Supplementary Material for "Robust Generalization against Photon-Limited Corruptions via Worst-Case Sharpness Minimization"

Zhuo Huang[1,†], Miaoxi Zhu[2,†], Xiaobo Xia[1], Li Shen[3], Jun Yu[4,✉],
Chen Gong[5], Bo Han[6], Bo Du[2], Tongliang Liu[1]

[1]Sydney AI Centre, The University of Sydney; [2]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University; [3]JD Explore Academy; [4]Department of Automation, University of Science and Technology of China;
[5]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology; [6]Department of Computer Science, Hong Kong Baptist University.

In the supplementary material, we first provide the details in the proof for the theoretical result in the main paper in Section. A. Then, we give details about our implementation details in Section B. Finally, we show more experimental results using different types of corruptions in Section C.

A. Convergence Analyses

A.1. Preliminaries

We first give some notations before we start our proof for the convergence.

1. We denote the expectation value for the loss function as $\mathbb{L}(\theta,\omega) := \mathbb{E}_{(x,y)\sim Q}\mathcal{L}(\theta,\omega;(x,y))$, and so as the SAM function that $\mathbb{R}(\theta,\omega) = \mathbb{E}_{(x,y)\sim Q}R(\theta,\omega;(x,y))$. So our objective can be turned into: $\min_\theta\{\max_\omega \mathbb{L}(\theta,\omega)\} + \mathbb{R}(\theta,\omega)$. And recalling our SharpDRO algorithm, we restate the meaning of the parameters: the model is parameterized by θ and ω means the weighted sampling.
2. κ is the condition number that $\kappa = \frac{l}{\mu}$, where l is the Lipschitz-smoothness in Assumption A.2 and μ means the PL condition in Assumption A.3.
3. We define $\mathbb{L}^*(\theta) = \max_\omega \mathbb{L}(\theta,\omega)$ and $\omega^*(\theta) = arg\max_\omega \mathbb{L}(\theta,\omega)$.

A.2. Update Rule

Before our theoretical analyses, we need to make the update rule for each variable explicit. We have to pay attention to the fact that our algorithm is stochastic that we can not directly get the real value of the gradient $\nabla\mathbb{L}(\theta,\omega)$, rather we estimate it by batches of samples $g_\theta(\theta,\omega) = \frac{1}{M}\sum_{i=1}^M \frac{\partial\mathcal{L}}{\partial\theta}(\theta,\omega;(x_i,y_i))$ and $g_\omega(\theta,\omega) = \frac{1}{M}\sum_{i=1}^M \frac{\partial\mathcal{L}}{\partial\omega}(\theta,\omega;(x_i,y_i))$, who hold some properties we will introduce in Assumption A.1. So the optimization iteration is executed as follows in reality:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta_\theta g_\theta(\theta_t + \rho g_\theta(\theta_t,\omega_t),\omega_t);\\ \omega_{t+1} &= \omega_t + \eta_\omega \nabla_\omega g_\omega(\theta_t,\omega_t).\end{aligned} \tag{1}$$

We further give a notation for brief that $\theta_{t+1/2} \triangleq \theta_t + \rho g_\theta(\theta_t,\omega_t)$, so the update for θ can be simplified as: $\theta_{t+1} = \theta_t - \eta_\theta g_\theta(\theta_{t+1/2},\omega_t)$.

A.3. Assumptions

We also have to make some necessary assumptions on our problem setting for this convergence proof:

Assumption A.1 (Bounded variance)**.** *The unbiased estimation about the gradient of the loss function also has bounded variance that:*

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim Q}[\frac{\partial\mathcal{L}}{\partial\theta}(\theta,\omega;(x,y))] = \nabla_\theta\mathbb{L}(\theta,\omega), \quad \mathbb{E}_{(x,y)\sim Q}\|\frac{\partial\mathcal{L}}{\partial\theta}(\theta,\omega;(x,y)) - \nabla_\theta\mathbb{L}(\theta,\omega)\|^2 \le \sigma^2;\\ &\mathbb{E}_{(x,y)\sim Q}[\frac{\partial\mathcal{L}}{\partial\omega}(\theta,\omega;(x,y))] = \nabla_\omega\mathbb{L}(\theta,\omega), \quad \mathbb{E}_{(x,y)\sim Q}\|\frac{\partial\mathcal{L}}{\partial\omega}(\theta,\omega;(x,y)) - \nabla_\omega\mathbb{L}(\theta,\omega)\|^2 \le \sigma^2.\end{aligned}$$

Remark. *Since g_θ and g_ω are the averaged samples that: $g_\theta = \frac{1}{M}\sum_{i=1}^{M}\frac{\partial \mathcal{L}}{\partial \theta}(\theta,\omega;(x_i,y_i))$ and $g_\omega = \frac{1}{M}\sum_{i=1}^{M}\frac{\partial \mathcal{L}}{\partial \omega}(\theta,\omega;(x_i,y_i))$ respectively, they also have the unbiased property and have bounded variance:*

$$\mathbb{E}_{(x,y)\sim Q}[g_\theta(\theta,\omega;(x,y))] = \nabla_\theta \mathbb{L}(\theta,\omega), \quad \mathbb{E}_{(x,y)\sim Q}\|g_\theta(\theta,\omega;(x,y)) - \nabla_\theta \mathbb{L}(\theta,\omega)\|^2 \le \frac{\sigma^2}{M};$$
$$\mathbb{E}_{(x,y)\sim Q}[g_\omega(\theta,\omega;(x,y))] = \nabla_\omega \mathbb{L}(\theta,\omega), \quad \mathbb{E}_{(x,y)\sim Q}\|g_\omega(\theta,\omega;(x,y)) - \nabla_\omega \mathbb{L}(\theta,\omega)\|^2 \le \frac{\sigma^2}{M}.$$

