1} (\delta\beta)^{n-l} \|P^l(x_l) - P_\Sigma(x_l)\|_2 \\ & + \|I - \mu A^T A\|_{\text{op}} \|P^l(x_l) - P_\Sigma(x_l)\|_2 \end{aligned}$$

We remark that $\|P^n(x_n) - P_\Sigma(x_n)\|_2 = (\delta\beta)^{n-n} \|P^n(x_n) - P_\Sigma(x_n)\|_2$, which allows to conclude

$$\|x_{n+1} - \hat{x}\|_2 \leq (\delta\beta)^{n+1} \|x_0 - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \sum_{l=0}^{n} (\delta\beta)^{n-l} \|P^l(x_l) - P_\Sigma(x_l)\|_2.$$

□

Lemma 2 bounds the recovery error by a geometric term and a term depending on the $\|P^l(x_l) - P_\Sigma(x_l)\|_2$. When these "projection errors" are small enough we are able to show convergence.

*Proof of Theorem 1.* First remark that the sequence $\|x_n - x_0\|_2$ is bounded, i.e. there exists $D \geq 0$ such that $\|x_n - x_0\|_2 \leq D$. Indeed, with the hypothesis, there exists $C' \geq 0$ such that $u_n = \|P^n(x_n) - P_\Sigma(x_n)\| \leq C'$, and from Lemma 2,

$$\begin{aligned} \|x_n - \hat{x}\|_2 &\leq \|x_0 - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} C' \sum_{l=0}^{n-1} (\delta\beta)^{n-l-1}. \\ &\leq \|x_0 - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} C' \sum_{l=0}^{\infty} (\delta\beta)^l =: D. \end{aligned}$$

We use Lemma 2 with initialization $x_{n_0}$ for some $n_0 \geq 0$. We have

$$\begin{aligned}
\|x_{n+n_0} - \hat{x}\|_2 &\leq (\delta\beta)^n \|x_{n_0} - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \sum_{l=0}^{n-1} (\beta\delta)^{n-1-l} \|P^{l+n_0}(x_{l+n_0}) - P_\Sigma\|_2 \\
&\leq D(\delta\beta)^n + \|I - \mu A^T A\|_{\text{op}} \sum_{l=0}^{n-1} (\beta\delta)^{n-1-l} \max_{l=0,n-1} \|P^{n_0+l}(x_{n_0+l}) - P_\Sigma(x_{n_0+l})\|_2 \\
&\leq D(\delta\beta)^n + \|I - \mu A^T A\|_{\text{op}} \frac{1}{1-\beta\delta} \max_{l=0,n-1} \|P^{n_0+l}(x_{n_0+l}) - P_\Sigma(x_{n_0+l})\|_2
\end{aligned}$$

Taking $n = n_0$, we have

$$\|x_{2n} - \hat{x}\|_2 \leq D(\delta\beta)^n + \frac{\|I - \mu A^T A\|_{\text{op}}}{1-\beta\delta} \max_{l=0,n-1} \|P^{n+l}(x_{n+l}) - P_\Sigma(x_{n+l})\|_2$$

Setting $C = \max(D, \frac{\|I-\mu A^T A\|_{\text{op}}}{1-\beta\delta})$, we have

$$\begin{aligned}
\|x_{2n} - \hat{x}\|_2 &\leq C((\delta\beta)^n + \max_{l=0,n-1} \|P^{n+l}(x_{n+l}) - P_\Sigma(x_{n+l})\|_2) \\
&\leq C((\delta\beta)^n + \max_{l=0,n} \|P^{n+l}(x_{n+l}) - P_\Sigma(x_{n+l})\|_2).
\end{aligned}$$

Taking $n = n_0 - 1$, we have

$$\begin{aligned}
\|x_{2n+1} - \hat{x}\|_2 &\leq C((\delta\beta)^n + \max_{l=0,n-1} \|P^{n+1+l}(x_{n+1+l}) - P_\Sigma(x_{n+1+l})\|_2) \\
&\leq C((\delta\beta)^n + \max_{l=1,n} \|P^{n+l}(x_{n+l}) - P_\Sigma(x_{n+l})\|_2) \\
&\leq C((\delta\beta)^n + \max_{l=0,n} \|P^{n+l}(x_{n+l}) - P_\Sigma(x_{n+l})\|_2).
\end{aligned}$$

Hence for any $m = 2n+1$ or $m = 2n$, we have $n = \lfloor m/2 \rfloor$, we deduce

$$\begin{aligned}
\|x_m - \hat{x}\|_2 &\leq C((\delta\beta)^{\lfloor m/2 \rfloor} + \max_{l=\lfloor m/2 \rfloor, \lfloor m/2 \rfloor + \lfloor m/2 \rfloor} \|P^l(x_l) - P_\Sigma(x_l)\|_2) \\
&\leq C((\sqrt{\delta\beta})^m + \max_{l=\lfloor m/2 \rfloor, m} \|P^l(x_l) - P_\Sigma(x_l)\|_2).
\end{aligned}$$

□

*Proof of Corollary 1.* Let $u_n := \|P^n(x_n) - P_\Sigma(x_n)\|_2$. By hypothesis, $(u_n)$ is a positive sequence that converges to 0. Hence $\lim_{n\to\infty} \sup_{l \geq n} u_l = \lim_{n\to\infty} u_n = 0$ and with Theorem 1,

$$\begin{aligned}
\|x_n - \hat{x}\| &\leq C((\sqrt{\delta\beta})^n + \max_{l=\lfloor n/2 \rfloor, n} \|P^l(x_l) - P_\Sigma(x_l)\|_2) \\
&\leq C((\sqrt{\delta\beta})^n + \sup_{l \geq \lfloor n/2 \rfloor} u_l) \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

□

*Proof of Corollary 2.* From the proof of Lemma 2, we have

$$\begin{aligned}
\|x_{n+1} - \hat{x}\|_2 &\leq \delta\beta \|x_n - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \|P^n(x_n) - P_\Sigma(x_n)\|_2 \\
&\leq \delta\beta \|x_n - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \|P^n - P_\Sigma\|_{\text{op}} \|x_n\|_2
\end{aligned}$$

and, by hypothesis, there exists $C'$ such that

$$\begin{aligned}\|x_{n+1} - \hat{x}\|_2 &\leq \delta\beta\|x_n - \hat{x}\|_2 + C'e^{-cn}\|x_n\|_2 \\ &\leq (\delta\beta + C'e^{-cn})\|x_n - \hat{x}\|_2 + C'e^{-cn}\|\hat{x}\|_2.\end{aligned}$$

We deduce that there is $n_1$, $\rho < 1$, such that for $n \geq n_1$

$$\|x_{n+1} - \hat{x}\|_2 \leq \rho\|x_n - \hat{x}\|_2 + C'e^{-cn}\|\hat{x}\|_2.$$

By induction,

$$\begin{aligned}\|x_{n_1+n} - \hat{x}\|_2 &\leq \rho^n\|x_{n_1} - \hat{x}\|_2 + \sum_{i=0}^{n-1}\rho^{n-i-1}C'e^{-c(n+n_1)}\|\hat{x}\|_2. \\ &\leq \rho^n\|x_{n_1} - \hat{x}\|_2 + \sum_{i=0}^{n-1}\rho^{n-i-1}C'\|\hat{x}\|_2. \\ &\leq \rho^n\|x_{n_1} - \hat{x}\|_2 + \frac{1}{1-\rho}C'\|\hat{x}\|_2.\end{aligned}$$

We deduce that $(x_n)$ is bounded and that $\lim_{n\to\infty}\|P_n(x_n) - P_\Sigma(x_n)\|_2 = 0$.

We conclude by remarking that there exists $D \geq 0$ such that

$$\max_{l=\lfloor n/2\rfloor,n}\|P^l(x_l) - P_\Sigma(x_l)\|_{\text{op}} \leq D\max_{l=\lfloor n/2\rfloor,n} e^{-cl} \leq De^{-\frac{c}{2}n}.$$

and by applying Theorem 1. □

## 8 Proofs for Section 4.3

To prove Proposition 1, we first require some additional notation and topological preliminaries. For a convex, compact set $\Sigma \subset \mathbb{R}^d$, let $\text{aff}(\Sigma)$ denote the affine hull of $\Sigma$, which corresponds to the smallest affine set that contains $\Sigma$:

$$\text{aff}(\Sigma) := \left\{\sum_{i=1}^k \lambda_i x_i : k > 0,\ x_i \in \Sigma,\ \lambda_i \in \mathbb{R},\ \sum_{i=1}^k \lambda_i = 1\right\}.$$

The intrinsic dimension of $\Sigma$ is then given by $s = \dim(\text{aff}(\Sigma))$. Note that when $\Sigma$ is a full-dimensional convex body (compact, convex with non-empty interior), the affine hull coincides with the entire ambient space, i.e., $\text{aff}(\Sigma) = \mathbb{R}^d$. Within $\text{aff}(\Sigma)$, we will consider several sets, such as the relative intrinsic ball in $\text{aff}(\Sigma)$:

$$B_{\text{aff}(\Sigma)}(x, r) := \{z \in \text{aff}(\Sigma) : \|z - x\|_2 < r\} = \text{aff}(\Sigma) \cap B(x, r).$$

We equip $\text{aff}(\Sigma)$ with the subspace topology, meaning that open sets in $\text{aff}(\Sigma)$ are of the form $\text{aff}(\Sigma) \cap U$ with $U$ open in $\mathbb{R}^d$. Then the relative interior of $\Sigma$, $\text{ri}(\Sigma)$, is

$$\text{ri}(\Sigma) := \{x \in \Sigma : \exists\epsilon > 0 \text{ s.t. } B_{\text{aff}(\Sigma)}(x, \epsilon) \subseteq \Sigma\}.$$

Then for compact $E$, we define the relative boundary of $E$ with respect to $\text{aff}(\Sigma)$ to be $\partial_{\text{aff}(\Sigma)}\Sigma := \Sigma \setminus \text{ri}(\Sigma)$. For our uniform measure, let $\lambda_s$ denote the $s$-dimensional Hausdorff measure restricted to $\text{aff}(\Sigma)$. Then the uniform probability measure on $E$ (relative to $\text{aff}(\Sigma)$) is

$$dp(x) = \frac{\mathbf{1}_{\{x\in\Sigma\}}}{\lambda_s(\Sigma)}d\lambda_s(x),\ x \in \text{aff}(\Sigma).$$

We are now ready to state the main result.

