GIANT: Experiments

Experiment Environment

- Spark 2.1.1   +   Scala 2.11.8

Experiment Environment

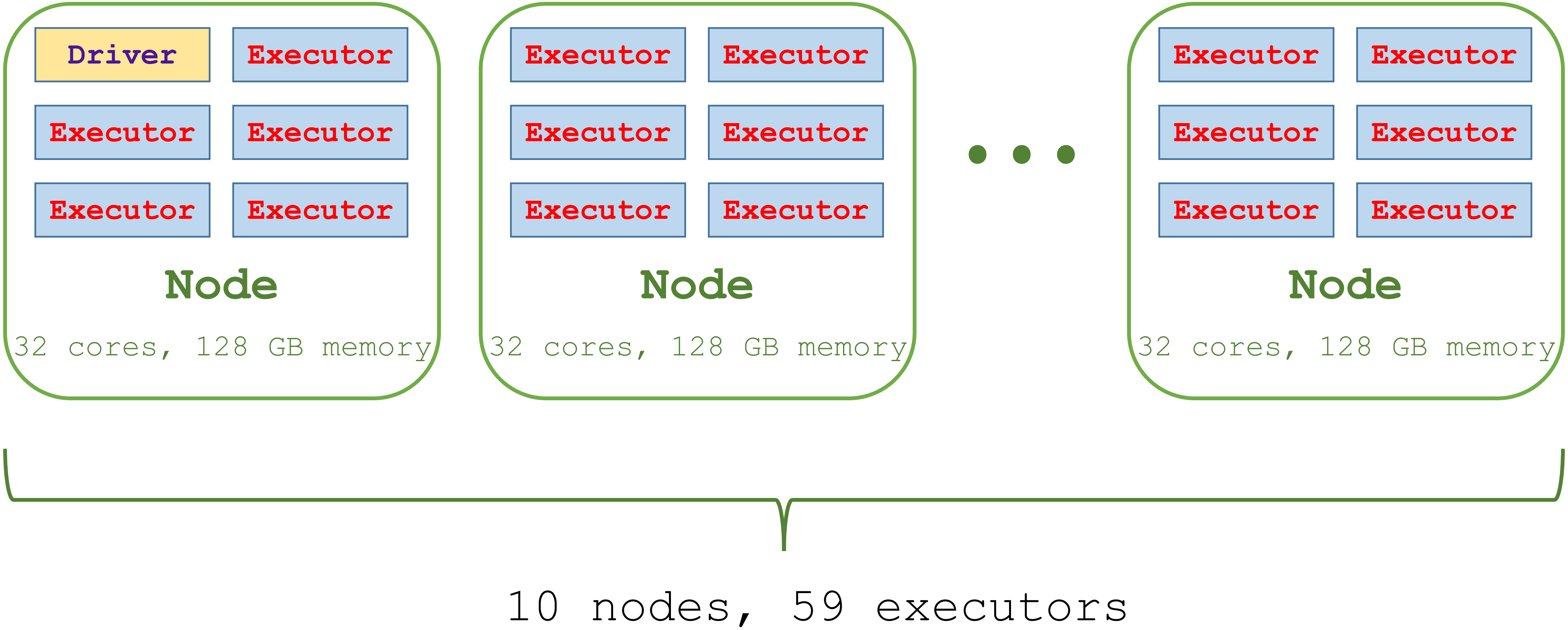
- Spark 2.1.1 + Scala 2.11.8

APACHE Spark Scala

- Cori Supercomputer
  - located at NERSC
  - each node has two 2.3GHz 16-core Haswell processors and 128GB DRAM
  - high-speed interconnect linking the compute nodes

NERSC National Energy Research Scientific Computing Center

Experiment Environment



Settings

- Solve the ℓ_2-regularized logistic regression:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad \left\{ f(\mathbf{w}) \triangleq \frac{1}{n}\sum_{j=1}^{n} \log\left(1+e^{-y_j \mathbf{x}_j^T \mathbf{w}}\right) + \frac{\gamma}{2}\|\mathbf{w}\|_2^2 \right\}$$

- Split $\mathbf{X} \in \mathbb{R}^{n\times d}$ (by data) to $m = 59$ parts.
- Local sample size $s = \frac{n}{m} \quad \gtrsim \quad$ number of features d.
  - Thus m cannot be over-large.
- We use dense $\mathbf{X}$.

