Course Materials for GEN-AI

Northeastern University

These materials have been prepared and sourced for the course **GEN-AI** at Northeastern University. Every effort has been made to provide proper citations and credit for all referenced works.

If you believe any material has been inadequately cited or requires correction, please contact me at:

Instructor: Ramin Mohammadi
r.mohammadi@northeastern.edu

Thank you for your understanding and collaboration.

Wasserstein

The Wasserstein distance is defined as:

$$D_w\left[\Pr(x)\|q(x)\right] = \min_{\mathbf{P}}\left[\sum_{i,j} P_{ij}\cdot|i-j|\right], \tag{15.10}$$

subject to the constraints that:

$$\begin{aligned}\sum_j P_{ij} &= \Pr(x=i) \quad \text{(initial distribution of } \Pr(x)),\\ \sum_i P_{ij} &= q(x=j) \quad \text{(initial distribution of } q(x)),\\ P_{ij} &\geq 0 \quad \text{(non-negative masses)}.\end{aligned} \tag{15.11}$$

In other words, the Wasserstein distance is the solution to a constrained minimization problem that maps the mass of one distribution to the other. This is inconvenient as we must solve this minimization problem over the elements P_{ij} every time we want to compute the distance.

To rewrite the problem in matrix form, let:

$$\mathbf{p} = \text{vec}(\mathbf{P}), \quad \mathbf{c} = \text{vec}(\mathbf{C}),$$

where $\mathbf{p}$ and $\mathbf{c}$ are vectorized forms of $\mathbf{P}$ and $\mathbf{C}$. Define the constraint matrix $\mathbf{A}$ and vector $\mathbf{b}$ such that:

$$\mathbf{A}\mathbf{p} = \mathbf{b},$$

where:

$$\mathbf{b} = \begin{bmatrix}\Pr(x=1)\\ \Pr(x=2)\\ \vdots\\ q(x=1)\\ q(x=2)\\ \vdots\end{bmatrix} \quad \mathbf{p} = \begin{bmatrix}P_{11}\\ P_{12}\\ \vdots\\ P_{1K}\\ P_{21}\\ P_{22}\\ \vdots\\ P_{2K}\\ \vdots\\ P_{K1}\\ P_{K2}\\ \vdots\\ P_{KK}\end{bmatrix} \quad \text{vec}(\mathbf{C}) = \begin{bmatrix}C_{11}\\ C_{12}\\ \vdots\\ C_{1K}\\ C_{21}\\ C_{22}\\ \vdots\\ C_{2K}\\ \vdots\\ C_{K1}\\ C_{K2}\\ \vdots\\ C_{KK}\end{bmatrix}.$$

The matrix $\mathbf{A}$ encodes the row and column constraints. It has dimensions $(2K)\times K^2$, where the first K rows correspond to the row constraints, and the next K rows correspond to the column constraints.

- **Row Marginal Constraints** ($\sum_j P_{ij} = \Pr(x = i)$):

$$\mathbf{A}_{\text{rows}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

- **Column Marginal Constraints** ($\sum_i P_{ij} = q(x = j)$):

$$\mathbf{A}_{\text{columns}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & \cdots & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix}.$$

Combine these into:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\text{rows}} \\ \mathbf{A}_{\text{columns}} \end{bmatrix}.$$

The primal problem becomes:

$$\min_{\mathbf{p}} \mathbf{c}^\top \mathbf{p},$$

$$\text{subject to } \mathbf{A}\mathbf{p} = \mathbf{b}, \quad \mathbf{p} \geq 0.$$

Lagrangian Formulation

Introduce the Lagrange multipliers $\mathbf{f}$ for the equality constraint $\mathbf{A}\mathbf{p} = \mathbf{b}$. The Lagrangian is:

$$\min_{\mathbf{p}} \max_{\mathbf{f}} L(\mathbf{p}, \mathbf{f}) = \mathbf{c}^\top \mathbf{p} + \mathbf{f}^\top (\mathbf{b} - \mathbf{A}\mathbf{p}).$$

Simplify the Lagrangian:

$$\min_{\mathbf{p}} \max_{\mathbf{f}} L(\mathbf{p}, \mathbf{f}) = \mathbf{f}^\top \mathbf{b} + \mathbf{p}^\top (\mathbf{c} - \mathbf{A}^\top \mathbf{f}).$$

To ensure L is finite as we are minimizing w.r.t p, we require:

$$\mathbf{c} - \mathbf{A}^\top \mathbf{f} \geq 0.$$

Dual Problem

Minimizing $L(\mathbf{p}, \mathbf{f})$ with respect to $\mathbf{p}$ gives:

$$L(\mathbf{p}, \mathbf{f}) = \mathbf{f}^\top \mathbf{b}, \quad \text{subject to } \mathbf{c} - \mathbf{A}^\top \mathbf{f} \geq 0.$$

Thus:

Primal Form	**Dual Form**
minimize $\mathbf{c}^T\mathbf{p}$, such that $\mathbf{A}\mathbf{p} = \mathbf{b}$, $\mathbf{p} \geq 0$	maximize $\mathbf{b}^T\mathbf{f}$, such that $\mathbf{A}^T\mathbf{f} \leq \mathbf{c}$

Interpreting the Dual Problem

Returning to the original Wasserstein problem, the dual variables $\mathbf{f}$ correspond to potentials f_i and f_j for the bins i and j. The dual constraint $\mathbf{A}^\top \mathbf{f} \leq \mathbf{c}$ implies:

$$f_i - f_j \leq C_{ij}, \quad \forall i, j.$$

The dual objective becomes:

$$D_w [\Pr(x) \| q(x)] = \max_{\mathbf{f}} \left[\sum_i \Pr(x = i) f_i - \sum_j q(x = j) f_j \right]$$

where f_i and f_j satisfy the constraints