

A statistical physics approach to the optimal transport problem

Patrice Koehl,¹ Marc Delarue,² and Henri Orland³

¹*Department of Computer Science and Genome Center,
University of California, Davis, CA 95616, USA*

²*Unité de Dynamique Structurale des Macromolécules,
Department of Structural Biology and Chemistry,*

UMR 3528 du CNRS, Institut Pasteur, 75015 Paris, France

³*Institut de Physique Théorique, CEA-Saclay, 91191 Gif/Yvette Cedex, France*

(Dated: July 2, 2019)

Originally defined for the optimal allocation of resources, optimal transport (OT) has found many theoretical and practical applications in multiple domains of science and physics. In this paper we develop a new method for solving the discrete version of this problem using techniques derived from statistical physics. We derive a strongly concave free energy function that captures the constraints of the OT problem at a finite temperature. Its maximum defines an optimal transport plan, or registration between the two discrete probability measures that are compared, as well as a **pseudodistance** between those measures that satisfies the triangular inequalities. The computation of this **pseudodistance** is fast and numerically stable. The temperature dependent OT **pseudodistance** is shown to decrease monotonically with respect to the inverse of the temperature and to converge to the standard OT distance at zero temperature, providing a robust framework for temperature annealing. We illustrate applications of this framework to the problem of image comparison.

Imagine that there are N_1 flour milling plants around Paris, France that serve N_2 bakeries within Paris, and let us assume balance, namely that there is as much flour produced by the plants as needed by the bakeries. A company in charge of the distribution of the flour will take into account the individual cost of transporting some amount of flour from one plant to one bakery to find an “optimal transport plan”, namely an assignment of how much flour needs to be transported between each plant and each bakery that leads to a minimal overall cost for the transport. Finding a solution to this seemingly simple practical problem has led to the development of a small gem in the mathematics and statistics communities, namely the optimal transport (OT) problem. What makes the OT problem so interesting is that its solution includes two essential components. First, it defines a distance between the distributions considered, with such distance being referred to as the Monge-Kantorovich distance, the Wasserstein distance, or the earth mover’s distance, depending on the field of applications. These distances have enabled statisticians and mathematicians to derive a geometric structure on the space of probability distributions [1, 2]. Second, it also provides the optimal transportation plan between the distributions; this optimal plan defines a registration, thereby enabling alignment between the distributions. Applications of OT have exploded in the recent years, in domains such as machine learning, computer vision, and linguistics. Multiple fields of physics are also impacted, from applications of OT to density functional theory [3], quantum mechanics [4], stochastic thermodynamics [5], and general relativity [6], among others. For extensive reviews of OT and its applications, see [1, 2].

Our focus in this paper is on the discrete version of the OT problem. We consider two sets of points S_1 of

size N_1 and S_2 of size N_2 . Each point k in S_1 (resp S_2) is assigned a “mass” $m_1(k)$ (resp $m_2(k)$). The balance condition implies that $\sum_k m_1(k) = \sum_l m_2(l)$. We assume that these two sums are equal to 1. We encode the cost of transport between S_1 and S_2 as a positive matrix C_{kl} with $k \in \{1 \dots, N_1\}$ and $l \in \{1 \dots, N_2\}$. The OT problem can then be formulated as finding a matrix G of correspondence between points in S_1 and points in S_2 that minimizes the transport cost U defined as

$$U(G) = \sum_{k,l} G_{kl} C_{kl} \quad (1)$$

where the summations extend over all k in S_1 and l in S_2 . The minimum of U is to be found for the values of G_{kl} that satisfy the following constraints

$$\forall(k, l), \quad G_{k,l} \geq 0 \quad (2a)$$

$$\forall k, \quad \sum_l G_{kl} = m_1(k) \quad (2b)$$

$$\forall l, \quad \sum_k G_{kl} = m_2(l) \quad (2c)$$

The solution to the OT problem provides an optimal transport plan G_{opt} and the corresponding minimum transport cost $U_{min} = U(G_{opt})$. The minimum cost defines a “distance” between the two sets of points. It has all the properties of a metric when the cost matrix C is a metric matrix, see [1]. This metric is often referred to as the Wasserstein distance $W(S_1, S_2)$ between the two sets of points. In this paper we will make the same assumption that C is a metric matrix.

