Optimal Transport

Instructor: John Thickstun

Discussion Board: Available on Ed

Zoom Link: Available on Canvas

Instructor Contact: thickstn@cs.washington.edu

Course Webpage: https://courses.cs.washington.edu/courses/cse599i/20au/

Kantorovich Rubinstein Duality

- Minimize Wasserstein distance between p and p_θ:

$$\theta_W = \arg\min_\theta W(p, p_\theta) = \arg\min_\theta \inf_{\pi \in \Pi(p, p_\theta)} \mathop{\mathbb{E}}_{(x,y)\sim\pi} \left[\|x - y\|_2 \right].$$

- $\Pi(p, q)$ is the set of probability distributions on $\mathcal{X} \times \mathcal{X}$ with marginals p, q.

- We can't enforce these constraints on the marginals.

- Instead, minimize a dual characterization of the Wasserstein distance:

$$W(p, q) = \inf_{\pi \in \Pi(p, q)} \mathop{\mathbb{E}}_{(x,y)\sim\pi} \left[\|x - y\|_2 \right] = \sup_{\|h\|_L \le 1} \left[\mathop{\mathbb{E}}_{x\sim p} h(x) - \mathop{\mathbb{E}}_{x\sim q} h(x) \right].$$

Gradient Penalty Relaxation

- Solve a saddle-point problem:

$$\theta_W = \arg\min_{\theta} \sup_{\varphi : \|h_\varphi\|_L \leq 1} \left[\mathop{\mathbb{E}}_{x\sim p} h_\varphi(x) - \mathop{\mathbb{E}}_{x\sim p_\theta} h_\varphi(x) \right].$$

- Idea: enforce $\|h_\varphi\|_L \leq 1$ as a soft constraint using Lagrange multipliers:

$$L(\theta, \varphi, \lambda) = \mathop{\mathbb{E}}_{x\sim p} h_\varphi(x) - \mathop{\mathbb{E}}_{x\sim p_\theta} h_\varphi(x) + \lambda \mathop{\mathbb{E}}_{x\sim ?} (\|\nabla_x h_\varphi(x)\| - 1)^2.$$

- Saddle point problem becomes $\theta_W^\lambda = \arg\min_\theta \sup_\varphi L(\theta, \varphi, \lambda)$.

- Technically need Lipschitz condition everywhere; where to enforce it?

- Uniformly along straight lines between points $x \sim p$ and $\tilde{x} \sim p_\theta$.

Wasserstein GAN

Algorithm 1 Wasserstein GAN

Initialize θ, φ.
while not converged **do**
  for t in range(1,num_critic) **do**
    Sample $x_1, \ldots, x_B \sim p$,
    Sample $y_1, \ldots, y_B \sim p_\theta$,

$$\varphi \leftarrow \varphi + \eta \nabla_\varphi \left(\frac{1}{B} \sum_{i=1}^{B} h_\varphi(x_i) - \frac{1}{B} \sum_{i=1}^{B} h_\varphi(y_i) + \lambda \text{Penalty}(\varphi) \right).$$

  end for
  Sample $z_1, \ldots, z_B \sim \mathcal{N}(0, I)$,

$$\theta \leftarrow \theta - \eta \nabla_\theta \frac{1}{B} \sum_{i=1}^{B} -h_\varphi(g_\theta(z_i)).$$

end while

GAN Evaluation

- For AR models and VAE, we calculated the test set log-likelihood.
- Can we do this for GAN's?

$$p_\theta(x) = r(g_\theta^{-1}(x))|\nabla_x g_\theta^{-1}(x)|.$$

- Inception score:
  - Using Inception v3 classifier $q(y|x)$, compute:

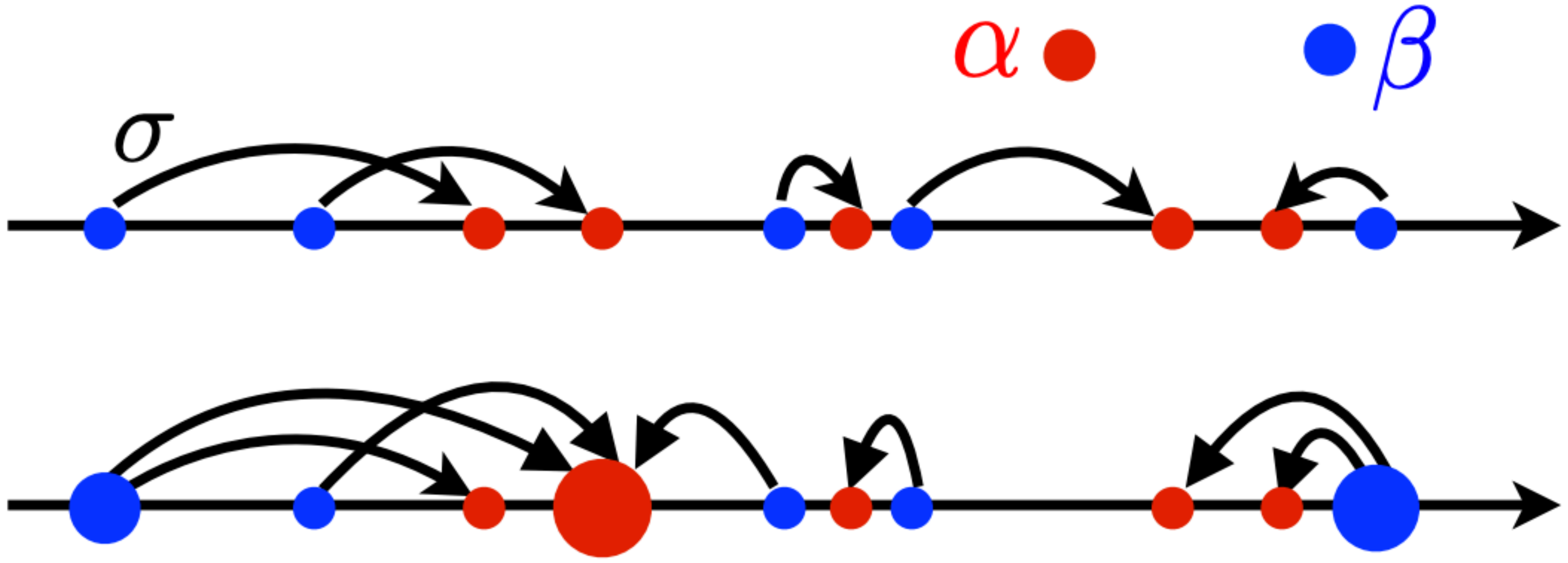
$$\mathrm{IS}(p) = \exp\left(\mathop{\mathbb{E}}_{x\sim p} D(q(y|x) \parallel q(y))\right)$$

  - Lower bound $\mathrm{IS}(p_\theta) = 1$; CIFAR-10 training data has $\mathrm{IS}(p) = 11.24.$
- Another popular variant is Frechet Inception Distance (FID).

The Monge Problem

- The optimal transport (sigma) maps blue probability mass onto red probability mass while minimizing the product:

  Distance x Mass



Peyré and Cuturi, 2019

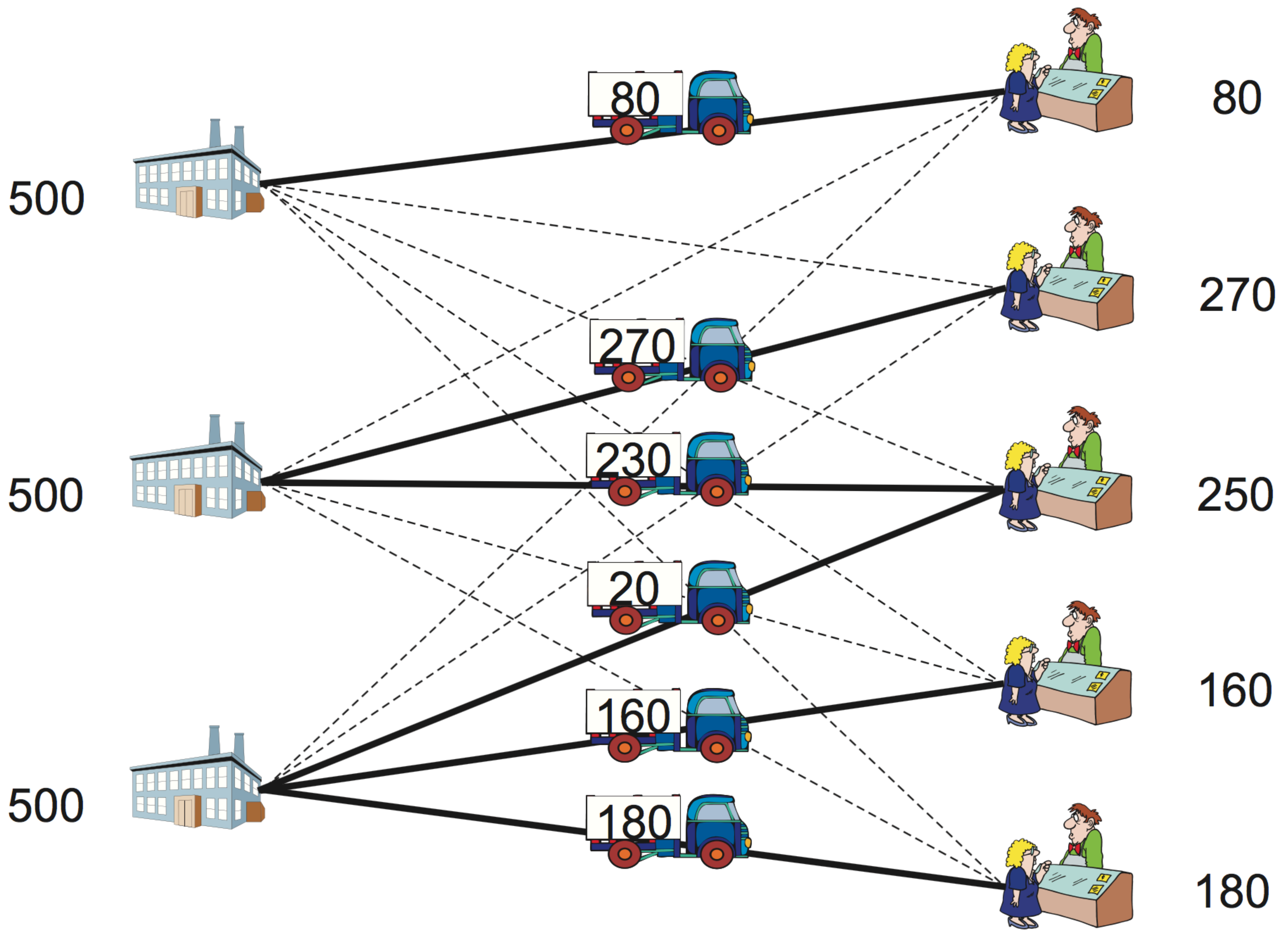
- The optimal transport sigma is called a “Monge map.”

- What if we can’t neatly pair up components of the blue and red distributions?

The Kantorovich Problem

- Relax the Monge map to a fractional map (i.e. a probabilistic map).
- This is a classical problem in logistics operations research!
- Formally, the problem is to find:

$$\arg\min_{\pi \in \Pi(p,q)} \sum_{x,y} c(x,y)\pi(x,y).$$



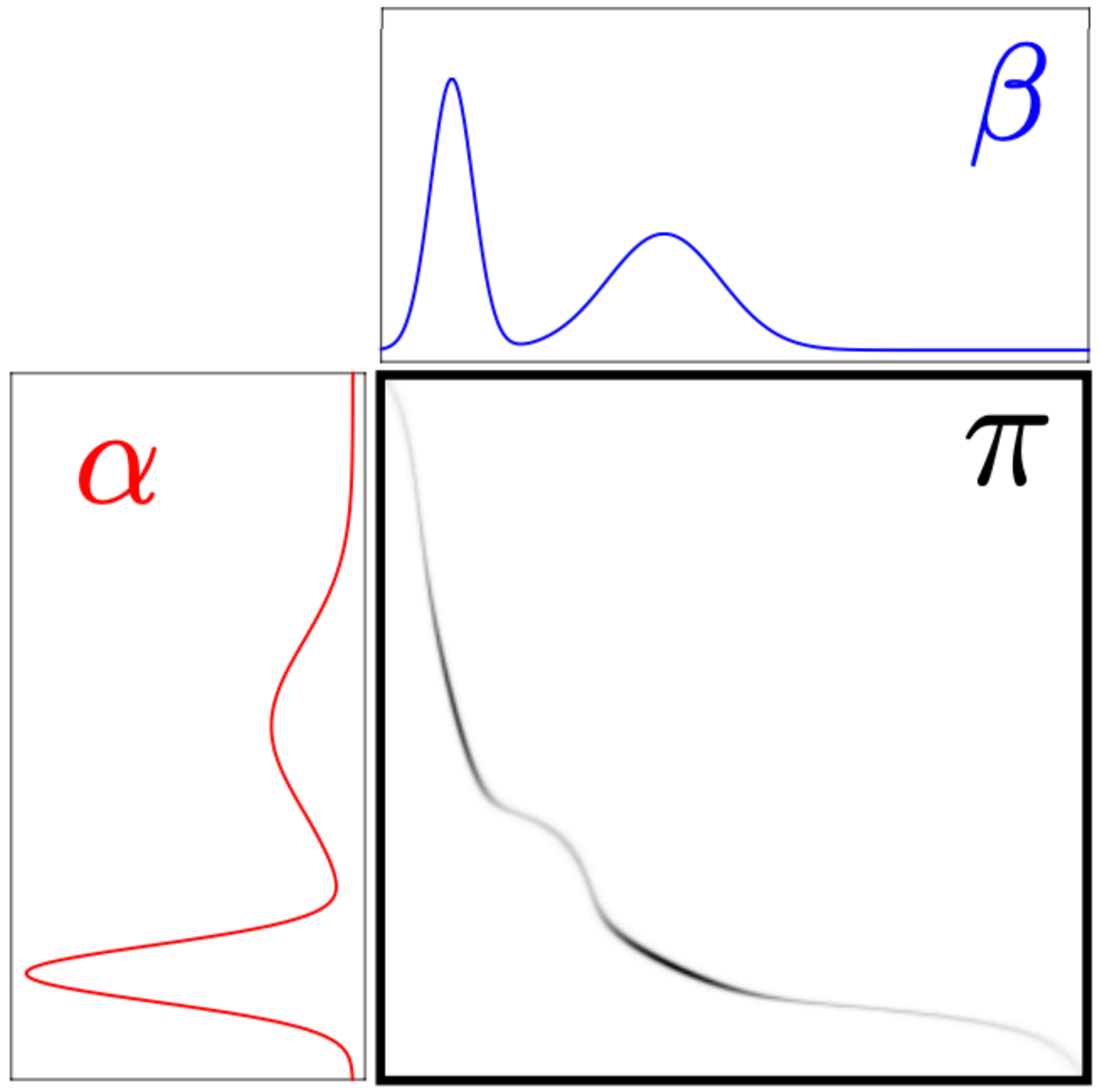
Pedroso, Rais, Kubo, and Muramatsu (2012)

