Optimal transport in machine learning

Quentin Berthet

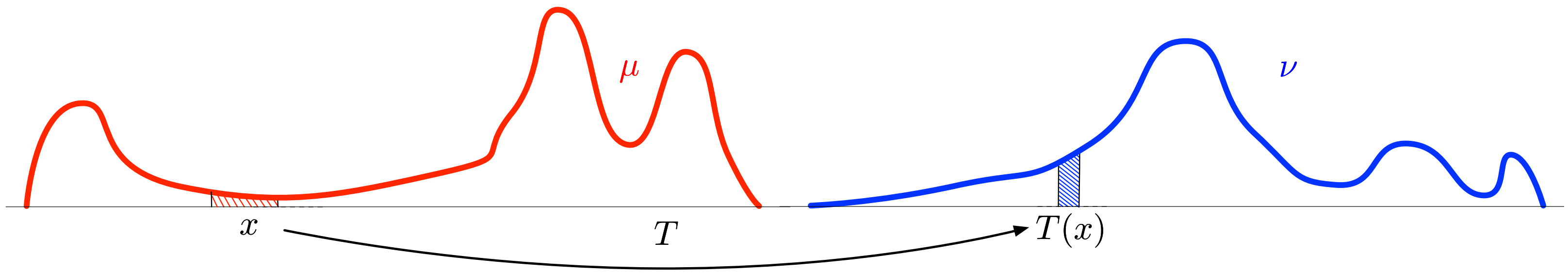
$$\xrightarrow{\quad T(\cdot) \quad}$$

Google AI
Brain Team

MLiTRW workshop - Critéo - 2019

Optimal transport - Monge (1781)

Transporting mass with measure μ to have measure ν with minimal effort.



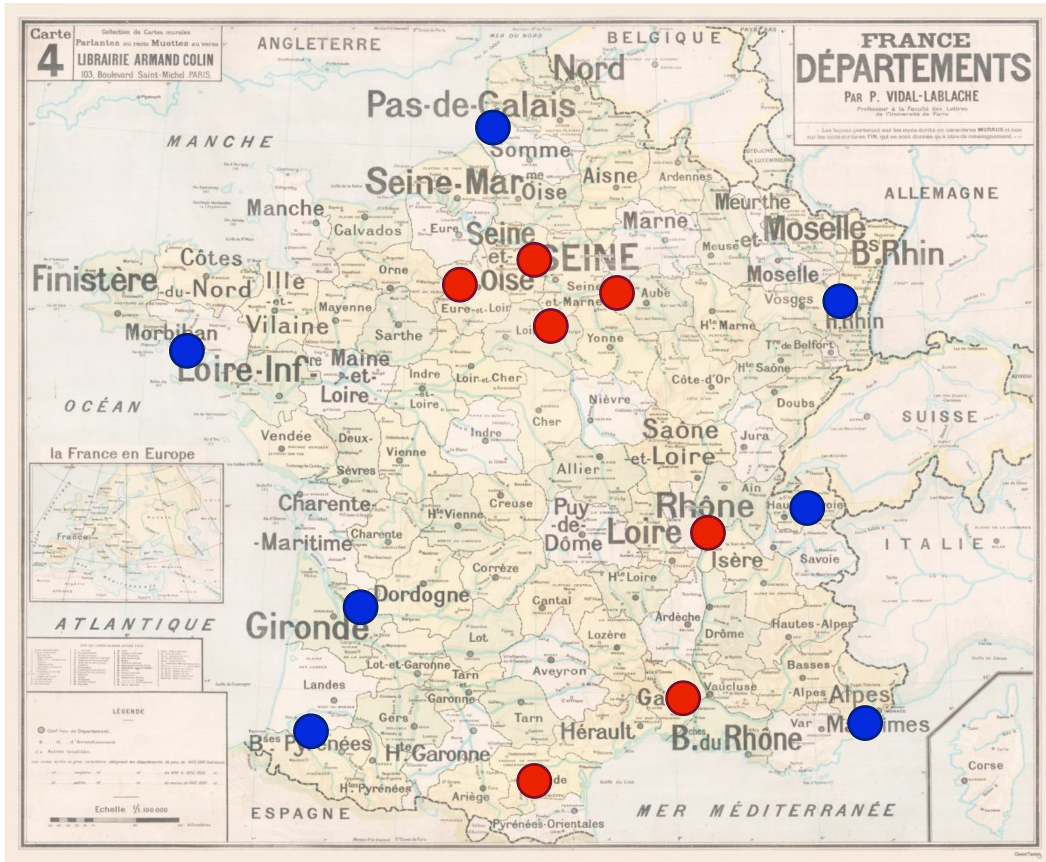
- All maps $T_{\#}\mu = \nu$ (transport from μ to ν): $T(X) \sim \nu$ when $X \sim \mu$.
- Finding map that minimizes the total transport cost.

$$W_p^p(\mu,\nu) = \inf_{T : T_{\#}\mu = \nu} \int \|T(x) - x\|^p \, \mathrm{d}\mu(x)\,.$$

Wasserstein distances between distributions based on optimal transport.

Measures "smallest" transformation between distributions.

