Supplementary material for the paper "Optimal Transport for Domain Adaptation"

Nicolas Courty, *Member, IEEE*, Remi Flamary, Devis Tuia, *Senior Member, IEEE*,
Alain Rakotomamonjy, *Member, IEEE*

I. Optimal transport for affine transformations

Our main hypothesis is that the transformation $\mathbf{T}(\cdot)$ preserves the conditional distribution

$$\mathbf{P}_s(y|\mathbf{x}^s) = \mathbf{P}_t(y|\mathbf{T}(\mathbf{x}^s))$$

Hence, if we are able to estimate a transformation $\mathbf{T}_0(\cdot)$ such that for all $\mathbf{x}$, $\mathbf{T}_0(\mathbf{x}) = \mathbf{T}(\mathbf{x})$, we can compute the exact conditional probability distribution of any example $\mathbf{x}$ in the target domain through

$$\mathbf{P}_t(\mathbf{y}|\mathbf{x}) = \mathbf{P}_s(\mathbf{y}|\mathbf{T}_0^{-1}(\mathbf{x}))$$

supposing that T_0 is invertible. We can remark that the ability of our approach to recover the conditional probability of an example essentially depends on how well the condition $\mathbf{T}_0(\mathbf{x}) = \mathbf{T}(\mathbf{x})\ \forall \mathbf{x}$ is respected. In the following, we prove that, for empirical distributions and under some mild conditions (positive definite $\mathbf{A}$), the estimated transportation map $\mathbf{T}_0$ obtained by minimizing the ℓ_2 ground metric does recover exactly the true transformation $\mathbf{T}$.

We want to show that, if $\mu^t(\mathbf{x}) = \mu^s(\mathbf{A}\mathbf{x}+\mathbf{b})$ where $\mathbf{A} \in \mathcal{S}^+$ and $\mathbf{b} \in \mathbb{R}^d$, then the estimated optimal transportation plan $\mathbf{T}_0(\mathbf{x}) = \mathbf{A}\mathbf{x} + b$. This boils down to proving the following theorem.

Theorem 1.1: Let μ^s and μ^t be two discrete distributions each composed of n diracs as defined in Equation (1) in the paper. If the following conditions hold

1) The source samples in μ^s are $\mathbf{x}_i^s \in \mathbb{R}^d, \forall i \in 1, \ldots, n$ such that $\mathbf{x}_i^s \neq \mathbf{x}_j^s$ if $i \neq j$.
2) All weights in the source and target distributions are equal to $\frac{1}{n}$.
3) The target samples are defined as $\mathbf{x}_i^t = \mathbf{A}\mathbf{x}_i^s + \mathbf{b}$ *i.e.* an affine tranformation of the source samples.
4) $\mathbf{b} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathcal{S}^+$ is a strictly positive definite matrix.
5) The cost function $c(\mathbf{x}^s, \mathbf{x}^t) = \|\mathbf{x}^s - \mathbf{x}^t\|^2$.

then the solution $\mathbf{T}_0$ of the optimal transport problem gives $\mathbf{T}_0(\mathbf{x}_i^s) = \mathbf{A}\mathbf{x}_i^s + \mathbf{b} = \mathbf{x}_i^t \quad \forall i \in 1, \ldots, n$.

First, note that the OT problem (and in particular the sum and positivity constraints) forces the solution $\boldsymbol{\gamma}$ to be a doubly stochastic matrix weighted by $\frac{1}{n}$. Since the objective function is linear, the solution will be a vertex of the set of doubly stochastic matrices which is a permutation matrix as stated by the Birkhoff-von Neumann Theorem [1].

To prove Theorem 1.1 we first show that, given the above conditions, if the solution of the discrete OT problem is the identity matrix $\boldsymbol{\gamma} = \frac{1}{n}\mathbf{I}_n$, then the corresponding transportation $\mathbf{T}_0(\mathbf{x})$ gives $\mathbf{T}_0(\mathbf{x}_i^s) = \mathbf{A}\mathbf{x}_i^s + \mathbf{b} = \mathbf{x}_i^t \quad \forall i \in 1, \ldots, n$. This property results from the interpolation formula in Equation (14), which states that an interpolation along the Wasserstein manifold from the source to target distribution with $t \in [0, 1]$ is defined as:

$$\hat{\mu} = \sum_{i,j} \gamma_0(i,j)\delta_{(1-t)\mathbf{x}_i^s + t\mathbf{x}_j^t}.$$

When $\boldsymbol{\gamma} = \frac{1}{n}\mathbf{I}_n$, it leads to

$$\hat{\mu} = \frac{1}{n}\sum_{i} \delta_{(1-t)\mathbf{x}_i^s + t\mathbf{x}_i^t}.$$

This equation shows that the mass from each sample $\mathbf{x}_i^s$ is transported without any splitting to its corresponding target sample $\mathbf{x}_i^t$ hence $T_0(\mathbf{x}_i^s) = \mathbf{x}_i^t$. Indeed, we have

$$\begin{aligned} (1-t)\mathbf{x}_i^s + t\mathbf{x}_i^t = & \quad (1-t)\mathbf{x}_i^s + t(\mathbf{A}\mathbf{x}_i^s + b) \\ = & \quad (1 - t - t\mathbf{A})\mathbf{x}_i^s + tb, \end{aligned} \tag{1}$$

which for $t = 1$ gives $T_0(\mathbf{x}_i^s) = \mathbf{x}_i^t$. Hence, the OT solution $\boldsymbol{\gamma} = \frac{1}{n}\mathbf{I}_n$ yields an exact reconstruction of the transformation on the samples and $\mathbf{T}_0(\mathbf{x}_i^s) = \mathbf{A}\mathbf{x}_i^s + \mathbf{b} = \mathbf{x}_i^t \quad \forall i \in 1, \ldots, n$.

