Appendix

A. Preliminaries

Assumptions Recall that we assumed the function f is L-smooth (or L-Lipschitz gradient) and ρ-Lipschitz Hessian. We define these two properties below.

Definition 1 (Smooth function)**.** A differentiable function f is L-smooth (or L-Lipschitz gradient) if

$$\|\nabla f(\mathbf{w}_1) - \nabla f(\mathbf{w}_2)\| \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|, \qquad \forall\, \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d \tag{26}$$

Definition 2 (Hessian Lipschitz)**.** A twice-differentiable function f is ρ-Lipschitz Hessian if

$$\|\nabla^2 f(\mathbf{w}_1) - \nabla^2 f(\mathbf{w}_2)\| \leq \rho\|\mathbf{w}_1 - \mathbf{w}_2\|, \qquad \forall\, \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d \tag{27}$$

Definition 3 (Bounded Gradient)**.** A differentiable function f is ℓ-bounded gradient [8] if

$$\|\nabla f_{\mathbf{z}}(\mathbf{w})\| \leq \ell, \qquad \forall\, \mathbf{w} \in \mathbb{R}^d \tag{28}$$

Lemma 5. *Let $\mathbf{w}_{t+1}$ be obtained from one stochastic gradient step at $\mathbf{w}_t$ on the L-smooth objective f, namely*

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla f_{\mathbf{z}}(\mathbf{w}_t)$$

where $\mathbf{E}_{\mathbf{z}}\left[\nabla f_{\mathbf{z}}(\mathbf{w}_t)\right] = \nabla f(\mathbf{w}_t)$ and $f_{\mathbf{z}}$ is ℓ-bounded gradient. Then the function value decreases in expectation as

$$\mathbf{E}_{\mathbf{z}}\left[f(\mathbf{w}_{t+1})\right] - f(\mathbf{w}_t) \leq -\eta\mathbf{E}\|\nabla f(\mathbf{w}_t)\|^2 + L\eta^2\ell^2/2. \tag{29}$$

Convergence of SGD on a smooth function

Proof. The proof is based on a straightforward application of smoothness:

$$\begin{aligned}
\mathbf{E}_{\mathbf{z}}\left[f(\mathbf{w}_{t+1})\right] - f(\mathbf{w}_t) &\leq -\eta(\nabla f(\mathbf{w}_t))^\top\mathbf{E}\left[\nabla f_{\mathbf{z}}(\mathbf{w}_t)\right] + L/2\eta^2\mathbf{E}\|\nabla f_{\mathbf{z}}(\mathbf{w}_t)\|^2 \\
&\leq -\eta\|\nabla f(\mathbf{w}_t)\|^2 + L\eta^2\|\nabla f_{\mathbf{z}}(\mathbf{w}_t)\|^2/2 \\
&\leq -\eta\|\nabla f(\mathbf{w}_t)\|^2 + L\eta^2\ell^2/2.
\end{aligned}$$

□

Bounded series

Lemma 6. *For all $1 > \beta > 0$, the following series are bounded as*

$$\sum_{i=1}^{t}(1+\beta)^{t-i} \leq 2\beta^{-1}(1+\beta)^t \tag{30}$$

$$\sum_{i=1}^{t}(1+\beta)^{t-i}i \leq 2\beta^{-2}(1+\beta)^t \tag{31}$$

$$\sum_{i=1}^{t}(1+\beta)^{t-i}i^2 \leq 6\beta^{-3}(1+\beta)^t \tag{32}$$

[8] This assumption guarantees ℓ-Lipschitzness of f.

Proof. The proof is based on the following bounds on power series for $|z| < 1$:

$$\sum_{k=1}^{\infty} z^k \leq 1/(1-z)$$

$$\sum_{k=1}^{\infty} z^k k = z/(1-z)^2$$

$$\sum_{k=1}^{\infty} z^k k^2 = z(1+z)/(1-z)^3.$$

Yet, for the sake of brevity, we omit the subsequent (straightforward) derivations needed to prove the statement. □

B. PGD analysis

B.1. Choosing the parameters

Table 4 represents the choice of parameters together with the collection of required constraints on the parameters. This table summarizes our approach for choosing the parameters of CNC-PGD presented in Algorithm 1.

Parameter	Value	Dependency to ϵ	Constraint	Source	constant
η	$1/L$	Independent	$\eta \leq 1/L$	Lemma 1	
r	$c_1(\delta\gamma\epsilon^{4/5})/(\ell^3 L^2)$	$\mathcal{O}(\epsilon^{4/5})$	$\gamma\epsilon^{4/5}/(16L\ell^3)$	Lemma 7 (Eq. (57))	$c_1 = 1/64$
t_{thres}	$c_2 L(\sqrt{\rho}\epsilon^{2/5})^{-1}\log(\ell L/(\gamma\delta\epsilon))$	$\mathcal{O}(\epsilon^{-2/5}\log(1/\epsilon))$	$cL(\sqrt{\rho}\epsilon^{2/5})^{-1}\log(\ell L/(\gamma r)))$	Lemma 7 (Eq. (59))	$c_2 = c$
f_{thres}	$c_3\delta\gamma^2\epsilon^{8/5}/(\ell^2 L)^2$	$\mathcal{O}(\epsilon^{8/5})$	$\leq \gamma\epsilon^{4/5} r/(32\ell)$	Lemma 7 (Eq. (58))	$c_3 = (64)^{-2}$
f_{thres}	”	”	$\geq 2L^2(\ell r)^2/\delta$	Lemma 15 (Eq. (60))	
g_{thres}	$f_{\text{thres}}/t_{\text{thres}}$	$\mathcal{O}(\epsilon^2/\log(1/\epsilon))$			
T	$4(f(\mathbf{w}_0) - f^*)/(\eta\delta g_{\text{thres}})$	$\mathcal{O}(\epsilon^{-2}\log(1/\epsilon))$			

Table 4. Parameters of CNC-PGD.(Restated Table 2)

B.2. Sharp negative curvature regime

Lemma 7 (Restated Lemma 2)**.** *Let Assumption 1 and 2 hold. Consider perturbed gradient steps (Algorithm 1 with parameters as in Table 2) starting from $\tilde{\mathbf{w}}_t$ such that $\|\nabla f(\tilde{\mathbf{w}}_t)\|^2 \leq g_{thres}$. Assume the Hessian matrix $\nabla^2 f(\tilde{\mathbf{w}}_t)$ has a large negative eigenvalue, i.e.*

$$\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{w}}_t)) \leq -\sqrt{\rho}\epsilon^{2/5}. \tag{33}$$

Then, after t_{thres} iterations the function value decreases as

$$\mathbf{E}\left[f(\mathbf{w}_{t+t_{thres}})\right] - f(\tilde{\mathbf{w}}_t) \leq -f_{thres}, \tag{34}$$

where the expectation is over the sequence $\{\mathbf{w}_k\}_{t+1}^{t+t_{thres}}$.

Notation Without loss of generality, we assume that $t = 0$. Let $\mathbf{v}$ be the eigenvector And we use the simplified notation $\xi := \nabla f_{\mathbf{z}}(\tilde{\mathbf{w}}_0)$, $\mathbf{v} := \mathbf{v}_0$. We also use the compact notations:

$$f_t := f(\mathbf{w}_t), \nabla f_t := \nabla f(\mathbf{w}_t), \tilde{f} := f(\tilde{\mathbf{w}}), \nabla \tilde{f} := \nabla f(\tilde{\mathbf{w}}_t), \mathcal{H} := \nabla^2 f(\tilde{\mathbf{w}}), \nabla g_t := g(\mathbf{w}_t),$$

Note that $\tilde{\mathbf{w}}$ denote parameter $\mathbf{w}_0$ before perturbation and $\mathbf{w}_i$ is obtained by i GD steps after perturbation. Recall the compact notation λ as

$$\lambda := |\min\{\lambda_{\min}\left(\nabla^2 f(\tilde{\mathbf{w}}), 0\}\right)|$$

Finally, we set $\kappa := 1 + \eta\lambda$.

Proof sketch The proof presented below proceeds by contradiction and is inspired by the analysis of accelerated gradient descent in non-convex settings as done in (Jin et al., 2017b). We first assume that the sufficient decrease condition is not met and show that this implies an upper bound on the distance moved over a given number of iterations. We then derive a lower bound on the iterate distance and show that - for the specific choice of parameters introduced earlier - this lower bound contradicts the upper bound for a large enough number of steps T. We therefore conclude that we get sufficient decrease for $t > T$.

