Online EM for Unsupervised Models

NAACL – June 3, 2009

Percy Liang   Dan Klein

University of California
Berkeley

Based on a true story

Part-of-speech induction:

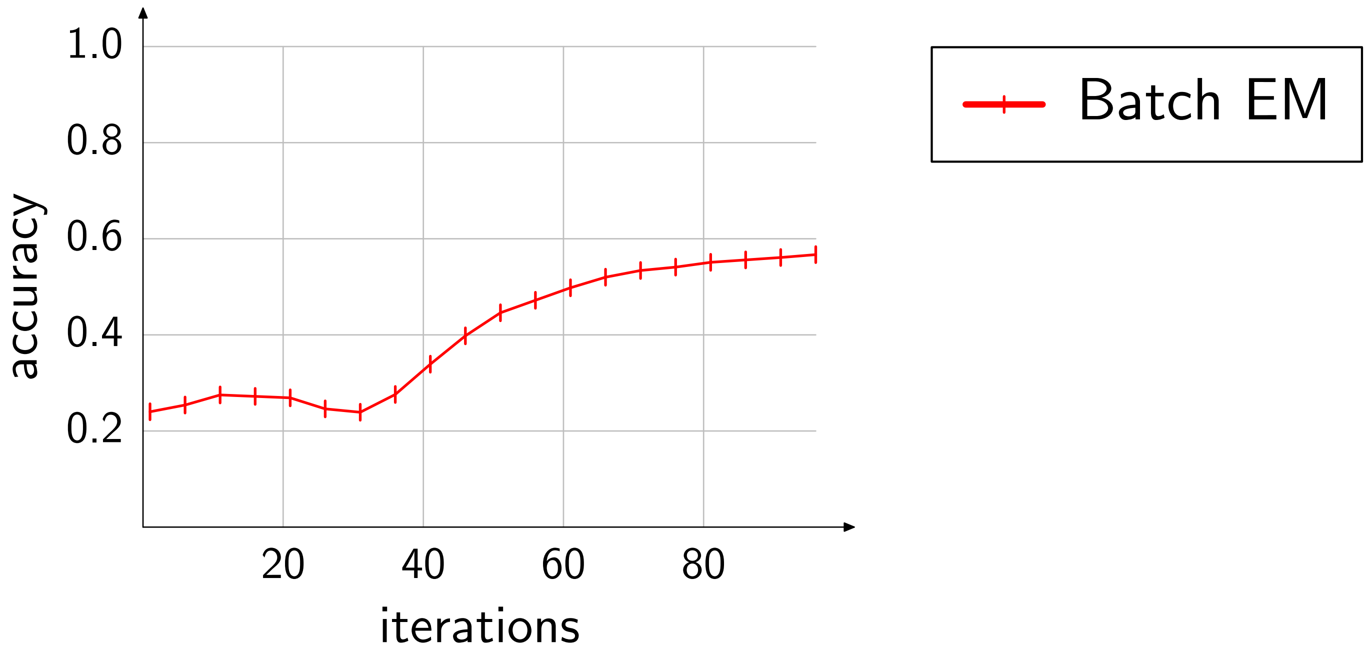
DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*

Based on a true story

Part-of-speech induction:

DT NNP NNP VBD

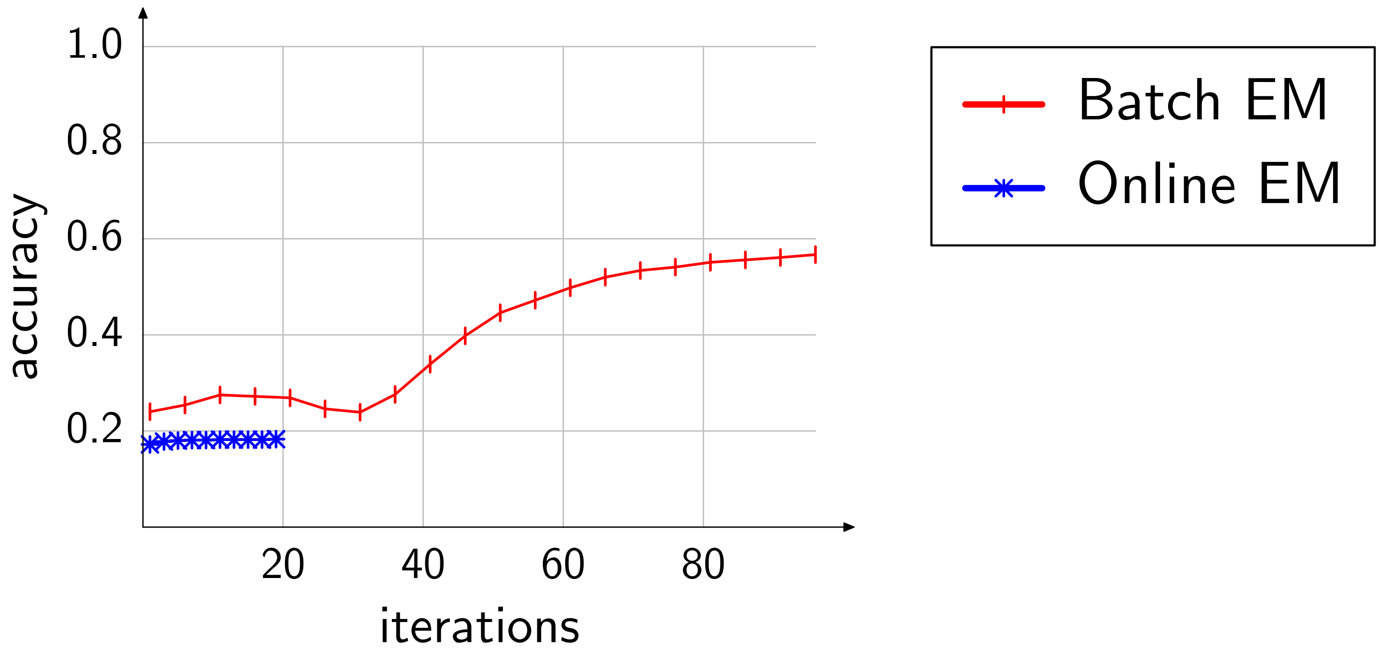
The European Commission agreed



Based on a true story

Part-of-speech induction:

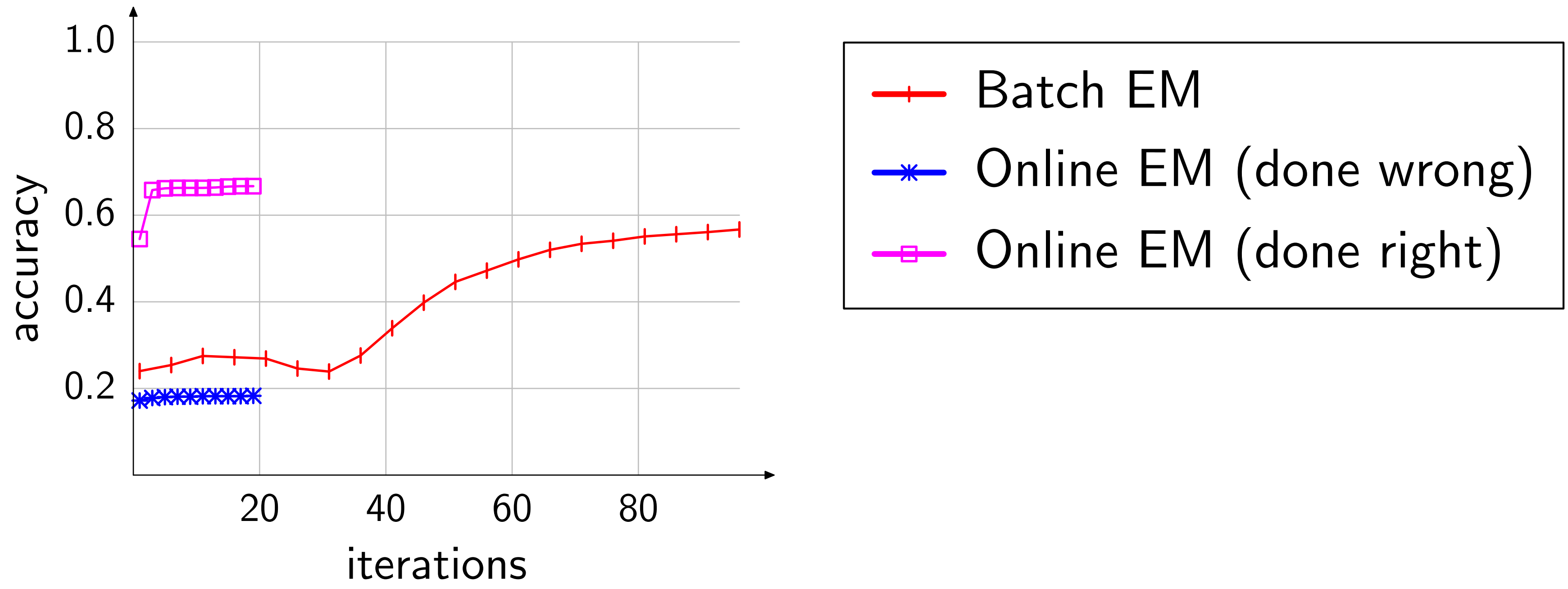
DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*



Based on a true story

Part-of-speech induction:

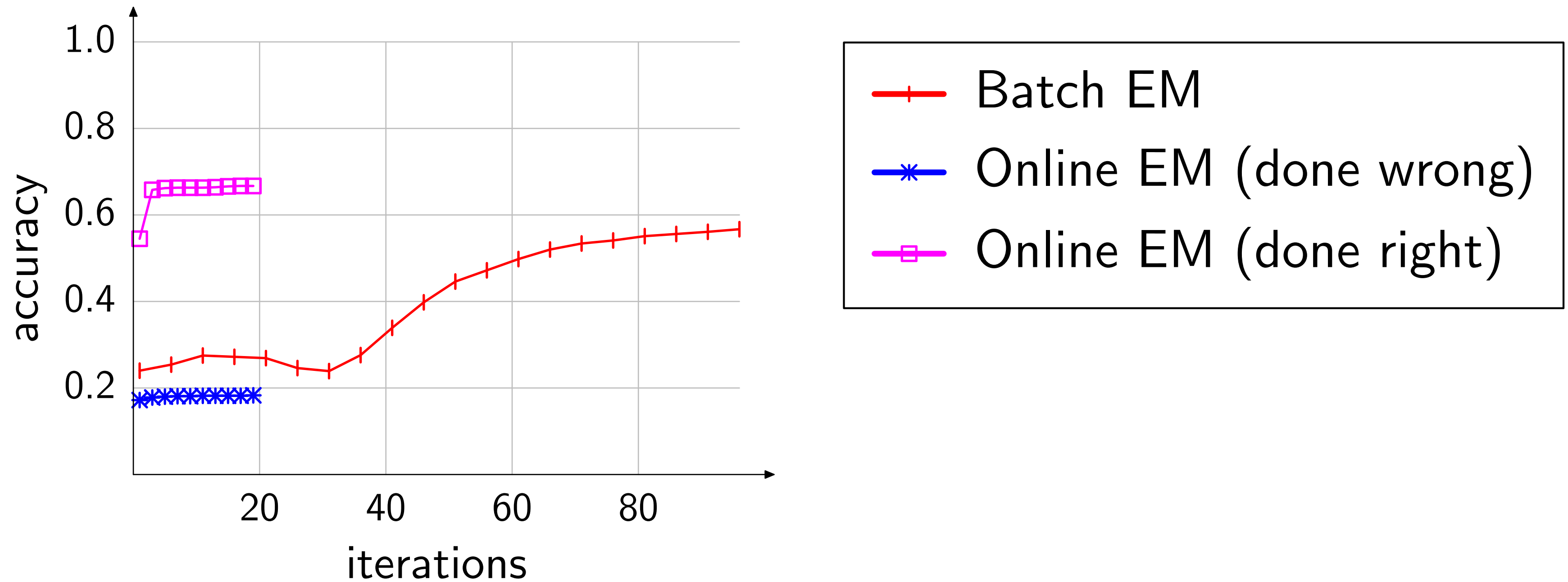
DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*



Based on a true story

Part-of-speech induction:

DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*



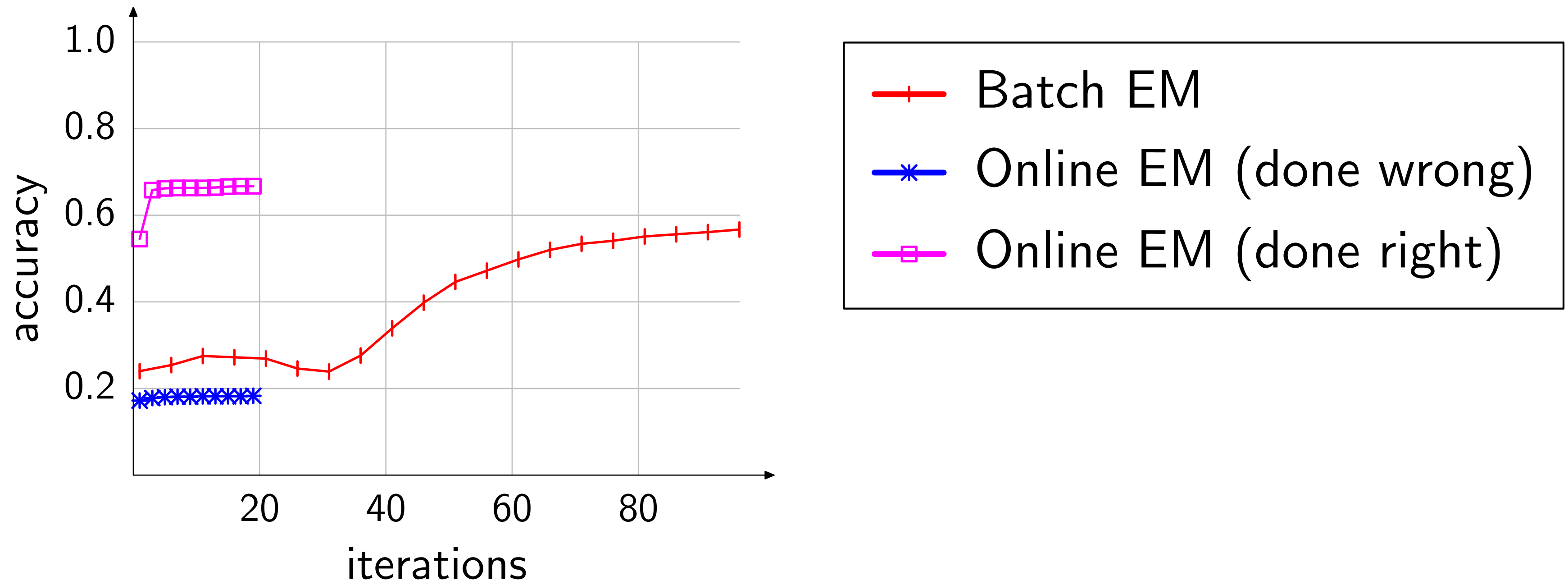
Observations:

1. Online EM is faster than batch EM

Based on a true story

Part-of-speech induction:

DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*



Observations:

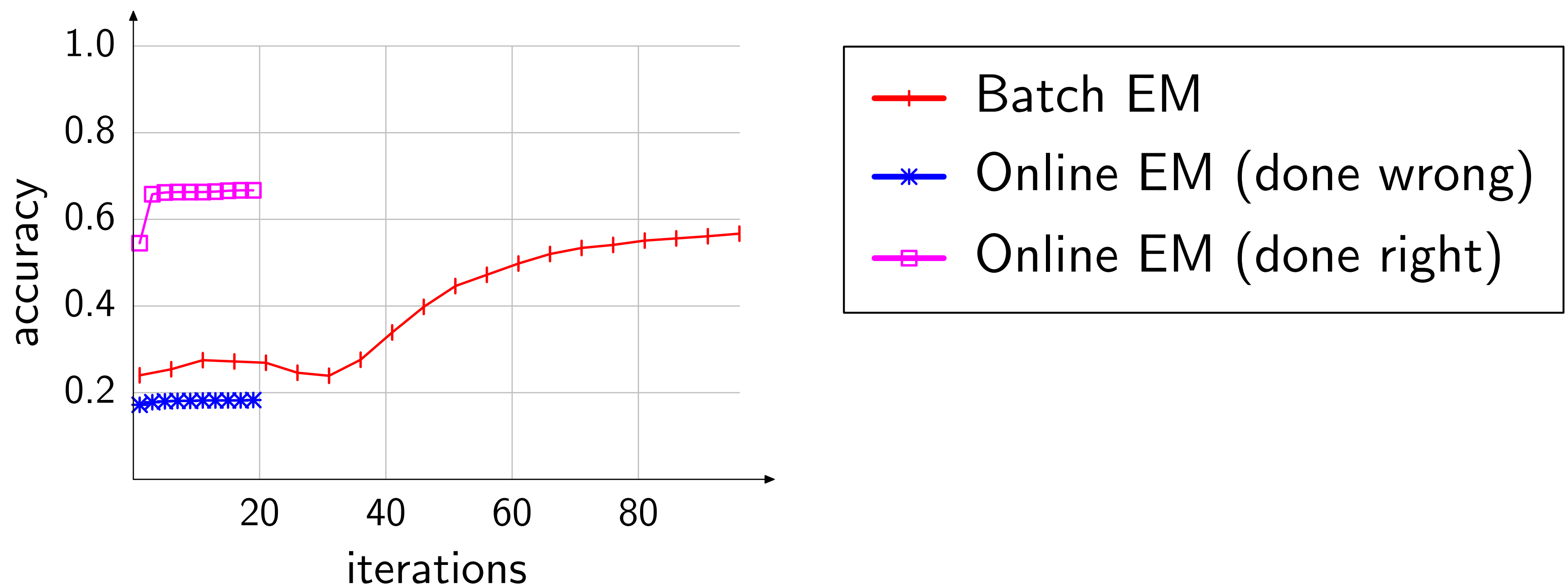
1. Online EM is faster than batch EM
2. Online EM improves accuracy(!)

Based on a true story

Part-of-speech induction:

DT NNP NNP VBD

The European Commission agreed



Observations:

1. Online EM is faster than batch EM
2. Online EM improves accuracy(!)
3. Details of online EM do matter

Four tasks

DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*

POS tagging

Four tasks

DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*

POS tagging

l o o k | *a t* | *t h e* | *b o o k*

Word segmentation

Four tasks

DT	NNP	NNP	VBD
The	*European*	*Commission*	*agreed*

POS tagging

l o o k | *a t* | *t h e* | *b o o k*

Word segmentation

Baseball

...Matt Williams has demonstrated throughout his career that he will NOT wait for good pitches to hit...

Document classification

Four tasks

DT NNP NNP VBD
The European Commission agreed

POS tagging

l o o k | a t | t h e | b o o k

Word segmentation

BASEBALL

...Matt Williams has demonstrated throughout his career that he will NOT wait for good pitches to hit...

Document classification

the European Commission

la Commission européenne

Word alignment

Unsupervised induction

Setting:

$$\mathbf{x}^{(1)} \qquad \mathbf{x}^{(2)} \qquad \cdots \qquad \mathbf{x}^{(n)}$$

Unsupervised induction

Setting:

$$\mathbf{x}^{(1)} \quad \mathbf{x}^{(2)} \quad \cdots \quad \mathbf{x}^{(n)}$$

$$\mathbf{z}^{(1)} \quad \mathbf{z}^{(2)} \quad \cdots \quad \mathbf{z}^{(n)}$$

Unsupervised induction

Setting:

$$\begin{array}{cccc} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(n)} \\ \mathbf{z}^{(1)} & \mathbf{z}^{(2)} & \cdots & \mathbf{z}^{(n)} \end{array}$$

Probabilistic model: $p(\mathbf{x}, \mathbf{z}; \theta)$

- $\mathbf{x}$: observed input
- $\mathbf{z}$: hidden output
- θ: parameters (multinomial probabilities)

Unsupervised induction

Setting:

$$\mathbf{x}^{(1)} \quad \mathbf{x}^{(2)} \quad \cdots \quad \mathbf{x}^{(n)}$$
$$\mathbf{z}^{(1)} \quad \mathbf{z}^{(2)} \quad \cdots \quad \mathbf{z}^{(n)}$$

Probabilistic model: $p(\mathbf{x}, \mathbf{z}; \theta)$

$\mathbf{x}$: observed input
$\mathbf{z}$: hidden output
θ: parameters (multinomial probabilities)

Training objective: likelihood

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{i=1}^{n} \log p(\mathbf{x}^{(i)}; \theta)$$

Unsupervised induction

Setting:

$$\mathbf{x}^{(1)} \quad \mathbf{x}^{(2)} \quad \cdots \quad \mathbf{x}^{(n)}$$

$$\mathbf{z}^{(1)} \quad \mathbf{z}^{(2)} \quad \cdots \quad \mathbf{z}^{(n)}$$

Probabilistic model: $p(\mathbf{x}, \mathbf{z}; \theta)$

- $\mathbf{x}$: observed input
- $\mathbf{z}$: hidden output
- θ: parameters (multinomial probabilities)

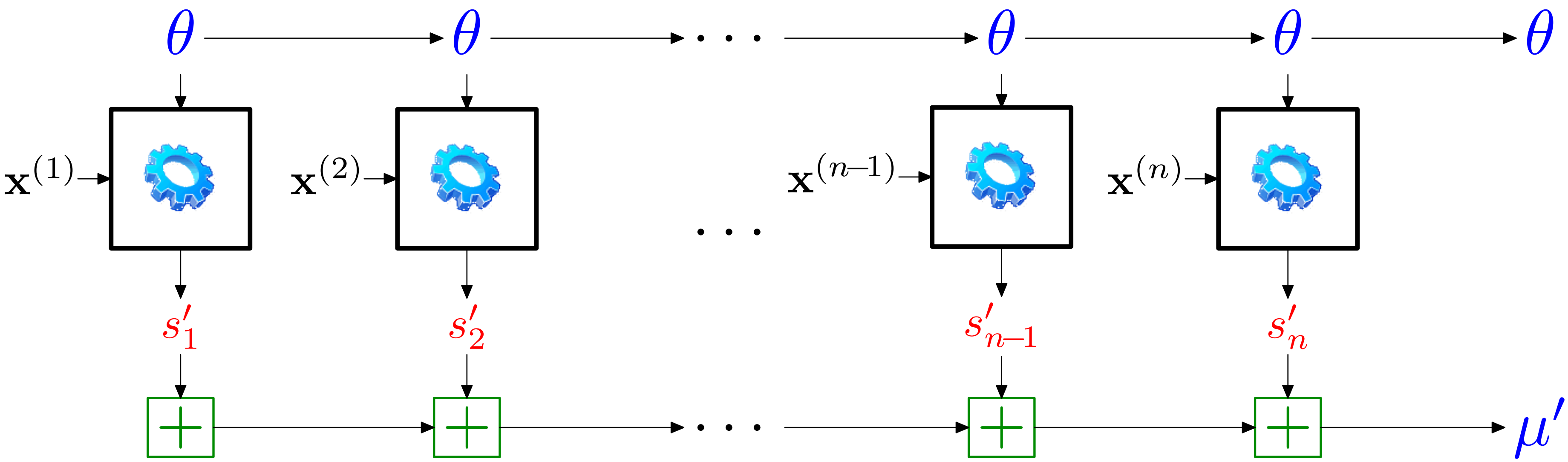
Training objective: likelihood

$$\theta^* = \operatorname*{argmax}_\theta \sum_{i=1}^n \log p(\mathbf{x}^{(i)}; \theta)$$

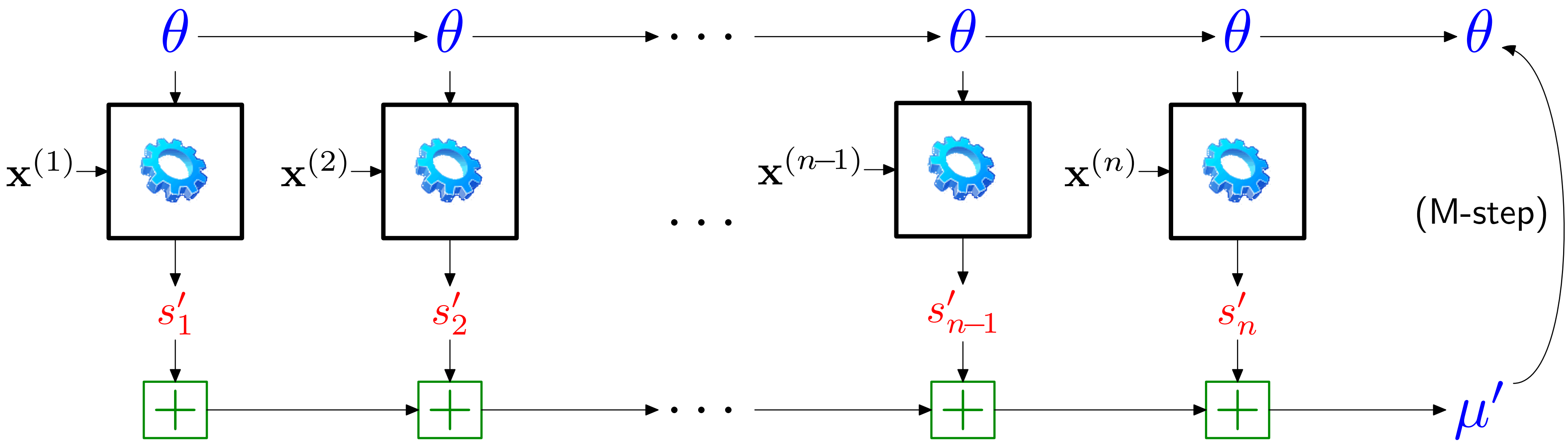
Evaluation: accuracy

gold $\mathbf{z}^{(i)}$ versus predicted $\text{argmax}_{\mathbf{z}}\, p(\mathbf{z} \mid \mathbf{x}^{(i)}; \theta^*)$

Batch EM

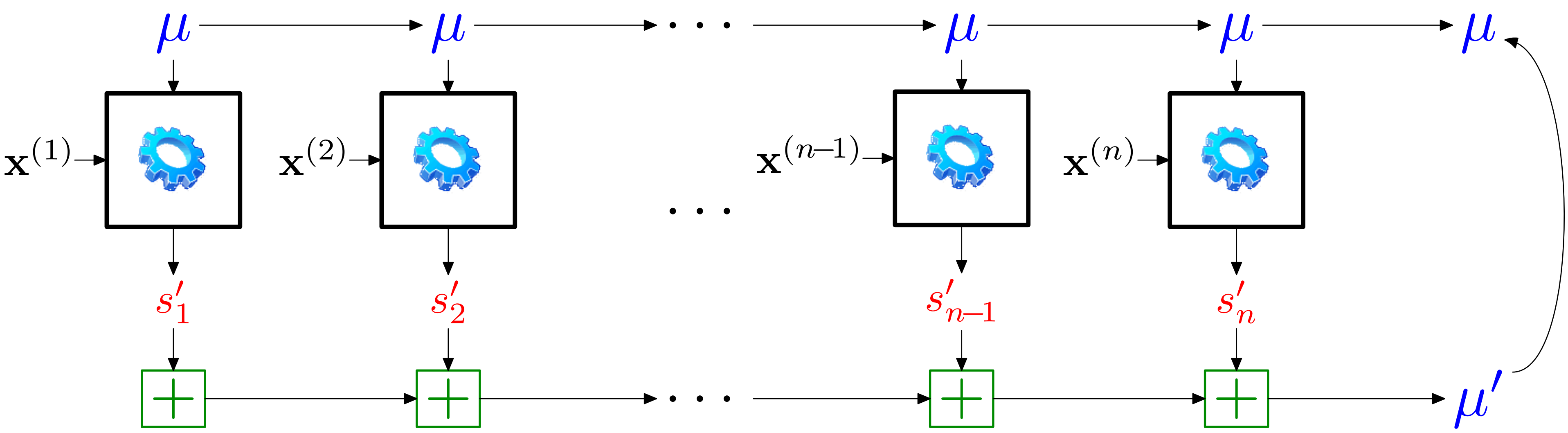

θ θ $\cdots$ θ θ θ
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\mathbf{x}^{(n-1)}$
$\mathbf{x}^{(n)}$
s'_1
s'_2
s'_{n-1}
s'_n
μ'

Batch EM

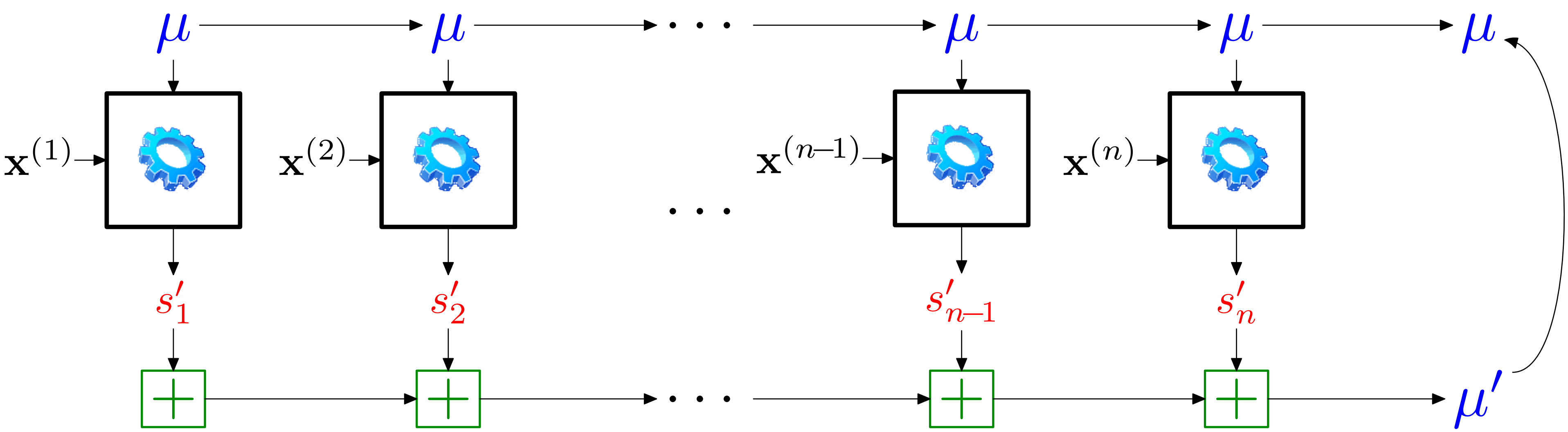


θ
θ
$\cdots$
θ
θ
θ
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\cdots$
$\mathbf{x}^{(n-1)}$
$\mathbf{x}^{(n)}$
(M-step)
s'_1
s'_2
s'_{n-1}
s'_n
$\cdots$
μ'

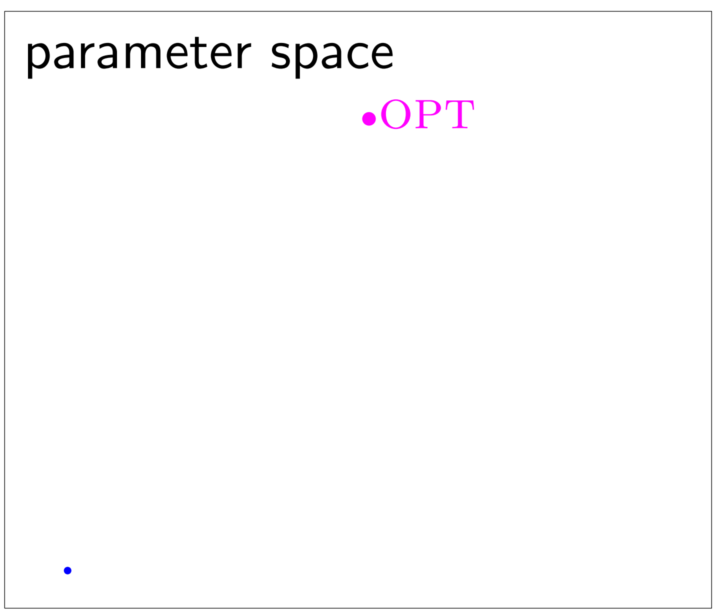
Batch EM


μ
μ
$\cdots$
μ
μ
μ
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\cdots$
$\mathbf{x}^{(n-1)}$
$\mathbf{x}^{(n)}$
s'_1
s'_2
s'_{n-1}
s'_n
$\cdots$
μ'

Batch EM

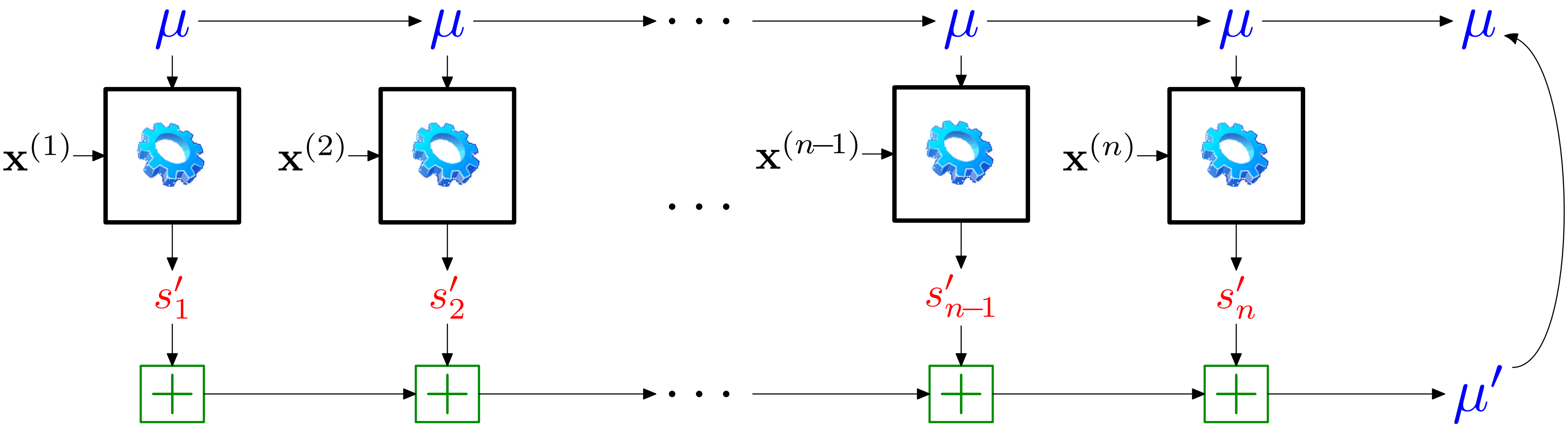


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

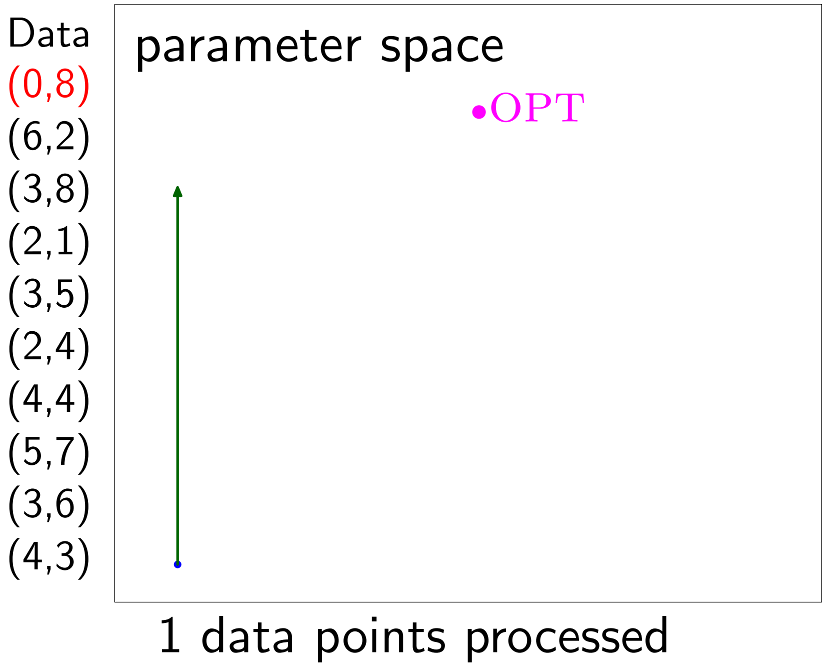


0 data points processed

Batch EM

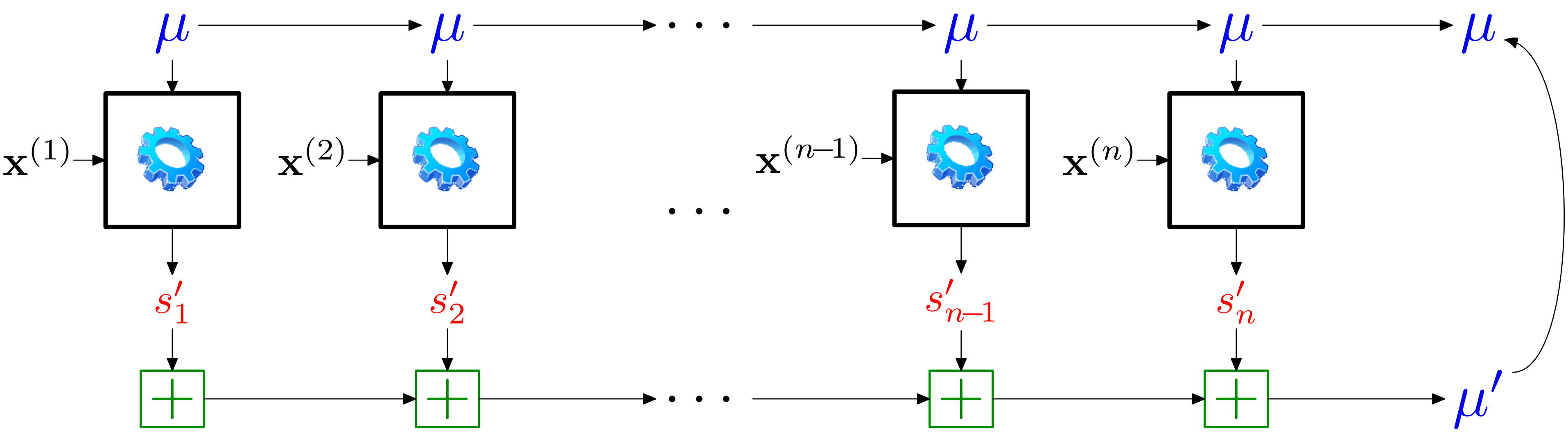


μ
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\mathbf{x}^{(n-1)}$
$\mathbf{x}^{(n)}$
s'_1
s'_2
s'_{n-1}
s'_n
μ'

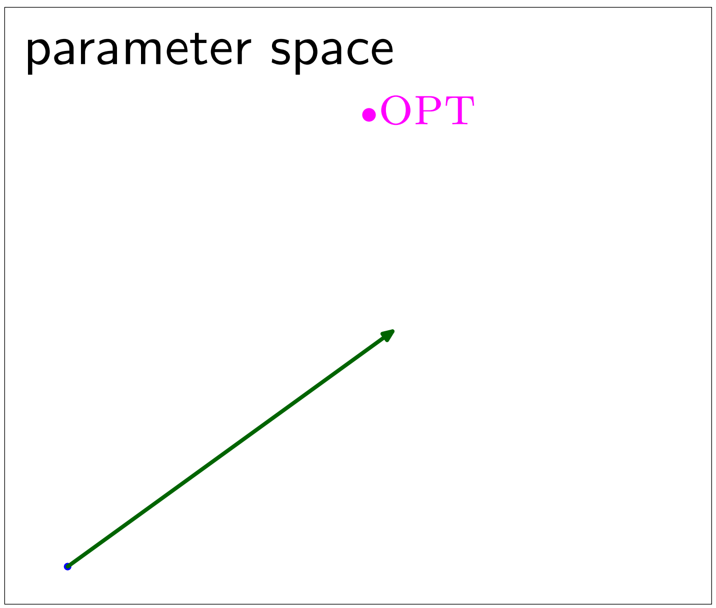


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)
parameter space
OPT
1 data points processed

