APPENDIX A: OMITTED PROOFS IN PAPER FOR ICML2018

XUDONG PAN, MI ZHANG, DAIZONG DING

Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science
Fudan University
China

Detailed proofs for all the theorems, lemmas and propositions omitted from our paper will be given here in a rigorous form. We provide them as a supplementary because we would like the audience of our paper to focus more on the development of our theory, and limit of space. Our proofs are mainly based on the texts of Chung [1], Rudin [2] and Nakahara [3].

1. Introduction

[No Theorems or Lemmas]

2. Preliminaries

Lemma 2.1. *Given a smooth manifold $\mathcal{M} = \{(U_i, \varphi_i)\}_{i=1}^K$ with pairwise disjointness and $\{\mu_k\}_{k=1}^K$ as the probability measures supported on $\{\varphi_i(U_i)\}_{i=1}^K$ correspondingly, a function $\mu_{\mathcal{M}} : \mathcal{B}(\mathcal{M}) \to [0, 1]$ is defined by*

$$d\mu_{\mathcal{M}}(s) = \frac{1}{K} \sum_{i=1}^{K} \mathbf{1}_{s \in U_i} d\mu_i \circ \varphi_i(s) \tag{1}$$

Then $\mu_{\mathcal{M}}$ is a probability measure defined on $\mathcal{M}$.

Proof. First claim $\mu_i \circ \varphi_i$ is a measure, which comes from the easy observation that for each $i \in [K]$, and countably many disjoint sets $\{A_n\}_{n=1}^{\infty} \subset \mathcal{B}(\mathcal{M})$, the Borel sets constructed over $\mathcal{M}$ as a topological space.

$$\varphi_i(\cup_{n=1}^{\infty} A_n) = \cup_{n=1}^{\infty} \varphi_i(A_n)$$

and since φ_i itself is a homeomorphism, which indicates the one-to-one property, we have the disjointness of sets $\{\varphi_i(A_n)\}_{n=1}^{\infty}$.

Thus from the assumption that μ_i is a probability measure, the countable additivity of $\mu_i \circ \varphi_i$ on $\mathcal{B}(\mathcal{M})$ is thus proved as

E-mail address: mi_zhang@fudan.edu.cn.

$$\mu_i \circ \varphi_i(\cup_{n=1}^{\infty} A_n) = \sum_{i=1}^{\infty} \mu_i(\varphi_i(A_n))$$

, which directly leads to the assertion that $\mu_i \circ \varphi_i$ is a measure.

Next, we would like to prove it is indeed a probability measure, which needs to prove the normalization condition.

We directly take integral over the manifold $\mathcal{M}$ with the derivative form of measure $\mu_{\mathcal{M}}$, as is defined.

$$\int_{\mathcal{M}} d\mu_{\mathcal{M}} \tag{2}$$

$$\overset{\text{substitute}}{=} \frac{1}{K} \int_{\mathcal{M}} \sum_{i=1}^{K} \mathbf{1}_{s \in U_i} d\mu_i \circ \varphi_i(s) \tag{3}$$

$$\overset{\text{exchange}}{=} \frac{1}{K} \sum_{i=1}^{K} \int_{U_i} d\mu_i \circ \varphi_i(s) \tag{4}$$

$$\overset{\text{change of variable}}{=} \frac{1}{K} \sum_{i=1}^{K} \int_{\varphi_i(U_i)} d\mu_i(s) \tag{5}$$

$$= 1 \tag{6}$$

Thus we have checked the normalization condition, which in turn proves the lemma. □

3. Natural Localization of cWGAN-Loss

Lemma 3.1. *Consider Riemmanian manifold $(\mathcal{N}, \tau)$ with curvature locally bounded above and below, $\tau \in C^{\infty}$ and its induced distance function denoted as $d_{\mathcal{N}}$, then for any path-independent function $f : \mathcal{N} \times \mathcal{N} \to \mathbb{R}^+ \cup \{0\}$, there exists a Riemmanian metric $\tau^{'}$ on $\mathcal{N}$, induced by the distance function*

$$d^{'}_{\mathcal{N}}(x, y) = f(x, y) d_{\mathcal{N}}(x, y) \quad \forall x, y \in \mathcal{N}$$