Assumption A.2 (Lipschitz smooth)**.** *$\mathcal{L}(\theta,\omega;(x,y))$ is differential and l-Lipschitz smooth for every given sample (x,y):*

$$\|\nabla_\theta \mathcal{L}(\theta_1,\omega;(x,y)) - \nabla_\theta \mathcal{L}(\theta_2,\omega;(x,y))\| \le l\|\theta_1-\theta_2\|, \quad \forall \omega,(x,y);$$
$$\|\nabla_\omega \mathcal{L}(\theta,\omega_1;(x,y)) - \nabla_\omega \mathcal{L}(\theta,\omega_2;(x,y))\| \le l\|\omega_1-\omega_2\|, \quad \forall \theta,(x,y).$$

Remark. *So the expectation function $\mathbb{L}$ also have the Lipschitz smooth property that:*

$$\|\nabla_\theta \mathbb{L}(\theta_1,\omega) - \nabla_\theta \mathbb{L}(\theta_2,\omega)\| \le \mathbb{E}\|\nabla_\theta \mathcal{L}(\theta_1,\omega;(x,y)) - \nabla_\theta \mathcal{L}(\theta_2,\omega;(x,y))\| \le l\|\theta_1-\theta_2\|, \quad \forall \omega;$$
$$\|\nabla_\omega \mathbb{L}(\theta,\omega_1) - \nabla_\omega \mathbb{L}(\theta,\omega_2)\| \le \mathbb{E}\|\nabla_\omega \mathcal{L}(\theta,\omega_1;(x,y)) - \nabla_\omega \mathcal{L}(\theta,\omega_2;(x,y))\| \le l\|\omega_1-\omega_2\|, \quad \forall \theta.$$

Assumption A.3 (PL condition)**.** *The loss function $\mathbb{L}(\theta,\cdot)$ satisfies PL condition on every given θ, i.e., there exists $\mu > 0$ such that $\|\nabla_\omega \mathbb{L}(\theta,\omega)\|^2 \ge 2\mu[\max_\omega \mathbb{L}(\theta,\omega) - \mathbb{L}(\theta,\omega)], \forall \theta,\omega$.*

A.4. Useful Lemmas

In this part, we will prove some necessary lemmas for us to prove the convergence bound. And we will give the definition of the stationary point of our problem at the beginning.

Definition A.1 (Stationary measure)**.** *θ is defined as the ϵ-stationary point of our problem if $\mathbb{E}\|\nabla \mathbb{L}^*(\theta)\| \le \epsilon$ for any $\epsilon \ge 0$.*

Remark. *For minmax problem, there are usually two ways to measure the stationary point. The other one is measured two-side that: when $\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta,\omega)\| \le \epsilon$ and $\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta,\omega)\| \le \epsilon$, we claim (θ,ω) is the (ϵ,ϵ)-stationary point. It has been proved in [4] that these two measures can be translated into each other when $\mathbb{L}^*$ is smooth which will be shown in Lemma A.1. But what we compute is the model parameter θ using the algorithm SharpDRO. So we choose the measure by $\mathbb{E}\|\mathbb{L}^*(\theta)\|$ here.*

Lemma A.1. *[3] Under Assumption A.2 and A.3, $\mathbb{L}^*(\theta)$ is $(l + \frac{l^2}{2\mu})$-Lipschitz smooth with the gradient:*

$$\nabla_\theta \mathbb{L}^*(\theta,\omega) = \nabla_\theta \mathbb{L}(\theta,\omega^*(\theta)).$$

Lemma A.2. *[3] Under Assumption A.2 and A.3, $\omega^*(\cdot)$ is smooth about its variable:*

$$\|\omega^*(\theta_1) - \omega^*(\theta_2)\| \le \frac{l}{2\mu}\|\theta_1-\theta_2\|, \quad \forall \theta_1,\theta_2.$$

Lemma A.3. *We give an estimation that $\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 \le (4\rho^2 l^2 + 2\rho l + 2)\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + (5\rho^2 l^2 + 2)\frac{\sigma^2}{M}$.*

Proof.

$$\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 = -\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t) - \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + 2\mathbb{E}\langle g_\theta(\theta_{t+1/2},\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle. \quad (2)$$

For the cross-product term, we divide it as follows:

$$\begin{aligned}
&\mathbb{E}\langle g_\theta(\theta_{t+1/2},\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
&= \mathbb{E}\langle g_\theta(\theta_{t+1/2},\omega_t) - g_\theta(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle + \mathbb{E}\langle g_\theta(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
&= \mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_{t+1/2},\omega_t) - \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle + \mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
&\overset{(i)}{\le} \frac{1}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2},\omega_t) - \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t)\|^2 + \frac{1}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad + \mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t) - \nabla_\theta \mathbb{L}(\theta_t,\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
&\overset{(ii)}{\le} \frac{\rho^2 l^2}{2}\mathbb{E}\|g_\theta(\theta_t,\omega_t) - \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \frac{3}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \rho l\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\overset{(iii)}{\le} (\rho l + \frac{3}{2})\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \frac{\rho^2 l^2 \sigma^2}{2M},
\end{aligned} \quad (3)$$

where the inequality (i) is due to the Cauchy-Schwarz inequality; the inequality (ii) is because of the Lipschitz-smoothness of $\mathbb{L}$ that $\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t + \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \omega_t)\|^2 \leq l^2 \mathbb{E}\|\theta_{t+1/2} - \theta_t - \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2$ and the property of Lipschitz-smoothness that $\langle \nabla_\theta \mathbb{L}(\theta_t + \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\rangle = \frac{1}{\rho}\langle \nabla_\theta \mathbb{L}(\theta_t + \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\rangle \leq \frac{l}{\rho}\|\rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2$; and the inequality (iii) makes use of the Assumption A.1.

