
Review: Momentum-Based Variance Reduction in Non-Convex SGD

Alan Nawoj

ECE 273 Final Project

University of California San Diego

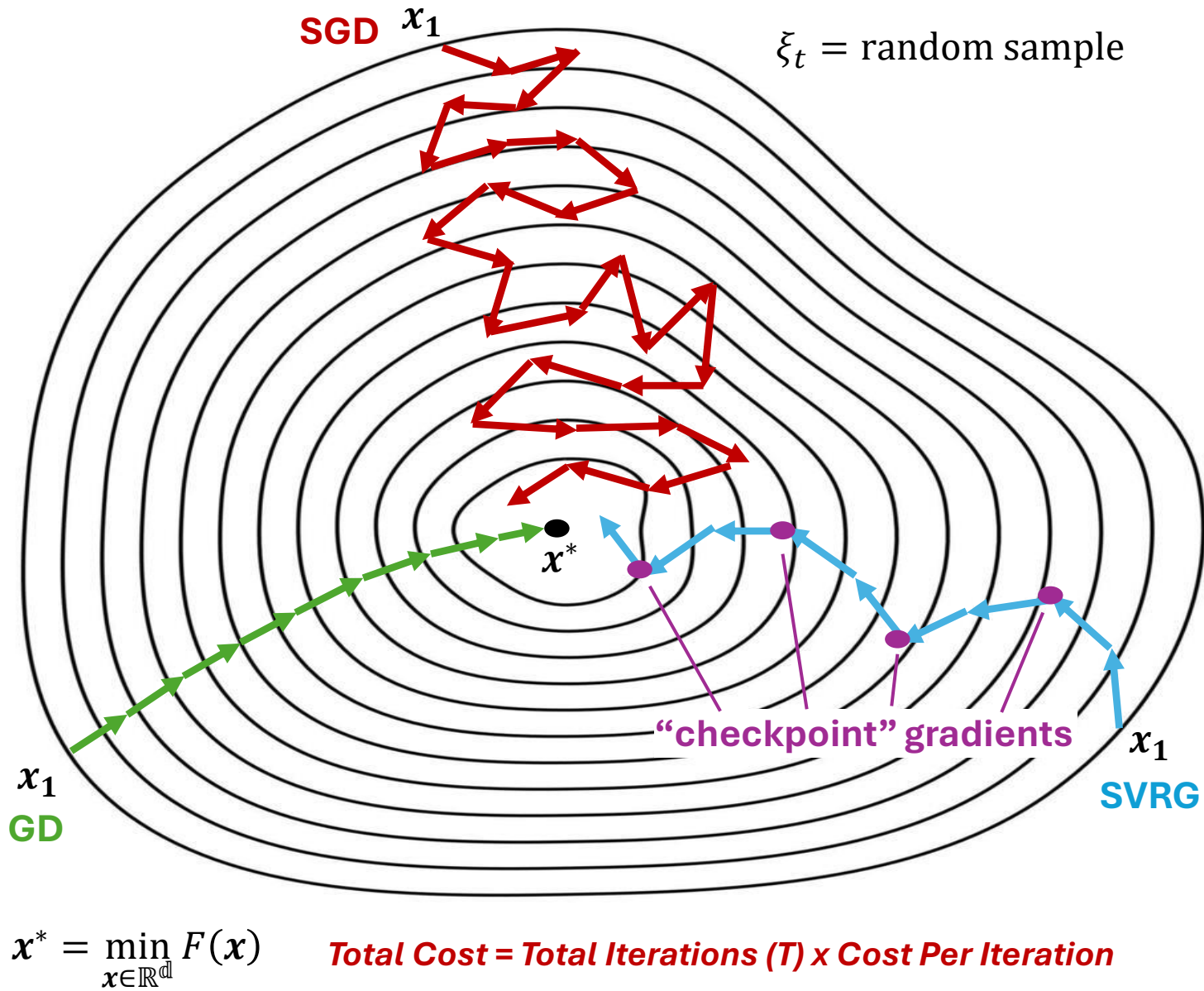
Spring 2026

Outline

- **Background and Motivation**
 - Standard SGD, SVRG, and SVRG Variants
 - SGD with Momentum
- **Description of STORM Algorithm**
 - Architectural Highlights
 - Key Algorithmic Steps
- **Underlying Design Principles**
- **Conclusion**

