arXiv:2303.01861v1 [stat.ML] 3 Mar 2023

D扩散模型是极小极大最优分布估计器 s

Kazusato Oko [1 2] Shunta Akiyama [1] Taiji Suzuki [1 2]

摘要

高效的分布学习无疑是扩散建模取得突破性成功的关键，但其理论保证相当有限。本文首次对扩散建模在知名函数空间中的近似能力与泛化能力进行了严格分析。本文的主要发现是，当真实密度函数属于 Besov空间，且经验得分匹配损失得到适当最小化时，生成的数据分布在总变差距离和一阶 Wasserstein 距离下，能够达到近乎极小极大最优的估计率。此外，我们将理论拓展到低维数据分布，展示了扩散模型如何适应这类分布。我们希望这些结果能推进对扩散建模及其生成逼真输出能力的理论认识。

1. 引言

扩散建模，也称为基于得分函数的生成建模（Sohl-Dickstein等人，2015；Song & Ermon，2019年；Song 等人，2020；Ho 等人，2020；Vahdat等人，2021），已在图像（Song 等人，2020；Dhariwal与Nichol，2021）、视频（Ho 等人，2022年）和音频（Chen 等人，2020；Kong等人，2020）领域取得最先进的性能。

借鉴Song 等人(2020的统一框架，扩散建模首先逐步向数据分布添加噪声，并将分布转化为预定义噪声分布。这种时间演化称为前向过程，可表示为与数据无关的随机微分方程（SDE）。另一方面，我们可以考虑该随机微分方程的时间反演，沿着这一所谓的反向过程，便可从噪声中生成数据。重要的是，反向过程的漂移项取决于数据分布，具体而言，取决于前向过程中每个时刻对数密度的梯度（得分函数）。

然而，在实践中，我们只能通过有限数量的样本来获取真实分布。因此，我们改用经验分布所对应的扩散过程的得分函数（Vincent，2011年；Sohl-Dickstein等人，2015；Song & Ermon，2019年）。此外，为提高计算效率，我们进一步使用神经网络（得分网络）替代经验得分函数，并通过得分匹配技术（Hyv¨arinen & Dayan，2005；Vincent，2011年）训练该网络，使其在某个损失函数下接近经验得分函数。由此，扩散建模通过学习经验得分函数，间接学习真实分布。

于是，一个自然而然的问题随即产生：扩散建模是否是一个良好的分布估计器？换言之，如何以依赖于数据结构的方式，明确地用训练数据的数量来界定生成数据分布的估计误差?

关于得分函数近似误差的影响 现有文献已在关于得分函数近似精度的两种假设下分析了估计误差。(i) 一种常见的假设是，分数匹配中损失函数的误差足够小，该假设最早由 Song 等人 (2020年) 提出，利用 Girsanov 定理对连续时间动力学的 Kullback–Leibler（KL 散度）进行界定。近期，离散时间情形中出现了多项式界，这意味着每一步评分估计误差和步数的多项式阶数足以得到全变差（TV）距离下最终的估计误差（Lee 等人，2022年b）。Lee 等人 (2022年b) 假设真实密度满足光滑性和对数 Sobolev 不等式（LSI），而 Chen 等人 (2022年) 和 Lee 等人 (2022年a) 消除了 LSI 假设，但仍保留了光滑性假设。此外，继 Song 等人 (2020年) 之后，Pidstrigach (2022年) 考虑了定义在流形上的真实分布。(ii) 另一种假设是在每个时间点和空间点上界定得分函数与网络之间的差异。De Bortoli 等人 (2021年)（同时具有耗散性）和 DeBortoli (2022年)（在流形假设下）分别在全变差和一阶 Wasserstein 距离（W_1）下推导出了非多项式界。

[1]东京大学数学信息学系，日本东京 [2]理化学研究所先进智能项目中心，日本东京。通信作者：Kazusato Oko < oko- kazusato@g.ecc.u-tokyo.ac.jp>。

泛化误差分析 然而，大多数文献假设真实得分函数可得，因此，对于有限数量的样本能否恰当地近似得分函数尚未得到解决；由此，对上述假设在现实中是否成立的疑虑削弱了所得估计误差界的价值。作为唯一的例外，De Bortoli (2022年) 针对 $n^{-1/d}$ 中的 W_1 界，推导了适用于 n 数据和一个 d维分布的结果。然而，在其分析中，假设神经网络几乎完美地拟合经验得分函数，且估计界依赖于经验分布收敛到真实分布的收敛速率（Weed &Bach，2019年）。由于经验测度收敛具有相同的下界（Dudley，1969年），他们的 $n^{-1/d}$ 界本质上无法通过对数据分布施加任何结构假设而得到改进。因此，无法将其结果扩展到正式密度估计问题；在此类问题中，人们预期可获得依赖于真实密度光滑性的更快收敛速率。我们还提及 Block 等人 (2020年) 主要针对每个离散化步骤进行的泛化误差分析，但他们没有明确给出最终估计误差，且其间的界依赖于未知的拉德马赫复杂度；该复杂度应足够大，以使假设类能够很好地近似真实得分函数。

因此，扩散模型作为分布学习器的性能这一根本问题，在很大程度上仍未得到解答。

1.1. 我们的贡献

本文建立了扩散建模的统计学习理论。假设真实密度属于众所周知的函数空间，并采用深度神经网络作为估计器，我们推导了估计误差的收敛速率。令人惊讶的是，我们发现扩散建模能够达到近乎极小极大的估计速率。本文的贡献详述如下：

(i) 我们给出了利用神经网络对得分函数进行近似的显式形式，并推导了每个 t 下 $L^2(p_t)$ 的误差界，其中初始密度的支撑集包含于 $[-1,1]^d$，属于 Besov空间 $B^s_{p,\,q}([-1,1]^d)$，且在边界处光滑。

(ii) 我们将近似误差分析转化为估计误差界。我们推导了 $n^{-\frac{s}{d+2s}}$ 在全变差下的界。此外，通过对随机微积分的细致讨论，我们推导了任意固定 $\delta>0$ 下 $n^{-\frac{s+1-\delta}{d+2s}}$ 在 W_1 中的速率，该速率基于修正后的分数匹配。由此得到的估计速率近乎极小极大最优，从理论上证明了扩散模型的成功。

（iii）通过扩展我们的理论，我们还证明了在流形假设下，当真实数据分布在低维平面上时，扩散模型可以避免维数灾难。这是DeBortoli （2022年）中的一个特例，但在这种情况下，我们的界限要紧得多。

1.2. 其他相关工作

近期，多项研究考察了 Wasserstein 距离下的极小极大估计速率（经验分布（Weed & Bach, 2019年；Singh & P´oczos, 2018；Lei, 2020）；光滑密度（Liang, 2017年；Singh et al., 2018；Schreuder et al., 2021）；Besov空间（Niles-Weed & Berthet, 2022年））。Niles-Weed & Berthet (2022) 使用小波基处理 Besov空间，而 Liang (2017) 则受生成对抗网络（GAN）（Goodfellow et al., 2020）的启发，采用神经网络作为估计量。

我们想强调的是，我们的工作并非用神经网络取代 Niles-Weed & Berthet (2022) 的小波展开。在扩散建模中，我们首先最小化类平方误差分数匹配损失，然后考察估计误差。因此， W_1 中已有的紧界无法直接应用。与生成对抗网络的分析不同，在后者中，最终目标的极小极大问题直接关联于 W_1，扩散模型的分析需要将得分函数近似误差转化为估计误差。

我们所依据的主要是深度神经网络范数下的函数估计理论（ L^p 范数（Barron，1993；Yarotsky，2017年；Petersen& Voigtlaender，2018；Suzuki，2018；Schmidt-Hieber，2020；Hayakawa& Suzuki，2020）。我们的近似结果可视为对 Suzuki （2018）中所用 B 样条基展开的推广。另一方面，我们的泛化界依赖于 Schmidt-Hieber（2020）；Hayakawa & Suzuki （2020）。

2. 预备知识

扩散建模 我们基本沿用De Bortoli（2022年）的记号。$(B_t)_{[0,\overline{T}]}$和 $\beta_t:\ [0,\overline{T}]\to\mathbb{R}_+$ 分别表示d维布朗运动和一个加权函数。我们用 p_t 表示 X_t的分布，因此 p_0 即为数据分布。作为定义在 $(X_t)_{[0,\overline{T}]}$ 上的前向过程 $\mathbb{R}^d$，我们考虑以下 Ornstein–Ulhenbeck（OU）过程：

$$\mathrm{d}X_t=-\beta_t X_t\mathrm{d}t+\sqrt{2\beta_t}\mathrm{d}B_t,\quad X_0\sim p_0.$$

于是我们有 $X_t|X_0\sim\mathcal{N}(m_tX_0,\sigma_t)$，其中 $m_t=\exp(-\int_0^t\beta_s\,\mathrm{d}s)$ $\sigma_t^2=1-\exp(-2\int_0^t\beta_s\mathrm{d}s)$。注意到$1-m_t\simeq t\wedge 1$ 并且 $\sigma_t\simeq\sqrt{t}\wedge 1$。在关于 p_0 (Haussmann& Pardoux，1986）的温和假设下，这些结论适用于我们的情形，

反向过程 $(Y_t)_{[0,T]}$ 满足 $Y_t = X_{\overline{T}-t}$

$$\mathrm{d}Y_t = \beta_{\overline{T}-t}(Y_t + 2\nabla \log p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t,$$
$$Y_0 \sim p_{\overline{T}}.$$

$\nabla \log\ p_t(x)$ 被称为得分函数，它由使用有限样本训练的得分网络 $\hat{s}(x,t)$ 所替代。此外，由于 p_t 趋近于 $\mathcal{N}(0, I_d)$，我们取 $\overline{T} = \mathcal{O}(1)$，并将 Y_0 的初始噪声分布替换为 $\mathcal{N}(0, I_d)$。于是，修改后的反向过程 $(Y_t)_{[0,\overline{T}]}$ 定义为

$$\mathrm{d}\hat{Y}_t = \beta_{\overline{T}-t}(\hat{Y}_t + 2\hat{s}(\hat{Y}_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t,$$
$$\hat{Y}_0 \sim \mathcal{N}(0, I_d).$$

分数匹配 得分网络理想情况下从假设类 $\mathcal{S}$ 中选取，以最小化去噪分数匹配损失

$$\mathbb{E}_t\left[\lambda(t)\left[\mathbb{E}_{x_0}\left[\mathbb{E}_{x_t|x_0}[\|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2]\right]\right]\right], \quad (1)$$

其中 t ～均匀分布$[0,\overline{T}]$、$x_0 \sim p_0$、$x_t|x_0 \sim p_t(x_t|x_0)$，$\lambda$ 是一个加权函数。使用有限数据 $\{x_{0,i}\}_{i=1}^n$ ($x_{0,i} \overset{\text{i.i.d.}}{\sim} p_0$) 进行训练时，会选择 $\hat{s}$ 以最小化以下损失，该损失将 $\mathbb{E}_{x_0}$ 替换为样本均值：

$$\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\substack{t\sim \mathrm{Unif}[\underline{T},\overline{T}] \\ x_t \sim p_t(x_t|x_{0,i})}} [\lambda(t)\|s(x_t,t) - \nabla \log p_t(x_t|x_{0,i})\|^2]. \quad (2)$$

这里 $p_t(x_t|x_{0,i})$ 对应于 $\mathcal{N}(m_t X_{0,i}, \sigma_t)$，并且可以以任意精度计算该经验损失。我们将积分区间裁剪为 $\underline{T} > 0$，因为通常当 $t \to 0$ 时，得分函数会发散，而对于任意神经网络，(1) 都会 ∞。若无其他说明，我们令 $\lambda(t) \equiv 1$。

我们指出，关于 t 和 x_t的期望可以用 t 和 x_t的有限样本替代，具体内容将在第4.1节详述。然而，对于每个 $x_{0,i}$，我们不可避免地需要多项式数量的样本（t, x_t），或者对 t的分布进行人为修改，这主要是由于得分函数无界。

神经网络类 与神经网络近似领域的惯例一致（Yarotsky，2017年；Liang，2017年），分数匹配中的假设类 $\mathcal{S}$ 是一个带有 ReLU 激活函数的深度神经网络类 $\mathrm{ReLU}(x) = \max\{0, x\}$（对向量逐元素运算）（Nair & Hinton，2010年；Glorot 等，2011年），并施加稀疏性约束（限制非零参数的数量）。得分网络是一个从 $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$ 到 $y \in \mathbb{R}^d$ 的函数。

定义 2.1. 深度为 L、宽度为 W、稀疏性约束为 S、范数约束为 B 的神经网络类 $\Phi(L, W, S, B)$ 定义为 $\Phi(L, W, S, B) :=$ $\{(A^{(L)}\mathrm{ReLU}(\cdot) + b^{(L)}) \circ \cdots \circ (A^{(1)}x + b^{(1)}) |\ A^{(i)} \in \mathbb{R}^{W_i \times W_{i+1}}, b^{(i)} \in \mathbb{R}^{W_{i+1}}, \sum_{i=1}^{L}(\|A^{(i)}\|_0 + \|b^{(i)}\|_0) S$ ≤ , 最大值 $_i \|A^{(i)}\|_\infty \vee \|b^{(i)}\|_\infty \le B\}$.

我们指出，关于全连接神经网络（FNN）的结果很容易推广到其他架构。例如，实践中常用的 U-Net 变体（Ronneberger 等人，2015年）（Song& Ermon，2019年；Ho 等人，2020；Ramesh 等人，2022年）属于卷积神经网络（CNN）。我们可以利用丰富的文献，将 FNN 的近似结果转换为 CNN 的近似结果（Oono 和 Suzuki，2019；Zhou，2020；Petersen &Voigtlaender，2020）。

Besov空间中的密度估计 作为真实密度的一类函数，Besov空间通过光滑模量引入。我们假设 Ω 是 $\mathbb{R}^d$ 中的一个立方体。

定义 2.2. 对于某个 $p \in (0, \infty]$ 的函数 $f \in L^p(\Omega)$，f 的 r 阶光滑模量定义为

$$w_{r,p}(f,t) = \sup_{\|h\|_2 \le t} \|\Delta_h^r(f)\|_p, \quad \text{where } \Delta_h^r(f)(x)$$
$$= \begin{cases} \sum_{j=0}^{r} \binom{r}{j}(-1)^{r-j} f(x + jh) & (\text{if } x + jh \in \Omega \text{ for all } j) \\ 0 & (\text{otherwise}). \end{cases}$$

定义 2.3（Besov空间 $B^s_{p,q}(\Omega)$）。对于 $0 < p, q \le \infty$，$s > 0, r := \lfloor s \rfloor + 1$，令半范数 $|\cdot|_{B^s_{p,q}}$ 为

$$|f|_{B^s_{p,q}} = \begin{cases} \left(\int_0^\infty (t^{-s} w_{r,p}(f,t))^q \frac{dt}{t}\right)^{\frac{1}{q}} & (q < \infty), \\ \sup_{t>0} t^{-s} w_{r,p}(f,t) & (q = \infty). \end{cases}$$

Besov空间的范数 $B^s_{p,q}$ 定义为 $\|f\|_{B^s_{p,q}} = \|f\|_p + |f|_{B^s_{p,q}}$ $B^s_{p,q} = \{f \in$., 并且有 , $L^p(\Omega)|\ \|f\|_{B^s_{p,q}} < \infty\}$.

考虑 Besov空间，许多众所周知的函数类都可以用统一的方式进行讨论。下面举几个例子。对于 $\alpha \in \mathbb{Z}_+^d$，令 $\partial^\alpha = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(x)$。对于 $s \in \mathbb{R}_{>0} \setminus \mathbb{Z}_+$，H¨older 空间是由 $\lfloor s \rfloor$ 次可微函数$\mathcal{C}^s(\Omega) = \{f$组成的集合：$\Omega \to \mathbb{R}|\ \|f\|_{\mathcal{C}^s} := \max_{|\alpha| \le s} \|\partial^\alpha f\|_\infty + \max_{|\alpha| = \lfloor s \rfloor} \sup_{x,y \in \Omega} \{\frac{\|\partial^\alpha f(x) - \partial^\alpha f(y)\|}{\|x-y\|^{s-\lfloor s \rfloor}} < \infty\}$) ( ), 其中 $s \in \mathbb{R}_{>0} \setminus \mathbb{Z}_+$。对于 $s \in \mathbb{N}$，$1 \le p \le \infty$，Sobolev 空间是由 s次可微函数 $W_p^s(\Omega) := \{f: \Omega \to \mathbb{R}|\ \|f\|_{W_p^s} := (\sum_{|\alpha| \le s} \|\partial^\alpha f\|_p^p)^{\frac{1}{p}} < \infty\}$组成的集合。随后，Amann 等人（1983）给出了以下关系：

- For $s \in \mathbb{N}$, $B^s_{p,1}(\Omega) \hookrightarrow W_p^s(\Omega) \hookrightarrow B^s_{p,\infty}(\Omega)$.
- $B^s_{2,2}(\Omega) = W_2^s(\Omega)$.
- For $s \in \mathbb{R}_{>0} \setminus \mathbb{Z}_+$, $\mathcal{C}^s(\Omega) = B^s_{\infty,\infty}(\Omega)$.

如果 $s > d/p$， $B^s_{p,q}(\Omega)$ 连续嵌入连续函数的集合中。否则，空间中的元素将不再连续。我们的结果对于$B^s_{p,q}(\Omega)$ 且 $s > d(1/p - 1/2)+$成立，因此可以包含不连续函数，这不同于现有的假设 Lipschitz 连续性的界（Lee 等人、2022b；a； Chen 等人、2022年）。

在这一问题设定中，我们评估分布$\hat{}$与真实分布 p_0之间的接近程度。作为分布估计器的性能度量，我们采用全变差距离 (TV) 和一阶 Wasserstein 距离(W_1)。在第 6 节中，假设数据位于低维流形上，我们重点考察 Wasserstein 距离。这是因为生成分布相对于真实分布绝不可能绝对连续，因此，Wasserstein 距离对分布微小平移的鲁棒性对于获得非平凡的界至关重要。∞。

2.1. 假设

这里我们正式陈述最小假设。设 d为该空间的维数， n 为样本数，$0 < p$、 $q \le \infty$、 $s > 0$ 以及 $s > (1/p - 1/2)+$ 为 Besov空间的参数。我们的主要假设如下。

假设2.4。真实密度 p_0 的支撑集为$[-1,1]^d$，在该支撑集上的上界和下界分别为 C_f 和 C_f^{-1}。此外，当 p_0 限制在 $[-1,1]^d$ 上时，属于 $U(B^s_{p,q}([-1,1]^d); C)$，其中某个常数为 C。

$U(\cdot; C)$ 表示半径为 C 的球，有时省略常数 C，写作 $U(\cdot)$。此外，我们还作出两个技术性假设。其一是 β_t 的光滑性。

假设 2.5。 $\beta_\cdot : [0,\overline{T}] \to \mathbb{R}_+ (t \mapsto \beta_t)$ 满足 $0 < \underline{\beta} \le \beta_\cdot \le \overline{\beta}$，且 $\beta_\cdot \in U(\mathcal{C}^\infty([0,\overline{T}]); 1)$ 是关于 $_t \in [0,T]$ 的函数。

另一个假设是真实密度 p_0 在边界区域上的光滑性。令 a_0 为之后定义的一个充分小的值，例如， $a_0 \approx n^{-\frac{1}{d+2s}}$ 在定理4.3中。

假设 2.6。当 p_0 限制于 $[-1,1]^d \setminus [-1+a_0, 1-a_0]^d$ 时，属于 $U(\mathcal{C}^\infty([-1,1]^d \setminus [-1+a_0, 1-a_0]^d))$。

这是为了在p_t 没有正下界的区域构造得分网络。这是必要的，因为在密度估计中通常会假设密度有正下界（Tsybakov，2009年），而没有正下界时，极小极大最优速率有时会变得更差（Niles-Weed &Berthet，2022年）。这一假设可以由密度衰减足够缓慢来替代，例如 Lee 等人 （2022b）中使用的对数 Sobolev 不等式（LSI）。我们还指出，这一修改不会损害极小极大速率。

3. 真实分数的近似

本节中，我们考虑通过深度神经网络逼近真实得分函数$\nabla \log p_t$，并推导逼近误差界。在本节中，我们固定 $\delta > 0$为任意值，并将 $N \gg 1$ 作为决定网络规模的参数。我们假设条件2.6 成立，其中$a_0 = N^{-\frac{1-\delta}{d}}$，并取 $\underline{T} = \mathrm{poly}(N^{-1})$，且 $\overline{T} \simeq \log N$。本节的主要贡献如下。

定理 3.1。存在一个神经网络 $\phi_{\mathrm{score}} \in \Phi(L, W, S, B)$，使得对所有 $t \in [\underline{T}, \overline{T}]$，

$$\int_x p_t(x) \|\phi_{\mathrm{score}}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d}} \log(N)}{\sigma_t^2}.$$

这里，L、 W、 S 和 B 可取为 $L = \mathcal{O}(\log^4 N)$ $\|W\|_\infty = \mathcal{O}(N \log^6 N)$ $S = \mathcal{O}(N \log^8 N)$，以及 $B = \exp(\mathcal{O}(\log^4 N))$。此外，我们可以取满足 $\|\phi_{\mathrm{score}}(\cdot, t)\|_\infty = \mathcal{O}(\sigma_t^{-1} \log^{\frac{1}{2}} N)$ 的 ϕ_{score}。

形式化证明见附录B。

3.1. 证明概述

为得到这一结果，近似应按以下方式构造。(i) 它应体现 $p_0(x)$ 的结构，尤其是 $p_0(x) \in U(B^s_{p,q})$ 这一事实。(ii) 它应在所有 $t \in [\underline{T}, \overline{T}]$ 上对得分函数作出良好的近似。为解决这些问题，我们在 $\mathbb{R}^d \times [\underline{T}, \overline{T}]$ 空间中构造了一种新颖的基分解，专门用于得分函数近似。此外，与逼近理论中的通常做法一样（Yarotsky，2017年；Schmidt-Hieber，2020），每个基都可以由神经网络高效实现，也就是说，相对于允许的误差，只需使用多对数规模的网络即可。

通过扩散 B 样条基进行近似 我们考虑对 $t \ll 1$ 的近似。首先回顾 Besov 函数的 B 样条基分解（DeVore & Popov，1988；Suzuki，2018）。令 $\mathcal{N}(x) = 1(x \in [0,1])$，0 (其他情况)。 阶数为 l 的基数 B 样条定义为 $\mathcal{N}_l(x) = \underbrace{\mathcal{N} * \mathcal{N} * \cdots * \mathcal{N}}_{l+1\text{ 次卷积}}(x)$，其中$(f * g)(x) = \int f(x-t) g(t) \mathrm{d}t$。然后， $\mathbb{R}^d$中的张量积 B 样条基针对 $k \in \mathbb{N}^d$和 $j \in \mathbb{Z}^d$定义为$M^d_{k,j}(x) = \prod_{i=1}^d \mathcal{N}(2^{k_i} x - j_i)$。已知，Besov空间中的函数$f$ 可由$M^d_{k,j}(x)$的线性叠加近似，其中$f_N = \sum_{(k,j)} \alpha_{(k,j)} M^d_{k,j}(x)$。

引理3.2（引理F.13的非正式版本；Suzuki（2018））。对于任意 $p_0 \in U(B^s_{p,q})$，存在由 N 个张量积B样条基构成的超位置 f_N，满足

$$\|p_0 - f_N\|_{L^2} \lesssim N^{-s/d} \|f\|_{B^s_{p,q}}.$$

受此启发，我们引入基分解。由于 $X_t|X_0 \sim \mathcal{N}(m_t X_0, \sigma_t)$，我们可以将 p_t 写为

$$p_t(x) = \int p_0(y) \underbrace{\frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)}_{=:K_t(x|y)} \mathrm{d}y.$$

由于转移核 $K_t(x|y)$ 对p_0 呈线性作用，且 p_0 可由 $f_N = \sum_{(k,j)} \alpha_{(k,j)} M^d_{k,j}(x)$ 近似，我们得到 p_t的如下近似：

$$p_t(x) \approx \sum_{(k,j)} \alpha_{(k,j)} \underbrace{\int M^d_{k,j}(y) K(x|y) \mathrm{d}y}_{=:E_{k,j}(x,t)}.$$

此外，$E_{k,j}$ 还可进一步分解为

$$E_{k,j}(x,t) = \prod_{i=1}^{d} \underbrace{\int \frac{\mathcal{N}(2^{k_i} x_i - j_i)}{\sigma_t \sqrt{2\pi}} \exp(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}) \mathrm{d}x_i}_{=:\mathcal{D}_{k,j}(x_i,t)}.$$

我们将$\mathcal{D}_{k,j}$ 称为扩散 B 样条基，将$E_{k,j}$ 称为张量积扩散B样条基。我们证明，存在一个神经网络，能够非常高效地近似$\mathcal{D}_{k,j}$ 和$E_{k,j}$。接下来，我们按如下方式进行构造：构造近似 m_t 和 σ_t的网络。

引理 3.3 (另见引理 B.1)。 在假设 2.6下，存在神经网络 $\phi_m(t)$ $\phi_\sigma(t) \in \Phi(L, W, B, S)$，可将 m_t 和 σ_t 近似到 ε，对所有 $t \geq 0$均成立，其中 $L = \mathcal{O}(\log^2(\varepsilon^{-1}))$ $\|W\|_\infty = \mathcal{O}(\varepsilon^{-1})$ $S = \mathcal{O}(\log^4(\varepsilon^{-1}))$ $B = \mathcal{O}(\varepsilon^{-2})$ (log())(log())(log())，且 exp((log()))。

接下来，我们截断$\mathcal{D}_{k的j}$ 积分区间，并用关于(x, m_t, σ_t)的有理函数近似被积函数。于是可得到引理B.3的非正式表述。

引理 3.4. 对于 $\varepsilon > 0$，存在一个神经网络ϕ_{TDB}：$\mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$，满足 $\|\phi_{\mathrm{TDB}}(x,t) - E_{k,j}(x,t)\|_\infty \leq \varepsilon$. 其中，$\phi_{\mathrm{TDB}} \in \Phi(L, W, S, B)$，且 $L = \mathcal{O}(\log^4(\varepsilon^{-1}))$ $\|W\|_\infty = \mathcal{O}(\log^6(\varepsilon^{-1}))$ $S = \mathcal{O}(\log^8(\varepsilon^{-1}))$ $B = \mathcal{O}(\exp(\mathcal{O}(\log^2(\varepsilon^{-1}))))$。

其中，ϕ_{TDB} 近似于给定 (x, m_t, σ_t) 的 $E_{k,j}(x,t)$。然后，我们使用 $\phi_{\mathrm{TDB}}(x, \phi_m(t)\ \phi_\sigma(t))$ 作为 $E_{k,j}(x,t)$ 的近似，并最终对 $p_t(x)$ 进行近似。类似的近似也可用于 $\nabla p_t(x)$，最后将得分函数与 $\nabla \log p_t(x) = \nabla p_t(x) / p_t(x)$ 一起进行近似，从而得到定理 3.1 中的界。

我们要指出的是，上述网络规模略大于 B样条基的规模（Suzuki （2018））， 因为近似积分和指数函数（附录 F.3）以及有理函数（附录F.2）比通过多项式实现 B样条基更加困难。

我们还要指出，在这一构造中，$\nabla p_t(x)$ 的近似误差在 $p_t(x) \ll 1$所在的区域会被放大。这就是为什么我们需要 p_0 在距离支撑集边缘小于 $\tilde{\mathcal{O}}(\sqrt{t})$ 的区域内具有高阶光滑性（假设 2.6）。该方法用于 $t \in [\underline{T}, 3N^{-\frac{2-\delta}{d}}]$期间，只需令 a_0为 $a_0 = N^{-\frac{1-\delta}{d}}$即可。

利用噪声诱导的光滑性 上述方法利用 p_0 的结构，在得分函数高度不光滑时，近似 $t \ll 1$ 中的得分函数。另一方面，经过一段时间后，p_t 的形状几乎变得像高斯分布一样，十分光滑且易于近似。本段扩展了前述方法，并基于噪声诱导的光滑性提出另一种近似方法，从而得到更紧的界。

我们首先计算 p_t 关于 t 的导数。

引理 3.5. 对任意 $k \in \mathbb{Z}_+$，存在一个常数 C_{a}，它仅依赖于 k、d 和 C_f，使得

$$\left|\partial_{x_{i_1}} \partial_{x_{i_2}} \cdots \partial_{x_{i_k}} p_t(x)\right| \leq \frac{C_{\mathrm{a}}}{\sigma_t^k}.$$

由此可得，对于 $t_* > 0$，有 $\|p_{t_*}\|_{W^k_p} = \mathcal{O}(t_*^{-\frac{k}{2}})$，且 $W^k_p \hookrightarrow B^k_p$、$\infty$。对于 $t > t_*$，将 p_t 视为由 p_{t_*} 得到的扩散分布，而非 p_0。我们可以证明，$\nabla \log p_t$ 可由一个规模为 N' 的神经网络近似，误差为 $\mathcal{O}\left(\frac{N'^{-2k/d}}{\sigma_t^2} \cdot t_*^{-k}\right)$ 的 L^2。如果 N' 和 k 足够大，这比先前的 $_{2s}$ 界 $\frac{N^{-\overline{d}}}{\sigma_t^2}$ 更紧。以下将该论证形式化。在附录 B 中，该结论以引理 B.7 的形式给出。

引理 3.6. 设 $N \gg 1$和 $N' \geq t_*^{-d/2} N^{\delta/2}$。假设 $t_* \geq N^{-(2-\delta)/d}$。则存在一个神经网络$\phi'_{\mathrm{score}} \in \Phi(L, W, S, B)$，满足

$$\int_x p_t(x) \|\phi'_{\mathrm{score}}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^2}$$

对于 $t \in [2t_*, \overline{T}]$ $L = \mathcal{O}(\log^4(N))$ $\|W\|_\infty = \mathcal{O}(N)$ $S = \mathcal{O}(N')$ $B = \mathcal{O}(N^4)$。具体而言，(log())()()，且 exp((log()))。

在引理中设定 $t_* = N^{-\frac{2-\delta}{d}}$和 $N' = N$，我们便可在不作假设 3.1 的情况下，得到定理 $t \gtrsim t_*$中的界。此外，稍后进一步利用此引理，对于在2.6距离下实现极小极大最优估计率发挥重要作用。 W_1。

4. 得分网络的泛化

本节将定理 3.1 转化为得分网络的泛化界。我们假设 $n \gg 1$ 以及假设 2.6 满足 $a_0 = n^{-\frac{1-\delta}{d+2s}}$，并取 $N =$

$n^{-d/(d+2s)}$，$\underline{T}$ = 多项式(N^{-1}) = 多项式(n^{-1})，且 $\overline{T}$ ≃对数 N ≃ 对数 n。形式化证明见附录 C。我们先给出以下事实（引理C.3；Vincent(2011年）。

引理 4.1. 对所有 $s(x,t)$ 和 $t>0$，以下结论成立：

$$\int_x\int_y \|s(x,t)-\nabla\log p_t(x|y)\|^2 p_t(x|y)p_0(x)\mathrm{d}y\mathrm{d}x$$
$$=\int_x \|s(x,t)-\nabla\log p_t(x)\|^2 p_t(x)\mathrm{d}x + C_t.$$

这里 C_t 是一个取决于 p_t 的常数。根据这一点，最小化（1）等价于最小化网络与得分函数在 $L^2(p_t)$ 之间的差异。

我们定义

$$\ell_s(x)=\int_{t=\underline{T}}^{\overline{T}}\int \|s(x_t,t)-\nabla\log p_t(x_t|x)\|^2 p_t(x_t|x)\mathrm{d}x_t\mathrm{d}t,$$

从而，期望损失（1）和经验损失分别写作 $\mathbb{E}_{x\sim p_0}[\ell(x)]$ 和 $\frac{1}{n}\sum_{i=1}^n \hat{}\ell(x_i)$。对于稍后指定的假设类 $\mathcal{S}$，我们定义 $\mathcal{L}=\{\ell_s|\ s\in\mathcal{S}\}$。定义经验损失最小化解 $\hat{s}\in \arg\min_{s\in\mathcal{S}}\frac{1}{n}\sum_i \ell_s(x_{0,i})$。本节评估 $\hat{s}$ 的经验损失与（1）之间的差异。为评估这一差异，我们需要界定：（i）在 $\mathcal{L}$ 上一致的 $\|\ell\|_\infty$，以及（ii） $\mathcal{L}$ 的覆盖数。

（i）上确界范数的界 根据定理 3.1，$_1\hat{s}(x,\ t)$ 可以取为使得 $\|\hat{s}(\cdot,\ t)\|_\infty \lesssim \log^{\overline{T}}_{\sigma_t} N$。因此，我们将定理 Φ（L，W，S，B）3.1 的限制为

$$\mathcal{S}:=\{\phi\in\Phi(L,W,S,B)|\ \|\phi(\cdot,t)\|_\infty \lesssim \frac{\log^{\frac{1}{2}} n}{\sigma_t}\}.$$

接着，附录C.1 表明，

$$\sup_{s\in\Phi'}\sup_{x_0\in[-1,1]^d}\ell_s(x_0)\lesssim \log^2 n.$$

（ii）覆盖数的计算 根据Suzuki（2018）中的引理 3，以及 $\|\ell_s\|_\infty$ 至多被 $\|s\|_\infty$乘以 poly（n）所界定这一事实，我们得到以下结果。

引理 4.2。 $\mathcal{L}$ 的覆盖数可由下式计算：

$$\log\mathcal{N}(\mathcal{L},\|\cdot\|_{L^\infty([-1,1]^d)},\delta)\lesssim SL\log(\delta^{-1}L\|W\|_\infty Bn).$$

证明见附录 C.2。将其应用于定理 3.1，覆盖数有如下上界：$\log\mathcal{N}\lesssim N(\log^{16}N+\log^{12}N\log\delta^{-1})$。

根据上述讨论，我们最终得到泛化界。下一个界是对 Schmidt-Hieber (2020)；Hayakawa& Suzuki (2020) 的扩展。他们考虑的是均方损失的最小化器，而我们考虑的是 $\ell(x_i)$ 的均值的最小化器。

定理 4.3。从 Φ' 中选出的经验得分函数最小化器满足

$$\mathbb{E}_{x_i}\left[\int_x\int_{t=\underline{T}}^{\overline{T}}\|\hat{s}(x,t)-\nabla\log p_t(x)\|^2 p_t(x)\mathrm{d}t\mathrm{d}x\right] \quad (3)$$
$$\lesssim \inf_{s\in\Phi'}\int_x\int_{\underline{T}}^{\overline{T}}\|s(x,t)-\nabla\log p_t(x)\|_2^2 p_t(x)\mathrm{d}x\mathrm{d}t$$
$$+\frac{\sup_{s\in\mathcal{S}}\|\ell_s\|_\infty\log\mathcal{N}}{n}+\delta.$$

应用 $\sup_{\ell\in\Phi'}\|\ell\|_\infty \lesssim N^{16}N+N^{12}\delta^{-1}$ 和 $N=\delta=n^{-d/(2s+d)}$ log 和 log（log log log）并结合，得到

$$(3)\lesssim n^{-\frac{2s}{d+2s}}\log^{18} n. \quad (4)$$

4.1. 采样 t 和 x_t，而非取期望

由于我们主要关注样本复杂度，为了简洁起见，我们考虑了 $\ell(x)$ 可以精确计算的情况。然而，在通常的实现中，Sohl-Dickstein 等人(2015)；Song & Ermon(2019)也会出于计算效率的考虑，将式（2）中关于 t 和 x_t的两个期望替换为采样。在这里，我们也介绍两种方法，用 t 和 x_t的有限样本替代期望。

通过多项式规模样本进行近似。让我们从 i_j ∼均匀分布({1, 2, ⋯ , n})、 t_j ∼均匀分布(T, T) 和 $x_j\sim p_{t_j}(x_j|x_{0,i})$ 中采样 (i_j,t_j,x_j)。然后我们将 $\hat{s}$ 定义为

$$\operatorname*{argmin}_{s\in\mathcal{S}}\frac{1}{M}\sum_{j=1}^M\|s(x_j,t_j)-\nabla\log p_{t_j}(x_j|x_{0,i_j})\|^2$$

并计算以下两者之间的差值

$$\frac{1}{n}\sum_{i=1}^n\ell_{\hat{s}}(x_i)-\operatorname*{argmin}_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^n\ell_s(x_i). \quad (5)$$

完整证明和正式陈述见附录 C.4 的定理 C.6，本文在此给出证明概要。我们首先证明$\|s(x_j,t_j)-\nabla\log p_{t_j}(x_j|x_{0,i_j})\|$ 是次高斯的（引理 C.5）。这里，为继续推导，我们只需将其理解为$\|s(x_j,t_j)-\nabla\log p_{t_j}(x_j|x_{0,i_j})\| = \tilde{\mathcal{O}}(t_j^{-\frac{1}{2}}) \lesssim \tilde{\mathcal{O}}(\underline{T}^{-\frac{1}{2}})$以高概率成立。随后，通过与推导定理 4.3 时类似的论证，我们可以将 (5) 上界为 $\tilde{\mathcal{O}}(\underline{T}^{-1}\cdot\log_M\mathcal{N})$。其中，$\mathcal{N}$满足 $\log\mathcal{N}\lesssim n^{\frac{d}{2s+d}}\log^8 n$。为了使 (5) 与 (4) 一样小，我们需要取 $M\gtrsim n\cdot\underline{T}^{-1}$。因此，对于每个$x_0$，应考虑 $(t_j,x_j|x_{0,i})$ 的, $\mathcal{O}(\underline{T}^{-1})$ =poly(n^{-1}) 样本。我们指出，需要多项式规模样本的主要原因在于$\|s(x_j,t_j)-\nabla\log p_{t_j}(x_j|x_{0,i_j})\|^2$的尺度。

修改 t 的分布。有人可能会认为，是否可以只为每个样本 $x_{0,i}$ 考虑一条路径。这里的主要问题是，随着 t_j 趋近于 0，$\|s(x_j,t_j)-\nabla \log p_{t_j}(x_j|x_{0,i_j})\|^2$ 的方差可能增长至无穷大。为解决这一问题，我们从 $\mu(t) \propto \frac{\mathbb{1}[\underline{T}\le t\le \overline{T}]}{t}$ 中采样 t_j，并将 $\lambda(t)$ 修改为 $\lambda(t) = t\log \frac{\overline{T}}{\overline{T}-\underline{T}}$，而 i_j, x_j 仍按此前的方式采样。于是，我们有

$$\mathbb{E}_{i_j,t_j,x_j}\big[\lambda(t_j)\|s(x_j,t_j)-\nabla \log p_{t_j}(x_j|x_{0,i})\|^2\big] = \frac{1}{n}\sum_{i=1}^{n}\ell(x_i),$$

并且 $\lambda(t_j)\|s(x_{t_j},t_j) - \nabla \log p_{t_i}(x_{t_j}|x_{0,i})\|^2 = \mathcal{O}(1)$ 以高概率成立（因为 $\|s(x_j,t_j) -\nabla \log p_{t_j}(x_j|x_{0,i_j})\|^3 = \tilde{\mathcal{O}}(t_j^{-1})$，且 $\lambda(t_j) \lesssim 1/t_j$）。在这种采样方式中，我们令 $\hat{s}$ 为

$$\operatorname*{argmin}_{s\in\mathcal{S}}\frac{1}{M}\sum_{j=1}^{M}\lambda(t_j)\|s(x_j,t_j) - \nabla \log p_{t_j}(x_j|x_{0,i_j})\|^2$$

并估计差值（5）。最后，使用定理 4.3 的类似论证，我们再次得到（5）被 $\tilde{O}(\frac{\log \mathcal{N}}{M}) \lesssim \tilde{O}(\frac{n^{\frac{d}{2s+d}}}{M})$ 所界定。取 $M = n$ 即可使该差值小到（4）。

5. 估计误差分析

本节最后评估扩散建模作为密度估计器的效果。稍作修改，若 $\|Y_{\overline{T}-\underline{T}}\|_\infty \ge 2$，则将其重置为 $Y_{\overline{T}-\underline{T}} = 0$。这不会增加估计误差，因为 $\|X_0\|_\infty \le 1$ 几乎必然成立。我们引入 $(\bar{Y}_t)_{t=0}^{\overline{T}-\underline{T}}$，用 $\bar{Y}_0 \sim p_t$ 替换 $(\hat{Y}_t)_{t=0}^{\overline{T}-\underline{T}}$ 的定义中的 $\hat{Y}_0 \sim \mathcal{N}(0, I_d)$。

5.1. 全变差下的估计速率

首先，我们以与 Song &Ermon (2019年)；Chen 等人 (2022年) 相同的方式，考虑全变差距离下的界。形式化证明见附录 D.2。全变差下的估计误差分解为

$$\mathbb{E}[\mathrm{TV}(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim \mathbb{E}[\mathrm{TV}(X_0, X_{\underline{T}})] + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d))] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})].$$

第一项来自对反向过程的截断，其上界为 $\sqrt{\underline{T}}n^{\mathcal{O}(1)}$，根据定理 D.2。第二项对应于对前向过程的截断，或 $Y_{\overline{T}-\underline{T}}$ 与 $Y_{\overline{T}-\underline{T}}$ 之间的差异，其上界为 $\exp(-\overline{T})$，这是由引理 D.3得出的。对于最后一项，经过一定修改的吉尔萨诺夫定理（命题D.1）可得

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\sqrt{\left[\int_{t=\underline{T}}^{\overline{T}}\mathbb{E}_{x\sim p_t}\left[\|\hat{s}(x,t)-\nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right]}. \tag{6}$$

利用 $\sqrt{\ }$ 的凸性以及得分网络的泛化界（4）可得（6）$\lesssim n^{-\frac{s}{d+2s}}\log^{\frac{5d+8s}{2d}} n$。现在，我们将估计误差界形式化。

定理 5.1. 设 $\underline{T} = n^{-\mathcal{O}(1)}$ 且 $\overline{T} = s\log\frac{n}{\underline{\beta}\ d+2s()}$。则，

$$\mathbb{E}[\mathrm{TV}(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-s/(2s+d)}\log^{\frac{5d+8s}{2d}} n.$$

另一方面，我们可以证明，Besov空间中的估计问题具有如下下界。证明见命题 D.4。

命题 5.2. 对于 $0 < p、 q \le \infty、 s > 0$ 和 $s > \max\{d(\frac{1}{p} - \frac{1}{2}) 0\}$，有

$$\inf_{\hat{\mu}}\sup_{p\in B^s_{p,q}}\mathbb{E}[\mathrm{TV}(\hat{\mu}, p)] \gtrsim n^{-s/(2s+d)},$$

其中，$\hat{\mu}$ 遍历所有基于 n 个观测的估计量。

因此，我们证明了扩散建模在全变差距离下，对Besov空间 $B^s_{p,\ q}$ 可达到极小极大估计率，误差仅相差一个对数因子。

5.2. W_1中的估计速率

我们还考虑 W_1中的估计率。由于 $\hat{X}_0$ 和 $Y_{\overline{T}-\underline{T}}$ 的支撑集均有界，定理 5.1 可直接给出 $n^{-s/(2s+d)}\log^{\frac{5d+8s}{2d}} n.$ 的收敛速率。然而，Niles-Weed& Berthet (2022年的研究表明，W_1 中的极小极大估计率比这一速率更快。

命题 5.3 (Niles-Weed &Berthet (2022年))。 设 $p,\ q \ge 1,$ $s > 0$ 和 $d \ge 2$。

$$\int_{\hat{\mu}}\sup_{p\in B^s_{p,q}}\mathbb{E}[W_1(\hat{\mu}, p)] \gtrsim n^{-(s+1)/(2s+d)},$$

