关于对抗鲁棒性及利用Wasserstein上升-下降动力学实现对抗鲁棒性

´ ´ ´ ´CAMILO ANDRES GARCIA TRILLOS 和 NICOLAS GARCIA TRILLOS

摘要。我们提出了迭代算法，用于解决多种值得关注的监督学习情形中的对抗问题。我们的算法可被解释为Wasserstein空间中的适当上升-下降动力学，其形式为一个相互作用粒子系统。研究表明，在粒子数量足够大的某些情形下，这些相互作用粒子动力学会收敛到相应的平均场极限方程。进一步地，我们证明，在满足一定正则性假设的条件下，这些平均场方程在长时间极限下会收敛到原始对抗学习问题的近似纳什均衡。我们给出了非凸非凹情形以及非凸凹情形下的结果。数值实验验证了我们的结果。

arXiv:2301.03662v1 [cs.LG] 9 Jan 2023

1. 引言

1.1. 问题。 在许多机器学习任务中，人们都需要求解概率空间上的极小极大问题，其形式如下

$$\min_{\nu\in\mathcal{P}(\Theta)} \max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} R(\tilde{\mu},\nu) - C(\mu,\tilde{\mu}). \tag{1.1}$$

例如，(1.1) 可以看作如下形式的（参数化）分布鲁棒优化（DRO）问题的泛函松弛：

$$\min_{\theta\in\Theta} \max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \hat{R}(\tilde{\mu},\theta) - C(\mu,\tilde{\mu}). \tag{1.2}$$

这里， Θ 表示学习模型的参数空间，例如分类器或回归函数； $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 是输入到输出样本的空间； R 是用于约束学习问题的风险函数； C 表示对手为将输入到输出的原始数据分布 μ 更改为新分布 $\tilde{\mu}$必须付出的成本 。简而言之，问题 (1.1) 和 (1.2) 都可以解释为学习者与对手之间进行的博弈：对于学习者而言，目标是选择一个参数/参数分布，使其在面对合理对手的攻击时仍能表现良好（对手的合理性由成本函数 C 建模）。在此背景下，从 (1.2) 到 (1.1) 的松弛过程可以通过多种合理方式实现。例如， R 可以定义为风险 R 的平均值，即$R(\tilde{\mu}, \nu) = \int_\Theta \hat{R}(\tilde{\mu}, \theta)d\nu(\theta)$， 也可以通过考虑某种平均控制的值函数来定义，即 $R(\tilde{\mu}, \nu) = \hat{R}\left(\tilde{\mu}, \int_\Theta \theta d\nu(\theta)\right)$. 这两类松弛方法均可用于改善原始目标函数在参数空间 Θ上的凸性，或者从方法论角度将其用作模型聚合技术。其他问题也可以直接用 (1.1) 建模，例如浅层神经网络的对抗训练，这也是本研究的一个动机示例；见下文第 1.2 节。

本文的总体目标是提出并分析上升-下降动力学，在对风险 R 和成本 C 作出合理假设的前提下，为问题 (1.1) 寻找近似解（解释为纳什均衡）。这些假设具有足够的普适性，能够涵盖应用中的重要场景，尤其是对抗机器学习。这些动力学将采用相互作用粒子系统的形式，我们尤其关注随着系统中的粒子数量增加，这些系统的行为如何变化。我们还将讨论该

关键词和短语。对抗鲁棒性、对抗训练、极小极大博弈、纳什均衡、Wasserstein梯度流、Wasserstein-Fisher-Rao度量、相互作用粒子系统、平均场极限。

极限平均场动力学的长期行为，并探究其能否为 (1.1) 产生近似纳什均衡。

为写出本文将要分析的具体粒子系统，并讨论其背后的几何结构为何会促使我们研究该系统，重新表述问题 (1.1) 会更为方便。我们假设 C 具有最优传输问题的结构；将在下文第 1.1.1 节讨论这种重构。同样，介绍一个具体的工作示例来说明问题 (1.1) 的动机也会很有帮助。具体而言，我们将赋予学习参数空间 Θ以明确含义，并定义特定的风险泛函 R和成本函数 C；我们将在下文第 1.2 节完成这些工作。

1.1.1. 用耦合重新表述问题 π。 本文将始终假设，式 (1.1) 中的对抗代价 C 可以写成如下形式的最优传输问题：

$$C(\mu,\tilde{\mu}) := \inf_{\pi\in\Gamma(\mu,\tilde{\mu})} \int c(z,\tilde{z})d\pi(z,\tilde{z}), \tag{1.3}$$

其中，$c: \mathcal{Z}\times\mathcal{Z}\to[0,\infty]$ 是相对于$\mathcal{Z}$所默认赋予的拓扑而言的下半连续成本函数，而 $\Gamma(\mu,\tilde{\mu})$ 表示 μ 与 $\tilde{\mu}$之间的耦合空间，即乘积空间 $\mathcal{Z}\times\mathcal{Z}$ 上的概率测度空间，其第一和第二边缘分布分别为 μ 和 $\tilde{\mu}$。成本函数 c 可以解释为对手将数据样本 z 移至新位置 $\tilde{z}$所需支付的边际成本。

(1.3) 中的成本结构相当宽泛，涵盖了文献中若干对抗鲁棒性模型。此外，当 C 具有 (1.3) 的结构时，(1.1) 中的优化变量$\tilde{\mu}$，即对应于（全局）对抗攻击的变量，可以替换为耦合变量 π，如 (1.3) 所示。为说明这一点，令 $\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})$，并定义

$$\mathcal{C}(\pi) := \int_{\mathcal{Z}\times\mathcal{Z}} c(z,\tilde{z})d\pi(z,\tilde{z})$$

我们用 π_z 和 $\pi_{\tilde{z}}$ 分别表示 π 的第一和第二个边缘分布。给定$\nu\in\mathcal{P}(\Theta)$ 和 $\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})$，我们定义：

$$\mathcal{U}(\pi,\nu) := \mathcal{R}(\pi,\nu) - \mathcal{C}(\pi), \tag{1.4}$$

其中 $\mathcal{R}(\pi,\nu) := R(\pi_{\tilde{z}},\nu)$，并且我们隐含地假设上述差值有意义，即当 $\mathcal{R}(\pi,\nu) = \mathcal{C}(\pi) = \infty$ 时，将 (π,ν) 排除在 $\mathcal{U}$ 的定义域之外。于是，问题 (1.1) 可等价地重构为：

$$\min_{\nu\in P(\Theta)} \max_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z});\pi_z=\mu} \mathcal{U}(\pi,\nu). \tag{1.5}$$

事实上，很容易看出，如果 (π^*,ν^*) 是问题 (1.5) 的纳什均衡，那么 $(\pi^*_{\tilde{z}},\nu^*)$ 就是问题 (1.1) 的纳什均衡。反过来，如果 $(\tilde{\mu}^*,\nu^*)$ 是问题 (1.1) 的解，那么 (π^*,ν^*) 是问题 (1.5) 的解，其中 π^* 是问题 (1.3) 在 $\tilde{\mu}=\tilde{\mu}^*$ 条件下的一个最优解。

重构 (1.5) 和变量替换 $\tilde{\mu}\mapsto\pi$ 使我们能够有效地用线性泛函 $\mathcal{C}(\cdot)$ 替代非线性目标函数 $C(\mu,\cdot)$。这种替换有助于定义我们的算法，在某些情况下还会产生凹性更好的极小极大问题（例如，参见第 1.4.1 节）。因此，我们将重点讨论重构 (1.5) 以及旨在求解该问题的上升-下降动力学。从概念上讲，(1.5) 中的变量 ν 指定了一个回归/分类函数（即学习者采取的行动/策略），而 π 则决定了对手修改原始数据的方式；更多细节见下文第 1.2 节。问题 (1.5) 可解释为一种带显式惩罚项而非显式约束的对抗训练分布鲁棒优化（DRO）版本；参见[45]。

1.2. 启发性示例：采用浅层神经网络的稳健回归或二分类。在继续介绍我们的算法和主要结果之前，先给出一些实际场景中值得关注的风险泛函 $\mathcal{R}$ 和成本函数$\mathcal{C}$ 的具体示例会很有帮助。本节将对此进行介绍。第 5 节将讨论另一种情形。

令 $\Theta \subseteq \mathbb{R} \times \mathbb{R}^{d'}$、$\mathcal{Z} = \mathbb{R}^{d'} \times \mathbb{R}$，并记 $\theta = (a, b)$ 和 $z = (x, y)$。我们考虑以下风险和对抗代价：

$$\mathcal{R}(\pi, \nu) := \int_{\mathcal{Z}\times\mathcal{Z}} \ell(h_\nu(\tilde{x}), \tilde{y}) d\pi(z, \tilde{z}), \quad h_\nu(x) := \int_\Theta a f(b \cdot x) d\nu(a, b), \tag{1.6}$$

其中 $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty]$ 是损失函数，f 是激活函数，并且

$$\mathcal{C}(\pi) := c_a \int_{\mathcal{Z}\times\mathcal{Z}} |z - \tilde{z}|^2 d\pi(z, \tilde{z}), \tag{1.7}$$

对于 c_a 这一正参数情形，采用平方欧氏距离，这是成本函数 c 的众多可能选择之一。

注意，函数 h_ν 可以解释为一个无偏置参数的两层（单隐藏层）神经网络的连续体类比（为简便起见）；参见 [14]。常见的激活函数示例如下：

i) Sigmoid：$f(t) = \frac{1}{1+e^{-t}}$。
ii) ReLU：$f(t) = (t)$+iii) 平方
ReLU：$f(t) = (t_+)^2$。
值得关注的损失函数示例如下：

i) 平方损失：$\ell(y, y') = |y - y'|^2$。ii) 逻辑损失：$\ell(y, y') = -\log(|1 - y' - y|)$，其中 y, y' 受限于属于 $[0, 1]$。

对抗代价中的参数 c_a 可解释为对抗预算的倒数，它决定了对抗攻击的强度。具体而言，当 c_a 较小时，攻击者可以实施更强的攻击；而当 c_a 较大时，情况则相反。在应用中，求解如 (1.5) 这样的问题，旨在提升学习模型的鲁棒性。换言之，除了简单地求解形如 $\min_\nu \mathcal{R}(\pi, \nu)$ 的优化问题（这对应于标准训练过程）之外，人们还会考虑问题 (1.5)，以将对手可能采取的行动纳入考量。然而，为避免以准确率大幅下降为代价来提升鲁棒性，适当调整对抗预算至关重要。研究鲁棒性与准确率之间权衡的一些论文包括 [47, 56]。

注1.1。在上述设定下，使用平方损失函数$\ell(y, y') = |y - y'|^2$时，损失函数 $\ell(h_\nu(\tilde{x}), \tilde{y})$ 展开平方后可写为

$$\ell(h_\nu(\tilde{x}), \tilde{y}) = \int_\Theta \int_\Theta a f(b \cdot \tilde{x}) a' f(b' \cdot \tilde{x}) d\nu(a', b') d\nu(a, b) - 2\tilde{y} \int_\Theta a f(b \cdot \tilde{x}) d\nu(a, b) + \tilde{y}^2.$$

特别地，我们可以写成

$$\begin{aligned}\mathcal{R}(\pi, \nu) = &\int_\Theta \int_\Theta \left(\int_{\mathcal{Z}^2} a f(b \cdot \tilde{x}) a' f(b' \cdot \tilde{x}) d\pi(z, \tilde{z}) \right) d\nu(a', b') d\nu(a, b) \\ &- 2 \int_\Theta \int_{\mathcal{Z}^2} \tilde{y} a f(b \cdot \tilde{x}) d\pi(z, \tilde{z}) d\nu(a, b) + \int_{\mathcal{Z}^2} \tilde{y}^2 d\pi(z, \tilde{z})\end{aligned}$$

这个表达式的优点是明确揭示了风险对测度的依赖关系 ν。

算法1 Wasserstein 上升-下降算法

要求： 一个集合 $\{z_{i,\ 0},\ \omega_{i,\ 0}\}_{i=1,\ \ldots,\ n}$ 使得 $\frac{1}{n}\sum_{i=1}^{n}\omega_{i,\ 0}\delta_{z_{i,\ 0}}$ 近似于 μ.

1: 设定 $t = 0$ 2: 选择 $\{\vartheta_{k,0}\}_{k=1,\ldots M}$、$\{\alpha_{k,0}\}_{k=1,\ldots,M}$、$\{\tilde{z}_{ij,0}\}_{i=1,\ldots,n;\ j=1,\ldots N}$ 以及 $\{\omega_{ij,0}\}_{i=1,\ldots,n;\ j=1,\ldots N}$，满足约束

$$\sum_{j=1}^{N}\omega_{ij,0} = \omega_{i,0} \text{ for all } i = 1,\ldots n.$$

3: **当** 尚未满足停止条件 **时** 4: 设定

$$\pi_t^{n,N} := \sum_{i=1}^{n}\sum_{j=1}^{N}\omega_{ij,t}\delta_{(z_{i,0},\tilde{z}_{ij,t})} \qquad \nu_t^{M} := \sum_{k=1}^{M}\alpha_{k,t}\delta_{\vartheta_{k,t}}$$

5: **对** $i = 1$ 至 n； $j = 1$ 至 N **执行**

6: $\tilde{z}_{ij,t+1} = P_{\mathcal{Z}}\left(\tilde{z}_{ij,t} + \eta_t\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi_t,\nu_t;z_{i,0}\tilde{z}_{ij,t})\right)$ $\hat{\omega}_{ij,\ t+1} := \omega_{ij,\ t}\exp\left(\eta_t\kappa\sum_{j'}\omega_{ij',\ t}\mathcal{U}_\pi(\pi_t,\ \nu_t,\ z_{i,\ 0},\right.$

7: $\left.\tilde{z}_{ij,\ t})\right)$

8: $\omega_{ij,t+1} := \frac{\hat{\omega}_{ij,t+1}}{\sum_{j'}\hat{\omega}_{ij',t+1}}$

9: **结束循环**

10: **对** $k = 1$ 至 M **执行**

11: $\vartheta_{k,t+1} = P_\Theta\left(\vartheta_{k,t} - \eta_t\nabla_\theta\mathcal{U}_\nu(\pi_t,\nu_t;\vartheta_{k,t})\right)$

12: $\hat{\alpha}_{k,\ t+1} := \alpha_{k,\ t}\exp\left(-\eta_t\kappa\sum_{k'}\alpha_{k',\ t}\mathcal{U}_\nu(\pi_t,\ \nu_t;\vartheta_{k,\ t})\right)$

13: $\alpha_{k,t+1} := \frac{\hat{\alpha}_{k,t+1}}{\sum_{k'}\hat{\alpha}_{k',t+1}}$

14: **结束 for 循环**

15: $t = t + 1$

16: **结束 while 循环**

17: ****计算时间平均值****

18: $\bar{z}_{ij} := \frac{1}{t}\sum_{s=0}^{t}\omega_{ij,\ s}\tilde{z}_{ij,\ s}$ 对于 $i = 1,\ \ldots\ldots,\ n;\ \ j = 1,\ \ldots\ldots,\ N$

19: $\bar{\vartheta}_k := \frac{1}{t}\sum_{s=0}^{t}\alpha_{k,s}\tilde{\vartheta}_{k,s}$ 对于 $k = 1,\ \ldots\ldots,\ M$

1.3. 算法。我们在算法1中引入一种基于粒子的方案，用于求解极小极大问题（1.5）。算法1的定义隐含地使用了泛函 $\mathcal{U}$ 沿方向 ν 和 π 的一阶变分。

根据 [42], 中的定义7.12，我们称可测函数 $\mathcal{U}_\pi : \mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ 为泛函 $\mathcal{U}$ 在点（π，ν）沿方向 π 的一阶变分，若对于任意 $\pi^* \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})$，都有

$$\frac{d}{d\epsilon}(\mathcal{U}(\pi + \epsilon(\pi^* - \pi),\nu))\Big|_{\epsilon=0} = \int_{\mathcal{Z}\times\mathcal{Z}}\mathcal{U}_\pi(\pi,\nu;z,\tilde{z})d(\pi^* - \pi).$$

一般而言，$\mathcal{U}_\pi$ 可能取决于计算一阶变分时所在的点（π，ν），但只要这样做不会引起混淆，我们就省略对这种依赖关系的明确说明；否则，我们会像这样写出 $\mathcal{U}_\pi$ 的所有自变量：$\mathcal{U}_\pi(\pi,\ \nu;\ z,\ \tilde{z})$。同样，若对于任意 $\nu^* \in \mathcal{P}(\Theta)$，均有下式成立，我们就称可测函数 $\mathcal{U}_\nu : \Theta\to\mathbb{R}$ 是 $\mathcal{U}$ 在点（π，ν）沿方向 ν 的一阶变分：

$$\frac{d}{d\epsilon}(\mathcal{U}(\pi,\nu + \epsilon(\nu^* - \nu)))\Big|_{\epsilon=0} = \int_\Theta\mathcal{U}_\nu(\pi,\nu;\theta)d(\nu^* - \nu).$$

全文中，我们假设 $\mathcal{U}$ 的一阶变分均有良好定义，并满足假设 1.4 中明确陈述的某些正则性条件。

在算法1中，η_t 可以解释为随时间变化的学习率，而 κ 则是一个固定参数。引入投影映射 $P_{\mathcal{Z}}$ 和 P_Θ，以确保迭代点始终处于集合 $\mathcal{Z}$ 和 Θ 内。第18–19行中的平均步骤将在第5节讨论；算法1与 [13, 51]中介绍的算法有关；第1.5节将比较那些论文与本文的内容。

算法1中的迭代点集合可以看作是以下常微分方程系统的时间离散化：

$$
\begin{aligned}
dZ_t^i &= 0 \\
d\tilde{Z}_t^i &= \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) dt \\
d\omega_t^i &= \kappa \omega_t^i \left(\mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) - \int \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{z}') d\pi_t^N(\tilde{z}'|Z_t^i) \right) dt \\
d\vartheta_t^i &= -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \vartheta_t^i) dt \\
d\alpha_t^i &= -\kappa \alpha_t^i \left(\mathcal{U}_\nu(\pi_N^N, \nu_t^N; \vartheta_t^i) - \int \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta') d\nu_t^N(\theta') \right) dt,
\end{aligned} \tag{1.8}
$$

给定初始条件（可能是随机的）$(Z_0^i,\ \tilde{Z}_0^i,\ \omega_0^i,\ \vartheta_0^i,\ \alpha_0^i)$，并且

$$
\pi_t^N := \frac{1}{N} \sum_{i=1}^N \omega_t^i \delta_{(Z_t^i, \tilde{Z}_t^i)}, \qquad \nu_t^N := \frac{1}{N} \sum_{i=1}^N \alpha_t^i \delta_{\vartheta_t^i}. \tag{1.9}
$$

请注意，在上述常微分方程中，以及在第 3 节的分析中，我们都考虑了~相同数量的粒子 Z，Z，ϑ 并且省去了双重索引。这样做是为了简化表达，并减少整个分析过程中繁琐的记号。只有在必要时，我们才会重新使用双重索引。

我们尤其关注测度 π_t^N，ν_t^N的演化，可以证明它们满足以下偏微分方程（弱形式）：

(1.10)

$$
\begin{cases}
\partial_t \pi_t &= -\eta_t \mathrm{div}_{z,\tilde{z}}(\pi_t(0, \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}))) + \kappa \pi_t \left(\mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}) - \int \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}') d\pi_t(\tilde{z}'|z) \right) \\
\partial_t \nu_t &= \eta_t \mathrm{div}_\theta(\nu_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \theta)) - \kappa \nu_t \left(\mathcal{U}_\nu(\pi_t, \nu_t; \theta) - \int \mathcal{U}_\nu(\pi_t, \nu_t; \theta') d\nu_t(\theta') \right),
\end{cases}
$$

初始条件为

$$
\pi_0 = \pi_0^N = \sum_{i=1}^N \omega_0^i \delta_{(Z_0^i, \tilde{Z}_0^i)}, \quad \nu_0 = \nu_0^N = \sum_{i=1}^N \alpha_0^i \delta_{\vartheta_0^i}. \tag{1.11}
$$

在我们的第一个主要结果——下文的定理 1.8——中，我们指出，在$\mathcal{U}$ 及其一阶变分$\mathcal{U}_\pi$ 和 $\mathcal{U}_\nu$ 满足某些条件时——这些条件最终由应用中风险函数$\mathcal{R}$ 和成本函数$\mathcal{C}$ 的选择决定——，以（1.11）所述方式初始化的系统（1.10）的（$N \to \infty$）大粒子数极限遵循相同的动力学（1.10），但具有更一般的初始化；作为第 3 节分析的一部分，我们证明该系统是适定的。第 2 节将讨论使用梯度上升-下降一词描述系统（1.10），并进一步描述算法1中的步骤的动机。

例1.2。 在第*1.2*节的启发性示例中，我们看到：

$$
\mathcal{U}_\pi(\pi, \nu; z, \tilde{z}) = \ell(h_\nu(\tilde{x}), \tilde{y}) - c(z, \tilde{z}) = \ell(h_\nu(\tilde{x}), \tilde{y}) - c_a |z - \tilde{z}|^2, \tag{1.12}
$$

and

$$
\mathcal{U}_\nu(\pi, \nu; \theta) = \int_{\mathcal{Z} \times \mathcal{Z}} \ell'(h_\nu(\tilde{x}), \tilde{y})(a f(b \cdot \tilde{x})) d\pi(z, \tilde{z}), \tag{1.13}
$$

其中 $\theta = (a, b)$。这里，ℓ' 表示 ℓ 关于其第一坐标的导数。

对于平方损失的情况，U_ν 可以改写为：

$$
\begin{aligned}
\mathcal{U}_\nu(\pi, \nu; \theta) =& 2 \int_{\mathcal{Z} \times \mathcal{Z}} \int_\Theta (a' f(b' \cdot \tilde{x}))(a f(b \cdot \tilde{x})) d\nu(\theta') d\pi(z, \tilde{z}) \\
& - 2 \int_{\mathcal{Z} \times \mathcal{Z}} \tilde{y}(a f(b \cdot \tilde{x})) d\pi(z, \tilde{z}).
\end{aligned} \tag{1.14}
$$

我们以如下说明结束本节：具有任意初始化的系统 (1.10) 满足某些质量守恒性质。

注 1.3. 注意，(1.10) 中的动力学意味着 π (π_z) 的第一边缘分布保持不变。考虑一个测试函数 ϕ: Z $\to \mathbb{R}$ 并观察以下事实，即可验证这一点：

$$\begin{aligned}\frac{d}{dt}\int_{\mathcal{Z}\times\mathcal{Z}}\phi(z)d\pi_t(z,\tilde{z}) &= \eta_t\int_{\mathcal{Z}\times\mathcal{Z}}\nabla_{z,\tilde{z}}\phi(z)\cdot(0,\nabla_{\tilde{z}}\mathcal{U}_\pi)d\pi_t(z,\tilde{z})\\ &\quad+\kappa\int_{\mathcal{Z}\times\mathcal{Z}}\phi(z)\left(\mathcal{U}_\pi(z,\tilde{z})-\int\mathcal{U}_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|z)\right)d\pi_t(z,\tilde{z})\\ &=\kappa\int_{\mathcal{Z}}\phi(z)\int_{\mathcal{Z}}\left(\mathcal{U}_\pi(z,\tilde{z})-\int\mathcal{U}_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|x)\right)d\pi_t(\tilde{z}|z)d\pi_{t,z}(z)\\ &=0.\end{aligned}$$

同样可以证明，若 ν_0、π_0 是概率测度，则对于所有时刻， ν_t 和 π_t 的总质量均为一。

1.4. **主要结果**。现在可以陈述我们的主要结果了。我们的第一个结果描述了系统 (1.10) 在按照 (1.11) 初始化时的大粒子数极限 ($N\to\infty$)，该结果是在以下关于$\mathcal{U}$ 及其一阶变分的假设下推导得到的。

假设 1.4。我们假设存在常数 M、$L>0$，使得

- U 有界，且关于1-*Wasserstein*距离满足 *Lipschitz* 条件，即 i s

$$|\mathcal{U}(\pi,\nu)|\le M;\qquad \mathcal{U}(\pi^1,\nu^1)-\mathcal{U}(\pi^2,\nu^2)\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)).$$

- U 的一阶变分有界且满足 *Lipschitz* 条件，即

$$|\mathcal{U}_\pi(\pi,\nu;\ z,\tilde{z})|+|\mathcal{U}_\nu(\pi,\nu;\ \theta)|\le M$$
$$|\mathcal{U}_\pi(\pi^1,\nu^1;z^1,\tilde{z}^1)-\mathcal{U}_\pi(\pi^2,\nu^2;z^2,\tilde{z}^2)|\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|z^1-z^2|+|\tilde{z}^1-\tilde{z}^2|)$$
$$|\mathcal{U}_\nu(\pi^1,\nu^1;\theta^1)-\mathcal{U}_\nu(\pi^2,\nu^2;\theta^2)|\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|\theta^1-\theta^2|).$$

- U的一阶变分的梯度有界且满足 *Lipschitz* 条件，即

$$|\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi,\nu;\ z,\tilde{z})|+|\nabla_\theta\mathcal{U}_\nu(\pi,\nu;\ \theta)|\le M$$
$$|\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi^1,\nu^1;z^1,\tilde{z}^1)-\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi^2,\nu^2;z^2,\tilde{z}^2)|\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|z^1-z^2|+|\tilde{z}^1-\tilde{z}^2|)$$
$$|\nabla_\theta\mathcal{U}_\nu(\pi^1,\nu^1;\theta^1)-\nabla_\theta\mathcal{U}_\nu(\pi^2,\nu^2;\theta^2)|\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|\theta^1-\theta^2|).$$

在上文中，$\pi,\pi^i\in P(Z^2)$，$\nu,\nu^i\in P(\Theta)$，$(z^i,\tilde{z}^i)\in Z^2$, 和 $\theta^i\in\Theta$。集合 Θ 和Z^2分别是欧几里得空间 $\mathbb{R}^d$ 和 $\mathbb{R}^{2d'}$中的紧致区域。我们假设这些定义域具有利普希茨边界。

注释1.5。下文中，我们使用 p-*Wasserstein*距离 W_p （其中 $p\ge1$）来比较各种度量空间$(\mathbb{M},d(\cdot,\cdot))$上的概率分布。回顾一下，对于给定的定义域 $\mathbb{M}$上的两个概率测度 υ,υ'，它们的 p-*Wasserstein*距离 $W_p(\upsilon,\upsilon')$ 定义如下：

$$W_p^p(\upsilon,\upsilon'):=\inf_{\Upsilon\in\Gamma(\upsilon,\upsilon')}\int_{\mathbb{M}\times\mathbb{M}}(d(u,u'))^p d\Upsilon(u,u'),$$

其中，$\Gamma(\upsilon,\upsilon')$ 表示 υ 与 υ'之间的耦合集合。每种情形中所用的度量 $d(\cdot,\cdot)$ 将在上下文中说明。例如，在假设 *1.4* 中，所考虑的 1-*Wasserstein* 距离是相应欧几里得空间中的欧几里得度量所诱导的距离。

此外，我们用 $P(\mathbb{M})$ 表示 $\mathbb{M}$上的博雷尔概率测度空间。

由于已假设集合 Θ 和 $\mathcal{Z}^2$ 有界，为了简化证明的表述，并确保本文研究的所有动力学都保持在定义域 Θ 和 $\mathcal{Z}^2$,内，我们作出以下技术性假设：

假设 1.6。在 Z 边界上的所有点 $\tilde{z}$ 以及 Θ 边界上的所有点 θ 处，向量 $\nabla_{\tilde{z}}U_\pi(\pi,\nu;\ z,\tilde{z})$ 均指向 Z 的内部，而不受 π,ν,z 的影响；向量 $\nabla_\theta U_\nu(\pi,\nu;\ \theta)$ 则指向 Θ 的内部，而不受 π,ν 的影响。

通过将研究范围限制在紧致区域 Θ，$\mathcal{Z}^2$ 内，我们可以更容易地验证假设 1.4 中的有界性和 Lipschitz 条件，因为这些条件可简化为局部利普希茨性等较弱的性质。请注意，在许多应用中，所求解[1]的支撑集具有自然的界限。另一方面，假设 1.6 保证本文讨论的所有动力学都保持在定义域 Θ 和$\mathcal{Z}^2$ 内（例如，动力学 (1.8)）。为满足这一假设施加的约束，我们可以使用一个经过修改的泛函 $\mathcal{U}$；随着我们逐渐接近定义域边界，该泛函会对离开定义域的行为施加强烈惩罚。具体而言，对于给定的满足假设 1.4 的 $\mathcal{U}$，我们可以在需要时添加一个形式为 $\int \varphi_2(\theta)d\nu(\theta) - \int \varphi_1(\tilde{z})d\pi_{\tilde{z}}$ 的外生项，其中势函数 φ_1，φ_2 在定义域边界以外为零，并在趋近边界时增大。我们再次强调，作出这一假设是为了简化证明的表述。全文中我们均采用假设 1.6，即使没有明确提及也是如此。

例 1.7。 在作为研究动机的例 *1.2* 中可以看到，对于紧致区域Θ和 Z，当考虑二阶可微的损失函数以及一阶导数为 *Lipschitz* 的激活函数时，假设 *1.4* 中的所有条件均得到满足。例如，平方损失以及平方*ReLU*或 *Sigmoid* 激活函数便符合这一情形。

