Supplementary: Train simultaneously, generalize better: Stability of gradient-based minimax learners

Anonymous Authors[1]

A. Additional Numerical Results

A.1. Convex Concave Minimax Settings

Here, we provide the results of the numerical experiments discussed in the main text for full-batch GDA and PPM algorithms as well as stochastic and full-batch GDmax algorithms. Note that in these experiments we use the same minimax objective and hyperparameters mentioned in the main text. Figure 1 shows the generalization risk in our experiments for the GDA algorithm. As seen in Figure 1 (right), the results for full-batch and stochastic GDA algorithms in the bilinear convex concave case look similar, with the only exception that the generalization risk in the full-batch case reached a slightly higher amplitude of 7.8. On the other hand, in the strongly-convex strongly-concave case, full-batch GDA demonstrated a vanishing generalization risk, whereas stochastic GDA could not reach below an amplitude of 0.2.

Figure 2 shows the results of our experiments for full-batch PPM. Observe that the generalization risk in both cases decreases to reach smaller values than those for stochastic PPM. Finally, Figures 3 and 4 include the results for ful-batch and stochastic GDmax algorithms. With the exception of the full-batch GDmax case for the bilinear objective (Figure 3-right), in all the other cases the generalization risk did not grow during the optimization, which is comparable to our results in the GDA experiments.

A.2. Non-convex Non-concave Minimax Settings

Here, we provide the image samples generated by the trained GANs discussed in the main text. Figure 5 shows the CIFAR-10 samples generated by the simultaneous 1,1 Adam training (Figure 5-left) and non-simultaneous 1,100-Adam optimization (Figure 5-right). While we observed that the simultaneous training experiment generated qualitatively sharper samples, the non-simultaneous optimization did not lead to any significant training failures. However, as we discussed in the main text the generalization risk in the non-simultaneous training was significantly larger than that of simultaneous training. Figure 6 shows the generated images in the CelebA experiments, which are qualitatively comparable between the two training algorithms. However, as discussed in the text the trained discriminator had a harder task in classifying the training samples from the generated samples than in classifying the test samples from the generated samples, suggesting a potential overfitting of the training samples in the non-simultaneous training experiment.

B. Proofs

B.1. The Expansivity Lemma for Minimax Problems

We will apply the following lemma to analyze the stability of gradient-based methods. We call an update rule G γ-expansive if for every $\mathbf{w}, \mathbf{w}' \in \mathcal{W}, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ we have

$$\|G(\mathbf{w}, \boldsymbol{\theta}) - G(\mathbf{w}', \boldsymbol{\theta}')\|_2 \leq \gamma \sqrt{\|\mathbf{w} - \mathbf{w}'\|_2^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2}. \tag{1}$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

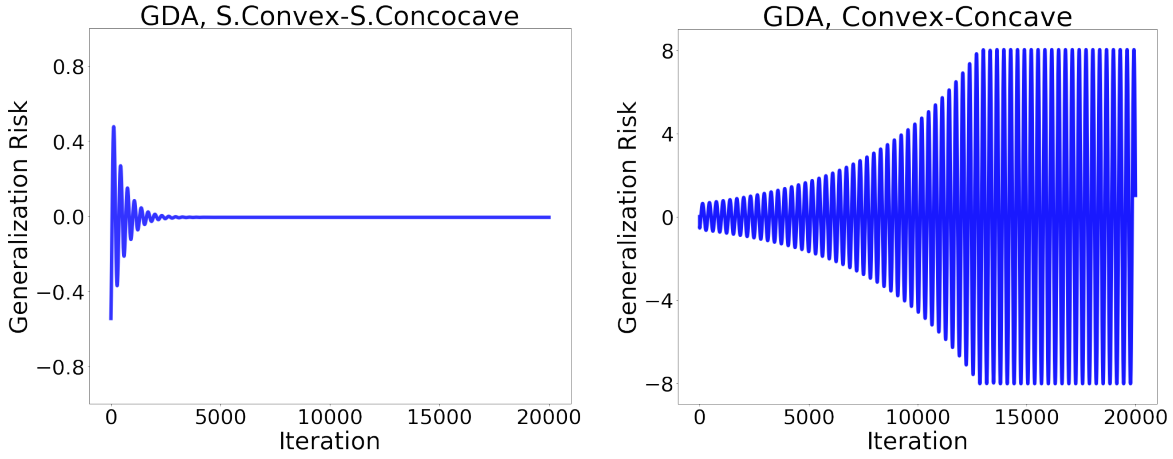

Figure 1: Generalization risk vs. iteration of full-batch GDA optimization in the (Left) strongly-convex strongly-concave setting and (Right) bilinear convex concave setting.

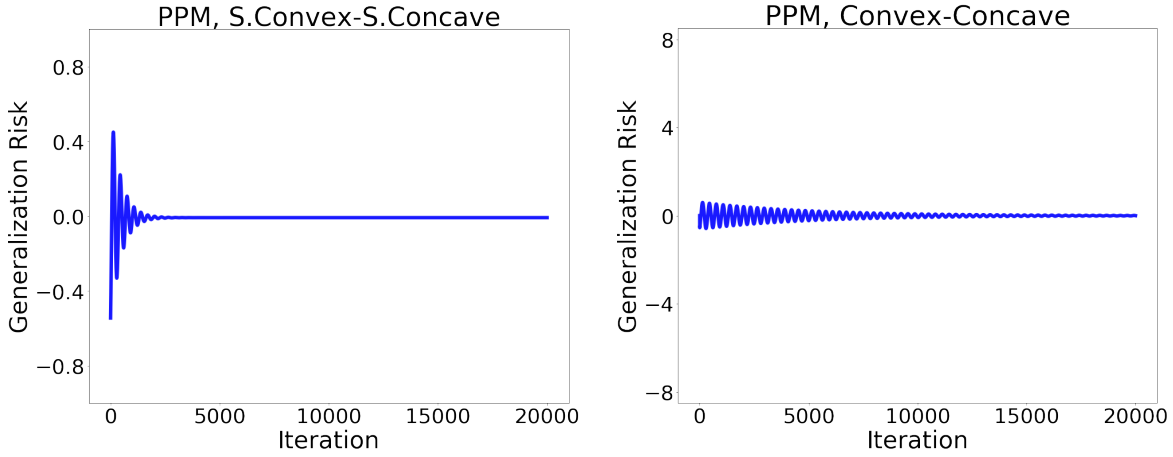

Figure 2: Generalization risk vs. iteration of full-batch PPM optimization in the (Left) strongly-convex strongly-concave setting and (Right) bilinear convex concave setting.

Lemma 1. *Consider the GDA and PPM updates for the following minimax problem*

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\boldsymbol{\theta}\in\Theta} f(\mathbf{w},\boldsymbol{\theta}), \tag{2}$$

where we assume objective $f(\mathbf{w},\boldsymbol{\theta})$ satisfies Assumptions 1 and 2. Then,

1. For a non-convex non-concave minimax problem, G_{GDA} is $(1+\ell\max\{\alpha_w,\alpha_\theta\})$-expansive. Assuming $\eta<\frac{1}{\ell}$, G_{PPM} will be $1/(1-\ell\eta)$-expansive.

2. For a convex concave minimax problem with $\alpha_w=\alpha_\theta$, G_{GDA} is $\sqrt{1+\ell^2\alpha_w^2}$-expansive and G_{PPM} will be 1-expansive.

3. For a μ-strongly-convex strongly-concave minimax problem, given that $\alpha_w=\alpha_\theta\leq\frac{2\mu}{\ell^2}$, G_{GDA} is $(1-\alpha_w\mu+\alpha_w^2\ell^2/2)$-expansive and G_{PPM} will be $1/(1+\mu\eta)$-expansive.

Proof. In Case 1 with non-convex non-concave minimax objective, f's smoothness property implies that for every $(\mathbf{w},\boldsymbol{\theta})$

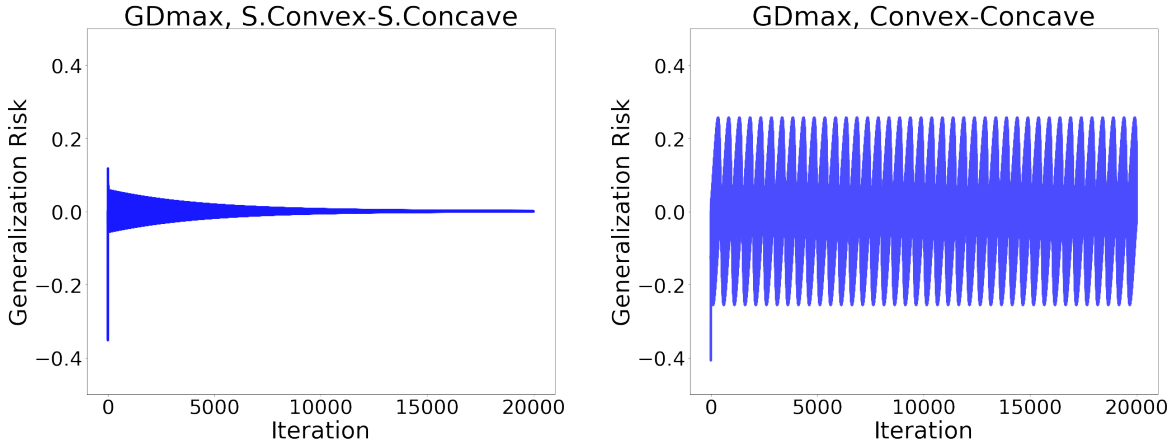

Figure 3: Generalization risk vs. iteration of full-batch GDmax optimization in the (Left) strongly-convex strongly-concave setting and (Right) bilinear convex concave setting.

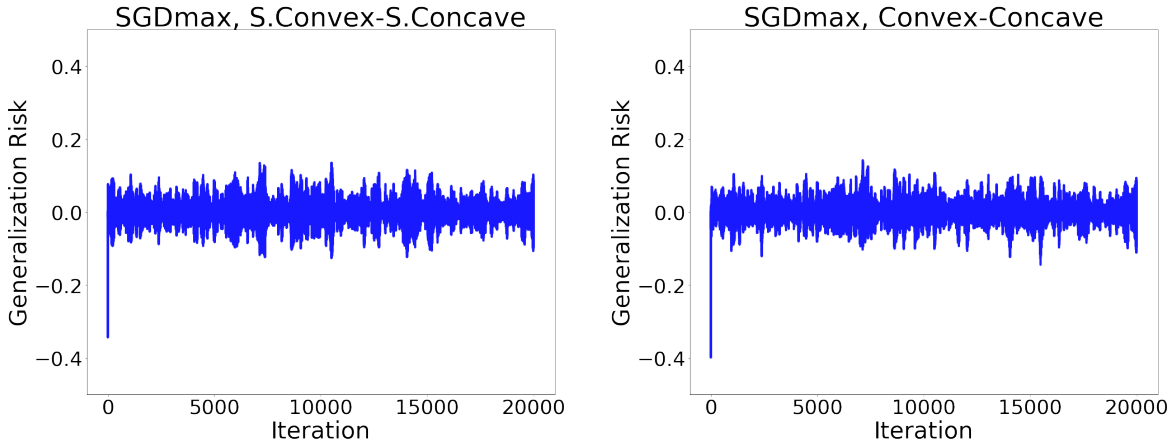

Figure 4: Generalization risk vs. iteration of stochastic GDmax optimization in the (Left) strongly-convex strongly-concave setting and (Right) bilinear convex concave setting.

and $(\mathbf{w}', \boldsymbol{\theta}')$:

$$\begin{aligned}\|G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) - G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}'\\ \boldsymbol{\theta}'\end{bmatrix})\| &= \| \begin{bmatrix}\mathbf{w} - \mathbf{w}' - \alpha_w(\nabla_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\theta}) - \nabla_{\mathbf{w}} f(\mathbf{w}', \boldsymbol{\theta}'))\\ \boldsymbol{\theta} - \boldsymbol{\theta}' + \alpha_\theta(\nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} f(\mathbf{w}', \boldsymbol{\theta}'))\end{bmatrix} \| \\ &\leq \| \begin{bmatrix}\mathbf{w} - \mathbf{w}'\\ \boldsymbol{\theta} - \boldsymbol{\theta}'\end{bmatrix} \| + \| \begin{bmatrix}\alpha_w(\nabla_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\theta}) - \nabla_{\mathbf{w}} f(\mathbf{w}', \boldsymbol{\theta}'))\\ \alpha_\theta(\nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} f(\mathbf{w}', \boldsymbol{\theta}'))\end{bmatrix} \| \\ &\leq (1 + \ell \max\{\alpha_w, \alpha_\theta\}) \| \begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix} - \begin{bmatrix}\mathbf{w}'\\ \boldsymbol{\theta}'\end{bmatrix} \|,\end{aligned}$$

which completes the proof for the GDA update. For the proximal operator, note that given $\eta \leq \frac{1}{\ell}$ the proximal optimization reduces to optimizing a strongly-convex strongly-concave minimax problem with a unique saddle solution and therefore at $(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}) = G_{\text{PPM}}(\mathbf{w}, \boldsymbol{\theta})$ we have

$$\mathbf{w}_{\text{PPM}} - \mathbf{w} = \eta \nabla_w f(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}), \quad \boldsymbol{\theta} - \boldsymbol{\theta}_{\text{PPM}} = \eta \nabla_\theta f(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}).$$

Figure 5: SN-GAN generated pictures in the CIFAR-10 experiments for (Left) simultaneous 1,1-Adam training (Right) non-simultaneous 1,100-Adam training.

Figure 6: SN-GAN generated pictures in the CelebA-10 experiments for (Left) simultaneous 1,1-Adam training (Right) non-simultaneous 1,100-Adam training.

