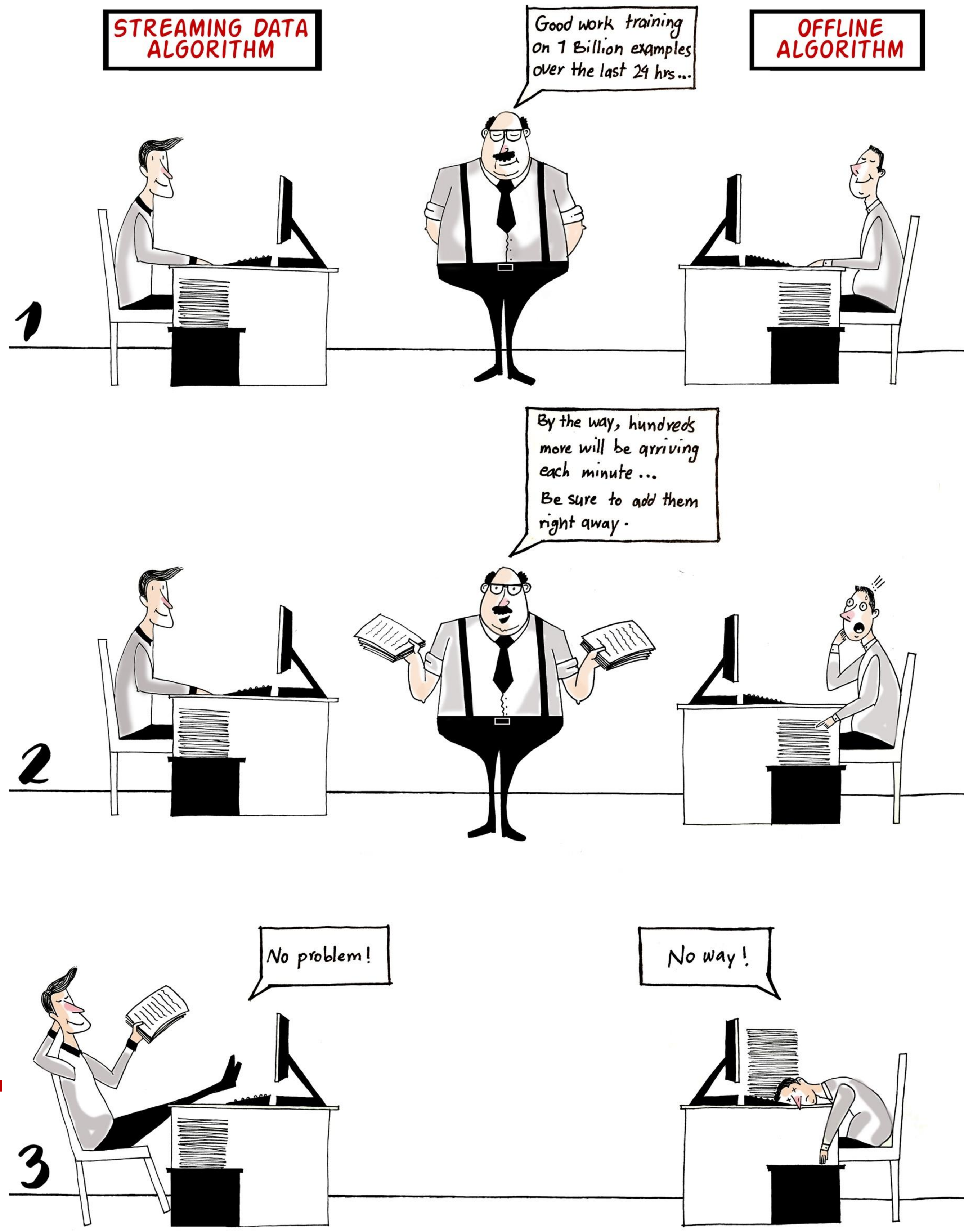
Variance-Reduced SGD on Streaming Data

Ellango Jothimurugesan[1], Ashraf Tahmasbi[2], Phillip B. Gibbons[1], Srikanta Tirthapura[2]
[1] Carnegie Mellon University, [2] Iowa State University

Problem Statement



Streaming data setting: training data arrives incrementally

- At each time $i = 1, 2, \ldots$, a set X_i of new points arrives. The number of new points varies with each time and is drawn from an unknown distribution.

