The Fourier Discrepancy Function

Gennaro Auricchio, Andrea Codegoni, Stefano Gualandi and Lorenzo Zambon

***Abstract*—In this paper, we propose the Fourier Discrepancy Function, a new discrepancy to compare discrete probability measures. We show that this discrepancy takes into account the geometry of the underlying space. We prove that the Fourier Discrepancy is convex, twice differentiable, and that its gradient has an explicit formula. We also provide a compelling statistical interpretation. Finally, we study the lower and upper tight bounds for the Fourier Discrepancy in terms of the Total Variation distance.**

***Index Terms*—Fourier metrics, discrepancy, weak convergence, maximum likelihood, tight bounds**

I. Introduction

COMPARING probability measures is a crucial task in several applied fields, such as computer vision [1], [2], [3], [4], [5], [6], [7], supervised learning [8], [9], [10], [11], [12] and generative models [13], [14], [15]. However, using different metrics for a given task can lead to different results [13]. For this reason, it is of crucial importance to have a wide range of mathematical tools and understand their features. For instance, in [16], a class of divergence measures based on the Shannon entropy has been introduced and studied. A relevant topic in information theory has then become to give a comparison between different discrepancy functions, especially in terms of tight bounds. The problem of finding tight bounds has been introduced in [17]; since then, many works have improved the constants of several known inequalities [18], [19], [20]. These bounds have also been proved to be useful for source coding [21], [22], [23].

In this paper, we introduce the Fourier Discrepancy Function, a distance between discrete probability measures inspired by the $1,2-$Periodic Fourier-Based Metric [24]. Metrics based on the Fourier Transform have been introduced in [25] and used in several fields, such as kinetic theory [26], [27], statistics [28], and, more recently, generative models [14]. The Fourier Discrepancy inherits the ability to capture the geometry of the underlying space, which is an appealing property in several applications [29], [30], [31]. Moreover, it is easy to compute using the Fast Fourier Transform [32].

In Section II, we recall the most commonly used discrepancy functions for discrete probability measures and the $1,2-$Periodic Fourier Based Metric [24]. Then, in Section III, we introduce the Fourier Discrepancy. We prove that the squared Fourier Discrepancy is twice differentiable. Unlike the Wasserstein distance [33], [34], its gradient has an explicit formula. Moreover, we prove that the Fourier Discrepancy is convex, and we provide an interesting statistical interpretation. Finally, in Section IV, we study the lower and upper tight bounds for the Fourier Discrepancy in terms of the Total Variation distance. We close our paper with an open conjecture on the value of the upper tight bound.

II. Discrepancy Functions for Probability Measures

A. Commonly used discrepancy functions

In this subsection, we recall the main distances and divergences used to compare discrete probability measures in computational applied mathematics. Here (X,d) denotes a discrete finite metric space, and $\mathcal{P}(X)$ denotes the set of all the probability measures over X. For a complete discussion, we refer to [35], [36].

- The Total Variation distance (TV) [35] is defined as
$$TV(\mu,\nu) := \frac{1}{2}\sum_{x\in X}|\mu_x-\nu_x|.$$
- The Kullback-Leibler divergence (KL) [37] is defined as
$$KL(\nu||\mu) := \sum_{x\in X}\log\Big(\frac{\nu_x}{\mu_x}\Big)\nu_x, \tag{1}$$
if $\nu_x = 0$ for every x such that $\mu_x = 0$, and otherwise $KL(\nu||\mu) := +\infty$. We follow the convention $0\cdot\log(0) = 0$.
- The Wasserstein distance (W_1) [36], [38] is defined as
$$W_1(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)}\Bigg\{\sum_{(x,y)\in X\times X}|x-y|\;\pi_{x,y}\Bigg\}, \tag{2}$$
where
$$\Pi(\mu,\nu) := \Bigg\{\pi \in\mathcal{P}(X\times X) \text{ s.t.} \\ \sum_{y\in X}\pi_{x,y}=\mu_x,\ \sum_{x\in X}\pi_{x,y}=\nu_y\Bigg\}.$$
Intuitively, $\pi_{x,y}$ denotes the amount of mass that is moved from the point x to the point y to reshape the configuration μ into the configuration ν. The cost of moving a unit of mass from x to y is given by $|x-y|$. The Wasserstein distance is then the minimum cost for performing the total reshape.

Although these three functions are commonly used to compare measures, their features, and thus their behaviours when applied to a given task, are different. Studying these features is crucial to choosing the right discrepancy to be used for the given task. For example, the Total Variation is robust against random noise when used as a loss function for classification tasks [39], the Kullback-Leibler divergence is closely related to likelihood maximisation [9], [10], while the Wasserstein distance performs well at capturing the geometry of the underlying space [31], [36], [38].

G. Auricchio, A. Codegoni, S. Gualandi, and L. Zambon are with the Department of Mathematics, University of Pavia, Pavia, PV, 27100 Italy e-mail: gennaro.auricchio@unipv.it, andrea.codegoni01@universitadipavia.it, stefano.gualandi@unipv.it, lorenzogianmar.zambon01@universitadipavia.it

B. The $1,2-$Periodic Fourier-based Metric

In this subsection, we review the main notions about the Fourier Transform of discrete measures (DFT) and about Fourier Based Metrics [26], [24], [25]. For a complete discussion on the DFT, we refer to [40].

In what follows, we fix $X = I_N$, where $I_N \subset [0,1]$ is defined as

$$I_N := \left\{0, \frac{1}{N}, \dots, \frac{N-1}{N}\right\},$$

for any given $N \in \mathbb{N}$. A discrete measure μ on I_N is then defined as

$$\mu := \sum_{j=0}^{N-1} \mu_j \delta_{\frac{j}{N}}, \tag{3}$$

where the values μ_j are non-negative real numbers such that $\sum_{j=0}^{N-1} \mu_j = 1$. Since any discrete measure supported on I_N is fully characterised by the $N-$uple of positive values $(\mu_0, \dots, \mu_{N-1})$, we refer to discrete measures and vectors interchangeably.

