Latent Representations of Financial Data for Market Regime Analysis - Intermediate Report

Anita Daliga, 6073379

1 Introduction

Financial markets exhibit complex behavior driven by economic events and periods of market stress. These changes are reflected in financial quantities such as asset returns and implied volatility surfaces. Their statistical properties often differ substantially across market conditions. Accurately modeling these dynamics is essential for derivative pricing, risk management, and scenario analysis. However, many classical financial models rely on assumptions of stationary dynamics and therefore struggle to capture structural changes observed in real markets.

To address this limitation, regime-switching models have become an important tool in quantitative finance. By allowing model parameters to change according to latent market states, these models provide an interpretable framework for representing different economic environments. At the same time, recent advances in machine learning have introduced deep generative models capable of learning complex, high-dimensional probability distributions directly from data. Among these approaches, Optimal Transport (OT) has emerged as a powerful mathematical framework for constructing stable and geometrically meaningful generative models.

Despite the success of both research directions, relatively little work has explored their combination. Regime-switching models offer economic interpretability, but rely on relatively restrictive parametric assumptions. On the other hand, OT-based generative models provide substantial flexibility without explicitly representing market regimes. Bridging these perspectives may improve both the interpretability and the generative capabilities of modern financial models.

This report first reviews the literature on regime-switching models and Optimal Transport, highlighting the motivation for combining these approaches. Then it presents the theoretical foundations of the AE-OT method, including the necessary concepts from optimal transport and convex geometry. Finally, preliminary experiments reproducing the original AE-OT results on benchmark datasets and synthetic examples are presented, providing the basis for future application of the method to implied volatility surface data.

2 Literature review

Financial markets are characterized by periods of different behavior, such as calm and stressed market conditions, which often manifest through immediate changes in returns, volatility, and option prices. Classical single-regime models assume constant parameters and therefore struggle to capture these structural shifts. As a result, much scientific work has been done around regime-switching models. Such models allow market dynamics to evolve according to latent states representing different economic environments.
More recently, advances in machine learning have introduced Optimal Transport (OT) as a powerful framework for learning and generating complex probability distributions. Together, these two research areas provide complementary perspectives on financial data modeling. Regime-switching models focus on economically meaningful state dynamics, while OT-based methods offer flexible tools for learning distributional transformations in high-dimensional settings.

2.1 Regime-Switching Models in Finance

Regime-switching models provide a flexible framework for capturing structural breaks, nonlinear dynamics, and state-dependent behavior in financial time series. The work of Hamilton [1] formalizes Markov regime-switching models, where model parameters evolve according to an unobserved finite-state Markov chain. Rather than treating structural changes as deterministic events, this framework models transitions probabilistically and allows the data to reveal the likelihood of different market regimes.

Authors of [2] developed a regime-switching model for European option pricing in which both the drift and volatility of the underlying asset depend on the current market state. They derived a risk-neutral measure and introduced a computationally efficient successive approximation method. By that, they show that even relatively simple multi-state models can reproduce important empirical features of option markets, including volatility smiles and volatility term structures. These results highlight the ability of regime-switching dynamics to capture behaviors that are difficult to explain within classical Black–Scholes-type frameworks.

Empirical evidence supporting regime-switching behavior is provided by [3]. In this paper authors investigate foreign exchange markets using Markov-switching models with independent shifts in mean and variance. Their findings indicate that multi-regime approach outperforms both single-regime and GARCH models in explaining exchange-rate dynamics and forecasting volatility. Moreover, the authors demonstrate that option prices do not fully incorporate regime-switching information, as trading strategies based on regime-dependent valuations outperform conventional Black–Scholes benchmarks. This suggests that latent market regimes contain economically relevant information not entirely reflected in observed market prices.

From an econometric perspective, in [4], authors address the important question of determining whether regime-switching behavior is statistically present. Because standard likelihood-ratio tests become invalid when additional regimes are introduced, the authors develop a quasi-likelihood ratio test that overcomes identification and boundary problems present in Markov-switching models. Their contribution provides a rigorous statistical foundation for detecting regime changes in empirical applications.

The broader landscape of regime-switching approaches is reviewed in [5]. In this article, authors compare four regime switching models defined by different mechanism. Namely, they compare threshold models, hidden Markov regime-switching models, hidden semi-Markov regime-switching models, and smooth transition models. The review highlights differences in switching mechanisms, distribution modeling, and estimation procedures. This comparison showed that depending on the target, one approach outperforms the other ones. For example, if the goal is to capture business cycle dynamics, choosing a hidden Markov or semi-Markov regime switching model would give the best results. However, if the main objective is to achieve a smooth regime switching process, smooth transitions models would be the best choice.

Collectively, these contributions establish the broad relevance of regime-switching models. Those models provide theoretical foundations, practical pricing tools and econometric testing procedures. Together they serve as an empirical evidence that representing structural changes, volatility clustering and state-dependent behavior are crucial for proper understanding of financial markets.

2.2 Optimal Transport and Generative Modeling

Regime-switching models provide a powerful framework for capturing state-dependent dynamics in financial markets. However, they rely on relatively structured parametric assumptions and are not designed to learn complex probability distributions directly. Recent developments in machine learning have therefore shifted attention toward distribution-based approaches. Optimal Transport (OT) has emerged as a particularly attractive framework for this purpose. Especially due to its ability to measure distances between probability distributions, while accounting for their geometric properties. OT has evolved from a theoretical concept in optimization and probability theory into a practical tool for deep generative modeling. The literature reveals an evolution of OT-based methods, from foundational theoretical developments to sophisticated generative models. At the same time, it identifies several important challenges, including computational complexity, robustness and scalability in high-dimensional settings, which continue to drive ongoing research.

A broad overview of OT in machine learning is provided by [6]. This article describes OT as a general-purpose framework for comparing and manipulating probability distributions. It is applicable across a variety of machine learning settings including supervised, unsupervised, transfer, and reinforcement learning. According to this work, OT provides a key concept for comparing probability distributions through the Wasserstein distance. This allows to overcome main limitations of traditional measures such as Kullback–Leibler divergence or Jensen–Shannon divergence. Especially when distributions have non-overlapping support, Wasserstein distance remains meaningful. Therefore, OT serves as a suitable approach for many applications, including generative modeling.

Building on this general foundation, in [7] the focus is narrowed to generative modeling. In this article, authors present the growing importance of OT in generative modeling and categorize OT-based generative models into dual, primal, relaxed, and approximated formulations. Dual formulations, exemplified by Wasserstein GANs (WGANs), leverage Kantorovich duality to improve training stability and mitigate mode collapse, which directly addresses issues identified in earlier GAN frameworks. Alternatively, primal formulations such as Wasserstein Autoencoders (WAEs), optimize transport costs more directly. Further, relaxed and approximated variants employ techniques

such as Sinkhorn regularization as an attempt to balance computational tractability with fidelity to the true OT objective. These developments demonstrate how OT evolved into a practical tool for modern deep learning. Nevertheless, classic OT faces several challenges, such as sensitivity to outliers, choice of cost functions and the curse of dimensionality. These limitations have motivated the development of more flexible OT formulations.

One important direction is the robust OT framework proposed by [8]. This work directly tackles the sensitivity of standard OT to noisy or corrupted data - a limitation acknowledged but not yet resolved. This method relaxes marginal constraints and introduces adaptive weighting mechanisms that reduce the influence of outliers. Importantly, by deriving an efficient dual formulation, this method makes it compatible with deep-learning optimization. This contribution demonstrates an improvement from WGAN-style dual formulations. While WGAN improves stability, robust OT further enhances data reliability and resilience, particularly in real-world noisy settings. Improved performance in GAN training and domain adaptation strengthen the practical importance of this robustness.

A closely related approach is presented in [9], where the authors build on unbalanced OT method, but focus on improving optimization stability and efficiency. Unlike the robust OT framework, which emphasizes theoretical robustness to outliers, this paper introduces a semi-dual formulation. By requiring only a single potential function, this idea simplifies optimization while keeping the advantages of relaxed transport constraints. This aligns with the broader trend identified in earlier surveys toward reducing computational and optimization complexity. Therefore, mentioned works share a common idea. Namely that relaxing strict marginal constraints is key to making OT viable in practical generative modeling. The semi-dual approach not only improves convergence stability but also empirical performance (for example, lower FID scores). Hence, unbalanced OT serves as a promising solution to many practical limitations of previous OT formulations.

An even more direct use of OT is proposed in [10]. This work presents a conceptual shift, as it proposes to learn the optimal transport map itself as the generative model, instead of minimizing transport distance. This idea directly addresses a limitation implicit in earlier approaches. In particular, distance-based objectives do not explicitly model the transformation between distributions. By learning the transport map directly, the framework captures structural relationships more explicitly. This leads to improved performance in tasks such as image generation and reconstruction. This approach can be viewed as a natural extension of primal OT formulations discussed in the survey literature, but taken one step further. Namely, from optimizing transport cost to explicitly modeling the transport mechanism. This approach also addresses concerns related to latent-space approximations by operating directly in high-dimensional spaces.

The literature reflects the rapid evolution of optimal transport in machine learning. From a theoretical distance metric to a versatile and increasingly practical framework for generative modeling. The shift from classic OT formulations to robust, unbalanced, and map-based approaches highlights ongoing efforts to address real-world challenges. Therefore, it shows that OT will continue to play an important role in the development of generative models.