现在可以陈述我们的第一个主要结果了。

定理 1.8 (粒子系统收敛)。 设 $T > 0$，并假设假设 *1.4* 和 *1.6* 成立。设 $\bar{\pi}_0, \bar{\nu}_0$ 为满足 $\bar{\pi}_{0,\ z} = \mu$ 的概率测度。设 γ_t^N, σ_t^N

$$\gamma_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{(Z_t^i, \tilde{Z}_t^i), \omega_t^i}, \qquad \sigma_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{\vartheta_t^i, \alpha_t^i},$$

初始值 ω_0^i, α_0^i 均由常数 D 从上方有界（对 N 一致），且 Z_0^i 属于 μ 的支撑集，并且演化满足 (1.8)。最后，假设当 $N \to \infty$ 时，

$$\inf_{\upsilon \in \Gamma_{Opt}(\pi_{0,z}^N, \bar{\pi}_{0,z})} \int W_1(\gamma_0^N(\cdot|z_0'), \gamma_0(\cdot|z_0))d\upsilon(z_0', z_0) \to 0, \ and\ W_1(\sigma_0^N, \sigma_0) \to 0, \tag{1.15}$$

对于满足 $F\gamma_0 = \bar{\pi}_0$ 和 $F\sigma_0 = \bar{\nu}_0$的两个概率测度 γ_0 和 σ_0，其中 F 定义见(2.1) （以及第 *2.0.6* 节）。上述的 $\Gamma_{Opt}(\pi_{0,z}^N, \bar{\pi}_{0,z})$ 表示实现$\pi_{0,z}^N$ 和$\bar{\pi}_{0,z}$之间1-*Wasserstein*距离的耦合集合。

于是，当 $N \to \infty$时，

$$\sup_{t\in[0,T]} \{W_1(\pi_t^N, \pi_t) + W_1(\nu_t^N, \nu_t)\} \to 0,$$

其中 π_t、ν_t 以初始条件 $\bar{\pi}_0$ 和 $\bar{\nu}_0$求解(1.10) 。

注 1.9 （定理 1.8 中近似初始化的构造）。 固定 π_0 和 ν_0，并定义 $\gamma_0 = \pi_0 \otimes \delta_1$。也就是说，$\gamma_0$ 是 π_0 与位于1处的狄拉克 δ 测度的乘积。同样，令$\sigma_0 = \nu_0 \otimes \delta_1$。显然，$F\gamma_0 = \pi_0$，$F\ \sigma_0 = \nu_0$。

我们使用随机化来构造满足定理 *1.8* 中假设的近似初始化。令 $\xi_1, \ldots, \xi_n, \ldots$ 为来自 $\pi_{0,z}$ 的一列独立同分布样本，并且对每个 $i \in \mathbb{N}$，令 $Z_{i1}, \ldots, Z_{im}, \ldots$ 为来自 $\pi_0(\cdot|\xi_i)$ 的独立同分布样本。令 $\theta_1, \ldots, \theta_n, \ldots$ 为来自 ν_0 的独立同分布样本。

对于固定的 n, m 和 $i \le n$ 以及 $j \le m$，设 $\omega_{ij} = \alpha_{ij} = 1$, $Z_{ij} = \xi_i$ 和 $\vartheta_{ij} = \theta_i$。考虑测度

$$\pi_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{(Z_{ij}, \tilde{Z}_{ij})}, \quad \gamma_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{(Z_{ij}, \tilde{Z}_{ij}, \omega_{ij})}$$

and

$$\nu_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{\vartheta_{ij}}, \quad \sigma_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{(\vartheta_{ij}, \alpha_{ij})}.$$

[1]举一个简单的例子，图像通常由像素表示，每个像素都有上下限值

显然，$F\gamma^{n}{}_{,0}{}^{m} = \pi^{n}{}_{,0}{}^{m}$ 以及 $F\sigma^{n}{}_{,0}{}^{m} = \nu^{n}{}_{,0}{}^{m}$，且可假设 Z_{ij} 属于 $\pi_{0,z}$的支撑集。显然，测度 $\gamma^{n}{}_{,0}{}^{m}$ 的支撑集位于 $Z^2 \times [0,1]$，而 $\sigma^{n}{}_{,0}{}^{m}$ 的支撑集位于 $\Theta \times [0,1]$。

根据附录 *A.2* 中的引理 *A.3*，我们可以得出，存在一个序列 $\{(n_k, m_k)\}_{k\in\mathbb{N}}$，使得当 $k \to \infty$ 时，测度 $\sigma_0^{N_k} := \sigma^{n_k}{}_{,0}{}^{m_k}$和 $\gamma_0^{N_k} := \gamma^{n_k}{}_{,0}{}^{m_k}$以概率 *1* 满足 (1.15) 。

注 1.10。我们强调，为了满足(1.15) 中的第一个条件，我们需要考虑注 *1.9* 中所示的变量 Z_{ij}，Z_{ij} 的迭代抽样；而通常，从 π_0 独立同分布抽样并不能为粒子系统提供有效的初始化。这是因为(1.15) 中的第一个条件比仅要求 $W_1(\gamma_0^N, \gamma_0) \to 0$ 更强；参见附录 *A* 中的注 *A.4*。

最后，我们强调，(1.15) 中对初始化时条件分布的假设，是用于控制系统随时间演化时的条件分布。

在讨论了粒子系统向平均场极限方程 **(1.10)** 的收敛之后，我们将注意力转向平均场的长时间行为，前提是平均场经过适当初始化。我们将证明，在假设 **1.4** 以及对 $\mathcal{U}$ 的一些额外凸凹性假设下（凸凹性按线性插值的意义理解，见下文假设 **1.13**），我们可以从 **(1.10)** 得到问题 **(1.5)** 的 ε-纳什均衡。

定义 1.11 (ε-纳什均衡)。我们称 (π^*, ν^*) 是问题 (1.5) 的 ε-纳什均衡，当且仅当 $\pi_z^* = \mu$ 和

$$\sup_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z}) \ s.t. \ \pi_z=\mu} \{\mathcal{U}(\pi, \nu^*)\} - \inf_{\nu\in\mathcal{P}(\Theta)} \{\mathcal{U}(\pi^*, \nu)\} \le \varepsilon. \tag{1.16}$$

注 1.12. 注意，条件 (1.16) 等价于：

$$\mathcal{U}(\pi, \nu^*) - \mathcal{U}(\pi^*, \nu) \le \varepsilon, \quad \forall \nu \in \mathcal{P}(\Theta), \quad \forall \pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z}) \ with \ \pi_z = \mu.$$

ε-纳什均衡可以理解为博弈(1.1) 的近似解，含义如下：任一博弈参与者（学习者或对手）偏离策略对(π^*, ν^*)，都不会使另一方的收益降低超过 ε。特别是，如果学习者采用策略 ν^*，那么 ν^* 就是鲁棒风险泛函 $F: \nu \in P(\Theta) \longmapsto \sup_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z}) \text{ 满足 } \pi_z=\mu} U(\pi, \nu)$的一个 ε-最小化解。

假设1.13。我们假设，按照线性插值的意义，U 关于 ν 是凸的，关于 π 是凹的。也就是说，

$$\mathcal{U}(\tau\pi + (1-\tau)\hat{\pi}, \nu) \ge \tau\mathcal{U}(\pi, \nu) + (1-\tau)\mathcal{U}(\hat{\pi}, \nu)$$

and

$$\mathcal{U}(\pi, \tau\nu + (1-\tau)\hat{\nu}) \le \tau\mathcal{U}(\pi, \nu) + (1-\tau)\mathcal{U}(\pi, \hat{\nu}),$$

对所有 $\tau \in [0,1]$ 以及所有概率测度 π，$\hat{}\pi \in P(Z \times Z)$，以及 ν，$\hat{}\nu \in P(\Theta)$。

假设 U 具有(1.4)的形式，由于 C 关于 π是线性的，上述条件等价于对 $R(\pi, \nu)$作出类似的凸凹性假设。

注释 1.14. 根据线性插值（即假设*1.13*中的定义），U 是凸凹函数，但这并不意味着相对于诱导动力学 (1.10) 的几何结构而言，U 是测地凸凹的（关于方程 (1.10)的几何解释，见第*2*节），因此无法直接得出其收敛到全局纳什均衡或近似纳什均衡的结论。鉴于此，尽管满足假设 *1.13*，我们仍将本节所讨论的情形称为非凸非凹情形。我们将该情形与第*1.4.1*节中的情形进行对比，并将后者称为非凸-凹情形。

例1.15. 在例*1.2*所述的动机背景下，我们可以看到，只要损失函数 ℓ 关于其第一坐标是凸函数，假设*1.13*就成立。平方损失和逻辑损失显然都满足这一条件。

现在可以陈述我们的第二个主要结果。

定理1.16 (长时间行为均值场偏微分方程)。 设 $\epsilon > 0$。假设假设*1.4*、*1.6*和*1.13*成立。设 ν_0 和 π_0 是概率测度（其中$\pi_{0,\ z} = \mu$），且存在 $k > 0$ 使得$\frac{d\nu_0}{d\theta}(\theta) > k$，并且对 μ支撑集中的所有 z，都有 $\frac{d\pi_0}{(\tilde{z}|z)\ d\tilde{z}}(\tilde{z}) > k$ 。最后，假设 η_t 的选取满足

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \int_0^s \eta_\tau d\tau ds = \bar{\eta}$$

对于满足以下条件的 $\bar{\eta}$

$$4(L+M)^2\bar{\eta} < \epsilon.$$

那么存在 T^*，使得对所有 $t > T^*$

$$\sup_{\pi^*\in\mathcal{P}(\mathcal{Z}^2)\ s.t\ \pi^*_z=\mu} \mathcal{U}(\pi^*, \bar{\nu}(t)) - \inf_{\nu^*\in\mathcal{P}(\Theta)} \mathcal{U}(\bar{\pi}(t), \nu^*) \le \epsilon,$$

其中 $\underline{\pi_t} := \frac{1}{t}\int_0^t \pi_s ds$ 且 $\underline{\nu_t} := \frac{1}{t}\int_0^t \nu_s ds$，并且$(\pi_t, \nu_t)$ 在初始值为 π_0, ν_0 时求解(1.10)，如上所述。

注1.17。 关于初始条件 π_0 和 ν_0 的假设实际上意味着，在时刻零，算法*1*中的粒子在各定义域内分布得相当均匀。这无疑是一个很强的假设，但与数学上研究神经网络训练过程的文献中其他理论假设并无二致；参见 *[10, 13, 51–53]*。

下一节将讨论，在第*1.2*节所述的设定中限制对手的对抗预算时，如何去除关于 π_0 的强假设。

1.4.1. 非凸且强凹情形。 与定理1.16不同，本节给出的结果不对初始化 π_0 作任何假设，但代价是需要对收益函数$\mathcal{U}$作出额外假设。如下文所述，在第1.2节所描述的设定中，这些额外假设并非不自然，因为它们与式(1.4)中对对抗代价泛函$\mathcal{C}$ 所赋予的强度有关。

假设1.18。 我们对函数 $U(\cdot\ \nu)$作如下统一 *PL*（*Polyak-Lojasiewicz*）条件假设：存在 $\lambda > 0$，使得对于所有 $\nu \in P(\Theta)$ 以及所有 $\pi \in P(Z^2)$，若$\pi_z = \mu$，则有

$$\int |\nabla_{\tilde{z}} \mathcal{U}_\pi(\pi, \nu; z, \tilde{z})|^2 d\pi(z, \tilde{z}) \ge \lambda(m^*_\nu - \mathcal{U}(\pi, \nu)),$$

其中 $m^*_\nu := \sup_{\tilde{\pi}\ 满足\ \tilde{\pi}_z=\mu} U(\tilde{\pi}, \nu)$。

注1.19。 为简便起见，我们将假设*1.18*成立的情形称为强凹情形，因为通常可以从强（测地）凹性推出 *PL* 条件；参见附录 *A.1* 中的命题*A.1*。

例1.20。 假设收益函数 U 具有(1.4) 的形式，其中 R 和 C 分别如(1.6)和(1.7)所示。正如我们在附录 *A.1* 的命题*A.1*中所证明的，如果集合 Z 是凸的（这在应用中是一个合理的假设），激活函数和损失函数二次连续可微，并且更重要的是，参数 c_a 足够大，那么假设(1.18)便成立。

为利用关于$\mathcal{U}(\cdot\ \nu)$的额外假设，考虑对 (1.10) 作一个细微变动会很有帮助：在下降方程中放慢时间，并从 π_t 的方程中去掉缩放因子 η。具体而言，给定 $K \ge 1$，我们考虑系统 (1.17)

$$\begin{cases} \partial_t \nu_t &= \frac{\eta_t}{K}\mathrm{div}_\theta(\nu_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \theta)) - \frac{\kappa}{K}\nu_t\left(\mathcal{U}_\nu(\pi_t, \nu_t; \theta)) - \int \mathcal{U}_\nu(\pi_t, \nu_t; \theta')d\nu_t(\theta')\right) \\ \partial_t \pi_t &= -\mathrm{div}_{z,\tilde{z}}(\pi_t(0, \nabla_{\tilde{z}}\mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}))) + \kappa\pi_t\left(\mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}) - \int \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}')d\pi_t(\tilde{z}'|z)\right), \end{cases}$$

该系统以任意的 $\pi_0 \in \mathcal{P}(\mathcal{Z}^2)$ 为初始值，且满足 $\pi_{0,z} = \mu$，以及某个 ν_0。在假设 1.4 和 1.6 下，该方程的适定性可参照方程 (1.10) 的处理方式加以证明；细节从略。为在算法（1）中体现 (1.17) 引入的变动，只需从变量 $\tilde{z}_{ij}$ 的更新中去掉 η，并允许在进入关于 k 的循环之前，for 循环 i, j 重复执行若干次（次数可调）。

我们证明以下结果。

定理 1.21。 假设假设 *1.4*、*1.6*、*1.13* 和 *1.18* 成立。进一步假设存在 $k > 0$，使得 $\frac{d\nu_0}{d\theta} > k$，并令 π_0 为任意满足 $\pi_{0,\overline{z}} = \mu$ 的概率测度。最后，假设

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \int_0^s \eta_\tau d\tau ds = \bar{\eta} < \infty.$$

固定 $\epsilon > 0$。那么存在 $K_0, r_0, r_1, t_0 > 0$，使得若 $K \geq K_0$ 且 $\eta/K \leq r_1$，则对所有 $t \geq$ 最大值$\{t_0, K/r_0\}$，都有

$$\sup_{\tilde{\pi}\in\mathcal{P}(\mathcal{Z}^2)\ s.t.\ \tilde{\pi}_z=\mu} \mathcal{U}(\tilde{\pi}, \bar{\nu}_t) - \inf_{\tilde{\nu}\in\mathcal{P}(\Theta)} \mathcal{U}(\bar{\pi}_t, \tilde{\nu}) \leq \epsilon.$$

在上述内容中，$\bar{\pi}_t := \frac{1}{t}\int_0^t \pi_s ds$ 且 $\bar{\nu}_t := \frac{1}{t}\int_0^t \nu_s ds$，并且 (π_t, ν_t) 求解以 ν_0, π_0 为初始值的 (1.17)，如上所述。

1.5. **文献综述**。 本节将简要回顾监督学习场景下对抗鲁棒性这一主题的相关文献，重点介绍近年来的一些进展。由于该领域的文献迅速增长，且涉及多个学科，我们的综述难免无法面面俱到。

对抗训练这一概念，即如何创建能够抵御数据对抗扰动的学习模型，几年前开始流行起来。当时，神经网络不久前才成为解决图像处理和自然语言处理任务的最先进技术。事实上，不同研究者注意到，神经网络模型（以及其他模型）虽然能够基于干净数据做出准确预测，效果显著，却对对抗攻击相当敏感——参见 [23, 27, 46]。为了增强对对抗攻击的防护，机器学习和计算机科学领域的研究者提出，应以一种新的训练方式取代模型的标准训练过程：标准训练过程以优化标准损失函数目标为基础，而新的训练方式则涉及优化一个纳入明确定义的对手行动的目标函数。例如，论文 [33] 提出了如下优化问题

$$\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x}\in\mathcal{B}_\varepsilon(x)} \ell(\theta, (\tilde{x}, y)) \right]. \tag{1.18}$$

在上述问题中，最小化针对学习模型的参数 θ （例如神经网络的参数）进行，内层最大化则针对给定输入数据点 x周围所有可能的“对抗攻击” $\tilde{x}$ 进行。 $\ell(\theta, (\tilde{x}, y))$表示参数为 θ的模型的数据失配。通过求解最小化问题（1.18），人们期望得到能够抵御由球$\mathcal{B}_\varepsilon(\cdot)$的选取所建模的特定攻击者的模型；也就是说，通过指定一个“距离”函数并设置对抗预算 ε来实现这一点。例如，在深度神经网络模型的研究中， [55] 等工作探讨了不同策略，以高效实现梯度上升-下降动力学（针对数据扰动进行上升，针对模型参数进行下降），从而求解类似（1.18）的问题。

类似 (1.18) 的问题推动了近期优化领域的研究，促使研究者重新关注有限维情形下一般极小极大博弈的分析（这些论文关注的设定更接近 (1.2)，而非 (1.1)）。这一方向的一些研究包括 [15, 30, 31, 36]。正如 [15, 30], 所讨论的，在对抗机器学习中，非凸-凹问题其实并不少见。

旨在增强学习模型鲁棒性的更一般目标，其形式与 (1.1) 或 (1.2) 更为接近：内层最大化是在一族分布上进行的，而不是像 (1.18) 那样逐点最大化。所有这些表述都可以纳入分布鲁棒优化（DRO）这一总称之下。DRO 表述的优点在于，它将监督学习中的对抗鲁棒性清晰地刻画为一个极小极大博弈。已有多项研究探讨了 DRO 设定下的对抗训练，所考虑的学习模型包括线性回归（或其他经典参数化设定）以及神经网络等模型；参见 [4, 5, 9, 26, 34, 45, 48]。

分析问题 (1.18) 时，以及在对抗鲁棒性的分布鲁棒优化（DRO）表述中，一种常用策略是用正则化风险代理目标替代与对手行动相关的内层最大化。例如，在 (1.18) 的情形下，可以通过围绕 $\varepsilon = 0$ 展开内层最大化目标，形式化地得到这一代理目标。采用这一策略的论文包括 [16, 32, 35, 40, 41, 54]。在他们的论文 [19], 中，作者探讨了学习者使用深度神经网络构建分类器/回归函数时的一般分布鲁棒优化问题中的这一策略；他们利用控制理论中的思想来实现对所得代理目标的优化。其他论文，如 [39] ，则探讨了通过以信息丰富的上界代理目标替代目标函数 (1.18) 来训练模型的学习情形。随后，使用上升-下降算法求解由此产生的问题。

近期有一些研究探讨了无先验假设学习者（即不对学习者作任何建模假设）情形下的对抗鲁棒性；这可以理解为学习模型族具有极强表达能力的问题的极限情形。这类情形之所以有用，是因为它们为更一般的对抗鲁棒性问题提供了下界。相关研究包括： [2, 3, 17, 37, 38]。具体而言， [3, 37, 38]推导并探讨了最优传输问题与 (1.18) 之间的等价性，其中学习模型空间是所有可测二元分类器的集合。其他研究，如 [7, 8, 21] ，则将二分类情形下的对抗鲁棒性与涉及周长和曲率的几何变分问题联系起来，并探讨鲁棒分类器的存在性及其正则性；研究 [2,17]则探索了满足特定“可认证性”属性的鲁棒二元分类器是否存在。研究 [20] 探讨了多类分类中的对抗问题，以及它们与测度空间中的多边最优传输问题和（广义）重心之间的等价性。

我们希望在本节结尾强调本研究的贡献，尤其是相对于 [13, 51], 而言；这些工作与我们的研究关系更为密切。

在文献 [13] 和最近的工作 [51] 中，作者研究了收益函数关于测度呈线性的极小极大问题。我们的设定更为宽泛，不仅涵盖非线性目标函数，还研究了其中一个组成部分被固定为输入函数时所产生的影响，这使我们能够研究测度空间中更广泛的对手情形（对抗训练的分布鲁棒优化版本）。值得指出的是，在文献 [51], 的较简单设定下，作者证明了一种具有与我们类似几何动机、但额外包含熵项的算法能够指数收敛（趋近于实际的纳什均衡）。在实际实现中，两种算法看起来非常相似。文献 [51] 中的收敛结论是在类似的正则性假设下得出的，但还额外假设（唯一）解的支撑集是一个离散测度。由于我们没有先验地假设唯一解存在，因此我们的结果在收敛速率方面较弱，而且只能恢复近似纳什均衡（精度可按任意指定值设定）。

我们强调，不将分析局限于在两个变量上均为线性的收益函数 u ，使我们能够涵盖比现有文献中更广泛的、与对抗机器学习研究相关的情形。一般性上的提升自然要以应对更多技术挑战为代价。举例来说，当收益函数为非线性时，其一阶变分依赖于测度，这已经表明，在证明粒子动力学收敛到平均场极限时，需要进行更为细致的分析。条件分布的存在进一步加剧了我们分析中的困难

。为了处理这些额外项，我们必须采用新的思路和构造。最终，我们提出的一般平均场分析也可以与下半连续性论证相结合，从而为论文第二部分中的某些步骤提供依据，即对平均场系统长时间行为的分析，并以此为文献中可能缺乏充分论证的方法提供替代方案；我们将在下文第 4 节讨论这一点。此外，我们认为，为了使模型达到预期的一般性而得到的一些辅助结果，本身也可能具有一定价值。

我们还重点研究了 1.4.1 节所述的非凸-凹情形。事实上，我们利用对抗问题的特定结构，并运用考虑预算较低的对手时所具有的强凹性，从而获得更强的收敛结果，使其收敛至该对抗问题的近似纳什均衡。其他探讨这一情形的论文包括 [45], ，但其中给出的结果仅能保证学习者收敛至驻点（不过值得指出的是，这些工作并未采用均值场机制）。

总而言之，我们的工作与 [13, 45, 51] 等其他论文互为补充。我们的结果可视为与 [10, 52], 等研究中的结果相类似；这些研究探讨了均值场机制下浅层神经网络（非鲁棒）训练的全局收敛性。

1.6. **概要**。本文其余部分安排如下。第 2 节讨论算法1的梯度上升-下降解释，并回顾理解这一解释所需的一些最优传输理论概念。第 3 节用于证明定理 1.8 及若干推论；这些推论将在第 4 节证明第二个主要结果定理1.16时用到。第 4.1 节给出定理1.16的证明，第 4.2 节给出定理 1.21 的证明。第5节讨论将我们的算法应用于实际机器学习任务时得到的一些数值结果。最后，第 6 节总结全文，给出一些结论并探讨未来的研究方向。

2. 算法1的梯度上升-下降诠释

除了引入一些数学定义和后文将用到的符号外，本节旨在为方程组 (1.10) 及其在算法1中的离散化提供直观的几何动机。因此，本节的讨论大体上是形式性的。有几篇文献为本节的讨论提供了启发，例如， [11, 18, 25, 29, 43]。

2.0.1. 一个提升空间。 下文中，我们用 $\mathcal{M}_+(\Theta)$ 表示 Θ 上有限正测度的空间。我们考虑一个投影映射 $\mathcal{F}$: $\mathcal{P}(\Theta \times [0,\infty)) \to \mathcal{M}_+(\Theta)$，它将每个 $\sigma \in \mathcal{P}(\Theta \times [0,\infty))$ 映射为一个由恒等式确定的测度 $\mathcal{F}\sigma \in \mathcal{M}_+(\Theta)$

$$\int \varphi(\theta) d(\mathcal{F}\sigma)(\theta) = \int \alpha\varphi(\theta) d\sigma(\theta,\alpha) \tag{2.1}$$

对所有足够正则的测试函数 $\varphi : \Theta \to \mathbb{R}$。如下文所示，映射 $\mathcal{F}$ 使我们能够将定义在 $\mathcal{M}_+(\Theta)$ 上的泛函提升为定义在 $\mathcal{P}(\Theta \times [0,\infty))$ 上的泛函。能量提升后，可以在提升空间中考虑梯度下降动力学。进而，这些提升动力学可以投影回原始测度空间，从而在其中生成演化。

注释 2.1. 注意，函数 F 是满射。事实上，令 $\nu \in M_+(\Theta)$ 并令$M = \nu(\Theta)$，我们首先假设它非零。考虑概率测度 $\sigma = \frac{\mu}{M} \otimes \delta_M$。不难证明 $F\sigma = \nu$。若 $M = 0$，这意味着 ν 是恒等于零的测度，则我们可以取任意满足 $\sigma(\Theta \times \{0\}) = 1$ 的、定义在 $\Theta \times [0,\infty)$ 上的概率测度作为 σ，从而得出 $F\sigma = \nu$。

Fin此外，尽管 F 是满射，但值得强调的是，它远非一一对应。 $e.$

要将泛函 $J\colon \mathcal{M}_+(\Theta) \to (-\infty,\infty]$ 提升为定义在 $\mathcal{P}(\Theta\times[0,\infty))$ 上的泛函，只需考虑 J 与投影映射 $\mathcal{F}$ 的复合，如下所示：

$$\mathcal{J}(\sigma) := J(\mathcal{F}\sigma), \quad \sigma\in\mathcal{P}(\Theta\times[0,\infty)). \tag{2.2}$$

接下来，我们讨论空间 $\mathcal{P}(\Theta\times[0,\infty))$ 上的一种黎曼结构。

2.0.2. 提升空间上的度量。我们首先按照以下方式定义空间 $\Theta\times(0,\infty)$ 上的度量张量：

$$g_{(\theta,\alpha)}((v,s),(\tilde{v},\tilde{s})) := \frac{\alpha}{\eta}\langle v,\tilde{v}\rangle + \frac{1}{\kappa\alpha}s\tilde{s},$$

其中，$\langle\cdot,\cdot\rangle$ 表示欧几里得空间中的标准内积，κ 和 η 是两个正参数。下文中，我们使用记号 $|(v,s)|^2_{(\theta,\alpha)} := g_{(\theta,\alpha)}((v,s),(v,s))$。

不难验证，标量函数 $\phi(\theta,\alpha)$ 关于内积 g的梯度（记为 $\overline{\nabla}\phi$）可写为

$$\overline{\nabla}\phi = (\frac{\eta}{\alpha}\nabla_\theta\phi, \kappa\alpha\partial_\alpha\phi), \tag{2.3}$$

其中，$\nabla_\theta\phi(\theta,\ \alpha)$是 ϕ 关于变量 θ 的通常梯度，而 $\partial_\alpha\phi(\theta,\ \alpha)$是 ϕ 关于 α的偏导数。注意，$\overline{\nabla}\phi$ 是 $\mathbb{R}^p\times\mathbb{R}$中的一个向量。

相对于 $\Theta\times(0,\ \infty)$上的基度量 g，我们在概率测度空间$\mathcal{P}(\Theta\times[0,\infty))$上定义一种动态形式的Wasserstein度量。更准确地说，对于 $\sigma,\ \sigma'\in\mathcal{P}(\Theta\times[0,\infty))$，我们考虑

$$W^2_{2,g}(\sigma,\hat{\sigma}) = \inf_{\{(\beta_t,\sigma_t)\}_{t\in[0,1]}\in \mathrm{CE}(\sigma,\tilde{\sigma})} \int_0^1\int |\overline{\nabla}\beta_t(\theta,\alpha)|^2_{\theta,\alpha} d\sigma_t(\theta,\alpha)dt, \tag{2.4}$$

其中，集合 $\mathrm{CE}(\sigma,\ \tilde{\sigma})$由所有满足（内在）连续性方程的解 $t\in[0,1]\mapsto(\beta_t,\ \sigma_t)$组成

$$\begin{cases} \partial_t\sigma_t + \overline{\mathrm{div}}(\sigma_t\overline{\nabla}\beta_t) = 0, \\ \sigma(0)=\sigma, \quad \sigma(1)=\sigma'; \end{cases} \tag{2.5}$$

特别地，div表示赋予度量g的空间 $\Theta\times(0,\ \infty)$中的散度。一般来说，方程(2.5)必须按弱意义来理解，即必须满足

$$\frac{d}{dt}\int\phi(\theta,\alpha)d\sigma_t(\theta,\alpha) = \int g_{(\theta,\alpha)}(\overline{\nabla}\beta_t(\theta,\alpha),\overline{\nabla}\phi(\theta,\alpha))d\sigma(\theta,\alpha)$$

