Optimal Transport + Gradient Flow
Volume 00, Number 0, Pages 000–000
S 0000-0000(XX)0000-0

AN OVERVIEW ON OPTIMAL TRANSPORT AND ITS APPLICATION TO MODEL FUSION

HANSEUL CHO*

Abstract. In this report, we briefly overview the theory of optimal transport (OT), entropically regularized OT, and the Sinkhorn algorithm, and their application to the deep learning model fusion technique. For the theory of OT, we delve into the derivation of the dual OT problem and the proof of the no-duality-gap result (i.e., strong duality). Through the lens of duality, we can analyze the entropic OT and its dual problem, and derive the Sinkhorn algorithm. We then provide a short survey of convergence results of the Sinkhorn algorithm and its variants. Lastly, we turn our attention to model fusion, which combines the power of several differently trained deep learning models (i.e., neural networks) into a single powerful model. We illustrate OTfusion [Singh and Jaggi, 2020], a method of aggregating several neural networks via optimal transport, and we offer a short discussion of further applications of it.

1. Preliminaries

1.1. **Notation.** Let μ and ν be probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively, i.e., $\int_{\mathcal{X}} \mathrm{d}\mu = \int_{\mathcal{Y}} \mathrm{d}\nu = 1$. We denote the set of transport plans between μ and ν as $\Pi(\mu,\nu) \triangleq \left\{\pi : \text{probability measure on } \mathcal{X}\times\mathcal{Y} \;\middle|\; \int_{\mathcal{Y}} \mathrm{d}\pi = \mu, \int_{\mathcal{X}} \mathrm{d}\pi = \nu \right\}$. A push-forward of a measure μ by a mapping f is denoted and defined as $f_{\#}\mu \triangleq \mu \circ f^{-1}$. Given a normed vector space $\mathcal{E}$, we denote $\mathcal{E}^* \triangleq \{L : \mathcal{E} \to \mathbb{R} \mid \text{linear and continuous}\}$ be its dual space. Given a function $\phi : \mathcal{E} \to \mathbb{R}\cup\{+\infty\}$ with $\phi \not\equiv \infty$, the Legendre-Fenchel transform of ϕ is a function ϕ^* defined on $\mathcal{E}$ by the formula $\phi^*(z^*) \triangleq \sup_{z\in\mathcal{E}} \langle z^*, z\rangle - \phi(z)$.

1.2. **Optimal Transport Problem.** Let μ, ν, and a cost function $c(x,y) \geq 0$ defined on $\mathcal{X}\times\mathcal{Y}$ be given. An *optimal transport (OT) problem*, also called *Monge-Kantorovich problem*, is described as follows:

$$\mathrm{OT}(\mu,\nu;c) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y),$$

or equivalently,

$$\text{(OT)}\qquad \begin{aligned} \mathrm{OT}(\mu,\nu;c) := \inf & \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y) \\ \text{s.t.} & \int_{\mathcal{Y}} \mathrm{d}\pi = \mu,\ \int_{\mathcal{X}} \mathrm{d}\pi = \nu,\ \pi \geq 0. \end{aligned}$$

In general, this is an infinite-dimensional linear programming (LP) problem. In a discrete setting, however, it reduces to a finite-dimensional LP. Consider discrete

Submitted date: Dec 2023.

*Kim Jaechul Graduate school of AI, KAIST. Student ID: 20235665.

probability measures $\mu = \sum_{i=1}^{m} \mu_i \delta_{\{x^{(i)}\}}$ and $\nu = \sum_{j=1}^{n} \nu_j \delta_{\{y^{(j)}\}}$ and corresponding cost values (c_{ij}). Then,

$$\text{OT}(\mu, \nu; c) = \min \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \pi_{ij} \quad \text{s.t.} \quad \sum_{j=1}^{n} \pi_{ij} = \mu_i, \; \sum_{i=1}^{m} \pi_{ij} = \nu_i, \; \pi_{ij} \geq 0. \tag{1.1}$$

Equipped with a natural topology (called *weak-$*$ topology*) for measures, the set of probability measures is guaranteed to be compact, thereby the existence of an optimal solution π^* of $\text{OT}(\mu, \nu; c)$ (called *optimal transport plan*) is ensured.

