Project Notes

This project builds upon [CLT24] to extend Remark 3.1 and Section 4.3 to the analysis of deep neural networks (DNNs) and transformers.

1 Notations

$\mathcal{A}$	σ-algebra on $\mathcal{Z}$ such that $\{z\} \in \mathcal{A}$ for any $z \in \mathcal{Z}$
$(\mathcal{Z}, \mathcal{A})$	Measurable space
$\mathcal{Z} \times \mathcal{Z}$	Cartesian product measurable space with σ-algebra $\mathcal{A} \times \mathcal{A}$
$\mathbb{P} : \mathcal{A} \to [0, \infty]$	Probability function with countably additivity, $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\mathcal{Z}) = 1$
$\mathcal{P}(\mathcal{Z})$	Space of all probabilities on $\mathcal{Z}$
$\mathrm{E}_{\mathbb{P}}[f(Z)] = \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z)$	Expectation of a measurable function f of a real-valued random variable Z on $(\mathcal{Z}, \mathcal{A}, \mathbb{P})$
$\boldsymbol{\delta}_S : \mathcal{Z} \to \mathbb{R}$	Indicator function of a set $S \subset \mathcal{Z}$, $\boldsymbol{\delta}_S(z) = 0$ if $z \in S$, and ∞ otherwise
$\boldsymbol{\chi}_{\{\hat{z}\}} \in \mathcal{P}(\mathcal{Z})$	Point mass function (Dirac measure) at point $\hat{z} \in \mathcal{Z}$ as $\boldsymbol{\chi}_{\{\hat{z}\}}(A) = 1$ if $\hat{z} \in A$, and 0 otherwise, for any measurable set $A \subset \mathcal{Z}$
$f \otimes g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$	$(x, y) \to f(x) \cdot g(y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$
$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$	Inner product on $\mathbb{R}^n$ for any $x, y \in \mathbb{R}^n$
$\|\cdot\|_{\mathbb{R}^n}$	An arbitrary norm on $\mathbb{R}^n$
$\|\cdot\|_{\mathbb{R}^n,*}$	Dual norm defined as $\|x\|_{\mathbb{R}^n,*} := \max_{y \in \mathbb{R}^n} \{\langle x, y \rangle \mid \|y\|_{\mathbb{R}^n} = 1\}$
$[A, B] \in \mathbb{R}^{n_1 \times (n_2 + n_3)}$	Horizontal concatenation of $A \in \mathbb{R}^{n_1 \times n_2}$ and $B \in \mathbb{R}^{n_1 \times n_3}$
$[A; C] \in \mathbb{R}^{(n_1 + n_3) \times n_2}$	Vertical concatenation of $A \in \mathbb{R}^{n_1 \times n_2}$ and $C \in \mathbb{R}^{n_3 \times n_2}$
sgn	Sign function as $\mathrm{sgn}(t) = -1$ if $t < 0$, and $\mathrm{sgn}(t) = 1$ otherwise
β	Decision variable from decision space $\mathcal{B}$
Z	Random variable in a given space $\mathcal{Z}$, with probability distribution $\mathbb{P}_{\text{true}}$
$\ell : \mathcal{Z} \times \mathcal{B} \to \mathbb{R}$	Loss function
$\mathbb{P}_N := \sum_{i=1}^{N} \mu_i \boldsymbol{\chi}_{\{Z^{(i)}\}}$	Empirical distribution
$\mathcal{Z}_N := \{Z^{(1)}, ..., Z^{(N)}\} \subset \mathcal{Z}$	Training dataset
$\{\mu_i\}_{i=1}^{N}$	Nonnegative weights satisfying $\sum_{i=1}^{N} \mu_i = 1$
$\mathfrak{M} \subset \mathcal{P}(\mathcal{Z})$	Ambiguity set
$d : \mathcal{Z} \times \mathcal{Z} \to [0, \infty]$	Extended nonnegative-valued function
$r \in [1, \infty)$	Exponent in $\mathcal{W}$
$\sigma(\mathcal{Z})$	Set of all measurable sets in $\mathcal{Z}$
$\Pi(\mathbb{P}, \mathbb{Q})$	Set of all joint probability distributions between $\mathbb{P}$ and $\mathbb{Q}$

$$\Pi(\mathbb{P}, \mathbb{Q}) = \{\pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z}) \text{ such that } \forall A, B \in \sigma(\mathcal{Z}),\ \pi(A \times \mathcal{Z}) = \mathbb{P}(A),\ \pi(\mathcal{Z} \times B) = \mathbb{Q}(B)\}$$

$$\mathcal{W}_{d,r}(\mathbb{P},\mathbb{Q}) := \left(\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{Z}\times\mathcal{Z}} d^r(z',z)\,\mathrm{d}\pi(z',z) \right)^{\frac{1}{r}} \tag{2}$$

$$\mathcal{S} := \sup_{\mathbb{P}:\mathcal{W}_{d,r}(\mathbb{P},\mathbb{P}_N)\le\delta} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \tag{4}$$

2 Remark 3.1

In Theorem 3.2, it is required that the condition (A2) holds for any $i = 1,\ldots,N$, with respect to the same Lipschitz constant $L_\beta^{\mathcal{Z}^N}$. To relax this condition, one might assume that Assumptions (A1 & A2) hold at each $Z^{(i)}$ with a Lipschitz constant $L_\beta^{\{Z^{(i)}\}}$, for $i = 1,\ldots,N$. Even though it might not guarantee that the lower bound and upper bound for $\mathcal{S}$ coincide as in Theorem 3.2, we show in Appendix C that one still has closed forms for the lower and upper bounds given by

$$\widehat{\mathcal{L}} = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + \sum_{i=1}^N \mu_i L_\beta^{\{Z^{(i)}\}}\delta,$$

$$\widehat{\mathcal{U}} = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + \max_{i=1,\ldots,N} L_\beta^{\{Z^{(i)}\}}\delta.$$

2.1 Theorem 3.2

Let $\mathcal{Z}_N := \{Z^{(1)},\ldots,Z^{(N)}\} \subset \mathcal{Z}$ be a given dataset and $\mathbb{P}_N := \sum_{i=1}^N \mu_i \boldsymbol{\chi}_{\{Z^{(i)}\}} \in \mathcal{P}(\mathcal{Z})$ be the corresponding empirical distribution. In addition, let $d(\cdot,\cdot)$ be a cost function on $\mathcal{Z}\times\mathcal{Z}$ and $\delta \in (0,\infty)$ be a scalar. Suppose the loss function $\ell : \mathcal{Z}\times\mathcal{B} \to \mathbb{R}$ takes the form as

$$\ell : (z;\beta) \mapsto \psi_\beta(z),$$

where the function $\psi_\beta : \mathcal{Z} \to \mathbb{R}$ satisfies the following assumptions:

(A1) ψ_β is $(L_\beta^{\mathcal{Z}_N}, d)$-Lipschitz at $\mathcal{Z}_N$ with $L_\beta^{\mathcal{Z}_N} \in (0,\infty)$;

(A2) For any $\epsilon \in (0, L_\beta^{\mathcal{Z}_N})$ and each $Z^{(i)} \in \mathcal{Z}_N$, there exists $\tilde{Z}_\epsilon^{(i)} \in \mathcal{Z}$ such that $\delta \le d(\tilde{Z}_\epsilon^{(i)}, Z^{(i)}) < \infty$ and

$$\psi_\beta(\tilde{Z}_\epsilon^{(i)}) - \psi_\beta(Z^{(i)}) \ge (L_\beta^{\mathcal{Z}_N} - \epsilon)d(\tilde{Z}_\epsilon^{(i)}, Z^{(i)}).$$

Then we have that $\mathcal{L} = \mathcal{S} = \mathcal{U}$ in Theorem 3.1, that is,

$$\sup_{\mathbb{P}:\mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_N)\le\delta} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + L_\beta^{\mathcal{Z}_N}\delta. \tag{7}$$

Proof. Since ψ_β is $(L_\beta^{\mathcal{Z}_N}, d)$-Lipschitz at $\mathcal{Z}_N$, by Theorem 3.1, we have that

$$\mathcal{L} \le \mathcal{S} = \sup_{\mathbb{P}:\mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_N)\le\delta} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \le \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + L_\beta^{\mathcal{Z}_N}\delta =: \mathcal{U}.$$

Hence, in order to prove (7), it suffices to show that $\mathcal{L} \ge \mathcal{U}$.

Let $\epsilon \in \left(0, \min\{L_\beta^{\mathcal{Z}_N}, \delta L_\beta^{\mathcal{Z}_N}\}\right)$ be an arbitrary scalar. By Assumption (A2), for any $Z^{(i)} \in \mathcal{Z}_N$, there exists $\tilde{Z}^{(i)} \in \mathcal{Z}$ such that $\delta \le d(\tilde{Z}^{(i)}, Z^{(i)}) < \infty$ and

$$\psi_\beta(\tilde{Z}^{(i)}) - \psi_\beta(Z^{(i)}) \ge \left(L_\beta^{\mathcal{Z}_N} - \frac{\epsilon}{\delta}\right) d(\tilde{Z}^{(i)}, Z^{(i)}).$$

Let $\eta^{(i)} := \delta / d(\tilde{Z}^{(i)}, Z^{(i)}) \in (0,1]$ and define

$$\tilde{\mathbb{P}}^{(i)} := \eta^{(i)}\boldsymbol{\chi}_{\{\tilde{Z}^{(i)}\}} + (1-\eta^{(i)})\boldsymbol{\chi}_{\{Z^{(i)}\}} \in \mathcal{P}(\mathcal{Z}).$$

Then we have

$$\mathcal{W}_{d,1}\left(\tilde{\mathbb{P}}^{(i)}, \boldsymbol{\chi}_{\{Z^{(i)}\}}\right) = \eta^{(i)} d(\tilde{Z}^{(i)}, Z^{(i)}) + (1-\eta^{(i)}) d(Z^{(i)}, Z^{(i)}) = \eta^{(i)} d(\tilde{Z}^{(i)}, Z^{(i)}) = \delta,$$

and

$$\begin{aligned}
\mathrm{E}_{\tilde{\mathbb{P}}^{(i)}}[\ell(Z;\beta)] &= \eta^{(i)}\psi_\beta(\tilde{Z}^{(i)}) + (1-\eta^{(i)})\psi_\beta(Z^{(i)}) \\
&= \psi_\beta(Z^{(i)}) + \eta^{(i)}\left[\psi_\beta(\tilde{Z}^{(i)}) - \psi_\beta(Z^{(i)})\right] \\
&\geq \psi_\beta(Z^{(i)}) + \eta^{(i)}\left(L_\beta^{\mathcal{Z}_N} - \frac{\epsilon}{\delta}\right) d(\tilde{Z}^{(i)}, Z^{(i)}) \\
&= \ell(Z^{(i)};\beta) + L_\beta^{\mathcal{Z}_N}\delta - \epsilon.
\end{aligned}$$

Letting $\epsilon \to 0$, we get for all $i = 1, \ldots, N$,

$$\mathcal{L}_i = \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \,\middle|\, \mathcal{W}_{d,1}(\mathbb{P}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \leq \delta \right\} \geq \ell(Z^{(i)};\beta) + L_\beta^{\mathcal{Z}_N}\delta.$$

Therefore, it holds that

$$\mathcal{L} = \sum_{i=1}^N \mu_i \mathcal{L}_i \geq \sum_{i=1}^N \mu_i \left(\ell(Z^{(i)};\beta) + L_\beta^{\mathcal{Z}_N}\delta\right) = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + L_\beta^{\mathcal{Z}_N}\delta = \mathcal{U}.$$

