Semi-Cyclic SGD

Hubert Eichner
Google

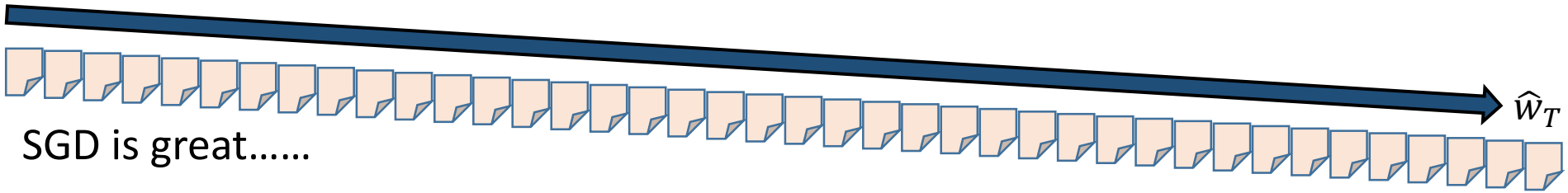
Tomer Koren
Google

Brendan McMahan
Google

Kunal Talwar
Google

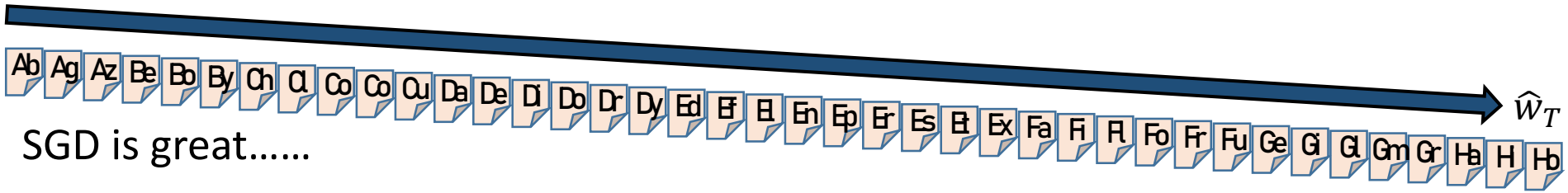
Nati Srebro
TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, z_t)$$



SGD is great……

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, z_t)$$



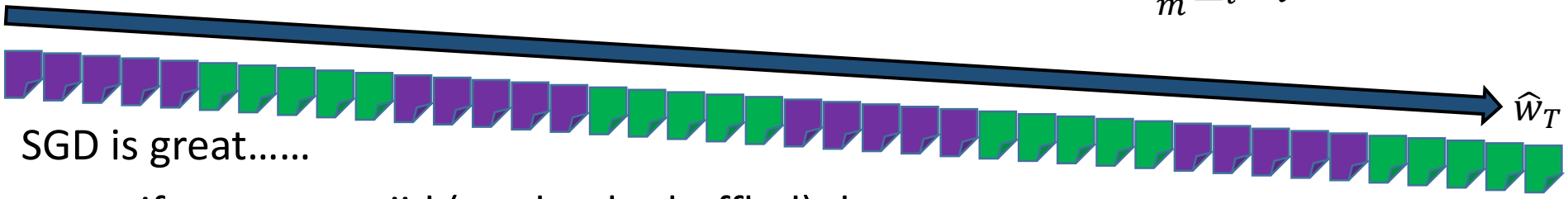
SGD is great……

if you run on iid (randomly shuffled) data

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, z_t)$$

Samples in block $i = 1..m$ are sampled from as $z_t \sim \mathcal{D}_i$

overall distribution: $\mathcal{D} = \frac{1}{m}\sum_i \mathcal{D}_i$



SGD is great……

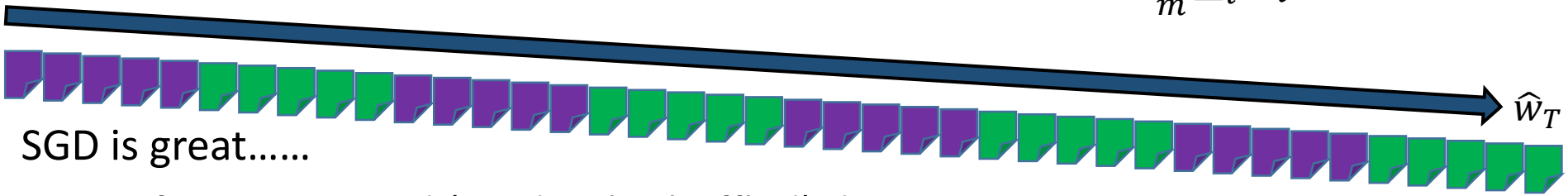
if you run on iid (randomly shuffled) data