Compared Methods

- Gradient descent with momentum
  - choose *step size* from {0.1, 1, 10, 100}
  - choose *momentum* from {0.5, 0.9, 0.95, 0.99, 0.999}

Compared Methods

- Gradient descent with momentum
- Limited memory BFGS (a quasi-Newton method)
  - choose *number of history* from {30, 100, 300}
  - line search is used

Compared Methods

- Gradient descent with momentum
- Limited memory BFGS
- Distributed ADMM [Boyd's review]
  - chose parameter ρ from $\{0.1\gamma, \gamma, 10\gamma\}$
  - local solver: SVRG
  - choose *step size of SVRG* from {0.1, 1, 10, 100}
  - choose *max iteration of SVRG* from {30, 100, 300}

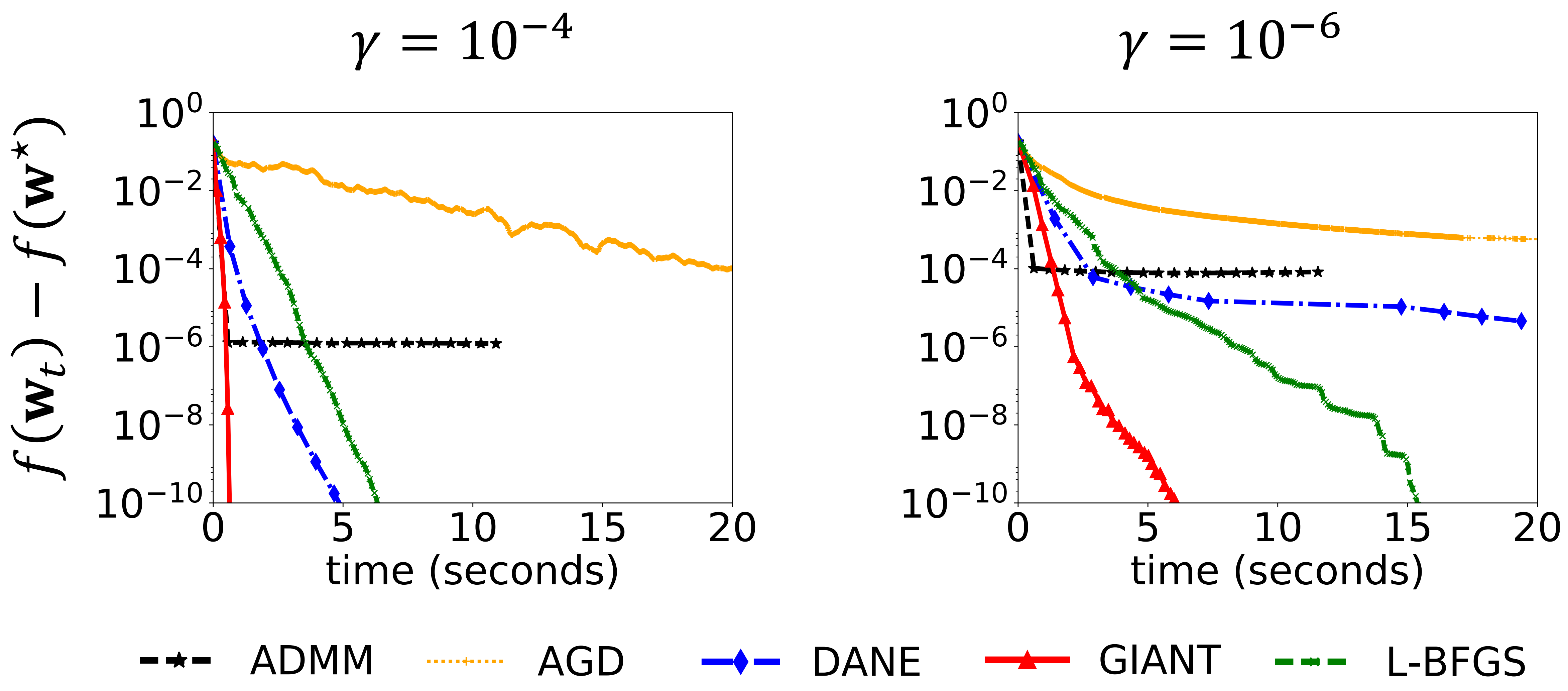
Compared Methods

- Gradient descent with momentum
- Limited memory BFGS
- Distributed ADMM [Boyd’s review]
- DANE (another Newton-type method) [Shamir, Srebro, Zhang 2014]
  - local solver: SVRG
  - choose *step size of SVRG* from {0.1, 1, 10, 100}
  - choose *max iteration of SVRG* from {30, 100, 300}

