First, we would like to thank all the reviewers for their valuable feedback. We sincerely believe that the document was substantially improved by taking their comments into consideration. Due to space limitations, we now address the main remarks regarding motivation, context, theoretical results and experimental setup.

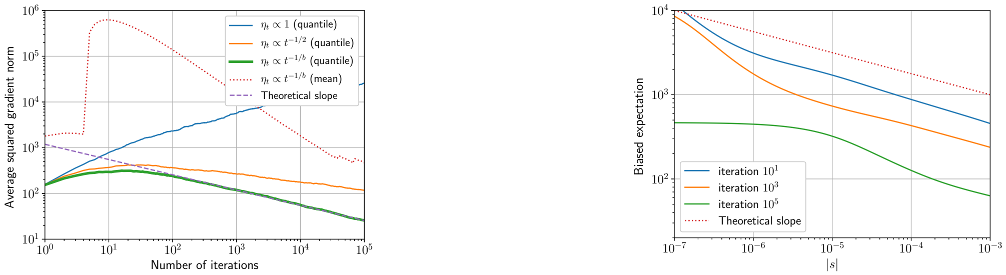
Motivations. In the camera-ready version, we will extend the introduction to include more information on the motivations of this work. More precisely, probabilistic bounds (i.e., on quantiles) provide three advantages over in-expectation bounds: **(1)** First, they allow to consider heavy-tailed noise distributions with infinite/undefined expectation. This setting was recently shown (Zhang et al., 2020 ; Şimşekli et al., 2020) to appear when training NLP models such as BERT over large corpora, and vision models such as AlexNet on Cifar10. **(2)** As a result, SGD may present instabilities that are often solved by running the optimization multiple times (a technique refered to as *multi-start*). Our analysis in probability explains why such a method works by showing that at least 1 out of X runs of SGD will exhibit good convergence and not be disrupted by extreme noise. **(3)** Finally, our work provides a simple and unified analysis under a large class of noise assumptions. We believe that its generality and simplicity could be useful for subsequent research.

Context and previous work. The additional ninth page of the camera-ready version will be used to provide a more extensive discussion of previous work and how our analysis differs from these works. More specifically, we will provide optimal convergence rates for the gradient norm $\|\nabla f(x_t)\| \leq \varepsilon$ for non-convex and smooth optimization ($O(\varepsilon^{-2})$, Carmon et al., 2019) and stochastic optimization ($\tilde{O}(\varepsilon^{-2})$ up to polylogarithmic factors, Foster et. al, 2019), as well as the standard (and suboptimal) rate for SGD ($O(\varepsilon^{-4})$, e.g. [20]) when the variance is bounded. We will also provide details on the methods to obtain fast convergence in the stochastic setting: SVRG (Reddi et al., 2016), SCSG (Lei et al., 2017), SGD4 (Allen-Zhu, 2018), NEON2 (Allen-Zhu and Li, 2018), along with their convergence rates. We will also mention the existence of algorithms with convergence guarantees to a local minimum instead of a stationary point (Allen-Zhu and Li, 2018 ; Fang et al., 2019). Our analysis allows to extend SGD convergence rates to heavy-tailed distributions, as well as quantiles instead of expectations (the motivation for both is discussed in the above paragraph). In particular, we extend the work [20] that also covers the sub-exponential and in-expectation cases, but not the heavy-tailed setting (note that Assumption A2 in [20] is equivalent to $\mu_{1/\sigma^2}(\|X_t\|^2) \leq \sigma^2$). We thank the reviewers for pointing out this weakness in our submission.

Bound on the minimum instead of current iterate. As pointed out by Rev. 1, obtaining tight convergence rates for the iterates $\|\nabla f(x_t)\|^2$ is hard, and most non-convex analyses focus on their minimum or average over time. However, we do agree that the iterate with minimum gradient norm can be hard to find in practice, and we have thus decided to extend all our results to the average $\frac{1}{t}\sum_{i\leq t}\|\nabla f(x_t)\|^2$. This extension is direct given our proofs and is a step towards bounds on the current iterate. We will also mention in a remark that all results imply convergence of the minimum.

Experiments. As pointed out by most reviewers, the experiments were not sufficiently conclusive, and potentially misleading. We have thus decided to replace them by more extensive experiments on a single ML application: ridge regression with a heavy-tail (Student's t) noise distribution of tail-index $b = 1.5$. The experiments were run 1000 times in order to better approximate expectations, quantiles, and biased expectations. Figure 1 shows several aspects of the experiments: (1) The expectation of $\frac{1}{t}\sum_{i\leq t}\|\nabla f(x_i)\|^2$ reaches extremely large values (infinite in theory) compared to quantiles. (2) The choice of $\eta_t \propto t^{-1/b}$ leads to the $t^{(b-1)/b}$ convergence rate of Theorem 5. (3) Standard step-sizes $\eta_t \propto t^{-1/2}$ and $\eta_t \propto 1$ (and independent of the desired precision ε) lead to suboptimal convergence rates, indicating that the choice $\eta_t \propto t^{-1/b}$ may be valuable for practitioners when the noise distribution is particularly fat-tailed. (4) Biased expectations $\mu_{-s}(\frac{1}{t}\sum_{i\leq t}\|\nabla f(x_i)\|^2)$ are well aligned with theory ($s^{(b-2)/2}$ in Eq. 13).

Additional remarks. **(1)** $s = 0$ will be carefully replaced by the notation $s \to 0$. **(2)** Far from a stationary point, SGD indeed converges in $O(1/\sqrt{t})$ for $\|\nabla f(x_t)\|$ (see for example [20, Corollary 2.2] that exhibits the exact same behavior).



(a) Mean and 50%-quantiles of $\frac{1}{t}\sum_{i\leq t}\|\nabla f(x_i)\|^2$.

(b) Biased expectation $\mu_{-s}(\frac{1}{t}\sum_{i\leq t}\|\nabla f(x_i)\|^2)$.

Figure 1: SGD on a ridge regression problem with heavy-tail (Student's t) distribution of tail-index $b = 1.5$.