A Auxiliary lemmas

In this section, we introduce auxiliary lemmas that are necessary for our proofs.

Lemma 10 ([21, Lemma 19]). *If $f(\cdot)$ is ℓ-smooth and ρ-Hessian Lipschitz, $\eta = 1/\ell$, then the gradient descent sequence $\{\mathbf{x}_t\}$ obtained by $\mathbf{x}_{t+1} := \mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)$ satisfies:*

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \eta\|\nabla f(\mathbf{x})\|^2/2, \tag{37}$$

for any step t in which Negative Curvature Finding is not called.

Lemma 11 ([21, Lemma 23]). *For a ℓ-smooth and ρ-Hessian Lipschitz function f with its stochastic gradient satisfying Assumption 1, there exists an absolute constant c such that, for any fixed $t, t_0, \iota > 0$, with probability at least $1 - 4e^{\iota}$, the stochastic gradient descent sequence in Algorithm 8 satisfies:*

$$f(\mathbf{x}_{t_0+t}) - f(\mathbf{x}_{t_0}) \leq -\frac{\eta}{8}\sum_{i=0}^{t-1}\|\nabla f(\mathbf{x}_{t_0+i})\|^2 + c\cdot\frac{\sigma^2}{\ell}(t+\iota), \tag{38}$$

if during $t_0 \sim t_0 + t$, Stochastic Negative Curvature Finding has not been called.

Lemma 12 ([21, Lemma 29]). *Denote $\alpha(t) := \left[\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{2(t-1-\tau)}\right]^{1/2}$ and $\beta(t) := (1+\eta\gamma)^t/\sqrt{2\eta\gamma}$. If $\eta\gamma \in [0,1]$, then we have*

1. *$\alpha(t) \leq \beta(t)$ for any $t \in \mathbb{N}$;*
2. *$\alpha(t) \geq \beta(t)/\sqrt{3}, \quad \forall t \geq \ln 2/(\eta\gamma)$.*

Lemma 13 ([21, Lemma 30]). *Under the notation of Lemma 12 and Appendix D.1, letting $-\gamma := \lambda_{\min}(\tilde{\mathcal{H}})$, for any $0 \leq t \leq \mathscr{T}_s$ and $\iota > 0$ we have*

$$\Pr\left(\|\mathbf{q}_p(t)\| \leq \beta(t)\eta r_s\cdot\sqrt{\iota}\right) \geq 1 - 2e^{-\iota}, \tag{39}$$

and

$$\Pr\left(\|\mathbf{q}_p(t)\| \geq \frac{\beta(t)\eta r_s}{10\sqrt{n}}\cdot\frac{\delta}{\mathscr{T}_s}\right) \geq 1 - \frac{\delta}{\mathscr{T}_s}, \tag{40}$$

$$\Pr\left(\|\mathbf{q}_p(t)\| \geq \frac{\beta(t)\eta r_s}{10\sqrt{n}}\cdot\delta\right) \geq 1 - \delta. \tag{41}$$

Lemma 14 ([21, Lemma 37]). *Given i.i.d. $\mathbf{X}_1, \ldots, \mathbf{X}_\tau \in \mathbb{R}^n$ all being zero-mean nSG(σ_i), then for any $\iota > 0$, and $B > b > 0$, there exists an absolute constant c such that, with probability at least $1 - 2n\log(B/b)\cdot e^{-\iota}$:*

$$\sum_{i=1}^n \sigma_i^2 \geq B \quad \text{or} \quad \left\|\sum_{i=1}^\tau \mathbf{X}_i\right\| \leq c\cdot\sqrt{\max\left\{\sum_{i=1}^\tau\sigma_i^2, b\right\}\cdot\iota}. \tag{42}$$

The next two lemmas are frequently used in the large gradient scenario of the accelerated gradient descent method:

Lemma 15 ([22, Lemma 7]). *Consider the setting of Theorem 21, define a new parameter*

$$\tilde{\mathscr{T}} := \frac{\sqrt{\ell}}{(\rho\epsilon)^{1/4}}\cdot c_A, \tag{43}$$

for some large enough constant c_A. If $\|\nabla f(\mathbf{x}_\tau)\| \geq \epsilon$ for all $\tau \in [0, \tilde{\mathscr{T}}]$, then there exists a large enough positive constant c_{A0}, such that if we choose $c_A \geq c_{A0}$, by running Algorithm 2 we have $E_{\tilde{\mathscr{T}}} - E_0 \leq -\mathscr{E}$, in which $\mathscr{E} = \sqrt{\frac{\epsilon^3}{\rho}}\cdot c_A^{-7}$, and E_τ is defined as:

$$E_\tau := f(\mathbf{x}_\tau) + \frac{1}{2\eta}\|\mathbf{v}_\tau\|^2. \tag{44}$$

Lemma 16 ([22, Lemma 4 and Lemma 5]). *Assume that the function f is ℓ-smooth. Consider the setting of Theorem 21, for every iteration that is not within $\mathscr{T}'$ steps after uniform perturbation, we have:*

$$E_{\tau+1} \leq E_\tau, \tag{45}$$

where E_τ is defined in (44) in Lemma 15.

B Proof details of negative curvature finding by gradient descent

B.1 Proof of Lemma 17

Lemma 17. *Under the setting of Proposition 3, we use α_t to denote*

$$\alpha_t = \|\mathbf{y}_{t,\|}\|/\|\mathbf{y}_t\|, \tag{46}$$

where $\mathbf{y}_{t,\|}$ is the component of $\mathbf{y}_t$ in the subspace $\mathfrak{S}_\|$ spanned by $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$. Then, during all the $\mathscr{T}$ iterations of Algorithm 1, we have $\alpha_t \geq \alpha_{\min}$ for

$$\alpha_{\min} = \frac{\delta_0}{4}\sqrt{\frac{\pi}{n}}, \tag{47}$$

given that $\alpha_0 \geq \sqrt{\frac{\pi}{n}}\delta_0$.

Proof. In this proof, we consider the worst case, where the initial value $\alpha_0 = \sqrt{\frac{\pi}{n}}\delta_0$ and the component $y_{0,1}$ along $\mathbf{u}_1$ equals 0. Also, the eigenvalues satisfy

$$\lambda_2 = \lambda_3 = \cdots = \lambda_p = -\sqrt{\rho\epsilon}, \qquad \lambda_{p+1} = \lambda_{p+2} = \cdots = \lambda_{n-1} = -\sqrt{\rho\epsilon} + \nu, \tag{48}$$

for an arbitrarily small positive constant ν, which can make components of $\mathbf{y}_t$ in $\mathfrak{S}_\perp$ as large as possible to make α_t smaller. Out of the same reason, we assume that each time we make a gradient call at point $\mathbf{x}$, the derivation term Δ from pure quadratic approximation

$$\Delta = \nabla h_f(\mathbf{x}) - \mathcal{H}(\mathbf{0})\mathbf{x}, \tag{49}$$

lies in the direction that can make α_t as small as possible. Then, the component $\Delta_\|$ in $\mathfrak{S}_\|$ should be in the opposite direction to $\mathbf{x}_\|$, and the component $\Delta_\perp$ in $\mathfrak{S}_\perp$ should be in the direction of $\mathbf{x}_\perp$. Hence in this case, we have $\|\mathbf{y}_{t,\perp}\|/\|\mathbf{y}_t\|$ being non-decreasing, since ν can be arbitrarily small. Also, it admits the following recurrence formula:

$$\begin{aligned}
\|\mathbf{y}_{t+1,\perp}\| &= (1 + (\sqrt{\rho\epsilon} - \nu)/\ell)\|\mathbf{y}_{t,\perp}\| + \|\Delta_\perp\|/\ell && (50)\\
&\leq (1 + \sqrt{\rho\epsilon}/\ell)\|\mathbf{y}_{t,\perp}\| + \|\Delta\|/\ell && (51)\\
&\leq \Big(1 + \sqrt{\rho\epsilon}/\ell + \frac{\|\Delta\|}{\ell\|\mathbf{y}_{t,\perp}\|}\Big)\|\mathbf{y}_{t,\perp}\|, && (52)
\end{aligned}$$

where the second inequality is due to the fact that ν can be an arbitrarily small positive number. Note that since $\|\mathbf{y}_{t,\perp}\|/\|\mathbf{y}_t\|$ is non-decreasing in this worst-case scenario, we have

$$\frac{\|\Delta\|}{\|\mathbf{y}_{t,\perp}\|} \leq \frac{\|\Delta\|}{\|\mathbf{y}_t\|} \cdot \frac{\|\mathbf{y}_0\|}{\|\mathbf{y}_{0,\perp}\|} \leq \frac{2\|\Delta\|}{\|\mathbf{y}_t\|} \leq 2\rho r, \tag{53}$$

which leads to

$$\|\mathbf{y}_{t+1,\perp}\| \leq (1 + \sqrt{\rho\epsilon}/\ell + 2\rho r/\ell)\|\mathbf{y}_{t,\perp}\|. \tag{54}$$

On the other hand, suppose for some value t, we have $\alpha_k \geq \alpha_{\min}$ with any $1 \leq k \leq t$. Then,

$$\begin{aligned}
\|\mathbf{y}_{t+2,\|}\| &= (1 + \sqrt{\rho\epsilon}/\ell)\|\mathbf{y}_{t+1,\perp}\| - \|\Delta_\|\|/\ell && (55)\\
&\geq \Big(1 + \sqrt{\rho\epsilon}/\ell - \frac{\|\Delta\|}{\ell\|\mathbf{y}_{t,\|}\|}\Big)\|\mathbf{y}_{t,\|}\|. && (56)
\end{aligned}$$

Note that since $\|\mathbf{y}_{t,\|}\|/\|\mathbf{y}_t\| \geq \alpha_{\min}$, we have

$$\frac{\|\Delta\|}{\|\mathbf{y}_{t,\|}\|} \geq \frac{\|\Delta\|}{\alpha_{\min}\|\mathbf{y}_t\|} = \rho r/\alpha_{\min}, \tag{57}$$

which leads to

$$\|\mathbf{y}_{t+1,\|}\| \geq (1 + \sqrt{\rho\epsilon}/\ell - \rho r/(\alpha_{\min}\ell))\|\mathbf{y}_{t,\|}\|. \tag{58}$$

Then we can observe that

$$\frac{\|\mathbf{y}_{t,\|}\|}{\|\mathbf{y}_{t,\perp}\|} \geq \frac{\|\mathbf{y}_{0,\|}\|}{\|\mathbf{y}_{0,\perp}\|} \cdot \Big(\frac{1 + \sqrt{\rho\epsilon}/\ell - \rho r/(\alpha_{\min}\ell)}{1 + \sqrt{\rho\epsilon}/\ell + 2\rho r/\ell}\Big)^t, \tag{59}$$

where

$$\begin{aligned}
\frac{1 + \sqrt{\rho\epsilon}/\ell - \rho r/(\alpha_{\min}\ell)}{1 + \sqrt{\rho\epsilon}/\ell + 2\rho r/\ell} &\geq (1 + \sqrt{\rho\epsilon}/\ell - \rho r/(\alpha_{\min}\ell))(1 - \sqrt{\rho\epsilon}/\ell - 2\rho r/\ell) && (60)\\
&\geq 1 - \rho\epsilon/\ell^2 - \frac{2\rho r}{\alpha_{\min}\ell} \geq 1 - \frac{1}{\mathscr{T}}, && (61)
\end{aligned}$$

by which we can derive that

$$\frac{\|\mathbf{y}_{t,\|}\|}{\|\mathbf{y}_{t,\perp}\|} \geq \frac{\|\mathbf{y}_{0,\|}\|}{\|\mathbf{y}_{0,\perp}\|}(1-1/\mathscr{T})^t \tag{62}$$

$$\geq \frac{\|\mathbf{y}_{0,\|}\|}{\|\mathbf{y}_{0,\perp}\|}\exp\Big(-\frac{t}{\mathscr{T}-1}\Big) \geq \frac{\|\mathbf{y}_{0,\|}\|}{2\|\mathbf{y}_{0,\perp}\|}, \tag{63}$$

indicating

$$\alpha_t = \frac{\|\mathbf{y}_{t,\|}\|}{\sqrt{\|\mathbf{y}_{t,\|}\|^2+\|\mathbf{y}_{t,\perp}\|^2}} \geq \frac{\|\mathbf{y}_{0,\|}\|}{4\|\mathbf{y}_{0,\perp}\|} \geq \alpha_{\min}. \tag{64}$$

Hence, as long as $\alpha_k \geq \alpha_{\min}$ for any $1 \leq k \leq t-1$, we can also have $\alpha_t \geq \alpha_{\min}$ if $t \leq \mathscr{T}$. Since we have $\alpha_0 \geq \alpha_{\min}$, we can thus claim that $\alpha_t \geq \alpha_{\min}$ for any $t \leq \mathscr{T}$ using recurrence. □

B.2 Proof of Proposition 5

To make it easier to analyze the properties and running time of Algorithm 2, we introduce a new Algorithm 4, which has a more straightforward structure and has the same effect as Algorithm 2 near any saddle point $\tilde{\mathbf{x}}$.

Algorithm 4: Accelerated Negative Curvature Finding without Renormalization($\tilde{\mathbf{x}}, r', \mathscr{T}'$).

