OPT2021: 13th Annual Workshop on Optimization for Machine Learning

Towards Modeling and Resolving Singular Parameter Spaces using Stratifolds

Pascal Mattia Esser ESSER@IN.TUM.DE
Technical University of Munich, Germany

Frank Nielsen FRANK.NIELSEN@ACM.ORG
Sony Computer Science Laboratories Inc., Japan

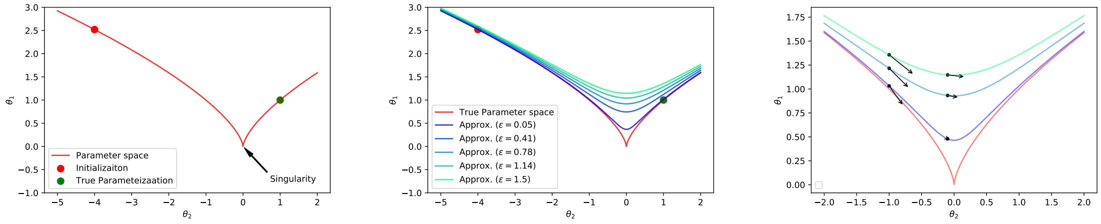
Abstract

When analyzing parametric statistical models, a useful approach consists in modeling geometrically the parameter space. However, even for very simple and commonly used hierarchical models like statistical mixtures or stochastic deep neural networks, the smoothness assumption of manifolds is violated at singular points which exhibit non-smooth neighborhoods in the parameter space. These singular models have been analyzed in the context of learning dynamics, where singularities can act as attractors on the learning trajectory and, therefore, negatively influence the convergence speed of models. We propose a general approach to circumvent the problem arising from singularities by using stratifolds, a concept from algebraic topology, to formally model singular parameter spaces. We use the property that specific stratifolds are equipped with a resolution method to construct a smooth manifold approximation of the singular space. We empirically show that using (natural) gradient descent on the smooth manifold approximation instead of the singular space allows us to avoid the attractor behavior and therefore improve the convergence speed in learning.

1. Introduction

A *parameter space* is a subspace of the Euclidean space equipped with the metric topology induced by the L_2-norm that includes all permissible parameters of a statistical model. This geometric viewpoint of the set of parameters allows us to analyze how specific topological properties given by the parameter space influence the learning dynamics. Specifically, regular parameter spaces are modeled as Fisher-Rao manifolds [14] to allow continuous and smooth update. However, even commonly used simple models like neural networks [21] violate the manifold assumption. We will refer to such points in the parameter space that are *continuous but not smooth* as *singular*. This setting results into two main problems: (1) the tangent space on the parameter space is not well defined *at the singularity* due to the lack of smoothness. Therefore, gradient-based learning algorithms in a gradient flow setting fail here. (2) several studies [4, 5, 27] have shown that *near the singularity* an attractor behavior [19] can be observed such that update steps gets smaller close to the singularity. This is due to the fact that the Hessian of the loss function becomes singular (also referred to as degenerate or higher-order saddles) when the parameters space is singular, leading to slower convergence when training the model [6, 24].

To overcome those problems we propose to model the parameter space as a specific topological space, namely Stratifolds [7, 11, 16] and then use a resolution of the space to obtain a smooth manifold approximation around the singularities.



(a) (b) (c)

Figure 1: (a) Singular parameter space in red. Initialization at the red point. True parameterization at the green point. The update has to go through the singularity. (b) Proposed smooth approximation. (c) Gradients at different points on the singular parameter space and the smooth approximations.

We illustrate the main idea on a specific example shown in Figure 1. Consider a function $f_{\theta_1,\theta_2}(x)$ such that the parameters lie on a topological space restricted to $\theta_1^2 + \theta_2^3 = 0$ as displayed in Figure 1 (a). A standard learning problem is now to start at some initialization (shown by the red dot) and try to reach a true parameterization (shown by the green dot) using a gradient decent (GD) approach. We note that at the singularity at $(\theta_1, \theta_2) = (0, 0)$ the space has a singularity, characterized by the lack of a unique tangent space. We can formally model such a space as a Stratifold by decomposing it into a set of smooth manifolds. To overcome the problem arising from the lack of smoothness we propose to instead consider a manifold approximation of the singular topological space as illustrated in Figure 1 (b), which can be formalized as $f_{\theta_1,\theta_2,\epsilon}(x)$ s.t. $\theta_1^2 + \theta_2^3 = \epsilon$. In the main part of the paper we will formalize this idea. Depending on how close the approximation is, we observe the following tradeoff: for a close approximation (in this case small ϵ) even points close to the singularity are well approximated however, as mentioned above, singularities admit an attractor behavior, which is stronger the closer the approximation is to the original model. We can see this illustrated in Figure 1 (c). Away from the singularity differences in gradients are very similar while close to the singularity the gradients on the approximation are large, allowing for the effective application of GD, while they degenerate near the singularity.

While there are a few works that consider Stratifolds or similar constructions in the context of learning theory [8, 21] and focus on local manifold structures [17], to the best of our knowledge this is the first approach to consider Stratifolds and their resolution in the context of learning dynamics. Furthermore there are several approaches on the analysis of specific singular models where the most general one, based on a topological viewpoint, is by Watanabe [24, 25, 26]. While [24] relies in principle on the same resolution theorems for algebraic varieties [12, 13] as we do, [24] focuses on information criteria, as standard approaches [1, 20] do not hold in the singular regime. The technique of [24] uses a *blowup* of the function space into higher dimensions and therefore follows a *functional perspective* where in contrast we consider a resolution that preserves the dimension of the manifold, motivated by a *topological viewpoint*.

The remainder of the paper is structured as follows. We start with Section 2.1 and Section 2.2 to introduce the construction and resolution of Stratifolds which are then applied to a simple toy model in Section 2.3. From there we illustrate that the above presented idea is relevant to the machine learning community by showing how using the resolution avoids the attractor behavior and improves the learning convergence rate.

2. Introduction to Stratifolds

Equipped with the above-given intuition, we can now define the central concept for this paper: *Stratifolds*. To do so, we follow [7, 11, 16] but restrict ourselves to outlining the main ideas. We provide further definitions and details in Appendix B and C.