1.3. **Wasserstein Distance and Wasserstein Barycenters.** If the spaces $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces of the same dimension, i.e., $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, and the cost function is composed with a metric D over $\mathbb{R}^d$, OT naturally induces a distance between probability measures. Specifically, the *p-Wasserstein distance* is defined as

$$\mathcal{W}_p(\mu, \nu) \triangleq \text{OT}(\mu, \nu; D(\cdot, \cdot)^p)^{1/p}.$$

Given the definition of distance between probability measures, we can also establish a notion of average between measures. In particular, a *Wasserstein barycenter* of given K measures $\mu_1, \ldots, \mu_K$, induced by $\mathcal{W}_p$, is also a probability measure, defined as

$$\mathcal{B}_p(\mu_1, \ldots, \mu_K) = \arg\min_{\nu} \sum_{k=1}^{K} \omega_k \mathcal{W}_p(\mu_k, \nu)^p,$$

where the weights ω_k's are known.

2. Duality of Optimal Transport

2.1. **Derivation of Dual Problem.** In this section, we derive the dual problem with respect to (OT) as a primal problem.

Consider functions $\phi(x)$, $\psi(y)$, and $\rho(x, y)$ defined on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$ where $\rho \geq 0$. For a measure π defined on $\mathcal{X} \times \mathcal{Y}$, define

$$P_{\phi,\psi,\rho}(\pi) \triangleq \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) - \int_{\mathcal{X} \times \mathcal{Y}} (\phi(x) + \psi(y) + \rho(x, y)) \, \mathrm{d}\pi(x, y).$$

Then if we let

$$F(\phi, \psi, \rho) \triangleq \inf_{\pi} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) + P_{\phi,\psi,\rho}(\pi),$$

we have $F(\phi, \psi, \rho) \leq \text{OT}(\mu, \nu)$. Thus,

$$D(\mu, \nu; c) \triangleq \sup_{\phi, \psi, \rho \colon \rho \geq 0} F(\phi, \psi, \rho) \leq \text{OT}(\mu, \nu; c).$$

Note that

$$F(\phi, \psi, \rho) = \begin{cases} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) & \text{if } c \equiv \phi + \psi + \rho, \\ -\infty & \text{otherwise.} \end{cases}$$

Thus, we arrive at the dual problem of (OT) as follows:

(OT-dual)

$$D(\mu, \nu; c) = \sup_{\phi, \psi} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) \quad \text{s.t. } \phi(x) + \psi(y) \leq c(x, y).$$

Naturally, as noted before, the following relationship holds: (OT-dual)$\leq$(OT). This relationship is called **weak duality**. At first glance, there seems to be a certain amount of *duality gap* between (OT-dual) and (OT). Surprisingly, as will be shown in the subsequent section, the duality gap is in fact zero under some assumptions.

2.2. **No Duality Gap.** For a finite-dimensional OT problem, it has been well-studied that there is no duality gap, due to the theory of LP. To study (possibly) infinite-dimensional problems, we need to bring some tools from functional analysis. To make the discussion succinct, let us introduce assumptions that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and the cost function c is continuous. Also, we will consider the following problem with stronger constraints:

$$\widetilde{D}(\mu, \nu; c) = \sup_{\phi,\psi} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) \quad \text{s.t. } \phi + \psi \leq c,\, \phi \in C(\mathcal{X}),\, \psi \in C(\mathcal{Y}).$$

Note that $\widetilde{D}(\mu, \nu; c) \leq D(\mu, \nu; c) \leq OT(\mu, \nu; c)$, where the second inequality is due to weak duality.