**Claim 1** (Proposition 1 in the main body)**.** *Suppose $\Sigma \subset \mathbb{R}^d$ is a compact, convex set containing the origin with non-empty relative interior in* $\mathrm{aff}(\Sigma)$*. Set $s = \dim(\mathrm{aff}(\Sigma))$ and let $p$ be the uniform measure on $\Sigma$. Then we have that there exists a positive absolute constant $C$ such that the following holds for any $y \in \mathbb{R}^d$:*

$$\|D_\sigma(y) - P_\Sigma^\perp(y)\|_2^2 \leq C\left(\frac{sR_\Sigma^2 \max\{1, \mathrm{dist}(y,\Sigma)\}^s \lambda_s(\Sigma)}{\kappa_{\Sigma,s}}\right) \cdot \sigma^2 \log(1/\sigma)$$

*where $R_\Sigma := \sup_{z\in\Sigma}\|z\|$ and $\kappa_{\Sigma,s}$ is the constant defined in Lemma 3. For a noise schedule $\sigma_n$, defining $P^n(x) := D_{\sigma_n}(x)$ as in (2), then we have for $n$ such that $\sigma_n < 1$,*

$$\sup_{\|x\|_2\leq R}\|P^n(x) - P_\Sigma^\perp(x)\|_2 \leq C_R\sqrt{\left(\frac{sR_\Sigma^2\lambda_s(\Sigma)}{\kappa_{\Sigma,s}}\right)\cdot\sigma_n^2\log(1/\sigma_n)}.$$

*Proof.* **Step 1: Connection between denoiser and truncated Gaussian:** We first show that the denoiser $D_\sigma(y)$ can be written as $P_\Sigma^\perp(y) = x^* = x^*(y)$ plus an additional term that is the conditional mean of a truncated Gaussian distribution. First, note that the posterior of $x \sim p$ given $y = x + \sigma z$ with $z \sim \mathcal{N}(0, I)$ is proportional to $\exp(-\|y-x\|_2^2/(2\sigma^2)\mathbf{1}_{\{x\in\Sigma\}}$ which gives

$$D_\sigma(y) = \mathbb{E}[x|y] = \frac{\int_\Sigma x\exp(-\|y-x\|_2^2/2\sigma^2)d\lambda_s(x)}{\int_\Sigma \exp(-\|y-u\|_2^2/2\sigma^2)d\lambda_s(u)}.$$

Using the substitutions $z = x - x^*$ and $z = u - x^*$ with $\Sigma_* := \Sigma - x^*$, we have that

$$D_\sigma(y) = x^* + \frac{\int_{\Sigma_*} z\exp(-\|y-z-x^*\|_2^2/2\sigma^2)d\lambda_s(z)}{\int_{\Sigma_*}\exp(-\|y-z-x^*)\|_2^2/2\sigma^2)d\lambda_s(z)}.$$

Thus, we get the following representation of the denoiser:

$$D_\sigma(y) - x^* = \frac{\int_{\Sigma_*} z\exp\left(-\frac{\|z-(y-x^*)\|_2^2}{2\sigma^2}\right)d\lambda_s(z)}{\int_{\Sigma_*}\exp\left(-\frac{\|z-(y-x^*)\|_2^2}{2\sigma^2}\right)d\lambda_s(z)}.$$

Hence this quantity has the interpretation as the mean of a degenerate, truncated Gaussian $\mathcal{N}(y-x^*, \sigma^2 I; \Sigma_*)$ which has measure $dp_\sigma(z) := \mathbf{1}_{\{z\in\Sigma_*\}}\exp(-\|z-(y-x^*)\|_2^2/2\sigma^2)d\lambda_s(z)/Z_\sigma$ with $Z_\sigma = \int_{\Sigma_*}\exp(-\|z-(y-x^*)\|_2^2/2\sigma^2)d\lambda_s(z)$, i.e., $D_\sigma(y) - x^* = \mathbb{E}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}[z]$. The next step decomposes an upper bound of $\|D_\sigma(y) - x^*\|_2^2$ depending on an arbitrary parameter $\delta > 0$.

**Step 2: General upper bound for $\delta > 0$:** Note that by Jensen's inequality, it suffices to bound

$$\|D_\sigma(y) - x^*\|_2^2 = \left\|\mathbb{E}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}[z]\right\|_2^2 \leq \mathbb{E}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}[\|z\|_2^2]. \tag{9}$$

Fix any $\delta \in (0,1)$ and consider the events $\mathbf{1}_{\|z\|_2>\delta}$ and $\mathbf{1}_{\|z\|_2\leq\delta}$. Set $R := \sup_{z\in\Sigma_*}\|z\|_2 < \infty$, which is finite since $\Sigma$ is compact. Then we note that

$$\begin{aligned}\mathbb{E}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}[\|z\|_2^2] &= \mathbb{E}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}[\|z\|_2^2\mathbf{1}_{\|z\|_2>\delta}] + \mathbb{E}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}[\|z\|_2^2\mathbf{1}_{\|z\|_2\leq\delta}] \\ &\leq R^2\mathbb{P}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}(\|z\|_2 > \delta) + \delta^2.\end{aligned}$$

We will estimate $\mathbb{P}(\|z\|_2 > \delta)$. Observe that

$$\mathbb{P}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}(\|z\|_2 > \delta) = \frac{\int_{z\in\Sigma_*,\|z\|_2>\delta}\exp\left(-\frac{\|z-(y-x^*)\|_2^2}{2\sigma^2}\right)d\lambda_s(z)}{\int_{\Sigma_*}\exp\left(-\frac{\|z-(y-x^*)\|_2^2}{2\sigma^2}\right)d\lambda_s(z)}.$$

First, note that for $z \in \Sigma_*$, we have by convexity of $\Sigma_*$ that $\langle z, y-x^*\rangle \leq 0$. This implies

$$\|z-(y-x^*)\|_2^2 = \|z\|_2^2 + \|y-x^*\|_2^2 - 2\langle z, y-x^*\rangle \geq \|z\|_2^2 + \|y-x^*\|_2^2.$$

Hence we have the following bound on the numerator:

$$\int_{z\in\Sigma_*,\|z\|_2>\delta} \exp\left(-\frac{\|z-(y-x^*)\|_2^2}{2\sigma^2}\right) d\lambda_s(z) \le \int_{z\in\Sigma_*,\|z\|_2>\delta} \exp\left(-\frac{\delta^2+\|y-x^*\|_2^2}{2\sigma^2}\right) d\lambda_s(z)$$
$$\le \lambda_s(\Sigma_*)\exp\left(-\frac{\delta^2+\|y-x^*\|_2^2}{2\sigma^2}\right).$$

For the denominator, note that $0\in\Sigma_*$ since $x^*\in\Sigma$ so $x^*-x^*=0\in\Sigma_*$. Consider a sufficiently small parameter $\varepsilon$ such that $\varepsilon<\frac{\delta^2}{4\max\{1,\|y-x^*\|_2\}}$. For $z\in B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*$, we have that $\|z-(y-x^*)\|\le\varepsilon+\|y-x^*\|$. Then we have that

$$\int_{\Sigma_*} \exp\left(-\frac{\|z-(y-x^*)\|_2^2}{2\sigma^2}\right) dz \ge \int_{B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*} \exp\left(-\frac{\varepsilon^2+2\varepsilon\|y-x^*\|_2+\|y-x^*\|_2^2}{2\sigma^2}\right) d\lambda_s(z)$$
$$= \exp\left(-\frac{\varepsilon^2+2\varepsilon\|y-x^*\|_2+\|y-x^*\|_2^2}{2\sigma^2}\right)\lambda_s(B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*).$$

Combining both bounds, we have that

$$\mathbb{P}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}(\|z\|_2>\delta) \le \frac{\lambda_s(\Sigma_*)}{\lambda_s(B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*)}\exp\left(-\frac{\delta^2+\|y-x^*\|_2^2}{2\sigma^2}+\frac{\varepsilon^2+2\varepsilon\|y-x^*\|+\|y-x^*\|_2^2}{2\sigma^2}\right)$$
$$= \frac{\lambda_s(\Sigma_*)}{\lambda_s(B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*)}\exp\left(\frac{-\delta^2+\varepsilon^2+2\varepsilon\|y-x^*\|_2}{2\sigma^2}\right)$$
$$\le \frac{\lambda_s(\Sigma_*)}{\lambda_s(B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*)}\exp\left(\frac{-\frac{\delta^2}{2}+\varepsilon^2}{2\sigma^2}\right)$$

where we used the fact that $\varepsilon<\delta^2/(4\|y-x^*\|_2)$ in the last inequality. By setting $\varepsilon:=\frac{\delta^2}{8\max\{1,\|y-x^*\|_2\}}$, we have that $-\delta^2/2+\varepsilon^2<0$. This implies that we have positive constants $C(\delta,y,x^*)$ and $c(\delta,y,x^*)$ that do not depend on $\sigma$ such that

$$\mathbb{P}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}(\|z\|_2>\delta)\le C(\delta,y,x^*)\exp\left(-\frac{c(\delta,y,x^*)}{2\sigma^2}\right).$$

Hence, for any fixed $\delta>0$ we have that

$$\mathbb{E}_{z\sim\mathcal{N}(y-x^*,\sigma^2 I;\Sigma_*)}[\|z\|_2^2]\le\delta^2+R^2C(\delta,y,x^*)\exp\left(-\frac{c(\delta,y,x^*)}{2\sigma^2}\right) \tag{10}$$

where $R:=\sup_{z\in\Sigma_*}\|z\|_2$, $c(\delta)=\delta^2/2-\varepsilon$ and $C(\delta):=\frac{\lambda_s(\Sigma_*)}{\lambda_s(B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*)}$.