Definition 1. *The Discrete Fourier Transform (DFT) of μ is the $N-$dimensional vector $\hat{\mu} := (\hat{\mu}_0, \dots, \hat{\mu}_{N-1})$ defined as*

$$\hat{\mu}_k := \sum_{j=0}^{N-1} \mu_j e^{-2\pi i \frac{j}{N} k}, \qquad k \in \{0, \dots, N-1\}. \tag{4}$$

Remark 1. *Since the complex exponential function $k \to e^{-2\pi i \frac{j}{N} k}$ is a $N-$periodic function for any integer j, we set*

$$\hat{\mu}_k := \hat{\mu}_{mod_N(k)}$$

for any $k \in \mathbb{Z}$, where $\text{mod}_N(k)$ is the $N-$modulo operation. In particular, $\hat{\mu}_{-k} = \hat{\mu}_{N-k}$ for any $k \in \{0, \dots, N-1\}$.

Remark 2. *The DFT of a discrete measure can be expressed as a linear map:*

$$(\hat{\mu}_0, \dots, \hat{\mu}_{N-1}) = \Omega \cdot (\mu_0, \dots, \mu_{N-1}), \tag{5}$$

where Ω is the $N \times N$ matrix defined as

$$\Omega := \begin{bmatrix} \omega_{0,0} & \omega_{0,1} & \cdots & \omega_{0,N-1} \\ \omega_{1,0} & \omega_{1,1} & \cdots & \omega_{1,N-1} \\ \cdots & \cdots & \cdots & \cdots \\ \omega_{N-1,0} & \omega_{N-1,1} & \cdots & \omega_{N-1,N-1} \end{bmatrix}, \tag{6}$$

and $\omega_{k,j} := e^{-2\pi i \frac{j}{N} k}$. Since the matrix Ω is invertible, the DFT is a bijective function.

We now introduce the $1,2-$Periodic Fourier-based Metric [24].

Definition 2. *Let μ and ν be two discrete measures over I_N. The $1,2-$Periodic Fourier-based Metric is defined as*

$$f_{1,2}^2(\mu,\nu) := \int_{[0,1]} \frac{\left|\sum_{j=0}^{N-1} (\mu_j - \nu_j) e^{-2\pi i j k}\right|^2}{|k|^2} dk. \tag{7}$$

In [24], it is proved that the integral in (7) converges for any pair of probability measures μ and ν, and that $f_{1,2}$ is equivalent to W_1.

III. The Fourier Discrepancy Function

In this section, we introduce the Fourier Discrepancy function, inspired by (7).

We compare the Fourier Discrepancy Function with other discrepancies, and we show with an example its ability to take into account the geometry of the underlying space. Then, we prove that the Fourier Discrepancy Function is convex, and we provide the explicit formula for the gradient and the Hessian matrix of its corresponding loss function. Finally, we present a statistical model with Gaussian noise in the space of frequencies, in which the minimisation of the Fourier Discrepancy is equivalent to the maximisation of the likelihood.

Remark 3. *Herein, we only consider one-dimensional discrete measures, but all the results may be extended to a multi-dimensional setting.*

Moreover, for the sake of simplicity, we assume that N is an even number.

Since $\hat{\mu}_k = \overline{\hat{\mu}_{N-k}}$, we have

$$\left|\hat{\mu}_k - \hat{\nu}_k\right| = \left|\hat{\mu}_{N-k} - \hat{\nu}_{N-k}\right|, \tag{8}$$

which means that the k-th discrete frequencies give us the same information of the $(N-k)$-th ones. Therefore, we only consider the first $\frac{N}{2} - 1$ frequencies and we take half of the $\frac{N}{2}$-th frequency. We then propose the following discrete version of the metric in (7).

Definition 3. *We define the Fourier Discrepancy function $\mathbb{F} : \mathcal{P}(I_N) \times \mathcal{P}(I_N) \to [0, +\infty)$ as*

$$\begin{aligned} \mathbb{F}^2(\mu,\nu) &:= \sum_{k=1}^{\frac{N}{2}-1} \frac{|\hat{\mu}_k - \hat{\nu}_k|^2}{|k|^2} + \frac{\left|\frac{1}{2}\left(\hat{\mu}_{\frac{N}{2}} - \hat{\nu}_{\frac{N}{2}}\right)\right|^2}{|\frac{N}{2}|^2} \\ &= \sum_{k=1}^{\frac{N}{2}-1} \frac{|\hat{\mu}_k - \hat{\nu}_k|^2}{|k|^2} + \frac{|\hat{\mu}_{\frac{N}{2}} - \hat{\nu}_{\frac{N}{2}}|^2}{|N|^2}. \end{aligned} \tag{9}$$

Remark 4. *The function $\mathbb{F}$ is a distance on $\mathcal{P}(I_N)$. Moreover, the following holds:*

$$\frac{1}{N} C_1 \cdot W_1 \leq \mathbb{F} \leq C_2 \cdot W_1, \tag{10}$$

where C_1, C_2 are positive constants that do not depend from N. This follows from the equivalence between the Fourier-based metric and the Wasserstein distance [24].

Example 1. *Figure 1 shows the behaviours of different discrepancy functions when comparing Dirac's delta distributions. We have omitted the KL, since it is always equal to $+\infty$ whenever the supports of the two distributions are disjoint. We highlight how the Fourier discrepancy, similarly to the W_1, is able to take into account the geometry of the underlying space.*

To conclude, we provide an upper bound for the Fourier Discrepancy with respect to the Total Variation and the Kullback-Leibler.