Batch EM

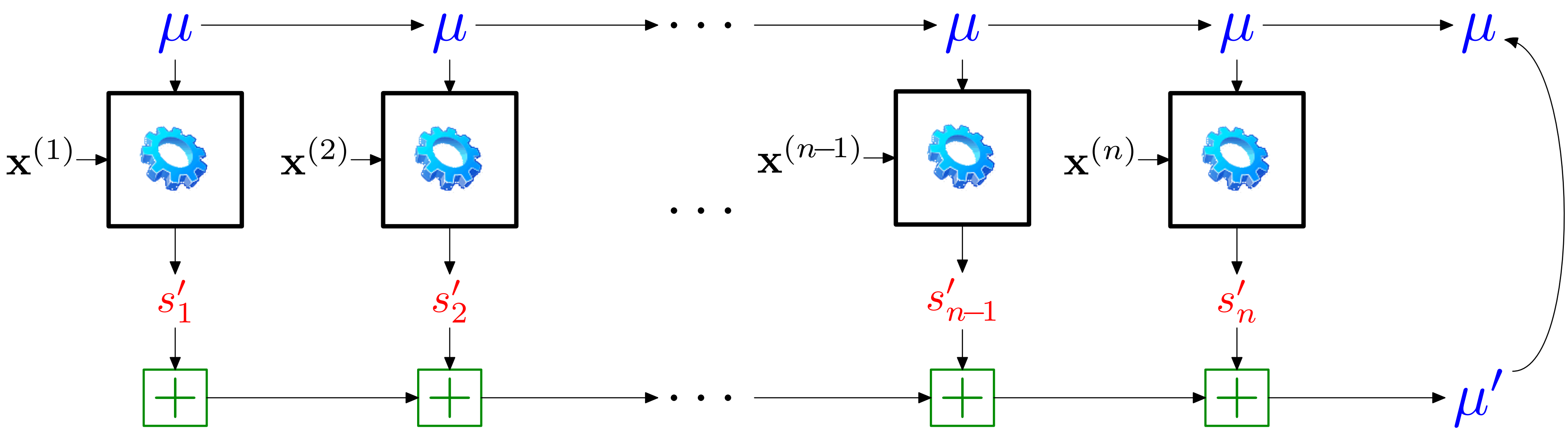


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

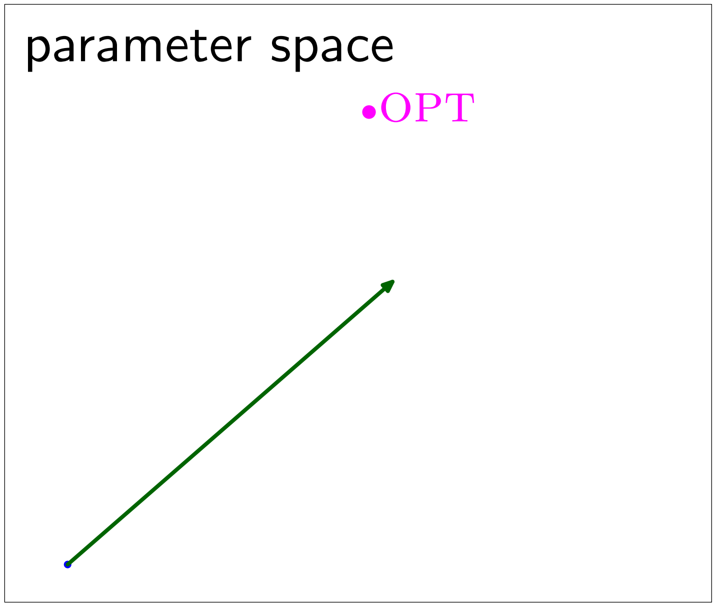


2 data points processed

Batch EM

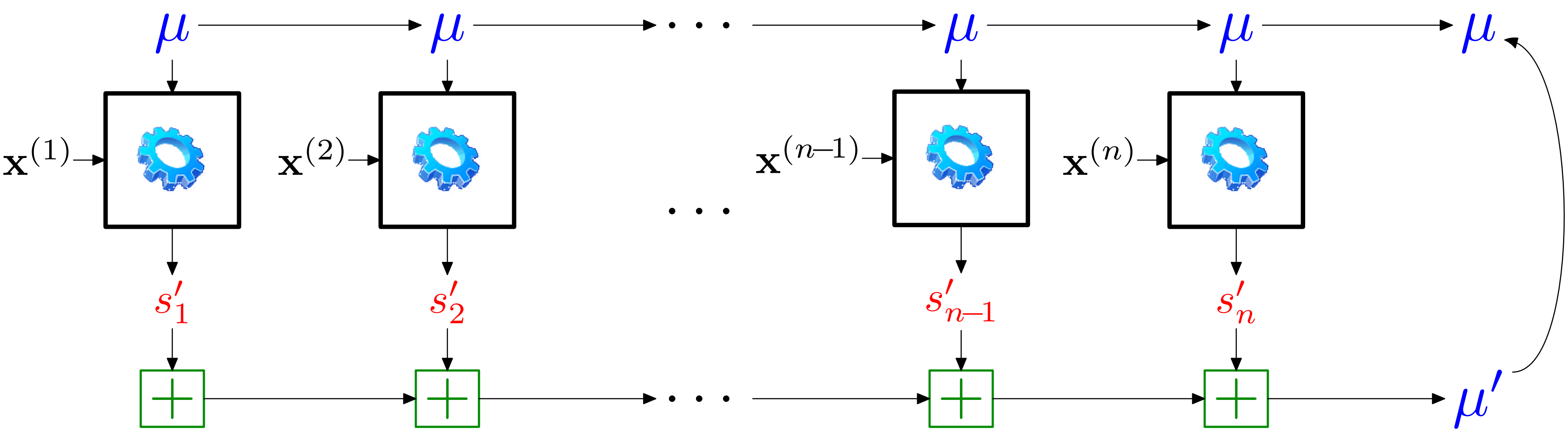


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

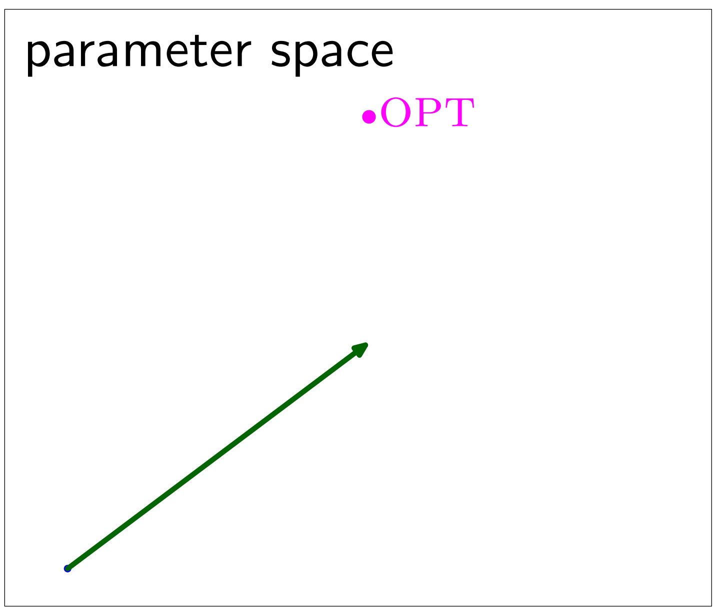


3 data points processed

Batch EM

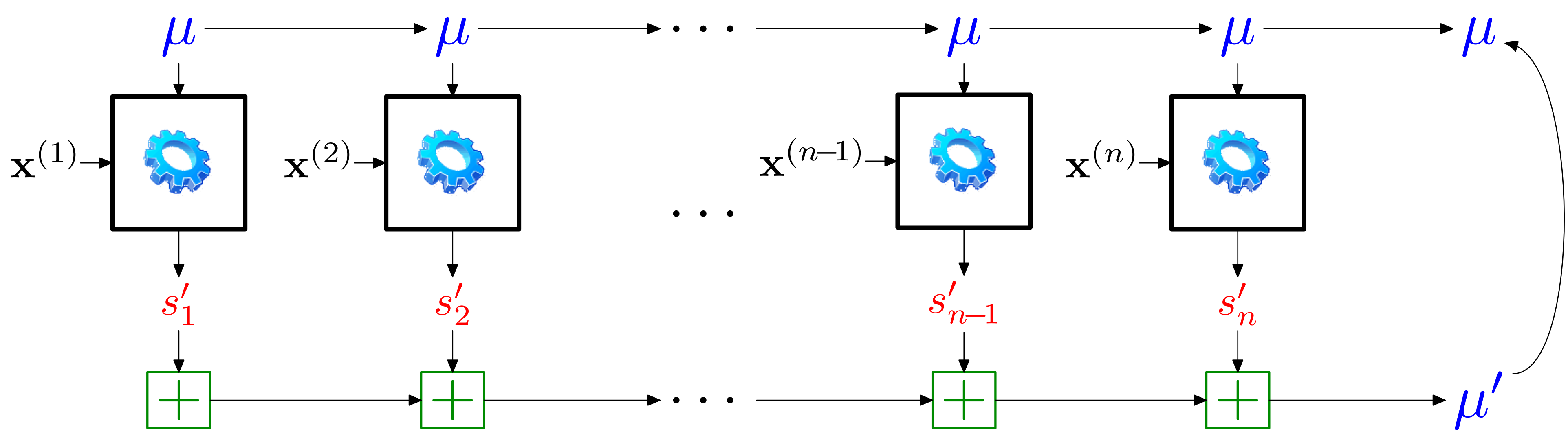


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

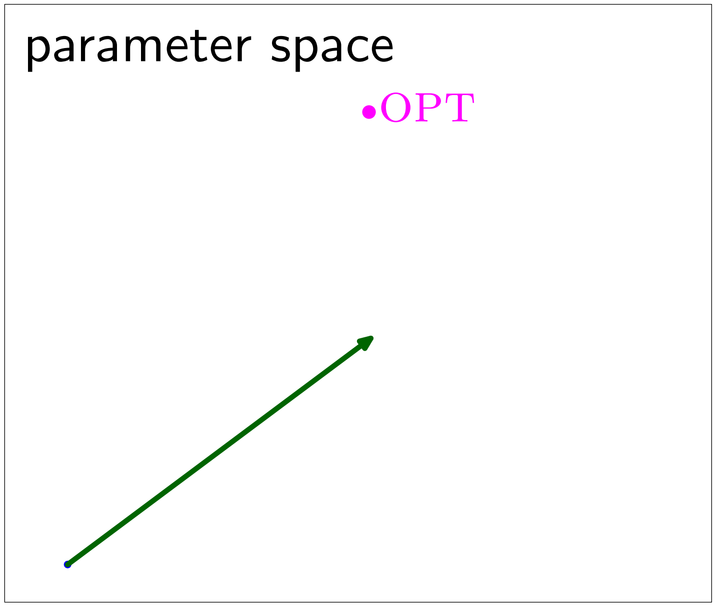


4 data points processed

Batch EM

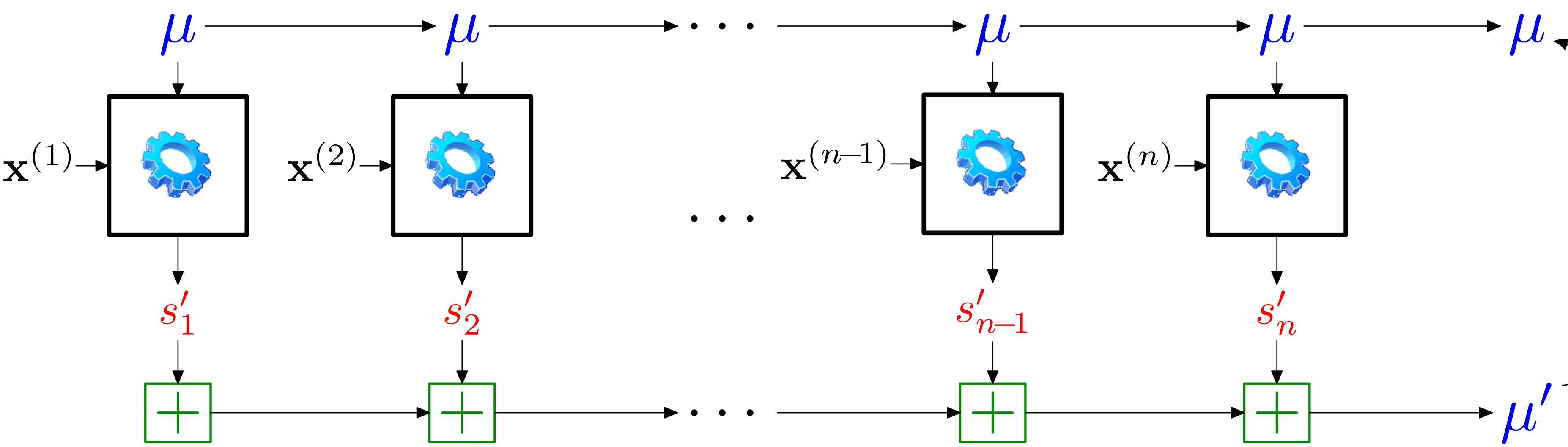


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

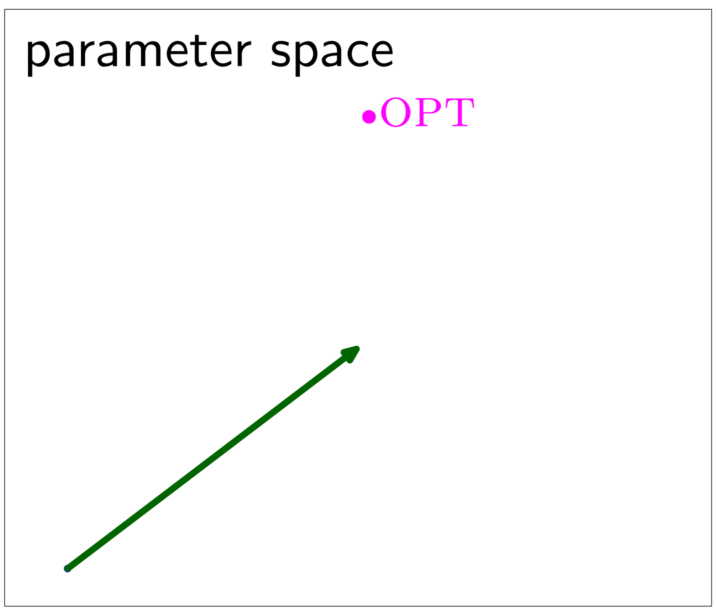


5 data points processed

Batch EM

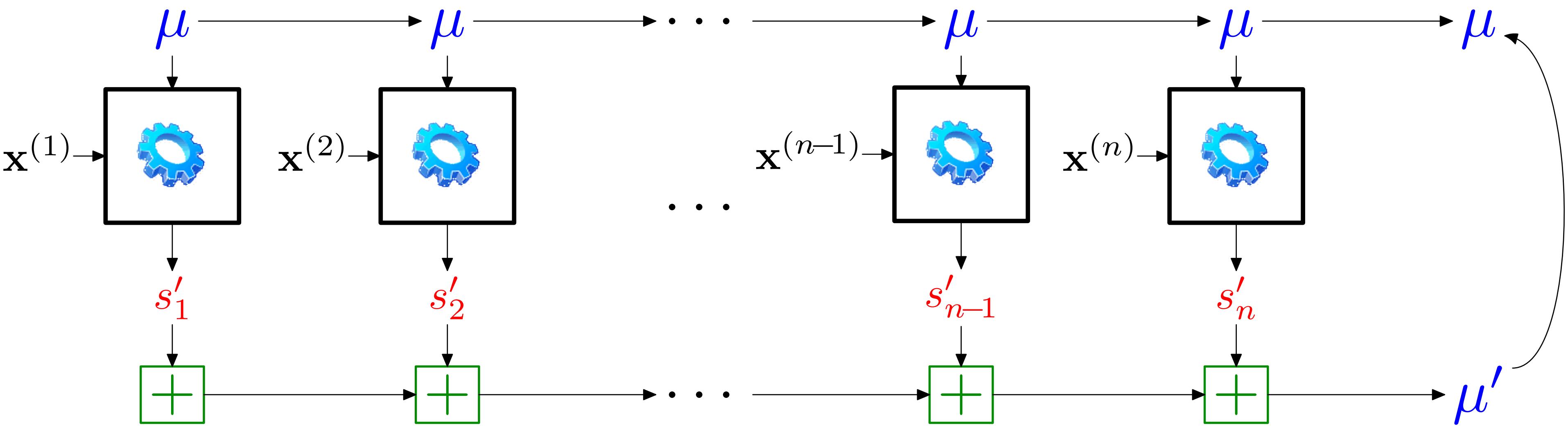


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

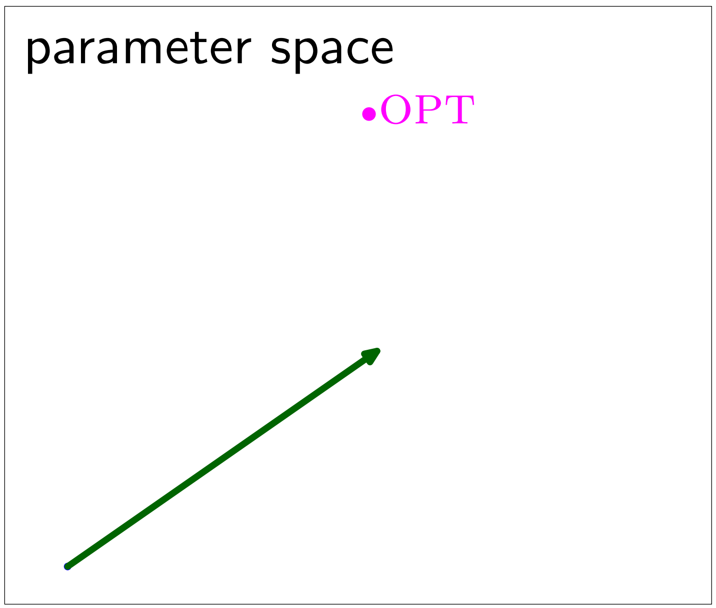


6 data points processed

Batch EM

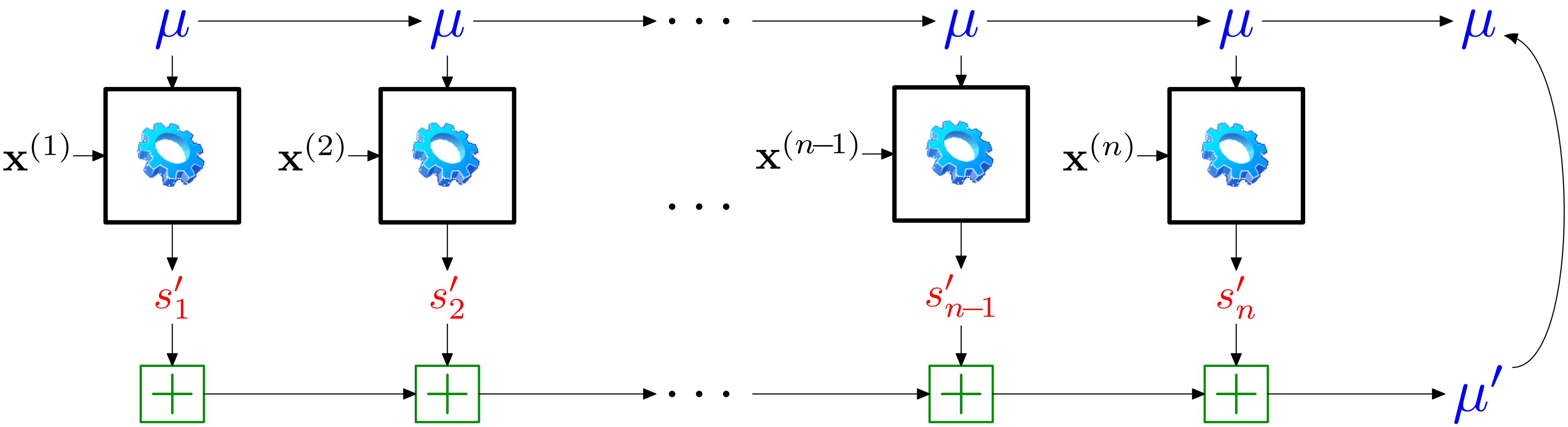


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

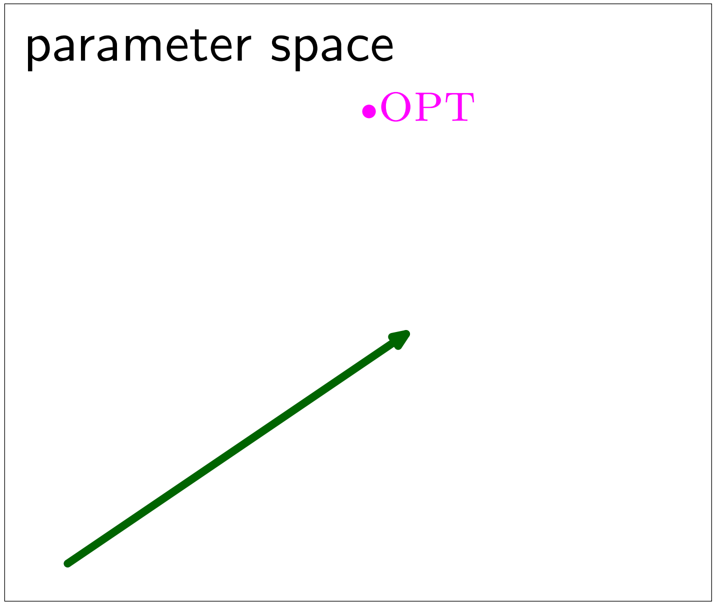


7 data points processed

Batch EM

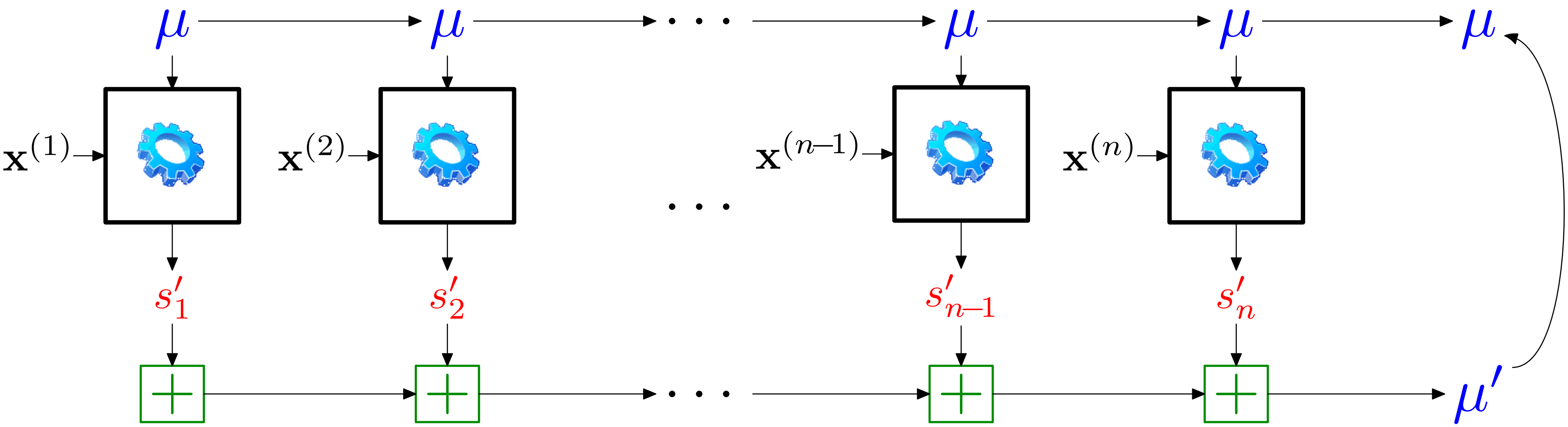


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

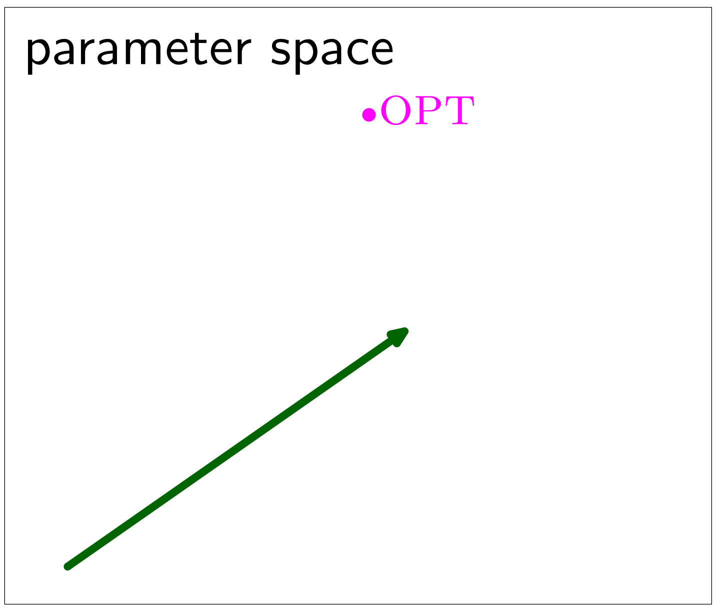


8 data points processed

Batch EM

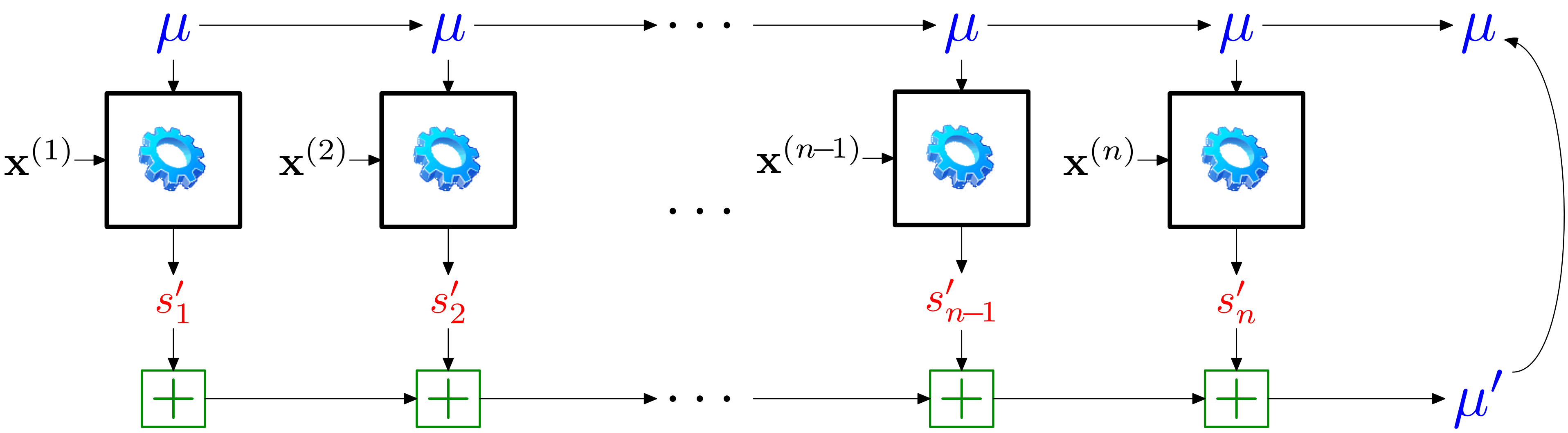


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

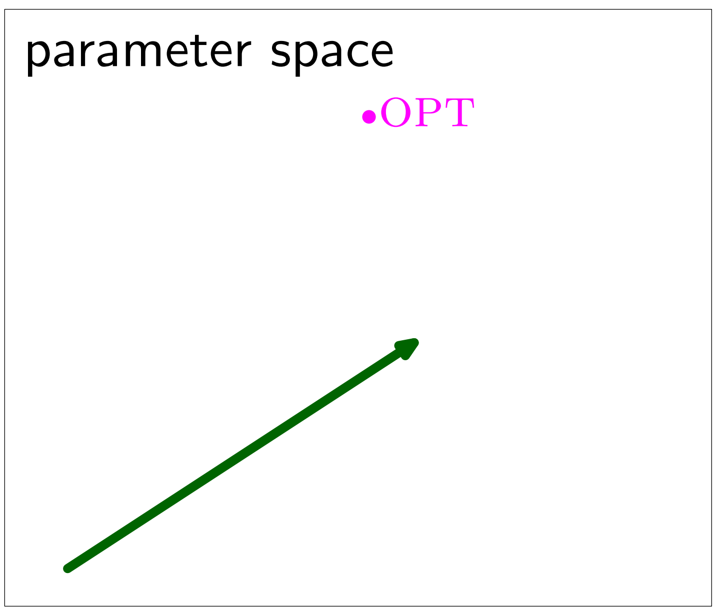


9 data points processed

Batch EM

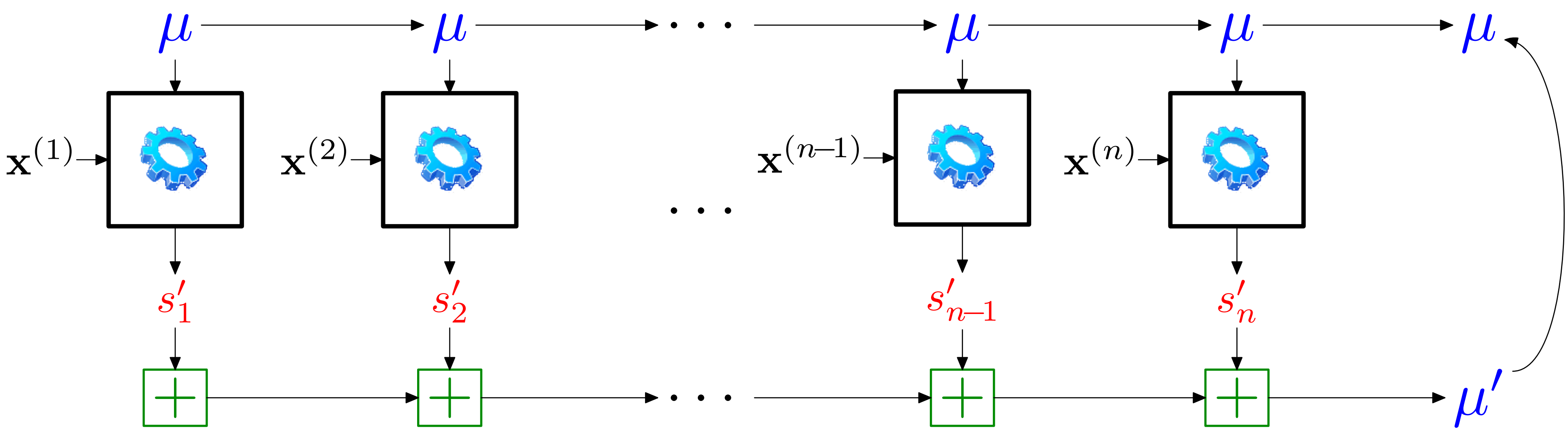


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

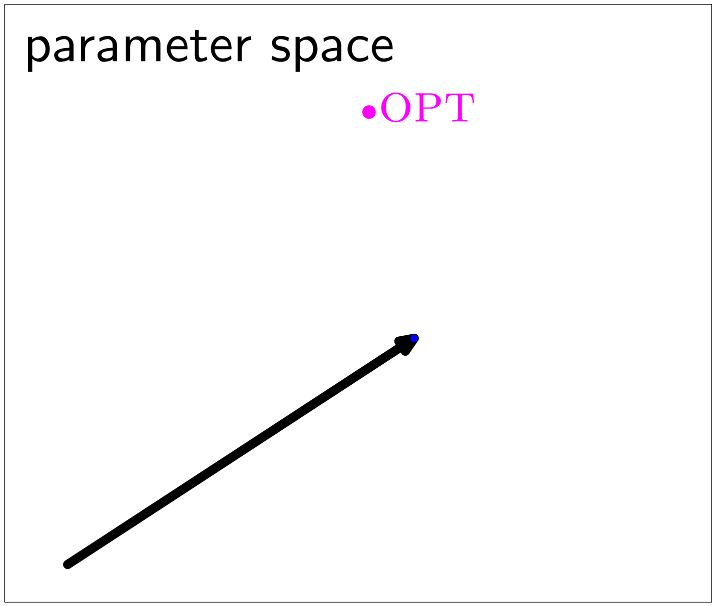


10 data points processed

Batch EM

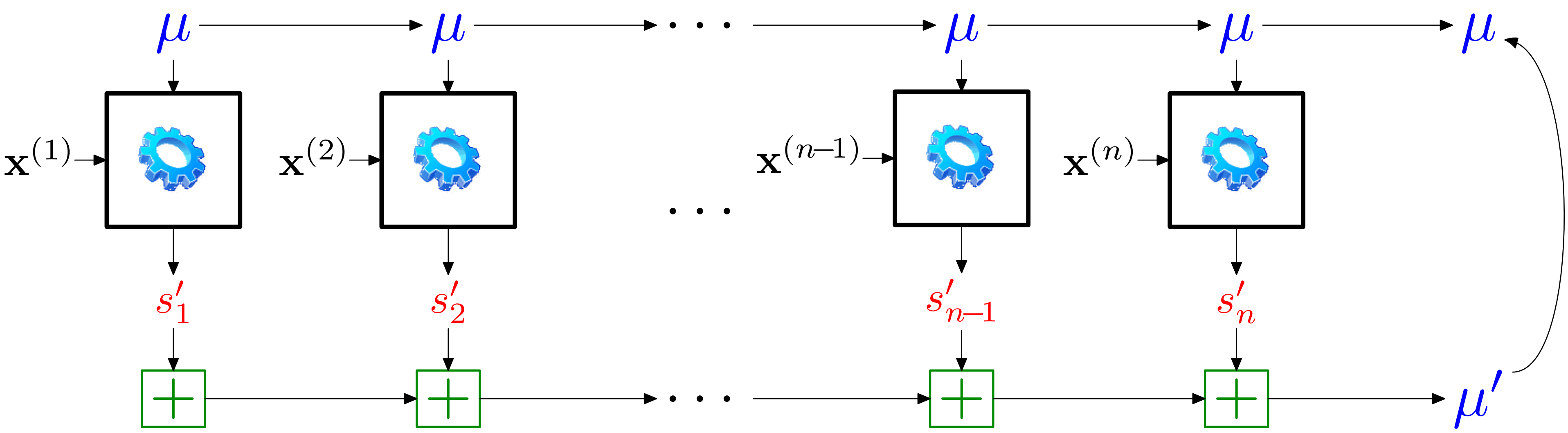


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

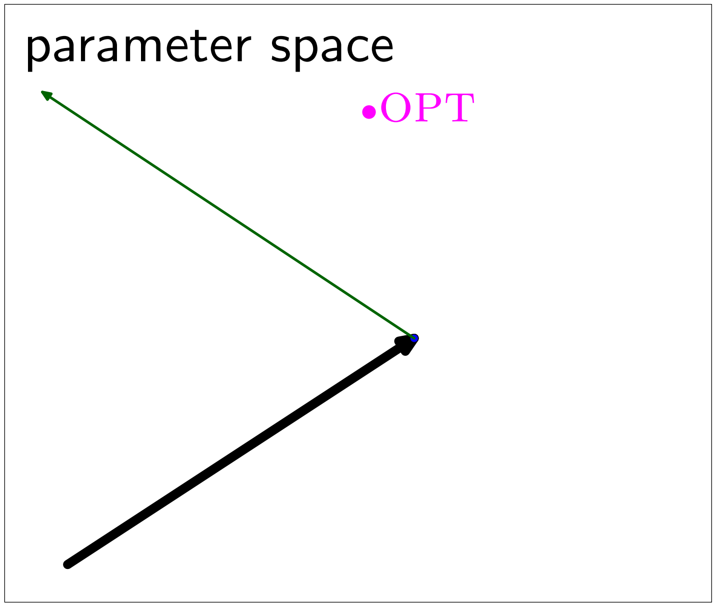


10 data points processed

Batch EM

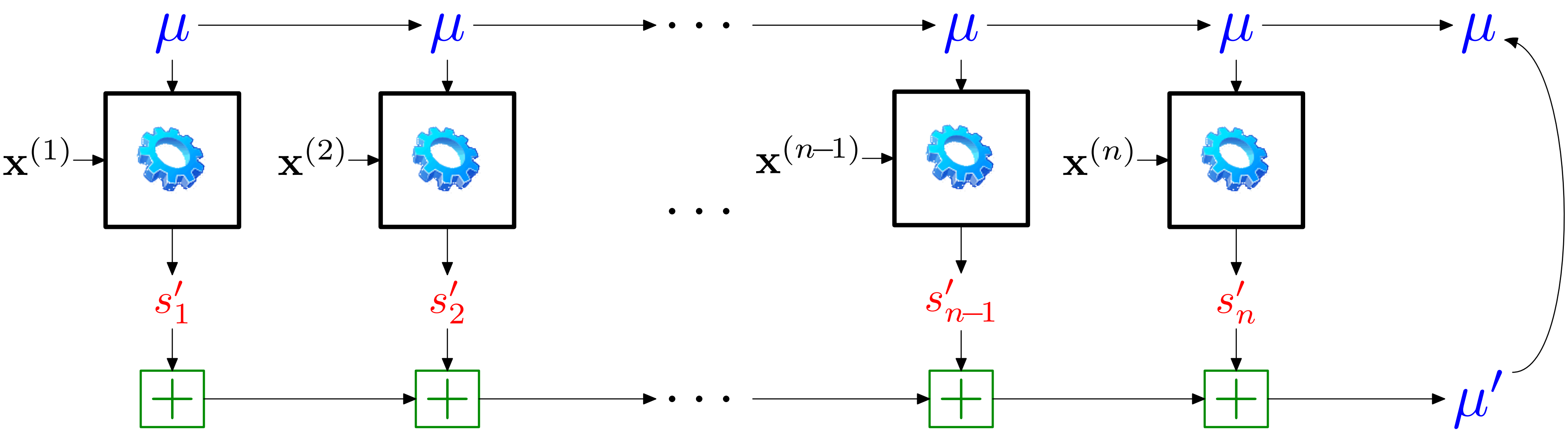


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

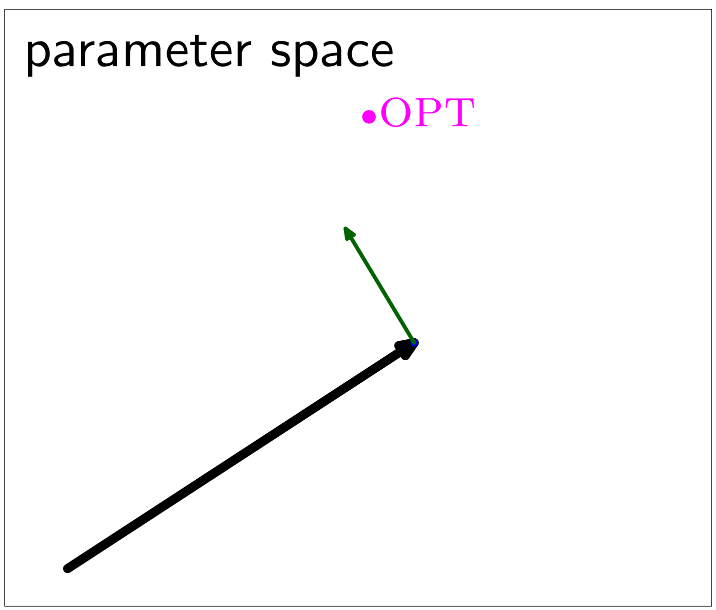


11 data points processed

Batch EM

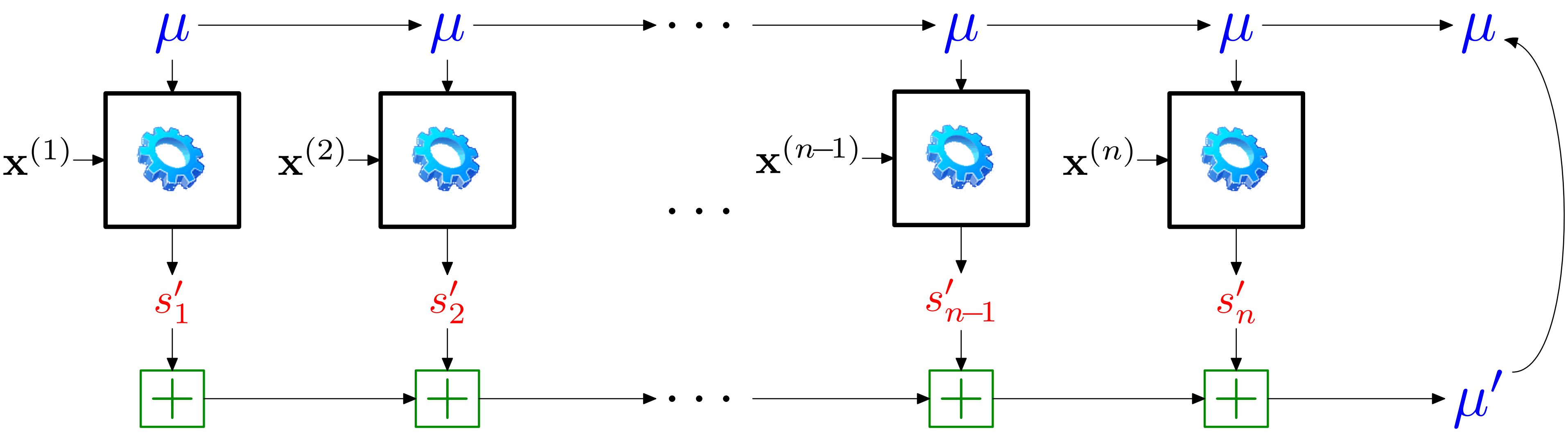


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

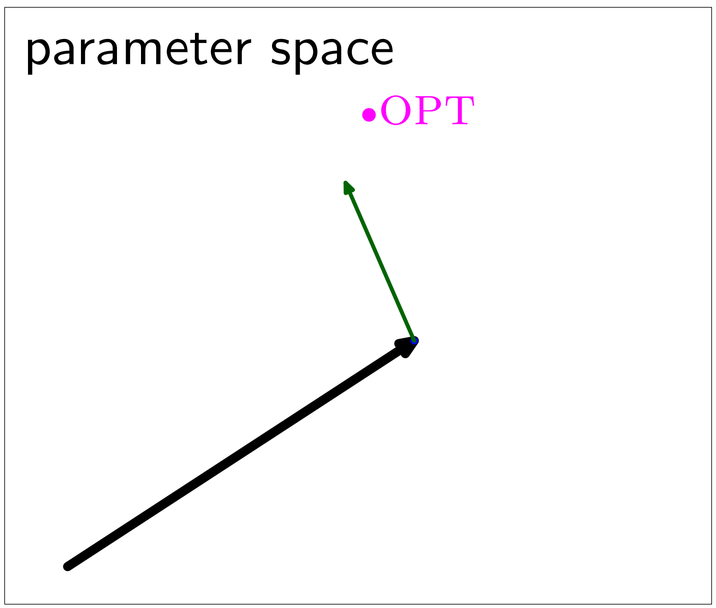


12 data points processed

Batch EM

