Towards a Statistical Theory of Learning to Learn In-context with Transformers

Youssef Mroueh
IBM Research
mroueh@us.ibm.com

Abstract

Classical learning theory focuses on supervised learning of functions via empirical risk minimization where labeled examples for a particular task are represented by the data distribution experienced by the model during training. Recently, in-context learning emerged as a paradigm shift in large pre-trained models. When conditioned with few labeled examples of potentially unseen tasks in the training, the model infers the task at hand and makes predictions on new points. Learning to learn in-context on the other hand, aims at training models in a meta-learning setup that generalize to new unseen tasks from only few shots of labeled examples. We present in this paper a statistical learning framework for the problem of in-context meta learning and define a function class that enables it. The meta-learner is abstracted as a function defined on the cross product of the probability space (representing "context") and the data space. The data distribution is sampled from a "meta distribution" on tasks. Thanks to the regularity we assume on the function class in the Wasserstein geometry, we leverage tools from optimal transport in order to study the generalization of the meta learner to unseen tasks. Finally, we show that encoder transformers exhibit this type of regularity and leverage our theory to analyze their generalization properties.

1 Introduction

Since their introduction transformer models [Vaswani et al., 2017] have reshaped the AI landscape and unveiled the power of Large Language Models (LLMs)[Bommasani et al., 2021]. Large Languages models such as GPT-J [Lieber et al., 2021], GPT2 [Radford et al., 2019], GPT3 [Brown et al., 2020], are generative pretrained transformers that are trained on massive text data on the web scale. The training objective of these models is next token prediction given previous tokens. Later these models were extended to the multi-modal setting thanks to massive interleaved text and image datasets for example the open source dataset [Zhu et al., 2023]. Such massive heterogeneous diverse and multimodal datasets enabled training Flamingo [Alayrac et al., 2022], open Flamingo [Awadalla et al., 2023] and GPT4 [OpenAI, 2023].

Brown et al. [2020] showed the emergence of *in-context learning* in such models. In this new paradigm, LLMs learn a new task without fine-tuning or learning by simply conditioning on a prompt containing a sequence of few shots of input/output pairs. For example for a sentiment analysis task, independent "in-context examples" are concatenated (e.g. "joyful positive, bad negative") with a test example ("happy"), to form a prompt on which an LLM is conditioned and the next token predicted by the LLM has high probability on the correct answer that is in this example "positive". Multiple works have been devoted to explaining this emergent behavior in decoder transformers and large LLMs, for example [Xie et al., 2022] explained it as an implicit bayesian inference performed by the LLM in discovering the latent concept of the task at hand, under the assumption of a Hidden Markov Model (HMM); [Hahn and Goyal, 2023] explained it as an implicit induction and [Wies et al., 2023] studied the PAC learnability of in-context learning.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

On the other hand, meta in-context learning and in-context tuning were recently introduced for decoder transformers in [Chen et al., 2021] [Min et al., 2022] where a single model is conditioned on in-context examples coming from different tasks and is trained to perform prediction on a test example. In a sense the model learns to learn in-context and the task is only inferred from the few shots presented to the model. A parallel line of works [Kirsch et al., 2022] [Müller et al., 2022] demonstrated that encoder transformers can be general purpose in-context meta learners and can do implicit bayesian inference. In this work we answer the following questions: *How do we formalize the problem of learning to learn in-context ? What kind of regularity on the function class considered in meta-learning will allow an interpretable control of the generalization of in-context meta-learning to unseen tasks?*

The main contributions of this paper are: (1) We introduce in Section 3 an idealized statistical learning framework for learning to learn in-context and relax it in Section 4 to the few shot setting (2) We show that under regularity in the Wasserstein geometry of in-context learners, we can bound the generalization of meta in-context learning with tools from optimal transport. Interestingly, the batched Wasserstein introduced for computational reasons in [Sommerfeld et al., 2019] [Fatras et al., 2020] [Fatras et al., 2021] appears in our bounds. (3) We show in Section 5 that encoder transformers satisfy this regularity and leverage our framework to analyze their in-context meta-learning, and their generalization properties. Our results are in line with empirical findings in [Kirsch et al., 2022] [Müller et al., 2022].

2 Preliminaries on Optimal Transport

Let $(\mathcal{X}, d_{\mathcal{X}})$, and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two compact metric spaces of finite diameters $\text{diam}(\mathcal{X})$ and $\text{diam}(\mathcal{Y})$ respectively. We assume $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, for $z, z' \in \mathcal{Z}$, $z = (x, y)$ and $z' = (x', y')$ define $d_{\mathcal{Z}}(z, z') = \sqrt{d^2_{\mathcal{X}}(x, x') + d^2_{\mathcal{Y}}(y, y')}$. $(\mathcal{Z}, d_{\mathcal{Z}})$ is a compact metric space. We denote $\mathcal{P}(\mathcal{Z})$ the set of Borel probability measures on $\mathcal{Z}$. We endow $\mathcal{P}(\mathcal{Z})$ with the Wasserstein distance of order 1 between two measures $\rho, \rho' \in \mathcal{P}(\mathcal{Z})$ defined on the metric space $(\mathcal{Z}, d_Z)$ as follows: $W_1(\rho, \rho') = \min_{\gamma \in \Gamma(\rho, \rho')} \int_{\mathcal{Z} \times \mathcal{Z}} d_{\mathcal{Z}}(z, z') d\gamma(z, z')$, where the minimum is taken over the set $\Gamma(\rho, \rho')$ of all couplings γ between ρ and ρ'. We equip $\mathcal{P}(\mathcal{Z})$ with the topology induced by W_1 and denote $\mathcal{P}(\mathcal{P}(\mathcal{Z}))$ the set of corresponding Borel probability measures on $\mathcal{P}(\mathcal{Z})$. A measure $\mathbb{P} \in \mathcal{P}(\mathcal{P}(\mathcal{Z}))$ is a distribution of distributions, i.e $\rho \sim \mathbb{P}$ is a Borel probability measure on $\mathcal{P}(\mathcal{Z})$.

3 Learning to Learn In-Context: Idealized Setting

In learning to learn in-context, or in general in meta learning we assume having access to multiple datasets defined on the same domain and multiple tasks that we want a single model f to solve. The task at hand will be inferred by the model f from only seeing few shots from the task it is asked to perform predictions on. These few shots are often referred to as demonstrations or context. This setup is different from the classical multi-task setup where a different model or a fine-tuned model is dedicated to each task. A single model that has the capacity to perform multiple tasks at once and to do in-context learning is referred to as a general purpose meta-learner [Kirsch et al., 2022].

Formally speaking, given k pairs of inputs/label examples of an eventually unseen task in meta-training $z_i = (x_i, y_i) \in \mathcal{Z}, i = 1 \dots k$, the meta model f is asked to predict the label $y \in \mathcal{Y}$ of a point $x \in \mathcal{X}$. Let $\hat{\rho}_k = \frac{1}{k} \sum_{i=1}^{k} \delta_{z_i}$, we formalize this problem as finding a predictor f, defined as follows :

$$\begin{aligned} f \colon \mathcal{P}(\mathcal{Z}) \times \mathcal{X} &\longrightarrow \mathcal{Y} \\ (\rho, x) &\longmapsto f(\rho, x), \end{aligned}$$

and the goal of the meta-training is to ensure that $f(\hat{\rho}_k, x) \approx y$, even if $\hat{\rho}_k$ does not correspond to a task seen in training. Let $\mathbb{P} \in \mathcal{P}(\mathcal{P}(\mathcal{Z}))$, be a mother distribution of tasks. Each measure $\rho \sim \mathbb{P}$ defines a task on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. In practice these tasks correspond to different datasets whose input domain is a subset of $\mathcal{X}$ and label or structured output domain is a subset of $\mathcal{Y}$. Let V be a loss that measures the fitness of the prediction of the meta-model f, $V \colon \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}^+$.

We are now ready to define the expected risk for learning to learn in-context :

$$\mathcal{E}^V_{\mathbb{P}}(f) = \int_{\mathcal{P}(\mathcal{Z})} \int_{\mathcal{Z}} V(y, f(\rho, x)) d\rho(x, y) d\mathbb{P}(\rho). \tag{1}$$

This expected risk is to be contrasted with the classicial expected risk in learning theory [Vapnik, 1995]:

$$\mathcal{E}^V_{\rho}(h) = \int_{\mathcal{Z}} V(y, h(x)) d\rho(x, y) \tag{2}$$

where a specialized function h is associated with each learning task ρ. A different setup from ours is the setup of multi-task learning, where we are given T tasks and the goal is to learn specialized functions $h_t \circ g$ that share a common representation g (see for e.g [Maurer et al., 2016] and references therein):

$$\mathcal{E}^V_{\rho}(h \circ g) = \frac{1}{T} \sum_{t=1}^{T} \int_{\mathcal{Z}} V(y, h_t \circ g(x)) d\rho_t(x, y) \tag{3}$$

And finally, we emphasize that our setup is also different from the distribution regression [Szabó et al., 2016, Poczos et al., 2013, Christmann and Steinwart, 2010] setup that learns function $f : \mathcal{P}(\mathcal{X}) \to \mathcal{Y}$, given a distribution $\mathbb{P} \in \mathcal{P}(\mathcal{P}(\mathcal{X}) \times \mathcal{Y})$:

$$\mathcal{E}^V_{\mathbb{P}}(f) = \int_{\mathcal{P}(\mathcal{X}) \times \mathcal{Y}} V(y, f(\rho)) d\mathbb{P}(\rho, y). \tag{4}$$

We refer to the new risk we define in Equation (1), as the lifted expected risk as it is lifted to the space of distribution of distributions, and a single meta model f learns all tasks at once. The introduction of the lifted risk alleviates the shortcomings of classical learning theory expected risk in i) dealing with heterogeneous data ii) allowing a single model to learn multiple tasks. For $d_{\mathcal{Y}}$ being the euclidean distance, consider $V(y, y') = ||y - y'||_2^2$[1], we see that the optimal function minimizing the objective in (1) (without any restriction on the function class except integrability $\int |f(x, \rho)|^2 d\rho_{\mathcal{X}}(x) d\mathbb{P}(\rho) < \infty$, where $\rho_{\mathcal{X}}$ is the marginalization on y of ρ) is the regression function

$$f^*(\rho, x) = \int y d\rho(y|x), \rho_{\mathcal{X}}\mathbb{P} \text{ almost surely,} \tag{5}$$

In order to estimate the regression function, a model needs to infer the right context "ρ" to perform prediction, this is in line with the implicit bayesian inference argument explored in [Xie et al., 2022][Müller et al., 2022].

Given a function class $\mathcal{F}$ of functions from $\mathcal{P}(\mathcal{Z}) \times \mathcal{X} \to \mathcal{Y}$ our goal is to solve for:

$$\min_{f \in \mathcal{F}} \mathcal{E}^V_{\mathbb{P}}(f)$$

We make the following assumptions on the lipschitzty of V and the function class $\mathcal{F}$ of f:

Assumption 1. *Assume that $V(0, 0) < \infty$ and for all (y_1, y_1') and $(y_2, y_2') \in \mathcal{Y} \times \mathcal{Y}$, there exists $L_V > 0$ such that*

$$|V(y_1, y_1') - V(y_2, y_2')| \leq L_V \sqrt{d^2_{\mathcal{Y}}(y_1, y_2) + d^2_{\mathcal{Y}}(y_1', y_2')}$$

Assumption 1 is for e.g satisfied for $V(y, y') = \varphi(d_{\mathcal{Y}}(y, y'))$ for a lipschitz loss φ.

Assumption 2. *The predictor $f \in \mathcal{F}$, satisfies for all $\rho, \rho' \in \mathcal{P}(\mathcal{Z})$, and $x, x' \in \mathcal{X}$:*

$$d_{\mathcal{Y}}(f(\rho, x), f(\rho', x')) \leq L_{\mathcal{F}} \left(d_{\mathcal{X}}(x, x') + W_1(\rho, \rho') \right).$$

Assumption 2 is a classical assumption in analysing diffusion processes see for e.g [Funaki, 1984].

In the reminder of this Section we aim at answering two questions:

1. Given a model trained using the lifted expected risk defined on a pre-training or a meta-training mother distribution $\mathbb{P} \in \mathcal{P}(\mathcal{P}(\mathcal{Z}))$, under which condition does the learning transfer to unseen tasks modeled with another mother distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{P}(\mathcal{Z}))$?