Optimal transport - discrete

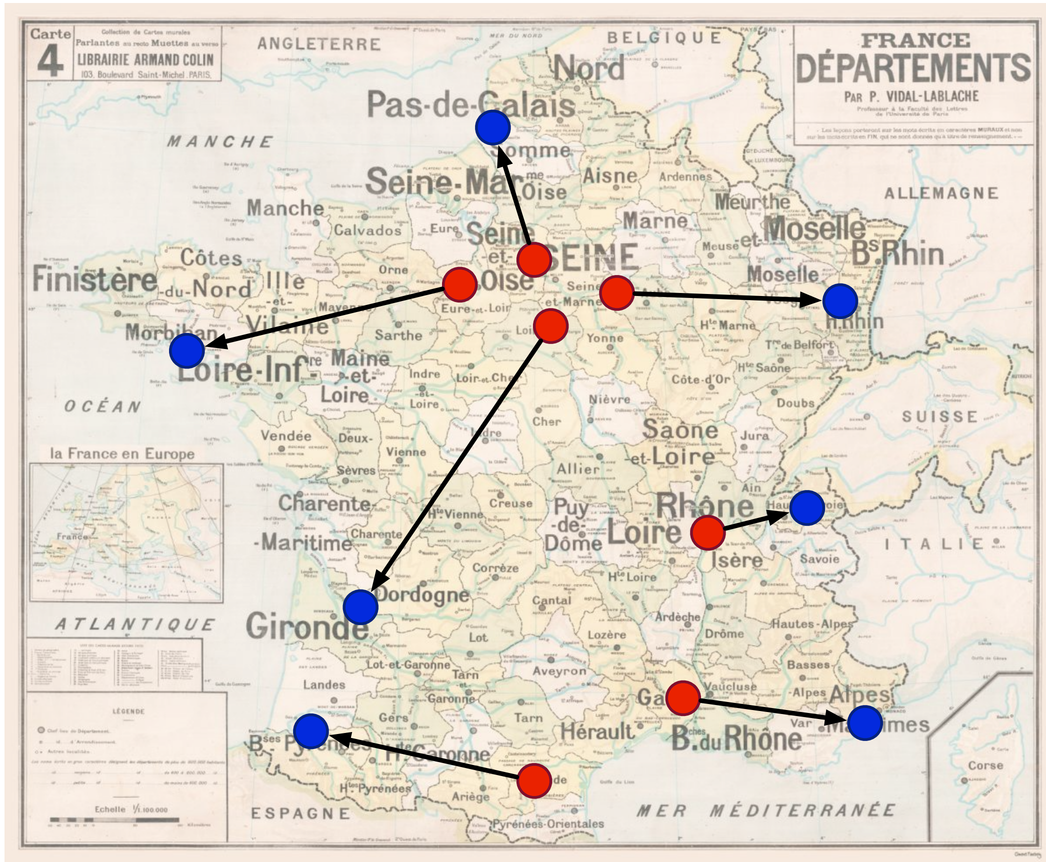


Transport measure μ to have measure ν with minimal effort. Monge (81)

$$\min_{T : T_{\#}\mu = \nu} \sum_x c\big(T(x), x\big)\, \mu(x) \quad \text{(discrete)}\,.$$

Complicated constraint, requires possible one-to-one mapping.

Optimal transport - discrete

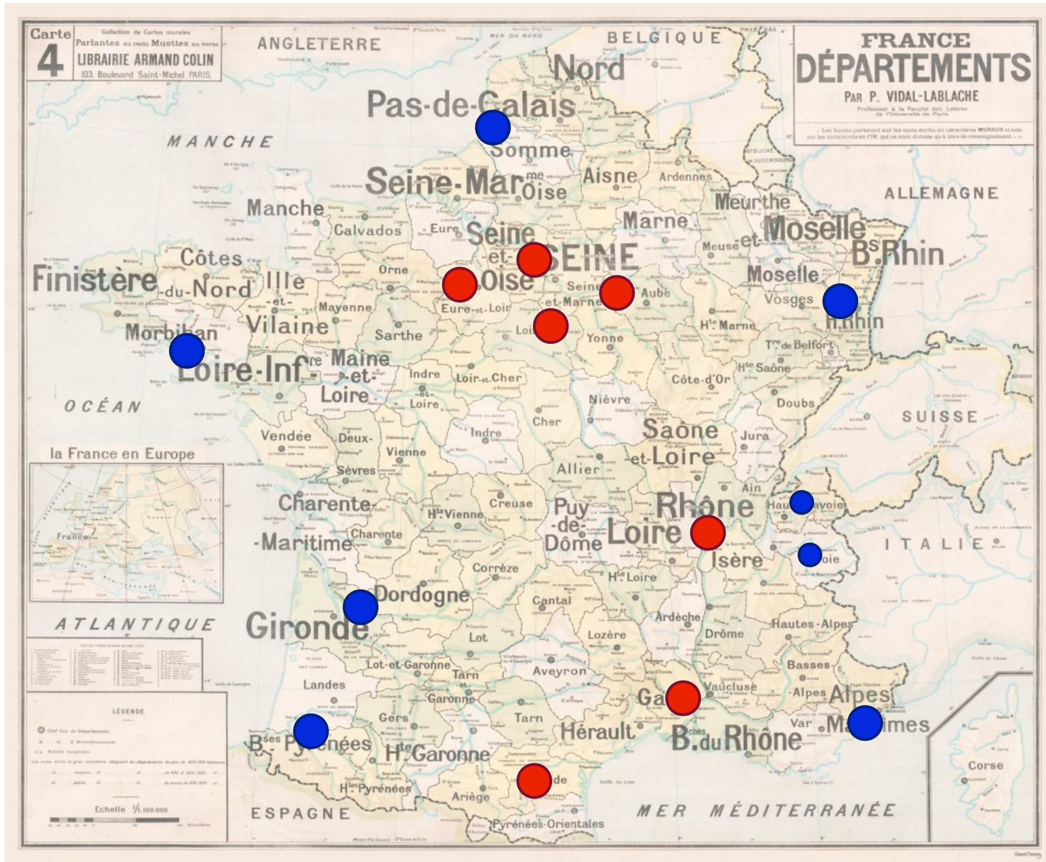


Transport measure μ to have measure ν with minimal effort. Monge (81)

$$\min_{T : T_{\#}\mu = \nu} \sum_x c(T(x), x)\, \mu(x) \quad \text{(discrete)}\,.$$

Complicated constraint, requires possible one-to-one mapping.