Consequently, to prove Theorem 1.1 we just need to prove that the OT solution $\boldsymbol{\gamma}$ is equal to $\frac{1}{n}\mathbf{I}_n$ given the above conditions. This is done in the following for some particular cases of affine transformations and for the general case where $\mathbf{A} \in \mathcal{S}^+$ and $\mathbf{b} \in \mathbb{R}^d$.

For better readability we will denote a sample in the source domain $\mathbf{x}_i^s$ as $\mathbf{x}_i$, whereas a sample in the target domain is $\mathbf{x}_i^t$.

A. Case with no transformation ($\mathbf{A} = \mathbf{I}_d$ and $\mathbf{b} = 0$, Figure 1)

In this case, there is no displacement.

$$\boldsymbol{\gamma}^* = \underset{\boldsymbol{\gamma} \in \mathcal{B}}{\operatorname{argmin}} \ \langle \boldsymbol{\gamma}, \mathbf{C}_0 \rangle_F \tag{2}$$

The solution is obviously $\boldsymbol{\gamma} = \frac{1}{n}\mathbf{I}_n$, since it does not violate the constraints and is the only possible solution with a 0 loss (any mass not on the diagonal of $\boldsymbol{\gamma}$ would imply an increase in the loss).

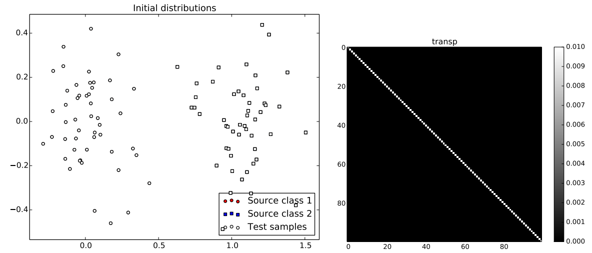

Fig. 1: Optimal transport in the affine transformation 1: case with no transformation

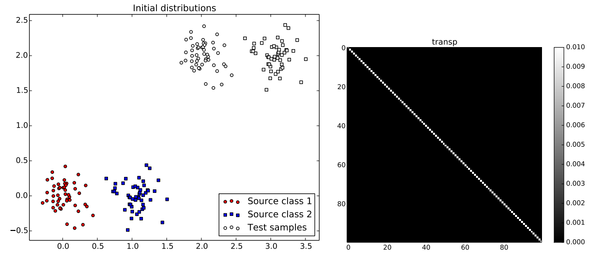

Fig. 2: Optimal transport in the affine transformation 2: case with translation

B. *Case with translation* ($\mathbf{A} = \mathbf{I}_d$ *and* $\mathbf{b} \in \mathbb{R}^d$, *Figure 2)*

In this case, the metric matrix becomes $\mathbf{C}$ such that

$$\begin{aligned} C(i,j) &= \|\mathbf{x}_i - \mathbf{x}_j - \mathbf{b}\|^2 \\ &= \|\mathbf{x}_i - \mathbf{x}_j\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{b}^\top(\mathbf{x}_i - \mathbf{x}_j). \end{aligned}$$

In this case the solution $\boldsymbol{\gamma} = \mathbf{I}$. The corresponding permutation $perm_{\mathbf{I}}$ leads to the loss

$$J(\mathbf{I}) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_i - \mathbf{b}\|^2 = \|\mathbf{b}\|^2.$$

In the general case, the loss for a permutation $perm$ can be expressed as

$$\begin{aligned} J(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_{perm(i)} - \mathbf{b}\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_{perm(i)}\|^2 + \\ &\qquad \|\mathbf{b}\|^2 - \frac{2}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{x}_{perm(i)})^\top \mathbf{b} \\ &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_{perm(i)}\|^2 + \|\mathbf{b}\|^2 - \frac{2}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{x}_i)^\top \mathbf{b} \\ &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_{perm(i)}\|^2 + \|\mathbf{b}\|^2 \\ &= \|\mathbf{b}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_{perm(i)}\|^2 \end{aligned}$$

Since $\mathbf{x}_i \neq \mathbf{x}_j$, if $i \neq j$, for any permutation term $perm(i) \neq i$ then $\|\mathbf{x}_i - \mathbf{x}_{perm(i)}\|^2 > 0$. This means that any permutation that is not the identity permutation will have a loss strictly larger that the loss of the identity permutation. Therefore, the solution of the optimization problem is the identity permutation.

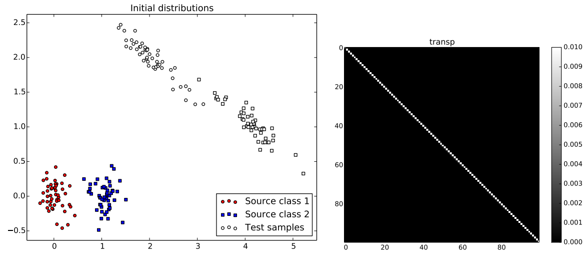

Fig. 3: Optimal transport in the affine transformation 3: general case

This shows that a translation of the data does not change the solution of the optimal transport.