Let $\boldsymbol{C} \subset C^0(\mathscr{S})$ be a locally detectable subalgebra then we can describe the decomposition for a differential space $(\mathscr{S}, \boldsymbol{C})$ over the subspace $\mathscr{S}^i := \{x \in \mathscr{S} | \; \dim(T_x\mathscr{S}) = i\}$ and the disjoint union $\mathscr{S} = \bigsqcup_i \mathscr{S}^i$. Furthermore we introduce the following terminology: Let $\mathscr{S}^i$ be the i-stratum of $\mathscr{S}$ and let $\bigcup_{i \leq r} \mathscr{S}^i =: \Sigma^r$ be the r-skeleton of $\mathscr{S}$. Now a n-dimensional Stratifold $\mathscr{S}$ is a topological space $\mathscr{S}$ together with a class of distinguished continuous (smooth) functions $\mathscr{S} \to \mathbb{R}$. This generalizes smooth manifolds $\mathcal{M}$ where the class of considered functions are in $\boldsymbol{C}^\infty$. An n-dimensional Stratifold is a smooth manifold iff $\mathscr{S}^i = \emptyset, \; \forall \; i < n$. For the Stratifold, this class of smooth functions offers the decomposition. A natural example of spaces with singularities occur in algebraic geometry as algebraic varieties, i.e., zero sets of a family of polynomials.

2.1. Inductive Construction of a Stratifold

Given the concept of Stratifolds, an immediate question that comes up is *given a parameter space, how can we construct a Stratifold that describes it?*

We consider an inductive construction, as defined in [16], where we start with the lowest dimensional stratum and then by induction iteratively glue higher dimensional strata on it until reaching the top stratum. Start by considering a n dimensional Stratifold $(\mathscr{S}, \boldsymbol{C})$ and a smooth manifold $\mathcal{W}$ together with a boundary with a collar $\boldsymbol{c} : \partial\mathcal{W} \times [0, \epsilon) \to \mathcal{W}$ and furthermore assume $k > n$. In addition we define the morphism $f : \partial\mathcal{W} \to \mathscr{S}$ to be the *attaching map*. Now we define $\mathscr{S}'$ by gluing $\mathcal{W}$ to $\mathscr{S}$ with f by: $\mathscr{S}' := \mathcal{W} \cup_f \mathscr{S}$. On this space we consider the algebra $\boldsymbol{C}'$ consisting of those functions $g : \mathscr{S}' \to \mathbb{R}$ who's restriction to $\mathscr{S}$ is in $\boldsymbol{C}$ who's restriction to the interior of $\mathcal{W}$, $\overset{\circ}{\mathcal{W}} := W - \partial\mathcal{W}$ is smooth and such that for some $\delta < \epsilon$ we have $g\mathbf{c}(x, t) = gf(x)$ for all $x \in \partial\mathcal{W}$ and $t < \delta$.

To illustrate this idea we can now apply the construction as described above on the example of a probability simplex. Consider the specific case of a simplex where the scalars of the individual points $u_0, \ldots, u_k \in \mathbb{R}^k$ are $k + 1$ standard unit vectors also known as a *Standard n-simplex*, denoted by Δ^n. An example of such a probability simplex in practice is the *multinomial distribution*: for n independent trials k possible mutually exclusive outcomes, with corresponding probabilities $p_1, \ldots, p_k \in \mathbb{R}^k | \sum_{i=0}^{k} p_i = 1, p_i \geq 0 \; \forall i \in [k]$, such that p_i lies on the simplex. We can now use the previously define inductive construction to obtain a Stratifold formulation. We have trivially for Δ^0 a 0-dimensional manifold and therefore also a 0-dimensional Stratifold: $\mathscr{S}^0 = \Delta^0 = \{(t_0) \in \mathbb{R}^1 | \sum_{i=0}^{0} t_i = 1\}$ which we use as a starting point and from there glue higher dimensional simplices on it. In addition the algebra $\boldsymbol{C}$ is all constant functions on $\mathscr{S}^0 \in \mathbb{R}^1$. Therefore in the spirit of induction, we can consider the Δ^0 and the gluing of Δ^1 on it as a kind of base case. Now consider $\partial\mathscr{S} = \mathscr{S} - \overset{\circ}{\mathscr{S}}$, where we define $\partial\mathscr{S} = \Delta^0$ and $\mathscr{S} = \Delta^1 = \{(t_0, t_1) \in \mathbb{R}^2 | \sum_{i=0}^{1} t_i = 1\}$ and $\overset{\circ}{\mathscr{S}} = \Delta^1 - \Delta^0 = \{(t_0, t_1) \in \mathbb{R}^2 | \sum_{i=0}^{1} t_i < 1\}$ where $t_i \geq 0 \forall i$. Then the algebra $\boldsymbol{C}$ is all functions that are smooth on $\overset{\circ}{\mathscr{S}}$ and constant on $\partial\mathscr{S}$. We can define the space by starting with initial Stratifold $(\mathscr{S}, \boldsymbol{C}) := \Delta^0$, considering a $k > n$ dimensional manifolds with boundary, in this case $\mathcal{W} := \Delta^1$. Finally the attaching map $f : \partial\mathcal{W} \to \mathscr{S}$ is an identity map as $\Delta^0 = \partial\mathcal{W}$ (the lower dimensional simplex lies on the boundary of the higher

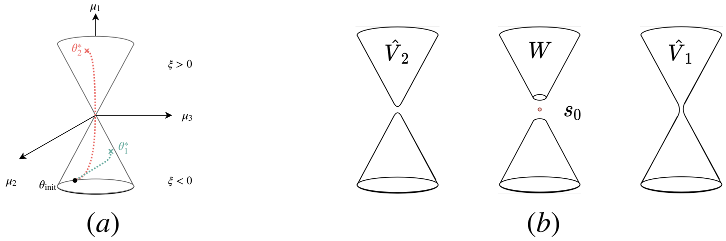


Figure 2: (a) Parameter space of the double cone model, together with examples of learning dynamics. Let $\boldsymbol{\theta}_{\text{init}}$ be the parameter initialization, then $\boldsymbol{\theta}_1^*$ is an example for a true parameterization that lies on the same cone, where as $\boldsymbol{\theta}_2^*$ is an example for a true parameterization that lies on the second cone. (b) *(left)* two sheet hyperboloid. *(middle)* Stratification. *(right)* one sheet hyperboloid.

dimensional one). From there, the inductive step would be to consider the newly constructed Δ^1 as a starting point and attach Δ^2 onto it.