An offline algorithm cannot handle incremental data and must recompute the model from scratch each time

Goal: A streaming data algorithm that efficiently updates the model, with accuracy similar to an offline algorithm

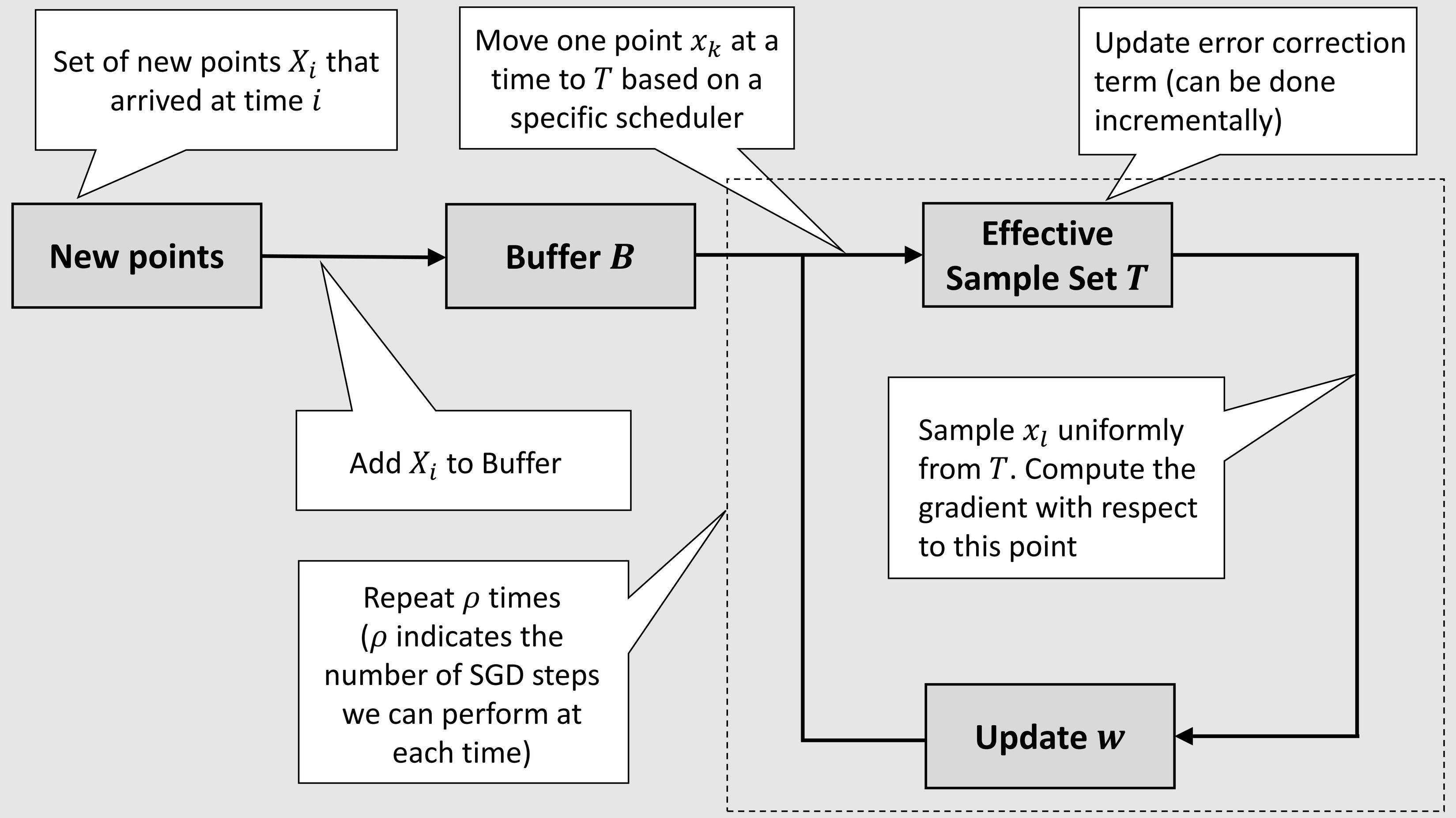
STRSAGA Algorithm

T: effective sample set

B: Buffer of training points not yet added to T

Algorithm at time i:

```
Add new points X_i to B
For j = 1 to ρ:
    if (j even) and (B not empty):
        Move a point from B to T
    Sample point x uniformly from T
    Do one iteration of variance-
     reduced SGD (SAGA[1]) using x
```



