APPENDIX

A Theoretical Analysis

A.1 Lemma 4

Lemma 4. *Suppose that f satisfies assumption 3; then, $\mathbf{z}^* := (\mathbf{x}^*, \mathbf{y}^*)$ is a locally optimal saddle point on $\mathcal{K}_\gamma^*$ if and only if the gradient is zero, i.e.*

$$\nabla f(\mathbf{x}^*, \mathbf{y}^*) = 0, \tag{33}$$

and the second derivative at $(\mathbf{x}^*, \mathbf{y}^*)$ is positive definite in $\mathbf{x}$ and negative definite in $\mathbf{y}$, i.e., there exist $\mu_\mathbf{x}, \mu_\mathbf{y} > 0$ such that

$$\nabla^2_\mathbf{x} f(\mathbf{x}^*, \mathbf{y}^*) \succ \mu_\mathbf{x} \mathbf{I}, \qquad \nabla^2_\mathbf{y} f(\mathbf{x}^*, \mathbf{y}^*) \prec -\mu_\mathbf{y} \mathbf{I}. \tag{34}$$

Proof. From definition 1 follows that a locally optimal saddle point $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{K}_\gamma^*$ is a point for which the following two conditions hold:

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}, \mathbf{y}) \quad \text{and} \quad f(\mathbf{x}, \mathbf{y}^*) \geq f(\mathbf{x}, \mathbf{y}) \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{K}_\gamma^* \tag{35}$$

Hence, $\mathbf{x}$ is a local minimizer of f and $\mathbf{y}$ is a local maximizer. We therefore, without loss of generality, prove the statement of the lemma only for the minimizer $\mathbf{x}$, namely that

(i) $\nabla_\mathbf{x} f(\mathbf{x}^*, \mathbf{y}) = 0 \quad \forall \mathbf{y}$ s.t. $\|\mathbf{y} - \mathbf{y}^*\| \leq \gamma$

(ii) $\nabla^2_\mathbf{x} f(\mathbf{x}^*, \mathbf{y}) \succ \mu_\mathbf{x} \mathbf{I} \quad \forall \mathbf{y}$ s.t. $\|\mathbf{y} - \mathbf{y}^*\| \leq \gamma, \mu_\mathbf{x} > 0$.

The proof for the maximizer $\mathbf{y}$ directly follows from this.

(i) If we assume that $\nabla_\mathbf{x} f(\mathbf{x}^*, \mathbf{y}) \neq 0$, then there exists a feasible direction $\mathbf{d} \in \mathbb{R}^k$ such that $\nabla_\mathbf{x}^\top f(\mathbf{x}^*, \mathbf{y})\mathbf{d} < 0$, and we can find a step size $\alpha > 0$ for $x(\alpha) = \mathbf{x}^* + \alpha \mathbf{d}$ s.t. $\alpha \|\mathbf{d}\| \leq \gamma$ with $\|\mathbf{d}\| = 1$. Using the smoothness assumptions (Assumption 2), we arrive at the following inequality

$$\Big| f(\mathbf{x}(\alpha), \mathbf{y}) - f(\mathbf{x}^*, \mathbf{y}) - \nabla_\mathbf{x}^\top f(\mathbf{x}^*, \mathbf{y})(\mathbf{x}(\alpha) - \mathbf{x}^*) - \frac{1}{2}(\mathbf{x}(\alpha) - \mathbf{x}^*)^\top \nabla^2_\mathbf{x} f(\mathbf{x}^*, \mathbf{y})(\mathbf{x}(\alpha) - \mathbf{x}^*) \Big| \leq \frac{\rho_\mathbf{x}}{6} \|(\mathbf{x}(\alpha) - \mathbf{x}^*)\| \tag{36}$$

Hence, it holds that:

$$f(\mathbf{x}(\alpha), \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}) + \alpha \left(\nabla_\mathbf{x}^\top f(\mathbf{x}^*, \mathbf{y})\mathbf{d} + \frac{\rho_\mathbf{x}}{6} + \frac{1}{2}\alpha L_x \right) \tag{37}$$

By choosing the gradient descent direction $\mathbf{d} = -\beta \nabla_\mathbf{x}^\top f(\mathbf{x}^*, \mathbf{y})$ (with $\beta > 0$ s.t. $\|\mathbf{d}\| = 1$), we can find a step size $0 < \alpha < \frac{2\beta}{L_x} \|\nabla_\mathbf{x} f(\mathbf{x}^*, \mathbf{y})\|^2 - \frac{\rho_\mathbf{x}}{3L_x}$ such that $f(\mathbf{x}(\alpha), \mathbf{y}) < f(\mathbf{x}^*, \mathbf{y})$,
which contradicts that $f(\mathbf{x}^*, \mathbf{y})$ is a local minimizer. Hence, $\nabla_\mathbf{x} f(\mathbf{x}^*, \mathbf{y}) = 0$ is a necessary condition for a local minimizer.

(ii) To prove the second statement, we again make use of inequality (36) coming from the smoothness assumption and the update $\mathbf{x}(\alpha) = \mathbf{x}^* + \alpha \mathbf{d}$ s.t. $\alpha \|\mathbf{d}\| \leq \gamma$ with $\|\mathbf{d}\| = 1$. From (i) we know that $\nabla_\mathbf{x} f(\mathbf{x}^*, \mathbf{y}) = 0$ and, therefore, we obtain:

$$\left| f(\mathbf{x}(\alpha), \mathbf{y}) - f(\mathbf{x}^*, \mathbf{y}) - \frac{1}{2}\mathbf{d}^\top \nabla^2_\mathbf{x} f(\mathbf{x}^*, \mathbf{y})\mathbf{d} \right| \leq \frac{\rho_\mathbf{x}}{6}\alpha \tag{38}$$

$$\Rightarrow f(\mathbf{x}(\alpha), \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}) + \frac{1}{2}\mathbf{d}^\top \nabla^2_\mathbf{x} f(\mathbf{x}^*, \mathbf{y})\mathbf{d} + \frac{\rho_\mathbf{x}}{6}\alpha \tag{39}$$