Proof of Lemma 7:

Part 1: Upper bounding the distance on the iterates in terms of function decrease. We assume that PGD does not obtain the desired function decrease in t_{thres} iterations, i.e.

$$\mathbf{E}\left[f(\mathbf{w}_{t_{\text{thres}}}) - f(\tilde{\mathbf{w}})\right] > -f_{\text{thres}}. \tag{35}$$

The above assumption implies the iterates $\mathbf{w}_t$ stay close to $\tilde{\mathbf{w}}$, for all $t \leq t_{\text{thres}}$. We formalize this result in the following lemma.

Lemma 8 (Distance Bound)**.** *Under the setting of Lemma 7, assume Eq (35) holds. Then the expected distance to the initial parameter can be bounded as*

$$\mathbf{E}\left[\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2\right] \leq 2\left(2\eta f_{thres} + \eta L(\ell r)^2\right)t + 2(\ell r)^2 \qquad \forall t \leq t_{thres}, \tag{36}$$

as long as $\eta \leq 1/L$.

Proof. Here, we use the proposed proof strategy of normalized gradient descent (Levy, 2016). First of all, we bound the effect of the noise in the first step. Recall the first update of Algorithm 1 under the above setting

$$\mathbf{w}_1 = \tilde{\mathbf{w}} - r\xi,\ \xi := \nabla f_{\mathbf{z}}(\tilde{\mathbf{w}}).$$

Then by a straightforward application of lemma 5, we have

$$\mathbf{E}\left[f_1 - \tilde{f}\right] \leq -r\|\nabla \tilde{f}\|^2 + \frac{L}{2}(\ell r)^2. \tag{37}$$

We proceed using the result of Lemma 1 that relates the function decrease to the norm of the visited gradients:

$$\begin{aligned}\mathbf{E}\left[f_{t_{\text{thres}}} - \tilde{f}\right] &= \sum_{t=1}^{t_{\text{thres}}} \mathbf{E}\left[f_t - f_{t-1}\right] \\ &\leq -\frac{\eta}{2}\sum_{t=1}^{t_{\text{thres}}-1} \mathbf{E}\|\nabla f_t\|^2 + \mathbf{E}\left[f_1 - \tilde{f}\right] \\ &\leq -\frac{\eta}{2}\sum_{t=1}^{t_{\text{thres}}-1} \mathbf{E}\|\nabla f_t\|^2 + \frac{L}{2}(\ell r)^2. \quad \text{[Eq. 37]}\end{aligned} \tag{38}$$

According to Eq. (35), the function value does not decrease too much. Plugging this bound into the above inequality yields an upper bound on the sum of the squared norm of the visited gradients, i.e.

$$\sum_{t=1}^{t_{\text{thres}}-1} \mathbf{E}\|\nabla f_t\|^2 \leq (2f_{\text{thres}} + L(\ell r)^2)/\eta. \tag{39}$$

Using the above result allows us to bound the expected distance in the parameter space as:

$$
\begin{aligned}
\mathbf{E}\left[\|\mathbf{w}_t - \mathbf{w}_1\|^2\right] &= \mathbf{E}\left[\|\sum_{i=2}^{t} \mathbf{w}_i - \mathbf{w}_{i-1}\|^2\right] \\
&\leq \mathbf{E}\left[\left(\sum_{i=2}^{t} \|\mathbf{w}_i - \mathbf{w}_{i-1}\|\right)^2\right] && \text{[Triangle inequality]} \\
&\leq \mathbf{E}\left[t\sum_{i=2}^{t} \|\mathbf{w}_i - \mathbf{w}_{i-1}\|^2\right] && \text{[Cauchy-Schwarz inequality]} \\
&\leq t\left(\mathbf{E}\left[\eta^2 \sum_{i=1}^{t-1} \|\nabla f_i\|^2\right]\right) \\
&\leq \left(2\eta f_{\text{thres}} + \eta L(\ell r)^2\right) t, \qquad \forall t \leq t_{\text{thres}}. && \text{[Eq. (39)]}
\end{aligned} \tag{40}
$$

Replacing the above inequality into the following bound completes the proof:

$$
\begin{aligned}
\mathbf{E}\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2 &\leq 2\mathbf{E}\|\mathbf{w}_t - \mathbf{w}_1\|^2 + 2\mathbf{E}\|\mathbf{w}_1 - \tilde{\mathbf{w}}\|^2 \\
&\leq 2\left(2\eta f_{\text{thres}} + \eta L(\ell r)^2\right) t + 2(\ell r)^2
\end{aligned}
$$

□

Part 2: Quadratic approximation Since the parameter vector stays close to $\tilde{\mathbf{w}}$ under the condition in Eq. (35), we can use a "stale" Taylor expansion approximation of the function f at $\tilde{\mathbf{w}}$:

$$g(\mathbf{w}) = \tilde{f} + (\mathbf{w} - \tilde{\mathbf{w}})^\top \nabla f(\tilde{\mathbf{w}}) + \frac{1}{2}(\mathbf{w} - \tilde{\mathbf{w}})^\top \mathcal{H}(\mathbf{w} - \tilde{\mathbf{w}}).$$

Using a stale Taylor approximation over all iterations is the essential part of the proof that is firstly proposed by (Ge et al., 2015) for analysis of PSGD method. The next lemma proves that the gradient of f can be approximated by the gradient of g as long as $\mathbf{w}$ is close enough to $\tilde{\mathbf{w}}$.

Lemma 9 (Taylor expansion bound for the gradient (Nesterov, 2013))**.** *For every twice differentiable, function $f : \mathbb{R}^d \to \mathbb{R}$ with ρ-Lipschitz Hessians the following bound holds true.*

$$\|\nabla f(\mathbf{w}) - \nabla g(\mathbf{w})\| \leq \frac{\rho}{2}\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 \tag{41}$$

Furthermore, the guaranteed closeness to the initial parameter allows us to use the gradient of the quadratic objective g in the GD steps as follows,

$$
\begin{aligned}
\mathbf{w}_{t+1} - \tilde{\mathbf{w}} &= \mathbf{w}_t - \eta\nabla f_t - \tilde{\mathbf{w}} \\
&= \mathbf{w}_t - \tilde{\mathbf{w}} - \eta\nabla g_t + \eta\left(\nabla g_t - \nabla f_t\right) \\
&= (\mathbf{I} - \eta\mathcal{H})(\mathbf{w}_t - \tilde{\mathbf{w}}) + \eta(\nabla g_t - \nabla f_t - \nabla f(\tilde{\mathbf{w}})) \\
&= \mathbf{u}_t + \eta(\boldsymbol{\delta}_t + \mathbf{d}_t),
\end{aligned} \tag{42}
$$

where the vectors $\mathbf{u}_t$, $\boldsymbol{\delta}_t$ and $\mathbf{d}_t$ are defined in Table 5.

As long as $\mathbf{w}_1 - \tilde{\mathbf{w}}$ is correlated with the negative curvature, the norm of $\mathbf{u}_t$ grows exponentially. In this case, the upper bound of Lemma 8 doesn't hold anymore after a certain number of iterations, as we formally prove in part 3. Indeed, the term $\mathbf{u}_t$ constitutes power iterations on the hessian matrix $\mathcal{H}$. The term $\boldsymbol{\delta}_t$ arises from the stale Taylor approximation errors through all iterations. Assuming that $\mathbf{w}_t$ stays close to $\tilde{\mathbf{w}}$, we will bound this term. Finally, the $\mathbf{d}_t$ terms depend on the initial gradient. We will show that the distance $\mathbf{E}\|\mathbf{w}_1 - \tilde{\mathbf{w}}\|^2$ is eventually dominated by the power iterates $\mathbf{u}_t$.

Vector	Formula	Indication
$\mathbf{u}_t$	$(\mathbf{I} - \eta\mathcal{H})^t (\mathbf{w}_1 - \tilde{\mathbf{w}})$	Power Iteration
$\boldsymbol{\delta}_t$	$\sum_{i=1}^t (\mathbf{I} - \eta\mathcal{H})^{t-i} (\nabla f_t - \nabla g_t)$	Stale Taylor Approximation Error
$\mathbf{d}_t$	$-\sum_{i=1}^t (\mathbf{I} - \eta\mathcal{H})^{t-i} \nabla f(\tilde{\mathbf{w}})$	Initial Gradient Dependency

Table 5. Components of CNC-PGD expanded steps.

Part 3: Lower bounding the iterate distance.

