The tractable formulation for a piecewise affine loss function $L(z) = \max_{j \in J} (a_j z + b_j)$

is given below :

$$\inf_{w} \sup_{Q \in B_\varepsilon(\hat{P}_N)} E^Q[\ell(\langle w, x\rangle, y)]$$

$$= \boxed{\begin{aligned} &\inf_{\substack{w, \lambda, s_i \\ p_{ij}^+, p_{ij}^-}} \quad \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i \\ &\text{s.t.} \\ &\quad s_x (a_j \hat{y}_i w - p_{ij}^+) + b_j + \langle p_{ij}^+, \hat{x}_i \rangle \le s_i \qquad i \in [N],\ j \in [J] \\ &\quad s_x (-a_j \hat{y}_i w - p_{ij}^-) + b_j + \langle p_{ij}^-, \hat{x}_i \rangle - K\lambda \le s_i \qquad j \in [J],\ i \in [N] \\ &\quad \| p_{ij}^- \|_* \le \lambda \\ &\quad \| p_{ij}^+ \|_* \le \lambda \\ &\quad s_i \ge 0 \end{aligned}}$$

Here $B_\varepsilon[\hat{P}_N]$ is the wasserstein ball centered around the emperical distribution, $\hat{P}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{\xi}_i}(\xi)$ and radius is ε

$\|\cdot\|_* \to$ dual Norm

$(\hat{x}_i, y_i) \to$ datapoint $\in \Xi = X \times \{-1, 1\} = Y$

$K \to$ cost of mislabelling

$\varepsilon \to$ wasserstein radius.

Proof For simplicing let us begin by considering the inner problem

$$\sup_{Q \in B_\varepsilon(\hat{P}_N)} E^Q[\ell(\langle w, x\rangle, y)]$$

$$= \sup_{\Pi} \int_{\Xi^2} \ell(\underbrace{\langle w, x\rangle, y}_{\xi}) \, \Pi(d\xi, d\xi')$$

s.t Π is a joint distribution of ξ and ξ' with marginals Q and $\hat{P}_N$

$$\int_{\Xi^2} d(\xi, \xi') \, \Pi(d\xi, d\xi') \leq \varepsilon$$

→ Here we used the defination of the wasserstein metric and ball and expanded the expectation.

Since we know that ξ' has marginal defined by the emperical distribution. We can make use of this

$$\Pi(d\xi, d\xi') = \mathbb{P}(d\xi) \cdot P(d\xi \mid d\xi')$$

$$= \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi'_i}(d\xi') \cdot Q^i(d\xi)$$

$Q^i(d\xi)$ = conditional p.d for $d\xi$ given $\xi' = \xi_i$

DATE:

From the above

$$\sup_{Q^i} \frac{1}{N} \sum_{i=1}^{N} \int_{\Xi} \ell(\xi) \, Q^i(d\xi)$$

$$\text{s.t} \quad \int_{\Xi} Q^i(d\xi) = 1 \qquad i \in [N]$$

$$\frac{1}{N} \sum_{i=1}^{N} \int_{\Xi} d(\xi, \hat{\xi}_i) \, Q^i(d\xi) < \varepsilon$$

Strong duality can be shown to hold for $\varepsilon > 0$
Hence we have the dual as.

$$\sup_{Q \in B_{\varepsilon}(\hat{P}_N)} E^{Q}[\ell(\xi)]$$

$$\boxed{\begin{aligned} = \inf_{\lambda, s_i} \quad & \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i \\ \text{st} \quad & \sup_{\xi \in \Xi} \ell(\xi) - \lambda d(\xi, \xi') \leq s_i \qquad i \in [N] \\ & \lambda \geqslant 0 \end{aligned}}$$

We define the transportation metric d as

$$d((x,y),(x',y')) = \|x - x'\| + \frac{K}{2} |y - \hat{y}|$$

where $\|\cdot\|$ is a norm in the input space
and K is the cost of switching a label.

Henceforth we will also consider

$\xi = (x, y)$

$$= \inf_{\lambda, s_i} \quad \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i$$

s.t

$$\sup_{(x,y) \in \Xi} \quad \ell(\langle w, x \rangle, y) - \lambda \| x - \hat{x}_i \| + \frac{K\lambda}{2} | y - \hat{y}_i | \le s_i \qquad i \in [N]$$

$$\lambda \geq 0$$

- Since we are dealing with a classification problem we have $y \in \{1, -1\}$.

We consider the two scenarios possible $y = \hat{y}_i$ or $y = -\hat{y}_i$

also $\ell(\langle w, x \rangle, y) = \ell(y w^T x)$ is used

Hence we can write the new constraints as

$$= \inf_{\lambda s_i} \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i$$

$$\sup_{x \in X} \ell(\hat{y}_i w^T x) - \lambda \| x - \hat{x}_i \| \le s_i \qquad i \in [N]$$

$$\sup_{x \in X} \ell(-\hat{y}_i w^T x) - \lambda \| x - \hat{x}_i \| + K\lambda \le s_i \qquad i \in [N]$$

$$\lambda \geq 0$$

Now using the defination of the piecewise affine function

$\ell(z) = \max\limits_{i \in J} (a_j z + b_j)$ we have the following

$$= \inf_{\lambda, s_i} \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i$$

$$\text{s.t} \quad \sup_{x \in X} a_j \hat{y}_i \langle w, x \rangle + b_j - \lambda \| x - \hat{x}_i \| \leq s_i \qquad i \in [N], j \in [J]$$

$$\sup_{x \in X} -a_j \hat{y}_i \langle w, x \rangle + b_j - \lambda \| x - \hat{x}_i \| + K\lambda \leq s_i \qquad i \in [N], j \in [J]$$