Proof. The proof is mainly based on a previous result in [4, 5], which asserts certain sufficient conditions for a synthetic distance function $d^{'}_{\mathcal{N}} : \mathcal{N} \times \mathcal{N} \to \mathbb{R}^+ \cup \{0\}$ on Riemmanian manifold $(\mathcal{N}, \tau)$ to be compatible with some Riemmanian metric on $\mathcal{N}$. That is, besides the conditions innate to the manifold

- curvature locally bounded above and below, i.e. $\forall s \in \mathcal{N}, \exists c_1, c_2,\ 0 < c_1 < c_2 < \infty$ and $c_1 < \Gamma^k_{ij}(s) < c_2$.
- $\tau \in C^{\infty}$, which means it is infinitely differentiable locally. In fact, the assumption can be relaxed to $\tau \in C^{1,\alpha}$, for any $\alpha > 0$.

, the condition imposed on the synthetic distance function is

- $d_{\mathcal{N}}^{'}$ is a *path-metric*, i.e. $\forall s_1, s_2 \in \mathcal{N}$, consider the set of paths connecting s_1, s_2, that is, the set of curves $\mathcal{P}_{s_1 \to s_2} = \{p : [0,1] \to \mathcal{M} | p(0) = s_1, p(1) = s_2\}$, there exists an functional $L : \mathcal{P}_{s_1 \to s_2} \to \mathbb{R}^+ \cup \{0\}$, s.t.

$$d_{\mathcal{N}}^{'}(s_1, s_2) = \inf_{p \in \mathcal{P}_{s_1 \to s_2}} L(p)$$

Thus let us turn back to our case, the synthetic distance function is actually expanded from an existing distance function on $\mathcal{N}$, induced by Riemmanian metric τ. As is well known, the induced distance $d_{\mathcal{N}}$ itself has the form

$$d_{\mathcal{N}}(s_1, s_2) = \inf_{p \in \mathcal{P}_{s_1 \to s_2}} L(p)$$

where L is called the length of curve p, defined as

$$L(p) \doteq \int_0^1 \sqrt{\sum_{i,j} g_{ij}(p(t)) \frac{\partial x^i}{\partial t} \frac{\partial x^j}{\partial t}} dt$$

Since from the assumption that $f(\bullet, \bullet)$ is path-independent, we are able to define the following functional L_f (easy to check its well-definedness),

$$L_f(p) \doteq f(p(0), p(1)) \ \forall p \in \mathcal{P}_{s_1 \to s_2}$$

Thus it is obvious that, by constructing $L^{'}$ as

$$L^{'}(p) = L_f(p) \bullet L(p) \ \forall p \in \mathcal{P}_{s_1 \to s_2}$$

, our synthetic distance function $d_{\mathcal{N}}^{'}$ is a path metric thus induced from some Riemmian metric on $\mathcal{N}$, which finishes our proof.

□

3.1. **Omitted Steps for Renormalization to Obtain Eq. 14.** After we rearrange $d_{\mathcal{N}}(G(s), t) d\gamma$ as $d_{\mathcal{N}}^{'}(G(s), t) d\gamma'$, the boundary condition $\int_{\mathcal{M}} \int_{\mathcal{N}} d\gamma^{'} = 1$ requires renormalization. By introducing an additional matrix $A \in \mathbf{H}(K)$ s.t. $\mathbf{H}(K) \doteq \{A \in \mathbb{R}^{K \times K} | \forall j \in K, \sum_i A^{ij} = K; \forall i, j \in [K], A^{ij} \geq 0\}$, the cWGAN-loss $\min_G \mathcal{L}_{adv}^{'}(G)$ can be reformulated as

$$\min_G \min_{A \in \mathbf{H}(K)} \sum_{i=1}^{K} \sum_{j=1}^{K} \int_{U_i} \int_{V_j} A^{ij} d_{\mathcal{N}}^{'}(G(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_j(t) \tag{7}$$