A lower-bound on the distance Our goal is to provide a lower-bound on $\mathbf{E}\|\mathbf{w}_{t_{\text{thres}}} - \mathbf{w}_0\|^2$ that contradicts the result of Lemma 8. To obtain a lower bound on the distance, we use the classical result $\|a+b\|^2 \geq \|a\|^2 + 2a^\top b$. Setting $a = \mathbf{u}_t$ and $b = \eta(\boldsymbol{\delta}_t + \mathbf{d}_t)$ yields

$$\begin{aligned} \mathbf{E}\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}\| &\geq \mathbf{E}\|\mathbf{u}_t\|^2 + 2\eta\mathbf{E}\left[\mathbf{u}_t^\top \boldsymbol{\delta}_t\right] + 2\eta\mathbf{E}\left[\mathbf{u}_t^\top\right]\mathbf{d}_t \\ &\geq \mathbf{E}\|\mathbf{u}_t\|^2 - 2\eta\mathbf{E}\left[\|\mathbf{u}_t\|\|\boldsymbol{\delta}_t\|\right] + 2\eta\mathbf{E}\left[\mathbf{u}_t^\top\right]\mathbf{d}_t \end{aligned} \tag{43}$$

Removing the initial gradient dependency The established lower-bound in Eq. (43) has a dependency to the gradient $\nabla f(\tilde{\mathbf{w}})$ through the term $\mathbf{E}\left[\mathbf{u}_t^\top\right]\mathbf{d}_t$. Intuitively, the initial gradient should not cause a problem for negative curvature exploration phase. More precisely, the third term of the lower bound of Eq. (43) should be positive. This result is proven in the next lemma.

Lemma 10 (Removing initial gradient dependency)**.** *Under the setting of Lemma 7,*

$$\mathbf{E}\left[\mathbf{u}_t^\top\right]\mathbf{d}_t \geq 0. \tag{44}$$

Proof. Assumption 1 (CNC) implies that $\mathbf{E}\left[\mathbf{w}_1 - \tilde{\mathbf{w}}\right] = -r\nabla f(\tilde{\mathbf{w}})$, hence the expectation of the power iteration term is

$$\mathbf{E}\left[\mathbf{u}_t\right] = (\mathbf{I} - \eta\mathcal{H})^t\,\mathbf{E}\left[\mathbf{w}_1 - \tilde{\mathbf{w}}\right] = -r\,(\mathbf{I} - \eta\mathcal{H})^t\,\nabla f(\tilde{\mathbf{w}}).$$

Using this result, as well as the fact that $(\mathbf{I} - \eta\mathcal{H}) \succeq 0$ for $\eta \leq 1/L$ we have

$$\begin{aligned} \mathbf{E}\left[\mathbf{u}_t^\top\right]\mathbf{d}_t &= r\left((\mathbf{I} - \eta\mathcal{H})^t \nabla f(\tilde{\mathbf{w}})\right)^\top \sum_{i=1}^t (\mathbf{I} - \eta\mathcal{H})^{t-i}\,\nabla f(\tilde{\mathbf{w}}) \\ &= r\sum_{i=1}^t \nabla f(\tilde{\mathbf{w}})^\top (\mathbf{I} - \eta\mathcal{H})^{2t-i} \nabla f(\tilde{\mathbf{w}}) \geq 0, \end{aligned}$$

which proves the assertion. □

Plugging the result of the last lemma into the lower-bound established in Eq. (43) yields

$$\mathbf{E}\left[\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2\right] \geq \mathbf{E}\|\mathbf{u}_t\|^2 - 2\eta\mathbf{E}\left[\|\mathbf{u}_t\|\|\boldsymbol{\delta}_t\|\right]. \tag{45}$$

To complete our lower bound, we need : (I) a lower bound on $\mathbf{E}\|\mathbf{u}_t\|^2$, (II) an upper bound on $\|\mathbf{u}_t\|$ and (III) an upper bound on $\mathbf{E}\|\boldsymbol{\delta}_t\|$.

Lemma 11 (Exponential Growing Power Iteration)**.** *Under the setting of Lemma 7, t steps of PGD yield an exponentially growing lower bound on the expected squared norm of $\mathbf{u}_t$, i.e.*

$$\mathbf{E}\left[\|\mathbf{u}_t\|^2\right] \geq \gamma r^2 \kappa^{2t}. \tag{46}$$

(I) Lower-bound on $\mathbf{E}\|\mathbf{u}_t\|^2$

Proof. We first use Cauchy-Schwarz inequality to derive the following lower bound:

$$\begin{aligned}\mathbf{E}\left[\|\mathbf{u}_t\|^2\right] &= \mathbf{E}\left[\|\mathbf{v}\|^2\|\mathbf{u}_t\|^2\right] \\ &\geq \mathbf{E}\left[(\mathbf{v}^\top \mathbf{u}_t)^2\right].\end{aligned} \tag{47}$$

Now, suppose $\mathcal{H} = U^\top \Sigma U$. Since any non-zero vector $\mathbf{u} \in \mathbb{R}^d$ is an eigenvector of the identity matrix we can decompose $I = U^\top I U$ and thus $(I - \eta\mathcal{H}) = U^\top(I - \eta\Sigma)U$. As a result, we have

$$\mathbf{v}^\top(I - \eta\mathcal{H}) = \mathbf{v}^\top(1 - \eta\lambda_{\min}(\mathcal{H})) = \mathbf{v}^\top(1 + \eta\lambda). \tag{48}$$

for the leftmost eigenvector $\mathbf{v}$ of the Hessian $\mathcal{H}$. Plugging Equation (48) into (47) and recalling $\kappa := 1 + \eta\lambda$ as well as the definition of $\mathbf{u}_t$ as given in Table 5 yields

$$\begin{aligned}\mathbf{E}\left[\|\mathbf{u}_t\|^2\right] &\geq (1 + \eta\lambda)^{2t}\mathbf{E}\left[(\mathbf{v}^\top(\mathbf{w}_1 - \tilde{\mathbf{w}}))^2\right] \\ &= r^2\kappa^{2t}\mathbf{E}\left[(\mathbf{v}^\top \xi)^2\right] \\ &= \gamma r^2 \kappa^{2t},\end{aligned}$$

where the last inequality follows from Assumption 1. □

(II) Upper-bound on $\|\mathbf{u}_t\|$ Using the definition of the scaling factor r and the fact that the noise lies inside the unit sphere the next lemma proves a deterministic bound on this term.

Lemma 12. *Under the setting of Lemma 7 the norm of $\mathbf{u}_t$ is **deterministically** bounded as*

$$\|\mathbf{u}_t\| \leq \kappa^t \ell r. \tag{49}$$

Proof. Starting from the definition of $\mathbf{u}_t$ and recalling that ξ has at most unit norm by Assumption 1, we have

$$\begin{aligned}\|\mathbf{u}_t\| &\leq \|\left(\mathbf{I} - \eta\mathcal{H}\right)^t(\mathbf{w}_1 - \tilde{\mathbf{w}})\| \\ &\leq \|\mathbf{I} - \eta\mathcal{H}\|^t\|\mathbf{w}_1 - \tilde{\mathbf{w}}\| \\ &\leq (1 + \eta\lambda)^t r\|\xi\| \\ &\leq \kappa^t r\ell\end{aligned}$$

□

(III) Upper bound on $\mathbf{E}\|\boldsymbol{\delta}_t\|$ To bound this term, we use the fact proved in Lemma 8 that $\mathbf{w}_t$ stays close to $\mathbf{w}_0$ for all $t \leq t_{\text{thres}}$.

Lemma 13. *Under the setting of Lemma 7, if*

$$\mathbf{E}\left[f_{t+1} - \tilde{f}\right] \geq -f_{thres},$$

then the norm of $\boldsymbol{\delta}_t$ is bounded in expectation:

$$\mathbf{E}\|\boldsymbol{\delta}_t\| \leq \left(\frac{4\eta f_{thres} + 2\eta L(\ell r)^2}{(\eta\lambda)^2} + \frac{2(\ell r)^2}{\eta\lambda}\right)\rho\kappa^t, \qquad \forall t \leq t_{thres}. \tag{50}$$

Proof. Using the result of Lemma 9 as well as the distance bound established in Lemma 8, we have

$$\begin{aligned}
\mathbf{E}\left[\|\boldsymbol{\delta}_t\|\right] &= \mathbf{E}\left[\|\sum_{i=1}^{t}\left(\mathbf{I}-\mathcal{H}\right)^{t-i}\left(\nabla g_i - \nabla f_i\right)\|\right] \\
&\leq \sum_{i=1}^{t} \|\mathbf{I}-\eta\mathcal{H}\|^{t-i}\mathbf{E}\|\nabla g_i - \nabla f_i\| \\
&\leq \frac{\rho}{2}\sum_{i=1}^{t}\kappa^{t-i}\mathbf{E}\|\mathbf{w}_i - \tilde{\mathbf{w}}\|^2 \qquad \text{[]Lemma 9]} \\
&\leq \rho\sum_{i=1}^{t}\kappa^{t-i}\left(\left(2\eta f_{\text{thres}} + \eta L(\ell r)^2\right)i + (\ell r)^2\right). \qquad \text{[Lemma 8]}
\end{aligned} \tag{51}$$

Now, the results on convergence of power series derived in Lemma 6 and the definition $\kappa := 1 + \eta\lambda$ give

$$\sum_{i=1}^{t}\kappa^{t-i} \leq \frac{2\kappa^t}{\eta\lambda} \quad \text{and} \quad \sum_{i=1}^{t}\kappa^{t-i}i \leq \frac{2\kappa^t}{(\eta\lambda)^2}. \tag{52}$$

By combining Equation (51) and (52) we can establish the desired bound on $\boldsymbol{\delta}_t$:

$$\mathbf{E}\left[\|\boldsymbol{\delta}_t\|\right] \leq \left(\frac{4\eta f_{\text{thres}} + 2\eta L(\ell r)^2}{(\eta\lambda)^2} + \frac{2(\ell r)^2}{\eta\lambda}\right)\rho\kappa^t. \tag{53}$$

□

We are now ready to combine the results of Lemma 11, 12, and 13, into Eq. (45) in order to obtain the desired lower bound on the distance travelled by the iterates of PGD.

