Takeaways:

1. Template of Variance Reduction analysis:
2. "Warm start": a good starting point matters (engineering trick)
3. Trade-offs (amount of computation VS performance) and how to benefit from them
   A. switching
   B. mixing (if time allows)

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Lecture notation $F(w,b) = \frac{1}{n} \sum_{i=1}^{n} f_i(w,b) + \psi(w)$

$f_i$: depend on $i^{th}$ sample $(x_i, y_i)$

Vector $x$: $x^t$ for iterates, $x_i$ for component

Scalar $\alpha$: $\alpha_t$ for iterates

GD: $x^t = x^{t-1} - \alpha_t \nabla F(x^{t-1})$

SGD: for $t = 1, 2, \cdots$ $\quad t > n$

$$i_t \leftarrow \text{Unif}\{1, 2, \cdots, n\}$$

$$x^t = x^{t-1} - \alpha_t \nabla f_{i_t}(x^{t-1})$$

1st order $\nabla F$

$F = (f_1 + f_2) \cdot \frac{1}{2}$

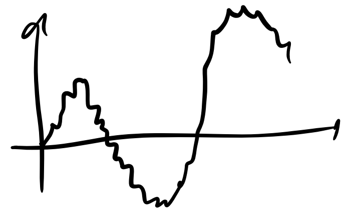
$x^*$ $\quad \nabla f_1(x^*) = 1$

$\nabla f_2(x^*) = -1$

$$\mathbb{E}\left[\nabla f_{i_t}(x^{t-1})\right] = \nabla F(x^{t-1}) \quad ?$$

$\alpha_t \approx$ constant

$$\mathbb{E}[\nabla f_{i_t}(x)] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x) = \nabla F(x^{t-1})$$



$\frac{1}{t}$

$\frac{1}{\sqrt{t}}$

VR template:

Consider an estimator $X$ for parameter $\theta \in \mathbb{R}^n$, where $X = \nabla f_{i_t}(x)$ and $\theta = \nabla F(x)$

$X$ is Unbiased if $\mathbb{E}X = \theta$

$$\text{Var}(X) := \mathbb{E}[(X - \theta)^2]$$

$Z := X - Y$. $X$ unbiased, $\mathbb{E}Y = 0$ $\Rightarrow$ $Z$ is unbiased $\mathbb{E}Z = \mathbb{E}X = \theta$

"Recall" $\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$

Fact: $\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y)$ ☆

$\mathrm{Var}(Z) \le \mathrm{Var}(X) \Leftrightarrow \mathrm{Cov}(X,Y) \ge \frac{1}{2}\mathrm{Var}(Y)$ $\ge 0$

$X, Y$ "positively correlated"

Example (midterm)

estimator $X$ for parameter $\theta$. $\mathbb{E}X = \theta = 1$

$Z := X - Y$. $\mathrm{Var}(Z) \le \mathrm{Var}(X)$

$\mathrm{Var}(X) = 0.5$, $\mathrm{Var}(Y) = 0.1$, $\mathbb{E}[Y] = 0$

$\mathrm{Cov}(X,Y) \ge \frac{1}{2}\mathrm{Var}(Y) = 0.05$

---

SAG (Stochastic averaged gradient, 2013)

Maintain a table of $g_i$ $i = 1, 2, \cdots, n$ ($g$ for gradient)

Initialize $x^0$, $g_i^0 = \nabla f_i(x^0)$ $\forall i$

for $t = 1, 2, \cdots$ do

$i_t \leftarrow \mathrm{Unif}\{1, \cdots, n\}$

$$g_{i_t}^t = \nabla f_i(x^{t-1}), \quad g_i^t = g_i^{t-1} \text{ for } i \neq i_t$$

$$x^t = x^{t-1} - \alpha_t \frac{1}{n} \sum_{i=1}^{n} g_i^t$$

| g | 1 | 2 | 3 | ... | n |
|---|---|---|---|---|---|
| Initialize | $\nabla f_1(x^0)$ | $\nabla f_2(x^0)$ | $\nabla f_3(x^0)$ | ... | $\nabla f_n(x^0)$ |
| | -//- | $\nabla f_2(x^{t-1})$ | -//- | | -//- |

$$g_1^0 := \nabla f_1(x^0)$$

Remark: ① Each $f_i$ contribute a part to gradient estimate

② $x^0$ — How to choose

③ $\frac{1}{n} \sum_{i=1}^{n} g_i^t$ — too much to average?

