



**HABILITATION À DIRIGER  
DES RECHERCHES**

**DE L'UNIVERSITÉ PSL**

Soutenue à l'École Normale Supérieure

**Statistical Theory for Manifold Learning & Generative  
Modeling**

Soutenue par

**Eddie Aamari**

Le 21 septembre 2026

École doctorale n°386

**Sciences Mathématiques  
de Paris Centre**

Spécialité

**Mathématiques**

Composition du jury :

|                                                         |                     |
|---------------------------------------------------------|---------------------|
| Benoît Cadre<br>Professeur, U. de Rennes                | <i>Président</i>    |
| Arnak Dalalyan<br>Professeur, ENSAE                     | <i>Rapporteur</i>   |
| Élisabeth Gassiat<br>Professeure, U. Paris-Saclay       | <i>Examinatrice</i> |
| Gabriel Peyré<br>Directeur de recherche, ENS – PSL      | <i>Examineur</i>    |
| Markus Reiß<br>Professor, Humboldt-U. zu Berlin         | <i>Rapporteur</i>   |
| Judith Rousseau<br>Professeure, U. Paris Dauphine – PSL | <i>Examinatrice</i> |
| Nick Whiteley<br>Professor, U. of Bristol               | <i>Rapporteur</i>   |



# Contents

|                                                                                  |            |
|----------------------------------------------------------------------------------|------------|
| <b>Acknowledgments</b>                                                           | <b>iii</b> |
| <b>Publications &amp; pre-publications</b>                                       | <b>iv</b>  |
| <b>Foreword</b>                                                                  | <b>v</b>   |
| <b>1 Geometric inference</b>                                                     | <b>1</b>   |
| 1.1 The reach and its geometric consequences . . . . .                           | 1          |
| 1.1.1 Reach as generalized convexity . . . . .                                   | 2          |
| 1.1.2 Equivalent formulations for submanifolds . . . . .                         | 2          |
| 1.1.3 The local vs. global dichotomy . . . . .                                   | 4          |
| 1.1.4 Properties conferred by a positive reach . . . . .                         | 4          |
| 1.2 Boundaryless manifold inference under reach condition . . . . .              | 8          |
| 1.2.1 Local geometry recovery . . . . .                                          | 8          |
| 1.2.2 Global geometry recovery . . . . .                                         | 12         |
| 1.3 Beyond the closed manifold model . . . . .                                   | 16         |
| 1.3.1 Manifolds with boundary . . . . .                                          | 16         |
| 1.3.2 Robustness and adversarial noise . . . . .                                 | 18         |
| 1.3.3 The union of manifolds hypothesis: stratifications . . . . .               | 20         |
| 1.4 Open directions . . . . .                                                    | 23         |
| 1.4.1 Implicit representations for geometric inference . . . . .                 | 23         |
| 1.4.2 A few conjectures . . . . .                                                | 24         |
| <b>2 Statistical foundations of generative modeling</b>                          | <b>27</b>  |
| 2.1 Historical perspective and statistical foundations . . . . .                 | 27         |
| 2.1.1 Beyond explicit density estimation . . . . .                               | 27         |
| 2.1.2 Main paradigms in generative modeling . . . . .                            | 28         |
| 2.1.3 Statistical foundations . . . . .                                          | 28         |
| 2.2 Generative adversarial networks as minimax distribution estimators . . . . . | 29         |
| 2.2.1 Integral probability metrics and the Wasserstein formulation . . . . .     | 29         |
| 2.2.2 The low-dimensional hypothesis and the statistical model . . . . .         | 30         |
| 2.2.3 Bias-variance trade-off in adversarial learning . . . . .                  | 31         |

|       |                                                                                    |    |
|-------|------------------------------------------------------------------------------------|----|
| 2.2.4 | From smooth discrimination to rougher IPMs: the interpolation inequality . . . . . | 32 |
| 2.2.5 | Constructive neural network architectures via wavelets . . . . .                   | 34 |
| 2.2.6 | Minimax optimality of the GAN estimator . . . . .                                  | 34 |
| 2.3   | Score-based generative models and diffusions . . . . .                             | 35 |
| 2.3.1 | Forward dynamics and marginal reversals . . . . .                                  | 35 |
| 2.3.2 | Denoising score matching . . . . .                                                 | 36 |
| 2.3.3 | The statistical model and the regularity of the score . . . . .                    | 37 |
| 2.3.4 | Quantitative SDE stability and the score-matching oracle inequality                | 39 |
| 2.3.5 | Minimax convergence rates for SGMs . . . . .                                       | 40 |
| 2.4   | Generator matching for Markovian generative models . . . . .                       | 41 |
| 2.4.1 | Probability paths . . . . .                                                        | 42 |
| 2.4.2 | Generators realizing a probability path . . . . .                                  | 43 |
| 2.4.3 | The conditional-to-marginal principle . . . . .                                    | 44 |
| 2.4.4 | Linear parameterizations of generators . . . . .                                   | 45 |
| 2.4.5 | Bregman matching . . . . .                                                         | 46 |
| 2.4.6 | The design space of Markov generators . . . . .                                    | 48 |
| 2.4.7 | Duality and stability . . . . .                                                    | 49 |
| 2.5   | Adaptive diffusion guidance . . . . .                                              | 51 |
| 2.5.1 | Classifier-free guidance and the tempering myth . . . . .                          | 51 |
| 2.5.2 | Rigorous probabilistic guarantees: the guiding field . . . . .                     | 52 |
| 2.5.3 | Adaptive guidance via stochastic optimal control . . . . .                         | 54 |
| 2.5.4 | Scalable optimization via the adjoint sensitivity method . . . . .                 | 54 |
| 2.6   | Perspectives . . . . .                                                             | 55 |

# Acknowledgments

Je remercie chaleureusement Arnak Dalalyan, Markus Reiß et Nick Whiteley d'avoir accepté d'être les rapporteurs de ce mémoire. Le temps qu'ils ont consacré à sa lecture et le soin apporté à leurs rapports m'ont beaucoup touché. Je remercie également Benoît Cadre, Élisabeth Gassiat, Gabriel Peyré et Judith Rousseau d'avoir accepté de me cuisiner au sein de ce jury. Benoît, c'est dans ton cours que j'ai découvert la statistique et que j'y ai pris goût. Merci d'avoir accepté de présider ce jury et d'assumer ainsi jusqu'au bout ta part de responsabilité.

Merci aux collègues des équipes administratives du LPSM, du DMA et d'ailleurs pour leur force et leur précision face à des procédures dont la régularité en temps reste, à ma connaissance, un problème ouvert.

Une mention particulière pour celles et ceux avec qui un sujet mal compris finit volontiers en groupe de travail. Je n'ai encore rien trouvé de plus plaisant ni de plus efficace pour avancer.

En cours comme en recherche, les étudiantes et étudiants m'ont beaucoup appris. Leurs questions m'obligent bien souvent à préciser ce que j'aurais préféré laisser vague. Les fonctions d'enseignant ou d'encadrant ne disent pas toujours qui apprend de qui.

Les projets m'ont aussi fourni d'excellents prétextes pour voyager et reprendre à plusieurs des raisonnements commencés seul, avec des discussions qui se prolongent loin du tableau sur des sujets qui ne figuraient pas au programme. Certaines de ces rencontres auraient eu bien peu de chances ( $p < 10^{-6}$ ) d'avoir lieu sans les mathématiques. Preuve, s'il en fallait, de leur utilité sociale.

Ces remerciements vont aussi à toutes celles et ceux avec qui j'ai pu discuter d'affaires scientifiques et de préoccupations plus terre à terre. Il se trouve que les premières occupent davantage de pages dans ce mémoire que les secondes. Qu'on n'en tire aucune conclusion quant à leur importance relative.

J'éprouve enfin un plaisir certain, et un soulagement non dénué de fierté, à ne développer désormais que des problèmes de reach. À ma famille.

# Publications & pre-publications

## Works presented in this manuscript

- [AS26] **Aamari, Eddie**; Stéphanovitch, Arthur. *Statistical Analysis of Markovian Generative Modeling*. 2026. [arXiv:2604.22712](#).
- [SAL25] Stéphanovitch, Arthur; **Aamari, Eddie**; Levrard, Clément. *Generalization bounds for score-based generative models: a synthetic proof*. 2025. [doi:10.48550/arXiv.2507.04794](#); [arXiv:2507.04794](#).
- [APL<sup>+</sup>25] Azangulov, Iskander; Potapchik, Peter; Li, Qinyu; **Aamari, Eddie**; Deligiannidis, George; Rousseau, Judith. *Adaptive Diffusion Guidance via Stochastic Optimal Control*. 2025. [doi:10.48550/arXiv.2505.19367](#); [arXiv:2505.19367](#).
- [AB24] **Aamari, Eddie**; Berenfeld, Clément. *A theory of stratification learning*. 2024. [doi:10.48550/arXiv.2405.20066](#); [arXiv:2405.20066](#).
- [SAL24] Stéphanovitch, Arthur; **Aamari, Eddie**; Levrard, Clément. *Wasserstein GANs are minimax optimal distribution estimators*. The Annals of Statistics **52**(5), 2167–2193, 2024. [doi:10.1214/24-AOS2430](#); [arXiv:2311.18613](#).
- [ABL23] **Aamari, Eddie**; Berenfeld, Clément; Levrard, Clément. *Optimal reach estimation and metric learning*. The Annals of Statistics **51**(3), 1086–1108, 2023. [doi:10.1214/23-AOS2281](#); [arXiv:2207.06074](#).
- [AAL23] **Aamari, Eddie**; Aaron, Catherine; Levrard, Clément. *Minimax boundary estimation and estimation with boundary*. Bernoulli **29**(4), 3334–3368, 2023. [doi:10.3150/23-BEJ1585](#); [arXiv:2108.03135](#).
- [AK22] **Aamari, Eddie**; Knop, Alexander. *Adversarial Manifold Estimation*. Foundations of Computational Mathematics **24**(1), 1–97, 2024. [doi:10.1007/s10208-022-09588-2](#); [arXiv:2011.04259](#).
- [AK21] **Aamari, Eddie**; Knop, Alexander. *Statistical query complexity of manifold estimation*. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC 2021)*, 116–122, 2021. [doi:10.1145/3406325.3451135](#); [arXiv:2011.04259](#).
- [AL19] **Aamari, Eddie**; Levrard, Clément. *Nonasymptotic rates for manifold, tangent space and curvature estimation*. The Annals of Statistics **47**(1), 177–204, 2019. [doi:10.1214/18-AOS1685](#); [arXiv:1705.00989](#).
- [AKC<sup>+</sup>19] **Aamari, Eddie**; Kim, Jisu; Chazal, Frédéric; Michel, Bertrand; Rinaldo, Alessandro; Wasserman, Larry. *Estimating the reach of a manifold*. Electronic Journal of Statistics **13**(1), 1359–1399, 2019. [doi:10.1214/19-EJS1551](#); [arXiv:1705.04565](#).
- [AL18] **Aamari, Eddie**; Levrard, Clément. *Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction*. Discrete & Computational Geometry **59**(4), 923–971, 2018. [doi:10.1007/s00454-017-9962-z](#); [arXiv:1512.02857](#).

## Other works

- [Aam26] **Aamari, Eddie**. *Seconder of the vote of thanks to Whiteley et al. and Contribution to the Discussion of “Statistical exploration of the manifold hypothesis”*. Journal of the Royal Statistical Society Series B: Statistical Methodology **88**(2), 388–389, 2026. [doi:10.1093/jrsssb/qkag040](#).
- [AACB24] **Aamari, Eddie**; Arias-Castro, Ery; Berenfeld, Clément. *The Coreness and H-Index of Random Geometric Graphs*. ALEA, Latin American Journal of Probability and Mathematical Statistics **21**, 963–983, 2024. [doi:10.30757/ALEA.v21-39](#); [arXiv:2105.03122](#).

# Foreword

High-dimensional data are now collected in ambient spaces  $\mathbb{R}^D$  whose dimension may be large. This is the case in image analysis, genomics, neuroscience, sensor networks, natural language processing, and many other fields. From a classical non-parametric viewpoint, such a setting requires further structure: the sample size needed for density estimation, regression, classification, or support recovery typically deteriorates exponentially with the ambient dimension. This phenomenon, often referred to as the *curse of dimensionality* [Bel61, Sto85], is not merely a technical issue, but a statistical limitation to unstructured high-dimensional inference.

One way to address this difficulty is to assume that the data-generating mechanism has a lower effective dimension than the ambient one. In parametric statistics, this idea often appears as sparsity: after choosing a coordinate system, only a few coordinates are relevant. In the non-parametric problems considered in this memoir, the analogous assumption is geometric and coordinate-free. The object to be learned has a small *intrinsic dimension*, measuring the number of local degrees of freedom, and some *intrinsic regularity*, describing how the local structure varies. The manifold hypothesis is one such model: the data are assumed to concentrate on, or near, a low-dimensional geometric object whose dimension is much smaller than that of the ambient representation.

This viewpoint is natural because it does not prescribe chosen ambient coordinates. A manifold encodes nonlinear constraints and correlations among the observed variables: locally, the data vary along a few directions and are constrained in the remaining ones, but these directions may rotate from point to point. Thus, the manifold hypothesis can be seen as a non-parametric, coordinate-free version of sparsity. It can describe curved supports, implicit constraints, and local charts, while allowing rates governed by intrinsic rather than ambient dimension.

There is also empirical motivation for such a model. In image datasets, variation may be driven by a small number of physical parameters such as pose, illumination, or camera position; in single-cell transcriptomics, high-dimensional expression profiles may reflect low-dimensional differentiation processes; in population genetics, genotype data may reveal geographic structure; and in neuroscience, neural activity has been observed to organize along low-dimensional topological patterns [GBS05, PSW<sup>+</sup>18, NJB<sup>+</sup>08, LLN<sup>+</sup>08, GHP<sup>+</sup>22]. Further examples arise in wireless sensor networks, visual speech recognition, drug discovery, RNA sequencing, and human motion synthesis [PH04, BO95, RS12, MSIB<sup>+</sup>18, LE06]. Such examples do not imply that all high-dimensional data lie on smooth manifolds, but they suggest that many observed datasets inherit geometry from lower-dimensional latent mechanisms.

Recent work also provides a structural explanation for this phenomenon. Whiteley [WGRD24] shows that manifold-like structure can emerge from latent metric models, where high-dimensional observations are indexed by an underlying latent space. In this setting, the geometry visible in the data is not imposed directly as a geometric axiom; it is produced by correlations across the latent domain. Under mild assumptions, the observed cloud retains the topology of the latent space, and under stronger stationarity-type assumptions it can even retain its metric structure, up to controlled deformation. This provides a statistical motivation for the manifold hypothesis: low-dimensional geometry may arise from basic ingredients such as latent variables, correlation, smoothness, and stationarity, rather than from an ad hoc geometric postulate [WGRD24, Aam26].

The purpose of this memoir is to present several contributions to the statistical analysis of such structured high-dimensional models. The guiding question is the following.

*Under which structural and regularity assumptions can features of a distribution be estimated, reconstructed, or generated, and when is the statistical complexity governed by intrinsic rather than ambient dimension?*

This question is broad. It appears in different forms depending on the object of interest. In geometric inference, the unknown object may be a support, a tangent bundle, a curvature tensor, a reach, a boundary, an intrinsic metric, or a stratification. In generative modeling, it may instead be a generator, a score field, Markovian dynamics, or a controlled probability path. In both cases, a similar statistical tension appears: one has to identify a representation that is rich enough to describe the object, but stable enough to be estimable non-asymptotically, i.e. from finitely many observations.

Here, the low-dimensional structure itself is unknown, and part of the statistical problem is to determine which of its features are estimable, at which rates, and under what regularity assumptions.

This habilitation thesis is organized into two parts.

- Chapter 1 is devoted to *geometric inference*.
- Chapter 2 is devoted to *generative modeling*.

This separation is expository. The common theme is the role played by intrinsic dimension and intrinsic regularity in non-parametric inference. In the first part, geometry is estimated explicitly. In the second, it is encoded implicitly through transport maps, vector fields, or stochastic dynamics.

## Geometric inference

The first part concerns estimating geometric structures from point clouds. A standard explanation for the success of many high-dimensional learning methods is the *manifold hypothesis*: observations in  $\mathbb{R}^D$  are assumed to concentrate on, or near, a  $d$ -dimensional submanifold  $M \subset \mathbb{R}^D$ , with  $d \ll D$  [BCV13, WGRD24]. In this setting, the support  $M$  represents the nonlinear constraints satisfied by the data, while  $d$  represents the effective number of local degrees of freedom.

The statistical problem is then to recover information on  $M$  from a finite sample. Depending on the question, this may mean estimating  $M$  in Hausdorff distance, recovering its topology, estimating tangent spaces, curvature, intrinsic distances, or global regularity parameters. A key parameter is the reach of Federer [Fed59]. Positive reach simultaneously controls curvature and rules out near self-intersections. It is therefore a scale parameter for geometric and topological stability, and appears in manifold estimation and homology inference [GPPVW12, NSW08, KZ15].

Several of the works presented here study what can be estimated under reach-type assumptions. In [AL18], Tangential Delaunay Complexes are made stable under statistical perturbations and yield minimax-optimal manifold reconstruction of  $\mathcal{C}^2$  submanifolds. Because of the instability of Delaunay constructions, the focus then shifts in [AK22, AK21] from sample complexity to access models, robustness, privacy, and statistical query complexity. In [AL19], non-asymptotic rates are derived for the estimation of the manifold, its tangent spaces, and its curvature under higher-order  $\mathcal{C}^k$  smoothness assumptions. The reach itself is studied as an unknown parameter in [AKC<sup>+</sup>19], where its estimation is related to a decomposition into local curvature and global bottleneck effects. In [ABL23], reach estimation is connected with intrinsic metric learning, showing that some global geometric quantities can be replaced by metric surrogates.

The assumptions of no boundary and positive reach are restrictive, and the chapter then turns to models where they are weakened. Manifolds with boundary are considered in [AAL23], where boundary detection and support reconstruction are treated jointly. Finally, [AB24] studies unions of manifolds with possible intersections and varying dimensions. In such models, the global reach may vanish, and several classical geometric quantities are no longer identifiable. Still, dimension-wise clustering and reconstruction remain possible under local curvature and volume assumptions.

Thus, this chapter is organized as a progression from smooth manifolds with positive reach to less regular geometric models. The aim is not to promote a single geometric assumption, but rather to clarify which assumptions make which statistical tasks possible.

## Generative modeling

The second part concerns generative modeling and distribution estimation. Here the geometry of the data is no longer necessarily reconstructed as a set. It can instead be represented by a map pushing forward a simple latent distribution, by a score field along a diffusion path, or by the infinitesimal generator of a Markov process. The target is then represented implicitly: one estimates a mechanism that produces samples from the target distribution.

Depending on the model, the relevant complexity may be governed by latent dimension, ambient dimension, smoothness, or the stability of the associated dynamics. The same question therefore reappears: which representation makes the distribution both expressive enough for modeling and stable enough for estimation from finitely many samples?

In [SAL24], Wasserstein GANs are studied as distribution estimators. The analysis

shows how minimax rates depend on the regularity of the target distribution and on the integral probability metric used to compare distributions. This provides a connection between geometric ideas and generative methods: the generator is an implicit low-dimensional representation of the distribution.

Score-based generative models give another connection. In these models, the target distribution is approximated by simulating a learned reverse-time dynamics, and the main statistical object is the score of the intermediate laws. In [SAL25], minimax guarantees are obtained by combining regularity properties of the score with stability estimates for stochastic differential equations. The same dynamical viewpoint is pursued in [APL<sup>+</sup>25], where diffusion guidance is formulated as a stochastic optimal control problem, and in [AS26], where Markovian generative procedures are studied through their infinitesimal generators.

These works suggest that generative modeling can be viewed as an implicit form of geometric inference: rather than estimating the data geometry directly, one estimates dynamics whose trajectories encode aspects of it.

# Chapter 1

## Geometric inference

This chapter studies the statistical recovery of geometric structure from point clouds. The question is which geometric quantities of an unknown low-dimensional support can be estimated at rates governed by its intrinsic dimension and regularity. Bridging the gap between the manifold hypothesis and noisy, discrete observations is usually called *geometric inference*. The goal can be to estimate the unknown support  $M$ , its topology, its intrinsic metric, or its differential quantities from a finite sample  $X_1, \dots, X_n \in \mathbb{R}^D$ , while establishing statistical limits of these tasks.

To make geometric inference statistically well-posed, bounding the curvature is insufficient; one also needs to rule out near self-intersections. A regularity parameter combining these constraints is the *reach*. A strictly positive reach  $\tau_M > 0$  acts as a generalized convexity parameter: it rules out narrow bottlenecks and uniformly bounds the maximal curvature, which is used in minimax manifold estimation.

This chapter tracks a progressive relaxation of these geometric and statistical constraints to better reflect less idealized data models. We begin within the classical framework of smooth boundaryless manifolds with positive reach, establishing minimax optimal estimators for local differential quantities and global metric properties. We then expand the theory to handle manifolds with boundaries, adversarial noise in the Statistical Query (SQ) framework, and finally stratifications of manifolds with mixed dimensions and intersections.

### 1.1 The reach and its geometric consequences

Among geometric constraints, the *reach* is one of the main regularity and scale parameters. It provides a generalized notion of convexity, restricts the maximal curvature of the manifold, and precludes the existence of arbitrarily narrow bottleneck structures. In this section, we review the origin of the reach, present several equivalent formulations and its main geometric properties, and detail why it is used in geometric inference, notably for deriving minimax rates.

### 1.1.1 Reach as generalized convexity

The concept of the reach was first introduced by Herbert Federer in his paper [Fed59] within the framework of geometric measure theory. Federer sought to generalize the classic Steiner formula—which expresses the volume of a tubular neighborhood of a convex set as a polynomial in the radius—to a larger class of non-convex sets. To achieve this, it was necessary to formalize a class of sets that behave *locally* like convex sets, meaning that points in their immediate ambient neighborhood possess a unique nearest metric projection on the set itself.

To formalize this notion, one must first define the medial axis, which identifies the ambient locations where the uniqueness of the projection is no longer directly applicable.

**Definition 1.1** (Medial Axis [Fed59]). Let  $A \subset \mathbb{R}^D$  be a closed subset. The *medial axis* of  $A$ , denoted by  $\text{Med}(A)$ , is defined as the set of points in the ambient space that have at least two distinct nearest neighbors on  $A$ . Formally,

$$\text{Med}(A) = \left\{ z \in \mathbb{R}^D \mid \exists p \neq q \in A, \|p - z\| = \|q - z\| = \text{dist}(z, A) \right\}.$$

The reach is subsequently defined as the minimal distance separating the set  $A$  from its medial axis.

**Definition 1.2** (Reach [Fed59]). The *reach* of a closed subset  $A \subset \mathbb{R}^D$ , denoted by  $\tau_A$ , is defined as:

$$\tau_A = \inf_{p \in A} \text{dist}(p, \text{Med}(A)) = \inf_{z \in \text{Med}(A)} \text{dist}(z, A).$$

By construction, the metric projection map  $\pi_A(x) = \arg \min_{p \in A} \|p - x\|$  onto  $A$  is well-defined and unique strictly outside the medial axis  $\text{Med}(A)$ . Consequently, the reach  $\tau_A$  corresponds to the largest radius  $r \geq 0$  such that  $\pi_A$  is uniquely defined on the entire open  $r$ -offset

$$A^r := \left\{ x \in \mathbb{R}^D \mid \text{dist}(x, A) \leq r \right\}.$$

The reach acts as a generalized convexity parameter: a closed subset  $A \subset \mathbb{R}^D$  is convex if and only if  $\tau_A = \infty$  [Fed59]. For general sets, possessing a strictly positive reach ( $\tau_A \geq \tau_{\min} > 0$ ) is often used as a regularity condition in geometric measure theory and integral geometry, allowing the definitions of generalized normal cones, tangent cones, and curvature measures [Thä08].

### 1.1.2 Equivalent formulations for submanifolds

While Definition 1.2 applies broadly to any closed subset, its reliance on the global medial axis, which is unstable under perturbations, is less convenient in statistical contexts. When  $M \subset \mathbb{R}^D$  is a smooth submanifold, Federer established an equivalent formulation that only involves the points of  $M$  and its tangent spaces [Fed59].

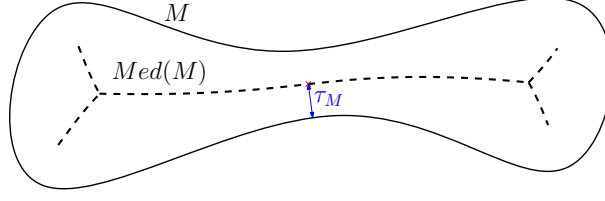


Figure 1.1: Illustration of the medial axis (dashed line) and the reach  $\tau_M$  of a closed planar curve  $M$ . The reach is the largest tubular radius below which every ambient point has a unique nearest point on  $M$ .

**Theorem 1.3** (Tangent Space Formulation [Fed59]). *For any  $\mathcal{C}^2$ -submanifold  $M \subset \mathbb{R}^D$ , the reach can be formulated equivalently as:*

$$\tau_M = \inf_{q \neq p \in M} \frac{\|q - p\|^2}{2 \operatorname{dist}(q - p, T_p M)}, \quad (1.1)$$

where  $T_p M$  denotes the tangent space of  $M$  at  $p$ .