2.3 Research Gap and Motivation

The reviewed literature highlights two complementary developments. On one hand, regime-switching models provide a theoretically grounded and empirically validated framework for modeling structural changes in financial markets, volatility dynamics, and option pricing behavior. On the other hand, Optimal Transport has emerged as a flexible and geometrically motivated framework for learning complex distributions and constructing powerful generative models.

However, a gap remains between these two strands of research. Traditional regime-switching models often rely on relatively restrictive parametric assumptions and can struggle to capture complex distributional structures. Conversely, OT-based generative models offer substantial flexibility but generally lack explicit mechanisms for representing economically interpretable market regimes. Bridging these approaches may therefore provide a promising direction for modeling financial distributions that exhibit both regime-dependent dynamics and high-dimensional nonlinear structure.

Recent work in [11] proposes a variational autoencoder (VAE) framework to address data scarcity in implied volatility surfaces (IVS). The resulting latent space provides a low-dimensional representation of IVS dynamics. This allows to generate economically interpretable surfaces in a controllable manner. Empirical results demonstrate that this model captures stylized features characteristic for stressed market regimes. While these results indicate that different market conditions are encoded within the latent space, the underlying geometric structure of this representation remains unexplored. In particular, it is unclear whether "normal" and "stressed" regimes correspond to distinct latent modes and whether boundaries between them can be detected and characterized.

Motivated by this gap, this thesis investigates whether regime transitions correspond to geometric structures such as mode boundaries. If the answer is positive, the next question is whether boundaries between market regimes can be detected and characterized. In order to answer these questions, the Autoencoder Optimal Transport (AE-OT) framework presented in [12] will be used. By leveraging OT geometry and the associated power diagram partitions of the latent space, the proposed approach aims to identify singular sets that separate latent modes. Such insights could not only improve the interpretability of generative models for IVS data, but also enable the controlled generation of synthetic samples representing stressed market scenarios.

3 Autoencoder Optimal Transport method

Autoencoder - Optimal Transport (AE-OT) is a method for image generation proposed by authors of [12]. It consists of two major steps. In the first step an Autoencoder (AE) is trained for two purposes. First, to encode the data manifold from the image space $\mathcal{X}$ to the latent space $\mathcal{Z}$, and second, to map the data distribution to the latent code distribution. Then the decoder (which is the mirror of the encoder) decodes the latent code back to the data manifold. The second step, which we will focus on in this section, is the optimal transportation (OT). Its goal is to transform one distribution into another and minimize the total transport cost. In this step, the optimal transportation map T from the noise distribution to the latent code distribution is computed. Then, by combining the extended OT map and the decoder, new images are generated from the noise distribution. This model allows to avoid representing discontinuous maps by Deep Neural Networks, hence it prevents mode collapse and mode mixture. This prevention is assured by modeling

globally continuous Brenier potential and locating the singularity set (the discontinuous points of the gradient map), based on Figalli's theory. To make every part of the OT step clear, all of its details and parts of optimal transport theory which are crucial for this setup will be explained in this section.

OT component of the method is constructed by several steps:

1. find the Brenier potential u_h by convex optimization process
2. gradient of Brenier potential is the semi-discrete OT map (target of this map is the discrete set of latent codes of training samples)
3. transport map T is piece-wise linearly extended to a global continuous map $\tilde{T}$
4. the target domain becomes a polytope obtained by triangulating the latent codes.
5. locate Figalli's singularity set in source domain
6. avoid singularity set when generating new samples

It is Worth noticing that when the support of the data distribution is non-convex, the transport map will be discontinuous. Moreover, this method converges to the unique global optimum with bounded error estimate. According to Figalli's theory [13], if there are multiple modes or the support of the target distribution is non-convex, there will be singular sets, where the Brenier potential is continuous but not differentiable, making its gradient map, i.e. the transport map, discontinuous.

3.1 Optimal Transport theory

This section is the basis for understanding optimal transport theory and geometrical interpretation that stands behind it. First, assumptions of considered Semi-Discrete Optimal transport problem have to be established.

3.1.1 Problem setup

Let $X \subset \mathbb{R}^d$ be a convex domain, and let μ be an absolutely continuous source measure (Gaussian or uniform distribution) on X. Let the target domain be a discrete set, $Y = \{y_1, y_2, \ldots, y_n\}, y_i \in \mathbb{R}^d$. Let the target measure ν be discrete and such that

$$\nu = \sum_{i=1}^{n} \nu_i \, \delta_{y_i}, \qquad y_i \in Y \subset \mathbb{R}^d, \quad \nu_i > 0, \quad \sum_{i=1}^{n} \nu_i = \mu(X). \tag{1}$$

We consider the quadratic transportation cost $c : X \times Y \mapsto \mathbb{R}$, which can be interpreted as the work needed to send one unit of mass from location $x \in X$ to location $y \in Y$:

$$c(x, y) = \frac{1}{2}\|x - y\|^2. \tag{2}$$

Our aim is to find the Optimal Transportation map, hence to find a map with a minimal transportation cost. In the literature this problem is known as the Monge problem [14] and is formulated

as follows.

Given measures μ and ν and a transportation cost function $c : X \times Y \mapsto \mathbb{R}$, one considers the infimum of the total transportation cost:

$$\inf \int_{X\times Y} c(x,y)d\pi(x,y), \tag{3}$$

where the infimum is over all joint probability measures π on $X \times Y$ with marginals μ and ν.

Because of the assumptions above, this problem can be simplified. However, in order to do that, some definitions and theorems are needed to fully understand why this simplification is valid.

Definition 1 (Push forward measure[15]) *A map $T : X \to Y$ is given, if for any measurable set $B \subset Y$, $\mu(T^{-1}(B)) = \nu(B)$, then ν is said to be the push-forward of μ by T and we write $\nu = T_{\#}\mu$. Map T is said to be measure preserving.*

Under a semi-discrete transportation map $T : X \mapsto Y$, a cell decomposition of the domain space X is induced, $X = \bigcup_{i=1}^{n} W_i$. This decomposition is such that every x in each cell W_i is mapped to the target point y_i, hence $T : x \in W_i \mapsto y_i$. The map T is measure preserving, denoted as $T\#\mu = \nu$, hence the μ-measure of each cell W_i equals to the ν-measure of the y_i. Therefore $w_i = \nu_i$, where $y_i = T(x)$, $\forall x \in W_i$, $\nu_i = \nu(y_i)$, $w_i = \mu(W_i)$.

Since the target domain is discrete and source domain is decomposed such that $X = \bigcup_{i=1}^{n} W_i$ and $W_i \cap W_j = \emptyset$, the total transportation cost of the transportation map T in this setup can be written as

$$\begin{aligned}\int_{X\times Y} c(x,y)\mathrm{d}\pi(x,y) &= \int_{\bigcup_{i=1}^{n} W_i \times Y} c(x,y)\mathrm{d}\pi(x,y) = \sum_{i=1}^{n}\int_{W_i} c(x,y_i)\mathrm{d}\mu(x) \\ &= \int_X c(x,T(x))\mathrm{d}\mu(x).\end{aligned} \tag{4}$$

Therefore, in order to find the optimal transport map $\hat{T}$, a transportation map T, which is measure preserving and minimizes total transportation cost must be found:

$$\hat{T} = \arg\min_{T\#\mu=\nu} \int_X c(x,T(x))\mathrm{d}\mu(x). \tag{5}$$

If cost function c is the power of the Euclidean distance, the total transportation cost is called the Wasserstein distance between the two measures μ and ν, and is denoted as $\mathcal{W}_c(\mu,\nu)$. As mentioned at the beginning of this section, the aim of the Monge problem is to find a mapping T which minimizes cost. This minimized cost is the Wasserstein distance:

$$\mathcal{W}_c(\mu,\nu) = \min_{T:X\mapsto Y} \left\{ \int_X \frac{1}{2}\|x - T(x)\|^2 \, d\mu(x) : T_{\#}\mu = \nu \right\}. \tag{6}$$

There is a substantial relation between optimal mass transport map and convex geometry, which was presented by Brenier in 1991 [16].

Theorem 1 (Brenier [16]) *Suppose $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}^d$ are the Euclidean spaces, the transportation cost is the quadratic Euclidean distance $c(x, y) = \|x - y\|^2$. If μ is absolutely continuous and μ and ν have finite second order moments, then there exists a convex function $u : X \to \mathbb{R}$, such that the gradient map ∇u gives the solution to the Monge's problem. Under these assumptions, the solution is unique (which is not true in general). Function u is called Brenier's potential, ∇u is called Brenier map or the optimal mass transportation map.*

Under assumptions in discussed setup, Brenier's theorem (Theorem 2.32 in [17]) states that there is exactly one measurable map T such that $T\#\mu = \nu$ and $T = \nabla u(x)$ for some convex function u. Uniqueness is in a sense that any two such maps coincide dμ-almost everywhere. Thus the transport geometry is encoded by a convex potential.

