Towards Noise-adaptive, Problem-adaptive (Accelerated) Stochastic Gradient Descent

Sharan Vaswani (Simon Fraser University)
Joint work with: Benjamin Dubois-Taine, Reza Babanezhad

SIOPT'23

Problem Setup

Unconstrained minimization: finite-sum objective.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

where n is the number of training examples.

Problem Setup

Unconstrained minimization: finite-sum objective.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

where n is the number of training examples.

- Smoothness and convexity: Each f_i is convex, differentiable and L_i-smooth, implying that f is L-smooth where $L := \max_i L_i$.
- Strong convexity: f is μ strongly-convex.
- Standard assumptions satisfied when training convex machine learning models such as regularized logistic regression or linear least squares.

Problem Setup

Unconstrained minimization: finite-sum objective.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

where n is the number of training examples.

- Smoothness and convexity: Each f_i is convex, differentiable and L_i-smooth, implying that f is L-smooth where $L := \max_i L_i$.
- Strong convexity: f is μ strongly-convex.
- Standard assumptions satisfied when training convex machine learning models such as regularized logistic regression or linear least squares.
- Define $w^* := \arg\min_{w \in \mathbb{R}^d} f(w)$; $f_i^* := \min_{w \in \mathbb{R}^d} f_i(w)$

Introduction

- For smooth, strongly-convex functions with condition number κ, deterministic gradient descent (GD) uses a constant step-size and has an $O(\exp(-T/\kappa))$ convergence rate.

Introduction

- For smooth, strongly-convex functions with condition number κ, deterministic gradient descent (GD) uses a constant step-size and has an $O(\exp(-T/\kappa))$ convergence rate.
- Can be further improved to $\Theta(\exp(-T/\sqrt{\kappa}))$ using Nesterov acceleration.

Introduction

- For smooth, strongly-convex functions with condition number κ, deterministic gradient descent (GD) uses a constant step-size and has an $O(\exp(-T/\kappa))$ convergence rate.
- Can be further improved to $\Theta(\exp(-T/\sqrt{\kappa}))$ using Nesterov acceleration.
- Stochastic gradient descent (SGD) requires a decreasing $O(1/k)$ step-size and has an $\Theta(1/T)$ convergence rate.

Introduction

- For smooth, strongly-convex functions with condition number κ, deterministic gradient descent (GD) uses a constant step-size and has an $O(\exp(-T/\kappa))$ convergence rate.
- Can be further improved to $\Theta(\exp(-T/\sqrt{\kappa}))$ using Nesterov acceleration.
- Stochastic gradient descent (SGD) requires a decreasing $O(1/k)$ step-size and has an $\Theta(1/T)$ convergence rate.
- The two regimes require a different step-size choice (constant vs decreasing) and the convergence rate is not adaptive to the noise (σ^2) in the stochastic gradients.

Introduction

- For smooth, strongly-convex functions with condition number κ, deterministic gradient descent (GD) uses a constant step-size and has an $O(\exp(-T/\kappa))$ convergence rate.
- Can be further improved to $\Theta(\exp(-T/\sqrt{\kappa}))$ using Nesterov acceleration.
- Stochastic gradient descent (SGD) requires a decreasing $O(1/k)$ step-size and has an $\Theta(1/T)$ convergence rate.
- The two regimes require a different step-size choice (constant vs decreasing) and the convergence rate is not adaptive to the noise (σ^2) in the stochastic gradients.
- Require noise-adaptivity – one step-size sequence that can achieve the optimal rate in both the deterministic and stochastic settings without knowledge of σ^2.

Related work towards noise-adaptivity

Work that attains the $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ convergence rate for,

Related work towards noise-adaptivity

Work that attains the $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ convergence rate for,

- smooth, strongly-convex functions using SGD that switches between two carefully designed step-sizes [Stich, 2019]. Requires knowledge of L, μ and σ^2.

Related work towards noise-adaptivity

Work that attains the $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ convergence rate for,

- smooth, strongly-convex functions using SGD that switches between two carefully designed step-sizes [Stich, 2019]. Requires knowledge of L, μ and σ^2.
- smooth functions satisfying the PL condition using SGD with a constant then decaying step-size [Khaled and Richtárik, 2020]. Noise adaptive but requires knowledge of L, μ.

Related work towards noise-adaptivity

Work that attains the $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ convergence rate for,

- smooth, strongly-convex functions using SGD that switches between two carefully designed step-sizes [Stich, 2019]. Requires knowledge of L, μ and σ^2.
- smooth functions satisfying the PL condition using SGD with a constant then decaying step-size [Khaled and Richtárik, 2020]. Noise adaptive but requires knowledge of L, μ.
- smooth functions satisfying the PL condition using SGD with an exponentially decreasing sequence of step-sizes [Li et al., 2020]. Noise adaptive but requires knowledge of L.

Motivation

- Problem 1: All noise-adaptive methods require knowledge of problem-dependent constants, and are not problem-adaptive.

Motivation

- Problem 1: All noise-adaptive methods require knowledge of problem-dependent constants, and are not problem-adaptive.
- None of the problem-adaptive methods [Duchi et al., 2011, Kingma and Ba, 2015, Vaswani et al., 2019b, Loizou et al., 2021] are noise-adaptive when minimizing smooth, strongly-convex functions.

