
Provable Sample-Efficient Transfer Learning Conditional Diffusion Models via Representation Learning

Ziheng Cheng¹, Tianyu Xie², Shiyue Zhang², Cheng Zhang^{2,3,*}

¹ Department of Industrial Engineering and Operations Research, University of California, Berkeley

² School of Mathematical Sciences, Peking University

³ Center for Statistical Science, Peking University

ziheng_cheng@berkeley.edu, tianyuxie@pku.edu.cn,
zhangshiyue@stu.pku.edu.cn, chengzhang@math.pku.edu.cn

Abstract

While conditional diffusion models have achieved remarkable success in various applications, they require abundant data to train from scratch, which is often infeasible in practice. To address this issue, transfer learning has emerged as an essential paradigm in small data regimes. Despite its empirical success, the theoretical underpinnings of transfer learning conditional diffusion models remain unexplored. In this paper, we take the first step towards understanding the sample efficiency of transfer learning conditional diffusion models through the lens of representation learning. Inspired by practical training procedures, we assume that there exists a low-dimensional representation of conditions shared across all tasks. Our analysis shows that with a well-learned representation from source tasks, the sample complexity of target tasks can be reduced substantially. Numerical experiments are also conducted to verify our results.

1 Introduction

Conditional diffusion models (CDMs) utilize a user-defined condition to guide the generative process of diffusion models (DMs) to sample from the desired conditional distribution. In recent years, CDMs have achieved groundbreaking success in various generative tasks, including text-to-image generation [Ho et al., 2020, Song et al., 2020, Ho and Salimans, 2022, Rombach et al., 2022], reinforcement learning [Janner et al., 2022, Chi et al., 2023, Wang et al., 2022, Reuss et al., 2023], time series [Tashiro et al., 2021, Rasul et al., 2021], and life science [Song et al., 2021, Watson et al., 2022, Gruver et al., 2024, Guo et al., 2024].

Training a CDM from scratch requires a large amount of data to achieve good generalization. However, in practical scenarios, users often have access to only limited data for the target distribution due to cost or risk concerns, making the model prone to over-fitting. In such small data regime, transfer learning has emerged as a predominant paradigm [Moon et al., 2022, Ruiz et al., 2023, Xie et al., 2023, Han et al., 2023]. By leveraging knowledge acquired during pre-training on large source datasets, transfer learning enhances the performance of fine-tuning on target tasks, facilitating few-shot learning and significantly improving practicality.

Among the successful applications of transfer learning CDMs, the conditions are typically high-dimensional vectors with embedded low-dimensional representations (features) that encapsulate all the information required for inference. In addition, these representations are likely to be task-agnostic,

*Corresponding Author.

Tasks	Backbone Score Network	Condition Encoder
Text-to-Image [Esser et al., 2024]	2-8B	4.7B
Text-to-Audio [Liu et al., 2024]	350-750M	750M
Robotic Control [Chi et al., 2023]	9M	20-45M

Table 1: Comparing the number of parameters of different parts in CDMs.

enabling effective knowledge transfer. For example, in text-to-image generation, the text input is inherently in high-dimensional space, but contains low-dimensional semantic information such as object attributes, spatial relationships, despite the differences of styles or contents in different image distributions. To take advantage of this structure, condition encoders are often frozen in the fine-tuning stage [Rombach et al., 2022, Esser et al., 2024], which typically constitutes a significant portion of the overall model (see Table 1).

While this paradigm has demonstrated remarkable empirical success, its theoretical underpinnings remain largely unexplored. The following fundamental question is still open:

Can transfer learning CDMs improve the sample efficiency of target tasks by leveraging the representation of conditions learned from source tasks?

There are some recent works attempting to study the theoretical underpinnings of CDMs [Fu et al., 2024, Jiao et al., 2024, Hu et al., 2024], but focus on single task training. Notably, Yang et al. [2024] investigates transfer learning DMs under the assumption that the data is a linear transformation of a low-dimensional latent variable following the same distribution across all tasks. However, fine-tuning merely the data encoder is not a widely adopted training approach in practice.

In this paper, we take the first step towards addressing the above question. Our key assumption is that there exists a generic low-dimensional representation of conditions shared across all distributions. Then we show that, with a well-learned representation from source tasks, the sample complexity of target tasks can be reduced substantially by training only the score network. The main contributions are summarized as follows:

- In Section 3, we establish the first generalization guarantee for transferring score matching error in CDMs, showing that transfer learning can reduce the sample complexity for learning condition encoder in the target task. This is aligned with existing transfer learning theory in supervised learning. Specifically, we present two results in Theorem 3.4 and Theorem 3.6, under the settings of task diversity assumption and meta-learning², respectively. On the technical side, we develop a novel approach to tackle Lipschitz continuity under weaker assumptions on data distribution in Lemma 3.1, which may be of independent interest for the analysis of even single-task diffusion models.
- In Section 4, we provide an end-to-end distribution estimation error bound in transfer learning CDMs. To obtain an L^2 accurate conditional score estimator, we construct a universal approximation theory using deep ReLU neural networks in Theorem 4.1. Then by combining both generalization error and approximation error, Theorem 4.2 and 4.3 provide sample complexity bounds for estimating conditional distribution. Notably, our results are *the state of the art* even when reduced to single-task learning setting.

In Section A, we further utilize our results to establish statistical guarantees in practical applications of CDMs. In particular, we investigate amortized variational inference (Theorem A.1) and behavior cloning (Theorem A.2), and present guarantees in terms of posterior estimation and optimality gap, laying the theoretical foundations of transfer learning CDMs in practice. We also conduct numerical experiments in Section 5 to verify our results.

²In practice, the terms such as transfer learning, meta-learning, learning-to-learn, *etc.*, often refer to the same training paradigm, *i.e.*, to fine-tune on target tasks with limited data using knowledge from source tasks. However, in the theoretical framework, we use meta-learning to emphasize that target tasks and source tasks are randomly sampled from a meta distribution [Baxter, 2000], whereas in transfer learning, the tasks are fixed.

1.1 Related Works

Score Approximation and Distribution Estimation Recently, some works analyze the score approximation theory via deep neural networks and corresponding sample complexity bounds for diffusion models. [Oko et al. \[2023\]](#) considers distributions with density in Besov space and supported on bounded domain. [Chen et al. \[2023b\]](#) assumes the data distribution lies in a low-dimensional linear subspace and obtains improved rates only depending on intrinsic dimension. [Fu et al. \[2024\]](#) studies conditional diffusion models for Hölder densities and [Hu et al. \[2024\]](#) further extends the framework to more advanced neural network architectures, *e.g.*, diffusion transformers. [Wibisono et al. \[2024\]](#) establishes a minimax optimal rate to estimate Lipschitz score by kernel methods. With an L^2 accurate score estimator, several works provide the convergence rate of discrete samplers for diffusion models [[Chen et al., 2022b, 2023a](#), [Lee et al., 2023](#), [Chen et al., 2024](#)]. Combining score matching error and convergence of samplers, one can obtain an end-to-end distribution estimation error bound.

Transfer Learning and Meta-learning Theory in Supervised Learning The remarkable empirical success of transfer learning, meta-learning, and multi-task learning across a wide range of machine learning applications has been accompanied by gradual progress in their theoretical foundations, especially from the perspective of representation learning. To the best of our knowledge, [Baxter \[2000\]](#) is the first theoretical work on meta-learning. It assumes a universal *environment* to generate tasks with some shared features. Following this setting, [Maurer et al. \[2016\]](#) provides sample complexity bound for general supervised learning problem and [Aliakbarpour et al. \[2024\]](#) studies very few samples per task regime. Another line of research replaces the *environment* assumption and instead establishes connections between source tasks and target tasks through various notions of task diversity [[Tripuraneni et al., 2020](#), [Du et al., 2020](#), [Tripuraneni et al., 2021](#), [Watkins et al., 2023](#), [Chua et al., 2021](#)]. However, theoretical understandings of transfer learning for unsupervised learning are much more limited.

Few-shot Fine-Tuning of Diffusion Models Adapting pre-trained conditional diffusion models to specific tasks with limited data remains a challenge in varied application scenarios. Few-shot fine-tuning aims to bridge this gap by leveraging various techniques to adapt those models to a novel task with minimal data requirements [[Ruiz et al., 2023](#), [Giannone et al., 2022](#)]. A promising paradigm is to use transfer (meta) learning by constructing a representation for conditions in all the tasks, which has been widely applied in image generation [[Rombach et al., 2022](#), [Ramesh et al., 2022](#), [Sinha et al., 2021](#)], reinforcement learning [[He et al., 2023](#), [Ni et al., 2023](#)], inverse problem [[Tewari et al., 2023](#), [Chung et al., 2023](#)], *etc.* Another work [Yang et al. \[2024\]](#) is closely related to this paper, proving that few-shot diffusion models can escape the curse of dimensionality by fine-tuning a linear encoder.

2 Preliminaries and Problem Setup

Notations We use x and y to denote the data and conditions, respectively. The blackboard bold letter $\mathbb{P}$ represents the joint distribution of (x, y) , while the lowercase p denotes its density function. The superscript k indicates the task index, and the subscript i means the sample index. The norm $\|\cdot\|$ refers to the ℓ_2 -norm for vectors and the spectral norm for matrices. For the hypothesis class $\mathcal{F}$, we use $\mathcal{F}^{\otimes K}$ to refer its K -fold Cartesian product. For any $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Finally, we use standard $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ to omit constant factors.

2.1 Conditional Diffusion Models

Let $\mathbb{R}^{d_x}$ denote the data space and $[0, 1]^{D_y}$ denote the condition space. Let $\mathbb{P}$ be any joint distribution over $\mathbb{R}^{d_x} \times [0, 1]^{D_y}$ with density p and $\mathbb{P}(\cdot|y)$ be the conditional distribution with density $p(\cdot|y)$. As in diffusion models, the forward process is defined as an Ornstein–Uhlenbeck (OU) process,

$$dX_t = -X_t dt + \sqrt{2}dW_t, X_0 \sim \mathbb{P}(\cdot|y). \quad (2.1)$$

where $\{W_t\}_{t \geq 0}$ is a standard Wiener process. We denote the distribution of X_t as $\mathbb{P}_t(\cdot|y)$. Note that the limiting distribution $\mathbb{P}_\infty(\cdot|y)$ is a standard Gaussian $\mathcal{N}(0, I)$.

To generate new samples, we can reverse the forward process (2.1) from any $T > 0$,

$$dX_t^\leftarrow = (X_t^\leftarrow + 2\nabla \log p_{T-t}(X_t^\leftarrow|y))dt + \sqrt{2}d\bar{W}_t, X_0^\leftarrow \sim \mathbb{P}_T(\cdot|y), 0 \leq t \leq T. \quad (2.2)$$

where $\{\bar{W}_t\}_{0 \leq t \leq T}$ is a time-reversed Wiener process. Unfortunately, we don't have access to the exact conditional score function $\nabla \log p_{T-t}$ and need to estimate it through neural networks. For any $(x, y) \sim \mathbb{P}$ and score estimator s , define the individual denoising score matching objective [Vincent, 2011] as

$$\ell(x, y, s) := \frac{1}{T - T_0} \int_{T_0}^T \mathbb{E}_{x_t \sim \phi_t(\cdot|x)} [\|s(x_t, y, t) - \nabla \log \phi_t(x_t|x)\|^2] dt, \quad (2.3)$$

where $\phi_t(x_t|x) = \mathcal{N}(x_t|\alpha_t x, \sigma_t^2 I)$, $\alpha_t = e^{-t}$, $\sigma_t^2 = 1 - e^{-2t}$, is the transition kernel of $x_t|x_0 = x$. And the population error of score matching is

$$L^\mathbb{P}(s) := \mathbb{E}_{(x,y) \sim \mathbb{P}} \mathbb{E}_{t, x_t} [\|s(x_t, y, t) - \nabla \log p_t(x_t|y)\|^2] = \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell(x, y, s) - \ell(x, y, s_*^\mathbb{P})]. \quad (2.4)$$

Here $s_*^\mathbb{P}$ denotes the true score function and $t \sim \text{Unif}([T_0, T])$. We also define $\ell^\mathbb{P}(x, y, s) := \ell(x, y, s) - \ell(x, y, s_*^\mathbb{P})$. In practice, with a score estimator $\hat{s}$, the generative process is to simulate

$$d\hat{X}_t^\leftarrow = (\hat{X}_t^\leftarrow + 2\hat{s}(\hat{X}_t^\leftarrow, y, T-t))dt + \sqrt{2}d\bar{W}_t, \hat{X}_0^\leftarrow \sim \mathcal{N}(0, I), 0 \leq t \leq T - T_0. \quad (2.5)$$

Here $T_0 > 0$ is the early-stopping time. And the distribution of $\hat{X}_{T-T_0}^\leftarrow$ is written as $\hat{\mathbb{P}}(\cdot|y)$.

Note that we don't apply the commonly used classifier-free guidance [Ho and Salimans, 2022] which has a tunable guidance strength since we mainly concentrate on sampling from conditional distribution instead of optimizing other objectives.

2.2 Transfer Diffusion Models via Learning Representation

Consider K source distributions over $\mathbb{R}^{d_x} \times [0, 1]^{D_y}$, $\mathbb{P}^1, \dots, \mathbb{P}^K$, and a target distribution $\mathbb{P}^0$. Suppose that for each source distribution $\mathbb{P}^k$, $1 \leq k \leq K$, we have n i.i.d. samples $\{(x_i^k, y_i^k)\}_{i=1}^n \sim \mathbb{P}^k$, and m i.i.d. samples $\{(x_i^0, y_i^0)\}_{i=1}^m \sim \mathbb{P}^0$ are available for the target distribution, where typically $m \ll n$. In transfer (meta) learning setup, we assume there exists a shared nonlinear representation of the condition y for all distributions, i.e., the conditional distribution $\mathbb{P}_{x|y}^k = \mathbb{P}_{x|h_*(y)}^k$ for some $h_* : [0, 1]^{D_y} \rightarrow [0, 1]^{d_y}$ (see also Assumption 3.2). Note that due to the shared features, the score of $p_t^k(\cdot|y)$ also has the form of $\nabla \log p_t^k(x_t|y) = f_*^k(x_t, h_*(y), t)$ for some f_*^k .

Similar to Tripuraneni et al. [2020], our transfer learning procedures consist of two phases. In the pre-training phase, the goal is to learn a representation map h_* through nK samples from K source distributions. Then during the fine-tuning phase, we learn the target distribution via m new samples and the representation map learned in the pre-training phase.

Formally, let $\mathcal{F}, \mathcal{H}$ be the hypothesis classes of score networks and representation maps, respectively. Further let $\mathcal{F}^0 \subseteq \mathcal{F}$ be the hypothesis class of score network in fine-tuning phase. In the pre-training phase, we solve the following Empirical Risk Minimization (ERM),

$$\hat{f}, \hat{h} = \arg \min_{f \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \ell(x_i^k, y_i^k, s_{f,h}^k). \quad (2.6)$$

Then for the fine-tuning task, we solve

$$\hat{f}^0 := \arg \min_{f \in \mathcal{F}^0} \frac{1}{m} \sum_{i=1}^m \ell(x_i^0, y_i^0, s_{f,\hat{h}}). \quad (2.7)$$

Here $s_{f,h}(x, y, t) := f(x, h(y), t)$ for $f : \mathbb{R}^{d_x} \times [0, 1]^{d_y} \times [T_0, T] \rightarrow \mathbb{R}^{d_x}$ and $h : [0, 1]^{D_y} \rightarrow [0, 1]^{d_y}$ and ℓ is defined in (2.3).

In the meta-learning setting, we further assume that all the distributions $\{\mathbb{P}^k\}_k$ are i.i.d. sampled from a meta distribution $\mathbb{P}_{\text{meta}}$. Here $\mathbb{P}_{\text{meta}}$ can be interpreted as a universal *environment* [Baxter, 2000, Maurer et al., 2016]. In this case, we posit the existence of a shared representation map that holds for all $\mathbb{P} \sim \mathbb{P}_{\text{meta}}$. And the performance benchmark is then defined as the expected error on the target distribution $\mathbb{P}^0 \sim \mathbb{P}_{\text{meta}}$.

2.3 Deep ReLU Neural Network Family

We use feedforward neural networks to approximate the score function and representation map. Let $\sigma(x) := \max\{x, 0\}$ be the ReLU activation. Define the score network family $NN_f(L, W, M, S, B, R, \gamma) := \{f(x, w, t) = (A_L \sigma(\cdot) + b_L) \circ \dots \circ (A_1 [x^\top, w^\top, t]^\top + b_1) :$

$A_i \in \mathbb{R}^{d_i \times d_{i+1}}, b_i \in \mathbb{R}^{d_{i+1}}, d_{L+1} = d_x, \max d_i \leq W, \|f\|_{L^\infty} \leq M, \sum_{i=1}^L (\|A_i\|_0 + \|b_i\|_0) \leq S, \max \|A_i\|_\infty \vee \|b_i\|_\infty \leq B, \|f(x, w, t) - f(x, w', t)\| \leq \gamma \|w - w'\|_\infty, \forall \|x\|_\infty \leq R, t \leq T\}$, and encoder network $NN_h(L, W, S, B) := \{h(y) = (A_L \sigma(\cdot) + b_L) \circ \dots \circ (A_1 y + b_1) :$
 $A_i \in \mathbb{R}^{d_i \times d_{i+1}}, b_i \in \mathbb{R}^{d_{i+1}}, d_{L+1} = d_y, \max d_i \leq W, \|h\|_{L^\infty([0,1]^{D_y})} \leq 1, \sum_{i=1}^L (\|A_i\|_0 + \|b_i\|_0) \leq S, \max \|A_i\|_\infty \vee \|b_i\|_\infty \leq B\}$. Throughout this paper, we let $\mathcal{F}^0 = \mathcal{F} = NN_f(L_f, W_f, M_f, S_f, B_f, R_f, \gamma_f)$ and $\mathcal{H} = NN_h(L_h, W_h, S_h, B_h)$ unless otherwise specified.

Remark 1. In practice, $\mathcal{F}^0 \subseteq \mathcal{F}$ may (and typically will) depend on $\hat{f}$ for parameter efficient fine-tuning (PEFT), e.g., LoRA [Hu et al., 2021]. This will substantially reduce the complexity of $\mathcal{F}^0$ and further improve sample efficiency. The analysis of PEFT is beyond the scope of this paper.

3 Statistical Guarantees for Transferring Score Matching Error

In this section, we present our main theoretical results, a statistical theory of transferring the conditional score matching loss. We provide two upper bounds of the score matching loss on target distribution, based on whether task diversity [Tripuraneni et al., 2020] is explicitly assumed. Our analysis introduces novel techniques to address the smoothness properties of the noised data distribution—a challenge that remains nontrivial even in single-task settings. Additionally, we extend the classical theory of local Rademacher complexity to quantify the empirical estimation error.

Throughout this paper, we make the following standard and mild regularity assumptions [Tripuraneni et al., 2020, Chen et al., 2023b] on the initial data distribution $\mathbb{P}$ and the representation map h_* .

Assumption 3.1 (Sub-gaussian tail). *For any source or target distribution $\mathbb{P}$, $\mathbb{P}$ is supported on $\mathbb{R}^{d_x} \times [0, 1]^{D_y}$ and admits a continuous density $p(x, y) \in \mathcal{C}^2(\mathbb{R}^{d_x} \times [0, 1]^{D_y})$. Moreover, the conditional distribution $p(x|y) \leq C_1 \exp(-C_2 \|x\|^2)$ for some constant C_1, C_2 .*

Assumption 3.2 (Shared low-dimensional representation). *There exists an L -Lipschitz function $h_* : [0, 1]^{D_y} \rightarrow [0, 1]^{d_y}$ with $d_y \leq D_y$, such that for any source and target distribution $\mathbb{P}$, the conditional density $p(x|y) = g_*^\mathbb{P}(x, h_*(y))$ for some $g_*^\mathbb{P} \in \mathcal{C}^2(\mathbb{R}^{d_x} \times [0, 1]^{d_y})$.*

Equivalently, $h_*(y)$ is a sufficient statistic for x , which indicates that $p_t(x|y) = p_t(x|h_*(y))$. Therefore, with a little abuse of notation, for any $w \in [0, 1]^{d_y}$, we define $p(x; w) = p(x|h_*(y) = w) = g_*^\mathbb{P}(x, w)$. Also note that by definition, for any x, y , we have $p(x; h_*(y)) = p(x|h_*(y)) = p(x|y)$.

Assumption 3.3 (Lipschitz score). *For any source and target distribution $\mathbb{P}$ and its density function p , the conditional score $\nabla_x \log p(x|y) = \nabla_x \log g_*^\mathbb{P}(x, h_*(y))$. The score function $\nabla_x \log g_*^\mathbb{P}(x, w)$ is L -Lipschitz in x and w . And $\|\nabla_x \log g_*^\mathbb{P}(0, w)\| \leq B$ for some constant B and any w .*

3.1 Tackling Lipschitz Continuity under Weaker Assumptions

Notice that we only impose smoothness assumption on the original data distribution $p(\cdot|y)$, instead of the entire trajectory $p_t(\cdot|y)$ in forward process. This is substantially weaker than the Lipschitzness assumption required in Chen et al. [2023b, 2022b], Yuan et al. [2024], Yang et al. [2024]. However, Lipschitzness of loss function ℓ and class $\mathcal{F}$ is a crucial hypothesis in theoretical analysis of transfer learning [Tripuraneni et al., 2020, Chua et al., 2021]. The intuition is that without Lipschitz continuity of the score network f , it is generally impossible to characterize the error from an imperfect representation map h . Hence it is inevitable to show the smoothness of $p_t(\cdot|y)$ to some extent.

Fortunately, even with assumptions merely on the initial data distribution, we are still able to prove smoothness of the forward process in any bounded region, as shown in the following lemma. The proof can be found in Appendix B.1.

Lemma 3.1. Under Assumption 3.1, 3.2, 3.3, for any $w \in [0, 1]^{d_y}$, denote the conditional score of forward process $\nabla_x \log p_t(x; w)$ by $f_*(x, w, t)$. There exist constants C_X, C'_X , such that for any $R > 0$, the function $f_*(x, w, t)$ is $(C_X + C'_X R^2)$ -Lipschitz in x , $(C_X + C'_X R)$ -Lipschitz in w , in the domain $\mathcal{B}_R \times [0, 1]^{d_y} \times [0, T]$. Here $\mathcal{B}_R$ denotes the ball with radius R centered at the origin.

3.2 Results under Task Diversity: Sample-Efficient Transfer Learning

In the literature of transfer learning, task diversity is an important assumption that connects target tasks with source tasks [Tripuraneni et al., 2020, Du et al., 2020, Chua et al., 2021]. In the context of conditional diffusion models, we state the formal definition as follows.

Definition 3.1 (Task diversity). Given hypothesis classes $\mathcal{F}, \mathcal{H}$, we say the source distributions $\mathbb{P}^1, \dots, \mathbb{P}^K$ are (ν, Δ) -diverse over target distribution $\mathbb{P}^0$, if for any representation $h \in \mathcal{H}$,

$$\inf_{f^0 \in \mathcal{F}^0} L^{\mathbb{P}^0}(s_{f^0, h}) \leq \frac{1}{\nu} \inf_{f \in \mathcal{F}^{\otimes K}} \frac{1}{K} \sum_{k=1}^K L^{\mathbb{P}^k}(s_{f^k, h}) + \Delta. \quad (3.1)$$

Here $L^{\mathbb{P}}$ is defined in (2.4). This notion of diversity ensures that the representation error on the target task caused by $\hat{h}$ can be controlled by the error on the source tasks, thereby establishing certain relationships in between. More detailed discussions are deferred to Appendix B.5.

We first present the generalization guarantee for each phase respectively.