**Step 3: Simplification for specific choice of $\delta(\sigma)$:** Since the above bound holds for any $\delta>0$, we will choose a specific value to minimize the above quantity. First, note that since $x^*\in\Sigma$, then we have $\lambda_s(B_{\mathrm{aff}(\Sigma)}(0,\varepsilon)\cap\Sigma_*)=\lambda_s(B_{\mathrm{aff}(\Sigma)}(x^*,\varepsilon)\cap\Sigma))$. Note that by Lemma 3, there exists an explicit $\kappa$, defined in (12), such that

$$C(\delta,y,x^*)\le\frac{\lambda_s(\Sigma_*)}{\kappa\varepsilon^s}=\frac{\lambda_s(\Sigma_*)8^s\max\{1,\|y-x^*\|_2\}^s}{\kappa\delta^{2s}}=:C_1\delta^{-2s}.$$

We also have that

$$c(\delta,y,x^*)=\frac{\delta^2}{2}-\frac{\delta^2}{8\max\{1,\|y-x^*\|_2\}}\ge\frac{\delta^2}{4} \tag{11}$$

so that, with (9) and (10), we get the bound

$$\|D_\sigma(y)-x^*\|_2^2\le\delta^2+R^2C_1\delta^{-2s}\exp\left(-\frac{\delta^2}{8\sigma^2}\right),\forall\delta>0.$$

Observe that the term on the right can be simplified as

$$(\delta^2)^{-s}\exp\left(-\frac{\delta^2}{8\sigma^2}\right) = \exp\left(-s\log(\delta^2) - \frac{\delta^2}{8\sigma^2}\right)$$

Note that if we choose

$$\delta^2 = 8(\alpha+2s)\sigma^2\log(1/\sigma),\ \alpha > 0$$

we obtain

$$\begin{aligned}\exp\left(-s\log(\delta^2) - \frac{\delta^2}{8\sigma^2}\right) &= \exp\left[-s\left(\log 8 + \log(\alpha+2s) + \log(\sigma^2) + \log\log(1/\sigma)\right)\right]\exp\left[-(\alpha+2s)\log(1/\sigma)\right] \\ &= 8^{-s}(\alpha+2s)^{-s}(\sigma^2)^{-s}[\log(1/\sigma)]^{-s}\cdot\sigma^{\alpha+2s} \\ &= (8(\alpha+2s))^{-s}\sigma^{\alpha}[\log(1/\sigma)]^{-s}.\end{aligned}$$

This gives the bound

$$\|D_\sigma(y) - x^*\|_2^2 \le 8(\alpha+2s)\sigma^2\log(1/\sigma) + R^2C_1(8(\alpha+2s))^{-s}\sigma^{\alpha}[\log(1/\sigma)]^{-s}.$$

Setting $\alpha = 2$ and using the definition of $C_1$ from (11) gives the following upper bound:

$$\|D_\sigma(y) - x^*(y)\|_2^2 \le 8(2+2s)\left(\frac{sR^2\lambda_s(\Sigma)\max\{1,\mathrm{dist}(y,\Sigma)\}^s}{\kappa}\right)\cdot\sigma^2\log(1/\sigma).$$

The final operator norm bound simply uses $\mathrm{dist}(y,\Sigma) = \min_{x\in\Sigma}\|x-y\|_2 \le \|0-y\|_2 = \|y\|_2 \le R$ since $0\in\Sigma$. □

**Lemma 3.** *Let $\Sigma\subset\mathbb{R}^d$ be a compact, convex set containing the origin with non-empty relative interior in* $\mathrm{aff}(\Sigma)$ *and set $s = \dim(\mathrm{aff}(\Sigma))$. Then there exists a constant $\kappa > 0$ that depends on $\Sigma$ such that for any $\varepsilon\in(0,1)$,*

$$\lambda_s(B_{\mathrm{aff}(\Sigma)}(x,\varepsilon)\cap E) \ge \kappa\epsilon^s\ \forall x\in\partial_{\mathrm{aff}(\Sigma)}\Sigma.$$

*A specific value for $\kappa$ can be taken as follows: if $\kappa_s$ denotes the volume of the unit Euclidean ball in $\mathbb{R}^s$ and we set $r := \sup\{t>0 : B_{\mathrm{aff}(\Sigma)}(0,t)\subseteq E\}$ and $R := \inf\{t>0 : \Sigma\subseteq B_{\mathrm{aff}(\Sigma)}(0,t)\}$, then*

$$\kappa := \kappa_s\cdot\left(\frac{\min\{r,1\}}{2\max\{R+r,1\}}\right)^s. \tag{12}$$

*Proof.* Fix $\varepsilon\in(0,1)$ and $x\in\partial_{\mathrm{aff}(\Sigma)}\Sigma$. Since $\Sigma$ is compact, convex with non-empty (relative) interior in $\mathrm{aff}(\Sigma)$, there exists $0<r<R<\infty$ such that $B_{\mathrm{aff}(\Sigma)}(0,r)\subseteq\Sigma\subseteq B_{\mathrm{aff}(\Sigma)}(0,R)$. Now, consider

$$\delta := \frac{\varepsilon}{2\max\{R+r,1\}}.$$

Note that by construction we have that $\delta\le\varepsilon/2 < 1/2$. Set $x_\delta := (1-\delta)x\in E$ and consider the ball $B_{\mathrm{aff}(\Sigma)}(x_\delta,\delta\min\{r,1\})$. We claim that $B_{\mathrm{aff}(\Sigma)}(x_\delta,\delta\min\{r,1\})\subseteq B_{\mathrm{aff}(\Sigma)}(x,\varepsilon)\cap\Sigma$. Indeed, let $y\in B_{\mathrm{aff}(\Sigma)}(x_\delta,\delta\min\{r,1\})$. To see that $y\in E$, note that we can write

$$y = (1-\delta)x + \delta z \text{ where } z\in B_{\mathrm{aff}(\Sigma)}(0,\min\{r,1\})\subseteq B_{\mathrm{aff}(\Sigma)}(0,r)\subseteq\Sigma.$$

Since $z,x\in\Sigma$ and $\Sigma$ is convex, we have that the linear combination $y=(1-\delta)x+\delta z\in\Sigma$. Moreover, we have that $y\in B_{\mathrm{aff}(\Sigma)}(x,\varepsilon)$ since $y\in\mathrm{aff}(\Sigma)$ and

$$\begin{aligned}\|y-x\|_2 &\le \|y-x_\delta\|_2 + \|x_\delta - x\|_2 \\ &\le \delta\min\{r,1\} + \delta\|x\|_2 \\ &\le \delta(1+R) \\ &\le \frac{\varepsilon}{2}\cdot 2 = \varepsilon.\end{aligned}$$

Hence we have that $B_{\mathrm{aff}(\Sigma)}(x_\delta, \delta\min\{r,1\}) \subseteq B_{\mathrm{aff}(\Sigma)}(x,\varepsilon)\cap\Sigma$. Let $\kappa_s := \mathrm{vol}(\{z\in\mathbb{R}^s : \|z\|_2 < 1\})$ be the $s$-dimensional volume of the unit Euclidean ball in $\mathbb{R}^s$. Using our above result, we have that

$$\begin{aligned}\lambda_s(B_{\mathrm{aff}(\Sigma)}(x,\varepsilon)\cap\Sigma) &\geq \lambda_s(B_{\mathrm{aff}(\Sigma)}(x_\delta,\delta\min\{r,1\}))\\ &= \mathrm{vol}(\{z\in\mathbb{R}^s : \|z\|_2 \leq \delta\min\{r,1\}\})\\ &= \kappa_s\cdot\delta^s\cdot\min\{r,1\}^s\\ &= \kappa_s\cdot\left(\frac{\min\{r,1\}}{2\max\{R+r,1\}}\right)^s\varepsilon^s\\ &=: \kappa\varepsilon^s.\end{aligned}$$

□

We now give a short proof of the corresponding convergence guarantee when the data distribution is uniform on a convex set.

*Proof of Corollary 3.* Note that following the proof of Theorem 1, since $P^n(x_n), P_\Sigma(x_n)\in\Sigma$, we have that the iterates are always bounded since $u_n \leq 2R_\Sigma$. Hence there exists a $D$ such that $\|x_n\|_2 \leq D$ for all $n$. Thus we can apply Proposition 1 with $R = D$ to (5) to get

$$\|x_n - \hat{x}\|_2 \leq C\left(\delta^{n/2} + \sigma_n\sqrt{\log(1/\sigma_n)}\right).$$

□

## 9 Proofs for Section 4.4

In the following, to make notations lighter, we use $t = t_n = \sigma_n^2$.