This formulation asserts that  $\tau_M \geq \tau_{\min} > 0$  if and only if  $\operatorname{dist}(q - p, T_p M) \leq \frac{\|q - p\|^2}{2\tau_{\min}}$  for all  $p, q \in M$ . Geometrically, the ratio in (1.1) represents the radius of an ambient Euclidean ball that is tangent to  $M$  at  $p$  and passes exactly through  $q$  (see Figure 1.2). Thus, the reach bounds how fast the submanifold deviates from its best linear approximation.

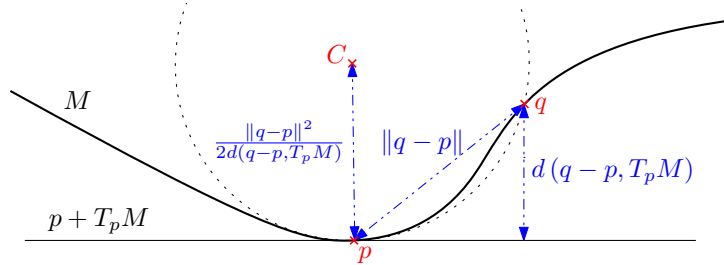


Figure 1.2: Geometric interpretation of the ratio  $\frac{\|q - p\|^2}{2 \operatorname{dist}(q - p, T_p M)}$ . It exactly corresponds to the radius of an ambient ball tangent to  $M$  at  $p$  whose boundary passes through  $q$ .

More recently, [BLW19] established a purely metric characterization of the reach based on geodesic distortion. Let  $d_M(p, q)$  denote the intrinsic geodesic distance between  $p$  and  $q$  on  $M$ . The reach controls the maximum scale at which the intrinsic metric is tightly controlled by the ambient Euclidean metric, specifically comparing it to the metric of a sphere.

**Theorem 1.4** (Metric Distortion Characterization [BLW19]). *Let  $M \subset \mathbb{R}^D$  be a closed submanifold. The reach of  $M$  is the supremum over all radii  $r > 0$  such that the intrinsic distance is bounded by the spherical distance at scale  $r$ :*

$$\tau_M = \sup \left\{ r > 0 \mid \forall p, q \in M, \|p - q\| < 2r \implies d_M(p, q) \leq 2r \operatorname{asin} \left( \frac{\|p - q\|}{2r} \right) \right\}.$$

This metric distortion framework does not rely on the ambient space's coordinate system, characterizing the reach through the relation between the intrinsic metric and the ambient embedding. This becomes particularly useful in graph-based manifold learning algorithms that operate directly on distance matrices (e.g., Isomap [TDSL00]).

### 1.1.3 The local vs. global dichotomy

By inspecting the infimum in Theorem 1.3, one sees that the reach is driven by two distinct geometric regimes: points that are infinitesimally close, and points that are macroscopically far apart. This dichotomy is formalized in the following structural theorem, establishing that the reach arises either from a global bottleneck or from maximal local curvature.

**Theorem 1.5** (Global vs. Local Reach Dichotomy [AKC<sup>+</sup>19]). *Let  $M \subset \mathbb{R}^D$  be a compact submanifold with reach  $\tau_M > 0$ . At least one of the following assertions holds:*

- **Global Case (Bottleneck):**  $M$  has a reach-attaining bottleneck pair  $(q_1, q_2) \in M^2$ . That is, there exists  $z_0 \in \text{Med}(M)$  such that  $q_1, q_2 \in \partial B(z_0, \tau_M)$  and  $\|q_1 - q_2\| = 2\tau_M$ .
- **Local Case (Curvature):** There exists  $q_0 \in M$  and an arc-length parametrized geodesic  $\gamma_0$  such that  $\gamma_0(0) = q_0$  and  $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$ .

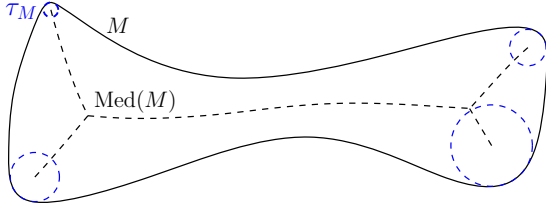


Figure 1.3: The local case: small reach  $\tau_M$  arises from an area of high curvature (corresponding to a small minimal radius of curvature).

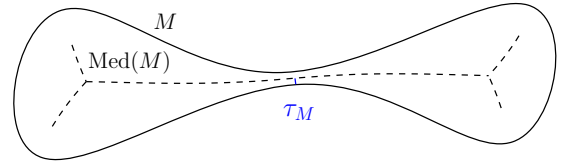


Figure 1.4: The global case: small reach  $\tau_M$  arises from a narrow bottleneck structure where  $M$  nearly self-intersects.

The distinction between local and global reach (Figure 1.3 and Figure 1.4) affects the statistical complexity of parameter estimation. Minimax optimal rates for estimating  $\tau_M$  depend on which regime determines the reach. We establish in [ABL23] that this yields slower rates for curvature-driven reach and faster rates for bottleneck-driven reach, which motivates local-global adaptive estimators.

### 1.1.4 Properties conferred by a positive reach

Submanifolds satisfying  $\tau_M \geq \tau_{\min} > 0$  satisfy several quantitative regularity guarantees. These guarantees are metric and coordinate-free, and apply uniformly across the manifold, allowing us to view the class of bounded-reach manifolds as a geometric analog to the Hölder class  $\mathcal{C}^2(1/\tau_{\min})$  for classical functional regularity, supplemented by a global separation condition preventing near self-intersections. We group these properties into several

categories: differential bounds, metric equivalences, properties of the metric projection, local topology, volume measure behavior, and stability constraints.

### Differential and curvature bounds

By analyzing the local limit of Theorem 1.3, one obtains uniform bounds on the principal curvatures of  $M$ . Let  $\mathbb{I}_p^M : T_p M \times T_p M \rightarrow (T_p M)^\perp$  denote the second fundamental form of  $M$  at  $p$ .

**Proposition 1.6** (Curvature Bound [NSW08]). *Let  $M \subset \mathbb{R}^D$  be a compact  $d$ -dimensional submanifold with reach  $\tau_M \geq \tau_{\min} > 0$ . For all  $p \in M$  and all unit vectors  $v \in T_p M$ :*

$$\|\mathbb{I}_p^M(v, v)\| \leq \frac{1}{\tau_{\min}}.$$

*Equivalently, if  $\gamma$  is any arc-length parametrized geodesic of  $M$ , then  $\|\gamma''(t)\| \leq \frac{1}{\tau_{\min}}$  for all  $t$ .*

A direct application of the Gauss equation to Proposition 1.6 bounds the sectional curvatures  $\kappa$  of  $M$ , which gives  $-\frac{2}{\tau_{\min}^2} \leq \kappa \leq \frac{1}{\tau_{\min}^2}$ . This limits the intrinsic warping of the manifold and provides a strict lower bound on the injectivity radius, which is guaranteed to be at least  $\pi\tau_{\min}$  [AB06].

### Metric equivalence: extrinsic vs. intrinsic distances

The absence of bottlenecks prevents the manifold from folding back on itself. Therefore, locally, the ambient Euclidean distance  $\|p - q\|$  is locally comparable to the intrinsic geodesic distance  $d_M(p, q)$ .

**Lemma 1.7** (Metric Equivalence [NSW08]). *Let  $M \subset \mathbb{R}^D$  be a compact submanifold with reach  $\tau_M \geq \tau_{\min} > 0$ . For all  $p, q \in M$  such that  $\|p - q\| \leq \tau_{\min}/2$ , the geodesic distance  $d_M(p, q)$  is uniquely defined, and we establish:*

$$\|q - p\| \leq d_M(p, q) \leq \tau_{\min} \left( 1 - \sqrt{1 - \frac{2\|p - q\|}{\tau_{\min}}} \right).$$

Lemma 1.7 yields  $d_M(p, q) \leq \|p - q\| + \frac{\|p - q\|^2}{2\tau_{\min}} + \mathcal{O}(\|p - q\|^3)$ . This equivalence is important for geometric algorithms (such as Isomap [TDSL00] or tangential Delaunay complexes [AL18]) that rely on ambient Euclidean distances as proxies for the unobservable intrinsic metric.

### Lipschitz continuity of the metric projection

The reach quantifies the maximal open neighborhood of a set on which the metric projection is uniquely defined. For any  $r > 0$ , we recall that  $A^r = \{x \in \mathbb{R}^D \mid \text{dist}(x, A) \leq r\}$  denote the open  $r$ -offset of  $A$ . When the reach is positive, Federer proved that the projection map is not only unique on  $A^r$  for  $r < \tau_A$ , but it is also locally Lipschitz.

**Proposition 1.8** (Lipschitz Projection [Fed59]). *Let  $A \subset \mathbb{R}^D$  be a closed set with positive reach  $\tau_A > 0$ . For any ambient radius  $r < \tau_A$ , the metric projection  $\pi_A : A^r \rightarrow A$  is globally Lipschitz. Specifically, for any two points  $x, y \in A^r$ , we have:*

$$\|\pi_A(x) - \pi_A(y)\| \leq \left( \frac{\tau_A}{\tau_A - r} \right) \|x - y\|.$$

In statistical frameworks where data is corrupted by ambient noise, this Lipschitz bound ensures that projecting points back onto the manifold does not arbitrarily inflate the noise. Furthermore, this regularity yields that the squared distance function  $x \mapsto \text{dist}^2(x, A)$  is continuously differentiable on the open tubular neighborhood  $A^{\tau_A}$ , with gradient  $\nabla \text{dist}^2(x, A) = 2(x - \pi_A(x))$ .

### Local topology and contractibility

A positive reach imposes geometric constraints that affect the local and global topology of the manifold. First, it implies that small tubular neighborhoods  $M^r$  ( $r < \tau_M$ ) are topologically equivalent to  $M$ , since  $M^r$  strongly deformation retracts onto  $M$  via the straight-line homotopy  $H(x, t) = (1 - t)x + t\pi_M(x)$ . Moreover, the intersections of  $M$  with small ambient Euclidean balls are topologically trivial.

**Lemma 1.9** (Local Contractibility [NSW08]). *Let  $M \subset \mathbb{R}^D$  be a compact submanifold with reach  $\tau_M \geq \tau_{\min} > 0$ . For any  $x \in \mathbb{R}^D$  and any radius  $r < \sqrt{3/5}\tau_{\min}$ , the ambient intersection  $M \cap B(x, r)$  is either empty or strongly contractible. In particular, for any  $p \in M$  and  $r \leq \tau_{\min}/2$ , the ambient intersection  $M \cap B(p, r)$  is simply connected.*

This contractibility property is used throughout topological data analysis. It guarantees that suitable covers of  $M$  by ambient Euclidean balls form a *good cover*. Invoking the Nerve Theorem, the Čech or Vietoris-Rips complexes built from a sufficiently dense sample will then recover the homology of the underlying manifold  $M$  [NSW08].

### Volume bounds and $(a, d)$ -standardness

The reach also restricts the distortion of the volume measure on the manifold. By bounding the second fundamental form, the reach controls the Jacobian of the exponential map  $\exp_p : T_p M \rightarrow M$ , relating the intrinsic Riemannian measure to the Lebesgue measure of the tangent space.

**Proposition 1.10** (Jacobian and Volume Measure [AL19, AAL23]). *Let  $M \subset \mathbb{R}^D$  be a  $d$ -dimensional compact submanifold with reach  $\tau_M \geq \tau_{\min} > 0$ . For any  $p \in M$ , let  $A_v = d_v \exp_p$  denote the differential of the exponential map at a vector  $v \in T_p M$ . If  $\|v\| \leq \frac{\pi\tau_{\min}}{2\sqrt{2}}$ , the Jacobian  $J(v) = \sqrt{\det(A_v^\top A_v)}$  satisfies:*

$$\left(1 - \frac{\|v\|^2}{6\tau_{\min}^2}\right)^d \leq J(v) \leq \left(1 + \frac{\|v\|^2}{\tau_{\min}^2}\right)^d.$$

Consequently, there exist absolute constants  $c_d, C_d > 0$ , depending only on the intrinsic dimension  $d$ , such that for all  $p \in M$  and all  $r \leq \tau_{\min}/4$ :

$$c_d r^d \leq \mathcal{H}^d(M \cap B(p, r)) \leq C_d r^d.$$

This property, often referred to as  $(a, d)$ -standardness, prevents singularities in the mass distribution of  $M$ . If the data-generating distribution  $P$  admits a density bounded away from zero with respect to  $\mathcal{H}^d$ , Proposition 1.10 implies that the probability mass of any ambient ball  $B(p, r)$  scales precisely as  $\Theta(r^d)$ . This standardness condition is common in non-parametric statistics on manifolds, as it controls the covering and packing numbers of  $M$ , thereby determining the optimal sampling density and convergence rates of estimators.

### Stability properties

As seen in Proposition 1.6, the inverse reach  $1/\tau_M$  plays the role of a curvature bound. It is stable under smooth changes of ambient coordinates in the following quantitative sense.

**Lemma 1.11** (Stability of reach under ambient diffeomorphisms [Fed59]). *Let  $A \subset \mathbb{R}^D$  be closed, with positive reach  $\tau_A > 0$ , and let  $\Phi : \mathbb{R}^D \rightarrow \mathbb{R}^D$  be a  $\mathcal{C}^1$ -diffeomorphism. Assume that  $\Phi$ ,  $\Phi^{-1}$ , and  $d\Phi$  are Lipschitz, with respective constants  $K$ ,  $N$ , and  $R$ . Then*

$$\tau_{\Phi(A)} \geq \frac{1}{N^2 (K/\tau_A + R)}.$$

Equivalently,

$$\frac{1}{\tau_{\Phi(A)}} \leq N^2 \left( K \frac{1}{\tau_A} + R \right).$$

This estimate is the geometric counterpart of the usual  $\mathcal{C}^2$  chain rule. Indeed, let  $h = \psi \circ f \circ \varphi$ , where  $f : U \subset \mathbb{R}^d \rightarrow \mathbb{R}^m$ ,  $\varphi : \tilde{U} \rightarrow U$  is a  $\mathcal{C}^2$ -diffeomorphism, and  $\psi$  is a  $\mathcal{C}^2$ -diffeomorphism of the codomain. Then

$$\|d^2 h\|_\infty \leq C_1(\varphi, \psi) \|d^2 f\|_\infty + C_2(f, \varphi, \psi),$$

where

$$C_1(\varphi, \psi) = \|d\psi\|_\infty \|d\varphi\|_\infty^2, \quad C_2(f, \varphi, \psi) = \|d\psi\|_\infty \|df\|_\infty \|d^2\varphi\|_\infty + \|d^2\psi\|_\infty \|df\|_\infty^2 \|d\varphi\|_\infty^2.$$

At the geometric level, write  $\Gamma_f = \{(x, f(x)) : x \in U\} \subset \mathbb{R}^d \times \mathbb{R}^m$ , for the graph of  $f$ , and define ambient warping

$$\Phi(x, y) = (\varphi^{-1}(x), \psi(y)).$$

Then  $\Phi(\Gamma_f) = \Gamma_h$ . Federer's estimate gives

$$\frac{1}{\tau_{\Gamma_h}} \leq N_\Phi^2 \left( K_\Phi \frac{1}{\tau_{\Gamma_f}} + R_\Phi \right),$$

where, for the product Euclidean norm,

$$K_\Phi = \max\{\|\mathrm{d}\varphi^{-1}\|_\infty, \|\mathrm{d}\psi\|_\infty\}, \quad N_\Phi = \max\{\|\mathrm{d}\varphi\|_\infty, \|\mathrm{d}\psi^{-1}\|_\infty\}, \quad R_\Phi = \max\{\mathrm{Lip}(\mathrm{d}\varphi^{-1}), \mathrm{Lip}(\mathrm{d}\psi)\}.$$

Equivalently,

$$\frac{1}{\tau_{\Gamma_h}} \leq C_1^{\mathrm{geom}}(\Phi) \frac{1}{\tau_{\Gamma_f}} + C_2^{\mathrm{geom}}(\Phi),$$

with

$$C_1^{\mathrm{geom}}(\Phi) = N_\Phi^2 K_\Phi, \quad C_2^{\mathrm{geom}}(\Phi) = N_\Phi^2 R_\Phi.$$

Thus both estimates have the same structure: a second-order size is multiplied by first-order distortion factors, then enlarged by the second-order size of the deformation. In the functional setting this size is  $\|\mathrm{d}^2 f\|_\infty$ ; in the geometric setting it is  $1/\tau_{\Gamma_f}$ .

## 1.2 Boundaryless manifold inference under reach condition

### 1.2.1 Local geometry recovery

#### $C^2$ regularity and triangulations

Reconstructing the support  $M$  in Hausdorff distance requires bridging the gap between a discrete point cloud  $\mathbb{X}_n = \{X_1, \dots, X_n\}$  and a continuous geometric structure. A standard approach in computational geometry relies on simplicial complexes. With exact samples and known tangent spaces, the Tangential Delaunay Complex (TDC) introduced by [BG14] provides a triangulation that is ambient isotopic to  $M$ .

Before studying the statistical limits of manifold learning, it is insightful to decouple the statistical randomness from the purely geometric reconstruction task. Our approach uses a deterministic approximation theorem, which establishes that if a point cloud satisfies specific local density and noise conditions, and is equipped with sufficiently accurate tangent space estimates, then one can deterministically rebuild the manifold with guaranteed topological and geometric fidelity.

**Theorem 1.12** (Deterministic Manifold Reconstruction [AL18, AK22]). *There exists a dimension-dependent constant  $\lambda_d > 0$  such that for all sampling scales  $\Delta \leq \lambda_d \tau_{\min}$  and all closed submanifolds  $M \subset \mathbb{R}^D$  with reach  $\tau_M \geq \tau_{\min} > 0$ , the following holds.*

*Let  $\mathcal{X} \subset \mathbb{R}^D$  be a finite point cloud and  $\mathbb{T}_{\mathcal{X}} = \{T_x\}_{x \in \mathcal{X}}$  be a family of  $d$ -dimensional linear subspaces of  $\mathbb{R}^D$  such that:*

- **Proximity:** *sample is  $\eta$ -close to  $M$ , meaning  $\max_{x \in \mathcal{X}} \mathrm{dist}(x, M) \leq \eta$ .*
- **Density:** *sample is  $\Delta$ -dense in  $M$ , meaning  $\sup_{p \in M} \mathrm{dist}(p, \mathcal{X}) \leq \Delta$ .*
- **Alignment:** *tangents are  $\theta$ -accurate, meaning  $\max_{x \in \mathcal{X}} \|\pi_{T_{\pi_M(x)} M} - \pi_{T_x}\|_{\mathrm{op}} \leq \sin \theta$ .*

Assume that the precision components scale as  $\sin \theta \lesssim \Delta/\tau_{\min}$  and  $\eta \lesssim \Delta^2/\tau_{\min}$ , and  $\mathcal{X}$  is additionally  $\Omega(\Delta)$ -sparse. Then one can build a Tangential Delaunay Complex  $\widehat{M}_{\text{TDC}} = \widehat{M}_{\text{TDC}}(\mathcal{X}, \mathbb{T}_{\mathcal{X}})$  with vertices in  $\mathcal{X}$  such that:

- **Geometric Approximation:**  $d_{\text{H}}(M, \widehat{M}_{\text{TDC}}) \lesssim \frac{\Delta^2}{\tau_{\min}}$ .
- **Topological Guarantees:**  $M$  and  $\widehat{M}_{\text{TDC}}$  are ambient isotopic.

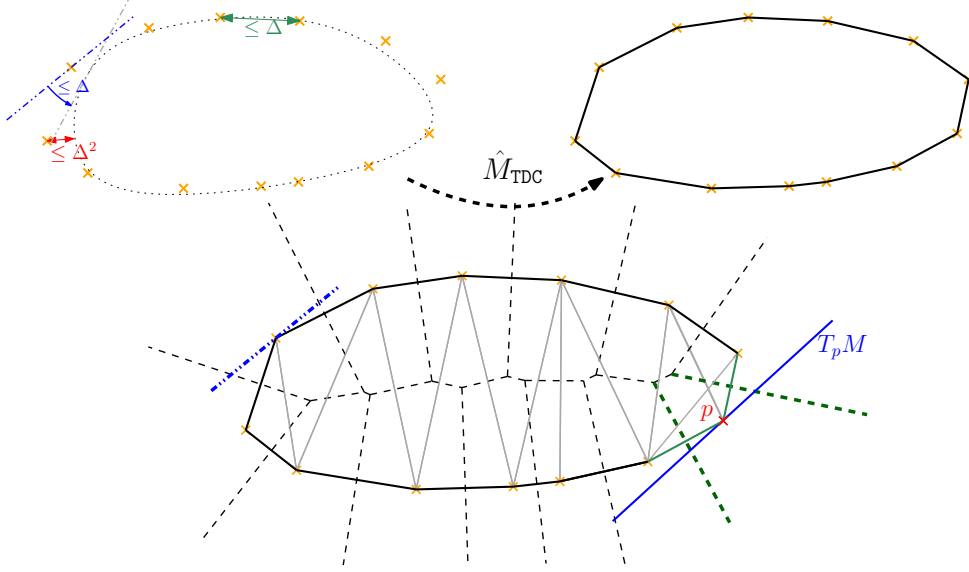


Figure 1.5: Geometric requirements of Theorem 1.12: the point cloud  $\mathcal{X}$  must densely cover the manifold at scale  $\Delta$  (green), lie  $\Delta^2$ -close to it (red), and be equipped with tangent estimates accurate up to angle  $\Delta$  (blue). The resulting Tangential Delaunay Complex  $\widehat{M}_{\text{TDC}}$  is built on top of  $\mathcal{X}$  and its estimated tangent spaces, and provides an ambient isotopic and geometrically tight reconstruction.

In [AL18], we establish that this topological faithfulness and geometric accuracy extend to the statistical setting with noisy samples and estimated tangent spaces. The analysis uses a global geometric interpolation result: given a  $\delta$ -sparse point cloud  $\mathcal{X}$  lying close to  $M$  and a collection of approximate tangent spaces, we construct a smooth proxy submanifold  $M'$  that exactly interpolates  $\mathcal{X}$  and its assigned tangent spaces, while tightly controlling the reach  $\tau_{M'} \gtrsim \tau_{\min}$ . This allows us to regard a noisy sample as an exact sample on the proxy manifold  $M'$ .

To translate this deterministic guarantee into a statistical framework, we simply need to evaluate the high-probability bounds of the quantities  $\Delta$ ,  $\eta$ , and  $\sin \theta$  for an empirical sample  $\mathbb{X}_n$ :

- **Proximity ( $\eta$ ):** Raw noiseless sample points are exactly on  $M$ , yielding  $\eta = 0$  in a noiseless setting. They could also be corrupted by small ambient additive noise.
- **Density ( $\Delta$ ):** A sample of size  $n$  drawn from a volume-equivalent distribution  $P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^2$  yields a uniform covering of  $M$  at scale  $\Delta \asymp (\log n/n)^{1/d}$  with high probability.

- **Alignment** ( $\sin \theta$ ): Using local *Principal Component Analysis* (PCA) over shrinking neighborhoods of radius  $h \asymp (\log n/n)^{1/d}$ , we can estimate the tangent spaces with an angular precision bounded by the maximal curvature,  $\sin \theta \lesssim h/\tau_{\min} \asymp \Delta/\tau_{\min}$ .

By combining the TDC with a farthest point sampling strategy for sparse covering and a local PCA for tangent space estimation, we establish that the resulting simplicial complex  $\widehat{M}_{\text{TDC}}$  achieves the minimax optimal Hausdorff rate over  $\mathcal{C}^2$  manifolds.

**Theorem 1.13** (Minimax Optimality of Tangential Delaunay Complexes [AL18]). *Let the sample  $\mathbb{X}_n$  be drawn from  $P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^2$ . Then, for  $n$  sufficiently large, the local PCA computed on neighborhoods of size  $h \asymp \left(\frac{\log n}{n}\right)^{1/d}$  around each data point yields tangent space estimators  $\widehat{T}_1, \dots, \widehat{T}_n$  satisfying*

$$\sup_P \mathbb{E} \left[ \max_{1 \leq i \leq n} \angle(T_{X_i} M, \widehat{T}_i) \right] \lesssim \left( \frac{\log n}{n} \right)^{1/d}$$

*Plugging them into the Tangential Delaunay Complex (see Theorem 1.12), we get an estimator  $\widehat{M}_{\text{TDC}} \subset \mathbb{R}^D$  which satisfies:*

$$\sup_{P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^2} \mathbb{E} \left[ d_H(M, \widehat{M}_{\text{TDC}}) \right] \lesssim \left( \frac{\log n}{n} \right)^{2/d}.$$

*Furthermore,  $\widehat{M}_{\text{TDC}}$  is ambient isotopic to  $M$  with high probability.*

This result shows that optimal Hausdorff recovery, together with topological recovery, can be achieved by computationally tractable triangulations. It should be emphasized, however, that for noisy samples with amplitude  $\sigma$ , this result deteriorates (actually with an additive term of order  $\sigma$ ). It hence has to be seen as an interpolation strategy rather than a denoising result.

