The Power of Adaptivity in SGD: Self-Tuning Step Sizes with Unbounded Gradients and Affine Variance

Speaker: Matthew Faw

Authors: F*, Isidoros Tziotis*, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, Rachel Ward
The University of Texas at Austin

TEXAS
The University of Texas at Austin

COLT

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

Stochastic Optimization with SGD

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

Stochastic Gradient Descent

$$w_{t+1} = w_t - \eta_t g_t$$

Achieves $\min_t \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate under the following assumptions:

Stochastic Optimization with SGD

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

Stochastic Gradient Descent

$$w_{t+1} = w_t - \eta_t g_t$$

Achieves $\min_t \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate under the following assumptions:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2$ (bounded variance)

Stochastic Optimization with SGD

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

Stochastic Gradient Descent

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \min\left\{\frac{1}{L}, \frac{c}{\sqrt{T}\sigma_0}\right\}$$

Achieves $\min_t \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate under the following assumptions:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2$ (bounded variance)

Stochastic Optimization with SGD

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

Stochastic Gradient Descent

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \min\left\{\frac{1}{L(1+\sigma_1^2)}, \frac{c}{\sqrt{T}\sigma_0}\right\}$$

Achieves $\min_t \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate under the following assumptions:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(w)\|^2$ (*affine* variance)

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

Stochastic Gradient Descent

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \min\left\{\frac{1}{L(1+\sigma_1^2)}, \frac{c}{\sqrt{T}\sigma_0}\right\}$$

Achieves $\min_t \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate under the following assumptions:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(w)\|^2$ (*affine* variance)

Can **oscillate** or **diverge** if L or σ_1^2 is underestimated!

Stochastic Optimization with SGD

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

Stochastic Gradient Descent

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \min\left\{\frac{1}{L(1+\sigma_1^2)}, \frac{c}{\sqrt{T}\sigma_0}\right\}$$

Achieves $\min_t \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate under the following assumptions:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(w)\|^2$ (*affine* variance)

Can **oscillate** or **diverge** if L or σ_1^2 is underestimated!

Does **not** recover improved $O(1/T)$ rate when σ_0^2 is small and unknown!

Stochastic Optimization with Adaptive Step Sizes

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

AdaGrad-Norm

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} \|g_s\|^2}}$$

Prior work shows that adaptive step sizes can overcome the drawbacks of SGD!*

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

AdaGrad-Norm

Similar to SGD

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} \|g_s\|^2}}$$

Prior work shows that adaptive step sizes can overcome the drawbacks of SGD!*

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

AdaGrad-Norm

Similar to SGD

$$w_{t+1} = w_t - \eta_t g_t$$

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} \|g_s\|^2}}$$

Sum of stochastic gradients

Prior work shows that adaptive step sizes can overcome the drawbacks of SGD!*

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

AdaGrad-Norm

Similar to SGD

$$w_{t+1} = w_t - \eta_t g_t$$

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} \|g_s\|^2}}$$

Free parameters

Sum of stochastic gradients

Prior work shows that adaptive step sizes can overcome the drawbacks of SGD!*

Problem Setup

Find a first-order stationary point of a non-convex, L-smooth function F.

AdaGrad-Norm

Similar to SGD

$$w_{t+1} = w_t - \eta_t g_t$$

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} \|g_s\|^2}}$$

Free parameters

Sum of stochastic gradients

Prior work shows that adaptive step sizes can overcome the drawbacks of SGD!*

*With major caveats

Unique challenges connected to adaptive methods!

Challenges for Adaptive Methods

Adaptive step-sizes present two unique challenges

Challenges for Adaptive Methods

Adaptive step-sizes present two unique challenges

i. **Challenge 1**: Bias + Affine Variance
   - Step size η_t depends on past and **current stochastic gradients**.

Challenges for Adaptive Methods

Adaptive step-sizes present two unique challenges

Descent direction $-\eta_t g_t$ is **biased!**

i. **Challenge 1**: Bias + Affine Variance
   - Step size η_t depends on past and **current stochastic gradients**.

Adaptive step-sizes present two unique challenges

Descent direction $-\eta_t g_t$ is **biased!**

i. **Challenge 1**: Bias + Affine Variance

- Step size η_t depends on past and **current stochastic gradients**.