$\mathbf{x}_0 \leftarrow$Uniform($\mathbb{B}_0(r')$) where $\mathbb{B}_0(r')$ is the ℓ_2-norm ball centered at $\tilde{\mathbf{x}}$ with radius r';
$\mathbf{z}_0 \leftarrow \mathbf{x}_0$;
for $t = 1, ..., \mathscr{T}'$ **do**
  $\mathbf{x}_{t+1} \leftarrow \mathbf{z}_t - \eta\Big(\frac{\|\mathbf{z}_t-\tilde{\mathbf{x}}\|}{r'}\nabla f\Big(r'\cdot\frac{\mathbf{z}_t-\tilde{\mathbf{x}}}{\|\mathbf{z}_t-\tilde{\mathbf{x}}\|}+\tilde{\mathbf{x}}\Big)-\nabla f(\tilde{\mathbf{x}})\Big)$;
  $\mathbf{v}_{t+1} \leftarrow \mathbf{x}_{t+1}-\mathbf{x}_t$;
  $\mathbf{z}_{t+1} \leftarrow \mathbf{x}_{t+1}+(1-\theta)\mathbf{v}_{t+1}$;
Output $\frac{\mathbf{x}_{\mathscr{T}'}-\tilde{\mathbf{x}}}{\|\mathbf{x}_{\mathscr{T}'}-\tilde{\mathbf{x}}\|}$.

Quantitatively, this is demonstrated in the following lemma:

Lemma 18. *Under the setting of Proposition 5, suppose the perturbation in Line 5 of Algorithm 2 is added at $t=0$. Then with the same value of r', $\mathscr{T}'$, $\tilde{\mathbf{x}}$ and $\mathbf{x}_0$, the output of Algorithm 4 is the same as the unit vector $\hat{\mathbf{e}}$ in Algorithm 2 obtained $\mathscr{T}'$ steps later. In other words, if we separately denote the iteration sequence of $\{\mathbf{x}_t\}$ in Algorithm 2 and Algorithm 4 as*

$$\{\mathbf{x}_{1,0}, \mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,\mathscr{T}'}\}, \qquad \{\mathbf{x}_{2,0}, \mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,\mathscr{T}'}\}, \tag{65}$$

we have

$$\frac{\mathbf{x}_{1,\mathscr{T}'}-\tilde{\mathbf{x}}}{\|\mathbf{x}_{1,\mathscr{T}'}-\tilde{\mathbf{x}}\|} = \frac{\mathbf{x}_{2,\mathscr{T}'}-\tilde{\mathbf{x}}}{\|\mathbf{x}_{2,\mathscr{T}'}-\tilde{\mathbf{x}}\|}. \tag{66}$$

Proof. Without loss of generality, we assume $\tilde{\mathbf{x}} = \mathbf{0}$ and $\nabla f(\tilde{\mathbf{x}}) = \mathbf{0}$. Use recurrence to prove the desired relationship. Suppose the following identities holds for all $k \leq t$ with t being some natural number:

$$\frac{\mathbf{x}_{2,k}}{\|\mathbf{x}_{2,k}\|} = \frac{\mathbf{x}_{1,k}}{r}, \qquad \frac{\mathbf{z}_{2,k}}{\|\mathbf{x}_{2,k}\|} = \frac{\mathbf{z}_{1,k}}{r'}. \tag{67}$$

Then,

$$\mathbf{x}_{2,t+1} = \mathbf{z}_{2,t} - \eta\cdot\frac{\|\mathbf{z}_{2,t}\|}{r'}\nabla f(\mathbf{z}_{2,t}\cdot r'/\|\mathbf{z}_{2,t}\|) = \frac{\|\mathbf{z}_{2,t}\|}{r'}(\mathbf{z}_{1,t}-\eta\nabla f(\mathbf{z}_{1,t})). \tag{68}$$

Adopting the notation in Algorithm 2, we use $\mathbf{x}'_{1,t+1}$ and $\mathbf{z}'_{1,t+1}$ to separately denote the value of $\mathbf{x}_{1,t+1}$ and $\mathbf{z}_{1,t+1}$ before renormalization:

$$\mathbf{x}'_{1,t+1} = \mathbf{z}_{1,t} - \eta\nabla f(\mathbf{z}_{1,t}), \quad \mathbf{z}'_{1,t+1} = \mathbf{x}'_{1,t+1} + (1-\theta)(\mathbf{x}'_{1,t+1}-\mathbf{x}_{1,t}). \tag{69}$$

Then,

$$\mathbf{x}_{2,t+1} = \frac{\|\mathbf{z}_{2,t}\|}{r'}(\mathbf{z}_{1,t}-\eta\nabla f(\mathbf{z}_{1,t})) = \frac{\|\mathbf{z}_{2,t}\|}{r'}\cdot\mathbf{x}'_{1,t+1}, \tag{70}$$

which further leads to

$$\mathbf{z}_{2,t+1} = \mathbf{x}_{2,t+1} + (1-\theta)(\mathbf{x}_{2,t+1}-\mathbf{x}_{2,t}) = \frac{\|\mathbf{z}_{2,t}\|}{r'}\cdot\mathbf{z}'_{1,t+1}. \tag{71}$$

Note that $\mathbf{z}_{1,t+1} = \frac{r'}{\|\mathbf{z}'_{1,t+1}\|} \cdot \mathbf{z}'_{1,t+1}$, we thus have

$$\frac{\mathbf{z}_{2,t+1}}{\|\mathbf{z}_{2,t+1}\|} = \frac{\mathbf{z}_{1,t+1}}{r'}. \tag{72}$$

Hence,

$$\mathbf{x}_{2,t+1} = \frac{\|\mathbf{z}_{2,t}\|}{r'} \cdot \mathbf{x}'_{1,t+1} = \frac{\|\mathbf{z}_{2,t}\|}{r'} \cdot \frac{\|\mathbf{z}_{1,t+1}\|}{\|\mathbf{z}_{1,t}\|} \cdot \mathbf{x}_{1,t+1} = \frac{\|\mathbf{z}_{2,t+1}\|}{r'} \cdot \mathbf{x}_{1,t+1}. \tag{73}$$

Since (67) holds for $k = 0$, we can now claim that it also holds for $k = \mathscr{T}'$. □

Lemma 18 shows that, Algorithm 4 also works in principle for finding the negative curvature near any saddle point $\tilde{\mathbf{x}}$. But considering that Algorithm 4 may result in an exponentially large $\|\mathbf{x}_t\|$ during execution, and it is hard to be merged with the AGD algorithm for large gradient scenarios. Hence, only Algorithm 2 is applicable in practical situations.

Use $\mathcal{H}(\tilde{\mathbf{x}})$ to denote the Hessian matrix of f at $\tilde{\mathbf{x}}$.Observe that $\mathcal{H}(\tilde{\mathbf{x}})$ admits the following eigen-decomposition:

$$\mathcal{H}(\tilde{\mathbf{x}}) = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \tag{74}$$

where the set $\{\mathbf{u}_i\}_{i=1}^n$ forms an orthonormal basis of $\mathbb{R}^n$. Without loss of generality, we assume the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ corresponding to $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ satisfy

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n, \tag{75}$$

in which $\lambda_1 \leq -\sqrt{\rho\epsilon}$. If $\lambda_n \leq -\sqrt{\rho\epsilon}/2$, Proposition 5 holds directly, since no matter the value of $\hat{\mathbf{e}}$, we can have $f(\mathbf{x}_{\mathscr{T}'}) - f(\tilde{\mathbf{x}}) \leq -\sqrt{\epsilon^3/\rho}/384$. Hence, we only need to prove the case where $\lambda_n > -\sqrt{\rho\epsilon}$, in which there exists some p with

$$\lambda_p \leq -\sqrt{\rho\epsilon} < \lambda_{p+1}. \tag{76}$$

We use $\mathfrak{S}_{\|}$ to denote the subspace of $\mathbb{R}^n$ spanned by $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$, and use $\mathfrak{S}_{\perp}$ to denote the subspace spanned by $\{\mathbf{u}_{p+1}, \mathbf{u}_{p+2}, \ldots, \mathbf{u}_n\}$. Then we can have the following lemma:

Lemma 19. *Under the setting of Proposition 5, we use α'_t to denote*

$$\alpha'_t = \frac{\|\mathbf{x}_{t,\|}\|}{\|\mathbf{x}_t\|}, \tag{77}$$

in which $\mathbf{x}_{t,\|}$ is the component of $\mathbf{x}_t$ in the subspace $\mathfrak{S}_{\|}$. Then, during all the $\mathscr{T}'$ iterations of Algorithm 4, we have $\alpha'_t \geq \alpha'_{\min}$ for

$$\alpha'_{\min} = \frac{\delta_0}{8}\sqrt{\frac{\pi}{n}}, \tag{78}$$

given that $\alpha'_0 \geq \sqrt{\frac{\pi}{n}}\delta_0$.

Proof. Without loss of generality, assume $\tilde{\mathbf{x}} = \mathbf{0}$ and $\nabla f(\tilde{\mathbf{x}}) = \mathbf{0}$. In this proof, we consider the worst case, where the initial value $\alpha'_0 = \sqrt{\frac{\pi}{n}}\delta_0$ and the component $x_{0,1}$ along $\mathbf{u}_1$ equals 0. Also, the eigenvalues satisfy

$$\lambda_2 = \lambda_3 = \cdots = \lambda_p = -\sqrt{\rho\epsilon}, \qquad \lambda_{p+1} = \lambda_{p+2} = \cdots = \lambda_{n-1} = -\sqrt{\rho\epsilon} + \nu, \tag{79}$$

for an arbitrarily small positive constant ν, which can make components of $\mathbf{x}_t$ in $\mathfrak{S}_{\perp}$ as large as possible to make α'_t smaller. Out of the same reason, we assume that each time we make a gradient call at point $\mathbf{z}_t$, the derivation term Δ from pure quadratic approximation

$$\Delta = \frac{\|\mathbf{z}_t\|}{r'} \cdot \Big(\nabla f(\mathbf{z}_t \cdot r'/\|\mathbf{z}_t\|) - \mathcal{H}(\mathbf{0}) \cdot \frac{r'}{\|\mathbf{z}_t\|} \cdot \mathbf{z}_t\Big), \tag{80}$$

lies in the direction that can make α'_t as small as possible. Then, the component $\Delta_{\|}$ in $\mathfrak{S}_{\|}$ should be in the opposite direction to $\mathbf{z}_{\|}$, and the component $\Delta_{\perp}$ in $\mathfrak{S}_{\perp}$ should be in the direction of $\mathbf{z}_{\perp}$. Hence in this case, we have both $\|\mathbf{x}_{t,\perp}\|/\|\mathbf{x}_t\|$ and $\|\mathbf{z}_{t,\perp}\|/\|\mathbf{z}_t\|$ being non-decreasing, since ν can be arbitrarily small. Also, it admits the following recurrence formula:

$$\begin{aligned}\|\mathbf{x}_{t+2,\perp}\| &\leq (1 + \eta(\sqrt{\rho\epsilon} - \nu))\big(\|\mathbf{x}_{t+1,\perp}\| + (1-\theta)(\|\mathbf{x}_{t+1,\perp}\| - \|\mathbf{x}_{t,\perp}\|)\big) + \eta\|\Delta_{\perp}\| &(81)\\ &\leq (1 + \eta\sqrt{\rho\epsilon})\big(\|\mathbf{x}_{t+1,\perp}\| + (1-\theta)(\|\mathbf{x}_{t+1,\perp}\| - \|\mathbf{x}_{t,\perp}\|)\big) + \eta\|\Delta_{\perp}\|, &(82)\end{aligned}$$

where the second inequality is due to the fact that ν can be an arbitrarily small positive number. Note that since $\|\mathbf{x}_{t,\perp}\|/\|\mathbf{x}_t\|$ is non-decreasing in this worst-case scenario, we have

$$\frac{\|\Delta_\perp\|}{\|\mathbf{x}_{t+1,\perp}\|} \le \frac{\|\Delta\|}{\|\mathbf{x}_{t+1}\|} \cdot \frac{\|\mathbf{x}_0\|}{\|\mathbf{x}_{0,\perp}\|} \le \frac{2\|\Delta\|}{\|\mathbf{x}_{t+1}\|} \le 2\rho r', \tag{83}$$

which leads to

$$\|\mathbf{x}_{t+2,\perp}\| \le (1+\eta\sqrt{\rho\epsilon}+2\eta\rho r')\big((2-\theta)\|\mathbf{x}_{t+1,\perp}\| - (1-\theta)\|\mathbf{x}_{t,\perp}\|\big). \tag{84}$$

On the other hand, suppose for some value t, we have $\alpha'_k \ge \alpha'_{\min}$ with any $1 \le k \le t+1$. Then,

$$\|\mathbf{x}_{t+2,\|}\| \ge (1+\eta(\sqrt{\rho\epsilon}-\nu))\big(\|\mathbf{x}_{t+1,\|}\| + (1-\theta)(\|\mathbf{x}_{t+1,\|}\| - \|\mathbf{x}_{t,\|}\|)\big) + \eta\|\Delta_\|\| \tag{85}$$

$$\ge (1+\eta\sqrt{\rho\epsilon})\big(\|\mathbf{x}_{t+1,\|}\| + (1-\theta)(\|\mathbf{x}_{t+1,\|}\| - \|\mathbf{x}_{t,\|}\|)\big) - \eta\|\Delta\|. \tag{86}$$

Note that since $\|\mathbf{x}_{t+1,\|}\|/\|\mathbf{x}_t\| \ge \alpha'_{\min}$, we have

$$\frac{\|\Delta\|}{\|\mathbf{z}_{t+1,\|}\|} \ge \frac{\|\Delta\|}{\alpha'_{\min}\|\mathbf{z}_{t+1}\|} = \rho r'/\alpha'_{\min}, \tag{87}$$

which leads to

$$\|\mathbf{x}_{t+2,\|}\| \ge (1+\eta\sqrt{\rho\epsilon}-\eta\rho r'/\alpha'_{\min})\big((2-\theta)\|\mathbf{x}_{t+1,\|}\| - (1-\theta)\|\mathbf{x}_{t,\|}\|\big). \tag{88}$$