Prior to this work, it was already established that if a sample forms a dense enough cover relative to  $\tau_M$ , the homology groups of  $M$  can be perfectly recovered by constructing a Čech or Vietoris-Rips complex over the data [NSW08, ALS13]. However, it is the first construction combining both an ambient isotopic reconstruction with tight Hausdorff approximation.

As will be described later (see Section 1.3.3), the (global) topology recovery of  $M$  does indeed require a global regularity assumption such as the reach. However, the (everywhere local)  $\mathcal{C}^2$  Hausdorff estimation rate can be achieved under much weaker curvature assumptions.

### $\mathcal{C}^k$ regularity and local polynomials

When the manifold  $M$  exhibits higher-order smoothness ( $k \geq 3$ ), one expects to achieve faster convergence rates for support estimation, as well as the ability to estimate higher-order differential quantities such as the tangent space  $T_x M$  and curvature structure such

as the second fundamental form  $\mathbb{I}_x^M$ . While local PCA optimally recovers tangent spaces for  $k = 2$ , it suffers from a structural bias induced by curvature, bounding its convergence rate to  $\mathcal{O}((\log n/n)^{1/d})$ .

To use the  $\mathcal{C}^k$  regularity, we develop a unified estimation scheme based on local polynomial fitting [AL19]. For a sample point  $X_i$ , we fit a  $d$ -dimensional polynomial  $\hat{\Psi}_i$  of degree  $k - 1$ ,

$$\hat{\Psi}_i(v) := X_i + v + \sum_{j=2}^{k-1} \hat{T}_{j,i}(v^{\otimes j}),$$

by minimizing a localized least-squares criterion over the Grassmannian of  $d$ -dimensional affine subspaces and the space of higher-order symmetric tensors. The zero, first, and second-order terms of the resulting polynomial respectively provide estimators for the support  $M$ , the tangent space  $T_{X_i}M$ , and the second fundamental form  $\mathbb{I}_{X_i}^M$ .

We define the quantitative statistical model  $\mathcal{P}_{\tau_{\min}, \mathbf{L}}^k$  as the class of probability distributions supported on  $d$ -dimensional manifolds  $M$  with  $\tau_M \geq \tau_{\min} > 0$ , endowed with local parameterizations whose derivatives up to order  $k$  are bounded by a tensor  $\mathbf{L}$ . Furthermore, we assume the density is bounded below by  $f_{\min} > 0$  and above by  $f_{\max} < \infty$ .

**Theorem 1.14** (Minimax Rates under Reach Assumptions [AL19]). *Let the sample  $\mathbb{X}_n$  be drawn independently from a distribution  $P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^k$  with support  $M$ . Then, for  $n$  sufficiently large, local polynomials of degree  $(k - 1)$  over neighborhoods of width  $h \asymp (\log n/n)^{1/d}$  allow to build the following estimators.*

- (0) **Support:** Local patches  $\widehat{M}_{\text{PATCH}} := \bigcup_{i=1}^n \hat{\Psi}_i(B_{\hat{T}_i}(0, h))$  estimate  $M$  in Hausdorff distance  $d_H$ , at rate

$$\sup_{P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^k} \mathbb{E} [d_H(M, \widehat{M}_{\text{PATCH}})] \lesssim \left( \frac{\log n}{n} \right)^{k/d}.$$

- (1) **Tangent Spaces:** The first order polynomials estimate tangent spaces  $T_{X_i}M$  at rate

$$\sup_{P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^k} \mathbb{E} \left[ \max_{1 \leq i \leq n} \angle(T_{X_i}M, \hat{T}_i) \right] \lesssim \left( \frac{\log n}{n} \right)^{\frac{k-1}{d}}.$$

- (2) **Curvature:** The second order polynomials estimate the second fundamental forms  $\mathbb{I}_{X_i}^M$  at rate

$$\inf_{\hat{\mathbb{I}}} \sup_{P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^k} \mathbb{E} \left[ \max_{1 \leq i \leq n} \left\| \mathbb{I}_{X_i}^M \circ \pi_{T_{X_i}M} - \hat{\mathbb{I}}_i \right\|_{\text{op}} \right] \lesssim \left( \frac{\log n}{n} \right)^{\frac{k-2}{d}}.$$

As for noiseless regression over  $\mathcal{C}^k$  functions, the rate for estimating a differential quantity of order  $\gamma$  on a  $\mathcal{C}^k$  object follows the expected rate shift  $(k - \gamma)/d$ .

**Remark 1.15** (About lower bounds at random base points). Although the rate shift  $(k - \gamma)/d$  is the same as in noiseless regression over  $\mathcal{C}^k$  functions, the lower bound for

random quantities such as  $T_{X_i}M$  or  $\mathbb{I}_{X_i}^M$  is not a direct fixed-point lower bound with  $x$  replaced by  $X_i$ . At a deterministic point  $x \in M$ , the target is a fixed functional of the underlying distribution, and standard Le Cam’s argument applies. By contrast, for a random base point, both the target  $\theta_{X_i}(P)$ , with  $\theta \in \{TM, \mathbb{I}^M\}$ , and the integration measure in the loss depend on  $P$ :

$$\mathbb{E}_{P^{\otimes n}} d(\theta_{X_1}(P), \hat{\theta}) = \mathbb{E}_{P^{\otimes(n-1)}} \left[ \mathbb{E}_{P_1} \left[ d(\theta(P), \hat{\theta}) \right] \right].$$

In the noiseless case this dependence is singular: different hypotheses are supported on different bumped manifolds, so comparing tangent spaces or curvatures “at the same point” is not canonical. The proof therefore uses a conditional version of Assouad’s lemma. Roughly speaking, one plants many disjoint local bumps, conditions on the event that  $X_1$  falls in one active bump while the remaining observations avoid it, and randomizes the bump parameters through mixture hypotheses to keep enough affinity between neighboring alternatives. This conditioning-and-mixture step is the main technical point, and is absent from the deterministic base-point lower bound.

This is analogous to noiseless random-design regression, where the rate is governed by the random holes left by the design points. For related minimax lower bounds in the model  $Y = f(X)$  with  $f \in \mathcal{C}^k$ , see for instance [BDK<sup>+</sup>17] for the sup-norm random-design interpolation rate, and [KK13] for an  $L^1(P_X)$  version.

## 1.2.2 Global geometry recovery

Having established the optimal recovery of local geometric quantities, we shift the focus to the global geometry of  $M$ , encompassing its intrinsic metric structure and its reach parameter  $\tau_M$ . These quantities describe global features of the manifold and are central to dimensionality reduction algorithms and topological inference.

### Intrinsic metric estimation

Unsupervised metric learning aims at recovering the intrinsic shortest-path (geodesic) distance  $d_M(x, y)$  between points on  $M$ , given only ambient Euclidean observations. This is the first step of algorithms such as Isomap [TDSL00], which construct a neighborhood graph to approximate geodesic paths. However, standard neighborhood graph methods are known to achieve slow convergence rates.

In [ABL23], we propose a plug-in strategy for metric estimation. By defining the intrinsic metric estimator  $\hat{d}(x, y)$  as the geodesic distance within an  $\varepsilon_n$ -offset of a support estimator  $\hat{M}_n$ , we establish a transfer of convergence rates from support estimation to metric estimation (see Figure 1.6).

**Theorem 1.16** (Minimax Geodesic Metric Learning [ABL23]). *Let  $\hat{M}_n$  be a support estimator achieving a Hausdorff error rate of  $\varepsilon_n$  over  $\mathcal{P}_{\tau_{\min}, \mathbf{L}}^k$ . Then, the metric estimator*

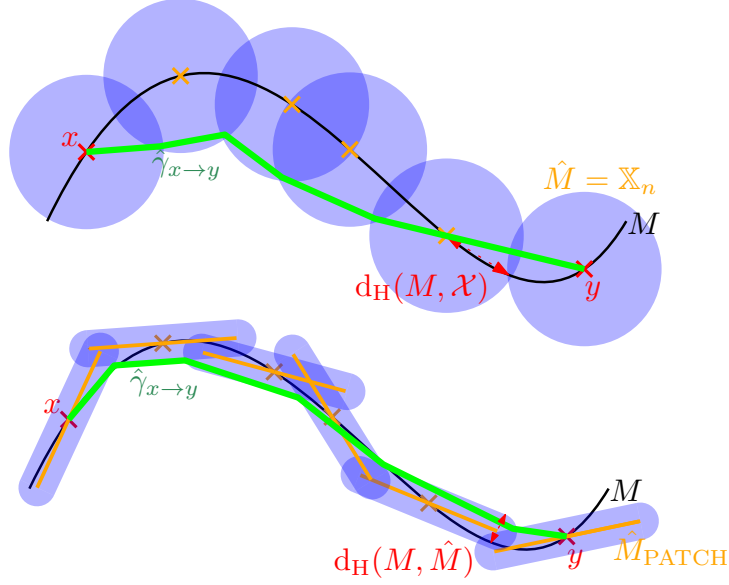


Figure 1.6: Comparison of the plug-in geodesic distance estimator of Theorem 1.16, for two estimators. (top) The trivial estimator  $\hat{\mathbb{X}}_n = \{X_1, \dots, X_n\}$  achieves Hausdorff rate  $\varepsilon_n \asymp (\log n/n)^{1/d}$  (bottom) The tangential Delaunay Complex achieves the faster Hausdorff rate  $\varepsilon_n \asymp (\log n/n)^{2/d}$ . Therefore, the geodesic distance estimator based on the later is better than the former.

$\hat{d}$  defined over the offset  $(\widehat{M}_n)^{\varepsilon_n}$  satisfies  $M \subset (\widehat{M}_n)^{\varepsilon_n}$ , and

$$\sup_P \mathbb{E} \left[ \sup_{x \neq y \in M} \left| 1 - \frac{\hat{d}(x, y)}{d_M(x, y)} \right| \right] \lesssim \frac{\varepsilon_n}{\tau_{\min}}.$$

In particular, using the local polynomial estimator for  $\widehat{M}_n$  of Theorem 1.14, the intrinsic metric is recovered at the (minimax optimal) rate  $\mathcal{O}((\log n/n)^{k/d})$ .

The proof relies on uniformly bounding the geodesic versus Euclidean metric distortions within tubular neighborhoods of positive reach. This result demonstrates that estimating the intrinsic metric can be reduced to estimating the set itself in Hausdorff distance.

### Estimating the reach parameter

While the reach  $\tau_M$  appears as an important prior parameter in the aforementioned theorems, estimating it directly requires a separate analysis. As established in Theorem 1.5, the value of the reach of a closed submanifold is dictated by two competing geometric obstructions: maximal curvature (local) and narrow bottlenecks (global). Consequently, an optimal estimator must adapt to the regime that determines the smallest scale. In [AKC<sup>+</sup>19] and [ABL23], we decouple the estimation of these two competing obstructions.

- In [AKC<sup>+</sup>19], we proposed a plug-in based on the tangent space formulation of the reach of Theorem 1.3. In addition to requiring a quite precise knowledge of tangent spaces, this estimate led to a suboptimal analysis, resulting in suboptimal rates.
- In [ABL23] we later proposed an estimator based on the metric characterization of Theorem 1.4. This allows us to estimate separately the local curvature constraint evaluating the maximal empirical curvature across all local polynomial patches of Theorem 1.14. For the global bottleneck constraint, we introduce the *Spherical Distortion Radius (SDR)*, an intermediate metric scale that characterizes the global reach in a metric way, and which we show to be Lipschitz with respect to perturbations in the Hausdorff metric. As a result, plugging in the optimal geodesic metric estimator of Theorem 1.16 yields optimal rates.

**Theorem 1.17** (Minimax Reach Estimation [ABL23]). *By combining the maximal curvature estimator  $\hat{R}_{loc}$  and the empirical Spherical Distortion Radius  $\widehat{\text{sdr}}_\delta$ , we construct an estimator  $\hat{\tau}_M = \hat{R}_{loc} \wedge \widehat{\text{sdr}}_\delta$  that achieves the following (minimax-optimal) adaptive rates:*

$$\sup_P \mathbb{E} [|\hat{\tau}_M - \tau_M|] \lesssim \begin{cases} \left(\frac{\log n}{n}\right)^{(k-2)/d} & \text{if } \tau_M \text{ attained locally (curvature),} \\ \left(\frac{\log n}{n}\right)^{k/d} & \text{if } \tau_M \text{ is attained globally (bottleneck).} \end{cases}$$

This shows the following statistical phenomenon: because of the double nature of the reach (Theorem 1.5) estimating a global topological bottleneck converges at the same rate as estimating the support, whereas estimating the reach driven by local curvature incurs the slower rate associated with second-order differential quantities.

### Minimax inconsistency without the reach

The results above may suggest that the reach assumption is mainly a quantitative device: it fixes the admissible scale in the TDC reconstruction of Theorems 1.12–1.13, controls the local polynomial parametrizations in Theorem 1.14, and enters the metric and reach estimators of Theorems 1.16–1.17. This interpretation is incomplete. The lower bound  $\tau_M \geq \tau_{\min} > 0$  is not only a regularity bound; it is an identifiability assumption. As recalled in Theorem 1.5, the reach simultaneously prevents large curvature and arbitrarily thin bottlenecks. Dropping it allows two distinct sheets of the same smooth manifold to come at an arbitrarily small Euclidean distance, while keeping all local  $\mathcal{C}^k$  bounds under control.

Let  $\mathcal{P}_{(x)}^k$  denote the class of distributions supported on compact  $\mathcal{C}^k$  submanifolds  $M$  passing through a fixed point  $x \in \mathbb{R}^D$ , with uniformly bounded local parametrizations and densities, but with no positive lower bound on  $\tau_M$ .

**Theorem 1.18** (Minimax inconsistency without reach [AL19]). *Assume that no reach lower bound is imposed. If  $D \geq d + 3$ , then for all  $k \geq 3$  and all sample sizes  $n \geq 1$ ,*

$$\inf_{\hat{T}_n} \sup_{P \in \mathcal{P}_{(x)}^k} \mathbb{E} [\angle(T_x M, \hat{T}_n)] \geq \frac{1}{2} > 0,$$

Similarly, if  $D \geq d + 1$ , then for some constant  $c > 0$ ,

$$\inf_{\hat{\mathbb{I}}_n} \sup_{P \in \mathcal{P}_{(x)}^k} \mathbb{E} \left[ \left\| \mathbb{I}_x^M \circ \pi_{T_x M} - \hat{\mathbb{I}}_n \right\|_{\text{op}} \right] \geq c > 0,$$

The mechanism is simple but leads to several consequences. For tangent estimation, one builds two  $\mathcal{C}^\infty$  manifolds  $M, M'$  that coincide outside an arbitrarily small ball  $B(x, \delta)$ , but whose tangent spaces at  $x$  differ by a fixed angle. For curvature estimation, one builds two manifolds that again coincide outside  $B(x, \delta)$ , but one is flat at  $x$  while the other has curvature of order one. The transition is made through a smooth bump whose width tends to zero, so that the  $\mathcal{C}^k$  bounds remain fixed while the reach collapses.

Le Cam's lemma then gives, for a geometric functional  $\theta_x(M)$ ,

$$\inf_{\hat{\theta}_n} \sup_P \mathbb{E} \left[ d(\theta(P), \hat{\theta}) \right] \geq \frac{1}{2} d(\theta_x(P), \theta_x(P')) (1 - \text{TV}(P, P'))^n.$$

Here the parameter separation  $d(\theta(P), \theta(P'))$  is bounded from below by a positive constant, whereas

$$\text{TV}(P, P') \lesssim P(B(x, \delta)) \lesssim \delta^d \rightarrow 0.$$

Thus, for every fixed  $n$ , the two statistical experiments become indistinguishable although their tangent spaces or curvatures at  $x$  remain separated. With high probability the sample only sees the common part of the two manifolds; the infinitesimal bottleneck deciding the local geometry at  $x$  is missed.

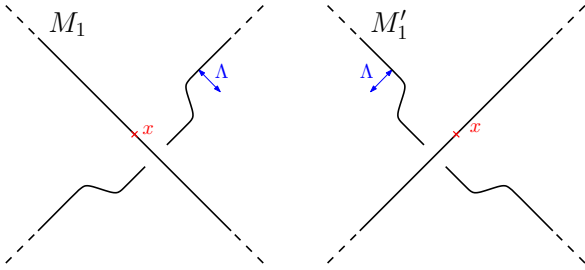


Figure 1.7: Tangent inconsistency: two smooth manifolds coincide except in an arbitrarily small neighborhood of  $x$ , but have distinct tangent spaces at  $x$ .

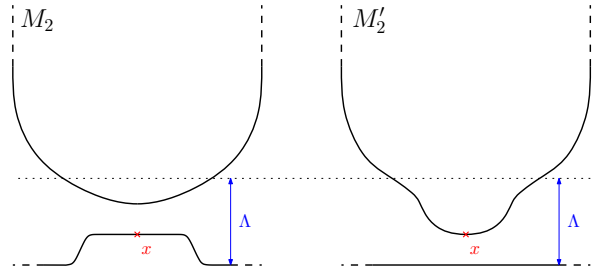


Figure 1.8: Curvature inconsistency: a flat germ and a curved cap differ only in a vanishing neighborhood of  $x$ .

This explains why the positive reach assumption is used throughout this section. It is what makes the local chart around  $x$  statistically visible at a scale comparable with the sampling radius; without it, local smoothness alone does not prevent hidden near self-intersections.

The same obstruction also affects all the global geometric or topological quantities. Without a reach-type constraint, homology is not uniformly identifiable [BRS<sup>+</sup>12]; connectedness, and hence finiteness of geodesic distances, cannot be tested uniformly; bottleneck-driven quantities such as the reach or the spherical distortion radius of Theorem 1.17

become unstable; and boundaries cannot be uniformly localized either [AAL23]. The stratified setting discussed later in Section 1.3.3 should not be interpreted as removing this obstruction, but as changing the target: exact tangent spaces or singular loci remain non-identifiable at unresolved intersections, while dimension-wise clustering and local tangent directions can still be recovered away from such ambiguous regions.

## 1.3 Beyond the closed manifold model

The previous section identified the role of the reach. Under the closed-manifold model, the lower bound  $\tau_M \geq \tau_{\min} > 0$  ties together the ambient and intrinsic geometries: it gives stable tangent tubes, makes local polynomial patches meaningful, enables the TDC reconstruction of Theorems 1.12–1.13, and underlies the global metric and reach estimators of Theorems 1.16–1.17. Conversely, Theorem 1.18 shows that removing this global constraint can destroy identifiability for several classical targets.

This section studies which tasks remain possible once the pristine manifold model is relaxed. The guiding principle is the following: when a structural assumption is removed, one must either accept a slower rate, restrict the observation model, or change the target of inference. We illustrate this in three different contexts: boundaries, creating a new lower-dimensional object that is not directly sampled; adversarial noise, forcing us to replace pointwise access by stable aggregate information; intersections and mixed dimensions, making classical manifold quantities ill-defined, but still allowing dimension-wise clustering and local support recovery.

### 1.3.1 Manifolds with boundary

The closed-manifold model used in Section 1.2.2 assumes a local Euclidean symmetry around every point. A boundary breaks this local symmetry: near  $\partial M$ , only one side of the tangent ball is populated. This has algorithmic consequences. While local PCA does not suffer any substantial bias near the rim, the Delaunay-type reconstructions may create spurious simplices across the missing side, and the local polynomial patches may point towards an empty zone of space. Thus, even though the manifold remains smooth and has positive reach away from singularities, the estimator must first learn where the boundary is.

In [AAL23], we address this problem for  $\mathcal{C}^2$  manifolds with boundary. The main difficulty is that  $\partial M$  has zero  $d$ -dimensional volume, hence it is never directly sampled. It must be inferred from an imbalance in the neighboring observations. In full-dimensional support estimation, boundary points are detected through large ambient Voronoi cells. For a low-dimensional manifold  $d < D$ , however, all ambient Voronoi cells are unbounded in the normal directions. We therefore compute Voronoi diagrams inside estimated tangent spaces.

For each sample point  $X_i$ , we project the local neighborhood  $\mathbb{X}_n \cap B(X_i, R_0)$  onto an estimated tangent space  $\hat{T}_i$ . A point is declared close to the boundary if its tangent-

projected Voronoi cell is elongated at scale  $\rho$ . The direction of this elongation estimates the outward normal  $\eta$  to the boundary. This produces a set of boundary observations  $\mathcal{Y}_{R_0, r, \rho}$  and associated normal estimates  $\hat{\eta}_i$ .

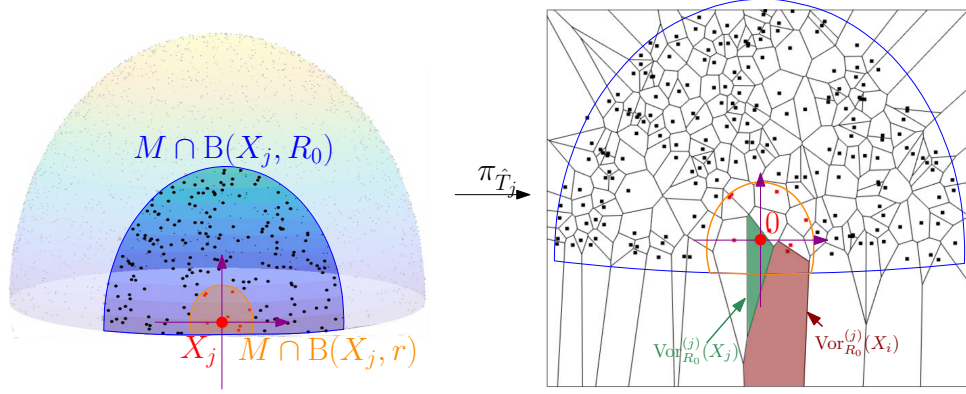


Figure 1.9: Boundary detection on a half-sphere ( $d = 2$ ,  $D = 3$ ). The Voronoi diagram is computed in the estimated tangent planes, not in the ambient space. Near  $\partial M$ , projected Voronoi cells become elongated in the outward normal direction, allowing both boundary detection and normal estimation.

With well-calibrated parameters,  $\mathcal{Y}_{R_0, r, \rho}$  concentrates around  $\partial M$ : it covers the boundary at scale  $(\log n/n)^{1/(d+1)}$  and consists of points at distance at most  $(\log n/n)^{2/(d+1)}$  from  $\partial M$ . Moreover, the estimated outward normals satisfy

$$\angle(\hat{\eta}_i, \eta_{\pi_{\partial M}(X_i)}) \lesssim \left(\frac{\log n}{n}\right)^{1/(d+1)}.$$

The exponent  $d + 1$  already reflects the statistical price of the boundary: one must infer a codimension-one object from a one-sided  $d$ -dimensional distribution.

The support estimator is then modified accordingly. Interior points are treated as in the local patch construction of Theorem 1.14 for  $k = 2$ . Boundary points instead generate half-patches, truncated by the estimated outward normal:

$$\widehat{M}_\partial = \bigcup_{X_i \in \mathcal{Y}_{R_0, r, \rho}} \left( X_i + B_{\hat{T}_i}(0, \varepsilon_\partial) \right) \cap \left\{ z \in \mathbb{R}^D : \langle z - X_i, \hat{\eta}_i \rangle \leq 0 \right\}.$$

Gluing these half-patches with the interior patches yields a boundary-adaptive estimator  $\widehat{M}$ .

**Theorem 1.19** (Minimax boundary and support recovery [AAL23]). *Let  $P$  be supported on a  $\mathcal{C}^2$  manifold  $M$  with boundary, with internal reach and boundary reach bounded from below. Then the tangent-Voronoi boundary detection and half-patch reconstruction procedure yields estimators  $\widehat{M}$  and  $\widehat{\partial M}$  such that*

$$\mathbb{E} \left[ d_H(M, \widehat{M}) \vee d_H(\partial M, \widehat{\partial M}) \right] \lesssim \left( \frac{\log n}{n} \right)^{2/(d+1)}.$$

*This rate is minimax optimal up to logarithmic factors.*

Thus, compared with the closed  $\mathcal{C}^2$  rate  $(\log n/n)^{2/d}$  of Theorem 1.13, the boundary changes the effective exponent as if one additional dimension were present. The lower bound follows by reducing the problem to the estimation of a convex domain with smooth boundary, where the same phenomenon is classical [DW96, RC07]. In the language of the previous section, the reach still prevents hidden bottlenecks, but the target now contains a zero-measure feature that must be detected before it can be estimated.

### 1.3.2 Robustness and adversarial noise

The estimators above use the sample as a geometric object: nearest neighbors, local PCA, Voronoi cells, and triangulations all rely on direct access to the point cloud. This can be unstable under adversarial outliers and incompatible with strict privacy or distributed-access constraints. The next question is therefore different from the previous one: assume the same regular manifold model as before, but restrict the information available to the learner. Can the manifold still be reconstructed?

In [AK21, AK22], we study this question in the *Statistical Query* (SQ) framework [Kear98]. The learner no longer observes samples. Instead, for every bounded measurable query  $r : \mathbb{R}^D \rightarrow [-1, 1]$ , it receives an adversarial approximation to  $\mathbb{E}r(X)$ :

$$|a - \mathbb{E}[r](X)| \leq \tau.$$

The tolerance  $\tau$  is part of the model. The oracle may answer adversarially within this interval and may adapt to the previous queries.