Optimal transport - discrete

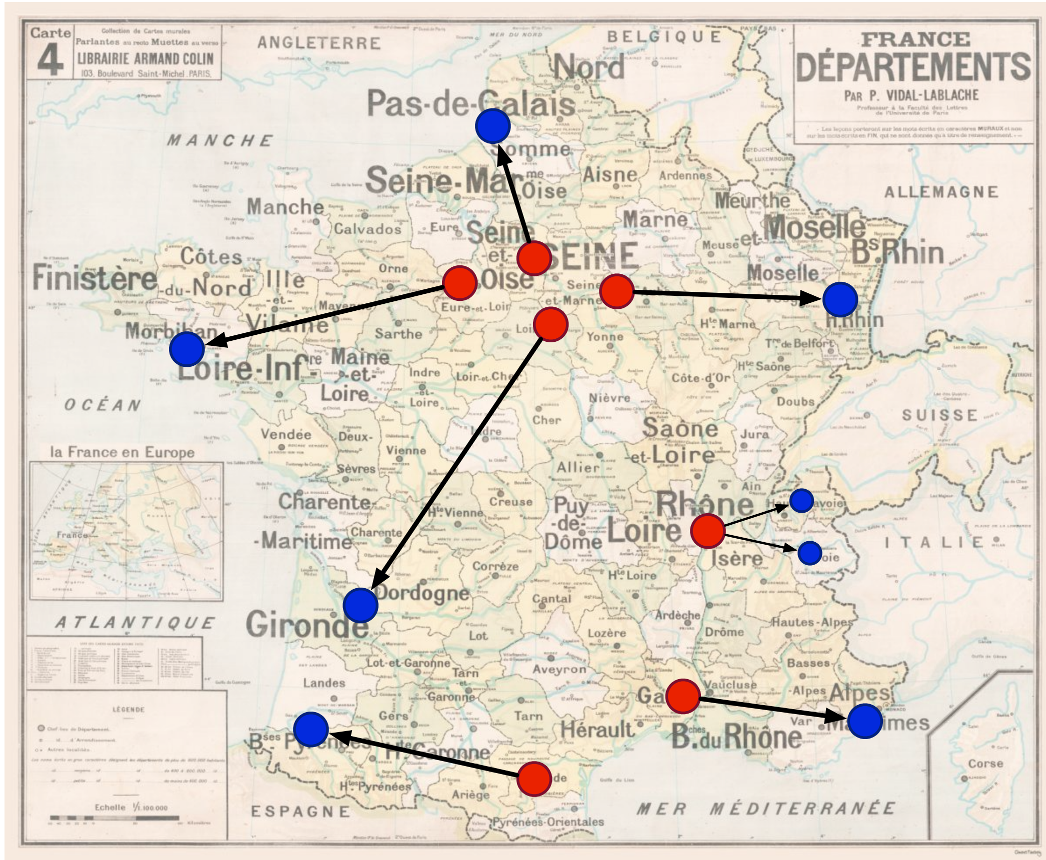


Transport measure μ to have measure ν with minimal effort. Monge (81)

$$\min_{T:T_{\#}\mu=\nu} \sum_{x} c\big(T(x), x\big)\, \mu(x) \quad \text{(discrete)}\,.$$

Complicated constraint, requires possible one-to-one mapping.

Optimal transport - discrete



Problem of transporting mass with measure μ to have measure ν.

$$\min_{\pi \in \mathcal{M}(\mu,\nu)} \sum_{x,y} c(x,y)\,\pi(x,y) = \min_{\pi \in \mathcal{M}(\mu,\nu)} \mathbf{E}_{(X,Y)\sim\pi}[c(X,Y)] \quad \text{Kantorovitch (42)}\,.$$

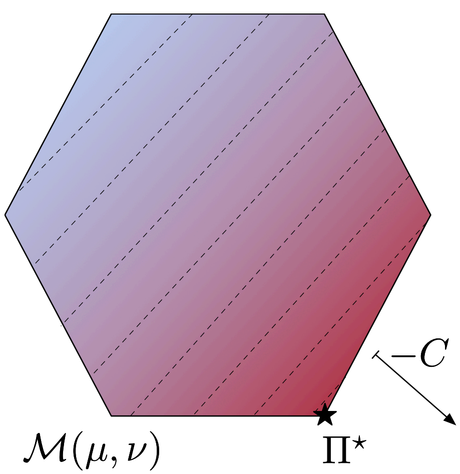
Distribution $\pi(x,\cdot)$ describes how the mass $\mu(x)$ is split. Linear constraints.

Optimal transport

In discrete case, with matrix $C_{x,y} = c(x,y)$ and $\Pi_{x,y} = \pi(x,y)$.

Forms a **linear program**, one of the foundational problems of optimization.

$$\begin{aligned} \min \quad & \langle \Pi, C \rangle \qquad \textbf{(OT)} \\ s.t. \quad & \mathbf{1}^\top \Pi = \mu^\top \\ & \Pi \, \mathbf{1} = \nu \\ & \Pi \geq 0 \,. \end{aligned}$$



- Linear objective and constraints.
- Size n problems: algorithm in $O(n^3)$.
- Linked to assignment problem.
- Solutions in extreme points: sparse.
- Uniform distributions:

  One-to-one transports

  Birkhoff polytope, relaxation tight.

Optimal transport - entropic regularized

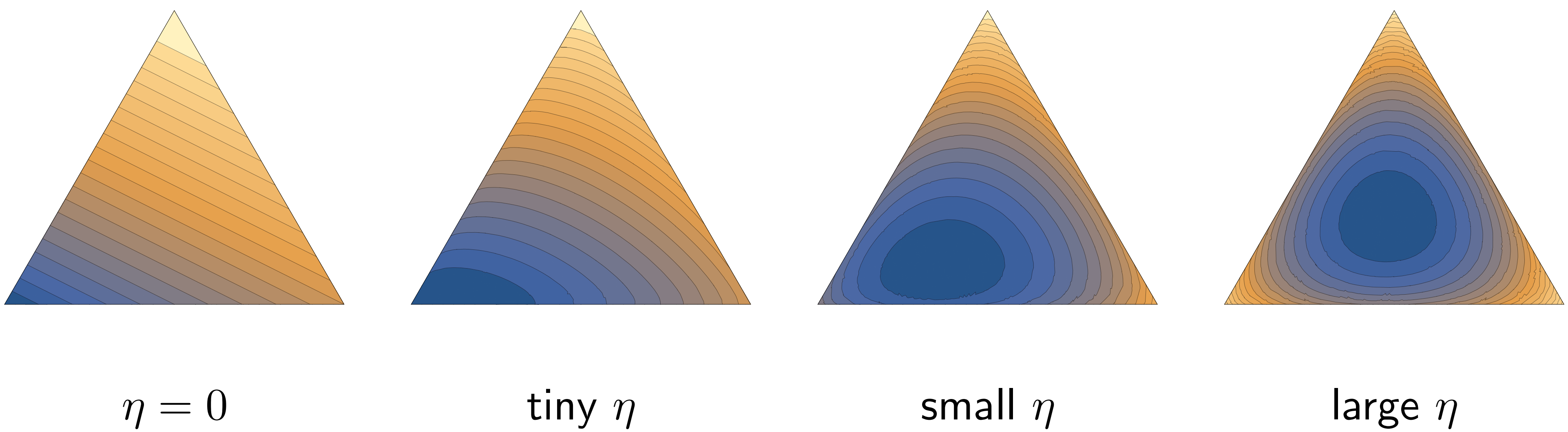
Regularized version, with **entropic penalty**, for $\eta > 0$ Wilson (62), Cuturi (13)

$$\min_{\Pi \in \mathcal{M}(\mu,\nu)} \langle \Pi, C \rangle - \eta H(\Pi)\,.$$

Computational speed-up Sinkhorn (64), strongly convex objective, influence of η.

