Direct-Search for a Class of Stochastic Min-Max Problems: Supplementary Materials

A ALGORITHMS

We present omitted algorithms. Algorithm 3 depicts the updates for the minimization problem. At each outer iteration of Algorithm 2, a single successful step for the minimization problem is performed. In contrast to standard direct-search algorithms, we do not increase the step size parameter immediately after a successful step, but instead, before the start of the next search for a new successful step, to simplify notation for the upcoming proofs.

```
Algorithm 3: One-Step-Direct-search($f, \mathbf{x}_0, \sigma_0$)
Input: $f$: objective function, with $f_k$ it's estimate at step $k$
       $\mathbf{x}$: initial point
       $\sigma_0$: step size value
$c$: forcing function constant
$\gamma > 1$: step size update parameter
Create the Positive Spanning Set $\mathcal{D}$ for the variables $\mathbf{x}$.
Update $\sigma_1 = \min\{\gamma\sigma_0, \sigma_{\max}\}$ as last update was successful.
for $k$ = 1, ... do
    1. Offspring generation:
    Generate the points
        $\mathbf{x}^i = \mathbf{x} + \sigma_k \mathbf{d}^i, \quad \forall \mathbf{d}^i \in \mathcal{D}.$
    2. Parent Selection:
    Choose $\mathbf{x}' = \arg\min_i f_k(\mathbf{x}^i)$.
    3. Sufficient Decrease: if $f_k(\mathbf{x}') < f_k(\mathbf{x}) - \rho(\sigma_k)$ then
        (Iteration is successful)
        return $\mathbf{x}', \sigma_k$.
    else
        (Iteration is unsuccessful)
        Decrease step size $\sigma_{k+1} = \gamma^{-1}\sigma_k$.
    end
end
```

B PROOFS OF SECTION 4

B.1 Proof of Lemma 1

Proof. The result follows by applying Holder's inequality.

$$\mathbb{E}\left[\frac{1_{J_k^c}|F_k^0 - f(\mathbf{X}_k)|}{l_f \Sigma_k^2} \mid \mathcal{F}_{k-1}\right] \leq \left(\mathbb{E}[1_{J_k^c}|\mathcal{F}_{k-1}]\right)^{1/2} \left(\mathbb{E}\left[\frac{|F_k^0 - f(\mathbf{X}_k)|^2}{l_f^2 \Sigma_k^4}\right]\right)^{1/2}.$$

By Assumption 1, it holds that $\left(\mathbb{E}\left[\frac{|F_k^0 - f(\mathbf{x}_k)|^2}{l_f^2 \Sigma_k^4}\right]\right)^{1/2} \leq 1$ and the result follows. Following the same steps, the second inequality of the Lemma holds as well.

□

B.2 Proof of Theorem 2

Proof. We begin by taking separate cases according to if the estimates are accurate or not and if the steps of Algorithm 1 are successful or not. We use 1_{Succ_k} to denote the event that step k is successful.

Case 1: Accurate estimates.

- **Successful step.**

  At a successful step with accurate estimates we have that:

$$\begin{aligned} &1_{\text{Succ}_k} 1_{J_k}(f(\mathbf{X}_{k+1}) - f(\mathbf{X}_k)) \\ &\quad = 1_{\text{Succ}_k} 1_{J_k}(f(\mathbf{X}_{k+1}) - f_k(\mathbf{X}_{k+1}) + f_k(\mathbf{X}_{k+1}) - f_k(\mathbf{X}_k) + f_k(\mathbf{X}_k) - f(\mathbf{X}_k)) \\ &\quad \le 1_{\text{Succ}_k} 1_{J_k}(-(c - 2\epsilon_f)\Sigma_k^2). \end{aligned}$$

  Therefore

$$\begin{aligned} &1_{\text{Succ}_k} 1_{J_k}(\Phi_{k+1} - \Phi_k) \\ &\quad = 1_{\text{Succ}_k} 1_{J_k}(v(f(\mathbf{X}_{k+1}) - f(\mathbf{X}_k)) + (1-v)\Sigma_{k+1}^2 - (1-v)\Sigma_k^2) \\ &\quad \le 1_{\text{Succ}_k} 1_{J_k}(-v(c - 2\epsilon_f)\Sigma_k^2 + (1-v)(\gamma^2 - 1)\Sigma_k^2). \end{aligned}$$

- **Unsuccessful step.**

$$\begin{aligned} 1_{\text{Succ}_k^c} 1_{J_k}(\Phi_{k+1} - \Phi_k) &= 1_{\text{Succ}_k^c} 1_{J_k}((1-v)\Sigma_{k+1}^2 - (1-v)\Sigma_k^2) \\ &= 1_{\text{Succ}_k^c} 1_{J_k}(-(1-v)(1 - \frac{1}{\gamma^2})\Sigma_k^2). \end{aligned}$$

Combining the above results and given that

$$\frac{v}{1-v} \ge \frac{1}{c - 2\epsilon_f}(\gamma^2 - \frac{1}{\gamma^2}) \implies -v(c - 2\epsilon_f) + (1-v)(\gamma^2 - 1) \le -(1-v)(1 - \frac{1}{\gamma^2}),$$

in the case of accurate estimates we have

$$\mathbb{E}[1_{J_k}(\Phi_{k+1} - \Phi_k) \mid \mathcal{F}_{k-1}] \le -p_f(1-v)(1 - \frac{1}{\gamma^2})\Sigma_k^2. \tag{11}$$

Case 2: Inaccurate estimates.

- **Successful step.**

$$\begin{aligned} 1_{\text{Succ}_k} 1_{J_k^c}(\Phi_{k+1} - \Phi_k) &= 1_{\text{Succ}_k} 1_{J_k^c}(v(f(\mathbf{X}_{k+1}) - f(\mathbf{X}_k)) + (1-v)\Sigma_{k+1}^2 - (1-v)\Sigma_k^2) \\ &= 1_{\text{Succ}_k} 1_{J_k^c}(v(f(\mathbf{X}_{k+1}) - f_k(\mathbf{X}_{k+1}) + f_k(\mathbf{X}_{k+1}) - f_k(\mathbf{X}_k) \\ &\qquad + f_k(\mathbf{X}_k) - f(\mathbf{X}_k)) + (1-v)\Sigma_{k+1}^2 - (1-v)\Sigma_k^2) \\ &\le 1_{\text{Succ}_k} 1_{J_k^c}(-vc\Sigma_k^2 + v|f(\mathbf{X}_{k+1}) - f_k(\mathbf{X}_{k+1})| + v|f(\mathbf{X}_k) - f_k(\mathbf{X}_k)| \\ &\qquad - (1-v)(\gamma^2 - 1)\Sigma_k^2), \end{aligned}$$

  where we will later bound terms $|f(\mathbf{X}_{k+1}) - f_k(\mathbf{X}_{k+1})|, |f(\mathbf{X}_k) - f_k(\mathbf{X}_k)|$ using Lemma 1.