#### From point clouds to stable aggregate information

The SQ model replaces individual samples by aggregate averages. This restriction is strong, but it captures several constraints at once. First, it is compatible with privacy, since the learner never accesses individual observations. Second, bounded queries are automatically robust to small adversarial contamination: if

$$P_{\text{cont}} = (1 - \alpha)P + \alpha Q,$$

then for all  $r \in [-1, 1]$ ,

$$|\mathbb{E}_{P_{\text{cont}}} r - \mathbb{E}_P r| \leq 2\alpha.$$

Thus contamination simply increases the effective oracle tolerance. Finally, SQ access models distributed settings in which only summary statistics can be transmitted.

Every SQ algorithm can be simulated from an ordinary sample: empirical averages answer  $q$  queries with tolerance  $\tau$  using  $n \gtrsim \tau^{-2} \log q$  samples, by Hoeffding's inequality. The converse is false in general. Thus SQ learnability is a more restrictive requirement than ordinary sample-based learnability.

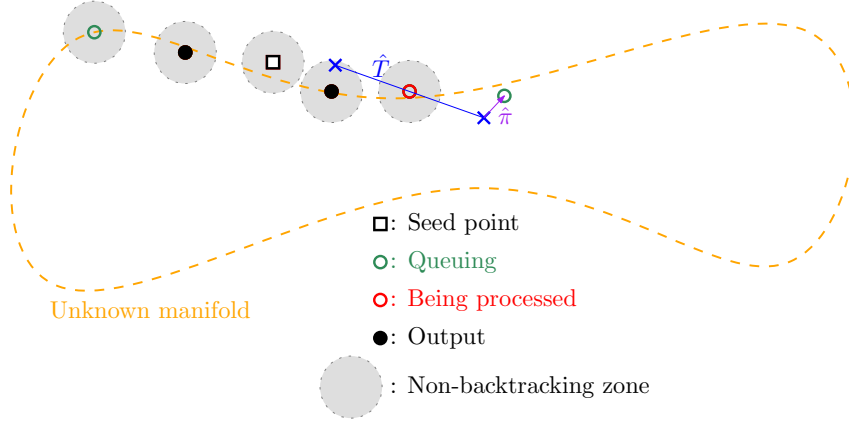


Figure 1.10: One propagation step of **Manifold Propagation**: tangent exploration, approximate projection back to  $M$ , and non-backtracking against the already explored region.

### Active exploration: the Manifold Propagation Algorithm

Since samples are unavailable, the geometric routines used earlier in this chapter must be replaced by query-based counterparts. The *Manifold Propagation Algorithm* (**Manifold Propagation**) implements this strategy. Its role is to construct, using only SQ access, a point cloud that is dense enough to be fed into the deterministic reconstruction theorem, Theorem 1.12.

The algorithm relies on three SQ primitives.

- **Seed detection.** An ambient binary search procedure, SQ Ambient Binary Search, locates a point close to  $M$  by querying masses of nested regions.
- **Tangent recovery.** At a point close to  $M$ , tangent estimation is performed through local covariance queries. A naive implementation would require  $O(D^2)$  queries, but the covariance is intrinsically low-rank. Using compressed-sensing ideas, one can recover the relevant tangent information with query complexity linear in  $D$ , up to constants depending on  $d$  for fixed intrinsic dimension  $d$ .
- **Projection.** Candidate points created during exploration may drift away from  $M$ . They are recentered by querying local conditional means, which play the role of a robust approximate projection onto the support.

The global algorithm is a breadth-first exploration of the manifold. Starting from a seed point, it estimates the tangent space, proposes new points on a sphere of radius  $\Delta$  in the tangent plane, discards candidates already close to the explored region, projects the remaining ones back toward  $M$ , and adds them to a queue.

Equivalently, outside the SQ formalism, the propagation loop is:

1. Start from a seed point  $\hat{x}_0$  close to  $M$ , with queue  $\mathcal{Q} = \{\hat{x}_0\}$  and output cloud  $\mathcal{O} = \emptyset$ .
2. Pop  $x_0$  from  $\mathcal{Q}$ , add it to  $\mathcal{O}$ , and estimate  $\hat{T}(x_0)$ .
3. Cover the sphere of radius  $\Delta$  in  $\hat{T}(x_0)$ , remove candidates already close to  $\mathcal{Q} \cup \mathcal{O}$ , approximately project the remaining candidates back to  $M$ , and add them to  $\mathcal{Q}$ .

4. Stop when  $\mathcal{Q} = \emptyset$ .

The non-backtracking test is needed: without it, even exact local routines would propagate forever. Once the resulting cloud is  $O(\Delta)$ -dense and equipped with tangent estimates, Theorem 1.12 turns it into a geometric and topological reconstruction.

**Theorem 1.20** (SQ complexity of manifold learning [AK22]). *Let the target Hausdorff precision be  $\varepsilon > 0$ . Under a  $\text{STAT}(\tau)$  oracle, manifold reconstruction can be achieved with tolerance*

$$\tau = \mathcal{O}(\varepsilon^{(d+1)/2})$$

and query complexity

$$q \asymp D \log \left( \frac{R}{\varepsilon} \right) + \frac{D}{\tau} \left( \frac{1}{\varepsilon} \right)^{d/2}.$$

Conversely, any randomized SQ algorithm achieving precision  $\varepsilon$  requires  $\Omega(\varepsilon^{-d/2})$  queries.

This result complements the minimax rates of the previous section. With clean samples, the statistical difficulty is governed by covering  $M$  at the correct intrinsic scale. With SQ access, the manifold must also be detectable through averages of tolerance  $\tau$ . If  $\tau$  is larger than the mass of the local geometric features to be discovered, those features cannot be detected, even though they would be identifiable from raw samples.

### 1.3.3 The union of manifolds hypothesis: stratifications

We now return to the obstruction identified in Theorem 1.18. If the support is allowed to self-intersect, or if several manifolds of different dimensions coexist, the global reach may collapse to zero. Then the classical targets of manifold inference are no longer stable: tangent spaces at singular points may not be uniquely defined, connectedness and homology may not be statistically identifiable, and the intrinsic metric may depend on unresolved bottlenecks.

The correct response is to avoid these non-identifiable targets, but to change the target. The *union of manifolds hypothesis* assumes that

$$M = \bigcup_{k=1}^K M_k,$$

where each  $M_k$  is the image of a  $\mathcal{C}^2$  immersed manifold of dimension

$$d_1 < d_2 < \cdots < d_K,$$

with bounded intrinsic curvature and volume. No global reach, separation, or transversality condition is imposed. Thus the assumptions are intrinsic to the strata, not ambient.

Despite the failure of exact tangent or topology estimation at singularities, we prove in [AB24] that the stratification can still be learned dimension by dimension. The procedure, called **Co-detection**, is an ascending peeling algorithm. Its basic test is local

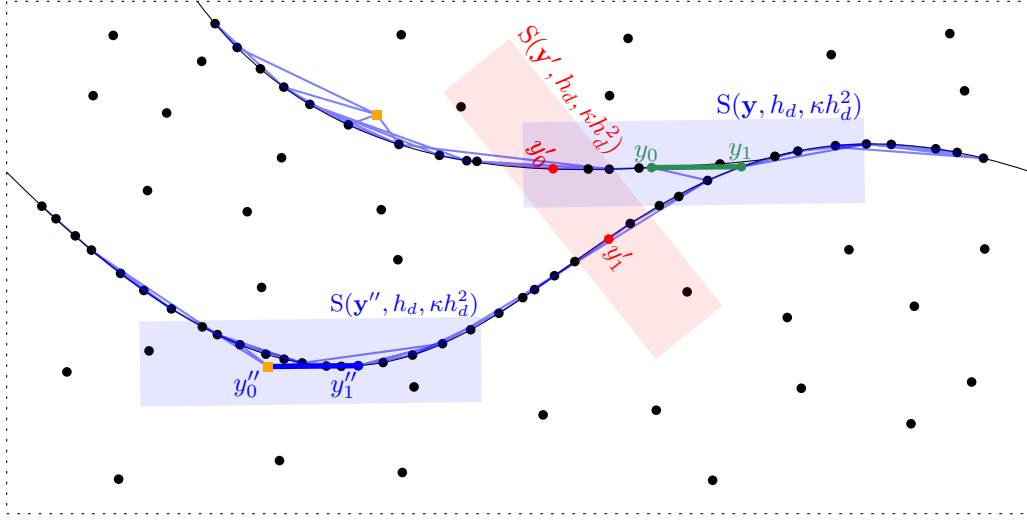


Figure 1.11: Co-detection at dimension  $d = 1$ . Pairs whose slabs contain enough mass are accepted. Higher-dimensional points may also trigger co-detection, but their convex hulls remain within the optimal  $O(h_d^2)$  neighborhood of the lower-dimensional stratum and therefore do not affect the Hausdorff rate.

and volumetric: it searches for configurations that carry enough empirical mass inside a thin slab.

For a candidate  $d$ -plane  $T$ , define the slab

$$S_T(x, h_{\parallel}, h_{\perp}) = x + B_T(0, h_{\parallel}) + B_{T^{\perp}}(0, h_{\perp}), \quad h_{\perp} \asymp h_{\parallel}^2.$$

The quadratic normal thickness matches the curvature scale of a  $\mathcal{C}^2$  manifold, exactly as in the locally linear constructions of Theorems 1.13 and 1.14 for smoothness  $k = 2$ . A  $(d + 1)$ -tuple spanning  $T$  is *co-detected* when the empirical mass of the associated slab exceeds the threshold predicted by VC concentration. Co-detected convex hulls are collected, nearby points are assigned to the current dimension, and then peeled away before moving to larger dimensions.

Let  $\hat{\mathcal{U}}_d$  be the set of co-detected  $(d + 1)$ -tuples at loop dimension  $d$ . Let  $\hat{d}_1, \dots, \hat{d}_{\hat{K}}$  be the dimensions for which  $\hat{\mathcal{U}}_d \neq \emptyset$ . For a sample point  $X_i$ , let  $\hat{d}(X_i)$  be the dimension assigned by the algorithm, and let  $Y_i$  be the latent dimension of the stratum from which  $X_i$  was sampled.

**Theorem 1.21** (Stratification learning [AB24]). *Let  $\mathbb{X}_n$  be drawn from a mixture supported on*

$$M = \bigcup_{k=1}^K M_k \subset \mathbb{R}^D,$$

*where the  $M_k$  are  $\mathcal{C}^2$  immersed manifolds of dimensions  $d_1 < \dots < d_K$ , with bounded curvature and volume. With well-chosen parameters, the ascending co-detection algorithm satisfies, with high probability:*

- **Dimension discovery:**

$$\widehat{K} = K, \quad \widehat{d}_k = d_k \quad (1 \leq k \leq K).$$

- **Dimension labeling:** for every  $i$ ,

$$\min \left\{ d_k : \text{dist}(X_i, M_k) \lesssim \left( \frac{\log n}{n} \right)^{2/d_k} \right\} \leq \widehat{d}(X_i) \leq Y_i.$$

- **Dimension-wise clustering:** the induced partition  $\mathbb{X}_n = \bigsqcup_{k=1}^K \widehat{\mathbb{X}}_{k,n}$  has relative misclassification error

$$\mathbb{E} \left[ \frac{\#(\widehat{\mathbb{X}}_{k,n} \triangle \mathbb{X}_{k,n})}{\#\mathbb{X}_{k,n}} \right] \lesssim \left( \frac{\log n}{n} \right)^{2\beta_k},$$

where

$$\beta_k = \min \left( \frac{d_k - d_{k-1}}{d_{k-1}}, \frac{d_{k+1} - d_k}{d_k} \right),$$

with convention  $d_0 = 0$  and  $d_{K+1} = \infty$ .

- **Stratum estimation:** the estimator

$$\widehat{M}_k = \bigcup_{\mathbf{y} \in \widehat{\mathcal{U}}_{d_k}} \text{conv}(\mathbf{y})$$

achieves the layer-adaptive Hausdorff rate

$$d_H(M_k, \widehat{M}_k) \lesssim \left( \frac{\log n}{n} \right)^{2/d_k}.$$

This theorem should be read together with Theorem 1.18. There is no contradiction. Without reach, exact tangent spaces at singular points, self-intersection loci, geodesic distances, and global topology remain non-identifiable. Theorem 1.21 works by changing the target: instead of asking for the exact local branch at every point, it detects the intrinsic dimensions present in the mixture, clusters the sample dimension by dimension, and estimates each stratum in Hausdorff distance. Thus, positive reach is needed for the global identifiability considered above, but not for local support approximation under intrinsic curvature control.

The same distinction appears at the level of tangent estimation. The non-identifiable target is to estimate the exact tangent space  $T_x M$  at a prescribed singular or near-singular point. The feasible problem is to estimate the tangent structure of each stratum locally, allowing a spatial tolerance at the same scale as the Hausdorff error. This leads naturally to a tangent-cone type estimator.

For  $x \in \widehat{M}_k$ , define the empirical tangent cone of the reconstructed stratum by

$$T_x \widehat{M}_k := \bigcup_{\substack{\sigma \in \widehat{\mathcal{U}}_{d_k} \\ x \in \text{conv}(\sigma)}} \text{Span}(\sigma).$$

This is not necessarily a single plane: near intersections, or near junctions of the empirical reconstruction, several candidate tangent directions may coexist. Accordingly, the usual Grassmannian loss must be localized. For a spatial tolerance  $\Delta > 0$ , set

$$\angle_{\Delta}(TM_k \mid T\widehat{M}_k) := \sup_{x \in M_k} \inf_{y \in \widehat{M}_k \cap B(x, \Delta)} \angle(T_x M_k \mid T_y \widehat{M}_k),$$

and

$$\angle_{\Delta}(T\widehat{M}_k \mid TM_k) := \sup_{y \in \widehat{M}_k} \inf_{x \in M_k \cap B(y, \Delta)} \angle(T_y \widehat{M}_k \mid T_x M_k).$$

Here  $\angle(V \mid \mathcal{C})$  denotes the one-sided angle from a plane  $V$  to a finite union of planes  $\mathcal{C}$ , obtained by minimizing over the planes in  $\mathcal{C}$ . We then use the symmetric localized loss

$$\angle_{\Delta}(T\widehat{M}_k, TM_k) := \angle_{\Delta}(T\widehat{M}_k \mid TM_k) \vee \angle_{\Delta}(TM_k \mid T\widehat{M}_k).$$

**Theorem 1.22** (Localized tangent estimation [AB24]). *With the same parameter choice as in Theorem 1.21, for  $n$  large enough, with high probability, the following holds simultaneously for all  $k \in \{1, \dots, K\}$ :*

$$\angle_{\Delta_k}(T\widehat{M}_k, TM_k) \lesssim \left( \frac{\log n}{n} \right)^{1/d_k},$$

where  $\Delta_k \asymp (\log n/n)^{2/d_k}$ .

In words, the  $d_k$ -dimensional tangent structure of the stratum  $M_k$  is recovered with angular accuracy  $(\log n/n)^{1/d_k}$  up to translations in neighborhoods of radius  $(\log n/n)^{2/d_k}$ . The two scales are the geometric scales already encountered in the positive-reach  $\mathcal{C}^2$  theory: tangent estimation has first-order rate, while support estimation has second-order Hausdorff rate; compare Theorems 1.13 and 1.14. For an embedded manifold with reach bounded below and no mixture ( $K = 1$ ), the tolerance  $\Delta_k$  can be removed, and Theorem 1.22 reduces to the usual tangent-space estimation rate. For general immersions, the tolerance is needed: at unresolved intersections there may be no statistically meaningful choice of a single tangent plane at the exact point.

Thus the obstruction of Theorem 1.18 is avoided by changing the target. Estimating  $T_x M$  at an exact point  $x$  may be impossible when another branch approaches at an arbitrarily small scale. But estimating a tangent direction of the correct stratum at some nearby point  $\|x - y\| \lesssim (\log n/n)^{2/d_k}$  is possible, with the optimal angular accuracy  $(\log n/n)^{1/d_k}$ . This result gives the corresponding tangent-space statement: without reach, exact pointwise geometry is not a stable target, whereas localized, stratum-wise geometry remains statistically learnable.

## 1.4 Open directions

### 1.4.1 Implicit representations for geometric inference

Throughout this chapter, we tracked a progressive relaxation of the rigid constraints historically imposed on geometric models. Beyond this theory of recovering shapes

from point clouds, geometric inference can serve as a preliminary step for downstream statistical learning. For instance, estimating the support is sometimes a prerequisite for performing distribution estimation [Div22] or defining meaningful generative models on manifolds [Sté25].

However, the implementation of these geometric estimators raises computational and representational questions. While a manifold  $M$  can be characterized mathematically as a subset of  $\mathbb{R}^D$ , this representation is difficult to use computationally in high dimensions. Alternative implicit representations such as membership functions  $\mathbb{1}_M$ , distance functions  $x \mapsto \text{dist}(x, M)$ , or orthogonal projection maps  $\pi_M$ , offer varying algorithmic advantages. Among these, the projection map  $\pi_M$  is particularly useful. Not only does it directly encode the normal bundle and local geometry, but it also naturally extends to noisy scenarios, where projecting off-manifold points back to the data support provides a natural denoising mechanism. From a computational geometry standpoint, access to a robust projection oracle enables the consistent meshing of surfaces [AK04] and the tracking of moving interfaces [Lev98]. This perspective connects explicit topological algorithms and modern representation learning.

### 1.4.2 A few conjectures

To conclude this chapter, we outline a few open mathematical questions and formulate specific conjectures that naturally extend the presented framework.

**Additive noise:** While the Statistical Query framework handles adversarial contamination, addressing *additive random noise* in the classical sample framework requires different tools. As data points themselves are displaced, the very concept of exact nearest-neighbor proximity is no longer directly applicable. Recent investigations into the robust geometry of noisy samples have revealed a range of noise models, each possibly carrying different statistical behaviors:

- **Convolutional Noise:** Observations are given by  $Y = X + \varepsilon$ , where the noise  $\varepsilon \in \mathbb{R}^D$  is a fully-dimensional ambient noise independent of  $X$ . In this regime, recovering a manifold currently requires carefully bounded noise variances that depend on the ambient dimension  $D$  [FILN18].
- **Orthogonal Noise:** The noise is confined to the normal bundle,  $Y = X + \varepsilon$  with  $\varepsilon \in (T_X M)^\perp$ , such that  $\mathbb{E}[\varepsilon \mid X] = 0$ . Here, the noise acts purely in the normal directions to the manifold [PS22].
- **Ambient Uniform/Tubular Noise:** Observations are uniformly distributed in a thickened  $\sigma$ -offset  $M^\sigma$  of the manifold [AS25].

One important phenomenon preventing direct estimation in these regimes is that the noise interacts with curvature to create bias in local estimators. As observed by [AS25], any geometric error in tangent space estimation yields a projection bias that translates into apparent non-centered noise, regardless of the initial noise distribution. This effect degrades classical non-parametric regression estimators and can push the reconstructed points away from the true underlying support.

To break this bias, the one current approach consists of *alternating* local PCA and nonparametric regression [AS25, PS22], which approaches asymptotic optimality provided the noise level  $\sigma \rightarrow 0$  fast enough as  $n \rightarrow \infty$ . However, clarifying what would be a minimax-optimal estimator when the noise amplitude  $\sigma$  is fixed remains largely open.

**Conjecture 1.1** (Manifold Denoising). Consider i.i.d. data  $Y_1, \dots, Y_n$  sampled as  $Y = X + \varepsilon$ , where  $X \sim P \in \mathcal{P}_k^{d,D}$  is uniform over its  $d$ -dimensional  $\mathcal{C}^k$ -smooth support  $M \subset \mathbb{R}^D$ , assumed to be a boundaryless submanifold with positive reach. Let the isotropic noise  $\varepsilon$  satisfy  $\|\varepsilon\| \leq \sigma$  and the density condition  $\mathbb{P}(\sigma - \delta \leq \|\varepsilon\| \leq \sigma) \gtrsim \delta^\alpha$  near its boundary. We conjecture that the minimax rate for estimating the support  $M$  satisfies:

$$\inf_{\widehat{M}_n} \sup_{Q=P_\varepsilon * P} \mathbb{E} \left[ d_H(M, \widehat{M}_n) \right] \asymp \left( \frac{\log n}{n} \right)^{\frac{k}{(\alpha \wedge 2)k + d}}.$$

**Geodesic Distance Estimation and Out-of-Sample Extensions:** In [ABL23], we established that estimating the support  $M$  in Hausdorff distance and recovering the geodesic metric  $d_M$  under the uniform  $L^\infty$  loss are statistically equivalent, both yielding the minimax rate  $\mathcal{O}((\log n/n)^{k/d})$  for boundaryless  $\mathcal{C}^k$ -submanifolds. However, the presence of a boundary  $\partial M \neq \emptyset$  disrupts this equivalence. Recent numerical evidence [ACC23] suggests that in “favorable” geometries—such as when  $M$  is isometric to a convex domain of  $\mathbb{R}^d$ —metric estimation can have faster rates than full support recovery in the  $\mathcal{C}^2$  case. This acceleration stems from the fact that shortest paths between generic interior points safely bypass the boundary, suggesting that metric learning could avoid the boundary-penalized rate  $\mathcal{O}(n^{-2/(d+1)})$  to achieve the interior rate  $\mathcal{O}(n^{-2/d})$ .

In applications, geodesic distance estimation is frequently coupled with Classical Scaling (e.g., Isomap) to map sample points to low-dimensional representations. Evaluating the precision of such Non-Linear Dimension Reduction methods requires tailored criteria:

- *Descriptive framework:* Precision is measured over the sample points via the empirical  $\ell^\infty$  loss  $\max_{i,j} |\widehat{d}(X_i, X_j) - d_M(X_i, X_j)|$ . Because interior geodesic paths only need to hold for empirical pairs with high probability, global isometry constraints could likely be relaxed to domains with positive boundary reach.
- *Predictive framework:* Out-of-sample generalization for new data should be evaluated via an expected integrated risk  $\|\widehat{d} - d_M\|_{L^2}$ . This relaxation would allow for weaker geometric assumptions, and could be generalized using Gromov-Wasserstein distances [BMNN24] to assess the global isometric fidelity of the embeddings.

**Conjecture 1.2** (Boundary Bypass and  $L^2$  Metric Learning). Let  $P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^k$  be a volume-equivalent distribution supported on a  $d$ -dimensional  $\mathcal{C}^k$  submanifold  $M \subset \mathbb{R}^D$  with a  $\mathcal{C}^k$  boundary  $\partial M \neq \emptyset$ . There exists a computationally tractable intrinsic metric estimator  $\widehat{d}$  based on approximate geodesic equations such that the integrated  $L^2$  risk avoids the boundary penalty, achieving the interior minimax optimal rate:

$$\sup_{P \in \mathcal{P}_{\tau_{\min}, \mathbf{L}}^k} \mathbb{E} \left[ \int_{M \times M} |\widehat{d}(x, y) - d_M(x, y)|^2 dP(x) dP(y) \right] \asymp \left( \frac{\log n}{n} \right)^{\frac{2k}{d}}.$$

**Intrinsic Volume and Uniform Measure Estimation:** While support estimation provides a Hausdorff-type geometric guarantee, a natural measure-theoretic extension is to estimate the intrinsic volume  $\text{vol}(M)$  and the associated uniform probability distribution  $\mathcal{U}_M$ . Generating synthetic samples distributed according to  $\mathcal{U}_M$  can be thought of as a pre-processing step to correct sampling biases in empirical data.

In different convexity-constrained models, a variety of works suggest that support estimation rates of  $M$  are closely linked with both the estimation rates of the uniform measure on  $M$  [Div22], and of the intrinsic volume of  $M$  [BR16, ACC23] and [Bru14, Section 4.1.1]. We believe that this suggests a connection between geometric measure theory and set estimation.

**Conjecture 1.3** (Equivalence of support, volume and uniform-measure estimation). Let  $\mathcal{P}$  be a geometric model of distributions  $P$  supported on compact  $d$ -dimensional sets  $M = M(P) \subset \mathbb{R}^D$ , with densities bounded above and below with respect to the intrinsic volume measure on  $M$ . Define

$$\begin{aligned} R_n^{\text{Haus}}(\mathcal{P}) &:= \inf_{\widehat{M}_n} \sup_{P \in \mathcal{P}} \mathbb{E} \left[ d_{\text{H}}(M, \widehat{M}_n) \right], \\ R_n^{\text{Vol}}(\mathcal{P}) &:= \inf_{\widehat{V}_n} \sup_{P \in \mathcal{P}} \mathbb{E} \left[ \left| \text{vol}(M) - \widehat{V}_n \right| \right], \\ R_n^{\text{Unif}}(\mathcal{P}) &:= \inf_{\widehat{\mu}_n} \sup_{P \in \mathcal{P}} \mathbb{E} \left[ W_1(\mathcal{U}_M, \widehat{\mu}_n) \right], \end{aligned}$$

where  $\mathcal{U}_M := \text{vol}(M)^{-1} \mathcal{H}_{|M}^d$  denotes the uniform probability measure on  $M$ . Then,

$$R_n^{\text{Unif}}(\mathcal{P}) \asymp R_n^{\text{Haus}}(\mathcal{P}), \quad R_n^{\text{Vol}}(\mathcal{P}) \lesssim \left( \frac{R_n^{\text{Haus}}(\mathcal{P})}{\sqrt{n}} \right)^{1/2}.$$

**Higher-order singularities and corners:** While we have treated the boundary detection problem for  $\mathcal{C}^2$  manifolds, extending this to higher-order  $\mathcal{C}^k$  boundary smoothness is open. Another question is how to treat generalized corners and open-book singularities within mixed dimensions—determining mathematically whether an intersection should be treated as a separate lower-dimensional stratum or as a specific boundary singularity of the higher-dimensional components.