2.2. Resolution of Stratifolds

While the previous section provided us with a framework for modeling parameter spaces that are not smooth manifolds, the descriptive element does not improve the model performance. Therefore we are interested in a *resolution* of the Stratifold that provides a smooth manifold approximation leaving the dimension of the top stratum untouched. For this paper we consider the setting of isolated singularities. [11] extends this concept but the details become quite technical and we therefore refer to this analysis as future work. To do so we start with defining the setup more formally [11].

Let $\mathscr{S}$ be an m-dimensional Stratifold. A resolution of $\mathscr{S}$ is a map $p : \hat{\mathscr{S}} \to \mathscr{S}$ such that: (1) $\hat{\mathscr{S}}$ is a smooth manifold, (2) p is a proper morphism, (3) the restriction of p to $p^{-1}(\mathscr{S}^m)$ is a diffeomorphism onto $\mathscr{S}^m$, (4) $p^{-1}(\mathscr{S}^m)$ is dense in $\hat{\mathscr{S}}$.

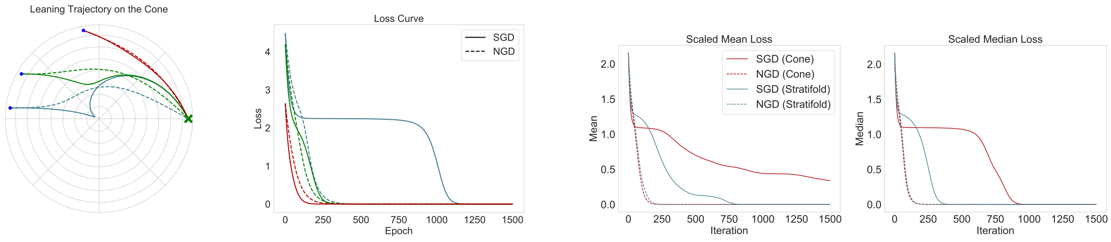
If $\mathscr{S}$ is an algebraic variety Hironaka [12, 13] shows that there is a resolution of the singularity using algebraic geometry. This link is especially interesting considering that Watanabe [24] is built on the same theorems but in the context of functional blowup and not in the context of topological resolutions. In the following we will illustrate the construction and resolution on a specific example.

2.3. The Cone Model as a Stratifold - Inductive Construction and Resolution

For our analysis we look at a simple model that is used to show the influence of singularities on learning trajectories and convergence speed [3–5, 27]. Consider a random variable $\boldsymbol{x} \in \mathbb{R}^3$, subject to a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{1}$. The hypothesis class over all such function is given by $\mathcal{H}_{\boldsymbol{\theta}} := \left\{ h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^3} \exp\left(-\frac{1}{2} \|\boldsymbol{x} - \boldsymbol{\mu}\|^2 \right) \middle| \, \boldsymbol{\mu} \in \mathbb{R}^3, \ \mu_1^2 + \mu_2^2 - \mu_3^2 = 0 \right\}$ where $\boldsymbol{\mu}$ is restricted to be on the surface of a double cone.

As the cone lies in $\mathbb{R}^3$ we can reparameterize the cone as: $\mu_1 = \xi, \ \mu_2 = \xi \cos\theta, \ \mu_3 = \xi \sin\theta$. This results in two cones, one for $\xi \geq 0$ and one for $\xi \leq 0$, connected at the apex $\xi = 0$, which we refer to as the singularity of the model. This means that the parameter surface is given by $\mathscr{S} = \left\{ (\mu_1, \mu_2, \mu_3) \in \mathbb{R}^3 \mid \mu_1^2 + \mu_2^2 = \mu_3^2 \right\}$. We can see this model illustrated in Figure 2 (a).

We are now interested in the applying the steps as described in Section 2.2 to define the model as a Stratifold and then also derive a resolution of it.



(*a*) (*b*)

Figure 3: (a) *(left)* top view onto the cone model. Displayed are the learning trajectories for three different initializations (blue dots) for the same target value (green cross). Solid lines: learning trajectories for the GD on the original cone model. Dotted lines: learning trajectories GD on the Stratifold resolution. *(right)* Loss functions corresponding to the plotted learning trajectories. (b) Loss for NGD and SGD on the Cone Model. Mean *(left)* and median *(right)* loss over different initializations.

Following directly from the definition it is easy to see that we can not define a unique tangent plane on the singularity at $\boldsymbol{\mu} = \{0, 0, 0\}$. Furthermore we also see that the point is closed. We define the zero set as $\Sigma := \{s_i\}_{i=0} = s_0 = (0, 0, 0) \quad \Rightarrow \quad \dim(\Sigma) = 0$. This gives us an algebraic variety $V \subset \mathbb{R}^3$ with one isolated singularity and in line with the earlier definitions the distance function becomes $\rho_i(x) := ||x - s_0||^2$, which gives a distance to the origin of the coordinate system. The ε enclosing ball is around the origin and we can see $V_{\varepsilon_i}(s_i) := V \cap D_{\varepsilon_i}(s_i)$ as the part of the double cone that is within the ε ball. We can see this illustrated in Figure 2 (b) (middle). Seeing D_{ε_i} to be the enclosing ball we can see $\mathring{D}_{\varepsilon_i}$ as the ε ball without the border. As we only have one singularity we can define W now as: $W := V - \mathring{D}_{\varepsilon_i}(s_0) \cup_{\mathrm{id}} \partial V_{\varepsilon_i}(s_i) \times [0, \varepsilon_i]$. Here W builds now the double cone structure away from the singularity where the enclosing ball builds the corners close to the singularity.

This gives a final Stratifold as one 0-dimensional manifold of the singularity, $\mathscr{S}^0$, and the manifold of the cone, with the corner build by the enclosing ball around the singularity $\mathscr{S}^0$ as: $\mathscr{S}^0 = s_0$, $\mathscr{S}^2 = W$. And the complete Stratifold as $\mathscr{S} = \bigsqcup_i \mathscr{S}^i = \mathscr{S}^0 \sqcup \mathscr{S}^2 = s_0 \sqcup W$. For the cone model define $h(\mu_1, \mu_2, \mu_3) = \mu_1^2 + \mu_2^2 - \mu_3^2 = 0$ and following Section 2.2 we have $\mathscr{V} = h^{-1}(0)$ with a not optimal resolution of $\widehat{\mathscr{V}_2} = h^{-1}(\varepsilon)$ which gives us a hyperboloid of two sheets (Figure 2 (b), left) and an optimal resolution of $\widehat{\mathscr{V}_1} = h^{-1}(-\varepsilon)$ which gives us a hyperboloid of one sheet as illustrated in Figure 2 (b) (right).