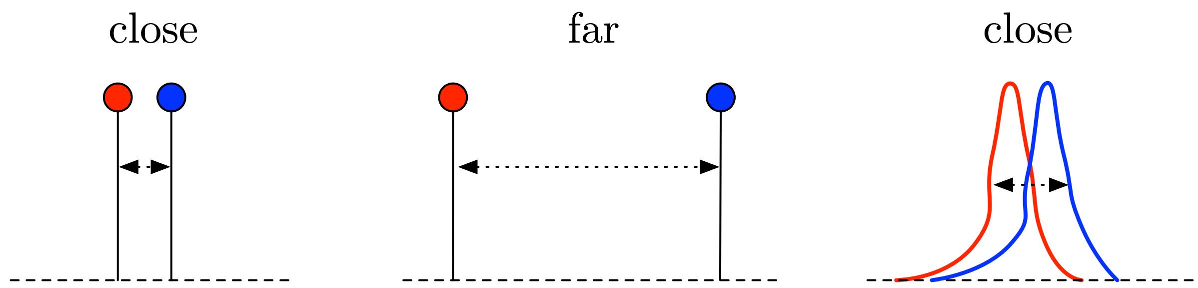
Guarantees for ε-approximation of **(OT)** in $O(n^2 \log(n)/\varepsilon^2)$ for all costs

Altschuler et al. (17), Dvurechensky et al. (17)

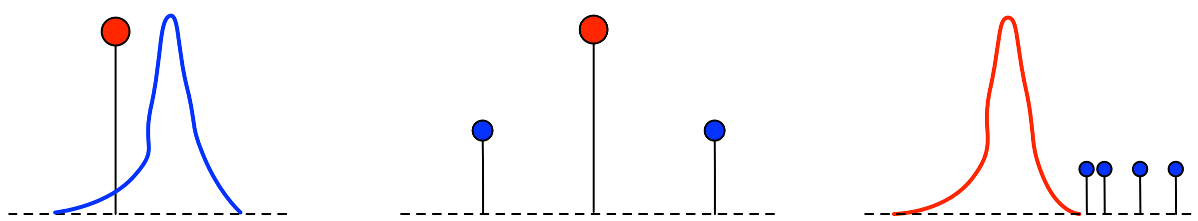


Optimal transport - statistics and ML

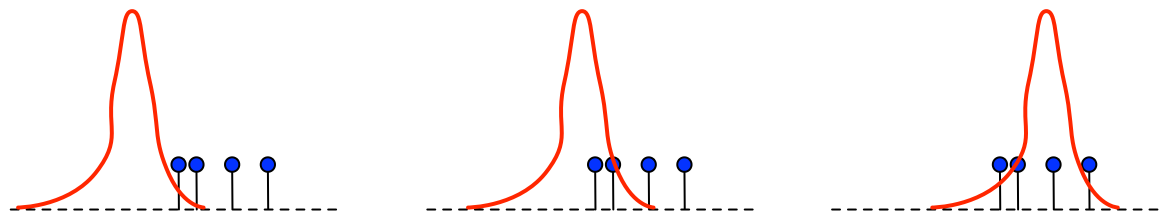
- Compares distributions taking geometric aspects in account.



- Polyvalent tool: compares continuous/atomic distributions



- Used as a loss $W(\alpha_\theta, \hat{\mu}_n)$ to fit between parametric α_θ and data $\hat{\mu}_n$



Optimal transport - statistics and ML

- Measures metric difference between random variables / datasets.
- Many applications in **statistics** and **machine learning** Peyré and Cuturi (18)
  - Wasserstein GANs Arjovsky et al. (17)
  - Wasserstein Autoencoders Tolstikhin et al. (18)
- Minimization of loss: Wasserstein variational problems

$$\min_{\theta \in \Theta} W_p(\alpha_\theta, \mu) \qquad \min_{\nu} \frac{1}{K} \sum_{i=1}^{K} W_p^p(\nu, \mu^{(i)}) \,.$$

  - Minimum Kantorovich estimators Bassetti et al. (06)
  - Wasserstein Barycenters Agueh and Carlier (11)
- In practice μ or $\mu^{(i)}$s based on samples, empirical $\hat{\mu}_n = (1/n) \sum_{j=1}^{n} \delta_{X_j}$
- When $n \to \infty$, over compact spaces $W_p(\mu, \hat{\mu}_n) \to 0$.

Unsupervised alignment of embeddings:

Wasserstein Procrustes

E. Grave (Facebook AI Research)

A. Joulin (Facebook AI Research)

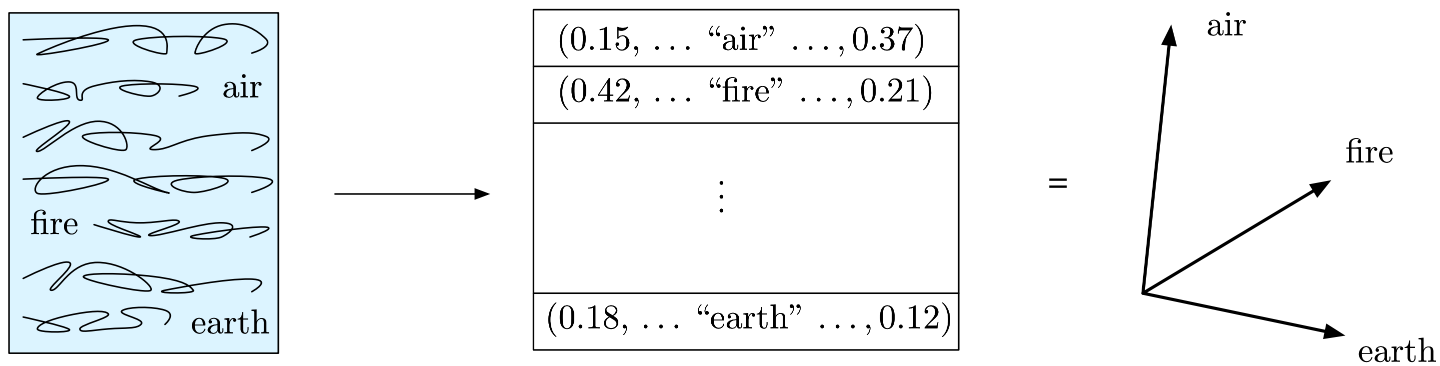
- **Unsupervised alignment of embeddings with Wasserstein Procrustes**