□

2.2 Appendix C

Let $\mathcal{Z}_N := \{Z^{(1)}, \ldots, Z^{(N)}\} \subset \mathcal{Z}$ be a given dataset and $\mathbb{P}_N := \sum_{i=1}^N \mu_i \boldsymbol{\chi}_{\{Z^{(i)}\}} \in \mathcal{P}(\mathcal{Z})$ be the corresponding empirical distribution. In addition, let $d(\cdot,\cdot)$ be a cost function on $\mathcal{Z}\times\mathcal{Z}$ and $\delta \in (0,\infty)$ be a scalar. Suppose the loss function $\ell : \mathcal{Z}\times\mathcal{B} \to \mathbb{R}$ takes the form

$$\ell : (z;\beta) \mapsto \psi_\beta(z),$$

where the function $\psi_\beta : \mathcal{Z} \to \mathbb{R}$ satisfies the following assumptions:

(C1) ψ_β is $(L_\beta^{\{Z^{(i)}\}}, d)$-Lipschitz at $\{Z^{(i)}\}$ with $L_\beta^{\{Z^{(i)}\}} \in (0,\infty)$ for each $1 \leq i \leq N$;

(C2) For any $\epsilon \in (0, \min_i L_\beta^{\{Z^{(i)}\}})$ and each $Z^{(i)} \in \mathcal{Z}_N$, there exists $\tilde{Z}_\epsilon^{(i)} \in \mathcal{Z}$ such that $\delta \leq d(\tilde{Z}_\epsilon^{(i)}, Z^{(i)}) < \infty$ and

$$\psi_\beta(\tilde{Z}_\epsilon^{(i)}) - \psi_\beta(Z^{(i)}) \geq (L_\beta^{\{Z^{(i)}\}} - \epsilon) d(\tilde{Z}_\epsilon^{(i)}, Z^{(i)}).$$

Then we have that $\widehat{\mathcal{L}} \leq \mathcal{S} \leq \widehat{\mathcal{U}}$, where

$$\widehat{\mathcal{L}} = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + \sum_{i=1}^N \mu_i L_\beta^{\{Z^{(i)}\}}\delta,$$

$$\widehat{\mathcal{U}} = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + \max_{i=1,\ldots,N} L_\beta^{\{Z^{(i)}\}}\delta,$$

which means that

$$\widehat{\mathcal{L}} = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + \sum_{i=1}^N \mu_i L_\beta^{\{Z^{(i)}\}}\delta \leq \sup_{\mathbb{P}:\mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_N)\leq\delta} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \leq \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + \max_{i=1,\ldots,N} L_\beta^{\{Z^{(i)}\}}\delta = \widehat{\mathcal{U}}.$$

Proof. Since ψ_β is $(L_\beta^{\{Z^{(i)}\}}, d)$-Lipschitz at $\{Z^{(i)}\}$ for each $1 \leq i \leq N$, it implies that ψ_β is $(L_\beta^{\mathcal{Z}_N}, d)$-Lipschitz at $\mathcal{Z}_N$, with

$$L_\beta^{\mathcal{Z}_N} = \max_{i=1,\dots,N} L_\beta^{\{Z^{(i)}\}}.$$

By Theorem 3.1, letting

$$\mathcal{L}_i := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \,\middle|\, \mathcal{W}_{d,1}(\mathbb{P}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \leq \delta \right\}, \quad i = 1, \dots, N,$$

we have

$$\sum_{i=1}^{N} \mu_i \mathcal{L}_i \leq \mathcal{S} = \sup_{\mathbb{P}:\mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_N)\leq\delta} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \leq \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + L_\beta^{\mathcal{Z}_N}\delta = \widehat{\mathcal{U}}.$$

Then by applying Theorem 3.2 to each $\mathcal{L}_i$, we obtain:

$$\mathcal{L}_i = \ell(Z^{(i)};\beta) + L_\beta^{\{Z^{(i)}\}}\delta,$$

which means that

$$\sum_{i=1}^{N} \mu_i \mathcal{L}_i = \sum_{i=1}^{N} \mu_i \ell(Z^{(i)};\beta) + \sum_{i=1}^{N} \mu_i L_\beta^{\{Z^{(i)}\}}\delta = \widehat{\mathcal{L}}.$$

□

2.3 Theorem 3.1

Let $\mathcal{Z}_N := \{Z^{(1)}, \dots, Z^{(N)}\} \subset \mathcal{Z}$ be a given dataset and $\mathbb{P}_N := \sum_{i=1}^N \mu_i \boldsymbol{\chi}_{\{Z^{(i)}\}} \in \mathcal{P}(\mathcal{Z})$ be the corresponding empirical distribution. In addition, let $r \in [1, \infty)$ be a scalar and $d(\cdot,\cdot)$ be a cost function on $\mathcal{Z} \times \mathcal{Z}$. Suppose the loss function $\ell : \mathcal{Z} \times \mathcal{B} \to \mathbb{R}$ takes the form as

$$\ell : (z;\beta) \mapsto \psi_\beta^r(z), \quad \text{with} \quad \begin{cases} \psi_\beta : \mathcal{Z} \to \mathbb{R} & \text{if } r = 1, \\ \psi_\beta : \mathcal{Z} \to \mathbb{R}_+ & \text{if } r > 1. \end{cases}$$

Let $\mathcal{S}$ be defined as in (4). Then the following statements hold for any $\delta \geq 0$.

(a) Let

$$\mathcal{L}_i := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \,\middle|\, \mathcal{W}_{d,r}(\mathbb{P}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \leq \delta \right\}, \quad \text{for } i = 1, \dots, N.$$

Then

$$\mathcal{S} \geq \mathcal{L} := \sum_{i=1}^{N} \mu_i \mathcal{L}_i \geq \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)].$$

(b) Suppose ψ_β is $(L_\beta^{\mathcal{Z}_N}, d)$-Lipschitz at $\mathcal{Z}_N$ with $L_\beta^{\mathcal{Z}_N} \in (0, \infty)$, then

$$\mathcal{S} \leq \mathcal{U} := \left((\mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)])^{1/r} + L_\beta^{\mathcal{Z}_N}\delta \right)^r.$$

(c) Suppose ψ_β is $(0, d)$-Lipschitz at $\mathcal{Z}_N$, then

$$\mathcal{S} = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)].$$

Proof. **(a)** For any collection $\left\{\tilde{\mathbb{P}}^{(i)}\right\}_{i=1}^{N} \subseteq \mathcal{P}(\mathcal{Z})$ such that $\mathcal{W}_{d,r}(\tilde{\mathbb{P}}^{(i)}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \leq \delta$ for all $i = 1, ..., N$. It follows from Lemma 3.1 that for each $i = 1, \dots, N$,

$$\mathcal{W}_{d,r}(\tilde{\mathbb{P}}^{(i)}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) = \left(\int_{\mathcal{Z}} d^r(z, Z^{(i)}) \, \mathrm{d}\tilde{\mathbb{P}}^{(i)}(z) \right)^{1/r} \leq \delta.$$

Define

$$\tilde{\mathbb{P}} := \sum_{i=1}^{N} \mu_i \tilde{\mathbb{P}}^{(i)}, \qquad \tilde{\pi} := \sum_{i=1}^{N} \left(\mu_i \tilde{\mathbb{P}}^{(i)} \otimes \boldsymbol{\chi}_{\{Z^{(i)}\}} \right). \tag{6}$$

Then $\tilde{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$, $\tilde{\pi} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$, and $\tilde{\pi} \in \Pi(\tilde{\mathbb{P}}, \mathbb{P}_N)$, since for any measurable sets $A, B \subset \mathcal{Z}$,

$$\tilde{\pi}(\mathcal{Z} \times B) = \sum_{i=1}^{N} \mu_i \tilde{\mathbb{P}}^{(i)}(\mathcal{Z}) \boldsymbol{\chi}_{\{Z^{(i)}\}}(B) = \sum_{i=1}^{N} \mu_i \boldsymbol{\chi}_{\{Z^{(i)}\}}(B) = \mathbb{P}_N(B),$$
$$\tilde{\pi}(A \times \mathcal{Z}) = \sum_{i=1}^{N} \mu_i \tilde{\mathbb{P}}^{(i)}(A) \boldsymbol{\chi}_{\{Z^{(i)}\}}(\mathcal{Z}) = \sum_{i=1}^{N} \mu_i \tilde{\mathbb{P}}^{(i)}(A) = \tilde{\mathbb{P}}(A).$$

Hence,

$$\begin{aligned}\mathcal{W}_{d,r}(\tilde{\mathbb{P}}, \mathbb{P}_N) &\le \left(\int_{\mathcal{Z} \times \mathcal{Z}} d^r(\tilde{z}, z) \, \mathrm{d}\tilde{\pi}(\tilde{z}, z) \right)^{1/r} = \left(\int_{\mathcal{Z} \times \mathcal{Z}} d^r(\tilde{z}, z) \, \mathrm{d} \sum_{i=1}^{N} \left(\mu_i \tilde{\mathbb{P}}^{(i)}(\tilde{z}) \boldsymbol{\chi}_{\{Z^{(i)}\}}(z) \right) \right)^{1/r} \\ &= \left(\sum_{i=1}^{N} \mu_i \int_{\mathcal{Z}} d^r(\tilde{z}, Z^{(i)}) \, \mathrm{d}\tilde{\mathbb{P}}^{(i)}(\tilde{z}) \right)^{1/r} = \left(\sum_{i=1}^{N} \mu_i \left(\mathcal{W}_{d,r}(\tilde{\mathbb{P}}^{(i)}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \right)^r \right)^{1/r} \le \delta.\end{aligned}$$

Moreover, from (6) we have:

$$\mathrm{E}_{\tilde{\mathbb{P}}}[\ell(Z; \beta)] = \sum_{i=1}^{N} \mu_i \mathrm{E}_{\tilde{\mathbb{P}}^{(i)}}[\ell(Z; \beta)].$$

By taking the supremum over all possible $\left\{ \widetilde{\mathbb{P}}^{(i)} \right\}_{i=1}^{N}$ such that $\mathcal{W}_{d,r}(\widetilde{\mathbb{P}}^{(i)}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \le \delta$ for all $i = 1, ..., N$, we have

$$\begin{aligned}\mathcal{S} &= \sup_{\mathbb{P}: \mathcal{W}_{d,r}(\mathbb{P}, \mathbb{P}_N) \le \delta} \mathrm{E}_{\mathbb{P}}[\ell(Z; \beta)] \\ &\ge \sum_{i=1}^{N} \mu_i \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathrm{E}_{\mathbb{P}}[\ell(Z; \beta)] \mid \mathcal{W}_{d,r}(\mathbb{P}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \le \delta \right\} = \sum_{i=1}^{N} \mu_i \mathcal{L}_i = \mathcal{L}.\end{aligned}$$

Besides, since $\mathcal{W}_{d,r}(\boldsymbol{\chi}_{\{Z^{(i)}\}}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) = 0 \le \delta$ by Lemma 3.1, we have that

$$\mathcal{L}_i = \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathrm{E}_{\mathbb{P}}[\ell(Z; \beta)] \mid \mathcal{W}_{d,r}(\mathbb{P}, \boldsymbol{\chi}_{\{Z^{(i)}\}}) \le \delta \right\} \ge \mathrm{E}_{\boldsymbol{\chi}_{\{Z^{(i)}\}}}[\ell(Z; \beta)] = \ell(Z^{(i)}; \beta),$$

and hence

$$\mathcal{L} = \sum_{i=1}^{N} \mu_i \mathcal{L}_i \ge \sum_{i=1}^{N} \mu_i \ell(Z^{(i)}; \beta) = \mathrm{E}_{\mathbb{P}_N}[\ell(Z; \beta)].$$