3. Experiments

Finally we can now analyze in the next section how such a resolution can be used in the context of machine learning — more specifically to improve convergence speed. Using the above concepts as a starting point we can conduct experiments that extend the analysis of the cone model in [27] to gradient decent on the previously derived smooth manifold approximation. We focus our analysis on the convergence speed and form of the learning trajectories. To do so we recall the two problems that can arise in the context of singularities: (1) slower convergence as result from attractor behaviour *near the singularity*. (2) the tangent space *at the singularity* is undefined.

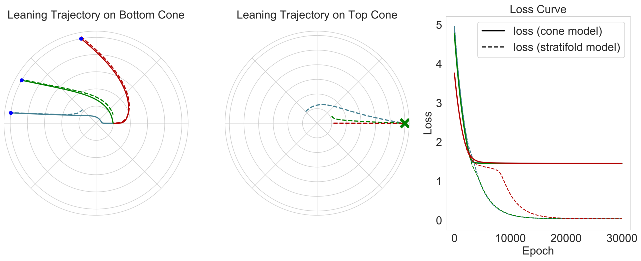


Figure 4: Comparing loss and learning trajectory where the initialization and true parameterization are on different cones. The parameters have to go through the singularity as $\xi = 0$. *(left)* cone one $\xi > 0$. *(middle)* cone two $\xi < 0$. *(right)* loss function.

To analyze the behavior near and at the singularity, we consider *gradient decent* over different initialization. Formally let $\boldsymbol{\theta}^{(t)}$ be the model parameters at time-step t, then the update rule for GD is defined as $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \vartheta \nabla_{\boldsymbol{\theta}^{(t)}} \mathcal{L}_c(x; \boldsymbol{\theta}^{(t)})$ where ϑ is the learning rate and $\nabla_{\boldsymbol{\theta}^{(t)}} \mathcal{L}_c(x; \boldsymbol{\theta}^{(t)})$ the gradient of the negative log-likelihood $\mathcal{L}(x; \boldsymbol{\theta})$ with respect to the parameters $\boldsymbol{\theta}^{(t)}$. Furthermore the NGD update rule is given by $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \vartheta \mathcal{I}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(x; \xi, \theta, \epsilon)$ where $\mathcal{I}$ is the Fisher information matrix (FIM). For readability we omit the indication of the time-step in the following.

Applying this to the double cone model we can write $\mathcal{L}_c(x; \boldsymbol{\mu}) := \frac{1}{2}\big(\sum_{i=1}^{3} (x_i - \mu_i)^2\big)$ with the following reparameterization $\boldsymbol{\mu} := (\mu_1, \mu_2, \mu_3) := (\xi, \sqrt{\xi^2 + \epsilon}\cos\theta, \sqrt{\xi^2 + \epsilon}\sin\theta)$. From there it is easy to obtain the learning dynamics under GD with $\overline{x}_i = \mathbb{E}[x_i]$ using

$$\nabla_x \mathcal{L}_c = \begin{bmatrix} \overline{x}_1 + \overline{x}_2 \cos\theta^{(t)} + \overline{x}_3 \sin\theta^{(t)} - 2\xi^{(t)} \\ -\xi^{(t)}(\overline{x}_2 \sin\theta^{(t)} + \overline{x}_3 \cos\theta^{(t)}) \end{bmatrix}, \quad \mathcal{I}_{cone} = \begin{bmatrix} 2 & 0 \\ 0 & \xi^2 \end{bmatrix}.$$

Similarly for NGD we obtain the learning dynamics for the smooth manifold approximation of the cone model by first defining the reparameterisation $(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3) = (\xi, \sqrt{\xi^2 + \epsilon}\cos\theta, \sqrt{\xi^2 + \epsilon}\sin\theta)$ so that we obtain the following gradient and FIM:

$$\nabla_x \mathcal{L}_{hyp} = \begin{bmatrix} -\overline{x}_1 + 2\xi - \xi\overline{x}_2 \cos\theta \frac{1}{\sqrt{\xi^2+\epsilon}} - \xi\overline{x}_3 \sin\theta \frac{1}{\sqrt{\xi^2+\epsilon}} \\ \sqrt{\xi^2 + \epsilon}\overline{x}_2 \sin\theta - \sqrt{\xi^2 + \epsilon}\overline{x}_3 \cos\theta \end{bmatrix}, \quad \mathcal{I}_{hyp} = \begin{bmatrix} \frac{\epsilon + 2\xi^2}{\epsilon + \xi^2} & 0 \\ 0 & \epsilon + \xi^2 \end{bmatrix}.$$

Now using the above update equations be obtain the learning trajectories and loss curves as illustrated in Figure 3. As we are interested in the comparison of GD on the initial model and the hyperboloid we note that for parameters that are initialized close to the true parameterization, the behavior for both settings is very similar as it does not come close to the singularity. In contrast if we focus on the learning dynamic plotted in blue, GD provides an update that gets trapped at the singularity and results in a plateau behaviour. On the other hand if we look at the same initialization on the hyperboloid, the influence of the singularity is less pronounced; the model does not get trapped and converges faster.

Interestingly if we consider *NGD*, the convergence speed on both the cone model and the hyperboloid is very similar. This first might seem to be surprising however previous results [27] showed that if we account for the FIM, the singularity has less influence on the learning trajectory.

This immediately brings up the question "*Why are we interested in the resolution of the model if the NGD seems to give a fast convergence?*" Answering this brings us to the the second question posed in the introduction: *what is the behavior of the learning dynamics at the singularity?* We note that the Fisher information may potentially be infinite when the integral diverges: For example, this happens for the parametric family of uniform distributions $\{U[a, a+1] \ : \ a \in \mathbb{R}\}$. In that case, Amari [2] proposed to used Finsler geometry instead of Riemannian geometry by defining the information using Hellinger divergence between infinitesimally close distributions. Furthermore we observe that the FIM degenerates at the singularity, therefore problems arise when we consider a setting where we have to go *through* the singularity (as illustrated in Figure 4). In the case of the cone model this happens when we initialize the parameters on one cone while the true parameters lie on the second cone. We indeed observe that the parameters get stuck at the singularity and we can not reach the true parameterization on the second cone. In contrast we see that for the hyperboloid, while we see a slowdown around $\xi = 0$, GD reaches the true parameterization. Therefore even in a NGD setting the consideration of the resolution can play a valuable role. Finally on an application level we note that the use of NGD also comes with limitations [18] like the requirement to computation of the inverse of the FIM which can become computationally intensive for complex models.