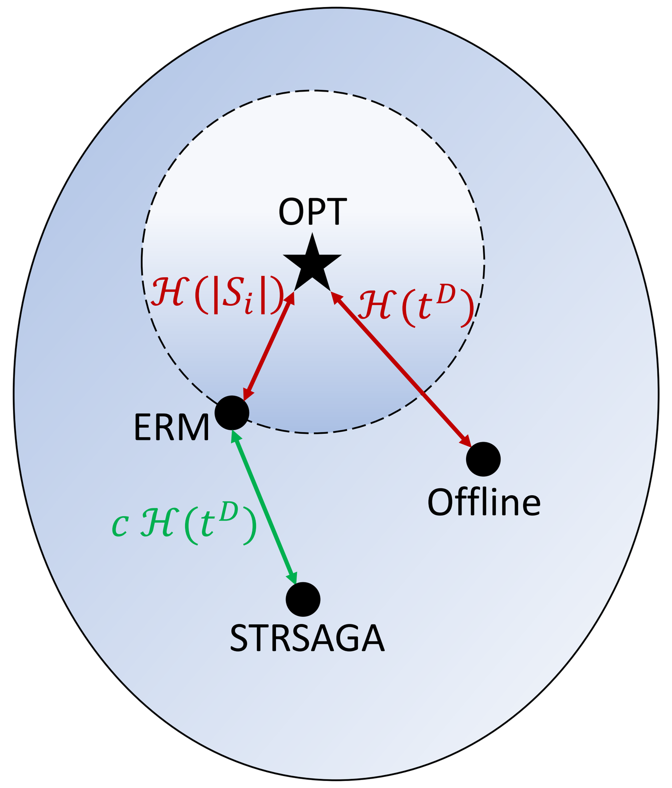
Theoretical Results

At each time i, compare STRSAGA to an offline algorithm

- DynaSAGA(ρ)[2], which has access to S_i at the beginning of time, but same computational power of $\rho \cdot i$ iterations. It optimizes over a subset of t^D points of S_i, and has error $\sim \mathcal{H}(t^D)$. (If $\rho \geq 2|S_i|/i$, then $t^D = |S_i|$.)

Under a variety of general arrival distributions, STRSAGA is c-risk-competitive to the offline algorithm w.h.p. at each time $i > i^*$; i.e., $\mathbf{E}[\mathrm{Subopt}_{S_i}] \leq c\mathcal{H}(t^D)$

- Holds for Poisson arrivals, bounded max, and unbounded max satisfying Bernstein's condition
- Shown in two steps. (1) Lemma: $\mathrm{Subopt}_T \sim \mathcal{H}(|T|)$ (2) Bound a sample-competitive ratio $|T|/t^D$ for various arrival patterns



Terminology

Let S_i be substream through time i, $S_i = \cup_{j=1}^{i} X_i$

Empirical risk $R_{S_i}(w) = \mathbf{E}_{x \sim S_i}[loss(w, x)]$, minimized by ERM $w_{S_i}^*$

Expected risk $R(w) = \mathbf{E}_{x \sim \mathcal{D}}[loss(w, x)]$, minimized by OPT w^*

$\mathrm{Error}(w) = R(w) - R(w^*)$

$\mathrm{Subopt}_{S_i}(w) = R_{S_i}(w) - R_{S_i}(w_{S_i}^*)$

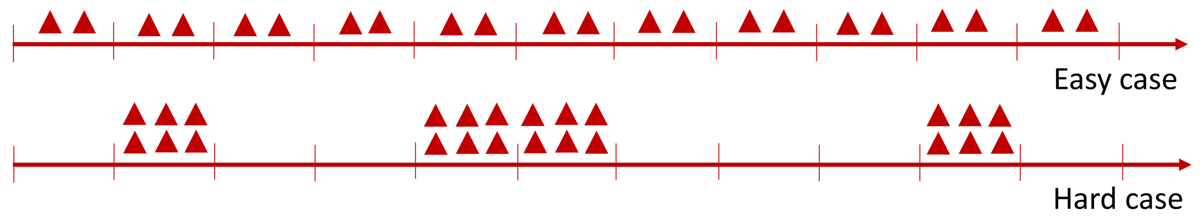
Statistical error $= R(w_{S_i}^*) - R(w^*) \leq \mathcal{H}(|S_i|)$

Assume $\mathcal{H}(n) = \kappa n^{-\alpha}, 1/2 \leq \alpha \leq 1$

If $\mathrm{Subopt}_{S_i}(w) \leq \varepsilon$, then $\mathrm{Error}(w) \leq \varepsilon + \mathcal{H}(|S_i|)$