**Conjecture 1.4** (Estimating Manifolds with Corners). Consider i.i.d. data  $X_1, \dots, X_n$  sampled from an unknown distribution  $P \in \mathcal{P}_{k,c}^{d,D}$ , with density bounded above and below on its  $d$ -dimensional support  $M \subset \mathbb{R}^D$ . Assume  $M$  is a  $\mathcal{C}^k$ -smooth submanifold with reach bounded below, but allow for boundary outer-corners defined by at most  $c \in \{0, \dots, d\}$  distinct transversally intersecting  $\mathcal{C}^k$  constraints. Then, the minimax rate for estimating  $M$  satisfies:

$$\inf_{\widehat{M}_n} \sup_P \mathbb{E} \left[ d_{\text{H}}(M, \widehat{M}_n) \right] \asymp \left( \frac{\log n}{n} \right)^{\frac{k}{d+(k-1)c}}.$$

# Chapter 2

## Statistical foundations of generative modeling

### 2.1 Historical perspective and statistical foundations

Modeling and sampling from complex, high-dimensional probability distributions is a central problem in modern machine learning. Given a dataset  $\mathbb{X}_n = \{X_1, \dots, X_n\}$  of observations drawn from an unknown target distribution  $p^\star$  defined over an ambient space  $\mathbb{R}^D$ , the problem of generative modeling consists in constructing an algorithmic sampling mechanism capable of producing a new, synthetic data point  $\widehat{X}_{n+1} := A(\mathbb{X}_n, Z)$  where  $Z$  is an independent random seed, that is as statistically or perceptually indistinguishable from the true data distribution as possible. In probabilistic terms, we seek to construct algorithms  $A$  such that

$$\mathcal{L}(A(\mathbb{X}_n, Z) \mid \mathbb{X}_n) \simeq p^\star \quad \text{with high probability over random } \mathbb{X}_n \sim (p^\star)^{\otimes n}$$

#### 2.1.1 Beyond explicit density estimation

Historically, the generative problem was first formulated as one of explicit density estimation. Classical non-parametric methods, such as the Parzen–Rosenblatt kernel density estimator [Ros56, Par62], provide a standard framework for estimating an unknown density from samples. Yet, for the reasons recalled in Chapter , this ambient-space perspective becomes inadequate in high dimension unless one imposes additional structure on the data-generating distribution.

The structural prior most relevant to this memoir is the *manifold hypothesis*: the target distribution  $p^\star$  is not thought of as filling the whole ambient space  $\mathbb{R}^D$ , but as being supported on, or concentrated near, a lower-dimensional geometric object  $M \subset \mathbb{R}^D$  of intrinsic dimension  $d \ll D$ . Classical geometric inference, discussed in Chapter 1, takes this hypothesis literally and aims to reconstruct  $M$  and its differential or topological invariants. Modern generative modeling follows a complementary, implicit route: instead of estimating the support or the density directly, it seeks to learn a push-forward map, a

vector field, or stochastic dynamics transporting a simple reference distribution, such as a standard Gaussian, toward the complex target law  $p^*$ .

### 2.1.2 Main paradigms in generative modeling

Deep learning led to a rapid succession of generative paradigms. In the early 2010s, Variational Autoencoders (VAEs) [KW13, RMW14] introduced probabilistic latent variable models optimized via amortized variational inference, providing a computationally tractable lower bound on the data log-likelihood (the ELBO). While VAEs offered a solid probabilistic foundation and stable training dynamics, the restrictive nature of the Gaussian prior and the surrogate reconstruction loss often led to generated samples that lacked high-frequency details.

Shortly thereafter, Generative Adversarial Networks (GANs) [GPAM<sup>+</sup>14] introduced a different approach. GANs reframed the generation problem as a zero-sum minimax game between two neural networks, bypassing explicit likelihood estimation entirely. This adversarial mechanism pushed generators to produce samples of high visual quality. Yet, this empirical success also introduced important difficulties: the adversarial optimization landscape is highly non-convex and unstable, frequently suffering from mode collapse and vanishing gradients. Furthermore, Normalizing Flows [DKB14, RM15] provided another path, constructing complex distributions via a sequence of invertible transformations to allow exact likelihood computation, though their architectural bijectivity requirement imposes strong topological constraints.

More recently, the field has increasingly focused on Score-Based Generative Models (SGMs) and Denoising Diffusion Probabilistic Models (DDPMs) [SDWMG15, HJA20, SSDK<sup>+</sup>21]. These models construct a forward stochastic differential equation (SDE) that gradually adds Gaussian noise to the data, and then learn to reverse this SDE by estimating the *score function* of the perturbed data distributions via denoising score matching [Vin11]. Diffusion models have often outperformed GANs in practice by offering stable, non-adversarial training, exact likelihood bounds, and superior coverage of the data distribution. Concurrently, Flow Matching [LCBH<sup>+</sup>22] has abstracted this framework, replacing SDEs with ordinary differential equations (ODEs) driven by learned vector fields.

### 2.1.3 Statistical foundations

In this progression, new methodologies frequently supplant their predecessors based purely on sample quality or engineering heuristics, often long before a rigorous understanding of their success is established. This rapid empirical pace leaves several theoretical questions unresolved:

- What is the statistical complexity of these methods? Do they rigorously adapt to the intrinsic dimension of the data manifold, or do they suffer fundamentally from the ambient curse of dimensionality?
- When do these models truly generalize, and under what conditions are they minimax optimal distribution estimators?

Establishing a rigorous statistical framework for generative models is therefore important. It allows us to derive minimax optimal rates, understand the propagation of approximation and estimation errors over continuous time, and guarantee that the generated samples remain compatible with the support and smoothness of the underlying data distribution. Furthermore, a unified framework can identify connections between different methods, paving the way for principled algorithmic improvements rather than heuristics.

In the remainder of this chapter, we present the formalization of modern generative models, based on our recent contributions. We begin by investigating GANs from a minimax optimal transport perspective (Section 2.2). We then study the continuous-time dynamics of Score-Based Generative Models, establishing their optimal convergence rates (Section 2.3). Seeking a broader abstraction, we unify continuous-time models into the estimation of Markovian infinitesimal generators in the *Generator Matching* framework (Section 2.4). Finally, we formalize the practice of conditional generation and diffusion guidance, proposing a stochastic optimal control formulation to adaptively govern the generative dynamics (Section 2.5).

## 2.2 Generative adversarial networks as minimax distribution estimators

The classical paradigm of density estimation relies heavily on the maximization of the likelihood, implicitly assuming that the target distribution  $p^*$  admits a tractable density with respect to the Lebesgue measure on the ambient space  $\mathbb{R}^D$ . In high-dimensional regimes, this assumption is often violated: data typically concentrate on lower-dimensional structures, making them mutually singular with respect to absolutely continuous ambient models. In such settings, classical discrepancies such as the Kullback-Leibler divergence and standard likelihood methods can become theoretically and numerically unstable [SBD21].

Generative Adversarial Networks (GANs) [GPAM<sup>+</sup>14] circumvent this limitation by reformulating the generative learning problem as a likelihood-free, adversarial competition between two functional entities: a generator  $g$  and a discriminator  $D$ . Given a known, easy-to-sample base distribution  $\mu_0$  (such as a standard Gaussian or a uniform measure) defined on a latent space  $\mathcal{Z}$ , the generator acts as a parameterized push-forward map  $g : \mathcal{Z} \rightarrow \mathbb{R}^D$ . It transforms the latent noise  $Z \sim \mu_0$  into a synthetic sample  $\widehat{X} = g(Z)$ . The objective is to match the induced push-forward measure  $g_{\#}\mu_0$  to the target distribution  $\mu = p^*$ . By relying exclusively on sampling rather than explicit density evaluation, GANs provide a natural framework for modeling distributions supported on low-dimensional manifolds.

### 2.2.1 Integral probability metrics and the Wasserstein formulation

To quantify the discrepancy between the generated distribution  $g_{\#}\mu_0$  and the target distribution  $\mu$ , GANs employ an Integral Probability Metric (IPM) [Mül97]. For a chosen

symmetric class of discriminator functions  $\mathcal{D} \subset L^1(\mu) \cap L^1(g_{\#}\mu_0)$ , the IPM is defined as:

$$d_{\mathcal{D}}(\mu, g_{\#}\mu_0) := \sup_{D \in \mathcal{D}} (\mathbb{E}_{X \sim \mu}[D(X)] - \mathbb{E}_{Z \sim \mu_0}[D(g(Z))]).$$

The original GAN formulation implicitly utilized a discriminator class that minimizes the Jensen-Shannon divergence. However, when the supports of  $\mu$  and  $g_{\#}\mu_0$  are disjoint—a typical scenario in high dimensions—this divergence saturates to a constant, leading to vanishing gradients and training stagnation [AB17].

To address this issue, the Wasserstein GAN (WGAN) [ACB17] proposes using the Kantorovich-Rubinstein duality of the 1-Wasserstein distance  $W_1$ . By selecting  $\mathcal{D}$  to be the set of 1-Lipschitz functions  $\text{Lip}_1(\mathbb{R}^D)$ , the IPM metrizes weak convergence, remaining sensitive to separation between disjoint supports; this motivates, but does not guarantee, improved optimization behavior, even for mutually singular distributions [Vil09]. Other prominent alternatives include Sobolev GANs [MLS<sup>+</sup>18], which constrain the discriminator within Sobolev spaces, and MMD GANs [LCC<sup>+</sup>17], where  $\mathcal{D}$  lies in a Reproducing Kernel Hilbert Space.

In this work [SAL24], to study the non-parametric minimax properties of these estimators comprehensively, we generalize the discriminator class to Hölder balls  $\mathcal{H}_1^{\gamma}(\mathbb{R}^D, \mathbb{R})$  for a given smoothness level  $\gamma \geq 1$ . The Hölder IPM is defined by:

$$d_{\mathcal{H}_1^{\gamma}}(\mu, \nu) := \sup_{D \in \mathcal{H}_1^{\gamma}(\mathbb{R}^D, \mathbb{R})} (\mathbb{E}_{X \sim \mu}[D(X)] - \mathbb{E}_{Y \sim \nu}[D(Y)]).$$

Note that for compactly supported distributions, the case  $\gamma = 1$  corresponds to the WGAN framework, as  $d_{\mathcal{H}_1^1}(\mu, \nu)$  is equivalent to the 1-Wasserstein distance  $W_1(\mu, \nu)$ .

Given a sample  $\mathbb{X}_n = \{X_1, \dots, X_n\}$  drawn i.i.d. from  $\mu$ , the empirical WGAN estimator restricts the generator to a parametric class  $\mathcal{G}$  (e.g., neural networks) and the discriminator to a parametric class  $\mathcal{D} \subset \mathcal{H}_1^{\gamma}(\mathbb{R}^D, \mathbb{R})$ , replacing the true expectation by the empirical measure. This yields the empirical saddle-point problem:

$$\hat{g} \in \arg \min_{g \in \mathcal{G}} \sup_{D \in \mathcal{D}} \left( \frac{1}{n} \sum_{i=1}^n D(X_i) - \frac{1}{n} \sum_{i=1}^n D(g(Z_i)) \right), \quad (2.1)$$

where  $Z_1, \dots, Z_n$  are independent (user-generated) samples with reference distribution  $\mu_0$ . While WGANs demonstrated significantly improved training empirically, their statistical properties regarding optimal distribution estimation long remained an open question. Specifically, our objective is to determine whether the empirical estimator  $\hat{g}_{\#}\mu_0$  acts as a minimax optimal distribution estimator of  $\mu$  for the distance  $d_{\mathcal{H}_1^{\gamma}}$ , and to describe how the complexities of the network classes  $\mathcal{G}$  and  $\mathcal{D}$  must be calibrated.

## 2.2.2 The low-dimensional hypothesis and the statistical model

To formally evaluate the adaptability of the empirical GAN estimator to the intrinsic dimension of the data, we model the low-dimensional structure by assuming that the target

measure  $\mu$  is generated by an unknown smooth mapping from a lower-dimensional latent space. We choose the  $d$ -dimensional flat torus  $\mathbb{T}^d = \mathbb{R}^d / \mathbb{Z}^d$  as the latent space, and  $\mu_0$  to be the uniform distribution  $\text{Unif}$  on  $\mathbb{T}^d$ . This choice ensures compactness and allows the application of harmonic analysis tools without the technicalities associated with boundary conditions.

In addition to smoothness, we enforce geometric regularity to ensure the support behaves as a well-defined topological space.

**Definition 2.1** (Manifold and density regularity conditions). Let  $g \in \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^D)$ .

1. The map  $g$  satisfies the  $K$ -manifold regularity condition if it is injective and its image  $M_g = g(\mathbb{T}^d)$  is a  $d$ -dimensional submanifold with reach  $\tau_{M_g} \geq K^{-1}$ .
2. The map  $g$  satisfies the  $K$ -density regularity condition if its differential satisfies

$$\inf_{u \in \mathbb{T}^d} \lambda_{\min} \left( \nabla g(u)^\top \nabla g(u) \right)^{1/2} \geq K^{-1}.$$

These conditions prevent self-intersections or cusps, and ensure that the push-forward measure is bounded away from zero and infinity relative to the intrinsic volume measure of the manifold. We now state the statistical model.

**Assumption A1** (Target distribution). There exists an unknown mapping  $g^* \in \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^D)$  satisfying the  $K$ -manifold and  $K$ -density regularity conditions such that the target measure  $\mu$  is given by:

$$\mu = g_{\#}^* U \quad \text{where } U := \text{Unif}(\mathbb{T}^d).$$

This formulation reflects the generation process: the intrinsic dimension  $d \ll D$  captures the effective degrees of freedom of the data, while the smoothness  $\beta + 1$  restricts the complexity of the underlying embedding. These assumptions deliberately restrict the target to a smooth embedded manifold without self-intersections or cusps. They are stronger than a generic smooth push-forward model and are used to control both the geometry of the support and the density induced by the parameterization. Under Assumption A1, standard non-parametric lower bounds [Tsy09, TZ22] indicate that the minimax optimal rate for estimating  $\mu$  under the  $d_{\mathcal{H}_1^\gamma}$  metric scales as  $n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-1/2}$ . Our objective is to establish that the empirical estimator  $\hat{g}$  achieves this rate.

### 2.2.3 Bias-variance trade-off in adversarial learning

The statistical performance of the estimator (2.1) is governed by a balance between the expressive power of the generator class  $\mathcal{G}$  and the discriminative capacity of  $\mathcal{D}$ . Unlike standard empirical risk minimization, the minimax nature of (2.1) introduces two competing approximation errors. For chosen classes  $\mathcal{G} \subset \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^D)$  and  $\mathcal{D} \subset \mathcal{H}_1^\gamma(\mathbb{R}^D, \mathbb{R})$ , we define the generator approximation error and the discriminator approximation gap

respectively as:

$$\Delta_{\mathcal{G}} := \inf_{g \in \mathcal{G}} d_{\mathcal{H}_1^\gamma}(g_{\#}U, g_{\#}^*U), \quad (2.2)$$

$$\Delta_{\mathcal{D}} := \sup_{g \in \mathcal{G}} \left\{ d_{\mathcal{H}_1^\gamma}(g_{\#}U, g_{\#}^*U) - d_{\mathcal{D}}(g_{\#}U, g_{\#}^*U) \right\}. \quad (2.3)$$

The term  $\Delta_{\mathcal{G}}$  measures the capacity of the generator class to approximate the target distribution in the prescribed metric. Conversely,  $\Delta_{\mathcal{D}}$  quantifies how much discriminative power is lost by restricting the supremum from the full Hölder ball  $\mathcal{H}_1^\gamma$  down to the parametric class  $\mathcal{D}$ .

By analyzing the concentration of the empirical processes indexed by  $\mathcal{G}$  and  $\mathcal{D}$  using Dudley's entropy bound, we establish a general oracle inequality that encapsulates the bias-variance trade-off for GANs.

**Proposition 2.2** (Empirical risk bound for GANs [BKS23, SAL24]). *Suppose Assumption A1 holds. For any compact classes  $\mathcal{G} \subset \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^D)$  and  $\mathcal{D} \subset \mathcal{H}_1^\gamma(\mathbb{R}^D, \mathbb{R})$ , the expected error of the empirical GAN estimator  $\hat{g}$  defined in (2.1) is bounded by:*

$$\mathbb{E} \left[ d_{\mathcal{H}_1^\gamma}(\hat{g}_{\#}U, g_{\#}^*U) \right] \lesssim \Delta_{\mathcal{G}} + \Delta_{\mathcal{D}} + \frac{1}{\sqrt{n}} \left[ \log \left( n |\mathcal{G}_{1/n}| |\mathcal{D}_{1/n}| \right) \right]^{(\frac{1}{2} - \frac{\gamma}{d}) \vee 0}.$$

where  $|\mathcal{G}_{1/n}|$  and  $|\mathcal{D}_{1/n}|$  denote the  $1/n$ -covering numbers of the respective classes under the supremum norm.

This proposition highlights the adversarial tension: a richer discriminator class  $\mathcal{D}$  reduces the bias  $\Delta_{\mathcal{D}}$  but inflates the variance term via the covering entropy  $\log |\mathcal{D}_{1/n}|$ . Conversely, a sufficiently expressive generator class  $\mathcal{G}$  is required to drive  $\Delta_{\mathcal{G}}$  down. Achieving the minimax rate requires ensuring that  $\Delta_{\mathcal{G}}$  and  $\Delta_{\mathcal{D}}$  decay at the optimal rate  $n^{-\frac{\beta+\gamma}{2\beta+d}}$ , while strictly controlling the metric entropy of both classes such that  $\log |\mathcal{G}_{1/n}| \vee \log |\mathcal{D}_{1/n}| \lesssim n^{\frac{d}{2\beta+d}}$ .

This result formalizes an empirical observation [SEKS21]: overly expressive discriminators may overfit finite samples. A discriminator class with unconstrained capacity would result in a large covering number, causing the statistical bound to diverge and forcing the generator to memorize the empirical measure.

## 2.2.4 From smooth discrimination to rougher IPMs: the interpolation inequality

The preceding oracle bound suggests a natural but somewhat inefficient methodology and proof strategy: for each target regularity  $\gamma$ , one could choose a discriminator class approximating  $\mathcal{H}_1^\gamma$  and analyze the corresponding GAN estimator. This would produce, for each value of  $\gamma$ , a different estimator adapted to the metric  $d_{\mathcal{H}_1^\gamma}$ . In particular, the Wasserstein case  $\gamma = 1$  would be handled by training against Lipschitz discriminators.

The approach in [SAL24] is different. Rather than training a separate GAN for each IPM, we train and analyze the estimator through a smoother discriminator class, of regularity  $\beta + 1$ , and then transfer this control to rougher IPMs. This is useful because smoother discriminator classes have smaller statistical complexity: they are less prone to overfitting the empirical measure and their covering numbers are easier to control at the minimax scale.

In the full-dimensional density setting, such a transfer is a classical and rather direct interpolation phenomenon. If two probability measures have  $\beta$ -smooth densities with respect to the same Lebesgue measure, then

$$\int D(x)(p(x) - q(x)) \, dx$$

can be analyzed by decomposing both the test function and the density difference into wavelets. Low frequencies are controlled by the smooth IPM, while high frequencies are controlled by the smoothness of the densities. This gives an interpolation inequality between Hölder IPMs.

The manifold setting is more delicate. The measures  $g_{\#}U$  and  $g_{\#}^*U$  are singular with respect to the Lebesgue measure on  $\mathbb{R}^D$ , and their supports may be two different  $d$ -dimensional submanifolds. There is therefore no common ambient density difference on which the usual full-dimensional argument could act. One must first compare the two manifolds, relate them through a sufficiently regular projection-like map, and only then compare the induced densities on a common support. This is the main geometric content behind the interpolation inequality.

This is also why the smoothness and regularity of the estimated generator are essential. The interpolation result is not a generic consequence of using smooth neural network activations. It applies only when both the target generator and the candidate generator remain in a controlled  $\mathcal{H}^{\beta+1}$  class and satisfy the required manifold and density regularity assumptions. In particular, when the theorem is applied to the empirical estimator  $\hat{g}$ , one must ensure that  $\hat{g}$  itself has controlled  $\mathcal{H}^{\beta+1}$  norm and does not create degenerate parametrizations. This explains the latter wavelet-based construction of the generator class  $\mathcal{G}$ : it is designed not only to approximate  $g^*$ , but also to keep the selected generator within the regularity regime where interpolation can be used.

**Theorem 2.3** (Hölder IPM interpolation [SAL24]). *Let  $g, g^* \in \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^D)$  satisfy the manifold and density regularity conditions of Definition 2.1. There exist constants  $C > 0$  such that, if*

$$d_{\mathcal{H}_1^{\beta+1}}(g_{\#}U, g_{\#}^*U) \leq C^{-1},$$

*then, for all  $\varepsilon \in (0, 1)$  and all  $\gamma \in [1, \beta + 1]$ ,*

$$d_{\mathcal{H}_1^{\gamma}}(g_{\#}U, g_{\#}^*U) \lesssim \left( d_{\mathcal{H}_1^{\beta+1}}(g_{\#}U, g_{\#}^*U)^{\frac{\beta+\gamma}{2\beta+1}} + \varepsilon \right) \text{polylog}(\varepsilon^{-1}).$$

The theorem shows that, within the regular manifold model, control in the smooth metric  $d_{\mathcal{H}_1^{\beta+1}}$  automatically yields control in all rougher metrics  $d_{\mathcal{H}_1^{\gamma}}$ ,  $\gamma \in [1, \beta + 1]$ . Thus

the GAN need not be trained against a different discriminator class for each target IPM. A single smooth discriminator class is enough, provided the generator class is also constructed so that the resulting estimator remains smooth and geometrically non-degenerate.

At the level of rates, this gives the desired transfer. Up to logarithmic factors, the smooth-metric analysis yields a bound of order  $n^{-\frac{2\beta+1}{2\beta+d}}$ . Applying Theorem 2.3 gives  $(n^{-\frac{2\beta+1}{2\beta+d}})^{\frac{\beta+\gamma}{2\beta+1}} = n^{-\frac{\beta+\gamma}{2\beta+d}}$ . This recovers the minimax rate for every  $\gamma \in [1, \beta + 1]$ , including the Wasserstein case  $\gamma = 1$ .

The role of Theorem 2.3 should therefore not be confused with the separate generator approximation argument. The approximation of  $g^*$  by elements of  $\mathcal{G}$  is handled by the wavelet neural network construction. The interpolation inequality has a different purpose: it turns convergence in a smooth, statistically manageable adversarial metric into convergence in the rougher IPMs in which the estimator is ultimately evaluated.

### 2.2.5 Constructive neural network architectures via wavelets

To implement the GAN estimator and achieve the bounds in Proposition 2.2, the neural network classes  $\mathcal{G}$  and  $\mathcal{D}$  must simultaneously exhibit approximation capabilities, controlled covering numbers but also the quite stringent regularity constraints  $\mathcal{G} \subset \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^D)$  and  $\mathcal{D} \subset \mathcal{H}_1^\gamma(\mathbb{R}^D, \mathbb{R})$ . Traditional fully connected networks with generic bounds often yield suboptimal complexity due to the curse of dimensionality over the ambient space  $\mathbb{R}^D$ .

To resolve this, our construction of  $\mathcal{G}$  and  $\mathcal{D}$  explicitly mimics a multi-resolution wavelet approximation. Drawing from harmonic analysis [Dau92], any smooth function  $f$  can be decomposed into a coarse approximation and a sum of high-frequency details localized in space and frequency. We design our neural networks (utilizing ReQU or tanh activation functions) to approximate the translated and dilated versions of Daubechies scaling functions and wavelets.

By explicitly parameterizing the networks to learn the wavelet coefficients up to a carefully calibrated resolution level  $J \asymp \log(n)$ , we obtain structured functional classes that mirror the approximation properties of periodized Besov spaces  $\mathcal{B}_{\infty,\infty}^{\beta+1}(\mathbb{T}^d, \mathbb{R}^D)$ . This architectural choice provides an explicit, theoretical prescription for setting the depth, width, and sparsity of the networks in generative adversarial training, proving that carefully scaled neural architectures are both sufficient for minimax optimal generative modeling.

### 2.2.6 Minimax optimality of the GAN estimator

When synthesizing these components, we demonstrate that the resulting empirical GAN estimator has rates governed by the intrinsic geometry of the data, with rates depending on the intrinsic rather than ambient dimension  $D$ .

**Theorem 2.4** (Minimax optimality of WGANs [SAL24]). *Suppose the target distribution satisfies Assumption A1. For any target IPM regularity  $\gamma \in [1, \beta + 1]$ , there exist explicitly constructible classes of neural networks  $\mathcal{G}$  and  $\mathcal{D}$  (with architectures scaled according to*

$n, d$ , and  $\beta$ ) such that the GAN estimator  $\hat{g}$  defined in (2.1) achieves the minimax optimal rate:

$$\mathbb{E} \left[ d_{\mathcal{H}_1^\gamma}(\mu, \hat{g}_\# U) \right] \lesssim \left( n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-1/2} \right) \text{polylog}(n).$$

We establish that GAN estimators, when calibrated with appropriately constrained neural networks, achieve the minimax optimal convergence rate for distribution estimation over non-parametric classes. The resulting rate exponent depends exclusively on the intrinsic dimension  $d$  and the regularities  $\beta$  and  $\gamma$ .