As for the second term, we have:

$$
\begin{aligned}
&\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\leq 2\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t)\|^2 + 2\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\leq \frac{2\sigma^2}{M} + 2l^2 \mathbb{E}\|\theta_{t+1/2} - \theta_t\|^2 \\
&= \frac{2\sigma^2}{M} + 2\rho^2 l^2 \mathbb{E}\|g_\theta(\theta_t, \omega_t)\|^2 \\
&\leq 2\frac{\sigma^2}{M}(2\rho^2 l^2 + 1) + 4\rho^2 l^2 \mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2,
\end{aligned}
\tag{4}
$$

where the last inequality comes from the fact that: $\mathbb{E}\|g_\theta(\theta_t, \omega_t)\|^2 \leq 2\mathbb{E}\|g_\theta(\theta_t, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 + 2\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2$

By combining the above inequalities, we can get:

$$
\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2 \leq (4\rho^2 l^2 + 2\rho l + 2)\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 + (5\rho^2 l^2 + 2)\frac{\sigma^2}{M}. \tag{5}
$$

□

Lemma A.4. *For the descending relationship of the function $\mathbb{L}^*$, we have:*

$$
\begin{aligned}
\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \leq \mathbb{E}[\mathbb{L}^*(\theta_t)] &- \frac{\eta_\theta}{2}(1 - 5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \\
&+ [\frac{\eta_\theta}{2}(1 + \frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)]\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 + (5\rho^2 l^2 + 2)\frac{L\eta_\theta^2 \sigma^2}{2M},
\end{aligned}
$$

where we use the brief notation that $L = l + \frac{l\kappa}{2}$.

Proof. Since $\mathbb{L}^*(\theta)$ is $(l + \frac{l\kappa}{2})$-Lipschitz smooth according to Lemma A.1, we have:

$$
\begin{aligned}
\mathbb{L}^*(\theta_{t+1}) &\leq \mathbb{L}^*(\theta_t) + \langle \nabla \mathbb{L}^*(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\|\theta_{t+1} - \theta_t\|^2 \\
&= \mathbb{L}^*(\theta_t) - \eta_\theta \langle \nabla \mathbb{L}^*(\theta_t), g_\theta(\theta_{t+1/2}, \omega_t)\rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\eta_\theta^2 \|g_\theta(\theta_{t+1/2}, \omega_t)\|^2.
\end{aligned}
\tag{6}
$$

Taking expectation conditioned on (θ_t, ω_t) and we get:

$$
\mathbb{E}[\mathbb{L}^*(\theta_{t+1})|\theta_t, \omega_t] \leq \mathbb{L}^*(\theta_t) - \eta_\theta \langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t)\rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\eta_\theta^2 \mathbb{E}[\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2|\theta_t, \omega_t]. \tag{7}
$$

We again take expectation on both side on above inequality so we have:

$$
\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \leq \mathbb{E}[\mathbb{L}^*(\theta_t)] - \eta_\theta \mathbb{E}\langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t)\rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\eta_\theta^2 \mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2. \tag{8}
$$

For the second term, we decompose it as follows:

$$
\begin{aligned}
&\mathbb{E}\langle\nabla\mathbb{L}^*(\theta_t), \nabla_\theta\mathbb{L}(\theta_{t+1/2}, \omega_t)\rangle \\
&= \mathbb{E}\langle\nabla\mathbb{L}^*(\theta_t), \nabla_\theta\mathbb{L}(\theta_t, \omega_t) + \nabla_\theta\mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\rangle \\
&\geq \mathbb{E}\langle\nabla\mathbb{L}^*(\theta_t), \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\rangle - \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\| \\
&\geq \mathbb{E}\langle\nabla\mathbb{L}^*(\theta_t), \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\rangle - \rho l\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|g_\theta(\theta_t, \omega_y)\| \\
&\geq \mathbb{E}\langle\nabla\mathbb{L}^*(\theta_t), \nabla\mathbb{L}^*(\theta_t) + \nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\rangle - \rho l\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|(\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t)\| + \|g_\theta(\theta_t, \omega_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|) \\
&\geq \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2 - \rho l\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t)\| \\
&\quad - \frac{1}{2}\rho l\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\rho l\mathbb{E}\|g_\theta(\theta_t, \omega_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\geq \frac{1-\rho l}{2}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2 - \rho l\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t)\| - \frac{\rho l\sigma^2}{2M}.
\end{aligned}
\tag{9}
$$

We continue estimating the last term in above inequality 9

$$
\begin{aligned}
&\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t)\| \\
&= \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t) + \nabla\mathbb{L}^*(\theta_t)\| \\
&\leq \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\| \\
&\overset{(i)}{\leq} \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \frac{1}{4}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2,
\end{aligned}
\tag{10}
$$

where the last inequality (i) is due to Young's inequality.