Consider the sequences with recurrence that can be written as

$$\xi_{t+2} = (1+\kappa)\big((2-\theta)\xi_{t+1} - (1-\theta)\xi_t\big) \tag{89}$$

for some $\kappa > 0$. Its characteristic equation can be written as

$$x^2 - (1+\kappa)(2-\theta)x + (1+\kappa)(1-\theta) = 0, \tag{90}$$

whose roots satisfy

$$x = \frac{1+\kappa}{2}\Big((2-\theta) \pm \sqrt{(2-\theta)^2 - \frac{4(1-\theta)}{1+\kappa}}\Big), \tag{91}$$

indicating

$$\xi_t = \Big(\frac{1+\kappa}{2}\Big)^t\big(C_1(2-\theta+\mu)^t + C_2(2-\theta-\mu)^t\big), \tag{92}$$

where $\mu := \sqrt{(2-\theta)^2 - \frac{4(1-\theta)}{1+\kappa}}$, for constants C_1 and C_2 being

$$\begin{cases} C_1 = -\dfrac{2-\theta-\mu}{2\mu}\xi_0 + \dfrac{1}{(1+\kappa)\mu}\xi_1, \\ C_2 = \dfrac{2-\theta+\mu}{2\mu}\xi_0 - \dfrac{1}{(1+\kappa)\mu}\xi_1. \end{cases} \tag{93}$$

Then by the inequalities (84) and (88), as long as $\alpha'_k \ge \alpha'_{\min}$ for any $1 \le k \le t-1$, the values $\|\mathbf{x}_{t,\perp}\|$ and $\|\mathbf{x}_{t,\|}\|$ satisfy

$$\|\mathbf{x}_{t,\perp}\| \le \Big(-\frac{2-\theta-\mu_\perp}{2\mu_\perp}\xi_{0,\perp} + \frac{1}{(1+\kappa_\perp)\mu_\perp}\xi_{1,\perp}\Big)\cdot\Big(\frac{1+\kappa_\perp}{2}\Big)^t\cdot(2-\theta+\mu_\perp)^t \tag{94}$$

$$+\Big(\frac{2-\theta+\mu_\perp}{2\mu_\perp}\xi_{0,\perp} - \frac{1}{(1+\kappa_\perp)\mu_\perp}\xi_{1,\perp}\Big)\cdot\Big(\frac{1+\kappa_\perp}{2}\Big)^t\cdot(2-\theta-\mu_\perp)^t, \tag{95}$$

and

$$\|\mathbf{x}_{t,\|}\| \ge \Big(-\frac{2-\theta-\mu_\|}{2\mu_\|}\xi_{0,\|} + \frac{1}{(1+\kappa_\|)\mu_\|}\xi_{1,\|}\Big)\cdot\Big(\frac{1+\kappa_\|}{2}\Big)^t\cdot(2-\theta+\mu_\|)^t \tag{96}$$

$$+\Big(\frac{2-\theta+\mu_\|}{2\mu_\|}\xi_{0,\|} - \frac{1}{(1+\kappa_\|)\mu_\|}\xi_{1,\|}\Big)\cdot\Big(\frac{1+\kappa_\|}{2}\Big)^t\cdot(2-\theta-\mu_\|)^t, \tag{97}$$

where

$$\kappa_\perp = \eta\sqrt{\rho\epsilon}+2\eta\rho r', \qquad \xi_{0,\perp} = \|\mathbf{x}_{0,\perp}\|, \qquad \xi_{1,\perp} = (1+\kappa_\perp)\xi_{0,\perp}, \tag{98}$$

$$\kappa_\| = \eta\sqrt{\rho\epsilon}-\eta\rho r'/\alpha'_{\min}, \qquad \xi_{0,\|} = \|\mathbf{x}_{0,\|}\|, \qquad \xi_{1,\|} = (1+\kappa_\|)\xi_{0,\|}. \tag{99}$$

Further we can derive that

$$\|\mathbf{x}_{t,\perp}\| \le \|\mathbf{x}_{0,\perp}\|\cdot\Big(\frac{1+\kappa_\perp}{2}\Big)^t\cdot(2-\theta+\mu_\perp)^t, \tag{100}$$

and

$$\|\mathbf{x}_{t,\|}\| \geq \frac{\|\mathbf{x}_{0,\|}\|}{2} \cdot \Big(\frac{1+\kappa_\|}{2}\Big)^t \cdot (2-\theta+\mu_\|)^t. \tag{101}$$

Then we can observe that

$$\frac{\|\mathbf{x}_{t,\|}\|}{\|\mathbf{x}_{t,\perp}\|} \geq \frac{\|\mathbf{x}_{0,\|}\|}{2\|\mathbf{x}_{0,\perp}\|} \cdot \Big(\frac{1+\kappa_\|}{1+\kappa_\perp}\Big)^t \cdot \Big(\frac{2-\theta+\mu_\|}{2-\theta+\mu_\perp}\Big)^t, \tag{102}$$

where

$$\begin{aligned}
\frac{1+\kappa_\|}{1+\kappa_\perp} &\geq (1+\kappa_\|)(1-\kappa_\perp) && (103)\\
&\geq 1-(2+1/\alpha'_{\min})\eta\rho r' - \kappa_\|\kappa_\perp && (104)\\
&\geq 1-2\eta\rho r'/\alpha'_{\min}, && (105)
\end{aligned}$$

and

$$\begin{aligned}
\frac{2-\theta+\mu_\|}{2-\theta+\mu_\perp} &= \frac{1+\mu_\|/(2-\theta)}{1+\mu_\perp/(2-\theta)} && (106)\\
&= \frac{1+\sqrt{1-\frac{4(1-\theta)}{(1+\kappa_\|)(2-\theta)^2}}}{1+\sqrt{1-\frac{4(1-\theta)}{(1+\kappa_\perp)(2-\theta)^2}}} && (107)\\
&\geq \Big(1+\frac{1}{2-\theta}\sqrt{\frac{\theta^2+\kappa_\|(2-\theta)^2}{1+\kappa_\|}}\Big)\Big(1-\frac{1}{2-\theta}\sqrt{\frac{\theta^2+\kappa_\perp(2-\theta)^2}{1+\kappa_\perp}}\Big) && (108)\\
&\geq 1-\frac{2(\kappa_\perp-\kappa_\|)}{\theta} \geq 1-\frac{3\eta\rho r'}{\alpha'_{\min}\theta}, && (109)
\end{aligned}$$

by which we can derive that

$$\begin{aligned}
\frac{\|\mathbf{x}_{t,\|}\|}{\|\mathbf{x}_{t,\perp}\|} &\geq \frac{\|\mathbf{x}_{0,\|}\|}{2\|\mathbf{x}_{0,\perp}\|} \cdot \Big(1-\frac{4\rho r'}{\alpha'_{\min}\theta}\Big)^t && (110)\\
&\geq \frac{\|\mathbf{x}_{0,\|}\|}{2\|\mathbf{x}_{0,\perp}\|}(1-1/\mathscr{T}')^t && (111)\\
&\geq \frac{\|\mathbf{x}_{0,\|}\|}{2\|\mathbf{x}_{0,\perp}\|}\exp\Big(-\frac{t}{\mathscr{T}'-1}\Big) \geq \frac{\|\mathbf{x}_{0,\|}\|}{4\|\mathbf{x}_{0,\perp}\|}, && (112)
\end{aligned}$$

indicating

$$\alpha'_t = \frac{\|\mathbf{x}_{t,\|}\|}{\sqrt{\|\mathbf{x}_{t,\|}\|^2+\|\mathbf{x}_{t,\perp}\|^2}} \geq \frac{\|\mathbf{x}_{0,\|}\|}{8\|\mathbf{x}_{0,\perp}\|} \geq \alpha'_{\min}. \tag{113}$$

Hence, as long as $\alpha'_k \geq \alpha'_{\min}$ for any $1 \leq k \leq t-1$, we can also have $\alpha'_t \geq \alpha'_{\min}$ if $t \leq \mathscr{T}'$. Since we have $\alpha'_0 \geq \alpha'_{\min}$ and $\alpha'_1 \geq \alpha'_{\min}$, we can claim that $\alpha'_t \geq \alpha'_{\min}$ for any $t \leq \mathscr{T}'$ using recurrence. □

Equipped with Lemma 19, we are now ready to prove Proposition 5.

Proof. By Lemma 18, the unit vector $\hat{\mathbf{e}}$ in Line 7 of Algorithm 2 obtained after $\mathscr{T}'$ iterations equals to the output of Algorithm 4 starting from $\tilde{\mathbf{x}}$. Hence in this proof we consider the output of Algorithm 4 instead of the original Algorithm 2.

If $\lambda_n \leq -\sqrt{\rho\epsilon}/2$, Proposition 5 holds directly. Hence, we only need to prove the case where $\lambda_n > -\sqrt{\rho\epsilon}/2$, in which there exists some p' with

$$\lambda'_p \leq -\sqrt{\rho\epsilon}/2 < \lambda_{p+1}. \tag{114}$$

We use $\mathfrak{S}'_\|$, $\mathfrak{S}'_\perp$ to denote the subspace of $\mathbb{R}^n$ spanned by $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{p'}\}$, $\{\mathbf{u}_{p'+1}, \mathbf{u}_{p+2}, \ldots, \mathbf{u}_n\}$. Furthermore, we define $\mathbf{x}_{t,\|'} := \sum_{i=1}^{p'} \langle \mathbf{u}_i, \mathbf{x}_t \rangle \mathbf{u}_i$, $\mathbf{x}_{t,\perp'} := \sum_{i=p'}^{n} \langle \mathbf{u}_i, \mathbf{x}_t \rangle \mathbf{u}_i$, $\mathbf{v}_{t,\|'} := \sum_{i=1}^{p'} \langle \mathbf{u}_i, \mathbf{v}_t \rangle \mathbf{u}_i$, $\mathbf{v}_{t,\perp'} := \sum_{i=p'}^{n} \langle \mathbf{u}_i, \mathbf{v}_t \rangle \mathbf{u}_i$ respectively to denote the component of $\mathbf{x}'_t$ and $\mathbf{v}'_t$ in Algorithm 4 in the subspaces $\mathfrak{S}'_\|$, $\mathfrak{S}'_\perp$, and let $\alpha'_t := \|\mathbf{x}_{t,\|}\|/\|\mathbf{x}_t\|$. Consider the case where $\alpha'_0 \geq \sqrt{\frac{\pi}{n}}\delta_0$, which can be achieved with probability

$$\Pr\Big\{\alpha'_0 \geq \sqrt{\frac{\pi}{n}}\delta_0\Big\} \geq 1-\sqrt{\frac{\pi}{n}}\delta_0 \cdot \frac{\mathrm{Vol}(\mathbb{B}_0^{n-1}(1))}{\mathrm{Vol}(\mathbb{B}_0^n(1))} \geq 1-\sqrt{\frac{\pi}{n}}\delta_0 \cdot \sqrt{\frac{n}{\pi}} = 1-\delta_0, \tag{115}$$

we prove that there exists some t_0 with $1 \leq t_0 \leq \mathscr{T}'$ such that

$$\frac{\|\mathbf{x}_{t_0,\perp'}\|}{\|\mathbf{x}_{t_0}\|} \leq \frac{\sqrt{\rho\epsilon}}{8\ell}. \tag{116}$$

Assume the contrary, for any t with $1 \leq t \leq \mathscr{T}'$, we all have $\frac{\|\mathbf{x}_{t,\perp'}\|}{\|\mathbf{x}_t\|} > \frac{\sqrt{\rho\epsilon}}{8\ell}$ and $\frac{\|\mathbf{z}_{t,\perp'}\|}{\|\mathbf{z}_t\|} > \frac{\sqrt{\rho\epsilon}}{8\ell}$. Focus on the case where $\|\mathbf{x}_{t,\perp'}\|$, the component of $\mathbf{x}_t$ in subspace $\mathfrak{S}'_\perp$, achieves the largest value possible. Then in this case, we have the following recurrence formula:

$$\|\mathbf{x}_{t+2,\perp'}\| \leq (1+\eta\sqrt{\rho\epsilon}/2)\big(\|\mathbf{x}_{t+1,\perp'}\| + (1-\theta)(\|\mathbf{x}_{t+1,\perp'}\| - \|\mathbf{x}_{t,\perp'}\|)\big) + \eta\|\Delta_{\perp'}\|. \tag{117}$$

Since $\frac{\|\mathbf{z}_{k,\perp'}\|}{\|\mathbf{z}_k\|} \geq \frac{\sqrt{\rho\epsilon}}{8\ell}$ for any $1 \leq k \leq t+1$, we can derive that

$$\frac{\|\Delta_\perp\|}{\|\mathbf{x}_{t+1,\perp}\| + (1-\theta)(\|\mathbf{x}_{t+1,\perp}\| - \|\mathbf{x}_{t,\perp}\|)} \leq \frac{\|\Delta\|}{\|\mathbf{z}_{t,\perp'}\|} \leq \frac{2\rho r'}{\sqrt{\rho\epsilon}}, \tag{118}$$

which leads to

$$\begin{aligned}
\|\mathbf{x}_{t+2,\perp'}\| &\leq (1+\eta\sqrt{\rho\epsilon}/2)\big(\|\mathbf{x}_{t+1,\perp'}\| + (1-\theta)(\|\mathbf{x}_{t+1,\perp'}\| - \|\mathbf{x}_{t,\perp'}\|)\big) + \eta\|\Delta_{\perp'}\| && (119)\\
&\leq (1+\eta\sqrt{\rho\epsilon}/2 + 2\rho r'/\sqrt{\rho\epsilon})\big((2-\theta)\|\mathbf{x}_{t+1,\perp'}\| - (1-\theta)\|\mathbf{x}_{t,\perp'}\|\big). && (120)
\end{aligned}$$