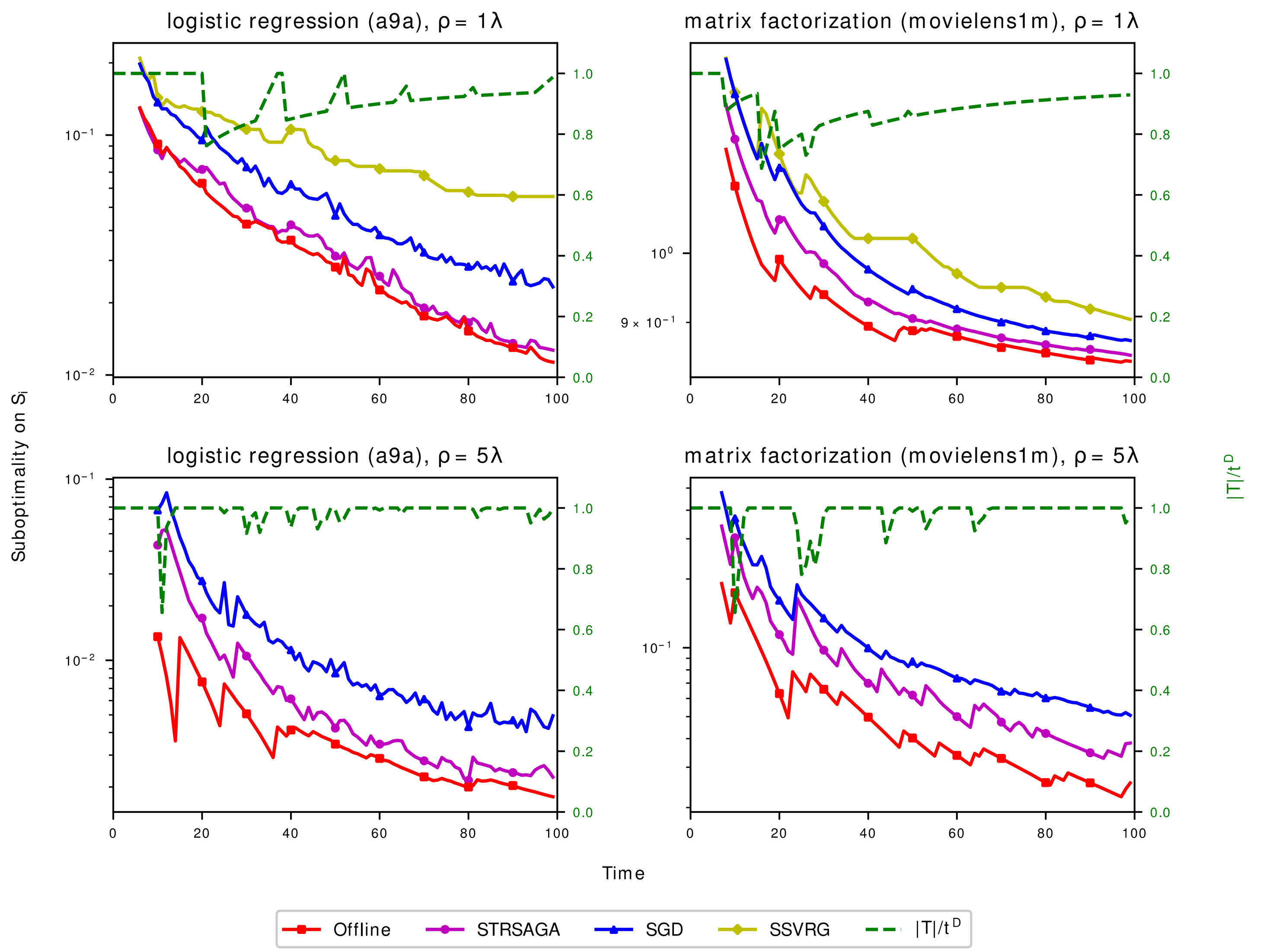
Experimental Results

The number of new points $|X_i|$ is drawn from an unknown arrival distribution, where $\mathbf{E}[|X_i|] = \lambda$.



We show results for the skewed arrival case. At each time, M points arrive w.p. λ/M, and 0 points w.p. $(1 - \lambda/M)$. Skew increases with parameter M. Below, $M = 8\,\lambda$.

STRSAGA is competitive with the offline algorithm DynaSAGA(ρ) run from scratch at each time i with $\rho \cdot i$ iterations over static S_i, and STRSAGA outperforms streaming data versions of SGD and SSVRG[3]



References

[1] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

[2] H. Daneshmand, A. Lucchi, and T. Hofmann. Starting small – learning with adaptive sample sizes. In *ICML*, pages 1463–1471, 2016.

[3] R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Competing with the empirical risk minimizer in a single pass. In *COLT*, pages 728–763, 2015.