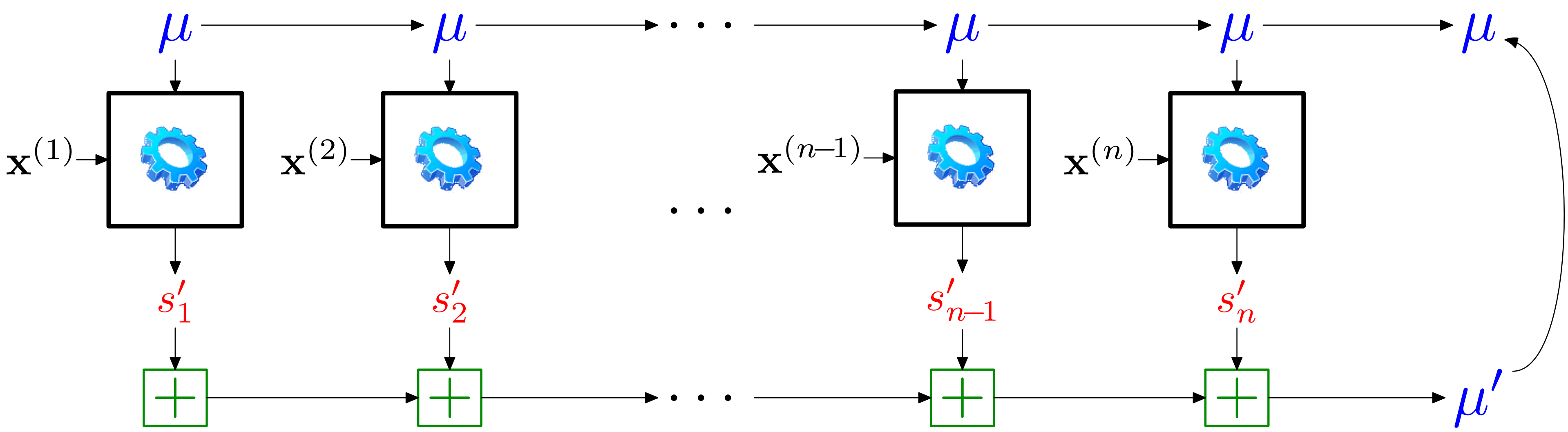


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

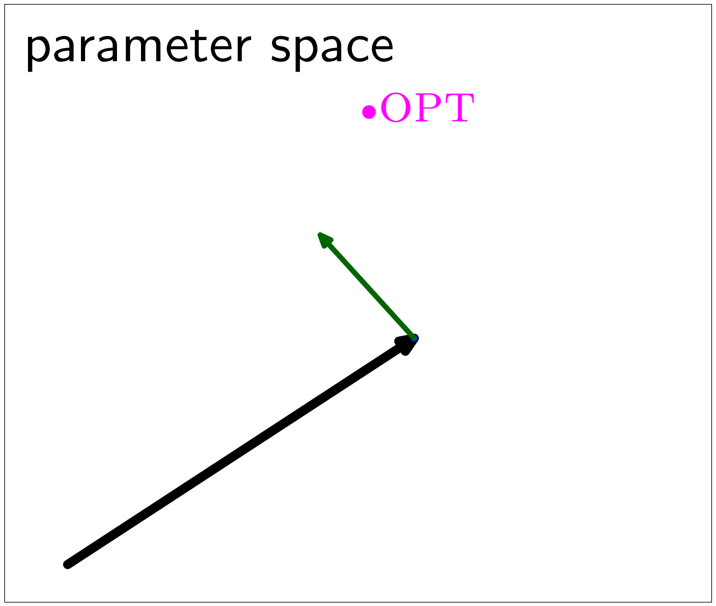


13 data points processed

Batch EM

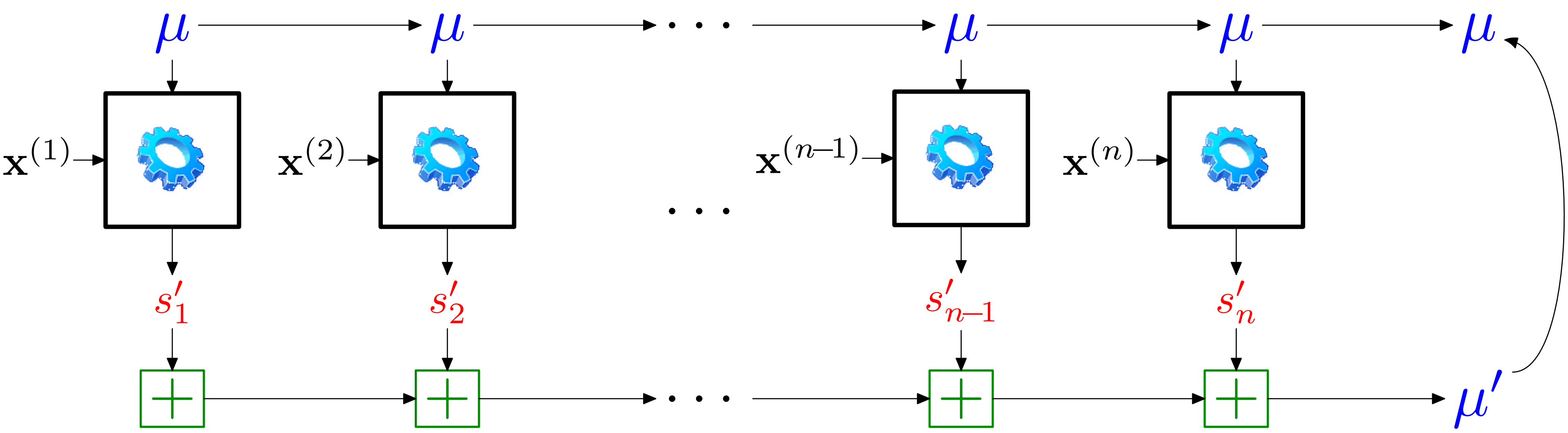


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

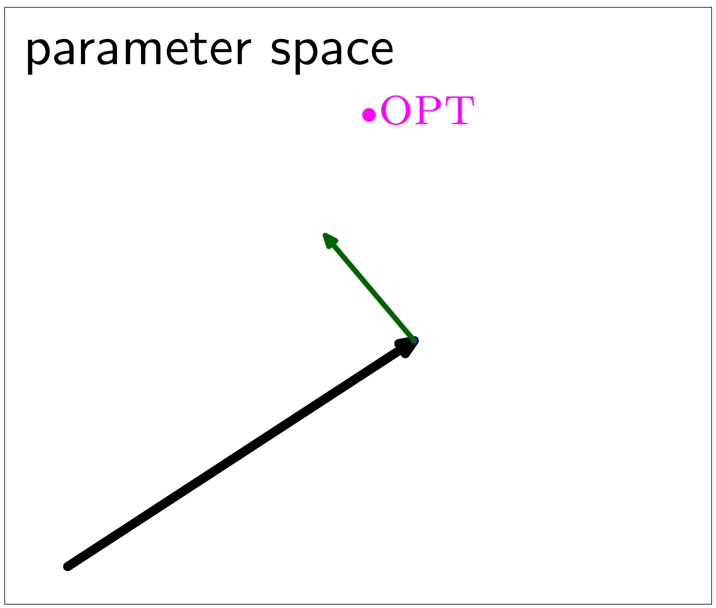


14 data points processed

Batch EM

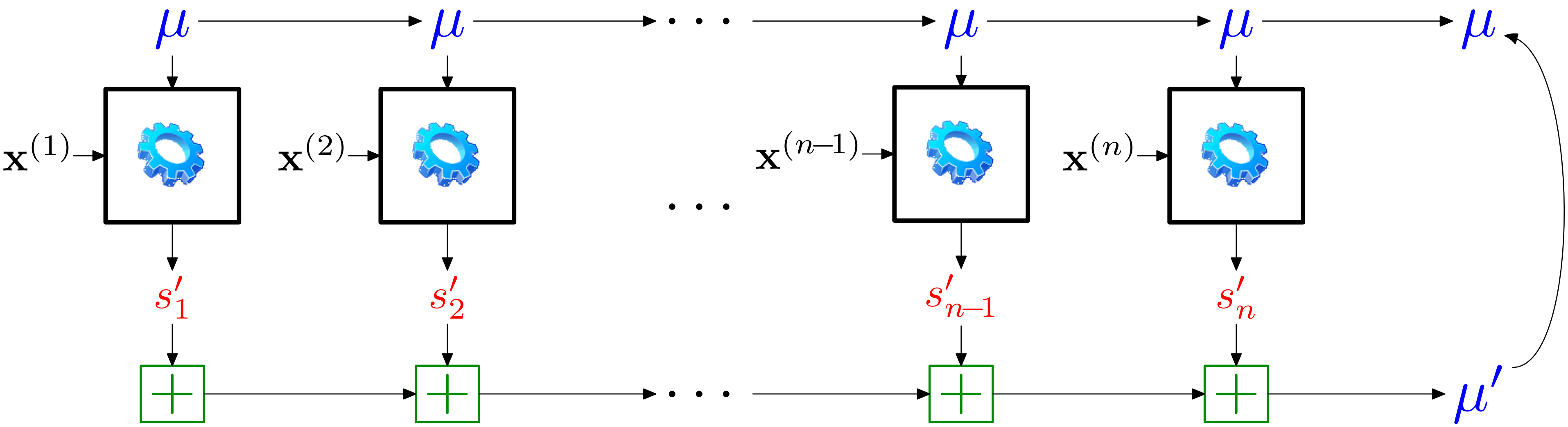


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

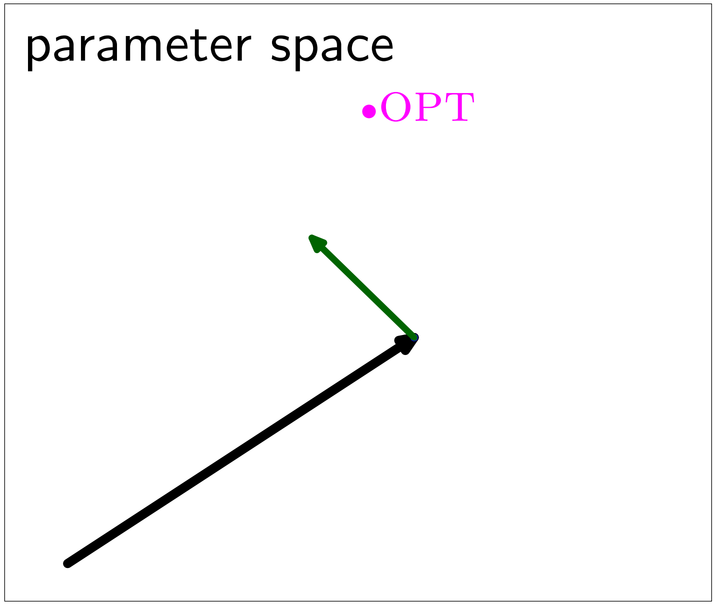


15 data points processed

Batch EM

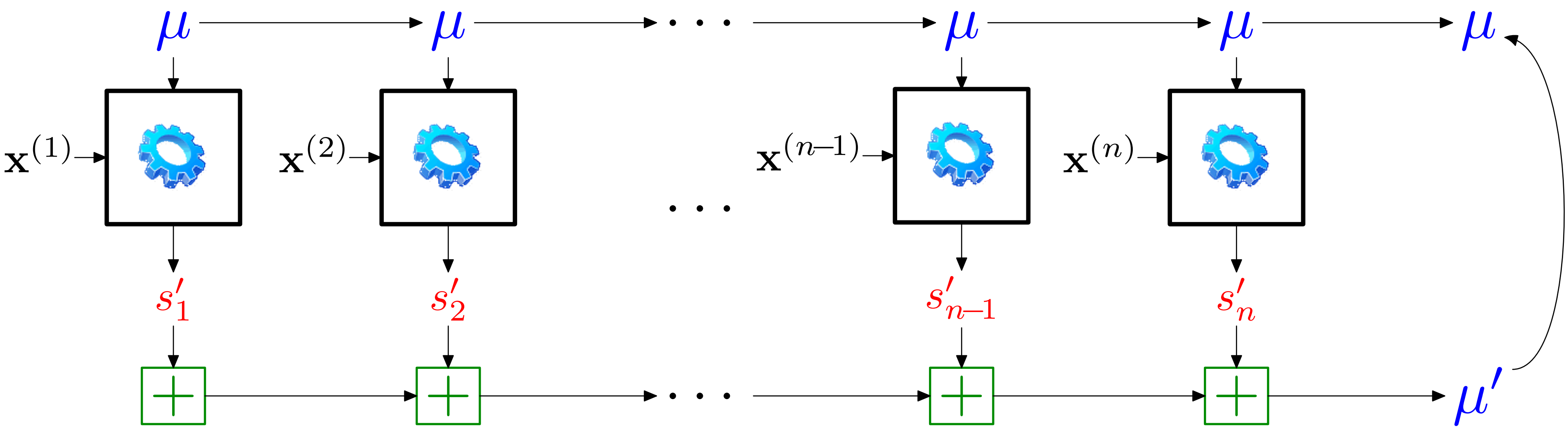


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

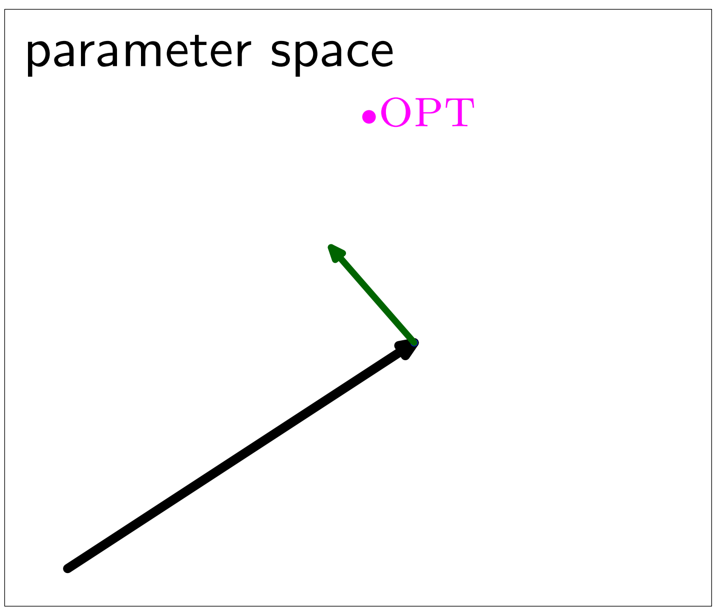


16 data points processed

Batch EM

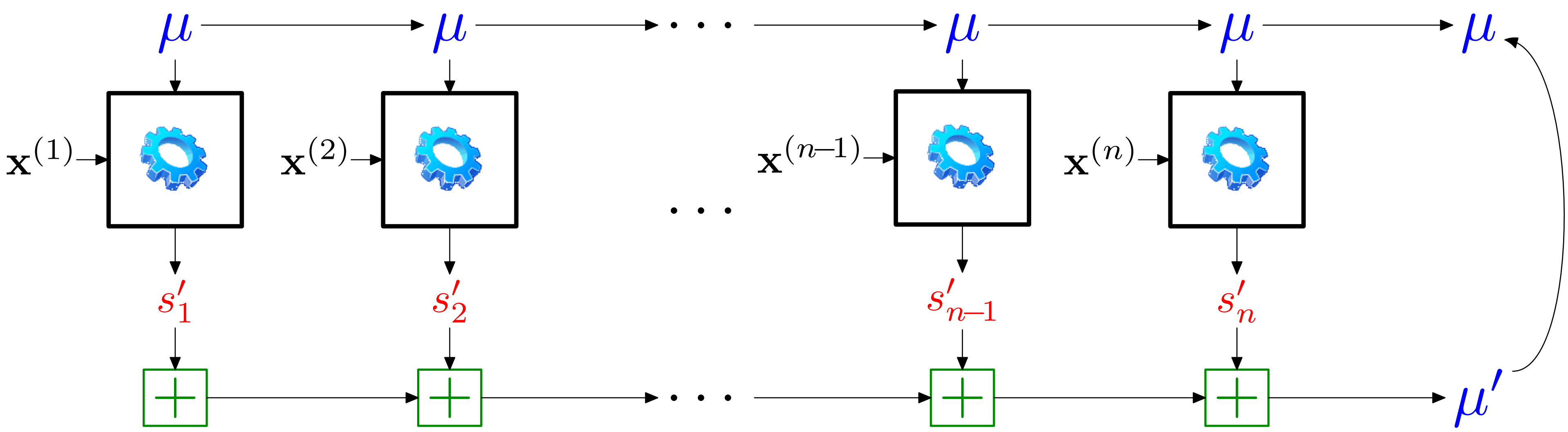


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

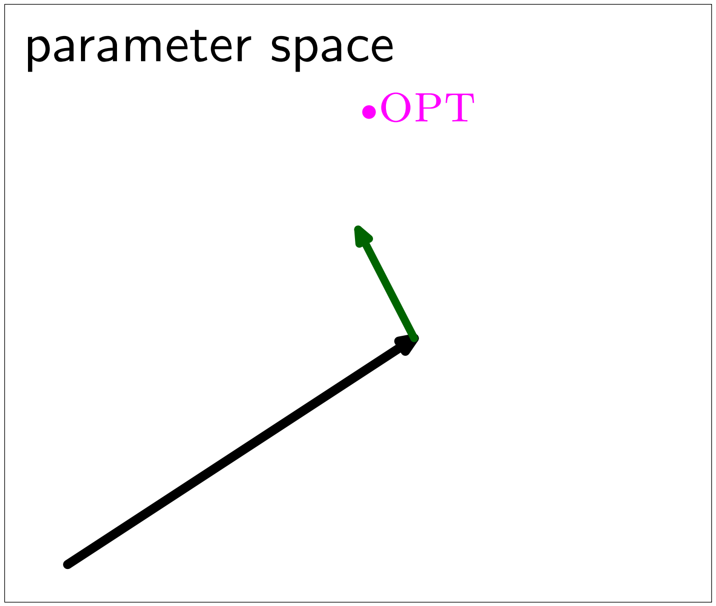


17 data points processed

Batch EM

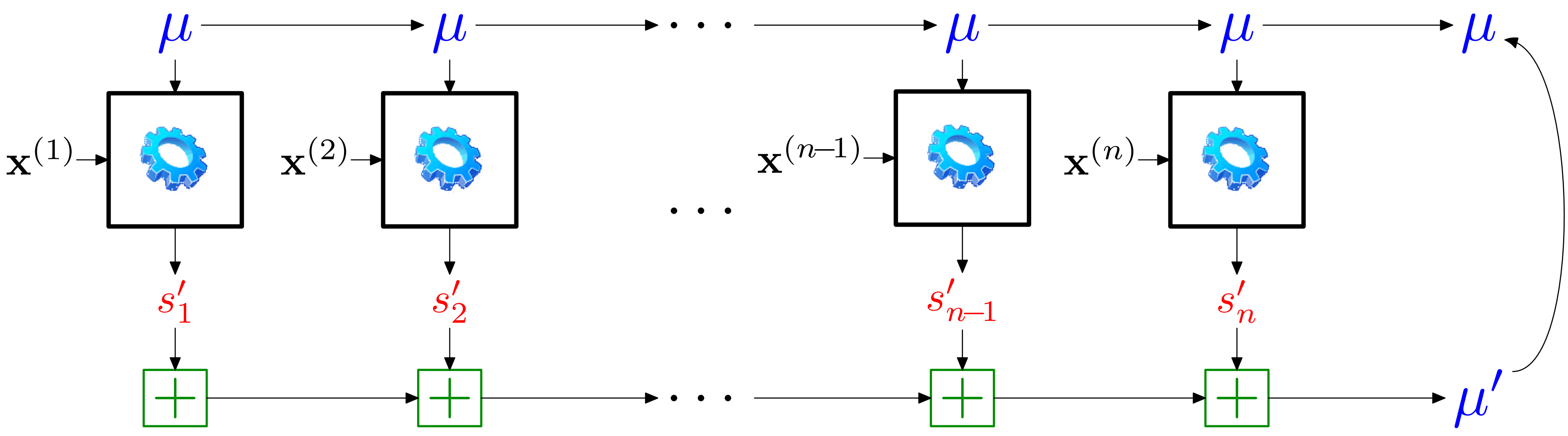


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

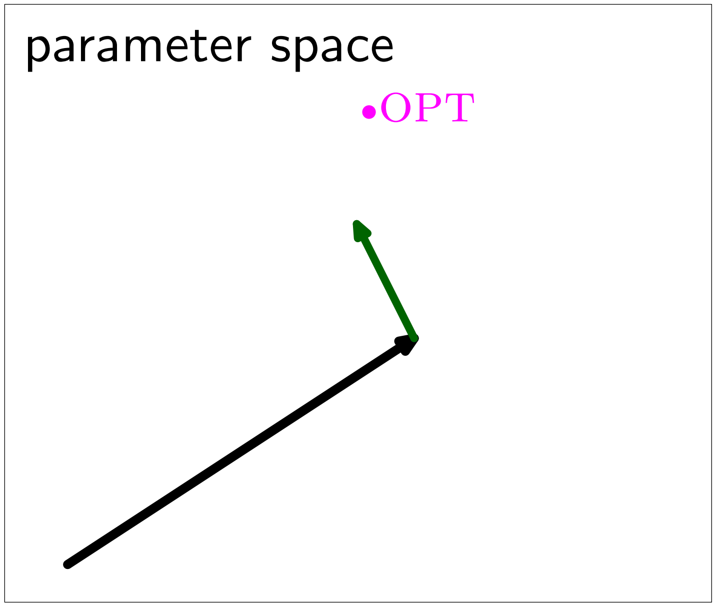


18 data points processed

Batch EM

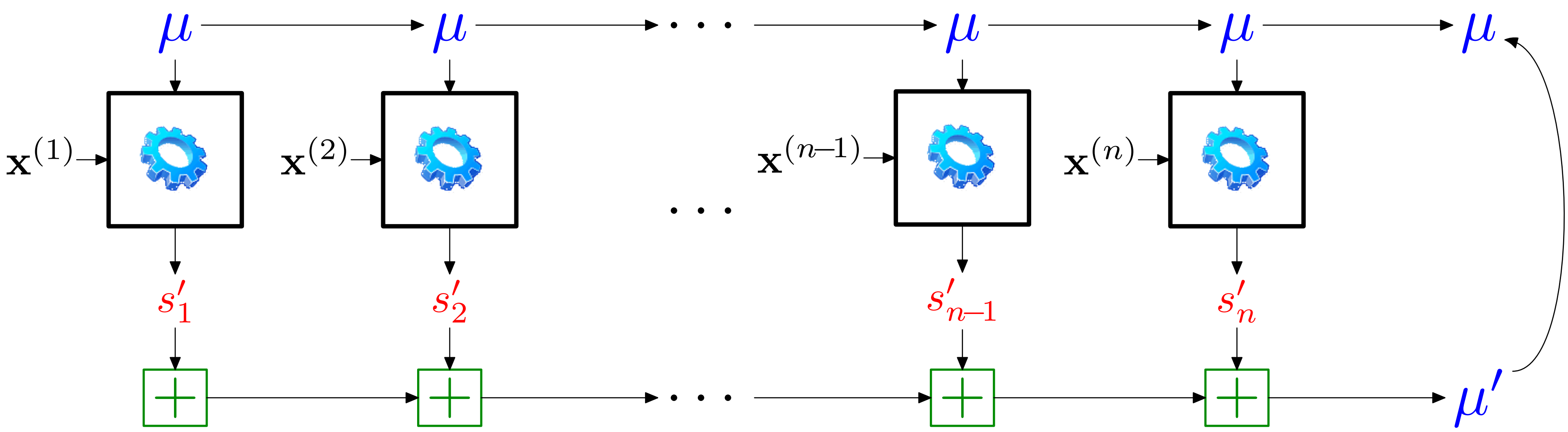


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

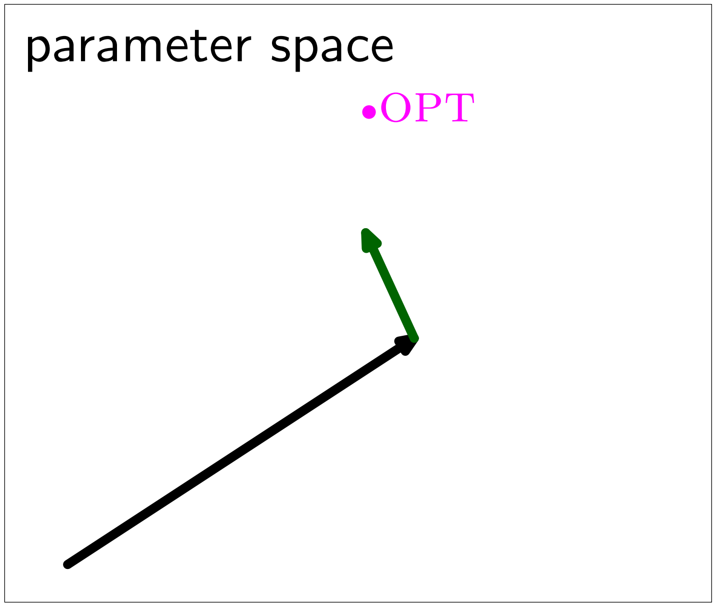


19 data points processed

Batch EM

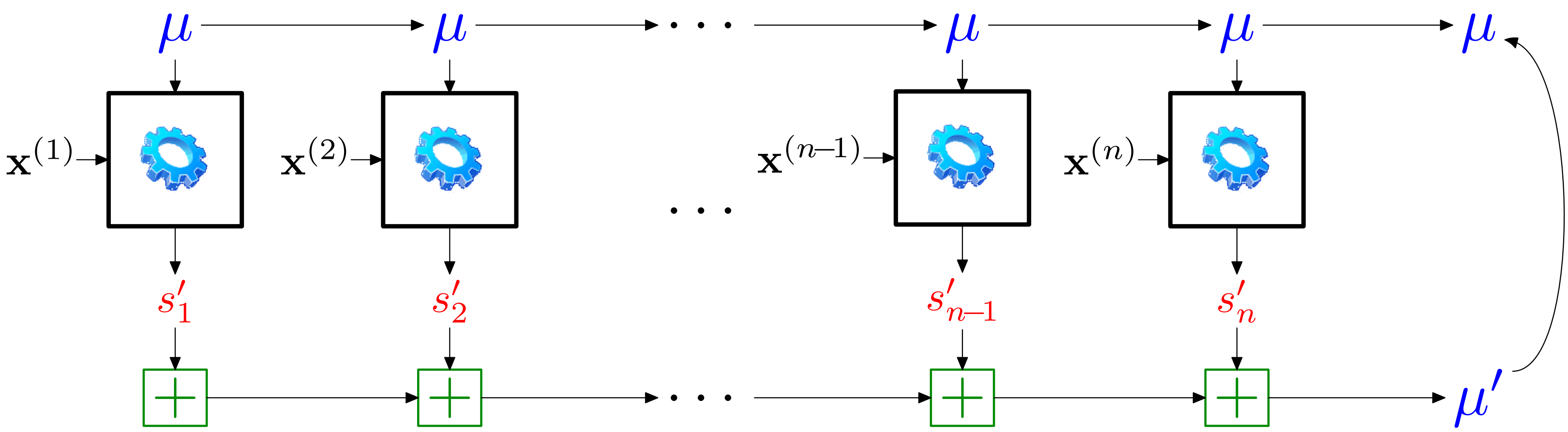


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

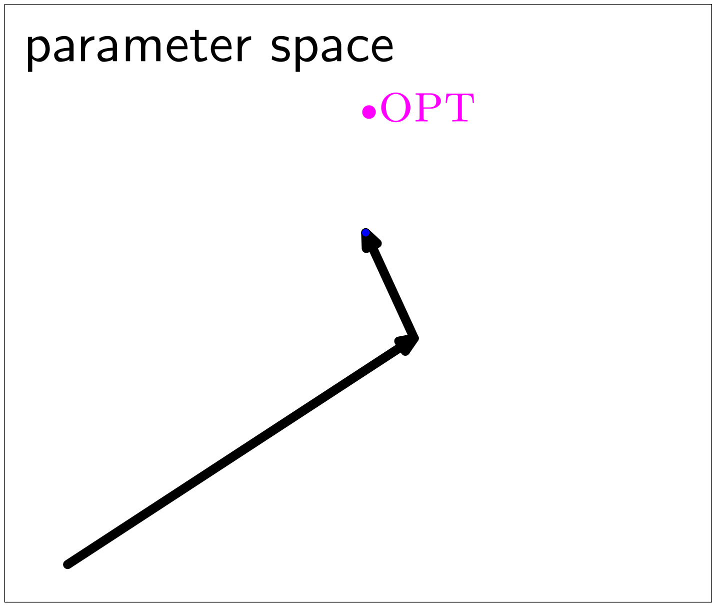


20 data points processed

Batch EM

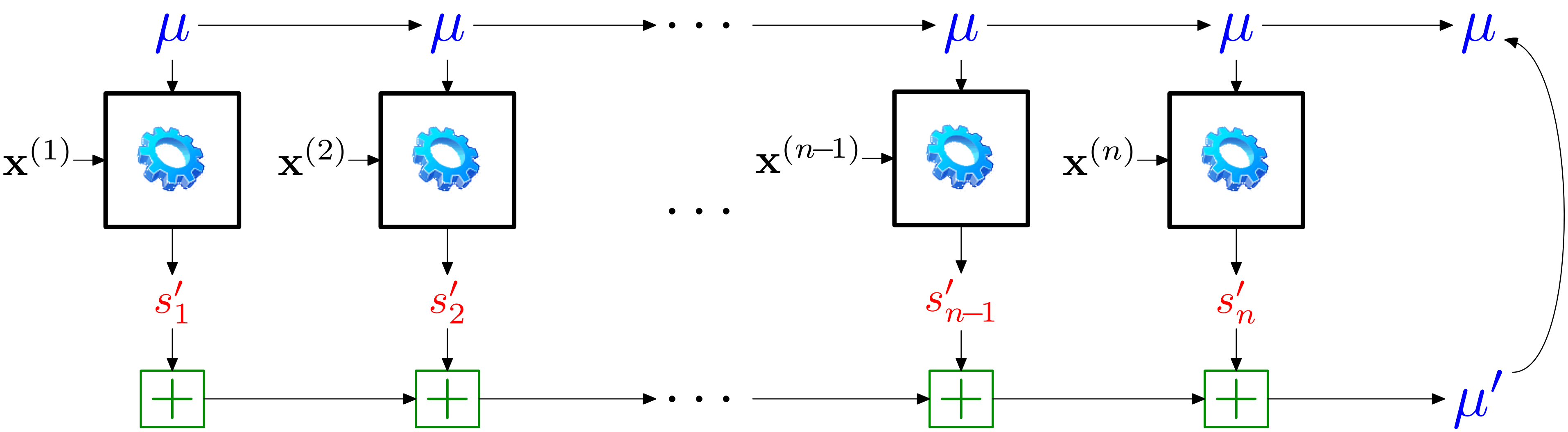


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

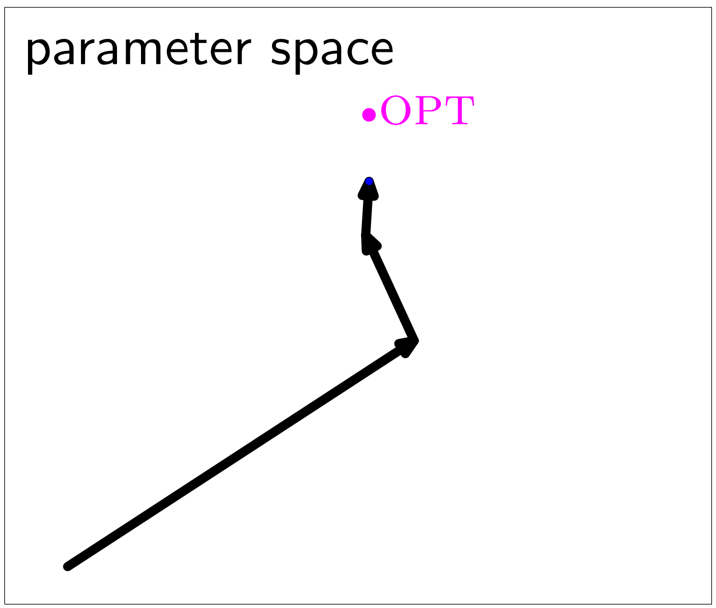


20 data points processed

Batch EM

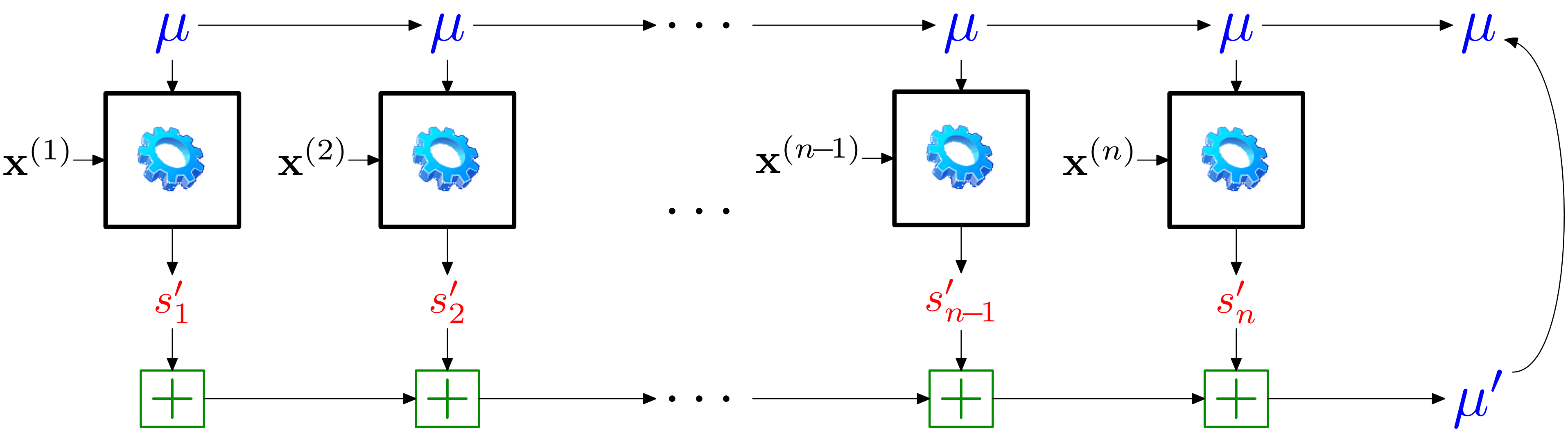


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

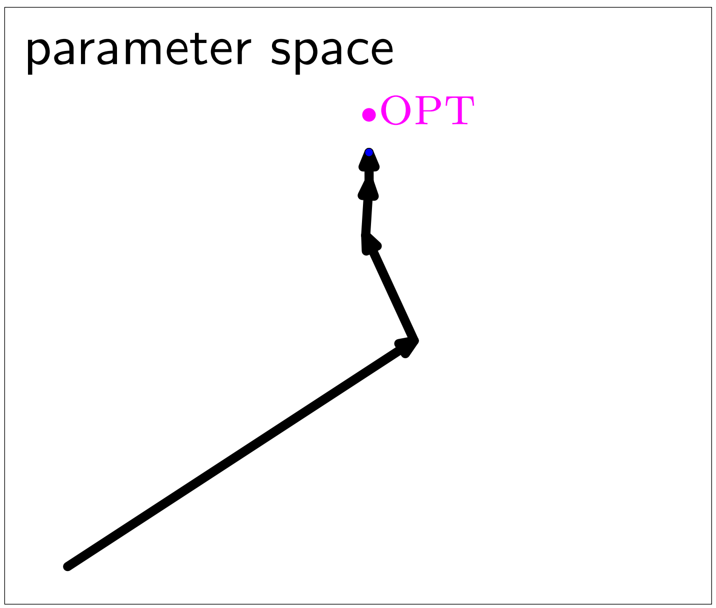


30 data points processed

Batch EM

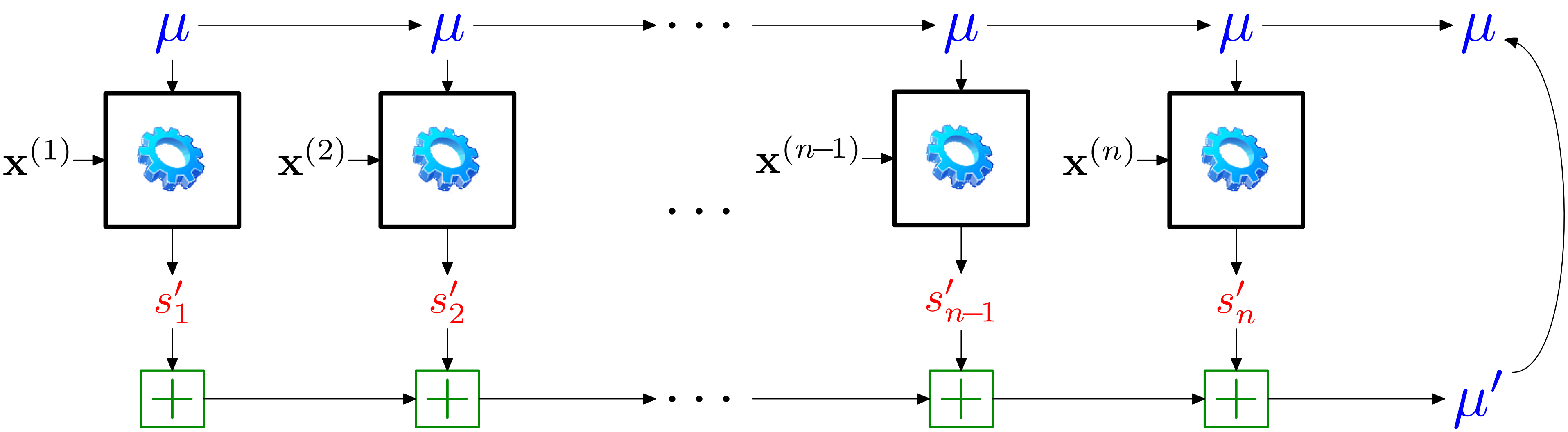


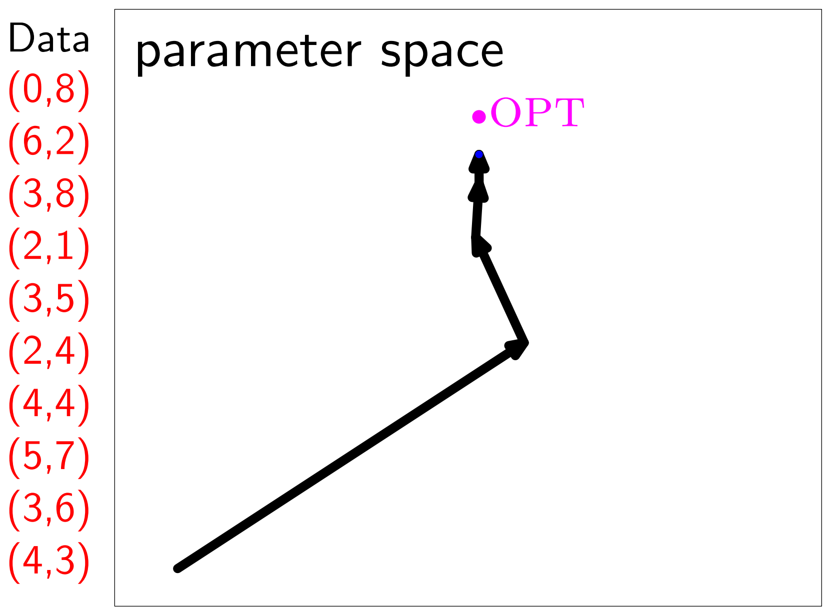
Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



40 data points processed

Batch EM

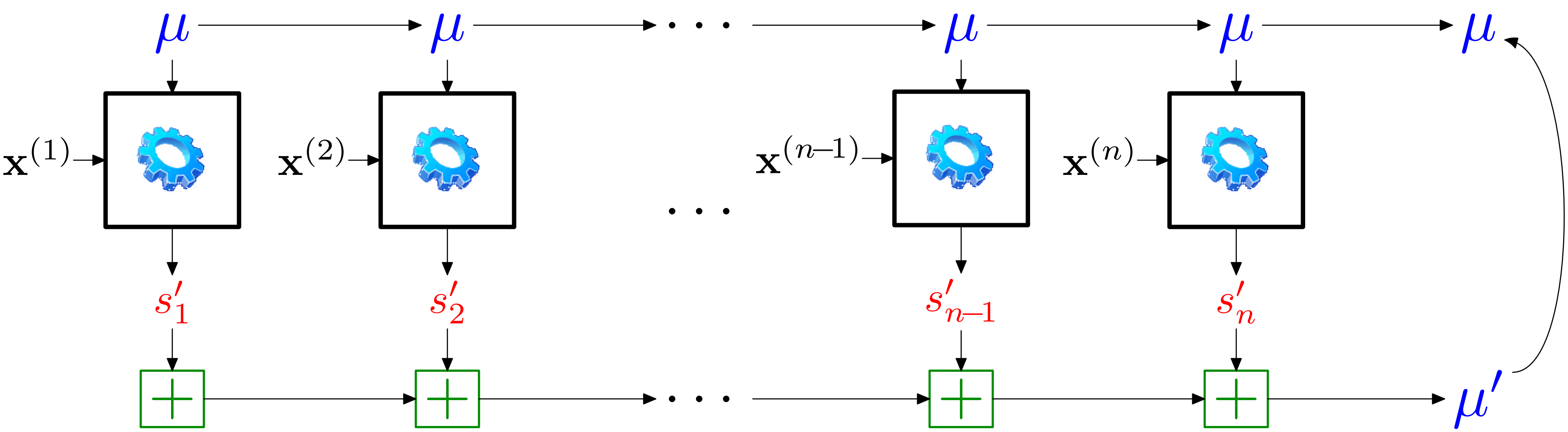


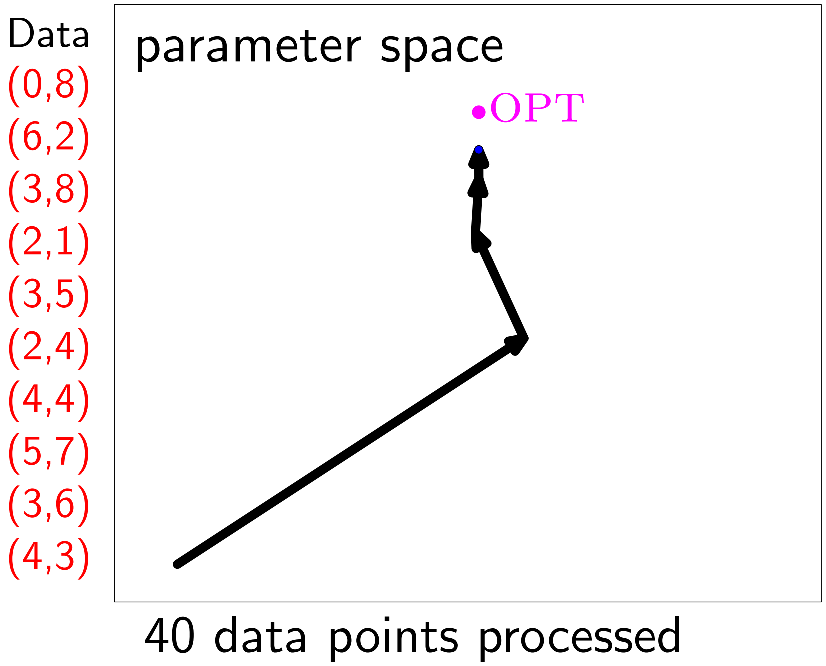


40 data points processed

- Spend a lot of time computing new parameters exactly, but have rough estimate much earlier