Proof. We start from the form,

$$\min_G \inf_{\gamma \in \Pi(\mu_{\mathcal{M}}, \nu_{\mathcal{N}})} \sum_{i=1}^{K} \sum_{j=1}^{K} \int_{U_i} \int_{V_j} d_{\mathcal{N}}(G(s), t) d\gamma(s, t) \tag{8}$$

when we rearrange the form with

$$d\gamma(s, t) = d\gamma(t|s) d\mu_{\mathcal{M}}(s) = \Delta(f_\gamma(s), t) d\mu_{\mathcal{M}}(s) d\nu_{\mathcal{N}}(t)$$

, we obtain

$$\min_{G}\min_{f_\gamma}\sum_{i=1}^{K}\sum_{j=1}^{K}\int_{U_i}\int_{V_j}\Delta(f_\gamma(s),t)d_{\mathcal{N}}(G(s),t)d\mu_{\mathcal{M}}(s)d\nu_{\mathcal{N}}(t) \tag{9}$$

And since we apply the equivalence of $\min_G \min_{f_\gamma}$ and $\min_G$, the boundary condition $\int_{\mathcal{M}}\int_{\mathcal{N}} d\gamma^{'} = 1$ may be broken. Thus we introduce additional variable $A \in \mathbf{H}(K)$ to maintain the normalization condition, as can be checked by

$$\int_M d\gamma^{'} \tag{10}$$

$$= \frac{1}{K^2}\sum_{j=1}^{K}\int_{U_i}(\int_{V_j}\sum_{i=1}^{K}A^{ij}d\nu_j)d\mu_i \tag{11}$$

$$= 1 \tag{12}$$

Note here we have applied the formulae for constructed probability measures on manifolds as

$$d\mu_{\mathcal{M}}(s) \doteq \frac{1}{K}\sum_{i=1}^{K}\mathbf{1}_{s\in U_i}d\mu_i \circ \varphi_i(s) \tag{13}$$

$$d\nu_{\mathcal{N}}(t) \doteq \frac{1}{K}\sum_{i=1}^{K}\mathbf{1}_{t\in U_i}d\nu_i \circ \psi_i(t) \tag{14}$$

Finally, by inserting the A^{ij} term into the original optimization problem above, we will obtain the final form as follows,

$$\min_{G}\min_{A\in\mathbf{H}(K)}\sum_{i=1}^{K}\sum_{j=1}^{K}\int_{U_i}\int_{V_j}A^{ij}d^{'}_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_j(t) \tag{15}$$

□

Theorem 3.1. *[Natural Localization of Adversarial Loss] For any p ∈ Sym(K), the optimization problem below*

$$\min_{G\in F_p}\min_{A\in\mathbf{H}(K)}\sum_{i=1}^{K}\sum_{j=1}^{K}\int_{U_i}\int_{V_j}A^{ij}d^{'}_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_j(t) \tag{16}$$

is equivalent to

$$\min_{G\in F_p}\sum_{i=1}^{K}\int_{U_i}\int_{V_{p(i)}}d^{'}_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_{p(i)}(t) \tag{17}$$

In other words, the optimal $A^* \in \mathbf{H}(K)$ has the closed form as

$$(A^*)^{ij} = K\delta_j^{p(i)} \tag{18}$$

where $\delta_j^{p(i)}$ is the Kronecker delta function.

Proof. Fix $i, j \in [K]$, s.t. $j \neq p(i)$ and arbitrary $G \in F_p$. We first compare the following two terms

$$T_{\text{non-paired}} = \int_{U_i}\int_{V_j} d'_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_j(t)$$

and

$$T_{\text{paired}} = \int_{U_i}\int_{V_{p(i)}} d'_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_{p(i)}(t)$$

Notice, for any $s \in U_i$, $G(s) \in V_{p(i)} \cap V_j = \varnothing$, which comes from the assumption that $G \in F_p$ and $j \neq p(i)$, which thus leads to $\forall t \in V_{p(i)}, t' \in V_j$, $d_{\mathcal{N}}(G(s),t) \leq d_{\mathcal{N}}(G(s),t')$, according to the compatibility of distance function with the assumed inner-relatedness.