Using similar characteristic function techniques shown in the proof of Lemma 19, it can be further derived that

$$\|\mathbf{x}_{t,\perp'}\| \leq \|\mathbf{x}_{0,\perp'}\| \cdot \Big(\frac{1+\kappa_{\perp'}}{2}\Big)^t \cdot (2-\theta+\mu_{\perp'})^t, \tag{121}$$

for $\kappa_{\perp'} = \eta\sqrt{\rho\epsilon}/2 + 2\rho r'/\sqrt{\rho\epsilon}$ and $\mu_{\perp'} = \sqrt{(2-\theta)^2 - \frac{4(1-\theta)}{1+\kappa_{\perp'}}}$, given $\frac{\|\mathbf{x}_{k,\perp'}\|}{\|\mathbf{x}_k\|} \geq \frac{\sqrt{\rho\epsilon}}{8\ell}$ and $\frac{\|\mathbf{z}_{k,\perp'}\|}{\|\mathbf{z}_k\|} \geq \frac{\sqrt{\rho\epsilon}}{8\ell}$ for any $1 \leq k \leq t-1$. Due to Lemma 19,

$$\alpha'_t \geq \alpha'_{\min} = \frac{\delta_0}{8}\sqrt{\frac{\pi}{n}}, \qquad \forall 1 \leq t \leq \mathscr{T}'. \tag{122}$$

and it is demonstrated in the proof of Lemma 19 that,

$$\|\mathbf{x}_{t,\|}\| \geq \frac{\|\mathbf{x}_{0,\|}\|}{2} \cdot \Big(\frac{1+\kappa_\|}{2}\Big)^t \cdot (2-\theta+\mu_\|)^t, \qquad \forall 1 \leq t \leq \mathscr{T}', \tag{123}$$

for $\kappa_\| = \eta\sqrt{\rho\epsilon} - \eta\rho r'/\alpha'_{\min}$ and $\mu_\| = \sqrt{(2-\theta)^2 - \frac{4(1-\theta)}{1+\kappa_\|}}$. Observe that

$$\begin{aligned}
\frac{\|\mathbf{x}_{\mathscr{T}',\perp'}\|}{\|\mathbf{x}_{\mathscr{T}',\|}\|} &\leq \frac{2\|\mathbf{x}_{0,\perp'}\|}{\|\mathbf{x}_{0,\|}\|} \cdot \Big(\frac{1+\kappa_{\perp'}}{1+\kappa_\|}\Big)^{\mathscr{T}'} \cdot \Big(\frac{2-\theta+\mu_{\perp'}}{2-\theta+\mu_\|}\Big)^{\mathscr{T}'} && (124)\\
&\leq \frac{2}{\delta_0}\sqrt{\frac{n}{\pi}}\Big(\frac{1+\kappa_{\perp'}}{1+\kappa_\|}\Big)^{\mathscr{T}'} \cdot \Big(\frac{2-\theta+\mu_{\perp'}}{2-\theta+\mu_\|}\Big)^{\mathscr{T}'}, && (125)
\end{aligned}$$

where

$$\frac{1+\kappa_{\perp'}}{1+\kappa_\|} \leq \frac{1}{1+(\kappa_\| - \kappa_{\perp'})} = 1 - \frac{1}{\eta\sqrt{\rho\epsilon}/2 + \rho r'(\eta/\alpha_{\min'} + 2/\sqrt{\rho\epsilon})} \leq 1 - \frac{\eta\sqrt{\rho\epsilon}}{4}, \tag{126}$$

and

$$\begin{aligned}
\frac{2-\theta+\mu_{\perp'}}{2-\theta+\mu_\|} &= \frac{1+\sqrt{1-\frac{4(1-\theta)}{(1+\kappa_{\perp'})(2-\theta)^2}}}{1+\sqrt{1-\frac{4(1-\theta)}{(1+\kappa_\|)(2-\theta)^2}}} && (127)\\
&\leq \frac{1}{1+\Big(\sqrt{1-\frac{4(1-\theta)}{(1+\kappa_{\perp'})(2-\theta)^2}} - \sqrt{1-\frac{4(1-\theta)}{(1+\kappa_\|)(2-\theta)^2}}\Big)} && (128)\\
&\leq 1 - \frac{\kappa_\| - \kappa_{\perp'}}{\theta} && (129)\\
&\leq 1 - \frac{\eta\sqrt{\rho\epsilon}}{4\theta} = 1 - \frac{(\rho\epsilon)^{1/4}}{16\sqrt{\ell}}. && (130)
\end{aligned}$$

Hence,

$$\frac{\|\mathbf{x}_{\mathscr{T}',\perp'}\|}{\|\mathbf{x}_{\mathscr{T}',\|}\|} \leq \frac{2}{\delta_0}\sqrt{\frac{n}{\pi}}\Big(1 - \frac{(\rho\epsilon)^{1/4}}{16\sqrt{\ell}}\Big)^{\mathscr{T}'} \leq \frac{\sqrt{\rho\epsilon}}{8\ell}. \tag{131}$$

Since $\|\mathbf{x}_{\mathscr{T}',\|}\| \leq \|\mathbf{x}_{\mathscr{T}'}\|$, we have $\frac{\|\mathbf{x}_{\mathscr{T}',\perp'}\|}{\|\mathbf{x}_{\mathscr{T}'}\|} \leq \frac{\sqrt{\rho\epsilon}}{8\ell}$, contradiction. Hence, there here exists some t_0 with $1 \leq t_0 \leq \mathscr{T}'$ such that $\frac{\|\mathbf{x}_{t_0,\perp'}\|}{\|\mathbf{x}_{t_0}\|} \leq \frac{\sqrt{\rho\epsilon}}{8\ell}$. Consider the normalized vector $\hat{\mathbf{e}} = \mathbf{x}_{t_0}/r$, we use $\hat{\mathbf{e}}_{\perp'}$ and $\hat{\mathbf{e}}_{\|'}$ to separately denote the component of $\hat{\mathbf{e}}$ in $\mathfrak{S}'_{\perp}$ and $\mathfrak{S}'_{\|}$. Then, $\|\hat{\mathbf{e}}_{\perp'}\| \leq \sqrt{\rho\epsilon}/(8\ell)$ whereas $\|\hat{\mathbf{e}}_{\|'}\| \geq 1 - \rho\epsilon/(8\ell)^2$. Then,

$$\hat{\mathbf{e}}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}} = (\hat{\mathbf{e}}_{\perp'} + \hat{\mathbf{e}}_{\|'})^T\mathcal{H}(\mathbf{0})(\hat{\mathbf{e}}_{\perp'} + \hat{\mathbf{e}}_{\|'}), \tag{132}$$

since $\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\perp'} \in \mathfrak{S}'_{\perp}$ and $\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\|'} \in \mathfrak{S}'_{\|}$, it can be further simplified to

$$\hat{\mathbf{e}}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}} = \hat{\mathbf{e}}_{\perp'}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\perp'} + \hat{\mathbf{e}}_{\|'}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\|'}, \tag{133}$$

Due to the ℓ-smoothness of the function, all eigenvalue of the Hessian matrix has its absolute value upper bounded by ℓ. Hence,

$$\hat{\mathbf{e}}_{\perp'}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\perp'} \leq \ell\|\hat{\mathbf{e}}_{\perp'}^T\|_2^2 = \frac{\rho\epsilon}{64\ell^2}. \tag{134}$$

Further according to the definition of $\mathfrak{S}_{\|}$, we have

$$\hat{\mathbf{e}}_{\|'}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\|'} \leq -\frac{\sqrt{\rho\epsilon}}{2}\|\hat{\mathbf{e}}_{\|'}\|^2. \tag{135}$$

Combining these two inequalities together, we can obtain

$$\hat{\mathbf{e}}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}} = \hat{\mathbf{e}}_{\perp}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\perp'} + \hat{\mathbf{e}}_{\|'}^T\mathcal{H}(\mathbf{0})\hat{\mathbf{e}}_{\|'} \leq -\frac{\sqrt{\rho\epsilon}}{2}\|\hat{\mathbf{e}}_{\|'}\|^2 + \frac{\rho\epsilon}{64\ell^2} \leq -\frac{\sqrt{\rho\epsilon}}{4}. \tag{136}$$

□

C Proof details of escaping from saddle points by negative curvature finding

C.1 Algorithms for escaping from saddle points using negative curvature finding

In this subsection, we first present algorithm for escaping from saddle points using Algorithm 1 as Algorithm 5.

Algorithm 5: Perturbed Gradient Descent with Negative Curvature Finding

Input: $\mathbf{x}_0 \in \mathbb{R}^n$;
for $t = 0, 1, ..., T$ **do**
  if $\|\nabla f(\mathbf{x}_t)\| \leq \epsilon$ **then**
    $\hat{\mathbf{e}} \leftarrow$NegativeCurvatureFinding$(\mathbf{x}_t, r, \mathscr{T})$;
    $\mathbf{x}_t \leftarrow \mathbf{x}_t - \frac{f'_{\hat{\mathbf{e}}}(\mathbf{x}_0)}{4|f'_{\hat{\mathbf{e}}}(\mathbf{x}_0)|}\sqrt{\frac{\epsilon}{\rho}} \cdot \hat{\mathbf{e}}$;
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{\ell}\nabla f(\mathbf{x}_t)$;

Observe that Algorithm 5 and Algorithm 2 are similar to perturbed gradient descent and perturbed accelerated gradient descent but the uniform perturbation step is replaced by our negative curvature finding algorithms. One may wonder that Algorithm 5 seems to involve nested loops since negative curvature finding algorithm are contained in the primary loop, contradicting our previous claim that Algorithm 5 only contains a single loop. But actually, Algorithm 5 contains only two operations: gradient descents and one perturbation step, the same as operations outside the negative curvature finding algorithms. Hence, Algorithm 5 is essentially single-loop algorithm, and we count their iteration number as the total number of gradient calls.

C.2 Proof details of escaping saddle points using Algorithm 1

In this subsection, we prove:

Theorem 20. *For any $\epsilon > 0$ and $0 < \delta \leq 1$, Algorithm 5 with parameters chosen in Proposition 3 satisfies that at least $1/4$ of its iterations will be ϵ-approximate second-order stationary point, using*

$$\tilde{O}\Big(\frac{(f(\mathbf{x}_0) - f^*)}{\epsilon^2} \cdot \log n\Big)$$

iterations, with probability at least $1 - \delta$, where f^* is the global minimum of f.

Proof. Let the parameters be chosen according to (2), and set the total step number T to be:

$$T = \max\left\{\frac{8\ell(f(\mathbf{x_0}) - f^*)}{\epsilon^2}, 768(f(\mathbf{x_0}) - f^*) \cdot \sqrt{\frac{\rho}{\epsilon^3}}\right\}, \tag{137}$$

similar to the perturbed gradient descent algorithm [21, Algorithm 4]. We first assume that for each $\mathbf{x}_t$ we apply negative curvature finding (Algorithm 1) with δ_0 contained in the parameters be chosen as

$$\delta_0 = \frac{1}{384(f(\mathbf{x}_0 - f^*)}\sqrt{\frac{\epsilon^3}{\rho}}\delta, \tag{138}$$

we can successfully obtain a unit vector $\hat{\mathbf{e}}$ with $\hat{\mathbf{e}}^T\mathcal{H}\hat{\mathbf{e}} \leq -\sqrt{\rho\epsilon}/4$, as long as $\lambda_{\min}(\mathcal{H}(\mathbf{x}_t)) \leq -\sqrt{\rho\epsilon}$. The error probability of this assumption is provided later.

Under this assumption, Algorithm 1 can be called for at most $384(f(\mathbf{x_0}) - f^*)\sqrt{\frac{\rho}{\epsilon^3}} \leq \frac{T}{2}$ times, for otherwise the function value decrease will be greater than $f(\mathbf{x_0}) - f^*$, which is not possible. Then, the error probability that some calls to Algorithm 1 fails is upper bounded by

$$384(f(\mathbf{x_0}) - f^*)\sqrt{\frac{\rho}{\epsilon^3}} \cdot \delta_0 = \delta. \tag{139}$$

For the rest of iterations in which Algorithm 1 is not called, they are either large gradient steps, $\|\nabla f(\mathbf{x}_t)\| \geq \epsilon$, or ϵ-approximate second-order stationary points. Within them, we know that the number of large gradient steps cannot be more than $T/4$ because otherwise, by Lemma 10 in Appendix A:

$$f(\mathbf{x}_T) \leq f(\mathbf{x}_0) - T\eta\epsilon^2/8 < f^*,$$

a contradiction. Therefore, we conclude that at least $T/4$ of the iterations must be ϵ-approximate second-order stationary points, with probability at least $1 - \delta$.