Proposition 3.2 (Fine-tuning phase generalization). Under Assumption 3.1, 3.2, 3.3, for any $\hat{h} \in \mathcal{H}$, the population loss of $\hat{f}^0$ can be bounded by

$$\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}^0} \mathbb{E}_{(x, y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}^0}(x, y, s_{\hat{f}^0, \hat{h}})] \lesssim \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}^0}(x, y, s_{f, \hat{h}})] + \log^3(m) r_x, \quad (3.2)$$

where $r_x = \frac{\log \tilde{\mathcal{N}}_{\mathcal{F}}}{m}$ and $\log \tilde{\mathcal{N}}_{\mathcal{F}}$ is some complexity measures of $\mathcal{F}$.

Proposition 3.3 (Pre-training phase generalization). Under Assumption 3.1, 3.2, 3.3, if $R_f \gtrsim \log^{\frac{1}{2}}(nK M_f / \delta)$, with probability no less than $1 - \delta$, the population loss can be bounded by

$$\frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} \ell^{\mathbb{P}^k}(x, y, s_{\hat{f}^k, \hat{h}}) \lesssim \inf_{f \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f^k, h})] + \log^3(nK / \delta) \left(r_z + \frac{\log(1/\delta)}{nK} \right), \quad (3.3)$$

where $r_z := \frac{K \log \tilde{\mathcal{N}}_{\mathcal{F}} + \log \tilde{\mathcal{N}}_{\mathcal{H}}}{nK}$ and $\log \tilde{\mathcal{N}}_{\mathcal{F}}, \log \tilde{\mathcal{N}}_{\mathcal{H}}$ are some complexity measures of $\mathcal{F}, \mathcal{H}$.

Combining these two propositions with the notion of task diversity in Definition 3.1, we are able to show the statistical rate of transfer learning as follows.

Theorem 3.4. Under Assumption 3.1, 3.2, 3.3, suppose $\mathbb{P}^1, \dots, \mathbb{P}^K$ are (ν, Δ) -diverse over target distribution $\mathbb{P}^0$ given $\mathcal{F}, \mathcal{H}$. If $R_f \gtrsim \log^{\frac{1}{2}}(nK M_f / \delta)$, then with probability no less than $1 - \delta$,

$$\begin{aligned} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{(x, y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}^0}(x, y, s_{\hat{f}^0, \hat{h}})] &\lesssim \frac{1}{\nu} \inf_{h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f, h})] + \Delta \\ &\quad + \frac{\log^3(m) \log \mathcal{N}_{\mathcal{F}}}{m} + \frac{\log^3(nK / \delta) (K \log \mathcal{N}_{\mathcal{F}} + \log(\mathcal{N}_{\mathcal{H}} / \delta))}{\nu nK}. \end{aligned} \quad (3.4)$$

where

$$\begin{aligned} \log \mathcal{N}_{\mathcal{F}} &:= M_f^2 S_f L_f \log(mn L_f W_f (B_f \vee 1) M_f T \log(1/\delta)), \\ \log \mathcal{N}_{\mathcal{H}} &:= S_h L_h \log(nK L_h W_h (B_h \vee 1) M_f \gamma \log(1/\delta)). \end{aligned} \quad (3.5)$$

The formal statements and proofs are provided in Appendix B.2.

Let $\varepsilon_{\text{approx}} = \inf_{h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f,h})]$ be the approximation error. The leading terms can be simplified to $\tilde{\mathcal{O}} \left(\varepsilon_{\text{approx}} + \frac{K \log \mathcal{N}_{\mathcal{F}} + \log \mathcal{N}_{\mathcal{H}}}{nK} + \frac{\log \mathcal{N}_{\mathcal{F}}}{m} \right)$, where $\log \mathcal{N}_{\mathcal{F}}$ and $\log \mathcal{N}_{\mathcal{H}}$ capture the complexity of the hypothesis classes.

Improving Sample Efficiency Theorem 3.4 demonstrates the sample efficiency of transfer learning. Compared to naively training the full CDM for target distribution, which has an error of $\tilde{\mathcal{O}} \left(\varepsilon_{\text{approx}} + \frac{\log \mathcal{N}_{\mathcal{F}} + \log \mathcal{N}_{\mathcal{H}}}{m} \right)$, transfer learning saves the complexity of learning $\mathcal{H}$ and thus the performance is much better when m is relatively small to n, K (i.e., in few-shot learning setting).

3.3 Results without Task Diversity: Meta-Learning Perspective

The results in previous section heavily depend on the task diversity assumption, which is hard to verify in practice. An alternative is to consider meta-learning setting, where all source and target distributions are sampled from the same *environment*, i.e., a meta distribution.

For any $h \in \mathcal{C}([0, 1]^{D_y}; [0, 1]^{d_y})$ and distribution $\mathbb{P}$ over $\mathbb{R}^{d_x} \times [0, 1]^{D_y}$, define the representation error as

$$\mathcal{L}(\mathbb{P}, h) := \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})] \geq 0. \quad (3.6)$$

We characterize the generalization bound of source tasks on the entire meta distribution as follows.

Proposition 3.5 (Generalization on meta distribution). *Under Assumption 3.1, 3.2, 3.3, there exists constant C_P such that for $\{\mathbb{P}^k\}_{k=1}^K \stackrel{i.i.d.}{\sim} \mathbb{P}_{\text{meta}}$, with probability no less than $1 - \delta$,*

$$\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathcal{L}(\mathbb{P}, h) \leq \frac{2}{K} \sum_{k=1}^K \mathcal{L}(\mathbb{P}^k, h) + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right), \quad (3.7)$$

$$\frac{1}{K} \sum_{k=1}^K \mathcal{L}(\mathbb{P}^k, h) \leq 2 \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathcal{L}(\mathbb{P}, h) + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right), \quad (3.8)$$

holds for any $h \in \mathcal{H}$, where $r_P = M_f^2 \exp(-\Omega(R_f^2)) + \frac{S_h L_h \log(K L_h W_h (B_h \vee 1) M_f \gamma_f)}{K}$.

Theorem 3.6. *Under Assumption 3.1, 3.2, 3.3, if $R_f \gtrsim \log^{\frac{1}{2}}(n K M_f / \delta)$, then with probability no less than $1 - \delta$, the expected population loss of new task can be bounded by*

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}^0 \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}^0} \mathbb{E}_{(x,y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}}(x, y, s_{\hat{f}_0, \hat{h}})] \\ & \lesssim \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})] + \frac{\log^3(m) \log \mathcal{N}_{\mathcal{F}}}{m} + \frac{\log^3(n K / \delta) \log \mathcal{N}_{\mathcal{F}}}{n} + \frac{\log(\mathcal{N}_{\mathcal{H}} / \delta)}{K}, \end{aligned} \quad (3.9)$$

where $\log \mathcal{N}_{\mathcal{F}}, \log \mathcal{N}_{\mathcal{H}}$ are defined in (3.5).

The formal statements and proofs are provided in Appendix B.3.

Let $\tilde{\varepsilon}_{\text{approx}} = \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})]$ be the approximation error in meta-learning.

The results above can be further simplified to $\tilde{\mathcal{O}} \left(\tilde{\varepsilon}_{\text{approx}} + \frac{\log \mathcal{N}_{\mathcal{F}}}{m \wedge n} + \frac{\log \mathcal{N}_{\mathcal{H}}}{K} \right)$. Different from transfer learning bound in Theorem 3.4, the leading term decays only in K and not in n . This is because that without task diversity assumption, the connection between source distributions and target distributions can only be constructed through meta distribution. And according to Proposition 3.5, the source distributions $\mathbb{P}^1, \dots, \mathbb{P}^K$ collectively form a K -shot empirical estimation of $\mathbb{P}_{\text{meta}}$, leading to an estimation error of $\mathcal{O}(1/K)$. Despite this, Theorem 3.6 still demonstrates the sample efficiency of meta-learning compared to naive training method when m is small and n, K are sufficient large.

4 End-to-End Distribution Estimation via Deep Neural Network

Section 3 provides a statistical guarantee for transferring score matching. In this section, we establish an approximation theory using deep neural network to quantify the misspecification error. Combining both results we are able to obtain an end-to-end distribution estimation error bound for transfer learning diffusion models.

4.1 Score Neural Network Approximation

The following theorem provides a guarantee for the ability of deep ReLU neural networks to approximate score and representation. The proof is provided in Appendix C.1.

Theorem 4.1. *Under Assumption 3.1, 3.2, 3.3, to achieve $R_f \gtrsim \log^{\frac{1}{2}}(nKM_f/\delta)$ and*

$$\inf_{h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f,h})] = \mathcal{O}(\log^2(nK/(\varepsilon\delta))\varepsilon^2), \text{ (transfer learning)} \quad (4.1)$$

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{meta}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})] = \mathcal{O}(\log^2(nK/(\varepsilon\delta))\varepsilon^2), \text{ (meta-learning)} \quad (4.2)$$

the configuration of $\mathcal{F}$ and $\mathcal{H}$ should satisfy

$$\begin{aligned} L_f &= \mathcal{O}\left(\log\left(\frac{\log(nK/(\varepsilon\delta))}{\varepsilon}\right)\right), W_f = \mathcal{O}\left(\frac{\log^{3(d_x+d_y)/2}(nK/(\varepsilon\delta))}{\varepsilon^{d_x+d_y+1}T_0^3}\right), \\ S_f &= \mathcal{O}\left(\frac{\log^{3(d_x+d_y)/2+1}(nK/(\varepsilon\delta))}{\varepsilon^{d_x+d_y+1}T_0^3}\right), B_f = \mathcal{O}\left(\frac{T \log^{\frac{3}{2}}(nK/(\varepsilon\delta))}{\varepsilon}\right), \\ R_f &= \mathcal{O}\left(\log^{\frac{1}{2}}(nK/(\varepsilon\delta))\right), M_f = \mathcal{O}(\log^3(nK/(\varepsilon\delta))), \gamma_f = \mathcal{O}(\log(nK/(\varepsilon\delta))), \end{aligned} \quad (4.3)$$

$$L_h = \mathcal{O}(\log(1/\varepsilon)), W_h = \mathcal{O}(\varepsilon^{-D_y} \log(1/\varepsilon)), S_h = \mathcal{O}(\varepsilon^{-D_y} \log^2(1/\varepsilon)), B_h = \mathcal{O}(1). \quad (4.4)$$

Here $\mathcal{O}(\cdot)$ hides all the polynomial factors of $d_x, d_y, D_y, C_1, C_2, L, B$.

Universal approximation of deep ReLU neural networks in a bounded region has been widely studied [Yarotsky, 2017, Schmidt-Hieber, 2020]. However, we have to deal with an unbounded domain here, hence more refined analysis is required, e.g. truncation arguments.

In addition, traditional approximation theories typically cannot provide Lipschitz continuity guarantees, which is crucial in transfer learning analysis. Following the constructions in Chen et al. [2023b], the Lipschitzness restriction doesn't compromise the approximation ability of neural networks, while ensuring validity of the generalization analysis in Section 3.

4.2 Distribution Estimation Error Bound

Given the approximation and generalization results, we are in the position of bounding the distribution estimation error of our transfer (meta) learning procedures. The formal statements and proofs can be found in Appendix C.2.

Theorem 4.2 (Transfer learning). *Under Assumption 3.1, 3.2, 3.3 and (ν, Δ) -diversity with proper configuration of neural network family and T, T_0 , it holds that with probability at least $1 - \delta$,*

$$\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}^0} \mathbb{E}_{y \sim \mathbb{P}_y^0} [\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0)] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3((m/\nu) \wedge n)}{\nu^{\frac{1}{2}}((m/\nu) \wedge n)^{\frac{1}{d_x+d_y+9}}} + \frac{\log^2(nK/\delta)}{\nu^{\frac{1}{2}}(nK)^{\frac{1}{D_y+2}}} + \sqrt{\Delta}. \quad (4.5)$$

Theorem 4.3 (Meta-learning). *Under Assumption 3.1, 3.2, 3.3 and meta-learning setting, with proper configuration of neural network family and T, T_0 , it holds that with probability at least $1 - \delta$,*

$$\mathbb{E}_{\mathbb{P}^0 \sim \mathbb{P}_{meta}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}^0} \mathbb{E}_{y \sim \mathbb{P}_y^0} [\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0)] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3(m \wedge n)}{(m \wedge n)^{\frac{1}{d_x+d_y+9}}} + \frac{\log^2(nK/\delta)}{K^{\frac{1}{D_y+2}}}. \quad (4.6)$$

m	10	20	30	40	50	100
fine-tuning	14.47	3.68	2.45	1.82	1.9	0.91
train-from-scratch	21.99	10.61	5.71	2.38	1.77	1.04

Table 2: MSEs for $\beta_0 = 5.5$.

m	10	20	30	40	50	100
fine-tuning	6.14	2.65	1.61	1.08	0.96	0.45
train-from-scratch	24.41	20.62	18.67	13.49	7.03	1.23

Table 3: MSEs for $\beta_0 = 15$.

Theorem 4.2 and 4.3 again unveil the benefits of transfer (meta) learning for conditional diffusion models, with a rate of $\tilde{\mathcal{O}}((m \wedge n)^{-\frac{1}{d_x+d_y+9}} + (nK)^{-\frac{1}{D_y+2}})$ or $\tilde{\mathcal{O}}((m \wedge n)^{-\frac{1}{d_x+d_y+9}} + K^{-\frac{1}{D_y+2}})$. To compare, naively learning the target distribution in isolation will yield $\tilde{\mathcal{O}}(m^{-\frac{1}{d_x+d_y+9}})$. When the condition dimension D_y is much larger than feature dimension d_y , transfer (meta) learning can substantially improve sample efficiency on target tasks, thanks to representation learning.

Comparison with Existing Complexity Bounds of CDMs Fu et al. [2024] studies conditional diffusion model for sub-gaussian distributions with β -Hölder density. Since the Lipschitzness of score is analogous to the requirement of twice differentiability of density [Wibisono et al., 2024], it is reasonable to let $\beta = 2$ for a fair comparison. In this case, the TV distance is bounded by $\tilde{\mathcal{O}}(m^{-\frac{1}{2(d_x+D_y+2)}})$ with sample size m according to Fu et al. [2024], which is worse than our naive bound $\tilde{\mathcal{O}}(m^{-\frac{1}{d_x+D_y+9}})$ due to the inefficiency of score approximation. We are also aware of another work [Jiao et al., 2024] that assumes Lipschitz density and score, obtaining a rate of $\tilde{\mathcal{O}}(m^{-\frac{1}{2(d_x+3)(d_x+D_y+3)}})$.

Relation to Yang et al. [2024] Unlike our setup, Yang et al. [2024] considers transfer learning unconditional diffusion models with only one source task, i.e., $D_y = d_y = 0, K = 1$. The unconditional distribution is assumed to be supported in a low-dimensional linear subspace, where the source task and the target task have the same latent variable distribution. Hence, *only* a linear encoder is trained for fine-tuning instead of the full score network. In this case, Yang et al. [2024] is able to bound the TV distance by $\tilde{\mathcal{O}}(m^{-\frac{1}{4}} + n^{-\frac{1-\alpha(n)}{d_x+5}})$, escaping the curse of dimensionality for target task. However, the assumption on shared latent variable distribution is stringent and we believe our analysis methods can be extended to this setting as well.

5 Experiments

Our theoretical results can be readily applied in various real world settings. In Appendix A, we investigate amortized variational inference and behavior cloning utilizing our theories, providing statistical guarantees of practical applications of CDMs. In addition, we conduct experiments on both synthetic and real world data to numerically verify the sample efficiency of transfer learning.

Conditioned Diffusion The first numerical example is the high-dimensional conditioned diffusion [Cui et al., 2016, Yu et al., 2023] arising from the following Langevin SDE

$$du_s = \beta u_s(1 - u_s^2)ds + dw_s, \quad u_0 = 0, \quad (5.1)$$

where $\beta > 0$ and w_s is a one-dimensional standard Brownian motion. The SDE (5.1) is discretized by the Euler-Maruyama scheme with a step size of 0.02, which defines the prior distribution $p_\beta(x)$ for the (discretized) trajectory $x = (u_{0.02}, u_{0.04}, \dots, u_{1.00})^\top \in \mathbb{R}^{50}$. We consider a conditional Gaussian likelihood function, $p(y|x) = \mathcal{N}(Mx, I_{100}/4)$, where $M \in \mathbb{R}^{100 \times 50}$ is a pre-defined projection matrix. Given a set of pre-selected $\{\beta_k; 1 \leq k \leq K\}$ with $\beta_k = k$ and $K = 10$, the k -th joint source distribution is given by $\mathbb{P}^k(x, y) = p_{\beta_k}(x)p(y|x)$. The target distribution $\mathbb{P}^0(x, y)$ is given by $\beta_0 = 5.5$ (in-domain) or $\beta_0 = 15$ (out-of-domain). More details are found in Appendix E.1.

We report the MSEs of the estimated posterior mean of $\mathbb{P}^0(x|y)$ on the test samples in Table 2 and 3. We see that across different values of β and m , the fine-tuned models can provide significantly more accurate posterior mean estimations in most cases, suggesting the effectiveness of the representation map $\hat{h}$ learned in the pre-training phase. Notably, as the number of fine-tuning samples m increases, the performance gaps between fine-tuned models and train-from-scratch models get smaller, since more training samples yield more generalization benefits and thus less dependence on the pre-trained

m	10	20	30	40	50	100
fine-tuning	0.3799	0.2846	0.2544	0.2406	0.2404	0.2268
train-from-scratch	0.4409	0.3180	0.2746	0.2551	0.2501	0.2344

Table 4: MSEs on the image restoration task.

model. This is aligned with our theoretical results. We also notice a large variance among the results of different replicates, and attribute the slightly worse performance of fine-tuned models at $m = 50, \beta_0 = 5.5$ to the potential randomness.

Image Restoration For a real data experiment, we consider the image restoration task on MNIST. We have $K = 9$ source tasks with $\mathbb{P}^k(x, y) = p_k(x)p(y|x)$, where the prior $p_k(x)$ is the data distribution of the digit k in the MNIST data set ($1 \leq k \leq K$) and $p(y|x) = \mathcal{N}(x, I_{784}/4)$. The target task is $\mathbb{P}^0(x, y) = p_0(x)p(y|x)$, where $p_0(x)$ is the data distribution of the digit 0. We use the full MNIST 1-9 data for pre-training which corresponds to $n = 5000$. For the finetuning phase, we consider $m = 10, 20, 30, 40, 50, 100$ training samples and 100 test samples from $\mathbb{P}^0(x, y)$. More details can be found in Appendix E.2.

We report the MSEs between estimated posterior mean of $\mathbb{P}^0(x, y) = p_0(x)p(y|x)$ and the ground truth sample x on the test samples in Table 4. We see that for all fine-tuning sample sizes m , the results obtained by fine-tuning consistently outperform those obtained by training from scratch, indicating the benefits of transfer learning. Similarly to the experiment on conditioned diffusion, we also observe a reduced performance gap as m increases.

6 Conclusion and Discussion

In this paper, we take the first step towards understanding the sample efficiency of transfer learning conditional diffusion models from the perspective of representation learning. We provide a generalization guarantee for transferring score matching in CDMs in different settings. We further establish an end-to-end distribution estimation error bound using deep neural networks. Two practical applications are investigated based on our theoretical results. We hope this work can motivate future theoretical study on the popular transfer learning paradigm in generative AIs.

Although this work provides the first statistical guarantee for transfer learning in CDMs, it has several limitations that we plan to address in future research. First, our theoretical results heavily rely on the task diversity notion introduced in Section 3.1, which can be challenging to verify in practice. While we provide some preliminary empirical evidence in Appendix B.5, a more fine-grained theoretical and empirical analysis will be essential for a deeper understanding of CDMs. Second, our analysis focuses on the ERM estimator, whereas in practice, fine-tuning typically starts from a pre-trained model and may employ techniques such as LoRA. Incorporating these settings would allow for an optimization-based perspective on the sample efficiency of transfer learning. Finally, in our current formulation, the sample efficiency gain arises from reducing the complexity associated with learning the conditional encoder. Consequently, our results primarily apply to CDMs in which the conditional encoder constitutes a substantial part of the overall model. Extending the theory to settings where this assumption does not hold is an important direction for future work.

Acknowledgements

This work was supported by National Natural Science Foundation of China (grant no. 12201014, grant no. 12292980 and grant no. 12292983). The research of Cheng Zhang was support in part by National Engineering Laboratory for Big Data Analysis and Applications, the Key Laboratory of Mathematics and Its Applications (LMAM) and the Key Laboratory of Mathematical Economics and Quantitative Finance (LMEQF) of Peking University. The authors are grateful for the computational resources provided by the High-performance Computing Platform of Peking University. The authors appreciate the anonymous NeurIPS reviewers for their constructive feedback.

References

- Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.
- Maryam Aliakbarpour, Konstantina Bairaktari, Gavin Brown, Adam Smith, Nathan Srebro, and Jonathan Ullman. Metalearning with very few samples per task. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 46–93. PMLR, 2024.
- Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.
- Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.
- Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023a.
- Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.
- Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022a.
- Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023b.
- Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022b.
- Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36, 2024.
- Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- Kurtland Chua, Qi Lei, and Jason D Lee. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871–8884, 2021.
- Hyungjin Chung, Dohoon Ryu, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Solving 3d inverse problems using pre-trained 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22542–22551, 2023.
- Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.
- Tiangang Cui, Kody JH Law, and Youssef M Marzouk. Dimension-independent likelihood-informed mcmc. *Journal of Computational Physics*, 304:109–137, 2016.
- Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.
- Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2022.
- Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36, 2024.
- Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2):136–154, 2024.
- Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.
- Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xue-long Li. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. *Advances in neural information processing systems*, 36:64896–64917, 2023.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Jerry Yao-Chieh Hu, Weimin Wu, Yi-Chen Lee, Yu-Chao Huang, Minshuo Chen, and Han Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. *arXiv preprint arXiv:2411.17522*, 2024.
- Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.
- Yuling Jiao, Lican Kang, Jin Liu, Heng Peng, and Heng Zuo. Model free prediction with uncertainty assessment. *arXiv preprint arXiv:2405.12684*, 2024.
- Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.
- Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- Pascal Massart. About the constants in talagrand’s concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

- Taehong Moon, Moonseok Choi, Gayoung Lee, Jung-Woo Ha, and Juho Lee. Fine-tuning diffusion models with limited data. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.
- Fei Ni, Jianye Hao, Yao Mu, Yifu Yuan, Yan Zheng, Bin Wang, and Zhixuan Liang. Metadiffuser: Diffusion model as conditional planner for offline meta-rl. In *International Conference on Machine Learning*, pages 26087–26105. PMLR, 2023.
- Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.
- Vitchyr H Pong, Ashvin V Nair, Laura M Smith, Catherine Huang, and Sergey Levine. Offline meta-reinforcement learning with online self-supervision. In *International Conference on Machine Learning*, pages 17811–17829. PMLR, 2022.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pages 8857–8868. PMLR, 2021.
- Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. *arXiv preprint arXiv:2304.02532*, 2023.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.
- Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL <https://doi.org/10.1214/19-AOS1875>.
- Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34: 12533–12548, 2021.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.
- Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csd: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.
- Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezhikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems*, 36: 12349–12362, 2023.
- Nilesh Tripathi, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.
- Nilesh Tripathi, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.
- Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.

- Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.
- Austin Watkins, Enayat Ullah, Thanh Nguyen-Tang, and Raman Arora. Optimistic rates for multi-task representation learning. *Advances in Neural Information Processing Systems*, 36:2207–2251, 2023.
- Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv*, pages 2022–12, 2022.
- Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*, 2024.
- Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Diffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2023.
- Ruofeng Yang, Bo Jiang, Cheng Chen, Ruinan Jin, Baoxiang Wang, and Shuai Li. Few-shot diffusion models escape the curse of dimensionality. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=JrraNaaZm5>.
- Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94: 103–114, 2017.
- Longlin Yu, Tianyu Xie, Yu Zhu, Tong Yang, Xiangyu Zhang, and Cheng Zhang. Hierarchical semi-implicit variational inference with application to diffusion model acceleration. *Advances in Neural Information Processing Systems*, 36, 2023.
- Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36, 2024.

A Applications

We explore two applications of transfer learning for conditional diffusion models, supported by theoretical guarantees derived from our earlier results. In particular, we study amortized variational inference and behavior cloning. These real-world use cases not only validate the applicability of our theoretical findings but also lay the foundations of transferring diffusion models in practice.

