Stochastic optimization under time drift

iterate averaging, step decay, and high probability guarantees

Joshua Cutler
Mathematics, University of Washington

Joint work with D. Drusvyatskiy (UW) and Z. Harchaoui (UW)

NeurIPS 2021

What this paper is about

Time-varying stochastic optimization:

$$\min_{x} \ \varphi_t(x) := f_t(x) + r_t(x)$$

indexed by time $t \in \mathbb{N}$, where

1. loss $f_t : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex;
2. regularizer $r_t : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and convex;
3. objective φ_t may evolve stochastically in time.

What this paper is about

Time-varying stochastic optimization:

$$\min_x \ \varphi_t(x) := f_t(x) + r_t(x)$$

indexed by time $t \in \mathbb{N}$, where

1. loss $f_t : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex;
2. regularizer $r_t \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and convex;
3. objective φ_t may evolve stochastically in time.

Goal: Track the optimum "as closely as possible" in "shortest amount of time".

▶ We build on extensive literature on the subject: Bartlett et al. '00, Besbes et al. '15, Guo-Ljung '95, Long '99, Madden et al. '21, Wilson et al. '18, . . .

What this paper is about

Time-varying stochastic optimization:

$$\min_{x} \ \varphi_t(x) := f_t(x) + r_t(x)$$

indexed by time $t \in \mathbb{N}$, where

1. loss $f_t : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex;
2. regularizer $r_t : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and convex;
3. objective φ_t may evolve stochastically in time.

Goal: Track the optimum "as closely as possible" in "shortest amount of time".

▶ We build on extensive literature on the subject: Bartlett et al. '00, Besbes et al. '15, Guo-Ljung '95, Long '99, Madden et al. '21, Wilson et al. '18, ...

Online proximal stochastic gradient method:

$$\text{Set } x_{t+1} = \text{prox}_{\eta_t r_t}\big(x_t - \eta_t \widetilde{\nabla} f_t(x_t)\big)$$

where $\widetilde{\nabla} f_t(x_t)$ is an unbiased estimator of $\nabla f_t(x_t)$.

Tracking the minimizer

Drift and noise: Suppose there exist $\Delta, \sigma > 0$ such that

$$\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \leq \Delta^2 \quad \text{and} \quad \mathbb{E}\|\nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)\|^2 \leq \sigma^2.$$

Tracking the minimizer

Drift and noise: Suppose there exist $\Delta, \sigma > 0$ such that

$$\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \le \Delta^2 \quad \text{and} \quad \mathbb{E}\|\nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)\|^2 \le \sigma^2.$$

Error decomposition: using step size $\eta \le 1/2L$ yields

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \underbrace{(1-\mu\eta)^t \cdot \|x_0 - x_0^\star\|^2}_{\text{optimization}} + \underbrace{\frac{\eta\sigma^2}{\mu}}_{\text{noise}} + \underbrace{\left(\frac{\Delta}{\mu\eta}\right)^2}_{\text{drift}}.$$

Tracking the minimizer

Drift and noise: Suppose there exist $\Delta, \sigma > 0$ such that

$$\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \leq \Delta^2 \quad \text{and} \quad \mathbb{E}\|\nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)\|^2 \leq \sigma^2.$$

Error decomposition: using step size $\eta \leq 1/2L$ yields

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \underbrace{(1-\mu\eta)^t \cdot \|x_0 - x_0^\star\|^2}_{\text{optimization}} + \underbrace{\frac{\eta\sigma^2}{\mu}}_{\text{noise}} + \underbrace{\left(\frac{\Delta}{\mu\eta}\right)^2}_{\text{drift}}.$$

Asymptotic error and optimal step size:

$$\mathcal{E} := \min_{\eta \in (0, 1/2L]} \left\{ \frac{\eta\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta}\right)^2 \right\} \quad \text{and} \quad \eta_\star := \min\left\{ \frac{1}{2L}, \left(\frac{2\Delta^2}{\mu\sigma^2}\right)^{1/3} \right\}.$$