(b) Let $\epsilon > 0$ be an arbitrary scalar. Fix any $\widetilde{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ such that $\mathcal{W}_{d,r}(\widetilde{\mathbb{P}}, \mathbb{P}_N) \le \delta$. Then by the definition of $\mathcal{W}_{d,r}(\cdot, \cdot)$, there exists $\widetilde{\pi} \in \Pi(\widetilde{\mathbb{P}}, \mathbb{P}_N)$ such that

$$\left(\int_{\mathcal{Z} \times \mathcal{Z}} d^r(\tilde{z}, z) \, \mathrm{d}\widetilde{\pi}(\tilde{z}, z) \right)^{1/r} \le \delta + \frac{\epsilon}{L_{\beta}^{\mathcal{Z}_N}}.$$

Besides, by the definition of the loss function $\ell(\cdot,\cdot)$, we have

$$\begin{aligned}\left(\mathrm{E}_{\widetilde{\mathbb{P}}}[\ell(Z;\beta)]\right)^{\frac{1}{r}} &= \left(\int_{\mathcal{Z}} \psi_\beta^r(\tilde{z})\,\mathrm{d}\widetilde{\mathbb{P}}(\tilde{z})\right)^{\frac{1}{r}} = \left(\int_{\mathcal{Z}\times\mathcal{Z}} \psi_\beta^r(\tilde{z})\,\mathrm{d}\widetilde{\pi}(\tilde{z},z)\right)^{\frac{1}{r}} \\ &= \left(\int_{\mathcal{Z}\times\mathcal{Z}} \left(\psi_\beta(z)+\psi_\beta(\tilde{z})-\psi_\beta(z)\right)^r\,\mathrm{d}\widetilde{\pi}(\tilde{z},z)\right)^{\frac{1}{r}} \\ &\leq^{(*)} \left(\int_{\mathcal{Z}\times\mathcal{Z}} \psi_\beta^r(z)\,\mathrm{d}\widetilde{\pi}(\tilde{z},z)\right)^{\frac{1}{r}} + \left(\int_{\mathcal{Z}\times\mathcal{Z}} |\psi_\beta(\tilde{z})-\psi_\beta(z)|^r\,\mathrm{d}\widetilde{\pi}(\tilde{z},z)\right)^{\frac{1}{r}} \\ &= \left(\int_{\mathcal{Z}} \psi_\beta^r(z)\,\mathrm{d}\mathbb{P}_N(z)\right)^{\frac{1}{r}} + \left(\int_{\mathcal{Z}\times\mathcal{Z}} |\psi_\beta(\tilde{z})-\psi_\beta(z)|^r\,\mathrm{d}\widetilde{\pi}(\tilde{z},z)\right)^{\frac{1}{r}} \\ &= \left(\mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)]\right)^{\frac{1}{r}} + \left(\int_{\mathcal{Z}\times\mathcal{Z}} |\psi_\beta(\tilde{z})-\psi_\beta(z)|^r\,\mathrm{d}\widetilde{\pi}(\tilde{z},z)\right)^{\frac{1}{r}},\end{aligned}$$

where the inequality $^{(*)}$ holds naturally if $r=1$, and follows from the Minkowski inequality if $r>1$. Since ψ_β is $(L_\beta^{\mathcal{Z}_N}, d)$-Lipschitz at $\mathcal{Z}_N$, it holds that

$$\left(\mathrm{E}_{\widetilde{\mathbb{P}}}[\ell(Z;\beta)]\right)^{\frac{1}{r}} \leq \left(\mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)]\right)^{\frac{1}{r}} + L_\beta^{\mathcal{Z}_N}\left(\int_{\mathcal{Z}\times\mathcal{Z}} d^r(\tilde{z},z)\,\mathrm{d}\widetilde{\pi}(\tilde{z},z)\right)^{\frac{1}{r}} \leq \left(\mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)]\right)^{\frac{1}{r}} + L_\beta^{\mathcal{Z}_N}\delta+\epsilon.$$

This means that for any $\epsilon>0$, we have

$$\mathcal{S}^{\frac{1}{r}} = \sup_{\mathbb{P}:\mathcal{W}_{d,r}(\mathbb{P},\mathbb{P}_N)\leq\delta} \left(\mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)]\right)^{\frac{1}{r}} \leq \left(\mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)]\right)^{\frac{1}{r}} + L_\beta^{\mathcal{Z}_N}\delta+\epsilon.$$

By letting $\epsilon\to 0$, we get the desired inequality.

(c) Since ψ_β is $(0,d)$-Lipschitz at $\mathcal{Z}_N$, by the convention that $0\cdot\infty=0$, one has $\psi_\beta(z')=\psi_\beta(z)$ for any $z'\in\mathcal{Z}, z\in\mathcal{Z}_N$. In particular, $\psi_\beta(\bar{z})=\psi_\beta(z)$ for any $\bar{z},z\in\mathcal{Z}_N$. Therefore, $\psi_\beta(\cdot)$ is a constant function on $\mathcal{Z}$, and so is $\ell(\cdot;\beta)$. Thus, we have

$$\mathcal{S} = \sup_{\mathbb{P}:\mathcal{W}_{d,r}(\mathbb{P},\mathbb{P}_N)\leq\delta} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)].$$

□

2.4 Lemma 3.1

Given any distribution $\mathbb{P}\in\mathcal{P}(\mathcal{Z})$ and any point $\hat{z}\in\mathcal{Z}$, for any scalar $r\geq 1$ and any extended nonnegative-valued measurable function $d:\mathcal{Z}\times\mathcal{Z}\to[0,\infty]$, we have

$$\mathcal{W}_{d,r}(\mathbb{P},\boldsymbol{\chi}_{\{\hat{z}\}}) = \left(\int_{\mathcal{Z}} d^r(z,\hat{z})\,\mathrm{d}\mathbb{P}(z)\right)^{1/r}.$$

Proof. For any $\pi\in\Pi(\mathbb{P},\boldsymbol{\chi}_{\{\hat{z}\}})$, we have $\pi(A\times\mathcal{Z})=\mathbb{P}(A), \pi(\mathcal{Z}\times B)=\boldsymbol{\chi}_{\{\hat{z}\}}(B)$ for any measurable sets $A,B\subset\mathcal{Z}$. In particular, it holds that

$$\pi(\mathcal{Z}\times(\mathcal{Z}\setminus\{\hat{z}\})) = \boldsymbol{\chi}_{\{\hat{z}\}}(\mathcal{Z}\setminus\{\hat{z}\}) = 0. \tag{11}$$

This implies that for any measurable set $A\subset\mathcal{Z}$, $\pi(A\times(\mathcal{Z}\setminus\{\hat{z}\}))=0$ and hence

$$\begin{aligned}\pi(A\times\{\hat{z}\}) &= \pi(A\times\mathcal{Z}) - \pi(A\times(\mathcal{Z}\setminus\{\hat{z}\})) \\ &= \pi(A\times\mathcal{Z}) = \mathbb{P}(A).\end{aligned}$$

Moreover, (11) also implies that

$$\int_{\mathcal{Z}\times(\mathcal{Z}\setminus\{\hat{z}\})} d^r(z',z)\,\mathrm{d}\pi(z',z) = 0.$$

Therefore, one has that

$$\int_{\mathcal{Z}\times\mathcal{Z}} d^r(z',z)\,\mathrm{d}\pi(z',z) = \int_{\mathcal{Z}\times\{\hat{z}\}} d^r(z',z)\,\mathrm{d}\pi(z',z)$$
$$= \int_{\mathcal{Z}} d^r(z',\hat{z})\,\mathrm{d}\mathbb{P}(z').$$

□

2.5 Definition 3.1

The function $d(\cdot,\cdot)$ defined on $\mathcal{Z}\times\mathcal{Z}$ is called a cost function if it is extended nonnegative-valued, measurable, and vanishes whenever two arguments are the same, that is, for any $z',z\in\mathcal{Z}$, $d(z',z)\in[0,\infty]$ and $d(z,z)=0$.

2.6 Definition 3.2

(Weak Lipschitz property). Given a function $f:\mathcal{Z}\to\mathbb{R}$, a cost function $d(\cdot,\cdot)$ on $\mathcal{Z}\times\mathcal{Z}$ and a subset $\mathcal{S}\subset\mathcal{Z}$, f is called $(L_f^{\mathcal{S}},d)$-Lipschitz at $\mathcal{S}$ if for any $z\in\mathcal{S}, z'\in\mathcal{Z}$, one has

$$|f(z')-f(z)|\le L_f^{\mathcal{S}} d(z',z),$$

where $L_f^{\mathcal{S}}\in[0,\infty)$ is a constant depending on f and $\mathcal{S}$.

3 Section 4.3

Apply our results to the cases where the cost function $d(\cdot,\cdot)$ is nonconvex, not positive definite, and the weak Lipschitz constant is not in the popular form of the norm of the regression vector β.

3.1 Example 4.4

(Ridge linear ordinary regression). For any $z'=(x',y'), z=(x,y)\in\mathbb{R}^n\times\mathbb{R}$, define $d(z',z)=\|z'-z\|_2\|z'+z\|_2$. Given any $\delta>0$, $\beta\in\mathbb{R}^n$ and any empirical distribution $\mathbb{P}_N$ on $\mathbb{R}^n\times\mathbb{R}$, for $\mathfrak{M}_1:=\{\mathbb{P}\in\mathcal{P}(\mathcal{Z})\mid \mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_N)\le\delta\}$, we have

$$\sup_{\mathbb{P}\in\mathfrak{M}_1} \mathrm{E}_{\mathbb{P}}[(Y+\langle\beta,X\rangle)^2] = \mathrm{E}_{\mathbb{P}_N}[(Y+\langle\beta,X\rangle)^2] + \|\beta\|_2^2\delta+\delta.$$

Proof. We first show that $\psi_\beta(z):=\langle[\beta;1],z\rangle^2$ for any $z\in\mathbb{R}^{n+1}$ satisfies Assumption (A1) with $L_\beta:=\|\beta\|_2^2+1$. For any $z',z\in\mathbb{R}^{n+1}$, we have

$$\begin{aligned}|\psi_\beta(z')-\psi_\beta(z)| &= |\langle[\beta;1],z'\rangle^2-\langle[\beta;1],z\rangle^2|\\ &= |\langle[\beta;1],z'-z\rangle||\langle[\beta;1],z'+z\rangle|\\ &\le \|[\beta;1]\|_2\|z'-z\|_2\cdot\|[\beta;1]\|_2\|z'+z\|_2\\ &= (\|\beta\|_2^2+1)\,d(z',z).\end{aligned}$$

Hence, ψ_β is (L_β,d)-Lipschitz on $\mathbb{R}^{n+1}$.

