A Proofs

Lemma 1. *For any normalized probability distributions $\boldsymbol{x}, \boldsymbol{x}' \in [0,1]^{n \times m}$, there exists at least one $\boldsymbol{\delta}$ such that $\boldsymbol{x}' = \Delta(\boldsymbol{x}, \boldsymbol{\delta})$. Furthermore:*

$$\min_{\boldsymbol{\delta}:\, \boldsymbol{x}' = \Delta(\boldsymbol{x}, \boldsymbol{\delta})} \|\boldsymbol{\delta}\|_1 = W_1(\boldsymbol{x}, \boldsymbol{x}') \tag{22}$$

Where W_1 denotes the 1-Wasserstein metric, using L_1 distance as the underlying distance metric.

Proof. We first show the equivalence of the above minimization problem with the linear program proposed by Ling and Okada (2007), restated here:

$$W_1(\boldsymbol{x}, \boldsymbol{x}') = \min_{\mathbf{g}} \sum_{(i,j)} \sum_{(i',j') \in \mathcal{N}((i,j))} \mathbf{g}_{(i,j),(i',j')} \tag{23}$$

where $\mathbf{g} \geq 0$ and $\forall (i,j)$,

$$\sum_{(i',j') \in \mathcal{N}((i,j))} \mathbf{g}_{(i,j),(i',j')} - \mathbf{g}_{(i',j'),(i,j)} = \mathbf{x}'_{i,j} - \mathbf{x}_{i,j}$$

It suffices to show that (1) there is a transformation from the variables $\mathbf{g}$ in Equation 23 to the variables $\boldsymbol{\delta}$ in Equation 22, such that all points which are feasible in Equation 23 are feasible in 22 and the minimization objective in Equation 22 is less than or equal to the minimization objective in Equation 23, and (2) there is a transformation from the variables $\boldsymbol{\delta}$ in Equation 22 to the variables $\mathbf{g}$ in Equation 23, such that all points which are feasible in Equation 22 are feasible in Equation 23 and the minimization objective in Equation 23 is less than or equal to the minimization objective in Equation 22.

We start with (1). We give the transformation as:

$$\begin{aligned} \boldsymbol{\delta}^{\text{vert.}}_{i,j} &:= \mathbf{g}_{(i,j),(i+1,j)} - \mathbf{g}_{(i+1,j),(i,j)} \\ \boldsymbol{\delta}^{\text{horiz.}}_{i,j} &:= \mathbf{g}_{(i,j),(i,j+1)} - \mathbf{g}_{(i,j+1),(i,j)} \end{aligned} \tag{24}$$

Where we let $\mathbf{g}_{(n,j),(n+1,j)} = \mathbf{g}_{(n+1,j),(n,j)} = \mathbf{g}_{(i,m+1),(i,m)} = \mathbf{g}_{(i,m),(i,m+1)} = 0$. To show feasibility, we write out fully the flow constraint of Equation 23:

$$\begin{aligned} &\mathbf{g}_{(i,j),(i+1,j)} - \mathbf{g}_{(i+1,j),(i,j)} + \\ &\mathbf{g}_{(i,j),(i-1,j)} - \mathbf{g}_{(i-1,j),(i,j)} + \\ &\mathbf{g}_{(i,j),(i,j+1)} - \mathbf{g}_{(i,j+1),(i,j)} + \\ &\mathbf{g}_{(i,j),(i,j-1)} - \mathbf{g}_{(i,j-1),(i,j)} = \mathbf{x}'_{i,j} - \mathbf{x}_{i,j} \end{aligned} \tag{25}$$

Substituting in Equation 24:

$$\boldsymbol{\delta}^{\text{vert.}}_{i,j} + -\boldsymbol{\delta}^{\text{vert.}}_{i-1,j} + \boldsymbol{\delta}^{\text{horiz.}}_{i,j} + -\boldsymbol{\delta}^{\text{horiz.}}_{i,j-1} = \mathbf{x}'_{i,j} - \mathbf{x}_{i,j} \tag{26}$$

But by Definition 3.1, this is exactly:

$$\Delta(\boldsymbol{x}, \boldsymbol{\delta})_{i,j} = \boldsymbol{x}'_{i,j} \tag{27}$$