3.1.2 Semi-discrete Optimal Transport problem

As previously set, the target measure is discrete $\nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}$, this is why this problem is called semi-discrete optimal transport. Because the target is supported on the finite set $Y = \{y_1, \dots, y_n\}$, and cost function is the L^2 distance, Brenier's theorem can be applied directly. Thus, the map $T = \nabla u(x)$ is defined by the function u such that:

$$u_h(x) = \max_{1 \le i \le n} \big(\langle x, y_i \rangle + h_i\big), \tag{7}$$

where the scalars h_i are height parameters.
Define the affine functions

$$\pi_i(x) := \langle x, y_i \rangle + h_i. \tag{8}$$

Then

$$u_h(x) = \max_{1 \le i \le n} \pi_i(x). \tag{9}$$

Therefore u_h is a convex, continuous, piecewise-linear function. In fact, the projection of Brenier potential induces a cell decomposition mentioned previously. Recall that

$$W_i(h) := \{x \in X : T(x) = y_i\} = \{x \in X : \nabla u_h(x) = y_i\}, \tag{10}$$

and μ-measure of $W_i(h)$ is denoted as $w_i(h) = \mu(W_i(h))$.

Now, the formulation of convex energy can be introduced.

Theorem 2 (Discrete Brenier Theorem [18]) *For any $\nu_1, \dots, \nu_n > 0$, with $\sum_{i=1}^{n} \nu_i = \mu(X)$, there exists vector of heights $h = (h_1, \dots, h_n) \in \mathbb{R}^n$, unique up to adding a constant, such that $w_i(h) = \nu_i$ for all i. The vector h is the unique minimum argument of the following convex energy*

$$E(h) = \int_0^h \sum_{i=1}^{n} w_i(\eta)\, d\eta_i - \sum_{i=1}^{n} h_i \nu_i, \tag{11}$$

defined on an open convex set $\mathcal{H} = \{h \in \mathbb{R}^n : w_i(h) > 0, i = 1, \dots, n\}$. Moreover, ∇u_h minimizes the quadratic cost $\int_X \|x - T(x)\|^2 d\mu(x)$ among all transport maps $T_\#\mu = \nu$. The gradient and Hessian of above energy is given respectively by

$$\nabla E(h) = (w_1(h) - \nu_1, \dots, w_n(h) - \nu_n)^T \tag{12}$$

and

$$\frac{\partial w_i}{\partial h_j} = -\frac{\mu(W_i \cap W_j)}{\|y_i - y_j\|}, \quad \frac{\partial w_i}{\partial h_i} = \sum_{j \neq i} \frac{\partial w_i}{\partial h_j}. \tag{13}$$

The optimal transport map can be obtained by convex optimization. It is easily observed, that for the optimal solution, the optimal height vector h satisfies $\nabla E(h) = 0$, which is exactly the mass-balance condition $w_i(h) = \nu_i, \forall i$. Since the function E is convex, minimizing it yields the correct heights. This optimization can be carried out using Newton's method. The global linear convergence rate is guaranteed by the Theorem 1.5 in [19].

Through the construction above, the semi-discrete OT map $\nabla u_h : X \mapsto Y$ maps all $x \in X$ to the latent codes of training samples. However, it will not generate new samples. In order to be able to use it as a generative model, a piece-wise linear (PL) extension of the OT map, $\tilde{T}$ is needed. The projection of u_h induces a cell decomposition of space $X = \bigcup_{i=1}^{n} W_i$. Each cell W_i is of μ-volume w_i and is mapped to the corresponding y_i of volume ν_i. By representing the cells by their μ-mass centers as $c_i := \int_{W_i(h)} x \mathrm{d}\mu(x)$, we can get the point-wise map $t : c_i \mapsto y_i$. The dual pair of the cell decomposition induces a triangulation of the centers $C = \{c_i\}$. If $W_i \cap W_j \neq \emptyset$, then c_i is connected with c_j to form an edge $[c_i, c_j]$. Similarly, if $W_{i_0} \cap W_{i_1} \cdots \cap W_{i_k} \neq \emptyset$, then there is a k-dimensional simplex $[c_{i_0}, c_{i_1}, \ldots, c_{i_k}]$. All these simplices form a triangulation (simplicial complex) of C, denoted as $T(C)$. Space Y can be triangulated in the same way to obtain the triangulation $T(Y)$. Once a random sample x is drawn from the distribution μ, a simplex σ in $T(C)$ containing x can be found.

Assume that simplex σ has $d+1$ vertices $\{c_{i_0}, c_{i_1}, \ldots, c_{i_d}\}$, the bary-centric coordinate of x in σ is defined as $x = \sum_{k=0}^{d} \lambda_k c_{i_k}$, and $\sum_{k=0}^{d} \lambda_k = 1$ and $\lambda_k \geq 0$ for all k. Then the generated latent code of x under the piece-wise linear OT map is given by $\tilde{T}(x) = \sum_{k=0}^{d} \lambda_k y_{i_k}$. Hence, all of the target points y_i are used to construct the simplicial complex $T(Y)$ in the support of the target distribution. Thus, there is a guarantee that no mode is lost.

3.1.3 Geometric interpretation

Introducing Minkowski problem from convex geometry, will help in understanding geometric interpretation of optimal transport map with Euclidean quadratic cost. Minkowski problem considers a question: "Given a collection of proposed facet normals and facet areas, is there a convex polytope in $\mathbb{R}^n$ whose facets fit the given data? If the answer is yes, is this polytope unique?" [20]. In order to find the answer for this question, the Minkowski theorem will be presented. Before that, some basic definitions and results of convex geometry are needed.

Definition 2 (Hyperplane) *A hyperplane is any codimension-1 vector subspace of a vector space.*

Note: Codimension indicates the difference between the dimensions of considered object and a smaller one, contained in it.

Definition 3 (Convex hull) *The convex hull of a set of points $S \subset \mathbb{R}^n$ is the intersection of all convex sets containing S.*

Definition 4 (Polytope) *A n-polytope P is the convex hull of finitely many points in $\mathbb{R}^n$. The set of convex polytopes in $\mathbb{R}^n$ is denoted as $\mathcal{P}^n$.*

Definition 5 (Face) *A face of the n-dimensional polytope is its intersection with a tangent hyperplane.*

Definition 6 (Facet) *If $dimP = n$ and P^u is a face of dimension $n-1$, then P^u is called a facet of P, where u is a corresponding facet normal.*

Definition 7 (Volume of a polytope) *A volume of a polytope $P \in \mathcal{P}^n$ is the Lebesgue measure of n-dimensional space enclosed by the polytope's boundaries in $\mathbb{R}^n$ and is denoted by $Vol(P)$.*

Definition 8 (Power distance [21]) *Let X be a Euclidean space with the distance function $d(x_1, x_2) = \|x_1 - x_2\|$.Given a point $x \in \mathbb{R}^d$ and a point $y_i \in \mathbb{R}^d$ with a power weight ψ_i, the power distance between x and y_i is given by $pow(x, y_i) = \|x - y_i\|^2 - \psi_i$.*

Definition 9 (Power diagram [21]) *Given the weighted points $(y_1, \psi_1), \ldots, (y_k, \psi_k)$, the power diagram is the cell decomposition of $\mathbb{R}^d$:*

$$\mathbb{R}^d = \bigcup_{i=1}^{k} W_i(\psi) \tag{14}$$

where each cell is a convex polytope:

$$W_i(\psi) = \{x \in \mathbb{R}^d : pow(x, y_i) \leq pow(x, y_j), \forall j \neq i\}. \tag{15}$$

With these basic definitions from the convex geometry, useful theorems can be now presented.

Theorem 3 (Minkowski [20]) *Suppose that $u_1, \ldots, u_k \in \mathbb{R}^n$ are unit vectors that span $\mathbb{R}^n$. Suppose that $\alpha_1, \ldots, \alpha_k > 0$. Then there exists a polytope $P \subset \mathcal{P}^n$ having facet unit normals $u_1, \ldots, u_k$ and corresponding facet areas $\alpha_1, \ldots, \alpha_k$ if and only if $\sum_{i=1}^{k} \alpha_i u_i = 0$. Moreover, this polytope is unique up to translation.*

The following theorem shows the connection between Discrete Brenier Theorem 2 and a convex geometry theory.

Theorem 4 (Alexandrov [22]) *Suppose X is a compact convex polytope with non-empty interior in $\mathbb{R}^n$, $n_1, \ldots, n_k \in \mathbb{R}^{n+1}$ are distinct k unit vectors with the negative (n+1)-th coordinates. Suppose that $\alpha_1, \ldots, \alpha_k > 0$ such that $\sum_{i=1}^{k} \alpha_i = Vol(X)$. Then there exists convex polytope $P \subset \mathbb{R}^{n+1}$ with exact k codimension-1 faces $F_1, \ldots, F_k$ so that n_i is the normal vector to F_i and the intersection between X and the projection of F_i is with volume α_i. Moreover, this polytope P is unique up to translation.*

Consider vectors $y_i \in \mathbb{R}^d$ for $i = 1, \ldots, k$ and heights vector $h = (h_1, \ldots, h_k) \in \mathbb{R}^k$. Define a piecewise linear convex function $u_h(x)$ as

$$u_h(x) = \max_{i=1}^{k} \{\langle x, y_i \rangle + h_i\} \tag{16}$$

Then the graph of u_h is a convex polytope in $\mathbb{R}^{d+1}$ and the projection induces a cell decomposition of $\mathbb{R}^d$. Each cell is a closed, convex polytope:

$$\begin{aligned} W_i(h) &= \{x \in X : \nabla u_h(x) = y_i\} \\ &= \{x \in X : \langle x, y_i \rangle + h_i \geq \langle x, y_j \rangle + h_j, \ \forall j = 1, \ldots, k\} \\ &= \{x \in X : \pi_i(x) \geq \pi_j(x), \forall j = 1, \ldots, k\}. \end{aligned} \tag{17}$$

Such cell is also called a region where the i-th plane is active.