If $\nabla^2_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y})$ is not positive semi-definite, then there exists at least one eigenvector $\mathbf{v}$ with negative curvature, i.e. $\mathbf{v}^\top \nabla^2_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y})\mathbf{v} = -\epsilon < 0$. This implies that for $\alpha > \frac{1}{3\epsilon}\rho_{\mathbf{x}}$ following the curvature vector $\mathbf{v}$ decreases the function value, i.e., $f(\mathbf{x}(\alpha), \mathbf{y}) < f(\mathbf{x}^*, \mathbf{y})$. This contradicts that $f(\mathbf{x}^*, \mathbf{y})$ is a local minimizer which proves the sufficient condition

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}) \succ \mu_{\mathbf{x}} \mathbf{I} \quad \text{, with } \mu_{\mathbf{x}} > 0. \tag{40}$$

□

A.2 Lemma 6

The following Lemma 11 proves that the gradient-based mapping for the saddle point problem is a diffeomorphism which will be needed in the proof for Lemma 6.

Lemma 11. *Suppose that assumption 2 holds; then the gradient mapping for the saddle point problem*

$$g(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y}) + \eta(-\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})) \tag{41}$$

with step size $\eta < \min\left(\frac{1}{L_x}, \frac{1}{L_y}, \frac{1}{\sqrt{2}L_z}\right)$ *is a diffeomorphism.*

Proof. The following proof is very much based on the proof of proposition 4.5 from [22].
A necessary condition for a diffeomorphism is bijectivity. Hence, we need to check that g is (i) injective, and (ii) surjective for $\eta < \min\left(\frac{1}{L_x}, \frac{1}{L_y}, \frac{1}{\sqrt{2}L_z}\right)$.

(i) Consider two points $\mathbf{z} := (\mathbf{x}, \mathbf{y}), \widetilde{\mathbf{z}} := (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in \mathcal{K}_\gamma$ for which

$$g(\mathbf{z}) = g(\widetilde{\mathbf{z}}) \tag{42}$$

holds. Then, we have that

$$\mathbf{z} - \widetilde{\mathbf{z}} = \eta \begin{bmatrix} -\nabla_{\mathbf{x}} f(\mathbf{z}) \\ \nabla_{\mathbf{y}} f(\mathbf{z}) \end{bmatrix} - \eta \begin{bmatrix} -\nabla_{\mathbf{x}} f(\widetilde{\mathbf{z}}) \\ \nabla_{\mathbf{y}} f(\widetilde{\mathbf{z}}) \end{bmatrix} \tag{43}$$

$$= \eta \begin{bmatrix} -\nabla_{\mathbf{x}} f(\mathbf{z}) + \nabla_{\mathbf{x}} f(\widetilde{\mathbf{z}}) \\ \nabla_{\mathbf{y}} f(\mathbf{z}) - \nabla_{\mathbf{y}} f(\widetilde{\mathbf{z}}) \end{bmatrix}. \tag{44}$$

Note that

$$\|\nabla_{\mathbf{x}} f(\mathbf{z}) - \nabla_{\mathbf{x}} f(\widetilde{\mathbf{z}})\| \leq \|\nabla f(\mathbf{z}) - \nabla f(\widetilde{\mathbf{z}})\| \leq L_z \|\mathbf{z} - \widetilde{\mathbf{z}}\| \tag{45}$$

$$\|\nabla_{\mathbf{y}} f(\mathbf{z}) - \nabla_{\mathbf{y}} f(\widetilde{\mathbf{z}})\| \leq \|\nabla f(\mathbf{z}) - \nabla f(\widetilde{\mathbf{z}})\| \leq L_z \|\mathbf{z} - \widetilde{\mathbf{z}}\|, \tag{46}$$

from which follows that

$$\|\mathbf{z} - \widetilde{\mathbf{z}}\| \leq \eta \sqrt{L_z^2 \|\mathbf{z} - \widetilde{\mathbf{z}}\|^2 + L_z^2 \|\mathbf{z} - \widetilde{\mathbf{z}}\|^2} \tag{47}$$

$$= \sqrt{2}\eta L_z \|\mathbf{z} - \widetilde{\mathbf{z}}\|. \tag{48}$$

For $0 < \eta < \frac{1}{\sqrt{2}L_z}$ this means $\mathbf{z} = \widetilde{\mathbf{z}}$, and therefore g is injective.

(ii) We will show that g is surjective by constructing an explicit inverse function for both optimization problems individually. As suggested by [22], we make use of the proximal point algorithm on the function $-f$ for the parameters $\mathbf{x}, \mathbf{y}$, individually.
For the parameter $\mathbf{x}$ the proximal point mapping of $-f$ centered at $\widetilde{\mathbf{x}}$ is given by

$$\mathbf{x}(\widetilde{\mathbf{x}}) = \arg\min_{\mathbf{x}} \underbrace{\frac{1}{2}\|\mathbf{x} - \widetilde{\mathbf{x}}\|^2 - \eta f(\mathbf{x}, \mathbf{y})}_{h(\mathbf{x})} \tag{49}$$

Moreover, note that $h(\mathbf{x})$ is strongly convex in $\mathbf{x}$ if $\eta < \frac{1}{L_x}$:

$$(\nabla_{\mathbf{x}} h(\mathbf{x}) - \nabla_{\mathbf{x}} h(\widehat{\mathbf{x}}))^\top (\mathbf{x} - \widehat{\mathbf{x}}) = (\mathbf{x} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \widehat{\mathbf{x}} + \eta \nabla_{\mathbf{x}} f(\widehat{\mathbf{x}}, \mathbf{y}))^\top (\mathbf{x} - \widehat{\mathbf{x}}) \tag{50}$$

$$= \|\mathbf{x} - \widehat{\mathbf{x}}\|^2 - \eta (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\widehat{\mathbf{x}}, \mathbf{y}))^\top (\mathbf{x} - \widehat{\mathbf{x}}) \geq (1 - \eta L_x) \|\mathbf{x} - \widehat{\mathbf{x}}\|^2 \tag{51}$$