Note that the first condition on G , (2a), extends to $0 \leq G_{kl} \leq 1$ for all k and l , based on our assumption that the sum of the discrete probability measures are 1

on both sets of points. Optimizing (1) under the constraints (2) is a linear programming (LP) problem. While much progress has been achieved for solving those problems [7], current practical implementations are roughly of order $O(n^3)$ (where $n = \max(N_1, N_2)$), with a quadratic complexity in the number of variables considered. Such complexity levels are considered problematic when n is larger than a few thousands.

Interestingly, the current successes of OT have been triggered by the idea of minimizing a regularized version of equation 1 :

$$U_\epsilon(G) = \sum_{kl} G_{kl} C_{kl} - \epsilon \sum_{k,l} G_{kl} \log(G_{kl}) \quad (3)$$

where ϵ is the regularization parameter, and the second term is an entropic barrier that enforces the positivity of the G_{kl} terms [8]. This regularized version of optimal transport is often called the Schrödinger problem [9]. It maps to the traditional OT problem as $\epsilon \rightarrow 0$; in addition, the optimal solution $U(G_{opt})$ where G_{opt} has been computed at a given ϵ is a **pseudodistance**, referred to as the Sinkhorn distance [8]. Note that **it is labeled as a pseudodistance as** it only satisfies the symmetry and triangular inequality properties of a metric [10]. The entropic penalization has the advantage that it defines a strongly convex problem with a unique solution [8]. Another advantage of the regularized OT problem is that its solution can be found efficiently using the Sinkhorn's algorithm [11, 12], with a time complexity of $O(n^2)$. This algorithm finds solutions for a given value of the relaxation parameter ϵ . For small values of this parameter, numerical issues may arise and a stabilization is necessary [13]. Convergence of a stabilized Sinkhorn algorithm can nevertheless be very slow when ϵ is small. Such small values are, however, desirable for finding good approximations to the solution of the non-regularized OT problem. A popular heuristic solution to this problem is the so-called ϵ -scaling, where one solves the regularized problem with gradually decreasing values for ϵ [14]. To our knowledge, no quantitative analyses of the convergence of such an ϵ -scaling method are available.

In this paper, we develop an alternate framework for solving the OT problem that is derived from a statistical physics point of view and report on its applications to computing the similarities between images. A more detailed theoretical and numerical analysis is presented in a companion paper [15]. In this framework, we exploit the formal analogy of the cost function in Eq. 3 to a free energy (ϵ is then the analogue of a temperature, T). It can be seen as a generalization of the so-called invisible hand algorithm, which used a similar approach for solving the assignment problem [16].

In statistical physics, a system that is in thermal equilibrium at finite temperature will sample many states. The corresponding Gibbs distribution represents the probability of this system to exist in any specific state.

The most probable state is then the one with lowest energy. Hence, minimizing an energy function can be reformulated as the problem of finding the most probable state of the system it defines. Let us consider two sets of weighted points, a source set S_1 and a target set S_2 equipped with masses $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively. To solve the OT problem between those two sets, the “system” is identified with the different transport plans between S_1 and S_2 that satisfy the constraints of mass balance and positivity. Those plans belong to a convex polytope denoted as $\mathcal{G}(S_1, S_2)$. Each state in this system is identified with a transport plan $G \in \mathcal{G}(S_1, S_2)$, and its corresponding energy $U(G)$ is defined in equation 1. The probability $P(G)$ associated with a transport plan G is derived as:

$$P(G) = \frac{1}{Z_\beta(S_1, S_2)} e^{-\beta U(G)} \quad (4)$$

In this equation, $\beta = 1/(k_B T)$ where k_B is the Boltzmann constant and T the temperature, and $Z_\beta(S_1, S_2)$ is the partition function computed over all states of the system. This partition function is given by

$$Z_\beta(S_1, S_2) = e^{-\beta \mathcal{F}_\beta(S_1, S_2)} = \int_{G \in \mathcal{G}(S_1, S_2)} e^{-\beta U(G)} d\mu_{12} \quad (5)$$

where $d\mu_{12}$ can be seen as the Lebesgue measure for the space of transport plans $\mathcal{G}(S_1, S_2)$ and $\mathcal{F}_\beta(S_1, S_2)$ is the free energy of the system. We note first two important properties of this free energy, namely that $\mathcal{F}_\beta(S_1, S_2)$ is symmetric and satisfies all triangle inequalities if the cost matrix C is metric, and that it is a monotonically decreasing function of β that converges to the traditional OT distance $W(S_1, S_2)$ (see the companion paper [15] for proofs). From those two properties, it is clear that this formulation of the optimal transport problem is appealing. It defines a temperature dependent free energy that satisfies metric properties when the cost function is metric, with a monotonic dependence on the temperature, and convergence to the actual OT distance at zero temperature. It is, unfortunately, of limited practical interest as the free energy cannot be computed explicitly. We propose a scheme for approximating the free energy using the saddle point approximation.