Next, we show that ψ_β satisfies Assumption (A2). For any $z\in\mathbb{R}^{n+1}$ and $k>0$, let $\tilde{z}:=z+k\Delta$ with

$$\Delta:=\frac{[\beta;1]}{\|[\beta;1]\|_2},$$

then $\|\tilde{z}-z\|_2=\|k\Delta\|_2=k$, $\|\tilde{z}+z\|_2=\|2z+k\Delta\|_2$ and

$$d(\tilde{z},z)=\|\tilde{z}-z\|_2\|\tilde{z}+z\|_2=k\|2z+k\Delta\|_2\ge k|k-2\|z\|_2|\to\infty \text{ as } k\to\infty.$$

On the other hand, we have

$$\frac{|\langle \Delta, \tilde{z}+z\rangle|}{\|\tilde{z}+z\|_2} \leq \|\Delta\|_2 = 1,$$

and

$$\frac{\langle \Delta, \tilde{z}+z\rangle}{\|\tilde{z}+z\|_2} = \frac{\langle \Delta, 2z+k\Delta\rangle}{\|2z+k\Delta\|_2} = \frac{\sum_{i=1}^{n+1}\Delta_i(2z_i/k+\Delta_i)}{\sqrt{\sum_{i=1}^{n+1}(2z_i/k+\Delta_i)^2}} \to 1, \text{ as } k\to\infty.$$

Thus, for the given δ and any $0<\epsilon<L_\beta$, there exists a positive integer k_ϵ such that for

$$\tilde{z}_\epsilon = z + k_\epsilon \Delta,$$

one has

$$d(\tilde{z}_\epsilon, z) = k_\epsilon\|\tilde{z}_\epsilon + z\|_2 \geq \delta,$$
$$\frac{\langle \Delta, \tilde{z}_\epsilon + z\rangle}{\|\tilde{z}_\epsilon+z\|_2} \geq 1 - \frac{\epsilon}{\|[\beta;1]\|_2^2}.$$

This implies that

$$\begin{aligned}
\psi_\beta(\tilde{z}_\epsilon) - \psi_\beta(z) &= \langle [\beta;1], \tilde{z}_\epsilon - z\rangle \cdot \langle [\beta;1], \tilde{z}_\epsilon + z\rangle \\
&= \|[\beta;1]\|_2^2 \langle \Delta, k_\epsilon\Delta\rangle \langle \Delta, \tilde{z}_\epsilon + z\rangle \\
&\geq \|[\beta;1]\|_2^2 k_\epsilon \left(1 - \frac{\epsilon}{\|[\beta;1]\|_2^2}\right)\|\tilde{z}_\epsilon + z\|_2 \\
&= \left(\|[\beta;1]\|_2^2 - \epsilon\right) k_\epsilon \|\tilde{z}_\epsilon + z\|_2 = (L_\beta - \epsilon)\, d(\tilde{z}_\epsilon, z).
\end{aligned}$$

Therefore, it satisfies Assumption (A2). By Theorem 3.2. □

4 Extension to DNN/Transformer

The main idea is that if a DNN or a Transformer's loss function is Lipschitz continuous (even just locally near the training data), WDRO theory provides upper/lower bound formulas for the worst-case loss, which is crucial for extending robust guarantees to high-dimensional models.

4.1 Literature Review

The notations used in the reviewed literature have been adapted to align with the notations in [CLT24].

4.1.1 [OSHL19]

Problem Statement Learn a language model $\mathbb{P}_\beta$ based on sentences sampled from the training distribution $Z \sim \mathbb{P}_N$, such that $\mathbb{P}_\beta$ performs well on unknown test distributions $\mathbb{P}_{\text{test}}$.

Language models $\mathbb{P}_\beta$ are generally trained to approximate $\mathbb{P}_N$ by minimizing the KL divergence $\mathrm{KL}(\mathbb{P}_N\|\mathbb{P}_\beta)$ via maximum likelihood estimation (MLE),

$$\inf_\beta \mathrm{E}[-\log \mathbb{P}_\beta(Z)]$$

Applying DRO, we optimize a model for loss ℓ and an ambiguity set of potential test distributions $\mathfrak{M}$ by minimizing the risk under the worst-case distribution in $\mathfrak{M}$,

$$\sup_{\mathbb{P}\in\mathfrak{M}} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)]$$

The above worst-case objective does not depend on the unknown quantity $\mathbb{P}_{\text{test}}$. The objective also upper bounds the test risk for all $\mathbb{P}_{\text{test}} \in \mathfrak{M}$ as

$$\mathrm{E}_{\mathbb{P}_{\text{test}}}[\ell(Z;\beta)] \leq \sup_{\mathbb{P}\in\mathfrak{M}} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)],$$

so optimizing the above objective gives guarantees on test performance whenever $\mathbb{P}_{\text{test}} \in \mathfrak{M}$.

4.1.2 [SV19]

Background and Notations A function f is said to be Lipschitz continuous if there exists a constant $L > 0$ such that

$$\|f(x) - f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.$$

The smallest such L is called the Lipschitz constant of f, denoted $L(f)$. When f is differentiable, we have $L(f) = \sup_x \|D_x f\|_2$.

$f \circ g$	Composition of functions $f : \mathbb{R}^k \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^k$
$D_x f \in \mathbb{R}^{m \times n}$	The Jacobian matrix at $x \in \mathbb{R}^n$, for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla f(x) = (D_x f)^\top$	The gradient in the scalar case ($m = 1$)
$\mathrm{diag}_{n,m}(x) \in \mathbb{R}^{n \times m}$	The rectangular diagonal matrix with entries of $x \in \mathbb{R}^{\min(n,m)}$ on the diagonal and zeros elsewhere (write $\mathrm{diag}(x)$ when unambiguous)

Multi-Layer Perceptron (MLP) A K-layer MLP $f_{\mathrm{MLP}} : \mathbb{R}^n \to \mathbb{R}^m$ is a function of the form:

$$f_{\mathrm{MLP}}(x) = T_K \circ \rho_{K-1} \circ \cdots \circ \rho_1 \circ T_1(x),$$

where $T_k(x) = M_k x + b_k$ is an affine function and $\rho_k : x \mapsto (g_k(x_i))_{i \in [\![1, n_k]\!]}$ is a non-linear activation function.

Computing the exact Lipschitz constant of a 2-layer MLP with ReLU activations is NP-hard.

Sequential Neural Networks and Lipschitz Bounds For an MLP, the Lipschitz constant has the following form:

$$L(f_{\mathrm{MLP}}) = \sup_{x \in \mathbb{R}^n} \left\| M_K \, \mathrm{diag}(g'_{K-1}(\theta_{K-1})) \cdots \mathrm{diag}(g'_1(\theta_1)) M_1 \right\|_2,$$

where $\theta_k = T_k \circ \rho_{k-1} \circ \cdots \circ \rho_1 \circ T_1(x)$ is the output after k layers.

The SeqLip algorithm approximates this by maximizing over all possible diagonal activations $\sigma_i \in [0, 1]^{n_i}$ and decomposing the expression via SVD. The resulting upper bound is:

$$\widehat{L}_{\mathrm{SL}} = \prod_{i=1}^{K-1} \max_{\sigma_i \in [0,1]^{n_i}} \left\| \widetilde{\Sigma}_{i+1} V_{i+1}^\top \mathrm{diag}(\sigma_i) U_i \widetilde{\Sigma}_i \right\|_2,$$

where $\widetilde{\Sigma}_i$ adjusts for boundary layers and U_i, V_i are singular vector matrices from M_i.

4.1.3 [BKM19]

The Robust Wasserstein Profile Function Given i.i.d. samples $\{\mathcal{W}1, \ldots, \mathcal{W}n\}$ and an estimating equation:

$$\mathrm{E}[h(W, \theta_*)] = 0,$$

the RWP function is defined as:

$$R_n(\theta) := \inf \left\{ D_c(\mathbb{P}, \mathbb{P}_n) : \mathrm{E}_{\mathbb{P}}[h(W, \theta)] = 0 \right\},$$

where D_c is a transport cost (typically Wasserstein), and $c(u, w) = \|w - u\|_q^\rho$.

Asymptotic Behavior Under regularity assumptions (A1-A4), the scaled RWP function converges in distribution:

$$n^{\rho/2} R_n(\theta_*) \xrightarrow{D} \bar{R}(\rho),$$

where, for $\rho > 1$,

$$\bar{R}(\rho) := \max_{\zeta \in \mathbb{R}^r} \left\{ \rho \zeta^\top H - (\rho - 1) \mathrm{E} \left\| \zeta^\top D_w h(W, \theta_*) \right\|_p^{\rho/(\rho-1)} \right\},$$

and $H \sim \mathcal{N}(0, \mathrm{Cov}[h(W, \theta_*)])$.

4.1.4 [SNVD20]

Introduction The key objective is to minimize the worst-case expected loss over a set of distributions $\mathfrak{M}$ around the empirical distribution $\mathbb{P}_N$:

$$\min_{\beta} \sup_{\mathbb{P}\in\mathfrak{M}} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)]$$

To ensure tractability, consider a Lagrangian relaxation with Wasserstein cost $\mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_N) \leq \delta$ and introduce a penalty parameter $\gamma \geq 0$:

$$\begin{aligned} \min_{\beta} F(\beta) &= \min_{\beta} \sup_{\mathbb{P}} \{\mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] - \gamma \mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_N)\} \\ &= \min_{\beta} \mathrm{E}_{\mathbb{P}_N}[\phi_\gamma(Z;\beta)] \end{aligned}$$

where $\phi_\gamma(Z;\beta)$ is the robust surrogate loss defined as:

$$\phi_\gamma(z_0;\beta) := \sup_{z\in\mathcal{Z}} \{\ell(z;\beta) - \gamma d(z_0,z)\}$$

Certified Robust Test Loss Bound Let $\mathbb{P}_{\text{test}}$ denote the empirical distribution on a test set $\{Z^i_{\text{test}}\}_{i=1}^{N_{\text{test}}}$. The robust test-time loss under worst-case perturbations within Wasserstein radius δ satisfies:

$$\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \sup_{z: d(z, Z^i_{\text{test}}) \leq \delta} \{\ell(z;\beta)\} \leq \sup_{\mathbb{P}: \mathcal{W}_{d,1}(\mathbb{P},\mathbb{P}_{\text{test}}) \leq \delta} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \leq \gamma\delta + \mathrm{E}_{\mathbb{P}_{\text{test}}}[\phi_\gamma(Z;\beta)]$$

This bound holds for all $\gamma, \delta \geq 0$. When γ is sufficiently large (so that the dual formulation is tight for small δ), this provides an efficient and certifiable upper bound on the worst-case test loss under Wasserstein perturbations.

4.1.5 [KKWP20]

Gradient-Based Approximation of DRO Objective Approximate the Wasserstein DRO worst-case risk by penalizing the expected norm of the gradient:

$$\sup_{\mathbb{P}: \mathcal{W}_{d,r}(\mathbb{P},\mathbb{P}_N)} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)] \approx \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + \alpha_n \|\nabla_z \ell(Z;\beta)\|_{\mathbb{P}_N, r^*},$$

where the norm $\|\cdot\|_{\mathbb{P}_N,r^*}$ is defined as:

$$\|g\|_{\mathbb{P}_N,r^*} := \left(\int_{\mathcal{Z}} \|g(z)\|_*^{r^*} \, \mathrm{d}\mathbb{P}_N(z) \right)^{1/r^*},$$

with $r^* := \frac{r}{r-1}$ denoting the Hölder conjugate of r. This surrogate loss avoids computing Lipschitz constants explicitly.