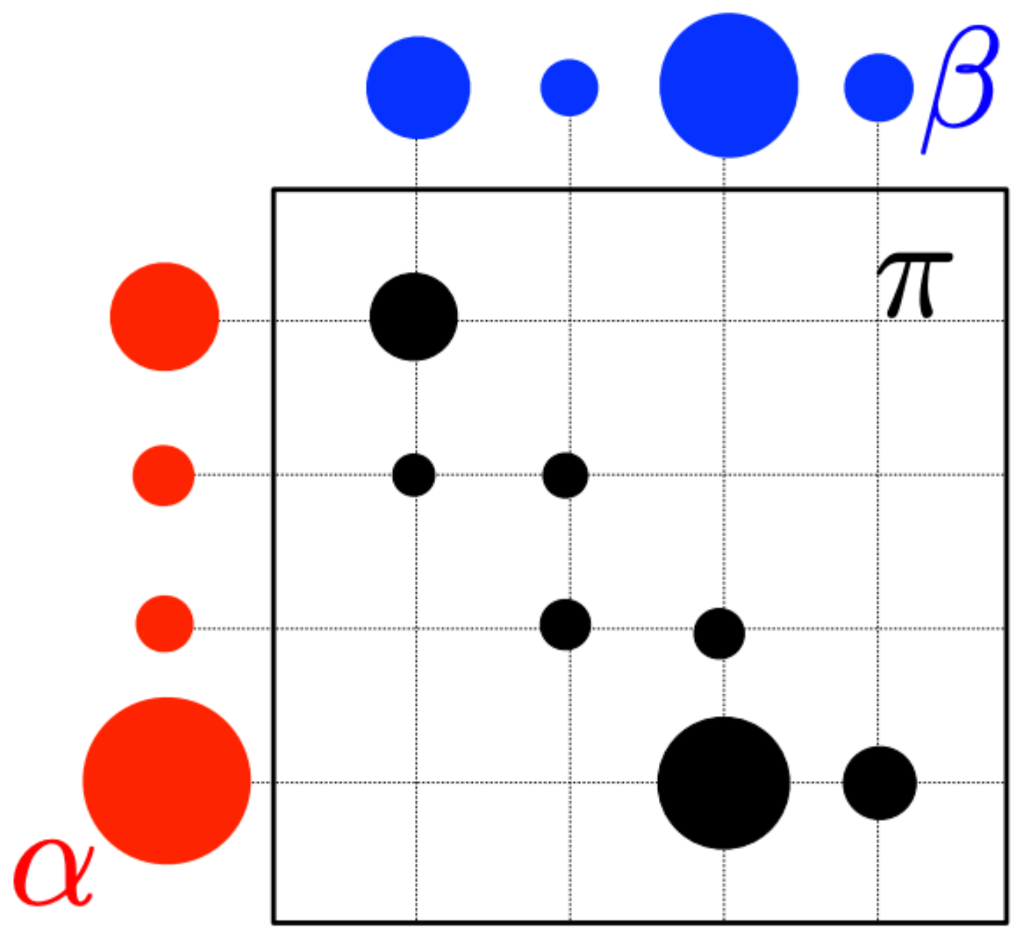
The Kantorovich Problem

- Want to find:

$$\underset{\pi \in \Pi(\alpha,\beta)}{\arg\min} \sum_{x,y} c(x,y)\pi(x,y).$$

- We can interpret the value of the minimizer as a distance between α and β.





$$\alpha, \beta \in \mathbb{R}^d, \text{ and } c, \pi \in \mathbb{R}^{d \times d}$$

Peyré and Cuturi, 2019

- Costs are not visualized here.

- If $x, y \in \mathbb{R}^d$ and $c(x,y) = \|x-y\|_2$ then

$$d_c(\alpha,\beta) = \min_{\pi \in \Pi(\alpha,\beta)} \sum_{x,y} \|x-y\|_2 \pi(x,y) = W_1(\alpha,\beta).$$

The Constraints

- We want to solve an optimal transport problem:

$$d_c(p, q) = \min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle = \min_{\pi \in \Pi(p,q)} \sum_{x,y} c(x, y)\pi(x, y).$$

- In the discrete setting, the constraints $\pi \in \Pi(p, q)$ become

$$p(x) = \sum_{i=1}^{m} \pi(x, y_i) = (\pi \mathbf{1}_m)_x, \text{ and } q(y) = \sum_{i=1}^{n} \pi(x_i, y) = (\mathbf{1}_n^T \pi)_y.$$

- The optimal transport problem is a linear program:

$$d_c(p, q) = \min_{\substack{\pi \mathbf{1}_m = p \\ \pi^\top \mathbf{1}_n = q}} \sum_{x,y} c(x, y)\pi(x, y).$$

Entropy Regularization

- We can compute an optimal transport faster if we regularize a bit:

$$d_c^\lambda(p,q) = \min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle - \lambda H(\pi).$$

- This problem has an interesting alternate interpretation:

$$\arg\min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle - \lambda H(\pi) = \arg\min_{\pi \in \Pi(p,q)} D(\pi \parallel p_k^\lambda).$$

- Analogous to how MLE is really KL minimization!

- Does the analogy run deeper?

KL Divergence Minimization

- Claim:

$$\arg\min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle - \lambda H(\pi) = \arg\min_{\pi \in \Pi(p,q)} D(\pi \parallel p_k^\lambda).$$

- Define $k(x,y) \equiv e^{-c(x,y)/\lambda}$ and $Z_\lambda = \sum_{x,y} k(x,y)$.
- Then $p_k^\lambda(x,y) \equiv \frac{1}{Z_\lambda} k(x,y)$ is a probability distribution and

$$D(\pi \parallel p_k^\lambda) = \sum_{x,y} \pi(x,y) \log \frac{\pi(x,y) Z_\lambda}{k(x,y)} = \frac{1}{\lambda} \langle c, \pi \rangle - H(\pi) + \log Z_\lambda.$$

5-Minute Break

A Conceptual Algorithm

- Our goal is to find:

$$\arg\min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle - \lambda H(\pi) = \arg\min_{\pi \in \Pi(p,q)} D(\pi \parallel p_k^\lambda).$$

- Two sets of constraints: $\pi \mathbf{1}_m = p$ and $\pi^\top \mathbf{1}_n = q$.

- How about alternating minimization? Initialize $\pi_\lambda^{(0)} = p_k^\lambda$ and iterate:

$$\pi_\lambda^{(\ell+1)} \equiv \begin{cases} \arg\min_{\pi \mathbf{1}_m = p} D(\pi \parallel \pi_\lambda^{(\ell)}) & \ell \text{ even}, \\ \arg\min_{\pi^\top \mathbf{1}_n = q} D(\pi \parallel \pi_\lambda^{(\ell+1)}) & \ell \text{ odd}. \end{cases}$$

- These sub-problems have a closed form!