- **Unsuccessful step.**

  As before:

$$\begin{aligned}1_{\mathrm{Succ}_k^c}1_{J_k^c}(\Phi_{k+1}-\Phi_k) &= 1_{\mathrm{Succ}_k^c}1_{J_k^c}((1-v)\Sigma_{k+1}^2-(1-v)\Sigma_k^2)\\ &= 1_{\mathrm{Succ}_k^c}1_{J_k^c}(-(1-v)(1-\frac{1}{\gamma^2})\Sigma_k^2).\end{aligned}$$

In total for inaccurate estimates and by using Assumption 1 and Lemma 1

$$\mathbb{E}[1_{J_k^c}(\Phi_{k+1}-\Phi_k)\mid\mathcal{F}_{k-1}]\leq 2v(1-p_f)^{1/2}l_f\Sigma_k^2. \tag{12}$$

Finally, integrating both successful and unsuccessful iterations

$$\begin{aligned}\mathbb{E}[\Phi_{k+1}-\Phi_k\mid\mathcal{F}_{k-1}] &\leq -p_f(1-v)(1-\frac{1}{\gamma^2})\Sigma_k^2+2v(1-p_f)^{1/2}l_f\Sigma_k^2\\ &\leq -p_f(1-v)(1-\frac{1}{\gamma^2})\frac{\Sigma_k^2}{2},\end{aligned}$$

for our requirement of $\frac{p_f}{\sqrt{1-p_f}}\geq\frac{4vl_f}{(1-v)(1-\gamma^{-2})}$.

□

B.3 Proof of Lemma 4

Proof. Similar to Conn et al. (2009), for an unsuccessful step with accurate estimates, we have that for some $\mathbf{d}_k\in\mathcal{D}$

$$\kappa(\mathcal{D})\|\nabla f(\mathbf{x}_k)\|\|\mathbf{d}_k\|\leq-\nabla f(\mathbf{x}_k)^\top\mathbf{d}_k. \tag{13}$$

By the mean value theorem, for some $\eta_k\in[0,1]$,

$$f(\mathbf{x}_k+\sigma_k\mathbf{d}_k)-f(\mathbf{x}_k)=\sigma_k\nabla f(\mathbf{x}_k+\eta_k\sigma_k\mathbf{d}_k)^\top\mathbf{d}_k.$$

Since k is the index of an unsuccessful iteration,

$$f_k(\mathbf{x}_k+\sigma_k\mathbf{d}_k)-f_k(\mathbf{x}_k)+\rho(\sigma_k)\geq 0$$

and since estimates are accurate

$$\begin{aligned}f(\mathbf{x}_k+\sigma_k\mathbf{d}_k)-f(\mathbf{x}_k) &= f(\mathbf{x}_k+\sigma_k\mathbf{d}_k)-f_k(\mathbf{x}_k+\sigma_k\mathbf{d}_k)\\ &\qquad+f_k(\mathbf{x}_k+\sigma_k\mathbf{d}_k)-f_k(\mathbf{x}_k)+f_k(\mathbf{x}_k)-f(\mathbf{x}_k)\\ &\geq -\epsilon_f\sigma_k^2-\rho(\sigma_k)-\epsilon_f\sigma_k^2\\ &= -(c+2\epsilon_f)\sigma_k^2.\end{aligned}$$

Combining the above equations,

$$\begin{aligned}&\sigma_k\nabla f(\mathbf{x}_k+\eta_k\sigma_k\mathbf{d}_k)^\top\mathbf{d}_k+(c+2\epsilon_f)\sigma_k^2\geq 0\\ \Longrightarrow\ &\nabla f(\mathbf{x}_k+\eta_k\sigma_k\mathbf{d}_k)^\top\mathbf{d}_k+(c+2\epsilon_f)\sigma_k\geq 0\\ \Longrightarrow\ &-\nabla f(\mathbf{x}_k)^\top\mathbf{d}_k\leq(\nabla f(\mathbf{x}_k+\eta_k\sigma_k\mathbf{d}_k)-\nabla f(\mathbf{x}_k))^\top\mathbf{d}_k+(c+2\epsilon_f)\sigma_k,\end{aligned} \tag{14}$$

where in the last inequality, we subtracted $\nabla f(\mathbf{x}_k)^\top \mathbf{d}_k$ from both sides.

Finally, Eq. (13) implies

$$\begin{aligned}
&\kappa(\mathcal{D})\|\nabla f(\mathbf{x}_k)\|\|\mathbf{d}_k\| \leq (\nabla f(\mathbf{x}_k + \eta_k\sigma_k\mathbf{d}_k) - \nabla f(\mathbf{x}_k))^\top \mathbf{d}_k + (c + 2\epsilon_f)\sigma_k \\
\implies &\kappa(\mathcal{D})\|\nabla f(\mathbf{x}_k)\| \leq \|\nabla f(\mathbf{x}_k + \eta_k\sigma_k\mathbf{d}_k) - \nabla f(\mathbf{x}_k)\| + (c + 2\epsilon_f)\sigma_k \\
&\leq \frac{L}{2}\sigma_k + (c + 2\epsilon_f)\sigma_k.
\end{aligned}$$

□

B.4 Proof of Theorem 5

Proof. By Theorem 2 and Lemma 4.10 from Paquette and Scheinberg (2018) we get that Assumption 2 is satisfied, for $\Sigma_\epsilon = C\epsilon$. Then by an application of Theorem 3 we get

$$\mathbb{E}[T_\epsilon] \leq \frac{p_f}{2p_f - 1} \frac{v(f(\mathbf{X}_0) - f^*) + (1-v)\Sigma_0^2}{p_f(1-v)(1 - \frac{1}{\gamma^2})\frac{C^2\epsilon^2}{2}}.$$

The result follows.

□

B.5 Proof of Theorem 6

We will also use the additional result holding for any function with Lipschitz-continuous gradients.

Lemma 8. *Let $f : \mathbf{x} \in \mathbb{R}^n \to \mathbb{R}$ be a continuous differentiable function with Lipschitz continuous gradient with a constant L and a minimum value achieved for $\mathbf{x}^*$. Then*

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2. \tag{15}$$

Proof. By smoothness and for $\mathbf{y} = \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})$ we have

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{x}^*) &\geq f(\mathbf{x}) - f(\mathbf{y}) \\
&\geq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle - \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\
&= \frac{1}{L}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2 \\
&= \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2.
\end{aligned}$$

□

We can now proceed with the proof of Theorem 6.