Proposition 1. *For any pair of probability measures μ and ν, we have that*

$$\mathbb{F}(\mu,\nu) \leq \frac{2}{\sqrt{6}} \pi \, TV(\mu,\nu). \tag{11}$$

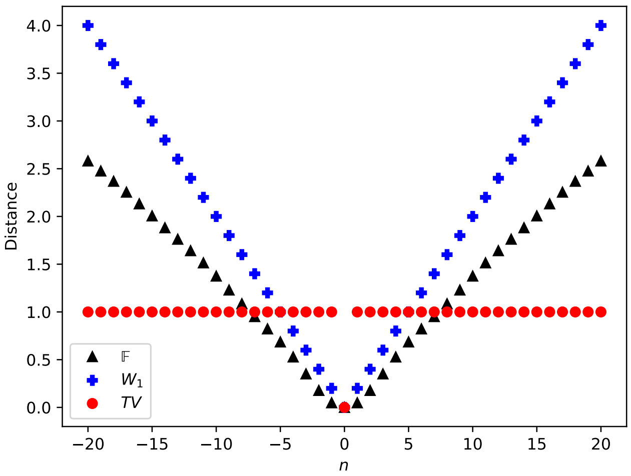

Fig. 1. Distance between δ_0 and δ_n. We have scaled the distances for visual convenience.

The proof is reported in Appendix A.

Remark 5. *Using the Pinsker's inequality [41], we also obtain the following bound:*

$$\mathbb{F}(\mu,\nu) \leq \frac{\pi}{\sqrt{3}}\sqrt{KL(\mu||\nu)}.$$

A. Analytical properties

In what follows, we study the analytical properties of the Fourier Discrepancy Function.

Let us introduce the matrix $\mathbb{K} := \mathrm{diag}(b)$, where the vector b is defined as

$$b := \frac{1}{2}\left(1, 1^{-2}, \dots, \left(\frac{N}{2}-1\right)^{-2}, \frac{2}{N^2}, \left(\frac{N}{2}-1\right)^{-2}, \dots, 1^{-2}\right) \tag{12}$$

We can express the Fourier Discrepancy function as a quadratic form:

$$\begin{aligned}\mathbb{F}^2(\mu,\nu) &= (\hat{\mu}-\hat{\nu})^T\mathbb{K}(\hat{\mu}-\hat{\nu})\\ &= (\mu-\nu)^T\Omega^T\mathbb{K}\Omega(\mu-\nu),\end{aligned} \tag{13}$$

where Ω is the DFT matrix defined in (6).

We now study the matrix $\mathbb{H} := \Omega^T\mathbb{K}\Omega$ to derive the analytical properties of the Fourier Discrepancy.

Proposition 2. *The matrix $\mathbb{H}$ is positive definite and its eigenvalues are given by*

$$\lambda_i = N \cdot b_i, \qquad i = 0, \dots, N-1,$$

where b is the vector in (12).

Since $\mathbb{H}$ is positive definite, there exists a matrix $\mathbb{L}$ such that $\mathbb{L}^T\mathbb{L} = \mathbb{H}$. Therefore, we can write

$$\mathbb{F}(\mu-\nu) = \|\mathbb{L}(\mu-\nu)\|.$$

Since $\mathbb{F}$ is given by the composition of a linear function with the norm operator, we have the following.

Proposition 3. *The Fourier Discrepancy is convex in $\mu-\nu$.*

In many applications, discrepancies are used to evaluate how different a given probability measure is from a target one. An established tool to perform this comparison is the loss function.

For any given $\nu \in \mathcal{P}(I_N)$, we define the Fourier Loss Function $L_\nu : \mathcal{P}(I_N) \to [0,\infty)$ as

$$L_\nu(\mu) := \mathbb{F}^2(\mu,\nu). \tag{14}$$

We are able to explicitly express the gradient and the Hessian matrix of this function.

Proposition 4. *For any probability measure ν, the function L_ν is twice differentiable. Moreover, its gradient and Hessian matrix are expressed through the explicit formulae:*

$$(\nabla L_\nu)_l(\mu) = \frac{\partial L_\nu}{\partial \mu_l}(\mu) = 2\sum_{j=0}^{N-1}(\mu_j-\nu_j)\cdot \mathrm{Re}\left(\hat{b}_{j-l}\right) \tag{15}$$

and

$$(HL_\nu)_{h,l}(\mu) = \frac{\partial^2 L_\nu}{\partial \mu_h \partial \mu_l}(\mu) = 2\,\mathrm{Re}\left(\hat{b}_{h-l}\right), \tag{16}$$

where $\hat{b}$ is the Fourier Transform of the vector b.

In particular, L_ν is a convex function for any $\nu \in \mathcal{P}(I_N)$.

B. Statistical interpretation

We now show how the minimisation of the Fourier Discrepancy is related to the maximum likelihood estimator in classification models with a random noise. This is a classic framework in machine learning, where we often assume the existence of an underlying probabilistic model that generates the data [9], [42]. This model is typically expressed as

$$y_i = f(x_i;\theta) + \epsilon_i, \qquad i = 1, \dots, m, \tag{17}$$

where $(x_1,y_1),\dots,(x_m,y_m)$ are the data, $\epsilon_1,\dots,\epsilon_m$ are i.i.d. random noises, f is a function that specifies the model structure, and θ is the parameter that has to be optimised.

Let us suppose that, for every $i = 1,\dots,m$, $\hat{\epsilon_i} \sim \mathcal{CN}(0,\Sigma)$, where $\mathcal{CN}(0,\Sigma)$ is the circularly-symmetric complex normal distribution with zero mean and covariance matrix Σ, defined as $\Sigma := \mathrm{diag}\left(2\sigma^2\beta\right)$, for some $\sigma > 0$, and where β is given by

$$\beta := \left(0, 1^2, \dots, \left|\frac{N}{2}-1\right|^2, \left|\frac{N^2}{2}\right|, \left|\frac{N}{2}-1\right|^2, \dots, 1^2\right).$$

For a complete discussion on complex normal distributions, we refer to [43].

