Supplementary Material

This document contains the proofs of the results presented in the paper: Robustness Analysis of Non-Convex Stochastic Gradient Descent using Biased Expectations.

Proof of Proposition 2. If $s = 0$, the result is trivial. Otherwise, we have $\mu_s(X + Y) = \frac{1}{2} \ln \mathbb{E} [e^{s(X + Y)}] = \frac{1}{2} \ln \left( \mathbb{E} [e^{sX}] \mathbb{E} [e^{sY}] \right) = \mu_s(X) + \mu_s(Y)$, where the second equality follows from the independence of $e^{sX}$ and $e^{sY}$.

Proof of Proposition 3. First, note that $\mu_s(X) = -\mu_{-s}(-X)$, and thus 2 implies 1 (as $-\text{ess sup}(-X) = \text{ess inf } X$). Moreover, $L_p$-norms of probability spaces are both non-decreasing and tend to the essential supremum (i.e., $p \mapsto \|Y\|_p$ is non-decreasing and $\lim_{p \to +\infty} \|Y\|_p = \text{ess sup } Y$). Hence, using the alternative formulation $\mu_s(X) = \ln \|e^X\|_s$, we get that $s \mapsto \mu_s(X)$ is non-decreasing, and $\lim_{s \to +\infty} \mu_s(X) = \ln(\text{ess sup}(e^X)) = \text{ess sup } X$. Finally, note that the function $(s, x) \mapsto \phi_s(x) = e^{sx} - 1$ is continuous. Let $s_0, s_1 \in I_X$, and $s_0 < s < s_1$. By definition of $I_X$, $\phi_{s_0}(X)$ and $\phi_{s_1}(X)$ are integrable. Moreover, $|\phi_s(X)| = \max\{\phi_{s_0}(X), \phi_{s_1}(X)\} \leq \max\{\phi_{s_0}(X), \phi_{s_1}(X)\} \leq \phi_{s_0}(X) + \phi_{s_1}(X) \leq |\phi_{s_0}(X)| + |\phi_{s_1}(X)|$ by monotonicity of $s \mapto \phi_s(x)$. As $|\phi_{s_0}(X)| + |\phi_{s_1}(X)|$ is integrable and independent of $s$, dominated convergence implies continuity of $\mathbb{E}[\phi_s(X)]$, and thus of $\mu_s(X)$, in $(s_1, s_2)$.

Proof of Proposition 4. A simple rewriting of $\mu_s(\mu_s(X)|F)$ leads to the desired result: $\mu_s(\mu_s(X)|F) = \phi_s^{-1}(\mathbb{E} [\phi_s \circ \phi_s^{-1}(\mathbb{E} [\phi_s(X)|F])] = \phi_s^{-1}(\mathbb{E} [\mathbb{E} [\phi_s(X)|F]]) = \phi_s^{-1}(\mathbb{E} [\phi_s(X)]) = \mu_s(X)$.

Proof of Proposition 5. Eq. (3) follows from the Chernoff bound $\mathbb{P}(X \geq a) \leq \mathbb{E} [e^{sX}] e^{-sa}$ for $a = \mu_s(X) + x$. Moreover, if $X \geq 0$ a.s., using Markov’s inequality on $\phi_s(X)$, we obtain

$$
\mathbb{P}(X \geq x) \leq \frac{\phi_s(\mu_s(X))}{\phi_s(x)}.
$$

When $s < 0$, we can further simplify Eq. (15) by using $\phi_s(\mu_s(X)) \leq \mu_s(X)$ (as $\phi_s$ is concave), and $\phi_s(x) \geq \frac{x}{1 - sx}$, which concludes the proof.

