A geometric study on optimal transport problem and stochastic machine learning

Daiji Tsutsui

Osaka University

January 2023

Contents

List of publications related to the thesis

1. D. Tsutsui. Center manifold analysis of plateau phenomena caused by degeneration of three-Layer perceptron. *Neural Computation*, **32**(4), 683–710, 2020.

2. Y. Sato, D. Tsutsui, A. Fujiwara. Noise-induced degeneration in online learning. *Physica D: Nonlinear Phenomena*, **430**, 133095, 2022.

3. D. Tsutsui. Optimal transport problems regularized by generic convex functions: a geometric and algorithmic approach. *Information Geometry*, **5**(1), 97–130, 2022.

Acknowledgments

I would like to express my gratitude to Professor Akio Fujiwara for his continuous support and numerous helpful comments.

I am deeply grateful to Professor Yuzuru Sato, who provides many suggestive comments from the viewpoint of random dynamical systems theory. I learned many important methodologies and ideas from his discussions.

I would also like to thank my dear friends, in particular Ryosei Imahashi, for their warm encouragements.

Chapter 1

Introduction

Information geometry originated in an attempt to understand the mechanisms of statistical inference from the standpoint of differential geometry. An early conception was provided by Rao in 1945 [28]. He discussed using the Fisher information matrix as a Riemannian metric on a statistical model, a family of probability distributions with finite parameters. This metric, called the Fisher metric, is shown to be a unique Riemannian metric that is invariant with respect to sufficient statistics by Chentsov's theorem [34]. Efron introduced a curvature of a statistical model [16]. It is a geometric expression of how "curved" the model is, compared to the exponential family, which is the most basic model in statistical estimation. This idea was later organized as an affine connection that is known nowadays as the e-connection. On the other hand, the natural affineness of the space of probabilities defines another affine connection, called the m-connection. On the probability simplex $\mathcal{S}$, the e-connection ∇^e and m-connection ∇^m are mutually dual with respect to the Fisher metric g_F, and they are both flat, where the probability simplex is a finite dimensional manifold consisting of non-singular probability distributions on finite events. Such a tuple $(\mathcal{S}, g_F, \nabla^e, \nabla^m)$ satisfying the duality and flatness is referred to as a dually flat manifold. A dually flat manifold admits an extension of the Pythagorean theorem. This is a remarkable result in information geometry and has been widely applied to statistical testing and estimation, information processing, and convex programming.

Optimal transport theory also studies a geometry on a space of probability distributions. It originates from solving the problem of transporting given resources to specified destinations at minimum cost. When a distance function is used as the cost for transporting an individual resource, the minimum transportation cost becomes a distance function between the resource distribution and target distribution, which is called the Wasserstein distance. The topology determined by the Wasserstein distance, unlike information geometry, reflects geometric properties such as the curvature of the underlying space [36, 35]. The Wasserstein distance itself is an important geometric object; however it has also become important in applications in recent years. In the context of machine learning, a method using the Wasserstein distance as a loss function was proposed [19]. Based on this idea, in 2017, the Wasserstein-GAN, a generative model minimizing while estimating the Wasserstein distance, was proposed [8]. This model, using the Kantorovich

duality for estimation, has been shown to produce stable, high-quality images compared to conventional generative models. Optimal transport theory is also applied to the analysis of the workings of learning machines. Sonoda and Murata showed that processing in a neural network with very deep hierarchical structure approximates a gradient flow with respect the Wasserstein distance [31]. In the field of image processing, the Wasserstein barycenter [1], a barycenter with respect to Wasserstein distance, has an interesting application. By treating grayscale images as probability distributions and computing the Wasserstein barycenter of them, one can obtain a smooth interpolation of those images [13, 9, 23]. The computation of the Wasserstein barycenter can be executed in a reasonable cost due to the contribution of Cuturi's entropic regularization [12]. Cuturi added the Shannon entropy as a regularization term to the objective function of the optimal transport problem, and showed that an algorithm called the Sinkhorn algorithm can solve the relaxed problem with significantly less computational cost than the original problem.

In the first part of the thesis, we extend the previous work [3, 4] by Amari *et al.* on the relationship of Wasserstein and information geometry. Amari *et al.* interpreted Cuturi's entropic regularization in the framework of information geometry, and proposed a divergence that is consistent with the Wasserstein distance in the limit where the regularization term tends to zero. We aim to extend their framework including computational algorithms to the regularization by general convex functions. In particular, we are interested in the optimal choice of convex function for each problem setting in practical applications. A straightforward generalization of the framework requires a tight assumption other than that the regularization term is strictly convex and smooth; the feasible domain of the dual problem must be unconstrained. Even a simple regularization term, such as a squared function, violates this assumption. We weaken this assumption by using subdifferentials on the boundary of the dual problem. As a demonstration, we constructed a generalized algorithm and computed Wasserstein barycenters regularized by the squared function. Compared to Cuturi's regularization, even when the regularization term is large, the numerical solution obtained in squared regularization has a smaller entropy, which results in less blurred in image processing applications. Unfortunately, the squared regularization could not outperform Cuturi's entropic regularization in terms of computational cost; however, we established a theoretical foundation for finding the optimal regularization term for computation.

We also study machine learning from a geometric perspective. Similar to statistical estimation, the goal of machine learning is to find a function that explains a given set of data well from a parametric family of functions. Thus, information geometry is also applied in machine learning. A well-known example is the Boltzmann machine, which models the function of a neuron to fire with a probability depending on condition of neighbor neurons. Learning a Boltzmann machine can be described as an iteration of alternating projections onto a pair of ∇^e- and ∇^m-autoparallel submanifolds [5]. Similar algorithms are applied in different contexts and known as the em-algorithm. Another example is the natural gradient method. The gradient descent method is used to train multilayer perceptrons (multilayer feedforward neural networks), the most basic model of deep learning. This method reduces some loss function by updating parameters successively in the opposite direction of the gradient using the Newton method. The parameter space of a multilayer perceptron is often referred to as a perceptron manifold; how-

ever, the term "gradient" in the gradient descent method simply means a partial derivative in a particular coordinate system, not a gradient vector in manifold theory. Amari introduced the Fisher metric into the perceptron manifold by regarding the model of a multilayer perceptron as a statistical model [2]. The natural gradient method is a learning method using the proper gradient vector field of the loss function with respect to the Fisher metric, and performs remarkably well compared to the vanilla gradient method.

The perceptron manifold is not a manifold in a strict sense, since it contains many singular regions composed of parameters that represent the same input-output relation. Due to this fact, the gradient-based learning process becomes a challenging dynamical system. In 1999, Amari and Fukumizu investigated the singular regions formed by the degeneration, in which several neurons act the same role and behave essentially as a single neuron. They showed that a singular region with both attractive and repulsive parts appears and causes serious stagnation of learning [20]. Such a region is called a Milnor-like attractor. Although it later became clear that such a structure is a very special case, the singular regions they examined are essential to elucidate how the learning machine acquires the ability to express a target function. The singular regions has a nested structure in the perceptron manifold, and the learning dynamical system reduces to a smaller dimensional subsystem on such a region. Deep learning involves tuning a huge number of parameters, and hence, it is usually not necessary to use all the parameters effectively for representing the target function. In some cases, the input data itself can be reduced to a smaller dimension, and it may be desirable to process it into a lower dimensional intermediate representation in the neural network [21]. In these cases, in order for a trained network to behave well with respect to a given input, the dynamical system needs to be trapped by an appropriate singular region.

In the second part of the thesis, we aim to analyze how degeneration occurs, focusing on the geometry of the perceptron manifold. In the author's thesis for master's degree, he analyzed the dynamics of the "averaged" gradient descent method around a Milnor-like attractor using a center manifold. With the help of the center manifold analysis, we show that the stochastic gradient descent (SGD), a stochastic learning method using training data randomly chosen for each instant, which exhibits different dynamical behavior from the averaged gradient method. Furthermore, for a basic model, we found numerically that the dynamics of SGD tends to cause stronger degeneration. We have considered that this novel type of degeneration suppresses overfitting. Overfitting is a phenomenon in which the system fits too closely to the training data and therefore fail to explain or predict unknown data. These results are consistent with previous studies [22, 39, 11]. Our approach is novel in that it deals with the transient dynamics of individual realization paths, whereas previous studies have been done from an asymptotic perspective.

The construction of the thesis is divided mainly into two parts. In Part I, we analyze the discrete optimal transport problem from the viewpoint of information geometry. In Chapter 2, we outline a regularized problem by Cuturi and the associated geometric structure. In Chapter 3, we give an information geometric analysis to more general regularized problems. In Chapter 4, we further weaken the assumption for regularization term. As a demonstration, we construct

an algorithm computing Wasserstein barycenters regularized the squared term. In Part II, we analyze dynamics of gradient descent learning of multilayer perceptrons. We investigate the learning dynamics in particular around singular regions in the parameter space. In Chapter 5, we outline a multilayer perceptron and its learning methods. In Chapter 6, we introduce the previous analysis around singular regions, including a remarkable example, a Milnor-like attractor. In Chapter 7, we compare a behavior of the stochastic gradient learning with the averaged learning around singular regions. We finally imply that the stochastic learning has a qualitatively different dynamics from the averaged one, which may bring an advantage in terms of generalization performance. Chapter 8 is devoted to the concluding remarks.

Part I

Geometry of optimal transport problem

Chapter 2

Preliminaries

2.1 Optimal transport problem

We treat the discrete setting of the optimal transport problem, that is, the sets of sources and targets are finite. Let $\mathcal{P}_{n-1}$ be the set of nonsingular probability distributions on $\{1, 2, \ldots, n\}$, which is given by

$$\mathcal{P}_{n-1} := \left\{ p \in \mathbb{R}^n \;\middle|\; \sum_{i=1}^{n} p_i = 1,\ p_i > 0, 1 \leq i \leq n \right\},$$

where the subscript $n-1$ describes the dimension as a manifold. We regard each element of $\mathcal{P}_{n-1}$ as a distribution of resources positioned on the set of n sources. We also assign $\mathcal{P}_{m-1}$ to the set of distributions of resources required in m targets.

In order to represent transportation, we introduce a set $\mathcal{P}_{nm-1}$ of marginal distributions defined by

$$\mathcal{P}_{nm-1} := \left\{ P = (P_{ij}) \in \mathbb{R}^{n \times m} \;\middle|\; \sum_{i,j} P_{ij} = 1,\ P_{ij} > 0, 1 \leq i \leq n, 1 \leq j \leq m \right\}.$$

Each entry P_{ij} means a quantity transported from the source i to the target j, and hence, each $P \in \mathcal{P}_{nm-1}$ is called a *transport plan*. Given $p \in \mathcal{P}_{n-1}, q \in \mathcal{P}_{m-1}$, each element of the subset

$$\Pi(p, q) := \left\{ P \in \mathcal{P}_{nm-1} \;\middle|\; \sum_{j=1}^{m} P_{ij} = p_i,\ \sum_{i=1}^{n} P_{ij} = q_j \right\}$$

is called a transport plan from p to q or a *coupling* of p and q.

Let $C = (C^{ij})$ be a given matrix with nonnegative entries. Each entry C^{ij} means a transport cost from a source i to a target j, and we call C a *cost matrix*. Then, the *optimal transport*

problem is presented as the minimization problem:

$$W(p,q) := \inf_{P \in \Pi(p,q)} \langle P, C \rangle = \inf_{P \in \Pi(p,q)} \sum_{i,j} P_{ij} C^{ij}. \tag{2.1}$$

Here, the target function $\langle P, C \rangle$ is the total cost of the transport plan P.

When $m = n$ and C is a distance matrix, it is known that the minimum $W(p,q)$ defines a distance on $\mathcal{P}_{n-1}$ that is often referred to as the *Wasserstein distance*.

2.2 Cuturi's entropic regularization

We herein outline the entropic regularization of the optimal transport problem by Cuturi. In his article [12], instead of the optimization problem (2.1), Cuturi considered the alternative problem:

$$W_\lambda(p,q) := \inf_{P \in \Pi(p,q)} \langle C, P \rangle - \lambda \mathcal{H}(P), \quad \lambda > 0, \tag{2.2}$$

$$\mathcal{H}(P) := -\sum_{ij} P_{ij} \log P_{ij}.$$

Here, the function $\mathcal{H}$ is known as the Shannon entropy and is smooth and concave with respect to the natural affine structure of $\mathcal{P}_{nm-1}$. Unlike that of the original problem (2.1), the target function

$$\Phi_\lambda(P) := \langle C, P \rangle - \lambda \mathcal{H}(P) \tag{2.3}$$

in the problem (2.2) is strictly convex under the natural affine structure of $\mathcal{P}_{nm-1}$. Since the function $\Phi_\lambda(P)$ is strictly convex on a convex affine subspace $\Pi(p,q)$, the problem (2.2) has a unique optimum $P^*(p,q)$, which is called the *optimal transport plan* from p to q, at least on the closure $\overline{\Pi(p,q)} \subset \mathbb{R}^{n \times m}$. This type of regularization is usually referred to as the *entropic regularization*.

The quantity $W_\lambda(p,q)$ does not define a distance on $\mathcal{P}_{n-1}$ in general; however it gives an approximation of the Wasserstein distance $W(p,q)$, that is, $W_\lambda(p,q)$ converges to $W(p,q)$ as λ tends to 0 [14].

Cuturi showed that one can thus compute $W_\lambda(p,q)$ much faster than $W(p,q)$. The optimal transport problem (2.1) can be solved by interior-point methods; however, it costs $O(n^3 \log n)$ time in the worst case for $m = n$ [26]. In contrast, $W_\lambda(p,q)$ is obtained by the *Sinkhorn algorithm*, which costs $O(\max\{n,m\}^2)$ time [12].

By the Sinkhorn algorithm (Algorithm 1), the optimal transport plan $P^*(p,q)$ is computed approximately as follows. Cuturi showed that there exists a solution $(\alpha, \beta) \in \mathbb{R}^{n+m}$ of the dual problem, which gives the optimal transport plan as

$$P^*(p,q)_{ij} = \exp\left(\frac{1}{\lambda}(\alpha_i + \beta_j - C^{ij})\right). \tag{2.4}$$

Leting $K_{ij} := \exp(-C^{ij}/\lambda), u_i = \exp(\alpha_i/\lambda), v_j = \exp(\beta_j/\lambda)$, it is also denoted as

$$P^*(p,q)_{ij} = u_i K_{ij} v_j. \tag{2.5}$$

By Algorithm 1, we obtain the vectors $u^{(t)}, v^{(t)}$, and $P^{(t)} \in \mathcal{P}_{nm-1}$ defined by

$$P^{(t)}_{ij} := u^{(t)}_i K_{ij} v^{(t)}_j.$$

The sequence $\{P^{(t)}\}$ converges to the optimal plan $P^*(p,q)$ as t tends to infinity [27, Section 4]. This algorithm requires $O(\max\{n,m\}^2)$ time for each iteration of the **while** loop.

Algorithm 1 Sinkhorn algorithm

$p \in \mathcal{P}_{n-1}, q \in \mathcal{P}_{m-1}, K = (K_{ij}) \in \mathbb{R}^{n\times m}, \lambda > 0$: given
$u^{(0)} \in \mathbb{R}^n, v^{(0)} \in \mathbb{R}^m$: initial values
while until converge **do**
  for $i = 1$ to n **do**
    $u_i^{(t+1)} \Leftarrow p_i / \left(\sum_{j=1}^m K_{ij} v_j^{(t)}\right)$
  end for
  for $j = 1$ to m **do**
    $v_j^{(t+1)} \Leftarrow q_j / \left(\sum_{i=1}^n u_i^{(t+1)} K_{ij}\right)$
  end for
  $t \Leftarrow t + 1$
end while

2.3 Geometry of entropic regularization

Let g_F be the Fisher metric, ∇^m the m-connection, and ∇^m the e-connection on the probability simplex $\mathcal{P}_{nm-1}$. Amari *et al.* [3] focused on the set of optimal transport plans

$$\mathcal{P}^{opt} := \{\, P^*(p,q) \in \mathcal{P}_{nm-1} \mid p \in \mathcal{P}_{n-1}, q \in \mathcal{P}_{m-1} \,\},$$

which is an exponential family with canonical parameters (α, β) as seen in (2.4). In terms of the information geometry [6], $\mathcal{P}^{opt}$ is a ∇^e-autoparallel submanifold of the dually flat manifold $(\mathcal{P}_{nm-1}, g_F, \nabla^m, \nabla^e)$. On the other hand, for each $p \in \mathcal{P}_{n-1}, q \in \mathcal{P}_{m-1}$,

$$M_{p,\cdot} := \left\{ P \in \mathcal{P}_{nm-1} \,\middle|\, \sum_{j=1}^m P_{ij} = p_i \right\}, \quad M_{\cdot,q} := \left\{ P \in \mathcal{P}_{nm-1} \,\middle|\, \sum_{i=1}^n P_{ij} = q_j \right\}$$

are ∇^m-autoparallel submanifolds. Moreover, they are orthogonal to $\mathcal{P}^{opt}$ with respect to g_F. From that fact, the Sinkhorn algorithm is interpreted as an iterative geometric operations.

Proposition 2.1 (Amari *et al.* [3])**.** *Let $\{u^{(t)}\}_t, \{v^{(t)}\}_t$ be a sequence given by Algorithm 1, and*

$P^{(t)}, Q^{(t)} \in \mathcal{P}_{nm-1}$ *defined as*

$$P_{ij}^{(t)} := u_i^{(t)} K_{ij} v_j^{(t)}, \qquad Q_{ij}^{(t)} := u_i^{(t+1)} K_{ij} v_j^{(t)},$$

for each $t \in \mathbb{N}$*. Then, for each instant* t*,* $Q^{(t)}$ *attains the e-projection of* $P^{(t)}$ *onto* $M_{p,\cdot}$*, and* $P^{(t+1)}$ *attains that of* $Q^{(t)}$ *onto* $M_{\cdot,q}$*.*

The Kullback-Leibler divergence, the canonical divergence on $(\mathcal{P}_{nm-1}, g_F, \nabla^m, \nabla^e)$, is given by

$$KL\left[P \| Q\right] := \sum_{i,j} P_{ij} \log \frac{P_{ij}}{Q_{ij}}, \quad P, Q \in \mathcal{P}_{nm-1}.$$

Due to the projection theorem [6, Theorem 3.9], the e-projection $Q^{(t)}$ of $P^{(t)}$ is given as

$$Q^{(t)} = \underset{R \in M_{p,\cdot}}{\arg\min}\, KL\left[R \| P^{(t)}\right]. \tag{2.6}$$

In this sense, $Q^{(t)}$ is the "nearest" point on $M_{p,\cdot}$ to $P^{(t)}$. In addition, from the Pythagorean theorem [6, Theorem 3.8], it holds that

$$\begin{aligned} KL\left[P^*(p,q) \| P^{(t)}\right] &= KL\left[P^*(p,q) \| Q^{(t)}\right] + KL\left[Q^{(t)} \| P^{(t)}\right] \\ &\geq KL\left[P^*(p,q) \| Q^{(t)}\right], \end{aligned}$$

and the equality holds if and only if $P^{(t)} = Q^{(t)}$. Combined with a similar argument for $P^{(t+1)}$, we obtain

$$KL\left[P^*(p,q) \| P^{(t)}\right] \geq KL\left[P^*(p,q) \| Q^{(t)}\right] \geq KL\left[P^*(p,q) \| P^{(t+1)}\right],$$

and thus, the Kullback-Leibler divergence between $P^*(p,q)$ and $P^{(t)}$ decreases monotonically during the Sinkhorn algorithm (Figure 2.1).

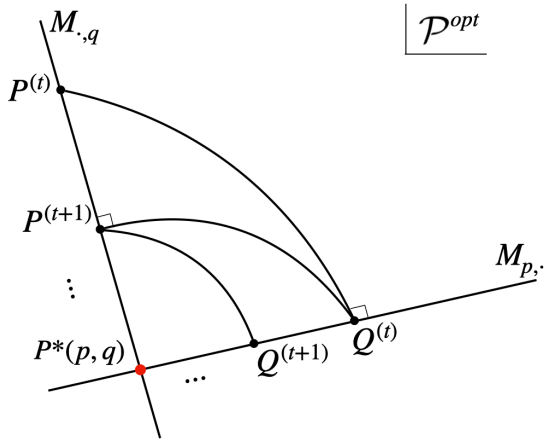


Figure 2.1: A schematic diagram for the geometric interpretation of the Sinkhorn algorithm. The point $Q^{(t)}$ is the e-projection of $P^{(t)}$ onto $M_{p,\cdot}$, and $P^{(t+1)}$ is that of $Q^{(t)}$ onto $M_{\cdot,q}$. All of $P^{(t)}$ and $Q^{(t)}$ belong to $\mathcal{P}^{opt}$ except for $P^{(0)}$. They converge to the optimal plan $P^*(p,q)$, which is located at the intersection of $\mathcal{P}^{opt}$ and $\Pi(p,q) = M_{p,\cdot} \cap M_{\cdot,q}$ (presented by the red point).