- Especially challenging in presence of **affine variance**.

$$\mathbb{E}_t[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

Adaptive step-sizes present two unique challenges

Descent direction $-\eta_t g_t$ is **biased!**

i. **Challenge 1**: Bias + Affine Variance
   - Step size η_t depends on past and **current stochastic gradients**.
   - Especially challenging in presence of **affine variance**.

$$(1 - \sigma_1 \cdot bias_t)\mathbb{E}_t[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

Large $\sigma_1 \implies$ techniques from bounded variance regimes **completely break**!

Challenges for Adaptive Methods

Adaptive step-sizes present two unique challenges

Descent direction $-\eta_t g_t$ is **biased!**

i. **Challenge 1**: Bias + Affine Variance
   - Step size η_t depends on past and **current stochastic gradients**.
   - Especially challenging in presence of **affine variance**.

$$(1 - \sigma_1 \cdot bias_t)\mathbb{E}_t[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

Large $\sigma_1 \implies$ techniques from bounded variance regimes **completely break**!

ii. **Challenge 2**: Step size scaling
   - Can no longer guarantee deterministically that $\eta_t \sim 1/\sqrt{T}$.
   - Depends inversely on $\sum_t \|g_t\|^2$.

Challenges for Adaptive Methods

Adaptive step-sizes present two unique challenges

Descent direction $-\eta_t g_t$ is **biased!**

i. **Challenge 1**: Bias + Affine Variance

- Step size η_t depends on past and **current stochastic gradients**.
- Especially challenging in presence of **affine variance**.

$$(1 - \sigma_1 \cdot bias_t)\mathbb{E}_t[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$$

Large $\sigma_1 \Longrightarrow$ techniques from bounded variance regimes **completely break**!

ii. **Challenge 2**: Step size scaling

- Can no longer guarantee deterministically that $\eta_t \sim 1/\sqrt{T}$.
- Depends inversely on $\sum_t \|g_t\|^2$.

Possibly **unbounded!**

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. “Bias-free” – [Li-Orabona’19,’20; Savarese-McAllester-Babu-Maire’21; …]

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. "Bias-free" – [Li-Orabona'19,'20; Savarese-McAllester-Babu-Maire'21; …]
    - $\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t-1} \|g_s\|^2}} \Longrightarrow \eta_t$ conditionally independent of g_t!
    - Automatically achieves faster convergence when σ_0 is small (unlike SGD).
    - **Needs knowledge of L** for convergence guarantee (like SGD).

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. "Bias-free" – [Li-Orabona'19,'20; Savarese-McAllester-Babu-Maire'21; …]
   - $\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t-1} \|g_s\|^2}} \Longrightarrow \eta_t$ conditionally independent of g_t!
   - Automatically achieves faster convergence when σ_0 is small (unlike SGD).
   - **Needs knowledge of L** for convergence guarantee (like SGD).

2. "Bounded gradients" – [Ward-Wu-Bottou'19; Kavis-Levy-Bach-Cevher'19; Défossez-Bottou-Bach-Usunier'20; …]

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. "Bias-free" – [Li-Orabona'19,'20; Savarese-McAllester-Babu-Maire'21; …]
   - $\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t-1} \|g_s\|^2}} \Longrightarrow \eta_t$ conditionally independent of g_t!
   - Automatically achieves faster convergence when σ_0 is small (unlike SGD).
   - **Needs knowledge of L** for convergence guarantee (like SGD).

2. "Bounded gradients" – [Ward-Wu-Bottou'19; Kavis-Levy-Bach-Cevher'19; Défossez-Bottou-Bach-Usunier'20; …]
   - Develop techniques to bound bias.
   - Converges w/o knowing L or σ_0.
   - But relies on **uniform gradient bounds** $\sup_{w \in \mathbb{R}^d} \|\nabla F(w)\|^2 \leq B < \infty$

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. "Bias-free" – [Li-Orabona'19,'20; Savarese-McAllester-Babu-Maire'21; …]
   - $\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t-1} \|g_s\|^2}} \Longrightarrow \eta_t$ conditionally independent of g_t!
   - Automatically achieves faster convergence when σ_0 is small (unlike SGD).
   - **Needs knowledge of L** for convergence guarantee (like SGD).