Proof of Theorem 6. The result follows from standard analysis of non-convex gradient descent. More specifically, using the $\beta$-smoothness of $f$, we have

$$
f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\beta}{2} \|x_{t+1} - x_t\|^2
\leq f(x_t) - \eta \|\nabla f(x_t), G_t\| + \frac{\beta \eta^2}{2} \|G_t\|^2
\leq f(x_t) - \eta \|\nabla f(x_t), X_t\| - \eta (1 - \beta \eta) \langle \nabla f(x_t), X_t \rangle + \frac{\beta \eta^2}{2} \|X_t\|^2
\leq f(x_t) - \eta \left(1 - \frac{\beta \eta}{2}\right) \|\nabla f(x_t)\|^2 - \eta (1 - \beta \eta) \langle \nabla f(x_t), X_t \rangle + \frac{\beta \eta^2}{2} \|X_t\|^2.
$$

Rearranging Eq. (16) and summing over all times $t \in \{0, T - 1\}$ leads to

$$
\eta \left(1 - \frac{\beta \eta}{2}\right) \sum_{t < T} \|\nabla f(x_t)\|^2 \leq \Delta - \eta (1 - \beta \eta) \sum_{t < T} \langle X_t, \nabla f(x_t) \rangle + \frac{\beta \eta^2}{2} \sum_{t < T} \|X_t\|^2,
$$

where $\Delta = f(x_0) - \min_{x \in \mathbb{R}^d} f(x)$. Finally, using the assumption $\eta \in (0, 1/\beta]$ we obtain $\frac{\eta}{2} \leq \eta \left(1 - \frac{\beta \eta}{2}\right)$, and thus dividing by $\eta T/2$ gives that

$$
\frac{1}{T} \sum_{t = 1, \ldots, T} \|\nabla f(x_t)\|^2 \leq \frac{2 \Delta}{\eta T} - \frac{2 (1 - \beta \eta)}{T} \sum_{t < T} \langle X_t, \nabla f(x_t) \rangle + \frac{\beta \eta}{T} \sum_{t < T} \|X_t\|^2.
$$

To conclude, we apply biased expectation to both sides of Eq. (18). As $\langle X_t, \nabla f(x_t) \rangle$ and $\|X_t\|^2$ are not independent, Proposition 2 does not apply. We thus use the following Lemma to decompose the error.
Lemma 18. Let $X, Y$ be two (possibly dependent) random variables and $s \in \mathbb{R}$. If $s \geq 0$, then $\mu_s(X + Y) \leq \mu_{2s}(X) + \mu_{2s}(Y)$. Otherwise, $\mu_s(X + Y) \leq \mu_s(X) + \frac{\mathbb{E}[Y e^{sX}]}{e^{s \mathbb{E}[X]}}$, whenever the right-hand sides are well-defined.

Proof. For $s > 0$, applying the Cauchy-Schwartz inequality to $e^{sX}$ and $e^{sY}$ gives $\mu_s(X + Y) = \frac{1}{s} \ln \left( \mathbb{E} \left[ e^{sX}e^{sY} \right] \right) \leq \frac{1}{s} \ln \left( \mathbb{E} \left[ e^{2sX} \right] \mathbb{E} \left[ e^{2sY} \right] \right) = \mu_{2s}(X) + \mu_{2s}(Y)$. For $s < 0$, we obtain that $\mu_s(X + Y) = \mu_s(X) + \frac{1}{s} \ln \mathbb{E} \left[ e^{sY} e^{sX} \mathbb{E} \left[ e^{sX} \right] \right]$ by a direct rewriting. Now, introducing the random variable $Y'$ with density $\frac{e^{sX}}{e^{s \mathbb{E}[X]}}$ w.r.t. the probability measure of $(X, Y)$ and using Jensen's inequality on the function $x \mapsto \frac{1}{s} \ln(x)$, we obtain that $\frac{1}{s} \ln \mathbb{E} \left[ e^{sY} e^{sX} \mathbb{E} \left[ e^{sX} \right] \right] = \frac{1}{s} \ln \mathbb{E} \left[ e^{sY} \right] \leq \mathbb{E} \left[ \frac{1}{s} \ln e^{sY} \right] = \mathbb{E} \left[ Y' \right] = \mathbb{E} \left[ Y \mathbb{E} \left[ e^{sX} \right] \right]$ which proves the result. 