2.4 Barycenter problem

The minimum $W_\lambda(p,q)$ in the problem (2.2) is a strictly convex function on $\mathcal{P}_{n-1} \times \mathcal{P}_{m-1}$ with respect to the natural (∇^m-) affine structure. In fact, letting $p_t := (1-t)p_0 + tp_1, q_t := (1-t)q_0 + tq_1$, we have

$$\begin{aligned} W_\lambda(p_t, q_t) &= \inf_{P \in \Pi(p_t, q_t)} \Phi_\lambda(P) \\ &\le \Phi_\lambda\left((1-t)P_0^* + tP_1^*\right) \\ &\le (1-t)\Phi_\lambda(P_0^*) + t\Phi_\lambda(P_1^*) \\ &= (1-t)W_\lambda(p_0, q_0) + tW_\lambda(p_1, q_1), \end{aligned}$$

where P_0^*, P_1^* are optimal plans for $(p_0, q_0), (p_1, q_1)$ respectively. We used the convexity of Φ_λ at the second inequality. The strict convexity of W_λ follows from that of Φ_λ.

As λ tends to 0, $W_\lambda(p,q)$ goes to the minimal cost $W(p,q)$ defined by (2.1). Assuming $m = n$ and some conditions for the cost matrix C, the quantity $W(p,q)$ can be regarded as a distance on $\mathcal{P}_{n-1}$, called the *Wasserstein distance* [35]. Then, one can consider the problem of computing the *Fréchet mean*

$$p^* := \arg \inf_{q \in \mathcal{P}_{n-1}} \sum_{k=1}^{N} r_k \, W(p^k, q), \tag{2.7}$$

of given points $p^1, \ldots, p^N \in \mathcal{P}_{n-1}$ and weights $r_1, \ldots, r_N > 0$ with $\sum_{k=1}^N r_k = 1$. Such a type of mean is called the *Wasserstein barycenter*.

Let us consider a relaxed variant of the problem (2.7): given $p^1, \ldots, p^N \in \mathcal{P}_{n-1}$ and $r_1, \ldots, r_N > 0$ with $\sum_{k=1}^N r_k = 1$,

$$\textbf{Minimize} \sum_{k=1}^{N} r_k \, W_\lambda(p^k, q) \text{ under } q \in \mathcal{P}_{n-1}. \tag{2.8}$$

Benamou *et al.* showed that this problem can be solved by a Sinkhorn-like algorithm [9], which is presented in Algorithm 2. In order to devise this algorithm, they enlarged the domain of the problem to $(\mathcal{P}_{n^2-1})^N$ and considered the problem

$$\textbf{Minimize} \sum_{k=1}^{N} r_k \, \Phi_\lambda(P^k) \text{ under } (P^1, \ldots, P^N) \in \mathbb{M}_1 \cap \mathbb{M}_2,$$

where

$$\mathbb{M}_1 = \left\{ (P^1, \ldots, P^N) \in (\mathcal{P}_{n^2-1})^N \;\middle|\; P^k \in M_{p^k,\cdot} \text{ for } {}^\forall k \right\},$$
$$\mathbb{M}_2 = \left\{ (P^1, \ldots, P^N) \in (\mathcal{P}_{n^2-1})^N \;\middle|\; {}^\exists q \in \mathcal{P}_{n-1} \text{ s.t. } P^k \in M_{\cdot,q}, \text{ for } {}^\forall k \right\}.$$

This problem is, in fact, equivalent to the original one (2.8). Algorithm 2 is obtained by computing the ∇^e-projection onto the pair of ∇^m-autoparallel submanifolds $\mathbb{M}_1$ and $\mathbb{M}_2$ of $(\mathcal{P}_{n^2-1})^N$, iteratively. Each iteration of the **while** loop costs $O(Nn^2)$ time.

Algorithm 2 Benamou *et al.*'s algorithm

$p^1, \ldots, p^N \in \mathcal{P}_{n-1}$, $r_1, \ldots, r_N > 0$ with $\sum_k^N r_k = 1, K \in \mathbb{R}^{n \times m}$: given
$(u^{(1;0)}, \ldots, u^{(N;0)}), (v^{(1;0)}, \ldots, v^{(N;0)}) \in \mathbb{R}^{n \times N}$: initial values
while until converge **do**
  for $k = 1$ to N **do**
    for $i = 1$ to n **do**
      $u_i^{(k;t+1)} \Leftarrow p_i^k / \left(\sum_{j=1}^n K_{ij} v_j^{(k;t)} \right)$
    end for
  end for
  for $j = 1$ to n **do**
    $\tilde{p}_j \Leftarrow \prod_{k=1}^N \left(\sum_{i=1}^n u_i^{(k;t+1)} K_{ij} \right)^{r_k}$
  end for
  for $k = 1$ to N **do**
    for $j = 1$ to n **do**
      $v_j^{(k;t+1)} \Leftarrow \tilde{p}_j / \left(\sum_{i=1}^n u_i^{(k;t+1)} K_{ij} \right)$
    end for
  end for
  $t \Leftarrow t + 1$
end while

Chapter 3

Generalized entropic regularization

In this chapter, for given $p \in \mathcal{P}_{n-1}$ and $q \in \mathcal{P}_{m-1}$, we consider the minimization problem

$$\varphi(p,q) = \inf_{P \in \Pi(p,q)} \Phi(P) \tag{3.1}$$

where Φ is a strictly convex smooth function on $\mathcal{P}_{nm-1}$. We herein assume that Φ can be presented as a restriction of a strictly convex smooth function $\tilde{\Phi}$ defined on an open convex set $U \subset \mathbb{R}_{++}^{n \times m}$ including $\mathcal{P}_{nm-1}$, where $\mathbb{R}_{++}$ denote the sets of positive real numbers. This problem includes Cuturi's regularized optimal transport problem as the case where $\Phi = \Phi_\lambda$ given by (2.3). We construct the dually flat structure suitable for the problem (3.1), and devise a procedure which generalizes the Sinkhorn algorithm. We also address the generalized barycenter problem:

$$\textbf{Minimize} \sum_{k=1}^{N} r_k \, \varphi(p^k, q) \text{ under } q \in \mathcal{P}_{n-1}. \tag{3.2}$$

3.1 Geometric framework

On $\mathcal{P}_{nm-1}$, the Bregman divergence D associated to Φ is given by

$$D(P||Q) := \Phi(\eta(P)) - \Phi(\eta(Q)) - \left\langle \frac{\partial \Phi}{\partial \eta}(\eta(Q)), \eta(P) - \eta(Q) \right\rangle,$$

where η is an affine coordinate system on $\mathcal{P}_{nm-1}$ compatible with the standard affine structure, which is specified later by the equation (3.5). The partial derivative $\partial\Phi/\partial\eta$ defines the dual affine coordinate system θ, i.e.,

$$\theta(P) := \frac{\partial \Phi}{\partial \eta}(\eta(P)), \quad {}^\forall P \in \mathcal{P}_{nm-1}. \tag{3.3}$$

The divergence D induces a dualistic structure (g, ∇, ∇^*) in the standard manner [6]:

$$
\begin{aligned}
g_P(X,Y) &= -X_P Y_Q D(P||Q)\big|_{Q=P}, \\
g_P(\nabla_X Y, Z) &= -X_P Y_P Z_Q D(P||Q)\big|_{Q=P}, \\
g_P(\nabla^*_X Y, Z) &= -X_Q Y_Q Z_P D(P||Q)\big|_{Q=P},
\end{aligned}
$$

for $X, Y, Z \in \mathscr{X}(\mathcal{P}_{nm-1})$ and $P \in \mathcal{P}_{nm-1}$, where $\mathscr{X}(\mathcal{P}_{nm-1})$ denotes the set of vector fields on $\mathcal{P}_{nm-1}$. The connections ∇ and ∇^* automatically become flat, and η and θ are their affine coordinate systems, respectively. Moreover, η and θ are mutually dual with respect to g, that is, they satisfy the relation

$$
g\left(\frac{\partial}{\partial \eta_i}, \frac{\partial}{\partial \theta^j}\right) = \delta^i_j, \tag{3.4}
$$

where δ^i_j denotes the Kronecker delta.

In order to illustrate a geometric presentation of the problem (3.1), we use the affine coordinate system η defined by

$$
\begin{cases}
\eta_{ij}(P) = P_{ij}, \quad 1 \leq {}^\forall i \leq n-1, 1 \leq {}^\forall j \leq m-1, \\
\eta_{im}(P) = \sum_{j=1}^m P_{ij}, \quad 1 \leq {}^\forall i \leq n-1, \\
\eta_{nj}(P) = \sum_{i=1}^n P_{ij}, \quad 1 \leq {}^\forall j \leq m-1.
\end{cases} \tag{3.5}
$$

Under this coordinate system, the subset $\Pi(p,q)$ is presented as

$$
\Pi(p,q) = M_{p,\cdot} \cap M_{\cdot,q}
$$

by using ∇-autoparallel submanifolds defined by

$$
\begin{aligned}
M_{p,\cdot} &:= \left\{ P \in \mathcal{P}_{nm-1} \;\middle|\; \eta_{im}(P) = p_i,\ 1 \leq {}^\forall i \leq n-1 \right\}, \\
M_{\cdot,q} &:= \left\{ P \in \mathcal{P}_{nm-1} \;\middle|\; \eta_{nj}(P) = q_j,\ 1 \leq {}^\forall j \leq m-1 \right\}.
\end{aligned}
$$

In particular, $\Pi(p,q)$ itself is also an ∇-autoparallel submanifold. Then, the problem (3.1) is reduced to

$$
\begin{cases}
\dfrac{\partial \Phi}{\partial \eta_{ij}}(\eta) = 0, \quad 1 \leq {}^\forall i \leq n-1, 1 \leq {}^\forall j \leq m-1, \\
\eta_{im} = p_i,\ \eta_{nj} = q_j, \quad 1 \leq {}^\forall i \leq n, 1 \leq {}^\forall j \leq m,
\end{cases} \tag{3.6}
$$

if its solution lies in the open simplex $\mathcal{P}_{nm-1}$.

From the inverse of (3.5)

$$\begin{cases} P_{ij} = \eta_{ij}, & 1 \leq {}^\forall i \leq n-1, 1 \leq {}^\forall j \leq m-1, \\ P_{im} = \eta_{im} - \sum_{j=1}^{m-1} \eta_{ij}, & 1 \leq {}^\forall i \leq n-1, \\ P_{nj} = \eta_{nj} - \sum_{i=1}^{n-1} \eta_{ij}, & 1 \leq {}^\forall j \leq m-1, \\ P_{nm} = 1 - \sum_{i=1}^{n-1} \eta_{im} - \sum_{j=1}^{m-1} \eta_{nj} + \sum_{i=1}^{n-1} \sum_{j=1}^{m-1} \eta_{ij}, \end{cases}$$

and the relation (3.3), one can check that the dual affine coordinate system θ is given by

$$\begin{cases} \theta^{ij} = S^{ij}(P) - S^{im}(P) - S^{nj}(P) + S^{nm}(P), \\ \qquad\qquad 1 \leq {}^\forall i \leq n-1, 1 \leq {}^\forall j \leq m-1 \\ \theta^{im} = S^{im}(P) - S^{nm}(P), \quad 1 \leq {}^\forall i \leq n-1, \\ \theta^{nj} = S^{nj}(P) - S^{nm}(P), \quad 1 \leq {}^\forall j \leq m-1. \end{cases} \tag{3.7}$$

Here, we denote by $S^{ij} := \partial\tilde{\Phi}/\partial A_{ij}$ for $1 \leq i \leq n, 1 \leq j \leq m$, where $\tilde{\Phi}$ is an extension of Φ to $U \subset \mathbb{R}^{n\times m}_{++}$ and $A = (A_{ij})$ indicates an element of $\mathbb{R}^{n\times m}_{++}$. Thus, the critical condition

$$\theta^{ij} = \frac{\partial \Phi}{\partial \eta_{ij}} = 0$$

in (3.6) is rewritten as

$$S^{ij}(P) - S^{im}(P) = S^{nj}(P) - S^{nm}(P),$$

which implies that $S^{ij}(P) - S^{im}(P)$ does not depend on i. By rearranging terms, one can also check that $S^{ij}(P) - S^{nj}(P)$ is independent of j. As a consequence, there exist $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^m$ satisfying

$$S^{ij}(P) = \alpha^i + \beta^j, \quad 1 \leq {}^\forall i \leq n, 1 \leq {}^\forall j \leq m.$$

From (3.7), we obtain

$$\begin{aligned} \theta^{im}(P) &= \alpha^i - \alpha^n, \quad 1 \leq {}^\forall i \leq n-1, \\ \theta^{nj}(P) &= \beta^j - \beta^m, \quad 1 \leq {}^\forall j \leq m-1, \end{aligned}$$

and thus, the quantities θ^{im} and θ^{nj} are equivalent variables to α^i and β^j up to additive constants α^n, β^m.

Let us consider a ∇^*-autoparallel submanifold defined by

$$\mathcal{P}^{opt} := \left\{ \theta^{ij} = 0, \quad 1 \leq i \leq n-1, 1 \leq j \leq m-1 \right\}.$$

Each element $P \in \mathcal{P}^{opt}$ lies in $\Pi(p, q)$ for some $p \in \mathcal{P}_{n-1}, q \in \mathcal{P}_{m-1}$, and then, P is the solution of the problem (3.1) for those p, q. Therefore, $\mathcal{P}^{opt}$ is the set of optimal solutions for some source and target distributions. From the duality (3.4) of the coordinate systems η and θ, $\mathcal{P}^{opt}$ is orthogonal to $\Pi(p, q)$ with respect to g. Now, the problem (3.1) is interpreted as the problem of finding the intersection point between the ∇^*-autoparallel submanifold $\mathcal{P}^{opt}$ and the

∇-autoparallel submanifold $\Pi(p,q)$, which are mutually orthogonal.

To describe such a type of problem, the concept of a mixed coordinate system is useful.

Proposition 3.1 (Mixed coordinate system)**.** *Let (M, g, ∇, ∇^*) be an n-dimensional dually flat manifold, and $\eta = (\eta_i)$ and $\theta = (\theta^i)$ be affine coordinate systems of ∇ and ∇^*, respectively. Suppose that η and θ are mutually dual. Then, for $1 \leq k \leq n$,*

$$\xi = (\theta^1, \dots, \theta^k, \eta_{k+1}, \dots, \eta_n)$$

becomes a coordinate system on M, which is called a mixed coordinate system.

We use the mixed coordinate system $\xi = (\xi^{ij})$ defined by

$$\begin{cases} \xi^{ij}(P) = \theta^{ij}(P), & 1 \leq {}^\forall i \leq n-1, 1 \leq {}^\forall j \leq m-1, \\ \xi^{im}(P) = \eta_{im}(P), & 1 \leq {}^\forall i \leq n-1, \\ \xi^{nj}(P) = \eta_{nj}(P), & 1 \leq {}^\forall j \leq m-1. \end{cases}$$

Under this coordinate system, the equation (3.6) is presented by

$$\xi^{ij}(P) = 0, \quad \xi^{im}(P) = p_i, \quad \xi^{nj}(P) = q_j, \quad 1 \leq {}^\forall i \leq n-1, 1 \leq {}^\forall j \leq m-1.$$

From the above discussion, the solution $P^*(p,q)$ of the problem (3.1) is, if it exists, given by

$$S^{ij}(P^*(p,q)) = (\alpha^*)^i + (\beta^*)^j,$$

and $\alpha^* \in \mathbb{R}^n$ and $\beta^* \in \mathbb{R}^m$ are determined by the conditions

$$\eta_{im}(P^*(p,q)) = p_i, \quad \eta_{nj}(P^*(p,q)) = q_j, \quad 1 \leq {}^\forall i \leq n, 1 \leq {}^\forall j \leq m,$$

due to Proposition 3.1.

3.2 Generalized Sinkhorn algorithm

We herein assume that the convex function Φ has a smooth extension $\tilde{\Phi}$ onto $U = \mathbb{R}^{n\times m}_{++}$, and that $S : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}^{n\times m}$ is surjective. For example, the function Φ_λ defined by (2.3) has the extension

$$\tilde{\Phi}_\lambda(A) := \langle A, C\rangle - \lambda\tilde{\mathcal{H}}(A), \tag{3.8}$$

$$\tilde{\mathcal{H}}(A) := -\sum_{i,j} A_{ij} \log A_{ij} + \left(\sum_{i,j} A_{ij} - 1\right), \quad A \in \mathbb{R}^{n\times m}_{++},$$

which satisfies the assumption. These assumptions are slightly too strong for introducing an information geometric structure; however, those assumptions are necessary for a straightforward generalization of the entropic regularization, including algorithms which solve it numerically.

Before we discuss the generalization of the Sinkhorn algorithm, let us introduce a dual problem of the problem (3.1). The next lemma is a relaxed version of the Kantorovich duality [35], which is a well-known theorem in the optimal transport theory. A proof of the lemma is located in Appendix A.

Lemma 3.2. *Let $\tilde{\Phi} : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}$ be a convex function. For $p \in \mathcal{P}_{n-1}$, $q \in \mathcal{P}_{m-1}$,*

$$\inf_{P\in\Pi(p,q)} \tilde{\Phi}(P) = \sup_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^m} \{\langle p,\alpha\rangle + \langle q,\beta\rangle - \tilde{\Phi}^*(\alpha\oplus\beta)\},$$

where $\tilde{\Phi}^* : \mathbb{R}^{n\times m} \to \mathbb{R}\cup\{+\infty\}$ is the Legendre transform of $\tilde{\Phi}$ defined by

$$\tilde{\Phi}^*(u) := \sup_{A\in\mathbb{R}^{n\times m}_{++}} \{\langle A,u\rangle - \tilde{\Phi}(A)\}, \quad u \in \mathbb{R}^{n\times m},$$

and $\alpha\oplus\beta \in \mathbb{R}^{n\times m}$ is defined by $(\alpha\oplus\beta)^{ij} := \alpha^i + \beta^j$.

In our setting, since $\tilde{\Phi}$ is an extension of Φ, the left hand side in Lemma 3.2 is equal to that of (3.1). Let us note that the dual problem always has a solution, which is guaranteed by Lemma A.2 in Appendix. We also note that the assumption that the domain of $\tilde{\Phi}$ is $\mathbb{R}^{n\times m}_{++}$ is essential for obtaining this lemma. Using the above lemma, we obtain the following theorem, which guarantees that the primal solution $P^*(p,q)$ is located on the interior of the domain, namely,

$$\inf_{P\in\Pi(p,q)} \Phi(P) = \min_{P\in\Pi(p,q)} \Phi(P).$$

Theorem 3.3. *Suppose that $\tilde{\Phi} : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}$ is smooth and strictly convex and that $S : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}^{n\times m}$ is surjective. There exists a unique solution $P^*(p,q) \in \Pi(p,q)$ of the minimization problem (3.1) for each $p \in \mathcal{P}_{n-1}, q \in \mathcal{P}_{m-1}$. Moreover, a pair of dual solutions $(\alpha^*,\beta^*) \in \mathbb{R}^n\times\mathbb{R}^m$ satisfies*

$$S^{ij}(P^*(p,q)) = (\alpha^*)^i + (\beta^*)^j.$$

Proof. Since Φ is strictly convex and the closure $\overline{\Pi(p,q)}$ is compact,

$$P^*(p,q) := \operatorname*{arg\,inf}_{P\in\Pi(p,q)} \Phi(P)$$

exists uniquely in $\overline{\Pi(p,q)}$. From the surjectivity of S, we can choose $A_* \in S^{-1}(\alpha^*\oplus\beta^*)$. We show that $P^*(p,q) = A_*$.

Due to Lemma 3.2, the primal and dual solutions $P^*(p,q)$, (α^*,β^*) attain the equality

$$\hat{\Phi}(P^*(p,q)) = \langle P^*, \alpha^*\oplus\beta^*\rangle - \tilde{\Phi}^*(\alpha^*\oplus\beta^*),$$

where $\hat{\Phi}(P^*(p,q)) := \lim_{\tilde{A}\to P^*(p,q)} \tilde{\Phi}(\tilde{A})$. This implies that $\alpha^*\oplus\beta^* \in \mathbb{R}^{n\times m}$ is a subgradient of $\tilde{\Phi}$ at $P^*(p,q)$. On the other hand, by the choice of A_*, $\alpha^*\oplus\beta^*$ is also a subgradient of $\tilde{\Phi}$ at A_*.

Hence, for any $t \in (0,1)$, we have

$$\begin{aligned}
&\tilde{\Phi}(tP^*(p,q) + (1-t)A_*) \\
&\geq \langle (tP^*(p,q) + (1-t)A_*) - P^*(p,q), \alpha^* \oplus \beta^* \rangle + \hat{\Phi}(P^*(p,q)) \\
&= \langle tP^*(p,q) + (1-t)A_*, \alpha^* \oplus \beta^* \rangle - \tilde{\Phi}^*(\alpha^* \oplus \beta^*) \\
&= t\left(\langle P^*(p,q), \alpha^* \oplus \beta^* \rangle - \tilde{\Phi}^*(\alpha^* \oplus \beta^*) \right) \\
&\qquad + (1-t)\left(\langle A_*, \alpha^* \oplus \beta^* \rangle - \tilde{\Phi}^*(\alpha^* \oplus \beta^*) \right) \\
&= t\hat{\Phi}(P^*(p,q)) + (1-t)\tilde{\Phi}(A_*).
\end{aligned}$$

From the strict convexity of $\tilde{\Phi}$, if $P^*(p,q) \neq A_*$, it holds that

$$\tilde{\Phi}(tP^*(p,q) + (1-t)A_*) < t\hat{\Phi}(P^*(p,q)) + (1-t)\tilde{\Phi}(A_*),$$

which leads to a contradiction. Hence, we obtain the former assertion

$$P^*(p,q) = A_* \in \overline{\Pi(p,q)} \cap \mathbb{R}^{n\times m}_{++} = \Pi(p,q).$$

The latter assertion follows from the smoothness of $\tilde{\Phi}$, since $\alpha^* \oplus \beta^*$ is a subgradient of $\tilde{\Phi}$ at $P^*(p,q)$. □

The Sinkhorn algorithm is generalized as the iterative procedure consisting of

(S-I) the ∇^*-projection onto the ∇-autoparallel submanifold $M_{p,\cdot}$ and

(S-II) the ∇^*-projection onto the ∇-autoparallel submanifold $M_{\cdot,q}$.