对于所有 $t\in(0,1)$ 以及所有 ϕ 足够正则的测试函数。

与度量 (2.4) 本身相比，我们更关注公式 (2.4) 中隐含的形式黎曼结构；我们可以在$\mathcal{P}(\Theta\times[0,\infty))$上赋予这种结构，并以启发式的方式用它来说明空间$\mathcal{P}(\Theta\times[0,\infty))$中的梯度下降或投影梯度下降动力学。按照从黎曼几何视角理解最优传输的惯常做法，可以将集合$\mathcal{T}_\sigma := \{\nabla\beta$ 满足 $\beta\colon \Theta\times(0,\infty)\ \mapsto\mathbb{R}\}$ 视为形式流形$\mathcal{P}(\Theta\times[0,\overline{\infty}))$在点 σ处的形式切平面，并可在这个形式切平面上根据以下方式定义内积$\langle\cdot,\cdot\rangle_\sigma$

$$\langle\overline{\nabla}\beta,\overline{\nabla}\beta'\rangle_\sigma := \int g_{(\theta,\alpha)}(\overline{\nabla}\beta(\theta,\alpha),\overline{\nabla}\beta'(\theta,\alpha))d\sigma_t(\theta,\alpha).$$

在结束本节之前，我们陈述一个后文将用到的结果，它使我们能够用基本的欧几里得散度和梯度算子来表示连续性方程 (2.5)。

命题 2.2。 内禀连续性方程 (2.5) 可用欧几里得散度 散度$_{\theta,\ \alpha}$ 在 $\mathbb{R}^p\times\mathbb{R}$ 中表示为

$$\partial_t\sigma_t + \mathrm{div}_{\theta,\alpha}(\sigma_t v_{\sigma_t}) = 0,$$

其中 v_σ 是向量场

$$v_\sigma(\theta,\alpha) := (\frac{\eta}{\alpha}\nabla_\theta\beta(\theta,\alpha), \kappa\alpha\partial_\alpha\beta(\theta,\alpha)).$$

证明。 这是以下简单观察的推论。对于所有足够正则的测试函数 ϕ，我们有

$$\begin{aligned}\frac{d}{dt}\int \phi d\sigma_t &= \int g_{(\theta,\alpha)}(\overline{\nabla}\beta, \overline{\nabla}\phi) d\sigma_t \\ &= \int \left(\frac{\eta}{\alpha}\nabla_\theta\phi\cdot\nabla_\theta\beta + \kappa\alpha\partial_\alpha\phi\partial_\alpha\beta\right) d\sigma_t \\ &= \int \langle \nabla_{\theta,\alpha}\phi, v_\sigma\rangle d\sigma_t,\end{aligned}$$

其中，上文用 $\langle\cdot,\cdot\rangle$ 表示 $\mathbb{R}^p\times\mathbb{R}$ 和∇_θ中的标准欧几里得内积，并用$_\alpha\phi$ 表示 $\mathbb{R}^p\times\mathbb{R}$中的标准梯度。

□

2.0.3. 垂直与水平向量场，定义于 $\mathcal{P}(\Theta\times[0,\infty))$。 现在我们引入并讨论形式切平面 T_σ 的一些相关子空间。稍后我们将用到这些子空间。

点 σ 处的水平空间 T^h_σ 定义为

$$T^h_\sigma := \{\overline{\nabla}\beta \text{ s.t. } \beta(\theta,\alpha) = \alpha\varphi(\theta) \text{ for some } \varphi\},$$

而垂直空间 T^v_σ 定义为

$$T^v_\sigma := \{\overline{\nabla}\beta \text{ s.t. } \langle\overline{\nabla}\beta, \overline{\nabla}\beta'\rangle_\sigma = 0, \quad \text{for all} \quad \overline{\nabla}\beta' \in T^h_\sigma\}.$$

垂直空间 T^v_σ 表示无穷小地保持 $\mathcal{F}\sigma$ 不变的方向，而水平空间是 T^v_σ 的正交补。

我们将$\mathcal{N} = \{\sigma$ 使得$\mathcal{F}\sigma\in\mathcal{P}(\Theta)\}$记作。对于 $\sigma\in\mathcal{N}$，我们考虑$\mathcal{T}_\sigma\mathcal{N}$的子空间$\mathcal{T}_\sigma$，定义如下:

$$\mathcal{T}_\sigma\mathcal{N} := \{\overline{\nabla}\phi \text{ s.t.} \int \alpha\partial_\alpha\phi(\theta,\alpha)d\sigma = 0\}.$$

子空间$\mathcal{T}_\sigma\mathcal{N}$ 可解释为所有经过 σ且始终处于$\mathcal{N}$中的曲线的切向量空间。

注 2.3. 空间 $M_+(\Theta)$ 可以赋予 *Wasserstein-Fisher-Rao*度量，使映射 F 成为黎曼浸没。事实上，注意，对于形如 $\alpha\varphi(\theta)$ 和 $\alpha\varphi'(\theta)$ 的两个势函数（即在点 σ处诱导水平向量场的两个势函数），我们有如下恒等式:

$$\langle\overline{\nabla}(\alpha\varphi), \overline{\nabla}(\alpha\varphi')\rangle_\sigma = \int_{\Theta\times[0,\infty)} \alpha(\eta\nabla_\theta\varphi\cdot\nabla_\theta\varphi' + \kappa\varphi\varphi')d\sigma(\theta,\alpha) = \int_\Theta (\eta\nabla_\theta\varphi\cdot\nabla_\theta\varphi' + \kappa\varphi\varphi')d\mathcal{F}\sigma(\theta).$$

换言之，上述内积实际上并不依赖于具体的 σ，而只依赖于 $F\,\sigma$。

关于 *Wasserstein-Fisher-Rao* 几何的详细信息，请参阅参考文献 *[11, 18, 25, 29, 43]* 。

2.0.4. 提升能量的梯度流。 给定任意能量 $\mathcal{J}$: $\mathcal{P}(\Theta\times[0,\infty)) \to (-\infty, \infty]$（不一定具有（2.2）的形式），相对于第 2.0.2 节中引入的黎曼几何，$\mathcal{J}$ 的梯度（下降）流具有如下形式:

$$\partial_t\sigma_t - \overline{\mathrm{div}}(\sigma_t\overline{\nabla}\mathcal{J}_{\sigma_t}) = 0, \tag{2.6}$$

其中，$\mathcal{J}_\sigma$ 是$\mathcal{J}$ 在点 σ处的一阶变分，其定义与第1.3节开头所述相同。关于将 (2.6) 解释为梯度流的更多细节，请参见 [50]中的第8.2章。

若$\mathcal{J}$具有如 (2.2) 中所示的提升能量结构，则其一阶变分可按如下方式计算。设 σ, σ^*，且 $\nu = \mathcal{F}\sigma$ 和 $\nu^* = \mathcal{F}\sigma^*$。利用映射 $\mathcal{F}$的线性（这可由其定义直接看出），得到:

$$\begin{aligned}\frac{d}{d\varepsilon}|_{\varepsilon=0}\mathcal{J}(\sigma+\varepsilon(\sigma^*-\sigma)) &= \frac{d}{d\varepsilon}|_{\varepsilon=0}J(\mathcal{F}(\sigma+\varepsilon(\sigma^*-\sigma)))\\ &= \frac{d}{d\varepsilon}|_{\varepsilon=0}J(\mathcal{F}\sigma+\varepsilon(\mathcal{F}\sigma^*-\mathcal{F}\sigma))\\ &= \int_\Theta J_\nu(\theta)d(\nu^*-\nu)\\ &= \int_{\Theta\times[0,\infty)} \alpha J_\nu(\theta)d(\sigma^*-\sigma),\end{aligned}$$

其中， J_ν 是 ν处 J 的一阶变分。换言之，在σ处，$\mathcal{J}$ 的一阶变分具有 αJ_ν的形式，其中 J_ν是 J的一阶变分；$\mathcal{J}_\nu$具有这种特定形式并不令人意外，因为函数$\mathcal{J}$ 沿垂直向量场保持常数，因此其梯度应为水平向量场。将此表达式代回式（2.6），可得提升能量$\mathcal{J}$ 的梯度流具有如下形式:

$$\partial\sigma_t - \overline{\mathrm{div}}(\sigma_t\overline{\nabla}(\alpha\mathcal{J}_{\nu_t})) = 0; \quad \mathcal{F}\sigma_t = \nu_t,$$

利用命题（2.2）后，也可将其写为

$$\begin{cases}\partial_t\sigma_t - \mathrm{div}_{\theta,\alpha}(\sigma_t v_\sigma) = 0;\\ v_\sigma(\theta,\alpha) = (\eta\nabla_\theta J_{\nu_t}(\theta), \kappa\alpha J_{\nu_t}(\theta))\,; \quad \nu_t = \mathcal{F}(\sigma_t).\end{cases} \tag{2.7}$$

2.0.5. 投影梯度。一般而言，对于 $t > 0$，式（2.7）中的 σ_t 可能不属于$\mathcal{N}$，即使初始值位于$\mathcal{N}$中的 $\sigma_0 \in$处也是如此。若要确保对所有 t都有 $\nu_t = \mathcal{F}\sigma_t \in \mathcal{P}(\Theta)$，则必须将驱动动力学（2.7）的能量$\mathcal{J}$的（Wasserstein）梯度投影到子空间$\mathcal{T}_\sigma\mathcal{N}$上。

给定 σ 和 $\nu = \mathcal{F}\sigma$，我们将势函数 αJ_ν 写为

$$\alpha J_\nu(\theta) = \alpha(J_\nu(\theta) - \int J_\nu(\theta')d\nu(\theta')) + \alpha\int J_\nu(\theta')d\nu(\theta').$$

直接计算可得

$$\langle\overline{\nabla}(\alpha\int J_\nu(\theta')d\nu(\theta'))), \overline{\nabla}\phi(\theta,\alpha)\rangle_\sigma = 0,$$

对所有$\nabla\phi \in \mathrm{T}_\sigma\mathrm{N}$ 均成立；这意味着 $\nabla(\alpha\int J_\nu(\theta')d\nu(\theta'))) \in \mathrm{T}_\sigma\mathrm{N}^{\perp}$。另一次直接计算可得 $\nabla(\alpha(J_\nu(\theta) - \int J_\nu(\theta')d\nu(\theta'))) \in \mathrm{T}_\sigma\mathrm{N}$。由此可见， $\nabla(\alpha(J_\nu(\theta) - \int J_\nu(\theta')d\nu(\theta')))$ 是 $\nabla(\alpha J_\nu)$ 在 $\mathrm{T}_\sigma\mathrm{N}$ 上的投影。

因此，利用命题 (2.2)，我们可以得出

$$\begin{cases}\partial_t\sigma_t - \mathrm{div}_{\theta,\alpha}(\sigma_t v_\sigma) = 0;\\ v_\sigma(\theta,\alpha) = (\eta\nabla_\theta J_{\nu_t}(\theta), \kappa\alpha(J_{\nu_t}(\theta) - \int J_{\nu_t}(\theta')\nu_t(\theta')))\,; \quad \nu_t = \mathcal{F}(\sigma_t),\end{cases} \tag{2.8}$$

表示提升能量 $\mathcal{J}$ 的梯度下降动力学在 $\mathcal{N}$ 上的投影。

2.0.6. $\mathrm{M}_+(\mathrm{Z}\times\mathrm{Z})$上的类似几何结构。空间 $\mathrm{M}_+(\mathrm{Z}\times\mathrm{Z})$可以赋予与我们在前几节讨论过的类似几何结构。下文中，我们用 γ表示提升空间 $\mathrm{P}(\mathrm{Z}\times\mathrm{Z}\times[0,\infty))$中的元素，并用形如 $(z,\tilde{z},\omega)$ 的三元组表示 $\mathrm{Z}\times\mathrm{Z}\times[0,\infty)$中的元素。空间 $\mathrm{P}(\mathrm{Z}\times\mathrm{Z}\times[0,\infty))$赋有一个 Wasserstein度量，形式与 (2.4) 相同，只需将所有出现的 θ替换为 $(z,\tilde{z})$，并将所有出现的 α 替换为 ω。我们用 F 表示投影映射 $\mathrm{F}: \mathrm{P}(\mathrm{Z}\times\mathrm{Z}\times[0,\infty)) \to \mathrm{M}_+(\mathrm{Z}\times\mathrm{Z})$（为简便起见，沿用第 2.0.1 节中的记号）。任意泛函 $J: \mathrm{M}_+(\mathrm{Z}\times\mathrm{Z}) \to (-\infty,\infty]$ 都可以通过与 F 复合提升到 $\mathrm{P}(\mathrm{Z}\times\mathrm{Z}\times[0,\infty))$（与前几节一样，我们用 J 表示这一复合）。J 的一阶变分具有如下形式： ωJ_π，其中 J_π 是 J 在 $\pi = \mathrm{F}(\gamma)$ 处的一阶变分。

由于问题 (1.5) 要求我们将范围限制在第一边缘分布等于 μ 的测度 π 上，因此我们考虑这样的演化方程：它们可以视为上述 Wasserstein度量下提升能量 $\mathcal{J}$ 的梯度上升 流的适当（投影）梯度上升 版本。这类演化方程具有如下形式：

$$\begin{cases}\partial_t\gamma_t + \mathrm{div}_{(z,\tilde{z}),\omega}(\gamma_t v_\gamma) = 0, \\ v_\gamma(z,\tilde{z},\omega) = \left(0, \eta\nabla_{\tilde{z}} J_\pi(z,\tilde{z}), \kappa\omega\left(J_\pi(z,\tilde{z}) - \int J_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|z)\right)\right); \quad \pi_t = \mathcal{F}\gamma_t.\end{cases} \tag{2.9}$$

为说明 $v_\gamma(z,\ \tilde{z},\ \omega)$ 的第一个分量为何为零，假设 $t \mapsto \pi_t$ 具有如下形式

$$\pi_t = \sum_{i,j=1}^{N} \omega_{ij,t}\delta_{(z_{ij,t},\tilde{z}_{ij,t})},$$

其中 π_t 满足演化方程

$$\partial_t\pi_t + \mathrm{div}_{z,\tilde{z}}(\pi_t\vec{V}_t) = 0$$

对于某个随时间平滑变化的向量场 $\vec{V}_t(z,\tilde{z}) = (V_{1,t}(z,\tilde{z})V_{2,t}(z,\tilde{z}))$。我们断言，如果 π_t^1 不随时间变化，那么 $V_{1,t}$ 必须在 π_t^1 的支撑集（因而也在 π_0^1 的支撑集）中的所有点上都等于零。事实上，只需注意到，如果 $V_{0,t}(z_{ij},\tilde{z}_{ij})$ 不等于 0，那么对于所有足够小的 $t > 0$，$z_{ij,t}$ 都不等于 $z_{i'j'}$,0 （对所有 $i'j'$ 均成立），这意味着当 $t > 0$ 足够小时，π_t^1 的支撑集不同于 π_0^1 的支撑集。这与 π_t^1 不随时间变化的假设相矛盾。

2.1. 提升空间中的上升-下降方程。根据前几节的讨论，方程 (2.7) 类似于欧几里得空间中针对固定能量 U_q 的（投影）梯度下降常微分方程

$$\dot{q}_t = -\nabla U_q(q_t),$$

而方程 2.9 则类似于如下形式的梯度上升方程

$$\dot{p} = \nabla U_p(p_t),$$

其中能量 U_p 固定。

对于极小极大问题，不存在一个固定的待最小化／最大化函数，而是给定一个目标收益函数，并希望找到它的鞍点；在欧几里得情形下，至少可以考虑如下形式的系统

$$\begin{cases}\dot{q}_t = -\nabla_q U(q_t,p_t) \\ \dot{p}_t = \nabla_p U(q_t,p_t),\end{cases}$$

如果变量 $p,\ q$ 存在额外约束，也可以考虑其投影版本。

以上启发我们在测度空间中建立如下梯度上升-下降方程：

$$\begin{cases}\partial_t\gamma_t &= -\mathrm{div}_{(z,\tilde{z}),\omega}(\gamma_t v_\gamma(z,\tilde{z},\omega)), \\ \partial_t\sigma_t &= \mathrm{div}_{\theta,\alpha}(\sigma_t v_\sigma(\theta,\alpha)),\end{cases} \tag{2.10}$$

其中

$$v_\gamma(z,\tilde{z},\omega) = \left(0, \eta_t\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde{z}), \kappa\omega\left(\mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde{z}) - \int \mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde{z}')d\pi_t(\tilde{z}'|z)\right)\right),$$

$$v_\sigma(\theta,\alpha) = \left(\eta_t\nabla_\theta\mathcal{U}(\pi_t,\nu_t;\theta), \kappa\alpha(\mathcal{U}_\nu(\pi_t,\nu_t;\theta) - \int \mathcal{U}_\nu(\pi_t,\nu_t;\theta')d\nu_t(\theta'))\right),$$

且 $\pi_t = \mathcal{F}\gamma_t, \quad \nu_t = \mathcal{F}\sigma_t$。

请注意，为方便起见，这里允许缩放因子 η 随时间变化。第 3 节专门研究方程 (2.10)。具体而言，我们证明其适定性，并表明系统 (2.10) 可以作为相互作用粒子系统的适当极限得到。为了着眼于第 4 节中的应用，我们在第 3 节实际研究一个比 (2.10) 略为一般的系统。

最后，要回到原系统 (1.10)，现在只需通过映射 $\mathcal{F}$ 将动力学 (2.10) 投影过去，正如下文所述。

命题 2.4。 假设(γ, σ) 满足提升动力学(2.10)。那么， $\pi_t = F\gamma_t$， $\nu_t = F\sigma_t$ 这一对满足系统(1.10)。

证明。 取一个测试函数 $\phi(\theta)$，我们看到

$$\begin{aligned}\frac{d}{dt}\int \phi(\theta)d\nu_t = \frac{d}{dt}\int \alpha\phi(\theta)d\sigma_t(\theta,\alpha) &= \eta_t \int \alpha\nabla_\theta\phi(\theta)\cdot\nabla_\theta\mathcal{U}(\pi_t,\nu_t;\theta)d\sigma_t(\theta,\alpha)\\ &\quad + \kappa\int \alpha(\mathcal{U}_\nu(\pi_t,\nu_t;\theta) - \int \mathcal{U}_\nu(\pi_t,\nu_t;\theta')d\nu_t(\theta'))d\sigma_t(\theta,\alpha)\\ &= \eta_t\int \nabla_\theta\phi(\theta)\cdot\nabla_\theta\mathcal{U}(\pi_t,\nu_t;\theta)d\nu_t(\theta)\\ &\quad + \kappa\int (\mathcal{U}_\nu(\pi_t,\nu_t;\theta) - \int \mathcal{U}_\nu(\pi_t,\nu_t;\theta')d\nu_t(\theta'))d\nu_t(\theta),\end{aligned}$$

这是 （1.10）中第二个方程的弱形式。 π 的方程可用类似的方法推导 得。 □

关键在于，通过研究系统（2.10）及其粒子系统近似，我们将隐式地研究系统（1.10）及其粒子系统近似。然而，系统（2.10）的优点是具有直接的拉格朗日表述，我们将加以利用。

2.2. **符号说明**。接下来，我们将使用以下符号：

– $\mu, \tilde{\mu}$ 是定义在 Z 上的概率测度。 μ 是观测数据分布， $\tilde{\mu}$ 表示对 μ 的对抗扰动。 – π 是 Z $\times$ 上的测度，我们将 π 支撑集中的点写作 $(z, \tilde{z})$。 z 可解释为观测数据点，而 $\tilde{z}$ 对应于扰动数据点。 – 每当 π 是概率测度时，我们将用 π_z 表示 π 的第一边缘分布。另一方面，$\pi(\cdot|z)$ 表示根据 π，在第一个变量等于 z 条件下第二个变量的条件分布。 – γ 表示提升空间 $\mathsf{Z}^2 \times \mathbb{R}_+$ 上的概率测度。 – ν 表示定义在 Θ 上的测度。 – σ 表示提升空间 $\Theta \times \mathbb{R}_+$ 上的测度。 – F 是投影映射，其定义域可以是 $\mathsf{P}(\Theta \times \mathbb{R}_+)$，值域为 $\mathsf{M}_+(\Theta)$；也可以是 $\mathsf{P}(\mathsf{Z}^2 \times \mathbb{R}_+)$，值域为 $\mathsf{M}_+(\mathsf{Z}^2)$。 – γ 表示空间 $\mathsf{C}([0,T], \mathsf{Z}\times\mathbb{R}_+)$ 上的概率测度， σ 表示空间 $\mathsf{C}([0,T], \Theta\times\mathbb{R}_+)$ 上的概率测度。空间 $\mathsf{C}([0,T], \Theta\times\mathbb{R}_+)$ 是从区间 $[0,T]$ 到 $\Theta\times\mathbb{R}_+$ 的连续函数所构成的空间，$\mathsf{C}([0,T], \mathsf{Z}^2\times\mathbb{R}_+)$ 的定义类似。这些空间赋予一致收敛度量。 – 我们用 $\check{}\gamma$ 表示提升空间 $\mathsf{Z}^2 \times \mathbb{R}_+^2$ （注意多了一个坐标）上的概率测度，并用 $\check{}\sigma$ 表示提升空间 $\Theta \times \mathbb{R}_+^2$ 上的概率测度。 – $\check{}\gamma$ 表示空间 $\mathsf{C}([0,T], \mathsf{Z} \times\mathbb{R}_+^2)$ 上的概率测度，$\check{}\sigma$ 表示空间 $\mathsf{C}([0,T], \Theta \times \mathbb{R}_+^2)$ 上的概率测度。 – U (π, ν) 表示与测度 π 和 ν 相关联的收益，U_π 和 U_ν 分别表示 U 关于坐标 π 和 ν 的一阶变分。 – 我们用 $\mathsf{H}(\cdot||\cdot)$ 表示定义在同一空间上的任意两个概率测度之间的 KL 散度或香农相对熵。也就是说，给定 υ, υ' 这两个概率测度，若 $\upsilon' \ll \upsilon$，则 $\mathsf{H}(\upsilon'||\upsilon)$ 定义为 $\int \log(\frac{d\upsilon'}{d\upsilon})d\upsilon'$；否则定义为 $+\infty$ 。

3. 从粒子系统到平均场偏微分方程

我们首先介绍一个扩展的常微分方程组，它与式 (1.8) 中的系统密切相关。对于 $i = 1, \dots, N$，令

$$
\begin{aligned}
dZ_t^i &= 0 \\
d\tilde{Z}_t^i &= \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) dt \\
d\omega_t^i &= \kappa \omega_t^i \left(\mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) - \int \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{z}') d\pi_t^N(\tilde{z}' | Z_t^i) \right) dt \\
d\vartheta_t^i &= -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \vartheta_t^i) dt \\
d\alpha_t^i &= -\kappa \alpha_t^i \left(\mathcal{U}_\nu(\pi_N^N, \nu_t^N; \vartheta_t^i) - \int \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta') d\nu_t^N(\theta') \right) dt \\
d\beta_t^i &= 0 \\
d\varrho_t^i &= 0;
\end{aligned}
\tag{3.1}
$$

给定初始条件（Z_0^i，$\tilde{Z}_0^i$，ω_0^i，ϑ_0^i，α_0^i，β_0^i，ϱ_0^i）（可能是随机的），并且

$$
\begin{aligned}
\check{\gamma}_t^N &:= \frac{1}{N} \sum_{i=1}^N \delta_{(Z_t^i, \tilde{Z}_t^i), \omega_t^i, \beta_t^i}, \quad \gamma_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{(Z_t^i, \tilde{Z}_t^i), \omega_t^i}, \quad \pi_t^N := \mathcal{F}[\gamma_t^N] = \frac{1}{N} \sum_{i=1}^N \omega_t^i \delta_{(Z_t^i, \tilde{Z}_t^i)}, \\
\check{\sigma}_t^N &:= \frac{1}{N} \sum_{i=1}^N \delta_{\vartheta_t^i, \omega_t^i, \varrho_t^i}, \quad \sigma_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{\vartheta_t^i, \alpha_t^i}, \quad \nu_t^N := \mathcal{F}[\sigma_t^N] = \frac{1}{N} \sum_{i=1}^N \alpha_t^i \delta_{\vartheta_t^i}.
\end{aligned}
\tag{3.2}
$$

为方便起见，我们将变量 β，$\varrho \in \mathbb{R}_+$ 加入了系统。尤其是，由这些变量不同的初始化方式带来的额外自由度，将在第 3.3 节后半部分派上用场。从 (3.1) 可以看出，这些变量不会影响其余变量随时间的演化。

本节的主要目标是针对 (3.1) 的 1-瓦瑟斯坦混沌初始化，获得混沌传播结果。具体而言，我们将假设，当 $N \to \infty$ 时，有

$$
W_1(\check{\gamma}_0^N, \check{\gamma}_0) \to 0, \qquad W_1(\check{\sigma}_0^N, \check{\sigma}_0) \to 0,
\tag{3.3}
$$

对于某些概率测度 $\check{}\gamma_0$ 和 $\check{}\sigma_0$。实际上，我们将作出一个更强的条件，以确保初始化时某些条件分布的一致性。鉴于动力学 (3.1) 显式依赖于条件分布，这个额外条件（我们在 (3.12) 中明确给出）对我们的分析是必要的。以上，我们使用$\Theta \times \mathbb{R}_+^2$（或 $\mathcal{Z}^2 \times \mathbb{R}_+^2$）中的欧几里得距离来定义 Wasserstein距离 W_1（见注释1.5）。在这些假设下，我们将刻画粒子位置的测度在时间演化过程中的大 N 极限。

事实上，在 N 极限下，系统 (3.1) 预期会表现得如同其中的相互作用被平均场 动力学取代后的系统。这样的系统如下所示。对于 $i = 1, \dots, N$，令

$$
\begin{aligned}
dZ_t^{mf,i} &= 0 \\
d\tilde{Z}_t^{mf,i} &= \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^{mf}, \nu_t^{mf}; Z_t^{mf,i}, \tilde{Z}_t^{mf,i}) dt \\
d\omega_t^{mf,i} &= \kappa \omega_t^{mf,i} \left(\mathcal{U}_\pi(\pi_t^{mf}, \nu_t^{mf}; Z_t^{mf,i}, \tilde{Z}_t^{mf,i}) - \int \mathcal{U}_\pi(\pi_t^{mf}, \nu_t^{mf}; Z_t^{mf,i}, \tilde{z}') d\pi_t^{mf}(\tilde{z}' | Z_t^{mf,i}) \right) dt \\
d\vartheta_t^{mf,i} &= -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t^{mf}, \nu_t^{mf}; \vartheta_t^{mf,i}) dt \\
d\alpha_t^{mf,i} &= -\kappa \alpha_t^{mf,i} \left(\mathcal{U}_\nu(\pi_N^{mf}, \nu_t^{mf}; \vartheta_t^{mf,i}) - \int \mathcal{U}_\nu(\pi_t^{mf}, \nu_t^{mf}; \theta') d\nu_t^{mf}(\theta') \right) dt \\
d\beta^{mf,i} &= 0 \\
d\varrho^{mf,i} &= 0;
\end{aligned}
\tag{3.4}
$$

其初始条件与 (3.1) 相同，并且 $\pi_t^{mf} = \mathcal{F}(\gamma_t)$ $\nu_t^{mf} = \mathcal{F}(\sigma_t)$，而 (γ, σ) 满足带初始条件 γ_0, σ_0 的方程 (2.10)（我们将在第 3.1 节证明该方程的适定性）；我们回顾一下，映射 $\mathcal{F}$ 已在第 2 节中引入。平均场系统 (3.4) 与原始粒子系统 (3.1) 之间的根本区别在于，决定 (3.4) 中动力学的测度可以视为固定的，且独立于演化中的粒子；而在系统 (3.1) 中，驱动动力学显式依赖于与底层演化中的粒子相关联的经验测度 π^N, ν^N。

为了推导系统 (3.1) 的混沌传播结果，我们需要证明平均场系统 (3.4) 定义良好，并控制演化 (3.1) 与 (3.4) 之间的差异；在假设 1.4 下，我们将证明这一点。为此，引入一些额外的数学结构会很方便，这将使我们能够使用标准的分析论证：我们将在$\mathcal{Z}^2 \times \mathbb{R}_+^2$ 上的连续路径的测度空间以及 $\Theta \times \mathbb{R}_+^2$上开展工作，并最终利用不动点论证来证明平均场方程的适定性。我们首先引入一族粒子演化，它们将在我们的分析中发挥关键作用。

固定 $T > 0$，令$\mathcal{A}_T$为满足以下条件的配对 $(\check{\gamma}, \check{\sigma}) \in \mathcal{P}(\mathcal{C}([0,T], \mathcal{Z}^2 \times \mathbb{R}_+^2)) \times \mathcal{P}(\mathcal{C}([0,T], \Theta \times \mathbb{R}_+^2))$:

i) $\mathcal{F}\sigma_t$和 $\mathcal{F}\gamma_t$对所有 $t \in [0,T]$均为概率测度。ii) $\mathcal{F}[\gamma_t](\cdot \times \mathcal{Z}) = \mathcal{F}[\gamma_0](\cdot \times \mathcal{Z})\ \forall t \in [0,T]$。

这里以及下文中，对于给定的 $\check{\gamma}$，我们用 γ 表示 $\check{\gamma}$ 在映射 $\{(z_t, \tilde{z}_t, \omega_t, \varrho_t)\} \mapsto \{(z_t, \tilde{z}_t, \omega_t)\}$下的推前。稍微滥用一下记号，下文中我们可能会混用 $\mathcal{F}\check{\gamma}_t$和 $\mathcal{F}\gamma_t$；同理，我们也可以关联 $\check{\sigma}$和 σ。