The likelihood of the observations $(x,y) = (x_i,y_i)_{i=1,\dots,m}$ is then given by

$$\begin{aligned}\mathbb{P}(y|x,\theta) &= \prod_{i=1}^m \mathbb{P}(y_i = f_i + \epsilon_i)\\ &= \prod_{i=1}^m \mathbb{P}(\hat{y_i} = \hat{f_i} + \hat{\epsilon_i})\\ &= C\cdot\prod_{i=1}^m \exp\left[-\frac{1}{2}(\hat{y_i}-\hat{f_i})^H\Sigma^{-1}(\hat{y_i}-\hat{f_i})\right]\\ &= C\cdot\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^m \mathbb{F}^2(y_i,f_i)\right],\end{aligned} \tag{18}$$

provided that the mass of y_i is the same as f_i, for every $i = 1, \dots, m$. The vector $(\hat{y}_i - \hat{f}_i)^H$ denotes the conjugate transpose of $(\hat{y}_i - \hat{f}_i)$ and C is a positive constant that does not depend on the data.

By taking the logarithm in (18), we obtain the following result.

Theorem 1. *Let us consider a model of the form* (17), *where* $(x_1, y_1), \dots, (x_m, y_m)$ *are the data and* $\epsilon_1, \dots, \epsilon_m$ *are distributed as described above. Then, the value of* θ *maximising the likelihood of the data is the one minimising the Fourier Discrepancy*

$$\frac{1}{m} \sum_{i=1}^{m} L_{\delta_{y_i}} \left(f(x_i; \theta) \right).$$

Notice that the structure of the covariance matrix Σ measures how the error on the k^{th} frequency is weighted. As the variance grows, we are more willing to accept discrepancies between the real value of the frequency and the predicted one. In particular, for $k = 0$, we have a null variance Gaussian (i.e. a Dirac's delta). Therefore the model does not admit any error on the null frequency: μ and ν must have the same mass.

IV. Tight Bounds

In this section, we study the tight bounds for the Fourier Discrepancy in terms of the Total Variation distance.

A first result is given in Proposition 1. However, what we aim to find are the lower and upper tight bounds, respectively $C_L(\theta)$ and $C_U(\theta)$, defined, for any given $\theta \in (0, 1]$, as

$$C_L(\theta) := \inf_{\mu,\nu : TV(\mu,\nu)=\theta} \mathbb{F}(\mu, \nu), \tag{19}$$

$$C_U(\theta) := \sup_{\mu,\nu : TV(\mu,\nu)=\theta} \mathbb{F}(\mu, \nu). \tag{20}$$

Due to the linearity of the DFT, we have that

$$\mathbb{F}(\mu, \nu)^2 = \sum_{k=1}^{\frac{N}{2}} \frac{|\widehat{(\mu - \nu)}_k|^2}{|k|^2},$$

we then set $\Delta := \mu - \nu$ and express both TV and $\mathbb{F}$ as functions of Δ, rather than μ and ν.

We now introduce and study the space of null sum measures.

Definition 4. *We say that a real measure* Δ *is a null sum measure if*

$$\sum_{i=0}^{N-1} \Delta_i = 0.$$

We denote by Θ *the set of all the null sum measures.*

Given any pair of probability measures μ and ν, their difference is a null sum measure. As the following result shows, up to a multiplicative constant, the converse is also true.

Proposition 5. *Given any non-zero* $\Delta \in \Theta$ *and* $\theta \in (0, 1]$, *there exists* $C > 0$ *and a pair of probability measures* (μ, ν) *such that*

$$\mu - \nu = C \cdot \Delta \quad \text{and} \quad TV(\mu, \nu) = \theta.$$

Proof. Let $C := \frac{\theta}{TV(\Delta)}$ and $\widetilde{\Delta} := C \cdot \Delta$, which are well-defined since $TV(\Delta) \neq 0$ for any non-zero Δ.

Then, for the $1-$homogeneity of TV, we have that $TV(\widetilde{\Delta}) = \frac{\theta}{TV(\Delta)} \cdot TV(\Delta) = \theta$.

Let $\widetilde{\mu}$ and $\widetilde{\nu}$ be, respectively, the positive and negative part of $\widetilde{\Delta}$. Therefore, $\widetilde{\Delta} = \widetilde{\mu} - \widetilde{\nu}$ and $\widetilde{\mu}_i, \widetilde{\nu}_i \geq 0$ for any i.

We have that

$$2\theta = \sum_i |\widetilde{\Delta}_i| = \sum_i \widetilde{\mu}_i + \sum_i \widetilde{\nu}_i, \tag{21}$$

and moreover, since $\widetilde{\Delta}$ is a null sum measure:

$$0 = \sum_i \widetilde{\Delta}_i = \sum_i \widetilde{\mu}_i - \sum_i \widetilde{\nu}_i. \tag{22}$$

From (21) and (22) follows easily that $\sum_i \widetilde{\mu}_i = \sum_i \widetilde{\nu}_i = \theta$.

We now define

$$\mu := \widetilde{\mu} + (1 - \theta)\delta_0, \quad \nu := \widetilde{\nu} + (1 - \theta)\delta_0.$$

We have that μ is a probability measure since $\mu_i \geq 0$ for any i and $\sum_i \mu_i = \sum_i \widetilde{\mu}_i + (1 - \theta) = 1$. The same holds for ν.

Moreover, $\mu - \nu = \widetilde{\Delta}$, hence $TV(\mu, \nu) = TV(\widetilde{\Delta}) = \theta$. □

Remark 6. *Thanks to Proposition 5, and for the* 1*-homogeneity of* $\mathbb{F}$, *we have that*

$$\begin{aligned} C_L(\theta) &= \inf_{\substack{\Delta \in \Theta: \\ \Delta \neq 0}} \mathbb{F}\left(\frac{\theta}{TV(\Delta)} \Delta \right) \\ &= \theta \cdot \inf_{\substack{\Delta \in \Theta: \\ \Delta \neq 0}} \frac{\mathbb{F}(\Delta)}{TV(\Delta)}, \end{aligned} \tag{23}$$

and, analogously,

$$C_U(\theta) = \theta \cdot \sup_{\substack{\Delta \in \Theta: \\ \Delta \neq 0}} \frac{\mathbb{F}(\Delta)}{TV(\Delta)}. \tag{24}$$

A. Lower tight bound

Let us define the complex vector $\omega_k \in \mathbb{C}^N$ as

$$\omega_k = \left(e^{i\frac{2\pi k}{N}0}, e^{i\frac{2\pi k}{N}1}, \dots, e^{i\frac{2\pi k}{N}(N-1)} \right).$$