By the pythagorean theorem [6], the Bregman divergence from the solution $P^*(p,q)$ will decrease monotonically for each iteration.

In terms of dual affine coordinate systems, this procedure is written as follows. Suppose that $P^{(t)}$ is given. Then, using the mixed coordinate system $(\eta_{im}, \theta^{ij}, \theta^{nj})$, the first projection $Q^{(t)}$ is designated by the coordinate

$$\begin{aligned}
&\eta_{im}(Q^{(t)}) = p_i, \quad 1 \leq i \leq n-1, \\
&\theta^{ij}(Q^{(t)}) = 0, \quad \theta^{nj}(Q^{(t)}) = \theta^{nj}(P^{(t)}), \quad 1 \leq i \leq n-1, 1 \leq j \leq m-1,
\end{aligned} \tag{3.9}$$

For $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m$, let $A(\alpha,\beta)$ denote an element of $\mathbb{R}^{n\times m}_{++}$ satisfying

$$S^{ij}(A(\alpha,\beta)) = \alpha^i + \beta^j, \quad 1 \leq i \leq n, 1 \leq j \leq m. \tag{3.10}$$

Then, letting $P^{(t)} = A(\alpha^{(t)}, \beta^{(t)})$, the projection is given by

$$Q^{(t)} = A(\alpha^{(t+1)}, \beta^{(t)}),$$

where $\alpha^{(t+1)}$ is determined by the equation

$$\eta_{im}(A(\alpha^{(t+1)}, \beta^{(t)})) = p_i, \quad 1 \leq i \leq n-1. \tag{S-I}'$$

Let us remark that the second equation $\theta^{ij} = 0$ in (3.9) is automatically satisfied because of the equation (3.10), and that the third equation $\theta^{nj}(Q^{(t)}) = \theta^{nj}(P^{(t)})$ means that $\beta^{(t)}$ is fixed. Similarly, the second projection $P^{(t+1)} = A(\alpha^{(t+1)}, \beta^{(t+1)})$ is obtained by solving the equation

$$\eta_{nj}(A(\alpha^{(t+1)}, \beta^{(t+1)})) = q_j, \quad 1 \leq j \leq m-1. \tag{S-II}'$$

Example 3.4 (Sinkhorn algorithm (Algorithm 1))**.** *The partial derivative of the function defined by* (3.8) *is*

$$S^{ij}(A) = C^{ij} + \lambda \log A_{ij}, \quad A \in \mathbb{R}^{n\times m}_{++}.$$

Due to Theorem 3.3, we obtain that

$$\begin{aligned} P^*(p,q)_{ij} &= \exp\left(\frac{1}{\lambda}(\alpha_i + \beta_j - C^{ij})\right) = A(\alpha, \beta) \\ &= u_i K_{ij} v_j, \end{aligned}$$

for some $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m$*, where we put* $u_i = \exp(\alpha_i/\lambda), v_j = \exp(\beta_j/\lambda)$*, and* $K_{ij} := \exp(-C^{ij}/\lambda)$*. Then, given* $P^{(t)} = (u^{(t)})^T K v^{(t)}$*, the equation (S-I)*$'$ *is reduced to*

$$u_i^{(t+1)}\left(\sum_j K_{ij} v_j^{(t)}\right) = p_i, \quad 1 \leq i \leq n,$$

and it is solved to

$$u_i^{(t+1)} = p_i/(Kv^{(t)})_i.$$

This shows that our algorithm actually gives a generalization of the Sinkhorn algorithm.

In the more general case, the equation (S-I)$'$ cannot be solved analytically; however, under some assumptions, one can solve it at least numerically. If each S^{ij} is a function of A_{ij} alone, the equation (S-I)$'$ is given as

$$\sum_{j=1}^{m} (S^{ij})^{-1}\left(\alpha_i^{(t+1)} + \beta_j^{(t)}\right) = p_i, \quad 1 \leq i \leq n-1.$$

In this case, its inverse function $(S^{ij})^{-1}$ becomes an increasing function, since so does the derivative S^{ij} of a convex function $\tilde{\Phi}$. Hence, in addition, if the function $(S^{ij})^{-1}$ is explicitly obtained, one can solve (S-I)$'$ by the Newton method (see Section 3.5, for example).

3.3 Generalized barycenter problem

We assume the surjectivity of $S : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}^{n\times m}$ also in this section. Then, a ∇-affine coordinate system on the submanifold $\mathcal{P}^{opt}$ is given by $(\eta_{im}, \eta_{nj}) = (p_i, q_j)$, since $\mathcal{P}^{opt}$ is characterized by

$\theta^{ij} = 0$. Let φ be a function on $\mathcal{P}_{n-1} \times \mathcal{P}_{m-1}$ defined by (3.1). Since Φ is strictly convex, φ is also strictly convex. Then, the function φ is the potential function for the dually flat structure on $\mathcal{P}^{opt}$. In fact, since $\theta^{ij}(P^*(p,q)) = 0$,

$$\begin{aligned}\frac{\partial \varphi}{\partial p_k}(p,q) &= \sum_{i=1}^{n-1}\sum_{j=1}^{m-1} \frac{\partial \Phi}{\partial \eta_{ij}}(P^*(p,q)) \frac{\partial P^*_{ij}(p,q)}{\partial p_k} \\ &\qquad + \sum_{i=1}^{n-1} \frac{\partial \Phi}{\partial \eta_{im}}(P^*(p,q)) \frac{\partial p_i}{\partial p_k} + \sum_{j=1}^{m-1} \frac{\partial \Phi}{\partial \eta_{nj}}(P^*(p,q)) \frac{\partial q_j}{\partial p_k} \\ &= \sum_{i,j} \theta^{ij}(P^*(p,q)) \frac{\partial P^*_{ij}(p,q)}{\partial p_k} + \sum_i \theta^{im}(P^*(p,q))\delta^k_i \\ &= \theta^{km}(P^*(p,q)). \end{aligned} \tag{3.11}$$

Similarly, the relation $\partial\varphi/\partial q_l = \theta^{nl}$ also follows.

We herein assume that $m = n$, and consider the barycenter problem (3.2). If the solution lies in the interior of the domain, this problem is interpreted as solving the critical condition

$$\frac{\partial}{\partial q_j}\left(\sum_{k=1}^N r_k\, \varphi(p^k, q)\right) = 0, \quad 1 \le j \le n,$$

due to the convexity of φ. Because of the relation (3.11), we can further interpret the problem as the system of equations on $(P^1, \dots, P^N) \in (\mathcal{P}_{n^2-1})^N$:

$$\begin{cases} \eta_{in}(P^k) = p^k_i, & \\ \eta_{nj}(P^1) = \cdots = \eta_{nj}(P^N)\ (= q_j), & \\ \displaystyle\sum_{k=1}^N r_k\, \theta^{nj}(P^k) = 0, & 1 \le i, j \le n-1, 1 \le k \le N. \\ \theta^{ij}(P^k) = 0, & \end{cases} \tag{3.12}$$

In order to illustrate the geometric view of the barycenter problem, we consider the dually flat structure $(\check{g}, \check{\nabla}, \check{\nabla}^*)$ on $(\mathcal{P}_{n^2-1})^N$ induced by the convex function

$$\check{\Phi}(P^1, \dots, P^N) := \sum_{k=1}^N r_k \Phi(P^k).$$

We choose a $\check{\nabla}$-affine coordinate system given by

$$\begin{cases} H^k_{ij}(P^1, \dots, P^N) = \eta_{ij}(P^k), \quad 1 \le k \le N, \\ H^k_{in}(P^1, \dots, P^N) = \eta_{in}(P^k), \quad 1 \le k \le N, \\ H^k_{nj}(P^1, \dots, P^N) = \eta_{nj}(P^k) - \eta_{nj}(P^{k+1}), \quad 1 \le k \le N-1, \\ H^N_{nj}(P^1, \dots, P^N) = \sum_{k=1}^N \eta_{nj}(P^k), \end{cases} \tag{3.13}$$

for $1 \leq i, j \leq n-1$, where $\eta_{ij}, \eta_{im}, \eta_{nj}$ are defined in (3.5). Then, the dual affine coordinate Θ is, from the general relation $\Theta = \partial\check{\Phi}/\partial H$, given by

$$\begin{cases} \Theta^{ij}_k = r_k \theta^{ij}(P^k), \quad 1 \leq k \leq N, \\ \Theta^{in}_k = r_k \theta^{in}(P^k), \quad 1 \leq k \leq N, \\ \Theta^{nj}_k(P) = r_k \theta^{nj}(P^k) - r_{k+1}\theta^{nj}(P^{k+1}), \quad 1 \leq k \leq N-1, \\ \Theta^{nj}_N(P) = \sum_{k=1}^N r_k \theta^{nj}(P^k). \end{cases} \tag{3.14}$$

We treat another minimization problem

$$\textbf{Minimize } \check{\Phi}(P^1, \ldots, P^N) \text{ under } (P^1, \ldots, P^N) \in \mathbb{M}_1 \cap \mathbb{M}_2, \tag{3.15}$$

and show that this problem is equivalent to (3.2). Here, $\mathbb{M}_1$ and $\mathbb{M}_2$ are $\check{\nabla}$-autoparallel submanifolds of $(\mathcal{P}_{n^2-1})^N$ defined by

$$\mathbb{M}_1 := \left\{ (P^1, \ldots, P^N) \in (\mathcal{P}_{n^2-1})^N \;\middle|\; H^k_{in}(P^1, \ldots, P^N) = p^k_i, \; \begin{array}{l} 1 \leq i \leq n-1, \\ 1 \leq k \leq N \end{array} \right\},$$

$$\mathbb{M}_2 := \left\{ (P^1, \ldots, P^N) \in (\mathcal{P}_{n^2-1})^N \;\middle|\; H^k_{nj}(P^1, \ldots, P^N) = 0, \; \begin{array}{l} 1 \leq j \leq n-1, \\ 1 \leq k \leq N-1 \end{array} \right\}.$$

Making use of the dual coordinate system defined in (3.13) and (3.14), the critical condition (3.12) is interpreted as

$$\begin{cases} H^k_{in}(P^1, \ldots, P^N) = p^k_i, \quad 1 \leq k \leq N, \\ H^k_{nj}(P^1, \ldots, P^N) = 0, \quad 1 \leq k \leq N-1, \\ \Theta^{nj}_N(P^1, \ldots, P^N) = 0, \\ \Theta^{ij}_k(P^1, \ldots, P^N) = 0, \quad 1 \leq k \leq N, \quad 1 \leq i, j \leq n-1, \end{cases} \tag{3.16}$$

which is no other than the critical condition

$$(P^1, \ldots, P^N) \in \mathbb{M}_1 \cap \mathbb{M}_2, \quad \frac{\partial \check{\Phi}}{\partial H^N_{nj}} = 0, \quad \frac{\partial \check{\Phi}}{\partial H^k_{ij}} = 0$$

of the problem (3.15). In this mean, the problem (3.15) is another form of (3.2). Hence, via the problem (3.15), the barycenter problem (3.2) is interpreted as the problem finding the intersection point between $\check{\nabla}$-autoparallel submanifold $\mathbb{M}_1 \cap \mathbb{M}_2$ and $\check{\nabla}^*$-autoparallel submanifold

$$\left\{ \Theta^{nj}_N = 0, \Theta^{ij}_k = 0, 1 \leq i, j \leq n-1, 1 \leq k \leq N \right\}.$$

3.4 Generalized algorithm for computing barycenter

Analogously to the Sinkhorn algorithm, the critical condition (3.16) can be solved by an iterative procedure, which is implemented as

(B-I) the $\check{\nabla}^*$-projection $Q^{(t)}$ of $P^{(t)}$ onto the $\check{\nabla}$-autoparallel submanifold $\mathbb{M}_1$, and

(B-II) the $\check{\nabla}^*$-projection $P^{(t+1)}$ of $Q^{(t)}$ onto the $\check{\nabla}$-autoparallel submanifold $\mathbb{M}_2$.

In fact, due to the pythagorean theorem, the Bregman divergence associated to $\check{\Phi}$ monotonically decreases in the procedure.

Suppose that $P^{(t)} = (P^{(1;t)}, \dots, P^{(N;t)})$ satisfies

$$\Theta_k^{ij}(P^{(t)}) = 0,\ \Theta_N^{nj}(P^{(t)}) = 0, \quad 1 \le i, j \le n-1, 1 \le k \le N.$$

Then, the $\check{\nabla}^*$-projection $Q^{(t)}$ of $P^{(t)}$ onto $\mathbb{M}_1$ is presented by a mixed coordinate system of (3.13) and (3.14) as

$$\begin{aligned} H_{in}^k(Q^{(t)}) = p_i^k, \quad \Theta_k^{ij}(Q^{(t)}) = 0 (= \Theta_k^{ij}(P^{(t)})), \quad \Theta_k^{nj}(Q^{(t)}) = \Theta_k^{nj}(P^{(t)}), \\ 1 \le i, j \le n-1, 1 \le k \le N. \end{aligned} \tag{B-I)$'$}$$

On the other hand, the second projection (B-II) onto $\mathbb{M}_2$ is given by

$$\begin{aligned} &H_{nj}^k(P^{(t+1)}) = 0, \quad 1 \le k \le N-1, \\ &\Theta_k^{ij}(P^{(t+1)}) = 0 (= \Theta_k^{ij}(Q^{(t)})), \quad \Theta_k^{in}(P^{(t+1)}) = \Theta_k^{in}(Q^{(t)}), \quad 1 \le k \le N, \\ &\Theta_N^{nj}(P^{(t+1)}) = 0 (= \Theta_N^{nj}(Q^{(t)})), \end{aligned} \tag{B-II)$'$}$$

where $1 \le i, j \le n-1$.

In terms of η and θ given in (3.5) and (3.7), these results are represented as follows:

- finding $Q^{(t)} = (Q^{(1;t)}, \dots, Q^{(N;t)})$ solving

$$\eta_{in}(Q^{(k;t)}) = p_i^k, \quad \theta^{ij}(Q^{(k;t)}) = 0, \quad \theta^{nj}(Q^{(k;t)}) = \theta^{nj}(P^{(k;t)}), \tag{B-I)$''$}$$

  for $1 \le i, j \le n-1, 1 \le k \le N$, and

- finding $P^{(t+1)} = (P^{(1;t+1)}, \dots, P^{(N;t+1)})$ solving

$$\begin{aligned} &\eta_{nj}(P^{(1;t+1)}) = \cdots = \eta_{nj}(P^{(N;t+1)}), \quad \sum_{l=1}^N r_l \theta^{nj}(P^{(l;t+1)}) = 0, \\ &\theta^{ij}(P^{(k;t+1)}) = 0, \quad \theta^{in}(P^{(k;t+1)}) = \theta^{in}(Q^{(k;t)}), \end{aligned} \tag{B-II)$''$}$$

  for $1 \le i, j \le n-1, 1 \le k \le N$.

By using the representation (3.10), we can further reduce each procedure. Letting

$$P^{(k;t)} = A\left(\alpha^{(k;t)}, \beta^{(k;t)}\right),$$

the algorithm is written as

- finding $(\alpha^{(1;t+1)}, \dots, \alpha^{(N;t+1)})$ solving

$$\eta_{in}\left(A\left(\alpha^{(k;t+1)}, \beta^{(k;t)}\right)\right) = p_i^k, \tag{B-I)$'''$}$$

for $1 \leq i \leq n-1, 1 \leq k \leq N$, and

- finding $(\beta^{(1;t+1)}, \ldots, \beta^{(N;t+1)})$ solving

$$\begin{aligned} &\eta_{nj}\left(A\left(\alpha^{(1;t+1)}, \beta^{(1;t+1)}\right)\right) = \cdots = \eta_{nj}\left(A\left(\alpha^{(N;t+1)}, \beta^{(N;t+1)}\right)\right), \\ &\sum_{k=1}^{N} r_k \left(\beta^{(k;t+1)}\right)^j = 0, \end{aligned} \tag{B-II}'''$$

for $1 \leq j \leq n-1, 1 \leq k \leq N$.

When the system of N equations (B-II)$'''$ is hard to solve, one can avoid that difficulty by splitting the projection onto $\mathbb{M}_2$. Let us consider the $\check{\nabla}^*$-projection onto

$$\mathbb{M}_{2;k} := \left\{ (P^1, \ldots, P^N) \in (\mathcal{P}_{n^2-1})^N \;\middle|\; H^k_{nj}(P^1, \ldots, P^N) = 0, 1 \leq j \leq n-1 \right\},$$

for $1 \leq k \leq N-1$. Since

$$\mathbb{M}_2 = \bigcap_{k=1}^{N-1} \mathbb{M}_{2;k},$$

due to the pythagorean theorem, a series of iterative $\check{\nabla}^*$-projections onto the $\check{\nabla}$-autoparallel submanifolds $\mathbb{M}_{2;k}$ decreases monotonically the divergence from $\mathbb{M}_2$. Hence, instead of computing (B-II)$'''$ directly, we can utilize alternative procedure by solving

$$\begin{aligned} &\eta_{nj}\left(A\left(\alpha^{(k;t+1)}, \beta^{(k;t+1)}\right)\right) = \eta_{nj}\left(A\left(\alpha^{(k+1;t+1)}, \beta^{(k+1;t+1)}\right)\right), \\ &r_k \left(\beta^{(k;t+1)}\right)^j + r_{k+1}\left(\beta^{(k+1;t+1)}\right)^j = -\sum_{l \neq k, k+1} r_l \left(\beta^{(l;t+1)}\right)^j, \end{aligned} \tag{B-II}_k$$

with fixing $(\alpha^{(1;t+1)}, \ldots, \alpha^{(N;t+1)})$ and $(\beta^{(1;t+1)}, \ldots, \beta^{(k-1;t+1)}, \beta^{(k+2;t+1)}, \ldots, \beta^{(N;t+1)})$.

3.5 Another perspective using 1-homogeneous extension

Let us introduce another generalization of Amari-Cuturi framework, using 1-homogeneous extension of the convex function Φ on $\mathcal{P}_{nm-1}$, which works well to solve the Tsallis entropic regularized optimal transport problem. Fix $\tilde{q} > 0$ with $\tilde{q} \neq 1$ and $\lambda > 0$. We consider the problem

$$\textbf{Minimize } \Phi(P) = \langle C, P \rangle - \lambda \mathcal{T}_{\tilde{q}}(P) \text{ under } P \in \Pi(p, q),$$

where $\mathcal{T}_{\tilde{q}}$ denotes the $\tilde{q}$-Tsallis entropy, which is given by

$$\mathcal{T}_{\tilde{q}}(P) = \frac{1}{\tilde{q}-1}\left(1 - \sum_{i,j} P_{ij}^{\tilde{q}}\right),$$

This problem is originally considered in [25].

With the simple extension $A \mapsto \sum_{i,j}(\tilde{q}-1)^{-1}(1-A_{ij}^{\tilde{q}})$ of the Tsallis entropy, the range of

$$S^{ij}(A) = \frac{\partial \tilde{\Phi}}{\partial A_{ij}} = C^{ij} + \frac{\lambda \tilde{q}}{\tilde{q}-1} A_{ij}^{\tilde{q}-1}$$

becomes the subset

$$\left\{ u \in \mathbb{R}^{n\times m} \;\middle|\; u_{ij} > C^{ij} \right\},$$

which violates the assumption that S is surjective. We herein consider the extension of $\mathcal{T}_{\tilde{q}}$ defined by

$$\tilde{\mathcal{T}}_{\tilde{q}}(A) = \frac{1}{\tilde{q}-1} \sum_{i,j} \left(A_{ij} - \left(\sum_{k,l} A_{kl} \right)^{1-\tilde{q}} A_{ij}^{\tilde{q}} \right), \quad A \in \mathbb{R}_{++}^{n\times m}.$$

which is 1-homogeneous, that is, $\tilde{\mathcal{T}}_{\tilde{q}}(tP) = t\mathcal{T}_{\tilde{q}}(P)$ for any $t > 0$ and $P \in \mathcal{P}_{nm-1}$. Then, the associated mapping $S : \mathbb{R}_{++}^{n\times m} \to \mathbb{R}^{n\times m}$ is given by

$$S^{ij}(tP) = S^{ij}(P) = C^{ij} + \frac{\lambda}{\tilde{q}-1} \left(\tilde{q} P_{ij}^{\tilde{q}-1} + (1-\tilde{q}) \sum_{k,l} P_{kl}^{\tilde{q}} - 1 \right), \tag{3.17}$$

for $t > 0, P \in \mathcal{P}_{nm-1}$.

In general, a 1-homogeneous convex function induces a dually flat structure, which is called Dawid's decision geometry [15]. For a 1-homogeneous $\tilde{\Phi}$, its derivative $S : \mathbb{R}_{++}^{n\times m} \to \mathbb{R}^{n\times m}$ induces a mapping from $\mathcal{P}_{nm-1}$ to $\mathbb{R}^{n\times m}/\langle 1_{nm} \rangle$. Here, $\mathbb{R}^{n\times m}/\langle 1_{nm} \rangle$ denotes a quotient vector space divided by

$$u \sim v \iff u - v = c\, 1_{nm} \text{ for some } c \in \mathbb{R},$$

where 1_{nm} denotes the matrix whose entries are all 1. Instead of Theorem 3.3, in this case, one can utilize the next theorem, whose proof is located in Appendix A.