Solving the Sub-Problems

- Suppose ℓ is even. We need to find $\arg\min_{\pi \mathbf{1}_m = p} D(\pi \parallel \pi_\lambda^{(\ell)})$.
- Introduce Lagrange multipliers:

$$\arg\min_{\pi \mathbf{1}_m = p} D(\pi \parallel \pi_\lambda^{(\ell)}) = \arg\min_{\pi} \max_{f} D(\pi \parallel \pi_\lambda^{(\ell)}) + \langle f, \pi \mathbf{1}_m - p \rangle.$$

- Appeal to strong duality to swap the min and max. We now need to find

$$\arg\min_{\pi} D(\pi \parallel \pi_\lambda^{(\ell)}) + \langle f, \pi \mathbf{1}_m - p \rangle.$$

- This is an unconstrained minimization problem.

Solving the Inner Optimization

- Want to find $\arg\min_{\pi} D(\pi \parallel \pi_\lambda^{(\ell)}) + \langle f, \pi \mathbf{1}_m - p \rangle$.

- Solve the first order optimality conditions:

$$\frac{\partial}{\partial \pi}\left[D(\pi \parallel \pi_\lambda^{(\ell)}) + \langle f, \pi \mathbf{1}_m - p \rangle\right] = 0.$$

- For a particular index (x,y), $1 + \log \pi_{\lambda,f}^{(\ell+1)}(x,y) - \log \pi_\lambda^{(\ell)}(x,y) + f_x = 0.$

- Solve for $\pi_{\lambda,f}^{(\ell+1)}(x,y)$:

$$\pi_\lambda^{(\ell+1)}(x,y) = e^{1-f_x}\pi_\lambda^{(\ell)}(x,y).$$

Solving the Outer Optimization

- Suppose ℓ is even. We need to find $\arg\min_{\pi \mathbf{1}_m = p} D(\pi \,\|\, \pi_\lambda^{(\ell)})$.

- Equivalent (strong duality) to finding

$$\arg\max_f \min_\pi D(\pi \,\|\, \pi_\lambda^{(\ell)}) + \langle f, \pi \mathbf{1}_m - p \rangle$$

$$= \arg\max_{f \,:\, \pi_{\lambda,f}^{(\ell+1)} \mathbf{1}_m = p} D(\pi_{\lambda,f}^{(\ell+1)} \,\|\, \pi_\lambda^{(\ell)}) + \langle f, \pi_{\lambda,f}^{(\ell+1)} \mathbf{1}_m - p \rangle.$$

- We previously saw that $\pi_\lambda^{(\ell+1)}(x,y) = e^{1-f_x} \pi_\lambda^{(\ell)}(x,y)$. So we need

$$p(x) = \sum_y \pi_{\lambda,f}^{(\ell+1)}(x,y) = e^{1-f_x} \sum_y \pi_\lambda^{(\ell)}(x,y) \implies e^{1-f_x} = \frac{p(x)}{\sum_y \pi_\lambda^{(\ell)}(x,y)}.$$

A Closed Form Solution

- Recall that our goal is to find $\arg\min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle - \lambda H(\pi) = \arg\min_{\pi \in \Pi(p,q)} D(\pi \parallel p_k^\lambda)$.
- We decomposed this into an alternating minimization problem:

$$\pi_\lambda^{(\ell+1)} \equiv \begin{cases} \arg\min_{\pi \mathbf{1}_m = p} D(\pi \parallel \pi_\lambda^{(\ell)}) & \ell \text{ even}, \\ \arg\min_{\pi^\top \mathbf{1}_n = q} D(\pi \parallel \pi_\lambda^{(\ell+1)}) & \ell \text{ odd}. \end{cases}$$

- The sub-problems have closed form:

$$\pi_\lambda^{(2\ell)} = \operatorname{diag}\left(\frac{p}{\pi_\lambda^{(2\ell-1)} \mathbf{1}_m}\right) \pi_\lambda^{(2\ell-1)}, \text{ and } \pi_\lambda^{(2\ell+1)} = \operatorname{diag}\left(\frac{q}{\mathbf{1}_n^\top \pi_\lambda^{(2\ell)}}\right) \pi_\lambda^{(2\ell)}.$$