Armijo Line-search Can Make (Stochastic) Gradient Descent Provably Faster

Sharan Vaswani (Simon Fraser University)
Joint work with: Reza Babanezhad (Samsung AI, Montreal)

ICML 2025

Introduction

Armijo line-search (Armijo-LS) [Armijo, 1966] is a standard method to set the step-size for gradient descent (GD).

- For uniformly L-smooth functions (for which $\left\|\nabla^2 f(\theta)\right\| \leq L$), Armijo-LS
  - Alleviates the need to know the global smoothness constant L.
  - Enables GD to adapt to the "local" smoothness and typically results in faster empirical convergence.

Introduction

Armijo line-search (Armijo-LS) [Armijo, 1966] is a standard method to set the step-size for gradient descent (GD).

- For uniformly L-smooth functions (for which $\|\nabla^2 f(\theta)\| \leq L$), Armijo-LS
  - Alleviates the need to know the global smoothness constant L.
  - Enables GD to adapt to the "local" smoothness and typically results in faster empirical convergence.
- Previous work [Scheinberg et al., 2014, Lu and Mei, 2023, Fox and Schmidt]
  - ✓ Propose different notions of local smoothness to formalize this intuition, and theoretically characterize the benefit of `GD-LS` over `GD(1/L)`.
  - × Only show that `GD-LS` can result in constant factor improvements over `GD(1/L)`.

Introduction

Armijo line-search (Armijo-LS) [Armijo, 1966] is a standard method to set the step-size for gradient descent (GD).

- For uniformly L-smooth functions (for which $\left\|\nabla^2 f(\theta)\right\| \leq L$), Armijo-LS
  - Alleviates the need to know the global smoothness constant L.
  - Enables GD to adapt to the "local" smoothness and typically results in faster empirical convergence.
- Previous work [Scheinberg et al., 2014, Lu and Mei, 2023, Fox and Schmidt]
  - ✓ Propose different notions of local smoothness to formalize this intuition, and theoretically characterize the benefit of `GD-LS` over `GD(1/L)`.
  - × Only show that `GD-LS` can result in constant factor improvements over `GD(1/L)`.
- **This paper**: Considers a class of non-uniform smooth objective functions and show that `GD-LS` can result in a provably faster rate of convergence compared to `GD(1/L)`.

Problem Formulation

Objective: $\min_{\theta \in \mathbb{R}^d} f(\theta)$ such that f satisfies the following assumptions:

(A1) f is non-negative and twice-differentiable.

Problem Formulation

Objective: $\min_{\theta \in \mathbb{R}^d} f(\theta)$ such that f satisfies the following assumptions:

(A1) f is non-negative and twice-differentiable.

(A2) (L_0, L_1) non-uniform smooth, i.e.,

- For all θ, $\left\|\nabla^2 f(\theta)\right\| \leq L_0 + L_1 f(\theta)$.
- For all x, y s.t. $\|x - y\| \leq q/L_1$, where $q \geq 1$ is a constant, if $A := 1 + e^q - \frac{e^q - 1}{q}$, $B := \frac{e^q - 1}{q}$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{(A L_0 + B L_1 f(x))}{2} \|y - x\|_2^2$$

Problem Formulation

Objective: $\min_{\theta \in \mathbb{R}^d} f(\theta)$ such that f satisfies the following assumptions:

(A1) f is non-negative and twice-differentiable.

(A2) (L_0, L_1) non-uniform smooth, i.e.,

- For all θ, $\left\|\nabla^2 f(\theta)\right\| \leq L_0 + L_1 f(\theta)$.
- For all x, y s.t. $\|x - y\| \leq q/L_1$, where $q \geq 1$ is a constant, if $A := 1 + e^q - \frac{e^q - 1}{q}$, $B := \frac{e^q - 1}{q}$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{(A L_0 + B L_1 f(x))}{2} \|y - x\|_2^2$$

(A3) There exists constants $\omega, \nu \geq 0$ s.t. for all θ, $\|\nabla f(\theta)\| \leq \nu f(\theta) + \omega$.