其中，$\hat{\mu}$ 遍历所有基于 n 个观测的估计量。此外，若 $1 \le p < \infty、 1 \le q \le \infty、 s > 0$和 $d \ge 3$，则存在一个估计量$\hat{\mu}_*$，可达到这一极小极大速率。

那么，在这种情况下，扩散模型是否是次优的？接下来，我们将展示一个令人惊讶的事实：只要进行一些修改，扩散模型训练也能达到这一近乎极小极大最优的速率。

定理 5.4. 对任意固定的 $\delta > 0$，我们可以用 $n(\gg 1)$ 个样本训练得分网络，并且有

$$\mathbb{E}[W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-\frac{(s+1-\delta)}{d+2s}}.$$

证明见附录D.2。

切换得分网络现在我们简要介绍我们的策略。首先，让我们仔细考察一下估计率损失在何处，并回到近似误差分析第 3 节。尽管对于所有 $\underline{T} \le t \le \overline{T}$ 我们都使用了定理 3.1，但引理 3.6指出，如果 $t \gtrsim N^{-\frac{2-\delta}{d}} \simeq n^{-\frac{2-\delta}{2s+d}}$，我们可以使用规模为 $N' \le N$ 的更小网络，使近似误差小于 $\frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^{-2}} = \overline{n^{-\frac{2(s+1)}{d+2s}}\sigma_t^{-2}}$。这意味着，就 $t \gtrsim_{2-\delta} n^{-\frac{2}{d+2s}}$ 而言，我们在近似误差和泛化误差两方面都使用了次优网络。

基于上述讨论，我们将时间划分为 $t_0 = \underline{T} < t_1 = 2n^{-\frac{2-\delta}{d+2s}} < \cdots < t_{K_*} = \overline{T} - \underline{T}$，其中 $t_{i+1}/t_i =$ 为常数。 ≤ 2 $(i \ge 1)$。区间数量为$K_* = \mathcal{O}(\log n)$。我们考虑为每个时间区间 $[t_i, t_{i+1}]$ 训练一个定制网络，并在不同区间切换使用。引理3.6 表明，对于 $i \ge 1$，存在一个网络 $s_i \in \Phi(L_i, W_i, S_i, W_i)$ 使得

$$\mathbb{E}_{x\sim p_t}[\|s_i(x,t) - \nabla \log p_t(x)\|^2] \lesssim \frac{n^{-\frac{2(s+1)}{d+2s}}}{\sigma_t^2} \quad (t \in [t_i, t_{i+1}]),$$

满足 $L = \mathcal{O}(\log^4(N))$ $\|W\|_\infty = \mathcal{O}(N)$ $S = \mathcal{O}(t_i^{-d/2}N^{\delta/2})$，且 $B = \exp(\mathcal{O}(\log^4 N))$。因此，我们选择一列得分网络$\hat{s}_i$，使得$\hat{s}_i$ 在限制于 $[t_i, t_{i+1}]$时最小化得分匹配损失:

$$\frac{1}{n}\sum_{j=1}^{n} \mathop{\mathbb{E}}_{\substack{t\sim \mathrm{Unif}[t_i, t_{j+1}] \\ x_t \sim p_t(x_t|x_{0,j})}} [\|s(x_t,t) - \nabla \log p_t(x_t|x_{0,j})\|^2].$$

与定理 4.3类似，定理C.4 给出了以下关于 $i \ge 1$ 的泛化误差界:

$$\mathbb{E}_{\{x_{0,j}\}_{i=j}^n}\left[\int_{t=t_i}^{t_{i+1}} \mathbb{E}_x\left[\|\hat{s}_i(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right] \quad (7)$$
$$\le \left(n^{-\frac{2(s+1)}{d+2s}} + \frac{t_i^{-d/2} n^{\frac{\delta d}{(d+2s)}} \log^{10} n}{n}\right) \cdot \underbrace{\tilde{\mathcal{O}}(t_i/\sigma_{t_i}^2)}_{=\tilde{\mathcal{O}}(1)}.$$

对于 $t \lesssim n^{-\frac{2-\delta}{d+2s}}$，我们使用通过得分匹配损失训练、并将其限制在 $[t_i, t_{i+1}]$上的网络。因此，关于 $i = 0$ 的（7）可类似于（4），由 $\tilde{\mathcal{O}}(n^{-\frac{2s}{d+2s}})$ 界定。人们可能会认为，上述改进毫无$^{2-\delta}$用，因为在 $t \le n^{-\frac{2}{d+2s}}$ 处产生的误差具有$n^{-2s/(d+2s)}$ 速率，并且会主导估计误差。然而，另一个重要的观察是，Wasserstein 距离是一种运输距离。在更接近 $t = 0$ 的时刻，分数估计误差对估 t 计误差的贡献更小，因为从该时刻开始，每条路径的演化幅度都很小。正如我们将看到的，提高较大 t 处的精度这一想法，确实能在 W_1中达到极小极大最优速率。

为利用这一观察，我们来考虑一个随机过程序列。令 $(Y_t)_{[0,\overline{T}]} = (\bar{Y}_{(0)\,t})_{[0,\overline{T}]}$，对于$i \ge 1$，令 $(\bar{Y}_{(i)})_{[0,\overline{T}]}$ 为一个随机过程：它在 $[0, \overline{T} - t_i]$ 期间使用真实分数，并在 $[\overline{T} - t_i, \overline{T} - \underline{T}]$期间使用估计得分$\hat{s}$，且 $\bar{Y}_{(i)0} \sim p_{\overline{T}}$。于是，我们有

$$\mathbb{E}[W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}})] \le \mathbb{E}[W_1(X_0, X_{\underline{T}})] \quad (8)$$
$$+ \mathbb{E}[W_1(\hat{Y}_{\overline{T}-\underline{T}}, \bar{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})].$$

根据（91），第一项由 $\sqrt{\underline{T}}$ 界定；根据引理 D.6，第二项由 $\exp(-\overline{T})$ 界定。最后一项 $\mathbb{E}[W_1(Y_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})]$的上界为$\sum_{i=1}^{K_*} \mathbb{E}[W_1(\bar{Y}_{(i-1)\overline{T}-\underline{T}}, \bar{Y}_{(i)\overline{T}-\underline{T}})]$. 接下来，我们使用以下引理，即引理 D.7的非正式版本。

引理 5.5。对于 $i = 1$、2、$\cdots$、K_*，我们有

$$W_1(\hat{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \hat{Y}^{(i)}_{\overline{T}-\underline{T}}) \le \tilde{\mathcal{O}}(1) \cdot \sqrt{t_{i-1}\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=t_{i-1}}^{t_i} \mathbb{E}_x\left[\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t\right]\right]}.$$

右端项分解为两个因子：在 $[t_{i-1}, t_i]$ 期间的分数匹配损失，以及 $\sqrt{t_i}$。后者对应于 Y_t 从 $t = \overline{T} - t_i$ 移动到 $\overline{T} - \underline{T}$的程度。该界表明，随着 $t_i \to 0$的增大，分数匹配变得更加困难，但其对 W_1 误差的贡献却会减小。形式化证明需要构造一个逐路径的输运映射；参见引理 D.7的证明。

综上所述，我们最终得到定理5.4，即在 W_1中的近似极小极大最优速率。具体而言，如果忽略对数因子，（8）有界于

$$\sqrt{\underline{T}} + \exp(-\overline{T}) + \sqrt{t_0} n^{-\frac{2s}{d+2s}} + \sum_{i=2}^{K_*} \sqrt{t_i}\sqrt{n^{-\frac{2(s+1)}{d+2s}} + \frac{t_i^{-d/2} n^{\frac{\delta d}{2(d+2s)}}}{n}} \lesssim n^{-\frac{s+1-\delta}{d+2s}},$$

其中我们设 $\underline{T} = n^{-\frac{2(s+1)}{d+2s}}$且 $\overline{T} = (s+1)\log \underline{\beta}\, d+2s^{\,n}()$.

注释 5.6。 尽管我们使用了经过不同优化的多个网络，但这种修改也可能在现实中被隐式地实现。第一个证据是隐式正则化，即学习过程会诱导解的稀疏化（Gunasekar 等人，2017年；Arora 等人，2019年；Soudry 等人，2018）。当通过分数匹配同时并行学习不同时间区间对应的子网络时（1），这些理论表明，无需像我们的切换过程那样进行显式正则化，也能获得良好的得分网络。另一个证据是，在实践中，权重函数 $\lambda(t)$ 有时会随着 t 变大而增加（Song& Ermon，2019年；Song 等人，2020），这表明得分网络在较大的 t 处的质量得到了更多重视。

5.3. 关于离散化误差的讨论

尽管本文主要关注连续时间随机微分方程以便于简单介绍，但我们也可以将离散化误差纳入考虑。这里我们只给出摘要，详细内容见附录D.3。设 $t_0 = \underline{T} < t_1 < \cdots < t_{K_*} = \overline{T}$为满足$\eta \equiv t_{k+1} - t_k$ 的时间步。我们将得分网络训练为下式的最小化器：

$$\sum_{i=1}^{n}\sum_{k=0}^{K-1}\eta\mathbb{E}[\|s(x_{t_k},t_k)-\nabla\log p_{\overline{T}-t_k}(x_{t_k}|x_{0,i})\|^2].$$

这里的期望是相对于 $x_{\overline{T}-t_k}\sim p_{\overline{T}-t_k}(x_{\overline{T}-t_k}|x_{0,\ i})$ 计算的。接下来考虑如下过程 $(Y_{\mathrm{d}t})_{t=0}^{\eta K}$，其中 $Y_{\mathrm{d}_0}\sim\mathcal{N}(0,I_d)$：对于 $t\in[\overline{T}-t_i,\overline{T}-t_{i+1}]$,

$$\mathrm{d}Y_t^{\mathrm{d}} = \beta_t(Y_t^{\mathrm{d}} + 2\hat{s}(Y_{\overline{T}-t_i}^{\mathrm{d}},\overline{T}-t_i))\mathrm{d}t + \beta_{\overline{T}-t}\mathrm{d}B_t$$

这只是在 t 处用¯最后一个离散化步处的漂移项替换该处的漂移项，我们可以像经典的欧拉–丸山离散化那样轻松地 $Y_{\eta(k+1)}$ 从– $Y_{\eta k}$ 得到¯ ¯，因为 $Y_{\eta(k+1)}$ 在给定 $Y_{\eta k}$ 的条件下服从高斯分布。De Bortoli （2022年）；Chen 等人（2022年）也采用了这种方法。然而，De Bortoli （2022年）要求 $\eta_i \le \exp(-n^{\mathcal{O}(1)})$，而 Chen 等人（2022年）假设得分函数具有 Lipschitz 连续性，但在我们的设定中这一条件未必成立。

我们可以证明以下离散化误差界：

定理 5.7。设 $\underline{T} = n^{-\mathcal{O}(1)}$ 和 $\overline{T} = {}^{s}\log\frac{n}{2s+d}$。则，

$$\mathbb{E}[\mathrm{TV}(X_0, Y_{\overline{T}-\underline{T}}^{\mathrm{d}})] \lesssim \tilde{\mathcal{O}}\left(\eta^2\underline{T}^{-3} + n^{-\frac{s}{d+2s}}\right).$$

因此，取 $\eta = \underline{T}^{-1.5}n^{-s/(2s+d)} = \mathrm{poly}(n^{-1})$ 足以忽略离散化误差。

6. 内在维数的误差分析

尽管第 5 节中得到的速率具有极小极大最优性，但仍受维数灾难的困扰：收敛速率的指数取决于d。在统计学中，避免维数灾难的一种方法是假设光滑性具有混合性或各向异性（Ibragimov & Khas'minskii, 1984; Meier et al., 2009; Suzuki, 2018; Suzuki & Nitanda, 2021），我们的理论可直接应用于这些情形。另一方面，人们提出了流形假设，即真实世界数据的分布位于低维流形上（Tenenbaum et al., 2000; Fefferman et al., 2016）；这是另一种能够避免维数灾难的假设：可以推导出收敛速率，使其取决于流形的维数 d'，而非空间本身的维数 d Schmidt-Hieber(2019); Nakada & Imaizumi (2020)。

对于扩散模型而言，尽管其在统计学上十分重要，但现有文献均未证明扩散模型能够缓解维数灾难；首先，密度估计问题本身也从未得到研究。

我们介绍近期几项在流形假设下研究扩散建模收敛性的工作。Pidstrigach 等人(2022年)讨论了得分函数近似的影响，但他们给出的界并非定量界，也未考虑估计率。DeBortoli (2022年)研究了估计速率，但近似误差必须相对于目标估计率小到指数级。Batzolis 等人(2022年)通过实验表明，扩散建模能够学习底层流形的维数，并且可以根据训练好的扩散模型估计流形的维数。

接下来，我们定义具有内在维数的特定密度函数类，并展示其估计率。

令 $d' \le d$ 为整数，且 $A\in\mathbb{R}^{d\times d'}$ 为由范数为 1 的正交列向量构成的矩阵。我们考察

d' $V = \{y\in\mathbb{R}^d \mid \exists x \in$ 维子 $\mathbb{R}^{d'}$ 空间：满足 $y = Ax\}$ 真实密度的支撑集所在的位置，即d' 表示内在维数。结合假设2.5，我们作出如下假设。

假设 6.1。真实密度 p_0 是一个概率测度，相对于限制在子空间 V 上的勒贝格测度绝对连续。其概率密度函数在子空间 V 的标准坐标系下表示为 q。

假设 6.2。 q 的上界和下界分别为 C_f 和 C_f^{-1} 。此外， q 属于 $U(B^s_{p,\ q};\ [-1,1]^{d'})$。

假设 6.3。 q属于带有 $a_0 = n^{-\frac{1-\delta}{d'}}$ 的 $U(\mathcal{C}^\infty($ $[-1,1]^{d'}\setminus[-1+a_0,1-a_0]^{d'}))$。

我们现在将结果陈述如下：

定理 6.4。对于任意固定的 $\delta>0$，我们可以使用 $n(\gg 1)$ 个样本训练得分网络，使得

$$\mathbb{E}[W_1(X_0,\hat{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-\frac{(s+1-\delta)}{d'+2s}}.$$

附录 E 给出了完整的证明。与定理 5.1不同，此处的上界取决于 d'（而不取决于 d）。因此，我们可以得出结论：扩散模型能够避免维数灾难。

7. 结论

本文从统计学习理论的角度分析了扩散建模作为分布学习器的表现，并推导出若干估计速率。当真实密度

属于 Besov空间且深度神经网络经过适当的最小化时，扩散建模能够在全变差和 W_1 下实现近乎极小极大最优的估计速率。

为近似得分函数，我们引入了一种新基，称为扩散 B 样条基。通过仔细权衡分数匹配的难度以及各个时刻分数匹配误差对 W_1 距离的影响程度，推导出了W_1 中的界。我们还证明，在流形假设下，扩散模型可以避免维数灾难。

致谢

KO 得到了富士通有限公司的部分资助。SA 得到了日本学术振兴会科研费（22J13388）的部分资助。TS 得到了日本学术振兴会科研费（20H00576）和 JST CREST 的部分资助。

参考文献

Amann, H., Bourguignon, J., Grove, K., Lions, P., Araki, H., Brezzi, F., Chang, K., Hitchin, N., Hofer, H., Kn¨orrer, H., 等。数学专著第 99 卷。1983 年。

Arora, S., Cohen, N., Hu, W. 和 Luo, Y. 深度矩阵分解中的隐式正则化。神经信息处理系统进展，32，2019年。

Bakry, D., Gentil, I., Ledoux, M. 等。《马尔可夫扩散算子的分析与几何》，第103卷。Springer，2014年。

Barron, A. R. S形函数叠加的通用逼近界。《IEEE信息论汇刊》，39(3):930–945，1993年。

Batzolis, G., Stanczuk, J. 和 Sch¨onlieb, C.-B. 你的扩散模型暗中知晓数据流形的维数。arXiv预印本 arXiv:2212.12611，2022年。

Block, A., Mroueh, Y. 和 Rakhlin, A. 使用去噪自编码器和朗之万采样进行生成建模。arXiv预印本 arXiv:2002.00107，2020年。

Boull´e, N., Nakatsukasa, Y. 和 Townsend, A. 有理神经网络。《神经信息处理系统进展》，33:14243–14253，2020。

Chang, S.-H., Cosman, P. C. 和 Milstein, L. B. 高斯误差函数的 Chernoff 型界。《 IEEE 通信汇刊》，59(11):2939–2944，2011年。

Chen, N.、Zhang, Y.、Zen, H.、Weiss, R. J.、Norouzi, M. 和 Chan, W. WaveGrad：用于波形生成的梯度估计。在《国际学习表征会议》上，2020年。

Chen, S.、Chewi, S.、Li, J.、Li, Y.、Salim, A. 和 Zhang, A. R. 采样与学习得分函数一样简单：关于仅需最少数据假设的扩散模型理论。arXiv 预印本 arXiv:2209.11215，2022年。De Bortoli, V. 流形假设下去噪扩散模型的收敛性。《机器学习研究汇刊》，2022年。网址：https://openreview.net/forum?id=MhK5aX o3gB。

De Bortoli, V., Thornton, J., Heng, J. 和 Doucet, A. 用于基于得分的生成建模的扩散 Schr¨odinger 桥。《神经信息处理系统进展》，34:17695–17709，2021年。

DeVore, R. A. 和 Popov, V. A. Besov 空间的插值。《美国数学会汇刊》，305(1):397–414，1988。

Dhariwal, P. 和 Nichol, A. 扩散模型在图像合成方面胜过生成对抗网络。《神经信息处理系统进展》，34:8780–8794，2021年。

Dudley, R. M. 均值 Glivenko-Cantelli 收敛的速度。《数学统计年刊》，40(1): 40–50，1969年。

Fefferman, C.、Mitter, S. 和 Narayanan, H. 检验流形假设。《美国数学学会杂志》，29(4):983–1049，2016年。

Glorot, X.、Bordes, A. 和 Bengio, Y. 深度稀疏整流神经网络。载于第十四届人工智能与统计国际会议论文集，第315–323页。JMLR 研讨会与会议论文集，2011年。

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 生成对抗网络。ACM通讯，63(11):139–144，2020年。

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. 矩阵分解中的隐式正则化。神经信息处理系统进展，30，2017年。

Haussmann, U. G. and Pardoux, E. 扩散过程的时间反演。 概率年刊，14(4): 1188–1205，1986年。doi: 10.1214/aop/1176992362。URLhttps://doi.org/10.1214/aop/1176992362 .

Hayakawa, S. 和 Suzuki, T. 关于深度神经网络学习在稀疏参数空间上的极小极大最优性与优越性。《 Neural Networks》，123：343–361，2020。

Ho, J.、Jain, A. 和 Abbeel, P. 去噪扩散概率模型。《Advances in Neural Information Processing Systems》，33：6840–6851，2020。

Ho, J.、Salimans, T.、Gritsenko, A.、Chan, W.、Norouzi, M. 和 Fleet, D. J. 视频扩散模型。arXiv:2204.03458，2022年。

Hyv¨arinen, A. 和 Dayan, P. 通过分数匹配估计非归一化统计模型。机器学习研究杂志，6(4)，2005年。

Ibragimov, I. A. 和 Khas'minskii, R. Z. 关于分布密度估计的更多论述。《苏联数学杂志》，25:1155–1165，1984年。

Karatzas, I., Karatzas, I., Shreve, S., and Shreve, S. E.《布朗运动与随机微积分》，第113卷。施普林格科学与商业媒体，1991年。Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave：一种用于音频合成的多功能扩散模型。发表于国际学习表征会议，2020年。Lee, H., Lu, J., and Tan, Y. 一般数据分布下基于得分的生成建模的收敛性。发表于2022年神经信息处理系统大会基于得分的方法研讨会，2022a。Lee, H., Lu, J., and Tan, Y. 具有多项式复杂度的基于得分的生成建模的收敛性。发表于神经信息处理系统进展，2022b。Lei, J. 无界函数空间中经验测度在Wasserstein距离下的收敛与集中性。《伯努利》，26(1):767–798，2020年。Liang, T. 生成对抗网络对密度的学习能力如何：一种非参数视角。arXiv预印本 arXiv:1712.08244，2017年。Meier, L., Van de Geer, S., and B¨uhlmann, P. 高维加性建模。2009年。Mhaskar, H. N. and Micchelli, C. A. 通过叠加S形函数和径向基函数进行近似。《应用数学进展》，13(3):350–373，1992年。Nair, V. and Hinton, G. E. 修正线性单元提升受限玻尔兹曼机的性能。发表于国际机器学习会议，2010年。Nakada, R. and Imaizumi, M. 深度神经网络的自适应近似与泛化及其内在维度。机器学习研究杂志，21(174):1–38，2020年。Niles-Weed, J. and Berthet, Q. Wasserstein距离下光滑密度的极小极大估计。《统计学年鉴》，50(3):1519–1540，2022年。

Oono, K. and Suzuki, T. ResNet类卷积神经网络的近似与非参数估计。发表于 国际机器学习会议，第4922–4931页。PMLR，2019年。Petersen, P. and Voigtlaender, F. 使用深度ReLU神经网络对分段光滑函数进行最优近似。神经网络，108:296–330，2018年。Petersen, P. and Voigtlaender, F. 卷积神经网络与全连接网络近似能力的等价性。美国数学学会会刊，148(4):1567–1581，2020年。Pidstrigach, J. 基于得分的生成模型能够检测流形。载于Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K.（编），神经信息处理系统进展，2022年。网址：https://openreview.net/forum?id=AiNrnIrDf

D9.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. 使用 CLIP潜变量进行分层文本条件图像生成。arXiv 预印本 arXiv:2204.06125，2022年。Ronneberger, O., Fischer, P., and Brox, T. U-Net：用于生物医学图像分割的卷积网络。载于《医学图像计算与计算机辅助干预国际会议》，第234–241页。Springer，2015年。Schmidt-Hieber, J. 深度ReLU 网络对流形上函数的逼近。arXiv 预印本 arXiv:1908.00695，2019年。Schmidt-Hieber, J. 使用带 ReLU激活函数的深度神经网络进行非参数回归。《统计年鉴》，48(4):1875–1897，2020年。Schreuder, N., Brunel, V.-E., and Dalalyan, A. 无支配生成模型的统计保证。载于《算法学习理论》，第1051–1071页。PMLR，2021年。Singh, S. and P´oczos, B. Wasserstein距离下的极小极大分布估计。arXiv 预印本 arXiv:1802.08855，2018年。Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and P´oczos, B. 对抗损失下的非参数密度估计。《神经信息处理系统进展》，31，2018年。Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. 使用非平衡热力学进行深度无监督学习。载于《国际机器学习会议》，第2256–2265页。PMLR，2015年。

Song, Y. and Ermon, S. 通过估计数据分布的梯度进行生成建模。神经信息处理系统进展，32，2019年。Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Er-mon, S., and Poole, B. 通过随机微分方程进行基于得分的生成建模。载于国际学习表征会议，2020年。Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. 梯度下降在可分数据上的隐式偏差。机器学习研究期刊，19(1):2822–2878，2018年。Suzuki, T. 深度 ReLU 网络在 Besov 和混合光滑 Besov 空间学习中的自适应性：最优速率与维数灾难。载于国际学习表征会议，2018年。Suzuki, T. and Nitanda, A. 深度学习能够适应各向异性 Besov 空间中模型光滑度的内在维数。神经信息处理系统进展，34:3609–3621，2021年。Telgarsky, M. 神经网络与有理函数。载于国际机器学习会议，第3387–3393页。PMLR，2017年。Tenenbaum, J. B., Silva, V. d., and Langford, J. C. 非线性降维的全局几何框架。《科学》，290(5500):2319–2323，2000年。Triebel, H. 混合可积性函数空间中的熵数。《马德里康普顿斯数学杂志 》，24(1): 169–188，2011年。Tsybakov, A. B. 《非参数估计导论》。施普林格统计学丛书。Springer，2009年。ISBN 978-0-387-79051-0。doi: 10.1007/b13794。URLhttps://doi.org/10.1007/b13794。Vahdat, A., Kreis, K., and Kautz, J. 潜在空间中的基于得分的生成建模。神经信息处理系统进展，34:11287–11302，2021年。Vincent, P. 分数匹配与去噪自编码器之间的联系。《神经计算》，23(7):1661–1674，2011年。Weed, J. and Bach, F. 经验测度在 Wasserstein距离下收敛的精确渐近速率与有限样本速率。《伯努利》，25(4A):2620–2648，2019年。Yang, Y. and Barron, A. 收敛极小极大速率的信息论判定。《统计年鉴》，第1564–1599页，1999年。Yarotsky, D. 深度 ReLU 网络逼近的误差界。《神经网络》，94:103–114，2017年。Zhou, D.-X. 深度卷积神经网络的通用性。《应用与计算调和分析》，48(2):787–794，2020。

A. 反向路径的若干高概率界

分析中的困难之一在于空间和得分函数值的无界性。本小节旨在为此类问题提供若干处理方法。这些不等式使我们能够专注于有界区域内的得分函数近似。然而，我们注意到，以下某些界仍然依赖于时间 t，因此得分函数的近似和估计难度会随 t 而变化。

在下文中，我们定义 若干常数 $C_{\mathrm{a},\ i}$。除本节外，为简洁起见，我们将它们简单记为 C_{a} .

A.1. $\|Y_t\|$ 和 $\|\Delta Y_t\|$ 的高概率界

我们首先给出若干高概率界，以保证大多数路径都在某个有界区域内移动。引理 A.1（$\|Y_t\|$ 和 $\|\Delta Y_t\|$ 的高概率界）。存在一个常数 $C_{\mathrm{a},\ 1}$ 使得

$$\mathbb{P}\left[\|Y_t\|_\infty \le m_{\overline{T}-t} + C_{\mathrm{a},1}\sigma_{\overline{T}-t}\sqrt{\log(\varepsilon^{-1}\underline{T}^{-1}\overline{T})}\ \textit{for all}\ t\in[0,\overline{T}-\underline{T}]\right]\ge 1-\varepsilon.$$

此外，对于任意固定的 $0<\tau\le 1$,

$$\mathbb{P}\left[\|Y_t - Y_{t+\tau}\|_\infty \le C_{\mathrm{a},1}\sqrt{\tau\log(\varepsilon^{-1}\tau^{-1}\overline{T})}\ \textit{for all}\ t\in[0,\overline{T}-\tau]\right]\ge 1-\varepsilon.$$

证明。回顾 $Y_t = X_{\overline{T}-t}$。因此，接下来我们讨论如何界定 X_t。

我们先证明第一个断言。令 t_1、t_2、$\cdots$、t_K 为满足 $\underline{T}=t_1<t_2<\cdots<t_K=\overline{T}$ 的时间步，其中 $t_i - t_{i-1}=\Delta t$ 是稍后指定的某个标量值。我们先证明，对于某个常数 C_1，有：

$$\mathbb{P}\left[\|X_t\|_\infty \le m_t + C_1\sigma_t\sqrt{\log\varepsilon^{-1}}\ \text{for all}\ t=t_i\ (i=1,2,\cdots,K)\right]\ge 1-\varepsilon K. \tag{9}$$

回顾 $X_t|X_0$ 满足 $\mathcal{N}(m_tX_0,\ \sigma_t^2)$ 且 $\|X_0\|_\infty\le 1$。引理 F.14 给出

$$\mathbb{P}\left[\|X\|_\infty \le m_t + C_1\sigma_t\sqrt{\log\varepsilon^{-1}}\ \text{for some fixed}\ t=t_i\right]\ge 1-\varepsilon,$$

这立即推出（9）。

接下来，我们考察每个粒子 X_t 从 $t=t_{i-1}$ 移动到 t_i 的距离。等价地，我们考察 X_t 并将其分解为

$$X_t = \exp\left(-\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s\right)X_{t_{i-1}} + B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s)}, \tag{10}$$

其中 B_s 表示一个 d维布朗运动。这是通过考虑从 $t=t_{i-1}$出发的奥恩斯坦-乌伦贝克过程得到的。根据引理 F.15，以至少 ε的概率，以下结论对 $t\in[t_{i-1},t_i]$一致成立：

$$\begin{aligned}\|X_t\|_\infty &\le \exp\left(-\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s\right)\|X_{t_{i-1}}\|_\infty + \sqrt{1-\exp(-2\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s)}\cdot 2\sqrt{\overline{\beta}2\log d\varepsilon^{-1}}\\ &\le \exp\left(-\int_{s=t_{i-1}}^{t_i}\beta_s\mathrm{d}s\right)\|X_{t_{i-1}}\|_\infty + \sqrt{2\underline{\beta}\Delta t}\cdot 2\sqrt{\overline{\beta}2\log d\varepsilon^{-1}}.\end{aligned}$$

若 $\|X_{t_{i-1}}\|_\infty \le m_{t_{i-1}} + C_1\sigma_{t_{i-1}}\sqrt{\log\ \varepsilon^{-1}}$，则可进一步将其界定为

$$\|X_t\|_\infty \le m_{t_{i-1}} + C_1\sigma_{t_{i-1}}\sqrt{\log\varepsilon^{-1}} + \sqrt{\Delta t}\cdot 4\sqrt{\overline{\beta}\underline{\beta}\log d\varepsilon^{-1}}.$$

由于我们可以验证 $\sigma_t\simeq\sqrt{t}\wedge 1\ge\sqrt{\underline{T}}$成立，若取 $\Delta\le\underline{T}$，则有

$$C_1\sigma_{t_{i-1}}\sqrt{\log\varepsilon^{-1}} + \sqrt{\Delta t}\cdot 4\sqrt{\overline{\beta}\underline{\beta}\log d\varepsilon^{-1}} \lesssim C_2\sigma_{t_{i-1}}\sqrt{\log\varepsilon^{-1}} \tag{11}$$

对所有 $t \in [t_{i-1}, t_i]$，其中某个常数为 C_2。

因此，以概率 $1-2K\varepsilon$，我们有（9），以及（11）对所有 i均成立。我们需要取 $K = \mathcal{O}(\overline{T}/\underline{T})$ 以满足 $\Delta \le \underline{T}$。我们将 $\frac{\varepsilon}{K}$ 重置为新的 ε，并相应地调整 C_2。现在，第一个断言得证。

接下来，我们考虑第二个断言。令我们考虑不同的时间离散化 $t_0 = 0$，$t_1 = \tau$，$t_2 = 2\tau$，$\cdots$，$t_K = K\tau$，其中 $K = \min\{i \in \mathbb{N} | K\tau \ge \overline{T}\}$。于是，根据第一个论证，我们有 $\|X_t\|_\infty \le m_t + C_2\sigma_t\sqrt{\log(\varepsilon^{-1}\tau^{-1}\overline{T})}$以至少 $1-\varepsilon$的概率成立，对所有 $t = t_0$，t_1，$\cdots$，t_K均成立。我们对由此条件化的事件进行条件化。根据（10），我们有，对于 $t \ge t_{i-1}$，

$$X_t - X_{t_{i-1}} = \left[\exp\left(-\int_{s=t_{i-1}}^{t_i} \beta_s \mathrm{d}s\right) - 1\right] X_{t_{i-1}} + B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i}\beta_s \mathrm{d}s)},$$

由此可得

$$\begin{aligned}\|X_t - X_{t_{i-1}}\|_\infty &\le \left|\exp\left(-\int_{s=t_{i-1}}^{t_i} \beta_s \mathrm{d}s\right) - 1\right| \|X_{t_{i-1}}\|_\infty + \left\|B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i}\beta_s \mathrm{d}s)}\right\|_\infty \\ &\le \tau\overline{\beta}(m_{t_{i-1}} + C_2\sigma_{t_{i-1}}\sqrt{\log(\varepsilon^{-1}\tau^{-1}\overline{T})}) + \left\|B_{1-\exp(-2\int_{s=t_{i-1}}^{t_i}\beta_s \mathrm{d}s)}\right\|_\infty\end{aligned}$$

我们对 $t \in [t_{i-1}, t_i]$ 上的最后一项进行界定。以至少 $1-\frac{\varepsilon}{K}$ 的概率，该项不超过 $\sqrt{2\underline{\beta}\tau}\cdot 2\sqrt{\overline{\beta}2\log dK\varepsilon^{-1}}$，根据引理 F.15。综上，以至少 $1-2\varepsilon$ 的概率，

$$\sup_{t\in[t_{i-1},t_i]} \|X_t - X_{t_{i-1}}\|_\infty \le \tau\overline{\beta}(m_{t_{i-1}} + C_2\sigma_{t_{i-1}}\sqrt{\log(\varepsilon^{-1}\tau^{-1}\overline{T})}) + \sqrt{2\underline{\beta}\tau}\cdot 2\sqrt{\overline{\beta}2\log dK\varepsilon^{-1}}$$

对所有 $i = 0, 1, \cdots, K-1$，该结论均成立。右侧由 $C_3\sqrt{\tau\log\varepsilon^{-1}\tau^{-1}\overline{T}}$界定，其中的常数 C_3 充分大。

那么，对于任意 t，存在 i使得 $t \le t_i \le t+\tau$。因此，以概率 $1-2\varepsilon$，$\|X_t - X_{t+\tau}\|_\infty \le \|X_t - X_{t_{i-1}}\|_\infty + \|X_{t_i} - X_{t_{i-1}}\|_\infty + \|X_{t+\tau} - X_{t_i}\|_\infty$ 对所有 t 均不超过 $3C_3\sqrt{\tau\log\varepsilon^{-1}\tau^{-1}\overline{T}}$。令 2ε 为 ε，即可得到第二个断言。□

A.2. $p_t(x)$ 的界

接下来我们给出 $p_t(x)$ 的上界和下界。

引理 A.2（密度 $p_t(x)$ 的上界和下界）。对于常数 $C_{a,2}$（取决于 C_f 和 d），以下关于 $p_t(x)$ 的上界和下界成立：

$$C_{a,2}^{-1}\exp\left(-\frac{d(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\right) \le p_t(x) \le C_{a,2}\exp\left(-\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2}\right). \quad (\textit{for all } t.)$$

证明。我们首先考虑 $x \in [-m_t, m_t]^d$ 的情形。上界相对容易得到。$f(y) \le C_f\mathbb{I}[y \in [-1,1]^d]$表示

$$p_t(x) = \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f(y)\exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y \le \int \frac{C_f\mathbb{I}[y\in[-1,1]^d]}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y = \frac{2^d C_f}{\sigma_t^d(2\pi)^{\frac{d}{2}}}. \quad (12)$$

与此同时，我们有

$$p_t(x) \le \int \frac{C_f}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y = \frac{C_f}{m_t^d}. \quad (13)$$

因此，根据（12）和（13），$p_t(x)$ 有界于 $\min\left\{\frac{2^d C_f}{\sigma_t^d(2\pi)^{\frac{d}{2}}}, \frac{C_f}{m_t^d}\right\}$。进一步地，它可以被一个仅依赖于 C_f 和 d 的常数所界定，因为 $m_t^2 + \sigma_t^2 = 1$ 对所有 t均成立。

下界可以理解如下。我们有

$$\begin{aligned} p_t(x) &= \int \frac{C_f^{-1}}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\ &\geq \frac{1}{(2\pi)^{\frac{d}{2}}} \int f(x/m_t - \sigma_t y) \exp\left(-\frac{\|m_t y\|^2}{2}\right) \mathrm{d}y \quad \text{(by letting } (x - m_t y)/\sigma_t \mapsto m_t y\text{).} \end{aligned} \tag{14}$$

由于 $x \in [-m_t, m_t]^d$，我们有 $x/m_t \in [-1,1]^d$。因此，$|\{y \in [-1,1]^d |\ x/m_t - \sigma_t y \in [-1,1]\}| \geq 1$。此外，$\exp\left(-\frac{\|m_t y\|^2}{2}\right) \geq \exp(-d^2/2)$ 在 $y \in [-1,1]^d$中。因此，积分（14）的下界为 $\exp(-d^2/2)$。

接下来考虑 $x \notin [-m_t, m_t]^d$ 的情况。对于这样的 x，令 $r = (\|x\|_\infty - m_t)/\sigma_t$ 并从 $\{1, 2, \cdots, d\}$ 中选择 i^*，使得 $|x_{i^*}| = \|x\|_\infty = m_t + r/\sigma_t$ 成立。于是，我们得到 $p_t(x)$ 的上界：

$$\begin{aligned} p_t(x) &= \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\ &\leq C_f \prod_{i=1}^d \int \frac{\mathbb{I}[-1 \leq y_i \leq 1]}{\sigma_t(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i \\ &\lesssim C_f \int_{y_{i^*} \in [-1,1]} \frac{1}{\sigma_t(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_{i^*} - m_t y_{i^*})^2}{2\sigma_t^2}\right) \mathrm{d}y \\ &\left(\text{because } \int \frac{\mathbb{I}[-1 \leq y_i \leq 1]}{\sigma_t(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y_i \text{ for } i \neq i^* \text{ is bounded by } \mathcal{O}(1), \text{ as } p_t(x) \text{ for } x \in [-m_t, m_t]^d.\right) \\ &\leq \frac{C_f}{m_t} \int_{a = r/\sqrt{2}}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}a \qquad \text{(by } a = x_{i*} - m_t y_{i_*}/\sqrt{2}\sigma_t\text{)} \\ &\leq \frac{C_f}{m_t} \exp\left(-r^2/2\right) = \frac{C_f}{m_t} \exp\left(-\frac{(\|x\|_\infty - m_t)^2}{2\sigma_t^2}\right) \end{aligned} \tag{15}$$

最后一个不等式中我们使用了 $\int_z^\infty e^{-a^2} \mathrm{d}a \leq e^{-z^2}$（例如，Chang 等人（2011年））。此外，(15) 也可以由 $\frac{2C_f}{\sigma_t\sqrt{2\pi}} \exp\left(-\frac{(\|x\|_\infty - m_t)^2}{2\sigma_t^2}\right)$ 进行界定。由于 $m_t^2 + \sigma_t^2 = 1$ 意味着 $\min\{m_t, \sigma_t\} \gtrsim 1$，因此有 $p_t(x) \lesssim C_f \exp\left(-\frac{(\|x\|_\infty - m_t)^2}{2\sigma_t^2}\right)$。

另一方面，

$$\begin{aligned} p_t(x) &= \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\ &\geq C_f^{-1} \prod_{i=1}^d \underbrace{\int_{y_i \in [-1,1]} \frac{1}{\sigma_t(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - m_t y_i)^2}{2\sigma_t^2}\right) \mathrm{d}y}_{\text{(a)}} \\ &= C_f^{-1} \left(\int_{y_{i^*} \in [-1,1]} \frac{1}{\sigma_t(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(x_{i^*} - m_t y_{i^*})^2}{2\sigma_t^2}\right) \mathrm{d}y\right)^d \quad \text{(because (a) is minimized when } i = i_*\text{)} \\ &\geq \frac{C_f^{-1}}{m_t^d} \left(\int_{a = r/\sqrt{2}}^{r/\sqrt{2} + \sqrt{2} m_t/\sigma_t} \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}y\right)^d \quad \text{(by } (x_{i^*} - m_t y_{i^*})/\sqrt{2}\sigma_t\text{)} \\ &\geq \frac{C_f^{-1}}{m_t^d} \left(\int_{a = r/\sqrt{2}}^{r/\sqrt{2} + \sqrt{2} m_t} \frac{1}{\sqrt{\pi}} \exp\left(-a^2\right) \mathrm{d}y\right)^d \end{aligned}$$

$$\geq \frac{C_f^{-1}}{m_t^d}\left(\frac{\sqrt{2}m_t}{\sqrt{\pi}}\exp\left(-(r/\sqrt{2}+\sqrt{2}m_t)^2\right)\right)^d$$

（通过对积分区间中的 $\exp(-a^2)$ 取下界，再乘以区间宽度）

$$\geq \frac{C_f^{-1}}{m_t^d}\left(\frac{\sqrt{2}m_t}{\sqrt{\pi}}\exp\left(-r^2-4\right)\mathrm{d}a\right)^d$$

$$\geq \frac{C_f^{-1}2^{d/2}}{e^{4d}\pi^{d/2}}\exp\left(-dr^2\right),$$

这给出了 $p_t(x)$ 的下界。 □

A.3. $p_t(x)$ 的导数与得分函数的界

本小节将评估 $p_t(x)$ 的导数与得分函数。一方面，通过直接论证可得，$p_t(x)$ 的导数由 $\partial^k p_t(x) = \mathcal{O}(1/\sigma_t^k) = \mathcal{O}(t^{-k/2})$ 所界定。另一方面，对于得分函数，一般而言，$\sup_{x\in\mathbb{R}^d}\|\nabla\log p_t(x)\| = \infty$ 成立，这使得我们无法使用神经网络构造得分函数的近似。这是因为当 $\|x\|\to\infty$ 时，$\nabla\log p_t(x) = \nabla p_t(x)/p_t(x)$ 和 $p_t(x)$ 可能任意小。然而，利用引理 A.2，我们可以在接下来的引理 A.3 中证明~~依赖于~~ x 和 t 的得分函数界。在引理 A.4 中，利用引理 A.3 证明 p_t 的衰减速度非常快，因此在 $p_t(x)$ 较小区域内的近似误差（该区域在某些 x 中可以是 $\gg 1$ ）不会对 $L^2(p_t)$ 的近似误差界造成太大影响；我们可以以高概率证明 $\|\nabla\log p_t(x)\| = \tilde{\mathcal{O}}(1/\sigma_t) = \tilde{\mathcal{O}}(1\vee 1/\sqrt{t})$（当 $x\sim p_t$ 时）。

引理 A.3（导数的有界性）。对于 $k\in\mathbb{Z}_+$，存在常数 $C_{\mathrm{a},3}$ 仅依赖于 k、d 和 C_f，使得

$$|\partial_{x_{i_1}}\partial_{x_{i_2}}\cdots\partial_{x_{i_k}}p_t(x)| \leq \frac{C_{\mathrm{a},3}}{\sigma_t^k}. \tag{16}$$

此外，我们有

$$\|\nabla\log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t}\cdot\left(\frac{(\|x\|_\infty - m_t)_+}{\sigma_t}\vee 1\right), \tag{17}$$

并且对于 $i\in\{1,2,\cdots,d\}$,

$$\|\partial_{x_i}\nabla\log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t^2}\left(\frac{(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\vee 1\right). \tag{18}$$

并且

$$\|\partial_t\nabla\log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t^3}\left[|\partial_t\sigma_t| + |\partial_t m_t|\right]\left(\frac{(\|x\|_\infty - m_t)_+^2}{\sigma_t^2}\vee 1\right)^{\frac{3}{2}}. \tag{19}$$

证明。首先，我们考虑 (16)。令 $g_1(x) = p_t(x) = \int_{\sigma_t^d}\frac{1}{(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y$. 对于 $s\in\mathbb{Z}_+^d$，我们将记号简写为 $g_1^{(s)}(x) = \partial_{x_1}^{s_1}\partial_{x_2}^{s_2}\cdots\partial_{x_d}^{s_d}g_1(x)$。对于 $s\in\overline{\mathbb{Z}_+^d}$，我们定义 $B_s = \{s'\in\mathbb{Z}_+^d | s'_i\leq s_i (i=1,\cdots,d)\}$ 以及常数 c_s，使得 $\partial_{x_1}^{s_1}\partial_{x_2}^{s_2}\cdots\partial_{x_d}^{s_d}e^{-\|x\|^2/2} = \sum_{s'\in B_s}c_{s'}x_1^{s'_1}x_2^{s'_2}\cdots x_d^{s'_d}e^{-\|x\|^2/2}$ 成立。然后，由于 $\partial_{x_i} = \frac{1}{\sigma}\partial_{\frac{x_i}{\sigma}}$，我们可以将 $g_1^{(s)}(x)$ 写作

$$g_1^{(s)}(x) = \frac{\sum_{s'\in B_s}c_{s'}}{\sigma_t^{\sum_{i=1}^d s_i}}\underbrace{\int\prod_{i=1}^d\left(\frac{x_i - m y_i}{\sigma_t}\right)^{s'_i}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y}_{\text{(a)}}. \tag{20}$$