4. Conclusion and Future Work

We take a first step towards constructing a general framework for modeling geometrically singular parameter spaces. By taking advantage of the nature of Stratifolds, we are able to describe parameter spaces that do not fulfill smoothness assumptions and use the resolution of this description of the parameter space to obtain a smooth manifold approximation of the singular space. Experimentally we show that using first-order methods (SGD) on the smooth manifold approximation decreases the effect of the singularity and therefore speeds up learning convergence. For natural gradient descent we additionally find that the Fisher information metric can no longer degenerate during training as the resolution offers an approximation at previously singular points.

As we usually consider extensions to vanilla first-order gradient-based optimization like [10, 15, 22] or optimizations that take the parameter surface into account [9, 23] as a step into considering the parameter space more explicitly, those approaches do not require explicit knowledge of the topological properties of the parameter space. While first steps are taken in this direction [17] a main focus of future research will be to develop a better understanding of the topology of the parameter spaces of commonly used models and formally characterize their singular points and their resolution. While the presented toy examples would allow for simpler geometric constructions we strongly expect that more complex structures such as Stratifolds will be necessary to describe more complex models.

In line with the this direction, we can recall that for this paper, we focused on the very simple case of isolated singularities. It is to be expected that studying the topology of parameter spaces of complex models will also reveal more complex singularities. While [11] provides first theoretical results on the resolution of more complex singularities, it is open how applicable those results are in practice. Nevertheless, we conclude by emphasizing that the goal of this paper is not to beat state-of-the-art models but rather to gain a better understanding of how the structure of the parameter space influences the learning behavior, and propose the idea of resolving singular parameter spaces into smooth manifolds.

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974.

[2] Shun-ichi Amari. Finsler geometry of non-regular statistical models. *RIMS Kokyuroku (in Japanese), Non-Regular Statistical Estimation, Ed. M. Akahira*, 538:81–95, 1984.

[3] Shun-ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Geometrical singularities in the neuromanifold of multilayer perceptrons. In *International Conference on Neural Information Processing Systems*, 2001.

[4] Shun-ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Singularities affect dynamics of learning in neuromanifolds. *Neural computation*, 2006.

[5] Shun-ichi Amari, Tomoko Ozeki, Ryo Karakida, Yuki Yoshida, and Masato Okada. Dynamics of Learning in MLP: Natural Gradient and Singularity Revisited. *Neural Computation*, 2018.

[6] Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *Computing Research Repository*, 2016.

[7] Toshiki Aoki and Katsuhiko Kuribayashi. On the category of stratifolds. 2016. URL https://arxiv.org/abs/1605.04142.

[8] Jean-Daniel Boissonnat and Mathijs Wintraecken. The Topological Correctness of PL-Approximations of Isomanifolds. In *International Symposium on Computational Geometry*, 2020.

[9] Silvère Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 2013.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

[11] Anna Grinberg. Resolution of Stratifolds and Connection to Mather's Abstract Pre-Stratified Spaces. 2003. URL http://www.ub.uni-heidelberg.de/archiv/3127.

[12] Heisuke Hironaka. Resolution of Singularities of an Algebraic Variety Over a Field of Characteristic Zero: I. *Annals of Mathematics*, 1964.

[13] Heisuke Hironaka. Resolution of Singularities of an Algebraic Variety Over a Field of Characteristic Zero: II. *Annals of Mathematics*, 1964.

[14] Harold Hotelling. Spaces of statistical parameters. *Bulletin of the American Mathematical Society*, 1930. First mention hyperbolic geometry for Fisher-Rao metric of location-scale family.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[16] Matthias Kreck. Differential algebraic topology: From stratifolds to exotic spheres. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 2012.

[17] Wu Lin, Frank Nielsen, Mohammad Emtiyaz Khan, and Mark Schmidt. Tractable structured natural-gradient descent using local parameterizations. In *International Conference on Machine Learning*, 2021.

[18] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 2020.

[19] John Milnor. On the concept of attractor. *Communications in Mathematical Physics*, 1985.

[20] Gideon Schwarz. Estimating the dimension of a model. 1978.

[21] Ke Sun and Frank Nielsen. Lightlike Neuromanifolds, Occam's Razor and Deep Learning, 2021. URL https://arxiv.org/abs/1905.11027.

[22] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[23] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I. Jordan. Averaging stochastic gradient descent on riemannian manifolds. In *Conference On Learning Theory*, 2018.

[24] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. 2009.

[25] Sumio Watanabe. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 2010.

[26] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 2013.

[27] Haikun Wei, Jun Zhang, Florent Cousseau, Tomoko Ozeki, and Shun ichi Amari. Dynamics of learning near singularities in layered networks. *Neural Computation*, 2008.

Appendix A. Notation and Illustration of Blowup

In this section we first give some notes on the general notation as well as some additional illustrations for conceps used in the main paper.

A.1. Notation

We denote algebras by $\boldsymbol{C}$, differential spaces as $(X, \boldsymbol{C})$ and tangent spaces as T_xX with the differential map $df_x : T_xX \to T_{f(x)}X'$. Manifolds are given by $\mathcal{M}, \mathcal{N}$ and Stratifolds by $\mathscr{S}$. For a topological space $\mathcal{S}$ we denote the interior by $\overset{\circ}{\mathcal{S}}$ and the boundary as $\partial\mathcal{S}$. The germ of a function at a point describes how the function behaves very close to the point. For a point $x \in X$, we consider the germs of functions at x, $\boldsymbol{C}_x$. For readability, we overload the notion of *resolution* as the function mapping from a singular space onto the smooth manifold as well as the manifold resulting from the resolution itself.

A.2. Additional Illustrations

- In Section 1 we discussed the comparison between the resolution as used in [24] and this paper. We further illustrate this in Figure 5.
- A standard n-simplex is the subset of $\mathbb{R}^{n+1}$ given by:

  $$\Delta^n = \left\{ (t_0, \dots, t_n) \in \mathbb{R}^{n+1} \middle| \sum_{i=0}^{n} t_i = 1, t_i \geq 0 \forall i \right\}$$

  For an illustration of the inductive construction see see Figure 7.
- The most natural examples of manifolds with singularities occur in algebraic geometry as algebraic varieties, i.e., zero sets of a family of polynomials. This example is mentioned in the main part and illustrated in Figure 6.

