**The Problem**

Goal: minimize the finite-sum problem

\[ \min_{x \in \mathbb{R}^d} \left\{ P(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\} \]

- each \( f_i(x) \) is convex and has Lipschitz continuous gradient with parameter \( L \)

Two special cases

- **CASE-A**: function \( P \) is \( \mu \) strongly convex
- **CASE-B**: each \( f_i \) is \( \mu \) strongly convex

**The Algorithm**

**Algorithm: SARAH vs. SVRG**

1. choose \( x_0 \)
2. for \( s = 0, 1, 2, \ldots \) do
3. \( \hat{x}_0 = x_s \)
4. for \( t = 0, 1, 2, \ldots, m \) do
5. choose random \( i_t \sim U([1, 2, \ldots, d]) \)
6. compute stochastic gradient \( \nu_t \)
7. update \( \hat{x}_{t+1} = \hat{x}_t - \eta \nu_t \)
8. end for
9. choose \( x_{s+1} \) randomly from \( \{\hat{x}_0, \ldots, \hat{x}_m\} \)
10. end for

**Stochastic gradient of SARAH**

This gradient is defined recursively as

\[ \nu_0 = \nabla P(\hat{x}_0) \]

\[ \nu_t = \nu_{t-1} + \nabla f_{i_t}(\hat{x}_t) - \nabla f_{i_t}(\hat{x}_{t-1}) \]

Remarks:
1. \( E[\nu_t] \neq \nabla P(\hat{x}_t) \), but \( E[\nu_t] = \nabla P(\hat{x}_t) \)
2. No need for extra storage as in SAG/SAGA!

**Stochastic gradient of SVRG**

\[ \nu_t = \nabla f_{i_t}(\hat{x}_t) - \nabla f_{i_t}(\hat{x}_0) + \nabla P(\hat{x}_0) \]

**Convergence Analysis**

**Theorem:** (CASE-A). For \( \eta \in (0, 2/L) \) it holds

\[ E[\|\nabla P(x_s)\|^2] \leq \left( \frac{1}{\eta(\mu + 1)} + \frac{n L \mu^2}{2 - \mu} \right)^{\frac{s}{2}} \left\| \nabla P(x_0) \right\| \]

Remarks:
- This is **better** than convergence of SVRG.
- **CASE-B** have a slightly better convergence rate.
- In paper we also analyze convex case.
- We have extented it to non-convex case \([1]\).

**Practical variant SARAH+**

Both SVRG and SARAH need \( m \) as an input! The performance is very sensitive on this choice.

**Facts:**
- **SARAH** is converging in each outerloop.
- It would not be efficient to take many tiny steps.

**SARAH+ Algorithm**

Let’s break the inner loop when \( \|\nu_t\|^2 > \gamma \|\nu_0\|^2 \) for some \( \gamma \in [0, 1] \) (usually \( \gamma = 0.1 \) is a good choice).

**Benefit:** No need to tune parameter \( m \)!