C. *General case* ($\mathbf{A} \in \mathcal{S}^+$ *and* $\mathbf{b} \in \mathbb{R}^d$, *Figure 3)*

In this case, we will set $\mathbf{b} = 0$ since, as proven previously, a translation does not change the optimal transport solution.

The loss can be expressed as

$$\begin{aligned} J(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{x}_{perm(i)}\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 + \|A\mathbf{x}_{perm(i)}\|^2 - 2\mathbf{x}_i^\top \mathbf{A}\mathbf{x}_{perm(i)} \\ &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 + \|A\mathbf{x}_i\|^2 - \frac{2}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{A}\mathbf{x}_{perm(i)}. \end{aligned}$$

The solution of the optimization will then be the permutation that maximizes the term $\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{A}\mathbf{x}_{perm(i)}$. $\mathbf{A}$ is a positive definite matrix, which means that the previous term can be seen as a sum of scalar products $\sum_{i=1}^{n}\langle \mathbf{A}^{1/2}\mathbf{x}_i, \mathbf{A}^{1/2}\mathbf{x}_{perm(i)}\rangle = \sum_{i=1}^{n}\langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{perm(i)}\rangle$, where $\tilde{\mathbf{x}}_i = \mathbf{A}^{1/2}\mathbf{x}_i$. It is easy to show that

$$0 \leq \sum_i \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{perm(i)}\|^2 = 2\sum_i \|\tilde{\mathbf{x}}_i\|^2 - 2\sum_i \tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_{perm(i)},$$

which means that

$$\sum_i \|\tilde{\mathbf{x}}_i\|^2 \geq \sum_i \tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_{perm(i)}.$$

The identity permutation is then a maximum of the cross scalar product and a minimum of J. Since $\mathbf{x}_i \neq \mathbf{x}_j$, if $i \neq j$ the equality stands only for the identity permutation. This means that once again the solution of the optimization problem is the identity matrix. Note that if $\mathbf{A}$ has negative eigenvalues, the transport fails in reconstructing the original transformation, as illustrated in Figure 4.

II. Influence of the choice of the cost function

We examine here role played by the type of cost function $\mathbf{C}$ used in the adaptation. This cost function allows naturally to take into account the specificity of the data space.

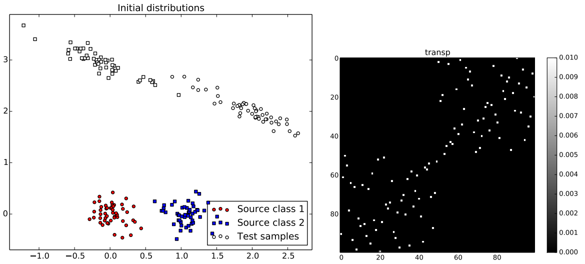

Fig. 4: Optimal transport in the affine transformation 4: case with negative eigenvalues

In the case of the Office-Caltech dataset, the features are histograms. We have tested our **OT-IT** adaptation strategy with alternative costs that operate directly on histograms (ℓ_2 and ℓ_1 norms), as it is a common practice to consider those metrics when comparing distributions. Also, as mostly done in the literature, we consider the pre-pocessed features (normalization + zscore) together with Euclidean and squared Euclidean norms. As it can be seen in Table I, the best results are not always obtained with the same metric. In the paper, we choose to only consider the ℓ_2 norm on preprocessed data, as it is the one that gave the best result on average, but it is relevant to question what is the best metric for an adaptation task. Several other metrics could be considered, like geodesic distances in the case where data live on a manifold. Also, a better tailored metric could be learnt from the data. Some recent works are considering this option [2], [3] for other tasks. This constitutes an interesting perspective for the following extensions of our methodology for the domain adaptation problem.

TABLE I: Results of changing the cost function matrix C on the *Office-Caltech* dataset.

Domains	ℓ_2	ℓ_2+preprocessing	ℓ_2^2+preprocessing	ℓ_1
C→A	36.01	39.14	43.01	39.77
C→W	25.08	35.59	37.63	37.63
C→D	31.21	43.31	42.68	42.68
A→C	32.41	34.64	34.46	38.29
A→W	32.88	34.92	32.20	34.24
A→D	35.03	36.94	33.76	36.94
W→C	22.35	32.59	31.79	25.56
W→A	27.56	39.98	38.00	31.84
W→D	84.71	90.45	89.81	92.99
D→C	26.36	31.79	32.32	30.63
D→A	29.54	32.36	30.69	32.67
D→W	74.92	87.80	88.81	87.12
mean	38.17	44.96	44.60	44.20

III. Remarks on Generalized conditional gradient

We first present the generalized conditional gradient algorithm as introduced by Bredies et al. [4] and, in the subsequent sections, we will provide additional properties that are of interest when applying this algorithm for regularized optimal transport problems.

A. The generalized conditional gradient algorithm

We are interested in the problem of minimizing under constraints a composite function such as

$$\min_{\gamma \in \mathcal{B}} \quad F(\gamma) = f(\gamma) + g(\gamma), \tag{3}$$

where both $f(\cdot)$ and $g(\cdot)$ are convex and differentiable functions and $\mathcal{B}$ is a compact set of $\mathbb{R}^n$. One might want to benefit from this composite structure during the optimization procedure. For instance, if we have an efficient solver for optimizing

$$\min_{\gamma \in \mathcal{B}} \quad \langle \nabla f, \gamma \rangle + g(\gamma) \tag{4}$$

we propose this solver in the optimization scheme instead of linearizing the whole objective function (as one would do with a conditional gradient algorithm [5]).