Theoretical Risk Consistency Assume ψ_β is differentiable and $\nabla_z \psi_\beta : \mathcal{Z} \to \mathbb{R}^d$ is (C_H, k)-Hölder continuous, i.e.,

$$\|\nabla_z \psi_\beta(z) - \nabla_z \psi_\beta(\tilde{z})\|_* \leq C_H \|z - \tilde{z}\|^k, \quad \forall z, \tilde{z} \in \mathcal{Z}.$$

Under smoothness assumptions (i.e., differentiability and Hölder continuity of gradients), the gradient-norm surrogate objective yields consistent minimizers, and the resulting estimator converges to the minimizer of the true worst-case risk.

4.1.6 [GFPC20]

Layerwise Composition of Lipschitz Constants A function $f : X \to Y$ is k-Lipschitz with respect to metrics D_X and D_Y if:

$$D_Y(f(\vec{x}_1), f(\vec{x}_2)) \leq k D_X(\vec{x}_1, \vec{x}_2), \quad \forall \vec{x}_1, \vec{x}_2 \in X.$$

For feedforward neural networks expressed as compositions of layers $f(\vec{x}) = (\phi_l \circ \cdots \circ \phi_1)(\vec{x})$, the Lipschitz constant satisfies:

$$L(f) \leq \prod_{i=1}^{l} L(\phi_i).$$

4.1.7 [GCK20]

Problem Formulation Given a family $\mathcal{F}$ of loss functions, a nominal distribution $\mathbb{Q} \in \mathcal{P}_r(\mathcal{Z})$, and a radius $\delta \geq 0$, the corresponding Wasserstein DRO problem is:

$$\min_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathrm{E}_{Z \sim \mathbb{P}}[f(Z)] : \mathcal{W}_{d,r}(\mathbb{P}, \mathbb{Q}) \leq \delta \right\}.$$

When $\mathbb{Q} = \mathbb{P}_N$ is the empirical distribution, the dual formulation of the inner supremum is:

$$\begin{cases} \min_{\lambda \geq 0} \left\{ \lambda \delta^r + \mathrm{E}_{Z \sim \mathbb{Q}} \left[\sup_{\tilde{z} \in \mathcal{Z}} \left\{ f(\tilde{z}) - \lambda d(\tilde{z}, Z)^r \right\} \right] \right\}, & r \in (1, \infty), \\ \mathrm{E}_{Z \sim \mathbb{Q}} \left[\sup_{\tilde{z} \in \mathcal{Z}} \left\{ f(\tilde{z}) : d(\tilde{z}, Z) \leq \delta \right\} \right], & r = \infty. \end{cases}$$

The Wasserstein regularizer is defined as:

$$\mathcal{R}_{\mathbb{Q},r}(\delta; f) := \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathrm{E}_{\mathbb{P}}[f(Z)] : \mathcal{W}_{d,r}(\mathbb{P}, \mathbb{Q}) \leq \delta \right\} - \mathrm{E}_{\mathbb{Q}}[f(Z)].$$

Theorem 1 (r-Wasserstein DRO)

(I) Let $r = \infty$. Assume Assumptions 1 and 2 hold. Then there exists $\bar{\delta} > 0$ such that for all $\delta < \bar{\delta}$ and $f \in \mathcal{F}$,

$$|\mathcal{R}_{\mathbb{P}_N,\infty}(\delta; f) - \delta \| \|\nabla f\|_* \|_{\mathbb{P}_N,1}| \leq \delta^2 \|H\|_{\mathbb{P}_N,1} + M \mathrm{E}_{\mathbb{P}_N}[(\delta - d(z, \mathcal{D}_f))_+].$$

(II) Let $r \in (1, \infty)$, $\delta_N = \delta_0/\sqrt{N}$, and assume Assumptions 1, 2, and 4 hold. Then there exist $\bar{\delta}, C_1, C_2 > 0$ such that for all $\delta_0 < \bar{\delta}$ and $f \in \mathcal{F}$,

$$|\mathcal{R}_{\mathbb{P}_N,r}(\delta_N; f) - \delta_N \| \|\nabla f\|_* \|_{\mathbb{P}_N,r^*}| \leq \delta_N^{2 \wedge r} \left(\|H\|_{\mathbb{P}_N, \frac{r}{r-2}} \mathbf{1}_{\{r>2\}} + C_1 \right) + M \mathrm{E}_{\mathbb{P}_N}[(C_2 \delta_N - d(z, \mathcal{D}_f))_+].$$

(III) Assume Assumption 3 holds. Let $t > 0$. Then there exists $\bar{\delta}, C > 0$ such that for all $\delta < \bar{\delta}$, with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$,

$$\mathrm{E}_{\mathbb{P}_N}[(\delta - d(z, \mathcal{D}_f))_+] \leq C\delta^2 + 2\delta \mathrm{E}_{\otimes}[\mathfrak{R}_N(\mathcal{I}_\delta)] + \delta \sqrt{\frac{t}{2N}},$$

and

$$\mathrm{E}_{\mathbb{P}_N}[(\delta - d(z, \mathcal{D}_f))_+] \leq 2C\delta^2 + \frac{48\delta}{\sqrt{N}} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon\delta; \mathcal{E}, \| \cdot \|_{\mathbb{P}_N,2})}\, d\epsilon + \delta \sqrt{\frac{t}{2N}}.$$

Under the condition $\delta_N = O(1/\sqrt{N})$ and $\mathrm{E}_{\mathbb{P}_N}[(\delta_N - d(z, \mathcal{D}_f))_+] = O(1/N)$, it follows that:

$$\min_{f \in \mathcal{F}} \sup_{\mathbb{P}: \mathcal{W}_{d,r}(\mathbb{P}, \mathbb{P}_N) \leq \delta_N} \mathrm{E}_{\mathbb{P}}[f(Z)] = \min_{f \in \mathcal{F}} \left\{ \mathrm{E}_{\mathbb{P}_N}[f(Z)] + \delta_N \mathcal{V}_{\mathbb{P}_N,r^*}(f) \right\} + O_p(1/N).$$

Neural Network Example Let $\theta = (W_1, W_2)$, and define a two-layer neural network with leaky ReLU activation and softmax cross-entropy loss:

$$f_\theta(z) := \ell(W_2\sigma(W_1x), y), \quad z = (x, y) \in \mathcal{Z},$$

where ℓ is the cross-entropy loss and $\sigma(z) = z \cdot \mathbf{1}\{z \geq 0\} + az \cdot \mathbf{1}\{z < 0\}$. Assume $\|W_2\|_{\text{op}} \cdot \|W_1\|_{\text{op}} \leq 1$, $\mathcal{X} \subset \mathbb{R}^d$ compact, and that $\mathbb{P}^x_{\text{true}}$ has continuous density. Then for differentiable points:

$$\|\nabla f_\theta(z)\|_2 = \|\nabla \ell(W_2\sigma(W_1x), y) \cdot W_2 \cdot \sigma'(W_1x) \cdot W_1\|_2 .$$

Using Theorem 1 and Lemma EC.21, there exists $\bar{\delta}, C > 0$ such that for all $\delta < \bar{\delta}$, with probability at least $1 - e^{-t}$, for all $\theta \in \Theta$,

$$\left|\mathcal{R}_{\mathbb{P}_N,2}(\delta_N; f_\theta) - \delta_N \, \|\|\nabla f_\theta\|_*\|_{\mathbb{P}_N,2}\right| \leq C_1\delta_N^2 + C_2 d_1\sqrt{\frac{d}{N}} + \delta_N\sqrt{\frac{t}{2N}}.$$

4.1.8 [KPM21]

Theorem 3.1 The dot-product multi-head attention (DP-MHA) mapping is not Lipschitz with respect to any p-norm $\|\cdot\|_p$ for $p \in [1, \infty]$.

Bounding the Lipschitz Constant of L2 Attention To address this limitation, analyze an alternative attention mechanism based on L2 similarity. For a Transformer block with:

- N: input sequence length,
- D: embedding dimension,
- H: number of attention heads,
- $W^{Q,h}, W^{K,h}, W^{V,h}$: query, key, value projection matrices for head h,
- W^O: output projection matrix,

Derive the following upper bound for the Lipschitz constant of the overall attention mapping F under the ℓ_2-norm:

$$\text{Lip}_2(F) \leq \frac{\sqrt{N}}{\sqrt{D/H}}\left(4\phi^{-1}(N-1)+1\right)\left(\sqrt{\sum_{h=1}^{H}\|W^{Q,h}\|_2^2\|W^{V,h}\|_2^2}\right)\|W^O\|_2$$

They also provide a bound under the ℓ_∞-norm:

$$\text{Lip}_\infty(F) \leq \left(4\phi^{-1}(N-1)+1\right) \cdot \max_h\left(\|W^{Q,h}\|_\infty\|W^{V,h}\|_\infty\right) \cdot \|W^O\|_\infty$$

where $\phi(x) = x\exp(x+1)$ and ϕ^{-1} is its inverse (which grows sub-logarithmically as $\mathcal{O}(\log N - \log\log N)$).

Specifically, for large N, these bounds simplify to:

$$\text{Lip}_\infty(F) = \mathcal{O}(\log N), \qquad \text{Lip}_2(F) = \mathcal{O}(\sqrt{N}\log N).$$

4.1.9 [ZYK+23]

Wasserstein DRO for CNN Classification. Propose a DRO framework for improving the generalization of CNNs on classification tasks under distribution shift. The goal is to minimize the worst-case expected loss over a Wasserstein ambiguity set $\mathfrak{M} := \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \mid \mathcal{W}_{d,1}(\mathbb{P}, \mathbb{P}_N) \leq \delta\}$:

$$\sup_{\mathbb{P}\in\mathfrak{M}} \mathrm{E}_{\mathbb{P}}[\ell(Z;\beta)],$$

where $\ell : \mathcal{Z} \times \mathcal{B} \to \mathbb{R}$ is the loss function and $\mathbb{P}_N$ is the empirical distribution.

Since directly solving the min–max problem is intractable, reformulate the objective into a regularized empirical risk minimization problem. Specifically, when the constraint

$$\left(\sum_{m=1}^{M} \frac{\|W_m\|}{M}\right)^M \leq \left(\frac{\theta}{M}\right)^M, \quad \text{for } \theta = \sum_{m=1}^{M} \|W_m^*\|$$

is satisfied by an optimal hypothesis $h^\star$, there exists a Lagrange multiplier $\tilde{\lambda} > 0$ such that $h^\star$ also solves the penalized problem:

$$\inf_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell(h(x^i), y^i) + \tilde{\lambda} \sum_{m=1}^{M} \|W_m\|.$$

In the experimental setup, the CNN feature extractor is pretrained and fixed, and DRO is applied only to the final classification layer (i.e., $M = 1$), making $\mathcal{H}$ a space of linear classifiers over frozen features. The regularization parameter $\tilde{\lambda}$ is treated as a hyperparameter and selected empirically via validation.

4.1.10 [LGL+23]

Wasserstein Penalty for Fair Representation Learning. Propose a fairness-aware classification method that promotes statistical independence between learned feature representations and sensitive attributes using a Wasserstein penalty. Rather than formulating distributional robustness via ambiguity sets, incorporate the 1-Wasserstein distance as a regularization term into the objective:

$$\min_{\theta} \mathcal{L}_{\text{cls}}(f_\theta(x), y) + \lambda \cdot \mathcal{W}(P_{f(x)}, P_{f(x)|A}),$$

where f_θ is the feature extractor, A denotes the sensitive attribute, and $\mathcal{W}$ is estimated using the Kantorovich–Rubinstein dual formulation.