Motivation

- Problem 1: All noise-adaptive methods require knowledge of problem-dependent constants, and are not problem-adaptive.
- None of the problem-adaptive methods [Duchi et al., 2011, Kingma and Ba, 2015, Vaswani et al., 2019b, Loizou et al., 2021] are noise-adaptive when minimizing smooth, strongly-convex functions.
- Problem 2: Current noise-adaptive methods do not match the optimal $\sqrt{\kappa}$ dependence and are therefore sub-optimal in the deterministic setting.

Motivation

- Problem 1: All noise-adaptive methods require knowledge of problem-dependent constants, and are not problem-adaptive.
- None of the problem-adaptive methods [Duchi et al., 2011, Kingma and Ba, 2015, Vaswani et al., 2019b, Loizou et al., 2021] are noise-adaptive when minimizing smooth, strongly-convex functions.
- Problem 2: Current noise-adaptive methods do not match the optimal $\sqrt{\kappa}$ dependence and are therefore sub-optimal in the deterministic setting.

1. Can we design SGD step-sizes that are simultaneously (i) problem-adaptive and (ii) noise-adaptive – achieve the $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ rate without knowledge of L, μ or σ^2?
2. Can we obtain the accelerated $\tilde{O}\left(\exp(-T/\sqrt{\kappa}) + \frac{\sigma^2}{T}\right)$ rate?

Outline

- Problem 1: SGD with exponential step-sizes
  - Known smoothness
  - Online estimation of unknown smoothness
  - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
  - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

Outline

- Problem 1: SGD with exponential step-sizes
  - Known smoothness
  - Online estimation of unknown smoothness
  - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
  - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

SGD with exponentially decreasing step-sizes

$$\textbf{Update}: \quad w_{k+1} = w_k - \underbrace{\gamma_k \alpha_k}_{:=\eta_k} \nabla f_{ik}(w_k) \qquad \text{(SGD)}$$

where γ_k is the problem-dependent scaling term that captures the smoothness and α_k that controls the decay of the step-size.

SGD with exponentially decreasing step-sizes

$$\textbf{Update}: \quad w_{k+1} = w_k - \underbrace{\gamma_k \alpha_k}_{:=\eta_k} \nabla f_{ik}(w_k) \tag{SGD}$$

where γ_k is the problem-dependent scaling term that captures the smoothness and α_k that controls the decay of the step-size.

Exponentially decreasing step-sizes [Li et al., 2020]: $\alpha := \left[\frac{\beta}{T}\right]^{1/T} \leq 1$ for $\beta \geq 1$ and $\alpha_k := \alpha^k$.

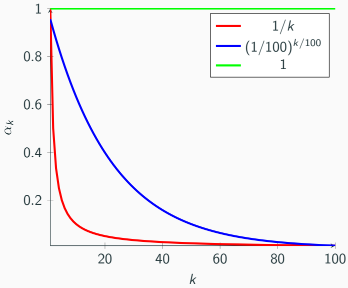
SGD with exponentially decreasing step-sizes

$$\textbf{Update}: \quad w_{k+1} = w_k - \underbrace{\gamma_k \alpha_k}_{:=\eta_k} \nabla f_{ik}(w_k) \tag{SGD}$$

where γ_k is the problem-dependent scaling term that captures the smoothness and α_k that controls the decay of the step-size.

Exponentially decreasing step-sizes [Li et al., 2020]: $\alpha := \left[\frac{\beta}{T}\right]^{1/T} \leq 1$ for $\beta \geq 1$ and $\alpha_k := \alpha^k$.

Exponential step-sizes lie between the constant and $1/k$ decreasing step-sizes, implying that for $k \in [T]$, $\alpha_k \in \left[\frac{1}{k}, 1\right]$.



Warm-up – known smoothness

- Assumption on the noise: $\sigma^2 := \mathbb{E}_i[f_i(w^*) - f_i^*] < \infty$.

Warm-up – known smoothness

- Assumption on the noise: $\sigma^2 := \mathbb{E}_i[f_i(w^*) - f_i^*] < \infty$.

SGD with known smoothness

Assuming (i) convexity and L-smoothness of each f_i, (ii) μ strong-convexity of f, SGD with $\gamma_k = \frac{1}{L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$ converges as,

Warm-up – known smoothness

- Assumption on the noise: $\sigma^2 := \mathbb{E}_i[f_i(w^*) - f_i^*] < \infty$.

SGD with known smoothness

Assuming (i) convexity and L-smoothness of each f_i, (ii) μ strong-convexity of f, SGD with $\gamma_k = \frac{1}{L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_2 \exp\left(-\frac{T}{\kappa}\frac{\alpha}{\ln({}^{T}/_{\beta})}\right)\left\|w_1 - w^*\right\|^2 + \frac{8\,c_2\kappa}{\mu e^2}\frac{(\ln({}^{T}/_{\beta}))^2}{\alpha^2}\frac{\sigma^2}{T},$$

where $\kappa = \frac{L}{\mu}$ and $c_2 = \exp\left(\frac{1}{\kappa}\cdot\frac{2\beta}{\ln({}^{T}/_{\beta})}\right)$.

Warm-up – known smoothness

- Assumption on the noise: $\sigma^2 := \mathbb{E}_i[f_i(w^*) - f_i^*] < \infty$.

SGD with known smoothness

Assuming (i) convexity and L-smoothness of each f_i, (ii) μ strong-convexity of f, SGD with $\gamma_k = \frac{1}{L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_2 \exp\left(-\frac{T}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \frac{8\, c_2 \kappa}{\mu e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T},$$

where $\kappa = \frac{L}{\mu}$ and $c_2 = \exp\left(\frac{1}{\kappa}\cdot\frac{2\beta}{\ln(T/\beta)}\right)$.