### 9.1 Computation of limiting projection for LR-GMM

For LR-GMM, we look at the limiting projection of

$$\begin{aligned}P^n(x) &= \frac{1}{1+t}\frac{\sum_{k=1}^K \pi_k\,\mathcal{N}\big(x;\,0,\,U_kU_k^T + tI\big)\,U_kU_k^T\,x}{\sum_{\ell=1}^K \pi_\ell\,\mathcal{N}\big(x;\,0,\,U_\ell U_\ell^T + tI\big)}\\ &= \frac{1}{1+t}\sum_{k=1}^K \omega_k(x,t)P^\perp_{E_k}(x)\end{aligned}$$

where $P^\perp_{E_k} = U_kU_k^T$ and where we defined

$$\omega_k(x,t) := \frac{\pi_k\,\mathcal{N}\big(x;\,0,\,U_kU_k^T + tI\big)}{\sum_{\ell=1}^K \pi_\ell\,\mathcal{N}\big(x;\,0,\,U_\ell U_\ell^T + tI\big)}.$$

Note that these quantities are not obviously defined at $t = 0$ because writing $\mathcal{N}(x;0,U_\ell U_\ell^T)$ makes the assumption that the probability measure associated with a low-dimensional Gaussian has a density in $\mathbb{R}^d$, but this is not the case.

However, the quantities $\omega_k(x,t)$ have a limit in $t = 0$.

**Lemma 4.** *Consider a low-rank GMM supported on $K$ subspaces $(E_\ell)_{\ell=1}^K$. We have for all $x\in\mathbb{R}^d$*

$$\lim_{t\to 0}\omega_k(x,t) = \begin{cases}1 & \text{if } \forall\ell\neq k, \|P^\perp_{E_k}(x)\|_2^2 > \|P^\perp_{E_\ell}(x)\|_2^2,\\ 0 & \text{if } \exists\ell\neq k, \|P^\perp_{E_k}(x)\|_2^2 < \|P^\perp_{E_\ell}(x)\|_2^2.\end{cases}$$

*Now assume that $\dim(E_k) = r$ for all $k$. Given $x\in\mathbb{R}^d$, define $L = \{\ell : \|P^\perp_{E_\ell}(x)\|_2^2 = \|P^\perp_\Sigma(x)\|_2^2\}$. Then we have*

$$\lim_{t\to 0}\omega_k(x,t) = \begin{cases}\frac{\pi_k}{\sum_{\ell\in L}\pi_\ell} & \text{if } k\in L,\\ 0 & \text{if } k\in L^c.\end{cases}$$

*where* $L^c = \{1, \dots, K\} \setminus L$.

*We consequently define* $\omega_k(x, 0) := \lim_{t \to 0} \omega_k(x, t)$.

*Proof.* We have, for $x \in \mathbb{R}^d$ and $t > 0$,

$$\omega_k(x,t) := \frac{\nu_k(x,t)}{\sum_{\ell=1}^K \nu_l(x,t)}$$

with

$$\nu_k(x,t) = \frac{\pi_k}{(2\pi)^{d/2} \det^{1/2}(U_k U_k^T + tI)} \cdot \exp\left(-\frac{1}{2} x^T (U_k U_k^T + tI)^{-1} x\right).$$

Let $r_k = \dim(E_k)$. We have $\det(U_k U_k^T + tI) = (1+t)^{r_k} t^{d-r_k}$.

Let $x = u + u^\perp$ with $u = P_{E_k}^\perp(x)$. Consider $V_k$ an orthonormal completion of $U_k$. We have

$$x^T (U_k U_k^T + tI)^{-1} x = x^T (V_k D_t V_k^T)^{-1} x$$

where $D_t$ is a diagonal matrix where $D_t(i,i) = (1+t)$ for $i \leq r_k$ and $D_t(i,i) = t$ for $i > r_k$. We deduce that (as $V_k$ is orthogonal)

$$\begin{aligned} x^T (U_k U_k^T + tI)^{-1} x = x^T V_k^T D_t^{-1} V_k x &= \frac{1}{1+t} u^T U_k U_k^T u + \frac{1}{t} (u^\perp)^T V_k^T V_k u^\perp \\ &= \frac{1}{1+t} \|u\|_2^2 + \frac{1}{t} \|u^\perp\|_2^2. \end{aligned}$$

We have

$$\omega_k(x,t) = \frac{1}{1 + \sum_{\ell \neq k} \frac{\nu_\ell(x,t)}{\nu_k(x,t)}}$$

and, using the Pythagorean identity $\|x\|_2^2 = \|P_{E_k}^\perp(x)\|_2^2 + \|P_{E_k}^\perp(x) - x\|_2^2$,

$$\begin{aligned} \frac{\nu_\ell(x,t)}{\nu_k(x,t)} =& \frac{\pi_\ell \sqrt{(1+t)^{r_k} t^{d-r_k}}}{\pi_k \sqrt{(1+t)^{r_\ell} t^{d-r_\ell}}} \\ & \times \exp\left(-\frac{1}{2}\left(\frac{1}{1+t}\|P_{E_\ell}^\perp(x)\|_2^2 + \frac{1}{t}\|P_{E_\ell}^\perp(x) - x\|_2^2 - \frac{1}{1+t}\|P_{E_k}^\perp(x)\|_2^2 - \frac{1}{t}\|P_{E_k}^\perp(x) - x\|_2^2\right)\right) \\ =& \frac{\pi_\ell \sqrt{(1+t)^{r_k} t^{d-r_k}}}{\pi_k \sqrt{(1+t)^{r_\ell} t^{d-r_\ell}}} \exp\left(-\frac{1}{2}\left(\frac{1}{t} - \frac{1}{1+t}\right)(\|P_{E_\ell}^\perp(x) - x\|_2^2 - \|P_{E_k}^\perp(x) - x\|_2^2)\right) \\ =& \frac{\pi_\ell \sqrt{(1+t)^{r_k} t^{d-r_k}}}{\pi_k \sqrt{(1+t)^{r_\ell} t^{d-r_\ell}}} \exp\left(-\frac{1}{2}\left(\frac{1}{t(1+t)}\right)(\|P_{E_\ell}^\perp(x) - x\|_2^2 - \|P_{E_k}^\perp(x) - x\|_2^2)\right) \end{aligned} \tag{13}$$

We deduce, as it is the ratio of (potentially) exponentially decreasing function over the square root of a polynomial function.

$$\lim_{t \to 0} \frac{\nu_\ell(x,t)}{\nu_k(x,t)} = \begin{cases} 0 & \text{if } \|P_{E_\ell}^\perp(x) - x\|_2^2 > \|P_{E_k}^\perp(x) - x\|_2^2, \\ \infty & \text{if } \|P_{E_\ell}^\perp(x) - x\|_2^2 < \|P_{E_k}^\perp(x) - x\|_2^2. \end{cases}$$

Using again the fact that $\|P_{E_k}^\perp(x) - x\|_2^2 = \|x\|_2^2 - \|P_{E_k}^\perp(x)\|_2^2$, we have equivalently

$$\lim_{t \to 0} \frac{\nu_\ell(x,t)}{\nu_k(x,t)} = \begin{cases} 0 & \text{if } \|P_{E_\ell}^\perp(x)\|_2^2 < \|P_{E_k}^\perp(x)\|_2^2, \\ \infty & \text{if } \|P_{E_\ell}^\perp(x)\|_2^2 > \|P_{E_k}^\perp(x)\|_2^2. \end{cases} \tag{14}$$

We deduce

$$\lim_{t \to 0} \omega_k(x,t) = \begin{cases} 1 & \text{if } \forall \ell \neq k, \|P_{E_k}^\perp(x)\|_2^2 > \|P_{E_\ell}^\perp(x)\|_2^2, \\ 0 & \text{if } \exists \ell \neq k, \|P_{E_k}^\perp(x)\|_2^2 < \|P_{E_\ell}^\perp(x)\|_2^2. \end{cases}$$

Now suppose $r_\ell = r$. Let $k, \ell \in L$, we have

$$\frac{\nu_\ell(x,t)}{\nu_k(x,t)} = \frac{\pi_\ell \sqrt{(1+t)^{r_k} t^{d-r_k}}}{\pi_k \sqrt{(1+t)^{r_\ell} t^{d-r_\ell}}} = \frac{\pi_\ell}{\pi_k}$$

and

$$\lim_{t\to 0} \omega_k(x,t) = \frac{1}{\sum_{\ell\in L} \pi_\ell/\pi_k} = \frac{\pi_k}{\sum_{\ell\in L}\pi_\ell}.$$

Let $k \in L^c$. It implies that there exists $\ell \in L$ such that $\|P^\perp_{E_k}(x)\|_2^2 < \|P^\perp_{E_\ell}(x)\|_2^2$ and, as in (14), $\lim_{t\to 0} \frac{\nu_\ell(x,t)}{\nu_k(x,t)} = \infty$, and $\lim_{t\to 0} \omega_k(x,t) = 0$.

□

Based on our understanding of $\omega(x,t)$, we can show the convergence to the limiting projection $P^\perp_\Sigma$.