4.1.11 [BHJO25]

First-Order W-DRO Fine-Tuning for Deep Neural Networks. Extend adversarial training to distributional threat models using Wasserstein DRO (W-DRO). Building on the sensitivity analysis of the inner adversary, derive a first-order approximation of the W-DRO objective and propose a two-step training method that can be applied to pretrained networks. The formulation,

$$\inf_{\theta} \left\{ \mathrm{E}_P[\mathcal{L}(f_\theta(x), y)] + \beta \sup_{\pi \in \Pi_2(P,\delta)} \mathrm{E}_\pi[\widetilde{\mathcal{L}}(f_\theta(x), f_\theta(x'))] \right\},$$

captures both clean accuracy and distributional robustness via a Wasserstein ball.

4.2 Analysis

This section extends [CLT24]'s analysis to deep neural networks (DNNs), residual networks (ResNets), and transformer architectures.

Given any $\delta > 0$ and any empirical distribution $\mathbb{P}_N$ on $\mathcal{Z}$, for $\mathfrak{M}_1 := \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \mid \mathcal{W}_{d,1}(\mathbb{P}, \mathbb{P}_N) \leq \delta\}$, our target is as follows:

$$\inf_{\beta \in \mathcal{B}} \left\{ \sup_{\mathbb{P} \in \mathfrak{M}_1} \mathrm{E}_{\mathbb{P}}[\psi_\beta(Z)] \right\}.$$

Following [ZYK+23], the sub-problem in the inner part of the above target can be rewritten in the following form:

$$\begin{aligned} \sup_{\pi \in \Pi(\mathbb{P}, \mathbb{P}_N)} \quad & \int_{\mathcal{Z} \times \mathcal{Z}} \psi_\beta(z') \mathrm{d}\pi(z', z) \\ \text{s.t.} \quad & \int_{\mathcal{Z} \times \mathcal{Z}} d(z', z) \mathrm{d}\pi(z', z) \leq \delta. \end{aligned}$$

Suppose $\mathcal{Z}_N := \{Z^{(1)}, ..., Z^{(N)}\} \subset \mathcal{Z}$ is a given dataset and we define $\mathbb{P}_N := \sum_{i=1}^{N} \mu_i \boldsymbol{\chi}_{\{Z^{(i)}\}} \in \mathcal{P}(\mathcal{Z})$ as the corresponding empirical distribution then we have:

$$\mathbb{P}(z') = \pi(z', \mathcal{Z}_N) = \sum_{i=1}^{N} \pi(z', z = Z^{(i)}) = \sum_{i=1}^{N} \pi(z'|z = Z^{(i)})\mathbb{P}_N(Z^{(i)}) = \sum_{i=1}^{N} \mathbb{P}^i(z')\mu_i$$

$$\pi(z', z) = \pi(z', z = Z^{(i)}) = \pi(z'|z = Z^{(i)})\mathbb{P}_N(Z^{(i)}) = \mathbb{P}^i(z')\mu_i \ ,$$

where $\mathbb{P}^i(z') = \pi(z'|z = Z^{(i)})$. Using these two equations, we can rewrite the sub-problem as:

$$\begin{aligned} \sup_{\{\mathbb{P}^i\}_{i=1}^N} \quad & \sum_{i=1}^{N} \mu_i \int_{\mathcal{Z}} \psi_\beta(z') \mathrm{d}\mathbb{P}^i(z') \\ \text{s.t.} \quad & \sum_{i=1}^{N} \mu_i \int_{\mathcal{Z}} d(z', Z^{(i)}) \mathrm{d}\mathbb{P}^i(z') \le \delta \\ & \int_{\mathcal{Z}} \mathrm{d}\mathbb{P}^i(z') = 1, \forall i \in [N] \ . \end{aligned}$$

Considering the Lagrangian of the above, we have:

$$\begin{aligned} \mathcal{L}(\{P^i\}, \lambda, \{\eta_i\}) &= \sum_{i=1}^{N} \mu_i \int_{\mathcal{Z}} \psi_\beta(z')\, \mathrm{d}\mathbb{P}^i(z') + \lambda \left[\delta - \sum_{i=1}^{N} \mu_i \int_{\mathcal{Z}} d(z', Z^{(i)}) \mathrm{d}\mathbb{P}^i(z')\right] \\ &\qquad + \sum_{i=1}^{N} \eta_i \left[1 - \int_{\mathcal{Z}} \mathrm{d}\mathbb{P}^i(z')\right] \\ &= \lambda\delta + \sum_{i=1}^{N} \left\{\eta_i + \int \left[\mu_i \psi_\beta(z') - \lambda \mu_i d(z', Z^{(i)}) - \eta_i\right] \mathrm{d}\mathbb{P}^i(z')\right\}, \end{aligned}$$

where $\lambda \ge 0$ and η_i are dual variables of the constraints. For a fixed i define:

$$f_i^{\lambda,\eta_i}(z') := \mu_i \psi_\beta(z') - \lambda \mu_i d(z', Z^{(i)}) - \eta_i.$$

Because $\mathbb{P}^i$ can put all its mass wherever f_i^{λ,η_i} is largest,

$$\sup_{\{\mathbb{P}^i\}_{i=1}^N} \int f_i^{\lambda,\eta_i}(z')\, \mathrm{d}\mathbb{P}^i = \sup_{z' \in \mathcal{Z}} f_i^{\lambda,\eta_i}(z').$$

Hence

$$\begin{aligned} \sup_{\{\mathbb{P}^i\}_{i=1}^N} \mathcal{L}(\{P^i\}, \lambda, \{\eta_i\}) &= \lambda\delta + \sum_{i=1}^{N} \left[\eta_i + \sup_{z' \in \mathcal{Z}} \left(\mu_i \psi_\beta(z') - \lambda \mu_i d(z', Z^{(i)}) - \eta_i\right)\right] \\ &= \lambda\delta + \sum_{i=1}^{N} \mu_i \sup_{z' \in \mathcal{Z}} \left[\psi_\beta(z') - \lambda d(z', Z^{(i)})\right]. \end{aligned}$$

The dual problem is

$$\inf_{\lambda \ge 0} \left\{\lambda\delta + \sum_{i=1}^{N} \mu_i \sup_{z' \in \mathcal{Z}} \left[\psi_\beta(z') - \lambda d(z', Z^{(i)})\right]\right\}. \tag{D}$$

Suppose ψ_β is $(L_\beta^{\mathcal{Z}_N}, d)$-Lipschitz at $\mathcal{Z}_N$ with $L_\beta^{\mathcal{Z}_N} \in (0, \infty)$ we have:

$$\psi_\beta(z') - \psi_\beta(z) \le |\psi_\beta(z') - \psi_\beta(z)| \le L_\beta^{\mathcal{Z}_N} d(z', z) \ \forall z \in \mathcal{Z}_N, z' \in \mathcal{Z} \ .$$

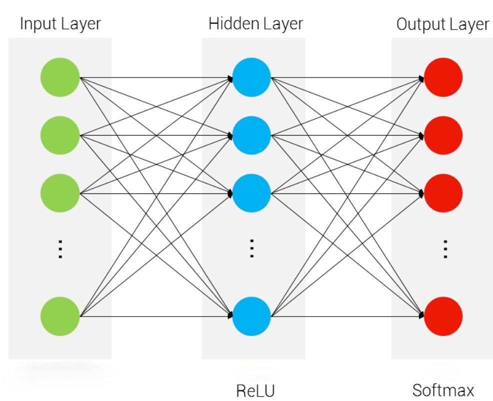
Using this assumption, we can approximate an upper bound for (D) as follows:

$$
\begin{aligned}
(\mathrm{D}) &\leq \inf_{\lambda \geq 0} \left\{ \lambda\delta + \sum_{i=1}^{N} \mu_i \sup_{z' \in \mathcal{Z}} \left[\psi_\beta(Z^{(i)}) + L_\beta^{\mathcal{Z}_N} d(z', Z^{(i)}) - \lambda d(z', Z^{(i)}) \right] \right\} \\
&= \inf_{\lambda \geq 0} \left\{ \lambda\delta + \sum_{i=1}^{N} \mu_i \sup_{z' \in \mathcal{Z}} \left[\psi_\beta(Z^{(i)}) + (L_\beta^{\mathcal{Z}_N} - \lambda) d(z', Z^{(i)}) \right] \right\} \\
&= L_\beta^{\mathcal{Z}_N} \delta + \sum_{i=1}^{N} \mu_i \psi_\beta(Z^{(i)}) = \mathrm{E}_{\mathbb{P}_N}[\ell(Z;\beta)] + L_\beta^{\mathcal{Z}_N} \delta \ .
\end{aligned}
$$

4.2.1 DNN

Toy Case Following the example given in [GCK20], we consider a two-layer network in the context of a K-class classification problem with ReLU activation function σ:

$$
\sigma(z) = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}
$$



Let $z' = (x', y')$, $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y}$ is a subset of the probability simplex in $\mathbb{R}^K$. Let $\beta = (W_1, W_2)$ where $W_1 \in \mathbb{R}^{n_1 \times n}$ and $W_2 \in \mathbb{R}^{K \times n_1}$ are weight matrices. Define the cost function d as:

$$
d(z', z) = \|x' - x\|_2 + \kappa \mathbf{1}_{\{y' \neq y\}},
$$

where κ is a positive constant, and define a two-layer ReLU network with cross-entropy loss as:

$$
\psi_\beta(z) := \ell(W_2 \sigma(W_1 x), y) = -\sum_{i=1}^{K} y_i \log \left(\frac{e^{W_{2,i}\sigma(W_1 x)}}{\sum_{k=1}^{K} e^{W_{2,k}\sigma(W_1 x)}} \right),
$$

where $W_{2,i}$ is the i-th row of W_2. We shall prove that ψ_β satisfies assumption (C1) and (C2).

Rewrite ψ_β as follows:

$$
\psi_\beta(z) = \log \left(\sum_{k=1}^{K} e^{W_{2,k}\sigma(W_1 x)} \right) - \sum_{i=1}^{K} y_i W_{2,i} \sigma(W_1 x) \ .
$$

Using the gradient of ψ_β with respect to $\theta(x) := W_2 \sigma(W_1 x) \in \mathbb{R}^K$ and applying the mean-value theorem, we have:

$$
\psi_\beta(z') - \psi_\beta(z) = \langle \nabla_\theta \ell(\theta + \tau(\theta' - \theta), y), \theta' - \theta \rangle, \quad \text{for some } \tau \in (0, 1).
$$

Using the Cauchy-Schwarz inequality, we have:

$$|\psi_\beta(z') - \psi_\beta(z)| \leq \|\nabla_\theta \ell(\theta + \tau(\theta' - \theta), y)\|_2 \|\theta' - \theta\|_2 \leq \sup_\theta \{\|\nabla_\theta \ell(\theta, y)\|_2\} \|\theta' - \theta\|_2 \ .$$

We have:

$$\begin{aligned}\|\nabla_\theta \ell(\theta, y)\|_2 &= \|\mathrm{softmax}(\theta) - y\|_2 \\ &= \sqrt{\|\mathrm{softmax}(\theta)\|_2^2 + \|y\|_2^2 - 2\mathrm{softmax}(\theta)^\top y} \\ &\leq \sqrt{1 + 1 - 0} = \sqrt{2}\ ,\end{aligned}$$

where the inequality holds because both softmax(θ) and y are probability vectors. Hence

$$|\psi_\beta(z') - \psi_\beta(z)| \leq \sqrt{2}\, \|\theta' - \theta\|_2 \ = \sqrt{2}\, \|W_2\sigma(W_1 x') - W_2\sigma(W_1 x)\|_2 .$$