Appendix: High-Level Summary of Theoretical Analysis

Variance Reduction for Stochastic Gradient Descent (SGD)



- **Full Gradient Descent (GD)**
 $x_{t+1} = x_t - \eta \nabla F(x_t)$
 $\mathbb{E}[\|\nabla F(x_t)\|] = \mathcal{O}(1/T^{1/2})$
- **Stochastic Gradient Descent (SGD)**
 $x_{t+1} = x_t - \eta \nabla f(x_t, \xi_t)$
 $\mathbb{E}[\|\nabla F(x_t)\|] = \mathcal{O}(1/T^{1/4})$
With decayed learning rate
- **SVRG (& Variants: SARAH*, SPIDER*)**
 $g_t = \nabla f(x_t, \xi_t) - \nabla f(v_s, \xi_t) + \nabla F(v_s)$
 $x_{t+1} = x_t - \eta g_t$
 $v_{s+1} = \frac{1}{K} \sum_{k=1}^K x_k$
Periodic large “checkpoint” gradients
 $\mathbb{E}[\|\nabla F(x_t)\|] = \mathcal{O}(1/T^{1/3})^*$

SGD with Momentum (“Heavy Ball”)

Core Update Rule:

$$d_t = \mathbf{a} \nabla f(x_t, \xi_t) + (1 - \mathbf{a}) d_{t-1} \quad \leftarrow \text{Momentum Vector}$$

$$x_{t+1} = x_t - \eta d_t$$

$\mathbf{a} = 0.1$ is typical

Exponential Moving Average (EMA) Example:

$$d_1 = \mathbf{a} \nabla f(x_1, \xi_1)$$

$$d_2 = \mathbf{a} \nabla f(x_2, \xi_2) + (1 - \mathbf{a}) [\mathbf{a} \nabla f(x_1, \xi_1)]$$

$$d_3 = \mathbf{a} \nabla f(x_3, \xi_3) + (1 - \mathbf{a}) [\mathbf{a} \nabla f(x_2, \xi_2) + (1 - \mathbf{a}) [\mathbf{a} \nabla f(x_1, \xi_1)]]$$

$$d_4 = \mathbf{a} \nabla f(x_4, \xi_4) + (1 - \mathbf{a}) [\mathbf{a} \nabla f(x_3, \xi_3) + (1 - \mathbf{a}) [\mathbf{a} \nabla f(x_2, \xi_2) + (1 - \mathbf{a}) [\mathbf{a} \nabla f(x_1, \xi_1)]]]$$

$$= \underline{\mathbf{a} \nabla f(x_4, \xi_4)} + \underline{(1 - \mathbf{a}) \mathbf{a} \nabla f(x_3, \xi_3)} + \underline{(1 - \mathbf{a})^2 \mathbf{a} \nabla f(x_2, \xi_2)} + \underline{(1 - \mathbf{a})^3 \mathbf{a} \nabla f(x_1, \xi_1)}$$