And thus $T_{\text{non-paired}} \geq T_{\text{paired}}$. Then we relieve the fixation of j. It is easy to see,

$$\sum_{j=1}^{K} A^{ij}\int_{U_i}\int_{V_j} d'_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_{p(i)}(t) \geq K\int_{U_i}\int_{V_{p(i)}} d'_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_{p(i)}(t)$$

, which is equivalent to say the optimal $(A^{ij})^* = K\delta^j_{p(i)}$ for each $j \in [K]$.

Similarly, we have for each $A \in \mathbf{H}(K)$,

$$\sum_{i=1}^{K}\sum_{j=1}^{K}\int_{U_i}\int_{V_j} A^{ij}d'_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_j(t) \geq \sum_{i=1}^{K}\int_{U_i}\int_{V_{p(i)}} d'_{\mathcal{N}}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_{p(i)}(t)$$

, which brings the equivalence between optimization problems above. □

4. Generalization for Conditional GAN

Theorem 4.1. *Consider generator $G : \mathbb{R}^d \to \mathbb{R}^d$ satisfying Lipschitz condition with constant M_G and μ_X, ν_Y are probability measures on $\mathbb{R}^d$ respectively with $\{x_i\}_{i=1}^{n_X} \overset{i.i.d.}{\sim} \mu_X$ and $\{y_i\}_{i=1}^{n_Y} \overset{i.i.d.}{\sim} \nu_Y$.*

Assume the classical generalization bound satisfies the following inequality with probability $1-\delta$

$$\mathbb{E}_{x\sim\mu_X, y\sim\nu_Y}\|G(x)-y\| - \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}\frac{\|G(x_i)-y_j\|}{n_X n_Y} < \epsilon_{classical} \tag{19}$$

where $\epsilon_{classical} \doteq \epsilon(n_X, n_Y, \mu_X, \nu_Y, \delta)$ the upper bound and ERM-principle [6] is satisfied with η (i.e. $\frac{1}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}\|G(x_i)-y_j\| < \eta$), then G generalizes with (n_X, n_Y) training samples and error ϵ_{adv} with probability $1-\delta$, i.e.

$$D_{LK}(G(\hat{\mu}_X^{n_X}), \nu_Y) - D_{LK}(\hat{\nu}_Y^{n_Y}, \nu_Y) < \epsilon_{adv} \tag{20}$$

if the following condition is satisfied

$$\epsilon_{classical} - \epsilon_{adv} + \eta < D_{LK}(\nu_Y, \hat{\nu}_Y^{n_Y}) - M_G D_{LK}(\mu_X, \hat{\mu}_X^{n_X}) \tag{21}$$

Proof. Let us start by bounding the term $D_{LK}(G(\hat{\mu}_X^{n_X}), \nu_Y)$,

$$D_{LK}(G(\hat{\mu}_X^{n_X}), \nu_Y) \tag{22}$$

$$\overset{\text{by def.}}{=} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|G(x) - y\| d\hat{\mu}_X^{n_X}(x) d\nu_Y(y) \tag{23}$$

$$\overset{\text{norm ineq.}}{\leq} \int\int \|G(x) - G(x^{'})\| d\hat{\mu}_X^{n_X}(x) d\mu_X(x^{'}) + \int\int \|G(x) - y\| d\mu_X(x) d\nu_Y(y) \tag{24}$$

$$\overset{\text{Lip.}}{\leq} M_G \int\int \|(x) - x^{'}\| d\hat{\mu}_X^{n_X}(x) d\mu_X(x^{'}) + \int\int \|G(x) - y\| d\mu_X(x) d\nu_Y(y) \tag{25}$$

$$\overset{\text{gen. bound, with probability 1-}\delta}{\leq} M_G D_{LK}(\mu_X, \hat{\mu}_X^{n_X}) + \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} \frac{\|G(x_i) - y_j\|}{n_X n_Y} + \epsilon_{\text{classical}} \tag{26}$$

$$\overset{\text{ERM}}{\leq} M_G D_{LK}(\mu_X, \hat{\mu}_X^{n_X}) + \eta + \epsilon_{\text{classical}} \tag{27}$$