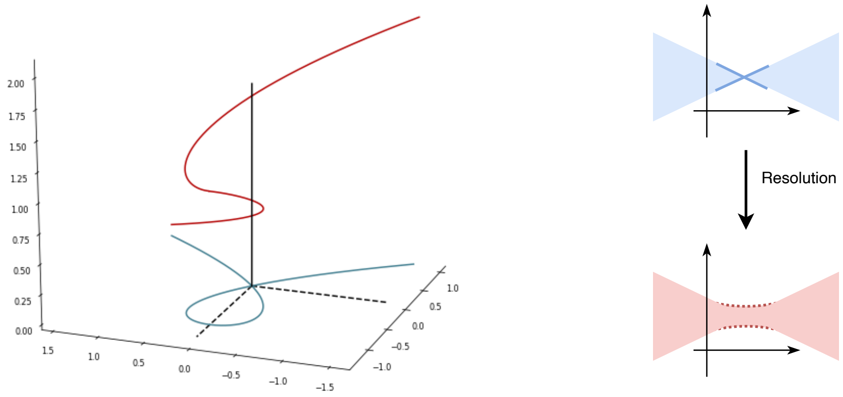


Figure 5: Illustration of the resolution of singular points. *(left)* Blowup following [24]. This illustrates a functional approach where the function is blown up into a higher dimension. *(right)* Our approach. Preserving the dimension, we propose to construct a smooth manifold approximation of the singular topological parameter space.

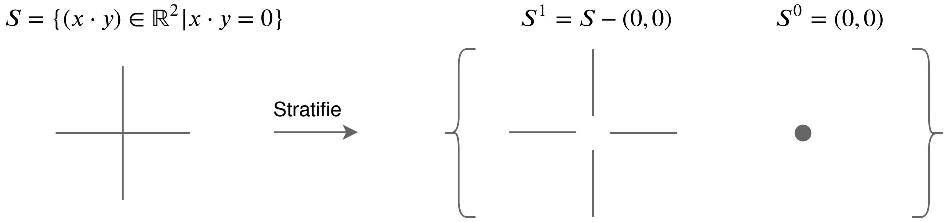


Figure 6: A simple example for a Stratifold. Considering a simple algebraic varietie $x \cdot y = 0$ with a stratification as described above.

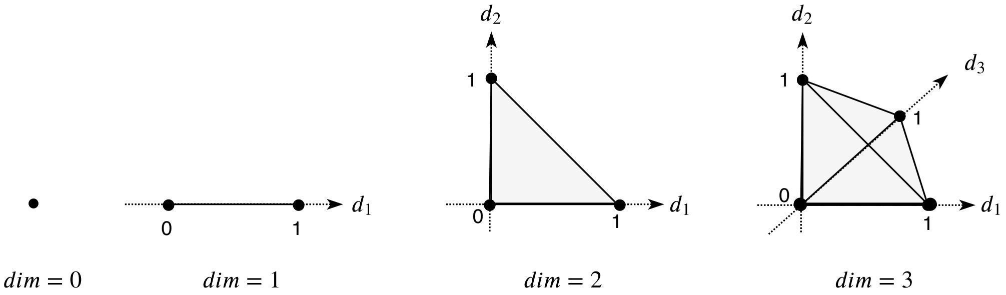


Figure 7: Probability simplex in a $\{d_1, d_2, d_3\}$ coordinate system: 0-simplex (point), 1-simplex (line segment), 2-simplex (triangle), 3-simplex (tetrahedron).

Appendix B. Additional Definitions — An Introduction to Topology

In this chapter we will recap some of the fundamental concepts from Topology, specifically related to smooth manifolds, as those concepts are essential for the later definition of Stratifolds. The specific notation and phrasing of the definitions of this chapter are taken from [11] but are not meant as a comprehensive introduction but as a reference for clarifying the terminology in the main part.

B.1. Differential Space

Definition 1 (Algebra) *Denote the set of* ***continuous functions*** *from* $X \to \mathbb{R}$ *by* $C^0(X)$

A subset $\boldsymbol{C} \subset C^0(X)$ *is called an* ***algebra*** *of continuous functions if for* $f, g \in \boldsymbol{C}$ *the sum* $f + g$ *the product* fg *and all constant functions are in* $\boldsymbol{C}$*.*

Example 1 *set of functions* $f : U \to \mathbb{R}$*, where* ***all partial derivatives*** *of all orders exist, given by* $C^\infty(U)$

Definition 2 (Locally Detectable) *Let* $\boldsymbol{C}$ *be a subalgebra of the algebra of continuous functions* $f : X \to \mathbb{R}$*.*

We say that $\boldsymbol{C}$ *is* ***locally detectable*** *if a function* $h : X \to \mathbb{R}$ *is contained in* $\boldsymbol{C}$ *if and only if for all* $x \in X$ *there is an open neighbourhood* U *of* x *and* $g \in \boldsymbol{C}$ *such that* $h|_U = g|_U$*.*

Definition 3 (Differential Space) *A* ***differential space*** *is a pair* $(\boldsymbol{X}, \boldsymbol{C})$*, where* $\boldsymbol{X}$ *is a topological space and* $\boldsymbol{C} \subset C^0(X)$ *is a locally detectable subalgebra of the algebra of continuous functions satisfying the condition:*

For all $f_1, ..., f_k \in \boldsymbol{C}$ *and smooth functions* $g : \mathbb{R}^k \to \mathbb{R}$*, the function*

$$x \mapsto g\left(f_1(x), \ldots, f_k(x)\right)$$

is in $\boldsymbol{C}$*.*

B.2. Smooth Manifolds

Definition 4 (homeomorphism) *A function* $f : X \to Y$ *between two topological spaces is a homeomorphism if it has the following properties:*

- f *is a bijection*
- f *is continuous and*
- *the inverse function* f^{-1} *is continuous*

Example 2 *A chart of a manifold is an homeomorphism between an open subset of the manifold and an open subset of a Euclidean space.*

Example 3 $\mathbb{R}^m$ *and* $\mathbb{R}^n$ *are not homeomorphic for* $m \neq n$*. (no bijaction)*

Definition 5 (Isomorphism) *Let* $(X, \boldsymbol{C})$ *and* $(X', \boldsymbol{C}')$ *be differential spaces. A homeomorphism* $f : X \to X'$ *is called an isomorphism if for each* $g \in \boldsymbol{C}'$ *and* $h \in \boldsymbol{C}$*, we have* $gf \in \boldsymbol{C}$ *and* $hf^{-1} \in \boldsymbol{C}'$*.*

Definition 6 (smooth manifold) *A k-dimensional smooth manifold is a **differential space** $(M, \boldsymbol{C})$ where M is a Hausdorff space with a countable basis of its topology, such that for each $x \in M$ there is an open neighbourhood $U \subseteq M$, an open subset $V \subset \mathbb{R}^k$ and an **isomorphism***

$$\varphi : (V, C^\infty(V)) \to (U, \mathbf{C}(U)).$$

a k-dimensional smooth manifold is a differential space which is locally isomorphic to $\mathbb{R}^k$

Definition 7 (Germ) *The germ of a function at a point describes how the **function behaves very close to the point***

For a point $x \in X$, we consider the germs of functions at $x, \boldsymbol{C}_x$. If $f \in \boldsymbol{C}$ and $g \in \boldsymbol{C}$ are representatives of germs at x, then the sum $f + g$ and the product $f \cdot g$ represent well-defined germs denoted $[f]_x + [g]_x \in \boldsymbol{C}_x$ and $[f]_x \cdot [g]_x \in \boldsymbol{C}_x$.