Lemma 14 (Distance lower bound)**.** *Under the setting of Lemma 7 and for each $t \leq t_{\mathit{thres}}$ and for the choice of parameters as in Table 4 we have*

$$\mathbf{E}\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2 \geq \frac{\gamma r^2\kappa^{2t}}{4}, \tag{54}$$

where $\kappa := 1 + \eta|\lambda_{\min}\left(\nabla^2 f(\tilde{\mathbf{w}})\right)|$.

Proof. To prove this statement we introduce each bound in Eq. (45) step by step:

$$\begin{aligned}
\mathbf{E}\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2 &\geq \mathbf{E}\|\mathbf{u}_t\|^2 - 2\eta\mathbf{E}\left[\|\mathbf{u}_t\|\|\boldsymbol{\delta}_t\|\right] \\
&\geq \gamma r^2\kappa^{2t} - 2\eta\mathbf{E}\left[\|\mathbf{u}_t\|\|\boldsymbol{\delta}_t\|\right] \qquad \text{[Lemma 11]} \\
&\geq \gamma r^2\kappa^{2t} - 2\eta\ell r\kappa^t\mathbf{E}\|\boldsymbol{\delta}_t\| \qquad \text{[Lemma 12]} \\
&\geq \left(\gamma r - \frac{8\rho\ell f_{\text{thres}} + 4\rho L\ell^3 r^2}{\lambda^2} - \frac{4\rho\ell^3 r^2}{\lambda}\right) r\kappa^{2t} \qquad \text{[Lemma 13]}
\end{aligned} \tag{55}$$

We need the lower bound to be positive to complete the proof. In this regard, we require the following condition to hold,

$$\gamma r - \underbrace{\frac{8\rho\ell f_{\text{thres}}}{\lambda^2}}_{\leq\gamma r/4} - \underbrace{\frac{4L\rho\ell^3 r^2}{\lambda^2}}_{\leq\gamma r/4} - \underbrace{\frac{4\rho\ell^3 r^2}{\lambda}}_{\leq\gamma r/4} \overset{!}{\geq} \frac{\gamma r}{4}. \tag{56}$$

Using the lower bound the absolute value of the minimum eigenvalue as $\lambda \geq \sqrt{\rho}\epsilon^{2/5}$ (in Eq. (33)), we choose parameters r, f_{thres}, and g_{thres} such that the above constraints are satisfied,[9] i.e.

$$r \leq \gamma\epsilon^{4/5}/(16L\ell^3) \leq \gamma\lambda^2/(16\rho L\ell^3) \overset{[\lambda<L]}{\leq} (\gamma\lambda)/(16\rho\ell^3) \tag{57}$$

$$f_{\text{thres}} \leq \gamma\epsilon^{4/5} r/(32\ell) \leq \gamma\lambda^2 r/(32\rho\ell) \tag{58}$$

These choices of parameters establish an exponential lower bound on the distance as

$$\mathbf{E}\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2 \geq \frac{\gamma r^2 \kappa^{2t}}{4}.$$

□

According to the result of Lemma 8, if the expected distance from the initial parameter is sufficiently large, then the assumption $\mathbf{E}\left[f_t - \tilde{f}\right] > -f_{\text{thres}}$ cannot be valid. To derive the contradiction, we have to choose the number of step such that the established lower-bound exceeds the upper-bound in Lemma 8, namely

$$\frac{1}{4}\gamma r^2\kappa^{2t} \overset{?}{\geq} 2\left(2\eta f_{\text{thres}} + \eta L(\ell r)^2\right)t + 2(\ell r)^2.$$

Since the left hand side is exponentially growing, we can derive the contradiction by choosing a large enough number of steps as:

$$t_{\text{thres}} \geq c(\eta\lambda)^{-1}\log\left(\ell L/(\gamma r)\right) \geq cL(\sqrt{\rho}\epsilon^{2/5})^{-1}\log(\ell L/(\gamma r))), \tag{59}$$

which completes the proof of Lemma 7.

□

B.3. Moderate negative curvature regime

Lemma 15 (Restate of Lemma 3)**.** *Let Assumption 1 and 2 hold. Consider perturbed gradient steps (Algorithm 1 with parameters as in Table 4) starting from* $\tilde{\mathbf{w}}_t$ *such that* $\|\nabla f(\tilde{\mathbf{w}}_t)\|^2 \leq g_{thres}$. *Then after* t_{thres} *iterations, the function value cannot increase by more than*

$$\mathbf{E}\left[f(\mathbf{w}_{t+t_{thres}})\right] - f(\tilde{\mathbf{w}}_t) \leq \frac{\eta\delta f_{thres}}{4},$$

where the expectation is over the sequence $\{\mathbf{w}_k\}_{t+1}^{t+t_{thres}}$.

Proof. Using the resulf of lemma 5, we bound the decrease in the function value as

$$\mathbf{E}\left[f(\mathbf{w}_1)\right] - f(\tilde{\mathbf{w}}_t) \leq L(\ell r)^2/2 \leq \delta\eta f_{\text{thres}}/4 \tag{60}$$

Since there is no perturbation in following t_{thres} steps, GD doesn't increase the function value in following t_{thres}-steps (according to the result of lemma 1). □

C. SGD analysis

C.1. Parameters and Constraints

Table 6 lists the parameters of CNC-SGD presented in Algorithm 2 together with the constraints that determines our choice of parameters. These constraints are driven by the theoretical analysis.

[9] Note that the second requirement in (56) is always more restrictive than the last since $\lambda < L$.

Parameter	Value	Dependency to ϵ	Constraint	Constraint Origin	Constant
r	$c_1\delta\gamma\epsilon^{4/5}/(\ell^3 L)$	$\mathcal{O}(\epsilon^{4/5})$	$\leq \gamma\epsilon^{4/5}/(24\ell^3 L)$	Lemma 16 (Eq. (77))	$c_1 = 1/34$
f_{thres}	$c_2\delta\gamma^2\epsilon^{8/5}/(\ell^4 L)$	$\mathcal{O}(\epsilon^{8/5})$	$\leq \gamma\epsilon^{4/5} r/(48\ell)$	Lemma 16(Eq. (79))	$c_2 = c_1/48$
f_{thres}	"	"	$\geq 2L(\ell r)^2/\delta$	Eq. (63)	
ω	$c_3 \log(\ell L/(\eta\epsilon r))$	$\mathcal{O}(\log(1/\epsilon))$			$c_3 = c$ (Eq.(81))
η	$c_4\delta^2\gamma^2\epsilon^2/(\ell^6 L^2\zeta)$	$\mathcal{O}(\epsilon^2/\log(1/\epsilon))$	$\leq \epsilon^2/(2L)$	Eq. (65)	
η	"	"	$\leq \gamma\sqrt{\rho}\epsilon^{6/5} r/(72\ell^3 L)$	Lemma 16(Eq.(78))	$c_4 = c_1/72$
t_{thres}	$(\eta\epsilon^{2/5})^{-1}\omega$	$\mathcal{O}(\epsilon^{-12/5}\log^2(1/\epsilon))$	$\geq c(\eta\lambda)^{-1}\log(\ell L/(\gamma r\eta\lambda))$	Lemma 16(Eq.(81))	
g_{thres}	$f_{\text{thres}}/t_{\text{thres}}$	$\mathcal{O}(\epsilon^4/\log^2(1/\epsilon))$	$\geq 2L(\ell\eta)^2/\delta$	Eq. (69)	
g_{thres}	"	"	$\leq \eta\epsilon^2/2$	Eq. (66)	
T	$2(f(\mathbf{w}_0) - f^*)/(\delta g_{\text{thres}})$	$\mathcal{O}(\epsilon^{-4}\log^2(1/\epsilon))$			

Table 6. Parameters of CNC-SGD (Restated Table 3)

C.2. Proof of the Main Theorem

Theorem 3 (Restated Theorem 2)**.** *Let the stochastic gradients $\nabla f_{\mathbf{z}}(\mathbf{w}_t)$ in CNC-SGD satisfy Assumption 1 and let f and $f_{\mathbf{z}}$ satisfy Assumption 2. Then algorithm 2 returns an $(\epsilon, \sqrt{\rho}\epsilon^{2/5})$-second order stationary point with probability at least $(1-\delta)$ after*

$$\mathcal{O}\left(\left(\frac{\delta\gamma\epsilon}{L\ell^{5/2}}\right)^{-4}\log^2\left(\frac{\ell L}{\epsilon\delta\gamma}\right)\right)$$

steps, where $\delta < 1$.