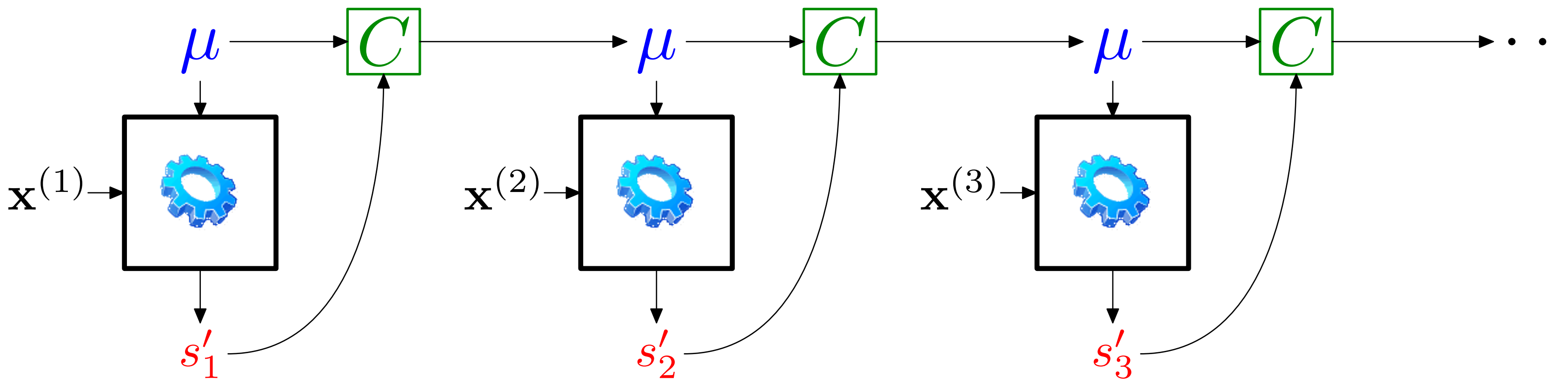
Batch EM



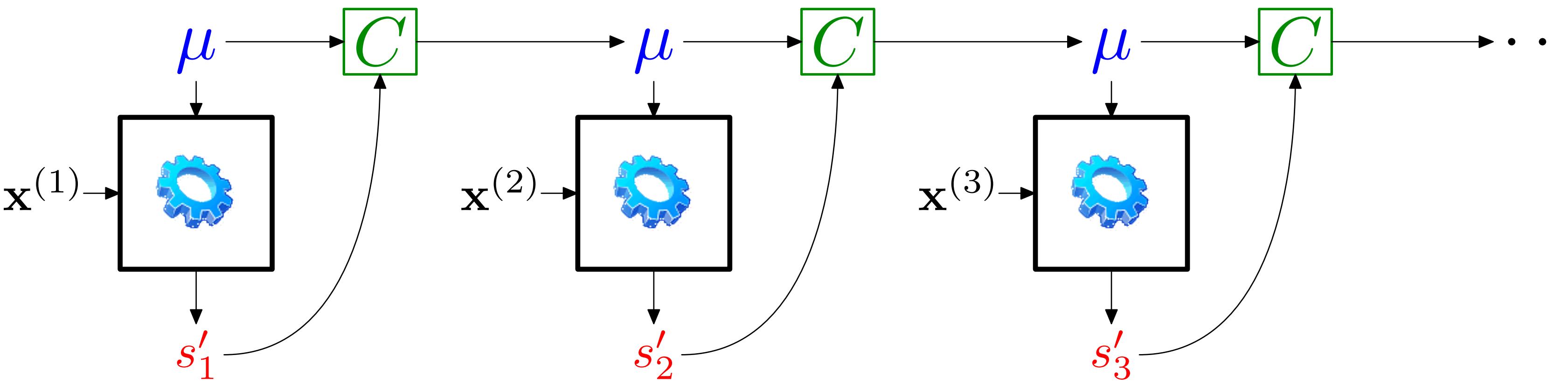


- Spend a lot of time computing new parameters exactly, but have rough estimate much earlier
- New parameters are intermediate, so don't need to obsess about the exact value

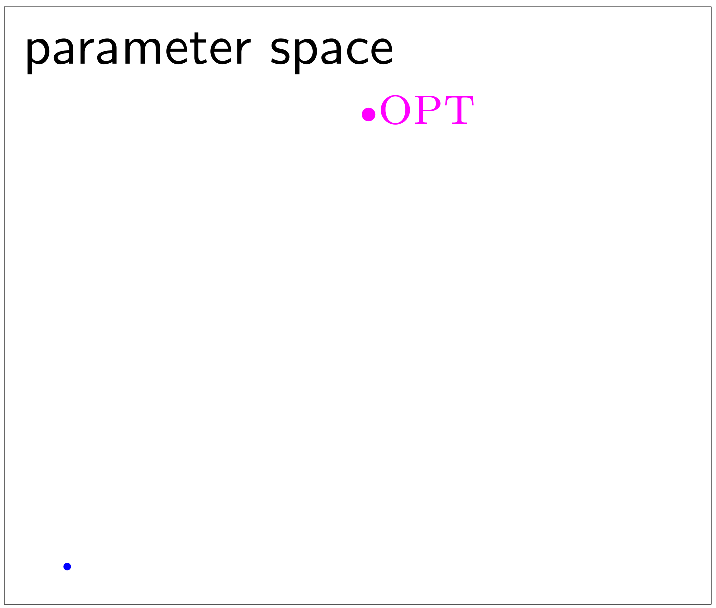
Online EM [Cappé & Moulines, 2009]



Online EM [Cappé & Moulines, 2009]

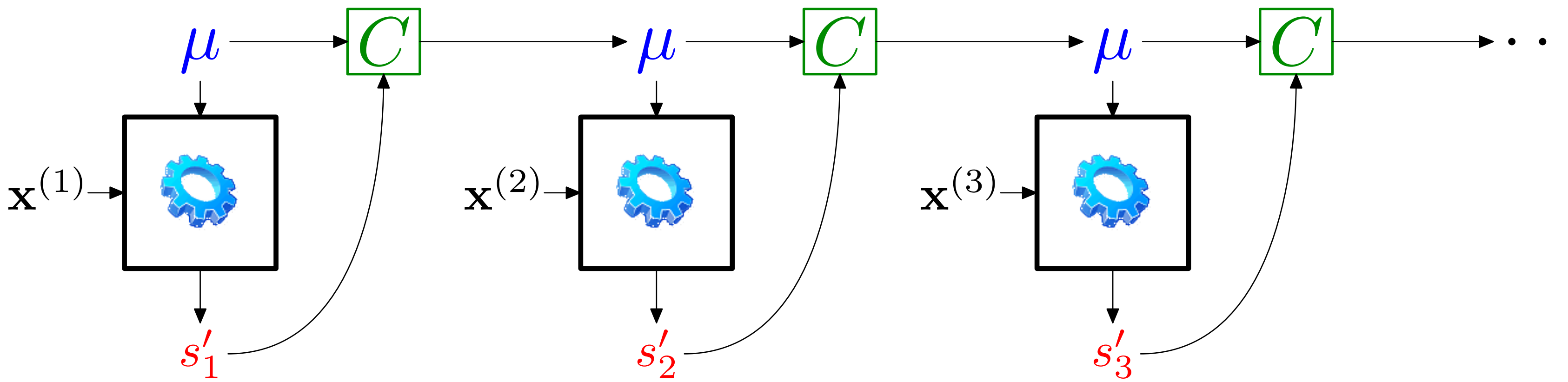


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

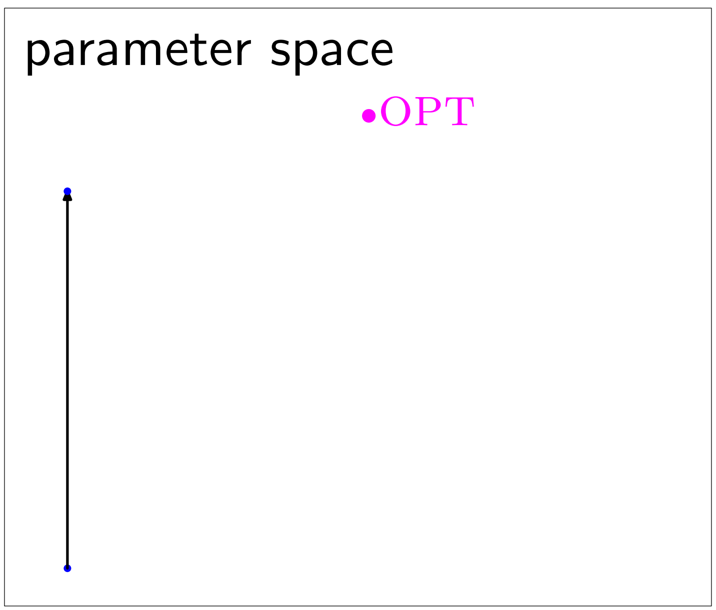


0 data points processed

Online EM [Cappé & Moulines, 2009]

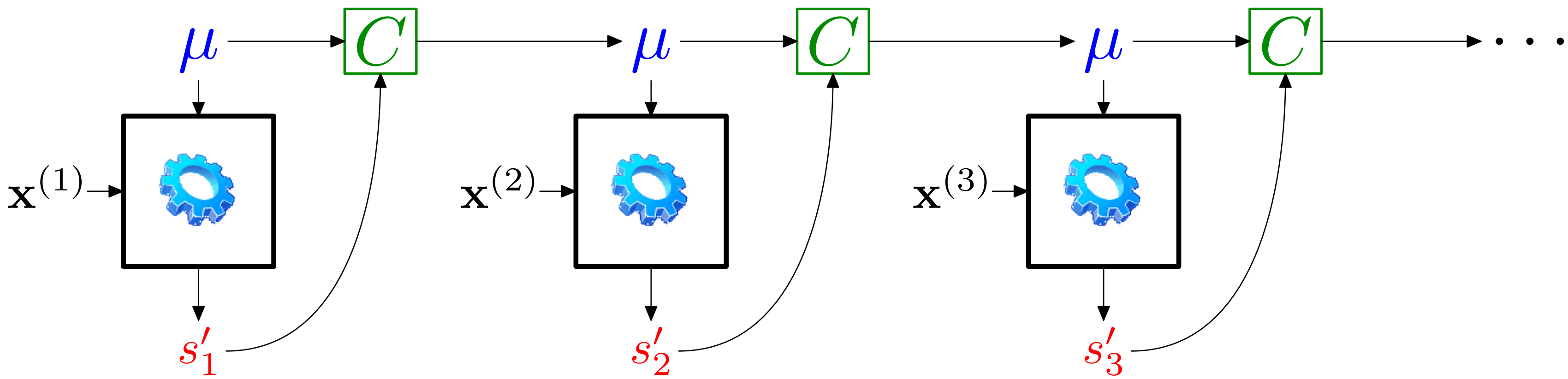


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

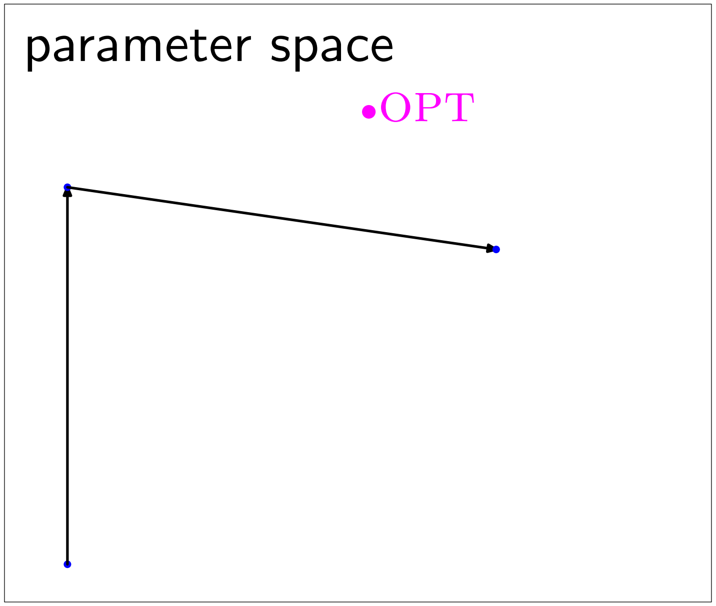


1 data points processed

Online EM [Cappé & Moulines, 2009]

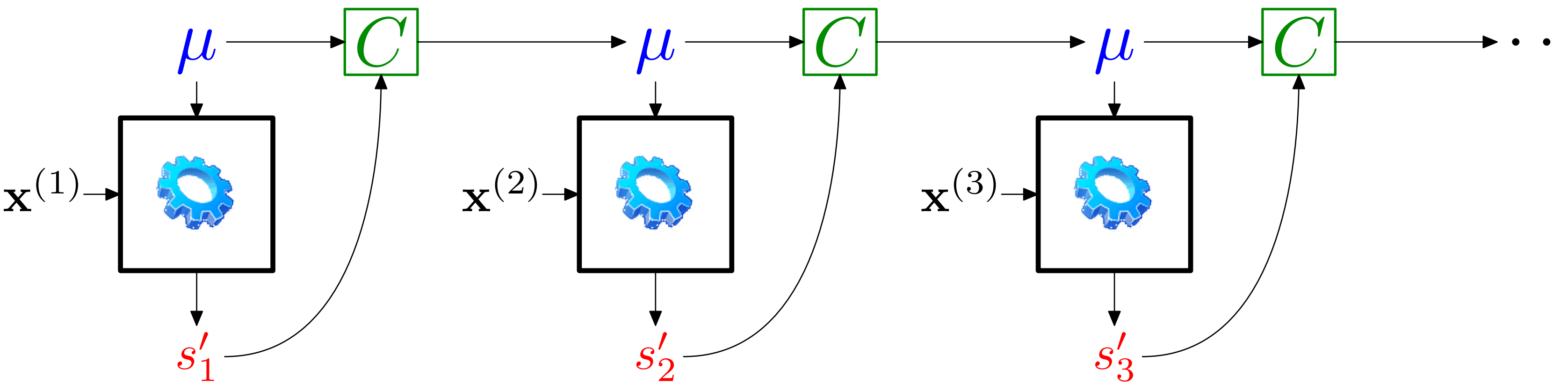


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

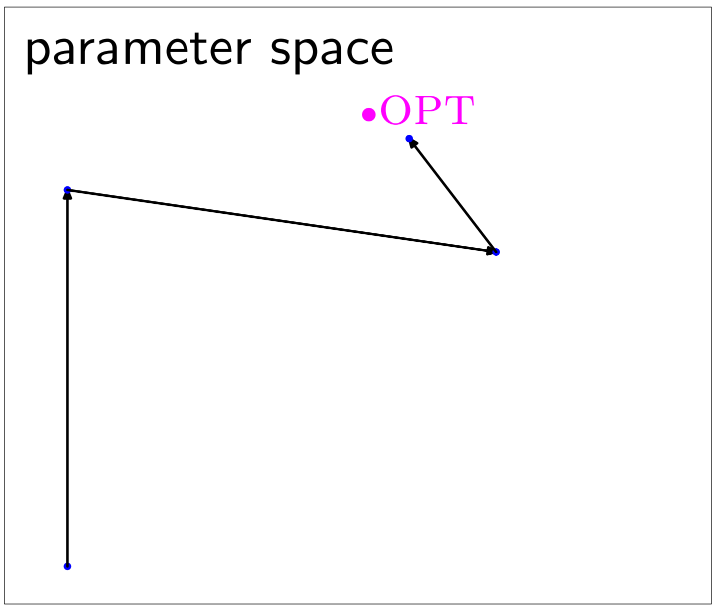


2 data points processed

Online EM [Cappé & Moulines, 2009]

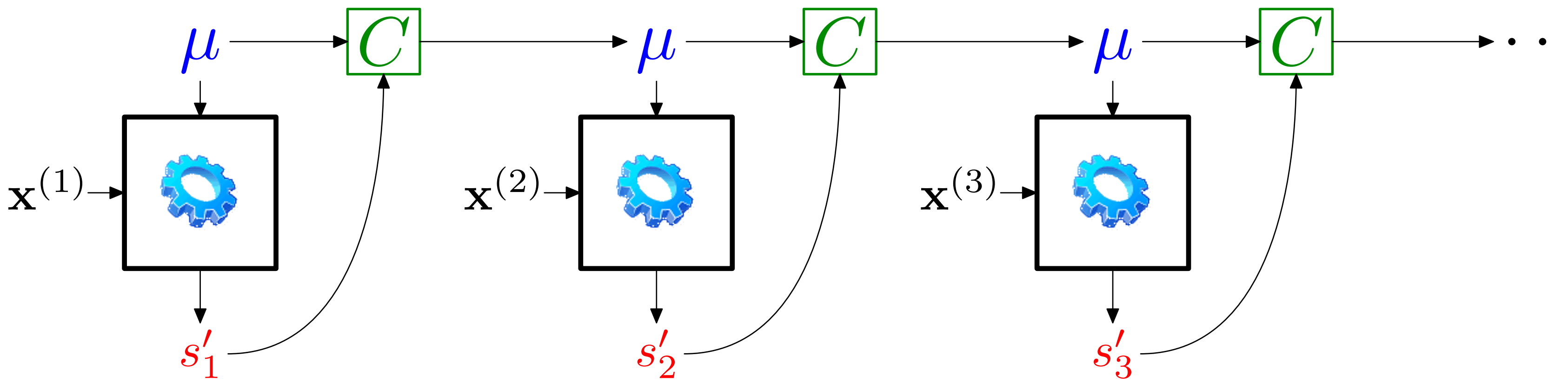


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

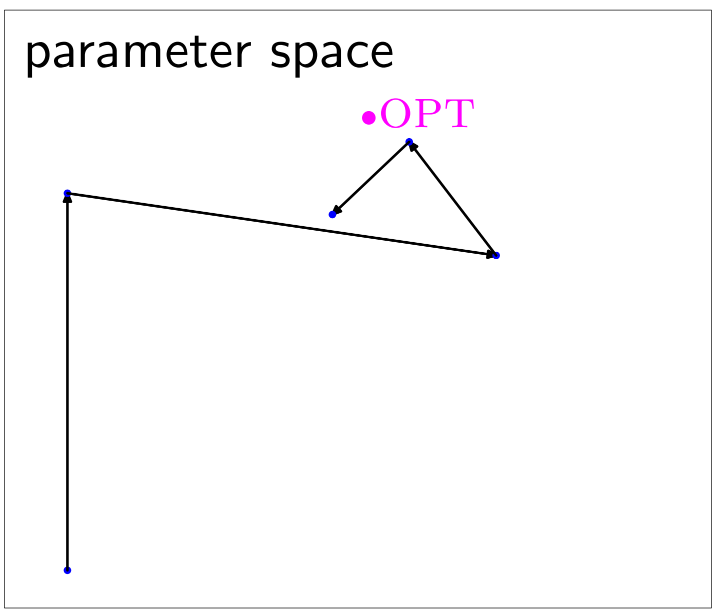


3 data points processed

Online EM [Cappé & Moulines, 2009]

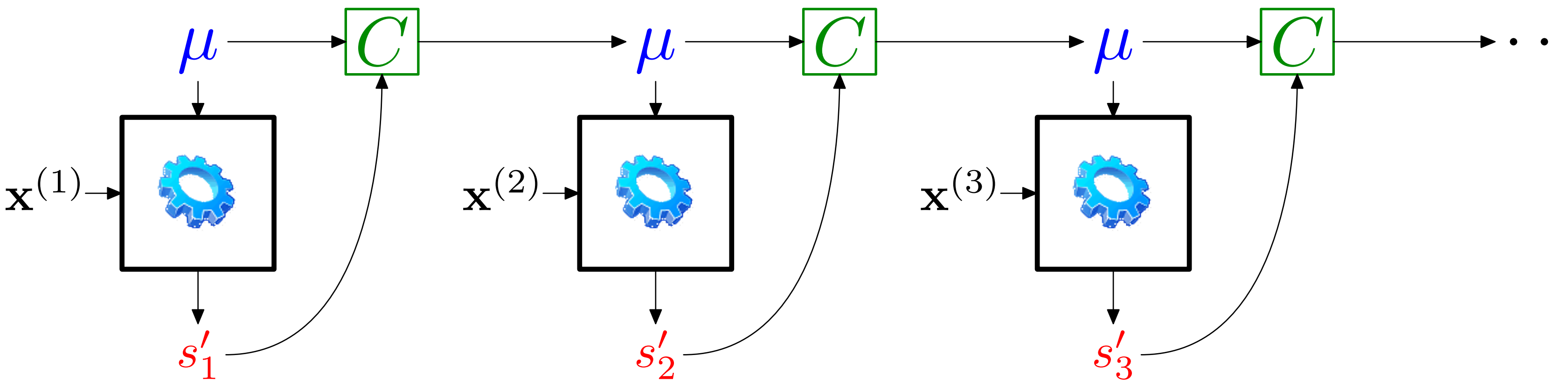


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

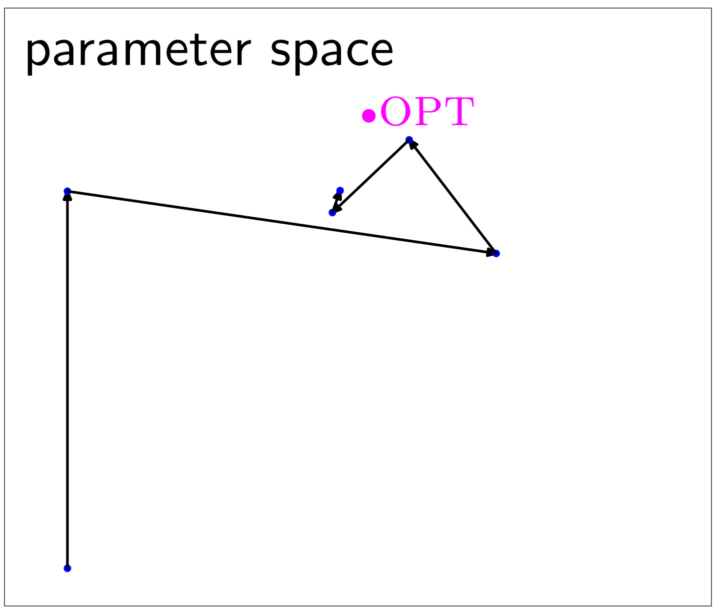


4 data points processed

Online EM [Cappé & Moulines, 2009]

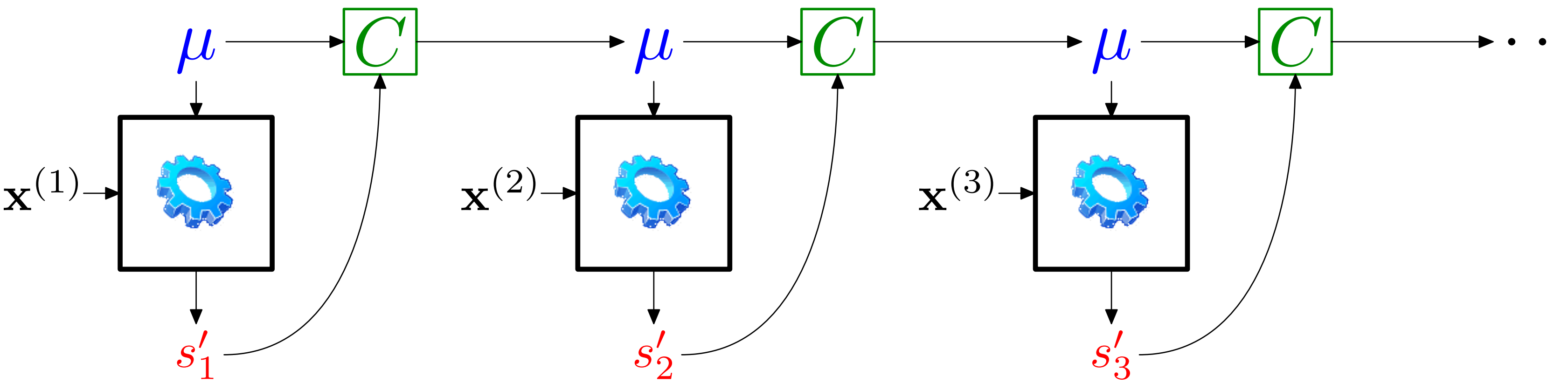


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

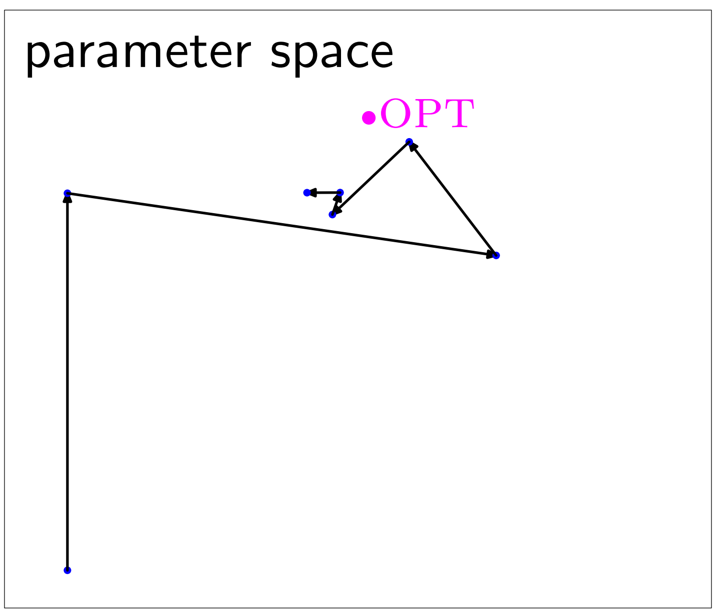


5 data points processed

Online EM [Cappé & Moulines, 2009]

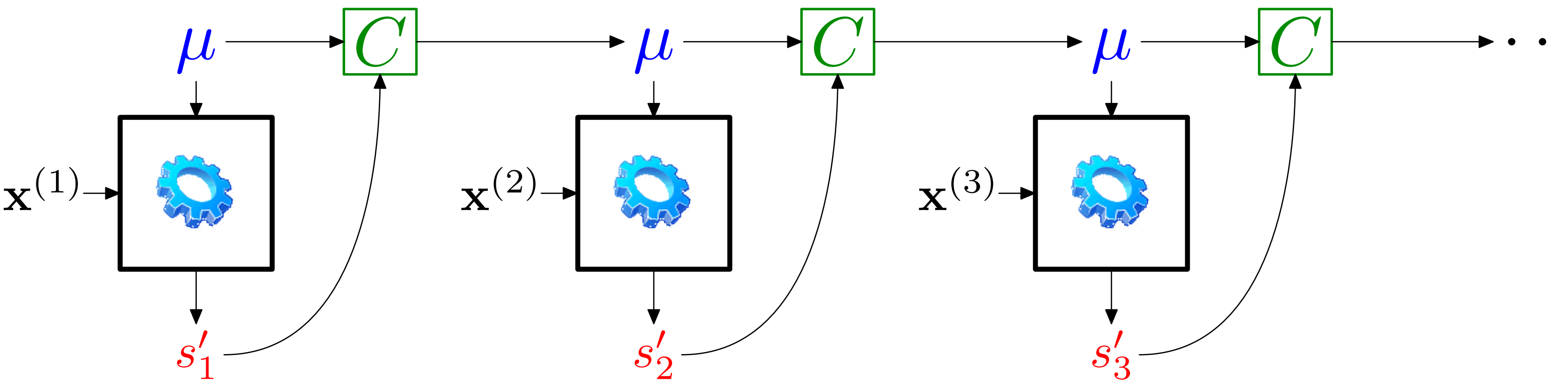


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

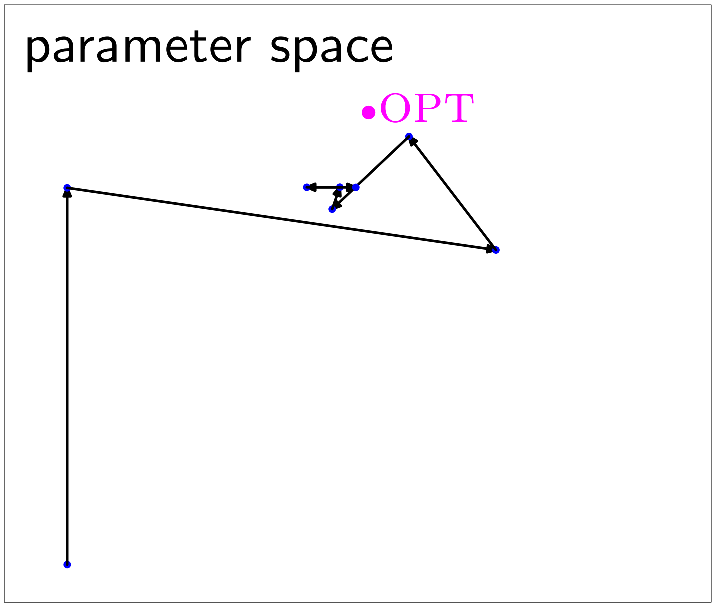


6 data points processed

Online EM [Cappé & Moulines, 2009]

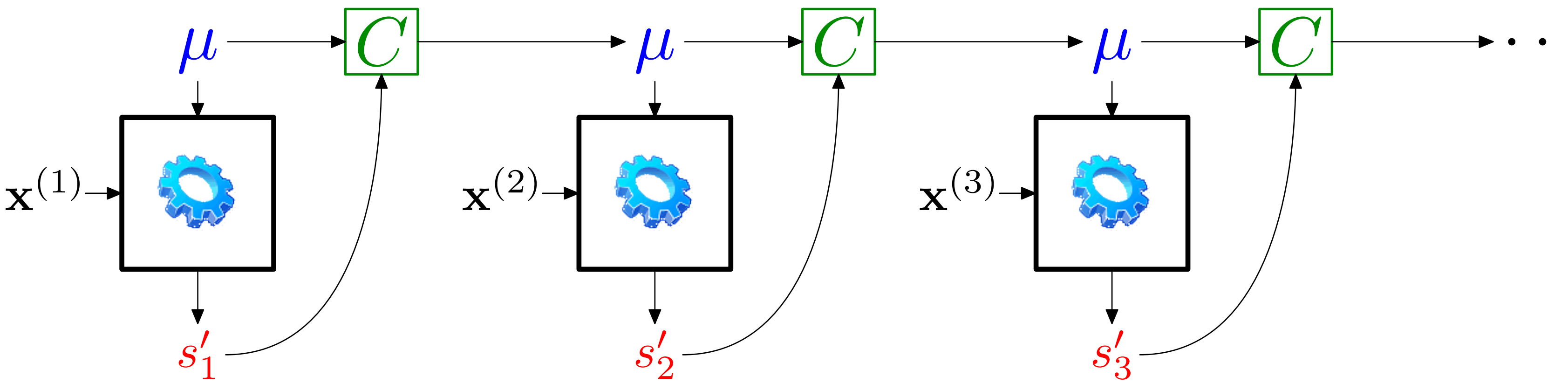


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

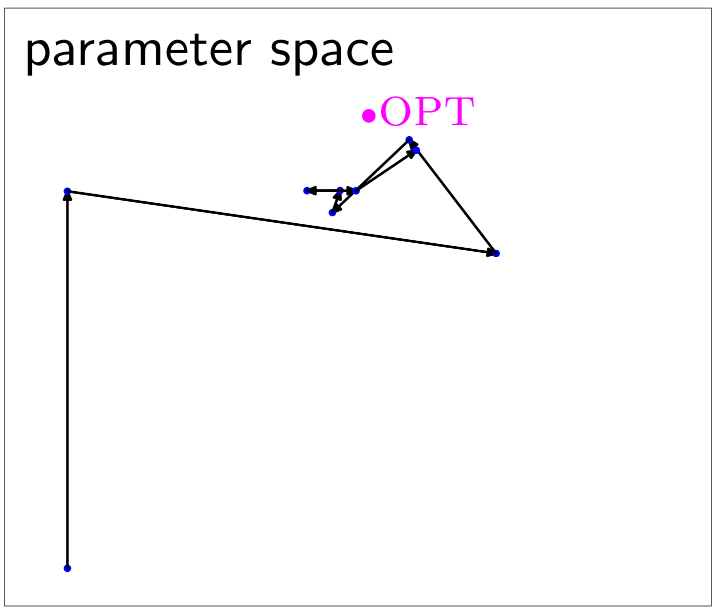


7 data points processed

Online EM [Cappé & Moulines, 2009]

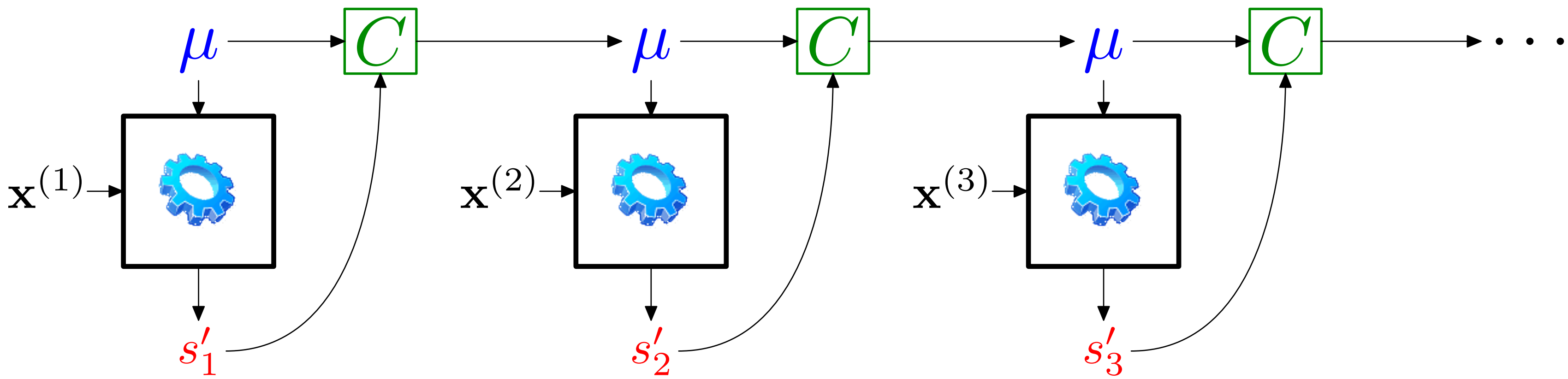


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

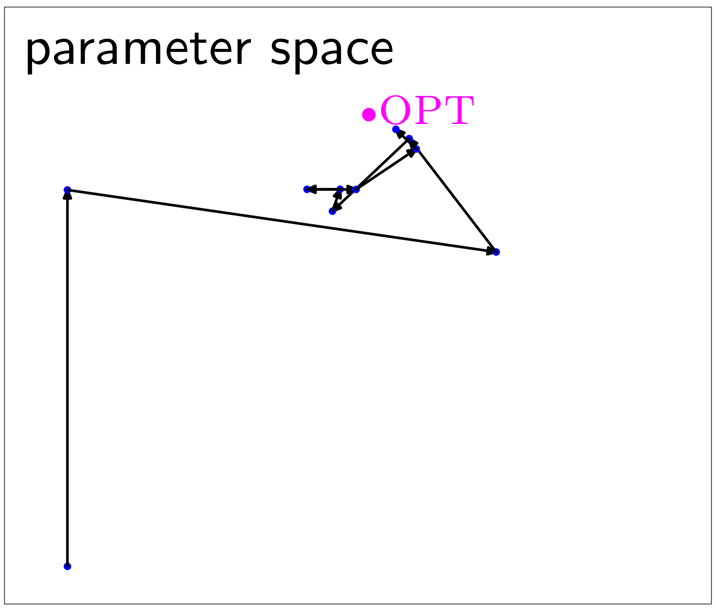


8 data points processed

Online EM [Cappé & Moulines, 2009]

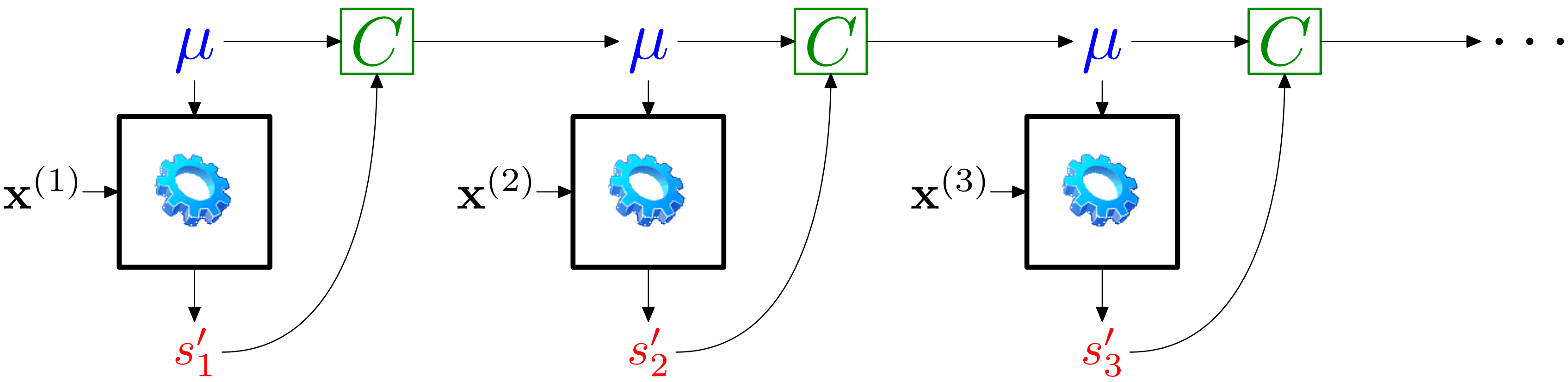


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

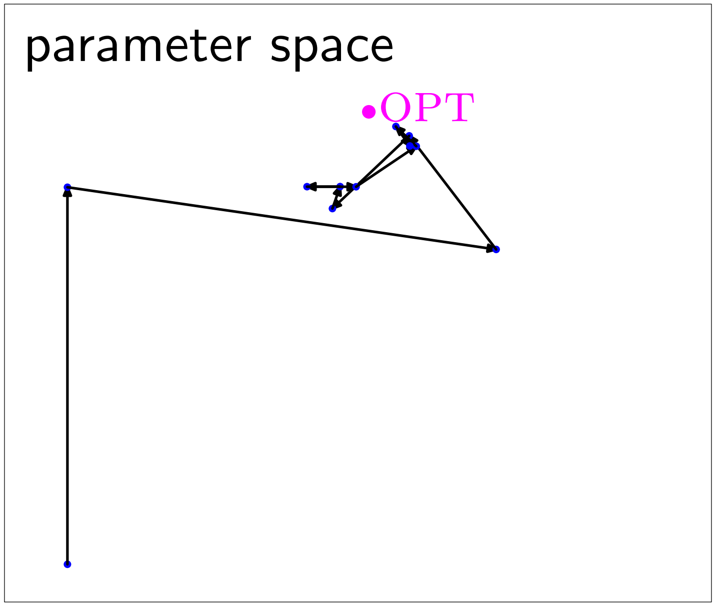


9 data points processed

Online EM [Cappé & Moulines, 2009]