注意到 $\max_{s:\ \sum s_i\leq k}\{\sum_{s'\in B_s}c_{s'}\}$ 被一个仅依赖于 k的常数所界定。因此，我们着重评估 (a)。当 $t\leq 1$时，(a) 在（20）中可由 $\mathcal{O}(1/m_t^d)\simeq\mathcal{O}(1)$（我们省略对 $\sum_{i=1}^d s'_i\leq k$和 C_f的依赖）。这

这是因为 $m_t \simeq 1$和 f（y） $\leq C_f$。另一方面，当 $t \geq 1$时， $\sigma_t \gtrsim 1$成立，因此我们可以通过注意到 (a) 可由 $\mathcal{O}$（1）界定；只在 f（y） $\neq 0$时才有 $y \in [-1,1]^d$。现在，第一个陈述（16）已得到证明。

接下来我们考虑 ∇log $p_t(x)$ 及其导数。我们可以聚焦于 $[\nabla \log p_t(x)]_1$,，得分函数的所有其他坐标也可以用同样的方式进行界定。令 $g_2(x) = \sigma_t[\nabla p_t(x)]_1 = -\int \frac{x_1 - m_t y_1}{\sigma_t^{d+1}\ 2\pi\ \overline{()}^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y$，并定义 $g_2^{(s)}$in 与 $g_1^{(s)}$ 的定义方式相同。

我们可以看到

$$[\nabla \log p_t(x)]_1 = \frac{1}{\sigma_t} \cdot \frac{g_2(x)}{g_1(x)}, \quad [\partial_{x_i} \nabla \log p_t(x)]_1 = \frac{1}{\sigma_t} \cdot \frac{\partial_{x_i} g_2(x)}{g_1(x)} - \frac{1}{\sigma_t} \cdot \frac{g_2(x)(\partial_{x_i} g_1(x))}{g_1^2(x)}. \tag{21}$$

此外,

$$\frac{g_2(x)}{g_1(x)} = \frac{-\int \frac{x_1 - m_t y_1}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}, \tag{22}$$

$$\frac{\partial_{x_i} g_1(x)}{g_1(x)} = \frac{1}{\sigma_t} \cdot \frac{-\int \frac{x_i - m_t y_i}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}, \tag{23}$$

$$\frac{\partial_{x_i} g_2(x)}{g_1(x)} = -\frac{1}{\sigma_t} \cdot \frac{\int \frac{\mathbb{I}^{[i=1]} - \frac{x_1 - m_t y_1}{\sigma_t} \frac{x_i - m_t y_i}{\sigma_t}}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}. \tag{24}$$

为了给它们设定界，我们考虑以下与 $\sum_{i=1}^d s_i \leq 2$ 有关的量。此外，令 ε 为稍后指定的标量值，暂且假设 $p_t(x)$ $\geq \varepsilon$ 成立。

$$\frac{\int \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{\sigma_t^2}\right) \mathrm{d}y} \tag{25}$$

根据引理 F.10，我们有

$$\left| \int_{A^x} \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma_t^2}\right) \mathrm{d}y \right.$$
$$\left. - \int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-my\|^2}{2\sigma_t^2}\right) \mathrm{d}y \right| \leq \frac{\varepsilon}{2}.$$

其中 $A^x = \prod_{i=1}^d a_i^x$含有 $a_i^x = [\frac{x_1}{m_t} - \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\text{对数 } 2\varepsilon^{-1}}$、 $\frac{x_1}{m_t} + \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\text{对数 } 2\varepsilon^{-1}}]$。注意， C_{f}仅取决于 $\sum_{i=1}^d s_i$、 d 和 C_f。

因此，当 $p_t(x) = g_1(x) \geq \varepsilon$,

$$(25) \leq \frac{2\int \prod_{i=1}^d \left(\frac{x_i - m_t y_i}{\sigma_t}\right)^{s_i} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y}{\int_{A^x} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f(y) \exp\left(-\frac{\|x-m_t y\|^2}{\sigma_t^2}\right) \mathrm{d}y}$$

$$\leq \frac{2\int_{A^x}\prod_{i=1}^d\left(\frac{x_i-m_t y_i}{\sigma_t}\right)^{s_i}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y}{\int_{A^x}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{\sigma_t^2}\right)\mathrm{d}y}+\frac{2\cdot\frac{\varepsilon}{2}}{\varepsilon}$$

(note that the denominator is larger than ε)

$$\leq 2\max_{y\in A_x}\left[\prod_{i=1}^d\left(\frac{x_i-m_t y_i}{\sigma_t}\right)^{s_i}\right]+1$$

$$\leq 2\left(C_{\mathrm{f}}^2\log\varepsilon^{-1}\right)^{(\sum_{i=1}^d s_i)/2}+1. \tag{26}$$

将此界应用于（22）、（23）和（24），$g_2(x)\,g_1(x)$、$\partial_{x_i}g_1(x)\,g_1(x)$，以及$\partial_{x_i}g_2(x)\,g_1(x)$均有界于

$$\log^{1/2}\varepsilon^{-1},\ \frac{\log^{1/2}\varepsilon^{-1}}{\sigma_t},\ \text{and}\ \frac{\log\varepsilon^{-1}}{\sigma_t},$$

分别相差至多一个常数因子。最后，我们将其应用于（21），得到

$$\|\nabla\log p_t(x)\|\lesssim\frac{\log^{1/2}\varepsilon^{-1}}{\sigma_t}\ \text{and},\ \|\partial_{x_i}\nabla\log p_t(x)\|\lesssim\frac{\log\varepsilon^{-1}}{\sigma_t^2}.$$

现在，我们将 ε 替换为一个具体值。记住，ε 应满足 $\varepsilon\leq p_t(x)$。根据引理 A.2，我们有$C_{\mathrm{a}}^{-1},2\exp\left(-\frac{d(\|x\|_\infty-m_t)_+^2}{\sigma_t^2}\right)\leq p_t(x)$，由此可得

$$\|\nabla\log p_t(x)\|\leq\frac{C_{\mathrm{a},3}}{\sigma_t}\cdot\frac{(\|x\|_\infty-m_t)_+}{\sigma_t}\vee 1,\ \text{and}\quad\|\partial_{x_i}\nabla\log p_t(x)\|\leq\frac{C_{\mathrm{a},3}}{\sigma_t^2}\left(\frac{(\|x\|_\infty-m_t)_+^2}{\sigma_t^2}\vee 1\right),$$

其中$C_{\mathrm{a},3}$取决于 k，d和 C_f。因此，我们得到（17）和（18）。

最后，我们考虑 $\partial_t\nabla\log p_t(x)$。

$$\partial_t\nabla\log p_t(x)=\partial_t\left(\frac{1}{\sigma_t}\cdot\frac{g_2(x)}{g_1(x)}\right)=\left(\partial_t\frac{1}{\sigma_t}\right)\frac{g_2(x)}{g_1(x)}-\frac{1}{\sigma_t}\cdot\frac{(\partial_t g_1(x))}{g_1(x)}\cdot\frac{g_2(x)}{g_1(x)}+\frac{1}{\sigma_t}\cdot\frac{\partial_t g_2(x)}{g_1(x)}$$

$$=\frac{(-\partial_t\sigma_t)}{\sigma_t}\nabla\log p_t(x)$$

$$-\frac{1}{\sigma_t}\cdot\frac{\int\frac{-d(\partial_t\sigma_t)\sigma_t^{-1}+\|x-m_t y\|^2(\partial_t\sigma_t)\sigma_t^{-3}-(\partial_t m_t)y^\top(m_t y-x)\sigma_t^{-2}}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y}{\int\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y}\cdot\nabla\log p_t(x).$$

$$+\frac{1}{\sigma_t}\cdot\frac{\int\frac{(\partial_t m_t)y_1+(x_1-m_t y_1)((d+1)(\partial_t\sigma_t)\sigma_t^{-1}-\|x-m_t y\|^2(\nabla_t\sigma_t)\sigma_t^{-3}+(\partial_t m_t)y^\top(m_t y-x)\sigma_t^{-2})}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y}{\int\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y}$$

仔细地将其分解为（25）中的各项之和，然后应用（26）和引理 A.2，即可得到最终界（19）。

□

现在，基于引理 A.3 我们证明只需在某个有界区域内，并且在 $\nabla\log p_t(x)$ 不太小的 x上近似 $p_t(x)$。

引理 A.4（裁剪操作导致的误差界）。设 $t\geq\underline{T}$。存在一个常数$C_{\mathrm{a},4}$取决于 d 和 C_f，我们有

$$\int_{\|x\|_\infty\geq m_t+C_{\mathrm{a},4}\sigma_t\sqrt{\log\varepsilon^{-1}\underline{T}^{-1}}}p_t(x)\|\nabla\log p_t(x)\|^2\mathrm{d}x\leq\varepsilon, \tag{27}$$

$$\int_{\|x\|_\infty\geq m_t+C_{\mathrm{a},4}\sigma_t\sqrt{\log\varepsilon^{-1}\underline{T}^{-1}}}p_t(x)\mathrm{d}x\leq\varepsilon \tag{28}$$

对所有 $t \geq \underline{T}$。

此外，存在一个常数 $C_{\mathrm{a},\ 5}$ 取决于 d 和 C_f，并且对于 x，使得 $\|x\|_\infty \leq m_t + C_{\mathrm{a},\ 4}\sigma_t\sqrt{\log \varepsilon^{-1}}$，我们有

$$\|\nabla \log p_t(x)\| \leq \frac{C_{\mathrm{a},5}}{\sigma_t}\sqrt{\log \varepsilon^{-1}}.$$

因此，

$$\int_{\|x\|_\infty \leq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}\underline{T}^{-1}}} p_t(x)\mathbb{I}[p_t(x) \leq \varepsilon]\|\nabla \log p_t(x)\|^2 \mathrm{d}x \leq \frac{C_{\mathrm{a},5}\varepsilon}{\sigma_t^2}\cdot \log^{\frac{d+2}{2}}(\varepsilon^{-1}\underline{T}^{-1}), \tag{29}$$

$$\int_{\|x\|_\infty \leq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}\underline{T}^{-1}}} p_t(x)\mathbb{I}[p_t(x) \leq \varepsilon]\mathrm{d}x \leq C_{\mathrm{a},5}\varepsilon\cdot \log^{\frac{d}{2}}(\varepsilon^{-1}\underline{T}^{-1}). \tag{30}$$

证明。根据引理A.2 和引理A.3，

$$\begin{aligned} p_t(x)\|\nabla \log p_t(x)\|^2 &\leq C_{\mathrm{a},2}\exp\left(-\frac{(\|x\|_\infty - m_t)_+^2}{2\sigma_t^2}\right)\cdot \frac{C_{\mathrm{a},3}^2}{\sigma_t^2}\frac{(\|x\|_\infty - m_t)_+^2}{\sigma_t^2} \\ &\leq \frac{C_{\mathrm{a},2}C_{\mathrm{a},3}^2}{\sigma_t^2}\exp\left(-\frac{r^2}{2}\right)r^2, \end{aligned}$$

其中令 $r := (\|x\|_\infty - m_t)_+/\sigma_t$。然后，

$$\begin{aligned} &\int_{\|x\|_\infty \geq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}}} p_t(x)\|\nabla \log p_t(x)\|^2 \mathrm{d}x \\ &\leq \int_{C_{\mathrm{a},4}\sqrt{\log \varepsilon^{-1}}}^\infty \frac{C_{\mathrm{a},2}C_{\mathrm{a},3}^2}{\sigma_t}\exp\left(-\frac{r^2}{2}\right)r^2(d-1)(\sigma_t r + m_t)^{d-1}\mathrm{d}r \\ &\lesssim \frac{1}{\sigma_t}\varepsilon\log^{d/2}\varepsilon^{-1}. \end{aligned}$$

通过分部积分即可得到最后一个不等式。因为$\sigma_t \gtrsim \sqrt{\underline{T}}$，若取 $\varepsilon' = \sqrt{\underline{T}}\cdot\varepsilon^2$，则有$\frac{1}{\sigma_t}\varepsilon'$ 对数$^{d/2}((\varepsilon')^{-1}) \lesssim \varepsilon$。因此，将 ε 替换为 ε' 并调整$C_{\mathrm{a},\ 4}$即可得到界（27）。同理，

$$\begin{aligned} \int_{\|x\|_\infty \geq m + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}}} p_t(x)\mathrm{d}x &\leq \int_{C_{\mathrm{a},4}\sqrt{\log \varepsilon^{-1}}}^\infty C_{\mathrm{a},2}\sigma_t\exp\left(-\frac{r^2}{2}\right)(d-1)(\sigma_t r + m)^{d-1}\mathrm{d}r \\ &\lesssim \sigma_t\varepsilon\log^{(d-2)/2}\varepsilon^{-1}, \end{aligned}$$

从而得到（28）。

接下来考虑引理的第二部分。式（28）是引理 A.3 的直接推论：对于 x 当 $\|x\|_\infty \leq m_t + C_{\mathrm{a},\ 5}\sigma_t\sqrt{\log \varepsilon^{-1}}$

$$\|\nabla \log p_t(x)\| \leq \frac{C_{\mathrm{a},3}}{\sigma_t}\cdot C_{\mathrm{a},4}\sqrt{\log \varepsilon^{-1}} \leq \frac{C_{\mathrm{a},5}}{\sigma_t}\sqrt{\log \varepsilon^{-1}}. \quad \text{(by taking } C_{\mathrm{a},5} \text{ larger than } C_{\mathrm{a},3}C_{\mathrm{a},4}.)$$

利用这一点，我们得到

$$\int_{\|x\|_\infty \leq m_t + C_{\mathrm{a},4}\sigma_t\sqrt{\log \varepsilon^{-1}}} p_t(x)\mathbb{I}[p_t(x) \leq \varepsilon]\|\nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \varepsilon\cdot\frac{C_{\mathrm{a},4}^2}{\sigma_t^2}\log\varepsilon^{-1}\cdot(m_t + C_{\mathrm{a},5}\sigma_t\sqrt{\log\varepsilon^{-1}})^d.$$

调整$C_{\mathrm{a}、\ 4}$、$C_{\mathrm{a}、\ 5}$ 并重置 ε 可得到（29）。方程（30）同理可得。 □

B. 得分函数的近似

本节分析（理想的）得分匹配损失最小化的近似误差。我们构造一个神经网络来近似 $\nabla \log p_t(x)$，并界定每个时间 t 处的近似误差。在本节中，我们始终将足够大的 N 作为决定神经网络规模的参数，并令 $\underline{T} = \text{poly}(N^{-1})$ 和 $\overline{T} = \mathcal{O}(\log N)$。

B.1. m_t 和 σ_t的近似

我们首先构造用于近似 m_t 和 σ_t的子网络。除了真实的数据分布 $p_0(x)$之外，得分函数 $\nabla \log p_t(x)$也依赖于 m_t 和 σ_t。事实上，在我们的构造中，每个扩散 B 样条基都被近似为关于 x、 m_t 和 σ_t的有理函数。这里， m_t 和 σ_t 与 x同样重要，因为我们会对 m_t 和 σ_t以及 x进行指数运算，而 m_t 和 σ_t 的精确值无法获得。换言之，由于 m_t和 σ_t 的近似误差会在这种指数运算中被放大，因此要获得紧致界，就必须高精度地近似 m_t 和 σ_t。因此，在本小节中，我们构造用于高效近似 m_t 和 σ_t的子网络。下面给出引理 3.3 的正式表述。

引理 B.1. 设 $0 < \varepsilon < \frac{1}{2}$。那么，存在一个神经网络 $\phi_m(t) \in \Phi(L, W, B, S)$，对于所有 $t \geq 0$，它都能以不超过 ε 的加性误差逼近 m_t，其中 $L = \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log \varepsilon^{-1})$ $S = \mathcal{O}(\log^2 \varepsilon^{-1})$，且$B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$。

此外，存在一个神经网络 $\phi_\sigma(t) \in \Phi(L, W, B, S)$，对于所有 $t \geq \varepsilon$，它都能以不超过 ε 的加性误差逼近 σ_t，其中 $L \leq \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1})$ $S = \mathcal{O}(\log^4 \varepsilon^{-1})$，且 $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$。

证明。首先考虑 $m_t = \exp(-\int_0^t \beta_s \mathrm{d}s)$。由于对所有 $t \geq A$ 都有 $\beta \geq \underline{\beta}$, $\int_0^t \beta_s \mathrm{d}s \geq \log 4\varepsilon^{-1} := \log 4\varepsilon^{-1}/\underline{\beta}$。我们将范围限制在 $[0, A]$ 内。然后，根据假设 2.5，我们可以将 β_s 展开为 $\beta_s = \sum_{i=0}^{k-1} \frac{\beta^{(i)}}{i!} s^i + \frac{\beta^{(k)}}{k!}(\theta s)^k$，其中 $|\beta^{(i)}| \leq 1$和 $0 < \theta < 1$，因此得到

$$\left| \int_0^t \beta_s \mathrm{d}s - \int_0^t \sum_{i=1}^{k-1} \frac{\beta^{(i)}}{i!} s^i \mathrm{d}s \right| \leq \frac{|\beta^{(k)}| A^{k+1}}{(k+1)!} \leq \frac{A^{k+1}}{(k+1)!}.$$

我们取 $k = \max\{2eA, \lceil \log_2 4\varepsilon^{-1} \rceil\} - 1$，从而有 $\frac{A^{k+1}}{(k+1)!} \leq \left(\frac{eA}{k+1}\right)^{k+1} \leq \frac{\varepsilon}{4}$。 $\int_0^t \sum_{i=1}^{k-1} \frac{\beta^{(i)}}{i}!$ $s^i = \sum_{i=1}^{k-1} \beta^{(i)}$ $(i+1)!$ t^{i+1} 可以由具有 $L = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}) = \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(A + \log \varepsilon^{-1}) = \overline{\mathcal{O}(\log}$ $\varepsilon^{-1})$ $S = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1}) = \mathcal{O}(\log^2 \varepsilon^{-1})$ $B = \exp(\log^2 \mathcal{O}(A + \log \varepsilon^{-1})) = \mathcal{O}(\log^2 \varepsilon^{-1})$ 的神经网络实现，其加性误差至多为 $\frac{\varepsilon}{4}$；这里用到了引理 F.3 和 F.6。根据 A 的定义，我们可以很容易地验证 $e^{-A} \leq \frac{\varepsilon}{4}$ 成立。我们用 $[0, A]$ 对输入进行截断，得到神经网络 ϕ_1，它以 $\frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}$ 的加性误差近似 $\int_0^t \beta_s \mathrm{d}s$（对于 $x \in [0, A]$），并且对所有 $x \geq A$ 都满足 $|\phi_1(x)| = |\phi_1(A)|$。

接下来，我们应用引理 F.12，其中 $\varepsilon = \frac{\varepsilon}{4}$。由此可得到所需规模的神经网络 ϕ_m，它以加性误差 $m_t = \exp(-\int_0^t \beta_s \mathrm{d}s)$ 近似 $\frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \frac{3\varepsilon}{4}$，其中 $x \in [0, A]$ 且 $|\phi_m(x) - e^{-x}| \leq |\phi_m(x) - \phi_m(A)| + |\phi_m(A)$
$- e^{-A}| + |e^{-A} - e^{-x}| \leq 0 + \frac{3\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon$，其中 $x \geq A$。

类似地，我们可以使用一个神经网络，以加性误差 $\mathcal{O}(\varepsilon^{1.5})$ 近似 $\sigma^2 = 1 - \exp(-2\int_0^t \beta_s \mathrm{d}s)$，其规模为$L = \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log \varepsilon^{-1})$ $S = \mathcal{O}(\log^2 \varepsilon^{-1})$ $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$。由于 $t \geq \varepsilon$，存在一个依赖于 $\underline{\beta}$ 的常数 c，使得 $\sigma_t^2 = 1 - \exp(-2\int_0^t \beta_s \mathrm{d}s) \geq c\varepsilon$。接下来，我们应用引理 F.9，其中 $\varepsilon = c\varepsilon$，最终得到一个神经网络 $\phi_\sigma(t)$，它以加性误差 $c\varepsilon + \frac{\varepsilon^{1.5}}{\sqrt{c\varepsilon}} = \mathcal{O}(\varepsilon)$ 近似 σ_t，其中 $L = \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1})$ $S = \mathcal{O}(\log^4 \varepsilon^{-1})$，且 $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$。调整隐藏常数可使近似误差小于 ε，从而完成证明。

□

B.2. 通过扩散 B 样条基进行近似

本小节介绍如何利用扩散 B 样条基和张量积扩散 B 样条基进行近似，从而使我们能够在 $\nabla \log p_t(x)$ 所处的$\mathbb{R}^d \times [\underline{T}, \overline{T}]$ 空间中近似得分函数。尽管我们考虑在一个$(d+1)$维空间中进行函数逼近，但所得速率（定理3.1）是一个 d维

空间。这是因为我们的基分解能够反映 p_0 在 $t>0$中的结构。在开始形式化证明之前，我们将详细阐述通过扩散 B 样条基和张量积扩散 B 样条基进行近似的证明思路，其内容比第 3 节中的描述更加详细。

请注意，阶数为 l 的基数 B 样条基可写为

$$\mathcal{N}_m(x)=\frac{1}{l!}\mathbb{I}[0\le x\le l+1]\sum_{l'=0}^{l}(-1)^j{}_{l+1}\mathrm{C}_{l'}(x-l')_+^l$$

（例如，参见Mhaskar & Micchelli （1992）的式（4.28））， 而 Besov空间中的函数可近似为若干项之和 $M_{k,j}^d($ $x)$

$$M_{k,j}^d(x)=\prod_{i=1}^{d}\mathcal{N}_m(2^{k_i}x_i-j_i)$$

其中 $k\in\mathbb{Z}_+^d$ 且 $j\in\mathbb{Z}^d$。
因此，得分函数的分母和分子

$$\nabla\log p_t(x)=\frac{\nabla p_t(x)}{p_t(x)}=-\frac{1}{\sigma_t}\cdot\frac{\int\frac{x-m_ty}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y}{\int\frac{1}{\sigma_t^{d}(2\pi)^{\frac{d}{2}}}f(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y}$$

可分解为以下各项之和

$$E_{k,j}^{(1)}(x,t):=\int\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\mathbb{I}[\|y\|_\infty\le C_{\mathrm{b},1}]M_{k,j}^d(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y \tag{31}$$

and

$$E_{k,j}^{(2)}(x,t):=\int\frac{x-m_ty}{\sigma^{d+1}(2\pi)^{\frac{d}{2}}}\mathbb{I}[\|y\|_\infty\le C_{\mathrm{b},1}]M_{k,j}^d(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y, \tag{32}$$

分别。这对应于我们在第 3 节中所称的张量积扩散 B 样条基。这里， $E_k^{(1)}{}_{,j}(x,t)$ 与第 3 节中的 $E_{k,j}(x,t)$ 相同，只是 $\mathbb{I}[\|y\|_\infty\le C_{\mathrm{b},1}]$ 项除外。注意，$C_{\mathrm{b},1}$ 是一个稍后会调整的标量值。接着，我们将结合子网络，分别近似 ∇ log $p_t(x)$ 的分母和分子；这些子网络分别近似每个 $E_k^{(1)}{}_{,j}(x,t)$ 或 $E_k^{(2)}{}_{,j}(x,t)$。

这里我们简要说明 $\mathbb{I}[\|y\|_\infty\le C_{\mathrm{b},1}]$ 出现的原因。假设 $C_{\mathrm{b},1}=1$，并基于 $p_0(x)$ 的基分解来近似 $p_t(x)$，尽管稍后我们还需要考虑其他情况。如果我们将基分解写成 $p_0(x)\approx f_N(x)=\sum M_{k,j}^d(x)$，那么诸如引理 F.13 之类的已有结果只能保证该近似在 $[-1,1]^d$ 内有效，而无法保证该区域之外的任何性质。当我们对整个 $\mathbb{R}^d$ 上的 $p_t(x)$ 近似进行积分时，这可能会损害近似精度。因此，如果 $\|x\|_\infty>1$，我们需要通过示性函数强制令 $f_N(x)=0$。

接下来，我们使用神经网络实现（修正后的）张量积扩散 B 样条基。以 $E_k^{(1)}{}_{,j}$ 为例， $E_k^{(2)}{}_{,j}$ 的步骤基本相同。提醒一下，在第 3 节中，我们将$E_{k,j}$ 分解为扩散 B 样条基的乘积:

$$\mathcal{D}_{k,j}(x_i,t)=\int\frac{\mathcal{N}(2^kx_i-j_i)}{\sigma_t\sqrt{2\pi}}\exp\left(-\frac{(x_i-m_ty_i)^2}{2\sigma_t^2}\right)\mathrm{d}x_i.$$

尽管我们处理 ed 与第 3 节中的基本相同；此处我们更正式地先截断积分 al

区间。我们将积分区间截取为

$$
\begin{aligned}
E_{k,j}^{(1)}(x,t) &\coloneqq \int_{y\in A^{x,t}} \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\mathbb{I}[\|y\|_\infty \le C_{\mathrm{b},1}] M_{k,j}^d(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y \\
&= \prod_{i=1}^d \left(\sum_{l'=0}^{l+1}\frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!}\int_{y_i\in a_i^x}\frac{1}{\sigma_t(2\pi)^{\frac12}}\mathbb{I}[|y_i|\le C_{\mathrm{b},1}]\mathbb{I}[0\le 2^{k_i}y_i-j_i\le l+1]\right. \\
&\qquad\left.\times (2^k y_i - l' - j_i)_+^l \exp\left(-\frac{(x_i-m_t y_i)^2}{2\sigma_t^2}\right)\mathrm{d}y_i\right),
\end{aligned}
\tag{33}
$$

其中 $A^{x,t}=\prod_{i=1}^d a^{x,t}_i$ 满足 $a^{x,t}_i=[\frac{x_i}{m_t}-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log}\,\varepsilon^{-1},\frac{x_i}{m_t}+\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log}\,\varepsilon^{-1}]$，$C_{\mathrm{f}}=\mathcal{O}(1)$，且 $0<\varepsilon<1$。根据引理 F.10 以及观察 $\mathbb{I}[\|y\|_\infty\le C_{\mathrm{b},1}]M_{k,j}^d(y)\le\big((l+1)^{l+1}2^{l+1}\big)^d$，这种截断造成的误差至多为 $\mathcal{O}(\varepsilon)$。总之，由于 $M_{k,j}^d(x)$ 是关于 $x_i(i=1,2,\cdots,d)$ 的单变量函数的乘积，现可将对 $\mathbb{R}^d$ 的积分分解为在有界区域上仅关于一个变量的积分；该积分对应的是第 3 节介绍的扩散 B 样条基的截断版本 $\mathcal{D}_{k,j}$。

现在我们通过以下引理开始形式化证明。我们近似

$$
\int_{y_i\in a_i^{x,t}}\frac{1}{\sigma_t(2\pi)^{\frac12}}\mathbb{I}[|y_i|\le C_{\mathrm{b},1}]\mathbb{I}[0\le 2^k y_i-j_i\le l+1](2^{k_i}y_i-l'-j_i)_+^l\exp\left(-\frac{(x_i-m_t y_i)^2}{2\sigma_t^2}\right)\mathrm{d}y_i \tag{34}
$$

（提醒（33））。注意 $\mathbb{I}[|y_i|\le C_{\mathrm{b},1}]\mathbb{I}[0\le 2^k y_i-j_i\le l+1]\equiv 0$ 或 $=\mathbb{I}[a\le 2^k y_i\le b]$ 成立，其中 a、b 满足

$$
-C2^k-l\le \min_i j_i\le j_i\le a<b\le j_i+l+1\le \max_i j_i+l+1\le C2^k+l+1, \tag{35}
$$

如果我们假设 $\mathrm{supp}(p_0)=[-C,C]^d$（13）（35）（34）$\mathbb{I}[|y_i|\le C_{\mathrm{b},1}]\mathbb{I}[0\le 2^k y_i-j_i\le l+1]\,(2^k y_i-l'-j_i)_+^l\not\equiv 0$（见引理 F.）。基于（）、（）（如果（））$^{l'}$，也可以改写为

$$
\begin{aligned}
&\int_{y_i\in a_i^{x,t}}\frac{1}{\sigma_t(2\pi)^{\frac12}}\mathbb{I}[\underline{j}\le 2^k y\le\overline{j}](2^k y_i-j')^l\exp\left(-\frac{(x_i-m_t y_i)^2}{2\sigma_t^2}\right)\mathrm{d}y_i, \\
&\text{with } \underline{j},\overline{j},j'\in\mathbb{R},\quad \overline{j}-l-1\le j'\le\underline{j}\le\overline{j},\quad -C2^k-l\le j',\underline{j},\overline{j}\le C2^k+l+1.
\end{aligned}
\tag{36}
$$

在以下引理中，我们考虑对（36）的近似。为简洁起见，我们省略坐标的下标 i。此外，（j' 36）中的 j记作 $j\in\mathbb{R}^d$，因为在以下引理中不会用到。

引理 B.2（扩散 B 样条基的近似）。设 j、k、$l\in\mathbb{Z}$、$\underline{j}$、$\overline{j}\in\mathbb{R}$ 满足 $\overline{j}-l-1\le j\le\underline{j}\le\overline{j}$、$-C2^k-l\le j$、$\underline{j}$、$\overline{j}\le C2^k+l+1$，且 k、$l\ge 0$。假设 $|\sigma'-\sigma_t|$、$|m'-m_t|\le\varepsilon_{\mathrm{error}}$，并任取 ε，使其取自 $0<\varepsilon<\frac12$ 和 $C>0$。那么，存在一个神经网络 $\phi^{j,\overline{j},\underline{j},k}_{\mathrm{dif},\overline{1}}\in\Phi(L,W,S,B)$，满足

$$
\begin{aligned}
L &= \mathcal{O}(\log^4\varepsilon^{-1}+\log^2 C+k),\\
\|W\|_\infty &= \mathcal{O}(\log^6\varepsilon^{-1}),\\
S &= \mathcal{O}(\log^8\varepsilon^{-1}+\log^2 C+k),\\
B &= \mathcal{O}(C^l 2^{kl})+\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}.
\end{aligned}
$$

使得

$$
\begin{aligned}
&\left|\phi^{j,\overline{j},\underline{j},k}_{\mathrm{dif},\overline{1}}(x,\sigma',m')-\int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}\frac{1}{\sqrt{2\pi}\sigma_t}\mathbb{I}[\underline{j}\le 2^k y\le\overline{j}](2^k y-j)^l\exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right)\mathrm{d}y\right| \\
&\le\tilde{\mathcal{O}}(\varepsilon)+\varepsilon_{\mathrm{error}}C^{4l}2^{k(4l+1)}\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}.
\end{aligned}
$$

对所有 x 属于 $-C\le x\le C$ 以及所有 $t\ge\varepsilon$均成立。

此外，在相同条件下，存在一个神经网络 $\phi^{j,\ \overline{j},\ \underline{j},\ k\,\mathrm{dif}\,\overline{2}} \in \Phi(L,W,S,B)$，其关于 L、$\|W\|_\infty$、S、B 的界与上述相同，使得

$$\left|\phi_{\mathrm{dif},\overline{2}}^{j,\overline{j},\underline{j},k}(x,\sigma',m') - \int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}} \frac{[x-m_t y]_i}{\sqrt{2\pi}\sigma_t^2}\,\mathbb{I}[\underline{j}\le 2^k y\le\overline{j}](2^k y-j)^l \exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right)\mathrm{d}y\right|$$
$$\le \tilde{\mathcal{O}}(\varepsilon)+\varepsilon_{\mathrm{error}} C^{4l}2^{k(4l+1)}\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}.$$

对所有 $x\in -C\le x\le C$ 以及所有 $t\ge\varepsilon$，均成立。

此外，我们可以选取这些网络，使得 $\|\phi^{j,\ \overline{j},\ \underline{j},\ k\,\mathrm{dif}\,\overline{1}}\|_\infty$、$\|\phi^{j,\ \overline{j},\ \underline{j},\ k\,\mathrm{dif}\,\overline{2}}\|_\infty = \mathcal{O}(1)$ 成立。

证明。这里我们只考虑 $\phi^{j,\ \overline{j},\ \underline{j},\ k\,\mathrm{dif},\ 1}$，因为关于 $\phi^{j,\ \overline{j},\ \underline{j},\ k\,\mathrm{dif},\ 2}$ 的断言，基本上可由关于 $\phi^{j,\ \overline{j},\ \underline{j},\ k\,\mathrm{dif},\ 1}$ 的论证推出。

首先，我们在闭区间内用次数至多为 $\mathcal{O}(\log\varepsilon^{-1})$ 的多项式来逼近指数函数。注意到 $\mathbb{I}[\underline{j}\le 2^k y\le\overline{j}](2^k y-j)^l$ 有界于 $(l+1)^l$，这是由 $\overline{j}-l-1\le j\le\underline{j}\le\overline{j}$ 的假设得出的。因此，根据引理 F.11，存在 $S=\mathcal{O}(\log\varepsilon^{-1})$，且有

$$\left|\exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right)-\sum_{s=0}^{S-1}\frac{(-1)^s}{s!}\frac{(x-m_t y)^{2s}}{2^s\sigma_t^{2s}}\right|\le\varepsilon^2$$

对于所有 $y\in[-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log}\ \varepsilon^{-1}+x,\ \frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log}\ \varepsilon^{-1}+x]$，我们有

$$\begin{aligned}&\left|\int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}\frac{1}{\sqrt{2\pi}\sigma_t}\mathbb{I}[\underline{j}\le 2^k y\le\overline{j}](2^k y-j)^l\exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right)\mathrm{d}y\right.\\&\quad\left.-\int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}\frac{1}{\sqrt{2\pi}\sigma_t}\mathbb{I}[\underline{j}\le 2^k y\le\overline{j}](2^k y-j)^l\left(\sum_{s=0}^{S-1}\frac{(-1)^s}{s!}\frac{(x-m_t y)^{2s}}{2^s\sigma_t^{2s}}\right)\mathrm{d}y\right|\\&\le\max\left\{\frac{2\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}},(l+1)\right\}\cdot\frac{1}{\sqrt{2\pi}\sigma_t^2}(l+1)^l\cdot\varepsilon\lesssim\varepsilon\log^{\frac{1}{2}}\varepsilon^{-1}.\end{aligned}$$

这里，$\frac{2\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\quad\varepsilon^{-1}}$ $l+1$ $\mathbb{I}[\underline{j}\le 2^k y\le\log$ 来自积分区间的长度，而 则来自其中 $\overline{j}]=1$ 成立的区间。

现在我们只需要近似计算闭区间上多项式的积分：

$$\begin{aligned}&\sum_{s=0}^{S-1}\int_{-\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log\varepsilon^{-1}}+\frac{x}{m_t}}\frac{1}{\sqrt{2\pi}\sigma_t}\mathbb{I}[\underline{j}\le 2^k y\le\overline{j}](2^k y-j)^l\cdot\frac{(-1)^s}{s!}\frac{(x-m_t y)^{2s}}{2^s\sigma_t^{2s}}\mathrm{d}y\\&=\sum_{s=0}^{S-1}\sum_{l'=0}^{l}\frac{-(-1)^{s+l}}{\sqrt{2\pi}m_t^{l+1}s!2^s}\left[{}_l\mathrm{C}_{l'}(2^k\sigma_t)^{l'}(jm_t-2^k x)^{l-l'}\int_{-C_{\mathrm{f}}\sqrt{\log\varepsilon^{-1}}}^{C_{\mathrm{f}}\sqrt{\log\varepsilon^{-1}}}\mathbb{I}\left[\frac{x-m_t 2^{-k}\overline{j}}{\sigma_t}\le y\le\frac{x-m_t 2^{-k}\underline{j}}{\sigma_t}\right]y^{l'+2s}\mathrm{d}y\right]\\&\qquad\left(\text{by resetting } y\leftarrow\frac{x-m_t y}{\sigma_t}\right)\\&=\sum_{s=0}^{S-1}\sum_{l'=0}^{l}\frac{-(-1)^{s+l}{}_l\mathrm{C}_{l'}2^{kl'}\sigma^{l'}(jm_t-2^k x)^{l-l'}}{\sqrt{2\pi}m_t^{l+1}s!2^s(l'+2s+1)}\left[\left(\min\left\{C_{\mathrm{f}}\sqrt{\log\varepsilon^{-1}},\max\left\{\frac{x-m_t 2^{-k}\underline{j}}{\sigma_t},-C_{\mathrm{f}}\sqrt{\log\varepsilon^{-1}}\right\}\right\}\right)^{l'+2s+1}\right.\\&\qquad\left.-\left(\min\left\{C_{\mathrm{f}}\sqrt{\log\varepsilon^{-1}},\max\left\{\frac{x-m_t 2^{-k}\overline{j}}{\sigma_t},-C_{\mathrm{f}}\sqrt{\log\varepsilon^{-1}}\right\}\right\}\right)^{l'+2s+1}\right].\end{aligned}\tag{37}$$

为方便起见，我们将（37）分解为以下子模块。令

$$
\begin{aligned}
f_1^{l',s}(x,\sigma,m) &= (\min\{C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1}),\max\{\frac{x-m2^{-k}\underline{j}}{\sigma},-C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1})\}\})^{l'+2s+1},\\
f_2^{l',s}(x,\sigma,m) &= (\min\{C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1}),\max\{\frac{x-m2^{-k}\overline{j}}{\sigma},-C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1})\}\})^{l'+2s+1},\\
f_3^{l',s}(x,\sigma,m) &= f_1^{l',s}(x,\sigma,m)-f_2^{l',s}(x,\sigma,m)\\
f_4^{l'}(x,m) &= (jm-2^kx)^{l-l'},\\
f_5^{l'}(\sigma) &= \sigma^{l'},\\
f_6(m) &= m^{-(l+1)},\\
f_7^{l',s}(x,\sigma,m) &= f_3^{l',s}(x,\sigma,m)f_4^{l'}(x,m)f_5^{l'}(\sigma)f_6(m).
\end{aligned}
$$

它们还取决于 j、$\underline{j}$、$\overline{j}$、k 和 l，但为简洁起见，我们省略对这些变量的依赖。我们取某个 $\varepsilon_1>0$，并在证明的最后部分对其进行调整。

我们首先考虑对 $f^{l',1,s}(x,\sigma,m)$ 的近似。我们将其实现为

$$
\begin{aligned}
&f_1^{l',s}(x,\sigma,m)\coloneqq\phi_1^{l',s}(x,\sigma,m)\\
&:=\phi_{\mathrm{mult}}(\cdot;l'+2s+1)\circ\phi_{\mathrm{clip}}(\cdot;-C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1}),-C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1}))\circ(\phi_{\mathrm{mult}}(x-m2^{-k}\underline{j},\phi_{\mathrm{rec}}(\sigma))).
\end{aligned}
$$

通过在推论 F.8 中令 $\varepsilon=\min\{\sigma_\varepsilon,\varepsilon_1\}$，在引理 F.6 中令 $\phi_{\mathrm{rec}},\varepsilon=\varepsilon_1 C=\max\{2C+l+1,\sigma_\varepsilon^{-1}\}\geq\max\{|x|+m2^{-k}\underline{j},\sigma_\varepsilon^{-1}\}$（对应第一个 ϕ_{mult}），在引理 F.4 中令 $a=-C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1})$ $b=C_{\mathrm{f}}\log^{\frac{1}{2}}(\varepsilon^{-1})$（对应 ϕ_{clip}），并在引理 F.6 中令 $\varepsilon=\varepsilon_1$，$C=C_{\mathrm{f}}\log^{\frac{1}{2}}(2\varepsilon^{-1})$（对应第二个 ϕ_{mult}）。注意，$\sigma_\varepsilon\simeq\sqrt{\varepsilon}$。然后，利用引理 F.1、F.4、F.6 和 F.7，网络的规模至多为

$$
\begin{aligned}
&L=\mathcal{O}(\log^2\varepsilon_1^{-1}+\log^2\varepsilon^{-1}+\log^2 C),\\
&\|W\|_\infty=\mathcal{O}(\log^3\varepsilon_1^{-1}+\log^3\varepsilon^{-1}),\\
&S=\mathcal{O}(\log^4\varepsilon_1^{-1}+\log^4\varepsilon^{-1}+\log^2 C),\\
&B=\mathcal{O}(\varepsilon_1^{-2}+C^2)+\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}.
\end{aligned}
\tag{38}
$$

$f^{l',1,s}(x,\sigma_t,m_t)$ 与 $\phi^{l',1,s}(x,\sigma',m')$ 之间的近似误差有界于

$$
\begin{aligned}
&\varepsilon_1+\mathcal{O}(\log\varepsilon^{-1})(C_{\mathrm{f}}\log^{\frac{1}{2}}\varepsilon^{-1})^{\mathcal{O}(\log\varepsilon^{-1})}\cdot(\varepsilon_1+\max\{C+l+2,\sigma_\varepsilon^{-1}\}^2\cdot(\varepsilon_1+\varepsilon_{\mathrm{error}}(\varepsilon_1^{-2}+\varepsilon^{-2})))\\
&=(\varepsilon_1+\varepsilon_{\mathrm{error}})\left(\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}+C^2\right).
\end{aligned}
$$

$f^{l',2,s}(x,\sigma_t,m_t)$ 也以相同方式进行近似，因此，将 $f^{l',1,s}(x,\sigma_t,m_t)$ 与 $f^{l',2,s}(x,\sigma_t,m_t)$ 聚合起来（使用引理 F.3），可得 $f^{l',3,s}(x,\sigma_t,m_t)$ 可由 $\phi^{l',3,s}(x,\sigma',m')$ 近似，误差至多为附加误差 $(\varepsilon_1+\varepsilon_{\mathrm{error}})\left(\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}+C^2\right)$，且所用神经网络的规模与式 (38) 相同。

接下来，我们考虑 $f_4^{l'}(x,m_t)$。由于 $2^kx=\mathcal{O}(C2^k)$ 和 $|jm_t-jm'|\leq\mathcal{O}(C2^k\varepsilon_{\mathrm{error}})$，我们使用神经网络 $\phi_4^{l'}(x,m')\in\Phi(L,W,S,B)$ 来近似 $f_4^{l'}(x,m_t)$，其中 $L,\|W\|_\infty,S,B$ 由引理 F.1 和 F.6（令 $\varepsilon=\varepsilon_1,C=\mathcal{O}(C2^k)$）计算得出，如下

$$
L=\mathcal{O}(\log\varepsilon_1^{-1}+k\log C),\quad W=\mathcal{O}(1),\quad S=\mathcal{O}(\log\varepsilon_1^{-1}+k\log C),\quad B=\mathcal{O}(C^l2^{kl}).
$$

之间的近似误差 $f_4^{l'}(x,\ m_t)$ 和 $\phi_4^{l'}(x,\ m')$ 可界定为 $\varepsilon_1+\mathcal{O}(C^l2^{kl})\varepsilon_{\mathrm{error}}$，根据引理 F.6。

对于 $f_5^{l'}(\sigma)$ 和 $f_6(m)$ 的论证，只需分别在引理 F.6 和推论 F.8 中设置适当的参数。对于 $f_5^{l'}(\sigma_t)$，通过在引理 F.6 中设置 $d=l'(\leq l)$ $\varepsilon=\varepsilon_1$，存在一个神经网络 $\phi_5^{l'}(\sigma')$，其 $L=\mathcal{O}(\log\varepsilon_1^{-1})$ $\|W\|_\infty=48l,S=\mathcal{O}(\log\varepsilon_1^{-1})$ $B=1$，且 $f_5^{l'}(\sigma)$ 与 $\phi_5^{l'}(\sigma')$ 之间的近似误差有界于 $\varepsilon_1+l\varepsilon_{\mathrm{error}}$。对于 $f_6(m_t)$，存在一个神经网络 $\phi_6(m')$，其 $L=\mathcal{O}(\log^2\varepsilon_1^{-1}+\log^2 m_\varepsilon^{-1})$ $\|W\|_\infty=\mathcal{O}(\log^3\varepsilon_1^{-1}+\log^3 m_\varepsilon^{-1})$ $S=$