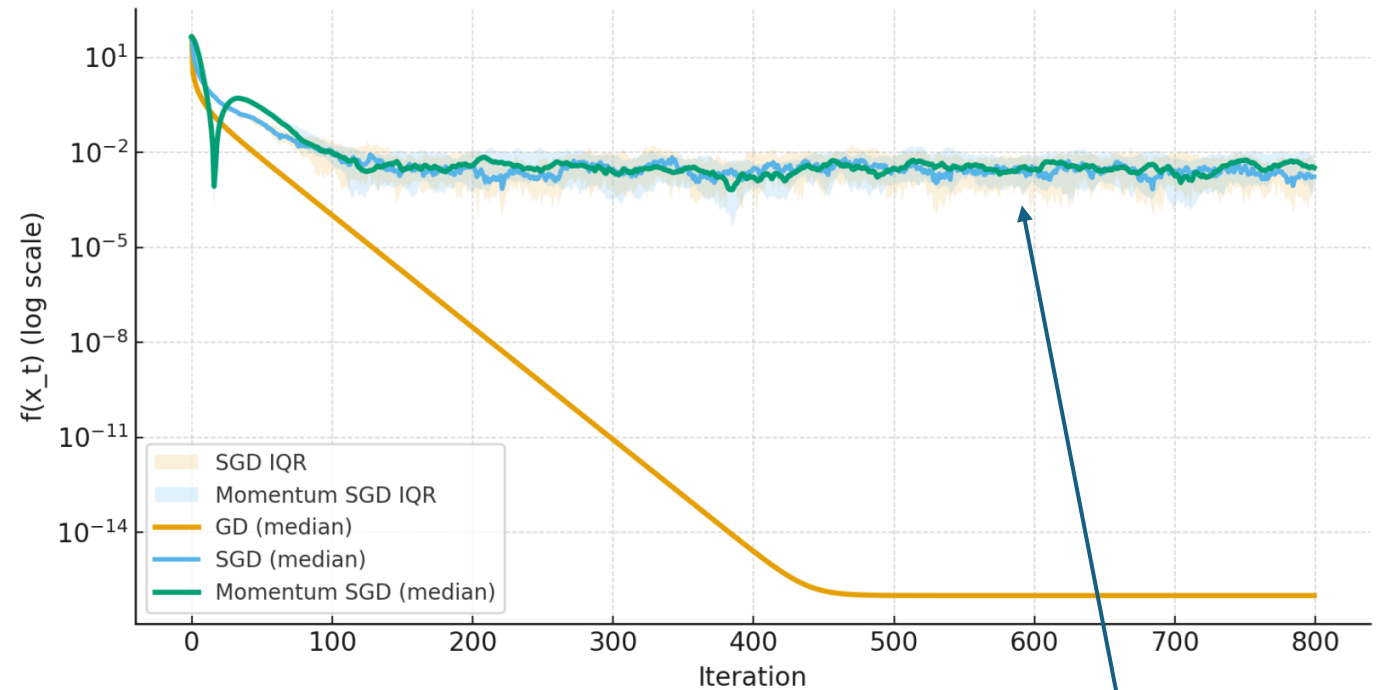
↑
0.1

↑
0.09

↑
0.081

↑
0.0729

Credit: Chen, Tianyi. Distributed Optimization for Machine Learning (Lecture 10), Cornell University, 2025.



SGD with momentum does **not** fundamentally eliminate variance

STOchastic Recursive Momentum (STORM) Overview

STORM achieves the optimal convergence rate of $\mathbb{E}[\|\nabla F(x_t)\|] = \mathcal{O}(1/T^{1/3})$ matched by SVRG algorithms (SARAH, SPIDER) but **without** the need for periodic large checkpoint gradients and hyperparameter tuning!

- **Modified Momentum Vector with Recursive Gradient Differencing**

$$\begin{aligned} d_t &= a_t \nabla f(x_t, \xi_t) + (1 - a_t) d_{t-1} + \boxed{(1 - a_t)(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t))} \\ &= \nabla f(x_t, \xi_t) + (1 - a_t)(d_{t-1} - \nabla f(x_{t-1}, \xi_t)) \end{aligned}$$