Taking into account the constraints defined in equation 2, the partition function can be written as

$$Z = \int_0^1 \prod_{kl} dG_{kl} e^{-\beta \sum_{kl} C_{kl} G_{kl}} \prod_k \delta\left(\sum_l G_{kl} - m_1(k)\right) \prod_l \delta\left(\sum_k G_{kl} - m_2(l)\right) \quad (6)$$

We use the Fourier representation of the delta functions, thereby introducing new auxiliary variables λ_k and μ_l , with $k \in \{1 \dots, N_1\}$ and $l \in \{1 \dots, N_2\}$, respectively. The partition function can then be written as (up to a

multiplicative constant),

$$Z = \int \prod_k d\lambda_k \int \prod_l d\mu_l \int_0^1 \prod_{k,l} dG_{kl} e^{-\beta \sum_{k,l} G_{kl} (C_{kl} + i\lambda_k + i\mu_l) + \beta \left(\sum_k i\lambda_k m_1(k) + \sum_l i\mu_l m_2(l) \right)} \quad (7)$$

Performing the integration over the variables G_{kl} , we get,

$$Z = \int \prod_k d\lambda_k \int \prod_l d\mu_l e^{-\beta F_{eff}(\beta, i\lambda_k, i\mu_l)} \quad (8)$$

where F_{eff} is a functional, or effective free energy defined by:

$$F_{eff}(\beta, \lambda, \mu) = - \left(\sum_k \lambda_k m_1(k) + \sum_l \mu_l m_2(l) \right) - \frac{1}{\beta} \sum_{kl} \ln \left[\frac{1 - e^{-\beta(C_{kl} + \lambda_k + \mu_l)}}{\beta(C_{kl} + \lambda_k + \mu_l)} \right]. \quad (9)$$

Let $\bar{G}_{kl}$ be the expected value of G_{kl} with respect to the Gibbs distribution given in equation 4. As mentioned above, it is unfortunately not possible to compute these expected values directly as the partition function defined in (8) is not known analytically. Instead, we derive a saddle point approximation (SPA) by looking for extrema of the effective free energy with respect to the variables λ and μ :

$$\frac{\partial \mathcal{F}_{eff}(\beta, i\lambda, i\mu)}{\partial \lambda_k} = 0 \quad \text{and} \quad \frac{\partial \mathcal{F}_{eff}(\beta, i\lambda, i\mu)}{\partial \mu_l} = 0 \quad (10)$$

After some rearrangements, those two equations can be written as

$$\forall k, \quad \sum_l \bar{G}_{kl} = m_1(k) \quad (11a)$$

$$\forall l, \quad \sum_k \bar{G}_{kl} = m_2(l) \quad (11b)$$

where,

$$\bar{G}_{kl} = \phi[\beta(C_{kl} + i\lambda_k + i\mu_l)] \quad (12)$$

and

$$\phi(x) = \frac{e^{-x}}{e^{-x} - 1} + \frac{1}{x}. \quad (13)$$

As is often the case, the saddle-point may be purely imaginary. In the present case, one can easily see from eq.(11) that the variables $i\lambda_k$ and $i\mu_l$ must be real and in the following, we will replace $\{i\lambda_k, i\mu_l\}$ by $\{\lambda_k, \mu_l\}$. We observe also that equations (9), (11), and (12) are invariant under the translation $\{\lambda_k + K, \mu_l - K\}$ where K is an arbitrary constant. This translational degree of freedom leaves the free energy F_{eff} unchanged.

To analyze the SPA, we need to check the existence and assess the unicity of the critical points of the free energy. In the companion paper [15], we have shown that the Hessian of $F_{eff}(\beta, \lambda, \mu)$ is negative semi-definite with $(N_1 + N_2 - 1)$ strictly negative eigenvalues and one zero eigenvalue. Furthermore, the eigenvector corresponding to the zero eigenvalue is $(1, \dots, 1, -1, \dots, -1)$ (with N_1 1s, and N_2 -1s), and thus corresponds to the constant translation invariance of this free energy. Setting one of the parameters λ_k or μ_k to zero, the free energy function on this restricted parameter space is strictly concave.

