LAMDA
Learning And Mining from DatA

Parameter-Free Stochastic Optimization

From **Yair Carmon** and **Oliver Hinder**

Presented by Yu-Hu Yan

2024.09.27

Making SGD Parameter-Free, COLT 2022

DOG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule, ICML 2023

Stochastic Optimization

Learning And Mining from DatA

Problem Setup

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^d \end{aligned}$$

In this talk, we consider: *(i)* ***unbounded*** domain $\mathbb{R}^d$;

(ii) ***convex G-Lipschitz*** function $f(\cdot)$.

Stochastic noise formulation:

$$\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla f(\mathbf{x}),$$
$$\mathbb{E}\left[\|\nabla f(\mathbf{x}) - \mathbf{g}(\mathbf{x})\|^2 \mid \mathbf{x}\right] \leq \sigma^2$$

Parameter-Freeness

*Achieving **parameter-freeness** is important in both **online** and **offline** optimization.*

Common parameters in ***stochastic*** optimization:

- G-Lipschitzness: $f(x) - f(y) \leq G\|x - y\|$
- L-smoothness: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
- initial optimality: $d_0 \triangleq \|x_0 - x^\star\|$
- stochastic noise: $\|g(x) - \nabla f(x)\| \leq \sigma$

Benchmark

LAMDA
Learning And Mining from DatA

Benchmark method

SGD: $x_{t+1} = x_t - \eta_t g_t$ *as simple as* ***OGD***

Although SGD is very simple, its analysis in stochastic optimization is far from complete.

Benchmark rate for ***convex Lipschitz*** case

- G-Lipschitzness: $f(x) - f(y) \leq G\|x - y\|$
- L-smoothness: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
- initial optimality: $d_0 \triangleq \|x_0 - x^\star\|$
- stochastic noise: $\|g(x) - \nabla f(x)\| \leq \sigma$

Tuned SGD achieves the ***optimal*** rate of

$$f(\bar{x}) - f(x_\star) \leq \mathcal{O}\left(\frac{d_0(G+\sigma)}{\sqrt{T}}\right) \quad (\bar{x} \text{ being some statistic of } \{x_t\}_{t=1}^T)$$

Can we obtain the optimal rate ***without these parameters****?*

Making SGD Parameter-Free

Oliver Hinder

University of Pittsburgh

Yair Carmon

Tel Aviv University

Ideal Step Size

LAMDA
Learning And Mining from DatA

- **With** parameters, what is the ***ideal*** step size?

To see this, we start with a preliminary analysis of SGD.

If we run SGD for T rounds and submit $\bar{x} \triangleq \frac{1}{T}\sum_{t=1}^{T} x_t$.

$$f(\bar{x}) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(x_t), x_t - x_\star\rangle$$

$$= \frac{1}{T}\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle + \frac{1}{T}\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star\rangle \qquad (\Delta_t \triangleq \nabla f(x_t) - g_t)$$

algorithm-related ***algorithm-irrelated*** due to ***unknown*** $\nabla f(x_t)$

Next step: analyze $\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle$ for ideal step size.

Ideal Step Size

LAMDA
Learning And Mining from DatA

- **With** parameters, what is the ***ideal*** step size?

To see this, we start with a preliminary analysis of SGD.

If we run SGD for T rounds and submit $\bar{x} \triangleq \frac{1}{T}\sum_{t=1}^{T} x_t$.

$$f(\bar{x}) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(x_t), x_t - x_\star\rangle$$

$$= \frac{1}{T}\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle + \frac{1}{T}\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star\rangle \qquad (\Delta_t \triangleq \nabla f(x_t) - g_t)$$

algorithm-related ***algorithm-irrelated*** due to ***unknown*** $\nabla f(x_t)$

Simplification: consider ***noiseless*** case of $g_t = \nabla f(x_t)$.

Ideal Step Size

LAMDA
Learning And Mining from DatA

- **With** parameters, what is the ***ideal*** step size?

Next step: analyze $\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle$. $\Rightarrow$ similar analysis to OGD

For simplicity, we denote by $d_t \triangleq \|x_t - x_\star\|$ and $G_t \triangleq \sum_{i\le t}\|g_i\|^2$.

$$d_{t+1}^2 = \|x_{t+1} - x_\star\|^2 = \|x_t - \eta g_t - x_\star\|^2 = d_t^2 - 2\eta\langle g_t, x_t - x_\star\rangle + \eta^2\|g_t\|^2 \quad \textit{(consider constant step size)}$$

$$\Longrightarrow \sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle \approx \frac{d_0^2 - d_T^2}{\eta} + \eta G_T \le \frac{d_0^2}{\eta} + \eta G_T \quad \Longrightarrow \quad \textbf{ideal step size: } \eta_\star \approx \frac{d_0}{\sqrt{G_T}}$$

Issue: to achieve the optimal rate $\mathcal{O}(d_0 G/\sqrt{T})$, the step size η must use initial optimality $d_0 \triangleq \|x_0 - x_\star\|$, which is ***impossible***.

Proxy for Ideal Step Size

LAMDA
Learning And Mining from DatA

- What kind of step size is implementable?

Back to $\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle \approx \frac{d_0^2 - d_T^2}{\eta} + \eta G_T$

Key insight: the negative term of $-d_T^2$ can be useful.