Theorem 3.5. *Suppose that $\tilde{\Phi} : \mathbb{R}_{++}^{n\times m} \to \mathbb{R}$ is 1-homogeneous and is strictly convex on $\mathcal{P}_{nm-1}$, and that its derivative $S : \mathbb{R}_{++}^{n\times m} \to \mathbb{R}^{n\times m}$ induces a bijection between*

$$\mathcal{P}_{nm-1} \cong \mathbb{R}^{n\times m}/\langle 1_{nm} \rangle .$$

Then, there exists a unique solution $P^*(p,q) \in \Pi(p,q)$ of the minimization problem (3.1) for each $p \in \mathcal{P}_{n-1}, q \in \mathcal{P}_{m-1}$. Moreover, there exists a pair $(\alpha^*, \beta^*) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfying

$$S(P^*(p,q))^{ij} = (\alpha^*)^i + (\beta^*)^j.$$

Applying Theorem 3.5 to the mapping (3.17), there are $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m$ satisfying

$$C^{ij} + \lambda \left(\frac{\tilde{q}}{\tilde{q}-1} P^*(p,q)_{ij}^{\tilde{q}-1} - \kappa \right) = \alpha^i + \beta^j,$$
$$\kappa := \sum_{k,l} P^*(p,q)_{kl}^{\tilde{q}} + \frac{1}{\tilde{q}-1}.$$

We can include κ in (α, β) by replacing α^i with $\alpha^i + \lambda\kappa$. Thus, the optimal plan has the form as

$$P^*(p,q)_{ij} = \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left(\alpha^i + \beta^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}}.$$

The first projection (S-I) in the generalized Sinkhorn algorithm is interpreted in this case as solving

$$\sum_{j=1}^{m} \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left(\alpha^i + \beta^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}} = p_i, \quad 1 \le i \le n,$$

for the variable α with fixing β. In practice, this can be solved by the Newton method, for example. The second projection (S-II) is similarly given by solving

$$\sum_{i=1}^{n} \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left(\alpha^i + \beta^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}} = q_j, \quad 1 \le j \le m,$$

with fixing α.

For the barycenter problem, the procedure (B-I) is similarly given as the equation

$$\sum_{j=1}^{n} \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left((\alpha^k)^i + (\beta^k)^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}} = p_i^k, \quad 1 \le i \le n, 1 \le k \le N,$$

for $(\alpha^1, \dots, \alpha^N)$. The procedure (B-II) is given by

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left((\alpha^1)^i + (\beta^1)^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}} = \cdots = \sum_{i=1}^{n} \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left((\alpha^N)^i + (\beta^N)^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}}, \\ \displaystyle\sum_{k=1}^{N} r_k (\beta^k)^j = 0, \qquad 1 \le j \le n. \end{cases}$$

However, this equation is hard to solve. As an alternative way, we can make use of the procedure (B-II)$_k$, which is the equation

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left((\alpha^k)^i + (\beta^k)^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}} = \sum_{i=1}^{n} \left(\frac{\tilde{q}-1}{\lambda\tilde{q}} \left((\alpha^{k+1})^i + (\beta^{k+1})^j - C^{ij} \right) \right)^{\frac{1}{\tilde{q}-1}}, \\ r_k (\beta^k)^j + r_{k+1} (\beta^{k+1})^j = - \displaystyle\sum_{l \ne k, k+1} r_l (\beta^l)^j, \qquad 1 \le j \le n, \end{cases}$$

only for $(\beta^k)^j$ and $(\beta^{k+1})^j$. Deleting the variable $(\beta^{k+1})^j$ by using the second relation, the solution $(\beta^k)^j$ can also be computed by the Newton method.

Chapter 4

Relaxation of assumptions

In this chapter, we consider a more general case. As in the previous chapter, we assume that the strictly convex function $\Phi : \mathcal{P}_{nm-1} \to \mathbb{R}$ has a smooth extension $\tilde{\Phi}$ to $\mathbb{R}^{n\times m}_{++}$. In this chapter, we weaken the assumption that the derivative $S : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}^{n\times m}$ of $\tilde{\Phi}$ is surjective, and consider a situation where each S^{ij} can be bounded from below. For example, $\tilde{\Phi}(A) = \frac{1}{2}\sum_{i,j} A_{ij}^2$ with its derivative $S^{ij}(A) = A_{ij} \geq 0$ is a target of this chapter.

The lack of surjectivity of S implies that the optimal plan can be located on the boundary of $\mathcal{P}_{nm-1}$. As seen in the above, the mapping S gives a correspondence between the primal and dual domains. Thus, when the image of S has a boundary in the dual domain, it corresponds to the boundary of the primal domain $\mathcal{P}_{nm-1}$.

Such a situation makes the problems difficult; however, it also can provide an advantage. It allows some masses of the optimal plan to be strictly zero, while all masses necessarily become positive in Cuturi's entropic regularization. In the application of optimal transportation in image processing, this means that a solution in our framework can have strictly white pixels. It can avoid a blurred image, which has been an issue of the original entropic regularization.

4.1 Duality and subdifferentials

Due to the convex analysis, we obtain the picture as follows. Let $\hat{\Phi} : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{+\infty\}$ be the continuous extension of $\tilde{\Phi} : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}$, that is,

$$\hat{\Phi}(A) = \begin{cases} \tilde{\Phi}(A), & A \in \mathbb{R}^{n\times m}_{++} \\ \lim\limits_{\tilde{A}\to A} \tilde{\Phi}(\tilde{A}), & A \in \mathbb{R}^{n\times m}_{+} \setminus \mathbb{R}^{n\times m}_{++} \\ +\infty, & A \notin \mathbb{R}^{n\times m}_{+} \end{cases},$$

where the symbol $\mathbb{R}_+$ denotes the set of nonnegative real numbers. Then, the convex function $\hat{\Phi}$ is not smooth only on

$$\partial\mathbb{R}^{n\times m}_{+} := \left\{ A \in \mathbb{R}^{n\times m} \;\middle|\; A_{ij} = 0 \text{ for some } (i,j) \right\}.$$

In contrast with the one-to-one correspondence

$$\eta \longleftrightarrow \theta = \frac{\partial \hat{\Phi}}{\partial \eta}(\eta)$$

on $\mathbb{R}^{n\times m}_{++}$, we make use of a one-to-many correspondence

$$\eta \longleftrightarrow \partial\hat{\Phi}(A(\eta))$$

on $\partial\mathbb{R}^{n\times m}_{+}$. Here, $\partial\hat{\Phi}(A)$ denotes the subdifferential of $\hat{\Phi}$ at $A \in \mathbb{R}^{n\times m}_{+}$, which is a convex subset of $\mathbb{R}^{n\times m}$ defined by

$$\partial\hat{\Phi}(A) := \left\{ S \in \mathbb{R}^{n\times m} \;\middle|\; \hat{\Phi}(\tilde{A}) \geq \hat{\Phi}(A) + \left\langle S, \tilde{A} - A \right\rangle, {}^{\forall}\tilde{A} \in \mathbb{R}^{n\times m}_{++} \right\}.$$

By using this type of one-to-many correspondence, we construct a pseudo-surjective mapping from the primal domain to the dual domain. The subdifferential on $\partial\mathbb{R}^{n\times m}_{+}$ is given in detail by the next lemma.

Lemma 4.1. *Suppose that the extension of $\hat{\Phi}$ is finite and of C^1 on $\mathbb{R}^{n\times m}_{+}$. Then, for $A \in \partial\mathbb{R}^{n\times m}_{+}$, letting*

$$\Lambda_A := \{\, (i,j) \mid 1 \leq i \leq n, 1 \leq j \leq m \ s.t.\ A_{ij} \neq 0 \,\},$$

the subdifferential of $\hat{\Phi}$ at A is given by

$$\partial\hat{\Phi}(A) = \left\{ S = (S^{ij}) \;\middle|\; \begin{array}{ll} S^{ij} = \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A), & (i,j) \in \Lambda_A, \\ S^{ij} \leq \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A), & (i,j) \notin \Lambda_A \end{array} \right\},$$

where

$$\frac{\partial\hat{\Phi}}{\partial A_{ij}}(A) = \lim_{\tilde{A}\to A,\ \tilde{A}\in\mathbb{R}^{n\times m}_{++}} \frac{\partial\tilde{\Phi}}{\partial A_{ij}}(\tilde{A})$$

Proof. Let $S = (S^{ij}) \in \mathbb{R}^{n\times m}$ satisfy

$$S^{ij} = \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A),\ (i,j) \in \Lambda_A, \qquad S^{ij} \leq \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A),\ (i,j) \notin \Lambda_A.$$

Then, for any $\tilde{A} \in \mathbb{R}^{n\times m}_{+}$, it holds that

$$(i,j) \in \Lambda_A \Longrightarrow \tilde{A}_{ij} - A_{ij} = \tilde{A}_{ij} \geq 0,$$

and thus,

$$\begin{aligned} \hat{\Phi}(A) + \left\langle S, \tilde{A} - A \right\rangle &= \hat{\Phi}(A) + \sum_{i,j} S^{ij} \left(\tilde{A}_{ij} - A_{ij} \right) \\ &\leq \hat{\Phi}(A) + \sum_{i,j} \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A) \left(\tilde{A}_{ij} - A_{ij} \right) \\ &\leq \hat{\Phi}(\tilde{A}). \end{aligned}$$

This implies that $S \in \partial\hat{\Phi}(A)$, where we used $(\partial\hat{\Phi}/\partial A)(A) \in \partial\hat{\Phi}(A)$ to obtain the last inequality.

Conversely, let $S \in \partial\hat{\Phi}(A)$. Since the restriction of $\hat{\Phi}$ onto the affine subspace

$$\mathbb{R}^{\Lambda_A}_{++} := \left\{ \tilde{A} = (\tilde{A}_{ij}) \middle| \begin{array}{ll} \tilde{A}_{ij} > 0, & (i,j) \in \Lambda_A, \\ \tilde{A}_{ij} = 0, & (i,j) \notin \Lambda_A \end{array} \right\}$$

is strictly convex and differentiable, the coordinates of the subgradient of the restriction must be $(\partial\hat{\Phi}/\partial A_{ij})(A)$, which leads

$$(i,j) \in \Lambda_A \Longrightarrow S^{ij} = \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A).$$

For $(k,l) \notin \Lambda_A$, we assume that $S^{kl} > (\partial\hat{\Phi}/\partial A_{kl})(A)$. Letting

$$\tilde{A} = \left\{ \begin{array}{l} \tilde{A}_{kl} = A_{kl} + \varepsilon, \\ \tilde{A}_{ij} = A_{ij}, \quad (i,j) \neq (k,l) \end{array} \right. ,$$

from the Taylor expansion

$$\begin{aligned} \tilde{\Phi}(\tilde{A}) &= \hat{\Phi}(A) + \sum_{i,j} \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A) \left(\tilde{A}_{ij} - A_{ij} \right) + O(\|\tilde{A} - A\|^2) \\ &= \hat{\Phi}(A) + \varepsilon \frac{\partial\hat{\Phi}}{\partial A_{kl}}(A) + O(\varepsilon^2), \end{aligned}$$

we obtain, for sufficiently small $\varepsilon > 0$,

$$\tilde{\Phi}(\tilde{A}) = \hat{\Phi}(A) + \varepsilon S^{kl} < \hat{\Phi}(A) + \sum_{i,j} S^{ij} \left(\tilde{A}_{ij} - A_{ij} \right).$$

This contradicts to $S \in \partial\hat{\Phi}(A)$, and hence,

$$(i,j) \notin \Lambda_A \Longrightarrow S^{ij} \leq \frac{\partial\hat{\Phi}}{\partial A_{ij}}(A).$$

□

4.2 Quadratic regularization of optimal transport

We derive an algorithm to compute the barycenter in a similar way to the previous chapter. However, the correspondence between η and $\partial\hat{\Phi}(A)$ is not one-to-one for $A \in \partial\mathbb{R}^{n\times m}_+$. To construct a practical algorithm, Lemma 4.1 helps to find a concrete correspondence. In this section, we introduce the optimal transport problem with quadratic regularization as an instance. It is also studied by Essid and Solomon [17]; however, they assumed that the cost matrix C is induced from a distance on a graph. In our framework, such an assumption is not required.

We consider the convex function

$$\tilde{\Phi}(A) = \langle A, C\rangle + \frac{\lambda}{2}\sum_{i,j} A_{ij}^2, \quad A \in \mathbb{R}_{++}^{n\times m},$$

for a fixed $\lambda > 0$. Its continuous extension is given by

$$\hat{\Phi}(A) = \begin{cases} \langle A, C\rangle + \dfrac{\lambda}{2}\displaystyle\sum_{i,j} A_{ij}^2, & A \in \mathbb{R}_{+}^{n\times m}, \\ +\infty, \quad \text{otherwise} \end{cases}.$$

Due to Lemma 4.1, for $A \in \partial\mathbb{R}_{+}^{n\times m}$, we obtain

$$\partial\hat{\Phi}(A) = \left\{ S = (S^{ij}) \,\middle|\, \begin{array}{l} S^{ij} = C^{ij} + \lambda A_{ij}, \quad (i,j) \in \Lambda_A, \\ S^{ij} \le C^{ij}, \quad (i,j) \notin \Lambda_A \end{array} \right\},$$

and hence, given $S = (S^{ij}) \in \mathbb{R}^{n\times m}$, the corresponding $A(S) \in \mathbb{R}_{+}^{n\times m}$ is presented explicitly by

$$A(S)_{ij} = \frac{1}{\lambda}\left(S^{ij} - C^{ij}\right)^+,$$

where $x^+ := \max\{x, 0\}$. Due to Lemma 3.2, there exists a subgradient $(\alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^m$ such that the optimal plan $P^*(p, q)$ is given as

$$P^*(p,q)_{ij} = \frac{1}{\lambda}\left(\alpha^i + \beta^j - C^{ij}\right)^+.$$

For the minimization problem (3.1), the procedure (S-I) to obtain the ∇^*-projection onto $M_{p,\cdot}$ is written as

$$\sum_j \frac{1}{\lambda}\left(\alpha^i + \beta^j - C^{ij}\right)^+ = p_i. \tag{4.1}$$

Solving this equation is implemented by Algorithm 3. In the algorithm, the function *sort* returns the vector obtained by rearranging the entries of γ so that $\gamma_1^{\downarrow} \ge \gamma_2^{\downarrow} \ge \cdots \ge \gamma_m^{\downarrow}$. This rearrangement is implemented, for example, by the merge sort, which costs $O(m \log m)$ time. The procedure (S-II) can be similarly implemented, and iterating these two procedures realizes a generalization of the Sinkhorn algorithm. The time required for (S-I) and (S-II) is $O(\max\{n, m\}^3)$ in the worst case. Although it requires more cost than Cuturi's algorithm, our algorithm works with reasonable cost, compared with the computational cost $O(n^3 \log n)$ required for solving the original problem (2.1).

Algorithm 3 Solver of (4.1) for each i

Require: $\lambda > 0, 1 \le i \le n, \beta = (\beta^j), p = (p_i), C = (C^{ij})$
Ensure: $\sum_j \frac{1}{\lambda}\left(\alpha^i + \beta^j - C^{ij}\right)^+ = p_i$
  $\gamma \Leftarrow (\beta^j - C^{ij})_j$
  $\gamma^\downarrow \Leftarrow \mathrm{sort}(\gamma)$
  for $J = 1$ to m **do**
    $\alpha^i \Leftarrow \left(\lambda p_i - \sum_{j=1}^J \gamma_j^\downarrow\right)/J$
    if $\sum_j \frac{1}{\lambda}\left(\alpha^i + \beta^j - C^{ij}\right)^+ = p_i$ **then**
      break
    end if
  end for

For the barycenter problem, the procedure (B-I) is presented in a similar form to (4.1), which is solved by Algorithm 3 for each k and i. On the other hand, the procedure (B-II) is hard to solve, and we make use of the alternative procedure $(\text{B-II})_k$, which is written as the equation

$$\begin{aligned} \sum_i \frac{1}{\lambda}\left((\alpha^k)^i + (\beta^k)^j - C^{ij}\right)^+ &= \sum_i \frac{1}{\lambda}\left((\alpha^{k+1})^i + (\beta^{k+1})^j - C^{ij}\right)^+, \\ r_k(\beta^k)^j + r_{k+1}(\beta^{k+1})^j &= -\sum_{l\neq k,k+1} r_l(\beta^l)^j =: \Sigma^{k;j}, \end{aligned} \tag{4.2}$$

of $\beta^k, \beta^{k+1} \in \mathbb{R}^n$, where $(\alpha^1, \dots, \alpha^N)$ and $(\beta^1, \dots, \beta^{k-1}, \beta^{k+2}, \dots, \beta^N)$ are fixed. We can solve this equation by Algorithm 4.

In summary, the barycenter for the quadratic regularized Wasserstein distance is obtained by applying Algorithm 3 to p^k for $1 \le k \le N$ and performing Algorithm 4 for $1 \le k \le N-1$, iteratively. Since Algorithm 4, in the worst case, costs $O(n^2)$ time for each j and k, each iteration of the main loop may require the cost $O(Nn^3)$.

Algorithm 4 Solver of (4.2) for each j and k

Require: $\lambda > 0, 1 \le j \le n, 1 \le k \le N-1, C = (C^{ij}), (\alpha^k, \alpha^{k+1}), \Sigma^{k;j}$
Ensure: $(\beta^k)^j, (\beta^{k+1})^j$ solves (4.2)
  $\gamma^1 \Leftarrow ((\alpha^k)^i - C^{ij})_i$; $\gamma^2 \Leftarrow ((\alpha^{k+1})^i - C^{ij})_i$
  $(\gamma^1)^\downarrow \Leftarrow \mathrm{sort}(\gamma^1)$; $(\gamma^2)^\downarrow \Leftarrow \mathrm{sort}(\gamma^2)$
  $I_1 \Leftarrow 1$; $I_2 \Leftarrow 1$
  while $I_1 < n$ and $I_2 < n$ **do**
    $(\beta^k)^j \Leftarrow \frac{1}{r_{k+1}I_1 + r_k I_2}\left(I_2\Sigma^{k;j} - r_{k+1}\left(\sum_{i=1}^{I_1}(\gamma^1)_i^\downarrow - \sum_{l=1}^{I_2}(\gamma^2)_l^\downarrow\right)\right)$
    $(\beta^{k+1})^j \Leftarrow \frac{1}{r_{k+1}}\left(\Sigma^{k;j} - r_k(\beta^k)^j\right)$
    if $\sum_i \left((\alpha^k)^i + (\beta^k)^j - C^{ij}\right)^+ = \sum_i \left((\alpha^{k+1})^i + (\beta^{k+1})^j - C^{ij}\right)^+$ **then**
      break
    else if $\left(\sum_i \left((\alpha^k)^i + (\beta^k)^j - C^{ij}\right)^+ - \sum_i \left((\alpha^{k+1})^i + (\beta^{k+1})^j - C^{ij}\right)^+\right) \cdot (\gamma^1)^\downarrow_{I_1+1} \le 0$ **then**
      $I_1 \Leftarrow I_1 + 1$
    else
      $I_2 \Leftarrow I_2 + 1$
    end if
  end while

4.3 Numerical Simulations

We performed numerical simulations of solving the barycenter problems with Cuturi's entropic regularization (2.8) and with quadratic regularization. We prepared two 32×32 pixel grayscale images as extreme points $p^1, p^2 \in \mathcal{P}_{n-1}$, where $n = 32 \times 32 = 1024$. They are presented in Figure 4.1. The cost matrix C is set to $C_{ij} = |i-j|^2$ for $1 \le i, j \le n$.

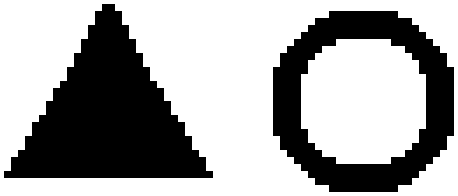


Figure 4.1: Two 32×32 pixel grayscale images prepared as $p^1, p^2 \in \mathcal{P}_{n-1}$. We regard each pixel as event $i \in \{1, 2, \ldots, n\}$ and brightness of each pixel as a density on i, that is, a white pixel has its density zero.