Now given a probability measure μ defined on space $X \subset \mathbb{R}^d$, the μ-volume of the polytope $W_i(h)$ is

$$Vol(W_i(h)) = \mu(W_i(h) \cap X) = \int_{W_i(h)\cap X} \mathrm{d}\mu = w_i(h). \tag{18}$$

The collection $\{W_i(h)\}_{i=1}^n$ is a partition of X. This is exactly a power diagram (also called a Laguerre diagram). Consider weights $w_i := -2h_i - \|y_i\|^2$. Using the equality $\|x-y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x, y\rangle$, it is easy to observe that the inequality from eqiation 17

$$\langle x, y_i\rangle + h_i \geq \langle x, y_j\rangle + h_j \tag{19}$$

is equivalent to

$$\|x - y_i\|^2 + w_i \leq \|x - y_j\|^2 + w_j. \tag{20}$$

Hence

$$W_i(h) = \left\{x \in X : \|x - y_i\|^2 + w_i \leq \|x - y_j\|^2 + w_j,\ \forall j\right\}, \tag{21}$$

is precisely the power cell associated with y_i and w_i. Hence, inside a cell $W_i(h)$, a single affine function π_i is active (attains the maximum), so the semi-discrete optimal transport is:

$$T(x) = \nabla u_h(x) = y_i \qquad \text{for } x \in \operatorname{int} W_i(h). \tag{22}$$

Thus this map is piecewise constant: $W_i(h) \longmapsto y_i$. The geometric interpretation of this map can be seen in figure **??** from [15]. The potential u_h is continuous everywhere, because it is the maximum of affine functions. However, it fails to be differentiable on the boundaries where two or more planes tie $\pi_i(x) = \pi_j(x) = u_h(x)$. Across such a boundary, the gradient jumps from y_i to y_j. Therefore u_h is continuous and ∇u_h is discontinuous on cell boundaries.

Theorem 5 (Aleksandrov [18]) *Let $X \subset \mathbb{R}^d$ be a compact convex domain, $\{y_1, \ldots, y_k\} \subset \mathbb{R}^d$ be a set of distinct points. Let μ be a probability measure on X. Then for any $\nu_1, \ldots, \nu_k > 0$ with $\sum_{i=1}^k \nu_i = \mu(X)$, there exists $h = (h_1, \ldots, h_k) \in \mathbb{R}^k$ unique up to a constant, so that $w_i(h) = \nu_i$ for all $i = 1, \ldots, k$. Vectors h are the maximum points of the concave function*

$$E(h) = \sum_{i=1}^k h_i\nu_i - \int_0^h \sum_{i=1}^k w_i(\eta) d\eta_i \tag{23}$$

on the open convex set

$$H = \{h \in \mathbb{R}^k : w_i(h) > 0, \forall i\}. \tag{24}$$

Moreover, ∇u_h minimizes the quadratic cost

$$\int_X \|x - T(x)\|^2 d\mu(x) \tag{25}$$

for all transport maps such that $T_{\#}\mu = \nu$, where the Dirac measure $\nu = \sum_{i=1}^k \nu_i \delta(y - y_i)$.

Therefore, the connection with Theorem 2 is explicit. The last object which needs to be translated into language of geometry is Brenier potential. The affine functions from equation 8 can be transleted into the affine hyperplanes in $\mathbb{R}^d$:

$$\pi_i(x) = \langle x, y_i \rangle + h_i, \tag{26}$$

and their upper envelope $\max_i \pi_i(x)$, which is a convex polyhedral surface in $\mathbb{R}^d$ is in fact the Brenier potential from equation 9. Thus each target point y_i determines one supporting hyperplane, the Brenier potential is the upper hull of these hyperplanes and geometry of OT is encoded by a convex polyhedron in one higher dimension $\mathbb{R}^{d+1}$.

3.2 Singular Set Detection

If there are multiple modes or the support of the target distribution is concave, Figalli's theory [13] indicates the existence of the singular set $\mathcal{S} \subset X$. The singular set contains such points, where the Brenier potential is continuous but not differentiable. This makes gradient map of the Brenier potential (thus the transport map) discontinuous.

The easiest way to understand interpretation of the singular set is to look at the example. One can consider the source distribution which is uniformly defined on X. Under the assumption that the target empirical distribution has two modes, there is the ridge on the Brenier potential u_h in the source space X. The projection of this ridge is the singular set $\mathcal{S}$. Then, the set $X \setminus \mathcal{S}$ consists of two connected components. Each of those components is mapped to a single mode. Singular set $\mathcal{S}$ consists of codimension 1 facets of cells. If $W_i(h) \cap W_j(h) \subset \mathcal{S}$, then the dihedral angle between the two supporting planes, $\pi_{h,i}$ and $\pi_{h,j}$ respectively, of u_h is large. This feature is useful for the next step. This is the example with two modes. However, the procedure is analogous for any case.

The threshold angle $\hat{\theta}$ is fixed. Then, from the graph of Brenier potential, pairs of facets whose dihedral angles exceed this threshold are chosen. The projection of intersection of those pairs, gives a co-dimension 1 cell in the singular set $\mathcal{S}$. If a random sample x is in the singular set $\mathcal{S}$, it is abandoned during the generation step. This approach prevents the mode mixture phenomenon.

In order to discover if $x \in \mathcal{S}$, the extended OT map is used. Given the extended OT map $\tilde{T}(x)$, some of the polyhedrons transverse the singular set. This means that vertices of the polyhedron belong to different modes. Therefore, if the sample x falls into such a polyhedron it is dropped during the generation step.

In order to check if $x \in \mathcal{S}$, the angles θ_{i_k} between π_{i_0} and π_{i_k}, $k = 1, 2, \ldots, d$ are computed by $\theta_{i_k} = \langle y_{i_0}, y_{i_k} \rangle / \|y_{i_0}\| \cdot \|y_{i_k}\|$. Then, if for all k the angles $\theta_{i_k} > \hat{\theta}$, it means that x belongs to the singular set. Other approach is to select, for a given x, the subset of planes $\{\pi_{i_k}\}$ such that $\theta_{i_k} \leq \hat{\theta}$, denoted as $\{\pi_{\hat{i}_k},\ k = 0, \ldots, d_1\}$.

In order to obtain extended OT map, first barycentric coordinates λ_k must be computed:

$$\lambda_k = \frac{d^{-1}(x, \hat{c}_{i_k})}{\sum_{j=0}^{d_1} d^{-1}(x, \hat{c}_{i_j})}. \tag{27}$$

Then the extended OT map is given by

$$\tilde{T}(x) = \sum_{k=0}^{d_1} \lambda_k T(\hat{c}_{i_k}). \tag{28}$$

This piece-wise linear extension of the transport map $\tilde{T}(\cdot)$ allows to smooth the discrete function $T(\cdot)$ in regions of dense latent codes. Moreover, it keeps the discontinuity of $T(\cdot)$ in regions where latent codes are very sparse. In this way the generation of spurious latent code is avoided. As a result, the generation quality is improved [12].

3.3 AE-OT: OT component implementation

The Optimal Transport theory and its geometrical interpretaion allows to move to the AE-OT method itself. The full algorithm can be found in [12]. Therefore, elements crucial for understanding the pipeline will be presented in this section while implementation details can be seen in the original paper.

As mentioned at the beginning of this section, the Optimal Transport component consists of 6 steps. These are divided into two separate algorithms in the original paper [12]. Algorithm 1 consists of first two steps, which involve finding Brenier potential u_h, and its gradient which is the semi-discrete OT map.
To accomplish this, four input objects are given:

- latent codes $Y = \{y_i\}_{i \in \mathcal{I}}$
- empirical distribution $\nu = \frac{1}{\#\mathcal{I}} \sum_{i \in \mathcal{I}} \delta_{y_i}$
- number of Monte Carlo samples N
- maximum number of iterations s

Main steps of this algorithm are as follows:

1. initialize heights vector: $h = (h_1, \ldots, h_n) = (0, \ldots, 0)$
2. generate N random samples from μ distribution (e.g. uniform): $\{x_j\}_{j=1}^N$
3. find cell W_i such that $x_j \in W_i$ by: $i = \arg\max_i \{\langle x_j, y_i \rangle + h_i\}, i = 1, 2, \ldots, n$
4. calculate the estimated $\mu-$volume of each cell $W_i(h) : \hat{w}_i(h) = \frac{1}{N} \#\{j \in \mathcal{I} : x_j \in W_i(h)\}$, (when N is large enough, $\hat{w}_i(h)$ converges to $w_i(h)$)
5. calculate $\nabla E(h) \approx (\hat{w}_i(h) - \nu_i)^T$
6. update h (to minimize $E(h)$), use Adam optimizer
7. check if convex energy $E(h)$ has decreased in s steps, if not then double the number of Monte Carlo samples
8. continue until the convergence is accomplished

As an output, the Optimal Transport map is given $T(\cdot) = \nabla(\max_i \langle \cdot, y_i \rangle + h_i)$.

