Stochastic Reweighted Gradient Descent (SRG)

Reporter:
廖修誼 (111652017)
吳泓諺 (111652040)

Lecture Videos, Slides, and Handout

- https://drive.google.com/drive/folders/1gKKlaybTYVS-bVmEWT8sod6z2_uI3PDf?usp=sharing

Outline

- Motivation
- SRG Algorithm
- SRG+ (Do SRG better)
- Experiment
- Theory Idea

Part I SRG Motivation

Could we do SGD better ?

SGD update: $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

Could we do SGD better ?

SGD update: $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

Motivation: use another sampling probability?

Could we do SGD better ?

SGD update: $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

SGD update generalization: $x_{k+1} = x_k - \alpha_k \frac{1}{np_k^{i_k}} \nabla f_{i_k}(x_k)$

Motivation: use another sampling probability?

Could we do SGD better ?

SGD update: $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

apply $p_k^{i_k} = \frac{1}{n}$

SGD update generalization: $x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$

Motivation: use another sampling probability? (instead of uniform sampling)

Could we do SGD better ?

SGD update: $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

apply $p_k^{i_k} = \frac{1}{n}$

SGD update generalization: $x_{k+1} = x_k - \alpha_k \boxed{\frac{1}{np_k^{i_k}} \nabla f_{i_k}(x_k)}$

Motivation: use other sampling probability.
Q: If we use other sampling probability, is this
gradient estimator still unbiased ? Yes, as long as p_k > 0. (why ?)

Importance sampling

SGD update: $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

apply $p_k^{i_k} = \frac{1}{n}$

SGD update generalization: $x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$

$$\mathbb{E}_p[f(x)] = \int p(x) f(x) dx = \int p(x) \frac{q(x)}{q(x)} f(x) dx = \int q(x) [f(x) \frac{p(x)}{q(x)}] dx = \mathbb{E}_q[f(x) \frac{p(x)}{q(x)}]$$

Old Distribution: 1/n

New Distribution: p_k^{i_k}

Ref: (Importance Sampling) https://medium.com/@liuec.jessica2000/importance-sampling-explained-end-to-end-a53334cb330b

Goal

SGD update generalization: $x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$

reduce variance of gradient estimator

improve asymptotic error

O(n) additional memory per iteration

O(log n) FP operations per iteration

Goal

SGD update generalization: $x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$

reduce variance of gradient estimator

improve asymptotic error

O(n) additional memory per iteration

O(log n) FP operations per iteration

How to choose sampling probability p_k?

Greedy strategy: choose p_k to
minimize the conditional variance of the gradient estimator

How to choose sampling probability p_k?

Greedy strategy: choose p_k to
minimize the conditional variance of the gradient estimator

$$\sigma^2(x_k, p) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{p_i} \|\nabla f_i(x_k)\|_2^2$$

How to choose sampling probability p_k?

Greedy strategy: choose p_k to
minimize the conditional variance of the gradient estimator

$$\sigma^2(x_k, p) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{p_i} \|\nabla f_i(x_k)\|_2^2$$

Zhao & Zhang, 2015

$$\arg\min_p \sigma^2(x_k, p) = \left(\frac{\|\nabla f_i(x_k)\|_2}{\sum_{j=1}^{n} \|\nabla f_j(x_k)\|_2} \right)_{i=1}^{n}$$

Computation cost Problem

$$\arg\min_{p} \sigma^2(x_k, p) = \left(\frac{\|\nabla f_i(x_k)\|_2}{\sum_{j=1}^{n} \|\nabla f_j(x_k)\|_2} \right)_{i=1}^{n}$$

Computation cost Problem

$$\arg\min_{p} \sigma^2(x_k, p) = \left(\frac{\|\nabla f_i(x_k)\|_2}{\sum_{j=1}^{n} \|\nabla f_j(x_k)\|_2} \right)_{i=1}^{n}$$

Take much time

Computation cost Problem

$$\arg\min_{p} \sigma^2(x_k, p) = \left(\frac{\|\nabla f_i(x_k)\|_2}{\sum_{j=1}^{n} \|\nabla f_j(x_k)\|_2}\right)_{i=1}^{n}$$