The number of iterations can be viewed as the sum of two parts, the number of iterations needed for gradient descent, denoted by T_1, and the number of iterations needed for negative curvature finding, denoted by T_2. with probability at least $1 - \delta$,

$$T_1 = T = \tilde{O}\Big(\frac{(f(\mathbf{x}_0) - f^*)}{\epsilon^2}\Big). \tag{140}$$

As for T_2, with probability at least $1 - \delta$, Algorithm 1 is called for at most $384(f(\mathbf{x_0}) - f^*)\sqrt{\frac{\rho}{\epsilon^3}}$ times, and by Proposition 3 it takes $\tilde{O}\Big(\frac{\log n}{\sqrt{\rho\epsilon}}\Big)$ iterations each time. Hence,

$$T_2 = 384(f(\mathbf{x_0}) - f^*)\sqrt{\frac{\rho}{\epsilon^3}} \cdot \tilde{O}\Big(\frac{\log n}{\sqrt{\rho\epsilon}}\Big) = \tilde{O}\Big(\frac{(f(\mathbf{x}_0) - f^*)}{\epsilon^2} \cdot \log n\Big). \tag{141}$$

As a result, the total iteration number $T_1 + T_2$ is

$$\tilde{O}\Big(\frac{(f(\mathbf{x}_0) - f^*)}{\epsilon^2} \cdot \log n\Big). \tag{142}$$

□

C.3 Proof details of escaping saddle points using Algorithm 2

We first present here the Negative Curvature Exploitation algorithm proposed in proposed in [22, Algorithm 3] appearing in Line 16 of Algorithm 2:

Algorithm 6: Negative Curvature Exploitation$(\mathbf{x}_t, \mathbf{v}_t, s)$

if $\|\mathbf{v}_t\| \geq s$ **then**
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t$;
else
  $\xi = s \cdot \mathbf{v}_t/\|\mathbf{v}\|_t$;
  $\mathbf{x}_t \leftarrow \text{argmin}_{\mathbf{x}\in\{\mathbf{x}_t+\xi, \mathbf{x}_t-\xi\}} f(\mathbf{x})$;
Output $(\mathbf{z}_{t+1}, \mathbf{0})$.

Now, we give the full version of Theorem 7 as follows:

Theorem 21. *Suppose that the function f is ℓ-smooth and ρ-Hessian Lipschitz. For any $\epsilon > 0$ and a constant $0 < \delta \leq 1$, we choose the parameters appearing in Algorithm 2 as follows:*

$$\delta_0 = \frac{\delta}{384\Delta_f}\sqrt{\frac{\epsilon^3}{\rho}}, \qquad \mathscr{T}' = \frac{32\sqrt{\ell}}{(\rho\epsilon)^{1/4}}\log\Big(\frac{\ell}{\delta_0}\sqrt{\frac{n}{\rho\epsilon}}\Big), \qquad \zeta = \frac{\ell}{\sqrt{\rho\epsilon}}, \tag{143}$$

$$r' = \frac{\delta_0\epsilon}{32}\sqrt{\frac{\pi}{\rho n}}, \qquad \eta = \frac{1}{4\ell}, \qquad \theta = \frac{1}{4\sqrt{\zeta}}, \tag{144}$$

$$\mathscr{E} = \sqrt{\frac{\epsilon^3}{\rho}}\cdot c_A^{-7}, \qquad \gamma = \frac{\theta^2}{\eta}, \qquad s = \frac{\gamma}{4\rho}, \tag{145}$$

where $\Delta_f := f(\mathbf{x}_0) - f^*$ and f^* is the global minimum of f, and the constant c_A is chosen large enough to satisfy both the condition in Lemma 15 and $c_A \geq (384)^{1/7}$. Then, Algorithm 2 satisfies that at least one of the iterations $\mathbf{z}_t$ will be an ϵ-approximate second-order stationary point in

$$\tilde{O}\Big(\frac{(f(\mathbf{x}_0) - f^*)}{\epsilon^{1.75}}\cdot\log n\Big) \tag{146}$$

iterations, with probability at least $1 - \delta$.

Proof. Set the total step number T to be:

$$T = \max\left\{\frac{4\Delta_f(\tilde{\mathscr{T}} + \mathscr{T}')}{\mathscr{E}}, 768\Delta_f\mathscr{T}'\sqrt{\frac{\rho}{\epsilon^3}}\right\} = \tilde{O}\Big(\frac{(f(\mathbf{x}_0) - f^*)}{\epsilon^{1.75}}\cdot\log n\Big), \tag{147}$$

where $\tilde{\mathscr{T}} = \sqrt{\zeta}\cdot c_A$ as defined in Lemma 15, similar to the perturbed accelerated gradient descent algorithm [22, Algorithm 2]. We first assert that for each iteration $\mathbf{x}_t$ that a uniform perturbation is added, after $\mathscr{T}'$ iterations we can successfully obtain a unit vector $\hat{\mathbf{e}}$ with $\hat{\mathbf{e}}^T\mathcal{H}\hat{\mathbf{e}} \leq -\sqrt{\rho\epsilon}/4$, as long as $\lambda_{\min}(\mathcal{H}(\mathbf{x}_t)) \leq -\sqrt{\rho\epsilon}$. The error probability of this assumption is provided later.

Under this assumption, the uniform perturbation can be called for at most $384(f(\mathbf{x_0}) - f^*)\sqrt{\frac{\rho}{\epsilon^3}}$ times, for otherwise the function value decrease will be greater than $f(\mathbf{x_0}) - f^*$, which is not possible. Then, the probability that at least one negative curvature finding subroutine after uniform perturbation fails is upper bounded by

$$384(f(\mathbf{x_0}) - f^*)\sqrt{\frac{\rho}{\epsilon^3}}\cdot\delta_0 = \delta. \tag{148}$$

For the rest of steps which is not within $\mathscr{T}'$ steps after uniform perturbation, they are either large gradient steps, $\|\nabla f(\mathbf{x}_t)\| \geq \epsilon$, or ϵ-approximate second-order stationary points. Next, we demonstrate that at least one of these steps is an ϵ-approximate stationary point.

Assume the contrary. We use $N_{\tilde{\mathscr{T}}}$ to denote the number of disjoint time periods with length larger than $\tilde{\mathscr{T}}$ containing only large gradient steps and do not contain any step within $\mathscr{T}'$ steps after uniform perturbation. Then, it satisfies

$$N_{\tilde{\mathscr{T}}} \geq \frac{T}{2(\tilde{\mathscr{T}} + \mathscr{T}')} - 384\Delta_f\sqrt{\frac{\rho}{\epsilon^3}} \geq (2c_A^7 - 384)\Delta_f\sqrt{\frac{\rho}{\epsilon^3}} \geq \frac{\Delta_f}{\mathscr{E}}. \tag{149}$$

From Lemma 15, during these time intervals the Hamiltonian will decrease in total at least $N_{\tilde{\mathscr{T}}}\cdot\mathscr{E} = \Delta_f$, which is impossible due to Lemma 16, the Hamiltonian decreases monotonically for every step except for the $\mathscr{T}'$ steps after uniform perturbation, and the overall decrease cannot be greater than Δ_f, a contradiction. Therefore, we conclude that at least one of the iterations must be an ϵ-approximate second-order stationary point, with probability at least $1 - \delta$. □

D Proofs of the stochastic setting

D.1 Proof details of negative curvature finding using stochastic gradients

In this subsection, we demonstrate that Algorithm 3 can find a negative curvature efficiently. Specifically, we prove the following proposition:

Proposition 22. *Suppose the function $f\colon \mathbb{R}^n \to \mathbb{R}$ is ℓ-smooth and ρ-Hessian Lipschitz. For any $0 < \delta < 1$, we specify our choice of parameters and constants we use as follows:*

$$\mathscr{T}_s = \frac{8\ell}{\sqrt{\rho\epsilon}}\cdot\log\Big(\frac{\ell n}{\delta\sqrt{\rho\epsilon}}\Big), \qquad \iota = 10\log\Big(\frac{n\mathscr{T}_s^2}{\delta}\log\Big(\frac{\sqrt{n}}{\eta r_s}\Big)\Big), \tag{150}$$

$$r_s = \frac{\delta}{480\rho n\mathscr{T}_s}\sqrt{\frac{\rho\epsilon}{\iota}}, \qquad m = \frac{160(\ell + \tilde{\ell})}{\delta\sqrt{\rho\epsilon}}\sqrt{\mathscr{T}_s\iota}, \tag{151}$$

Then for any point $\tilde{\mathbf{x}} \in \mathbb{R}^n$ satisfying $\lambda_{\min}(\mathcal{H}(\tilde{\mathbf{x}})) \leq -\sqrt{\rho\epsilon}$, with probability at least $1-3\delta$, Algorithm 3 outputs a unit vector $\hat{\mathbf{e}}$ satisfying

$$\hat{\mathbf{e}}^T \mathcal{H}(\tilde{\mathbf{x}}) \hat{\mathbf{e}} \leq -\frac{\sqrt{\rho\epsilon}}{4}, \tag{152}$$

where $\mathcal{H}$ stands for the Hessian matrix of function f, using $O(m \cdot \mathscr{T}_s) = \tilde{O}\Big(\frac{\log^2 n}{\delta \epsilon^{1/2}}\Big)$ iteartions.

Similarly to Algorithm 1 and Algorithm 2, the renormalization step Line 6 in Algorithm 3 only guarantees that the value $\|\mathbf{y}_t\|$ would not scales exponentially during the algorithm, and does not affect the output. We thus introduce the following Algorithm 7, which is the no-renormalization version of Algorithm 3 that possess the same output and a simpler structure. Hence in this subsection, we analyze Algorithm 7 instead of Algorithm 3.

Algorithm 7: Stochastic Negative Curvature Finding without Renormalization($\tilde{\mathbf{x}}, r_s, \mathscr{T}_s, m$).

1 $\mathbf{z}_0 \leftarrow 0$;
2 **for** $t = 1, ..., \mathscr{T}_s$ **do**
3 Sample $\big\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}\big\} \sim \mathcal{D}$;
4 $\mathbf{g}(\mathbf{z}_{t-1}) \leftarrow \frac{\|\mathbf{z}_{t-1}\|}{r_s} \cdot \frac{1}{m} \sum_{j=1}^m \Big(\mathbf{g}\Big(\tilde{\mathbf{x}} + \frac{r_s}{\|\mathbf{z}_{t-1}\|}\mathbf{z}_{t-1}; \theta^{(j)}\Big) - \mathbf{g}(\tilde{\mathbf{x}}; \theta^{(j)})\Big)$;
5 $\mathbf{z}_t \leftarrow \mathbf{z}_{t-1} - \frac{1}{\ell}(\mathbf{g}(\mathbf{z}_{t-1}) + \xi_t), \qquad \xi_t \sim \mathcal{N}\Big(0, \frac{r_s^2}{d} I\Big)$;
6 **Output** $\mathbf{z}_{\mathscr{T}} / \|\mathbf{z}_{\mathscr{T}}\|$.

Without loss of generality, we assume $\tilde{\mathbf{x}} = \mathbf{0}$ by shifting $\mathbb{R}^n$ such that $\tilde{\mathbf{x}}$ is mapped to $\mathbf{0}$. As argued in the proof of Proposition 3, $\mathcal{H}(\mathbf{0})$ admits the following following eigen-decomposition:

$$\mathcal{H}(\mathbf{0}) = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \tag{153}$$

where the set $\{\mathbf{u}_i\}_{i=1}^n$ forms an orthonormal basis of $\mathbb{R}^n$. Without loss of generality, we assume the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ corresponding to $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ satisfy

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n, \tag{154}$$

where $\lambda_1 \leq -\sqrt{\rho\epsilon}$. If $\lambda_n \leq -\sqrt{\rho\epsilon}/2$, Proposition 22 holds directly. Hence, we only need to prove the case where $\lambda_n > -\sqrt{\rho\epsilon}/2$, where there exists some p, p' with

$$\lambda_p \leq -\sqrt{\rho\epsilon} < \lambda_{p+1}, \quad \lambda_{p'} \leq -\sqrt{\rho\epsilon}/2 < \lambda_{p'+1}. \tag{155}$$

Notation: Throughout this subsection, let $\tilde{\mathcal{H}} := \mathcal{H}(\tilde{\mathbf{x}})$. Use $\mathfrak{S}_\|$, $\mathfrak{S}_\perp$ to separately denote the subspace of $\mathbb{R}^n$ spanned by $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$, $\{\mathbf{u}_{p+1}, \mathbf{u}_{p+2}, \ldots, \mathbf{u}_n\}$, and use $\mathfrak{S}'_\|$, $\mathfrak{S}'_\perp$ to denote the subspace of $\mathbb{R}^n$ spanned by $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{p'}\}$, $\{\mathbf{u}_{p'+1}, \mathbf{u}_{p+2}, \ldots, \mathbf{u}_n\}$. Furthermore, define $\mathbf{z}_{t,\|} := \sum_{i=1}^p \langle \mathbf{u}_i, \mathbf{z}_t \rangle \mathbf{u}_i$, $\mathbf{z}_{t,\perp} := \sum_{i=p}^n \langle \mathbf{u}_i, \mathbf{z}_t \rangle \mathbf{u}_i$, $\mathbf{z}_{t,\|'} := \sum_{i=1}^{p'} \langle \mathbf{u}_i, \mathbf{z}_t \rangle \mathbf{u}_i$, $\mathbf{z}_{t,\perp'} := \sum_{i=p'}^n \langle \mathbf{u}_i, \mathbf{z}_t \rangle \mathbf{u}_i$ respectively to denote the component of $\mathbf{z}_t$ in Line 5 of Algorithm 7 in the subspaces $\mathfrak{S}_\|$, $\mathfrak{S}_\perp$, $\mathfrak{S}'_\|$, $\mathfrak{S}'_\perp$, and let $\gamma = -\lambda_1$.