For simplicity, we denote by $r_t \triangleq \|x_t - x_0\|$ and $\bar{r}_t \triangleq \max_{i\le t}\|x_i - x_0\|$.

$$\frac{d_0^2 - d_T^2}{\eta} + \eta G_T = \frac{(d_0 + d_T)(d_0 - d_T)}{\eta} + \eta G_T$$

Consider two cases:

- Case $d_T \le d_0$: then $d_0 + d_T \le 2d_0$ and $d_0 - d_T = \|x_0 - x_\star\| - \|x_T - x_\star\| \le \|x_T - x_0\| = r_T \le \bar{r}_T$.
- Case $d_T > d_0$: then $d_0^2 - d_T^2 < 0 \le 2d_0\bar{r}_T$.

$\Rightarrow$ $d_0^2 - d_T^2 \lesssim d_0\bar{r}_T$. *What is different?*

Proxy for Ideal Step Size

Learning And Mining from DatA

- What kind of step size is implementable?

$$d_0^2 - d_T^2 \lesssim d_0 \bar{r}_T \implies \sum_{t=1}^{T} \langle g_t, x_t - x_\star \rangle \approx \frac{d_0^2 - d_T^2}{\eta} + \eta G_T \leq \frac{d_0 \bar{r}_T}{\eta} + \eta G_T$$

$\Rightarrow$ We can choose a ***proxy*** for the ideal step size: $\eta_\dagger \approx \dfrac{\bar{r}_T}{\sqrt{\alpha G_T}}$ $(\bar{r}_T \triangleq \max_{t \leq T} \|x_t - x_0\|)$ (α only for scalability)

Since x_0 is known, the quantity $\{\bar{r}_t\}_{t \in [T]}$ is ***trackable***.

$\Rightarrow$ $\displaystyle\sum_{t=1}^{T} \langle g_t, x_t - x_\star \rangle \lesssim \frac{d_0 \bar{r}_T}{\eta} + \eta G_T \lesssim (d_0 + \bar{r}_T)\sqrt{G_T}$ very close to the optimal rate $\mathcal{O}(d_0 G/\sqrt{T})$.

Next step: connect $\bar{r}_T \triangleq \max_{t \leq T} \|x_t - x_0\|$ to $d_0 \triangleq \|x_0 - x_\star\|$ with $\eta = \eta_\dagger$.

Proxy for Ideal Step Size

LAMDA
Learning And Mining from DatA

Proof

Next step: relate $\bar{r}_T \triangleq \max_{t \le T} \|x_t - x_0\|$ to $d_0 \triangleq \|x_0 - x_\star\|$ with $\eta = \eta_\dagger$.

From standard analysis: $0 \le \sum_{t=1}^{T} f(x_t) - \sum_{t=1}^{T} f(x_\star) \le \sum_{t=1}^{T} \langle g_t, x_t - x_\star \rangle \le \frac{d_0^2 - d_T^2}{\eta} + \eta G_T$

$\Rightarrow$ $d_T^2 \le d_0^2 + \eta^2 G_T$ $\Rightarrow$ Taking $\eta = \eta_\dagger = \frac{\bar{r}_T}{\sqrt{\alpha G_T}}$ leads to $d_T^2 \le d_0^2 + \frac{1}{\alpha}\bar{r}_T^2$.

$\bar{r}_T \triangleq \max_{t \le T} \|x_t - x_0\| \le \max_{t \le T} \|x_t - x_\star\| + \|x_0 - x_\star\| = \bar{d}_T + d_0$ $(\bar{d}_T \triangleq \max_{t \le T} \|x_t - x_\star\|)$

$\Rightarrow$ $\bar{d}_T^2 \le d_0^2 + \frac{1}{\alpha}(\bar{d}_T + d_0)^2$ $\Rightarrow$ $\bar{d}_T \le \frac{\alpha + 1}{\alpha - 1} \cdot d_0$ & $\bar{r}_T \le \frac{2\alpha}{\alpha - 1} \cdot d_0$

(taking max on both sides)

One More Thing

- Is this step size implementable? $\eta_\dagger \approx \frac{\bar{r}_T}{\sqrt{\alpha G_T}}$

  No! Because $\bar{r}_T \triangleq \max_{t \le T} \|x_t - x_0\|$ and $G_T \triangleq \sum_{t \le T} \|g_t\|^2$ are only available ***after*** the algorithm ends.

- An observation

  Consider ***noiseless*** case (i.e., $g_t = \nabla f(x_t)$), given step size η, the decision x_t is a function of η, marked as $x_t(\eta)$. As a result, $\bar{r}_T(\eta)$ and $G_T(\eta)$ are both functions of η.

  fixed-point problem: $\eta = \frac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta)}}$ ⇨ **Solution:** ***bisection*** method

Bisection

Common bisection

Given initial bound $[a^{(1)}, b^{(1)}]$, w.l.o.g., choosing $a^{(2)} = \frac{a^{(1)}+b^{(1)}}{2}$ and $b^{(2)} = b^{(1)}$,

$$b^{(2)} - a^{(2)} = b^{(1)} - \frac{a^{(1)} + b^{(1)}}{2} = 1/2 \cdot (b^{(1)} - a^{(1)}) \quad \Rightarrow \text{Each bisection halves } (b^{(i)} - a^{(i)}).$$

With $(b^{(1)} - a^{(1)})$, it runs $\log T$ times to make it $1/T$ times smaller.