[1] On bounded domain $\mathcal{Y}$ this loss satisfies Assumption 1 with lipschitz constant $2\text{diam}(\mathcal{Y})$.

2. What is the sample complexity of learning with the lifted risk given $(\rho_i)_{1\leq i\leq m} \sim \mathbb{P}^{\otimes m}$, and $(z_{ij} = (x_{ij}, y_{ij}))_{1\leq j\leq n} \sim \rho_i$?

In order to answer the first question, we need a metric on $\mathcal{P}(\mathcal{P}(\mathcal{Z}))$ in order to compare $\mathbb{P}$ and $\mathbb{Q}$. Following [Carlier et al., 2023], we define the following lifted Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$:

$$\mathcal{W}_{\mathbf{D}}(\mathbb{P}, \mathbb{Q}) = \min_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{P}(\mathcal{Z})\times\mathcal{P}(\mathcal{Z})} \mathbf{D}(\rho, \rho')d\gamma(\rho, \rho'), \tag{6}$$

where $\mathbf{D}$ is a metric on $\mathcal{P}(\mathcal{Z})$. In Carlier et al. [2023], $\mathbf{D}$ is considered to be the W_2 distance as a cost of the optimal transport in (6). We will see that different learning problems we consider in this paper will result in different transport cost $\mathbf{D}$ of interest in the lifted Wasserstein distance given in Equation (6).

Theorem 1 (Transferability of in-context Learning)**.** *For any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(Z))$, under Assumptions 1 and 2, for all $f \in \mathcal{F}$ we have:*

$$\left|\mathcal{E}^V_{\mathbb{Q}}(f) - \mathcal{E}^V_{\mathbb{P}}(f)\right| \leq L\, \mathcal{W}_{\mathbf{W_1}}(\mathbb{P}, \mathbb{Q}),$$

where $L = L_V(1 + L_{\mathcal{F}})(1 + \sqrt{2})$, and $\mathcal{W}_{W_1}$ is the lifted Wasserstein distance $\mathcal{W}_{\mathbf{D}}$, for $\mathbf{D}$ being the Wasserstein 1 distance W_1 on $\mathcal{P}(\mathcal{Z})$.

Let $f_{\mathbb{P}} \in \mathcal{F}$ be a minimizer of $\mathcal{E}^V_{\mathbb{P}}(f)$, Theorem 1 implies that the expected risk of $f_{\mathbb{P}}$ on unseen tasks sampled from a mother distribution $\mathbb{Q}$ is bounded by the risk of $f_{\mathbb{P}}$ on the pre-training distribution $\mathbb{P}$ and the lifted Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$. In order to ensure transfer learning to new unseen tasks, we need to have $\mathcal{W}_{\mathbf{W_1}}(\mathbb{P}, \mathbb{Q}) \leq \varepsilon/L$, so that : $\mathcal{E}^V_{\mathbb{Q}}(f_{\mathbb{P}}) \leq \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}) + \varepsilon$.

Lifted Empirical Risk Minimization We turn now to the problem of estimation of the lifted expected risk (Equation (1)) from samples. Let $(\rho_i)_{1\leq i\leq m} \sim \mathbb{P}^{\otimes m}$, $(z_{ij} = (x_{ij}, y_{ij}))_{1\leq j\leq n} \sim \rho_i$. We note $\hat{\mathbb{P}}_m = \frac{1}{m}\sum_{i=1}^m \delta_{\rho_i}$, and $\hat{\rho}^n_i = \frac{1}{n}\sum_{j=1}^n \delta_{z_{ij}}$ and lastly $\hat{\mathbb{P}}^n_m = \frac{1}{m}\sum_{i=1}^m \delta_{\hat{\rho}^n_i}$. The *lifted Empirical Risk* is therefore:

$$\mathcal{E}^V_{\hat{\mathbb{P}}^n_m}(f) = \frac{1}{m}\frac{1}{n}\sum_{i=1}^m\sum_{j=1}^n V(y_{ij}, f(\hat{\rho}^n_i, x_{ij})) \tag{7}$$

Remark 1. *This setup is not of practical interest, since the predictor sees all the empirical distribution, including the point on which the prediction is performed. The next section will introduce a more practical setup.*

Let $\widehat{f}_{\mathbb{P}^n_m}$ be a minimizer of $\mathcal{E}^V_{\hat{\mathbb{P}}^n_m}(.)$ that we assume exists in $\mathcal{F}$. In order to derive sample complexity bounds we need to make assumptions on the structure of the meta-distribution $\mathbb{P}$, as well as a structure that holds on all probability measures ρ, $\mathbb{P}$ almost surely.

For $\mathbb{P}$, we use the notion of upper Wasserstein distance introduced in [Weed and Bach, 2019].

Definition 1 (Upper Wasserstein Dimension [Weed and Bach, 2019])**.** *Given a metric space (X, d). Given a measure μ on X , the (ε, τ) covering number of μ in (X, d) is:*

$$\mathcal{N}_\varepsilon(\mu, \tau) = \inf\{\mathcal{N}_\varepsilon(S) : \mu(S) \geq \tau\},$$

where $\mathcal{N}_\epsilon(S)$ is the ε covering number of S, and the (ε, τ) dimension is :

$$d_\varepsilon(\mu, \tau) := \frac{\mathcal{N}_\varepsilon(\mu, \tau)}{-\log(\varepsilon)}.$$

The upper Wasserstein distance is defined as follows:

$$d^*_p(\mu) = \inf\{s \in (2p, \infty) : \lim_{\varepsilon\to 0}\sup d_\varepsilon(\mu, \varepsilon^{\frac{sp}{s-2p}}) \leq s\}$$

Assumption 3. *We assume that $\mathbb{P}$ satisfies in the metric space $(\mathcal{P}(\mathcal{P}(Z)), W_1)$ a bounded upper Wasserstein dimension: $d^*_1(\mathbb{P}) \leq s$.*

For $\rho \sim \mathbb{P}$ we assume that these distributions are clusterable in the following sense defined in [Weed and Bach, 2019]:

Definition 2 (Clusterable Distribution [Weed and Bach, 2019]). *A distribution is* (q, Δ) *if supp*(μ) *lies in the union of* q *balls of radius at most* Δ.

Assumption 4. *We assume that all* $\rho \sim \mathbb{P}$ *are* (q, Δ)*-clusterable .*

Theorem 2 (Sample complexity of in-context learning). *Under Assumptions 1 and 2, the following bound holds for the generalization of in-context learning:*

$$\mathbb{E}\left(\mathcal{E}^V_{\mathbb{P}}(\widehat{f}_{\mathbb{P}^n_m})\right) \leq \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}) + 2L\ \left(\mathbb{E}\mathcal{W}_{\mathbf{W_1}}(\mathbb{P}, \mathbb{P}_m) + \mathbb{E}\mathcal{W}_{\mathbf{W_1}}(\mathbb{P}_m, \mathbb{P}^n_m)\right))$$

Under Assumptions 3 and 4, we have for any $m \geq 1$ *and* $n \leq q(2\Delta)^{-2}$*:*

$$\mathbb{E}\left(\mathcal{E}^V_{\mathbb{P}}(\widehat{f}_{\mathbb{P}^n_m})\right) \lesssim \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}) + 2L\left(m^{-\frac{1}{s}} + 12\sqrt{\frac{q}{n}}\right).$$

Without Assumption 4 the learning is cursed in dimension and we have instead of $\sqrt{\frac{q}{n}}$ a rate of $n^{-\frac{1}{d_x+d_y}}$. For the squared euclidean loss, if we assume that f^* defined in Equation 5 $\in \mathcal{F}$, we have $f_{\mathbb{P}} = f^*$, and we see that as n the number of in-context examples grows and the number of tasks m grows we recover the bayesian inference regression function f^*, and hence the estimator $\widehat{f}_{\mathbb{P}^n_m}$ performs implicit bayesian inference.

4 Learning to Learn In Context: Few Shots or Batched Setting

One crucial shortcoming of the theory presented in the previous Section, is that we require in the lifted ERM to see all samples from context distributions, whereas in practice we only see few shots, i.e. only k shots or demonstrations. Hence we define the k- shots expected risk as follows:

$$\begin{aligned}\mathcal{R}^V_{k,\mathbb{P}}(f) &= \mathbb{E}_{\rho\sim\mathbb{P}}\mathbb{E}_{z=(x,y)\sim\rho}\mathbb{E}_{z_1,\dots z_k\sim\rho} V\left(y, f\left(\frac{1}{k}\sum_{\ell=1}^k \delta_{z_\ell}, x\right)\right)\\ &= \int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}} V(y, f(\hat{\rho}_k, x))d\rho(x,y)\Pi_{\ell=1}^k d\rho(x_\ell, y_\ell)d\mathbb{P}(\rho) \qquad (8)\end{aligned}$$

where $\hat{\rho}_k = \frac{1}{k}\sum_{\ell=1}^k \delta_{z_\ell}$, $(z_\ell = (x_\ell, y_\ell))_{1\leq\ell\leq k} \sim \rho^{\otimes k}$. Intuitively as $k \to \infty$, $\mathcal{R}^V_{k,\mathbb{P}}(f)$ recovers $\mathcal{E}^V_{\mathbb{P}}(f)$.

Lemma 1. *Under Assumptions 1 and 2 we have:*

$$\left|\mathcal{R}^V_{k,\mathbb{P}}(f) - \mathcal{E}^V_{\mathbb{P}}(f)\right| \leq L_V(1+L_{\mathcal{F}})\mathbb{E}_{\rho\sim\mathbb{P}}\mathbb{E}_{z_\ell\sim\rho^{\otimes k}} W_1(\hat{\rho}_k, \rho) \lesssim L_V(1+L_{\mathcal{F}})k^{-\frac{1}{d_x+d_y}},$$

and $\lim_{k\to\infty}\mathcal{R}^V_{k,\mathbb{P}}(f) = \mathcal{E}^V_{\mathbb{P}}(f)$.

Define the batched Wasserstein [Sommerfeld et al., 2019] [Fatras et al., 2020] [Fatras et al., 2021]:

$$W_1^{k_1,k_2}(\rho,\rho') = \mathbb{E}_{\substack{Z_1\dots Z_{k_1}\sim\rho\\ Z'_1\dots Z'_{k_2}\sim\rho'}}\left[W_1\left(\frac{1}{k_1}\sum_{i=1}^{k_1}\delta_{Z_i}, \frac{1}{k_2}\sum_{i=1}^{k_2}\delta_{Z'_i}\right)\right], \qquad (9)$$

Note that this is a not a metric, since $W_1^{k_1,k_2}(\rho,\rho) \geq 0$, and this quantity is related to the notion of $k-$ variance of ρ introduced in [Solomon et al., 2022].

We study first transferability in this setting:

Theorem 3 (Transferability of in few shot Learning). *For any* $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(Z))$*, under Assumptions 1 and 2, for all* $f \in \mathcal{F}$ *we have:*

$$\left|\mathcal{R}^V_{k_1,\mathbb{P}}(f) - \mathcal{R}^V_{k_2,\mathbb{Q}}(f)\right| \leq \sqrt{2}L_V(1+L_{\mathcal{F}})\ \mathcal{W}_{\mathbf{W_1}+\mathbf{W_1^{k_1,k_2}}}(\mathbb{P},\mathbb{Q}), \qquad (10)$$

where $\mathcal{W}_{\mathbf{W_1}+\mathbf{W_1^{k_1,k_2}}}$ *is the lifted Wasserstein distance* $\mathcal{W}_{\mathbf{D}}$ *(defined in Equation (6)) for* $\mathbf{D}$ *being the sum of the Wasserstein 1 distance* W_1 *and the batched Wasserstein* W_{k_1,k_2} *defined in Equation (9), i.e. for* $\mathbf{D}(\rho,\rho') = W_1(\rho,\rho') + W_1^{k_1,k_2}(\rho,\rho')$.