与 $(\check{\gamma}, \check{\sigma}) \in \mathcal{A}_T$相对应，我们考虑以下多维常微分方程:

$$
\begin{aligned}
dZ_t^{\check{\gamma},\check{\sigma}} &= 0 \\
d\tilde{Z}_t^{\check{\gamma},\check{\sigma}} &= \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t, \nu_t; Z_t^{\check{\gamma},\check{\sigma}}, \tilde{Z}_t^{\check{\gamma},\check{\sigma}}) dt \\
d\omega_t^{\check{\gamma},\check{\sigma}} &= \kappa\omega_t \left(\mathcal{U}_\pi(\pi_t, \nu_t; Z_t^{\check{\gamma},\check{\sigma}}, \tilde{Z}_t^{\check{\gamma},\check{\sigma}}) - \int \mathcal{U}_\pi(\pi_t, \nu_t; Z_t^{\check{\gamma},\check{\sigma}}, \tilde{z}') d\pi_t(\tilde{z}' | Z_t^{\check{\gamma},\check{\sigma}}) \right) dt \\
d\vartheta_t^{\check{\gamma},\check{\sigma}} &= -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \vartheta_t^{\check{\gamma},\check{\sigma}}) dt \\
d\alpha_t^{\check{\gamma},\check{\sigma}} &= -\kappa\alpha_t \left(\mathcal{U}_\nu(\pi_t, \nu_t; \vartheta_t^{\check{\gamma},\check{\sigma}}) - \int \mathcal{U}_\nu(\pi_t, \nu_t; \theta') d\nu_t(\theta') \right) dt \\
d\beta_t^{\check{\gamma},\check{\sigma}} &= 0 \\
d\varrho_t^{\check{\gamma},\check{\sigma}} &= 0 \\
\pi_t = \mathcal{F}(\gamma_t), &\qquad \nu_t = \mathcal{F}(\sigma_t),
\end{aligned}
\tag{3.5}
$$

初始条件为

$$
((Z_0^{\check{\gamma},\check{\sigma}}, \tilde{Z}_0^{\check{\gamma},\check{\sigma}}), \omega_0^{\check{\gamma},\check{\sigma}}, \varrho_0^{\check{\gamma},\check{\sigma}}) = ((\xi, \tilde{\xi}), \omega_0, \varrho_0) \sim \check{\gamma}_0, \qquad (\vartheta_0^{\check{\gamma},\check{\sigma}}, \alpha_0^{\check{\gamma},\check{\sigma}}, \beta_0^{\check{\gamma},\check{\sigma}}) = (\vartheta, \alpha_0, \beta_0) \sim \check{\sigma}_0. \tag{3.6}
$$

$(\check{\gamma}, \check{\sigma}) \in \mathcal{A}_T$ 这一事实尤其用于使方程 (3.5) 中的项 $\pi_t(\cdot | Z^{\gamma}{}_{,t}{}^{\sigma})$ 有意义。事实上，令 $\pi_{t,z}$ 表示 π_t 在 z 坐标上的边缘分布。根据 $\check{\gamma}$ 假设，$\pi_{t,z} = \pi_{0,z}$；而不失一般性，可以假定 $Z^{\gamma}{}_{,t}{}^{\sigma} = Z^{\gamma}{}_{,0}{}^{\sigma}$ 属于 $\pi_{0,z}$ 的支撑集。条件分布 $\pi_t(\cdot | Z^{\gamma}{}_{,t}{}^{\sigma})$ 因此根据解体定理定义良好。

方程 (3.5) 是一个多维经典常微分方程，描述了一个孤立粒子在外生测度驱动的动力学下的运动。一个关键观察是，在假设 1.4 下，方程 (3.5) 由 Lipschitz 系数驱动，因此根据 Caratheodory 存在定理（见 [24] 中的定理 5.3），该方程是适定的。假设 1.4 和格朗沃尔不等式进一步$\check{\gamma}$给出了关于 $\omega^{\gamma,\sigma}$ 和 $\alpha^{\gamma,\sigma}$的界。为便于查阅，我们在下一个命题中总结这些观察结果。

命题 3.1. 在假设 *1.4* 下，对于任意给定的初始值，(3.5) 存在唯一解。此外，我们有

$$\omega_t^{\check{\gamma},\check{\sigma}} \in [0, \omega_0 e^{2\kappa M t}], \quad \alpha_t^{\check{\gamma},\check{\sigma}} \in [0, \alpha_0 e^{2\kappa M t}]; \qquad \forall T \geq t > 0.$$

对于给定的 $T > 0$，现在考虑映射：

$$\Psi_T : \mathcal{A}_T \mapsto \mathcal{P}(\mathcal{C}([0,T], \mathcal{Z}^2 \times \mathbb{R}_+^2)) \times \mathcal{P}(\mathcal{C}([0,T], \Theta \times \mathbb{R}_+^2))$$

定义为

$$\Psi_T(\check{\gamma}, \check{\sigma}) = (\Psi_T^1(\check{\gamma}, \check{\sigma}), \Psi_T^2(\check{\gamma}, \check{\sigma})) := (\mathrm{Law}[(Z^{\check{\gamma},\check{\sigma}}, \tilde{Z}^{\check{\gamma},\check{\sigma}}), \omega^{\check{\gamma},\check{\sigma}}, \varrho^{\check{\gamma},\check{\sigma}}], \mathrm{Law}[\vartheta^{\check{\gamma},\check{\sigma}}, \alpha^{\check{\gamma},\check{\sigma}}, \beta^{\check{\gamma},\check{\sigma}}]),$$

即， Ψ_T 将提升空间中测度空间的路径映射到其自身。此外， Ψ_T 将$\mathcal{A}_T$ 映射到其自身，正如我们将在下一个引理中所述。

引理 3.2. 在假设 *1.4* 下，可得

$$\Psi_T(\mathcal{A}_T) \subseteq \mathcal{A}_T.$$

此外，对于每个 $(\check{\gamma}, \check{\sigma}) \in \mathcal{A}_T$，我们有

$$\mathcal{F}[(\Psi_T^1(\check{\gamma}, \check{\sigma}))_t](\cdot \times \mathcal{Z}) = \mathcal{F}[\check{\gamma}_t](\cdot \times \mathcal{Z}).$$

证明。 该结果可由注 1.3 以及 $dZ^{\gamma,\ t}{}^{\sigma} = 0$ 这一事实直接得出。 □

出于技术上的考虑，引入集合$\mathcal{A}_T$的一个版本会很方便，其中元素的支撑集满足某个有界性条件。具体而言，对于给定的 $T > 0$ 和$D > 0$，令$\mathcal{A}_{T,D}$ 表示集合

$$\mathcal{A}_{T,D} := \{(\check{\gamma}, \check{\sigma}) \in \mathcal{A}_T \text{ s.t. } \check{\gamma}_t(\mathcal{Z}^2 \times [0, De^{2\kappa M t}] \times [0, D]) = 1, \quad \check{\sigma}_t(\Theta \times [0, De^{2\kappa M t}] \times [0, D]) = 1, \quad \forall t \in [0,T]\}.$$

特别地，对于 $(\check{\gamma}, \check{\sigma}) \in \mathcal{A}_{T,D}$，由 (3.6) 得到的权重 (ω_0, ϱ_0) 和 (α_0, β_0) 可以假定属于 $[0, D]^2$。结合命题 3.1，我们可以推得，对于 $(\check{\gamma}, \check{\sigma}) \in \mathcal{A}_{T,D}$，动力学 (3.5) 中的权重 $\omega^{\check{\gamma},\ t}{}^{\check{\sigma}}, \alpha^{\check{\gamma},\ t}{}^{\check{\sigma}}$ 可以由 $De^{2\kappa M t}$ 从上方界定。我们将此总结为以下引理。

引理 3.3. 对于每个 $T, D > 0$，都有 $\Psi_T(\mathcal{A}_{T,D}) \subseteq \mathcal{A}_{T,D}$。

在假设 1.4 下，我们证明可以用两对元素自身的距离，控制其像之间的距离： $(\check{\gamma}^1, \check{\sigma}^1)$ 和 $(\check{\gamma}^2, \check{\sigma}^2)$ 属于 $\mathcal{A}_{T,\ D}$。给定 $p \geq 1$，我们使用

$$W_{t,p}^p(\upsilon, \upsilon') := \inf_{\Upsilon \in \Gamma(\upsilon, \upsilon')} \int \sup_{s \in [0,t]} |u_s - v_s|^p d\Upsilon(x, y), \tag{3.7}$$

来比较同一路径空间上的两个概率测度。特别地，我们将使用这些距离来比较任意提升空间中路径上的测度。

首先，在限制映射 $\mathcal{F}$ 定义域时，我们证明其连续性。

引理 3.4. 设 σ、σ' 是 $\Theta \times [0, D]$ 上的两个概率测度，其中 D 是一个固定常数，并假设 $F\sigma$ 和 $F\sigma'$ 也是概率测度。

Then

$$W_p^p(\mathcal{F}(\sigma), \mathcal{F}(\sigma')) \leq C_{\Theta,p,D} W_p(\sigma, \sigma'),$$

其中常数 $C_{\Theta,\ p,\ D}$ 可写为 $C_{\Theta,\ p,\ D} = \mathrm{diam}(\Theta)^{p-1}(\mathrm{diam}(\Theta) + D)$。

特别是，当其定义域受到限制时，映射 F 在1-*Wasserstein* 意义下是 *Lipschitz* 的。

证明。 首先注意到，使 σ Fσ 成为概率测度的测度，恰好是满足 $\int \alpha d\sigma(\theta,\alpha) = 1$ 的测度。

我们首先证明 $p=1$ 的结果。

假设 σ 和 σ' 具有 $\sigma=\sigma_n=\frac{1}{n}\sum_{i=1}^n \delta_{(\theta_i,\alpha_i)}$ 和 $\sigma'=\sigma'_n=\frac{1}{n}\sum_{i=1}^n \delta_{(\theta'_i,\alpha'_i)}$ 的形式。众所周知，在这种情况下，存在一个置换 T: $\{1,\dots,n\}\mapsto\{1,\dots,n\}$，使得 $W_1(\sigma_n,\sigma'_n)=\frac{1}{n}\sum_{i=1}^n |(\theta_i,\alpha_i)-(\theta'_{T(i)}\ \alpha'_{T(i)})|$. 现在，我们可以将测度 $\mathcal{F}\sigma_n$ 和 $\mathcal{F}\sigma'_n$ 写为

$$\mathcal{F}\sigma_n=\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i,\alpha'_{T(i)}\}\delta_{\theta_i}+\frac{1}{n}\sum_{i=1}^n(\alpha_i-\min\{\alpha_i,\alpha'_{T(i)}\})\delta_{\theta_i}$$

and

$$\mathcal{F}\sigma'_n=\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i,\alpha'_{T(i)}\}\delta_{\theta'_i}+\frac{1}{n}\sum_{i=1}^n(\alpha'_{T(i)}-\min\{\alpha_i,\alpha'_{T(i)}\})\delta_{\theta'_i}.$$

注意，可以用来自 $\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i,\alpha'_{T(i)}\}\delta_{\theta_i}$ 的质量来补足 $\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i,\alpha'_{T(i)}\}\delta_{\theta'_i}$ 中所需的质量。我们进行如下质量转移：对于每个 i，从 θ_i 向 $\theta'_T(i)$ 转移 $\min\{\alpha_i,\alpha'_{T(i)}\}$ 个单位的质量。此次质量转移的总成本为 $\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i,\alpha'_{T(i)}\}|\theta_i-\theta'_{T(i)}|\le DW_1(\sigma_n,\sigma'_n)$。最后，质量 $\frac{1}{n}\sum_{i=1}^n(\alpha_i-\min\{\alpha_i,\alpha'_T(i)\})\delta_{\theta_i}$ 可用于补足 $\frac{1}{n}\sum_{i=1}^n(\alpha'_{T(i)}-\min\{\alpha_i,\alpha'_{T(i)}\})\delta_{\theta'_i}$ 中所需的质量。这种质量转移可以通过任意方式进行，关键在于，此类质量转移的总成本不会超过待转移的质量总量 $\frac{1}{n}\sum_{i=1}^n(\alpha_i-\min\{\alpha_i,\alpha'_{T(i)}\})$ 乘以集合 Θ 的直径，其中该质量总量小于 $W_1(\sigma_n,\sigma'_n)$。归根结底，$W_1(\mathrm{F}\sigma_n,\mathrm{F}\sigma'_n)\le(D+\mathrm{diam}(\Theta))W_1(\sigma_n,\sigma_n)$。

我们可以将其推广到任意概率测度 σ，σ'，只需注意到：1）任意概率测度 σ 都可以在弱意义下由经验概率测度 σ_n 逼近，其中 n 逐渐增大；2）映射 $\mathcal{F}$ 在弱意义下连续（可直接验证）；3）鉴于所有测度都支撑于一个固定的有界集合上，Wasserstein度量关于弱收敛是连续的。

最后，要推广到任意 $p\ge 1$，注意到 $W_1(\sigma,\sigma')\le W_p(\sigma,\sigma')$，而 $W_p^p(\mathcal{F}\sigma,\mathcal{F}\sigma')\le\mathrm{diam}(\Theta)^{p-1}W_1(\mathcal{F}\sigma,\mathcal{F}\sigma')$。□

注 3.5。 引理 *3.4* 对作用于测度 $\gamma\in\mathcal{Z}^2\times[0,D]$ 上的 F 也同样成立，只需作相应修改。

现在，我们可以先验地控制两组不同测度对 $(\check{\gamma}^i,\check{\sigma}^i)$ 和 $i=1,2$ 对应的方程 (3.5) 的解之间的差异。

引理 3.6. 设 $T,D>0$。假设假设 *1.4* 成立。对于 $i=1,2$，令 $(\check{\gamma}^i,\check{\sigma}^i)\in\mathcal{A}_{T,D}$，并记

$$\zeta^i=(Z^{\check{\gamma}^i,\check{\sigma}^i},\tilde{Z}^{\check{\gamma}^i,\check{\sigma}^i},\omega^{\check{\gamma}^i,\check{\sigma}^i},\vartheta^{\check{\gamma}^i,\check{\sigma}^i},\alpha^{\check{\gamma}^i,\check{\sigma}^i},\beta^{\check{\gamma}^i,\check{\sigma}^i},\varrho^{\check{\gamma}^i,\check{\sigma}^i})$$

由(3.5)确定的相应演化。我们假设 $Z^{\check{\gamma}^i,\check{}_0\ \sigma^i}$（尽管它可能是随机的）属于 $\pi^i_{0,z}$ 的支撑集，其中 π^i_0 在 z 坐标上的边缘分布。我们还假设 $\omega^i_0,\varrho^i_0,\alpha^i_0,\beta^i_0\in[0,D]$.

那么存在一个常数 $K_{T,D}$，它仅依赖于 T, D、函数 η, κ 以及假设 *1.4* 中的常数，使得对于所有 $t\in[0,T]$，我们有

$$\mathbb{E}|\zeta^1_t-\zeta^2_t|\le\mathbb{E}[\sup_{0\le s\le t}|\zeta^1_s-\zeta^2_s|]$$

$$\le K_{T,D}\left(\mathbb{E}|\zeta^1_0-\zeta^2_0|+\int_0^t\{W_1(\check{\gamma}^1_s,\check{\gamma}^2_s)+W_1(\check{\sigma}^1_s,\check{\sigma}^2_s)+\mathbb{E}\left(W_1(\pi^2_s(\cdot|Z_0^{\check{\gamma}^2,\check{\sigma}^2}),\pi^1_s(\cdot|Z_0^{\check{\gamma}^1,\check{\sigma}^1}))\right)\}ds\right).$$

上述期望是针对两个系统预先规定的（联合）初始条件取的。

证明。Fr 由式 (3.5) 以及假设 1.4 中的 Lipschitz 连续性和有界性条件可得 t

$$|\frac{d}{dt}(\tilde{Z}_t^{\check{\gamma}^1,\check{\sigma}^1}-\tilde{Z}_t^{\check{\gamma}^2,\check{\sigma}^2})|=\eta_t\Big|\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi_t^1,\nu_t^1;Z_t^{\check{\gamma}^1,\check{\sigma}^1},\tilde{Z}_t^{\check{\gamma}^1,\check{\sigma}^1})-\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi_t^2,\nu_t^2;Z_t^{\check{\gamma}^2,\check{\sigma}^2},\tilde{Z}_t^{\check{\gamma}^2,\check{\sigma}^2}))\Big|$$
$$\leq \eta_t L\{|Z_t^{\check{\gamma}^1,\check{\sigma}^1}-Z_t^{\check{\gamma}^2,\check{\sigma}^2}|+|\tilde{Z}_t^{\check{\gamma}^1,\check{\sigma}^1}-\tilde{Z}_t^{\check{\gamma}^2,\check{\sigma}^2}|+W_1(\pi_t^1,\pi_t^2)+W_1(\nu_t^1,\nu_t^2)\}.$$

对系统的其他分量进行类似分析，并利用假设 $(\check{\gamma}^i,\check{\sigma}^i)\in\mathcal{A}_{T,D}$ 以及假设 1.4，我们可以推出存在一个常数 $C_{T,D}$，使得对所有 $t\in[0,T]$

$$|\zeta_t^1-\zeta_t^2|\leq|\zeta_0^1-\zeta_0^2|$$
$$+C_{T,D}\int_0^t\Big\{|\zeta_s^1-\zeta_s^2|+W_1(\pi_s^1,\pi_s^2)+W_1(\nu_s^1,\nu_s^2)+W_1(\pi_s^1(\cdot|Z_s^{\check{\gamma}^1,\check{\sigma}^1}),\pi_s^2(\cdot|Z_s^{\check{\gamma}^2,\check{\sigma}^2}))\Big\}ds.$$

因此，利用格朗沃尔不等式，我们得到对所有 $t\in[0,T]$

(3.8)

$$|\zeta_t^1-\zeta_t^2|\leq\sup_{0\leq s\leq t}|\zeta_s^1-\zeta_s^2|$$
$$\leq e^{C_{T,D}T}\left(|\zeta_0^1-\zeta_0^2|+C_{T,D}\int_0^t\Big\{W_1(\pi_s^1,\pi_s^2)+W_1(\nu_s^1,\nu_s^2)+W_1(\pi_s^1(\cdot|Z_s^{\check{\gamma}^1,\check{\sigma}^1}),\pi_s^2(\cdot|Z_s^{\check{\gamma}^2,\check{\sigma}^2}))\Big\}ds\right).$$

现在，由于 $Z^{\check{\gamma}^i,\check{}}{}_s{}^{\sigma^i}=Z^{\check{\gamma}^i,\check{}}{}_0{}^{\sigma^i}$，可得

$$\mathbb{E}\left(W_1(\pi_s^2(\cdot|Z_s^{\check{\gamma}^2,\check{\sigma}^2}),\pi_s^1(\cdot|Z_s^{\check{\gamma}^1,\check{\sigma}^1}))\right)=\mathbb{E}\left(W_1(\pi_s^2(\cdot|Z_0^{\check{\gamma}^2,\check{\sigma}^2}),\pi_s^1(\cdot|Z_0^{\check{\gamma}^1,\check{\sigma}^1}))\right).$$

由此以及 (3.8) 可推出，对所有 $t\in[0,T]$

$$\mathbb{E}[|\zeta_t^1-\zeta_t^2|]\leq\mathbb{E}[\sup_{0\leq s\leq t}|\zeta_s^1-\zeta_s^2|]$$
$$\leq e^{C_{T,D}T}\left(|\zeta_0^1-\zeta_0^2|+C_{T,D}\int_0^t W_1(\pi_s^1,\pi_s^2)+W_1(\nu_s^1,\nu_s^2)+\mathbb{E}\left(W_1(\pi_s^2(\cdot|Z_0^{\check{\gamma}^2,\check{\sigma}^2}),\pi_s^1(\cdot|Z_0^{\check{\gamma}^1,\check{\sigma}^1}))\right)ds\right).$$

最后，我们应用命题 3.1 和引理 3.4。 □

一般而言，$\mathbb{E}\left(W_1(\pi_s^2(\cdot|Z^{\gamma^2,}{}_0{}^{\sigma^2})\ \pi_s^1(\cdot|Z^{\gamma^1,}{}_0{}^{\sigma^1}))\right)$无法由 π_s^2 与 π_s^1之间的Wasserstein距离从上界定。为了能够更直接地处理该项，而无需对原始测度作出进一步假设，我们引入以下概念，该概念与 Knothe 输运的概念相关（参见 [42], 或 [6]中的第 2.3 节）：给定两组 $\pi^1:=\{\pi_s^1\}_{0\leq s\leq T}$和 $\pi^2:=\{\pi_s^2\}_{0\leq s\leq T}$，我们将它们的成本定义为$\tilde{W}_{t,\ 1}$，即

$$\tilde{W}_{t,1}(\pi^1,\pi^2):=\sup_{s\in[0,t]}\inf_{\upsilon_s\in\Gamma_{\mathrm{Opt}}(\pi_{s,z}^1,\pi_{s,z}^2)}\{\int W_1(\pi_s^2(\cdot|z^2),\pi_s^1(\cdot|z^1))d\upsilon_s(z^1,z^2)\};\tag{3.9}$$

在上文中，我们将$\pi_{s,\ z}^i$ 解释为测度 π_s^i在 z 坐标上的边缘分布，而$\Gamma_{\mathrm{Opt}}(\pi_{s,\ z}^1,\ \pi_{s,\ z}^2$)是实现$\pi_{s,\ z}^1$与$\pi_{s,\ z}^2$之间的 1-瓦瑟斯坦距离的最优耦合集合。

有了这一概念，我们便可以陈述并证明引理 3.6 的以下推论。

推论 3.7。 假设引理 *3.6* 中的假设成立。进一步假设 $\check{}\gamma_0^1=\check{}\gamma_0^2$ 和$\check{}\sigma_0^1=\check{}\sigma_0^2$。那么存在一个常数 $\tilde{C}_{T,D}$，它仅依赖于 M, L, T, D, κ, η，使得对所有$t\in[0,T]$

$$W_{t,1}(\Psi_T^1(\check{\gamma}^1,\check{\sigma}^1),\Psi_T^1(\check{\gamma}^2,\check{\sigma}^2))+W_{t,1}(\Psi_T^2(\check{\gamma}^1,\check{\sigma}^1),\Psi_T^2(\check{\gamma}^2,\check{\sigma}^2))+\tilde{W}_{t,1}(\Psi_T^1(\pi^1),\Psi_T^1(\pi^2))$$
$$\leq t\tilde{C}_{T,D}\{W_{t,1}(\check{\gamma}^1,\check{\gamma}^2)+W_{t,1}(\check{\sigma}^1,\check{\sigma}^2)+\tilde{W}_{t,1}(\pi^1,\pi^2)\}.$$

这里我们稍微滥用一下记号，用 $\Psi_T^1(\pi^i)$ 表示测度的集合 $\{F((\Psi_T^1(\check{\gamma}^i, \check{\sigma}^i))_s)\}_{s\in[0,T]}$。

证明。取 ζ^i 和 $i=1$，在引理（3.6）中令二者具有相同的初始条件，分别从 $\check{\gamma}_0$ 中采样 $(Z_0^1, \tilde{Z}_0^1, \omega_0^1, \varrho_0^1)$，从 $\check{\sigma}_0$ 中采样 $(\vartheta_0^1, \alpha_0^1, \beta_0^1)$。根据Wasserstein距离的定义，可得

$$W_{t,1}(\Psi_T^1(\check{\gamma}^1,\check{\sigma}^1),\Psi_T^1(\check{\gamma}^2,\check{\sigma}^2))+W_{t,1}(\Psi_T^2(\check{\gamma}^1,\check{\sigma}^1),\Psi_T^2(\check{\gamma}^2,\check{\sigma}^2))\le \mathbb{E}\sup_{0\le s\le t}|\zeta_s^1-\zeta_s^2|.$$

现在，由于 $\check{\gamma}_0^1=\check{\gamma}_0^2$ 和 $(\check{\gamma}^1, \check{\sigma}^1)$，$(\check{\gamma}^2, \check{\sigma}^2)\in\mathcal{A}_T$，$\tilde{W}_{t,1}(\pi^1, \pi^2)$ 中的最优耦合 υ_s 全都是测度 $\pi^1_{0,z}$ 的恒等耦合；对于 $\tilde{W}_{t,1}(\Psi_T^1(\pi^1)\ \Psi_T^1(\pi^2))$，情况完全相同。因此

$$\begin{aligned}
&W_{t,1}(\Psi_T^1(\check{\gamma}^1,\check{\sigma}^1),\Psi_T^1(\check{\gamma}^2,\check{\sigma}^2))+W_{t,1}(\Psi_T^2(\check{\gamma}^1,\check{\sigma}^1),\Psi_T^2(\check{\gamma}^2,\check{\sigma}^2))\\
&\le \mathbb{E}\sup_{0\le s\le t}|\zeta_s^1-\zeta_s^2|\le K_{T,D}\int_0^t\{W_1(\check{\gamma}_s^1,\check{\gamma}_s^2)+W_1(\check{\sigma}_s^1,\check{\sigma}_s^2)+\tilde{W}_{s,1}(\pi^1,\pi^2)\}ds\\
&\le K_{T,D}\int_0^t\{W_{s,1}(\check{\gamma}^1,\check{\gamma}^2)+W_{s,1}(\check{\sigma}^1,\check{\sigma}^2)+\tilde{W}_{s,1}(\pi^1,\pi^2)\}ds\\
&\le tK_{T,D}\{W_{t,1}(\check{\gamma}^1,\check{\gamma}^2)+W_{t,1}(\check{\sigma}^1,\check{\sigma}^2)+\tilde{W}_{t,1}(\pi^1,\pi^2)\}.
\end{aligned}$$

同理，

$$\tilde{W}_{t,1}(\Psi_T^1(\pi^1),\Psi_T^1(\pi^2))\le \mathbb{E}\sup_{0\le s\le t}|\zeta_s^1-\zeta_s^2|.$$

综合以上估计，即得所需结果。 □

3.1. **平均场偏微分方程的适定性**。现在我们来寻找刻画系统（2.10）解的常微分方程组。自然的候选方程是平均场方程

$$\begin{aligned}
&(Z^{mf},\tilde{Z}^{mf},\omega^{mf},\vartheta^{mf},\alpha^{mf},\beta^{mf},\varrho^{mf}):=(Z^{\check{\gamma},\check{\sigma}},\tilde{Z}^{\check{\gamma},\check{\sigma}},\omega^{\check{\gamma},\check{\sigma}},\vartheta^{\check{\gamma},\check{\sigma}},\alpha^{\check{\gamma},\check{\sigma}},\beta^{\check{\gamma},\check{\sigma}},\varrho^{\check{\gamma},\check{\sigma}});\\
&\qquad\text{with}\quad \check{\gamma}=\mathrm{Law}[((Z^{mf},\tilde{Z}^{mf}),\omega^{mf},\varrho^{mf})],\quad \check{\sigma}=\mathrm{Law}[(\vartheta^{mf},\alpha^{mf},\beta^{mf})].
\end{aligned}\tag{3.10}$$

事实上，假设这样的平均场方程存在，我们可以验证，令

$$\gamma=\mathrm{Law}[((Z^{mf},\tilde{Z}^{mf}),\omega^{mf})],\ \text{and}\ \sigma=\mathrm{Law}[(\vartheta^{mf},\alpha^{mf})]$$

我们满足 (2.10)。任取两个测试函数 ϕ，φ。我们有

$$\frac{d}{dt}\mathbb{E}[\phi(Z_t^{mf},\tilde{Z}_t^{mf}),\omega_t^{mf})]=\mathbb{E}\left[\nabla_{(z,\tilde{z}),\omega}\phi((Z_t^{mf},\tilde{Z}_t^{mf}),\omega_t^{mf})\cdot[(\frac{d}{dt}Z_t^{mf},\frac{d}{dt}\tilde{Z}_t^{mf}),\frac{d}{dt}\omega_t^{mf}]^\top\right],$$

and

$$\frac{d}{dt}\mathbb{E}[\varphi(\vartheta_t^{mf},\omega_t^{mf})]=\mathbb{E}\left[\nabla_{\theta,\alpha}\varphi(\vartheta_t^{mf},\alpha_t^{mf})\cdot[\frac{d}{dt}\vartheta_t^{mf},\frac{d}{dt}\alpha_t^{mf}]^\top\right].$$

利用按定义给定的 π，ν 所对应的动力学 (3.5)，我们便可在弱意义下得到 (2.10)。注意，由于驱动该动力学的测度来自动力学自身的分布，因此先前的论证不能直接适用。然而，通过证明常微分方程组 (3.10) 的适定性，这个问题便迎刃而解。

定理3.8。设 $D>0$，并假设 $\check{\gamma}_0$ 和 $\check{\sigma}_0$ 是两个概率测度，满足 $\check{\gamma}_0(Z^2\times[0,D]^2)=1$、$\check{\sigma}_0(\Theta\times[0,D]^2)=1$，且 $F\check{\sigma}_0$ 和 $F\check{\gamma}_0$ 均为概率测度。那么，在假设 *1.4* 下，对于初始分布 $(\check{\gamma}_0,\check{\sigma}_0)$，平均场系统(3.10) 存在一个（逐路径）唯一解。