Take much time

Q: Could we construct efficient approximations of the conditional variances ?
A: Yes. (variance reduction method)

Minimize Approximator

$$\sigma^2(x_k, p) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{p_i} \|\nabla f_i(x_k)\|_2^2$$

Minimize Approximator

$$\sigma^2(x_k, p) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{p_i}\|\nabla f_i(x_k)\|_2^2$$

g_k^i : component gradients of $\nabla f_i(x_k)$

$$\tilde{\sigma}^2(x_k, p) := \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{p_i}\|g_k^i\|_2^2$$

Minimize Approximator

$$\sigma^2(x_k, p) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{p_i} \|\nabla f_i(x_k)\|_2^2$$

g_k^i : component gradients of $\nabla f_i(x_k)$

$$\tilde{\sigma}^2(x_k, p) := \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{p_i} \|g_k^i\|_2^2$$

Zhao & Zhang, 2015

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \|g_k^j\|_2} \right)_{i=1}^{n}$$

Why such Approximator is good ?

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \|g_k^j\|_2} \right)_{i=1}^n$$

$$\Rightarrow \quad \sigma^2(x_k, p) \leq \frac{2}{n^2} \sum_{i=1}^n \frac{\left\|\nabla f_i(x_k) - g_k^i\right\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p) \quad (7)$$

How to calculate g_k^i

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \|g_k^j\|_2} \right)_{i=1}^n$$

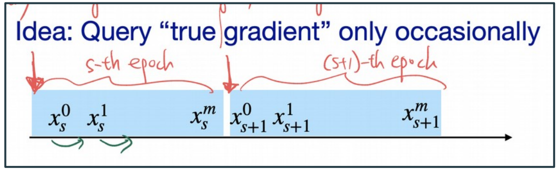
How to calculate

Ref: Lecture 7-Stochastic Gradient Descent and Variance Reduction

How to calculate g_k^i

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \|g_k^j\|_2} \right)_{i=1}^n$$

How to calculate



Like in Variance Reduction, we maintain an array of gradient norms, where g_k^i : component gradients of $\nabla f_i(x_k)$.

Ref: Lecture 7-Stochastic Gradient Descent and Variance Reduction

How to ensure RHS is small ?

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \|g_k^j\|_2} \right)_{i=1}^n$$

$$\sigma^2(x_k, p) \leq \frac{2}{n^2} \sum_{i=1}^n \frac{\left\|\nabla f_i(x_k) - g_k^i\right\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p) \quad (7)$$

How to ensure RHS is small ?

$$q_k = \arg\min_{p} \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^{n} \|g_k^j\|_2} \right)_{i=1}^{n}$$

$$\Rightarrow \quad \sigma^2(x_k, p) \leq \frac{2}{n^2} \sum_{i=1}^{n} \frac{\left\| \nabla f_i(x_k) - g_k^i \right\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p) \quad (7)$$

How to ensure RHS is small ?

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \|g_k^j\|_2} \right)_{i=1}^n$$

$$\Rightarrow \quad \sigma^2(x_k, p) \leq \frac{2}{n^2} \sum_{i=1}^n \frac{\|\nabla f_i(x_k) - g_k^i\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p) \quad (7)$$

$\Rightarrow$ **Mixing Distribution Method**

Mixing Distribution Method

θ_k in (0, 1]

$$p_k = (1 - \theta_k)q_k + \frac{\theta_k}{n}$$

Greedy strategy

Uniformly sampling

Mixing Distribution Method

θ_k in (0, 1]

$$p_k = (1 - \theta_k) q_k + \frac{\theta_k}{n}$$

Greedy strategy

Uniformly sampling

Mixing Distribution Method

θ_k in (0, 1]

$$p_k = (1 - \theta_k)q_k + \frac{\theta_k}{n}$$

Greedy strategy

Uniformly sampling

Mixing Distribution Method

to bound conditional variance (Lemma 4.1)