Which is the sole constraint in Equation 22: then any solution which is feasible in Equation 23 is feasible in Equation 22. Also note that:

$$\begin{aligned} \|\boldsymbol{\delta}\|_1 &= \sum_{i,j} |\boldsymbol{\delta}^{\text{vert.}}_{i,j}| + |\boldsymbol{\delta}^{\text{horiz.}}_{i,j}| \\ &\leq \sum_{i,j} |\mathbf{g}_{(i,j),(i+1,j)}| + |\mathbf{g}_{(i+1,j),(i,j)}| \\ &\quad + |\mathbf{g}_{(i,j),(i,j+1)}| + |\mathbf{g}_{(i,j+1),(i,j)}| \\ &= \sum_{i,j} \mathbf{g}_{(i,j),(i+1,j)} + \mathbf{g}_{(i+1,j),(i,j)} \\ &\quad + \mathbf{g}_{(i,j),(i,j+1)} + \mathbf{g}_{(i,j+1),(i,j)} \\ &= \sum_{i,j} \mathbf{g}_{(i,j),(i+1,j)} + \mathbf{g}_{(i,j),(i,j+1)} \\ &\quad + \sum_{i,j} \mathbf{g}_{(i+1,j),(i,j)} + \mathbf{g}_{(i,j+1),(i,j)} \\ &= \sum_{i,j} \mathbf{g}_{(i,j),(i+1,j)} + \mathbf{g}_{(i,j),(i,j+1)} \\ &\quad + \sum_{i,j} \mathbf{g}_{(i,j),(i-1,j)} + \mathbf{g}_{(i,j),(i,j-1)} \\ &= \sum_{(i,j)} \sum_{(i',j') \in \mathcal{N}((i,j))} \mathbf{g}_{(i,j),(i',j')} \end{aligned} \tag{28}$$

Where the inequality follows from triangle inequality applied to Equation 24, and in the second sum in the fourth line, we exploit the fact that $\mathbf{g}_{(n,j),(n+1,j)} = \mathbf{g}_{(n+1,j),(n,j)} = \mathbf{g}_{(i,m+1),(i,m)} = \mathbf{g}_{(i,m),(i,m+1)} = 0$ to shift indices. This shows that the minimization objective in Equation 22 is less than or equal to the minimization objective in Equation 23.

Moving on to (2), we give the transformation as:

$$\begin{aligned} \mathbf{g}_{(i,j),(i+1,j)} &:= \max(\boldsymbol{\delta}^{\text{vert.}}_{i,j}, 0) \\ \mathbf{g}_{(i,j),(i-1,j)} &:= \max(-\boldsymbol{\delta}^{\text{vert.}}_{i-1,j}, 0) \\ \mathbf{g}_{(i,j),(i,j+1)} &:= \max(\boldsymbol{\delta}^{\text{horiz.}}_{i,j}, 0) \\ \mathbf{g}_{(i,j),(i,j-1)} &:= \max(-\boldsymbol{\delta}^{\text{horiz.}}_{i,j-1}, 0) \end{aligned} \tag{29}$$

Note that the non-negativity constraint of Equation 23 is automatically satisfied by the form of these definitions. Shifting indices, we also have:

$$\begin{aligned} \mathbf{g}_{(i-1,j),(i,j)} &= \max(\boldsymbol{\delta}^{\text{vert.}}_{i-1,j}, 0) \\ \mathbf{g}_{(i+1,j),(i,j)} &= \max(-\boldsymbol{\delta}^{\text{vert.}}_{i,j}, 0) \\ \mathbf{g}_{(i,j-1),(i,j)} &= \max(\boldsymbol{\delta}^{\text{horiz.}}_{i,j-1}, 0) \\ \mathbf{g}_{(i,j+1),(i,j)} &= \max(-\boldsymbol{\delta}^{\text{horiz.}}_{i,j}, 0) \end{aligned} \tag{30}$$

From the constraint on Equation 22, we have:

$$\begin{aligned}
\mathbf{x}'_{i,j} - \mathbf{x}_{i,j} =& \boldsymbol{\delta}^{\text{vert.}}_{i,j} + \\
& - \boldsymbol{\delta}^{\text{vert.}}_{i-1,j} + \\
& \boldsymbol{\delta}^{\text{horiz.}}_{i,j} + \\
& - \boldsymbol{\delta}^{\text{horiz.}}_{i,j-1} \\
=& \max(\boldsymbol{\delta}^{\text{vert.}}_{i,j}, 0) - \max(-\boldsymbol{\delta}^{\text{vert.}}_{i,j}, 0) + \\
& \max(-\boldsymbol{\delta}^{\text{vert.}}_{i-1,j} - \max(\boldsymbol{\delta}^{\text{vert.}}_{i-1,j}, 0), 0) + \\
& \max(\boldsymbol{\delta}^{\text{horiz.}}_{i,j}, 0) - \max(-\boldsymbol{\delta}^{\text{horiz.}}_{i,j}, 0) + \\
& \max(-\boldsymbol{\delta}^{\text{horiz.}}_{i,j-1}, 0) - \max(\boldsymbol{\delta}^{\text{horiz.}}_{i,j-1}, 0) \\
=& \mathbf{g}_{(i,j),(i+1,j)} - \mathbf{g}_{(i+1,j),(i,j)} + \\
& \mathbf{g}_{(i,j),(i-1,j)} - \mathbf{g}_{(i-1,j),(i,j)} + \\
& \mathbf{g}_{(i,j),(i,j+1)} - \mathbf{g}_{(i,j+1),(i,j)} + \\
& \mathbf{g}_{(i,j),(i,j-1)} - \mathbf{g}_{(i,j-1),(i,j)}
\end{aligned} \tag{31}$$

