# Tight analyses of non-smooth stochastic gradient descent

Nicholas J.A. Harvey    Christopher Liaw    Yaniv Plan    Sikander Randhawa

University of British Columbia

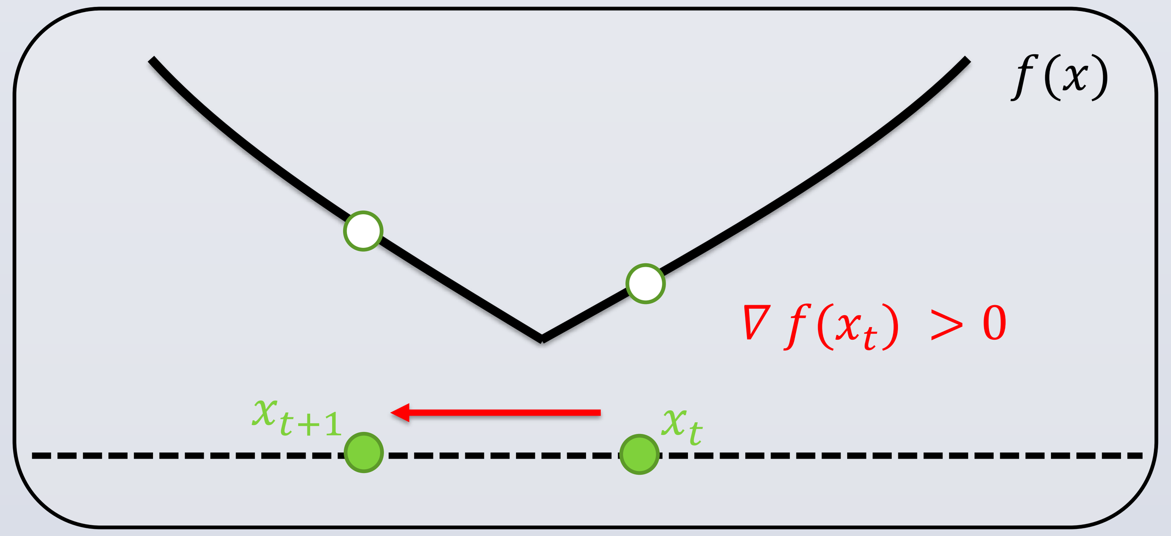
## Textbook non-smooth gradient descent

**Input:** $X \subset R^n, x_1 \in R^n, \eta_1, \eta_2, \ldots$
**For** $t = 1, \ldots, T$, **do:**
- Query the gradient oracle to obtain $g_t \in \partial f(x_t)$
- $y_{t+1} \leftarrow x_t - \eta_t g_t$
- $x_{t+1} \leftarrow \Pi_X(y_{t+1})$

**Endfor**

| Setting | Standard Convergence Rates | Optimal |
|---|---|---|
| Non-Smooth and Strongly Convex | $f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = O(\log(t)/t)$ | $O(1/t)$ |
| Non-Smooth and Lipschitz | $f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = O(1/\sqrt{t})$ | $O(1/\sqrt{t})$ |

**Because of non-monotonicity, standard results for non-smooth gradient descent require averaging.**




## Stochastic gradient descent

Useful when it is infeasible to compute a true gradient.

**Input:** $X \subset R^n, x_1 \in R^n, \eta_1, \eta_2, \ldots$
**For** $t = 1, \ldots, T$, **do:**
- Query the gradient oracle to obtain $\boxed{\hat{g}_t}$
- $y_{t+1} \leftarrow x_t - \eta_t \boxed{\hat{g}_t}$
- $x_{t+1} \leftarrow \Pi_X(y_{t+1})$

**Endfor**

**Assumptions:**
$\mathbb{E}[\hat{g}_t \mid x_1, \ldots, x_t] \in \partial f(x_t)$.
$\| \hat{g}_t \|$ is a.s. bounded.

## Prior Work: Lipschitz functions

| Strategy | Expected UB | High Prob. UB | Expected LB |
|---|---|---|---|
| Uniform Averaging | $O(1/\sqrt{t})$ [Nemirovski-Yudin `83] | $O(1/\sqrt{t})$ [Azuma] | $\Omega(1/\sqrt{t})$ [Nemirovski-Yudin `83] |
| Last Iterate | $O(\log(t)/\sqrt{t})$ [Shamir-Zhang `13] | ? [Main Question 2] | ? [Main Question 1] |

## Our Contribution: Lipschitz functions

| Strategy | Expected UB | High Prob. UB | Expected LB |
|---|---|---|---|
| Uniform Averaging | $O(1/\sqrt{t})$ [Nemirovski-Yudin `83] | $O(1/\sqrt{t})$ [Azuma] | $\Omega(1/\sqrt{t})$ [Nemirovski-Yudin `83] |
| Last Iterate | $O(\log(t)/\sqrt{t})$ [Shamir-Zhang `13] | $O(\log(t)/\sqrt{t})$ [This work] | $\Omega(\log(t)/\sqrt{t})$ [This work] |