  E. Grave, A. Joulin, Q.B.

  AIStats 2019

Word embeddings

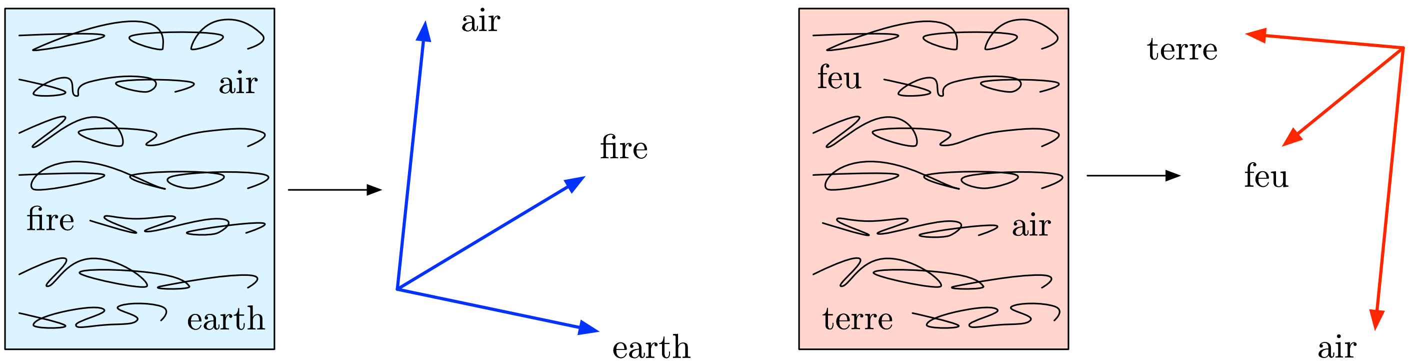
- Vectors representing words, obtained in data-driven manner from corpus



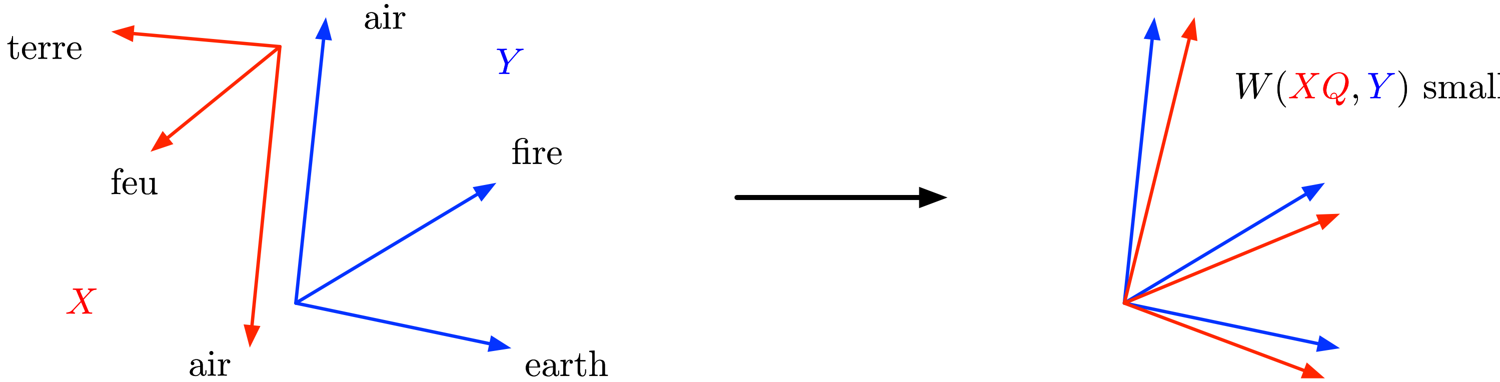
- Word embeddings with fastText: similar to word2vec with n-gram information.
- Obtained from wikipedia pages in several languages.
- Loss is invariant by rotation, relies on relative placement of vectors.

Word embeddings alignment

- Different corpora in different languages yield embeddings $X, Y \in \mathbf{R}^X$

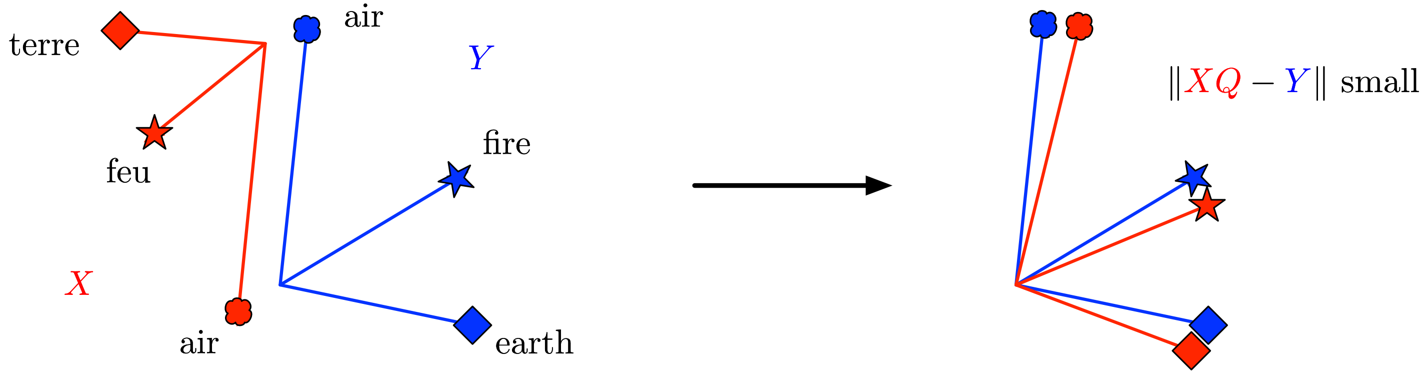


- **Embedding alignment:** Transformation $Q \in \mathcal{O}_d$ matching elements of XQ, Y



Embedding alignment

- **Supervised alignment:** Transformation $Q \in \mathcal{O}_d$ fitting $XQ \approx Y$.