Which is exactly the second constraint of Equation 23: then any solution which is feasible in Equation 23 is feasible in Equation 22. Also note that:

$$\begin{aligned}
& \sum_{(i,j)} \sum_{(i',j') \in \mathcal{N}((i,j))} \mathbf{g}_{(i,j),(i',j')} \\
=& \sum_{i,j} \mathbf{g}_{(i,j),(i+1,j)} + \mathbf{g}_{(i,j),(i,j+1)} \\
& + \sum_{i,j} \mathbf{g}_{(i,j),(i-1,j)} + \mathbf{g}_{(i,j),(i,j-1)} \\
=& \sum_{i,j} \max(\boldsymbol{\delta}^{\text{vert.}}_{i,j}, 0) + \max(\boldsymbol{\delta}^{\text{horiz.}}_{i,j}, 0) \\
& + \sum_{i,j} \max(-\boldsymbol{\delta}^{\text{vert.}}_{i-1,j}, 0) + \max(-\boldsymbol{\delta}^{\text{horiz.}}_{i,j-1}, 0) \\
=& \sum_{i,j} \max(\boldsymbol{\delta}^{\text{vert.}}_{i,j}, 0) + \max(-\boldsymbol{\delta}^{\text{vert.}}_{i,j}, 0) \\
& + \sum_{i,j} \max(\boldsymbol{\delta}^{\text{horiz.}}_{i,j}, 0) + \max(-\boldsymbol{\delta}^{\text{horiz.}}_{i,j}, 0) \\
=& \sum_{i,j} |\boldsymbol{\delta}^{\text{vert.}}_{i,j}| + |\boldsymbol{\delta}^{\text{horiz.}}_{i,j}| \\
=& \|\boldsymbol{\delta}\|_1
\end{aligned} \tag{32}$$

Where we again exploit the fact that $\mathbf{g}_{(n,j),(n+1,j)} = \mathbf{g}_{(n+1,j),(n,j)} = \mathbf{g}_{(i,m+1),(i,m)} = \mathbf{g}_{(i,m),(i,m+1)} = 0$ to shift indices, in the fourth line. This shows that the minimization objective in Equation 23 is less than or equal to the minimization objective in Equation 22, completing (2).

Finally, now that we have shown that Equations 22 and 23 are in fact equivalent minimizations (i.e., we have proven Equation 22 correct), we would like to show that there is always a feasible solution to 22, as claimed. By the above transformations, it suffices to show that there is always a feasible solution to Equation 23. Ling and Okada (2007) show that any feasible solution the the general Wasserstein minimization LP (Definition 1) can be transformed into a solution to Equation 23, so it suffices to show that the LP in Definition 1 always has a feasible solution. This is trivially satisfied by taking $\Pi = \boldsymbol{x}(\boldsymbol{x}')^T$, where we note that $\boldsymbol{x}$, a probability distribution, is non-negative. □

Theorem 1. *Consider a normalized probability distribution $\boldsymbol{x} \in [0,1]^{n \times m}$, and a classification score function $\boldsymbol{f} : \mathbb{R}^{n \times m} \to [0,1]^k$. Let $\bar{\boldsymbol{f}}$ refer to the Wasserstein-smoothed classification function:*

$$\bar{\boldsymbol{f}}(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\delta} \sim \mathcal{L}(\sigma)} [\boldsymbol{f}(\Delta(\boldsymbol{x}, \boldsymbol{\delta}))] \tag{33}$$

Let i be the class assignment of $\boldsymbol{x}$ using the smoothed classifier $\bar{\boldsymbol{f}}$ (i.e. $i = \arg\max_{i'} \bar{\boldsymbol{f}}_{i'}(\boldsymbol{x})$). If

$$\bar{\boldsymbol{f}}_i(\boldsymbol{x}) \geq e^{2\sqrt{2}\rho/\sigma} \max_{i' \neq i} \bar{\boldsymbol{f}}_{i'}(\boldsymbol{x}) \tag{34}$$