$$x^t = x^{t-1} - \alpha_t \left( \frac{1}{n} \sum_{i \neq i_t} g_i^t + \frac{1}{n} g_{i_t}^t \right)$$

$$= x^{t-1} - \alpha_t \left( \underbrace{\frac{1}{n} \sum_{i \neq i_t} g_i^{t-1} + \frac{1}{n} g_{i_t}^{t-1}} - \frac{1}{n} g_{i_t}^{t-1} + \underline{\frac{1}{n} g_{i_t}^t} \right)$$

old average

$$= x^{t-1} - \alpha_t \left( \frac{1}{n} g_{i_t}^t - \frac{1}{n} g_{i_t}^{t-1} + \frac{1}{n} \sum_{i=1}^{n} g_i^{t-1} \right)$$

$\nabla f_{i_t}(x^t)$

$\parallel$

$$\underbrace{g_{i_t}^t}_{} - \underbrace{\left( g_{i_t}^{t-1} - \sum_{i=1}^{n} g_i^{t-1} \right)}_{}$$

$$Z := X - Y$$

$\mathbb{E}X = \nabla F(x^t)$ — unbiased

$\mathbb{E}Y \neq 0 \Rightarrow Z$ is biased

$X, Y$ correlated

As $t \to \infty$, $X - Y \to 0$

$$g_{i_t}^t \approx g_{i_t}^{t-1}$$

$$\sum_{i=1}^{n} g_i \approx 0$$

SAGA (2014)

If $\mathbb{E}Y=0$, $Z$ is unbiased

$$\underbrace{g_{i_t}^{t}}_{X} - \underbrace{\left(g_{i_t}^{t-1} - \frac{1}{n}\sum_{i=1}^{n} g_i^{t-1}\right)}_{Y}$$

$i_t \leftarrow \mathrm{Unif}\{1, \cdots, n\}$

$\mathbb{E}Y = 0$ because $i_t$ chosen u.a.r from $\{1, 2, \cdots, n\}$

Exercise: Var SAGA $n^2$ time larger than SAG

Note: SAGA still needs a careful choice of $x^0$

$\boxed{\text{Choose } x^{00}\text{, run 1 epoch of SGD to give } x^0}$

"Warm start"

(2013)

Stochastic variance reduction gradient (SVRG)

Template: $\underbrace{g_{i_t}^t}_{X} - \underbrace{\nabla f_{i_t}(x^{old}) + \nabla F(x^{old})}_{Y}$

Exercise: ① $\mathbb{E}Y = 0$

② $X, Y$ positively correlated

$X - Y \to 0$, because . . .

③ Need to compute $\nabla F(x^{old})$ ocassionally

Algorithm: Initialize $x$

$F = \frac{1}{n}\sum_{i=1}^{n} f_i$

Until convergence do

$x^{old} \leftarrow$ newest iterate $x$

$x^0 \leftarrow x^{old}$, Compute $\nabla F(x^{old})$

for $t=0,1,\cdots,m-1$ do

$i_t \leftarrow \text{Unif}\{1,\cdots,n\}$

$$\hat{\nabla}^t = \nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{old}) + \nabla F(x^{old})$$

$$x^{t+1} = x^t - \eta \hat{\nabla}^t$$

Remark: ① Constant $\eta$

② Small $m \rightarrow$ Compute $\nabla F$ more often

③ $m \gg n$

# Computation SVRG $\approx$ SGD

SARAH:

Initialize $x$

Until convergence do

$\cancel{x^{old} \leftarrow \text{newest iterate } x}$ $\quad \hat{\nabla}^0 = \nabla F(\text{newest } x)$

$\cancel{x^0 \leftarrow x^{old}, \text{ Compute } \nabla F(x^{old})}$

for $t=0,1,\cdots,m-1$ do

$i_t \leftarrow \text{Unif}\{1,\cdots,n\}$

$$\hat{\nabla}^t = \nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1}) + \hat{\nabla}^{t-1}$$

$$x^{t+1} = x^t - \eta \hat{\nabla}^t$$

Note that I misspoke in the lecture - SARAH still computes full gradient, the same as SVRG. It just doesn't use it again and again in the inner loop update, but use it as an occasional correction