2. "Bounded gradients" – [Ward-Wu-Bottou'19; Kavis-Levy-Bach-Cevher'19; Défossez-Bottou-Bach-Usunier'20; …]
   - Develop techniques to bound bias.
   - Converges w/o knowing L or σ_0.
   - But relies on **uniform gradient bounds** $\sup_{w \in \mathbb{R}^d} \|\nabla F(w)\|^2 \leq B < \infty$

Rules out *strongly convex* functions: $F(w) = \|w\|^2$!

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. "Bias-free" – [Li-Orabona'19,'20; Savarese-McAllester-Babu-Maire'21; …]
   - $\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t-1} \|g_s\|^2}} \Longrightarrow \eta_t$ conditionally independent of g_t!
   - Automatically achieves faster convergence when σ_0 is small (unlike SGD).
   - **Needs knowledge of L** for convergence guarantee (like SGD).

2. "Bounded gradients" – [Ward-Wu-Bottou'19; Kavis-Levy-Bach-Cevher'19; Défossez-Bottou-Bach-Usunier'20; …]
   - Develop techniques to bound bias.
   - Converges w/o knowing L or σ_0.
   - But relies on **uniform gradient bounds** $\sup_{w \in \mathbb{R}^d} \|\nabla F(w)\|^2 \le B < \infty$

Rules out *strongly convex* functions: $F(w) = \|w\|^2$!

3. "Bounded variance with light noise" – [Kavis-Levy-Cevher'22]

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. "Bias-free" – [Li-Orabona'19,'20; Savarese-McAllester-Babu-Maire'21; …]
   - $\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t-1} \|g_s\|^2}} \Longrightarrow \eta_t$ conditionally independent of g_t!
   - Automatically achieves faster convergence when σ_0 is small (unlike SGD).
   - **Needs knowledge of L** for convergence guarantee (like SGD).

2. "Bounded gradients" – [Ward-Wu-Bottou'19; Kavis-Levy-Bach-Cevher'19; Défossez-Bottou-Bach-Usunier'20; …]
   - Develop techniques to bound bias.
   - Converges w/o knowing L or σ_0.
   - But relies on **uniform gradient bounds** $\sup_{w \in \mathbb{R}^d} \|\nabla F(w)\|^2 \leq B < \infty$

Rules out *strongly convex* functions: $F(w) = \|w\|^2$!

3. "Bounded variance with light noise" – [Kavis-Levy-Cevher'22]
   - Develop techniques to bound bias.
   - Converges w/o knowing L or σ_0.
   - Requires noise $\|g - \nabla F(w)\|^2$ to be **uniformly sub-Gaussian** ($\Longrightarrow$ bounded variance).

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Three main lines of Prior Work (**uniformly-bounded** variance):

1. "Bias-free" – [Li-Orabona'19,'20; Savarese-McAllester-Babu-Maire'21; …]
   - $\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t-1} \|g_s\|^2}} \Longrightarrow \eta_t$ conditionally independent of g_t!
   - Automatically achieves faster convergence when σ_0 is small (unlike SGD).
   - **Needs knowledge of L** for convergence guarantee (like SGD).

2. "Bounded gradients" – [Ward-Wu-Bottou'19; Kavis-Levy-Bach-Cevher'19; Défossez-Bottou-Bach-Usunier'20; …]
   - Develop techniques to bound bias.
   - Converges w/o knowing L or σ_0.
   - But relies on **uniform gradient bounds** $\sup_{w \in \mathbb{R}^d} \|\nabla F(w)\|^2 \leq B < \infty$

Rules out *strongly convex* functions: $F(w) = \|w\|^2$!

3. "Bounded variance with light noise" – [Kavis-Levy-Cevher'22]
   - Develop techniques to bound bias.
   - Converges w/o knowing L or σ_0.
   - Requires noise $\|g - \nabla F(w)\|^2$ to be **uniformly sub-Gaussian** ($\Longrightarrow$ bounded variance).

No prior proof techniques extend to the **affine variance** case
$\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(w)\|^2$

Stochastic Optimization with Adaptive Step Sizes – Prior Work

Question

Is there an adaptive method which:

- Converges at an $\tilde{O}\left(\frac{1}{\sqrt{T}}\right)$ rate under same assumptions as SGD?
  - $\mathbb{E}[\boldsymbol{g}] = \nabla F(w)$ (unbiased stochastic gradient)
  - $\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(w)\|^2$ (*affine* variance)
- Requires no knowledge of L, σ_0, or σ_1?