- Variance decreases as the mini-batch size increases. Equal to zero under interpolation for over-parameterized models [Loizou et al., 2021].

Warm-up – known smoothness

- Assumption on the noise: $\sigma^2 := \mathbb{E}_i[f_i(w^*) - f_i^*] < \infty$.

SGD with known smoothness

Assuming (i) convexity and L-smoothness of each f_i, (ii) μ strong-convexity of f, SGD with $\gamma_k = \frac{1}{L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_2 \exp\left(-\frac{T}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \frac{8\, c_2 \kappa}{\mu e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T},$$

where $\kappa = \frac{L}{\mu}$ and $c_2 = \exp\left(\frac{1}{\kappa} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

- Variance decreases as the mini-batch size increases. Equal to zero under interpolation for over-parameterized models [Loizou et al., 2021].
- Similar result in Li et al. [2020], but we do not require the growth condition and use a different proof technique that helps handle unknown smoothness later.

SGD with online estimation of unknown smoothness

- Since L is difficult to estimate, use stochastic line-search (SLS) [Vaswani et al., 2019b] to automatically set γ_k, the problem-dependent part of the step-size.

SGD with online estimation of unknown smoothness

- Since L is difficult to estimate, use stochastic line-search (SLS) [Vaswani et al., 2019b] to automatically set γ_k, the problem-dependent part of the step-size.
- Starting from a guess ($\gamma_{\max}$) of the step-size, SLS uses a backtracking procedure and returns the largest step-size γ_k that satisfies the following conditions: $\gamma_k \leq \gamma_{\max}$ and

$$f_{ik}(w_k - \gamma_k \nabla f_{ik}(w_k)) \leq f_{ik}(w_k) - c\gamma_k \left\|\nabla f_{ik}(w_k)\right\|^2 .$$

SGD with online estimation of unknown smoothness

- Since L is difficult to estimate, use stochastic line-search (SLS) [Vaswani et al., 2019b] to automatically set γ_k, the problem-dependent part of the step-size.
- Starting from a guess (γ_{max}) of the step-size, SLS uses a backtracking procedure and returns the largest step-size γ_k that satisfies the following conditions: $\gamma_k \leq \gamma_{\text{max}}$ and

$$f_{ik}(w_k - \gamma_k \nabla f_{ik}(w_k)) \leq f_{ik}(w_k) - c\gamma_k \left\|\nabla f_{ik}(w_k)\right\|^2 .$$

- Ensures that $\gamma_k \in \left[\min\left\{\frac{2(1-c)}{L_{ik}}, \gamma_{\text{max}}\right\}, \gamma_{\text{max}}\right]$.

SGD with online estimation of unknown smoothness

- Since L is difficult to estimate, use stochastic line-search (SLS) [Vaswani et al., 2019b] to automatically set γ_k, the problem-dependent part of the step-size.
- Starting from a guess ($\gamma_{\max}$) of the step-size, SLS uses a backtracking procedure and returns the largest step-size γ_k that satisfies the following conditions: $\gamma_k \leq \gamma_{\max}$ and

$$f_{ik}(w_k - \gamma_k \nabla f_{ik}(w_k)) \leq f_{ik}(w_k) - c\gamma_k \left\|\nabla f_{ik}(w_k)\right\|^2 .$$

- Ensures that $\gamma_k \in \left[\min\left\{\frac{2(1-c)}{L_{ik}}, \gamma_{\max}\right\}, \gamma_{\max}\right]$.
- When $\sigma = 0$, SGD with $\alpha_k = 1$ for all k and γ_k set according to SLS (with $c \geq 1/2$) has an $O(\exp(-T/\kappa))$ convergence to the minimizer [Vaswani et al., 2019b].

SGD with online estimation of unknown smoothness

- Since L is difficult to estimate, use stochastic line-search (SLS) [Vaswani et al., 2019b] to automatically set γ_k, the problem-dependent part of the step-size.
- Starting from a guess ($\gamma_{\max}$) of the step-size, SLS uses a backtracking procedure and returns the largest step-size γ_k that satisfies the following conditions: $\gamma_k \leq \gamma_{\max}$ and

$$f_{ik}(w_k - \gamma_k \nabla f_{ik}(w_k)) \leq f_{ik}(w_k) - c\gamma_k \left\|\nabla f_{ik}(w_k)\right\|^2 .$$

- Ensures that $\gamma_k \in \left[\min\left\{\frac{2(1-c)}{L_{ik}}, \gamma_{\max}\right\}, \gamma_{\max}\right]$.
- When $\sigma = 0$, SGD with $\alpha_k = 1$ for all k and γ_k set according to SLS (with $c \geq 1/2$) has an $O(\exp(-T/\kappa))$ convergence to the minimizer [Vaswani et al., 2019b].
- When $\sigma \neq 0$, this method converges to a neighbourhood that depends on $\gamma_{\max}\sigma^2$.