PAGE: Each step do (2020)

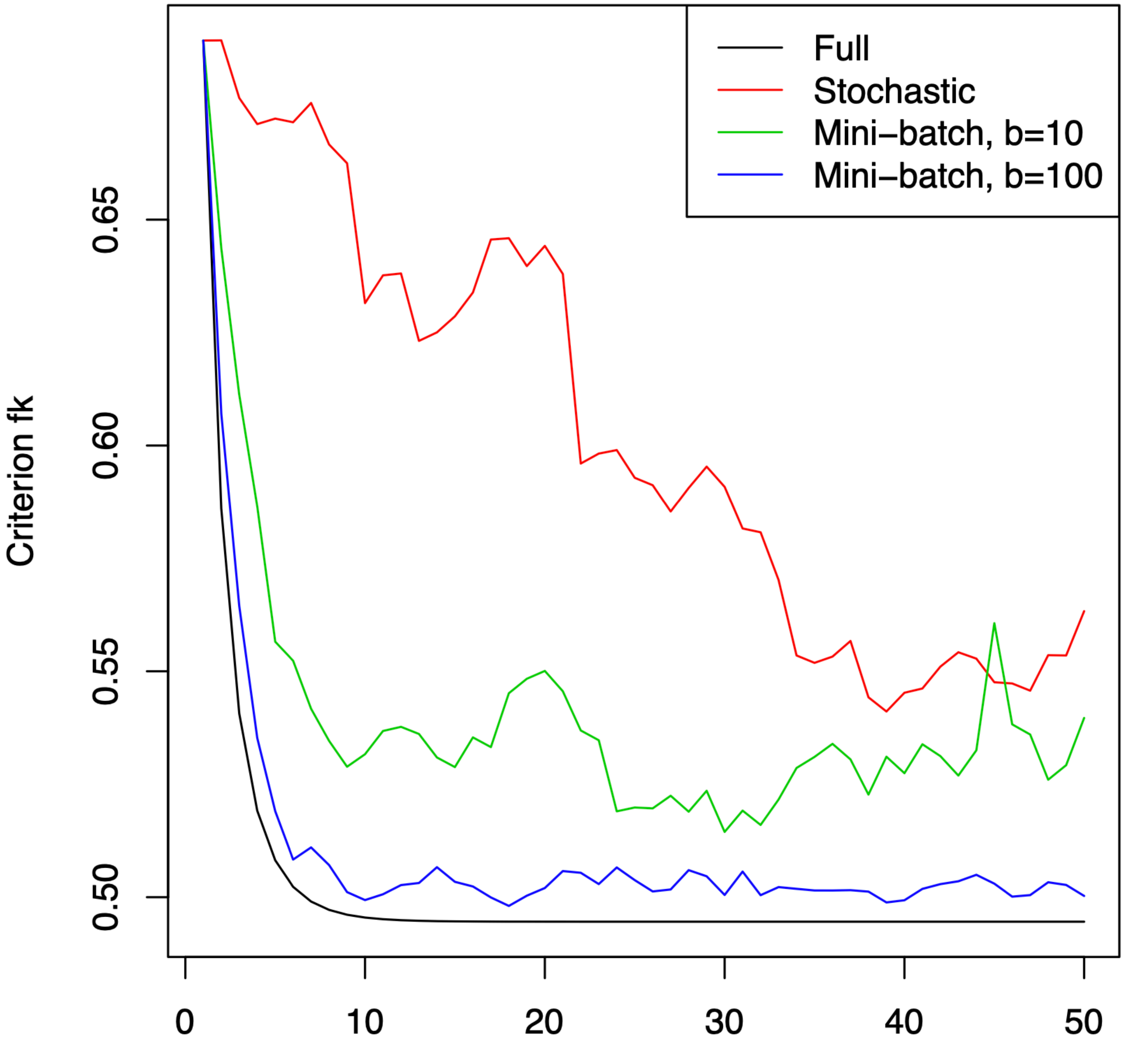
$b'$ - mini batch SARAH w.p. $p_t$

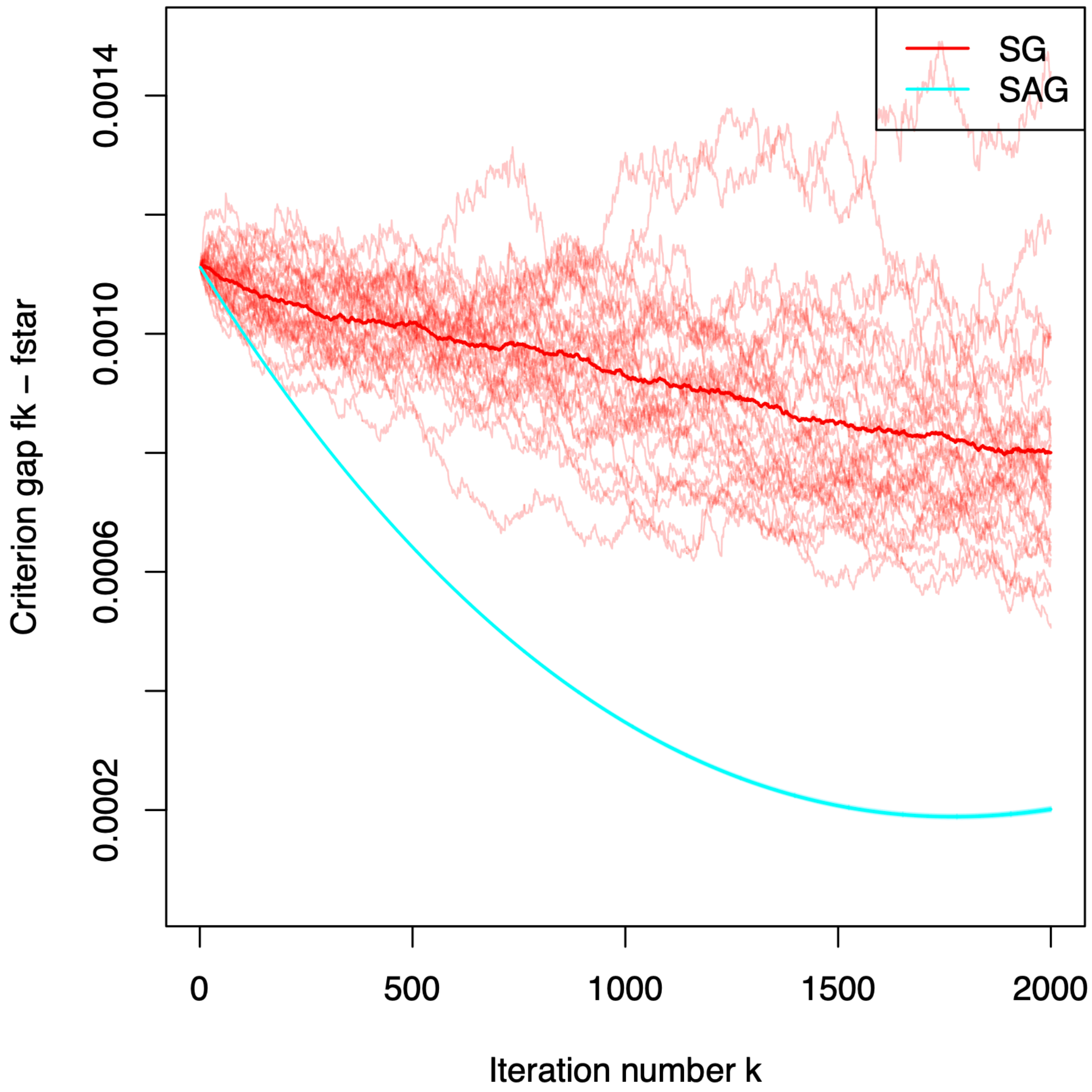
$b$ - mini batch SGD w.p. $1-p_t$

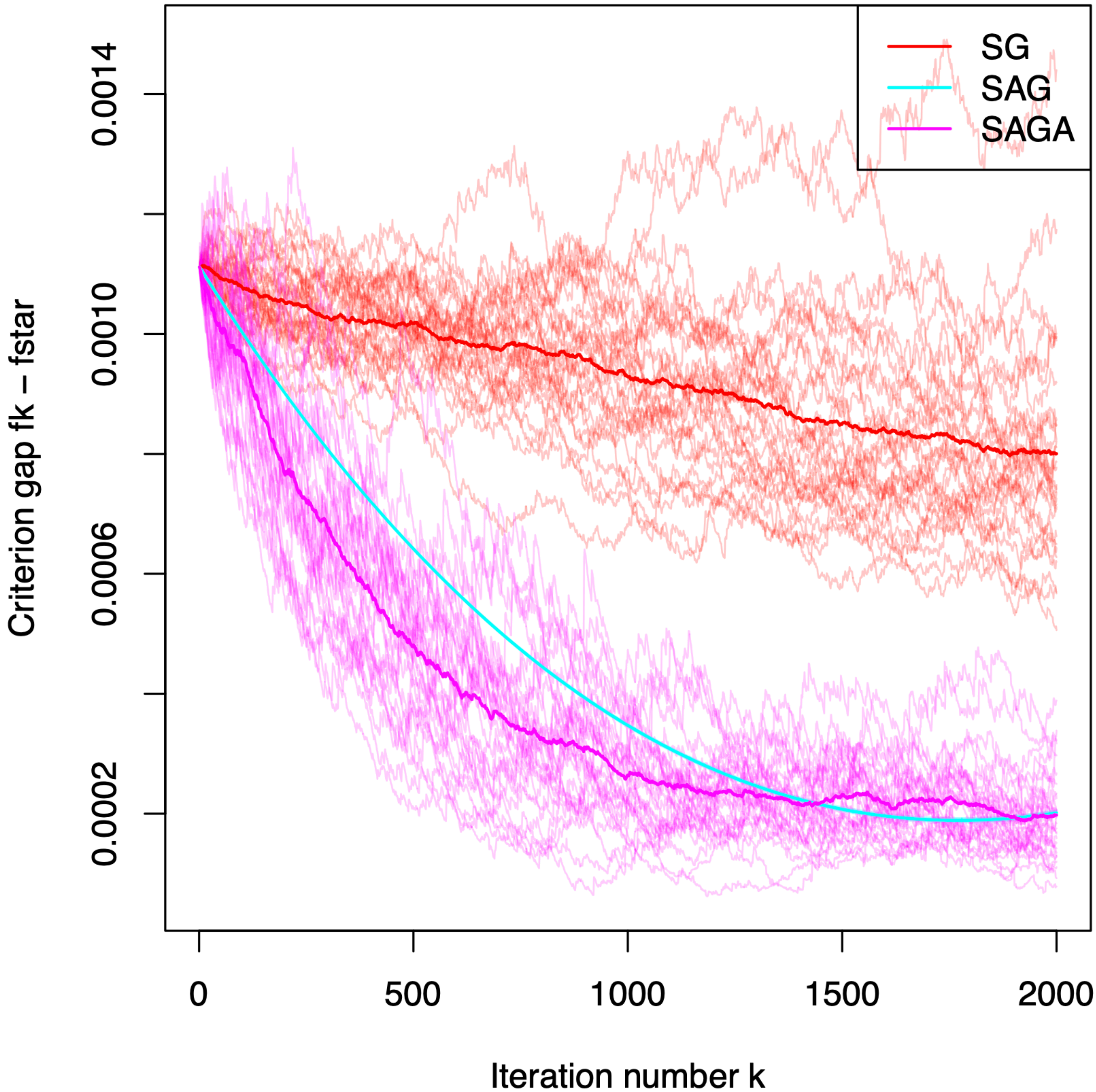
$$p_t \equiv \frac{b}{b+b'}, \quad b' \ll b.$$

The following pictures are from http://www.stat.cmu.edu/~ryantibs/convexopt/lectures/modern-sgd.pdf.

This note is prepared bas http://www.stat.cmu.edu/~ryantibs/convexopt/lectures/modern-sgd.pdf and http://www.princeton.edu/~yc5/ele522_optimization/lectures/variance_reduction.pdf




SG
SAG
Criterion gap fk − fstar
0.0002
0.0006
0.0010
0.0014
0
500
1000
1500
2000
Iteration number k



SG
SAG
SAGA
Criterion gap fk − fstar
0.0014
0.0010
0.0006
0.0002
0
500
1000
1500
2000
Iteration number k