$\mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^4 m_\varepsilon^{-1})$ $B = \mathcal{O}(\varepsilon_1^{-l-1} + \underline{m}^{-l-1})$，且 $f_6(m_t)$ 与 $\phi_6(m')$ 之间的近似误差有界于 $\varepsilon_1 + (l+1)\varepsilon_1^{-l-2}\varepsilon_{\text{error}} + (l+1)$ $m_\varepsilon^{-l-2}\varepsilon_{\text{error}}$。通过在推论 F.8 中设置 $d = l+1, \varepsilon = \min\{\varepsilon_1, m_\varepsilon\}$，即可得到上述结论。注意，$m_\varepsilon \gtrsim 1$。

因此，根据引理 F.6 和 $\varepsilon = \varepsilon_1$，存在一个神经网络 $\phi^{l',\ 7\ s}(x, m, \sigma)$，使得

$$\begin{aligned} L &= \mathcal{O}(\log^2 \varepsilon_1^{-1} + \log^2 \varepsilon^{-1} + \log^2 C + k), \\ \|W\|_\infty &= \mathcal{O}(\log^3 \varepsilon_1^{-1} + \log^3 \varepsilon^{-1}), \\ S &= \mathcal{O}(\log^4 \varepsilon_1^{-1} + \log^4 \varepsilon^{-1} + \log^2 C + k), \\ B &= \mathcal{O}(\varepsilon_1^{-2} + C^2) + \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1} + C^l 2^{kl}. \end{aligned}$$

其中，$f^{l',\ 7\ s}(x, m_t, \sigma_t)$ 与 $\phi^{l',\ 7\ s}(x, m', \sigma')$ 之间的近似误差有如下界：

$$\left| f_7^{l',s}(x, \sigma, m) - \phi_7^{l',s}(x, m', \sigma') \right| \le (\varepsilon_1 + \varepsilon_{\text{error}}(\varepsilon_1^{-l-2} + C^{4l} 2^{4kl})) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}.$$

最后，根据式 (37) 并利用引理 F.3，我们对 (l', s) 上的 $\phi^{l',\ 7\ s}(x, m', \sigma')$ 乘以 $\frac{-\ -1\ \ s+l}{\sqrt{2\pi}s\ 2^s\ \ l'+2s+2}$( ) C ! ( ) 求和。这里，系数有界于 $2^{(k+1)l}$，可能组合 (l', s) 的总数有界于 $\mathcal{O}(lS) = \mathcal{O}(\log \varepsilon^{-1})$。于是，式 (37) 的近似误差有如下界：

$$2^{(k+1)l}(\varepsilon_1 + \varepsilon_{\text{error}}(\varepsilon_1^{-l-2} + C^{4l} 2^{4kl})) \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}.$$

为了将与 ε_1 相关的项界定为 $\mathcal{O}(\varepsilon)$，我们取 $\varepsilon_1 = \mathcal{O}(2^{-(k+1)l} \log^{-\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1})$。于是，总近似误差有界为 $\tilde{\mathcal{O}}(\varepsilon)$ $+ \varepsilon_{\text{error}} C^{4l} 2^{k(4l+1)} \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}$，且该误差可由一个神经网络实现，该网络具有

$$\begin{aligned} L &= \mathcal{O}(\log^4 \varepsilon^{-1} + \log^2 C + k), \\ \|W\|_\infty &= \mathcal{O}(\log^6 \varepsilon^{-1}), \\ S &= \mathcal{O}(\log^8 \varepsilon^{-1} + \log^2 C + k), \\ B &= \mathcal{O}(C^l 2^{kl}) + \log^{\mathcal{O}(\log \varepsilon^{-1})} \varepsilon^{-1}. \end{aligned}$$

最后，由于

$$\begin{aligned} &\left| \int_{-\frac{\sigma_t C_{\mathrm{f},1}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m}}^{\frac{\sigma_t C_{\mathrm{f}}}{m_t}\sqrt{\log \varepsilon^{-1}} + \frac{x}{m_t}} \frac{1}{\sqrt{2\pi}\sigma_t} \mathbb{I}[\underline{j} \le 2^k y \le \overline{j}](2^k y - j)^l \exp\left(-\frac{(x - m_t y)^2}{2\sigma_t^2}\right) \mathrm{d}y \right| \\ &\le \int \frac{1}{\sqrt{2\pi}\sigma_t} \mathbb{I}[\underline{j} \le 2^k y \le \overline{j}](l+1)^l \exp\left(-\frac{(x - m_t y)^2}{2\sigma_t^2}\right) \mathrm{d}y \lesssim C_f, \end{aligned}$$

我们可以截断 $\phi^{j}_{,\overline{j},\underline{j},k\ \mathrm{dif},1}$，使其有界于 $\mathcal{O}(1)$。 □

现在我们来近似（修改后的）张量积扩散B样条基。以下给出引理 3.4 的正式表述。去掉 $\mathbb{I}[\|y\|_\infty \le C_{\mathrm{b},1}]$ 这一项后，该表述与引理 3.4 一致。这个网络 $\phi_{\mathrm{dif},3}$ 对应于引理 ϕ_{TDB} 中的 3.4。

引理 B.3（张量积扩散 B 样条基的近似）。令 $k \in \mathbb{Z}_+$、$j \in \mathbb{Z}^d$、$l \in \mathbb{Z}_+$ with $-C2^k - l \le j_i \le C2^k (i = 1, 2, \cdots, d)$、$\varepsilon (0 < \varepsilon < \frac{1}{2})$ 和 $C > 0$。存在一个神经网络 $\phi_{\mathrm{dif},\ 3}(x, t) \in \Phi(L, W, S, B)$，满足

$$\begin{aligned} L &= \mathcal{O}(\log^4 \varepsilon^{-1} + \log^2 C + k^2), \\ \|W\|_\infty &= \mathcal{O}(\log^6 \varepsilon^{-1} + \log^3 C + k^3), \\ S &= \mathcal{O}(\log^8 \varepsilon^{-1} + \log^4 C + k^4), \\ B &= \exp\left(\log^4 \varepsilon^{-1} + \log C + k\right), \end{aligned}$$

使得

$$\left|\phi^{k,j}_{\mathrm{dif},3}(x,t)-\int_{\mathbb{R}^d}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\mathbb{I}[\|y\|_\infty\le C_{\mathrm{b},1}]M^d_{k,j}(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y\right|\le\varepsilon$$

对所有 $x\in[-C,C]^d$ 均成立。

此外，在相同条件下，存在一个神经网络 $\phi_{\mathrm{dif},4}\in\Phi(L,W,S,B)$，其$L$、$\|W\|_\infty$、$S$、$B$的界与上文相同，并且满足

$$\left\|\phi^{k,j}_{\mathrm{dif},4}(x,\sigma',m')-\int_{\mathbb{R}^d}\frac{x-m_ty}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}}\mathbb{I}[\|y\|_\infty\le C_{\mathrm{b},1}]M^d_{k,j}(y)\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\mathrm{d}y\right\|\le\varepsilon.$$

对所有 $x\in[-C,C]^d$ 均成立。

此外，我们可以选择这些网络，使得 $\|\phi^{k,j\ \mathrm{dif}\ 3}\|_\infty,\ \|\phi^{k,j\ \mathrm{dif}\ 4}\|_\infty=\mathcal{O}(1)$ 成立。

证明。这里我们只证明第一部分，因为第二部分可用同样的方法推出。我们假设 $|\sigma'-\sigma_t|\vee|m'-m_t|\le\varepsilon_{\mathrm{error}}$。

根据讨论（33），我们近似

$$\prod_{i=1}^d\left(\sum_{l'=0}^{l+1}\frac{(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}}{l!}\int_{y_i\in a_i^x}\frac{1}{\sigma(2\pi)^{\frac12}}\mathbb{I}[|y_i|\le C_{\mathrm{b},1}]\mathbb{I}[0\le2^ky_i-j_i\le l+1]\right.$$
$$\left.\times(2^{k_i}y_i-l'-j_i)_+^l\exp\left(-\frac{(x_i-my_i)^2}{2\sigma^2}\right)\mathrm{d}y_i\right),\qquad(39)$$

它等于$D^d_{k,\ j}(x)$误差在加性意义上不超过$\mathcal{O}(\varepsilon)$，因此我们近似（39）。这里$a_i^x=[\frac{x_i}{m_t}-\frac{\sigma_tC_{\mathrm{f}}}{m_t}\sqrt{\text{对数}\,\varepsilon^{-1}},\frac{x_i}{m_t}+\frac{\sigma_tC_{\mathrm{f}}}{m_t}\sqrt{\text{对数}\,\varepsilon^{-1}}]$。

我们令 $f_i(y_i;\ j_i,k,l'):=\mathbb{I}[|y_i|\le C_{\mathrm{b},1}]\mathbb{I}[0\le2^ky_i-j_i\le l+1](2^ky_i-l'-j_i)_+^l\exp\left(-\frac{(x_i-m_ty_i)}{2\sigma_t^2}{}^2\right)\mathrm{d}y_i$。

首先，$\sum_{l'=0}^{l+1}(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}{}^{l'}!\,f_i(y_i;\ j_i,k,l')$ 由 $\sum_{l'=0}^{l+1}(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}{}^{l'}!\,\phi^{j_i-l',\bar j_{l'},\underline{j}_{l'},k}\mathrm{dif},1\ (y_i,\sigma',m')$ 近似（网络的聚合见引理 F.3）。这里，$\bar j_{l'}$和 $\underline{j}_{l'}$的定义使得 $\mathbb{I}[\underline{j}_{l'}\le2^ky\le\bar j_{l'}]=\mathbb{I}[|y_i|\le C_{\mathrm{b},1}]\mathbb{I}[0\le2^ky_i-j_i\le l+1]$ 成立。

现在我们将 $\sum_{l'=0}^{l+1}(-1)^{l'}{}_{l+1}\mathrm{C}_{l'}{}^{l'}!\,\phi^{j_i,\bar j_{l'},\underline{j}_{l'},k}\mathrm{dif},1\ (y_i,\sigma',m')$ 与 $i=1,2,\cdots,d$ 使用 ϕ_{mult} 相乘，以得到所需的网络$\phi^{k,j}_{\mathrm{dif},3}$。根据引理 B.2（使用 $\varepsilon=\varepsilon$）以及引理 F.6（使用 $\varepsilon=\varepsilon$和 $C=\mathcal{O}(1)$）（因为 $\|\phi^{j_i,\bar j_{l'},\underline{j}_{l'},k}\mathrm{dif}_1\|_\infty=\mathcal{O}(1)$），存在一个神经网络 $\phi_1(x,m',\sigma')\in\Phi(L,W,S,B)$，满足

$$\begin{aligned}L&=\mathcal{O}(\log^4\varepsilon^{-1}+\log^2C+k),\\\|W\|_\infty&=\mathcal{O}(\log^6\varepsilon^{-1}),\\S&=\mathcal{O}(\log^8\varepsilon^{-1}+\log^2C+k),\\B&=\mathcal{O}(C^l2^{kl})+\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}\end{aligned}$$

并且我们可以界定 ϕ_1（x，m'，σ'）与（39）之间的近似误差，其上界为

$$\tilde{\mathcal{O}}(\varepsilon)+\varepsilon_{\mathrm{error}}C^{4l}2^{k(4l+1)}\log^{\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}.\qquad(40)$$

现在，我们考虑 $\phi_{\mathrm{dif},3}=\phi_1(x,\phi_m(t)\ \phi_\sigma(t))$。我们应用引理 B.1 并取 $\varepsilon=C^{-4l}2^{-k(4l+1)}\log^{-\mathcal{O}(\log\varepsilon^{-1})}\varepsilon^{-1}$，因此

t ˜hat $\varepsilon_{\text{error}}$ 足够小，且 (40) 被 $\mathcal{O}(\varepsilon)$ 所界。然后，$\phi_{\text{dif},3}$ 的大小被 y

$$
\begin{aligned}
L &= \mathcal{O}(\log^4 \varepsilon^{-1} + \log^2 C + k^2),\\
\|W\|_\infty &= \mathcal{O}(\log^6 \varepsilon^{-1} + \log^3 C + k^3),\\
S &= \mathcal{O}(\log^8 \varepsilon^{-1} + \log^4 C + k^4),\\
B &= \exp\left(\log^4 \varepsilon^{-1} + \log C + k\right).
\end{aligned}
$$

˜现在，调整 ε 以将 $\mathcal{O}(\varepsilon)$ 替换为 ε，即可得到第一个断言。

We can make $\|\phi^{k,j}_{\text{dif},3}\|_\infty$ hold, because $\int_{\mathbb{R}^d} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \mathbb{I}[\|y\|_\infty \le C_{\text{b},1}] M^d_{k,j}(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y = \mathcal{O}(1)$.

□

B.3. 基于 p_0 的近似误差界

现在我们将所有结果结合起来，推导出定理 3.1。在本小节及下一小节中，我们取 $N \gg 1, T_1 = \underline{T} = \text{poly}(N^{-1})$ 以及 $T_5 = \overline{T} = \mathcal{O}(\log N)$。此外，令 $T_2 = N^{-(2-\delta)/d}, T_3 = 2T_2, T_4 = 3T_2$。本小节讨论 $t \in [T_1, T_4]$ 的近似。

我们从以下引理开始，该引理给出了 Besov 函数的基分解 .

引理 B.4 （基分解）。在 $N \gg 1$、假设 *2.4*、*2.5*、*2.6* 及 $a_0 = N^{-(1-\delta)/d}$下，存在 f_N 满足

$$
\begin{aligned}
&\|p_0 - f_N\|_{L^2([-1,1]^d)} \lesssim N^{-s/d},\\
&\|p_0 - f_N\|_{L^2([-1,1]^d \setminus [-1+N^{-(1-\delta)/d}, 1-N^{-(1-\delta)/d}]^d)} \lesssim N^{-(3s+2)/d},
\end{aligned}
$$

且对于所有满足 x 的 $f_N(x) = 0$ $\|x\|_\infty \ge 1$，并具有以下形式：

$$
f_N(x) = \sum_{i=1}^{N} \alpha_i \mathbb{I}[\|x\|_\infty \le 1] M^d_{k,j_i}(x) + \sum_{i=N+1}^{3N} \alpha_i \mathbb{I}[\|x\|_\infty \le 1 - N^{-(1-\delta)/d}] M^d_{k,j_i}(x), \tag{41}
$$

其中 $-2^{(k)_m} - l \le (j_i)_m \le 2^{(k)_m} (i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$，$|k| \le K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$，对于 $\delta = d(1/p - 1/r)_+$ 和 $\nu = (2s - \delta)/(2\delta)$。此外，$|\alpha_i| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)_+}$。

证明。由于 $p_0 \in \mathcal{C}^{3s+2}([-1,1]^d \setminus [-1 + N^{-(1-\delta)/d}, 1 - N^{-(1-\delta)/d}]^d)$，根据引理 F.13，我们有 f_1 使得

$$
\|p_0 - f_1\|_{L^2([-1,1]^d \setminus [-1+N^{-(1-\delta)/d}, 1-N^{-(1-\delta)/d}]^d)} \lesssim N^{-(3s+2)/d}.
$$

且具有以下形式：

$$
f_1(x) = \sum_{i=1}^{N} \alpha_i M^d_{k,j_i}(x),
$$

其中 $-2^{(k)_m} - l \le (j_i)_m \le 2^{(k)_m} (i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$，$|k| \le K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ 对于 $\delta = d(1/p - 1/r)_+$ 和 $\nu = (2s - \delta)/(2\delta)$ 成立。此外，$|\alpha_{1,\ i}| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-2s)_+}$。

接下来，我们来近似 $f \in [-1,1]^d$。由于 $\|p_0\|_{B^s_{p,\ q}} \lesssim 1$，我们有 f_2 使得

$$
\|p_0 - f_2\|_{L^2([-1,1]^d)} \lesssim N^{-s/d}.
$$

其形式如下：

$$
f_2(x) = \sum_{i=N+1}^{2N} \alpha_i M^d_{k,j_i}(x),
$$

其中 $-2^{(k)_j} - l \le (j_i)_j \le 2^{(k)_j} (i = 1, 2, \cdots, N,\ j = 1, 2, \cdots, d)$，$|k| \le K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1}\log N)$，对于 $\delta = d(1/p - 1/r)_+$ 和 $\nu = (s-\delta)/(2\delta)$。此外，$|\alpha_{2,i}| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)}$。

因此，

$$
\begin{aligned}
&\mathbb{I}[\|x\|_\infty \le 1] f_1(x) - \mathbb{I}[\|x\|_\infty \le 1 - N^{-(1-\delta)/d}] f_1(x) + \mathbb{I}[\|x\|_\infty \le 1 - N^{-(1-\delta)/d}] f_2(x) \\
&= \sum_{i=1}^{N} \alpha_i M^d_{k_i,j_i}(x) - \sum_{i=1}^{N} \alpha_i \mathbb{I}[\|x\|_\infty \le 1 - N^{-(1-\delta)/d}] M^d_{k_i,j_i}(x) + \sum_{i=N+1}^{2N} \alpha_i \mathbb{I}[\|x\|_\infty \le 1 - N^{-(1-\delta)/d}] M^d_{k_i,j_i}(x)
\end{aligned}
$$

成立，对基底重新编号即可得到结论。 □

下面的引理给出了一个在 $[T_1, T_4]$ 中近似 $\nabla \log p_t(x)$ 的神经网络。

引理 B.5（$T_1 \le t \le T_4$ 的得分函数近似）。存在一个满足以下条件的神经网络 $\phi_{\text{score},1} \in \Phi(L, W, S, B)$

$$
\int p_t(x) \|\phi_{\text{score},1}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t \lesssim \frac{N^{-2s/d} \log N}{\sigma_t^2} \tag{42}
$$

其中，$L, \|W\|_\infty, S, B$ 的取值为

$$
L = \mathcal{O}(\log^4 N), \quad \|W\|_\infty = \mathcal{O}(N \log^6 N), \quad S = \mathcal{O}(N \log^8 N), \quad \textit{and } B = \exp(\mathcal{O}(\log^4 N)).
$$

证明。 在进入证明的主要部分之前，我们将讨论范围限制在有界区域内。根据引理A.4，我们有

$$
\int_{\|x\|_\infty \ge m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x) \|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{T}{N^{(2s+1)/d}} \left(1 + \|s(\cdot,t)\|_\infty^2\right), \tag{43}
$$

1，其中 $\mathcal{O}(1)$ 中的隐藏常数足够大。由于 $\|\nabla \log p_t(x)\|$ 在 $\log_{\sigma_t}^{\frac{s}{2}} N$ 属于 $\|x\|_\infty \ge m_t + 1\mathcal{O}(1)\sigma_t\sqrt{\log N}$ 时是有界的，根据引理 A.3，s 可以取为使得 $\|s(\cdot,t)\|_\infty \lesssim \log_{\sigma_t}^{\frac{s}{2}} N$，因此 (43) 被 $\frac{T}{N^{(2s+1)}} \cdot \log_T^N = N^{-(2s+1)/d} \log N$ 所界定，而该值小于 (42) 的上界。因此，我们可以专注于在 $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$ 内对得分函数 $\nabla \log p_t(x)$ 的近似。此外，我们还可以排除 $p_t(x) \le N^{-(2s+1)/d}$ 的情况，因为引理 A.4 可以界定误差

$$
\begin{aligned}
\int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x) \mathbb{I}[p_t(x) \le \varepsilon] \|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x &\lesssim \frac{\varepsilon}{\sigma_t^2} \log^{\frac{d+2}{2}}(\varepsilon^{-1}T^{-1}) + \varepsilon \|s(x,t)\| \\
&\lesssim \frac{\varepsilon}{\sigma_t^2} \log^{\frac{d+2}{2}}(\varepsilon^{-1}T^{-1}) + \frac{\varepsilon}{\sigma_t^2} \log N,
\end{aligned} \tag{44}
$$

并设定 $\varepsilon = N^{-(2s+1)/d}$ 使得（44）小于界（42）。

因此，在下文中，我们考虑满足 $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$ 且 $p_t(x) \ge N^{-(2s+1)/d}$ 成立的 x。在这种情况下，我们有 $\|\nabla \log p_t(x)\| \lesssim \log_{\sigma_t}^{\frac{s}{2}} N$。

构造很直接。根据引理 B.4 中的 (41)，令

$$
\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \coloneqq \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\
&= \sum_{i=1}^{N} \alpha_i E^{(1)}_{k_i,j_i}(x,t) =: \tilde{f}_1(x,t), \\
f_1(x,t) &:= \tilde{f}_1(x,t) \vee N^{-(2s+1)/d},
\end{aligned}
$$

and

$$\sigma_t\nabla p_t(x) = \int \frac{x-m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} p_0(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y \eqqcolon \int \frac{x-m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f_N(y)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\mathrm{d}y$$
$$= \sum_{i=1}^{N}\alpha_i E^{(2)}_{k_i,j_i}(x,t) =: f_2(x,t),$$
$$f_3(x,t) := \frac{f_2(x,t)}{f_1(x,t)}\mathbb{I}\left[\left\|\frac{f_2(x,t)}{f_1(x,t)}\right\| \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}\right]$$

使得 α_i, $E^{(1)}_{k_i},_{j_i}(x,t)$ 和 $E^{(2)}_{k_i},_{j_i}(x,t)$ 与引理 B.4 中的基分解相对应。因此，$|\alpha_i| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)^+}$ 以及 $|k_i| = \mathcal{O}(\log N)$。我们说明，$C_{\mathrm{b}}$,1 在 (31) 和 (32) 中设为 1 或 $1-N^{-(1-\delta)/d}$。我们根据引理 B.3，通过令 $\varepsilon=\varepsilon_1$ 和 $C = m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$（因为 $\sigma_t \le \sigma_{T_2} \lesssim \log^{-\frac{1}{2}} N$），用 $\phi^{k_i},_{j_i}$dif,3 和 $\phi^{k_i},_{j_i}$dif,4 近似 $E^{(1)}_{k}$,$_{j_i}$ 和 $E^{(2)}_{k}$,$_{j_i}$，其中 $\varepsilon_1 =$poly(N^{-1}) 是一个将在下文调整的标量值。然后，我们利用引理 F.3 将这些子网络相加，得到神经网络 ϕ_{dif},5(x,t) 和 ϕ_{dif},6(x,t)，分别近似 $f_1(x,t)$ 和 $f_2(x,t)$。

由于我们可以将误差分解为

$$\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x)\mathbb{I}[p_t(x)\ge N^{-\frac{2s+1}{d}}]\|s(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}x$$
$$\lesssim \int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{I}[p_t(x)\ge N^{-\frac{2s+1}{d}}]p_t(x)\left\|\phi_{\mathrm{score},1}(x,t)-\frac{f_3(x,t)}{\sigma_t}\right\|^2\mathrm{d}x \tag{45}$$
$$+\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{I}[p_t(x)\ge N^{-\frac{2s+1}{d}}]p_t(x)\left\|\frac{f_3(x,t)}{\sigma_t}-\nabla\log p_t(x)\right\|^2\mathrm{d}x, \tag{46}$$

我们暂时考虑对 $f_3(x,t)/\sigma_t$ 的近似，而不是 $\nabla\log p_t(x) = \nabla p_t(x,t)/f_1(x,t)$，并对 (45) 进行界定。根据网络的构造，我们有以下界：

$$|f_1(x,t)-\phi_{\mathrm{dif},5}(x,t)| \lesssim N\cdot\max|\alpha_i|\cdot\varepsilon_1, \tag{47}$$
$$\|f_2(x,t)-\phi_{\mathrm{dif},6}(x,t)\| \lesssim N\cdot\max|\alpha_i|\cdot\varepsilon_1. \tag{48}$$

对所有 x 满足 $\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$。注意，$\max|\alpha_i|$ 有界于 $N^{(\nu^{-1}+d^{-1})(d/p-s)^+}$。因此，我们取 $\varepsilon_1 \lesssim N^{-1}\cdot N^{-(\nu^{-1}+d^{-1})(d/p-s)^+}\cdot N^{-\frac{9s+3}{d}}$，使得引理 47) 和 (48) 中的量有界于 $N^{-\frac{9s+3}{d}}$。F.6。

然后我们将 $\phi_{\mathrm{dif},\ 7}$ 定义为

$$[\phi_{\mathrm{dif},7}(x,t)]_i := \phi_{\mathrm{clip}}(\phi_{\mathrm{mult}}(\phi_{\mathrm{rec}}(\phi_{\mathrm{clip}}(\phi_{\mathrm{dif},5}(x,t);N^{-(2s+1)/d},\mathcal{O}(1)))),[\phi_{\mathrm{dif},6}(x,t)]_i);-\mathcal{O}(\log^{\frac{1}{2}}N),\mathcal{O}(\log^{\frac{1}{2}}N)).$$

以近似 $\sigma_t\nabla\log p_t(x)$。这里我们利用了 $p_t(x)$ 在 $[N^{-(2s+1)/d},\mathcal{O}(1)]$ 下的有界性来截断 ϕ_{dif},5(x,t)，并利用 $\sigma_t\nabla\log p_t(x)$ 在 $[-\mathcal{O}(\log^{\frac{1}{2}}N)\ \mathcal{O}(\log^{\frac{1}{2}}N)]$ 下的有界性来截断整个输出。对于 ϕ_{rec}，我们在引理 F.7 中令 $\varepsilon = N^{-(3s+1)/d}$；对于 ϕ_{mult}，我们令 $\varepsilon = N^{-s/d}$和 $C = N^{(2s+1)/d}$。于是，

$$\|\phi_{\mathrm{dif},7}(x,t)-f_3(x,t)\| = \left\|\phi_{\mathrm{dif},7}(x,t)-\frac{f_2(x,t)}{f_1(x,t)}\mathbb{I}\left[\left\|\frac{f_2(x,t)}{f_1(x,t)}\right\|\lesssim\frac{\log^{\frac{1}{2}}N}{\sigma_t}\right]\right\|$$
$$\lesssim N^{-s/d}+N^{(2s+1)/d}\cdot(N^{-(3s+1)/d}+N^{2(3s+1)/d}|f_1(x,t)-\phi_{\mathrm{dif},5}(x,t)|+\|f_2(x,t)-\phi_{\mathrm{dif},6}(x,t)\|)$$
$$\lesssim N^{-s/d}+N^{(8s+3)/d}|f_1(x,t)-\phi_{\mathrm{dif},5}(x,t)|+N^{(2s+1)/d}\|f_2(x,t)-\phi_{\mathrm{dif},6}(x,t)\|. \tag{49}$$

应用（47）$\le N^{-\frac{9s+3}{d}}$和（48）$\le N^{-\frac{9s+3}{d}}$可得（49）$\le N^{-\frac{s}{d}}$。

最后，我们令

$$\phi_{\mathrm{score},1}(x,t) := \phi_{\mathrm{mult}}(\phi_{\mathrm{dif},7}(x,t),\phi_\sigma(t)).$$

通过令 $\varepsilon = N^{-s/d}$以及 $C \simeq$最大值{对数$^{\frac{1}{2}} N$， σ_T} $\lesssim$多项式(N)在引理 F.6中针对 ϕ_{mult}以及 $\varepsilon = N^{-s/d}$/多项式(N)在引理B.1 中针对 ϕ_{σ}。然后，

$$\left\|\phi_{\mathrm{score},1}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\| \lesssim N^{-s/d} + \mathrm{poly}(N)\cdot N^{-s/d}/\mathrm{poly}(N) \lesssim N^{-s/d},$$

由此可得

$$(45) = \int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{I}[p_t(x) \ge N^{-\frac{2s+1}{d}}] p_t(x) \left\|\phi_{\mathrm{score},1}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\|^2 \mathrm{d}x \lesssim N^{-2s/d}.$$

$\phi_{\mathrm{dif},7}$ 和 $\phi_{\mathrm{score},1}$ 的结构评估为

$$L = \mathcal{O}(\log^4 N),\ \|W\|_\infty = \mathcal{O}(N\log^6 N),\ S = \mathcal{O}(N\log^8 N),\ \text{and}\ B = \exp\left(\log^4 N\right).$$

这里我们使用了 $|k_i| = \mathcal{O}(\log\ N)$ 和 $C = \mathcal{O}(1)$。

接下来，我们分析 $f_3(x,t)\ \sigma_t$ 与 $\nabla \log\ p_t(x)$ 之间的误差，以界定式 (46)。请注意，我们考虑满足 1 $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(1)$ 且 $p_t(x) \ge N^{-(2s+1)/d}$成立的 x。在这种情况下，我们有 $\|\nabla \log\ p_t(x)$ $\| \lesssim \log\ ^{\frac{2}{\sigma_t}} N$。首先，考虑 $x \in [-m_t, m_t]^d$的情形。根据引理 A.2， $p_t(x)$ 的下界为 C_{a}^{-1}；只要 $|f_1(x,t) - p_t(x)$ $| \le C_{\mathrm{a}}^{-1}/2$，就可以说近似误差有界于 $\lesssim |f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t\nabla p_t(x)\|$。另一方面，如果 $|f_1(x,t) - p_t$ $(x)| \ge C_{\mathrm{a}}^{-1}/2$，我们就不再有这样的界，但这时可以利用以下事实： $f_2(x,t)\ f_1(x,t)$，且 $\sigma_t\ ^{\sigma_t\nabla p_t}(x)\ p_t(x)$ 有界于 l og $^{\frac{1}{2}} N$。因此，当 $x \in [-m_t, m_t]^d$ 时，我们可以将近似误差界定为

$$\left\|f_3(x,t) - \sigma_t\frac{\nabla p_t(x)}{p_t(x)}\right\| \le \left\|\frac{f_2(x,t)}{f_1(x,t)} - \sigma_t\frac{\nabla p_t(x)}{p_t(x)}\right\| \lesssim \log^{\frac{1}{2}} N(|f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t\nabla p_t(x)\|).$$

接下来，我们考虑以下情形 当 $x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}]^d \setminus [-m_t, m_t]^d$ 时。于是，我们有 t

$$\left\|f_3(x,t) - \sigma_t\frac{\nabla p_t(x)}{p_t(x)}\right\| \le \left\|\frac{f_2(x,t)}{f_1(x,t)} - \sigma_t\frac{\nabla p_t(x)}{p_t(x)}\right\| \lesssim \frac{\|f_2(x,t) - \sigma_t\nabla p_t(x)\|}{f_1(x,t)} + \|\sigma_t\nabla p_t(x)\|\left|\frac{1}{f_1(x,t)} - \frac{1}{p_t(x)}\right|. \tag{50}$$

第一项由 $N^{(2s+1)/d}\|f_2(x,t)(x,t) - \sigma_t\nabla p_t(x)\|$ 控制，因为我们关注的是 $p_t(x) \ge N^{-(2s+1)/d}$ 的情形。对于第二项，由于 $\|\nabla$ log $p_t(x)\| = \left\|\sigma_t\ ^{\nabla p_t(x)\,(x)}_{p_t}\right\| \lesssim_{\log\ \frac{1}{2}}$，我们有 $\|\sigma_t\nabla p_t(x)\| \lesssim p_t(x)\log^{\frac{1}{2}} N$。利用这一点，我们可以将第二项控制为

$$\begin{aligned}\|\sigma_t\nabla p_t(x)\|\left|\frac{1}{f_1(x,t)} - \frac{1}{p_t(x)}\right| &\lesssim \log^{\frac{1}{2}} N p_t(x)\left|\frac{1}{f_1(x,t)} - \frac{1}{p_t(x)}\right| \\ &\lesssim \log^{\frac{1}{2}} N\frac{|p_t(x) - f_1(x,t)|}{f_1(x,t)} \\ &\lesssim N^{\frac{2s+1}{d}}\log^{\frac{1}{2}} N\,|p_t(x) - f_1(x,t)|,\end{aligned}$$

其中我们用到了 $f_1(x,t) \ge N^{-(2s+1)/d}$。因此，对于 $x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log}$ $N]^d \setminus [-m_t, m_t]^d$以及$p_t(x) \ge N^{-\frac{2s+1}{d}}$,(50) 的上界为

$$\left\|\phi_{\mathrm{dif},7}(x,t) - \frac{\sigma_t\nabla p_t(x)}{p_t(x)}\right\| \lesssim N^{\frac{2s+1}{d}}\log^{\frac{1}{2}} N(|\phi_{\mathrm{dif},5}(x,t) - p_t(x)| + \|\phi_{\mathrm{dif},6}(x,t) - \sigma_t\nabla p_t(x)\|).$$

因此，我们有

$$\begin{aligned}&\left\|\frac{f_2(x,t)}{\sigma_t f_1(x,t)} - \frac{\nabla p_t(x)}{p_t(x)}\right\| \\ &\lesssim \begin{cases}\log^{\frac{1}{2}} N(|f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t\nabla p_t(x)\|)/\sigma_t & (\|x\|_\infty \le m_t) \\ N^{\frac{2s+1}{d}}\log^{\frac{1}{2}} N(|f_1(x,t) - p_t(x)| + \|f_2(x,t) - \sigma_t\nabla p_t(x)\|)/\sigma_t \\ \qquad\qquad (x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}]^d \setminus [-m_t, m_t]^d).\end{cases}\end{aligned} \tag{51}$$

我们考虑 $L^2(p_t)$ 的损失（51）。首先，我们考虑 $\|x\|_\infty \le m_t$的情况。

$$
\begin{aligned}
&\int_{\|x\|_\infty \le m_t} p_t(x) \left\| \frac{f_2(x,t)}{\sigma_t f_1(x,t)} - \frac{\nabla p_t(x)}{p_t(x)} \right\|^2 \mathrm{d}x \\
&\lesssim \int_{\|x\|_\infty \le m_t} (|f_1(x,t) - p_t(x)|^2 + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|^2) \log N/\sigma_t^2 \mathrm{d}x \text{ (we used(51) and } p_t(x) = \mathcal{O}(1) \text{ by Lemma A.2.)} \\
&\lesssim \int_{\|x\|_\infty \le m_t} \left(\left| \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right|^2 \right. \\
&\quad \left. + \left\| \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right\|^2 \right) \log N/\sigma_t^2 \mathrm{d}x \\
&\lesssim \log N/\sigma_t^2 \cdot \int_{\|x\|_\infty \le m_t} \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x \\
&\quad + \log N/\sigma_t^2 \cdot \int_{\|x\|_\infty \le m_t} \int \frac{|x - m_t y|}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x \\
&= \log N/\sigma_t^2 \cdot \int \int_{\|x\|_\infty \le m_t} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}x \mathrm{d}y \\
&\quad + \log N/\sigma_t^2 \cdot \int \int_{\|x\|_\infty \le m_t} \frac{|x - m_t y|}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}x \mathrm{d}y \\
&\lesssim \log N/\sigma_t^2 \cdot \int |p_0(y) - f_N(y)|^2 \mathrm{d}y + \log N/\sigma_t^2 \cdot \int |p_0(y) - f_N(y)|^2 \mathrm{d}y \lesssim \log N/\sigma_t^2 \cdot N^{-2s/d}.
\end{aligned}
$$

对于第三个不等式，我们使用了詹森不等式。对于倒数第二个不等式，我们使用了 f_N 的构造以及引理 B.4。

接着，我们考虑 $x \in [-m_t - \mathcal{O}(1)\sigma_t\sqrt{\log N}, m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}]^d \setminus [-m_t, m_t]^d$ 的情形。大部分内容与之前相同。

$$
\begin{aligned}
&\int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x) \mathbb{I}[p_t(x) \ge N^{-\frac{2s+1}{d}}] \left\| \frac{f_2(x,t)}{\sigma_t f_1(x,t)} - \frac{\nabla p_t(x)}{p_t(x)} \right\|^2 \mathrm{d}x \\
&\lesssim \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} (|f_1(x,t) - p_t(x)|^2 + \|f_2(x,t) - \sigma_t \nabla p_t(x)\|^2) N^{\frac{4s+2}{d}} \log N/\sigma_t^2 \mathrm{d}x \\
&\lesssim \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \left(\left| \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right|^2 \right. \\
&\quad \left. + \left\| \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y - \int \frac{x - m_t y}{\sigma_t^{d+1} (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \right\|^2 \right) N^{\frac{4s+2}{d}} \log N/\sigma_t^2 \mathrm{d}x \\
&\lesssim N^{\frac{4s+2}{d}} \log N/\sigma_t^2 \cdot \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x \\
&\quad + N^{\frac{4s+2}{d}} \log N/\sigma_t^2 \cdot \int_{m_t \le \|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} \int \frac{|x - m_t y|^2}{\sigma_t^{d+2} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) |p_0(y) - f_N(y)|^2 \mathrm{d}y \mathrm{d}x
\end{aligned}
$$

$$\lesssim \left[\int_{m_t\le\|x\|_\infty\le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\left[\int_{\|\frac{x}{m_t}-y\|_\infty\le\mathcal{O}(1)\sigma_t\sqrt{\log N}}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}y+N^{-\frac{6s+2}{d}}\right]\mathrm{d}x\right.$$
$$\left.+\int_{m_t\le\|x\|_\infty\le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\left[\int_{\|\frac{x}{m_t}-y\|_\infty\le\mathcal{O}(1)\sigma_t\sqrt{\log N}}\frac{|x-m_ty|^2}{\sigma_t^{d+2}(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}y+N^{-\frac{6s+2}{d}}\right]\mathrm{d}x\right]$$
$$\cdot N^{\frac{4s+2}{d}}\log N/\sigma_t^2 \quad \text{(we used Lemma F.10.)}$$
$$\lesssim N^{\frac{4s+2}{d}}\log N/\sigma_t^2\cdot\left[\int_{m_t\le\|x\|_\infty\le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\left[\int_{\|\frac{x}{m_t}-y\|_\infty\le\mathcal{O}(1)\sigma_t\sqrt{\log N}}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}y\right]\mathrm{d}x\right.$$
$$\left.+\int_{m_t\le\|x\|_\infty\le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\left[\int_{\|\frac{x}{m_t}-y\|_\infty\le\mathcal{O}(1)\sigma_t\sqrt{\log N}}\frac{\log N}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}y\right]\mathrm{d}x+N^{-\frac{6s+2}{d}}\right]$$
$$\lesssim N^{\frac{4s+2}{d}}\log^2 N/\sigma_t^2\int_{m_t\le\|x\|_\infty\le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\int_{\|\frac{x}{m_t}-y\|_\infty\le\mathcal{O}(1)\sigma_t\sqrt{\log N}}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}x\mathrm{d}y$$
$$+N^{-\frac{2s}{d}}\log N/\sigma_t^2 \tag{52}$$

对于第三个不等式，我们使用了詹森不等式。这里，我们注意到，若 (x,y) 满足 $m_t\le\|x\|_\infty\le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}=\mathcal{O}(1)$ 和 $\|\frac{x}{m_t}-y\|_\infty\le\mathcal{O}(1)\sigma_t\sqrt{\log N}$，则有 $1-\mathcal{O}(1)\sigma_t\sqrt{\log N}\le\|y\|_\infty\le1+\mathcal{O}(1)\frac{\sigma_t}{m_t}\sqrt{\log N}$，并且 $1-\mathcal{O}(1)\sqrt{t}\le\|y\|_\infty\le1+\mathcal{O}(1)\sqrt{t}$。由于我们考虑时间 $t\le T_4=3N^{-\frac{2-\delta}{d}}$，对于充分大的 N，$\mathcal{O}(1)\sqrt{t}\lesssim N^{-\frac{1-\delta}{d}}$成立。因此，(52) 进一步被上界为

(52)
$$\lesssim N^{\frac{4s+2}{d}}\log^2 N/\sigma_t^2\int_x\int_{1-N^{-\frac{1-\delta}{d}}\le\|y\|_\infty\le1+N^{-\frac{1-\delta}{d}}}\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}x\mathrm{d}y$$
$$+N^{-\frac{2s}{d}}\log N/\sigma_t^2$$
$$=N^{\frac{4s+2}{d}}\log^2 N/\sigma_t^2\int_{1-N^{-\frac{1-\delta}{d}}\le\|y\|_\infty\le1+N^{-\frac{1-\delta}{d}}}\int_x\frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}y\mathrm{d}x$$
$$+N^{-\frac{2s}{d}}\log N/\sigma_t^2$$
$$\lesssim N^{\frac{4s+2}{d}}\log^2 N/\sigma_t^2\cdot N^{-\frac{6s+4}{d}}+N^{-\frac{2s}{d}}\log N/\sigma_t^2\lesssim N^{-\frac{2s}{d}}\log N/\sigma_t^2,$$

其中我们使用了 f_N 的构造以及引理B.4 来证明倒数第二个不等式。现在我们成功地对（46）进行了界定，结论由此得出。 □

B.4. 近似误差界：利用诱导光滑性

接下来，我们考虑对 $t\gtrsim T_2=N^{-(2-\delta)/d}$ 的近似。这可以通过考虑从 $t=t_*>0$ 开始的扩散过程来证明。我们先给出以下引理。

引理 B.6（{v1} 的基 于 $t=t_*$ 的分解）。若 N、$N'\gg1$ 和 $N'\ge t_*^{-\frac{d}{2}}N^{\frac{\delta}{2}}$，则存在 $f_{N'}$ 使得

$$\|p_{t_*}-f_{N'}\|_{L^2(\mathbb{R}^d)}\lesssim N^{-(3s+5)/d},$$

$f_{N'}(x)=0$ 针对 x，其 $\|x\|_\infty\gtrsim\mathcal{O}(\sqrt{\log N})$，形式如下：

$$f_N(x)=\sum_{i=1}^{N'}\mathbb{I}[\|x\|_\infty\lesssim\mathcal{O}(\sqrt{\log N})]M_{k_i,j_i}^d(x),$$

其中 $-\sqrt{\log N}2^{(k_i)_m}-l\lesssim(j_i)_l\lesssim\sqrt{\log N}2^{(k_i)_l}(i=1,2,\cdots,N,\ m=1,2,\cdots,d)$，$\|k_i\|_\infty\le K=\mathcal{O}(d^{-1}\log N)$，且 $|\alpha_i|\lesssim N^{\frac{(3s+6)(2-\delta)}{\delta}}$。

证明。设 $\alpha=\frac{2}{(3s+6)\delta}+1$。根据引理 A.3，对任意 x，有

$$\|\partial_{x_{i_1}}\partial_{x_{i_2}}\cdots\partial_{x_{i_k}}p_{T_2}(x)\|\le\frac{C_{\mathrm{a}}}{\sigma_{t_*}^k}.$$

所有直至 α 阶的导数均由 $\sigma_{t_*}^{-\alpha} \lesssim t_*^{-\frac{\alpha}{2}} \vee 1$有界， $p_{t_*}(x)\overline{t_*^{-\frac{\alpha}{2}} \vee a}$属于 W_∞^α，且其在 W_∞^α 中的范数由依赖于 α的某个常数控制，因此也属于 $B_{\infty,\ \infty}^\alpha$。因此，根据引理 F.13，存在一个 B 样条基阶数为 $l = \alpha + 2$的基分解：

$$f_{N'}(x) = (t_*^{-\frac{\alpha}{2}} \vee 1) \sum_{i=1}^{N} \alpha_i M_{k_i,j_i}^d(x).$$

使得

$$\begin{aligned}
&\|p_{t_*} - f_{N'}\|_{L^2([-\mathcal{O}(\sqrt{\log N}),\mathcal{O}(\sqrt{\log N})]^d)} \lesssim (\sqrt{\log N})^\alpha N'^{-\alpha/d} t_*^{-\frac{\alpha}{2}} \\
&= (\sqrt{\log N})^\alpha N^{\alpha\delta/2d} = (\sqrt{\log N})^\alpha N^{-(3s+6)/d} \lesssim N^{-(3s+5)/d},
\end{aligned}$$