证明。考虑集合 $\mathrm{A}_{T,D}(\check{\gamma}_0,\check{\sigma}_0):=\{(\check{\gamma},\check{\sigma})\in\mathrm{A}_{T,D}$ 使得 $\check{\gamma}_0=\check{\gamma}_0,\check{\sigma}_0=\check{\sigma}_0\}$。显然，赋予度量 $W_{T,1}(\check{\gamma}^1,\check{\gamma}^2)+W_{T,1}(\check{\sigma}^1,\check{\sigma}^2)$ 的 $\mathrm{A}_{T,D}(\check{\gamma}_0,\check{\sigma}_0)$ 是一个完备度量空间。还要注意，根据引理3.7，可以取足够小的 $T>0$，使得Ψ 在空间 $\mathrm{A}_{T,D}(\check{\gamma}_0,\check{\sigma}_0)$ 中将量 $W_{T,1}(\check{\gamma}^1,\check{\gamma}^2)+W_{T,1}(\check{\sigma}^1,\check{\sigma}^2)+\tilde{W}_{T,1}(\pi^1,\pi^2)$ 压缩;

注意，后一个量支配空间 $\mathcal{A}_{T,D}(\check{}\gamma_0,\check{}\sigma_0)$ 中的度量。因此，固定点方程存在唯一解 $(\check{}\gamma,\check{}\sigma) \in \mathcal{A}_{T,D}(\check{}\gamma_0, \check{}\sigma_0)$

$$\Psi(\check{\gamma},\check{\sigma}) = (\check{\gamma},\check{\sigma}).$$

根据定义，平均场系统 (3.10) 在区间$[0,T]$上成立且适定。通过延拓，适定性可以任意延长。□

注 3.9. 由于(2.10) 定义良好，我们也可以得出结论：平均场粒子系统(3.4) 定义良好，因为它可以通过代入平均场概率律得到；但初始条件并非从$\check{}\gamma_0,\check{}\sigma_0$ 中抽样，而是按照系统(3.1)取定。

3.2. **混沌传播**。 在陈述混沌传播结果之前，我们先给出一个引理。

引理 *3.10.* 设 $(\check{}\gamma_0, \check{}\sigma_0)$ 满足 $\check{}\gamma_0(Z^2\times[0,D]^2) = 1$，$\check{}\sigma_0(\Theta\times[0,D]^2) = 1$，且 $F\check{}\sigma_0$ 和 $F\check{}\gamma_0$ 均为概率测度。设 $(\check{}\gamma, \check{}\sigma)$ 是在 $(\check{}\gamma_0, \check{}\sigma_0) = (\check{}\gamma_0, \check{}\sigma_0)$ 条件下，*(3.10)* 的概率律。设 z_0' 和 z_0 是 $\pi_{0,z}$ 的支撑集中的任意两点。那么，对每个 $t\in[0,T]$，我们有

$$\sup_{s\in[0,t]} W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0)) \le K_{T,D}(W_1(\pi_0(\cdot|z_0'),\pi_0(\cdot|z_0)) + |z_0 - z_0'|). \tag{3.11}$$

证明。考虑一个如（3.4）所示的粒子，将其记为 ζ，并将其初始化为 $Z_0 = \tilde{z_0}$ 和 $(Z_0,\omega_0,\varrho_0) \sim \check{}\gamma_0(\cdot|z_0)$ 以及 $(\vartheta_0,\alpha_0,\beta_0) \sim \check{}\sigma_0$。同样，考虑另一个如（3.4）所示的粒子，将其记为 ζ'，并将其初始化为 $Z_0' = z_0'$、$(\tilde{Z_0'},\omega_0',\varrho_0') \sim \check{}\gamma_0(\cdot|z_0)$ 和 $(\vartheta_0',\alpha_0',\beta_0') = (\vartheta_0,\alpha_0,\beta_0)$。此时，我们不具体说明变量 $\tilde{z}$、ω、ϱ 的初始条件的联合分布，但可以理解为，在下文的计算中已固定了这样一种耦合。

由引理（3.6）可得，对每个 $t\in[0,T]$

$$\mathbb{E}[\sup_{0\le s\le t}|\zeta'-\zeta|] \le K_{T,D}\mathbb{E}|\zeta_0'-\zeta_0| + K_{T,D}\int_0^t W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0))ds.$$

根据 Wasserstein距离 的定义，上述表达式的左侧可由 $W_1(\pi_s(\cdot|z_0')\ \pi_s(\cdot|z_0))$ 从下方界定，其中任意 $s\in[0,t]$ 均成立，因此

$$\sup_{s\in[0,t]} W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0)) \le K_{T,D}\mathbb{E}|\zeta_0'-\zeta_0| + K_{T,D}\int_0^t W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0))ds.$$

利用变量 $(\tilde{z},\omega,\varrho)$ 的分布之间的耦合是任意的这一事实，我们可以得出

$$\sup_{s\in[0,t]} W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0)) \le K_{T,D}(W_1(\pi_0(\cdot|z_0'),\pi_0(\cdot|z_0))+|z_0-z_0'|)+K_{T,D}\int_0^t W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0))ds.$$

此时，我们可以运用 Gronwall型论证 得到所需结果。

□

定理 3.11 (混沌传播)**.** 设 $T, D>0$，并假设假设 *1.4* 成立。设 $(\check{}\gamma_0, \check{}\sigma_0)$ 满足 $\check{}\gamma_0(Z^2\times[0,D]^2) = 1$ 且 $\check{}\sigma_0(\Theta\times[0,D]^2) = 1$，并假设 $F\check{}\sigma_0$ 和 $F\check{}\gamma_0$ 是概率测度。

对于 $N\in\mathbb{N}$，考虑与满足$\check{}\gamma_0^N(Z^2\times[0,D]^2) = 1$ 且 $\check{}\sigma_0^N(\Theta\times[0,D]^2) = 1$ 的序列 $\{(\check{}\gamma_0^N, \check{}\sigma_0^N)\}_{N\in\mathbb{N}}$ 相对应的系统 (3.1)，其中上述条件对所有足够大的 N 均成立；并假设 $F\check{}\sigma_0^N$ 和 $F\check{}\gamma_0^N$ 是概率测度。我们还假设 Z_0^i 属于测度 $\pi_{0,z}$ 的支撑集。

进一步假设，当 $N\to\infty$ 时，我们有

$$\inf_{\upsilon_z\in\Gamma_{Opt}(\pi_{0,z}^N,\pi_{0,z})}\int W_1(\check{\gamma}_0^N(\cdot|z_0'),\check{\gamma}_0(\cdot|z_0))d\upsilon_z(z_0',z_0)\to 0,\ \textit{and}\ W_1(\check{\sigma}_0^N,\check{\sigma}_0)\to 0. \tag{3.12}$$

Then

$$W_{T,1}(\check{\gamma}^N,\check{\gamma})\to 0,\qquad W_{T,1}(\check{\sigma}^N,\check{\sigma})\to 0,\quad \tilde{W}_{T,1}(\pi^N,\pi)\to 0,$$

其中 $(\check{\gamma},\check{\sigma})$ 是初始条件取自$(\check{\gamma}_0,\check{\sigma}_0)$的平均场系统(3.10) 的概率律。

证明。 根据假设，不失一般性地，我们可以假设，对于每个 N以及每个 $i=1,\dots,N$，权重 ω_0^i,α_0^i属于 $[0,D]$。由格朗沃尔不等式和假设 (1.4)，我们可以得出权重 ω_t^i,α_t^i 属于 $[0,De^{2M\kappa t}]$。因此 $(\check{\gamma}^N,\check{\sigma}^N)\in\mathcal{A}_{T,D}$。

下文中，我们用 ζ^i 表示系统 (3.1) 中第 i 个粒子的所有变量的路径，并用 $\zeta^{mf,i}$ 表示 (3.4) 中对应的粒子；我们回顾一下，这些粒子假设初始化于同一位置。我们考虑

$$\check{\gamma}_t^{N,mf}:=\frac{1}{N}\sum_{i=1}^N\delta_{(Z_t^{mf,i},\tilde{Z}_t^{mf,i}),\omega_t^{mf,i},\varrho_t^{mf,i}}\quad\text{and}\quad \check{\sigma}_t^{N,mf}:=\frac{1}{N}\sum_{i=1}^N\delta_{\vartheta_t^{mf,i},\alpha_t^{mf,i},\beta_t^{mf,i}},$$

也就是说，粒子平均场系统的经验测度。

由三角不等式可得

(3.13)

$$W_{t,1}(\check{\gamma},\check{\gamma}^N)+W_{t,1}(\check{\sigma},\check{\sigma}^N)\le\{W_{t,1}(\check{\gamma}^{N,mf},\check{\gamma}^N)+W_{t,1}(\check{\sigma}^{N,mf},\check{\sigma}^N)\}+\{W_{t,1}(\check{\gamma},\check{\gamma}^{N,mf})+W_{t,1}(\check{\sigma},\check{\sigma}^{N,mf})\}$$

现在我们声称，对于项 $\tilde{W}_{t,1}(\pi,\pi^N)$，类似的不等式也成立。具体而言，我们证明

$$\tilde{W}_{t,1}(\pi,\pi^N)\le\tilde{W}_{t,1}(\pi,\pi^{N,mf})+\tilde{W}_{t,1}(\pi^{N,mf},\pi^N).\tag{3.14}$$

注意到，所有动力学的 z 坐标都保持不变，并且 ζ^i 与 $\zeta^{mf,i}$ 的初始化相同。由此可得 π_s^N，$_z=\pi_{0,z}^N=\pi_{0,z}^{N,mf}=\pi_{s,z}^{N,mf}$，因此 $\Gamma_{\mathrm{Opt}}(\pi_{s,z}^{N,mf},\pi_{s,z}^N)$ 仅由恒等耦合构成。令 $\upsilon\in\Gamma_{\mathrm{Opt}}(\pi_{s,z},\pi_{s,z}^N)=\Gamma_{\mathrm{Opt}}(\pi_{0,z},\pi_{0,z}^N)$。由 W_1 的三角不等式可推得

$$\begin{aligned}\int W_1(\pi_s(\cdot|z),\pi_s^N(\cdot|z'))d\upsilon(z,z')&\le\int(W_1(\pi_s(\cdot|z),\pi_s^{N,mf}(\cdot|z'))+W_1(\pi_s^{N,mf}(\cdot|z'),\pi_s^N(\cdot|z')))d\upsilon(z,z')\\&\le\int W_1(\pi_s(\cdot|z),\pi_s^{N,mf}(\cdot|z'))d\upsilon(z,z')+\int W_1(\pi_s^{N,mf}(\cdot|z'),\pi_s^N(\cdot|z'))d\pi_{s,z}^N(z')\\&\le\int W_1(\pi_s(\cdot|z),\pi_s^{N,mf}(\cdot|z'))d\upsilon(z,z')+\tilde{W}_{t,1}(\pi^{N,mf},\pi^N),\end{aligned}$$

对每个 $s\in[0,t]$ 均成立。对两边关于 υ 取下确界，再关于 $s\in[0,t]$ 取上确界，即得 (3.14)。

我们再次利用粒子 ζ^i 和 $\zeta^{mf,i}$ 具有相同初始化这一事实，按照推论 3.7 的证明进行推导，从而得出对每个 $t\in[0,T]$

$$W_{t,1}(\check{\gamma}^{N,mf},\check{\gamma}^N)+W_{t,1}(\check{\sigma}^{N,mf},\check{\sigma}^N)\le K_{T,D}\int_0^t\{W_{s,1}(\check{\gamma},\check{\gamma}^N)+W_{s,1}(\check{\sigma},\check{\sigma}^N)+\tilde{W}_{s,1}(\pi,\pi^N)\}ds,$$

以及

$$\tilde{W}_{t,1}(\pi^{N,mf},\pi^N)\le K_{T,D}\int_0^t\{W_{s,1}(\check{\gamma},\check{\gamma}^N)+W_{s,1}(\check{\sigma},\check{\sigma}^N)+\tilde{W}_{s,1}(\pi,\pi^N)\}ds.$$

我们可以 现在将前面的两个不等式与（3.13）和（3.14）结合起来，可得

$$\begin{aligned}W_{t,1}(\check{\gamma},\check{\gamma}^N)+W_{t,1}(\check{\sigma},\check{\sigma}^N)+\tilde{W}_{t,1}(\pi,\pi^N)\le&\\&\{W_{t,1}(\check{\gamma}^{N,mf},\check{\gamma})+W_{t,1}(\check{\sigma}^{N,mf},\check{\sigma})+\tilde{W}_{t,1}(\pi^{N,mf},\pi)\}\\&+K_{T,D}\int_0^t\{W_{s,1}(\check{\gamma},\check{\gamma}^N)+W_{s,1}(\check{\sigma},\check{\sigma}^N)+\tilde{W}_{s,1}(\pi,\pi^N)\}ds.\end{aligned}$$

结合格朗沃尔不等式，上述结果意味着

$$W_{t,1}(\check{\gamma},\check{\gamma}^N)+W^1_{t,1}(\check{\sigma},\check{\sigma}^N)+\tilde{W}_{t,1}(\pi,\pi^N)\le e^{tK_{T,D}}\{W_{t,1}(\check{\gamma},\check{\gamma}^{N,mf})+W_{t,1}(\check{\sigma},\check{\sigma}^{N,mf})+\tilde{W}_{t,1}(\pi,\pi^{N,mf})\}.$$

为完成证明，我们必须证明当 $N\to\infty$ 时，上述表达式右侧趋于零。为此，我们比较 $\zeta^{mf,i}_{\mathsf{Z}}$ $\mathsf{Z}:=(Z^{mf,i},\tilde{Z}^{mf,i},\omega^{mf,i},\varrho^{mf,i})$ 和 ζ^{mf}_{Z} $\mathsf{Z}:=(Z^{mf},\tilde{Z}^{mf},\omega^{mf},\varrho^{mf})$ 的演化，然后分别比较 $\zeta^{mf,i}_{\Theta}$ $:=(\vartheta^{mf,i},\alpha^{mf,i},\beta^{mf,i})$ 和 $\zeta^{mf}_{\Theta}:=(\vartheta^{mf},\alpha^{mf},\beta^{mf})$ 的演化。对于第一对演化，我们按照引理 3.6 的证明进行推导，得到

$$\sup_{0\le s\le t}|\zeta^{mf,i}_{\mathcal{Z},s}-\zeta^{mf}_{\mathcal{Z},s}|\le K_{T,D}|\zeta^{mf,i}_{\mathcal{Z},0}-\zeta^{mf}_{\mathcal{Z},0}|+K_{T,D}\int_0^t W_1(\pi_s(\cdot|Z^{mf}_0),\pi_s(\cdot|Z^{mf,i}_0))ds.$$

然后我们可以利用引理 3.10 得到

$$\sup_{0\le s\le t}|\zeta^{mf,i}_{\mathcal{Z},s}-\zeta^{mf}_{\mathcal{Z},s}|\le K_{T,D}|\zeta^{mf,i}_{\mathcal{Z},0}-\zeta^{mf}_{\mathcal{Z},0}|+K_{T,D}W_1(\pi_0(\cdot|Z^{mf}_0),\pi_0(\cdot|Z^{mf,i}_0)).$$

结合上述逐路径估计，以及初始化时选择耦合的自由度，我们可以得出

$$\begin{aligned}W_{T,1}(\check{\gamma},\check{\gamma}^{N,mf}),W_{T,1}(\pi,\pi^{N,mf})&\le K_{T,D}W_1(\pi_{0,z},\pi^N_{0,z})\\&\quad+K_{T,D}\inf_{\upsilon_z\in\Gamma_{\mathrm{Opt}}(\pi^N_{0,z},\pi_{0,z})}\int W_1(\check{\gamma}^N_0(\cdot|z_0'),\check{\gamma}_0(\cdot|z_0))d\upsilon_z(z_0',z_0).\end{aligned}$$

根据假设 (3.12)、注 A.4 和引理 3.4，可知上述表达式右侧当 $N\to\infty$ 时趋于零。对于演化对 $\zeta^{mf,i}_{\Theta}$ 和 ζ^{mf}_{Θ}，我们按照引理 3.6 的证明进行推导，这次注意到我们可以写成

$$\sup_{0\le s\le t}|\zeta^{mf,i}_{\Theta,s}-\zeta^{mf}_{\Theta,s}|\le K_{T,D}|\zeta^{mf,i}_{\Theta,0}-\zeta^{mf}_{\Theta,0}|.$$

结合上述逐路径估计以及为初始化选择耦合的自由度，我们可以得出

$$W_{T,1}(\check{\sigma},\check{\sigma}^{N,mf})\le K_{T,D}W_1(\check{\sigma}_0,\check{\sigma}^N_0)\to 0.$$

□

3.3. **定理 3.11 的推论。** 本节将证明由定理 3.11 推出的若干重要结果。第一个结果是定理 1.8。

定理 *1.8* 的证明。 令 (γ_0,σ_0) 满足 $\mathcal{F}\gamma_0=\pi_0$ 和 $\mathcal{F}\sigma_0=\overline{\nu_0}$，且满足 (1.15)。我们引入测度 $\check{}\gamma_0:=\gamma_0\otimes\delta_1(d\varrho)$ 和 $\check{}\sigma_0:=\sigma_0\otimes\delta_1(d\beta)$。也就是说，$\check{}\gamma_0$ 是 γ_0 与在 ϱ 坐标取值为 1 处的狄拉克 δ 测度的乘积；$\check{}\sigma_0$ 的定义类似。同样，我们定义 $\check{}\gamma^N_0:=\gamma^N_0\otimes\delta_1$ 和 $\check{}\sigma^N_0:=\sigma^N_0\otimes\delta_1$。显然，根据这些定义，(3.12) 成立，因此可以应用定理 3.11 得出

$$W_{T,1}(\check{\gamma}^N,\check{\gamma})+W_{T,1}(\check{\sigma}^N,\check{\sigma})\to 0,$$

其中，$(\check{}\gamma^N,\check{}\sigma^N)$ 是由粒子系统 (3.1) 在定理陈述中所述的初始条件以及对所有 i 均满足 $\beta^i_0=\varrho^i_0=1$ 的条件下，在路径空间上诱导的测度；而 $(\check{}\gamma,\check{}\sigma)$ 则是 (3.10) 中初始条件为 $(\check{}\gamma_0,\check{}\sigma_0)=(\check{}\gamma_0,\check{}\sigma_0)$ 的概率律。

利用引理 3.4，我们得出，对于每个 $t\in[0,T]$

$$W_1(\pi^N_t,\pi_t)=W_1(\mathcal{F}\gamma^N_t,\mathcal{F}\gamma_t)\le K_{T,D}W_1(\gamma^N_t,\sigma^N_t)\le K_{T,D}W_1(\check{\gamma}^N_t,\check{\gamma}_t)\le K_{T,D}W_{T,1}(\check{\gamma}^N,\check{\gamma}).$$

同理，

$$W_1(\nu^N_t,\nu_t)\le K_{T,D}W_{T,1}(\check{\sigma}^N,\check{\sigma}).$$

对上面两个表达式之和中的所有 $t\in[0,T]$ 取上确界，我们得到

$$\sup_{t\in[0,T]}\{W_1(\pi^N_t,\pi_t)+W_1(\nu^N_t,\nu_t)\}\le K_{T,D}(W_{T,1}(\check{\gamma}^N,\check{\gamma})+W_{T,1}(\check{\sigma}^N,\check{\sigma})),$$

由此即可得到所需结果。

□

推论 3.16 和下文的注 3.17 是我们在本节讨论的定理 3.11 的另外两个重要推论，我们将在第 4 节中用到它们。在那里，我们考虑一种与 (1.10) 密切相关的演化$\{(\hat{\nu}_t, \hat{\pi}_t)\}_t$，其定义如下:

$$\partial_t \hat{\nu}_t = \eta_t \mathrm{div}(\hat{\nu}_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \theta)), \quad \partial_t \hat{\pi}_t = -\eta_t \mathrm{div}(\hat{\pi}_t(0, \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t, \nu_t; (z, \tilde{z})))), \tag{3.15}$$

其初始条件$\hat{\nu}_0, \hat{\pi}_0$分别关于 ν_0 和 π_0绝对连续。正是在这一阶段，我们使用了 (3.1) 中的额外坐标 β, ϱ。事实上，引入这些变量是为了适应$\hat{\nu}_0$与 ν_0 之间以及$\hat{\pi}_0$与 π_0之间的测度变换。我们将能够使用通用定理 3.11，证明系统 (3.15) 的粒子近似的一致性。

我们先给出一个预备结果。

命题 3.12. 设 $\nu_t^N = \frac{1}{N}\sum_{i=1}^n \alpha_i(t)\delta_{\vartheta_i(t)}$ 和 $\pi_t^N = \frac{1}{N}\sum_{i=1}^N \omega_i(t)\delta_{(Z_i(t)\ \tilde{Z}_i(t))}$ 如 *(3.2)* 所示。设 $\beta_1, \ldots, \beta_N$ 和 $\varrho_1, \ldots, \varrho_N$ 是两组非负标量，满足

$$\frac{1}{N}\sum_{i=1}^N \beta_i \alpha_i(0) = 1, \quad \frac{1}{N}\sum_{i=1}^N \varrho_i \omega_i(0) = 1.$$

令$\hat{\nu}_t^N$和$\hat{\pi}_t^N$为按如下方式定义的概率测度

$$\hat{\nu}_t^N := \frac{1}{N}\sum_{i=1}^n \beta_i \alpha_i(0)\delta_{\vartheta_i(t)}, \quad \pi_t^N := \frac{1}{N}\sum_{i=1}^N \varrho_i \omega_i(0)\delta_{(Z_i(t), \tilde{Z}_i(t))} \quad t \geq 0.$$

Then

$$\partial_t \hat{\nu}_t^N = \eta_t \mathrm{div}(\hat{\nu}_t^N \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta)), \quad \partial_t \hat{\pi}_t^N = -\eta_t \mathrm{div}(\hat{\pi}_t^N(0, \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^N, \nu_t^N; (z, \tilde{z})))) \tag{3.16}$$

在弱意义下。

证明。令 $\phi(\theta)$为任意测试函数。由 (3.1) 可知

$$\begin{aligned}
\frac{d}{dt}\int \phi(\theta) d\hat{\nu}_t^N(\theta) &= \frac{1}{N}\sum_{i=1}^N \beta_i \alpha_i(0) \frac{d}{dt}\phi(\vartheta_i(t)) = \frac{1}{N}\sum_{i=1}^N \beta_i \alpha_i(0) \nabla\phi(\vartheta_i(t)) \cdot \dot{\vartheta}_i(t) \\
&= -\frac{\eta_t}{N}\sum_{i=1}^N \beta_i \alpha_i(0) \nabla\phi(\vartheta_i(t)) \cdot \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \vartheta_i(t)) \\
&= \eta_t \int \nabla\phi(\theta) \cdot \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta) d\hat{\nu}_t^N(\theta).
\end{aligned}$$

这表明，$\hat{\nu}^N$ 在弱意义下满足方程 (3.16)。$\hat{\pi}^N$ 的方程可类似地推得。□

现在我们将探讨 (3.16) 与 (3.15) 之间的联系。首先，我们引入一些额外的数学工具，以帮助实现这一目标。

令$\check{\mathcal{F}}: \mathcal{P}(\mathcal{Z}^2 \times \mathbb{R}_+^2) \to \mathcal{M}_+(\mathcal{Z}^2)$ 为通过以下恒等式定义的映射

$$\int \phi(\theta) d(\check{\mathcal{F}}\check{\sigma})(\theta) = \int \alpha\beta\phi(\theta) d\check{\sigma}(\theta, \alpha, \beta),$$

对所有测试函数 ϕ。类似地，将 $\check{F}$定义为映射 $\check{\mathcal{F}}: \mathcal{P}(\Theta \times \mathbb{R}_+^2) \to \mathcal{M}_+(\Theta)$，将上述内容中所有出现的 $\check{\sigma}, \theta, \alpha, \beta$ 替换为 $\check{\gamma}, (z, \tilde{z})$ ω, ϱ。注意，只要 $\int \alpha\beta d\check{\sigma}(\theta, \alpha, \beta) = 1$，$\check{\mathcal{F}}\sigma$ 就是一个概率测度；当 $\check{\mathcal{F}}$作用于 $\check{\gamma}$时，也有类似结论。

现在引入映射 $\mathcal{G}: \mathcal{C}([0,T], \mathcal{Z}^2 \times \mathbb{R}_+^2) \to \mathcal{C}([0,T], \mathcal{Z}^2 \times \mathbb{R}_+^2)$，定义如下：

$$\mathcal{G} : \{(z_t, \tilde{z}_t, \omega_t, \varrho_t)\}_{t\in[0,T]} \mapsto \{(z_t, \tilde{z}_t, \omega_0, \varrho_0)\}_{t\in[0,T]}. \tag{3.17}$$

也就是说，$\mathcal{G}$ 是一个映射，它会冻结给定路径的坐标 ω, ϱ，使其等于各自的初始条件。自然地，$\mathcal{G}$ 通过推前诱导出一个从 $\mathcal{P}(\mathcal{C}([0,T], \mathcal{Z}^2 \times \mathbb{R}_+^2))$

到其自身的映射；我们稍微滥用一下记号，也用 $\mathcal{G}$ 表示这个诱导映射。此外，我们还将 $\mathcal{G}$ 视为映射 $\mathcal{G}\colon \mathcal{C}([0,T],\Theta\times\mathbb{R}_+^2) \to \mathcal{C}([0,T],\Theta\times\mathbb{R}_+^2)$，它会冻结给定路径的坐标 α,β，使其等于各自的初始条件；我们也用 $\mathcal{G}$ 表示由推前诱导出的从 $\mathcal{P}(\mathcal{C}([0,T],\Theta\times\mathbb{R}_+^2))$ 到其自身的映射。在不同情况下使用 $\mathcal{G}$ 的哪种解释，应可由上下文清楚判断。

注 3.13. 注意到，命题 *3.12* 中的$\hat{\pi}_t^N$和$\hat{\nu}_t^N$分别可写为$\check{F}((G\check{\gamma}^N)_t)$和$\check{F}((G\check{\sigma}^N)_t)$。

引理 *3.14.* 设 $(\check{\gamma},\check{\sigma})$ 是从一对初值 $(\check{\gamma}_0,\check{\sigma}_0)$ 出发的过程 *(3.10)* 的概率律。那么，$\{\hat{\nu}_t := \check{F}((G\check{\sigma})_t)\}_{t\in[0,T]}$ 和 $\{\hat{\pi}_t := \check{F}((G\check{\gamma})_t)\}_{t\in[0,T]}$ 满足偏微分方程 *(3.15)*，其中 $\pi_t = F(\gamma_t)$ 和 $\nu_t = F(\sigma_t)$。

证明。 考虑平均场常微分方程 (3.10)。对于每个光滑测试函数 ϕ，我们有

$$\int \phi(\theta)d\hat{\nu}_t(\theta) = \int \alpha\beta\phi(\theta)d(\mathcal{G}\sigma)_t(\theta,\alpha,\beta) = \mathbb{E}[\alpha_0\beta_0\phi(\vartheta_t)].$$

特别地，

$$\begin{aligned}\frac{d}{dt}\int \phi(\theta)d\hat{\nu}_t(\theta) = \frac{d}{dt}\mathbb{E}[\alpha_0\beta_0\phi(\vartheta_t)] &= -\mathbb{E}[\eta_t\alpha_0\beta_0\nabla\phi(\vartheta_t)\cdot\nabla_\theta\mathcal{U}_\nu(\pi_t,\nu_t;\vartheta_t)]\\ &= -\eta_t\int \nabla\phi(\theta)\cdot\nabla_\theta\mathcal{U}_\nu(\pi_t,\nu_t;\theta)d\hat{\nu}_t(\theta).\end{aligned}$$

这证 明了 $\hat{\nu}$ 满足所需的方程。$\hat{\pi}$ 的方程可类似地得到。

□

注 3.15。 请注意，如果$\hat{\nu}_0$ 和$\hat{\pi}_0$ 是概率测度，那么$\hat{\nu}_t$ 和$\hat{\pi}_t$ 也是概率测度。

接下来，我们将使用定理 3.11 证明，在对初始化作出适当假设的条件下，可以通过合适的粒子近似恢复式 (3.16) 中的系统。

推论 3.16。 令 ν_0 和 π_0 任意取值，并令$\hat{\nu}_0$ 和$\hat{\pi}_0$ 为满足$\hat{\nu}_0 \ll \nu_0$、$\hat{\pi}_0 \ll \pi_0$的概率测度，其中 $\frac{d\hat{\nu}_0}{d\nu_0} \in L^\infty(\nu_0)$ 且 $\frac{d\hat{\pi}_0}{d\pi_0} \in L^\infty(\pi_0)$。令$\check{\gamma}_0$ 和$\check{\sigma}_0$ 如定理 *3.11* 中所述，并进一步假设它们满足 $F\check{\gamma}_0 = \pi_0$、$F\check{\sigma}_0 = \nu_0$，以及 $F\check{\gamma}_0 = \hat{\pi}_0$、$F\check{\sigma}_0 = \hat{\nu}_0$。