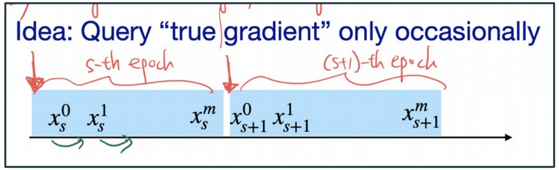
$$p_k = (1 - \theta_k) q_k + \frac{\theta_k}{n}$$

$$\sigma^2(x_k, p) \leq \frac{2}{n^2} \sum_{i=1}^{n} \frac{\left\| \nabla f_i(x_k) - g_k^i \right\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p) \quad (7)$$

What time to update g_k^i

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \|g_k^j\|_2}\right)_{i=1}^n$$

only update when the index i_k is drawn from the uniform mixture component.
(need in Lemma 4.1)



Ref: Lecture 7-Stochastic Gradient Descent and Variance Reduction

Part II SRG Algorithm

SRG algorithm

Algorithm 1 SRG

1: **Parameters:** step sizes $(\alpha_k)_{k=0}^{\infty} > 0$, mixture coefficients $(\theta_k)_{k=0}^{\infty} \in (0, 1]$
2: **Initialization:** $x_0 \in \mathbb{R}^d$, $(\|g_0^i\|_2)_{i=1}^n \in \mathbb{R}^n$
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: $\quad p_k = (1 - \theta_k)q_k + \theta_k/n$ $\quad$ $\{q_k$ is defined in (6)$\}$
5: $\quad b_k \sim \text{Bernoulli}(\theta_k)$
6: $\quad$ **if** $b_k = 1$ **then** $i_k \sim 1/n$ **else** $i_k \sim q_k$
7: $\quad x_{k+1} = x_k - \alpha_k \frac{1}{np_k^{i_k}} \nabla f_{i_k}(x_k)$
8: $\quad \|g_{k+1}^i\|_2 = \begin{cases} \|\nabla f_i(x_k)\|_2 & \text{if } b_k = 1 \text{ and } i_k = i \\ \|g_k^i\|_2 & \text{otherwise} \end{cases}$
9: **end for**

Mixing distribution

Ref: (SRG) https://proceedings.mlr.press/v162/hanchi22a.html

SRG algorithm

Algorithm 1 SRG

Parameters: step sizes $(\alpha_k)_{k=0}^{\infty} > 0$, mixture coefficients $(\theta_k)_{k=0}^{\infty} \in (0, 1]$

Initialization: $x_0 \in \mathbb{R}^d$, $(\|g_0^i\|_2)_{i=1}^n \in \mathbb{R}^n$

for $k = 0, 1, 2, \ldots$ **do**

  $p_k = (1 - \theta_k)q_k + \theta_k/n$ $\{q_k$ is defined in (6)$\}$

  $b_k \sim \text{Bernoulli}(\theta_k)$

  if $b_k = 1$ **then** $i_k \sim 1/n$ **else** $i_k \sim q_k$

  $x_{k+1} = x_k - \alpha_k \frac{1}{np_k^{i_k}} \nabla f_{i_k}(x_k)$

  $$\left\|g_{k+1}^i\right\|_2 = \begin{cases} \|\nabla f_i(x_k)\|_2 & \text{if } b_k = 1 \text{ and } i_k = i \\ \left\|g_k^i\right\|_2 & \text{otherwise} \end{cases}$$

end for

Mixing distribution

SRG update

Ref: (SRG) https://proceedings.mlr.press/v162/hanchi22a.html

SRG algorithm

Algorithm 1 SRG

Parameters: step sizes $(\alpha_k)_{k=0}^{\infty} > 0$, mixture coefficients $(\theta_k)_{k=0}^{\infty} \in (0, 1]$