A.1 Amortized Variational Inference

Diffusion models have exhibited groundbreaking success in probabilistic inference, especially latent variable models. We study a simple amortized variational inference model, where the observation y given latent variable x is distributed according to an exponential family $\mathcal{F}_\Psi$ with density

$$p_\psi(y|x) = \psi(y) \exp(\langle x, h_*(y) \rangle - A_\psi(x)), \quad (\text{A.1})$$

where $\psi \in \Psi$ is non-negative and supported on $[0, 1]^D$ and $h_*(y) \in [0, 1]^d$. Note that we also have $d_x = d$ in this case. The prior distribution of variable x is denoted as p_ϕ for some $\phi \in \Phi$. Let $\theta = (\psi, \phi)$ and we aim to sample from the posterior distribution of $p_\theta(x|y) \propto p_\phi(x)p_\psi(y|x) \propto p_\phi(x) \exp(\langle x, h_*(y) \rangle - A_\psi(x))$. Due to the special structure, the posterior $p_\theta(x|y)$ only depends on the low-dimensional feature $h_*(y)$, shared across all $\theta \in \Theta := \Psi \times \Phi$. This formulation encompasses various applications including independent component analysis [Comon, 1994], inverse problem [Song et al., 2021, Ajay et al., 2022] and variational Bayesian inference [Kingma, 2013].

Consider source tasks consisting of $\theta^1, \dots, \theta^K \in \Theta$, and for each θ^k we have n i.i.d. samples $\{(x_i^k, y_i^k)\}_{i=1}^n$. For the target task θ^0 , we only have m samples $\{(x_i^0, y_i^0)\}_{i=1}^m$. We conduct our transfer learning procedures to train a conditional diffusion models $\hat{\mathbb{P}}_{\theta^0}(\cdot|y)$. For theoretical analysis, we further impose some assumptions on the probabilistic model as follows.

Assumption A.1. *The prior distribution satisfies $p_\phi(x) \leq C_1 \exp(-C_2\|x\|^2)$ and $\nabla_x \log p_\phi(x)$ is L -Lipschitz in x , $\|\nabla_x \log p_\phi(0)\| \leq B$ for any $\phi \in \Phi$. The representation h_* is L -Lipschitz. The integral $\int \psi(y)dy \in [1/C, C]$ for any $\psi \in \Psi$.*

Theorem A.1. *Suppose Assumption A.1 holds. Then under meta-learning setting, we have with probability no less than $1 - \delta$,*

$$\mathbb{E}_{\theta^0} \mathbb{E}_{\{(x_i^0, y_i^0)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_{\theta^0}} [\text{TV}(\hat{\mathbb{P}}_{\theta^0}(\cdot|y), \mathbb{P}_{\theta^0}(\cdot|y))] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3(m \wedge n)}{(m \wedge n)^{\frac{1}{2d+9}}} + \frac{\log^2(nK/\delta)}{K^{\frac{1}{d+2}}}. \quad (\text{A.2})$$

If (ν, Δ) -diversity holds, then we have with probability no less than $1 - \delta$,

$$\mathbb{E}_{\{(x_i^0, y_i^0)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_{\theta^0}} [\text{TV}(\hat{\mathbb{P}}_{\theta^0}(\cdot|y), \mathbb{P}_{\theta^0}(\cdot|y))] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3((m/\nu) \wedge n)}{\nu^{\frac{1}{2}}((m/\nu) \wedge n)^{\frac{1}{2d+9}}} + \frac{\log^2(nK/\delta)}{\nu^{\frac{1}{2}}(nK)^{\frac{1}{d+2}}} + \sqrt{\Delta}. \quad (\text{A.3})$$

The proof is deferred to Appendix D.1. We show that under mild assumptions, transfer (meta) learning diffusion models can improve the sample efficiency for target task in the context of amortized variational inference. This error bound can be further extended to establish guarantees for statistical inference such as moment prediction, uncertainty assessment, etc.

A.2 Behavior Cloning via Meta-Diffusion Policy

Although originally developed for image generation tasks, diffusion models have recently been extended to reinforcement learning (RL) [Janner et al., 2022, Chi et al., 2023, Wang et al., 2022], enabling the modeling of complex distributions of dynamics and policies. In the context of meta-RL, some works have further utilized diffusion models for planning and synthesis tasks [Ni et al., 2023, He et al., 2023]. In this application, we focus on a popular framework of behavior cloning, *diffusion policy* [Chi et al., 2023], which uses conditional diffusion models to learn multi-modal expert policies in high-dimensional state spaces. In such settings, the state often corresponds to visual observations of the robot's surroundings, such as high resolution images, and thus typically share a low-dimensional underlying representation.

Let $\mathcal{M}$ be the space of decision-making environments, where each $M \in \mathcal{M}$ is an infinite horizon Markov Decision Process (MDP) sharing the same state space $\mathcal{S}$, action space $\mathcal{A}$, discount factor γ and initial distribution $\rho \in \Delta(\mathcal{S})$. And each $M \in \mathcal{M}$ has its own transition kernel $\mathcal{T}_M : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$, and reward function $r_M : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$. The policy is defined as a map $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$. The value function of MDP M under policy π is

$$\begin{aligned} V_M(\pi, s_0) &:= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_M(s_t, a_t) \right], a_t \sim \pi(\cdot | s_t), s_{t+1} \sim \mathcal{T}_M(\cdot | s_t, a_t), \\ V_M(\pi) &:= \mathbb{E}_{s_0 \sim \rho} [V_M(\pi, s_0)]. \end{aligned} \quad (\text{A.4})$$

Denote the visitation measure as $d_M^\pi(s, a) := (1 - \gamma) \mathbb{E}_{s_0 \sim \rho} \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s | \pi, s_0) \pi(a | s)$.

Suppose there are K source tasks $M^1, \dots, M^K \in \mathcal{M}$, and the expert policy of each task is denoted as π_*^k . In behavior cloning, for each source task M^k , we have n pairs of $\{(s_i^k, a_i^k)\}_{i=1}^n \stackrel{i.i.d.}{\sim} d_*^k := d_{M^k}^{\pi_*^k}$. The goal is to imitate the expert policy of target task $M^0 \in \mathcal{M}$, of which the sample size is only $m \ll n$.

To unify the notation, let $x = a, y = s$ and assume $\mathcal{A} = \mathbb{R}^{d_a}, \mathcal{S} = [0, 1]^{D_s}$ and representation space $[0, 1]^{d_s}$. Our meta diffusion-policy framework aims to learn a state encoder $h : \mathcal{S} \rightarrow [0, 1]^{d_s}$ during pre-training, which acts as a shared representation map in different MDPs and consequently enhances sample efficiency on fine-tuning tasks. Let $\hat{\pi}^0$ be the learned policy in fine-tuning phase. The following theorem shows the optimality gap between the learned policy and the expert policy.

Theorem A.2. *Suppose the expert policy π_*^k satisfies Assumption 3.1, 3.2, 3.3. Then under meta-learning setting, it holds that with probability no less than $1 - \delta$,*

$$\mathbb{E}_{M^0} \mathbb{E}_{\{(s_i^0, a_i^0)\}_{i=1}^m \sim d_*^0} [V_{M^0}(\pi_*^0) - V_{M^0}(\hat{\pi}^0)] \lesssim \frac{1}{(1 - \gamma)^2} \left[\frac{\log^{\frac{5}{2}}(nK/\delta) \log^3(m \wedge n)}{(m \wedge n)^{\frac{1}{d_a + d_s + 9}}} + \frac{\log^2(nK/\delta)}{K^{\frac{1}{D_s + 2}}} \right]. \quad (\text{A.5})$$

If we further assume $\pi_*^1, \dots, \pi_*^K$ are (ν, Δ) -diverse over π_*^0 , then the gap can be improved by

$$\mathbb{E}_{\{(s_i^0, a_i^0)\}_{i=1}^m \sim d_*^0} [V_{M^0}(\pi_*^0) - V_{M^0}(\hat{\pi}^0)] \lesssim \frac{1}{(1 - \gamma)^2} \left[\frac{\log^{\frac{5}{2}}(nK/\delta) \log^3((m/\nu) \wedge n)}{\nu^{\frac{1}{2}}((m/\nu) \wedge n)^{\frac{1}{d_a + d_s + 9}}} + \frac{\log^2(nK/\delta)}{\nu^{\frac{1}{2}}(nK)^{\frac{1}{D_s + 2}}} + \sqrt{\Delta} \right]. \quad (\text{A.6})$$

The proof can be found in Appendix D.2. This provides the first statistical guarantee of diffusion policy in behavior cloning. Notably, in both cases, the number of source tasks K has an exponential dependence on D_s , further suggesting the importance of data coverage when tackling distribution shift in offline meta-RL [Pong et al., 2022].

B Proofs in Section 3

B.1 Preliminaries

Lemma B.1. *If $x_0 \sim p(x_0 | y)$, the density of forward process $p_t(x | y)$ can be written as*

$$p_t(x | y) = \int \phi_t(x | x_0) p(x_0 | y) dx_0, \quad \phi_t(x | x_0) = \frac{1}{(2\pi\sigma_t^2)^{\frac{d_x}{2}}} \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right). \quad (\text{B.1})$$

Besides, the score function has the form of

$$\nabla_x \log p_t(x | y) = \int \nabla_x \log \phi_t(x | x_0) \frac{\phi_t(x | x_0) p(x_0 | y)}{\int \phi_t(x | z) p(z | y) dz} dx_0 \quad (\text{B.2})$$

$$= \frac{1}{\alpha_t} \int \nabla_x \log p(x_0 | y) \frac{\phi_t(x | x_0) p(x_0 | y)}{\int \phi_t(x | z) p(z | y) dz} dx_0. \quad (\text{B.3})$$

Proof. (B.1) can be directly implied by the definition of forward process. And it yields

$$\begin{aligned}\nabla_x \log p_t(x|y) &= \frac{\nabla_x p_t(x|y)}{p_t(x|y)} \\ &= \frac{\int \nabla_x \phi_t(x|x_0) p(x_0|y) dx_0}{\int \phi_t(x|x_0) p(x_0|y) dx_0} \\ &= \int \nabla_x \log \phi_t(x|x_0) \frac{\phi_t(x|x_0) p(x_0|y)}{\int \phi_t(x|z) p(z|y) dz} dx_0,\end{aligned}\tag{B.4}$$

which is (B.2). Moreover, noticing that $\nabla_x \phi_t(x|x_0) = -\frac{1}{\alpha_t} \nabla_{x_0} \phi_t(x|x_0)$, then by integration by parts,

$$\begin{aligned}\frac{\int \nabla_x \phi_t(x|x_0) p(x_0|y) dx_0}{\int \phi_t(x|x_0) p(x_0|y) dx_0} &= -\frac{1}{\alpha_t} \frac{\int \nabla_{x_0} \phi_t(x|x_0) p(x_0|y) dx_0}{\int \phi_t(x|x_0) p(x_0|y) dx_0} \\ &= \frac{1}{\alpha_t} \frac{\int \phi_t(x|x_0) \nabla_{x_0} p(x_0|y) dx_0}{\int \phi_t(x|x_0) p(x_0|y) dx_0} \\ &= \frac{1}{\alpha_t} \int \nabla_x \log p(x_0|y) \frac{\phi_t(x|x_0) p(x_0|y)}{\int \phi_t(x|z) p(z|y) dz} dx_0.\end{aligned}\tag{B.5}$$

Hence (B.3) is proved. $\square$

Lemma B.2. [Lem. 3.1] For any $w \in [0, 1]^{d_y}$, denote the conditional score of forward process $\nabla_x \log p_t(x; w)$ by $f_*(x, w, t)$. Then there exist constants C_X, C'_X , such that for any $R > 0$, the function $f_*(x, w, t)$ is $(C_X + C'_X R^2)$ -Lipschitz in x , $(C_X + C'_X R)$ -Lipschitz in w , in the domain $\mathcal{B}_R \times [0, 1]^{d_y} \times [0, T]$. Here $\mathcal{B}_R$ denotes the ball with radius R centered at the origin.

Proof. Define density function $q_t(x_0|x, w) \propto \phi_t(x|x_0) p(x_0; w)$. Our proof strategy will depend on whether $t \geq \frac{1}{2(L+1)}$.

When $t \geq \frac{1}{2(L+1)}$, according to (B.2), we have

$$\begin{aligned}\nabla_x f_*(x, w, t) &= \nabla_x^2 \log p_t(x; w) \\ &= \mathbb{E}_{q_t(x_0|x, w)} [\nabla_x^2 \log \phi_t(x|x_0)] + \text{Var}_{q_t(x_0|x, w)} (\nabla_x \log \phi_t(x|x_0)) \\ &= -\frac{I}{\sigma_t^2} + \text{Var}_{q_t(x_0|x, w)} \left(\frac{\alpha_t x_0 - x}{\sigma_t^2} \right)\end{aligned}\tag{B.6}$$

For any $R > 0$, we have

$$\begin{aligned}\text{Var}_{q_t(x_0|x, w)} \left(\frac{\alpha_t x_0 - x}{\sigma_t^2} \right) &\leq \frac{1}{\sigma_t^2} \int \left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^2 \frac{\phi_t(x|x_0) p(x_0|y)}{\int \phi_t(x|z) p(z|y) dz} dx_0 \\ &\leq \frac{R^2}{\sigma_t^2} + \frac{\int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \geq R} \left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^2 \exp \left(-\frac{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^2}{2\sigma_t^2} \right) p(x_0; w) dx_0}{\sigma_t^2 \int \exp \left(-\frac{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^2}{2\sigma_t^2} \right) p(x_0; w) dx_0} \\ &\leq \frac{R^2}{\sigma_t^2} + \frac{\int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \geq R} \exp \left(-\frac{R^2}{4} \right) p(x_0; w) dx_0}{\sigma_t^2 \int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \leq R/2} \exp \left(-\frac{R^2}{8} \right) p(x_0; w) dx_0}.\end{aligned}\tag{B.7}$$

Let $R = \frac{2\|x\| + 2C_0}{\sigma_t}$, then the domain $\left\{ x_0 : \left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \leq R/2 \right\}$ includes $\left\{ x_0 : \|x_0\| \leq C_0 \right\}$, indicating

$$\begin{aligned}\int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \leq R/2} p(x_0; w) dx_0 &\geq \int_{\|x_0\| \leq C_0} p(x_0; w) dx_0 \geq 1 - 2 \exp(-C'_1 C_0^2) \geq \frac{1}{2}, \\ \int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \geq R} p(x_0; w) dx_0 &\leq \int_{\|x_0\| \geq C_0} p(x_0; w) dx_0 \leq \frac{1}{2}.\end{aligned}\tag{B.8}$$

and

$$\|\nabla_x f_*(x, w, t)\| \leq \frac{1}{\sigma_t^2} + \left\| \text{Var}_{q_t(x_0|x, w)} \left(\frac{\alpha_t x_0 - x}{\sigma_t^2} \right) \right\| \leq \frac{R^2}{\sigma_t^2} + \frac{2}{\sigma_t^2} \leq \frac{8\|x\|^2 + 8C_0^2 + 2\sigma_t^2}{\sigma_t^4}. \quad (\text{B.9})$$

Similarly, for w we have

$$\begin{aligned} \nabla_w f_*(x, w, t) &= \text{Cov}_{q_t(x_0|x, w)}(\nabla_x \log \phi_t(x|x_0), \nabla_w \log p(x_0; w)) \\ &= \text{Cov}_{q_t(x_0|x, w)}\left(\frac{\alpha_t x_0}{\sigma_t^2}, \nabla_w \log p(x_0; w)\right) \end{aligned} \quad (\text{B.10})$$

which implies

$$\begin{aligned} \|\nabla_w f_*(x, w, t)\| &\leq B \sqrt{\left\| \text{Var}_{q_t(x_0|x, w)} \left(\frac{\alpha_t x_0 - x}{\sigma_t^2} \right) \right\|} \\ &\leq \frac{B(2\|x\| + 2C_0 + 1)}{\sigma_t} \end{aligned} \quad (\text{B.11})$$

When $t \leq \frac{1}{2(L+1)}$, we have $\sigma_t^2 \leq \frac{\alpha_t^2}{2L}$ and

$$\begin{aligned} \nabla_x f_*(x, w, t) &= \nabla_x^2 \log p_t(x; w) \\ &= \frac{\nabla_x^2 p_t(x; w)}{p_t(x; w)} - \nabla_x \log p_t(x; w) (\nabla_x \log p_t(x; w))^\top \\ &= \frac{1}{\alpha_t^2} \frac{\int \phi_t(x|x_0) \nabla_x^2 p(x_0; w) dx_0}{p_t(x; w)} - \nabla_x \log p_t(x; w) (\nabla_x \log p_t(x; w))^\top \\ &= \frac{1}{\alpha_t^2} \mathbb{E}_{q_t(x_0|x, w)} \left[\frac{\nabla_x^2 p(x_0; w)}{p(x_0; w)} \right] - \nabla_x \log p_t(x; w) (\nabla_x \log p_t(x; w))^\top \\ &= \frac{1}{\alpha_t^2} \mathbb{E}_{q_t(x_0|x, w)} [\nabla_x^2 \log p(x_0; w) + \nabla_x \log p(x_0; w) (\nabla_x \log p(x_0; w))^\top] \\ &\quad - \nabla_x \log p_t(x; w) (\nabla_x \log p_t(x; w))^\top \\ &\stackrel{(\text{B.3})}{=} \frac{1}{\alpha_t^2} \mathbb{E}_{q_t(x_0|x, w)} [\nabla_x^2 \log p(x_0; w)] + \frac{1}{\alpha_t^2} \text{Var}_{q_t(x_0|x, w)}(\nabla_x \log p(x_0; w)). \end{aligned} \quad (\text{B.12})$$

Note that when $\sigma_t^2 \leq \frac{\alpha_t^2}{2L}$, the distribution $q_t(x_0|x, w) \propto \exp(-\frac{\|\alpha_t x_0 - x\|^2}{2\sigma_t^2})p(x_0; w)$ is L -strongly log-concave, and thus satisfies the Poincare inequality with a constant L^{-1} [Chen et al., 2023a],

$$\text{Var}_{q_t(x_0|x, w)}(\nabla_x \log p(x_0; w)) \preceq L^{-1} \mathbb{E} [\nabla_x^2 \log p(x_0; w) (\nabla_x \log p(x_0; w))^\top] \preceq L. \quad (\text{B.13})$$

And thus

$$\|\nabla_x f_*(x, w, t)\| \leq \frac{2L}{\alpha_t^2}. \quad (\text{B.14})$$

Analogously,

$$\begin{aligned} \nabla_w f_*(x, w, t) &= \frac{1}{\alpha_t} \mathbb{E}_{q_t(x_0|x, w)} [\nabla_w \nabla_x \log p(x_0; w)] + \frac{1}{\alpha_t} \text{Cov}_{q_t(x_0|x, w)}(\nabla_x \log p(x_0; w), \nabla_w \log p(x_0; w)) \\ &\leq \frac{L}{\alpha_t} + \frac{B}{\alpha_t} \sqrt{\text{Var}_{q_t(x_0|x, w)}(\nabla_x \log p(x_0; w))} \\ &\leq \frac{L + B\sqrt{L}}{\alpha_t} \end{aligned} \quad (\text{B.15})$$

Combine all the arguments in (B.9), (B.11), (B.14), (B.15) and we complete the proof. $\square$

Lemma B.3 (Lemma 7, Chen et al. [2022a]). *The covering number of $\mathcal{F} = NN_f(L_f, W_f, M_f, S_f, B_f, R_f, \gamma_f)$ can be bounded by*

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty([-R, R]^{d_x+d_y+1})}, \varepsilon) \lesssim S_f L_f \log \left(\frac{L_f W_f (B_f \vee 1) R}{\varepsilon} \right). \quad (\text{B.16})$$

The covering number of $\mathcal{H} = NN_h(L_h, W_h, S_h, B_h)$ can be bounded by

$$\log \mathcal{N}(\mathcal{H}, \|\cdot\|_{L^\infty([0,1]^{D_y}), \epsilon}) \lesssim S_h L_h \log \left(\frac{L_h W_h (B_h \vee 1)}{\epsilon} \right). \quad (\text{B.17})$$

B.2 Proofs of Transfer Learning

Proposition B.4 (Prop. 3.2). *Under Assumption 3.1, 3.2, 3.3, there exists some constant C_{xy} such that the following holds. For any $h \in \mathcal{H}$ and $(x_1, y_1), \dots, (x_m, y_m) \stackrel{i.i.d.}{\sim} \mathbb{P}$, define the empirical minimizer*

$$\hat{f} := \arg \min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(x_i, y_i, s_{f,h}). \quad (\text{B.18})$$

The population loss of $\hat{f}$ can be bounded by

$$\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{\hat{f},h})] \leq 4 \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})] + C_{xy} \log^3(m) r_x, \quad (\text{B.19})$$

$$\text{where } r_x = \frac{M_f^2 S_f L_f \log(m L_f W_f (B_f \vee 1) M_f T)}{m}.$$

Proof. Consider the truncated function class defined on $\mathbb{R}^{d_x} \times [0, 1]^{D_y}$,

$$\Phi = \{(x, y) \mapsto \tilde{\ell}(x, y, f) := (\ell(x, y, s_{f,h}) - \ell(x, y, s_*^{\mathbb{P}})) \cdot \mathbb{1}_{\|x\|_\infty \leq R} : f \in \mathcal{F}\}, \quad (\text{B.20})$$

where the truncation radius $R \geq 1$ will be defined later. It is easy to show that with probability no less than $1 - 2m \exp(-C'_1 R^2)$, it holds that $\|x_i\|_\infty \leq R$ for all $1 \leq i \leq m$. Hence by definition, the empirical minimizer also satisfies $\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \tilde{\ell}(x_i, y_i, f)$. Below we reason conditioned on this event and verify the conditions required in Lemma B.11.

Step 1. To bound the individual loss,

$$\tilde{\ell}(x, y, f) \leq \mathbb{E}_{t, x_t | x} \|s_{f,h}(x_t, y, t) - \nabla_x \log \phi_t(x_t | x)\|^2 \lesssim M_f^2 + d_x \left(\frac{\log(1/T_0)}{T - T_0} + 1 \right). \quad (\text{B.21})$$

And by Lemma B.10,

$$-\tilde{\ell}(x, y, f) \leq \mathbb{E}_{t, x_t | x} \|s_*^{\mathbb{P}}(x_t, y, t) - \nabla_x \log \phi_t(x_t | x)\|^2 \cdot \mathbb{1}_{\|x\|_\infty \leq R} \lesssim C_X'' R^6 + d_x \left(\frac{\log(1/T_0)}{T} + 1 \right). \quad (\text{B.22})$$

Let $M := C \left(C_X'' R^6 + M_f^2 + d_x \left(\frac{\log(1/T_0)}{T} + 1 \right) \right)$ and thus $|\tilde{\ell}(x, y, f)| \leq M$.