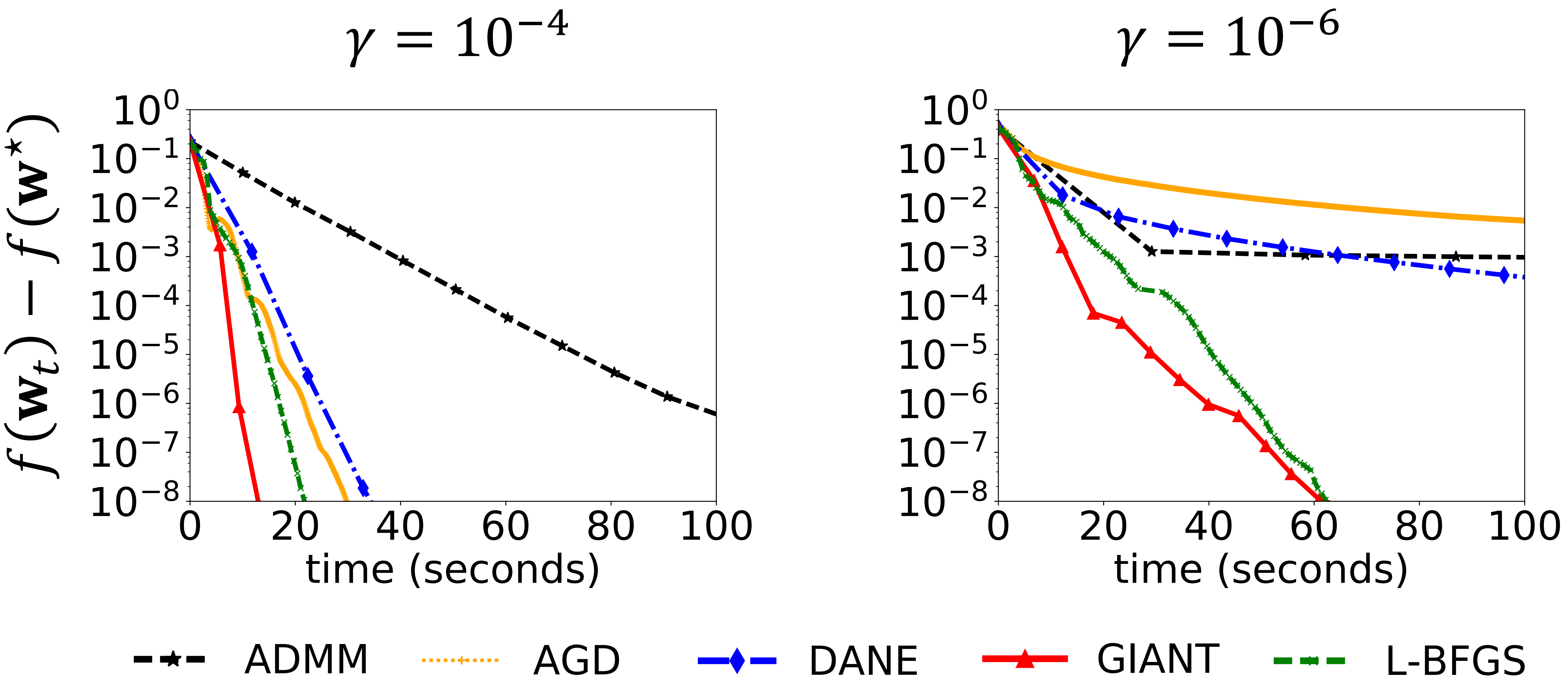
Compared Methods

- Gradient descent with momentum
- Limited memory BFGS
- Distributed ADMM [Boyd's review]
- DANE (another Newton-type method) [Shamir, Srebro, Zhang 2014]