Since $\{\omega_k\}_{k=0,\dots,N-1}$ is an orthogonal basis of $\mathbb{C}^n$ [40], for any $\Delta \in \Theta$ there exists a unique N-tuple of complex coefficients $\left(\lambda^{(k)}\right)_k$ such that

$$\Delta = \sum_{k=0}^{N-1} \lambda^{(k)} \omega_k.$$

We define

$$\Xi := \left\{ \Delta \in \Theta : \sum_{k=0}^{N-1} |\lambda^{(k)}| = 1 \right\}. \tag{25}$$

From (23), and for the 1-homogeneity of both TV and $\mathbb{F}$, we have that:

$$\begin{aligned} C_L(\theta) &= \theta \cdot \inf_{\substack{\Delta\in\Theta:\\ \Delta\neq 0}} \frac{\mathbb{F}\left(\frac{\Delta}{\sum|\lambda^{(k)}|}\right)}{TV\left(\frac{\Delta}{\sum|\lambda^{(k)}|}\right)} \frac{\sum|\lambda^{(k)}|}{\sum|\lambda^{(k)}|}, \\ &= \theta \cdot \inf_{\Delta\in\Xi} \frac{\mathbb{F}(\Delta)}{TV(\Delta)}. \end{aligned} \tag{26}$$

Lemma 2. *We have that*

$$\sup_{\Delta\in\Xi} TV(\Delta) = \frac{N}{2}, \tag{27}$$

and the supremum is attained at $\Delta = \omega_{\frac{N}{2}}$.

Proof. Since $|(\omega_k)_j| = |e^{i\frac{2\pi jk}{N}}| = 1$ for all j and k, we have that $TV(\omega_k) = \frac{N}{2}$. Then, for any $\Delta\in\Theta$ such that $\sum|\lambda^{(k)}| = 1$, we have

$$\begin{aligned} TV(\Delta) &= TV\Big(\sum_{k=0}^{N-1}\lambda^{(k)}\omega_k\Big) = \frac{1}{2}\sum_{j=0}^{N-1}\Big|\sum_{k=0}^{N-1}\lambda^{(k)}(\omega_k)_j\Big| \\ &\leq \frac{1}{2}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1}\Big|\lambda^{(k)}(\omega_k)_j\Big| = \frac{1}{2}\sum_{k=0}^{N-1}|\lambda^{(k)}|\sum_{j=0}^{N-1}|(\omega_k)_j| \\ &= \frac{N}{2}\sum_{k=0}^{N-1}|\lambda^{(k)}| = \frac{N}{2}. \end{aligned}$$

Finally, notice that $\omega_{\frac{N}{2}}\in\Theta$ since $\left(\omega_{\frac{N}{2}}\right)_j = e^{i\pi j} = (-1)^j$, therefore $\omega_{\frac{N}{2}}$ is real and $\sum_{j=0}^{N-1}\left(\omega_{\frac{N}{2}}\right)_j = 0$. □

Lemma 3. *For any* $\Delta\in\Theta$*, the Fourier Discrepancy is given by*

$$\mathbb{F}^2(\Delta) = N^2\Big(\sum_{k=1}^{\frac{N}{2}-1}\frac{|\lambda^{(k)}|^2}{k^2} + \frac{|\lambda^{(\frac{N}{2})}|^2}{|N|^2}\Big). \tag{28}$$

Proof. For any $j = 0,\dots,N-1$, we have that the DFT of ω_j is given by

$$\widehat{(\omega_j)}_k = \sum_{l=0}^{N-1} e^{-i\frac{2\pi}{N}lk}(\omega_j)_l = \sum_{l=0}^{N-1} e^{-i\frac{2\pi}{N}l(k-j)} = N\delta_{k-j}.$$

Hence, for the linearity of the DFT:

$$\widehat{\Delta}_k = \sum_{j=0}^{N-1}\lambda^{(j)}\widehat{(\omega_j)}_k = N\sum_{j=0}^{N-1}\lambda^{(j)}\delta_{k-j} = N\lambda^{(k)}.$$

□

Lemma 4. *We have that*

$$\inf_{\Delta\in\Xi}\mathbb{F}(\Delta) = 1,$$

and the infimum is attained at $\Delta = \omega_{\frac{N}{2}}$.

Proof. Let $\Delta\in\Xi$. Then $\lambda^{(0)} = \sum_j \Delta_j = 0$. Moreover, since Δ is real, we have that $\widehat{\Delta}_k = \overline{\widehat{\Delta}_{N-k}}$ for any $k = 1,\dots,N-1$, hence $|\lambda^{(k)}| = |\lambda^{(N-k)}|$. If we define

$$\gamma_j := \begin{cases} 2|\lambda^{(j)}| & \text{for } j = 1,\dots,\frac{N}{2}-1, \\ |\lambda^{(\frac{N}{2})}| & \text{for } j = \frac{N}{2}, \end{cases}$$

from (28) we obtain

$$\mathbb{F}^2(\Delta) = \Big(\frac{N}{2}\Big)^2\sum_{k=1}^{\frac{N}{2}}\frac{\gamma_k^2}{k^2},$$

while the constraint (25) is written as

$$\sum_{j=1}^{\frac{N}{2}}\gamma_j = 1.$$

It is easy to see that the minimum is achieved when $\gamma_{\frac{N}{2}} = 1$ and $\gamma_j = 0$ for $j = 1,\dots,\frac{N}{2}-1$. Therefore $\Delta = \omega_{\frac{N}{2}}\in\Xi$, and $\mathbb{F}(\Delta) = 1$.

□

Combining (26) with Lemma 2 and Lemma 4, we infer that the lower tight bound is attained at $\Delta = [-1,1,-1,1,\dots,1]$. Thanks to Proposition 5, we can conclude with the following theorem.