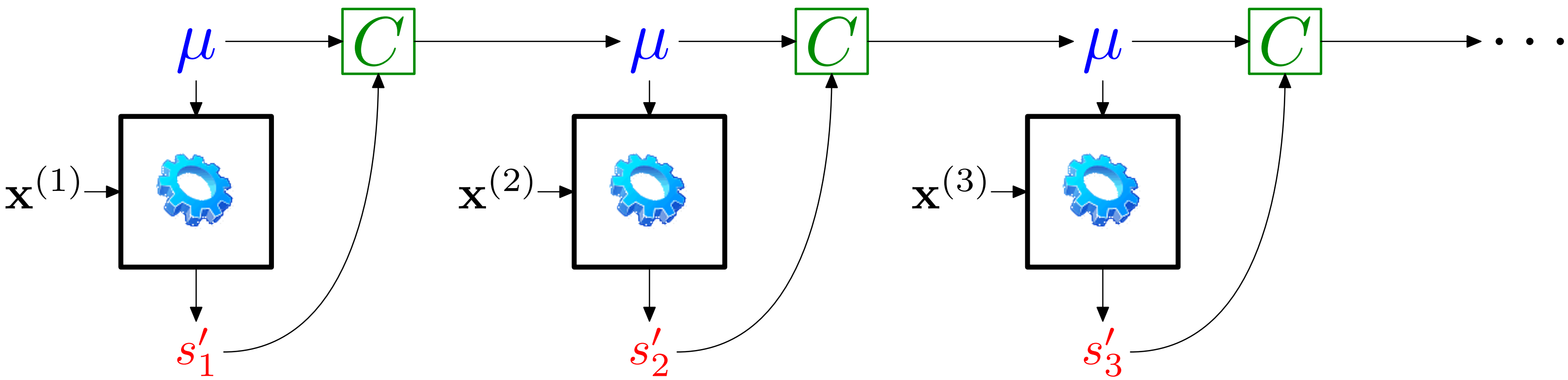


Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

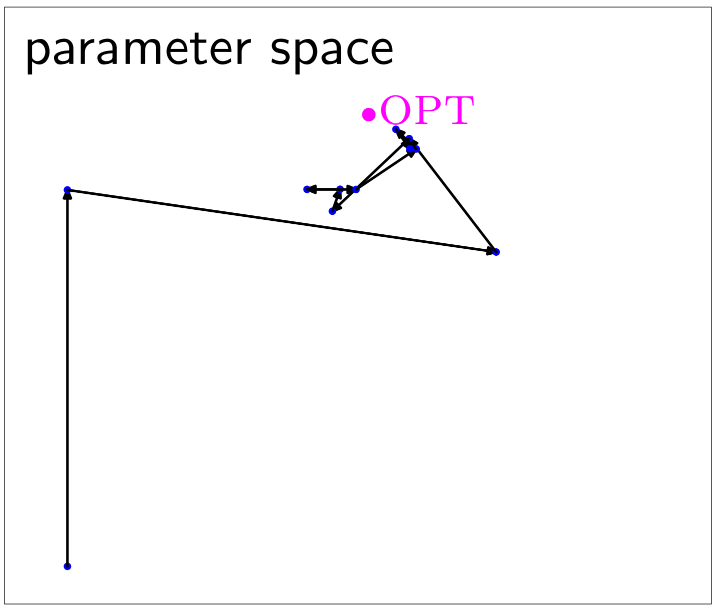


10 data points processed

Online EM [Cappé & Moulines, 2009]



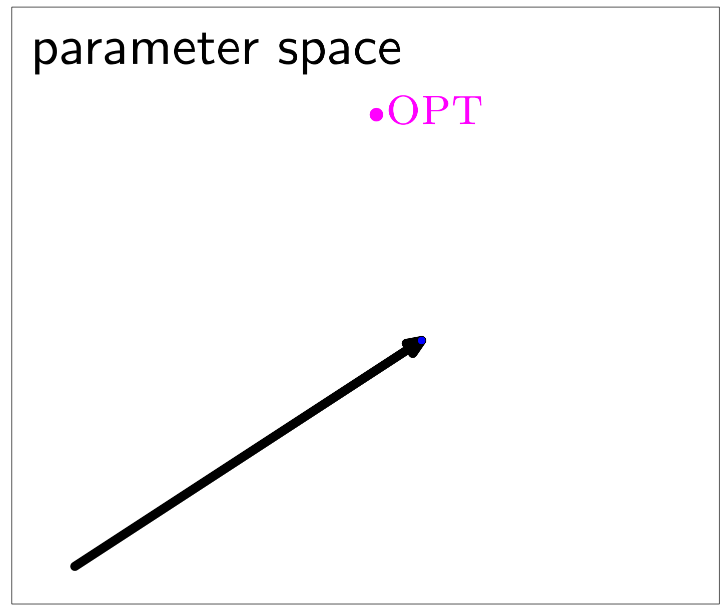
Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



10 data points processed

Online (fast, unstable)

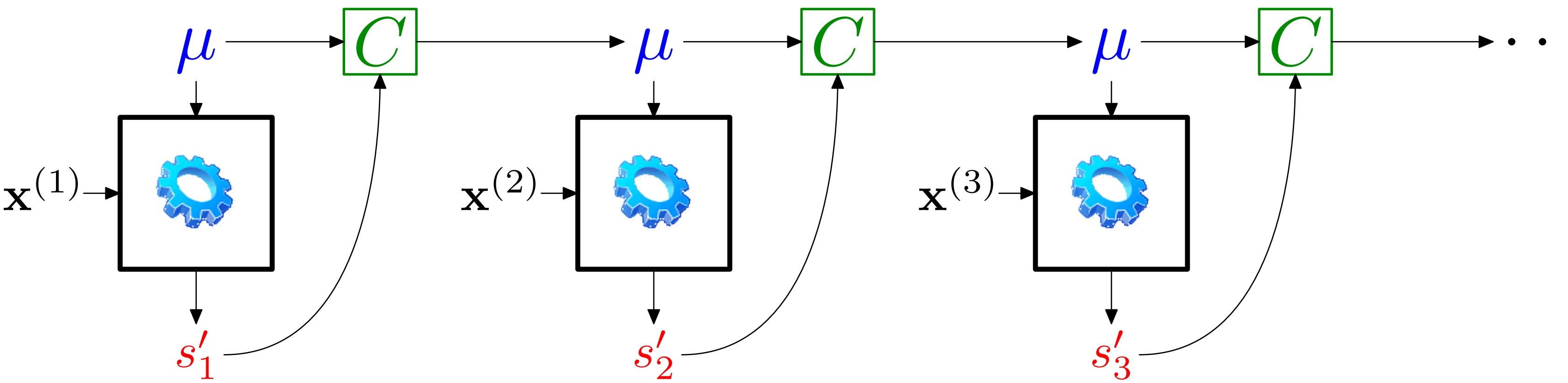
Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)

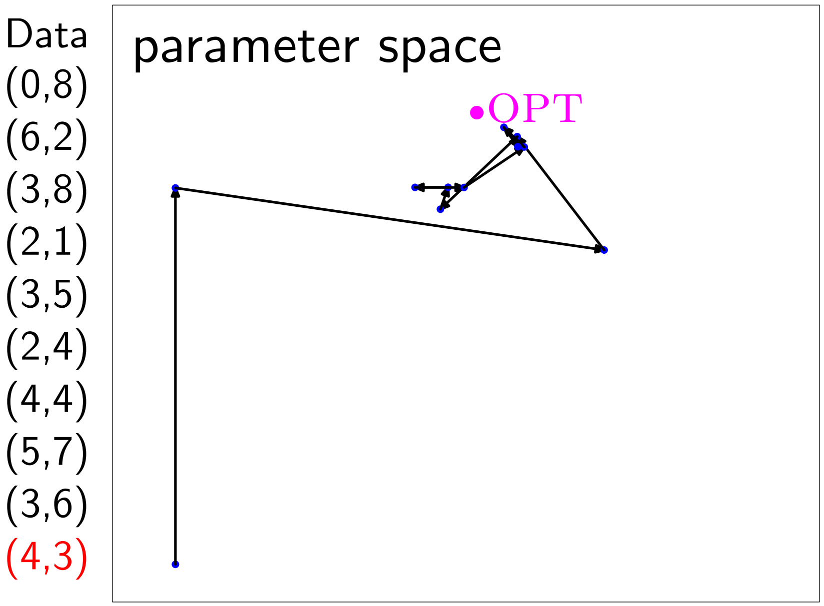


10 data points processed

Batch (slow, stable)

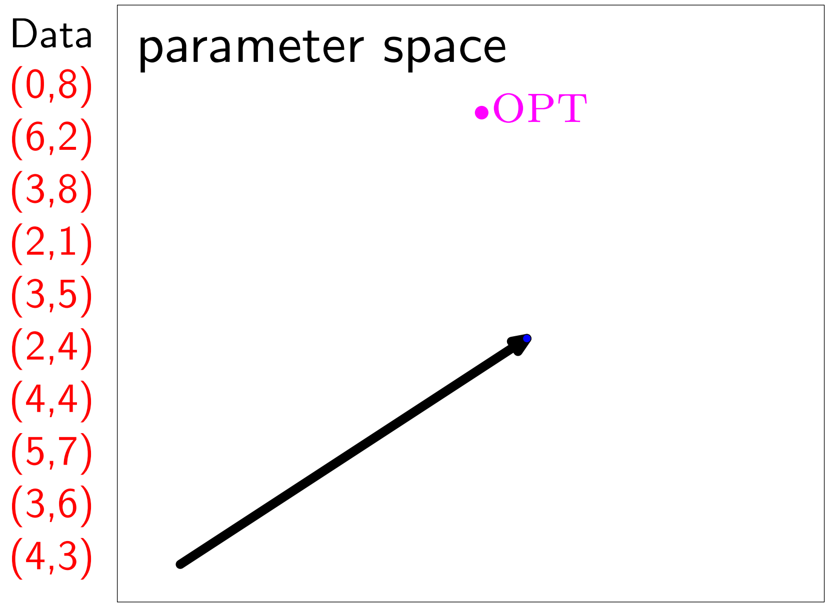
Online EM [Cappé & Moulines, 2009]





10 data points processed

Online (fast, unstable)

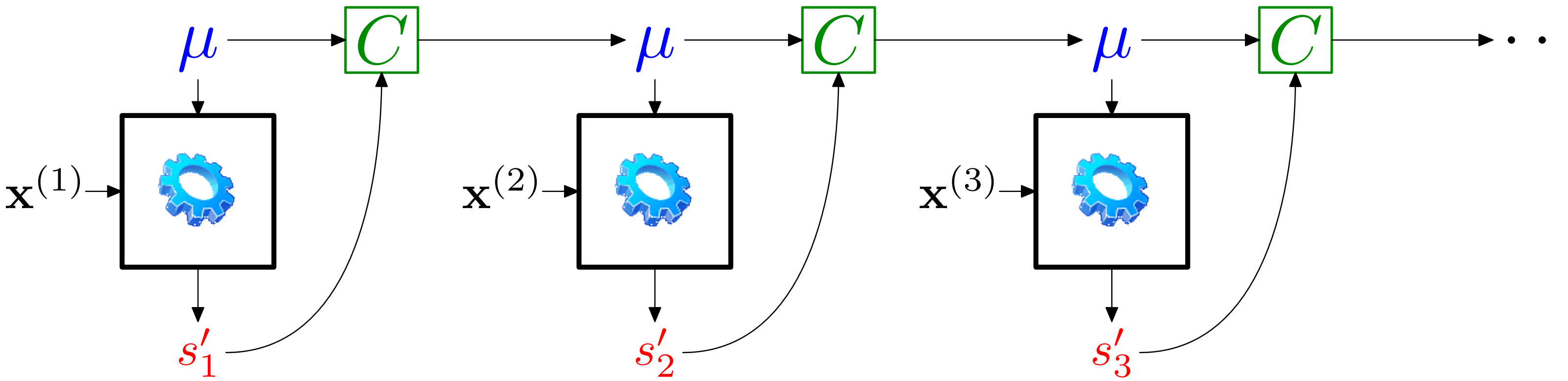


10 data points processed

Batch (slow, stable)

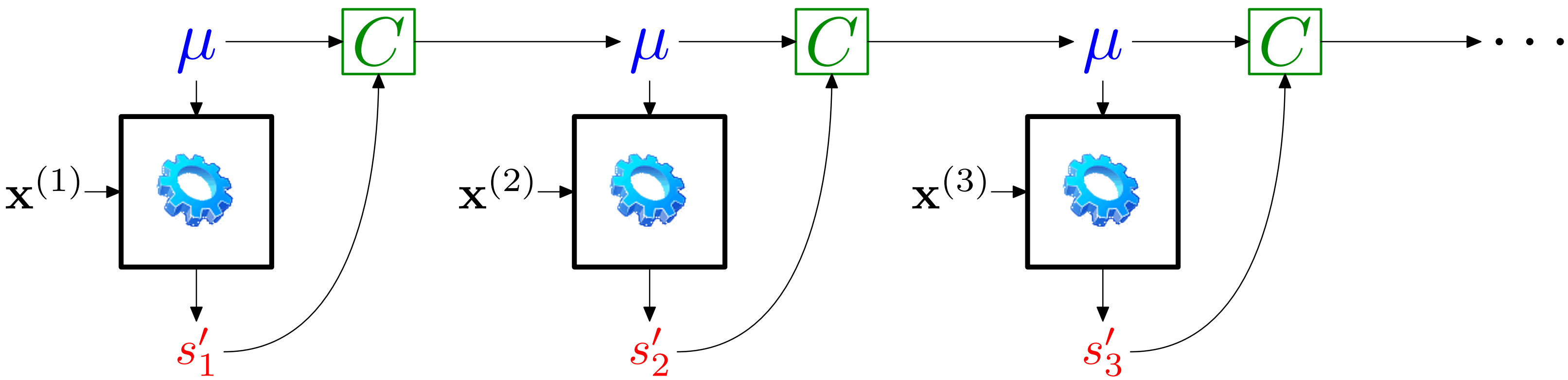
Next: stabilize online EM by modifying optimization parameters

Optimization parameter 1 of 2: stepsize



Combine old μ and new s'_i:

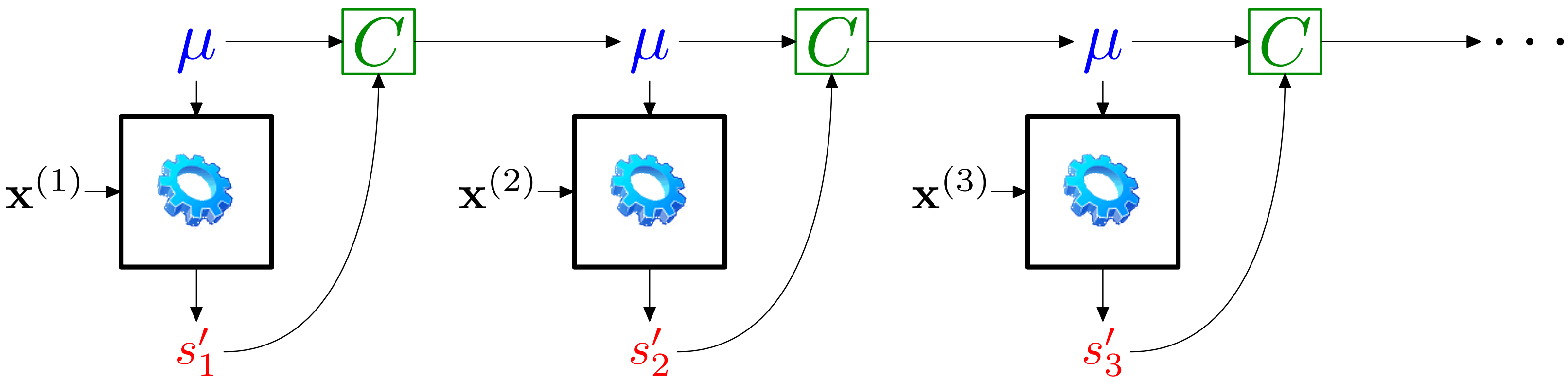
Optimization parameter 1 of 2: stepsize



Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

Optimization parameter 1 of 2: stepsize

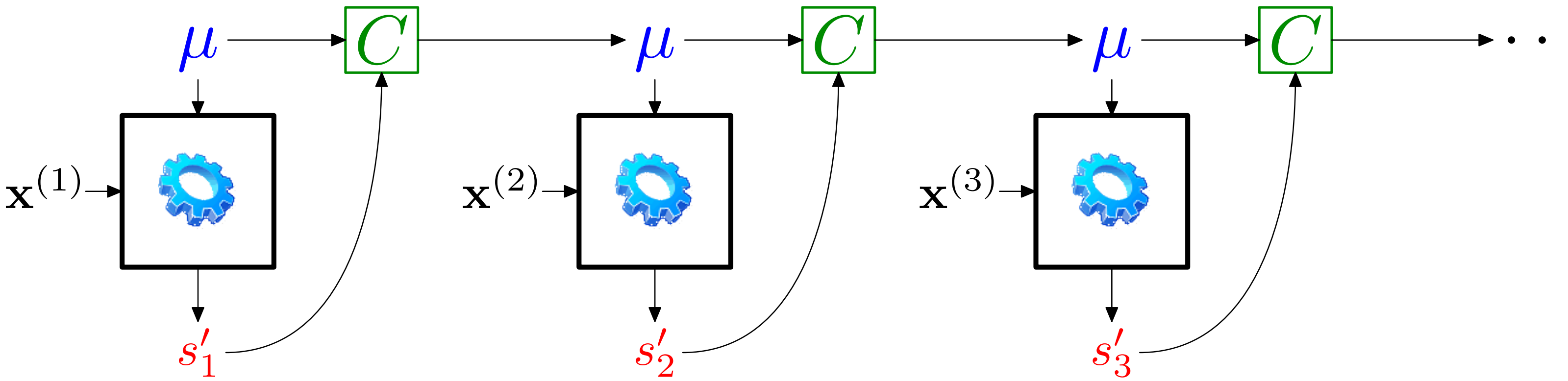


Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

Optimization parameter 1 of 2: stepsize



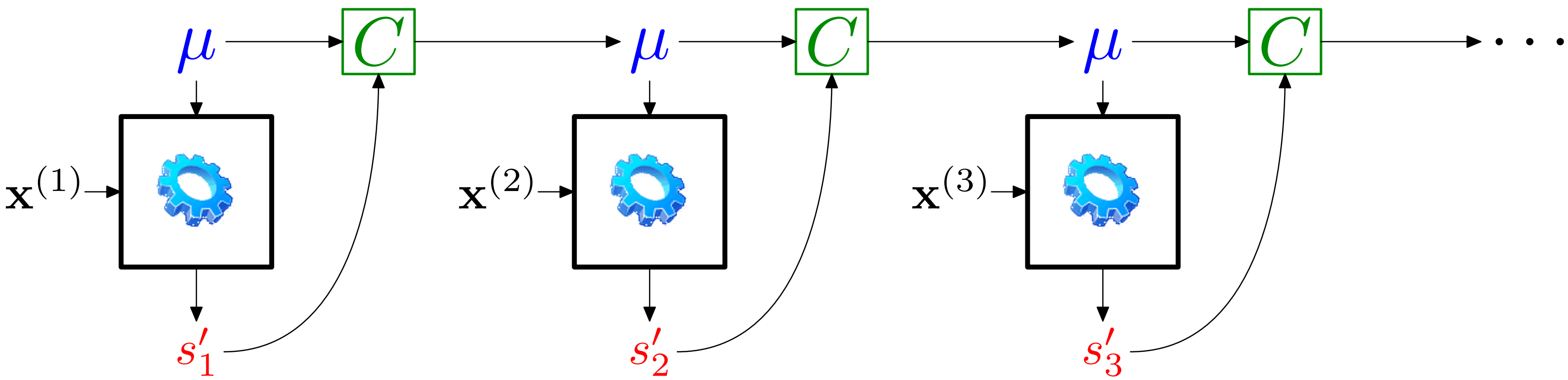
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

large updates, unstable — small updates, stable

Optimization parameter 1 of 2: stepsize



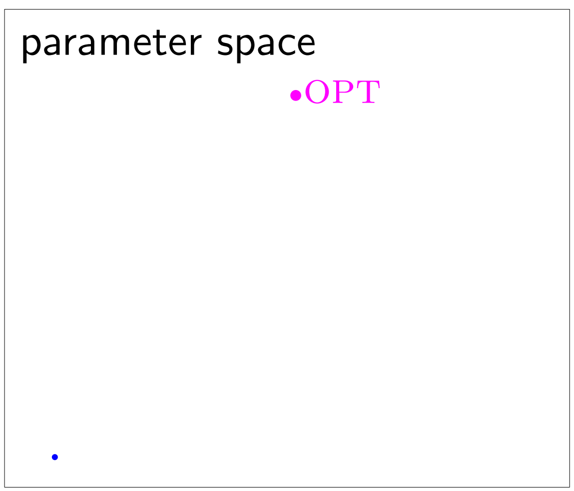
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

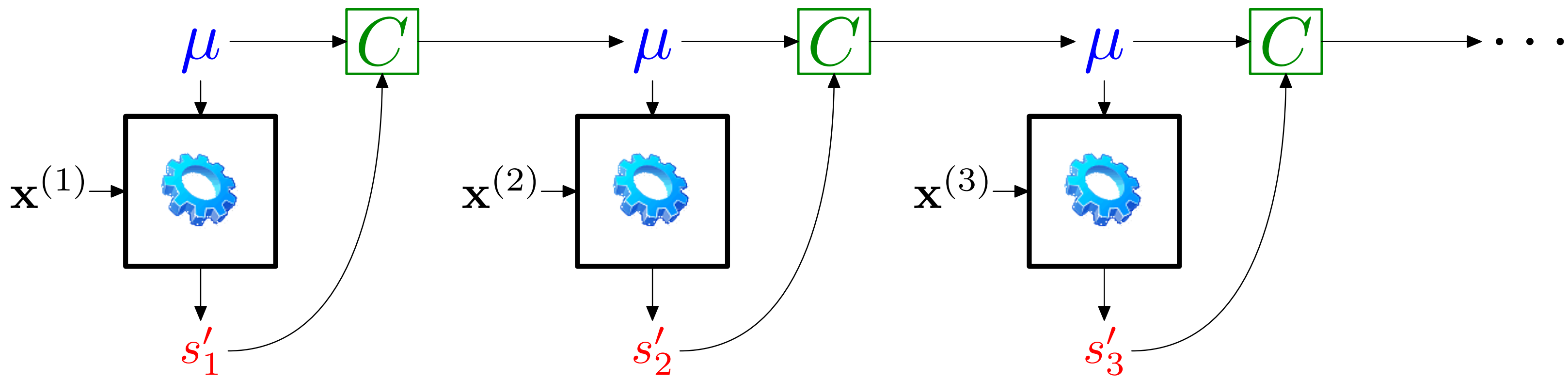
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



0 data points processed

Optimization parameter 1 of 2: stepsize



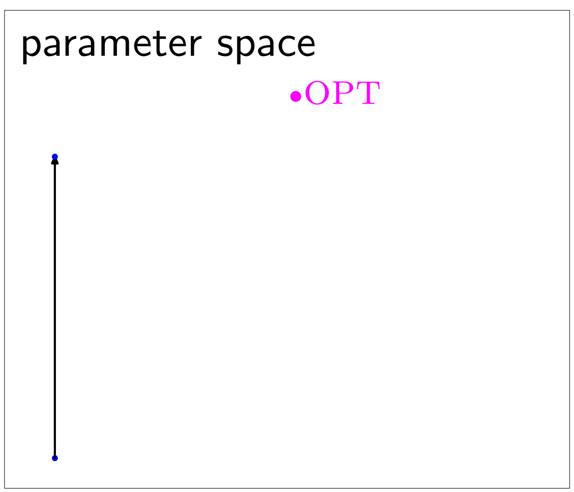
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

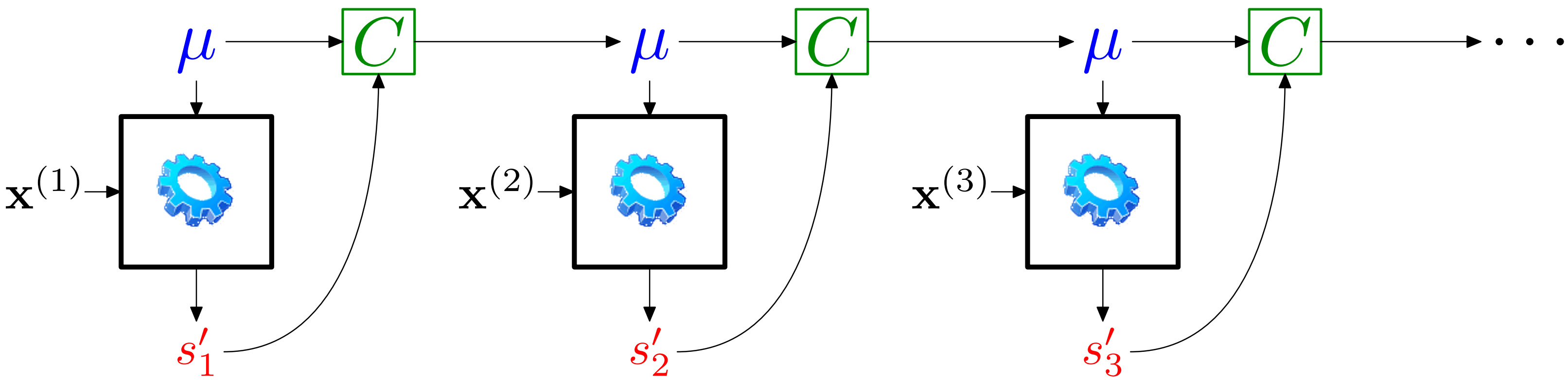
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



1 data points processed

Optimization parameter 1 of 2: stepsize



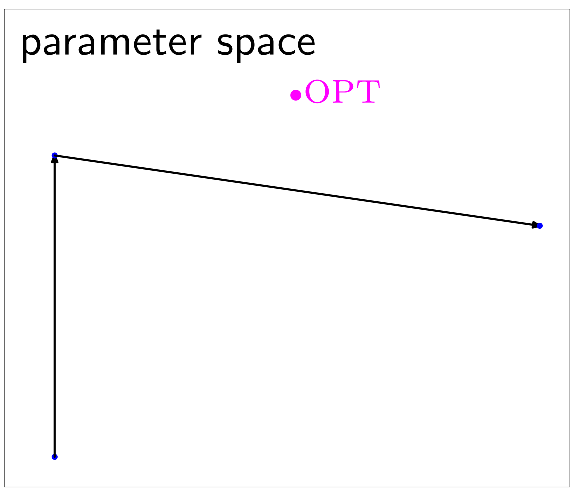
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

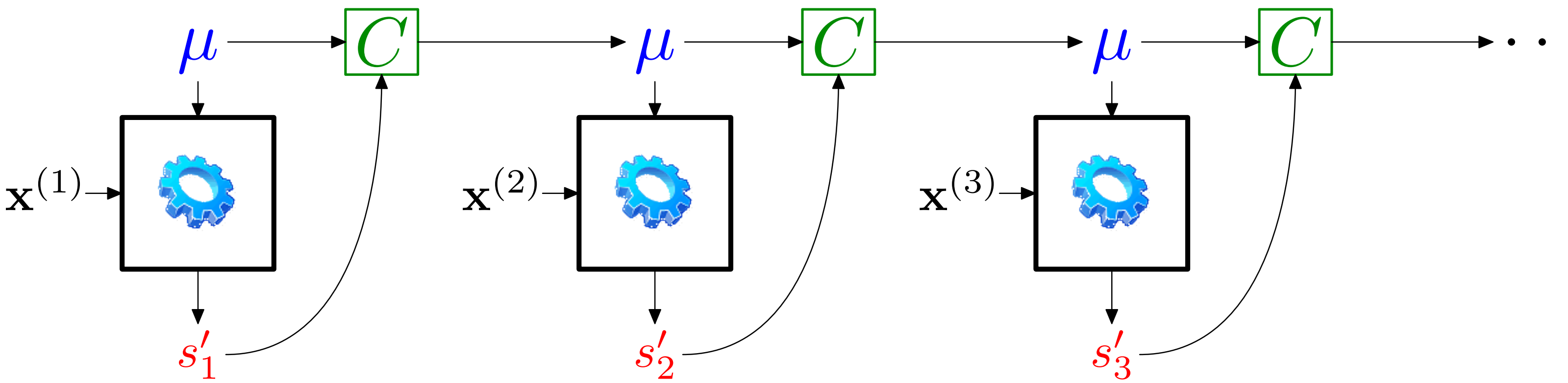
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



2 data points processed

Optimization parameter 1 of 2: stepsize



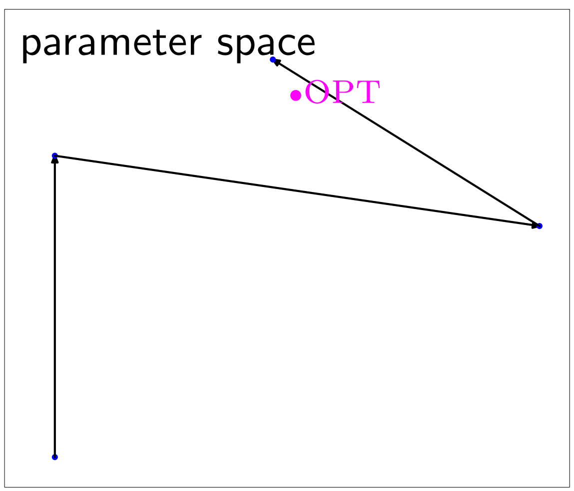
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

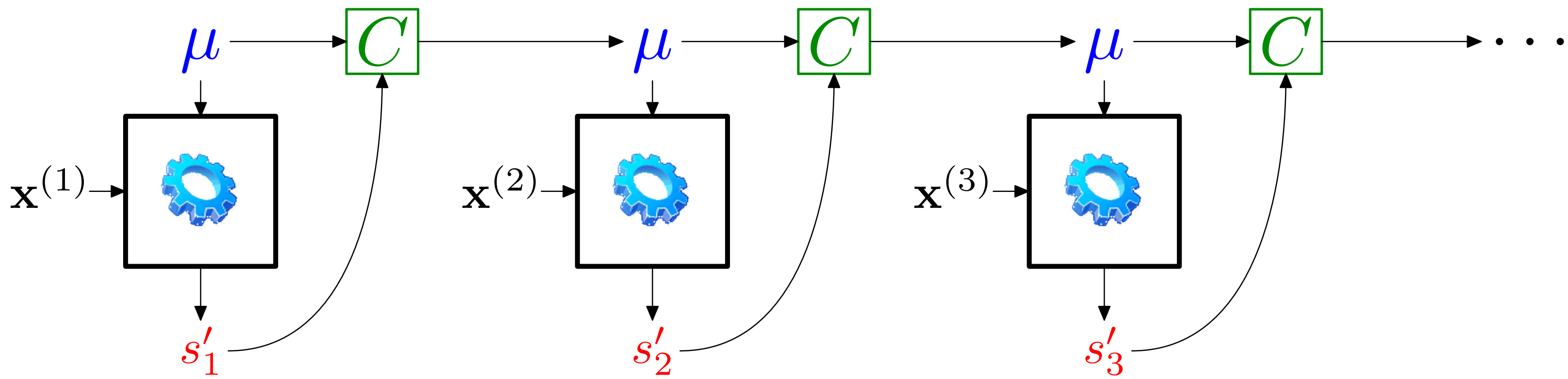
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



3 data points processed

Optimization parameter 1 of 2: stepsize



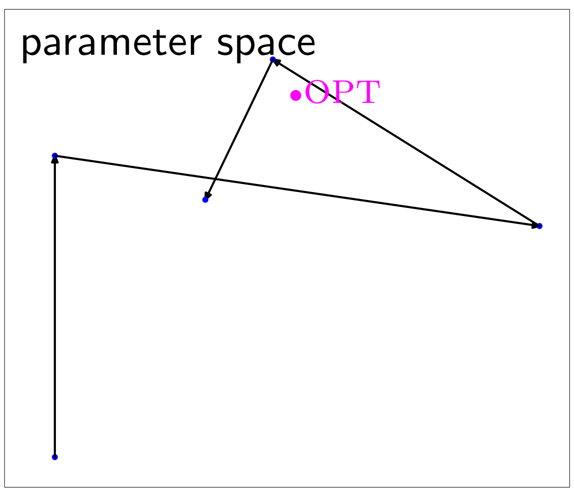
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

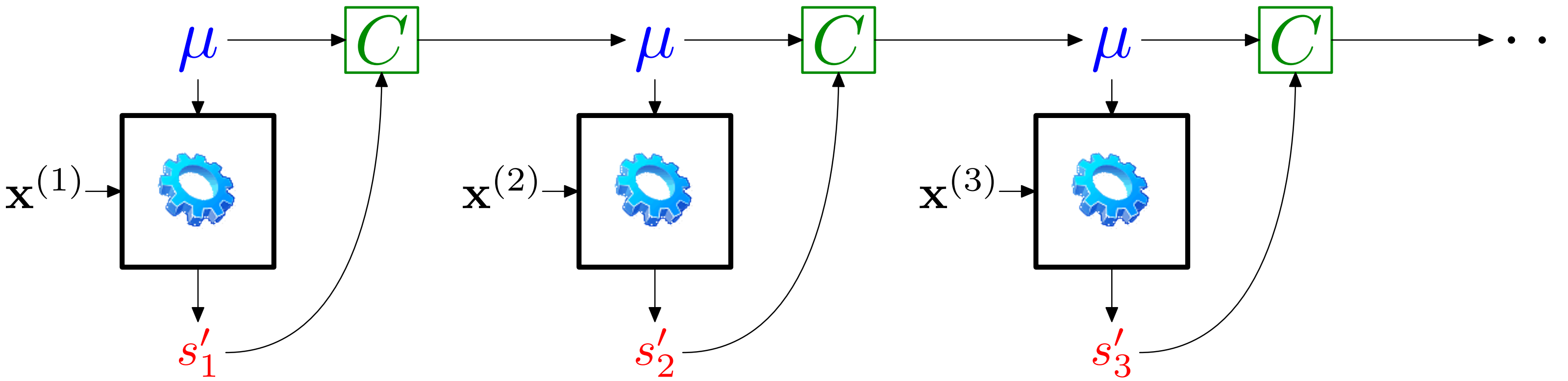
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



4 data points processed

Optimization parameter 1 of 2: stepsize



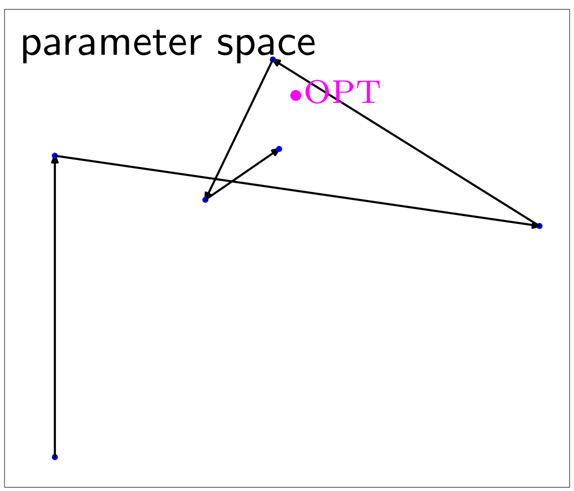
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

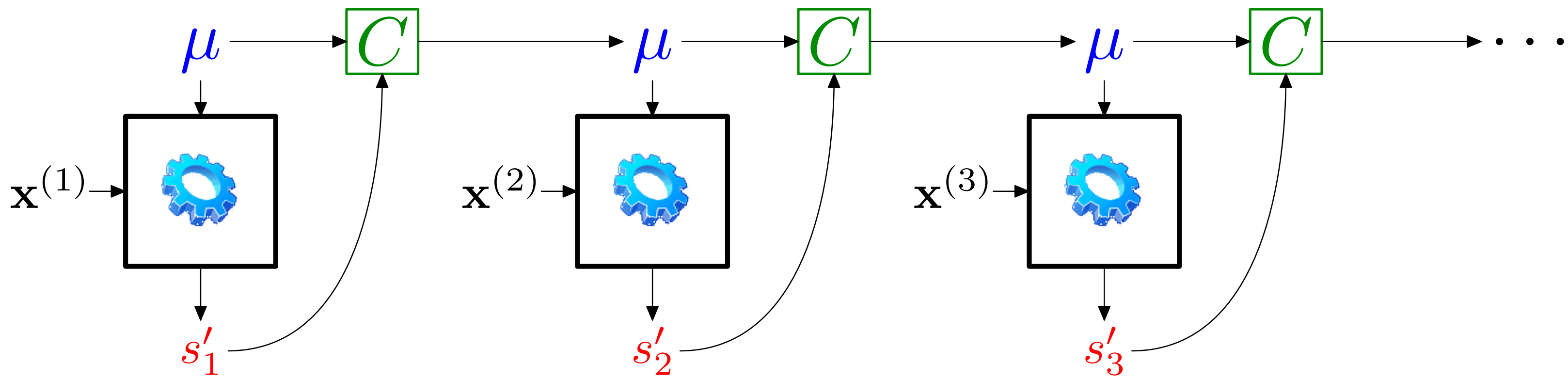
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



5 data points processed

Optimization parameter 1 of 2: stepsize



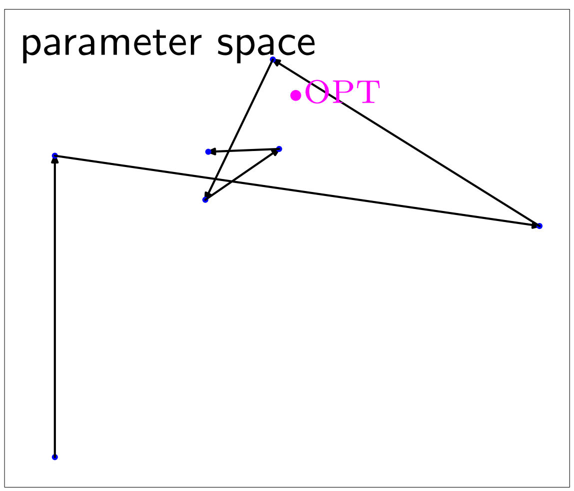
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

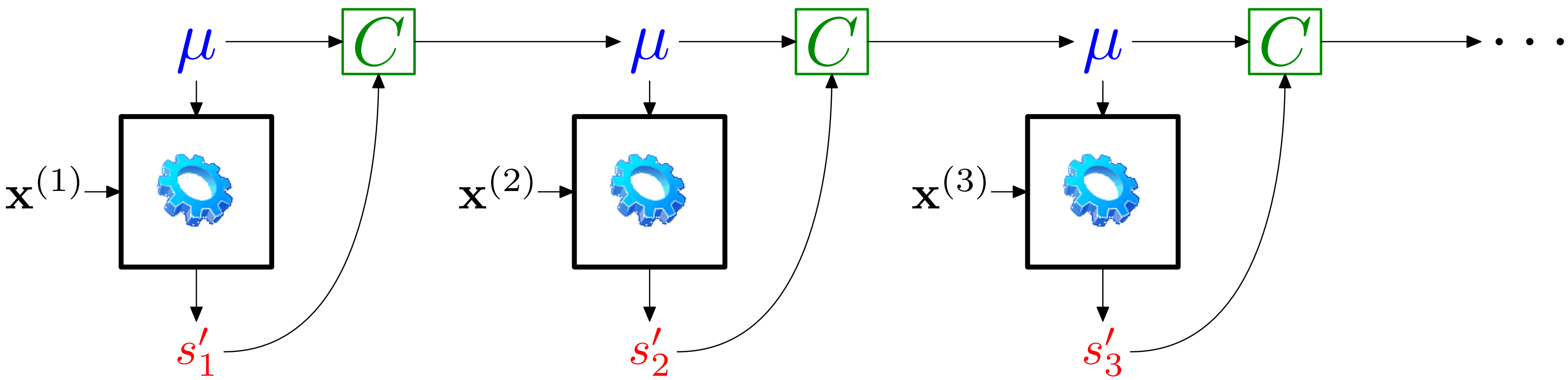
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



6 data points processed

Optimization parameter 1 of 2: stepsize



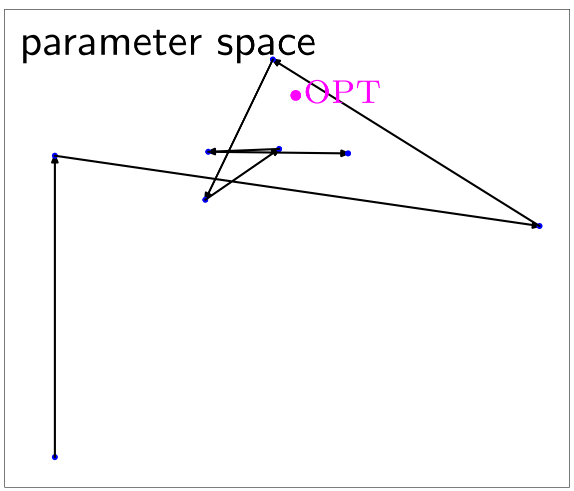
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

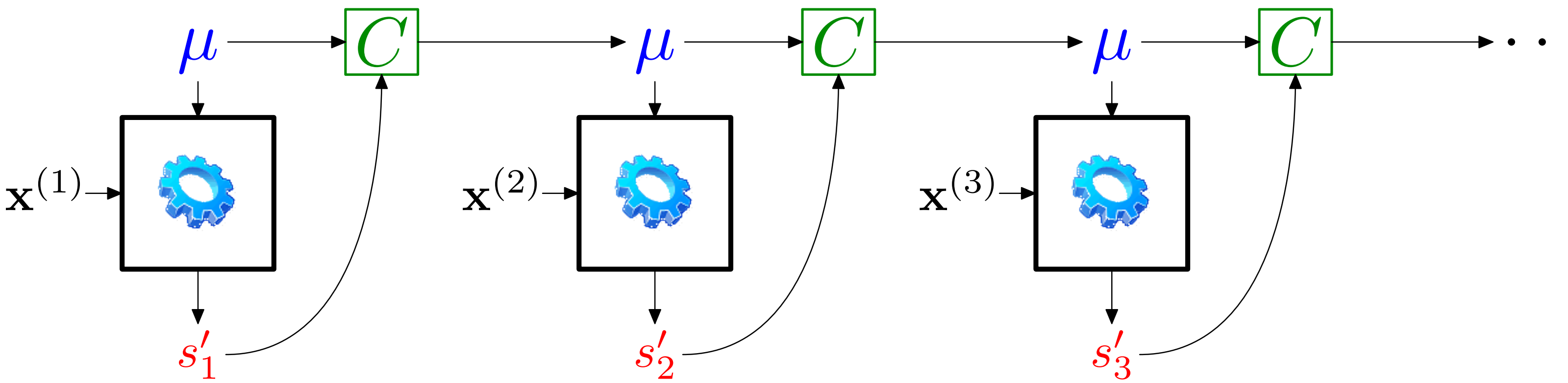
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