Then for any perturbed probability distribution $\tilde{\boldsymbol{x}}$ such that $W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \leq \rho$:

$$\bar{\boldsymbol{f}}_i(\tilde{\boldsymbol{x}}) \geq \max_{i' \neq i} \bar{\boldsymbol{f}}_{i'}(\tilde{\boldsymbol{x}}) \tag{35}$$

Proof. Let $\boldsymbol{u}$ be the uniform probability vector. As a consequence of Lemma 1, for any distribution $\boldsymbol{x}$, there exists a nonempty set of local flow plans $S_{\boldsymbol{x}}$:

$$S_{\boldsymbol{x}} = \{\boldsymbol{\delta} | \boldsymbol{x} = \Delta(\boldsymbol{u}, \boldsymbol{\delta})\} \tag{36}$$

Also, we may define a version of the classifier $\boldsymbol{f}$ on the local flow plan domain:

$$\boldsymbol{f}^{\text{flow}}(\boldsymbol{\delta}) = \boldsymbol{f}(\Delta(\boldsymbol{u}, \boldsymbol{\delta})) \tag{37}$$

Let $\boldsymbol{\delta}_x$ be an arbitrary element in S_x, and consider any perturbed $\tilde{\boldsymbol{x}}$ such that $W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \leq \rho$. By Theorem 1:

$$\min_{\boldsymbol{\delta}:\, \tilde{\boldsymbol{x}} = \Delta(\boldsymbol{x}, \boldsymbol{\delta})} \|\boldsymbol{\delta}\|_1 = W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \tag{38}$$

Then, using Equation 6:

$$\min_{\boldsymbol{\delta}:\, \tilde{\boldsymbol{x}} = \Delta(\boldsymbol{u}, \boldsymbol{\delta}_{\boldsymbol{x}} + \boldsymbol{\delta})} \|\boldsymbol{\delta}\|_1 = W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \tag{39}$$

Let the minimum be achieved at $\boldsymbol{\delta}^*$. Making a change of variables ($\boldsymbol{\delta}_{\tilde{\boldsymbol{x}}} = \boldsymbol{\delta}^* + \boldsymbol{\delta}_x$), we have:

$$\|\boldsymbol{\delta}_{\tilde{\boldsymbol{x}}} - \boldsymbol{\delta}_{\boldsymbol{x}}\|_1 = W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \quad \text{where } \tilde{\boldsymbol{x}} = \Delta(\boldsymbol{u}, \boldsymbol{\delta}_{\tilde{\boldsymbol{x}}}) \tag{40}$$

Note that for any $\boldsymbol{x}'$ (for $\boldsymbol{\delta}' \sim \mathcal{L}(\sigma)$) :

$$\begin{aligned}
\bar{\boldsymbol{f}}(\boldsymbol{x}') =& \mathbb{E}\left[\boldsymbol{f}(\Delta(\boldsymbol{x}', \boldsymbol{\delta}')\right] \\
=& \mathbb{E}\left[\boldsymbol{f}(\Delta(\boldsymbol{u}, \boldsymbol{\delta}_{\boldsymbol{x}'} + \boldsymbol{\delta}'))\right] \\
=& \mathbb{E}\left[\boldsymbol{f}^{\text{flow}}(\boldsymbol{\delta}_{\boldsymbol{x}'} + \boldsymbol{\delta}'))\right]
\end{aligned} \tag{41}$$

We can now apply Proposition 1 from Lecuyer et al. (2019), restated here:

Proposition. *Consider a vector $\boldsymbol{v} \in \mathbb{R}^d$, and a classification score function $\boldsymbol{h} : \mathbb{R}^d \to [0,1]^k$. Let $\epsilon \sim \text{Laplace}(0,\sigma)^d$, and let i be the class assignment of $\boldsymbol{v}$ using a Laplace-smoothed version of the classifier $\boldsymbol{h}$:*

$$i = \arg\max_{i'} \mathop{\mathbb{E}}_{\epsilon}[\boldsymbol{h}_{i'}(\boldsymbol{v}+\epsilon)] \tag{42}$$

If:

$$\mathop{\mathbb{E}}_{\epsilon}[\boldsymbol{h}_i(\boldsymbol{v}+\epsilon)] \geq e^{2\sqrt{2}\rho/\sigma} \max_{i' \neq i} \mathop{\mathbb{E}}_{\epsilon}[\boldsymbol{h}_{i'}(\boldsymbol{v}+\epsilon)] \tag{43}$$

Then for any perturbed probability distribution $\tilde{\boldsymbol{v}}$ such that $\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|_1 \leq \rho$:

$$\mathop{\mathbb{E}}_{\epsilon}[\boldsymbol{h}_i(\tilde{\boldsymbol{v}}+\epsilon)] \geq \max_{i' \neq i} \mathop{\mathbb{E}}_{\epsilon}[\boldsymbol{h}_{i'}(\tilde{\boldsymbol{v}}+\epsilon)] \tag{44}$$