Theorem 5. *The lower tight buond* $C_L(\theta)$ *is given by*

$$C_L(\theta) = \frac{2\theta}{N}, \tag{29}$$

and is attained at

$$\begin{aligned} \mu &= \frac{2\theta}{N}[1,0,1,0,\dots,0] + (1-\theta)\delta_0, \\ \nu &= \frac{2\theta}{N}[0,1,0,1,\dots,1] + (1-\theta)\delta_0. \end{aligned}$$

B. Upper tight bound

First, we introduce a suitable class of null sum measures.

Definition 5. *For any* $i,j\in\{0,\dots,N-1\}$*, we define the measure* $\eta_{i,j}$ *as*

$$\eta_{i,j} := \delta_i - \delta_j.$$

Theorem 6. *Let* Δ *be a null sum measure on* $\{0,\dots,N-1\}$*. Then, we can express* Δ *as* $\Delta = TV(\Delta)\cdot\Delta'$*, where* Δ' *is a convex combination of* $\{\eta_{i_k,j_k}\}_k$ *such that, for any pair* η_{i_k,j_k} *and* $\eta_{i_{k'},j_{k'}}$*, we have*

$$i_k \neq j_{k'} \tag{30}$$

for any $k\neq k'$.

Proof. Let Δ be a null sum measure. Without loss of generality, we can reorder the values of Δ as follows:

$$\Delta = (\alpha_1,\dots,\alpha_r,-\beta_1,\dots,-\beta_l,0,\dots,0),$$

where $r+l\leq N$, α_i, $\beta_j > 0$, $\alpha_i\leq\alpha_{i+1}$, $\beta_j\leq\beta_{j+1}$, for any i and j, and $\sum\alpha_i = \sum\beta_j$.

Without loss of generality, we assume that

$$\alpha_1\leq\beta_1.$$

Hence, we can write

$$\Delta = \alpha_1\eta_{0,r} + \Delta^{(1)},$$

where

$$\begin{aligned} \Delta^{(1)} &= (0,\alpha_2^{(1)},\dots,\alpha_r^{(1)},-\beta_1^{(1)},\dots,-\beta_l^{(1)},0,\dots,0) \\ &: = (0,\alpha_2,\dots,\alpha_r,-(\beta_1-\alpha_1),-\beta_2,\dots,-\beta_l,0,\dots,0). \end{aligned}$$

Next, we compare $\alpha_2^{(1)}$ and $\beta_1^{(1)}$ and repeat the process until every entry vanishes. At the end, we find

$$\begin{aligned}\Delta &= \lambda_1\eta_{0,r} + \cdots + \lambda_k\eta_{r-1,N-1} \\ &=: \sum_k \lambda_k \eta_{i_k,j_k}. \end{aligned} \tag{31}$$

Notice that each $\eta_{i,j}$ in (31) is such that $i < r$ and $j \geq r$ by construction, which implies condition (30).

Since by hypothesis, for any $l = 0, \dots, N-1$, all the l-th entries $(\eta_{i_k,j_k})_i$ have the same sign, we can write

$$|\Delta_l| = \Big|\sum_k \lambda_k(\eta_{i_k,j_k})_l\Big| = \sum_k \lambda_k|(\eta_{i_k,j_k})_l|.$$

Therefore:

$$\begin{aligned} TV(\Delta) &= \frac{1}{2}\sum_l |\Delta_l| = \frac{1}{2}\sum_l\sum_k \lambda_k|(\eta_{i_k,j_k})_l| \\ &= \frac{1}{2}\sum_k\sum_l \lambda_k|(\eta_{i_k,j_k})_l| \\ &= \frac{1}{2}\sum_k \lambda_k \sum_l |(\eta_{i_k,j_k})_l| = \sum_k \lambda_k, \end{aligned}$$

since $\sum_l |(\eta_{i,j})_l| = 2$ for any i, j. To conclude, it suffices to set

$$\Delta' := \frac{1}{TV(\Delta)}\Delta = \sum_k \widetilde{\lambda}_k \eta_{i_k,j_k},$$

where $\widetilde{\lambda}_k := \frac{\lambda_k}{\sum_l \lambda_l} > 0$, and $\sum_k \widetilde{\lambda}_k = 1$. □

Theorem 7. *There exist $i^\star, j^\star \in \{0, \dots, N-1\}$ such that, for any $\theta \in (0,1]$:*

$$\theta \cdot \eta_{i^\star,j^\star} = \underset{TV(\Delta)=\theta}{\operatorname{argmax}}\ \mathbb{F}(\Delta). \tag{32}$$

Proof. First, let us define

$$(i^\star, j^\star) := \underset{i,j\in\{0,\dots,N-1\}}{\operatorname{argmax}}\ \mathbb{F}(\eta_{i,j}), \tag{33}$$

which exists since the maximum is taken over a finite set. For any $\theta \in (0,1]$ and any null sum measure Δ with $TV(\Delta) = \theta$, thanks to Theorem 6, we can write $\Delta = \theta \cdot \sum_k \lambda_k \eta_{i_k,j_k}$.

From the 1-homogeneity and the convexity of $\mathbb{F}$, we obtain:

$$\begin{aligned} \mathbb{F}(\Delta) &= \mathbb{F}\left(\theta \cdot \sum_k \lambda_k \eta_{i_k,j_k}\right) = \theta \cdot \mathbb{F}\left(\sum_k \lambda_k \eta_{i_k,j_k}\right) \\ &\leq \theta \cdot \sum_k \lambda_k \mathbb{F}\left(\eta_{i_k,j_k}\right) \leq \theta \cdot \sum_k \lambda_k \mathbb{F}\left(\eta_{i^\star,j^\star}\right) \\ &= \theta \cdot \mathbb{F}\left(\eta_{i^\star,j^\star}\right) = \mathbb{F}\left(\theta \cdot \eta_{i^\star,j^\star}\right). \end{aligned}$$

□

As a straightforward consequence, we get the following result.