Next steps are included in algorithm 2 in [12]. This part is responsible for latent code generation. It consists of the PL extension of the OT map, triangularization of the latent codes, locating Figalli's singularity set and avoiding it during generation of new samples.
The procedure starts with 3 inputs:

- OT map: $T(\cdot)$
- number of samples to generate: n
- angle threshold: $\hat{\theta}$

Main steps of the algorithm are as follows:

1. sample $x \sim \mu$
2. compute approximated mass centers using Monte Carlo method: $\hat{c}_i = \dfrac{\sum_{x_j \in W_i} x_j}{\#\{j : x_j \in W_i\}}$
3. compute and sort in the ascending order Euclidean distances between x and mass centers: $d(x, \hat{c}_i)$, $i = 1, 2, \ldots, n$
4. first $d+1$ distances are then given by: $\{d(x, \hat{c}_{i_0}), \ldots, d(x, \hat{c}_{i_1}), \ldots, d(x, \hat{c}_{i_d})\}$
5. simplex σ that contains x has $d+1$ vertices: $\{\hat{c}_{i_0}, \hat{c}_{i_1}, \ldots, \hat{c}_{i_d}\}$
6. barycentric coordinates are estimated as: $\hat{\lambda}_{i_k} = \frac{d^{-1}(x, \hat{c}_{i_k})}{\sum_{k=0}^{d} d^{-1}(x, \hat{c}_{i_k})}$ for $k = 0, \ldots, d$
7. compute dihedral angles θ_{i_k} between π_{i_0} and π_{i_k}, $k = 1, \ldots, d$ by $\theta_{i_k} = \frac{\langle y_{i_0}, y_{i_k} \rangle}{\|y_{i_0}\| \cdot \|y_{i_k}\|}$.
8. select θ_{i_k} with $\theta_{i_k} \leq \hat{\theta}$.
   This results in indices $\hat{i}_k$, with $k = 0, 1, \ldots, d_1$ such that $\theta_{\hat{i}_k} \leq \hat{\theta}$ for all $\hat{i}_k$.
9. For new indices λ_k beomes: $\lambda_k = \frac{d^{-1}(x, \hat{c}_{\hat{i}_k})}{\sum_{j=0}^{d_1} d^{-1}(x, \hat{c}_{\hat{i}_j})}$.
10. for new indices $\hat{i}_k = 0, 1, \ldots, d_1$, generate latent code using $\tilde{T}(x) = \sum_{k=0}^{d_1} \lambda_k T(\hat{c}_{\hat{i}_k})$
11. in this manner generate n new latent codes

As an output, n new latent codes P are generated.

4 Preliminary results

In this section some preliminary results will be presented. First, the experiment on the MNIST dataset from the original paper [12] was reproduced. Then, three toy examples were introduced and experiment using AE-OT method was conducted on these toy examples.

4.1 AE-OT method performed on MNIST dataset

As a first step in numerical work during the literature review phase, results of experiment on the MNIST dataset of handwritten digits have been reproduced from [12]. Using method proposed by the authors, which combines Autoencoders (AE) and Optimal Transport (OT), the key idea is to first learn a meaningful latent representation of the data using an autoencoder and then model the structure of this latent space using Optimal Transport. New samples are generated by interpolating within the latent space according to the learned transport structure. Then the resulting latent vectors are decoded back into the image space. Using the original CelebA-based code provided by the authors, the same generative pipeline has been adapted to the MNIST dataset.

4.1.1 MNIST dataset

The experiments are conducted on the MNIST dataset, which consists of grayscale images of handwritten digits ranging from 0 to 9. Each image has a resolution of 28 by 28 pixels and contains a single channel. Dataset consists of 60,000 training samples and 10,000 test samples. Images are loaded using the `torchvision.datasets.MNIST`, converted into tensors and normalized to have values between 0 and 1.

4.1.2 Model Architecture

The core of the pipeline is the autoencoder, which is responsible for learning a compact and meaningful latent representation of the input images. The encoder is implemented as a series of convolutional layers that progressively reduce the spatial dimensions of the input while increasing the number of feature channels. Precise architecture can be seen in Table 1.

The decoder mirrors the process using transposed convolutions layers gradually reconstructing the spatial structure of the image from the latent representation. Precise architecture can be seen in Table 2. Unlike the architecture proposed in the original paper, the implemented autoencoder is fully convolutional and does not contain any fully connected layers. The latent representation is obtained directly by the final convolutional layer of the encoder and expanded by a transposed convolutional layer in the decoder Instead of tanh activation function, as in the original code, a sigmoid activation function is used in the final layer, which ensures that output values lie in $[0, 1]$.

The training objective for the autoencoder consists of two components. The first is the mean squared error between the input image and its reconstruction, the second is an L_1 regularization term applied to the latent vector, weighted by a small coefficient λ:

$$\mathcal{L} = \text{MSE}(x, \hat{x}) + \lambda \|z\|_1$$

where x is an input image, $\hat{x}$ is its reconstruction, $\lambda = 1e-5$, z is the latent vector and $\|\cdot\|_1$ is L_1 norm.

Once the autoencoder is trained, all training images are passed through the encoder to obtain their corresponding latent vectors: $x \to z$. These vectors are stored and serve as the input to the Optimal Transport algorithm. Parameters used in training stage can be seen in Table 3.

Table 1: Encoder architecture for MNIST experiment

layer	**number of outputs**	**kernel size**	**stride**	**BN**	**activation**
Input $x \sim P_{data}$	$28 \times 28 \times 1$	–	–	–	–
Convolution	$14 \times 14 \times 64$	4×4	2	No	LeakyReLU
Convolution	$7 \times 7 \times 128$	4×4	2	Yes	LeakyReLU
Convolution	$3 \times 3 \times 256$	4×4	2	Yes	LeakyReLU
Convolution	$1 \times 1 \times 2$	3×3	2	No	LeakyReLU

Table 2: Decoder architecture for MNIST experiment

layer	**number of outputs**	**kernel size**	**stride**	**BN**	**activation**
Input $z \sim P_{latent}$	$1 \times 1 \times 2$	–	–	–	–
Transposed Convolution	$3 \times 3 \times 256$	3×3	2	Yes	ReLU
Transposed Convolution	$7 \times 7 \times 128$	4×4	2	Yes	ReLU
Transposed Convolution	$14 \times 14 \times 64$	4×4	2	Yes	ReLU
Transposed Convolution	$28 \times 28 \times 1$	4×4	2	No	Sigmoid

Table 3: MNIST experiment parameters

Parameter	**Value**
train epochs	150
refinement epochs	50
batch size	256
learning rate	10^{-3}
latent dimension z	2
coefficient λ	10^{-5}
optimizer	Adam
angle threshold θ	$\arccos(0.8)$
interpolation parameter w	0.75

4.1.3 Optimal Transport

The Optimal Transport component operates on the set of latent vectors $\{z_i\}_{i=1}^N$ obtained from the autoencoder. The objective is to learn a function that partitions the latent space such that randomly sampled points are evenly assigned to the latent vectors. This is achieved by introducing a scalar value h_i for each latent point. During training, random samples are generated using a Sobol sequence. For each random sample, a score is computed with respect to each latent point, and the sample is assigned to the point with the highest score. By counting how many samples are assigned to each latent point, a histogram is obtained. The goal is to adjust the scalar values such that this histogram becomes uniform. The scalar values are updated using an optimization procedure similar to the Adam optimizer. Over multiple iterations, this process leads to a balanced partitioning of the latent space and trained OT model is stored.

4.1.4 Feature generation and decoding

After the Optimal Transport model has been trained, new latent vectors are generated by interpolating between pairs of existing latent vectors. The selection of pairs is guided by a measure of angular distance, ensuring that only sufficiently distinct pairs (z_i and z_j) are used. This encourages diversity in the generated samples. The interpolation is controlled by a parameter w that determines how far the generated point lies between the two selected latent vectors. The resulting latent vectors $z_{\text{new}} = (1 - w)z_i + wz_j$ are then passed through the decoder to produce new images. What's important, only interpolated latent vectors are used for generation. The inclusion of original latent vectors in the generated set is avoided to ensure that the evaluation reflects true generative performance.

4.1.5 Experiments and Results

The model was evaluated under different configurations to assess its behavior. In small configurations with few training epochs and reduced model capacity, the generated digits were recognizable but often noisy or distorted. These experiments were useful for verifying the correctness of the pipeline. In the targeted configuration with increased latent dimension and more training epochs, the generated digits became clearer and more consistent. The model was able to capture the structure of handwritten digits and produce plausible variations.

The results can be seen in figure 1. This qualitative evaluation shows that the model generates smooth interpolations between digits and visually generated digits are as good as digits from the real data. Moreover, the importance of localizing the singular set can be observed in the next comparison. In figure 2 one can see unclear and ambiguous images. Some of them are a mixture of two different digits, thus they are difficult to be assigned to a one specific digit. This happens when points from the singular set are accepted and during generation the model is allowed to interpolate between those points. Then, the interpolation between two different modes can occur and the results can be seen in mentioned figure. On the other hand, in figure 3 the correct results are presented. When the angle threshold is set correctly, and points from the singular set are ommited, then the quality of generated samples is extremely better. That happens because the problem of mode mixture is avoided.

