A. Proofs for Section 3

A.1. Proofs for Section 3.1

A.1.1. Proof of Theorem 1

To start the proof of Theorem 1, we need the following lemma.

Lemma 1. *For any $\boldsymbol{w}$ and r, we have*

$$\sup_{P\in B_{\mathsf{W}_\infty}(P_0,r)} R_P(\boldsymbol{w}) = \mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right]. \tag{16}$$

Proof. Let $T_r^{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{x} + \arg\max_{\{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty\le r\}} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})$ with $\boldsymbol{x}$ is an input data. The existence of $T_r^{\boldsymbol{w}}(\boldsymbol{x})$ is guaranteed by the continuity of $f(\boldsymbol{w},\boldsymbol{x})$. P_r is the distribution of $T_r^{\boldsymbol{w}}(\boldsymbol{x})$ with $\boldsymbol{x}\sim P_0$. Then

$$\mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right] = \mathbb{E}_{P_r}[f(\boldsymbol{w},\boldsymbol{x})]. \tag{17}$$

Since

$$\mathsf{W}_\infty(P_0,P_r) \le \mathbb{E}_{P_0}[\|\boldsymbol{x} - T_r^{\boldsymbol{w}}(\boldsymbol{x})\|_\infty] \le r, \tag{18}$$

we have

$$\mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right] \le \sup_{P\in B_{\mathsf{W}_\infty}(P_0,r)} R_P(\boldsymbol{w}). \tag{19}$$

On the other hand, let $P_r^* \in \arg\max_{P\in B_{\mathsf{W}_\infty}(P_0,r)} R_P(\boldsymbol{w})$. Due to Kolmogorov's theorem, P_r^* can be distribution of some random vector $\boldsymbol{z}$, due to the definition of W_∞-distance, we have $\|\boldsymbol{z}-\boldsymbol{x}\|_\infty \le r$ holds almost surely. Then we conclude

$$\sup_{P\in B_{\mathsf{W}_\infty}(P_0,r)} R_P(\boldsymbol{w}) = R_{P_r^*}(\boldsymbol{w}) = \mathbb{E}_{P_r^*}[f(\boldsymbol{w},\boldsymbol{z})] \le \mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right]. \tag{20}$$

Thus, we get the conclusion. □

This lemma shows that the distributional perturbation measured by W_∞-distance is equivalent to input perturbation. Hence we can study $\mathsf{W}_{\inf}$-distributional robustness through $\ell_{\inf}$-input-robustness. The basic tool for our proof is the covering number, which is defined as follows.

Definition 2. *(Wainwright, 2019) A r-cover of $(\mathcal{X},\|\cdot\|_p)$ is any point set $\{\boldsymbol{u}_i\}\subseteq\mathcal{X}$ such that for any $\boldsymbol{u}\in\mathcal{X}$, there exists $\boldsymbol{u}_i$ satisfies $\|\boldsymbol{u}-\boldsymbol{u}_i\|_p\le r$. The covering number $\mathcal{N}(r,\mathcal{X},\|\cdot\|_p)$ is the cardinality of the smallest r-cover.*

Now we are ready to give the proof of Theorem 1 which is motivated by (Xu & Mannor, 2012).

Proof of Theorem 1. We can construct a r-cover to $(\mathcal{X},\|\cdot\|_2)$ then $\mathcal{N}(r,\mathcal{X},\|\cdot\|_2)\le(2d_0)^{(2D/r^2+1)}=N$, because the $\mathcal{X}$ can be covered by a polytope with ℓ_2-diameter smaller than $2D$ and $2d_0$ vertices, see (Vershynin, 2018) Theorem 0.0.4 for details. Due to the geometrical structure, we have $\mathcal{N}(r,\mathcal{X},\|\cdot\|_\infty)\le(2d_0)^{(2D/r^2+1)}$. Then, there exists $(C_1,\cdots,C_N)$ covers $(\mathcal{X},\|\cdot\|_\infty)$ where C_i is disjoint with each other, and $\|\boldsymbol{u}-\boldsymbol{v}\|_\infty\le r$ for any $\boldsymbol{u},\boldsymbol{v}\in C_i$. This can be constructed by $C_i=\hat{C}_i\bigcap\left(\bigcup_{j=1}^{i-1}\hat{C}_j\right)^c$ with $(\hat{C}_1,\cdots,\hat{C}_N)$ covers $(\mathcal{X},\|\cdot\|_\infty)$, and the diameter of each $\hat{C}_i$ is smaller than r since

$\mathcal{N}(r, \mathcal{X}, \|\cdot\|_\infty) \le N$. Let $A_j = \{\boldsymbol{x}_i : \boldsymbol{x}_i \in C_j\}$, and $|A_j|$ is the cardinality of A_j. Due to Lemma 1, we have

$$\begin{aligned}
\left|\sup_{P\in B_{\mathsf{W}_\infty}(P_0,r_0)} R_P(\boldsymbol{w}) - R_{P_n}(\boldsymbol{w})\right| &= \left|\mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r_0} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right] - R_{P_n}(\boldsymbol{w})\right| \\
&= \left|\sum_{j=1}^N \mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r_0} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{x}\in C_j\right] P_0(C_j) - R_{P_n}(\boldsymbol{w})\right| \\
&\le \left|\sum_{j=1}^N \mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r_0} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{x}\in C_j\right] \frac{|A_j|}{n} - \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{w},\boldsymbol{x}_i)\right| \\
&+ \left|\sum_{j=1}^N \mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty\le r_0} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{x}\in C_j\right] \left(\frac{|A_j|}{n} - P_0(C_j)\right)\right| \\
&\le \left|\frac{1}{n}\sum_{j=1}^N \sum_{\boldsymbol{x}_i\in C_j} \sup_{\boldsymbol{x}\in C_j + B_\infty(\boldsymbol{0},r_0)} |f(\boldsymbol{w},\boldsymbol{x}) - f(\boldsymbol{w},\boldsymbol{x}_i)|\right| + M\sum_{j=1}^N \left|\frac{|A_j|}{n} - P_0(C_j)\right| \\
&\overset{a}{\le} \frac{1}{n}\sum_{i=1}^n \sup_{\|\boldsymbol{\delta}\|_\infty\le 2r} |f(\boldsymbol{w},\boldsymbol{x}_i+\boldsymbol{\delta}) - f(\boldsymbol{w},\boldsymbol{x}_i)| + M\sum_{j=1}^N \left|\frac{|A_j|}{n} - P_0(C_j)\right| \\
&\le \epsilon + M\sum_{j=1}^N \left|\frac{|A_j|}{n} - P_0(C_j)\right|.
\end{aligned} \tag{21}$$

Here a is due to $C_j + B_\infty(\boldsymbol{0}, r) \subseteq B_\infty(\boldsymbol{x}_i, 2r)$ when $\boldsymbol{x}_i \in C_j$, since ℓ_∞-diameter of C_j is smaller than r. The last inequality is due to $(2r, \epsilon, P_n, \infty)$-robustness of $f(\boldsymbol{w}, \boldsymbol{x})$. On the other hand, due to Proposition A6.6 in (van der Vaart & Wellner, 2000), we have

$$\mathbb{P}\left(\sum_{j=1}^N \left|\frac{|A_j|}{n} - P_0(C_j)\right| \ge \theta\right) \le 2^N \exp\left(\frac{-n\theta^2}{2}\right). \tag{22}$$

Combine this with (21), due to the value of N, we get the conclusion. □

A.1.2. Proof of Theorem 2

There is a little difference of proving Theorem 2 compared with Theorem 1. Because the out-distribution P constrained in $B_{\mathsf{W}_\infty}(P_0, r)$ only correspond with OOD data that contained in a ℓ_∞-ball of in-distribution data almost surely, see Lemma 1 for a rigorous description. Hence, we can utilize ℓ_∞-robustness of model to derive the OOD generalization under W_∞-distance by Theorem 1. However, in the regime of W_2-distance, roughly speaking, the transformed OOD data $T_r^{\boldsymbol{w}}(\boldsymbol{x})$ is contained in a ℓ_2-ball of $\boldsymbol{x}$ in expectation. Thus, Lemma 1 is invalid under W_2-distance.

To discuss the OOD generalization under W_2-distance, we need to give a delicate characterization to the distribution $P \in B_{\mathsf{W}_2}(P_0, r)$. First, we need the following lemma.