Problem Formulation

Objective: $\min_{\theta \in \mathbb{R}^d} f(\theta)$ such that f satisfies the following assumptions:

(A1) f is non-negative and twice-differentiable.

(A2) (L_0, L_1) non-uniform smooth, i.e.,

- For all θ, $\left\|\nabla^2 f(\theta)\right\| \leq L_0 + L_1 f(\theta)$.
- For all x, y s.t. $\|x - y\| \leq q/L_1$, where $q \geq 1$ is a constant, if $A := 1 + e^q - \frac{e^q - 1}{q}$, $B := \frac{e^q - 1}{q}$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{(A\, L_0 + B\, L_1\, f(x))}{2} \|y - x\|_2^2$$

(A3) There exists constants $\omega, \nu \geq 0$ s.t. for all θ, $\|\nabla f(\theta)\| \leq \nu\, f(\theta) + \omega$.

Examples

- Logistic regression satisfies (A1)-(A3) with $L_0 = 0$, $L_1 = 8$, $\nu = 8$, $\omega = 0$.
- Generalized linear model with a logistic link function satisfies (A1)-(A3) with $L_0 = 9/16$, $L_1 = 9$, $\nu = 9$, $\omega = 1$.
- Softmax policy gradient objective for multi-armed bandits satisfies (A1)-(A3) with $L_0 = 0$, $L_1 = 72$, $\nu = 24$, $\omega = 0$.
- **Others**: Linear multi-class classification with the cross-entropy loss, 2 layer neural networks with the exponential loss, Softmax policy gradient for tabular MDPs.

Algorithm

Algorithm: At iteration t of GD, use back-tracking to choose the (approximately) “largest” step-size that satisfies the Armijo condition: $f(\theta_t - \eta_t \nabla f(\theta_t)) \leq f(\theta_t) - c\eta_t \left\|\nabla f(\theta_t)\right\|_2^2$.

Algorithm 1 GD with Armijo Line-search (GD−LS)

```
Input: θ_0, η_max, c ∈ (0, 1), β ∈ (0, 1)
for t = 0, . . . , T − 1 do
    η̃_t ← η_max
    while f(θ_t − η̃_t ∇f(θ_t)) > f(θ_t) − c η̃_t ‖∇f(θ_t)‖_2^2 do
        η̃_t ← η̃_t β
    end while
    η_t ← η̃_t
    θ_{t+1} = θ_t − η_t ∇f(θ_t)
end for
```

Algorithm

Algorithm: At iteration t of GD, use back-tracking to choose the (approximately) "largest" step-size that satisfies the Armijo condition: $f(\theta_t - \eta_t \nabla f(\theta_t)) \leq f(\theta_t) - c\eta_t \|\nabla f(\theta_t)\|_2^2$.

Algorithm 1 GD with Armijo Line-search (GD-LS)

1: **Input:** θ_0, $\eta_{\max}$, $c \in (0,1)$, $\beta \in (0,1)$
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad \tilde{\eta}_t \leftarrow \eta_{\max}$
4: $\quad$ **while** $f(\theta_t - \tilde{\eta}_t \nabla f(\theta_t)) > f(\theta_t) - c\,\tilde{\eta}_t \|\nabla f(\theta_t)\|_2^2$ **do**
5: $\quad\quad \tilde{\eta}_t \leftarrow \tilde{\eta}_t\, \beta$
6: $\quad$ **end while**
7: $\quad \eta_t \leftarrow \tilde{\eta}_t$
8: $\quad \theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t)$
9: **end for**

Lemma: If f satisfies (A1)-(A3), then, at iteration t, GD-LS (with "exact" backtracking) returns a step-size η_t s.t.

$$\eta_t \geq \min\left\{\eta_{\max}, \frac{1}{\lambda_0 + \lambda_1 f(\theta_t)}\right\},$$

where $\lambda_0 := 3\,\frac{L_0 + L_1\,\omega}{(1-c)}$ and $\lambda_1 := 3\,\frac{L_1(\nu+1)}{(1-c)}$.

Algorithm: At iteration t of GD, use back-tracking to choose the (approximately) "largest" step-size that satisfies the Armijo condition: $f(\theta_t - \eta_t \nabla f(\theta_t)) \leq f(\theta_t) - c\eta_t \|\nabla f(\theta_t)\|_2^2$.