Quantitative evaluation was performed using the Fréchet Inception Distance (FID), comparing generated images with real MNIST test images. The FID result obtained via torch-fidelity is equal to 5.73. This agrees with the values presented in the paper, where FID for autoencoder is reported to be 5.5 and for the presented method 6.4.

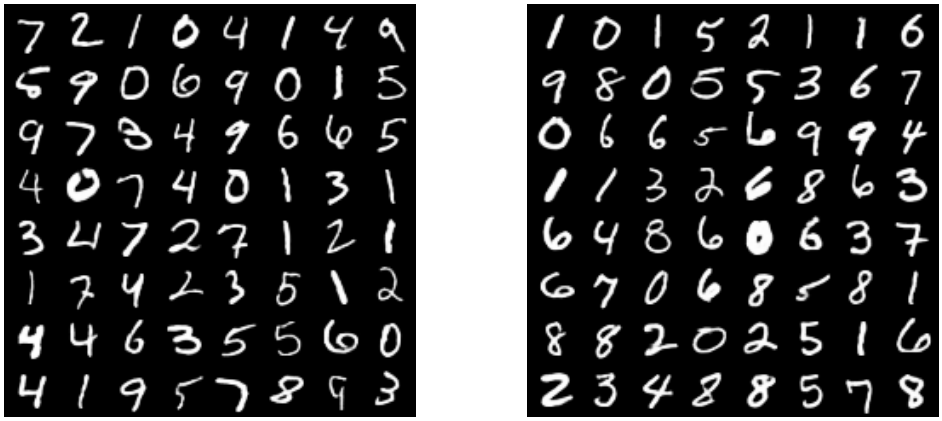

Figure 1: The visual comparison of real data (left) and result of used method (right)

Figure 2: Examples of rejected samples

Figure 3: Examples of accepted samples

4.2 AE-OT method performed on toy examples

In this experiment three toy setups were made. Similarly as in the MNIST experiment, the original code was adapted to this certain experiment, but the main idea follows the algorithm presented in [12].

4.2.1 Toy data

For this experiment three synthetic datasets were constructed. In these datasets the true latent space contains a two-dimensional structured distribution embedded in higher dimensions.

Synthetic latent codes are constructed as follows. Let $z^* \in \mathbb{R}^d$ denote the ground-truth latent variable $z^* = (z_1^*, z_2^*, z_{3:d}^*)$. The first two dimensions follow a structured toy distribution. The remaining $d-2$ dimensions are nuisance noise $z_{3:d}^* \sim \mathcal{N}(0, \sigma_n^2 I)$. Three toy geometries are evaluated.

(i) Asymmetric Two-Gaussian Mixture: $p(z^*_{1:2}) = 0.8\mathcal{N}((-2,0), 0.2^2 I) + 0.2\mathcal{N}((2,0), 1.0^2 I)$

(ii) Two-Moons: $X_1(t) = (\cos t, \sin t),\ X_2(t) = (1 - \cos t, -\sin t - 0.5)$, where $t \sim\ U(0, \pi)$, with equal mixture weights and $z^*_{1:2} = X_k(t) + 0.06\epsilon,\ \epsilon \sim \mathcal{N}(0, I)$

(iii) Ring+Center: with center $z^*_{1:2} \sim \mathcal{N}(0, 0.2^2 I)$ and ring $r = 3 + \epsilon,\ \epsilon \sim \mathcal{N}(0, 0.1^2)$, $\theta \sim U(0, 2\pi),\ z^*_{1:2} = r(\cos\theta, \sin\theta)$

With these two-dimensional structured distributions, observed data in higher dimension is generated. Thus $x \in \mathbb{R}^D$ is generated in two ways.

(1) Via linear decoder: $x = Az^* + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma_x^2 I)$, where $A \in \mathbb{R}^{D\times d}$.

(2) Via nonlinear decoder: $x = g_\psi(z^*) + \epsilon$, where g_ψ is a fixed 3-layer MLP with GELU activations.

4.2.2 Model Architecture

The encoder consists of a sequence of fully connected layers that progressively compress the input vector from dimension D into a low-dimensional latent representation z.The final encoder layer maps the learned feature representation directly into the latent space of dimension $\dim_z$. The precise encoder architecture is summarized in Table 4.

The decoder mirrors the encoder structure, gradually reconstructing the original input vector from the latent representation. Starting from z, a sequence of fully connected layers with ReLU activations expands the representation through intermediate dimensions until the original dimensionality D is recovered. The complete decoder architecture is shown in Table 5.

Table 4: Encoder architecture for toy-data experiment

layer	number of outputs	BN	activation
Input $x \sim P_{data}$	128	–	–
Linear	512	Yes	LeakyReLU
Linear	256	Yes	LeakyReLU
Linear	128	Yes	LeakyReLU
Linear	16	No	Linear

Table 5: Decoder architecture for toy-data experiment

layer	number of outputs	BN	activation
Input z	16	–	–
Linear	128	Yes	ReLU
Linear	256	Yes	ReLU
Linear	512	Yes	ReLU
Linear	128	No	Linear

Unlike the convolutional architecture used in the original AE-OT implementation, the proposed autoencoder is entirely composed of fully connected layers. This design is more appropriate for vector-valued toy datasets where no spatial structure is present. The latent representation is obtained directly from the final encoder layer and serves as a compact embedding of the input data.

After training, all samples from the training set are passed through the encoder to obtain their corresponding latent vectors, $x \to z$. These latent representations are subsequently used as inputs to the Optimal Transport stage, where transport maps are learned directly in the low-dimensional latent space. The training hyperparameters used for the autoencoder are reported in Table 6.

Table 6: Parameters for toy-data experiments

Parameter	Value
Latent dimension ($\dim_z$)	16
Hidden layer width	256
Number of hidden layers	3
Batch normalization	Yes
Optimizer	Adam
Learning rate	10^{-3}
Batch size	512
Number of epochs	30

4.2.3 Experiments and Results

The objective of these experiments is to verify whether AE-OT can recover correct mode partitions, align OT-induced boundaries with true separators, and remain stable under high-dimensional nuisance variations. In order to answer these questions, settings presented in Table 7 were tested. Stability summaries are shown per toy-example. On each of those three datasets, AE-OT method is compared with baselines such as k-means and Gaussian Mixture Model (GMM) by computing mean and standard deviation for tested settings.

Table 7: Toy-data experiments settings

Parameter	Value
Number of samples (n)	10 000
Intrinsic dimension (d)	$\{8, 64, 784\}$
Observation dimension (D)	128
Nuisance noise (σ_n)	$\{0.1, 0.5, 1.0\}$
Decoder type	Linear / Nonlinear
Decoder noise (σ_x)	0.01
Number of target points (N)	$\{8, 16, 64\}$
Random seeds	$\{0, 2, 3, 5, 42\}$

After testing number of settings and five different random seeds, Table 8 shows that AE-OT method is stable for computing mode purity and achieves similar results as baseline methods for Asymmetric Two-Gaussian Mixture example. However, this method obtained worse results than k-means

and GMM in boundary localization metrics.

metric	method	mean ± std dev
Mode purity	AE-OT	0.910 ± 0.081
Mode purity	k-means	0.915 ± 0.098
Mode purity	GMM	0.915 ± 0.097
Chamfer distance	AE-OT	1.287 ± 0.396
Chamfer distance	k-means	1.156 ± 0.402
Chamfer distance	GMM	1.018 ± 0.277
Mean absolute signed distance	AE-OT	1.860 ± 0.620
Mean absolute signed distance	k-means	1.709 ± 0.728
Mean absolute signed distance	GMM	1.658 ± 0.797

Table 8: Stability summary and comparison to baselines for Gaussian Mixture toy dataset

For Two-Moons dataset ,Table 9 shows that mode purity is noticeably lower than for Gaussian Mixture example. Nevertheless, results are similar across all methods. Moreover similar results and precision are obtained for chamfer distance and mean absolute signed distance, which was not the case for the previous toy example. Standard deviation for mode purity and mean absolute signed distance results in similar values for all three methods. On the other hand, AE-OT obtains less stable results when computing Chamfer distance.

metric	method	mean ± std dev
Mode purity	AE-OT	0.769 ± 0.170
Mode purity	k-means	0.771 ± 0.188
Mode purity	GMM	0.762 ± 0.193
Chamfer distance	AE-OT	0.956 ± 0.119
Chamfer distance	k-means	0.956 ± 0.076
Chamfer distance	GMM	0.959 ± 0.091
Mean absolute signed distance	AE-OT	0.841 ± 0.023
Mean absolute signed distance	k-means	0.835 ± 0.020
Mean absolute signed distance	GMM	0.837 ± 0.020

Table 9: Stability summary and comparison to baselines for Two-Moons toy dataset

For the last example, Ring+Center, results from Table 10 shows that while standard deviation is similar for all methods it is higher than for other examples. This dataset has more complex geometry,which may be the reason for such poor results in terms of stability. What deserves attention is mode purity for AE-OT method. It shows a noticeably better result than for other methods.

metric	method	mean ± std dev
Mode purity	AE-OT	0.821 ± 0.161
Mode purity	k-means	0.706 ± 0.128
Mode purity	GMM	0.757 ± 0.149
Chamfer distance	AE-OT	1.190 ± 0.572
Chamfer distance	k-means	1.117 ± 0.590
Chamfer distance	GMM	0.914 ± 0.460
Mean absolute signed distance	AE-OT	1.630 ± 0.368
Mean absolute signed distance	k-means	1.525 ± 0.430
Mean absolute signed distance	GMM	1.308 ± 0.483

Table 10: Stability summary and comparison to baselines for Ring+Center toy dataset

Since above results show that AE-OT obtains the best results for mode purity, sensitivity of this metric to different parameters was tested. Firstly, in Figure 4, one can see that in every case, the mode purity gets worse if nuisance noise raises. Intuitively, the more noise is presented, the more difficult it is for model to find a real mode distinction. What is also worth mentioning is that Ring+Center example for higher σ_n does not present any results for the highest value of dimension $d = 784$. Because all of those experiments were conducted with the same threshold $\theta = 0.7$, in this setting model could not find pairs of latent points which survived filtering. This observation raises a question whether one can find a suitable threshold for each geometry.