Corollary 1. *The upper tight bound $C_U(\theta)$ is given by*

$$C_U(\theta) = \theta \cdot \mathbb{F}(\eta_{i^\star,j^\star}). \tag{34}$$

Corollary 1 allows to search for the upper tight bound over a finite set of points. By explicit computation of the Fourier Discrepancy (see Appendix B), we have that

$$\mathbb{F}^2(\eta_{j,l}) = \sum_{k=1}^{\frac{N}{2}-1} \frac{2 - 2\cos\left(\frac{2\pi|j-l|}{N}k\right)}{k^2} + \frac{2-2\cos(\pi|j-l|)}{N^2}.$$

Notice that $\mathbb{F}(\eta_{j,l})$ depends on j and l only through $d := |j-l|$. Hence, we can further restrict to measures of the form $\eta_{0,d}$, with $d \in \{1, \dots, N-1\}$. By studying the derivatives with respect to d, it is possible to show that $d^* = \frac{N}{2}$ is a local minimum for the function $g : [0, N] \to \mathbb{R}$, defined as:

$$g(d) := \sum_{k=1}^{\frac{N}{2}-1} \frac{\cos\left(\frac{2\pi d}{N}k\right)}{k^2} + \frac{\cos(\pi d)}{N^2}. \tag{35}$$

We close our paper with the following open conjecture.

Conjecture 1. *$d^* = \frac{N}{2}$ is a global minimum for g.*

If our conjecture was true, we would have

$$C_U(\theta) = \theta \cdot \sqrt{\sum_{k=1}^{\frac{N}{2}-1} \frac{2-2(-1)^k}{k^2} + \frac{2-2(-1)^{\frac{N}{2}}}{N^2}}.$$

Acknowledgment

We are deeply indebted to Giuseppe Toscani for several stimulating discussions and valuable suggestions on the Fourier Based Metrics. The research was partially supported by the Italian Ministry of Education, University and Research (MIUR): Dipartimenti di Eccellenza Program (2018–2022) - Dept. of Mathematics "F. Casorati", University of Pavia. The PhD scholarship of Andrea Codegoni is funded by Sea Vision S.r.l..

References

[1] S. Angenent, S. Haker, A. Tannenbaum, and L. Zhu, "Optimal mass transport for registration and warping," *International Journal of computer vision*, vol. 60, no. 3, pp. 225–240, 2004.

[2] N. Bonneel and D. Coeurjolly, "SPOT: Sliced Partial Optimal Transport," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–13, 2019.

[3] G. Auricchio, M. Gualandi, S.and Veneroni, and F. Bassetti, "Computing Kantorovich-Wasserstein distances on d-dimensional histograms using $(d+1)$-partite graphs." in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 5798–5808.

[4] F. Bassetti, S. Gualandi, and M. Veneroni, "On the computation of Kantorovich-Wasserstein distances between two-dimensional histograms by uncapacitated minimum cost flows," *SIAM Journal on Optimization*, vol. 30, no. 3, pp. 2441–2469, 2020.

[5] G. Auricchio, F. Bassetti, S. Gualandi, and M. Veneroni, "Computing Wasserstein Barycenters via Linear Programming," in *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, 2019, pp. 355–363.

[6] M. Cuturi and A. Doucet, "Fast computation of Wasserstein barycenters," in *International Conference on Machine Learning*, 2014, pp. 685–693.

[7] N. Papadakis, "Optimal transport for image processing," Ph.D. dissertation, Université de Bordeaux; Habilitation thesis, 2015.

[8] K. Janocha and W. M. Czarnecki, "On Loss Functions for Deep Neural Networks in Classification," *Schedae Informaticae*, vol. 25, pp. 49–59, 2016.

[9] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*. MIT press, 2017.

[10] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[11] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.

[12] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio, “Learning with a wasserstein loss,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 2, 2015, pp. 2053–2061.

[13] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 214–223.

[14] A. F. Ansari, J. Scarlett, and H. Soh, “A characteristic function approach to deep implicit generative modeling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7478–7487.

[15] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, “Mmd gan: Towards deeper understanding of moment matching network,” *arXiv preprint arXiv:1705.08584*, 2017.

[16] J. Lin, “Divergence measures based on the shannon entropy,” *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[17] G. L. Gilardoni, “On the minimum f-divergence for given total variation,” *Comptes Rendus Mathematique*, vol. 343, no. 11-12, pp. 763–766, 2006.

[18] F. Topsoe, “Some inequalities for information divergence and related measures of discrimination,” *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1602–1609, 2000.

[19] A. Guntuboyina, S. Saha, and G. Schiebinger, “Sharp inequalities for f-divergences,” *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 104–121, 2013.

[20] G. L. Gilardoni, “On pinsker’s and vajda’s type inequalities for csiszár’s f-divergences,” *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5377–5386, 2010.

[21] I. Csiszár, “Two remarks to noiseless coding,” *Information and Control*, vol. 11, no. 3, pp. 317–322, 1967.

[22] ——, “Information-type measures of difference of probability distributions and indirect observation,” *studia scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[23] I. Sason, “Tight bounds for symmetric divergence measures and a refined bound for lossless source coding,” *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 701–707, 2014.

[24] G. Auricchio, A. Codegoni, S. Gualandi, G. Toscani, and M. Veneroni, “The equivalence of Fourier-based and Wasserstein metrics on imaging problems,” *Rendiconti Lincei - Matematica e Applicazioni*, vol. 31, pp. 627–649, 2020.

[25] G. Gabetta, G. Toscani, and B. Wennberg, “Metrics for probability distributions and the trend to equilibrium for solutions of the Boltzmann equation,” *Journal of statistical physics*, vol. 81, no. 5, pp. 901–934, 1995.

[26] J. Carrillo and G. Toscani, “Contractive probability metrics and asymptotic behavior of dissipative kinetic equations,” *Rivista Matematica Università di Parma*, vol. 7, no. 6, pp. 75–198, 2007.

[27] L. Baringhaus and R. Grübel, “On a class of characterization problems for random convex combinations,” *Annals of the Institute of Statistical Mathematics*, vol. 49, no. 3, pp. 555–567, 1997.

[28] C. R. Heathcote, “The integrated squared error estimation of parameters,” *Biometrika*, vol. 64, no. 2, pp. 255–264, 1977.