This work's bisection

Given initial bound $[a^{(1)}, b^{(1)}]$, w.l.o.g., choosing $a^{(2)} = \sqrt{a^{(1)}b^{(1)}}$ and $b^{(2)} = b^{(1)}$,

$$\log \frac{b^{(2)}}{a^{(2)}} = \log \frac{b^{(1)}}{\sqrt{a^{(1)}b^{(1)}}} = \log \sqrt{\frac{b^{(1)}}{a^{(1)}}} = \frac{1}{2} \cdot \log \frac{b^{(1)}}{a^{(1)}} \quad \Rightarrow \text{Each bisection halves } \log \frac{b^{(i)}}{a^{(i)}}.$$

With $\frac{b^{(1)}}{a^{(1)}}$, it runs $\log \log T$ times to make it $1/T$ times smaller.

Bisection

- This work's bisection

fixed-point problem: $\eta = h(\eta) \triangleq \dfrac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta)}}$

If we can ensure that there exists an interval $[\eta_-, \eta_+]$ such that $\eta_- \leq h(\eta_-)$ and $\eta_+ \geq h(\eta_+)$, the bisection method can find $\eta_\dagger$ with $\log\log \frac{\eta_+}{\eta_-}$ steps.

Next step: how to find a valid interval $[\eta_-, \eta_+]$ with $\eta_- \leq h(\eta_-)$ and $\eta_+ \geq h(\eta_+)$.

Solution: start with $\eta_- = \eta_\epsilon$ with some small enough user-defined η_ϵ and $\eta_+ = 2^{2^k}\eta_\epsilon$ as upper bound.

Lemma 1. *If* $\eta \geq \eta_{\max} \triangleq \frac{2\alpha}{\alpha-1} \cdot \frac{d_0}{\sqrt{\alpha\|g_0\|^2}}$ *then* $\eta > h(\eta)$. *Then* $\mathcal{O}(\log\log \frac{\eta_{\max}}{\eta_\epsilon})$ *bisection steps are enough.*

The proof is by ***contradiction***, simple, and thus omitted here.

Useful in Online Ensemble?

- Can we use a similar step-size pool here?

*We take **full-information dynamic regret minimization** as an example.*

Set the step size pool as

$$\left\{\eta_1 = \frac{1}{\sqrt{T}}, \eta_i = 2^{2^i}\eta_1, \eta_N = \mathcal{O}(1)\right\} \quad \Longrightarrow \quad N \approx \mathcal{O}(\log\log T)$$

More efficient than the classic $\mathcal{O}(\log T)$ size.

How will this new step size pool affect the regret bound?

Dynamic regret of the i-th base learner can be bounded by $\mathcal{O}\left(\frac{1+P_T}{\eta_i} + \eta_i T\right)$.

Suppose there exist $\eta_{i^\star}$ such that $\eta_{i^\star} \leq \eta_\star \leq \eta_{i^\star+1}$. $\Longrightarrow$ $\eta_{i^\star} T \leq \eta_\star T$

(this term remains to be tuned)

Next step: find a *lower* bound for $\eta_{i^\star}$ to handle the term $\frac{1+P_T}{\eta_{i^\star}}$.

Useful in Online Ensemble?

LAMDA
Learning And Mining from DatA

- Can we use a similar step-size pool here?

*We take **full-information dynamic regret minimization** as an example.*

Set the step size pool as

$$\left\{\eta_1 = \frac{1}{\sqrt{T}}, \eta_i = 2^{2^i}\eta_1, \eta_N = \mathcal{O}(1)\right\} \Rightarrow N \approx \mathcal{O}(\log\log T)$$

More efficient than the classic $\mathcal{O}(\log T)$ size.

How will this new step size pool affect the regret bound?

Suppose there exist $\eta_{i^\star}$ such that $\eta_{i^\star} \leq \eta_\star \leq \eta_{i^\star+1}$.

Next step: find a *lower* bound for $\eta_{i^\star}$ to handle the term $\frac{1+P_T}{\eta_{i^\star}}$.

$$\eta_{i+1} = 2^{2^{i+1}}\eta_1 \text{ and } \eta_i = 2^{2^i}\eta_1 \Rightarrow \frac{\eta^{i+1}}{\eta^i} = \frac{2^{2^{i+1}}}{2^{2^i}} = 2^{2^{i+1}-2^i} = 2^{2^i} \Rightarrow \eta_{i^\star} \geq \frac{\eta_\star}{2^{2^{i^\star}}}$$

$$\Rightarrow \frac{1+P_T}{\eta_{i^\star}} \leq \frac{1+P_T}{\eta_\star} \cdot 2^{2^{i^\star}}$$

Last question: How large is $2^{2^{i^\star}}$?

Useful in Online Ensemble?

LAMDA
Learning And Mining from DatA

- Can we use a similar step-size pool here?

*We take **full-information dynamic regret minimization** as an example.*

Set the step size pool as

$$\left\{\eta_1 = \frac{1}{\sqrt{T}}, \eta_i = 2^{2^i}\eta_1, \eta_N = \mathcal{O}(1)\right\} \quad \Longrightarrow N \approx \mathcal{O}(\log\log T)$$

More efficient than the classic $\mathcal{O}(\log T)$ size.

How will this new step size pool affect the regret bound?

