ESAIM: COCV 28 (2022) 10
https://doi.org/10.1051/cocv/2022002

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

FORMULATION AND PROPERTIES OF A DIVERGENCE USED TO COMPARE PROBABILITY MEASURES WITHOUT ABSOLUTE CONTINUITY*,**

Paul Dupuis[1,***] and Yixiang Mao[2]

Abstract. This paper develops a new divergence that generalizes relative entropy and can be used to compare probability measures without a requirement of absolute continuity. We establish properties of the divergence, and in particular derive and exploit a representation as an infimum convolution of optimal transport cost and relative entropy. Also included are examples of computation and approximation of the divergence, and the demonstration of properties that are useful when one quantifies model uncertainty.

Mathematics Subject Classification. 60A10,62B10,93E15,94A17.

Received December 4, 2020. Accepted January 6, 2022.

1. Introduction

To compare different probabilistic models for a given application, one needs a notion of "distance" between the distributions. The specification of this distance is a subtle issue. Probability models are typically large or infinite dimensional, and the usefulness of the distance will depend on its mathematical properties. Is it convenient for analysis and optimization? Does it scale well with system size?

For situations that require an analysis of (probabilistic) model form uncertainly, the quantity known as relative entropy (or Kullback-Leibler divergence) is the most widely used such distance. This is true because relative entropy has all the attractive properties asked for in the last paragraph, and many more. (Relative entropy is not a true metric since it is not symmetric in its arguments, but owing to its other attributes it is more widely used for these purposes than any legitimate metric.)

The definition of relative entropy is as follows. Suppose S is a Polish space with metric $d(\cdot,\cdot)$ and associated Borel σ-algebra $\mathcal{B}$. Let $\mathcal{P}(S)$ be the space of probability measures over $(S,\mathcal{B})$. If $\mu,\nu\in\mathcal{P}(S)$ and μ is absolutely continuous with respect to ν (denoted $\mu\ll\nu$), then

$$R(\mu\|\nu)\doteq\int_S\left(\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu$$

*Research supported in part by the National Science Foundation (NSF-DMS-1904992).
**Research supported in part by the Air Force Office of Scientific Research (FA-9550-18-1-0214).

Keywords and phrases: Relative entropy, optimal transport theory, convex duality, calculus of variation, information-theoretic divergence, risk-sensitive control.

[1] Division of Applied Mathematics, Brown University, Providence, RI 02912, USA.
[2] Quantitative Research, Susquehanna International Group, Bala Cynwyd, PA 19004, USA.

*** Corresponding author: paul_dupuis@brown.edu

(even though $\log \mathrm{d}\mu/\mathrm{d}\nu$ can take both positive and negative values, as we discuss in the beginning of section 2, the definition is never ambiguous). Otherwise, we set $R(\mu\|\nu) = \infty$.

While we cannot go into all the reasons why relative entropy is so useful, it is essential that we describe why it is convenient for the analysis of model form uncertainty. This is due to a dual pair of variational formulas which relate $R(\mu\|\nu)$, integrals with respect μ, and what are called **risk-sensitive** integrals with respect to ν. Let $C_b(S)$ denote the set of bounded and continuous functions on S. Then Proposition 1.4.2 and Lemma 1.4.3 of [9] give

$$R(\mu\,\|\nu) = \sup_{g\in C_b(S)} \left\{ \int_S g\mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\}, \tag{1.1}$$

and for any $g \in C_b(S)$,

$$\log \int_S e^g \mathrm{d}\nu = \sup_{\mu\in\mathcal{P}(S)} \left\{ \int_S g\mathrm{d}\mu - R(\mu\,\|\nu) \right\}. \tag{1.2}$$

It is immediate from either of these that for $\mu, \nu \in \mathcal{P}(S)$ and $g \in C_b(S)$,

$$\int_S g\mathrm{d}\mu \le R(\mu\,\|\nu) + \log \int_S e^g \mathrm{d}\nu$$

(in fact these expressions hold with $C_b(S)$ replaced by the bounded and measurable functions on S). If we interpret ν as the **nominal** or **design** model (chosen perhaps on the basis of data or for computational tractability) and μ as the **true** model (or at least a more accurate model), then according to the last display one obtains a bound on an integral with respect to the true model. (In fact by introducing a parameter one can obtain bounds that are in some sense optimal [11].) We typically interpret the integral $\int_S g\mathrm{d}\mu$ as a **performance measure**, and so we have a bound on the performance of the system under the true distribution in terms of the relative entropy distance $R(\mu\,\|\nu)$, plus a risk-sensitive performance measure under the design model. From this elementary but fundamental inequality, and by exploiting the helpful qualitative and quantitative properties of relative entropy, there has emerged a set of tools that can be used to answer many questions where probabilistic model form uncertainty is important, including [3, 7, 8, 10–13, 15, 16, 18, 19].

However, relative entropy has one important shortcoming: for the bound to be meaningful we must have $R(\mu\,\|\nu) < \infty$, which imposes the requirement of absolute continuity of the true model with respect to the design model. For various uses, such as model building and model simplification, this restriction can be significant. In the context of model building, it can happen that one attempts to fit distributions to data by comparing an empirical measure constructed using data with the elements of a parameterized family, such as a collection of Gaussian distributions. In this case the two distributions one would compare are singular, and so relative entropy cannot be used. A second example, and one that occurs frequently in the physical sciences, operations research and elsewhere, is that a detailed model (such as the population process of a chemical reaction network, which takes values in a lattice) is approximated by a simpler process that takes values in the continuum (for example a diffusion process). For exactly the same reason as in the previous example, these processes, as well as their corresponding stationary distributions, are not absolutely continuous.

Because relative entropy is not directly applicable to such problems, significant effort has been put into investigating alternatives ([4, 5] and references therein). A class that has attracted some attention (*e.g.*, in the machine learning community) are the *type-1 Wasserstein* or, more generally, *optimal transport distances* [14, 20, 25]. These distances, which are true metrics, have certain attractive properties but also some shortcomings. One is that the distances do not have an interpretation as the dual of a strictly convex function. To be a little

more concrete, it is the strict concavity of the mapping $g \to H[g;\mu,\nu]$ with

$$H[g;\mu,\nu] \doteq \int_S g\mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \tag{1.3}$$

in the variational representation for $R(\mu\,\|\nu)$ that leads to tight bounds when applied to problems of control or optimization of stochastic uncertain systems [10]. As an elementary example, given a fixed bound M on $R(\mu\,\|\nu)$, it follows from (1.2) that for any $c > 0$

$$\int_S g\mathrm{d}\mu \le \frac{1}{c}\left[M + c\log \int_S e^{cg}\mathrm{d}\nu\right],$$

and bounds that are tight for the collection $\{\mu : R(\mu\,\|\nu) \le M\}$ can be obtained by optimizing on $c > 0$. This is not possible for the analogous variational representation for Wasserstein type distances which involves

$$H[g;\mu,\nu] \doteq \int_S g\mathrm{d}\mu - \int_S g\mathrm{d}\nu. \tag{1.4}$$

Also, in some problems of learning, one encounters optimization problems such as $\inf_\theta M(\mu,\nu_\theta)$ where M is a "distance" and ν_θ is a parameterized family. For $M(\mu,\nu_\theta)$ corresponding to relative entropy one obtains a min/max problem of the form

$$\inf_{\theta\in\Theta} \sup_{g\in C_b(S)} H[g;\mu,\nu_\theta]$$

that is solved iteratively, with H as in (1.3). Although we would prefer to avoid the restriction $\mu \ll \nu$, the (strong) concavity/convexity properties of the mapping $(g,\nu) \to H[g;\mu,\nu]$ appear preferable to those of the analogous affine mapping that corresponds to (1.4).

A second limitation of Wasserstein distances is that, owing to the absence of a chain rule, they do not in general scale well with respect to system dimension. This is an issue in applications to problems from the physical sciences, where large time horizons and large dimensions are common.

Rather than give up entirely the attractive features of the dual pair $(R(\mu\,\|\nu), \log\int_S e^g\mathrm{d}\nu)$, an alternative is to be more restrictive regarding the class of costs or performance measures for which bounds are required. Indeed, the requirement of absolute continuity in relative entropy is entirely due to the very large class of functions, $C_b(S)$, appearing in (1.1). For a collection $\Gamma \subset C_b(S)$ one can consider in lieu of $R(\mu\,\|\nu)$ what we call the Γ-divergence, which is defined by

$$G_\Gamma(\mu\,\|\nu) \doteq \sup_{g\in\Gamma}\left\{\int_S g\mathrm{d}\mu - \log\int_S e^g\mathrm{d}\nu\right\}. \tag{1.5}$$

By imposing regularity conditions on Γ (*e.g.*, Lipschitz continuity, additional smoothness) one generates (under mild additional conditions on Γ) divergences which relax the absolute continuity condition. Thus one is trading restrictions on the class of performance measures or observables for which bounds are valid, for the enlargement of the class of distributions to which the bounds apply. These divergences are of course not as nice as relative entropy, but one can prove that they retain versions of its most important properties (in particular, the sense in which a version of the chain rule persists is discussed in Sect. 6.2). In addition, the dual function remains $\log\int_S e^g\mathrm{d}\nu$. As noted this is useful owing to its convexity properties, and it is also useful when considering problems of optimization or control since the corresponding risk-sensitive optimization and optimal control problems are well studied in the literature.

In our formulation of the Γ-divergence the underlying idea is that to extend the range of probability measures that can be compared, one must restrict the class of integrands that will be considered. However, this leads directly to an interesting connection with the Wasserstein distance mentioned previously, which is that for suitable collections Γ we will prove the inf-convolution expression

$$G_\Gamma(\mu\,\|\nu) = \inf_{\gamma\in\mathcal{P}(S)} \{W_\Gamma(\mu-\gamma) + R(\gamma\,\|\nu)\}\,,$$

where W_Γ is the Wasserstein metric whose dual (sup) formulation uses the set of functions Γ. Moreover one recovers relative entropy by taking the limit $b\to\infty$ in $G_{b\Gamma}(\mu\,\|\nu)$, which may be useful if one wants to allow relatively small violations of the absolute continuity restriction, while at the same time taking advantage of simple approximations for the Wasserstein distance in the high transportation cost limit. The sup formulation (1.5) can also be used as the basis for sampling based computation, by adapting the approach of [17].

The organization of the paper is as follows. In Section 2 we define the Γ-divergence, and prove the first main result of this paper, which is the inf-convolution formula described above (Thm. 2.4). In Section 3, we show several properties of the Γ-divergence, and establish a convex duality formula for the Γ-divergence. Section 4 investigates the Γ-divergence for a special choices of Γ, which are sets of bounded Lipschitz continuous functions. We establish a relation between Γ-divergence and optimal transport cost, and prove existence and uniqueness for optimizers of variational representations of Γ-divergence (Thm. 4.9), and also formulas for directional derivatives of the Γ-divergence (Thm. 4.16). Section 5 considers limits for the Γ-divergence, and in Section 6 there is a preliminary discussion on how one can apply the Γ-divergence to obtain uncertainty quantification bounds.

As last remarks we note that the paper [1] defines a "relaxation" of Wasserstein distance by putting in an entropy term of the mass-transfer matrix. The new divergence so defined is easier to compute than the original Wasserstein distance, but is not the same as the divergences we develop here. Also, [24] makes use of an inf-convolution formula analogous to the one presented above to extend type-1 Wasserstein distances to positive measures.

2. Definition of the Γ-divergence

Throughout this section, S is a Polish space with metric $d(\cdot,\cdot)$ and associated Borel σ-algebra $\mathcal{B}$. $C_b(S)$ denotes the space of all bounded continuous functions from S to $\mathbb{R}$. Let $\mathcal{P}(S)$ be the space of probability measures over $(S,\mathcal{B})$, $\mathcal{M}(S)$ be the space of finite signed (Borel) measures over $(S,\mathcal{B})$, and $\mathcal{M}_0(S)$ be the subspace of $\mathcal{M}(S)$ whose total mass is 0. $\overline{\mathbb{R}} \doteq \mathbb{R}\cup\{\infty\}$ is the extended real numbers. Throughout this section, we consider $C_b(S)$ equipped with weak topology induced by $\mathcal{M}(S)$. Thus for $f_n, f\in C_b(S)$, $f_n\to f$ if $\int_S f_n \mathrm{d}\mu \to \int_S f\,\mathrm{d}\mu$ for all $\mu\in\mathcal{M}(S)$.

We recall that

$$R(\mu\|\nu) \doteq \int_S \left(\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu$$

whenever μ is absolutely continuous with respect to ν. For $t\in\mathbb{R}$ define $t^- \doteq -(t\wedge 0)$. Since the function $s(\log s)^-$ is bounded for $s\in[0,\infty)$, whenever $\mu\ll\nu$,

$$\int_S \left(\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^- \mathrm{d}\mu = \int_S \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\left(\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^- \mathrm{d}\nu < \infty.$$

Thus $R(\mu\|\nu)$ is always well defined.

We recall the Donsker-Varadhan variational representation (1.1) for relative entropy. We will use equation (1.1) as an equivalent characterization of $R(\cdot\|\nu)$ on $\mathcal{P}(S)$, and consider an extension to $\mathcal{M}(S)$. With an abuse of notation, we will also call the extended function R. The following lemma states basic properties of the extension. Its proof appears in the Appendix.

Lemma 2.1. *Consider $R : \mathcal{M}(S) \times \mathcal{P}(S) \to (-\infty, \infty]$ defined by (1.1). Then*

(1) $R(\mu\|\nu) \geq 0$ and $R(\mu\|\nu) = 0$ if and only if $\mu = \nu$,
(2) $R(\cdot\|\cdot)$ is convex,
(3) $R(\mu\|\nu) = \infty$ if $\mu \in \mathcal{M}(S)\backslash\mathcal{P}(S)$.

Though relative entropy has very attractive regularity and optimization properties, as noted $R(\mu\|\nu)$ is finite if and only if $\mu \ll \nu$. As such, it cannot be used to give a meaningful notion of "distance" without this absolute continuity restriction. In order to define a meaningful divergence for a pair of probability measures that are not mutually absolute continuous, but at the same time not to lose the useful properties of the "dual" function $g \to \log \int_S e^g \mathrm{d}\nu$ appearing in (1.1), a natural approach is to restrict the set of test functions in the variational formula. We define a criterion for the classes of "admissible" test functions we want to use.

Definition 2.2. Let Γ be a subset of $C_b(S)$ endowed with the inherited weak topology. We call Γ **admissible** if the following hold.
1) Γ is convex and closed.
2) Γ is symmetric in that $g \in \Gamma$ implies $-g \in \Gamma$, and Γ contains all constant functions.
3) Γ is determining for $\mathcal{P}(S)$, *i.e.*, for any $\mu, \nu \in \mathcal{P}(S)$ with $\mu \neq \nu$, there exists $g \in \Gamma$ such that

$$\int_S g \mathrm{d}\mu \neq \int_S g \mathrm{d}\nu.$$

We next define a new divergence by restricting the class of test functions in the definition of relative entropy. Let Γ^c denote the complement of Γ.

Definition 2.3. Fix $\nu \in \mathcal{P}(S)$. For $\mu \in \mathcal{M}(S)$, we define the **Γ-divergence** associated with the admissible set Γ by

$$G_\Gamma(\mu\|\nu) \doteq \sup_{g\in\Gamma} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\}.$$

We also define the following related quantity. For $\eta \in \mathcal{M}(S)$ let

$$W_\Gamma(\eta) \doteq \sup_{g\in\Gamma} \left\{ \int_S g \mathrm{d}\eta \right\} = \sup_{g\in\mathcal{C}_b(S)} \left\{ \int_S g \mathrm{d}\eta - \infty 1_{\{g\in\Gamma^c\}} \right\}.$$

When Γ is clear based on context, we will drop the subscript from G_Γ and W_Γ. Using a similar argument as in Lemma 2.1, one can show that $G_\Gamma(\mu\|\nu) = \infty$ if $\mu(S) \neq 1$. The next theorem states an important property of the Γ-divergence, which is that it can be written as a convolution involving relative entropy and W_Γ.