Hence, the function $h(\mathbf{x})$ has a unique minimizer, given by

$$0 \stackrel{!}{=} \nabla_{\mathbf{x}} h(\mathbf{x}) = \mathbf{x} - \widetilde{\mathbf{x}} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \tag{52}$$

$$\Rightarrow \widetilde{\mathbf{x}} = \mathbf{x} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \tag{53}$$

which means that there is a unique mapping from $\mathbf{x}$ to $\widetilde{\mathbf{x}}$ under the gradient mapping g if $\eta < \frac{1}{L_x}$.
The same line of reasoning can be applied to the parameter $\mathbf{y}$ with the negative proximal point mapping of $-f$ centered at $\widetilde{\mathbf{y}}$, i.e.

$$\mathbf{y}(\widetilde{\mathbf{y}}) = \arg\max_{\mathbf{y}} \underbrace{-\frac{1}{2} \|\mathbf{y} - \widetilde{\mathbf{y}}\|^2 - \eta f(\mathbf{x}, \mathbf{y})}_{h(\mathbf{y})} \tag{54}$$

Similarly as before, we can observe that $h(\mathbf{y})$ is strictly concave for $\eta < \frac{1}{L_y}$ and that the unique minimizer of $h(\mathbf{y})$ yields the $\mathbf{y}$ update step of g. This let's us conclude that the mapping g is surjective for $(\mathbf{x}, \mathbf{y})$ if $\eta < \min\left(\frac{1}{L_x}, \frac{1}{L_y}\right)$

Observing that for $\eta < \frac{1}{L_{\mathbf{z}}}$, g^{-1} is continuously differentiable concludes the proof that g is a diffeomorphism. □

Lemma 6 (Random Initialization)**.** *Suppose that assumption 2 holds. Consider gradient iterates of Eq.* (3) *with step size* $\eta < \min\left(\frac{1}{L_x}, \frac{1}{L_y}, \frac{1}{\sqrt{2}L_z}\right)$ *starting from a random initial point. If the iterates converge to a stationary point, then the stationary point is almost surely stable.*

Proof. From lemma 11 follows that the gradient update from Eq. (3) for the saddle point problem is a diffeomorphism. The remaining part of the proof follows directly from theorem 4.1 from [22]. □

A.3 Lemma 7

Lemma 7. *The point* $\mathbf{z} := (\mathbf{x}, \mathbf{y})$ *is a stationary point of the iterates in Eq.* (22) *if and only if* $\mathbf{z}$ *is a locally optimal saddle point.*

Proof. The point $\mathbf{z}^*$ is a stationary point of the iterates if and only if $\mathbf{v}_{\mathbf{z}^*} + \eta(-\nabla_{\mathbf{x}} f(\mathbf{z}^*), \nabla_{\mathbf{y}} f(\mathbf{z}^*)) = 0$. Let's consider w.l.o.g. only the stationary point condition with respect to $\mathbf{x}$, i.e.

$$\mathbf{v}_{\mathbf{z}^*} = \eta \nabla_{\mathbf{x}} f(\mathbf{z}^*) \tag{55}$$

We prove that the above equation holds only if $\nabla f(\mathbf{z}^*) = \mathbf{v}_{\mathbf{z}^*} = 0$. This can be proven by a simple contradiction; suppose that $\nabla f(\mathbf{z}^*) \neq 0$, then multiplying both sides of the above equation by $\nabla f(\mathbf{z}^*)$ yields

$$\underbrace{\lambda_{\mathbf{x}^*}/(2\rho_{\mathbf{x}})}_{<0} \underbrace{\operatorname{sgn}(\mathbf{v}_{\mathbf{x}^*}^\top \nabla_{\mathbf{x}} f(\mathbf{z}^*)) \mathbf{v}_{\mathbf{x}^*}^\top \nabla_{\mathbf{x}} f(\mathbf{z}^*)}_{>0} = \eta \|\nabla_{\mathbf{x}} f(\mathbf{z}^*)\|^2 \tag{56}$$

Since the left-hand side is negative and the right-hand side is positive, the above equation leads to a contradiction. Therefore, $\nabla f(\mathbf{z}^*) = 0$ and $\mathbf{v}_{\mathbf{z}^*} = 0$. This means that $\lambda_{\mathbf{x}^*} \geq 0$ and $\lambda_{\mathbf{y}^*} \leq 0$ and therefore according to lemma 4, $\mathbf{z}^*$ is a locally optimal saddle point. □

A.4 Lemma 8

Lemma 8. *Suppose that assumptions 2 and 3 hold. Let* $\mathbf{z}^* := (\mathbf{x}^*, \mathbf{y}^*)$ *be a locally optimal saddle point, i.e.*

$$\nabla f(\mathbf{z}) = 0,\ \nabla^2_{\mathbf{x}} f(\mathbf{z}^*) \succeq \mu_{\mathbf{x}} \mathbf{I},\ \nabla^2_{\mathbf{y}} f(\mathbf{z}^*) \preceq -\mu_{\mathbf{y}} \mathbf{I},\ (\mu_{\mathbf{x}}, \mu_{\mathbf{y}} > 0) \tag{57}$$

Then iterates of Eq. (22) *are stable in* $\mathcal{K}^*_\gamma$ *as long as*

$$\gamma \leq \min\{\mu_{\mathbf{x}}/(\sqrt{2}\rho_{\mathbf{x}}), \mu_{\mathbf{y}}/(\sqrt{2}\rho_{\mathbf{y}})\} \tag{58}$$