This provides a rigorous proof that WGANs are minimax optimal distribution estimators over manifolds [TZ22], connecting explicit geometric reconstruction methods and implicit generative adversarial representations. It improves upon previous suboptimal GAN rates found in the literature [CJLZ20, SBD21, CK22, BKS23]. Furthermore, because our constructive discriminator class  $\mathcal{D}$  specifically targets the smoother  $\mathcal{H}_1^{\beta+1}$  space via Theorem 2.3, our results show that a *single* unified GAN architecture is simultaneously minimax optimal for all evaluation metrics  $\gamma \in [1, \beta + 1]$ , highlighting the adaptivity of the adversarial estimator when structured properly.

## 2.3 Score-based generative models and diffusions

Score-Based Generative Models (SGMs) [SSDK<sup>+</sup>21] and Denoising Diffusion Probabilistic Models (DDPMs) [HJA20] replace the adversarial training problem of GANs by a regression problem. The object to be learned is the score of a family of noisy versions of the data distribution. Once this score is estimated, it is used to run a reverse-time dynamics from a simple reference distribution back to the data distribution.

In [SAL25], we study this construction from a non-asymptotic minimax point of view. The main question is whether denoising score matching, combined with a suitable reverse diffusion or probability-flow ODE, can achieve the minimax Wasserstein rate, up to logarithmic factors, for estimating a smooth distribution. The answer relies on three ingredients: regularity properties of the score induced by assumptions on the target density, a stability estimate for the backward dynamics, and an oracle inequality for the empirical score-matching estimator.

Here, for simplicity, we will focus on the full-dimensional case, where the population distribution is absolutely continuous with respect to the ambient Lebesgue measure on  $\mathbb{R}^D$ . The model being full-dimensional, the convergence rates hence naturally depend on the ambient dimension.

### 2.3.1 Forward dynamics and marginal reversals

We consider the variance-preserving Ornstein–Uhlenbeck forward process

$$\begin{cases} dX_t = -X_t dt + \sqrt{2}\sigma dB_t, \\ X_0 \sim p^*(x) dx, \end{cases} \quad (2.4)$$

where  $\sigma > 0$  and  $(B_t)_{t \geq 0}$  is a standard Brownian motion in  $\mathbb{R}^D$ . The solution is explicit:

$$X_t \sim e^{-t} X_0 + \sigma_t Z, \quad \sigma_t = \sigma \sqrt{1 - e^{-2t}}, \quad (2.5)$$

with  $Z \sim \mathcal{N}(0, \text{Id})$  independent of  $X_0$ . We write  $p_t$  for the density of  $X_t$ . Thus  $p_0 = p^*$ , while  $p_t$  converges in distribution as  $t \rightarrow \infty$  to the Gaussian equilibrium

$$p_\infty(x) dx = \mathcal{N}(0, \sigma^2 \text{Id}).$$

The score of the noisy marginal is

$$s^*(t, x) = \nabla \log p_t(x). \quad (2.6)$$

It is the unknown quantity appearing in the reverse-time drift. More precisely, let  $\bar{T} > 0$  be a terminal time and let  $b_t \geq 0$  be a deterministic backward diffusion coefficient. Anderson's time reversal formula gives the backward dynamics

$$\begin{cases} d\widehat{X}_t = [\widehat{X}_t + (\sigma^2 + b_t^2) s^*(\bar{T} - t, \widehat{X}_t)] dt + \sqrt{2} b_t d\bar{B}_t, \\ \widehat{X}_0 \sim p_{\bar{T}}(x) dx. \end{cases} \quad (2.7)$$

Then  $\widehat{X}_t \sim p_{\bar{T}-t}$  for all  $t \in [0, \bar{T}]$ . The choice  $b_t = \sigma$  corresponds to a stochastic DDPM-type sampler, while  $b_t = 0$  gives the deterministic probability-flow ODE.

In practice, three modifications are made. The exact score  $s^*$  is replaced by an estimator  $\hat{s}$ , the exact initial distribution  $p_{\bar{T}}$  is replaced by the limiting Gaussian  $p_\infty$ , and the reverse dynamics is stopped at time  $\bar{T} - \underline{T}$  for some  $\underline{T} > 0$ . The generated law is therefore obtained from

$$\begin{cases} d\widehat{X}_t = [\widehat{X}_t + (\sigma^2 + b_t^2) \hat{s}(\bar{T} - t, \widehat{X}_t)] dt + \sqrt{2} b_t d\bar{B}_t, \\ \widehat{X}_0 \sim p_\infty(x) dx, \end{cases} \quad (2.8)$$

and we set

$$\hat{p}_{\text{SGM}} := \mathcal{L}(\widehat{X}_{\bar{T}-\underline{T}} \mid \mathbf{X}_n).$$

This decomposition separates three errors: the initialization error, due to replacing  $p_{\bar{T}}$  by  $p_\infty$ ; the early-stopping error, due to stopping at  $p_{\underline{T}}$  rather than at  $p_0 = p^*$ ; and the score error, due to replacing  $s^*$  by  $\hat{s}$ . The initialization and early-stopping errors are controlled using the explicit Ornstein–Uhlenbeck dynamics. The score error is handled through score estimation and SDE stability.

### 2.3.2 Denoising score matching

The natural risk for score estimation is the time-integrated Fisher loss

$$\mathcal{E}_{\underline{T}, \bar{T}}(s) = \int_{\underline{T}}^{\bar{T}} \mathbb{E} [\|s(t, X_t) - s^*(t, X_t)\|^2] dt. \quad (2.9)$$

Although this expression involves the unknown score  $s^*$ , the Ornstein–Uhlenbeck structure gives an equivalent denoising objective. If  $X \sim p^*$  and  $Z \sim \mathcal{N}(0, \text{Id})$  are independent, then

$$X_t = e^{-t}X + \sigma_t Z.$$

The conditional score of  $X_t$  given  $X$  is explicit and equal to  $-Z/\sigma_t$ . The denoising score identity [Vin11] yields, up to an additive constant independent of  $s$ ,

$$\mathbb{E} \left[ \|s(t, X_t) - s^*(t, X_t)\|^2 \right] = \mathbb{E} \left[ \left\| s(t, e^{-t}X + \sigma_t Z) + \frac{Z}{\sigma_t} \right\|^2 \right] + \text{const.} \quad (2.10)$$

Given observations  $\mathbb{X}_n = \{X_1, \dots, X_n\}$  drawn independently from  $p^*$ , the empirical denoising score matching estimator is defined by

$$\hat{s} \in \arg \min_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^n \int_{\underline{T}}^{\bar{T}} \mathbb{E}_{Z \sim \mathcal{N}(0, \text{Id})} \left[ \left\| s(t, e^{-t}X_i + \sigma_t Z) + \frac{Z}{\sigma_t} \right\|^2 \right] dt, \quad (2.11)$$

where  $\mathcal{S}$  is a class of candidate time-dependent vector fields. The lower cutoff  $\underline{T} > 0$  is needed because  $\sigma_t \rightarrow 0$  as  $t \rightarrow 0$ , so the conditional score  $-Z/\sigma_t$  diverges as  $t \rightarrow 0$ .

### 2.3.3 The statistical model and the regularity of the score

The score should not be assumed regular independently of the data distribution. It is a nonlinear function of the noisy density

$$p_t(x) = \int_{\mathbb{R}^D} \varphi_{\sigma_t}(x - e^{-t}y) p^*(y) dy.$$

Its regularity depends on the smoothness of  $p^*$ , but also on the behavior of  $p^*$  near the boundary of its support or at infinity. If one only assumes that  $p^*$  is  $\beta$ -Hölder, the score may still fail to satisfy useful uniform bounds in regions where  $p_t$  is very small. This matters because the reverse dynamics requires a one-sided Lipschitz control of the drift, and because score estimation requires quantitative approximation bounds.

The following assumption is used to obtain these two controls. It imposes a  $\beta$ -Hölder perturbation of a strongly convex reference potential. The Hölder part contains the non-parametric smoothness assumption, while the convex potential controls the geometry of the support and the decay of the density.

**Assumption A2** (Target distribution). Let  $\beta > 0$  and  $K \geq 1$ . The target density  $p^* : \mathbb{R}^D \rightarrow \mathbb{R}_+$  belongs to  $\mathcal{A}_K^\beta$  if it can be written as

$$p^*(x) = \exp(-u(x) + a(x)),$$

where  $u : \mathbb{R}^D \rightarrow \mathbb{R} \cup \{\infty\}$  and  $a : \mathbb{R}^D \rightarrow \mathbb{R}$  satisfy the following properties, with  $\Omega = \text{Support}(p^*)$ :

- i)  $a \in \mathcal{H}_K^\beta(\Omega)$  and  $u \in \mathcal{C}^{1+\max\{1, \lfloor \beta \rfloor\}}(\Omega)$ ;

- ii)  $\nabla^2 u(x) \succeq K^{-1} \text{Id}$  for all  $x \in \Omega$ ;
- iii)  $\|\arg \min u\| \leq K$ ;
- iv) for all  $x, y \in \Omega$  such that

$$\|x - y\| \leq (K(1 \vee |u(x)|))^{-K},$$

and for all  $k \in \{0, \dots, \lfloor \beta \rfloor + 1\}$ ,

$$\|\nabla^k u(y)\| \leq K(1 + |u(x)|^K).$$

The decomposition  $p^* = \exp(-u + a)$  can be interpreted as follows. The strongly convex part  $u$  gives concentration and prevents uncontrolled decay near the boundary or at infinity. The perturbation  $a$  contains the  $\beta$ -Hölder non-parametric component. This class includes compactly supported distributions on convex sets and full-support distributions with controlled tails. It also includes non-log-concave examples, such as bounded Hölder perturbations of finite mixtures of non-degenerate Gaussians with controlled means and covariances.

The consequence used below is that the score inherits the regularity needed for the analysis of the backward dynamics and of score matching.

To state the estimates in a more compact form, we use the noise scale  $\sigma_t$  defined in (2.5). For  $\alpha \geq 0$ , set

$$\text{Reg}_{\beta,t}(\alpha) = (1 + \sigma_t^{-1})^{(\alpha-\beta)_+}.$$

We call this quantity the regularity factor. It is equal to 1 when the available smoothness  $\beta$  is at least the requested order  $\alpha$ . Otherwise, it records the small-time singularity introduced by the missing smoothness. Since  $\sigma_t \asymp \sqrt{t}$  as  $t \rightarrow 0$ , one has

$$\text{Reg}_{\beta,t}(\alpha) \asymp 1 + t^{-\frac{(\alpha-\beta)_+}{2}}.$$

**Theorem 2.5** (Space-time regularity of the score [SAL25]). *Assume that  $p^* \in \mathcal{A}_K^\beta$ . Let  $s^*(t, x) = \nabla \log p_t(x)$  be the score of the Ornstein–Uhlenbeck marginal (2.4). Then the following properties hold.*

- The score has controlled growth:

$$\|s^*(t, x)\| \leq C(1 + t^{-1})(1 + \|x\|), \quad t > 0, \quad x \in \mathbb{R}^D. \quad (2.12)$$

- The score satisfies the one-sided Lipschitz estimate

$$\sup_{x \in \mathbb{R}^D} \lambda_{\max} \left( \nabla s^*(t, x) + \frac{1}{\sigma^2} \text{Id} \right) \lesssim e^{-2t} \text{Reg}_{\beta \wedge 1, t}(2), \quad t > 0. \quad (2.13)$$

*In particular, the right-hand side is integrable on  $(0, \infty)$ .*

- For every  $\varepsilon \in (0, 1/4)$  and every  $t > 0$ , there exists a convex subset  $A_t^\varepsilon \subset B(0, \text{polylog}(\varepsilon^{-1}))$  such that  $p_t(A_t^\varepsilon) \geq 1 - \varepsilon$ . On this set, the score has the following Hölder bounds. For every  $k \geq 0$  and every  $x \in A_t^\varepsilon$ ,

$$\left\| s^*(\cdot, x) + \frac{x}{\sigma^2} \right\|_{\mathcal{H}^k([t, \infty))} \lesssim e^{-t} \text{Reg}_{\beta, t}(1 + 2k), \quad (2.14)$$

and, for every  $\gamma \geq 0$  and every  $k \geq 0$ ,

$$\left\| \partial_t^k \left[ s^*(t, \cdot) + \frac{\cdot}{\sigma^2} \right] \right\|_{\mathcal{H}^\gamma(A_t^\varepsilon)} \lesssim e^{-t} \text{Reg}_{\beta, t}(1 + 2k + \gamma). \quad (2.15)$$

Together, (2.12) and (2.13) turn out to be the basic well-posedness conditions for the true backward dynamic (2.7) on  $[0, \bar{T}]$ , especially when  $b_t = 0$ .

Note that the one-sided Lipschitz estimate (2.13) is much weaker than global Lipschitz continuity, but it is the property used in the stability estimate for the backward equation. The Hölder estimates (2.14) and (2.15) show how the smoothness of  $p^*$  appears along the forward diffusion path. For small  $t$ , the bounds deteriorate because  $\sigma_t^2$  is of order  $t$ . For large  $t$ , the score approaches the linear Gaussian score  $x \mapsto -x/\sigma^2$ , and the nonlinear part decreases exponentially.

### 2.3.4 Quantitative SDE stability and the score-matching oracle inequality

We now state the two estimates that connect score estimation to the error of the generated law. The first is analytic: it controls the Wasserstein distance between two diffusions by an integrated drift error. The second is statistical: it controls the integrated score error of the empirical score-matching estimator.

**Proposition 2.6** ( $W_1$  stability of generative diffusions [SAL25]). *Let  $(X_t)_{\underline{T} \leq t \leq \bar{T}}$  and  $(\bar{X}_t)_{\underline{T} \leq t \leq \bar{T}}$  solve*

$$\begin{aligned} dX_t &= a_t(X_t) dt + r_t dB_t, & X_{\underline{T}} &\sim p_{\underline{T}}, \\ d\bar{X}_t &= \bar{a}_t(\bar{X}_t) dt + r_t dB_t, & \bar{X}_{\underline{T}} &\sim p_{\underline{T}}, \end{aligned}$$

*with the same space-homogeneous diffusion coefficient  $r_t$ . Let  $X_t \sim p_t$  and  $\bar{X}_t \sim \bar{p}_t$ . Assume that  $\bar{a}_t$  satisfies the one-sided Lipschitz bound*

$$\sup_{x \in \mathbb{R}^D} \lambda_{\max}(\nabla \bar{a}_t(x)) \leq \bar{L}_t.$$

*Then*

$$W_1(p_{\bar{T}}, \bar{p}_{\bar{T}}) \leq \int_{\underline{T}}^{\bar{T}} \int_{\mathbb{R}^D} \exp\left(\int_t^{\bar{T}} \bar{L}_u du\right) \|a_t(x) - \bar{a}_t(x)\| p_t(x) dx dt. \quad (2.16)$$

Applied to the backward SDEs (2.7) and (2.8), the drift difference is proportional to  $\hat{s} - s^*$ . The exponential factor can be bounded uniformly in  $n$  if the estimated score is chosen in a class satisfying the same type of one-sided Lipschitz bound as the true score. Thus the network class is used both for approximation and for stability.

The statistical input is an oracle inequality for empirical denoising score matching. Let

$$\mathcal{S}_{[\underline{T}, \bar{T}]} = \left\{ s_{[\underline{T}, \bar{T}] \times \mathbb{R}^D} \mid s \in \mathcal{S} \right\}$$

and let  $\mathcal{N}(\mathfrak{S}, \|\cdot\|_\infty, \delta)$  denote the covering number of  $\mathfrak{S}$  at scale  $\delta$  for the uniform norm.

**Proposition 2.7** (Oracle inequality for denoising score matching [SAL25]). *Assume that  $p^*$  is sub-Gaussian and that its score satisfies the regularity properties stated in Theorem 2.5. Let  $\hat{s}$  be an empirical risk minimizer of the denoising score matching criterion (2.11) over a class  $\mathcal{S}$  such that, for some  $V > 0$ ,*

$$\sup_{s \in \mathcal{S}} \sup_{t \in [\underline{T}, \bar{T}]} \left\| s(t, \cdot) + \frac{\cdot}{\sigma^2} \right\|_\infty \leq V.$$

Then, for  $n^{-1} \leq \underline{T} < \bar{T}$ ,

$$\mathbb{E} \left[ \mathcal{E}_{\underline{T}, \bar{T}}(\hat{s}) \right] \lesssim \inf_{s \in \mathcal{S}} \mathcal{E}_{\underline{T}, \bar{T}}(s) + \frac{c_n}{n} \log \left( \mathcal{N} \left( \mathcal{S}_{[\underline{T}, \bar{T}]}, \|\cdot\|_\infty, n^{-1} \right) \vee n \right), \quad (2.17)$$

where  $c_n \lesssim (1 + \underline{T}^{-1})(\bar{T} - \underline{T})\text{polylog}(n)$ .

The first term in (2.17) is the approximation error of the class  $\mathcal{S}$ . The second term is a complexity term expressed through the metric entropy of  $\mathcal{S}$ . The dependence on  $\underline{T}$  reflects the singular behavior of the denoising target near  $t = 0$ .

It is therefore enough to construct a function class  $\mathcal{S}$  with good approximation properties, controlled entropy, and a controlled one-sided Lipschitz constant.

### 2.3.5 Minimax convergence rates for SGMs

The class  $\mathcal{S}$  is built from tanh neural networks whose architecture depends on time. Near  $t = 0$ , the score is less regular and may have larger norm, so the approximation error is larger. At larger times, the nonlinear part of the score is smoother and smaller. The interval  $[\underline{T}, \bar{T}]$  is therefore divided into subintervals, and different network sizes are used on different subintervals.

The early-stopping and initialization parameters are chosen to balance the three errors described above. The initialization error decreases exponentially with  $\bar{T}$ , the early-stopping error is of order  $\sqrt{\underline{T}}$ , and the score estimation error is controlled by the oracle inequality. This leads to the following bound.

**Theorem 2.8** (Minimax convergence of score-based generative models [SAL25]). *Let  $D \geq 3$ ,  $\beta > 0$ ,  $K \geq 1$ , and let  $X_1, \dots, X_n$  be drawn independently from  $p^* \in \mathcal{A}_K^\beta$ . Assume that  $K^{-1} \leq \sigma \leq K$  and that the backward diffusion coefficient satisfies*

$$\|b\|_{L^\infty([0, \bar{T} - \underline{T}])} \leq K.$$

There exist time cutoff satisfying

$$\underline{T} \asymp n^{-\frac{2(\beta+1)}{2\beta+D}}, \quad \overline{T} \asymp \log n,$$

and a time-dependent class of tanh neural networks  $\mathcal{S}$  such that, if  $\hat{s}$  is obtained by denoising score matching (2.11) and  $\hat{p}_{\text{SGM}}$  denotes the generated law (2.8), then

$$\sup_{p^* \in \mathcal{A}_K^\beta} \mathbb{E} [W_1(p^*, \hat{p}_{\text{SGM}})] \lesssim n^{-\frac{\beta+1}{2\beta+D}} \text{polylog}(n). \quad (2.18)$$

Moreover, this rate matches the minimax lower bound up to logarithmic factors:

$$\inf_{\tilde{p}_n} \sup_{p^* \in \mathcal{A}_K^\beta} \mathbb{E} [W_1(p^*, \tilde{p}_n)] \geq cn^{-\frac{\beta+1}{2\beta+D}}, \quad (2.19)$$

where the infimum is taken over all estimators based on  $X_1, \dots, X_n$ .

This theorem concerns an exact continuous-time simulation and an exact empirical score-matching minimizer. Numerical time-discretization and optimization errors are not included. The rate  $n^{-(\beta+1)/(2\beta+D)}$  is the minimax rate for estimating a  $\beta$ -smooth density in  $\mathbb{R}^D$  under the Wasserstein-1 distance. The theorem shows that, in this model, denoising score matching followed by a reverse diffusion or probability-flow ODE satisfies this rate up to logarithmic factors.

The argument does not assume that the score is smooth a priori. Instead, the assumptions on the target density imply quantitative regularity of the score along the Ornstein–Uhlenbeck path. This regularity allows score approximation by neural networks, while the one-sided Lipschitz estimate ensures that score estimation errors are propagated stably by the backward dynamics.

## 2.4 Generator matching for Markovian generative models

The previous section described Score-Based Generative Models through one specific mechanism: an Ornstein–Uhlenbeck forward diffusion, a denoising score-matching objective, and a reverse-time diffusion or probability-flow ODE. This mechanism is important, but it is only one way of producing a time-dependent transformation from a simple distribution to the data distribution. Flow Matching [LCBH<sup>+</sup>22, AVE23], continuous-time discrete diffusion [CBDB<sup>+</sup>22], and jump-based generative models [LME24] can be described by a similar principle: fix a probability path and learn a dynamical system that realizes it *marginally*.

In [AS26], following the Generator Matching formalism of [H<sup>+</sup>24], we study this idea at the level of infinitesimal generators. One fixes a probability path

$$(p_t)_{0 \leq t \leq 1}, \quad p_0 \text{ simple}, \quad p_1 = p^*,$$

and looks for a continuous-time Markov process whose marginal law at time  $t$  is  $p_t$ . The score-based models of Section 2.3 correspond to a particular diffusion generator. Generator Matching keeps the same conditional-regression structure as denoising score matching, but replaces the score by the parameter field of a general Markov generator.

Throughout this section, when  $\mu$  is a measure and  $f$  is a test function, we write

$$\langle \mu, f \rangle = \int f \, d\mu.$$

The same bracket notation will also be used for vector and matrix inner products; the meaning will always be clear from the arguments.

### 2.4.1 Probability paths

Let  $\mathcal{S}$  be a measurable state space. In the score-based setting of Section 2.3, one has  $\mathcal{S} = \mathbb{R}^D$ . For discrete generative modeling,  $\mathcal{S}$  may instead be finite, for instance a vocabulary, a finite graph state space, or a product of categorical variables. The first modeling choice is a probability path  $(p_t)_{0 \leq t \leq 1}$  joining a tractable reference distribution  $p_0$  to the target distribution  $p^*$ .

As in flow matching and denoising score matching, the path is usually specified through conditional paths. Given a latent data point  $Z^\circ \sim p^*$ , one chooses an explicit conditional path

$$p_t(\cdot \mid Z^\circ = z^\circ), \quad p_0(\cdot \mid Z^\circ = z^\circ) = p_0, \quad p_1(\cdot \mid Z^\circ = z^\circ) = \delta_{z^\circ}.$$

The marginal path is then obtained by mixing over  $Z^\circ$ :

$$p_t(A) = \int_{\mathcal{S}} p_t(A \mid Z^\circ = z^\circ) p^*(dz^\circ), \quad A \subseteq \mathcal{S}. \quad (2.20)$$

Equivalently, conditionally on  $Z^\circ$ , one samples a static random variable

$$X_t^\circ \sim p_t(\cdot \mid Z^\circ),$$

so that  $X_t^\circ$  has marginal law  $p_t$ . The superscript  $\circ$  indicates that  $(X_t^\circ)_t$  is a static interpolation, not necessarily the trajectory of a Markov process.

On  $\mathbb{R}^D$ , a standard class of interpolations is

$$X_t^\circ = m_t(Z^\circ) + \sigma_t \xi^\circ, \quad \xi^\circ \sim p_0.$$

The usual flow-matching path is obtained by taking  $m_t(z) = tz$  and  $\sigma_t = 1 - t$ , namely

$$X_t^\circ = (1 - t)\xi^\circ + tZ^\circ.$$

This path is close to the Gaussian noising paths of Section 2.3, but it is written in the generative direction, from  $p_0$  to  $p^*$ . When the conditional law is Gaussian, the conditional score is explicit:

$$\nabla \log p_t(x \mid Z^\circ) = -\frac{x - m_t(Z^\circ)}{\sigma_t^2}.$$

The denoising identity (2.10) says that the marginal score is the conditional expectation of this explicit conditional score.

On an arbitrary state space, a simple alternative is the mixture path

$$p_t(dx \mid Z^\circ = z^\circ) = (1 - \kappa_t)p_0(dx) + \kappa_t\delta_{z^\circ}(dx), \quad (2.21)$$

where  $\kappa_0 = 0$ ,  $\kappa_1 = 1$ , and  $\kappa_t$  is non-decreasing. This path does not transport mass continuously in space. It keeps a fraction  $1 - \kappa_t$  of the reference law and moves the remaining mass to the endpoint  $z^\circ$ . It is therefore suitable for finite or non-Euclidean state spaces.