Lemma 2. *For any r and $\boldsymbol{w}$, let $P_r^* \in \arg\max_{P\in B_{\mathsf{W}_2}(P_0,r)} R_P(\boldsymbol{w})$. Then, there exists a mapping $T_r^{\boldsymbol{w}}(\boldsymbol{x})$ such that $T_r^{\boldsymbol{w}}(\boldsymbol{x}) \sim P_r^*$ with $\boldsymbol{x} \sim P_0$.*

Proof. The proof of Theorem 6 in (Sinha et al., 2018) shows that

$$R_{P_r^*}(\boldsymbol{w}) = \sup_{P\in B_{\mathsf{W}_2}(P_0,r)} R_P(\boldsymbol{w}) = \inf_{\lambda\ge 0}\sup_{P,\pi\in(P,P_0)}\left(\int_{\mathcal{X}\times\mathcal{X}} f(\boldsymbol{w},\boldsymbol{x}) - \lambda\|\boldsymbol{x}-\boldsymbol{z}\|^2 d\pi(\boldsymbol{x},\boldsymbol{z}) + \lambda r\right). \tag{23}$$

We next show that the supremum over π in the last equality is attained by the joint distribution $(T_r^{\boldsymbol{w}}(\boldsymbol{x}), \boldsymbol{x})$, which implies our conclusion. For any $\lambda > 0$, we have

$$\sup_{P,\pi\in(P,P_0)}\left(\int_{\mathcal{X}\times\mathcal{X}} f(\boldsymbol{w},\boldsymbol{x}) - \lambda\|\boldsymbol{x}-\boldsymbol{z}\|^2 d\pi(\boldsymbol{x},\boldsymbol{z})\right) \le \int_{\mathcal{X}} \sup_{\boldsymbol{x}}\left(f(\boldsymbol{w},\boldsymbol{x}) - \lambda\|\boldsymbol{x}-\boldsymbol{z}\|^2\right) dP_0(\boldsymbol{z}), \tag{24}$$

due to the supremum in the left hand side is taken over P and π. On the other hand, let $P(\cdot \mid \boldsymbol{z})$ and $\boldsymbol{x}(\cdot)$ respectively be the regular conditional distribution on $\mathcal{X}$ with $\boldsymbol{z}$ given and the function on $\mathcal{X}$. Since $P(\cdot \mid \boldsymbol{z})$ is measurable,

$$\begin{aligned}\sup_{P,\pi\in(P,P_0)}\left(\int_{\mathcal{X}\times\mathcal{X}} f(\boldsymbol{w},\boldsymbol{x})-\lambda\|\boldsymbol{x}-\boldsymbol{z}\|^2 d\pi(\boldsymbol{x},\boldsymbol{z})\right) &\geq \sup_{P(\cdot|\boldsymbol{z})}\left(\int_{\mathcal{X}\times\mathcal{X}} f(\boldsymbol{w},\boldsymbol{x})-\lambda\|\boldsymbol{x}-\boldsymbol{z}\|^2 dP(\boldsymbol{x}\mid\boldsymbol{z})dP_0(\boldsymbol{z})\right)\\ &\geq \sup_{\boldsymbol{x}(\cdot)}\left(\int_{\mathcal{X}} f(\boldsymbol{w},\boldsymbol{x}(\boldsymbol{z}))-\lambda\|\boldsymbol{x}(\boldsymbol{z})-\boldsymbol{z}\|^2 dP_0(\boldsymbol{z})\right)\\ &\geq \int_{\mathcal{X}}\sup_{\boldsymbol{x}}\left(f(\boldsymbol{w},\boldsymbol{x})-\lambda\|\boldsymbol{x}-\boldsymbol{z}\|^2\right)dP_0(\boldsymbol{z}).\end{aligned} \tag{25}$$

Thus, we get the conclusion. □

Proof of Theorem 2. Similar to the proof of Theorem 1, we can construct a disjoint cover $(C_1,\cdots,C_N)$ to $(\mathcal{X},\|\cdot\|_2)$ such that $N\leq(2d_0)^{(2\epsilon^2D/r^2+1)}$, and the l_2-diameter of each C_i is smaller than r/ϵ. Let $P_r^*\in\arg\max_{P\in B_{\mathsf{W}_2}(P_0,r)}R_P(\boldsymbol{w})$, by Lemma 2, we have

$$\begin{aligned}\sup_{P\in B_{\mathsf{W}_2}(P_0,r)} R_P(\boldsymbol{w}) &= R_{P_r^*}(\boldsymbol{w})\\ &= \mathbb{E}_{P_0}\left[f(\boldsymbol{w},T_r^{\boldsymbol{w}}(\boldsymbol{x}))\right]\\ &= \mathbb{E}_{P_0}\left[f(\boldsymbol{w},T_r^{\boldsymbol{w}}(\boldsymbol{x}))\left(\mathbf{1}_{T_r^{\boldsymbol{w}}(\boldsymbol{x})\in B_2(\boldsymbol{x},r/\epsilon)}+\mathbf{1}_{T_r^{\boldsymbol{w}}(\boldsymbol{x})\notin B_2(\boldsymbol{x},r/\epsilon)}\right)\right]\\ &\leq \mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_2\leq r/\epsilon} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right]+M\mathbb{P}(T_r^{\boldsymbol{w}}(\boldsymbol{x})\notin B_2(\boldsymbol{x},r/\epsilon)).\end{aligned} \tag{26}$$

Due to the definition of $T_r^{\boldsymbol{w}}(\boldsymbol{x})$, by Markov's inequality, we have

$$\left(\frac{r}{\epsilon}\right)\mathbb{P}(T_r^{\boldsymbol{w}}(\boldsymbol{x})\notin B_2(\boldsymbol{x},r/\epsilon))\leq\int_{\mathcal{X}}\|T_r^{\boldsymbol{w}}(\boldsymbol{x})-\boldsymbol{x}\|^2dP_0(\boldsymbol{x})=\mathsf{W}_2(P_0,P_r^*)\leq r. \tag{27}$$

Plugging this into (26), and due to the definition of Wasserstein distance, we have

$$\mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_2\leq r/\epsilon} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right]\leq\sup_{P\in B_{\mathsf{W}_2}(P_0,r)}R_P(\boldsymbol{w})\leq\mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_2\leq r/\epsilon} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right]+M\epsilon. \tag{28}$$

Similar to the proof of Theorem 1, due to the model is $(2r/\epsilon,\epsilon,P_n,2)$-robust, we have

$$\left|\mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_2\leq r/\epsilon} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})\right]-R_{P_n}(\boldsymbol{w})\right|\leq\epsilon+M\sqrt{\frac{(2d_0)^{(2\epsilon^2D/r^2+1)}\log 2+2\log(1/\theta)}{n}} \tag{29}$$

holds with probability at least $1-\theta$. Combining this with (28), we get the conclusion. □

A.2. Proofs for Section 3.2

The proof of Theorem 3 is same for $p\in\{2,\infty\}$, we take $p=\infty$ as an example. Before providing the proof, we first give a lemma to characterize the convergence rate of the first inner loop in Algorithm 1.

Lemma 3. *For any $\boldsymbol{w},\boldsymbol{x}\in\{\boldsymbol{x}_i\}$, and r, there exists $\boldsymbol{\delta}^*\in\arg\max_{\{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty\leq r\}} f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta})$ such that*

$$\|\boldsymbol{\delta}_{K+1}-\boldsymbol{\delta}^*\|^2\leq\left(1-\frac{\mu_{\boldsymbol{x}}}{L_{22}}\right)^K\|\boldsymbol{\delta}_1-\boldsymbol{\delta}^*\|^2 \tag{30}$$

when $\boldsymbol{\delta}_{k+1}=\mathrm{Proj}_{B_\infty(\mathbf{0},r)}\left(\boldsymbol{\delta}_k+\eta_{\boldsymbol{x}}\nabla_{\boldsymbol{x}}f(\boldsymbol{w},\boldsymbol{x}+\boldsymbol{\delta}_k)\right)$ with $\eta_{\boldsymbol{x}}=1/L_{22}$.