Initialization: $x_0 \in \mathbb{R}^d$, $(\|g_0^i\|_2)_{i=1}^{n} \in \mathbb{R}^n$

for $k = 0, 1, 2, \ldots$ **do**

  $p_k = (1 - \theta_k) q_k + \theta_k / n$ $\{q_k$ is defined in (6)$\}$

  $b_k \sim \text{Bernoulli}(\theta_k)$

  if $b_k = 1$ **then** $i_k \sim 1/n$ **else** $i_k \sim q_k$

  $x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$

  $$\|g_{k+1}^i\|_2 = \begin{cases} \|\nabla f_i(x_k)\|_2 & \text{if } b_k = 1 \text{ and } i_k = i \\ \|g_k^i\|_2 & \text{otherwise} \end{cases}$$

end for

Mixing distribution

SRG update

Update gradient norms table

Ref: (SRG) https://proceedings.mlr.press/v162/hanchi22a.html

Part III SRG+ Motivation

Could we do SRG better ?

$$\boxed{\sigma^2(x_k, p) \leq \frac{2}{n^2} \sum_{i=1}^{n} \frac{\left\|\nabla f_i(x_k) - g_k^i\right\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p) \quad (7)}$$

Motivation 1: use shaper bound ? Yes.

Could we do SRG better ?

$$\sigma^2(x_k, p) \leq \frac{2}{n^2} \sum_{i=1}^{n} \frac{\left\| \nabla f_i(x_k) - g_k^i \right\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p) \quad (7)$$

Motivation 1: use shaper bound ? Yes.

$$x_{k+1} = x_k - \alpha_k \frac{1}{np_k^{i_k}} \nabla f_{i_k}(x_k)$$

$$\left\| g_{k+1}^i \right\|_2 = \begin{cases} \left\| \nabla f_i(x_k) \right\|_2 & \text{if } b_k = 1 \text{ and } i_k = i \\ \left\| g_k^i \right\|_2 & \text{otherwise} \end{cases}$$

Motivation 2: decoupling the table update and gradient update ? (maximal couplings)

Choose a better Distribution

$$p'_k = (1 - \eta_k - \theta_k)q_k + \eta_k v + \frac{\theta_k}{n} \qquad (11)$$

$$v = \left(\frac{L_i}{n\bar{L}}\right)_{i=1}^{n} \qquad (12)$$

Choose a better Distribution

$$p'_k = (1 - \eta_k - \theta_k)q_k + \eta_k v + \frac{\theta_k}{n} \qquad (11)$$

$$v = \left(\frac{L_i}{n\bar{L}}\right)_{i=1}^{n} \qquad (12)$$

Q: Why we use such distribution ?

Choose a better Distribution

$$p'_k = (1 - \eta_k - \theta_k)q_k + \eta_k v + \frac{\theta_k}{n} \quad (11)$$

$$v = \left(\frac{L_i}{n\bar{L}}\right)_{i=1}^{n} \quad (12)$$

Q: Why we use such distribution ?

$$\sigma^2(x_k, p) \leq \frac{3}{n^2}\sum_{i=1}^{n}\frac{L_i}{p^i}\langle\nabla f_i(x_k) - \nabla f_i(x^*), x_k - x^*\rangle$$
$$+ \frac{3}{n^2}\sum_{i=1}^{n}\frac{1}{p^i}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + 3\tilde{\sigma}^2(x_k, p) \quad (10)$$

Choose a better Distribution

$$p'_k = (1 - \eta_k - \theta_k)q_k + \eta_k v + \frac{\theta_k}{n} \qquad (11)$$

$$v = \left(\frac{L_i}{n\bar{L}}\right)_{i=1}^{n} \qquad (12)$$

Q: Why we use such distribution ?

$$\sigma^2(x_k, p) \le \frac{3}{n^2}\sum_{i=1}^{n}\frac{L_i}{p^i}\langle \nabla f_i(x_k) - \nabla f_i(x^*), x_k - x^* \rangle$$
$$+ \frac{3}{n^2}\sum_{i=1}^{n}\frac{1}{p^i}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + 3\tilde{\sigma}^2(x_k, p) \qquad (10)$$

L_i ∝ P^i

Choose a better Distribution

$$p'_k = (1 - \eta_k - \theta_k) q_k + \eta_k v + \frac{\theta_k}{n} \qquad (11)$$

$$v = \left(\frac{L_i}{n\bar{L}} \right)_{i=1}^{n} \qquad (12)$$

Q: Why we use such distribution ?