Convergence of SGD with SLS

SGD with SLS – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and the SLS condition with $c = 1/2$ converges as,

Convergence of SGD with SLS

SGD with SLS – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and the SLS condition with $c = {}^1\!/_2$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_1 \exp\left(-\frac{T}{\kappa'}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \frac{8\,c_1(\kappa')^2\gamma_{\max}}{e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T} + \frac{2c_1\kappa'\ln(T/\beta)\,\sigma^2\left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)}{e\alpha},$$

where $\kappa' := \max\left\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

Convergence of SGD with SLS

SGD with SLS – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and the SLS condition with $c = 1/2$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_1 \exp\left(-\frac{T}{\kappa'}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \frac{8\, c_1 (\kappa')^2 \gamma_{\max}}{e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T} + \frac{2 c_1 \kappa' \ln(T/\beta)\, \sigma^2 \left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)}{e\alpha},$$

where $\kappa' := \max\left\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

- $O\left(\exp(-T/\kappa) + \sigma^2/T\right)$ convergence to a neighbourhood determined by σ^2 and $\left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$.

Convergence of SGD with SLS

SGD with SLS – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and the SLS condition with $c = 1/2$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_1 \exp\left(-\frac{T}{\kappa'}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \frac{8\, c_1 (\kappa')^2 \gamma_{\max}}{e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T} + \frac{2 c_1 \kappa' \ln(T/\beta)\, \sigma^2 \left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)}{e\alpha},$$

where $\kappa' := \max\left\{\frac{L}{\mu}, \frac{1}{\mu \gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

- $O\left(\exp(-T/\kappa) + \sigma^2/T\right)$ convergence to a neighbourhood determined by σ^2 and $\left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$.
- If $\sigma = 0$, recovers the rate in [Vaswani et al., 2019b] upto log factors.

Convergence of SGD with SLS

SGD with SLS – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and the SLS condition with $c = 1/2$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_1 \exp\left(-\frac{T}{\kappa'}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \frac{8\, c_1 (\kappa')^2 \gamma_{\max}}{e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T} + \frac{2 c_1 \kappa' \ln(T/\beta)\, \sigma^2 \left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)}{e\alpha},$$

where $\kappa' := \max\left\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

- $O\left(\exp(-T/\kappa) + \sigma^2/T\right)$ convergence to a neighbourhood determined by σ^2 and $\left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$.
- If $\sigma = 0$, recovers the rate in [Vaswani et al., 2019b] upto log factors.
- If $\gamma_{\max} \leq \frac{1}{L}$, equivalent to constant step-size SGD with convergence to the minimizer.

Convergence of SGD with SLS

SGD with SLS – Lower Bound

When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w)+f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w-1)^2$ and $f_2(w) = \frac{1}{2}\left(2w + {}^1\!/_2\right)^2$, setting the step-size using SLS with $\gamma_{\text{max}} \geq 1$ and $c \geq {}^1\!/_2$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution.

Convergence of SGD with SLS

SGD with SLS – Lower Bound

When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w)+f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w-1)^2$ and $f_2(w) = \frac{1}{2}\left(2w + {}^1\!/_2\right)^2$, setting the step-size using SLS with $\gamma_{\max} \geq 1$ and $c \geq {}^1\!/_2$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution. Specifically, if $w_1 > 0$, then,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right).$$

Convergence of SGD with SLS

SGD with SLS – Lower Bound

When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w)+f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w-1)^2$ and $f_2(w) = \frac{1}{2}\left(2w + {}^1/_2\right)^2$, setting the step-size using SLS with $\gamma_{\max} \geq 1$ and $c \geq {}^1/_2$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution. Specifically, if $w_1 > 0$, then,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right).$$

- Lower-bound is not specific to SLS and will work for other methods [Loizou et al., 2021, Berrada et al., 2020] that set the step-size in an online fashion.

Convergence of SGD with SLS

SGD with SLS – Lower Bound

When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w)+f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w-1)^2$ and $f_2(w) = \frac{1}{2}\left(2w + 1/2\right)^2$, setting the step-size using SLS with $\gamma_{\max} \geq 1$ and $c \geq 1/2$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution. Specifically, if $w_1 > 0$, then,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right).$$

- Lower-bound is not specific to SLS and will work for other methods [Loizou et al., 2021, Berrada et al., 2020] that set the step-size in an online fashion.
- Lower-bound is not specific to exponential step-sizes and works for any α_k sequence.

Convergence of SGD with SLS

SGD with SLS – Lower Bound

When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w)+f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w-1)^2$ and $f_2(w) = \frac{1}{2}\left(2w + 1/2\right)^2$, setting the step-size using SLS with $\gamma_{\max} \geq 1$ and $c \geq 1/2$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution. Specifically, if $w_1 > 0$, then,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right).$$

- Lower-bound is not specific to SLS and will work for other methods [Loizou et al., 2021, Berrada et al., 2020] that set the step-size in an online fashion.
- Lower-bound is not specific to exponential step-sizes and works for any α_k sequence.
- **Neighbourhood term is the price of misestimation of the smoothness.**

Convergence of SGD with SLS

SGD with SLS – Lower Bound

When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w)+f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w-1)^2$ and $f_2(w) = \frac{1}{2}\left(2w + 1/2\right)^2$, setting the step-size using SLS with $\gamma_{\max} \geq 1$ and $c \geq 1/2$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution. Specifically, if $w_1 > 0$, then,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right).$$

- Lower-bound is not specific to SLS and will work for other methods [Loizou et al., 2021, Berrada et al., 2020] that set the step-size in an online fashion.
- Lower-bound is not specific to exponential step-sizes and works for any α_k sequence.
- **Neighbourhood term is the price of misestimation of the smoothness.**

Idea: Estimate the smoothness ensuring that there is no correlation between γ_k and i_k.