Proof. The existence of $\boldsymbol{\delta}^*$ is due to the continuity of $f(\boldsymbol{w}, \cdot)$. Then

$$\begin{aligned} f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}^*) - f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_{k+1}) &= f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}^*) - f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_k) + f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_k) - f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_{k+1}) \\ &\overset{a}{\leq} \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_k), \boldsymbol{\delta}^* - \boldsymbol{\delta}_k \rangle - \frac{\mu_{\boldsymbol{x}}}{2} \|\boldsymbol{\delta}_k - \boldsymbol{\delta}^*\|^2 \\ &+ \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_k), \boldsymbol{\delta}_k - \boldsymbol{\delta}_{k+1} \rangle + \frac{L_{22}}{2} \|\boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k\|^2 \\ &= \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_k), \boldsymbol{\delta}^* - \boldsymbol{\delta}_{k+1} \rangle - \frac{\mu_{\boldsymbol{x}}}{2} \|\boldsymbol{\delta}_k - \boldsymbol{\delta}^*\|^2 + \frac{L_{22}}{2} \|\boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k\|^2 \\ &\overset{b}{\leq} L_{22} \langle \boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k, \boldsymbol{\delta}^* - \boldsymbol{\delta}_{k+1} \rangle - \frac{\mu_{\boldsymbol{x}}}{2} \|\boldsymbol{\delta}_k - \boldsymbol{\delta}^*\|^2 + \frac{L_{22}}{2} \|\boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k\|^2 \\ &= L_{22} \langle \boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k, \boldsymbol{\delta}^* - \boldsymbol{\delta}_k \rangle - \frac{\mu_{\boldsymbol{x}}}{2} \|\boldsymbol{\delta}_k - \boldsymbol{\delta}^*\|^2 - \frac{L_{22}}{2} \|\boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k\|^2, \end{aligned} \tag{31}$$

where a is due to the L_{22}-Lipschitz continuity of $\nabla_{\boldsymbol{x}} f(\boldsymbol{w}, \boldsymbol{x})$ and strongly convexity, b is because the property of projection (see Lemma 3.1 in (Bubeck, 2014)). Then we get

$$\begin{aligned} \|\boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}^*\|^2 &= \|\boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k\|^2 + \|\boldsymbol{\delta}_k - \boldsymbol{\delta}^*\|^2 + 2\langle \boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k, \boldsymbol{\delta}_k - \boldsymbol{\delta}^* \rangle \\ &\leq \left(1 - \frac{\mu_{\boldsymbol{x}}}{L_{22}}\right) \|\boldsymbol{\delta}_k - \boldsymbol{\delta}^*\|^2 \end{aligned} \tag{32}$$

by plugging (31) into the above equality and $f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}^*) - f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta}_{k+1}) \geq 0$. Thus, we get the conclusion. □

This lemma shows that the inner loop in Algorithm 1 can efficiently approximate the worst-case perturbation for any $\boldsymbol{w}_t$ and $\boldsymbol{x}_i$. Now we are ready to give the proof of Theorem 3.

We need the following lemma, which is Theorem 6 in (Rakhlin et al., 2012).

Lemma 4. *Let $\{\xi_1, \cdots, \xi_t\}$ be a martingale difference sequence with a uniform upper bound b. Let $V_t = \sum_{j=1}^t \mathsf{Var}(\xi_j \mid \mathcal{F}_{j-1})$ with $\mathcal{F}_j$ is the σ-field generated by $\{\xi_1, \cdots, \xi_j\}$. Then for every a and $v > 0$,*

$$\mathbb{P}\left(\bigcup_{s \leq t} \left(\left\{ \sum_{j=1}^t \xi_j \geq a \right\} \bigcap \{V_t \leq v\} \right) \right) \leq \exp\left(\frac{-a^2}{2(v + ba)} \right). \tag{33}$$

This is a type of Bennett's inequality which is sharper compared with Azuma-Hoeffding's inequality when the variance v is much smaller than uniform bound b.

A.2.1. Proof of Theorem 3

Proof. With a little abuse of notation, let $r(p) = r$ and define $g(\boldsymbol{w}, \boldsymbol{x}) = \sup_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq r} f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta})$. Lemma A.5 in (Nouiehed et al., 2019) implies $g(\boldsymbol{w}, \boldsymbol{x})$ has $L_{11} + \frac{L_{12}L_{21}}{\mu_{\boldsymbol{x}}}$-Lipschitz continuous gradient with respect to $\boldsymbol{w}$ for any specific $\boldsymbol{x}$. Then $\tilde{R}_{P_n}(\boldsymbol{w})$ has $L = L_{11} + \frac{L_{12}L_{21}}{\mu_{\boldsymbol{x}}}$-Lipschitz continuous gradient. Let $\boldsymbol{x}^* \in \boldsymbol{x} + \arg\max_{\{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq r\}} f(\boldsymbol{w}, \boldsymbol{x}+\boldsymbol{\delta})$, due to the Lipschitz gradient of $\tilde{R}_{P_n}(\boldsymbol{w})$,

$$\begin{aligned} \tilde{R}_{P_n}(\boldsymbol{w}_{t+1}) - \tilde{R}_{P_n}(\boldsymbol{w}_t) &\leq \langle \nabla \tilde{R}_{P_n}(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle + \frac{L}{2} \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \\ &= -\eta_{\boldsymbol{w}_t} \langle \nabla \tilde{R}_{P_n}(\boldsymbol{w}_t), \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_K) \rangle + \frac{\eta_{\boldsymbol{w}_t}^2 L}{2} \|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_K)\|^2 \\ &= -\eta_{\boldsymbol{w}_t} \|\nabla \tilde{R}_{P_n}(\boldsymbol{w}_t)\|^2 + \eta_{\boldsymbol{w}_t} \langle \nabla \tilde{R}_{P_n}(\boldsymbol{w}_t), \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t}^*) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_K) \rangle \\ &+ \eta_{\boldsymbol{w}_t} \langle \nabla \tilde{R}_{P_n}(\boldsymbol{w}_t), \nabla \tilde{R}_{P_n}(\boldsymbol{w}_t) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t}^*) \rangle + \frac{\eta_{\boldsymbol{w}_t}^2 L}{2} \|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_K)\|^2. \end{aligned} \tag{34}$$

Here the last equality is due to $\nabla_{\boldsymbol{w}} g(\boldsymbol{w}, \boldsymbol{x}) = \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x}^*)$ (Similar to Danskin's theorem, see Lemma A.5 in (Nouiehed et al., 2019)), and $\boldsymbol{x}_{i_t}^*$ is the local maxima approximated by $\boldsymbol{x}_{i_t} + \boldsymbol{\delta}_K$ in Lemma 3. By taking expectation to $\boldsymbol{w}_{t+1}$ with $\boldsymbol{w}_t$

given in the both side of the above equation, Jesen's inequality, combining Lemma 3 and $\eta_{\boldsymbol{w}_t} = 1/\mu_{\boldsymbol{w}} t$,

$$\begin{aligned}
\mathbb{E}[\tilde{R}_{P_n}(\boldsymbol{w}_{t+1})] - \tilde{R}_{P_n}(\boldsymbol{w}^*) &\leq \tilde{R}_{P_n}(\boldsymbol{w}_t) - \tilde{R}_{P_n}(\boldsymbol{w}^*) - \eta_{\boldsymbol{w}_t}\|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_t)\|^2 \\
&\quad + \mathbb{E}\left[\eta_{\boldsymbol{w}_t}\|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_t)\|\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}^*_{i_t}) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x}_{i_t} + \boldsymbol{\delta}_K)\|\right] + \frac{\eta^2_{\boldsymbol{w}_t} G^2 L}{2} \\
&\leq \tilde{R}_{P_n}(\boldsymbol{w}_t) - \tilde{R}_{P_n}(\boldsymbol{w}^*) - \eta_{\boldsymbol{w}_t}\|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_t)\|^2 + \eta_{\boldsymbol{w}_t}\|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_t)\|\left(1 - \frac{\mu_{\boldsymbol{x}}}{L_{22}}\right)^K \mathbb{E}\left[\|\boldsymbol{\delta}_1 - \boldsymbol{\delta}^*_{i_t}\|^2\right] + \frac{\eta^2_{\boldsymbol{w}_t} G^2 L}{2} \\
&\leq (1 - 2\mu_{\boldsymbol{w}}\eta_{\boldsymbol{w}_t})\left(\tilde{R}_{P_n}(\boldsymbol{w}_t) - \tilde{R}_{P_n}(\boldsymbol{w}^*)\right) + \eta^2_{\boldsymbol{w}_t} G^2 L \\
&= \left(1 - \frac{2}{t}\right)\left(\tilde{R}_{P_n}(\boldsymbol{w}_t) - \tilde{R}_{P_n}(\boldsymbol{w}^*)\right) + \frac{G^2 L}{\mu^2_{\boldsymbol{w}} t^2}.
\end{aligned} \tag{35}$$