By combining inequality 8 with 10, we can get:

$$
\begin{aligned}
&\mathbb{E}\langle\nabla\mathbb{L}^*(\theta_t), \nabla_\theta\mathbb{L}(\theta_{t+1/2}, \omega_t)\rangle \\
&\geq \frac{1-\rho l}{2}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2 - 2\rho l\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{\rho l}{4}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{\rho l\sigma^2}{2M} \\
&= \frac{1}{2}(1-5\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}(1+\frac{1}{2}\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|^2 - \frac{\rho l\sigma^2}{2M}.
\end{aligned}
\tag{11}
$$

Finally, we combine inequality 8 with Lemma A.3 and inequality 11:

$$
\begin{aligned}
&\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \\
&\leq \mathbb{E}[\mathbb{L}^*(\theta_t)] - \frac{\eta_\theta}{2}(1-5\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\quad + \frac{1}{2}(l+\frac{l\kappa}{2})\eta_\theta^2((4\rho^2l^2 + 2\rho l + 2)\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|^2 + (5\rho^2l^2+2)\frac{\sigma^2}{M}) \\
&\overset{(i)}{\leq} \mathbb{E}[\mathbb{L}^*(\theta_t)] - \frac{\eta_\theta}{2}(1-5\rho l - \eta_\theta(2l+l\kappa)(4\rho^2l^2+2\rho l+2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \\
&\quad + [\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l) + \eta_\theta^2(l+\frac{l\kappa}{2})(4\rho^2l^2+2\rho l+2)]\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|^2 + \frac{1}{2}(l+\frac{l\kappa}{2})(5\rho^2l^2+2)\frac{\eta_\theta^2\sigma^2}{M},
\end{aligned}
\tag{12}
$$

where the last inequality (i) uses the Cauchy-Schwarz inequality that $\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|^2 \leq 2\|\nabla\mathbb{L}^*(\theta_t)\|^2 + 2\|\nabla_\theta\mathbb{L}(\theta_t, \omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2$. □

A.5. Theorem

Theorem 1. *Under Assumption A.1,A.2,A.3, and the learning rate satisfy that* $\eta_\theta \leq \min\{\frac{1}{128\kappa^2 l}, \sqrt{\frac{M(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])}{132T\kappa^4 l\sigma^2}}\}$, $\eta_\omega \leq 64\kappa^2\eta_\theta$ *and* $\rho \leq \frac{\eta_\theta}{2l}$, *we have the convergence bound for our problem:*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \leq 320\sqrt{\frac{3\kappa^4 l(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])\sigma^2}{11MT}} = \mathcal{O}(\frac{\kappa^2}{\sqrt{MT}}).
\tag{13}
$$

Proof. First recall the descending relationship of the function $\mathbb{L}^*$ in Lemma A.4:

$$\begin{aligned}
&\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \\
&\leq \mathbb{E}[\mathbb{L}^*(\theta_t)] - \frac{\eta_\theta}{2}(1 - 5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \\
&\quad + [\frac{\eta_\theta}{2}(1 + \frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)]\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 + (5\rho^2 l^2 + 2)\frac{L\eta_\theta^2 \sigma^2}{2M}.
\end{aligned} \tag{14}$$

Then, using the smoothness of the variables θ and ω respectively, we can get:

$$\begin{aligned}
\mathbb{L}(\theta_{t+1}, \omega_t) &\geq \mathbb{L}(\theta_t, \omega_t) + \langle \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \theta_{t+1} - \theta_t \rangle - \frac{l}{2}\|\theta_{t+1} - \theta_t\|^2; \\
\mathbb{L}(\theta_{t+1}, \omega_{t+1}) &\geq \mathbb{L}(\theta_{t+1}, \omega_t) + \langle \nabla_\omega \mathbb{L}(\theta_{t+1}, \omega_t), \omega_{t+1} - \omega_t \rangle - \frac{l}{2}\|\omega_{t+1} - \omega_t\|^2.
\end{aligned}$$

Taking expectation we can get:

$$\begin{aligned}
\mathbb{E}[\mathbb{L}(\theta_{t+1}, \omega_t)] &\geq \mathbb{E}[\mathbb{L}(\theta_t, \omega_t)] - \eta_\theta \mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t)\rangle - \frac{l\eta_\theta^2}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_t, \omega_t)] - \eta_\theta \mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 - \frac{\eta_\theta}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\quad - \frac{\eta_\theta}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 - \frac{l\eta_\theta^2}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_t, \omega_t)] - \frac{3\eta_\theta}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 - \frac{l^2\rho^2\eta_\theta}{2}\mathbb{E}\|g_\theta(\theta_t, \omega_t)\|^2 - \frac{l\eta_\theta^2}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_t, \omega_t)] - (\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1))\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\quad - (\frac{l^2\rho^2\eta_\theta}{2} + \frac{l\eta_\theta^2}{2}(5\rho^2 l^2 + 2))\frac{\sigma^2}{M}; \\
\mathbb{E}[\mathbb{L}(\theta_{t+1}, \omega_{t+1})] &\geq \mathbb{E}[\mathbb{L}(\theta_{t+1}, \omega_t)] + \eta_\omega \mathbb{E}\langle \nabla_\omega \mathbb{L}(\theta_{t+1}, \omega_t), \nabla_\omega \mathbb{L}(\theta_t, \omega_t)\rangle - \frac{l\eta_\omega^2}{2}\mathbb{E}\|g_\omega(\theta_t, \omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_{t+1}, \omega_t)] + \frac{\eta_\omega}{2}\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_t, \omega_t)\|^2 - \frac{\eta_\omega}{2}\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_{t+1}, \omega_t) - \nabla_\omega \mathbb{L}(\theta_t, \omega_t)\|^2 - \frac{l\eta_\omega^2}{2}\mathbb{E}\|g_\omega(\theta_t, \omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_{t+1}, \omega_t)] + \frac{\eta_\omega}{2}\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_t, \omega_t)\|^2 - \frac{l\eta_\theta^2\eta_\omega}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2 - \frac{l\eta_\omega^2}{2}\mathbb{E}\|g_\omega(\theta_t, \omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_{t+1}, \omega_t)] + (\frac{\eta_\omega}{2} - \frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_t, \omega_t)\|^2 - (l\eta_\theta^2\eta_\omega(2\rho^2 l^2 + \rho l + 1))\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\quad - (\frac{l\eta_\omega^2}{2} + \frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2 l^2 + 2))\frac{\sigma^2}{M}.
\end{aligned} \tag{15}$$

Then we construct a potential function in the same way as [4]:

$$V_t = V(\theta_t, \omega_t) = \mathbb{L}^*(\theta_t) + \alpha[\mathbb{L}^*(\theta_t) - \mathbb{L}(\theta_t, \omega_t)],$$

where $\alpha > 0$ is a preset parameter. Then we come to evaluate the descending relationship of the potential function V_t.