Suppose there exist $\eta_{i^\star}$ such that $\eta_{i^\star} \le \eta_\star \le \eta_{i^\star+1}$. **Last question:** How large is $2^{2^{i^\star}}$?

$$\left(\eta_{i^\star} \triangleq 2^{2^{i^\star}}\eta_1\right) \le \eta_\star \quad \Longrightarrow 2^{2^{i^\star}} \le \frac{\eta_\star}{\eta_1} \approx \eta_\star \cdot \sqrt{T}$$

$$\frac{1+P_T}{\eta_{i^\star}} + \eta_{i^\star}T \le \frac{1+P_T}{\eta_\star}\cdot 2^{2^{i^\star}} + \eta_\star T \le \frac{1+P_T}{\eta_\star}\cdot\eta_\star\cdot\sqrt{T} + \eta_{i^\star}T \lesssim \sqrt{T}(1+P_T)$$

Useless. $\mathcal{O}(\log T)$ seems necessary.

Stochasticity Analysis

Learning And Mining from DatA

- Recall the starting analysis

If we run SGD for T rounds and submit $\bar{x} \triangleq \frac{1}{T}\sum_{t=1}^{T} x_t$.

$$f(\bar{x}) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(x_t), x_t - x_\star\rangle$$

$$= \frac{1}{T}\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle + \frac{1}{T}\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star\rangle \qquad (\Delta_t \triangleq \nabla f(x_t) - g_t)$$

algorithm-related *algorithm-irrelated* due to *unknown* $\nabla f(x_t)$

Next step: analyze the stochastic gap $\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star\rangle$. $\Rightarrow$ *equals zero for expectation bounds.*

For *high-probability* result $\Rightarrow$ **Solution:** *concentration*

Stochasticity Analysis

Learning And Mining from DatA

- Concentration $\sum_{t=1}^{T}\langle \nabla f(x_t) - g_t, x_t - x_\star\rangle$

Stochastic part: gap between the true gradient $\nabla f(x_t)$ and its stochastic sample g_t.

⇨ ***concentration*** inequality

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \le c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le c$ with probability 1,*

$$\Pr\left[\exists t \le T : \left|\sum_{i=1}^{t} X_i\right| \ge 4\sqrt{\theta_{t,\delta}\sum_{i=1}^{t}\left(X_i - \hat{X}_i\right)^2 + c^2\theta_{t,\delta}^2}\right] \le \delta,$$

where $\theta_{t,\delta} \triangleq \log\frac{60\log(6t)}{\delta}$.

Stochasticity Analysis

Learning And Mining from DatA

- How to handle ***boundedness***?

$$\sum_{t=1}^{T}\langle \nabla f(x_t) - g_t, x_t - x_\star \rangle$$

$x_t - x_\star$ is ***unbounded***!

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq c$ with probability 1,*

$$\Pr\left[\exists t \leq T : \left|\sum_{i=1}^{t} X_i\right| \geq 4\sqrt{\theta_{t,\delta}\sum_{i=1}^{t}\left(X_i - \hat{X}_i\right)^2 + c^2\theta_{t,\delta}^2}\right] \leq \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60\log(6t)}{\delta}$.

How to construct a bounded sequence? $\nabla f(x_t) - g_t$ is bounded by assuming ***bounded noise***.

First try to import boundedness: $\left(d_t \triangleq \|x_t - x_\star\|\right) \leq \left(\bar{d}_t \triangleq \max_{i \leq t} \|x_i - x_\star\|\right)$

$$\Rightarrow \sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star \rangle = \bar{d}_T \sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{\bar{d}_T}\right\rangle$$

Is the concentration applicable?

No! $X_t = \left\langle \Delta_t, \frac{x_t - x_\star}{\bar{d}_T}\right\rangle$ is not adapted to $\mathcal{F}_t$.

Stochasticity Analysis

Learning And Mining from DatA

- How to handle ***boundedness***?

$$\sum_{t=1}^{T}\langle \nabla f(x_t) - g_t, x_t - x_\star \rangle$$

$x_t - x_\star$ is ***unbounded***!

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \le c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le c$ with probability 1,*

$$\Pr\left[\exists t \le T : \left|\sum_{i=1}^{t} X_i\right| \ge 4\sqrt{\theta_{t,\delta}\sum_{i=1}^{t}\left(X_i - \hat{X}_i\right)^2 + c^2\theta_{t,\delta}^2}\right] \le \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60\log(6t)}{\delta}$.

How to construct a bounded sequence? $\nabla f(x_t) - g_t$ is bounded by assuming ***bounded noise***.

Second try to import boundedness: $\left(d_t \triangleq \|x_t - x_\star\|\right) \le \left(\bar{d}_t \triangleq \max_{i \le t}\|x_i - x_\star\|\right)$

$$\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star\rangle = \sum_{t=1}^{T} d_t \cdot \left\langle \Delta_t, \frac{x_t - x_\star}{d_t}\right\rangle \le \bar{d}_T \sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{d_t}\right\rangle$$

$X_t = \langle \Delta_t, (x_t - x_\star)/d_t\rangle$ is now adapted to $\mathcal{F}_t$.

Is this correct? No! $\langle \Delta_t, (x_t - x_\star)/d_t\rangle$ can be negative.

Stochasticity Analysis

Learning And Mining from DatA

- How to handle ***boundedness***?

$$\sum_{t=1}^{T}\langle \nabla f(x_t) - g_t, x_t - x_\star \rangle$$

$x_t - x_\star$ is ***unbounded***!

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \le c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le c$ with probability 1,*

$$\Pr\left[\exists t \le T : \left|\sum_{i=1}^{t} X_i\right| \ge 4\sqrt{\theta_{t,\delta}\sum_{i=1}^{t}\left(X_i - \hat{X}_i\right)^2 + c^2\theta_{t,\delta}^2}\right] \le \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60\log(6t)}{\delta}$.