Question

Is there an adaptive method which:

- Converges at an $\tilde{O}\left(\frac{1}{\sqrt{T}}\right)$ rate under same assumptions as SGD?
  - $\mathbb{E}[\boldsymbol{g}] = \nabla F(w)$ (unbiased stochastic gradient)
  - $\mathbb{E}[\|g - \nabla F(w)\|^2] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(w)\|^2$ (*affine* variance)
- Requires no knowledge of L, σ_0, or σ_1?

Yes!

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - A key inequality may be *vacuous:*

$$(1-\sigma_1 \cdot bias_t)\mathbb{E}[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t)-F(w_{t+1})]+c\cdot\mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - A key inequality may be *vacuous:*

$$(1 - \sigma_1 \cdot bias_t)\mathbb{E}[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

Key Idea 1:
Focus on the ***"good" times***
when bound is non-vacuous

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - A key inequality may be *vacuous:*

$$\underbrace{(1 - \sigma_1 \cdot bias_t)}_{\geq \frac{1}{2}} \mathbb{E}[\eta_t] \|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$$

Key Idea 1:
Focus on the ***"good" times***
when bound is non-vacuous

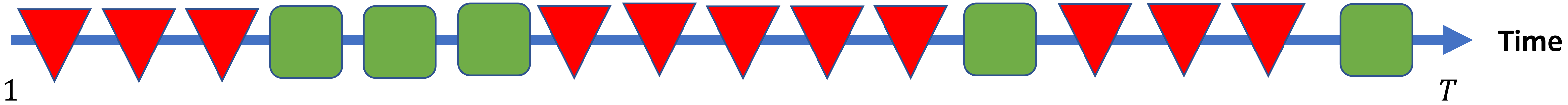
Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - A key inequality may be *vacuous:*

$$\underbrace{(1 - \sigma_1 \cdot bias_t)}_{\geq \frac{1}{2}} \mathbb{E}[\eta_t] \|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$$

Key Idea 1:
Focus on the ***"good" times***
when bound is non-vacuous



Overcoming the challenges of **adaptive** step sizes

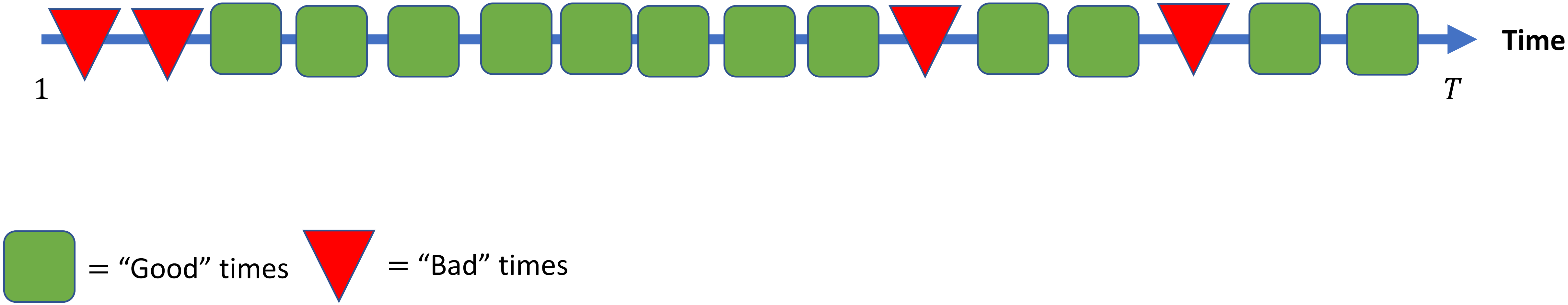
- **Challenge 1**: Bias + affine variance
  - A key inequality may be *vacuous:*

$$(1 - \sigma_1 \cdot bias_t)\mathbb{E}[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$$

$(1 - \sigma_1 \cdot bias_t) \geq \frac{1}{2}$

Key Idea 1:
Focus on the ***"good" times***
when bound is non-vacuous

- Most times are **"good"** (typically)



