Tight Analyses for Non-Smooth Stochastic Gradient Descent

Nick Harvey
(UBC)

Chris Liaw
(UBC)

Yaniv Plan
(UBC)

Sikander Randhawa
(UBC)

Gradient Descent: *basics.*

<u>Definition</u> (Gradient Descent): $x_{t+1} \leftarrow x_t - \eta_t g_t$ where $g_t \in \partial f\,(x_t)$.

Gradient Descent: *basics.*

<u>Definition</u> (Gradient Descent): $x_{t+1} \leftarrow x_t - \eta_t g_t$ where $g_t \in \partial f\,(x_t)$.

There are still basic questions that have yet to be answered.

Gradient Descent: *basics.*

Definition (Gradient Descent): $x_{t+1} \leftarrow x_t - \eta_t g_t$ where $g_t \in \partial f\,(x_t)$.

There are still basic questions that have yet to be answered.

Assumption on f	Standard Convergence Rates
Smooth and **Strongly Convex**	$f(x_t) - OPT = \mathcal{O}(\exp(-t)\,)$
Smooth	$f(x_t) - OPT = \mathcal{O}(1/t)$
Non-Smooth and **Strongly Convex**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(\log(t)\,/t)$
Non-Smooth and **Lipschitz**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(1/\sqrt{t})$

Gradient Descent: *basics.*

Definition (Gradient Descent): $x_{t+1} \leftarrow x_t - \eta_t g_t$ where $g_t \in \partial f\,(x_t)$.

There are still basic questions that have yet to be answered.

Assumption on f	Standard Convergence Rates
Smooth and **Strongly Convex**	$f(x_t) - OPT = \mathcal{O}(\exp(-t)\,)$
Smooth	$f(x_t) - OPT = \mathcal{O}(1/t)$
Non-Smooth and **Strongly Convex**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(\log(t)\,/t)$
Non-Smooth and **Lipschitz**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(1/\sqrt{t})$

We focus on non-smooth functions.

Gradient Descent: *basics.*

Definition (Gradient Descent): $x_{t+1} \leftarrow x_t - \eta_t g_t$ where $g_t \in \partial f\,(x_t)$.

There are still basic questions that have yet to be answered.

Assumption on f	Standard Convergence Rates
Smooth and **Strongly Convex**	$f(x_t) - OPT = \mathcal{O}(\exp(-t)\,)$
Smooth	$f(x_t) - OPT = \mathcal{O}(1/t)$
Non-Smooth and **Strongly Convex**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(\log(t)\,/t)$
Non-Smooth and **Lipschitz**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(1/\sqrt{t})$

We focus on non-smooth functions.

Standard results in non-smooth setting require averaging of iterates.

Gradient Descent: *basics.*

Definition (Gradient Descent): $x_{t+1} \leftarrow x_t - \eta_t g_t$ where $g_t \in \partial f\,(x_t)$.

There are still basic questions that have yet to be answered.

Assumption on f	Standard Convergence Rates
Smooth and **Strongly Convex**	$f(x_t) - OPT = \mathcal{O}(\exp(-t))$
Smooth	$f(x_t) - OPT = \mathcal{O}(1/t)$
Non-Smooth and **Strongly Convex**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(\log(t)/t)$
Non-Smooth and **Lipschitz**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(1/\sqrt{t})$

We focus on non-smooth functions.

Standard results in non-smooth setting require averaging of iterates.

Averaging is optimal for non-smooth Lipschitz functions.

Gradient Descent: *basics.*

Definition (Gradient Descent): $x_{t+1} \leftarrow x_t - \eta_t g_t$ where $g_t \in \partial f\,(x_t)$.

There are still basic questions that have yet to be answered.