Proof. We note that for the conditions on the constants c, v and p_f, requirements of Theorem 2 are also satisfied. We define as $T_i = \inf\{k \geq 0 : f(\mathbf{X}_k) - f^* \leq \frac{f(\mathbf{X}_0) - f^*}{2^i}\}$, with $T_0 = 0$. We will also use the random variable $\Lambda_i = T_i - T_{i-1}$.

We will assume without loss of generality that

$$\Sigma_0^2 \leq \frac{9\gamma^2}{c}(f(\mathbf{X}_0) - f^*) \triangleq A(f(\mathbf{X}_0) - f^*), \tag{16}$$

for $A = \frac{9\gamma^2}{c}$. We apply Theorem 3. Given that $f(\mathbf{X}_{T_{i-1}}) - f^* \leq \frac{f(\mathbf{X}_{T_0})-f^*}{2^{i-1}}$, and that $f(\mathbf{X}_k) - f^* > \frac{f(\mathbf{X}_{T_0})-f^*}{2^i}$ for $k \in [T_{i-1}, T_i)$ (possibly an empty set), Lemma 4 and the Definition 5, then for step sizes $\Sigma_k^2 \leq C^2\mu\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}$ and accurate estimates, steps are successful. Then by Theorem 3 for an application of the results from Theorem 2 as before, we have

$$\begin{aligned}
\mathbb{E}[\Lambda_i \mid \mathcal{F}_{T_{i-1}-1}] &\leq \frac{p_f}{2p_f - 1}\frac{v(f(\mathbf{X}_{T_{i-1}}) - f^*) + (1-v)\Sigma_{T_{i-1}}^2}{p_f(1-v)(1-\gamma^{-2})\frac{1}{2}C^2\mu\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}} \\
&= \frac{2}{(2p_f-1)(1-\gamma^{-2})C^2\mu}\left(\frac{v}{1-v}\frac{f(\mathbf{X}_{T_{i-1}}) - f^*}{\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}} + \frac{\Sigma_{T_{i-1}}^2}{\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}}\right) \\
&\leq \frac{2}{(2p_f-1)(1-\gamma^{-2})C^2\mu}\left(\frac{v}{1-v}\frac{\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}}{\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}} + \frac{\Sigma_{T_{i-1}}^2}{\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}}\right) \\
&= \frac{2}{(2p_f-1)(1-\gamma^{-2})C^2\mu}\left(\frac{v}{1-v} + \frac{\Sigma_{T_{i-1}}^2}{\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}}\right)
\end{aligned} \tag{17}$$

We will further show with induction that $\mathbb{E}[\Sigma_{T_i}^2] \leq A\frac{f(\mathbf{X}_0)-f^*}{2^i}$. As a result

$$\begin{aligned}
\mathbb{E}[\Lambda_i] &\leq \frac{2}{(2p_f-1)(1-\gamma^{-2})C^2\mu}\left(\frac{v}{1-v} + \frac{\mathbb{E}[\Sigma_{T_{i-1}}^2]}{\frac{f(\mathbf{X}_0)-f^*}{2^{i-1}}}\right) \\
&\leq \frac{2}{(2p_f-1)(1-\gamma^{-2})C^2\mu}\left(\frac{v}{1-v} + A\right).
\end{aligned} \tag{18}$$

The final complexity will be:

$$\begin{aligned}
\mathbb{E}[T_{\lceil \log \frac{2L(f(\mathbf{X}_0)-f^*)}{\epsilon^2}\rceil}] &= \mathbb{E}[\Lambda_1 + \Lambda_2 + \cdots + \Lambda_{\lceil \log \frac{2L(f(\mathbf{X}_0)-f^*)}{\epsilon^2}\rceil}] \\
&\leq \frac{2}{(2p_f-1)(1-\gamma^{-2})C^2\mu}\left(\frac{v}{1-v} + A\right)\lceil \log\left(\frac{2L(f(\mathbf{X}_0)-f^*)}{\epsilon^2}\right)\rceil.
\end{aligned} \tag{19}$$

Getting $\|\nabla f(\mathbf{x}_0)\|^2 \leq 2L(f(\mathbf{x}_0) - f^*)$ from Lemma 8, the result follows.

It remains to show the result that $\mathbb{E}[\Sigma_{T_i}^2] \leq A\frac{f(\mathbf{X}_0)-f^*}{2^i}$. By assumption, as aforementioned, it holds for T_0. We then assume that it holds for T_{i-1} and show that it also holds for T_i. For each T_i, the last step $T_i - 1$ was a successful one as the parameter $\mathbf{X}$ was updated to satisfy the goal $f(\mathbf{X}_{T_i}) - f^* \leq \frac{f(\mathbf{X}_0)-f^*}{2^i}$. As in Theorem 2 we differentiate between the events of this step being accurate or not.

Since we have a successful step $f_{T_i-1}(\mathbf{X}_{T_i}) - f_{T_i-1}(\mathbf{X}_{T_i-1}) \leq -c\Sigma_{T_i-1}^2$. Then

$$\begin{aligned}
f(\mathbf{X}_{T_i}) - f(\mathbf{X}_{T_i-1}) = f(\mathbf{X}_{T_i}) - f_{T_i-1}(\mathbf{X}_{T_i}) + \\
f_{T_i-1}(\mathbf{X}_{T_i}) - f_{T_i-1}(\mathbf{X}_{T_i-1}) + f_{T_i-1}(\mathbf{X}_{T_i-1}) - f(\mathbf{X}_{T_i-1})
\end{aligned} \tag{20}$$

We denote with p_{Acc} the probability of this last step being accurate. Note that this is not the same as p_f as we are conditioning on a successful step. Then we distinguish the two cases.