We see that the number of shots k_1 and k_2 goes to infinity in Theorem 3, we recover up to constants the stability Theorem 1 of the idealized setting. In the few shot setting, the transfer learning between two meta distributions of tasks is bounded by a stricter distance than the idealized setting since it takes into accounts the batched Wasserstein that measures the closeness of the distribution when seeing only few samples. It is easy to see that :

$$\mathcal{W}_{\mathbf{W_1}+\mathbf{W_1^{k_1,k_2}}}(\mathbb{P},\mathbb{Q}) \geq \mathcal{W}_{\mathbf{W_1}}(\mathbb{P},\mathbb{Q}) + \mathcal{W}_{\mathbf{W_1^{k_1,k_2}}}(\mathbb{P},\mathbb{Q}) \geq \mathcal{W}_{\mathbf{W_1}}(\mathbb{P},\mathbb{Q}).$$

This means that for transfer learning to be feasible in the few shot setting we need also a small Lifted batched Wasserstein distance $\mathcal{W}_{\mathbf{W_{k_1,k_2}}}(\mathbb{P},\mathbb{Q})$.

Few Shots Lifted Empirical Risk Minimization Let $(\rho_i)_{1\leq i\leq m} \sim \mathbb{P}^{\otimes m}$, and $(z_{ij} = (x_{ij}, y_{ij}))_{1\leq j\leq n} \sim \rho_i^{\otimes n}$ and let $\hat{\rho}_i^n = \frac{1}{n}\sum_{j=1}^n \delta_{z_{ij}}$. Define $\hat{\rho}_i^{b,k} = \frac{1}{k}\sum_{\ell=1}^k \delta_{z_{i\ell}^b}$, where $(z_{i\ell}^b)_{\ell=1\dots k, b=1\dots B} \sim \rho_i^{\otimes kB}$. We note $S = \left\{\rho_i, \hat{\rho}_i^n, \hat{\rho}_i^{b,k}, i = 1\dots m, b = \dots B\right\}$.We consider for the simplicity the following empirical unbiased estimator :

$$\hat{\mathcal{R}}_{k,S}^V(f) = \frac{1}{m}\frac{1}{n}\frac{1}{B}\sum_{i=1}^m\sum_{j=1}^n\sum_{b=1}^B V(y_{ij}, f(\hat{\rho}_i^{b,k}, x_{ij}))), \tag{11}$$

we recover here the training cost in [Kirsch et al., 2022] and [Müller et al., 2022].

Remark 2. *Note that we have decoupled all random variables in the expectation to ease the analysis and to get interpretable geometric bounds. Other U-statistics estimators can be derived for sampling $(k+1)$ tuples from n samples, and decoupling techniques involving Rademacher chaos [De la Pena and Giné, 2012] can be used, but we decided to state the results with easier to interpret bounds in this simplified setup.*

Let $f_{\mathcal{R},\mathbb{P}}$ be the optimal function in $\mathcal{F}$ minimizing the expected risk given (8) and let $\hat{f}_{m,n,B}^k$ be the minimizer in $\mathcal{F}$ of the empirical risk given in (11).

Theorem 4 (Generalization of Few Shot In-Context Meta-Learning)**.** *Let $\mathbb{P}_m = \frac{1}{m}\sum_{i=1}^m \delta_{\rho_i}$ (tasks seen in the training) and $\mathbb{P}'_m = \frac{1}{m}\sum_{i=1}^m \delta_{\rho'_i}$ (unseen tasks during the training), where $\rho'_i \sim \mathbb{P}^{\otimes m}$ independent from ρ_i. Define S' similarly to S from points sampled from ρ'_i. Under Assumptions 1 and 2 we have:*

$$\mathbb{E}_S\left(\mathcal{R}_{k,\mathbb{P}}^V(\hat{f}_{m,n,B}^k)\right) \leq \mathcal{R}_{k,\mathbb{P}}^V(f_{\mathcal{R},\mathbb{P}}) + 2\sqrt{2}L_V(1+L_{\mathcal{F}})\mathbb{E}_S\mathbb{E}_{S'}\widehat{\mathcal{W}}_{\mathbf{W}_1+\mathbf{W}_1^{k,k}}(\mathbb{P}_m, \mathbb{P}'_m),$$

where:

$$\widehat{\mathcal{W}}_{\mathbf{W}_1+\mathbf{W}_1^{k,k}}(\mathbb{P}_m, \mathbb{P}'_m) = \min_{\gamma\in\Gamma(\mathbb{P}_m,\mathbb{P}'_m)}\sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1 i_2}\left(W_1(\hat{\rho}_{i_1}^n, \hat{\rho}_{i_2}^{'n}) + \frac{1}{B^2}\sum_{b_1,b_2=1}^B W_1(\hat{\rho}_{i_1}^{b_1,k}, \hat{\rho}_{i_2}^{',b_2,k})\right).$$

Note that the generalization in the few shot learning depends on the similarity of two independent empirical samples from the meta distribution $\mathbb{P}$ corresponding to seen tasks in the training and unseen tasks. If this similarity is low this leads to good generalization guarantees. This similarity is measured by an estimate of the lifted Wasserstein for $\mathbf{D}$ being the sum of the Wasserstein distance and the batched Wasserstein that assesses the few shot scenario. The first term can be seen as the variance on unseen test samples and the second term as the variance of the context or few shots examples as tasks change. This notion of variance is related to the k-variance notion introduced in [Solomon et al., 2022] for measures in $\mathcal{P}(\mathcal{Z})$ and was linked to generalization in supervised learning in [Chuang et al., 2021]. As the number of tasks m increases the "borrowing of strength" between unseen tasks in the training ($\hat{\rho}'_i$) and seen tasks ($\hat{\rho}_i$) increases, leading to better generalization as this reduces the Lifted Wasserstein distance. This is in line with empirical findings in [Kirsch et al., 2022] and [Müller et al., 2022].

The generalization bound given in Theorem 4 can be evaluated empirically given seen and unseen tasks in meta-training. We give here non-sharp quantitative bounds :

Corollary 1 (Quantitative Bound)**.** *The following bounds holds for few shot learning:*

$$\mathbb{E}_S\left(\mathcal{R}_{k,\mathbb{P}}^V(\hat{f}_{m,n,B}^k)\right) \lesssim \mathcal{R}_{k,\mathbb{P}}^V(f_{\mathcal{R},\mathbb{P}}) + 2\sqrt{2}L_V(1+L_{\mathcal{F}})\left(2\mathbb{E}_{\rho,\rho'\sim\mathbb{P}}W_1(\rho,\rho') + 2k^{-\frac{1}{d_x+d_y}} + 2n^{-\frac{1}{d_x+d_y}}\right)$$

The constant term $\mathbb{E}_{\rho,\rho'\sim\mathbb{P}}W_1(\rho,\rho')$ accounts for out of distribution tasks, and is not a sharp bound. The learning is cursed in the dimension unless we consider Assumption 4, then the rate in n and k are replaced by $n^{-\frac{1}{2}}$ and $k^{-\frac{1}{2}}$. Note that m and B does not enter in the bound for the expectation and will only reduce the variance of the estimator.

We will now relate the empirical solution of the few shot problem (11) to the idealized setting ((1)):

Corollary 2 (From Few Shots Back To the Idealized Setting)**.** *We have the following bound relating the few shot setting to the idealized in-context learning objective:*

$$\mathbb{E}_S\left(\mathcal{E}^V_{\mathbb{P}}(\hat{f}^k_{m,n,B})\right) \lesssim \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}) + 2\sqrt{2}L_V(1+L_{\mathcal{F}})\mathbb{E}_S\mathbb{E}_{S'}\widehat{\mathcal{W}}_{\mathbf{W}_1+\mathbf{W}_1^{k,k}}(\mathbb{P}_m,\mathbb{P}'_m) + 2k^{-\frac{1}{(d_x+d_y)}} + \left(\mathcal{E}^V_{\mathbb{P}}(f_{\mathcal{R},\mathbb{P}}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}})\right).$$

When we consider the square loss and assume that the regression function f^* given in Equation (5) is in the function space $\mathcal{F}$ so that $f_{\mathbb{P}} = f^*$. Recall that f^* performs bayesian inference, plugging f^* in the equation above we obtain:

$$\mathbb{E}_S\left(\mathcal{E}^V_{\mathbb{P}}(\hat{f}^k_{m,n,B})\right) \lesssim \mathcal{E}^V_{\mathbb{P}}(f^*) + 2\sqrt{2}L_V(1+L_{\mathcal{F}})\mathbb{E}_S\mathbb{E}_{S'}\widehat{\mathcal{W}}_{\mathbf{W}_1+\mathbf{W}_1^{k,k}}(\mathbb{P}_m,\mathbb{P}'_m) + 2k^{-\frac{1}{(d_x+d_y)}} + L_V\int d\mathbb{P}(\rho)\int d_{\mathcal{Y}}(f^*(\rho,x), f_{\mathcal{R},\mathbb{P}}(\rho,x))d\rho_{\mathcal{X}}(x).$$

We see that $\hat{f}^k_{m,n,B}$ will learn to do bayesian inference, as long as the few shot numbers k is large, the generalization bound as measured by the lifted Wasserstein distance is small, and the function space $\mathcal{F}$ is large to approximate the regression function f^*. Hence we can think of $\hat{f}^k_{m,n,B}$ as performing implicitly a bayesian inference.

5 Meta-Learning Transformers and Implicit Bayesian Inference

[Kirsch et al., 2022] and [Müller et al., 2022] showed emprically that meta-learning transformers within a few shot in-context learning setup (same setup as ours in Section 4) enable general purpose in-context learning [Kirsch et al., 2022] or implicit bayesian inference [Müller et al., 2022]. The central assumption in our theory is the regularity of the function class in terms of its Lipschitzness given in Assumption 2 with respect to measures (in ρ) and points (in x). Note that in the few shot setting it is enough to have regularity on empirical measures only. Hence in order to apply the results of the previous section to transformers, we need to understand the regularity of transformers and attention models. This question was addressed recently in two concurrent works: [Kim et al., 2021] studied the regularity of attention in the euclidean geometry and [Vuckovic et al., 2021] studied it in the Wasserstein geometry.

Regularity of Self-Attention In Encoder Transformers We start here by reviewing the transformer architecture and focus on the setup considered in [Müller et al., 2022] i.e encoder only models (no masking) and without positional encoding (to ensure permutation invariance). Given an input the sequence $x = (x_1 \dots x_T) \in \mathcal{X} \subset B(R,\mathbb{R}^d)$, following [Vuckovic et al., 2021] and [Sander et al., 2022] we think of attention as a push forward map applied to the empirical measure $m(x) = \frac{1}{t}\sum_{t=1}^T \delta_{x_t}$. A self-attention block [2] acts on m as follows: for $t = 1 \dots T$ $\mathcal{T}_{m(x)}(x_t) = \sum_{s=1}^T \frac{\exp(a(W_Q x_t, W_K x_s))}{\sum_{s'=1}^T \exp(a(W_Q x_t, W_k x_{s'}))} W_V x_s$, where W_Q, W_K and W_V are matrices in $\mathbb{R}^{d\times d}$ with bounded spectral norms. For $u, v \in \mathcal{X}$, a in the original transformer is the dot product $a(u,v) = \langle u, v\rangle$. If $\mathcal{X}$ is convex and compact, [Vuckovic et al., 2021] showed that the self-attention with a being the dot product, has Lipschitz regularity in the Wasserstein geometry. For unbounded domains, this Wasserstein regularity is satisfied only for $a(u,v) = -\|u-v\|^2$. We set here some notations that will be used in the remainder of the paper. Let $\mu = \frac{1}{T}\sum_{t=1}^T \delta_{b_t}$ we note $(\mathcal{T}_\mu)_\sharp(\mu) = \frac{1}{T}\sum_{t=1}^T \delta_{\mathcal{T}_\mu(b_t)}$, the point cloud that result after applying self attention. Denote by $\psi(\mu)$ the inverse map from a measure to a sequence $\psi(\mu) = (b_1, \dots b_T)$. It follows that the resulting sequence after self-attention is : $\psi(\mathcal{T}_\mu)_\sharp(\mu)) = (\mathcal{T}_\mu(b_1), \dots \mathcal{T}_\mu(b_T))$.