Assumption on f	Standard Convergence Rates
Smooth and **Strongly Convex**	$f(x_t) - OPT = \mathcal{O}(\exp(-t))$
Smooth	$f(x_t) - OPT = \mathcal{O}(1/t)$
Non-Smooth and **Strongly Convex**	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(\log(t)/t)$
Non-Smooth and **Lipschitz** (Needs $\eta_t = {}^1/_{\sqrt{t}}$)	$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - OPT = \mathcal{O}(1/\sqrt{t})$

We focus on non-smooth functions.

Standard results in non-smooth setting require averaging of iterates.

Averaging is optimal for non-smooth Lipschitz functions.

Non-smooth functions: *why average?*

Smoothness $\Rightarrow$ iterates are monotonically decreasing in value.

If f is smooth, one can choose η_t such that $f(x_{t+1}) \le f(x_t)$ for every t.

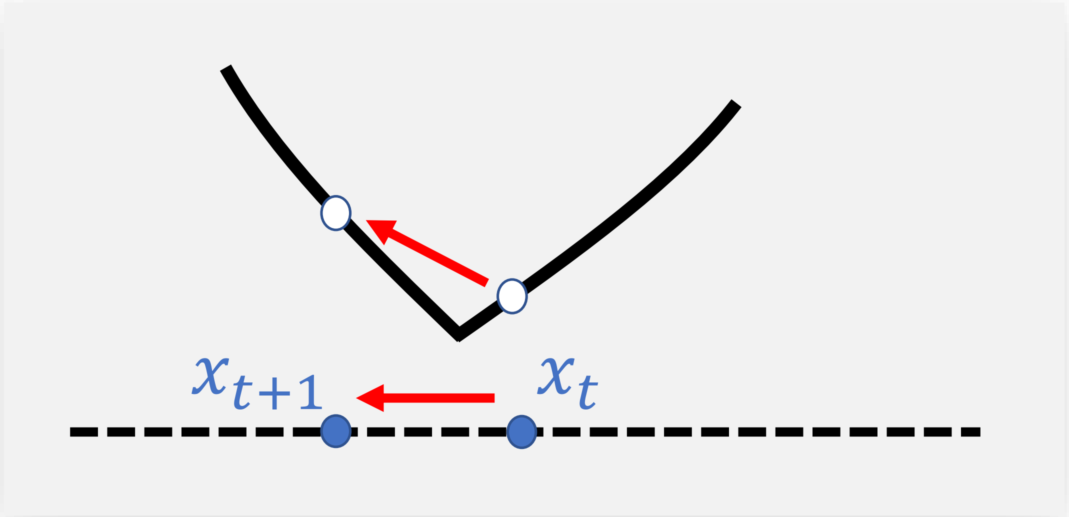
Non-smooth functions: *why average?*

Smoothness ⇒ iterates are monotonically decreasing in value.

If f is smooth, one can choose η_t such that $f(x_{t+1}) \leq f(x_t)$ for every t.

Non-smooth functions can change direction rapidly.

It is possible that $f(x_{t+1}) > f(x_t)$.



Non-smooth functions: *why average?*

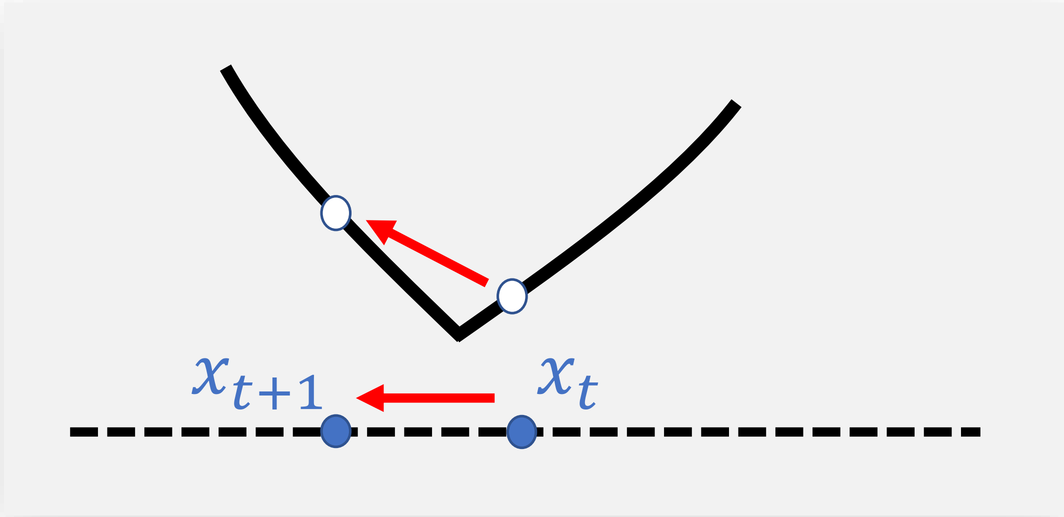
Smoothness ⇒ iterates are monotonically decreasing in value.