$$\sigma^2(x_k, p) \le \frac{3}{n^2} \sum_{i=1}^{n} \frac{L_i}{p^i} \langle \nabla f_i(x_k) - \nabla f_i(x^*), x_k - x^* \rangle$$
$$+ \frac{3}{n^2} \sum_{i=1}^{n} \frac{1}{p^i} \left\| g_k^i - \nabla f_i(x^*) \right\|_2^2 + 3\tilde{\sigma}^2(x_k, p) \qquad (10)$$

L_i ∝ P^i

uniformly sample

Choose a better Distribution

$$p'_k = (1 - \eta_k - \theta_k)q_k + \eta_k v + \frac{\theta_k}{n} \quad (11)$$

$$v = \left(\frac{L_i}{n\bar{L}}\right)_{i=1}^{n} \quad (12)$$

Q: Why we use such distribution ?

$$\sigma^2(x_k, p) \leq \frac{3}{n^2}\sum_{i=1}^{n}\frac{L_i}{p^i}\langle \nabla f_i(x_k) - \nabla f_i(x^*), x_k - x^* \rangle$$
$$+ \frac{3}{n^2}\sum_{i=1}^{n}\frac{1}{p^i}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + 3\tilde{\sigma}^2(x_k, p) \quad (10)$$

L_i ∝ P^i

uniformly
sample

q_k

Part IV SRG+ Algorithm

SRG+ algorithm

Algorithm 2 SRG+

Parameters: step sizes $(\alpha_k)_{k=0}^{\infty} > 0$, mixture coefficients $(\theta_k)_{k=0}^{\infty} \in (0, 1]$

Initialization: $x_0 \in \mathbb{R}^d$, $(\|g_0^i\|_2)_{i=1}^n \in \mathbb{R}^n$

for $k = 0, 1, 2, \ldots$ **do**

$p_k = (1 - \theta_k)q_k + \theta_k v$

$\{q_k$ is given by (6), v is given by (12)$\}$

$b_k \sim \text{Bernoulli}(\theta_k)$

if $b_k = 1$ **then** $(i_k, j_k) \sim \pi$ **else** $i_k \sim q_k$

$\{\pi$ maximally couples $(v, 1/n)\}$

$x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$

$$\left\|g_{k+1}^j\right\|_2 = \begin{cases} \|\nabla f_j(x_k)\|_2 & \text{if } b_k = 1 \text{ and } j = j_k \\ \left\|g_k^j\right\|_2 & \text{otherwise} \end{cases}$$

end for

Better Distribution

Ref: (SRG+) https://proceedings.mlr.press/v162/hanchi22a.html

SRG+ algorithm

Algorithm 2 SRG+

Parameters: step sizes $(\alpha_k)_{k=0}^{\infty} > 0$, mixture coefficients $(\theta_k)_{k=0}^{\infty} \in (0,1]$

Initialization: $x_0 \in \mathbb{R}^d$, $(\|g_0^i\|_2)_{i=1}^n \in \mathbb{R}^n$

for $k = 0, 1, 2, \ldots$ **do**

$p_k = (1-\theta_k)q_k + \theta_k v$ ← Better Distribution

$\{q_k$ is given by (6), v is given by (12)$\}$

$b_k \sim \text{Bernoulli}(\theta_k)$

if $b_k = 1$ **then** $(i_k, j_k) \sim \pi$ **else** $i_k \sim q_k$

$\{\pi$ maximally couples $(v, 1/n)\}$ ← Decouple update

$x_{k+1} = x_k - \alpha_k \frac{1}{np_k^{i_k}} \nabla f_{i_k}(x_k)$

$$\left\|g_{k+1}^j\right\|_2 = \begin{cases} \|\nabla f_j(x_k)\|_2 & \text{if } b_k = 1 \text{ and } j = j_k \\ \left\|g_k^j\right\|_2 & \text{otherwise} \end{cases}$$