- **Procrustes:** Closed form solution with SVD of X and Y, gradients

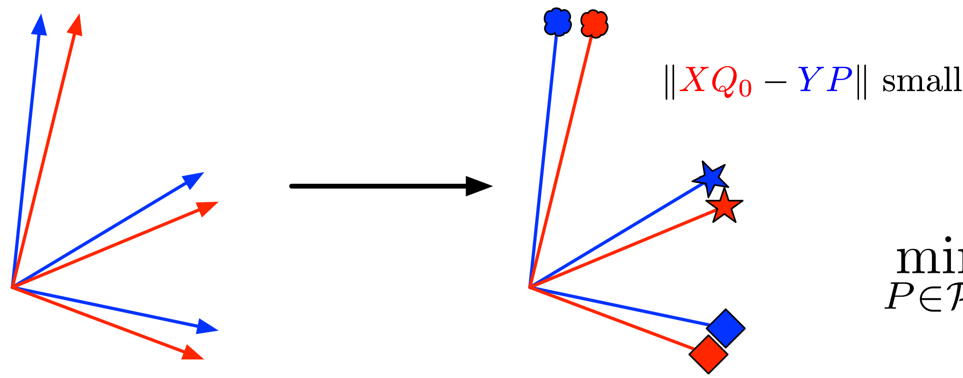
$$\min_{Q \in \mathcal{O}_d} \|XQ - Y\|^2 \quad \text{with} \quad Q^* = UV^\top \quad \text{for } X^\top Y = USV^\top .$$

Requires an existing dictionary, unreasonable expectation in many applications.

Sometimes finding the correspondence is the objective: point registration.

Unsupervised embedding alignment: Wasserstein Procrustes

- **Correspondence:** Once aligned Q_0, finding $P \in \mathcal{P}_n$ such that $XQ_0 \approx YP$.



$$\min_{P \in \mathcal{P}_n} \|XQ_0 - YP\|^2 = W_2^2(XQ_0, Y)$$

Equivalent to assignment problem (OT), minimum distance = Wasserstein

- **Wasserstein Procrustes:** Optimizing jointly alignment and correspondance

$$\min_{P \in \mathcal{P}_n} \min_{Q \in \mathcal{O}_d} \|XQ - YP\|^2 = \min_{Q \in \mathcal{O}_d} W_2^2(XQ, Y)$$

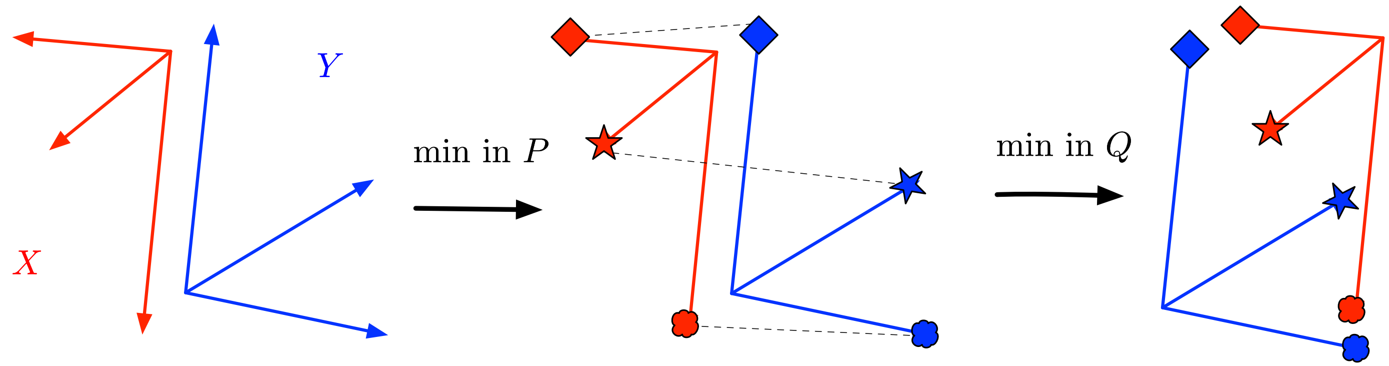
Equivalent to Wasserstein loss minimization between XQ and Y

Gold and Rangarajan (96), Zhang et al. (17).

No joint convexity, problem computationally NP-hard.

Wasserstein Procrustes

- **Alternated minimization:** Solving each min. problem, iteratively



Requires a large number of initializations, slow convergence. Zhang et al (17)

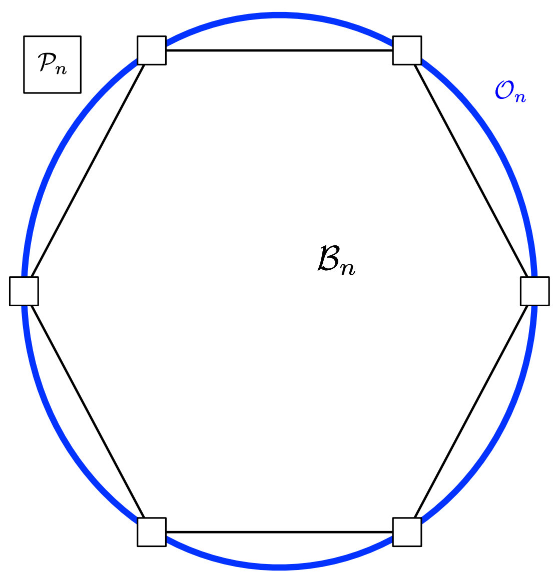
- **Related work:** Other approaches to alignment and Wassertein minimization.
  - Minimization with other techniques Conneau et al. (17), Artetxe et al. (18)
  - Regularization with entropic penalty Alvarez-Melis et al. (19)

Wasserstein Procrustes - our approach

- **Symmetry exploitation:** Gram matrix $K_X = XX^\top = (XQ)(XQ)^\top$
  - Finding row/column permutation P between $K_X = XX^\top$ and $K_Y = YY^\top$.
  - Permutation not fooled by initial local placement of X and Y.

$$\min_{P \in \mathcal{P}_n} \|K_X - PK_YP^\top\|_2^2 = \min_{P \in \mathcal{P}_n} \|K_XP - PK_Y\|_2^2$$

  - Convex relaxation, over the Birkhoff polytope (convex hull of permutations).