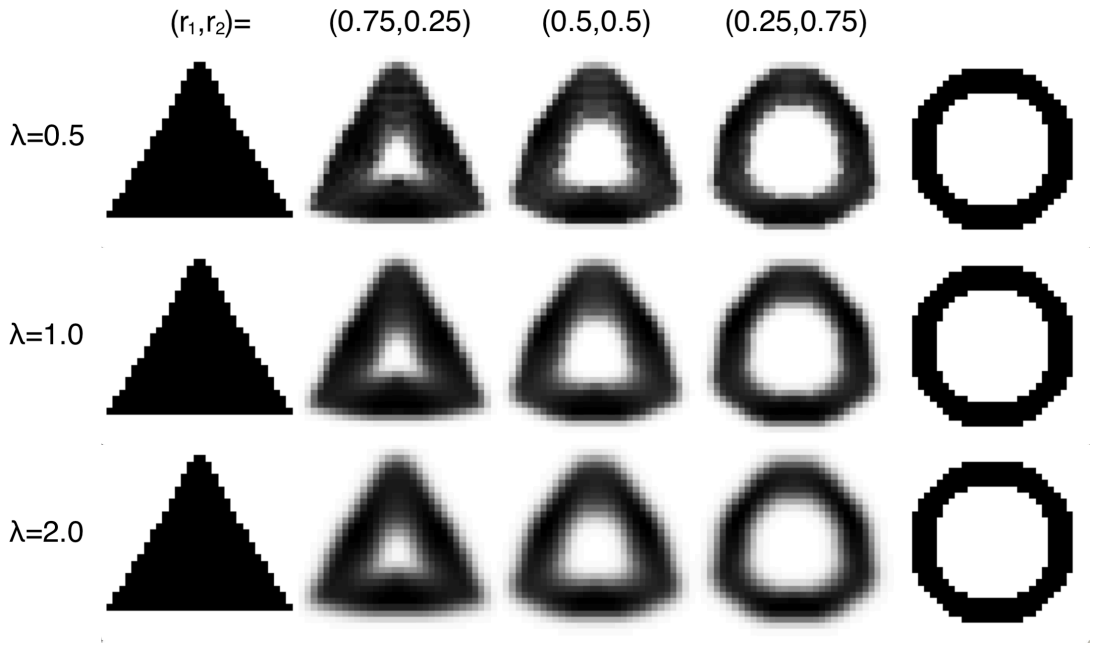


Figure 4.2: Wasserstein barycenters regularized by Shannon entropy computed by Benamou *et al.*'s algorithm. The number of iteration is 3,000, and the regularization constant λ is 0.5 (top), 1.0 (middle), and 2.0 (bottom). The ratio (r_1, r_2) is set to (0.75,0.25), (0.5,0.5), and (0.25,0.75), for each simulation.

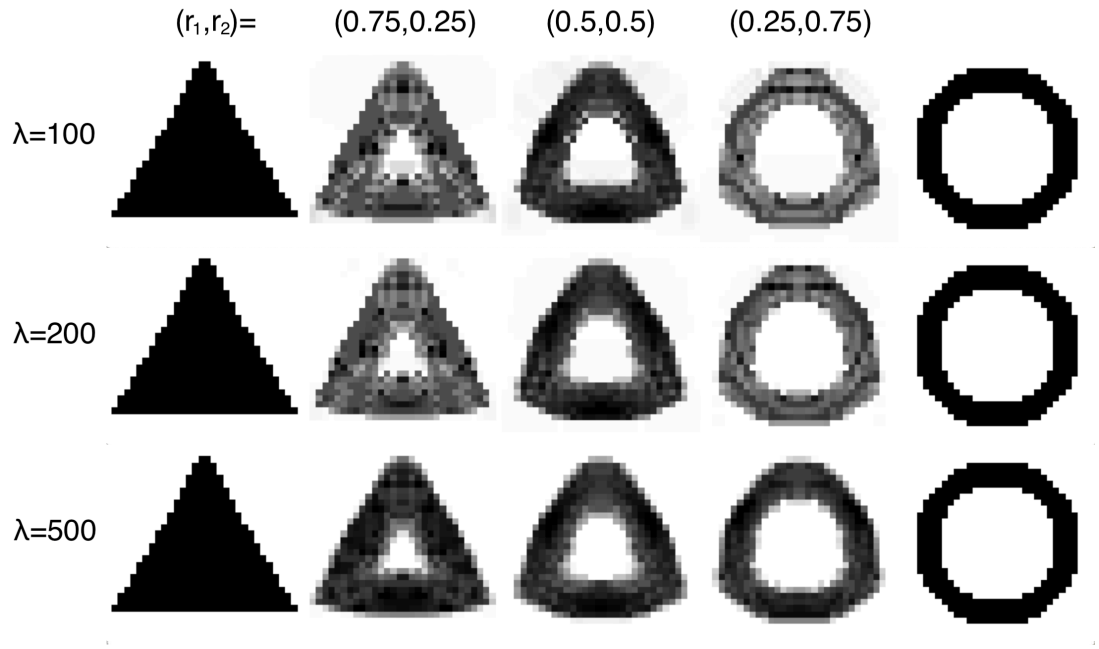


Figure 4.3: Wasserstein barycenters regularized by quadratic term computed by our algorithm. The number of iteration is 3,000, and the regularization constant λ is 100 (top), 200 (middle), and 500 (bottom). The ratio (r_1, r_2) is set to (0.75,0.25), (0.5,0.5), and (0.25,0.75), for each simulation.

We set the number of iteration to 3,000, and the ratio (r_1, r_2) of the Fréchet mean to (0.75,0.25), (0.5,0.5), and (0.25,0.75). Figure 4.2 presents the results of numerical simulations for Cuturi's entropic regularization solved by Benamou *et al.*'s algorithm (Algorithm 2), and the regularization constant λ is set to 0.5, 1.0, and 2.0. Figure 4.3 does that of quadratic regularization solved by our algorithm (Algorithm 3 and Algorithm 4), where λ is set to 100, 200, and 500.

One can see that the barycenters for smaller λ is less blurred in both Figure 4.2 and Figure 4.3. However, the barycenters with Cuturi's regularization has blurred images even for $\lambda = 0.5$, and Benamou *et al.*'s algorithm cannot perform well for smaller λ because of underflow. In fact, the values $K_{ij} = \exp(-C^{ij}/\lambda)$ become too small for practical computation when λ is small. On the contrary, our algorithm uses only summation, and never causes the underflow. Thus, the barycenters with quadratic regularization can be computed even for smaller λ, although the convergence will be slower.

Part II

Geometry of stochastic gradient descent method

Chapter 5

Preliminaries

In this chapter, we introduce notations used throughout the thesis, and outline the basic concepts of machine learning theory. We also introduce well-known issues on machine learning, plateau phenomena and overfitting.

5.1 Perceptron

A *perceptron* is a kind of artificial neural network that models the structure and functions of a biological brain. It consists of a number of units called *artificial neurons*, or *simple perceptrons*, each of which is defined as

$$y = \varphi\left(\sum_{i=1}^{n} w^i x_i - b\right) = \varphi(\boldsymbol{w} \cdot \boldsymbol{x} - b).$$

The vector $\boldsymbol{x} \in \mathbb{R}^n$ represents input signals of a neuron, and $y \in \mathbb{R}$ represents an output signal. The vector $\boldsymbol{w} \in \mathbb{R}^n$ and real number $b \in \mathbb{R}$ are system parameters that represent a weighting factor for input signals and a threshold of activation, respectively. A non-linear function φ called an *activation function*.

At an early stage of the study of neural networks, McCulloch-Pitts [24] used the step function

$$\varphi(z) := \begin{cases} 1 & (z \geq 0) \\ 0 & (z < 0) \end{cases}$$

as an activation function. With this choice of φ, the neuron y returns the output value 1 only when the weighted accumulation $\boldsymbol{w} \cdot \boldsymbol{x}$ of input signals exceeds the threshold b. Such a choice of φ was natural at that time, since the input and output signals were taken to be binary. Later, the use of analog-valued inputs and outputs became popular, and the class of activation functions was enlarged accordingly. Nowadays, a variety of functions are used as activation functions. We only assume that an activation function φ is differentiable (at all but finitely many points), which allows us to use gradient descent algorithms in learning.

To simplify the notation, it is convenient to enlarge the vectors $\boldsymbol{x}$ and $\boldsymbol{w}$ as

$$\boldsymbol{x} = (x_0, x_1, \ldots, x_n),$$
$$\boldsymbol{w} = (w^0, w^1, \ldots, w^n),$$

where $x_0 := 1$ and $w_0 := -b$ are dummy variables, so that the argument of the activation function φ is simply written as

$$\sum_{i=1}^{n} w^i x_i - b = \boldsymbol{w} \cdot \boldsymbol{x}.$$

Unless otherwise noted, we will use this abridged notation throughout the thesis.

A widely used class of perceptrons is the *multilayer perceptron* (MLP), a hierarchical model in which each layer consists of several artificial neurons. A set of input signals is given to the neurons in the first layer, and their output signals are passed to the next layer. The neurons in the next layer receive the previous output signals as inputs and send their outputs to the further next layer. Finally, the output values of the last layer is obtained as the output of the whole system for the given input. A schematic diagram of a multilayer perceptron is shown in Figure 5.1.

In the thesis, we mainly deal with a three-layer perceptron, which consists of an input layer, a hidden layer, and an output layer (Figure 5.2). Since a multilayer perceptron includes a three-layer perceptron as a subnetwork, the phenomena that occur in learning of a three-layer perceptron also occur in that of a multilayer perceptron. The three-layer perceptron is a minimal target for the discussion of learning around singular regions caused by the degeneration of the hidden layer, which is the subject of this thesis. Mathematically, it is defined as follows.

Definition 5.1 (Three-layer perceptron)**.** *A three-layer perceptron composed of n input units, d hidden units and m output units, which is called an (n-d-m)-perceptron, is defined by the following input-output relation:*

$$\boldsymbol{f}^{(d)}(\boldsymbol{x}; \boldsymbol{\theta}) := \sum_{j=1}^{d} \boldsymbol{v}_j \varphi(\boldsymbol{w}_j \cdot \boldsymbol{x}) + \boldsymbol{\eta}. \tag{5.1}$$

Here, $\boldsymbol{x} \in \mathbb{R}^{n+1}$ is an input vector, and

$$\boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_d, \boldsymbol{\eta})$$

is the system parameter with $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d \in \mathbb{R}^{n+1}$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d, \boldsymbol{\eta} \in \mathbb{R}^m$.

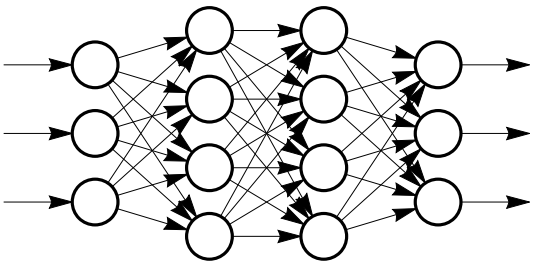


Figure 5.1: Multilayer perceptron.

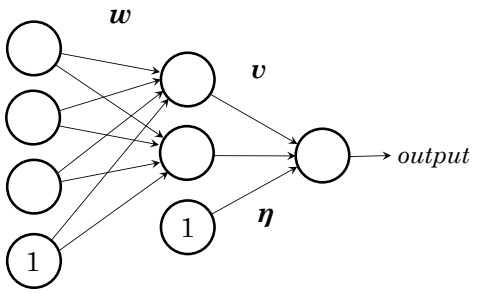


Figure 5.2: Three-layer perceptron.

In what follows, the word "perceptron" means either the set $\{\boldsymbol{f}^{(d)}(\boldsymbol{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ of functions or each function $\boldsymbol{f}^{(d)}(\boldsymbol{x};\boldsymbol{\theta})$ designated by a certain value $\boldsymbol{\theta}$ according to the context.

5.2 Supervised machine learning

In this section, we introduce the supervised learning. For each problem, we set an input-output relation which is called a teacher function $T(\boldsymbol{x})$. The purpose of learning is to adjust the system parameters of a perceptron so that the perceptron emulates the given teacher function as accurately as possible.

To achieve that goal, we consider the minimization problem for a function $L(\boldsymbol{\theta})$ constructed as follows. Let $\ell(\boldsymbol{x},\boldsymbol{y})$ be a function which satisfies $\ell(\boldsymbol{x},\boldsymbol{y}) \geq 0$ with equality if and only if $\boldsymbol{y} = \boldsymbol{T}(\boldsymbol{x})$. Such a function is called an (instantaneous) *loss function*. A typical example is the *squared loss function*:

$$\ell(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{T}(\boldsymbol{x})||^2. \tag{5.2}$$

Once an instantaneous loss function is chosen, the *averaged loss function* (or the *risk function*) $L(\boldsymbol{\theta})$ is defined as

$$L(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}))\right],$$

where $\mathbb{E}_{\boldsymbol{x}}$ denotes the expectation over the input vectors $\boldsymbol{x}$, which are generated according to an unknown probability distribution. Note that if there is a parameter $\boldsymbol{\theta}^*$ that satisfies $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}^*) = \boldsymbol{T}(\boldsymbol{x})$ for all $\boldsymbol{x}$, then $L(\boldsymbol{\theta}^*) = 0$ holds.

To minimize the averaged loss function $L(\boldsymbol{\theta})$, we make use of the differential equation

$$\dot{\boldsymbol{\theta}} = -\frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}). \tag{5.3}$$

A method of finding a minimizer of $L(\boldsymbol{\theta})$ by using such a dynamical system is often referred to as a *gradient decent method*. Since the gradient vector $\partial L/\partial\boldsymbol{\theta}$ tells us the direction in which $L(\boldsymbol{\theta})$ increases, we can decrease $L(\boldsymbol{\theta})$ by changing $\boldsymbol{\theta}$ to the reverse direction of the gradient.

To implement this method on a computer, we often make use of the Euler method, in which one changes the value of the parameter $\boldsymbol{\theta}$ successively according to the recursion formula:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \varepsilon \frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_t). \tag{5.4}$$

However, the calculation of the exact averaged loss function $L(\boldsymbol{\theta})$ is a computationally demanding task. Therefore, as an alternative to (5.4), a *stochastic gradient descent* (SGD) method has been proposed, in which one changes the value of $\boldsymbol{\theta}$ successively according to the formula:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \varepsilon \sum_{s=1}^{S} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{x}_s, \boldsymbol{f}(\boldsymbol{x}_s; \boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}, \tag{5.5}$$

where $\{\boldsymbol{x}_s\}_{s=1}^{S}$ be a set of realization of input vectors chosen at random. A learning strategy based on this recursion is called the *batch mode* of the SGD, and the size S of the set of input data is called the *batch size*. When S is large enough, the behavior of the batch mode dynamics (5.5) imitates the dynamics of (5.4) well. On the other hand, when $S = 1$, the formula (5.5) reduces to

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \varepsilon \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}, \tag{5.6}$$

with a single realization of $\boldsymbol{x}$ at time t. This method is called the *on-line mode* of the SGD.

5.3 Known difficulties

5.3.1 Plateau phenomena

In the actual learning process, one sometimes finds that the averaged loss function does not decrease for a long period of steps and starts decreasing again thereafter. Generally, this temporary stagnation of the loss function decrease occurs repeatedly. Such a phenomenon is called a *plateau* phenomenon. In the gradient descent learning, plateau phenomena occur near a critical point of the averaged loss $L(\boldsymbol{\theta})$, since its derivative becomes small there. The learning starts to progress normally again after the parameter $\boldsymbol{\theta}$ gets away from the critical point. As described in the next chapter, due to the structure of a multilayer perceptron, there are subspaces consisting of critical points in the parameter space, and which construct a complex structure. These subspaces trap the dynamics many times during the learning process.

5.3.2 Overfitting

For practical purposes, a learning machine is trained using a prepared training dataset. The goal of training, however, is for the model to be able to successfully represent the teacher function even for unknown data outside of the training dataset. When a highly complex model with a huge number of degrees of freedom is sufficiently trained, the model will be able to explain the training dataset very well, but will often return significantly different values from the teacher function for unknown data. This phenomenon is often referred to as *overfitting*.

Overfitting is usually formulated as a discrepancy between generalization error and empirical error. The averaged loss function we defined above is averaged over all possible data including unknown data. We refer to this type of averaged loss function as the generalization error. On the other hand, the loss function

$$L^{emp}(\boldsymbol{\theta}) := \sum_{m=1}^{M} \ell(\boldsymbol{x}_m, \boldsymbol{f}(\boldsymbol{x}_m; \boldsymbol{\theta}))$$

averaged over the training dataset $\{\boldsymbol{x}_m\}_{m=1}^M$ is called the empirical error. Although the formula is very similar to the summation in the batch mode (5.5) of SGD, note that the batch $\{\boldsymbol{x}_s\}_{s=1}^S$ in the batch mode is a subset randomly chosen from the training dataset at an instance. Overfitting is treated as phenomena in which once can attain very small empirical error while the generalization error does not decrease.

Chapter 6

Singular regions and Milnor-like attractors

In this chapter, we introduce singular regions of multilayer perceptron, which is a key concept in the dynamics of machine learning. As a remarkable class of singular regions, we also explain a Milnor-like attractor found by Fukumizu and Amari [20]. We presents a center manifold analysis of a Milnor-like attractor provided by the author [33].

6.1 Singular regions

The parameter space of a multilayer perceptron is sometimes called a "perceptron manifold." However, in many cases, it is not really a manifold since it usually contains a subset whose points correspond to the same input-output relation. Such a subset is usually referred to as a *singular region*. In general, there are many singular regions due to degeneration of hidden units. For example, let us consider an (n-2-m)-perceptron. Then, for arbitrary $\boldsymbol{w} \in \mathbb{R}^{n+1}, \boldsymbol{v} \in \mathbb{R}^m$, the subset

$$R(\boldsymbol{w}, \boldsymbol{v}) := \{\, \boldsymbol{\theta} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}_1, \boldsymbol{v}_2) \mid \boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}, \boldsymbol{v}_1 + \boldsymbol{v}_2 = \boldsymbol{v} \,\}$$

of the parameter space forms a typical singular region. In fact, on the subset $R(\boldsymbol{w}, \boldsymbol{v})$, an ($n$-2-$m$)-perceptron $\boldsymbol{f}^{(2)}(\boldsymbol{x}; \boldsymbol{\theta})$ is reduced to the following (n-1-m)-perceptron:

$$\boldsymbol{f}^{(1)}(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{v}) := \boldsymbol{v}\, \varphi\, (\boldsymbol{w} \cdot \boldsymbol{x})\,.$$

On such a singular region, the loss function $L^{(2)}$ inherits the criticality of $L^{(1)}$, as the following proposition shows.

Proposition 6.1 (Fukumizu and Amari [20])**.** *Let $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{v}^*)$ be a critical point of $L^{(1)}$. Then, the parameter $\boldsymbol{\theta} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}_1, \boldsymbol{v}_2) = (\boldsymbol{w}^*, \boldsymbol{w}^*, \lambda \boldsymbol{v}^*, (1-\lambda)\boldsymbol{v}^*)$ is a critical point of $L^{(2)}$ for any $\lambda \in \mathbb{R}$.*

Proof.

$$\frac{\partial L^{(2)}}{\partial \boldsymbol{w}_i}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{f}^{(1)}(\boldsymbol{x};\boldsymbol{\theta}^*)) \cdot \boldsymbol{v}_i\right) \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}^T\right] = \lambda_i \frac{\partial L^{(1)}}{\partial \boldsymbol{w}}(\boldsymbol{\theta}^*),$$
$$\frac{\partial L^{(2)}}{\partial \boldsymbol{v}_i}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \ell}{\partial \boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{f}^{(1)}(\boldsymbol{x};\boldsymbol{\theta}^*))\varphi(\boldsymbol{w}^* \cdot \boldsymbol{x})\right] = \frac{\partial L^{(1)}}{\partial \boldsymbol{v}}(\boldsymbol{\theta}^*), \qquad i = 1, 2,$$

where $\lambda_1 := \lambda$ and $\lambda_2 := 1 - \lambda$. Since $\boldsymbol{\theta}^*$ is a critical point of $L^{(1)}$, these are all zero. □

The original form provided in [20] is for an (n-d-1)-perceptron which contains an (n-(d-1)-1)-perceptron as a subnetwork, for general $d \geq 2$. Furthermore, for any $d > d_1 \geq 1$, a similar argument can be made for the inheritance of criticality from $L^{(d)}$ to $L^{(d_1)}$. This suggests that singular regions consisting of critical points universally exist in the parameter space and form a nested structure with respect to the number d of hidden units.

Near a critical point, since the gradient $\partial L/\partial \boldsymbol{\theta}$ almost vanishes, the dynamics slows down. Because of the universality of singular regions, such stagnation is frequently observed, and is referred to as *vanishing gradient phenomena*.

6.2 Milnor-like attractors

When $m = 1$, *i.e.* the output layer is one-dimensional, every point $\boldsymbol{\theta} \in R(\boldsymbol{w}^*, v^*)$ is a critical point of $L^{(2)}$, since $R(\boldsymbol{w}^*, v^*)$ is one-dimensional. In this case, the second-order property of $L^{(1)}$ is also inherited in $L^{(2)}$, and the singular region $R(\boldsymbol{w}^*, v^*)$ can have a remarkable structure which causes serious stagnation of learning.

Proposition 6.2 (Fukumizu and Amari [20])**.** *Let $m = 1$ and $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, v^*)$ be a strict local minimizer of $L^{(1)}$ with $v^* \neq 0$. Define an $(n+1) \times (n+1)$ matrix H by*

$$H := \mathbb{E}_{\boldsymbol{x}}\left[v^* \frac{\partial \ell}{\partial y}(\boldsymbol{x}, f^{(1)}(\boldsymbol{x};\boldsymbol{\theta}^*))\, \varphi''(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}\, \boldsymbol{x}^T\right], \tag{6.1}$$

and let for $\lambda \in \mathbb{R}$

$$\boldsymbol{\theta}_\lambda := (\boldsymbol{w}^*, \boldsymbol{w}^*, \lambda v^*, (1-\lambda)v^*).$$

If the matrix H is positive (resp. negative) definite, then the point $\boldsymbol{\theta} = \boldsymbol{\theta}_\lambda$ is a local minimizer (resp. saddle point) of $L^{(2)}$ for any $\lambda \in (0, 1)$, and is a saddle point (resp. local minimizer) for any $\lambda \in \mathbb{R} \setminus [0, 1]$. On the other hand, if the matrix H is indefinite, then the point $\boldsymbol{\theta}_\lambda$ is a saddle point of $L^{(2)}$ for all $\lambda \in \mathbb{R} \setminus \{0, 1\}$.