其中 $-\sqrt{\log N}2^{(k_i)_m} - l \lesssim (j_i)_l \lesssim \sqrt{\log N}2^{(k_i)_l}(i = 1, 2, \cdots, N,\ m = 1, 2, \cdots, d)$, $\|k_i\|_\infty \le K = \mathcal{O}(d^{-1}\log N)$, 且 $|\alpha_i| \lesssim 1$。此外，若 $\varepsilon = N^{-\frac{6s+10}{d}}$and $m_{t_*} + \mathcal{O}(1)\sigma_{t_*}\sqrt{\log N} \lesssim \sqrt{\log N}$，引理 A.4 保证 $\|p_{T_2} - f_N\|_{L^2(\mathbb{R}^d \subseteq [-\mathcal{O}(\sqrt{\log N}) \mathcal{O}(\sqrt{\log N})]^d)} \lesssim N^{-(3s+5)/d}$。因此，通过重新设定 $\alpha_i \leftarrow (t_*^{-\frac{\alpha}{2}} \vee 1)\alpha_i$，该断言成立。（$\alpha_iis_\alpha$随后有 T_2 $^{\frac{\alpha}{2}}$的界。） □

引理 B.6 为 $T_3 \le t \le T_5$ 给出了神经网络的具体构造。

引理 B.7（$T_3 \le t \le T_5$ 的得分函数近似；引理 3.6）。设 $N \gg 1$和 $N' \ge t_*^{-d/2} N^{\delta/2}$。假设$t_* \ge N^{-(2-\delta)/d}$。则存在一个神经网络 $\phi_{\text{score},2} \in \Phi(L, W, S, B)$，满足

$$\int_x p_t(x)\|\phi_{\text{score},2}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^2}$$

对于 $t \in [2t_*, \overline{T}]$。具体而言， $L = \mathcal{O}(\log^4(N))$ $\|W\|_\infty = \mathcal{O}(N)$ $S = \mathcal{O}(N')$，且 $B = \exp(\mathcal{O}(\log^4 N))$。此外，我们可以取 $\phi_{\text{score},2}$，满足$\|\phi_{\text{score},2}\|_\infty = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} N)$。

证明。证明与引理 B.5 的证明基本相同。这里的细微差别在于：(i) p_t、ϕ_{dif}、8 和 f_1 的下界为 $N^{-(2s+3)/d}$，而不是 $N^{-(2s+1)/d}$；(ii) $L^2(p_t)$ 的误差应由 $\frac{N^{-\frac{2(s+1)}{d}}}{\sigma_t^2}$界定，而不是由 $\frac{N^{-\frac{2s}{d}}}{\sigma_t^2}$界定；(iii) p_{t_*}的支撑集为 $\mathbb{R}^d$，而不是 $[-1, 1]^d$。界定 Observe 之间的差异。注意到 $t_* \ge T_1 = N^{-\frac{2-\delta}{d}}$成立，这是应用引理 B.5 中论证所必需的。

在以下证明中，我们重新设定时间 $t \leftarrow t - t_*$ 并考虑从 p_0 （在新的定义下）开始的扩散过程，以简化记号。我们有 $t \ge t_* \gtrsim \mathrm{poly}(N^{-1})$ 在新的定义下。根据引理 A.4，可得

$$\int_{\|x\|_\infty \ge m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x)\|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{t_*}{N^{(2s+2)/d}} \left(1 + \|s(\cdot,t)\|_\infty^2\right), \tag{53}$$

其中 $\mathcal{O}(1)$ 中的隐藏常数足够大。我们将 x 的定义域限制为 $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(\sqrt{\log N})$。在此区域内，引理 A.3 给出 $\|\nabla \log p_t(x)\| \lesssim \sqrt{\log \frac{N}{\sigma_t}}$，因此我们可以取 s，使得 $\|s(\cdot,t)\|_\infty \le \sqrt{\log \frac{N}{\sigma_t}} \lesssim \sqrt{\log \frac{N}{\sqrt{t_* \wedge 1}}}$ 成立。于是，(53) 被 $N^{-2(s+1)/d}$ 所界定。此外，

$$\begin{aligned}
&\int_{\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}} p_t(x) \mathbb{I}[p_t(x) \le N^{-(2s+3)/d}]\|s(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{\varepsilon}{\sigma_t^2} \log^{\frac{d+2}{2}}(N) + \varepsilon\|s(x,t)\| \\
&\lesssim \left(\frac{N^{-(2s+3)/d}}{\sigma_t^2} \log^{\frac{d+2}{2}}(N) + \frac{N^{-(2s+3)/d}}{\sigma_t^2} \log N\right) \log^{\frac{d}{2}} N \lesssim N^{-2(s+1)/d}.
\end{aligned}$$

这意味着我们只需考虑满足 $p_t(x) \ge N^{-(2s+3)/d}$ 的 x。

利用前一个引理中的基分解，我们令

$$\begin{aligned} p_t(x) &= \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \coloneqq \int \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\ &= \sum_{i=1}^{N'} \alpha_i E^{(1)}_{k_i,j_i}(x,t) =: \tilde{f}_1(x,t), \\ f_1(x,t) &:= \tilde{f}_1(x,t) \vee N^{-(2s+3)/d}, \end{aligned}$$

and

$$\begin{aligned} \sigma_t \nabla p_t(x) &= \int \frac{x-m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \coloneqq \int \frac{x-m_t y}{\sigma_t^{d+1}(2\pi)^{\frac{d}{2}}} f_N(y) \exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y \\ &= \sum_{i=1}^{N'} \alpha_i E^{(2)}_{k_i,j_i}(x,t) =: f_2(x,t), \\ f_3(x,t) &:= \frac{f_2(x,t)}{f_1(x,t)} \mathbb{I}\left[\left\|\frac{f_2(x,t)}{f_1(x,t)}\right\| \lesssim \frac{\log^{\frac{1}{2}} N}{\sigma_t}\right] \end{aligned}$$

（定义与引理 B.5 中的完全相同，除了 $f_1(x,t) := \tilde{f}_1(x,t) \vee N^{-(2s+3)/d}$ 。接着，我们使用引理 B.3 分别近似每个 $\alpha_i E^{(1)}_{k_i,j_i}(x,t)$ 和 $\alpha_i E^{(2)}_{k_i,j_i}(x,t)$，其中 $\varepsilon \lesssim N'^{-1} \cdot N^{-(3s+6)(2-\delta)} \cdot N^{-\frac{9s+10}{d}}$ 和 $C = m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N} = \mathcal{O}(\sqrt{\log N})$，再根据引理 F.3 将它们聚合，得到 $\phi_{\mathrm{dif},8}(x,t)$ 和 $\phi_{\mathrm{dif},9}(x,t)$，它们分别近似 f_1 和 f_2，并满足

$$|f_1(x,t) - \phi_{\mathrm{dif},8}(x,t)| \lesssim N^{-\frac{9s+3}{d}}, \quad \|f_2(x,t) - \phi_{\mathrm{dif},9}(x,t)\| \lesssim N^{-\frac{9s+10}{d}}.$$

对于所有满足 $\|x\|_\infty = \mathcal{O}(\sqrt{\log N})$ 的 x。现在，我们将 $\phi_{\mathrm{dif},7}$ 定义为

$$[\phi_{\mathrm{dif},10}(x,t)]_i := \phi_{\mathrm{clip}}(\phi_{\mathrm{mult}}(\phi_{\mathrm{rec}}(\phi_{\mathrm{clip}}(\phi_{\mathrm{dif},8}(x,t); N^{-(2s+3)/d}, \mathcal{O}(1))), [\phi_{\mathrm{dif},9}(x,t)]_i); -\mathcal{O}(\log^{\frac{1}{2}} N), \mathcal{O}(\log^{\frac{1}{2}} N)),$$

其中我们令 $\varepsilon = N^{-(3s+4)/d}$ 在引理 F.7 中取 ϕ_{rec}，并令 $\varepsilon = N^{-(s+1)/d}$ 以及 $C = N^{(2s+3)/d}$ 在引理 F.6 中取。最后，我们令

$$\phi_{\mathrm{score},2}(x,t) := \phi_{\mathrm{mult}}(\phi_{\mathrm{dif},10}(x,t), \phi_\sigma(t)).$$

其中 $\varepsilon = N^{-(s+1)/d}$ 以及 $C \simeq$ 最大值{对数 $\frac{1}{2} N$， σ_T} $\lesssim$ 多项式（N） 在引理 F.6 中取 ϕ_{mult}，且 $\varepsilon = N^{-(s+1)/d}$/多项式（$N$） 在引理 B.1 中取 ϕ_σ。综上，我们可以验证

$$\left\|\phi_{\mathrm{score},2}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\| \lesssim N^{-(s+1)/d}$$

对所有满足 x 且 $\|x\|_\infty \lesssim \sqrt{\text{对数 } N}$ 的情形均成立，因此

$$\int_{\|x\|_\infty \lesssim \sqrt{\log N}} p_t(x) \left\|\phi_{\mathrm{score},2}(x,t) - \frac{f_3(x,t)}{\sigma_t}\right\|^2 \lesssim N^{-(s+1)/d}. \tag{54}$$

此外，$\phi_{\mathrm{score},\,2}$ 的大小有界于

$$L = \mathcal{O}(\log^4 N),\ \|W\|_\infty = \mathcal{O}(N' \log^6 N) \lesssim \mathcal{O}(N),\ S = \mathcal{O}(N' \log^8 N),\ \text{and}\ B = \exp\left(\log^4 N\right). \tag{55}$$

现在，我们考察 $f_3(x,t)/\sigma_t$ 与 $\nabla \log p_t(x)$ 之间的差异。其在 $\|x\|_\infty \le m_t + \mathcal{O}(1)\sigma_t\sqrt{\log N}$ 中的 L^2 误差为

其界与之前相同，最终得到

$$
\begin{aligned}
&\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}} \mathbb{I}[p_t(x)\ge N^{-\frac{2s+3}{d}}]p_t(x)\left\|\frac{f_3(x,t)}{\sigma_t}-\frac{\nabla p_t(x)}{p_t(x)}\right\|^2 \mathrm{d}x\\
&\lesssim N^{\frac{4s+6}{d}}\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}(|f_1(x,t)-p_t(x)|^2+\|f_2(x,t)-\sigma_t\nabla p_t(x)\|^2)\log N/\sigma_t^2\mathrm{d}x\\
&\lesssim N^{\frac{4s+6}{d}}\log N/\sigma_t^2\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\left|\int_y \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)(p_0(y)-f_N(y))\mathrm{d}y\right|^2\mathrm{d}x\\
&\quad+N^{\frac{4s+6}{d}}\log N/\sigma_t^2\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\left|\int_y \frac{x-m_t y}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)(p_0(y)-f_N(y))\mathrm{d}y\right|^2\mathrm{d}x\\
&\lesssim N^{\frac{4s+6}{d}}\log N/\sigma_t^2\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\int_y \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}y\mathrm{d}x\\
&\quad+N^{\frac{4s+6}{d}}\log N/\sigma_t^2\int_{\|x\|_\infty \le m_t+\mathcal{O}(1)\sigma_t\sqrt{\log N}}\int_y \frac{|x-m_t y|}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}y\mathrm{d}x\\
&\lesssim N^{\frac{4s+6}{d}}\log N/\sigma_t^2\int_y\int_x \frac{1}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}x\mathrm{d}y\\
&\quad+N^{\frac{4s+6}{d}}\log N/\sigma_t^2\int_y\int_x \frac{|x-m_t y|}{\sigma_t^d(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)|p_0(y)-f_N(y)|^2\mathrm{d}x\mathrm{d}y\\
&\lesssim N^{\frac{4s+6}{d}}\log N/\sigma_t^2\int_y|p_0(y)-f_N(y)|^2\mathrm{d}y\lesssim N^{\frac{4s+6}{d}}\log N/\sigma_t^2\cdot N^{-\frac{6s+10}{d}}\lesssim N^{-\frac{2(s+1)}{d}}/\sigma_t^2. \qquad (56)
\end{aligned}
$$

此处最后一个不等式使用了前一个引理的结果。由等式（54）和（55）、（56）可得结论。

□

结合引理B.5 和B.7，其中对B.5 使用引理 $T_1 \le t \le T_4$，对B.7 使用引理 $T_3 \le t \le T_5$，即可立即得到定理3.1。

定理 3.1 的证明。注意，在引理 B.7 中可以设定 $N' = N$和 $t_* = N^{-(2-\delta)/d}$。根据引理 B.5 和 B.7，我们有两个神经网络 $\phi_{\text{score},1}(x,t)$ 和 $\phi_{\text{score},2}(x,t)$，它们分别在 $[T_1,T_4]$和 $[T_3,T_5]$中近似得分函数。因此，在引理 F.5 中令 $\overline{t}_1 = T_4$和 $\underline{t}_2 = T_3$， $\phi_{\text{score}}(x,t) = \phi^1_{\text{swit}}(t;\underline{t}_2,\overline{t}_1)\phi_{\text{score},1}(x,t) + \phi^2_{\text{swit}}(t;\underline{t}_2,\overline{t}_1)\phi_{\text{score},2}(x,t)$ 可在 $L^2(p_t)$ 中近似近似误差，且加性误差为 $N^{-2s/d}\log_{\sigma_t^2}{}^{N}$。乘法（$\phi^1_{\text{swit}}\phi_{\text{score},1}$ 和 $\phi^2_{\text{swit}}\phi_{\text{score},2}$）以及聚合$\overline{\phi^1_{\text{swit}}\phi_{\text{score},1} + \phi^2_{\text{swit}}\phi_{\text{score},2}}$ 的实现是显而易见的。最后，根据引理 B.5 和 B.7，网络的大小有如下界：

$$
L=\mathcal{O}(\log^4(N)),\|W\|_\infty=\mathcal{O}(N\log^6 N),S=\mathcal{O}(N\log^8 N),\quad \text{and } B=\exp(\mathcal{O}(\log^4 N)),
$$

至此，证明完成。 □

我们还准备了近似定理的一种积分形式。

定理 B.8 (近似定理)。 假设 *2.4*、 *2.5*、 *2.6* 满足 $a_0 = N^{-(1-\delta)/d}$， $N \gg 1$， $\underline{T}$ =poly(N^{-1})，且 $\overline{T} \simeq \log N$。则存在一个神经网络 $\phi_{\text{score}} \in \Phi(L,\ W,\ S,\ B)$满足

$$
\int_{t=\underline{T}}^{\overline{T}}\int_x p_t(x)\|\phi_{\text{score}}(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}x\mathrm{d}t \lesssim N^{-2s/d}\log N(\log(\overline{T}/\underline{T})+(\overline{T}-\underline{T})).
$$

此处， $L, \|W\|_\infty, S, B$定义为

$$
L=\mathcal{O}(\log^4 N),\quad \|W\|_\infty=\mathcal{O}(N),\quad S=\mathcal{O}(N),\quad \textit{and } B=\exp(\mathcal{O}(\log^4 N)).
$$

此外，假设 $N' \geq t_*^{-d/2} N^{\delta/2}$, $t_* \geq N^{-(2-\delta)/d}$，且 $\underline{T} \geq 2t_*$，则存在一个神经网络 $\phi_{\text{score}} \in \Phi(L, W, S, B)$，满足

$$\int_{t=\underline{T}}^{\overline{T}} \int_x p_t(x) \|\phi_{\text{score}}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t \lesssim N^{-\frac{2(s+1)}{d}} (\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T})).$$

具体而言，$L = \mathcal{O}(\log^4(N))$ $\|W\|_\infty = \mathcal{O}(N)$ $S = \mathcal{O}(N')$，且 $B = \exp(\mathcal{O}(\log^4 N))$ 。

证明。我们只证明第一部分；第二部分可用相同的方法由引理 B.7 得出。根据定理 3.1，存在一个网络 ϕ_{score}，其规模满足所需条件，且满足

$$\int_x p_t(x) \|\phi_{\text{score}}(x,t) - s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d}} \log(N)}{\sigma_t^2}.$$

注意到 $\sigma_t \gtrsim t \wedge 1$。因此，

$$\int_{t=\underline{T}}^{\overline{T}} \frac{N^{-\frac{2s}{d}} \log(N)}{\sigma_t^2} \mathrm{d}t \lesssim \int_{t=\underline{T}}^{\overline{T}} N^{-\frac{2s}{d}} \log(N)(1 \vee 1/t) \mathrm{d}t \leq N^{-\frac{2s}{d}} \log(N)(\log(\overline{T}/\underline{T}) + (\overline{T} - \underline{T})),$$

由此可得定理的第一部分。 □

C. 得分网络的泛化

现在我们考虑泛化误差。与第 4 节一样，我们首先考虑 ℓ 的上确界范数，并计算覆盖数。

C.1. 上确界范数的界

引理 C.1. 假设 $\|s(\cdot,t)\|_\infty = \mathcal{O}(\sigma_t^{-1} \log^{\frac{1}{2}} n)$、$\underline{T} = \text{poly}(n^{-1})$ 且 $\overline{T} \simeq \log n$。那么，我们有 tha t

$$\int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t \mathrm{d}t \lesssim \log^2 n.$$

证明。计算过程大多很直接。

$$\begin{aligned}
&\int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t \mathrm{d}t \\
&\leq 2 \int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t|x_0) \mathrm{d}x\mathrm{d}t + 2 \int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|\log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t \mathrm{d}t \\
&\lesssim \int_{t=\underline{T}}^{\overline{T}} \frac{\log n}{\sigma_t^2} \mathrm{d}t + \int_{t=\underline{T}}^{\overline{T}} \frac{1}{\sigma_t^2} \mathrm{d}t \\
&\lesssim \int_{t=\underline{T}}^{\overline{T}} \frac{\log n}{t \wedge 1} \mathrm{d}t \leq (\log n) \cdot (\log \underline{T}^{-1} + \overline{T}) \lesssim \log^2 n
\end{aligned}$$

在计算 $\int_{x_t} \|\log p_t(x_t|x_0)\|^2 p_t(x_t|x_0) \mathrm{d}x_t$ 时，我们用到了 $p_t(x_t|x_0)$ 是 $\mathcal{N}(m_t x_0, \sigma_t^2)$ 的密度函数这一事实。此外，在最后一个不等式中，我们用到了 $\underline{T} = \text{poly}(n^{-1})$ 和 $\overline{T} \simeq \log n$。 □

C.2. 覆盖数计算

引理 **C.2**（$\mathcal{L}$的覆盖数）。对于神经网络 $s \cdot \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$，我们将 $\ell \cdot \mathbb{R}^d \to \mathbb{R}$ 定义为

$$\ell_s(x) = \int_{t=\underline{T}}^{\overline{T}} \int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x) \mathrm{d}x\mathrm{d}t.$$

对于假设网络类 $\mathcal{S}\in\Phi(L,W,S,B)$，我们定义一个函数类 $\mathcal{L}=\{\ell_s|\ s\in\mathcal{S}\}$。如果对于某个 ℓ_s，相应的 s显而易见，我们有时将 ℓ_s 简称为 ℓ。

假设 $s(x,t)$ 被 $\|\|s(\cdot,t)\|_2\|_{L^\infty}=\mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}}n)$ 在所有 $s\in\mathcal{S}$ 和 $C\geq 1$ 上一致地有界。那么，$\mathcal{S}$ 的覆盖数可表示为

$$\log\mathcal{N}(\mathcal{S},\|\|\cdot\|_2\|_{L^\infty([-C,C]^{d+1})},\delta)\lesssim 2SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)C),\tag{57}$$

基于此，$\mathcal{L}$ 的覆盖数可表示为

$$\log\mathcal{N}(\mathcal{L},\|\cdot\|_{L^\infty([-1,1]^d)},\delta)\lesssim SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)n)\tag{58}$$

当 δ^{-1}、$\underline{T}^{-1}$、$\overline{T}$、$N=$ 为 poly(n) 时。

证明。第一个界（57）可直接由Suzuki （2018）得到，只需对输入区域稍作修改。按照他们的证明，我们可以看到，他们的 δ-网对于 $L^\infty([0,1]^d)$-范数而言，是 $C\delta$-网对于 $L^\infty([-C,C]^d)$-范数而言的。因此，我们只需将他们界中的 $\delta\leftarrow C^{-1}\delta$ 设为该值，即可得到（57）。

接下来考虑（58）。首先，我们截断 ℓ 定义中的积分区间。

$$\begin{aligned}
&\left|\ell_s(x)-\int_{t=\underline{T}}^{\overline{T}}\int_{\|x_t\|_\infty\leq\mathcal{O}(\sqrt{\log n})}\|s(x_t,t)-\nabla\log p_t(x_t|x)\|^2p_t(x_t|x)\mathrm{d}x_t\mathrm{d}t\right|\\
&\leq\int_{t=\underline{T}}^{\overline{T}}\int_{\|x_t\|_\infty\geq\mathcal{O}(\sqrt{\log n})}\|s(x_t,t)-\nabla\log p_t(x_t|x)\|^2p_t(x_t|x)\mathrm{d}x_t\mathrm{d}t\mathrm{d}t\\
&\leq\|\|s(\cdot,\cdot)\|_2\|_{L^\infty}^2\int_{t=\underline{T}}^{\overline{T}}\int_{\|x_t\|_\infty\geq\mathcal{O}(\sqrt{\log n})}p_t(x_t|x)\mathrm{d}x_t\mathrm{d}t+\int_{t=\underline{T}}^{\overline{T}}\int_{\|x_t\|_\infty\geq\mathcal{O}(\sqrt{\log n})}\|\nabla\log p_t(x_t|x)\|^2p_t(x_t|x)\mathrm{d}x_t\mathrm{d}t.
\end{aligned}\tag{59}$$

由于 $p_t(x_t|x)$ 是 $\mathcal{N}(m_tx|\sigma_t^2)$ 的密度函数，我们可以证明，当 $\delta^{-1},\underline{T}^{-1},\overline{T},N=$poly$(n)$，且 $\mathcal{O}(\sqrt{\log n})$ 中的隐藏常数足够大时（见引理 F.14），$\int_{\|x_t\|_\infty\geq\mathcal{O}(\sqrt{\log\ n})}p_t(x_t|x)\mathrm{d}x_t$ 和$\int_{\|x_t\|_\infty\geq\mathcal{O}(\sqrt{\log\ n})}\|\nabla\log p_t(x_t|x)\|^2p_t(x_t|x)\mathrm{d}x_t$ 均由 $\frac{\delta}{3\overline{T}\cdot\cdot(\|\|s(\,,\,)\|_2\|_{L^\infty}^2\vee 1)}$ 界定。因此，(59) 由以下式子界定

$$\|\|s(\cdot,\cdot)\|_2\|_{L^\infty}(\overline{T}-\underline{T})\cdot\frac{\delta}{3\overline{T}\|\|s(\cdot,\cdot)\|_2\|_{L^\infty}}+(\overline{T}-\underline{T})\cdot\frac{\delta}{3\overline{T}}\leq\frac{2}{3}\delta.\tag{60}$$

接下来，我们取 $C=$poly$(n)\gtrsim\sqrt{\log\ n}$，并为以下集合构造一个 $\frac{\delta}{3}$-网：

$$\ell'(x):=\int_{t=\underline{T}}^{\overline{T}}\int_{\|x_t\|_\infty\leq C}\|s(x_t,t)-\nabla\log p_t(x_t|x)\|^2p_t(x_t|x)\mathrm{d}x_t\mathrm{d}t\tag{61}$$

在所有 $s\in\mathcal{S}$上。为此，我们取 $\frac{\delta}{n^{\mathcal{O}(1)}}$-网，其覆盖 $\mathcal{S}$，并采用 $L^\infty([-C,C]^{d+1})$-范数。根据 (57)，覆盖数可计算为

$$\log\mathcal{N}\left(\mathcal{S},\|\|\cdot\|_2\|_{L^\infty([-C,C]^{d+1})},\frac{\delta}{n^{\mathcal{O}(1)}}\right)\lesssim 2SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)n).$$

对于不同的 s 和 s'，由于 $\|\nabla\log\ p_t(x_t|x)\|\lesssim\frac{C}{\sigma_t^2}$对于 $\|x_t\|_\infty\leq C$，我们有

$$\begin{aligned}
&|\|s(x_t,t)-\nabla\text{对数}\ p_t(x_t|x)\|^2-\|s'(x_t,t)-\nabla\text{对数}\ p_t(x_t|x)\|^2| &(62)\\
&\leq(\|s(x_t,t)-\nabla\log\ p_t(x_t|x)\|+\|s'(x_t,t)-\nabla\log\ p_t(x_t|x)\|^2)|\|s(x_t,t)-\nabla\log\ p_t(x_t|x)\|-\|s'(x_t,t)-\nabla\log\ p_t(x_t|x)\|| &(63)\\
&\leq(\|\|s(\cdot,\cdot)\|_2\|_{L^\infty}+\|\|s'(\cdot,\cdot)\|_2\|_{L^\infty}+2C/\sigma_t^2)\cdot\frac{\delta}{n^{\mathcal{O}(1)}}.
\end{aligned}$$

通过将 $\frac{\delta}{n^{\mathcal{O}(1)}}$ 中的隐藏常数取足够大，当 C、$\underline{T}^{-1}$、$\overline{T}=$poly(n) 时，该式可进一步由 $\frac{\delta}{3\overline{T}(2C)^d}$ 上界。对 $\int_{t=\underline{T}}^{\overline{T}}\int_{\|x_t\|_\infty\leq C}\mathrm{d}x_t\mathrm{d}t$ 积分（62）和（63）可得，这个 $\frac{\delta}{n^{\mathcal{O}(1)}}$-net of $\mathcal{S}$ 实际上给出了（61）中集合的 $\frac{\delta}{3}$-net；最后，结合（60），我们得到了 δ-net for $\mathcal{L}$。 □

C.3. 得分匹配损失的泛化误差界

本小节给出定理 4.3 的完整证明。首先，下面的关系十分有用。它展示了显式分数匹配与去噪分数匹配之间的等价性，并可用于证明经验去噪分数匹配的最小化器也近似最小化显式分数匹配损失。

引理 C.3 （显式分数匹配与去噪分数匹配的等价性（Vincent （2011年））） 。以下等式对所有 s（x_t， t）和 $t > 0$均成立：

$$\int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t = \int_{x_0}\int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0)p_0(x_0)\mathrm{d}x_0\mathrm{d}x_0 + C,$$

其中 $C = \int_{x_t} \|\nabla \log\ p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t - \int_{x_0}\int_{x_t} \|\nabla \log p_t(x_t|x_0)\|^2 p_t(x_t|x_0)p_0(x_0)\mathrm{d}x_t\mathrm{d}x_0$。

证明。证明沿用Vincent （2011年）。

$$\begin{aligned}
&\int_{x_t} \|s(x_t,t) - \nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_t} p_t(x_t)s(x_t,t)^\top \nabla \log p_t(x_t)\mathrm{d}x + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x \\
&= -2\int_{x_t} s(x_t,t)^\top \nabla p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x \\
&= -2\int_{x_t} s(x_t,t)^\top \nabla \left(\int_{x_0} p_t(x_t|x_0)p_0(x_0)\mathrm{d}x_0\right)\mathrm{d}x_t + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_t} s(x_t,t)^\top \left(\int_{x_0} p_0(x_0)\nabla p_t(x_t|x_0)\mathrm{d}x_0\right)\mathrm{d}x_t + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_t} p_t(x_t|y)p_0(x_0)s(x_t,t)^\top \left(\int_{x_0} \nabla \log p_t(x_t|x_0)\mathrm{d}x_0\right)\mathrm{d}x_t + \int_{x_t} \|s(x_t,t)\|^2 p_t(x_t)\mathrm{d}x_t + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= -2\int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)s(x_t,t)^\top \nabla \log p_t(x_t|x_0)\mathrm{d}x_t\mathrm{d}x_0 + \int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)\|s(x_t,t)\|^2\mathrm{d}x_t\mathrm{d}x_0 \\
&\qquad + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&= \int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)\|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2\mathrm{d}x_t\mathrm{d}x_0 + \int_{x_t} \|\nabla \log p_t(x_t)\|^2 p_t(x_t)\mathrm{d}x_t \\
&\qquad - \int_{x_0}\int_{x_t} p_t(x_t|x_0)p_0(x_0)\|\nabla \log p_t(x_t|x_0)\|^2\mathrm{d}x_t\mathrm{d}x_0,
\end{aligned}$$

其中第二个等式使用了 $\nabla\log\ p_t(x_t)\ = (\nabla p_t(x_t))/p_t(x_t)$，第三个等式使用了 $p_t(x_t)\ =\ \int_{x_0} p_t(x_t|x_0)p_0(x_0)\mathrm{d}x_0$，第五个等式使用了$\nabla\log\ p_t(x_t|x_0)\ = (\nabla p_t(x_t|x_0))/p_t(x_t|x_0)$。

□

现在，我们评估泛化误差，以下定理是定理 4.3 的正式表述。

定理 C.4（基于覆盖数的泛化误差界）。设 $\hat{s}$ 为以下目标的最小化器

$$\frac{1}{n}\sum_{i=1}^{n}\int_{t=\underline{T}}^{\overline{T}}\int_x \|s(x,t) - \nabla \log p_t(x|x_i)\|_2^2 p_t(x|x_{0,i})\mathrm{d}x\mathrm{d}t, \tag{64}$$

取值于 $\mathcal{S} \subset L^2(\mathbb{R}^d \times [\underline{T},\overline{T}])$。对于每个 $s \in \mathcal{S}$，令 $\ell(x)\ =\ \int_{t=\underline{T}}^{\overline{T}}\int_x \|s(x,t)\ - \nabla\log\ p_t(y|x)\|_2^2 p_t(y|x)\mathrm{d}y\mathrm{d}t$ and$\mathcal{L}$ 为与每个 $s \in \mathcal{S}$对应的一组 ℓ。假设每个元素 $\ell \in \mathcal{L}$ 都满足 $\|\ell\|_{L^\infty([-1,1]^d)} \le C_\ell$，其中某个固定

$0 < C_{\ell\circ}$ 对于任意的 $\delta > 0$，如果 $N := N(\mathcal{L}, \|\cdot\|_{L^\infty([-1,1]^d)}\, \delta) \geq 3$，那么我们有

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x\right]$$
$$\leq 2 \inf_{s\in\mathcal{S}} \int_x \int_{\underline{T}}^{\overline{T}} \|s(x,t) - \nabla \log p_t(x)\|_2^2 p_t(x) \mathrm{d}x\mathrm{d}t + \frac{2C_\ell}{n}\left(\frac{37}{9}\log N + 32\right) + 3\delta.$$

证明。 在以下证明中，$x_{0,\ i}$ 记作 x_i 以简化记号。（64）写作$\frac{1}{n}\sum_{i=1}^n \ell\ (x_i)$ 。此外，对于 $s^\circ\ (x,\ t)$ $=\nabla\log\ p_t$ (x)，我们记为

$$\begin{aligned}
R(\hat{\ell}, \ell^\circ) &:= \int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x \\
&= \int_x \int_{t=\underline{T}}^{\overline{T}} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x - \underbrace{\int_x \int_{t=\underline{T}}^{\overline{T}} \|s^\circ(x,t) - \nabla \log p_t(x)\|^2 p_t(x) \mathrm{d}t\mathrm{d}x}_{=0} \\
&= \int_y \int_{t=\underline{T}}^{\overline{T}} \int_x \|s(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y) p_0(x) \mathrm{d}y\mathrm{d}t\mathrm{d}x + C(\overline{T} - \underline{T}) \\
&\qquad - \int_y \int_{t=\underline{T}}^{\overline{T}} \int_x \|s^\circ(x,t) - \nabla \log p_t(x|y)\|^2 p_t(x|y) p_0(x) \mathrm{d}y\mathrm{d}t\mathrm{d}x - C(\overline{T} - \underline{T}) \\
&= \mathbb{E}_{\{x_i'\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\ell}(x_i') - \ell^\circ(x_i'))\right] \qquad (65)
\end{aligned}$$

其中 $\{x_i'\}_{i=1}^n$，它是来自 p_0 且与 $\{x_i\}_{i=1}^n$独立的独立同分布样本。对于第二个等式，我们使用了引理C.3。首先，我们计算

$$D := \left|\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\ell}(x_i) - \ell^\circ(x_i))\right] - R(\hat{\ell}, \ell^\circ)\right|.$$

利用（65），我们得到

$$D = \left|\mathbb{E}_{x_i, x_i'}\left[\frac{1}{n}\sum_{i=1}^n ((\hat{\ell}(x_i) - \ell^\circ(x_i)) - (\hat{\ell}(x_i') - \ell^\circ(x_i')))\right]\right| \leq \frac{1}{n}\mathbb{E}_{x_i, x_i'}\left[\left|\sum_{i=1}^n ((\hat{\ell}(x_i) - \ell^\circ(x_i)) - (\hat{\ell}(x_i') - \ell^\circ(x_i')))\right|\right].$$

令 $\mathcal{L}_d = \{\ell_1, \ell_2, \cdots, \ell_N\}$ 为 $\mathcal{L}$ 在 L^∞（$[-1,1]^d$）度量下基数最小的 δ-覆盖。根据 N（$\mathcal{L}, \|\cdot\|_\infty, \delta$）$\geq 3$ 的假设，我们有 $\log\ N \geq 1$。我们定义 $g_j\ (x, x')\ =\ (\ell_j\ (x)\ -\ell^\circ\ (x)\ )\ -\ (\ell_j\ (x')\ -\ell^\circ\ (x')\ )$，并定义一个取值于 $\{1, 2, \cdots, N\}$ 的随机变量 J，使得 $\|\ell - f_J\|_\infty \leq \delta$，于是我们有

$$D \leq \frac{1}{n}\mathbb{E}_{x_i, x_i'}\left[\left|\sum_{i=1}^n g_J(x_i, x_i')\right|\right] + \|(\hat{\ell}_j(x) - \ell_J(x)) - (\hat{\ell}_j(x') - \ell_J(x')\|_\infty \leq \frac{1}{n}\mathbb{E}_{x_i, x_i'}\left[\left|\sum_{i=1}^n g_J(x_i, x_i')\right|\right] + \delta. \quad (66)$$

T然后我们定义 r_j:=max$\{A, \sqrt{\mathbb{E}_{x'}[\ell_j(x')\ -\ \ell^\circ(x')]}\}(j = 1, 2, \cdots, N)$，以及一个随机变量 e

$$G := \max_{1\leq j\leq N}\left|\sum_{i=1}^n \frac{g_j(x_i, x_i')}{r_j}\right|,$$

其中 $A > 0$ 是稍后调整的常数。接着，我们进一步将（66）估算为

$$D \leq \frac{1}{n}\mathbb{E}_{x_i, x_i'}[r_J G] + \delta \leq \frac{1}{n}\sqrt{\mathbb{E}_{x_i, x_i'}[r_J^2]\mathbb{E}_{x_i, x_i'}[G^2]} + \delta \leq \frac{1}{2}\mathbb{E}_{x_i, x_i'}[r_J^2] + \frac{1}{2n^2}\mathbb{E}_{x_i, x_i'}[G^2] + \delta, \qquad (67)$$

根据柯西-施瓦茨不等式和算术-几何平均不等式。根据 J 的定义可得 t

$$\mathbb{E}_{x_i,x_i'}[r_J^2] \le A^2 + \mathbb{E}_{x'}[\ell_J(x') - \ell^\circ(x')] \le A^2 + \mathbb{E}_{x'}[\hat{\ell}(x') - \ell^\circ(x')] + \delta = R(\hat{\ell}, \ell^\circ) + A^2 + \delta. \tag{68}$$

由于 x_i 和 x_i'相互独立，我们有

$$\begin{aligned}\mathbb{E}_{x_i,x_i'}\left[\left(\sum_{i=1}^n \frac{g_j(x_i,x_i')}{r_j}\right)^2\right] &\le \sum_{i=1}^n \mathbb{E}_{x_i,x_i'}\left[\left(\frac{g_j(x_i,x_i')}{r_j}\right)^2\right] \\ &= \sum_{i=1}^n \left(\mathbb{E}_{x_i,x_i'}\left[\frac{(\ell_j(x_i)-\ell^\circ(x_i))^2}{r_j^2}\right] + \mathbb{E}_{x_i,x_i'}\left[\frac{(\ell_j(x_i')-\ell^\circ(x_i'))^2}{r_j^2}\right]\right) \\ &\le 2C_\ell n \end{aligned} \tag{69}$$

成立，其中我们用到了以下事实：$g_j(x_i, x_i')$是中心化的，且 $|\ell_j(x) - \ell^\circ(x)|$ 被 C_ℓ所界。此外，$\frac{g_j(x_i,x_i')}{r_j}$ 被 C_ℓ/A所界。然后，利用伯恩斯坦不等式，我们有

$$\mathbb{P}[G^2 \ge t] = \mathbb{P}[G \ge \sqrt{t}] \le 2N \exp\left(-\frac{t}{2C_\ell(2n + \frac{\sqrt{t}}{3A})}\right),$$

对于任意 $t \ge 0$，均成立。这给出了对 $\mathbb{E}_{x_i,x_i'}[G^2]$的评估。对于任意 $t_0 > 0$，我们有

$$\begin{aligned}\mathbb{E}_{x_i,x_i'}[G^2] &= \int_0^\infty \mathbb{P}[G^2 \ge t]\mathrm{d}t \\ &\le t_0 + \int_{t_0}^\infty \mathbb{P}[G^2 \ge t]\mathrm{d}t \\ &\le t_0 + 2N\int_{t_0}^\infty \exp\left(-\frac{t}{8C_\ell n}\right)\mathrm{d}t + 2N\int_{t_0}^\infty \exp\left(-\frac{3A\sqrt{t}}{4C_\ell}\right)\mathrm{d}t.\end{aligned}$$

这两个积分计算如下

$$\begin{aligned}\int_{t_0}^\infty \exp\left(-\frac{t}{8C_\ell n}\right)\mathrm{d}t &= \left[-8C_\ell n \exp\left(-\frac{t}{8C_\ell n}\right)\right]_{t_0}^\infty = 8C_\ell n \exp\left(-\frac{t_0}{8C_\ell n}\right) \\ \int_{t_0}^\infty \exp\left(-\frac{3A\sqrt{t}}{4C_\ell}\right)\mathrm{d}t &= \int_{t_0}^\infty \exp\left(-a\sqrt{t}\right)\mathrm{d}t \qquad (a := 3A/4C_\ell) \\ &= \left[-\frac{2(a\sqrt{t}+1)}{a^2}\exp(-a\sqrt{t})\right]_{t_0}^\infty \\ &= \frac{8C_\ell\sqrt{t_0}}{3A}\exp\left(-\frac{3A\sqrt{t_0}}{4C_\ell}\right) + \frac{32C_\ell}{9A^2}\exp\left(-\frac{3A\sqrt{t_0}}{4C_\ell}\right).\end{aligned}$$

我们取 $A = \sqrt{t_0}6n$，使得

$$\begin{aligned}\mathbb{E}_{x_i,x_i'}[G^2] &\le t_0 + 2N\left(8C_\ell n + 16C_\ell n + \frac{128C_\ell n^2}{t_0}\right)\exp\left(-\frac{t_0}{8C_\ell n}\right) \\ &\le t_0 + 16N\exp\left(-\frac{3A\sqrt{t_0}}{4C_\ell}\right)n(3 + 16n/t_0)\exp\left(-\frac{t_0}{8C_\ell n}\right)\end{aligned}$$

成立。此外，我们取 $t_0 = 8C_\ell n \log N$，于是可得

$$\mathbb{E}_{x_i,x_i'}[G^2] \le 18C_\ell n\left(\log N + 6 + \frac{2}{C_\ell \log N}\right). \tag{70}$$

现在，我们结合（67）、（68）、（70）和 $A^2 = \frac{2C_\ell}{\log_{9/7} N}$，得到

$$\begin{aligned} D &\le \left(\frac{1}{2}R(\hat{\ell},\ell^\circ) + \frac{1}{2}A^2 + \frac{1}{2}\delta\right) + \frac{4C_\ell}{n}\left(\log N + 6 + \frac{2}{C_\ell \log N}\right) + \delta \\ &\le \frac{1}{2}R(\hat{\ell},\ell^\circ) + \frac{C_\ell}{n}\left(\frac{37}{9}\log N + 32\right) + \frac{3}{2}\delta, \end{aligned}$$

其中我们用到了 $\log N \ge 1$。因此，我们得到

$$R(\hat{\ell},\ell^\circ) \le 2\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n(\hat{\ell}(x_i) - \ell^\circ(x_i))\right] + \frac{2C_\ell}{n}\left(\frac{37}{9}\log N + 32\right) + 3\delta. \tag{71}$$

对于任意固定的 $\ell \in \mathcal{L}$,

$$\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n(\hat{\ell}(x_i) - \ell^\circ(x_i))\right] \le \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n(\ell(x_i) - \ell^\circ(x_i))\right] = \mathbb{E}_x[\ell(x) - \ell^\circ(x)].$$

右侧在 $\inf_{\ell\in\mathcal{L}} \mathbb{E}_x[\ell(x) - \ell^\circ(x)]$时取最小值。最后，将其与 (71) 结合，我们得到

$$R(\hat{\ell},\ell^\circ) \le 2\inf_{\ell\in\mathcal{L}} \mathbb{E}_x[\ell(x) - \ell^\circ(x)] + \frac{2C_\ell}{n}\left(\frac{37}{9}\log N + 32\right) + 3\delta.$$

根据引理 C.3，我们有

$$R(\hat{\ell},\ell^\circ) \le 2\inf_{s\in\mathcal{S}} \int_{\underline{T}}^{\overline{T}} \int_x \|s(x,t) - \nabla \log p_t(x)\|_2^2 p_t(x)\mathrm{d}x\mathrm{d}t + \frac{2C_\ell}{n}\left(\frac{37}{9}\log N + 32\right) + 3\delta.$$

□

C.4. 对 t 和 x_t 进行采样，而非取期望

本节论证第 4.1 节中提出的两种方法。我们假设 $\delta^{-1}, \underline{T}^{-1}, \overline{T}, N = \mathsf{poly}(n)$。首先给出以下引理。该引理表明，$\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\|$ 是次高斯的。

引理 C.5. 令我们从 $i_j \sim \mathsf{Unif}(\{1,2,\cdots,n\})$、$t_j \sim \mathsf{Unif}(\mathrm{T},\mathrm{T})$ 和 $x_j \sim p_{t_j}(x_j|x_{0,i_j})$ 中采样 (i_j,t_j,x_j)。那么，对所有 $t > 0$,

$$\mathbb{P}\left[\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| \ge \sup_{(x,t)} \|s(x,t)\| + \frac{\sqrt{dt}}{\sigma_{\underline{T}}}\right] \le 2\exp\left(-t^2/2\right).$$

证明。首先注意到

$$\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| \le \|s(x_j,t_j)\| + \|\nabla p_{t_j}(x_j|x_{0,i_j})\| \le \sup_{x,t}\|s(x,t)\| + \|\nabla p_{t_j}(x_j|x_{0,i_j})\|.$$

由于 $\nabla p_{t_j}(x_j|x_{0,i_j}) = \frac{x_j - m_t x_{0,i_j}}{\sigma_t^2}$ 且 $x_j \sim p_{t_j}(x_j|x_{0,i_j}) = \mathcal{N}\left(m_t x_{0,i_j}, \sigma_t^2\right)$，我们有 $[\nabla p_{t_j}(x_j|x_{0,i_j})]_i$ 是次高斯的，且 σ_t^{-1}。因此，$\|\nabla p_{t_j}(x_j|x_{0,i_j})\|$ 是次高斯的，且 $\sqrt{d}\sigma_t^{-1}$。现在应用 $\sigma_t \ge \sigma_{\underline{T}}$，即可得到该断言。