The resulting approach is defined in Algorithm 1. Conceptually, our algorithm lies in-between the original optimization problem and the conditional gradient. Indeed, if we do not consider any linearization, step 3 of the algorithm is equivalent to solving the original problem and one iterate will suffice for convergence. If we use a full linearization as in the conditional gradient approach, step 3 is equivalent to solving a rough approximation of the original problem. By linearizing only a part of the objective function, we thus optimize a better approximation of that function. This will lead to a provably better certificate of optimality than the one of the conditional gradient algorithm [6]. This makes us thus believe that if an efficient solver of the partially linearized problem is available, our algorithm is of strong interest.

Note that this partial linearization idea has already been introduced by Bredies et al. [4] for solving problem (3). Their theoretical results related to the resulting algorithm apply when $f(\cdot)$ is differentiable, $g(\cdot)$, is convex and both f and g satisfy some others mild conditions like coercivity. These results state that the generalized conditional gradient algorithm is a descent method and that any limit point of the algorithm is a stationary point of $f + g$.

In what follows, we provide some results when f and g are differentiable. Some of these results provide novel insights on the generalized gradient algorithms (relation between optimality and minimizer of the search direction, convergence rate, and optimality certificate) while some are redundant to those proposed by Bredies (convergence).

B. Convergence of the algorithm

1) Reformulating the search direction: Before discussing convergence of the algorithm, we first reformulated its step 3 so as to make its properties more accessible and its convergence analysis more amenable.

The reformulation we propose is

$$\mathbf{s}^k = \underset{\mathbf{s} \in \mathcal{B}}{\operatorname{argmin}} \quad \langle \nabla f(\gamma^k), \mathbf{s} - \gamma^k \rangle + g(\mathbf{s}) - g(\gamma^k) \tag{5}$$

and it is easy to note that the problem in line 3 of Algorithm 1 is equivalent to this one and leads to the same solution.

Algorithm 1 Generalized Conditional Gradient (GGG)

```
Initialize $k = 0$ and $\boldsymbol{\gamma}^0 \in \mathcal{P}$
repeat
    Solve problem
        $\mathbf{s}^k = \underset{\boldsymbol{\gamma} \in \mathcal{B}}{\text{argmin}} \quad \langle \nabla f(\boldsymbol{\gamma}^k), \boldsymbol{\gamma} \rangle + g(\boldsymbol{\gamma})$
    Find the optimal step $\alpha^k$ with $\Delta\boldsymbol{\gamma} = \mathbf{s}^k - \boldsymbol{\gamma}^k$
        $\alpha^k = \underset{0 \le \alpha \le 1}{\text{argmin}} \quad f(\boldsymbol{\gamma}^k + \alpha\Delta\boldsymbol{\gamma}) + g(\boldsymbol{\gamma}^k + \alpha\Delta\boldsymbol{\gamma})$
    or choose $\alpha^k$ so that it satisfies the Armijo rule.
    $\boldsymbol{\gamma}^{k+1} \leftarrow \boldsymbol{\gamma}^k + \alpha^k\Delta\boldsymbol{\gamma}$, set $k \leftarrow k + 1$
until Convergence
```

2) Relation between minimizers of Problems 3 and 5 : The above formulation allows us to derive a property that highligths the relation between problems 3 and 5.

Proposition 3.1: $\mathbf{x}^\star$ is a minimizer of problem (3) if and only if

$$\boldsymbol{\gamma}^\star = \underset{\mathbf{s} \in \mathcal{B}}{\text{argmin}} \quad \langle \nabla f(\boldsymbol{\gamma}^\star), \mathbf{s} - \boldsymbol{\gamma}^\star \rangle + g(\mathbf{s}) - g(\boldsymbol{\gamma}^\star). \tag{6}$$

Proof: The proof relies on optimality conditions of constrained convex optimization problem. Indeed, for a convex and differentiable f and g, $\boldsymbol{\gamma}^\star$ is solution of problem (3) if and only if [7]

$$-\nabla f(\boldsymbol{\gamma}^\star) - \nabla g(\boldsymbol{\gamma}^\star) \in N_{\mathcal{B}}(\boldsymbol{\gamma}^\star), \tag{7}$$

where $N_{\mathcal{B}}(\boldsymbol{\gamma})$ is the normal cone of $\mathcal{B}$ at $\boldsymbol{\gamma}$. In the same way, a minimizer $\mathbf{s}^\star$ of problem (5) at $\boldsymbol{\gamma}^k$ can also be characterized as

$$-\nabla f(\boldsymbol{\gamma}^k) - \nabla g(\mathbf{s}^\star) \in N_{\mathcal{B}}(\mathbf{s}^\star). \tag{8}$$

Now suppose that $\boldsymbol{\gamma}^\star$ is a minimizer of problem (3). It is easy to see that if we choose $\boldsymbol{\gamma}^k = \boldsymbol{\gamma}^\star$, then because $\boldsymbol{\gamma}^\star$ satisfies Equation (7), Equation (8) also holds. Conversely, if $\boldsymbol{\gamma}^\star$ is a minimizer of problem (5) at $\boldsymbol{\gamma}^\star$, then $\boldsymbol{\gamma}^\star$ also satisfies Equation (7). ■

3) Intermediate results and gap certificate: We prove several lemmas and exhibit a gap certificate that provides a bound on the difference of the objective value along the iterations to the optimal objective value.