And the definition of generation in adversarial learning sense requires

$$D_{LK}(G(\hat{\mu}_X^{n_X}), \nu_Y) - D_{LK}(\hat{\nu}_Y^{n_Y}, \nu_Y) < \epsilon_{\text{adv}}$$

By direct inserting the last expressions during the estimation above, we have obtained the generic inequality to guarantee generalization sufficiently,

$$\epsilon_{\text{classical}} - \epsilon_{\text{adv}} + \eta < D_{LK}(\nu_Y, \hat{\nu}_Y^{n_Y}) - M_G D_{LK}(\mu_X, \hat{\mu}_X^{n_X})$$

□

5. Benefits of Localization and Conditions of Generalization

Proposition 5.1. *Consider the probability measure underlying the global task as* $\mu_X = \frac{1}{K}\sum_{i=1}^K \mu_i$ *and* $\nu_Y = \frac{1}{K}\sum_{i=1}^K \nu_i$ *in Euclidean sense and*

$$\epsilon_{adv}^{local} = \frac{1}{K}\sum_{i=1}^K D_{LK}(\mu_i, \hat{\mu}_i^m) \tag{28}$$

$$\epsilon_{adv}^{global} = D_{LK}(\frac{1}{K}\sum_{i=1}^K \mu_i, \hat{\mu}_X^{Km}) \tag{29}$$

, if the compatibility with inner-relatedness $(\forall i, j \in [K], D_{LK}(\mu_i, \mu_j) \geq D_{LK}(\mu_i, \mu_i))$ *is satisfied, then*

$$\epsilon_{adv}^{local} < \epsilon_{adv}^{global}$$

Proof. First let us consider the situation when $m \to \infty$, which correspondingly leads to $\hat{\mu}_X^{Km} \to \frac{1}{K}\sum_{i=1}^K \mu_i$ and $\forall i \in [K]$, $\hat{\mu}_i^m \to \mu_i$.

Thus by honestly inserting the term into the definition of D_{LK}, we have

$$\epsilon^{global}_{\text{adv},\, n_X \to \infty} = D_{LK}(\frac{1}{K}\sum_{i=1}^K \mu_i, \frac{1}{K}\sum_{i=1}^K \mu_i) \tag{30}$$

$$= \frac{1}{K^2}\sum_{i=1}^K D_{LK}(\mu_i, \mu_i) + \frac{1}{K(K-1)}\sum_{i<j\in[K]} D_{LK}(\mu_i, \mu_j) \tag{31}$$

$$> \frac{1}{K}\sum_{i=1}^K D_{LK}(\mu_i, \mu_i) = \epsilon^{local}_{\text{adv},\, n_X \to \infty} \tag{32}$$

The last inequality comes from the observation that, $\forall i, j \in [K]$

$$D_{LK}(\mu_i, \mu_i) = \|\mu_i - \mu_i\| + 2tr(\Sigma_{\mathcal{M}}) \leq \|\mu_i - \mu_j\| + 2tr(\Sigma_{\mathcal{M}}) = D_{LK}(\mu_i, \mu_j)$$