- GIANT
    - local solver: conjugate gradient
    - choose *max iteration of CG* from {30, 100, 300}

Covtype (n=581K, d=54)



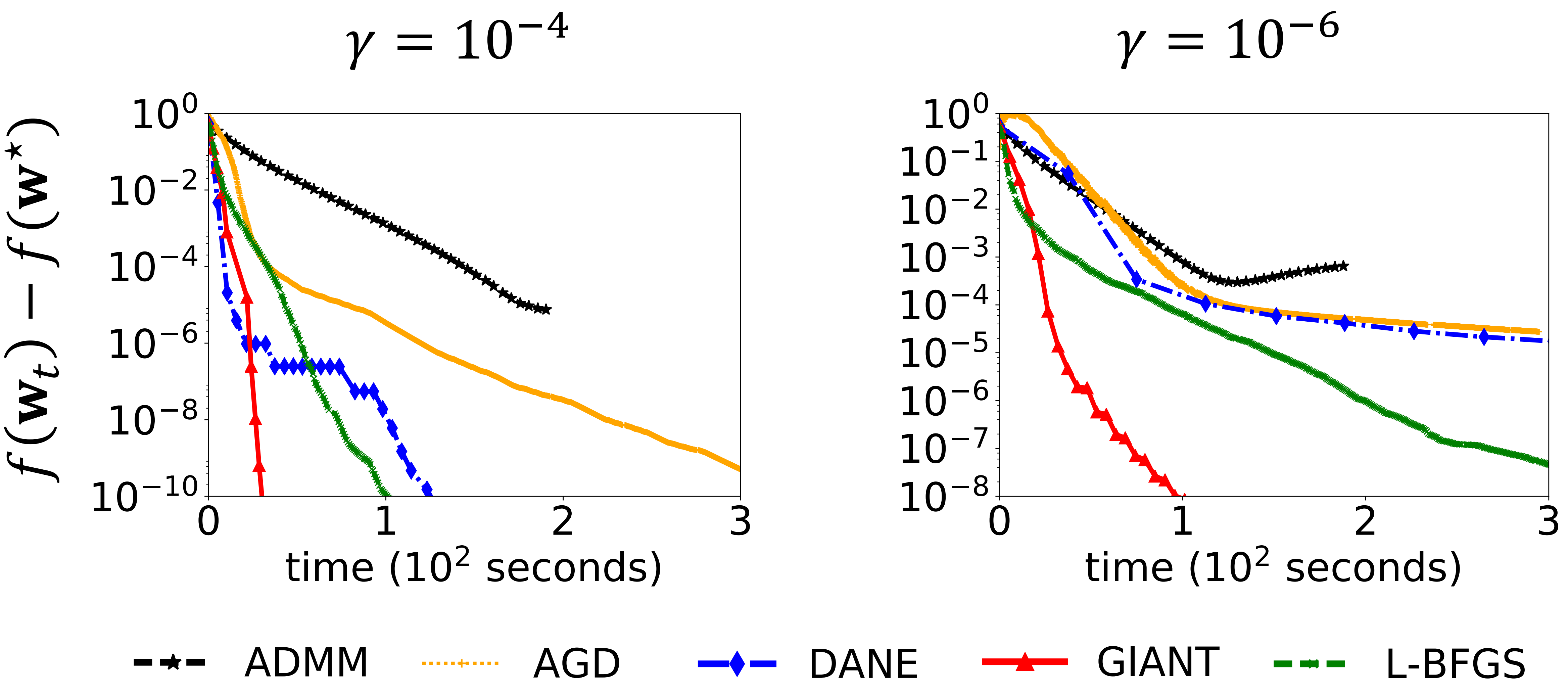
Epsilon (n=400K, d=2K)