Theorem 2.4. *Assume Γ is an admissible set. Then for $\mu \in \mathcal{M}(S)$, $\nu \in \mathcal{P}(S)$,*

$$G_\Gamma(\mu\|\nu) = \inf_{\gamma\in\mathcal{P}(S)} \{R(\gamma\|\nu) + W_\Gamma(\mu - \gamma)\}$$

Remark 2.5. It will be pointed out in Section 4 that if Γ is taken to be the Lipschitz functions with respect to a cost function $c(x, y)$ that satisfies some specified conditions, $W_\Gamma(\mu - \nu)$ will be the corresponding optimal transport cost from μ to ν. If Γ is also admissible then the theorem tells us that by restricting the set of test functions in the variational representation of relative entropy to Γ, we get a quantity which is an inf-convolution of relative entropy and a metric.

The rest of this section is focused on the proof of Theorem 2.4. In order to do this, we need a few definitions and also will find it convenient to consider a more general setting.

Definition 2.6. Points x and y in a topological space Y can be **separated** if there exists an open neighborhood U of x and an open neighborhood V of y such that U and V are disjoint ($U \cap V = \varnothing$). Y is a **Hausdorff** space if all distinct points in Y are pairwise separable.

Definition 2.7. A subset C of a topological vector space Y over the number field $\mathbb{R}$ is

1. **convex** if for any $x, y \in C$ and any $t \in [0,1]$, $tx + (1-t)y \in C$,
2. **balanced** if for all $x \in C$ and any $\lambda \in \mathbb{R}$ with $|\lambda| \leq 1$, $\lambda x \in C$,
3. **absorbant** if for all $y \in Y$, there exists $t > 0$ and $x \in C$ such that $y = tx$.

A topological vector space Y is called **locally convex** if the origin has a local topological basis of convex, balanced and absorbent sets.

Definition 2.8. For a topological vector space Y over the number field $\mathbb{R}$, its **topological dual space** Y^* is defined as the space of all continuous linear functionals $\varphi : Y \to \mathbb{R}$.

The **weak* topology** on Y^* is the topology induced by Y. In other words, it is the coarsest topology such that functional $y : Y^* \to \mathbb{R}$, $y(\varphi) = \varphi(y)$ is continuous in Y^*.

For $y \in Y$ and $\varphi \in Y^*$, we also write $\langle y, \varphi \rangle \doteq \varphi(y) = y(\varphi)$.

Now let Y be a Hausdorff locally convex space with Y^* being its topological dual space and endowed with the weak* topology.

Definition 2.9. For a function $f : Y \to \overline{\mathbb{R}}$, its **convex dual** $f^* : Y^* \to \overline{\mathbb{R}}$ is defined by

$$f^*(z) = \sup_{y \in Y} \{\langle y, z \rangle - f(y)\}.$$

Definition 2.10. Let $f_1, f_2 : Y \to \overline{\mathbb{R}}$ be two functions. We define the inf-convolution of f_1 and f_2 by

$$[f_1 \square f_2](y) \doteq \inf_{y_1 \in Y} \{f_1(y_1) + f_2(y - y_1)\}.$$

Definition 2.11. For a function $f : Y \to \overline{\mathbb{R}}$ the **lower semicontinuous hull** $\overline{f}$ is defined by

$$\overline{f}(x) \doteq \sup\{g(x) : g \leq f, g : Y \to \overline{\mathbb{R}} \textit{ is continuous}\}.$$

Definition 2.12. A convex function $f : Y \to \overline{\mathbb{R}}$ is **proper** if there exists $y \in Y$ such that $f(y) < \infty$. The **domain** of a convex, proper function f is defined by

$$\mathrm{dom}(f) \doteq \{y \in Y : f(y) < \infty\}.$$

Now let us introduce an important lemma.

Lemma 2.13. *([6], Thm. 2.3.10) Let $f_i : Y \to \overline{\mathbb{R}}$ be convex, proper and lower-semicontinuous functions fulfilling $\bigcap_{i=1}^m \mathrm{dom}(f_i) \neq \varnothing$. Then one has*

$$\left(\sum_{i=1}^m f_i\right)^* = \overline{f_1^* \square \cdots \square f_m^*}.$$

In our use we take $Y = C_b(S)$ equipped with topology induced by $\mathcal{M}(S)$, *i.e.*, the topological basis around $g \in Y$ is taken as sets of the form

$$\left\{f \in Y : \int_S f \mathrm{d}\mu_k \in \left(\int_S g \mathrm{d}\mu_k - \epsilon_k, \int_S g \mathrm{d}\mu_k + \epsilon_k\right), k = 1, 2, \ldots, m\right\},$$

where $m \in \mathbb{N}, \{\mu_k\}_{k=1,2,\ldots,m} \subset \mathcal{M}(S)$ and $\epsilon_k > 0, k = 1, 2, \ldots, m$ are arbitrary. It can be easily verified that under this topology, $C_b(S)$ is a Hausdorff locally convex space, with $C_b(S)^* = \mathcal{M}(S)$ ([22], Thm. 3.10). For $g \in C_b(S)$ and $\mu \in \mathcal{M}(S)$, we define the bilinear form

$$\langle g, \mu \rangle \doteq \int_S g \mathrm{d}\mu.$$

We are now ready to prove the main theorem.

Proof of Theorem 2.4. Define $H_1, H_2 : C_b(S) \to \overline{\mathbb{R}}$ by

$$H_1(g) \doteq \log \int_S e^g \mathrm{d}\nu \text{ and } H_2(g) \doteq \infty 1_{\Gamma^c}(g).$$

Then

$$\begin{aligned} G_\Gamma(\mu\|\nu) &= \sup_{g\in\Gamma} \left\{ \int_S g\mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\} \\ &= \sup_{g\in C_b(S)} \left\{ \int_S g\mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu - \infty 1_{\Gamma^c}(g) \right\} \\ &= (H_1 + H_2)^* (\mu). \end{aligned}$$

Notice that $\{0\} \in \mathrm{dom}(H_1) \cap \mathrm{dom}(H_2) \neq \varnothing$, and both H_1 and H_2 are proper and convex. For lower-semicontinuity, under the topology induced by $\mathcal{M}(S)$, H_1 is lower semicontinuous because of (1.2) and the fact that supremum of continuous functions are lower semicontinuous, and H_2 is lower semicontinuous since Γ is closed. Thus, by Lemma 2.13

$$G_\Gamma(\mu\|\nu) = (H_1 + H_2)^*(\mu) = [\overline{H_1^* \square H_2^*}](\mu).$$

By equation (1.1) and the definition of W_Γ, we know that

$$R(\mu\|\nu) = H_1^*(\mu) \text{ and } W_\Gamma(\eta) = H_2^*(\eta).$$

In the following display, the first equality is due to the definition of inf-convolution, and the second is since $R(\gamma\|\nu) < \infty$ only when $\gamma \in \mathcal{P}(S)$:

$$H_1^* \square H_2^*(\mu) = \inf_{\gamma\in\mathcal{M}(S)} \{R(\gamma\|\nu) + W_\Gamma(\mu - \gamma)\} = \inf_{\gamma\in\mathcal{P}(S)} \{R(\gamma\|\nu) + W_\Gamma(\mu - \gamma)\}.$$

Thus the last thing we need to prove is that $H_1^* \square H_2^*$ is lower semicontinuous. Note that relative entropy is lower semicontinuous in the first argument in the weak topology ([9], Lem. 1.4.3 (b)), and W_Γ is lower semicontinuous in the weak topology since it is the supremum of a collection of linear functionals. Let

$$F(\mu) \doteq H_1^* \square H_2^*(\mu) = \inf_{\gamma\in\mathcal{P}(S)} \{R(\gamma\|\nu) + W_\Gamma(\mu - \gamma)\}.$$

Consider any sequence $\mu_n \Rightarrow \mu$ with $\mu_n, \mu \in \mathcal{M}(S)$. Here "$\Rightarrow$" means convergence in the weak* topology, *i.e.*, for any $f \in C_b(S)$, $\int f\mathrm{d}\mu_n \to \int f\mathrm{d}\mu$. Let $\varepsilon > 0$, and for each μ_n let γ_n satisfy

$$R(\gamma_n\|\nu) + W_\Gamma(\mu_n - \gamma_n) \leq F(\mu_n) + \varepsilon.$$

We want to show that

$$\liminf_{n\to\infty} F(\mu_n) \geq F(\mu). \tag{2.1}$$

If $\liminf_{n\to\infty} F(\mu_n) = \infty$, the inequality above holds automatically. Assuming $\liminf_{n\to\infty} F(\mu_n) < \infty$, let n_k be a subsequence such that

$$\lim_{k\to\infty} F(\mu_{n_k}) = \liminf_{n\to\infty} F(\mu_n).$$

Notice that

$$R(\gamma_{n_k}\|\nu) \leq R(\gamma_{n_k}\|\nu) + W_\Gamma(\mu_{n_k} - \gamma_{n_k}) \leq F(\mu_{n_k}) + \varepsilon.$$

Since $\{F(\mu_{n_k})\}_{k\geq 1}$ is bounded, we know that $\{\gamma_{n_k}\}_{k\geq 1}$ is tight ([9], Lem. 1.4.3(c)). Then we can take a further subsequence that converges weakly. For simplicity of notation, let n_k denote this subsequence, and let γ_∞ denote the weak limit of γ_{n_k}. Then using the lower semicontinuity of $R(\cdot\|\nu)$ on $\mathcal{P}(S)$ and the lower semicontinuity of W_Γ on $\mathcal{M}(S)$,

$$\begin{aligned}
\liminf_{n\to\infty} F(\mu_n) + \varepsilon &= \lim_{k\to\infty} F(\mu_{n_k}) + \varepsilon \\
&\geq \lim_{k\to\infty} [R(\gamma_{n_k}\|\nu) + W(\mu_{n_k} - \gamma_{n_k})] \\
&\geq R(\gamma_\infty\|\nu) + W(\mu - \gamma_\infty) \\
&\geq \inf_{\gamma\in\mathcal{P}(S)} \{R(\gamma\|\nu) + W_\Gamma(\mu - \gamma)\} \\
&= F(\mu).
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary this establishes (2.1), and thus F is lower semicontinuous in $\mathcal{M}(S)$. The theorem is proved. □

3. Properties of the Γ-divergence

Theorem 2.4 provides an interesting characterization of the Γ-divergence. Before we continue to specific choices of Γ, we first state some general properties associated with Γ-divergence. Throughout this section we fix an admissible set Γ, and thus drop the subscript from G_Γ and W_Γ in this section. Also, now that we have established the expression for G as an inf-convolution as in Theorem 2.4, we no longer need to consider G as a function on $\mathcal{M}(S) \times \mathcal{P}(S)$, and instead can consider it just on $\mathcal{P}(S) \times \mathcal{P}(S)$, since we want to use G as a measure of how two probability distributions differ.

Lemma 3.1. *For $(\mu, \nu) \in \mathcal{P}(S) \times \mathcal{P}(S)$ define $G(\mu\|\nu)$ by Definition 2.3 and assume Γ is admissible. Then the following properties hold.*

1) $G(\mu\|\nu) \geq 0$, with $G(\mu\|\nu) = 0$ if and only if $\mu = \nu$.

2) $G(\mu\|\nu)$ is a convex and lower semicontinuous function of (μ, ν). In particular, $G(\mu\|\nu)$ is a convex, lower semicontinuous function of each variable μ or ν separately.

3) $G(\mu\|\nu) \leq R(\mu\|\nu)$ and $G(\mu\|\nu) \leq W(\mu - \nu)$.

Remark 3.2. 1) The first property justifies our calling G a divergence as the term is used in information theory.

2) Relative entropy has the property that for each fixed $\nu \in \mathcal{P}(S)$, $R(\cdot\|\nu)$ is strictly convex on $\{\mu \in \mathcal{P}(S) : R(\mu\|\nu) < \infty\}$. However, $G(\cdot\|\nu)$ in general is not strictly convex.

Proof of Lemma 3.1. 1) As noted in Lemma 2.1, $R(\cdot\|\cdot)$ is non-negative ([9], Lem. 1.4.1), and for any $\mu \in \mathcal{P}(S)$

$$W(\mu) = \sup_{g\in\Gamma}\left\{\int_S g\mathrm{d}\mu\right\} \geq \int_S 0\mathrm{d}\mu = 0.$$

Thus

$$G(\mu\|\nu) = \inf\{R(\mu_1\|\nu) + W(\mu_2) : \mu_1 + \mu_2 = \mu\} \geq 0.$$

Also by Lemma 2.1, $R(\mu_1\|\nu) = 0$ if and only if $\mu_1 = \nu$. Thus $G(\mu\|\nu) = 0$ if and only if

$$W(\mu - \nu) = \sup_{g\in\Gamma}\left\{\int_S gd(\mu - \nu)\right\} = 0,$$

which tells us $\mu = \nu$ since Γ is admissible.

2) This is a straightforward corollary of Theorem 2.4, since the supremum of a collection of linear and continuous functionals is both convex and lower semicontinuous.

3) This follows from Theorem 2.4 and that $R(\nu\|\nu) = W(0) = 0$. □

For relative entropy we have the following lemma ([9], Prop. 1.4.2).

Lemma 3.3. *For all $g \in C_b(S)$*

$$\log\int_S e^g\mathrm{d}\nu = \sup_{\mu\in\mathcal{P}(S)}\left\{\int_S g\mathrm{d}\mu - R(\mu\|\nu)\right\},$$

where the supremum is achieved uniquely at μ_0 satisfying

$$\frac{\mathrm{d}\mu_0}{\mathrm{d}\nu}(x) \doteq \frac{e^{g(x)}}{\int_S e^g\mathrm{d}\nu}.$$

A similar duality formula holds for the Γ-divergence when $g \in \Gamma$.

Theorem 3.4. *If Γ is admissible then for $g \in \Gamma$*

$$\log\int_S e^g\mathrm{d}\nu = \sup_{\mu\in\mathcal{P}(S)}\left\{\int_S g\mathrm{d}\mu - G(\mu\|\nu)\right\}.$$

Proof. Using the definition of the Γ-divergence

$$\begin{aligned}\sup_{\mu\in\mathcal{P}(S)}\left\{\int_S g\mathrm{d}\mu - G(\mu\|\nu)\right\} &= \sup_{\mu\in\mathcal{P}(S)}\left\{\int_S g\mathrm{d}\mu - \sup_{f\in\Gamma}\left\{\int_S f\mathrm{d}\mu - \log\int_S e^f\mathrm{d}\nu\right\}\right\}\\ &\leq \sup_{\mu\in\mathcal{P}(S)}\left\{\int_S g\mathrm{d}\mu - \left\{\int_S g\mathrm{d}\mu - \log\int_S e^g\mathrm{d}\nu\right\}\right\}\\ &= \log\int_S e^g\mathrm{d}\nu.\end{aligned}$$

On the other hand, we know for relative entropy that

$$\log \int_S e^g \mathrm{d}\nu = \sup_{\mu \ll \nu} \left\{ \int_S g \mathrm{d}\mu - R(\mu\|\nu) \right\}.$$

Since $G(\mu\|\nu) \leq R(\mu\|\nu)$,

$$\log \int_S e^g \mathrm{d}\nu = \sup_{\mu \ll \nu} \left\{ \int_S g \mathrm{d}\mu - R(\mu\|\nu) \right\} \leq \sup_{\mu \ll \nu} \left\{ \int_S g \mathrm{d}\mu - G(\mu\|\nu) \right\} \leq \sup_{\mu \in \mathcal{P}(S)} \left\{ \int_S g \mathrm{d}\mu - G(\mu\|\nu) \right\}.$$

The statement of the theorem follows from the two inequalities. □

The last theorem has two important implications. The first is related to the fact that Lemma 3.3 implies bounds for $\int_S g \mathrm{d}\mu$ when $R(\mu\|\nu)$ is bounded, an observation that has served as the basis for the analysis of various aspects of model form uncertainty [8, 11]. Using Theorem 3.4, we obtain analogous bounds on $\int_S g \mathrm{d}\mu$ for $g \in \Gamma$ when $G(\mu\|\nu)$ is bounded. Applications of these bounds will be further developed elsewhere. The second is that for $g \in \Gamma$, if we take μ_0 as defined in Lemma 3.3, then

$$\begin{aligned} \log \int_S e^g \mathrm{d}\nu &= \int_S g \mathrm{d}\mu_0 - R(\mu_0\|\nu) \\ &\leq \int_S g \mathrm{d}\mu_0 - G(\mu_0\|\nu) \\ &\leq \sup_{\mu \in \mathcal{P}(S)} \left\{ \int_S g \mathrm{d}\mu - G(\mu\|\nu) \right\} \\ &= \log \int_S e^g \mathrm{d}\nu, \end{aligned}$$

where the first inequality comes from $G(\mu_0\|\nu) \leq R(\mu_0\|\nu)$. Since both inequalities above must be equalities, we must have

$$R(\mu_0\|\nu) = G(\mu_0\|\nu).$$

The next lemma gives a more detailed picture of $G(\mu\|\nu)$ when $\mu \ll \nu$.