If f is smooth, one can choose η_t such that $f(x_{t+1}) \leq f(x_t)$ for every t.

Non-smooth functions can change direction rapidly.

It is possible that $f(x_{t+1}) > f(x_t)$.

Typically use averaging to get around this issue.



A basic question: *what about the individual iterates?*

Shamir's Open Questions [COLT '12]:

A basic question: *what about the individual iterates?*

Shamir's Open Questions [COLT '12]:

($50) "What is the suboptimality of the last iterate returned by GD?"

A basic question: *what about the individual iterates?*

Shamir's Open Questions [COLT '12]:

($50) "What is the suboptimality of the last iterate returned by GD?"

($20) "will be awarded for a tight *high probability* bound on the suboptimality of [the last iterate]."

A basic question: *what about the individual iterates?*

Shamir's Open Questions [COLT '12]:

($50) "What is the suboptimality of the last iterate returned by GD?"

($20) "will be awarded for a tight *high probability* bound on the suboptimality of [the last iterate]."

We answer both of these questions.

For strongly-convex and Lipschitz functions.

For Lipschitz functions.

Setting for today: Lipschitz and Non-Smooth functions

Main Question 1: *convergence of individual iterates?*

[Nesterov-Shikhman '15]: give algorithm where iterates' values converge to OPT at the optimal rate.

Main Question 1: *convergence of individual iterates?*

[Nesterov-Shikhman ’15]: give algorithm where iterates’ values converge to OPT at the optimal rate.

Motivates another question:

- do the individual iterates’ values of GD converge to OPT? If so, at what rate?

Main Question 1: *convergence of individual iterates?*

[Nesterov-Shikhman '15]: give algorithm where iterates' values converge to OPT at the optimal rate.

Motivates another question:

- do the individual iterates' values of GD converge to OPT? If so, at what rate?
  - Best upper bound [Shamir-Zhang '13]: $\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$ (sub-optimal).

Main Question 1: *convergence of individual iterates?*

[Nesterov-Shikhman ’15]: give algorithm where iterates’ values converge to OPT at the optimal rate.

Motivates another question:

- do the individual iterates’ values of GD converge to OPT? If so, at what rate?
  - Best upper bound [Shamir-Zhang ’13]: $\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$ (sub-optimal).

The question we address:

- Is $\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$ the right rate of convergence for the iterates of GD? [Shamir ’12]

Main Result 1: *lower bound, Lipschitz case.*

Theorem: Fix $T \in \mathbb{N}$.

Main Result 1: *lower bound, Lipschitz case.*

Theorem: Fix $T \in \mathbb{N}$. Then $\exists$ 1-Lipschitz $f : B_2^T \to \mathbb{R}$ s.t. $OPT = f(0) = 0$ and executing GD from $x_1 = 0$ with $\eta_t = c/\sqrt{t}$ yields:

Remark: It is possible to make the function independent of T by working with a function from $\ell_2 \to \mathbb{R}$.