Step 2. To bound the second order moment, we have

$$\begin{aligned} & \mathbb{E}_{(x, y) \sim \mathbb{P}} \left[\mathbb{1}_{\|x\|_\infty \leq R} (\ell(x, y, s_{f,h}) - \ell(x, y, s_*^{\mathbb{P}}))^2 \right] \\ &= \mathbb{E}_{(x, y) \sim \mathbb{P}} \left[\mathbb{1}_{\|x\|_\infty \leq R} (\mathbb{E}_{t, x_t | x} \|s_{f,h}(x_t, y, t) - \nabla_x \log \phi_t(x_t | x)\|^2 - \|s_*^{\mathbb{P}}(x_t, y, t) - \nabla_x \log \phi_t(x_t | x)\|^2)^2 \right] \\ &\leq \mathbb{E}_{(x, y) \sim \mathbb{P}} \left[\mathbb{1}_{\|x\|_\infty \leq R} (\mathbb{E}_{t, x_t | x} \|s_{f,h}(x_t, y, t) - s_*^{\mathbb{P}}(x_t, y, t)\|^2) \right. \\ &\quad \cdot (\mathbb{E}_{t, x_t | x} \|s_{f,h}(x_t, y, t) + s_*^{\mathbb{P}}(x_t, y, t) - 2\nabla_x \log \phi_t(x_t | x)\|^2) \left. \right] \\ &\leq 4M \mathbb{E}_{(x, y) \sim \mathbb{P}} \left[\mathbb{1}_{\|x\|_\infty \leq R} (\mathbb{E}_{t, x_t | x} \|s_{f,h}(x_t, y, t) - s_*^{\mathbb{P}}(x_t, y, t)\|^2) \right] \\ &\leq 4M \mathbb{E}_{(x, y) \sim \mathbb{P}} (\ell(x, y, s_{f,h}) - \ell(x, y, s_*^{\mathbb{P}})) \\ &\leq 4M \mathbb{E}_{(x, y) \sim \mathbb{P}} [\tilde{\ell}(x, y, f)] + 8M^2 \exp(-C'_1 R^2). \end{aligned} \quad (\text{B.23})$$

Step 3. To bound the local Rademacher complexity, note that

$$\left\| \frac{1}{\sqrt{m}} \sum_{i=1}^m \sigma_i \tilde{\ell}(x_i, y_i, f_1) - \frac{1}{\sqrt{m}} \sum_{i=1}^m \sigma_i \tilde{\ell}(x_i, y_i, f_2) \right\|_{\psi_2} \leq 4 \|\tilde{\ell}(\cdot, \cdot, f_1) - \tilde{\ell}(\cdot, \cdot, f_2)\|_{L^2(\widehat{\mathbb{P}}_m)}, \quad (\text{B.24})$$

where $\widehat{\mathbb{P}}_m := \frac{1}{m} \sum_{i=1}^m \delta_{(x_i, y_i)}$. Define $\Phi_r := \{\varphi \in \Phi : \frac{1}{m} \sum_{i=1}^m \varphi(x_i, y_i)^2 \leq r\}$ and it is easy to show that $\text{diam}(\Phi_r, \|\cdot\|_{L^2(\widehat{\mathbb{P}}_m)}) \leq 2\sqrt{r}$. By Dudley's bound [Van Handel, 2014, Wainwright, 2019], there exists an absolute constant C_0 such that for any $\theta > 0$,

$$\mathcal{R}_m(\Phi_r) \leq C_0 \left(\theta + \int_{\theta}^{2\sqrt{r}} \sqrt{\frac{\log \mathcal{N}(\Phi_r, \|\cdot\|_{L^2(\widehat{\mathbb{P}}_m)}, \varepsilon)}{m}} d\varepsilon \right). \quad (\text{B.25})$$

Since $\|x_i\| \leq R$,

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^m (\tilde{\ell}(x_i, y_i, f_1) - \tilde{\ell}(x_i, y_i, f_2))^2 &= \frac{1}{m} \sum_{i=1}^m (\ell(x_i, y_i, s_{f_1, h}) - \ell(x_i, y_i, s_{f_2, h}))^2 \\ &\leq \frac{1}{m} \sum_{i=1}^m [\mathbb{E}_{t, x_t | x_i} \|f_1 - f_2\|^2] \cdot [\mathbb{E}_{t, x_t | x_i} \|f_1 + f_2 - 2\nabla_x \log \phi_t\|^2] \\ &\leq \frac{4M}{m} \sum_{i=1}^m \mathbb{E}_{t, x_t | x_i} \|f_1(x_t, h(y_i), t) - f_2(x_t, h(y_i), t)\|^2. \end{aligned} \quad (\text{B.26})$$

Let $R_1 = 2R$. Since $x_t | x_i \sim \mathcal{N}(x_t; \alpha_t x_i, \sigma_t^2 I)$, we have $\mathbb{P}(\|x_t\|_{\infty} \geq R_1) \leq d_x \mathbb{P}(|\mathcal{N}(0, 1)| \leq R) \leq 2d_x \exp(-C'_0 R^2)$ for some absolute constant C'_0 . Therefore,

$$\begin{aligned} \mathbb{E}_{t, x_t | x_i} \|f_1(x_t, h(y_i), t) - f_2(x_t, h(y_i), t)\|^2 &\leq \mathbb{E}_{t, x_t | x_i} [\mathbb{1}_{\|x_t\| \leq R_1} \|\ell(x_t, h(y_i), t) - \ell(x_t, h(y_i), t)\|^2] + 8d_x M_f^2 \exp(-C'_0 R^2) \\ &\leq \|f_1 - f_2\|_{L^\infty(\Omega_{R_1})}^2 + 8d_x M_f^2 \exp(-C'_0 R^2) \end{aligned} \quad (\text{B.27})$$

where $\Omega_{R_1} := [-R_1, R_1]^{d_x} \times [0, 1]^{d_y} \times [T_0, T]$. Plug in the bound above,

$$\sqrt{\frac{1}{m} \sum_{i=1}^m (\tilde{\ell}(x_i, y_i, f_1) - \tilde{\ell}(x_i, y_i, f_2))^2} \leq 4M^{\frac{1}{2}} \|f_1 - f_2\|_{L^\infty(\Omega_{R_1})} + 8d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2). \quad (\text{B.28})$$

For any $\varepsilon \geq 16d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2)$, according to B.3,

$$\begin{aligned} \log \mathcal{N}(\Phi_r, \|\cdot\|_{L^2(\widehat{\mathbb{P}}_m)}, \varepsilon) &\leq \log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty(\Omega_{R_1})}, \varepsilon/(8M^{\frac{1}{2}})) \\ &\leq C_4 S_f L_f \log \left(\frac{L_f W_f (B_f \vee 1)(R \vee T) M}{\varepsilon} \right). \end{aligned} \quad (\text{B.29})$$

Plug in (B.25) and let $\theta = 16d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2)$,

$$\begin{aligned} \mathcal{R}_m(\Phi_r) &\leq C_0 \left(\theta + \int_{\theta}^{2\sqrt{r}} \sqrt{\frac{C_4 S_f L_f \log \left(\frac{L_f W_f (B_f \vee 1)(R \vee T) M}{\varepsilon} \right)}{m}} d\varepsilon \right) \\ &\leq C_0 \left(16d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2) + \sqrt{\frac{C'_4 S_f L_f \log \left(\frac{L_f W_f (B_f \vee 1)(R \vee T) M}{r} \right) \cdot r}{m}} \right) \\ &=: \tilde{\mathcal{R}}_m(r) \end{aligned} \quad (\text{B.30})$$

Combine the three steps above, by Lemma B.11 with $B_0 = 8M^2 \exp(-C'_1 R^2)$, $B = 4M$, $b = M$, it holds that with probability no less than $1 - 2m \exp(-C'_1 R^2) - \delta/2$, for any $f \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathbb{P}}[\tilde{\ell}(x, y, f)] &\leq \frac{2}{m} \sum_{i=1}^m \tilde{\ell}(x_i, y_i, f) + C_5 M \left(r_m^* + \frac{\log(\log(m)/\delta)}{m} \right) \\ &\quad + C_5 \sqrt{\frac{M^2 \log(\log(m)/\delta)}{m}} \exp(-C'_1 R^2), \end{aligned} \quad (\text{B.31})$$

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^m \tilde{\ell}(x_i, y_i, f) &\leq 2\mathbb{E}_{(x,y) \sim \mathbb{P}}[\tilde{\ell}(x, y, f)] + C_5 M \left(r_m^* + \frac{\log(\log(m)/\delta)}{m} \right) \\ &\quad + C_5 \sqrt{\frac{M^2 \log(\log(m)/\delta)}{m}} \exp(-C'_1 R^2). \end{aligned} \quad (\text{B.32})$$

where r_m^* is the largest fixed point of $\tilde{\mathcal{R}}_m$, and it can be bounded as

$$r_m^* \leq C_6 \left(d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2) + \frac{S_f L_f \log(m L_f W_f (B_f \vee 1)(R \vee T)M)}{m} \right), \quad (\text{B.33})$$

for some absolute constant C_6 . Moreover, we have

$$\left| \mathbb{E}_{(x,y) \sim \mathbb{P}}[\ell(x, y, s_{f,h}) - \ell(x, y, s_*^{\mathbb{P}})] - \mathbb{E}_{(x,y) \sim \mathbb{P}}[\tilde{\ell}(x, y, f)] \right| \leq 2M \exp(-C'_1 R^2). \quad (\text{B.34})$$

Combine this with (B.31), (B.32),

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathbb{P}}[\ell(x, y, s_{f,h}) - \ell(x, y, s_*^{\mathbb{P}})] &\leq \frac{2}{m} \sum_{i=1}^m [\ell(x_i, y_i, s_{f,h}) - \ell(x_i, y_i, s_*^{\mathbb{P}})] \\ &\quad + C_5 M \left(r_m^* + \frac{\log(\log(m)/\delta)}{m} + \exp(-C'_1 R^2) \right), \end{aligned} \quad (\text{B.35})$$

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^m [\ell(x_i, y_i, s_{f,h}) - \ell(x_i, y_i, s_*^{\mathbb{P}})] &\leq 2\mathbb{E}_{(x,y) \sim \mathbb{P}}[\ell(x, y, s_{f,h}) - \ell(x, y, s_*^{\mathbb{P}})] \\ &\quad + C_5 M \left(r_m^* + \frac{\log(\log(m)/\delta)}{m} + \exp(-C'_1 R^2) \right), \end{aligned} \quad (\text{B.36})$$

Plug in the definition of $M = C \left(C_X'' R^6 + M_f^2 + d_x \left(\frac{\log(1/T_0)}{T} + 1 \right) \right)$ and let $R = C \log^{\frac{1}{2}}(m d_x M_f / \delta)$ for some large constant C . Hence (B.35) and (B.36) reduce to

$$\mathbb{E}_{(x,y) \sim \mathbb{P}}[\ell^{\mathbb{P}}(x, y, s_{f,h})] \leq \frac{2}{m} \sum_{i=1}^m [\ell(x_i, y_i, s_{f,h}) - \ell(x_i, y_i, s_*^{\mathbb{P}})] + C_7 M_f^2 \log^3(m/\delta) \left(r_m^{\dagger} + \frac{\log(\log(m)/\delta)}{m} \right), \quad (\text{B.37})$$

$$\frac{1}{m} \sum_{i=1}^m [\ell(x_i, y_i, s_{f,h}) - \ell(x_i, y_i, s_*^{\mathbb{P}})] \leq 2\mathbb{E}_{(x,y) \sim \mathbb{P}}[\ell^{\mathbb{P}}(x, y, s_{f,h})] + C_7 M_f^2 \log^3(m/\delta) \left(r_m^{\dagger} + \frac{\log(\log(m)/\delta)}{m} \right), \quad (\text{B.38})$$

where $r_m^{\dagger} := \frac{S_f L_f \log(m L_f W_f (B_f \vee 1) T M_f \log(1/\delta))}{m}$.

Therefore, we obtain that with probability no less than $1 - \delta$, the population loss of the empirical minimizer $\hat{f}$ can be bounded by

$$\begin{aligned}\mathbb{E}_{(x,y) \sim \mathbb{P}}[\ell^{\mathbb{P}}(x, y, s_{\hat{f},h})] &\leq \frac{2}{m} \sum_{i=1}^m [\ell(x_i, y_i, s_{\hat{f},h}) - \ell(x_i, y_i, s_{*}^{\mathbb{P}})] + 2C_7 M_f^2 \log^3(m/\delta) \left(r_m^{\dagger} + \frac{\log(1/\delta)}{m} \right) \\ &\leq \inf_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^m [\ell(x_i, y_i, s_{f,h}) - \ell(x_i, y_i, s_{*}^{\mathbb{P}})] + 2C_7 M_f^2 \log^3(m/\delta) \left(r_m^{\dagger} + \frac{\log(1/\delta)}{m} \right) \\ &\leq 4 \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}}[\ell^{\mathbb{P}}(x, y, s_{f,h})] + 6C_7 M_f^2 \log^3(m/\delta) \left(r_m^{\dagger} + \frac{\log(1/\delta)}{m} \right),\end{aligned}\tag{B.39}$$

We conclude the proof by noticing that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x) dx$ and plugging in the bound above. $\square$

Proposition B.5 (Prop. 3.3). *There exists some constant C_Z, C_R such that the following holds. For any $\mathbb{P}^1, \dots, \mathbb{P}^K$, let $x_1^k, \dots, x_n^k \stackrel{i.i.d.}{\sim} \mathbb{P}^k$ for any k and $(x_i^k)_{i,k}$ are all independent. Consider the empirical minimizer*

$$\hat{\mathbf{f}}, \hat{h} = \arg \min_{\mathbf{f} \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \ell(x_i^k, y_i^k, s_{\mathbf{f}^k, h}).\tag{B.40}$$

For any $\delta \in (0, 1)$, if the configuration of $\mathcal{F}$ satisfies $R_f \geq C_R \log^{\frac{1}{2}}(nK M_f / \delta)$, then with probability no less than $1 - \delta$, the population loss of $\hat{\mathbf{f}}, \hat{h}$ can be bounded by

$$\begin{aligned}\frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim \mathbb{P}^k} \ell^{\mathbb{P}^k}(x, y, s_{\hat{\mathbf{f}}^k, \hat{h}}) &\leq \inf_{\mathbf{f} \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{4}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{\mathbf{f}^k, h})] + C_Z \log^3(nK/\delta) \left(r_z + \frac{\log(1/\delta)}{nK} \right), \\ \text{where } r_z &:= \frac{M_f^2 [K S_f L_f \log(n L_f W_f (B_f \vee 1) M_f T \log(1/\delta)) + S_h L_h \log(n K L_h W_h (B_h \vee 1) M_f \gamma_f \log(1/\delta))]}{nK}.\end{aligned}\tag{B.41}$$

Proof. Throughout the proof, we will use $z = (k, x, y)$ to denote the tuple of task index k and data (x, y) . With a little abuse of notation, we will also let $s_*^k = s_*^{\mathbb{P}^k}$. Consider the function class defined on $[K] \times \mathbb{R}^{d_x} \times [0, 1]^{D_y}$,

$$\Phi = \left\{ z = (k, x, y) \mapsto \tilde{\ell}(z, \mathbf{f}, h) := (\ell(x, y, s_{\mathbf{f}^k, h}) - \ell(x, y, s_*^k)) \cdot \mathbb{1}_{\|x\|_\infty \leq R} : \mathbf{f} \in \mathcal{F}^{\otimes K}, h \in \mathcal{H} \right\},\tag{B.42}$$

where $1 \leq R \leq \frac{R_f}{2}$ will be specified later. It is easy to show that with probability no less than $1 - 2nK \exp(-C'_1 R^2)$, it holds that $\|x_i^k\|_\infty \leq R$ for all i, k . Hence by definition, the empirical minimizer also satisfies

$$\hat{\mathbf{f}}, \hat{h} = \arg \min_{\mathbf{f} \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \tilde{\ell}(z_i^k, \mathbf{f}, h).\tag{B.43}$$

where $z_i^k = (k, x_i^k, y_i^k)$. Below we reason conditioned on this event and verify the conditions in Lemma B.12.

Following Step 1 and 2 in Proposition B.4, we have for any $\mathbf{f} \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}$,

$$|\tilde{\ell}(z, \mathbf{f}, h)| \leq M := C \left(C_X'' R^6 + M_f^2 + d_x \left(\frac{\log(1/T_0)}{T} + 1 \right) \right).\tag{B.44}$$

$$\frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\tilde{\ell}(z^k, \mathbf{f}, h)^2] \leq \frac{4M}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\tilde{\ell}(z^k, \mathbf{f}, h)] + 8M^2 \exp(-C'_1 R^2).\tag{B.45}$$

For the local Rademacher complexity bound, note that

$$\left\| \frac{1}{\sqrt{nK}} \sum_{k=1}^K \sum_{i=1}^n \sigma_i^k \tilde{\ell}(z_i^k, \mathbf{f}_1, h_1) - \frac{1}{\sqrt{nK}} \sum_{k=1}^K \sum_{i=1}^n \sigma_i^k \tilde{\ell}(z_i^k, \mathbf{f}_2, h_2) \right\|_{\psi_2} \leq 4 \|\tilde{\ell}(\cdot, \mathbf{f}_1, h_1) - \tilde{\ell}(\cdot, \mathbf{f}_2, h_2)\|_{L^2(\widehat{\mathbb{P}}_n^{(K)})}, \quad (\text{B.46})$$

where $\widehat{\mathbb{P}}_n^{(K)} := \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \delta_{z_i^k}$ and $\mathbf{diam}(\Phi_r, \|\cdot\|_{L^2(\widehat{\mathbb{P}}_n^{(K)})}) \leq 2\sqrt{r}$. By Dudley's bound [Van Handel, 2014, Wainwright, 2019], there exists an absolute constant C_0 such that for any $\theta > 0$,

$$\mathcal{R}_{K,n}(\Phi_r) \leq C_0 \left(\theta + \int_{\theta}^{2\sqrt{r}} \sqrt{\frac{\log \mathcal{N}(\Phi_r, \|\cdot\|_{L^2(\widehat{\mathbb{P}}_n^{(K)})}, \varepsilon)}{nK}} d\varepsilon \right). \quad (\text{B.47})$$

Since $\|x_i^k\|_{\infty} \leq R$,

$$\begin{aligned} & \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (\tilde{\ell}(z_i^k, \mathbf{f}_1, h_1) - \tilde{\ell}(z_i^k, \mathbf{f}_2, h_2))^2 \\ &= \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (\ell(x_i^k, y_i^k, s_{f_1^k, h_1}) - \ell(x_i^k, y_i^k, s_{f_2^k, h_2}))^2 \\ &\leq \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \left[\mathbb{E}_{t, x_t | x_i^k} \|f_1^k - f_2^k\|^2 \right] \cdot \left[\mathbb{E}_{t, x_t | x_i^k} \|f_1^k + f_2^k - 2\nabla_x \log \phi_t\|^2 \right] \\ &\leq \frac{4M}{nK} \sum_{k=1}^K \sum_{i=1}^n \mathbb{E}_{t, x_t | x_i^k} \|f_1^k(x_t, h_1(y_i^k), t) - f_2^k(x_t, h_2(y_i^k), t)\|^2 \\ &\leq \frac{8M}{nK} \sum_{k=1}^K \sum_{i=1}^n \mathbb{E}_{t, x_t | x_i^k} \|f_1^k(x_t, h_1(y_i^k), t) - f_2^k(x_t, h_1(y_i^k), t)\|^2 \\ &\quad + \frac{8M}{nK} \sum_{k=1}^K \sum_{i=1}^n \mathbb{E}_{t, x_t | x_i^k} \|f_2^k(x_t, h_1(y_i^k), t) - f_2^k(x_t, h_2(y_i^k), t)\|^2. \end{aligned} \quad (\text{B.48})$$

Let $R_1 = 2R$. Since $x_t | x_i^k \sim \mathcal{N}(x_t; \alpha_t x_i^k, \sigma_t^2 I)$, we have $\mathbb{P}(\|x_t\|_{\infty} \geq R_1) \leq d_x \mathbb{P}(|\mathcal{N}(0, 1)| \leq R) \leq 2d_x \exp(-C'_0 R^2)$ for some absolute constant C'_0 . Therefore,

$$\begin{aligned} & \mathbb{E}_{t, x_t | x_i^k} \|f_1^k(x_t, h_1(y_i^k), t) - f_2^k(x_t, h_1(y_i^k), t)\|^2 \\ &\leq \mathbb{E}_{t, x_t | x_i^k} [\mathbb{1}_{\|x_t\| \leq R_1}] [\|f_1^k(x_t, h_1(y_i^k), t) - f_2^k(x_t, h_1(y_i^k), t)\|^2] + 8d_x M_f^2 \exp(-C'_0 R^2) \\ &\leq \|f_1^k - f_2^k\|_{L^\infty(\Omega_{R_1})}^2 + 8d_x M_f^2 \exp(-C'_0 R^2), \end{aligned} \quad (\text{B.49})$$

where $\Omega_{R_1} := [-R_1, R_1]^{d_x} \times [0, 1]^{d_y} \times [T_0, T]$. Moreover, notice that $R_f \geq 2R = R_1$,

$$\begin{aligned} & \mathbb{E}_{t, x_t | x_i^k} \|f_2^k(x_t, h_1(y_i^k), t) - f_2^k(x_t, h_2(y_i^k), t)\|^2 \\ &\leq \mathbb{E}_{t, x_t | x_i^k} [\mathbb{1}_{\|x_t\| \leq R_f}] [\|f_2^k(x_t, h_1(y_i^k), t) - f_2^k(x_t, h_2(y_i^k), t)\|^2] + 8d_x M_f^2 \exp(-C'_0 R^2) \\ &\leq \gamma_f^2 \|h_1 - h_2\|_{L^\infty([0, 1]^{d_y})}^2 + 8d_x M_f^2 \exp(-C'_0 R^2). \end{aligned} \quad (\text{B.50})$$

Plug in the bound above,

$$\begin{aligned} & \sqrt{\frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (\tilde{\ell}(z_i^k, \mathbf{f}_1, h_1) - \tilde{\ell}(z_i^k, \mathbf{f}_2, h_2))^2} \\ &\leq 8M^{\frac{1}{2}} \left(\max_k \|f_1^k - f_2^k\|_{L^\infty(\Omega_{R_1})} + \gamma_f \|h_1 - h_2\|_{L^\infty([0, 1]^{d_y})} \right) + 16d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2). \end{aligned} \quad (\text{B.51})$$

For any $\varepsilon \geq 32d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2)$, according to Lemma B.3,

$$\begin{aligned}
& \log \mathcal{N}(\Phi_r, \|\cdot\|_{L^2(\widehat{\mathbb{P}}_n^{(K)})}, \varepsilon) \\
& \leq K \log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty(\Omega_{R_1})}, \varepsilon/(16M^{\frac{1}{2}})) + \log \mathcal{N}(\mathcal{H}, \|\cdot\|_{L^\infty([0,1]^{D_y})}, \varepsilon/(16\gamma_f M^{\frac{1}{2}})) \\
& \leq C_4 K S_f L_f \log \left(\frac{L_f W_f (B_f \vee 1)(R \vee T) M}{\varepsilon} \right) + C_4 S_h L_h \log \left(\frac{L_h W_h (B_h \vee 1) M \gamma_f}{\varepsilon} \right).
\end{aligned} \tag{B.52}$$

Plug in (B.25) and let $\theta = 32d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2)$,

$$\begin{aligned}
\mathcal{R}_{K,n}(\Phi_r) & \leq C_0 \left(\theta + \int_\theta^{2\sqrt{r}} \sqrt{\frac{C_4 K S_f L_f \log \left(\frac{L_f W_f (B_f \vee 1)(R \vee T) M}{\varepsilon} \right) + C_4 S_h L_h \log \left(\frac{L_h W_h (B_h \vee 1) M \gamma_f}{\varepsilon} \right)}{nK}} d\varepsilon \right) \\
& \leq C_0 \sqrt{\frac{C_4 \left[K S_f L_f \log \left(\frac{L_f W_f (B_f \vee 1)(R \vee T) M}{r} \right) + S_h L_h \log \left(\frac{L_h W_h (B_h \vee 1) M \gamma_f}{r} \right) \right]}{nK}} \\
& \quad + C_0 32d_x^{\frac{1}{2}} M \exp(-C'_0 R^2/2) \\
& =: \widetilde{\mathcal{R}}_{K,n}(r).
\end{aligned} \tag{B.53}$$