Cyclically varying (not fully shuffled) data

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, z_t)$$

Samples in block $i = 1..m$ are sampled from as $z_t \sim \mathcal{D}_i$

overall distribution: $\mathcal{D} = \frac{1}{m} \sum_i \mathcal{D}_i$

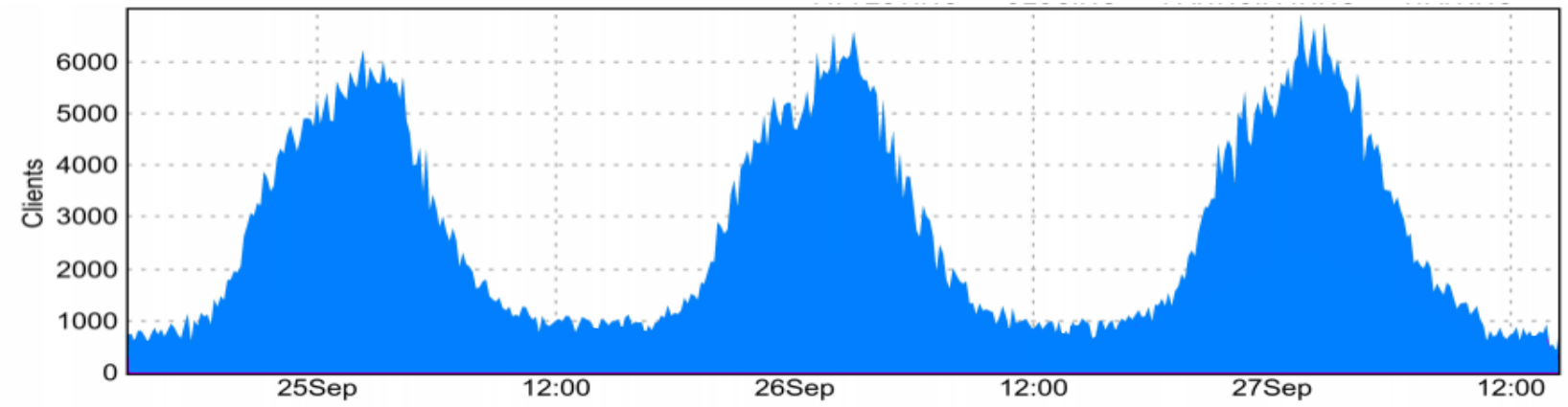


SGD is great……

if you run on iid (randomly shuffled) data

Cyclically varying (not fully shuffled) data, e.g. in **Federated Learning**

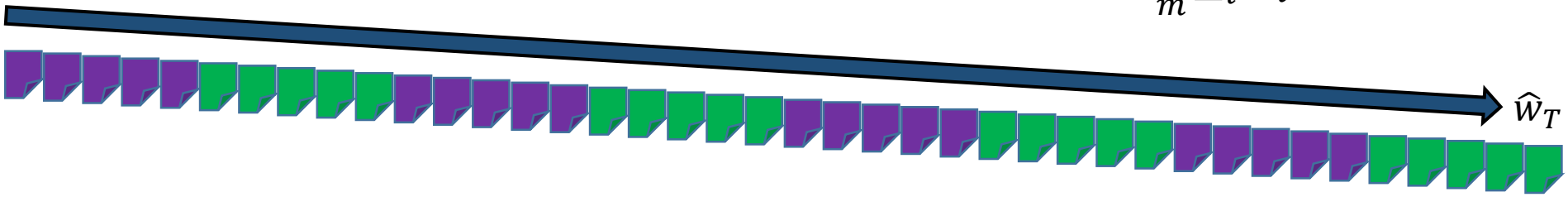
- Train model by executing SGD steps on user devices ***when device available*** (plugged in, idle, on WiFi)
- Diurnal variations (e.g. Day vs night available devices; US vs UK vs India)