□

现在，我们给出第一种方法的以下定理。

定理 C.6. 令我们从 $i_j \sim$均匀分布$(\{1,2,\cdots,n\})$、$t_j \sim$均匀分布(T,T) 和 $x_j \sim p_{t_j}(x_j|x_{0,\ i})$ 中抽取样本 (i_j,t_j,x_j)。令 s_1 为

$$\frac{1}{M}\sum_{j=1}^M \|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i})\|^2$$

并令 s_2 为

$$\frac{1}{n}\sum_{i=1}^{n}\ell(x_i)=\frac{1}{n}\sum_{i=1}^{n}\int_{t=\underline{T}}^{\overline{T}}\|s(x_t,t)-\nabla p_t(x_t|x_{0,i})\|^2 p_t(x_t|x_{0,i})\mathrm{d}x_t\mathrm{d}t,$$

关于 $\mathcal{S}\subseteq\Phi(L,W,S,B)$，其中 $s\in\mathcal{S}$满足 $\|\|s(\cdot,t)\|_2\|_{L^\infty}=\mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}}n)\lesssim\mathcal{O}(\sigma_{\underline{T}}^{-1}\log^{\frac{1}{2}}n)=:C_s$。于是，我们有

$$\mathbb{E}_{\{(i_j,t_j,x_j)\}_{i=1}^n}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_1(x_i)-\frac{1}{n}\sum_{i=1}^{n}\ell_2(x_i)\right|\lesssim\frac{C_s^2+\sigma_{\underline{T}}^{-2}}{M}2SL\log(\delta^{-1}L\|W\|_\infty(B\vee1)(C_s))+\delta.$$

证明。为简洁起见，我们记 (i_j,t_j,x_j) 为 $=y_j$，并令 $Y=\{(i_j,t_j,x_j)\}_{j=1}^M=\{y_j\}_{j=1^\circ}^M$ 令 $Y'=\{(i'_j,t'_j,x'_j)\}_{j=1}^M=\{y'_j\}_{j=1}^M$为 Y 的一个副本，它与 Y 独立。我们记 $\kappa(y_j)=\|s(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,\ i_j})\|^2$。于是，我们有

$$\begin{aligned}&\mathbb{E}_Y\left|\frac{1}{M}\sum_{j=1}^{M}\kappa_1(y_j)-\frac{1}{M}\sum_{j=1}^{M}\kappa_2(y_j)-\frac{1}{n}\sum_{i=1}^{n}\ell_1(x_i)-\frac{1}{n}\sum_{i=1}^{n}\ell_2(x_i)\right| \qquad (72)\\ &=\mathbb{E}_Y\left|\frac{1}{M}\sum_{j=1}^{M}(\kappa_1(y_j)-\kappa_2(y_j))-\mathbb{E}_{Y'}\left[\frac{1}{M}\sum_{j=1}^{M}(\kappa_1(y'_j)-\kappa_2(y'_j))\right]\right|\\ &\le\mathbb{E}_{Y,Y'}\left|\frac{1}{M}\sum_{j=1}^{M}((\kappa_1(y_j)-\kappa_2(y_j))-(\kappa_1(y'_j)-\kappa_2(y'_j)))\right|. \qquad (73)\end{aligned}$$

接下来，令 C_s 为满足 $C_s\ge\sup_{s\in\mathcal{S}}\sup_{x,t}\|s(x,t)\|$ 的最小整数；对于 $i=1,2,\cdots$，我们将 $\mathcal{E}_i$ 定义为一个事件，在该事件中 $C_s+\frac{\sqrt{d}(i-1)}{\sigma_{\underline{T}}}\le\sup_{s\in\mathcal{S}}\max_j\max\{\|s(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,i_j})\|,\|s(x'_j,t'_j)-\nabla p_{t'_j}(x'_j|x_{0,i'_j})\|\}<C_s+\frac{\sqrt{d}i}{\sigma_{\underline{T}}}$ 成立。对于 $i=0$，我们将 $\mathcal{E}_0$ 定义为一个事件，在该事件中 $\sup_{s\in\mathcal{S}}\max_j\max\{\|s(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,i_j})\|,\|s(x'_j,t'_j)-\nabla p_{t'_j}(x'_j|x_{0,i'_j})\|\}<C_s$ 成立。我们令 $a_i=\mathbb{P}[\mathcal{E}_i]$ 和 $\mathbb{E}_i$ 为在事件 $\mathcal{E}_i$ 条件下的期望。那么，(73) 有如下界：

$$\mathbb{E}_0\left|\frac{1}{M}\sum_{j=1}^{M}((\kappa_1(y_j)-\kappa_2(y_j))-(\kappa_1(y'_j)-\kappa_2(y'_j)))\right|+\sum_{i=1}^{\infty}a_i\mathbb{E}_i\left|\frac{1}{M}\sum_{j=1}^{M}((\kappa_1(y_j)-\kappa_2(y_j))-(\kappa_1(y'_j)-\kappa_2(y'_j)))\right|. \qquad (74)$$

我们注意到，对于每个 $\mathbb{E}_i$，$\frac{1}{M}\sum_{j=1}^{M}((\kappa_1(y_j)-\kappa_2(y_j))-(\kappa_1(y'_j)-\kappa_2(y'_j)))$ 有界于 $8C_s^2+\frac{8di^2}{\sigma_t^2}$。这里，$\kappa_1$ 是 $\frac{1}{M}\sum_{j=1}^{M}\kappa(y_j)$ 的最小化器，而 κ_2 是 $\mathbb{E}[\kappa(y)]$ 的最小化器。此外，由于 $\|(x_j-x_{0,i_j})/\sigma_t\|=\|\nabla p_{t_j}(x_j|x_{0,i_j})\|\le\|s(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,i_j})\|+\|s(x_j,t_j)\|$，我们有 $\|s(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,i_j})\|\le C_s+\frac{\sqrt{d}i}{\sigma_{\underline{T}}}$ 蕴含 $\|x_j\|\le2C_s+\sqrt{d}i$。我们应用与定理 C.4 中相同的论证，得到

$$\begin{aligned}&\mathbb{E}_i\left|\frac{1}{M}\sum_{j=1}^{M}\kappa_1(y_j)-\frac{1}{M}\sum_{j=1}^{M}\kappa_2(y_j)-\frac{1}{n}\sum_{i=1}^{n}\ell_1(x_i)-\frac{1}{n}\sum_{i=1}^{n}\ell_2(x_i)\right|\\ &\lesssim\frac{C_s^2+\sigma_{\underline{T}}^{-2}i^2}{M}\log\mathcal{N}(\mathcal{S},L^\infty([-(2C_s+\sqrt{d}i),2C_s+\sqrt{d}i]^{d+1}),\delta/(C_s+i\sigma_{\underline{T}}^{-1}))+\delta.\\ &\lesssim\frac{C_s^2+\sigma_{\underline{T}}^{-2}i^2}{M}2SL\log(\delta^{-1}L\|W\|_\infty(B\vee1)(C_s+i))+\delta.\end{aligned}$$

我们指出，y_j 和 y'_j 在以 $\mathcal{E}_i$为条件时并不独立。然而，在（69）中，类似的论证仍然成立。原证明中我们利用了 x_i 和 x'_i 的独立性，这是因为 y_j 和 y'_j 的对称性并未被破坏

通过取条件期望。基于此，以及 $a_i \le 2\exp(-(i-1)^2/2)\,(i \ge 1)$，根据引理C.5，我们将（74）计算为

(74)

$$
\begin{aligned}
&\lesssim \frac{C_s^2+\sigma_{\underline{T}}^{-2}}{M} SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \delta + \sum_{i=1}^{\infty} a_i \left[\frac{C_s^2+\sigma_{\underline{T}}^{-2} i^2}{M} SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s+i)) + \delta\right] \\
&\lesssim \frac{C_s^2+\sigma_{\underline{T}}^{-2}}{M} SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \delta \\
&+ \sum_{i=1}^{\infty} \exp\left(-\frac{(i-1)^2}{2}\right)\left[\frac{C_s^2+\sigma_{\underline{T}}^{-2} i^2}{M} 2SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s+i)) + \delta\right] \\
&\lesssim \frac{C_s^2+\sigma_{\underline{T}}^{-2}}{M} SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \delta.
\end{aligned}
$$

这给出了（72）的上界。因此，我们最终得到

$$
\begin{aligned}
&\mathbb{E}_{\{y_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n \ell_1(x_i) - \frac{1}{n}\sum_{i=1}^n \ell_2(x_i)\right] \\
&\le \mathbb{E}_{\{y_i\}_{j=1}^M}\left[\frac{1}{M}\sum_{j=1}^M \kappa_1(y_j) - \sum_{j=1}^M \kappa_2(y_j)\right] + \frac{C_s^2+\sigma_{\underline{T}}^{-2}}{M} SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \delta \\
&\le \frac{C_s^2+\sigma_{\underline{T}}^{-2}}{M} SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \delta,
\end{aligned}
$$

因为 κ_1 是 $\frac{1}{M}\sum_{j=1}^M \kappa(y_j)$ 的最小化器。现在，我们得到该断言。 □

注释 C.7. 当 $\|s(x,t)\| = \sqrt{\log N/\sigma_t}$成立时，$\underline{T} = \mathrm{poly}(N^{-1})$ $\overline{T} = \mathcal{O}(\log N)$，我们有 $\sup_{(x,t)}\|s(x,t)\| = C_s \lesssim \sqrt{\underline{T}^{-1}\log N}$。我们设 $N = n^{\frac{d}{2s+d}}, \delta = n^{-\frac{2s}{d+2s}}$并使用定理 3.1 中的网络类，得到

$$
\begin{aligned}
\mathbb{E}_{(i_j,t_j,x_j)}\left[\frac{1}{n}\sum_{i=1}^n \ell_1(x_i)\right] - \int_{\ell_s:\, s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^n \ell_s(x_i) &\lesssim \frac{C_s^2+\sigma_{\underline{T}}^{-2}}{M} 2SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \delta \\
&\lesssim \frac{\underline{T}^{-1}\log n + \underline{T}^{-1}}{M} n^{-\frac{d}{2s+d}}\log^{16} n \lesssim \frac{n^{-\frac{d}{2s+d}}\log^{17} n}{\underline{T}M}.
\end{aligned}
$$

接下来，我们给出第二种方法的证明。

定理 C.8. 我们从 $\mu(t)$ 中采样 $t_j \propto \frac{\mathbb{1}[\underline{T}\le t\le \overline{T}]}{t}$，并将 $\lambda(t)$ 修改为 $\lambda(t) = {}^{t}\log\frac{\overline{T}/\underline{T}}{\overline{T}-\underline{T}}$ 同时将 i_j、x_j采样为$i_j \sim$ 均匀分布$(\{1,2,\cdots,n\})$ 和 $x_j \sim p_{t_j}(x_j|x_{0,i})$。然后，$\mathcal{S} \subseteq \Phi(L,W,S,B)$ 上的最小化器 s_1 满足

$$
\frac{1}{M}\sum_{j=1}^M \lambda(t_j)\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i})\|^2
$$

满足

$$
\mathbb{E}_{(i_j,t_j,x_j)}\left[\frac{1}{n}\sum_{i=1}^n \ell_1(x_i)\right] - \int_{\ell_s:\, s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^n \ell_s(x_i) \lesssim \frac{C_s^2+\overline{T}}{M} SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s)) + \delta,
$$

这里，$C_s = \sup_{t,x}\sqrt{\lambda(t)}\|s(x,t)\|$。

证明。我们只需将前一个引理中的 $\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i})\|$替换为 $\sqrt{\lambda(t_j)}\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i})\|$。与引理 C.5类似，我们有，对于所有 $t>0$,

$$
\mathbb{P}\left[\lambda^{\frac{1}{2}}(t_j)\|s(x_j,t_j) - \nabla p_{t_j}(x_j|x_{0,i_j})\| \ge \sup_{(x,t)}\lambda^{\frac{1}{2}}(t)\|s(x,t)\| + \frac{\sqrt{d}\lambda^{\frac{1}{2}}(t_j)t}{\sigma_{t_j}}\right] \le 2\exp\left(-t^2/2\right).
$$

然后，我们分别将 $\sup_{(x,t)}\|s(x,t)\|$ 替换为 $\sup_{(x,t)}\lambda^{\frac{1}{2}}(t)\|s(x,t)\|$，并将 $\frac{\sqrt{d}}{\sigma_T}$ 替换为 $\sup_t \sqrt{d}\lambda^{\frac{1}{2}}(t)\sigma_t$，从而得到

$$\mathbb{E}_{i_j,t_j,x_j}\mathbb{E}_{i'_j,t'_j,x'_j}\left[\lambda(t_j)\|s_1(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,i_j})\|^2\right]-\inf_{s\in\mathcal{S}}\mathbb{E}_{i_j,t_j,x_j}\left[\lambda(t_j)\|s(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,i_j})\|^2\right]$$
$$\lesssim \frac{C_s^2+\overline{T}}{M}SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s))+\delta, \tag{75}$$

其中，(i'_j,t'_j,x'_j) 是 (i_j,t_j,x_j) 的独立副本。注意，

$$\mathbb{E}_{i_j,t_j,x_j}\left[\lambda(t_j)\|s(x_j,t_j)-\nabla p_{t_j}(x_j|x_{0,i_j})\|^2\right]=\frac{1}{n}\sum_{i=1}^n \ell(x_i) \tag{76}$$

对所有（固定的）s。（75）和（76）可得

$$\mathbb{E}_{(i_j,t_j,x_j)}\left[\frac{1}{n}\sum_{i=1}^n \ell_1(x_i)\right]-\int_{\ell_s:\, s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^n \ell_s(x_i)\le \frac{C_s^2+\overline{T}}{M}SL\log(\delta^{-1}L\|W\|_\infty(B\vee 1)(C_s))+\delta,$$

由此完成证明。 □

注 C.9。当 $\|s(x,t)\|=\sqrt{\log N/\sigma_t}$ 成立时，$\underline{T}=\text{poly}(N^{-1})$ $\overline{T}=\mathcal{O}(\log N)$，我们有 $\sup_{(x,t)}\sqrt{\lambda(t)}\|s(x,t)\|=C_s\lesssim\sqrt{\log N}$。令 $N=n^{\frac{d}{2s+d}}$, $\delta=n^{-\frac{2s}{d+2s}}$，并使用定理 3.1 中的网络类，可得

$$\mathbb{E}_{(i_j,t_j,x_j)}\left[\frac{1}{n}\sum_{i=1}^n \ell_1(x_i)\right]-\int_{\ell_s:\, s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^n \ell_s(x_i)\lesssim n^{-\frac{2s}{d+2s}}\log^{17} n.$$

D. 估计误差分析

以下 Girsanov 定理在将分数匹配误差转换为估计误差时很有用。命题 D.1 (吉尔萨诺夫定理（Karatzas 等人，1991））。设 p_0 为任意概率分布，且 $Z=(Z_t)^{t\in[0,T]}$、$Z'=(Z'_t)^{t\in[0,T]}$ 为满足以下条件的两个不同过程

$$\begin{aligned}\mathrm{d}Z_t&=b(Z_t,t)\mathrm{d}t+\sigma(t)\mathrm{d}B_t,\quad Z_0\sim p_0,\\ \mathrm{d}Z'_t&=b'(Z'_t,t)\mathrm{d}t+\sigma(t)\mathrm{d}B_t,\quad Z'_0\sim p_0.\end{aligned}$$

我们将 Z_t 和 Z'_t 的分布分别记为 p_t 和 p'_t，将 Z 和 Z' 的路径测度分别记为 $\mathbb{P}$ 和 $\mathbb{P}'$。假设满足以下诺维科夫条件：

$$\mathbb{E}_{\mathbb{P}}\left[\exp\left(\int_0^T\frac{1}{2}\int_x\sigma^{-2}(t)\|(b-b')(x,t)\|^2\mathrm{d}x\mathrm{d}t\right)\right]<\infty. \tag{77}$$

那么，$\mathbb{P}$ 相对于 $\mathbb{P}'$ 的拉东-尼科迪姆导数为

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}'}(Z)=\exp\left\{-\frac{1}{2}\int_0^T\sigma(t)^{-2}\|(b-b')(Z_t,t)\|^2\mathrm{d}t-\int_0^T\sigma(t)^{-1}(b-b')(Z_t,t)\mathrm{d}B_t\right\},$$

因此，我们有

$$\mathrm{KL}(p_T|p'_T)\le\mathrm{KL}(\mathbb{P}|\mathbb{P}')=\int_0^T\frac{1}{2}\int_x p_t(x)\sigma(t)^{-2}\|(b-b')(x,t)\|^2\mathrm{d}x\mathrm{d}t.$$

此外，Chen 等人（2022年）证明，若 $\int_x p_t(x)\sigma^{-2}(t)\|(b-b')(x,\ t)\|^2\mathrm{d}x\le C$ 对某个常数 C 在所有 t上均成立，则有

$$\mathrm{KL}(p_T|p'_T)\le\int_0^T\frac{1}{2}\int_x p_t(x)\sigma(t)^2\|(b-b')(x,t)\|^2\mathrm{d}x\mathrm{d}t,$$

即使不满足诺维科夫条件（77）。

D.1. 全变差距离下的估计界

我们将在本小节中展示全变差距离下的上界和下界估计速率。令 Y 为将 $\hat{Y}_0 \sim \mathcal{N}(0, I_d)$ 替换为 $\bar{Y}_0 \sim p_t$ 后的 Y。首先注意到

$$\begin{aligned}\mathbb{E}[\mathrm{TV}(\mathrm{X}_0, \hat{\mathrm{Y}}_{\overline{T}-\underline{T}})] &\lesssim \mathbb{E}[\mathrm{TV}(\mathrm{Y}_{\overline{T}}, \mathrm{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\ &\lesssim \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{T}}) + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0)] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\ &= \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{T}}) + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d))] + \mathbb{E}[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \end{aligned} \tag{78}$$

这里，$\mathbb{E}[\mathrm{TV}(\mathrm{Y}_{\overline{T}}, \mathrm{Y}_{\overline{T}-\underline{T}})] = \mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{T}}) + \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0)]$ 可由正向过程与反向过程之间的对应关系推出，而 $\mathbb{E}[\mathrm{TV}(Y_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \le \mathbb{E}[\mathrm{TV}(X_{\overline{T}}, Y_0)]$ 则可由 Y 和 Y 的定义推出（唯一的区别在于初始分布）。接下来，我们依次对式（78）中的三项进行界定。先从第一项开始。

定理 D.2. 我们有

$$\mathrm{TV}(\mathrm{X}_0, \mathrm{X}_{\underline{T}}) \lesssim \sqrt{\underline{T}} n^{\mathcal{O}(1)} \lesssim n^{-s/(d+2s)}$$

对于 $\underline{T} = n^{-\mathcal{O}(1)}$。

证明。我们需要计算 $\|p_0 - p_{\underline{T}}\|_{L^1}$。记住，$p_0$ 可分解为

$$f_N(x) = \sum_{i=1}^{N} \alpha_i \mathbb{I}[\|x\|_\infty \le 1] M^d_{k_i, j_i}(x)$$

在引理 F.13 中，其中 $\|k\|_\infty \le K^* = (\mathcal{O}(1) + \log N)\nu^{-1} + \mathcal{O}(d^{-1} \log N)$ 对于 $\delta = d(1/p - 1)_+$ 和 $\nu = (2s - \delta)/(2\delta)$，且 $\|p_0 - f_N\|_{L^1([-1,1]^d)} \lesssim N^{-s/d} \simeq n^{-s/(2s+d)}$ 成立。由于我们取 $N = n^{d/(2s+d)} = n^{\mathcal{O}(1)}$，因此可以说每个 $M^d_{k_i,j_i}(x)$ 都是 $n^{\mathcal{O}(1)}$-利普希茨的。此外，$|\alpha_i| \lesssim N^{(\nu^{-1}+d^{-1})(d/p-s)_+} = n^{\mathcal{O}(1)}$。因此，$f_N$ 是 $n^{\mathcal{O}(1)}$-利普希茨的。

我们将 p_0 分解为 $p_0 = f_N + (p_0 - f_N)$，其中使用上述 f_N。于是我们有

$$\begin{aligned}&\left| p_{\underline{T}}(x) - \int \frac{f_N(y)}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \right| \\ &= \left| \int \frac{(p_0(y) - f_N(y))}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \right| \\ &\le \int \frac{|p_0(y) - f_N(y)|}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y. \end{aligned} \tag{79}$$

对所有 x 积分可得

$$\begin{aligned}\int \left| p_{\underline{T}}(x) - \int \frac{f_N(y)}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \right| \mathrm{d}x &\le \int \int \frac{|p_0(y) - f_N(y)|}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \mathrm{d}x \\ &= \int |p_0(y) - f_N(y)| \int \frac{1}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}x \mathrm{d}y \\ &\le \int |p_0(y) - f_N(y)| \, \mathrm{d}y = \|p_0 - f_N\|_{L^1([-1,1]^d)}. \end{aligned}$$

因此，$\|p_0 - p_{\underline{T}}\|_{L_1}$ 的上界为

$$\|p_0 - f_N\|_{L^1([-1,1]^d)} + \underbrace{\int \left| f_N(x) - \int \frac{f_N(y)}{\sigma^d_{\underline{T}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma^2_{\underline{T}}} \right) \mathrm{d}y \right| \mathrm{d}x}_{\text{if } f_N \text{ is replaced by } p_0, \text{ this is equal to } \|p_0 - p_t\|_{L_1}} + \underbrace{\|p_0 - f_N\|_{L^1([-1,1]^d)}}_{(79)}. \tag{80}$$

由于 $\|p_0 - f_N\|_{L^1([-1,1]^d)}$ 由 $n^{-s/(2s+d)}$ 界定，我们将重点放在第二项上。

注意，在每个 x，

$$\left| \int \frac{f_N(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - \int_{A^x} \frac{f_N(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \right| \lesssim n^{-s/(d+2s)}, \tag{81}$$

其中 $A^x = \prod_{i=1}^d a_i^x$ 满足 $a_i^x = [\frac{x_i}{m_{\underline{T}}} - \frac{\sigma_{\underline{T}}\mathcal{O}(1)}{m_{\underline{T}}}\sqrt{\log n}, \frac{x_i}{m_{\underline{T}}} + \frac{\sigma_{\underline{T}}\mathcal{O}(1)}{m_{\underline{T}}}\sqrt{\log n}]$，根据引理 F.10。由于 $\sigma_{\underline{T}} = \mathcal{O}(\sqrt{\underline{T}})$ 且对于充分小的 $\underline{T}$，$m_{\underline{T}} = \mathcal{O}(1)$，因此 $p_{\underline{T}}(x)$ 的值几乎完全由距离 x 仅为 $\mathcal{O}(\sqrt{\underline{T}\log n})$ 的点决定。由于 p_0 的 Lipschitz 连续性，对于每个 $x \in [-m_{\underline{T}} - \mathcal{O}(\sqrt{\underline{T}\log n})\ m_{\underline{T}} + \mathcal{O}(\sqrt{\underline{T}\log n})]^d$，

$$\left| \int_{A^x} \frac{f_N(y)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - \int_{A^x} \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \right| \le n^{\mathcal{O}(1)} \cdot \sqrt{\underline{T}\log n}. \tag{82}$$

其中我们用到了 f_N 的利普希茨连续性。令 $\underline{T}$ 相对于 n 多项式级小，则有 (82) $\lesssim n^{-s/(d+2s)}$。此外，

$$\begin{aligned} &\left| \int_{A^x} \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - f_N(x) \right| \\ &= \left| \int_{A^x} \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y - \int \frac{f_N(x)}{\sigma_{\underline{T}}^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - m_{\underline{T}} y\|^2}{2\sigma_{\underline{T}}^2}\right) \mathrm{d}y \right| \lesssim n^{-s/(d+2s)}, \end{aligned} \tag{83}$$

再次使用引理F.10。

因此，结合（80）、（81）、（82）和（83），我们得到

$$\|p_0 - p_{\underline{T}}\|_{L_1} \lesssim \sqrt{\underline{T}} n^{\mathcal{O}(1)} \lesssim n^{-s/(d+2s)}.$$

对于 $\underline{T} = n^{-\mathcal{O}(1)}$。 □

接下来我们考虑第二项。

引理 D.3。我们可以如下界定全变差$(X_{\overline{T}}, \mathcal{N}(0, I_d))$。

$$\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d)) \lesssim \exp(-\underline{\beta}\overline{T}).$$

证明。奥恩斯坦–乌伦贝克过程的指数收敛（Bakry 等人，2014年）推出

$$\mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d)) \lesssim \sqrt{\mathrm{KL}(p_{\overline{T}} \| \mathcal{N}(0, I_d))} \le \exp(-\underline{\beta}\overline{T}) \sqrt{\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d))} \lesssim \exp(-\underline{\beta}\overline{T}),$$

因为 $C_f^{-1} \le p_0 \le C_f$ 成立，且 $\mathcal{N}(0, I_d)$ 的密度在 p_0 的支集 $\mathrm{supp}(p_0)$ 上以下界 $\gtrsim 1$ 为界 $= [-1, 1]^d$，这意味着 $\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d)) = \mathcal{O}(1)$。 □

¯第三项 $\mathbb{E}[$全变差$(Y_{\overline{T}} - \underline{T},\ Y_{\overline{T}-\underline{T}})]$ 在 (78) 中由吉尔萨诺夫定理、命题 D.1 以及第4节中的 (4:

$$\begin{aligned} \mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}}) &\lesssim \mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \sqrt{\int_{t=\underline{T}}^{\overline{T}} p_t(x) \beta_t^{-2} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t} \\ &\lesssim \sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \int_{t=\underline{T}}^{\overline{T}} p_t(x) \beta_t^{-2} \|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t} \\ &\lesssim \sqrt{n^{-\frac{2s}{d+2s}} \log^{18} n} \\ &\lesssim n^{-\frac{s}{d+2s}} \log^9 n. \end{aligned}$$

因此，中的三个项 （78）与上述内容有界，因此定理 5.1 得证。我们还展示如下下界 ws.

命题 D.4。假设 $0 < p、\ q \le \infty、\ s > 0$ 和

$$s > \left\{ d\left(\frac{1}{p} - \frac{1}{2}\right), d\left(\frac{1}{p} - 1\right), 0 \right\}$$

成立。因此，我们有

$$\inf_{\hat{\mu}} \sup_{p \in B^s_{p,q}([-1,1]^d)} \mathbb{E}[\mathrm{TV}(\hat{\mu}, p)] \gtrsim n^{-s/(d+2s)},$$

其中期望是针对样本取的，下确界是在所有基于 n 个观测的估计量中取得的 .

证明。 Triebel 的定理 10 Triebel (2011年) 指出，对于有界区域 $\Omega \subset \mathbb{R}^d$,

$$\log N(U(B^s_{p,q}(\Omega)), \|\cdot\|_r, \varepsilon) \simeq \varepsilon^{-d/s}, \tag{84}$$

对于满足以下条件的 $0 < p、\ q \le \infty、\ 1 \le r < \infty$ 和 $s > 0$

$$s > \max\left\{ d\left(\frac{1}{p} - \frac{1}{r}\right), d\left(\frac{1}{p} - 1\right), 0 \right\}.$$

尽管他们考虑了所有不满足 $\int f \mathrm{d}\mu = 1$ 的 Besov 函数，但我们可以根据他们的证明验证，对函数进行有界约束并不会影响熵数的阶。现在我们使用 Yang 与 Barron（1999）的定理 4。注意，覆盖数与熵等价，是因为 $\|\cdot\|_r$ 是一个距离，因此（84）可以转化为熵的结果。定理的条件 2 可直接由（84）验证。此外，若对所有 $\alpha \in (0,1)$，取 $f_*(x) = 1/2^d (x \in [-1,1]^d)$，0（否则），则条件 3 成立。最后，若 $s > \left\{ d(\frac{1}{p} - \frac{1}{2})\ d(\frac{1}{p} - 1)\ 0 \right\}$，则 $\log N(U(B^s_{p,q}(\Omega)) \|\cdot\|_2, \varepsilon) \simeq \log N(U(B^s_{p,q}(\Omega)) \|\cdot\|_1, \varepsilon)$ 成立。因此，应用 Yang 与 Barron（1999）的定理 4（i），可得

$$\min_{\hat{\mu}} \max_{p \in B^s_{p,q}} \mathbb{E}[\|\hat{\mu} - p\|_1] \simeq \varepsilon_n,$$

其中 ε_n 取为 为 $\log N(U(B^s_{p,q}(\Omega)) \|\cdot\|_r, \varepsilon_n) = n\varepsilon_n^2$ 成立。结合（84），我们得到该断言。 □

D.2. W_1 距离下的估计率

类似于（78），我们有如下分解：

$$\begin{aligned}\mathbb{E}[W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}})] &\le \mathbb{E}[W_1(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\ &\le \mathbb{E}[W_1(X_0, X_{\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})]. \end{aligned} \tag{85}$$

首先，我们对（85）中的第一项进行界定。

引理 D.5. 我们可以如下界定 $W_1(X_0, X_{\underline{T}})$。

$$W_1(X_0, X_{\underline{T}}) \lesssim \sqrt{\underline{T}}$$

证明。令 $X \sim p_0$ 和 $Z \sim N(0, I_d)$。那么，

$$\begin{aligned} W_1(X_0, X_{\underline{T}}) &\le \mathbb{E}[\|X - m_{T_1} X + \sigma_{T_1} Z\|] \le (1 - m_{\underline{T}})\mathbb{E}[\|X\|] + \sigma_{\underline{T}}\mathbb{E}[\|Z\|] \\ &\le (1 - m_{\underline{T}})\sqrt{d} + \sigma_{\underline{T}}\sqrt{d} \lesssim \sqrt{\underline{T}}, \end{aligned}$$

至此，证明完成。 □

接下来，我们对式（85）的第二项进行界定。

引理 D.6。我们可以将 $\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})]$ 界定如下。

$$\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] \lesssim \mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) \lesssim \exp(-\underline{\beta}\overline{T}).$$

Proof. 奥恩斯坦–乌伦贝克过程的指数收敛（Bakry 等人，2014年）表明 t

$$\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) = \mathrm{TV}(p_{\overline{T}}, \mathcal{N}(0, I_d)) \le \sqrt{2\mathrm{KL}(p_{\overline{T}} \| \mathcal{N}(0, I_d))} \le 2\exp(-\overline{T}\underline{\beta})\sqrt{\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d))} \lesssim \exp(-\underline{\beta}\overline{T})$$

因为 $C_f^{-1} \le p_0 \le C_f$ 成立，且 $\mathcal{N}(0, I_d)$的密度在 $\mathrm{supp}(p_0) = [-1, 1]^d$上以下界 $\mathcal{O}(1)$ 为界，这意味着 $\mathrm{KL}(p_0 \| \mathcal{N}(0, I_d)) = \mathcal{O}(1)$。此外，因为 $\|\hat{Y}^{(k)}_{\overline{T}-\underline{T}}\|_\infty, \|\hat{Y}_{\overline{T}-\underline{T}}\|_\infty \le 2 = \mathcal{O}(1)$，并且 $\hat{Y}^{(k)}$与 $\hat{Y}$之间唯一的差别是初始分布，所以我们有 $W_1(\hat{Y}^{(k)}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})$。$\mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) = \mathrm{TV}(p_{\overline{T}}, \mathcal{N}(0, I_d))$。综合以上结果，我们得到

$$W_1(\hat{Y}^{(k)}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}}) \lesssim \mathrm{TV}(X_{\overline{T}}, \hat{Y}_0) = \mathrm{TV}(p_{\overline{T}}, \mathcal{N}(0, I_d)) \lesssim \exp(-\underline{\beta}\overline{T}),$$

由此可得该断言。 □

最后，我们对第三项（85）进行界定。正如我们在第 5.2 节中所见，

$$\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \le \sum_{i=1}^{K_*} \mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})]. \tag{86}$$

回顾随机过程序列 $\{(\hat{Y}^{(i)\,t})_{t=0}^{\overline{T}-\underline{T}}\}_{i=0}^{K_*}$.的定义。首先，$\bar{Y}^{(0)} = (\bar{Y}^{(0)\,t})_{t\in[0,\overline{T}]} = Y = (Y_t)_{t\in[0,\overline{T}]}$ 定义为满足以下条件的过程：

$$\mathrm{d}Y_t = \beta_{\overline{T}-t}(Y_t + 2\nabla \log p_t(Y_t, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [0, \overline{T}]), \quad Y_0^{(0)} \sim p_{\overline{T}}.$$

然后，$Y_{\overline{T}-t} \sim p_t$对所有 $t \in [0, \overline{T}]$均成立。接下来，对于 $i = 1, 2, \cdots, K_*$，令 $\bar{Y}^{(i)} = (\bar{Y}^{(i)\,t})_{t\in[0,\overline{T}-\underline{T}]}$ 满足：

$$\begin{aligned}
\bar{Y}_0^{(i)} \sim p_{\overline{T}}, \quad \mathrm{d}\bar{Y}_t^{(i)} &= \beta_{\overline{T}-t}(\bar{Y}_t^{(i)} + 2\nabla \log p_t(\bar{Y}_t^{(i)}, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [0, \overline{T}-t_i]), \\
\mathrm{d}\bar{Y}_t^{(i)} &= \beta_{\overline{T}-t}(\bar{Y}_t^{(i)} + 2\hat{s}(\bar{Y}_t^{(i)}, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_i, \overline{T}-\underline{T}]).
\end{aligned}$$

注意，$t_0 = \underline{T}, t_1 = N^{-\frac{2-\delta}{d}} = n^{-\frac{2-\delta}{d+2s}}$, $1 < \frac{t_{i+1}}{t_i} =$为常数。 ≤ 2，且 $t_{K_*} = \overline{T} - \underline{T}$。于是，$\bar{Y}^{(K_*)} = \bar{Y}$成立。这里，$\bar{Y}^{(i)}{}_{\overline{T}-t} \sim p_t$ 对所有 $t \in [0, \overline{T}-t_i]$均成立，但在 $t = \overline{T} - t_i$之后，真实得分函数会被估计得分函数取代。如果原始定义中的$\|\bar{Y}^{(i)}{}_{\overline{T}-\underline{T}}\|_\infty > 2$，则我们将 $\bar{Y}^{(i)}{}_{\overline{T}-\underline{T}}$重置为 $\bar{Y}^{(i)}{}_{\overline{T}-\underline{T}} := 0$。

此外，我们还引入另一个随机过程 $\bar{Y}^{(i)\prime}$。我们将 $d+1$ 维集合 $A \subseteq \mathbb{R}^{d+1}$ 定义为

$$A = \left\{(x, t) \in \mathbb{R}^d \times \mathbb{R} \,\middle|\, \|x\|_\infty \le m_t + C_{\mathrm{a},1}\sigma_t\sqrt{\log(n)}, \ \underline{T} \le t \le \overline{T}\right\}.$$

根据引理 A. 1, $1 - n^{-\mathcal{O}(1)}$ Y_t $_{t=0}^{\overline{T}}$ Y_t $\overline{T}-t$ $\in$以至少 的概率，反向过程（ ）的一条路径满足（ , ）, A $\underline{T} \le t \le \overline{T}$ 对所有 均成立。基于此，对于 $i = 0$, 1 , $\cdots$, $K_* - 1$, $\bar{Y}^{(i)\prime}$ ()定义为

$$\begin{aligned}
\bar{Y}_0^{(i)\prime} &\sim p_{\overline{T}}, \\
\mathrm{d}\bar{Y}_t^{(i)\prime} &= \beta_{\overline{T}-t}(\bar{Y}_t^{(i)\prime} + 2\nabla \log p_t(\bar{Y}_t^{(i)\prime}, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [0, \overline{T}-t_i]), \\
\mathrm{d}\bar{Y}_t^{(i)\prime} &= \beta_{\overline{T}-t}\Big(\bar{Y}_t^{(i)\prime} + 2_{\mathbb{I}}[(\bar{Y}_s^{(i)\prime}, \overline{T}-s) \notin A \text{ for some } s \le t]\nabla \log p_t(\bar{Y}_t^{(i)\prime}) \\
&\qquad + 2_{\mathbb{I}}[(\bar{Y}_s^{(i)\prime}, \overline{T}-s) \in A \text{ for all } s \le t]\hat{s}(\bar{Y}_t^{(i)\prime}, \overline{T}-t)\Big)\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_{i+1}, \overline{T}-t_i]), \\
\mathrm{d}\bar{Y}_t^{(i)\prime} &= \beta_{\overline{T}-t}(\bar{Y}_t^{(i)\prime} + 2\hat{s}(\bar{Y}_t^{(i)\prime}, \overline{T}-t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_i, \overline{T}-\underline{T}]).
\end{aligned}$$

引理 D.7。假设 $\|\hat{s}(\cdot,t)\|_\infty \lesssim$ 对数$\frac{\overline{T}}{\sqrt{t}\wedge 1}$成立。则对所有 $i=1,2,\cdots,K_*$，以下结论均成立：

$$W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}},\bar{Y}^{(i)}_{\overline{T}-\underline{T}}) \lesssim \sqrt{t_i\log n}\sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=t_{i-1}}^{t_i}\mathbb{E}_x\left[\|\hat{s}(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}t\right]\right]}+n^{-\frac{s+1}{d+2s}}. \tag{87}$$

因此，我们有

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}},\bar{Y}^{(i)}_{\overline{T}-\underline{T}})] \lesssim \sqrt{t_i\log n}\sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\int_{t=t_{i-1}}^{t_i}\mathbb{E}_x\left[\|\hat{s}(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}t\right]\right]}+n^{-\frac{s+1}{d+2s}}. \tag{88}$$

证明。我们构造 $\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}$与 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$之间的传输映射。我们的方法着眼于每条路径。

由于 $\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}$和 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$不满足诺维科夫条件，因此无法使用命题 D.1 来考察这两条路径之间的全变差距离；命题 D.1 只能给出 KL($\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}},\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$)，而非 KL($\bar{Y}^{(i-1)}$ $\bar{Y}^{(i)}$，这个界不足以支持我们的讨论。因此，我们首先对 $\mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}},\bar{Y}^{(i-1)\prime})]$进行界定。根据引理 A.1，以至少 $1-n^{-\mathcal{O}(1)}$ 的概率，过程 $(\bar{Y}^{(i-1)}_t)_{t=0}^{\overline{T}}$和 $(\bar{Y}^{(i-1)\prime})_{t=0}^{\overline{T}}$的一条路径满足 $(\bar{Y}^{(i-1)}_t,\overline{T}-t)$，$(\bar{Y}^{(i-1)\prime},\overline{T}-t)\in A$对所有 $0\le t\le \overline{T}-t_{i-1}$ 均成立。因此，$\mathbb{E}[\mathrm{TV}(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}},\bar{Y}^{(i-1)\prime})]$由 $n^{-\mathcal{O}(1)}$ 界定（$\mathcal{O}(1)$中的常数取充分大）。这意味着 $\mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}},\bar{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}})]\lesssim n^{-\mathcal{O}(1)}$，因为$\mathrm{TV}(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}},\bar{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}})=\mathcal{O}(1)$（几乎处处成立）。

我们现在讨论 $\mathbb{E}[W_1(\bar{Y}^{(i-1)\prime},\bar{Y}^{(i)}_{\overline{T}-\underline{T}})]$。设 $\bar{Y}^{(i-1)\prime}$和 $\bar{Y}^{(i)}$的路径测度分别为 $\mathbb{P}$和 $\mathbb{P}'$，并取一条路径 p，它在$t=\overline{T}-\underline{T}$处为 y，在$t=\overline{T}-t_i$处为 z。若 $\mathrm{d}\mathbb{P}[p]>\mathrm{d}\mathbb{P}'[p]$，那么我们沿路径 p 反向时间移动，直到 $t=\overline{T}-t_i$，将 $\bar{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}}=y$中质量为$\mathrm{d}\mathbb{P}[p]-\mathrm{d}\mathbb{P}'[p]$的部分移至 z。对所有路径 p应用此操作后，被移动的 $\bar{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}}$的总质量至多为

$$\frac{1}{2}\mathrm{TV}((\bar{Y}^{(i-1)\prime}),(\bar{Y}^{(i)})) \le \frac{1}{2}\sqrt{\int_{t=t_{i-1}}^{t_i}\int_x p_t(x)\beta_t^{-2}\|\hat{s}(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}x\mathrm{d}t}. \tag{89}$$

根据 命题D.1。这里我们指出，在这种情况下，诺维科夫条件显然成立 e.
到目前为止，$\hat{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}}$ 的一部分质量沿着各条对应的路径移动，但此时尚未构造出耦合测度。为了实现该耦合测度，我们对 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$ 考虑相同的过程。也就是说，对于每条路径 p with$\bar{Y}^{(i)}_{\overline{T}-\underline{T}}=y$and $\bar{Y}^{(i)}_{\overline{T}-t_i}=z$，如果 $\mathrm{d}\mathbb{P}[p]<\mathrm{d}\mathbb{P}'[p]$，那么我们将 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}=y$ 的质量沿路径 p 移动至 z，移动量为 $\mathrm{d}\mathbb{P}'[p]-\mathrm{d}\mathbb{P}[p]$。受影响的 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$ 的总质量不超过 1 2 全变差($(\bar{Y}^{(i-1)\prime})$，$(\bar{Y}^{(i)\prime})$)，而后者由（89）式界定。

现在我们可以看到，从 $\bar{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}}$ 和 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$运送到 $t=\overline{T}-t_i$的质量相同。因此，在每个 z处，我们可以任意将来自 $\bar{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}}$ 的质量与来自 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$的质量配对。利用这一点，最多有 1 2 TV($(\bar{Y}^{(i-1)\prime})$, $(\bar{Y}^{(i)\prime})$) 的质量会从 $\bar{Y}^{(i-1)\prime}_{\overline{T}-\underline{T}}$ 运送到 $\bar{Y}^{(i)}_{\overline{T}-\underline{T}}$，方法是将路径反向至 $t=\overline{T}-t_i$。

现在我们的 interest 是每次运输平均需要移动多远。首先，我们考虑 $t_i\lesssim 1$.
首先，我们对 $\|\bar{Y}^{(i)}_{\overline{T}-\underline{T}}-\bar{Y}^{(i)}_{\overline{T}-t_i}\|$进行界定。根据引理 A.1，对于所有 $t\in[\overline{T}-t_i,\overline{T}-\underline{T}]$，我们有 $\|\int_{\overline{T}-t_i}^{\overline{T}-\underline{T}}2\beta_{\overline{T}-t}\mathrm{d}B_t\|\lesssim\sqrt{t_i\log n}$，并且 $\bar{Y}^{(i)}_{\overline{T}-t_i}\lesssim m_{\overline{T}-t_i}+\sigma_{\overline{T}-t_i}\sqrt{\log n}\lesssim\sqrt{\log n}$ 以概率 $1-n^{-\mathcal{O}(1)}$ 成立。我们考虑事件

以.....为条件 m。注意，$\|s(x,t)\| \lesssim \sqrt{\log \frac{n}{\sigma_t}} \lesssim \sqrt{\log \frac{n}{\sqrt{t}}}$成立。于是，对于所有 $\overline{T}-t_i \le t \le \overline{T}-\underline{T}$，我们有 ，

$$\begin{aligned}
\|\bar{Y}_t^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\| &= \left\| \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \beta_{\overline{T}-s}(\bar{Y}_s^{(i)} + 2\nabla \log p_t(\bar{Y}_s^{(i)}, \overline{T}-s))\mathrm{d}t + \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \sqrt{2\beta_{\overline{T}-s}}\mathrm{d}B_s \right\| \\
&\lesssim \overline{\beta} \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)}\|\mathrm{d}s + 2\overline{\beta} \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \frac{\sqrt{\log n}}{\sqrt{s}}\mathrm{d}s + \sqrt{t_i \log n}, \\
&\lesssim \overline{\beta} \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)}\|\mathrm{d}s + \sqrt{t_i \log n} + \sqrt{t_i \log n}. \\
&\lesssim \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\|\mathrm{d}s + \sqrt{t_i \log n} + t_i \|\bar{Y}_{\overline{T}-t_i}^{(i)}\| \\
&\lesssim \int_{\overline{T}-t_i}^{\overline{T}-\underline{T}} \|\bar{Y}_s^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\|\mathrm{d}s + \sqrt{t_i \log n} + t_i \sqrt{\log n}
\end{aligned}$$