Combine the arguments above, by Lemma B.12 with $B_0 = 8M^2 \exp(-C'_1 R^2)$, $B = 4M$, $b = M$, it holds that with probability no less than $1 - 2nK \exp(-C'_1 R^2) - \delta/2$, for any $\mathbf{f} \in \mathcal{F}^{\otimes K}$, $h \in \mathcal{H}$,

$$\begin{aligned}
\mathbb{E}_{z \sim \widehat{\mathbb{P}}^{(K)}} [\widetilde{\ell}(z, \mathbf{f}, h)] & \leq \frac{2}{nK} \sum_{k=1}^K \sum_{i=1}^n \widetilde{\ell}(z_i^k, \mathbf{f}, h) + C_5 M \left(r_{K,n}^* + \frac{\log(\log(nK)/\delta)}{nK} \right) \\
& \quad + C_5 \sqrt{\frac{M^2 \log(\log(nK)/\delta)}{nK}} \exp(-C'_1 R^2),
\end{aligned} \tag{B.54}$$

$$\begin{aligned}
\frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \widetilde{\ell}(z_i^k, \mathbf{f}, h) & \leq 2\mathbb{E}_{z \sim \widehat{\mathbb{P}}^{(K)}} [\widetilde{\ell}(z, \mathbf{f}, h)] + C_5 M \left(r_{K,n}^* + \frac{\log(\log(nK)/\delta)}{nK} \right) \\
& \quad + C_5 \sqrt{\frac{M^2 \log(\log(nK)/\delta)}{nK}} \exp(-C'_1 R^2).
\end{aligned} \tag{B.55}$$

where $r_{K,n}^*$ is the largest fixed point of $\widetilde{\mathcal{R}}_{K,n}$, and it can be bounded by

$$r_{K,n}^* \leq C_6 \left(d_x^{\frac{1}{2}} M_f \exp(-C'_0 R^2/2) + \frac{K S_f L_f \log(n L_f W_f (B_f \vee 1)(R \vee T) M) + S_h L_h \log(n K L_h W_h (B_h \vee 1) M \gamma_f)}{nK} \right), \tag{B.56}$$

for some absolute constant C_6 . Moreover, we have

$$\left| \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\ell(x, y, s_{f^k, h}) - \ell(x, y, s_*^k)] - \mathbb{E}_{z \sim \widehat{\mathbb{P}}^{(K)}} [\widetilde{\ell}(z, \mathbf{f}, h)] \right| \leq 2M \exp(-C'_1 R^2). \tag{B.57}$$

Combine this with (B.54), (B.55),

$$\begin{aligned}
\frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\ell(x, y, s_{f^k, h}) - \ell(x, y, s_*^k)] & \leq \frac{2}{nK} \sum_{k=1}^K \sum_{i=1}^n [\ell(x_i^k, y_i^k, s_{f^k, h}) - \ell(x_i^k, y_i^k, s_*^k)] \\
& \quad + C_5 M \left(r_{K,n}^* + \frac{\log(\log(nK)/\delta)}{nK} + \exp(-C'_1 R^2) \right),
\end{aligned} \tag{B.58}$$

$$\begin{aligned} \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n [\ell(x_i^k, y_i^k, s_{f^k, h}) - \ell(x_i^k, y_i^k, s_*^k)] &\leq \frac{2}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell(x, y, s_{f^k, h}) - \ell(x, y, s_*^{\mathbb{P}})] \\ &\quad + C_5 M \left(r_{K, n}^* + \frac{\log(\log(nK)/\delta)}{nK} + \exp(-C'_1 R^2) \right), \end{aligned} \quad (\text{B.59})$$

Plug in the definition of $M = C \left(C''_X R^6 + M_f^2 + d_x \left(\frac{\log(1/T_0)}{T} + 1 \right) \right)$ and define $R = C' \log^{\frac{1}{2}}(nK d_x M_f / \delta)$ for some large constant C' . Hence (B.58) and (B.59) reduce to

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell(x, y, s_{f^k, h}) - \ell(x, y, s_*^{\mathbb{P}})] &\leq \frac{2}{nK} \sum_{k=1}^K \sum_{i=1}^n [\ell(x_i^k, y_i^k, s_{f^k, h}) - \ell(x_i^k, y_i^k, s_*^k)] \\ &\quad + C_7 M_f^2 \log^3(nK/\delta) \left(r_{K, n}^\dagger + \frac{\log(\log(nK)/\delta)}{nK} \right), \end{aligned} \quad (\text{B.60})$$

$$\begin{aligned} \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n [\ell(x_i^k, y_i^k, s_{f^k, h}) - \ell(x_i^k, y_i^k, s_*^k)] &\leq \frac{2}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell(x, y, s_{f^k, h}) - \ell(x, y, s_*^{\mathbb{P}})] \\ &\quad + C_7 M_f^2 \log^3(nK/\delta) \left(r_{K, n}^\dagger + \frac{\log(\log(nK)/\delta)}{nK} \right), \end{aligned} \quad (\text{B.61})$$

$$\text{where } r_{K, n}^\dagger := \frac{K S_f L_f \log(n L_f W_f (B_f \vee 1) M_f T \log(1/\delta)) + S_h L_h \log(n K L_h W_h (B_h \vee 1) M_f \gamma_f \log(1/\delta))}{nK}.$$

Therefore, we obtain that with probability no less than $1 - \delta$, the population loss of the empirical minimizer $\hat{f}, \hat{h}$ can be bounded by

$$\begin{aligned} &\frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{\hat{f}^k, \hat{h}})] \\ &\leq \frac{2}{nK} \sum_{k=1}^K \sum_{i=1}^m [\ell(x_i^k, y_i^k, s_{\hat{f}, h}) - \ell(x_i^k, y_i^k, s_*^k)] + 2C_7 M_f^2 \log^3(nK/\delta) \left(r_{K, n}^\dagger + \frac{\log(1/\delta)}{nK} \right) \\ &\leq \inf_{f \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{2}{nK} \sum_{k=1}^K \sum_{i=1}^n [\ell(x_i^k, y_i^k, s_{f^k, h}) - \ell(x_i^k, y_i^k, s_*^k)] + 2C_7 M_f^2 \log^3(nK/\delta) \left(r_{K, n}^\dagger + \frac{\log(1/\delta)}{nK} \right) \\ &\leq \inf_{f \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{4}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}}(x, y, s_{f^k, h})] + 6C_7 M_f^2 \log^3(nK/\delta) \left(r_{K, n}^\dagger + \frac{\log(1/\delta)}{nK} \right), \end{aligned} \quad (\text{B.62})$$

which concludes the proof. $\square$

Theorem B.6 (Thm. 3.4). *Under Assumption 3.1, 3.2, 3.3, suppose $\mathbb{P}^1, \dots, \mathbb{P}^K$ are (ν, Δ) -diverse over target distribution $\mathbb{P}^0$ given $\mathcal{F}, \mathcal{H}$. There exists some constant C, C_R such that the following holds. Define the empirical minimizer of training task and new task as*

$$\hat{f}, \hat{h} = \arg \min_{f \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \ell(x_i^k, y_i^k, s_{f^k, h}), \quad (\text{B.63})$$

$$\hat{f}^{\mathbb{P}^0} := \arg \min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(x_i^0, y_i^0, s_{f, \hat{h}}). \quad (\text{B.64})$$

If $R_f \geq C_R \log^{\frac{1}{2}}(nK M_f / \delta)$, then with probability no less than $1 - \delta$, the expected population loss of new task can be bounded by

$$\begin{aligned} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{(x, y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}^0}(x, y, s_{\hat{f}^0}, \hat{h})] &\lesssim \frac{1}{\nu} \inf_{h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f, h})] + \Delta \\ &+ C \left(\frac{\log^3(m) \log \mathcal{N}_{\mathcal{F}}}{m} + \frac{\log^3(nK/\delta)(K \log \mathcal{N}_{\mathcal{F}} + \log(\mathcal{N}_{\mathcal{H}}/\delta))}{\nu n K} \right). \end{aligned} \quad (\text{B.65})$$

where

$$\log \mathcal{N}_{\mathcal{F}} := M_f^2 S_f L_f \log(mn L_f W_f (B_f \vee 1) M_f T \log(1/\delta)), \quad (\text{B.66})$$

$$\log \mathcal{N}_{\mathcal{H}} := S_h L_h \log(nK L_h W_h (B_h \vee 1) M_f \gamma_f \log(1/\delta)). \quad (\text{B.67})$$

Proof.

$$\begin{aligned} &\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{(x, y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}^0}(x, y, s_{\hat{f}^0}, \hat{h})] \\ &\lesssim \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}}(x, y, s_{f, \hat{h}})] + C_{xy} \log^3(m) r_x \\ &\lesssim \frac{1}{\nu K} \sum_{k=1}^K \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f, \hat{h}})] + \Delta + C_{xy} \log^3(m) r_x \\ &\lesssim \frac{1}{\nu K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{\hat{f}^k}, \hat{h})] + \Delta + C_{xy} \log^3(m) r_x \\ &\lesssim \frac{1}{\nu} \inf_{f \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f^k, h})] + \frac{1}{\nu} C_Z \log^3(nK/\delta) \left(r_z + \frac{\log(1/\delta)}{nK} \right) \\ &\quad + \Delta + C_{xy} \log^3(m) r_x. \end{aligned} \quad (\text{B.68})$$

Here we apply Proposition B.4 in the first inequality, task diversity in the second inequality, and Proposition B.5 in the fourth. Plug in the definition of r_z, r_x and $\log \mathcal{N}_{\mathcal{F}}, \log \mathcal{N}_{\mathcal{H}}$ and we complete the proof. $\square$

B.3 Proofs of Meta-Learning

Proposition B.7 (Prop. 3.5). *There exists some constants C'_1, C_P , such that for $\mathbb{P}^1, \dots, \mathbb{P}^K \stackrel{i.i.d.}{\sim} \mathbb{P}_{\text{meta}}$, with probability no less than $1 - \delta$, we have for any $h \in \mathcal{H}$,*

$$\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathcal{L}(\mathbb{P}, h) \leq \frac{2}{K} \sum_{k=1}^K \mathcal{L}(\mathbb{P}^k, h) + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right), \quad (\text{B.69})$$

$$\frac{1}{K} \sum_{k=1}^K \mathcal{L}(\mathbb{P}^k, h) \leq 2 \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathcal{L}(\mathbb{P}, h) + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right), \quad (\text{B.70})$$

$$\text{where } r_P = M_f^2 \exp(-C'_1 R_f^2) + \frac{S_h L_h \log(K L_h W_h (B_h \vee 1) M_f \gamma_f)}{K}.$$

Proof. Given $\mathbb{P}^1, \dots, \mathbb{P}^K \stackrel{i.i.d.}{\sim} \mathbb{P}_{\text{meta}}$, we define the empirical Rademacher complexity of a function class Φ defined on the set of distribution $\mathcal{P}(\mathbb{R}^{d_x} \times [0, 1]^{d_y})$ as

$$\mathcal{R}_K(\Phi) := \mathbb{E}_{\sigma} \sup_{\varphi \in \Phi} \left| \frac{1}{K} \sum_{k=1}^K \sigma_k \varphi(\mathbb{P}^k) \right|, \quad \sigma \sim \text{Unif}(\{-1, 1\}^K). \quad (\text{B.71})$$

For any $r > 0$, let $\mathcal{H}_r := \left\{ h \in \mathcal{H} : \frac{1}{K} \sum_{k=1}^K (\mathcal{L}(\mathbb{P}^k, h))^2 \leq r \right\}$ and $\Phi_r := \{\mathcal{L}(\cdot, h) : h \in \mathcal{H}_r\}$. Note that for any $\varphi_1, \varphi_2 \in \Phi_r$,

$$\begin{aligned} \left\| \frac{1}{\sqrt{K}} \sum_{k=1}^K \sigma_k \varphi_1(\mathbb{P}^k) - \frac{1}{\sqrt{K}} \sum_{k=1}^K \sigma_k \varphi_2(\mathbb{P}^k) \right\|_{\psi_2} &\leq 4 \sqrt{\frac{1}{K} \sum_{k=1}^K \|\varphi_1(\mathbb{P}^k) - \varphi_2(\mathbb{P}^k)\|^2} \\ &= 4 \|\varphi_1 - \varphi_2\|_{L^2(\mathbb{P}_{\text{meta}}^{(K)})}, \end{aligned} \quad (\text{B.72})$$

where $\mathbb{P}_{\text{meta}}^{(K)} := \frac{1}{K} \sum_{k=1}^K \delta_{\mathbb{P}^k}$ and $\text{diam}(\Phi_r, \|\cdot\|_{L^2(\mathbb{P}_{\text{meta}}^{(K)})}) \leq 2\sqrt{r}$. Then by Dudley's bound [Van Handel, 2014, Wainwright, 2019], there exists an absolute constant C_0 such that for any $\theta \geq 0$,

$$\mathcal{R}_K(\Phi_r) \leq C_0 \left(\theta + \int_{\theta}^{2\sqrt{r}} \sqrt{\frac{\log \mathcal{N}(\Phi_r, \|\cdot\|_{L^2(\mathbb{P}_{\text{meta}}^{(K)})}, \varepsilon)}{K}} d\varepsilon \right). \quad (\text{B.73})$$

For any $\mathbb{P}$ and $h_1, h_2 \in \mathcal{H}_r$, denote the minimizer of (3.6) in $\mathcal{F}$ as f_1, f_2 , respectively. Without loss of generality, suppose $\mathcal{L}(\mathbb{P}, h_1) \geq \mathcal{L}(P, h_2)$. Then

$$\begin{aligned} \mathcal{L}(\mathbb{P}, h_1) - \mathcal{L}(P, h_2) &\leq \mathbb{E}_{t, x_t, y} \left[\left\| f_2(x_t, h_1(y), t) - \nabla_x \log p_t(x_t|y) \right\|^2 - \left\| f_2(x_t, h_2(y), t) - \nabla_x \log p_t(x_t|y) \right\|^2 \right] \\ &\leq \mathbb{E}_{t, x_t, y} \left[\left\| f_2(x_t, h_1(y), t) - f_2(x_t, h_2(y), t) \right\| \right. \\ &\quad \left. \times \left\| f_2(x_t, h_1(y), t) + f_2(x_t, h_2(y), t) - 2\nabla_x \log p_t(x_t|y) \right\| \right] \\ &\leq \sqrt{\mathbb{E}_{t, x_t, y} \left[\left\| f_2(x_t, h_1(y), t) - f_2(x_t, h_2(y), t) \right\|^2 \right]} \cdot 8(M_f + C_L^{1/2}) \end{aligned} \quad (\text{B.74})$$

In the last inequality we apply $\|f_i\| \leq M_f$ and $\mathbb{E}_{t, x_t, y} \|\nabla_x \log p_t(x_t|y)\|^2 \leq C_L$ by Lemma B.9. Moreover,

$$\begin{aligned} &\mathbb{E}_{(t, x_t, y)} \left[\left\| f_2(x_t, h_1(y), t) - f_2(x_t, h_2(y), t) \right\|^2 \right] \\ &\leq \mathbb{E}_{t, y} \left[\int \left\| f_2(x_t, h_1(y), t) - f_2(x_t, h_2(y), t) \right\|^2 p_t(x_t|y) dx_t \right] \\ &\leq \mathbb{E}_{t, y} \left[\int_{\|x_t\|_{\infty} \leq R_f} \left\| f_2(x_t, h_1(y), t) - f_2(x_t, h_2(y), t) \right\|^2 p_t(x_t|y) dx_t + 4M_f^2 \mathbb{P}(\|x_t\|_{\infty} > R_f|y) \right] \\ &\leq \gamma_f^2 \mathbb{E}_y [\|h_1(y) - h_2(y)\|^2] + 8M_f^2 \exp(-C_1' R_f^2) \\ &\leq \gamma_f^2 \|h_1 - h_2\|_{L^{\infty}([0,1]^{D_y})}^2 + 8M_f^2 \exp(-C_1' R_f^2) \end{aligned} \quad (\text{B.75})$$

Therefore, let $C_3 = 32(M_f + C_L^{1/2})M_f \leq 64M_f^2$ and we have

$$|\mathcal{L}(\mathbb{P}, h_1) - \mathcal{L}(P, h_2)| \leq C_3 \left(\gamma_f \|h_1 - h_2\|_{L^{\infty}([0,1]^{D_y})} + \exp(-C_1' R_f^2) \right), \quad (\text{B.76})$$

which implies that when $\varepsilon \geq 2C_3 \exp(-C_1' R_f^2)$, by Lemma B.3,

$$\begin{aligned} \log \mathcal{N}(\Phi_r, \|\cdot\|_{L^2(\mathbb{P}_{\text{meta}}^{(K)})}, \varepsilon) &\leq \log \mathcal{N}(\mathcal{H}_r, \|\cdot\|_{L^{\infty}([0,1]^{D_y})}, \varepsilon / (2C_3 \gamma_f)) \\ &\leq C_4 S_h L_h \log \left(\frac{L_h W_h (B_h \vee 1) C_3 \gamma_f}{\varepsilon} \right). \end{aligned} \quad (\text{B.77})$$

Plug in (B.73) and let $\theta = 2C_3 \exp(-C'_1 R_f^2)$,

$$\begin{aligned} \mathcal{R}_K(\Phi_r) &\leq C_0 \left(\theta + \int_{\theta}^{2\sqrt{r}} \sqrt{\frac{C_4 S_h L_h \log\left(\frac{L_h W_h (B_h \vee 1) C_3 \gamma_f}{\varepsilon}\right)}{K}} d\varepsilon \right) \\ &\leq C_0 \left(2C_3 \exp(-C'_1 R_f^2) + \sqrt{\frac{C'_4 S_h L_h \log\left(\frac{L_h W_h (B_h \vee 1) M_f \gamma_f}{r}\right)}{K}} \cdot r \right) \\ &=: \tilde{\mathcal{R}}_K(r). \end{aligned} \quad (\text{B.78})$$

According to Lemma B.11 (by setting $B_0 = 0, B = b = C_L$), for some absolute constant C_5 , with probability no less than $1 - \delta$, we have for any $h \in \mathcal{H}$,

$$\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathcal{L}(\mathbb{P}, h) \leq \frac{2}{K} \sum_{k=1}^K \mathcal{L}(\mathbb{P}^k, h) + C_5 C_L \left(r_K^* + \frac{\log(\log(K)/\delta)}{K} \right), \quad (\text{B.79})$$

$$\frac{1}{K} \sum_{k=1}^K \mathcal{L}(\mathbb{P}^k, h) \leq 2\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathcal{L}(\mathbb{P}, h) + C_5 C_L \left(r_K^* + \frac{\log(\log(K)/\delta)}{K} \right), \quad (\text{B.80})$$

where r_K^* is the unique fixed point of $\tilde{\mathcal{R}}_K$. And it is easy to show that for some absolute constant C_6 ,

$$r_K^* \leq C_6 \left(C_3 \exp(-C'_1 R_f^2) + \frac{S_h L_h \log(K L_h W_h (B_h \vee 1) M_f \gamma_f)}{K} \right). \quad (\text{B.81})$$

which concludes the proof. $\square$

Theorem B.8 (Thm. 3.6). *Under Assumption 3.1, 3.2, 3.3, there exists some constant C, C_R such that the following holds. Define the empirical minimizer of training task and new task as*

$$\hat{\mathbf{f}}, \hat{h} = \arg \min_{\mathbf{f} \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \ell(x_i^k, y_i^k, s_{f^k, h}), \quad (\text{B.82})$$

$$\hat{f}^{\mathbb{P}} := \arg \min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(x_i, y_i, s_{f, \hat{h}}). \quad (\text{B.83})$$

If $R_f \geq C_R \log^{\frac{1}{2}}(nK M_f / \delta)$, then with probability no less than $1 - \delta$, the expected population loss of new task can be bounded by

$$\begin{aligned} &\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{\hat{\mathbf{f}}, \hat{h}})] \\ &\lesssim \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f, h})] + C \left(\frac{\log^3(m) \log \mathcal{N}_{\mathcal{F}}}{m} + \frac{\log^3(nK/\delta) \log \mathcal{N}_{\mathcal{F}}}{n} + \frac{\log(\mathcal{N}_{\mathcal{H}}/\delta)}{K} \right), \end{aligned} \quad (\text{B.84})$$

where

$$\log \mathcal{N}_{\mathcal{F}} := M_f^2 S_f L_f \log(mn L_f W_f (B_f \vee 1) M_f T \log(1/\delta)), \quad (\text{B.85})$$

$$\log \mathcal{N}_{\mathcal{H}} := S_h L_h \log(nK L_h W_h (B_h \vee 1) M_f \gamma_f \log(1/\delta)). \quad (\text{B.86})$$

Proof.

$$\begin{aligned}
& \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{\hat{f}^{\mathbb{P}}, \hat{h}})] \\
& \lesssim \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f, \hat{h}})] + C_{xy} \log^3(m) r_x \\
& \lesssim \frac{1}{K} \sum_{k=1}^K \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f, \hat{h}})] + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right) + C_{xy} \log^3(m) r_x \\
& \lesssim \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{\hat{f}^k, \hat{h}})] + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right) + C_{xy} \log^3(m) r_x \\
& \lesssim \inf_{f \in \mathcal{F}^{\otimes K}, h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x, y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f^k, h})] + C_Z \log^3(nK/\delta) \left(r_z + \frac{\log(1/\delta)}{nK} \right) \\
& \quad + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right) + C_{xy} \log^3(m) r_x \\
& \lesssim \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f, h})] + C_Z \log^3(nK/\delta) \left(r_z + \frac{\log(1/\delta)}{nK} \right) \\
& \quad + C_P \left(r_P + \frac{\log(1/\delta)}{K} \right) + C_{xy} \log^3(m) r_x.
\end{aligned} \tag{B.87}$$

Here we apply Proposition B.4 in the first inequality, Proposition B.7 in the second and last inequality, Proposition B.5 in the fourth. Plugging in the definition of r_z, r_P, r_x and $\log \mathcal{N}_{\mathcal{F}}, \log \mathcal{N}_{\mathcal{H}}$ and noticing that $R_f \geq C_R \log^{\frac{1}{2}}(nK d_x M_f / \delta) \geq C'_R \log^{\frac{1}{2}} \left(\frac{M_f K}{\log \mathcal{N}_{\mathcal{H}}} \right)$, we have with probability no less than $1 - \delta$,

$$\begin{aligned}
& \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m \sim \mathbb{P}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{\hat{f}^{\mathbb{P}}, \hat{h}})] \\
& \lesssim \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f, h})] + C \left(\frac{\log^3(m) \log \mathcal{N}_{\mathcal{F}}}{m} + \frac{\log^3(nK/\delta) \log \mathcal{N}_{\mathcal{F}}}{n} + \frac{\log(\mathcal{N}_{\mathcal{H}}/\delta)}{K} \right).
\end{aligned} \tag{B.88}$$

□

B.4 Auxiliary Lemmas

Lemma B.9. *There exists some constant C_L such that for any $h, \mathbb{P}$,*

$$\mathcal{L}(\mathbb{P}, h) \leq \mathbb{E}_{t, x_t, y} \|\nabla_x \log p_t(x_t|y)\|^2 \leq C_L. \tag{B.89}$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f, h})] &= \mathbb{E}_{(x, y) \sim \mathbb{P}} \mathbb{E}_{t, x_t | x} [\|f(x_t, h(y), t) - \nabla_x \log p_t(x_t|y)\|^2] \\
&= \mathbb{E}_{t, x_t, y} [\|f(x_t, h(y), t) - \nabla_x \log p_t(x_t|y)\|^2]
\end{aligned} \tag{B.90}$$

and $0 \in \mathcal{F}$, it suffices to show that $\mathbb{E}_{t, x_t, y} [\|\nabla_x \log p_t(x_t|y)\|^2]$ is uniformly bounded for any $\mathbb{P}, h$. According to (B.2),

$$\begin{aligned}
\mathbb{E}_{x_t, y} [\|\nabla_x \log p_t(x_t|y)\|^2] &\leq \mathbb{E}_{x_t, y} \mathbb{E}_{x_0 | (x_t, y)} [\|\nabla_x \log \phi_t(x_t|x_0)\|^2] \\
&= \mathbb{E}_{x_0, y} \mathbb{E}_{x_t | x_0} [\|\nabla_x \log \phi_t(x_t|x_0)\|^2] \\
&= \frac{d_x}{\sigma_t^2} = \frac{d_x}{1 - e^{-2t}}.
\end{aligned} \tag{B.91}$$