- **Accurate estimates.**

  By Assumption 3 we get that

$$1_{\text{Acc}}(f(\mathbf{X}_{T_i}) - f(\mathbf{X}_{T_i-1})) \leq 1_{\text{Acc}}(-(c - 2\epsilon_f)\Sigma_{T_i-1}^2). \tag{21}$$

- **Inaccurate.**

  In this case, similarly to the proof of Lemma 1 we get

$$\mathbb{E}[1^c_{\text{Acc}}(f(\mathbf{X}_{T_i}) - f(\mathbf{X}_{T_i-1})) \mid \mathcal{F}_{T_i-2}] \leq -(1-p_{\text{Acc}})c\Sigma^2_{T_i-1} + 2\sqrt{1-p_{\text{Acc}}}l_f\Sigma^2_{T_i-1}. \tag{22}$$

Combining the above cases, we get

$$\begin{aligned}
\mathbb{E}[f(\mathbf{X}_{T_i}) - f(\mathbf{X}_{T_i-1}) \mid \mathcal{F}_{T_i-2}] &\leq -p_{\text{Acc}}(c-2\epsilon_f)\Sigma^2_{T_i-1} - (1-p_{\text{Acc}})c\Sigma^2_{T_i-1} + 2\sqrt{1-p_{\text{Acc}}}l_f\Sigma^2_{T_i-1} \\
&\leq -c\Sigma^2_{T_i-1}(1 - \frac{p_{\text{Acc}}}{2} - \sqrt{\frac{1-p_{\text{Acc}}}{2}}), \text{ for } c > \max\{4\epsilon_f, 2\sqrt{2}l_f\} \\
&\leq -\frac{c}{4}\Sigma^2_{T_i-1} \\
\implies \Sigma^2_{T_i-1} &\leq \frac{4\mathbb{E}[f(\mathbf{X}_{T_i-1}) - f^*|\mathcal{F}_{T_i-2}]}{c}.
\end{aligned} \tag{23}$$

Furthermore, by Theorem 2 we get that

$$\begin{aligned}
\mathbb{E}[\Phi_{T_i-1} \mid \mathcal{F}_{T_{i-1}-1}] &\leq \Phi_{T_{i-1}} \\
\mathbb{E}[v(f(\mathbf{X}_{T_i-1}) - f^*) \mid \mathcal{F}_{T_{i-1}-1}] &\leq v(f(\mathbf{X}_{T_{i-1}}) - f^*) + (1-v)\Sigma^2_{T_{i-1}} \\
\mathbb{E}[f(\mathbf{X}_{T_i-1}) - f^*] &\leq \frac{f(\mathbf{X}_0) - f^*}{2^{i-1}} + \frac{(1-v)}{v}\mathbb{E}[\Sigma^2_{T_{i-1}}] \\
&\leq \frac{f(\mathbf{X}_0) - f^*}{2^{i-1}} + \frac{(1-v)}{v}A\frac{f(\mathbf{X}_0) - f^*}{2^{i-1}} \\
&\leq \frac{f(\mathbf{X}_0) - f^*}{2^i}2(1 + \frac{1-v}{v}A).
\end{aligned} \tag{24}$$

By combining (23) and (24) and using the law of iterated expectation we have that

$$\begin{aligned}
\mathbb{E}[\Sigma^2_{T_i}] &= \mathbb{E}[\gamma^2\Sigma^2_{T_i-1}] \\
&\leq \frac{f(\mathbf{X}_0) - f^*}{2^i}\frac{8\gamma^2}{c}(1 + \frac{1-v}{v}A) \\
&\leq A\frac{f(\mathbf{X}_0) - f^*}{2^i},
\end{aligned} \tag{25}$$

for $\frac{v}{1-v} \geq \frac{72\gamma^2}{c}$ and $A = \frac{9\gamma^2}{c}$. The proof is complete.

□

C PROOFS OF SECTION 5

We first present some additional results, required for our proof.

From Karimi and Schmidt (2015), for a function that satisfies the PL condition, it additionally satisfies the Quadratic Growth (GQ) condition.

Lemma 9. *A differentiable function f that satisfies the PL condition with parameter μ, also satisfies the QG condition with parameter 4μ:*

$$f(\mathbf{x}) - f^* \geq 2\mu\|\mathbf{x}^* - \mathbf{x}\|^2,$$

where $\mathbf{x}^*$ belongs to the solution set $\mathcal{X}^*$.

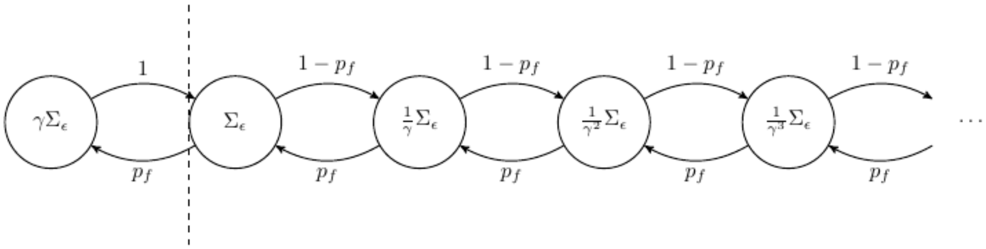
1
$1-p_f$
$1-p_f$
$1-p_f$
$1-p_f$
$\gamma\Sigma_\epsilon$
Σ_ϵ
$\frac{1}{\gamma}\Sigma_\epsilon$
$\frac{1}{\gamma^2}\Sigma_\epsilon$
$\frac{1}{\gamma^3}\Sigma_\epsilon$
$\cdots$
p_f
p_f
p_f
p_f
p_f

Figure 3: Worst case scenario for step sizes. Ignoring steps for $\Sigma > \Sigma_\epsilon$, it corresponds to a biased reflected random walk. The dotted line indices the barrier at position 0, indicating a step size of Σ_ϵ.

Based on the previous Lemma, we can easily prove the following result.

Lemma 10. *Let a differentiable μ-PL function f and also $\mathbf{x}^* \in \arg\min_{\mathbf{x}} f(\mathbf{x})$. If we know that $\|\nabla f(\mathbf{x})\| \leq \epsilon$ then:*

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{1}{2\mu}\epsilon.$$

Proof. By Lemma 9 and the definition of the PL condition we have that:

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \sqrt{\frac{1}{2\mu}(f(\mathbf{x}) - f^*)} \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x})\| \leq \frac{1}{2\mu}\epsilon.$$

□

Lemma 11. *(Lemma A.3 from Nouiehed et al. (2019)) Assume that $-f(\mathbf{x}, \mathbf{y})$ for a specific $\mathbf{x}$, is a class of μ-PL functions in $\mathbf{y}$. Define the set of optimal solutions $\mathbb{Y}(\mathbf{x}) = \arg\max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. Then for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1^* \in \mathbb{Y}(\mathbf{x}_1), \mathbf{y}_2^* \in \mathbb{Y}(\mathbf{x}_2)$ it holds that:*

$$\|\mathbf{y}_1^* - \mathbf{y}_2^*\| \leq L_{xy}\|\mathbf{x}_1 - \mathbf{x}_2\|,$$

where we denote with $L_{xy} = \frac{L_{12}}{2\mu}$.

Next, we will need to establish a lower bound on the step size Σ. In the deterministic case, Lemma 4 establishes such a lower bound for unsuccessful steps, guaranteeing that if $\Sigma = \Sigma_\epsilon$, then $\|\nabla f(\mathbf{x})\| \leq \epsilon$. However, in the stochastic case, inaccurate steps may occur. We want to ensure a lower bound on the step size parameter with high probability.