We summarize in the following proposition results implied by the work in [Vuckovic et al., 2021]:

[2] We present it here with a single head

Proposition 1 ((Informal)). *Consider two sequences $x = (x_1, \dots, x_T)$ and $y = (y_1 \dots y_T)$. There exist constants L and L' that depends on d, T spectral norms of W_Q, W_K, and W_V (in the bounded case for a being dot product and in the unbounded case for $a(u, v) = -\|u - v\|^2$) such that:*

1. $W_1\Big((\mathcal{T}_{m(x)})_\sharp(m(x)), (\mathcal{T}_{m(y)})_\sharp(m(y))\Big) \leq L\, W_1(m(x), m(y))$
2. *For all $t = 1 \dots T$ we have:*

$$d_{\mathcal{X}}(\mathcal{T}_{m(x)}(x_t), \mathcal{T}_{m(y)}(y_t)) \leq L' \left(d_{\mathcal{X}}(x_t, y_t) + W_1(m(x), m(y))\right)$$

Note that in the transformer architecture, Self-Attention blocks ($\mathcal{T}^j_{\mu_j}, j = 1 \dots D$) are followed by residual feedforward networks ($\text{FN}_j, j = \dots D$) that preserve the Lipschitz regularity. Iterating through multiple layers of attentions results therefore in a Lipschtiz constant that is the multiplication of Lipschitz constant of the consecutive blocks.

Using Transformers in-context Learning For learning to learn using transformers, given k in-context examples $\{(x_\ell, y_\ell)_{1 \leq \ell \leq k}\}$and a point x, we define $f(\frac{1}{k}\sum_{\ell=1}^k \delta_{(x_\ell, y_\ell)}, x)$, as follows:

1. For the case ($\mathcal{X} \neq \mathcal{Y}$), for example x is an image and y is the label encoded as one hot, define a sequence $a = (z_1 = (x_1, y_1), \dots z_k = (x_k, y_k), (x, 0))$ of length $T = k + 1$, this is the setup considered in [Kirsch et al., 2022]. For the case $\mathcal{X} = \mathcal{Y}$, this is the case in natural language define the sequence $a = (x_1, y_1 \dots x_k, y_k, x)$ of length $T = 2k + 1$, this is the setup considered in in-context-learning in [Brown et al., 2020][3]. Let $m(a)$ be the corresponding empirical measure to the sequence a in both cases.
2. Define a deep transformer by iterating for $j = 0 \dots D - 1$, $\mu_{j+1} = (\text{FN}_j)_\sharp\left[(\mathcal{T}^j_{\mu_j})_\sharp(\mu_j)\right]$, where $\mu_0 = m(a)$. We obtain a sequence from the μ_{j+1} by using the operator $\psi(\mu_{j+1}) = (h_1^{j+1} \dots h_T^{j+1})$.
3. Extract h_T^D from μ_D and define $f(\rho, x) = W_O h_T^D$ (W_O is a linear map to the output space $\mathcal{Y}$ that has bounded spectral norm).

The following proposition uses the regularity of self-attention in the Wasserstein geometry (proposition 1) and shows that encoder transformers satisfy the regularity assumptions we considered in Section 4.

Proposition 2. *f defined above by an encoder transformer satisfies Assumption 2 in both cases for $\mathcal{X} \neq \mathcal{Y}$ and and for $\mathcal{X} = \mathcal{Y}$.*

We conclude that encoder transformers are meta in-context learners thanks to their regularity in the Wasserstein geometry and their generalization properties are given in Section 4. The generalization to unseen tasks is governed by the lifted Wasserstein distance with a cost involving the Wasserstein distance and the batched Wasserstein. Proposition 3 in the Appendix gives generalization bounds in the attention feature space, that further links generalization of ICL to the cluster-ability of the contexts in the lifted wasserstein geometry.

6 Conclusion

We presented in this paper a statistical learning framework for learning to learn in-context and showed that encoder transformers are in-context meta learners and they can infer the task at hand from few shots and perform in a way an implicit bayesian inference. We based our analysis on the regularity of self-attention in the Wasserstein geometry [Vuckovic et al., 2021].

We leave for future work the study of decoder transformers and their emergent in-context learning [Brown et al., 2020] when pre-trained for next token prediction. We conjecture that their emergent in-context learning is also due to some form of regularity. The main challenge is reconciling the sequential regularity of decoder models and the permutation invariance of in-context examples, a problem that was alluded to in [Xie et al., 2022]. Another venue for future work is the analysis of the complexity of encoder transformers using coverings in the Wasserstein geometry, this would lead to a tighter analysis of the generalization of in-context learning.

[3] [Brown et al., 2020] consider a decoder transformer and our analysis does not apply to their results.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL `https://doi.org/10.5281/zenodo.7733589`.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433. PMLR, 2019.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 235–246. PMLR, 22–24 Mar 2019.

Guillaume Carlier, Alex Delalande, and Quentin Merigot. Quantitative stability of barycenters in the wasserstein space. 2023.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*, 2021.

Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL `https://proceedings.neurips.cc/paper_files/paper/2010/file/4e0cb6fb5fb446d1c92ede2ed8780188-Paper.pdf`.

Ching-Yao Chuang, Youssef Mroueh, Kristjan Greenewald, Antonio Torralba, and Stefanie Jegelka. Measuring generalization with optimal transport. *Advances in Neural Information Processing Systems*, 34:8294–8306, 2021.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1566–1575. PMLR, 09–15 Jun 2019.

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2131–2141. PMLR, 26–28 Aug 2020. URL `http://proceedings.mlr.press/v108/fatras20a.html`.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.

Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *arXiv preprint arXiv:1710.11622*, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019.

Tadahisa Funaki. A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67(3):331–348, 1984.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pages 1704–1713. PMLR, 2018.

Michael Hahn and Navin Goyal. A theory of emergent in-context learning as implicit structure induction, 2023.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571. PMLR, 2021.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. Jurassic-1: Technical details and evaluation. Technical report, AI21 Labs, August 2021.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *NAACL-HLT*, 2022.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=KSugKcbNf9`.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International conference on machine learning*, pages 2554–2563. PMLR, 2017.

OpenAI. Gpt-4 technical report, 2023.

Barnabas Poczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 507–515, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3515–3530. PMLR, 28–30 Mar 2022.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1842–1850, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/santoro16.html.

Justin Solomon, Kristjan Greenewald, and Haikady Nagaraja. k-variance: A clustered notion of variance. *SIAM Journal on Mathematics of Data Science*, 4(3):957–978, 2022.

Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23, 2019. URL http://jmlr.org/papers/v20/18-079.html.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16:2023–2049, 2015.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv:1706.03762*, 2017.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. On the regularity of attention. *arXiv preprint arXiv:2102.05628*, 2021.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning, 2023.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.

A Supplementary Material

Contents

A.1 Related Work: Discussion of Meta-Learning

We discuss our work in the context of the learning to learn framework of [Baxter, 2000]. Baxter [2000] introduces also the notion of an environment that is a meta distribution on tasks η. For a task $t \sim \eta$, we draw an iid training set $x^t = (x_1^t \dots x_n^t) \sim \mu_t^n$, and one wants to learn a function $f \circ h$, where h is a feature map. The meta learning problem can be written as follows (see for e.g [Maurer, 2016]):

$$\psi_t(h) = \min_f \frac{1}{n} \sum_{i=1}^n V(y_i^t, f(h(x_i^t)))$$

$$h = \arg\min_{h \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^T \psi_t(h)$$

The feature map is meta learned across all tasks.

Black-Box Meta-Learning Our setting does not fall within this category of bi-level optimization and is in line with works on black-box meta learning methods, where a single model is jointly learned on all environments: RNNs were considered in [Santoro et al., 2016], conditional neural process in [Garnelo et al., 2018], meta networks in [Munkhdalai and Yu, 2017] and attention in [Mishra et al., 2017]. Our work can be seen as a study of black-box meta learning with transformers.

Gradient Based Meta Learning The bilevel optimization problem in meta learning is expensive as it appeals to computing hessians and this lead many advances via unrolled optimization pioneered by the MAML approach [Finn et al., 2017] and other gradient based approaches studied in [Finn and Levine, 2017, Balcan et al., 2019, Bullins et al., 2019, Denevi et al., 2019, Finn et al., 2019].

A.2 Proofs of Section 3

Proof of Theorem 1. Assume without loss of generality that all measures have densities. Let γ be the optimal coupling between $\mathbb{P}$ and $\mathbb{Q}$, i.e the minimizer of (6), γ satisifies:

$$\int \gamma(\rho, \rho') d\rho' = \mathbb{P}(\rho) \text{ and } \int \gamma(\rho, \rho') d\rho = \mathbb{Q}(\rho').$$

For $\rho, \rho' \in \mathcal{P}(\mathcal{Z})$, let $\tilde{\gamma}_{\rho,\rho'}$ be the optimal coupling in $\Gamma(\rho, \rho')$ i.e the minimizer of wasserstein distance between ρ and ρ':

$$\int_{\mathcal{Z}} \tilde{\gamma}_{\rho,\rho'}(x, y, x', y') dx' dy' = \rho(x, y) \text{and} \int_{\mathcal{Z}} \tilde{\gamma}_{\rho,\rho'}(x, y, x', y') dz = \rho'(x', y').$$

From Assumption 1 and the compactness of $\mathcal{Y}$, we deduce that the loss V is bounded : $|V(y, y')| \leq V(0, 0) + L_V \sqrt{2}\, \text{diam}(\mathcal{Y})$ and hence using Fubini's Theorem we obtain:

$$\begin{aligned}
\mathcal{E}_{\mathbb{P}}^V(f) &= \int_{\mathcal{P}(\mathcal{Z})} \mathbb{P}(\rho) d\rho \int_{\mathcal{Z}} V(y, f(\rho, x)) \rho(x, y) dx dy \\
&= \int_{\mathcal{P}(\mathcal{Z})} d\rho \left(\int_{\mathcal{P}(\mathcal{Z})} \gamma(\rho, \rho') d\rho' \right) \int_{\mathcal{Z}} V(y, f(\rho, x)) \left(\int_{\mathcal{Z}} \gamma_{\rho,\rho'}(x, y, x', y') dx' dy' \right) dx dy \\
&= \int_{\mathcal{P}(\mathcal{Z})} \int_{\mathcal{P}(\mathcal{Z})} \int_{\mathcal{Z}} \int_{\mathcal{Z}} V(y, f(\rho, x)) \gamma(\rho, \rho') \tilde{\gamma}_{\rho,\rho'}(x, y, x', y') d\rho d\rho' dx dy dx' dy',
\end{aligned}$$

the last equality is another application of Fubini's theorem. A similar expression holds for $\mathcal{E}^V_{\mathbb{Q}}(f)$:

$$\mathcal{E}^V_{\mathbb{Q}}(f) = \int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}}\int_{\mathcal{Z}} V(y', f(\rho', x'))\gamma(\rho,\rho')\tilde{\gamma}_{\rho,\rho'}(x,y,x',y')d\rho d\rho' dxdydx'dy'$$

Hence we have:

$$\begin{aligned}&\mathcal{E}^V_{\mathbb{Q}}(f) - \mathcal{E}^V_{\mathbb{P}}(f)\\ &= \int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}}\int_{\mathcal{Z}} \Big(V(y', f(\rho', x')) - V(y, f(\rho, x))\Big)\gamma(\rho,\rho')\tilde{\gamma}_{\rho,\rho'}(x,y,x',y')d\rho d\rho' dxdydx'dy'\end{aligned} \tag{12}$$