On the other hand, by (B.3) and Assumption 3.3,

$$\begin{aligned}
\mathbb{E}_{x_t, y} [\|\nabla_x \log p_t(x_t|y)\|^2] &\leq \mathbb{E}_{x_t, y} \mathbb{E}_{x_0|(x_t, y)} [\|\nabla_x \log p(x_0|y)\|^2 \cdot e^{2t}] \\
&= \mathbb{E}_{x_0, y} \mathbb{E}_{x_t|x_0} [\|\nabla_x \log p(x_0|y)\|^2 \cdot e^{2t}] \\
&= \mathbb{E}_{x_0, y} [\|\nabla_x \log p(x_0|y)\|^2 / \alpha_t^2] \\
&\leq \mathbb{E}_{x_0, y} [(B + L\|x_0\|)^2 \cdot e^{2t}] \\
&\leq C'_2 e^{2t}
\end{aligned} \tag{B.92}$$

Therefore, we have

$$\begin{aligned}
\mathcal{L}(\mathbb{P}, h) &\leq \mathbb{E}_{t, x_t, y} [\|\nabla_x \log p_t(x_t|y)\|^2] \\
&\leq \mathbb{E}_t \left[\frac{d_x}{1 - e^{-2t}} \wedge C'_2 e^{2t} \right] \\
&\leq 2(C'_2 + d_x) =: C_L.
\end{aligned} \tag{B.93}$$

□

Lemma B.10. *There exists some constant C_X'' such that for any $t \in [0, T]$ and $x \in \mathbb{R}^{d_x}$, $y \in [0, 1]^{D_y}$,*

$$\mathbb{E}_{x_t|x} \|\nabla_x \log p_t(x_t|y)\|^2 \leq C_X'' (\|x\|^6 + 1). \tag{B.94}$$

Proof. Note that $x_t|x \sim \mathcal{N}(x_t|\alpha_t x, \sigma_t^2 I)$ and by Lemma B.2,

$$\mathbb{E}_{x_t|x} \|\nabla_x \log p_t(x_t|y)\|^2 \leq \mathbb{E}_{x_t|x} 2 \left[\|\nabla_x \log p_t(0|y)\|^2 + (C_X + C_X' \|x_t\|^2)^2 \|x_t\|^2 \right] \tag{B.95}$$

Let $q_t(x_0|x_t, y) \propto \phi_t(x_t|x_0)p(x_0|y)$. Since $\phi_t(0|x_0) \propto \exp\left(-\frac{\alpha_t^2 \|x\|^2}{2\sigma_t^2}\right)$ is decreasing in $\|x\|$, by Fortuin–Kasteleyn–Ginibre inequality,

$$\mathbb{E}_{q_t(x_0|0, y)} \|x_0\|^2 \leq \mathbb{E}_{p(x_0|y)} \|x_0\|^2 \leq C_0. \tag{B.96}$$

According to (B.2),

$$\|\nabla_x \log p_t(0|y)\|^2 \leq \frac{\alpha_t^2}{\sigma_t^4} \mathbb{E}_{q_t(x_0|0, y)} \|x_0\|^2 \leq \frac{C_0 \alpha_t^2}{\sigma_t^4}. \tag{B.97}$$

By (B.3), we also have

$$\|\nabla_x \log p_t(0|y)\|^2 \leq \frac{1}{\alpha_t^2} \mathbb{E}_{q_t(x_0|0, y)} \|\nabla_x \log p(x_0|y)\|^2 \leq \frac{1}{\alpha_t^2} \mathbb{E}_{q_t(x_0|0, y)} [(B + L\|x_0\|)^2] \leq \frac{2(B^2 + L^2 C_0)}{\alpha_t^2}. \tag{B.98}$$

Combine the two inequalities,

$$\|\nabla_x \log p_t(0|y)\|^2 \leq (B^2 + (L^2 + 1)C_0) \cdot \left(\frac{\alpha_t^2}{\sigma_t^4} \wedge \frac{1}{\alpha_t^2} \right) \leq 2(B^2 + (L^2 + 1)C_0). \tag{B.99}$$

Plug in (B.95) and we obtain for some constant C_X'' ,

$$\begin{aligned}
\mathbb{E}_{x_t|x} \|\nabla_x \log p_t(x_t|y)\|^2 &\leq \mathbb{E}_{x_t|x} 2 \left[(C_X + C_X' \|x_t\|^2)^2 \|x_t\|^2 \right] + 2(B^2 + (L^2 + 1)C_0) \\
&\leq C_X'' (\|x\|^6 + 1).
\end{aligned} \tag{B.100}$$

□

Lemma B.11. *Let Φ be a class of functions on domain Ω and $\mathbb{P}$ be a probability distribution over Ω . Suppose that for any $\varphi \in \Phi$, $\|\varphi\|_{L^\infty(\Omega)} \leq b$, $\mathbb{E}_{\mathbb{P}}[\varphi] \geq 0$, and $\mathbb{E}_{\mathbb{P}}[\varphi^2] \leq B\mathbb{E}_{\mathbb{P}}[\varphi] + B_0$ for some $b, B, B_0 \geq 0$. Let $x_1, \dots, x_n \stackrel{i.i.d.}{\sim} \mathbb{P}$ and ϕ_n be a positive, non-decreasing and sub-root function such that*

$$\mathcal{R}_n(\Phi_r) := \mathbb{E}_{\sigma} \sup_{\varphi \in \Phi_r} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \varphi(x_i) \right| \leq \phi_n(r). \tag{B.101}$$

where $\Phi_r := \left\{ \varphi \in \Phi : \frac{1}{n} \sum_{i=1}^n (\varphi(x_i))^2 \leq r \right\}$. Define the largest fixed point of ϕ_n as r_n^* . Then for some absolute constant C' , with probability no less than $1 - \delta$, it holds that for any $\varphi \in \Phi$,

$$\mathbb{E}_{\mathbb{P}}[\varphi] \leq \frac{2}{n} \sum_{i=1}^n \varphi(x_i) + C'(B \vee b) \left(r_n^* + \frac{\log((\log n)/\delta)}{n} \right) + C' \sqrt{\frac{B_0 \log((\log n)/\delta)}{n}}, \quad (\text{B.102})$$

$$\frac{1}{n} \sum_{i=1}^n \varphi(x_i) \leq 2\mathbb{E}_{\mathbb{P}}[\varphi] + C'(B \vee b) \left(r_n^* + \frac{\log((\log n)/\delta)}{n} \right) + C' \sqrt{\frac{B_0 \log((\log n)/\delta)}{n}}. \quad (\text{B.103})$$

Proof. We follow the procedures in [Bousquet \[2002\]](#). Let $\epsilon_j = b2^{-j}$ and consider a sequence of classes

$$\Phi^{(j)} := \{ \varphi \in \Phi : \epsilon_{j+1} < \mathbb{E}_{\mathbb{P}}[\varphi] \leq \epsilon_j \}. \quad (\text{B.104})$$

Note that $\Phi = \cup_{j \geq 0} \Phi^{(j)}$ and for $\varphi \in \Phi^{(j)}$, $\mathbb{E}_{\mathbb{P}}[\varphi^2] \leq B\epsilon_j + B_0$. Let $j_0 = \lfloor \log_2 n \rfloor$. Then by [Bousquet \[2002, Lemma 6.1\]](#), it holds that with probability no less than $1 - \delta$, for any $j \leq j_0$ and $\varphi \in \Phi^{(j)}$,

$$\left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \mathbb{E}_{\mathbb{P}}[\varphi] \right| \lesssim \mathcal{R}_n(\Phi^{(j)}) + \sqrt{\frac{(B\epsilon_j + B_0) \log(\log(b/\epsilon_j)/\delta)}{n}} + \frac{b \log(\log(b/\epsilon_j)/\delta)}{n}, \quad (\text{B.105})$$

$$\left| \frac{1}{n} \sum_{i=1}^n (\varphi(x_i))^2 - \mathbb{E}_{\mathbb{P}}[\varphi^2] \right| \lesssim b\mathcal{R}_n(\Phi^{(j)}) + \sqrt{\frac{b^2(B\epsilon_j + B_0) \log(\log(b/\epsilon_j)/\delta)}{n}} + \frac{b^2 \log(\log(b/\epsilon_j)/\delta)}{n}. \quad (\text{B.106})$$

Besides, for $\varphi \in \cup_{k > k_0} \Phi^{(j)} =: \Phi^{(j_0 \cdot)}$,

$$\left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \mathbb{E}_{\mathbb{P}}[\varphi] \right| \lesssim \mathcal{R}_n(\Phi^{(j_0 \cdot)}) + \sqrt{\frac{(B\epsilon_{j_0} + B_0) \log(\log(n)/\delta)}{n}} + \frac{b \log((\log n)/\delta)}{n} \quad (\text{B.107})$$

From now on we reason on the conjunction of (B.105), (B.106) and (B.107). Define

$$U_j = B\epsilon_j + B_0 + b\mathcal{R}_n(\Phi^{(k)}) + \sqrt{\frac{b^2(B\epsilon_j + B_0) \log(\log(b/\epsilon_j)/\delta)}{n}} + \frac{b^2 \log(\log(b/\epsilon_j)/\delta)}{n}. \quad (\text{B.108})$$

and thus for any $\varphi \in \Phi^{(j)}$, we have $\frac{1}{n} \sum_{i=1}^n (\varphi(x_i))^2 \leq CU_j$ for some absolute constant C by (B.106), indicating that $\mathcal{R}_n(\Phi^{(j)}) \leq \phi_n(CU_j) \leq \sqrt{C}\phi_n(U_j)$. For any $j \leq j_0$,

$$U_j \leq 2(B\epsilon_j + B_0) + b\sqrt{C}\phi_n(U_j) + \frac{2b^2 \log((\log n)/\delta)}{n}. \quad (\text{B.109})$$

Since ϕ_n is non-decreasing and sub-root, the inequality above implies that

$$U_j \lesssim b^2 r_n^* + B\epsilon_j + B_0 + \frac{b^2 \log((\log n)/\delta)}{n} =: r_n(\epsilon_j). \quad (\text{B.110})$$

Therefore, for any $\varphi \in \Phi^{(j)}$, $j \leq j_0$, by (B.105),

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \mathbb{E}_{\mathbb{P}}[\varphi] \right| &\lesssim \phi_n(r_n(\epsilon_j)) + \sqrt{\frac{(B\epsilon_j + B_0) \log((\log n)/\delta)}{n}} + \frac{b \log((\log n)/\delta)}{n} \\ &=: F_n(\epsilon_j). \end{aligned} \quad (\text{B.111})$$

Noticing that $\mathbb{E}_{\mathbb{P}}[\varphi] \leq \epsilon_j \leq 2\mathbb{E}_{\mathbb{P}}[\varphi]$, it reduces to

$$\left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \mathbb{E}_{\mathbb{P}}[\varphi] \right| \lesssim F_n(\mathbb{E}_{\mathbb{P}}[\varphi]). \quad (\text{B.112})$$

Hence we have by noting that F_n is also a non-decreasing sub-root function,

$$\mathbb{E}_{\mathbb{P}}[\varphi] \leq \frac{2}{n} \sum_{i=1}^n \varphi(x_i) + C'(B \vee b) \left(r_n^* + \frac{\log((\log n)/\delta)}{n} \right) + C' \sqrt{\frac{B_0 \log((\log n)/\delta)}{n}}, \quad (\text{B.113})$$

$$\frac{1}{n} \sum_{i=1}^n \varphi(x_i) \leq 2\mathbb{E}_{\mathbb{P}}[\varphi] + C'(B \vee b) \left(r_n^* + \frac{\log((\log n)/\delta)}{n} \right) + C' \sqrt{\frac{B_0 \log((\log n)/\delta)}{n}}. \quad (\text{B.114})$$

Here C' is an absolute constant. Moreover, when $\varphi \in \Phi^{(j)}$ for $j > j_0$, we have $\mathbb{E}_{\mathbb{P}}[\varphi] \leq \frac{b}{n}$, and according to (B.107),

$$\left| \frac{1}{n} \sum_{i=1}^n \varphi(x_i) - \mathbb{E}_{\mathbb{P}}[\varphi] \right| \lesssim F_n(\epsilon_{j_0}). \quad (\text{B.115})$$

Hence the same bounds apply, which completes the proof. $\square$

Lemma B.12. *Let Φ be a class of functions on domain Ω , $\mathbb{P}^1, \dots, \mathbb{P}^K$ be probability distributions over Ω , and $\widehat{\mathbb{P}}^{(K)} = \frac{1}{K} \sum_{k=1}^K \delta_{\mathbb{P}^k}$. Suppose that for any $\varphi \in \Phi$, $\|\varphi\|_{L^\infty(\Omega)} \leq b$, $\mathbb{E}_{\widehat{\mathbb{P}}^{(K)}}[\varphi] \geq 0$, and $\mathbb{E}_{\widehat{\mathbb{P}}^{(K)}}[\varphi^2] \leq B\mathbb{E}_{\widehat{\mathbb{P}}^{(K)}}[\varphi] + B_0$ for some $b, B, B_0 \geq 0$. Let $x_1^k, \dots, x_n^k \stackrel{i.i.d.}{\sim} \mathbb{P}^k$ for any k and all $(x_i^k)_{i,k}$ are independent. Let $\phi_{K,n}$ be a positive, non-decreasing and sub-root function such that*

$$\mathcal{R}_{K,n}(\Phi_r) := \mathbb{E}_{\sigma} \sup_{\varphi \in \Phi_r} \left| \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \sigma_i^k \varphi(x_i^k) \right| \leq \phi_{K,n}(r). \quad (\text{B.116})$$

where $\Phi_r := \left\{ \varphi \in \Phi : \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (\varphi(x_i^k))^2 \leq r \right\}$. Define the largest fixed point of $\phi_{K,n}$ as $r_{K,n}^*$.

Then for some absolute constant C' , with probability no less than $1 - \delta$, it holds that for any $\varphi \in \Phi$,

$$\mathbb{E}_{\widehat{\mathbb{P}}^{(K)}}[\varphi] \leq \frac{2}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i) + C'(B \vee b) \left(r_{K,n}^* + \frac{\log((\log nK)/\delta)}{nK} \right) + C' \sqrt{\frac{B_0 \log((\log nK)/\delta)}{nK}}, \quad (\text{B.117})$$

$$\frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) \leq 2\mathbb{E}_{\widehat{\mathbb{P}}^{(K)}}[\varphi] + C'(B \vee b) \left(r_{K,n}^* + \frac{\log((\log nK)/\delta)}{nK} \right) + C' \sqrt{\frac{B_0 \log((\log nK)/\delta)}{nK}}. \quad (\text{B.118})$$

Proof. We follow the procedures in Bousquet [2002]. Let $\epsilon_k = b2^{-k}$ and consider a sequence of classes

$$\Phi^{(j)} := \{\varphi \in \Phi : \epsilon_{j+1} < \mathbb{E}_{\widehat{\mathbb{P}}^{(K)}}[\varphi] \leq \epsilon_j\}. \quad (\text{B.119})$$

Note that $\Phi = \cup_{j \geq 0} \Phi^{(j)}$ and for $\varphi \in \Phi^{(j)}$, $\mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi^2] \leq B\epsilon_j + B_0$. Let $j_0 = \lfloor \log_2(nK) \rfloor$. Then by [Massart \[2000, Theorem 3\]](#), with probability no less than $1 - \delta$, for any $j \leq j_0$ and $\varphi \in \Phi^{(j)}$,

$$\left| \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) - \mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi] \right| \lesssim \mathcal{R}_{K,n}(\Phi^{(j)}) + \sqrt{\frac{(B\epsilon_j + B_0) \log(\log(b/\epsilon_j)/\delta)}{nK}} + \frac{b \log(\log(b/\epsilon_j)/\delta)}{nK}, \quad (\text{B.120})$$

$$\left| \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (\varphi(x_i^k))^2 - \mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi^2] \right| \lesssim b\mathcal{R}_{K,n}(\Phi^{(j)}) + \sqrt{\frac{b^2(B\epsilon_j + B_0) \log(\log(b/\epsilon_j)/\delta)}{nK}} + \frac{b^2 \log(\log(b/\epsilon_j)/\delta)}{nK}. \quad (\text{B.121})$$

Besides, for any $\varphi \in \cup_{j > j_0} \Phi^{(j)} =: \Phi^{(j_0 \cdot)}$,

$$\left| \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) - \mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi] \right| \lesssim \mathcal{R}_{K,n}(\Phi^{(j_0 \cdot)}) + \sqrt{\frac{(B\epsilon_{j_0} + B_0) \log((\log nK)/\delta)}{nK}} + \frac{b \log((\log nK)/\delta)}{nK}. \quad (\text{B.122})$$

From now on we reason on the conjunction of (B.120), (B.121) and (B.122). Define

$$U_j = B\epsilon_j + B_0 + b\mathcal{R}_{K,n}(\Phi^{(j)}) + \sqrt{\frac{b^2(B\epsilon_j + B_0) \log(\log(b/\epsilon_j)/\delta)}{nK}} + \frac{b^2 \log(\log(b/\epsilon_j)/\delta)}{nK}. \quad (\text{B.123})$$

and thus for any $\varphi \in \Phi^{(j)}$, we have $\frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (\varphi(x_i^k))^2 \leq CU_j$ for some absolute constant C by (B.121), indicating that $\mathcal{R}_{K,n}(\Phi^{(j)}) \leq \phi_{K,n}(CU_j) \leq \sqrt{C}\phi_{K,n}(U_j)$. For any $j \leq j_0$,

$$U_j \leq 2(B\epsilon_j + B_0) + b\sqrt{C}\phi_{K,n}(U_j) + \frac{2b^2 \log((\log nK)/\delta)}{nK}. \quad (\text{B.124})$$

Since $\phi_{K,n}$ is non-decreasing and sub-root, the inequality above implies that

$$U_j \lesssim b^2 r_{K,n}^* + B\epsilon_j + B_0 + \frac{b^2 \log((\log nK)/\delta)}{nK} =: r_{K,n}(\epsilon_j). \quad (\text{B.125})$$

Therefore, for any $\varphi \in \Phi^{(j)}$, $j \leq j_0$, by (B.120),

$$\begin{aligned} \left| \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) - \mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi] \right| &\lesssim \phi_{K,n}(r_{K,n}(\epsilon_j)) + \sqrt{\frac{(B\epsilon_j + B_0) \log((\log nK)/\delta)}{nK}} + \frac{b \log((\log nK)/\delta)}{nK} \\ &=: F_{K,n}(\epsilon_j). \end{aligned} \quad (\text{B.126})$$

Noticing that $\mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi] \leq \epsilon_j \leq 2\mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi]$, it reduces to

$$\left| \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) - \mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi] \right| \lesssim F_{K,n}(\mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi]). \quad (\text{B.127})$$

Hence we have by noting that $F_{K,n}$ is also a non-decreasing sub-root function,

$$\mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi] \leq \frac{2}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) + C'(B \vee b) \left(r_{K,n}^* + \frac{\log((\log nK)/\delta)}{nK} \right) + C' \sqrt{\frac{B_0 \log((\log nK)/\delta)}{nK}}, \quad (\text{B.128})$$

$$\frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) \leq 2\mathbb{E}_{\widehat{\mathbb{P}}(K)}[\varphi] + C'(B \vee b) \left(r_{K,n}^* + \frac{\log((\log nK)/\delta)}{nK} \right) + C' \sqrt{\frac{B_0 \log((\log nK)/\delta)}{nK}}. \quad (\text{B.129})$$

Here C' is an absolute constant. Moreover, when $\varphi \in \Phi^{(j)}$ for $j > j_0$, we have $\mathbb{E}_{\hat{\mathbb{P}}(K)}[\varphi] \leq \frac{b}{nK}$, and according to (B.122),

$$\left| \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \varphi(x_i^k) - \mathbb{E}_{\hat{\mathbb{P}}(K)}[\varphi] \right| \lesssim F_{K,n}(\varepsilon_{j_0}). \quad (\text{B.130})$$

Hence the same bounds apply, which completes the proof. $\square$

B.5 Verifying Task Diversity Assumption

When $\mathcal{F}$ is linear function class, Tripuraneni et al. [2020] provides an explicit bound on (ν, Δ) . However, in general, performing a fine-grained analysis is challenging, especially for complex function classes such as neural networks. In the following proposition, we present a very pessimistic bound for (ν, Δ) based on density ratio, which is independent of the specific choice of hypothesis classes $\mathcal{F}$ and $\mathcal{H}$.

Proposition B.13. Suppose $\mathcal{F} = \text{conv}(\mathcal{F})$, and $\inf_{x,y} \frac{p^k(x,y)}{p^0(x,y)} \geq \lambda_k$ for any $1 \leq k \leq K$. Let $\lambda = \sum_{k=1}^K \lambda_k$. Then $\mathbb{P}^1, \dots, \mathbb{P}^K$ are $(\tilde{\nu}, \tilde{\Delta})$ -diverse over $\mathbb{P}^0$ with $\tilde{\nu} = \lambda/(2K)$,

$$\tilde{\Delta} = 2\mathbb{E}_{(x,y) \sim \mathbb{P}^0} \mathbb{E}_{t,x_t} \left[\left\| \frac{1}{\lambda} \sum_{k=1}^K \lambda_k \nabla \log p_t^k(x_t|y) - \nabla \log p_t^0(x_t|y) \right\|^2 \right]. \quad (\text{B.131})$$

We mention that the only requirement is $\mathcal{F}$ is a convex hull of itself, which can be easily satisfied by most hypothesis classes such as neural networks. More refined analysis on specific neural network class is an interesting future work.

Proof. For any $h \in \mathcal{H}$, let $f^k \in \mathcal{F}$ be the corresponding minimizer for $1 \leq k \leq K$. Further define

$$\lambda = \sum_{k=1}^K \lambda_k \text{ and } \tilde{f}^0 = \frac{1}{\lambda} \sum_{k=1}^K \lambda_k f^k \in \text{conv}(\mathcal{F}) \in \mathcal{F}. \text{ Then we have}$$

$$\begin{aligned} L^{\mathbb{P}^0}(s_{\tilde{f}^0,h}) &= \mathbb{E}_{\mathbb{P}^0} \left[\|\tilde{f}^0(x_t, h(y), t) - \nabla \log p_t^0(x_t|y)\|^2 \right] \\ &\leq 2\mathbb{E}_{\mathbb{P}^0} \left[\|\tilde{f}^0(x_t, h(y), t) - \sum_{k=1}^K \frac{\lambda_k}{\lambda} \nabla \log p_t^k(x_t|y)\|^2 + \left\| \sum_{k=1}^K \frac{\lambda_k}{\lambda} \nabla \log p_t^k(x_t|y) - \nabla \log p_t^0(x_t|y) \right\|^2 \right] \\ &\leq \frac{2}{\lambda} \sum_{k=1}^K \mathbb{E}_{\mathbb{P}^0} \lambda_k [\|f^k(x_t, h(y), t) - \nabla \log p_t^k(x_t|y)\|^2] + \tilde{\Delta} \\ &\leq \frac{2}{\lambda} \sum_{k=1}^K \mathbb{E}_{\mathbb{P}^k} [\|f^k(x_t, h(y), t) - \nabla \log p_t^k(x_t|y)\|^2] + \tilde{\Delta} \\ &= \frac{1}{\tilde{\nu}} \inf_{f \in \mathcal{F}^{\otimes K}} \frac{1}{K} \sum_{k=1}^K L^{\mathbb{P}^k}(s_{f^k,h}) + \tilde{\Delta}. \end{aligned} \quad (\text{B.132})$$

We conclude the proof by noticing that $\inf_{f \in \mathcal{F}} L^{\mathbb{P}^0}(s_{f,h}) \leq L^{\mathbb{P}^0}(s_{\tilde{f}^0,h})$. $\square$

C Proofs in Section 4

C.1 Proofs of Score Network Approximation

Theorem C.1 (Thm. 4.1). *Under Assumption 3.1, 3.2, 3.3, to achieve $R_f \geq C_R \log^{\frac{1}{2}}(nKM_f/\delta)$ and*

$$\inf_{h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^K \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}^k} [\ell^{\mathbb{P}^k}(x, y, s_{f,h})] = \mathcal{O}(\log^2(nK/(\varepsilon\delta))\varepsilon^2), \text{ (transfer learning)} \quad (\text{C.1})$$

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})] = \mathcal{O}(\log^2(nK/(\varepsilon\delta))\varepsilon^2), \text{ (meta-learning)} \quad (\text{C.2})$$

the configuration of $\mathcal{F} = NN_f(L_f, W_f, M_f, S_f, B_f, R_f, \gamma_f)$, $\mathcal{H} = NN_h(L_h, W_h, S_h, B_h)$ should satisfy

$$\begin{aligned} L_f &= \mathcal{O}\left(\log\left(\frac{\log(nK/(\varepsilon\delta))}{\varepsilon}\right)\right), W_f = \mathcal{O}\left(\frac{\log^{3(d_x+d_y)/2}(nK/(\varepsilon\delta))}{\varepsilon^{d_x+d_y+1}T_0^3}\right), \\ S_f &= \mathcal{O}\left(\frac{\log^{3(d_x+d_y)/2+1}(nK/(\varepsilon\delta))}{\varepsilon^{d_x+d_y+1}T_0^3}\right), B_f = \mathcal{O}\left(\frac{T \log^{\frac{3}{2}}(nK/(\varepsilon\delta))}{\varepsilon}\right), \\ R_f &= \mathcal{O}\left(\log^{\frac{1}{2}}(nK/(\varepsilon\delta))\right), M_f = \mathcal{O}(\log^3(nK/(\varepsilon\delta))), \gamma_f = \mathcal{O}(\log(nK/(\varepsilon\delta))), \end{aligned} \quad (\text{C.3})$$

$$\begin{aligned} L_h &= \mathcal{O}(\log(1/\varepsilon)), W_h = \mathcal{O}(\varepsilon^{-D_y} \log(1/\varepsilon)), \\ S_h &= \mathcal{O}(\varepsilon^{-D_y} \log^2(1/\varepsilon)), B_h = \mathcal{O}(1). \end{aligned} \quad (\text{C.4})$$

Here $\mathcal{O}(\cdot)$ hides all the polynomial factors of $d_x, d_y, D_y, C_1, C_2, L, B$.