## Prior Work: Strongly-convex functions

| Strategy | Expected UB | High Prob. UB | Expected LB |
|---|---|---|---|
| Uniform Averaging | $O(\log(t)/t)$ [Nemirovski-Yudin `83] | $O(\log(t)/t)$ [Kakade-Tewari `08] | $\Omega(\log(t)/t)$ [Rakhlin-Shamir-Sridaran `12] |
| Epoch Averaging | $O(1/t)$ [Hazan-Kale `11] | $O(\log\log t/t)$ [Hazan-Kale `11] | $\Omega(1/t)$ [Nemirovski-Yudin `83] |
| Suffix Averaging | $O(1/t)$ [Rakhlin-Shamir-Sridaran `12] | $O(\log\log t/t)$ [Rakhlin-Shamir-Sridaran `12] [Question 3] | $\Omega(1/t)$ [Nemirovski-Yudin `83] |
| Last Iterate | $O(\log(t)/t)$ [Shamir-Zhang `13] | ? [Main Question 2] | ? [Main Question 1] |

## Our Contribution: Strongly-convex functions

| Strategy | Expected UB | High Prob. UB | Expected LB |
|---|---|---|---|
| Uniform Averaging | $O(\log(t)/t)$ [Nemirovski-Yudin `83] | $O(\log(t)/t)$ [Kakade-Tewari `08] | $\Omega(\log(t)/t)$ [Rakhlin-Shamir-Sridaran `12] |
| Epoch Averaging | $O(1/t)$ [Hazan-Kale `11] | $O(\log\log t/t)$ [Hazan-Kale `11] | $\Omega(1/t)$ [Nemirovski-Yudin `83] |
| Suffix Averaging | $O(1/t)$ [Rakhlin-Shamir-Sridaran `12] | $O(1/t)$ [This work] | $\Omega(1/t)$ [Nemirovski-Yudin `83] |
| Last Iterate | $O(\log(t)/t)$ [Shamir-Zhang `13] | $O(\log(t)/t)$ [This work] | $\Omega(\log(t)/t)$ [This work] |

## The main questions

**Main Question 1: What is the expected sub-optimality of the last iterate returned by gradient descent? [Shamir '12]**

**Main Question 2: Can one obtain a high probability convergence rate for the sub-optimality of the last iterate which matches the expected rate? [Shamir '12]**

## Optimal high probability bounds

**Question 3: Is there an algorithm which achieves the optimal $O(1/t)$ rate with high probability?**

1. **Lipschitz functions:** uniform averaging achieves optimal $O(1/\sqrt{t})$ rate whp.
2. **Strongly convex functions**: various algorithms achieve $O(1/t)$ in expectation, but not whp.

## High Probability Upper Bounds:

**Lipschitz case:**

**Theorem:** Let $f : X \to \mathbb{R}$ be convex and 1-Lipschitz with $diam(X)$ bounded.
Then, for every $\delta \in (0,1)$:
$$f(x_T) - OPT \leq O\left(\frac{\log(T) \cdot \log(1/\delta)}{\sqrt{T}}\right) \text{ w.p.} \geq 1 - \delta.$$

**Strongly convex case:**

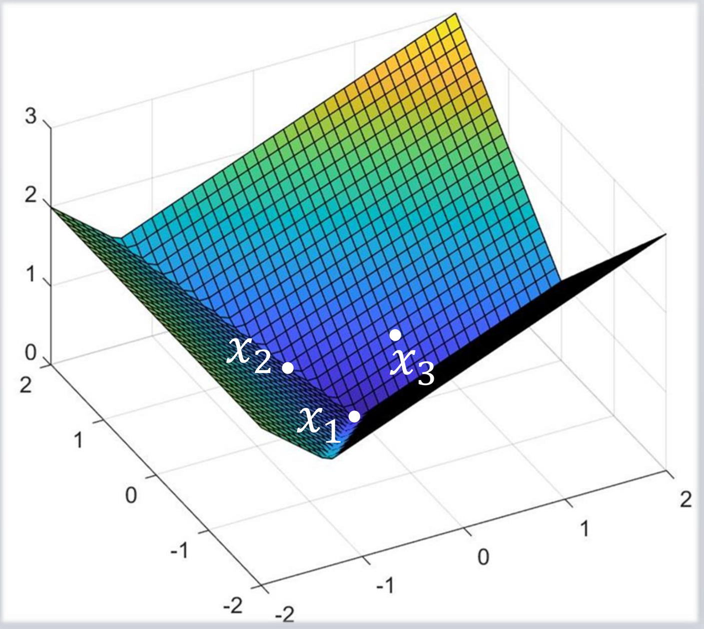
**Theorem:** Let $f : X \to \mathbb{R}$ be convex and 1-Lipschitz and 1-strongly convex. Then, for every $\delta \in (0,1)$:
$$f(x_T) - OPT \leq O\left(\frac{\log(T) \cdot \log(1/\delta)}{T}\right) \text{ w.p.} \geq 1 - \delta.$$