To prove Proposition 22, we first introduce the following lemma:

Lemma 23. *Under the setting of Proposition 22, for any point $\tilde{\mathbf{x}} \in \mathbb{R}^n$ satisfying $\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}})) \leq -\sqrt{\rho\epsilon}$, with probability at least $1-3\delta$, Algorithm 3 outputs a unit vector $\hat{\mathbf{e}}$ satisfying*

$$\|\hat{\mathbf{e}}_{\perp'}\| := \Big\| \sum_{i=p'}^n \langle \mathbf{u}_i, \hat{\mathbf{e}} \rangle \mathbf{u}_i \Big\| \leq \frac{\sqrt{\rho\epsilon}}{8\ell} \tag{156}$$

using $O(m \cdot \mathscr{T}_s) = \tilde{O}\Big(\frac{\log^2 n}{\delta \epsilon^{1/2}}\Big)$ iteartions.

D.1.1 Proof of Lemma 23

In the proof of Lemma 23, we consider the worst case, where $\lambda_1 = -\gamma = -\sqrt{\rho\epsilon}$ is the only eigenvalue less than $-\sqrt{\rho\epsilon}/2$, and all other eigenvalues equal to $-\sqrt{\rho\epsilon}/2 + \nu$ for an arbitrarily small constant ν. Under this scenario, the component $\mathbf{z}_{t,\perp'}$ is as small as possible at each time step.

The following lemma characterizes the dynamics of Algorithm 7:

Lemma 24. *Consider the sequence $\{\mathbf{z}_i\}$ and let $\eta = 1/\ell$. Further, for any $0 \le t \le \mathscr{T}_s$ we define*

$$\zeta_t := \mathbf{g}(\mathbf{z}_{t-1}) - \frac{\|\mathbf{z}_t\|}{r_s}\Big(\nabla f\Big(\tilde{\mathbf{x}} + \frac{r_s}{\|\mathbf{z}_t\|}\mathbf{z}_t\Big) - \nabla f(\tilde{\mathbf{x}})\Big), \tag{157}$$

to be the errors caused by the stochastic gradients. Then $\mathbf{z}_t = -\mathbf{q}_h(t) - \mathbf{q}_{sg}(t) - \mathbf{q}_p(t)$, where:

$$\mathbf{q}_h(t) := \eta\sum_{\tau=0}^{t-1}(I - \eta\tilde{\mathcal{H}})^{t-1-\tau}\Delta_\tau\hat{\mathbf{z}}_\tau, \tag{158}$$

for $\Delta_\tau = \int_0^1 \mathcal{H}_f\big(\psi\frac{r_s}{\|\mathbf{z}_\tau\|}\mathbf{z}_\tau\big)\mathrm{d}\psi - \tilde{\mathcal{H}}$, and

$$\mathbf{q}_{sg}(t) := \eta\sum_{\tau=0}^{t-1}(I - \eta\tilde{\mathcal{H}})^{t-1-\tau}\zeta_\tau, \quad \mathbf{q}_p(t) := \eta\sum_{\tau=0}^{t-1}(I - \eta\tilde{\mathcal{H}})^{t-1-\tau}\xi_\tau. \tag{159}$$

Proof. Without loss of generality we assume $\tilde{\mathbf{x}} = \mathbf{0}$. The update formula for $\mathbf{z}_t$ can be written as

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta\Big(\frac{\|\mathbf{z}_t\|}{r_s}\Big(\nabla f\Big(\frac{r_s}{\|\mathbf{z}_t\|}\mathbf{z}_t\Big) - \nabla f(\mathbf{0})\Big) + \zeta_t + \xi_t\Big), \tag{160}$$

where

$$\frac{\|\mathbf{z}_t\|}{r_s}\Big(\nabla f\Big(\frac{r_s}{\|\mathbf{z}_t\|}\mathbf{z}_t\Big) - \nabla f(\mathbf{0})\Big) = \frac{\|\mathbf{z}_t\|}{r_s}\int_0^1 \mathcal{H}_f\Big(\psi\frac{r_s}{\|\mathbf{z}_t\|}\mathbf{z}_t\Big)\frac{r_s}{\|\mathbf{z}_t\|}\mathbf{z}_t\mathrm{d}\psi = (\tilde{\mathcal{H}} + \Delta_t)\mathbf{z}_t, \tag{161}$$

indicating

$$\mathbf{z}_{t+1} = (I - \eta\tilde{\mathcal{H}})\mathbf{x}_t - \eta(\Delta_t\mathbf{z}_t + \zeta_t + \xi_t) \tag{162}$$

$$= -\eta\sum_{\tau=0}^{t}(I - \eta\tilde{\mathcal{H}})^{t-\tau}(\Delta_t\mathbf{z}_t + \zeta_t + \xi_t), \tag{163}$$

which finishes the proof. □

At a high level, under our parameter choice in Proposition 22, $\mathbf{q}_p(t)$ is the dominating term controlling the dynamics, and $\mathbf{q}_h(t) + \mathbf{q}_{sg}(t)$ will be small compared to $\mathbf{q}_p(t)$. Quantitatively, this is shown in the following lemma:

Lemma 25. *Under the setting of Proposition 22 while using the notation in Lemma 12 and Lemma 24, we have*

$$\Pr\Big(\|\mathbf{q}_h(t) + \mathbf{q}_{sg}(t)\| \le \frac{\beta(t)\eta r_s\delta}{20\sqrt{n}}\cdot\frac{\sqrt{\rho\epsilon}}{16\ell},\ \forall t \le \mathscr{T}_s\Big) \ge 1 - \delta, \tag{164}$$

where $-\gamma := \lambda_{\min}(\tilde{\mathcal{H}}) = -\sqrt{\rho\epsilon}$.

Proof. Divide $\mathbf{q}_p(t)$ into two parts:

$$\mathbf{q}_{p,1}(t) := \langle\mathbf{q}_p(t), \mathbf{u}_1\rangle\,\mathbf{u}_1, \tag{165}$$

and

$$\mathbf{q}_{p,\perp'}(t) := \mathbf{q}_p(t) - \mathbf{q}_{p,1}(t). \tag{166}$$

Then by Lemma 13, we have

$$\Pr\Big(\|\mathbf{q}_{p,1}(t)\| \le \frac{\beta(t)\eta r_s}{\sqrt{n}}\cdot\sqrt{\iota}\Big) \ge 1 - 2e^{-\iota}, \tag{167}$$

and

$$\Pr\Big(\|\mathbf{q}_{p,1}(t)\| \ge \frac{\beta(t)\eta r_s}{20\sqrt{n}}\cdot\delta\Big) \ge 1 - \delta/4. \tag{168}$$

Similarly,

$$\Pr\Big(\|\mathbf{q}_{p,\perp'}(t)\| \le \beta_{\perp'}(t)\eta r_s\cdot\sqrt{\iota}\Big) \ge 1 - 2e^{-\iota}, \tag{169}$$

and

$$\Pr\Big(\|\mathbf{q}_{p,\perp'}(t)\| \ge \frac{\beta_{\perp'}(t)\eta r_s}{20}\cdot\delta\Big) \ge 1 - \delta/4, \tag{170}$$

where $\beta_{\perp'}(t) := \frac{(1+\eta\gamma/2)^t}{\sqrt{\eta\gamma}}$. Set $t_{\perp'} := \frac{\log n}{\eta\gamma}$. Then for all $\tau \le t_{\perp'}$, we have

$$\frac{\beta(\tau)}{\beta_{\perp'}(\tau)} \le \sqrt{n}, \tag{171}$$

which further leads to

$$\Pr\Big(\|\mathbf{q}_{p,\perp'}(\tau)\| \le 2\beta_{\perp'}(t)\eta r_s \cdot \sqrt{\iota}\Big) \ge 1 - 2e^{-\iota}. \tag{172}$$

Next, we use induction to prove that the following inequality holds for all $t \le t_{\perp'}$:

$$\Pr\Big(\|\mathbf{q}_h(\tau) + \mathbf{q}_{sg}(\tau)\| \le \beta_{\perp'}(\tau)\eta r_s \cdot \frac{\delta}{20},\ \forall \tau \le t\Big) \ge 1 - 10nt^2 \log\Big(\frac{\sqrt{n}}{\eta r_s}\Big)e^{-\iota}. \tag{173}$$

For the base case $t = 0$, the claim holds trivially. Suppose it holds for all $\tau \le t$ for some t. Then due to Lemma 13, with probability at least $1 - 2t_{\perp'}e^{-\iota}$, we have

$$\|\mathbf{z}_t\| \le \eta\|\mathbf{q}_p(t)\| + \eta\|\mathbf{q}_h(t) + \mathbf{q}_{sg}(t)\| \le 3\beta_{\perp'}(\tau)\eta r_s \cdot \sqrt{\iota}. \tag{174}$$

By the Hessian Lipschitz property, Δ_τ satisfies:

$$\|\Delta_\tau\| \le \rho r_s. \tag{175}$$

Hence,

$$\|\mathbf{q}_h(t+1)\| \le \big\|\eta\sum_{\tau=0}^{t}(I - \eta\tilde{\mathcal{H}})^{t-\tau}\Delta_\tau \mathbf{z}_\tau\big\| \tag{176}$$

$$\le \eta\rho r_s \sum_{\tau=0}^{t}(I - \eta\tilde{\mathcal{H}})^{t-\tau}\|\mathbf{z}_\tau\| \tag{177}$$

$$\le (\eta\rho r_s n\mathscr{T}_s) \cdot (3\beta_{\perp'}(t)\eta r_s) \cdot \sqrt{\iota} \tag{178}$$

$$\le \frac{\beta_{\perp'}(t+1)\eta r_s}{10\sqrt{n}} \cdot \frac{\delta\sqrt{\rho\epsilon}}{16\ell}. \tag{179}$$

As for $\mathbf{q}_{sg}(t)$, note that $\hat{\zeta}_\tau|\mathcal{F}_{\tau-1} \sim \mathrm{nSG}((\ell+\tilde{\ell})\|\hat{\mathbf{z}}_\tau\|/\sqrt{m})$. By applying Lemma 14 with $b = \alpha^2(t)\cdot\eta^2(\ell+\tilde{\ell})^2/m$ and $b = \alpha^2(t)\eta^2(\ell+\tilde{\ell})^2\eta^2 r_s^2/(mn)$, with probability at least

$$1 - 4n \cdot \log\Big(\frac{\sqrt{n}}{\eta r_s}\Big) \cdot e^{-\iota}, \tag{180}$$

we have

$$\|\mathbf{q}_{sg}(t+1)\| \le \frac{\eta(\ell+\tilde{\ell})\sqrt{t}}{m} \cdot (\beta_\perp(t)\eta r_s) \cdot \sqrt{\iota} \le \frac{\beta_{\perp'}(t+1)\eta r_s}{20} \cdot \frac{\delta\sqrt{\rho\epsilon}}{8\ell}. \tag{181}$$

Then by union bound, with probability at least

$$1 - 10n(t+1)^2 \log\Big(\frac{\sqrt{n}}{\eta r_s}\Big)e^{-\iota}, \tag{182}$$

we have

$$\|\mathbf{q}_h(t+1) + \mathbf{q}_{sg}(t+1)\| \le \beta_{\perp'}(t+1)\eta r_s \cdot \frac{\delta}{20} \cdot \frac{\sqrt{\rho\epsilon}}{8\ell}, \tag{183}$$

indicating that (173) holds. Then with probability at least

$$1 - 10nt_{\perp'}^2 \log\Big(\frac{\sqrt{n}}{\eta r_s}\Big)e^{-\iota} - \delta/4, \tag{184}$$

we have

$$\|\mathbf{q}_h(t_{\perp'}) + \mathbf{q}_{sg}(t_{\perp'})\| \le \|\mathbf{q}_{p,1}(t_{\perp'})\| \cdot \frac{\sqrt{\rho\epsilon}}{16\ell}. \tag{185}$$

Based on this, we prove that the following inequality holds for any $t_{\perp'} \le t \le \mathscr{T}_s$:

$$\Pr\Big(\|\mathbf{q}_h(\tau) + \mathbf{q}_{sg}(\tau)\| \le \frac{\beta(\tau)\eta r_s}{20\sqrt{n}} \cdot \frac{\delta\sqrt{\rho\epsilon}}{16\ell},\ \forall t_{\perp'} \le \tau \le t\Big) \ge 1 - 10nt^2 \log\Big(\frac{\sqrt{n}}{\eta r_s}\Big)e^{-\iota}. \tag{186}$$

We still use recurrence to prove it. Note that its base case $\tau = t_{\perp'}$ is guaranteed by (173). Suppose it holds for all $\tau \le t$ for some t. Then with probability at least $1 - 2te^{-\iota}$, we have

$$\|\mathbf{z}_t\| \le \eta\|\mathbf{q}_p(t)\| + \eta\|\mathbf{q}_h(t) + \mathbf{q}_{sg}(t)\| \tag{187}$$

$$\le 2\|\mathbf{q}_{p,1}(t)\| + \eta\|\mathbf{q}_h(t) + \mathbf{q}_{sg}(t)\| \tag{188}$$

$$\le \frac{3\beta(\tau)\eta r_s}{\sqrt{n}} \cdot \sqrt{\iota}. \tag{189}$$

Then following a similar procedure as before, we can claim that

$$\|\mathbf{q}_h(t+1)+\mathbf{q}_{sg}(t+1)\| \le \frac{\beta(t+1)\eta r_s}{\sqrt{n}}\cdot\frac{\delta}{20}\cdot\frac{\sqrt{\rho\epsilon}}{8\ell}, \tag{190}$$

holds with probability

$$1-10n(t+1)^2\log\Big(\frac{\sqrt{n}}{\eta r_s}\Big)e^{-\iota}-\frac{\delta}{4}, \tag{191}$$

indicating that (186) holds. Then under our choice of parameters, the desired inequality

$$\|\mathbf{q}_h(t)+\mathbf{q}_{sg}(t)\| \le \frac{\beta(t)\eta r_s\delta}{20\sqrt{n}}\cdot\frac{\sqrt{\rho\epsilon}}{16\ell} \tag{192}$$

holds with probability at least $1-\delta$. □

Equipped with Lemma 25, we are now ready to prove Lemma 23.