Proof. With a little abuse of notation, in transfer learning setting, we define $\mathbb{P}_{\text{meta}} := \frac{1}{K} \sum_{k=1}^K \delta_{\mathbb{P}^k}$ and it directly reduces to meta-learning case. Therefore, we only focus on the proof in meta-learning.

We first decompose the misspecification error into two components: representation error and score approximation error.

$$\begin{aligned} &\inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})] \\ &= \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} \mathbb{E}_{t, x_t | x} [\|f(x_t, h(y), t) - f_{*}^{\mathbb{P}}(x_t, h_{*}(y), t)\|^2] \\ &\leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} \mathbb{E}_{t, x_t | x} 2 [\|f(x_t, h(y), t) - f(x_t, h_{*}(y), t)\|^2 + \|f(x_t, h_{*}(y), t) - f_{*}^{\mathbb{P}}(x_t, h_{*}(y), t)\|^2]. \end{aligned} \quad (\text{C.5})$$

Further note that for any $f \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathbb{P}} \mathbb{E}_{t, x_t | x} [\|f(x_t, h(y), t) - f(x_t, h_{*}(y), t)\|^2] &\leq \mathbb{E}_{t, x_t, y} \|f(x_t, h(y), t) - f(x_t, h_{*}(y), t)\|^2 \cdot \mathbb{1}_{\|x_t\| \leq R_f} \\ &\quad + 8M_f^2 \exp(-C'_1 R_f^2) \\ &\leq \mathbb{E}_{y \sim \mathbb{P}} [\gamma_f^2 \|h(y) - h_{*}(y)\|^2] + 8M_f^2 \exp(-C'_1 R_f^2), \end{aligned} \quad (\text{C.6})$$

where $\Omega_{R_f} = [-R_f, R_f]^{d_x} \times [0, 1]^{d_y} \times [T_0, T]$. By Proposition C.2, C.3,

$$\begin{aligned}
& \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{f,h})] \\
& \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{y \sim \mathbb{P}} [2\gamma_f^2 \|h(y) - h_*(y)\|^2] + 16M_f^2 \exp(-C'_1 R_f^2) \\
& \quad + \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} 2\|f(x_t, h_*(y), t) - f_*^{\mathbb{P}}(x_t, h_*(y), t)\|^2 \\
& \leq 2 \inf_{h \in \mathcal{H}} \gamma_f^2 \|h - h_*\|_{L^\infty([0,1]^{D_y})}^2 + 16M_f^2 \exp(-C'_1 R_f^2) \\
& \quad + 2\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \inf_{f \in \mathcal{F}} \|f(x_t, h_*(y), t) - f_*^{\mathbb{P}}(x_t, h_*(y), t)\|^2 \\
& \lesssim (\log^2(nK/(\varepsilon\delta))d_y + d_x) \varepsilon^2 \\
& = \mathcal{O}(\log^2(nK/(\varepsilon\delta))\varepsilon^2).
\end{aligned} \tag{C.7}$$

□

Proposition C.2. To achieve $R_f \geq C_R \log^{\frac{1}{2}}(nKM_f/\delta)$ and

$$\inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathbb{P}} \mathbb{E}_{t, x_t | x} [\|f(x_t, h_*(y), t) - f_*^{\mathbb{P}}(x_t, h_*(y), t)\|^2] \leq d_x \varepsilon^2, \tag{C.8}$$

the configuration of $\mathcal{F} = NN_f(L_f, W_f, M_f, S_f, B_f, R_f, \gamma_f)$ should satisfy

$$\begin{aligned}
L_f &= \mathcal{O}\left(\log\left(\frac{\log(nK/(\varepsilon\delta))}{\varepsilon}\right)\right), W_f = \mathcal{O}\left(\frac{\log^{3(d_x+d_y)/2}(nK/(\varepsilon\delta))}{\varepsilon^{d_x+d_y+1}T_0^3}\right), \\
S_f &= \mathcal{O}\left(\frac{\log^{3(d_x+d_y)/2+1}(nK/(\varepsilon\delta))}{\varepsilon^{d_x+d_y+1}T_0^3}\right), B_f = \mathcal{O}\left(\frac{T \log^{\frac{3}{2}}(nK/(\varepsilon\delta))}{\varepsilon}\right), \\
R_f &= \mathcal{O}\left(\log^{\frac{1}{2}}(nK/(\varepsilon\delta))\right), M_f = \mathcal{O}(\log^3(nK/(\varepsilon\delta))), \gamma_f = \mathcal{O}(\log(nK/(\varepsilon\delta))).
\end{aligned} \tag{C.9}$$

Here $\mathcal{O}(\cdot)$ hides all the polynomial factors of $d_x, d_y, D_y, C_1, C_2, L, B$.

Proof. For notation simplicity, we will $f_* = f_*^{\mathbb{P}}$ throughout the proof. Our procedures consist of two main steps. The first is to clip the whole input space to a bounded set $\Omega_{R_f} := [-R_f, R_f]^{d_x} \times [0, 1]^{d_y} \times [T_0, T]$ thanks to the light tail property of $\mathbb{P}$. Then we approximate $f_*^{\mathbb{P}}$ on Ω_{R_f} .

By Lemma B.2 and C.6, f_* is γ_1 -Lipschitz in x , γ_2 -Lipschitz in w , and γ_3 -Lipshcitz in t in a bounded domain Ω_{R_f} , where $\gamma_1 = C_X + C'_X R_f^2$, $\gamma_2 = C_X + C'_X R_f$, $\gamma_3 = \frac{C_s R_f^3}{T_0^3}$.

We first rescale the input domain by $x' = \frac{x}{2R_f} + \frac{1}{2}$, $w' = w$, $t' = t/T$, which can be implemented by a single ReLU layer. Denote $v = (x', w', t')$. We only need to approximate $g(v) := f_*(R_f(2x' - 1), w', Tt')$ defined on $\Omega := [0, 1]^{d_x+d_y} \times [T_0/T, 1]$. And g is $\gamma_x := 2\gamma_1 R_f$ -Lipschitz in x' , $\gamma_w := \gamma_2$ -Lipschitz in w' and $\gamma_t := \gamma_2 T$ -Lipschitz in t' . We will approximate each coordinate of $g = [g_1, \dots, g_{d_x}]^\top$ separately and then concatenate them together.

Now we partition the domain Ω into non-overlapping regions. For the first $d_x + d_y$ dimensions, the space $[0, 1]^{d_x+d_y}$ is uniformly divided into hypercubes with an edge length of e_1 . For the last dimension, the interval $[T_0/T, 1]$ is divided into subintervals of length e_2 , where the values of e_1 and e_2 will be specified later. Let the number of intervals in each partition be $N_1 = \lceil 1/e_1 \rceil$ and $N_2 = \lceil 1/e_2 \rceil$, respectively.

Let $u = [u_1, \dots, u_{d_x+d_y}] \in \{0, \dots, N_1 - 1\}^{d_x+d_y}$ be a multi-index. Define

$$\bar{g}_i(x', w', t') = \sum_{u,j} g_i(u/N_1, j/N_2) \Psi_{u,j}(x', w', t'), \tag{C.10}$$

where Ψ is the coordinate-wise product of trapezoid function:

$$\Psi_{u,j}(x', w', t') := \psi(3N_2(t' - j/N_2)) \prod_{r=1}^{d_x} \psi(3N_1(x'_r - u_r/N_1)) \prod_{r=1}^{d_y} \psi(3N_1(w'_r - u_{r+d_x}/N_1)), \quad (\text{C.11})$$

$$\psi(a) := \begin{cases} 1, & |a| < 1 \\ 2 - |a|, & 1 \leq |a| < 2 \\ 0, & |a| \geq 2 \end{cases} \quad (\text{C.12})$$

We claim that $\bar{g}_i$ is an approximation to g_i since for any $o' = (x', w') \in [0, 1]^{d_x+d_y}, t' \in [T_0/T, 1]$,

$$\begin{aligned} \sup_{o', t'} |\bar{g}_i(o', t') - g_i(o', t')| &\leq \sup_{o', t'} \left| \sum_{u,j} (g_i(\frac{u}{N_1}, \frac{j}{N_2}) - g_i(o', t')) \Psi_{u,j}(o', t') \right| \\ &\leq \sup_{o', t'} \sum_{u: |\frac{u}{N_1} - o'_i| \leq \frac{2N_1}{3}, j: |\frac{j}{N_2} - t'| \leq \frac{2N_2}{3}} \left| g_i(\frac{u}{N_1}, \frac{j}{N_2}) - g_i(o', t') \right| \Psi_{u,j}(o', t') \\ &\leq \frac{2\gamma_x}{3N_1} + \frac{2\gamma_t}{3N_2}. \end{aligned} \quad (\text{C.13})$$

Below we construct a ReLU neural network to approximate $\bar{g}_i$. Let σ be ReLU activation and $r(a) = 2\sigma(a) - 4\sigma(a - 0.5) + 2\sigma(a - 1)$ for any scalar $a \in [0, 1]$. Define

$$\phi_{\text{square}}^l(a) = a - \sum_{m=1}^l 2^{-2m} r_m(a), \quad r_m = \underbrace{r \circ \dots \circ r}_{m \text{ compositions}} \quad (\text{C.14})$$

$$\phi_{\text{mul}}^l(a, b) = \phi_{\text{square}}^l\left(\frac{a+b}{2}\right) - \phi_{\text{square}}^l\left(\frac{a-b}{2}\right) \quad (\text{C.15})$$

According to [Yarotsky \[2017\]](#),

$$|\phi_{\text{mul}}^l(a, b) - ab| \leq 2^{-2l-2}, \quad \forall a, b \in [0, 1]. \quad (\text{C.16})$$

Then we approximate $\Psi_{u,j}$ by recursively apply ϕ_{mul}^l :

$$\widehat{\Psi}_{u,j}(x', w', t') := \phi_{\text{mul}}^l(\psi(3N_2(t' - j/N_2)), \phi_{\text{mul}}^l(\psi(3N_1(x'_1 - u_1/N_1)), \dots)) \quad (\text{C.17})$$

And we construct the final neural network approximation as

$$\widehat{g}_i(x', w', t') := \sum_{u,j} g_i(u/N_1, j/N_2) \widehat{\Psi}_{u,j}(x', w', t'). \quad (\text{C.18})$$

The approximation error of $\widehat{g}_i$ can be bounded by

$$\begin{aligned} \|\widehat{g}_i - g_i\|_{L^\infty(\Omega)} &\leq \|\widehat{g}_i - \bar{g}_i\|_{L^\infty(\Omega)} + \|\bar{g}_i - g_i\|_{L^\infty(\Omega)} \\ &\leq 2^{d_x+d_y+1} \|g_i\|_{L^\infty(\Omega)} \sup_{u,j} \|\widehat{\Psi}_{u,j} - \Psi_{u,j}\|_{L^\infty(\Omega)} + \frac{2\gamma_x(d_x+d_y)^{\frac{1}{2}}}{3N_1} + \frac{2\gamma_t}{3N_2} \\ &\leq (d_x+d_y+1) 2^{d_x+d_y+1} \|g_i\|_{L^\infty(\Omega)} 2^{-(2l+2)} + \frac{2\gamma_x(d_x+d_y)^{\frac{1}{2}}}{3N_1} + \frac{2\gamma_t}{3N_2}. \end{aligned} \quad (\text{C.19})$$

Besides, by [Chen et al. \[2020, Lemma 15\]](#), for $l \gtrsim d_x + d_y$ and $\forall x', w', w'', t'$,

$$|\widehat{g}_i(x', w', t') - \widehat{g}_i(x', w'', t')| \lesssim (d_x + d_y) (\gamma_w + N_1 \|g_i\|_{L^\infty(\Omega)} 2^{-l+d_x+d_y}) \|w' - w''\|_\infty. \quad (\text{C.20})$$

Let $l = \mathcal{O}\left(d_x + d_y + \log \frac{\gamma_w(\|g\|_{L^\infty(\Omega)} + 1)}{\varepsilon}\right)$, $N_1 = \mathcal{O}\left(\frac{\gamma_x}{\varepsilon}\right)$, $N_2 = \mathcal{O}\left(\frac{\gamma_t}{\varepsilon}\right)$. Then

$$\|\widehat{g}_i - g_i\|_{L^\infty(\Omega)} \leq \varepsilon/2, \quad |\widehat{g}_i(x', w', t') - \widehat{g}_i(x', w'', t')| \lesssim \gamma_w(d_x + d_y) \|w' - w''\|_\infty. \quad (\text{C.21})$$

Define $\widehat{g} := [\widehat{g}_1, \dots, \widehat{g}_{d_x}]$ and $\widehat{f}(x, w, t) := \widehat{g} \left(\frac{x}{2R_f} + \frac{1}{2}, w, t/T \right)$. Then the approximation error of $\widehat{f}$ in Ω_{R_f} can be bounded by

$$\|\widehat{f} - f\|_{L^\infty(\Omega_{R_f})} \leq \|\widehat{g} - g\|_{L^\infty(\Omega)} \leq \sqrt{d_x} \varepsilon / 2, \text{ and } \widehat{f}(x, w, t) = 0, \forall \|x\|_\infty > R_f. \quad (\text{C.22})$$

Therefore, when $R_f \geq C_R \log^{\frac{1}{2}}((M_f^2 + C_L)/\varepsilon)$, the overall approximation error is

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathbb{P}} \mathbb{E}_{t, x_t | x} [\|f(x_t, h_*(y), t) - f_*^\mathbb{P}(x_t, h_*(y), t)\|^2] &\leq \mathbb{E}_{t, x_t, y} \|f(x_t, h(y), t) - f(x_t, h_*(y), t)\|^2 \cdot \mathbb{1}_{\|x_t\| \leq R_f} \\ &\quad + 4(M_f^2 + C_L) \exp(-C'_1 R_f^2) \\ &\leq \|f - f_*^\mathbb{P}\|_{L^\infty(\Omega_{R_f})}^2 + 4(M_f^2 + C_L) \exp(-C'_1 R_f^2) \\ &\leq d_x \varepsilon^2. \end{aligned} \quad (\text{C.23})$$

Now we characterize the configuration of neural network $\widehat{f}(x, w, t)$. For boundedness, by Lemma B.10,

$$\|\widehat{f}(x, w, t)\| \leq \|f_*\|_{L^\infty(\Omega_{R_f})} + \varepsilon \leq 2C_X'' R_f^6 =: M_f. \quad (\text{C.24})$$

Hence we can let $R_f = \mathcal{O} \left(\log^{\frac{1}{2}} \left(\frac{nK}{\varepsilon \delta} \right) \right)$ to ensure the lower bound of R_f mentioned above and in Theorem B.8. For Lipschitzness, by (C.21),

$$\begin{aligned} \|\widehat{f}(x, w, t) - \widehat{f}(x, \widetilde{w}, t)\| &\lesssim \gamma_w(d_x + d_y) \|w - \widetilde{w}\|_\infty \\ &\lesssim (C_X + C_X' R_f^2)(d_x + d_y) \|w - \widetilde{w}\|_\infty. \end{aligned} \quad (\text{C.25})$$

Hence $\gamma_f = \mathcal{O}((C_X + C_X' R_f^2)(d_x + d_y)) = \mathcal{O} \left(\log \left(\frac{nK}{\varepsilon \delta} \right) \right)$.

For the size of neural network, for each coordinate, by the construction in (C.18), the neural network $\widehat{g}_i$ consists of $N_1^{d_x+d_y} N_2$ parallel subnetworks, i.e., $g_i(u/N_1, j/N_2) \widehat{\Psi}_{u,j}(\cdot, \cdot, \cdot)$. By definition in (C.17), the subnetwork consists of $\mathcal{O} \left((d_x + d_y)(d_x + d_y + \log \frac{R_f}{\varepsilon}) \right)$ layers and the width is bounded by $\mathcal{O}(d_x + d_y)$. Therefore, the whole neural network $\widehat{g}_i$ can be implemented by $\mathcal{O}((d_x + d_y)(d_x + d_y + \log(R_f/\varepsilon)))$ layers with width $\mathcal{O}(N_1^{d_x+d_y} N_2(d_x + d_y)) = \mathcal{O} \left(\frac{R_f^{3(d_x+d_y)}}{\varepsilon^{d_x+d_y+1} T_0^3} \right)$, and the number of parameter is bounded by $\mathcal{O} \left(\frac{R_f^{3(d_x+d_y)} \log(R_f/\varepsilon)}{\varepsilon^{d_x+d_y+1} T_0^3} \right)$. Combine these arguments together, we can claim that the size of neural network $\widehat{f}$ is

$$\begin{aligned} L &= \mathcal{O}((d_x + d_y)(d_x + d_y + \log(R_f/\varepsilon))) = \mathcal{O} \left(\log \left(\frac{\log(nK/(\varepsilon \delta))}{\varepsilon} \right) \right), \\ W &= \mathcal{O} \left(\frac{R_f^{3(d_x+d_y)}}{\varepsilon^{d_x+d_y+1} T_0^3} \right) = \mathcal{O} \left(\frac{\log^{3(d_x+d_y)/2}(nK/(\varepsilon \delta))}{\varepsilon^{d_x+d_y+1} T_0^3} \right), \\ S &= \mathcal{O} \left(\frac{(d_x + d_y) R_f^{3(d_x+d_y)} \log(R_f/\varepsilon)}{\varepsilon^{d_x+d_y+1} T_0^3} \right) = \mathcal{O} \left(\frac{\log^{3(d_x+d_y)/2+1}(nK/(\varepsilon \delta))}{\varepsilon^{d_x+d_y+1} T_0^3} \right). \end{aligned} \quad (\text{C.26})$$

To bound of the neural network parameters, note that the trapezoid function ψ is rescaled by $3N_1$ or $3N_2$ and the weight parameter of ϕ_{mul}^l is bounded by a constant. Moreover, the input of $\widehat{f}$ is first rescaled by R_f or T . Hence we have

$$B = \mathcal{O}(N_1 R_f + N_2 T) = \mathcal{O} \left(\frac{R_f^3 T}{\varepsilon} \right) = \mathcal{O} \left(\frac{T \log^{\frac{3}{2}}(nK/(\varepsilon \delta))}{\varepsilon} \right), \quad (\text{C.27})$$

which concludes the proof. $\square$

Proposition C.3. *To achieve*

$$\inf_{h \in \mathcal{H}} \|h - h_*\|_{L^\infty([0,1]^{D_y})} \leq \sqrt{d_y} \varepsilon, \quad (\text{C.28})$$

the configuration of $\mathcal{H} = NN_h(L_h, W_h, S_h, B_h)$ should satisfy

$$\begin{aligned} L_h &= \mathcal{O}(\log(1/\varepsilon)), W_h = \mathcal{O}(\varepsilon^{-D_y} \log(1/\varepsilon)), \\ S_h &= \mathcal{O}(\varepsilon^{-D_y} \log^2(1/\varepsilon)), B_h = \mathcal{O}(1). \end{aligned} \quad (\text{C.29})$$

Here $\mathcal{O}(\cdot)$ hides all the polynomial factors of d_x, d_y, L .