To do so, we will consider the worst-case scenario where step sizes get as small as possible. This corresponds to the case where for all step sizes $\Sigma > \Sigma_\epsilon$, unsuccessful steps occur. So do all of the inaccurate estimates, with probability $1 - p_f$. For convenience, we will ignore steps above the value Σ_ϵ since we only require a bound. This corresponds to a random walk with a reflection barrier at position 0 (which corresponds to the step size Σ_ϵ) and an increment probability $1 - p_f$, where p_f is the probability of accurate estimates. We, therefore, use the following Lemma to get a probabilistic lower bound on the step sizes.

Lemma 12. *Let a random walk starting at position 0, with a reflection barrier at position 0 and a transition probability matrix*

$$\begin{bmatrix} p_f & 1-p_f & & & \\ p_f & 0 & 1-p_f & & \\ & p_f & 0 & 1-p_f & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

for $p_f > \frac{1}{2}$. Then for $k \geq \frac{\log(1-e^{\frac{1}{n}\log(\delta)})}{\log\left(\frac{1-p_f}{p_f}\right)} - 1$, the random walk of length n, stays confined within the space $[0, k]$ with a probability at least $\delta > 0$.

Proof. Let a random walk $S_n = \max\{S_{n-1} + X_n, 0\}$, with $S_0 = 0$ and $\mathsf{P}(X_n = 1) = 1 - p_f$, $\mathsf{P}(X_n = -1) = p_f$, for $p_f > \frac{1}{2}$. The probability that the random walk stays until position k, $\mathsf{P}(S_i \leq k, \forall i \leq n)$, is bounded below by the probability of n randomly chosen points from the stationary distribution to be at positions lower or equal to k.

Let us denote with $p_{i,n}$ the probability that the random walk is at position i after n total steps. We first prove by induction that

$$p_{i,n} \geq \frac{p_f}{1 - p_f} p_{i+1,n}. \tag{26}$$

It obviously holds for $n = 0$, as $p_{0,0} = 1$ and $p_{i,0} = 0$, $\forall i \geq 1$. Assume that it holds for n. As shown in Fig. 3, with probability $(1 - p_f)$, position i is incremented, therefore for $i \geq 1$

$$\begin{aligned}
p_{i,n+1} &= p_{i-1,n}(1 - p_f) + p_{i+1,n} p_f \\
&\geq \frac{p_f}{1 - p_f} p_{i,n}(1 - p_f) + \frac{p_f}{1 - p_f} p_{i+2,n} p_f, \text{ by induction} \\
&= \frac{p_f}{1 - p_f} (p_{i,n}(1 - p_f) + p_{i+2,n} p_f) \\
&= \frac{p_f}{1 - p_f} p_{i+1,n+1}
\end{aligned}$$

and for $i = 0$

$$\begin{aligned}
p_{0,n+1} &= p_{0,n} p_f + p_{1,n} p_f \\
&\geq p_{0,n} p_f + \frac{p_f}{1 - p_f} p_{2,n} p_f, \text{ by induction} \\
&= \frac{p_f}{1 - p_f} (p_{0,n}(1 - p_f) + p_{2,n} p_f) \\
&= \frac{p_f}{1 - p_f} p_{1,n+1}.
\end{aligned}$$

Let us now consider the probability that the random walk resides in the first k positions. Then:

$$\begin{aligned}
\sum_{i=0}^{k} p_{i,n+1} &= p_{0,n} p_f + p_{1,n} p_f + \sum_{i=1}^{k} (p_{i-1,n}(1 - p_f) + p_{i+1,n} p_f) \\
&= \sum_{i=0}^{k} p_{i,n} - p_{k,n}(1 - p_f) + p_{k+1,n} p_f \\
&\leq \sum_{i=0}^{k} p_{i,n}, \text{ by (26)},
\end{aligned}$$

where the equality in the second line is due to the terms telescoping in the sum in the first line.

As a result, we can lower bound the probability $\sum_{i=0}^{k} p_{i,n}$ with the corresponding one for $n \to \infty$, which corresponds to a stationary distribution. Also

$$\begin{aligned}
\mathsf{P}(S_i \leq k \mid S_{i-1} \leq k) &= \mathsf{P}(S_i \leq k \mid S_{i-1} = k)\, \mathsf{P}(S_{i-1} = k) + \mathsf{P}(S_i \leq k \mid S_{i-1} < k)\, \mathsf{P}(S_{i-1} < k) \\
&= p_f\, \mathsf{P}(S_{i-1} = k) + \mathsf{P}(S_{i-1} < k) \geq p_f
\end{aligned}$$

and

$$\begin{aligned}\mathsf{P}(S_i \le k \mid S_{i-1} > k) &= \mathsf{P}(S_i \le k \mid S_{i-1} = k+1)\,\mathsf{P}(S_{i-1} = k+1)\\ &\qquad + \mathsf{P}(S_i \le k \mid S_{i-1} > k+1)\,\mathsf{P}(S_{i-1} > k+1)\\ &= p_f\,\mathsf{P}(S_{i-1} = k+1) + 0\,\mathsf{P}(S_{i-1} > k+1)\\ &\le p_f \ \le \mathsf{P}(S_i \le k \mid S_{i-1} \le k).\end{aligned}$$

As a result

$$\begin{aligned}\mathsf{P}(S_i \le k) = \mathsf{P}(S_i \le k \mid S_{i-1} \le k)\,\mathsf{P}(S_{i-1} \le k) + \mathsf{P}(S_i \le k \mid S_{i-1} > k)\,\mathsf{P}(S_{i-1} > k)\\ \le \mathsf{P}(S_i \le k | S_{i-1} \le k).\end{aligned} \tag{27}$$

The probability of a random walk of length n to stay between the first $k \ge 0$ positions is thus

$$\begin{aligned}\mathsf{P}(S_i \le k,\, \forall i \le n) &= \mathsf{P}(S_0 \le k) \prod_{i=1}^{n} \mathsf{P}(S_i \le k \mid S_j \le k,\, \forall j \in [0, i-1])\\ &= \prod_{i=1}^{n} \mathsf{P}(S_i \le k \mid S_{i-1} \le k),\ \text{with } \mathsf{P}(S_0 \le k) = 1,\, \forall k \ge 0\\ &\ge \prod_{i=1}^{n} \mathsf{P}(S_i \le k),\ \text{by Eq. (27)}\\ &= \prod_{j=1}^{n} \left(\sum_{i=0}^{k} p_{i,j}\right)\\ &\ge \left(\sum_{i=0}^{k} \pi_i\right)^{n},\end{aligned}$$

where π_i denotes the stationary probability of the random walk for $i \in \mathbb{N}$. From the recursive relation, we get $\pi_i = (\frac{p_f}{1-p_f})\pi_{i+1}$, which means $\pi_i = (\frac{1-p_f}{p_f})^i \pi_0$.