Proof. First note that under our choice of $\mathscr{T}_s$, we have

$$\Pr\Big(\frac{\|\mathbf{q}_{p,\perp'}(\mathscr{T}_s)\|}{\|\mathbf{q}_{p,1}(\mathscr{T}_s)\|}\le\frac{\sqrt{\rho\epsilon}}{16\ell}\Big)\ge 1-\delta. \tag{193}$$

Further by Lemma 25 and union bound, with probability at least $1-2\delta$,

$$\frac{\|\mathbf{q}_h(\mathscr{T}_s)+\mathbf{q}_{sg}(\mathscr{T}_s)\|}{\|\mathbf{q}_p(\mathscr{T}_s)\|}\le\|\mathbf{q}_h(\mathscr{T}_s)+\mathbf{q}_{sg}(\mathscr{T}_s)\|\cdot\frac{20\sqrt{n}}{\delta\beta(t)\eta r_s}\le\frac{\sqrt{\rho\epsilon}}{16\ell}. \tag{194}$$

For the output $\hat{\mathbf{e}}$, observe that its component $\hat{\mathbf{e}}_{\perp'}=\hat{\mathbf{e}}-\hat{\mathbf{e}}_1$, since $\mathbf{u}_1$ is the only component in subspace $\mathfrak{S}_{\|'}$. Then with probability at least $1-3\delta$,

$$\|\hat{\mathbf{e}}_{\perp'}\|\le\sqrt{\rho\epsilon}/(8\ell). \tag{195}$$

□

D.1.2 Proof of Proposition 22

Based on Lemma 23, we present the proof of Proposition 22 as follows:

Proof. By Lemma 23, the component $\hat{\mathbf{e}}_{\perp'}$ of output $\mathbf{e}$ satisfies

$$\|\hat{\mathbf{e}}_{\perp'}\|\le\frac{\sqrt{\rho\epsilon}}{8\ell}. \tag{196}$$

Since $\hat{\mathbf{e}}=\hat{\mathbf{e}}_{\|'}+\hat{\mathbf{e}}_{\perp'}$, we can derive that

$$\|\hat{\mathbf{e}}_{\|'}\|\ge\sqrt{1-\frac{\rho\epsilon}{(8\ell)^2}}\ge 1-\frac{\rho\epsilon}{(8\ell)^2}. \tag{197}$$

Note that

$$\hat{\mathbf{e}}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}=(\hat{\mathbf{e}}_{\perp'}+\hat{\mathbf{e}}_{\|'})^T\tilde{\mathcal{H}}(\hat{\mathbf{e}}_{\perp'}+\hat{\mathbf{e}}_{\|'}), \tag{198}$$

which can be further simplified to

$$\hat{\mathbf{e}}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}=\hat{\mathbf{e}}_{\perp'}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}_{\perp'}+\hat{\mathbf{e}}_{\|'}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}_{\|'}. \tag{199}$$

Due to the ℓ-smoothness of the function, all eigenvalue of the Hessian matrix has its absolute value upper bounded by ℓ. Hence,

$$\hat{\mathbf{e}}_{\perp}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}_{\perp}\le\ell\|\hat{\mathbf{e}}_{\perp}^T\|_2^2=\frac{\rho\epsilon}{64\ell^2}, \tag{200}$$

whereas

$$\hat{\mathbf{e}}_{\|'}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}_{\|'}\le-\frac{\sqrt{\rho\epsilon}}{2}\|\hat{\mathbf{e}}_{\|'}\|^2. \tag{201}$$

Combining these two inequalities together, we can obtain

$$\hat{\mathbf{e}}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}=\hat{\mathbf{e}}_{\perp'}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}_{\perp'}+\hat{\mathbf{e}}_{\|'}^T\tilde{\mathcal{H}}\hat{\mathbf{e}}_{\|'}\le-\frac{\sqrt{\rho\epsilon}}{2}\|\hat{\mathbf{e}}_{\|'}\|^2+\frac{\rho\epsilon}{64\ell^2}\le-\frac{\sqrt{\rho\epsilon}}{4}. \tag{202}$$

□

Algorithm 8: Stochastic Gradient Descent with Negative Curvature Finding.

Input: $\mathbf{x}_0 \in \mathbb{R}^n$;
for $t = 0, 1, ..., T$ **do**
  Sample $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}\} \sim \mathcal{D}$;
  $\mathbf{g}(\mathbf{x}_t) = \frac{1}{M}\sum_{j=1}^{M} \mathbf{g}(\mathbf{x}_t; \theta^{(j)})$;
  if $\|\mathbf{g}(\mathbf{x}_t)\| \le 3\epsilon/4$ **then**
    $\hat{\mathbf{e}} \leftarrow$StochasticNegativeCurvatureFinding$(\mathbf{x}_t, r_s, \mathscr{T}_s, m)$;
    $\mathbf{x}_t \leftarrow \mathbf{x}_t - \frac{f'_{\hat{\mathbf{e}}}(\mathbf{x}_0)}{4|f'_{\hat{\mathbf{e}}}(\mathbf{x}_0)|}\sqrt{\frac{\epsilon}{\rho}} \cdot \hat{\mathbf{e}}$;
    Sample $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}\} \sim \mathcal{D}$;
    $\mathbf{g}(\mathbf{x}_t) = \frac{1}{M}\sum_{j=1}^{M} \mathbf{g}(\mathbf{x}_t; \theta^{(j)})$;
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{\ell}\mathbf{g}(\mathbf{x}_t; \theta_t)$;

D.2 Proof details of escaping saddle points using Algorithm 3

In this subsection, we demonstrate that Algorithm 3 can be used to escape from saddle points in the stochastic setting. We first present the explicit Algorithm 8, and then introduce the full version Theorem 9 with proof.

Theorem 26 (Full version of Theorem 9)**.** *Suppose that the function f is ℓ-smooth and ρ-Hessian Lipschitz. For any $\epsilon > 0$ and a constant $0 < \delta_s \le 1$, we choose the parameters appearing in Algorithm 8 as*

$$\delta = \frac{\delta_s}{2304\Delta_f}\sqrt{\frac{\epsilon^3}{\rho}} \qquad \mathscr{T}_s = \frac{8\ell}{\sqrt{\rho\epsilon}} \cdot \log\Big(\frac{\ell n}{\delta\sqrt{\rho\epsilon}}\Big), \qquad \iota = 10\log\Big(\frac{n\mathscr{T}_s^2}{\delta}\log\Big(\frac{\sqrt{n}}{\eta r_s}\Big)\Big), \tag{203}$$

$$r_s = \frac{\delta}{480\rho n\mathscr{T}_s}\sqrt{\frac{\rho\epsilon}{\iota}}, \qquad m = \frac{160(\ell+\tilde{\ell})}{\delta\sqrt{\rho\epsilon}}\sqrt{\mathscr{T}_s\iota}, \qquad M = \frac{16\ell\Delta_f}{\epsilon^2} \tag{204}$$

where $\Delta_f := f(\mathbf{x}_0) - f^*$ and f^* is the global minimum of f. Then, Algorithm 8 satisfies that at least $1/4$ of the iterations $\mathbf{x}_t$ will be ϵ-approximate second-order stationary points, using

$$\tilde{O}\Big(\frac{(f(\mathbf{x}_0) - f^*)}{\epsilon^4} \cdot \log^2 n\Big) \tag{205}$$

iterations, with probability at least $1 - \delta_s$.

Proof. Let the parameters be chosen according to (2), and set the total step number T to be:

$$T = \max\Big\{\frac{8\ell(f(\mathbf{x_0}) - f^*)}{\epsilon^2}, 768(f(\mathbf{x_0}) - f^*) \cdot \sqrt{\frac{\rho}{\epsilon^3}}\Big\}. \tag{206}$$

We will show that the following two claims hold simultaneously with probability $1 - \delta_s$:

1. At most $T/4$ steps have gradients larger than ϵ;
2. Algorithm 3 can be called for at most $384\Delta_f\sqrt{\frac{\rho}{\epsilon^3}}$ times.

Therefore, at least $T/4$ steps are ϵ-approximate secondary stationary points. We prove the two claims separately.

Claim 1. Suppose that within T steps, we have more than $T/4$ steps with gradients larger than ϵ. Then with probability $1 - \delta_s/2$,

$$f(\mathbf{x}_T) - f(\mathbf{x}_0) \le -\frac{\eta}{8}\sum_{i=0}^{T-1}\|\nabla f(\mathbf{x}_i)\|^2 + c \cdot \frac{\sigma^2}{M\ell}(T + \log(1/\delta_s)) \le f^* - f(\mathbf{x}_0), \tag{207}$$

contradiction.

Claim 2. We first assume that for each $\mathbf{x}_t$ we apply negative curvature finding (Algorithm 3), we can successfully obtain a unit vector $\hat{\mathbf{e}}$ with $\hat{\mathbf{e}}^T\mathcal{H}(\mathbf{x}_t)\hat{\mathbf{e}} \le -\sqrt{\rho\epsilon}/4$, as long as $\lambda_{\min}(\mathcal{H}(\mathbf{x}_t)) \le -\sqrt{\rho\epsilon}$. The error probability of this assumption is provided later.

Under this assumption, Algorithm 3 can be called for at most $384(f(\mathbf{x_0})-f^*)\sqrt{\frac{\rho}{\epsilon^3}} \leq \frac{T}{2}$ times, for otherwise the function value decrease will be greater than $f(\mathbf{x_0})-f^*$, which is not possible. Then, the error probability that some calls to Algorithm 3 fails is upper bounded by

$$384(f(\mathbf{x_0})-f^*)\sqrt{\frac{\rho}{\epsilon^3}}\cdot(3\delta)=\delta_s/2. \tag{208}$$

The number of iterations can be viewed as the sum of two parts, the number of iterations needed in large gradient scenario, denoted by T_1, and the number of iterations needed for negative curvature finding, denoted by T_2. With probability at least $1-\delta_s$,

$$T_1 = O(M\cdot T)=\tilde{O}\Big(\frac{(f(\mathbf{x}_0)-f^*)}{\epsilon^4}\Big). \tag{209}$$

As for T_2, with probability at least $1-\delta_s$, Algorithm 3 is called for at most $384(f(\mathbf{x_0})-f^*)\sqrt{\frac{\rho}{\epsilon^3}}$ times, and by Proposition 22 it takes $\tilde{O}\Big(\frac{\log^2 n}{\delta\sqrt{\rho\epsilon}}\Big)$ iterations each time. Hence,

$$T_2 = 384(f(\mathbf{x_0})-f^*)\sqrt{\frac{\rho}{\epsilon^3}}\cdot\tilde{O}\Big(\frac{\log^2 n}{\delta\sqrt{\rho\epsilon}}\Big)=\tilde{O}\Big(\frac{(f(\mathbf{x}_0)-f^*)}{\epsilon^4}\cdot\log^2 n\Big). \tag{210}$$

As a result, the total iteration number T_1+T_2 is

$$\tilde{O}\Big(\frac{(f(\mathbf{x}_0)-f^*)}{\epsilon^4}\cdot\log^2 n\Big). \tag{211}$$

□

E More numerical experiments

In this section, we present more numerical experiment results that support our theoretical claims from a few different perspectives compared to Section 4. Specifically, considering that previous experiments all lies in a two-dimensional space, and theoretically our algorithms have a better dependence on the dimension of the problem n, it is reasonable to check the actual performance of our algorithm on high-dimensional test functions, which is presented in Appendix E.1. Then in Appendix E.2, we introduce experiments on various landscapes that demonstrate the advantage of Algorithm 2 over PAGD [22]. Moreover, we compare the performance of our Algorithm 2 with the NEON$^+$ algorithm [30] on a few test functions in Appendix E.3. To be more precise, we compare the negative curvature extracting part of NEON$^+$ with Algorithm 2 at saddle points in different types of nonconvex landscapes.

E.1 Dimension dependence

Recall that n is the dimension of the problem. We choose a test function $h(x)=\frac{1}{2}x^T\mathcal{H}x+\frac{1}{16}x_1^4$ where $\mathcal{H}$ is an n-by-n diagonal matrix: $\mathcal{H}=\text{diag}(-\epsilon,1,1,...,1)$. The function $h(x)$ has a saddle point at the origin, and only one negative curvature direction. Throughout the experiment, we set $\epsilon=1$. For the sake of comparison, the iteration numbers are chosen in a manner such that the statistics of Algorithm 1 and PGD in each category of the histogram are of similar magnitude.

E.2 Comparison between Algorithm 2 and PAGD on various nonconvex landscapes

Quartic-type test function Consider the test function $f(x_1,x_2)=\frac{1}{16}x_1^4-\frac{1}{2}x_1^2+\frac{9}{8}x_2^2$ with a saddle point at $(0,0)$. The advantage of Algorithm 2 is illustrated in Figure 4.

Triangle-type test function. Consider the test function $f(x_1,x_2)=\frac{1}{2}\cos(\pi x_1)+\frac{1}{2}\Big(x_2+\frac{\cos(2\pi x_1)-1}{2}\Big)^2-\frac{1}{2}$ with a saddle point at $(0,0)$. The advantage of Algorithm 2 is illustrated in Figure 5.

Exponential-type test function. Consider the test function $f(x_1,x_2)=\frac{1}{1+e^{x_1^2}}+\frac{1}{2}\big(x_2-x_1^2e^{-x_1^2}\big)^2-1$ with a saddle point at $(0,0)$. The advantage of Algorithm 2 is illustrated in Figure 6.