Combining the above inequalities we can get the descending relationship of the potential function:

$$\begin{aligned}
&\mathbb{E}[V_{t+1}] - \mathbb{E}[V_t] \\
&= (1 + \alpha)(\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] - \mathbb{E}[\mathbb{L}^*(\theta_t)]) - \alpha(\mathbb{E}[\mathbb{L}(\theta_{t+1}, \omega_{t+1})] - \mathbb{E}[\mathbb{L}(\theta_t, \omega_t)]) \\
&\leq (1 + \alpha)\{-\frac{\eta_\theta}{2}(1 - 5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \\
&\quad + [\frac{\eta_\theta}{2}(1 + \frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)]\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 + (5\rho^2 l^2 + 2)\frac{L\eta_\theta^2 \sigma^2}{2M}\} \\
&\quad - \alpha\{-(\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1))\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 - (\frac{l^2\rho^2\eta_\theta}{2} + \frac{l\eta_\theta^2}{2}(5\rho^2 l^2 + 2))\frac{\sigma^2}{M}
\end{aligned}$$

$$+(\frac{\eta_\omega}{2}-\frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega\mathbb{L}(\theta_t,\omega_t)\|^2-(l\eta_\theta^2\eta_\omega(2\rho^2l^2+\rho l+1))\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2-(\frac{l\eta_\omega^2}{2}+\frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2l^2+2))\frac{\sigma^2}{M}\} \quad (16)$$

$$=-\frac{\eta_\theta}{2}(1+\alpha)(1-5\rho l-2L\eta_\theta(4\rho^2l^2+2\rho l+2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2$$
$$+(1+\alpha)(\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l)+L\eta_\theta^2(4\rho^2l^2+2\rho l+2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)-\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2$$
$$+\alpha[(\frac{3\eta_\theta}{2}+\frac{l^2\rho^2\eta_\theta}{2}+l\eta_\theta^2(2\rho^2l^2+\rho l+1))+l\eta_\theta^2\eta_\omega(2\rho^2l^2+\rho l+1)]\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2$$
$$-\alpha(\frac{\eta_\omega}{2}-\frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega\mathbb{L}(\theta_t,\omega_t)\|^2$$
$$+[(1+\alpha)(5\rho^2l^2+2)\frac{L\eta_\theta^2}{2}+\alpha(\frac{l^2\rho^2\eta_\theta}{2}+\frac{l\eta_\theta^2}{2}(5\rho^2l^2+2))+\alpha(\frac{l\eta_\omega^2}{2}+\frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2l^2+2))]\frac{\sigma^2}{M}$$
$$\leq-\{\frac{\eta_\theta}{2}(1+\alpha)(1-5\rho l-2L\eta_\theta(4\rho^2l^2+2\rho l+2))-2\alpha[(\frac{3\eta_\theta}{2}+\frac{l^2\rho^2\eta_\theta}{2}+l\eta_\theta^2(2\rho^2l^2+\rho l+1))+l\eta_\theta^2\eta_\omega(2\rho^2l^2+\rho l+1)]\}$$
$$\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2$$
$$+\{(1+\alpha)(\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l)+L\eta_\theta^2(4\rho^2l^2+2\rho l+2))+2\alpha[(\frac{3\eta_\theta}{2}+\frac{l^2\rho^2\eta_\theta}{2}+l\eta_\theta^2(2\rho^2l^2+\rho l+1))+l\eta_\theta^2\eta_\omega(2\rho^2l^2+\rho l+1)]\}$$
$$\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)-\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2$$
$$-\alpha(\frac{\eta_\omega}{2}-\frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega\mathbb{L}(\theta_t,\omega_t)\|^2$$
$$+[(1+\alpha)(5\rho^2l^2+2)\frac{L\eta_\theta^2}{2}+\alpha(\frac{l^2\rho^2\eta_\theta}{2}+\frac{l\eta_\theta^2}{2}(5\rho^2l^2+2))+\alpha(\frac{l\eta_\omega^2}{2}+\frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2l^2+2))]\frac{\sigma^2}{M}. \quad (17)$$

Since we have the following property according to Lemma A.1 and the PL condition A.3:

$$\|\nabla\mathbb{L}^*(\theta_t)-\nabla_\theta f(\theta_t,\omega_t)\|\leq l\|\omega^*(\theta_t)-\omega_t\|\leq\kappa\|\nabla_\omega f(\theta_t,\omega_t)\|.$$

So we can further the above inequality as follows:

$$\mathbb{E}[V_{t+1}]-\mathbb{E}[V_t]$$
$$\leq-\{\frac{\eta_\theta}{2}(1+\alpha)(1-5\rho l-2L\eta_\theta(4\rho^2l^2+2\rho l+2))-2\alpha[(\frac{3\eta_\theta}{2}+\frac{l^2\rho^2\eta_\theta}{2}+l\eta_\theta^2(2\rho^2l^2+\rho l+1))+l\eta_\theta^2\eta_\omega(2\rho^2l^2+\rho l+1)]\}$$
$$\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2$$
$$-\{\alpha(\frac{\eta_\omega}{2}-\frac{l\eta_\omega^2}{2})-\kappa^2[(1+\alpha)(\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l)+L\eta_\theta^2(4\rho^2l^2+2\rho l+2))$$
$$+2\alpha[(\frac{3\eta_\theta}{2}+\frac{l^2\rho^2\eta_\theta}{2}+l\eta_\theta^2(2\rho^2l^2+\rho l+1))+l\eta_\theta^2\eta_\omega(2\rho^2l^2+\rho l+1)]]\}\mathbb{E}\|\nabla_\omega\mathbb{L}(\theta_t,\omega_t)\|^2$$
$$+[(1+\alpha)(5\rho^2l^2+2)\frac{L\eta_\theta^2}{2}+\alpha(\frac{l^2\rho^2\eta_\theta}{2}+\frac{l\eta_\theta^2}{2}(5\rho^2l^2+2))+\alpha(\frac{l\eta_\omega^2}{2}+\frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2l^2+2))]\frac{\sigma^2}{M}. \quad (18)$$