• Using the definition of the dual Norm we introduce p_{ij}^+ and p_{ij}^-

$$\lambda \| x - \hat{x}_i \| = \inf_{\underbrace{\| p_{ij}^+ \|_* \leq \lambda}_{\text{dual Norm}}} \langle p_{ij}^+, x - \hat{x}_i \rangle$$

→ Let us consider some simplification to the 1st constraint

$$\sup_{x \in X} a_j \hat{y}_i \langle w, x \rangle + b_j - \lambda \| x - \hat{x}_i \| \leq s_i$$

with dual Norm

$$\Rightarrow \sup_{x \in X} a_j \hat{y}_i \langle w, x \rangle + b_j - \inf_{\| p_{ij}^+ \|_* \leq \lambda} \langle p_{ij}^+, x - \hat{x}_i \rangle \leq s_i$$

Switching the sup and inf.

$$= \inf_{\|p_{ij}\|^* \leq \lambda} \sup_{x \in X} a_j \hat{y}_i \langle w, x \rangle - \langle p_{ij}{}^{+}, x - \hat{x}_i \rangle + b_j \leq s_i$$

rearranging the terms

$$= \inf_{\|p_{ij}{}^{+}\|^* \leq \lambda} \sup_{x \in X} \langle a_j \hat{y}_i w - p_{ij}{}^{+}, x \rangle + \langle p_{ij}{}^{+}, \hat{x}_i \rangle + b_j \leq s_i$$

We introduce δ_X which the indicator for the set X

$$\delta_X(t) = \begin{cases} 0 & t \in X \\ \infty & t \notin X \end{cases}$$

$$= \inf_{\|p_{ij}{}^{+}\|^* \leq \lambda} \sup_{x \in \mathbb{R}^d} \left(\langle a_j \hat{y}_i w - p_{ij}{}^{+}, x \rangle + \delta_X(x) \right) + \langle p_{ij}{}^{+}, \hat{x}_i \rangle + b_j \leq s_i$$

- The defination of support function on the set X can be used (It is the conjugate of the Indicator)

$$S_X(t) = \sup_{x \in \mathbb{R}^d} \langle t, x \rangle - \delta_X(x)$$

$$= \inf_{\|p_{ij}\|^* \leq \lambda} S_X(a_j \hat{y}_i w - p_{ij}{}^{+}) + b_j + \langle p_{ij}{}^{+}, \hat{x}_i \rangle \leq s_i$$

Similarly we can do the same for the second constraint.

$$\Rightarrow \inf_{\|p_{ij}^-\|_* \le \lambda} S_x(-a_j \hat{y}_i w - p_{ij}^-) + b_j + \langle p_{ij}^-, \hat{x}_i \rangle - K\lambda \le s_i$$

The complete formulation now is:

$$\boxed{\begin{aligned}
&\inf_{\lambda, s_i} \quad \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i \\
&\text{s.t} \quad \inf_{\|p_{ij}^+\|_* \le \lambda} S_x(a_j \hat{y}_i w - p_{ij}^+) + b_j + \langle p_{ij}^+, \hat{x}_i \rangle \le s_i \qquad i \in [N],\ j \in [J] \\
&\inf_{\|p_{ij}^-\|_* \le \lambda} S_x(-a_j \hat{y}_i w - p_{ij}^-) + b_j + \langle p_{ij}^-, \hat{x}_i \rangle - K\lambda \le s_i
\end{aligned}}$$

- Now we consider the $\inf_{w} \sup_{Q \in B_\varepsilon(\hat{P}_N)} E^Q(\ell(\xi))$ complete form and include the decision variable for w. For the constraints we can argue that that if $\exists\ p_{ij}^+$ s.t $\|p_{ij}^+\|_* \le \lambda$ and satisfies the 1st constraint it holds for the inf case of p_{ij} as well. Therefore we can do away with the inf and add p_{ij}^+ and p_{ij}^- as a decision variable.

The final formulation we get is,

$$\inf_{w} \sup_{Q \in B_\varepsilon[\hat{P}_N]} E^Q[\ell(\xi)]$$

$$= \inf_{\substack{w, \lambda, s_i \\ p_{ij}^+, p_{ij}^-}} \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i$$

s.t

$$\left.\begin{aligned} & S_x(a_j \hat{y}_i w - p_{ij}^+) + b_j + \langle p_{ij}^+, \hat{x}_i \rangle \leq s_i \\ & S_x(-a_j \hat{y}_i w - p_{ij}^-) + b_j + \langle p_{ij}^-, \hat{x}_i \rangle - \kappa\lambda \leq s_i \end{aligned}\right\} \begin{matrix} i \in [N \\ j \in [\end{matrix}$$

$$\| p_{ij}^+ \|_* \leq \lambda$$

$$\| p_{ij}^- \|_* \leq \lambda$$

$$s_i \geq 0$$

Hence Proved $\square$