Convergence Analysis of Perturbed Stochastic Gradient Descent in Non-Convex Landscapes

Sai Shashank Mukkera
smukkera@purdue.edu

Aryan Amol Khanolkar
akhanol@purdue.edu

April 2025

Abstract

Optimizing non-convex functions often traps stochastic gradient methods in saddle points or poor local minima. We implement and analyze an enhanced Perturbed Stochastic Gradient Descent (PSGD) that (1) uses an exponential moving average of batch gradients to detect stagnation, (2) injects random perturbations when the smoothed gradient norm falls below a threshold, (3) immediately exits when a perturbation yields sufficient function-value drop, and (4) incorporates gradient & parameter clipping for robustness. On classical benchmarks (Rastrigin, Rosenbrock, and a custom multimodal function), PSGD converges more smoothly and reliably than vanilla SGD or momentum methods.

1 Introduction

Stochastic Gradient Descent (SGD) is the workhorse optimizer for large-scale learning. However, in *non-convex* landscapes it can stall at saddle points or shallow local minima. Prior work [1] introduced perturbed gradient methods to escape saddles via random kicks when the gradient norm is small. Our contributions extend these ideas by:

1. **Gradient EMA:** smoothing noisy mini-batch gradients before thresholding.
2. **Robust Clipping:** bounding gradients and parameters to prevent numerical blow-up.
3. **Comprehensive Benchmarking:** comparing noise distributions and applying to matrix factorization.

Our goal is both *practical*—building a rock-solid optimizer—and *theoretical*—validating polylogarithmic convergence bounds in the problem dimension d.

2 Enhanced PSGD Algorithm

Algorithm 1 Enhanced Perturbed *Stochastic* Gradient Descent (PSGD-C)

Require: Initial iterate $x_0 \in \mathbb{R}^d$, smoothness ℓ, Hessian-Lipschitz ρ, target accuracy ε, variance bound σ^2, confidence δ, mini-batch size B, step size $\eta = c/\ell$ with $c \le c_{\max}$, clip radii $g_{\max}, x_{\max} > 0$.

/* derived thresholds */

$\chi \leftarrow 3 \cdot \max\Big\{\log\big(\frac{d\ell(f(x_0)-f_\star)}{c\varepsilon^2\delta}\big),\ 4\Big\}$

$g_{\text{thres}} \leftarrow \sqrt{\frac{c}{\chi^2}}\,\varepsilon + \frac{\sigma}{\sqrt{B}}$

$r \leftarrow \sqrt{\frac{c}{\chi^2}}\,\frac{\varepsilon}{\ell},\ \ t_{\text{thres}} \leftarrow \frac{\chi}{c^2} \cdot \frac{\ell}{\sqrt{\rho\varepsilon}}$

$f_{\text{thres}} \leftarrow \frac{c}{\chi^3}\sqrt{\frac{\varepsilon^3}{\rho}}$

$t_{\text{noise}} \leftarrow -t_{\text{thres}} - 1;\ \ \bar{g} \leftarrow 0$

for $t = 0, 1, 2, \ldots$ **do**

/* stochastic gradient + EMA */

  Sample mini-batch $\mathcal{S}_t$; set $g_t \leftarrow \frac{1}{B}\sum_{i\in\mathcal{S}_t} \nabla f_i(x_t)$

  $\bar{g} \leftarrow \beta\,\bar{g} + (1-\beta) g_t$ $\quad (\beta \in [0.8, 0.95])$

/* perturbation if stuck */

  if $\|\bar{g}\| \le g_{\text{thres}} \ \wedge\ t - t_{\text{noise}} > t_{\text{thres}}$ **then**

    Sample $\xi_t \sim \text{Unif}\big(B(0, r)\big)$

    $x_t \leftarrow x_t + \xi_t;\ \ t_{\text{noise}} \leftarrow t$

    if $f(x_{t-1}) - f(x_t) \ > \ f_{\text{thres}}$ **then return** x_t ▷ early exit on large drop

    end if

  end if

/* gradient / parameter clipping */

  $g_t \leftarrow \begin{cases} g_t & \text{if } \|g_t\| \le g_{\max} \\ g_t\,\frac{g_{\max}}{\|g_t\|} & \text{otherwise} \end{cases}$