考虑定理 *3.11* 中的粒子系统近似，并额外假设$\hat{\nu}_0^N$、$\hat{\pi}_0^N$是概率测度。则

$$\sup_{t\in[0,T]}\{W_1(\hat{\nu}_t^N,\hat{\nu}_t) + W_1(\hat{\pi}_t^N,\hat{\pi}_t)\} \to 0, \quad \sup_{t\in[0,T]}\{W_1(\nu_t^N,\nu_t) + W_1(\pi_t^N,\pi_t)\} \to 0,$$

as $N\to\infty$。以上使用的记号与引理 *3.14* 和注释 *3.1* 中相同 *3.*

证明。 首先请注意，条件 $\frac{d\hat{\nu}_0}{d\nu_0} \in L^\infty(\nu_0)$ 和 $\frac{d\hat{\pi}_0}{d\pi_0} \in L^\infty(\pi_0)$ 用于确保我们确实可以构造具有有界支撑的 $\check{\sigma}_0$ 和 $\check{\gamma}_0$；见下文注释 3.17 的第一部分。

不难验证$\mathcal{G}$ 是一个利普希茨映射，即

$$W_{T,1}(\mathcal{G}\check{\sigma},\mathcal{G}\check{\sigma}') \le 2W_{T,1}(\check{\sigma},\check{\sigma}'), \quad W_{T,1}(\mathcal{G}\check{\gamma},\mathcal{G}\check{\gamma}') \le 2W_{T,1}(\check{\gamma},\check{\gamma}').$$

此外，我们可以找到一个常数 $C_{T,D}$，使得对每个 $t\in[0,T]$

$$W_1(\check{\mathcal{F}}((\mathcal{G}\check{\sigma})_t),\check{\mathcal{F}}((\mathcal{G}\check{\sigma}^N)_t)) \le C_{T,D}W_1((\mathcal{G}\check{\sigma})_t),(\mathcal{G}\check{\sigma}^N)_t) \le C_{T,D}W_{T,1}(\mathcal{G}\check{\sigma},\mathcal{G}\check{\sigma}^N),$$

其中，第一个不等式可由与引理 3.4 中非常相似的方法得出。同理，

$$W_1(\check{\mathcal{F}}((\mathcal{G}\check{\gamma})_t),\check{\mathcal{F}}((\mathcal{G}\check{\gamma}^N)_t)) \le C_{T,D}W_{T,1}(\mathcal{G}\check{\gamma},\mathcal{G}\check{\gamma}^N).$$

我们可以将上述不等式与定理 3.11 结合，从而得到

$$W_{T,1}(\check{\gamma}^N,\check{\gamma}) + W_{T,1}(\check{\sigma}^N,\check{\sigma}) \to 0,$$

以推导出所需的收敛。

□

注 3.17（构造初始条件）。任取 π_0 和 ν_0，并令 $\rho_\nu = \frac{d\hat{\nu}_0}{d\nu_0}$ 和 $\rho_\pi = \frac{d\hat{\pi}_0}{d\pi_0}$ 满足假设 —— $\rho_\nu \in L^\infty(\nu_0)$ 和 $\rho_\pi \in L^\infty(\pi_0)$；我们进一步假设 $\hat{\pi}_{0,z} = \pi_{0,z}$。后一假设意味着，对于支撑集 $\pi_{0,z}$ 中的所有 z，均有 $\int \rho_\pi(z,\tilde{z})d\pi_0(\tilde{z}|z) = 1$。

令 $\check{\gamma}_0$ 和 $\check{\sigma}_0$ 为测度 $\check{\gamma}_0 := h_{\gamma\sharp}\pi_0$、$\check{\sigma}_0 := h_{\sigma\sharp}\overline{\nu}_0$、$h_\gamma : (z,\tilde{z}) \mapsto (z,\tilde{z},1,\rho_\pi(z,\tilde{z}))$、$h_\sigma : \theta \mapsto (\theta,1,\rho_\nu(\theta))$。注意，$F\check{\gamma}_0 = \pi_0$ 和 $\hat{F}\check{\gamma}_0 = \hat{\pi}_0$，而 $F\check{\sigma}_0 = \overline{\nu}_0$ 和 $\hat{F}\check{\sigma}_0 = \hat{\nu}_0$。

我们采用与注 *1.9* 中相同的对象和记号，并引入额外变量 $\beta_{ij} = \rho_\nu(\vartheta_{ij})$ 和 $\varrho_{ij} = \rho_\pi(Z_{ij}, Z_{ij})$；注意，对 ρ_ν 和 ρ_π 施加一致有界性条件，是为了确保权重 ϱ_{ij} 和 β_{ij} 一致有界。考虑测度

$$\check{\gamma}_0^{n,m} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{(Z_{ij},\tilde{Z}_{ij},\omega_{ij},\varrho_{ij})}, \quad \check{\sigma}_0^{n,m} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{(\vartheta_{ij},\alpha_{ij},\beta_{ij})}.$$

根据引理 *A.3*，我们可以找到一个序列 $\{(n_k,m_k)\}_{k\in\mathbb{N}}$，使得几乎必然，所诱导的成对序列 $\check{\gamma}_0^{N_k} := \check{\gamma}^{n_k,m_k}_0$，$\check{\sigma}_0^{N_k} := \check{\sigma}^{n_k,m_k}_0$ 满足条件 (3.12)。此外，根据大数定律和附录 *A* 中的引理 *A.5*，可以假定该子序列满足

$$\lim_{k\to\infty}\frac{1}{n_k}\sum_{i=1}^{n_k}\left|\frac{1}{\frac{1}{m_k}\sum_{j=1}^{m_k}\rho_\pi(Z_{ij},\tilde{Z}_{ij})}-1\right| = 0, \quad \lim_{k\to\infty}\frac{1}{n_k}\sum_{i=1}^{n_k}\left|\frac{1}{\frac{1}{m_k}\sum_{j=1}^{m_k}\rho_\nu(\vartheta_{ij})}-1\right| = 0. \tag{3.18}$$

我们对权重 ϱ_{ij} 和 β_{ij} 稍作修改，将其归一化，使得对所有 i，$\frac{1}{m_k}\sum_j \varrho_{ij} = 1$，并且 $\frac{1}{n_k m_k}\sum_{ij}\beta_{ij} = 1$。根据 (3.18)，我们可以直接证明，权重归一化后，条件(3.12) 仍然成立。由此得到的测度 $\hat{\pi}_0^{N_k} = \sum_j \varrho_{ij}\delta_{(Z_{ij},\tilde{Z}_{ij})}$ 和 $\hat{\nu}_0^{N_k} = \sum_{ij}\beta_{ij}\delta_{\vartheta_{ij}}$ 可以看出分别在 *Wasserstein* 意义下收敛到 $\hat{\nu}_0$ 和 $\hat{\pi}_0$；而测度 $\pi_0^{N_k} = \frac{1}{n_k m_k}\sum_{ij}\delta_{(Z_{ij},\tilde{Z}_{ij})}$ 和 $\nu_0^{N_k} = \frac{1}{n_k m_k}\sum_{ij}\delta_{\vartheta_{ij}}$ 则分别收敛到 π_0 和 ν_0。

此外，再次应用大数定律可知

$$\mathcal{H}(\hat{\nu}_0^{N_k}||\nu_0^{N_k}) = \left(\frac{1}{n_k m_k}\sum_{ij}\rho_\nu(\vartheta_{ij})\right)^{-1}\frac{1}{n_k m_k}\sum_{ij}\log(\rho_\nu(\vartheta_{ij}))\rho_\nu(\vartheta_{ij}) - \log\left(\frac{1}{n_k m_k}\sum_{ij}\rho_\nu(\vartheta_{ij})\right)$$

当 $k\to\infty$ 趋于 $\int_\Theta$ 时，收敛至 $\int_\Theta \log(\rho_\nu(\theta))\rho_\nu(\theta)d\nu_0(\theta)$，这恰好就是 $H(\hat{\nu}_0||\nu_0)$。同样，我们可以看到 $H(\hat{\pi}_0^{N_k}||\pi_0^{N_k}) \to H(\hat{\pi}_0||\pi_0)$，当 $k\to\infty$ 时。

上述相对熵的收敛将在下一节中用到。

4. 平均场方程的长期行为及（1.5）的近似纳什均衡

本节研究从任意测度 (π_0,ν_0) 出发的方程组（1.10）的长期行为。我们的目标是研究系统（1.10）为问题（1.5）生成近似纳什均衡的能力。我们将讨论分为两种不同情形：1）相当一般的非凸非凹情形，以及 2）非凸凹情形。请注意，我们所说的非凸／非凹，实际上是指相对于驱动动力学的某种最优传输几何而言，不是测地凸／凹；与此同时，我们确实假设其在线性插值意义下具有凸性／凹性，如假设1.13所述。

为开始我们的分析，我们首先讨论系统 (1.10) 与引理 3.14 中所述的一个相关带帽过程之间的关系。类似系统的研究已在诸如 [13] 的工作中得到探讨。然而，本文提出一种替代方法，使我们能够充分论证推导过程；更多详情请参见下文的注 4.4。我们的方法利用了第 3 节中研究的更大结构，具体而言，即注 3.17 的内容。事实上，对于任意初始条件，我们将利用注 3.17 中的粒子近似来理解 $\hat{\nu}$ 与 ν 以及 $\hat{\pi}$ 与 π 之间相对熵的时间演化。作为第一步，我们

研究测度 $(\nu_t^{N_k}, \pi_t^{N_k})$ 和 $(\hat{\nu}_t^{N_k}, \hat{\pi}_t^{N_k})$ 在按照注 3.17 初始化为经验测度时，相对熵的时间演化。

命题 4.1。 设 π_0 和 ν_0 任意，且 $\hat{\pi}_0$ 和 $\hat{\nu}_0$ 如注 *3.17* 所定义。对于固定的 $k \in \mathbb{N}$，设 $\nu_t^{N_k}$、$\hat{\nu}_t^{N_k}$、$\pi_t^{N_k}$和$\hat{\pi}_t^{N_k}$如命题 *3.12* 所定义，并按注 *3.17* 的方式初始化。则

$$\frac{d}{dt}\mathcal{H}(\hat{\nu}_t^{N_k} \| \nu_t^{N_k}) = \kappa \int_\Theta \mathcal{U}_\nu(\pi_t^{N_k}, \nu_t^{N_k}; \theta) d(\hat{\nu}_t^{N_k} - \nu_t^{N_k}) \tag{4.1}$$

and

$$\frac{d}{dt}\mathcal{H}(\hat{\pi}_t^{N_k} \| \pi_t^{N_k}) = -\kappa \int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_t^{N_k}, \nu_t^{N_k}; z, \tilde{z}) d(\hat{\pi}_t^{N_k} - \pi_t^{N_k}).$$

证明。 注意到

$$\begin{aligned}
\frac{d}{dt}\mathcal{H}(\hat{\nu}_t^{N_k} \| \nu_t^{N_k}) &= \frac{d}{dt}\left(\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \log\left(\frac{\beta_{ij}(0)\alpha_{ij}(0)}{\alpha_{ij}(t)} \right) \beta_{ij}(0)\alpha_{ij}(0) \right) \\
&= -\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \frac{d}{dt} \log(\alpha_{ij}(t))\beta_{ij}(0)\alpha_{ij}(0) \\
&= \frac{\kappa}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} (\mathcal{U}_\nu(\pi_t^{N_k}, \nu_t^{N_k}; \vartheta_{ij}(t)) - \overline{\mathcal{U}}_\nu)\beta_{ij}(0)\alpha_{ij}(0),
\end{aligned}$$

从第二行到第三行时，我们对 $\alpha_{ij}(t)$使用了方程 (3.1)。我们还使用了简记 $\mathcal{U}_\nu = \int_\Theta \mathcal{U}_\nu(\pi_t^{N_k}, \nu_t^{N_k}; \theta) d\nu_t^{N_k}(\theta)$。 恒等式 (4.1) 由此可得。

$\frac{d}{dt}\mathcal{H}(\hat{\pi}_t^N \| \pi_t^N)$的恒等式可通过类似的推理得到，但此处我们依赖于权重 ϱ_{ij} 的每一行都经过归一化这一事实：

$$\begin{aligned}
\frac{d}{dt}\mathcal{H}(\hat{\pi}_t^{N_k} \| \pi_t^{N_k}) &= \frac{d}{dt}\left(\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \log\left(\frac{\varrho_{ij}(0)\omega_{ij}(0)}{\omega_{ij}(t)} \right) \varrho_{ij}(0)\omega_{ij}(0) \right) \\
&= -\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \frac{d}{dt} \log(\omega_{ij}(t))\varrho_{ij}(0)\omega_{ij}(0) \\
&= -\frac{\kappa}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} (\mathcal{U}_\pi(\pi_t^{N_k}, \nu_t^{N_k}; Z_{ij}, \tilde{Z}_{ij}) - \overline{\mathcal{U}}_{\pi,i})\varrho_{ij}(0)\omega_{ij}(0).
\end{aligned}$$

以上我们使用了简记 $\mathcal{U}_{\pi,\,i} = \int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_t^{N_k},\ \nu_t^{N_k};\ Z_{ij},\ \tilde{z}) d\pi_t^{N_k}(\tilde{z}|Z_{ij})$；请注意，在我们的构造中， Z_{ij} 不依赖于 j。□

接下来，为了寻找一种类似于 (4.1) 但适用于一般初始条件的关系，我们在推论 3.16 的近似结果中加入一个要素。

命题 4.2. 任取 π_0 和 ν_0，并令 $\hat{\pi}_0$ 和 $\hat{\nu}_0$ 如注 *3.17* 中所述。令$(\hat{\nu},\hat{\pi})$ 为引理 *3.14* 中的动力学，其初始化方式如注 *3.17* 所述。对于任意 $k \in \mathbb{N}$，令 $\nu_t^{N_k}$、$\hat{\nu}_t^{N_k}$、$\pi_t^{N_k}$、$\hat{\pi}_t^{N_k}$ 如命题 *3.12* 中所述，其初始化方式如注 *3.17* 所述。

Then

$$\lim_{k\to\infty} \int \mathcal{U}_\nu(\pi_s^{N_k}, \nu_s^{N_k}; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k}) = \int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s - \nu_s) \tag{4.2}$$

以及

$$\lim_{k\to\infty} \int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_s^{N_k}, \nu_s^{N_k}; z, \tilde{z}) d(\hat{\pi}_s^{N_k} - \pi_s^{N_k}) = -\int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}) d(\hat{\pi}_s - \pi_s). \tag{4.3}$$

证明。 由假设 1.4 和推论 3.16 可得

$$\left|\int \mathcal{U}_\nu(\pi_s^{N_k},\nu_s^{N_k};\theta)d(\hat{\nu}_s^{N_k}-\nu_s^{N_k})-\int \mathcal{U}_\nu(\pi_s,\nu_s;\theta)d(\hat{\nu}_s^{N_k}-\nu_s^{N_k})\right| \le L(W_1(\nu_s,\nu_s^{N_k})+W_1(\pi_s,\pi_s^{N_k}))\to 0,$$

如 $k\to\infty$。另一方面，由于函数 $\mathcal{U}_\nu(\pi_s,\ \nu_s,\ \cdot)$ 关于 $|\theta|$ 连续且至多线性增长，并且根据推论 3.16，当 $k\to\infty$ 时，$W_1(\nu_s^{N_k},\ \nu_s)\to 0$, $W_1(\hat{}\nu_s^{N_k},\ \hat{}\nu_s)\to 0$，因此

$$\lim_{k\to\infty}\left|\int \mathcal{U}_\nu(\pi_s,\nu_s;\theta)d(\hat{\nu}_s^{N_k}-\nu_s^{N_k})-\int \mathcal{U}_\nu(\pi_s,\nu_s;\theta)d(\hat{\nu}_s-\nu_s)\right|=0.$$

方程 (4.2) 可直接得出。(4.3) 的推导类似。

□

命题 4.3。 任取 π_0 和 ν_0¯，并令 $\hat{}\pi_0$ 和 $\hat{}\nu_0$ 如注 *3.17* 所述。令($\hat{}\nu$,$\hat{}\pi$) 为引理 *3.14* 中的动力学，其初始值如注 *3.17* 所述。

则以下不等式成立：

$$\mathcal{H}(\hat{\nu}_t||\nu_t)-\mathcal{H}(\hat{\nu}_0||\overline{\nu}_0)\le\kappa\int_0^t\left(\int \mathcal{U}_\nu(\pi_s,\nu_s;\theta)d(\hat{\nu}_s-\nu_s)(\theta)\right)ds,\quad \forall t\ge 0, \tag{4.4}$$

and

$$\mathcal{H}(\hat{\pi}_t||\pi_t)-\mathcal{H}(\hat{\pi}_0||\overline{\pi}_0)\le-\kappa\int_0^t\left(\int \mathcal{U}_\pi(\pi_s,\nu_s;z,\tilde{z})d(\hat{\pi}_s-\pi_s)(z,\tilde{z})\right)ds,\quad \forall t\ge 0. \tag{4.5}$$

证明。 对于每个 $k\in\mathbb{N}$，考虑 $\nu_t^{N_k}$，$\hat{}\nu_t^{N_k}$，$\pi_t^{N_k}$，$\hat{}\pi_t^{N_k}$，它们如命题 3.12 所述，且初始值如注 3.17 所述。注意，根据推论 3.16，当 $k\to\infty$ 时，我们有 $W_1(\nu_s^{N_k},\ \nu_s)\to 0$, $W_1(\hat{}\nu_s^{N_k},\ \hat{}\nu_s)\to$0。

根据命题 (4.1)，我们有

$$\mathcal{H}(\hat{\nu}_t^{N_k}\|\nu_t^{N_k})=\mathcal{H}(\hat{\nu}_0^{N_k}\|\nu_0^{N_k})+\int_0^t\int_\Theta \mathcal{U}_\nu(\pi_s^{N_k},\nu_s^{N_k};\theta)d(\hat{\nu}_s^{N_k}-\nu_s^{N_k})ds.$$

现在我们可以利用相对熵关于弱收敛的联合下半连续性，得到：

$$\begin{aligned}\mathcal{H}(\hat{\nu}_t||\nu_t)\le\liminf_{k\to\infty}\mathcal{H}(\hat{\nu}_t^{N_k}||\nu_t^{N_k})&=\liminf_{k\to\infty}\kappa\int_0^t\left(\int \mathcal{U}_\nu(\pi_s^{N_k},\nu_s^{N_k};\theta)d(\hat{\nu}_s^{N_k}-\nu_s^{N_k})(\theta)\right)ds\\&\quad+\lim_{k\to\infty}\mathcal{H}(\hat{\nu}_0^{N_k}||\nu_0^{N_k}).\\&=\liminf_{k\to\infty}\kappa\int_0^t\left(\int \mathcal{U}_\nu(\pi_s^{N_k},\nu_s^{N_k};\theta)d(\hat{\nu}_s^{N_k}-\nu_s^{N_k})(\theta)\right)ds\\&\quad+\mathcal{H}(\hat{\nu}_0||\overline{\nu}_0).\end{aligned} \tag{4.6}$$

利用命题 (4.2) 以及注 3.17 中讨论的逼近性质，我们得到 (4.4)。不等式 (4.5) 的推导类似。□

注 4.4。 不同于 *[13],* 中所呈现的分析，我们在此采用了第 *3* 节中的平均场极限结果以及相对熵的下半连续性，从而充分证明了单侧恒等式 （4.4）和（4.5）。正如我们将在下文看到的，这些单侧恒等式足以满足我们的分析需要。沿用我们的方法，我们可以避开 *[13]*中为分析类似问题而考虑的策略。他们的策略依赖于这样一个假设：存在描述类似于此处所考虑的 ν 和$\hat{}\nu$ 之间测度变换之演化的某个偏微分方程的解，并且该解具有正则性。遗憾的是，一般而言，这样的偏微分方程甚至没有明确定义；当考虑以经验测度为初始测度的流时，这一点就变得很明显。尽管 *[13],* 承认了这一技术困难，却没有为其提供解决方案；参见 *[13]*中的第 *29* 页。

4.1. 非凸非凹情形。 有了命题（4.3），并按照 [13], 中的类似步骤，我们现在可以在假设 1.4 和 1.13 下推导出以下结果。

引理 4.5. 设 π, ν 是方程 (1.10) 从概率测度 π_0, ν_0出发的解，其中 $\pi_{0,z} = \mu$。设 π^*, ν^* 分别是 $Z \times Z$ 上任意的概率测度，且假设 $\pi^*_z = \mu$。令

$$\mathcal{Q}_\pi(\pi_0, \pi^*; \tau) := \inf_{\hat{\pi} \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z}),\, \hat{\pi}_z = \mu} \{\|\pi^* - \hat{\pi}\|^*_{BL} + \frac{1}{\tau}\mathcal{H}(\hat{\pi}||\pi_0)\},$$

其中 $k\cdot k^*_{BL}$表示有界利普希茨范数的对偶范数。另将 $Q_\nu(\nu_0, \nu^*; \tau)$定义为

$$\mathcal{Q}_\nu(\nu_0, \nu^*; \tau) := \inf_{\hat{\nu} \in \mathcal{P}(\Theta)} \{\|\nu^* - \hat{\nu}\|^*_{BL} + \frac{1}{\tau}\mathcal{H}(\hat{\nu}||\nu_0)\}.$$

假设假设 *1.4* 和 *1.13* 成立，则

$$\mathcal{U}(\pi^*, \bar{\nu}(t)) - \mathcal{U}(\bar{\pi}(t), \nu^*) \le B(\mathcal{Q}_\pi(\pi_0, \pi^*; \kappa Bt) + \mathcal{Q}_\nu(\nu_0, \nu^*; \kappa Bt)) + \frac{2B^2}{t}\int_0^t \int_0^s \eta_\tau d\tau ds,$$

其中 $B := M + L$ （ L 和 M的含义见假设 *1.4*）。上文中，$\pi_t := \frac{1}{t}\int_0^t \pi_s ds$ 且 $\nu_t := \frac{1}{t}\int_0^t \nu_s ds$。

证明. 取任意两个概率测度 $\hat{\pi}_0$ 和 $\hat{\nu}_0$ ，满足 $\hat{\pi}_0 \ll \pi_0, \hat{\nu}_0 \ll \nu_0$， $\hat{\pi}_{0,z} = \mu$， $\frac{d\hat{\nu}_0}{d\nu_0} \in L^\infty(\nu_0)$，且 $\frac{d\hat{\pi}_0}{d\pi_0} \in L^\infty(\pi_0)$。我们考虑命题 4.3 中的动力学 $(\hat{\pi}_t, \hat{\nu}_t)$ 和 (π_t, ν_t)。

i) 步骤 1：由$\mathcal{U}$ 关于第一坐标的凹性（相对于线性插值）可得

$$\begin{aligned}
\mathcal{U}(\pi^*, \nu_t) \le& \mathcal{U}(\pi_t, \nu_t) + \int \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}) d(\pi^* - \pi_t) \\
=& \mathcal{U}(\pi_t, \nu_t) + \int \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}) d(\pi^* - \hat{\pi}_t) \\
&+ \int \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}) d(\hat{\pi}_t - \pi_t).
\end{aligned}$$

使用有界利普希茨范数，由命题 4.3 可得

$$\begin{aligned}
\int_0^t \mathcal{U}(\pi^*, \nu_s) ds - \int_0^t \mathcal{U}(\pi_s, \nu_s) ds \le& \int_0^t (\|\mathcal{U}_\pi(\pi_s, \nu_s; \cdot)\|_{BL} \|\pi^* - \hat{\pi}_s\|^*_{BL}) ds \\
&+ \int_0^t \int \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}) d(\hat{\pi}_s - \pi_s) ds \\
\le& \int_0^t (\|\mathcal{U}_\pi(\pi_s, \nu_s; \cdot)\|_{BL} \|\pi^* - \hat{\pi}_s\|^*_{BL}) ds \\
&- \frac{1}{\kappa}(\mathcal{H}(\hat{\pi}_t||\pi_t) - \mathcal{H}(\hat{\pi}_0||\pi_0)).
\end{aligned} \tag{4.7}$$

通过类似的论证，利用$\mathcal{U}$ 关于第二坐标的凸性，可推出

$$\begin{aligned}
\int_0^t \mathcal{U}(\pi_s, \nu^*) ds - \int_0^t \mathcal{U}(\pi_s, \nu_s) ds \ge& -\int_0^t (|\mathcal{U}_\nu(\pi_s, \nu_s; \cdot)\|_{BL} \|\nu^* - \hat{\nu}_s\|^*_{BL}) ds \\
&+ \frac{1}{\kappa}(\mathcal{H}(\hat{\nu}_t||\nu_t) - \mathcal{H}(\hat{\nu}_0||\nu_0)).
\end{aligned} \tag{4.8}$$

再次利用$\mathcal{U}$的凹性和凸性，我们得到：

$$\mathcal{U}(\bar{\pi}_t, \nu^*) \ge \frac{1}{t}\int_0^t \mathcal{U}(\pi_s, \nu^*) ds, \qquad \mathcal{U}(\pi^*, \bar{\nu}_t) \le \frac{1}{t}\int_0^t \mathcal{U}(\pi^*, \nu_s) ds.$$

结合上述结果、(4.7)、(4.8)，以及$\mathcal{H}(\hat{}\nu_t||\nu_t)$,$\mathcal{H}(\hat{}\pi_t||\pi_t) \geq 0$ 这一事实，我们得出

$$\begin{aligned}\mathcal{U}(\pi^*,\bar{\nu}_t)-\mathcal{U}(\bar{\pi}_t,\nu^*) \leq & \frac{1}{t}\int_0^t (\|\mathcal{U}_\nu(\pi_s,\nu_s;\cdot)\|_{BL}\|\nu^*-\hat{\nu}_s\|_{BL}^* + \|\mathcal{U}_\pi(\pi_s,\nu_s;\cdot)\|_{BL}\|\pi^*-\hat{\pi}_s\|_{BL}^*)ds \\ & + \frac{1}{\kappa t}\left(\mathcal{H}(\hat{\nu}_0||\nu_0)+\mathcal{H}(\hat{\pi}_0||\pi_0)\right) \\ \leq & \frac{B}{t}\int_0^t (\|\nu^*-\hat{\nu}_s\|_{BL}^* + \|\pi^*-\hat{\pi}_s\|_{BL}^*)ds + \frac{1}{\kappa t}\left(\mathcal{H}(\hat{\nu}_0||\nu_0)+\mathcal{H}(\hat{\pi}_0||\pi_0)\right). \end{aligned} \tag{4.9}$$

ii) 步骤 2：注意到$\mathcal{U}_\pi$ 和$\mathcal{U}_\nu$ 的 BL 范数都不超过 $B = M + L$。要得出结论，只需指出

$$\begin{aligned}\frac{1}{t}\int_0^t \|\pi^*-\hat{\pi}_s\|_{BL}^* ds \leq & \|\pi^*-\hat{\pi}_0\|_{BL}^* + \frac{1}{t}\int_0^t \|\hat{\pi}_0-\hat{\pi}_s\|_{BL}^* ds \\ = & \|\pi^*-\hat{\pi}_0\|_{BL}^* + \frac{1}{t}\int_0^t \left\{\sup_{\|f\|_{BL}\leq 1; f\in C^1}\int f d(\hat{\pi}_s-\hat{\pi}_0)\right\} ds \\ \leq & \|\pi^*-\hat{\pi}_0\|_{BL}^* + \frac{B}{t}\int_0^t\int_0^s \eta_\tau d\tau ds,\end{aligned}$$

类似地，

$$\frac{1}{t}\int_0^t \|\nu^*-\hat{\nu}_s\|_{BL}^* ds \leq \|\nu^*-\hat{\nu}_0\|_{BL}^* + \frac{B}{t}\int_0^t\int_0^s \eta_\tau d\tau ds.$$

代入 (4.9)，可得

$$\begin{aligned}\mathcal{U}(\pi^*,\bar{\nu}_t)-\mathcal{U}(\bar{\pi}_t,\nu^*) \leq & B(\|\nu^*-\hat{\nu}_0\|_{BL}^* + \|\pi^*-\hat{\pi}_0\|_{BL}^*) \\ & + \frac{1}{\kappa t}\left(\mathcal{H}(\hat{\nu}_0||\nu_0)+\mathcal{H}(\hat{\pi}_0||\pi_0)\right) + \frac{2B^2}{t}\int_0^t\int_0^s \eta_\tau d\tau ds.\end{aligned} \tag{4.10}$$

回想一下，$\hat{}\pi_0$ 和 $\hat{}\nu_0$ 是任意的测度，其密度分别相对于 π_0 和ν_0 定义，且属于 L^∞。通过一个简单的稠密性论证，现在可以得出所需的估计。

□

下面的引理摘自 [13] ，而其中采用的论证又沿用了 [1中的论证。 0].