As a result, we have

$$
\begin{aligned}
&\|G_{\text{PPM}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) - G_{\text{PPM}}(\begin{bmatrix}\mathbf{w}'\\ \boldsymbol{\theta}'\end{bmatrix})\| \\
=&\| \begin{bmatrix} \mathbf{w}-\mathbf{w}' + \eta(\nabla_{\mathbf{w}} f(G_{\text{PPM}}(\mathbf{w},\boldsymbol{\theta})) - \nabla_{\mathbf{w}} f(G_{\text{PPM}}(\mathbf{w}',\boldsymbol{\theta}'))) \\ \boldsymbol{\theta}-\boldsymbol{\theta}' - \eta(\nabla_{\boldsymbol{\theta}} f(G_{\text{PPM}}(\mathbf{w},\boldsymbol{\theta})) - \nabla_{\boldsymbol{\theta}} f(G_{\text{PPM}}(\mathbf{w}',\boldsymbol{\theta}')) \end{bmatrix} \| \\
\leq&\| \begin{bmatrix} \mathbf{w}-\mathbf{w}' \\ \boldsymbol{\theta}-\boldsymbol{\theta}' \end{bmatrix} \| + \| \begin{bmatrix} \eta(\nabla_{\mathbf{w}} f(G_{\text{PPM}}(\mathbf{w},\boldsymbol{\theta})) - \nabla_{\mathbf{w}} f(G_{\text{PPM}}(\mathbf{w}',\boldsymbol{\theta}'))) \\ \eta(\nabla_{\boldsymbol{\theta}} f(G_{\text{PPM}}(\mathbf{w},\boldsymbol{\theta})) - \nabla_{\boldsymbol{\theta}} f(G_{\text{PPM}}(\mathbf{w}',\boldsymbol{\theta}')) \end{bmatrix} \| \\
\leq&\| \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} \mathbf{w}' \\ \boldsymbol{\theta}' \end{bmatrix} \| + \frac{\eta}{\ell} \|G_{\text{PPM}}(\mathbf{w},\boldsymbol{\theta}) - G_{\text{PPM}}(\mathbf{w}',\boldsymbol{\theta}')\|.
\end{aligned}
$$

The final result of the above inequalities implies that

$$
(1-\frac{\eta}{\ell})\|G_{\text{PPM}}(\mathbf{w},\boldsymbol{\theta}) - G_{\text{PPM}}(\mathbf{w}',\boldsymbol{\theta}')\| \leq \| \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} \mathbf{w}' \\ \boldsymbol{\theta}' \end{bmatrix} \|,
$$

which completes the proof for the case of non-convex non-concave case.

For convex-concave objectives, the proof is mainly based on the monotonicity of convex concave objective's gradients (Rockafellar, 1976), implying that for every $\mathbf{w}, \mathbf{w}', \boldsymbol{\theta}, \boldsymbol{\theta}'$:

$$
(\begin{bmatrix} \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} \mathbf{w}' \\ \boldsymbol{\theta}' \end{bmatrix})^T (\begin{bmatrix} \nabla_{\mathbf{w}} f(\mathbf{w},\boldsymbol{\theta}) \\ -\nabla_{\boldsymbol{\theta}} f(\mathbf{w},\boldsymbol{\theta}) \end{bmatrix} - \begin{bmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}',\boldsymbol{\theta}') \\ -\nabla_{\boldsymbol{\theta}} f(\mathbf{w}',\boldsymbol{\theta}') \end{bmatrix}) \geq 0. \tag{3}
$$

As shown by (Rockafellar, 1976), the above property implies that the proximal operator for a convex-concave minimax objective will also be monotone and 1-expansive for any positive choice of η. For the GDA update, note that due to the

monotonicity property

$$\begin{aligned}
&\big\|G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) - G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}'\\ \boldsymbol{\theta}'\end{bmatrix})\big\|_2^2 \\
=&\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2 - 2\alpha_w \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}^T \begin{bmatrix}\nabla_{\mathbf{w}} f(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} f(\mathbf{w}',\boldsymbol{\theta}')\\ -\nabla_{\boldsymbol{\theta}} f(\mathbf{w},\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} f(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix} \\
&\quad + \alpha_w^2 \big\| \begin{bmatrix}\nabla_{\mathbf{w}} f(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} f(\mathbf{w}',\boldsymbol{\theta}')\\ \nabla_{\boldsymbol{\theta}} f(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} f(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix}\big\|_2^2 \\
\leq& (1+\alpha_w^2\ell^2)\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2,
\end{aligned}$$

which results in the following inequality and completes the proof for the convex-concave case:

$$\big\|G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) - G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}'\\ \boldsymbol{\theta}'\end{bmatrix})\big\|_2 \leq \sqrt{1+\alpha_w^2\ell^2}\, \big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2.$$

Finally, for the strongly-convex strongly-concave case, note that $\tilde{f}(\mathbf{w},\boldsymbol{\theta}) = f(\mathbf{w},\boldsymbol{\theta}) + \frac{\mu}{2}(\|\boldsymbol{\theta}\|^2 - \|\mathbf{w}\|^2)$ will be convex-concave and hence the proximal update $(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}) = G_{\text{PPM}}(\mathbf{w},\boldsymbol{\theta})$ will satisfy

$$\begin{aligned}
\frac{1}{1+\mu\eta}\mathbf{w} &= \mathbf{w}_{\text{PPM}} + \frac{\eta}{1+\mu\eta}\nabla_{\mathbf{w}}\tilde{f}(\mathbf{w}_{\text{PPM}},\boldsymbol{\theta}_{\text{PPM}}),\\
\frac{1}{1+\mu\eta}\boldsymbol{\theta} &= \boldsymbol{\theta}_{\text{PPM}} - \frac{\eta}{1+\mu\eta}\nabla_{\boldsymbol{\theta}}\tilde{f}(\mathbf{w}_{\text{PPM}},\boldsymbol{\theta}_{\text{PPM}}),
\end{aligned}$$

where the right-hand side follows from the proximal update for $\tilde{f}$ with stepsize $\eta/(1+\mu\eta)$ and hence 1-expansive. Therefore, the proximal update for f will be $1/(1+\mu\eta)$-expansive. Furthemore, for GDA udpates note that

$$\begin{aligned}
&\big\|G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) - G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}'\\ \boldsymbol{\theta}'\end{bmatrix})\big\|_2^2 \\
=&(1-\mu\alpha_w)^2\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2 - 2(1-\mu\alpha_w)\alpha_w \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}^T \begin{bmatrix}\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\\ -\nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix} \\
&\quad + \alpha_w^2 \big\| \begin{bmatrix}\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\\ \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix}\big\|_2^2 \\
\leq& ((1-\mu\alpha_w)^2 + \alpha_w^2(\ell^2-\mu^2))\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2 \\
\leq& (1-2\mu\alpha_w + \alpha_w^2\ell^2)\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2.
\end{aligned}$$

Note that the above result finishes the proof because $\sqrt{1-t} \leq 1 - t/2$ holds for every $t \leq 1$, which is based on the lemma's assumption $\alpha_w \leq 2\mu/\ell^2$. Also, the last inequality in the above holds since $\tilde{f}$ will be $\sqrt{\ell^2-\mu^2}$-smooth. This is because f is assumed to be ℓ-smooth, implying that for every $\mathbf{w}, \mathbf{w}', \boldsymbol{\theta}, \boldsymbol{\theta}'$ we have

$$\begin{aligned}
\ell^2\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2 &\geq \big\| \begin{bmatrix}\nabla_{\mathbf{w}} f(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} f(\mathbf{w}',\boldsymbol{\theta}')\\ \nabla_{\boldsymbol{\theta}} f(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} f(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix}\big\|_2^2 \\
&= \mu^2\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2 + 2\mu \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}'-\boldsymbol{\theta}\end{bmatrix}^T \begin{bmatrix}\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\\ \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix} \\
&\quad + \big\| \begin{bmatrix}\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\\ \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix}\big\|_2^2 \\
&\geq \mu^2\big\| \begin{bmatrix}\mathbf{w}-\mathbf{w}'\\ \boldsymbol{\theta}-\boldsymbol{\theta}'\end{bmatrix}\big\|_2^2 + \big\| \begin{bmatrix}\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\\ \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \tilde{f}(\mathbf{w}',\boldsymbol{\theta}')\end{bmatrix}\big\|_2^2,
\end{aligned}$$

where the inequality uses the monotonicity of the gradient operator. The final inequality shows that $\tilde{f}$ will be $\sqrt{\ell^2-\mu^2}$-smooth and hence finishes the proof. □

B.2. Proof of Theorem 1

Theorem. *(a) Assume minimax learner A is ϵ-uniformly stable in minimization. Then, A's expected minimax generalization risk is bounded as $\epsilon_{\text{gen}}^{\text{mm}}(A) \leq \epsilon$.*

(b) Assume minimax learner A is ϵ-uniformly stable in minimization. If the maximization problem over $\boldsymbol{\theta} \in \Theta$ can be swapped with the expectation over $\mathbf{Z}$, A's expected generalization risk will be bounded as $\epsilon_{\text{gen}}(A) \leq \epsilon$.

(c) Assume that minimax learner A is ϵ-uniformly stable in the minimization solution and the minimax objective is μ-strongly-concave in $\boldsymbol{\theta}$ over a convex feasible set Θ and satisfies Assumptions 1,2. Then, defining the condition number $\kappa := \ell/\mu$, A's expected generalization risk is bounded as $\epsilon_{\text{gen}}(A) \leq \sqrt{\kappa^2+1}\, L\, \epsilon$.

Proof. We start by proving the following lemma which is also shown in (Lin et al., 2019).

Lemma 2. *Consider a non-convex μ-strongly convex minimax objective $f(\mathbf{w}, \boldsymbol{\theta})$ satisfying Assumption 2 over a convex feasible set Θ. Then, the maximized objective $f_{\max}(\mathbf{w}) := \max_{\boldsymbol{\theta}\in\Theta} f(\mathbf{w}, \boldsymbol{\theta})$ will be $(\ell + \ell^2/2\mu)$-smooth, i.e., for every $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ it satisfies*

$$\big\|\nabla f_{\max}(\mathbf{w}_2) - \nabla f_{\max}(\mathbf{w}_1)\big\|_2 \leq \big(\ell + \frac{\ell^2}{2\mu}\big)\big\|\mathbf{w}_2 - \mathbf{w}_1\big\|_2, \tag{4}$$

Furthermore, defining $\boldsymbol{\theta}^*(\mathbf{w})$ as the optimal solution $\boldsymbol{\theta} \in \Theta$ for $\mathbf{w}$, we have that $\boldsymbol{\theta}^*(\mathbf{w})$ is ℓ/μ-Lipschitz.

Proof. Consider two arbitrary points $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ and define $\boldsymbol{\theta}^*(\mathbf{w}_1), \boldsymbol{\theta}^*(\mathbf{w}_2)$ as the optimal maximizers over Θ for $f(\mathbf{w}_1, \cdot), f(\mathbf{w}_2, \cdot)$, respectively. Since, $f(\mathbf{w}, \cdot)$ is ℓ-smooth and μ-strongly-convex, there exists a unique solution $\boldsymbol{\theta}^*(\mathbf{w})$ for every $\mathbf{w}$. Then, the μ-strongly concavity implies that

$$\mu\big\|\boldsymbol{\theta}^*(\mathbf{w}_1) - \boldsymbol{\theta}^*(\mathbf{w}_2)\big\|_2^2 \leq \big(\boldsymbol{\theta}^*(\mathbf{w}_2) - \boldsymbol{\theta}^*(\mathbf{w}_1)\big)^T\big(\nabla_{\boldsymbol{\theta}} f(\mathbf{w}_1, \boldsymbol{\theta}^*(\mathbf{w}_1)) - \nabla_{\boldsymbol{\theta}} f(\mathbf{w}_1, \boldsymbol{\theta}^*(\mathbf{w}_2))\big).$$

Due to the optimality of $\boldsymbol{\theta}^*(\mathbf{w}_1), \boldsymbol{\theta}^*(\mathbf{w}_2)$ over the convex feasible set Θ we further have

$$\begin{aligned}
&\big(\boldsymbol{\theta}^*(\mathbf{w}_2) - \boldsymbol{\theta}^*(\mathbf{w}_1)\big)^T\big(\nabla_{\boldsymbol{\theta}} f(\mathbf{w}_1, \boldsymbol{\theta}^*(\mathbf{w}_1)) - \nabla_{\boldsymbol{\theta}} f(\mathbf{w}_2, \boldsymbol{\theta}^*(\mathbf{w}_2))\big)\\
&= \big(\boldsymbol{\theta}^*(\mathbf{w}_2) - \boldsymbol{\theta}^*(\mathbf{w}_1)\big)^T\nabla_{\boldsymbol{\theta}} f(\mathbf{w}_1, \boldsymbol{\theta}^*(\mathbf{w}_1)) + \big(\boldsymbol{\theta}^*(\mathbf{w}_1) - \boldsymbol{\theta}^*(\mathbf{w}_2)\big)^T\nabla_{\boldsymbol{\theta}} f(\mathbf{w}_2, \boldsymbol{\theta}^*(\mathbf{w}_2))\\
&\leq 0.
\end{aligned}$$

Combining the above two equations, we obtain

$$\begin{aligned}
\mu\big\|\boldsymbol{\theta}^*(\mathbf{w}_1) - \boldsymbol{\theta}^*(\mathbf{w}_2)\big\|_2^2 &\leq \big(\boldsymbol{\theta}^*(\mathbf{w}_2) - \boldsymbol{\theta}^*(\mathbf{w}_1)\big)^T\big(\nabla_{\boldsymbol{\theta}} f(\mathbf{w}_2, \boldsymbol{\theta}^*(\mathbf{w}_2)) - \nabla_{\boldsymbol{\theta}} f(\mathbf{w}_1, \boldsymbol{\theta}^*(\mathbf{w}_2))\big)\\
&\leq \ell\big\|\boldsymbol{\theta}^*(\mathbf{w}_1) - \boldsymbol{\theta}^*(\mathbf{w}_2)\big\|_2\|\mathbf{w}_2 - \mathbf{w}_1\|_2.
\end{aligned}$$