Main Result 1: *lower bound, Lipschitz case.*

Theorem: Fix $T \in \mathbb{N}$. Then $\exists$ 1-Lipschitz $f : B_2^T \to \mathbb{R}$ s.t. $OPT = f(0) = 0$ and executing GD from $x_1 = 0$ with $\eta_t = c/\sqrt{t}$ yields:

$$f(x_{t+1}) \geq f(x_t) + \frac{1}{32\sqrt{T}(T-t)} \text{ for } t < T.$$

(Monotonic **increase**.)

Remark: It is possible to make the function independent of T by working with a function from $\ell_2 \to \mathbb{R}$.

Main Result 1: *lower bound, Lipschitz case.*

Theorem: Fix $T \in \mathbb{N}$. Then $\exists$ 1-Lipschitz $f : B_2^T \to \mathbb{R}$ s.t. $OPT = f(0) = 0$ and executing GD from $x_1 = 0$ with $\eta_t = c/\sqrt{t}$ yields:

$$f(x_{t+1}) \geq f(x_t) + \frac{1}{32\sqrt{T}(T-t)} \text{ for } t < T. \qquad \text{(Monotonic **increase**.)}$$

Thus, $f(x_T) - OPT \geq \frac{\log T}{32\sqrt{T}}$. (Suboptimal convergence.)

Remark: It is possible to make the function independent of T by working with a function from $\ell_2 \to \mathbb{R}$.

Main Result 1: *lower bound, Lipschitz case.*

Theorem: Fix $T \in \mathbb{N}$. Then $\exists$ 1-Lipschitz $f : B_2^T \to \mathbb{R}$ s.t. $OPT = f(0) = 0$ and executing GD from $x_1 = 0$ with $\eta_t = c/\sqrt{t}$ yields:

$$f(x_{t+1}) \geq f(x_t) + \frac{1}{32\sqrt{T}(T-t)} \text{ for } t < T. \qquad \text{(Monotonic **increase**.)}$$

Thus, $f(x_T) - OPT \geq \frac{\log T}{32\sqrt{T}}$. (Suboptimal convergence.)

1. We use $\eta_t = c/\sqrt{t}$ because it is only choice of step size that gives the optimal $O(1/\sqrt{t})$ convergence rate.

Remark: It is possible to make the function independent of T by working with a function from $\ell_2 \to \mathbb{R}$.

Main Result 1: *lower bound, Lipschitz case.*

Theorem: Fix $T \in \mathbb{N}$. Then $\exists$ 1-Lipschitz $f : B_2^T \to \mathbb{R}$ s.t. $OPT = f(0) = 0$ and executing GD from $x_1 = 0$ with $\eta_t = c/\sqrt{t}$ yields:

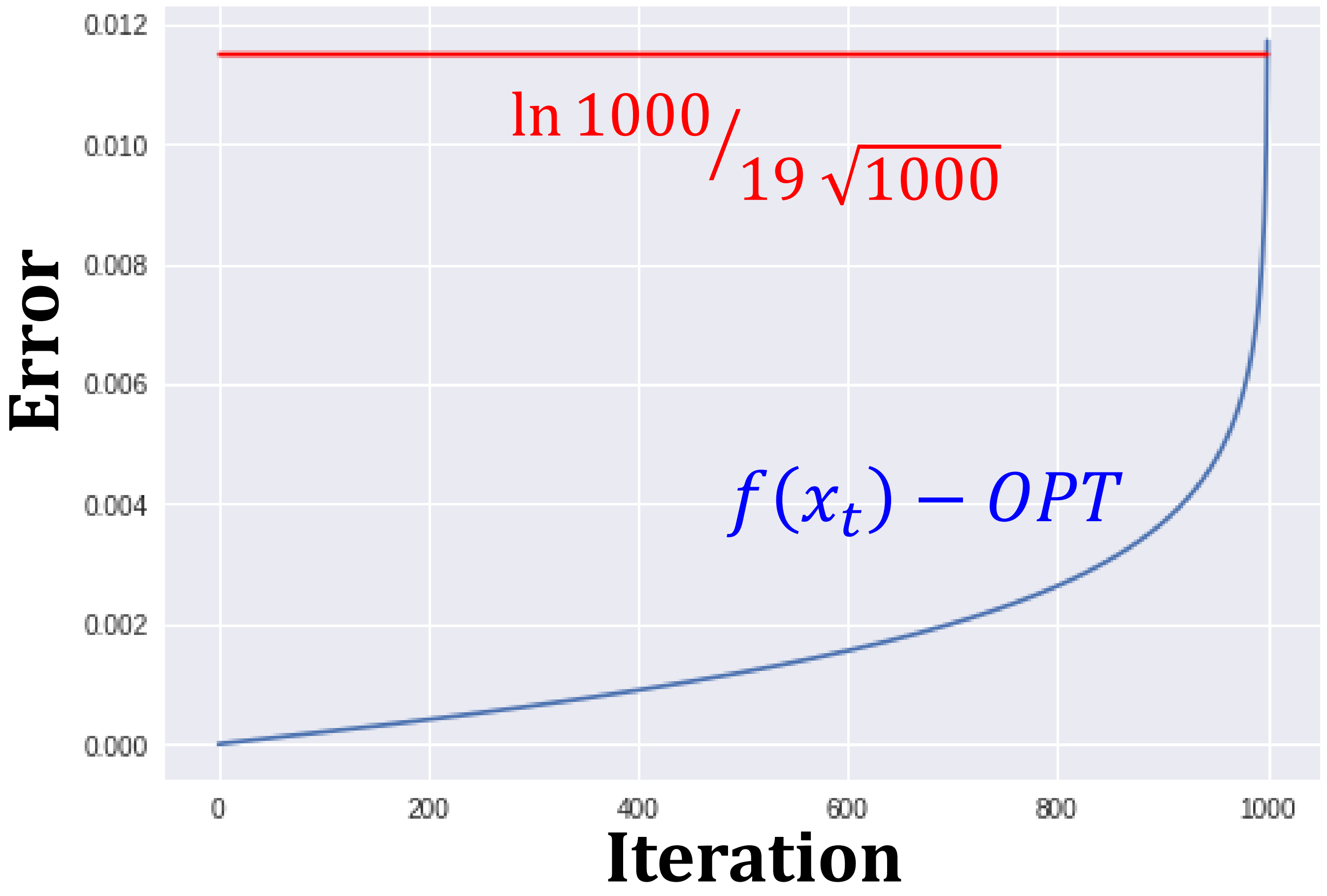
$$f(x_{t+1}) \geq f(x_t) + \frac{1}{32\sqrt{T}(T-t)} \text{ for } t < T. \qquad \text{(Monotonic **increase**.)}$$

Thus, $f(x_T) - OPT \geq \frac{\log T}{32\sqrt{T}}$. (Suboptimal convergence.)

1. We use $\eta_t = c/\sqrt{t}$ because it is only choice of step size that gives the optimal $O\left(1/\sqrt{t}\right)$ convergence rate.
2. If T is known ahead of time, other step sizes can be used [Jain-Nagaraj-Netrapalli '19].

Remark: It is possible to make the function independent of T by working with a function from $\ell_2 \to \mathbb{R}$.

Main Result 1: *lower bound, Lipschitz case.*



Fix $T = 1000$.

Python simulation of GD for this f.

T consecutive iterations of **increase**!

At step T, error is $\Omega\left(\log(T)/\sqrt{T}\right)$.

Stochastic Gradient Descent: *definition.*

Input: $X \subset R^n, x_1 \in R^n, \eta_1, \eta_2, \dots$

For $t = 1, \dots, T,$ **do:**

- Query the gradient oracle to obtain $\hat{\boldsymbol{g}}_{\mathbf{t}}$
- $y_{t+1} \leftarrow x_t - \eta_t \hat{g}_{\mathrm{t}}$
- $x_{t+1} \leftarrow \Pi_X \left(y_{t+1} \right)$

Endfor

Stochastic Gradient Descent: *definition.*

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 $\hat{g}_t$
- $y_{t+1} \leftarrow x_t - \eta_t \hat{g}_t$
- $x_{t+1} \leftarrow \Pi_X \left(y_{t+1} \right)$

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.