Lemma 2.1 (Fenchel-Rockafellar duality)**.** *Let $\mathcal{E}$ be a normed vector space, $\mathcal{E}^*$ its topological dual space, and. Θ, Ξ two convex functions on $\mathcal{E}$ with values in $\mathbb{R} \cup \{\infty\}$. Let Θ^*, Ξ^* be the Legendre-Fenchel transforms of Θ, Ξ respectively. Assume that there exists $z_0 \in \mathcal{E}$ such that*

$$\Theta(z_0) < \infty, \quad \Xi(z_0) < \infty, \quad \textit{and } \Theta \textit{ is continuous at } z_0.$$

Then,

$$\inf_{z \in \mathcal{E}} \left[\Theta(z) + \Xi(z)\right] = \max_{z^* \in \mathcal{E}^*} \left[-\Theta^*(-z^*) - \Xi^*(z^*)\right].$$

With this duality theorem, we can prove the following strong duality result. In contrast to the lecture notes, the following proof will contain as many details as possible. We remark that the plot of the proof is almost taken (but not exactly copied) from Villani [2021].

Theorem 2.2. *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact normed spaces and the cost function c is continuous. Then,*

$$\widetilde{D}(\mu, \nu; c) = D(\mu, \nu; c) = OT(\mu, \nu; c).$$

As a result, the duality gap is zero.

Proof. Let $\mathcal{E} = C(\mathcal{X} \times \mathcal{Y})$ be the set of all (bounded) continuous real-valued functions on $\mathcal{X} \times \mathcal{Y}$. Because of compactness, by the Riesz representation theorem, its topological dual space $\mathcal{E}^*$ is the space of (regular) Radon measures on $\mathcal{X} \times \mathcal{Y}$.

To apply Fenchel-Rockafellar duality, we introduce functions of functions: for $u \in \mathcal{E}$,

$$\Theta(u) = \begin{cases} 0 & \text{if } u(x,y) \geq -c(x,y), \\ \infty & \text{otherwise,} \end{cases}$$

$$\Xi(u) = \begin{cases} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) & \text{if } u(x,y) = \phi(x) + \psi(y) \text{ for some } \phi \in C(\mathcal{X}),\, \psi \in C(\mathcal{Y}), \\ \infty & \text{otherwise.} \end{cases}$$

Note that Ξ is a well-defined function, since μ and ν are probability measures: if $\phi(x)+\psi(y)=\tilde{\phi}(x)+\tilde{\psi}(y)$, $s:=\phi-\tilde{\phi}=\tilde{\psi}-\psi$ must be a constant function. Hence,

$$\int_{\mathcal{X}}\left(\phi-\tilde{\phi}\right)\mathrm{d}\mu=\int_{\mathcal{X}}s\mathrm{d}\mu=s=\int_{\mathcal{Y}}s\mathrm{d}\nu=\int_{\mathcal{Y}}\left(\tilde{\psi}-\psi\right)\mathrm{d}\nu,$$

and thus $\int_{\mathcal{X}}\phi\mathrm{d}\mu+\int_{\mathcal{Y}}\psi\mathrm{d}\nu=\int_{\mathcal{X}}\tilde{\phi}\mathrm{d}\mu+\int_{\mathcal{Y}}\tilde{\psi}\mathrm{d}\nu$.

Also, note that Θ and Ξ are convex. The convexity of Θ naturally follows from the fact that its domain is a convex subset of $\mathcal{E}$. On the other hand, the convexity of Ξ can be proved by showing that Ξ is linear on its domain.