SGD with offline estimation of unknown smoothness

- Using **any** method, set γ_k *before* sampling i_k. For the analysis, consider a fixed $\gamma_k = \gamma$ and assume that $\gamma = \frac{\nu}{L}$ where $\nu > 0$ is the misestimation in $1/L$.

SGD with offline estimation of unknown smoothness

- Using **any** method, set γ_k *before* sampling i_k. For the analysis, consider a fixed $\gamma_k = \gamma$ and assume that $\gamma = \frac{\nu}{L}$ where $\nu > 0$ is the misestimation in $1/L$.

SGD with offline estimation of the smoothness – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ converges as,

SGD with offline estimation of unknown smoothness

- Using **any** method, set γ_k *before* sampling i_k. For the analysis, consider a fixed $\gamma_k = \gamma$ and assume that $\gamma = \frac{\nu}{L}$ where $\nu > 0$ is the misestimation in $1/L$.

SGD with offline estimation of the smoothness – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_2 \exp\left(\frac{-T\min\{\nu, 1\}}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2$$
$$+ \max\{\nu^2, 1\}\frac{8c_2\kappa\ln(T/\beta)}{\mu\, e^2\, \alpha^2}\,\frac{\left[2\ln(T/\beta)\sigma^2 + G\left[\ln(\nu)\right]_+\right]}{T}$$

where $c_2 = \exp\left(\frac{1}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$, $[x]_+ = \max\{x, 0\}$, $G = \max_{j\in[k_0]}\{f(w_j) - f^*\}$ and $k_0 := \tilde{O}(T[\ln(\nu)]_+)$.

SGD with offline estimation of unknown smoothness

- Using **any** method, set γ_k *before* sampling i_k. For the analysis, consider a fixed $\gamma_k = \gamma$ and assume that $\gamma = \frac{\nu}{L}$ where $\nu > 0$ is the misestimation in $1/L$.

SGD with offline estimation of the smoothness – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_2 \exp\left(\frac{-T\min\{\nu, 1\}}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \max\{\nu^2, 1\}\frac{8c_2\kappa\ln(T/\beta)}{\mu\, e^2\, \alpha^2}\frac{\left[2\ln(T/\beta)\sigma^2 + G\,[\ln(\nu)]_+\right]}{T}$$

where $c_2 = \exp\left(\frac{1}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$, $[x]_+ = \max\{x, 0\}$, $G = \max_{j\in[k_0]}\{f(w_j) - f^*\}$ and $k_0 := \tilde{O}(T[\ln(\nu)]_+)$.

- Ensures convergence to the minimizer, but the rate is slowed down proportional to ν.

SGD with offline estimation of unknown smoothness

- Using **any** method, set γ_k *before* sampling i_k. For the analysis, consider a fixed $\gamma_k = \gamma$ and assume that $\gamma = \frac{\nu}{L}$ where $\nu > 0$ is the misestimation in $1/L$.

SGD with offline estimation of the smoothness – Upper Bound

Under the same assumptions, SGD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ converges as,

$$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq c_2 \exp\left(\frac{-T\min\{\nu,1\}}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\left\|w_1 - w^*\right\|^2 + \max\{\nu^2, 1\}\frac{8c_2\kappa\ln(T/\beta)}{\mu\, e^2\,\alpha^2}\frac{\left[2\ln(T/\beta)\sigma^2 + G\left[\ln(\nu)\right]_+\right]}{T}$$

where $c_2 = \exp\left(\frac{1}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$, $[x]_+ = \max\{x, 0\}$, $G = \max_{j\in[k_0]}\{f(w_j) - f^*\}$ and $k_0 := \tilde{O}(T[\ln(\nu)]_+)$.

- Ensures convergence to the minimizer, but the rate is slowed down proportional to ν.
- For polynomial α_k sequences, Moulines and Bach [2011] show an $O(\exp(\nu))$ dependence on the rate $\implies$ exponential step-sizes are more robust towards misspecification.

SGD with offline estimation of unknown smoothness

SGD with offline estimation of the smoothness – Lower Bound

When minimizing a one-dimensional quadratic function $f(w) = \frac{1}{2}(xw - y)^2$, GD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ for $\nu > 3$, satisfies

SGD with offline estimation of unknown smoothness

SGD with offline estimation of the smoothness – Lower Bound

When minimizing a one-dimensional quadratic function $f(w) = \frac{1}{2}(xw - y)^2$, GD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ for $\nu > 3$, satisfies

$$w_{k+1} - w^* = (w_1 - w^*) \prod_{i=1}^{k} (1 - \nu\alpha_i).$$

After $k' := \frac{T}{\ln(T/\beta)} \ln\left(\frac{\nu}{3}\right)$ iterations, we have that

$$|w_{k'+1} - w^*| \geq 2^{k'} |w_1 - w^*|.$$

SGD with offline estimation of unknown smoothness

SGD with offline estimation of the smoothness – Lower Bound

When minimizing a one-dimensional quadratic function $f(w) = \frac{1}{2}(xw - y)^2$, GD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ for $\nu > 3$, satisfies

$$w_{k+1} - w^* = (w_1 - w^*) \prod_{i=1}^{k} (1 - \nu\alpha_i).$$

After $k' := \frac{T}{\ln(T/\beta)} \ln\left(\frac{\nu}{3}\right)$ iterations, we have that

$$|w_{k'+1} - w^*| \geq 2^{k'} |w_1 - w^*|.$$

- If $\nu = 10$, then $k' \geq \lfloor \frac{T}{\ln(T/\beta)} \rfloor \implies$ divergence in the first $\frac{T}{\ln(T/\beta)}$ iterations, and the optimality gap has been increased by a factor of $2^{T/\ln(T/\beta)}$.