How to construct a bounded sequence? $\nabla f(x_t) - g_t$ is bounded by assuming ***bounded noise***.

$$\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star \rangle = s_{k_T}\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_{k_T}} \right\rangle$$

The same issue that $X_t = \left\langle \Delta_t, \frac{x_t - x_\star}{s_{k_T}} \right\rangle$ is not adapted to $\mathcal{F}_t$.

Solution: $s_{k_t} \triangleq 2^{k_t} d_0$, where $k_t \triangleq \left\lceil \log \frac{d_t}{d_0} \right\rceil$ $\Rightarrow$ Intuitively, $d_t \le s_{k_t} \le 2d_t$

Key idea: use ***discrete*** s_{k_t} to replace the ***coutinuous*** d_t.

Stochasticity Analysis

Learning And Mining from DatA

- How to handle ***boundedness***?

$$\sum_{t=1}^{T}\langle \nabla f(x_t) - g_t, x_t - x_\star \rangle$$

$x_t - x_\star$ is ***unbounded***!

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \le c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le c$ with probability 1,*

$$\Pr\left[\exists t \le T : \left|\sum_{i=1}^{t} X_i\right| \ge 4\sqrt{\theta_{t,\delta}\sum_{i=1}^{t}\left(X_i - \hat{X}_i\right)^2 + c^2\theta_{t,\delta}^2}\right] \le \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60\log(6t)}{\delta}$.

How to construct a bounded sequence? $\nabla f(x_t) - g_t$ is bounded by assuming ***bounded noise***.

$$\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star \rangle = s_{k_T}\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_{k_T}} \right\rangle$$

Solution: $s_{k_t} \triangleq 2^k d_0$, where $k_t \triangleq \left\lceil \log \frac{d_t}{d_0} \right\rceil$

$\Rightarrow$ Intuitively, $d_t \le s_{k_t} \le 2d_t$

Key idea: use ***discrete*** s_{k_t} to replace the ***coutinuous*** d_t.

$$\Pr\left[\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_{k_T}} \right\rangle \ge X\right] \le \sum_{k}\Pr\left[\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_k} \right\rangle \ge X\right]$$

(union bound)

s_k is now ***unrelated*** to the filtration.

Stochasticity Analysis

Learning And Mining from DatA

- How to handle ***boundedness***?

$$\sum_{t=1}^{T}\langle \nabla f(x_t) - g_t, x_t - x_\star \rangle$$

$x_t - x_\star$ is ***unbounded***!

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \le c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le c$ with probability 1,*

$$\Pr\left[\exists t \le T : \left|\sum_{i=1}^{t} X_i\right| \ge 4\sqrt{\theta_{t,\delta}\sum_{i=1}^{t}\left(X_i - \hat{X}_i\right)^2 + c^2\theta_{t,\delta}^2}\right] \le \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60\log(6t)}{\delta}$.

How to construct a bounded sequence? $\nabla f(x_t) - g_t$ is bounded by assuming ***bounded noise***.

$$\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star \rangle = s_{k_T}\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_{k_T}} \right\rangle$$

Solution: $s_{k_t} \triangleq 2^k d_0$, where $k_t \triangleq \left\lceil \log \frac{d_t}{d_0} \right\rceil$

$\Rightarrow$ Intuitively, $d_t \le s_{k_t} \le 2d_t$

Key idea: use ***discrete*** s_{k_t} to replace the ***coutinuous*** d_t.

$$\Pr\left[\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_{k_T}} \right\rangle \ge X\right] \le \sum_{k}\Pr\left[\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_k} \right\rangle \ge X\right]$$

Next step: how many possible s_k in the T-th round.

Stochasticity Analysis

Learning And Mining from DatA

- How to handle ***boundedness***?

$$\sum_{t=1}^{T}\langle \nabla f(x_t) - g_t, x_t - x_\star \rangle$$

$x_t - x_\star$ is ***unbounded***!

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \le c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le c$ with probability 1,*

$$\Pr\left[\exists t \le T : \left|\sum_{i=1}^{t} X_i\right| \ge 4\sqrt{\theta_{t,\delta}\sum_{i=1}^{t}\left(X_i - \hat{X}_i\right)^2 + c^2\theta_{t,\delta}^2}\right] \le \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60\log(6t)}{\delta}$.

How to construct a bounded sequence? $\nabla f(x_t) - g_t$ is bounded by assuming ***bounded noise***.

Solution: $s_{k_t} \triangleq 2^k d_0$, where $k_t \triangleq \left\lceil \log \frac{d_t}{d_0} \right\rceil$ ⇨ Intuitively, $d_t \le s_{k_t} \le 2d_t$

Key idea: use ***discrete*** s_{k_t} to replace the ***coutinuous*** d_t.

$$\Pr\left[\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_{k_T}}\right\rangle \ge X\right] \le \sum_{k}\Pr\left[\sum_{t=1}^{T}\left\langle \Delta_t, \frac{x_t - x_\star}{s_k}\right\rangle \ge X\right]$$

Next step: how many possible s_k in the T-th round.

$$d_{t+1} = \|x_{t+1} - x_\star\| = \|x_t - \eta g_t - x_\star\| \le d_t + \eta(G + \sigma) \quad \Rightarrow d_t = \mathcal{O}(t) \quad \Rightarrow k_t = \mathcal{O}(\log t)$$