MNIST8M (n=1.6M, d=784)

- Digit “4” versus “9”: 1.6M samples out of the total 8M samples

MNIST8M (n=1.6M, d=784)



How about Larger d?

- Split $\mathbf{X} \in \mathbb{R}^{n \times d}$ (by data) to $m = 59$ parts.
- Previously, local sample size $s = \frac{n}{m} \gg$ number of features d.
- Does GIANT work if $s \approx d$?

Random Feature Maps (RFM)

- Generate 10K *random Fourier features* *[Rahimi & Recht, 07]* of the *RBF kernel*

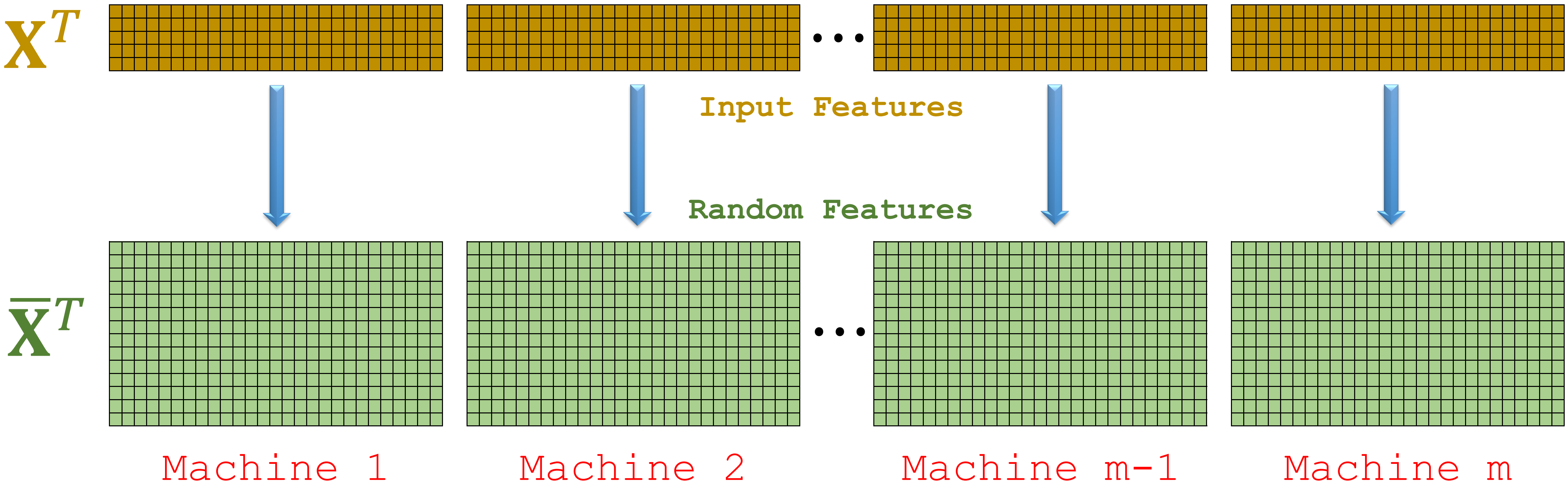
$$k(\mathbf{x}_i, \mathbf{x}_j) \;=\; \exp\Big(-\tfrac{1}{2\sigma}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\Big)$$