As one may remark, our algorithm is very similar to a conditional gradient algorithm, the Frank-Wolfe algorithm. As such, our proof of convergence of the algorithm will follow similar lines as those used by Bertsekas. Our convergence results is based on the following proposition and definition given in [5].

Proposition 3.2: from [5]. Let $\{\boldsymbol{\gamma}^k\}$ be a sequence generated by the feasible direction method $\boldsymbol{\gamma}^{k+1} = \boldsymbol{\gamma}^k + \alpha_k\Delta\boldsymbol{\gamma}$ with $\Delta\boldsymbol{\gamma}^k = \mathbf{s}^k - \boldsymbol{\gamma}^k$. Assume that $\{\Delta\boldsymbol{\gamma}^k\}$ is gradient-related and that α^k is chosen by the limited minimization or the Armijo rule, then every limit point of $\{\boldsymbol{\gamma}^k\}$ is a stationary point.

Definition A sequence $\Delta\boldsymbol{\gamma}^k$ is said to be gradient-related to the sequence $\boldsymbol{\gamma}^k$ if, for any subsequence of $\{\boldsymbol{\gamma}^k\}_{k\in K}$ that converges to a non-stationary point, the corresponding subsequence $\{\Delta\boldsymbol{\gamma}^k\}_{k\in K}$ is bounded and satisfies

$$\limsup_{k\to\infty, k\in K} \nabla F(\boldsymbol{\gamma}^k)^\top \Delta\boldsymbol{\gamma}^k < 0$$

Basically, this property says that if a subsequence converges to a non-stationary point, then at the limit point the feasible direction defined by $\Delta\boldsymbol{\gamma}$ is still a descent direction. Before proving that the sequence defined by $\{\Delta\boldsymbol{\gamma}^k\}$ is gradient-related, we prove useful lemmas.

Lemma 3.3: For any $\boldsymbol{\gamma}^k \in \mathcal{B}$, each $\Delta\boldsymbol{\gamma}^k = \mathbf{s}^k - \boldsymbol{\gamma}^k$ defines a feasible descent direction.

Proof: By definition, $\mathbf{s}^k$ belongs to the convex set $\mathcal{B}$. Hence, for any $\alpha^k \in [0, 1]$, $\boldsymbol{\gamma}^{k+1}$ defines a feasible point. Hence $\Delta\boldsymbol{\gamma}^k$ is a feasible direction.

Now let us show that it also defines a descent direction. By definition of the minimizer $\mathbf{s}^k$, we have for all $\mathbf{s} \in \mathcal{B}$

$$\begin{aligned}\langle \nabla f(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k \rangle + g(\mathbf{s}^k) - g(\boldsymbol{\gamma}^k) \le &\langle \nabla f(\boldsymbol{\gamma}^k), \mathbf{s} - \boldsymbol{\gamma}^k \rangle \\ &+ g(\mathbf{s}) - g(\boldsymbol{\gamma}^k).\end{aligned}$$

Because the above inequality also holds for $\mathbf{s} = \boldsymbol{\gamma}^k$, we have

$$\langle \nabla f(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k \rangle + g(\mathbf{s}^k) - g(\boldsymbol{\gamma}^k) \le 0. \tag{9}$$

By convexity of $g(\cdot)$ we have

$$g(\mathbf{s}^k) - g(\boldsymbol{\gamma}^k) \ge \langle \nabla g(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k \rangle,$$

which, plugged in equation, (9) leads to

$$\langle \nabla f(\boldsymbol{\gamma}^k) + \nabla g(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k \rangle \le 0$$

and thus $\langle \nabla F(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k \rangle \le 0$, which proves that $\Delta\boldsymbol{\gamma}^k$ is a descent direction. ■

The next lemma provides an interesting feature of our algorithm. Indeed, the lemma states that the difference between the optimal objective value and the current objective value can be easily monitored.

Lemma 3.4: For all $\boldsymbol{\gamma}^k \in \mathcal{B}$, the following property holds

$$\min_{\mathbf{s} \in \mathcal{B}} \quad \left[\langle \nabla f(\boldsymbol{\gamma}^k), \mathbf{s} - \boldsymbol{\gamma}^k \rangle + g(\mathbf{s}) - g(\boldsymbol{\gamma}^k)\right] \le F(\boldsymbol{\gamma}^\star) - F(\boldsymbol{\gamma}^k) \le 0,$$

where $\boldsymbol{\gamma}^\star$ is a minimizer of F. In addition, if $\boldsymbol{\gamma}^k$ does not belong to the set of minimizers of $F(\cdot)$, then the second inequality is strict.