Here the third inequality is because

$$\eta_{\boldsymbol{w}_t}\|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_t)\|\left(1 - \frac{\mu_{\boldsymbol{x}}}{L_{22}}\right)^K \|\boldsymbol{\delta}_1 - \boldsymbol{\delta}^*_{i_t}\|^2 \leq \eta_{\boldsymbol{w}_t} G\left(1 - \frac{\mu_{\boldsymbol{x}}}{L_{22}}\right)^K 4 d_0 r^2 \leq \frac{\eta^2_{\boldsymbol{w}_t} G^2 L}{2}, \tag{36}$$

for any $\boldsymbol{\delta}^*_{i_t}$, since

$$K \log\left(1 - \frac{\mu_{\boldsymbol{x}}}{L_{22}}\right) \leq -K\frac{\mu_{\boldsymbol{x}}}{L_{22}} \leq \log\left(\frac{GL}{8T\mu_{\boldsymbol{w}} d_0 r^2}\right). \tag{37}$$

Then by induction,

$$\begin{aligned}
\mathbb{E}[\tilde{R}_{P_n}(\boldsymbol{w}_{t+1})] - \tilde{R}_{P_n}(\boldsymbol{w}^*) &\leq \frac{G^2 L}{\mu^2_{\boldsymbol{w}}} \sum_{j=2}^{t} \frac{1}{j^2} \prod_{k=j+1}^{t}\left(1 - \frac{2}{k}\right) \\
&= \frac{G^2 L}{\mu^2_{\boldsymbol{w}}} \sum_{j=2}^{t} \frac{1}{j^2} \frac{(j-1)j}{(t-1)t} \\
&\leq \frac{G^2 L}{t\mu^2_{\boldsymbol{w}}}.
\end{aligned} \tag{38}$$

Thus we get the first conclusion of convergence in expectation by taking $t = T$ for $t \geq 2$. For the second conclusion, let us define $\xi_t = \langle \nabla\tilde{R}_{P_n}(\boldsymbol{w}_t), \nabla\tilde{R}_{P_n}(\boldsymbol{w}_t) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}^*_{i_t})\rangle$. Then Schwarz inequality implies that

$$|\xi_t| \leq \|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_t)\|\|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_t) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t, \boldsymbol{x}^*_{i_t})\| \leq 2G^2. \tag{39}$$

Similar to (35), for $t \geq 2$,

$$\begin{aligned}
\tilde{R}_{P_n}(\boldsymbol{w}_{t+1}) - \tilde{R}_{P_n}(\boldsymbol{w}^*) &\leq (1 - 2\mu_{\boldsymbol{w}}\eta_{\boldsymbol{w}_t})\left(\tilde{R}_{P_n}(\boldsymbol{w}_t) - \tilde{R}_{P_n}(\boldsymbol{w}^*)\right) + \eta^2_{\boldsymbol{w}_t} G^2 L + 2\eta_{\boldsymbol{w}_t}\xi_t \\
&\leq \frac{G^2 L}{t\mu^2_{\boldsymbol{w}}} + \frac{2}{\mu_{\boldsymbol{w}}} \sum_{j=2}^{t} \frac{\xi_j}{j} \prod_{k=j+1}^{t}\left(1 - \frac{2}{k}\right) \\
&= \frac{G^2 L}{t\mu^2_{\boldsymbol{w}}} + \frac{2}{\mu_{\boldsymbol{w}}} \sum_{j=2}^{t} \frac{1}{j} \frac{(j-1)j}{(t-1)t} \xi_j \\
&= \frac{G^2 L}{t\mu^2_{\boldsymbol{w}}} + \frac{2}{\mu_{\boldsymbol{w}}} \sum_{j=2}^{t} \frac{(j-1)}{(t-1)t} \xi_j.
\end{aligned} \tag{40}$$

Since the second term in the last inequality is upper bonded by $\sum_{j=2}^{t} \xi_j$ which is a sum of martingale difference, and $|\xi_j| \leq 2G^2$, a simple Azuma-Hoeffding's inequality based on bounded martingale difference (Corollary 2.20 in (Wainwright, 2019)) can give a $\mathcal{O}(1/\sqrt{t})$ convergence rate in the high probability. However, we can sharpen the convergence rate via a Bennett's inequality (Proposition 3.19 in (Duchi, 2016)), because the conditional variance of ξ_j will decrease across training. We consider the conditional variance of $\sum_{j=2}^{t}(j-1)\xi_j$, let $\mathcal{F}_j$ be the σ-field generated by $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_j\}$, since $\mathbb{E}[\xi_j] = 0$

we have

$$\begin{aligned}\mathsf{Var}\left(\sum_{j=2}^{t}(j-1)\xi_j \mid \mathcal{F}_{j-1}\right) &= \sum_{j=2}^{t}(j-1)^2\mathsf{Var}\left(\xi_j \mid \mathcal{F}_{j-1}\right)\\ &= \sum_{j=2}^{t}(j-1)^2\mathbb{E}\left[\xi_j^2 \mid \mathcal{F}_{j-1}\right]\\ &\le 4G^2\sum_{j=2}^{t}(j-1)^2\|\nabla\tilde{R}_{P_n}(\boldsymbol{w}_j)\|^2\\ &\le 8G^2L\sum_{j=2}^{t}(j-1)^2\left(\tilde{R}_{P_n}(\boldsymbol{w}_j)-\tilde{R}_{P_n}(\boldsymbol{w}^*)\right),\end{aligned} \tag{41}$$

where first inequality is from Schwarz's inequality and the last inequality is because

$$\begin{aligned}\tilde{R}_{P_n}(\boldsymbol{w}^*)-\tilde{R}_{P_n}(\boldsymbol{w}) &\le \tilde{R}_{P_n}\left(\boldsymbol{w}-\frac{1}{L}\nabla\tilde{R}_{P_n}(\boldsymbol{w})\right)-\tilde{R}_{P_n}(\boldsymbol{w})\\ &\le -\left\langle\nabla\tilde{R}_{P_n}(\boldsymbol{w}),\frac{1}{L}\nabla\tilde{R}_{P_n}(\boldsymbol{w})\right\rangle+\frac{L}{2}\left\|\frac{1}{L}\nabla\tilde{R}_{P_n}(\boldsymbol{w})\right\|^2\\ &= -\frac{1}{2L}\left\|\nabla\tilde{R}_{P_n}(\boldsymbol{w})\right\|^2,\end{aligned} \tag{42}$$

for any $\boldsymbol{w}$. By applying Lemma 4, as long as $T \ge 4$ and $0 < \theta < 1/e$, then with probability at least $1-\theta$, for all $t \le T$,

$$\begin{aligned}&\tilde{R}_{P_n}(\boldsymbol{w}_{t+1})-\tilde{R}_{P_n}(\boldsymbol{w}^*)\\ &\le \frac{8G}{\mu_{\boldsymbol{w}}(t-1)t}\max\left\{\sqrt{2L\sum_{j=2}^{t}(j-1)^2\left(\tilde{R}_{P_n}(\boldsymbol{w}_j)-\tilde{R}_{P_n}(\boldsymbol{w}^*)\right)},G(t-1)\sqrt{\log\left(\frac{\log T}{\theta}\right)}\right\}\sqrt{\log\left(\frac{\log T}{\theta}\right)}+\frac{G^2L}{t\mu_{\boldsymbol{w}}^2}\\ &\le \frac{8G\sqrt{\log\left(\log\left(T/\theta\right)\right)}}{\mu_{\boldsymbol{w}}(t-1)t}\sqrt{2L\sum_{j=2}^{t}(j-1)^2\left(\tilde{R}_{P_n}(\boldsymbol{w}_j)-\tilde{R}_{P_n}(\boldsymbol{w}^*)\right)}+\frac{(8\mu_{\boldsymbol{w}}G^2\log\left(\log\left(T/\theta\right)\right)+G^2L)}{t\mu_{\boldsymbol{w}}^2}.\end{aligned} \tag{43}$$