*Proof of Lemma 1.* Note that as $t_n \to 0$, $P^n(x) = \frac{1}{1+t_n}\sum_{\ell=1}^K \omega_\ell(x,t_n) P^\perp_{E_\ell}(x) \to_{n\to\infty} \sum_{\ell=1}^K \omega_\ell(x,0) P^\perp_{E_\ell}(x)$. If $x$ is strictly closer to one subspace $E_k$ (with respect to the orthogonal projection), the previous Lemma shows that $\lim_{n\to\infty} P^n(x) = P^\perp_{E_k}(x)$. This is exactly the definition of the orthogonal projection onto the union-of-subspaces $\Sigma = \bigcup_\ell E_\ell$, i.e. $P^\perp_{E_k}(x) = P^\perp_\Sigma(x)$

When the dimensions of all subspaces are equal and $x$ is equidistant to subspaces $(E_\ell)_{\ell\in L}$, by Lemma 4, we have that

$$\lim_{n\to\infty} P^n(x) = \tilde{P}_\Sigma(x) := \sum_{k\in L} \frac{\pi_k}{\sum_{\ell\in L}\pi_\ell} P^\perp_{E_k}(x).$$

□

Note that this is a point-wise convergence result. Uniform convergence (for the operator norm) is not verified in general. Indeed, on the frontier between subspaces, $\tilde{P}_\Sigma(x)$ averages the projection over the different subspaces instead of selecting a subspace as $P^\perp_\Sigma$ does.

### 9.2 Convergence for the LR-GMM

We will first prove a local convergence result when iterates lie in a neighborhood of $\hat{x}$ away from the frontier between subspaces, and then obtain global convergence by showing that the iterates escape the frontier.

#### 9.2.1 Local convergence

To discuss our local convergence result, we will focus on regions that lie away from the frontier $F$, which will be defined shortly. For $x \in \mathbb{R}^d$, we denote by $k_x$ an index such that $P^\perp_{E_{k_x}}(x) \in P^\perp_\Sigma(x)$. We define the frontier between subspaces $F$ at $x$ to be

$$F := \left\{ x : \exists \ell \neq k_x,\ \|P^\perp_{E_\ell}(x) - x\|_2^2 = \|P^\perp_{E_{k_x}}(x) - x\|_2^2 \right\} = \left\{ x : \#P^\perp_\Sigma(x) > 1 \right\}.$$

where $\#P^\perp_\Sigma(x)$ denotes the number of elements in $P^\perp_\Sigma(x)$. When $k_x$ is unique, we define the set $\Omega_\eta$ to be

$$\begin{aligned}\Omega_\eta &:= \left\{ x : \left| \|P^\perp_{E_\ell}(x) - x\|_2^2 - \|P^\perp_{E_{k_x}}(x) - x\|_2^2 \right| \geq \eta \text{ for all } \ell \neq k_x \right\} \\ &= \left\{ x : \left| \|P^\perp_{E_\ell}(x)\|_2^2 - \|P^\perp_\Sigma(x)\|_2^2 \right| \geq \eta \text{ for all } \ell \neq k_x \right\}.\end{aligned}$$

This a set of points that lie away from the frontier $F$: when $\eta = 0$, $\Omega_\eta$ is the whole space, while when $\eta$ increases, elements $x$ of $\Omega_\eta$ are closer to $E_{k_x}$ than other subspaces $E_\ell$. A visual representation of $\Omega_\eta$ and $F$ can be found in Figure 9.2.1. We bound the rate of convergence of the operator norm of $P^n - P^\perp_\Sigma$ restricted to $\Omega_\eta$ denoted by $\|P^n - P^\perp_\Sigma\|_{\mathrm{op},\Omega_\eta}$.

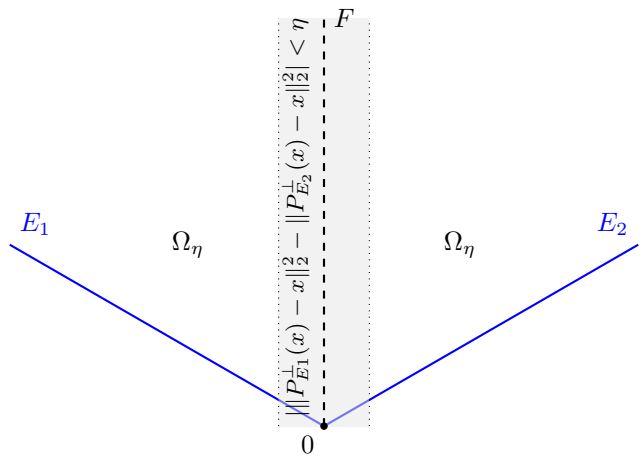


Figure 3: A representation of the set $\Omega_\eta$. Note that boundary between $\Omega_\eta$ and $\Omega_\eta^c$ might not be linear.

**Lemma 5.** *Consider $\Sigma = \bigcup_{\ell=1}^K E_\ell$ with $\dim(E_\ell) = r$ for all $\ell$. Consider $P^n$ defined by (6). Let $\eta > 0$ and consider $x \in \Omega_\eta$. Then we have*

$$\frac{\|P^n(x) - P^{\perp}_{\Sigma}(x)\|_2}{\|x\|_2} \leq \|P^n - P^{\perp}_{\Sigma}\|_{\mathrm{op},\Omega_\eta} \leq 2 \sum_{\ell=1,\ell\neq k_x}^{K} \frac{\pi_\ell}{\pi_{k_x}} \exp\left(-\frac{1}{2}\frac{\eta}{t_n(1+t_n)}\right) + t_n.$$

*As a consequence, for $t_n = \sigma_n^2$,*

$$\|P^n - P^{\perp}_{\Sigma}\|_{\mathrm{op},\Omega_\eta} = O(e^{-\frac{\eta}{4\sigma_n^2}} + \sigma_n^2).$$

*Proof.* Consider $t = t_n \in [0,1]$. Let $x \in \Omega_\eta$ and consider the unique $k \in \{1,\dots,K\}$ such that $P_\Sigma(x) = P^{\perp}_{E_k}(x)$. We have, for $\ell \neq k$, $\|P^{\perp}_{E_\ell}(x) - x\|_2^2 - \|P^{\perp}_{E_k}(x) - x\|_2^2 \geq \eta$ and, using (13),

$$\begin{aligned}\frac{\nu_\ell(x,t)}{\nu_k(x,t)} &= \frac{\pi_\ell}{\pi_k}\exp\left(-\frac{1}{2}\frac{1}{t(1+t)}(\|P^{\perp}_{E_\ell}(x) - x\|_2^2 - \|P^{\perp}_{E_k}(x) - x\|_2^2)\right) \\ &\leq \frac{\pi_\ell}{\pi_k}\exp\left(-\frac{1}{2}\frac{\eta}{t(1+t)}\right).\end{aligned} \tag{15}$$

Using the expression of $P^n$ from (13), with the triangle inequality, the fact that $\|P^{\perp}_{E_\ell}(x)\|_2 \leq \|x\|_2$ for any $x$ and $\frac{1}{1+t} \leq 1$, we deduce that

$$\begin{aligned}\|P^n(x) - P^{\perp}_{\Sigma}(x)\|_2 &= \Big\|(\frac{1}{1+t}\omega_k(x,t) - 1)P^{\perp}_{E_k}(x) + \frac{1}{1+t}\sum_{\ell=1;\ell\neq k}^{K}\omega_\ell(x,t)P^{\perp}_{E_\ell}(x)\Big\|_2 \\ &\leq \|x\|_2\left(\left|\frac{1}{1+t}\omega_k(x,t) - 1\right| + \frac{1}{1+t}\sum_{\ell=1;\ell\neq k}^{K}\omega_\ell(x,t)\right) \\ &\leq \|x\|_2\left(\left|\frac{\omega_k(x,t) - 1}{1+t}\right| + \frac{t}{1+t} + \frac{1}{1+t}\sum_{\ell=1;\ell\neq k}^{K}\omega_\ell(x,t)\right) \\ &\leq \|x\|_2\left(|\omega_k(x,t) - 1| + t + \sum_{\ell=1;\ell\neq k}^{K}\omega_\ell(x,t)\right).\end{aligned}$$

We have, with (15),

$$|\omega_k(x,t)-1| = \left|\frac{\nu_k(x,t)}{\sum_{\ell=1}^K \nu_\ell(x,t)} - 1\right| = \left|\frac{\sum_{\ell=1;\ell\neq k}^K \nu_\ell(x,t)}{\sum_{\ell=1}^K \nu_\ell(x,t)}\right| \leq \sum_{\ell=1;\ell\neq k}^K \frac{\nu_\ell(x,t)}{\nu_k(x,t)} \leq \sum_{\ell=1;\ell\neq k}^K \frac{\pi_\ell}{\pi_k}\exp\left(-\frac{1}{2}\frac{\eta}{t(1+t)}\right).$$

For $\ell \neq k$, we have

$$\omega_\ell(x,t) = \frac{\nu_\ell(x,t)}{\sum_{\ell'=1}^K \nu_{\ell'}(x,t)} \leq \frac{\nu_\ell(x,t)}{\nu_k(x,t)} \leq \frac{\pi_\ell}{\pi_k}\exp\left(-\frac{1}{2}\frac{\eta}{t(1+t)}\right).$$

We deduce overall

$$\frac{\|P^n(x) - P_\Sigma^\perp(x)\|_2}{\|x\|_2} \leq 2\sum_{\ell=1,;0\ell\neq k}^K \frac{\pi_\ell}{\pi_k}\exp\left(-\frac{1}{2}\frac{\eta}{t(1+t)}\right) + t.$$

□

Using this bound, we establish local convergence guarantees: there exists a neighborhood around $\hat{x}$ such that any iterate entering this region remains within it and converges at the following rate.