Algorithm 1 GD with Armijo Line-search (GD−LS)

1: **Input:** $\theta_0, \eta_{\max}, c \in (0,1), \beta \in (0,1)$
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad \tilde{\eta}_t \leftarrow \eta_{\max}$
4: $\quad$ **while** $f(\theta_t - \tilde{\eta}_t \nabla f(\theta_t)) > f(\theta_t) - c\,\tilde{\eta}_t \|\nabla f(\theta_t)\|_2^2$ **do**
5: $\quad\quad \tilde{\eta}_t \leftarrow \tilde{\eta}_t \beta$
6: $\quad$ **end while**
7: $\quad \eta_t \leftarrow \tilde{\eta}_t$
8: $\quad \theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t)$
9: **end for**

Lemma: If f satisfies (A1)-(A3), then, at iteration t, GD−LS (with "exact" backtracking) returns a step-size η_t s.t.

$$\eta_t \geq \min\left\{\eta_{\max}, \frac{1}{\lambda_0 + \lambda_1 f(\theta_t)}\right\},$$

where $\lambda_0 := 3\,\frac{L_0 + L_1\,\omega}{(1-c)}$ and $\lambda_1 := 3\,\frac{L_1(\nu+1)}{(1-c)}$.

- Hence, for functions satisfying (A1)-(A3), given a large $\eta_{\max}$, η_t increases as $f(\theta_t)$ decreases, and consequently, GD−LS results in faster convergence.

Theoretical Results – Meta Theorem

Theorem: For a fixed $\epsilon > 0$, if f satisfies (A1)-(A3), and if for a constant $R > 0$, $\|\nabla f(\theta_t)\|_2^2 \geq \frac{[f(\theta_t) - f^*]^2}{R}$ for all iterations $t \in [T]$, then, `GD-LS` with $\eta_{\max} = \infty$ requires

$$T \geq \begin{cases} \max\{2\,R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right) \text{ if } f^* \geq \frac{\lambda_0}{\lambda_1} - \epsilon & \textbf{(Case (1))} \\ \frac{2\lambda_0\,R}{\epsilon} + \max\{2\,R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right) \text{ otherwise} & \textbf{(Case (2))} \end{cases}$$

iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.

Theoretical Results – Meta Theorem

Theorem: For a fixed $\epsilon > 0$, if f satisfies (A1)-(A3), and if for a constant $R > 0$, $\|\nabla f(\theta_t)\|_2^2 \geq \frac{[f(\theta_t) - f^*]^2}{R}$ for all iterations $t \in [T]$, then, `GD-LS` with $\eta_{\text{max}} = \infty$ requires

$$T \geq \begin{cases} \max\{2\,R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right) \text{ if } f^* \geq \frac{\lambda_0}{\lambda_1} - \epsilon & \textbf{(Case (1))} \\ \frac{2\lambda_0\,R}{\epsilon} + \max\{2\,R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right) \text{ otherwise} & \textbf{(Case (2))} \end{cases}$$

iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.

- If $L_1 = 0 \implies \lambda_1 = 0$, `GD-LS` converges at an $O(1/\epsilon)$ rate matching the `GD(1/L)` rate for uniformly-smooth functions.

Theoretical Results – Meta Theorem

Theorem: For a fixed $\epsilon > 0$, if f satisfies (A1)-(A3), and if for a constant $R > 0$, $\|\nabla f(\theta_t)\|_2^2 \geq \frac{[f(\theta_t) - f^*]^2}{R}$ for all iterations $t \in [T]$, then, GD-LS with $\eta_{\max} = \infty$ requires

$$T \geq \begin{cases} \max\{2R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right) & \text{if } f^* \geq \frac{\lambda_0}{\lambda_1} - \epsilon \quad \textbf{(Case (1))} \\ \frac{2\lambda_0 R}{\epsilon} + \max\{2R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right) & \text{otherwise} \quad \textbf{(Case (2))} \end{cases}$$

iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.