Remark: Can relax bounded noise assumption to a sub-gaussian noise assumption.

Stochastic Gradient Descent: *definition.*

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 $\hat{g}_t$
- $y_{t+1} \leftarrow x_t - \eta_t \hat{g}_t$
- $x_{t+1} \leftarrow \Pi_X (y_{t+1})$

Endfor

$f(x_T) - OPT$ is now a random quantity.

Assumptions:

$\mathbb{E}[\hat{g}_t \mid x_1, \ldots, x_t] \in \partial f(x_t).$

$\| \hat{g}_t \|$ is a.s. bounded.

Remark: Can relax bounded noise assumption to a sub-gaussian noise assumption.

Stochastic Gradient Descent: *definition.*

Input: $X \subset R^n, x_1 \in R^n, \eta_1, \eta_2, \dots$
For $t = 1, \dots, T$, **do:**

- Query the gradient oracle to obtain $\hat{g}_t$
- $y_{t+1} \leftarrow x_t - \eta_t \hat{g}_t$
- $x_{t+1} \leftarrow \Pi_X (y_{t+1})$

Endfor

$f(x_T) - OPT$ is now a random quantity.

[Shamir-Zhang '12]: $\mathbb{E}[f(x_T) - OPT] \le \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$

Assumptions:

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

Remark: Can relax bounded noise assumption to a sub-gaussian noise assumption.

Stochastic Gradient Descent: *definition.*

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 $\hat{\boldsymbol{g}}_{\mathbf{t}}$
- $y_{t+1} \leftarrow x_t - \eta_t \hat{g}_t$
- $x_{t+1} \leftarrow \Pi_X\left(y_{t+1} \right)$

Endfor

$f(x_T) - OPT$ is now a random quantity.

[Shamir-Zhang '12]: $\mathbb{E}[f(x_T) - OPT] \leq \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$

Our lower bound shows this is tight, *in expectation.*

Assumptions:

$\mathbb{E}[\hat{g}_t \mid x_1, \ldots, x_t] \in \partial f(x_t)$.

$\| \hat{g}_t \|$ is a.s. bounded.

Remark: Can relax bounded noise assumption to a sub-gaussian noise assumption.

Main Question 2: *high probability bounds?*

Shamir's Open Question [COLT '12]:

"Another issue is obtaining a bound which holds with probability $\mathbf{1-\delta}$ and **logarithmic dependence on $\mathbf{1/\delta}$**. An extra **$20** will be awarded for proving a tight bound on the suboptimality of **[the last iterate]** which holds in **high probability.**"

Main Result 2: *high probability upper bound, Lipschitz case.*

<u>Theorem:</u> Let $f : X \to \mathbb{R}$ be convex and 1-Lipschitz with $diam(X)$ bounded.

Main Result 2: *high probability upper bound, Lipschitz case.*

Theorem: Let $f : X \to \mathbb{R}$ be convex and 1-Lipschitz with $diam(X)$ bounded. Let the stochastic gradient oracle return $\hat{g}_{\mathrm{t}}$ such that

$$\mathbb{E}[\hat{g}_t \mid x_1, \dots, x_t] \in \partial f(x_t) \quad \text{and} \quad \|\hat{g}_t\| \text{ is a.s. bounded.}$$

Main Result 2: *high probability upper bound, Lipschitz case.*

Theorem: Let $f : X \to \mathbb{R}$ be convex and 1-Lipschitz with $diam(X)$ bounded. Let the stochastic gradient oracle return $\hat{g}_{\mathrm{t}}$ such that

$$\mathbb{E}[\hat{g}_t \mid x_1, \dots, x_t] \in \partial f(x_t) \quad \text{and} \quad \|\hat{g}_t\| \text{ is a.s. bounded.}$$

Then, for every $\delta \in (0,1)$:

$$f(x_T) - OPT \leq O\left(\frac{\log(T) \cdot \log(1/\delta)}{\sqrt{T}}\right) \text{ w.p.} \geq 1 - \delta.$$

Remark: It is not clear whether the dependence of the upper bound on $\log\left({}^{1}/_{\delta}\right)$ is completely tight.