Lemma 3.5. *For $\mu, \nu \in \mathcal{P}(S)$, if $\mu \ll \nu$ then*

$$G(\mu\|\nu) = \sup_{\gamma \in \mathcal{A}(S)} \left\{ \int_S \log \left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu} \right) \mathrm{d}\mu \right\},$$

where

$$\mathcal{A}(S) \doteq \left\{ \gamma \in \mathcal{P}(S) : \gamma \ll \nu, \exists g \in \Gamma \text{ such that } \frac{\mathrm{d}\gamma}{\mathrm{d}\nu}(x) = e^{g(x)} \text{ for } x \in \mathrm{supp}(\nu) \right\}.$$

Proof. We use the definition

$$G(\mu\|\nu) = \sup_{g \in \Gamma} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\}$$

to prove this lemma. For any $g \in \Gamma$, we define $\gamma_g \in \mathcal{P}(S)$ by the relation

$$\frac{\mathrm{d}\gamma_g}{\mathrm{d}\nu}(x) = \frac{e^{g(x)}}{\int_S e^g \mathrm{d}\nu}$$

for $x \in \mathrm{supp}(\nu)$, and $\gamma_g(\mathrm{supp}(\nu)^c) = 0$. Then for $x \in \mathrm{supp}(\nu)$,

$$\log\left(\frac{\mathrm{d}\gamma_g}{\mathrm{d}\nu}(x)\right) = g(x) - \log\int_S e^g \mathrm{d}\nu.$$

Since $\mu \ll \nu$, we have

$$\int_S \log\left(\frac{\mathrm{d}\gamma_g}{\mathrm{d}\nu}\right)\mathrm{d}\mu = \int_S g\mathrm{d}\mu - \log\int_S e^g \mathrm{d}\nu,$$

and thus

$$G(\mu\|\nu) = \sup_{g\in\Gamma}\left\{\int_S g\mathrm{d}\mu - \log\int_S e^g\mathrm{d}\nu\right\} \le \sup_{\gamma\in\mathcal{A}(S)}\left\{\int_S \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right)\mathrm{d}\mu\right\}.$$

On the other hand, for any $\gamma \in \mathcal{A}(S)$, by definition, we can find a $g_\gamma \in \Gamma$ such that

$$g_\gamma(x) = \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}(x)\right)$$

for $x \in \mathrm{supp}(\nu)$. Then

$$\int_S g_\gamma \mathrm{d}\mu - \log\int_S e^{g_\gamma}\mathrm{d}\nu = \int_S \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right)\mathrm{d}\mu.$$

Thus

$$\sup_{\gamma\in\mathcal{A}(S)}\left\{\int_S \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right)\mathrm{d}\mu\right\} \le \sup_{g\in\Gamma}\left\{\int_S g\mathrm{d}\mu - \log\int_S e^g\mathrm{d}\nu\right\} = G(\mu\|\nu).$$

Combining the two inequalities completes the proof. □

Remark 3.6. When $\mu \in \mathcal{A}(S)$ we always have $G(\mu\|\nu) = R(\mu\|\nu)$. This is because if $\gamma \in \mathcal{A}(S)$ then $\mu \ll \gamma$, and therefore

$$\int_S \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu - \int_S \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right)\mathrm{d}\mu = \int_S \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}\right)\mathrm{d}\mu = R(\mu\|\gamma) \ge 0.$$

Rearranging gives

$$\int_S \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right)\mathrm{d}\mu = R(\mu\|\nu) - R(\mu\|\gamma),$$

and so

$$G(\mu\|\nu) = \sup_{\gamma\in\mathcal{A}(S)} \left\{ \int_S \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right) \mathrm{d}\mu \right\} = R(\mu\|\nu).$$

This statement is not valid when $\mu \ll \nu$ does not hold, since then $\log(\mathrm{d}\gamma/\mathrm{d}\nu)$ is not defined in $\mathrm{supp}(\mu)\backslash\mathrm{supp}(\nu)$, thus

$$\int_S \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right) \mathrm{d}\mu$$

is not well defined.

4. Connection with optimal transport theory

In the proceeding sections, we discussed general properties for the Γ-divergence with an admissible set $\Gamma \subset C_b(S)$. In this section, we discuss specific choices of Γ which relate the Γ-divergence with optimal transport theory. First we state some well known results in optimal transport theory.

4.1. Preliminary results from optimal transport theory

The results in this section are from Chapter 4 of [20]. The general Monge-Kantorovich mass transfer problem with given marginals $\mu, \nu \in \mathcal{P}(S)$ and cost function $c : S \times S \to \mathbb{R}_+$ is

$$\mathcal{C}(c;\mu,\nu) \doteq \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{S\times S} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) \right\},$$

where $\Pi(\mu,\nu)$ denotes the collection of all probability measures on $S \times S$ with first and second marginals being μ and ν, respectively.

A natural dual problem with respect to this is

$$\mathcal{B}(c;\rho) \doteq \sup_{f\in\mathrm{Lip}(c,S;C_b(S))} \left\{ \int_S f(x)\rho(\mathrm{d}x) \right\},$$

where $\rho = \mu - \nu$, $C_b(S)$ denotes the set of bounded continuous functions mapping S to $\mathbb{R}$ and

$$\mathrm{Lip}(c,S;C_b(S)) \doteq \{f \in C_b(S) : f(x) - f(y) \le c(x,y) \text{ for all } x,y \in S\}. \tag{4.1}$$

We want to know when

$$\mathcal{C}(c;\mu,\nu) = \mathcal{B}(c,\rho) \tag{4.2}$$

holds. The following is a necessary and sufficient condition. As with many results in this section, one can extend in a trivial way to the case where costs are bounded from below, rather than non-negative. Recall that S is a Polish space.

Condition 4.1. There is a nonempty subset $Q \subset C_b(S)$ such that the cost $c : S \times S \to [0,\infty]$ has the representation

$$c(x,y) = \sup_{u\in Q} (u(x) - u(y)) \quad \text{for all } (x,y) \in S \times S. \tag{4.3}$$

Theorem 4.2. *([20], Thm. 4.6.6) Under Condition 4.1, (4.2) holds.*

Remark 4.3. Condition 4.1 implies that c satisfies the triangle inequality, *i.e.*, for all $x, y, z \in S$

$$c(x,z) \le c(x,y) + c(y,z).$$

This follows easily from

$$\begin{aligned}\sup_{u\in Q}(u(x)-u(z)) &= \sup_{u\in Q}((u(x)-u(y))+(u(y)-u(z))) \\ &\le \sup_{u\in Q}(u(x)-u(y)) + \sup_{u\in Q}(u(y)-u(z)).\end{aligned}$$

On the other hand, Condition 4.1 also allows for a wide range of choices of $c(x,y)$. For example, suppose that c is a continuous metric on S, where continuity is with respect to the underlying metric of S. Then we can choose

$$Q = \{\min(c(x,x_0), n) : x_0 \in S, n \in \mathbb{N}\}.$$

It is easily verified that $Q \subset C_b(S)$, and that with this choice of Q (4.3) holds.

4.2. Γ-divergence with the choice $\Gamma = \mathrm{Lip}(c, S; C_b(S))$

Suppose $\Gamma = \mathrm{Lip}(c, S; C_b(S))$, with $c : S \times S \to [0,\infty]$ satisfying Condition 4.1. To make the presentation simple, we have assumed that c is non-negative, and further assume it is symmetric, meaning $c(x,y) = c(y,x) \ge 0$ for any $x, y \in S$. To distinguish from $W_\Gamma(\mu-\nu)$ for general Γ, we denote the transport cost for $\mu, \nu \in \mathcal{P}(S)$ by

$$W_c(\mu,\nu) \doteq \sup_{g\in \mathrm{Lip}(c,S;C_b(S))} \left\{ \int_S g d(\mu-\nu) \right\}.$$

Then by Theorem 4.2

$$W_c(\mu,\nu) = \sup_{g\in \mathrm{Lip}(c,S;C_b(S))} \left\{ \int_S g d(\mu-\nu) \right\} = \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{S\times S} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) \right\}.$$

Condition 4.4. Suppose $\mathrm{Lip}(c, S; C_b(S))$ is measure determining, *i.e.*, for all $\mu, \nu \in \mathcal{P}(S)$, $\mu \ne \nu$, there exists $f \in \mathrm{Lip}(c, S; C_b(S))$ such that

$$\int_S f \mathrm{d}\mu \ne \int_S f \mathrm{d}\nu.$$

Remark 4.5. It is well known that if $d(x,y)$ denotes the metric on S, then $\mathrm{Lip}(d, S; C_b(S))$ is determining. Hence a simple sufficient assumption for Condition 4.4 is that for some $\theta > 0$, $\theta d(x,y) \le c(x,y)$ for $x, y \in S$. In fact, it is enough that for each compact set $K \subset S$ there is θ with $\theta d(x,y) \le c(x,y)$ for $x, y \in S$. To see this, let $f \in \mathrm{Lip}(d, S; C_b(S))$ satisfy

$$\int_S f \mathrm{d}\mu \ge \int_S f \mathrm{d}\nu + \delta,$$

where $\delta > 0$ and we can assume that $0 \le f \le 1$. Since a single probability measure is always tight we can find a compact set K such that $\mu(K^c) \le \delta/8$ and $\nu(K^c) \le \delta/8$. Then under the assumption f is bounded and Lipschitz

continuous with respect to c on K, and using

$$f(z) = \min\{c(z,x) + f(x), x \in K\} \wedge 1$$

to redefine f off K, we obtain $f \in \mathrm{Lip}(c, S; C_b(S))$ such that

$$\int_S f \mathrm{d}\mu \geq \int_S f \mathrm{d}\nu + \delta/2.$$

Under Condition 4.4, Γ is admissible (see Def. 2.2), and by Theorem 2.4

$$G_\Gamma(\mu\|\nu) = \sup_{g\in\Gamma} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\} = \inf_{\gamma\in\mathcal{P}(S)} \left\{ W_c(\mu,\gamma) + R(\gamma\|\nu) \right\}. \tag{4.4}$$

Hence by choosing Γ properly, we get that the Γ-divergence is an infimal convolution of relative entropy, which is a convex function of likelihood ratios, and an optimal transport cost, which depends on a cost structure on the space S. Natural questions to raise here are the following.

i) Do there exist optimizers γ^* and g^* in the variational problem (4.4)? If so, are they unique?

ii) How can one characterize γ^* and g^*?

iii) For a fixed $\nu \in \mathcal{P}(S)$ (resp., $\mu \in \mathcal{P}(S)$), what is the effect of a perturbation of μ (resp., ν) on $G_\Gamma(\mu\|\nu)$?

We will address these questions sequentially in this section. From now on, we will drop the subscript Γ in this section for the simplicity of writing. We consider the case where $G(\mu\|\nu) < \infty$. To impose additional constraints on μ and ν such that $G(\mu\|\nu) < \infty$ holds, we make a further assumption on c.

Condition 4.6. There exists $a : S \to \mathbb{R}_+$ such that

$$c(x,y) \leq a(x) + a(y).$$

Now consider $\mu, \nu \in L^1(a) \doteq \{\theta \in \mathcal{P}(S) : \int_S a(x)\theta(\mathrm{d}x) < \infty\}$. Then

$$\begin{aligned}
G(\mu\|\nu) &= \inf_{\gamma\in\mathcal{P}(S)} \{W_c(\mu,\gamma) + R(\gamma\|\nu)\} \\
&\leq W_c(\mu,\nu) \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{S\times S} c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) \right\} \\
&\leq \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{S\times S} [a(x)+a(y)]\,\pi(\mathrm{d}x,\mathrm{d}y) \right\} \\
&= \int_S a(x)\mu(\mathrm{d}x) + \int_S a(y)\nu(\mathrm{d}y) \\
&< \infty.
\end{aligned}$$

We will assume the following mild conditions on the space S and cost c to make $\mathrm{Lip}(c, S; C_b(S))$ precompact.

Condition 4.7. There exists $\{K_m\}_{m\in\mathbb{N}}$ such that $K_m \subset S$ is compact, $K_m \subset K_{m+1}$ for all $m \in \mathbb{N}$, and $S = \cup_{m\in\mathbb{N}} K_m$. For each m, there exists $\theta_m : \mathbb{R}_+ \to \mathbb{R}_+$, such that $\lim_{a\to 0} \theta_m(a) = 0$, and $\delta_m > 0$, such that for any $x, y \in K_m$ satisfying $d(x,y) \leq \delta_m$,

$$c(x,y) \leq \theta_m(d(x,y)).$$

Recalling the definition (4.1), we define the unbounded version as follows

$$\mathrm{Lip}(c, S) \doteq \{f \in C(S) : f(x) - f(y) \leq c(x, y) \text{ for all } x, y \in S\},$$

where $C(S)$ is the set of continuous functions mapping S to $\mathbb{R}$. Before we proceed, we state the following lemma, which will be used repeatedly in this section.

Lemma 4.8. *If $g \in \mathrm{Lip}(c, S)$ and $\theta, \nu \in \mathcal{P}(S)$ satisfy $\int_S |g| \mathrm{d}\theta < \infty$, then*

$$\int_S g \mathrm{d}\theta - \log \int_S e^g \mathrm{d}\nu \leq G(\theta \| \nu) \leq R(\theta \| \nu).$$

Proof. We use a standard truncation argument. Since by Lemma 3.1 we already have $G(\theta\|\nu) \leq R(\theta\|\nu)$, we only need to prove the first inequality in the statement of the lemma. If $\int_S e^g \mathrm{d}\nu = \infty$, then

$$\int_S g \mathrm{d}\theta - \log \int_S e^g \mathrm{d}\nu = -\infty < 0 \leq G(\theta \| \nu).$$

Hence we only need consider the case $\int_S e^g \mathrm{d}\nu < \infty$. Let $g_n = \min(\max(g, -n), n) \in \mathrm{Lip}(c, S; C_b(S)) = \Gamma$ for $n \in \mathbb{N}$. We have $|g_n(x)| \leq |g(x)|$ and

$$\lim_{n\to\infty} g_n(x) = g(x) \quad x \in S.$$

Thus by the dominated convergence theorem $\lim_{n\to\infty} \int_S g_n \mathrm{d}\theta = \int_S g \mathrm{d}\theta$. Also we have

$$e^{g_n(x)} \leq e^{g(x)} + 1 \text{ and } \lim_{n\to\infty} e^{g_n(x)} = e^{g(x)}.$$

Then again using the dominated convergence theorem, $\lim_{n\to\infty} \int_S e^{g_n} \mathrm{d}\nu = \int_S e^g \mathrm{d}\nu$. Together with (1.1), this gives

$$\begin{aligned}
\int_S g \mathrm{d}\theta - \log \int_S e^g \mathrm{d}\nu &= \lim_{n\to\infty} \left(\int_S g_n \mathrm{d}\theta - \log \int_S e^{g_n} \mathrm{d}\nu \right) \\
&\leq \sup_{f \in \Gamma} \left\{ \int_S f \mathrm{d}\theta - \log \int_S e^f \mathrm{d}\nu \right\} \\
&= G(\theta \| \nu).
\end{aligned}$$

□

Now we are ready to state the first main theorem of this section.