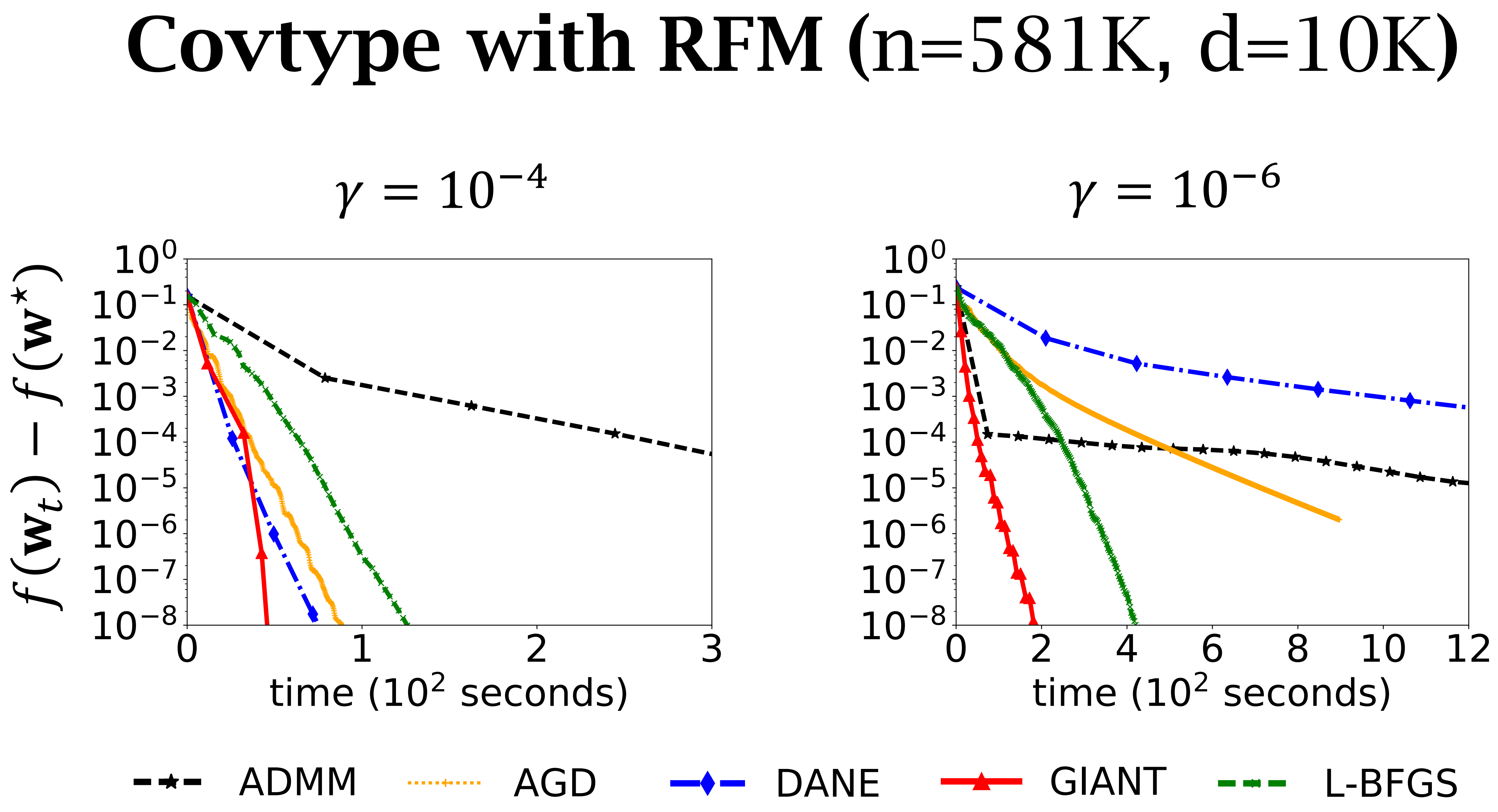
- Setting of RBF *kernel width parameter* σ:

$$\sigma \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$$

- Replace the original features by the higher-dim random features.