- **Adaptive Momentum Coefficient**

$$a_{t+1} = c\eta_t^2$$

$$a \in (0,1)$$

- **Adaptive Learning Rate**

$$\eta_t = \frac{k}{(w + \sum_{i=1}^t \|\nabla f(x_i, \xi_i)\|^2)^{1/3}}$$

$$x_{t+1} = x_t - \eta_t d_t$$

Assumptions

$$\|\nabla F(x_{t+1}) - \nabla F(x_t)\| \leq L\|x_{t+1} - x_t\|$$

$$\|\nabla F(x)\| \leq G$$

$$\mathbb{E}[\|\nabla f(x, \xi_t) - \nabla F(x)\|^2] \leq \sigma^2$$

$$\mathbb{E}[\nabla f(x, \xi_t)] = \nabla F(x)$$

STOchastic Recursive Momentum (STORM) Algorithm

Algorithm 1 STORM: STOchastic Recursive Momentum

```
1: Input: Parameters  $k, w, c$ , initial point  $\mathbf{x}_1$ 
2: Sample  $\xi_1$ 
3:  $G_1 \leftarrow \|\nabla f(\mathbf{x}_1, \xi_1)\|$ 
4:  $\mathbf{d}_1 \leftarrow \nabla f(\mathbf{x}_1, \xi_1)$ 
5:  $\eta_0 \leftarrow \frac{k}{w^{1/3}}$ 
6: for  $t = 1$  to  $T$  do
7: ★  $\eta_t \leftarrow \frac{k}{(w + \sum_{i=1}^t G_i^2)^{1/3}}$  ← Adaptive Step Size, Inversely Proportional to Sum of Stochastic Gradient Norms Seen Thus Far
8: ★  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta_t \mathbf{d}_t$ 
9: ★  $a_{t+1} \leftarrow c\eta_t^2$  ← Adaptive Momentum Coefficient, Proportional to Square of Learning Rate
10: Sample  $\xi_{t+1}$ 
11:  $G_{t+1} \leftarrow \|\nabla f(\mathbf{x}_{t+1}, \xi_{t+1})\|$ 
12: ★  $\mathbf{d}_{t+1} \leftarrow \nabla f(\mathbf{x}_{t+1}, \xi_{t+1}) + (1 - a_{t+1})(\mathbf{d}_t - \nabla f(\mathbf{x}_t, \xi_{t+1}))$  ← Momentum Vector Update (Recursive Variance Reduction)
13: end for
14: Choose  $\hat{\mathbf{x}}$  uniformly at random from  $\mathbf{x}_1, \dots, \mathbf{x}_T$ . (In practice, set  $\hat{\mathbf{x}} = \mathbf{x}_T$ ).
15: return  $\hat{\mathbf{x}}$ 
```

Final Iterate

Underlying Algorithmic Design Principles

- Modified Momentum Vector Recursively Achieves Variance Reduction Without Large Checkpoint Gradients

$$d_t = a_t \nabla f(x_t, \xi_t) + (1 - a_t) d_{t-1} + \boxed{(1 - a_t)(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t))} \leftarrow \text{Same sample, two consecutive points}$$

Decompose stochastic gradient into **signal** plus **noise** terms: $\nabla f(x, \xi) = \nabla F(x) + \epsilon(x, \xi)$

$\rightarrow$ Rewrite difference as: $(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t)) = \underbrace{(\nabla F(x_t) - \nabla F(x_{t-1}))}_{\text{True gradient difference}} + \underbrace{(\epsilon(x_t, \xi_t) - \epsilon(x_{t-1}, \xi_t))}_{\text{Error difference}}$

Same sample evaluated at two *nearby* points implies *high correlation* in noise between the two gradients:

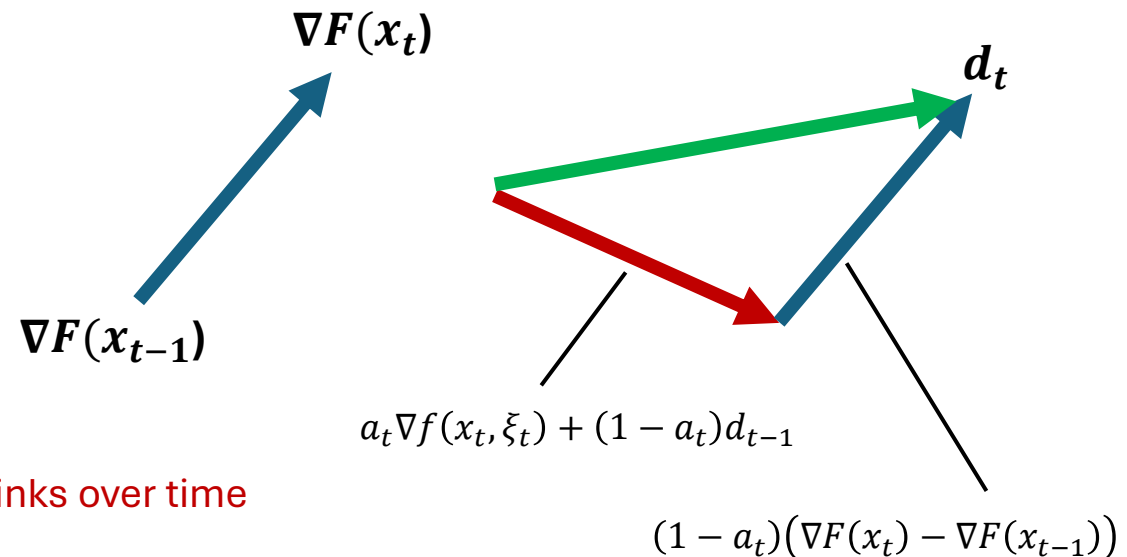
$$(\epsilon(x_t, \xi_t) - \epsilon(x_{t-1}, \xi_t)) \approx 0$$

$$\underbrace{(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t))}_{\substack{\downarrow \\ \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y)}} \approx \underbrace{(\nabla F(x_t) - \nabla F(x_{t-1}))}_{\substack{\downarrow \\ \text{Cov}(X, Y) \approx \text{Var}(X) \approx \text{Var}(Y) \rightarrow \text{Var}(X - Y) \approx 0}}$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y)$$

$$\text{Cov}(X, Y) \approx \text{Var}(X) \approx \text{Var}(Y) \rightarrow \text{Var}(X - Y) \approx 0$$

$$\text{Var}(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t)) \leq \boxed{L^2 \|x_t - x_{t-1}\|^2} \leftarrow \text{Shrinks over time}$$



Underlying Algorithmic Design Principles

- Adaptive Learning Rate

$$\eta_t = \frac{k}{\left(w + \sum_{i=1}^t \|\nabla f(x_i, \xi_i)\|^2\right)^{1/3}}$$