Theorem 4.9. *Suppose Conditions 4.1, 4.4, 4.6 and 4.7 are satisfied. Fix $\mu, \nu \in L^1(a)$. Then the following conclusions hold.*
1) There exists a unique optimizer γ^* in the expression (4.4).
2) There exists an optimizer $g^* \in \mathrm{Lip}(c, S)$ in the expression (4.4), which is unique up to an additive constant in $\mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)$.
3) g^* and γ^* satisfy the following conditions:

i)

$$\frac{\mathrm{d}\gamma^*}{\mathrm{d}\nu}(x) = \frac{e^{g^*(x)}}{\int_S e^{g^*(y)}\mathrm{d}\nu}, \quad \nu - a.s.$$

ii)

$$W_c(\mu, \gamma^*) = \int_S g^* d(\mu - \gamma^*).$$

Remark 4.10. With many analogous expressions related to relative entropy, one can only conclude the uniqueness of γ^* and g^* (up to constant addition) almost everywhere according to either the measure μ or ν. However, because of the regularity condition $g^* \in \mathrm{Lip}(c, S; C(S))$ and Condition 4.7, the uniqueness of g^* (up to constant addition) on $\mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)$ will follow.

Proof. For $n \in \mathbb{N}$ consider $\gamma_n \in \mathcal{P}(S)$ that satisfies

$$R(\gamma_n \| \nu) + W_c(\mu, \gamma_n) \leq G(\mu \| \nu) + 1/n.$$

Then by Lemma 1.4.3(c) in [9] $\{\gamma_n\}_{n\geq 1}$ is precompact in the weak topology, and thus has a convergent subsequence $\{\gamma_{n_k}\}_{k\geq 1}$. Denote $\gamma^* \doteq \lim_{k\to\infty} \gamma_{n_k}$. Then by the lower semicontinuity of both $R(\cdot\|\nu)$ and $W_c(\mu, \cdot)$, we have

$$R(\gamma^* \| \nu) + W_c(\mu, \gamma^*) \leq \liminf_{k\to\infty} \left(R(\gamma_{n_k} \| \nu) + W_c(\mu, \gamma_{n_k})\right) \leq G(\mu \| \nu).$$

Since

$$G(\mu \| \nu) = \inf_{\gamma \in \mathcal{P}(S)} \{R(\gamma \| \nu) + W_c(\mu, \gamma)\} \leq R(\gamma^* \| \nu) + W_c(\mu, \gamma^*)$$

it follows that

$$G(\mu \| \nu) = R(\gamma^* \| \nu) + W_c(\mu, \gamma^*),$$

which shows that γ^* is an optimizer in expression (4.4). If there exist two optimizers $\gamma_1 \neq \gamma_2$, the strict convexity of $R(\cdot\|\nu)$ and convexity of $W_c(\mu, \cdot)$ imply that for $\gamma_3 = \frac{1}{2}(\gamma_1 + \gamma_2)$

$$\begin{aligned} R(\gamma_3 \| \nu) + W_c(\mu, \gamma_3) &< \frac{1}{2}\left((R(\gamma_1 \| \nu) + W_c(\mu, \gamma_1)) + (R(\gamma_2 \| \nu) + W_c(\mu, \gamma_2))\right) \\ &= G(\mu \| \nu) \leq R(\gamma_3 \| \nu) + W_c(\mu, \gamma_3), \end{aligned}$$

a contradiction. Thus the existence and uniqueness of an optimizer γ^* of (4.4) is proved, which establishes 1) in the statement of the theorem. Before proceeding, we establish the following lemma, whose proof appears in the Appendix.

Lemma 4.11. *If $g \in \mathrm{Lip}(c, S)$, then $\int_S g \mathrm{d}\gamma^* < \infty$.*

Now we consider the other variational representation of $G(\mu\|\nu)$, which is

$$G(\mu \| \nu) = \sup_{g \in \mathrm{Lip}(c, S; C_b(S))} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\}.$$

Take $g_n \in \mathrm{Lip}(c, S; C_b(S))$ such that

$$G(\mu\|\nu) - 1/n \leq \int_S g_n \mathrm{d}\mu - \log \int_S e^{g_n} \mathrm{d}\nu \leq G(\mu\|\nu).$$

Without loss of generality, we can assume $g_n(x_0) = 0$ for some fixed $x_0 \in K_0 \subset S$. Since for any $m \in \mathbb{N}$ $K_m \subset S$ is compact, we have that $\{g_n\}_{n\in\mathbb{N}}$ is bounded and equicontinuous on K_m by Condition 4.7. By the Arzelà-Ascoli theorem, there exists a subsequence of $\{g_n\}_{n\in\mathbb{N}}$ that converges uniformly in K_m. Using a diagonalization argument, by taking subsequences sequentially along $\{K_m\}_{m\in\mathbb{N}}$, where the next subsequence is a subsequence of the former one, and taking one element from each sequence, we conclude there exists a subsequence $\{g_{n_j}\}_{j\in\mathbb{N}}$, that converges uniformly in any K_m. Since $S = \cup_{m\in\mathbb{N}} K_m$, we conclude that $\{g_{n_j}\}_{j\in\mathbb{N}}$ converges pointwise in S. Denotes its limit by g^*. It can be easily verified that $g^* \in \mathrm{Lip}(c, S)$.

Since $g_{n_j}(x) \leq g_{n_j}(x_0) + c(x_0, x) \leq a(x_0) + a(x)$ and $\int_S (a(x_0) + a(x)) \mathrm{d}\mu < \infty$, by the dominated convergence theorem $\lim_{j\to\infty} \int_S g_{n_j} \mathrm{d}\mu = \int_S g^* \mathrm{d}\mu$. By Fatou's lemma, we have $\liminf_{j\to\infty} \int_S e^{g_{n_j}} \mathrm{d}\nu \geq \int e^{g^*} \mathrm{d}\nu$, and therefore

$$-\log \int e^{g^*} \mathrm{d}\nu \geq \limsup_{j\to\infty} -\int_S e^{g_{n_j}} \mathrm{d}\nu.$$

Putting these together, we have

$$\begin{aligned}
G(\mu\|\nu) &= \sup_{g\in\mathrm{Lip}(c,S;C_b(S))} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\} \\
&\leq \limsup_{j\to\infty} \left\{ \int_S g_{n_j} \mathrm{d}\mu - \log \int_S e^{g_{n_j}} \mathrm{d}\nu \right\} \\
&\leq \int_S g^* \mathrm{d}\mu - \log \int_S e^{g^*} \mathrm{d}\nu \\
&= \left(\int_S g^* \mathrm{d}\mu - \int_S g^* \mathrm{d}\gamma^* \right) + \left(\int_S g^* \mathrm{d}\gamma^* - \log \int_S e^{g^*} \mathrm{d}\nu \right).
\end{aligned}$$

We can add and subtract $\int_S g^* \mathrm{d}\gamma^*$ because we have proved in Lemma 4.11 that γ^* is integrable with respect to functions in $\mathrm{Lip}(c, S)$, and $g^* \in \mathrm{Lip}(c, S)$. By Lemma 4.8 we have

$$\int_S g^* \mathrm{d}\gamma^* - \log \int_S e^{g^*} \mathrm{d}\nu \leq R(\gamma^*\|\nu).$$

We also have

$$\int_S g^* \mathrm{d}\mu - \int_S g^* \mathrm{d}\gamma^* \leq W_c(\mu, \gamma^*),$$

which is due to

$$\begin{aligned}
W_c(\mu, \gamma^*) &= \sup_{g\in\mathrm{Lip}(c,S;C_b(S))} \int_S g d(\mu - \gamma^*) \\
&\geq \limsup_{n\to\infty} \int_S \max(\min(g^*, n), -n) d(\mu - \gamma^*) \\
&= \int_S g^* d(\mu - \gamma^*),
\end{aligned}$$

where the last equality is because of the dominated convergence theorem and integrability of $|g^*|$ with respect to μ and γ^* (Lem. 4.11). We can therefore continue the calculation above as

$$\left(\int_S g^* \mathrm{d}\mu - \int_S g^* \mathrm{d}\gamma^*\right) + \left(\int_S g^* \mathrm{d}\gamma^* - \log \int_S e^{g^*} \mathrm{d}\nu\right) \leq W_c(\mu, \gamma^*) + R(\gamma^* \| \nu) = G(\mu \| \nu).$$

Since both the upper and lower bounds on the inequalities coincide, we must have all inequalities to be equalities, and therefore

$$G(\mu \| \nu) = \int_S g^* \mathrm{d}\mu - \log \int_S e^{g^*} \mathrm{d}\nu, \qquad \int_S g^* \mathrm{d}\mu - \int_S g^* \mathrm{d}\gamma^* = W_c(\mu, \gamma^*),$$

and

$$\int_S g^* \mathrm{d}\gamma^* - \log \int_S e^{g^*} \mathrm{d}\nu = R(\gamma^* \| \nu).$$

The last equation gives us the relationship

$$\frac{\mathrm{d}\gamma^*}{\mathrm{d}\nu}(x) = \frac{e^{g^*(x)}}{\int_S e^{g^*} \mathrm{d}\nu} \quad \nu - a.s.$$

Thus we have shown the existence of optimizer $g^* \in \mathrm{Lip}(c, S)$ and its relationship with γ^*. Lastly, for any other optimizer $\bar{g} \in \mathrm{Lip}(c, S)$ the analogous argument shows

$$\frac{\mathrm{d}\gamma^*}{\mathrm{d}\nu}(x) = \frac{e^{\bar{g}(x)}}{\int_S e^{\bar{g}} \mathrm{d}\nu} \quad \nu - a.s.$$

Hence uniqueness of the optimizer g^* in $\mathrm{supp}(\nu)$ up to $\nu - a.s.$ is also proved.

To determine the uniqueness of the optimizer g^* in $\mathrm{supp}(\mu)$, we take an optimal transport plan between μ and γ^*, $\pi^* \in \Pi(\mu, \gamma^*)$ for $W_c(\mu, \gamma^*)$, which means

$$W_c(\mu, \gamma^*) = \inf_{\pi \in \Pi(\mu, \gamma^*)} \left\{ \int_{S \times S} c(x, y) \pi(\mathrm{d}x, \mathrm{d}y) \right\} = \int_{S \times S} c(x, y) \pi^*(\mathrm{d}x, \mathrm{d}y).$$

(Note that c satisfying Condition 4.1 is lower semicontinuous, and therefore ([2], Thm. 1.5) shows the existence of an optimal transport plan π^*.)

Since $g^*(x) - g^*(y) \leq c(x, y)$,

$$\begin{aligned} W_c(\mu, \gamma^*) &= \int_{S \times S} c(x, y) \pi^*(\mathrm{d}x, \mathrm{d}y) \\ &\geq \int_{S \times S} \left[g^*(x) - g^*(y) \right] \pi^*(\mathrm{d}x, \mathrm{d}y) \\ &= \int_S g^*(x) (\mu - \gamma^*)(\mathrm{d}x) \\ &= W_c(\mu, \gamma^*). \end{aligned}$$

Then the only inequality above must be equality, which implies that for $(x, y) \in \mathrm{supp}(\gamma^*)$, $g^*(x) - g^*(y) = c(x, y)$, $\pi^* - a.s.$ This is also true for any other optimizer $\bar{g} \in \mathrm{Lip}(c, S)$ for (4.4). Thus we are able to determine g^*

uniquely in $\mathrm{supp}(\mu)$ $\mu - a.s.$ with the help of π^* and data of g^* in $\mathrm{supp}(\nu)$. Lastly, since $g^* \in \mathrm{Lip}(c, S)$ and by Condition 4.7, we conclude the uniqueness of g^* in $\mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)$ by the continuity of g^*. □

Remark 4.12. When $\mu \ll \nu$ Theorem 4.9 implies that for some constant c_0

$$g^*(x) = \log\left(\frac{\mathrm{d}\gamma^*}{\mathrm{d}\nu}(x)\right) - c_0 \quad \nu - a.s.$$

Hence

$$G(\mu\|\nu) = \int_S g^* \mathrm{d}\mu - \log \int_S e^{g^*}\mathrm{d}\nu = \int_S \log\left(\frac{\mathrm{d}\gamma^*}{\mathrm{d}\nu}(x)\right)\mathrm{d}\mu,$$

and so the Γ-divergence of μ with respect to ν looks like a "modified" version of relative entropy.

The next theorem tells us that 3) of Theorem 4.9 is not only a description of the pair of optimizer (g^*, γ^*), but also a characterization of it.

Theorem 4.13. *Suppose Conditions 4.1, 4.4, 4.6 and 4.7 are satisfied. Fix $\mu, \nu \in L^1(a)$. If $g_1 \in \mathrm{Lip}(c, S)$ and $\gamma_1 \in \mathcal{P}(S)$ satisfy condition 3) in Theorem 4.9, then (g_1, γ_1) are optimizers in the corresponding variational problem (4.4):*

$$G_\Gamma(\mu\|\nu) = \int_S g_1 \mathrm{d}\mu - \log \int_S e^{g_1}\mathrm{d}\nu = W_c(\mu, \gamma_1) + R(\gamma_1\|\nu).$$

Proof. The theorem follows from the two variational characterization of Γ-divergence in (4.4). Condition 3) of Theorem 4.9 implies

$$R(\gamma_1\|\nu) = \int_S g_1 \mathrm{d}\gamma_1 - \log \int_S e^{g_1}\mathrm{d}\nu \text{ and } W_c(\mu, \gamma_1) = \int_S g_1 d(\mu - \gamma_1),$$

and therefore

$$R(\gamma_1\|\nu) + W_c(\mu, \gamma_1) = \int_S g_1 \mathrm{d}\mu - \log \int_S e^{g_1}\mathrm{d}\nu.$$

This implies

$$\begin{aligned} G(\mu\|\nu) &= \inf_{\gamma\in\mathcal{P}(S)} \{R(\gamma\|\nu) + W_c(\mu, \gamma)\} \\ &\le R(\gamma_1\|\nu) + W_c(\mu, \gamma_1) \\ &= \int_S g_1 \mathrm{d}\mu - \log \int_S e^{g_1}\mathrm{d}\nu \\ &\le \sup_{g\in\Gamma}\left\{\int_S g\mathrm{d}\mu - \log\int_S e^g \mathrm{d}\nu\right\} \\ &= G(\mu\|\nu). \end{aligned}$$

The first inequality comes from the fact that $\gamma_1 \in \mathcal{P}(S)$, while the second needs a little more discussion, which will be given below. Assuming this, the last display shows that (g_1, γ_1) are optimizers. The second inequality

follows from Lemma 4.8 and the fact that

$$\int_S |g_1(x)|\mu(\mathrm{d}x) \le \int_S [|g_1(0)| + c(0,x)]\,\mu(\mathrm{d}x) \le \int_S [|g_1(0)| + a(0) + a(x)]\,\mu(\mathrm{d}x) < \infty.$$

The proof is complete. □

The last theorem answers questions i) and ii) raised earlier in this section, now we want to answer iii), which is to characterize the directional derivatives of $G(\mu\|\nu)$ in the one variable when fixing the other, *e.g.*,

$$\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\left(G(\mu+\varepsilon\rho\|\nu) - G(\mu\|\nu)\right)$$

for $\rho \in \mathcal{M}_0(S)$ which satisfies certain conditions. From Theorem 4.9 and remarks following it we know that any optimizer g^* of expression (4.4) is unique in $\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)$. However, there is still freedom to choose g^* in $S\backslash\{\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)\}$, since the variational problem in (4.4) does not take into account of the information of g^* outside $\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)$, other than requiring that g^* belong to $\mathrm{Lip}(c,S)$. We will define a special g^* that is uniquely defined not only in $\mathrm{supp}(\mu)$ and $\mathrm{supp}(\nu)$, but also on $S\backslash\{\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)\}$. For $x \in S\backslash\{\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)\}$, set

$$g_+^*(x) \doteq \inf_{y\in\mathrm{supp}(\nu)} \{g^*(y) + c(x,y)\}, \tag{4.5}$$

also known as the "c-transform" in the optimal transport literature. The following lemma confirms that this construction of g_+^* still lies in $\mathrm{Lip}(c,S)$. While part 1 is standard, we could not find a reference for part 2, and so the proof appears in the Appendix.