- If $L_1 = 0 \implies \lambda_1 = 0$, GD-LS converges at an $O(1/\epsilon)$ rate matching the GD(1/L) rate for uniformly-smooth functions.
- If $L_0 = 0, \omega = 0 \implies \lambda_0 = 0$, GD-LS converges at an $O\left(R\left(\frac{f^*}{\epsilon}\right) \ln\left(\frac{1}{\epsilon}\right)\right)$ rate. If $\epsilon = \Theta(f^*)$, this implies a faster $O\left(R \ln\left(\frac{1}{\epsilon}\right)\right)$ compared to the $O(1/\epsilon)$ rate for GD(1/L).

Theoretical Results – Meta Theorem

Theorem: For a fixed $\epsilon > 0$, if f satisfies (A1)-(A3), and if for a constant $R > 0$, $\|\nabla f(\theta_t)\|_2^2 \geq \frac{[f(\theta_t)-f^*]^2}{R}$ for all iterations $t \in [T]$, then, GD-LS with $\eta_{\max} = \infty$ requires

$$T \geq \begin{cases} \max\{2R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0)-f^*}{\epsilon}\right) & \text{if } f^* \geq \frac{\lambda_0}{\lambda_1} - \epsilon \quad \textbf{(Case (1))} \\ \frac{2\lambda_0 R}{\epsilon} + \max\{2R\lambda_1, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0)-f^*}{\epsilon}\right) & \text{otherwise} \quad \textbf{(Case (2))} \end{cases}$$

iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.

- If $L_1 = 0 \implies \lambda_1 = 0$, GD-LS converges at an $O(1/\epsilon)$ rate matching the GD(1/L) rate for uniformly-smooth functions.
- If $L_0 = 0, \omega = 0 \implies \lambda_0 = 0$, GD-LS converges at an $O\left(R\left(\frac{f^*}{\epsilon}\right)\ln\left(\frac{1}{\epsilon}\right)\right)$ rate. If $\epsilon = \Theta(f^*)$, this implies a faster $O\left(R\ln\left(\frac{1}{\epsilon}\right)\right)$ compared to the $O(1/\epsilon)$ rate for GD(1/L).
- In general, if $\lambda_0, \lambda_1 \neq 0$, GD-LS has a two-phase behaviour, fast convergence until the loss becomes smaller than the threshold $(\frac{\lambda_0}{\lambda_1})$, followed by slower convergence to the minimizer.

Theoretical Results – Convex Losses

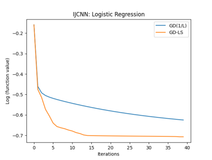
Examples: Logistic regression, multi-class classification with the cross-entropy loss.

.

Theoretical Results – Convex Losses

Examples: Logistic regression, multi-class classification with the cross-entropy loss.

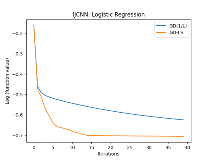
Corollary: For a fixed $\epsilon > 0$, assuming $f(\theta)$ is convex and satisfies (A1)-(A3) with $L_0 = 0$ and $\omega = 0$, GD-LS with $\eta_{\max} = \infty$, requires $T \geq \max\{2\lambda_1 \left\|\theta_0 - \theta^*\right\|_2^2, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right)$ iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.



Theoretical Results – Convex Losses

Examples: Logistic regression, multi-class classification with the cross-entropy loss.

Corollary: For a fixed $\epsilon > 0$, assuming $f(\theta)$ is convex and satisfies (A1)-(A3) with $L_0 = 0$ and $\omega = 0$, GD-LS with $\eta_{\max} = \infty$, requires $T \geq \max\{2\lambda_1 \|\theta_0 - \theta^*\|_2^2, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right)$ iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.

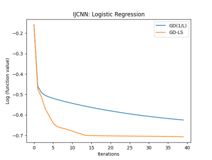


- Matches the rate of normalized gradient descent [Axiotis and Sviridenko, 2023].

Theoretical Results – Convex Losses

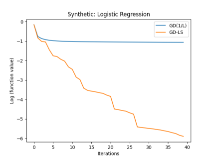
Examples: Logistic regression, multi-class classification with the cross-entropy loss.