Main Result 2: *high probability upper bound, Lipschitz case.*

Theorem: Let $f : X \to \mathbb{R}$ be convex and 1-Lipschitz with $diam(X)$ bounded.
Let the stochastic gradient oracle return $\hat{g}_t$ such that

$$\mathbb{E}[\hat{g}_t \mid x_1, \dots, x_t] \in \partial f(x_t) \quad \text{and} \quad \|\hat{g}_t\| \text{ is a.s. bounded.}$$

Then, for every $\delta \in (0,1)$:

$$f(x_T) - OPT \le O\left(\frac{\log(T) \cdot \log(1/\delta)}{\sqrt{T}}\right) \text{ w.p. } \ge 1 - \delta.$$

Uses a generalization of Freedman's inequality to handle a special class of martingales.

Remark: It is not clear whether the dependence of the upper bound on $\log\left({}^{1}/_{\delta}\right)$ is completely tight.

The main tool: *a generalization of Freedman's inequality.*

<u>Generalized Freedman Inequality:</u> a martingale concentration inequality useful when the **variance** is bounded by the **martingale** itself.

The Generalized Freedman Inequality is a key tool in several high probability bounds (last iterate, suffix averaging, …).

The main tool: *a generalization of Freedman's inequality.*

Generalized Freedman Inequality: a martingale concentration inequality useful when the **variance** is bounded by the **martingale** itself.

Theorem [HLPR2018]**:** Let $D_1, D_2, \ldots$ be the increments of a martingale.

The Generalized Freedman Inequality is a key tool in several high probability bounds (last iterate, suffix averaging, …).

The main tool: *a generalization of Freedman's inequality.*

Generalized Freedman Inequality: a martingale concentration inequality useful when the **variance** is bounded by the **martingale** itself.

Theorem [HLPR2018]**:** Let $D_1, D_2, \ldots$ be the increments of a martingale. Suppose, whp:

Total conditional variance $\quad \sum_{i=1}^{T} D_i^2 \leq \left(\alpha^2 + \alpha \sum_{i=1}^{T} D_i\right).$ $\quad$ The martingale

The Generalized Freedman Inequality is a key tool in several high probability bounds (last iterate, suffix averaging, ...).

The main tool: *a generalization of Freedman's inequality.*

Generalized Freedman Inequality: a martingale concentration inequality useful when the **variance** is bounded by the **martingale** itself.

Theorem [HLPR2018]: Let $D_1, D_2, \ldots$ be the increments of a martingale. Suppose, whp:

Total conditional variance

$$\sum_{i=1}^{T} D_i^2 \leq \left(\alpha^2 + \alpha \sum_{i=1}^{T} D_i\right).$$

The martingale

Then, with high probability

The martingale

$$\sum_{i=1}^{T} D_i \leq \alpha.$$

The Generalized Freedman Inequality is a key tool in several high probability bounds (last iterate, suffix averaging, ...).