Lemma 4.14. *The following two statements hold.*

1) For $x \in \mathrm{supp}(\mu)$, the expression (4.5) also holds. In other words, for $x \in S\backslash\mathrm{supp}(\nu)$, we have

$$g^*(x) = \inf_{y\in\mathrm{supp}(\nu)} \{g^*(y) + c(x,y)\}.$$

2) g_+^* defined by equation (4.5) is in $\mathrm{Lip}(c,S)$. In addition,

$$g_+^*(x) = \sup\{h(x) : h \in \mathrm{Lip}(c,S), h(y) = g^*(y) \text{ for } y \in \mathrm{supp}(\nu)\} \tag{4.6}$$

Remark 4.15. We also will make use of the function

$$g_-^*(x) = \inf\{h(x) : h \in \mathrm{Lip}(c,S), h(y) = g^*(y) \text{ for } y \in \mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)\}. \tag{4.7}$$

Then based on these constructions, we have the following result. A sufficient condition for the requirement that $\int_S g_-^*\mathrm{d}\rho$ be well defined and finite and the related assumption regarding convergence is $\int e^{c(x,x_0)}\rho(\mathrm{d}x) < \infty$.

Theorem 4.16. *Take $\Gamma = \mathrm{Lip}(c,S;C_b(S))$ where c satisfies the conditions of Theorem 4.9 and $\mu,\nu \in L^1(a)$. Take $\rho = \rho_+ - \rho_- \in \mathcal{M}_0(S)$ where $\rho_+,\rho_- \in \mathcal{P}(S)$ are mutually singular probability measures, $\rho_+ \in L^1(a)$, and assume there exists $\varepsilon_0 > 0$ such that $\mu + \varepsilon\rho \in \mathcal{P}(S)$ for $0 < \varepsilon \le \varepsilon_0$. Then*

$$\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\left(G(\mu+\varepsilon\rho\|\nu) - G(\mu\|\nu)\right) = \int_S g_+^*\mathrm{d}\rho.$$

where g_+^* is given by (4.6). Suppose that $\int_S e^{g_-^*}\mathrm{d}\rho$ is well defined and finite, where g_-^* is given by (4.7), that if $g_n \in \mathrm{Lip}(c,S)$ converges to g_-^* pointwise then $\int_S e^{g_n}\mathrm{d}\rho \to \int_S e^{g_-^*}\mathrm{d}\rho$, and that there is $\varepsilon_0 > 0$ such that $\nu + \varepsilon\rho \in \mathcal{P}(S)$ for $0 < \varepsilon \le \varepsilon_0$. Then

$$\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\left(G(\mu\|\nu+\varepsilon\rho) - G(\mu\|\nu)\right) = -\int_S g_-^*\mathrm{d}\rho \Big/ \int_S g_-^*\mathrm{d}\nu.$$

Proof. We use the variational formula (4.4) for $G(\mu+\varepsilon\rho\|\nu)$, where $\mu+\varepsilon\rho \in \mathcal{P}(S)$ and $\rho_+ \in L^1(a)$. Recall that g_+^* is an optimizer for (4.4). Using Lemma 4.8 with $\theta = \mu + \varepsilon\rho$,

$$\begin{aligned} G(\mu+\varepsilon\rho\|\nu) &= \sup_{g\in\Gamma}\left\{\int_S g d(\mu+\varepsilon\rho) - \log\int_S e^g \mathrm{d}\nu\right\} \\ &\ge \int_S g_+^* d(\mu+\varepsilon\rho) - \log\int_S e^{g_+^*}\mathrm{d}\nu \\ &= \varepsilon\int_S g_+^*\mathrm{d}\rho + \int_S g_+^*\mathrm{d}\mu - \log\int_S e^{g_+^*}\mathrm{d}\nu \\ &= \varepsilon\int_S g_+^*\mathrm{d}\rho + G(\mu\|\nu). \end{aligned}$$

Thus

$$\liminf_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\left(G(\mu+\varepsilon\rho\|\nu) - G(\mu\|\nu)\right) \ge \int_S g_+^*\mathrm{d}\rho. \tag{4.8}$$

The other direction is more delicate. Take $f(\varepsilon) = G(\mu+\varepsilon\rho\|\nu)$. From Lemma 3.1 we know that f is convex, lower semicontinuous and finite on $[0,\varepsilon_0]$. Using a property of convex functions in one dimension, we know f is differentiable on $(0,\varepsilon_0)$ except for a countable number of points. Take $\varepsilon \in (0,\varepsilon_0)$ to be a place where f is differentiable, and $\delta > 0$ small. Take $g_\varepsilon^* \in \mathrm{Lip}(c,S)$ to be the optimizer for $G(\mu+\varepsilon\rho\|\nu)$ satisfying $g_\varepsilon^*(x_0) = 0$ for some x_0 in the support of ν, so that

$$G(\mu+\varepsilon\rho\|\nu) = \int_S g_\varepsilon^* d(\mu+\varepsilon\rho) - \log\int_S e^{g_\varepsilon^*}\mathrm{d}\nu.$$

Then using an argument that already appeared in this proof, we have

$$G(\mu+(\varepsilon\pm\delta)\rho\|\nu) - G(\mu+\varepsilon\rho\|\nu) \ge \pm\delta\int_S g_\varepsilon^*\mathrm{d}\rho.$$

It follows that

$$\begin{aligned} \int_S g_\varepsilon^*\mathrm{d}\rho &\le \lim_{\delta\to 0}\frac{1}{\delta}\left(G(\mu+(\varepsilon+\delta)\rho\|\nu) - G(\mu+\varepsilon\rho\|\nu)\right) \\ &= f'(\varepsilon) \\ &= \lim_{\delta\to 0}\frac{1}{\delta}\left(G(\mu+\varepsilon\rho\|\nu) - G(\mu+(\varepsilon-\delta)\rho\|\nu)\right) \\ &\le \int_S g_\varepsilon^*\mathrm{d}\rho. \end{aligned}$$

and therefore

$$f'(\varepsilon) = \int_S g_\varepsilon^* \mathrm{d}\rho. \tag{4.9}$$

If we denote

$$f'_+(0) = \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}(f(\varepsilon) - f(0)),$$

then by a property of convex functions ([21], Thm. 24.1), for any sequence of $\{\varepsilon_n\}_{n\in\mathbb{N}}$ such that $\varepsilon_0 > \varepsilon_n \downarrow 0$ and f is differentiable at $\varepsilon_n > 0$, we have

$$f'_+(0) = \lim_{n\to\infty} f'(\varepsilon_n) = \lim_{n\to\infty} \int_S g^*_{\varepsilon_n} \mathrm{d}\rho.$$

By the same argument used in the proof of Theorem 4.9 (paragraphs following Lem. 4.11), *i.e.*, by applying the Arzelà-Ascoli theorem to $\{g_{\varepsilon_n}\}$ on each compact set $K_m \subset S$, and then doing a diagonalization argument, there exists a subsequence of $\{n_k\}_{k\geq 0} \subset \{n\}_{n\geq 0}$, such that $g^*_{\varepsilon_{n_k}}$ converges pointwise to a function that we denote by $g_0^* \in \mathrm{Lip}(c, S)$. To simplify the notation, let n denote the convergent subsequence.

Since $\rho = \rho_+ - \rho_-$, where $\rho_+ \in L^1(a)$ and $\mu + \varepsilon_0\rho \in \mathcal{P}(S)$, $\mu \in L^1(a)$ implies $\rho_- \in L^1(a)$, and therefore

$$\int_S a d|\rho| < \infty.$$

Here $|\rho| = \rho_+ + \rho_-$. Recall that for any $\varepsilon \in (0, \varepsilon_0)$, $g^*_\varepsilon(0) = 0$. For any $x \in S$,

$$g^*_\varepsilon(x) \leq g^*_\varepsilon(0) + c(0, x) \leq a(0) + a(x).$$

Thus by the dominated convergence theorem

$$f'_+(0) = \lim_{n\to\infty} \int_S g^*_{\varepsilon_n} \mathrm{d}\rho = \int_S g_0^* \mathrm{d}\rho.$$

Lastly, to connect g_0^* back to g_+^*, note that by the lower semicontinuity of $G(\cdot\|\nu)$,

$$\begin{aligned}
G(\mu\|\nu) &\leq \liminf_{n\to\infty} G(\mu + \varepsilon_n\rho\|\nu) \\
&= \liminf_{n\to\infty} \left(\int_S g^*_{\varepsilon_n} d(\mu + \varepsilon_n\rho) - \log \int_S e^{g^*_{\varepsilon_n}} \mathrm{d}\nu \right) \\
&\leq \limsup_{n\to\infty} \int_S g^*_{\varepsilon_n} d(\mu + \varepsilon_n\rho) - \liminf_{n\to\infty} \log \int_S e^{g^*_{\varepsilon_n}} \mathrm{d}\nu \\
&\leq \int_S g_0^* \mathrm{d}\mu - \log \int_S e^{g_0^*} \mathrm{d}\nu \\
&\leq G(\mu\|\nu).
\end{aligned}$$

The third inequality uses dominated convergence, (4.9), and Fatou's lemma. The fourth inequality uses Lemma 4.8.

Since both sides of the inequality coincide, g_0^* must be an optimizer for the variational expression (4.4). By Theorem 4.9 and equation (4.6), we have $g_0^*(x) \leq g_+^*(x)$ for all $x \in S$ and $g_0^*(x) = g_+^*(x)$ for all $x \in \mathrm{supp}(\rho_-) \subset$

$\mathrm{supp}(\mu)$. Thus

$$f'_+(0) = \int_S g_0^* \mathrm{d}\rho \leq \int_S g_+^* \mathrm{d}\rho, \tag{4.10}$$

and the other direction of the inequality is proved. Combining (4.10) and (4.8) gives

$$\lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(G(\mu + \varepsilon\rho \| \nu) - G(\mu \| \nu) \right) = \int_S g_+^* \mathrm{d}\rho.$$

We next consider the second statement, and now use that we have

$$G_\Gamma(\mu\|\nu) = \sup_{g \in \Gamma} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\} = \int_S g_-^* \mathrm{d}\mu - \log \int_S e^{g_-^*} \mathrm{d}\nu.$$

For the given $\rho \in \mathcal{M}_0(S)$ and $\varepsilon \in (0, \varepsilon_0)$ we have

$$\begin{aligned} G_\Gamma(\mu\|\nu + \varepsilon\rho) &= \sup_{g \in \Gamma} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g d(\nu + \varepsilon\rho) \right\} \\ &\geq \int_S g_-^* \mathrm{d}\mu - \log \int_S e^{g_-^*} d(\nu + \varepsilon\rho) \\ &= G_\Gamma(\mu\|\nu) - \varepsilon \frac{\int_S e^{g_-^*} \mathrm{d}\rho}{\int_S e^{g_-^*} \mathrm{d}\nu} + O(\varepsilon^2), \end{aligned}$$

and thus

$$\liminf_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(G_\Gamma(\mu\|\nu + \varepsilon\rho) - G_\Gamma(\mu\|\nu) \right) \geq - \frac{\int_S e^{g_-^*} \mathrm{d}\rho}{\int_S e^{g_-^*} \mathrm{d}\nu}.$$

For the reverse direction the line of argument parallels the previous case. With now $f(\varepsilon) = G_\Gamma(\mu\|\nu + \varepsilon\rho)$, we again have a right derivative at $\varepsilon = 0$. Let $g_\varepsilon^* \in \mathrm{Lip}(c, S)$ satisfy $g_\varepsilon^*(x^*) = 0$ for some point x_0 in the support of ν, and

$$G_\Gamma(\mu\|\nu + \varepsilon\rho) = \int_S g_\varepsilon^* \mathrm{d}\mu - \log \int_S e^{g_\varepsilon^*} d(\nu + \varepsilon\rho).$$

Without loss we can assume $f(\varepsilon)$ is differentiable at $\varepsilon > 0$, and for $\delta > 0$ have the bounds

$$G_\Gamma(\mu\|\nu + (\varepsilon \pm \delta)\rho) \geq \int_S g_\varepsilon^* \mathrm{d}\mu - \log \int_S e^{g_\varepsilon^*} d(\nu + (\varepsilon \pm \delta)\rho).$$

Therefore

$$\begin{aligned} \liminf_{\delta \to 0} \frac{1}{\delta} \left[G_\Gamma(\mu\|\nu + (\varepsilon \pm \delta)\rho) - G_\Gamma(\mu\|\nu + \varepsilon\rho) \right] &\geq - \liminf_{\delta \to 0} \frac{1}{\delta} \log \left(1 \pm \delta \frac{\int_S e^{g_\varepsilon^*} \mathrm{d}\rho}{\int_S e^{g_\varepsilon^*} d(\nu + \varepsilon\rho)} \right) \\ &= \mp \frac{\int_S e^{g_\varepsilon^*} \mathrm{d}\rho}{\int_S e^{g_\varepsilon^*} d(\nu + \varepsilon\rho)}, \end{aligned}$$

which implies

$$f'(\varepsilon) = -\frac{\int_S e^{g^*_\varepsilon} \mathrm{d}\rho}{\int_S e^{g^*_\varepsilon} d(\nu + \varepsilon\rho)}.$$

We can assume that there is a sequence $\varepsilon_n \to 0$ and g^*_0 such that $g^*_{\varepsilon_n}$ converges pointwise to g^*_0, and so under the assumptions of the theorem

$$f'(0) = \lim_{n\to\infty} f'(\varepsilon) = -\frac{\int_S e^{g^*_0} \mathrm{d}\rho}{\int_S e^{g^*_0} \mathrm{d}\nu}.$$

Using lower semicontinuity of $G_\Gamma(\mu\|\cdot)$, as in the proof of the first part

$$G_\Gamma(\mu\|\nu) = \int_S g^*_0 \mathrm{d}\mu - \log \int_S e^{g^*_0} \mathrm{d}\nu,$$

and so g^*_0 is an optimizer. Again we have that $g^*_0 = g^*_-$ in $\mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)$ and hence in $\mathrm{supp}(\rho_-)$. Since $g^*_0 \geq g^*_-$ otherwise, this implies

$$f'(0) = -\frac{\int_S e^{g^*_0} \mathrm{d}\rho}{\int_S e^{g^*_0} \mathrm{d}\nu} = -\frac{\int_S e^{g^*_0} d(\rho_+ - \rho_-)}{\int_S e^{g^*_0} \mathrm{d}\nu} \leq -\frac{\int_S e^{g^*_-} \mathrm{d}\rho}{\int_S e^{g^*_-} \mathrm{d}\nu}.$$

□

Remark 4.17. When $\rho \in \mathcal{M}_0(S)$ is taken such that there exists $\varepsilon_0 > 0$ such that for $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$, $\mu + \varepsilon\rho \in \mathcal{P}(S)$, then by applying the above theorem to ρ and $-\rho$ respectively, we can conclude $G_\Gamma(\mu + \varepsilon\rho\|\nu)$ as a function of ε is differentiable at $\varepsilon = 0$ with derivative $\int_S g^*_+ \mathrm{d}\rho$. A similar statement applies to $G_\Gamma(\mu\|\nu + \varepsilon\rho)$.

Remark 4.18. One can consider g^*_+ defined in (4.5) the unique potential associated with $G_\Gamma(\mu\|\nu)$. This g^*_+ is similar to the Kantorovich potential in the optimal transport literature. However, for the optimal transport cost $W_c(\mu, \nu)$ more conditions are needed (*e.g.*, [23], Prop. 7.18) to ensure the uniqueness of the Kantorovich potential. Here under very mild conditions we are able to confirm the uniqueness of the potential, and prove that it is the directional derivative of the corresponding Γ-divergence, as is case of the Kantorovich potential for optimal transport cost when its uniqueness is established.

5. Limits and approximations of Γ-divergence

In this section, we consider limits that are obtained as the admissible set gets large or small, and the Γ-divergence will be approximated by relative entropy or a transport distance, respectively. We also consider in special cases more informative expansions. Throughout the section we assume the conditions of Theorem 4.9.

Fix an admissible set of Γ_0 of the form $\mathrm{Lip}(c, S; C_b(S))$ as in (4.1). Then the conditions of Theorem 4.9 hold for $\Gamma = b\Gamma_0 = \{b \cdot f : f \in \Gamma_0\}$ if $b > 0$, and the following proposition holds.

Proposition 5.1. *For $\mu, \nu \in \mathcal{P}(S)$,*

$$\lim_{b\to\infty} G_{b\Gamma_0}(\mu\|\nu) = R(\mu\|\nu).$$

Proof. We separate the proof into two cases, $R(\mu\|\nu) < \infty$ and $R(\mu\|\nu) = \infty$.