Definition 8 (Derivation) *Let $(X, \boldsymbol{C})$ be a differential space. A derivation at $x \in X$ is a map from the germs of functions at x*

$$\alpha : \mathbf{C}_x \longrightarrow \mathbb{R}$$

s.t.

$$\begin{aligned} \alpha\left([f]_x + [g]_x\right) &= \alpha\left([f]_x\right) + \alpha\left([g]_x\right) \\ \alpha\left([f]_x \cdot [g]_x\right) &= \alpha\left([f]_x\right) \cdot g(x) + f(x) \cdot \alpha\left([g]_x\right) \\ \alpha\left([c]_x \cdot [f]_x\right) &= c \cdot \alpha\left([f]_x\right) \end{aligned}$$

for all $f, g \in \boldsymbol{C}$ and $[c]_x$ the germ of the constant function which maps all $y \in X$ to $c \in \mathbb{R}$.

Definition 9 (Tangent Space) *Let $(X, \boldsymbol{C})$ be a differential space and $x \in X$. The **vector space of derivations** at x is called the tangent space of X at x and denoted by $T_x X$.*

Definition 10 (Differential) *Let $f : (X, \boldsymbol{C}) \to (X', \boldsymbol{C}')$ be a morphism. Then for each $x \in X$ the differential*

$$df_x : T_x X \to T_{f(x)} X'$$

is the map which sends a derivation α to α' where α' assigns to $[g]_{f(x)} \in \boldsymbol{C}'_{x'}$ the value $\alpha([gf]_x)$.

B.3. Boundaries

Besides the very standard definitions presented above, we have to consider some concepts that are related

Let $(\mathscr{S}, \partial\mathscr{S})$ be a pair of topological spaces. We denote $\mathscr{S} - \partial\mathscr{S}$ by $\overset{\circ}{\mathscr{S}}$ and call it the **interior**. We assume that $\overset{\circ}{\mathscr{S}}$ and $\partial\mathscr{S}$ are stratifolds of dimension n and $n-1$ and that $\partial\mathscr{S}$ is a closed subspace.

Now we want to look at the map onto a neighborhood of the boundary:

[1] As the illustration might be misleading: An important note here is that at p we have to to be able to define the tangent space $T_p\mathcal{M}$. Therfore the point can not lie on the boundary of a topological space. We will further discuss this in the following, Section B.3.

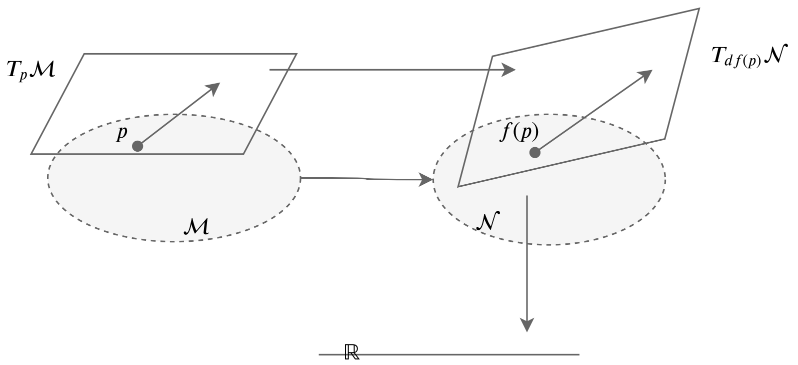


Figure 8: Illustration of a Differential map $df_x : T_xX \to T_{f(x)}X'$. Consider $\mathcal{M}$ and $\mathcal{N}$ be smooth manifolds, displayed are only a part of it [1]

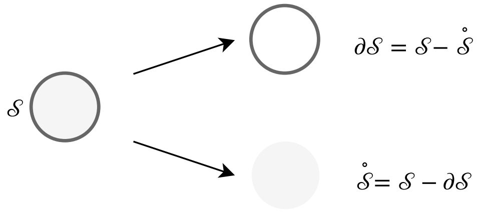

$\mathcal{S}$
$\partial\mathcal{S} = \mathcal{S}- \mathring{\mathcal{S}}$
$\mathring{\mathcal{S}}= \mathcal{S} - \partial\mathcal{S}$

Figure 9: boundary: For a topological space we consider the **interior** and **boundary** or the space, as illustrated above

Definition 11 (Collar) *Let $(\mathscr{S}, \partial\mathscr{S})$ be a pair as above. A collar is a homeomorphism*

$$\mathbf{c} : U_\epsilon \to V$$

where $\epsilon > 0, U_\epsilon := \partial\mathscr{S} \times [0, \epsilon)$ and V is an open neighbourhood of $\partial\mathscr{S}$ in $\mathscr{S}$ such that $\mathbf{c}|_{\partial\mathscr{S}\times\{0\}}$ is the identity map to $\partial\mathscr{S}$ and $\mathbf{c}|_{U_\varepsilon-(\partial\mathscr{S}\times\{0\})}$ is an isomorphism of stratifolds onto $V - \partial\mathscr{S}$

Intuitively: the boundary cross an interval open on one side, equivalently a small open neighborhood of the boundary.

Appendix C. More Formal setup of Section 2

While in Section 2 we generally refer to Stratifolds we note that there are technical more detailed to consider. Therefore we state the following definitions.