We apply this proposition to $\boldsymbol{f}^{\text{flow}}$, noting that $\|\boldsymbol{\delta}_{\tilde{\boldsymbol{x}}} - \boldsymbol{\delta}_{\boldsymbol{x}}\|_1 = W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \leq \rho$:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{\delta}'}\left[\boldsymbol{f}_i^{\text{flow}}(\boldsymbol{\delta}_{\boldsymbol{x}} + \boldsymbol{\delta}'))\right] \geq e^{2\sqrt{2}\rho/\sigma} \max_{i' \neq i} \mathop{\mathbb{E}}_{\boldsymbol{\delta}'}\left[\boldsymbol{f}_{i'}^{\text{flow}}(\boldsymbol{\delta}_{\boldsymbol{x}} + \boldsymbol{\delta}'))\right] \\ &\implies \mathop{\mathbb{E}}_{\boldsymbol{\delta}'}\left[\boldsymbol{f}_i^{\text{flow}}(\boldsymbol{\delta}_{\tilde{\boldsymbol{x}}} + \boldsymbol{\delta}'))\right] \geq \max_{i' \neq i} \mathop{\mathbb{E}}_{\boldsymbol{\delta}'}\left[\boldsymbol{f}_{i'}^{\text{flow}}(\boldsymbol{\delta}_{\tilde{\boldsymbol{x}}} + \boldsymbol{\delta}'))\right] \end{aligned} \tag{45}$$

Then, using Equation 41:

$$\begin{aligned} \bar{\boldsymbol{f}}_i(\boldsymbol{x}) \geq e^{2\sqrt{2}\rho/\sigma} \max_{i' \neq i} \bar{\boldsymbol{f}}_{i'}(\boldsymbol{x}) \implies \\ \bar{\boldsymbol{f}}_i(\tilde{\boldsymbol{x}}) \geq \max_{i' \neq i} \bar{\boldsymbol{f}}_{i'}(\tilde{\boldsymbol{x}}) \end{aligned} \tag{46}$$

Which was to be proven. □

Corollary 1. *For any normalized probability distributions $\boldsymbol{x}, \boldsymbol{x}' \in [0,1]^{n \times m}$, if $W_1(\boldsymbol{x}, \boldsymbol{x}') \leq \rho/2$, then $\|\boldsymbol{x} - \boldsymbol{x}'\|_1 \leq \rho$, where W_1 is the 1-Wasserstein metric using any L_p norm as the underlying distance metric. Furthermore, there exist distributions where these inequalities are tight.*

Proof. Let Π indicate the optimal transport plan between $\boldsymbol{x}$ and $\boldsymbol{x}'$. From Definition 1, we have $\Pi \mathbf{1} = \boldsymbol{x}$ and $\Pi^T \mathbf{1} = \boldsymbol{x}'$. Then:

$$(\Pi^T - \Pi)\mathbf{1} = \boldsymbol{x}' - \boldsymbol{x} \tag{47}$$

Let Π' represent a modified version of Π, with the diagonal elements set to zero. Note that $< \Pi', C >=< \Pi, C >$ and $\Pi^T - \Pi = (\Pi')^T - \Pi'$. Then, using triangle inequality:

$$\begin{aligned} &\|(\Pi')^T \mathbf{1}\|_1 + \|(\Pi')\mathbf{1}\|_1 \\ \geq &\|((\Pi')^T - \Pi')\mathbf{1}\|_1 \\ = &\|\boldsymbol{x}' - \boldsymbol{x}\|_1 \end{aligned} \tag{48}$$

Because the elements of Π' are non-negative, this is simply:

$$2\sum_{i,j} \Pi'_{i,j} \geq \|((\Pi')^T - \Pi')\mathbf{1}\|_1 = \|\boldsymbol{x}' - \boldsymbol{x}\|_1 \tag{49}$$

Then, because the (non-diagonal) elements of C are at least 1 for any L_p norm, we have,

$$2 < \Pi', C > \geq 2\sum_{i,j} \Pi'_{i,j} \geq \|\boldsymbol{x}' - \boldsymbol{x}\|_1 \tag{50}$$