Next, we would like to consider the case for arbitrary m and the inequality with corresponding optimal empirical estimators $\{\hat{\mu}_i^m\}_{i=1}^K$. Thus with Kn samples, the optimal estimator for the global distribution as a mixture of gaussians with the mixture coefficients priorly known is $\frac{1}{K}\sum_{i=1}^K \hat{\mu}_i^m$. With a similar procedure as above,

$$\epsilon^{global}_{\text{adv}} = D_{LK}(\frac{1}{K}\sum_{i=1}^K \mu_i, \frac{1}{K}\sum_{i=1}^K \hat{\mu}_i^m) \tag{33}$$

$$= \frac{1}{K^2}\sum_{i=1}^K D_{LK}(\mu_i, \hat{\mu}_i^m) + \frac{1}{K(K-1)}\sum_{1\leq i<j\leq K} D_{LK}(\mu_i, \hat{\mu}_j^m) \tag{34}$$

$$\geq \epsilon^{local}_{\text{adv}} \tag{35}$$

with the following observation

$$D_{LK}(\mu_i, \hat{\mu}_j^m) = \|\mu_i - \hat{\mu}_j + \hat{\mu}_j - \hat{\mu}_j^m\| \tag{36}$$

$$\overset{\text{consider } \mu_i - \hat{\mu}_j \text{ white noise}}{=} \|\mu_i - \hat{\mu}_j\| + \|\hat{\mu}_j - \hat{\mu}_j^m\| \tag{37}$$

$$> \|\hat{\mu}_j - \hat{\mu}_j^m\| = D_{LK}(\mu_j, \hat{\mu}_j^m) \tag{38}$$

□

Lemma 5.1. *$\forall i \in [K]$, consider a measureable mapping $\tilde{f} : U_i \to V_i$ with $f \doteq \psi_i \circ \tilde{f} \circ \varphi_i^{-1}$ satisfies Lipschitz condition, then $\int_{U_i}\int_{V_i} d'_{\mathcal{N}}(\tilde{f}(s), t)d\tilde{\mu}_i(s)d\tilde{\nu}_i(t) \simeq D_{LK}(f(\mu_i), \nu_i)$, i.e. there exists constants $0 < C_l < C_u < \infty$ such that*

$$C_l < \frac{\int_{U_i}\int_{V_i} d'_{\mathcal{N}}(\tilde{f}(s), t)d\tilde{\mu}_i(s)d\tilde{\nu}_i(t)}{D_{LK}(f(\mu_i), \nu_i)} < C_u \tag{39}$$

Proof. With the measureability of $\tilde{f}$ and smoothness of φ_i, ψ_i, the induced mapping $\tilde{\nu}^{'} = \frac{1}{\tilde{E}}\tilde{f}(\mu_i)$ and $\nu^{'} = \frac{1}{E}(\psi_i \circ \tilde{f})(\mu_i)$ are also probability measures respectively on $\tilde{f}(U_i) \subset V_i$ and $(\psi_i \circ \tilde{f})(U_i) \subset \psi_i(V_i)$ (with $\tilde{E}, E$ some normalizing factor).

Observe the following bounds, which comes from the inclusion relations above,

$$\int_{\tilde{f}(U_i)} \int_{V_i} d_{\mathcal{N}}^{'}(\tilde{f}(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_i(t) \leq \int_{V_i} \int_{V_i} d_{\mathcal{N}}^{'}(t^{'}, t) d\tilde{\nu}^{'}(t^{'}) d\tilde{\nu}_i(t)$$

$$\int_{(\psi_i \circ \tilde{f})(U_i)} \int_{\psi_i(V_i)} \|f(s) - t\| d\mu_i(s) d\nu_i(t) \leq \int_{\psi_i(V_i)} \int_{\psi_i(V_i)} \|t^{'} - t\| d\nu^{'}(t^{'}) d\nu_i(t)$$

With the Lipschitz condition of f, it can be asserted that $supp(\tilde{\nu}^{'})$ and $supp(\nu^{'}(t^{'}))$ is bounded by a finite disk respectively on $V_i, \psi_i(V_i)$. Together with the gaussian assumption, we have $tr(\Sigma_{\mathcal{M}})$, $tr(\Sigma_{\mathcal{N}}) < \infty$, which leads to the boundedness of $supp(\nu_i)$, $supp(\tilde{\nu}_i)$ as well.