Turning now to :

$$\begin{aligned}|V(y', f(\rho', x')) - V(y, f(\rho, x)| &\leq L_V\sqrt{d^2_{\mathcal{Y}}(y,y') + d^2_{\mathcal{Y}}(f(\rho,x), f(\rho',x'))}\\ &\leq L_V(d_{\mathcal{Y}}(y,y') + d_{\mathcal{Y}}(f(\rho,x), f(\rho',x')))\\ &\leq L_V\left(d_{\mathcal{Y}}(y,y') + L_{\mathcal{F}}\left(d_{\mathcal{X}}(x,x') + W_1(\rho,\rho')\right)\right)\\ &\leq L_V(1+L_{\mathcal{F}})\left(d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y') + W_1(\rho,\rho')\right)\\ &\leq L_V(1+L_{\mathcal{F}})\left(\sqrt{2}\sqrt{d^2_{\mathcal{X}}(x,x') + d^2_{\mathcal{Y}}(y,y')} + W_1(\rho,\rho')\right)\end{aligned} \tag{13}$$

where in the first inequality we used Assumption 1; in the second inequality we used that for $a, b > 0$, $\sqrt{a^2+b^2} \leq a + b$; in the third inequality we used Assumption 2; and in the last inequality we used the fact that $|a+b| \leq \sqrt{2(a^2+b^2)}$. It follows from equations (12) and (13) that:

$$\left|\mathcal{E}^V_{\mathbb{Q}}(f) - \mathcal{E}^V_{\mathbb{P}}(f)\right| \leq L_V(1+L_{\mathcal{F}})(A+B), \tag{14}$$

where :

$$\begin{aligned}A &= \sqrt{2}\int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}}\int_{\mathcal{Z}}\sqrt{d^2_{\mathcal{X}}(x,x') + d^2_{\mathcal{Y}}(y,y')}\gamma(\rho,\rho')\tilde{\gamma}_{\rho,\rho'}(x,y,x',y')d\rho d\rho' dxdydx'dy'\\ &= \sqrt{2}\int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})}\gamma(\rho,\rho')\left(\int_{\mathcal{Z}}\int_{\mathcal{Z}}\sqrt{d^2_{\mathcal{X}}(x,x') + d^2_{\mathcal{Y}}(y,y')}\tilde{\gamma}_{\rho,\rho'}(x,y,x',y')dxdx'dydy'\right)d\rho d\rho'\\ &= \sqrt{2}\int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})}\gamma(\rho,\rho')W_1(\rho,\rho')d\rho d\rho'\\ &= \sqrt{2}\mathcal{W}_{\mathbf{W_1}}(\mathbb{P},\mathbb{Q}),\end{aligned}$$

in the second equality we used Fubini Theorem, in the third and last equality we used the optimality of $\tilde{\gamma}_{\rho,\rho'}$ and γ respectively. And,

$$\begin{aligned}B &= \int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}}\int_{\mathcal{Z}} W_1(\rho,\rho')\gamma(\rho,\rho')\tilde{\gamma}_{\rho,\rho'}(x,y,x',y')d\rho d\rho' dxdydx'dy'\\ &= \int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})} W_1(\rho,\rho')\gamma(\rho,\rho')\underbrace{\left(\int_{\mathcal{Z}}\int_{\mathcal{Z}}\tilde{\gamma}_{\rho,\rho'}(x,y,x',y')d\rho d\rho' dxdydx'dy'\right)}_{=1}d\rho d\rho'\\ &= \int_{\mathcal{P}(\mathcal{Z})}\int_{\mathcal{P}(\mathcal{Z})} W_1(\rho,\rho')\gamma(\rho,\rho')d\rho d\rho'\\ &= \mathcal{W}_{\mathbf{W}_1}(\mathbb{P},\mathbb{Q}),\end{aligned}$$

where we used Fubini and the fact that $\int_{\mathcal{Z}}\int_{\mathcal{Z}}\tilde{\gamma}_{\rho,\rho'}(x,y,x',y')d\rho d\rho' dxdydx'dy' = 1$ since $\gamma_{\rho,\rho'}$ is a coupling and the optimality of γ in the last equality. Finally,

$$\left|\mathcal{E}^V_{\mathbb{Q}}(f) - \mathcal{E}^V_{\mathbb{P}}(f)\right| \leq L_V(1+L_{\mathcal{F}})(1+\sqrt{2})\mathcal{W}_{\mathbf{W}_1}(\mathbb{P},\mathbb{Q}). \tag{15}$$

□

Proof of Theorem 2. Our goal is to bound:

$$\begin{aligned}\mathcal{E}^V_{\mathbb{P}}(\widehat{f}_{\mathbb{P}^n_m}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}) &= \left(\mathcal{E}^V_{\mathbb{P}}(\widehat{f}_{\mathbb{P}^n_m}) - \mathcal{E}^V_{\mathbb{P}^n_m}(\widehat{f}_{\mathbb{P}^n_m})\right) + \left(\mathcal{E}^V_{\mathbb{P}^n_m}(\widehat{f}_{\mathbb{P}^n_m}) - \mathcal{E}^V_{\mathbb{P}^n_m}(f_{\mathbb{P}})\right) + \left(\mathcal{E}^V_{\mathbb{P}^n_m}(f_{\mathbb{P}}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}})\right) \\ &\leq \left(\mathcal{E}^V_{\mathbb{P}}(\widehat{f}_{\mathbb{P}^n_m}) - \mathcal{E}^V_{\mathbb{P}^n_m}(\widehat{f}_{\mathbb{P}^n_m})\right) + \left(\mathcal{E}^V_{\mathbb{P}^n_m}(f_{\mathbb{P}}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}})\right) \\ &\leq 2L\, \mathcal{W}_{\mathbf{W}_1}(\mathbb{P}, \mathbb{P}^n_m),\end{aligned}$$

where we used in the first inequality $\mathcal{E}^V_{\mathbb{P}^n_m}(\widehat{f}_{\mathbb{P}^n_m}) \leq \mathcal{E}^V_{\mathbb{P}^n_m}(f_{\mathbb{P}})$, since $\widehat{f}_{\mathbb{P}^n_m}$ is a minimizer of $\mathcal{E}^V_{\mathbb{P}^n_m}(.)$ and the last inequality is a direct application of the stability Theorem 1. Taking the expectation on the randomness in both $\hat{\mathbb{P}}_m$ and $\hat{\mathbb{P}}^n_m$ we obtain by applying the triangle inequality:

$$\mathbb{E}\left(\mathcal{E}^V_{\mathbb{P}}(\widehat{f}_{\mathbb{P}^n_m}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}})\right) \leq 2L\, \mathbb{E}\mathcal{W}_{\mathbf{W}_1}(\mathbb{P}, \mathbb{P}^n_m) \leq 2L\ \left(\mathbb{E}\mathcal{W}_{\mathbf{W}_1}(\mathbb{P}, \mathbb{P}_m) + \mathbb{E}\mathcal{W}_{\mathbf{W}_1}(\mathbb{P}_m, \mathbb{P}^n_m)\right)) \tag{16}$$

Under Assumption 3, we have by Theorem 1 in [Weed and Bach, 2019]:

$$\mathbb{E}\mathcal{W}_{\mathbf{W}_1}(\mathbb{P}, \mathbb{P}_m) \lesssim m^{-\frac{1}{s}}. \tag{17}$$

Note that $\mathcal{W}_{\mathbf{W}_1}(\mathbb{P}_m, \mathbb{P}^n_m)) = \sum_{ij} \gamma^*_{ij} W_1(\rho_i, \hat{\rho}^n_j)$, where γ^* is the optimal coupling in $\Gamma(\mathbb{P}_m, \mathbb{P}^n_m)$. Let $\gamma^u_{ii} = \frac{1}{m}$ for $i = 1 \dots m$ and $\gamma^u_{ij} = 0$ for $i \neq j$. By optimality of γ^* we have:

$$\mathcal{W}_{\mathbf{W}_1}(\mathbb{P}_m, \mathbb{P}^n_m)) \leq \sum_{ij} \gamma^u_{ij} W_1(\rho_i, \hat{\rho}^n_j) = \frac{1}{m}\sum_{i=1}^m W_1(\rho_i, \hat{\rho}^n_i)$$

Hence taking expectations we obtain:

$$\mathbb{E}\mathcal{W}_{\mathbf{W}_1}(\mathbb{P}_m, \mathbb{P}^n_m)) \leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}W_1(\rho_i, \hat{\rho}^n_i),$$

without further assumptions W_1 is cursed in dimension and we have:

$$\mathbb{E}W_1(\rho_i, \hat{\rho}^n_i) \lesssim n^{-\frac{1}{d_x + d_y}}.$$

Under Assumption 4, by proposition 13 in [Weed and Bach, 2019] we have for $n \leq q(2\Delta)^{-2}$, for all $i = 1 \dots m$:

$$\mathbb{E}W_1(\rho_i, \hat{\rho}^n_i) \leq 12\sqrt{\frac{q}{n}}. \tag{18}$$

Hence under Assumptions 3 and 4, combining Equations (16), (17) and (18) we have for $n \leq q(2\Delta)^{-2}$:

$$\mathbb{E}\left(\mathcal{E}^V_{\mathbb{P}}(\widehat{f}_{\mathbb{P}^n_m}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}})\right) \lesssim 2L\left(m^{-\frac{1}{s}} + 12\sqrt{\frac{q}{n}}\right).$$

□

A.3 Proofs of Section 4

Proof of Lemma 1.

$$\mathcal{R}^V_{k,\mathbb{P}}(f) - \mathcal{E}^V_{\mathbb{P}}(f) = \mathbb{E}_{\rho\sim\mathbb{P}}\mathbb{E}_{(x,y)\sim\rho}\mathbb{E}_{(x_\ell,y_\ell)\sim\rho^{\otimes k}} V(y, f(\hat{\rho}_k, x) - V(y, f(\rho, x))) \tag{19}$$

It follows from Equation (13) in proof of Theorem 1 that:

$$|V(y, f(\hat{\rho}_k, x) - V(y, f(\rho, x)))| \leq L_V(L_{\mathcal{F}} + 1)W_1(\hat{\rho}_k, \rho)$$

Taking expectation, absolute values and using Jensen inequality we get the results, using classical bounds on W_1, see for example [Weed and Bach, 2019]. □

Proof of Theorem 3. The proof is similar to the proof of Theorem 1. Let γ be a coupling between $\mathbb{P}$ and $\mathbb{Q}$. For $\rho \sim \mathbb{P}$ and $\rho' \sim \mathbb{Q}$, $\tilde{\gamma}_{\rho,\rho'}$ is optimal W_1 coupling between ρ and ρ'.

$$\begin{aligned}
&\mathcal{R}^V_{k_1,\mathbb{P}}(f) - \mathcal{R}^V_{k_2,\mathbb{Q}}(f) \\
&= \mathbb{E}_{(\rho,\rho')\sim\gamma}\mathbb{E}_{((x,y),(x',y'))\sim\tilde{\gamma}_{\rho,\rho'}}\mathbb{E}_{\substack{Z_1\dots Z_{k_1}\sim\rho \\ Z'_1\dots Z'_{k_2}\sim\rho'}}\left[V\left(y, f\left(\frac{1}{k_1}\sum_{\ell=1}^{k_1}\delta_{Z_\ell}, x\right)\right) - V\left(y', f\left(\frac{1}{k_2}\sum_{\ell=1}^{k_2}\delta_{Z'_\ell}, x'\right)\right)\right]
\end{aligned} \tag{20}$$

Under assumptions Assumption 1 and Assumption 2 by equation (13), we have:

$$\begin{aligned}
&\left|V\left(y, f\left(\frac{1}{k_1}\sum_{\ell=1}^{k_1}\delta_{Z_\ell}, x\right)\right) - V\left(y', f\left(\frac{1}{k_2}\sum_{\ell=1}^{k_2}\delta_{Z'_\ell}, x'\right)\right)\right] \\
&\leq L_V(1+L_{\mathcal{F}})\left(\sqrt{2}\sqrt{d^2_{\mathcal{X}}(x,x') + d^2_{\mathcal{Y}}(y,y')} + W_1(\hat{\rho}_{k_1}, \hat{\rho}'_{k_2})\right)
\end{aligned}$$

Hence taking absolute value in Equation (20), using Jensen inequality and the above bound on the loss, we have:

$$\begin{aligned}
\left|\mathcal{R}^V_{k_1,\mathbb{P}}(f) - \mathcal{R}^V_{k_2,\mathbb{Q}}(f)\right| &\leq L_V(1+L_{\mathcal{F}})\Big\{\sqrt{2}\mathbb{E}_{(\rho,\rho')\sim\gamma}\mathbb{E}_{((x,y),(x',y'))\sim\tilde{\gamma}_{\rho,\rho'}}\sqrt{d^2_{\mathcal{X}}(x,x') + d^2_{\mathcal{Y}}(y,y')} \\
&\quad .... + \mathbb{E}_{(\rho,\rho')\sim\gamma}\mathbb{E}_{\substack{Z_1\dots Z_{k_1}\sim\rho \\ Z'_1\dots Z'_{k_2}\sim\rho'}}\left[W_1\left(\frac{1}{k_1}\sum_{i=1}^{k_1}\delta_{Z_i}, \frac{1}{k_2}\sum_{i=1}^{k_2}\delta_{Z'_i}\right)\right]\Big\} \\
&= L_V(1+L_{\mathcal{F}})\left(\sqrt{2}\mathbb{E}_{(\rho,\rho')\sim\gamma}W_1(\rho,\rho') + \mathbb{E}_{(\rho,\rho')\sim\gamma}W_1^{k_1,k_2}(\rho,\rho')\right) \\
&\leq L_V(1+L_{\mathcal{F}})\sqrt{2}\mathbb{E}_{(\rho,\rho')\sim\gamma}\left(W_1(\rho,\rho') + W_1^{k_1,k_2}(\rho,\rho')\right)
\end{aligned} \tag{21}$$

where we used that $\tilde{\gamma}_{\rho,\rho'}$ is the W_1 optimal coupling, i.e $W_1(\rho,\rho') = \mathbb{E}_{((x,y),(x',y'))\sim\tilde{\gamma}_{\rho,\rho'}}\sqrt{d^2_{\mathcal{X}}(x,x') + d^2_{\mathcal{Y}}(y,y')}$, and the definition of the batched Wasserstein distance given in Equation (9).