Moreover, note that, for any $a \in \mathbb{R}$, $\mu_s(aX) = a \mu_{as}(X)$. Then, using Proposition 2 to remove the deterministic error, we have

$$\mu_s \left( \frac{1}{T} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \right) \leq \mu_s \left( \frac{2\Delta}{T} - \frac{2(1-\delta \eta)}{T} \sum_{t<T} (X_t, \nabla f(x_t)) + \frac{\beta \eta}{T} \sum_{t<T} \| X_t \|^2 \right) \leq \mu_s \left( \sum_{t<T} A_t \right),$$

(19)

where $A_t = -\frac{2(1-\beta \eta)}{T} (X_t, \nabla f(x_t)) + \frac{\beta \eta}{T} \| X_t \|^2$. Using Lemma 18 with $X = \frac{\beta \eta}{T} \| X_t \|^2$ and $Y = -\frac{2(1-\beta \eta)}{T} (X_t, \nabla f(x_t))$, we have $\mu_s(A_t \mid \mathcal{F}_t) \leq \frac{2(1-\beta \eta)}{T} m_u + \frac{\beta \eta}{T} \sigma_v^2$, where $u, v$ are defined as in Theorem 8 and $m_s, \sigma_v^2$ as in Assumption 7. Finally, we use Proposition 3 to bound the sums over iterations:

$$\mu_s \left( \sum_{t<T} A_t \right) = \mu_s \left( \sum_{t<T} A_t \mid \mathcal{F}_{T-1} \right) \leq \mu_s \left( \sum_{t<T-1} A_t \mid \mathcal{F}_{T-1} \right) \leq \mu_s \left( \sum_{t<T-1} A_t + \frac{2(1-\beta \eta)}{T} m_u + \frac{\beta \eta}{T} \sigma_v^2 \right) \leq 2 \left( 1 - \beta \eta \right) m_u + \beta \eta \sigma_v^2,$$

(20)

which concludes the proof.

In order to simplify our convergence rates, we will use the following lemma.

Lemma 19. Let $a, b, c, p > 0$ and $f(x) = ax^p + bx$ Then, with $x^* = \min \left\{ \left( \frac{b}{pa} \right)^{\frac{1}{p+1}}, c \right\}$, we have

$$f(x^*) \leq (1 + p^{-1})bc^{-1} + (1 + p)p^{\frac{1}{p+1}} a^{\frac{1}{p+1}} b^{\frac{1}{p+1}}.$$

(21)

Proof. When $b < pac^{1+p}$, we have $x^* = \left( \frac{b}{pa} \right)^{\frac{1}{p+1}}$ and $f(x^*) = \left( \frac{b}{pa} \right)^{\frac{p+1}{p}}$. Otherwise, we have $x^* = c$ and $f(x^*) = ac^p + bc \leq (1 + p^{-1})b/c$. Hence, $f(x^*)$ is inferior to the sum of both terms. 

Proof of Theorem 11. First, note that all the r.v. are integrable since the variance of the noise is bounded. Hence, for all the considered r.v. $X$, we have $\mu_0(X) = \mathbb{E}[X]$ (see Proposition 3), and Theorem 8 gives us that $\mathbb{E} \left[ (1/T) \cdot \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \right] \leq \frac{2\Delta}{T} + \beta \eta \sigma_v^2$ when $s = 0$. Minimizing the right-hand side term over $\eta \in (0, 1/\beta]$ using Lemma 19 leads to the desired result.

Proof of Theorem 12. Using Proposition 3, we have $\lim_{a \to +\infty} \mu_s (\| X_t \|^2 \mid \mathcal{F}_t) = \text{ess sup} \| X_t \|^2 \leq B^2$. Theorem 8 with $s \to +\infty$ and $\eta = 1/\beta$ thus gives $\text{ess sup} \left\{ (1/T) \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \right\} \leq \frac{2\Delta}{T} + B^2$. 

