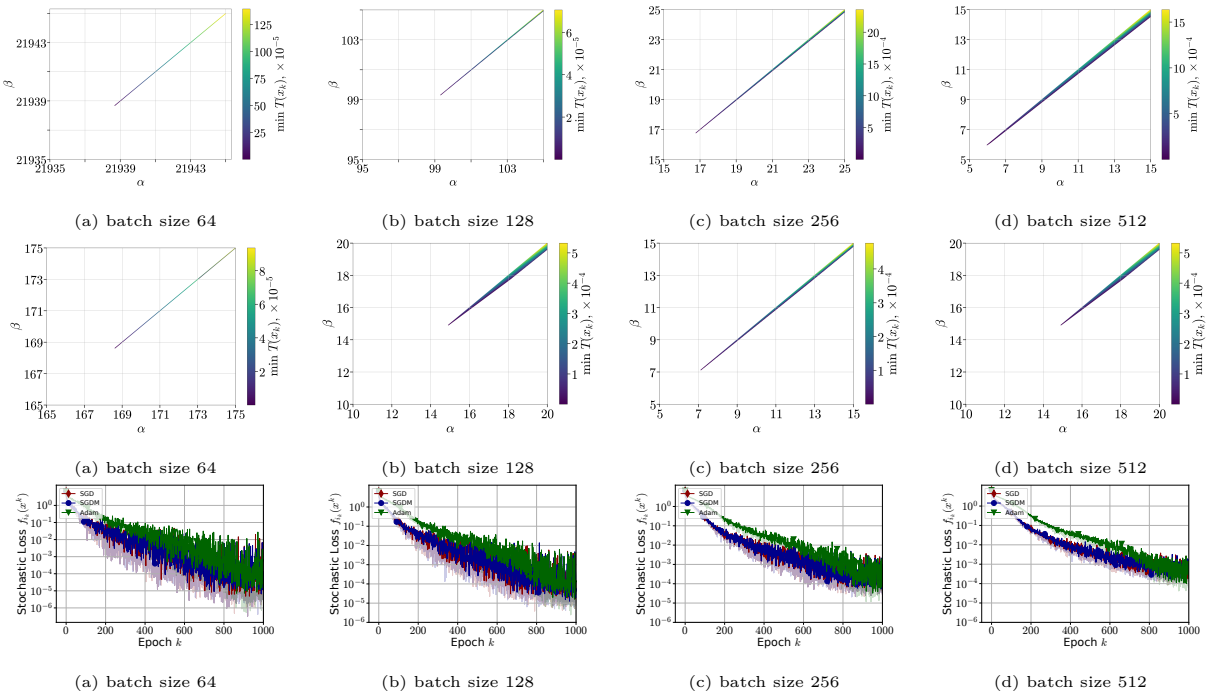

(a) batch size 64 (b) batch size 128 (c) batch size 256 (d) batch size 512

(a) batch size 64 (b) batch size 128 (c) batch size 256 (d) batch size 512

(a) batch size 64 (b) batch size 128 (c) batch size 256 (d) batch size 512

Figure 1: Training of ResNet9 model on CIFAR100 dataset varying the batch size SGD (stepsize 0.01 and momentum 0.9 with OneCycle learning rate scheduler), for SGDM (stepsize 0.01 and momentum 0.9 with OneCycle learning rate scheduler) and Adam (stepsize 0.0001 with default momentum parameters with OneCycle learning rate scheduler) optimizers. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$. Minimum is taken across all runs and iterations for given pair of (α, β). We plot values of α,β in α-β-condition for SGDM (**first row**) and Adam (**second row**), and the average stochastic loss along with minimum/maximum fluctuations across 3 runs (**third row**).

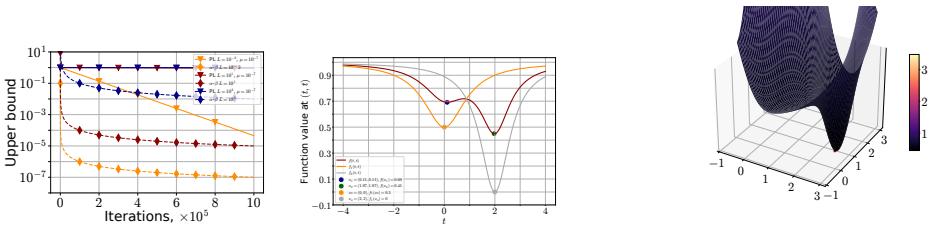

Figure 2: Left: Theoretical convergence of optimization terms in the rates of SGD under PL and α-β-condition ($\frac{L}{K(\alpha-\beta)}$ under α-β-condition and $(1-\frac{\mu}{L})^K$ under PL condition). Center and right: loss landscape for new toy example provided in the general rebuttal.

Size of 2nd layer in MLP	32	128	512	1024	2048	4096
$\beta\sigma_{\text{int}}^2$	16.6	1.9	6.4	2.9	1.1	0.7

# of convolutions in CNN	32	64	128	256	512	1024
$\beta\sigma_{\text{int}}^2$	0.42	0.16	0.11	0.06	0.0007	0.0006

Batch size for Resnet	64	128	256	512
$\beta\sigma_{\text{int}}^2$	0.5	0.03	0.001	0.0008

Table 1: We can very roughly estimate the value of $\beta\sigma_{\text{int}}^2$ based on experiments. The value of σ_{int}^2 can be approximated as the value of the stochastic loss at the end of the training (since we use $x^K \approx x_p$, then $f_i(x^K) \approx f_i(x_p)$). Therefore, assuming that $f_i^* = 0$ we get that $\sigma_{\text{int}}^2 \approx f_i(x^K)$. Using such approximations, we compute the value $\beta\sigma_{\text{int}}^2$ following the experimental setup of the paper.