SGD with offline estimation of unknown smoothness

SGD with offline estimation of the smoothness – Lower Bound

When minimizing a one-dimensional quadratic function $f(w) = \frac{1}{2}(xw - y)^2$, GD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ for $\nu > 3$, satisfies

$$w_{k+1} - w^* = (w_1 - w^*)\prod_{i=1}^{k}(1 - \nu\alpha_i).$$

After $k' := \frac{T}{\ln(T/\beta)} \ln\left(\frac{\nu}{3}\right)$ iterations, we have that

$$|w_{k'+1} - w^*| \geq 2^{k'}|w_1 - w^*|.$$

- If $\nu = 10$, then $k' \geq \lfloor \frac{T}{\ln(T/\beta)} \rfloor \implies$ divergence in the first $\frac{T}{\ln(T/\beta)}$ iterations, and the optimality gap has been increased by a factor of $2^{T/\ln(T/\beta)}$.
- **Slowdown in rate is the price of misestimation of the smoothness**.

Outline

- Problem 1: SGD with exponential step-sizes
  - Known smoothness
  - Online estimation of unknown smoothness
  - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
  - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

Outline

- Problem 1: SGD with exponential step-sizes
  - Known smoothness
  - Online estimation of unknown smoothness
  - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
  - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

Accelerated SGD with exponentially decreasing step-sizes

$$\textbf{Update}: \quad \begin{aligned} y_k &= w_k + b_k\,(w_k - w_{k-1}), \\ w_{k+1} &= y_k - \gamma_k \alpha_k \nabla f_{ik}(y_k). \end{aligned} \tag{ASGD}$$

where $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, and $(r_k,\ b_k)$ satisfy $r_k^2 = (1 - r_k) r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}} + r_k \mu \eta_k$,
$b_k = \frac{(1-r_{k-1}) r_{k-1} \frac{\eta_k}{\eta_{k-1}}}{r_k + r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}}}$.

Accelerated SGD with exponentially decreasing step-sizes

$$\textbf{Update}: \qquad \begin{aligned} y_k &= w_k + b_k\left(w_k - w_{k-1}\right), \\ w_{k+1} &= y_k - \gamma_k \alpha_k \nabla f_{ik}(y_k). \end{aligned} \tag{ASGD}$$

where $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, and $(r_k,\, b_k)$ satisfy $r_k^2 = (1 - r_k) r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}} + r_k \mu \eta_k$,
$b_k = \frac{(1-r_{k-1}) r_{k-1} \frac{\eta_k}{\eta_{k-1}}}{r_k + r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}}}$.

- Equivalent to Nesterov acceleration if we use a deterministic gradient $\nabla f(y_k)$ and $\gamma_k = \gamma = \frac{1}{L}$ and $\alpha_k = 1$ for all k.

Accelerated SGD with exponentially decreasing step-sizes

$$\textbf{Update}: \quad \begin{aligned} y_k &= w_k + b_k(w_k - w_{k-1}), \\ w_{k+1} &= y_k - \gamma_k \alpha_k \nabla f_{ik}(y_k). \end{aligned} \tag{ASGD}$$

where $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, and (r_k, b_k) satisfy $r_k^2 = (1 - r_k) r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}} + r_k \mu \eta_k$, $b_k = \frac{(1-r_{k-1}) r_{k-1} \frac{\eta_k}{\eta_{k-1}}}{r_k + r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}}}$.

- Equivalent to Nesterov acceleration if we use a deterministic gradient $\nabla f(y_k)$ and $\gamma_k = \gamma = \frac{1}{L}$ and $\alpha_k = 1$ for all k.
- Assumption on the noise: $\mathbb{E}_i \|\nabla f_i(w)\|^2 \leq \rho \|\nabla f(w)\|^2 + \sigma^2$.

Convergence of ASGD

Under the same assumptions as before and (iii) the growth condition on the stochastic gradients, if $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$, ASGD with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}$ converges as,

Convergence of Accelerated SGD – Known smoothness & strong-convexity

Convergence of ASGD

Under the same assumptions as before and (iii) the growth condition on the stochastic gradients, if $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$, ASGD with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\,\alpha}{r_k + r_{k-1}^2\,\alpha}$ converges as,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2\, c_3 \exp\left(\frac{-T}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln(T/\beta)}\right) \mathbb{E}[f(w_1) - f^*] + \frac{2\, c_3}{\rho\mu e^2} \frac{(\ln(T/\beta))^2}{\alpha^2} \frac{\sigma^2}{T},$$

- In the deterministic setting, $\rho = 1$ and $\sigma = 0$, and ASGD is near-optimal.

Convergence of Accelerated SGD – Known smoothness & strong-convexity

Convergence of ASGD

Under the same assumptions as before and (iii) the growth condition on the stochastic gradients, if $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$, ASGD with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\,\alpha}{r_k + r_{k-1}^2\,\alpha}$ converges as,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2\, c_3 \exp\left(\frac{-T}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln(T/\beta)}\right) \mathbb{E}[f(w_1) - f^*] + \frac{2\, c_3}{\rho\mu e^2} \frac{(\ln(T/\beta))^2}{\alpha^2} \frac{\sigma^2}{T},$$

- In the deterministic setting, $\rho = 1$ and $\sigma = 0$, and ASGD is near-optimal.
- In the worst-case, $\rho = O(\kappa)$ and ASGD has the same rate as SGD. When using a mini-batch size b, $\rho = O(\kappa/b)$. If b is large enough s.t. $\rho < \kappa$, ASGD is faster than SGD.