Then, an upper bound to the first term in the last inequality can give our conclusion. Note that if $\tilde{R}_{P_n}(\boldsymbol{w}_j)-\tilde{R}_{P_n}(\boldsymbol{w}^*)$ is smaller than $\mathcal{O}(1/j-1)$, the conclusion is full-filled. To see this, we should find a large constant a such that $\tilde{R}_{P_n}(\boldsymbol{w}_{t+1})-\tilde{R}_{P_n}(\boldsymbol{w}^*) \le a/t$. This is clearly hold when $a \ge G^2/2\mu_{\boldsymbol{w}}$ for $t=1$ due to the PL inequality and bounded gradient. For $t \ge 2$, we find this a by induction. Let $b = 8G\sqrt{2L\log\left(\log\left(T/\theta\right)\right)}/\mu_{\boldsymbol{w}}$ and $c = (8\mu_{\boldsymbol{w}}G^2\log\left(\log\left(T/\theta\right)\right)+G^2L)/\mu_{\boldsymbol{w}}^2$. A satisfactory a yields

$$\frac{a}{t} \ge \frac{b}{(t-1)t}\sqrt{a\sum_{j=2}^{t}(j-1)}+\frac{c}{t} = \frac{b}{(t-1)t}\sqrt{\frac{at(t-1)}{2}}+\frac{c}{t} \ge \frac{1}{t}\left(b\sqrt{\frac{a}{2}}+c\right). \tag{44}$$

By solving a quadratic inequality, we conclude that $a - b\sqrt{a/2}-c \ge 0$. Then

$$a \ge \left(\frac{b+\sqrt{b^2+8c}}{2\sqrt{2}}\right)^2. \tag{45}$$

By taking

$$a \ge 2\left(\frac{2b^2+8c}{8}\right) \ge \left(\frac{b+\sqrt{b+8c}}{2\sqrt{2}}\right)^2, \tag{46}$$

we get

$$a \ge \frac{64G^2L\log\left(\log\left(T/\theta\right)\right)}{\mu_{\boldsymbol{w}}^2}+\frac{(16\mu_{\boldsymbol{w}}G^2\log\left(\log\left(T/\theta\right)\right)+G^2L)}{\mu_{\boldsymbol{w}}^2} = \frac{G^2\log\left(\log\left(T/\theta\right)\right)(64L+16\mu_{\boldsymbol{w}})+G^2L}{\mu_{\boldsymbol{w}}^2}, \tag{47}$$

due to the value of b and c. Hence, we get the conclusion by taking $t=T$. □

A.2.2. PROOF OF PROPOSITION 1

Proof. From the definition of $\tilde{R}_{P_n}(\boldsymbol{w})$, for any $r \geq 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\|\boldsymbol{\delta}\|_p \leq r}\left(f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta}) - f(\boldsymbol{w}, \boldsymbol{x}_i)\right) \leq \tilde{R}_{P_n}(\boldsymbol{w}) \leq \epsilon. \tag{48}$$

On the other hand

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\|\boldsymbol{\delta}\|_p \leq r}\left(f(\boldsymbol{w}, \boldsymbol{x}_i) - f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta})\right) \leq R_{P_n}(\boldsymbol{w}) \leq \tilde{R}_{P_n}(\boldsymbol{w}) \leq \epsilon. \tag{49}$$

Take a sum to the two above inequalities, we get

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\|\boldsymbol{\delta}\|_p \leq r}\left|f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta}) - f(\boldsymbol{w}, \boldsymbol{x}_i))\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left(\sup_{\|\boldsymbol{\delta}\|_p \leq r} f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta}) - \inf_{\|\boldsymbol{\delta}\|_p \leq r} f(\boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{\delta}))\right) \leq 2\epsilon. \tag{50}$$

Then the conclusion is verified. □

B. Proofs for Section 4

B.1. Proof of Theorem 4

Proof. We have $r(\infty) = r$ in this theorem. The key is to bound the $|\sup_{P \in B_{\mathsf{W}_\infty}(P_0, r)} R_P(\boldsymbol{w}_{\text{pre}}) - \sup_{Q \in B_{\mathsf{W}_\infty}(Q_0, r)} R_Q(\boldsymbol{w}_{\text{pre}})|$, then triangle inequality and Hoeffding's inequality imply the conclusion. Let $P_r^* \in \arg\max_{\{P \in B_{\mathsf{W}_\infty}(P_0, r)\}} R_P(\boldsymbol{w}_{\text{pre}})$. For any given $\boldsymbol{x}$, due to the continuity of $f(\boldsymbol{w}_{\text{pre}}, \cdot)$, similar to Lemma 1, we can find the $T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x}) = \boldsymbol{x} + \arg\max_{\{\boldsymbol{\delta}: \|\boldsymbol{\delta}\|_\infty \leq r\}} f(\boldsymbol{w}_{\text{pre}}, \boldsymbol{x} + \boldsymbol{\delta})$. Then due to Lemma 1,

$$R_{P_r^*}(\boldsymbol{w}_{\text{pre}}) = \mathbb{E}_{P_0}\left[\sup_{\|\boldsymbol{\delta}\|_\infty \leq r} f(\boldsymbol{w}_{\text{pre}}, \boldsymbol{x} + \boldsymbol{\delta})\right]. \tag{51}$$

Thus, $T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x}) \sim P_r^*$ when $\boldsymbol{x} \sim P_0$. We can find $\boldsymbol{z} \sim Q_0$ due to the Kolmogorov's Theorem, and let $T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{z}) \sim Q_r^*$. By the definition of W_∞-distance, one can verify $\mathsf{W}_\infty(Q_0, Q_r^*) \leq r$ as well as $R_{Q_r^*}(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}}$. Note that $0 \leq f(\boldsymbol{w}_{\text{pre}}, \cdot) \leq M$, then

$$\begin{aligned}
\left|R_{P_r^*}(\boldsymbol{w}_{\text{pre}}) - R_{Q_r^*}(\boldsymbol{w}_{\text{pre}})\right| &= \left|\int_{\mathcal{X}} f(\boldsymbol{w}_{\text{pre}}, \boldsymbol{x}) dP_r^*(\boldsymbol{x}) - \int_{\mathcal{X}} f(\boldsymbol{w}_{\text{pre}}, \boldsymbol{x}) dQ_r^*(\boldsymbol{x})\right| \\
&= \left|\int_{\mathcal{X}} f(\boldsymbol{w}_{\text{pre}}, T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x})) dP_0(\boldsymbol{x}) - \int_{\mathcal{X}} f(\boldsymbol{w}_{\text{pre}}, T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x})) dQ_0(\boldsymbol{x})\right| \\
&\leq \int_{\mathcal{X}} \left|f(\boldsymbol{w}_{\text{pre}}, T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x}))\right| |dP_0(\boldsymbol{x}) - dQ_0(\boldsymbol{x})| \\
&\leq M \int_{\mathcal{X}} |dP_0(\boldsymbol{x}) - dQ_0(\boldsymbol{x})| \\
&= 2M\mathsf{TV}(P_0, Q_0).
\end{aligned} \tag{52}$$

The last equality is from the definition of total variation distance (Villani, 2008). Thus a simple triangle inequality implies that

$$R_{P_r^*}(\boldsymbol{w}_{\text{pre}}) \leq \left|R_{P_r^*}(\boldsymbol{w}_{\text{pre}}) - R_{Q_r^*}(\boldsymbol{w}_{\text{pre}})\right| + R_{Q_r^*}(\boldsymbol{w}_{\text{pre}}) \leq \epsilon_{\text{pre}} + 2M\mathsf{TV}(P_0, Q_0). \tag{53}$$

Next we give the concentration result of $\tilde{R}_{P_n}(\boldsymbol{w}_{\text{pre}})$. Due to the definition of $\tilde{R}_{P_n}(\boldsymbol{w}_{\text{pre}})$, it can be rewritten as $R_{P_n^*}(\boldsymbol{w}_{\text{pre}})$ where P_n^* is the empirical distribution on $\{T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x}_i)\}$. Since $0 \leq f(\boldsymbol{w}_{\text{pre}}, \cdot) \leq M$ and $\{T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x}_i)\}$ are i.i.d draws from P_r^*. Azuma-Hoeffding's inequality (Corollary 2.20 in (Wainwright, 2019)) shows that with probability at least $1 - \theta$,

$$\tilde{R}_{P_n}(\boldsymbol{w}_{\text{pre}}) - R_{P_r^*}(\boldsymbol{w}_{\text{pre}}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}_{\text{pre}}, T(\boldsymbol{x}_i)) - R_{P_r^*}(\boldsymbol{w}_{\text{pre}}) \leq M\sqrt{\frac{\log(1/\theta)}{2n}}. \tag{54}$$