Proof: By convexity of f, we have

$$f(\boldsymbol{\gamma}^\star) - f(\boldsymbol{\gamma}^k) \ge \nabla f(\boldsymbol{\gamma}^k)^\top(\boldsymbol{\gamma}^\star - \boldsymbol{\gamma}^k).$$

By adding $g(\boldsymbol{\gamma}^\star) - g(\boldsymbol{\gamma}^k)$ to both sides of the inequality, we obtain

$$F(\boldsymbol{\gamma}^\star) - F(\boldsymbol{\gamma}^k) \ge \nabla f(\boldsymbol{\gamma}^k)^\top(\boldsymbol{\gamma}^\star - \boldsymbol{\gamma}^k) + g(\boldsymbol{\gamma}^\star) - g(\boldsymbol{\gamma}^k)$$

and because $\boldsymbol{\gamma}^\star$ is a minimizer of F, we also have $0 \ge F(\boldsymbol{\gamma}^\star) - F(\boldsymbol{\gamma}^k)$. Hence, the following holds

$$\langle \nabla f(\boldsymbol{\gamma}^k), \boldsymbol{\gamma}^\star - \boldsymbol{\gamma}^k \rangle + g(\boldsymbol{\gamma}^\star) - g(\boldsymbol{\gamma}^k) \le F(\boldsymbol{\gamma}^\star) - F(\boldsymbol{\gamma}^k) \le 0$$

and we also have

$$\min_{s\in\mathcal{B}} \quad \langle \nabla f(\boldsymbol{\gamma}^k), \mathbf{s}-\boldsymbol{\gamma}^k\rangle + g(\mathbf{s}) - g(\boldsymbol{\gamma}^k) \leq F(\boldsymbol{\gamma}^\star) - F(\boldsymbol{\gamma}^k) \leq 0,$$

which concludes the first part of the proof.
Finally, if $\boldsymbol{\gamma}^k$ is not a minimizer of F, then we naturally have

$$0 > F(\boldsymbol{\gamma}^\star) - F(\boldsymbol{\gamma}^k).$$

■

4) Proof of convergence: Now that we have all the pieces of the proof, let us show the key ingredient.

Lemma 3.5: The sequence $\{\Delta\boldsymbol{\gamma}^k\}$ of our algorithm is gradient-related.

Proof: For showing that our direction sequence is gradient-related, we have to show that, given a subsequence $\{\boldsymbol{\gamma}^k\}_{k\in K}$ that converges to a non-stationary point $\tilde{\boldsymbol{\gamma}}$, the sequence $\{\Delta\boldsymbol{\gamma}^k\}_{k\in K}$ is bounded and that

$$\limsup_{k\to\infty, k\in K} \nabla F(\boldsymbol{\gamma}^k)^\top \Delta\boldsymbol{\gamma}^k < 0.$$

Boundedness of the sequence naturally derives from the facts that $\mathbf{s}^k \in \mathcal{B}$, $\boldsymbol{\gamma}^k \in \mathcal{B}$ and the set $\mathcal{B}$ is bounded.

The second part of the proof starts by showing that

$$\begin{aligned}\langle \nabla F(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k\rangle &= \langle \nabla f(\boldsymbol{\gamma}^k) + \nabla g(\boldsymbol{\gamma}), \mathbf{s}^k - \boldsymbol{\gamma}^k\rangle \\ &\leq \langle \nabla f(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k\rangle + g(\mathbf{s}^k) - g(\boldsymbol{\gamma}^k),\end{aligned}$$

where the last inequality is obtained owing to the convexity of g. Because that inequality holds for the minimizer, it also holds for any vector $s \in \mathcal{B}$:

$$\langle \nabla F(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k\rangle \leq \langle \nabla f(\boldsymbol{\gamma}^k), \mathbf{s} - \boldsymbol{\gamma}^k\rangle + g(\mathbf{s}) - g(\boldsymbol{\gamma}^k).$$

Taking limit yields to

$$\limsup_{k\to\infty, k\in K} \langle \nabla F(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k\rangle \leq \langle \nabla f(\tilde{\boldsymbol{\gamma}}), \mathbf{s} - \tilde{\boldsymbol{\gamma}}\rangle + g(\mathbf{s}) - g(\tilde{\boldsymbol{\gamma}})$$

for all $\mathbf{s} \in \mathcal{B}$. As such, this inequality also holds for the minimizer

$$\begin{aligned}&\limsup_{k\to\infty, k\in K} \langle \nabla F(\boldsymbol{\gamma}^k), \mathbf{s}^k - \boldsymbol{\gamma}^k\rangle \\ &\qquad \leq \min_{\mathbf{s}\in\mathcal{B}} \langle \nabla f(\tilde{\boldsymbol{\gamma}}), \mathbf{s} - \tilde{\boldsymbol{\gamma}}\rangle + g(\mathbf{s}) - g(\tilde{\boldsymbol{\gamma}}).\end{aligned}$$

Now, since $\tilde{\boldsymbol{\gamma}}$ is not stationary, it is not optimal and it does not belong to the minimizer of F, hence according to the above lemma 3.4,

$$\min_{\mathbf{s}\in\mathcal{B}} \quad \langle \nabla f(\tilde{\boldsymbol{\gamma}}), \mathbf{s} - \tilde{\boldsymbol{\gamma}}\rangle + g(\mathbf{s}) - g(\tilde{\boldsymbol{\gamma}}) < 0,$$

which concludes the proof. ■

This latter lemma proves that our direction sequence is gradient-related, thus proposition 3.2 applies.

5) Rate of convergence: We can show that the objective value $F(\mathbf{x}_k)$ converges towards $F(\mathbf{x}^\star)$ in a linear rate if we have some additional smoothness condition/DTs of $F(\cdot)$. We can easily prove this statement by following the steps proposed by Jaggi et al. [6] for the conditional gradient algorithm.

We make the hypothesis that there exists a constant C_F so that for any $\mathbf{x}, \mathbf{y} \in \mathcal{B}$ and any $\alpha \in [0, 1]$, the inequality

$$F((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq F(\mathbf{x}) + \alpha\nabla F(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{C_F}{2}\alpha^2$$

holds.