Overall

Final result overview

Theorem 7 *In the noiseless setting Algorithm 1, with parameters* $\alpha^{(k)} = 3$, $\beta^{(k)} = 0$, $\eta_\varepsilon > 0$, $B \in \mathbb{N}$, *and* $x_0 \in \mathbb{R}^d$, *performs at most* B *subgradient queries and returns* $\bar{x} = \frac{1}{T}\sum_{i<T} x_i(\eta) \in \mathbb{R}^d$

for some $\eta \geq \eta_\varepsilon$ *and integer* T *satisfying*

$$T \geq \max\left\{\frac{B}{12\log\log_+ \frac{\|x_0 - x_\star\|}{\eta_\varepsilon \|g_0\|}}, 1\right\} \tag{10}$$

such that either

$$\|\bar{x} - x_\star\| \leq 4\|x_0 - x_\star\| \;\; \textit{and} \;\; f(\bar{x}) - f(x_\star) \leq \sqrt{27}\,\frac{\|x_0 - x_\star\|\sqrt{G_T(\eta')}}{T} \tag{11}$$

for some $\eta' \in [\eta, 2\eta]$, *or* $\eta = \eta_\varepsilon$ *and*

$$\|\bar{x} - x_\star\| \leq \eta_\varepsilon\sqrt{3G_T(\eta_\varepsilon)} \;\; \textit{and} \;\; f(\bar{x}) - f(x_\star) \leq 2\,\frac{\eta_\varepsilon G_T(\eta_\varepsilon)}{T}. \tag{12}$$

Remarks:

(i) Need to consider the corner case where η_ϵ cannot serve as the lower bound for bisection.

(ii) Not fully parameter-free since the learner requires T.

(iii) Only with a $\log\log T$ overhead compared with the tuned optimal bound.

Overall

- Three key ideas overview
  - Formulate the problem as bisection
  - An aggressive bisection method
  - Construct bounded sequence for concentration

One Question

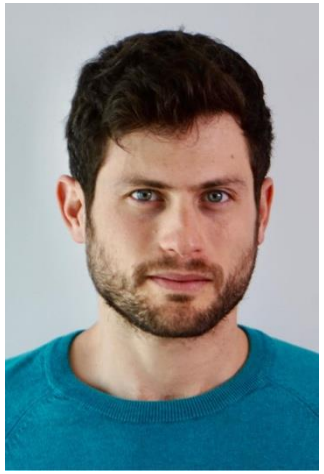
- Still fixed-point problem with stochastic gradients?

$$\textbf{\textit{fixed-point}} \text{ problem: } \eta = \frac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta)}}$$

For ***noiseless*** (deterministic) case, i.e., $g_t = \nabla f(x_t)$, given η, $G_T \triangleq \sum_{t \leq T} \|g_t\|^2$ can be ***uniquely*** decided, and we can consider G_T as $G_T(\eta)$.

For ***stochastic*** case, i.e., $\mathbb{E}[g_t] = \nabla f(x_t)$, given η, $G_T \triangleq \sum_{t \leq T} \|g_t\|^2$ ***cannot*** be uniquely decided due to the additional ***noise in*** $\{g_t\}_{t=1}^T$. Can G_T still be represented as $G_T(\eta)$?

DOG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule



Maor Ivgi
Tel Aviv University

Oliver Hinder
University of Pittsburgh

Yair Carmon
Tel Aviv University

Difference

- What's different from COLT'22?

COLT 2022	This work
time-invariant step size	*time-varying* step size **DOG (Distance over Gradients)**
$\eta_\dagger \approx \frac{\bar{r}_T}{\sqrt{\alpha G_T}}$ $(\bar{r}_T \triangleq \max_{t \le T} \|x_t - x_0\|)$ $(G_T \triangleq \sum_{t \le T} \|g_t\|^2)$	$\eta_t = \frac{\bar{r}_t}{\sqrt{\alpha G_t}}$ $(\bar{r}_t \triangleq \max_{i \le t} \|x_i - x_0\|)$ $(G_t \triangleq \sum_{i \le t} \|g_i\|^2)$
Bisection is needed	Bisection is ***not*** needed
Analysis is ***complicated***	Analysis is ***much simpler***

Analysis

Starting from the begining

If we run SGD for T rounds and submit $\bar{x} \triangleq \frac{1}{T}\sum_{t=1}^{T} x_t$.

$$f(\bar{x}) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x_\star) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(x_t), x_t - x_\star\rangle$$

$$= \frac{1}{T}\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle + \frac{1}{T}\sum_{t=1}^{T}\langle \Delta_t, x_t - x_\star\rangle \qquad (\Delta_t \triangleq \nabla f(x_t) - g_t)$$

algorithm-related *algorithm-irrelated* due to *unknown* $\nabla f(x_t)$

Next step: analyze $\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle$.

Algorithm-Related Analysis

LAMDA
Learning And Mining from DatA

- What's different when using **DOG** step size?