2
Proof of Theorem 14. By definition of sub-exponential r.v., we have, \( \forall s \in (0, 1/c), \mu_s(-\langle X_t, \nabla f(x_t) \rangle | F_t) \leq a \sigma^2 s/2 \) and \( \mu_s(\|X_t\|^2 | F_t) \leq (1 + b/2c)\sigma^2 \). Using Proposition 5 and Theorem 8 we thus have, \( \forall x, s > 0 \) such that \( u = \frac{4s}{T} \leq 1/c \) and \( v = \frac{2\beta s}{T} \leq 1/c \),

\[
P \left( \frac{1}{T} \sum_{t=0}^{T-1} \| \nabla f(x_t) \|^2 \geq \frac{2\Delta}{\eta T} + 2(1 - \beta \eta)m_u + \beta \eta \sigma_u^2 + x \right) \leq e^{-sx}. \tag{22}
\]

where \( m_u = a \sigma^2 u/2 \) and \( \sigma_u^2 = (1 + b/2c)\sigma^2 \). Hence, if \( \eta \in (0, 1/\beta) \), we have, with probability at least \( 1 - \delta \),

\[
\frac{1}{T} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \leq \frac{2\Delta}{\eta T} + \frac{4a \sigma^2 s}{T} + (1 + b/2c)\beta \eta \sigma^2 + \frac{1}{s} \ln(1/\delta). \tag{23}
\]

Optimizing over \( \eta \) and \( s \) gives \( \beta \eta = \min \left\{ \sqrt{\frac{2\beta \Delta}{(1 + b/2c)\sigma^2}}, 1 \right\} \) and \( \frac{4s}{T} = \min \left\{ \sqrt{\frac{4c^2 \ln(1/\delta)}{a \sigma^2 T}}, 1 \right\} \).

Using Lemma 19, Eq. (23) thus becomes

\[
\frac{1}{T} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \leq \frac{4\beta \Delta + 8c \ln(1/\delta)}{T} + \frac{8(1 + b/2c)\beta \Delta \sigma^2}{T} + 4\sigma \sqrt{\frac{a \ln(1/\delta)}{T}}. \tag{24}
\]

Proof of Proposition 16. Using the second concentration inequality of Proposition 5 we have, \( \forall x \geq \sqrt{\mathcal{E}} \),

\[
P(\|X_t\| \geq x | F_t) = P(\|X_t\|^2 \geq x^2 | F_t) \leq \frac{2\mu_{-1/x^2}(\|X_t\|^2 | F_t)}{x^2} \leq 2ax^{-b}. \tag{25}
\]

Proof of Theorem 17. We first bound the biased mean in both settings. If the noise is symmetric, then \( \mathbb{E} \left[ -\langle X_t, \nabla f(x_t) \rangle e^{-s\|X_t\|^2} | F_t \right] = 0 \) and, for \( s > 0, m_{-s} = 0 \) verifies Assumption 7. Otherwise, we use the following Lemma.

Lemma 20. If \( f \) is L-Lipschitz and Assumption 7 is verified, then, \( \forall s \in [0, 1/c] \),

\[
\frac{\mathbb{E} \left[ -\langle X_t, \nabla f(x_t) \rangle e^{-s\|X_t\|^2} | F_t \right]}{\mathbb{E} \left[ e^{-s\|X_t\|^2} | F_t \right]} \leq \kappa_6 Ls^{b-1} \left( e^\frac{a}{2} + \frac{4ab}{(b-1)(3-b)} \right). \tag{26}
\]

where \( \kappa_6 = (1 - ac^{-b/2})^{-1} \left( e^\frac{b}{2} + \frac{4ab}{(b-1)(3-b)} \right) \).