Based on this inequality, for a sequence $\{\mathbf{x}_k\}$ obtained from the generalized conditional gradient algorithm we have

$$\begin{aligned}F(\mathbf{x}_{k+1}) - F(\mathbf{x}^\star) &= F((1-\alpha_k)\mathbf{x}_k + \alpha_k\mathbf{s}_k) - F(\mathbf{x}^\star) \\ &\leq F(\mathbf{x}_k) - F(\mathbf{x}^\star) + \alpha_k\nabla F(\mathbf{x}_k)^\top(s_k - \mathbf{x}_k) \\ &\quad + \frac{C_F}{2}\alpha_k^2. \qquad (10)\end{aligned}$$

Let us denote $h(\mathbf{x}_k) = F(\mathbf{x}_k) - F(\mathbf{x}^\star)$. Now by adding to both sides of the inequality $\alpha_k[g(\mathbf{s}_k) - g(\mathbf{x}_k)]$, we have

$$\begin{aligned}&h(\mathbf{x}_{k+1}) + \alpha_k[g(\mathbf{s}_k) - g(\mathbf{x}_k)] \\ &\leq h(\mathbf{x}_k) + \alpha_k[\nabla f(\mathbf{x}_k)^\top(\mathbf{s}_k - \mathbf{x}_k) + g(\mathbf{s}_k) - g(\mathbf{x}_k)] \\ &\quad + \alpha_k\nabla g(\mathbf{x}_k)^\top(\mathbf{s}_k - \mathbf{x}_k) + \frac{C_F}{2}\alpha_k^2 \\ &\leq h(\mathbf{x}_k) + \alpha_k[\nabla f(\mathbf{x}_k)^\top(\mathbf{x}^\star - \mathbf{x}_k) + g(\mathbf{x}^\star) - g(\mathbf{x}_k)] \\ &\quad + \alpha_k\nabla g(\mathbf{x}_k)^\top(\mathbf{s}_k - \mathbf{x}_k) + \frac{C_F}{2}\alpha_k^2,\end{aligned}$$

where the second inequality comes from the definition of the search direction $\mathbf{s}_k$. Because $f(\cdot)$ is convex, we have $f(\mathbf{x}^\star) - f(\mathbf{x}_k) \geq \nabla f(\mathbf{x}_k)^\top(\mathbf{x}^\star - \mathbf{x}_k)$. Thus, we have

$$\begin{aligned}&h(\mathbf{x}_{k+1}) + \alpha_k[g(\mathbf{s}_k) - g(\mathbf{x}_k)] \\ &\leq h(\mathbf{x}_k) + \alpha_k[f(\mathbf{x}^\star) - f(\mathbf{x}_k) + g(\mathbf{x}^\star) - g(\mathbf{x}_k)] \\ &\quad + \alpha_k\nabla g(\mathbf{x}_k)^\top(\mathbf{s}_k - \mathbf{x}_k) + \frac{C_F}{2}\alpha_k^2 \\ &\leq (1-\alpha_k)h(\mathbf{x}_k) + \alpha_k\nabla g(\mathbf{x}_k)^\top(\mathbf{s}_k - \mathbf{x}_k) + \frac{C_F}{2}\alpha_k^2.\end{aligned}$$

Now, again owing to the convexity of $g(\cdot)$, we have $0 \geq -g(\mathbf{s}_k) + g(\mathbf{x}_k)\nabla g(\mathbf{x}_k)^\top(\mathbf{s}_k - \mathbf{x}_k)$. Using this fact in the last above inequality leads to

$$h(\mathbf{x}_{k+1}) \leq (1-\alpha_k)h(\mathbf{x}_k) + \frac{C_F}{2}\alpha_k^2. \qquad (11)$$

Based on this result, we can now state the following

Theorem 3.6: For each $k \geq 1$, the iterates $\mathbf{x}_k$ of Algorithm 1 satisfy

$$F(\mathbf{x}_k) - F(\mathbf{x}^\star) \leq \frac{2C_F}{k+2}.$$

Proof: The proof stands on Equation (11) and on the same induction as the one used by Jaggi et al [6]. ■
Note that this convergence property would also hold if we choose the step size as $\alpha_k = \frac{2}{k+2}$.

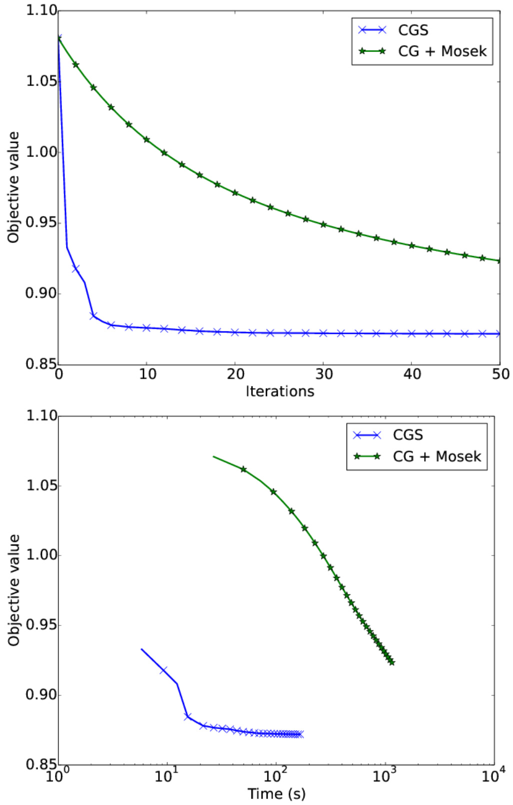

Fig. 5: Example of the evolution of the objective value along (top) the iterations and (bottom) the running times for the generalized conditional gradient (CGS) and the conditional gradient (CG + Mosek) algorithms.