end for

Ref: (SRG+) https://proceedings.mlr.press/v162/hanchi22a.html

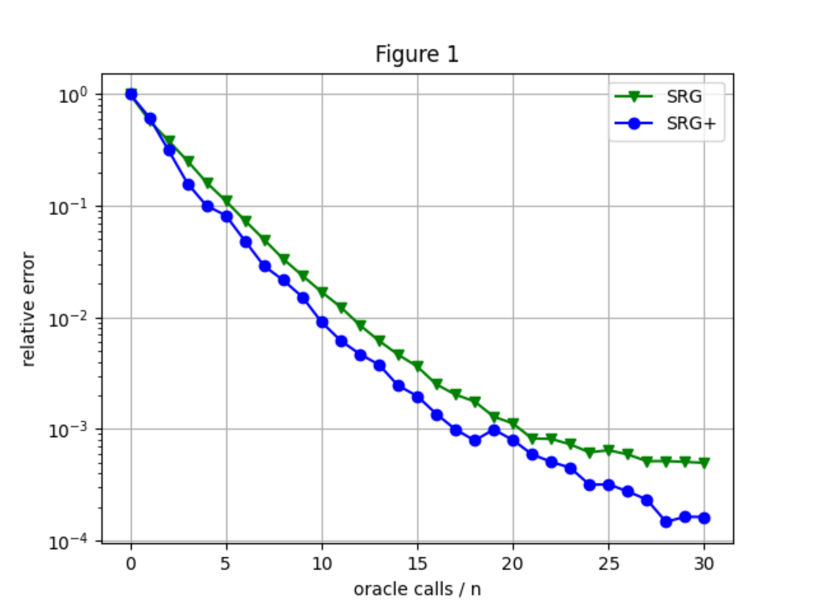
Part V Comparison

Comparation

Item	SGD	SRG	SRG+
Complexity	$O\left(\kappa_{\max}+\frac{\sigma^2}{\mu^2\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)$	$O\left(n+\sqrt{\frac{n\sigma_*^2}{\mu^2\varepsilon}}+\kappa_{\max}+\frac{\sigma_*^2}{\mu^2\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)$	$O\left(n+\sqrt{\frac{n\sigma_*^2}{\mu\varepsilon}}+\bar{\kappa}+\frac{\sigma_*^2}{\mu^2\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)$
Gradient Computation times	1	1	1 or 2

Part VI Experiment

Experiment 1

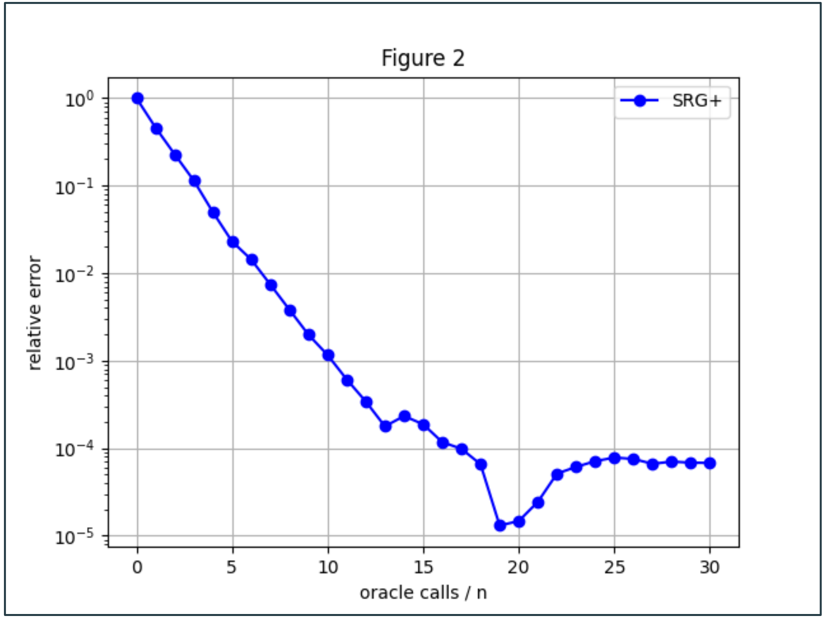


Objective function
- f_i(x) = (x - a_i) ^ 2 / 2
- a_i = 1 if i = n - 1
- a_i = 0 else