Lastly, the assumptions of Fenchel-Rockafellar duality hold with $z_0\equiv 1$. Since $c\geq 0$, $\Theta(z_0)=0<\infty$. Also, Θ is constant on its domain, so it is continuous at z_0. Besides, we can check $\Xi(z_0)=1<\infty$. Therefore, now we can apply Fenchel-Rockafellar duality:

$$\inf_{u\in\mathcal{E}}\left[\Theta(u)+\Xi(u)\right]=\sup_{\pi\in\mathcal{E}^*}\left[-\Theta^*(-\pi)-\Xi^*(\pi)\right].$$

Let us reckon both sides. For the left-hand side,

$$\begin{aligned}
&\inf_{u\in\mathcal{E}}\left[\Theta(u)+\Xi(u)\right]\\
&=\inf_{\phi\in C(\mathcal{X}),\,\psi\in C(\mathcal{Y})}\left\{\int_{\mathcal{X}}\phi(x)\mathrm{d}\mu(x)+\int_{\mathcal{Y}}\psi(y)\mathrm{d}\nu(y);\quad \phi(x)+\psi(y)\geq -c(x,y)\right\}\\
&=-\sup_{\phi\in C(\mathcal{X}),\,\psi\in C(\mathcal{Y})}\left\{\int_{\mathcal{X}}\phi(x)\mathrm{d}\mu(x)+\int_{\mathcal{Y}}\psi(y)\mathrm{d}\nu(y);\quad \phi(x)+\psi(y)\leq c(x,y)\right\}\\
&=-\widetilde{D}(x,y;c).
\end{aligned}$$

To compute the right-hand side, we need the Fenchel-Legendre transform of Θ and Ξ. First,

$$\Theta^*(-\pi)=\sup_{u\in\mathcal{E},u\geq -c}\int u(-\mathrm{d}\pi)=\sup_{u\in\mathcal{E},u\leq c}\int u\mathrm{d}\pi=\begin{cases}\int c\,\mathrm{d}\pi & \text{if } \pi\geq 0,\\ \infty & \text{otherwise,}\end{cases}$$

because if π is not non-negative, then there exists a function $v\in C(\mathcal{X}\times\mathcal{Y})$ such that $v\leq 0$ and $\int v\mathrm{d}\pi>0$; scaling v yields the supremum infinity. With a similar logic, one can show that

$$\begin{aligned}
\Xi^*(\pi)&=\sup_{u\in\mathcal{E},u(x,y)=\phi(x)+\psi(y)}\int u\mathrm{d}\pi-\left(\int\phi\mathrm{d}\mu+\int\psi\mathrm{d}\nu\right)\\
&=\begin{cases}0 & \text{if } \pi \text{ has marginals } \mu \text{ and } \nu,\\ \infty & \text{otherwise.}\end{cases}
\end{aligned}$$

We therefore get

$$\begin{aligned}
\sup_{\pi\in\mathcal{E}^*}\left[-\Theta^*(-\pi)-\Xi^*(\pi)\right]&=\sup_{\pi\in\Pi(\mu,\nu)}\left(-\int c\,\mathrm{d}\pi\right)\\
&=-\inf_{\pi\in\Pi(\mu,\nu)}\int c\,\mathrm{d}\pi=-\operatorname{OT}(\mu,\nu;c).
\end{aligned}$$

This proves the desired result because of weak duality. □

3. Fast Optimal Transport

Now we turn our attention to the efficient computation of (OT). In fact, we are interested in its finite-dimensional version (1.1).

It is known that the computational complexity of solving LP is polynomial in the number of variables, because of inevitable matrix-matrix products [Cohen et al., 2021, van den Brand, 2020]. This makes solving the OT problem computationally infeasible if the size (number of variables/constraints) of the LP is huge. We will particularly get closer to the entropic regularization technique, proposed and studied by Marco Cuturi, Gabriel Peyré, and many other researchers [Cuturi, 2013, Peyré et al., 2017]. With this "slightly perturbed" formulation of the OT problem, we can have exponentially fast convergence to the (yet approximate, but unique!) solution thanks to the power of the Sinkhorn algorithm (or its variants).