Hence we get our conclusion. □

B.2. Proof of Theorem 5

With a little abuse of notation, let $r(2) = r/\epsilon_{\text{pre}}$ denoted by r in the proof, and $P_r^* \in \arg\max_{P \in B_{\mathsf{W}_2}(P_0, r)} R_P(\boldsymbol{w})$. By Lemma 2, there exists $T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{x}) \sim P_r^*$ with $\boldsymbol{x} \sim P_0$. Then we can find $\boldsymbol{z} \sim Q_0$ due to Kolmogorov's Theorem. Let $T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{z}) \sim Q_r^*$, we see

$$\begin{aligned}
\mathsf{W}_2(Q_0, Q_r^*)^2 &\le \int_{\mathcal{X}} \|\boldsymbol{z} - T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{z})\|^2 dQ_0(\boldsymbol{z}) \\
&\le \int_{\mathcal{X}} \|\boldsymbol{z} - T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{z})\|^2 \, |dQ_0(\boldsymbol{z}) - dP_0(\boldsymbol{z})| + \int_{\mathcal{X}} \|\boldsymbol{z} - T_r^{\boldsymbol{w}_{\text{pre}}}(\boldsymbol{z})\|^2 dP_0(\boldsymbol{z}) \\
&\le D^2 \int_{\mathcal{X}} |dQ_0(\boldsymbol{z}) - dP_0(\boldsymbol{z})| + r^2 \\
&= 2D^2 \mathsf{TV}(P_0, Q_0) + r^2.
\end{aligned} \tag{55}$$

Thus $R_{Q_r^*}(\boldsymbol{w}_{\text{pre}}) \le \epsilon_{\text{pre}}$. Similar to (52) and (53) we get the conclusion.

C. Hyperparameters

Table 4: Hyperparameters of adversarial training on `CIFAR10`.

Hyperparam	Std	Adv-ℓ_2	Adv-ℓ_∞
Learning Rate	0.1	0.1	0.1
Momentum	0.9	0.9	0.9
Batch Size	128	128	128
Weight Decay	5e-4	5e-4	5e-4
Epochs	200	200	200
Inner Loop Steps	-	8	8
Perturbation Size	-	2/12	2/255
Perturbation Step Size	-	1/24	1/510

Table 5: Hyperparameters of adversarial training on `ImageNet`.

Hyperparam	Std	Adv-ℓ_2	Adv-ℓ_∞
Learning Rate	0.1	0.1	0.1
Momentum	0.9	0.9	0.9
Batch Size	512	512	512
Weight Decay	5e-4	5e-4	5e-4
Epochs	100	100	100
Inner Loop Steps	-	3	3
Perturbation Size	-	0.25	2/255
Perturbation Step Size	-	0.05	1/510

Table 6: Hyperparameters of adversarial training on BERT base model.

Hyperparam	Std	Adv-ℓ_2	Adv-ℓ_∞
Learning Rate	3e-5	3e-5	3e-5
Batch Size	32	32	32
Weight Decay	0	0	0
Hidden Layer Dropout Rate	0.1	0.1	0.1
Attention Probability Dropout Rate	0.1	0.1	0.1
Max Epochs	10	10	10
Learning Rate Decay	Linear	Linear	Linear
Warmup Ratio	0	0	0
Inner Loop Steps	-	3	3
Perturbation Size	-	1.0	0.001
Perturbation Step Size	-	0.1	0.0005

D. Ablation Study

D.1. Effect of Perturbation Size

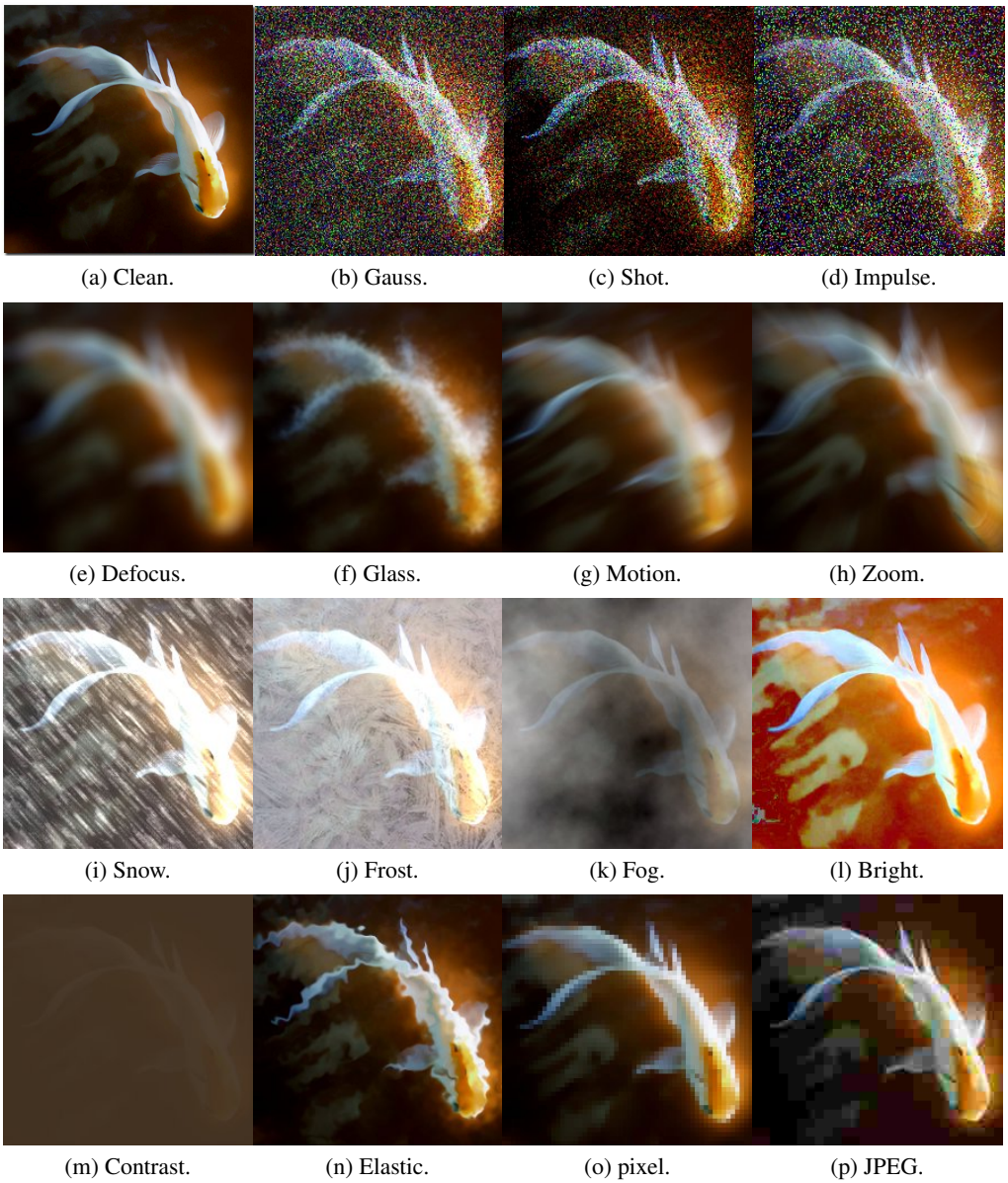
We study the effect of perturbation size r in adversarial training in bounds (5) and (6). We vary the perturbation size r in $\{2^{-5}/12, 2^{-4}/12, 2^{-3}/12, 2^{-2}/12, 2^{-1}/12, 2^{0}/12, 2^{1}/12, 2^{2}/12, 2^{3}/12, 2^{4}/12, 2^{5}/12, 2^{6}/12, 2^{7}/12\}$ for Adv-ℓ_2 and in $\{2^{-4}/255, 2^{-3}/255, 2^{-2}/255, 2^{-1}/255, 2^{0}/255, 2^{1}/255, 2^{2}/255, 2^{3}/255, 2^{4}/255\}$ for Adv-ℓ_∞. The perturbation step size $\eta_{\boldsymbol{x}}$ in Algorithm 1 is set to be $r/4$ (Salman et al., 2020a). Experiments are conducted on `CIFAR10` and the settings follow those in Section 5.1.1.

The results are shown in Figures 3 and 4. In the studied ranges, the accuracy on the OOD data from all categories exhibits similar trend, i.e., first increases and then decreases, as r increases. This is consistent with our discussion in Section 5.1.1 that there is an optimal perturbation size r for improving OOD generalization via adversarial training. For data corrupted

under types Fog, Bright and Contrast, adversarial training degenerates the performance in Table 1. We speculate this is because the three corruption types rescale the input pixel values to smaller values and the same perturbation size r leads to relatively large perturbation. Thus according to the discussion in Section 5.1.1 that there is an optimal r for improving OOD generalization, we suggest conducting adversarial training with a smaller perturbation size to defend these three types of corruption. Figures 3 and 4 also show that smaller optimal perturbation sizes have better performances for these three types of corruption.