Because $< \Pi', C >=< \Pi, C >= W_1(\boldsymbol{x}, \boldsymbol{x}')$, this means that $\|\boldsymbol{x}' - \boldsymbol{x}\|_1 \leq 2W_1(\boldsymbol{x}, \boldsymbol{x}') \leq \rho$, which was to be proven. Note that this inequality can be tight. For example, let $\boldsymbol{x}$ be the distribution where the entire probability mass is at position (i,j), and $\boldsymbol{x}'$ be the distribution where the probability mass is equally split between at positions (i,j) and $(i+1,j)$. (In other words, $\boldsymbol{x}_{(i,j)} = 1, \boldsymbol{x}'_{(i,j)} = .5, \boldsymbol{x}'_{(i+1,j)} = .5$). In this case, $\|\boldsymbol{x}' - \boldsymbol{x}\|_1 = 1$, $W_1(\boldsymbol{x}, \boldsymbol{x}') = .5$. □

Corollary 2. *Consider a color image with three channels, denoted $\mathbf{x} = [\mathbf{x}^R, \mathbf{x}^G, \mathbf{x}^B]$, normalized such that $\sum_{(i,j)} \mathbf{x}^R_{(i,j)} + \mathbf{x}^G_{(i,j)} + \mathbf{x}^B_{(i,j)} = 1$. Consider a perturbed image $\tilde{\boldsymbol{x}}$ such that $\forall K \in \{R,G,B\}$, $\sum_{(i,j)} \mathbf{x}^K_{(i,j)} = \sum_{(i,j)} \tilde{\boldsymbol{x}}^K_{(i,j)}$. Let $W_1(\mathbf{x}, \tilde{\boldsymbol{x}})$ denote the 1-Wasserstein distance (with L_1 distance metric) between $\mathbf{x}$ and $\tilde{\boldsymbol{x}}$, where, when determining the minimum transport plan, transport between channels is not permitted. Using this definition, let $W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \leq \rho$. Define:*

$$\begin{aligned} \boldsymbol{\delta} &= \{\boldsymbol{\delta}^R, \boldsymbol{\delta}^G, \boldsymbol{\delta}^B\} \\ \Delta(\boldsymbol{x}, \boldsymbol{\delta}) &= \{\Delta(\boldsymbol{x}^R, \boldsymbol{\delta}^{\boldsymbol{R}}), \Delta(\boldsymbol{x}^G, \boldsymbol{\delta}^{\boldsymbol{G}}), \Delta(\boldsymbol{x}^B, \boldsymbol{\delta}^{\boldsymbol{B}})\} \end{aligned} \tag{51}$$

and let $\mathcal{L}^{color}(\sigma)$ represent independent draws of Laplace noise each with standard deviation σ in the shape of $\boldsymbol{\delta}$. Then if

$$\bar{\boldsymbol{f}}_i(\boldsymbol{x}) \geq e^{2\sqrt{2}\rho/\sigma} \max_{i' \neq i} \bar{\boldsymbol{f}}_{i'}(\boldsymbol{x}) \tag{52}$$

then

$$\bar{\boldsymbol{f}}_i(\tilde{\boldsymbol{x}}) \geq \max_{i' \neq i} \bar{\boldsymbol{f}}_{i'}(\tilde{\boldsymbol{x}}). \tag{53}$$

Proof. Let the mass in each channel be denoted s_K:

$$s_K \coloneqq \sum_{(i,j)} \mathbf{x}^K_{(i,j)} = \sum_{(i,j)} \tilde{\boldsymbol{x}}^K_{(i,j)} \tag{54}$$

Consider the formulation of Wasserstein distance given in Definition 1. If we represent the elements of $\boldsymbol{x}$ as a vector by concatenating the elements of $\boldsymbol{\delta}^R, \boldsymbol{\delta}^G$, and $\boldsymbol{\delta}^B$, then the restriction that there is no flow between channels amounts to the requirement that Π is block-diagonal:

$$\Pi = \begin{bmatrix} \Pi^R & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Pi^G & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Pi^B \end{bmatrix} \tag{55}$$

Let $C^{1,1}$ represent the standard cost matrix for 1-Wasserstein transport (with L_1 distance metric). Because the cost of transport within each channel is the

same for standard 1-Wasserstein transport (with L_1 distance metric), we have:

$$C = \begin{bmatrix} C^{1,1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & C^{1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & C^{1,1} \end{bmatrix} \tag{56}$$

Then we have:

$$< \Pi, C >=< \Pi^R, C^{1,1} > + < \Pi^G, C^{1,1} > + < \Pi^B, C^{1,1} > \tag{57}$$

And by Equation 55, the constraints also factorize out:

$$\begin{aligned} \Pi^R \mathbf{1} = \boldsymbol{x}^R,\ (\Pi^R)^T \mathbf{1} = \tilde{\boldsymbol{x}}^R, \\ \Pi^B \mathbf{1} = \boldsymbol{x}^B,\ (\Pi^B)^T \mathbf{1} = \tilde{\boldsymbol{x}}^B, \\ \Pi^G \mathbf{1} = \boldsymbol{x}^G,\ (\Pi^G)^T \mathbf{1} = \tilde{\boldsymbol{x}}^G \end{aligned} \tag{58}$$