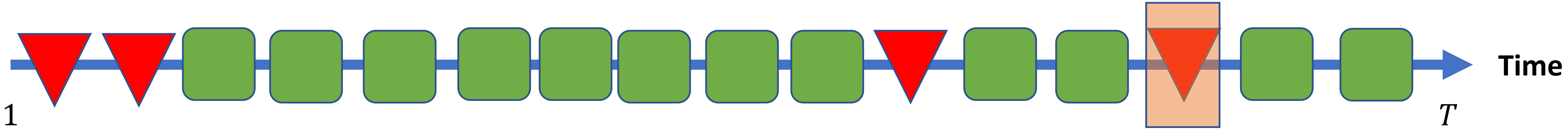
Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - A key inequality may be *vacuous:*

$$\underbrace{(1 - \sigma_1 \cdot bias_t)}_{\geq \frac{1}{2}}\mathbb{E}[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

Key Idea 1:
Focus on the ***"good" times***
when bound is non-vacuous

- Most times are **"good"** (typically)
- Large "bad" times can still ruin analysis



Overcoming the challenges of **adaptive** step sizes

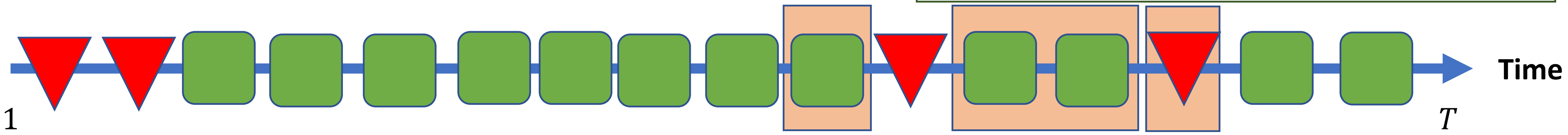
- **Challenge 1**: Bias + affine variance
  - A key inequality may be *vacuous:*

$$(1 - \sigma_1 \cdot bias_t)\mathbb{E}[\eta_t]\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

$$\underbrace{(1 - \sigma_1 \cdot bias_t)}_{\geq \frac{1}{2}}$$

Key Idea 1:
Focus on the ***"good" times***
when bound is non-vacuous

- Most times are **"good"** (typically)
- Large "bad" times can still ruin analysis
- **Compensate** for "bad" times with a few nearby, earlier "good" ones



= "Good" times = "Bad" times

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on $\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$

Start with a crude, polynomial bound
$\sum_t \mathbb{E}[\|\nabla F(w_t)\|^2] \lesssim T^x \log(T)^y$

Obtained via *smoothness* + *unit-step* property of AdaGrad

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on $\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$

Start with a crude, polynomial bound $\sum_t \mathbb{E}[\|\nabla F(w_t)\|^2] \lesssim T^x \log(T)^y$

Bound $\eta_t \gtrsim \frac{1}{\sqrt{T^{x'} \log(T)^{y'}}}$ w.h.p.

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$

Start with a crude, polynomial bound
$\sum_t \mathbb{E}[\|\nabla F(w_t)\|^2] \lesssim T^x \log(T)^y$

Bound $\eta_t \gtrsim \frac{1}{\sqrt{T^{x'} \log(T)^{y'}}}$ w.h.p.

Invariant (from **Challenge 1**):
$\mathbb{E}[\sum_t \eta_t \|\nabla F(w_t)\|^2] \leq F(w_1) - F^* + \text{poly}\log(T)$

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$

Start with a crude, polynomial bound
$\sum_t \mathbb{E}[\|\nabla F(w_t)\|^2] \lesssim T^x \log(T)^y$

Bound $\eta_t \gtrsim \frac{1}{\sqrt{T^{x'} \log(T)^{y'}}}$ w.h.p.