For a given value of β , the expected values $\bar{G}_{kl}$ form a transport plan G_{opt} between S_1 and S_2 that is optimal with respect to the effective free energy. We can associate to this transport plan an optimum mean field energy $U_\beta^{MF}(G_{opt})$. This energy satisfies some important properties. Namely, for all $\beta > 0$ and cost metric matrix C , $d_\beta(S_1, S_2) = U_\beta^{MF}(G_{opt})$ is a temperature dependent **pseudodistance** between S_1 and S_2 that satisfies the symmetry and triangular inequality properties of a metric (**it is not a true distance as $d_\beta(S_1, S_1) > 0$ for β finite**). Furthermore, $d_\beta(S_1, S_2)$ is a monotonic decreasing function of the parameter β that converges to the transport distance $W(S_1, S_2)$. The validity of those two assertions is proved in the companion paper [15].

The properties of the free energy functional and of the optimized mean field energy of the system highlight a number of advantages of the proposed framework. First, at each temperature the OT problem is turned into a strongly concave problem with a unique solution. This problem has a linear complexity in the number of variables. The concavity allows for the use of simple algorithms for finding a maximum of the free energy functional (equation 9). We have used an iterative Newton method to solve those equations. In the companion paper [15], we have shown that these equations can be solved with a time complexity of $O(n^2)$, thereby making this approach competitive with the entropy regularized OT method. We note also that equation 12 has better numerical stability than the operations involved in the Sinkhorn algorithm, even at low temperatures, because of the ratio of exponentials in the definition of the function ϕ . Second, the modified problem defines an optimal **pseudodistance** at each temperature, that converges to the traditional OT distance when $T \rightarrow 0$. Finally, the convergence as a function of temperature is monotonic.

We present some computational examples that illustrates the use of our framework. We used the publicly available Japanese Female Facial Expression (JAFPE) database [17, 18]. This database comprises 213 images of facial expressions posed by 10 Japanese female models. We refer to the images of one of the model as a class. Each class contains several different poses, namely 6 basic facial expressions + 1 neutral. We characterized each image by selecting a set of “keypoints” using the SURF procedure. SURF is an image feature detector

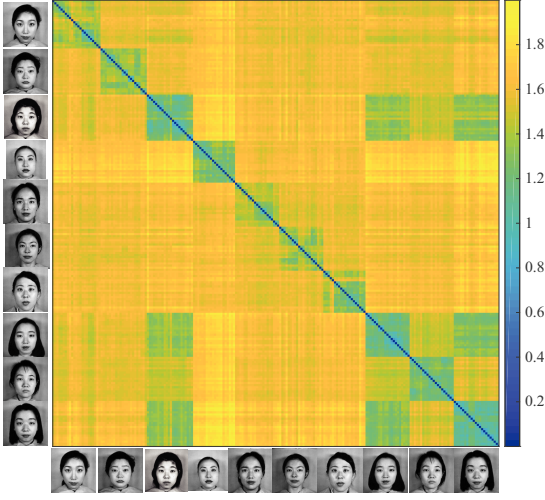


FIG. 1. **Optimal transport distance matrix between all images in the JAFFE facial expression database.** The distance $D(\infty)(k, l)$ is the converged transport energy between the SURF keypoints of the two images k and l (see text for details). The different images for the same model appear closer to each other than to images of other models. Interestingly $D(\infty)$ shows similarity between models 3, 8, and 10; those are the only 3 models whose hairs cover their ears.

and descriptor used for object recognition and image registration [19]. Within SURF, a keypoint is a pixel within the image that is expected to be significant, i.e. a signature feature of the image. The significance is defined from a local neighborhood of the pixel of interest, characterized by a vector of 64 features. A pair of images is represented with their sets of keypoints, S_1 and S_2 , the cost matrix C between those keypoints, such that C_{kl} between a keypoint k on image 1 and a keypoint l on image 2 is equal to the Euclidean distance between their feature vectors. The masses of the keypoints are set uniform. We computed a set of matrices $D(\beta)$ for β ranging between 1000 and 10^{10} , such that $D(\beta)(k, l)$ is the optimized transport energy $d_\beta(S_k, S_l) = U_\beta^{MF}(G_{opt})$, i.e. the temperature-based **pseudodistance** between the sets of keypoints S_k and S_l of the images k and l . We also computed $D(\infty)$, the matrix of distances at convergence. See figure 1 for a graphical representation of $D(\infty)$.