Then the variables of each Π^K are separable (in that they appear together in the objective only in the sum and share no constraints). We can then factorize the minimization:

$$W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_K \min_{\Pi^K \in \mathbb{R}_+^{(n\cdot m)\times(n\cdot m)}} < \Pi^K, C^{(1,1)} >, \tag{59}$$

$$\forall K,\ \Pi^K \mathbf{1} = \boldsymbol{x}^K,\ (\Pi^K)^T \mathbf{1} = \tilde{\boldsymbol{x}}^K \tag{60}$$

$$\tag{61}$$

We can transform each x^K into a normalized probability distribution by scaling it by a factor of $1/s_K$. We similarly scale each Π^K :

$$\boldsymbol{x}^K_{\text{sc.}} := \frac{\boldsymbol{x}^K}{s_K} \quad \Pi^K_{\text{sc.}} := \frac{\Pi^K}{s_K} \tag{62}$$

Then we have:

$$W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_K s_K \cdot \min_{\Pi^K_{\text{sc.}} \in \mathbb{R}_+^{(n\cdot m)\times(n\cdot m)}} < \Pi^K_{\text{sc.}}, C^{(1,1)} >, \tag{63}$$

$$\forall K,\ \Pi^K_{\text{sc.}} \mathbf{1} = \boldsymbol{x}^K_{\text{sc.}},\ (\Pi^K_{\text{sc.}})^T \mathbf{1} = \tilde{\boldsymbol{x}}^K_{\text{sc.}} \tag{64}$$

$$\tag{65}$$

But note that this is simply:

$$W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_K s_K \cdot W_1(\boldsymbol{x}^K_{\text{sc.}}, \tilde{\boldsymbol{x}}^K_{\text{sc.}}) \tag{66}$$

By Lemma 1, this is:

$$W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_K s_K \cdot \min_{\boldsymbol{\delta}^K_{\text{sc.}}:\ \tilde{\boldsymbol{x}}^K_{\text{sc.}} = \Delta(\boldsymbol{x}^K_{\text{sc.}}, \boldsymbol{\delta}^K_{\text{sc.}})} \|\boldsymbol{\delta}^K_{\text{sc.}}\|_1 \tag{67}$$

By the linearity to scaling of Δ and the L_1 norm, this is simply:

$$W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_K \min_{\boldsymbol{\delta}^K:\ \tilde{\boldsymbol{x}}^K = \Delta(\boldsymbol{x}^K, \boldsymbol{\delta}^K)} \|\boldsymbol{\delta}^K\|_1 \tag{68}$$

Which, by Equation 51, is simply,

$$W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \min_{\boldsymbol{\delta}:\ \tilde{\boldsymbol{x}} = \Delta(\boldsymbol{x}, \boldsymbol{\delta})} \|\boldsymbol{\delta}\|_1 \tag{69}$$

Then all of the mechanics of the proof of Theorem 1 apply, and (avoiding unnecessary repetition), we conclude the result. □

Corollary 3. *Let W^1 denote the L_1 1-Wasserstein distance, and W^2 denote the L_2 1-Wasserstein distance. For a radius ρ_2, define $\rho_1 := \sqrt{2}\rho_2$. Then, for any classifier f and input $\boldsymbol{x}$, if there does not exist any adversarial example $\tilde{\boldsymbol{x}}$ with $W_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \leq \rho_1$, then there are also no adversarial examples $\tilde{\boldsymbol{x}}'$ with $W_2(\boldsymbol{x}, \tilde{\boldsymbol{x}}') \leq \rho_2$.*

Proof. We show the contrapositive: If there is an adversarial example $\tilde{\boldsymbol{x}}'$ with $W^2(\boldsymbol{x}, \tilde{\boldsymbol{x}}') \leq \rho_2$, then there is an adversarial example $\tilde{\boldsymbol{x}}$ with $W^1(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \leq \rho_1$. It is sufficient to show that for any arbitrary $\boldsymbol{x}'$, if $W^2(\boldsymbol{x}, \boldsymbol{x}') \leq \rho_2$, then $W^1(\boldsymbol{x}, \boldsymbol{x}') \leq \rho_1$. (The predicate is then satisfied with $\tilde{\boldsymbol{x}}' = \tilde{\boldsymbol{x}} = \boldsymbol{x}'$). In other words, we need to show that

$$\sqrt{2} W^2(\boldsymbol{x}, \boldsymbol{x}') \geq W^1(\boldsymbol{x}, \boldsymbol{x}'), \quad \forall \boldsymbol{x}, \boldsymbol{x}'. \tag{70}$$