Proof. The proof is based on a simple idea: in a $\mathcal{K}^*_\gamma$ neighborhood of a locally optimal saddle point, f can not have extreme curvatures, i.e., $\mathbf{v}_{\mathbf{z}} = \mathbf{0}$. Hence, within $\mathcal{K}^*_\gamma$ the update of Eq. (22) reduces to the gradient update in Eq. (3), which is stable according to [27, 26].
To prove our claim that negative curvature doesn't exist in $\mathcal{K}^*_\gamma$, we make use of the smoothness assumption. Suppose that $\mathbf{z} := (\mathbf{x}, \mathbf{y}) \in \mathcal{K}^*_\gamma$, then the smoothness assumption 2 implies

$$\begin{aligned}
\nabla^2_{\mathbf{x}} f(\mathbf{z}) &= \nabla^2_{\mathbf{x}} f(\mathbf{z}_*) - \left(\nabla^2_{\mathbf{x}} f(\mathbf{z}_*) - \nabla^2_{\mathbf{x}} f(\mathbf{z})\right) && (59)\\
&\succeq \nabla^2_{\mathbf{x}} f(\mathbf{z}_*) - \rho_{\mathbf{x}} \|\mathbf{z} - \mathbf{z}_*\| \mathbf{I} && (60)\\
&\succeq \nabla^2_{\mathbf{x}} f(\mathbf{z}_*) - \sqrt{2}\rho_{\mathbf{x}} \gamma \mathbf{I} && (61)\\
&\succeq (\mu_{\mathbf{x}} - \sqrt{2}\rho_{\mathbf{x}}\gamma)\mathbf{I} && (62)\\
&\succ 0 \qquad [\gamma < \mu_{\mathbf{x}}/(\sqrt{2}\rho_{\mathbf{x}})] && (63)
\end{aligned}$$

Similarly, one can show that

$$\nabla^2_{\mathbf{y}} f(\mathbf{z}) \prec 0 \qquad [\gamma < \mu_{\mathbf{y}}/(\sqrt{2}\rho_{\mathbf{y}})]. \tag{64}$$

Therefore, the extreme curvature direction is zero according to the definition in Eq. (20). □

A.5 Lemma 9

Lemma 9. *Suppose that* $\mathbf{z}^* := (\mathbf{x}^*, \mathbf{y}^*)$ *is an undesired stationary point of the gradient dynamics, namely*

$$\nabla f(\mathbf{z}^*) = 0, \|\mathbf{v}_{\mathbf{z}^*}\| > 0. \tag{65}$$

Consider the iterates of Eq. (22) *starting from* $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{y}_0)$ *in a* γ*-neighbourhood of* $\mathbf{z}^*$*. After one step the iterates escape the* γ*-neighbourhood of* $\mathbf{z}^*$*, i.e.*

$$\|\mathbf{z}_1 - \mathbf{z}^*\| \geq \gamma \tag{66}$$

for a sufficiently small $\gamma = \mathcal{O}(\|\mathbf{v}_{\mathbf{z}^*}\|)$.

Proof. **Preliminaries:** Consider compact notations

$$\begin{aligned}
\nabla_0 &:= (-\nabla_{\mathbf{x}} f(\mathbf{z}_0), \nabla_{\mathbf{y}} f(\mathbf{z}_0)), \mathbf{v}_0 := \mathbf{v}_{\mathbf{z}_0}, \mathbf{v}_* = \mathbf{v}_{\mathbf{z}^*} && (67)\\
\lambda^{(-)} &:= \lambda_{\min}\left(\nabla^2_{\mathbf{x}} f(\mathbf{z}^*)\right) < 0, \lambda^{(+)} := \lambda_{\max}\left(\nabla^2_{\mathbf{y}} f(\mathbf{z}^*)\right) > 0 && (68)\\
\lambda_0^{(-)} &:= \lambda_{\min}\left(\nabla^2_{\mathbf{x}} f(\mathbf{z}_0)\right) < 0, \lambda_0^{(+)} := \lambda_{\max}\left(\nabla^2_{\mathbf{y}} f(\mathbf{z}_0)\right) > 0 && (69)
\end{aligned}$$

Characterizing extreme curvature: The choice of $\mathbf{v}_0$ ensures that

$$\nabla_0^\top \mathbf{v}_0 > 0 \tag{70}$$

holds. Since $\mathbf{z}_0$ lies in a γ-neighbourhood of $\mathbf{z}^*$, we can use the smoothness of f to relate the negative curvature at $\mathbf{z}_0$ to negative curvature in $\mathbf{z}^*$:

$$\begin{aligned}
\nabla^2_{\mathbf{x}} f(\mathbf{z}_0) &\preceq \nabla^2_{\mathbf{x}} f(\mathbf{z}^*) + \rho_{\mathbf{x}} \|\mathbf{z}_0 - \mathbf{z}^*\| \mathbf{I} && (71)\\
&\preceq \nabla^2_{\mathbf{x}} f(\mathbf{z}^*) + \sqrt{2}\rho_{\mathbf{x}} \gamma \mathbf{I}. && (72)
\end{aligned}$$

Therefore

$$\lambda_0^{(-)} \leq \lambda^{(-)} + \sqrt{2}\rho_{\mathbf{x}}\gamma \tag{73}$$

Similarly, one can show that

$$\lambda_0^{(+)} \geq \lambda^{(+)} - \sqrt{2}\rho_{\mathbf{y}}\gamma \tag{74}$$

Combining these two bounds yields

$$\begin{aligned}
\|\mathbf{v}_0\| &= \sqrt{(\lambda_0^{(-)}/(2\rho_{\mathbf{x}}))^2 + (\lambda_0^{(+)}/(2\rho_{\mathbf{y}}))^2} && (75)\\
&\geq \frac{1}{4}|\lambda_0^{(-)}/\rho_{\mathbf{x}}| + \frac{1}{4}|\lambda_0^{(+)}/\rho_{\mathbf{y}}| && (76)\\
&\geq \frac{1}{4}\left(|\lambda^{(-)}/\rho_{\mathbf{x}}| + \lambda^{(+)}/\rho_{\mathbf{y}} - 2\sqrt{2}\gamma\right) && (77)
\end{aligned}$$