**The proof of this result requires a high probability bound of $O(1/T)$ on the error of the suffix average.**

## Lower Bounds

**Lipschitz case:**

**Theorem:** Fix $T \in \mathbb{N}$. $\exists$ 1-Lipschitz $f_T : B_2^T \to \mathbb{R}$ s.t. executing GD from $x_1 = 0$ with $\eta_t = c/\sqrt{t}$ yields:
$$f_T(x_T) - OPT \geq \frac{\log T}{32\sqrt{T}}. \quad \text{(Suboptimal convergence.)}$$




Instantiation of $f_T$ with $T = 3$.

**The function $f_T$:**
$$\max\left\{ x^T\begin{pmatrix} -1/2\sqrt{T} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, x^T\begin{pmatrix} 1/8T \\ -\sqrt{2}/2\sqrt{T} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, x^T\begin{pmatrix} 1/8T \\ 1/8(T-1) \\ -\sqrt{2}/2\sqrt{T} \\ \vdots \\ 0 \end{pmatrix}, \ldots, x^T\begin{pmatrix} 1/8T \\ 1/8(T-1) \\ 1/8(T-2) \\ \vdots \\ 1/8 \end{pmatrix} \right\}$$

**GD on $f_T$ produces:**
$$x_T = \Theta(1/\sqrt{T})\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

**Strongly convex case:**

**Theorem:** Fix $T \in \mathbb{N}$. $\exists$ 3-Lipschitz and 1-strongly convex function $f : B_2^T \to \mathbb{R}$ s.t. executing GD from $x_1 = 0$ with $\eta_t = c/t$ yields:
$$f(x_T) - OPT \geq \frac{\log T}{4T}. \quad \text{(Suboptimal convergence.)}$$

$f_T$ is defined similarly to the definition of $f_T$ in the Lipschitz case, with an additional regularization term to ensure strong convexity.

## The Generalized Freedman Inequality

**Theorem:** Let $d_1, d_2, \ldots$ be the increments of a martingale. Suppose $d_i^2 \leq v_i \in \mathcal{F}_{i-1}$. Let $M_n = \sum_{i=1}^n D\_i$ and $V_n = \sum_{i=1}^n v_i$. Then,
$$\Pr[M_n \geq x \text{ and } V_n \leq \boxed{\alpha M_n} + \beta] \leq \exp\left(-\frac{x}{\boxed{4\alpha} + 8\beta}\right).$$

1. Key tool in proving high probability upper bound for error of final iterate in strongly convex and non-strongly convex case, as well as for optimal high probability bound for suffix averaging.
2. Can recover Freedmans' inequality by setting $\alpha = 0$.

## High Probability Upper Bound Sketch

1. Can split analysis into analysis of final iterate for deterministic GD and the analysis of the total accumulated noise.
2. Deterministic analysis is handled by [Shamir-Zhang '13].
3. Suffices to analyze the total amount of noise accumulated after $T$ steps. Call this $Z_T$.
4. The noise, $Z_T$, is a martingale. Write: $Z_T = \sum_{i=1}^T d_i$
5. Can show whp:
$$V_T(Z_T) \leq \frac{\log^2(T)}{T} + \frac{\log(T)}{\sqrt{T}} Z_T.$$
6. Apply Generalized Freedman Inequality.

## Optimal High Prob. Strongly Convex Algorithm

**Theorem:** Let $f : X \to \mathbb{R}$ be convex and 1-Lipschitz and 1-strongly convex. Then, for every $\delta \in (0,1)$:
$$f\left(\sum_{t=T/2}^{T} x_t\right) - OPT \leq O\left(\frac{\log(1/\delta)}{T}\right) \text{ w.p.} \geq 1 - \delta.$$

First known result which obtains the optimal $O(1/T)$ rate with high probability for strongly convex functions.

## References

**[Hazan-Kale '11]** Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 69(2-3):169–192, 2007.

**[Shamir-Zhang '13]** Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. Proceedings of the 30th International Conference on Machine Learning, PMLR, 28(1):71–79, 2013.

**[Shamir '12]** Ohad Shamir. Open problem: Is averaging needed for strongly convex stochastic gradient descent? Proceedings of the 25th Annual Conference on Learning Theory, PMLR, 23:47.1–47.3, 2012.

**[Rakhlin-Shamir-Sridaran '12]** Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In Proceedings of ICML, 2012.

**[Nemirovski-Yudin '83]** A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. Wiley, 1983.

**[Kakade-Tewari '08]** Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In NIPS, pages 801–808, 2008.