引理 4.6. 假设 *1.4* 和 *1.13* 成立。进一步假设存在$k > 0$ ，使得 $\frac{d\nu_0}{d\theta}(\theta) > k$，并假设 $|B_{\theta,\epsilon}\cap\Theta| \geq k'\epsilon^d$ 关于$\theta \in \Theta$一致成立，其中 $B_{\theta,\epsilon}$是以 θ 为中心、半径为 ϵ 的欧几里得球。那么，

$$\mathcal{Q}_\nu(\nu_0,\nu^*;\tau) \leq \frac{d}{\tau}\left\{1-\log\left(\frac{d}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k)-\log(k')\}.$$

证明。我们为$\mathcal{Q}_\nu$定义中的最小值求得一个界。对于固定的 $\epsilon > 0$ ，我们引入由下式给出的概率测度 ν^ϵ

$$\nu^\epsilon(A) := \int_\Theta \frac{|\mathcal{B}_{\theta,\epsilon}\cap A|}{|\mathcal{B}_{\theta,\epsilon}\cap\Theta|} d\nu^*(\theta).$$

现在我们计算 $W_1(\nu^*, \nu^\epsilon)$。考虑以下耦合：

$$\Upsilon(A,A') := \int_A \frac{|\mathcal{B}_{\theta,\epsilon}\cap A'|}{|\mathcal{B}_{\theta,\epsilon}\cap\Theta|} d\nu^*(\theta).$$

事实上，容易验证 $\Upsilon(\Theta, A') = \nu^\epsilon(A')$ 且 $\Upsilon(A, \Theta) = \nu^*(A)$。因此，

$$W_1(\nu^*,\nu^\epsilon) \leq \int_\Theta\int_\Theta |\theta-\theta'| d\zeta(\theta,\theta') = \int_\Theta \int_{\theta'\in\mathcal{B}_{\theta,\epsilon}\cap\Theta} |\mathcal{B}_{\theta,\epsilon}\cap\Theta|^{-1}|\theta-\theta'| d\theta' d\nu^*(\theta) \leq \epsilon.$$

由于对于任意测度 $\nu \in \mathcal{P}(\Theta)$，都有 $\|\nu^* - \nu\|_{BL}^* \leq W_1(\nu^*, \nu)$，因此我们得到 $\mathcal{Q}_\nu$ 中第一项的一个上界 ϵ。

现在我们转向相对熵项。注意到 ν^ϵ 的定义以及 Fubini 定理可得

$$\frac{d\nu^\epsilon}{d\theta}(\theta) = \int_\Theta |\mathcal{B}_{\theta',\epsilon} \cap \Theta|^{-1} \mathbb{1}_{\mathcal{B}_{\theta',\epsilon} \cap \Theta}(\theta) d\nu^*(\theta');$$

因此，由函数 $x \mapsto x\log(x)$ 的凸性、Jensen 不等式和 Fubini 定理可得

$$\begin{aligned} \int_\Theta \frac{d\nu^\epsilon}{d\theta}(\theta) \log\left(\frac{d\nu^\epsilon}{d\theta}(\theta)\right) d\theta &\leq \int_\Theta \int_\Theta |\mathcal{B}_{\theta',\epsilon} \cap \Theta|^{-1} \mathbb{1}_{\mathcal{B}_{\theta',\epsilon}}(\theta) \log(|\mathcal{B}_{\theta',\epsilon} \cap \Theta|^{-1}) d\nu^*(\theta') d\theta \\ &\leq -\int_\Theta \log(|\mathcal{B}_{\theta',\epsilon} \cap \Theta|) d\nu^*(\theta') \leq -\log(k') - d\log(\epsilon), \end{aligned} \tag{4.11}$$

w这里我们采用了约定 $0 \times -\infty = 0$。另一方面，根据假设，

$$\int_\Theta \frac{d\nu^\epsilon}{d\theta}(\theta) \log\left(\frac{d\nu_0}{d\theta}(\theta)\right) d\theta = \int_\Theta \log\left(\frac{d\nu_0}{d\theta}(\theta)\right) d\nu^\epsilon(\theta) \geq \log(k). \tag{4.12}$$

由上述内容可得

$$\mathcal{H}(\nu^\epsilon || \nu^0) \leq -\log(k) - \log(k') - d\log(\epsilon).$$

因此，

$$\mathcal{Q}(\nu_0, \nu^*; \tau) \leq \epsilon - \frac{1}{\tau}(d\log(\epsilon) + \log(k') + \log(k)),$$

对于每个 $\epsilon > 0$ 都成立。令 $\epsilon = \frac{d}{\tau}$ 为上述表达式右侧的最小值点，便可得到所需结果。□

备注 4.7。由假设 Θ 的边界是 *Lipschitz* 的这一事实可推出条件 $|\mathcal{B}_{\theta,\epsilon} \cap \Theta| \geq k'\epsilon^d$ 对 $\theta \in \Theta$ 一致成立；参见假设 *1.4*。

引理 4.8。假设 *1.4* 和 *1.13* 成立。设 π_0 满足 $\pi_{0,z} = \mu$，并假设存在 $k > 0$ 使得对于 μ 的支撑集中的所有 z，均有 $\frac{d\pi_0(\tilde{z}|z)}{d\tilde{z}}(\tilde{z}) > k$。进一步假设 $|\mathcal{B}_{\tilde{z},\epsilon} \cap Z| \geq k'\epsilon^{d'}$ 对 $\tilde{z} \in Z$ 一致成立。那么，对于所有满足 $\pi_z^* = \mu$ 的 π^*，我们有

$$\mathcal{Q}_\pi(\pi_0, \pi^*; \tau) \leq \frac{d'}{\tau}\left\{1 - \log\left(\frac{d'}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\}.$$

证明。由于所有相关测度都必须具有相同的第一边缘分布（即 μ），我们按照引理 4.6 的方法进行，但这次只对条件分布进行正则化。更确切地说，对于支撑集中的给定 $z \mathcal{Z}$，我们将测度 $\pi^\epsilon(\cdot|z)$ 定义如下：

$$\pi^\epsilon(A|z) := \int_{\mathcal{Z}} \frac{|\mathcal{B}_{\tilde{z},\epsilon} \cap A|}{|\mathcal{B}_{\tilde{z},\epsilon} \cap \mathcal{Z}|} d\pi^*(\tilde{z}|z).$$

接下来，我们定义测度 π^ϵ 如下：

$$d\pi^\epsilon(z, \tilde{z}) = d\pi^\epsilon(\tilde{z}|z) d\mu(z).$$

注意，测度 π^ϵ 满足 $\pi_z^\epsilon = \mu$。此外，显然可以证明（重复引理 4.6 证明中的类似计算），$W_1(\pi^\epsilon, \pi^*) \leq \epsilon$ 且 $\mathcal{H}(\pi^\epsilon||\pi_0) \leq -\log(k) - \log(k') - d'\log(\epsilon)$。所需结果现可由引理 4.6 同理推出。□

定理*1.16*的证明。一方面，根据假设，我们可以找到 T_1，使得对所有 $t > T^*$

$$|2B^2 \frac{1}{t} \int_0^t \int_0^s \eta_u du ds| \leq \frac{3}{4}\epsilon.$$

另一方面，引理 4.6 和 4.8 表明，存在 T_2，使得对于所有 $t > T^*$ 以及满足 $\pi_z^* = \mu$ 和 ν^* 的任意 π^*，都有

$$(\mathcal{Q}(\pi_0, \pi^*; \kappa B t) + (\mathcal{Q}(\nu_0, \nu^*; \kappa B t) \leq \frac{\epsilon}{4B}.$$

最后，取 $T^* = T_1 \vee T_2$ 并使用引理 4.5，即可得出结论。 □

4.2. 非凸且强凹情形。 本节给出定理 1.21 的证明。

定理 1.21 的证明。 在整个证明中，我们用 m_t^* 表示以下量

$$m_t^* := \sup_{\pi \text{ s.t. } \pi_z = \mu} \mathcal{U}(\pi, \nu_t).$$

由$\mathcal{U}$ 在意义下的凹凸性（线性插值），对于任意 $\tilde{\pi}$（其中 $\tilde{\pi}_z = \mu$）和 $\tilde{\nu}$，有：

$$\begin{aligned}
\mathcal{U}(\tilde{\pi}, \overline{\nu}_t) - \mathcal{U}(\overline{\pi}_t, \tilde{\nu}) &\le \frac{1}{t}\int_0^t (\mathcal{U}(\tilde{\pi}, \nu_s) - \mathcal{U}(\pi_s, \tilde{\nu}))ds \\
&= \frac{1}{t}\int_0^t (\mathcal{U}(\tilde{\pi}, \nu_s) - m_s^*)ds + \frac{1}{t}\int_0^t (m_s^* - \mathcal{U}(\pi_s, \tilde{\nu}))ds \\
&\le \frac{1}{t}\int_0^t (m_s^* - \mathcal{U}(\pi_s, \tilde{\nu}))ds \\
&= \frac{1}{t}\int_0^t (m_s^* - \mathcal{U}(\pi_s, \nu_s))ds + \frac{1}{t}\int_0^t (\mathcal{U}(\pi_s, \nu_s) - \mathcal{U}(\pi_s, \tilde{\nu}))ds. \\
&=: \mathcal{I}_1 + \mathcal{I}_2
\end{aligned} \tag{4.13}$$

在上式中，第二个不等式由 m_s^* 的定义得出。现在我们将控制上式右侧的各项 $\mathcal{I}_1$ 和 $\mathcal{I}_2$ 。

为了控制$\mathcal{I}_1$，，我们先利用链式法则（例如，参见 [1]中的第 10.1.2 节），得到 $\frac{d}{ds}\mathcal{U}(\pi_s, \nu_s)$ 的表达式：

$$\begin{aligned}
\frac{d}{ds}\mathcal{U}(\pi_s, \nu_s) = &\int |\nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z})|^2 d\pi_s(z, \tilde{z}) + \kappa \int \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z})(\mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}) - \overline{\mathcal{U}}_{\pi,z}) d\pi_s(z, \tilde{z}) \\
&- \frac{\eta_t}{K}\int |\nabla_\theta \mathcal{U}_\nu(\pi_s, \nu_s; \theta)|^2 d\nu_s(\theta) - \frac{\kappa}{K}\int \mathcal{U}_\nu(\pi_s, \nu_s; \theta)(\mathcal{U}_\nu(\pi_s, \nu_s; \theta) - \mathcal{U}_\nu) d\nu_s(\theta);
\end{aligned} \tag{4.14}$$

在上式中，我们用简记 $\mathcal{U}_{\pi,z}$ 表示 $\int \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}')d\pi_s(\tilde{z}'|z)$，并用$\mathcal{U}_\nu$ 表示 $\int \mathcal{U}_\nu(\pi_s, \nu_s; \theta')d\nu_s(\theta')$。假设1.18意味着，(4.14) 右侧的第一项有下界 $\lambda(m_s^* - \mathcal{U}(\pi_s, \nu_s))$。另一方面，由 Jensen 不等式可知，第二项非负。最后，由假设 1.4 可知，最后几项有下界 $-\|\eta\|_K^\infty M^2 - \frac{2\kappa}{K}M^2$。因此，对于所有 $t \ge 0$

$$\begin{aligned}
\mathcal{U}(\pi_t, \nu_t) =& \mathcal{U}(\pi_0, \nu_0) + \int_0^t \frac{d}{dr}\mathcal{U}(\pi_r, \nu_r)dr \\
\ge& \mathcal{U}(\pi_0, \nu_0) - \frac{t}{K}\tilde{B} + \lambda\int_0^t (m_s^* - \mathcal{U}(\pi_s, \nu_s))ds,
\end{aligned}$$

其中 $\tilde{B} := (\|\eta\|_\infty + 2\kappa)M^2$。将 m_t^* 从上述不等式两边相减，借助假设 1.4，我们得到，

$$\mathcal{U}(\pi_t, \nu_t) - m_t^* \ge -2M - \frac{t}{K}\tilde{B} + \lambda\int_0^t (m_s^* - \mathcal{U}(\pi_s, \nu_s))ds.$$

等价地，

$$m_t^* - \mathcal{U}(\pi_t, \nu_t) \le 2M + \frac{t}{K}\tilde{B} - \lambda\int_0^t (m_s^* - \mathcal{U}(\pi_s, \nu_s))ds.$$

由此可见，函数 $f(t) := m_t^* - \mathcal{U}(\nu_t, \pi_t)$ 满足

$$f(t) \le 2M + \frac{\tilde{B}}{K}t - \lambda \int_0^t f(s)ds,$$

并由附录 A.3 中的引理 A.6 可得

$$\mathcal{I}_1 \le \frac{\tilde{B}}{K\alpha} + \frac{A}{t},$$

对于 $A := \frac{1}{\lambda}|2M - \frac{B}{K\lambda}|$。

为了估计 (4.15) 中的 $\mathcal{I}_2$，我们按照引理 4.5 的证明进行类似计算，可得

$$\begin{aligned}\int_0^t \mathcal{U}(\pi_s, \nu_s)ds - \int_0^t \mathcal{U}(\pi_s, \tilde{\nu})ds \le & \int_0^t (|\mathcal{U}_\nu(\pi_s, \nu_s; \cdot)\|_{BL}\|\tilde{\nu} - \hat{\nu}_s\|_{BL}^*)ds \\ & - \int_0^t \left(\int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s - \nu_s)(\theta) \right) ds,\end{aligned} \tag{4.15}$$

其中现在我们使用满足以下条件的修改后的帽过程 $\hat{\nu}$

$$\partial_t \hat{\nu}_t = \frac{\eta_t}{K} \mathrm{div}_\theta(\hat{\nu}_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \theta)),$$

其初始值为任意一个密度在 $L^\infty(\nu_0)$ 中的 $\hat{\nu}_0 \ll \nu_0$。对命题 4.3 作简单调整后，我们便可得

$$\mathcal{H}(\hat{\nu}_t||\nu_t) - \mathcal{H}(\hat{\nu}^0||\nu^0) \le \frac{\kappa}{K} \int_0^t \left(\int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s - \nu_s)(\theta) \right) ds, \quad \forall t \ge 0,$$

由此可得

$$\begin{aligned}\mathcal{I}_2 &\le \frac{1}{t} \int_0^t (\|\mathcal{U}_\nu(\pi_s, \nu_s; \cdot)\|_{BL}\|\tilde{\nu} - \hat{\nu}_s\|_{BL}^*)ds + \frac{K}{\kappa t}\mathcal{H}(\hat{\nu}_0||\nu_0) \\ &\le B\|\tilde{\nu} - \hat{\nu}_0\|_{BL}^* + \frac{B^2}{Kt}\int_0^t \int_0^s \eta_\tau d\tau ds + \frac{K}{\kappa t}\mathcal{H}(\hat{\nu}_0||\nu_0).\end{aligned}$$

由上述内容可推得

$$\mathcal{I}_2 \le B\mathcal{Q}_\nu(\nu_0, \tilde{\nu}; \frac{\kappa}{K}Bt)) + \frac{B^2}{Kt}\int_0^t \int_0^s \eta_\tau d\tau ds.$$

综合所有估计，我们得到

$$\mathcal{U}(\tilde{\pi}, \overline{\nu}_t) - \mathcal{U}(\overline{\pi}_t, \tilde{\nu}) \le \frac{\tilde{B}}{K\lambda} + \frac{A}{t} + B\mathcal{Q}_\nu(\nu_0, \tilde{\nu}; \frac{\kappa}{K}Bt)) + \frac{B^2}{Kt}\int_0^t \int_0^s \eta_\tau d\tau ds.$$

此时，我们可以利用 ν_0 的具体性质，并应用引理 4.6 得出存在 $r_0(\epsilon)$ $K_0(\epsilon)$ $t_0(\epsilon)$ $r_1(\epsilon)$，使得若 $\frac{K}{t} \le r_0(\epsilon)$、$K \ge K_0(\epsilon)$ $t \ge t_0(\varepsilon)$、$\eta/K \le r_1(\epsilon)$，则

$$\sup_{\tilde{\pi} \in \mathcal{P}(\mathcal{Z}^2) \text{ s.t. } \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi}, \overline{\nu}_t) - \inf_{\tilde{\nu} \in \mathcal{P}(\Theta)} \mathcal{U}(\overline{\pi}_t, \tilde{\nu}) \le \epsilon.$$

□

5. 数值示例

我们以 MNIST 数据集上的图像分类为背景，通过数值实验展示我们的结果。

在这一框架中，我们将表示分布 ν 的粒子设为具有固定宽度和深度[2]的简单卷积网络的训练参数（即权重和偏置）；而表示分布 π 的粒子则是图像对，其中第一幅

[2]两个卷积核大小为 5、输出通道数分别为 32 和 64 的卷积层，采用 ReLU 激活函数和最大池化；末尾还有两个线性层，其输出尺寸分别为 1000 和 10。

图像来自原始数据库，第二幅则是在训练过程中生成的对抗图像。我们将损失函数定义为

$$\mathcal{R}(\pi,\nu)=\int_{\mathcal{Z}\times\mathcal{Z}}\int_{\Theta}|h_\theta(\tilde{x})-\tilde{y}|^2 d\nu(\theta)d\pi(z,\tilde{z});\quad \mathcal{C}(\pi)=c_a\int_{\mathcal{Z}\times\mathcal{Z}}|z-\tilde{z}|^2 d\pi(z,\tilde{z})$$

其中， $h_\theta(x)$ 是将网络参数设为 θ 时，卷积网络对输入 x 的输出。

鉴于卷积架构具有非线性结构，直接应用算法1中的第19步时间平均会极其耗费内存，因为这要求我们在训练过程中跟踪所有中间网络的副本。一种可能的解决方案（例如 [13]，中提出的方案）是仅对权重计算时间平均，同时保留网络参数粒子的最后位置。我们还实现了另一种基于粒子滤波器所用重采样方法的方案（综述见 [28] ）：内存中最多保留一定数量的网络参数（M'）。在每个更新时刻，我们使用残差系统重采样（RSR），从内存中已有的 M' 个参数以及新加入的一批 M 个粒子中，选出 M' 个参数予以保留。关于（RSR）方法的详细信息见 [28] （例如表2中的代码4）。对抗图像的时间平均计算也采用类似方法。

为阐明我们的主要结果，我们计算以下比率的近似值

$$r_a:=\frac{\sup_{\tilde{\pi}\in\mathcal{P}(\mathcal{Z}^2)\text{ s.t. }\tilde{\pi}_z=\mu}\mathcal{U}(\tilde{\pi},\nu^\star)}{\mathcal{U}(\pi^\star,\nu^\star)},\text{ and }r_m:=\frac{\inf_{\tilde{\nu}\in\mathcal{P}(\Theta)}\mathcal{U}(\pi^\star,\tilde{\nu})}{\mathcal{U}(\pi^\star,\nu^\star)},$$

其中，$(\pi^\star,\ \nu^\star)$ 是训练后得到的网络和对抗图像的时间平均分布。根据我们的结果，我们应达到近似的纳什均衡，因此预期这两个比率都接近于零。近似值的计算方法如下：我们通过在训练每个表示 $\pi^\star$ 的网络时固定 $\nu^\star$ ，来近似求取 r_m 中的上确界；训练采用随机梯度下降，训练轮次数固定（权重保持为常数）。随后，我们计算此过程后总风险的相对变化。 r_a 的近似值以类似方式计算。表1汇总了数值实验所用的参数及所得结果。

模型参数	
N	4
M	2
η_t	0。1 $(t+1)^{-1}$
κ	0.25
c_a	10

测试实现参数	
数据集	MNIST
批次大小	64
训练轮次	4

结果		
准确率		
	权重时间平均值	重采样
干净样本	96.34%	93.53%
PGD（20步）	62.21%	58.49 %
解处损失的相对变化 - 5 个额外训练轮次		
	权重时间平均值	重采样
r_a	0.21%	0.03 %
r_m	1.82%	3.5%

表1. 数值测试的参数和结果

直观而言，我们预期，网络最终时间平均分布所提供的分类既有效又稳健。为了检验这一想法，我们使用一个干净样本 测试样本（独立于原始分布）来评估该分类器的准确率，并使用一个

对抗 版本进行评估，该版本通过使用投影梯度下降（PGD）对后者进行修改而构造，迭代20步，步长为0.04。PGD通过反复以固定幅度扰动图像中的每个像素来构造对抗图像，扰动方向取为损失函数相对于该像素值的梯度符号。示例见 [33]。该测试结果也列于表1。我们观察到，整个流程使网络在干净样本上的表现略有下降，但显著增强了其抵御对抗攻击的能力。作为参考，使用相同特征的网络进行简单训练所得到的基线，在相同训练轮次下，干净样本准确率为98.41%，但经过PGD攻击后的准确率仅为0.68%（另请比较 [19] 中的结果）。表1显示，在所测试的情形中，仅对权重计算时间平均不仅更简单，而且效果也优于重采样过程。不过，也可能存在本文未探讨的设置，使后一种方法更具优势。对此进行探索将是一个有趣的研究方向。

6. 结论

本文研究了概率测度空间上的极小极大问题，其收益结构受到监督学习情境中的对抗问题启发；我们还研究了旨在求解这些问题的梯度上升-下降动力学。我们所研究的动力学具有偏微分方程演化系统的形式，可通过有限多个相互作用粒子构成的系统进行离散化。在对博弈收益结构作出一些合理假设后，我们可以证明，所提出的粒子系统具有一致性；随着系统中的粒子数量增加，该系统能够恢复原偏微分方程的解。我们还讨论了偏微分方程演化系统在时间趋于无穷时的行为，并证明在某种意义下（见下文），该系统能够为对抗博弈生成近似纳什均衡。我们的结果是在两种值得关注的情形下，在对初始化作出适当假设后得出的：1）非凸非凹收益（凸性和凹性是在适当的最优传输意义下理解的）；2）非凸强凹问题（同样是在适当的最优传输意义下）。这两种情形都符合监督学习任务中的对抗学习博弈；而一般来说，只有在收益函数中引入额外的（外生）正则化惩罚项，才能确保凸性。

由于我们问题中的收益不具凸性（相对于诱导上升-下降动力学的度量而言），我们只能保证偏微分方程系统生成的测度的时间平均在 $t \to \infty$极限下成为近似纳什均衡。因此，为了使我们的算法更贴近理论结果，讨论如何构造不会造成内存过载、且仍能在至少一些基准学习任务中为该博弈恢复近似纳什均衡的替代时间平均策略，就显得尤为重要。

我们的工作启发了若干研究方向。下面列举其中几个。

首先，本文的理论分析假设优化更新会考虑所有（经过扰动的）数据点，但在实践中，一种自然的策略是在每次迭代时使用数据批次来计算损失（及其梯度）。因此，我们认为，研究随机梯度下降（SGD）的使用会如何影响由此得到的偏微分方程系统，是很有意义的。

未来研究的另一个有趣方向，是探索更广泛的对抗学习框架，将多类分类 情形纳入其中（而非本文所考虑的回归问题或仅仅是二分类问题）。原则上，甚至可以考虑这样一种情形：在学习问题中，已有关于类别相似性的先验信息（例如，“吉他”类可能被认为比“棒球”类与“小提琴”类更相似）；在这种情况下，利用此类信息构建更精细的风险模型和允许的对抗攻击模型可能会有所裨益。例如，文献 [44] 讨论了在不同学习任务中利用类别间相似性或层次结构的优势；文献 [12] 则明确探讨了如何利用标签的语义内容来构建标签之间的相似性。

我们的框架似乎确实更适用于回归问题，因为在这种情形下，可以很自然地利用特征空间与标签空间上的 Wasserstein距离之类的度量来定义对手的成本函数 C，而后者就是实数轴。当响应变量具有离散结构时，如何以一种合理（从建模角度而言）的方式为对手定义成本结构，就不那么显而易见了；同时还要确保由此得到的对抗博弈仍能像本文所探讨的那样，通过上升-下降方案求解。

致谢

作者感谢 Theodor Misiakiewicz、Katy Craig、Matt Jacobs 和 L´ena¨ıc Chizat 就本文相关主题展开的启发性讨论。NGT 得到了 NSF-DMS 项目资助（编号 2005797），并感谢威斯康星大学麦迪逊分校的 IFDS 以及 NSF 通过 TRIPODS 项目（编号 2023239）提供的支持。

附录 A. 辅助结果与计算

A.1. 关于假设1.18的PL条件。

命题A.1. 假设 *Z* 是一个凸集，并在 *1.2* 的设定中选取一个二阶连续可微的激活函数和损失函数。那么，按 *1.6* 和 *1.7* 中的定义取 *R* 和 *C* 的函数 *U*，对于所有足够大的 c_a，都满足假设*1.18*中的条件。

证明。 我们回顾一下，在这种情况下有

$$\mathcal{U}_\pi(\pi,\nu;z,\tilde{z}) = \ell(h_\nu(\tilde{x}),\tilde{y}) - c_a|z-\tilde{z}|^2 =: \mathcal{U}(\nu;z,\tilde{z});$$

参见例1.2。假设损失函数 ℓ 和激活函数 f 至少二次连续可微，则可得函数 $\tilde{z} \in \mathcal{Z} \mapsto \mathcal{U}(\nu; z, \tilde{z})$（对于固定的 z 和 ν）是强凹的（对于所有 z 和 ν），前提是 c_a 足够大。事实上，这是因为我们可以对 z, ν 上的情况统一地界定 $\mathcal{U}(\nu; z, \tilde{z})$. 中第一项的二阶导数。由此以及 [50], 中的定理5.15 ii)，我们可推得存在 $\lambda > 0$，使得对于每个 $z \in \mathcal{Z}$ 和 $\Upsilon \in \mathcal{P}(\mathcal{Z})$，都有

$$\int_{\mathcal{Z}} |\nabla_{\tilde{z}} \mathcal{U}(\nu;z,\tilde{z})|^2 d\Upsilon(\tilde{z}) \geq \lambda\Big(\sup_{\hat{\Upsilon}\in\mathcal{P}(\mathcal{Z})} \int_{\mathcal{Z}} \mathcal{U}(\nu;z,\tilde{z}) d\hat{\Upsilon}(\tilde{z}) - \int_{\mathcal{Z}} \mathcal{U}(\nu;z,\tilde{z}) d\Upsilon(\tilde{z})\Big).$$

特别地，对于给定的 $\pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$ 且满足 $\pi_{0,z} = \mu$，我们有

$$\int_{\mathcal{Z}} |\nabla_{\tilde{z}} \mathcal{U}(\nu;z,\tilde{z})|^2 d\pi(\tilde{z}|z) \geq \lambda\Big(\sup_{\hat{\Upsilon}\in\mathcal{P}(\mathcal{Z})} \int_{\mathcal{Z}} \mathcal{U}(\nu;z,\tilde{z}) d\hat{\Upsilon}(\tilde{z}) - \int_{\mathcal{Z}} \mathcal{U}(\nu;z,\tilde{z}) d\pi(\tilde{z}|z)\Big),$$

对所有 $z \in \mathcal{Z}$ 以及所有 $\nu \in \mathcal{P}(\Theta)$。对上述不等式两边关于 μ 在 z 上积分，可得

$$\begin{aligned}
\int_{\mathcal{Z}\times\mathcal{Z}} |\nabla_{\tilde{z}} \mathcal{U}_\pi(\pi,\nu;z,\tilde{z})|^2 d\pi(z,\tilde{z}) &= \int_{\mathcal{Z}\times\mathcal{Z}} |\nabla_{\tilde{z}} \mathcal{U}(\nu;z,\tilde{z})|^2 d\pi(z,\tilde{z}) \\
&\geq \lambda\left(\int_{\mathcal{Z}} \left(\sup_{\hat{\Upsilon}\in\mathcal{P}(\mathcal{Z})} \int_{\mathcal{Z}} \mathcal{U}(\nu;z,\tilde{z}) d\hat{\Upsilon}(\tilde{z})\right) d\mu(z) - \mathcal{U}(\pi,\nu)\right) \\
&\geq \lambda\left(\sup_{\hat{\pi}\in\mathcal{P}(\mathcal{Z}^2) \text{ s.t. } \hat{\pi}_z=\mu} \mathcal{U}(\hat{\pi},\nu) - \mathcal{U}(\pi,\nu)\right).
\end{aligned}$$

□

A.2. **定理1.8和3.11中近似初始化构造的辅助引理。**

命题 A.2. 设 A 和 B 分别是 $\mathbb{R}^d$ 和 $\mathbb{R}^{d'}$ 的有界博雷尔子集。设 $\mu \in P(A)$，且 $u \in A \mapsto \mu_u(\cdot) \in P(B)$ 是一个可测映射。

对于满足以下条件的任意序列 $\{\Upsilon_n\}_{n\in\mathbb{N}} \subseteq \Gamma(\mu,\mu)$

$$\lim_{n\to\infty} \int_{A\times A} |u-u'| d\Upsilon_n(u,u') = 0,$$

我们有

$$\lim_{n\to\infty} \int_{A\times A} W_1(\mu_u, \mu_{u'}) d\Upsilon_n(u,u') = 0.$$