### 2.4.2 Generators realizing a probability path

Once the path  $(p_t)_t$  is fixed, the next question is whether it can be realized as the family of marginals of a continuous-time Markov process. Let  $(X_t)_{0 \leq t \leq 1}$  be such a process, with transition kernels  $k_{s,t}(x, dy)$ . For a class  $\mathcal{T}$  of test functions  $f : \mathcal{S} \rightarrow \mathbb{R}$ , its infinitesimal generator is

$$L_t f(x) = \lim_{h \downarrow 0} \frac{\int_{\mathcal{S}} f(y) k_{t,t+h}(x, dy) - f(x)}{h}, \quad (2.22)$$

whenever the limit exists. If the marginal law of  $X_t$  is  $p_t$ , then the weak Kolmogorov forward equation reads

$$\frac{d}{dt} \langle p_t, f \rangle = \langle p_t, L_t f \rangle, \quad f \in \mathcal{T}. \quad (2.23)$$

When the measures have densities, this is the weak form of  $\partial_t p_t = L_t^* p_t$ . For a deterministic flow, it is the continuity equation. For a diffusion, it is the Fokker–Planck equation. For a pure-jump process, it is the master equation.

The finite-state case is the simplest setting in which gradients do not appear. Suppose that  $\mathcal{S}$  is finite. A continuous-time Markov chain is described by transition rates

$$q_t(x, y) \geq 0, \quad x \neq y.$$

For small  $h > 0$ ,

$$\mathbb{P}(X_{t+h} = y \mid X_t = x) = h q_t(x, y) + o(h), \quad y \neq x,$$

and

$$\mathbb{P}(X_{t+h} = x \mid X_t = x) = 1 - h \sum_{y \neq x} q_t(x, y) + o(h).$$

The generator is therefore

$$L_t f(x) = \sum_{y \neq x} q_t(x, y) (f(y) - f(x)). \quad (2.24)$$

Writing  $p_t(x)$  for the mass of state  $x$ , the weak equation (2.23) is equivalent to the master equation

$$\dot{p}_t(y) = \sum_{x \neq y} p_t(x) q_t(x, y) - p_t(y) \sum_{x \neq y} q_t(y, x). \quad (2.25)$$

Thus, on a finite state space, learning the generator means learning the transition rates of a time-inhomogeneous Markov chain. This is the discrete counterpart of learning a drift field in flow matching or a score-dependent drift in score-based diffusion.

**Definition 2.9** (Weak realization of a probability path). Let  $(p_t)_{0 \leq t \leq 1}$  be a probability path on  $\mathcal{S}$ . A family of generators  $(L_t)_{0 \leq t \leq 1}$  is said to realize  $(p_t)_t$  on the test class  $\mathcal{T}$  if

$$\frac{d}{dt} \langle p_t, f \rangle = \langle p_t, L_t f \rangle, \quad f \in \mathcal{T}.$$

This definition is weak by design. It does not require us to write the transition kernels of the process explicitly. It only asks that the infinitesimal evolution of expectations along the path is generated by  $L_t$ . A weak realization need not, by itself, determine a well-posed Markov process. Existence and uniqueness require additional domain, positive-maximum-principle, conservativity, and martingale-problem assumptions.

### 2.4.3 The conditional-to-marginal principle

The conditional construction of the path is useful because the conditional dynamics are often explicit. Suppose that, for each endpoint  $z^\circ$ , there is a generator  $L_t^{z^\circ}$  that realizes the conditional path  $p_t(\cdot \mid Z^\circ = z^\circ)$ . The marginal generator is then obtained by conditioning on the current interpolated state.

**Proposition 2.10** (Conditional-to-marginal generator formula [H<sup>+</sup>24, AS26]). *Let  $(X_t^\circ)_{0 \leq t \leq 1}$  be the static interpolating process defined above. Assume that, conditionally on  $Z^\circ$ , the path admits a conditional generator  $L_t^{Z^\circ}$ . Under suitable regularity assumptions, the marginal path  $(p_t)_t$  is realized by the generator*

$$L_t f(x) = \mathbb{E} \left[ L_t^{Z^\circ} f(x) \mid X_t^\circ = x \right]. \quad (2.26)$$

Equivalently,

$$\frac{d}{dt} \mathbb{E}[f(X_t^\circ)] = \mathbb{E}[L_t f(X_t^\circ)].$$

*Proof.* For a test function  $f$ , the tower property gives

$$\mathbb{E}[f(X_t^\circ)] = \mathbb{E}[\mathbb{E}[f(X_t^\circ) \mid Z^\circ]].$$

Differentiating in time and using the conditional generator,

$$\frac{d}{dt} \mathbb{E}[f(X_t^\circ)] = \mathbb{E} \left[ L_t^{Z^\circ} f(X_t^\circ) \right].$$

Conditioning on  $X_t^\circ$  gives

$$\mathbb{E} \left[ L_t^{Z^\circ} f(X_t^\circ) \right] = \mathbb{E} \left[ \mathbb{E} \left[ L_t^{Z^\circ} f(X_t^\circ) \mid X_t^\circ \right] \right],$$

which is exactly (2.26).  $\square$

In the finite-state case, if the conditional generators have rates  $q_t^{z^\circ}(x, y)$ , the formula becomes

$$q_t(x, y) = \mathbb{E} \left[ q_t^{Z^\circ}(x, y) \mid X_t^\circ = x \right], \quad x \neq y. \quad (2.27)$$

Thus continuous-time discrete diffusion can be written as a regression problem for transition rates.

The mixture path (2.21) gives a concrete example. For  $t < 1$ , set

$$\lambda_t = \frac{\dot{\kappa}_t}{1 - \kappa_t}.$$

Then the conditional generator

$$L_t^{z^\circ} f(x) = \lambda_t (f(z^\circ) - f(x)) \quad (2.28)$$

realizes the conditional mixture path. On a finite state space, this means that the chain jumps to  $z^\circ$  at rate  $\lambda_t$  and makes no other jumps. Indeed,

$$\begin{aligned} \langle p_t(\cdot \mid z^\circ), L_t^{z^\circ} f \rangle &= \frac{\dot{\kappa}_t}{1 - \kappa_t} \int_{\mathcal{S}} (f(z^\circ) - f(x)) ((1 - \kappa_t)p_0(dx) + \kappa_t \delta_{z^\circ}(dx)) \\ &= \dot{\kappa}_t (f(z^\circ) - \langle p_0, f \rangle) \\ &= \frac{d}{dt} \langle p_t(\cdot \mid z^\circ), f \rangle. \end{aligned}$$

This example shows that jump generators are not only discrete analogues of diffusion generators. They are a Markovian representation of paths that move mass by jumps rather than continuously.

#### 2.4.4 Linear parameterizations of generators

The preceding formula defines the marginal generator, but it does not yet say what the neural network should output. A generator is an operator acting on test functions, whereas a neural network produces a finite-dimensional object. Generator Matching resolves this by using a linear parameterization of the generator.

**Definition 2.11** (Linear parameterization). Let  $V$  be an inner-product space, let  $\Omega \subseteq V$  be convex, and let  $K$  be a linear map acting on test functions. A family of generators  $(L_t)_t$  admits a linear parameterization if there exists a parameter field

$$F_t : \mathcal{S} \rightarrow \Omega$$

such that

$$L_t f(x) = \langle K f(x), F_t(x) \rangle_V, \quad f \in \mathcal{T}. \quad (2.29)$$

For a deterministic flow on  $\mathbb{R}^D$ ,

$$L_t f(x) = \langle a_t(x), \nabla f(x) \rangle.$$

Thus  $Kf = \nabla f$  and  $F_t = a_t$ . For an additive diffusion

$$dX_t = a_t(X_t) dt + \sqrt{2}b_t dB_t,$$

the generator is

$$L_t f(x) = \langle a_t(x), \nabla f(x) \rangle + b_t^2 \Delta f(x).$$

In this case, the unknown parameter may be the drift  $a_t$ , or the pair  $(a_t, b_t^2)$  if the diffusion coefficient is also learned.

For a finite-state Markov chain, the parameter field is the vector of transition rates:

$$Kf(x) = \left( f(y) - f(x) \right)_{y \neq x}, \quad F_t(x) = \left( q_t(x, y) \right)_{y \neq x}.$$

The natural constraint is then  $F_t(x) \in \mathbb{R}_+^{\mathcal{S} \setminus \{x\}}$ .

For a pure-jump process on a general state space,

$$L_t f(x) = \int_{\mathcal{S}} \left( f(y) - f(x) \right) Q_{t,x}(dy),$$

where  $Q_{t,x}$  is a nonnegative jump measure. Its total mass is the jump intensity, and its normalization, when the total mass is positive, is the jump distribution. In particular,  $Q_{t,x}$  needs not be a probability measure.

The conditional-to-marginal formula transfers to the parameter fields:

$$F_t(x) = \mathbb{E} \left[ F_t^{Z^\circ}(x) \mid X_t^\circ = x \right].$$

The learning problem is therefore a conditional regression problem. The target  $F_t$  is unknown, but the conditional target  $F_t^{Z^\circ}$  is explicit.

### 2.4.5 Bregman matching

Let  $F_t^\theta(x)$  be a neural-network approximation of the marginal parameter field. The ideal loss would be

$$\mathcal{L}_{\text{GM}}(\theta) = \mathbb{E} \left[ D_\Phi \left( F_T(X_T^\circ) \| F_T^\theta(X_T^\circ) \right) \right], \quad T \sim \text{Unif}(0, 1),$$

but  $F_t$  is not directly available. The conditional-to-marginal identity shows that a computable replacement is the conditional loss

$$\mathcal{L}_{\text{CGM}}(\theta) = \mathbb{E} \left[ D_\Phi \left( F_T^{Z^\circ}(X_T^\circ) \| F_T^\theta(X_T^\circ) \right) \right].$$

The loss class for which this replacement changes the loss only by an additive constant is the class of Bregman divergences.

**Definition 2.12** (Bregman divergence). Let  $\Phi : \Omega \rightarrow \mathbb{R}$  be differentiable and strictly convex on a convex set  $\Omega \subseteq V$ . The associated Bregman divergence is

$$D_\Phi(u||v) = \Phi(u) - \Phi(v) - \langle \nabla \Phi(v), u - v \rangle_V. \quad (2.30)$$

For  $\Phi(u) = \frac{1}{2} \|u\|_V^2$ , one obtains

$$D_\Phi(u||v) = \frac{1}{2} \|u - v\|_V^2.$$

On the probability simplex, taking  $\Phi(u) = \sum_i u_i \log u_i$  gives the Kullback–Leibler divergence. More generally, on the cone of nonnegative finite measures dominated by a reference measure  $\mu$ , writing  $\nu = r\mu$  and  $\eta = s\mu$ , the same entropy functional gives

$$D_{\text{KL}}(\nu||\eta) = \int \left[ r(y) \log \frac{r(y)}{s(y)} - r(y) + s(y) \right] d\mu(y). \quad (2.31)$$

This last example is relevant for jump models, since it compares nonnegative measures and therefore applies to  $Q_{t,x}(dy)$ , whose total mass is an intensity rather than one.

**Proposition 2.13** (Bregman divergence matching [H<sup>+</sup>24, AS26]). *Let  $D_\Phi$  be a Bregman divergence on  $\Omega$ , and let  $T \sim \text{Unif}(0, 1)$ . Assume that*

$$F_t(x) = \mathbb{E} \left[ F_t^{Z^\circ}(x) \mid X_t^\circ = x \right].$$

*Then*

$$\mathcal{L}_{\text{CGM}}(\theta) = \mathcal{L}_{\text{GM}}(\theta) + \mathbb{E} \left[ D_\Phi(F_T^{Z^\circ}(X_T^\circ) || F_T(X_T^\circ)) \right].$$

*In particular, the two losses have the same gradients with respect to  $\theta$ .*

*Proof.* A Bregman divergence is affine in its first argument, up to a term that does not depend on the second one. Indeed,

$$D_\Phi(u||v) = \Phi(u) + \langle u, -\nabla \Phi(v) \rangle_V - \Phi(v) + \langle \nabla \Phi(v), v \rangle_V.$$

In the only term where  $u$  interacts with  $v = F_T^\theta(X_T^\circ)$ , we may replace  $F_T(X_T^\circ)$  by  $F_T^{Z^\circ}(X_T^\circ)$  after conditioning on  $X_T^\circ$ . The remaining difference is

$$\mathbb{E} \left[ \Phi(F_T(X_T^\circ)) - \Phi(F_T^{Z^\circ}(X_T^\circ)) \right],$$

which does not depend on  $\theta$  and yields the proof.  $\square$

This proposition is an analogue of denoising score matching in Section 2.3. There, the unknown marginal score is the conditional expectation of an explicit Gaussian conditional score, and the squared loss gives a computable denoising objective. Here, the score is replaced by the parameter field of a Markov generator. Depending on the generator class, the supervised target may be a drift vector, a transition-rate vector, or a jump measure.

### 2.4.6 The design space of Markov generators

The generator viewpoint also describes which types of Markov dynamics can be used. On finite spaces, the parameters are rate matrices. On Euclidean spaces, a standard result, related to Courrège’s theorem, says that Markov generators satisfying the positive maximum principle decompose into drift, diffusion, and jump parts. In the following statement, we restrict the jump part to finite activity, which is enough for the present discussion.

**Theorem 2.14** (Basic forms of Markov generators [AS26]). *Under standard regularity assumptions, the following descriptions hold.*

*If  $\mathcal{S}$  is finite, every continuous-time Markov generator has the form*

$$L_t f(x) = \sum_{y \neq x} q_t(x, y) (f(y) - f(x)), \quad q_t(x, y) \geq 0.$$

*If  $\mathcal{S} = \mathbb{R}^D$ , every sufficiently regular finite-activity generator has the form*

$$L_t f(x) = \langle a_t(x), \nabla f(x) \rangle + \frac{1}{2} \langle \sigma_t(x) \sigma_t(x)^\top, \nabla^2 f(x) \rangle_{\text{HS}} + \int_{\mathbb{R}^D} (f(y) - f(x)) Q_{t,x}(\mathrm{d}y), \quad (2.32)$$

where  $a_t$  is a drift field,  $\sigma_t$  is a diffusion matrix, and  $Q_{t,x}$  is a finite nonnegative jump measure.

The score-based models of Section 2.3 lie in the drift-diffusion part of (2.32). The learned score enters the drift of the reverse SDE (2.8). Flow matching removes the diffusion term and learns a deterministic drift. Continuous-time discrete diffusion uses the finite-state part of the theorem. Jump models use the last term in (2.32), either on a discrete state space or on a continuous one.

A single probability path may also admit several generator realizations. For the Gaussian interpolation

$$X_t^\circ = (1 - t)\xi^\circ + tZ^\circ,$$

the conditional flow generator is

$$L_t^{Z^\circ} f(x) = \left\langle \frac{Z^\circ - x}{1 - t}, \nabla f(x) \right\rangle.$$

Hence the conditional-to-marginal formula gives the usual flow-matching drift

$$a_t(x) = \mathbb{E} \left[ \frac{Z^\circ - x}{1 - t} \mid X_t^\circ = x \right].$$

For the mixture path (2.21), the natural realization is instead the pure-jump generator (2.28). Thus the probability path and the generator class should be chosen together: some paths are represented by flows, others by jumps, and some admit several realizations.

### 2.4.7 Duality and stability

Once a generator  $\hat{L}_t$  has been learned, one must relate the generator error to the final generated law. This is the same issue as in Section 2.3: the empirical loss controls an error along the probability path, whereas the final performance is measured at time 1. The backward Kolmogorov equation gives the following dual identity.

**Proposition 2.15** (Duality identity for generator errors [AS26]). *Let  $(p_t)_{0 \leq t \leq 1}$  and  $(\hat{p}_t)_{0 \leq t \leq 1}$  solve the weak Kolmogorov forward equations associated with  $L_t$  and  $\hat{L}_t$ , with the same initial condition  $p_0 = \hat{p}_0$ . Let  $f \in \mathcal{T}$ , and let  $\hat{f}_t$  solve the backward Kolmogorov equation*

$$\partial_t \hat{f}_t + \hat{L}_t \hat{f}_t = 0, \quad \hat{f}_1 = f.$$

Then

$$\langle p_1 - \hat{p}_1, f \rangle = \int_0^1 \langle p_t, (L_t - \hat{L}_t) \hat{f}_t \rangle dt. \quad (2.33)$$

*Proof.* Differentiate  $t \mapsto \langle p_t, \hat{f}_t \rangle$ . Using the forward equation for  $p_t$  and the backward equation for  $\hat{f}_t$ ,

$$\frac{d}{dt} \langle p_t, \hat{f}_t \rangle = \langle p_t, L_t \hat{f}_t \rangle + \langle p_t, \partial_t \hat{f}_t \rangle = \langle p_t, (L_t - \hat{L}_t) \hat{f}_t \rangle.$$

Similarly,

$$\frac{d}{dt} \langle \hat{p}_t, \hat{f}_t \rangle = \langle \hat{p}_t, \hat{L}_t \hat{f}_t + \partial_t \hat{f}_t \rangle = 0.$$

Integrating from 0 to 1, using  $p_0 = \hat{p}_0$  and  $\hat{f}_1 = f$ , gives (2.33).  $\square$

The identity separates the statistical error  $L_t - \hat{L}_t$  from the propagation of this error by the learned dynamics. For flows and diffusions, this propagation can be bounded by a one-sided Lipschitz constant, as in the stability estimates used for SGMs in Section 2.3.

**Theorem 2.16** (Backward Lipschitz propagation [AS26]). *Assume that*

$$\hat{L}_t \phi(x) = \langle \hat{a}_t(x), \nabla \phi(x) \rangle + b_t^2 \Delta \phi(x),$$

corresponding to the SDE

$$dX_t = \hat{a}_t(X_t) dt + \sqrt{2} b_t dB_t,$$

where  $b_t$  depends only on time. Assume that  $\hat{a}_t$  is differentiable in space and that

$$\lambda_{\max}(\nabla \hat{a}_t(x)) \leq \hat{\ell}_t, \quad x \in \mathbb{R}^D,$$

for some integrable function  $\hat{\ell}$ . If  $\hat{f}_t$  solves

$$\partial_t \hat{f}_t + \hat{L}_t \hat{f}_t = 0, \quad \hat{f}_1 = f,$$

then

$$\|\hat{f}_t\|_{\text{Lip}} \leq \exp\left(\int_t^1 \hat{\ell}_u \, du\right) \|f\|_{\text{Lip}}.$$

Consequently, if  $L_t$  and  $\hat{L}_t$  have the same diffusion coefficient and drifts  $a_t$  and  $\hat{a}_t$ , then

$$W_1(p_1, \hat{p}_1) \leq \int_0^1 \exp\left(\int_t^1 \hat{\ell}_u \, du\right) \mathbb{E}_{X_t \sim p_t} [\|a_t(X_t) - \hat{a}_t(X_t)\|] \, dt. \quad (2.34)$$

*Proof.* Rather than using Alekseev–Gröbner perturbation formulas, the proof uses the Markov representation of the backward equation and a synchronous coupling.

Fix  $t \in [0, 1]$  and  $x, y \in \mathbb{R}^D$ . Let  $X_s^{t,x}$  and  $X_s^{t,y}$  solve, for  $s \in [t, 1]$ ,

$$X_s^{t,x} = x + \int_t^s \hat{a}_u(X_u^{t,x}) \, du + \int_t^s \sqrt{2} b_u \, dB_u,$$

and

$$X_s^{t,y} = y + \int_t^s \hat{a}_u(X_u^{t,y}) \, du + \int_t^s \sqrt{2} b_u \, dB_u,$$

with the same Brownian motion. By the Feynman–Kac representation,

$$\hat{f}_t(x) = \mathbb{E}[f(X_1^{t,x})], \quad \hat{f}_t(y) = \mathbb{E}[f(X_1^{t,y})].$$

Set  $Z_s = X_s^{t,x} - X_s^{t,y}$ . Since the diffusion coefficient is additive and the Brownian motion is shared, the noise cancels:

$$Z_s = x - y + \int_t^s (\hat{a}_u(X_u^{t,x}) - \hat{a}_u(X_u^{t,y})) \, du.$$

Therefore

$$\frac{d}{ds} \|Z_s\|^2 = 2 \langle Z_s, \hat{a}_s(X_s^{t,x}) - \hat{a}_s(X_s^{t,y}) \rangle.$$

The one-sided Lipschitz assumption gives

$$\langle Z_s, \hat{a}_s(X_s^{t,x}) - \hat{a}_s(X_s^{t,y}) \rangle \leq \hat{\ell}_s \|Z_s\|^2.$$

By Gronwall’s lemma,

$$\|X_1^{t,x} - X_1^{t,y}\| \leq \exp\left(\int_t^1 \hat{\ell}_u \, du\right) \|x - y\|.$$

Hence

$$|\hat{f}_t(x) - \hat{f}_t(y)| \leq \|f\|_{\text{Lip}} \exp\left(\int_t^1 \hat{\ell}_u \, du\right) \|x - y\|,$$

which proves the Lipschitz estimate. The Wasserstein bound follows by applying Proposition 2.15 to 1-Lipschitz test functions and using the Kantorovich–Rubinstein duality.  $\square$

This theorem is the stability estimate used in the Wasserstein control of score-based models in Section 2.3. In that case, the generator error is the drift error induced by replacing the true score by  $\hat{s}$ , and the preceding bound has the same form as the SDE stability estimate used in the minimax analysis.

In summary, the construction is close to that of SGMs. A probability path specifies the desired marginal evolution. A Markov generator specifies how this evolution is realized infinitesimally. Conditional Generator Matching replaces the unknown marginal generator by an explicit conditional regression target. Backward Kolmogorov duality then propagates the learned generator error to a final distributional bound. The results in this section identify population generators, computable conditional objectives, and stability identities. They do not yet provide a finite-sample generalization theorem for an empirically learned generator; obtaining such a theorem is an open part of the program.

## 2.5 Adaptive diffusion guidance

While learning to sample from the unconditional data distribution  $p^*$  is useful, many applications require user control over the generative process. *Conditional generation* aims to sample exclusively from the conditional distribution  $p^*(\cdot | c)$ , where  $c$  represents semantic conditioning information, ranging from discrete class labels to physical constraints or complex textual prompts.

### 2.5.1 Classifier-free guidance and the tempering myth

A direct approach to conditional sampling in score-based diffusion models would require defining a conditional forward process and retraining a specific conditional score network  $\nabla \log p_t(x | c)$  entirely from scratch for each condition. However, Bayes' rule yields a precise algebraic decomposition of the conditional score:

$$\nabla \log p_t(x | c) = \nabla \log p_t(x) + \nabla \log p_t(c | x). \quad (2.35)$$

This formulation, known as *Classifier Guidance* [DN21], allows practitioners to reuse an unconditionally trained diffusion model and simply add the gradient of an auxiliary, independently trained noisy classifier  $p_t(c | x)$  during the reverse SDE simulation. While theoretically sound, training noise-conditioned classifiers that remain robust across all noise levels  $t$  is difficult and computationally expensive.

To bypass the need for an explicit classifier, while improving sample quality and adherence to the condition  $c$ , practitioners rely on *Classifier-Free Guidance* (CFG) [HS22]. CFG co-trains the conditional score  $\nabla \log p_t(x | c)$  and the unconditional score  $\nabla \log p_t(x)$  within a single neural network by randomly dropping the conditioning information during training (using a dummy token  $\emptyset$ ). The classifier gradient is then inferred algebraically.

Furthermore, an artificial *guidance weight*  $w > 0$  (often denoted  $\gamma = 1 + w$  in the

literature) is introduced to extrapolate the score. The guided score is thus defined as:

$$\begin{aligned} s_t^\gamma(x \mid c) &:= \nabla \log p_t(x \mid c) + w (\nabla \log p_t(x \mid c) - \nabla \log p_t(x)) \\ &= (1 - \gamma) \nabla \log p_t(x) + \gamma \nabla \log p_t(x \mid c). \end{aligned} \quad (2.36)$$

Despite its ubiquity and use in models such as Stable Diffusion [RBL<sup>+</sup>22], the exact mathematical mechanism of guidance is not fully explained theoretically. A common but generally inaccurate intuition suggests that integrating the SDE driven by (2.36) effectively samples from a “tempered” conditional distribution proportional to the geometric average:

$$\tilde{p}_{\gamma,t}(x) \propto p_t(x) p_t(c \mid x)^\gamma = p_t(x)^{1-\gamma} p_t(x \mid c)^\gamma.$$

As recent geometric and probabilistic literature shows [C<sup>+</sup>24, B<sup>+</sup>24], this is not correct in general: the resulting non-linear SDE marginals do not correspond to the tempered target. The obstruction is that the geometric average of the marginals is not the marginal of the geometric average. In fact, CFG behaves more like a so-called *predictor-corrector* mechanism [B<sup>+</sup>24]. Attempts to characterize it via Feynman-Kac correctors reveal that exact tempered sampling would actually require simulating a complex sequential Monte Carlo scheme along the trajectories [S<sup>+</sup>25].

Another widespread concern in the literature is that imposing a large, constant weight  $\gamma \in [4, 10]$  might push the Langevin dynamics outside the true data manifold  $\text{supp}(p^\star(\cdot \mid c))$ , resulting in distorted, out-of-distribution samples and generating visual artifacts.

## 2.5.2 Rigorous probabilistic guarantees: the guiding field

In our work [APL<sup>+</sup>25], we provide a theoretical formulation of the guidance mechanism that addresses these points. For simplicity, we restrict the analysis to the DDPM-like backward diffusion where the backward diffusion coefficient is  $b_t = 1$  (see Section 2.3).