hyperparam.
mixture coefficient theta= 1/2
n = 20

Optimal point
x^\star = 1 / n

Experiment 2



Objective function
- f_i(x) = L_i(x - a_i)^2 / 2
- a_i = 0 if i = 1, 2,..., n - 1
- a_i = 1 if i = n
- L_1 = n - 1
- L_n = 1/n
- L_i = n(n - 1) / [n(n-2)]
- L^\bar = 1
- L_max = n - 1

hyperparam.
the same

Optimal point
x^\star = 1 / (n^2)

Part VII Idea of Prove

Recall : Nesterov in Lecture 5

Lyapunov Function for Solving ODEs

To motivate the proof idea, let's take a slightly simpler ODE as

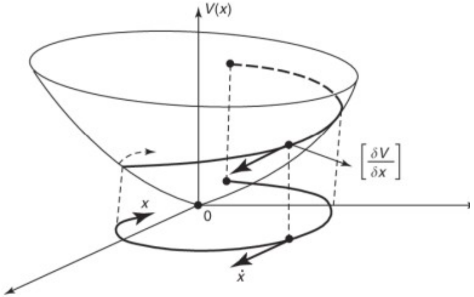
$$\dot{X}(\tau) + \nabla f(X(\tau)) = 0, \quad \tau > 0$$

Construct $\mathscr{E}$ Lyapunov function (or an energy function)

$$V(t) := \big(f(X(t)) - f(x^*)\big) + \frac{\|X(t) - x^*\|^2}{2}$$

If we can show that $V(t)$ is decreasing with t, then we have a convergence rate

$$f(X(t)) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2t}$$



Lyapunov function

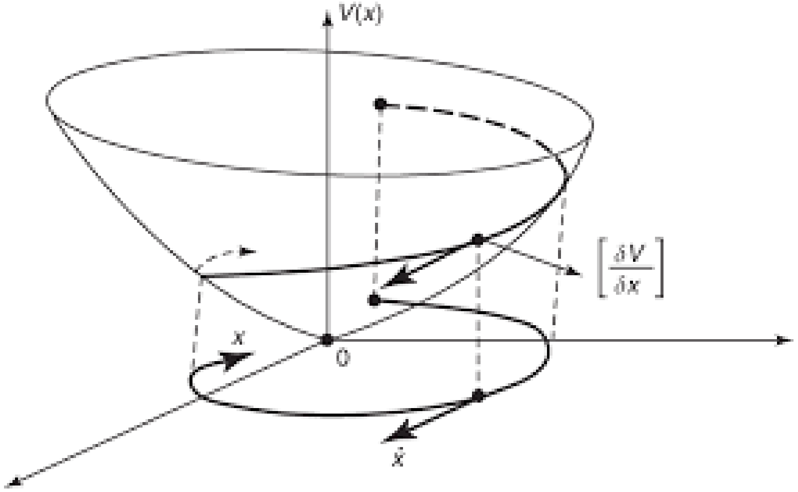
$$T^k := \frac{\alpha_k}{\theta_k} \frac{a}{L_{\max}} \sum_{i=1}^{n} \|g_k^i - \nabla f_i(x^*)\|_2^2 + \|x_k - x^*\|_2^2$$

1. Decrease monotically

$$\mathbb{E}\left[T^{k+1}\right] \leq (1-\rho_k)\mathbb{E}\left[T^k\right] + (1+2\theta_k)6\alpha_k^2\sigma_*^2$$

2. Convergence

$$O\left(n + \sqrt{\frac{n\sigma_*^2}{\mu^2\varepsilon}} + \kappa_{\max} + \frac{\sigma_*^2}{\mu^2\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)$$