Thus we have

$$\int_{V_i} \int_{V_i} d_{\mathcal{N}}^{'}(t^{'}, t) d\tilde{\nu}^{'}(t^{'}) d\tilde{\nu}_i(t) < \infty$$

$$\int_{\psi_i(V_i)} \int_{\psi_i(V_i)} \|t^{'} - t\| d\nu^{'}(t^{'}) d\nu_i(t) < \infty$$

With the finiteness of right side, we are able to claim

$$\int_{\tilde{f}(U_i)} \int_{V_i} d_{\mathcal{N}}^{'}(\tilde{f}(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_i(t) \simeq \int_{V_i} \int_{V_i} d_{\mathcal{N}}^{'}(t^{'}, t) d\tilde{\nu}^{'}(t^{'}) d\tilde{\nu}_i(t)$$

$$\int_{(\psi_i \circ \tilde{f})(U_i)} \int_{\psi_i(V_i)} \|f(s) - t\| d\mu_i(s) d\nu_i(t) \simeq \int_{\psi_i(V_i)} \int_{\psi_i(V_i)} \|t^{'} - t\| d\nu^{'}(t^{'}) d\nu_i(t)$$

, which directly leads to the lemma since $\frac{\int_{V_i} \int_{V_i} d_{\mathcal{N}}^{'}(t^{'}, t) d\tilde{\nu}^{'}(t^{'}) d\tilde{\nu}_i(t)}{\int_{\psi_i(V_i)} \int_{\psi_i(V_i)} \|t^{'} - t\| d\nu^{'}(t^{'}) d\nu_i(t)} < \infty$ □

Theorem 5.1. *Under the assumptions above, consider a generator $G \in F_e$ and a hypothesis space $\mathcal{H}$ with VC-dimension bound by constant Λ. Assume for each $i \in [K]$, the restriction of G to a pair of charts $f_i \doteq G_{\downarrow(U_i, V_i)} \in \mathcal{H}$ with $(\psi_i \circ G \circ \varphi_i^{-1})$ satisfies Lipschitz condition with constant M_G, then G generalizes globally with (Kn, Km) samples only if the following inequality is satisfied with probability $1 - C(\epsilon, \Lambda)(nm\epsilon^2)^{\tau(\Lambda)} e^{-nm\alpha\epsilon^2}$,*

$$\epsilon + \frac{1}{nm} \max\{\sum_{i=1}^{n} \sum_{j=1}^{m} d_{\mathcal{N}}(G(s_k^i), t_k^j)\}_{k=1}^{K} <$$

$$\frac{1}{\sqrt{m}} \sqrt{tr(\Sigma_N)} + 2tr(\Sigma_N) - M_G(\frac{1}{\sqrt{n}} \sqrt{tr(\Sigma_M)} + 2tr(\Sigma_M)) \tag{40}$$

where $C(\epsilon, \Lambda)$ and $\tau(\Lambda)$ are positive functions independent from n, m and $\alpha \in [1, 2]$ a constant.

Proof. For K independent local tasks, with Lma. 5.1, the global generalization condition in Thm. 4.1 will thus be written as

$$\max\{\epsilon^i_{\text{classical}} - \epsilon^i_{\text{adv}} + \eta^i\}_{i=1}^K < \min\{D_{LK}(\nu_i, \hat{\nu}_i^n) - M_G D_{LK}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^K$$

, which serves as a sufficient condition (note it is not a necessary condition) in the worst case.