Denominator monotonically increases,
therefore η_t monotonically decreases

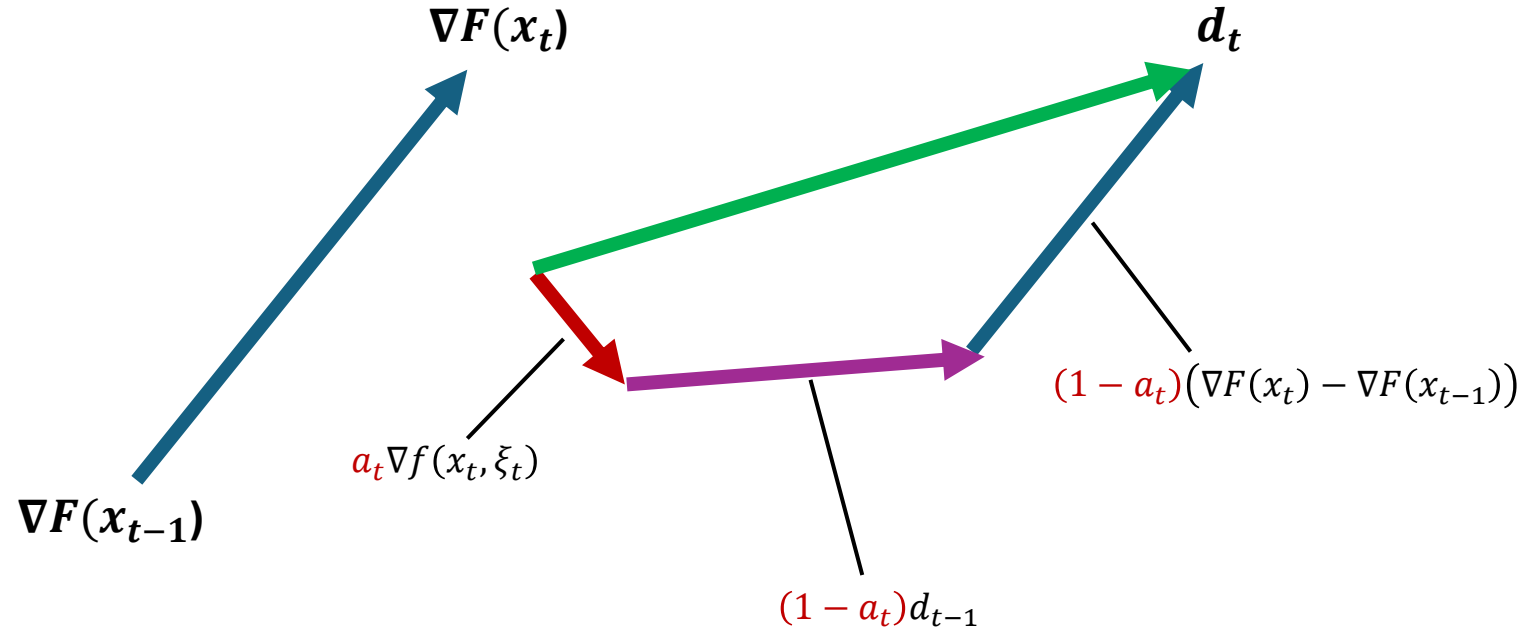
$$x_{t+1} = x_t - \eta_t d_t$$

- Adaptive Momentum Coefficient

$$a_{t+1} = c\eta_t^2 \quad \leftarrow a_{t+1} \text{ drops as } \eta_t \text{ drops}$$

$$a \in (0,1)$$

$$d_t = \underbrace{a_t \nabla f(x_t, \xi_t)}_{\text{Current}} + \underbrace{(1 - a_t) d_{t-1}}_{\text{Previous}} + \underbrace{(1 - a_t) (\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t))}_{\text{Variance Reduction}}$$



Conclusion

STORM (STOchastic Recursive Momentum) is a landmark and paradigm-shifting algorithm that was officially published in NeurIPS 2019 by Ashok Cutkosky and Francesco Orabona and spurred significant follow-up research



Ashok Cutkosky
ECE Assistant Professor
Boston University



Francesco Orabona
ECE Associate Professor
KAUST Saudi Arabia

Key Advancements from **STORM**:

- Achievement of stochastic gradient variance reduction through unique variant on SGD with momentum
- Simpler than SVRG (and variants) due to elimination of periodic massive checkpoint gradients and associated hyperparameters
- Achieves same optimal convergence rate of $\mathcal{O}(1/T^{1/3})$ as some SVRG-based algorithms
- Adaptive learning rate and momentum coefficient enables self-tuning without global knowledge of stochastic gradient variance

Thank You!

Appendix: High-Level Summary of Theoretical Analysis

Goal of Mathematical Proof: $\mathbb{E}[\|\nabla F(x_t)\|] = \mathcal{O}(1/T^{1/2} + \sigma^{1/3}/T^{1/3})$

Step 1: Define Lyapunov Function

$$\Phi_t = \underbrace{[F(\mathbf{x}_t) - F(\mathbf{x}^*)]}_{\text{Optimization Progress}} + \underbrace{\hat{\Phi}_t}_{\text{Noise / Variance}}$$