1) If $R(\mu\|\nu) < \infty$, then for any $b > 0$,

$$G_{b\Gamma_0}(\mu\|\nu) = \inf_{\gamma\in\mathcal{P}(S)} \{W_{b\Gamma_0}(\mu,\gamma) + R(\gamma\|\nu)\} \le R(\mu\|\nu) < \infty. \tag{5.1}$$

From Theorem 4.9 we know there exists a unique optimizer γ^* for each b, which we write as γ_b^*. Note that

$$R(\gamma_b^*\|\nu) \le R(\mu\|\nu) < \infty,$$

and therefore $\{\gamma_b^*\}_{b>0}$ is precompact in the weak topology ([9], Lem. 1.4.3(c)). Given any subsequence b_k, there exists a further subsequence (again denoted by b_k) and $\gamma_\infty^* \in \mathcal{P}(S)$ such that $\gamma_{b_k}^* \Rightarrow \gamma_\infty^*$. On the other hand,

$$\begin{aligned} W_{b\Gamma_0}(\mu,\gamma_b^*) &= \sup_{f\in b\Gamma_0} \left\{ \int_S f d(\mu - \gamma_b^*) \right\} \\ &= b \sup_{f\in\Gamma_0} \left\{ \int_S f d(\mu - \gamma_b^*) \right\} = bW_{\Gamma_0}(\mu,\gamma_b^*), \end{aligned}$$

and $W_{b\Gamma_0}(\mu,\gamma_b^*) \le G_{b\Gamma_0}(\mu\|\nu) \le R(\mu\|\nu) < \infty$. Thus

$$W_{\Gamma_0}(\mu,\gamma_\infty^*) \le \liminf_{k\to\infty} W_{\Gamma_0}(\mu,\gamma_{b_k}^*) = \liminf_{k\to\infty} \frac{1}{b_k} W_{b_k\Gamma_0}(\mu,\gamma_{b_k}^*) \le \liminf_{k\to\infty} \frac{1}{b_k} R(\mu\|\nu) = 0,$$

and since Γ_0 is admissible, $\gamma_\infty^* = \mu$. We thus conclude that

$$\liminf_{k\to\infty} G_{b_k\Gamma_0}(\mu\|\nu) = \liminf_{k\to\infty} \left(W_{b_k\Gamma_0}(\mu,\gamma_{b_k}^*) + R(\gamma_{b_k}^*\|\nu) \right) \ge \liminf_{k\to\infty} R(\gamma_{b_k}^*\|\nu) \ge R(\mu\|\nu),$$

and since the original subsequence was arbitrary

$$\liminf_{b\to\infty} G_{b\Gamma_0}(\mu\|\nu) \ge R(\mu\|\nu).$$

On the other hand, we have by (5.1) that

$$\limsup_{b\to\infty} G_{b\Gamma_0}(\mu\|\nu) \le R(\mu\|\nu),$$

and the statement is proved.

2) $R(\mu\|\nu) = \infty$. For this case, we want to prove that $\liminf_{b\to\infty} G_{b\Gamma_0}(\mu\|\nu) = \infty$. If not, then there exists a subsequence $\{b_k\}_{b\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} G_{b_k\Gamma_0}(\mu\|\nu) < \infty.$$

For this subsequence, we can apply the argument used in part 1) to conclude there exists $\gamma_{b_k}^*$ such that

$$G_{b_k\Gamma_0}(\mu\|\nu) = W_{b_k\Gamma_0}(\mu,\gamma_{b_k}^*) + R(\gamma_{b_k}^*\|\nu).$$

Moreover there exists a further subsequence of this sequence, which for simplicity we also denote by $\{b_k\}_{k\in\mathbb{N}}$, which satisfies $\gamma_{b_k}^* \Rightarrow \mu$. Then by the same argument as in 1), we would conclude

$$\lim_{k\to\infty} G_{b_k\Gamma_0}(\mu\|\nu) \ge R(\mu\|\nu) = \infty.$$

This contradiction proves the statement. □

On the other hand, if $\Gamma = \delta\Gamma_0$ for small $\delta > 0$, we can approximate the Γ-divergence in terms of the W_{Γ_0}.

Proposition 5.2. *For $\mu, \nu \in \mathcal{P}(S)$*

$$\lim_{\delta\to 0} \frac{1}{\delta} G_{\delta\Gamma_0}(\mu\|\nu) = W_{\Gamma_0}(\mu,\nu).$$

Proof. For any $\delta > 0$, Jensen's inequality implies

$$\begin{aligned}
\frac{1}{\delta} G_{\delta\Gamma_0}(\mu\|\nu) &= \frac{1}{\delta} \sup_{g\in\delta\Gamma_0} \left\{ \int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu \right\} \\
&\le \frac{1}{\delta} \sup_{g\in\delta\Gamma_0} \left\{ \int_S g \mathrm{d}\mu - \int_S g \mathrm{d}\nu \right\} \\
&= \sup_{g\in\Gamma_0} \left\{ \int_S g \mathrm{d}\mu - \int_S g \mathrm{d}\nu \right\} \\
&= W_{\Gamma_0}(\mu,\nu),
\end{aligned}$$

and therefore

$$\limsup_{\delta\to 0} \frac{1}{\delta} G_{\delta\Gamma_0}(\mu\|\nu) \le W_{\Gamma_0}(\mu,\nu).$$

For the reverse inequality we consider two cases.

1) $W_{\Gamma_0}(\mu,\nu) < \infty$. For $0 < \delta < 1$ the argument used above shows

$$G_{\delta\Gamma_0}(\mu\|\nu) \le \delta W_{\Gamma_0}(\mu,\nu) \le W_{\Gamma_0}(\mu,\nu) < \infty.$$

By Theorem 4.9, we know there exists $\gamma^*_\delta \in \mathcal{P}(S)$, such that

$$G_{\delta\Gamma_0}(\mu\|\nu) = W_{\delta\Gamma_0}(\mu,\gamma^*_\delta) + R(\gamma^*_\delta\|\nu).$$

Since $R(\gamma^*_\delta\|\nu) < G_{\delta\Gamma_0}(\mu\|\nu) \le W_{\Gamma_0}(\mu,\nu)$ for $\delta \in (0,1)$, for any sequence $\delta_k \subset (0,1)$ there a further a subsequence (again denoted δ_k) such that δ_k is decreasing, $\lim_{k\to\infty}\delta_k = 0$, and $\gamma^*_{\delta_k}$ converges weakly to a probability measure, which we denote as γ^*_0. Then by the lower semicontinuity of $R(\cdot\|\nu)$

$$R(\gamma^*_0\|\nu) \le \liminf_{k\to\infty} R(\gamma^*_{\delta_k}\|\nu) \le \liminf_{k\to\infty} G_{\delta_k\Gamma_0}(\mu,\nu) \le \lim_{k\to\infty} \delta_k W_{\Gamma_0}(\mu,\nu) = 0.$$

Since $R(\gamma^*_0\|\nu) \ge 0$ with equality if and only if $\gamma^*_0 = \nu$, we conclude $R(\gamma^*_0\|\nu) = 0$ and $\gamma^*_0 = \nu$. Therefore

$$\begin{aligned}
\liminf_{k\to\infty} \frac{1}{\delta_k} G_{\delta_k\Gamma_0}(\mu\|\nu) &\ge \liminf_{k\to\infty} \frac{1}{\delta_k} W_{\delta_k\Gamma_0}(\mu,\gamma^*_{\delta_k}) \\
&= \liminf_{k\to\infty} W_{\Gamma_0}(\mu,\gamma^*_{\delta_k}) \\
&\ge W_{\Gamma_0}(\mu,\gamma^*_0) = W_{\Gamma_0}(\mu,\nu),
\end{aligned}$$

and since the original sequence was arbitrary

$$\liminf_{\delta\to 0}\frac{1}{\delta}G_{\delta\Gamma_0}(\mu\|\nu)\geq W_{\Gamma_0}(\mu,\nu).$$

2) $W_{\Gamma_0}(\mu,\nu)=\infty$. If $\liminf_{\delta\to 0}\frac{1}{\delta}G_{\delta\Gamma_0}(\mu\|\nu)<\infty$, then there is a subsequence $\{\delta_l\}_{l\in\mathbb{N}}\subset(0,1)$ that achieves this lim inf. From essentially the same proof above applied to this subsequence, it can be shown there exists a further subsequence (again denoted $\{\delta_l\}$) and $\gamma_0^*\in\mathcal{P}(S)$ such that

$$G_{\delta_l\Gamma_0}(\mu\|\nu)=W_{\delta_l\Gamma_0}(\mu,\gamma_\delta^*)+R(\gamma_{\delta_l}^*\|\nu),$$

and $\gamma_l^*\Rightarrow\gamma_0^*$. Denote $M\doteq\liminf_{\delta\to 0}\frac{1}{\delta}G_{\delta\Gamma_0}(\mu\|\nu)=\lim_{l\to\infty}\frac{1}{\delta_l}G_{\delta_l\Gamma_0}(\mu\|\nu)<\infty$. Since for l large enough

$$R(\gamma_{\delta_l^*}\|\nu)\leq G_{\delta_l\Gamma_0}(\mu\|\nu)\leq\delta_l(M+1),$$

we have

$$R(\gamma_0^*\|\nu)\leq\liminf_{l\to\infty}R(\gamma_{\delta_l}^*\|\nu)\leq\lim_{l\to\infty}\delta_l(M+1)=0,$$

and thus $\gamma_0^*=\nu$. However this leads to

$$\begin{aligned}M=\lim_{l\to\infty}\frac{1}{\delta_l}G_{\delta_l\Gamma_0}(\mu\|\nu)&\geq\lim_{l\to\infty}\frac{1}{\delta_l}W_{\delta_l\Gamma_0}(\mu,\gamma_{\delta_l}^*)\\&=\lim_{l\to\infty}W_{\Gamma_0}(\mu,\gamma_{\delta_l}^*)\geq W_{\Gamma_0}(\mu,\nu)=\infty.\end{aligned}$$

This contradiction implies

$$\liminf_{\delta\to 0}\frac{1}{\delta}G_{\delta\Gamma_0}(\mu\|\nu)=\infty=W_{\Gamma_0}(\mu,\nu).$$

□

We now consider more refined approximations when b is large. Previously we described the limiting behavior when we vary the size of Γ. From Proposition 5.1, we know that when $\mu\not\ll\nu$, $\lim_{b\to\infty}G_{b\Gamma_0}(\mu\|\nu)=\infty$. In some applications one might use a large transport cost as "penalty" so that while allowing non-absolutely continuous perturbations, control on $G_\Gamma(\mu\|\nu)$ will ensure that μ is not too far away from ν.

In the rest of this section, we investigate the behavior when $b\to\infty$, and in particular how $G_{b\Gamma_0}(\mu\|\nu)$ will behave for fixed μ and ν. We only consider the case that $\Gamma_0=\mathrm{Lip}(c,S;C_b(S))$ for some function c satisfies the condition of Theorem 4.2, Assumption 4.4 and Assumption 4.6, and $\mu,\nu\in L^1(a)$ with a in Assumption 4.6. We separate the cases depending on whether μ and ν are discrete or continuous. The results presented here are only for special cases, and further development of these sorts of expansions would be useful.

5.1. Finitely supported discrete measures

We will consider the case where $\mathrm{supp}(\nu)$ has finite cardinality, and μ is also discrete with finite support. The proof of the following appears in the Appendix.

Theorem 5.3. *Suppose ν and μ are discrete with finite support, where* $\mathrm{supp}(\nu) = \{x_i\}_{1\le i\le N}$ *and* $\mathrm{supp}(\mu) = \{y_j\}_{1\le j\le M}$. *Then there exists $\tilde\gamma \in \mathcal{P}(S)$ with $\tilde\gamma \ll \nu$ such that*

$$G_{b\Gamma_0}(\mu\|\nu) = bW_{\Gamma_0}(\mu,\tilde\gamma) + R(\tilde\gamma\|\nu) + e(b), \tag{5.2}$$

where $e(b) \le 0$ satisfies $e(b) \to 0$ as $b \to \infty$. Furthermore, we can characterize $\tilde\gamma$ as the measure that minimizes $R(\gamma\|\nu)$ over the collection of $\gamma \in \mathcal{P}(S)$ that satisfy the constraint

$$W_{\Gamma_0}(\mu,\gamma) = \inf_{\theta\ll\nu} W_{\Gamma_0}(\mu,\theta). \tag{5.3}$$

If to simplify the statement below we further assume that

$$c(y_j,x_i) \ne c(y_j,x_l)$$

for $1 \le j \le M$ and $1 \le i \ne l \le N$, then $\tilde\gamma$ has the following form. Let S_i be the indices j in $\{1,\dots,M\}$ for which x_i is the point in $\{x_l\}_{1\le l\le N}$ closest to y_j. Then for $1 \le i \le N$,

$$\tilde\gamma(\{x_i\}) = \sum_{j\in S_i} \mu(\{y_j\}).$$

Remark 5.4. In discrete case, it is easily checked that the infimum in (5.3) is achieved. Take a sequence of $\theta_n \ll \nu$ such that

$$W_{\Gamma_0}(\mu,\theta_n) \le \inf_{\theta\ll\nu} W_{\Gamma_0}(\mu,\theta) + 1/n.$$

Since θ_n is supported on the compact set $\mathrm{supp}(\nu) = \{x_i\}_{1\le i\le N}$ $\{\theta_n\}_{n\in\mathbb{N}}$ is compact, and hence there exist $\theta^* \ll \nu$ and a subsequence $\{\theta_{n_k}\}_{k\in\mathbb{N}}$ that converges to θ^* weakly. By the lower semicontinuity of W_{Γ_0}

$$W_{\Gamma_0}(\mu,\theta^*) \le \liminf_{n\to\infty} W_{\Gamma_0}(\mu,\theta_n) \le \inf_{\theta\ll\nu} W_{\Gamma_0}(\mu,\theta),$$

and therefore θ^* achieves the infimum of (5.3).

5.2. An example with ν is continuous

To illustrate an interesting scaling phenomenon, here we consider the example with $S = \mathbb{R}$, $c(x,y) = |x-y|$, $\nu = \mathrm{Unif}([0,1])$, $\mu = \delta_0$. Consider $\gamma^*(\mathrm{d}x) = c_0 e^{-bx}\mathrm{d}x$ and $g^*(x) = -bx$ for $0 \le x \le 1$, where c_0 is the normalizing constant. For this example $\Gamma_0 = \mathrm{Lip}(c,S;C_b(S))$ is the set of bounded functions over $\mathbb{R}$ with Lipschitz constant 1. It is easily checked using Theorem 4.13 that γ^* and g^* are the optimizers in

$$G_{b\Gamma_0}(\mu\|\nu) = \inf_{\gamma\in\mathcal{P}(S)} \{W_{b\Gamma_0}(\mu,\gamma) + R(\gamma\|\nu)\} = \sup_{g\in b\Gamma_0} \left\{ \int_S g\mathrm{d}\mu - \log \int_S e^g \nu \right\}.$$

Thus we have

$$G_{b\Gamma_0}(\mu\|\nu) = -\int_0^1 bx\mathrm{d}\mu - \log\int_0^1 e^{-bx}\mathrm{d}\nu = -\log\int_0^1 e^{-bx}\mathrm{d}x = \log\left(\frac{b}{1-e^{-b}}\right),$$

and in this case, $G_{b\Gamma_0}(\mu\|\nu)$ scales as $\log(b) + o(\log(b))$.

For comparison we consider the optimal transport cost between μ and ν. We have

$$W_{bc}(\mu,\nu) \doteq \sup_{g\in b\Gamma_0}\left\{\int_S g\mathrm{d}\mu - \int_S g\mathrm{d}\nu\right\} = b\sup_{g\in\Gamma_0}\left\{\int_S g\mathrm{d}\mu - \int_S g\mathrm{d}\nu\right\} = bW_c(\mu,\nu)$$

and one can calculate that $W_c(\mu,\nu) = 1/2$. Thus $W_{bc}(\mu,\nu) = b/2$, and so $G_{b\Gamma_0}(\mu\|\nu)$ gives a much smaller divergence between non absolutely continuous measures μ and ν than the corresponding optimal transport cost when the admissible $\Gamma = b\Gamma_0$ is becoming large.