**Lemma 6.** *Consider the hypotheses of Lemma 5. Suppose $\delta\beta < 1$ with $\delta = \delta(\mu A^T A), \beta = \beta_\Sigma(P_\Sigma^\perp)$. Let $C > 0$ and $x \in B(\hat{x}, C)$. Suppose $B(\hat{x}, C) \subset \Omega_\eta$. Then there is $n_0$ such that for $n \geq n_0$, defining $z = P^n(x) - \mu A^T A(P^n(x) - \hat{x})$, we have $z \in B(\hat{x}, C)$ .*

*Proof.* We have

$$\begin{aligned}\|z-\hat{x}\|_2 &\leq \delta\beta\|x-\hat{x}\|_2 + \|I-\mu A^T A\|_{\mathrm{op}}\|P^n(x) - P_\Sigma^\perp(x))\|_2 \\ &\leq \delta\beta C + \|I-\mu A^T A\|_{\mathrm{op}}\left(2C'\sum_{\ell=1,\ell\neq K}^K \frac{\pi_\ell}{\pi_k}\exp\left(-\frac{1}{2}\frac{c}{t(1+t)}\right) + C'\frac{t}{1+t}\right)\end{aligned}$$

Hence, as $\delta\beta < 1$, there is $t = \sigma_n^2$ sufficiently small such that $\|z - \hat{x}\|_2 \leq C$. □

We have just guaranteed the stability of iterates in $B(\hat{x}, C) \subset \Omega_\eta$. Using the previous results, we can now provide a convergence rate if the iterates fall in a basin of attraction of $\hat{x}$. Note that we require $\hat{x}$ to belong to only one subspace $E_k \subset \Sigma$. In practice this is not a problem since the set of intersections of subspaces is of measure 0 (being of dimension strictly lower than $r = \dim(E_k)$, $k \in \{1, \ldots, K\}$).

**Theorem 3.** *Let $\hat{x} \in \Sigma$ such that there exists a unique $k \in \{1, \ldots, K\}$, such that $\hat{x} \in E_k$. Suppose $\sigma_n^2$ decreases to 0. There exists $c, C, C_1, C_2$, $n_0$ such that, if there exists $n_1 \geq n_0$ such that $x_{n_1} \in B(\hat{x}, C)$, then, for all $n \geq 0$*

$$\|x_{n_1+n} - \hat{x}\|_2 \leq \; C_1(\sqrt{\delta\beta})^n\|x_{n_1} - \hat{x}\|_2 + C_2(e^{-\frac{c}{\sigma_{n/2}^2}} + \sigma_{n/2}^2).$$

*Proof.* As $\hat{x} \neq 0$, taking $C$ sufficiently small, we have $B(\hat{x}, C) \subset \Omega_\eta$. With Lemma 6, there is $n_0$ such that for $n_1 \geq n_0$, we have $x_{n_1+1} \in B(\hat{x}, C)$. By induction, for any $n \geq 0$, $x_{n_1+n} \in B(\hat{x}, C) \subset \Omega_\eta$, we deduce that $x_n$ is bounded, and with Theorem 1 and the bound on $\|P^l - P_\Sigma\|_{\mathrm{op},\Omega_\eta}$ from Lemma 5, we have that there exists $C'$,

$$\begin{aligned}\|x_{n_1+n} - \hat{x}\|_2 &\leq C'((\sqrt{\delta\beta})^n\|x_{n_1} - \hat{x}\|_2 + \max_{l=\lfloor n/2\rfloor, n}\|P^l(x_l) - P_\Sigma(x_l)\|_2) \\ &\leq C'(\sqrt{\delta\beta})^n\|x_{n_1} - \hat{x}\|_2 + O(e^{-\frac{c}{\sigma_{n/2}^2}} + \sigma_{n/2}^2).\end{aligned}$$

□

### 9.3 Global convergence

In this Section we show global convergence of $(x_n)$. We guarantee that after a given "burn-in" time period, the iterates $x_n$ converge linearly to $\hat{x} \in \Sigma$ (with the right noise schedule) under the condition that $\delta\beta < 1$. The main technical difficulty of the proof is when $x_n$ is potentially close to the frontiers between subspaces $F_L = \{x : \|P_\Sigma(x)\|_2 = \|P^\perp_{E_\ell}(x)\|_2, \ell \in L\}$, $L \subset \{1, \dots, K\}$ (see Figure 2). Note that we define $L^c := \{1, \dots, K\} \backslash L$. In this case, we use the fact that the iterations pull the iterates away from this frontier.

**Lemma 7.** *Consider $\Sigma = \bigcup_{\ell=1}^K E_\ell$ and $P^n$ defined in (6). Let $\delta := \delta(\mu A^T A)$. Let $\beta = \beta_\Sigma(P^\perp_\Sigma)$. Let $L \subset \{1, \dots, K\}$ and $v \in F_L$. Then we have*

$$\|x_{n+1} - \hat{x}\|_2 \leq \delta\beta\|x_n - \hat{x}\|_2 + \|v\|_2\delta\left(\sum_{\ell \in L^c} \omega_\ell(x_n, t_n)\right) + \delta(1+\beta)\|x_n - v\|_2 + \delta t_n\|\hat{x}\|_2.$$

*Proof.* Consider $t = t_n > 0$. Letting $v \in F_L$, we have, using the fact that $P^n(x) = \frac{1}{1+t}\sum_{\ell=1}^K \omega_\ell(x,t)P^\perp_{E_\ell}(x)$ for any $x$ and $\sum_\ell \omega_\ell(x_n, t) = 1$, by convexity,

$$\begin{aligned}
\|x_{n+1} - \hat{x}\|_2 &= \|(I - \mu A^T A)(P^n(x_n) - \hat{x})\|_2 \\
&= \|(I - \mu A^T A)(\sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}\left(P^\perp_{E_\ell}(x_n) - P^\perp_{E_\ell}(v) + P^\perp_{E_\ell}(v)\right) - \hat{x})\|_2 \\
&\leq \sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}\|(I - \mu A^T A)(P^\perp_{E_\ell}(v) - (1+t)\hat{x})\|_2 + \|(I - \mu A^T A)(\sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}(P^\perp_{E_\ell}(x_n) - P^\perp_{E_\ell}(v))\|_2 \\
&\leq \sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}\|(I - \mu A^T A)(P^\perp_{E_\ell}(v) - \hat{x} - t\hat{x})\|_2 + \sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}\|(I - \mu A^T A)P^\perp_{E_\ell}(x_n - v))\|_2.
\end{aligned}$$

With the triangle inequality and the RIC, and the fact that $1 + t \geq 1$

$$\begin{aligned}
\|x_{n+1} - \hat{x}\|_2 &\leq \delta\sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}(\|P^\perp_{E_\ell}(v) - \hat{x}\|_2 + t\|\hat{x}\|) + \sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}\delta\|P^\perp_{E_\ell}(x_n - v)\|_2 \\
&\leq \delta\sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}\|P^\perp_{E_\ell}(v) - \hat{x}\|_2 + \sum_{\ell=1}^K \frac{\omega_\ell(x_n,t)}{1+t}\delta\|x_n - v\|_2 + \frac{\delta t}{1+t}\|\hat{x}\|_2 \\
&\leq \delta\sum_{\ell=1}^K \omega_\ell(x_n,t)\|P^\perp_{E_\ell}(v) - \hat{x}\|_2 + \delta\|x_n - v\|_2 + \delta t\|\hat{x}\|_2.
\end{aligned}$$

As $P^\perp_{E_\ell}(v) \in P^\perp_\Sigma(v)$ for $\ell \in L$ by definition of $F_L$, using the restricted Lipschitz condition of $P^\perp_\Sigma$, and the triangle inequality, we have

$$\begin{aligned}
\|x_{n+1} - \hat{x}\|_2 &\leq \delta\beta\sum_{\ell \in L}\omega_\ell(x_n,t)\|v - \hat{x}\|_2 + \delta\sum_{\ell \in L^c}\omega_\ell(x_n,t)\|P^\perp_{E_\ell}(v) - \hat{x}\|_2 + \delta\|x_n - v\|_2 + \delta t\|\hat{x}\|_2 \\
&= \delta\beta\|v - \hat{x}\|_2 + \delta\sum_{\ell \in L^c}\omega_\ell(x_n,t)(\|P^\perp_{E_\ell}(v) - \hat{x}\|_2 - \beta\|v - \hat{x}\|_2) + \delta\|x_n - v\|_2 + \delta t\|\hat{x}\|_2 \\
&\leq \delta\beta\|x_n - \hat{x}\|_2 + \delta\sum_{\ell \in L^c}\omega_\ell(x_n,t)(\|P^\perp_{E_\ell}(v) - \hat{x}\|_2 - \beta\|v - \hat{x}\|_2) + (\delta + \delta\beta)\|x_n - v\|_2 + \delta t\|\hat{x}\|_2.
\end{aligned}$$

We remark that $\|P^\perp_{E_\ell}(v) - \hat{x}\|_2 - \beta\|v - \hat{x}\|_2 \leq \|(I - P^\perp_{E_\ell})(v)\|_2 + (1-\beta)\|v - \hat{x}\|_2 \leq \|(I - P^\perp_{E_\ell})(v)\|_2 \leq \|v\|_2$ and we conclude

$$\|x_{n+1} - \hat{x}\|_2 \leq \delta\beta\|x_n - \hat{x}\|_2 + \delta\sum_{\ell \in L^c}\omega_\ell(x_n,t)\|v\|_2 + (\delta + \delta\beta)\|x_n - v\|_2 + \delta t\|\hat{x}\|_2.$$

□

We need the following Lemma to bound the quantities $\omega_\ell(x_n, t)$ given in Lemma 7.