Proof. We decompose the SGD step as

$$\tilde{\mathbf{w}}_t = \mathbf{w}_{t-1} - r\xi_t \quad \text{[Large Step-Size]} \tag{61}$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla f(\mathbf{w}_t) + \eta\underbrace{(\nabla f(\mathbf{w}_t) - \nabla f_{\mathbf{z}}(\mathbf{w}_t))}_{\zeta_t}, \quad \text{[Small Step-Size]} \tag{62}$$

where the noise term ζ_ts are i.i.d and zero-mean and the noise term ξ_t satisfies CNC assumption 1. Our analysis relay on the CNC assumption only at steps with a larger step-size r. Indeed, we only exploit the negative curvature in the steps with a large step size r. In this regard, we need to use the larger step size $r > \eta$ in these steps. This is different from Perturbed SGD – with isotropic noise– (Ge et al., 2015) where the variance of perturbations in all steps is exploited in the analysis.

Amortized increase due enlarging step size Recall Algorithm 2 increases the step size every t_{thres} step. The increase in the function value in this step is bounded as

$$\mathbf{E}\left[f(\mathbf{w}_{t+1})\right] - f(\mathbf{w}_t) \overset{\text{lemma 5}}{\leq} (L/2)(\ell r)^2 \leq \delta f_{\text{thres}}/4 \tag{63}$$

which leads to a per step increase of

$$\frac{\mathbf{E}\left[f(\mathbf{w}_{t+1})\right] - f(\mathbf{w}_t)}{t_{\text{thres}}} \leq (\delta/4) f_{\text{thres}}/t_{\text{thres}} = \delta g_{\text{thres}}/4 \tag{64}$$

Large gradient regime: If the norm of the gradient is large, i.e.

$$\|\nabla f(\mathbf{w}_t)\|^2 \geq \epsilon^2 \geq 2L\eta, \tag{65}$$

then the result on convergence of one step of SGD in Lemma 5 guarantees the desired decrease

$$\begin{aligned}\mathbf{E}_{\mathbf{z}}\left[f(\mathbf{w}_{t+1})\right] - f(\mathbf{w}_t) &\leq -\eta\|\nabla f(\mathbf{w}_t)\|^2 + L\eta^2 \\ &\leq -\eta\epsilon^2/2 \\ &\leq -g_{\text{thres}}.\end{aligned} \tag{66}$$

Sharp curvature regime: When the minimum eigenvalue is significantly less than zero, SGD steps with a large step-size r provides enough variance for following SGD steps –with a smaller step size η– to exploit the negative curvature direction. This estatement is formally proved in the next lemma.

Lemma 16 (Negative curvature exploration by CNC-SGD)**.** *Suppose Assumptions 1 and 2 hold. If the Hessian matrix at $\tilde{\mathbf{w}}_t$ has a small negative eigenvalue, i.e.*

$$\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{w}}_t)) \leq -\sqrt{\rho}\epsilon^{2/5}. \tag{67}$$

Then there exists a $k < t_{thres}$ such that the expectation of the function value decreases as

$$\mathbf{E}\left[f(\mathbf{w}_{t+k})\right] - f(\tilde{\mathbf{w}}_t) \leq -f_{thres}, \tag{68}$$

where the expectation is taken over the sequence $\{\mathbf{w}_k\}_{t+1}^{t+k}$.

Moderate curvature and gradient regime: Suppose that the minimum eigenvalue of the Hessian is quite small and visited gradients has also a small norm. In this case, we need to bound increase in the function caused by the variance of SGD. A straight-forward application of lemma 5 obtains the desired bound on the increase of function value caused by the variance of SGD:

$$\mathbf{E}\left[f(\mathbf{w}_{t+1})\right] - f(\mathbf{w}_t) \leq L(\ell\eta)^2/2 \leq \delta g_{\text{thres}}/4 \tag{69}$$

Probabilistic bound The probabilistic lower bound on returning the desired second stationary point can be derived from Eq.s (66) and (69) as well as Lemma 16 using exactly the same argument as the probabilistic argument on perturbed gradient descent. We define the event $\mathcal{A}_t$ as

$$\mathcal{A}_t := \{\|\nabla f(\mathbf{w}_t)\| \geq \epsilon \text{ or } \lambda_{\min}(\nabla^2 f(\mathbf{w}_t)) \leq -\sqrt{\rho}\epsilon^{2/5}\}.$$

According to the result fpr the large gradient regime (in Eq. (66)) and the large curvature result (in Lemma 16), SGD obtains the guaranteed decrease in function value –amortized per step– conditional on $\mathcal{A}_t$:

$$\mathbf{E}\left[f(\mathbf{w}_{t+1}) - f(\mathbf{w}_t)|\mathcal{A}_t\right] \leq -g_{\text{thres}}.$$

Furthermore, the increase of the function value due to the stochastic gradient steps is controlled by using our choice of steps sizes, according to the result of Eq. (64) and (69):

$$\mathbf{E}\left[f(\mathbf{w}_{t+1}) - f(\mathbf{w}_t)|\mathcal{A}_t^c\right] \leq \delta g_{\text{thres}}/2.$$

Let $\mathcal{P}_t$ is the probability associated with $\mathcal{A}_t$, hence $1 - \mathcal{P}_t$ is the probability associated with its complement event $\mathcal{A}_t^c$. Note that computing the probabilities $\mathcal{P}_t$ is very hard due to the dependency of $\mathbf{w}_t$ to all stochastic gradient steps before iteration t. Plugging these probabilities into the above conditional expectation results yields

$$\mathbf{E}\left[f(\mathbf{w}_{t+1}) - f(\mathbf{w}_t)\right] \leq (1 - \mathcal{P}_t)\delta g_{\text{thres}}/2 - \mathcal{P}_t g_{\text{thres}}.$$

Summing the above inequalities over the T steps obtains the following upper-bound on the average of $\mathcal{P}_t$s

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{P}_t \leq \frac{f(\mathbf{w}_0) - f^*}{T g_{\text{thres}}} + \frac{\delta}{2}.$$

The above bound allows us to lower-bound the probability of retrieving an $(\epsilon, \sqrt{\rho}\epsilon^{2/5})$-second order stationary point (which is equivalent to the occurrence of the complement event $\mathcal{A}_t^c$) uniformly over T steps:

$$\sum_{t=1}^{T}(1 - \mathcal{P}_t)/T \geq 1 - \delta.$$

This concludes the proof of the convergence guarantee of CNC-SGD under Assumption 1. □

C.3. Proof of the main Lemma 16

Lemma 17 (Restated Lemma 16)**.** *Suppose Assumptions 1 and 2 hold. If the Hessian matrix at $\tilde{\mathbf{w}}_t$ has a small negative eigenvalue, i.e.*

$$\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{w}}_t)) \leq -\sqrt{\rho}\epsilon^{2/5}. \tag{70}$$

Then there exists a $k < t_{thres}$ such that the expectation of the function value decreases as

$$\mathbf{E}\left[f(\mathbf{w}_{t+k})\right] - f(\tilde{\mathbf{w}}_t) \leq -f_{thres} \tag{71}$$

where the expectation is taken over the sequence $\{\mathbf{w}_k\}_{t+1}^{t+t_{thres}}$.

Proof. Our analysis for the large curvature case in CNC-PGD (lemma 7) can be extended to SGD. Here, we borrow the compact notations from Lemma 7. Similar to the proof scheme of lemma 7, our proof is based on contradiction. We assume that for all $t < t_{\text{thres}}$ the desired decrease in the function value is not obtained, namely

$$\mathbf{E}\left[f(\mathbf{w}_t)\right] - \tilde{f} \geq -f_{\text{thres}}, \quad \forall t \leq t_{\text{thres}}. \tag{72}$$

A direct result of the above assumption is that we can establish a bound on the expectation of the distance to $\tilde{\mathbf{w}}$ for all iterates $\mathbf{w}_t$ such that $t < t_{\text{thres}}$.