[29] L. Guibas, Y. Rubner, and C. Tomasi, “The Earth Mover’s Distance as a metric for image retrieval,” *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[30] M. A. Ruzon and C. Tomasi, “Edge, junction, and corner detection using color distributions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1281–1295, 2001.

[31] G. Peyré and M. Cuturi, “Computational optimal transport: With applications to data science,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[32] J. W. Cooley, P. A. Lewis, and P. D. Welch, “The fast fourier transform and its applications,” *IEEE Transactions on Education*, vol. 12, no. 1, pp. 27–34, 1969.

[33] G. Peyré, “Entropic approximation of wasserstein gradient flows,” *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2323–2351, 2015.

[34] B. Lévy and E. L. Schwindt, “Notions of optimal transport theory and how to implement them on a computer,” *Computers & Graphics*, vol. 72, pp. 135–148, 2018.

[35] E. Çınlar, *Probability and stochastics*. Springer Science & Business Media, 2011.

[36] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[37] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[38] F. Santambrogio, “Optimal transport for applied mathematicians,” *Birkäuser, NY*, vol. 55, no. 58-63, p. 94, 2015.

[39] A. Ghosh, H. Kumar, and P. Sastry, “Robust loss functions under label noise for deep neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[40] K. R. Rao and P. C. Yip, *The transform and data compression handbook*. CRC press, 2018.

[41] S. Kullback, “A lower bound for discrimination information in terms of variation (corresp.),” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 126–127, 1967.

[42] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[43] N. R. Goodman, “Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction),” *The Annals of mathematical statistics*, vol. 34, no. 1, pp. 152–177, 1963.

Appendix A
Proof of Proposition 1

Proof. Let us fix two probability measures μ and ν over I_N. By definition, we have

$$\begin{aligned}
\mathbb{F}(\mu,\nu)^2 &= \sum_{k=1}^{\frac{N}{2}-1} \frac{\left|\sum_{j=0}^{N-1}(\mu_j-\nu_j)(e^{-2\pi i \frac{j}{N}k})\right|^2}{|k|^2} \\
&\quad + \frac{\left|\sum_{j=0}^{N-1}(\mu_j-\nu_j)(e^{-i\pi j})\right|^2}{|N|^2} \\
&\leq \sum_{k=1}^{\frac{N}{2}} \frac{\left|\sum_{j=0}^{N-1}|\mu_j-\nu_j||e^{-2\pi i \frac{j}{N}k}|\right|^2}{|k|^2} \qquad (36)\\
&= \sum_{k=1}^{\frac{N}{2}} \frac{\left|\sum_{j=0}^{N-1}|\mu_j-\nu_j|\right|^2}{|k|^2} \\
&= 4TV(\mu,\nu)^2 \sum_{k=1}^{\frac{N}{2}} \frac{1}{k^2} \\
&\leq 4TV(\mu,\nu)^2 \sum_{k=1}^{+\infty} \frac{1}{k^2} \\
&= \frac{4\pi^2}{6} TV(\mu,\nu)^2,
\end{aligned}$$

where inequality (36) follows from the fact that $N^{-2} \leq \left(\frac{N}{2}\right)^{-2}$. The proof is concluded by taking the square root on both sides. □

Appendix B
Computing $\mathbb{F}(\eta_{j,l})$

Let us consider null sum measures of the form $\eta_{l,j}$. We recall that $\eta_{l,j} := \delta_l - \delta_j$. Since

$$\widehat{\eta_{l,j}} = \Omega \cdot \eta_{l,j},$$

we have

$$\widehat{\eta_{l,j}} = \Theta_l - \Theta_j, \qquad (37)$$

where Θ_k is the $k-$th column of the matrix Ω. By the definition of Ω we have

$$\Theta_l = \left(e^{i\frac{2\pi l}{N}0}, e^{i\frac{2\pi l}{N}1}, \dots, e^{i\frac{2\pi l}{N}(N-1)}\right),$$

therefore, the value $\mathbb{F}(\eta_{l,j})^2$ is then given by

$$\mathbb{F}(\eta_{l,j})^2 = \sum_{k=1}^{\frac{N}{2}-1} \frac{|(\Theta_l-\Theta_j)_k|^2}{k^2} + \frac{|(\Theta_l-\Theta_j)_{\frac{N}{2}}|^2}{|N|^2}. \qquad (38)$$

Let us now compute explicitly $|(\Theta_l-\Theta_j)_k|^2$ for a given k. We have

$$\begin{aligned}
(\Theta_l-\Theta_j)_k &= \cos\left(\frac{2\pi l}{N}k\right) - \cos\left(\frac{2\pi j}{N}k\right) \\
&\quad + i\sin\left(\frac{2\pi l}{N}k\right) - i\sin\left(\frac{2\pi j}{N}k\right),
\end{aligned}$$

therefore

$$\begin{aligned}
|(\Theta_l-\Theta_j)_k|^2 &= \left(\cos\left(\frac{2\pi l}{N}k\right) - \cos\left(\frac{2\pi j}{N}k\right)\right)^2 \\
&\quad + \left(\sin\left(\frac{2\pi l}{N}k\right) - \sin\left(\frac{2\pi j}{N}k\right)\right)^2 \\
&= 2 - 2\Bigg(\cos\left(\frac{2\pi l}{N}k\right)\cos\left(\frac{2\pi j}{N}k\right) \\
&\quad + \sin\left(\frac{2\pi l}{N}k\right)\sin\left(\frac{2\pi j}{N}k\right)\Bigg) \\
&= 2 - 2\cos\left(\frac{2\pi(j-l)}{N}k\right), \qquad (39)
\end{aligned}$$

where the equality in (39) comes from the following trigonometric identity:

$$\cos(\alpha-\beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta).$$

Therefore

$$\mathbb{F}^2(\eta_{j,l}) = \sum_{k=1}^{\frac{N}{2}-1} \frac{2-2\cos\left(\frac{2\pi|j-l|}{N}k\right)}{k^2} + \frac{2-2\cos(\pi|j-l|)}{N^2}. \qquad (40)$$