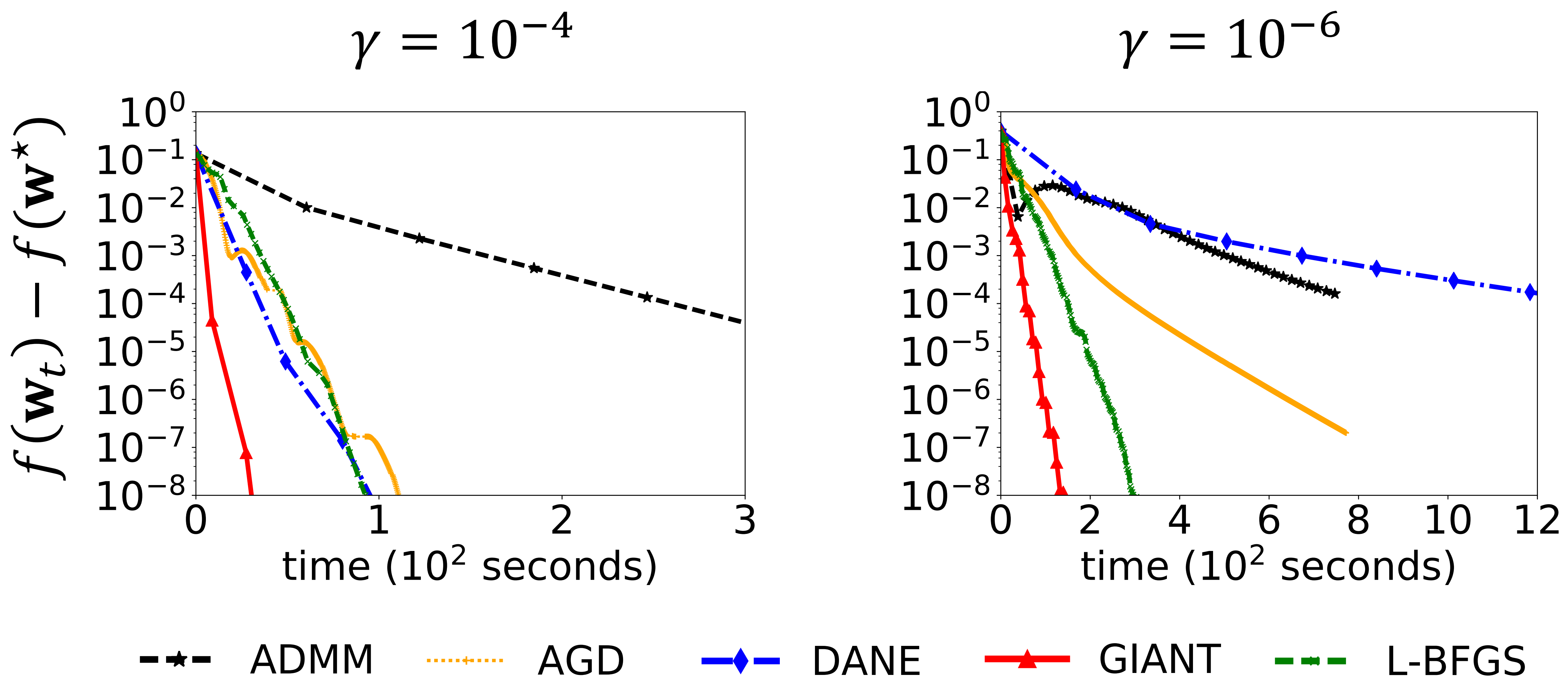
Random Feature Maps (RFM)





Covtype with RFM (n=581K, d=10K)
$\gamma = 10^{-4}$
$\gamma = 10^{-6}$
$f(\mathbf{w}_t) - f(\mathbf{w}^\star)$
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
0
1
2
3
4
6
8
10
12
time (10^2 seconds)
ADMM
AGD
DANE
GIANT
L-BFGS

Epsilon with RFM (n=400K, d=10K)



MNIST8M with RFM (n=1.6M, d=10K)