Corollary: For a fixed $\epsilon > 0$, assuming $f(\theta)$ is convex and satisfies (A1)-(A3) with $L_0 = 0$ and $\omega = 0$, GD-LS with $\eta_{\text{max}} = \infty$, requires $T \geq \max\{2\lambda_1 \, \|\theta_0 - \theta^*\|_2^2, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right)$ iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.



- Matches the rate of normalized gradient descent [Axiotis and Sviridenko, 2023].

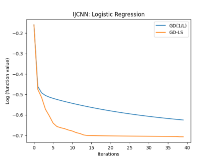
Corollary: For logistic regression on linearly separable data with margin γ, if, for all i, $\|x_i\| \leq 1$, for an initialization θ_0, an $\epsilon \in (0, f(\theta_0))$, GD-LS with $\eta_{\text{max}} = \infty$ requires $T \geq O\left(\frac{1}{\gamma^2} \left[\ln\left(\frac{1}{\epsilon}\right)\right]^2\right)$ iterations to ensure that $f(\theta_T) \leq 2\,\epsilon$.



Theoretical Results – Convex Losses

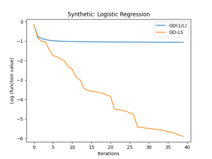
Examples: Logistic regression, multi-class classification with the cross-entropy loss.

Corollary: For a fixed $\epsilon > 0$, assuming $f(\theta)$ is convex and satisfies (A1)-(A3) with $L_0 = 0$ and $\omega = 0$, GD-LS with $\eta_{\max} = \infty$, requires $T \geq \max\{2\lambda_1 \, \|\theta_0 - \theta^*\|_2^2 \, , 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right)$ iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.



- Matches the rate of normalized gradient descent [Axiotis and Sviridenko, 2023].

Corollary: For logistic regression on linearly separable data with margin γ, if, for all i, $\|x_i\| \leq 1$, for an initialization θ_0, an $\epsilon \in (0, f(\theta_0))$, GD-LS with $\eta_{\max} = \infty$ requires $T \geq O\left(\frac{1}{\gamma^2} \, \left[\ln\left(\frac{1}{\epsilon}\right)\right]^2\right)$ iterations to ensure that $f(\theta_T) \leq 2\,\epsilon$.

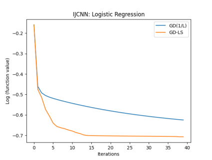


- GD(1/L) cannot have a convergence faster than $\Omega(1/\epsilon)$ [Wu et al., 2024].

Theoretical Results – Convex Losses

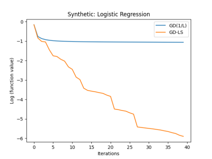
Examples: Logistic regression, multi-class classification with the cross-entropy loss.

Corollary: For a fixed $\epsilon > 0$, assuming $f(\theta)$ is convex and satisfies (A1)-(A3) with $L_0 = 0$ and $\omega = 0$, GD-LS with $\eta_{\max} = \infty$, requires $T \geq \max\{2\lambda_1 \, \|\theta_0 - \theta^*\|_2^2, 1\} \left(\frac{f^*}{\epsilon} + 1\right) \ln\left(\frac{f(\theta_0) - f^*}{\epsilon}\right)$ iterations to ensure that $f(\theta_T) - f^* \leq \epsilon$.



- Matches the rate of normalized gradient descent [Axiotis and Sviridenko, 2023].

Corollary: For logistic regression on linearly separable data with margin γ, if, for all i, $\|x_i\| \leq 1$, for an initialization θ_0, an $\epsilon \in (0, f(\theta_0))$, GD-LS with $\eta_{\max} = \infty$ requires $T \geq O\left(\frac{1}{\gamma^2} \, \left[\ln\left(\frac{1}{\epsilon}\right)\right]^2\right)$ iterations to ensure that $f(\theta_T) \leq 2\,\epsilon$.



- GD(1/L) cannot have a convergence faster than $\Omega(1/\epsilon)$ [Wu et al., 2024].
- Additional result: GD with the Polyak step-size can match the linear convergence of GD-LS.