This proposition implies that the one-dimensional region $R(\boldsymbol{w}^*, v^*) = \{\boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R}\}$ may have both attractive parts and repulsive parts in the gradient descent method (Figure 6.1). Such a region is referred to as a *Milnor-like attractor* [37]. The parameter $\boldsymbol{\theta}$ near the attractive part flows into the Milnor-like attractor and fluctuates in the region for a long time, until it reaches the repulsive part.

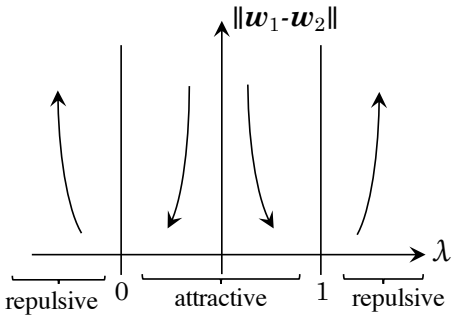


Figure 6.1: Flow near the singular region $R(\boldsymbol{w}^*, v^*)$.

We remark that the phenomenon itself is universal with respect to the number d of hidden units. In fact, the original theorem is given in [20] for an (n-d-1)-perceptron, which contains an (n-(d-1)-1)-perceptron as a subnetwork. The proposition above for (n-2-1)-perceptron is a minimal version.

We also remark that the point $\boldsymbol{\theta}_\lambda$ cannot be a strict local minimizer. In fact, $L^{(2)}$ takes the same value on the singular region $\{\boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R}\}$. In particular, the second derivative of $L^{(2)}$ along the direction λ is always zero.

Let us introduce a practically important case outside the scope of Proposition 6.2. Suppose that a three-layer perceptron has some redundant hidden units to represent the teacher function $T(\boldsymbol{x})$. Mathematically, we suppose that a true parameter $\boldsymbol{\theta}_{true}$ exists (*i.e.* $T(\boldsymbol{x}) = f^{(2)}(\boldsymbol{x}; \boldsymbol{\theta}_{true})$), and that it lies in the singular region $R(\boldsymbol{w}^*, v^*)$. In this case, the function $L^{(2)}$ takes the same value $L^{(1)}(\boldsymbol{w}^*, v^*) = 0$ on $R(\boldsymbol{w}^*, v^*)$. Therefore, every point of $R(\boldsymbol{w}^*, v^*)$ becomes a global minimizer of $L^{(2)}$, and a Milnor-like attractor does not appear. In fact, one can check that the assumption of Proposition 6.2 fails as follows. For each $\boldsymbol{x} \in \mathbb{R}^n$, we obtain

$$\frac{\partial \ell}{\partial y}(\boldsymbol{x}, f^{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*)) = 0,$$

since a loss function $\ell(\boldsymbol{x}, y)$ takes its minimum 0 at $y = T(\boldsymbol{x}) = f^{(1)}(\boldsymbol{x}; \boldsymbol{w}^*, v^*)$. This implies that the matrix H becomes the zero matrix. Thus, in particular, H is neither positive nor negative definite.

We next treat the case when $m \geq 2$. There also exists a one-dimensional region consisting of critical points due to Proposition 6.1. However, in this case, the region becomes simply repulsive, and does not have an attractive part, as the following theorem asserts.

Theorem 6.3 (Tsutsui [33])**.** *Let $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{v}^*)$ be a local minimizer of $L^{(1)}$. If the $m \times (n+1)$ matrix*

$$\mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \ell}{\partial \boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{f}^{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))\varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}^T\right] \tag{6.2}$$

is non-zero, then $\boldsymbol{\theta}_\lambda = (\boldsymbol{w}^*, \boldsymbol{w}^*, \lambda \boldsymbol{v}^*, (1-\lambda)\boldsymbol{v}^*)$ is a saddle point of $L^{(2)}$ for any $\lambda \in \mathbb{R}$, where we regard the derivative $\partial \ell / \partial \boldsymbol{y}$ as a column vector.

In their article [7], Amari *et al.* stated a prototype of Theorem 6.3, although they did not give a full proof. In fact, we found that some additional assumption was necessary to prove their assertion. In Theorem 6.3, we have added a mild assumption that the matrix (6.2) is non-zero. Note that since $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{v}^*)$ is a local minimizer of $L^{(1)}$, it holds that

$$\mathbf{0} = \frac{\partial L^{(1)}}{\partial \boldsymbol{w}}(\boldsymbol{\theta}^*) = (\boldsymbol{v}^*)^T \, \mathbb{E}_{\boldsymbol{x}} \left[\frac{\partial \ell}{\partial \boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{f}^{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*)) \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x}) \, \boldsymbol{x}^T \right].$$

Thus, the matrix (6.2) has a kernel whose dimension is greater than one. Hence, the assumption automatically fails when $m = 1$. This is an underlying mechanism for Proposition 6.2.

By Theorem 6.3, Milnor-like attractors are not observed in the parameter space when $m \geq 2$. However, also in this case, plateau phenomena can occur because of the nested structure of singular regions. Since the singular region is composed of saddle points, it is not repulsive in all directions, but is attractive in some directions, and the dynamics of learning can be attracted to the nested singular regions. To the best of our knowledge, theoretically rigorous study of learning dynamics near the nested structure of singular regions is not known. In Chapter 7 of this thesis, we explain, based on numerical simulation results, that the dynamics of stochastic learning is qualitatively different from those of averaged learning around singular regions.

6.3 Center manifold of Milnor-like attractor

The content described in this section is the main result of master's thesis by the author. The dynamics of the gradient descent method around a Milnor-like attractor can be analyzed by the center manifold analysis. About the definition of a center manifold, please refer to Appendix B. Concretely, under a coordinate system $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$ given by

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1 (\boldsymbol{w}_1 - \boldsymbol{w}^*) + v_2 (\boldsymbol{w}_2 - \boldsymbol{w}^*)}{v^*} + \boldsymbol{w}^* \\ v = v_1 + v_2 \\ \boldsymbol{u} = \dfrac{v_2 (\boldsymbol{w}_1 - \boldsymbol{w}^*) - v_1 (\boldsymbol{w}_2 - \boldsymbol{w}^*)}{v^*} \\ z = v_1 - v_2 \end{cases}, \tag{6.3}$$

the center manifold theorem for a Milnor-like attractor is described as follows.

Theorem 6.4 (Tsutsui [33])**.** *In the coordinate system* (6.3)*, the dynamical system* (5.3) *admits a center manifold structure around the critical points* $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ *in which* $(\boldsymbol{w}, v)$ *converge exponentially fast.*

Here, the points $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ are the boundary points of attractive and repulsive parts in the Milnor-like attractor $R(\boldsymbol{w}^*, v^*) = \{\boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R}\}$. By applying the center manifold theory, near those points, we can assume that the dynamics (5.3) is on the center manifold. In other words, we can reduce the dynamical system into that of slow parameters $(\boldsymbol{u}, z)$. Under the assumption that the dynamics is on the center manifold $(\boldsymbol{w}, v) = \boldsymbol{h}(\boldsymbol{u}, z)$, calculating the Taylor expansion

of $(\dot{\boldsymbol{u}}, \dot{z})$, we obtain an approximated dynamics around $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ as

$$\begin{aligned}\dot{\boldsymbol{u}} &= \frac{1}{2v^*}(z - v^*)H\boldsymbol{u} + O(\|\boldsymbol{u}, z - v^*\|^3),\\ \dot{z} &= \frac{1}{2v^*}\boldsymbol{u}^T H\boldsymbol{u} + O(\|\boldsymbol{u}, z - v^*\|^3).\end{aligned} \tag{6.4}$$

Neglecting the higher order terms, we can integrate this equation to obtain

$$\|\boldsymbol{u}\|^2 = (z - v^*)^2 + C, \tag{6.5}$$

where C is an integral constant.

We remark that Theorem 6.4 is valid even when there exists a true parameter in the singular region $R(\boldsymbol{w}^*, v^*)$; however, in this case, such a simple form of the reduced dynamical system as (B.10) is not obtained. As mentioned above, this case implies that H becomes the zero matrix. Then, the second order terms of the reduced dynamical system vanish, and the third order terms become dominant. It needs to calculate the center manifold $(\boldsymbol{w}, v) = \boldsymbol{h}(\boldsymbol{u}, z)$ up to the second order, which makes the analysis complicated.

6.4 Numerical simulations

As mentioned in the previous section, the dynamics of $(\boldsymbol{w}, v)$ are fast and those of $(\boldsymbol{u}, z)$ are slow under the coordinate system (6.3). In this section, we shall verify this fact by numerical simulations.

As an example, we set the input dimension to be $n = 1$, and choose the teacher function $T : \mathbb{R} \to \mathbb{R}$ defined by

$$T(x) := 2\tanh(x) - \tanh(4x),$$

where tanh is the hyperbolic tangent function. The shape of T is shown in Figure 6.2 by the solid black line. We set the activation function φ as tanh. Thus, the teacher function T can be represented by the (1-2-1)-perceptron with no bias terms

$$f^{(2)}(x; \boldsymbol{\theta}) = v_1\varphi(w_1 x) + v_2\varphi(w_2 x),$$

and the true parameter $\boldsymbol{\theta}_{true}$ is $(w_1, w_2, v_1, v_2) = (1, 4, 2, -1)$. We also discard the bias terms w_1^0 and w_2^0 of the student perceptron. This makes the matrix H scalar valued, and hence the assumption of Proposition 6.2 holds trivially.

Let $\{x_s\}_{s=1}^S$ be a dataset given at random. In order to simulate the averaged learning (5.4), we use the batch learning (5.5) with the batch size $S = 1000$. In this simulation, we draw the dataset $\{x_s\}_{s=1}^S$ *i.i.d.* from $N(0, 2^2)$. Here, $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Under the distribution $N(0, 2^2)$, we obtained a local minimizer $\boldsymbol{\theta}^* = (w^*, v^*) \approx (0.459, 1.15)$ of $L^{(1)}$. The shape of the function $f^{(1)}(x; \boldsymbol{\theta}^*)$ that corresponds to the local minimizer is shown by the dashed blue line in Figure 6.2. The value of H is approximately 0.0472. Since $H > 0$, the attractive region is $\{\boldsymbol{\theta}_\lambda \mid \lambda \in (0, 1)\}$, due to Proposition 6.2.

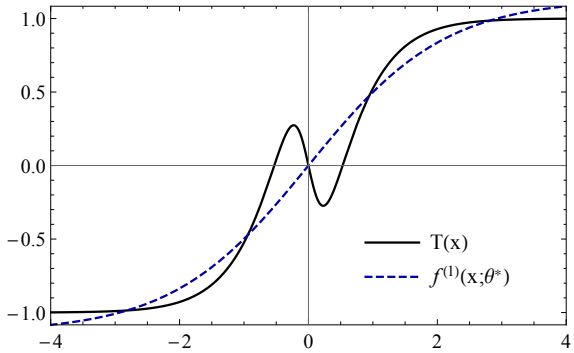
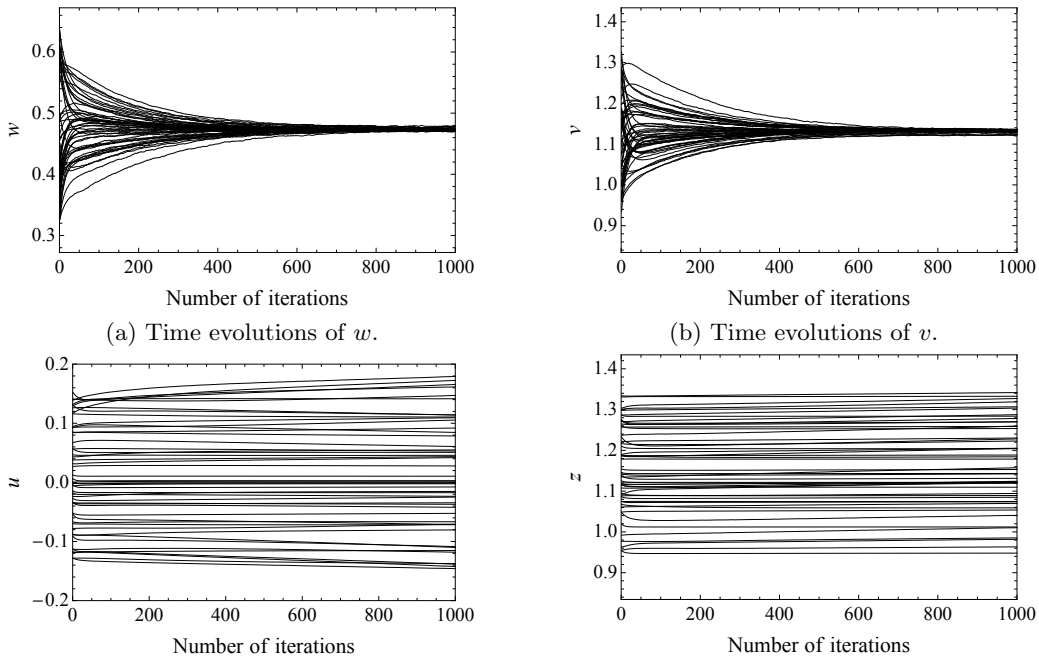


Figure 6.2: The teacher function $T(x)$ and the (1-1-1)-perceptron $f^{(1)}(x;\boldsymbol{\theta}^*)$ which corresponds to the local minimizer $\boldsymbol{\theta}^* = (0.459, 1.15)$.

Figures 6.3(a-d) display time evolutions of each parameter in the first 1,500 iterations from 50 different initial points. We chose an initial parameter $\boldsymbol{\theta}^{(0)} = (w_1^{(0)}, w_2^{(0)}, v_1^{(0)}, v_2^{(0)})$ by

$$w_1^{(0)} = w^* + \zeta_1, \quad w_2^{(0)} = w^* + \zeta_2,$$
$$v_1^{(0)} = v^* + \frac{1}{2}(\zeta_3 + \zeta_4), \quad v_2^{(0)} = \frac{1}{2}(\zeta_3 - \zeta_4),$$

so that $v = v^* + \zeta_3$, and $z = v^* + \zeta_4$, where $\zeta_1, \zeta_2 \sim U(-0.2, 0.2)$, and $\zeta_3, \zeta_4 \sim U(-0.2, 0.2)$. Here, $U(a, b)$ denotes the uniform distribution on the interval $[a, b] \subset \mathbb{R}$. We set the learning rate ε to be 0.05 and the number of iterations to be 20,000. We can see that the parameters w and v converge to their equilibriums exponentially fast ((a) and (b)), while u and z evolve slowly ((c) and (d)).



(a) Time evolutions of w.

(b) Time evolutions of v.

(c) Time evolutions of u.

(d) Time evolutions of z.

Figure 6.3: Time evolutions of each parameter for the first 1,500 iterations. Each trajectory of w or v quickly converges to the equilibrium point $w^* = 0.459, v^* = 1.15$ respectively. On the other hand, trajectories of u and z evolve very slowly compared with w and v.

Figure 6.4 shows evolutions on the $(z, ||u||^2)$-plane. The red circles in the figure represent initial points. When $w = w^*$ and $v = v^*$, the z-axis is a Milnor-like attractor, and the region $|z| < v^*$ is the attractive part of it. We can check that parameters near the attractive region are trapped, and those near the repulsive region are escaping. The intersection point of the line $z = v^*$ and z-axis corresponds to the point $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, the boundary of the attractive and repulsive parts of the Milnor-like attractor. The analytical trajectories (B.11) are plotted as dashed blue curves. Numerical evolutions of the parameter follow the analytical trajectories considerably well around $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. We can find in the figure that some instances of time evolutions change their direction sharply. This is because the fast dynamics of w and v are the main dynamics in the beginning of the learning while the slow dynamics of u and z become dominant after w and v converge to the center manifold.

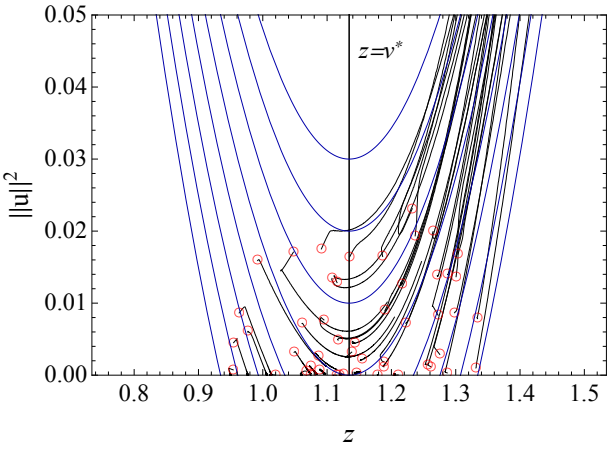


Figure 6.4: Trajectories on the $(z, ||u||^2)$-plane obtained by learning for 20,000 iterations (solid black curves) and analytical trajectories (dashed blue curves) near $\boldsymbol{\theta} = \boldsymbol{\theta}_1 = (w^*, w^*, v^*, 0)$. Red circles represent initial points.

Chapter 7

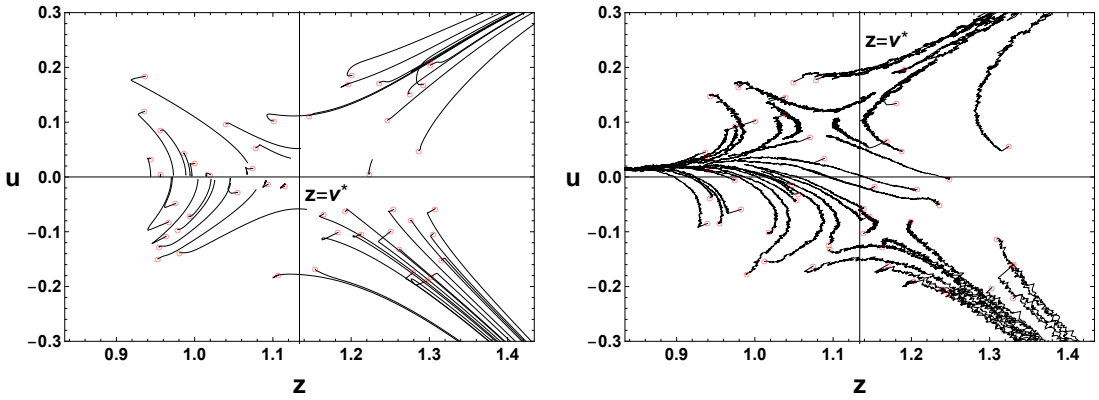
Stochastic gradient descent

In this chapter, we discuss stochastic effects in the learning process. So far, we discussed the deterministic dynamical system (5.3) driven by the averaged gradient. In practice, the dynamical system is simulated as the batch mode (5.5) of SGD over a large number of input data as seen in Section 6.4. In order to reduce the computational cost, the summation is often replaced by the online mode (5.6) of SGD. We use the term "SGD" as the online mode in this chapter unless otherwise noted. Unlike the averaged gradient descent, the learning process by the SGD is a random dynamical system. It is reported that the dynamics of the SGD behaves differently from than the averaged one and results better generalization performance [22, 39], which is called *implicit regularization* [11, 38]. We herein analyze the two instances of difference between these dynamics: dynamics around a Milnor-like attractor and attraction to a strong degenerated subspace.

The content of this chapter includes joint research [30] with Sato and Fujiwara.

7.1 Centre manifold analysis of SGD

In numerical simulations, we found that sample paths of the SGD seems quite different from trajectories obtained in the averaged gradient descent method. We carried out numerical simulations of the SGD in the same setting as Section 6.4. First, Figure 7.1 (a) shows numerical trajectories of the averaged gradient descent on the (z, u)-plane around $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. In order to approximate the averaged gradient descent sufficiently, we used the empirical distribution on a dataset of 10,000 data drawn *i.i.d.* according to $N(0, 2^2)$. Compared to the above, Figure 7.1 (b) shows sample paths of the SGD for a common input data sequence $\{x_t\}_t$. In contrast to the averaged gradient descent, in the SGD, some sample paths move from the region $\{|z| > v^*\}$ to $\{|z| < v^*\}$. Such sample paths are observed even when we use another realization of the input data sequence.

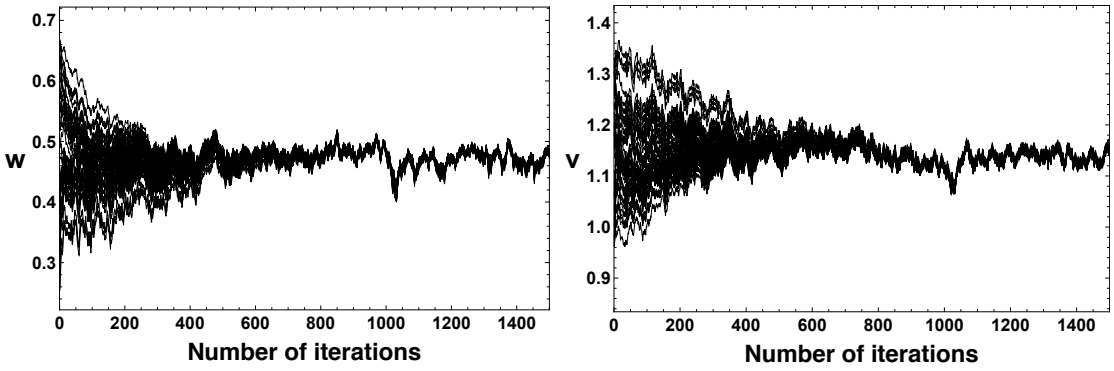


(a) Averaged gradient descent method.