By the definition of 1-Wasserstein distance, can rewrite this goal as

$$\sqrt{2} \min_{\Pi} < \Pi, C^2 > \ \geq \ \min_{\Pi} < \Pi, C^1 > \tag{71}$$

where in both minimizations Π is non-negative and subject to $\Pi \mathbf{1} = \boldsymbol{x}$, $\Pi^T \mathbf{1} = \boldsymbol{x}'$. Here, C^2 and C^1 are the weight matrices for L_2 and L_1 Wasserstein distances. Note that Π is subject to the same constraints in both minimizations: therefore any Π that is feasible in one is feasible in the other. Let Π_2^* be the minimum of the first (L_2) minimization. Recall that

$$C^2_{(i,j),(i',j')} = \sqrt{(i-i')^2 + (j-j')^2} \tag{72}$$

while

$$C^1_{(i,j),(i',j')} = |i-i'| + |j-j'|. \tag{73}$$

By equivalence of norms, we have:

$$\sqrt{2} C^2_{(i,j),(i',j')} \geq C^1_{(i,j),(i',j')}. \tag{74}$$

Then by linearity (and using Π non-negative),

$$\sqrt{2} < \Pi, C^2 > \ \geq \ < \Pi, C^1 > \quad \forall \Pi \geq 0 \tag{75}$$

So

$$\begin{aligned} \sqrt{2} \min_{\Pi} < \Pi, C^2 > \ &= \ \sqrt{2} < \Pi_2^*, C^2 > \\ &\geq \ < \Pi_2^*, C^1 > \\ &\geq \ \min_{\Pi} < \Pi, C^1 >, \end{aligned} \tag{76}$$

as desired. □

B Training Parameters

In this paper, network architectures models used were identical to those used in Wong et al. (2019). Unless stated otherwise, all parameters of attacks are the same as used in that paper for each data set. For training smoothed models, we train the base classifier using standard cross-entropy loss on individual noised sample images, using the same noise distribution as used when performing smoothed classification. However, during training, rather than using the same image repeatedly while adding different noise (as at test time), we instead train with each image only once per epoch, with one noise draw. In fact, for computational efficiency and as suggested by Lecuyer et al. (2019), we re-use the same noise for each image in a batch. Training parameters are as follows (Tables 3, 4):

Table 3 Training Parameters for MNIST Experiments

Training Epochs	200
Batch Size	128
Optimizer	Stochastic Gradient Descent with Momentum
Learning Rate	.001
Momentum	0.9
L_2 Weight Penalty	0.0005

Table 4 Training Parameters for CIFAR-10 Experiments

Training Epochs	200
Batch Size	128
Training Set Preprocessing	Normalization, Random Cropping (Padding:4) and Random Horizontal Flip
Optimizer	Stochastic Gradient Descent with Momentum
Learning Rate	.01 (Epochs 1-200) .001 (Epochs 201-400)
Momentum	0.9
L_2 Weight Penalty	0.0005

C Comparison to other Defenses in Wong et al. (2019)

In addition to proposing adversarial training as a defense against Wasserstein Adversarial attacks, Wong et al. (2019) also tests other defenses. On MNIST, binarization of the input and using a provably L_∞-robust classifier were also tested as defenses: our randomized smoothing method is more effective than these methods at all attack magnitudes (see Figure 7). On CIFAR-10, Wong et al. (2019) only tested a provably

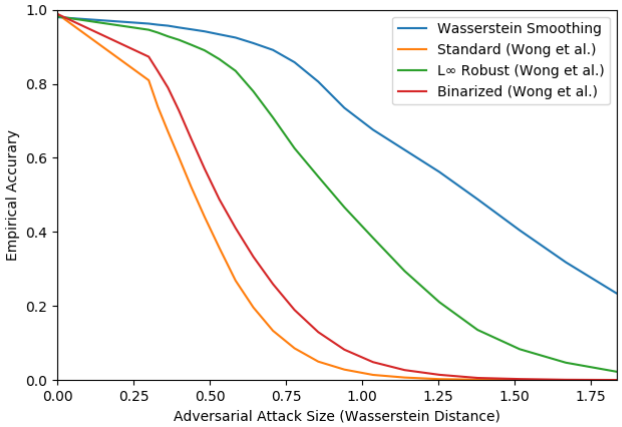

Figure 7: Comparison of empirical robustness on MNIST to additional defenses from (Wong et al., 2019), other than adversarial training. Randomized Smoothing shown here is Wasserstein smoothing with $\sigma = 0.01$. (This is the amount of noise which maximizes certified robustness, as seen in Table 1.)

L_∞-robust classifier as an additional defense: unfortunately, code was not provided for this model, so we did not attempt to replicate the results.