Intermediate Lemma

Lemma 4.1. *Let $k \in \mathbb{N}$ and suppose that $(g_k^i)_{i=1}^n$ evolves as in Algorithm 1. Taking expectation with respect to (b_k, i_k), conditional on $(b_t, i_t)_{t=0}^{k-1}$, we have:*

$$\mathbb{E}\left[\sum_{i=1}^{n} \left\|g_{k+1}^i - \nabla f_i(x^*)\right\|_2^2\right] \le 2\theta_k L_{max}\left[F(x_k) - F(x^*)\right]$$
$$+ \left(1 - \frac{\theta_k}{n}\right)\sum_{i=1}^{n}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2$$

Lemma 4.2. *Let $k \in \mathbb{N}$ and assume that $\theta_k \in (0, 1/2]$. Taking expectation with respect to (b_k, i_k), conditional on $(b_t, i_t)_{t=0}^{k-1}$, we have, for all $\beta, \gamma, \delta, \eta > 0$:*

$$\mathop{\mathbb{E}}_{i_k \sim p_k}\left[\left\|\frac{1}{np_k^{i_k}}\nabla f_{i_k}(x_k)\right\|_2^2\right] \le \frac{2D_1 L_{max}}{\theta_k}\left[F(x_k) - F^*\right]$$
$$+\frac{D_2}{\theta_k n}\sum_{i=1}^{n}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + D_3(1 + 2\theta_k)\sigma_*^2$$

Thm : Convergence

Theorem 4.3. *Suppose that $(x_k, (g_k^i)_{i=1}^n)$ evolves according to Algorithm 1. Further, assume that for all $k \in \mathbb{N}$: (i) α_k/θ_k is non-increasing. (ii) $\theta_k \in (0, 1/2]$. (iii) $\alpha_k \leq \theta_k/12L_{max}$. Then:*

$$\mathbb{E}\left[T^{k+1}\right] \leq (1-\rho_k)\mathbb{E}\left[T^k\right] + (1+2\theta_k)6\alpha_k^2\sigma_*^2$$

for all $k \in \mathbb{N}$, and where:

$$\rho_k := \min\left\{\frac{\theta_k}{12n}, \alpha_k\mu\right\}$$

Corollary 4.4. *Suppose that $(x_k, (g_k^i)_{i=1}^n)$ evolves according to Algorithm 1 with a constant mixture coefficient $\theta_k = \theta \in (0, 1/2]$ and a constant step size $\alpha_k = \alpha \leq \theta/12L_{max}$. Then for any $k \in \mathbb{N}$:*

$$\mathbb{E}\left[T^k\right] \leq (1-\rho)^k T^0 + (1+2\theta)\frac{6\alpha^2\sigma_*^2}{\rho}$$

where $\rho = \rho_k$ is as defined in Theorem 4.3. For any $\varepsilon > 0$ and $\theta \in (0, 1/2]$, choosing:

$$\alpha = \min\left\{\frac{\theta}{12L_{max}}, \frac{\varepsilon\mu}{(1+2\theta)12\sigma_*^2}, \sqrt{\frac{\theta}{1+2\theta}\frac{\varepsilon}{144n\sigma_*^2}}\right\}$$

and:

$$k \geq \max\left\{\frac{12n}{\theta}, \frac{1}{\alpha\mu}\right\}\log\left(\frac{2T^0}{\varepsilon}\right)$$

guarantees $\mathbb{E}\left[\|x_k - x^*\|_2^2\right] \leq \varepsilon$

Reference

Reference

- Stochastic Reweighted Gradient Descent:
https://proceedings.mlr.press/v162/hanchi22a.html

- Estimating Convergence of Markov chains with L-Lag Couplings:
https://www.semanticscholar.org/paper/Estimating-Convergence-of-Markov-chains-with-L-Lag-Biswas-Jacob/5363fdc254ee28230a7eef1e4aec642ceb8a7749

- Importance Sampling Explained End-to-End:
https://medium.com/@liuec.jessica2000/importance-sampling-explained-end-to-end-a53334cb330b

- Lecture 7-Stochastic Gradient Descent and Variance Reduction

Thank you for your attention