$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, z_t)$$

Samples in block $i = 1..m$ are sampled from as $z_t \sim \mathcal{D}_i$

overall distribution: $\mathcal{D} = \frac{1}{m}\sum_i \mathcal{D}_i$

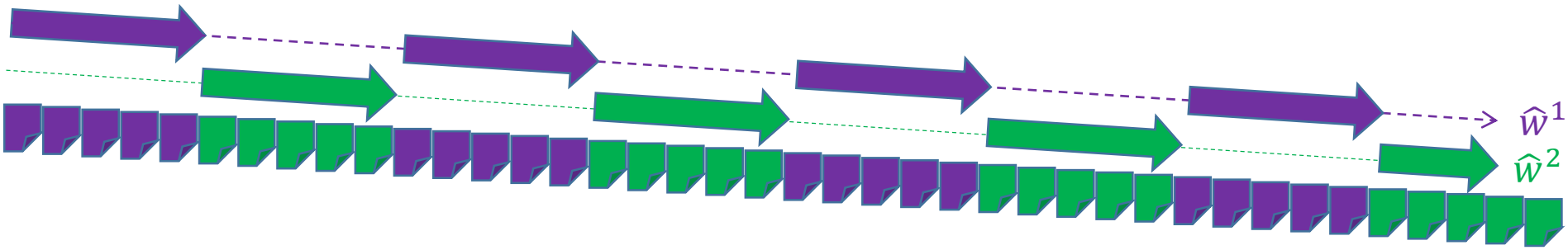


- Train $\widehat{w}_T$ by running block-cyclic SGD

  ➔ could be MUCH slower, by an arbitrary large factor

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, z_t)$$

Samples in block $i = 1..m$ are sampled from as $z_t \sim \mathcal{D}_i$



- Train $\widehat{w}_T$ by running block-cyclic SGD

  ➔ could be MUCH slower, by an arbitrary large factor

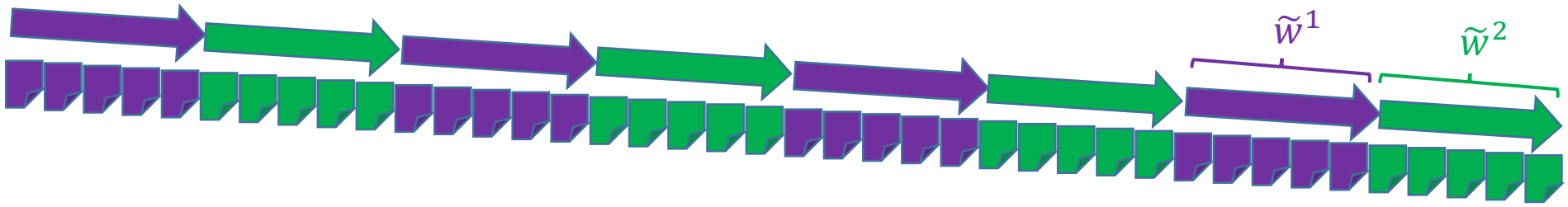
Pluralistic approach: learn different $\widehat{w}^i$ for each block $i = 1..m$

- Train each $\widehat{w}^i$ separately on data from that block (across all cycles)

  ➔ could be slower/less efficient by a factor of m

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, z_t)$$

Samples in block $i = 1..m$ are sampled from as $z_t \sim \mathcal{D}_i$



- Train $\widehat{w}_T$ by running block-cyclic SGD

  ➔ could be MUCH slower, by an arbitrary large factor

Pluralistic approach: learn different $\widehat{w}^i$ for each block $i = 1..m$

- Train each $\widehat{w}^i$ separately on data from that block (across all cycles)

  ➔ could be slower/less efficient by a factor of m

- Our solution: train $\widetilde{w}^i$ using single SGD chain+"pluralistic averaging"

  ➔ **exactly same guarantee as if using random shuffling (no degradation)**

  ➔ **no extra comp. cost, no assumptions about $\mathcal{D}_i$ nor relatedness**