Then we require the parameters satisfy: $\alpha=\frac{1}{16}$, $\rho l\leq\frac{1}{16}$, $\eta_\theta(2\rho l+1)^2\kappa l\leq\frac{1}{64}$, $\kappa^2\eta_\theta l\leq\frac{1}{128}$, $\rho\leq\frac{\eta_\theta}{2l}$ and $\eta_\omega\leq 64\kappa^2\eta_\theta$. So the inequality can be further simplified as:

$$\mathbb{E}[V_{t+1}]-\mathbb{E}[V_t]$$
$$\leq-\frac{11}{80}\eta_\theta\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2-\frac{41}{32}\eta_\theta\kappa^2\mathbb{E}\|\nabla_\omega f(\theta_t,\omega_t)\|^2+129\kappa^4l\eta_\theta^2\frac{\sigma^2}{M}. \quad (19)$$

Telescoping the above inequality we can get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2\leq\frac{80}{11\eta_\theta T}(\mathbb{E}[V_0]-\mathbb{E}[V_T])+960\kappa^4l\eta_\theta\frac{\sigma^2}{M}. \quad (20)$$

Further, we can evaluate the first term that:

$$\begin{aligned}\mathbb{E}[V_0] - \mathbb{E}[V_T] &\leq \mathbb{E}[V_0] - \min_{\theta,\omega} \mathbb{E}[V(\theta,\omega)] \\ &\leq \mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_{\theta} \mathbb{E}[\mathbb{L}^*(\theta)] + \frac{1}{16}(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \mathbb{E}[\mathbb{L}(\theta_0,\omega_0)]) \\ &= \mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_{\theta} \mathbb{E}[\mathbb{L}^*(\theta)] + \frac{1}{16}\Delta_0,\end{aligned}$$

where we denote the initial error as: $\Delta_0 = \mathbb{E}[\mathbb{L}^*(\theta_0)] - \mathbb{E}[\mathbb{L}(\theta_0,\omega_0)]$.

Therefore, the inequality 20 can be further evaluated as:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \leq \frac{80}{11\eta_\theta T}(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_{\theta} \mathbb{E}[\mathbb{L}^*(\theta)]) + \frac{5}{11\eta_\theta T}\Delta_0 + 960\kappa^4 l\eta_\theta \frac{\sigma^2}{M}, \tag{21}$$

when we select $\eta_\theta = \sqrt{\frac{M(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])}{132T\kappa^4 l\sigma^2}}$, and samples can be minibatch, the convergence can be bounded by:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \leq 320\sqrt{\frac{3\kappa^4 l(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])\sigma^2}{11MT}} = \mathcal{O}(\frac{\kappa^2}{\sqrt{MT}}). \tag{22}$$

□

Table 1. Quantitative comparisons on distribution-aware robust generalization setting. Averaged accuracy (%) with standard deviations are computed over three independent trails.