Now choosing γ to be the optimal coupling γ^* of $\mathcal{W}_{\mathbf{D}}$, for

$$\mathbf{D}(\rho,\rho') = W_1(\rho,\rho') + W_1^{k_1,k_2}(\rho,\rho')$$

we obtain:

$$\begin{aligned}
\left|\mathcal{R}^V_{k_1,\mathbb{P}}(f) - \mathcal{R}^V_{k_2,\mathbb{Q}}(f)\right| &\leq L_V(1+L_{\mathcal{F}})\sqrt{2}\mathbb{E}_{(\rho,\rho')\sim\gamma^*}\left(W_1(\rho,\rho') + W_1^{k_1,k_2}(\rho,\rho')\right) \\
&= L_V(1+L_{\mathcal{F}})\sqrt{2}\mathcal{W}_{\mathbf{W_1}+\mathbf{W_1^{k_1,k_2}}}(\mathbb{P},\mathbb{Q}).
\end{aligned}$$

Note that :

$$\mathcal{W}_{\mathbf{W_1}}(\mathbb{P},\mathbb{Q}) + \mathcal{W}_{\mathbf{W_1^{k_1,k_2}}}(\mathbb{P},\mathbb{Q}) \leq \mathcal{W}_{\mathbf{W_1}+\mathbf{W_1^{k_1,k_2}}}(\mathbb{P},\mathbb{Q}).$$

□

Proof of Theorem 4.

$$\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{R}^V_{k,\mathbb{P}}(f_{\mathbb{P}}) \leq \mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \hat{\mathcal{R}}^V_{k,S}(\hat{f}^k_{m,n,B}) + \hat{\mathcal{R}}^V_{k,S}(f_{\mathbb{P}}) - \mathcal{R}^V_{k,\mathbb{P}}(f_{\mathbb{P}})$$

Taking expectations on $(\rho_i)_{1\leq i\leq m} \sim \mathbb{P}^{\otimes m}$, $((x_{ij},y_{ij}))_{1\leq j\leq n}) \sim \rho_i^{\otimes n}$ and $(z^b_{i\ell})_{\ell=1\dots k, b=1\dots B} \sim \rho_i^{\otimes kB}$, we denote all those random variables on which expectation is taken by $S = \left\{\rho_i, \hat{\rho}^n_i, \hat{\rho}^{b,k}_i, i = 1\dots m, b = \dots B\right\}$:

$$\begin{aligned}\mathbb{E}_S\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{R}^V_{k,\mathbb{P}}(f_{\mathbb{P}}) &\leq \mathbb{E}_S\left(\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \hat{\mathcal{R}}^V_{k,S}(\hat{f}^k_{m,n,B})\right) + \mathbb{E}_S\left(\hat{\mathcal{R}}^V_{k,S}(f_{\mathbb{P}}) - \mathcal{R}^V_{k,\mathbb{P}}(f_{\mathbb{P}})\right)\\ &\leq 2\sup_{f\in\mathcal{F}}\mathbb{E}_S\left|\mathcal{R}^V_{k,\mathbb{P}}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right|\\ &\leq 2\mathbb{E}_S\sup_{f\in\mathcal{F}}\left|\mathcal{R}^V_{k,\mathbb{P}}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right|\end{aligned}$$

Note that:

$$\begin{aligned}\mathcal{R}^V_{k,\mathbb{P}}(f) &= \mathbb{E}_{\rho\sim\mathbb{P}}\mathbb{E}_{(x,y)\sim\rho}\mathbb{E}_{(z_\ell=(x_\ell,y_\ell)_{1\leq\ell\leq k})\sim\rho^{\otimes(k)}}V\left(y, f\left(\frac{1}{k}\sum_{\ell=1}^k\delta_{z_\ell}, x\right)\right)\\ &= \mathbb{E}_{\rho_i'\sim\mathbb{P}^{\otimes m}}\mathbb{E}_{(x'_{ij},y'_{ij})_{1\leq j\leq n}\sim\rho_i'^{\otimes n}}\mathbb{E}_{(z'^b_{i\ell})_{\ell=1\ldots k,b=1\ldots B}\sim\rho_i'^{\otimes kB}}\frac{1}{m}\frac{1}{n}\frac{1}{B}\sum_{i=1}^m\sum_{j=1}^n\sum_{b=1}^B V\left(y'_{ij}, f\left(\frac{1}{k}\sum_{\ell=1}^k\delta_{z'^b_{i\ell}}, x'_{ij}\right)\right)\\ &= \mathbb{E}_{S'}\hat{\mathcal{R}}^V_{k,S'}(f),\end{aligned}$$

where S' is an iid set of random variables constructed similar to S, $S' = \left\{\rho_i' \sim \mathbb{P}^{\otimes m}, \hat{\rho}_i'^n = \frac{1}{N}\sum_{j=1}^n\delta_{(x'_{ij},y'_{ij})}, \hat{\rho}_i^{'b,k} = \frac{1}{k}\sum_{\ell=1}^k\delta_{z'^b_{i\ell}}, i = 1\ldots n, b = \ldots B\right\}$.
Now:

$$\mathcal{R}^V_{k,\mathbb{P}}(f) - \hat{\mathcal{R}}^V_{k,S}(f) = \mathbb{E}_{S'}\left(\hat{\mathcal{R}}^V_{k,S'}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right) \tag{22}$$

It follows that:

$$\begin{aligned}\mathbb{E}_S\left(\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{R}^V_{k,\mathbb{P}}(f_{\mathbb{P}})\right) &\leq 2\mathbb{E}_S\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{S'}\left(\hat{\mathcal{R}}^V_{k,S'}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right)\right|\\ &\leq 2\mathbb{E}_S\sup_{f\in\mathcal{F}}\mathbb{E}_{S'}\left|\hat{\mathcal{R}}^V_{k,S'}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right| \text{ Jensen inequality convexity of absolute value}\\ &\leq 2\mathbb{E}_S\mathbb{E}_{S'}\sup_{f\in\mathcal{F}}\left|\hat{\mathcal{R}}^V_{k,S'}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right| \end{aligned}\tag{23}$$

Let $\gamma \in \mathbb{R}^{+m\times m}$ be a coupling between $\mathbb{P}_m = \frac{1}{m}\sum_{i_1=1}^m\delta_{\rho_{i_1}}$ and $\mathbb{P}'_m = \frac{1}{m}\sum_{i_2=1}^m\delta_{\rho'_{i_2}}$, and let $\tilde{\gamma}^{i_1i_2} \in \mathbb{R}^{+n\times n}$ be an optimal coupling (in W_1) between $\hat{\rho}^n_{i1}$ and $\hat{\rho}'^n_{i_2}$.

$$\begin{aligned}&\hat{\mathcal{R}}^V_{k,S}(f) - \hat{\mathcal{R}}^V_{k,S'}(f)\\ &= \frac{1}{m}\frac{1}{n}\frac{1}{B}\sum_{i_1=1}^m\sum_{j_1=1}^n\sum_{b_1=1}^B V(y_{i_1j_1}, f(\hat{\rho}^{b_1,k}_{i_1}, x_{i_1j_1}))) - \frac{1}{m}\frac{1}{n}\frac{1}{B}\sum_{i_2=1}^m\sum_{j_2=1}^n\sum_{b_2=1}^B V(y'_{i_2j_2}, f(\hat{\rho}^{'b_2,k}_{i_2}, x'_{i_2j_2})))\\ &= \sum_{i_1=1}^m\sum_{i_2=1}^m\gamma_{i_1,i_2}\sum_{j_1=1}^n\sum_{j_2=1}^n\tilde{\gamma}^{i_1i_2}_{j_1j_2}\frac{1}{B^2}\sum_{b_1,b_2=1}^B\left(V(y_{i_1j_1}, f(\hat{\rho}^{b_1,k}_{i_1}, x_{i_1j_1}))) - V(y'_{i_2j_2}, f(\hat{\rho}^{'b_2,k}_{i_2}, x'_{i_2j_2})))\right)\end{aligned}$$

Hence using Equation (13) we have

$$\begin{aligned}
&\left|\hat{\mathcal{R}}^V_{k,S'}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right| \\
&\leq L_V(1+L_{\mathcal{F}})\left(\sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1,i_2}\sum_{j_1=1}^n\sum_{j_2=1}^n \tilde{\gamma}^{i_1i_2}_{j_1j_2}\frac{1}{B^2}\sum_{b_1,b_2=1}^B \sqrt{2}\sqrt{d^2_{\mathcal{X}}(x_{i_1j_1},x'_{i_2j_2})+d^2_{\mathcal{Y}}(y_{i_1j_1},y'_{i_2j_2})} + W_1(\hat{\rho}^{b_1,k}_{i_1},\hat{\rho}^{'b_2,k}_{i_2})\right) \\
&= L_V(1+L_{\mathcal{F}})\Big\{\sqrt{2}\sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1,i_2}\sum_{j_1=1}^n\sum_{j_2=1}^n \tilde{\gamma}^{i_1i_2}_{j_1j_2}\sqrt{d^2_{\mathcal{X}}(x_{i_1j_1},x'_{i_2j_2})+d^2_{\mathcal{Y}}(y_{i_1j_1},y'_{i_2j_2})} \\
&+ \sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1,i_2}\underbrace{\sum_{j_1=1}^n\sum_{j_2=1}^n \tilde{\gamma}^{i_1i_2}_{j_1j_2}}_{=1}\frac{1}{B^2}\sum_{b_1,b_2=1}^B W_1(\hat{\rho}^{b_1,k}_{i_1},\hat{\rho}^{'b_2,k}_{i_2})\Big\} \\
&= L_V(1+L_{\mathcal{F}})\Big\{\sqrt{2}\sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1,i_2}W_1(\hat{\rho}^n_{i_1},\hat{\rho}^{'n}_{i_2}) + \sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1,i_2}\frac{1}{B^2}\sum_{b_1,b_2=1}^B W_1(\hat{\rho}^{b_1,k}_{i_1},\hat{\rho}^{'b_2,k}_{i_2})\Big\} \text{ (by optimality of } \tilde{\gamma}^{i_1i_2}) \\
&\leq L_V(1+L_{\mathcal{F}})\sqrt{2}\Big\{\sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1i_2}\left(W_1(\hat{\rho}^n_{i_1},\hat{\rho}^{'n}_{i_2}) + \frac{1}{B^2}\sum_{b_1,b_2=1}^B W_1(\hat{\rho}^{b_1,k}_{i_1},\hat{\rho}^{'b_2,k}_{i_2})\right)\Big\} \quad (A)
\end{aligned}$$

Let $\mathbb{P}_m = \frac{1}{m}\sum_{i=1}^m \delta_{\rho_i}$ and $\mathbb{P}'_m = \frac{1}{m}\sum_{i=1}^m \delta_{\rho'_i}$, define:

$$\widehat{\mathcal{W}}_{\mathbf{W}_1+\mathbf{W}_1^{k,k}}(\mathbb{P}_m,\mathbb{P}'_m) = \min_{\gamma\in\Gamma(\mathbb{P}_m,\mathbb{P}'_m)}\sum_{i_1=1}^m\sum_{i_2=1}^m \gamma_{i_1i_2}\left(W_1(\hat{\rho}^n_{i_1},\hat{\rho}^{'n}_{i_2}) + \frac{1}{B^2}\sum_{b_1,b_2=1}^B W_1(\hat{\rho}^{b_1,k}_{i_1},\hat{\rho}^{'b_2,k}_{i_2})\right),$$