7 data points processed

Optimization parameter 1 of 2: stepsize



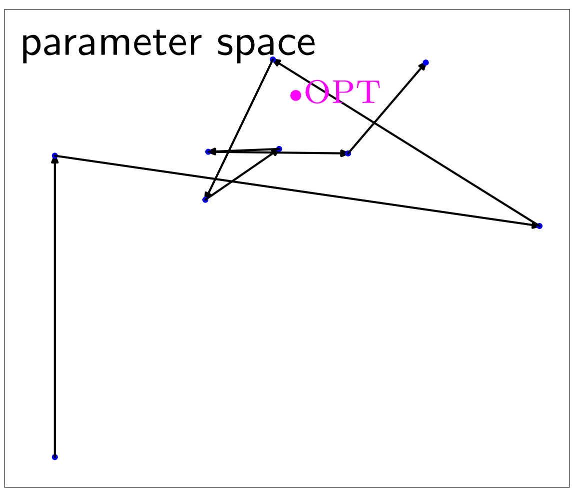
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

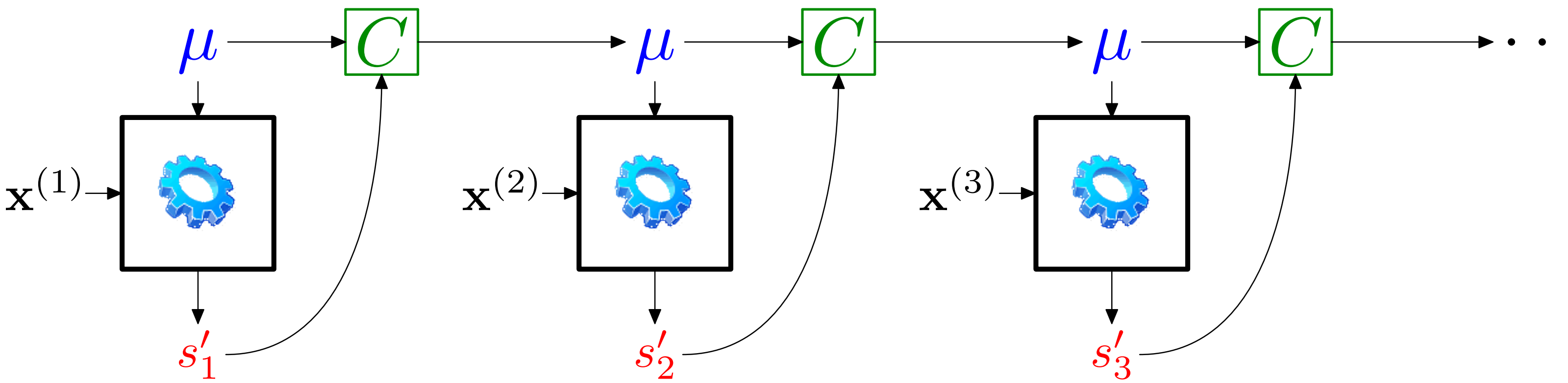
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



8 data points processed

Optimization parameter 1 of 2: stepsize



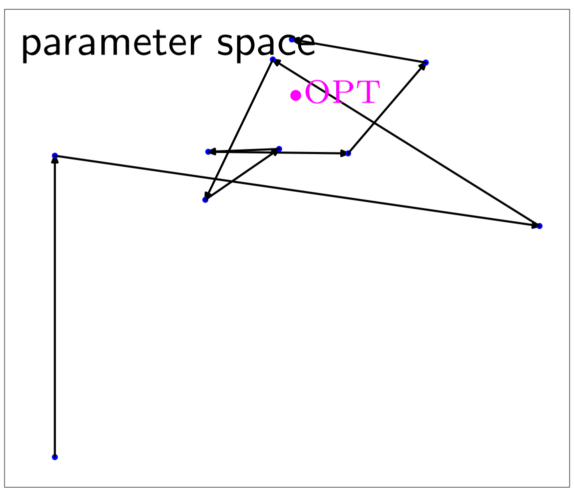
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

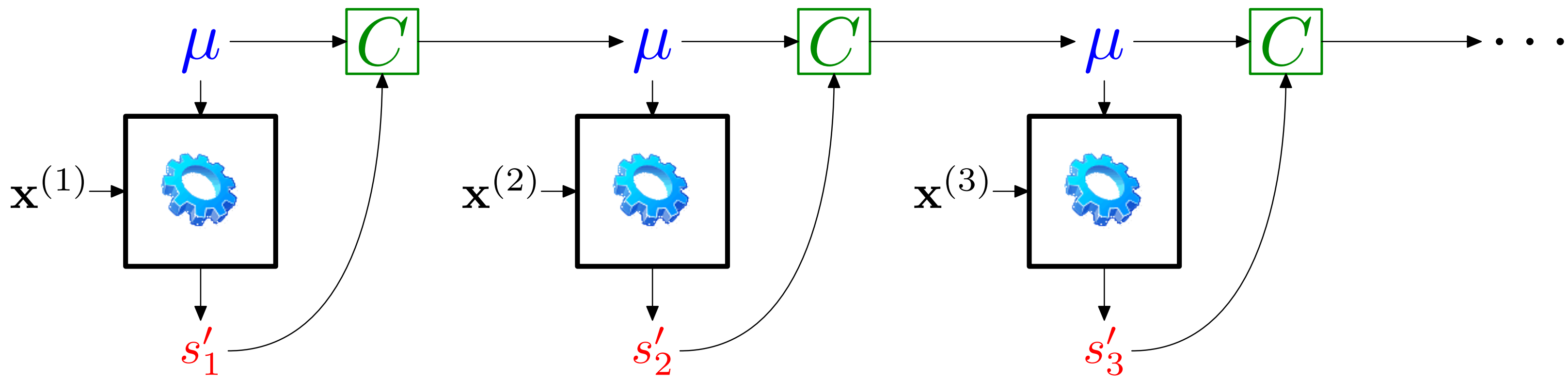
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



9 data points processed

Optimization parameter 1 of 2: stepsize



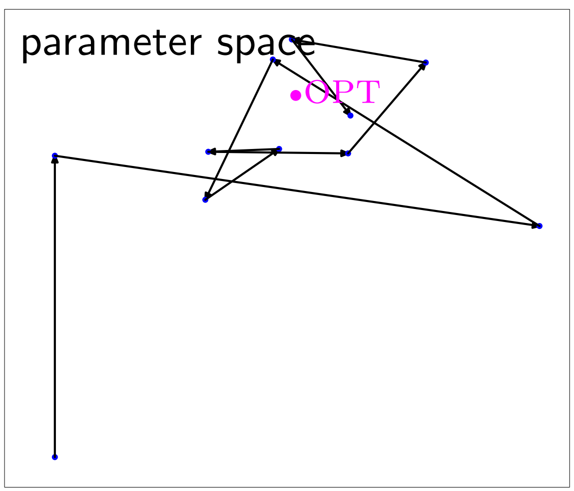
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

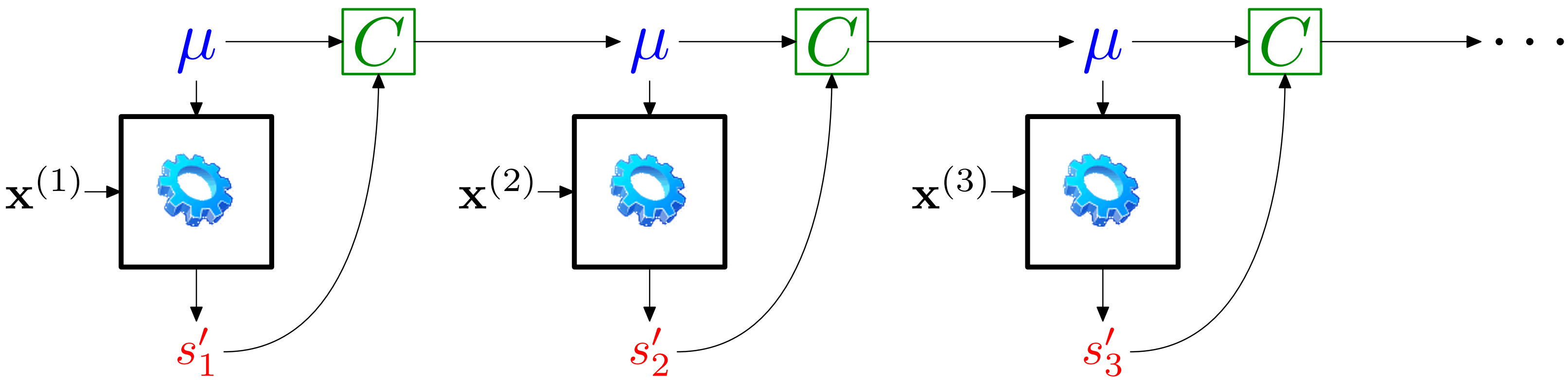
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



10 data points processed

Optimization parameter 1 of 2: stepsize

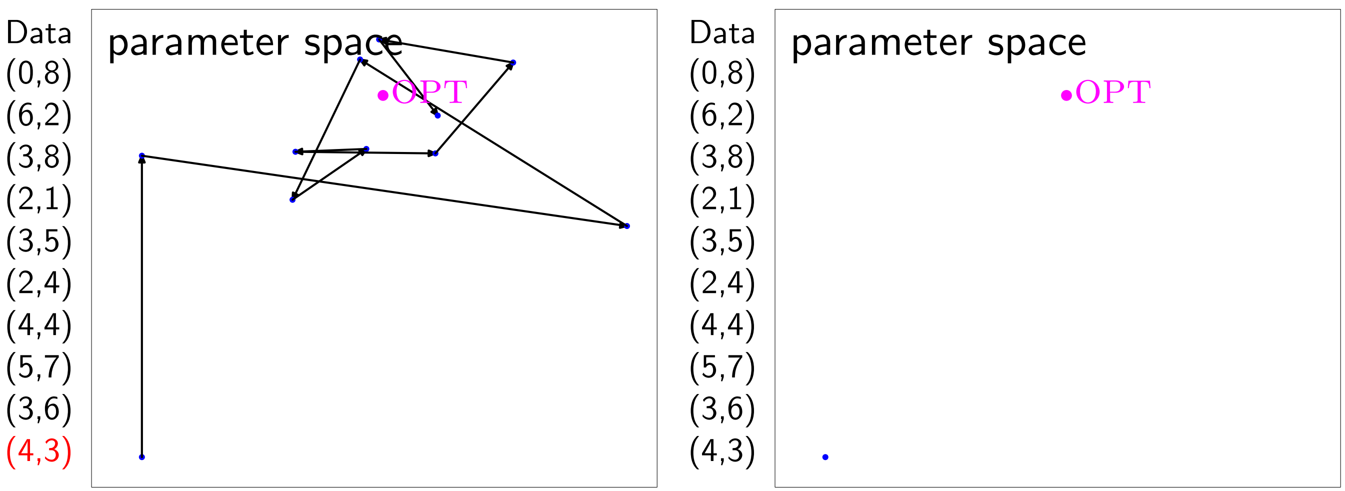


Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

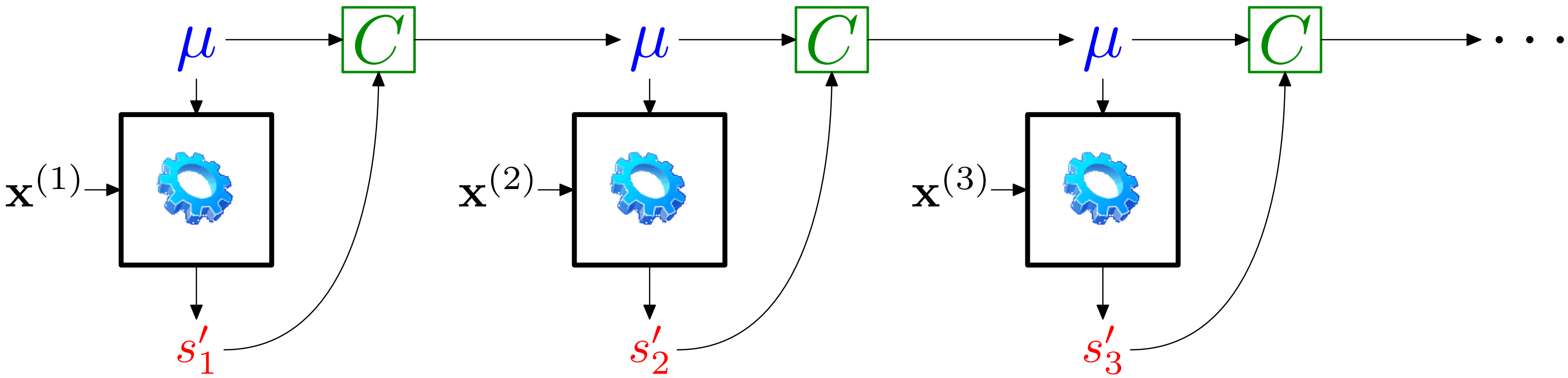
$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

large updates, unstable — small updates, stable



10 data points processed — 0 data points processed

Optimization parameter 1 of 2: stepsize

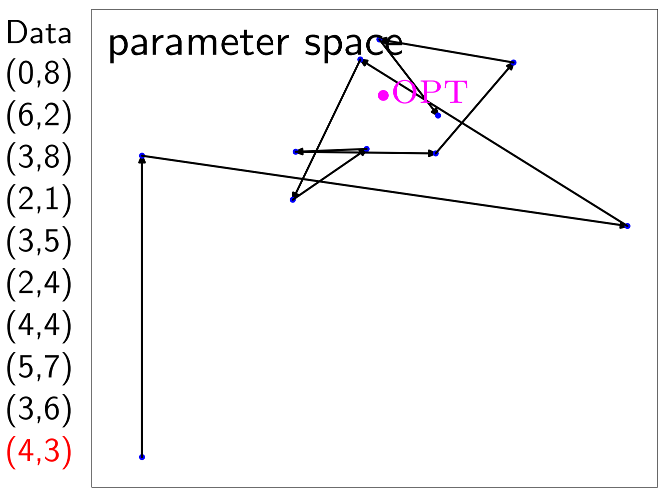


Combine old μ and new s'_i:

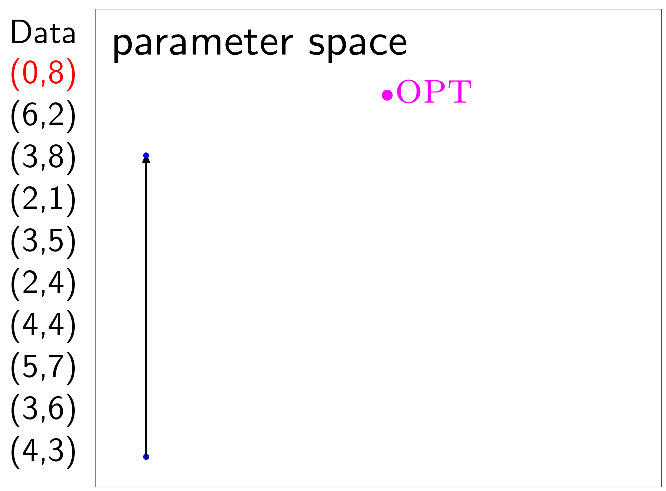
$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

large updates, unstable | small updates, stable

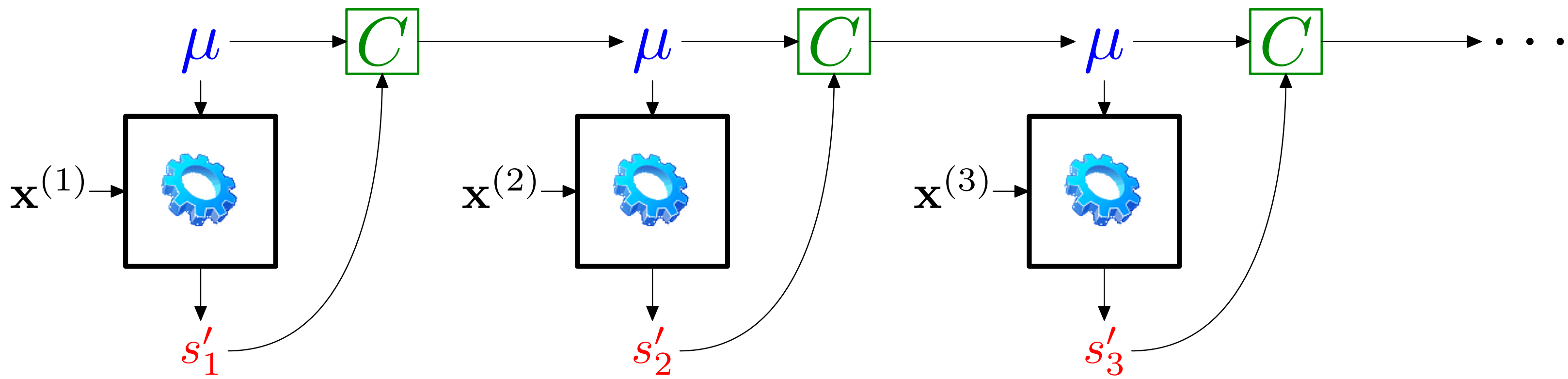


10 data points processed



1 data points processed

Optimization parameter 1 of 2: stepsize



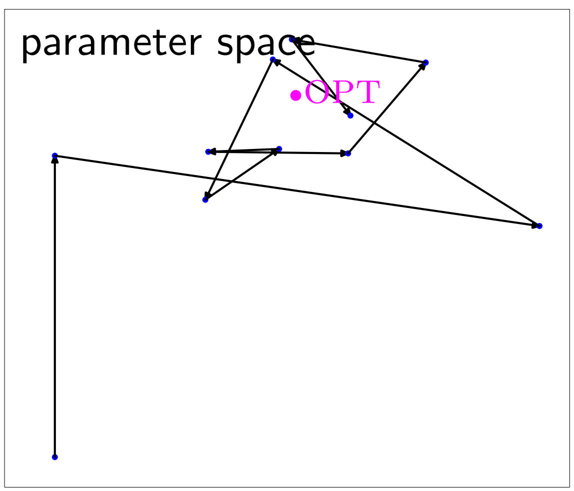
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

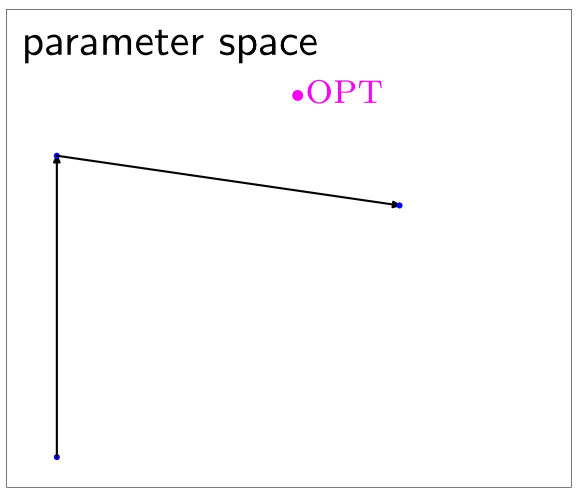
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



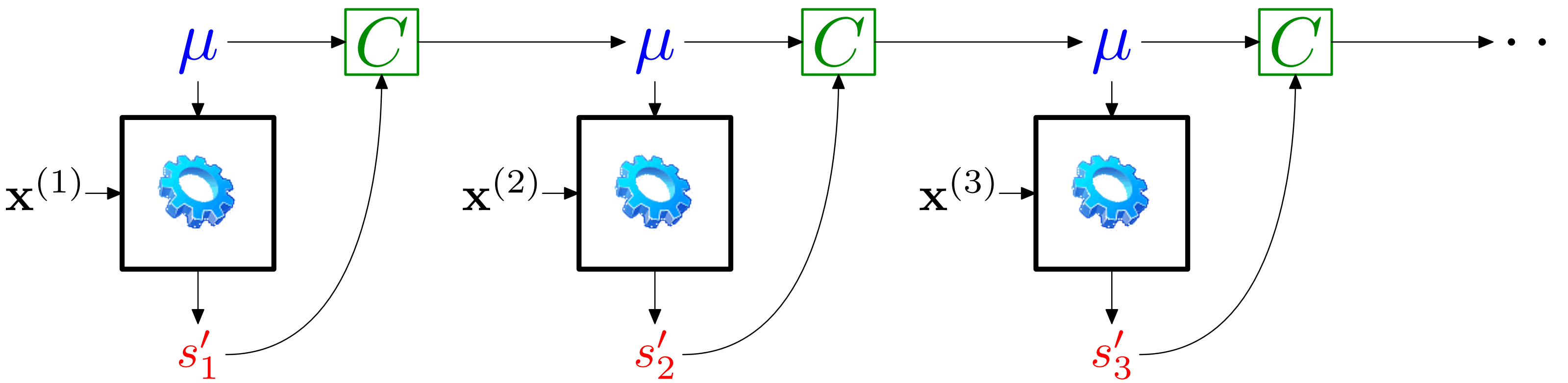
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



2 data points processed

Optimization parameter 1 of 2: stepsize



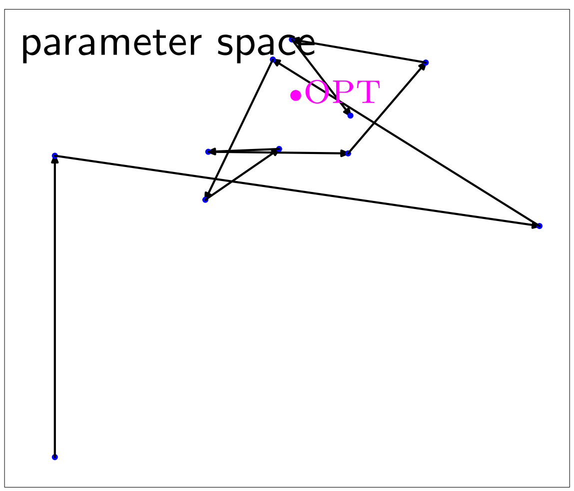
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

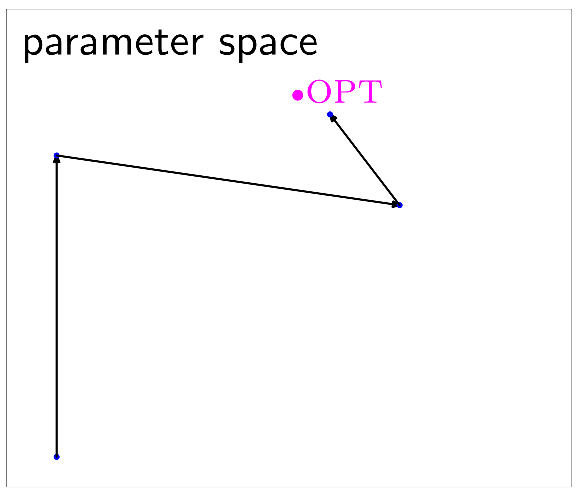
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



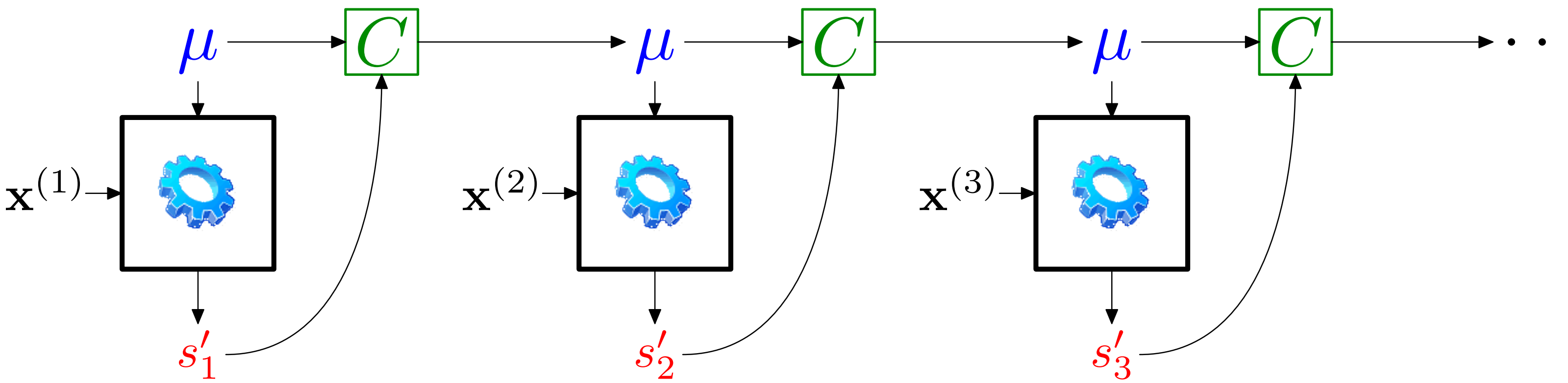
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



3 data points processed

Optimization parameter 1 of 2: stepsize

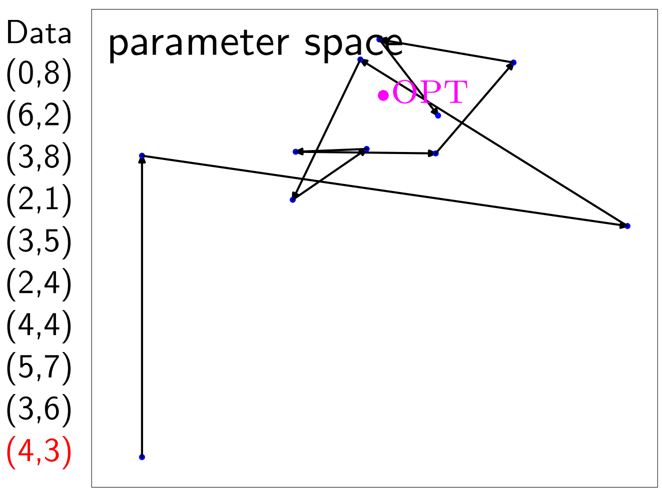


Combine old μ and new s'_i:

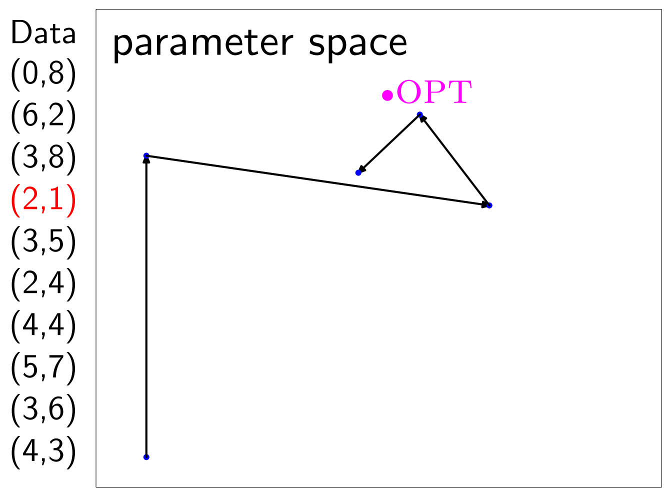
$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

large updates, unstable — small updates, stable

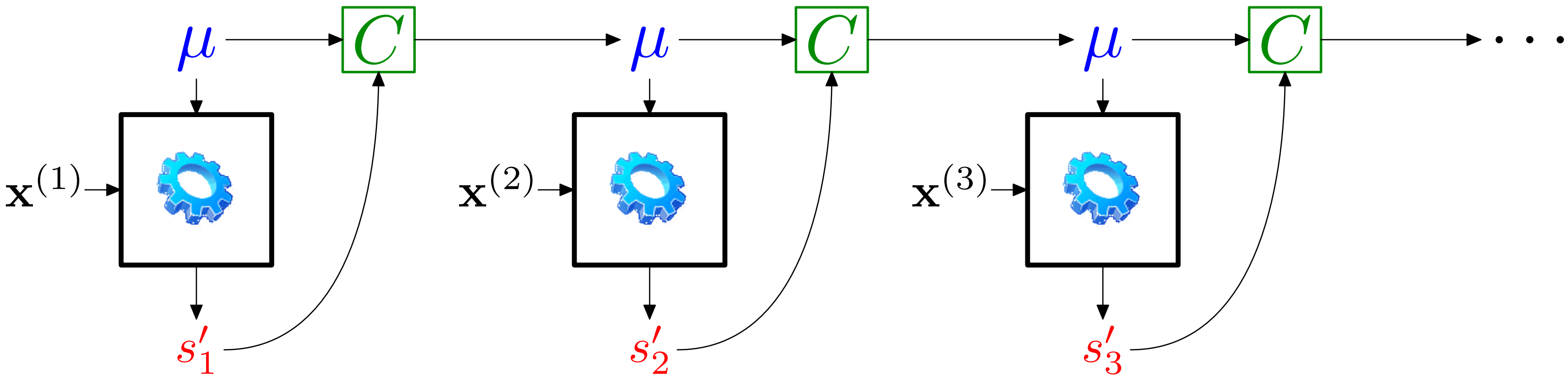


10 data points processed



4 data points processed

Optimization parameter 1 of 2: stepsize



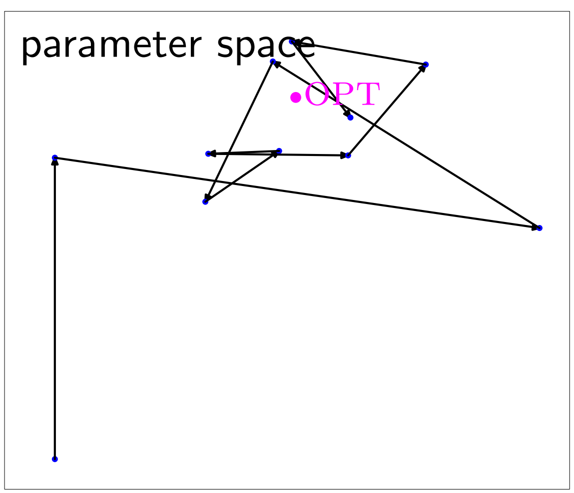
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

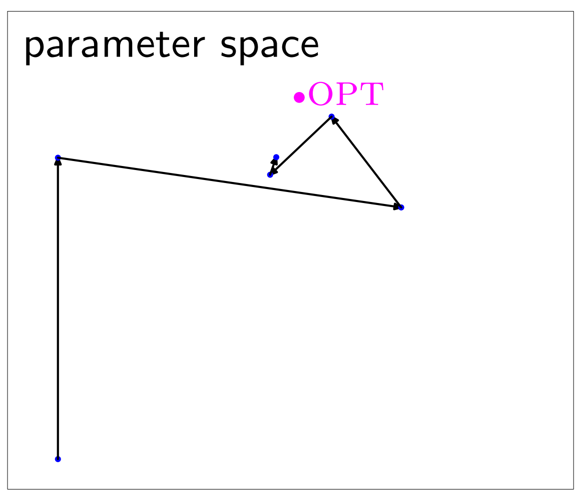
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



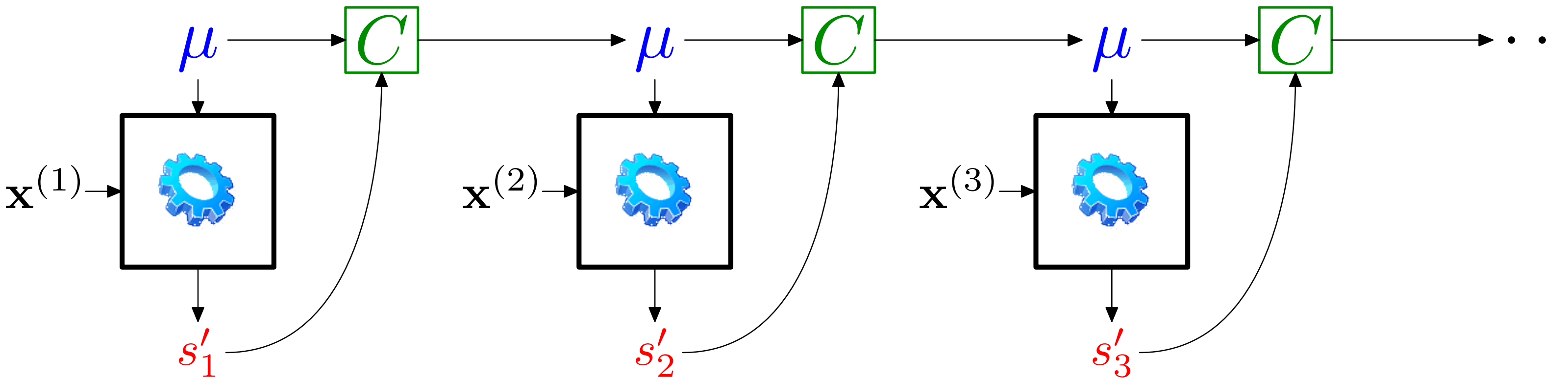
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



5 data points processed

Optimization parameter 1 of 2: stepsize



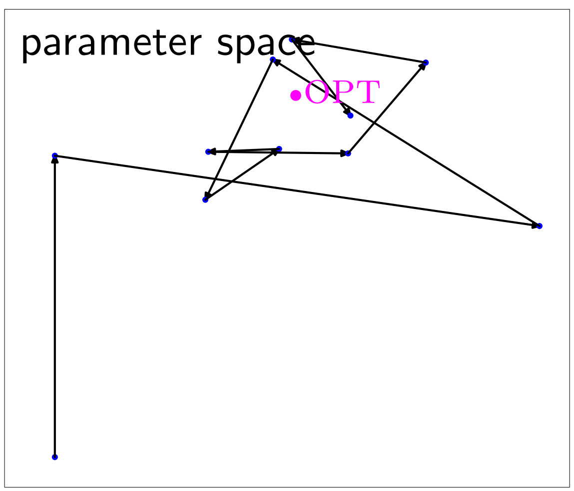
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

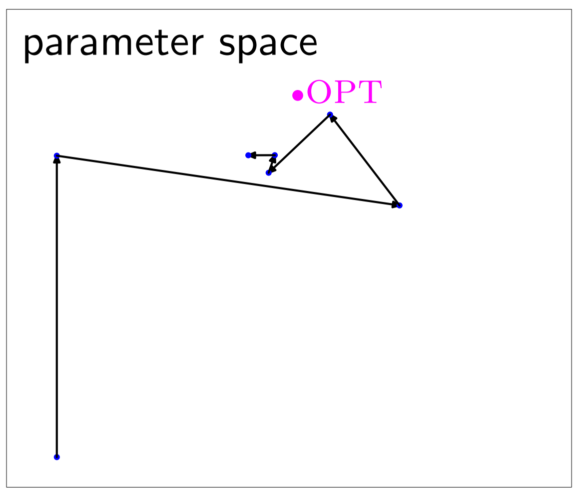
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



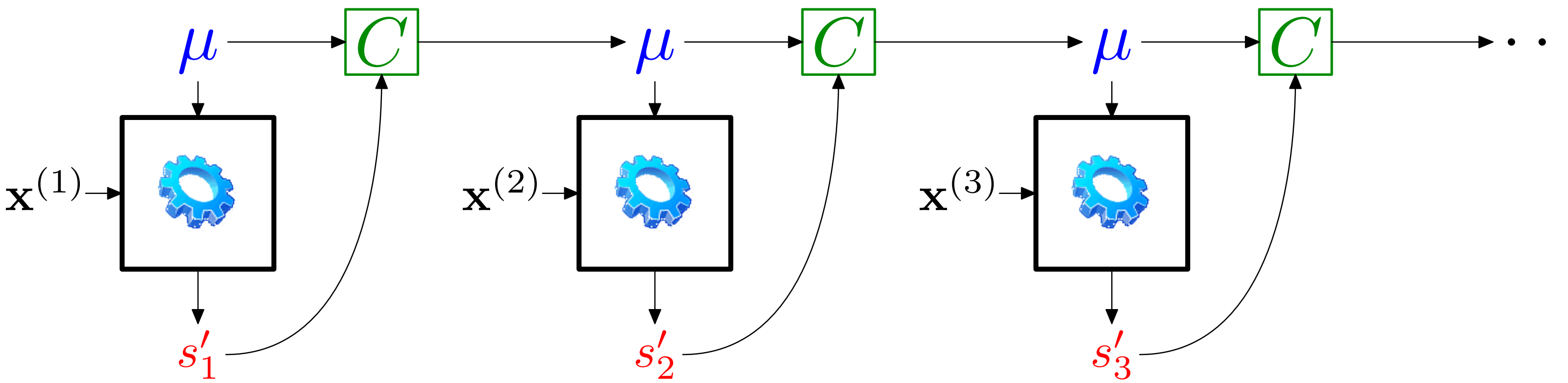
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



6 data points processed

Optimization parameter 1 of 2: stepsize

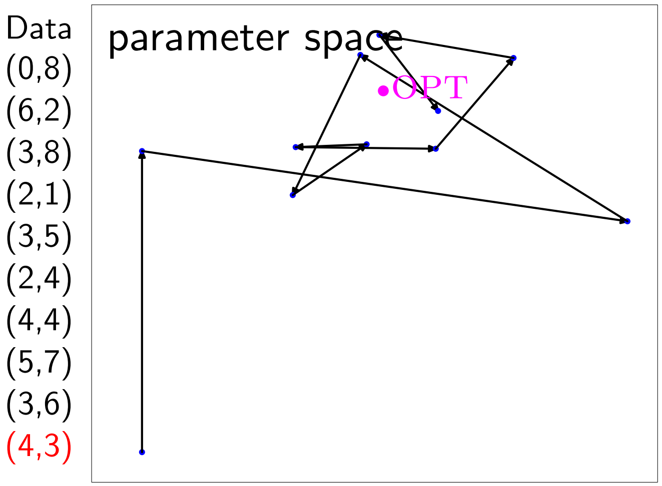


Combine old μ and new s'_i:

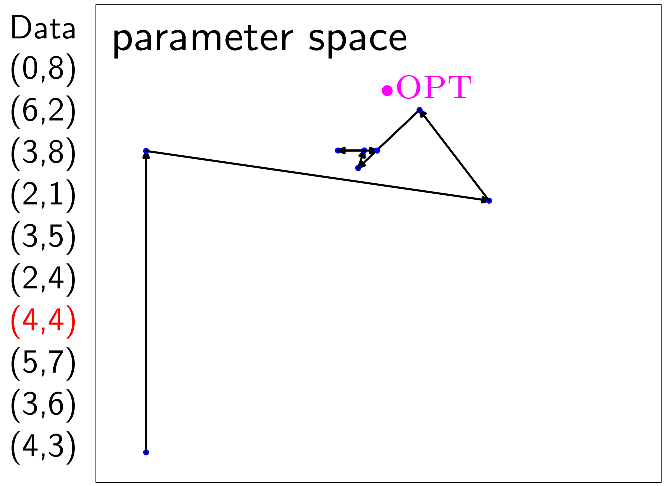
$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

large updates, unstable — small updates, stable

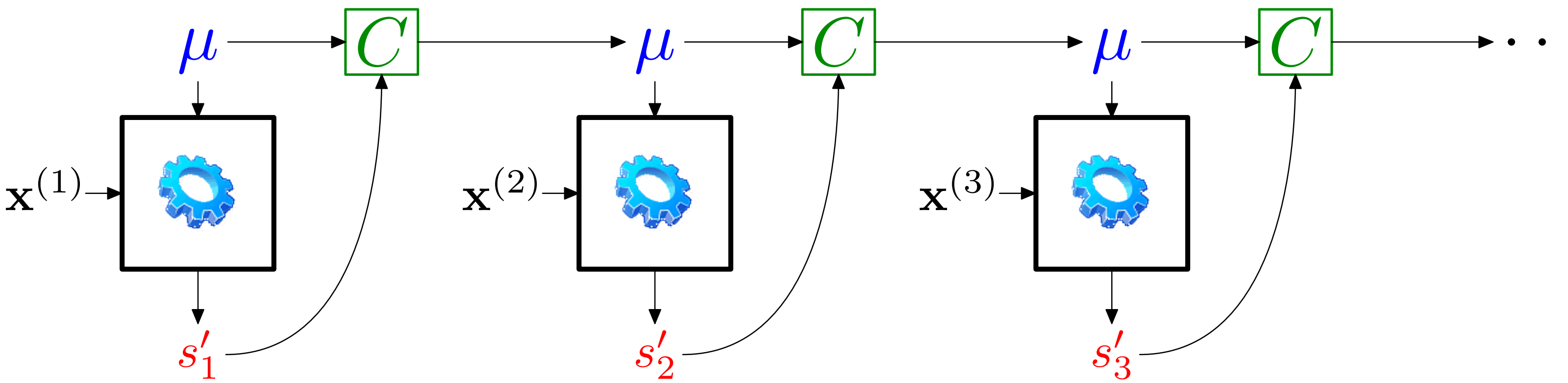


10 data points processed



7 data points processed

Optimization parameter 1 of 2: stepsize



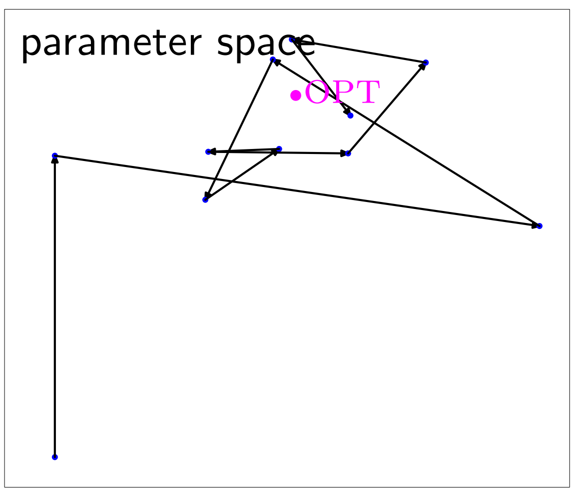
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