For $Z^{(i)} = (x^{(i)}, y^{(i)}) \in \mathcal{Z}_N$, we introduce assumption

(D1): no training point lies on a ReLU facet,

and let

$$D^{(i)} = \mathrm{diag}\left(\mathbf{1}_{\left\{\left(W_1 x^{(i)}\right)_j > 0\right\}}\right) \in \mathbb{R}^{n_1 \times n_1},$$

then for $z \in \{Z^{(i)}\}$

$$\theta(x) = W_2 D^{(i)} W_1 x\ ,$$

and the Jacobian is given by

$$J^{(i)} := \nabla_x \theta(x) = W_2 D^{(i)} W_1\ .$$

For any z' that has x' in the same ReLU cell as $x^{(i)}$, we have:

$$\begin{aligned}|\psi_\beta(z') - \psi_\beta(z)| &\leq \sqrt{2}\left\|W_2 D^{(i)} W_1 x' - W_2 D^{(i)} W_1 x\right\|_2 = \sqrt{2}\left\|J^{(i)}(x' - x)\right\|_2 \\ &\leq \sqrt{2}\left\|J^{(i)}\right\|_2 \|x' - x\|_2 \\ &\leq \sqrt{2}\left\|J^{(i)}\right\|_2 d(z', z)\ .\end{aligned}$$

For any z' that has x' in a different ReLU cell as $x^{(i)}$, we define the straight path from $x^{(i)}$ to x' as

$$x(t) = x^{(i)} + t(x' - x^{(i)}), \quad t \in [0, 1]$$

and the hidden-unit pre-activations along the path as

$$h_j(t) := (W_1 x(t))_j, \quad \text{where each is an affine function of t and } j \in [n_1].$$

For each hidden unit j with $h_j(0)h_j(1) < 0$, there exist a unique $t \in (0, 1)$ such that $h_j(t) = 0$. Collect all such t, then we can sort them as

$$0 = t_0 < t_1 < ... < t_{m-1} < t_m = 1, \quad m - 1 \leq n_1.$$

For every $k \in \{0, ..., m-1\}$ define

$$x^{(k)} := x(t_k), \quad \therefore \quad D^{(k)} = \mathrm{diag}\left(\mathbf{1}_{\left\{\left(W_1 x^{(k)}\right)_j > 0\right\}}\right).$$

Because the activation mask is constant and equal to $D^{(k)}$ on the sub-segment $[x^{(k)}, x^{(k+1)}]$, we have:

$$
\begin{aligned}
|\psi_\beta(z') - \psi_\beta(z)| &= \left| \sum_{k=0}^{m-1} \left[\psi_\beta((x^{(k+1)}, y)) - \psi_\beta((x^{(k)}, y)) \right] \right| \\
&\le \sum_{k=0}^{m-1} \left| \psi_\beta((x^{(k+1)}, y)) - \psi_\beta((x^{(k)}, y)) \right| \\
&\le \sqrt{2} \left(\max_{k \in \{0,\ldots,m-1\}} \left\| J^{(k)} \right\|_2 \right) \sum_{k=0}^{m-1} \left\| x^{(k+1)} - x^{(k)} \right\|_2 \overset{(*)}{\le} L_{(i)} \left\| x' - x^{(i)} \right\|_2 \\
&\le L_{(i)} d(z', z) \ ,
\end{aligned}
$$

where

$$
L_{(i)} := \sqrt{2} \left(\max_{D \in \mathcal{D}_{(i)}} \|W_2 D W_1\|_2 \right),
$$

$$
\mathcal{D}_{(i)} \ := \ \left\{ D \ \middle|\ \exists\, x \in \mathbb{R}^n,\ \exists\, 0 \le t_1 < t_2 \le 1 :\ x(t) = x^{(i)} + t\big(x - x^{(i)}\big) \in \mathcal{C}_D \ \forall t \in [t_1, t_2] \right\},
$$

$$
\mathcal{C}_D := \{x :\ D\, W_1 x > 0, (I - D) W_1 x < 0\},
$$

$$
(*) : \sum_{k=0}^{m-1} \left\| x^{(k+1)} - x^{(k)} \right\|_2 = \sum_{k=0}^{m-1} \left\| (t_{k+1} - t_k)(x' - x^{(i)}) \right\|_2 = \sum_{k=0}^{m-1} (t_{k+1} - t_k) \left\| x' - x^{(i)} \right\|_2 = \left\| x' - x^{(i)} \right\|_2
$$

Therefore, ψ_β satisfies Assumption (C1) with $L_\beta^{\{Z^{(i)}\}} = L_{(i)}$. Let

$$
L := \sqrt{2} \max_{D \in \mathcal{D}} \|W_2 D W_1\|_2 , \quad \mathcal{D} := \{ D(x) \,|\, x \in \mathbb{R}^n \} ,
$$

where $D(x) = \operatorname{diag}\big(\mathbf{1}_{\,W_1 x > 0}\big)$ is the usual ReLU mask. Fix an arbitrary $D \in \mathcal{D}$. By definition there exists a point $x^D \in \mathbb{R}^n$ that strictly satisfies the inequalities of that mask:

$$
D\, W_1 x^D > 0, \qquad (I - D) W_1 x^D < 0.
$$

Because the inequalities are strict, the cell $\mathcal{C}_D := \{x :\ D W_1 x > 0,\ (I - D) W_1 x < 0\}$ is an open set that contains x^D. By continuity, there is an $\varepsilon > 0$ such that every point within Euclidean distance ε of x^D still lies in $\mathcal{C}_D$. Now consider the straight segment joining the reference point $x^{(i)}$ to x^D:

$$
x(t) \ = \ x^{(i)} + t\big(x^D - x^{(i)}\big), \qquad t \in [0, 1].
$$

Choose $t_1 := 1 - \dfrac{\varepsilon}{\|x^{(i)} - x^D\|_2}$ and $t_2 := 1$. For every $t' \in [t_1, t_2]$ we have

$$
\|x(t') - x^D\|_2 = \|x^{(i)} + t'\big(x^D - x^{(i)}\big) - x^D\|_2 \le (1 - t_1)\, \|x^{(i)} - x^D\|_2 \le \varepsilon,
$$

hence $x(t') \in \mathcal{C}_D$. That means the mask stays constant and equal to D on the entire sub-segment $[t_1, t_2]$. Therefore $D \in \mathcal{D}_{(i)}$. Because the choice of D was arbitrary, we have shown the opposite inclusion $\mathcal{D} \subseteq \mathcal{D}_{(i)}$. The other direction $\mathcal{D}_{(i)} \subseteq \mathcal{D}$ is immediate from the definitions, so

$$
\mathcal{D}_{(i)} = \mathcal{D} \quad \therefore \quad L_{(i)} = L \ .
$$

Hence ψ_β satisfies Assumption (A1) on the whole input space with

$$
L_\beta^{\mathcal{Z}_N} = L = \sqrt{2} \max_{D \in \mathcal{D}} \|W_2 D W_1\|_2
$$

Next, we show that ψ_β satisfies Assumption (A2). Let

$$
D^* := \arg\max_{D \in \mathcal{D}} \|W_2 D W_1\|_2 , \quad J^* := W_2 D^* W_1, \quad \|J^*\|_2 = L/\sqrt{2}.
$$

For the strict ReLU cell

$$\mathcal{C}^* = \{x : D^* W_1 x > 0, (I - D^*) W_1 x < 0\},$$

write its recession cone

$$\operatorname{rec}(\mathcal{C}^*) = \{u \neq 0 : D^* W_1 u \geq 0, (I - D^*) W_1 u \leq 0\}.$$

Denote

$$\Omega := \operatorname{int}(\operatorname{rec}(\mathcal{C}^*)),$$
$$V := J^*(\Omega) \subset \mathbb{R}^K,$$
$$S_*^{K-1} := \{z : \|z\|_2 = \|J^*\|_2\}, \quad V_* := V \cap S_*^{K-1}.$$

We introduce an assumption:

(D2) Fix the training label $c \in \{1, ..., K\}$, there exists $j_+ \neq c$ and a unit vector $u \in \Omega$, such that

$$J^* u = w := \frac{\|J^*\|_2}{\sqrt{2}} (e_{j_+} - e_c), \quad \|w\|_2 = \|J^*\|_2, \quad w \in V_*$$

with e_j being the j-th canonical basis vector.

We have,

$$w_{j_+} - w_c \;:=\; \big(e_{j_+} - e_c\big)^\top w \;=\; \frac{\|J^*\|_2}{\sqrt{2}}\, (e_{j_+} - e_c)^\top (e_{j_+} - e_c) = L$$

Because $u \in \Omega$, each coordinate of $W_1(x^{(i)} + tu)$ moves strictly toward the sign prescribed by D^*. For every j put

$$t_j \;:=\; \max\Big\{0,\; -\frac{(W_1 x^{(i)})_j}{(W_1 u)_j}\Big\},$$

each denominator is non-zero and finite, hence $t_j < \infty$. Define the first hit time

$$\tau := \max_j t_j + \eta\ ,$$

where η is an infinitesimal amount and set

$$x^* := x^{(i)} + \tau u \in \mathcal{C}^*,\; z^* := (x^*, y^{(i)}).$$

By Assumption (A1), we have

$$\psi_\beta(z^*) - \psi_\beta(Z^{(i)}) \geq -L\tau.$$

For $t \geq 0$, stay on the ray

$$x(t) \;=\; x^* + t\, u, \qquad z(t) = (x(t), y^{(i)}),$$

which never leaves the strict cell $\mathcal{C}^*$ (u is in its recession cone). Inside that cell, the network reduces to an affine map; hence, the loss and the directional derivative along any internal ray exist and are continuous. The directional derivative of the loss in direction u is

$$\dot{\psi}_\beta\big(z(t)\big) := u^\top \nabla_x \psi_\beta(z(t)) = w^\top \big(\operatorname{softmax}\theta(t) - y^{(i)}\big).$$

Using the Cauchy-Schwarz inequality, we have for all $t \geq 0$

$$\Big|\dot{\psi}_\beta\big(z(t)\big)\Big| \leq \|w\|_2 \|\operatorname{softmax}\theta(t) - y^{(i)}\|_2 \leq \sqrt{2}\|J^*\|_2 = L \quad \therefore \quad \dot{\psi}_\beta\big(z(t)\big) \geq -L.$$

Because $w_{j_+} - w_c > 0$, $\theta_{j_+}(t) - \theta_c(t) = \big(\theta_{j_+}^{(i)} - \theta_c^{(i)}\big) + t\,(w_{j_+} - w_c) \to \infty$ as $t \to \infty$ hence, $\operatorname{softmax}_{j_+} \to 1$, $\operatorname{softmax}_c \to 0$ as $t \to \infty$. Because the label vector $y^{(i)}$ is one-hot with a 1 in position c,

$$\dot{\psi}_\beta(z(t)) = w_{j_+}\big[\operatorname{softmax}_{j_+}(\theta(t))\big] + w_c\big[\operatorname{softmax}_c(\theta(t)) - 1\big] + \sum_{k \notin \{j_+, c\}} w_k \operatorname{softmax}_k(\theta(t)).$$

Therefore $\dot{\psi}_\beta(z(t)) \to w_{j_+} - w_c$ as $t \to \infty$. Fix an accuracy level $0 < \varepsilon < L$ where $L = \sqrt{2}\,\|J^*\|_2$. By continuity, there exist $t_\varepsilon > 0$ such that

$$\dot{\psi}_\beta(z(t)) \geq L - \varepsilon/2 \quad \forall\, t \geq t_\varepsilon.$$