Let  $(Y_t^w)_{t \in [0, T]}$  denote the reverse generative process driven by the guided score  $s_{T-t}^{1+w}$ . To isolate the effect of guidance, we define the backward-running *guiding field*  $G_t(x)$  as the log-likelihood ratio:

$$G_t(x) := \log p_{T-t}(c \mid x) - \log p(c) = \log p_{T-t}(x \mid c) - \log p_{T-t}(x).$$

Under this notation, the guided drift perturbation corresponds precisely to  $2w_t(x) \nabla G_t(x)$ , where we explicitly allow the guidance weight  $w_t(x)$  to depend dynamically on both time and space. The backward guided dynamics are governed by the SDE:

$$dY_t^w = [Y_t^w + 2 \nabla \log p_{T-t}(Y_t^w \mid c) + 2w_t(Y_t^w) \nabla G_t(Y_t^w)] dt + \sqrt{2} d\bar{B}_t. \quad (2.37)$$

First, we resolve the ambiguity regarding support preservation. A natural question is whether the arbitrary drift perturbation introduced by  $w_t(x) \nabla G_t(x)$  can destroy the support recovery property of the generative SDE. To ensure absolute continuity of the path measures, we bound the norm of the guiding field  $\|\nabla G_t(x)\|$  using Tweedie’s formula [Rob56]. The marginal score can be expressed exactly as the conditional expectation

of the clean data given the noisy observation, implying that the gradient of the guiding field is purely a function of the difference between the conditional and unconditional posterior expectations of the data.

**Theorem 2.17** (Population support recovery [APL<sup>+</sup>25]). *Assume the unconditional data distribution has compact support  $\text{supp}(p^*) \subseteq B(0, R)$  for some  $R < \infty$ , and the dynamic guidance weight is uniformly bounded,  $|w_t(x)| \leq C_{\max}$  and that  $p(c) > 0$ . Then, by utilizing Girsanov's theorem and the Donsker-Varadhan variational formula, the Kullback-Leibler divergence between the guided process  $Y^w$  and the exact conditional process  $Y^0$  satisfies for any  $\varepsilon > 0$ :*

$$\text{KL}(Y_{[0, T-\varepsilon]}^w \parallel Y_{[0, T-\varepsilon]}^0) \leq \mathbb{E} \left[ \int_0^{T-\varepsilon} w_t^2 \|\nabla G_t(Y_t^w)\|^2 dt \right] \leq -C_{\max} \log p(c) < \infty.$$

Consequently, almost surely, the terminal state of the guided process remains within the true conditional support:

$$\mathbb{P}(Y_T^w \in \text{supp}(p^*(\cdot | c))) = 1.$$

This theorem confirms that, contrary to certain empirical beliefs, guidance with  $w \neq 0$  does not by itself destroy support recovery at the population level, provided the guidance weight remains bounded.

Furthermore, we can characterize how the CFG mechanism optimizes the alignment with the condition  $c$ . By applying Itô's lemma and the Fokker-Planck equation to the backward running guiding field  $G_t(Y_t^w)$ , we establish a differential identity:

$$dG_t(Y_t^w) = (1 + 2w_t) \|\nabla G_t(Y_t^w)\|^2 dt + \sqrt{2} \langle \nabla G_t(Y_t^w), d\bar{B}_t \rangle.$$

Taking the expectation over the trajectories directly establishes that guidance monotonically increases the expected log-likelihood of the condition.

**Theorem 2.18** (Likelihood enhancement and martingale property [APL<sup>+</sup>25]). *The expected log-likelihood of the conditioning class increases according to of the applied guidance strength along the paths. Specifically:*

$$\mathbb{E} [\log p_0(c | Y_T^w)] = C + \mathbb{E} \left[ \int_0^T (1 + 2w_t(Y_t^w)) \|\nabla G_t(Y_t^w)\|^2 dt \right], \quad (2.38)$$

where the constant  $C = \log p(c) + \mathbb{E}[\log p_T(Y_0^w | c) - \log p_T(Y_0^w)]$  is independent of the choice of  $w$ . Furthermore, evaluating the quadratic variation of  $G_t$  reveals that the Doléans-Dade stochastic exponential associated with the guiding field,

$$S_t := \frac{p(c)}{p_{T-t}(c | Y_t^w)} \exp \left( \int_0^t 2w_s(Y_s^w) \|\nabla G_s(Y_s^w)\|^2 ds \right),$$

is a super-martingale on  $[0, T]$ . By Doob's maximal inequality, this gives a high-probability control on the increase of the adherence to the class label.

### 2.5.3 Adaptive guidance via stochastic optimal control

These theoretical insights reveal a limitation of current practices: a large constant guidance weight  $w \in [3, 10]$  is applied across all time steps  $t \in [0, T]$  and spatial regions. As (2.38) demonstrates, maximizing  $\mathbb{E}[\log p_0(c \mid Y_T^w)]$  without regularization, with a large constant  $w$  may drive trajectories into low-probability regions of the unconditional distribution  $p_t$ . In these regions, the underlying score network  $\nabla \log p_t(x)$  is used outside its well-estimated regime, thereby reducing sample diversity and causing the well-documented visual artifacts.

To resolve this, we recast the guidance scheduling problem into the framework of *Stochastic Optimal Control* (SOC) [FS06]. Rather than imposing a fixed heuristic schedule, we formulate the objective as adaptively maximizing the terminal classifier alignment while concurrently penalizing the trajectory’s deviation from the true unguided conditional process. We seek the optimal dynamic policy  $w^*$  that maximizes the reward functional:

$$\mathcal{R}(w) := \mathbb{E}[\log p_0(c \mid Y_T^w)] - \alpha \text{KL}(Y_{[0,T]}^w \parallel Y_{[0,T]}^0), \quad (2.39)$$

where  $\alpha > 0$  sets the statistical trade-off between conditional alignment (alignment with condition  $c$ ) and generative diversity (proximity to the true conditional measure).

By expanding the KL divergence via Girsanov’s theorem and substituting our likelihood enhancement representation from Theorem 2.18, the complex objective reduces to a tractable path integral over the guided trajectories:

$$\mathcal{R}(w) = \mathbb{E} \left[ \int_0^T \left( 1 + 2w_t(Y_t^w) - \alpha w_t(Y_t^w)^2 \right) \|\nabla G_t(Y_t^w)\|^2 dt \right] + C', \quad (2.40)$$

where  $C'$  is a constant independent of the guidance  $w$ .

This continuous-time control problem is governed by the Hamilton-Jacobi-Bellman (HJB) equation. Defining the optimal value function  $V_t(x)$ , the optimal adaptive guidance strategy can be derived formally by maximizing the Hamiltonian, yielding:

$$w_t^*(x) = \frac{\langle \nabla G_t(x), \nabla V_t(x) \rangle + \|\nabla G_t(x)\|^2}{\alpha \|\nabla G_t(x)\|^2}. \quad (2.41)$$

This equation reveals that the optimal guidance is explicitly state- and time-dependent.

### 2.5.4 Scalable optimization via the adjoint sensitivity method

To implement this in high dimensions where solving the HJB PDE directly via grid-based methods is computationally prohibitive, we parameterize the scalar guidance strength using a shallow neural network  $w_\theta(t, x, c)$ . We optimize the parameters  $\theta$  using an *adjoint sensitivity method* [DE+25].

Because changing  $w_\theta$  alters both the instantaneous reward in (2.40) and the distribution of the trajectories  $Y_t^w$  themselves, standard differentiation requires tracking the sensitivity process  $Z_t(\tau) := \frac{\partial Y_t^w}{\partial w_\tau}$ , which propagates a perturbation in the guidance weight at time  $\tau$  to the state at time  $t > \tau$ .

To avoid simulating  $Z_t$  forward for every possible  $\tau$ , we define the backward-in-time *adjoint process*  $\lambda_t$ , governed by the linear SDE:

$$\begin{cases} d\lambda_t = - [\text{Id} + 2\nabla^2 \log p_{T-t}(Y_t^w | c) + 2w_t \nabla^2 G_t(Y_t^w)]^\top \lambda_t dt \\ \quad - 2(1 + 2w_t - \alpha w_t^2) \nabla G_t(Y_t^w)^\top \nabla^2 G_t(Y_t^w) dt, \\ \lambda_T = 0. \end{cases}$$

By integrating the inner product  $\langle \lambda_t, Z_t \rangle$  by parts along the trajectories, the dependence on the forward sensitivity process drops out. The total gradient of the reward functional with respect to the network parameters is then computed efficiently as an expectation over simulated guided trajectories:

$$\nabla_{\theta} \mathcal{R}(w_{\theta}) = \mathbb{E} \left[ \int_0^T \left( 2(1 - \alpha w_t) \|\nabla G_t\|^2 + 2 \langle \lambda_t, \nabla G_t \rangle \right) \nabla_{\theta} w_t(Y_t^w, c) dt \right].$$

This formulation allows us to compute the parameter gradients  $\nabla_{\theta} \mathcal{R}$  efficiently by running a single forward pass to simulate the trajectories, followed by a single backward pass to compute the adjoint process  $\lambda_t$ . This SOC framework [APL<sup>+</sup>25] thus provides an adaptive mechanism for conditional generation. By learning to modulate the scalar guidance strength based on the local geometry of the score field and the temporal dynamics of diffusion, it provides an adaptive alternative to fixed heuristic constants: a dynamically responsive guidance mechanism that balances condition fulfillment with the geometric fidelity of the generated samples.

## 2.6 Perspectives

The analysis of generative modeling presented in this chapter—from distribution estimation to dynamical generators and optimal control—describes some statistical limits of learning complex distributions with continuous-time techniques. We outline several open mathematical directions extending this framework, bridging generative models with geometric inference.

**Submanifold minimax rates for diffusions:** We established minimax optimal rates for SGMs in the ambient space. However, if the target distribution is supported on a lower-dimensional submanifold  $M \subset \mathbb{R}^D$ , the smoothed density  $p_t$  develops singularities (e.g., the score field may vanish symmetrically at the origin for a sphere). Therefore, the score cannot exhibit global Hölder continuity in the ambient space.

**Conjecture 2.1** (SGM minimax rates on submanifolds). Let  $p^*$  be a  $\mathcal{C}^{\beta}$  density strictly bounded away from zero relative to the intrinsic volume measure of a  $d$ -dimensional  $\mathcal{C}^{\beta+1}$  compact submanifold  $M \subset \mathbb{R}^D$  with positive reach. We conjecture that the appropriate regularity notion is that the exact score satisfies  $s^*(t, \cdot) \in W_{C^{t-(1+\gamma)/2}}^{\beta+\gamma, 2}(\mathbb{R}^D, p_t)$ , where  $W$  is the  $L^2(p_t)$ -weighted Sobolev space. Consequently, score estimators trained via denoising

score matching on submanifolds achieve the intrinsic minimax optimal rate for distribution estimation in Wasserstein-1 distance:

$$\inf_{\hat{P}_{\text{SGM}}} \sup_{p^*} \mathbb{E} [W_1(P^*, \hat{P}_{\text{SGM}})] \asymp \left( \frac{\log n}{n} \right)^{\frac{\beta+1}{2\beta+d}}.$$

**Computational gains of jump generators:** The Generator Matching framework unifies continuous ODE/SDE dynamics with pure jump processes via mixture paths. An open question is whether formulating generative dynamics as a jump process provides *algorithmic or numerical* advantages, potentially bypassing the stiffness and truncation errors associated with integrating highly-curved vector fields across low-density regions.

**Geometric guarantees for sampled distributions:** Our current statistical analyses evaluate generative models using integral probability metrics such as  $W_1$ , which measure an average-case transport cost. A small Wasserstein error does not guarantee that the effective support of the generated distribution  $\hat{P}_{\text{SGM}}$  inherits the geometric and topological properties (e.g., homology, reach) of the target support  $M$ . Ensuring that synthetic samples satisfy underlying physical or geometric constraints requires shifting the analysis from average mass transport to Hausdorff and geometric fidelity.

**Influence of boundaries in Wasserstein estimation:** In Chapter 1, we demonstrated that the presence of a boundary  $\partial M \neq \emptyset$  structurally breaks the local Euclidean symmetry, inherently slowing down the optimal estimation rates of the support in Hausdorff distance. A natural inquiry is whether this geometric boundary effect translates to distribution estimation via generative models.

**Conjecture 2.2** (Boundary independence in  $W_1$ ). Because the  $W_1$  distance metrizes weak convergence and measures an average transport cost, it is less sensitive to highly localized geometric singularities. We conjecture that for distributions supported on submanifolds with a sufficiently regular boundary, the minimax rate for distribution estimation under the  $W_1$  metric remains the same as in the boundaryless case. The presence of the boundary degrades  $L^\infty$  geometric recovery but should not incur a statistical penalty for average optimal transport costs.

**Generative drifting and score matching:** A recent development in the generative modeling literature relies on *drifting* algorithms, which dynamically alternate between adding noise and denoising during sampling. As shown by [TO26], generative drifting can be interpreted as a score matching mechanism. Formalizing the non-asymptotic statistical complexities of drifting algorithms—specifically how alternating steps mitigate accumulated drift errors without introducing singularities near  $t = 0$ —is a promising open problem.

**Information geometry and guidance:** Guidance operates by extrapolating scores via a non-convex affine combination:  $s_t^\gamma = (1 - \gamma)\nabla \log p_t(x) + \gamma\nabla \log p_t(x \mid c)$ . In Information Geometry, affine combinations of logarithmic derivatives correspond to Amari’s  $\alpha$ -connections, which interpolate between exponential and mixture families. This raises a natural geometric question.

**Conjecture 2.3** (Guidance and information geometry). Guided trajectories can be related to geodesics under the Fisher information metric in the statistical manifold of the latent space, with the guidance weight dictating the projection onto the conditioning submanifold.



# Bibliography

- [AACB24] Eddie Aamari, Ery Arias-Castro, and Clément Berenfeld. The coreness and H-index of random geometric graphs. *ALEA, Latin American Journal of Probability and Mathematical Statistics*, 21:963–983, 2024.
- [AAL23] Eddie Aamari, Catherine Aaron, and Clément Levrard. Minimax boundary estimation and estimation with boundary. *Bernoulli*, 29(4):3334–3368, 2023.
- [Aam26] Eddie Aamari. Second order of the vote of thanks to Whiteley et al. and contribution to the discussion of “Statistical exploration of the manifold hypothesis”. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(2):388–389, 2026.
- [AB06] Stephanie B Alexander and Richard L Bishop. Cut loci, minimizers, and absolute retracts. *Journal of Differential Geometry*, 74(3):331–352, 2006.
- [AB17] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.
- [AB24] Eddie Aamari and Clément Berenfeld. A theory of stratification learning, 2024.
- [ABL23] Eddie Aamari, Clément Berenfeld, and Clément Levrard. Optimal reach estimation and metric learning. *The Annals of Statistics*, 51(3):1086–1108, 2023.
- [ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [ACC23] Ery Arias-Castro and P. A. Chau. Minimax estimation of distances on a surface and minimax manifold learning in the isometric-to-convex setting. *Information and Inference: A Journal of the IMA*, 2023.
- [AK04] Nina Amenta and Yong Joo Kil. Defining point-set surfaces. *ACM Transactions on Graphics*, 23(3):264–270, 2004.

- [AK21] Eddie Aamari and Alexander Knop. Statistical query complexity of manifold estimation. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, pages 116–122, 2021.
- [AK22] Eddie Aamari and Alexander Knop. Adversarial manifold estimation. *Foundations of Computational Mathematics*, 24(1):1–97, 2022.
- [AKC<sup>+</sup>19] Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019.
- [AL18] Eddie Aamari and Clément Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry*, 59(4):923–971, 2018.
- [AL19] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.
- [ALS13] Dominique Attali, André Lieutier, and David Salinas. Vietoris–Rips complexes also provide topologically correct reconstructions of sampled shapes. *Computational Geometry*, 46(4):448–465, 2013.
- [APL<sup>+</sup>25] Iskander Azangulov, Peter Potaptchik, Qinyu Li, Eddie Aamari, George Deligiannidis, and Judith Rousseau. Adaptive diffusion guidance via stochastic optimal control, 2025.
- [AS25] Yariv Aizenbud and Barak Sober. Non-parametric estimation of manifolds from noisy data. *Journal of Machine Learning Research*, 26(183):1–60, 2025. Initially appeared on arXiv in 2021.
- [AS26] Eddie Aamari and Arthur Stéphanovitch. Statistical analysis of markovian generative modeling, 2026.
- [AVE23] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.
- [B<sup>+</sup>24] A. Bradley et al. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2401.00000*, 2024.
- [BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- [BDK<sup>+</sup>17] Benedikt Bauer, Luc Devroye, Michael Kohler, Adam Krzyżak, and Harro Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017.

- [Bel61] Richard Bellman. *Adaptive control processes: a guided tour*. Princeton university press, 1961.
- [BG14] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential Delaunay complexes. *Discrete & Computational Geometry*, 51(1):221–267, 2014.
- [BKS23] Denis Belomestny, Klaus Krumbiegel, and John Schoenmakrs. Rates of convergence for generative adversarial networks. *Journal of Machine Learning Research*, 24(1):1–31, 2023.
- [BLW19] Jean-Daniel Boissonnat, André Lieutier, and Mathijs Wintraecken. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *Journal of Applied and Computational Topology*, 3(1):29–58, 2019.
- [BMNN24] M. Bauer, F. Mémoli, T. Needham, and M. Nishino. The z-gromov-wasserstein distance. *arXiv preprint arXiv:2403.02329*, 2024.
- [BO95] Christoph Bregler and Stephen M Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proceedings of IEEE International Conference on Computer Vision*, pages 494–499. IEEE, 1995.
- [BR16] Nikolay Baldin and Markus Reiß. Unbiased estimation of the volume of a convex body. *Stochastic Processes and their Applications*, 126(12), 2016.
- [BRS<sup>+</sup>12] Sivaraman Balakrishnan, Alesandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. In *Artificial Intelligence and Statistics*, pages 64–72. PMLR, 2012.
- [Bru14] Victor Emmanuel Brunel. *Non parametric estimation of convex bodies and convex polytopes*. PhD thesis, Université Pierre et Marie Curie-Paris VI; University of Haifa, 2014.
- [C<sup>+</sup>24] Muthu Chidambaram et al. Does classifier-free guidance really perform tempered sampling? *arXiv preprint arXiv:2405.00000*, 2024.
- [CBDB<sup>+</sup>22] Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, Lorenzo Siciliano, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.
- [CJLZ20] Minshuo Chen, Wenru Jiang, Wenjing Liao, and Tuo Zhao. Distribution estimation with Wasserstein generative adversarial networks. *arXiv preprint arXiv:2009.04382*, 2020.
- [CK22] Minwoo Chae and Dong-kyoung Kim. Rates of convergence for Wasserstein GANs on linear subspaces and submanifolds. *The Annals of Statistics*, 50(1):389–413, 2022.

- [Dau92] Ingrid Daubechies. *Ten lectures on wavelets*, volume 61. SIAM, 1992.
- [DE<sup>+</sup>25] Carles Domingo-Enrich et al. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. *arXiv preprint arXiv:2501.00000*, 2025.
- [Div22] Vincent Divol. Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, 183(1):581–647, 2022.
- [DKB14] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- [DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [DW96] Lutz Dümbgen and Günther Walther. Rates of convergence for random approximations of convex sets. *Advances in Applied Probability*, 28(2):384–393, 1996.
- [Fed59] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.
- [FILN18] Charles Fefferman, Sergei Ivanov, Matti Lassas, and Hariharan Narayanan. Fitting a putative manifold to noisy data. *Conference On Learning Theory*, pages 688–720, 2018.
- [FS06] Wendell H Fleming and Halil Mete Soner. *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media, 2006.
- [GBS05] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.
- [GHP<sup>+</sup>22] Richard J Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A Baas, Benjamin A Dunn, May-Britt Moser, and Edvard I Moser. Toroidal topology of population activity in grid cells. *Nature*, pages 1–6, 2022.
- [GPAM<sup>+</sup>14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [GPPVW12] Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13(1):1263–1291, 2012.

- [H<sup>+</sup>24] Peter Holderrieth et al. Generator matching for generalized continuous-time generative models. In *International Conference on Learning Representations*, 2024.
- [HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [HS22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.
- [KK13] Michael Kohler and Adam Krzyżak. Optimal global rates of convergence for interpolation problems with random design. *Statistics & Probability Letters*, 83(8):1871–1879, 2013.
- [KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [KZ15] Arlene KH Kim and Harrison H Zhou. Tight minimax rates for manifold estimation under hausdorff loss. *Electronic Journal of Statistics*, 9(1):1562–1582, 2015.
- [LCBH<sup>+</sup>22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [LCC<sup>+</sup>17] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [LE06] Chan-Su Lee and Ahmed Elgammal. Human motion synthesis by motion manifold learning and motion primitive segmentation. In *International Conference on Articulated Motion and Deformable Objects*, pages 464–473. Springer, 2006.
- [Lev98] David Levin. The approximation power of moving least-squares. *Mathematics of Computation*, 67(224):1517–1531, 1998.
- [LLN<sup>+</sup>08] Oscar Lao, Timothy T Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A Bindoff, David Comas, et al. Correlation between genetic and geographic structure in europe. *Current Biology*, 18(16):1241–1248, 2008.

- [LME24] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete flow matching. *arXiv preprint arXiv:2407.13822*, 2024.
- [MLS<sup>+</sup>18] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- [MSIB<sup>+</sup>18] Kevin R Moon, Jay S Stanley III, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.
- [Mül97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.
- [NJB<sup>+</sup>08] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.
- [NSW08] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.
- [Par62] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.
- [PH04] Neal Patwari and Alfred O Hero. Manifold learning algorithms for localization in wireless sensor networks. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–857. IEEE, 2004.
- [PS22] Nikita Puchkin and Vladimir Spokoiny. Structure-adaptive manifold estimation. *Journal of Machine Learning Research*, 23(40):1–62, 2022.
- [PSW<sup>+</sup>18] Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391):eaq1723, 2018.
- [RBL<sup>+</sup>22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

- [RC07] Alberto Rodríguez-Casal. Set estimation under convexity type assumptions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 43(6):763–774, 2007.
- [RM15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
- [RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.
- [Rob56] Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163. University of California Press, 1956.
- [Ros56] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pages 832–837, 1956.
- [RS12] Michael Reutlinger and Gisbert Schneider. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *Journal of Molecular Graphics and Modelling*, 34:108–117, 2012.
- [S<sup>+</sup>25] M. Skreta et al. Feynman-kac correctors in diffusion models. *arXiv preprint arXiv:2501.00000*, 2025.
- [SAL24] Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Wasserstein GANs are minimax optimal distribution estimators. *The Annals of Statistics*, 52(5):2167–2193, 2024.
- [SAL25] Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Generalization bounds for score-based generative models: a synthetic proof, 2025.
- [SBD21] Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. *Algorithmic Learning Theory*, pages 1051–1071, 2021.
- [SDWMG15] 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.
- [SEKS21] Jan Stanczuk, Christian Etmann, Lisa Kamiya, and Carola-Bibiane Schönlieb. Wasserstein GANs work because they fail (to approximate the Wasserstein distance). In *International Conference on Machine Learning*, 2021.

- [SSDK<sup>+</sup>21] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Arvind Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.
- [Sté25] Arthur Stéphanovitch. Generative model for optimal density estimation on unknown manifold. *arXiv preprint arXiv:2506.19587*, 2025.
- [Sto85] Charles J Stone. Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705, 1985.
- [TDSL00] Joshua B Tenenbaum, Vin de De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [Thä08] Christoph Thäle. 50 years sets with positive reach—a survey. *Surveys in Mathematics and its Applications*, 3:123–165, 2008.
- [TO26] Erkan Turan and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026.
- [Tsy09] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2009.
- [TZ22] Yuting Tang and Yun Zhao. Minimax rates for distribution estimation on manifolds under the Wasserstein metric. *arXiv preprint arXiv:2207.03960*, 2022.
- [Vil09] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2009.
- [Vin11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.
- [WGRD24] Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy. Statistical exploration of the manifold hypothesis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):353–379, 2024.





## RÉSUMÉ

---

Ce mémoire d'habilitation étudie comment la structure géométrique, la régularité et la stabilité régissent la complexité statistique de l'inférence en grande dimension. La première partie est consacrée à l'inférence géométrique et établit des vitesses minimax optimales pour l'estimation de supports, de quantités différentielles et d'invariants globaux à partir de nuages de points. La seconde partie étudie la modélisation générative comme une forme implicite d'inférence géométrique, en établissant des garanties statistiques pour les GANs de Wasserstein, les modèles de diffusion basés sur le score, l'appariement de générateurs de Markov et le guidage adaptatif des diffusions.

## MOTS CLÉS

---

dimension intrinsèque, régularité géométrique, statistique non paramétrique, inférence géométrique, modélisation générative, vitesses minimax, hypothèse de variété, modèles de diffusion

## ABSTRACT

---

This habilitation thesis studies how geometric structure, smoothness, and stability govern statistical complexity in high-dimensional inference. The first part is devoted to geometric inference and establishes minimax optimal rates for estimating supports, differential quantities, and global invariants from point clouds. The second part studies generative modeling as an implicit form of geometric inference, establishing statistical guarantees for Wasserstein GANs, score-based diffusion models, Markov generator matching, and adaptive diffusion guidance.

## KEYWORDS

---

intrinsic dimension, geometric regularity, nonparametric statistics, geometric inference, generative modeling, minimax rates, manifold hypothesis, diffusion models