Definition 12 (k-dimensional Stratifold [16])
A k-dimensional Stratifold is a differential space $(\mathscr{S}, \boldsymbol{C})$, where S is a locally compact Hausdorff space with countable basis, and the skeleta Σ^i are closed subspaces. In addition we assume:

1. *Restriction gives a smooth structure on $\mathscr{S}^i$ and for each $x \in \mathscr{S}^i$ restriction gives an isomorphism*
$$i^* : \mathbf{C}_x \overset{\cong}{\to} C^\infty\left(\mathscr{S}^i\right)_x$$
2. *All tangent spaces have dimension $\leq k$,*
$$\dim T_x \mathscr{S} \leq k \ \ \forall\, x \in \mathscr{S}$$
3. *(Bump Function) for each $x \in \mathscr{S}$ and open neighbourhood $U \subset \mathscr{S}$ there is a non-negative function $\rho \in \boldsymbol{C}$ such that $\rho(x) \neq 0$ and supp $\rho \subseteq U$*

A natural examples of spaces with singularities occur in algebraic geometry as algebraic varieties, i.e., zero sets of a family of polynomials. To get a better intuition, consider the following example: an algebraic variety of the form $x \cdot y = 0$ as Stratifold. Consider the parameter space defined over $\{x, y\}$ s.t. $x \cdot y = 0$ we can stratify this space into a 0-dimensional manifold as $\{0, 0\}$ and four axis elements. In addition we define the following specific class of Stratifolds.

Definition 13 (c-stratifold [16]) *An n-dimensional c-stratifold $\mathscr{S}$ (a collared Stratifold) is a pair of topological spaces $(\mathscr{S}, \partial\mathscr{S})$ together with a germ of collars $[\mathbf{c}]$ where $\mathscr{S} = \overset{\circ}{\mathscr{S}} - \partial\mathscr{S}$ is an n-dimensional Stratifold and $\partial\mathscr{S}$ is an $(n-1)$-dimensional Stratifold, which is a closed subspace of $\mathscr{S}$. We call $\partial\mathscr{S}$ the boundary of $\mathscr{S}$.*

A smooth map from $\mathscr{S}$ to a smooth manifold $\mathcal{M}$ is a continuous function f who's restriction to $\overset{\circ}{\mathscr{S}}$ and to $\partial\mathscr{S}$ is smooth and which commutes with an appropriate representative of the germ of collars, i.e., there is a $\gamma > 0$ such that $f\mathbf{c}(x, t) = f(x)$ for all $x \in \partial\mathscr{S}$ and $t < \delta$.

Those refined definitions are not central for the simple example considered in the main paper but will play a central role when transferring the idea to more complex models.

Remark 14 (Uniqueness of the decomposition) *We can easily construct examples for the same stratification but different Stratifolds as well as one Space different Stratifolds.*

For the main paper we consider the simple case of an isolated singularly as defined below. Again for more complex models an extension to not singular singularities will be necessary. [11] provides a theoretical framework for such cases but we formalize the isolated case below.

Definition 15 (Isolated Singularity) *Let $\{x\}_{i\in\mathbb{N}}$ be a countable set of points and let $g : \partial\mathcal{M} \to \{x\}_{i\in\mathbb{N}}$ be a a proper map from an m-dimensional smooth manifold $\mathcal{M}$ onto a set of a 0-dimensional manifold. Writing it as a Stratifold gives us*

$$\mathscr{S} = \mathcal{M} \cup_g \{x_i\}_i .$$

A m-dimensional Stratifold $\mathscr{S}$ is said to have isolated singularity iff

$$\mathscr{S}^i = \emptyset \quad \forall i \in \{1, \cdots, m-1\}$$

C.1. Resolution of Algebraic Varieties with Isolates Singularities [11]

Consider an algebraic variety $V \subset \mathbb{R}^n$ with isolated singularities $s_i \in \Sigma$ where Σ is the singular set, all points $\{s_i\}_i$ are 0-dimensional manifold. In the case where s_i is open in V we do not need to resolve the point.

Therefore consider the other option. Let s_i be closed in V for the following. From there we define a distance function $\rho_i(x) := ||x - s_i||^2$ on $\mathbb{R}^n$ and let $D_{\varepsilon_i}(s_i)$ be the enclosing ball in $\mathbb{R}^n$ with radius ε_i around s_i. Now the idea is that we define an enclosing ball around the singularity and define a union of the initial algebraic variety and the enclosing ball. We can define this as

$$\partial V_{\varepsilon_i}(s_i) := V_{\varepsilon_i}(s_i) \cap \partial D_{\varepsilon_i}(s_i)$$

such that the restriction $\rho_i|_{V_{\varepsilon_i}-\{s_i\}}$ has no critical value. With that we can define a continuous map

$$\overline{h} : \partial V_{\varepsilon_i}(s_i) \times [0, \varepsilon_i] \to V_{\varepsilon_i}$$

this gives us a c-manifold W with obvious collar over the enclosing ball:

$$W := V - \left(\bigsqcup_i \mathring{D}_{\varepsilon_i}(s_i) \right) \cup_{\text{id}} \partial V_{\varepsilon_i}(s_i) \times [0, \varepsilon_i]$$

finally the map

$$f = \text{id} \cup \overline{h} : W \to \mathcal{V}$$

gives $\mathcal{V}$ the structure of a Stratifold with isolated singularities

C.2. Resolution of Hypersurfaces [11]

Again we assume an algebraic variety, a polynomial $p : \mathbb{R}^{n+1} \to \mathbb{R}$ with singularity $\{s_i\}_i$

$$s_i \in V := p^{-1}(0).$$

We are interested in the link to the singularity $\partial V_{\varepsilon_i}(s_i)$. Choose a $\delta > 0$ such that for all c with $|c| \leq \delta$ are regular values of p and pick c such that

$$p^{-1}(c) \neq \emptyset$$

then $p^{-1}(c)$ is a smooth manifold. In the case of a complex polynomial $p : \mathbb{C}^{n+1} \to \mathbb{C}(n > 0)$, every deformation $p^{-1}(c)$ gives an optima resolution given $||c||$ is small enough [11].

Remark 16 *In the above definitions and also the resolution considered afterward, we only consider a specific class of Stratifolds. To be specific, we have to distinguish between p-stratifolds and c-stratifolds as defined in [11]. For a rigorous definition of the Stratifolds spaces and the resolution, those distinctions are technically important as they define exact restrictions on allowed mappings and the properties of the space. In particular, Stratifolds constructed inductively by attaching manifolds together using the data: germs of collars and attaching maps, are called parameterized Stratifolds or p-stratifolds [16]. For this paper and to illustrate the main ideas we omit those distinctions there but note that revisiting them will become important when modeling more complex models.*