Proof. First, we have

\[
\mathbb{E} \left[ e^{-s\|X_t\|^2} | F_t \right] = e^{-s\mu_{-s}(\|X_t\|^2 | F_t)} \geq 1 - s\mu_{-s}(\|X_t\|^2 | F_t) \geq 1 - as^{b/2} \geq 1 - ac^{-b/2}. \tag{27}
\]

Then, let \( Y = -\langle X_t, \nabla f(x_t) \rangle \). As \( \mathbb{E} [Y | F_t] = 0 \), we have

\[
\mathbb{E} \left[ Ye^{-s\|X_t\|^2} | F_t \right] = \mathbb{E} \left[ Y_+ e^{-s\|X_t\|^2} | F_t \right] - \mathbb{E} \left[ Y_- e^{-s\|X_t\|^2} | F_t \right] \leq \mathbb{E} \left[ Y_+ | F_t \right] - \mathbb{E} \left[ Y_- e^{-s\|X_t\|^2} | F_t \right] \leq \mathbb{E} \left[ Y_- \left( 1 - e^{-s\|X_t\|^2} \right) | F_t \right] \leq \mathbb{E} \left[ \|X_t\| \left( 1 - e^{-s\|X_t\|^2} \right) | F_t \right], \tag{28}
\]

3
as \( Y_\leq |\langle X_t, \nabla f(x_t) \rangle| \leq L \|X_t\| \). Finally, we bound \( \mathbb{E} \left[ \|X_t\| \left( 1 - e^{-s\|X_t\|^2} \right) \mid \mathcal{F}_t \right] \) by using the function \( g(x) = x \left( 1 - e^{-sx^2} \right) \). As \( g \) is monotonically increasing, we have

\[
\mathbb{E} \left[ \|X_t\| \left( 1 - e^{-s\|X_t\|^2} \right) \mid \mathcal{F}_t \right] = \mathbb{E} \left[ g(\|X_t\|) \right] = \int_0^{+\infty} \mathbb{P} \left( g(X) > x \right) dx = \int_0^{+\infty} \mathbb{P} \left( X > g^{-1}(x) \right) dx = \int_0^{g(\sqrt{s})} \mathbb{P} \left( X > g^{-1}(x) \right) dx + \int_{g(\sqrt{s})}^{+\infty} \mathbb{P} \left( X > g^{-1}(x) \right) dx \leq g(\sqrt{c}) + 2a \int_0^{+\infty} \frac{1}{\sqrt{T}} dx \leq \frac{se^{3/2}}{\frac{b}{(b-1)(3-b)}} s^{\frac{b-1}{2}},
\]

where the second inequality comes from \( g(x) \leq sx^3 \) and \( g(x) \geq \min \{x, (x/s)^{1/3}\} \), and the last inequality from \( s \leq 1/c \).

Hence, we can use \( m_{-s} = \kappa_6 L s^{\frac{b-1}{2}} \), with the special case \( L = 0 \) if \( X_t \) is symmetric, in order to describe both settings. Using Theorem 8 and Assumption 15, we thus have, for \( s \in \left[ 0, \frac{T}{\beta \eta T} \right] \),

\[
\mu_{-s} \left( \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \right) \leq \frac{2\Delta}{\eta T} + \beta \eta a \left( \frac{\beta \eta s}{T} \right)^{\frac{b-2}{2}} + \kappa_6 L \left( \frac{\beta \eta s}{T} \right)^{\frac{b-1}{2}}.
\]

We obtain a concentration inequality using Proposition 5, leading to, \( \forall x > 0 \),

\[
\mathbb{P} \left( \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \geq x \right) \leq \frac{1 + sx}{x} \left[ \frac{2\Delta}{\eta T} + \beta \eta a \left( \frac{\beta \eta s}{T} \right)^{\frac{b-2}{2}} + \kappa_6 L \left( \frac{\beta \eta s}{T} \right)^{\frac{b-1}{2}} \right].
\]