The above equation results in

$$\big\|\boldsymbol{\theta}^*(\mathbf{w}_1) - \boldsymbol{\theta}^*(\mathbf{w}_2)\big\|_2 \leq \frac{\ell}{\mu}\|\mathbf{w}_2 - \mathbf{w}_1\|_2. \tag{5}$$

As a result, applying the Danskin's theorem for smooth objectives with a unique solution (Bernhard & Rapaport, 1995) implies that

$$\begin{aligned}
\big\|\nabla f_{\max}(\mathbf{w}_2) - \nabla f_{\max}(\mathbf{w}_1)\big\|_2 &= \big\|\nabla_{\mathbf{w}} f(\mathbf{w}_2, \boldsymbol{\theta}^*(\mathbf{w}_2)) - \nabla_{\mathbf{w}} f(\mathbf{w}_1, \boldsymbol{\theta}^*(\mathbf{w}_1))\big\|_2\\
&\leq \ell\sqrt{\big\|\mathbf{w}_2 - \mathbf{w}_1\big\|_2^2 + \big\|\boldsymbol{\theta}^*(\mathbf{w}_1) - \boldsymbol{\theta}^*(\mathbf{w}_2)\big\|_2^2}\\
&\leq \ell\sqrt{\big(1 + (\ell/\mu)^2\big)\|\mathbf{w}_2 - \mathbf{w}_1\|_2^2}\\
&= \ell\sqrt{1 + (\ell/\mu)^2}\|\mathbf{w}_2 - \mathbf{w}_1\|_2\\
&\leq \big(\ell + \frac{\ell^2}{2\mu^2}\big)\|\mathbf{w}_2 - \mathbf{w}_1\|_2,
\end{aligned}$$

where the last line holds since $\sqrt{1+t} \leq 1 + t/2$ for every $t \geq -1$. The proof is hence complete. □

Here, we provide a proof based on standard techniques in stability-based generalization theory (Bousquet & Elisseeff, 2002). Consider two independent datasets $S = (z_1, \ldots, z_n)$ and $S' = (z'_1, \ldots, z'_n)$. We use the notation $S^{(i)} = (z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n)$ to denote the dataset with the ith sample replaced with z'_i.

To show part (a), note that for every $\boldsymbol{\theta} \in \Theta$ we have

$$\begin{aligned}
\mathbb{E}_S\mathbb{E}_A[R_S(A_w(S), \boldsymbol{\theta})] &= \mathbb{E}_S\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A_w(S), \boldsymbol{\theta}; z_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A_w(S^{(i)}), \boldsymbol{\theta}; z'_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A_w(S), \boldsymbol{\theta}; z'_i)\right] + \zeta \\
&= \mathbb{E}_S\mathbb{E}_A[R(A_w(S), \boldsymbol{\theta})] + \zeta,
\end{aligned}$$

where we define

$$\zeta := \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A_w(S^{(i)}), \boldsymbol{\theta}; z'_i) - \frac{1}{n}\sum_{i=1}^{n} f(A_w(S), \boldsymbol{\theta}; z'_i)\right]$$

Note that according to the uniform-stability assumption we have $|f(A_w(S^{(i)}), \boldsymbol{\theta}; z'_i) - f(A_w(S), \boldsymbol{\theta}; z'_i)| \leq \epsilon$ which shows that $|\zeta| \leq \epsilon$ and shows that for every $\boldsymbol{\theta} \in \Theta$:

$$\left|\mathbb{E}_S\mathbb{E}_A[R(A_w(S), \boldsymbol{\theta}) - R_S(A_w(S), \boldsymbol{\theta})]\right| \leq \epsilon,$$

which completes the proof of part (a).

To show part (b), note that under the swapping condition we can move the maximization inside the summation since the max subproblems are independently solved for different data points. Then,

$$\begin{aligned}
\mathbb{E}_S\mathbb{E}_A[R_S(A_w(S))] &= \mathbb{E}_S\mathbb{E}_A\left[\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} f(A_w(S), \boldsymbol{\theta}; z_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} \max_{\boldsymbol{\theta}\in\Theta} f(A_w(S), \boldsymbol{\theta}; z_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} \max_{\boldsymbol{\theta}\in\Theta} f(A_w(S^{(i)}), \boldsymbol{\theta}; z'_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} \max_{\boldsymbol{\theta}\in\Theta} f(A_w(S), \boldsymbol{\theta}; z'_i)\right] + \zeta \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} f(A_w(S), \boldsymbol{\theta}; z'_i)\right] + \zeta \\
&= \mathbb{E}_S\mathbb{E}_A[R(A_w(S))] + \zeta.
\end{aligned}$$

In the above, ζ is defined as

$$\zeta := \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n}\left[\max_{\boldsymbol{\theta}\in\Theta} f(A_w(S^{(i)}), \boldsymbol{\theta}; z'_i) - \max_{\boldsymbol{\theta}'\in\Theta} f(A_w(S), \boldsymbol{\theta}'; z'_i)\right]\right].$$

Note that due to the uniform stability assumption for every data point z and datasets S, S' with only one different sample we have

$$\max_{\boldsymbol{\theta}\in\Theta} f(A_w(S), \boldsymbol{\theta}; z) - \max_{\boldsymbol{\theta}'\in\Theta} f(A_w(S'), \boldsymbol{\theta}'; z) \leq \max_{\boldsymbol{\theta}\in\Theta}\big\{f(A_w(S), \boldsymbol{\theta}; z) - f(A_w(S'), \boldsymbol{\theta}; z)\big\} \leq \epsilon.$$

Therefore, replacing the order of S, S' in the above inequality we obtain

$$\left|\max_{\boldsymbol{\theta}\in\Theta} f(A_w(S),\boldsymbol{\theta};z) - \max_{\boldsymbol{\theta}'\in\Theta} f(A_w(S'),\boldsymbol{\theta}';z)\right| \leq \epsilon.$$

As a result, we conclude that $|\zeta| \leq \epsilon$ which shows that

$$\left|\mathbb{E}_S\mathbb{E}_A[R_S(A_w(S))] - \mathbb{E}_S\mathbb{E}_A[R(A_w(S))]\right| \leq \epsilon.$$

The proof of part (b) is hence complete.

Finally, to prove part (c) we use $\boldsymbol{\theta}^*_S(\mathbf{w})$ to denote the optimal maximization solution for dataset S and minimization variable $\mathbf{w}$. Similarly, we use $\boldsymbol{\theta}^*(\mathbf{w})$ to denote the the optimal maximization solution for the underlying distribution P_Z and minimization variable $\mathbf{w}$. Then, we have

$$\begin{aligned}
\mathbb{E}_S\mathbb{E}_A[R_S(A_w(S))] &= \mathbb{E}_S\mathbb{E}_A\left[\max_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^n f(A_w(S),\boldsymbol{\theta};z_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n f(A_w(S),\boldsymbol{\theta}^*_S(A_w(S));z_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n f(A_w(S^{(i)}),\boldsymbol{\theta}^*_{S^{(i)}}(A_w(S^{(i)}));z'_i)\right] \\
&\geq \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n f(A_w(S^{(i)}),\boldsymbol{\theta}^*(A_w(S^{(i)}));z'_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n f(A_w(S),\boldsymbol{\theta}^*(A_w(S));z'_i)\right] + \zeta \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n \max_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{z\sim P_Z}\left[f(A_w(S),\boldsymbol{\theta}^*(A_w(S));z)\right]\right] + \zeta \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\max_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{z\sim P_Z}\left[f(A_w(S),\boldsymbol{\theta}^*(A_w(S));z)\right]\right] + \zeta \\
&= \mathbb{E}_S\mathbb{E}_A[R(A_w(S))] + \zeta.
\end{aligned}$$

In the above, we define

$$\zeta := \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^n\left[f(A_w(S^{(i)}),\boldsymbol{\theta}^*(A_w(S^{(i)}));z'_i) - f(A_w(S),\boldsymbol{\theta}^*(A_w(S));z'_i)\right]\right]$$

Lemma 2 implies that for every S, $\boldsymbol{\theta}^*_S(\mathbf{w})$ is κ-Lipschitz in $\mathbf{w}$, and hence for every z $f(\mathbf{w},\boldsymbol{\theta}^*(\mathbf{w});z)$ is $L\sqrt{\kappa^2+1}$-Lipschitz in $\mathbf{w}$. As a result, based on the uniform stability assumption we have

$$|\zeta| \leq \frac{L\sqrt{\kappa^2+1}}{n}\sum_{i=1}^n \left\|A_w(S) - A_w(S^{(i)})\right\|_2 \leq L\sqrt{\kappa^2+1}\epsilon,$$

which makes the proof of part (c) complete. □

B.3. Proof of Theorem 2

Note that in the following discussion we define PPmax as a proximal point method which fully optimizes the maximization variable at every iteration with the following update rule:

$$G_{\text{PPmax}}\left(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}\right) := \operatorname*{argmin}_{\widetilde{\mathbf{w}}\in\mathcal{W}} \operatorname*{argmax}_{\widetilde{\boldsymbol{\theta}}\in\Theta}\left\{f(\widetilde{\mathbf{w}},\widetilde{\boldsymbol{\theta}}) + \frac{1}{2\eta_w}\|\widetilde{\mathbf{w}}-\mathbf{w}\|_2^2\right\}.$$

Theorem. *Let minimax learning objective $f(\cdot,\cdot;\mathbf{z})$ be μ-strongly convex strongly-concave and satisfy Assumption 2 for every $\mathbf{z}$. Assume that Assumption 1 holds for convex-concave $\widetilde{f}(\mathbf{w},\boldsymbol{\theta};\mathbf{z}) := f(\mathbf{w},\boldsymbol{\theta};\mathbf{z}) + \frac{\mu}{2}(\|\boldsymbol{\theta}\|_2^2 - \|\mathbf{w}\|_2^2)$ and every $\mathbf{z}$. Then,*

1. *Full-batch and Stochastic GDA and GDmax with constant stepsize $\alpha_w = \alpha_\theta \leq \frac{2\mu}{\ell^2}$ for T iterations will satisfy*

$$\epsilon_{\text{gen}}(\text{GDA}), \epsilon_{\text{gen}}(\text{SGDA}) \leq \frac{2L^2\sqrt{\kappa^2+1}}{(\mu - \frac{\alpha_w \ell^2}{2})n}, \quad \epsilon_{\text{gen}}(\text{GDmax}), \epsilon_{\text{gen}}(\text{SGDmax}) \leq \frac{2L^2\sqrt{\kappa^2+1}}{\mu n}. \tag{6}$$

2. *Full-batch and stochastic PPM and PPmax with constant parameter η for T iterations will satisfy*

$$\epsilon_{\text{gen}}(\text{PPM}), \epsilon_{\text{gen}}(\text{SPPM}) \leq \frac{2L^2\sqrt{\kappa^2+1}}{\mu n}, \quad \epsilon_{\text{gen}}(\text{PPmax}), \epsilon_{\text{gen}}(\text{SPPmax}) \leq \frac{2L^2\sqrt{\kappa^2+1}}{\mu n}. \tag{7}$$

Proof. We start by proving the following lemmas.

Lemma 3 (Growth Lemma)**.** *Consider two sequences of updates $G_1,\ldots,G_T$ and $G'_1,\ldots.G'_t$ with the same starting point $\mathbf{w}_0 = \mathbf{w}'_0, \boldsymbol{\theta}_0 = \boldsymbol{\theta}'_0$. We define $\delta_t := \sqrt{\|\mathbf{w}_t - \mathbf{w}'_t\|^2 + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}'_t\|^2}$. Then, if G_T, G'_T is ξ-expansive we have $\delta_{t+1} \leq \xi\delta_t$ for identical $G_t = G'_t$, and in general we have*

$$\delta_{t+1} \leq \min\{\xi,1\}\delta_t + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|[\mathbf{w},\boldsymbol{\theta}] - G_t([\mathbf{w},\boldsymbol{\theta}])\|\} + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|[\mathbf{w},\boldsymbol{\theta}] - G'_t([\mathbf{w},\boldsymbol{\theta}])\|\}.$$

Furthermore, for any constant r we have

$$\delta_{t+1} \leq \xi\delta_t + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - G_t([\mathbf{w},\boldsymbol{\theta}])\|\} + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - G'_t([\mathbf{w},\boldsymbol{\theta}])\|\}.$$

Finally, if $G_t = G + \tilde{G}_t$ and $G'_t = G + \tilde{G}'_t$ for ξ_0-expansive G and ξ_1-expansive $\tilde{G}_t$ and $\tilde{G}'_t$, then for any constant r we have

$$\delta_{t+1} \leq (\xi_0+\xi_1)\delta_t + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - \tilde{G}_t([\mathbf{w},\boldsymbol{\theta}])\|\} + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - \tilde{G}'_t([\mathbf{w},\boldsymbol{\theta}])\|\}.$$

Proof. The first part of the theorem is a direct consequence of the definition of ξ-expansive operators. For the second part, note that

$$\begin{aligned}
\delta_{t+1} &= \|G_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&= \|G_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - [\mathbf{w}_t,\boldsymbol{\theta}_t] + [\mathbf{w}_t,\boldsymbol{\theta}_t] - [\mathbf{w}'_t,\boldsymbol{\theta}'_t] + [\mathbf{w}'_t,\boldsymbol{\theta}'_t] - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \|G_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - [\mathbf{w}_t,\boldsymbol{\theta}_t]\| + \|[\mathbf{w}_t,\boldsymbol{\theta}_t] - [\mathbf{w}'_t,\boldsymbol{\theta}'_t]\| + \|[\mathbf{w}'_t,\boldsymbol{\theta}'_t] - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \delta_t + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|[\mathbf{w},\boldsymbol{\theta}] - G_t([\mathbf{w},\boldsymbol{\theta}])\|\} + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|[\mathbf{w},\boldsymbol{\theta}] - G'_t([\mathbf{w},\boldsymbol{\theta}])\|\}.
\end{aligned}$$