For any $T > t_\varepsilon$, using the mean-value theorem, we have:

$$\begin{aligned}
\psi_\beta\big(z(T)\big) - \psi_\beta\big(Z^{(i)}\big) &= \big[\psi_\beta\big(z(T)\big) - \psi_\beta\big(z(t_\varepsilon)\big)\big] + \big[\psi_\beta\big(z(t_\varepsilon)\big) - \psi_\beta(z^*)\big] + \Big[\psi_\beta(z^*) - \psi_\beta\big(Z^{(i)}\big)\Big] \\
&\geq (L - \varepsilon/2)\,(T - t_\varepsilon) - L\,t_\varepsilon - L\,\tau \\
&= (L - \varepsilon)(T + \tau) + \varepsilon(T/2 + t_\varepsilon/2 + \tau) - 2L(t_\varepsilon + \tau).
\end{aligned}$$

By choosing $T \geq \dfrac{4L(t_\varepsilon + \tau)}{\varepsilon} - t_\varepsilon - 2\tau + \delta$, we have

$$\psi_\beta\big(z(T)\big) - \psi_\beta\big(Z^{(i)}\big) \geq (L - \varepsilon)(T + \tau)$$

Finally, define the witness

$$\widetilde{Z}_\varepsilon^{(i)} := \big(x^* + T\,u,\; y^{(i)}\big)\,,$$

where the distance is

$$d(\widetilde{Z}_\varepsilon^{(i)}, Z^{(i)}) = \|x^* + T\,u - x^{(i)}\|_2 = \|x^{(i)} + \tau u + T\,u - x^{(i)}\|_2 = (\tau + T)\|u\|_2 = \tau + T \geq \delta.$$

Then we have:

$$\psi_\beta\big(\widetilde{Z}_\varepsilon^{(i)}\big) - \psi_\beta\big(Z^{(i)}\big) \;\geq\; (L - \varepsilon)\,d\big(\widetilde{Z}_\varepsilon^{(i)}, Z^{(i)}\big).$$

All in all, Assumption (A2) holds under Assumptions (D1 & D2). Therefore, by Theorem 3.2,

$$\sup_{\mathbb{P}: W_{d,1}(\mathbb{P}, \mathbb{P}_N) \leq \delta} \mathrm{E}_{\mathbb{P}}\big[\psi_\beta(Z)\big] \;=\; \mathrm{E}_{\mathbb{P}_N}\big[\psi_\beta(Z)\big] + L\,\delta$$

General Case Fix an integer $H \geq 1$. We consider an $(H+1)$-layer ReLU network with parameters

$$\beta \;=\; (W_1, \dots, W_{H+1}), \quad W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}\ (n_0 = n,\ n_{H+1} = K),$$

and define

$$x_0 = x, \qquad x_\ell = \sigma\big(W_\ell\, x_{\ell-1}\big) \quad (\ell = 1, \dots, H), \qquad \theta(x) = W_{H+1}\, x_H.$$

The loss is

$$\psi_\beta(z) \;=\; \ell\big(\theta(x), y\big) \;=\; -\sum_{k=1}^{K} y_k \log \frac{e^{\theta_k(x)}}{\sum_{j=1}^{K} e^{\theta_j(x)}}.$$

On each strict ReLU cell, we freeze the activation masks

$$D_\ell(x) \;=\; \mathrm{diag}\big(\mathbf{1}_{\{W_\ell\, x_{\ell-1} > 0\}}\big), \quad \ell = 1, \dots, H,$$

so that

$$J(x) \;=\; \nabla_x \theta(x) \;=\; W_{H+1}\, D_H(x)\, W_H\; \cdots\; D_1(x)\, W_1$$

is constant on that cell.

Let $Z^{(i)} = (x^{(i)}, y^{(i)})$ be any training sample, and let $\mathcal{D}_{(i)}$ be the collection of all mask-tuples $(D_1, \dots, D_H)$ that arise along the straight segment from $x^{(i)}$ to any x'. Then, exactly as in the two-layer case, splitting into sub-segments, each lying inside one cell, and applying the mean-value theorem plus $\|\nabla_\theta \ell\|_2 \leq \sqrt{2}$ gives

$$\big|\psi_\beta(z') - \psi_\beta(z)\big| \;\leq\; \sqrt{2} \max_{(D_1, \dots, D_H) \in \mathcal{D}_{(i)}} \big\|W_{H+1} D_H \cdots D_1 W_1\big\|_2\, \|x' - x\|_2 \;\leq\; L_{(i)}\, d(z', z),$$

where

$$L_{(i)} := \sqrt{2} \max_{(D_1,\dots,D_H)\in\mathcal{D}_{(i)}} \big\|W_{H+1}D_H\cdots D_1W_1\big\|_2.$$

A standard "ε-ball" argument then shows $\mathcal{D}_{(i)}$ exhausts all global mask-tuples $\mathcal{D}$, so in fact

$$L_\beta^{\mathcal{Z}_N} \;=\; L := \sqrt{2} \max_{(D_1,\dots,D_H)\in\mathcal{D}} \big\|W_{H+1}D_H\cdots D_1W_1\big\|_2$$

and Assumption (A1) holds on the whole input space.

Let $(D_1^*,\dots,D_H^*)$ attain the maximum defining L, and set

$$J^* = W_{H+1}D_H^*W_H\cdots D_1^*W_1, \qquad \|J^*\|_2 = \frac{L}{\sqrt{2}}.$$

Denote the corresponding strict cell $\mathcal{C}^* = \{x : D_\ell^*W_\ell x_{\ell-1} > 0,\, (I - D_\ell^*)W_\ell x_{\ell-1} < 0\ \forall\ell\}$ and its recession cone $\operatorname{rec}(\mathcal{C}^*)$. Under the assumptions

(D1) $x^{(i)}$ not on any ReLU facet,

(D2) $\exists$ unit vector $u \in \operatorname{int}(\operatorname{rec}(\mathcal{C}^*))$ with $J^*u = \dfrac{\|J^*\|_2}{\sqrt{2}}(e_{j_+} - e_c)$,

the ray $x(t) = x^* + t\,u$ (starting from some $x^* \in \mathcal{C}^*$) keeps all masks frozen and makes the directional derivative $\dot\psi_\beta(z(t)) \to L$. Integrating as before yields, for any $0 < \varepsilon < L$, a witness $\widetilde{Z}_\varepsilon^{(i)}$ such that

$$\psi_\beta\big(\widetilde{Z}_\varepsilon^{(i)}\big) - \psi_\beta\big(Z^{(i)}\big) \;\ge\; (L-\varepsilon)\, d\big(\widetilde{Z}_\varepsilon^{(i)}, Z^{(i)}\big),$$

so Assumption (A2) holds under (D1 & D2).

By Theorem 3.2, combining (A1) and (A2) gives the formula

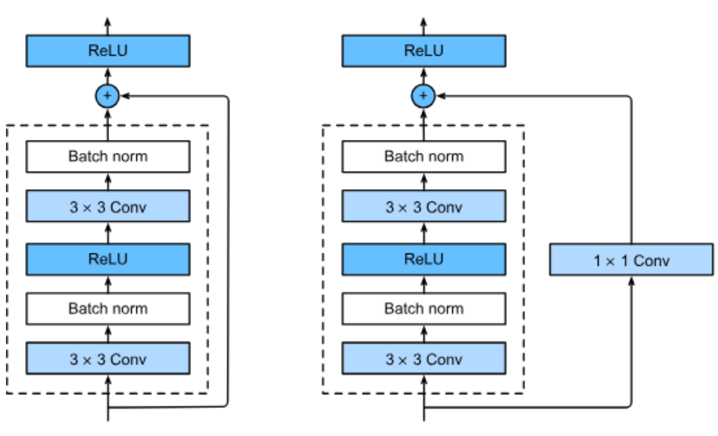
$$\sup_{P:\,W_{d,1}(P,P_N)\le\delta} \mathrm{E}_P\big[\psi_\beta(Z)\big] \;=\; \mathrm{E}_{P_N}\big[\psi_\beta(Z)\big] \;+\; L\,\delta.$$

4.2.2 ResNet

Toy Case Let $s \in \mathbb{N}$ be the stride of the first 3×3 convolution. Assume, for simplicity, that s divides both H and W. Then

$$x^\flat \in \mathbb{R}^{d_{\text{in}}}, \qquad z^\flat \in \mathbb{R}^{d_{\text{out}}},$$
$$d_{\text{in}} \;=\; C_{\text{in}}\,H\,W, \qquad d_{\text{out}} \;=\; C_{\text{out}}\,\tfrac{H}{s}\,\tfrac{W}{s},$$

and, in typical ResNet implementations, the channel dimension expands in proportion to the stride, e.g. $C_{\text{out}} = s\,C_{\text{in}}$ (this choice is not required for the analysis; any $C_{\text{out}} \ge 1$ works).



For our analysis, each "3×3 Conv → BN" pair is absorbed into a single matrix:

$$W_1 \in \mathbb{R}^{d_{\text{mid}} \times d_{\text{in}}}, \qquad W_2 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{mid}}},$$

where W_1 incorporates stride s and W_2 has stride 1. The skip path is a 1×1 projection

$$P \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}},$$

which defaults to the identity map whenever $d_{\text{out}} = d_{\text{in}}$ (in practice, this is the case $s = 1$ and $C_{\text{out}} = C_{\text{in}}$). With ReLU masks $D_1 = \operatorname{diag}\big(\mathbf{1}_{W_1 x^\flat > 0}\big)$, define

$$\mathcal{F}(x^\flat) \;=\; W_2\, W_1\, D_1\, x^\flat.$$

$$z^\flat \;=\; P\, x^\flat + \mathcal{F}(x^\flat), \qquad \tilde{z}^\flat := \sigma(z^\flat).$$

For a single-block toy analysis we identify $\tilde{z}^\flat \equiv z^\flat$. Holding D_1 fixed,

$$\nabla_{x^\flat} z^\flat \;=\; P \;+\; W_2\, D_2\, W_1\, D_1.$$

General Case

4.2.3 Transformer

Toy Case

General Case

References

[BHJO25] Xingjian Bai, Guangyi He, Yifan Jiang, and Jan Obloj. Wasserstein distributional adversarial training for deep neural networks, 2025.

[BKM19] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, September 2019.

[CLT24] Hong T. M. Chu, Meixia Lin, and Kim-Chuan Toh. Wasserstein distributionally robust optimization and its tractable regularization formulations, 2024.

[GCK20] Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization, 2020.

[GFPC20] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing lipschitz continuity, 2020.

[KKWP20] Yongchan Kwon, Wonyoung Kim, Joong-Ho Won, and Myunghee Cho Paik. Principled learning method for wasserstein distributionally robust optimization with local perturbations, 2020.

[KPM21] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention, 2021.

[LGL+23] Thibaud Leteno, Antoine Gourru, Charlotte Laclau, Rémi Emonet, and Christophe Gravier. Fair text classification with wasserstein independence, 2023.

[OSHL19] Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust language modeling, 2019.

[SNVD20] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training, 2020.

[SV19] Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation, 2019.

[ZYK+23] Qing Zhang, Yi Yan, Fannie Kong, Shifei Chen, and Linfeng Yang. A wasserstein-based distributionally robust neural network for non-intrusive load monitoring. *Frontiers in Energy Research*, 11, 04 2023.