Choosing \( s = \min \left\{ \frac{1}{x}, \frac{T}{\beta \eta T} \right\} \) gives \( \frac{1+sx}{x} \leq \frac{\beta \eta x}{x} \left( \frac{\beta \eta s}{T} \right)^{\frac{b-1}{2}} \leq \left( \frac{\beta \eta}{T} \right)^{\frac{b-2}{2}} \leq \left( \frac{\beta \eta}{T} \right)^{\frac{b-2}{2}} + c^{\frac{2\Delta}{T}} x. \)

Hence, with \( y = \frac{\beta \eta}{T} \), we have

\[
\mathbb{P} \left( \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \geq x \right) \leq \frac{2\beta \Delta}{T^2 T} + aTxy^2 + ac^{\frac{2\Delta}{T}} Txy + \kappa_6 L y^{\frac{b-1}{2}},
\]

and optimizing the first two terms over \( y \) (and thus \( \eta \)) gives \( y = \min \left\{ \left( \frac{4\beta \Delta}{aT^2x} \right)^{\frac{2\Delta}{T}} \right\} \). Lemma 19 then gives

\[
\mathbb{P} \left( \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \geq x \right) \leq A + B + C + D,
\]

where

\[
\begin{align*}
A &= \frac{4(2b+1)\beta \Delta}{bT^2} \\
B &= (2 + b) \left( \frac{\beta \Delta}{T} \right)^{\frac{2b}{T}} x^{-1}(aT) \frac{2\beta \Delta}{T^2 T} \left( \frac{4\beta \Delta}{aT^2x} \right)^{\frac{1}{2\Delta}} \\
C &= 2ac^{\frac{2\Delta}{T}} T \left( \frac{4\beta \Delta}{aT^2x} \right)^{\frac{1}{2\Delta}} \\
D &= 2\kappa_6 L \left( \frac{4\beta \Delta}{aT^2x} \right)^{\frac{1}{2\Delta}}.
\end{align*}
\]

Using \( A + B + C + D \leq \max\{4A, 4B, 4C, 4D\} \) and bounding the previous term by \( \delta \), we get, with probability \( 1 - \delta \),

\[
\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{\kappa_2 \beta \Delta}{T^2} + \frac{\kappa_3 \sqrt{\beta \Delta}}{T^{1+\frac{\delta}{2}} \delta^{\frac{2\Delta}{T}}} + \frac{\kappa_4 L^{\frac{2\Delta}{T^2}} (\beta \Delta)^{\frac{b-1}{2}}}{T^{1+\frac{\delta}{2}} \delta^{\frac{2\Delta}{T}}} + \frac{\kappa_5 \sqrt{\beta \Delta}}{T^{1+\frac{\delta}{2}} \delta^{\frac{2\Delta}{T}}},
\]

as desired.
where
\[
\begin{align*}
\kappa_2 &= 16(2 + b)/b & \leq 36 \\
\kappa_3 &= 2 \cdot 8^{2+h} c^{4-\frac{h}{2}} b^{-\frac{1}{2}} a^\frac{b}{4} & \leq 12 c^{\frac{4-h}{2}} a^\frac{b}{4} \\
\kappa_4 &= 2^{\frac{2+h}{2+h}} \kappa_6^{\frac{2+h}{4}} (ab)^{\frac{1-h}{4}} & \leq 8 \kappa_6^{\frac{2+h}{4}} a^\frac{1-h}{4} \\
\kappa_5 &= 2 \cdot (4 \cdot (2 + b))^{\frac{2+h}{2}} b^{-1/2} a^{1/b} & \leq 84 a^{1/b} \\
\kappa_6 &= (1 - ac^{-b/2})^{-1} \left( c^{\frac{b}{2}} + \frac{4ab}{(b-1)(3-b)} \right) & \leq (1 - ac^{-b/2})^{-1} \left( c^{\frac{b}{2}} + \frac{8a}{b-1} \right)
\end{align*}
\]