现在我们应用格朗沃尔不等式，得到

$$\|\bar{Y}_t^{(i)} - \bar{Y}_{\overline{T}-t_i}^{(i)}\| \lesssim e^{\overline{\beta} t_i} \sqrt{t_i \log n} \lesssim \sqrt{t_i \log n}.$$

对所有 $\overline{T}-t_i \le t \le \overline{T}-\underline{T}$均成立。因此，以概率 $1-n^{-\mathcal{O}(1)}$，$\|\bar{Y}_{(i)\,t} - \bar{Y}_{(i)\,\overline{T}-t_i}\|$至多为$\sqrt{t_i \log n}$乘以一个常数因子（对所有 $\overline{T}-t_i \le t \le \overline{T}-\underline{T}$均成立）。

接下来我们对 $\|\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)'} - \bar{Y}_{\overline{T}-t_i}^{(i-1)'}\|$ 进行界定。将其分解为

$$\|\bar{Y}_{\overline{T}-t_i}^{(i-1)'} - \bar{Y}_{\overline{T}-t_{i-1}}^{(i-1)'}\| + \|\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)'} - \bar{Y}_{\overline{T}-t_{i-1}}^{(i-1)'}\|.$$

第一项以至少 $1-n^{-\mathcal{O}(1)}$ 的概率被$\lesssim \sqrt{t_i \log n}$ 所界定。这是因为，根据引理 A.1 的第一部分，$\bar{Y}_{\overline{T}-t_{i-1}}^{(i-1)'} \in A$ 以概率 $1-n^{-\mathcal{O}(1)}$成立；对于这类路径，$\bar{Y}_{\overline{T}-t_{i-1}}^{(i-1)'}$的演化与$Y_t$相同，在其中我们应用引理 A.1 的第二部分。根据关于 $\|\bar{Y}_{(i)\,t} - \bar{Y}_{(i)\,\overline{T}-t_i}\|$的讨论，第二项以概率$1-n^{-\mathcal{O}(1)}$被 $\sqrt{t_{i-1}\log n}$所界定。综上，以概率 $1-n^{-\mathcal{O}(1)}$，我们可以将 $\|\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)'} - \bar{Y}_{\overline{T}-t_i}^{(i-1)'}\|$ 界定为 $\sqrt{t_{i-1}\log n}(\le \sqrt{t_i \log n})$ 乘以一个常数因子。

总之，当 $t_i \lesssim 1$时，传输映射以概率 $1-n^{-\mathcal{O}(1)}$至多移动 $\mathcal{O}(\sqrt{t_i \log n})$。由于 $\bar{Y}_{(i-1)'}$和 $\bar{Y}_{(i)\overline{T}-\underline{T}}$的支撑集均有界，对于移动距离超过 $\sqrt{t_i}\log n$的质量，其对Wasserstein距离的影响至多为 $n^{-\mathcal{O}(1)}$。因此，我们得到了关于 $t_i \lesssim 1$ 的所需界（87）。

对于 $t_i \gtrsim 1$，由于 $\bar{Y}_{(i-1)\overline{T}-\underline{T}}$和 $\bar{Y}_{(i)\overline{T}-\underline{T}}$的支撑集均有界，

$$W_1(\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)}, \bar{Y}_{\overline{T}-\underline{T}}^{(i)}) \lesssim \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}^{(i-1)}, \bar{Y}_{\overline{T}-\underline{T}}^{(i)}) \lesssim \frac{1}{2}\sqrt{\int_{t=t_{i-1}}^{t_i} \int_x p_t(x)\beta_t^{-2}\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}x\mathrm{d}t}$$

成立。因此，我们同样得到（87）。

由（87），根据詹森不等式可轻易得到（88）。

□

此外，我们对每个网络 s_i的泛化误差进行界定。

引理 D.8. 对于 $1 \le i \le K_* - 1$，设 s_i 是从 $\Phi(L, W, S, B)$ 中选取的一个网络，满足

$$L = \mathcal{O}(\log^4 n),\ \|W\|_\infty = \mathcal{O}(n^{\frac{d}{d+2s}}),\ S = \mathcal{O}(t_i^{-d/2} n^{\frac{\delta d}{2(2s+d)}}),\ \textit{and}\ B = \exp(\mathcal{O}(\log^4 n)),$$

1</style}且 (,)$\log_{\sigma_t}^{\overline{2}} n$。那么，我们有

$$\mathbb{E}_{\{x_{0,j}\}_{i=j}^n}\left[\int_{t=t_i}^{t_{i+1}}\mathbb{E}_x\left[\|\hat{s}_i(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}t\right]\right]\lesssim n^{-\frac{2(s+1)}{d+2s}}\log n+\frac{t_i^{-d/2}n^{\frac{\delta d}{2(d+2s)}}\log^{10}n}{n}.$$

此外，对于 $i=0$，令 s_0 为从 $\Phi(L,W,S,B)$ 中选取的一个网络，满足

$$L=\mathcal{O}(\log^4 n),\ \|W\|_\infty=\mathcal{O}(n^{\frac{d}{d+2s}}\log^6 n),\ S=\mathcal{O}(n^{\frac{d}{2s+d}}\log^8 n),\ \textit{and}\ B=\exp(\mathcal{O}(\log^4 n)),$$

1且 $\|s_0(\cdot,t)\|_{L^\infty}\lesssim\log_{\sigma_t}^{\overline{2}} n$。那么，我们有

$$\mathbb{E}_{\{x_{0,j}\}_{i=j}^n}\left[\int_{t=t_i}^{t_{i+1}}\mathbb{E}_x\left[\|\hat{s}_0(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}t\right]\right]\lesssim n^{-\frac{2s}{d+2s}}\log^{18}n.$$

证明。首先考虑第一部分。我们取 $N=n^d d+2s$ 和 $t_*=t_i/2$，见引理 3.6。注意，N 和 $t_*(\geq n^{\frac{2-\delta}{d+2s}})$满足 $t_*\geq N^{-(2-\delta)/d}$(，这正是定理 B.8中的假设)。那么，存在一个神经网络 $\phi\in\Phi(L,\ W,\ S,\ B)$ 满足

$$\int_{t=t_i}^{t_{i+1}}\int_x p_t(x)\|\phi(x,t)-s(x,t)\|^2\mathrm{d}x\mathrm{d}t\lesssim N^{-\frac{2(s+1)}{d}}\log n=N^{-\frac{2(s+1)}{d+2s}}\log n.$$

具体而言，$L=\mathcal{O}(\log^4(n))$ $\|W\|_\infty=\mathcal{O}(n^{\frac{d}{d+2s}})$ $S=\mathcal{O}(t_i^{-d/2}n^{\frac{\delta d}{(d+2s)}})$，且 $B=\exp(\mathcal{O}(\log^4 n))$。因此，我们通过将 和 $\overline{T}$分别替换为 t_i 和 t_{i+1}，并取 $\delta=n^{-\frac{2(s+1)}{d+2s}}$，应用 (64)，得到第一条断言如下：

$$\begin{aligned}\mathbb{E}_{\{x_{0,j}\}_{i=j}^n}\left[\int_{t=t_i}^{t_{i+1}}\mathbb{E}_x\left[\|\hat{s}_i(x,t)-\nabla\log p_t(x)\|^2\mathrm{d}t\right]\right]&\lesssim N^{-\frac{2(s+1)}{d}}\log n+\frac{C_\ell}{n}\log\mathcal{N}+\delta\\&\lesssim n^{-\frac{2(s+1)}{d+2s}}\log n+\frac{\log^2 n}{n}\left(t_i^{-d/2}n^{\frac{\delta d}{2(d+2s)}}\log^8\right)+n^{-\frac{2(s+1)}{d+2s}}\\&\lesssim n^{-\frac{2(s+1)}{d+2s}}\log n+\frac{t_i^{-d/2}n^{\frac{\delta d}{2(d+2s)}}\log^{10}n}{n}.\end{aligned}$$

对于第二部分，我们只需沿用推导出（4）的讨论，将 $\overline{T}$ 替换为 t_1（$\overline{T}$），这不会增加泛化误差。

□

定理 *5.4* 的证明。我们使用引理 D.8 中给出的网络序列。具体而言，我们考虑以下过程。

$$\hat{Y}_0^{(i)}\sim\mathcal{N}(0,I),\quad \mathrm{d}\hat{Y}_t^{(i)}=\beta_{\overline{T}-t}(\hat{Y}_t^{(i)}+2\hat{s}(\hat{Y}_t^{(i)},\overline{T}-t))\mathrm{d}t+\sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t\ (t\in[\overline{T}-t_i,\overline{T}-t_{i+1}],i=0,1,\cdots,K_*),$$

并且，当 $\|\hat{Y}^{(i)}_{\overline{T}-\underline{T}}\|_\infty>2$ 时，我们将 $\hat{Y}^{(i)}_{\overline{T}-\underline{T}}$ 修改为 0。

最后，我们对上述过程中的误差进行求和。式（86）可进一步界定为

$$
\begin{aligned}
&\mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\
&\le \sum_{i=1}^{K_*} \mathbb{E}[W_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})]. \\
&\lesssim \sum_{i=1}^{K_*} \left[\sqrt{t_{i-1}\log n} \sqrt{\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \left[\int_{t=t_i}^{t_i} \mathbb{E}_x \left[\|\hat{s}(x,t) - \nabla \log p_t(x)\|^2 \mathrm{d}t \right] \right]} + n^{-\frac{s+1}{d+2s}} \right] \quad \text{(by Lemma D.7)} \\
&\lesssim \sum_{i=2}^{K_*} \left[\sqrt{t_i \log n} \left(n^{-\frac{(s+1)}{d+2s}} \sqrt{\log n} + \frac{t_i^{-d/4} n^{\frac{\delta d}{4(d+2s)}} \log^5 n}{\sqrt{n}} \right) + n^{-\frac{(s+1)}{d+2s}} \right] \\
&\quad + \sqrt{t_1 \log n} \left[n^{-\frac{s}{d+2s}} \log^9 n + n^{-\frac{s}{d+2s}} \right] \quad \text{(by Lemma D.8)} \\
&\lesssim \left[\sqrt{t_1} n^{-\frac{s}{d+2s}} + \sqrt{t_1} \frac{t_1^{-d/4} n^{\frac{\delta d}{4(d+2s)}}}{\sqrt{n}} \right] \cdot \tilde{\mathcal{O}}(1) \\
&\quad \text{(because } K_* = \mathcal{O}(\log n) \text{ and } t_1 \le \cdots t_{K_*} = \mathcal{O}(\log N) \text{ with } 1 < t_{i+1}/t_i = \text{const.} \le 2 \ (i \ge 1).) \\
&= \left[(n^{-\frac{2-\delta}{d+2s}})^{\frac{1}{2}} n^{-\frac{s}{d+2s}} + (n^{-\frac{2-\delta}{d+2s}})^{\frac{1}{2}} \frac{(n^{-\frac{2-\delta}{d+2s}})^{-d/4} n^{\frac{\delta d}{4(d+2s)}}}{\sqrt{n}} \right] \cdot \tilde{\mathcal{O}}(1) \\
&\lesssim n^{-\frac{(s+1-\delta)}{d+2s}}.
\end{aligned} \tag{90}
$$

因此，取 $\underline{T} \lesssim n^{-\frac{2(s+1)}{d+2s}}$和 $\overline{T} = (s+1)\log n \cdot \underline{T} \cdot \frac{d+2s}{\delta} \cdot \log n$，我们得到

$$
\begin{aligned}
W_1(X_0, \hat{Y}_{\overline{T}-\underline{T}}) &\le \mathbb{E}[W_1(X_0, X_{\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, \hat{Y}_{\overline{T}-\underline{T}})] + \mathbb{E}[W_1(\bar{Y}_{\overline{T}-\underline{T}}, Y_{\overline{T}-\underline{T}})] \\
&\lesssim \sqrt{\underline{T}} + \exp(-\underline{\beta}\overline{T}) + n^{-\frac{(s+1-\delta)}{d+2s}} \quad \text{(by Lemmas D.5 and D.6 and (90))} \\
&\lesssim n^{-\frac{(s+1-\delta)}{d+2s}} + n^{-\frac{(s+1-\delta)}{d+2s}} + n^{-\frac{(s+1-\delta)}{d+2s}} \lesssim n^{-\frac{(s+1-\delta)}{d+2s}},
\end{aligned}
$$

由此完成定理 5.4的证明。 □

D.3. 离散化误差讨论

与第 5.3 节一样，$t_0 = \underline{T} < t_1 < \cdots < t_{K_*} = \overline{T}$为满足 $t_{k+1} - t_k \equiv \eta \ll 1$ 的时间步。考虑以下过程 $(Y_{\mathrm{d}t})_{t=0}^{\eta K} = (Y_{\mathrm{d}t})_{t=0}^{\overline{T}-\underline{T}}$，其中 $Y_{\mathrm{d}_0} \sim \mathcal{N}(0, I_d)$：

$$
\mathrm{d}Y_t^{\mathrm{d}} = \beta_t (Y_t^{\mathrm{d}} + 2\hat{s}(Y^{\mathrm{d}}_{\overline{T}-t_i}, \overline{T} - t_i))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \quad (t \in [\overline{T} - t_i,\ \overline{T} - t_{i-1}]).
$$

这里，$\hat{s}$ 是通过分数匹配得到的得分网络：

$$
\hat{s} \in \operatorname{argmin} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \eta \mathbb{E}[\|s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k}|x_{0,i})\|^2]. \tag{91}
$$

这里，每个期望都是相对于 $x_{\overline{T}-t_k} \sim p_{\overline{T}-t_k}(x_{\overline{T}-t_k}|x_{0,i})$ 取的。

定理 D.9. 设 $\underline{T} = n^{-\mathcal{O}(1)}, \overline{T} = s\log\frac{n}{2s+d}$，且 $\eta = \mathrm{poly}(n^{-1})$。那么，

$$
\mathbb{E}[\mathrm{TV}(X_0, \bar{Y}_{\overline{T}-\underline{T}})] \lesssim n^{-\frac{2s}{d+2s}} \log^{18} n + \eta^2 \underline{T}^{-3} \log^3 n + \eta \underline{T}^{-1} \log^3 n + \eta \log^4 n.
$$

证明。 我们首先证明，最小化器 $\hat{s}$ 在 Φ'（见第 4节）

$$
\hat{s} \in \operatorname{argmin} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=} \eta \mathbb{E}[\|s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k}|x_{0,i})\|^2].
$$

满足

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\sum_{k=1}^K \eta \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|\hat{s}(x_{t_k},t_k)-\nabla \log p_{t_k}(x_{t_k})\|^2]\right] \lesssim n^{-2s/(2s+d)}\log^{18} n. \tag{92}$$

我们取 $N = n^{\frac{d}{d+2s}}$ 根据定理3.1，对于 $N \gg 1$，存在一个神经网络 ϕ_{score}，其 $L = \mathcal{O}(\log^4 N)$ $\|W\|_\infty = \mathcal{O}(N\log^6 N)$ $S = \mathcal{O}(N\log^8 N)$，并且 $B = \exp(\mathcal{O}(\log^4 N))$ 满足

$$\int_x p_t(x)\|\phi_{\text{score}}(x,t)-s(x,t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d}}\log(N)}{\sigma_t^2}. \tag{93}$$

对所有 $t \in [\underline{T}, \overline{T}]$ 均成立。将其对所有 $t = t_k$ 求和，可得

$$\begin{aligned}&\sum_{k=1}^K \eta \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|\phi_{\text{score}}(x_{t_k},t_k)-\nabla \log p_{\eta k}(X_{t_k})\|^2] \lesssim \sum_{k=1}^K \eta \frac{N^{-\frac{2s}{d}}\log(N)}{1\wedge t_k} \\ &\le N^{-\frac{2s}{d}}\log(N)\left(\eta K + \eta\sum_{k=1}^K \frac{1}{t_k}\right) \lesssim N^{-\frac{2s}{d}}\log(N)(\overline{T}+\log(\overline{T}/\underline{T})) \lesssim N^{-\frac{2s}{d}}\log^2(N).\end{aligned} \tag{94}$$

为了将其转化为泛化界，我们需要评估以下两项。首先，$\hat{s}$ 可以取为使得

$$\sup_x \|\phi_{\text{score}}(x,t)\|\mathrm{d}x \lesssim \frac{\log^{\frac{1}{2}}(N)}{\sigma_t},$$

因此，我们按照第 s 节中的做法对其进行截断 4。由于这样的 s 满足

$$\int_x p_t(x)\|\phi_{\text{score}}(x,t)-\nabla \log p_t(x)\|^2 \mathrm{d}x \lesssim \frac{\log(N)}{\sigma_t^2},$$

我们可得

$$\sum_{k=1}^K \eta \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|\phi_{\text{score}}(x_{t_k},t_k)-\nabla \log p_{t_k}(x_{t_k})\|^2] \le C_\ell = \mathcal{O}(\log^2(n))$$

（沿用引理 C.1 的论证，以及我们如何从（94）推导出（93））。其次，$\ell(x) = \sum_{k=1}^K \eta\mathbb{E}[\|s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k}|x)\|^2]$ 在所有满足 s 的 $\delta = n^{-\frac{2s}{d+2s}}$ 网络类的覆盖数有界于 $n^{\frac{d}{d+2s}}\log^{16} n$，这可通过附录C.2中的论证得到。因此，定理 C.4 可修改为适用于此情形，并得到

$$\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\sum_{k=1}^K \eta \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|s(x_{t_k},t_k)-\nabla \log p_{t_k}(x_{t_k})\|^2]\right] \lesssim n^{-s/(2s+d)}\log^2 n.$$

成立。因此，根据第 4 节中的讨论，我们有

$$\begin{aligned}&\mathbb{E}_{\{x_{0,i}\}_{i=1}^n}\left[\sum_{k=1}^K \eta_k \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|s(x_{t_k},t_k)-\nabla \log p_{t_k}(x_{t_k})\|^2]\right] \\ \lesssim &\sum_{k=1}^K \eta \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|\phi_{\text{score}}(x_{t_k},t_k)-\nabla \log p_{\eta k}(X_{t_k})\|^2] + \frac{C_\ell}{n}\log\mathcal{N} + \delta \\ \lesssim &n^{\frac{d}{d+2s}}\log^2 n + \frac{\log^2 n}{n}\cdot n^{\frac{d}{d+2s}}\log^{18} n + n^{-\frac{2s}{d+2s}} \lesssim n^{-\frac{2s}{d+2s}}\log^{18} n,\end{aligned}$$

这证明了（92）。

从现在起，我们界定全变差 $\mathrm{TV}(Y_0, Y_{\mathrm{d}\overline{T}-\underline{T}})$。我们引入以下过程。 $\bar{Y}_{\mathrm{d}} = (\bar{Y}_{\mathrm{d}t})_{t=0}^{\overline{T}-\underline{T}}$的定义方式与 Y_{d} 相同，区别仅在于 $\bar{Y}_{\mathrm{d}0} \sim p_{\overline{T}}$ 的初始分布。在 $t = \overline{T} - \underline{T}$，如果 $\int$-范数大于 2，则将其重置为 0。$\bar{Y} = (\bar{Y}_t)_{t=0}^{\overline{T}-\underline{T}}$定义为 $\bar{Y}_0 \sim p_{\overline{T}}$，并

$$
\begin{aligned}
&\bar{Y}_0 \sim p_{\overline{T}},\\
&\mathrm{d}\bar{Y}_t = \beta_{\overline{T}-t}\big(Y_t + 2_{\mathbb{I}}[(\bar{Y}_s, \overline{T}-s) \notin A \text{ for some } s \le t]\nabla \log p_t(\bar{Y}_t)\\
&\qquad + 2_{\mathbb{I}}[(Y_s, \overline{T}-s) \in A \text{ for all } s \le t]\hat{s}(\bar{Y}_{\overline{T}-t_k}, \overline{T}-t_k)\big)\,\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t \ (t \in [\overline{T}-t_i,\ \overline{T}-t_{i-1}]).
\end{aligned}
$$

在 $t = \overline{T} - \underline{T}$ 处，如果 ∞-范数大于 2，我们就将其重置为 0。这里， $A \subseteq \mathbb{R}^{d+1}$ 定义为

$$
A = \left\{(x,t) \in \mathbb{R}^d \times \mathbb{R} \,\middle|\, \|x\|_\infty \le m_t + C_{\mathrm{a}}\sigma_t\sqrt{\log(n)},\ \underline{T} \le t \le \overline{T}\right\}.
$$

于是，我们有

$$
\begin{aligned}
\mathrm{TV}(Y_{\overline{T}}, Y^{\mathrm{d}}_{\overline{T}-\underline{T}}) &\le \mathrm{TV}(Y_{\overline{T}}, Y_{\overline{T}-\underline{T}}) + \mathrm{TV}(Y_0, \bar{Y}_{\overline{T}-\underline{T}}) + \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}}) + \mathrm{TV}(\bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}})\\
&\le \mathrm{TV}(X_0, X_{\underline{T}}) + \mathrm{TV}(Y_0, \bar{Y}_{\overline{T}-\underline{T}}) + \mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}}) + \mathrm{TV}(X_{\overline{T}}, \mathcal{N}(0, I_d)).
\end{aligned}
$$

第一项由 $n^{-\frac{2s}{d+2s}}$ 控制，方法是在定理 $\underline{T} = n^{-\mathcal{O}(1)}$ 中取 D.2。根据引理 $n^{-\frac{2s}{d+2s}}$，取 C_{a} 充分大，则第二项由 A.1 控制。根据引理 $\exp(-\underline{\beta}\overline{T})$ D.3，第四项由 $\underline{T} = \mathcal{O}(\log\, n)$ 控制，因此取 $\exp(-\underline{\beta}\overline{T}) \lesssim n^{-\frac{2s}{d+2s}}$ 即可。

现在，我们来界定第三项。命题D.1 给出

$$
\begin{aligned}
&\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}})\\
&\lesssim \sum_{k=1}^{K}\int_{t=\overline{T}-t_k}^{\overline{T}-t_{k-1}} \mathbb{E}_{\bar{Y}}[_{\mathbb{I}}[(\bar{Y}_s, \overline{T}-s) \in A \text{ for all } s \le t]\|\hat{s}(\bar{Y}_{\overline{T}-t_k}, \overline{T}-t_k) - \nabla \log p_t(\bar{Y}_t)\|^2]\mathrm{d}t\\
&\le \sum_{k=1}^{K}\int_{t=\overline{T}-t_k}^{\overline{T}-t_{k-1}} \mathbb{E}_{\bar{Y}}[_{\mathbb{I}}[(\bar{Y}_t, \overline{T}-t) \in A, (\bar{Y}_{\overline{T}-t_k}, t_k) \in A]\|\hat{s}(\bar{Y}_{\overline{T}-t_k}, \overline{T}-t_k) - \nabla \log p_t(\bar{Y}_t)\|^2]\mathrm{d}t\\
&\le \sum_{k=1}^{K}\int_{t=t_{k-1}}^{t_k} \mathbb{E}_{X}[_{\mathbb{I}}[(X_t, t) \in A, (X_{t_k}, t_k) \in A]\|\hat{s}(X_{t_k}, t_k) - \nabla \log p_t(X_t)\|^2]\mathrm{d}t\\
&\lesssim \sum_{k=1}^{K}\int_{t=t_{k-1}}^{t_k} \mathbb{E}_{x_{t_k}\sim p_{t_k}}[\|\hat{s}(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k})\|^2]\mathrm{d}t \qquad (95)\\
&\quad + \sum_{k=1}^{K}\int_{t=t_{k-1}}^{t_k} \mathbb{E}_{X}[_{\mathbb{I}}[(X_t, t) \in A, (X_{t_k}, t_k) \in A]\|\nabla \log p_t(X_t) - \nabla \log p_{t_k}(X_t)\|^2]\mathrm{d}t \qquad (96)\\
&\quad + \sum_{k=1}^{K}\int_{t=t_{k-1}}^{t_k} \mathbb{E}_{X}[_{\mathbb{I}}[(X_t, t) \in A, (X_{t_k}, t_k) \in A]\|\nabla \log p_{t_k}(X_t) - \nabla \log p_{t_k}(X_{t_k})\|^2]\mathrm{d}t \qquad (97)
\end{aligned}
$$

首先，我们考虑（96）。由于 $(X_t,\ t) \in A$，$(\|X_t\|_\infty - m_t)_+ \lesssim \sigma_t\sqrt{\log}(n)$。对所有 $t \le s \le t_k$，$|\partial_s\sigma_s| \lesssim \frac{1}{\sqrt{t}}$，$|\partial_s m_s| \lesssim 1$，以及

$$
\|\partial_s \nabla \log p_s(x)\| \lesssim \frac{|\partial_s\sigma_s| + |\partial_s m_s|}{\sigma_s^3}\left(\frac{(\|x\|_\infty - m_s)_+^2}{\sigma_s^2} \vee 1\right)^{\frac{3}{2}} \lesssim \frac{|\partial_t\sigma_{t_k}| + |\partial_t m_{t_k}|}{\sigma_{t_k}^3}\left(\frac{(\|x\|_\infty - m_{t_k})_+^2}{\sigma_{t_k}^2} \vee 1\right)^{\frac{3}{2}},
$$

根据 引理 A.3。因此，(96) 可由 $\sum_{k=1}^{K}\eta(\eta(t_k^{-2} \vee 1)$ 对数$^{\frac{3}{2}}\, n)^2 = \eta^2(t_k^{-4} \vee 1)$ 对数$^3\, n$.

接下来，对于（97），我们首先注意到 $\|X_t\|_\infty - m_{t_k}, \|X_{t_k}\|_\infty - m_{t_k} \lesssim \sigma_{t_k}\sqrt{\log(n)} = \tilde{\mathcal{O}}(1)$。因此，根据引理 A.3， $\|\partial_{x_i}\nabla \log\, p_{t_k}(x)\|$可由 $\frac{1}{\sigma_{t_k}^2}\left(\frac{(\|X_{t_k}\|_\infty - m_{t_k})_+^2}{\sigma_{t_k}^2} \vee 1\right) \lesssim t_k^{-1}\log\, n$上界。其概率至少为 $1 -$

$n^{-\mathcal{O}(1)}$, $\|X_t - X_{t_k}\|_\infty \lesssim \sqrt{\eta \log n}$，根据引理 F.15。因此，

$$(97) \lesssim \sum_{k=1}^{K} \eta(\sqrt{\eta \log n} \cdot (t_k^{-1} \vee 1) \log n)^2 + n^{-\mathcal{O}(1)} \cdot \tilde{\mathcal{O}}(1) \lesssim \sum_{k=1}^{K} \eta^2 (t_k^{-2} \vee 1) \log^3 n.$$

最后，对于 (97)，我们应用 (92)。现在，(95)、(96) 和 (97) 的三项均有界，因此我们得到 th at

$$\begin{aligned}
\mathbb{E}_{\{x_{0,i}\}_{i=1}^n} \left[\mathrm{TV}(\bar{Y}_{\overline{T}-\underline{T}}, \bar{Y}^{\mathrm{d}}_{\overline{T}-\underline{T}}) \right] &\lesssim n^{-\frac{2s}{d+2s}} \log^{18} n + \sum_{k=1}^{K} (\eta^3 (t_k^{-4} \vee 1) \log^3 n + \eta^2 (t_k^{-2} \vee 1) \log^3 n) \\
&\lesssim n^{-\frac{2s}{d+2s}} \log^{18} n + \eta^2 \underline{T}^{-3} \log^3 n + \eta \underline{T}^{-1} \log^3 n + \eta \overline{T} \log^3 n \\
&\lesssim n^{-\frac{2s}{d+2s}} \log^{18} n + \eta^2 \underline{T}^{-3} \log^3 n + \eta \underline{T}^{-1} \log^3 n + \eta \log^4 n.
\end{aligned}$$

因此，令 $\eta = \underline{T}^{-1.5} n^{-\frac{s}{d+2s}}$ 即可得出该断言。

□

E. 内在维数下的误差分析

E.1. 证明概要

得分网络的泛化误差分析，以及得分估计误差对定理中最终估计速率的影响 6.4 只需在先前的分析中将 d 替换为 d' 即可。因此，我们重点关注近似误差界。为得到定理 3.1 和引理 3.6的对应结果，我们旨在将得分函数分解为两部分：一部分由内在结构分量（位于 V中）决定，另一部分由其他分量（位于 $V^\perp$中）决定。我们将 z用作一个与 d'的典范系相对应的 V维向量。对此目标的第一个观察是

$$\begin{aligned}
p_t(x) &= \int \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} p_0(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2} \right) \mathrm{d}y \\
&= \int_V \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} q(z) \exp\left(-\frac{\|A^\top x - m_t z\|^2 + \|(I_d - A^\top)x\|^2}{2\sigma_t^2} \right) \mathrm{d}z \\
&\quad (z \text{ is a } d'\text{-dimensional vector corresponding to the canonical system of } V.) \\
&= \underbrace{\int_V \frac{q(z)}{\sigma_t^{d'} (2\pi)^{\frac{d'}{2}}} \exp\left(-\frac{\|A^\top x - m_t z\|^2}{2\sigma_t^2} \right) \mathrm{d}z}_{p_t^{(1)}(x)} \cdot \underbrace{\frac{1}{\sigma_t^{d-d'} (2\pi)^{\frac{d-d'}{2}}} \exp\left(-\frac{\|(I_d - A^\top)x\|^2}{2\sigma_t^2} \right)}_{p_t^{(2)}(x)}.
\end{aligned}$$

这里，$p_t^{(1)}(x)$ 和 $p_t^{(2)}(x)$ 可分别视为关于内在分量和剩余空间的密度函数。注意，

$$\nabla \log p_t(x) = \nabla \log(p_t^{(1)}(x) p_t^{(2)}(x)) = \nabla \log p_t^{(1)}(x) + \nabla \log p_t^{(2)}(x).$$

因此，我们只需构造近似各项的神经网络，并将它们拼接起来。此外，$p_t^{(1)}(x)$ 可视为 $A^\top x$ 处的密度，对应于 d' 维空间上的扩散过程，其初始密度由 q 定义。因此，我们令

$$q_t(z') = \int_V \frac{q(z)}{\sigma_t^{d'} (2\pi)^{\frac{d'}{2}}} \exp\left(-\frac{\|z' - m_t z\|^2}{2\sigma_t^2} \right) \mathrm{d}z$$

对于 $z' \in \mathbb{R}^{d'}$。这里 $p_t^{(1)}(x) = q_t(A^\top x)$成立。

E.2. 定理 6.4的证明

我们首先考虑 $p_t^{(1)}(x)$的近似。我们有定理 3.1 和引理 3.6的对应结论，唯一的区别在于，这里 d 被替换为 d'。

引理 E.1. 设 $N \gg 1$, $\underline{T} = \mathrm{poly}(N^{-1})$ 和 $\overline{T} = \mathcal{O}(\log N)$。则存在一个神经网络 $\phi_{\mathrm{score},\,3} \in \Phi(L, W, S, B)$，使得对于所有 $t \in [\underline{T}, \overline{T}]$,

$$\int_{x\in\mathbb{R}^d} p_t(x)\|\nabla \log p_t^{(1)}(x) - \phi_{\mathrm{score},3}(A^\top x, t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2s}{d'}}\log(N)}{\sigma_t^2}. \tag{98}$$

这里， L、W、S和 B分别取为 $L = \mathcal{O}(\log^4 N)$ $\|W\|_\infty = \mathcal{O}(N\log^6 N)$ $S = \mathcal{O}(N\log^8 N)$，以及 $B = \exp(\mathcal{O}(\log^4 N))$。我们可以取 $\phi_{\mathrm{score},\,3}$，使其满足 $\|\phi_{\mathrm{score},\,3}(\cdot, t)\|_\infty = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} N)$。

此外，设 $N' \geq t_*^{-d'/2}N^{\delta/2}$和 $t_* \geq N^{-(2-\delta)/d'}$。则存在一个神经网络 $\phi_{\mathrm{score},\,4} \in \Phi(L, W, S, B)$，使其满足

$$\int_{x\in\mathbb{R}^d} p_t(x)\|\nabla \log p_t^{(1)}(x) - A\phi_{\mathrm{score},4}(A^\top x, t)\|^2 \mathrm{d}x \lesssim \frac{N^{-\frac{2(s+1)}{d'}}}{\sigma_t^2} \tag{99}$$

对于 $t \in [2t_*, \overline{T}]$。具体而言， $L = \mathcal{O}(\log^4(N))$ $\|W\|_\infty = \mathcal{O}(N)$ $S = \mathcal{O}(N')$，且 $B = \exp(\mathcal{O}(\log^4 N))$。我们可以取$\phi_{\mathrm{score},\,4}$满足 $\|\phi_{\mathrm{score},\,4}(\cdot, t)\|_\infty = \mathcal{O}(\sigma_t^{-1}\log^{\frac{1}{2}} N)$。

证明。令 ϕ_{score}: $\mathbb{R}^{d'} \times \mathbb{R}_+ \to \mathbb{R}^{d'}$，它近似于 $q_t(z)$。注意

$$\nabla \log p_t^{(1)}(x) = A\nabla \log q_t(A^\top x)$$

因此

$$\begin{aligned}
\int_{x\in\mathbb{R}^d} p_t(x)\|\nabla \log p_t^{(1)}(x) - A\phi_{\mathrm{score}}(A^\top x, t)\|^2 \mathrm{d}x &= \int_{x\in\mathbb{R}^d} p_t^{(1)}(x)p_t^{(2)}(x)\|A\nabla \log p_t^{(1)}(A^\top x) - A\phi_{\mathrm{score}}(A^\top x, t)\|^2 \mathrm{d}x \\
&= \int_{x\in\mathbb{R}^d} q_t(A^\top x)\|A\nabla \log p_t^{(1)}(A^\top x) - A\phi_{\mathrm{score}}(A^\top x, t)\|^2 \mathrm{d}x \\
&= \int_{z\in\mathbb{R}^{d'}} q_t(z)\|\nabla \log q_t(z) - \phi_{\mathrm{score}}(z, t)\|^2 \mathrm{d}z,
\end{aligned}$$

这里我们用到了以下事实： $p_t^{(1)}$ 和 $p_t^{(2)}$ 分别依赖于 $A^\top x$ 和 $(I - A^\top)x$，且 $A^\top x$ 和 $(I - A^\top)x$ 是正交的。此外，我们还用到了 $\det(A^\top A) = 1$ 以及 A的列的正交性。因此，我们可以转换定理 3.1和引理3.6。

□

接下来，我们考虑对 $p_t^{(2)}(x)$的近似。与附录 A中的做法相同，我们首先证明只需考虑有界区域内的近似。

引理 E.2。对于 $\varepsilon > 0$，我们定义$B_{t,\varepsilon}$ 为

$$B_{t,\varepsilon} = \left\{x \in \mathbb{R}^d \,\middle|\, \|(I_d - A^\top)x\| \leq C_{\mathrm{e}}\sigma_t\sqrt{\log \varepsilon^{-1}}.\right\}$$

我们有时将其简写为 B_ε。于是，我们有

$$\int_{x\in\bar{B}_\varepsilon} p_t(x)\left[1 \vee \|\nabla \log(p_t^{(2)}(x))\|^2\right]\mathrm{d}x \lesssim \varepsilon.$$

证明。 A 的列是正交的。 $p_t^{(1)}$ 和 $p_t^{(2)}$ 分别取决于 $A^\top x$ 和$(I - A^\top)x$，而 $A^\top x$ 和

$(I - A^\top)x$ 是正交的。因此，我们有

$$\begin{aligned}\int_{x\in\bar{B}_{t,\varepsilon}} p_t(x)\left[1\vee\|\nabla\log(p_t(x))\|^2\right]\mathrm{d}x &= \int_{x\in\bar{B}_{t,\varepsilon}} p_t^{(1)}(x)p_t^{(2)}(x)\left[1\vee\|\nabla\log(p_t(x))\|^2\right]\mathrm{d}x \\ &= \int_{x\in\bar{B}_{t,\varepsilon}} p_t^{(2)}(x)\left[1\vee\|\nabla\log(p_t(x))\|^2\right]\mathrm{d}x \\ &= \int_{w\in\mathbb{R}^{d-d'}:\,\|w\|\geq C_{\mathrm{e}}\sigma_t\sqrt{\log\varepsilon^{-1}}} \frac{1\vee\|w\|^2/\sigma_t^2}{\sigma_t^{d-d'}(2\pi)^{\frac{d-d'}{2}}}\exp\left(-\frac{\|w\|^2}{2\sigma_t^2}\right)\mathrm{d}w.\end{aligned} \tag{100}$$

应用推论 F.8，(100) 有界于 ε，其中常数 C_{e}足够大。 □

现在我们只需考虑在 $B_{t,\varepsilon}$ 内对 $\nabla\log\ p_t^{(2)}(x)$ 的近似。

引理 E.3. 设 $N \gg 1, \underline{T}, \varepsilon = \mathrm{poly}(N^{-1})$ 且 $\overline{T} \simeq \log\ N$。存在一个神经网络 $\phi_{\mathrm{score},4} \in \Phi(L, W, S, B)$，使得

$$\sup_{t\in[\underline{T},\overline{T}]}\int_x p_t(x)\|\nabla\log p_t^{(2)}(x) - \phi_{\mathrm{score},4}(x,t)\|^2\mathrm{d}x \lesssim \frac{N^{-\frac{2(s+1)}{d'}}}{\sigma_t^2}. \tag{101}$$

具体而言，$\phi_{\mathrm{score},4} \in \Phi(L, W, S, B)$ 成立，其中

$$L = \mathcal{O}(\log^2 N)), \|W\|_\infty = \mathcal{O}(\log^3 N), S = \mathcal{O}(\log^4 N),\ \textit{and}\ B = \exp(\mathcal{O}(\log^2 N)). \tag{102}$$

证明。首先注意 在 $\nabla\log\ p_t^{(2)}(x) = -\frac{1}{\sigma_t^2}(I_d - A)(I_d - A^\top)x_\circ$ 我们通过以下四个步骤对其进行近似 s.