Theoretical Results – Non-Convex Losses

Multi-armed Bandits (Exact Setting)

Corollary: Given an MAB problem with K arms and known deterministic rewards $r \in [0,1]^K$, consider the class of softmax policies $\pi_\theta \in \Delta_K$ parameterized by $\theta \in \mathbb{R}^K$ s.t. $\pi_\theta(a) \propto \exp(\theta(a))$ and the softmax policy gradient objective: $f(\theta) := r(a^*) - \langle \pi_\theta, r \rangle$, where $a^* := \arg\max_{a \in [K]} r(a)$.

Theoretical Results – Non-Convex Losses

Multi-armed Bandits (Exact Setting)

Corollary: Given an MAB problem with K arms and known deterministic rewards $r \in [0,1]^K$, consider the class of softmax policies $\pi_\theta \in \Delta_K$ parameterized by $\theta \in \mathbb{R}^K$ s.t. $\pi_\theta(a) \propto \exp(\theta(a))$ and the softmax policy gradient objective: $f(\theta) := r(a^*) - \langle \pi_\theta, r \rangle$, where $a^* := \arg\max_{a \in [K]} r(a)$. GD-LS with a uniform initialization i.e. $\forall a$, $\pi_{\theta_0}(a) = 1/K$, $c = \frac{1}{2}$, $\eta_{\max} = \infty$ requires $T \geq O\left(K^2 \ln\left(1/\epsilon\right)\right)$ iterations to guarantee $\langle \pi_{\theta_T}, r \rangle \geq r(a^*) - \epsilon$.

Theoretical Results – Non-Convex Losses

Multi-armed Bandits (Exact Setting)

Corollary: Given an MAB problem with K arms and known deterministic rewards $r \in [0,1]^K$, consider the class of softmax policies $\pi_\theta \in \Delta_K$ parameterized by $\theta \in \mathbb{R}^K$ s.t. $\pi_\theta(a) \propto \exp(\theta(a))$ and the softmax policy gradient objective: $f(\theta) := r(a^*) - \langle \pi_\theta, r \rangle$, where $a^* := \arg\max_{a \in [K]} r(a)$. `GD-LS` with a uniform initialization i.e. $\forall a$, $\pi_{\theta_0}(a) = 1/K$, $c = \frac{1}{2}$, $\eta_{\max} = \infty$ requires $T \geq O\left(K^2 \ln\left(1/\epsilon\right)\right)$ iterations to guarantee $\langle \pi_{\theta_T}, r \rangle \geq r(a^*) - \epsilon$.

- Above linear rate is provably better than the $\Omega(1/\epsilon)$ rate of `GD(1/L)` [Mei et al., 2020].

Multi-armed Bandits (Exact Setting)

Corollary: Given an MAB problem with K arms and known deterministic rewards $r \in [0,1]^K$, consider the class of softmax policies $\pi_\theta \in \Delta_K$ parameterized by $\theta \in \mathbb{R}^K$ s.t. $\pi_\theta(a) \propto \exp(\theta(a))$ and the softmax policy gradient objective: $f(\theta) := r(a^*) - \langle \pi_\theta, r \rangle$, where $a^* := \arg\max_{a \in [K]} r(a)$. GD-LS with a uniform initialization i.e. $\forall a,\ \pi_{\theta_0}(a) = {}^1/_K$, $c = \frac{1}{2}$, $\eta_{\max} = \infty$ requires $T \geq O\left(K^2 \ln\left({}^1/_\epsilon\right)\right)$ iterations to guarantee $\langle \pi_{\theta_T}, r \rangle \geq r(a^*) - \epsilon$.

- Above linear rate is provably better than the $\Omega(1/\epsilon)$ rate of GD(1/L) [Mei et al., 2020].
- GD-LS can match the convergence rate of specialized algorithms (natural policy gradient, normalized GD, GD with increasing step-sizes) for the softmax policy gradient objective.