6. Application to uncertainty bounds

6.1. Extension to unbounded functions

From

$$G_\Gamma(\mu\|\nu) \doteq \sup_{g\in\Gamma}\left\{\int_S g\mathrm{d}\mu - \log\int_S e^g\mathrm{d}\nu\right\}$$

we get for all $g\in\Gamma$,

$$\int_S g\mathrm{d}\mu \le G_\Gamma(\mu\|\nu) + \log\int_S e^g\mathrm{d}\nu.$$

The inequality above with relative entropy in place of $G_\Gamma(\mu\|\nu)$ is the key to uncertainty bounds in [11]. We would like to extend this inequality to unbounded functions. Define

$$\hat\Gamma_+ = \{f:\ \text{there exist } g_i\in\Gamma \text{ with } c\le g_i(x)\uparrow f(x) \text{ for } x\in S\},$$

and

$$\hat\Gamma_- = \{f:\ \text{there exist } g_i\in\Gamma \text{ with } c\ge g_i(x)\downarrow f(x) \text{ for } x\in S\}.$$

Proposition 6.1. *For $g\in\hat\Gamma_+\cup\hat\Gamma_-$, we have*

$$\int_S g\mathrm{d}\mu \le G_\Gamma(\mu\|\nu) + \log\int_S e^g\mathrm{d}\nu. \tag{6.1}$$

Proof. The proof is straightforward. Take $g\in\hat\Gamma_+$. Then there exist $g_i\in\Gamma$, which are bounded below, and increase to g pointwise in S. By monotone convergence theorem,

$$\lim_{i\to\infty}\int_S g_i\mathrm{d}\mu = \int_S g\mathrm{d}\mu,$$

and

$$\lim_{i\to\infty}\int_S e^{g_i}\mathrm{d}\nu = \int_S e^g\mathrm{d}\nu.$$

Since $g_i\in\Gamma$, for all i

$$\int_S g_i\mathrm{d}\mu \le G_\Gamma(\mu\|\nu) + \log\int_S e^{g_i}\mathrm{d}\nu.$$

Taking $i \to \infty$ in the last display gives (6.1). For $g \in \hat{\Gamma}_-$ the reasoning is essentially the same. □

In the case when $\Gamma = \mathrm{Lip}(c, S; C_b(S))$, where c satisfies the conditions introduced in Section 4, we can get a stronger version of the result. The proof is essentially the same as in Lemma 4.8, and is omitted.

Proposition 6.2. *Assume $c : S \times S \to \mathbb{R} \cup \{+\infty\}$ satisfies Conditions 4.1, 4.4, 4.6 and 4.7. Fix $\mu, \nu \in L^1(a)$. Then for $g \in \mathrm{Lip}(c, S)$*

$$\int_S g \mathrm{d}\mu \le G_\Gamma(\mu \| \nu) + \log \int_S e^g \mathrm{d}\nu.$$

6.2. Decomposition and scaling properties

A property of great importance in applications of relative entropy is the chain rule. When probability measures can be decomposed, such as when Markov measures on a path space are written as the repeated integration with respect to transition kernels, the chain rule allows one to decompose the relative entropy of two such measures on path space in terms of the simpler relative entropies of the transition kernels. This decomposition also exhibits important scaling properties of relative entropy, *e.g.*, that for such Markov measures on path space the relative entropy scales proportionate to the number of time steps.

Except in special circumstances, optimal transport metrics do not possess a property like the chain rule, and it is therefore not to be expected that Γ-divergence would either. However, if one considers certain classes of functions on path space, then one can show there are analogous decomposition and scaling properties. In this section we will discuss a setting relevant to many applications, though the results have many analogues and possible generalizations.

As usual, we assume that S is a Polish space, and let $p : S \times \mathcal{B}(S)$ be a probability transition kernel:

- for every $A \in \mathcal{B}(S)$ the map $x \to p(x, A)$ is Borel measurable, and
- for every $x \in S$, $p(x, \cdot)$ is in $\mathcal{P}(S)$.

The quantities of interest are large and infinite time averages, both with respect to time and the underlying distribution, and we wish to bound in a tight fashion the error in such quantities due to model misspecification. Thus if q is some other transition kernel, then we seek useful bounds on differences of the form

$$\frac{1}{cT} \log E^{\gamma,p}\left[e^{c\sum_{i=1}^T f(X_i)}\right] - E^{\theta,q}\left[\frac{1}{T}\sum_{i=1}^T f(X_i)\right],$$

where $E^{\gamma,p}$ indicates that the chain uses transition kernel p and initial distribution γ, and similarly for $E^{\theta,q}$. Under conditions, relative entropy can provide useful bounds when $q(x, \cdot) \ll p(x, \cdot)$ for a suitable set of $x \in S$. One question then is under what conditions will the Γ-divergence allow one to weaken the absolute continuity restriction. It is also worth noting that even when $q(x, \cdot) \ll p(x, \cdot)$ the bounds obtained using the Γ-divergence (when applicable) are tighter, since it is never greater than relative entropy, and in some cases the improvement can be dramatic. These issues will be explored in greater detail elsewhere.

It follows directly from discussion in earlier sections that even in the setting of product measures that one must restrict the class of functions f under consideration. When considering Markov measures, the following definition is relevant.

Definition 6.3. For a transition kernel p, let

$$\mathcal{R}(\Gamma, p) = \left\{ -\log \int_S e^{-g(y)} p(x, \mathrm{d}y) - g(x) + a : g \in \Gamma \text{ and } a \in \mathbb{R} \right\}.$$

Then $\mathcal{R}(\Gamma, p)$ will determine the set of costs f such that bounds can be obtained using the Γ-divergence. In particular, we have the following.

Theorem 6.4. *Suppose that $f \in \mathcal{R}(\Gamma, p)$ for some g and a. Consider any transition kernel q on S and any stationary probability measure π_q of q. Then*

$$\int_S f(x)\pi_q(\mathrm{d}x) \leq \int_S G_\Gamma(q(x,\cdot)\|p(x,\cdot))\pi_q(\mathrm{d}x) + a.$$

Remark 6.5. If p is ergodic then we recognize

$$f(x) = -\log \int_S e^{-g(y)} p(x, \mathrm{d}y) - g(x) + a$$

as the equation that uniquely characterizes the multiplicative cost

$$a = \lim_{M\to\infty} \frac{1}{M} \log E^p e^{-\sum_{i=0}^{M-1} f(X_i)},$$

with g a type of cost potential. Note that for a given f the function g plays no role in the bound. We need to check that f is in the range of Γ (which of course imposes restrictions on f), but the bound does not depend on knowing the specific form of g.

Proof. Since $g \in \Gamma$

$$\begin{aligned} g(x) &= -f(x) - \log \int_S e^{-g(y)} p(x, \mathrm{d}y) + a \\ &= -f(x) + \inf_{q(x,\mathrm{d}y)} \left[G_\Gamma(q(x,\cdot)\|p(x,\cdot)) + \int_S g(y) q(x, \mathrm{d}y) \right] + a. \end{aligned}$$

For the given transition kernel q

$$g(x) \leq -f(x) + \left[G_\Gamma(q(x,\cdot)\|p(x,\cdot)) + \int_S g(y) q(x, \mathrm{d}y) \right] + a,$$

and integrating both sides with respect to $\pi_q(\mathrm{d}x)$ and using $\int_S q(x, \mathrm{d}y)\pi_q(\mathrm{d}x) = \pi_q(\mathrm{d}y)$ gives the result. □

We next consider two examples to illustrate Definition 6.3.

Example 6.6. $S = \mathbb{R}$, $p(x,\cdot) \sim N(\alpha x, \sigma^2)$ is normal distribution with mean αx and variance σ^2, where $0 < \alpha < 1$. Let $g(x) = -bx^2 - cx - d$, for $b, c, d \in \mathbb{R}$.

Then direct computation gives that when $1 - 2b\sigma^2 > 0$

$$\begin{aligned} &-\log \int_S e^{-g(y)} p(x, \mathrm{d}y) - g(x) + a \\ &= -\frac{b\alpha^2 x^2 + c\alpha x + c^2\sigma^2/2}{1 - 2b\sigma^2} + bx^2 + cx + a \\ &= b\left(1 - \frac{\alpha^2}{1 - 2b\sigma^2}\right) x^2 + c\left(1 - \frac{\alpha}{1 - 2b\sigma^2}\right) x + a. \end{aligned}$$

Letting $k(b) = b(1 - \frac{\alpha^2}{1-2b\sigma^2})$,

$$k'(b) = 1 - \frac{\alpha^2}{(1-2b\sigma^2)^2}.$$

Since $1 - 2b\sigma^2 > 0$, we can conclude k reaches its maximum at $1 - 2b\sigma^2 = \alpha$, *i.e.*, $b = \frac{1-\alpha}{2\sigma^2}$, where $k(b) = \frac{(1-\alpha)^2}{2\sigma^2}$.If $b \to \frac{1}{2\sigma^2}$ then $k(b) \to -\infty$. Also notice that when $b \neq \frac{1-\alpha}{2\sigma^2}$, we can pick c to make the coefficient of x to be any given number. Thus with $p(x,\cdot) \sim N(\alpha x, \sigma^2)$ and $\Gamma = \left\{bx^2 + cx + d : b, c, d \in \mathbb{R}\right\}$,

$$\mathcal{R}(\Gamma, p) = \left\{bx^2 + cx + d : b < \frac{(1-\alpha)^2}{2\sigma^2}, c, d \in \mathbb{R}\right\} \cup \left\{\frac{(1-\alpha)^2}{2\sigma^2}x^2 + d : d \in \mathbb{R}\right\}.$$

Example 6.7. $S = \{x_1, x_2, \ldots, x_n\}$ is a finite space, and there is a cost function $c : S \times S \to \mathbb{R}_+$ associated with this space. Take $\Gamma = \mathrm{Lip}(c, S; C_b(S))$. Since p is a transition matrix we denote $p_{ij} = p(x_i, x_j)$ and $P = (p_{ij})_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$.

A question we ask here is whether there exists $\sigma > 0$ such that $\sigma\Gamma \in R(\Gamma, p)$. In other words, does there exist $\sigma > 0$ such that for any $f \in \sigma\Gamma$ we can find $g \in \Gamma$ and $a \in \mathbb{R}$ such that

$$f(x_i) = -g(x_i) - \log \sum_{j=1}^{n} p(x_i, x_j) e^{-g(x_j)} + a, \quad i = 1, 2, \ldots, n.$$

If $R(\Gamma, p)$ includes such a neighborhood of zero, then when combined with Theorem 6.4 it would allow for sensitivity bounds, *i.e.*, bounds on quantities of the form

$$\frac{d}{\mathrm{d}\theta} \sum_{x \in S} \pi(\theta, x) f(x),$$

where $f \in \Gamma$, $\pi(\theta, \cdot)$ is the stationary distribution of $P(\theta)$, $P(0) = P$, and $P(\theta)$ depends smoothly on a vector of parameters θ (see [11]). In contrast with [11], we would not need that the transition matrices be mutually absolutely continuous.

Since S is finite we write f_i for $f(x_i)$ and let $f = (f_1, \ldots, f_n)$, and similarly for g. Then the relation above defines a mapping from (g, a) to f, which we denote it by $f = \varphi(g, a)$. Note that

$$(0, 0, \ldots, 0) = \varphi((0, 0, \ldots, 0), 0),$$

The $(n, n+1)$ dimensional matrix of partial derivatives takes the form

$$J = \left[(P - I), \mathbf{1}\right],$$

where I is the $n \times n$ identity matrix and $\mathbf{1}$ is a column vector of ones. If we can show that J is of full rank then the range of the mapping defined by J, *i.e.*, the linearization of φ will be onto $\mathbb{R}^n$. Then by the implicit function theorem there will be an open neighborhood U of $\mathbf{0} \in \mathbb{R}^n$ and a continuous function $\gamma : \mathbb{R}^n \to \mathbb{R}^n$ such that for all $f \in U$,

$$f = \varphi(0, \gamma(f)).$$

Since $O \doteq \{(y_1, y_2, \ldots, y_n) | (0, y_1, \ldots, y_{n-1}) \in \text{int}(\text{Lip}(c, S)), y_n \in \mathbb{R}\}$ is open, $\mathbf{0} \in U \cap \gamma^{-1}(O) \subset \mathbb{R}^n$ is also open. Thus we can pick $\sigma > 0$ such that $\mathbf{0} \in \sigma\Gamma \subset U \cap \phi^{-1}(O)$. So we have shown the existence of $\sigma > 0$ such that $\sigma\Gamma \in R(\Gamma, p)$.

Whether or not J is of full rank will depend on the structure of P. We have the following lemma.

Lemma 6.8. *Suppose that $S = \bar{S} \cup M$, where M consists of the transient states, and that when restricted to $\bar{S}$, P is ergodic. Then J is of full rank.*

Proof. Let π denote the stationary distribution of P. Then interpreting π as a column vector, it is the unique vector in the null space of $(P - I)^T$. According to the Fredholm alternative, the range of $(P - I)$ is the $n - 1$ dimensional collection of vectors $b \in \mathbb{R}^n$ such that $\langle b, \pi \rangle = 0$. Now $\langle \mathbf{1}, \pi \rangle > 0$, which shows that $\mathbf{1}$ is *not* in the range of $(P - I)$. Therefore the range of J is all of $\mathbb{R}^n$. □

To give a simple example of how the Γ-divergence could be used for model simplification, consider the situation where we are given an ergodic chain P with state space $\bar{S}$, and would like to replace P by a chain Q with state space $S = \bar{S} \cup M$, where the new states are intended to replace a (possibly large) number of states in $\bar{S}$, with the goal being to maintain good approximation of certain functionals of the stationary distribution. If π_q denotes the stationary distribution of Q on S and π_p that of P on $\bar{S}$, then one could not use relative entropy to obtain any bounds. Suppose we were to extend P to $\bar{S} \cup M$ (while keeping P as the transition matrix), by making all states in M transient. Then one could use the Γ-divergence as long as the functionals of interest are in $R(\Gamma, p)$ (with respect to the extended transition probabilities). Note that the location of the new states would be relevant to this question, since the costs f depend on these locations. Similarly, one could do sensitivity bounds for non-absolutely continuous transitions by using such a device.

7. Conclusion

In this paper, we defined a new divergence by starting with a variational representation for relative entropy and placing additional restrictions on the collection of test functions used in the representation, so as to relax the requirement of absolute continuity. Basic qualitative properties of the divergence were investigated, as well as its relationship with optimal transport metrics. Future work will use the divergence to develop uncertainty quantification bounds, sensitivity bounds and methods for model approximation and simplification for stochastic for models without the absolute continuity requirement. Also needed is further investigation of qualitative and computational aspects of the Γ-divergence.

Appendix A.

In this appendix we collect proofs of some intermediate results.

Proof of Lemma 2.1. If we prove item 3, then items 1 and 2 will follow from the corresponding statements when μ is restricted to $\mathcal{P}(S)$ [9]. If $m = \mu(S) \neq 1$, then taking $g(x) \equiv c$ a constant,

$$\int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu = c\mu(S) - c = c(m - 1).$$

Since $m \neq 1$ and $c \in \mathbb{R}$, the right hand side of equation (1.1) is ∞.