(b) Stochastic gradient descent method.

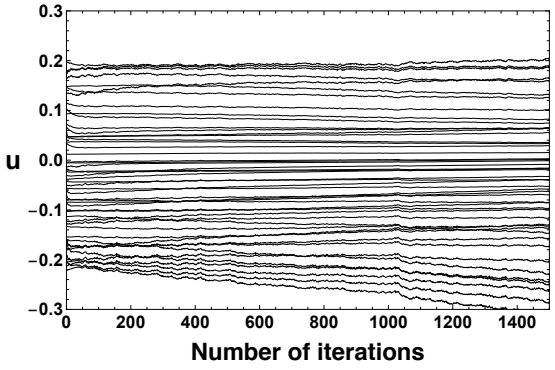
Figure 7.1: Trajectories on the (z, u)-plane obtained by the averaged gradient descent method (5.3) and the SGD (5.6) for 20,000 iterations. Red circles represent initial values.

In order to investigate this phenomenon, we observe the evolution of the parameters, again in the coordinate system (6.3). Figure 7.2 (a-d) show time evolutions of each parameter in the first 1,500 iterations of the SGD. The parameters (w, v) evolve very fast compared with (u, z) also in this case. However, in this case, (w, v) does not converge to its equilibrium point $(w^*, v^*) \approx (0.472, 1.13)$, but fluctuate stochastically around (w^*, v^*).

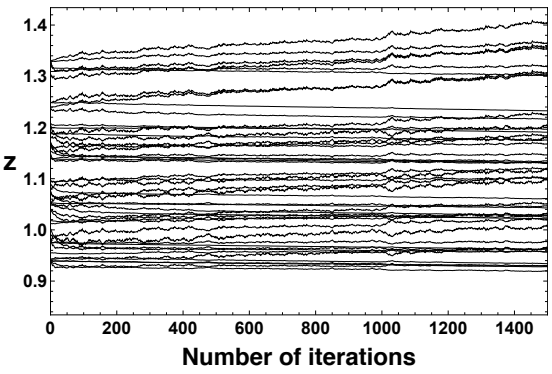


(a) Time evolutions of w.

(b) Time evolutions of v.



(c) Time evolutions of u.



(d) Time evolutions of z.

Figure 7.2: Time evolutions of each parameter for the first 1,500 iterations of the SGD. Each trajectory of w or v fluctuates intensively around the equilibrium point $w^* \approx 0.472, v^* \approx 1.134$ respectively. Trajectories of u and z evolve very slowly compared with w and v also in this case.

Based on these observations, we suppose that w and v run over sufficiently wide range of their values to be integrated while u and z move in a small range. Then, we assume that the dynamics of (u, z) is integrated with respect to (w, v) according to some stationary distribution. We further assume that (w, v) is distributed around (w^*, v^*) with finite variance. By integrating the dynamics with the "random variables" w and v, we obtain the following dynamical system near $\boldsymbol{\theta} = \boldsymbol{\theta}_1$:

$$\begin{aligned} \dot{\boldsymbol{u}} &= \frac{1}{2}(z - v^*)H\boldsymbol{u} + C_1, \\ \dot{z} &= \frac{1}{2}\boldsymbol{u}^T H\boldsymbol{u} + C_2. \end{aligned} \tag{7.1}$$

Here, C_1 and C_2 are constants resulting from the variance and covariance of (w, v). Figure 7.3 is the analytical trajectories of the dynamical system (7.1), where $C_1 = 1.71 \times 10^{-4}$ and $C_2 =$

-3.06×10^{-4} are determined heuristically. One can find that the deterministic dynamical system (7.1) gives similar trajectories to sample paths of the SGD presented in Figure 7.1 (b).

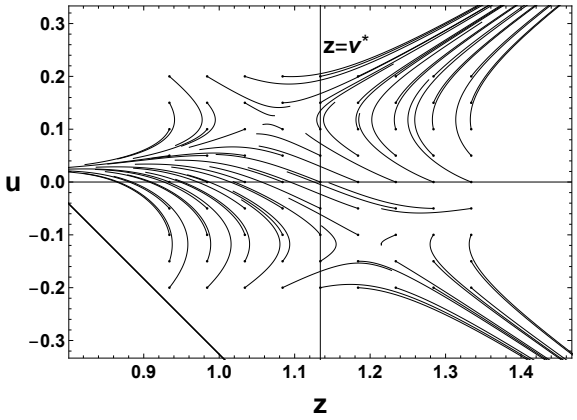


Figure 7.3: Analytical trajectories on the (z, u)-plane given by the dynamical system (7.1) with $C_1 = 1.71 \times 10^{-4}$ and $C_2 = -3.06 \times 10^{-4}$.

From the above, we deduce that a fluctuation of the parameter around a centre manifold causes constants C_1 and C_2 working as drift terms, and that it makes the SGD qualitatively different from the averaged gradient descent. This example suggests that stochastic effects can influence a macroscopic flow of the learning process via a centre manifold structure.

7.2 Noise-induced degeneration

From the viewpoint of random dynamical systems theory, we herein look into the dynamic (5.6) of the SGD. Let us consider again the dynamics around the singular region $R(\boldsymbol{w}, \boldsymbol{v})$. In this section, we impose no constraints on the dimensions of $\boldsymbol{w}$ and $\boldsymbol{v}$. We assume that ℓ is the squared loss function, and that $R(\boldsymbol{w}, \boldsymbol{v})$ has a local minimizer of L. In particular, a minimizer $\theta^* \in R(\boldsymbol{w}, \boldsymbol{v})$ of L satisfies an equilibrium condition

$$\mathbf{0} = \frac{\partial L}{\partial \boldsymbol{w}_1}(\theta^*) = \mathbb{E}_{\boldsymbol{x}}\left[(\boldsymbol{v}\varphi(\boldsymbol{w}\cdot\boldsymbol{x}) - T(\boldsymbol{x}))\,\varphi'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}\right].$$

Around the minimizer, the dynamics is attracted to $R(\boldsymbol{w}, \boldsymbol{v})$.

Here, we consider a more strongly degenerated subspace

$$R_m(\boldsymbol{w}, \boldsymbol{v}) := \left\{ \boldsymbol{\theta} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}_1, \boldsymbol{v}_2) \;\middle|\; \boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}, \boldsymbol{v}_1 = \boldsymbol{v}_2 = \frac{\boldsymbol{v}}{2} \right\} \subset R(\boldsymbol{w}, \boldsymbol{v}).$$

Note that the dynamics is closed on $R_m(\boldsymbol{w}, \boldsymbol{v})$ even with stochastic effects of the SGD. In other words, once the dynamics is on the region $R_m(\boldsymbol{w}, \boldsymbol{v})$, it cannot escape from the region even with stochastic effects.

The dynamics of the SGD shows a type of synchronization and tends to be attracted to the strongly generated subspace $R_m(\boldsymbol{w}, \boldsymbol{v})$. Using a coordinate system

$$\begin{cases} p = \dfrac{\boldsymbol{w}_1 + \boldsymbol{w}_2}{2}, & q = \dfrac{\boldsymbol{v}_1 + \boldsymbol{v}_2}{2}, \\ r = \dfrac{\boldsymbol{w}_1 - \boldsymbol{w}_2}{2}, & s = \dfrac{\boldsymbol{v}_1 - \boldsymbol{v}_2}{2}, \end{cases} \tag{7.2}$$

the singular regions are described as

$$R(\boldsymbol{w}, \boldsymbol{v}) = \{\, p = \boldsymbol{w}, q = \boldsymbol{v}, r = \boldsymbol{0} \,\},$$
$$R_m(\boldsymbol{w}, \boldsymbol{v}) = \{\, p = \boldsymbol{w}, q = \boldsymbol{v}, r = \boldsymbol{0}, s = \boldsymbol{0} \,\}.$$

We focus on the dynamics of s

$$s_{t+1} = s_t - \frac{\varepsilon}{2}\left(\frac{\partial}{\partial \boldsymbol{v}_1}\ell(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})) - \frac{\partial}{\partial \boldsymbol{v}_2}\ell(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}))\right)$$

with an *ad hoc* assumption that p, q approaches $\boldsymbol{w}, \boldsymbol{v}$ quickly. The is calculated as

$$\begin{aligned} s_{t+1} &= s_t - \frac{\varepsilon}{2}\left(\frac{\partial \ell}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{v}_1}(\boldsymbol{x};\boldsymbol{\theta}) - \frac{\partial \ell}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{v}_2}(\boldsymbol{x};\boldsymbol{\theta})\right) \\ &= s_t - \frac{\varepsilon}{2}\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}) - T(\boldsymbol{x})\right)\left(\varphi(\boldsymbol{w}_1 \cdot \boldsymbol{x}) - \varphi(\boldsymbol{w}_2 \cdot \boldsymbol{x})\right). \end{aligned}$$

By calculation, we can see that the Taylor expansion with respect to r around $\boldsymbol{0}$ is given by

$$s_{t+1} = -\varepsilon\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}) - T(\boldsymbol{x})\right)(r \cdot \boldsymbol{x})\varphi'(p \cdot \boldsymbol{x}) + \left(1 - 2\varepsilon(r \cdot \boldsymbol{x})^2\varphi'(p \cdot \boldsymbol{x})^2\right) s_t + O(\|r\|^3).$$

This implies that s_t is contracting, since the coefficient of s_t is smaller than 1. On the other hand, the constant term is, in the average, equal to the zero vector, since

$$O = \frac{\partial L}{\partial p}\Big|_{\boldsymbol{w}_1 = \boldsymbol{w}_2} = \mathbb{E}_{\boldsymbol{x}}\left[\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}) - T(\boldsymbol{x})\right)\varphi'(p \cdot \boldsymbol{x})\boldsymbol{x}\right].$$

Hence, the parameter s_t tends to approach zero in average.

This type of degeneration to $R_m(\boldsymbol{w}, \boldsymbol{v})$ is a characteristic behavior of the dynamics of the SGD. In the deterministic gradient dynamics (5.3), r approaches the zero vector monotonically in many cases. By contrast, in the SGD, r fluctuates around the zero vector, which causes the contraction of s. This phenomenon is comparable with noise-induced synchronization [29, 32] in random dynamical systems.

Since the dynamics is attracted to the sub-dynamics on $R_m(\boldsymbol{w}, \boldsymbol{v})$ in the SGD, it is more difficult to escape from the singular region compared to averaged learning. This implies that the stochastic learning process of the SGD attempts to estimate the teacher function using fewer hidden units, and thus the stagnation of learning can be much longer than in averaged learning. This can be seen as the SGD causing a more serious plateau phenomenon. On the other hand, it can also be interpreted as a spontaneous suppression of overfitting due to learning dynamics. Overfitting is caused by the large degrees of freedom of the training model relative to the size of

the training dataset. Therefore, if the degrees of freedom can be systematically adjusted during the learning process, overfitting may be avoided. In the SGD, we expect trapping in the strong degenerated singular region to result in keeping the degrees of freedom at a certain size.

Chapter 8

Concluding remarks

In this thesis, we addressed two issues; the optimal transport problem and the stochastic gradient descent machine learning. In Part I, we generalized the entropic regularization of the optimal transport problem by Cuturi [12], and studied the minimization problem of a strictly convex smooth function Φ on the set $\mathcal{P}_{nm-1}$ of joint distributions under the constraint that marginal distributions are fixed. We clarified that the solution of the problem is represented in a specific form using a pair of dual affine coordinate systems (Theorem 3.3), and proposed an iterative method for obtaining the solution. We also studied the barycenter problem [13] from an information geometric point of view and provided generalized algorithms to compute the solution of the regularized barycenter problem. As a demonstration, we showed numerically that our method works for $\Phi(P) = \langle C, P\rangle + \frac{\lambda}{2}\|P\|^2$. The framework treated in this paper is a maximal extension of Amari-Cuturi's one. Our framework subsumes some important problems, such as the Tsallis entropic regularized transport problem [25], which is represented in our framework with $\Phi(P) = \langle C, P\rangle - \lambda\mathcal{T}_{\tilde{q}}(P)$, and the quadratic regularized one on a graph [17], which corresponds to $\Phi(P) = \langle C, P\rangle + \frac{\lambda}{2}\|P\|^2$ with C being a metric matrix induced from a graph. Given the generality of our framework, it is expected that there exist many other applications. Finding another practical problem to which our framework can be applied is a future task.

We proved that the Bregman divergence associated to the convex function Φ as measured from the optimal solution monotonically decreases with our method; however, its convergence property is not revealed. For the Sinkhorn algorithm, a special case where $\Phi(P) = \langle C, P\rangle - \lambda\mathcal{H}(P)$, Franklin and Lorenz [18] studied the convergence rate, and showed the exponentially fast convergence of the sequence generated from the algorithm with respect to the Hilbert metric. However, their analysis is specialized to the Sinkhorn algorithm, and it is difficult to extend their result to a generic case, since the Hilbert metric has, to the best of our knowledge, no relation with the information geometry. For a generic Φ, evaluating the convergence rate of our method is an open problem.

The properties of numerical solutions for the choice of regularization term also remains to be investigated in future work. Compared to Cuturi's method, our algorithm using the squared regularization gives a barycenter with smaller Shannon entropy, which results in a less blurred

image in the application to image processing. However, each iteration in our algorithm demands more computational cost than Cuturi's one. We hope to find a regularization that provides a less blurred barycenter, with faster convergence and lower computational cost.

In Part II, we first gave a quick review of multilayer perceptrons, gradient descent learning, and singular regions. We explained how degeneration of hidden units gives rise to a Milnor-like attractor consisting of both attractive and repulsive parts and causes plateau phenomena in a three-layer perceptron. We next gave a review for the center manifold analysis for the gradient descent learning around a Milnor-like attractor, which is first provided in the author's thesis for master's degree. Then, we showed numerically that the dynamics of stochastic gradient descent (SGD) is qualitatively different from that of the averaged one and gave an explanation for the characteristic behavior of SGD with the aid of the center manifold analysis. We also found in our numerical experiments that learning in SGD tends to cause stronger degeneration of hidden units than averaged learning. This type of degeneration causes stagnation of learning. On the other hand, degeneration to a smaller dimensional subsystem can suppress overfitting. This observation is consistent with previous work claiming that SGD achieves higher generalization performance, which is a key concept of deep learning. Unfortunately, we only gave a conceptual explanation for the phenomena, and could not give any mathematically rigorous formulation in the present thesis. Future work should provide the theoretical groundwork for a rigid analysis of the stronger degeneration in SGD.

Appendix

A Dual problem and proof of Theorem 3.5

In this section, we give a proof of Lemma 3.2 and Theorem 3.5, and prove the existence of a dual solution (Lemma A.2). Lemma 3.2 follows from the Fenchel-Rockafellar duality, which is a prominent result in convex analysis. In order to prove Theorem 3.5 analogously to Theorem 3.3, we prepare Lemma A.3, which provides the dual problem for a 1-homogeneous Φ.

Proposition A.1 (Fenchel-Rockafellar duality)**.** *Let $\Theta, \Xi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex functions which are proper, i.e., the sets $\{\Theta(u) < +\infty\}$ and $\{\Xi(u) < +\infty\}$ are not empty. Let Θ^*, Ξ^* be their Legendre transformations, respectively. Then, the equality*

$$\inf_{u \in \mathbb{R}^n} \{ \Theta(u) + \Xi(u) \} = \sup_{A \in \mathbb{R}^n} \{ -\Theta^*(-A) - \Xi^*(A) \}$$

holds.

Proof of Lemma 3.2. Applying Proposition A.1 to the convex functions

$$\Theta(u) = \tilde{\Phi}^*(-u), \qquad \Xi(u) = \begin{cases} -\langle p, \alpha \rangle - \langle q, \beta \rangle & \text{if } u = \alpha \oplus \beta \\ +\infty & \text{otherwise} \end{cases},$$

since one can check that

$$\Theta^*(-A) = \begin{cases} \hat{\Phi}(A) & \text{if } A \in \mathbb{R}^{n \times m}_+ \\ +\infty & \text{otherwise} \end{cases},$$

$$\Xi^*(A) = \begin{cases} 0 & \text{if } \sum_{j=1}^m A_{ij} = p_i \ \sum_{i=1}^n A_{ij} = q_j \\ +\infty & \text{otherwise} \end{cases},$$

the conclusion follows, where $\hat{\Phi}$ denotes the continuous extension of $\tilde{\Phi}$ onto $\mathbb{R}^{n \times m}_+$. □

Lemma A.2. *Let $\tilde{\Phi} : \mathbb{R}^{n \times m}_{++} \to \mathbb{R}$ be a convex function and $\tilde{\Phi}^* : \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{+\infty\}$ be its Legendre transform. Then, there exists a solution (α^*, β^*) of the optimization problem*

$$\sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} \{ \langle p, \alpha \rangle + \langle q, \beta \rangle - \tilde{\Phi}^*(\alpha \oplus \beta) \} \tag{A.1}$$

for any $p \in \mathcal{P}_{n-1}$, $q \in \mathcal{P}_{m-1}$.

Proof. Let Ψ be a function on $\mathbb{R}^n \times \mathbb{R}^m$ defined by

$$\Psi(\alpha, \beta) = \tilde{\Phi}^*(\alpha \oplus \beta).$$

Then, Ψ becomes a convex function on the standard affine structure on $\mathbb{R}^n \times \mathbb{R}^m$. The problem (A.1) is no other than the Legendre transformation

$$\Psi^*(p, q) = \sup_{(\alpha,\beta)\in\mathbb{R}^n\times\mathbb{R}^m} \{\langle (p, q), (\alpha, \beta)\rangle - \Psi(\alpha, \beta)\},$$

and its supremum (α^*, β^*) is given by a subgradient of Ψ^* at (p, q). Hence, unless $\Psi^*(p, q) = +\infty$, we can conclude that (α^*, β^*) exists. Due to Lemma 3.2, we can see that

$$\Psi^*(p, q) = \inf_{P\in\Pi(p,q)} \tilde{\Phi}(P),$$

and it cannot be infinite, since $\tilde{\Phi}$ is finite-valued on $\mathbb{R}^{n\times m}_{++}$. □

Before we prove Theorem 3.5, we observe the dual problem for a 1-homogeneous function $\tilde{\Phi}$. The next lemma is a direct consequence of Lemma 3.2, and thus we omit a proof.

Lemma A.3. *Let* $\Phi : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}$ *be a 1-homogeneous convex function. For* $p \in \mathcal{P}_{n-1}$, $q \in \mathcal{P}_{m-1}$,

$$\inf_{P\in\Pi(p,q)} \Phi(P) = \sup \left\{ \langle p, \alpha\rangle + \langle q, \beta\rangle \;\middle|\; \begin{array}{l} \alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m, \\ \Phi^*(\alpha \oplus \beta) = 0 \end{array} \right\}.$$

The next result is also important when treating a 1-homogeneous convex function.

Lemma A.4. *Let* $\tilde{\Phi} : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}$ *be a 1-homogeneous convex function, and* $S : \mathbb{R}^{n\times m}_{++} \to \mathbb{R}^{n\times m}$ *be its derivative. Then,*

$$\tilde{\Phi}(A) = \langle A, S(A)\rangle \geq \left\langle A, S(\tilde{A})\right\rangle,$$

for any $A, \tilde{A} \in \mathbb{R}^{n\times m}_{++}$.

Proof. First, for any $\rho \neq 1$, since $S(A) \in \partial\tilde{\Phi}(A)$, we have

$$H(\rho A) \geq \langle \rho A - A, S(A)\rangle + H(A).$$

Then, since $\tilde{\Phi}$ is 1-homogeneous, we can write

$$(\rho - 1)H(A) \geq (\rho - 1)\left\langle A, S(A)\right\rangle.$$

We can choose $\rho \neq 1$ arbitrarily, and hence, the first equality is shown.

Second, since $S(\tilde{A}) \in \partial\tilde{\Phi}(\tilde{A})$, we have

$$\tilde{\Phi}(A) \geq \left\langle A - \tilde{A}, S(\tilde{A})\right\rangle + \tilde{\Phi}(\tilde{A})$$

for any $A \in \mathbb{R}_{++}^{n\times m}$. Then, since $\tilde{\Phi}(\tilde{A}) = \left\langle \tilde{A}, S(\tilde{A}) \right\rangle$ as seen in the above, we finally obtain

$$\tilde{\Phi}(A) \geq \left\langle A, S(\tilde{A}) \right\rangle.$$

□

Now, we arrive at the proof of Theorem 3.5.

Proof of Theorem 3.5. The proof proceeds in three steps. First, let $P^*(p,q) \in \overline{\Pi(p,q)}$ be a solution of (3.1), (α^*, β^*) be a dual solution, and we show that $\alpha^* \oplus \beta^* \in \partial\tilde{\Phi}(P^*)$. Second, we construct $P_* \in \mathcal{P}_{nm-1}$ such that $\nabla\tilde{\Phi}(P_*) = \alpha^* \oplus \beta^*$. Finally, we show that $P^*(p,q)$ and P_* are equal, and thus it is located in $\Pi(p,q)$. Here, we can assume that $\Phi(P^*(p,q)) < \infty$ without loss of generality. Thus, if necessary, let $\Phi(P^*(p,q)), S(P^*(p,q))$ denote the continuous extension of Φ, S to $P^*(p,q) \in \overline{\Pi(p,q)}$, respectively.