Dataset	Type	Method	Corruption Severity 0	1	2	3	4	5
CIFAR10	Snow	ERM	90.8 ± 0.01	90.1 ± 0.02	88.1 ± 0.02	88.1 ± 0.02	85.7 ± 0.02	82.6 ± 0.01
		IRM	91.1 ± 0.02	90.7 ± 0.01	89.7 ± 0.02	88.0 ± 0.03	84.6 ± 0.02	83.2 ± 0.03
		REx	91.8 ± 0.02	$\mathbf{91.9 \pm 0.01}$	88.4 ± 0.01	88.3 ± 0.01	88.6 ± 0.01	83.0 ± 0.02
		GroupDRO	91.5 ± 0.02	91.0 ± 0.01	88.7 ± 0.02	88.6 ± 0.02	85.2 ± 0.03	83.5 ± 0.02
		SharpDRO	$\mathbf{93.1 \pm 0.01}$	91.8 ± 0.01	$\mathbf{90.5 \pm 0.02}$	$\mathbf{90.8 \pm 0.02}$	$\mathbf{87.9 \pm 0.01}$	$\mathbf{84.3 \pm 0.02}$
	Shot	ERM	$\mathbf{92.5 \pm 0.02}$	91.1 ± 0.02	89.9 ± 0.01	85.6 ± 0.03	85.7 ± 0.01	78.8 ± 0.01
		IRM	90.4 ± 0.01	90.3 ± 0.02	89.4 ± 0.02	86.3 ± 0.01	84.3 ± 0.02	79.1 ± 0.02
		REx	91.1 ± 0.02	90.6 ± 0.02	90.2 ± 0.03	86.8 ± 0.02	84.7 ± 0.02	80.5 ± 0.01
		GroupDRO	92.2 ± 0.01	$\mathbf{91.4 \pm 0.01}$	89.4 ± 0.02	84.0 ± 0.01	84.7 ± 0.02	78.3 ± 0.01
		SharpDRO	92.4 ± 0.02	91.1 ± 0.02	$\mathbf{90.3 \pm 0.02}$	$\mathbf{87.5 \pm 0.02}$	$\mathbf{86.4 \pm 0.02}$	$\mathbf{83.3 \pm 0.02}$
CIFAR100	Snow	ERM	67.7 ± 0.01	68.1 ± 0.01	64.7 ± 0.01	63.1 ± 0.01	60.5 ± 0.02	57.3 ± 0.01
		IRM	69.3 ± 0.01	67.5 ± 0.02	64.9 ± 0.02	61.0 ± 0.01	58.2 ± 0.01	55.1 ± 0.01
		REx	66.4 ± 0.01	65.9 ± 0.01	62.4 ± 0.01	61.2 ± 0.02	57.5 ± 0.03	56.0 ± 0.02
		GroupDRO	68.0 ± 0.02	68.2 ± 0.01	65.1 ± 0.01	60.9 ± 0.03	59.8 ± 0.01	58.1 ± 0.02
		SharpDRO	$\mathbf{71.5 \pm 0.01}$	$\mathbf{70.8 \pm 0.03}$	$\mathbf{67.5 \pm 0.02}$	$\mathbf{65.5 \pm 0.01}$	$\mathbf{62.3 \pm 0.01}$	$\mathbf{59.2 \pm 0.03}$
	Shot	ERM	67.6 ± 0.03	65.1 ± 0.01	62.9 ± 0.01	56.0 ± 0.01	55.1 ± 0.01	47.3 ± 0.01
		IRM	67.5 ± 0.02	65.7 ± 0.01	62.7 ± 0.01	59.5 ± 0.01	55.8 ± 0.01	48.3 ± 0.01
		REx	65.7 ± 0.01	63.8 ± 0.02	61.9 ± 0.01	59.3 ± 0.03	53.8 ± 0.01	48.1 ± 0.01
		GroupDRO	67.0 ± 0.02	65.8 ± 0.01	63.1 ± 0.01	58.9 ± 0.01	57.5 ± 0.01	49.3 ± 0.01
		SharpDRO	$\mathbf{69.2 \pm 0.01}$	$\mathbf{67.3 \pm 0.02}$	$\mathbf{65.4 \pm 0.03}$	$\mathbf{62.5 \pm 0.01}$	$\mathbf{57.7 \pm 0.02}$	$\mathbf{51.6 \pm 0.01}$
ImageNet30	Snow	ERM	86.7 ± 0.03	85.2 ± 0.01	83.4 ± 0.01	81.1 ± 0.01	75.3 ± 0.01	75.6 ± 0.01
		IRM	85.6 ± 0.01	84.0 ± 0.02	82.1 ± 0.03	79.7 ± 0.01	75.0 ± 0.01	75.6 ± 0.01
		REx	85.4 ± 0.01	84.6 ± 0.02	82.7 ± 0.02	80.5 ± 0.03	75.7 ± 0.03	75.9 ± 0.03
		GroupDRO	86.7 ± 0.01	85.5 ± 0.03	83.4 ± 0.01	81.2 ± 0.02	76.3 ± 0.01	76.6 ± 0.01
		SharpDRO	$\mathbf{88.2 \pm 0.02}$	$\mathbf{88.2 \pm 0.01}$	$\mathbf{85.4 \pm 0.02}$	$\mathbf{81.9 \pm 0.01}$	$\mathbf{79.8 \pm 0.03}$	$\mathbf{79.5 \pm 0.02}$
	Shot	ERM	86.9 ± 0.01	84.8 ± 0.01	83.6 ± 0.01	79.7 ± 0.01	75.4 ± 0.01	64.6 ± 0.01
		IRM	86.8 ± 0.01	85.1 ± 0.03	81.5 ± 0.01	73.5 ± 0.02	68.5 ± 0.03	62.5 ± 0.03
		REx	83.8 ± 0.01	86.3 ± 0.03	82.5 ± 0.02	73.9 ± 0.01	70.6 ± 0.03	64.0 ± 0.02
		GroupDRO	86.7 ± 0.01	85.6 ± 0.03	84.5 ± 0.01	80.7 ± 0.01	76.2 ± 0.04	65.4 ± 0.01
		SharpDRO	$\mathbf{88.1 \pm 0.01}$	$\mathbf{87.2 \pm 0.02}$	$\mathbf{84.7 \pm 0.01}$	$\mathbf{82.2 \pm 0.01}$	$\mathbf{78.2 \pm 0.01}$	$\mathbf{67.9 \pm 0.02}$

B. More Details

In this section, we first give a practical implementation of our SharpDRO. Then, we provide more experimental details.

B.1. Practical Implementation

Our SharpDRO requires two backward phases, so the time complexity is twice as much as plain training, for efficient sharpness computation, please refer to [1,2,6–8]. In the first step, we record the label prediction p of each data during inference and simultaneously compute the loss $\mathcal{L}$. Additionally, in the first backward pass, we store the computed gradient $\nabla\mathcal{L}(\theta)$. Further, by adding ϵ^*, we use the perturbed model to compute the second label prediction $\hat{p}$, which is further leveraged to compute the sharpness regularization $\mathcal{R}$. Moreover, in the distribution-agnostic setting, the predictions p and $\hat{p}$ from two forward steps are used to compute the OOD score ω_i. Then, we add the recorded gradient $\nabla\mathcal{L}(\theta)$ back to the model parameter and conduct sharpness minimization over the selected worst-case data. In this way, our SharpDRO can be correctly performed.

B.2. Experimental Details

In our experiments, we choose Wide ResNet-28-2 [5] as our backbone model, using stochastic gradient descent with learning rate $3e-2$ as the base optimizer. The momentum and weight decay factor of the optimizer is set to 0.9 and $5e-4$, respectively. We run all experiments for 200 epochs with three independent trials and report the average test accuracy with standard deviation.

Table 2. Quantitative comparisons on distribution-agnostic robust generalization setting. Averaged accuracy (%) with standard deviations are computed over three independent trails.