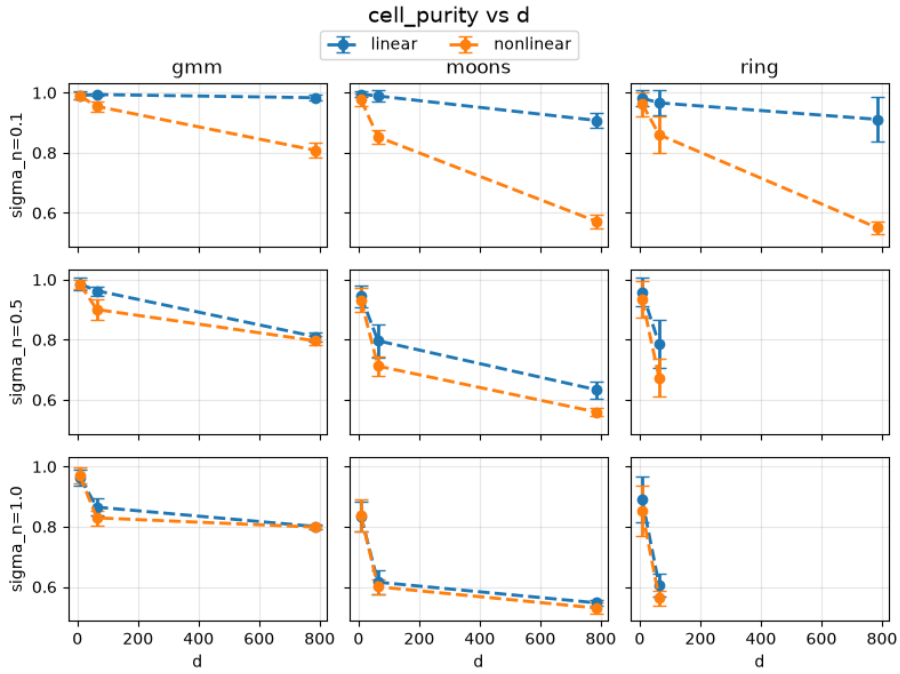

Figure 4: Plot of mode purity against dimension of ground-truth latent variable for three toy examples and different values of nuisance noise.

Secondly, the behavior of the mode purity value for different number of target points is illustrated in Figure 5. As can be seen, for each toy geometry values differs a little, but the graph remains similar for different number of target points.

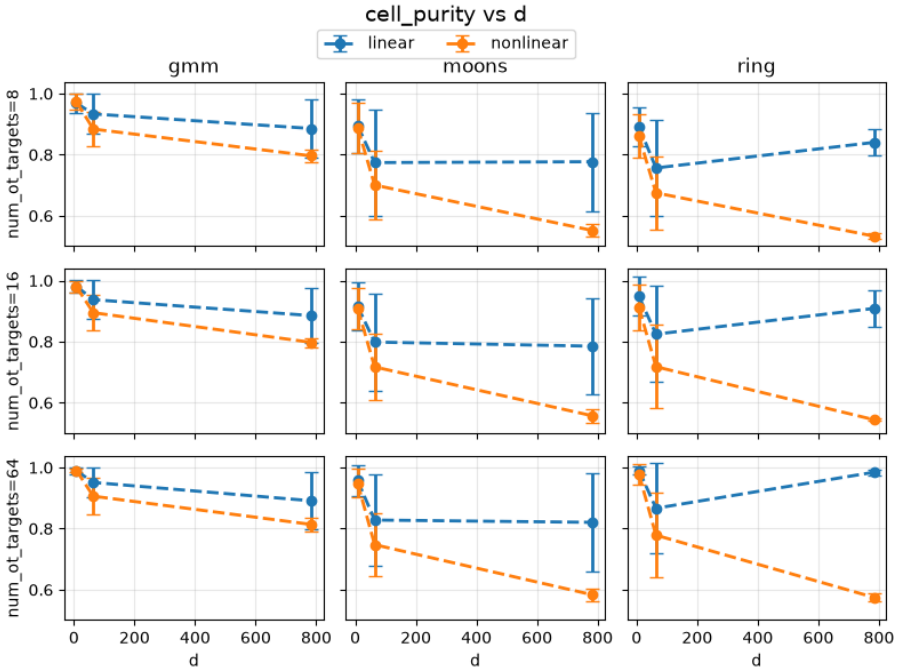

Figure 5: Plot of mode purity against dimension of ground-truth latent variable for three toy examples and different number of OT target points.

Lastly, values of mode purity against nuisance noise σ_n was tested and results can be found in Figure 6. As expected, the increase of parameter σ_n plays bigger role in decrease of mode purity. Even big increae in number of target points does not influence this trend.

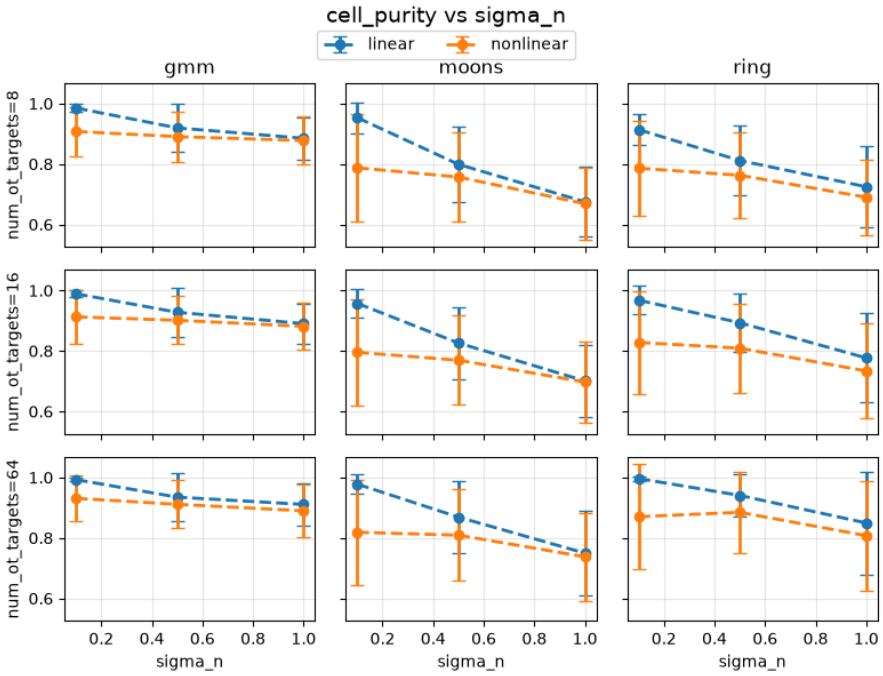

Figure 6: Plot of mode purity against nuisance noise for three toy examples and different number of OT target points.

5 Summary

From comparisons in the previous section, one can conclude that stability across all three methods is similar. However AE-OT method shows better performance for mode assigning than aligning with ground truth boundaries. One of the reasons for such a behavior can be seen in Figure 7. In this figure one can observe a "boundary" found via AE-OT method. However, as can be also seen in Figures 8 and 9 the method's main advantage is finding singularity set, thus generating new samples only for latent points within the same mode. Therefore AE-OT method seems to be a good choice for avoiding mode mixture.

Even if this method is not able to explicitly find the ground-truth boundary between different modes, it still can assign true modes better than other tested methods. However, much depends on the chosen parameters. As previous results have shown, the decoder used during data generation and value of the nuisance noise can have a big influence on the performance. Additionally, choosing a specific angle threshold for specific task might also improve the results.

Once the examples of the known geometry of the latent space were tested, it is time to test IVS data. What can be taken from these preliminary results is that multiple tests for choosing suitable parameters, especially angle threshold might improve performance. Moreover the attention might be drawn more to mode clustering than geometrical description of the AE-OT boundaries in future

testing.