The existence of $P^*(p,q)$ follows from the compactness of $\overline{\Pi(p,q)}$. Let (α^*, β^*) be a solution guaranteed in Lemma A.2. Due to the duality in Lemma A.3,

$$\Phi(P^*(p,q)) = \langle p, \alpha^* \rangle + \langle q, \beta^* \rangle = \langle P^*(p,q), \alpha^* \oplus \beta^* \rangle - \tilde{\Phi}^*(\alpha^* \oplus \beta^*), \tag{A.2}$$

which implies that $\alpha^* \oplus \beta^* \in \partial\tilde{\Phi}(P^*(p,q))$.

Since $S : \mathcal{P}_{nm-1} \to \mathbb{R}^{n\times m}/\langle 1_{nm} \rangle$ is surjective by assumption, there exist $c \in \mathbb{R}$ and $P_* \in \mathcal{P}_{nm-1}$ such that

$$S(P_*) = \frac{\partial\tilde{\Phi}}{\partial A}(P_*) = \alpha^* \oplus \beta^* + c\, 1_{nm}. \tag{A.3}$$

If we assume that $c > 0$, from Lemma A.4, for $A \in \mathbb{R}_{+}^{n\times m}$,

$$\begin{aligned}\tilde{\Phi}(A) &\geq \langle A, S(P_*) \rangle \\ &= \langle A, \alpha^* \oplus \beta^* + c1_{nm} \rangle .\end{aligned}$$

Letting A tend to $P^*(p,q)$, compared with (A.2), it yields $0 \geq c$, which leads to a contradiction. On the other hand, if we assume that $c < 0$, since $\tilde{\Phi}^*(\alpha^* \oplus \beta^*) = 0$, we have

$$\begin{aligned}\langle P_*, \alpha^* \oplus \beta^* \rangle &\leq \langle P_*, S(P_*) \rangle \\ &= \langle P_*, \alpha^* \oplus \beta^* + c\, 1_{nm} \rangle ,\end{aligned}$$

which also leads to a contradiction. Hence, we obtain $c = 0$ or $S(P_*) = \alpha^* \oplus \beta^*$ from (A.3).

Finally, we show that $P^*(p,q) = P_* \in \Pi(p,q)$ by contradiction. Otherwise, since Φ is strictly convex on $\mathcal{P}_{nm-1}$, for $t \in (0,1)$,

$$\begin{aligned}\Phi(tP_* + (1-t)P^*(p,q)) &< t\Phi(P_*) + (1-t)\Phi(P^*(p,q)) \\ &= \langle tP_* - (1-t)P^*(p,q), \alpha^* \oplus \beta^* \rangle .\end{aligned}$$

On the other hand, since $\alpha^* \oplus \beta^*$ is a subgradient of $\tilde{\Phi}$ at $P^*(p,q)$,

$$\begin{aligned}
&\Phi(tP_* + (1-t)P^*(p,q)) \\
&\geq \langle (tP_* - (1-t)P^*(p,q)) - P^*(p,q), \alpha^* \oplus \beta^* \rangle + \Phi(P^*(p,q)) \\
&= \langle tP_* - (1-t)P^*(p,q), \alpha^* \oplus \beta^* \rangle .
\end{aligned}$$

Hence, we obtain a contradiction. This completes the proof. □

B Center manifold analysis for averaged gradient learning

B.1 Brief review of center manifold

We herein give a quick review of center manifold according to the textbook [10]. Suppose that a dynamical system

$$\begin{cases} \dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{y}(t)) \\ \dot{\boldsymbol{y}}(t) = B\boldsymbol{y}(t) + \boldsymbol{g}(\boldsymbol{x}(t), \boldsymbol{y}(t)) \end{cases} \tag{B.4}$$

is given, where the parameters $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, A and B are constant matrices, and $\boldsymbol{f}$ and $\boldsymbol{g}$ are C^2 functions such that they, along with their first derivatives, vanish at the origin. We assume that all the eigenvalues of A have zero real parts while all the eigenvalues of B have negative real parts. This assumption means that the parameter $\boldsymbol{y}$ converges to the origin exponentially fast, and the parameter $\boldsymbol{x}$ is driven only by the higher order terms of $\boldsymbol{f}$ and evolves very slowly compared with $\boldsymbol{y}$. Since $\boldsymbol{f}$ and $\boldsymbol{g}$ are of the second order with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$, the assumptions implies that the coefficient matrix of the linearization of the system has the form as

$$\begin{pmatrix} A & O \\ O & B \end{pmatrix}.$$

Definition B.5. *A set $S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ is said to be a local invariant manifold of* (B.4) *if for $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in S$, the solution $(\boldsymbol{x}(t), \boldsymbol{y}(t))$ of* (B.4) *with $(\boldsymbol{x}(0), \boldsymbol{y}(0)) = (\boldsymbol{x}_0, \boldsymbol{y}_0)$ is in S for $|t| < T$ with some $T > 0$.*

Definition B.6. *A local invariant manifold represented in the form of $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ is called a local center manifold (or simply a center manifold) if $\boldsymbol{h}$ is differentiable and satisfies $\boldsymbol{h}(\boldsymbol{0}) = \boldsymbol{0}$ and $\dfrac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}(\boldsymbol{0}) = O$.*

The following center manifold theorems give us a method of simplifying a dynamical system around an equilibrium point.

Proposition B.7 (Center manifold theorem 1 [10])**.** *The equation* (B.4) *has a center manifold $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ for $||\boldsymbol{x}|| < \delta$, for some $\delta > 0$ and C^2 function $\boldsymbol{h}$.*

Proposition B.8 (Center manifold theorem 2 [10]). *Suppose that the origin* $\boldsymbol{u} = \mathbf{0}$ *is a stable equilibrium point of the reduced dynamical system*

$$\dot{\boldsymbol{u}}(t) = A\boldsymbol{u}(t) + \boldsymbol{f}(\boldsymbol{u}(t), \boldsymbol{h}(\boldsymbol{u}(t))). \tag{B.5}$$

Let $(\boldsymbol{x}(t), \boldsymbol{y}(t))$ *be a solution of* (B.4) *with the initial value* $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. *Then, if* $||(\boldsymbol{x}_0, \boldsymbol{y}_0)||$ *is sufficiently small, there exists a solution* $\boldsymbol{u}(t)$ *of* (B.5) *such that*

$$\begin{aligned}\boldsymbol{x}(t) &= \boldsymbol{u}(t) + O(e^{-\gamma t}),\\ \boldsymbol{y}(t) &= \boldsymbol{h}(\boldsymbol{u}(t)) + O(e^{-\gamma t}),\end{aligned}$$

as $t \to \infty$, *where* γ *is a positive constant.*

Proposition B.8 asserts that the parameter $(\boldsymbol{x}, \boldsymbol{y})$ approaches the center manifold $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ quickly, and then evolves along it. Thus, the dynamical system (B.4) around the origin is essentially controlled by the slow parameter $\boldsymbol{x}$, and reduced to the lower dimensional system.

B.2 Proof of Theorem 6.4

In the following sections, we assume that the matrix H mentioned in Theorem 6.2 is positive definite or negative definite so that a Milnor-like attractor exists. In a column vector representation, the dynamical system (5.4) for the (n-2-1)-perceptron is written as

$$\dot{\boldsymbol{\theta}} = -\left(\frac{\partial L^{(2)}}{\partial \boldsymbol{\theta}}\right)^T.$$

By the coordinate transformation to another coordinate system $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\theta})$, the dynamical system above is transformed to

$$\dot{\boldsymbol{\xi}} = -\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}}\left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}}\right)^T\left(\frac{\partial L^{(2)}}{\partial \boldsymbol{\xi}}\right)^T. \tag{B.6}$$

Thus, the coefficient matrix of its linearization at a critical point $\boldsymbol{\xi} = \boldsymbol{\xi}^*$ is

$$\begin{aligned}\frac{\partial \dot{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}^*) &= -\frac{\partial}{\partial \boldsymbol{\xi}}\left\{\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}}\left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}}\right)^T\left(\frac{\partial L^{(2)}}{\partial \boldsymbol{\xi}}\right)^T\right\}\Bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}\\ &= -\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}}\left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}}\right)^T\frac{\partial^2 L^{(2)}}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}}(\boldsymbol{\xi}^*),\end{aligned}$$

where we used $(\partial L^{(2)}/\partial \boldsymbol{\xi})(\boldsymbol{\xi}^*) = \mathbf{0}$. This relation implies that the coefficient matrix has the same rank as the Hessian matrix $(\partial^2 L^{(2)}/\partial \boldsymbol{\xi}\partial \boldsymbol{\xi})(\boldsymbol{\xi}^*)$. In particular, the rank of the coefficient matrix of the linearization does not depend on the choice of a coordinate system.

To prove the Theorem 6.4, we make use of the following lemma.

Lemma B.9. *If the matrix* X *is positive definite and* Y *is positive semi-definite, all the eigenvalues of the matrix* XY *are non-negative.*

Proof. The matrix XY is rewritten as

$$XY = X^{\frac{1}{2}}(X^{\frac{1}{2}}Y\,X^{\frac{1}{2}})X^{-\frac{1}{2}},$$

where $X^{\frac{1}{2}}$ is a unique positive definite matrix such that $(X^{\frac{1}{2}})^2 = X$. Here, the matrix $Z := X^{\frac{1}{2}}Y\,X^{\frac{1}{2}}$ is positive semi-definite. Hence, for each eigenvector $\boldsymbol{a}$ of Z, the vector $X^{\frac{1}{2}}\boldsymbol{a}$ is an eigenvector of the matrix XY, and the corresponding eigenvalue is non-negative. □

Then, Theorem 6.4 is shown as follows.

Proof of Theorem 6.4. The proof is essentially based on a straightforward calculation. The coefficient matrix of the linearization of the dynamical system (B.6) under the coordinate system $\boldsymbol{\xi}$ defined by (6.3) splits into $(\boldsymbol{w}, v)$ and $(\boldsymbol{u}, z)$ parts at $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$. In fact, the negative of the coefficient matrix is written as

$$\tilde{A}_0 = \tilde{A}_1 = \begin{array}{c} \\ \boldsymbol{w}, v \\ \boldsymbol{u} \\ z \end{array} \begin{array}{c} \overbrace{}^{\boldsymbol{w},v} \quad \overbrace{}^{\boldsymbol{u}} \quad \overbrace{}^{z} \\ \left\{ \left(\begin{array}{cccc} Q+H & P & 0 & 0 \\ 2P^T & 2R & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) \right. \end{array}. \tag{B.7}$$

and thus the system (B.6) is written as

$$\dot{\boldsymbol{\xi}} = -\tilde{A}_\lambda(\boldsymbol{\xi} - \boldsymbol{\xi}_\lambda) + \tilde{\boldsymbol{g}}_\lambda(\boldsymbol{\xi}), \quad \lambda = 0, 1,$$

where $\tilde{\boldsymbol{g}}_\lambda$ is the higher order term, which vanish at the $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$ together with its first derivative. Here,

$$\begin{aligned} P &:= \mathbb{E}_{\boldsymbol{x}}\left[(\partial^2\ell)\, v^*\, \varphi(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}\right], \\ Q &:= \mathbb{E}_{\boldsymbol{x}}\left[(\partial^2\ell)\, (v^*)^2\, \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})^2\, \boldsymbol{x}\, \boldsymbol{x}^T\right], \\ R &:= \mathbb{E}_{\boldsymbol{x}}\left[(\partial^2\ell)\, \varphi(\boldsymbol{w}^* \cdot \boldsymbol{x})^2\right], \\ \partial\ell &:= \frac{\partial\ell}{\partial y}(\boldsymbol{x}, f^{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*)), \\ \partial^2\ell &:= \frac{\partial^2\ell}{\partial y^2}(\boldsymbol{x}, f^{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*)), \end{aligned}$$

and H is the matrix defined by (6.1). Remark that H and Q are matrices, P is a column vectors, and R is a scalar.

Then, we show that all the eigenvalues of $(\boldsymbol{w}, v)$-block of the coefficient matrix $-\tilde{A}_0$ are strictly negative. Recall that the coefficient matrix at a critical point $\boldsymbol{\xi}^*$ of the dynamical system (5.4) is given by

$$-\frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\theta}}\left(\frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\theta}}\right)^T \frac{\partial^2 L^{(2)}}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}}(\boldsymbol{\xi}^*).$$

Applying Lemma B.9 to $X = (\partial\boldsymbol{\xi}/\partial\boldsymbol{\theta})(\partial\boldsymbol{\xi}/\partial\boldsymbol{\theta})^T$ and $Y = (\partial^2 L^{(2)}/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi})(\boldsymbol{\xi}_0)$, all the eigenvalues of the coefficient matrix $-\tilde{A}_0$ are non-positive. One can check that the Hessian matrix is positive semi-definite and has rank $n+2$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$, due to the assumption that $(\boldsymbol{w}^*, v^*)$ is a strict local minimizer of $L^{(1)}$. The coefficient matrix $-\tilde{A}_0$ has the same rank as the Hessian matrix, which implies that the $(\boldsymbol{w}, v)$-block is of full-rank. Hence, all the eigenvalues of $(\boldsymbol{w}, v)$-block are strictly negative. At $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, a similar assertion for $-\tilde{A}_1$ holds.

Finally, due to Proposition B.7, there are center manifolds parametrized by $(\boldsymbol{u}, z)$ around $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ respectively. This completes the proof. □

B.3 Reduced dynamical system

By virtue of Proposition B.8 and Theorem 6.4, we can assume that the dynamics (5.4) of the gradient descent is on the center manifold near the points $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$. Thus, we can reduce the dynamical system into that of $(\boldsymbol{u}, z)$. Recalling the coefficient matrix (B.7), we can see that $\dot{\boldsymbol{u}}$ and $\dot{z}$ have no first order terms. In more detail, calculating the Taylor expansion of $(\dot{\boldsymbol{u}}, \dot{z})$ up to the second order around $\boldsymbol{\xi} = \boldsymbol{\xi}_1$, we obtain

$$\begin{aligned}
\dot{\boldsymbol{u}} = \frac{1}{v^*}\Big\{ & -(P \cdot (\boldsymbol{w} - \boldsymbol{w}^*))(\boldsymbol{u} + (\boldsymbol{w} - \boldsymbol{w}^*)) \\
& - (v - v^*)(R\, I + \frac{1}{2}H)(\boldsymbol{u} + (\boldsymbol{w} - \boldsymbol{w}^*)) \\
& + \frac{1}{2}(z - v^*)H(\boldsymbol{u} + (\boldsymbol{w} - \boldsymbol{w}^*))\Big\} + O(||\boldsymbol{\xi} - \boldsymbol{\xi}_1||^3), \qquad \text{(B.8)} \\
\dot{z} = \frac{1}{v^*}\Big\{ & -(\boldsymbol{w} - \boldsymbol{w}^*)^T Q(\boldsymbol{u} + (\boldsymbol{w} - \boldsymbol{w}^*)) \\
& - (v - v^*)(P \cdot (\boldsymbol{u} + (\boldsymbol{w} - \boldsymbol{w}^*))) \\
& - \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^T H(\boldsymbol{w} - \boldsymbol{w}^*) + \frac{1}{2}\boldsymbol{u}^T H \boldsymbol{u}\Big\} + O(||\boldsymbol{\xi} - \boldsymbol{\xi}_1||^3), \qquad \text{(B.9)}
\end{aligned}$$

where I denotes the $(n+1) \times (n+1)$ identity matrix. Now we consider the reduced dynamical system on the center manifold. Here, the center manifold $(\boldsymbol{w}, v) = \boldsymbol{h}(\boldsymbol{u}, z)$ satisfies that

$$\boldsymbol{h}(\boldsymbol{u}, z) = \begin{bmatrix} \boldsymbol{w}(\boldsymbol{u}, z) \\ v(\boldsymbol{u}, z) \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^* \\ v^* \end{bmatrix} + O(\|\boldsymbol{u}, z - v^*\|^2),$$

by definition. This gives an approximation of the dynamics on the center manifold near $\boldsymbol{\xi} = \boldsymbol{\xi}_1$ as

$$\begin{aligned}
\dot{\boldsymbol{u}} &= \frac{1}{2v^*}\,(z - v^*)H\boldsymbol{u} + O(\|\boldsymbol{u}, z - v^*\|^3), \\
\dot{z} &= \frac{1}{2v^*}\,\boldsymbol{u}^T H \boldsymbol{u} + O(\|\boldsymbol{u}, z - v^*\|^3). \qquad \text{(B.10)}
\end{aligned}$$

Neglecting the higher order terms, we can integrate this equation to obtain

$$\|\boldsymbol{u}\|^2 = (z - v^*)^2 + C, \qquad \text{(B.11)}$$

where C is an integral constant.

Around the point $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, we obtain the similar dynamics

$$\dot{\boldsymbol{u}} = -\frac{1}{2v^*}(z + v^*)H\boldsymbol{u} + O(\|\boldsymbol{u}, z + v^*\|^3),$$
$$\dot{z} = -\frac{1}{2v^*}\boldsymbol{u}^T H\boldsymbol{u} + O(\|\boldsymbol{u}, z + v^*\|^3).$$

and the relation

$$\|\boldsymbol{u}\|^2 = (z + v^*)^2 + C.$$

We remark that Theorem 6.4 is valid even when there exists a true parameter in the singular region $R(\boldsymbol{w}^*, v^*)$; however, in this case, such a simple form of the reduced dynamical system as (B.10) is not obtained. Since this case implies that H becomes the zero matrix, the second order terms of the reduced dynamical system (B.10) vanish, and the third order terms become dominant. Thus, we have to take into account the cross terms between $(\boldsymbol{w} - \boldsymbol{w}^*, v - v^*)$ and $(\boldsymbol{u}, z - v^*)$. It needs to calculate the center manifold $(\boldsymbol{w}, v) = \boldsymbol{h}(\boldsymbol{u}, z)$ up to the second order, which makes the analysis complicated.

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] S.-i. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[3] S.-i. Amari, R. Karakida, and M. Oizumi. Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem. *Information Geometry*, 1:13–37, 2018.

[4] S.-i. Amari, R. Karakida, M. Oizumi, and M. Cuturi. Information geometry for regularized optimal transport and barycenters of patterns. *Neural Computation*, 31(5):827–848, 2019.

[5] S.-i. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, 3(2):260–271, 1992.

[6] S.-i. Amari and H. Nagaoka. *Method of information geometry.* American Mathematical Society, 2000.

[7] S.-i. Amari, T. Ozeki, R. Karakida, Y. Yoshida, and M. Okada. Dynamics of learning in MLP: Natural gradient and singularity revisited. *Neural Computation*, 30(1):1–33, 2018.

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 70:214–223, 2017.

[9] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transport problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[10] J. Carr. *Applications of centre manifold theory.* Springer New York, 1981.

[11] P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *In International Conference on Learning Representations (ICLR)*, 2018.

[12] M. Cuturi. Sinkhorn distances: light speed computation of optimal transport. *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, 2:2292–2300, 2013.

[13] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 32(2):685–693, 2014.

[14] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[15] A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

[16] B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, 3(6):1189–1242, 1975.

[17] M. Essid and J. Solomon. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

[18] J. N. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.

[19] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein loss. *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, 2:2053–2061, 2015.

[20] K. Fukumizu and S.-i. Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.

[21] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[22] S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, A. Storkey, and Y. Bengio. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

[23] H. Lavenant, S. Claici, E. Chien, and J. Solomon. Dynamical optimal transport on discrete surfaces. *ACM Transactions on Graphics*, 37(6):1–16, 2018.

[24] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[25] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen. Tsallis regularized optimal transport and ecological inference. *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

[26] O. Pele and M. Werman. Fast and robust earth mover's distances. *In 2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pages 460–467, 2009.

[27] G. Peyré and M. Cuturi. Computational optimal transport: with applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

[28] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37(3):81–91, 1945.

[29] Y. Sato, D. T. Son, N. T. The, and H. T. Tuan. An analytical proof for synchronization of stochastic phase oscillator. *arXiv preprint arXiv:1801.02761*, 2018.

[30] Y. Sato, D. Tsutsui, and A. Fujiwara. Noise-induced degeneration in online learning. *Physica D: Nonlinear Phenomena*, 430:133095, 2022.

[31] S. Sonoda and N. Murata. Transport analysis of infinitely deep neural network. *Journal of Machine Learning Research*, 20(2):1–52, 2019.

[32] J.-n. Teramae and D. Tanaka. Robustness of the noise-induced phase synchronization in a general class of limit cycle oscillators. *Physical Review Letters*, 93:204103, 2004.

[33] D. Tsutsui. Center manifold analysis of plateau phenomena caused by degeneration of three-layer perceptron. *Neural Computation*, 32(4):683–710, 2020.

[34] N. N. Čencov. *Statistical decision rules and optimal inference*. American Mathematical Society, 1981.

[35] C. Villani. *Topics in optimal transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

[36] M.-K. von Renesse and K.-T. Strum. Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on Pure and Applied Mathematics*, 58(7):923–940, 2005.

[37] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, and S.-i. Amari. Dynamics of learning near singularities in layered networks. *Neural Computation*, 20(3):813–843, 2008.

[38] L. Wu and W. J. Su. The implicit regularization of dynamical stability in stochastic gradient descent. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 37656–37684, 2023.

[39] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The anisotropic noise in stochastic gradient descent: its behavior of escaping from sharp minima and regularization effects. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7654–7663, 2019.