3.1. **Entropic Regularization.**

Let us delve into the entropically regularized optimal transport problem. Recall that we are considering the discrete probability measures $\mu = \sum_{i=1}^m \mu_i \delta_{\{\boldsymbol{x}_i\}}$ and $\nu = \sum_{j=1}^n \nu_j \delta_{\{\boldsymbol{y}_j\}}$ and corresponding cost values (c_{ij}). For a small hyperparameter $\epsilon > 0$, the entropic OT problem is stated as

$$\text{(EOT)} \qquad \mathrm{OT}_\epsilon(\mu,\nu;c) := \inf_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^m \sum_{j=1}^n c_{ij}\pi_{ij} - \epsilon\, S(\pi),$$

where $S(\pi) = -\sum_{i=1}^m \sum_{j=1}^n \pi_{ij}(\log \pi_{ij} - 1)$, with the convention that $0 \log 0 = 0$. This problem is no longer an LP but a convex problem. In particular, the objective function is an ϵ-strongly convex function because $S(\pi)$ is 1-strongly concave, so the problem has a unique solution.

3.2. **Dual of Entropic OT.**

To study the entropic OT, which is a slightly perturbed problem of the original OT, it is natural to study its dual. Let us first derive the dual entropic OT problem.

Writing the Lagrangian, for $\phi_i, \psi_j \in \mathbb{R}$ and $\rho_{ij} \in [0,\infty)$ $(i \in [m], j \in [n])$,

$$\begin{aligned}\mathcal{L}(\pi,\phi,\psi,\rho) &:= \sum_{i=1}^m \sum_{j=1}^n (c_{ij} - \rho_{ij})\pi_{ij} + \epsilon S(\pi) + \sum_{i=1}^m \phi_i \left(\mu_i - \sum_{j=1}^n \pi_{ij}\right) + \sum_{j=1}^n \psi_j \left(\nu_j - \sum_{i=1}^m \pi_{ij}\right) \\ &= \sum_{i=1}^m \sum_{j=1}^n \{c_{ij} - \phi_i - \psi_j - \rho_{ij} + \epsilon(\log \pi_{ij} - 1)\}\, \pi_{ij} + \sum_{i=1}^m \phi_i\mu_i + \sum_{j=1}^n \psi_j\nu_j.\end{aligned}$$

Note that the function $h(s) = (C + \epsilon(\log s - 1))s$ has the unique minimum $-\epsilon \exp(-C/\epsilon)$ at $s = \exp(-C/\epsilon)$. Using this fact, we can compute the Lagrangian dual

$$\begin{aligned}g(\phi,\psi,\rho) &= \min_\pi \mathcal{L}(\pi,\phi,\psi,\rho) \\ &= -\epsilon \sum_{i=1}^m \sum_{j=1}^n \exp\left(-\frac{c_{ij} - \phi_i - \psi_j - \rho_{ij}}{\epsilon}\right) + \sum_{i=1}^m \phi_i\mu_i + \sum_{j=1}^n \psi_j\nu_j\end{aligned}$$

We finally get the dual problem by maximizing g: $D_\epsilon(\mu,\nu) := \sup_{\phi,\psi,\rho:\rho\geq 0} g(\phi,\psi,\rho)$, and thus

(EOT-dual)

$$D_\epsilon(\mu,\nu;c) = \sup_{\phi,\psi}\left\{ G(\phi,\psi) := \sum_{i=1}^m \phi_i\mu_i + \sum_{j=1}^n \psi_j\nu_j - \epsilon\sum_{i=1}^m\sum_{j=1}^n \exp\left(-\frac{c_{ij}-\phi_i-\psi_j}{\epsilon}\right)\right\},$$

since the optimal ρ is zero among $\rho \geq 0$. Note that, it is easy to verify that $D_\epsilon(\mu,\nu;c) \leq \mathrm{OT}_\epsilon(\mu,\nu;c)$, since $\sup_{\phi,\psi,\rho:\rho\geq 0}\inf_\pi \mathcal{L}(\pi,\phi,\psi,\rho) \leq \inf_\pi \sup_{\phi,\psi,\rho:\rho\geq 0}\mathcal{L}(\pi,\phi,\psi,\rho)$ (weak duality).