To simplify the above bound, we use the compact notation $\lambda := 0.25(|\lambda^{(-)}/\rho_{\mathbf{x}}| + \lambda^{(+)}/\rho_{\mathbf{y}})$:

$$\|\mathbf{v}_0\| \geq \lambda - \frac{\sqrt{2}}{2}\gamma \tag{78}$$

Proof of escaping: The squared norm of the update can be computed as

$$\begin{aligned}
\|\mathbf{z}_1 - \mathbf{z}^*\|^2 &= \|\mathbf{z}_0 - \mathbf{z}^* + \eta\nabla_0 + \mathbf{v}_0\|^2 && (79)\\
&= \|\mathbf{z}_0 - \mathbf{z}^*\|^2 + \|\eta\nabla_0 + \mathbf{v}_0\|^2 + 2(\mathbf{z}_0 - \mathbf{z}^*)^\top(\eta\nabla_0 + \mathbf{v}_0) && (80)\\
&\geq \|\eta\nabla_0 + \mathbf{v}_0\|\left(\|\eta\nabla_0 + \mathbf{v}_0\| - 2\sqrt{2}\gamma\right) && (81)
\end{aligned}$$

Now, we plug the results obtained from the smoothness assumption in the above inequality. First, we provide a lower-bound on the sum of gradients and the extreme curvature:

$$\begin{aligned}
\|\eta\nabla_0 + \mathbf{v}_0\| &= \left(\eta^2\|\nabla_0\|^2 + \|\mathbf{v}_0\|^2 + 2\mathbf{v}_0^\top\nabla_0\right)^{1/2} && (82)\\
&\overset{(70)}{\geq} \|\mathbf{v}_0\| && (83)\\
&\overset{(78)}{\geq} \lambda - \frac{\sqrt{2}}{2}\gamma && (84)
\end{aligned}$$

Under the condition $\|\eta\nabla_0 + \mathbf{v}_0\| - 2\sqrt{2}\gamma > 0$, we can use the above lower-bound for inequality (81) yielding

$$\|\mathbf{z}_1 - \mathbf{z}^*\|^2 \geq (\lambda - \frac{\sqrt{2}}{2}\gamma)(\lambda - \frac{5}{\sqrt{2}}\gamma) \tag{85}$$

Choice of $\boldsymbol{\gamma}$: To complete our inductive argument, we need to choose γ such that the derived lower-bound of Eq. (85) is greater than γ^2, i.e.

$$(\lambda - \frac{\sqrt{2}}{2}\gamma)(\lambda - \frac{5}{\sqrt{2}}\gamma) \geq \gamma^2 \tag{86}$$

which holds for

$$\gamma \leq \lambda(\sqrt{2} - \frac{2}{\sqrt{3}}) = \mathcal{O}\left(\|\mathbf{v}_{\mathbf{z}^*}\|\right). \tag{87}$$

Observing that for this choice for the bound of γ it holds that

$$\begin{aligned}
\|\eta\nabla_0 + \mathbf{v}_0\| - 2\sqrt{2}\gamma &\geq \lambda - \frac{5}{\sqrt{2}}\gamma && (88)\\
&\geq \lambda - \frac{5}{\sqrt{2}}\lambda(\sqrt{2} - \frac{2}{\sqrt{3}}) && (89)\\
&= \lambda\frac{5\sqrt{6} - 12}{3} > 0 && (90)
\end{aligned}$$

concludes the proof. □

A.6 Guaranteed decrease/increase

The gradient update step of Eq. (3) has the property that an update with respect to $\mathbf{x}$ ($\mathbf{y}$) decreases (increases) the function value. The next lemma proves that CESP shares the same desirable property in regions where extreme curvature exists. Note that in regions without extreme curvature, the CESP method reduces to gradient based optimization and therefore inherits its theoretical properties.

Lemma 12. *In each iteration of Eq.* (22)*, f decreases in $\mathbf{x}$ with*

$$f(\mathbf{x}_{t+1}, \mathbf{y}_t) \leq f(\mathbf{x}_t, \mathbf{y}_t) - (\eta/2)\|\nabla_{\mathbf{x}} f(\mathbf{z}_t)\|^2 + \lambda_{\mathbf{x}}^3/(24\rho_{\mathbf{x}}^2), \tag{91}$$

and increases in $\mathbf{y}$ with

$$f(\mathbf{x}_t, \mathbf{y}_{t+1}) \geq f(\mathbf{x}_t, \mathbf{y}_t) + (\eta/2)\|\nabla_{\mathbf{y}} f(\mathbf{z}_t)\|^2 + \lambda_{\mathbf{y}}^3/(24\rho_{\mathbf{y}}^2). \tag{92}$$

as long as the step size is chosen as

$$\eta \leq \min\left\{ \frac{\sqrt{9L_{\mathbf{x}}^2 + 48\rho_{\mathbf{x}}\ell_{\mathbf{x}}} - 3L_{\mathbf{x}}}{8\rho_{\mathbf{x}}\ell_{\mathbf{x}}}, \frac{\sqrt{9L_{\mathbf{y}}^2 + 48\rho_{\mathbf{y}}\ell_{\mathbf{y}}} - 3L_{\mathbf{y}}}{8\rho_{\mathbf{y}}\ell_{\mathbf{y}}} \right\} \tag{93}$$