Choosing γ to be the optimal coupling to the above problem we have combining Equations (23) and the bound in (A):

$$\begin{aligned}
\mathbb{E}_S\left(\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{R}^V_{k,\mathbb{P}}(f_{\mathbb{P}})\right) &\leq 2\mathbb{E}_S\mathbb{E}_{S'}\sup_{f\in\mathcal{F}}\left|\hat{\mathcal{R}}^V_{k,S'}(f) - \hat{\mathcal{R}}^V_{k,S}(f)\right| \\
&\leq 2\sqrt{2}L_V(1+L_{\mathcal{F}})\mathbb{E}_S\mathbb{E}_{S'}\widehat{\mathcal{W}}_{\mathbf{W}_1+\mathbf{W}_1^{k,k}}(\mathbb{P}_m,\mathbb{P}'_m).
\end{aligned}$$

□

Proof of Corollary 1. By reverse triangle inequality we have:

$$|W_1(\hat{\rho}^n_{i_1},\hat{\rho}^{'n}_{i_2}) - W_1(\rho_{i_1},\rho'_{i_2})| \leq W_1(\hat{\rho}^n_{i_1},\rho_{i_1}) + W_1(\hat{\rho}^{'n}_{i_2},\rho'_{i_2}) \lesssim 2n^{-1/(d_x+d_y)}$$

Writing:

$$\begin{aligned}
&\widehat{\mathcal{W}}_{\mathbf{W}_1+\mathbf{W}_1^{k,k}}(\mathbb{P}_m,\mathbb{P}'_m) \\
&= \sum_{i_1=1}^m\sum_{i_2=1}^m \gamma^*_{i_1i_2}\left(2W_1(\rho_{i_1},\rho'_{i_2}) + (W_1(\hat{\rho}^n_{i_1},\hat{\rho}^{'n}_{i_2}) - W_1(\rho_{i_1},\rho'_{i_2})) + \frac{1}{B^2}\sum_{b_1,b_2=1}^B (W_1(\hat{\rho}^{b_1,k}_{i_1},\hat{\rho}^{'b_2,k}_{i_2}) - W_1(\rho_{i_1},\rho'_{i_2}))\right) \\
&\lesssim 2\sum_{i_1=1}^m\sum_{i_2=1}^m \gamma^*_{i_1i_2}W_1(\rho_{i_1},\rho'_{i_2}) + 2n^{-1/(d_x+d_y)} + 2k^{-1/(d_x+d_y)} \\
&\lesssim 2\frac{1}{m}\sum_{i=1}^m W_1(\rho_i,\rho'_i) + 2n^{-1/(d_x+d_y)} + 2k^{-1/(d_x+d_y)}
\end{aligned}$$

where the last inequality follows from suboptimality of coupling with 0 off diagonal and $1/m$ on diagonal. Taking expectation we obtain:

$$\begin{aligned}
&\frac{2}{m}\sum_{i=1}^m \mathbb{E}_{S,S'}W_1(\rho_i,\rho'_i) + 2k^{-\frac{1}{d_x+d_y}} + 2n^{-\frac{1}{d_x+d_y}} \\
&= 2\mathbb{E}_{\rho,\rho'\sim\mathbb{P}}W_1(\rho,\rho') + 2k^{-\frac{1}{d_x+d_y}} + 2n^{-\frac{1}{d_x+d_y}}
\end{aligned}$$

□

Proof of Corollary 2.

$$\begin{aligned}\mathcal{E}^V_{\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}) &= (\mathcal{E}^V_{\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{R}^{V,k}_{\mathbb{P}}(\hat{f}^k_{m,n,B})) + (\mathcal{R}^{V,k}_{\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{R}^V_{\mathbb{P}}(f^{\mathcal{R}}_{\mathbb{P}})) \\ &+ (\mathcal{R}^V_{\mathbb{P}}(f^{\mathcal{R}}_{\mathbb{P}}) - \mathcal{E}^V_{\mathbb{P}}(f^{\mathcal{R}}_{\mathbb{P}})) + (\mathcal{E}^V_{\mathbb{P}}(f^{\mathcal{R}}_{\mathbb{P}}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}})) \\ &\lesssim k^{-\frac{1}{d_x+d_y}} + (\mathcal{R}^{V,k}_{\mathbb{P}}(\hat{f}^k_{m,n,B}) - \mathcal{R}^V_{\mathbb{P}}(f^{\mathcal{R}}_{\mathbb{P}})) + k^{-\frac{1}{d_x+d_y}} + (\mathcal{E}^V_{\mathbb{P}}(f^{\mathcal{R}}_{\mathbb{P}}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}))\end{aligned}$$

where we used Lemma 1. Taking expectation gives the result. To bound we use the lipschitzty of V and get :

$$\mathcal{E}^V_{\mathbb{P}}(f^{\mathcal{R}}_{\mathbb{P}}) - \mathcal{E}^V_{\mathbb{P}}(f_{\mathbb{P}}) \leq L_V \mathbb{E}_{\rho\sim\mathbb{P}}\mathbb{E}_{x\sim\rho} d_{\mathcal{Y}}(f_{\mathbb{P}}, f^{\mathcal{R}}_{\mathbb{P}}).$$

□

A.4 Proofs of Section 5

Proof of Proposition 1. Point 1) is proved in [Vuckovic et al., 2021]. For point 2) We use the same notations as in [Vuckovic et al., 2021] (attention kernel defined in Definition 7 in [Vuckovic et al., 2021]):

$$\begin{aligned}d_{\mathcal{X}}(\mathcal{T}_{m(x)}(x_t), \mathcal{T}_{m(y)}(y_t)) &= W_1(\delta_{x_t} A_{m(x)}, \delta_{y_t} A_{m(y)}) \\ &= W_1(\Pi[\Psi_{G(x_t,.)}(m(x))]L], (\Pi[\Psi_{G(y_t,.)}(m(y))]L]) \\ &\leq \tau_1(\Pi)\tau_1(L) W_1(\Psi_{G(x_t,.)}(m(x)), \Psi_{G(y_t,.)}(m(y)))\end{aligned}$$

where $\tau_1(\Pi) = d$, and $\tau_1(L) = ||W_V||_2$. By proposition 27 in [Vuckovic et al., 2021] we have:

$$W_1(\Psi_{G(x_t,.)}(m(x)), \Psi_{G(y_t,.)}(m(y))) \leq [\sqrt{d}\sqrt{\log T + \frac{1}{2e}||G||_{\text{lip}} + ||G||_\infty + \sqrt{d} + 2}](d_{\mathcal{X}}(x_t, y_t) + W_1(m(x), m(y)))$$

it follows that there exists a constant L' such that:

$$d_{\mathcal{X}}(\mathcal{T}_{m(x)}(x_t), \mathcal{T}_{m(y)}(y_t)) \leq L'(d_{\mathcal{X}}(x_t, y_t) + W_1(m(x), m(y)))$$

□

Proof of Proposition 2. We consider here euclidean distances. Define $\psi_T(\frac{1}{T}\sum_{j=1}^T \delta_{a_j}) = a_T$. Consider a and a' are two sequences constructed as in step 1 ($a = (x_1, y_1, \dots x_k, y_k, x)$ and $a' = (x'_1, y'_1, \dots x'_k, y'_k, x')$) in the case $\mathcal{X} = \mathcal{Y}$.

By iterating Lipschitzity of the push forward maps (we ignored in the computation of lipchitz constant the residual and feedforward that will result in multiplicative lipchitz constant on each layer) we have :

$$\begin{aligned}|f(\rho, x) - f(\rho', x')| &= |W_O(h^D_T - h'^D_T)| \\ &\leq ||W_O||_2 \left\|h^D_T - h'^D_T\right\| \\ &\leq ||W_O||_2 L'_D \left(d_{\mathcal{X}}(\psi_T(\mu_{D-1}), \psi_T(\mu'_{D-1})) + W_1(\mu_{D-1}, \mu'_{D-1})\right) \\ &\leq ||W_O||_2 L'_D \left(L'_{D-1}(d_{\mathcal{X}}(\psi_T(\mu_{D-2}), \psi_T(\mu'_{D-2})) + W_1(\mu_{D-2}, \mu'_{D-2})) + L_{D-1} W(\mu_{D-2}, \mu'_{D-2})\right) \\ &\leq ||W_O||_2 L'_D 2\max(L_{D-1}, L'_{D-1}) \left(d_{\mathcal{X}}(\psi_T(\mu_{D-2}), \psi_T(\mu'_{D-2})) + W(\mu_{D-2}, \mu'_{D-2})\right)\end{aligned}$$

Hence we obtain a recurrence:

$$\begin{aligned}&\left(d_{\mathcal{X}}(\psi_T(\mu_{D-1}), \psi_T(\mu'_{D-1})) + W_1(\mu_{D-1}, \mu'_{D-1})\right) \\ &\leq 2\max(L_{D-1}, L'_{D-1}) \left(d_{\mathcal{X}}(\psi_T(\mu_{D-2}), \psi_T(\mu'_{D-2})) + W(\mu_{D-2}, \mu'_{D-2})\right)\end{aligned}$$

Note that residual connection and feedforward networks will result with a multiplicative Lipschitz constant:

$$\begin{aligned}&\left(d_{\mathcal{X}}(\text{FN}_{D-1}[\psi_T(\mu_{D-1})], \text{FN}_{D-1}[\psi_T(\mu'_{D-1})]) + W_1((\text{FN}_{D-1})_\sharp \mu_{D-1}, (\text{FN}_{D-1})_\sharp \mu'_{D-1})\right) \\ &\leq 2L^{D-1}_{\text{FN}} \max(L_{D-1}, L'_{D-1}) \left(d_{\mathcal{X}}(\psi_T(\mu_{D-2}), \psi_T(\mu'_{D-2})) + W(\mu_{D-2}, \mu'_{D-2})\right)\end{aligned}$$

We can iterate this recurrence up to $\mu_0 = m(a)$ and $\mu'_0 = m(a')$, and $\psi_T(\mu_0) = x$ and $\psi_T(\mu'_0) = x'$. Hence there exists $L = ||W_O||_2 L'_D 2^{D-1} \Pi_{j=1}^{D-1} L_{\text{FN}}^j \max(L_j, L'_j)$ such that:

$$|f(\rho, x) - f(\rho', x')| \leq L \left(d(x, x') + W_1(m(a), m(a')) \right),$$

Let γ^* optimal coupling between ρ and ρ':

$$\begin{aligned}
&\sqrt{2}\frac{k}{2k+1} W_1(\rho, \rho') + \frac{1}{2k+1} d_{\mathcal{X}}(x, x') \\
&= \sqrt{2}\frac{k}{2k+1} \sum_{i,j} \gamma^*_{ij} \sqrt{(d^2_{\mathcal{X}}(x_i, x'_j) + d^2(y_i, y'_j)} + \frac{1}{2k+1} d_{\mathcal{X}}(x, x') \\
&\geq \sum_{i,j} \frac{k}{2k+1} \gamma^*_{ij} (d_{\mathcal{X}}(x_i, x'_j) + d_{\mathcal{X}}(y_i, y'_j)) + \frac{1}{2k+1} d_{\mathcal{X}}(x, x')
\end{aligned}$$

Define for $i = 1 \dots k, j = 1 \dots k$: $\tilde{\gamma}_{2i-1,2j-1} = \frac{k}{2k+1}\gamma^*_{ij}$ and $\tilde{\gamma}_{2i-1,2j} = 0$ and $\tilde{\gamma}_{2i,2j-1} = 0$ and $\tilde{\gamma}_{2i,2j} = \frac{k}{2k+1}\gamma^*_{ij}$, $\tilde{\gamma}_{2k+1,\ell} = 0$ and $\tilde{\gamma}_{\ell,2k+1} = 0$ for $\ell = 1 \dots 2k$ and $\tilde{\gamma}_{2k+1,2k+1} = \frac{1}{2k+1}$. It is easy to see that $\tilde{\gamma}$ is coupling between $m(a)$ and $m(a')$:

$$\tilde{\gamma} = \begin{array}{c} \\ x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_k \\ y_k \\ x \end{array} \begin{array}{c} \begin{array}{cccccccc} x'_1 & y'_1 & x'_2 & y'_2 & \cdots & x'_k & y'_k & x' \end{array} \\ \begin{pmatrix}
\frac{k}{2k+1}\gamma^*_{11} & 0 & \frac{k}{2k+1}\gamma^*_{12} & 0 & \cdots & \frac{k}{2k+1}\gamma^*_{1k} & 0 & 0 \\
0 & \frac{k}{2k+1}\gamma^*_{11} & 0 & \frac{k}{2k+1}\gamma^*_{12} & \cdots & 0 & \frac{k}{2k+1}\gamma^*_{1k} & 0 \\
\frac{k}{2k+1}\gamma^*_{21} & 0 & \frac{k}{2k+1}\gamma^*_{22} & 0 & \cdots & \frac{k}{2k+1}\gamma^*_{2k} & 0 & 0 \\
0 & \frac{k}{2k+1}\gamma^*_{21} & 0 & \frac{k}{2k+1}\gamma^*_{22} & \cdots & 0 & \frac{k}{2k+1}\gamma^*_{2k} & 0 \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\
\frac{k}{2k+1}\gamma^*_{k1} & 0 & \frac{k}{2k+1}\gamma^*_{k2} & 0 & \cdots & \frac{k}{2k+1}\gamma^*_{kk} & 0 & 0 \\
0 & \frac{k}{2k+1}\gamma^*_{k1} & 0 & \frac{k}{2k+1}\gamma^*_{k2} & \cdots & 0 & \frac{k}{2k+1}\gamma^*_{kk} & 0 \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{2k+1}
\end{pmatrix} \end{array}$$