- $\mathcal{P}_n = \mathcal{B}_n \cap \mathcal{O}_n$, exact quadratic reformulation.
- Gromov-Wasserstein problem.
- Relaxation over convex hull $\mathcal{B}_n$

$$\min_{P \in \mathcal{B}_n} \|K_XP - PK_Y\|_2^2$$

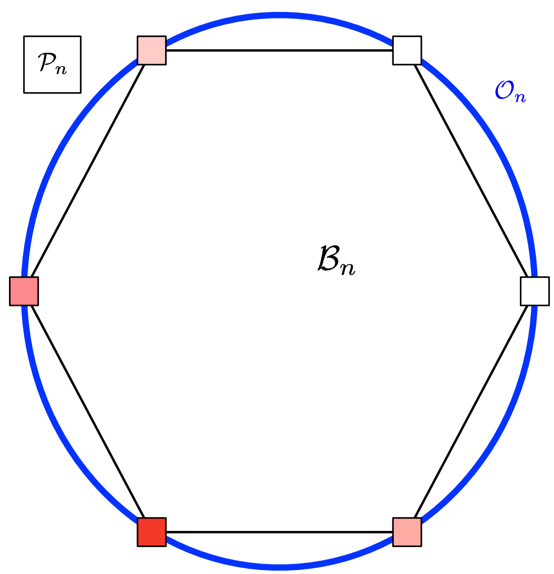
- Exact* for identical* clouds of vectors.

Wasserstein Procrustes - our approach

- **Symmetry exploitation:** Gram matrix $K_X = XX^\top = (XQ)(XQ)^\top$
  - Finding row/column permutation P between $K_X = XX^\top$ and $K_Y = YY^\top$.
  - Permutation not fooled by initial local placement of X and Y.

$$\min_{P \in \mathcal{P}_n} \|K_X - PK_YP^\top\|_2^2 = \min_{P \in \mathcal{P}_n} \|K_XP - PK_Y\|_2^2$$

  - Convex relaxation, over the Birkhoff polytope (convex hull of permutations).



- $\mathcal{P}_n = \mathcal{B}_n \cap \mathcal{O}_n$, exact quadratic reformulation.
- Gromov-Wasserstein problem.
- Relaxation over convex hull $\mathcal{B}_n$

$$\min_{P \in \mathcal{B}_n} \|K_XP - PK_Y\|_2^2$$

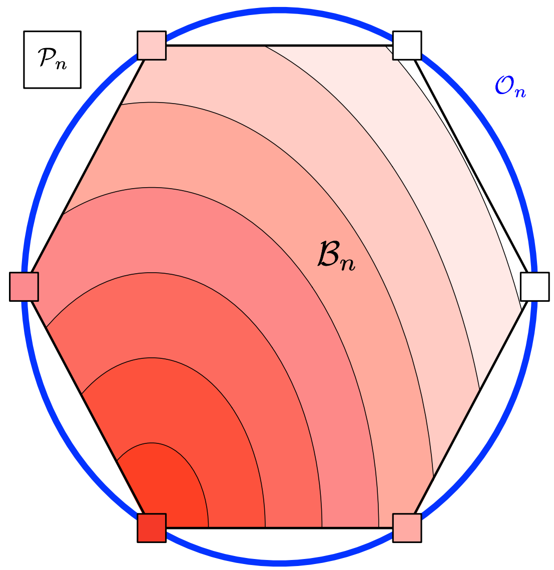
- Exact* for identical* clouds of vectors.

Wasserstein Procrustes - our approach

- **Symmetry exploitation:** Gram matrix $K_X = XX^\top = (XQ)(XQ)^\top$
  - Finding row/column permutation P between $K_X = XX^\top$ and $K_Y = YY^\top$.
  - Permutation not fooled by initial local placement of X and Y.

$$\min_{P \in \mathcal{P}_n} \|K_X - PK_Y P^\top\|_2^2 = \min_{P \in \mathcal{P}_n} \|K_X P - PK_Y\|_2^2$$

  - Convex relaxation, over the Birkhoff polytope (convex hull of permutations).



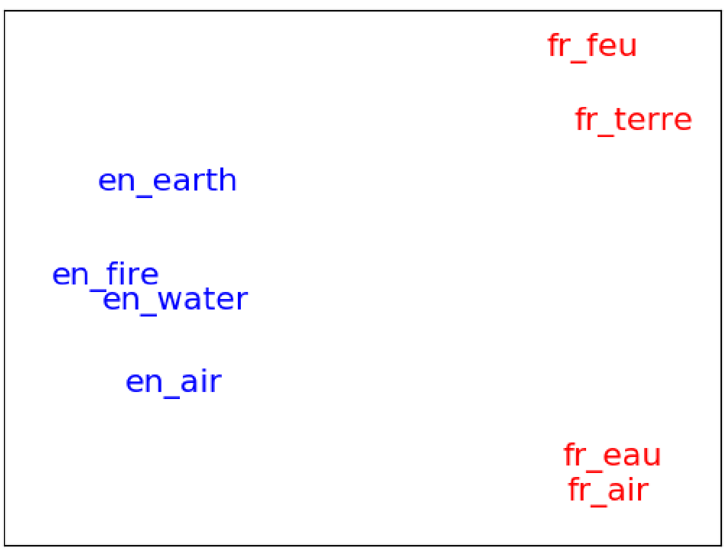
- $\mathcal{P}_n = \mathcal{B}_n \cap \mathcal{O}_n$, exact quadratic reformulation.
- Gromov-Wasserstein problem.
- Relaxation over convex hull $\mathcal{B}_n$

$$\min_{P \in \mathcal{B}_n} \|K_X P - PK_Y\|_2^2$$

- Exact* for identical* clouds of vectors.

Wasserstein Procrustes - full algorithm

- **Two central ideas**
  - Initialize (P_0, Q_0) with convex relaxation, not sensitive to relative placement.
  - Use mini-batches of vectors at each step: stochastic optimization.