Theoretical Results – Non-Convex Losses

Multi-armed Bandits (Exact Setting)

Corollary: Given an MAB problem with K arms and known deterministic rewards $r \in [0,1]^K$, consider the class of softmax policies $\pi_\theta \in \Delta_K$ parameterized by $\theta \in \mathbb{R}^K$ s.t. $\pi_\theta(a) \propto \exp(\theta(a))$ and the softmax policy gradient objective: $f(\theta) := r(a^*) - \langle \pi_\theta, r \rangle$, where $a^* := \arg\max_{a \in [K]} r(a)$. `GD-LS` with a uniform initialization i.e. $\forall a$, $\pi_{\theta_0}(a) = {}^1/_K$, $c = \frac{1}{2}$, $\eta_{\max} = \infty$ requires $T \geq O\left(K^2 \ln\left({}^1/_\epsilon\right)\right)$ iterations to guarantee $\langle \pi_{\theta_T}, r \rangle \geq r(a^*) - \epsilon$.

- Above linear rate is provably better than the $\Omega(1/\epsilon)$ rate of `GD(1/L)` [Mei et al., 2020].
- `GD-LS` can match the convergence rate of specialized algorithms (natural policy gradient, normalized GD, GD with increasing step-sizes) for the softmax policy gradient objective.
- Under additional assumptions, similar linear rate holds for tabular MDPs.

Theoretical Results – Non-Convex Losses

Multi-armed Bandits (Exact Setting)

Corollary: Given an MAB problem with K arms and known deterministic rewards $r \in [0,1]^K$, consider the class of softmax policies $\pi_\theta \in \Delta_K$ parameterized by $\theta \in \mathbb{R}^K$ s.t. $\pi_\theta(a) \propto \exp(\theta(a))$ and the softmax policy gradient objective: $f(\theta) := r(a^*) - \langle \pi_\theta, r \rangle$, where $a^* := \arg\max_{a \in [K]} r(a)$. `GD-LS` with a uniform initialization i.e. $\forall a$, $\pi_{\theta_0}(a) = {}^1/_K$, $c = \frac{1}{2}$, $\eta_{\max} = \infty$ requires $T \geq O\left(K^2 \ln\left({}^1/_\epsilon\right)\right)$ iterations to guarantee $\langle \pi_{\theta_T}, r \rangle \geq r(a^*) - \epsilon$.

- Above linear rate is provably better than the $\Omega(1/\epsilon)$ rate of `GD(1/L)` [Mei et al., 2020].
- `GD-LS` can match the convergence rate of specialized algorithms (natural policy gradient, normalized GD, GD with increasing step-sizes) for the softmax policy gradient objective.
- Under additional assumptions, similar linear rate holds for tabular MDPs.

Additional Results:

- For generalized linear model with a logistic link, `GD-LS` has a convergence rate better than or equal to `GD(1/L)` and variants of normalized GD [Mei et al., 2021, Hazan et al., 2015].

Theoretical Results – Non-Convex Losses

Multi-armed Bandits (Exact Setting)

Corollary: Given an MAB problem with K arms and known deterministic rewards $r \in [0,1]^K$, consider the class of softmax policies $\pi_\theta \in \Delta_K$ parameterized by $\theta \in \mathbb{R}^K$ s.t. $\pi_\theta(a) \propto \exp(\theta(a))$ and the softmax policy gradient objective: $f(\theta) := r(a^*) - \langle \pi_\theta, r \rangle$, where $a^* := \arg\max_{a \in [K]} r(a)$. `GD-LS` with a uniform initialization i.e. $\forall a$, $\pi_{\theta_0}(a) = 1/K$, $c = \frac{1}{2}$, $\eta_{\max} = \infty$ requires $T \geq O\left(K^2 \ln(1/\epsilon)\right)$ iterations to guarantee $\langle \pi_{\theta_T}, r \rangle \geq r(a^*) - \epsilon$.

- Above linear rate is provably better than the $\Omega(1/\epsilon)$ rate of `GD(1/L)` [Mei et al., 2020].
- `GD-LS` can match the convergence rate of specialized algorithms (natural policy gradient, normalized GD, GD with increasing step-sizes) for the softmax policy gradient objective.
- Under additional assumptions, similar linear rate holds for tabular MDPs.

Additional Results:

- For generalized linear model with a logistic link, `GD-LS` has a convergence rate better than or equal to `GD(1/L)` and variants of normalized GD [Mei et al., 2021, Hazan et al., 2015].
- For two layer neural networks, when minimizing the exponential loss, `GD-LS` can match the linear convergence rate of normalized GD [Taheri and Thrampoulidis, 2023].