Numerical illustration

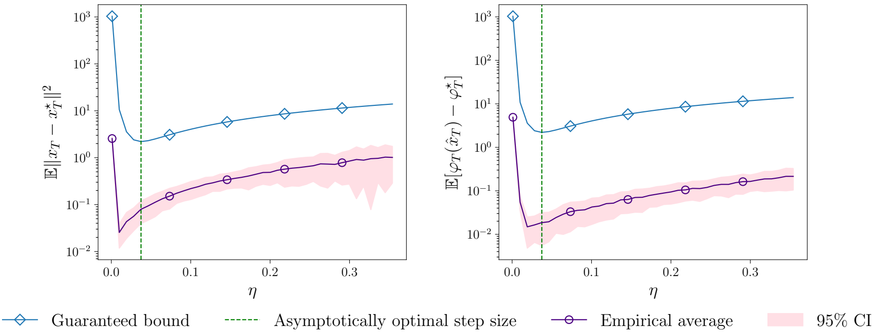

Figure: Semilog plots of guaranteed bounds and empirical tracking errors at horizon T with respect to step size η for logistic regression with stochastically evolving labels.

Two regimes of variation

Asymptotically optimal step size:

$$\eta_\star = \begin{cases} \frac{1}{2L} & \text{if } \frac{\Delta}{\sigma} \geq \sqrt{\frac{\mu}{16L^3}} \\ \left(\frac{2\Delta^2}{\mu\sigma^2}\right)^{1/3} & \text{otherwise.} \end{cases}$$

Two regimes of variation

Asymptotically optimal step size:

$$\eta_\star = \begin{cases} \frac{1}{2L} & \text{if } \frac{\Delta}{\sigma} \geq \sqrt{\frac{\mu}{16L^3}} \\ \left(\frac{2\Delta^2}{\mu\sigma^2}\right)^{1/3} & \text{otherwise.} \end{cases}$$

- The high drift-to-noise regime $\Delta/\sigma \geq \sqrt{\mu/16L^3}$ is uninteresting.

Two regimes of variation

Asymptotically optimal step size:

$$\eta_\star = \begin{cases} \frac{1}{2L} & \text{if } \frac{\Delta}{\sigma} \geq \sqrt{\frac{\mu}{16L^3}} \\ \left(\frac{2\Delta^2}{\mu\sigma^2}\right)^{1/3} & \text{otherwise.} \end{cases}$$

- The high drift-to-noise regime $\Delta/\sigma \geq \sqrt{\mu/16L^3}$ is uninteresting.

Thm (C-Drusvyatskiy-Harchaoui '21): In the low drift-to-noise regime, a step-decay schedule $\{\eta_t\}$ ensures:

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E} \quad \text{after time} \quad t \lesssim \frac{L}{\mu} \log\left(\frac{\|x_0 - x_0^\star\|^2}{\mathcal{E}}\right) + \frac{\sigma^2}{\mu^2 \mathcal{E}}.$$

Two regimes of variation

Asymptotically optimal step size:

$$\eta_\star = \begin{cases} \frac{1}{2L} & \text{if } \frac{\Delta}{\sigma} \geq \sqrt{\frac{\mu}{16L^3}} \\ \left(\frac{2\Delta^2}{\mu\sigma^2}\right)^{1/3} & \text{otherwise.} \end{cases}$$

- The high drift-to-noise regime $\Delta/\sigma \geq \sqrt{\mu/16L^3}$ is uninteresting.

Thm (C-Drusvyatskiy-Harchaoui '21): In the low drift-to-noise regime, a step-decay schedule $\{\eta_t\}$ ensures:

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E} \quad \text{after time} \quad t \lesssim \frac{L}{\mu} \log\left(\frac{\|x_0 - x_0^\star\|^2}{\mathcal{E}}\right) + \frac{\sigma^2}{\mu^2 \mathcal{E}}.$$

- This is analogous to the static setting with $\mathcal{E}$ in place of target accuracy ε.

High probability guarantees

Settings in which an online algorithm can only be executed once call for efficiency estimates that hold with high probability.

High probability guarantees

Settings in which an online algorithm can only be executed once call for efficiency estimates that hold with high probability.

Sub-Gaussian drift and noise: Suppose there exist $\Delta, \sigma > 0$ such that

1. $\|x_t^\star - x_{t+1}^\star\|$ is Δ-sub-Gaussian (conditioned on $\mathcal{F}_t$);
2. $\|\nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)\|$ is σ-sub-Gaussian (conditioned on $\mathcal{F}_t$).