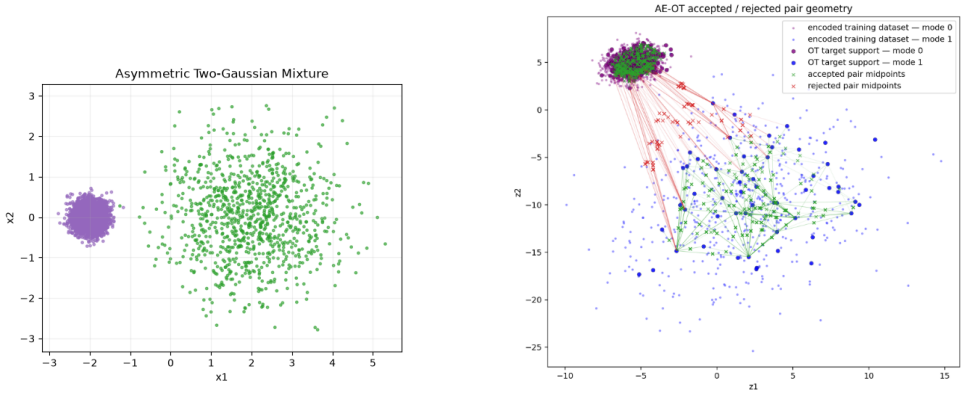

Figure 7: Asymmetric Two-Gaussian Mixture: Real latent space geometry (left) and result of used method (right)

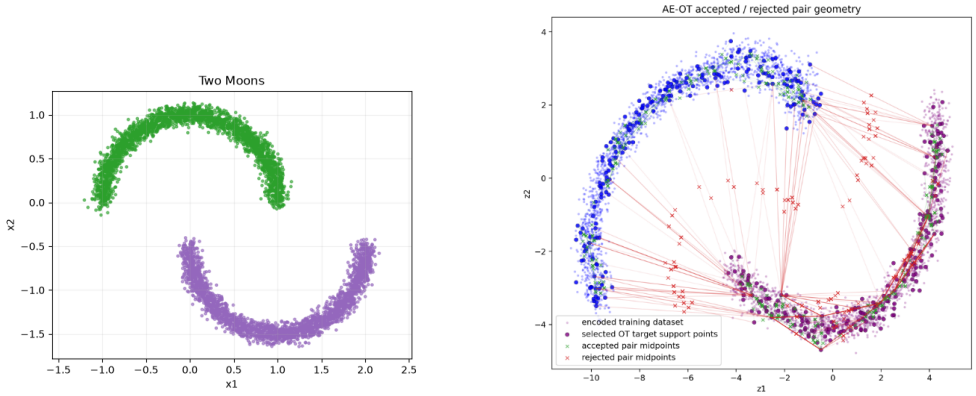

Figure 8: Two Moons: Real latent space geometry (left) and result of used method (right)

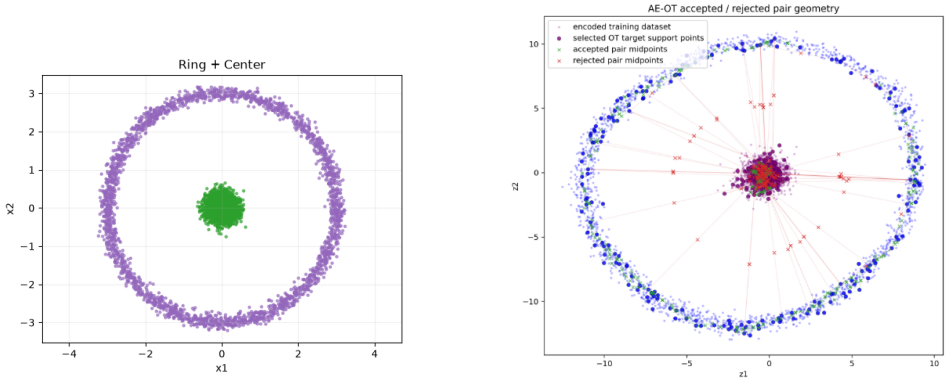

Figure 9: Two Moons: Real latent space geometry (left) and result of used method (right)

References

[1] James D. Hamilton. "Regime Switching Models". In: *The New Palgrave Dictionary of Economics*. 2016, pp. 1–7. ISBN: 978-1-349-95121-5. DOI: 10.1057/978-1-349-95121-5_2459-1.

[2] David D. Yao, Qing Zhang, and Xun Yu Zhou. "A Regime-Switching Model for European Options". In: *Stochastic Processes, Optimization, and Control Theory: Applications in Financial Engineering, Queueing Networks, and Manufacturing Systems: A Volume in Honor of Suresh Sethi*. 2006, pp. 281–300. ISBN: 978-0-387-33815-6. DOI: 10.1007/0-387-33815-2_14.

[3] Nicolas P.B. Bollen, Stephen F. Gray, and Robert E. Whaley. "Regime switching in foreign exchange rates:: Evidence from currency option prices". In: *Journal of Econometrics* 94.1 (2000), pp. 239–276. ISSN: 0304-4076. DOI: https://doi.org/10.1016/S0304-4076(99)00022-6.

[4] Jin Seo Cho and Halbert White. "Testing for Regime Switching". In: *Econometrica* 75.6 (2007), pp. 1671–1720. DOI: https://doi.org/10.1111/j.1468-0262.2007.00809.x.

[5] Zhenni Tan and Yuehua Wu. "On Regime Switching Models". In: *Mathematics* 13.7 (2025). ISSN: 2227-7390. DOI: 10.3390/math13071128.

[6] Eduardo Fernandes Montesuma, Fred Maurice Ngolè Mboula, and Antoine Souloumiac. "Recent Advances in Optimal Transport for Machine Learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47.2 (2025), pp. 1161–1180. DOI: 10.1109/TPAMI.2024.3489030.

[7] Viet Huynh and Dinh Phung. "Optimal transport for deep generative models: state of the art and research challenges". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. IJCAI International Joint Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (AAAI), 2021, pp. 4450–4457. DOI: 10.24963/ijcai.2021/607.

[8] Yogesh Balaji, Rama Chellappa, and Soheil Feizi. "Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 12934–12944. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/9719a00ed0c5709d80dfef33795dcef3-Paper.pdf.

[9] Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. "Generative Modeling through the Semi-dual Formulation of Unbalanced Optimal Transport". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 42433–42455. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/84706cdfc192cd0351daf48f379847e6-Paper-Conference.pdf.

[10] Litu Rout, Alexander Korotin, and Evgeny Burnaev. *Generative Modeling with Optimal Transport Maps*. Oct. 2021. DOI: 10.48550/arXiv.2110.02999.

[11] Jing Wang, Shuaiqiang Liu, and Cornelis Vuik. *Controllable Generation of Implied Volatility Surfaces with Variational Autoencoders*. 2025. URL: https://arxiv.org/abs/2509.01743.

[12] Dongsheng An, Yang Guo, Na Lei, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. "AE-OT: A NEW GENERATIVE MODEL BASED ON EXTENDED SEMI-DISCRETE OPTIMAL TRANSPORT". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=HkldyTNYwH.

[13] Alessio Figalli. "Regularity Properties of Optimal Maps Between Nonconvex Domains in the Plane". In: *Communications in Partial Differential Equations* 35.3 (2010), pp. 465–479. DOI: 10.1080/03605300903307673.

[14] Cedric Villani. *Optimal Transport Old and New*. 1st ed. Berlin: Springer Berlin, Heidelberg, 2009. DOI: https://doi.org/10.1007/978-3-540-71050-9.

[15] Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng David Gu. "A geometric view of optimal transportation and generative model". In: *Computer Aided Geometric Design* 68 (2019), pp. 1–21. ISSN: 0167-8396. DOI: https://doi.org/10.1016/j.cagd.2018.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0167839618301249.

[16] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* 44.4 (1991), pp. 375–417. DOI: https://doi.org/10.1002/cpa.3160440402. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.3160440402. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160440402.

[17] Cédric Villani. *Topics in Optimal Transportation Theory*. Vol. 58. Mar. 2003. ISBN: 9780821833124. DOI: 10.1090/gsm/058. URL: https://www.researchgate.net/publication/239230352_Topics_in_Optimal_Transportation_Theory.

[18] Xianfeng Gu, Feng Luo, Jian Sun, and Shing-Tung Yau. "Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge–Ampère equations". In: *Asian Journal of Mathematics* 20.2 (2026), pp. 383–398. DOI: 10.4310/AJM.2016.v20.n2.a7.

[19] Jun Kitagawa, Quentin Mérigot, and Boris Thibert. "Convergence of a Newton algorithm for semi-discrete optimal transport". In: *Journal of the European Mathematical Society* 21.9 (2019), pp. 2603–2651. DOI: 10.4171/JEMS/889.

[20] Daniel A. Klain. "The Minkowski problem for polytopes". In: *Advances in Mathematics* 185.2 (2004), pp. 270–288. ISSN: 0001-8708. DOI: https://doi.org/10.1016/j.aim.2003.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0001870803002263.

[21] Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. "A Geometric Understanding of Deep Learning". In: *Engineering* 6.3 (2020), pp. 361–374. ISSN: 2095-8099. DOI: https://doi.org/10.1016/j.eng.2019.09.010.

[22] A.D. Alexandrov. *Convex polyhedra*. Springer, 2005. URL: https://scholar.google.com/scholar_lookup?title=Convex%20Polyhedra&author=A.D.%20Alexandrov&publication_year=2005.