We now calculate the probability $\pi_{\le k}$ of a randomly chosen point to be part of the first k positions for the stationary distribution

$$\pi_{\le k} = \sum_{i=0}^{k} \pi_i = \frac{\pi_0 \sum_{i=0}^{k} (\frac{1-p_f}{p_f})^i}{\pi_0 \sum_{i=0}^{\infty} (\frac{1-p_f}{p_f})^i} = 1 - \left(\frac{1-p_f}{p_f}\right)^{k+1}.$$

Thus the required probability must be lower bounded by

$$\begin{aligned}\pi_{\le k}^{n} \ge \delta \implies \left(1 - \left(\frac{1-p_f}{p_f}\right)^{k+1}\right)^{n} &\ge \delta\\ \log\left(1 - \left(\frac{1-p_f}{p_f}\right)^{k+1}\right) &\ge \frac{1}{n}\log(\delta)\\ \left(\frac{1-p_f}{p_f}\right)^{k+1} &\le 1 - e^{\frac{1}{n}\log(\delta)}\\ k &\ge \frac{\log\left(1 - e^{\frac{1}{n}\log(\delta)}\right)}{\log\left(\frac{1-p_f}{p_f}\right)} - 1,\end{aligned} \tag{28}$$

where for the last step we used the fact that $\log\left(\frac{1-p_f}{p_f}\right) < 0$ since $p_f > \frac{1}{2}$ implies $\frac{1-p_f}{p_f} < 1$. □

We can now move to the proof of Theorem 7.

Proof. We denote with c_x the constant used for the sufficient decrease condition of the min problem. We first prove the deterministic case. In the deterministic case, Theorems 5 and 6 hold deterministically, meaning that we can reduce the norm of the gradient below a threshold ϵ for the nonconvex case in $\mathcal{O}(\epsilon^{-2})$ iterations and for the case that the function satisfies our PL condition in $\mathcal{O}(\log(\epsilon^{-1}))$ iterations.

At each step, the max problem is solved almost exactly, which is guaranteed by Theorem 6 and Algorithm 1. Then

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_t)\| \leq \epsilon^{\max},$$

for an accuracy $\epsilon^{\max}$ to be specified later. In the proof, we will show that for a particular choice of a forcing function constant, the improvement on the minimization problem is better than possible deterioration caused by the updates of the max problem. By Assumption 3 of Lipschitz continuity

$$\begin{aligned} \|\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_t)\| &\leq L_{12}\|\mathbf{x}_t - \mathbf{x}_{t-1}\| = L_{12}\sigma_t \\ \implies \|\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)\| &\leq L_{12}\sigma_t + \epsilon^{\max}, \end{aligned} \tag{29}$$

for a successful update. Here σ_t is used to denote the step size used for the minimization step throughout Algorithms 2 and 3. We note that σ_t always belongs to a successful step, by the notation used in Algorithm 2. Also by triangle inequality we have that (let $\mathbf{y}_t^*$ and $\mathbf{y}_{t+1}^*$ belong to the optimal solution sets at iterations t and $t+1$ respectively)

$$\begin{aligned} \|\mathbf{y}_{t+1} - \mathbf{y}_t\| &= \|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^* + \mathbf{y}_{t+1}^* - \mathbf{y}_t^* + \mathbf{y}_t^* - \mathbf{y}_t\| \\ &\leq \|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^*\| + \|\mathbf{y}_{t+1}^* - \mathbf{y}_t^*\| + \|\mathbf{y}_t^* - \mathbf{y}_t\|. \end{aligned} \tag{30}$$

By Lemma 11 we have that $\|\mathbf{y}_{t+1}^* - \mathbf{y}_t^*\| \leq L_{xy}\sigma_t$, since $\mathbf{y}_{t+1}^* \in \mathbb{Y}(\mathbf{x}_t)$ and $\mathbf{y}_t^* \in \mathbb{Y}(\mathbf{x}_{t-1})$ (we remind that $\mathbb{Y}(\mathbf{x}) = \arg\max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$). Also, as a consequence of Definition 5 and Lemma 10 we have that both

$$\|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^*\| \leq \frac{\epsilon^{\max}}{2\mu} \quad \text{and} \quad \|\mathbf{y}_t - \mathbf{y}_t^*\| \leq \frac{\epsilon^{\max}}{2\mu}.$$

As a result

$$\|\mathbf{y}_{t+1} - \mathbf{y}_t\| \leq \frac{\epsilon^{\max}}{\mu} + L_{xy}\sigma_t. \tag{31}$$

Finally, for a successful update of the Algorithm 3 we have

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}_{t+1}) - f(\mathbf{x}_t, \mathbf{y}_t) &\leq \langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_{t+1} - \mathbf{x}_t \rangle + \frac{L_{22}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\leq (L_{12}\sigma_t + \epsilon^{\max})(L_{xy}\sigma_t + \frac{\epsilon^{\max}}{\mu}) \\ &\quad + \frac{L_{22}}{2}(L_{xy}\sigma_t + \frac{\epsilon^{\max}}{\mu})^2 \\ &= D_1\sigma_t^2 + D_2\sigma_t\epsilon^{\max} + D_3(\epsilon^{\max})^2, \end{aligned} \tag{32}$$

for $D_1 \triangleq L_{12}L_{xy} + \frac{L_{22}}{2}L_{xy}^2$, $D_2 = \frac{L_{12}}{\mu} + L_{xy} + \frac{L_{22}L_{xy}}{\mu}$ and $D_3 = \frac{1}{\mu}(1 + \frac{L_{22}}{2\mu})$. During the updates of the minimization problem we have that

$$\sigma \geq \sigma_{\min} = \sigma_\epsilon, \text{ with } \sigma_\epsilon = C\epsilon.$$