Invariant (from **Challenge 1**):
$\mathbb{E}[\sum_t \eta_t \|\nabla F(w_t)\|^2] \leq F(w_1) - F^* + \text{poly}\log(T)$

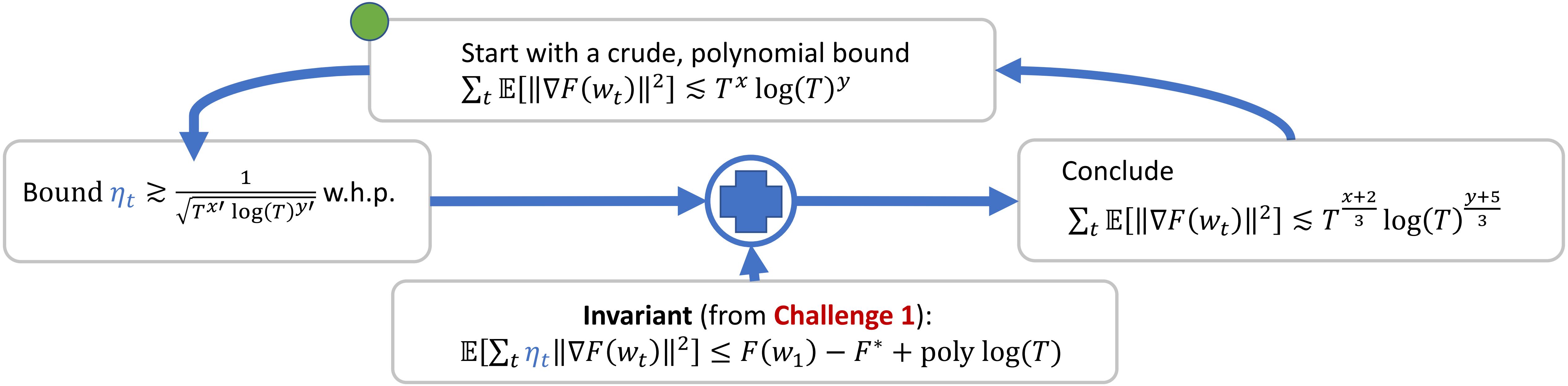
Conclude
$\sum_t \mathbb{E}[\|\nabla F(w_t)\|^2] \lesssim T^{\frac{x+2}{3}} \log(T)^{\frac{y+5}{3}}$