By standard analysis as before: $\sum_{t=1}^{T}\langle g_t, x_t - x_\star\rangle \lesssim \sum_{t=1}^{T}\frac{d_t^2 - d_{t+1}^2}{\eta_t} + \sum_{t=1}^{T}\eta_t\|g_t\|^2$

$$\sum_{t=1}^{T}\frac{d_t^2 - d_{t+1}^2}{\eta_t} \Rightarrow \boxed{\eta_t = \frac{\bar{r}_t}{\sqrt{G_t}} \quad (\bar{r}_t \triangleq \max_{i\le t}\|x_i - x_0\|) \quad (G_t \triangleq \sum_{i\le t}\|g_i\|^2)} \Rightarrow \sum_{t=1}^{T}\frac{d_t^2 - d_{t+1}^2}{\bar{r}_t}\cdot\sqrt{G_t}$$

*This term is hard to analyze due to the **DOG** step size.*

Solution: use a weighted combination $\bar{x} \triangleq \frac{1}{\sum_{t=1}^{T}\bar{r}_t}\sum_{t=1}^{T}\bar{r}_t x_t$

Algorithm-Related Analysis

Learning And Mining from DatA

$$\eta_t = \frac{\bar{r}_t}{\sqrt{G_t}} \quad (\bar{r}_t \triangleq \max_{i \le t} \|x_i - x_0\|) \quad (G_t \triangleq \sum_{i \le t} \|g_i\|^2)$$

- What's different when using **DOG** step size?

$$f(\bar{x}) - f(x_\star) \le \frac{1}{\sum_{t=1}^T \bar{r}_t} \left(\sum_{t=1}^T \bar{r}_t \cdot \langle g_t, x_t - x_\star \rangle + \sum_{t=1}^T \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle \right) \quad (\Delta_t \triangleq \nabla f(x_t) - g_t)$$

algorithm-related ***algorithm-irrelated*** due to ***unknown*** $\nabla f(x_t)$

$$\Rightarrow \sum_{t=1}^T \bar{r}_t \cdot \langle g_t, x_t - x_\star \rangle \lesssim \sum_{t=1}^T \frac{\bar{r}_t}{\eta_t} (d_t^2 - d_{t+1}^2) + \sum_{t=1}^T \eta_t \bar{r}_t \|g_t\|^2$$

$$\Rightarrow \sum_{t=1}^T \frac{\bar{r}_t}{\eta_t} (d_t^2 - d_{t+1}^2) \le \sum_{t=1}^T (d_t^2 - d_{t+1}^2) \cdot \sqrt{G_t} \le \cdots \lesssim \bar{r}_T \bar{d}_T \sqrt{G_T} \quad \textit{(standard derivation)}$$

$$\Rightarrow \sum_{t=1}^T \eta_t \bar{r}_t \|g_t\|^2 \le \sum_{t=1}^T \bar{r}_t{}^2 \cdot \frac{\|g_t\|^2}{G_t} \lesssim \bar{r}_T^2 \sqrt{G_T} \quad \textit{(self-confident tuning)}$$

Algorithm-Related Analysis

Learning And Mining from DatA

- What's different when using **DOG** step size?

$$\eta_t = \frac{\bar{r}_t}{\sqrt{G_t}} \quad (\bar{r}_t \triangleq \max_{i \le t} \|x_i - x_0\|) \quad (G_t \triangleq \sum_{i \le t} \|g_i\|^2)$$

$$f(\bar{x}) - f(x_\star) \le \frac{1}{\sum_{t=1}^T \bar{r}_t} \left(\sum_{t=1}^T \bar{r}_t \cdot \langle g_t, x_t - x_\star \rangle + \sum_{t=1}^T \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle \right) \quad (\Delta_t \triangleq \nabla f(x_t) - g_t)$$

algorithm-related *algorithm-irrelated* due to *unknown* $\nabla f(x_t)$

Algorithm-related part:

$$\sum_{t=1}^T \bar{r}_t \cdot \langle g_t, x_t - x_\star \rangle \lesssim \sum_{t=1}^T \frac{\bar{r}_t}{\eta_t} (d_t^2 - d_{t+1}^2) + \sum_{t=1}^T \eta_t \bar{r}_t \|g_t\|^2 \lesssim \bar{r}_T (\bar{d}_T + \bar{r}_T) \sqrt{G_T}$$

Next step: analyze the stochastic gap $\sum_{t=1}^T \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle$.

Stochasticity Analysis

Weighted Concentration

To handle the stochastic gap $\sum_{t=1}^{T} \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle$, we need a ***weighted*** version of concentration inequality.

Lemma 2 (Uniform). *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq c$ with probability 1 for all t. Then for all $\delta \in (0, 1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq c$ with probability 1,*

$$\Pr\left[\exists t \leq T : \left|\sum_{i=1}^{t} X_i\right| \geq 4\sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}\right] \leq \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60 \log(6t)}{\delta}$.

Stochasticity Analysis

Weighted Concentration

To handle the stochastic gap $\sum_{t=1}^{T} \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle$, we need a ***weighted*** version of concentration inequality.

Lemma 3 (Weighted). *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq c$ with probability 1 for all t. Then for all $\delta \in (0, 1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq c$ with probability 1,*

$$\Pr\left[\exists t \leq T : \left|\sum_{i=1}^{t} Y_i X_i\right| \geq 8 Y_t \sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}\right] \leq \delta,$$

where $\{Y_t\}$ is a non-negetive and non-decreasing sequence and $\theta_{t,\delta} \triangleq \log \frac{60 \log(6t)}{\delta}$.

Stochasticity Analysis

Weighted Concentration

The proof is simple with the ***uniform concentration*** and a ***useful technical lemma***.

Lemma 3. *Let $c > 0$ and X_t be a* ***martingale difference*** *sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq c$ with probability 1,*

$$\Pr\left[\exists t \leq T : \left|\sum_{i=1}^{t} Y_i X_i\right| \geq 8 Y_t \sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}\right] \leq \delta,$$

where $\{Y_t\}$ is a non-negetive and non-decreasing sequence and $\theta_{t,\delta} \triangleq \log \frac{60 \log(6t)}{\delta}$.