High probability guarantees

Settings in which an online algorithm can only be executed once call for efficiency estimates that hold with high probability.

Sub-Gaussian drift and noise: Suppose there exist $\Delta, \sigma > 0$ such that

1. $\|x_t^\star - x_{t+1}^\star\|$ is Δ-sub-Gaussian (conditioned on $\mathcal{F}_t$);
2. $\|\nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)\|$ is σ-sub-Gaussian (conditioned on $\mathcal{F}_t$).

Thm (C-Drusvyatskiy-Harchaoui '21): For any specified $t \in \mathbb{N}$ and $\delta \in (0,1)$, using step size $\eta \le 1/2L$ yields the following bound with probability at least $1-\delta$:

$$\|x_t - x_t^\star\|^2 \lesssim \left(1 - \frac{\mu\eta}{2}\right)^t \|x_0 - x_0^\star\|^2 + \left(\frac{\eta\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta}\right)^2\right) \log\left(\frac{e}{\delta}\right).$$

High probability guarantees

Settings in which an online algorithm can only be executed once call for efficiency estimates that hold with high probability.

Sub-Gaussian drift and noise: Suppose there exist $\Delta, \sigma > 0$ such that

1. $\|x_t^\star - x_{t+1}^\star\|$ is Δ-sub-Gaussian (conditioned on $\mathcal{F}_t$);
2. $\|\nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)\|$ is σ-sub-Gaussian (conditioned on $\mathcal{F}_t$).

Thm (C-Drusvyatskiy-Harchaoui '21): For any specified $t \in \mathbb{N}$ and $\delta \in (0,1)$, using step size $\eta \leq 1/2L$ yields the following bound with probability at least $1 - \delta$:

$$\|x_t - x_t^\star\|^2 \lesssim \left(1 - \frac{\mu\eta}{2}\right)^t \|x_0 - x_0^\star\|^2 + \left(\frac{\eta\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta}\right)^2\right) \log\left(\frac{e}{\delta}\right).$$

- Proof uses techniques from Harvey et al. '19.

High probability guarantees

Settings in which an online algorithm can only be executed once call for efficiency estimates that hold with high probability.

Sub-Gaussian drift and noise: Suppose there exist $\Delta, \sigma > 0$ such that

1. $\|x_t^\star - x_{t+1}^\star\|$ is Δ-sub-Gaussian (conditioned on $\mathcal{F}_t$);
2. $\|\nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)\|$ is σ-sub-Gaussian (conditioned on $\mathcal{F}_t$).

Thm (C-Drusvyatskiy-Harchaoui '21): For any specified $t \in \mathbb{N}$ and $\delta \in (0, 1)$, using step size $\eta \leq 1/2L$ yields the following bound with probability at least $1 - \delta$:

$$\|x_t - x_t^\star\|^2 \lesssim \left(1 - \frac{\mu\eta}{2}\right)^t \|x_0 - x_0^\star\|^2 + \left(\frac{\eta\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta}\right)^2\right) \log\left(\frac{e}{\delta}\right).$$

- Proof uses techniques from Harvey et al. '19.
- With this result in hand, implementing a step-decay schedule as before yields a high-probability efficiency estimate.

Tracking the minimal value

Using the running average

$$\hat{x}_0 := x_0 \quad \text{and} \quad \hat{x}_{t+1} := \left(1 - \frac{\mu\eta_t}{2-\mu\eta_t}\right)\hat{x}_t + \frac{\mu\eta_t}{2-\mu\eta_t}x_{t+1}$$

of the iterates $\{x_t\}$, we obtain analogous results for tracking the minimal value.

Tracking the minimal value

Using the running average

$$\hat{x}_0 := x_0 \quad \text{and} \quad \hat{x}_{t+1} := \left(1 - \frac{\mu\eta_t}{2 - \mu\eta_t}\right)\hat{x}_t + \frac{\mu\eta_t}{2 - \mu\eta_t} x_{t+1}$$

of the iterates $\{x_t\}$, we obtain analogous results for tracking the minimal value.