D.2. Effect of the the Number of Training Samples

We study the effect of the number of training samples, as bounds (5) and (6) suggest that more training samples lead to better OOD generalization. We split `CIFAR10` into 5 subsets, each of which has 10000, 20000, 30000, 40000 and 50000 training samples. The other settings follow those in Section 5.1.1. The results are in shown Figures 5 and 6.



(a) Clean. (b) Gauss. (c) Shot. (d) Impulse.

(e) Defocus. (f) Glass. (g) Motion. (h) Zoom.

(i) Snow. (j) Frost. (k) Fog. (l) Bright.

(m) Contrast. (n) Elastic. (o) pixel. (p) JPEG.

Figure 2: 15 types of artificially constructed corruptions from four categories: Noise, Blur, Weather, and Digital from the `ImageNet-C` dataset (Hendrycks & Dietterich, 2018). Each corruption has five levels of severity with figures under severity 5 are shown here.

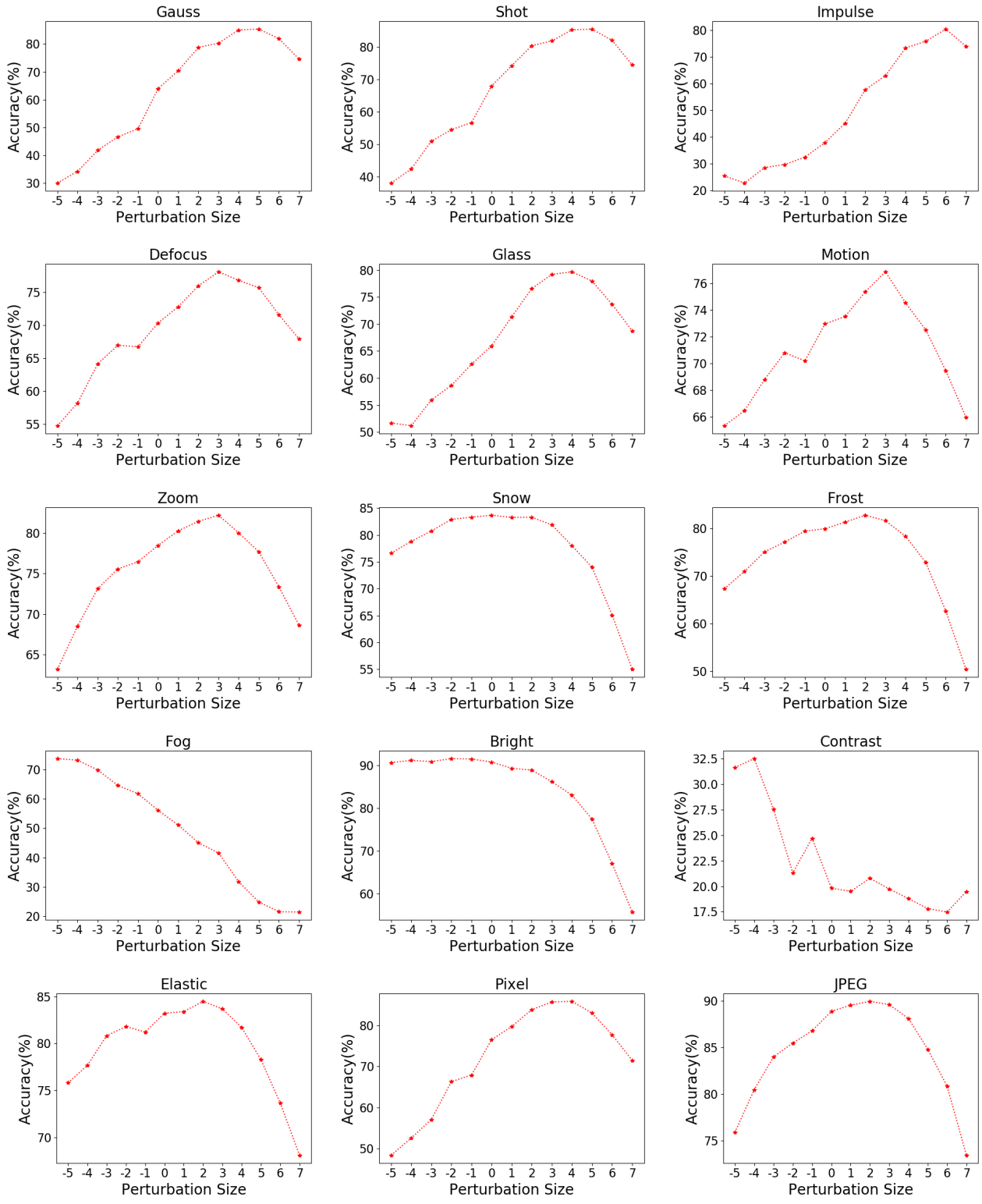


Figure 3: Accuracy of Adv-ℓ_2 on `CIFAR10-C` over various perturbation sizes. The x-axis means the perturbation size is $2^x/12$.

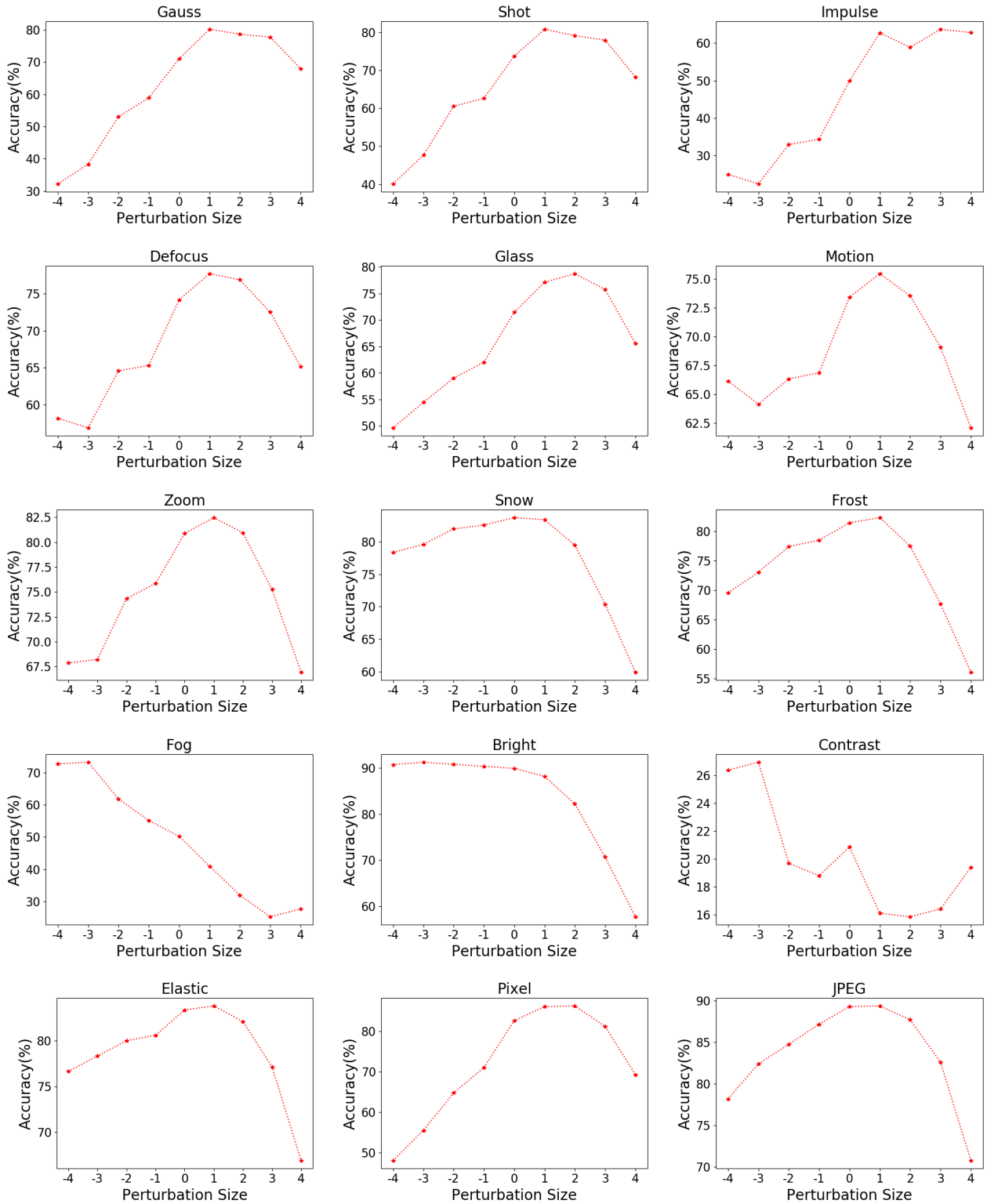


Figure 4: Accuracy of Adv-ℓ_∞ on `CIFAR10-C` over various perturbation sizes. The x-axis means the perturbation size is $2^x/255$.