In addition, we can bound δ_{t+1} as

$$\begin{aligned}
\delta_{t+1} &= \|G_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&= \|G_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - G_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t]) + G_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t]) - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \|G_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - G_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| + \|G_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t]) - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \xi\delta_t + \|G_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t]) - r[\mathbf{w}'_t,\boldsymbol{\theta}'_t]\| + \|r[\mathbf{w}'_t,\boldsymbol{\theta}'_t] - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \xi\delta_t + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - G_t([\mathbf{w},\boldsymbol{\theta}])\|\} + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - G'_t([\mathbf{w},\boldsymbol{\theta}])\|\}.
\end{aligned}$$

The above result for general constant r and also combined with the previous result with $r=1$ finishes the proof of the first two parts. For the final segment of the lemma, note that

$$\begin{aligned}
\delta_{t+1} &= \|G_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - G'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&= \|G([\mathbf{w}_t,\boldsymbol{\theta}_t]) + \tilde{G}_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - G([\mathbf{w}'_t,\boldsymbol{\theta}'_t]) - \tilde{G}'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \|G([\mathbf{w}_t,\boldsymbol{\theta}_t]) - G([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| + \|\tilde{G}_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - \tilde{G}'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \xi_0\delta_t + \|\tilde{G}_t([\mathbf{w}_t,\boldsymbol{\theta}_t]) - \tilde{G}'_t([\mathbf{w}'_t,\boldsymbol{\theta}'_t])\| \\
&\leq \xi_0\delta_t + \xi_1\delta_t + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - \tilde{G}_t([\mathbf{w},\boldsymbol{\theta}])\|\} + \sup_{\mathbf{w},\boldsymbol{\theta}}\{\|r[\mathbf{w},\boldsymbol{\theta}] - \tilde{G}'_t([\mathbf{w},\boldsymbol{\theta}])\|\}.
\end{aligned}$$

In the above equations, the last line follows from the second part of the lemma which finishes the proof. □

In order to show the Theorem for SGDA updates, note that given two datsets S, S' of size n with only one different sample at every iteration of stochastic GDA the update rule will be the same with probability $1-1/n$ and with probability $1/n$ we have two different $(1-\alpha_w\mu+\alpha_w^2\ell^2/2)$-expansive operators both of which satisfy

$$\sup_{\mathbf{w},\boldsymbol{\theta}}\{\|(1-\alpha_w\mu)[\mathbf{w},\boldsymbol{\theta}]-G_{\text{SGDA}}([\mathbf{w},\boldsymbol{\theta}])\|\}\leq L\alpha_w.$$

The above inequality holds, because $\tilde{f}$ is assumed to be continuously differentiable and L-Lipschitz. As a result, Lemmas 3,1 together with the law of total probability imply that the expected norm of $\delta_t^{\text{SGDA}}=\sqrt{\|\mathbf{w}_t-\mathbf{w}_t'\|^2+\|\boldsymbol{\theta}_t-\boldsymbol{\theta}_t'\|^2}$ for the SGDA updates applied to the two datasets will satisfy

$$\begin{aligned}\mathbb{E}[\delta_{t+1}^{\text{SGDA}}]&\leq(1-\frac{1}{n})(1-\alpha_w\mu+\frac{\alpha_w^2\ell^2}{2})\mathbb{E}[\delta_t^{\text{SGDA}}]+\frac{1}{n}\big((1-\alpha_w\mu+\frac{\alpha_w^2\ell^2}{2})\mathbb{E}[\delta_t^{\text{SGDA}}]+2\alpha_wL\big)\\&=(1-\alpha_w\mu+\frac{\alpha_w^2\ell^2}{2})\mathbb{E}[\delta_t^{\text{SGDA}}]+\frac{2\alpha_wL}{n}.\end{aligned}$$

Note that in the above upper-bound $1-\frac{1}{n}$ is the probability that the stochastic GDA algorithm chooses a shared sample between the two datasets and $\frac{1}{n}$ is the probability of picking the index of the different sample.

Similarly, the update rule for the full-batch GDA algorithm can be written as the sum of the updates for the shared samples, i.e., $\sum_{i=1}^{n-1}\frac{1}{n}G_{\text{GDA}}([\mathbf{w},\boldsymbol{\theta}];\mathbf{z}_i)$, and the different sample $\mathbf{z}_n$'s update $\frac{1}{n}G_{\text{GDA}}([\mathbf{w},\boldsymbol{\theta}];\mathbf{z}_n)$. As a result, the last part of Lemma 3 together with Lemma 1 implies that

$$\begin{aligned}\delta_{t+1}^{\text{GDA}}&\leq(1-\frac{1}{n}+\frac{1}{n})(1-\alpha_w\mu+\frac{\alpha_w^2\ell^2}{2})\delta_t^{\text{GDA}}+\frac{1}{n}(2\alpha_wL)\\&=(1-\alpha_w\mu+\frac{\alpha_w^2\ell^2}{2})\mathbb{E}[\delta_t^{\text{GDA}}]+\frac{2\alpha_wL}{n},\end{aligned}$$

which is the same bound we derived for stochastic GDA. Therefore, given that $\delta_0=0$, for SGDA updates we have

$$\begin{aligned}\mathbb{E}[\delta_t^{\text{SGDA}}]&\leq\frac{2\alpha_wL}{n}\sum_{i=0}^{t}(1-\alpha_w\mu+\frac{\alpha_w^2\ell^2}{2})^i\\&\leq\frac{2\alpha_wL}{n}\sum_{i=0}^{\infty}(1-\alpha_w\mu+\frac{\alpha_w^2\ell^2}{2})^i\\&=\frac{2\alpha_wL}{n(\alpha_w\mu-\frac{\alpha_w^2\ell^2}{2})}\\&=\frac{2L}{n(\mu-\frac{\alpha_w\ell^2}{2})}.\end{aligned}$$

Note that $\|\mathbf{w}_t-\mathbf{w}_t'\|\leq\delta_t$. As a result, the SGDA algorithm applied for T iterations will be $(2L/n(\mu-\alpha_w\ell^2/2))$-uniformly stable in the minimization variable, and the result follows from Theorem 1. The result for the GDA algorithm will follow from the same steps, since it shares the same growth rule with the SGDA algorithm.

Similarly, the SPPM updates will be $1/(1+\mu\eta)$-expansive due to Lemma 1. Furthermore, they will satisfy

$$\sup_{\mathbf{w},\boldsymbol{\theta}}\{\|\frac{1}{1+\eta\mu}[\mathbf{w},\boldsymbol{\theta}]-G_{\text{SPPM}}([\mathbf{w},\boldsymbol{\theta}])\|\}\leq\frac{L\eta}{1+\eta\mu}.$$

The above equation holds, because for a SPPM update $[\mathbf{w}_{\text{SPPM}},\boldsymbol{\theta}_{\text{SPPM}}]=G_{\text{SPPM}}([\mathbf{w},\boldsymbol{\theta}])$ at sample $\mathbf{z}$ we have

$$\begin{aligned}\begin{bmatrix}\mathbf{w}\\\boldsymbol{\theta}\end{bmatrix}&=\begin{bmatrix}(1+\eta\mu)\mathbf{w}_{\text{SPPM}}+\eta\nabla_w\tilde{f}(\mathbf{w}_{\text{SPPM}},\boldsymbol{\theta}_{\text{SPPM}};\mathbf{z})\\(1+\eta\mu)\boldsymbol{\theta}_{\text{SPPM}}-\eta\nabla_\theta\tilde{f}(\mathbf{w}_{\text{SPPM}},\boldsymbol{\theta}_{\text{SPPM}};\mathbf{z})\end{bmatrix}\\\Rightarrow\ \frac{1}{1+\eta\mu}\begin{bmatrix}\mathbf{w}\\\boldsymbol{\theta}\end{bmatrix}-\begin{bmatrix}\mathbf{w}_{\text{SPPM}}\\\boldsymbol{\theta}_{\text{SPPM}}\end{bmatrix}&=\frac{\eta}{1+\eta\mu}\begin{bmatrix}\nabla_w\tilde{f}(\mathbf{w}_{\text{SPPM}},\boldsymbol{\theta}_{\text{SPPM}};\mathbf{z})\\-\nabla_\theta\tilde{f}(\mathbf{w}_{\text{SPPM}},\boldsymbol{\theta}_{\text{SPPM}};\mathbf{z}).\end{bmatrix}\end{aligned}$$

Therefore, applying the law of total probability we will have

$$\begin{aligned}\mathbb{E}[\delta_{t+1}^{\text{SPPM}}] &\leq (1-\frac{1}{n})\frac{1}{1+\mu\eta}\mathbb{E}[\delta_t^{\text{SPPM}}] + \frac{1}{n}\big(\frac{1}{1+\mu\eta}\mathbb{E}[\delta_t^{\text{SPPM}}] + 2\frac{L\eta}{1+\eta\mu}\big)\\ &= \frac{1}{1+\mu\eta}\mathbb{E}[\delta_t^{\text{SPPM}}] + \frac{2L\eta}{(1+\eta\mu)n}.\end{aligned}$$

Also, for the PPM algorithm given that $\mathbf{z}_n$ denotes the different sample between datasets S, S' note that

$$\begin{aligned}\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix} &= \begin{bmatrix}(1+\eta\mu)\mathbf{w}_{\text{PPM}} + \frac{\eta}{n}\sum_{i=1}^{n-1}\nabla_w \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{SPPM}}; \mathbf{z}_i) + \frac{\eta}{n}\nabla_w \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{SPPM}}; \mathbf{z}_n)\\ (1+\eta\mu)\boldsymbol{\theta}_{\text{PPM}} - \frac{\eta}{n}\sum_{i=1}^{n-1}\nabla_\theta \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_i) + \frac{\eta}{n}\nabla_\theta \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_n)\end{bmatrix}\\ \Rightarrow \frac{1}{1+\eta\mu}\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix} - \begin{bmatrix}\mathbf{w}_{\text{SPPM}}\\ \boldsymbol{\theta}_{\text{SPPM}}\end{bmatrix} &= \frac{\eta}{1+\eta\mu}\Bigg(\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n-1}\nabla_w \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_i)\\ -\frac{1}{n}\sum_{i=1}^{n-1}\nabla_\theta \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_i)\end{bmatrix}\\ &\quad + \begin{bmatrix}\frac{1}{n}\nabla_w \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_n)\\ -\frac{1}{n}\nabla_\theta \tilde{f}(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_n)\end{bmatrix}\Bigg).\end{aligned}$$

In the above, the last line shows the sum of the updates for shared samples between the two datasets, i.e., $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}$, and the different sample $\mathbf{z}_n$. Therefore, Lemma 3 together with Lemma 1 implies that

$$\begin{aligned}\delta_{t+1}^{\text{PPM}} &\leq (1-\frac{1}{n}+\frac{1}{n})\frac{1}{1+\mu\eta}\delta_t^{\text{PPM}} + \frac{2L\eta}{n(1+\eta\mu)}\\ &= \frac{1}{1+\mu\eta}\delta_t^{\text{PPM}} + \frac{2L\eta}{(1+\eta\mu)n},\end{aligned}$$

which proves the same growth rule shown for SPPM also applies to the PPM algorithm. Since $\delta_0 = 0$, the above discussion implies the following for SPPM updates:

$$\begin{aligned}\mathbb{E}[\delta_t^{\text{SPPM}}] &\leq \frac{2L\eta}{(1+\eta\mu)n}\sum_{i=0}^{t}\big(\frac{1}{1+\mu\eta}\big)^i\\ &\leq \frac{2L\eta}{(1+\eta\mu)n}\sum_{i=0}^{\infty}\big(\frac{1}{1+\mu\eta}\big)^i\\ &= \frac{2L\eta}{(1+\eta\mu)n(1-1/(1+\mu\eta))}\\ &= \frac{2L}{n\mu}.\end{aligned}$$

Since $\|\mathbf{w}_t - \mathbf{w}_t'\| \leq \delta_t$, the SPPM algorithm applied for T iterations will be $(2L/n\mu)$-uniformly stable in the minimization variable. Therefore, the theorem's result is a corollary of Theorem 1. We can prove the result for the PPM algorithm by repeating the same steps we did for SPPM, as the two algorithms were shown to share the same growth rule.