Dataset	Type	Method	Corruption Severity 0	1	2	3	4	5
CIFAR10	Snow	JTT	88.6 ± 0.02	87.8 ± 0.03	86.5 ± 0.02	87.2 ± 0.02	84.2 ± 0.02	83.2 ± 0.03
		EIIL	88.3 ± 0.02	87.8 ± 0.01	85.6 ± 0.02	87.3 ± 0.03	85.2 ± 0.04	82.3 ± 0.01
		SharpDRO	$\mathbf{91.6 \pm 0.01}$	$\mathbf{91.1 \pm 0.02}$	$\mathbf{90.8 \pm 0.01}$	$\mathbf{89.7 \pm 0.02}$	$\mathbf{86.2 \pm 0.01}$	$\mathbf{83.8 \pm 0.02}$
	Shot	JTT	91.3 ± 0.02	90.5 ± 0.03	89.3 ± 0.01	86.5 ± 0.02	83.1 ± 0.02	79.8 ± 0.02
		EIIL	90.3 ± 0.03	90.1 ± 0.02	88.3 ± 0.01	86.2 ± 0.02	82.3 ± 0.03	78.5 ± 0.02
		SharpDRO	$\mathbf{91.6 \pm 0.01}$	$\mathbf{90.5 \pm 0.02}$	$\mathbf{89.8 \pm 0.02}$	$\mathbf{88.7 \pm 0.01}$	$\mathbf{86.0 \pm 0.02}$	$\mathbf{81.7 \pm 0.01}$
CIFAR100	Snow	JTT	67.5 ± 0.01	68.1 ± 0.02	65.3 ± 0.02	64.3 ± 0.02	60.2 ± 0.02	57.8 ± 0.02
		EIIL	68.2 ± 0.03	69.1 ± 0.03	65.2 ± 0.02	64.0 ± 0.02	61.0 ± 0.04	57.5 ± 0.04
		SharpDRO	$\mathbf{70.6 \pm 0.02}$	$\mathbf{69.9 \pm 0.03}$	$\mathbf{66.7 \pm 0.03}$	$\mathbf{64.4 \pm 0.02}$	$\mathbf{61.9 \pm 0.03}$	$\mathbf{60.7 \pm 0.03}$
	Shot	JTT	66.3 ± 0.02	65.3 ± 0.03	63.4 ± 0.02	56.6 ± 0.04	55.5 ± 0.04	48.6 ± 0.04
		EIIL	66.5 ± 0.02	65.3 ± 0.03	62.8 ± 0.04	57.5 ± 0.02	56.5 ± 0.01	49.5 ± 0.01
		SharpDRO	$\mathbf{68.9 \pm 0.02}$	$\mathbf{66.2 \pm 0.03}$	$\mathbf{64.9 \pm 0.03}$	$\mathbf{60.1 \pm 0.02}$	$\mathbf{58.4 \pm 0.03}$	$\mathbf{52.7 \pm 0.02}$
ImageNet30	Snow	JTT	86.0 ± 0.04	85.8 ± 0.02	82.3 ± 0.03	80.4 ± 0.02	74.6 ± 0.02	73.5 ± 0.02
		EIIL	87.5 ± 0.01	85.4 ± 0.02	83.5 ± 0.04	$\mathbf{81.6 \pm 0.01}$	76.3 ± 0.01	75.8 ± 0.02
		SharpDRO	$\mathbf{87.5 \pm 0.03}$	$\mathbf{86.7 \pm 0.02}$	$\mathbf{85.4 \pm 0.02}$	81.5 ± 0.03	$\mathbf{78.9 \pm 0.02}$	$\mathbf{78.5 \pm 0.03}$
	Shot	JTT	86.5 ± 0.02	85.4 ± 0.03	82.6 ± 0.04	79.6 ± 0.04	77.2 ± 0.04	65.0 ± 0.01
		EIIL	85.5 ± 0.01	86.3 ± 0.04	81.6 ± 0.02	80.2 ± 0.03	75.3 ± 0.02	64.4 ± 0.03
		SharpDRO	$\mathbf{87.3 \pm 0.02}$	$\mathbf{87.2 \pm 0.03}$	$\mathbf{84.6 \pm 0.03}$	$\mathbf{83.2 \pm 0.06}$	$\mathbf{79.6 \pm 0.03}$	$\mathbf{68.0 \pm 0.03}$

C. Additional Experiments

In the main paper, we have provided the results using "Gaussian Noise" corruption and "JPEG compression" corruption, here we conduct additional experiments to show the effectiveness of SharpDRO under "Snow" and "Shot Noise" corruptions. The results on CIFAR10, CIFAR100, and ImageNet30 datasets in both distribution-aware and distribution-agnostic scenarios are shown in Tables 1 and 2. We can see that SharpDRO still performances effectively and surpasses other methods with large margin. Especially, on ImageNet30 dataset in both two problem settings, SharpDRO outperforms second-best method about 3%, which indicates the capability of SharpDRO on generalization against different corruptions.

References

[1] Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent YF Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *ICLR*, 2022. 8

[2] Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent YF Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *arXiv preprint arXiv:2205.14083*, 2022. 8

[3] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, volume 32, 2019. 2

[4] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *AISTATS*, pages 5485–5517. PMLR, 2022. 2, 5

[5] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 8

[6] Zhiyuan Zhang, Ruixuan Luo, Qi Su, and Xu Sun. Ga-sam: Gradient-strength based adaptive sharpness-aware minimization for improved generalization. *arXiv preprint arXiv:2210.06895*, 2022. 8

[7] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *ICML*, 2022. 8

[8] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Ss-sam: Stochastic scheduled sharpness-aware minimization for efficiently training deep neural networks. *arXiv preprint arXiv:2203.09962*, 2022. 8