Lemma 18 (Distance Bound for SGD)**.** *Suppose that expectation of the decrease in function value is lower-bounded as stated in (72). Then, the expectation of the distance from the current iterate to $\tilde{\mathbf{w}}$ is bounded as*

$$\mathbf{E}\left[\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2\right] \leq \left(4f_{thres}\eta + 2L\eta(\ell r)^2 + 4(\ell\eta)^2\right)t + 2L\eta^3\ell^2t^2 + 2(\ell r)^2, \quad \forall t \leq t_{thres},$$

as long as Assumption 2 holds.

We postpone the proof of the last lemma to section C.4. The proposed bound in the last lemma is larger than the established distance bound for PGD steps , in lemma 8. This is due to the variance of stochastic gradients. In the rest of the proof, we will construct a lower-bound that contradicts to the above upper-bound using the CNC assumption 1. To this end, we expansion SGD steps (in Eq. (61)) using the gradient of the stale Taylor expansion $g(\tilde{\mathbf{w}})$:

$$\begin{aligned}
\mathbf{w}_{t+1} - \tilde{\mathbf{w}} &= \mathbf{w}_t - \tilde{\mathbf{w}} - \eta\nabla f_t + \eta\zeta_t \\
&= \mathbf{w}_t - \tilde{\mathbf{w}} - \eta\nabla g_t + \eta\left(\nabla f_t - \nabla g_t - \nabla f(\tilde{\mathbf{w}}) + \zeta_t\right) \\
&= \left(\mathbf{I} - \eta\mathcal{H}\right)\left(\mathbf{w}_t - \tilde{\mathbf{w}}\right) + \eta\left(\nabla f_t - \nabla g_t - \nabla f(\tilde{\mathbf{w}}) + \zeta_t\right) \\
&= \mathbf{u}_t + \eta\left(\boldsymbol{\delta}_t + \mathbf{d}_t + \boldsymbol{\zeta}_t\right)
\end{aligned}$$

where the vectors $\mathbf{u}_t$, $\boldsymbol{\delta}_t$, $\mathbf{d}_t$, and $\boldsymbol{\zeta}_t$ are defined in Table 7. The only new term in the expansion is the noise of the stochastic gradient steps $\boldsymbol{\zeta}_t$s. Similarly to PGD, the power iterations $\mathbf{u}_t$ plays an essential rule in the negative curvature exploration.

Vector	Formula	Indication	Included in PGD analysis?
$\mathbf{u}_t$	$\left(\mathbf{I} - \eta\mathcal{H}\right)^t\left(\mathbf{w}_1 - \tilde{\mathbf{w}}\right)$	Power Iteration	Yes
$\boldsymbol{\delta}_t$	$\sum_{i=1}^t \left(\mathbf{I} - \eta\mathcal{H}\right)^{t-i}\left(\nabla f_t - \nabla g_t\right)$	Stale Taylor Approximation Error	Yes
$\mathbf{d}_t$	$-\sum_{i=1}^t \left(\mathbf{I} - \eta\mathcal{H}\right)^{t-i}\nabla f(\tilde{\mathbf{w}})$	Initial Gradient Dependency	Yes
$\boldsymbol{\zeta}_t$	$\sum_{i=1}^t \left(\mathbf{I} - \eta\mathcal{H}\right)^{t-i}\zeta_i$	Noise of Stochastic Gradients	No

Table 7. Components of CNC-SGD expanded steps.

For this term, we can reuse our analysis in lemmas 12, and 11. The term $\boldsymbol{\delta}_t$ is caused by using a stale Taylor approximation in all iterates $t \leq t_{\text{thres}}$. We need to bound the perturbation effect of this term to guarantee that power iterates $\mathbf{u}_t$ exploit the negative curvature. To this end, we required a bound on $\mathbf{E}\|\boldsymbol{\delta}_t\|$. This bound is established in the next lemma using the distance bound of Lemma 18.

Lemma 19. *Under the condition of Lemma 18, the bound*

$$\mathbf{E}\|\boldsymbol{\delta}_t\| \leq \rho\left(\frac{2(\ell r)^2}{\eta\lambda} + \frac{4\eta f_{thres} + 2L\eta(\ell r)^2 + 4(\ell\eta)^2}{(\lambda\eta)^2} + \frac{6L\eta^3\ell^2}{(\lambda\eta)^3}\right) \tag{73}$$

holds true.

Proof.

$$\begin{aligned}
\mathbf{E}\|\boldsymbol{\delta}_t\| &= \mathbf{E}\|\sum_{k=1}^{t}(\mathbf{I}-\eta\mathcal{H})^{t-k}\left(\nabla f_k - \nabla g_k\right)\| \\
&\leq \sum_{k=1}^{t}(1+\eta\lambda)^{t-k}\mathbf{E}\|\nabla f_k - \nabla g_k\| \\
&\leq (\rho/2)\sum_{k=1}^{t}\kappa^{t-k}\mathbf{E}\|\mathbf{w}_k - \tilde{\mathbf{w}}\|^2 \\
&\leq (\rho/2)\sum_{k=1}^{t}\kappa^{t-k}\left(\left(4f_{\text{thres}}\eta + 2L\eta(\ell r)^2 + 4(\ell\eta)^2\right)k + 2L\eta^3\ell^2k^2 + 2(\ell r)^2\right) \quad \text{[Lemma 18]} \\
&\leq \rho\left(\frac{2(\ell r)^2}{\eta\lambda} + \frac{4\eta f_{\text{thres}} + 2L\eta(\ell r)^2 + 4(\ell\eta)^2}{(\lambda\eta)^2} + \frac{6L\eta^3\ell^2}{(\lambda\eta)^3}\right)\kappa^t \quad \text{[Lemma 6]}
\end{aligned} \tag{74}$$

□

Lower-bound on the distance Using the step expansion, we lower-bound the distance from the pivot $\tilde{\mathbf{w}}$ as

$$\begin{aligned}
\mathbf{E}\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}\|^2 &\geq \mathbf{E}\|\mathbf{u}_t\|^2 - 2\eta\|\mathbf{u}_t\|\mathbf{E}\|\boldsymbol{\delta}_t\| + 2\eta\mathbf{E}\left[\mathbf{u}_t^\top\mathbf{d}_t\right] + 2\eta\mathbf{E}\left[\mathbf{u}_t\right]\underbrace{\mathbf{E}\left[\boldsymbol{\zeta}_t\right]}_{=0} \\
&\geq \mathbf{E}\|\mathbf{u}_t\|^2 - 2\eta\|\mathbf{u}_t\|\mathbf{E}\|\boldsymbol{\delta}_t\| + 2\eta\mathbf{E}\left[\mathbf{u}_t^\top\mathbf{d}_t\right] \\
&\geq \mathbf{E}\|\mathbf{u}_t\|^2 - 2\eta\|\mathbf{u}_t\|\mathbf{E}\|\boldsymbol{\delta}_t\| \quad \text{[Lemma 10]} \\
&\geq \gamma r^2\kappa^{2t} - 2\eta\ell r\kappa^t\mathbf{E}\left[\|\boldsymbol{\delta}_t\|\right] \quad \text{[Lemma 11 \& 12]} \\
&\geq \left(\gamma r - \frac{4\rho\ell^3r^2}{\lambda} - \frac{8\rho\ell f_{\text{thres}}}{\lambda^2} - \frac{4L\rho\ell^3r^2}{\lambda^2} - \frac{8\rho\eta\ell^3}{\lambda^2} - \frac{12L\rho\eta\ell^3}{\lambda^3}\right)r\kappa^{2t} \quad \text{[Lemma 19]}
\end{aligned} \tag{75}$$

Constraints on the parameter To derived the desired contradiction, i.e. $\mathbf{E}\left[f_t\right] - \tilde{f} \leq -f_{\text{thres}}$, we need to prove that the distance $\mathbf{E}\left[\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2\right]$ is larger than the upper-bound established in lemma 18. To this end, we have to choose parameters such that the established lower-bound on the distance in Eq. (75) be positive, i.e.

$$\gamma r - \underbrace{\frac{4\rho\ell^3r^2}{\lambda}}_{\leq\gamma r/6} - \underbrace{\frac{8\rho\ell f_{\text{thres}}}{\lambda^2}}_{\leq\gamma r/6} - \underbrace{\frac{4\rho L\ell^3r^2}{\lambda^2}}_{\leq\gamma r/6} - \underbrace{\frac{8\rho\eta\ell^3}{\lambda^2}}_{\leq\gamma r/6} - \underbrace{\frac{12\rho L\eta\ell^3}{\lambda^3}}_{\leq\gamma r/6} \overset{!}{\geq} \gamma r/6 \tag{76}$$