Convergence of ASGD

Under the same assumptions as before and (iii) the growth condition on the stochastic gradients, if $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$, ASGD with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}$ converges as,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2\, c_3 \exp\left(\frac{-T}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln(T/\beta)}\right) \mathbb{E}[f(w_1) - f^*] + \frac{2\, c_3}{\rho\mu e^2} \frac{(\ln(T/\beta))^2}{\alpha^2} \frac{\sigma^2}{T},$$

- In the deterministic setting, $\rho = 1$ and $\sigma = 0$, and ASGD is near-optimal.
- In the worst-case, $\rho = O(\kappa)$ and ASGD has the same rate as SGD. When using a mini-batch size b, $\rho = O(\kappa/b)$. If b is large enough s.t. $\rho < \kappa$, ASGD is faster than SGD.
- When $\sigma = 0$, the rate improves over Vaswani et al. [2019a] and matches [Mishkin, 2020].

Convergence of Accelerated SGD – Known smoothness & strong-convexity

Convergence of ASGD

Under the same assumptions as before and (iii) the growth condition on the stochastic gradients, if $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$, ASGD with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}$ converges as,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2\, c_3\, \exp\left(\frac{-T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\mathbb{E}[f(w_1) - f^*] + \frac{2\, c_3}{\rho\mu e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T},$$

- In the deterministic setting, $\rho = 1$ and $\sigma = 0$, and ASGD is near-optimal.
- In the worst-case, $\rho = O(\kappa)$ and ASGD has the same rate as SGD. When using a mini-batch size b, $\rho = O(\kappa/b)$. If b is large enough s.t. $\rho < \kappa$, ASGD is faster than SGD.
- When $\sigma = 0$, the rate improves over Vaswani et al. [2019a] and matches [Mishkin, 2020].
- Aybat et al. [2019] use a more complicated algorithm and prove this rate when $T \geq 2\sqrt{\kappa}$.

Convergence of Accelerated SGD – Known smoothness & strong-convexity

Convergence of ASGD

Under the same assumptions as before and (iii) the growth condition on the stochastic gradients, if $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$, ASGD with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\,\alpha}{r_k + r_{k-1}^2\,\alpha}$ converges as,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2\, c_3 \exp\left(\frac{-T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\mathbb{E}[f(w_1) - f^*] + \frac{2\, c_3}{\rho\mu e^2}\frac{(\ln(T/\beta))^2}{\alpha^2}\frac{\sigma^2}{T},$$

- In the deterministic setting, $\rho = 1$ and $\sigma = 0$, and ASGD is near-optimal.
- In the worst-case, $\rho = O(\kappa)$ and ASGD has the same rate as SGD. When using a mini-batch size b, $\rho = O(\kappa/b)$. If b is large enough s.t. $\rho < \kappa$, ASGD is faster than SGD.
- When $\sigma = 0$, the rate improves over Vaswani et al. [2019a] and matches [Mishkin, 2020].
- Aybat et al. [2019] use a more complicated algorithm and prove this rate when $T \geq 2\sqrt{\kappa}$.
- Similar to SGD, we can quantify the effect of misestimating L, μ on the convergence rate.

Outline

- Problem 1: SGD with exponential step-sizes
    - Known smoothness
    - Online estimation of unknown smoothness
    - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
    - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

Outline

- Problem 1: SGD with exponential step-sizes
  - Known smoothness
  - Online estimation of unknown smoothness
  - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
  - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

Experimental evaluation

- **Conservative decorrelated SLS**: At iteration k, we use a stochastic line-search starting from γ_{k-1} (with $\gamma_0 = \gamma_{\max}$) for the previously sampled function ($j_k = i_{k-1}$), find the largest step-size γ_k that satisfies,

$$f_{j_k}(w_k - \gamma_k \nabla f_{j_k}(w_k)) \leq f_{j_k}(w_k) - c\gamma_k \left\|\nabla f_{j_k}(w_k)\right\|^2 .$$

Experimental evaluation

- **Conservative decorrelated SLS**: At iteration k, we use a stochastic line-search starting from γ_{k-1} (with $\gamma_0 = \gamma_{\max}$) for the previously sampled function ($j_k = i_{k-1}$), find the largest step-size γ_k that satisfies,

$$f_{j_k}(w_k - \gamma_k \nabla f_{j_k}(w_k)) \leq f_{j_k}(w_k) - c\gamma_k \left\| \nabla f_{j_k}(w_k) \right\|^2 .$$

- Ensures that there is no correlation between γ_k and i_k, and that ν is reasonable in practice.

Experimental evaluation

- **Conservative decorrelated SLS**: At iteration k, we use a stochastic line-search starting from γ_{k-1} (with $\gamma_0 = \gamma_{\max}$) for the previously sampled function ($j_k = i_{k-1}$), find the largest step-size γ_k that satisfies,

$$f_{j_k}(w_k - \gamma_k \nabla f_{j_k}(w_k)) \leq f_{j_k}(w_k) - c\gamma_k \left\|\nabla f_{j_k}(w_k)\right\|^2 .$$

- Ensures that there is no correlation between γ_k and i_k, and that ν is reasonable in practice.