For four words, before alignment

Algorithm 1 Stochastic optimization

for $t = 1$ to T **do**

Draw $\mathbf{X}_t, \mathbf{Y}_t$ from $\mathbf{X}, \mathbf{Y}$, of size b

Optimal matching $\mathbf{P}_t$ between $\mathbf{X}_t\mathbf{Q}_t$ and $\mathbf{Y}_t$

$$\mathbf{P}_t = \operatorname*{argmax}_{\mathbf{P} \in \mathcal{P}_b} \mathbf{Tr}\, \mathbf{Y}_t \mathbf{Q}_t^\top \mathbf{X}_t^\top \mathbf{P}.$$

Gradient $\mathbf{G}_t$ with respect to $\mathbf{Q}$:

$$\mathbf{G}_t = -2\mathbf{X}_t^\top \mathbf{P}_t \mathbf{Y}_t.$$

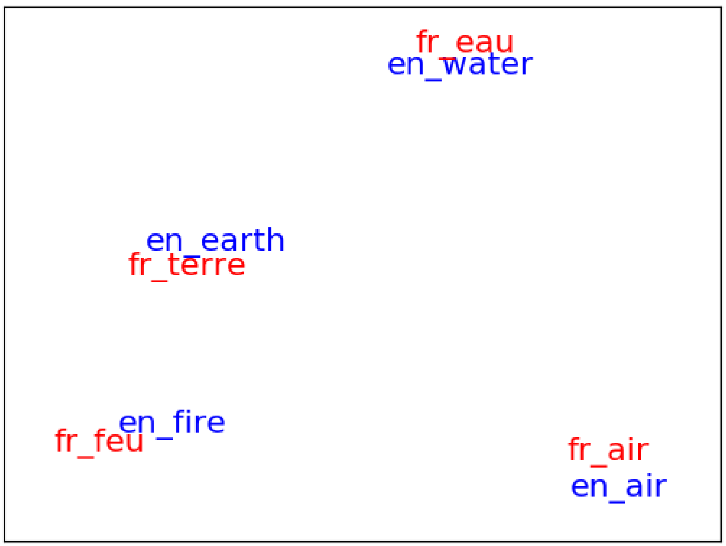
Projected gradient step:

$$\mathbf{Q}_{t+1} = \Pi_{\mathcal{O}_d}\left(\mathbf{Q}_t - \alpha \mathbf{G}_t\right).$$

end for

Wasserstein Procrustes - full algorithm

- **Two central ideas**
  - Initialize (P_0, Q_0) with convex relaxation, not sensitive to relative placement.
  - Use mini-batches of vectors at each step: stochastic optimization.



For four words, after alignment

Algorithm 2 Stochastic optimization

for $t = 1$ to T **do**

  Draw $\mathbf{X}_t, \mathbf{Y}_t$ from $\mathbf{X}, \mathbf{Y}$, of size b

  Optimal matching $\mathbf{P}_t$ between $\mathbf{X}_t\mathbf{Q}_t$ and $\mathbf{Y}_t$

$$\mathbf{P}_t = \operatorname*{argmax}_{\mathbf{P} \in \mathcal{P}_b} \mathbf{Tr}\,\mathbf{Y}_t \mathbf{Q}_t^\top \mathbf{X}_t^\top \mathbf{P}.$$

  Gradient $\mathbf{G}_t$ with respect to $\mathbf{Q}$:

$$\mathbf{G}_t = -2\mathbf{X}_t^\top \mathbf{P}_t \mathbf{Y}_t.$$

  Projected gradient step:

$$\mathbf{Q}_{t+1} = \Pi_{\mathcal{O}_d}\left(\mathbf{Q}_t - \alpha \mathbf{G}_t\right).$$

end for

Results

- Embeddings obtained with fastText from wikipedia pages with 200k words.
- Alignement on 20k words, convex relaxation on 2.5k words.

	EN-ES	EN-FR	EN-DE	EN-RU
Procrustes	82.7	82.7	74.8	51.3
Adversarial*	81.7	82.3	74.0	44.0
Iterative Closest Point*	82.1	**82.3**	74.7	**47.5**
Gromov-Wasserstein	81.7	81.3	71.9	45.1
Ours	**82.8**	**82.3**	**75.6**	45.2

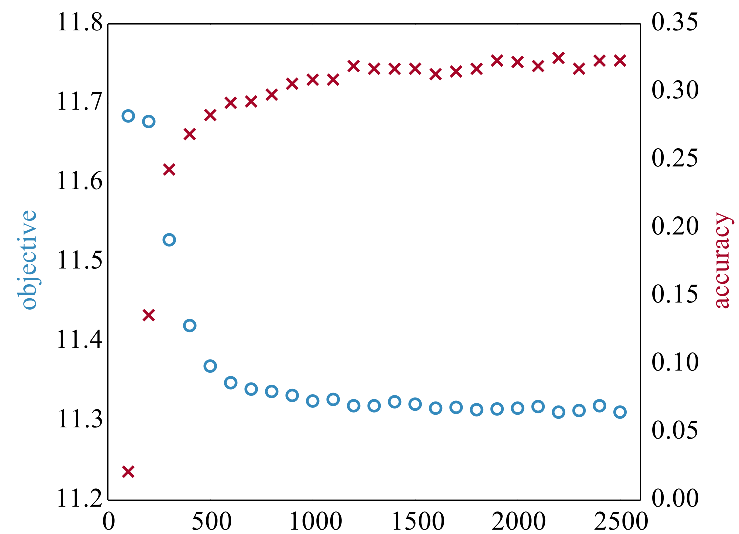
Comparison with supervised and unsupervised state-of-the-art approaches. In bold, the best among unsupervised methods, * indicates unnormalized vectors.

- Experiments on languages also indicate degree of proximity.

Discussion

- Stochastic element: correspondence between X_t and Y_t - suprising
- Can be interpreted as gradient step on $W_2^2(QX_t, Y_t)$, proxy for $W_2^2(QX, Y)$
- Empirical measure of X_t and Y_t: sampling from measure of X and Y
- Question of convergence $W_2(\hat{\mu}_b, \hat{\nu}_b)$ to $W_2(\mu, \nu)$.

	100	200	400	800	1600
Time	1m47s	2m07s	2m54s	5m34s	22m13s
EN-ES	68.5	73.8	74.9	75.0	76.3
EN-FR	67.4	71.9	74.5	75.6	75.7
EN-DE	59.1	63.0	64.4	65.8	66.4
EN-RU	23.7	27.9	29.9	32.3	33.2



Left: precision as a function of the batch size, 4k iterations.

Right: Accuracy and objective function value in EN-RU, batch-size 2k.

Merci