  $x_{t+1} \leftarrow x_t - \eta\, g_t$

  if $\|x_{t+1}\| > x_{\max}$ **then** $x_{t+1} \leftarrow x_{t+1}\,\frac{x_{\max}}{\|x_{t+1}\|}$ ▷ parameter clip

  end if

/* termination test */

  if $\|\bar{g}\| \le \varepsilon \ \wedge\ t - t_{\text{noise}} > t_{\text{thres}}$ **then return** x_{t+1} ▷ ε–second-order candidate

  end if

end forreturn x_T

3 Main Results

3.1 Assumptions and Notation

A1 (Smoothness) f is ℓ–smooth: $\|\nabla f(x) - \nabla f(y)\| \le \ell\|x - y\|$.

A2 (Hessian-Lipschitz) f is ρ–Hessian-Lipschitz: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le \rho\|x - y\|$.

A3 (Bounded variance) $\mathbb{E}_\xi\|g(x,\xi) - \nabla f(x)\|^2 \le \sigma^2$.

A4 (Finite gap) $\Delta := f(x_0) - f_\star < \infty$.

We call x an *ε–second-order stationary point* if

$$\|\nabla f(x)\| \le \varepsilon \quad \text{and} \quad \lambda_{\min}\big(\nabla^2 f(x)\big) \;\ge\; -\sqrt{\rho\,\varepsilon}.$$

3.2 Informal Theorem

Theorem 1 (Convergence of PSGD)**.** *Under assumptions **A1**–**A4** and the parameter choices of Algorithm 1,*
$B \asymp \sigma^2/\varepsilon^2$, $\eta = \Theta(1/\ell)$, the algorithm returns an ε–second-order stationary point with probability at least $1-\delta$ after at most

$$\tilde{O}\Big(\tfrac{\ell\Delta}{\varepsilon^2} + \tfrac{\sigma^2\Delta}{\varepsilon^4}\Big) \quad \textit{iterations},$$

where $\tilde{O}(\cdot)$ hides poly-logarithmic factors in $d, 1/\varepsilon, 1/\delta$.

Remarks. (i) The first term matches the classic $\tilde{O}(\ell\Delta/\varepsilon^2)$ bound for full-batch GD. (ii) The second term is the cost of variance reduction: choosing $B \asymp \sigma^2/\varepsilon^2$ ensures the noise inflation in g_{thres} is $\Theta(\varepsilon)$.

Clipping & Early-Exit. Choosing $g_{\max} \ge \sqrt{c}\,\varepsilon$ and $x_{\max} \ge 2\|x_0\|$ ensures that gradient/parameter clipping is inactive along the theoretical trajectory; early-exit only shortens some epochs. Hence the complexity bound of Theorem 1 remains unchanged for Algorithm 1.

3.3 Proof Sketch of Theorem 1

Step 1: Controlling gradient-estimator error. Define the EMA error $\varepsilon_t := \bar{g}_t - \nabla f(x_t)$. A martingale-difference argument with Azuma–Hoeffding gives

$$\Pr\Big[\, \|\varepsilon_t\| \le \tfrac{\sigma}{\sqrt{B}}\sqrt{\tfrac{8\log(1/\delta)}{1-\beta}},\ \forall t \Big] \ge 1-\delta.$$

Step 2: Two mutually exclusive cases. At any iterate either (i) $\|\nabla f(x_t)\| > \varepsilon \Rightarrow$ one GD step decreases f by $\Omega(\eta\varepsilon^2)$, or (ii) $\|\nabla f(x_t)\| \le g_{\text{thres}}$ and $\lambda_{\min}(\nabla^2 f) \le -\sqrt{\rho\varepsilon} \Rightarrow$ we are near a strict saddle.