For GDmax and PPmax algorithms, note that $\tilde{f}_{\max}(\mathbf{w}; S) := \max_{\boldsymbol{\theta}} \mathbb{E}_S[\tilde{f}(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})] - \frac{\mu}{2}\|\boldsymbol{\theta}\|^2$ will be convex and $L\sqrt{\kappa^2+1}$-Lipschitz in $\mathbf{w}$. Therefore, summing this function with $\frac{\mu}{2}\|\mathbf{w}\|^2$ will be μ-strongly convex. Since GDmax and SGDmax apply gradient descent to the maximized function, the theorem's result for GDmax and SGDmax follows from Theorem 3.9 in (Hardt et al., 2016). For SPPmax, we note that similar to Lemma 1 it can be seen that the proximal point updates will be $1/(1+\mu\eta)$-expansive. Moreover for the update $\mathbf{w}_{\text{SPPmax}} = G_{\text{SPPmax}}(\mathbf{w})$, we will have

$$\begin{aligned}&\mathbf{w} = (1+\eta\mu)\mathbf{w}_{\text{SPPmax}} + \eta\nabla_w \tilde{f}_{\max}(\mathbf{w}_{\text{SPPmax}}; \mathbf{z})\\ \Rightarrow\ &\frac{1}{1+\eta\mu}\mathbf{w} - \mathbf{w}_{\text{SPPmax}} = \frac{\eta}{1+\eta\mu}\nabla_w \tilde{f}_{\max}(\mathbf{w}_{\text{SPPmax}}; \mathbf{z}).\end{aligned}$$

As a result of Lemma 2.5 in (Hardt et al., 2016), defining $\delta_t^{\text{SPPmax}} = \|\mathbf{w}_t - \mathbf{w}_t'\|$ for datatsets S, S' we will have:

$$\begin{aligned}\mathbb{E}[\delta_{t+1}^{\text{SPPmax}}] &\leq (1-\frac{1}{n})\frac{1}{1+\mu\eta}\mathbb{E}[\delta_t^{\text{SPPmax}}] + \frac{1}{n}\big(\frac{1}{1+\mu\eta}\mathbb{E}[\delta_t^{\text{SPPmax}}] + 2\frac{L_w\eta}{1+\eta\mu}\big)\\ &= \frac{1}{1+\mu\eta}\mathbb{E}[\delta_t^{\text{SPPmax}}] + \frac{2L_w\eta}{(1+\eta\mu)n}.\end{aligned}$$

Furthermore for PPmax, we will have the following for $\mathbf{w}_{\text{PPM}} = G_{\text{PPM}}(\mathbf{w})$ when applied to the two datasets different in only the $\mathbf{z}_n$ sample:

$$\mathbf{w} = (1+\eta\mu)\mathbf{w}_{\text{PPmax}} + \frac{\eta}{n}\sum_{i=1}^{n-1}\nabla_w \tilde{f}_{\max}(\mathbf{w}_{\text{PPmax}};\mathbf{z}_i) + \frac{\eta}{n}\nabla_w \tilde{f}_{\max}(\mathbf{w}_{\text{PPmax}};\mathbf{z}_n)$$

$$\Rightarrow \quad \frac{1}{1+\eta\mu}\mathbf{w} - \mathbf{w}_{\text{PPM}} = \frac{\eta}{1+\eta\mu}\big(\frac{1}{n}\sum_{i=1}^{n-1}\nabla_w \tilde{f}_{\max}(\mathbf{w}_{\text{PPmax}};\mathbf{z}_i)\big) + \frac{\eta}{1+\eta\mu}\big(\frac{1}{n}\nabla_w \tilde{f}_{\max}(\mathbf{w}_{\text{PPmax}};\mathbf{z}_n)\big).$$

Applying Lemma 2.5 from (Hardt et al., 2016) and defining $\delta_t^{\text{PPmax}} = \|\mathbf{w}_t - \mathbf{w}_t'\|$ for datatsets S, S' we will have:

$$\begin{aligned}\delta_{t+1}^{\text{PPmax}} &\le (1 - \frac{1}{n} + \frac{1}{n})\frac{1}{1+\mu\eta}\delta_t^{\text{PPmax}} + \frac{1}{n}\big(2\frac{L\sqrt{\kappa^2+1}\eta}{1+\eta\mu}\big)\\ &= \frac{1}{1+\mu\eta}\delta_t^{\text{PPmax}} + \frac{2L\sqrt{\kappa^2+1}\eta}{(1+\eta\mu)n}.\end{aligned}$$

Note that $\delta_0^{\text{PPmax}} = \delta_0^{\text{SPPmax}} = 0$ which implies that:

$$\begin{aligned}\mathbb{E}[\delta_t^{\text{SPPmax}}] &\le \frac{2L\sqrt{\kappa^2+1}\eta}{(1+\eta\mu)n}\sum_{i=0}^{t}\big(\frac{1}{1+\mu\eta}\big)^i\\ &\le \frac{2L\sqrt{\kappa^2+1}\eta}{(1+\eta\mu)n}\sum_{i=0}^{\infty}\big(\frac{1}{1+\mu\eta}\big)^i\\ &= \frac{2L\sqrt{\kappa^2+1}\eta}{(1+\eta\mu)n(1-1/(1+\mu\eta))}\\ &= \frac{2L\sqrt{\kappa^2+1}}{n\mu}.\end{aligned}$$

Therefore, the SPPMax algorithm applied for T iterations will be $(2L^2\sqrt{\kappa^2+1}/n\mu)$-uniformly stable according to (Hardt et al., 2016)'s Definition 2.1. The result is hence a consequence of Theorem 2.2 in (Hardt et al., 2016). We can prove the result for the PPmax algorithm by repeating the same steps. □

B.4. Proof of Remark 1

Remark. *Consider a convex concave minimax objective $f(\cdot,\cdot;\mathbf{z})$ satisfying Assumptions 1 and 2. Given constant stepsizes $\alpha_w = \alpha_\theta = \alpha$, the GDA's generalization risk over T iterations will be bounded as:*

$$\epsilon_{\text{gen}}(\text{GDA}) \le O\big(\frac{\alpha L L_w(1+\alpha^2\ell^2)^{T/2}}{n}\big).$$

In particular, the bound's exponential dependence on T is tight for the GDA's generalization risk in the special case of $f(\mathbf{w},\boldsymbol{\theta};\mathbf{z}) = \mathbf{w}^\top(\mathbf{z}-\boldsymbol{\theta})$.

Proof. As shown in Lemma 1, the GDA's update will be $\sqrt{1+\alpha^2\ell^2}$-expansive in this case. As a result, in learning over two datasets S, S' which are different in only one sample, Lemma 3 shows the following growth rule for $\delta_t = \sqrt{\|\mathbf{w}_t - \mathbf{w}_t'\|_2^2 + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t'\|_2^2}$:

$$\begin{aligned}\delta_{t+1} &\le (\frac{n-1}{n} + \frac{1}{n})\sqrt{1+\alpha^2\ell^2}\delta_t + \frac{2\alpha L}{n}\\ &= \sqrt{1+\alpha^2\ell^2}\delta_t + \frac{2\alpha L}{n}.\end{aligned}$$

Considering that $\delta_0 = 0$, we get the following exponentially growing bound in T for δ_T:

$$\begin{aligned}\delta_T &\leq \sum_{t=1}^{T} \big(1+\alpha^2\ell^2\big)^{t/2} \frac{2\alpha L}{n} \\ &= \frac{2\alpha L}{n} \frac{\big(1+\alpha^2\ell^2\big)^{(T+1)/2} - 1}{\sqrt{1+\alpha^2\ell^2} - 1} \\ &= O\big(\frac{\alpha L \big(1+\alpha^2\ell^2\big)^{T/2}}{n}\big),\end{aligned}$$

which considering that $f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})$ is L_w-Lipschitz in $\mathbf{w}$ together with Theorem 1 shows that

$$\epsilon_{\text{gen}}(GDA) \leq O\big(\frac{\alpha L L_w (1+\alpha^2\ell^2)^{T/2}}{n}\big).$$

Also, note that for the special convex-concave case $f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z}) = \mathbf{w}^T(\mathbf{z} - \boldsymbol{\theta})$ given that $\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i$ the GDA's update rule will satisfy the following

$$\begin{aligned}&\begin{bmatrix} \mathbf{w}_{t+1} \\ \boldsymbol{\theta}_{t+1} - \bar{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \alpha\mathbf{I} \\ -\alpha\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \boldsymbol{\theta}_t - \bar{\mathbf{z}} \end{bmatrix}, \\ \Rightarrow \quad &\begin{bmatrix} \mathbf{w}_{t+1} \\ \boldsymbol{\theta}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \alpha\mathbf{I} \\ -\alpha\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \boldsymbol{\theta}_t \end{bmatrix} - \begin{bmatrix} \alpha\bar{\mathbf{z}} \\ \mathbf{0} \end{bmatrix}.\end{aligned}$$

As a result, for the updates on the two datasets S, S' with size n differing in only the sample $\mathbf{z}_n$ we have:

$$\begin{bmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}'_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \alpha\mathbf{I} \\ -\alpha\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \boldsymbol{\theta}_t - \boldsymbol{\theta}'_t \end{bmatrix} + \begin{bmatrix} \frac{\alpha}{n}(\mathbf{z}'_n - \mathbf{z}_n) \\ \mathbf{0} \end{bmatrix}.$$

Hence, knowing that $\mathbf{w}_0 = \mathbf{w}'_0$, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}'_0$ we have

$$\begin{bmatrix} \mathbf{w}_T - \mathbf{w}'_T \\ \boldsymbol{\theta}_T - \boldsymbol{\theta}'_T \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \alpha\mathbf{I} \\ -\alpha\mathbf{I} & \mathbf{I} \end{bmatrix}^T \begin{bmatrix} \frac{\alpha}{n}(\mathbf{z}'_n - \mathbf{z}_n) \\ \mathbf{0} \end{bmatrix}.$$

Since the matrix $\begin{bmatrix} 1 & \alpha \\ -\alpha & 1 \end{bmatrix}$ has the conjugate complex eigenvalues $1 \pm \alpha\mathrm{i}$, we will have

$$\Big\| \begin{bmatrix} \mathbf{w}_T - \mathbf{w}'_T \\ \boldsymbol{\theta}_T - \boldsymbol{\theta}'_T \end{bmatrix} \Big\|_2 = \big(\sqrt{1+\alpha^2}\big)^T \Big\| \begin{bmatrix} \frac{\alpha}{n}(\mathbf{z}'_n - \mathbf{z}_n) \\ \mathbf{0} \end{bmatrix} \Big\|_2 = \frac{\alpha\big(\sqrt{1+\alpha^2}\big)^T}{n} \|\mathbf{z}'_n - \mathbf{z}_n\|_2$$

As a consequence of the conjugate eigenvalues and the resulting iterative rotations in the complex space, the above equality shows that as long as $\alpha \neq 0$ for any constant $0 < C < 1$ there will exist arbitrarily large T values such that

$$\big\|\mathbf{w}_T - \mathbf{w}'_T\big\|_2 \geq \frac{C\alpha\big(\sqrt{1+\alpha^2}\big)^T}{n} \|\mathbf{z}'_n - \mathbf{z}_n\|_2.$$

Equivalently, we have

$$\|\mathbf{w}_T - \mathbf{w}'_T\|_2 = \Omega_T\big(\frac{\alpha(1+\alpha^2)^{T/2}}{n} \|\mathbf{z}'_n - \mathbf{z}_n\|_2\big), \tag{8}$$

which proves the exponential dependence of the expected generalization risk on T and completes the proof. □

B.5. Proof of Theorem 3

Here we prove the following generalized version of Theorem 3 in the text for general time-varying stepsize values.

Theorem. *Consider a convex-concave minimax learning objective $f(\cdot, \cdot; \mathbf{z})$ satisfying Assumptions 1 and 2 for every $\mathbf{z}$. Then, stochastic PPM with stepsizes η_t at iteration t over T iterations will satisfy*

$$\epsilon^{\text{mm}}_{\text{gen}}(\text{PPM}), \epsilon^{\text{mm}}_{\text{gen}}(\text{SPPM}) \leq \frac{2LL_w}{n} \sum_{t=1}^{T} \eta_t, \quad \epsilon^{\text{mm}}_{\text{gen}}(\text{PPmax}), \epsilon^{\text{mm}}_{\text{gen}}(\text{SPPmax}) \leq \frac{2L_w^2}{n} \sum_{t=1}^{T} \eta_t. \tag{9}$$

Proof. Consider two datasets S, S' with size n which have only one different sample. As a result of Lemma 1, the proximal point updates will be 1-expansive. Therefore, according to Lemma 3 and the Law of total probability, defining $\delta_t^{\text{SPPM}} = \sqrt{\|\mathbf{w}_t - \mathbf{w}_t'\|^2 + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t'\|^2}$ we will have

$$\begin{aligned}\mathbb{E}[\delta_{t+1}^{\text{SPPM}}] &\leq (1 - \frac{1}{n})\mathbb{E}[\delta_t^{\text{SPPM}}] + \frac{1}{n}\big(\mathbb{E}[\delta_t^{\text{SPPM}}] + 2\eta_t L\big) \\ &= \mathbb{E}[\delta_t^{\text{SPPM}}] + \frac{2\eta_t L}{n}\end{aligned}$$

Given that $\delta_0^{\text{SPPM}} = 0$, we reach the following inequality for every T

$$\mathbb{E}[\delta_T^{\text{SPPM}}] \leq \frac{2L}{n}\sum_{t=1}^{T}\eta_t.$$

Note that for every $\boldsymbol{\theta}, \mathbf{z}$, $f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})$ is L_w-Lipschitz, which with the above inequality implies that SPPM will be uniformly-stable in minimization with the following degree

$$\frac{2LL_w}{n}\sum_{t=1}^{T}\eta_t.$$

The theorem's result for SPPM then becomes a consequence of Theorem 1. Furthermore, regarding the PPM algorithm applying Lemma 3 and Lemma 1 implies that

$$\begin{aligned}\delta_{t+1}^{\text{PPM}} &\leq (1 - \frac{1}{n} + \frac{1}{n})\delta_t^{\text{PPM}} + \frac{2\eta_t L}{n} \\ &= \delta_t^{\text{PPM}} + \frac{2\eta_t L}{n}.\end{aligned}$$

The above equation holds because the update rule of PPM can be written in the following way where $\mathbf{z}_n$ denotes the only different sample between the two datasets,

$$\begin{bmatrix}\mathbf{w} \\ \boldsymbol{\theta}\end{bmatrix} - \begin{bmatrix}\mathbf{w}_{\text{PPM}} \\ \boldsymbol{\theta}_{\text{PPM}}\end{bmatrix} = \eta\left(\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n-1}\nabla_w f(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_i) \\ -\frac{1}{n}\sum_{i=1}^{n-1}\nabla_\theta f(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_i)\end{bmatrix} + \begin{bmatrix}\frac{1}{n}\nabla_w f(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_n) \\ -\frac{1}{n}\nabla_\theta f(\mathbf{w}_{\text{PPM}}, \boldsymbol{\theta}_{\text{PPM}}; \mathbf{z}_n)\end{bmatrix}\right).$$