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$

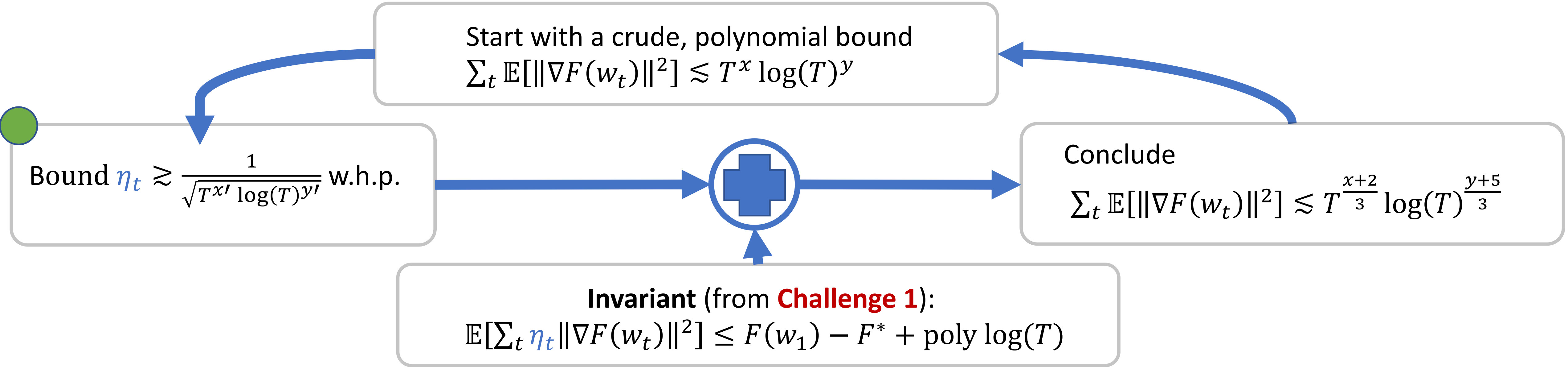


Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

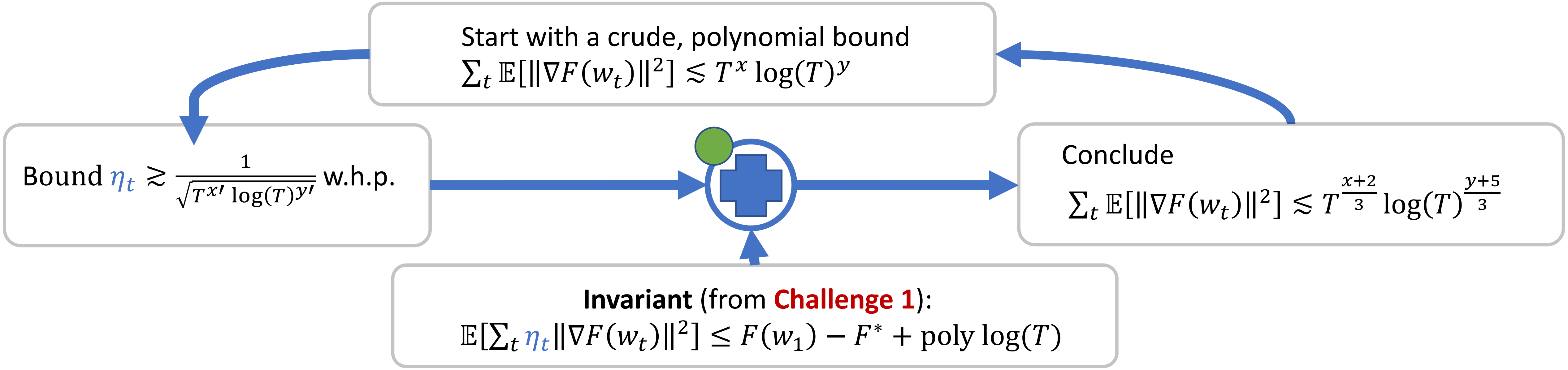
Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$



Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\,\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\,\sum \|\nabla F(w_t)\|^2]$