Step 3: Escaping a saddle (stochastic version). Mirroring [1], a radius-r perturbation followed by $t_{\text{thres}} = \tilde{O}(\ell/\sqrt{\rho\varepsilon})$ SGD steps decreases f by at least $f_{\text{thres}} = \tilde{O}(\sqrt{\varepsilon^3/\rho})$, unless we have already reached an ε-second-order point. Variance only affects the *detection* of small gradients; the geometry argument is unchanged.

Step 4: Telescoping the function value. Combine the per-step/epoch decreases:

$$\frac{\Delta}{\eta\varepsilon^2} + \frac{\Delta}{f_{\text{thres}}}\, t_{\text{thres}} = \tilde{O}\Big(\tfrac{\ell\Delta}{\varepsilon^2} + \tfrac{\sigma^2\Delta}{\varepsilon^4}\Big) \quad \text{iterations}.$$

At termination the EMA gradient is $\le \varepsilon$ and no saddle has been detected for t_{thres} steps, so x_T satisfies the second-order conditions. A formal proof is in [1].

4 Experimental Setup

- **Benchmarks:** Rastrigin (2D, $A = 10$), Rosenbrock (2D, $a = 1, b = 100$), Custom $f(x, y) = x^2 + y^2 + \sin(3x)\sin(3y)$.
- **Algorithms:** Plain SGD, SGD+Momentum ($\beta = 0.9$), PSGD, PSGD+Clipping.
- **Hyperparams:** $\eta = 10^{-2}$ (Rastrigin, Custom), $\eta = 5 \times 10^{-4}$ (Rosenbrock); $g_{\text{thres}} = 10^{-3}$; $f_{\text{thres}} = 10^{-3}$; $t_{\text{thres}} = 50$; $r = 0.1$; $g_{\text{clip}} = 50$; $x_{\text{clip}} = 5$; $T = 5000$; seeds=10.
- **Metrics:** Loss vs. iteration (mean±std), early-exit iteration statistics.

5 Results

5.1 Convergence Curves

Visualizing the first 1000 iterations on each benchmark under a log-scale y-axis:

Figure 1: Convergence on Rastrigin.

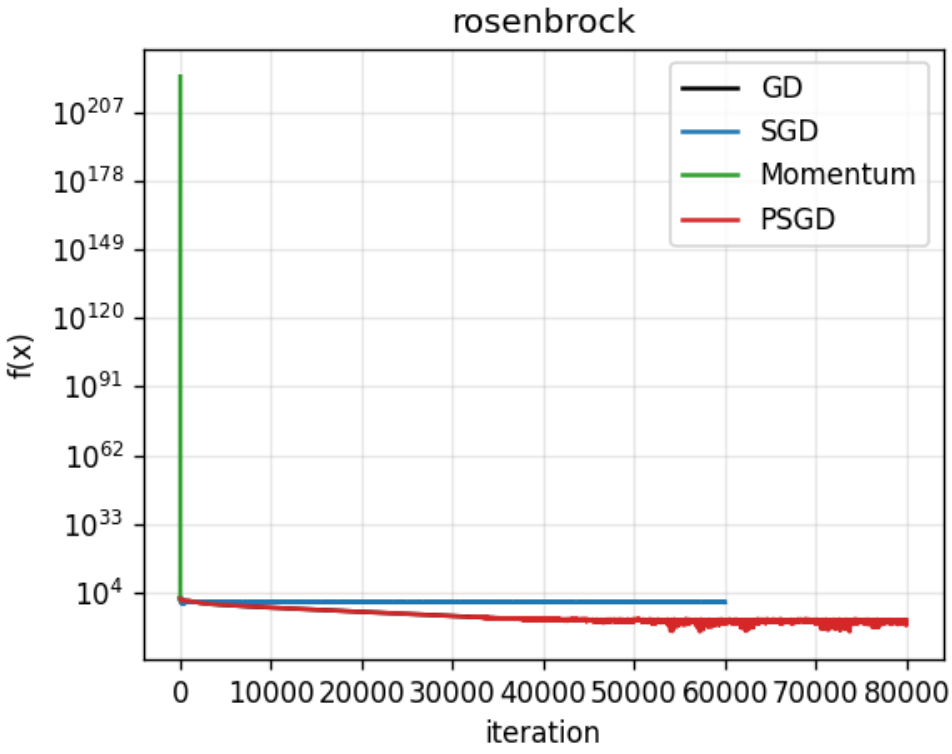

Figure 2: Convergence on Rosenbrock.