Since $\delta_0^{\text{PPM}} = 0$, at iteration T we have

$$\delta_T^{\text{PPM}} \leq \frac{2L}{n}\sum_{t=1}^{T}\eta_t.$$

As a result, we can repeat the last step of our proof for the case of SPPM to complete the proof for the PPM case. For the PPmax and SPPmax algorithms, note that $f_{\max}(\mathbf{w}; \mathbf{z}) := \max_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})$ will be convex and L_w-Lipschitz in $\mathbf{w}$. The result is therefore a corollary of Theorem 3.8 and Lemma 4.6 in (Hardt et al., 2016). □

B.6. Proof of Theorem 4

Theorem. *Given a differentiable minimax objective $f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})$ the average iterate updates $\bar{\mathbf{w}}^{(T)} := \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$, $\bar{\boldsymbol{\theta}}^{(T)} := \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)}$ of SPPM and SPPmax with setpsize parameter η will satisfy the following optimality gaps for a saddle solution $[\mathbf{w}_S^*, \boldsymbol{\theta}_S^*]$ of the empirical risk for dataset S:*

$$\text{SPPM}: \ \mathbb{E}\big[\, R_S(\bar{\mathbf{w}}^{(T)})\,\big] - R_S(\mathbf{w}_S^*) \leq \frac{\|\mathbf{w}^{(0)} - \mathbf{w}_S^*\|^2 + \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_S^*\|^2}{2\eta T},$$

$$\text{SPPmax}: \ \mathbb{E}\big[R_S(\bar{\mathbf{w}}^{(T)})\big] - R_S(\mathbf{w}_S^*) \leq \frac{\|\mathbf{w}^{(0)} - \mathbf{w}_S^*\|^2}{2\eta T}.$$

Proof. Note that for any proximal operator F_k such that $\mathbf{v}_{k+1} = \mathbf{v}_k - \eta F_k(\mathbf{v}_{k+1})$ we will have the following for every $\mathbf{v}$:

$$\begin{aligned}
&\frac{1}{2\eta}\|\mathbf{v}_k - \mathbf{v}\|^2 - \frac{1}{2\eta}\|\mathbf{v}_{k+1} - \mathbf{v}\|^2 - \frac{1}{2\eta}\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 \\
&= -\frac{1}{\eta}\big(\|\mathbf{v}_{k+1}\|^2 - \mathbf{v}_k^T\mathbf{v}_{k+1} - \mathbf{v}^T\mathbf{v}_{k+1} + \mathbf{v}^T\mathbf{v}_k\big) \\
&= -\frac{1}{\eta}(\mathbf{v}_{k+1} - \mathbf{v}_k)^T(\mathbf{v}_{k+1} - \mathbf{v}) \\
&= F_k(\mathbf{v}_{k+1})^T(\mathbf{v}_{k+1} - \mathbf{v}).
\end{aligned}$$

As a result, we have

$$\begin{aligned}
\frac{1}{T}\sum_{k=0}^{T} F_k(\mathbf{v}_k)^T(\mathbf{v}_k - \mathbf{v}) &= \frac{1}{2\eta T}\|\mathbf{v}_0 - \mathbf{v}\|^2 - \frac{1}{2\eta T}\|\mathbf{v}_T - \mathbf{v}\|^2 - \frac{1}{T}\sum_{k=0}^{T}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 \\
&\leq \frac{1}{2\eta T}\|\mathbf{v}_0 - \mathbf{v}\|^2.
\end{aligned}$$

Given that every F_k is a stochastic proximal rule for a uniformly random training sample, the law of iterated expectation conditioned to random update $\mathbf{v}_t$ at iteration t implies that

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{T}\sum_{k=0}^{T} F_k(\mathbf{v}_k)^T(\mathbf{v}_k - \mathbf{v})\Big] &= \frac{1}{T}\sum_{k=0}^{T}\mathbb{E}\Big[F_k(\mathbf{v}_k)^T(\mathbf{v}_k - \mathbf{v})\Big] \\
&= \frac{1}{T}\sum_{k=0}^{T}\mathbb{E}\Big[\mathbb{E}\big[F_k(\mathbf{v}_k)^T(\mathbf{v}_k - \mathbf{v})\big|\mathbf{v}_k\big]\Big] \\
&= \frac{1}{T}\sum_{k=0}^{T}\mathbb{E}\Big[\mathbb{E}\big[F_k(\mathbf{v}_k)\big|\mathbf{v}_k\big]^T(\mathbf{v}_k - \mathbf{v})\Big] \\
&= \frac{1}{T}\sum_{k=0}^{T}\mathbb{E}\Big[\mathbb{E}\big[\bar{F}(\mathbf{v}_k)\big|\mathbf{v}_k\big]^T(\mathbf{v}_k - \mathbf{v})\Big] \\
&= \frac{1}{T}\sum_{k=0}^{T}\mathbb{E}\Big[\bar{F}(\mathbf{v}_k)^T(\mathbf{v}_k - \mathbf{v})\Big] \\
&= \mathbb{E}\Big[\frac{1}{T}\sum_{k=0}^{T}\bar{F}(\mathbf{v}_k)^T(\mathbf{v}_k - \mathbf{v})\Big]
\end{aligned}$$

where $\bar{F}$ denotes the gradient update for the averaged loss over the training samples. Therefore, we have

$$\mathbb{E}\Big[\frac{1}{T}\sum_{k=0}^{T}\bar{F}(\mathbf{v}_k)^T(\mathbf{v}_k - \mathbf{v})\Big] \leq \frac{1}{2\eta T}\|\mathbf{v}_0 - \mathbf{v}\|^2.$$

Considering the optimal saddle solution $\mathbf{v} = [\mathbf{w}_S^*, \boldsymbol{\theta}_S^*]$ for the SPPM algorithm, combining the above result with Lemma 2 in (Mokhtari et al., 2019) proves the theorem's result on the convergence of SPPM's average iterates. For the convergence result on SPPmax updates, note that given a convex function f and its gradient F and minimizer $\mathbf{v}^*$ we have

$$f\big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{v}^{(t)}\big) - f(\mathbf{v}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\big[f(\mathbf{v}^{(t)}) - f(\mathbf{v}^*)\big] \leq \frac{1}{T}\sum_{t=1}^{T}F(\mathbf{v}^{(t)})^T(\mathbf{v}^{(t)} - \mathbf{v}^*). \tag{10}$$

The above equation together with the property shown for the stochastic updates of SPPmax completes the theorem's proof. □

B.7. Proof of Corollary 1

Corollary. *Consider a convex concave minimax objective that satisfies the sawpping condition in Theorem 1b, which we optimize via PPM and PPmax with setpsize parameter η. Then, given that $\|\mathbf{w}^{(0)} - \mathbf{w}_S^*\|^2 + \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_S^*\|^2 \leq D^2$ for PPM and $\|\mathbf{w}^{(0)} - \mathbf{w}_S^*\| \leq D$ for PPmax holds with probability 1, it will take $T_{\text{SPPM}} = \sqrt{\frac{nD^2}{2\eta^2 LL_w}}$ and $T_{\text{SPPmax}} = \sqrt{\frac{nD^2}{2\eta^2 L_w^2}}$ iterations for the average iterates to achieve the following bounded excess risks where $\mathbf{w}^*$ denotes the optimal learner minimizing the true risk $R(\mathbf{w})$:*

$$\text{PPM, SPPM}: \ \mathbb{E}\big[R(\bar{\mathbf{w}}^{(T_{\text{SPPM}})})\big] - R(\mathbf{w}^*) \leq \sqrt{\frac{2D^2LL_w}{n}},$$
$$\text{PPmax, SPPmax}: \ \mathbb{E}\big[R(\bar{\mathbf{w}}^{(T_{\text{SPPmax}})})\big] - R(\mathbf{w}^*) \leq \sqrt{\frac{2D^2L_w^2}{n}}.$$

Proof. First, we show that using a constant stepsize parameter η the average iterates reach $1/2$ of the generalization bound for the final iterates in Theorem 3. For the average iterates $(\bar{\mathbf{w}}_t, \bar{\boldsymbol{\theta}}_t)$ and $(\bar{\mathbf{w}}'_t, \bar{\boldsymbol{\theta}}'_t)$ we have the following application of Jensen's inequality on the convex norm function for the difference of average iterates $\bar{\delta}_t = \sqrt{\|\bar{\mathbf{w}}^{(t)} - \bar{\mathbf{w}}'^{(t)}\|^2 + \|\bar{\boldsymbol{\theta}}^{(t)} - \bar{\boldsymbol{\theta}}'^{(t)}\|^2}$

$$\begin{aligned}
\bar{\delta}_t :&= \sqrt{\|\bar{\mathbf{w}}^{(t)} - \bar{\mathbf{w}}'^{(t)}\|^2 + \|\bar{\boldsymbol{\theta}}^{(t)} - \bar{\boldsymbol{\theta}}'^{(t)}\|^2} \\
&\leq \frac{1}{t}\sum_{k=0}^{t-1} \sqrt{\|\mathbf{w}_k - \mathbf{w}'_k\|^2 + \|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|^2} \\
&= \frac{1}{t}\sum_{k=0}^{t-1} \delta_t.
\end{aligned}$$

Similarly, one can show that $\bar{\delta}_{w,t} \leq \frac{1}{t}\sum_{k=1}^{t}\delta_{w,t}$. Therefore, knowing that $\mathbb{E}[\delta_t] \leq \frac{2LL_w t\eta}{n}$ implies that

$$\mathbb{E}[\bar{\delta}_t] \leq \frac{1}{t}\sum_{k=1}^{t}\mathbb{E}[\delta_t] \leq \frac{1}{t}\sum_{k=0}^{t-1}\frac{2LL_w k\eta}{n} \leq \frac{LL_w t\eta}{n}.$$

Hence, at the Tth average iterate of PPM and SPPM we will have

$$\mathbb{E}_A[R(\bar{\mathbf{w}}^{(T)})] - R_S[\bar{\mathbf{w}}^{(T)}] \leq \frac{LL_w T\eta}{n}$$

which together with (Mokhtari et al., 2019)'s Theorem 1 for the PPM and and our generalization of that theorem to stochastic PPM in Theorem 4 shows that

$$\mathbb{E}_{A,S}[R(\bar{\mathbf{w}}^{(T)})] - \mathbb{E}_S[R_S[\bar{\mathbf{w}}^{(T)}]] \leq \frac{LL_w\eta T}{n} + \frac{D^2}{2\eta T}.$$

Note that $\mathbb{E}_S[R_S(\mathbf{w}_S)] \leq \mathbb{E}_S[R_S(\mathbf{w}^*)] = R(\mathbf{w}^*)$, indicating that

$$\mathbb{E}_{A,S}[R(\bar{\mathbf{w}}^{(T)})] - R(\mathbf{w}^*) \leq \frac{LL_w\eta T}{n} + \frac{D^2}{2\eta T}.$$

The above upper-bound will be minimized when $\eta T = \sqrt{\frac{nD^2}{2LL_w}}$ and the optimized excess risk upper-bound for PPM and SPPM will be

$$\mathbb{E}_{A,S}[R(\bar{\mathbf{w}}^{(T)})] - R(\mathbf{w}^*) \leq \sqrt{\frac{2LL_wD^2}{n}}.$$

Similarly, it can be seen that for PPmax and SPPmax the optimal bound will be achieved at $\eta T = \sqrt{\frac{nD^2}{2L_w^2}}$ which suggests the following excess risk bound:

$$\mathbb{E}_{A,S}[R(\bar{\mathbf{w}}^{(T)})] - R(\mathbf{w}^*) \leq \sqrt{\frac{2L_w^2D^2}{n}}.$$

The proof is therefore complete. □

B.8. Proof of Theorem 5

Theorem. *Let learning objective $f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})$ be non-convex μ-strongly-concave and satisfy Assumptions 1 and 2. Also, we assume that $f_{\max}(\mathbf{w}; \mathbf{z}) := \max_{\boldsymbol{\theta} \in \Theta} f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})$ is bounded as $0 \leq f_{\max}(\mathbf{w}; \mathbf{z}) \leq 1$ for every $\mathbf{w}, \mathbf{z}$. Then, defining $\kappa := \ell/\mu$ we have*

1. The SGDA algorithm with vanishing stepsizes $\alpha_{w,t} = c/t$, $\alpha_{\theta,t} = cr^2/t$ for constants $c > 0, 1 \leq r \leq \kappa$ satisfies the following bound over T iterations:

$$\epsilon_{\text{gen}}(\text{SGDA}) \leq \frac{1 + \frac{1}{(r+1)c\ell}}{n} \big(12cL^2(r+1)\sqrt{\kappa^2+1}\big)^{\frac{1}{(r+1)c\ell+1}} T^{\frac{(r+1)c\ell}{(r+1)c\ell+1}}. \tag{11}$$

2. The SGDmax algorithm with vanishing stepsize $\alpha_{w,t} = c/t$ for constant $c > 0$ satisfies the following bound over T iterations:

$$\epsilon_{\text{gen}}(\text{SGDmax}) \leq \frac{1 + \frac{2}{(\kappa+2)\ell c}}{n-1} \big(2cL^2\sqrt{\kappa^2+1}\big)^{\frac{2}{(\kappa+2)\ell c+2}} T^{\frac{(\kappa+2)\ell c}{(\kappa+2)\ell c+2}}. \tag{12}$$

Proof. We start by proving the following lemmas.