**Lemma 8.** *Let $L \subset \{1, \dots, K\}$, $x \in \mathbb{R}^d$ such that $P^\perp_{E_k}(x) \in P^\perp_\Sigma(x)$. Suppose for $\ell \in L^c$ we have $\|P^\perp_{E_\ell}(x_n) - x_n\|_2^2 - \|P^\perp_{E_k}(x_n) - x_n\|_2^2 \geq \alpha$, then*

$$\sum_{\ell \in L^c} \omega_\ell(x,t) \leq \left( \sum_{\ell \in L^c} \frac{\pi_\ell}{\pi_k} \right) \exp\left( -\frac{1}{2} \frac{\alpha}{t(1+t)} \right)$$

*Proof.* We have, similarly as in the proof of Lemma 5, for $\ell \in L^c$,

$$\omega_\ell(x,t) \leq \frac{\pi_\ell}{\pi_k} \frac{\nu_\ell(x,t)}{\nu_k(x,t)} \leq \frac{\pi_\ell}{\pi_k} \exp\left( -\frac{1}{2} \frac{\alpha}{t(1+t)} \right).$$

We deduce

$$\sum_{\ell \in L^c} \omega_\ell(x,t) \leq \sum_{\ell \in L^c} \frac{\pi_\ell}{\pi_k} \exp\left( -\frac{1}{2} \frac{\alpha}{t(1+t)} \right).$$

□

Suppose $P^\perp_{E_k}(x_n) \in P_\Sigma(x_n)$. Let $L \subset \{1, \dots, K\}$. Define

$$\eta_n^L := \min_{\ell \in L, \ell \neq k} \left( \|P^\perp_{E_\ell}(x_n) - x_n\|_2^2 - \|P^\perp_{E_k}(x_n) - x_n\|_2^2 \right) = \min_{\ell \in L, \ell \neq k} \left( \|P_\Sigma(x_n)\|_2^2 - \|P^\perp_{E_\ell}(x_n)|_2^2 \right)$$

Before giving the main convergence result, we need to link the definition of $\eta_n^{L_n}$ with the accumulation of points near $F_{L_n}$, where $L_n$ is a sequence of indices defining a sequence of frontier sets.

**Lemma 9.** *Consider a sequence $(x_n)$ and a sequence of indices $L_n \subset \{1, \dots, K\}$. Suppose $\lim_{n\to\infty} \eta_n^{L_n} = 0$. Then $d(x_n, F_{L_n}) = \min_{v \in F_{L_n}} \|v - x_n\|_2 \to_{n\to\infty} 0$.*

*Proof.* Define $k_x$ the index of a subspace closest to $x$. Consider the functions

$$f_{L_n}(x) = \min_{\ell \in L_n} (\|P^\perp_{E_\ell}(x) - x\|_2^2 - \|P^\perp_{E_{k_x}}(x) - x\|_2^2) \geq 0.$$

We have that $f_{L_n}$ is continuous. Indeed it is the minimum of $|L_n|$ continuous functions $x \mapsto \|P^\perp_{E_\ell}(x) - x\|_2^2 - \|P^\perp_{E_{k_x}}(x) - x\|_2^2 = \|P^\perp_\Sigma(x)\|_2^2 - \|P^\perp_{E_\ell}(x)\|_2^2$ (because $x \mapsto \|P^\perp_\Sigma(x)\|_2^2$ is continuous, see [Traonmilin et al., 2024, Lemma 3.7]).

Hence, for a fixed $n$, any sequence $x_i$ such that $\lim_{i\to\infty} f_{L_n}(x_i) = 0$ is such that $d(x_i, F_{L_n}) \to_{i\to\infty} 0$ as $F_{L_n} = \{x : f_{L_n}(x) = 0\}$. As there is only a finite number of functions $f_{L_n}$, the sequence $f_{L_n}(x_n)$ can be partitioned into a finite number of subsequences where $L_n = L$. For each of these subsequences, we have $\lim_{n\to\infty} d(x_n, F_L) = 0$, which concludes the proof. □

We now give the main Theorem controlling global convergence.

*Proof of Theorem 2.* First, we use Lemma 7 with $v = 0$ and $L = \{1, \dots, K\}$ this gives

$$\|x_{n+1} - \hat{x}\|_2 \leq \delta\beta \|x_n - \hat{x}\|_2 + \delta(1+\beta)\|x_n\|_2 + \delta t \|\hat{x}\|_2.$$

We observe that there exists $b > 0$ and a rank $n_0$ such that for $n \geq n_0$, $\|x_n\|_2 > b$, otherwise we can find a subsequence of $x_n$ converging to 0 and, as $t \to 0$ taking the above inequality to the limit leads to $\|\hat{x}\|_2 \leq \delta\beta\|\hat{x}\|_2$ which contradicts the hypothesis $\delta\beta < 1$.

At each iteration, we can use two bounds:

- Lemma 7 guarantees that for any choice of $L \subset \{1, \dots, K\}$, $v \in F_L$, we have, as $\sum_{\ell \in L^c} \omega_\ell(x_n, t) \leq 1$,

$$\begin{aligned}
\|x_{n+1} - \hat{x}\|_2 &\leq \delta\beta\|x_n - \hat{x}\|_2 + \|v\|_2 \delta \left( \sum_{\ell \in L^c} \omega_\ell(x_n, t) \right) + \delta(1+\beta)\|x_n - v\|_2 + \delta t \|\hat{x}\|_2 \\
&\leq \delta\beta\|x_n - \hat{x}\|_2 + \|x_n + v - x_n\|_2 \delta \left( \sum_{\ell \in L^c} \omega_\ell(x_n, t) \right) + \delta(1+\beta)\|x_n - v\|_2 + \delta t \|\hat{x}\|_2 \\
&\leq \delta\beta\|x_n - \hat{x}\|_2 + \|x_n\|_2 \delta \left( \sum_{\ell \in L^c} \omega_\ell(x_n, t) \right) + \delta(2+\beta)\|x_n - v\|_2 + \delta t \|\hat{x}\|_2;
\end{aligned}$$

- Lemma 2 gives that between $x_n$ and $x_{n+1}$, we have the bound

$$\|x_{n+1} - \hat{x}\|_2 \leq \delta\beta\|x_n - \hat{x}\|_2 + \|I - \mu A^T A\|_{\text{op}} \|P^n(x_n) - P_\Sigma(x_n)\|_2.$$

This gives

$$\|x_{n+1} - \hat{x}\|_2 \leq \delta\beta\|x_n - \hat{x}\|_2 + D \min(\|P^n(x_n) - P_\Sigma(x_n)\|_2, \|x_n\|_2 (\sum_{\ell \in L^c} \omega_\ell(x_n, t)) + \|v - x_n\|_2 + t))$$

where $D = \max(\|I - \mu A^T A\|_{\text{op}}, \delta(2+\beta), \delta\|\hat{x}\|_2)$.

We can specifically choose $v_n \in F_{L_n}$ at each step $n$ where $L_n$ and $v_n$ are to be set later. We have

$$\|x_{n+1} - \hat{x}\|_2 \leq \delta\beta\|x_n - \hat{x}\|_2 + \|x_n\|_2 D U_n$$

where $U_n = \min\left( \frac{\|P^n(x_n) - P_\Sigma(x_n)\|_2}{\|x_n\|_2}, (\sum_{\ell \in L_n^c} \omega_\ell(x_n, t)) + \frac{\|v_n - x_n\|_2}{\|x_n\|} + \frac{t}{\|x_n\|_2} \right)$.

Define $[K] = \{1, \dots, K\}$ and $k_n$ such that $P_{E_{k_n}}(x_n) \in P_\Sigma^\perp(x_n)$. Let $\ell_n \in \arg\min_{\ell \neq k_n} \|P_{E_\ell}^\perp(x_n) - x_n\|_2^2 - \|P_{E_{k_n}}^\perp(x_n) - x_n\|_2^2$ , consider $L_n \supset \{k_n, \ell_n\}$.