1. σ_t 由 ϕ_σ 根据引理 3.3进行近似。这里设 $\varepsilon \leftarrow (\underline{T}^4 \wedge \varepsilon^4)\varepsilon^4$。
2. 根据近似 σ_t， σ_t^{-2}由 $\phi_{\mathrm{rec}}(\cdot$； 2) 根据推论 F.8进行近似。这里设 $\varepsilon \leftarrow (\underline{T} \wedge \varepsilon)\varepsilon$.
3. $(I_d - A)(I_d - A^\top)$由 $\mathrm{ReLU}((I_d - A)(I_d - A^\top)\ \cdot x + 0) - \mathrm{ReLU}(-(I_d - A)(I_d - A^\top)\ \cdot x + 0)$实现。
4. 根据引理 F.6 以及 $\varepsilon \leftarrow \varepsilon$ 和 $C \leftarrow \underline{T}^{-1} \vee \sqrt{\log\ \varepsilon^{-1}}$， 构造出 σ_t^{-2}和 $(I_d - A)(I_d - A^\top)$ 的乘积。

通过将这些网络串联起来（使用引理 F.1），所得网络的规模有界为 a s

$$\begin{aligned}&L = \mathcal{O}(\log^2\varepsilon^{-1} + \log^2\underline{T}^{-1})), \|W\|_\infty = \mathcal{O}(\log^3\varepsilon^{-1} + \log^3\underline{T}^{-1}), S = \mathcal{O}(\log^4\varepsilon^{-1} + \log^4\underline{T}^{-1}), \\ &\text{and } B = \exp(\mathcal{O}(\log^2\varepsilon^{-1} + \log^2\underline{T}^{-1})).\end{aligned}$$

那么，对于 $x \in B_{t,\ \varepsilon}$ 且满足 $t \geq \underline{T}$，我们有

$$\|\nabla\log p_t^{(2)}(x) - \phi_{\mathrm{score},4}\| \lesssim \varepsilon.$$

由此可得

$$\int_{B_{t,\varepsilon}} p_t(x)\|\nabla\log p_t^{(2)}(x) - \phi_{\mathrm{score},4}\|\mathrm{d}x \lesssim \varepsilon.$$

结合引理 E.2，取 $\varepsilon = \mathrm{poly}(N^{-1})$，即可得到该断言。 □

定理 6.4 的证明。注意，虽然引理 E.3 中的误差界 (101) 比引理 E.1 中的界 (98) 和 (99) 更紧，但引理 E.3 所需的网络规模 (102) 小于引理 E.1 中的规模界。还要注意，除将 d 替换为 d' 外，引理 E.1 中的界与定理 3.1 和引理 3.6 中的界相同。因此，只需汇总 ϕ_{score}、3 和 ϕ_{score}、4，即可得到近似定理定理 3.1 和引理 3.6 的对应结果，其余分析与 d 维情形相同。因此，我们得到所述结论。

□

F. 辅助引理

本节最后总结已有结果，并为证明的主要部分准备基本工具。本节很大一部分（附录 F.1 至 F.4）旨在介绍使用神经网络进行函数逼近的基本工具，因此，熟悉这些主题的读者（Yarotsky，2017年；Petersen & Voigtlaender，2018；Schmidt-Hieber，2019年）可以跳过这些小节（尽管其中包含一些细化和扩展）。引理 F.14 给出了关于高斯分布和布朗运动命中时刻的基本界。

在下文中，我们将定义常数 $C_{\mathrm{f},1}$ 和 $C_{\mathrm{f},2}$。除本节外，它们记作 C_{f}；有时，本节中产生的其他常数也可以记作 C_{f}。

F.1. 更大神经网络的构造

在构造所需的神经网络时，我们经常需要组合能够近似更简单函数的子网络，以实现更复杂的函数。我们准备了以下引理，其直接来源是 Nakada& Imaizumi (2020)，但类似的思想可以追溯到更早的文献，例如 Yarotsky (2017年)；Petersen & Voigtlaender(2018)。

首先，我们考虑复合函数的构造。尽管原始版本中未给出稀疏度 S 的界，但通过仔细检查其证明，我们可以验证这一点。

引理 F.1（神经网络的串接（Nakada & Imaizumi（2020）的备注 13））。对于任意神经网络$\phi^1: \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$、$\phi^2: \mathbb{R}^{d_2} \to \mathbb{R}^{d_3}$、$\cdots$、$\phi^k: \mathbb{R}^{d_k} \to \mathbb{R}^{d_{k+1}}$，满足 $\phi^i \in \Psi(L^i, W^i, S^i, B^i)$ $(i = 1, 2, \cdots, d)$，存在一个神经网络 $\phi \in \Phi(L, W, S, B)$，使得对于所有 $x \in \mathbb{R}^{d_1}$，$\phi(x) = \phi^k \circ \phi^{k-1} \cdots \circ \phi^1(x)$，其中

$$L = \sum_{i=1}^{k} L^i, \quad W \le 2\sum_{i=1}^{k} W^i, \quad S \le \sum_{i=1}^{k} S^i + \sum_{i=1}^{k-1}(\|A^i_{L^i}\|_0 + \|b^i_{L^i}\|_0 + \|A^{i+1}_1\|_0) \le 2\sum_{i=1}^{k} S^i, \quad \text{and } B \le \max_{1\le i\le k} B^i.$$

这里，A^i_j 是第 j 层第 i 个神经网络 ϕ^i 的参数矩阵，b^i_j 是偏置向量。

接下来我们介绍恒等函数。

引理 F.2（恒等函数（Nakada & Imaizumi （2020）第 19 页））。对于 $L \ge 2$ 和 $d \in \mathbb{N}$，存在一个神经网络 $\phi^{d,\,L\,\mathrm{Id}} \in \Phi(L, W, S, B)$，其参数为 $(A_1, b_1) = ((I_d, -I_d)^\top, 0)$，$(A_i, b_i) = (I_{2d}, 0)(i = 1, 2, \cdots, L-2)$，$(A_L) = ((I_d, -I_d), 0)$，能够实现 d 维恒等映射。这里，

$$\|W\|_\infty = 2d, \quad S = 2dL, \quad B = 1.$$

对于 $L = 1$，具有参数 (A_1, b_1) 的神经网络 $\phi^{d,\,1\,\mathrm{Id}} \in \Phi(1, (d)\ d, 1) = (I_d, 0)$ 实现了 d 维恒等映射。

接下来我们考虑神经网络并行化。以下引理是 Nakada & Imaizumi（2020）中的注释 14 和 15，并进行了修改，以允许子网络具有不同的深度。

引理 F.3（神经网络并行化）。对于任意具有 $\cdot\,\phi^i$ 的神经网络 $\phi^1, \phi^2, \cdots, \phi^k$：$\mathbb{R}^{d_i} \to \mathbb{R}^{d'_i}$ 以及$\phi^i \in \Psi(L^i, W^i, S^i, B^i)$ $(= i = 1, 2, [\phi^1\ x^1\ \top\ \phi^2\ x^2\ \top\ \cdots\ \phi^k\ x^k\ \top]^\top \mathbb{R}^{d_1+d_2+\cdots+d_k} \to \mathbb{R}^{d'_1+d'_2+\cdots+d'_k} \quad x = x_1^\top\ x_2^\top\ \cdots\ x_k^\top\ \top \in \mathbb{R}^{d_1+d_2+\cdots+d_k}, d)$，

存在一个满足 $\phi(\,)(\,)(\,)(\,)$ 的神经网络 $\phi \in \Phi(L, W, S, B)$：对于所有 $(\,)$（此处可以共享 x_i），其中

$$L = L, \quad \|W\|_\infty \le \sum_{i=1}^{k} \|W^i\|_\infty, \quad S \le \sum_{i=1}^{k} S^i, \quad \text{and } B \le \max_{1\le i\le k} B^i \quad \textit{(when } L = L_i \textit{ holds for all } i\textit{)},$$

$$L = \max_{1\le i\le k} L^i, \quad \|W\|_\infty \le 2\sum_{i=1}^{k} \|W^i\|_\infty, \quad S \le 2\sum_{i=1}^{k}(S^i + LW^i_L), \quad \text{and } B \le \max\{\max_{1\le i\le k} B^i, 1\} \quad \textit{(otherwise)}.$$

此外，存在一个网络 $\phi_{\mathrm{sum}}(x) \in \Phi(L, W, S, B)$，它实现 $= \sum_{i=1}^{k} \phi^i(x)$，其中

$$L = \max_{1\le i\le k} L^i + 1, \quad \|W\|_\infty \le 4\sum_{i=1}^{k} \|W^i\|_\infty, \quad S \le 4\sum_{i=1}^{k}(S^i + LW_L) + 2W_L, \quad \text{and } B \le \max\{\max_{1\le i\le k} B^i, 1\}.$$

引理 F.3 的证明。我们先考虑第一部分。对于 $L = L_i$ 对所有 i都成立的情况，这些断言与 Nakada & Imaizumi（2020）的注释 14 和 15 完全相同。否则，我们先构造一个网络 ϕ'^i，使其对所有 i实现 $\phi^d,_{L-L_i}\mathrm{Id} \circ \phi^i$，从而使每个网络具有相同的深度，同时不改变各网络的输出。根据引理 F.1 和 F.2，
$\phi'^i \in \Phi\ L\ W'^i\ S'^i\ B'^i \qquad L = \quad _{1\le i\le k} L^i\ \|W'^i\|_\infty = \quad \{\|W^i\|_\infty, 2W_L\} \le 2\|W^i\|_\infty\ S'^i \le 2S^i + (,,,)$ 成立，其中 $\max , \max , 2(L-L_i)W_L^i \le 2(S^i + LW_L^i)$，并且 $B'^i = \max\{B^i, 1\}$。然后，我们应用 $L = L_i (i = 1, 2, \cdots, k)$ 情况下的结果。

对于第二部分，由于 k 神经网络输出的求和可以由一个宽度为 1 的层神经网络实现，因此 k，引理 F.3 与引理 F.1 共同给出了实现 $\sum_{i=1}^k \phi^i(x)$的界。 □

在对基于分数的扩散模型进行分析时，我们经常会遇到无界函数。为了解决由无界性带来的困难，通常会采用截断操作。

引理 F.4（截断函数）。对于任意 $a, b \in \mathbb{R}^d$，若满足 $a_i \le b_i (i = 1, 2, \cdots, d)$，则存在一个截断函数$\phi_{\mathrm{clip}}(x; a, b) \in \Phi(2, (d, 2d, d)^\top, 7d, \max^{1\le i\le d} \max\{|a_i|, b_i\})$，使得

$$\phi_{\mathrm{clip}}(x; a, b)_i = \min\{b_i, \max\{x_i, a_i\}\} \quad (i = 1, 2, \cdots, d)$$

成立。当对所有 i 均有 $a_i = c$和 $b_i = C$时，我们有时使用标量值 c和 C，将 $\phi_{\mathrm{clip}}(x; a, b)$ 记作 $\phi_{\mathrm{clip}}(x; c, C)$。

证明。由于对每个坐标 i，$\min\{b_i, \max\{x_i, a_i\}\}$ 可表示为

$$\min\{b_i, \max\{x_i, a_i\}\} = \mathrm{ReLU}(x_i - a_i) - \mathrm{ReLU}(x_i - b_i) + a_i \in \Phi(2, (1, 2, 1), 7, \max\{|a_i|, b_i\}),$$

对所有 i 应用引理F.3 进行并行化，即可得到该断言。 □

利用上述截断函数，我们构造切换函数，从而得到一种方法：当两个不同的近似分别在不同区域内有效时，可在这些区域的交叠部分构造近似。

引理 F.5（切换函数）。设 $\underline{t}_1 < \underline{t}_2 < \overline{t}_1 < \overline{t}_2$，且 $f(x, t)$ 是某个标量值函数（对于向量值函数，只需逐坐标应用）。假设 $\phi^1(x, t)$ 和 $\phi^2(x, t)$ 对 $f(x, t)$ 的近似误差至多为 ϵ，但利用 $\phi^1(x, t)$ 和 $\phi^2(x, t)$ 的近似分别在 $[\underline{t}_1, \overline{t}_1]$ 和 $[\underline{t}_2, \overline{t}_2]$ 上有效。那么，存在神经网络 $\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1)$ $\phi^2_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1) \in \Phi(3, (1, 2, 1, 1)^\top, 8, \max\{\overline{t}_1, (\overline{t}_1 - \underline{t}_2)^{-1}\})$，以及 $\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1)\phi^1(x, t) + \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1)\phi^2(x, t)$，使其在 $[\underline{t}_1, \overline{t}_2]$ 中对 $f(x, t)$ 的近似误差至多为 ϵ。

证明。我们定义

$$\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1) = \frac{1}{\overline{t}_1 - \underline{t}_2}\mathrm{ReLU}(\phi_{\mathrm{clip}}(t; \underline{t}_2, \overline{t}_1) - \underline{t}_2), \quad \text{and } \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1) = \frac{1}{\overline{t}_1 - \underline{t}_2}\mathrm{ReLU}(\overline{t}_1 - \phi_{\mathrm{clip}}(t; \underline{t}_2, \overline{t}_1)).$$

这里，对于所有 t，$\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1) = 0$；对于所有 $t \ge \overline{t}_1$，以及 $\phi^2_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1)$，当 $t \le \underline{t}_2$ 时，$\phi^1_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1)$ $\phi^2_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1) \in [0, 1], \phi^1_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1) + \phi^2_{\mathrm{swit}}(t; \underline{t}_2, \overline{t}_1)$。由此构造可得出断言。 □

F.2. 逼近有理函数的基本神经网络结构

使用神经网络逼近 Besov空间中的函数时，网络最基本的结构是用于逼近多项式的结构（Suzuki，2018年）。在我们构造扩散 B 样条基时，需要逼近有理函数。

我们从单项式开始。尽管使用神经网络以任意加性误差 ϵ 仅使用 $\mathcal{O}(\log \varepsilon^{-1})$ 个非零参数即可逼近单项式这一传统结论已广为人知（Yarotsky，2017年；Petersen &Voigtlaender，2018；Schmidt-Hieber，2020），但我们没有找到明确说明参数依赖关系的结果，其中包括输入的次数和范围。因此，为确保严谨，我们重新考察了 Schmidt-Hieber(2020)中的引理 A.3，并在此给出该引理的扩展版本。

引理 F.6（单项式的近似）。设 $d \ge 2, C \ge 1, 0 < \varepsilon_{\mathrm{error}} \le 1$。对于任意 $\varepsilon > 0$，存在一个神经网络 $\phi_{\mathrm{mult}}(x_1, x_2, \cdots, x_d) \in \Psi(L, W, S, B)$，其 $L = \mathcal{O}(\log d(\log \varepsilon^{-1} + d \log C))$ $\|W\|_\infty = 48d, S = \mathcal{O}(d \log \varepsilon^{-1} + d \log C))$ $B = C^d$，使得

$$\left|\phi_{\mathrm{mult}}(x'_1, x'_2, \cdots, x'_d) - \prod_{d'=1}^d x_{d'}\right| \le \varepsilon + dC^{d-1}\varepsilon_{\mathrm{error}}, \quad \textit{for all } x \in [-C, C]^d \textit{ and } x' \in \mathbb{R} \textit{ with } \|x - x'\|_\infty \le \varepsilon_{\mathrm{error}},$$

$|\phi_{\text{mult}}(x)| \le C^d$ 对所有 $x \in \mathbb{R}^d$ 都成立，并且如果 x'_i 中至少有一个为 0，则 $\phi_{\text{mult}}(x'_1, x'_2, \cdots, x'_d) = 0$。

我们注意到，x_i、x_j（$i \neq j$）中的某些部分可以共享。对于 $\prod_{i=1}^I x_i^{\alpha_i}$ 与 $\alpha_i \in \mathbb{Z}_+$（$i = 1$、2、$\cdots$、$I$）以及 $\sum_{i=1}^I \alpha_i = d$，存在一个满足上述相同界限的神经网络，该网络记为 $\phi_{\text{mult}}(x;\ \alpha)$。

证明。首先，根据 Schmidt-Hieber (2020) 的研究可知，存在一个神经网络 $\phi'_{\text{mult}}(x,\ y) \in \Psi(L,\ W,\ S,\ B)$ 满足 $L = i + 5$，$\|W\|_\infty = 6$，$B = 1$ 使得

$$|\bar{\phi}'_{\text{mult}}(x, y) - xy| \le 2^{-i}, \quad \text{for all } (x, y) \in [0, 1]^2,$$

并且对于所有 $(x,\ y) \in \mathbb{R}^2$，$|\bar{\phi}'_{\text{mult}}(x,\ y)| \le 1$，以及当 x 或 y 中任一个为 0 时，$\bar{\phi}'_{\text{mult}}(x,\ y) = 0$。利用该网络，我们可以看到，对于所有 $(x,\ y) \in [-1, 1]^2$，$|\text{sign}(xy)\bar{\phi}'_{\text{mult}}(|x|,\ |y|) - xy| \le 2^{-i}$ 成立；对于所有 $(x,\ y) \in \mathbb{R}^2$，$|\bar{\phi}'_{\text{mult}}(x,\ y)| \le 1$；并且当 x 或 y 中任一个为 0 时，$\phi_{\text{mult}}(x,\ y) = 0$。因为

$$\begin{aligned}
\text{sign}(xy)\bar{\phi}'_{\text{mult}}(|x|, |y|) = {} & \text{ReLU}(\bar{\phi}'_{\text{mult}}(\text{ReLU}(x), \text{ReLU}(y)) + \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x), \text{ReLU}(-y)) \\
& - \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x), \text{ReLU}(y)) - \bar{\phi}'_{\text{mult}}(\text{ReLU}(x), \text{ReLU}(-y))) \\
& - \text{ReLU}(-\bar{\phi}'_{\text{mult}}(\text{ReLU}(x), \text{ReLU}(y)) - \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x), \text{ReLU}(-y)) \\
& + \bar{\phi}'_{\text{mult}}(\text{ReLU}(-x), \text{ReLU}(y)) + \bar{\phi}'_{\text{mult}}(\text{ReLU}(x), \text{ReLU}(-y))) \\
=: {} & \bar{\phi}_{\text{mult}}(x, y)
\end{aligned}$$

成立时，我们可以通过神经网络 $\bar{\phi}_{\text{mult}}(x, y) \in \Psi(L, W, S, B)$，在 $L = i + 7, \|W\|_\infty = 48, S \le L\|W\|_\infty(\|W\|_\infty + 1) = 48(i + 7)$ $B = 1$ 的条件下，实现函数 xy for $[-1, 1]^d$，其近似误差最大为 2^{-i}。

接下来，按照 Schmidt-Hieber (2020，我们递归构造 $\phi_{\text{mult}}(x_1, x_2, \cdots, x_{2^{j+1}})$，使用

$$\bar{\phi}_{\text{mult}}(x_1, x_2, \cdots, x_{2^{j+1}}) = \bar{\phi}_{\text{mult}}(\bar{\phi}_{\text{mult}}(x_1, x_2, \cdots, x_{2^j}), \bar{\phi}_{\text{mult}}(x_{2^j+1}, x_{2^j+2}, \cdots, x_{2^{j+1}})).$$

将 $(x_1, x_2, \cdots, x_{2^j})$ 的额外维度填为 1，可得神经网络 $\phi_{\text{mult}}(x_1, x_2, \cdots, x_d) \in \Psi(L, W, S, B)$，对于所有 $d \ge 2$ and $L = \mathcal{O}(\log d(\log \varepsilon^{-1} + \log d))$ $\|W\|_\infty = 48d, S = \mathcal{O}(d(\log \varepsilon^{-1} + \log d))$ $B = 1$，均满足

$$\left|\bar{\phi}_{\text{mult}}(x_1, x_2, \cdots, x_d) - \prod_{d'=1}^d x_{d'}\right| \le \varepsilon, \quad \text{for all } x \in [-1, 1]^d.$$

然后我们按如下方式构造 ϕ_{mult}：

$$\phi_{\text{mult}}(x) = C^d \bar{\phi}_{\text{mult}}(\phi_{\text{clip}}(x; -C, C)/C).$$

这里，$[-C, C]^d$ 上的近似误差有界于 $C^{-d}\varepsilon$。我们重新设置 $\varepsilon \leftarrow C^{-d}\varepsilon$，使近似误差小于 ε，于是有 $\phi_{\text{mult}} \in \Phi(L, W, S, B)$，其中 $L = \mathcal{O}(\log d(\log d + \log \varepsilon^{-1} + d \log C))$ $\|W\|_\infty = 48d, S = \mathcal{O}(d(\log d + \log \varepsilon^{-1} + d \log C))$ $B = 1$。因此，断言中关于 L、$\|W\|_\infty$、B、S 的界由引理 F.1 和 F.4 得出。

当输入发生波动时，我们有

$$\begin{aligned}
& \left|C^d \bar{\phi}_{\text{mult}}(\phi_{\text{clip}}(x'; -C, C)/C) - \prod_{i=1}^d x_i\right| \\
\le & \left|C^d \bar{\phi}_{\text{mult}}(\phi_{\text{clip}}(x'; -C, C)/C) - \prod_{i=1}^d \min\{C, \max\{x'_i, -C\}\}\right| + \left|\prod_{i=1}^d \min\{C, \max\{x'_i, -C\}\} - \prod_{i=1}^d x_i\right| \\
\le & C^d \cdot C^{-d}\varepsilon + C^{d-1} \sum_{i=1}^d |x_i - \min\{C, \max\{x'_i, -C\}\}| = \varepsilon + dC^{d-1}\varepsilon_{\text{error}},
\end{aligned}$$

由此得到断言的第一部分。

最后，我们注意到，由于我们只需识别 x_i, x_j $(i \neq j)$ 的第一层中对应于同一坐标的列，因此其中一些可以共享。

□

接下来，我们说明如何近似倒数函数 $y = \frac{1}{x}$。有理函数的近似已在（Telgarsky，2017年；Boull´e 等人，2020）中研究过。然而，我们发现，通过谨慎使用局部泰勒展开，其（Telgarsky（2017年）中的引理 3.5）关于 $L = \mathcal{O}(\log^7 \varepsilon^{-1})$ 和 $\mathcal{O}(\log^4 \varepsilon^{-1})$ 个节点的界可以改进至 $\mathcal{O}(\log \varepsilon^{-1})$ 阶，因此我们给出自己的证明。

引理 F.7（近似倒数函数）。对于任意 $0 < \varepsilon < 1$，存在 $\phi_{\text{rec}} \in \Psi(L, W, S, B)$，满足$L \le \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1})$ $S = \mathcal{O}(\log^4 \varepsilon^{-1})$，以及 $B = \mathcal{O}(\varepsilon^{-2})$，使得

$$\left|\phi_{\text{rec}}(x') - \frac{1}{x}\right| \le \varepsilon + \frac{|x' - x|}{\varepsilon^2}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

证明。我们用分段多项式函数近似反函数 $y = \frac{1}{x}$ $x_i = 1.5^i \cdot \varepsilon$ $i =$。我们取 $0\ 1\ \cdots\ i^* = \lceil 2 \quad {}_{1.5}\varepsilon^{-1} \rceil$ $x_{i^*} \ge \varepsilon^{-1}$ $y = \frac{1}{x}$ (, , , : log)，使得 ，并按以下方式进行近似:

$$\frac{1}{x} \coloneqq \sum_{i=1}^{i^*} f_i(\phi_{\text{clip}}(x; x_{i-1}, x_i)) + \frac{1}{\varepsilon},$$

where $f_i(x)$ 是满足以下条件的函数： $f_i(x) = 0$ 当 $x \le x_{i-1}$， $f_i(x) = -\frac{1}{x_{i-1}} + \frac{1}{x_i}$ 当 $x_i \le x$，且d

$$\max_{x_{i-1} \le x \le x_i} |f_i(x) - 1/x + 1/x_{i-1}| \le \frac{\varepsilon}{2}.$$

现在我们给出这类函数的构造。首先，根据 $\frac{1}{x} = \frac{1}{x_{i-1}} \frac{x_{i-1}}{x} = \frac{1}{x_{i-1}} \sum_{l'=1}^{\infty} (-\frac{x}{x_{i-1}} + 1)^{l'}$ $(1 \le \frac{x}{x_{i-1}} \le 1.5)$，令

$$\tilde{f}_i(x) = \frac{1}{x_{i-1}} \sum_{l'=1}^{l} (-x/x_{i-1} + 1)^{l'} - \frac{1}{x_{i-1}}.$$

$\tilde{f}_i(x)$ 与 $\frac{1}{x} - \frac{1}{x_{i-1}}$ 之间的差为 $((x_{i-1} - x)/x_{i-1})^{l+1}/x$，其上界为 $2^{-l-1}/x$。此外，通过向 $\tilde{f}_i(x)$ 添加$(\frac{1}{x_i} - \tilde{f}_i(x_i))(x - x_{i-1})$ $x_i - x_{i-1}$ $= ((x_{i-1} - x_i)/x_{i-1})^{l+1}(x - x_{i-1})\, x_i(x_i - x_{i-1})$，我们得到 $f_i(x)$，其中 $f_i(x_{i-1}) = 0, f_i(x_i) = -\frac{1}{x_{i-1}} + \frac{1}{x_i}$，并且

$$\max_{x_{i-1} \le x \le x_i} |f_i(x) - 1/x + 1/x_{i-1}| \le 2^{-l}/x \le 2^{-l}\varepsilon^{-1}.$$

因此，我们取 $l = \lceil \log_2 2\varepsilon^{-1} \rceil$，使得右端小于$\frac{\varepsilon}{2}$。因此，我们最终得到 $y = \frac{1}{x}$ 的显式近似:

$$f(x) = \underbrace{\sum_{i=1}^{i^*} \frac{1}{x_{i-1}} \sum_{l'=1}^{l} (-\phi_{\text{clip}}(x; x_{i-1}, x_i))/x_{i-1} + 1)^{l'}}_{\text{(a)}} - \sum_{i=1}^{i^*} \frac{1}{x_{i-1}} \tag{103}$$
$$+ \underbrace{\sum_{i=1}^{i^*} \frac{((x_{i-1} - x_i)/x_{i-1})^{l+1}(\phi_{\text{clip}}(x; x_{i-1}, x_i)) - x_{i-1})}{x_i(x_i - x_{i-1})}}_{\text{(b)}} + \frac{1}{\varepsilon}.$$

根据引理 F.6，$(-\phi_{\text{clip}}(x;\ x_{i-1}, x_i))/x_{i-1} + 1)^{l'}$ 可由 $L = \mathcal{O}((\log\log\varepsilon^{-1} + \log\varepsilon^{-1})\log\log\varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log\varepsilon^{-1})$ $S = \mathcal{O}(\log\varepsilon^{-1}(\log\log\varepsilon^{-1} + \log\varepsilon^{-1}))$ $B = 1.5^{\lceil \log_2 2\varepsilon^{-1} \rceil} = \mathcal{O}(\varepsilon^{-1})$ 实现，使得每一项的近似误差均不超过 $\mathcal{O}(\varepsilon^2/li^*)$。由于 (103) 的 (a) 中有 $\mathcal{O}(li^*)$ 项，根据引理 F.1 和 F.3，使用神经网络 ϕ_{rec} 对 $f(x)$ 进行近似的最终近似误差为$\frac{\varepsilon}{2}$，其中 $\phi_{\text{rec}} \in \Phi(L, W, S, B)$，且 $L \le \mathcal{O}((\log\log\varepsilon^{-1} + \log\varepsilon^{-1})\log\log\varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log^3\varepsilon^{-1})$ $S = \mathcal{O}(\log^3\varepsilon^{-1}(\log\log\varepsilon^{-1} + \log\varepsilon^{-1}))$，并且 $B = \mathcal{O}(\varepsilon^{-2})$。（这里计算 $B = \mathcal{O}(\varepsilon^{-2})$ 是因为在 (b) 中我们需要将系数 $((x_{i-1} - x_i)$ $/x_{i-1})^{l+1}$) ( ) 控制在 ε^{-2} 以内。）

敏感性分析由 $|\phi_{\text{rec}}(x') - \frac{1}{x}| \le |\phi_{\text{rec}}(x') - {}^{1}_{\text{最大值}\{x',\varepsilon\}}| + |\frac{1}{\text{最大值}\{x',\varepsilon\}} - \frac{1}{x}|$ 得出。 □

结合引理 F.6 和 F.7，我们得到以下推论。

推论 F.8. 对于任意 $0 < \varepsilon < 1$，存在 $\phi_{\mathrm{rec}} \in \Psi(L, W, S, B)$，满足 $L \le \mathcal{O}(\log^2 l + \log^2 \varepsilon)$ $\|W\|_\infty = \mathcal{O}(l + \log^3 \varepsilon^{-1})$ $S = \mathcal{O}(l \log l + l \log \varepsilon^{-1} + \log^4 \varepsilon^{-1})$，以及 $B = \mathcal{O}(\varepsilon^{-(2 \vee l)})$，使得

$$\left| \phi_{\mathrm{rec}}(x'; l) - \frac{1}{x^l} \right| \le \varepsilon + l \frac{|x' - x|}{\varepsilon^{l+1}}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

证明。 考虑 $\phi_{\mathrm{mult}}(\cdot;\ l) \circ \phi_{\mathrm{rec}}$。该结果直接由引理 F.6 和引理 F.7得出。 □

同理，通过对上述证明中定义的每个区间内的 $\sqrt{1+x}$ 进行泰勒展开，我们可以得到关于 $y = \sqrt{x}$ 的类似结果。

引理 F.9（逼近根函数）。对于任意 $0 < \varepsilon < 1$，存在 $\phi_{\mathrm{root}} \in \Psi(L, W, S, B)$，满足 $L \le \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1})$ $S = \mathcal{O}(\log^4 \varepsilon^{-1})$，以及 $B = \mathcal{O}(\varepsilon^{-1})$，使得

$$\left| \phi_{\mathrm{root}}(x') - \sqrt{x} \right| \le \varepsilon + \frac{|x' - x|}{\sqrt{\varepsilon}}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

F.3. 如何处理指数函数

有时我们需要近似某些积分，其被积函数包含某个高斯分布的密度函数，且积分区间为 $\mathbb{R}^d$。例如，扩散 B 样条基就是这类积分的典型例子。为处理这类积分，我们采用以下两步论证：首先截断积分区间，然后用有理函数近似被积函数。我们需要使用有理函数，因为密度函数取决于方差平方根的倒数，而方差又取决于 t，因此需要对其进行近似。第一个引理对应第一步，第二个和第三个引理则分别对应第二步。

引理 F.10（积分截断）。设 $x \in \mathbb{R}^d$、$0 < m_t \le 1$、$\alpha \in \mathbb{Z}_+^d$ 满足 $\sum_{i=1}^d \alpha_i \le k$，且 f 是定义在 $\mathbb{R}^d$ 上的任意函数，其绝对值不超过 C_f。对于任意 $0 < \varepsilon < \frac{1}{2}$，存在一个常数 $C_{\mathrm{f},\,1}$，它仅取决于 k 和 d，使得

$$\left| \int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t} \right)^{\alpha_i} f(y) \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2} \right) dy \right.$$
$$\left. - \int_{A^x} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t} \right)^{\alpha_i} f(y) \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2} \right) dy \right| \lesssim \varepsilon,$$

其中 $A^x = \prod_{i=1}^d a_i^x$ 满足 $a_i^x = [\frac{x_i}{m_t} - \frac{\sigma_t C_{\mathrm{f},\,1}}{m_t} \sqrt{\text{对数} \overline{\varepsilon^{-1} \frac{x_i}{m_t}}} + \frac{\sigma_t C_{\mathrm{f},}}{m_t} \frac{1}{\sqrt{\text{对数} \overline{\varepsilon^{-1}}}}]$。

证明。

$$\frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \left| \int_{\mathbb{R}^d} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t} \right)^{\alpha_i} f(y) \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2} \right) dy - \int_{A^x} \prod_{i=1}^d \left(\frac{m_t y_i - x_i}{\sigma_t} \right)^{\alpha_i} f(y) \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2} \right) dy \right|$$
$$\le \frac{C_f}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d \setminus A^x} \prod_{i=1}^d \left(\frac{|m_t y_i - x_i|}{\sigma_t} \right)^{\alpha_i} \mathbb{I}[\|y\|_\infty \le 1] \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2} \right) dy \quad (\text{by } |f(y)| \le C_f)$$
$$\le \frac{C_f}{\sigma^d (2\pi)^{\frac{d}{2}}} \sum_{i=1}^d \int_{\underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{i-1 \text{ times}} \times (\mathbb{R} \setminus a_i^x) \times \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{d-i \text{ times}}} \prod_{j=1}^d \left(\frac{|m_t y_j - x_j|}{\sigma_t} \right)^{\alpha_j} \mathbb{I}[|y_j| \le 1] \exp\left(-\frac{\|m_t y - x\|^2}{2\sigma_t^2} \right) dy$$
$$= C_f \sum_{i=1}^d \prod_{j=1}^d \left(\mathbb{I}[i \ne j] \int_{\mathbb{R}} \left(\frac{|m_t y_j - x_j|}{\sigma_t} \right)^{\alpha_j} \frac{\mathbb{I}[|y_j| \le 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2} \right) dy_j \right.$$
$$\left. + \mathbb{I}[i = j] \int_{\mathbb{R} \setminus a_i^x} \left(\frac{|m_t y_j - x_j|}{\sigma_t} \right)^{\alpha_j} \frac{\mathbb{I}[|y_j| \le 1]}{\sigma_t (2\pi)^{\frac{1}{2}}} \exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2} \right) dy_j \right). \quad (104)$$

现在我们对每一项进行界定。首先，

$$\int_{\mathbb{R}}\left(\frac{|m_t y_j - x_j|}{\sigma_t}\right)^{\alpha_j}\frac{\mathbb{I}[|y_j|\le 1]}{\sigma_t(2\pi)^{\frac{1}{2}}}\exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2}\right)dy_j \le \begin{cases}\frac{1}{m_t}\int_{\mathbb{R}}|y_j'|^{\alpha_j}\frac{1}{(2\pi)^{\frac{1}{2}}}\exp\left(-\frac{y_j'^2}{2}\right)dy_j' & \left(\frac{m_t y_j - x_j}{\sigma_t} = y_j'\right)\\ \frac{2^{d+\alpha_j}}{\sigma_t^{\alpha_j+1}(2\pi)^{\frac{1}{2}}} & (\text{because of the term of } \mathbb{I}[|y_j|\le 1].)\end{cases}$$

因此，左端可由 $.\max\left\{\frac{1}{m_t}, \frac{1}{\sigma_t^{\alpha_j+1}}\right\} \lesssim 1.$ 界定。

接下来，

$$\begin{aligned}&\int_{\mathbb{R}\setminus a_i^x}\left(\frac{|m_t y_j - x_j|}{\sigma_t}\right)^{\alpha_j}\frac{\mathbb{I}[|y_j|\le 1]}{\sigma_t(2\pi)^{\frac{1}{2}}}\exp\left(-\frac{(m_t y_j - x_j)^2}{2\sigma_t^2}\right)dy_j \qquad (105)\\ &\le \frac{2}{m_t}\int_{C_{\mathrm{f},1}\sqrt{\log\varepsilon^{-1}}}^{\infty}|y_j|^{\alpha_j}\exp\left(-\frac{y_j^2}{2}\right)\mathrm{d}y_i \quad \left(\text{by letting } \frac{m_t y_j - x_j}{\sigma_t}\mapsto y_j\right)\\ &\le \begin{cases}\frac{2}{m_t}\sum_{l=0}^{\frac{\alpha_j-1}{2}}\frac{(\alpha_j-1)!!}{(2l)!!}(C_{\mathrm{f},1}^2\log\varepsilon^{-1})^l\varepsilon^{\frac{C_{\mathrm{f},1}}{2}} & (\text{ if } \alpha_j \text{ is odd})\\ \frac{2}{m_t}\sum_{l=1}^{\frac{\alpha_j}{2}}\frac{(\alpha_j-1)!!}{(2l-1)!!}(C_{\mathrm{f},1}^2\log\varepsilon^{-1})^l\varepsilon^{\frac{C_{\mathrm{f},1}}{2}} + \frac{2}{m_t}\int_{C_{\mathrm{f},1}\sqrt{\log\varepsilon^{-1}}}^{\infty}\exp\left(-\frac{y_j^2}{2}\right)\mathrm{d}y_j & (\text{ if } \alpha_j \text{ is even}).\end{cases}\end{aligned}$$

因此，通过设定 $C_{\mathrm{f},\ 1}$ 足够大，使得 $C_{\mathrm{f},\ 1}$ 取决于 $\alpha_j(\le k)$ 和 d，可将其界定为 $\frac{\varepsilon}{m_t}$。此外，若 $m_t \gtrsim 1$，则积分区间与 $-1 \le y_j \le 1$不重叠，此时（105）也可由 0界定。

因此，（104）可进一步界定为

$$(104) \lesssim \sum_{i=1}^{d}\prod_{j=1}^{d} 1^{d-1}\cdot\varepsilon \lesssim \varepsilon,$$

由此得出该断言。 □

接下来，我们分别介绍用多项式（引理 F.11）和神经网络（引理 F.12）对指数函数进行泰勒展开的方法。

引理 F.11（用多项式逼近指数函数）。设 $A > 0$ 和 $0 \le m_t \le 1$。对于$t \ge \max\{4eA^2, \lceil\log_2\varepsilon^{-1}\rceil\}$，有

$$\left|\exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right) - \sum_{s=0}^{t-1}\frac{(-1)^s}{s!}\frac{(x-m_t y)^{2s}}{2^s\sigma_t^{2s}}\right| \le \varepsilon$$

对所有 $y \in [\frac{-\sigma_t A + x}{m_t}, \frac{\sigma_t A + x}{m_t}]$均成立。

证明。根据对 e^z 进行的次数不超过 $t-1$的标准泰勒展开，有

$$\exp\left(-\frac{(x-m_t y)^2}{2\sigma_t^2}\right) = \sum_{s=0}^{t-1}\frac{(-1)^s}{s!}\frac{(x-m_t y)^{2s}}{2^s\sigma_t^{2s}} + \frac{(-1)^t}{t!}\frac{(\theta(x-m_t y))^{2t}}{2^t\sigma_t^{2t}}$$

其中某个 $\theta \in (0,1)$。我们对余项的第二项进行估计。当 $y \in [\frac{-\sigma_t A + x}{m_t}, \frac{\sigma_t A + x}{m_t}]$ 且 t 是满足 $t \ge \max\{4eA^2, \lceil\log_2\varepsilon^{-1}\rceil\}$的最小整数时，有

$$\frac{1}{t!}\frac{(\theta(x-m_t y) + (1-\theta)x)^{2t}}{2^t\sigma_t^{2t}} \le \frac{(2\sigma_t A)^{2t}}{t!2^t\sigma_t^{2t}} \le \frac{(2\sigma_t A)^{2t}}{(t/e)^t\cdot 2^t\sigma_t^{2t}} \le \frac{2^t A^{2t}}{(4A^2)^t} \le \frac{1}{2^t} \le \varepsilon,$$

其中我们使用了以下事实：$t! \ge (t/e)^t$。 □

引理 F.12（用神经网络逼近指数函数）。任意取 $\varepsilon > 0$。存在一个神经网络 $\phi_{\exp} \in \Phi(L, W, S, B)$，使得

$$\sup_{x,x' \geq 0} \left| e^{-x'} - \phi_{\exp}(x) \right| \leq \varepsilon + |x - x'|$$

成立，其中 $L = \mathcal{O}(\log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(\log \varepsilon^{-1})$ $S = \mathcal{O}(\log^2 \varepsilon^{-1})$ $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$。此外，对于所有 $x \geq \log 3\varepsilon^{-1}$，$|\phi_{\exp}(x)| \leq \varepsilon$。

证明。令 $A = \log 3\varepsilon^{-1}$。根据泰勒展开，对于所有 x in $0 \leq x \leq A$，我们有

$$\left| e^{-x} - \sum_{i=0}^{k-1} \frac{(-1)^i}{i!} x^i \right| \leq \frac{A^k}{k!}.$$

此外，我们可以将右侧计算为 $\frac{A^k}{k}! \leq \left(\frac{eA}{k}\right)^k$，因此取 $k = \max\{2eA, \lceil \log_2 3\varepsilon^{-1} \rceil\}$ 后，可以用 $\frac{\varepsilon}{3}$ 对右侧进行界定。现在，我们使用引理 F.6，以 $d = \mathcal{O}(A + \log \varepsilon^{-1})$ $C = \mathcal{O}(A)$ $\varepsilon = \frac{\varepsilon}{3k}$ 为依据，近似每个 x^i，并使用引理 F.3 将它们聚合起来。由此得到一个具有 $L = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1})$ $\|W\|_\infty = \mathcal{O}(A + \log \varepsilon^{-1})$ $S = \mathcal{O}(A^2 + \log^2 \varepsilon^{-1})$ $B = \exp(\log A \cdot \mathcal{O}(A + \log \varepsilon^{-1}))$ 的神经网络。最后，在该神经网络之前添加两层 $\phi_{\text{clip}}(x; 0, A)$，将输入限制在 $x > 0$ 范围内。于是，我们得到一个神经网络 $\phi_{\exp}$，它对 e^{-x} 的近似在 $[0, A]$ 中的加性误差至多为 $\frac{2\varepsilon}{3}$。此外，对于 $x > A$，有 $|\phi_{\exp}(x) - e^{-x}| \leq |e^{-x} - e^{-A}| + |\phi_{\exp}(A) - e^{-A}| \leq \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon$。

敏感性分析由 $|\phi_{\exp}(x') - e^{-x}| \leq |\phi_{\exp}(\max\{x', 0\}) - e^{-x}| \leq |\phi_{\exp}(\max\{x', 0\}) - e^{-\max\{x', 0\}}| + |e^{-\max\{x', 0\}} - e^{-x}| \leq \varepsilon + |\max\{x', 0\} - x| \leq \varepsilon + |x' - x|$得出。 □

F.4. 近似的已有结果

我们的扩散 B 样条基分解（第 3 节和附录B）建立在 Besov空间的 B 样条基分解之上（DeVore& Popov，1988；Suzuki，2018）。以下事实可见于 Suzuki （2018）的引理 2（尽管原始版本采用 $\Omega = [0, 1]^d$，但我们只需将定义域划分为边长为 1 的立方体，即可轻松调整这一差异）。$|\alpha_{k\text{的大小}j}|$ 在 Suzuki （2018）的第 17 页中进行了评估。

引理 F.13（Besov空间的可逼近性（Suzuki （2018））。设 $C > 0$。在 $s > d(1/p - 1/r)_+$ 以及 $0 < s < \min\{l, l - 1 + 1/p\}$ 的条件下，其中 $l \in \mathbb{N}$ 为基数 B 样条基的阶数，对于任意 $f \in B^s_{p,q}([-C, C]^d)$，存在满足以下条件的 f_N

$$\|f - f_N\|_{L^r([-C,C]^d)} \lesssim C^s N^{-s/d} \|f\|_{B^s_{p,q}([-C,C]^d)}$$

对于 $N \gg 1$，其形式如下：

$$f_N(x) = \sum_{k=0}^{K} \sum_{j \in J(k)} \alpha_{k,j} M^d_{k,j}(x) + \sum_{k=K+1}^{K^*} \sum_{i=1}^{n_k} \alpha_{k,j_i} M^d_{k,j_i}(x) \quad \textit{with} \sum_{k=0}^{K} |J(k)| + \sum_{k=K+1}^{K^*} n_k = N,$$

其中 $J(k) = \{-C2^k - l, -C2^k - l + 1, \cdots C2^k - 1, C2^k\}$，$(j_i)_{i=1}^{n_k} \subseteq J(k)$，$K = \mathcal{O}(d^{-1} \log(N/C^d))$，$K^* = (\mathcal{O}(1) + \log(N/C^d))\nu^{-1} + K$，$n_k = \mathcal{O}((N/C^d)2^{-\nu(k-K)})$ $(k = K + 1, \cdots, K^*)$，当 $\delta = d(1/p - 1/r)_+$ 且 $\nu = (s - \delta)/(2\delta)$时。此外，$|\alpha_{k,\ j}| \lesssim N_{(\nu^{-1} + d^{-1})(d/p - s)_+}$。

F.5. 高斯分布与命中时刻的基本界

引理 F.14. 设 $0 < \varepsilon \ll 1$、$l \in \mathbb{Z}^d_+$ 和 $p(x)$ 为 $\mathcal{N}(0, \sigma_t^2 I_d)$ 的密度函数，即 $p(x) = {}_{\sigma_t^d}{}^{1}{}_{(2\pi)^{\frac{d}{2}}} \underline{\exp\left(-\frac{\|x\|^2}{\sigma_t^2}\right)}$。那么，以下界成立：

$$\int_{\|x\|_\infty \geq \sigma_t \sqrt{4 \log dl\varepsilon^{-1}}} \frac{\prod_{i=1}^d x_i^{l_i}}{\sigma^{\sum_{i=1}^d l_i}} p(x) \mathrm{d}x \lesssim \varepsilon.$$

我们有时将 $\sqrt{4\log dl\varepsilon^{-1} = C_{\mathrm{f},\ 2}}\sqrt{\log \varepsilon^{-1}}$ 写作。

证明。令 为便于简洁表示，记 $x^l = \prod_{i=1}^d x_i^{l_i}$和 $|l| = \sum_{i=1}^d l_i$。令 $r = \|x\|_\infty$，我们得到 t

$$
\begin{aligned}
&\int_{\|x\|_\infty \geq \sigma_t \sqrt{4\log \varepsilon^{-1}}} \frac{x^l}{\sigma_t^{|l|}} p(x)\mathrm{d}x \\
&\int_{\|x\|_1 \geq \sigma_t \sqrt{4\log \varepsilon^{-1}}} \frac{x^l}{\sigma_t^{|l|}} p(x)\mathrm{d}x \\
&\leq \int_{r=\sigma_t \sqrt{4\log \varepsilon^{-1}}}^{\infty} \frac{r^{|l|}}{\sigma_t^{|l|}} \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{r^2}{2\sigma^2}\right)(d-1) r^{d-1}\mathrm{d}r \\
&= \int_{s=\sqrt{4\log \varepsilon^{-1}}}^{\infty} s^{|l|+d-1} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s \quad \text{(by letting } s = r/\sigma_t\text{)} \\
&= \frac{(4\log \varepsilon^{-1})^{(|l|+d-1)/2}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{4\log \varepsilon^{-1}}{2}\right)(d-1) + \int_{s=\sqrt{4\log \varepsilon^{-1}}}^{\infty} \frac{(|l|+d-1)s^{|l|+d-2}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s \\
&= \cdots = \sum_{0 \leq i \leq \lfloor \frac{|l|+d-1}{2} \rfloor} \frac{\frac{(|l|+d-1)!!}{(|l|+d-1-2i)!!}(4\log \varepsilon^{-1})^{(|l|+d-1-2i)/2}(d-1)}{(2\pi)^{\frac{d}{2}}} \varepsilon^2 \\
&\qquad + \begin{cases} \int_{s=\sqrt{4\log \varepsilon^{-1}}}^{\infty} \frac{(|l|+d-1)!!}{(2\pi)^{\frac{d}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{s^2}{2}\right)(d-1)\mathrm{d}s & (|l|+d\text{: even}) \\ 0 & (|l|+d\text{: odd}) \end{cases} \quad \text{(by iterating integration by 部分)} \\
&\lesssim \varepsilon^2 \log^{\frac{d+|l|-1}{2}} \varepsilon^{-1}. \qquad (106)
\end{aligned}
$$

将 ε 替换为 ε/dl，则（106）的右端由下式界定

$$\frac{\varepsilon^2}{d^2 l^2} \log^{\frac{dn+|l|-1}{2}} (\varepsilon/dl)^{-1} \lesssim \varepsilon,$$

由此得出结论。 □

引理 F.15。设 $(B_s)_{[0,t]}$ 为一维布朗运动，且 $X_t = \int_0^t \beta_s \mathrm{d}B_s$，其中 $\beta_s \leq \overline{\beta}$。则有

$$\mathbb{P}\left[\sup_{s\in[0,t]} |X_t| \geq 2\sqrt{\overline{\beta} t \log(2\varepsilon^{-1})}\right] \leq \varepsilon.$$

证明。 我们对以下情形进行界定： $\beta_s \equiv \overline{\beta}$ 因为该情形使命中概率最大。根据 Karatzas 等人 (1991)，对于$x > 0$,

$$\mathbb{P}\left[\sup_{s\in[0,t]} |X_t| \geq x\right] = \frac{4}{\sqrt{2\pi}} \int_{\frac{x}{\sqrt{2\overline{\beta}t}}}^{\infty} e^{-y^2/2}\mathrm{d}y = \frac{4}{\sqrt{2\pi}} \int_{\frac{x}{\sqrt{4\overline{\beta}t}}}^{\infty} e^{-z^2}\sqrt{2}\mathrm{d}z \leq 2e^{-x^2/4\overline{\beta}t}.$$

对于第二个等式，我们只需用 z 替换 $y/\sqrt{2}$。对于最后一个不等式，我们使用了 $\frac{4}{\sqrt{2\pi}} \cdot \sqrt{2} \leq 2$ 和 $\int_x^\infty e^{-y^2}\mathrm{d}y \leq e^{-x^2}$。因此，令 $x = 2\sqrt{\overline{\beta}t\log(2\varepsilon^{-1})}$，即可得到该断言。

□