Conclusion

- For specific problems in machine learning including convex losses (logistic regression, linear multi-class classification) and non-convex losses (softmax policy gradient, generalized linear models), `GD-LS` can
  - either match or provably improve upon the sublinear rate of `GD(1/L)`,
  - do so without relying on the knowledge of problem-dependent constants,
  - match the fast convergence of algorithms tailored for these problems.

Conclusion

- For specific problems in machine learning including convex losses (logistic regression, linear multi-class classification) and non-convex losses (softmax policy gradient, generalized linear models), `GD-LS` can
  - either match or provably improve upon the sublinear rate of `GD(1/L)`,
  - do so without relying on the knowledge of problem-dependent constants,
  - match the fast convergence of algorithms tailored for these problems.
- **Additional results**:
  - For logistic regression on separable data, SGD with a stochastic line-search [Vaswani et al., 2019] can match the fast linear convergence of `GD-LS`.

Conclusion

- For specific problems in machine learning including convex losses (logistic regression, linear multi-class classification) and non-convex losses (softmax policy gradient, generalized linear models), `GD-LS` can
  - either match or provably improve upon the sublinear rate of `GD(1/L)`,
  - do so without relying on the knowledge of problem-dependent constants,
  - match the fast convergence of algorithms tailored for these problems.
- **Additional results**:
  - For logistic regression on separable data, SGD with a stochastic line-search [Vaswani et al., 2019] can match the fast linear convergence of `GD-LS`.
  - The non-uniform smoothness assumption in Zhang et al. [2019] implies (A1)-(A3), and hence our results also apply to this class of non-uniform smooth functions. This reduction implies that `GD-LS` can match the convergence of adaptive methods [Vankov et al., 2024, Gorbunov et al., 2024] for this class of non-uniform smooth functions.

Conclusion

- For specific problems in machine learning including convex losses (logistic regression, linear multi-class classification) and non-convex losses (softmax policy gradient, generalized linear models), `GD-LS` can
  - either match or provably improve upon the sublinear rate of `GD(1/L)`,
  - do so without relying on the knowledge of problem-dependent constants,
  - match the fast convergence of algorithms tailored for these problems.
- **Additional results**:
  - For logistic regression on separable data, SGD with a stochastic line-search [Vaswani et al., 2019] can match the fast linear convergence of `GD-LS`.
  - The non-uniform smoothness assumption in Zhang et al. [2019] implies (A1)-(A3), and hence our results also apply to this class of non-uniform smooth functions. This reduction implies that `GD-LS` can match the convergence of adaptive methods [Vankov et al., 2024, Gorbunov et al., 2024] for this class of non-uniform smooth functions.

Poster: Wed 16 July, 11 a.m. PDT - 1:30 p.m. PDT

Paper: https://arxiv.org/abs/2503.00229

Contact: vaswani.sharan@gmail.com, babanezhad@gmail.com

References i

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 1966.

Kyriakos Axiotis and Maxim Sviridenko. Gradient descent converges linearly for logistic regression on separable data. In *International Conference on Machine Learning*, pages 1302–1319. PMLR, 2023.

Curtis Fox and Mark Schmidt. Glocal smoothness: Line search can really help! In *OPT 2024: Optimization for Machine Learning*.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, 28, 2015.

Zhaosong Lu and Sanyou Mei. Accelerated first-order methods for convex optimization with locally lipschitz continuous gradient. *SIAM J. Optim.*, 33(3):2275–2310, 2023.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR, 2020.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021.

Katya Scheinberg, Donald Goldfarb, and Xi Bai. Fast first-order methods for composite convex optimization with backtracking. *Found. Comput. Math.*, 14(3):389–417, 2014.

Hossein Taheri and Christos Thrampoulidis. Fast convergence in learning two-layer neural networks with separable data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9944–9952, 2023.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. *arXiv preprint arXiv:2410.10800*, 2024.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32:3732–3745, 2019.

Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. *arXiv preprint arXiv:2402.15926*, 2024.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.