Proof. The Lipschitzness of the Hessian (Assumption 6) implies that for $\Delta \in \mathbb{R}^d$

$$|f(\mathbf{x}+\Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) - \Delta^\top \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \frac{1}{2}\Delta^\top \nabla_{\mathbf{x}}^2 f(\mathbf{x}, \mathbf{y})\Delta| \leq \frac{\rho_x}{6}\|\Delta\|^3 \tag{94}$$

holds. The update in Eq. (22) for $\mathbf{x}$ is given by $\Delta = \alpha\mathbf{v} - \eta\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$, where $\alpha = -\lambda/(2\rho_{\mathbf{x}})$ (where we assume w.l.o.g. that $\mathbf{v}^\top \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) < 0$) and $\mathbf{v}$ is the eigenvector associated with the minimum eigenvalue λ of the Hessian matrix $\nabla_{\mathbf{x}}^2 f(\mathbf{x}_t, \mathbf{y}_t)$. In the following we use the shorter notation: $\nabla f := \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$, and $\mathbf{H} := \nabla_{\mathbf{x}}^2 f(\mathbf{x}, \mathbf{y})$. We can construct a lower bound on the left hand side of Eq. (94) as

$$|f(\mathbf{x}+\Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) - \Delta^\top \nabla f - \frac{1}{2}\Delta^\top \mathbf{H}\Delta| \tag{95}$$

$$\geq f(\mathbf{x}+\Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) - \Delta^\top \nabla f - \frac{1}{2}\Delta^\top \mathbf{H}\Delta \tag{96}$$

$$\geq f(\mathbf{x}+\Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) - \alpha\mathbf{v}^\top \nabla f + (\eta - \frac{\eta^2}{2}L_x)\|\nabla f\|^2 - \frac{1}{2}\alpha^2\lambda + \alpha\eta\lambda\mathbf{v}^\top \nabla f \tag{97}$$

which leads to the following inequality

$$f(\mathbf{x}+\Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) \leq \alpha\mathbf{v}^\top \nabla f - (\eta - \frac{\eta^2}{2}L_x)\|\nabla f\|^2 + \frac{1}{2}\alpha^2\lambda - \alpha\eta\lambda\mathbf{v}^\top \nabla f + \frac{\rho_x}{6}\|\Delta\|^3 \tag{98}$$

$$\leq \frac{1}{2}\alpha^2\lambda - (\eta - \frac{\eta^2}{2}L_x)\|\nabla f\|^2 + \frac{\rho_x}{6}\|\Delta\|^3 \tag{99}$$

By using the triangular inequality we obtain the following bound on the cubic term

$$\|\Delta\|^3 \leq (\eta\|\nabla f\| + \alpha\|\mathbf{v}\|)^3 \tag{100}$$

$$\leq 4\eta^3\|\nabla f\|^3 + 4\alpha^3\|\mathbf{v}\|^3 = 4\eta^3\|\nabla f\|^3 + 4\alpha^3. \tag{101}$$

Replacing this bound into the upper bound of Eq. (98) yields

$$f(\mathbf{x}+\Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) \leq \frac{1}{2}\alpha^2\lambda - (\eta - \frac{\eta^2}{2}L_x)\|\nabla f\|^2 + \frac{\rho_x}{6}\left(4\eta^3\|\nabla f\|^3 + 4\alpha^3\right) \tag{102}$$

The choice of $\alpha = -\lambda/(2\rho_{\mathbf{x}})$ leads to further simplification of the above bound:

$$f(\mathbf{x}+\Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) \leq \frac{\lambda^3}{24\rho_{\mathbf{x}}^2} - \eta(1 - \frac{\eta}{2}L_x - \frac{2}{3}\rho_{\mathbf{x}}\eta^2\ell_{\mathbf{x}})\|\nabla f\|^2 \tag{103}$$

Now, we choose step size η such that

$$1 - \frac{\eta}{2} L_x - \frac{2}{3}\rho_{\mathbf{x}}\eta^2 \ell_{\mathbf{x}} \geq 1/2 \tag{104}$$

For $\eta \leq \frac{\sqrt{9L_{\mathbf{x}}^2 + 48\rho_{\mathbf{x}}\ell_{\mathbf{x}}} - 3L_{\mathbf{x}}}{8\rho_{\mathbf{x}}\ell_{\mathbf{x}}}$, the above inequality holds. Therefore, the following decrease in the function value is guaranteed

$$f(\mathbf{x} + \Delta, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) \leq \lambda^3/(24\rho_{\mathbf{x}}^2) - (\eta/2)\|\nabla f\|^2 \tag{105}$$

Similarly, one can derive the lower-bound for the function increase in $\mathbf{y}$. □

Within a region of extreme curvature, this lemma guarantees a larger decrease (increase) in $\mathbf{x}$ (in $\mathbf{y}$) compared to gradient descent (ascent). However, we are not claiming that these decrements accelerate the global convergence.

A.7 Lemma 10

Lemma 10. *The set of locally optimal saddle points as defined in Def. 1 and the set of stable points of the linear-transformed CESP update method in Eq.* (29) *are the same.*

Proof. As a direct consequence of lemma 7 and the positive definiteness property of the linear transformation matrix follows that a locally optimal saddle point is a stationary point of the linear-transformed updates.

In the following, we prove stability of locally optimal saddles. Let's consider a locally optimal saddle point $\mathbf{z}^* := (\mathbf{x}^*, \mathbf{y}^*)$ in the sense of Def. 1. From lemma 4 follows that

$$\nabla_{\mathbf{x}}^2 f(\mathbf{x}^*, \mathbf{y}^*) \succeq \mu_{\mathbf{x}}\mathbf{I}, \qquad \nabla_{\mathbf{y}}^2 f(\mathbf{x}^*, \mathbf{y}^*) \preceq -\mu_{\mathbf{y}}\mathbf{I}. \tag{106}$$

for $\mu_{\mathbf{x}}, \mu_{\mathbf{y}} > 0$. As a direct consequence the extreme curvature direction is zero, i.e. $\mathbf{v}_{\mathbf{z}} = \mathbf{0}$. Hence, the Jacobian of the update in Eq. (29) is given by