Using the lower-bound on the absolute value of minimum eigenvalue, i.e. $\lambda \geq \sqrt{\rho}\epsilon^{2/5}$, we choose parameters such that the above constraints are satisfied:

$$r \leq \gamma \epsilon^{4/5}/(24\ell^3 L) \leq (\gamma\lambda^2)/(24L\rho\ell^3) \leq \lambda\gamma/(24\rho\ell^3) \tag{77}$$

$$\eta \leq \gamma\sqrt{\rho}\epsilon^{6/5} r/(72\ell^3 L) \leq \gamma\lambda^3 r/(72L\rho\ell^3) \tag{78}$$

$$f_{\text{thres}} \leq \gamma\epsilon^{4/5} r/(48\ell) \leq \gamma\lambda^2 r/(48\rho\ell). \tag{79}$$

Our choice of parameters fulfills the above constraints. Plugging the above result into Eq. (75) obtains the exponential growing lower-bound on the distance

$$\mathbf{E}\left[\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}\right] \geq \gamma r^2 \kappa^{2t}/6 \tag{80}$$

Contradiction by choosing the number of iterations t_{thres} Using the lower bound of Eq. (80), we can establish a contradictory result with the upperbound on the distance proposed in lemma 18

$$\gamma r^2 \kappa^{2t_{\text{thres}}}/6 \overset{!}{\geq} \left(2f_{\text{thres}}\eta + L\eta(\ell r)^2 + 2(\ell\eta)^2\right) t + L\eta^3\ell^2 t^2 + 2(\ell r)^2.$$

Since the left-side of the above inequality is exponentially growing, one can choose the number iterations t_{thres} large enough to derive the contradiction:

$$t_{\text{thres}} \geq c(\eta\lambda)^{-1}\log(L\ell/(\gamma r\eta\lambda)) \tag{81}$$

where c is a constant independent of parameters λ,γ,L and ρ. □

C.4. Bound on the expectation of distance

Here, we complete the proof of lemma 7 by proving the following lemma, which is used in lemma 15.

Lemma 20 (Restated Lemma 18). *Suppose that expectation of the decrease in function value is lower-bounded as*

$$\mathbf{E}\left[f(\mathbf{w}_t)\right] - \tilde{f} \geq -f_{\textit{thres}}, \quad \forall t \leq t_{\textit{thres}}. \tag{82}$$

Then, the expectation of the distance from the current iterate to $\tilde{\mathbf{w}}$ is bounded as

$$\mathbf{E}\left[\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2\right] \leq \left(4f_{\textit{thres}}\eta + 2L\eta(\ell r)^2 + 4(\ell\eta)^2\right)t + 2L\eta^3\ell^2 t^2 + 2(\ell r)^2, \quad \forall t \leq t_{\textit{thres}}, \tag{83}$$

as long as Assumption 2 holds.

Proof. We use the result of lemma 5

$$-f_{\text{thres}} \leq \mathbf{E}\left[f_{t+1} - \tilde{f}\right] = \mathbf{E}\left[\sum_{i=1}^{t} f_{i+1} - f_i\right]$$
$$\leq -\eta\sum_{i=1}^{t}\mathbf{E}\|\nabla f_i\|^2 + L(\ell\eta)^2 t/2 + L(\ell r)^2/2 \quad \text{[Lemma 5]}.$$

Rearranging terms obtains a bound on the sum of the squared norm of visited gradients:

$$\sum_{i=1}^{t}\mathbf{E}\|\nabla f_i\|^2 \leq f_{\text{thres}}/\eta + L\ell^2\eta t/2 + L(\ell r)^2/(2\eta) \tag{84}$$

Using the Telescopic expansion of the difference $\mathbf{w}_{t+1} - \mathbf{w}_1$, we relate the distance to the visited stochastic gradients:

$$\begin{aligned}\mathbf{E}\left[\|\mathbf{w}_{t+1} - \mathbf{w}_1\|^2\right] &= \mathbf{E}\left[\|\sum_{i=1}^{t}\mathbf{w}_{i+1} - \mathbf{w}_i\|^2\right] \\ &\leq \eta^2\mathbf{E}\|\sum_{i=1}^{t}\left(\zeta_i - \nabla f_i\right)\|^2. \quad \text{[SGD-step decomposition, Eq. (61)]}\end{aligned} \tag{85}$$

To upper bound the right-side of the above inequality, we rely on i.i.d and zero-mean assumption of ζ_ts:

$$\begin{aligned}
\mathbf{E}\|\sum_{i=1}^{t}(\zeta_i - \nabla f_i)\|^2 &\leq 2\mathbf{E}\|\sum_{i=1}^{t}\nabla f_i\|^2 + 2\mathbf{E}\|\sum_{i=1}^{t}\zeta_i\|^2 \quad \text{[Parallelogram law]} \\
&= 2\mathbf{E}\|\sum_{i=1}^{t}\nabla f_i\|^2 + 2\sum_{i\neq j}\mathbf{E}\underbrace{\left[\zeta_i^\top\zeta_j\right]}_{\text{Independent}} + 2\sum_{i=1}^{t}\mathbf{E}\left[\zeta_i^\top\zeta_i\right] \\
&= 2\mathbf{E}\|\sum_{i=1}^{t}\nabla f_i\|^2 + 2\sum_{i=1}^{t}\mathbf{E}\|\zeta_i\|^2 \\
&\leq 2\mathbf{E}\|\sum_{i=1}^{t}\nabla f_i\|^2 + 2t\ell^2 \\
&\leq 2\mathbf{E}\left(\sum_{i=1}^{t}\|\nabla f_i\|\right)^2 + 2t\ell^2 \quad \text{[Triangle inequality]} \\
&\leq 2t\sum_{i=1}^{t}\mathbf{E}\|\nabla f_i\|^2 + 2t\ell^2 \quad \text{[CauchySchwarz inequality]} \\
&\overset{(84)}{\leq} 2t\left(f_{\text{thres}}/\eta + L\eta\ell^2 t/2 + L(\ell r)^2/(2\eta) + \ell^2\right).
\end{aligned} \tag{86}$$

Replacing the above bound into Eq. (85) yields:

$$\mathbf{E}\|\mathbf{w}_{t+1} - \mathbf{w}_1\|^2 \leq t\left(2f_{\text{thres}}\eta + L\eta(\ell r)^2 + 2(\ell\eta)^2\right) + L\eta^3\ell^2 t^2.$$

Using the above result, we bound the distance as:

$$\begin{aligned}
\mathbf{E}\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}\|^2 &\leq 2\mathbf{E}\|\mathbf{w}_{t+1} - \mathbf{w}_1\|^2 + 2\mathbf{E}\left[\|\mathbf{w}_1 - \tilde{\mathbf{w}}\|^2\right] \quad \text{[Parallelogram law]} \\
&\leq 2\mathbf{E}\|\mathbf{w}_{t+1} - \mathbf{w}_1\|^2 + 2(\ell r)^2 \\
&\leq \left(4f_{\text{thres}}\eta + 2L\eta(\ell r)^2 + 4(\ell\eta)^2\right)t + 2L\eta^3\ell^2 t^2 + 2(\ell r)^2
\end{aligned} \tag{87}$$

Finally, replacing $t+1$ by t concludes the proof. □

D. Analysis of Learning Half-spaces

Lemma 21 (Restated 4)**.** *Consider the problem of learning half-spaces as stated in Eq. (21), where φ satisfies Assumption 3. Furthermore, assume that the support of $\mathcal{P}$ is a subset of the unit sphere. Let $\mathbf{v}$ be a unit length eigenvector of $\nabla^2 f(\mathbf{w})$ with corresponding eigenvalue $\lambda < 0$. Then*

$$\mathbf{E}_{\mathbf{z}}\left[(\nabla f_{\mathbf{z}}(\mathbf{w})^\top\mathbf{v})^2\right] \geq (\lambda/c)^2. \tag{88}$$