We have, using Lemma 5, Lemma 8 and the fact that $\|x_n\| \geq b > 0$ for $n \geq n_0$,

$$\begin{aligned}
U_n &= O\left( \min\left( \exp(-\frac{1}{4}\frac{\eta_n^{[K]}}{t}) + t, \exp\left( -\frac{1}{4}\frac{\eta_n^{L_n^c}}{t} \right) + \frac{\|x_n - v_n\|_2}{\|x_n\|_2} + \frac{t}{\|x_n\|_2} \right) \right) \\
&= O\left( \min\left( \exp(-\frac{1}{4}\frac{\eta_n^{[K]}}{t}), \exp\left( -\frac{1}{4}\frac{\eta_n^{L_n^c}}{t} \right) + \|x_n - v_n\|_2 \right) + t \right),
\end{aligned}$$

i.e. there exists $D' > 0$ such that

$$U_n \leq D'(\min(\exp(-\frac{1}{4}\frac{\eta_n^{[K]}}{t}), \exp\left( -\frac{1}{4}\frac{\eta_n^{L_n^c}}{t} \right) + \|x_n - v_n\|_2) + t).$$

Let $c' > 0$, we distinguish two cases:

- if $\exp(-\frac{1}{4}\frac{\eta_n^{[K]}}{t}) \leq c't$ we have

$$U_n \leq D'(c't + t) = D'(1 + c')t;$$

- Otherwise, we have $\exp(-\frac{1}{4}\frac{\eta_n^{[K]}}{t}) > c't$, $\eta_n^{[K]} \leq 4t\log(\frac{1}{c't}) \to_{t \to 0} 0$. Subsequently we choose

$$L_n = \left\{ \ell : \exp\left( -\frac{1}{4}\frac{\|P_{E_\ell}^\perp(x_n) - x_n\|_2^2 - \|P_{E_{k_n}}^\perp(x_n) - x_n\|_2^2}{t} \right) > c't \right\}$$

which contains $\ell_n$ (by definition of $\ell_n$ and using the condition $\exp(-\frac{1}{4}\frac{\eta_n^{[K]}}{t}) > c't$). This guarantees that $\exp\left(-\frac{1}{4}\frac{\eta_n^{L_n^c}}{t}\right) \leq c't$ and

$$U_n \leq D'(c't + \|x_n - v_n\|_2 + t) = D'((1+c')t + \|x_n - v_n\|_2)$$

Notice that $\eta_n^{L_n} = \eta_n^{[K]} \to_{n\to\infty} 0$, taking $v_n \in \arg\min_{v\in F_{L_n}} \|v - x_n\|_2$, with Lemma 9, we have $\lim_{n\to\infty} \|v_n - x_n\|_2 = 0$.

As $t \to 0$, we deduce that any subsequence of $U_n$ converges to 0 and $\lim_{n\to\infty} U_n = 0$.

We have

$$\|x_{n+1} - \hat{x}\|_2 \leq \delta\beta\|x_n - \hat{x}\| + D\|x_n\|U_n \leq (\delta\beta + DU_n)\|x_n - \hat{x}\| + D\|\hat{x}\|U_n$$

with $\lim_{n\to\infty} U_n = 0$. For sufficiently large $n$, $\delta\beta + DU_n \leq \rho < 1$ and as $U_n$ is bounded , e.g. $D\|\hat{x}\|U_n \leq D''$, for some constant $D''$,we have $\|x_{n+1} - \hat{x}\|_2 \leq \rho\|x_n - \hat{x}\| + D''$ and

$$\|x_{n+1} - \hat{x}\|_2 \leq \rho^n\|x_0 - \hat{x}\|_2 + D''\frac{1}{1-\rho}.$$

We deduce that $\|x_n\|_2$ is bounded and $V_n = D\|x_n\|U_n \to_{n\to\infty} 0$.

Similarly to Corollary 3, we deduce that $\lim_{n\to\infty} \|x_{n+1} - \hat{x}\|_2 = 0$ and that there exists $n_1$ large enough such that $n \geq n_1$ implies $x_n \in B(\hat{x}, C) \subset \Omega_\eta$ for some $\eta > 0$ and

$$U_n = O(\exp(-\frac{1}{4}\frac{\eta}{t}) + t). \tag{16}$$

We have, by induction as in Lemma 2,

$$\begin{aligned}\|x_n - \hat{x}\|_2 &\leq \delta\beta\|x_n - \hat{x}\|_2 + V_n \\ &\leq (\delta\beta)^n\|x_n - \hat{x}\|_2 + \sum_{l=0}^{n-1}(\delta\beta)^{n-l-1}V_l\end{aligned}$$

We deduce as in Theorem 1, that there is $C' > 0$ such that

$$\|x_n - \hat{x}\|_2 \leq C'((\sqrt{\delta\beta})^n + \max_{l=\lfloor n/2\rfloor,n} V_l).$$

Using the fact that $V_n = O(U_n)$ and (16), we deduce that there exists $C > 0$ such that

$$\|x_n - \hat{x}\|_2 \leq C((\sqrt{\delta\beta})^n + \max_{l=\lfloor n/2\rfloor,n}(\exp(-\frac{1}{4}\frac{\eta}{t_l}) + t_l)$$

Setting $c = \frac{1}{4}\eta$ and recalling that $t_l = \sigma_l^2$ leads to the result.

□

### 9.4 Influence of subspace geometry on convergence rate

Inspecting the proof of Theorem 2, we see that the distance to the frontier can be shown to go to zero as the quantity $\eta_n^{L_n}$ goes to zero. If it is possible to achieve a bound of the form $d(x_n, F_{L_n}) \leq O(\eta_n^{L_n})$, then we can also control the burn-in period for which we enter a region of the ground-truth $\hat{x}$. This is highly dependent on the geometry of the underlying union-of-subspaces. In certain cases, however, we can show that such a bound indeed holds. A particularly instructive case is when the model set is the union of two orthogonal subspaces.

**Proposition 2.** *Suppose that the underlying model set $\Sigma = E_1 \cup E_2$ where $E_1$ and $E_2$ are orthogonal subspaces in $\mathbb{R}^d$. Then we have that for any $x \in \mathbb{R}^d$,*

$$\min_{v\in F_{\{1,2\}}} \|x - v\|_2^2 \leq \frac{1}{2}\min_{\ell\neq k_x : P^\perp_{E_{k_x}}(x)\in P_\Sigma(x)}(\|x - P^\perp_{E_\ell}(x)\|_2^2 - \|x - P_\Sigma(x)\|_2^2).$$

*Proof of Proposition 2.* Fix $x \neq 0$. If $x \in F_{\{1,2\}}$, the left-hand-side is zero, so we consider the case $x \notin F_{\{1,2\}}$. Suppose, without loss of generality, that $k_x = 1$ so that $P_{E_1}^{\perp}(x) \in P_{\Sigma}(x)$. Our goal is to show that

$$\min_{v \in F_{\{1,2\}}} \|x - v\|_2^2 \leq \frac{1}{2}(\|x - P_{E_2}^{\perp}(x)\|_2^2 - \mathrm{dist}(x, \Sigma)^2). \tag{17}$$

As $E_1$ and $E_2$ are orthogonal subspaces, consider the following orthogonal decomposition of $x$ where $u_i = P_{E_i}^{\perp}(x)$ for $i = 1, 2$:

$$x = u_1 + u_2 + w_{12},\ u_1 \in E_1,\ u_2 \in E_2,\ w_{12} \in (E_1 \oplus E_2)^{\perp}.$$

Then observe that we have $\mathrm{dist}(x, E_i) = \|x - P_{E_i}^{\perp}(x)\|_2 = \|u_j + w_{12}\|_2$ for $j \neq i$, which means

$$\begin{aligned}
\|x - P_{E_2}^{\perp}(x)\|_2^2 - \mathrm{dist}(x, \Sigma)^2 = \|x - P_{E_2}^{\perp}(x)\|_2^2 - \|x - P_{E_1}^{\perp}(x)\|_2^2 &= \|u_1 + w_{12}\|_2^2 - \|u_2 + w_{12}\|_2^2 \\
&= \|u_1\|_2^2 + \|w_{12}\|_2^2 - \|u_2\|_2^2 - \|w_{12}\|_2^2 \\
&= \|u_1\|_2^2 - \|u_2\|_2^2
\end{aligned} \tag{18}$$

where we used the fact that $u_1, u_2 \perp w_{12}$ in the second-to-last equality. For the left-hand-side of (17), note that the frontier admits the following description:

$$\begin{aligned}
F_{\{1,2\}} &= \left\{x \in \mathbb{R}^d : \|x - P_{E_1}^{\perp}(x)\|_2^2 = \|x - P_{E_2}^{\perp}(x)\|_2^2\right\} \\
&= \left\{x \in \mathbb{R}^d : x = u_1 + u_2 + w_{12},\ u_i \in E_i,\ w_{12} \in (E_1 \oplus E_2)^{\perp},\ \|u_2 + w_{12}\|_2^2 = \|u_1 + w_{12}\|_2^2\right\} \\
&= \left\{x \in \mathbb{R}^d : x = u_1 + u_2 + w_{12},\ u_i \in E_i,\ w_{12} \in (E_1 \oplus E_2)^{\perp},\ \|u_2\|_2^2 = \|u_1\|_2^2\right\}.
\end{aligned}$$

Consider the following feasible point $\tilde{v} \in F_{\{1,2\}}$. Let $\gamma := (\|u_1\|_2 + \|u_2\|_2)/2$ and set $\tilde{u}_i = \gamma u_i / \|u_i\|_2 \in E_i$. We then put

$$\tilde{v} := \tilde{u}_1 + \tilde{u}_2 + w_{12} \in F_{\{1,2\}}.$$

We can then bound the distance to the frontier as

$$\begin{aligned}
\min_{v \in F_{\{1,2\}}} \|x - v\|_2^2 \leq \|x - \tilde{v}\|_2^2 &= \|u_1 + u_2 + w_{12} - (\tilde{u}_1 + \tilde{u}_2 + w_{12})\|_2^2 \\
&= \|u_1 - \tilde{u}_1 + u_2 - \tilde{u}_2\|_2^2 \\
&= \|u_1 - \tilde{u}_1\|_2^2 + \|u_2 - \tilde{u}_2\|_2^2 \\
&= (\|u_1\|_2 - \gamma)^2 + (\|u_2\|_2 - \gamma)^2 \\
&= \frac{1}{2}(\|u_1\|_2 - \|u_2\|_2)^2 \\
&\leq \frac{1}{2}(\|u_1\|_2 - \|u_2\|_2)(\|u_1\|_2 + \|u_2\|_2) \\
&= \frac{1}{2}(\|u_1\|_2^2 - \|u_2\|_2^2).
\end{aligned}$$

Recognizing (18) concludes the proof. □