$$\mathbf{I} + \eta \mathbf{A}_{\mathbf{z}^*}\mathbf{H}(\mathbf{z}^*). \tag{107}$$

where $\mathbf{H}(\mathbf{z}^*)$ is the partial derivative of $(-\nabla_{\mathbf{x}} f(\mathbf{z}^*), \nabla_{\mathbf{y}} f(\mathbf{z}^*))$. The point $\mathbf{z}^*$ is a stable point of the dynamic if all eigenvalues of its Jacobian lie within the unit disk. This condition can be fulfilled, with a sufficiently small step size, if and only if all the real parts of the eigenvalues of $\mathbf{A}_{\mathbf{z}^*}\mathbf{H}(\mathbf{z}^*)$ are negative.
Hence, to prove stability of the update for a locally optimal saddle point $\mathbf{z}^*$, we have to show that the following expression is a Hurwitz matrix [17]:

$$\mathbf{J}(\mathbf{z}^*) = \underbrace{\mathbf{A}_{\mathbf{z}}}_{:=\mathbf{A}} \underbrace{\begin{bmatrix} -\nabla_{\mathbf{x}}^2 f(\mathbf{z}^*) & -\nabla_{\mathbf{x},\mathbf{y}} f(\mathbf{z}^*) \\ \nabla_{\mathbf{y},\mathbf{x}} f(\mathbf{z}^*) & \nabla_{\mathbf{y}}^2 f(\mathbf{z}^*) \end{bmatrix}}_{=\mathbf{H}} := \mathbf{J} \tag{108}$$

Since $\mathbf{A}$ is a symmetric, positive definite matrix, we can construct its square root $\mathbf{A}^{\frac{1}{2}}$ such that $\mathbf{A} = \mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}$. The matrix product $\mathbf{AH}$ can be re-written as

$$\mathbf{AH} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{H}\mathbf{A}^{\frac{1}{2}})\mathbf{A}^{-\frac{1}{2}}. \tag{109}$$

Since we are multiplying the matrix $\tilde{\mathbf{J}} = \mathbf{A}^{\frac{1}{2}}\mathbf{H}\mathbf{A}^{\frac{1}{2}}$ from the left with the inverse of the matrix from which we are multiplying from the right side, we can observe that $\mathbf{J}$ has the same eigenvalues as $\tilde{\mathbf{J}}$. The symmetric part of $\tilde{\mathbf{J}}(\mathbf{z}^*)$ is given by

$$\frac{1}{2}\left(\tilde{\mathbf{J}} + \tilde{\mathbf{J}}^\top\right) = \frac{1}{2}(\mathbf{A}^{\frac{1}{2}}\mathbf{H}\mathbf{A}^{\frac{1}{2}} + \mathbf{A}^{\frac{1}{2}}\mathbf{H}^\top\mathbf{A}^{\frac{1}{2}}) = \mathbf{A}^{\frac{1}{2}}(\mathbf{H} + \mathbf{H}^\top)\mathbf{A}^{\frac{1}{2}} \tag{110}$$

$$= \mathbf{A}^{\frac{1}{2}} \begin{bmatrix} -\nabla_{\mathbf{x}}^2 f(\mathbf{z}^*) & 0 \\ 0 & \nabla_{\mathbf{y}}^2 f(\mathbf{z}^*) \end{bmatrix} \mathbf{A}^{\frac{1}{2}} \tag{111}$$

From the assumption in Eq. (106) follows that the block diagonal matrix $(\mathbf{H} + \mathbf{H}^\top)$ is a symmetric, negative definite matrix, for which it therefore holds that $x^\top(\tilde{\mathbf{J}} + \tilde{\mathbf{J}}^\top)x \leq 0$ for any $x \in \mathrm{R}^{k+d}$. The remaining part of the

proof follows the argument from [5] Theorem 3.6.

Let (λ, v) be an eigenpair of $\tilde{\mathbf{J}}$. Then, the following two equalities hold:

$$v^* \tilde{\mathbf{J}} v = \lambda \tag{112}$$

$$(v^* \tilde{\mathbf{J}} v)^* = v^* \tilde{\mathbf{J}}^\top v = \bar{\lambda} \tag{113}$$

Therefore, we can re-write the real part of the eigenvalue λ as:

$$\mathrm{Re}(\lambda) = \frac{\lambda + \bar{\lambda}}{2} = \frac{1}{2} v^* (\tilde{\mathbf{J}} + \tilde{\mathbf{J}}^\top) v. \tag{114}$$

By observing that

$$v^* (\tilde{\mathbf{J}} + \tilde{\mathbf{J}}^\top) v = \mathrm{Re}(v)^\top (\tilde{\mathbf{J}} + \tilde{\mathbf{J}}^\top) \mathrm{Re}(v) + \mathrm{Im}(v)^\top (\tilde{\mathbf{J}} + \tilde{\mathbf{J}}^\top) \mathrm{Im}(v) \tag{115}$$

is a real, negative quantity, we can be sure that the real part of any eigenvalue of J is negative. Therefore it directly follows that, with a sufficiently small step size $\eta > 0$, any locally optimal saddle point $\mathbf{z}^*$ is a stable stationary point of the linear-transformed update method in Eq. (29). □

B Transformed Gradient Updates

Table 1 shows the update matrices for commonly used optimization methods on the saddle point problem.

Table 1: Update matrices of the different optimization schemes.

	Formula	positive definite?
Gradient Descent	$\mathcal{A}_t = I$ $\mathcal{B}_t = I$	Yes.
Adagrad [11]	$\mathcal{A}_{t,ii} = \left(\sqrt{\sum_{\tau=1}^{t} \left(\nabla_{\mathbf{x}_i} f(\mathbf{x}_\tau, \mathbf{y}_\tau) \right)^2 + \epsilon} \right)^{-1}$ $\mathcal{B}_{t,ii} = \left(\sqrt{\sum_{\tau=1}^{t} \left(\nabla_{\mathbf{y}_i} f(\mathbf{x}_\tau, \mathbf{y}_\tau) \right)^2 + \epsilon} \right)^{-1}$	Yes.
Saddle-Free Newton [10]	$\mathcal{A}_t = \left\lvert \nabla^2_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) \right\rvert^{-1}$ $\mathcal{B}_t = \left\lvert \nabla^2_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) \right\rvert^{-1}$	Yes.