Lemma 4. *Let $f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})$ satisgy Assumptions 1,2 and assume that $f_{\max}(\mathbf{w}; S) := \max_{\boldsymbol{\theta}} \mathbb{E}_S[f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z})]$ is bounded $0 \leq f_{\max}(\mathbf{w}; \mathbf{z}) \leq 1$. Then, in applying SGDA for learning over two datasets S, S' which differ in only one sample the updated variables $\mathbf{w}_t, \mathbf{w}'_t$ will satisfy the following inequality for every $t_0 \in \{1, \ldots, n\}$ where $\delta_t := \sqrt{\|\mathbf{w}_t - \mathbf{w}'_t\|^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}'_t\|^2}$:*

$$\forall S: \quad \mathbb{E}\big[|f_{\max}(\mathbf{w}_t; S) - f_{\max}(\mathbf{w}'_t; S)|\big] \leq \frac{t_0}{n} + L\sqrt{\kappa^2+1}\,\mathbb{E}[\delta_t | \delta_{t_0} = 0].$$

Proof. Define the event $E_{t_0} = \mathbb{I}(\delta_{t_0} = 0)$ as the indicator of the outcome $\delta_{t_0} = 0$. Then, due to the law of total probability

$$\begin{aligned}
\mathbb{E}\big[|f_{\max}(\mathbf{w}_t; \mathbf{z}) - f_{\max}(\mathbf{w}'_t; \mathbf{z})|\big] &= \Pr(E_{t_0})\mathbb{E}\big[|f_{\max}(\mathbf{w}_t; \mathbf{z}) - f_{\max}(\mathbf{w}'_t; \mathbf{z})| \,\big|E_{t_0}\big] \\
&\quad + \Pr(E^c_{t_0})\mathbb{E}\big[|f_{\max}(\mathbf{w}_t; \mathbf{z}) - f_{\max}(\mathbf{w}'_t; \mathbf{z})| \,\big|E^c_{t_0}\big] \\
&\overset{(a)}{\leq} \mathbb{E}\big[|f_{\max}(\mathbf{w}_t; \mathbf{z}) - f_{\max}(\mathbf{w}'_t; \mathbf{z})| \,\big|E_{t_0}\big] + \Pr(E^c_{t_0}) \\
&\overset{(b)}{\leq} L\sqrt{\kappa^2+1}\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}'_t\| \,\big|E_{t_0}\big] + \Pr(E^c_{t_0}) \\
&\overset{(c)}{\leq} L\sqrt{\kappa^2+1}\mathbb{E}\big[\delta_t \,\big|\delta_{t_0} = 0\big] + \frac{t_0}{n}.
\end{aligned}$$

In the above equations, (a) follows from the boundedness assumption on $f_{\max}$. (b) is the consequence of L_w-Lipschitzness of f which also transfers to $f_{\max}$. Finally, (c) holds because $\|\mathbf{w}_t - \mathbf{w}'_t\| \leq \delta_t$ according to the definition. Then, using the union bound on the outcome $I = I_t$ where I is the index of different samples in S, S' and I_t is the index of sample used by SGDA at iteration t we obtain that

$$\Pr(E^c_{t_0}) = \Pr(\delta_{t_0} > 0) \leq \sum_{i=1}^{t_0} \Pr(I = I_i) = \frac{t_0}{n}.$$

The lemma's proof is therefore complete. □

In order to prove the theorem for SGDA updates, we provide an extension of Lemma 1 for non-convex concave minimax objectives.

Lemma 5. *Consider a non-convex μ-strongly concace objective $f(\mathbf{w}, \boldsymbol{\theta})$ satisfying Assumption 2. Then, for every two pairs $(\mathbf{w}, \boldsymbol{\theta})$, $(\mathbf{w}', \boldsymbol{\theta}')$ the GDA updates $[\mathbf{w}_{\text{GDA}}, \boldsymbol{\theta}_{\text{GDA}}] = G_{\text{GDA}}([\mathbf{w}, \boldsymbol{\theta}])$, $[\mathbf{w}'_{\text{GDA}}, \boldsymbol{\theta}'_{\text{GDA}}] = G_{\text{GDA}}([\mathbf{w}', \boldsymbol{\theta}'])$ with stepsizes $\alpha_w, \alpha_\theta \leq \frac{1}{\ell}$ will satisfy the following expansivity equation:*

$$\begin{bmatrix} \|\mathbf{w}_{\text{GDA}} - \mathbf{w}'_{\text{GDA}}\| \\ \|\boldsymbol{\theta}_{\text{GDA}} - \boldsymbol{\theta}'_{\text{GDA}}\| \end{bmatrix} \leq \begin{bmatrix} 1 + \alpha_w\ell & \alpha_w\ell \\ \alpha_\theta\ell & 1 - \frac{\alpha_\theta\mu}{2} \end{bmatrix} \begin{bmatrix} \|\mathbf{w} - \mathbf{w}'\| \\ \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \end{bmatrix}.$$

Proof. Note that

$$\begin{aligned}\|\mathbf{w}_{\text{GDA}} - \mathbf{w}'_{\text{GDA}}\| &= \|\mathbf{w} - \alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\theta}) - \mathbf{w}' + \alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}', \boldsymbol{\theta}')\| \\ &\leq \|\mathbf{w} - \alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\theta}) - \mathbf{w}' + \alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}', \boldsymbol{\theta})\| \\ &\quad + \|\alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}', \boldsymbol{\theta}) - \alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}', \boldsymbol{\theta}')\| \\ &\leq (1 + \alpha_w \ell)\|\mathbf{w} - \mathbf{w}'\| + \alpha_w \ell \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\|\boldsymbol{\theta}_{\text{GDA}} - \boldsymbol{\theta}'_{\text{GDA}}\| &= \|\boldsymbol{\theta} + \alpha_\theta \nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta}) - \boldsymbol{\theta}' - \alpha_\theta \nabla_{\boldsymbol{\theta}} f(\mathbf{w}', \boldsymbol{\theta}')\| \\ &\leq \|\boldsymbol{\theta} + \alpha_\theta \nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta}) - \boldsymbol{\theta}' - \alpha_\theta \nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta}')\| \\ &\quad + \|\alpha_\theta \nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta}') - \alpha_\theta \nabla_{\boldsymbol{\theta}} f(\mathbf{w}', \boldsymbol{\theta}')\| \\ &\leq \big(1 - \frac{\alpha_\theta \mu}{2}\big)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| + \alpha_\theta \ell \|\mathbf{w} - \mathbf{w}'\|,\end{aligned}$$

where the last inequality follows from Lemma 3.7 in (Hardt et al., 2016) knowing that $\mu \leq \ell$. Therefore, the lemma's proof is complete. □

Lemma 6. *Consider two sequence of updates $G_1, \ldots, G_T$ and $G'_1, \ldots, G'_T$ for minimax objective $f(\mathbf{w}, \boldsymbol{\theta})$. Define $\delta_{w,t} = \|\mathbf{w}_t - \mathbf{w}'_t\|$ and $\delta_{\theta,t} = \|\boldsymbol{\theta}_t - \boldsymbol{\theta}'_t\|$. Assume that G_t is $\boldsymbol{\eta}$-expansive for matrix $\eta_{2\times 2}$, i.e. it satisfies the following inequality for every $[\mathbf{w}_{G_t}, \boldsymbol{\theta}_{G_t}] := G_t(\mathbf{w}, \boldsymbol{\theta})$, $[\mathbf{w}'_{G_t}, \boldsymbol{\theta}'_{G_t}] := G_t(\mathbf{w}', \boldsymbol{\theta}')$*

$$\begin{bmatrix}\|\mathbf{w}_{G_t} - \mathbf{w}'_{G_t}\| \\ \|\boldsymbol{\theta}_{G_t} - \boldsymbol{\theta}'_{G_t}\|\end{bmatrix} \leq \boldsymbol{\eta} \begin{bmatrix}\|\mathbf{w} - \mathbf{w}'\| \\ \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|\end{bmatrix}.$$

Also, suppose that for every $[\mathbf{w}_{G_t}, \boldsymbol{\theta}_{G_t}] := G_t(\mathbf{w}, \boldsymbol{\theta})$, $[\mathbf{w}_{G_t}, \boldsymbol{\theta}_{G'_t}] := G'_t(\mathbf{w}, \boldsymbol{\theta})$ we have

$$\begin{aligned}&\sup_{\mathbf{w},\boldsymbol{\theta}} \|\mathbf{w}_{G_t} - \mathbf{w}\| \leq \sigma_w, \quad \sup_{\mathbf{w},\boldsymbol{\theta}} \|\boldsymbol{\theta}_{G_t} - \boldsymbol{\theta}\| \leq \sigma_\theta, \\ &\sup_{\mathbf{w},\boldsymbol{\theta}} \|\mathbf{w}_{G'_t} - \mathbf{w}\| \leq \sigma_w, \quad \sup_{\mathbf{w},\boldsymbol{\theta}} \|\boldsymbol{\theta}_{G'_t} - \boldsymbol{\theta}\| \leq \sigma_\theta.\end{aligned}$$

Then, we have

$$\begin{bmatrix}\delta_{w,t+1} \\ \delta_{\theta,t+1}\end{bmatrix} \leq \boldsymbol{\eta} \begin{bmatrix}\delta_{w,t} \\ \delta_{\theta,t}\end{bmatrix} + 2 \begin{bmatrix}\sigma_w \\ \sigma_\theta\end{bmatrix}.$$

Proof. Note that

$$\begin{aligned}\begin{bmatrix}\delta_{w,t+1} \\ \delta_{\theta,t+1}\end{bmatrix} &= \begin{bmatrix}\|G_{t,w}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G'_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\| \\ \|G_{t,\theta}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G'_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\|\end{bmatrix} \\ &= \begin{bmatrix}\|G_{t,w}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) + G_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) - G'_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\| \\ \|G_{t,\theta}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) + G_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) - G'_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\|\end{bmatrix} \\ &= \begin{bmatrix}\|G_{t,w}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\| \\ \|G_{t,\theta}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\|\end{bmatrix} + \begin{bmatrix}\|G_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) - G'_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\| \\ \|G_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) - G'_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\|\end{bmatrix} \\ &= \begin{bmatrix}\|G_{t,w}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\| \\ \|G_{t,\theta}(\mathbf{w}_t, \boldsymbol{\theta}_t) - G_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\|\end{bmatrix} + \begin{bmatrix}\|G_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) - \mathbf{w}'_t\| \\ \|G_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t) - \boldsymbol{\theta}'_t\|\end{bmatrix} \\ &\quad + \begin{bmatrix}\|\mathbf{w}'_t - G'_{t,w}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\| \\ \|\boldsymbol{\theta}'_t - G'_{t,\theta}(\mathbf{w}'_t, \boldsymbol{\theta}'_t)\|\end{bmatrix} \\ &\leq \boldsymbol{\eta} \begin{bmatrix}\delta_{w,t} \\ \delta_{\theta,t}\end{bmatrix} + 2 \begin{bmatrix}\sigma_w \\ \sigma_\theta\end{bmatrix},\end{aligned}$$

which makes the proof complete. □

To prove the theorem's result on SGDA note that Lemma 5 suggests that the SGDA update at iteration t for non-convex non-concave problems will be expansive with the following matrix:

$$B_t := \begin{bmatrix} 1+\alpha_{w,t}\ell & \alpha_{w,t}\ell \\ \alpha_{\theta,t}\ell & 1-\frac{\alpha_{\theta,t}\mu}{2} \end{bmatrix} = I + \alpha_{w,t}\ell \begin{bmatrix} 1 & \frac{1}{\alpha_{\theta,t}} \\ \frac{\alpha_{\theta,t}}{\alpha_{w,t}} & -\frac{\mu\alpha_{\theta,t}}{\ell\alpha_{w,t}} \end{bmatrix} = I + \frac{c\ell}{t}\begin{bmatrix} 1 & 1 \\ r^2 & -r^2/\kappa \end{bmatrix}.$$

For analyzing the powers of the above matrix, we diagonalize it using its eigenvalues λ_1, λ_2 and corresponding eigenvectors $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2$. Note that the product of the eigenvalues of $\begin{bmatrix} 1 & 1 \\ r^2 & -r^2/\kappa \end{bmatrix}$, i.e. the matrix's determinant, is negative and hence the matrix has two different real eigenvalues with opposite signs. This implies that the matrix is diagonlizable and so is a linear combination of the matrix with the identity matrix. As a result, given the invertible matrix $\boldsymbol{\nu} = [\boldsymbol{\nu}_1, \boldsymbol{\nu}_2]$ we have

$$B_t = \begin{bmatrix} 1+\alpha_{w,t}\ell & \alpha_{w,t}\ell \\ \alpha_{\theta,t}\ell & 1-\frac{\alpha_{\theta,t}\mu}{2} \end{bmatrix} = \boldsymbol{\nu}^{-1}\begin{bmatrix} 1+\frac{c\ell\lambda_1}{t} & 0 \\ 0 & 1+\frac{c\ell\lambda_2}{t} \end{bmatrix}\boldsymbol{\nu}.$$

Also, notice that we have the following closed-form solution for λ_1, λ_2:

$$\lambda_1 = \frac{\kappa - r^2 + \sqrt{4\kappa^2 r^2 + (\kappa + r^2)^2}}{2\kappa}, \quad \lambda_2 = \frac{\kappa - r^2 - \sqrt{4\kappa^2 r^2 + (\kappa + r^2)^2}}{2\kappa}.$$

Therefore, since we assume $1 \leq r \leq \kappa$,

$$\max\{\lambda_1, \lambda_2\} \leq \frac{1 - \frac{r^2}{\kappa} + (2r + (\frac{r^2}{\kappa} + 1))}{2} = r + 1.$$