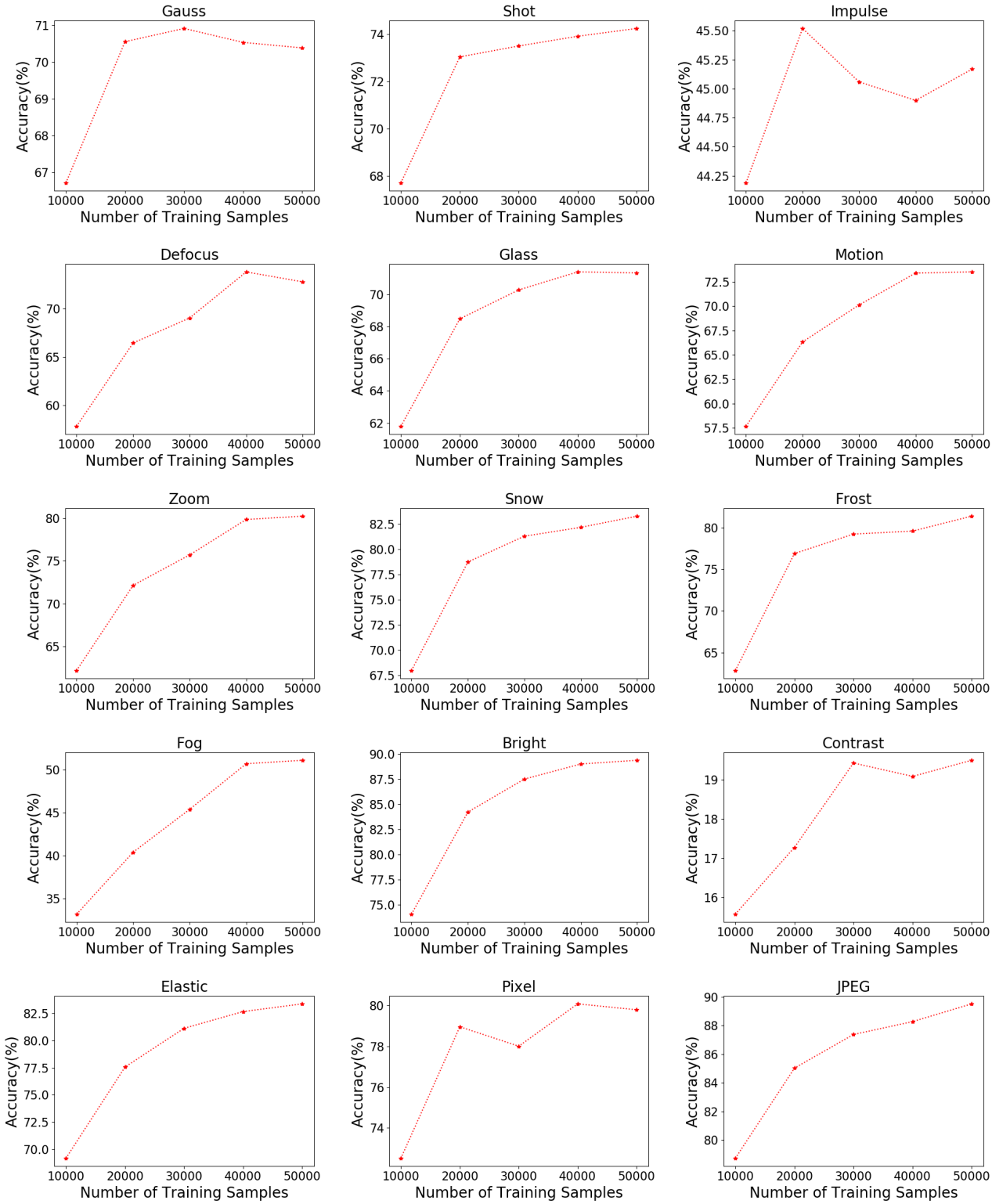


Figure 5: Accuracy of Adv-ℓ_2 on `CIFAR10-C` over various numbers of training samples.

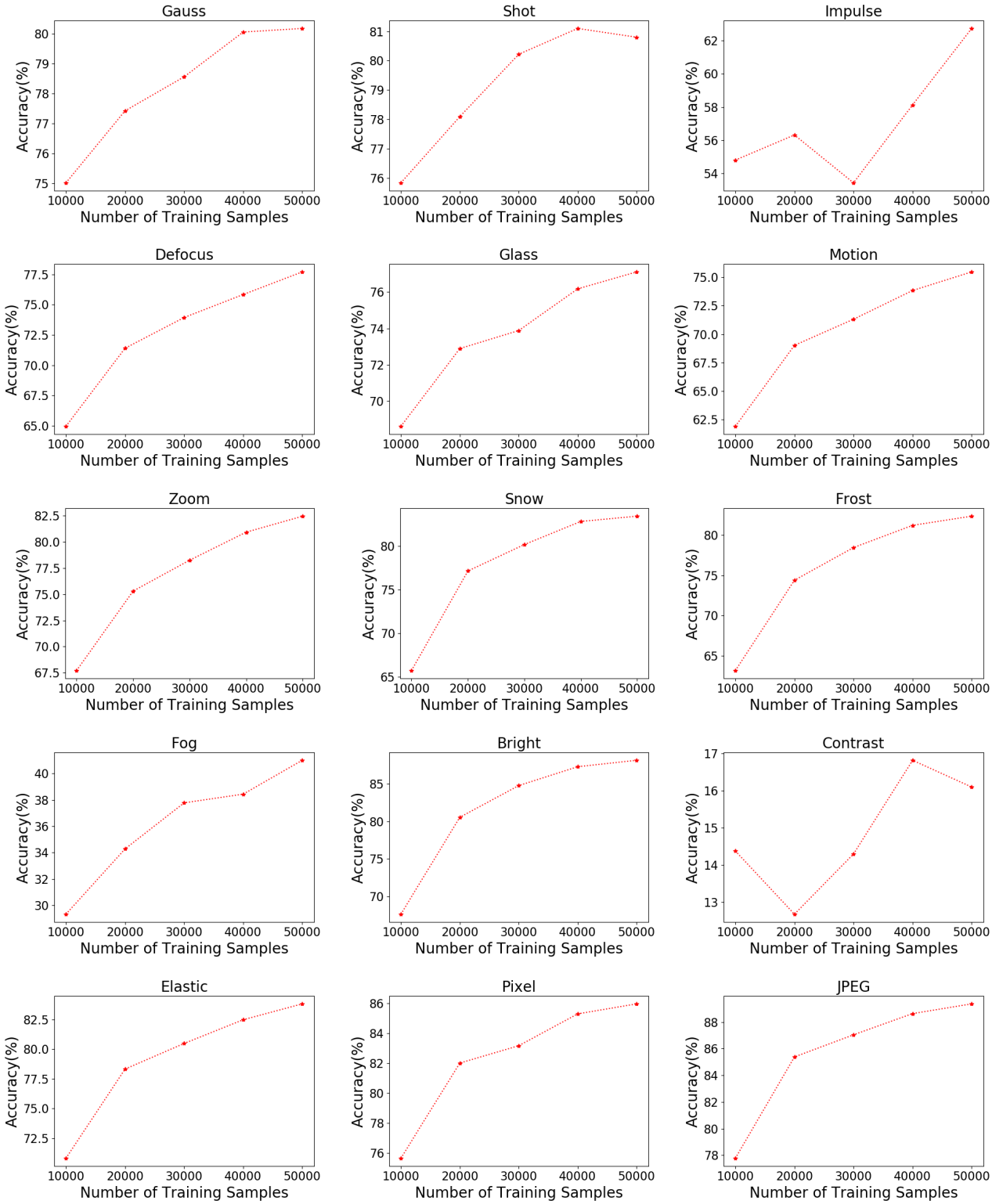


Figure 6: Accuracy of Adv-ℓ_∞ on `CIFAR10-C` over various numbers of training samples.