In order to assess the discriminative power contained in the different matrices $D(\beta)$, we considered a set of classification task as follows: we randomly selected half the images from each class to form a training set and use it for performing 1-nearest neighbor classification (where nearest is with respect to the **pseudodistance** d_β to the remaining images). By simple comparison between the class predicted by the classifier and the actual class to which the image belongs we obtain an estimate for the probability of correct classification $P(\beta)$ using d_β . We then repeat this procedure for 10000 random choices of the training set. In figure 2, we plot $P(\beta)$ as a function of

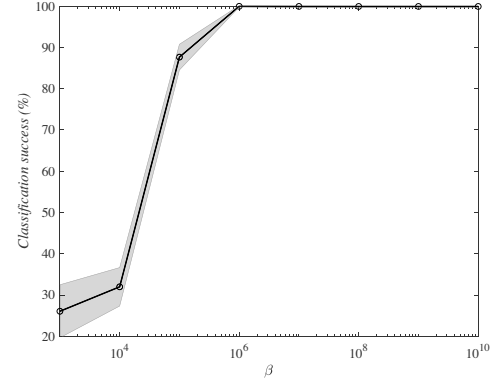


FIG. 2. **Discriminative power of the temperature-based OT **pseudodistances** for images in the JAFFE database.** The probability of correct classification using the **pseudodistance** measure d_β (see text for details) is plotted against $\beta = 1/T$. The solid line corresponds to the arithmetic means computed over 10000 classification experiments (see text for details). Shaded areas represent standard deviations.

β . Note that the lower the temperature (or alternatively the higher the parameter β), the more discriminative the **pseudodistance** d_β . Optimal classification (99.9 %) however is already obtained for $\beta = 10^6$, i.e. much before convergence to the optimal transport distance, usually reached for $\beta > 10^9$. This high success rate should be compared to a success rate of 37% when the Hausdorff distance [20] is used to compare two images (using the same keypoint representation of the image, with the same Euclidean-based distance between keypoints).

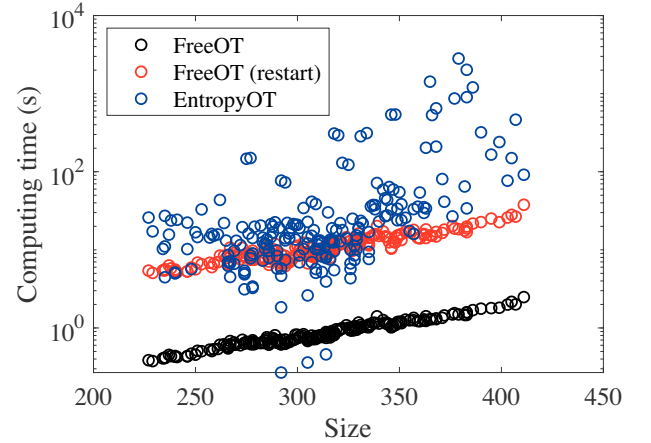


FIG. 3. The average running times for FreeOT, without (black circles) and with reset (red circles) and for EntropyOT used for comparing two images is plotted against the number of keypoints in the images (see text for details).

We have claimed above that the temperature-based OT method enables a fast and robust solution to the OT problem. To check that it is indeed the case, we have compared our implementation of this method, FreeOT,

with our own implementation, EntropyOT, of the entropy regularized approach to the OT problem to compare images, as described above. EntropyOT is based on a log-domain stabilization and eta-scaling heuristic [14] and an overrelaxation scheme [21]; both modifications to the original algorithm of Cuturi [8] are expected to improve convergence and robustness. We have compared each image in the JAFFE dataset against five other images that have similar numbers of keypoints. Each comparison is performed until convergence, i.e. until the relative change in the energy falls below a tolerance of 10^{-6} . Such convergence is usually reached for $\beta = 10^{11}$ (or equivalently for $\epsilon = 10^{-11}$ for EntropyOT). The computing times for FreeOT and EntropyOT, averaged over the 5 comparisons, are plotted against the number of keypoints in the images in Figure 3. With only a few exceptions, FreeOT is always found to be faster than EntropyOT, as the latter is found to slow down significantly for very small ϵ values. While convergence with high precision may not be needed, we observe that FreeOT is free of those convergence problems. Both FreeOT and EntropyOT include a scaling of the regularization parameter. For FreeOT, the values of the converged parameters λ and μ at one value of β are used as input to the next value of β considered. This is expected to improve convergence. To check if this is true, we repeated the calculations with FreeOT by resetting λ and μ to zero for each β value. The results are shown in Figure 3, as FreeOT (reset). The reset does lead to less efficient convergence. Of significance, however, the computations remain feasible, even for very large β values. In contrast, similar experiments with EntropyOT failed for most image comparisons, due to numerical instabilities for $\epsilon < 10^{-5}$.