Rows and columns of $\tilde{\gamma}$ sum to $\frac{1}{2k+1}$. It follows that:

$$\sqrt{2}\frac{k}{2k+1} W_1(\rho, \rho') + \frac{1}{2k+1} d_{\mathcal{X}}(x, x') \geq \sum_{\ell,\ell'=1}^{2k+1} \tilde{\gamma}_{\ell,\ell'} d_{\mathcal{X}}(a_\ell, a'_{\ell'}) \geq W_1(m(a), m(a'))$$

and therefore we have:

$$W_1(m(a), m(a')) \leq \sqrt{2}\frac{k}{2k+1} W_1(\rho, \rho') + \frac{1}{2k+1} d_{\mathcal{X}}(x, x') \leq W_1(\rho, \rho') + d_{\mathcal{X}}(x, x')$$

and hence :

$$|f(\rho, x) - f(\rho', x')| \leq L d(x, x') + W_1(m(a), m(a')) \leq L(2 d_{\mathcal{X}}(x, x') + W_1(\rho, \rho')) \leq 2L(d_{\mathcal{X}}(x, x') + W_1(\rho, \rho'))$$

it follows that f is lipschitz in both argument. for the case $\mathcal{X} \neq \mathcal{Y}$, a similar proof relates $W_1(m(a), m(a'))$ to $W_1(\rho, \rho')$. □

Proposition 3. *For two independent samples from the meta distributions $\mathbb{P}_m = \frac{1}{m}\sum_{i=1}^m \delta_{\rho_i}$ and $\mathbb{P}'_m = \frac{1}{m}\sum_{i=1}^m \delta_{\rho'_i}$. Let $\hat{\rho} = \frac{1}{k}\sum_{i=1}^k \delta_{(x_i, y_i)}$ and $x \in \mathcal{X}$ define $a = (x_1, y_1 \dots x_k, y_k, x)$. Define attention blocks $\mathcal{A}_j(\mu_j) = (FN_j)_\sharp \left[(\mathcal{T}^j_{\mu_j})_\sharp (\mu_j) \right]$ and $\psi_T(\frac{1}{T}\sum_{j=1}^T \delta_{a_j}) = a_T$. The encoder transformer for few shot learning can be written as:*

$$f(\hat{\rho}, x) = \underbrace{g \circ \psi_T \circ \mathcal{A}_D \circ \dots \mathcal{A}_{r-1}}_{h} \circ \underbrace{\mathcal{A}_r \circ \dots \mathcal{A}_1}_{\mathcal{A}}(m(a))$$

we have:

$$\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) \leq \mathbb{E}_S \hat{\mathcal{R}}^V_{k,S}(\hat{f}^k_{m,n,B}) + \sqrt{2} L_V (1 + L_{\mathcal{F}}) \widehat{\mathcal{W}}(\mathcal{A}_\sharp \mathbb{P}_m, \mathcal{A}_\sharp \mathbb{P}'_m),$$

Where

$$\widehat{\mathcal{W}}(\mathcal{A}_\sharp \mathbb{P}_m, \mathcal{A}_\sharp \mathbb{P}'_m)$$
$$= \min_{\gamma \in \Gamma(\mathbb{P}_m, \mathbb{P}'_m)} \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \gamma_{i_1 i_2} \left(\frac{1}{B^2} \sum_{b_1, b_2 = 1}^{B} \left(W^{\mathcal{A},k}(\hat{\rho}^n_{i_1}, \hat{\rho}^n_{i_2}) + W_1(\mathcal{A}_\sharp \hat{\rho}^{b_1,k}_{i_1}, \mathcal{A}_\sharp \hat{\rho}^{'b_2,k}_{i_2}) \right) \right), \quad (24)$$

and:

$$W^{\mathcal{A},k}(\hat{\rho}^n_{i_1}, \hat{\rho}^n_{i_2}) = \inf_{\tilde{\gamma} \in \Gamma(\hat{\rho}^n_{i_1}, \hat{\rho}^n_{i_2})} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \tilde{\gamma}^{i_1 i_2}_{j_1 j_2} \frac{1}{B^2} \sum_{b_1, b_2 = 1}^{B} c(x_{i_1 j_1}, x'_{i_2, j_2}, y_{i_1 j_1}, y'_{i_2 j_2}, \hat{\rho}^{'b_2,k}_{i_2}, \hat{\rho}^{b_1,k}_{i_1})$$

where:

$$c(x_{i_1 j_1}, x'_{i_2, j_2}, y_{i_1 j_1}, y'_{i_2 j_2}, \hat{\rho}^{'b_2,k}_{i_2}, \hat{\rho}^{b_1,k}_{i_1})$$
$$= \sqrt{d^2_{\mathcal{X}}(\psi_T(\mathcal{A}_\sharp(m(\psi(\hat{\rho}^{b_1,k}_{i_1}) \oplus x_{i_1 j_1})), \psi_T(\mathcal{A}_\sharp(m(\psi(\hat{\rho}^{b_2,k}_{i_2}) \oplus x'_{i_2 j_2}))) + d^2_{\mathcal{Y}}(y_{i_1 j_1}, y'_{i_2 j_2})}$$

For a sequence $a = (a_1, \ldots, a_T)$, we define $a \oplus x = (a_1, \ldots, a_T, x)$, and ψ maps the empirical measure to a sequence.

Note that the lifted wasserstein is evaluated in the hidden space following an intermediate self attention block. Note that in (25) the test points are now not contributing to the generalization independent of the context since $W^{\mathcal{A},k}$ depends on the context. The generalization bound given in the attention latent space can be much smaller then the one in the input domain , by lipschitzity the input domain bound is an upper bound. Note that this type of bounds is not uniform on the function class and hold only for a given function. The input space bound is uniform on the function class.

Proof of Proposition 3. Let $a = (x_1, y_1 \ldots x_k, y_k, x)$, $\mathcal{A}_j(\mu_j) = (\text{FN}_j)_\sharp \left[(\mathcal{T}^j_{\mu_j})_\sharp (\mu_j) \right]$ and $\psi_T(\frac{1}{T} \sum_{j=1}^{T} \delta_{a_j}) = a_T$.

$$f(\hat{\rho}, x) = \underbrace{g \circ \psi_T \circ \mathcal{A}_D \circ \ldots \mathcal{A}_{r-1}}_{h} \circ \underbrace{\mathcal{A}_r \circ \ldots \mathcal{A}_1}_{\mathcal{A}}(m(a))$$

Note that is h lipschitz in $\mathcal{A}(m(a))$ in the sense that there exists $L_{\mathcal{F}}$:

$$|h(\mathcal{A}(m(a))) - h(\mathcal{A}(m(b)))| \leq L_{\mathcal{F}} \left(W_1(\mathcal{A}(m(a)), \mathcal{A}(m(b)) + d_{\mathcal{X}}(\psi_T(\mathcal{A}(m(a))), \psi_T(\mathcal{A}(m(b))))) \right)$$

The proof follows from the proof of Theorem 4. The main difference with theorem 4 is that the lipchiizity is not in the input $x_{i_1 j_1}$ but in $\psi_T(\mathcal{A}_\sharp(m(\psi(\hat{\rho}^{b_1,k}_{i_1}) \oplus x_{i_1 j_1})$ that is the last output in the tensor output of the attention block at layer r. The lipschitzty is also in the empirical measure of output of the attention block $\mathcal{A}_\sharp \hat{\rho}^{b_1,k}_{i_1}$ and not in $\hat{\rho}^{b_1,k}_{i_1}$. Note that in the test points $x_{i_1 j_1}$ and the contexts don't appear any more in independent contributions.

$$\mathbb{E}_S \left(\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \hat{\mathcal{R}}^V_{k,S}(\hat{f}^k_{m,n,B}) \right) \leq \mathbb{E}_S \mathbb{E}_{S'} \sup_{h \in \{h | f = h \circ \mathcal{A}\}} \left| \hat{\mathcal{R}}^V_{k,S'}(f) - \hat{\mathcal{R}}^V_{k,S}(f) \right|$$
$$\leq \sqrt{2} L_V (1 + L_{\mathcal{F}}) \widehat{\mathcal{W}}(\mathcal{A}_\sharp \mathbb{P}_m, \mathcal{A}_\sharp \mathbb{P}'_m)$$

$$\widehat{\mathcal{W}}(\mathcal{A}_\sharp \mathbb{P}_m, \mathcal{A}_\sharp \mathbb{P}'_m)$$
$$= \min_{\gamma \in \Gamma(\mathbb{P}_m, \mathbb{P}'_m)} \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \gamma_{i_1 i_2} \left(\frac{1}{B^2} \sum_{b_1, b_2 = 1}^{B} \left(W^{\mathcal{A},k}(\hat{\rho}^n_{i_1}, \hat{\rho}^n_{i_2}) + W_1(\mathcal{A}_\sharp \hat{\rho}^{b_1,k}_{i_1}, \mathcal{A}_\sharp \hat{\rho}^{'b_2,k}_{i_2}) \right) \right), \quad (25)$$

where:

$$W^{\mathcal{A},k}(\hat{\rho}^n_{i_1}, \hat{\rho}^n_{i_2}) = \inf_{\tilde{\gamma}\in\Gamma(\hat{\rho}^n_{i_1}, \hat{\rho}^n_{i_2})} \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} \tilde{\gamma}^{i_1 i_2}_{j_1 j_2} \frac{1}{B^2} \sum_{b_1,b_2=1}^{B} c(x_{i_1 j_1}, x'_{i_2,j_2}, y_{i_1 j_1}, y'_{i_2 j_2}, \hat{\rho}^{'b_2,k}_{i_2}, \hat{\rho}^{b_1,k}_{i_1})$$

where:

$$\begin{aligned} &c(x_{i_1 j_1}, x'_{i_2,j_2}, y_{i_1 j_1}, y'_{i_2 j_2}, \hat{\rho}^{'b_2,k}_{i_2}, \hat{\rho}^{b_1,k}_{i_1}) \\ &= \sqrt{d^2_{\mathcal{X}}(\psi_T(\mathcal{A}_\sharp(m(\psi(\hat{\rho}^{b_1,k}_{i_1}) \oplus x_{i_1 j_1})), \psi_T(\mathcal{A}_\sharp(m(\psi(\hat{\rho}^{b_2,k}_{i_2}) \oplus x'_{i_2 j_2}))) + d^2_{\mathcal{Y}}(y_{i_1 j_1}, y'_{i_2 j_2})} \end{aligned}$$

For a sequence $a = (a_1, \ldots, a_T)$, we define $a \oplus x = (a_1, \ldots, a_T, x)$.

Hence we have for encoder attention:

$$\mathbb{E}_S\left(\mathcal{R}^V_{k,\mathbb{P}}(\hat{f}^k_{m,n,B}) - \hat{\mathcal{R}}^V_{k,S}(\hat{f}^k_{m,n,B})\right) \leq \sqrt{2} L_V (1 + L_{\mathcal{F}}) \widehat{\mathcal{W}}(\mathcal{A}_\sharp \mathbb{P}_m, \mathcal{A}_\sharp \mathbb{P}'_m).$$

□