To seek a dual optimum $(\phi^\epsilon,\psi^\epsilon)$, we now can use partial derivatives: where $G(\phi,\psi)$ is defined in Equation (EOT-dual),

$$\begin{aligned}\frac{\partial}{\partial\phi_i}G(\phi,\psi) &= \mu_i - \sum_{j=1}^n \exp\left(-\frac{c_{ij}-\phi_i-\psi_j}{\epsilon}\right) = 0,\\ \frac{\partial}{\partial\psi_j}G(\phi,\psi) &= \nu_j - \sum_{i=1}^m \exp\left(-\frac{c_{ij}-\phi_i-\psi_j}{\epsilon}\right) = 0.\end{aligned}\tag{3.1}$$

Thus, $(\phi^\epsilon,\psi^\epsilon)$ is the dual optimum if and only if

$$\pi^\epsilon := \left(\exp\left(-\frac{1}{\epsilon}\left[c_{ij}-\phi_i^\epsilon-\psi_j^\epsilon\right]\right)\right)_{ij}\tag{3.2}$$

belongs to $\Pi(\mu,\nu)$. Interestingly, if we plug in $\pi = \pi^\epsilon$ to the objective function of the problem (EOT), we get

$$\sum_{i=1}^m\sum_{j=1}^n c_{ij}\pi_{ij}^\epsilon - \epsilon\, S(\pi^\epsilon) = G(\phi^\epsilon,\psi^\epsilon).$$

In this case, the primal and the dual match. Thus, if we have a pair of dual optimal variables, they directly determine the primal optimal solution π^ϵ. This is the main motivation of the Sinkhorn algorithm.

One last remark is that the dual optimal point is not necessarily unique; instead, they are unique up to additive constants. Namely, if $(\phi^\epsilon,\psi^\epsilon)$ is a dual optimal point, then $(\tilde{\phi}^\epsilon,\tilde{\psi}^\epsilon) := (\phi^\epsilon + a, \psi^\epsilon - a)$ is also dual optimal, for any constant vector a.

3.3. **Sinkhorn Algorithm.**

To illustrate the Sinkhorn algorithm for approximately solving the entropic OT problem, recall that we aim to find (ϕ,ψ) satisfying (3.1). Instead of finding (ϕ,ψ), consider (u,v) such that $u_i = \exp(\phi_i/\epsilon)$ and $v_j = \exp(\psi_j/\epsilon)$. Then, Equation (3.1) can be written as

$$\mu_i = u_i\sum_{j=1}^n \exp\left(-\frac{c_{ij}}{\epsilon}\right)v_j, \quad \nu_j = v_j\sum_{i=1}^m \exp\left(-\frac{c_{ij}}{\epsilon}\right)u_i.$$

If we define the matrix $K = [K_{ij}] := \left[\exp\left(-\frac{c_{ij}}{\epsilon}\right)\right]$, the equations above can be rewritten as the following matrix-vector calculations, where the division is calculated elementwise.

$$u = \frac{\mu}{Kv}, \quad v = \frac{\nu}{K^\top u}.$$

Thus, the solution (u, v) is the fixed point of the mapping $(u, v) \mapsto \left(\frac{\mu}{Kv}, \frac{\nu}{K^\top u}\right)$. To find such a point, theoretically, any kind of fixed-point algorithm can be applied. The Sinkhorn algorithm proposes one of the simplest methods of such.

$$\text{(Sinkhorn)} \qquad u^{(\ell+1)} = \frac{\mu}{Kv^{(\ell)}}, \quad v^{(\ell+1)} = \frac{\nu}{K^\top u^{(\ell+1)}}$$

In the rest of the section, we will briefly look into the global convergence of the Sinkhorn algorithm. The following theorem is excerpted and simplified from Peyré et al. [2017] (see references therein).