In summary, we have used statistical physics to derive an alternative representation of the discrete optimal transport problem. We have constructed a strongly concave effective free energy function that captures the constraints of the OT problem. This effective free energy function is parameterized by temperature. Its maximum defines an optimal transport plan as well as a **pseudodistance** between the two sets of points considered. In the companion paper [15], we have shown that this formalism can be implemented in an algorithm with the same time complexity as those of the implementations of the regularized OT algorithms in time complexity, making it a competitive approach to solving the OT problem and therefore amenable to applications in data sciences. In addition, the temperature dependent OT **pseudodistance** is shown to decrease monotonically with respect to the parameter β to the standard optimal transport or Wasserstein distance, thereby providing a robust framework for temperature annealing, a process that is still elusive for the entropy regularized optimal transport problem. Addressing issues of partial transport, as well as extensions to other transport problems such as the

computation of the Gromov-Wasserstein distance [22] are promising directions for future work.

The work discussed here originated from a visit by P.K. at the Institut de Physique Théorique, CEA Saclay, France, during the fall of 2018. He thanks them for their hospitality and financial support.

-
- [1] C. Villani, *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften (Springer, 2008).
 - [2] G. Peyré and M. Cuturi, arXiv e-prints , arXiv:1803.00567 (2018), arXiv:1803.00567 [stat.ML].
 - [3] G. Buttazzo, L. D. Pascale, and P. Gori-Giorgi, *Physical Review A* **85**, 062502 (2012).
 - [4] Y. Chen, T. T. Georgiou, and A. Tannenbaum, *IEEE Trans. Automatic Control* **63**, 2612 (2018).
 - [5] E. Aurell, C. Mejía-Monasterio, and P. Muratore-Ginanneschi, *Phys. Rev. Lett.* **106**, 250601 (2011).
 - [6] A. Mondino and S. Suhr, arXiv e-prints , arXiv:1810.13309 (2018), 1810.13309 [math-ph].
 - [7] G. Sierksma and Y. Zwols, *Linear and Integer Optimization: Theory and Practice, Third Edition* (CRC Press, 2015).
 - [8] M. Cuturi, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., 2013) pp. 2292–2300.
 - [9] C. Léonard, *Discrete Continuous Dyn. Syst. Series A* **34**, 1533 (2014).
 - [10] S.-I. Amari, R. Karakida, M. Oizumi, and M. Cuturi, *Neural Computation* **31**, 827 (2019).
 - [11] R. Sinkhorn, *Ann. Math. Statistics* **35**, 876 (1964).
 - [12] R. Sinkhorn and P. Knopp, *Pacific J. Math.* **21**, 343 (1967).
 - [13] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *Math. Comp.* **87**, 2563 (2018).
 - [14] B. Schmitzer, arXiv e-prints , arXiv:1610.06519 (2016), 1610.06519 [math.OC].
 - [15] P. Koehl, M. Delarue, and H. Orland, *Phys. Rev. E* (2019).
 - [16] J. Kosowsky and A. Yuille, *Neural Networks* **7**, 477 (1994).
 - [17] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition* (1998) pp. 200–205.
 - [18] M. Lyons, J. Budynek, and S. Akamatsu, *IEEE Trans. Pattern Anal. Mach. Intell.* **21**, 1357 (1999).
 - [19] H. Bay, A. Ess, T. Tuytelaars, and L. VanGool, *Computer Vision and Image Understanding* **110**, 346 (2008).
 - [20] R. Rockafellar and R. J.-B. Wets, *Variational Analysis* (Springer-Verlag, 2005).
 - [21] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis, in *NIPS’17 Workshop on Optimal Transport & Machine Learning* (Long Beach, United States, 2017).
 - [22] F. Mémoli, in *Eurographics Symposium on Point-Based Graphics* (2007) pp. 81–90.