Proof. Using the definition of an eigenvector, $\nabla^2 f(\mathbf{w})\mathbf{v} = \lambda\mathbf{v}$ and since $\nabla^2 f(\mathbf{w}) = \varphi''(\mathbf{w}^\top\mathbf{z})\mathbf{z}\mathbf{z}^\top$ we have:

$$\begin{aligned}
\lambda &= \mathbf{v}^\top\nabla^2 f(\mathbf{w})\mathbf{v} \\
&= \mathbf{E}\left[\varphi''(\mathbf{w}^\top\mathbf{z})(\mathbf{z}^\top\mathbf{v})^2\right] \\
&\geq -\mathbf{E}\left[|\varphi''(\mathbf{w}^\top\mathbf{z})|(\mathbf{z}^\top\mathbf{v})^2\right] \\
&\geq -c\mathbf{E}\left[|\varphi'(\mathbf{w}^\top\mathbf{z})|(\mathbf{z}^\top\mathbf{v})^2\right] \quad \text{[Eq. (23)]} \\
&\geq -c\mathbf{E}\left[|\varphi'(\mathbf{w}^\top\mathbf{z})||\mathbf{z}^\top\mathbf{v}|\right] \quad [\|\mathbf{z}\| \leq 1] \\
&\geq -c\mathbf{E}\left[|\varphi'(\mathbf{w}^\top\mathbf{z})\mathbf{z}^\top\mathbf{v}|\right].
\end{aligned} \tag{89}$$

Using the above result and as well as Jensen's inequality, we derive the desired result:

$$\begin{aligned}\mathbf{E}\left[(\nabla f_{\mathbf{z}}(\mathbf{w})^\top \mathbf{v})^2\right] &= \mathbf{E}\left[(\varphi'(\mathbf{w}^\top \mathbf{z})\mathbf{z}^\top \mathbf{v})^2\right] \\ &= \left(\sqrt{\mathbf{E}\left[(\varphi'(\mathbf{w}^\top \mathbf{z})\mathbf{z}^\top \mathbf{v})^2\right]}\right)^2 \\ &\geq \left(\mathbf{E}\left[\sqrt{(\varphi'(\mathbf{w}^\top \mathbf{z})\mathbf{z}^\top \mathbf{v})^2}\right]\right)^2 \\ &\geq \left(\mathbf{E}|\varphi'(\mathbf{w}^\top \mathbf{z})\mathbf{z}^\top \mathbf{v})|\right)^2 \\ &\geq (\lambda/c)^2,\end{aligned} \tag{90}$$

where the last inequality follows from Eq. (89) and the fact that $\lambda < 0$. □

E. Additional experimental results

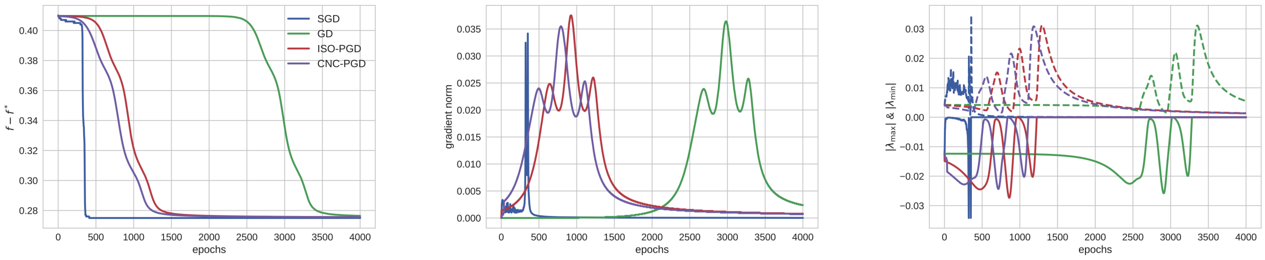
Learning halfspaces From each of two multivariat gaussian distributions we draw $n/2 = 20$ samples $x_i \in \mathbb{R}^4$ and assign them the labels $y_i \in \{0, 1\}$ respectively. We then optimize the loss function

$$f(\mathbf{w}) = \text{sigmoid}\left(-y_i \mathbf{x}_i^\top \mathbf{w}\right) + \frac{1}{2}\|\mathbf{w}\|^2$$

with the following methods and hyperparameters:

Gradient Descent, Stochastic Gradient Descent, PGD as in (Jin et al., 2017a) with perturbation radius $r = 0.1$ and PGD-CNC with a stochastic gradient step as perturbation. All methods use the step size $\alpha = 1/4$, the stochastic gradient steps are performed with batch size 1 and the perturbed gradient descent methods perturb as soon as $\nabla f(\mathbf{w}) < g_{\text{thres}} := 0.01$.

To complete the picture of Figure 2 we here also present the gradient norms and minimum/maximum eigenvalues along the trajectories of the different methods. It becomes apparent that all of them indeed started at a saddle and eventually move towards (and along) the flat end of the sigmoid. However, Gradient Descent is much slower in finding regions of significant negative curvature than the stochastic methods.



1. Suboptimality

2. Gradient norm

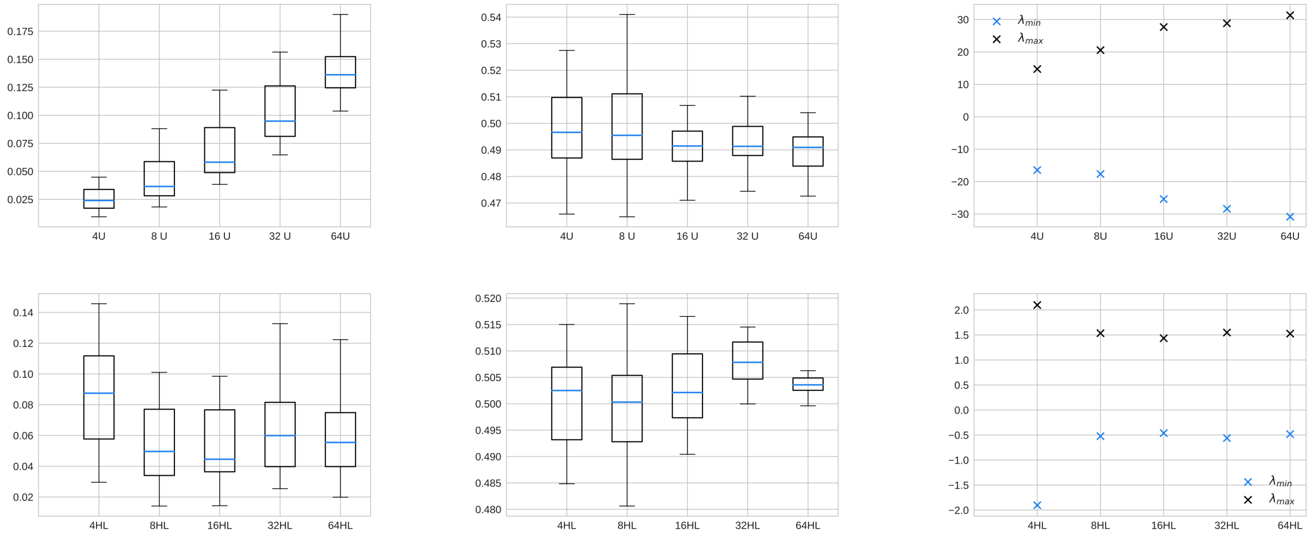
3. min and max eigenvalues

Figure 4. Learning halfspaces: more details.

Neural Networks The neural network experiments were implemented using the Pytorch library and conducted on a GPU server. Note that we downsized the mnist dataset to an image size of 10×10 and applied sigmoid acivations in the hidden layers as well as a cross-entropy loss over the 10 classes.

While we present covariances between the stochastic gradients/isotropic noise vectors with the *leftmost* Eigenvectors in the main paper, Figure 5 plots the covariances with the entire negative eigenspectrum.

In Figure 3 we show that the correlation of eigenvectors and stochastic gradients increases with the magnitude of the associated eigenvalues. As expected, this is not the case for noise vectors that are drawn randomly from the unit sphere. Furthermore, these correlations show a decrease with an increasing dimension as can be seen in Figure 6.



1. SGD covariance w/ left eigenvectors
2. isotropic noise covariance w/ left eigenvectors
3. min and max eigenvalues

Figure 5. Average covariances and eigenvalues of 30 random parameters in Neural Networks with increasing width (top) and depth (bottom).

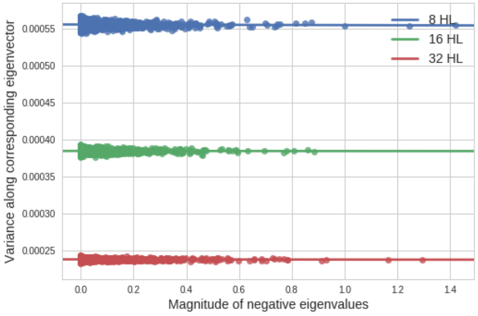


Figure 6. Correlation of stochastic gradients with eigenvectors corresponding to eigenvalues of different magnitudes on Neural Nets with 8, 16 and 32 hidden layers. Scatterplot and fitted linear model with 95% confidence interval.