Theorem 3.1 (Theorem 4.2, Peyré et al. [2017])**.** *Denote $\pi^{(\ell)}$ as $\pi_{ij}^{(\ell)} \triangleq u_i^{(\ell)} K_{ij} v_j^{(\ell)}$. Then, one has*

$$\left\| \log(\pi^{(\ell)}) - \log(\pi^\epsilon) \right\|_\infty = \mathcal{O}\left(\left(1 - \frac{2}{\sqrt{\eta(K)} + 1} \right)^{2\ell} \right),$$

$$\textit{where } \eta(K) = \max_{i,j,k,l} \frac{K_{ik}K_{jl}}{K_{jk}K_{il}} \geq 1.$$

The theorem states that the Sinkhorn algorithm converges exponentially fast to the unique minimum of the primal problem (EOT). One caveat is that it does not necessarily imply the exponential convergence to a solution of the original OT; only the sublinear (slower than exponential) convergence is guaranteed with small non-decaying values of ϵ [Altschuler et al., 2017, Dvurechensky et al., 2018]. Adaptive scheduling (e.g., adaptively halving ϵ) can make the Sinkhorn algorithm achieve exponential convergence to the original OT [Chen et al., 2023].

4. Model Fusion via Discrete Optimal Transport

In this section, we introduce the concept of aggregation of deep learning models and recent advances in this field using discrete optimal transport.

4.1. **Model Fusion.**

In practice, there are several well-trained machine learning models, even for each of the tasks. Can we empower our prediction capability by combining the capabilities of several models? If so, how?

Ensemble methods are the most common approach to do such a task, which combine the *outputs* of different models for better test-time performance and robustness of the prediction. Despite the practical benefits of ensemble methods, they require storing K different models and running each of them, thereby the memory and computation cost scale with the number of models. If each model is huge (and typically larger models perform better), this quickly becomes infeasible to apply.

A clever detour is the *model fusion*: combine the *model weights/parameters* into a single model. For the sake of simplicity, assume that the models have similar architectures (e.g., having the same number of hidden layers, or the same 'depth'). The simplest way of combining several models is direct weight average; we call it VanillaFusion. Although it is easy to apply, it has two main weaknesses. First, it is applicable only when the model architectures are identical (though each of the weight values may be different). Second, in practice, the prediction performance is poor. This is mainly because there is no one-to-one correspondence of weights between two distinct models, especially for neural networks, even if their prediction performance would be similar.

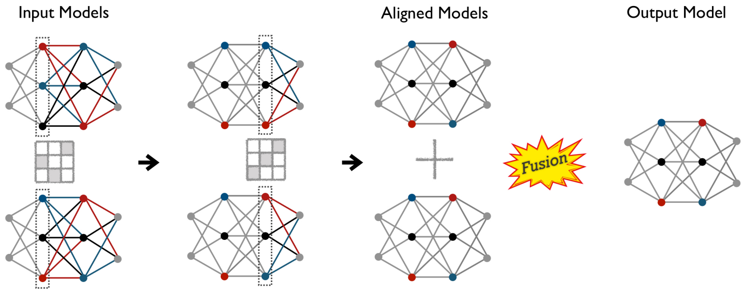

FIGURE 1. **Model Fusion Procedure of OTFUSION.**

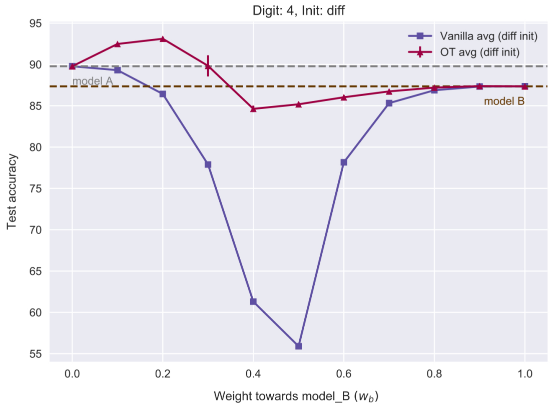

FIGURE 2. **OTFUSION v.s. VANILLAFUSION: One-shot skill transfer performance** when the models A and B are fused in varying proportions.