Here C, which is defined in Lemma 4, entails the constants for the min problem. We want to ensure that

$$f(\mathbf{x}_t, \mathbf{y}_{t+1}) - f(\mathbf{x}_{t-1}, \mathbf{y}_t) < -K\sigma_t^2, \tag{33}$$

for some $K > 0$ and for $\sigma_t \geq \sigma_{\text{min}}$, to then apply Theorem 2, for f^* the minimum of f at each $\mathbf{y}_t$. Taking also into account our sufficient decrease condition, we want to make sure that the following holds for the polynomial p

$$p(\sigma_t) \triangleq K\sigma_t^2 - c_x\sigma_t^2 + D_1\sigma_t^2 + D_2\sigma_t\epsilon^{\max} + D_3(\epsilon^{\max})^2 \leq 0 \tag{34}$$

for every $\sigma_t \geq \sigma_{\text{min}}$. To establish this we just need to ensure that for the quadratic with negative second degree coefficient (for $c_x > D_1 + K$) the maximum occurs at position:

$$\frac{D_2\epsilon^{\max}}{2(c_x - K - D_1)} \leq C\epsilon \implies \epsilon^{\max} \leq \epsilon\frac{2C(c_x - K - D_1)}{D_2} \tag{35}$$

and also that

$$\begin{aligned} p(C\epsilon) \leq 0 \iff \\ (-c_x + K + D_1)C^2\epsilon^2 + D_2C\epsilon\epsilon^{\max} + D_3(\epsilon^{\max})^2 \leq 0. \end{aligned} \tag{36}$$

For the final condition to hold

$$\epsilon^{\max} \leq \epsilon\frac{C(-D_2 + \sqrt{D_2^2 + 4(c_x - K - D_1)D_3}}{2D_3}. \tag{37}$$

In the stochastic case, we apply Theorems 5 and 6 as is to get the expected number of steps. In this case however, the step size may become smaller than the pre-specified σ_{min} parameter, due to inaccurate estimates. We can then use Lemma 12, to get a bound with high probability, regarding this minimum step size value. More specifically for the number of iterates n specified by Theorem 5, for $k \geq \frac{\log(1-e^{\frac{1}{n}\log(\delta)})}{\log(\frac{1-p_x}{p_x})} - 1$, throughout the updates

$$\sigma \geq \sigma'_{\text{min}} = \frac{1}{\gamma^k}\sigma_{\text{min}},$$

with probability at least $\delta > 0$, where γ is the update parameter for the min problem in Algorithm 3. We then get the similar bounds

$$\epsilon^{\max} \leq \epsilon \ \min\left\{\frac{2C(c_x - K - D_1)}{\gamma^k D_2}, \ \frac{C(-D_2 + \sqrt{D_2^2 + 4(c_x - K - D_1)D_3}}{2D_3\gamma^k}\right\}.$$

We note that K acts as a new sufficient decrease constant and should be taken into account for all assumptions of Theorem 5, namely $K > 2\epsilon_x$, which holds for the constant $c_x > D_1 + K > D_1 + 2\epsilon_x$.

□

D EXPERIMENTAL SETUP

D.1 Robust Optimization

The Wisconsin breast cancer data set, is a binary classification task with 569 samples in total, each having 30 attributes. We use a simple neural network with a hidden layer of size 50 and a LeakyReLU activation. This choice of activation accommodates the GDA baseline providing additional gradient information. All networks across methods and folds are initialized with the same weights. For the GDA method we tried a range of different learning rates from the set $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005\}$, but only present results for the cases that converged.

In Fig. 4 we present the evolution of the zero-one error across epochs for each method. We stress that one epoch for the GDA approach corresponds to one update each for the max and the min problem, whereas one epoch for DR corresponds to a series of updates for the max problem (at most 10) followed by a single update for the min problem. GDA was run for a total of 10000 epochs and DR for a maximum of 2000 epochs but usually converges a lot faster than that. GDA suffers considerably more by poor initializations compared to DR. In Fig. 4 constant large errors correspond to a constant output of the network for a specific class of the problem (for this unbalanced dataset with rates 0.63 and 0.37).

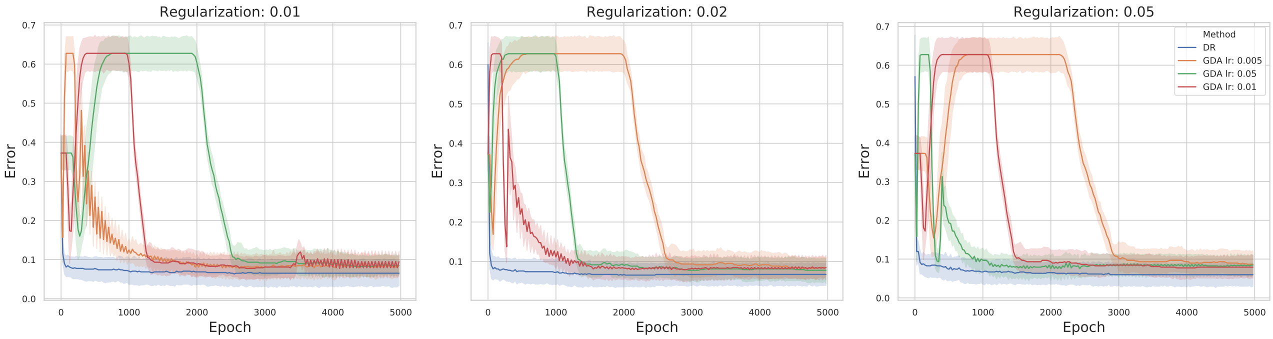

Figure 4: Misclassification error across epochs for each method.

D.2 Toy Examples

Although in the examples following, the objective of the max player is nonconcave and does not satisfy the PL condition, empirical results demonstrate that the proposed algorithm can be successful. We begin by illustrating examples of GANs learning different 2D underlying distributions for a continuous case in Fig. 5. Both the generator and the discriminator have 2 hidden layers of size 20 (64 for learning a mixture of Gaussian in a grid formation) with Tanh activations, while we also use spectral normalization for the discriminator. In all scenarios, we sample the latent code from a lower-dimensional space $N(0, I_2)$, such that it matches the data dimensionality, allowing the generator to learn a simpler mapping (as in Grnarova et al. (2017)).

Motivated by encouraging results, we proceed in a discrete setting, where each of the 2 dimensions of the underlying distributions is parametrized by a categorical variable. The choice of this categorical variable makes the objective function of the generator nondifferentiable. As aforementioned, our algorithm can support multi-categorical data. In the current literature, the most popular methods to deal with this kind of scenario are baselines based on the Gumbel-softmax or the REINFORCE algorithm. Due to their sampling techniques though, dependence on the number of parameters of the model is exponential for these baselines.

We describe shortly how training is performed for each of the baselines used. Based on the output logits o of size n, the result of a projection layer, each method samples a new point y.