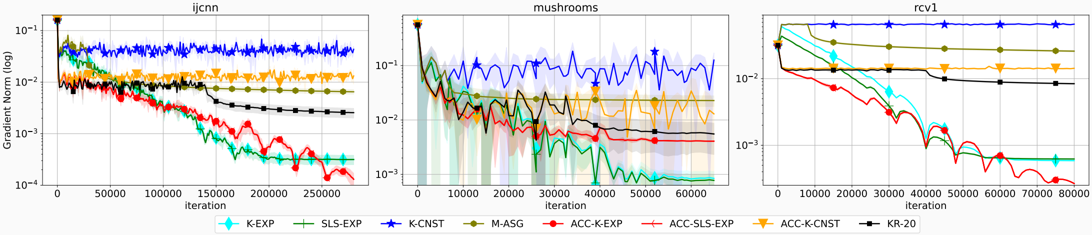


Figure 1: Regularized logistic regression

Outline

- Problem 1: SGD with exponential step-sizes
  - Known smoothness
  - Online estimation of unknown smoothness
  - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
  - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

Outline

- Problem 1: SGD with exponential step-sizes
  - Known smoothness
  - Online estimation of unknown smoothness
  - Offline estimation of unknown smoothness
- Problem 2: Accelerated SGD with exponential step-sizes
  - Known smoothness & strong-convexity
- Experimental evaluation
- Conclusions and Future Work

Conclusions and Future Work

- ✓ Used exponentially decreasing step-sizes to make SGD noise-adaptive.
- ✓ Quantified the price of problem-adaptivity – estimating the smoothness in an online fashion results in convergence to a neighbourhood of the solution, while an offline estimation results in a slower convergence to the minimizer.
- ✓ Developed an accelerated variant of SGD (ASGD) and proved that it achieves the near-optimal noise-adaptive convergence rate.

Conclusions and Future Work

- ✓ Used exponentially decreasing step-sizes to make SGD noise-adaptive.
- ✓ Quantified the price of problem-adaptivity – estimating the smoothness in an online fashion results in convergence to a neighbourhood of the solution, while an offline estimation results in a slower convergence to the minimizer.
- ✓ Developed an accelerated variant of SGD (ASGD) and proved that it achieves the near-optimal noise-adaptive convergence rate.
- Algorithms without any price of misestimation?

Conclusions and Future Work

- ✓ Used exponentially decreasing step-sizes to make SGD noise-adaptive.
- ✓ Quantified the price of problem-adaptivity – estimating the smoothness in an online fashion results in convergence to a neighbourhood of the solution, while an offline estimation results in a slower convergence to the minimizer.
- ✓ Developed an accelerated variant of SGD (ASGD) and proved that it achieves the near-optimal noise-adaptive convergence rate.
- Algorithms without any price of misestimation?
- × Exponential step-sizes do not seem to be noise-adaptive for convex functions (without strong-convexity) [Upper-bound]. Results showing that it is unlikely any oblivious exponential/polynomial step-size will be noise-adaptive in this case.
- Oblivious step-size schemes that are noise-adaptive for convex functions?

Questions?

Paper: https://arxiv.org/abs/2110.11442

Code: https://github.com/R3za/expsls

Contact: vaswani.sharan@gmail.com

Backup Slides

ASGD with offline estimation of the smoothness & strong-convexity

- Assume $\gamma_k = \gamma = {}^1/_{\rho\tilde{L}} = \frac{\nu_L}{\rho L}$ and $\tilde{\mu} = \nu_\mu \mu$ where $\nu_\mu \leq 1$.

Convergence of ASGD

Under the same assumptions and $\nu = \nu_L \nu_\mu \leq \rho\kappa$, ASGD with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho\tilde{L}} = \frac{\nu_L}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\tilde{\mu} = \nu_\mu \mu \leq \mu$, $r_k = \sqrt{\frac{\nu}{\rho\kappa}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}$ converges as,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2c_3 \exp\left(\frac{-T\sqrt{\min\{\nu, 1\}}}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln({}^T/_\beta)}\right) \mathbb{E}[f(w_1) - f^*]$$
$$+ \frac{2c_3(\ln({}^T/_\beta))^2}{e^2\alpha^2\mu} \frac{\left[\frac{\sigma^2}{\rho} + \frac{G^2\,[\ln(\nu_L)]_+}{\ln(T/\beta)}\right]}{T} \max\{\frac{\nu_L}{\nu_\mu}, \nu_L^2\},$$

where $c_3 = \exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{2\beta}{\ln({}^T/_\beta)}\right)$, $k_0 := \lfloor T \frac{[\ln(\nu_L)]_+}{\ln(T/\beta)} \rfloor$, $G = \max_{j\in[k_0]} \|\nabla f(y_j)\|$.

- Implies an $\tilde{O}\left(\exp\left(\frac{-T\sqrt{\min\{\nu,1\}}}{\sqrt{\kappa\rho}}\right) + \left[\frac{\sigma^2 + G^2[\ln(\nu_L)]_+}{T}\right] \max\{\frac{\nu_L}{\nu_\mu}, \nu_L^2\}\right)$ rate.

References i

Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. *Advances in neural information processing systems*, 32: 8525–8536, 2019.

Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Training neural networks for and by interpolation. In *International Conference on Machine Learning*, pages 799–809. PMLR, 2020.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, convergence, and performance. *arXiv preprint arXiv:2002.05273*, 2020.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

Aaron Mishkin. *Interpolation, growth conditions, and stochastic gradient descent*. PhD thesis, University of British Columbia, 2020.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24:451–459, 2011.

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019a.

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, 2019b.