C Experiments

C.1 Toy Example

Figure 6 shows the basin of attraction for GD and CESP on the toy saddle point problem of Eq. (17).

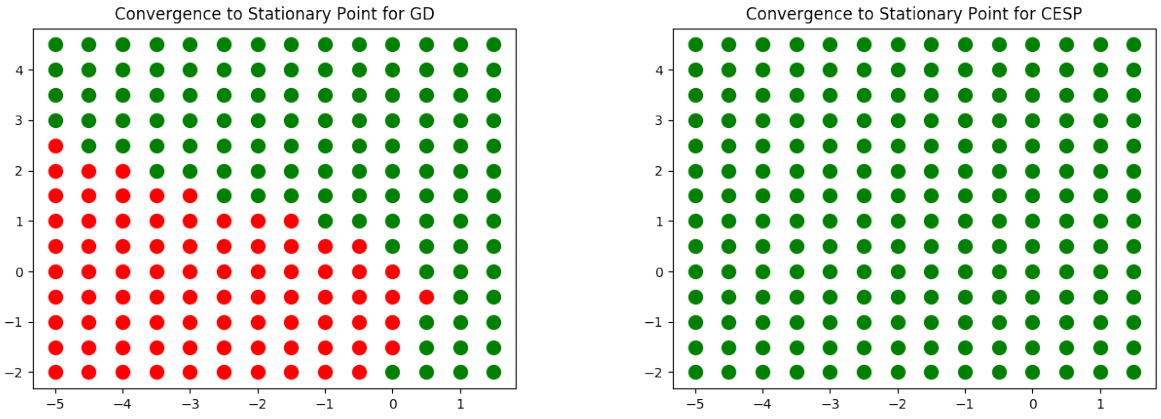
C.2 Generative Adversarial Networks

C.2.1 Single-layer GAN

Using two individual loss functions It is common practice in GAN training to not consider the saddle point problem as defined in Eq. (32), but rather split the training into two individual optimization problems over different functions. In particular, one usually considers

$$\min_{\mathbf{x}} (f_1(\mathbf{x}, \mathbf{y}) = -\mathbb{E}_{\mathbf{z} \sim p_z} \log D_{\mathbf{x}}(G_{\mathbf{y}}(\mathbf{z}))) \tag{116}$$

$$\max_{\mathbf{y}} (f_2(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\theta \sim p_d} \log D_{\mathbf{x}}(\theta) + \mathbb{E}_{\mathbf{z} \sim p_z} \log(1 - D_{\mathbf{x}}(G_{\mathbf{y}}(\mathbf{z})))) \tag{117}$$



(a) (b)

Figure 6: Comparison of the basin of attraction for GD and CESP to the locally optimal saddle point (green area) and the undesired critical point (red area).

Table 2: Parameters of the single-layer GAN model.

	Discriminator	Generator
Input Dimension	784	10
Hidden Layers	1	1
Hidden Units / Layer	100	100
Activation Function	Leaky ReLU	Leaky ReLU
Output Dimension	1	784
Batch Size	1000	
Learning Rate η	0.01	
Learning Rate $\alpha := \frac{1}{2\rho_{\mathbf{x}}} = \frac{1}{2\rho_{\mathbf{y}}}$	0.05	

Our CESP optimization method is defined individually for the two parameter sets $\mathbf{x}$ and $\mathbf{y}$ and can therefore also be applied on such a setting with two individual objectives. Figure 7 shows the results on the single-layer GAN problem, trained with two individual losses. In this experiment, CESP decreases the negative curvature of $\nabla^2_{\mathbf{x}} f_1$, while the gradient method can not exploit the negative curvature appropriately (the value of the smallest eigenvalue is oscillating in the negative area).

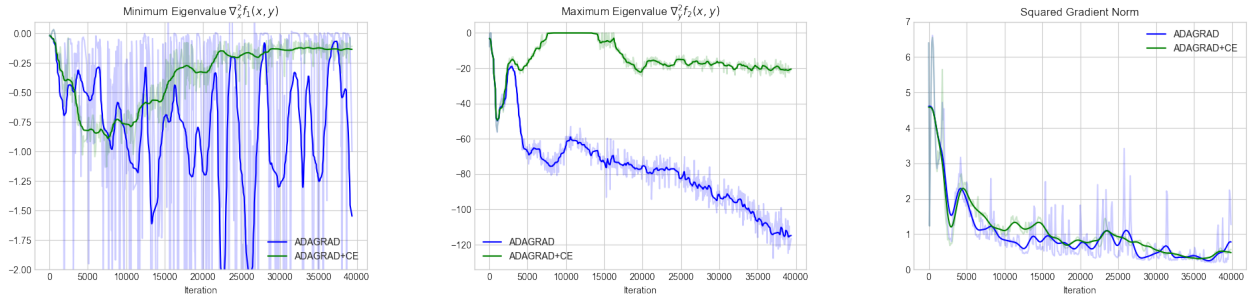


Figure 7: Results of the single-layer GAN with individual loss functions on MNIST data. The first two plots show the minimum eigenvalue of $\nabla^2_{\mathbf{x}} f_1(\mathbf{x}, \mathbf{y})$ and the maximum eigenvalue of $\nabla^2_{\mathbf{y}} f_2(\mathbf{x}, \mathbf{y})$, respectively. The third plot shows $\|\nabla f(\mathbf{z}_t)\|^2$. The transparent graph shows the original values, whereas the solid graph is smoothed with a Gaussian filter.