证明。 满足命题假设的序列 $\{\Upsilon_n\}_{n\in\mathbb{N}}$ 称为传输计划的停滞序列； 参见 [22]。

设 $\varepsilon > 0$。对于这样的 $\varepsilon > 0$，我们可以构造集合 B 的一个有限划分 $\{B_l\}_{l=1,\dots,L}$，使得每个集合 B_l 的直径都小于 $\varepsilon/3$；只需将 $\mathbb{R}^{k'}$ 中直径小于 $\varepsilon/3$ 的盒子网格与集合 B 相交，即可构造出该划分。现在，在每个 B_l 中选取一点 v_l。对于每个 $l = 1,\dots,L$，定义函数 $h_l \in L^1(\mu)$ 如下：对于支撑集为 μ 的任意 u，定义 $h_l(u) := \mu_u(B_l)$。现在考虑测度 $\hat{}\mu_u := \sum_{l=1}^L h_l(u)\delta_{v_l}$。注意，这些都是满足 $W_1(\hat{}\mu_u, \mu_u) \le \varepsilon/3$ 的概率测度。特别地，利用 W_1 的三角不等式，我们可得

$$\begin{aligned}\int_{A\times A} W_1(\mu_u,\mu_{u'}) d\Upsilon_n(u,u') &\le \int_{A\times A} W_1(\mu_u,\hat{\mu}_u) d\Upsilon_n(u,u') + \int_{A\times A} W_1(\hat{\mu}_u,\hat{\mu}_{u'}) d\Upsilon_n(u,u') \\ &\quad + \int_{A\times A} W_1(\hat{\mu}_{u'},\mu_{u'}) d\Upsilon_n(u,u'). \\ &\le \frac{2}{3}\varepsilon + \int_{A\times A} W_1(\hat{\mu}_u,\hat{\mu}_{u'}) d\Upsilon_n(u,u').\end{aligned}$$

现在，我们来为项 $\int_{A\times A} W_1(\hat{}\mu_u, \hat{}\mu_{u'}) d\Upsilon_n(u, u')$ 求一个上界。根据 W_1 距离的康托罗维奇对偶性，我们有

$$W_1(\hat{\mu}_u,\hat{\mu}_{u'}) = \sup_{Lip(f)\le 1} \{\int f(v) d\hat{\mu}_u(v) - \int f(v) d\hat{\mu}_{u'}(v)\}.$$

由于集合 B 有界，且上确界中的表达式在给定的 f 加上一个常数后保持不变，因此我们还可以假设该上确界是在上确界范数受某个固定常数 C 限制的函数 f 上取到的。对于这样的函数 f，我们有

$$\int f(v) d\hat{\mu}_u(v) - \int f(v) d\hat{\mu}_{u'}(v) = \sum_{l=1}^L (h_l(u) - h_l(u')) f(v_l) \le C \sum_{l=1}^L |h_l(u) - h_l(u')|.$$

由上可得

$$\int_{A\times A} W_1(\hat{\mu}_u,\hat{\mu}_{u'}) d\Upsilon_n(u,u') \le C \sum_{l=1}^L \int_{A\times A} |h_l(u) - h_l(u')| d\Upsilon_n(u,u').$$

现在，我们应用 [22] 中的引理 3.10，得出上述表达式右侧在 $n \to \infty$ 时趋于零。特别地，存在足够大的 N，使得对于所有 $n \ge N$，都有 $C\sum_{l=1}^L \int_{A\times A} |h_l(u) - h_l(u')| d\Upsilon_n(u, u') \le \frac{\varepsilon}{3}$。进而，我们得出，如果 $n \ge N$，那么

$$\int_{A\times A} W_1(\mu_u,\mu_{u'}) d\Upsilon_n(u,u') \le \varepsilon.$$

这就得到了所需的结果。 □

引理 A.3. 设 A 和 B 分别是 $\mathbb{R}^d$ 和 $\mathbb{R}^{d'}$ 的有界博雷尔子集。设 $\mu \in P(A)$，且 $u \in A \mapsto \mu_u(\cdot) \in P(B)$ 是可测映射。

设 $u_1, \ldots, u_n, \ldots$ 是来自 μ 的独立同分布样本，并且对每个 $i \in \mathbb{N}$，设 $v_{i1}, \ldots, v_{im}, \ldots$ 是来自 $\mu_{u_i}(\cdot)$ 的独立同分布样本。对每个 n 和 m，考虑如下（随机）测度

$$\mu^{n,m} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{(u_i,v_{ij})}, \quad \mu^n := \frac{1}{n}\sum_{i=1}^{n}\delta_{u_i},$$

并令 $\mu^{n,m}(\cdot|u)$ 为根据 $\mu^{n,m}$ 得到的条件分布，即在给定第一个坐标的取值 u 时，变量 v 的条件分布。

Then

$$\lim_{n\to\infty}\lim_{m\to\infty}\mathbb{E}\left[\inf_{\Upsilon_n\in\Gamma_{Opt}(\mu^n,\mu)}\int W_1(\mu^{n,m}(\cdot|u),\mu_{u'})d\Upsilon_n(u,u')\right]=0.$$

特别地，存在一个序列 $\{(n_k, m_k)\}_{k\in\mathbb{N}}$，使得

$$\lim_{k\to\infty}\mathbb{E}\left[\inf_{\Upsilon_k\in\Gamma_{Opt}(\mu^{n_k},\mu)}\int W_1(\mu^{n_k,m_k}(\cdot|u),\mu_{u'})d\Upsilon_k(u,u')\right]=0,$$

以及一个子序列（不重新标记），使得

$$\lim_{k\to\infty}\inf_{\Upsilon_k\in\Gamma_{Opt}(\mu^{n_k},\mu)}\int W_1(\mu^{n_k,m_k}(\cdot|u),\mu_{u'})d\Upsilon_k(u,u')=0,$$

几乎必然。

证明。令 $\Upsilon_n \in \Gamma_{\mathrm{Opt}}(\mu^n, \mu)$。根据 [49] 中的推论 5.22，可以用一种可测的方式选取这个随机测度，定义在为问题中的随机变量提供支撑集的、隐含定义的样本空间上。

由 W_1 的三角不等式可得

(A.1)

$$\begin{aligned}\int W_1(\mu^n(\cdot|u),\mu_{u'})d\Upsilon_n(u,u') &\le \int W_1(\mu^n(\cdot|u),\mu_u)d\Upsilon_n(u,u') + \int W_1(\mu_u,\mu_{u'})d\Upsilon_n(u,u')\\ &= \frac{1}{n}\sum_{i=1}^{n}W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i}) + \int W_1(\mu_u,\mu_{u'})d\Upsilon_n(u,u').\end{aligned}$$

接下来，我们分析上述表达式右侧的各项。先来看第二项。

引入 $\hat{\Upsilon}_n := \mathbb{E}[\Upsilon_n]$，这是一个（确定性）测度，其对测试函数 ϕ 的作用方式如下：

$$\int \phi(u,u')d\hat{\Upsilon}_n(u,u') = \mathbb{E}[\int \phi(u,u')d\Upsilon_n(u,u')].$$

显然，$\hat{\Upsilon}_n \in \Gamma(\mu, \mu)$。现在，由于空间 A 有界，且 μ^n 几乎必然弱收敛到 μ，我们知道，几乎必然有

$$\lim_{n\to\infty}\int |u-u'|d\Upsilon_n(u,u') = \lim_{n\to\infty} W_1(\mu^n,\mu) = 0.$$

因此，由控制收敛定理可得

$$\lim_{n\to\infty}\int |u-u'|d\hat{\Upsilon}_n(u,u') = \lim_{n\to\infty}\mathbb{E}[\int |u-u'|d\Upsilon_n(u,u')] = 0.$$

特别地，$\{\hat{\Upsilon}_n\}_{n\in\mathbb{N}}$ 是针对 μ 的传输计划停滞序列，因此由引理 A.2 可知

$$\lim_{n\to\infty}\mathbb{E}[\int W_1(\mu_u,\mu_{u'})d\Upsilon_n(u,u')] = \lim_{n\to\infty}\int W_1(\mu_u,\mu_{u'})d\hat{\Upsilon}_n(u,u') = 0.$$

现在我们研究（A.1）右端的第一项。为避免引入繁琐的记号，为简单起见，我们假设所有 u_i 均不相同，因此特别有$\mu^{n,m}(\cdot|u_i) = \frac{1}{m}\sum_{j=1}^{m}\delta_{v_{ij}}$。于是我们得到

$$\begin{aligned}\lim_{m\to\infty}\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i})] &= \mathbb{E}[\lim_{m\to\infty}\frac{1}{n}\sum_{i=1}^{n}W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i})] \\ &= \mathbb{E}[\mathbb{E}[\lim_{m\to\infty}\frac{1}{n}\sum_{i=1}^{n}W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i})|u_1,\ldots,u_n]] \\ &= 0,\end{aligned} \tag{A.2}$$

其中，第一行使用了控制收敛定理；最后一行则利用了$\frac{1}{m}\sum_{j=1}^{m}\delta_{v_{ij}}$ 在 Wasserstein 意义下几乎必然收敛到 μ_{u_i} 这一事实。□

注 A.4. 设 $\{\mu^n\}_{n\in\mathbb{N}}$ 是 $A\times B$ 上的一列概率测度，且 μ 是一个概率测度。我们证明，条件

$$\inf_{\Upsilon_n\in\Gamma_{Opt}(\mu_u^n,\mu_u)}\int W_1(\mu^n(\cdot|u),\mu(\cdot|u'))d\Upsilon_n(u,u')\to 0$$

蕴含

$$W_1(\mu^n,\mu)\to 0,$$

而反过来一般并不成立；上文中，μ_u^n 和 μ_u 分别表示 μ^n 和 μ的第一边缘分布。事实上，假设第一个条件成立，并对每个 u，u'取 $\Upsilon^{u,u'}$ 为 $\mu^n(\cdot|u)$与 $\mu(\cdot|u')$之间实现 W_1 距离的一个耦合。此外，选取Υ_n属于 $\Gamma_{Opt}(\mu^n,\mu)$，使得 $\int W_1(\mu^n(\cdot|u)\ \mu(\cdot|u'))d\Upsilon_n(u,u')\to 0$，并考虑测度$d\pi_n((u,v),(u',v')) := d\Upsilon^{u,u'}(v,v')d\Upsilon_n(u,u')$。不难验证，$\pi_n\in\Gamma(\mu^n,\mu)$，且 $\int|(u,v)-(u',v')|d\pi_n\to 0$。这意味着 $W_1(\mu^n,\mu)\to 0$。

正如我们先前所述，逆命题并不成立。例如，取 $A=[0,1]$, $B=[0,1]$, μ 为 $[0,1]^2$, 上的均匀测度，并令 $(u_1,v_1),\ldots,(u_n,v_n)$ 为来自 μ的独立同分布样本，我们看到 $W_1(\mu^n,\mu)\to 0$，而对所有 n，$\inf_{\Upsilon_n\in\Gamma_{Opt}(\mu_u^n,\mu_u)}\int W_1(\mu^n(\cdot|u)\ \mu(\cdot|u'))d\Upsilon_n(u,u')=1$。

引理 A.5. 考虑与引理 *A.3* 相同的设定和记号。设 $\rho:\ A\times B\to[0,D]$是满足以下条件的可测函数

$$\int\rho(u,v)d\mu_u(v)=1,$$

对 μ的支撑集中的所有 u均成立。那么，以概率 *1*，

$$\lim_{n\to\infty}\lim_{m\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{\frac{1}{m}\sum_{j=1}^{m}\rho(u_i,v_{ij})}-1\right|=0.$$

证明。这是大数定律的直接推论。 □

A.3. **第 4 节的辅助引理**。以下结果可由 Gronwall型论证得出。

Lemma A.6. 设 $B,M,K,\lambda>0$，并假设 $h:\,[0,\infty)\to[0,\infty)$ 是一个满足以下条件的函数

$$h(t)\le 2M+\frac{\tilde{B}}{K}t-\lambda\int_0^t h(s)ds,$$

对所有 t 均成立。那么，对所有 $T>0$,

$$\frac{1}{T}\int_0^T h(s)ds\le\frac{\tilde{B}}{K\lambda}+\frac{A}{T},$$

其中 $A:=\frac{1}{\lambda}|2M-\frac{B}{K\lambda}|$。

证明。 h 上的条件可等价地写为

$$h(t) - \frac{\tilde{B}}{K\lambda} \leq (2M - \frac{\tilde{B}}{K\lambda}) - \lambda \int_0^t (h(s) - \frac{\tilde{B}}{K\lambda}) ds.$$

令 $H(t) := \int_0^t (h(s) - \frac{\tilde{B}}{K\lambda}) ds$。因此，上述条件可写为

$$H'(t) \leq (2M - \frac{\tilde{B}}{K\lambda}) - \lambda H(t).$$

由此可得

$$\frac{d}{dt}(H(t)e^{\lambda t}) \leq (2M - \frac{\tilde{B}}{K\lambda}) e^{\lambda t}.$$

对上述表达式积分，得到:

$$H(t)e^{\lambda t} \leq (2M - \frac{\tilde{B}}{K\lambda}) \frac{1}{\lambda} (e^{\lambda t} - 1),$$

或者等价地

$$H(t) \leq (2M - \frac{\tilde{B}}{K\lambda}) \frac{1}{\lambda} - (2M - \frac{\tilde{B}}{K\lambda}) \frac{1}{\lambda} e^{-\lambda t}.$$

回忆定义 o，可推出 f H d d h

$$\frac{1}{T} \int_0^T h(s) ds \leq \frac{\tilde{B}}{K\lambda} + \frac{1}{T} (2M - \frac{\tilde{B}}{K\lambda}) \frac{1}{\lambda} - \frac{1}{T} (2M - \frac{\tilde{B}}{K\lambda}) \frac{1}{\lambda} e^{-\lambda T} \leq \frac{\tilde{B}}{K\lambda} + \frac{A}{T}.$$

□

参考文献

[1] Luigi Ambrosio、Nicola Gigli 和 Giuseppe Savar´e。《度量空间及概率测度空间中的梯度流》。第二版。苏黎世联邦理工学院数学讲义。Birkh¨auser 出版社，巴塞尔，2008年，第 x+334 页。isbn：978-3-7643-8721-1。[2] Pranjal Awasthi、Natalie Frank 和 Mehryar Mohri。“对抗贝叶斯分类器的存在性”。载于：《神经信息处理系统进展》34（2021年），第 2978–2990 页。[3] Arjun Nitin Bhagoji、Daniel Cullina 和 Prateek Mittal。“基于最优传输的对抗鲁棒性下界”。载于：《神经信息处理系统进展》。H. Wallach 等编。第 32 卷。Curran Associates, Inc.，2019年。网址： https : //proceedings.neurips.cc/paper/2019/file/02bf86214e264535e3412283e817deaa-Paper.pdf.

[4] Jose Blanchet、Yang Kang 和 Karthyek Murthy。“鲁棒 Wasserstein 剖面推断及其在机器学习中的应用”。载于：《应用概率杂志》56.3（2019年），第 830–857 页。[5] Jose Blanchet、Karthyek Murthy 和 Viet Anh Nguyen。“Wasserstein 分布鲁棒估计量的统计分析”。载于：（2021年）。电子预印本： arXiv:2108.02120。[6] V I Bogachev、A V Kolesnikov 和 K V Medvedev。“测度的三角变换”。载于：《数学文集》196.3（2005年4月），第 309 页。doi: 10.1070/SM2005v196n03ABEH000882.

网址： https://dx.doi.org/10.1070/SM2005v196n03ABEH000882.[7] Leon Bungert 和 Kerrek Stinson。《对抗机器学习中出现的非局部周长的伽马收敛》。载于：（2022年）。arXiv： arXiv : 2211 . 15223。网址： https ://arxiv.org/abs/2211.15223.[8] Leon Bungert、Nicol´as Garc´ıa Trillos 和 Ryan Murray。《二分类中对抗训练的几何》。载于：（2021年）。即将在《信息与推断：*IMA*期刊》发表： arXiv:2111.13613。网址： https://arxiv.org/abs/2111.13613.[9] Ruidi Chen 和 Ioannis Ch. Paschalidis。《分布鲁棒学习》。载于：优化基础与趋势® 专辑 4.1-2（2020年），第1–243页。ISSN: 2167-3888。DOI：10.1561/2400000026。网址： http://dx.doi.org/10.1561/2400000026。

[10] L´ena¨ıc Chizat 和 Francis Bach。《利用最优传输研究过参数化模型中梯度下降的全局收敛性》。载于：神经信息处理系统进展。S. Bengio 等编。第31卷。Curran Associates, Inc.，2018年。网址：https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.[11] L´ena¨ıc Chizat 等. “非平衡最优传输：动态表述与 Kantorovich 表述”。载于：《函数分析杂志》274.11（2018年），第3090–3123页。ISSN：0022-1236。DOI：https://doi.org/10.1016/j.jfa.2018.03.008。网址：https://www.sciencedirect.com/science/article/pii/S0022123618301058...Ur¨un Dogan 等. “标签相似度课程学习”。载于：计算机视觉——ECCV 2020。Andrea Vedaldi 等编。Cham：Springer International Publishing，2020年，第174–190页。ISBN：978-3-030-58526-6。[13] Carles Domingo-Enrich 等. “双人零和博弈的平均场分析”。载于：神经信息处理系统进展33（2020年），第20215–20226页。[14] Weinan E、Chao Ma 和 Lei Wu. “神经网络模型的 Barron 空间与流诱导函数空间 ”。载于：《构造性逼近》 55.1 （ 2022 年 2 月），第 369–406页。ISSN：1432-0940。DOI：10.1007/s00365-021-09549-y。网址：https://doi.org/10.1007/s00365-021-09549-y.[15] Tanner Fiez 等. “平滑算法对手下的极小极大优化”。载于：CoRR abs/2106.01488（2021年）。arXiv：2106.01488。网址：https://arxiv.org/abs/2106.01488。[16] Chris Finlay 和 Adam M. Oberman. “用于对抗鲁棒性的可扩展输入梯度正则化”。载于：《机器学习与应用》3（2021年），第100017页。ISSN：2666-8270。DOI：https://doi.org/10.1016/j.mlwa.2020.100017。网址：https://www.sciencedirect.com/science/article/pii/S2666827020300177.[17] Natalie S. Frank. “二分类中对抗代理风险的存在性与极小极大定理”。载于：（2022年）。arXiv: arXiv:2206.09098。网址：https://arxiv.org/abs/2206.09098.[18] Thomas O. Gallou¨et 和 L´eonard Monsaingeon. “用于 Kantorovich–Fisher–Rao 梯度流的 JKO 分裂格式”。载于：SIAM数学分析杂志49.2（2017年），第1100–1130页。DOI：10.1137/16M106666X。电子印本：https://doi.org/10.1137/16M106666X。网址：https://doi.org/10.1137/16M106666X。[19] Camilo Andr´es Garc´ıa Trillos 和 Nicol´as Garc´ıa Trillos. “深度神经网络上分布鲁棒学习的正则化风险”。英文。载于：Research in the Mathematical Sciences 9.3（2022年9月），第54页。ISSN：2522-0144，2197-9847。DOI：10.1007/s40687-022-00349-9。网址：https://link.springer.com/10.1007/s40687-022-00349-9（访问日期：2022年12月19日）。[20] Nicol´as Garc´ıa Trillos、Matt Jacobs 和 Jakwang Kim. “对抗多类分类的多边际最优传输表述”。载于：（2022年）。即将在 JMLR 发表。 arXiv:2204.12676。网址：https://arxiv.org/abs/2204.12676。[21] Nicol´as Garc´ıa Trillos 和 Ryan Murray. “对抗分类：必要条件与几何流”。载于：《机器学习研究杂志》23.187（2022年），第1–38页。网址：http://jmlr.org/papers/v23/21-0222.html。[22] Nicol´as Garc´ıa Trillos 和 Dejan Slep˘cev. “点云上全变差的连续极限”。载于：《理性力学与分析档案》220.1（2016年4月），第193–241页。ISSN：1432-0673。DOI：10.1007/s00205-015-0929-z。网址：https://doi.org/10.1007/s00205-015-0929-z。[23] Ian Goodfellow、Jonathon S hlens 和 Christian Szegedy. “解释并利用对抗样本”。载于：国际学习表征会议。2015年。网址：http://arxiv.org/abs/1412.6572。[24] Jack K Hale. 常微分方程。第二版。马拉巴尔，佛罗里达州：R. E. Krieger，1980年。

[25] Stanislav Kondratyev、L´eonard Monsaingeon 和 Dmitry Vorotnikov. “有限拉东测度空间上的一种新最优传输距离”。载于：Advances in Differential Equations 21.11/12（2016年），第1117–1164页。DOI： ade/1476369298。网址： https://doi.org/10.57262/ade/1476369298。[26] Daniel Kuhn 等. “Wasserstein分布鲁棒优化：理论及其在机器学习中的应用”。2019年10月，第130–166页。ISBN：978-0-9906153-3-0。DOI：10.1287/educ.2019.0198。[27] Alexey Kurakin、Ian J. Goodfellow 和 Samy Bengio. “物理世界中的对抗样本”。载于：第五届学习表征国际会议（ICLR 2017），法国土伦，2017年4月24日至26日，研讨会论文集。OpenReview.net，2017年。网址：https://openreview.net/forum?id=HJGU3Rodl。[28] Tiancheng Li、Miodrag Bolic 和 Petar M. Djuric. “粒子滤波的重采样方法：分类、实现与策略”。载于：IEEE信号处理杂志32.3（2015年5月）。会议名称：IEEE信号处理杂志，第70–86页。ISSN：1558-0792。DOI：10.1109/MSP.2014.2330626。[29] Matthias Liero、Alexander Mielke 和 Giuseppe Savar´e. “与反应竞争中的最优传输：海灵格–康托罗维奇距离与测地线”。载于：SIAM数学分析杂志48.4（2016年），第2869–2911页。[30] Tianyi Lin、Chi Jin 和 Michael Jordan. “非凸-凹极小极大问题的梯度下降上升”。载于：第37届国际机器学习会议论文集。Hal Daum´e III 和 Aarti Singh 编。第119卷。机器学习研究会议录。PMLR，2020年7月13日至18日，第6083–6093页。网址： https://proceedings.mlr.press/v119/lin20a.html。[31] Tianyi Lin、Chi Jin 和 Michael I. Jordan. “极小极大优化的近最优算法”。载于：第三十三届学习理论会议论文集。Jacob Abernethy 和 Shivani Agarwal 编。第125卷。机器学习研究会议录。PMLR，2020年9月，第2738–2779页。网址： https://proceedings.mlr.press/v125/lin20a.html。[32] Chunchuan Lyu、Kaizhu Huang 和 Hai-Ning Liang. “对抗样本的统一梯度正则化族”。载于：2015年IEEE国际数据挖掘会议。2015年，第301–309页。DOI：10.1109/ICDM.2015.84。[33] Aleksander Madry 等. “迈向抵御对抗攻击的深度学习模型”。载于：第六届国际学习表征会议（ICLR 2018），加拿大不列颠哥伦比亚省温哥华，2018年4月30日—5月3日，会议论文集。OpenReview.net，2018年。网址：https://openreview.net/forum?id=rJzIBfZAb。[34] Laurent Meunier 等. “对抗样本博弈中的混合纳什均衡”。载于：第38届国际机器学习会议论文集。Marina Meila 和 Tong Zhang 编。第139卷。机器学习研究会议录。PMLR，2021年7月18日至24日，第7677–7687页。网址： https://proceedings.mlr.press/v139/meunier21a.html。[35] Seyed-Mohsen Moosavi-Dezfooli 等. “通过曲率正则化实现鲁棒性，反之亦然”。载于：2019年IEEE/CVF计算机视觉与模式识别会议（CVPR）。2019年，第9070–9078页。DOI：10.1109/CVPR.2019.00929。[36] Maher Nouiehed 等. “使用迭代一阶方法求解一类非凸极小极大博弈”。载于：第33届神经信息处理系统国际会议论文集。美国纽约州雷德胡克：Curran Associates Inc.，2019年。[37] Muni Sreenivas Pydi 和 Varun Jog. “通过最优传输和最优耦合研究对抗风险”。载于：第37届国际机器学习会议论文集。Hal Daum´e III 和 Aarti Singh 编。第119卷。机器学习研究会议录。PMLR，2020年7月13日至18日，第7814–7823页。网址： https://proceedings.mlr.press/v119/pydi20a.html。[38] Muni Sreenivas Pydi 和 Varun Jog. “对抗风险的多种面貌”。载于：神经信息处理系统进展。M. Ranzato 等编。第34卷。Curran

Associates, Inc., 2021年, pp. 10000–10012. url: https://proceedings.neurips.cc/paper/2021/file/52c4608c2f126708211b9e0a60eaf050-Paper.pdf.[39] Aditi Raghunathan、Jacob Steinhardt 和 Percy Liang。“针对对抗样本的认证防御”。载于：第六届国际学习表征会议，ICLR 2018，温哥华，不列颠哥伦比亚省，加拿大，2018年4月30日—5月3日，会议论文集。OpenReview.net，2018年。url: https://openreview.net/forum?id=Bys4ob-Rb.[40] Andrew Slavin Ross 和 Finale Doshi-Velez。“通过正则化输入梯度提升深度神经网络的对抗鲁棒性和可解释性”。载于：2018年。[41] Kevin Roth 等。“通过结构化梯度正则化进行对抗鲁棒训练”。载于：（2018年）。电子印本：arXiv:1805.08736.[42] Filippo Santambrogio。“面向应用数学家的最优传输”。载于：Birk¨auser，纽约州 55.58-63（2015年），第94页。[43] Thibault S´ejourn´e、Gabriel Peyr´e 和 Fran¸cois-Xavier Vialard。“非平衡最优传输：从理论到数值方法”。载于：（2022年）。arXiv：1711.07667。[44] Carlos N. Silla 和 Alex A. Freitas。“跨不同应用定义域的层次分类综述”。载于：《数据挖掘与知识发现》22.1-2（2010年4月），第31–72页。doi: 10.1007/s10618-010-0175-9。url: https://doi.org/10.1007/s10618-010-0175-9.[45] Aman Sinha、Hongseok Namkoong 和 John C. Duchi。“通过有理论依据的对抗训练认证部分分布鲁棒性”。载于：第六届国际学习表征会议，ICLR 2018，温哥华，不列颠哥伦比亚省，加拿大，2018年4月30日—5月3日，会议论文集。OpenReview.net，2018年。url: https://openreview.net/forum?id=Hk6kPgZA-.[46] Christian Szegedy 等。“神经网络的有趣特性”。载于：第二届国际学习表征会议，ICLR 2014，班夫，艾伯塔省，加拿大，2014年4月14日至16日，会议论文集。Yoshua Bengio 和 Yann LeCun 编。2014年。url:http://arxiv.org/abs/1312.6199.[47] Dimitris Tsipras 等。“鲁棒性可能与准确率相悖”。载于：第七届国际学习表征会议，ICLR 2019，新奥尔良，路易斯安那州，美国，2019年5月6日至9日。OpenReview.net，2019年。url: https://openreview.net/forum?id=SyxAb30cY7.[48] Zhuozhuo Tu、Jingwei Zhang 和 Dacheng Tao。“对抗学习的理论分析：一种极小极大方法”。载于：《神经信息处理系统进展》。H. Wallach 等编。第32卷。Curran Associates, Inc.，2019年。url: https://proceedings.neurips.cc/paper/2019/file/16bda725ae44af3bb9316f416bd13b1b-Paper.pdf.[49] C´edric Villani。《最优传输》。第338卷。《数学科学基础丛书》[《数学科学基本原理》]。新旧理论。Springer-Verlag，柏林，2009年，第xxii+973页。isbn：978-3-540-71049-3。doi: 10.1007/978-3-540-71050-9。url: https://doi-org.ezproxy.library.wisc.edu/10.1007/978-3-540-71050-9.[50] C´edric Villani。《最优传输专题》。第58卷。《数学研究生教材》。美国数学学会，普罗维登斯，罗得岛州，2003年，第xvi+370页。isbn：0-8218-3312-X。doi: 10.1090/gsm/058。url: https://doi-org.ezproxy.library.wisc.edu/10.1090/gsm/058.[51] Guillaume Wang 和 L´ena¨ic Chizat。《一种用于连续博弈混合纳什均衡的指数收敛粒子方法》。arXiv:2211.01280 [cs, math]。2022年11月。url: http://arxiv.org/abs/2211.01280（访问日期：2022年12月14日）。[52] Stephan Wojtowytsch。“平均场机制下双层ReLU网络梯度下降训练的收敛性”。载于：CoRR abs/2005.13530（2020年）。arXiv：2005.13530。url: https://arxiv.org/abs/2005.13530.[53] Stephan Wojtowytsch 和 Weinan E。“浅层神经网络能否克服维数灾难？平均场训练视角”。载于：《IEEE人工智能汇刊》1.2（2020年），第121–129页。doi: 10.1109/TAI.2021.3051357。

[54] Eric C Yeats、Yiran Chen 和 Hai Li。“使用复值神经网络改进梯度正则化”。载于：第*38*届国际机器学习会议论文集。Marina Meila 和 Tong Zhang 编。第139卷。《机器学习研究会议录》。PMLR，2021年7月18日至24日，第11953–11963页。url: https://proceedings.mlr.press/v139/yeats21a.html.[55] Dinghuai Zhang 等。“只传播一次：通过极大值原理加速对抗训练”。载于：神经信息处理系统进展。H. Wallach 等编。第32卷。Curran Associates, Inc.，2019年。url: https://proceedings.neurips.cc/paper/2019/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf.[56] Hongyang Zhang 等。“鲁棒性与准确率之间具有理论依据的权衡”。载于：第*36*届国际机器学习会议论文集。Kamalika Chaudhuri 和 Ruslan Salakhutdinov 编。第97卷。《机器学习研究会议录》。PMLR，2019年9月，第7472–7482页。url: https://proceedings.mlr.press/v97/zhang19p.html。

电子邮箱地址： camilo.garcia@ucl.ac.uk

电子邮箱地址： garciatrillo@wisc.edu