Lemma 4. *Let $\{a_t\}_{t=1}^T$ be a non-negative and non-decreasing sequence and $\{b_t\}_{t=1}^T$ be any sequence. Then it holds that*

$$\left|\sum_{i=1}^{t} a_i b_i\right| \leq 2 a_t \max_{i \leq t} \left|\sum_{j=1}^{i} b_j\right|.$$

Stochasticity Analysis

LAMDA
Learning And Mining from DatA

Weighted Concentration

The proof is simple with the ***uniform concentration*** and a ***useful technical lemma***.

Lemma 3. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq c$ with probability 1,*

$$\Pr\left[\exists t \leq T : \left|\sum_{i=1}^{t} Y_i X_i\right| \geq 8 Y_t \sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}\right] \leq \delta,$$

where $\{Y_t\}$ is a non-negetive and non-decreasing sequence and $\theta_{t,\delta} \triangleq \log \frac{60 \log(6t)}{\delta}$.

Lemma 4. *Let $\{a_t\}_{t=1}^T$ be a non-negative and non-decreasing sequence and $\{b_t\}_{t=1}^T$ be any sequence. Then it holds that*

$$\left|\sum_{i=1}^{t} a_i b_i\right| \leq 2 a_t \max_{i \leq t} \left|\sum_{j=1}^{i} b_j\right|.$$

+

Lemma 2. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq c$ with probability 1,*

$$\Pr\left[\exists t \leq T : \left|\sum_{i=1}^{t} X_i\right| \geq 4 \sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}\right] \leq \delta,$$

where $\theta_{t,\delta} \triangleq \log \frac{60 \log(6t)}{\delta}$.

$\Rightarrow$ With probability $1 - \delta$,

$$\left|\sum_{i=1}^{t} Y_i X_i\right| \leq 2 Y_t \max_{i \leq t} \left|\sum_{j=1}^{i} X_j\right| \leq 2 Y_t \cdot \max_{i \leq t} 4 \sqrt{\theta_{i,\delta} \sum_{j=1}^{i} \left(X_j - \hat{X}_j\right)^2 + c^2 \theta_{i,\delta}^2} \leq 8 Y_t \sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}$$

Stochasticity Analysis

Learning And Mining from DatA

Weighted Concentration

Lemma 3. *Let $c > 0$ and X_t be a **martingale difference** sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq c$ with probability 1 for all t. Then for all $\delta \in (0,1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq c$ with probability 1,*

$$\Pr\left[\exists t \leq T : \left|\sum_{i=1}^{t} Y_i X_i\right| \geq 8 Y_t \sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}\right] \leq \delta,$$

where $\{Y_t\}$ is a non-negetive and non-decreasing sequence and $\theta_{t,\delta} \triangleq \log \frac{60 \log(6t)}{\delta}$.

$$\sum_{t=1}^{T} \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle = \sum_{t=1}^{T} \underbrace{\bar{r}_t \bar{d}_t}_{Y_t} \cdot \underbrace{\left\langle \Delta_t, \frac{x_t - x_\star}{\bar{d}_t} \right\rangle}_{X_t} \qquad (\bar{d}_t \triangleq \max_{i \leq t} \|x_i - x_\star\|)$$

Using **Lemma 3** accordingly, with probability $1 - \delta$,

$$\sum_{t=1}^{T} \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle \lesssim \bar{r}_T \bar{d}_T \sqrt{\theta_{T,\delta} G_T + \theta_{T,\delta}^2 G^2}$$

Overall Analysis

- Combining two parts

Algorithm-related part:
$$\sum_{t=1}^{T} \bar{r}_t \cdot \langle g_t, x_t - x_\star \rangle \lesssim \bar{r}_T (\bar{d}_T + \bar{r}_T) \sqrt{G_T}$$

Stochastic part:
$$\sum_{t=1}^{T} \bar{r}_t \cdot \langle \Delta_t, x_t - x_\star \rangle \lesssim \bar{r}_T \bar{d}_T \sqrt{\theta_{T,\delta} G_T + \theta_{T,\delta}^2 G^2}$$

$$\Rightarrow \quad f(\bar{x}) - f(x_\star) \leq \frac{1}{\sum_{t=1}^{T} \bar{r}_t} \left(\sum_{t=1}^{T} \bar{r}_t \langle g_t, x_t - x_\star \rangle + \sum_{t=1}^{T} \bar{r}_t \langle \Delta_t, x_t - x_\star \rangle \right)$$

$$\lesssim \frac{\bar{r}_T (\bar{d}_T + \bar{r}_T) \sqrt{\theta_{T,\delta} G_T + \theta_{T,\delta}^2 G^2}}{\sum_{t=1}^{T} \bar{r}_t} \lesssim \frac{(d_0 + \bar{r}_T) \sqrt{\theta_{T,\delta} G_T + \theta_{T,\delta}^2 G^2}}{\sum_{t=1}^{T} \bar{r}_t / \bar{r}_T}$$

(Later details omitted…)

Overall

- Three key ideas overview
  - **DOG** step size (time-varying version of COLT 2022)
  - Weighted combination for final decision
  - Weighted concentration

Summary

- **Problem:** Parameter-free stochastic optimization
- **Two Works from Yair Carmon & Oliver Hinder**
  - Time-invariant step size with bisection method [COLT 2022]
    - Key technique: ***bisection formulation***
  - Time-varying step size (**DOG**) [ICML 2023]
    - Key technique: ***weighted concentration***
- **Following Work:** para.-free $1/T^2$ rate of convex & smooth function [COLT 2024]

Thanks!