Gumbel-softmax Using the Gumbel-max trick, the sampling can be parametrized as

$$y = \text{one-hot}(\arg\max_{1 \leq i \leq n}(o^{(i)} + g^{(i)})), \tag{38}$$

where g are sampled from the *i.i.d.* Gumbel distribution. To enable the calculation of gradients, this is relaxed to the result of a softmax operation

$$\hat{y} = \sigma(\frac{o + g}{T}), \tag{39}$$

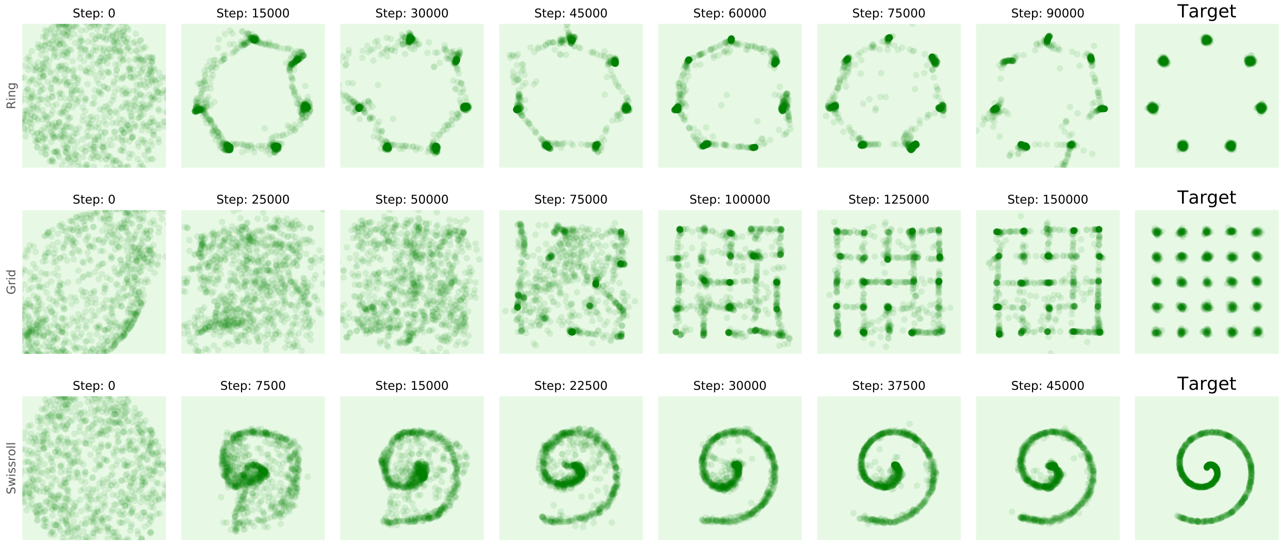

Figure 5: Mode collapse check for direct-search methods for three different problems in the continuous setting.

with $T > 0$, the temperature, controlling the softness of the sampling. A high initial temperature value forces more exploration. As the temperature decreases, $\hat{y}$ becomes a better approximation of y, leading however to steeper gradients and more instabilities. This is also the reason why gradient clipping is crucial for the stability of this method. Untimely updates of the temperature have been known to bolster mode collapse. For all experiments, we use an exponential decay update scheme, decreasing the temperature after a predefined number of steps.

REINFORCE We sample a new point s from the output logits and appoint a specific reward according to the output of the discriminator $r = 2 * (\mathcal{D}(s) - 0.5)$ (we remind that the discriminator uses a sigmoid output), rewarding positively samples that manage to fool the discriminator and negatively those that fail to do so. Subtracting the baseline value of 0.5 helps reduce variance. To alleviate the large variance introduced by the sampling, we increase the number of steps taken by the generator compared to the other methods.

Direct-search Direct-search method just chooses the output with the highest probability

$$y = \text{one-hot}(\arg\max_{1 \leq i \leq n} o^{(i)}). \tag{40}$$

For the discrete toy example illustrated, we draw samples from an evenly weighted and evenly spaced mixture of Gaussians with 7 components in a 2-dimensional space. The 2d sample is then discretized, according to a specific level, leading to 51 possible values for each of the two dimensions of the problem. These categorical datapoints are then transformed to the corresponding continuous one, by a linear mapping, and given as input to the discriminator. This leads to an ordinal relationship between the categorical points (we only display ordinal data for visualization purposes). To monitor the learning curve of the generator compared to the true distribution, we also calculate the Hellinger distance, which is a nonparametric method that calculates the difference of two discrete distributions as

$$\mathcal{H}(\mathcal{P}, \mathcal{Q}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}, \tag{41}$$

as well as the maximum mean discrepancy, comparing the generated samples with the underlying data. Note that this example suffers from stochasticity due to the way samples are collected and batches are created. Results for all baselines are provided in Figure 6.

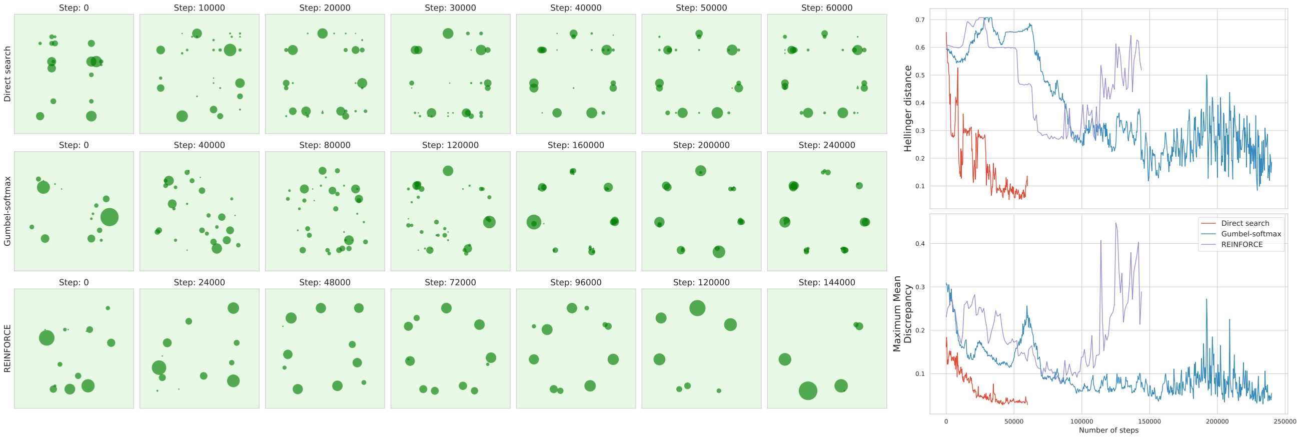

Figure 6: Comparison of direct-search with baselines for the discrete toy example.

When using DR, to overcome the nonsmoothness of the objective function, increasing the ℓ_2 regularization can help with the convergence. Too large of an increase inevitably leads to mode collapse. On the other hand, too small of a regularization parameter requires a large enough step size parameter to enable progress and thus can make learning more difficult. In general, we found that DR converges to similar solutions for a wide range of regularization parameter choices without significant variation in terms of the number of steps required to do so.