Proof. The main idea replicates Yarotsky [2017, Theorem 1]. We approximate each coordinate of $h_* = [h_{*1}, \dots, h_{*d_y}]$ respectively and then concatenate all them together. By Yarotsky [2017, Theorem 1], h_{*i} can be approximated up to ε by a network $\hat{h}_i$ with $\mathcal{O}(\log(1/\varepsilon))$ layers and $\mathcal{O}(\varepsilon^{-D_y} \log(1/\varepsilon))$ width. Besides, the range of all the parameters are bounded by some constant, and the number of parameters is $\mathcal{O}(\varepsilon^{-D_y} \log^2(1/\varepsilon))$. Then we concatenate all the subnetworks to get $\hat{h} = [\hat{h}_1, \dots, \hat{h}_{d_y}]$ and $\|\hat{h} - h_*\|_{L^\infty([0,1]^{D_y})} \leq \sqrt{d_y} \varepsilon$. $\square$

C.2 Proofs of Distribution Estimation

Theorem C.4 (Thm. 4.2). *Suppose Assumption 3.1, 3.2, 3.3 hold. For sufficiently large integers n, K, m and $\delta > 0$, further suppose that $\mathbb{P}^1, \dots, \mathbb{P}^K$ are (ν, Δ) -diverse over target distribution $\mathbb{P}^0$ with proper configuration of neural network family and T, T_0 . It holds that with probability no less than $1 - \delta$,*

$$\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y^0} [\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0)] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3((m/\nu) \wedge n)}{\nu^{\frac{1}{2}}((m/\nu) \wedge n)^{\frac{1}{d_x+d_y+9}}} + \frac{\log^2(nK/\delta)}{\nu^{\frac{1}{2}}(nK)^{\frac{1}{D_y+2}}} + \sqrt{\Delta}. \quad (\text{C.30})$$

Proof. Combine Theorem C.1 and Theorem B.6 and plug in the configuration of $\mathcal{F}, \mathcal{H}$, we have with probability no less than $1 - \delta$

$$\begin{aligned} &\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{(x, y) \sim \mathbb{P}^0} [\ell^{\mathbb{P}^0}(x, y, s_{\hat{f}^{\mathbb{P}^0}}, \hat{h})] \\ &\lesssim \frac{1}{\nu} \log^2(nK/(\varepsilon\delta)) \varepsilon^2 + \Delta + \frac{\log^{\frac{3(d_x+d_y)+15}{2}}(nK/\varepsilon\delta) \log(T/T_0)}{(m \wedge (\nu n)) \varepsilon^{d_x+d_y+1} T_0^3} + \frac{\log^4(1/\varepsilon) \log(nK/(\varepsilon\delta))}{\nu n K \varepsilon^{D_y}} \end{aligned} \quad (\text{C.31})$$

By Lemma C.7,

$$\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0) \lesssim \sqrt{T_0} \log^{\frac{d_x+1}{2}}(1/T_0) + e^{-T} + \sqrt{\mathbb{E}_{\mathbb{P}_{x|y}^0}[\ell^{\mathbb{P}^0}(x, y, s_{\hat{f}^{\mathbb{P}^0}}, \hat{h})]} \quad (\text{C.32})$$

Taking expectation of $y, \hat{f}^{\mathbb{P}}, \mathbb{P}$, we have

$$\begin{aligned} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y^0} [\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0)] &\lesssim \sqrt{T_0} \log^{\frac{d_x+1}{2}}(1/T_0) + e^{-T} + \nu^{-\frac{1}{2}} \log(nK/(\varepsilon\delta)) \varepsilon + \sqrt{\Delta} \\ &\quad + \frac{\log^{\frac{3(d_x+d_y)+15}{4}}(\frac{nK}{\varepsilon\delta}) \log^{\frac{1}{2}}(\frac{T}{T_0})}{(m \wedge (\nu n))^{\frac{1}{2}} \varepsilon^{\frac{d_x+d_y+1}{2}} T_0^{\frac{3}{2}}} + \frac{\log^2(\frac{1}{\varepsilon}) \log^{\frac{1}{2}}(\frac{nK}{\varepsilon\delta})}{(\nu n K)^{\frac{1}{2}} \varepsilon^{\frac{D_y}{2}}}. \end{aligned} \quad (\text{C.33})$$

Let $T_0 = \mathcal{O}(\varepsilon_0^2 / \log^{d_x+1}(1/\varepsilon_0))$, $T = \mathcal{O}(\log(1/\varepsilon_0))$, $\varepsilon = \mathcal{O}(\varepsilon_0 / \log(nK/(\varepsilon_0\delta_0)))$ for some small $\varepsilon_0 > 0$ defined later. Then it reduces to

$$\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y^0} [\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0)] \lesssim \frac{\varepsilon_0}{\nu^{\frac{1}{2}}} + \sqrt{\Delta} + \frac{\log^{\frac{5(d_x+d_y)+17}{4}}(\frac{nK}{\varepsilon_0\delta}) \log^{\frac{3d_x+5}{2}}(\frac{1}{\varepsilon_0})}{(m \wedge (\nu n))^{\frac{1}{2}} \varepsilon_0^{\frac{d_x+d_y+7}{2}}} + \frac{\log^2(\frac{1}{\varepsilon_0}) \log^{D_y+\frac{1}{2}}(\frac{nK}{\varepsilon_0\delta})}{(\nu n K)^{\frac{1}{2}} \varepsilon_0^{\frac{D_y}{2}}}. \quad (\text{C.34})$$

Let $\varepsilon_0 = C \max \left\{ \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3((m/\nu) \wedge n)}{((m/\nu) \wedge n)^{\frac{1}{d_x+d_y+9}}}, \frac{\log^2(nK/\delta)}{(nK)^{\frac{1}{D_y+2}}} \right\}$, and we can conclude that

$$\mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y} [\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0)] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3((m/\nu) \wedge n)}{\nu^{\frac{1}{2}}((m/\nu) \wedge n)^{\frac{1}{d_x+d_y+9}}} + \frac{\log^2(nK/\delta)}{\nu^{\frac{1}{2}}(nK)^{\frac{1}{D_y+2}}} + \sqrt{\Delta}. \quad (\text{C.35})$$

Theorem C.5 (Thm. 4.3). *Suppose Assumption 3.1, 3.2, 3.3 hold. For sufficiently large integers n, K, m and $\delta > 0$, with proper configuration of neural network family and T, T_0 , it holds that with probability no less than $1 - \delta$,*

$$\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y} [\text{TV}(\hat{\mathbb{P}}_{x|y}, \mathbb{P}_{x|y})] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3(m \wedge n)}{(m \wedge n)^{\frac{1}{d_x+d_y+9}}} + \frac{\log^2(nK/\delta)}{K^{\frac{1}{D_y+2}}}. \quad (\text{C.36})$$

Proof. Combine Theorem C.1 and Theorem B.8 and plug in the configuration of $\mathcal{F}, \mathcal{H}$, we have with probability no less than $1 - \delta$

$$\begin{aligned} & \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell^{\mathbb{P}}(x, y, s_{\hat{f}^{\mathbb{P}}, \hat{h}})] \\ & \lesssim \log^2(nK/(\varepsilon\delta))\varepsilon^2 + \frac{\log^{\frac{3(d_x+d_y)+15}{2}}(nK/\varepsilon\delta) \log(T/T_0)}{(m \wedge n)^{\varepsilon^{d_x+d_y+1}} T_0^3} + \frac{\log^4(1/\varepsilon) \log(nK/(\varepsilon\delta))}{K \varepsilon^{D_y}} \end{aligned} \quad (\text{C.37})$$

By Lemma C.7,

$$\text{TV}(\hat{\mathbb{P}}_{x|y}, \mathbb{P}_{x|y}) \lesssim \sqrt{T_0} \log^{\frac{d_x+1}{2}}(1/T_0) + e^{-T} + \sqrt{\mathbb{E}_{\mathbb{P}_{x|y}}[\ell^{\mathbb{P}}(x, y, s_{\hat{f}^{\mathbb{P}}, \hat{h}})]} \quad (\text{C.38})$$

Taking expectation of $y, \hat{f}^{\mathbb{P}}, \mathbb{P}$, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y} [\text{TV}(\hat{\mathbb{P}}_{x|y}, \mathbb{P}_{x|y})] & \lesssim \sqrt{T_0} \log^{\frac{d_x+1}{2}}(1/T_0) + e^{-T} + \log(nK/(\varepsilon\delta))\varepsilon \\ & + \frac{\log^{\frac{3(d_x+d_y)+15}{4}}(\frac{nK}{\varepsilon\delta}) \log^{\frac{1}{2}}(\frac{T}{T_0})}{(m \wedge n)^{\frac{1}{2}} \varepsilon^{\frac{d_x+d_y+1}{2}} T_0^{\frac{3}{2}}} + \frac{\log^2(\frac{1}{\varepsilon}) \log^{\frac{1}{2}}(\frac{nK}{\varepsilon\delta})}{K^{\frac{1}{2}} \varepsilon^{\frac{D_y}{2}}}. \end{aligned} \quad (\text{C.39})$$

Let $T_0 = \mathcal{O}(\varepsilon_0^2 / \log^{d_x+1}(1/\varepsilon_0))$, $T = \mathcal{O}(\log(1/\varepsilon_0))$, $\varepsilon = \mathcal{O}(\varepsilon_0 / \log(nK/(\varepsilon_0\delta_0)))$ for some small $\varepsilon_0 > 0$ defined later. Then it reduces to

$$\begin{aligned} \mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y} [\text{TV}(\hat{\mathbb{P}}_{x|y}, \mathbb{P}_{x|y})] & \lesssim \varepsilon_0 + \frac{\log^{\frac{5(d_x+d_y)+17}{4}}(\frac{nK}{\varepsilon_0\delta}) \log^{\frac{3d_x+5}{2}}(\frac{1}{\varepsilon_0})}{(m \wedge n)^{\frac{1}{2}} \varepsilon_0^{\frac{d_x+d_y+7}{2}}} \\ & + \frac{\log^2(\frac{1}{\varepsilon_0}) \log^{D_y+\frac{1}{2}}(\frac{nK}{\varepsilon_0\delta})}{K^{\frac{1}{2}} \varepsilon_0^{\frac{D_y}{2}}}. \end{aligned} \quad (\text{C.40})$$

Let $\varepsilon_0 = C \max \left\{ \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3(m \wedge n)}{(m \wedge n)^{\frac{1}{d_x+d_y+9}}}, \frac{\log^2(nK/\delta)}{K^{\frac{1}{D_y+2}}} \right\}$, and we can conclude that

$$\mathbb{E}_{\mathbb{P} \sim \mathbb{P}_{\text{meta}}} \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^m} \mathbb{E}_{y \sim \mathbb{P}_y} [\text{TV}(\hat{\mathbb{P}}_{x|y}, \mathbb{P}_{x|y})] \lesssim \frac{\log^{\frac{5}{2}}(nK/\delta) \log^3(m \wedge n)}{(m \wedge n)^{\frac{1}{d_x+d_y+9}}} + \frac{\log^2(nK/\delta)}{K^{\frac{1}{D_y+2}}}. \quad (\text{C.41})$$

C.3 Auxiliary Lemmas

Lemma C.6. Let $\Omega_{R_f} = [-R_f, R_f]^{d_x} \times [0, 1]^{d_y} \times [T_0, T]$ for some $R_f \geq 1$. Then there exists some constant C_s , such that the score function $f_*^{\mathbb{P}}(x, w, t)$ is $\frac{C_s R_f^3}{T_0^3}$ -Lipschitz with respect to t in Ω_{R_f} .

Proof. According to (B.2),

$$f_*^{\mathbb{P}}(x, w, t) = -\frac{x}{\sigma_t^2} + \frac{\alpha_t}{\sigma_t^2} \int x_0 \frac{\phi_t(x|x_0)p(x_0;w)}{\phi_t(x|z)p(z;w)} dz. \quad (\text{C.42})$$

Define density function $q_t(x_0|x, w) \propto \phi_t(x|x_0)p(x_0;w)$. Then

$$\frac{\partial}{\partial t} f_*^{\mathbb{P}}(x, w, t) = -\frac{2\alpha_t^2 x}{\sigma_t^2} + \frac{\alpha_t}{\sigma_t^2} \text{Cov}_{q_t(x_0|x, w)} \left(x_0, \frac{\partial}{\partial t} \log \phi_t(x|x_0) \right) - \frac{\alpha_t(1 + \alpha_t^2)}{\sigma_t^4} \mathbb{E}_{q_t(x_0|x, w)}[x_0]. \quad (\text{C.43})$$

Note that

$$\begin{aligned} \text{Cov}_{q_t(x_0|x, w)} \left(x_0, \frac{\partial}{\partial t} \log \phi_t(x|x_0) \right) &= -\text{Cov}_{q_t(x_0|x, w)} \left(x_0, \frac{\partial}{\partial t} \frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2} \right) \\ &= \text{Cov}_{q_t(x_0|x, w)} \left(x_0, \frac{\alpha_t(x - \alpha_t x_0)^\top \mathbf{1}}{\sigma_t^2} - \frac{2\alpha_t^2 \|x - \alpha_t x_0\|^2}{\sigma_t^4} \right) \end{aligned} \quad (\text{C.44})$$

Hence for any $x \in [-R_f, R_f]^{d_x}, w \in [0, 1]^{d_y}$,

$$\left\| \frac{\partial}{\partial t} f_*^{\mathbb{P}}(x, w, t) \right\|_\infty \lesssim \frac{\alpha_t^2 R_f}{\sigma_t^2} + \mathbb{E}_{q_t(x_0|x, w)} \left\| \frac{x - \alpha_t x_0}{\sigma_t^2} \right\|^3 + \frac{\alpha_t(1 + \alpha_t^2)}{\sigma_t^4} \mathbb{E}_{q_t(x_0|x, w)}[\|x_0\|_\infty] \quad (\text{C.45})$$

Let $R = \frac{2R_f + 2C_0}{\sigma_t}$. We have

$$\begin{aligned} \mathbb{E}_{q_t(x_0|x, w)} \left\| \frac{\alpha_t x_0 - x}{\sigma_t^2} \right\|^3 &\leq \frac{1}{\sigma_t^3} \int \left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^3 \frac{\phi_t(x|x_0)p(x_0;w)}{\int \phi_t(x|z)p(z;w) dz} dx_0 \\ &\leq \frac{R^3}{\sigma_t^3} + \frac{\int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \geq R} \left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^2 \exp\left(-\frac{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^2}{2\sigma_t^2}\right) p(x_0;w) dx_0}{\sigma_t^3 \int \exp\left(-\frac{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\|^2}{2\sigma_t^2}\right) p(x_0;w) dx_0} \\ &\leq \frac{R^3}{\sigma_t^3} + \frac{\int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \geq R} \exp\left(-\frac{R^2}{4}\right) p(x_0;w) dx_0}{\sigma_t^3 \int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \leq R/2} \exp\left(-\frac{R^2}{8}\right) p(x_0;w) dx_0}. \end{aligned} \quad (\text{C.46})$$

The domain $\left\{x_0 : \left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \leq R/2\right\}$ includes $\left\{x_0 : \|x_0\| \leq C_0\right\}$, indicating

$$\begin{aligned} \int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \leq R/2} p(x_0;w) dx_0 &\geq \int_{\|x_0\| \leq C_0} p(x_0;w) dx_0 \geq 1 - 2\exp(-C_1' C_0^2) \geq \frac{1}{2}, \\ \int_{\left\| \frac{\alpha_t x_0 - x}{\sigma_t} \right\| \geq R} p(x_0;w) dx_0 &\leq \int_{\|x_0\| \geq C_0} p(x_0;w) dx_0 \leq \frac{1}{2}. \end{aligned} \quad (\text{C.47})$$

Therefore, for any $(x, w, t) \in \Omega_{R_f}$,

$$\left\| \frac{\partial}{\partial t} f_*^{\mathbb{P}}(x, w, t) \right\|_\infty \lesssim \frac{R_f^2}{\sigma_t^2} + \frac{R_f^3 + C_0^3}{\sigma_t^6} + \frac{R_f + C_0}{\sigma_t^3} \lesssim \frac{R_f^3}{T_0^3}. \quad (\text{C.48})$$

□

Lemma C.7. Suppose $\text{KL}(\mathbb{P}_{x|y}^0 \| \mathcal{N}(0, I)) \leq C_{\text{KL}}$ for some constant C_{KL} . Then

$$\text{TV}(\hat{\mathbb{P}}_{x|y}^0, \mathbb{P}_{x|y}^0) \lesssim \sqrt{T_0} \log^{\frac{d_x+1}{2}}(1/T_0) + e^{-T} + \sqrt{\mathbb{E}_{\mathbb{P}_{x|y}^0}[\ell^{\mathbb{P}^0}(x, y, s_{\hat{f}, \hat{h}})]}. \quad (\text{C.49})$$

Proof. With a little abuse of notation, we will use $p_t(x_t|y)$ to denote the conditional density of $x_t|y$ under $\mathbb{P}_{x|y}^0$. Consider the following two backward processes

$$d\tilde{x}_t = (\tilde{x}_t + 2\nabla \log p_{T-t}(\tilde{x}_t|y))dt + \sqrt{2}dW_t, \quad \tilde{x}_0 \sim \mathcal{N}(0, I), 0 \leq t \leq T - T_0, \quad (\text{C.50})$$

$$d\bar{x}_t = (\bar{x}_t + 2\nabla \log p_{T-t}(\bar{x}_t|y))dt + \sqrt{2}dW_t, \quad \bar{x}_0 \sim p_T, 0 \leq t \leq T - T_0. \quad (\text{C.51})$$

Denote the distribution of $\tilde{x}_t$ as $\tilde{\mathbb{P}}_{T-t}$. And note that $\bar{x}_t \sim p_{T-t}$ by classic reverse-time SDE results [Anderson, 1982]. Then by Fu et al. [2024, Lemma D.5],

$$\text{TV}(\mathbb{P}_{T_0}, \mathbb{P}_0) \lesssim \sqrt{T_0} \log^{\frac{d_x+1}{2}}(1/T_0). \quad (\text{C.52})$$

At the same time, we apply Data Processing inequality and Pinsker's inequality to get

$$\text{TV}(\mathbb{P}_{T_0}, \tilde{\mathbb{P}}_{T_0}) \leq \text{TV}(\mathbb{P}_T, \mathcal{N}(0, I)) \lesssim \sqrt{\text{KL}(\mathbb{P}_T \| \mathcal{N}(0, I))} \lesssim \sqrt{\text{KL}(\mathbb{P}_0 \| \mathcal{N}(0, I))} e^{-T}. \quad (\text{C.53})$$

Again according to Pinsker's inequality and Oko et al. [2023, Proposition D.1],

$$\text{TV}(\hat{\mathbb{P}}, \tilde{\mathbb{P}}_{T_0}) \lesssim \sqrt{\text{KL}(\tilde{\mathbb{P}}_{T_0} \| \hat{\mathbb{P}})} \lesssim \sqrt{\mathbb{E}_{x|y}[\ell^{\mathbb{P}}(x, y, s_{\hat{f}, \hat{h}})]}. \quad (\text{C.54})$$

Combine three inequalities above and we complete the proof. $\square$

D Proofs in Section A

D.1 Proof of Theorem A.1

Proof. Due to the structure of exponential family, Assumption 3.2 holds obviously. To apply previous results, we only need to verify Assumption 3.1 and 3.3. Recall that a basic property of exponential family is

$$\nabla_x A_\psi(x) = \mathbb{E}_{p_\psi(y|x)}[h_*(y)] \in [0, 1]^d, \quad (\text{D.1})$$

$$0 \preceq \nabla_x^2 A_\psi(x) = \text{Var}_{p_\psi(y|x)}(h_*(y)) \preceq I. \quad (\text{D.2})$$

Hence by Assumption A.1, $A_\psi(x) \leq A_\psi(0) + \|x\|_1 \leq \log \left(\int \psi(y) dy \right) + \|x\|_1 \leq \log C + \|x\|_1$. And $A_\psi(x) \geq A_\psi(0) - \|x\|_1 \geq -\log C - \|x\|_1$. Further note that the posterior density $p_\theta(x|y) = \frac{p_\phi(x) \exp(\langle x, h_*(y) \rangle - A_\psi(x))}{Z_\theta}$, where the normalizing constant $Z_\theta(y)$ is lower bounded by

$$\begin{aligned} Z_\theta(y) &= \int p_\phi(x) \exp(\langle x, h_*(y) \rangle - A_\psi(x)) dx \\ &\geq \int p_\phi(x) \exp(-2\|x\|_1)/C dx \\ &\geq \exp(-2\sqrt{d}R)(1 - 2\exp(-C'_1 R^2))/C =: C_0. \end{aligned} \quad (\text{D.3})$$

where in the second inequality we apply $\mathbb{P}_\phi(\|x\| \geq R) \leq 2\exp(-C'_1 R^2)$ and let $R = 1/\sqrt{C'_1}$ to get C_0 . Therefore, by Assumption A.1,

$$p_\theta(x|y) \leq C_1 \exp(-C_2\|x\|^2 + 2\|x\|_1 + \log C)/C_0 \leq C'_1 \exp(-C'_2\|x\|^2), \quad (\text{D.4})$$

and thus Assumption 3.1 holds. At the same time, let $w = h_*(y)$, then the score function is

$$\nabla_x \log p_\theta(x|y) = \nabla_x \log p_\theta(x, w) = \nabla_x \log p_\phi(x) + w - \nabla_x A_\psi(x). \quad (\text{D.5})$$

Since $\nabla_x \log p_\phi(x)$ is L -Lipschitz, $\nabla A_\psi(x)$ is also 1-Lipschitz, the score function $\nabla_x \log p_\theta(x, w)$ is $(L+1)$ -Lipschitz in x and 1-Lipschitz in w . And $\|\nabla_x \log p_\theta(0, w)\| \leq \|\nabla_x \log p_\phi(0)\| + 2\sqrt{d} = B + 2\sqrt{d}$, indicating that Assumption 3.3 holds with $L' = L+1$, $B' = B + 2\sqrt{d}$.

We conclude the proof by applying Theorem 4.3 under meta-learning setting or Theorem 4.2 under (ν, Δ) -diversity. $\square$

D.2 Proof of Theorem A.2

Proof. Let $A_M^\pi(s, a) = Q_M^\pi(s, a) - V_M(\pi, s)$ be the advantage function of policy π . Note that the reward function $r_M \in [0, 1]$, we have $|A_M^\pi(s, a)| \leq \frac{2}{1-\gamma}$ for any M, π . According to performance difference lemma,

$$\begin{aligned} V_{M^0}(\pi_*^0) - V_{M^0}(\hat{\pi}^0) &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_*^0} [A_{M^0}^{\hat{\pi}^0}(s, a)] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_*^0} \left[\mathbb{E}_{a \sim \pi_*^0(\cdot|s)} [A_{M^0}^{\hat{\pi}^0}(s, a)] - \mathbb{E}_{a \sim \hat{\pi}^0(\cdot|s)} [A_{M^0}^{\hat{\pi}^0}(s, a)] \right] \quad (\text{D.6}) \\ &\leq \frac{2}{(1-\gamma)^2} \mathbb{E}_{s \sim d_*^0} [\text{TV}(\pi_*^0(\cdot|s), \hat{\pi}^0(\cdot|s))]. \end{aligned}$$

Hence in meta-learning setting, we plug in Theorem 4.3 to obtain

$$\mathbb{E}_{M^0} \mathbb{E}_{\{(s_i^0, a_i^0)\}_{i=1}^m \sim d_*^0} [V_{M^0}(\pi_*^0) - V_{M^0}(\hat{\pi}^0)] \lesssim \frac{1}{(1-\gamma)^2} \left[\frac{\log^{\frac{5}{2}}(nK/\delta) \log^3(m \wedge n)}{(m \wedge n)^{\frac{1}{d_a+d_s+9}}} + \frac{\log^2(nK/\delta)}{K^{\frac{1}{D_s+2}}} \right]. \quad (\text{D.7})$$

If we further assume (ν, Δ) -diversity holds, then we plug in Theorem 4.2,

$$\mathbb{E}_{\{(s_i^0, a_i^0)\}_{i=1}^m \sim d_*^0} [V_{M^0}(\pi_*^0) - V_{M^0}(\hat{\pi}^0)] \lesssim \frac{1}{(1-\gamma)^2} \left[\frac{\log^{\frac{5}{2}}(nK/\delta) \log^3((m/\nu) \wedge n)}{\nu^{\frac{1}{2}}((m/\nu) \wedge n)^{\frac{1}{d_a+d_s+9}}} + \frac{\log^2(nK/\delta)}{\nu^{\frac{1}{2}}(nK)^{\frac{1}{D_s+2}}} + \sqrt{\Delta} \right]. \quad (\text{D.8})$$

□

E Experiment Details

E.1 Conditioned Diffusion

Each f^k and f^0 are implemented as a 2-layer MLP with 128 internal channels and 60 input channels. The representation map h is implemented as a 5-layer MLP with 512 internal channels and 10 output channels. We have $n = 1000$ pre-training samples from each source distribution $\mathbb{P}^k$, $m \in \{10, 20, 30, 40, 50, 100\}$ fine-tuning samples from the target distribution $\mathbb{P}^0$. We run Langevin Monte Carlo for sufficiently long time to obtain 100 test samples from the target distribution $\mathbb{P}^0$ for evaluating the test error of different models. In the pre-training phase, the $\{\hat{f}^k; 1 \leq k \leq K\}$ and $\hat{h}$ are trained on the $K = 10$ source distributions with 400K iterations and a batch size of 512. In the fine-tuning phase, the pre-trained representation map $\hat{h}$ is fixed, and the $\hat{f}^0$ is trained on the target distribution with 200K iterations and a batch size of m . As an important baseline, we also consider jointly training h and f^0 on the target distribution from scratch, using the same fine-tuning samples.

E.2 Image Restoration on MNIST

Each f^k and f^0 are implemented as a 3-layer MLP with 512 internal channels and 784 input channels. The representation map h is implemented as a 5-layer MLP with 512 internal channels and 64 output channels. We have $n = 5000$ pre-training samples from each source distribution $\mathbb{P}^k$, and $m \in \{10, 20, 30, 40, 50, 100\}$ fine-tuning samples from the target distribution $\mathbb{P}^0$. For evaluation, we directly compute the mean squared error between the posterior samples and the ground truth images, based on 100 test samples from $\mathbb{P}^0$. In the pre-training phase, the $\{\hat{f}^k; 1 \leq k \leq K = 9\}$ and $\hat{h}$ are 2K epochs and a batch size of 512. The initial learning rate is 0.0003 and is annealed according to a cosine annealing schedule. In the fine-tuning phase, the pre-trained representation map $\hat{h}$ is fixed, and the $\hat{f}^0$ is trained on the target distribution with 20K iterations and a batch size of m . As an important baseline, we also consider jointly training h and f^0 on the target distribution from scratch, using the same fine-tuning samples.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction accurately reflect the paper's contributions, i.e., proposing a data-efficient training method for machine learning models.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper discusses the limitations of the work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The paper provides the full set of assumptions and complete (and correct) proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper fully discloses all the information needed to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will provide complete codes upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all the training and test details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports experimental results based on the average of independent random trials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides sufficient information on the computer resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets currently.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.