Suppose next that $\mu(S) = 1$ but $\mu \in \mathcal{M}(S) \backslash \mathcal{P}(S)$. Then there exist sets $A, B \in \mathcal{B}$ such that $A \cap B = \varnothing, A \cup B = S, \mu(A) < 0$ and $\mu(B) > 0$. For $c > 0$, let $g(x) = -c$ for $x \in A$ and $g(x) = 0$ for $x \in B$. Then

$$\int_S g \mathrm{d}\mu - \log \int_S e^g \mathrm{d}\nu = c\,|\mu(A)| - C_c,$$

where $C_c \in (\log \nu(B), 0)$ for all c. Letting $c \to \infty$ and using (1.1) (or more precisely the analogous statement using bounded measurable functions) shows $R(\mu\|\nu) = \infty$. □

Proof of Lemma 4.11. This can be shown by contradiction. Assume there exists $h \in \mathrm{Lip}(c, S)$ such that $\int_S |h| \mathrm{d}\gamma^* = \infty$. By symmetry, we can just consider h to be non-negative, since $\max(h, 0) \in \mathrm{Lip}(c, S)$ and $h = \max(h, 0) - \max(-h, 0)$. Thus we can assume there exists non-negative $h \in \mathrm{Lip}(c, S)$ satisfying

$$\int_S h \mathrm{d}\gamma^* = \infty,$$

and by the fact that $\mu \in L^1(a)$ together with Condition 4.6,

$$\int_S h \mathrm{d}\mu \le \int_S [h(0) + c(x, 0)]\, \mu(\mathrm{d}x) = h(0) + a(0) + \int_S a(x)\mu(\mathrm{d}x) < \infty.$$

Then

$$\begin{aligned} W_c(\mu, \gamma^*) &= \sup_{g \in \mathrm{Lip(c,S)}} \int_S g d(\mu - \gamma^*) \\ &\ge \limsup_{n\to\infty} \int_S \max(-h, -n) d(\mu - \gamma^*) \\ &= \limsup_{n\to\infty} \left[\int_S \max(-h, -n) \mathrm{d}\mu + \int_S \min(h, n) \mathrm{d}\gamma^* \right] \\ &= \int_S -h \mathrm{d}\mu + \int_S h \mathrm{d}\gamma^* \\ &= \infty, \end{aligned}$$

where the second to last equation comes from dominated and monotone convergence theorems applied to the first and second terms respectively. However, since γ^* is the optimizer, we have

$$W_c(\mu, \gamma^*) \le W_c(\mu, \gamma^*) + R(\gamma^*\|\nu) = G(\mu\|\nu) < \infty.$$

This contradiction shows the integrability of γ^* with respect to any $\mathrm{Lip}(c, S)$ function. □

Proof of Lemma 4.14. 1) For $x \in \mathrm{supp}(\mu)$, from an optimal transport plan between μ and γ^*, $\pi^* \in \Pi(\mu, \gamma^*)$ for $W_c(\mu, \gamma^*)$, we know there exists $y_x \in \mathrm{supp}(\nu)$ such that $(x, y_x) \in \mathrm{supp}(\pi^*)$. Thus by Remark 1.15 in [2],

$$g^*(x) = g^*(y_x) + c(x, y_x).$$

On the other hand, by Theorem 4.9, $g^*|_{\mathrm{supp}(\nu)\cup\mathrm{supp}(\mu)} \in \mathrm{Lip}(c, S)$. Thus, for other $y \in \mathrm{supp}(\nu)$, $g^*(x) \le c(x, y) + g^*(y)$, which in turn gives

$$g^*(x) \le \inf_{y \in \mathrm{supp}(\nu)} \{g^*(y) + c(x, y)\}.$$

By combining the two expressions above, we have for $x \in \mathrm{supp}(\mu)$, (4.5) also holds. In other words, g^* is totally characterized by $g^*|_{\mathrm{supp}}(\nu)$ and (4.5).

2) We check the Lipschitz condition for g_+^* for pairs of points according to whether they are in $\mathrm{supp}(\nu)$. First, since $g_+^*|_{\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)}$ is an optimizer for (4.4), by Theorem 4.9, $g_+^*|_{\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)}$ satisfies the Lipchitz condition, *i.e.*, for $y_1, y_2 \in \mathrm{supp}(\nu)$,

$$g_+^*(y_2) - c(y_1, y_2) \le g_+^*(y_1) \le g_+^*(y_2) + c(y_1, y_2). \tag{A.1}$$

For $x \notin \mathrm{supp}(\nu)$ and $y \in \mathrm{supp}(\nu)$, by (4.5) we have

$$g_+^*(x) \le g_+^*(y) + c(x, y).$$

On the other hand, for any $0 < n < \infty$, there exists $y_1 \in \mathrm{supp}(\nu)$ such that

$$g_+^*(x) \ge g_+^*(y_1) + c(x, y_1) - 1/n.$$

Notice that both y and y_1 are from $\mathrm{supp}(\mu)$, so from (A.1), we have $g_+^*(y_1) \ge g_+^*(y) - c(y, y_1)$, thus we have

$$\begin{aligned} g_+^*(x) &\ge g_+^*(y_1) + c(x, y_1) - 1/n \\ &\ge g_+^*(y) - c(y, y_1) + c(x, y_1) - 1/n \\ &\ge g_+^*(y) - c(y, x) - 1/n, \end{aligned}$$

where the last equation uses the triangle inequality property of c. Now since $n > 0$ is arbitrary, by getting $n \to \infty$, we have

$$g_+^*(x) \ge g_+^*(y) - c(x, y).$$

Combining, we have for $x \notin \mathrm{supp}(\nu)$, $y \in \mathrm{supp}(\mu)$,

$$g_+^*(y) - c(x, y) \le g_+^*(x) \le g_+^*(y) + c(x, y).$$

Lastly, we check for $x_1, x_2 \notin \mathrm{supp}(\nu)$ the Lipschitz constraint is satisfied. From the definition (4.5), we know for any $n < \infty$ there exists $y_1 \in \mathrm{supp}(\nu)$ such that

$$c(x_1, y_1) - 1/n \le g_+^*(x_1) - g_+^*(y_1).$$

Also, because $y_1 \in \mathrm{supp}(\nu)$,

$$g_+^*(x_2) - g_+^*(y_1) \le c(x_2, y_1).$$

Therefore

$$g_+^*(x_2) - g_+^*(x_1) \le (c(x_2, y_1) - c(x_1, y_1)) + 1/n \le c(x_1, x_2) + 1/n,$$

where the last inequality uses the triangle inequality property of c. Since $n > 0$ is arbitrary and we can swap the roles of x_1 and x_2, we have proved the Lipschitz condition of g_+^* for $x_1, x_2 \notin \mathrm{supp}(\nu)$. Thus the statement that $g_+^* \in \mathrm{Lip}(c, S)$ is proven.

For (4.6), notice that for $h \in \mathrm{Lip}(c, S)$, $x \in S$ and $y \in \mathrm{supp}(\nu)$,

$$h(x) \le h(y) + c(x, y).$$

So if $h(y) = g_+^*(y)$ for $y \in \mathrm{supp}(\nu)$, then for $x \in S \backslash \mathrm{supp}(\nu)$,

$$h(x) \le \inf_{y \in \mathrm{supp}(\nu)} \{h(y) + c(x,y)\} = \inf_{y \in \mathrm{supp}(\nu)} \{g_+^*(y) + c(x,y)\} = g_+^*(x).$$

Since g_+^* is also in $\mathrm{Lip}(c, S)$, this proves (4.6). □

Proof of Theorem 5.3. We use the representation $G_\Gamma(\mu\|\nu) = \inf_{\gamma \in \mathcal{P}(S)} \{R(\gamma\|\nu) + W_\Gamma(\mu,\gamma)\}$. First note that

$$G_{b\Gamma_0}(\mu\|\nu) = \inf_{\gamma \in \mathcal{P}(S)} \{R(\gamma\|\nu) + W_{b\Gamma_0}(\mu,\gamma)\} \le R(\tilde\gamma\|\nu) + W_{b\Gamma_0}(\mu,\tilde\gamma) = R(\tilde\gamma\|\nu) + bW_{\Gamma_0}(\mu,\tilde\gamma).$$

Next, fix any $\varepsilon > 0$, and take a near optimizer γ_b, so that for each b

$$G_{b\Gamma_0}(\mu\|\nu) \ge R(\gamma_b\|\nu) + W_{b\Gamma_0}(\mu,\gamma_b) - \varepsilon.$$

We must have $\gamma_b \ll \nu$. By the characterization (5.3), we know

$$W_{b\Gamma_0}(\mu,\gamma_b) = bW_{\Gamma_0}(\mu,\gamma_b) \ge bW_{\Gamma_0}(\mu,\tilde\gamma) = W_{b\Gamma_0}(\mu,\tilde\gamma).$$

Thus

$$\begin{aligned} R(\tilde\gamma\|\nu) + W_{b\Gamma_0}(\mu,\tilde\gamma) &\ge \inf_{\gamma \in \mathcal{P}(S)} \{R(\gamma\|\nu) + W_{b\Gamma_0}(\mu,\gamma)\} \\ &= G_{b\Gamma_0}(\mu,\nu) \\ &\ge R(\gamma_b\|\nu) + W_{b\Gamma_0}(\mu,\gamma_b) - \varepsilon \\ &\ge R(\gamma_b\|\nu) + W_{b\Gamma_0}(\mu,\tilde\gamma) - \varepsilon. \end{aligned} \tag{A.2}$$

Since $W_{b\Gamma_0}(\mu,\tilde\gamma)$ is finite we can subtract it on both sides, and get

$$R(\gamma_b\|\nu) \le R(\tilde\gamma\|\nu) + \varepsilon$$

for any $b < \infty$. Then by Lemma 1.4.3(c) in [9] $\{\gamma_b\}_{b\in(0,\infty)}$ is tight. Take a convergent subsequence $\{\gamma_{b_k}\}$, and denote its limit by γ_∞. It is easily checked that $\gamma_\infty \ll \nu$, so $W_{\Gamma_0}(\mu,\gamma_\infty) \ge W_{\Gamma_0}(\mu,\tilde\gamma)$. On the other hand, by (A.2)

$$\begin{aligned} W_{\Gamma_0}(\mu,\gamma_\infty) - W_{\Gamma_0}(\mu,\tilde\gamma) &\le \liminf_{k\to\infty} W_{\Gamma_0}(\mu,\gamma_{b_k}) - W_{\Gamma_0}(\mu,\tilde\gamma) \\ &= \liminf_{k\to\infty} \frac{1}{b_k}(W_{b_k\Gamma_0}(\mu,\gamma_{b_k}) - W_{b_k\Gamma_0}(\mu,\tilde\gamma)) \\ &\le \liminf_{k\to\infty} \frac{1}{b_k}(R(\tilde\gamma\|\nu) - R(\gamma_{b_k}\|\nu) + \varepsilon) \\ &\le \liminf_{k\to\infty} \frac{1}{b_k}(R(\tilde\gamma\|\nu) + \varepsilon) \\ &= 0. \end{aligned}$$

Thus we conclude that $W_{\Gamma_0}(\mu,\gamma_\infty) = W_{\Gamma_0}(\mu,\tilde\gamma)$. By the definition of $\tilde\gamma$ we must have $R(\gamma_\infty\|\nu) \ge R(\tilde\gamma\|\nu)$. Choose k_0 such that $b_{k_0} \ge 1$. Then

$$\liminf_{k\to\infty} (G_{b_k\Gamma_0}(\mu\|\nu) - [R\tilde\gamma\|\nu) + b_k W_{\Gamma_0}(\mu,\tilde\gamma)])$$

$$
\begin{aligned}
&\geq \liminf_{k\to\infty} \left(R(\gamma_{b_k}\|\nu) + b_k W_{\Gamma_0}(\mu,\gamma_{b_k}) - \varepsilon - (R(\tilde{\gamma}\|\nu) + b_k W_{\Gamma_0}(\mu,\tilde{\gamma}))\right)\\
&\geq \liminf_{k\to\infty} (R(\gamma_{b_k}\|\nu) - R(\tilde{\gamma}\|\nu)) + \liminf_{k\to\infty} b_k (W_{\Gamma_0}(\mu,\gamma_{b_k}) - W_{\Gamma_0}(\mu,\tilde{\gamma})) - \varepsilon\\
&\geq (R(\gamma_\infty\|\nu) - R(\tilde{\gamma}\|\nu)) + \liminf_{k\to\infty} (W_{\Gamma_0}(\mu,\gamma_{b_k}) - W_{\Gamma_0}(\mu,\tilde{\gamma})) - \varepsilon\\
&\geq 0 + (W_{\Gamma_0}(\mu,\gamma_\infty) - W_{\Gamma_0}(\mu,\tilde{\gamma})) - \varepsilon\\
&= -\varepsilon,
\end{aligned}
$$

where the fourth inequality is because $R(\gamma_\infty\|\nu) \geq R(\tilde{\gamma}\|\nu)$ and the lower semi-continuity of $W_{\Gamma_0}(\mu,\cdot)$. Since $\varepsilon > 0$ is arbitrary, this establishes (5.2) along the given subsequence. For any other sequence $\{b_k\}_{k\in\mathbb{N}}$ along which $\lim_{k\to\infty} (G_{b_k\Gamma_0}(\mu\|\nu) - [R(\tilde{\gamma}\|\nu) + b_k W_{\Gamma_0}(\mu,\tilde{\gamma})])$ has a limit, we can also take a subsequence from it according to the discussion above. Thus the statement is proved.

The proof of the claimed form for $\tilde{\gamma}$ is as follows. Let θ be any probability measure with $\theta \ll \nu$, and assume that for some i

$$
\theta(\{x_i\}) < \sum_{j\in S_i} \mu(\{y_j\}). \tag{A.3}
$$

Then there exists $j \in S_i$ for which some of the mass is sent to a point x_k with $c(x_k, y_j) > c(x_i, y_k)$. By taking this mass from x_k and assigning it to x_i, while keeping all other assignments the same, we get a strictly lower cost. Thus (A.3) cannot hold for any i at an optimizer, and therefore equality must hold for all i. □

References

[1] S.-I. Amari, R. Karakida and M. Oizumi, Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem. *Inf. Geometry* **1** (2018) 13–37.

[2] L. Ambrosio and N. Gigli, A User's Guide to Optimal Transport. Springer, Berlin, Heidelberg (2013).

[3] G. Bayraksan and D.K. Love, Chapter 1 of *Data-Driven Stochastic Programming Using Phi-Divergences* (2015) 1–19.

[4] J. Blanchet, Y. Kang and K. Murthy, Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* **56** (2016) 10

[5] J. Blanchet and K. Murthy, Quantifying distributional model risk via optimal transport. *SSRN Electr. J.* (2016).

[6] R.I. Bot, S.-M. Grad and G. Wanka, Duality in Vector Optimization. Springer-Verlag, Berlin Heidelberg (2009).

[7] T. Breuer and I. Csiszár, Measuring distribution model risk. *Math. Finance* **26** (2013) 395–411.

[8] K. Chowdhary and P. Dupuis, Distinguishing and integrating aleatoric and epistemic variationi in uncertainty quantification. *ESAIM: M2AN* **47** (2013) 635–662.

[9] P. Dupuis and R.S. Ellis, A Weak Convergence Approach to the Theory of Large Deviations. John Wiley & Sons, New York (1997).

[10] P. Dupuis, M.R. James and I.R. Petersen, Robust properties of risk–sensitive control. *Math. Control Signals Syst.* **13** (2000) 318–332.

[11] P. Dupuis, M.A. Katsoulakis, Y. Pantazis and P. Plechac, Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA J. Uncert. Quantific.* **4** (2016) 80–111.

[12] P. Glasserman and X. Xu, Robust risk measurement and model risk. *Quantit. Finance* **14** (2014) 29–58.

[13] L.P. Hansen and T.J. Sargent, Robust control and model uncertainty. *Am. Econ. Rev.* **91** (2001) 60–66.

[14] S. Kolouri, S. Park, M. Thorpe, D. Slepcev and G. Rohde, Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.* **34** (2017) 43–59.

[15] H. Lam, Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* **41** (2016) 1248–1275.

[16] A.E.B. Lim, J.G. Shanthikumar and T. Watewai, Robust intensity control with multiple levels of model uncertainty and the dual risk-sensitive problem. In *49th IEEE Conference on Decision and Control (CDC)* (2010) 4305–4310.

[17] X. Nguyen, H.J. Wainwright and M.I. Jordan, Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inform. Theory* **56** (2010) 5847–5861.

[18] A. Nilim and L. El Ghaoui, Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.* **53** (2005) 780–798.

[19] I.R. Petersen, M.R. James and P. Dupuis, Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Automatic Control* **45** (2000) 398–412.

[20] S.T. Rachev and L. Rüschendorf, Mass Transportation Problems. *Probability and Its Applications*. Springer-Verlag New York (1998).

[21] R.T. Rockafellar, Convex Analysis. Princeton University Press, Princeton (1970).
[22] W. Rudin, Functional Analysis. McGraw-Hill, New York (1991).
[23] F. Santambrogio, Optimal Transport for Applied Mathematicians. *Progress in Nonlinear Differential Equations and Their Applications*. Birkhauser Basel (2015).
[24] B. Schmitzer and B. Wirth, A framework for Wasserstein-1-type metrics. *J. Convex Anal.* **26** (2019) 353–396.
[25] C. Villani, Optimal Transport: Old and New. Springer-Verlag, Berlin Heidelberg (2009).