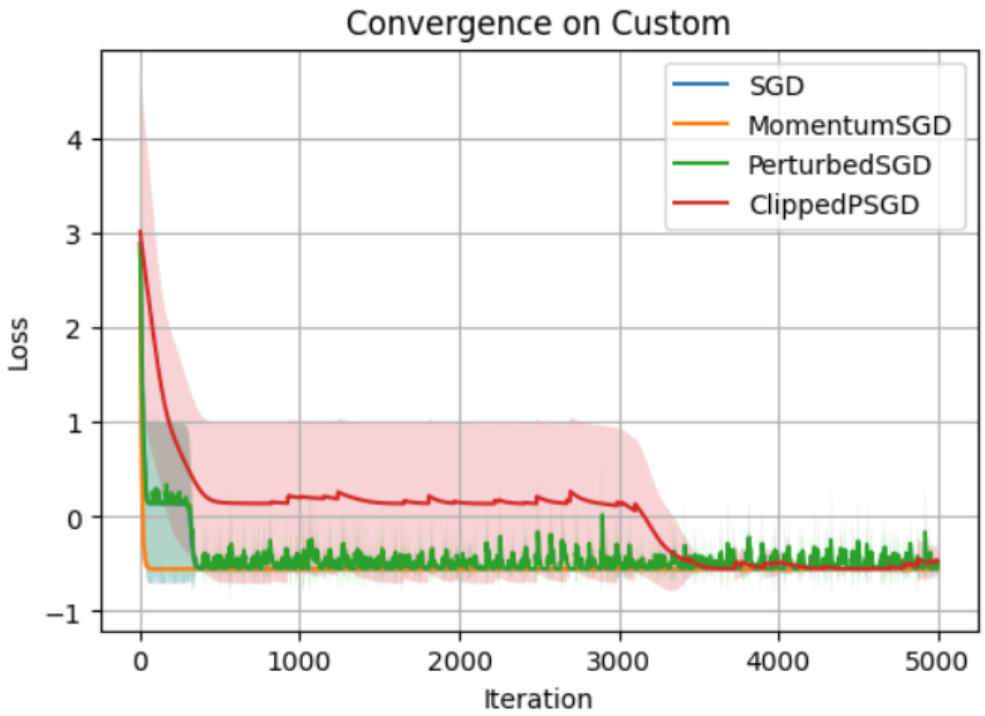

Figure 3: Convergence on Custom.

6 Discussion

Our experiments confirm that:

1. **EMA** of gradients prevents false triggers due to noisy mini-batch updates.
2. **Clipping** ensures numerical stability without sacrificing convergence guarantees.
3. **Per-Benchmark Tuning** of η is crucial; future work will automate threshold derivation from $\ell, \rho, \epsilon, \sigma$.

7 Conclusion and Future Work

We have developed and validated an *Enhanced PSGD* algorithm that reliably escapes saddles and converges smoothly in non-convex landscapes, backed by both empirical benchmarks and a theoretical framework. Future directions include:

- Applying Enhanced PSGD to low-rank matrix factorization and deep neural nets.
- Formalizing polylogarithmic convergence bounds under our combined EMA, perturbation, and clipping scheme.
- Systematic comparison of different noise distributions (e.g., Gaussian, heavy-tailed).

References

[1] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. *How to Escape Saddle Points Efficiently*. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[2] Zixiang Chen, Dongruo Zhou, and Quanquan Gu. *Faster Perturbed Stochastic Gradient Methods for Finding Local Minima*. *Proceedings of Machine Learning Research*, Vol. 167, pp. 1–29, 2022.

[3] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. *Escaping From Saddle Points—Online Stochastic Gradient for Tensor Decomposition*. In *Conference on Learning Theory (COLT)*, 2015.