We would like to consider the situation when $\epsilon^i_{\text{adv}} = 0$ and since the classical generalization error is equivalent with the assumption that the observed samples on each pair of charts are identical, we can reformulate the inequality as

$$\epsilon_{\text{classical}} + \max\{\eta^i\}_{i=1}^K < \min\{D_{LK}(\nu_i, \hat{\nu}_i^n) - M_G D_{LK}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^K$$

Apply the result from [7], we could bound the left side by ϵ with probability $1 - C(\epsilon, \Lambda)(nm\epsilon^2)^{\tau(\Lambda)} e^{-nm\alpha\epsilon^2}$, that is

$$\epsilon_{\text{classical}} + \max\{\eta^i\}_{i=1}^K < \epsilon + \frac{1}{nm} \max\{\sum_{i=1}^n \sum_{j=1}^m d_{\mathcal{N}}(G(s_k^i), t_k^j)\}_{k=1}^K$$

The next step is to deal with the right side, with a honest calculation, we could deduce

$$\min\{D_{LK}(\nu_i, \hat{\nu}_i^n) - M_G D_{LK}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^K \tag{41}$$

$$= \min\{\mathbb{E}\|\nu_i - \nu_i^n\| - M_G \mathbb{E}\|\mu_i - \mu_i^n\|\}_{i=1}^K + 2\text{tr}(\Sigma_N) - 2\text{tr}(\Sigma_M) \tag{42}$$

In order to write the first minimization term in a closed form, we use the following theorem from the theory of information geometry of Amari [8]

Theorem 1. [8, Theorem 4.4] *The mean square error of a biased-corrected first-order efficient estimator is given asymptotically by the expansion (with N observed samples):*

$$\mathbb{E}[(\hat{u}^a - u^a)(\hat{u}^b - u^b)] = \frac{1}{N} g^{ab} + O(\frac{1}{N^2})$$

where g^{ab} denotes the Fisher metric on the manifold constructed from a parametrized family of probability.

We thus apply such an estimation to figure out $\mathbb{E}\|\nu_i - \nu_i^n\|$ and $\mathbb{E}\|\mu_i - \mu_i^n\|$. As is well known, the matrix of fisher metric for a gaussian $\mathcal{N}(x, \Sigma)$ is directly Σ, the covariance matrix itself.

By observing $\mathbb{E}\|\nu_i - \hat{\nu}_i^n\| = \sqrt{tr(\mathbb{E}[(\nu_i - \hat{\nu}_i^n)(\nu_i - \hat{\nu}_i^n)^T])}$ and $\mathbb{E}\|\mu_i - \hat{\mu}_i^m\| = \sqrt{tr(\mathbb{E}[(\mu_i - \hat{\mu}_i^n)(\mu_i - \hat{\mu}_i^n)^T])}$, we have (with $O(N^{-2})$ term omitted)

$$\min\{D_{LK}(\nu_i, \hat{\nu}_i^n) - M_G D_{LK}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^K = \frac{1}{\sqrt{m}}\sqrt{\text{tr}(\Sigma_N)} + 2\text{tr}(\Sigma_N) - M_G(\frac{1}{\sqrt{n}}\sqrt{\text{tr}(\Sigma_M)} + 2\text{tr}(\Sigma_M))$$

, which thus gives the condition for generalization above. □

References

[1] Chung, Kai Lai. A course in probability theory. Academic press, 2001.
[2] Rudin, Walter. Real and complex analysis. Tata McGraw-Hill Education, 1987.
[3] Nakahara, Mikio. Geometry, topology and physics. CRC Press, 2003.
[4] Nikolaev, I. G. "Smoothness of the metric of spaces with bilaterally bounded curvature in the sense of AD Aleksandrov." Sibirsk. Mat. Zh 24.2 (1983): 114-132.
[5] Nikolaev, I. G. "A metric characterization of Riemannian spaces." Siberian Advances in Mathematics 9.4 (1999): 1-58.
[6] Vapnik, Vladimir Naumovich, and Vlamimir Vapnik. Statistical learning theory. Vol. 1. New York: Wiley, 1998.
[7] Vayatis, Nicolas, and Robert Azencott. "Distribution-dependent vapnik-chervonenkis bounds." Euro-COLT. Vol. 99. 1999.
[8] Amari, Shun-ichi, and Hiroshi Nagaoka. Methods of information geometry. Vol. 191. American Mathematical Soc., 2007.