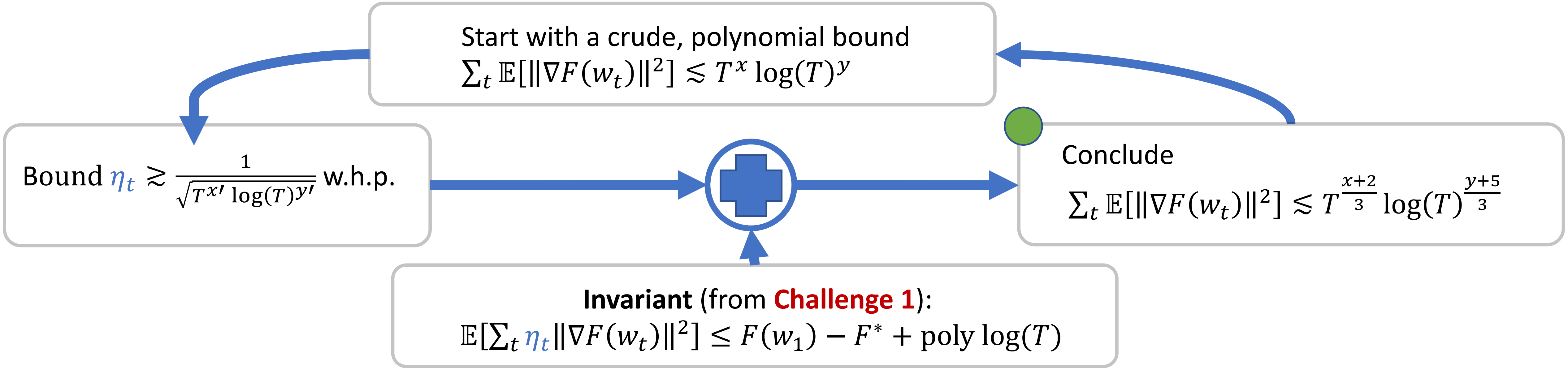


Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

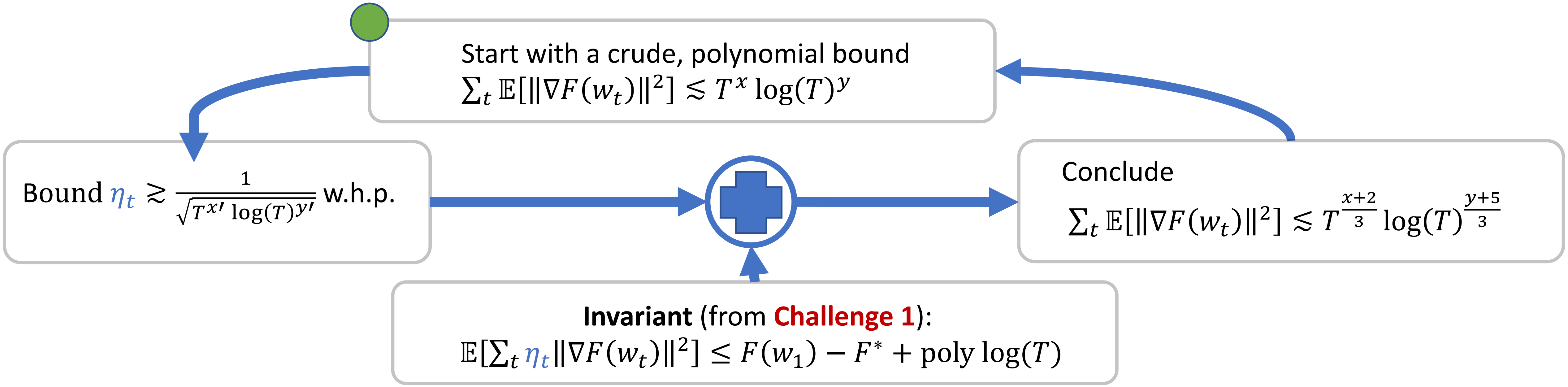
Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$



Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - Cannot guarantee directly that $\eta_t \gtrsim \frac{1}{\sqrt{T}}$ (even in expectation).
  - Crucial step is to bound $\mathbb{E}[\sum_t \|\nabla F(w_t)\|^2] = \tilde{O}(T)$.

Key Idea 2:
Recursively improve crude bound on
$\mathbb{E}[\sum \|\nabla F(w_t)\|^2]$



Best of Both Worlds

Theorem

AdaGrad-Norm enjoys a $\min_t \|\nabla F(w_t)\|^2 = \tilde{O}(1/\sqrt{T})$ convergence rate in the same setting as SGD (smooth + affine).

- ✓ Without a uniform upper bound on the gradients or variance.
- ✓ For any parameter choices $\eta, b_0 > 0$ (no knowledge of L, σ_0 or σ_1 is required).

Best of Both Worlds

Theorem

AdaGrad-Norm enjoys a $\min_t \|\nabla F(w_t)\|^2 = \tilde{O}(1/\sqrt{T})$ convergence rate in the same setting as SGD (smooth + affine).

- ✓ Without a uniform upper bound on the gradients or variance.
- ✓ For any parameter choices $\eta, b_0 > 0$ (no knowledge of L, σ_0 or σ_1 is required).

Remark

- ✓ We show that AdaGrad-Norm automatically obtains a $\tilde{O}(1/T)$ convergence rate in the small noise regime.

Best of Both Worlds

Theorem

AdaGrad-Norm enjoys a $\min_t \|\nabla F(w_t)\|^2 = \tilde{O}(1/\sqrt{T})$ convergence rate in the same setting as SGD (smooth + affine).

- ✓ Without a uniform upper bound on the gradients or variance.
- ✓ For any parameter choices $\eta, b_0 > 0$ (no knowledge of L, σ_0 or σ_1 is required).

Remark

- ✓ We show that AdaGrad-Norm automatically obtains a $\tilde{O}(1/T)$ convergence rate in the small noise regime.

Remark

✓ "Best of both worlds" result!

AdaGrad-Norm achieves order optimal convergence (similar to SGD) without tuning any hyperparameters!

Best of Both Worlds

Theorem

AdaGrad-Norm enjoys a $\min_t \|\nabla F(w_t)\|^2 = \tilde{O}(1/\sqrt{T})$ convergence rate in the same setting as SGD (smooth + affine).

- ✓ Without a uniform upper bound on the gradients or variance.
- ✓ For any parameter choices $\eta, b_0 > 0$ (no knowledge of L, σ_0 or σ_1 is required).

Remark

- ✓ We show that AdaGrad-Norm automatically obtains a $\tilde{O}(1/T)$ convergence rate in the small noise regime.

Remark

✓ "Best of both worlds" result!

AdaGrad-Norm achieves order optimal convergence (similar to SGD) without tuning any hyperparameters!

Thanks for listening!