4.2. **OTFUSION: Model Fusion via Optimal Transport.** One can borrow the power of optimal transport to cleverly fuse several differently trained neural networks. For simplicity, we consider the fusion of two models, A and B, and we assume that the models have the same number of hidden layers (same depth), but each of the hidden layers might have a different number of neurons (different width). Then, we can apply OT to "align" the neurons and weights of different neural networks, and then the networks can be directly averaged. Note that this method, called OTFUSION [Singh and Jaggi, 2020], works layer-wise: the model averaging is applied after performing (soft) alignment via OT layer by layer.

The crux of the mechanism hidden in OTFUSION is that we would like to aggregate neurons or weights that have similar roles. If we average two arbitrarily different neurons, their effect would be averaged out. To avoid this phenomenon, appropriate alignment between neurons in different networks is important. To this end, fix a layer ℓ and consider a transportation cost between neurons or weights of that layer in each of models A and B. Fix model B, and consider the scenario that we re-align the weights in model A. If we carefully design the way of transporting the weights of the layer ℓ in model A to those in model B, we have the same size of weight matrices, and therefore they can be averaged. The detailed mechanism is illustrated in Singh and Jaggi [2020].

The empirical result is promising. To simulate the model fusion, Singh and Jaggi [2020] train the two fully-connected neural networks A and B with identical structure on a couple of distinct chunks (data chunk A for model A, chunk B for model B) of the MNIST digit classification dataset. Then, by fusing two models with VanillaFusion and OTfusion, they observe how well the skill transfers from model A to B. The result is shown in Figure 2. While the performance of vanilla-fused models is degraded, the performance degradation is much less for OTfusion. Somewhat surprisingly, some OT-fused models *outperform* the individual models A and B. This suggests that a combined model through OT may have multiple capabilities that each of the individual models possesses.

Recently, Imfeld et al. [2023] carefully applied the techniques from OTfusion to aggregate renowned Transformer models [Vaswani et al., 2017]. In their application, the Sinkhorn algorithm is the main workhorse to achieve promising performances, which is different from the original OTfusion paper, where they utilized an exact OT solver.

5. Conclusion

We discussed various aspects related to optimal transport, including duality theory, entropic regularization, and the Sinkhorn algorithm, and their application to model fusion. Specifically regarding model fusion, we did not delve deeply into the technical intricacies of OTfusion because there is no theoretical assurance that it represents an optimal weight transportation.

Numerous avenues for future research in model fusion plus optimal transport exist. First, there is a lack of theoretical understanding regarding the optimal or most efficient fusion methods and their underlying mechanisms. Exploring more effective ways to fuse models using optimal transport would be intriguing. Additionally, a more profound theoretical exploration of model fusion is necessary. Second, the applicability of OTfusion is restricted to cases where two models share the same architecture and depth. Investigating methods for fusing models not only in the weight space but also in the function space could be valuable, allowing the fusion of models with entirely different structures. Third, creating a multi-lingual large language model by fusing several uni-lingual (i.e., single-language) language models without any fine-tuning could be a promising area of study, with optimal transport serving as a crucial tool in the design process.

References

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

Jingbang Chen, Li Chen, Yang P Liu, Richard Peng, and Arvind Ramaswami. Exponential convergence of sinkhorn under regularization scheduling. In *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA23)*, pages 180–188. SIAM, 2023.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)*, 68(1):1–39, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

Moritz Imfeld, Jacopo Graldi, Marco Giordano, Thomas Hofmann, Sotiris Anagnostidis, and Sidak Pal Singh. Transformer fusion with optimal transport. *arXiv preprint arXiv:2310.05719*, 2023.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017-86), 2017.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020.

Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 259–278. SIAM, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Email address: jhs4015@kaist.ac.kr