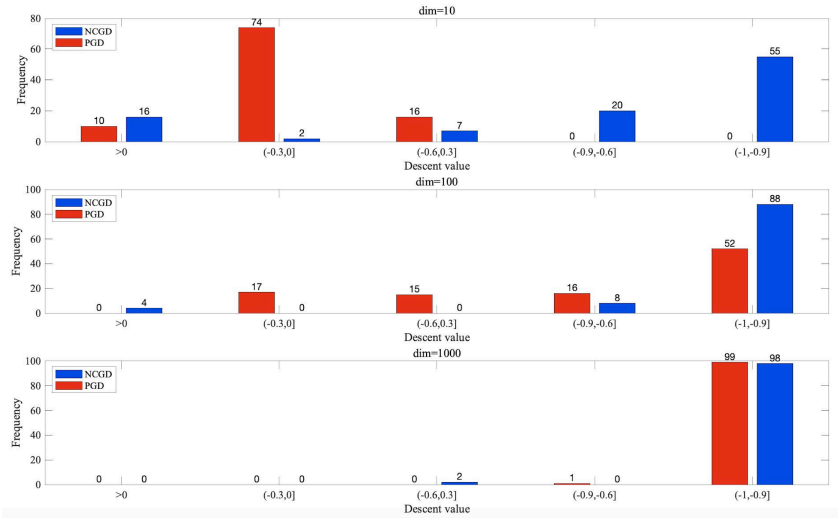

Figure 3: Dimension dependence of Algorithm 1 and PGD. We set $\epsilon = 0.01$, $r = 0.1$, $n = 10^p$ for $p = 1, 2, 3$. The iteration number of Algorithm 1 and PGD are separately set to be $30p$ and $20p^2 + 10$, and the sample size $M = 100$. As we can see, to maintain the same performance, the number of iterations in PGD grows faster than the number of iterations in Algorithm 1.

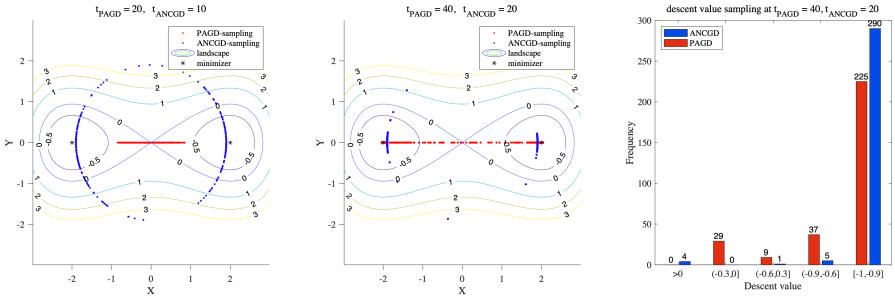

Figure 4: Run Algorithm 2 and PAGD on landscape $f(x_1, x_2) = \frac{1}{16}x_1^4 - \frac{1}{2}x_1^2 + \frac{9}{8}x_2^2$. Parameters: $\eta = 0.05$ (step length), $r = 0.08$ (ball radius in PAGD and parameter r in Algorithm 2), $M = 300$ (number of samplings).
Left: The contour of the landscape is placed on the background with labels being function values. Blue points represent samplings of Algorithm 2 at time step $t_{\text{ANCGD}} = 10$ and $t_{\text{ANCGD}} = 20$, and red points represent samplings of PAGD at time step $t_{\text{PAGD}} = 20$ and $t_{\text{PAGD}} = 40$. Similarly to Algorithm 1, Algorithm 2 transforms an initial uniform-circle distribution into a distribution concentrating on two points indicating negative curvature, and these two figures represent intermediate states of this process. It converges faster than PAGD even when $t_{\text{ANCGD}} \ll t_{\text{PAGD}}$.
Right: A histogram of descent values obtained by Algorithm 2 and PAGD, respectively. Set $t_{\text{ANCGD}} = 20$ and $t_{\text{PAGD}} = 40$. Although we run two times of iterations in PAGD, there are still over 20% of PAGD paths with function value decrease no greater than 0.9, while this ratio for Algorithm 2 is less than 5%.

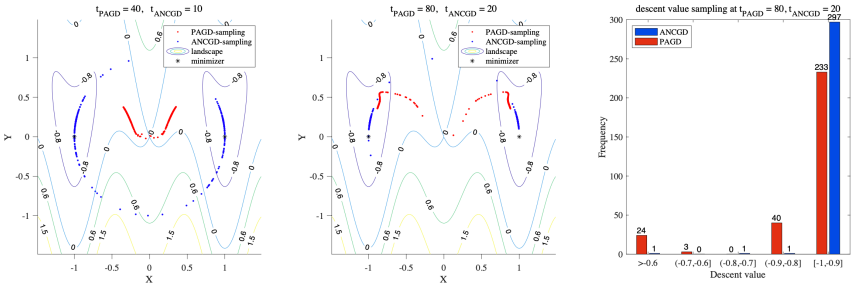

Figure 5: Run Algorithm 2 and PAGD on landscape $f(x_1, x_2) = \frac{1}{2}\cos(\pi x_1) + \frac{1}{2}\left(x_2 + \frac{\cos(2\pi x_1)-1}{2}\right)^2 - \frac{1}{2}$. Parameters: $\eta = 0.01$ (step length), $r = 0.1$ (ball radius in PAGD and parameter r in Algorithm 2), $M = 300$ (number of samplings).
Left: The contour of the landscape is placed on the background with labels being function values. Blue points represent samplings of Algorithm 2 at time step $t_{\text{ANCGD}} = 10$ and $t_{\text{ANCGD}} = 20$, and red points represent samplings of PAGD at time step $t_{\text{PAGD}} = 40$ and $t_{\text{PAGD}} = 80$. Algorithm 2 converges faster than PAGD even when $t_{\text{ANCGD}} \ll t_{\text{PAGD}}$.
Right: A histogram of descent values obtained by Algorithm 2 and PAGD, respectively. Set $t_{\text{ANCGD}} = 20$ and $t_{\text{PAGD}} = 80$. Although we run four times of iterations in PAGD, there are still over 20% of gradient descent paths with function value decrease no greater than 0.9, while this ratio for Algorithm 2 is less than 5%.

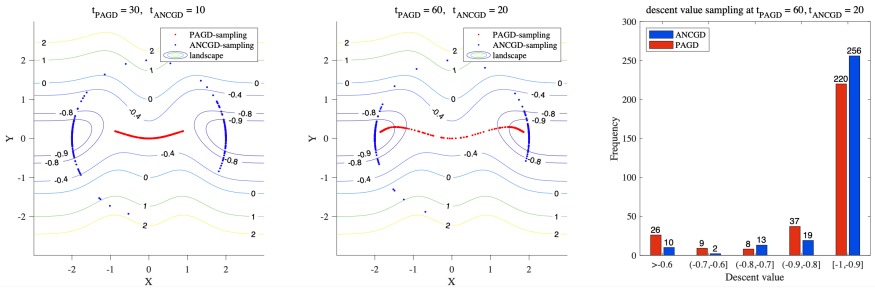

Figure 6: Run Algorithm 2 and PAGD on landscape $f(x_1, x_2) = f(x_1, x_2) = \frac{1}{1+e^{x_1^2}} + \frac{1}{2}\left(x_2 - x_1^2 e^{-x_1^2}\right)^2 - 1$. Parameters: $\eta = 0.03$ (step length), $r = 0.1$ (ball radius in PAGD and parameter r in Algorithm 2), $M = 300$ (number of samplings).
Left: The contour of the landscape is placed on the background with labels being function values. Blue points represent samplings of Algorithm 2 at time step $t_{\text{ANCGD}} = 10$ and $t_{\text{ANCGD}} = 20$, and red points represent samplings of PAGD at time step $t_{\text{PAGD}} = 30$ and $t_{\text{PAGD}} = 60$. Algorithm 2converges faster than PAGD even when $t_{\text{ANCGD}} \ll t_{\text{PAGD}}$.
Right: A histogram of descent values obtained by Algorithm 2 and PAGD, respectively. Set $t_{\text{ANCGD}} = 20$ and $t_{\text{PAGD}} = 60$. Although we run three times of iterations in PAGD, its performance is still dominated by our Algorithm 2.

Compared to the previous experiment on Algorithm 1 and PGD shown as Figure 1 in Section 4, these experiments also demonstrate the faster convergence rates enjoyed by the general family of "momentum methods". Specifically, using fewer iterations, Algorithm 2 and PAGD achieve larger function value decreases separately compared to Algorithm 1 and PGD.

E.3 Comparison between Algorithm 2 and NEON$^+$ on various nonconvex landscapes

Triangle-type test function. Consider the test function $f(x_1, x_2) = \frac{1}{2}\cos(\pi x_1) + \frac{1}{2}\left(x_2 + \frac{\cos(2\pi x_1)-1}{2}\right)^2 - \frac{1}{2}$ with a saddle point at $(0, 0)$. The advantage of Algorithm 2 is illustrated in Figure 7.

Exponential-type test function. Consider the test function $f(x_1, x_2) = \frac{1}{1+e^{x_1^2}} + \frac{1}{2}\left(x_2 - x_1^2 e^{-x_1^2}\right)^2 - 1$ with a saddle point at $(0, 0)$. The advantage of Algorithm 2 is illustrated in Figure 8.

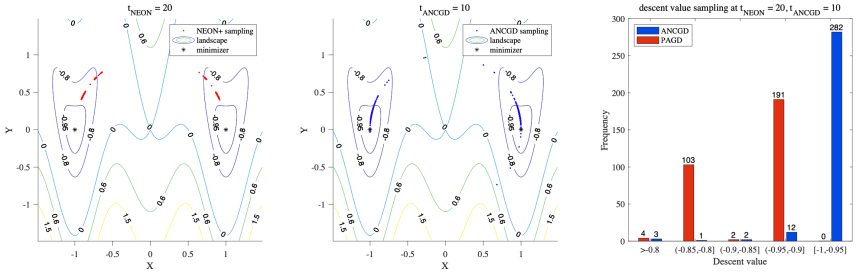

Figure 7: Run Algorithm 2 and NEON$^+$ on landscape $f(x_1, x_2) = \frac{1}{2}\cos(\pi x_1) + \frac{1}{2}\left(x_2 + \frac{\cos(2\pi x_1)-1}{2}\right)^2 - \frac{1}{2}$. Parameters: $\eta = 0.04$ (step length), $r = 0.1$ (ball radius in NEON$^+$ and parameter r in Algorithm 2), $M = 300$ (number of samplings).
Left: The contour of the landscape is placed on the background with labels being function values. Red points represent samplings of NEON$^+$ at time step $t_{\text{NEON}} = 20$, and blue points represent samplings of Algorithm 2 at time step $t_{\text{ANCGD}} = 10$. Algorithm 2 and the negative curvature extracting part of NEON$^+$ both transform an initial uniform-circle distribution into a distribution concentrating on two points indicating negative curvature. Note that Algorithm 2 converges faster than NEON$^+$ even when $t_{\text{ANCGD}} \ll t_{\text{NEON}}$.
Right: A histogram of descent values obtained by Algorithm 2 and NEON$^+$, respectively. Set $t_{\text{ANCGD}} = 10$ and $t_{\text{NEON}} = 20$. Although we run two times of iterations in NEON$^+$, none of NEON$^+$ paths has function value decrease greater than 0.95, while this ratio for Algorithm 2 is larger than 90%.

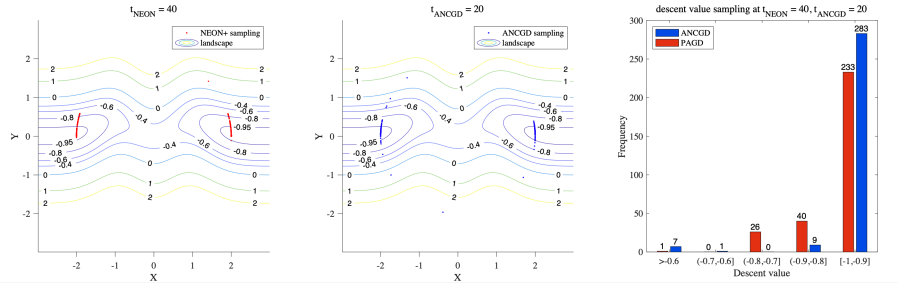

Figure 8: Run Algorithm 2 and NEON$^+$ on landscape $f(x_1, x_2) = \frac{1}{1+e^{x_1^2}} + \frac{1}{2}\left(x_2 - x_1^2 e^{-x_1^2}\right)^2 - 1$. Parameters: $\eta = 0.03$ (step length), $r = 0.1$ (ball radius in NEON$^+$ and parameter r in Algorithm 2), $M = 300$ (number of samplings).
Left: The contour of the landscape is placed on the background with labels being function values. Red points represent samplings of NEON$^+$ at time step $t_{\text{NEON}} = 40$, and blue points represent samplings of Algorithm 2 at time step $t_{\text{ANCGD}} = 20$. Algorithm 2 converges faster than NEON$^+$ even when $t_{\text{ANCGD}} \ll t_{\text{NEON}}$.
Right: A histogram of descent values obtained by Algorithm 2 and NEON$^+$, respectively. Set $t_{\text{ANCGD}} = 20$ and $t_{\text{NEON}} = 40$. Although we run two times of iterations in NEON$^+$, there are still over 20% of NEON$^+$ paths with function value decrease no greater than 0.9, while this ratio for Algorithm 2 is less than 10%.

Compared to the previous experiments on Algorithm 2 and PAGD in Appendix E.2, these two experiments also reveal the faster convergence rate of both NEON$^+$ and Algorithm 2 against PAGD [22] at small gradient regions.