TABLE II: Computational time (seconds) for the best set of regularization parameters

	OT-exact	OT-IT	OT-GL	OT-Laplace
U→M	86.0	4.6	92.5	55.6
M→U	85.0	2.3	75.4	20.5
P1→P2	131.0	30.8	432.7	333.2
P1→P3	133.9	27.9	456.6	296.2
P1→P4	132.5	21.3	276.5	153.4
P2→P1	130.9	38.3	666.8	413.1
P2→P3	65.1	14.3	418.9	182.8
P2→P4	67.8	11.7	270.1	120.0
P3→P1	135.7	36.4	519.0	389.0
P3→P2	67.6	14.3	297.8	156.6
P3→P4	62.2	9.8	248.3	108.8
P4→P1	134.6	22.9	540.4	272.6
P4→P2	66.7	11.1	262.5	123.0
P4→P3	65.3	12.6	269.6	127.5
C→A	26.9	1.2	22.5	17.8
C→W	6.8	0.3	6.6	7.4
C→D	3.4	0.2	3.5	2.4
A→C	26.5	1.1	29.4	23.8
A→W	5.6	0.3	6.0	6.0
A→D	2.9	0.2	3.2	4.1
W→C	6.8	0.3	16.2	7.6
W→A	5.7	0.3	4.1	7.4
W→D	0.8	0.1	2.2	1.1
D→C	3.6	0.2	14.7	5.9
D→A	3.0	0.2	12.5	5.1
D→W	0.8	0.1	3.8	1.1
mean	86.8	20.4	349.1	246.4

6) Computational performances: Let us first show that the GCG algorithm is more efficient than a classical conditional gradient method as the one used in [8]. We illustrate this in Figure 5, showing the convergence (top panel) and the corresponding computational times (bottom panel). We take as an example the case of computing the **OT-GL** transport plan of digits 1 and 2 in USPS to those in MNIST. For this example, we have allowed a maximum of 50 iterations. Regarding convergence of the cost function along the iterations, we can clearly see that, while GCG reaches nearly-optimal objective value in around 10 iterations, the CG approach is still far from convergence after 50 iterations. In addition, the *per-iteration* cost is significantly higher for the conditional gradient algorithm. For this example, we save an order of magnitude of running time yet CG has not converged.

We now study the computational performances of the different optimal transport strategies in the visual object recognition tasks considered above. We use Python implementations of the different OT methods. The test were run on a simple Macbook pro station, with a 2.4 Ghz processor. The original **OT-Exact** solution to the optimal transport problem is computed with the MOSEK [9] linear programming solver[1], whereas the other strategies follow our own implementation based on the Sinkhorn-Knopp method. For the regularized optimal transport **OT-GL** and **OT-Laplace** we used the conditional gradient splitting algorithm (the source code will be made available upon the acceptance of the article).

We report in Table II the computational time needed by the OT methods. As expected, **OT-IT** is the less computationally intensive. The solution of the exact optimal transport (**OT-exact**) is longer to compute by a factor 4. Also, as expected, the two regularized versions **OT-GL** and **OT-Laplace** are the most demanding methods. We recall here that the maximum number of inner loop of the GCG approach was set to 10, meaning that each of those methods made 10 calls to the Sinkhorn-Knopp solver used by **OT-IT**. However, the added computational cost is mostly due to the line search procedure (line 4 in Algorithm 1), which involves several computations of the cost function. We explain the difference between **OT-GL** and **OT-Laplace** by the difference of computation time needed by this procedure. Summing up, one can notice that, even for large problems (case P1→P4 for instance, involving 3332×3329 variables), the computation time is not prohibitive and remains tractable.

References

[1] J. von Neumann, "A certain zero-sum two-person game equivalent to the optimal assignment problem," *Contributions to the Theory of Games*, vol. 2, 1953.

[2] G. Zen, E. Ricci, and N. Sebe, "Simultaneous ground metric learning and matrix factorization with earth mover's distance," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, Aug 2014, pp. 3690–3695.

[1] other publicly available solvers were considered, but it turned out this particular one was an order of magnitude faster than the others

[3] M. Cuturi and D. Avis, “Ground metric learning,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 533–564, Jan. 2014.

[4] Kristian Bredies, Dirk Lorenz, and Peter Maass, *Equivalence of a generalized conditional gradient method and the method of surrogate functionals*, Zentrum für Technomathematik, 2005.

[5] Dimitri P Bertsekas, *Nonlinear programming*, Athena scientific Belmont, 1999.

[6] Martin Jaggi, “Revisiting frank-wolfe: Projection-free sparse convex optimization,” in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 427–435.

[7] D. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, 2003.

[8] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol, “Regularized discrete optimal transport,” *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, 2014.

[9] Erling Andersen and Knud Andersen, “The mosek interior point optimizer for linear programming: An implementation of the homogeneous algorithm,” in *High Performance Optimization*, vol. 33 of *Applied Optimization*, pp. 197–232. Springer US, 2000.