Stronger control on drift and noise: Suppose the regularizers $r_t \equiv r$ are identical and there exist $\Delta, \sigma > 0$ such that for all $0 \leq i < t$,

1. the gradient drift $G_{i,t} := \sup_x \|\nabla f_i(x) - \nabla f_t(x)\|$ satisfies

$$\mathbb{E}[G_{i,t}^2] \leq (\mu\Delta|i - t|)^2;$$

2. the gradient noise $z_t := \nabla f_t(x_t) - \widetilde{\nabla} f_t(x_t)$ satisfies

$$\mathbb{E}\|z_t\|^2 \leq \sigma^2 \quad \text{and} \quad \mathbb{E}\langle z_i, x_t^\star \rangle = 0.$$

Tracking the minimal value

Thm (C-Drusvyatskiy-Harchaoui '21): Using step size $\eta \le 1/2L$ yields

$$\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \underbrace{\left(1 - \frac{\mu\eta}{2}\right)^t \cdot (\varphi_0(x_0) - \varphi_0^\star)}_{\text{optimization}} + \underbrace{\eta\sigma^2}_{\text{noise}} + \underbrace{\frac{\Delta^2}{\mu\eta^2}}_{\text{drift}}.$$

Tracking the minimal value

Thm (C-Drusvyatskiy-Harchaoui '21): Using step size $\eta \leq 1/2L$ yields

$$\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \underbrace{\left(1 - \frac{\mu\eta}{2}\right)^t \cdot (\varphi_0(x_0) - \varphi_0^\star)}_{\text{optimization}} + \underbrace{\eta\sigma^2}_{\text{noise}} + \underbrace{\frac{\Delta^2}{\mu\eta^2}}_{\text{drift}}.$$

Asymptotic error and optimal step size: $\mathcal{G} := \mu\mathcal{E}$ and same $\eta_\star$ as before.

Tracking the minimal value

Thm (C-Drusvyatskiy-Harchaoui '21): Using step size $\eta \leq 1/2L$ yields

$$\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \underbrace{\left(1 - \frac{\mu\eta}{2}\right)^t \cdot (\varphi_0(x_0) - \varphi_0^\star)}_{\text{optimization}} + \underbrace{\eta\sigma^2}_{\text{noise}} + \underbrace{\frac{\Delta^2}{\mu\eta^2}}_{\text{drift}}.$$

Asymptotic error and optimal step size: $\mathcal{G} := \mu\mathcal{E}$ and same $\eta_\star$ as before.

Thm (C-Drusvyatskiy-Harchaoui '21): In the low drift-to-noise regime, a step-decay schedule $\{\eta_t\}$ ensures:

$$\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \mathcal{G} \quad \text{after time} \quad t \lesssim \frac{L}{\mu} \log\left(\frac{\varphi_0(x_0) - \varphi_0^\star}{\mathcal{G}}\right) + \frac{\sigma^2}{\mu\mathcal{G}}.$$

Tracking the minimal value

Thm (C-Drusvyatskiy-Harchaoui '21): Using step size $\eta \leq 1/2L$ yields

$$\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \underbrace{\left(1 - \frac{\mu\eta}{2}\right)^t \cdot (\varphi_0(x_0) - \varphi_0^\star)}_{\text{optimization}} + \underbrace{\eta\sigma^2}_{\text{noise}} + \underbrace{\frac{\Delta^2}{\mu\eta^2}}_{\text{drift}}.$$

Asymptotic error and optimal step size: $\mathcal{G} := \mu\mathcal{E}$ and same $\eta_\star$ as before.

Thm (C-Drusvyatskiy-Harchaoui '21): In the low drift-to-noise regime, a step-decay schedule $\{\eta_t\}$ ensures:

$$\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \mathcal{G} \quad \text{after time} \quad t \lesssim \frac{L}{\mu}\log\left(\frac{\varphi_0(x_0) - \varphi_0^\star}{\mathcal{G}}\right) + \frac{\sigma^2}{\mu\mathcal{G}}.$$

- Under light-tail assumptions, analogous guarantees hold with high probability. Caveat: analysis is more complicated than for distance tracking.

Thank you!

Further details are in the paper:

- “Stochastic optimization under time drift: iterate averaging, step decay, and high probability guarantees”, https://arxiv.org/abs/2108.07356.