Now, applying the law of total probability as well as Lemma 6 shows that

$$\begin{aligned} \begin{bmatrix} \mathbb{E}[\delta_{w,t+1}] \\ \mathbb{E}[\delta_{\theta,t+1}] \end{bmatrix} &\leq (1-\frac{1}{n})B_t \begin{bmatrix} \mathbb{E}[\delta_{w,t}] \\ \mathbb{E}[\delta_{\theta,t}] \end{bmatrix} + \frac{1}{n}\big(B_t \begin{bmatrix} \mathbb{E}[\delta_{w,t}] \\ \mathbb{E}[\delta_{\theta,t}] \end{bmatrix} + 2\begin{bmatrix} \alpha_{w,t}L_w \\ \alpha_{\theta,t}L_\theta \end{bmatrix}\big) \\ &= B_t \begin{bmatrix} \mathbb{E}[\delta_{w,t}] \\ \mathbb{E}[\delta_{\theta,t}] \end{bmatrix} + \begin{bmatrix} \frac{2cL_w}{nt} \\ \frac{2cr^2L_\theta}{nt} \end{bmatrix}. \end{aligned}$$

Therefore, over T iterations we will have

$$\begin{aligned} \begin{bmatrix} \mathbb{E}[\delta_{w,T}] \\ \mathbb{E}[\delta_{\theta,T}] \end{bmatrix} &\leq \sum_{t=t_0+1}^{T}\{\prod_{k=t+1}^{T} B_k\}\begin{bmatrix} \frac{2cL_w}{nt} \\ \frac{2cr^2L_\theta}{nt} \end{bmatrix} \\ &= \sum_{t=t_0+1}^{T}\boldsymbol{\nu}^{-1}\{\prod_{k=t+1}^{T}\begin{bmatrix} 1+\frac{c\ell\lambda_1}{k} & 0 \\ 0 & 1+\frac{c\ell\lambda_2}{k} \end{bmatrix}\}\boldsymbol{\nu}\begin{bmatrix} \frac{2cL_w}{nt} \\ \frac{2cr^2L_\theta}{nt} \end{bmatrix}. \end{aligned}$$

Hence, denoting the minimum and maximum singular values of $\boldsymbol{\nu}$ with $\sigma_{\min}(\boldsymbol{\nu}), \sigma_{\max}(\boldsymbol{\nu})$ and noting that $\boldsymbol{\nu}^{-1}$'s operator

norm is equal to $1/\sigma_{\min}(\boldsymbol{\nu})$ we will have

$$\begin{aligned}
\left\| \begin{bmatrix} \mathbb{E}[\delta_{w,T}] \\ \mathbb{E}[\delta_{\theta,T}] \end{bmatrix} \right\|_2 &\leq \frac{\sigma_{\max}(\boldsymbol{\nu})}{\sigma_{\min}(\boldsymbol{\nu})} \sum_{t=t_0+1}^{T} \left\| \{ \prod_{k=t+1}^{T} \begin{bmatrix} 1+\frac{c\ell\lambda_1}{k} & 0 \\ 0 & 1+\frac{c\ell\lambda_2}{k} \end{bmatrix} \} \begin{bmatrix} \frac{2cL_w}{nt} \\ \frac{2cr^2L_\theta}{nt} \end{bmatrix} \right\|_2 \\
&\leq \frac{\sigma_{\max}(\boldsymbol{\nu})}{\sigma_{\min}(\boldsymbol{\nu})} \sum_{t=t_0+1}^{T} \left\| \prod_{k=t+1}^{T} \begin{bmatrix} \exp(\frac{c\ell\lambda_1}{k}) & 0 \\ 0 & \exp(\frac{c\ell\lambda_2}{k}) \end{bmatrix} \right\|_2 \| \begin{bmatrix} \frac{2cL_w}{nt} \\ \frac{2cr^2L_\theta}{nt} \end{bmatrix} \|_2 \\
&= \frac{\sigma_{\max}(\boldsymbol{\nu})}{\sigma_{\min}(\boldsymbol{\nu})} \sum_{t=t_0+1}^{T} \left\| \begin{bmatrix} \exp(\sum_{k=t+1}^{T} \frac{c\ell\lambda_1}{k}) & 0 \\ 0 & \exp(\sum_{k=t+1}^{T} \frac{c\ell\lambda_2}{k}) \end{bmatrix} \right\|_2 \| \begin{bmatrix} \frac{2cL_w}{nt} \\ \frac{2cr^2L_\theta}{nt} \end{bmatrix} \|_2 \\
&\leq \frac{\sigma_{\max}(\boldsymbol{\nu})}{\sigma_{\min}(\boldsymbol{\nu})} \sum_{t=t_0+1}^{T} \exp(\sum_{k=t+1}^{T} \frac{c\ell(r+1)}{k}) \| \begin{bmatrix} \frac{2cL_w}{nt} \\ \frac{2cr^2L_\theta}{nt} \end{bmatrix} \|_2 \\
&\leq \frac{2crL\sigma_{\max}(\boldsymbol{\nu})}{n\sigma_{\min}(\boldsymbol{\nu})} \sum_{t=t_0+1}^{T} \frac{\exp(\sum_{k=t+1}^{T} \frac{c\ell(r+1)}{k})}{t} \\
&= \frac{2crL\sigma_{\max}(\boldsymbol{\nu}) T^{c\ell(r+1)}}{n\sigma_{\min}(\boldsymbol{\nu})} \sum_{t=t_0+1}^{T} t^{-c\ell(r+1)-1} \\
&\leq \frac{2rL\sigma_{\max}(\boldsymbol{\nu})}{(r+1)\ell n\sigma_{\min}(\boldsymbol{\nu})} (\frac{T}{t_0})^{c\ell(2r+1)} \\
&\leq \frac{12L}{n\ell} (\frac{T}{t_0})^{c\ell(r+1)}.
\end{aligned}$$

We note that assuming $r \geq 1$ we have $\boldsymbol{\nu}$'s condition number $\sigma_{\max}(\boldsymbol{\nu})/\sigma_{\min}(\boldsymbol{\nu}) \leq (\sqrt{2}+1)/(\sqrt{2}-1) \leq 6$. This is because given an eigenvalue λ of $\begin{bmatrix} 1 & 1 \\ r^2 & -r^2/\kappa \end{bmatrix}$ and its corresponding eigenvector $[\nu_1, \nu_2]$ we have $\nu_2 = (\lambda - 1)\nu_1$ and hence the eigenvector aligns with $[1, \lambda - 1]$. Therefore, we can bound the condition number of the following symmetric matrix, because we can consider any vector column along the eigenvector's direction:

$$\begin{bmatrix} 1 & \lambda_1 - 1 \\ \lambda_1 - 1 & (\lambda_1 - 1)(\lambda_2 - 1) \end{bmatrix} = \begin{bmatrix} 1 & \frac{-1-\frac{r^2}{\kappa}+\sqrt{4r^2+(1+r^2/\kappa)^2}}{2} \\ \frac{-1-\frac{r^2}{\kappa}+\sqrt{4r^2+(1+r^2/\kappa)^2}}{2} & -r \end{bmatrix}.$$

Since the above matrix is symmetric, its eigenvalues have the same absolute value as its singular values, and therefore the condition number will be bounded as

$$\begin{aligned}
\frac{\sigma_{\max}(\boldsymbol{\nu})}{\sigma_{\min}(\boldsymbol{\nu})} &\leq \frac{\sqrt{(r-1)^2 + 4(r + (\lambda_1 - 1)^2)} + (r-1)}{\sqrt{(r-1)^2 + 4(r + (\lambda_1 - 1)^2)} - (r-1)} \\
&\leq \frac{\sqrt{(r-1)^2 + 4(r + (r - \frac{r+1}{2})^2)} + (r-1)}{\sqrt{(r-1)^2 + 4(r + (r - \frac{r+1}{2})^2)} - (r-1)} \\
&\leq \frac{\sqrt{(r-1)^2 + 4(r - \frac{r+1}{2})^2} + (r-1)}{\sqrt{(r-1)^2 + 4(r - \frac{r+1}{2})^2} - (r-1)} \\
&= \frac{\sqrt{2(r-1)^2} + (r-1)}{\sqrt{2(r-1)^2} - (r-1)} \\
&= \frac{\sqrt{2}+1}{\sqrt{2}-1}.
\end{aligned}$$

As a result, we showed that conditioned to $\delta_{t_0} = 0$ we will have

$$\mathbb{E}\big[\delta_{w,T} \big| \delta_{t_0} = 0\big] \leq \frac{12L}{n\ell} (\frac{T}{t_0})^{c\ell(r+1)}.$$

Combining the above equation with Lemma 4, we obtain that

$$\forall \mathbf{z}, t_0: \ \ \mathbb{E}\big[|f_{\max}(\mathbf{w}_T;\mathbf{z}) - f_{\max}(\mathbf{w}'_T;\mathbf{z})|\big] \leq \frac{t_0}{n} + \frac{12L^2\sqrt{\kappa^2+1}}{n\ell}\big(\frac{T}{t_0}\big)^{c\ell(r+1)}.$$

The above bound will be approaximately minimized at

$$t_0 = \big(12(r+1)cLL^2\sqrt{\kappa^2+1}\big)^{\frac{1}{(r+1)c\ell+1}} T^{\frac{(r+1)c\ell}{(r+1)c\ell+1}}$$

which leads to the following bound

$$\forall \mathbf{z}: \ \ \mathbb{E}\big[|f_{\max}(\mathbf{w}_T;\mathbf{z}) - f_{\max}(\mathbf{w}'_T;\mathbf{z})|\big] \leq \frac{1+\frac{1}{(r+1)c\ell}}{n}\big(12(r+1)cLL^2\sqrt{\kappa^2+1}\big)^{\frac{1}{(r+1)c\ell+1}} T^{\frac{(r+1)c\ell}{(r+1)c\ell+1}}.$$

The theorem's bound on SGDA updates is then a consequence of Theorem 2.2 in (Hardt et al., 2016).

For the theorem's bound on SGDmax updates, note that $f_{\max}(\mathbf{w};S)$ will be $L\sqrt{\kappa^2+1}$-Lipschitz. Also, Lemma 2 implies that $f_{\max}(\mathbf{w};S)$ will be $\ell(\frac{\kappa}{2}+1)$-smooth in $\mathbf{w}$. Therefore, the result follows from Theorem 3.12 in (Hardt et al., 2016). □

B.9. Proof of Theorem 6

Theorem. *Let minimax cost $0 \leq f(\cdot,\cdot;\mathbf{z}) \leq 1$ be a bounded non-convex non-concave objective which satisfies Assumptions 1 and 2. Then, the SGDA algorithm with vanishing stepsizes $\max\{\alpha_{w,t},\alpha_{\theta,t}\} \leq c/t$ for constant $c > 0$ satisfies the following bound over T iterations:*

$$\epsilon^{\text{mm}}_{\text{gen}}(\text{SGDA}) \leq \frac{1+\frac{1}{\ell c}}{n}\big(2cLL_w\big)^{\frac{1}{\ell c+1}} T^{\frac{\ell c}{\ell c+1}}. \tag{13}$$

Proof. To show this result, we apply Lemma 4. Defining $\delta_t = \sqrt{\|\mathbf{w}_t - \mathbf{w}'_t\|^2 + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}'_t\|^2}$ for the norm difference of parameters learned by SGDA over two datasets S, S' with one different sample, according to the law of total probability we have:

$$\begin{aligned}\mathbb{E}[\delta_{t+1}] &\leq (1-\frac{1}{n})(1+\frac{c\ell}{t})\mathbb{E}[\delta_t] + \frac{1}{n}\big((1+\frac{c\ell}{t})\mathbb{E}[\delta_t] + \frac{2cL}{t}\big)\\ &= (1+\frac{c\ell}{t})\mathbb{E}[\delta_t] + \frac{2cL}{nt}.\end{aligned}$$

As a result, conditioned on $\delta_{t_0} = 0$ we will have

$$\begin{aligned}\mathbb{E}[\delta_T|\delta_{t_0}=0] &\leq \sum_{t=t_0+1}^{T}\prod_{k=t+1}^{T}\{1+\frac{c\ell}{k}\}\frac{2cL}{nt}\\ &\leq \sum_{t=t_0+1}^{T}\prod_{k=t+1}^{T}\{\exp(\frac{c\ell}{k})\}\frac{2cL}{nt}\\ &= \sum_{t=t_0+1}^{T}\exp\big(\sum_{k=t+1}^{T}\frac{c\ell}{k}\big)\frac{2cL}{nt}\\ &\leq \sum_{t=t_0+1}^{T}\exp(c\ell\log(T/t))\frac{2cL}{nt}\\ &= \frac{2cLT^{c\ell}}{n}\sum_{t=t_0+1}^{T}t^{-c\ell-1}\\ &\leq \frac{2L}{n\ell}\big(\frac{T}{t_0}\big)^{c\ell}.\end{aligned}$$

Therefore, Lemma 4 shows that for every t_0 and $\mathbf{z}$:

$$\mathbb{E}\big[|f_{\max}(\mathbf{w}_t;\mathbf{z}) - f_{\max}(\mathbf{w}'_t;\mathbf{z})|\big] \leq \frac{t_0}{n} + \frac{2LL_w}{n\ell}\big(\frac{T}{t_0}\big)^{c\ell}.$$

The above upper-bound will be approximately minimized at

$$t_0 = (2cLL_w)^{\frac{1}{\ell c+1}} T^{\frac{\ell c}{\ell c+1}}.$$

Plugging in the above t_0 to the upper-bound we obtain the following bound for every $\mathbf{z}$:

$$\mathbb{E}\big[|f_{\max}(\mathbf{w}_t;\mathbf{z}) - f_{\max}(\mathbf{w}'_t;\mathbf{z})|\big] \leq \frac{1+\frac{1}{\ell c}}{n}(2cLL_w)^{\frac{1}{\ell c+1}} T^{\frac{\ell c}{\ell c+1}}.$$

The above result combined with Theorem 2.2 from (Hardt et al., 2016) proves the theorem. □

References

Bernhard, P. and Rapaport, A. On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24(8):1163–1181, 1995.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. Convergence rate of o(1/k) for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. *arXiv preprint arXiv:1906.01115*, 2019.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5): 877–898, 1976.