Lipschitz Functions

Return Scheme	**Deterministic & Expected UB**	**High Probability UB**	**Deterministic 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]	**???**	**???**

Strongly Convex & Lipschitz Functions

Return Scheme	**Deterministic & Expected UB**	**High Probability UB**	**Deterministic LB**
Uniform Averaging	$O(\log(T)/T)$ [Hazan-Agarwal-Kale '07]	$O(\log(T)/T)$ [Kakade-Tewari '08]	$\Omega(\log(T)/T)$ (expectation) [Rakhlin-Shamir-Sridharan '12]
Epoch-based 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-Sridharan '12]	$O(\log(\log T)/T)$ [Rakhlin-Shamir-Sridharan '12]	$\Omega(1/T)$ [Nemirovski–Yudin '83]
Last Iterate	$O(\log(T)/T)$ [Shamir-Zhang '13]	**???**	**???**

Lipschitz Functions

Return Scheme	Deterministic & Expected UB	High Probability UB	Deterministic LB
Uniform Averaging	$O(1/\sqrt{T})$ [Nemirovski-Yudin '83] **Tight**	$O(1/\sqrt{T})$ [Azuma] **Tight**	$\Omega(1/\sqrt{T})$ [Nemirovski-Yudin '83]
Last Iterate	$O(\log(T)/\sqrt{T})$ [Shamir-Zhang '13]	**???**	**???**

Strongly Convex & Lipschitz Functions

Return Scheme	Deterministic & Expected UB	High Probability UB	Deterministic LB
Uniform Averaging	$O(\log(T)/T)$ [Hazan-Agarwal-Kale '07] **Tight**	$O(\log(T)/T)$ [Kakade-Tewari '08] **Tight**	$\Omega(\log(\mathrm{T})/T)$ (expectation) [Rakhlin-Shamir-Sridharan '12]
Epoch-based Averaging	$O(1/T)$ [Hazan-Kale '11] **Tight**	$O(\log(\log T)/T)$ [Hazan-Kale '11] **Tight**	$\Omega(1/T)$ [Nemirovski–Yudin '83]
Suffix Averaging	$O(1/T)$ [Rakhlin-Shamir-Sridharan '12] **Tight**	$O(\log(\log \mathrm{T})/T)$ [Rakhlin-Shamir-Sridharan '12] **Tight**	$\Omega(1/T)$ [Nemirovski–Yudin '83]
Last Iterate	$O(\log(T)/T)$ [Shamir-Zhang '13]	**???**	**???**

Lipschitz Functions

* dependence on $\log(1/\delta)$ is not completely tight

Return Scheme	Deterministic & Expected UB	High Probability UB	Deterministic LB
Uniform Averaging	$O(1/\sqrt{T})$ [Nemirovski-Yudin '83] **Tight**	$O(1/\sqrt{T})$ [Azuma] **Tight**	$\Omega(1/\sqrt{T})$ [Nemirovski-Yudin '83]
Last Iterate	$O(\log(T)/\sqrt{T})$ [Shamir-Zhang '13] **Tight***	$O(\log(T)/\sqrt{T})$ **[This work]** **Tight***	$\Omega(\log(T)/\sqrt{T})$ **[This work]**

Strongly Convex & Lipschitz Functions

Return Scheme	Deterministic & Expected UB	High Probability UB	Deterministic LB
Uniform Averaging	$O(\log(T)/T)$ [Hazan-Agarwal-Kale '07] **Tight**	$O(\log(T)/T)$ [Kakade-Tewari '08] **Tight**	$\Omega(\log(T)/T)$ (expectation) [Rakhlin-Shamir-Sridharan '12]
Epoch-based Averaging	$O(1/T)$ [Hazan-Kale '11] **Tight**	$O(\log(\log T)/T)$ [Hazan-Kale '11] **Tight**	$\Omega(1/T)$ [Nemirovski–Yudin '83]
Suffix Averaging	$O(1/T)$ [Rakhlin-Shamir-Sridharan '12] **Tight**	$O(1/T)$ **[This work]** **Tight**	$\Omega(1/T)$ [Nemirovski–Yudin '83]
Last Iterate	$O(\log(T)/T)$ [Shamir-Zhang '13] **Tight***	$O(\log(T)/T)$ **[This work]** **Tight***	$\Omega(\log(T)/T)$ **[This work]**

Thank you!
Questions?

Come see us at poster 168!