Optimization Progress

Noise / Variance

$$\text{STORM: } \Phi_t = F(\mathbf{x}_t) + z_t \|\epsilon_t\|^2$$

$$z_t \propto \frac{1}{\eta_{t-1}}$$

$$\|\epsilon_t\|^2 = \|\mathbf{d}_t - \nabla F(\mathbf{x}_t)\|^2$$

Step 2a: Bound Optimization Progress Term $F(\mathbf{x}_t)$

Use L-smoothness, update step, and d_t definitions

$$F(x_{t+1}) \leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2$$

$$x_{t+1} - x_t = -\eta_t d_t$$

$$d_t = \nabla F(x_t) + \epsilon_t$$

Use Young's Inequalities

$$\langle a, b \rangle \leq \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|^2 \quad \text{and} \quad \|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$$

$$\mathbb{E}[F(x_{t+1}) - F(x_t)] \leq \mathbb{E}[-\eta_t/4 \|\nabla F(x_t)\|^2 + 3\eta_t/4 \|\epsilon_t\|^2]$$

Appendix: High-Level Summary of Theoretical Analysis

Step 2b: Bound Variance Term $z_t \|\epsilon_t\|^2$

Seeking to calculate $\mathbb{E}[z_t \|\epsilon_t\|^2] = \mathbb{E}[\|\epsilon_t\|^2 / \eta_{t-1}]$

Substitute $\epsilon_t = d_t - \nabla F(x_t)$ then substitute STORM update $d_t = \nabla f(x_t, \xi_t) + (1 - a_t)(d_{t-1} - \nabla f(x_{t-1}, \xi_t))$

$$\mathbb{E} \left[\frac{\|\epsilon_t\|^2}{\eta_{t-1}} \right] \leq \mathbb{E} \left[2c^2 \eta_{t-1}^3 \sigma^2 + (1 - a_t)^2 (1 + 4L^2 \eta_{t-1}^2) \frac{\|\epsilon_{t-1}\|^2}{\eta_{t-1}} + 4(1 - a_t)^2 L^2 \eta_{t-1} \|\nabla F(\mathbf{x}_{t-1})\|^2 \right]$$

Step 3: Bound on Expected Lyapunov Function Value Rate of Contraction for Single Iteration

Combine bounds from (2a) and (2b) to calculate bound on $\mathbb{E}[\Phi_{t+1} - \Phi_t]$

$$\mathbb{E}[\Phi_{t+1} - \Phi_t] = \mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] + \mathbb{E}[\|\epsilon_{t+1}\|^2 / \eta_t - \|\epsilon_t\|^2 / \eta_{t-1}]$$

$$\mathbb{E}[\Phi_{t+1} - \Phi_t] \leq \mathbb{E} \left[-\eta_t / 4 \|\nabla F(\mathbf{x}_t)\|^2 + 3\eta_t / 4 \|\epsilon_t\|^2 + \frac{1}{32L^2\eta_t} \|\epsilon_{t+1}\|^2 - \frac{1}{32L^2\eta_{t-1}} \|\epsilon_t\|^2 \right]$$

Appendix: High-Level Summary of Theoretical Analysis

Step 4: Summation of Bound on Lyapunov Function Decrease Across All Iterations (1 to T)

Sum both sides of (3) from $t = 1 \rightarrow T$ to calculate final convergence rate of STORM

Telescoping terms on RHS cancel out and LHS terms also telescope, enabling removal of Φ_t terms

$$\sum_{i=1}^T \mathbb{E}[\Phi_{t+1} - \Phi_t] = \mathbb{E}[\Phi_{T+1} - \Phi_1] \geq F^* - F(\mathbf{x}_1)$$

Combine repeating components into new variable M and move expected gradient norm to LHS



$$\mathbb{E}[\|\nabla F(\hat{\mathbf{x}})\|] \leq \frac{w^{1/6}\sqrt{2M} + 2M^{3/4}}{\sqrt{T}} + \frac{2\sigma^{1/3}}{T^{1/3}} = \mathcal{O}(1/T^{1/2} + \sigma^{1/3}/T^{1/3})$$