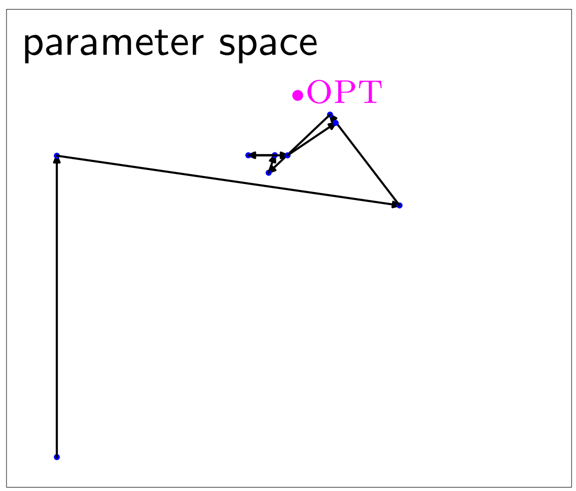
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



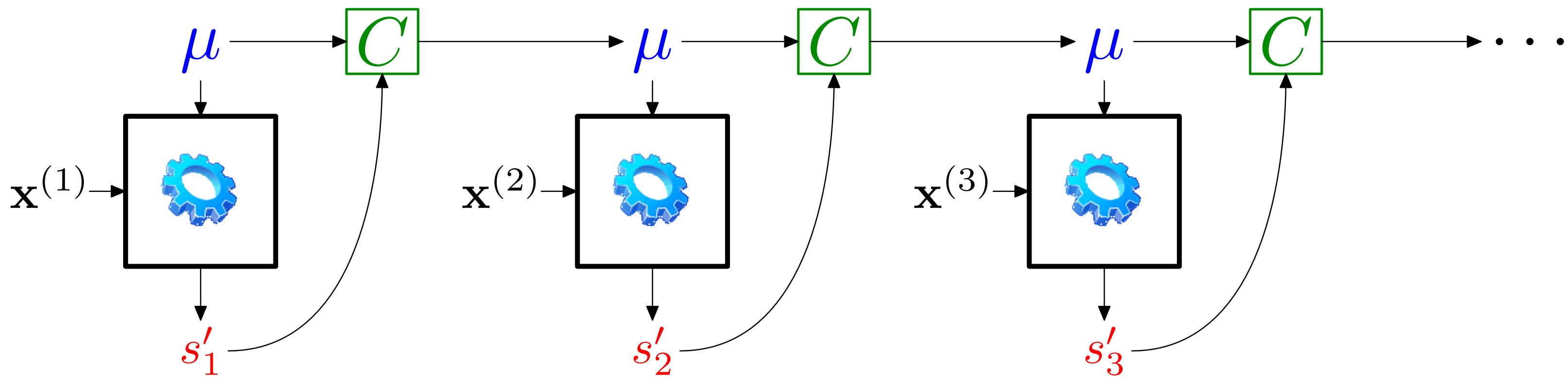
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



8 data points processed

Optimization parameter 1 of 2: stepsize

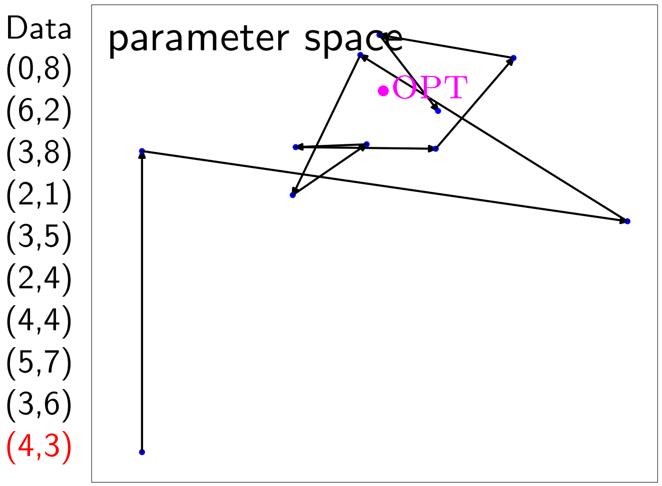


Combine old μ and new s'_i:

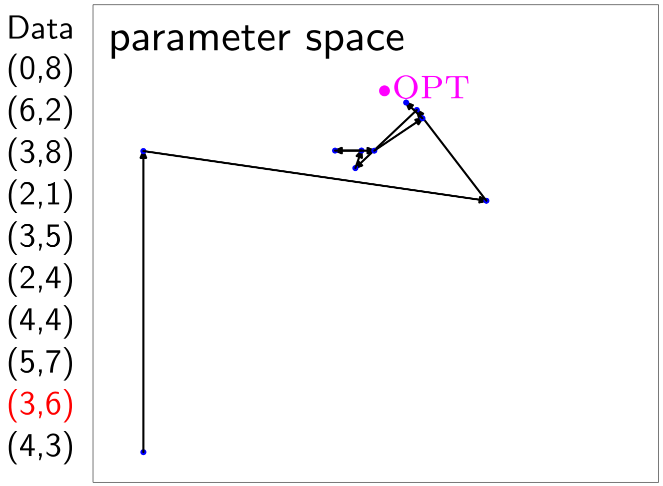
$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

large updates, unstable — small updates, stable

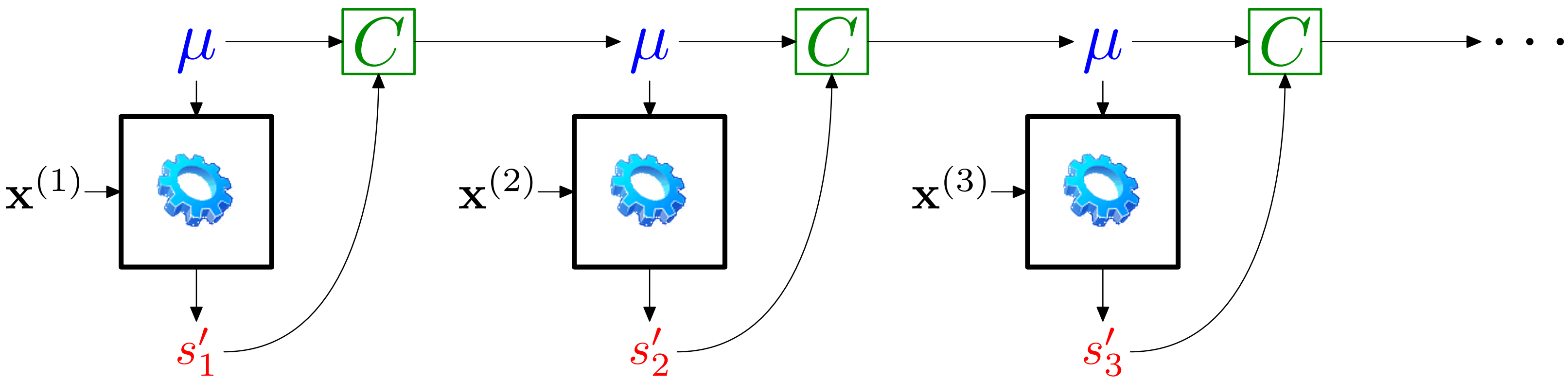


10 data points processed



9 data points processed

Optimization parameter 1 of 2: stepsize

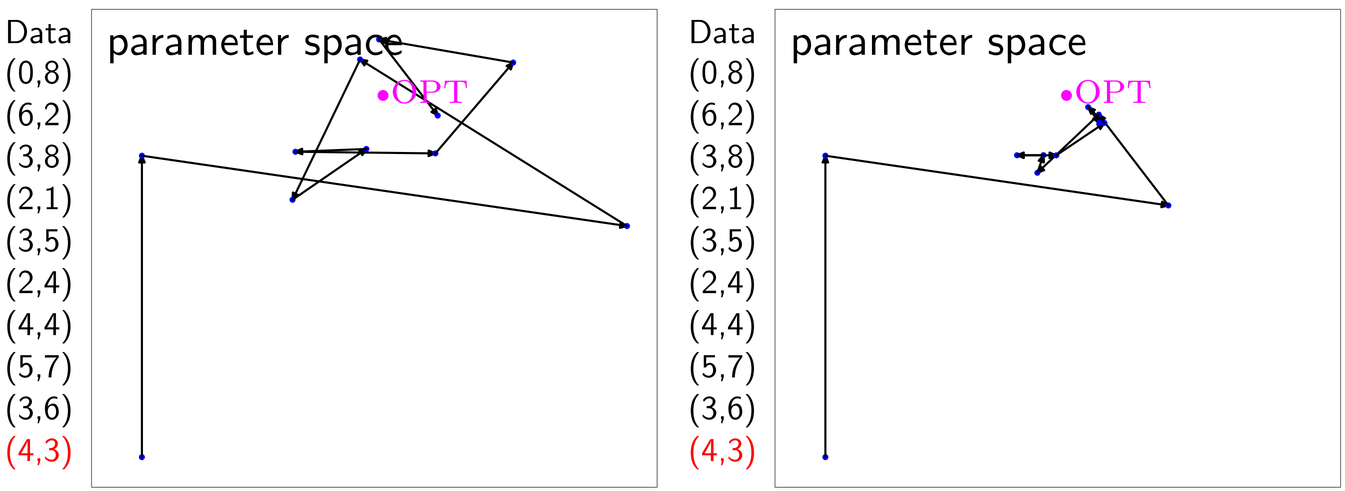


Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

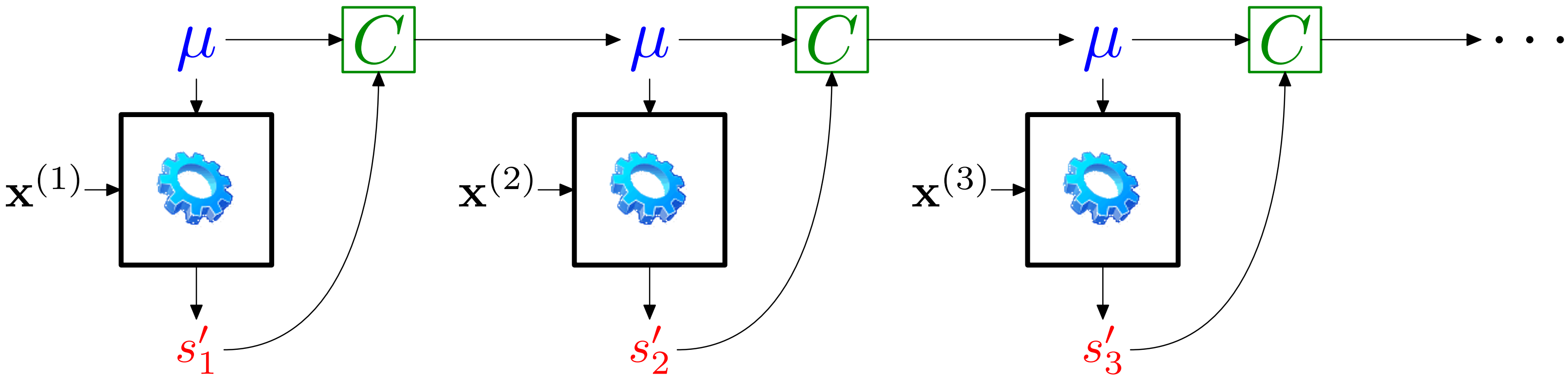
large updates, unstable — small updates, stable



10 data points processed

10 data points processed

Optimization parameter 1 of 2: stepsize

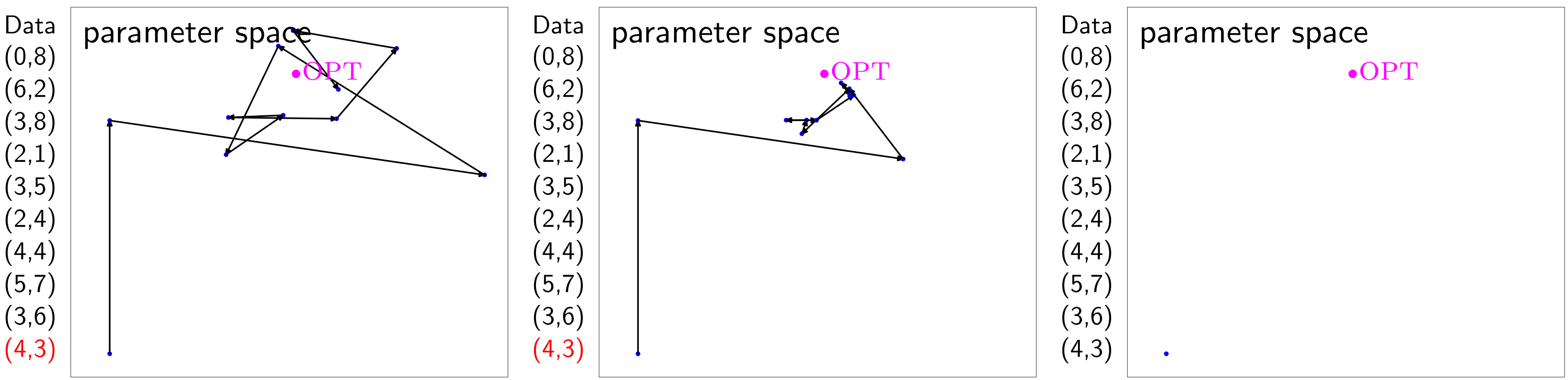


Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

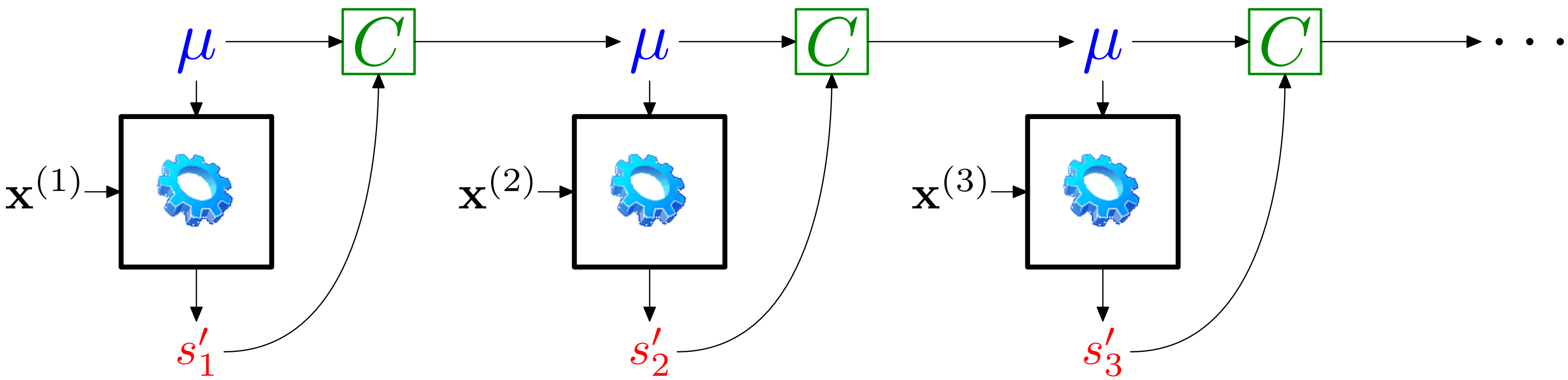
$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

large updates, unstable — small updates, stable



10 data points processed — 10 data points processed — 0 data points processed

Optimization parameter 1 of 2: stepsize

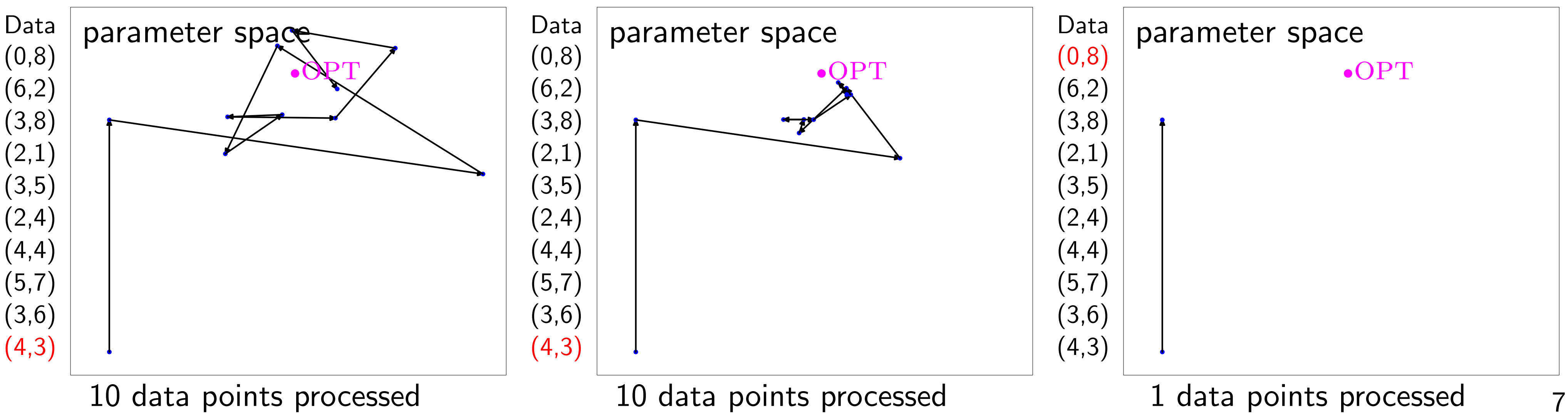


Combine old μ and new s'_i:

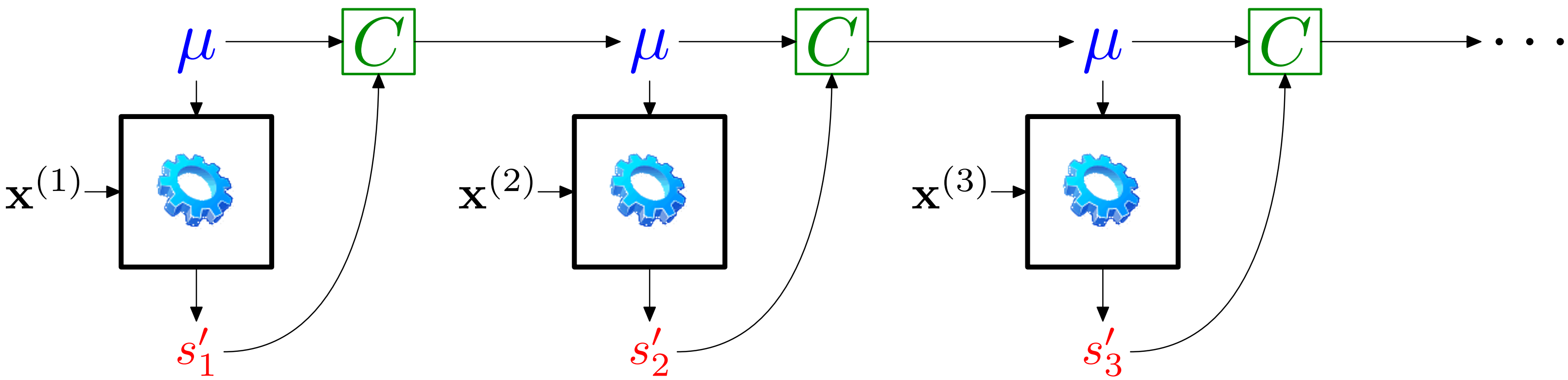
$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟵⟶ $\alpha = 1$

large updates, unstable — small updates, stable



Optimization parameter 1 of 2: stepsize

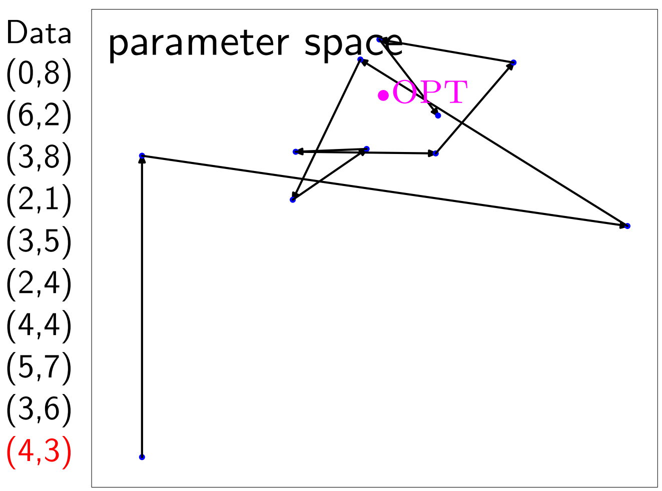


Combine old μ and new s'_i:

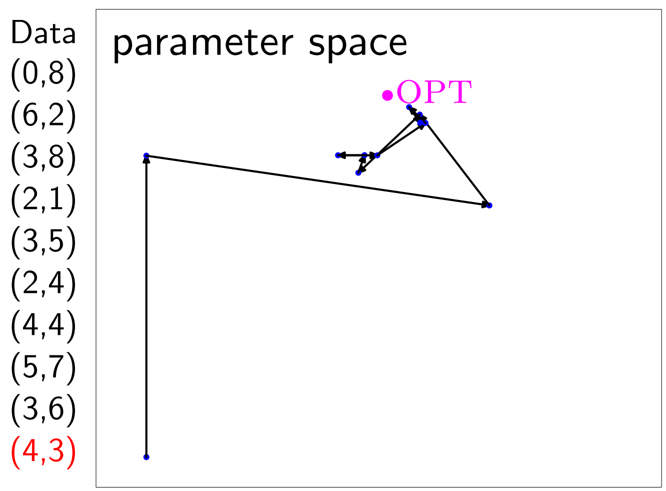
$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

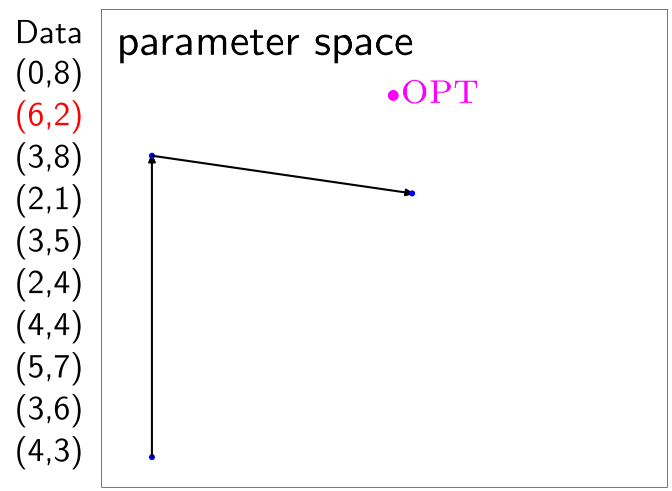
large updates, unstable — small updates, stable



10 data points processed

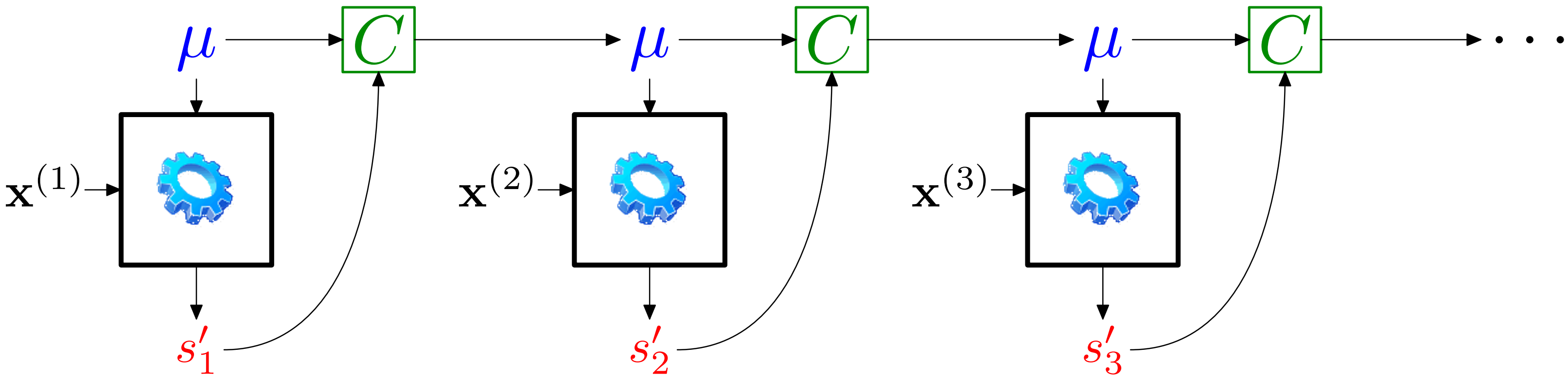


10 data points processed



2 data points processed

Optimization parameter 1 of 2: stepsize



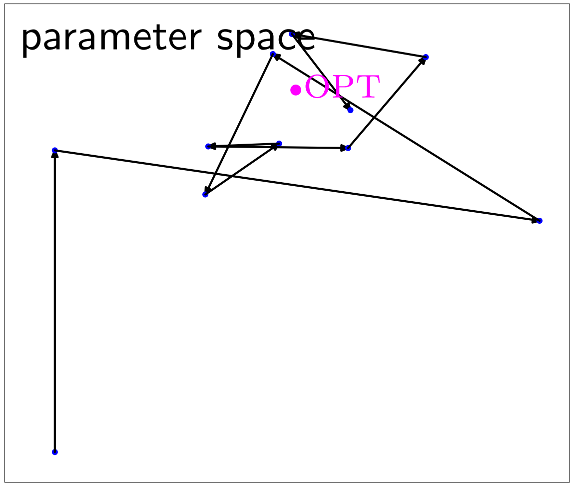
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

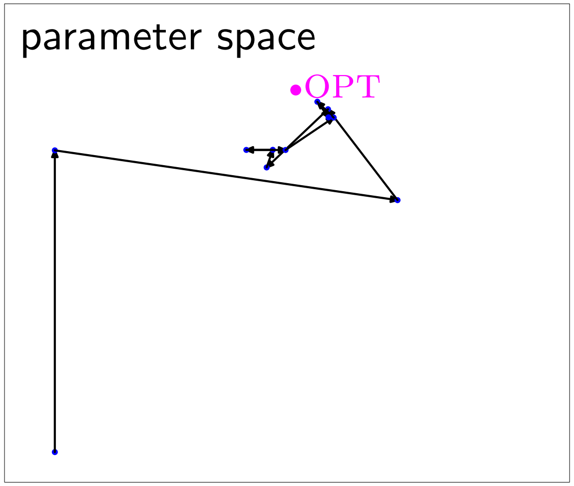
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



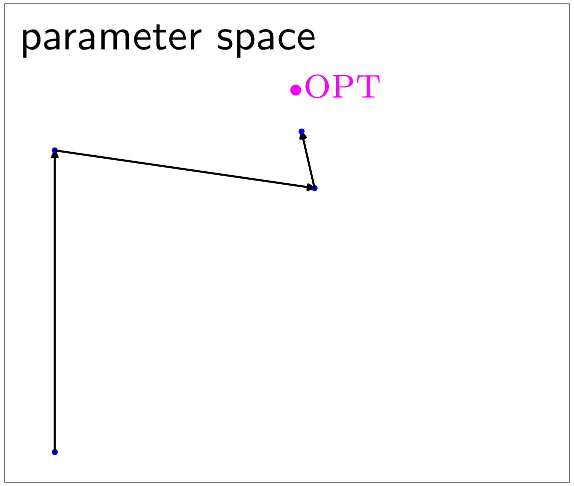
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



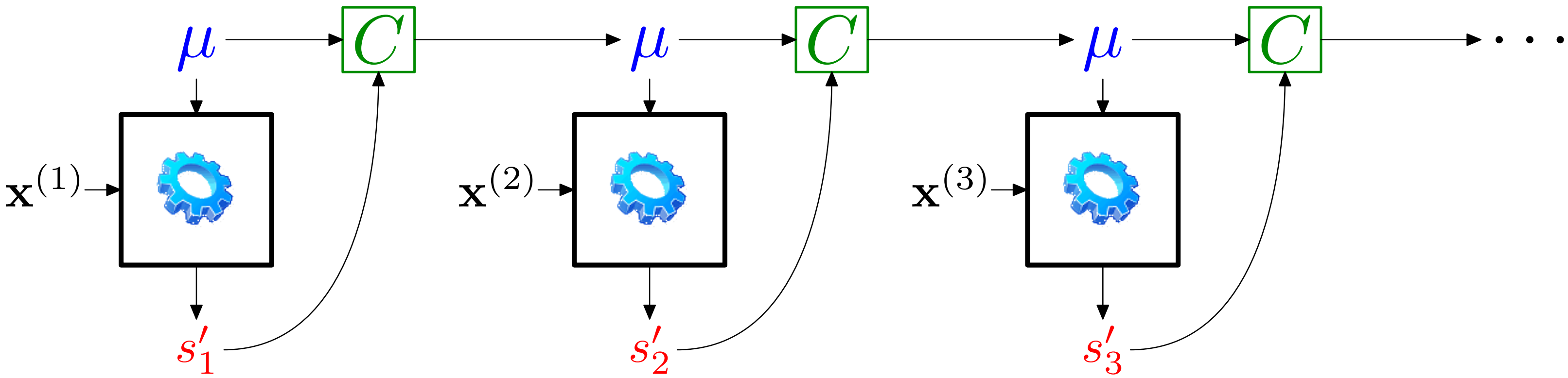
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



3 data points processed

Optimization parameter 1 of 2: stepsize

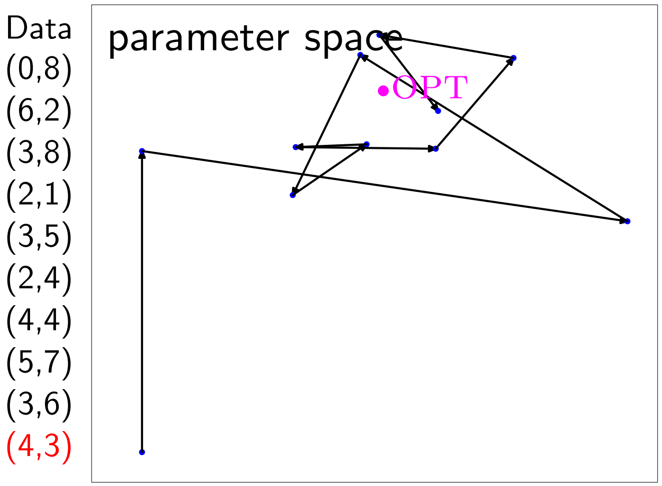


Combine old μ and new s'_i:

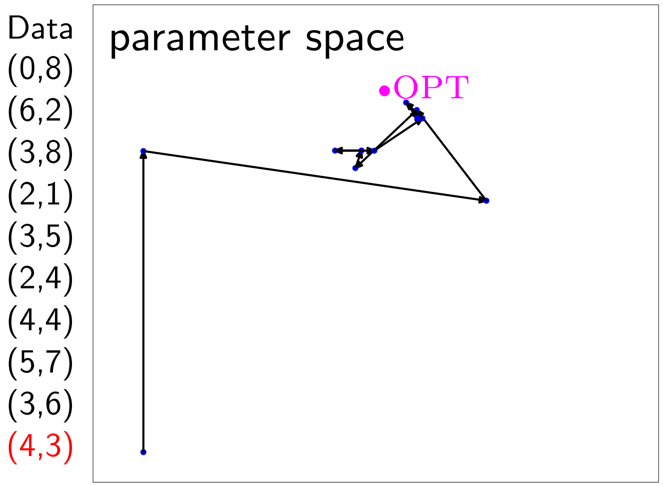
$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$

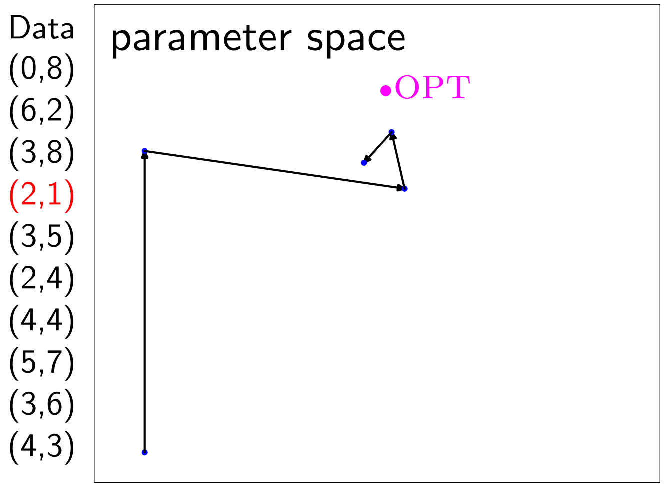
large updates, unstable — small updates, stable



10 data points processed

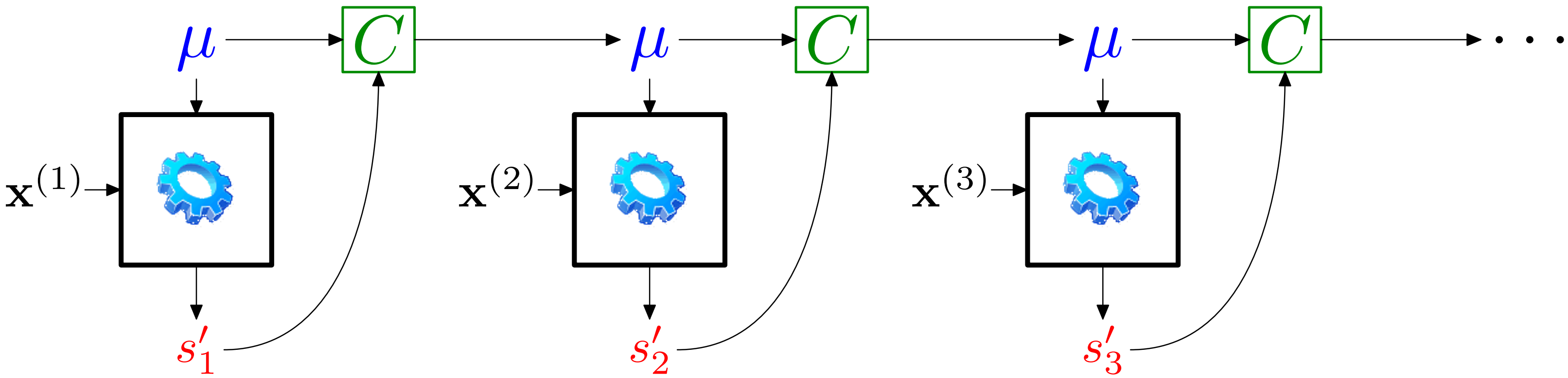


10 data points processed



4 data points processed

Optimization parameter 1 of 2: stepsize

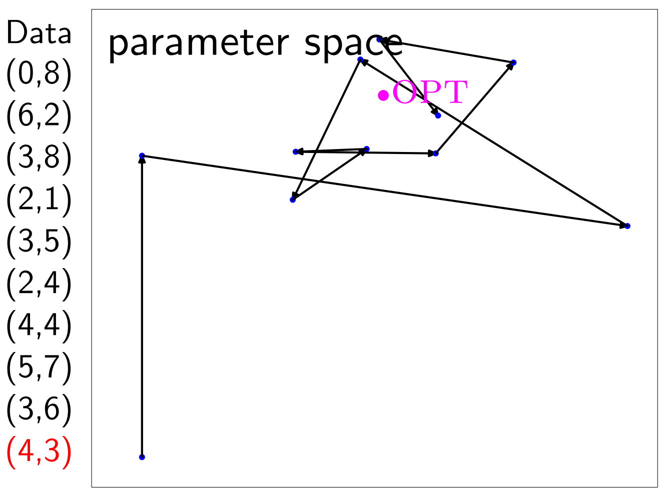


Combine old μ and new s'_i:

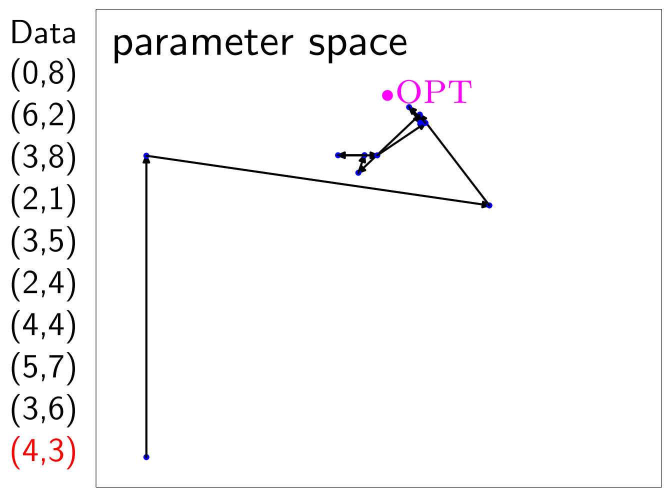
$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

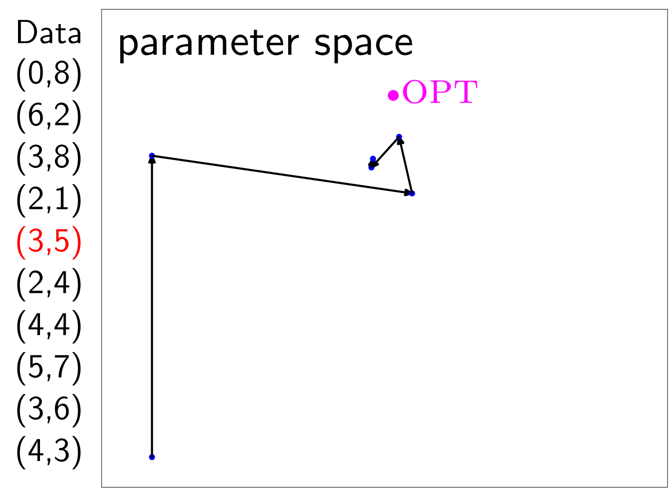
large updates, unstable — small updates, stable



10 data points processed

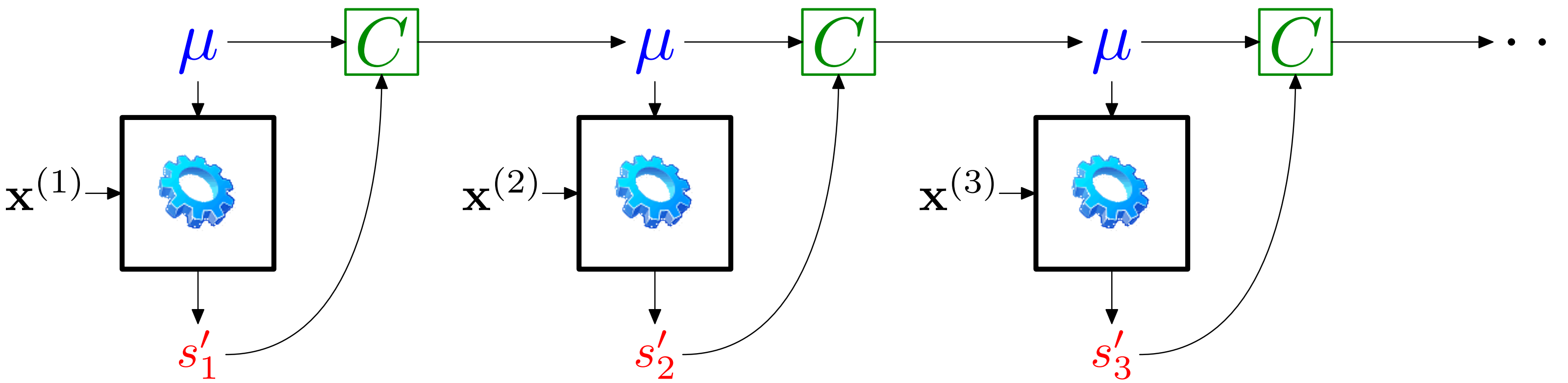


10 data points processed



5 data points processed

Optimization parameter 1 of 2: stepsize

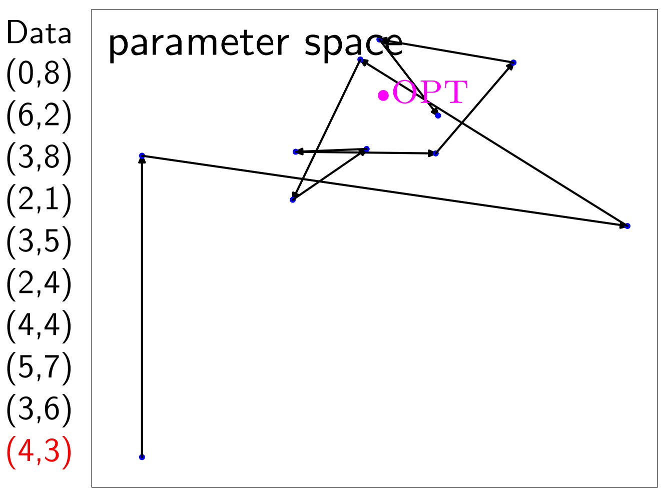


Combine old μ and new s'_i:

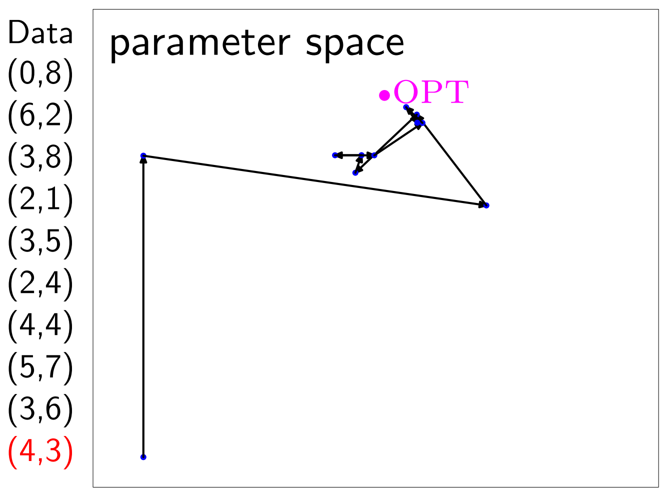
$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

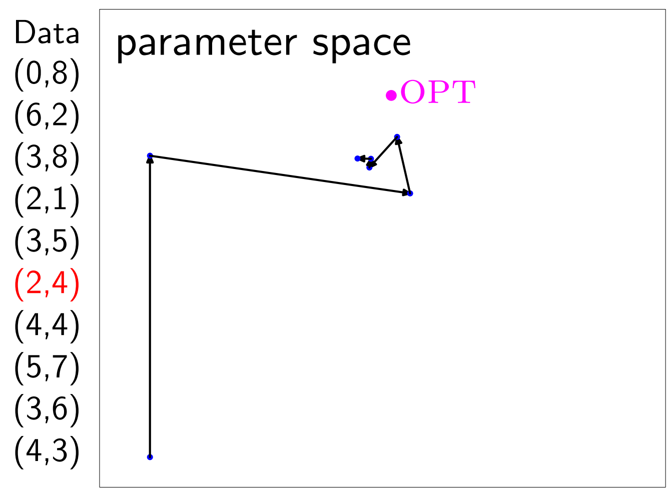
large updates, unstable — small updates, stable



10 data points processed

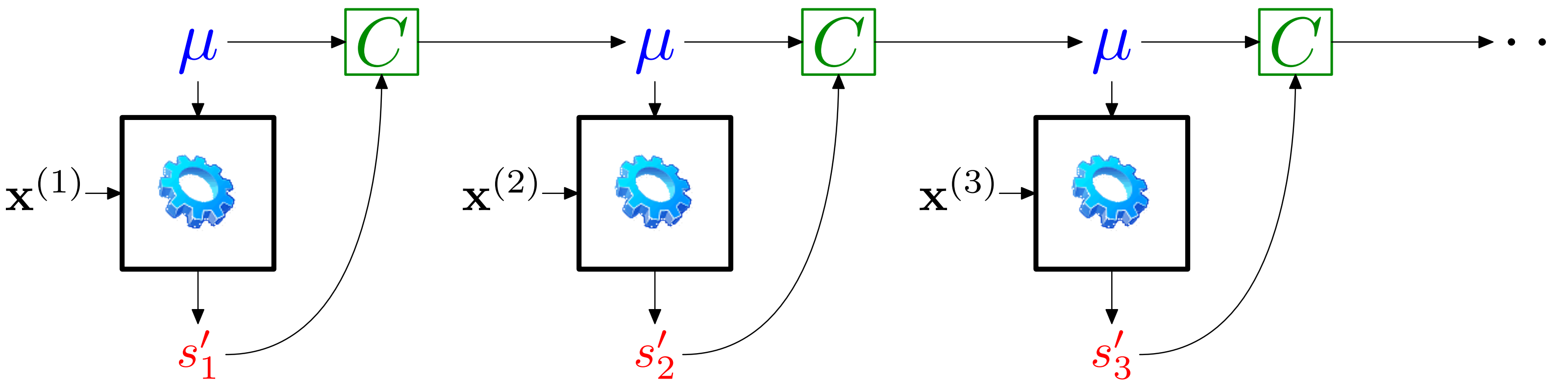


10 data points processed



6 data points processed

Optimization parameter 1 of 2: stepsize

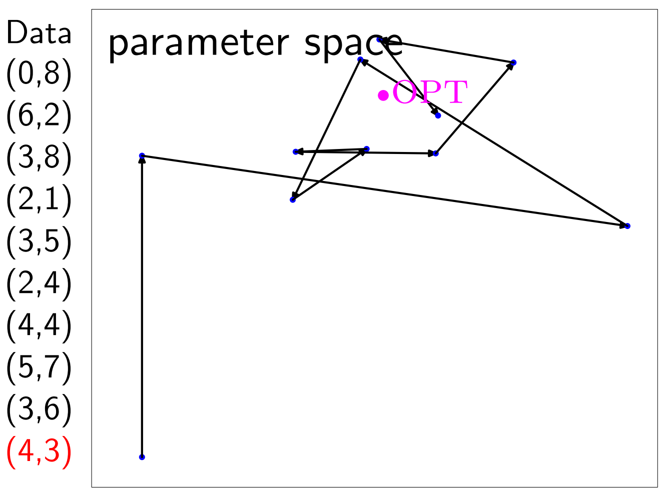


Combine old μ and new s'_i:

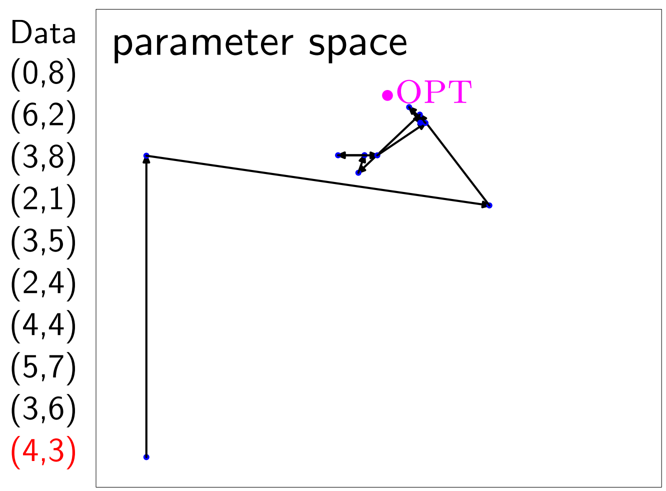
$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ $\longleftrightarrow$ $\alpha = 1$

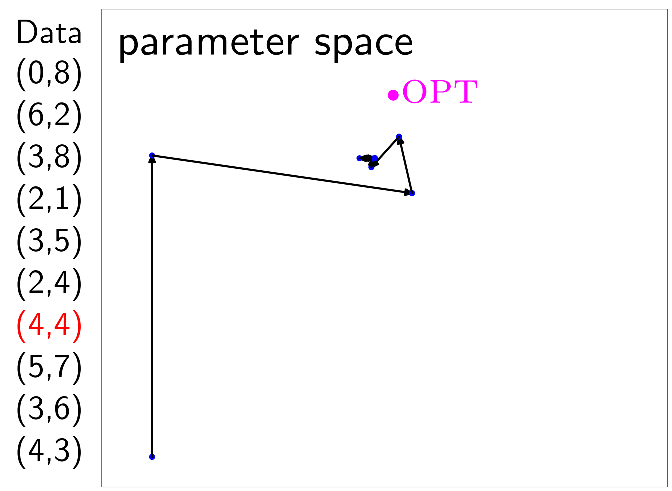
large updates, unstable — small updates, stable



10 data points processed

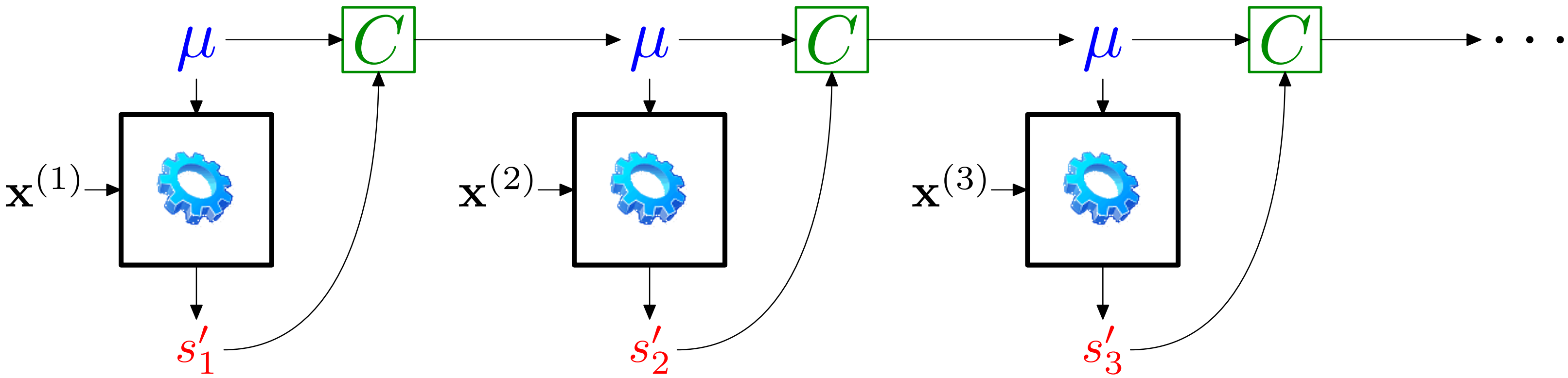


10 data points processed



7 data points processed

Optimization parameter 1 of 2: stepsize



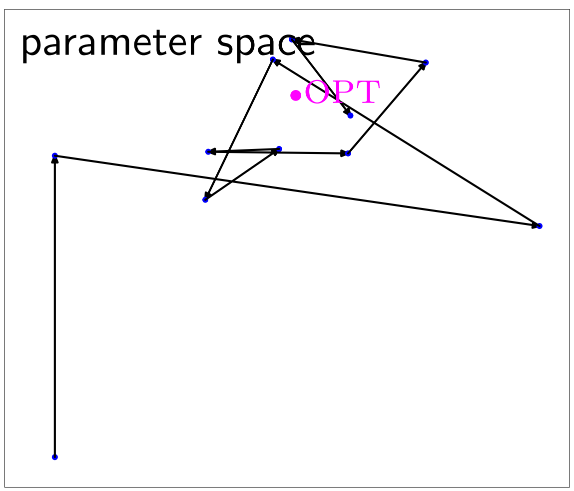
Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

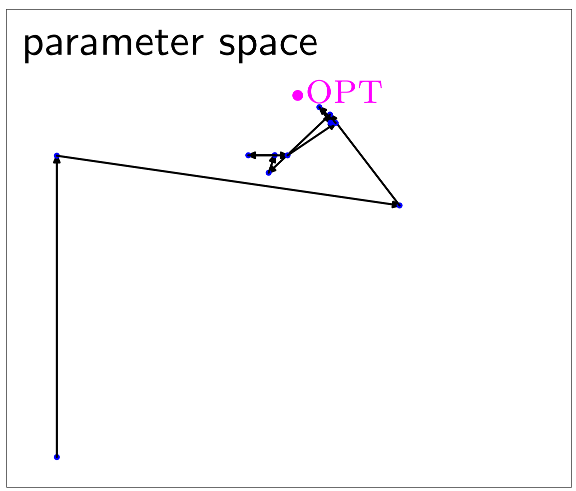
large updates, unstable — small updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



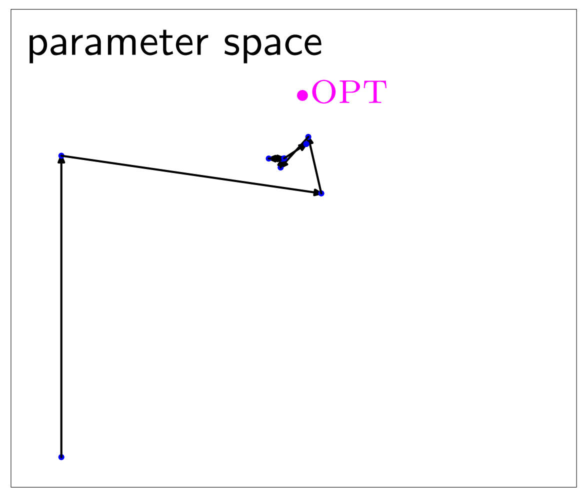
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



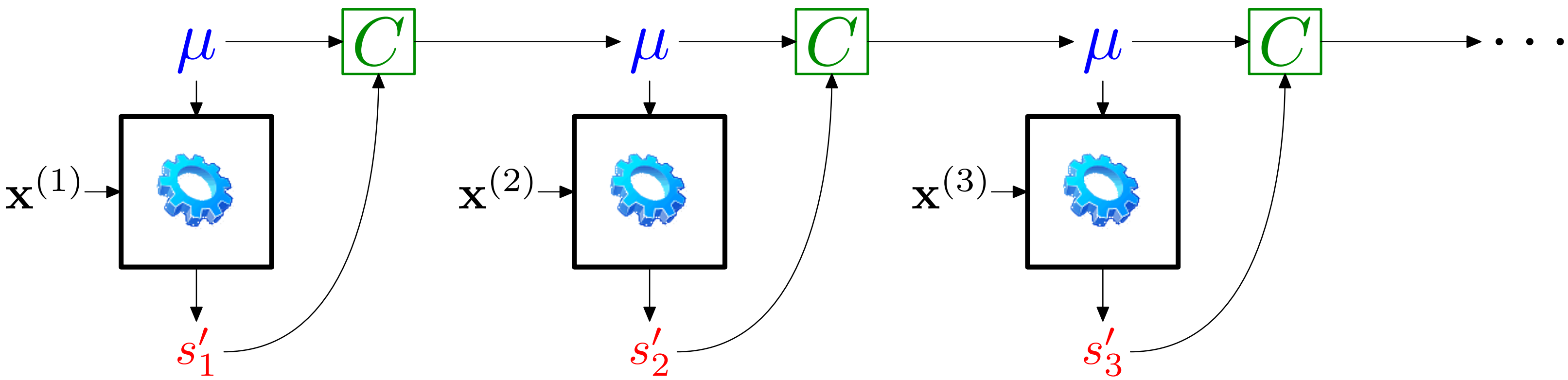
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



8 data points processed

Optimization parameter 1 of 2: stepsize

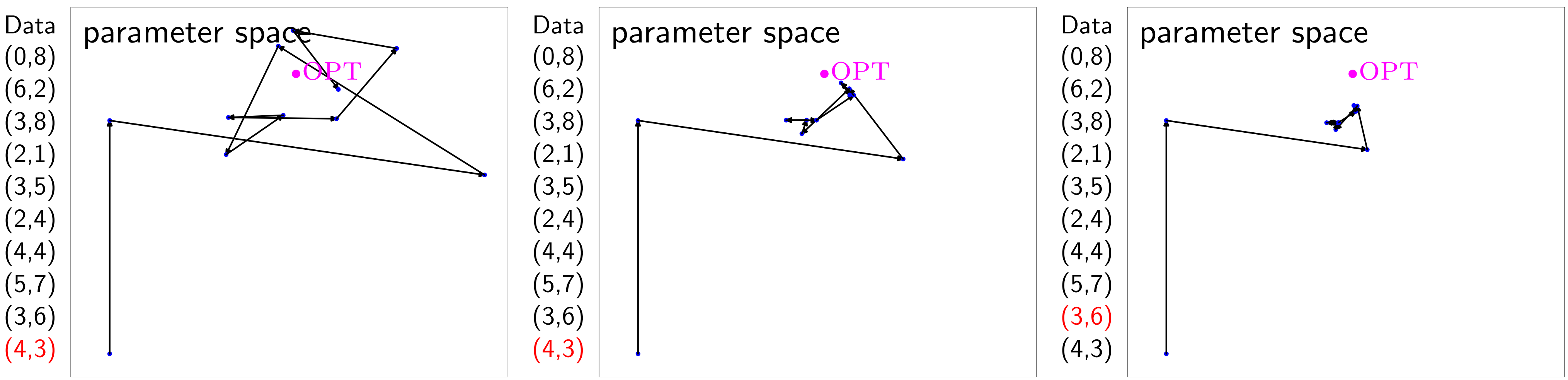


Combine old μ and new s'_i:

$$C(\mu, s'_i) = (1-\eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

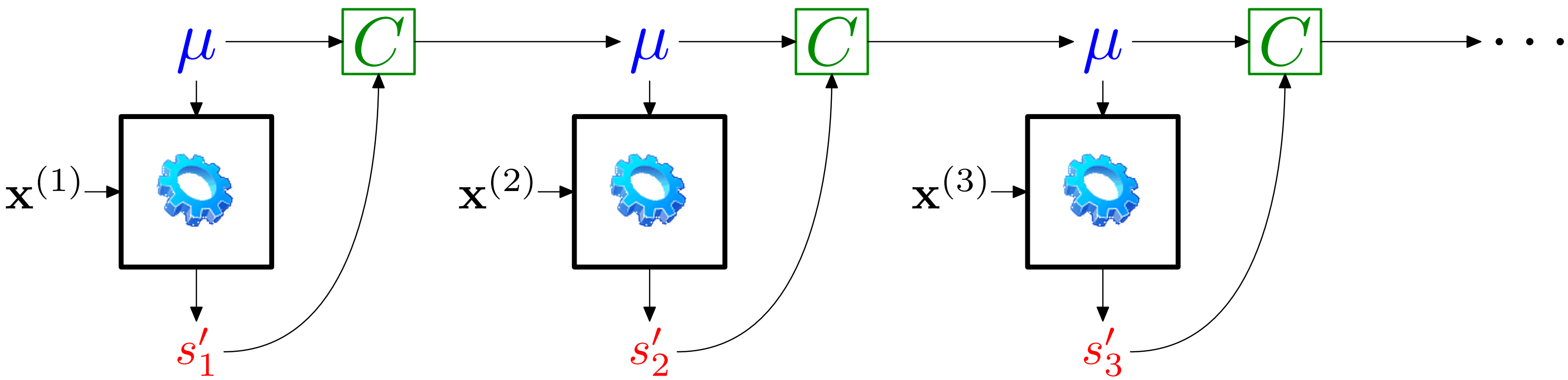
$\alpha = \frac{1}{2}$ ⟷ $\alpha = 1$

large updates, unstable — small updates, stable



10 data points processed — 10 data points processed — 9 data points processed

Optimization parameter 1 of 2: stepsize

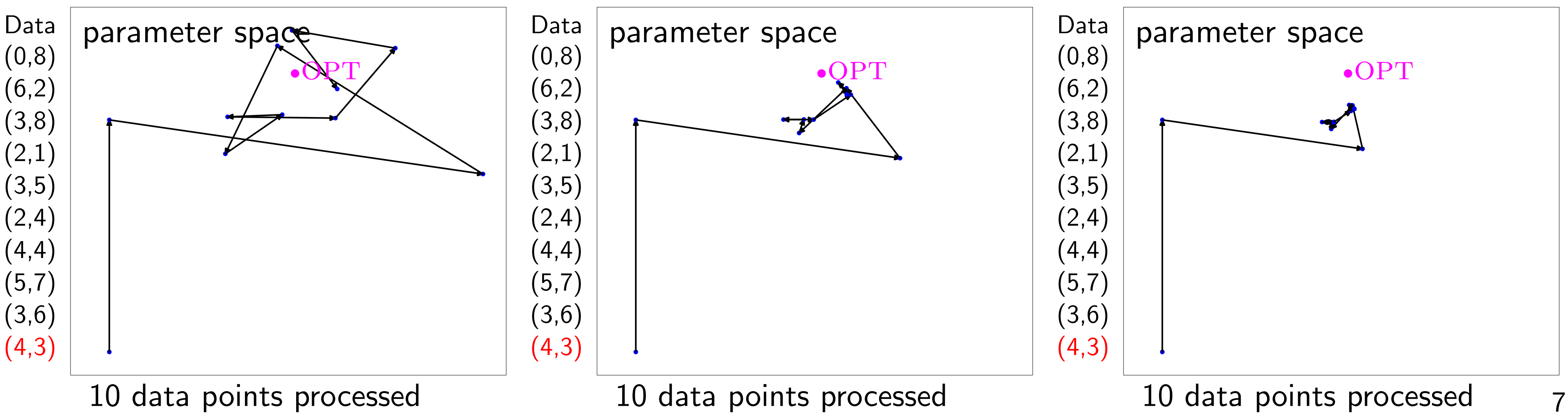


Combine old μ and new s'_i:

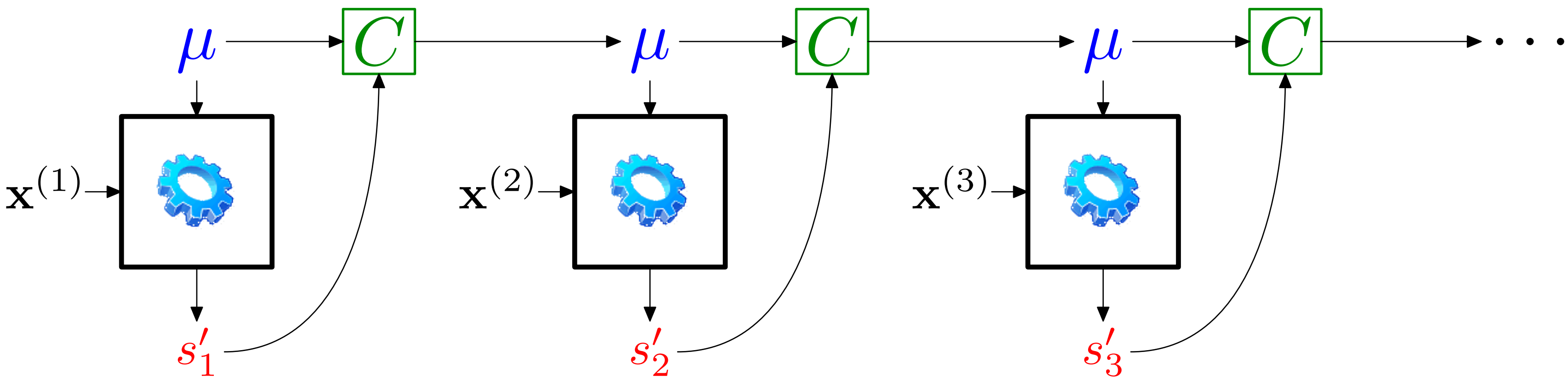
$$C(\mu, s'_i) = (1 - \eta_k)\mu + \eta_k s'_i, \quad \eta_k = \frac{1}{k^\alpha} \text{ on } k\text{-th update}$$

$$\alpha = \frac{1}{2} \longleftrightarrow \alpha = 1$$

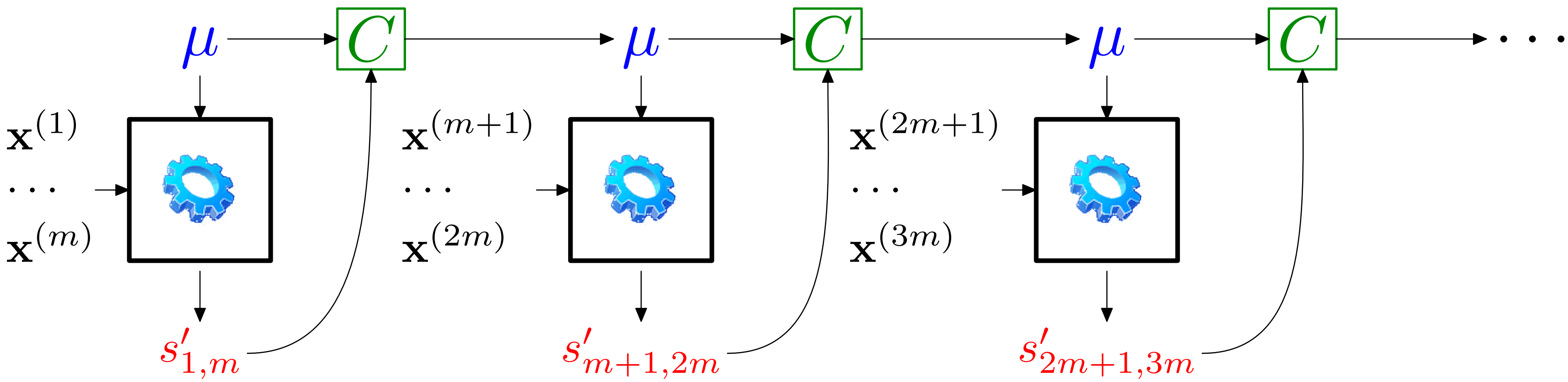
large updates, unstable — small updates, stable



Optimization parameter 2 of 2: minibatch size

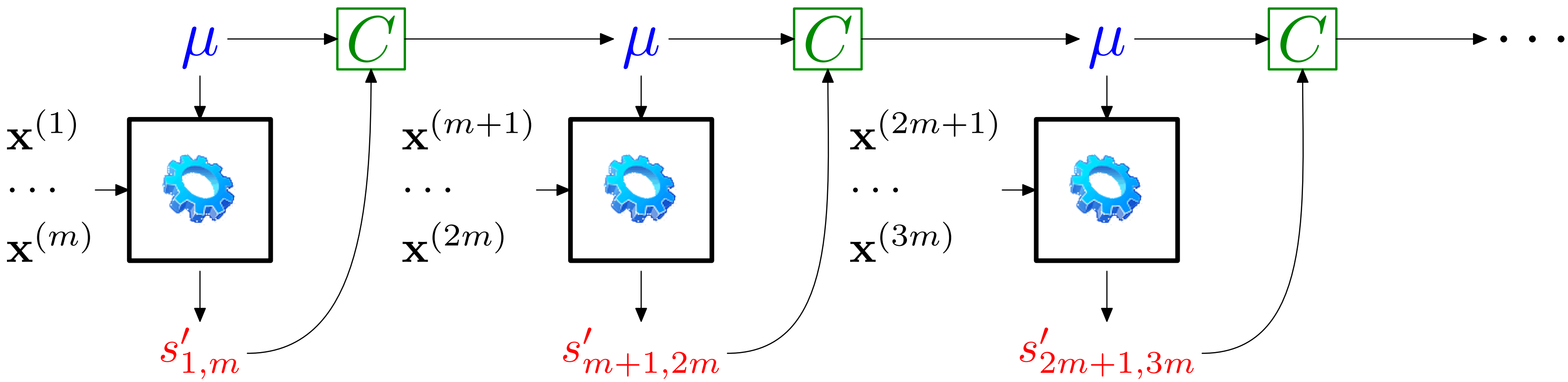


Optimization parameter 2 of 2: minibatch size



$m =$ size of a mini-batch

Optimization parameter 2 of 2: minibatch size

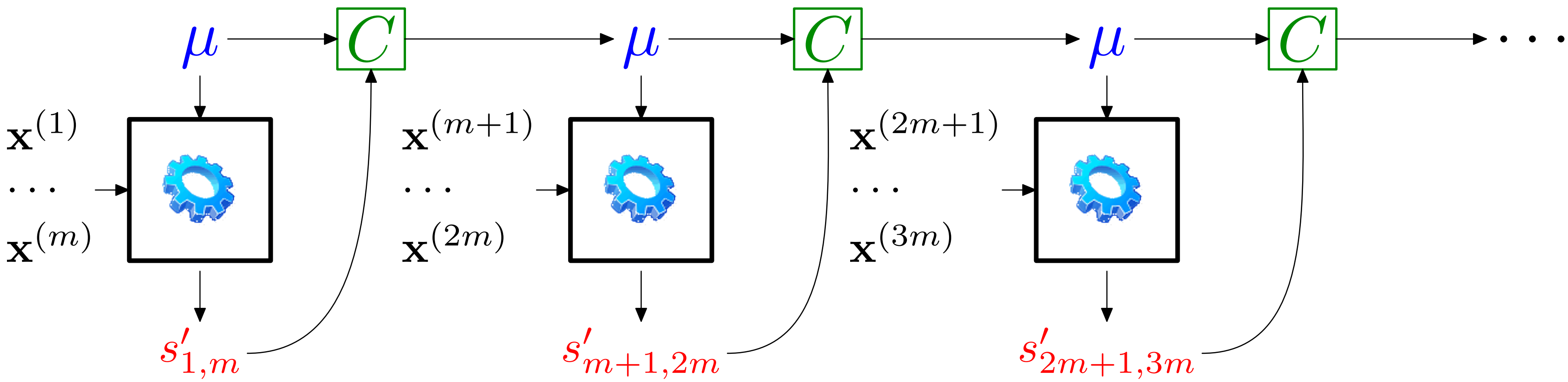


$m =$ size of a mini-batch

$m = 1 \longleftrightarrow m = n$

frequent updates, unstable — infrequent updates, stable

Optimization parameter 2 of 2: minibatch size

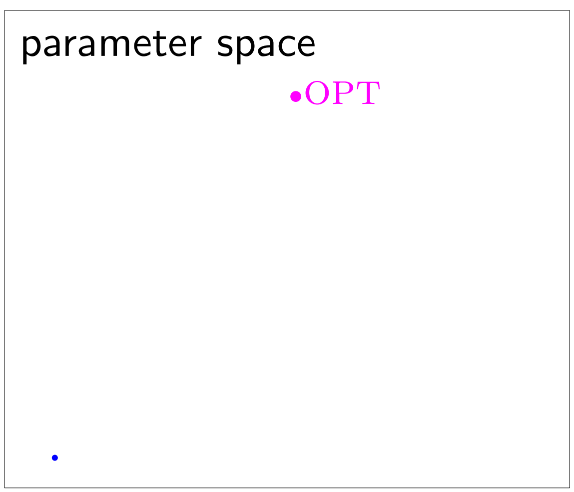


m = size of a mini-batch

$m = 1 \longleftrightarrow m = n$

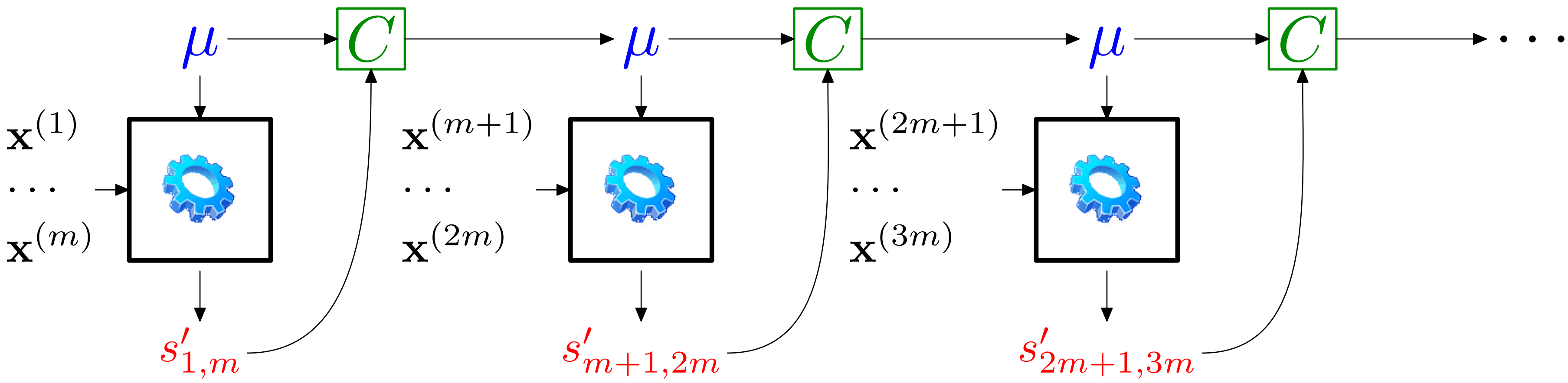
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



0 data points processed

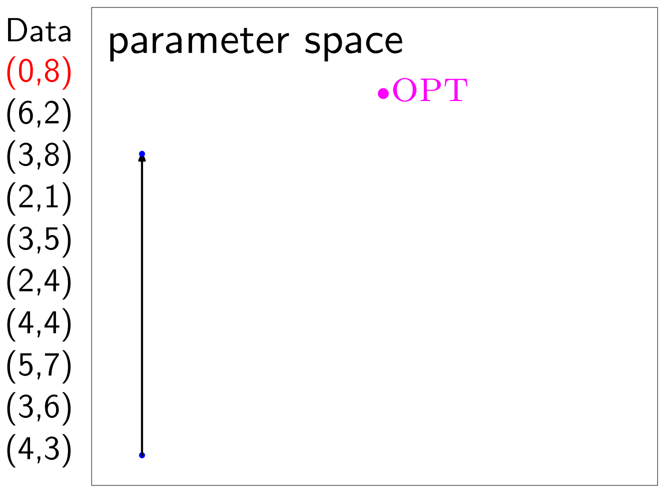
Optimization parameter 2 of 2: minibatch size



m = size of a mini-batch

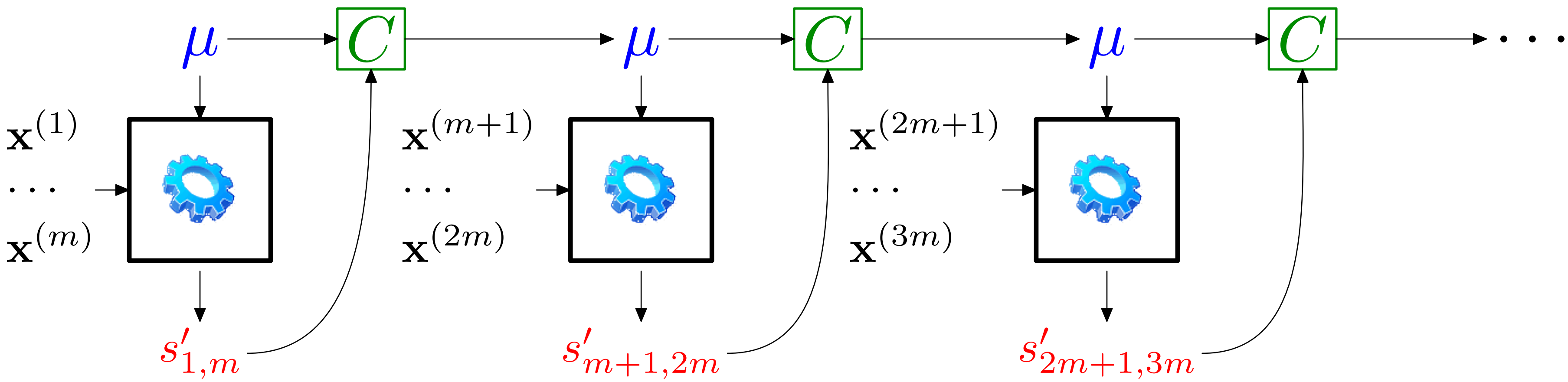
$m = 1 \longleftrightarrow m = n$

frequent updates, unstable — infrequent updates, stable



1 data points processed

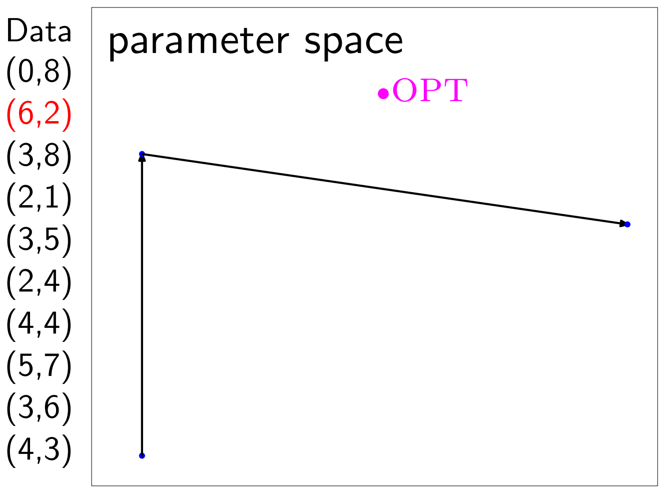
Optimization parameter 2 of 2: minibatch size



m = size of a mini-batch

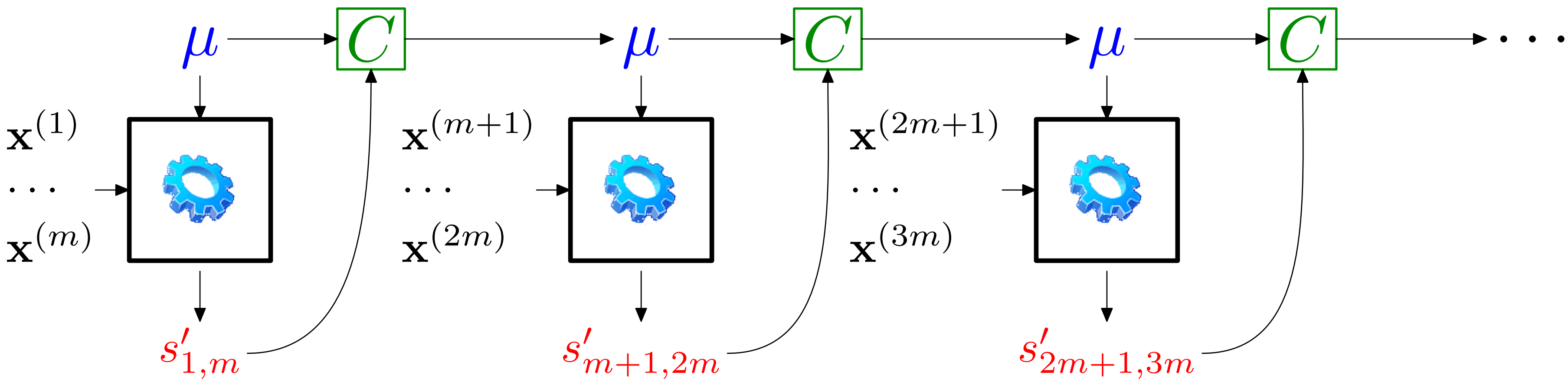
$m = 1 \longleftrightarrow m = n$

frequent updates, unstable — infrequent updates, stable



2 data points processed

Optimization parameter 2 of 2: minibatch size

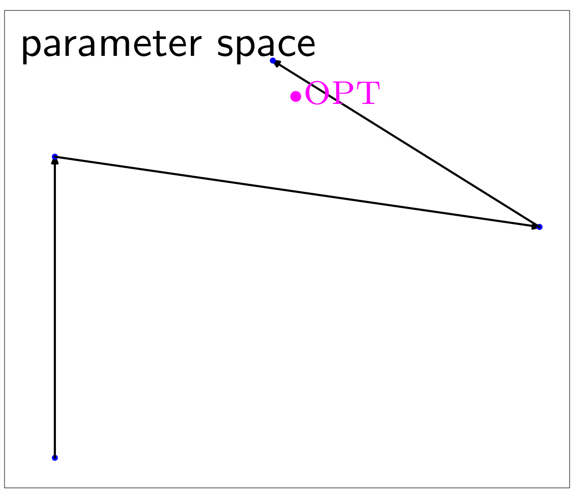


$m =$ size of a mini-batch

$m = 1 \longleftrightarrow m = n$

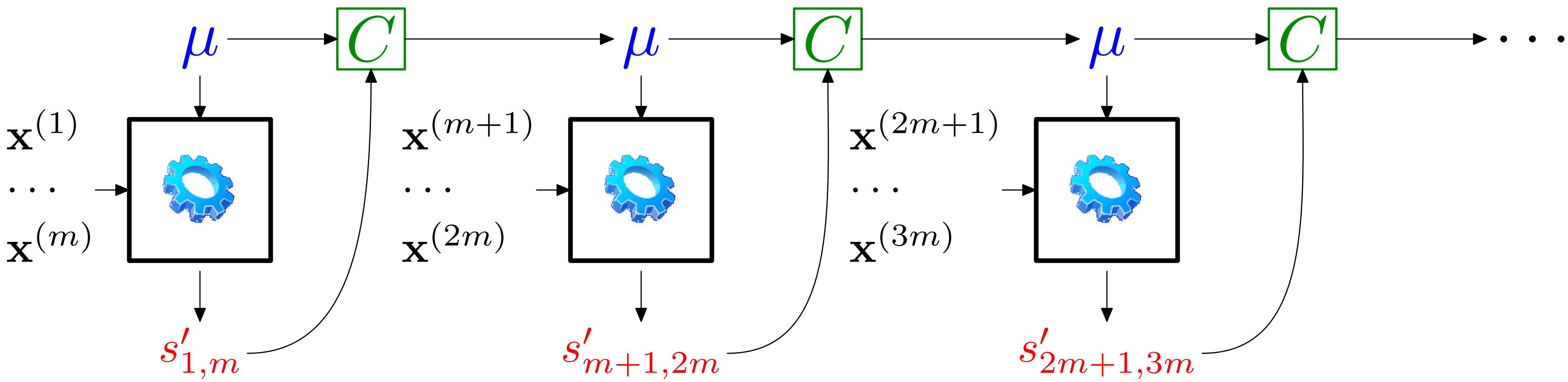
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



3 data points processed

Optimization parameter 2 of 2: minibatch size

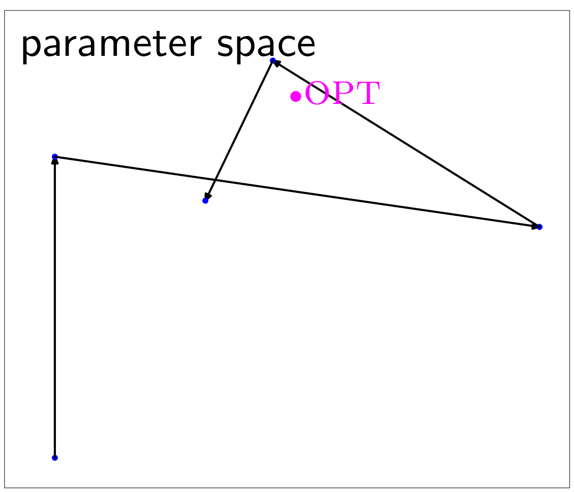


$m =$ size of a mini-batch

$m = 1$ ⟷ $m = n$

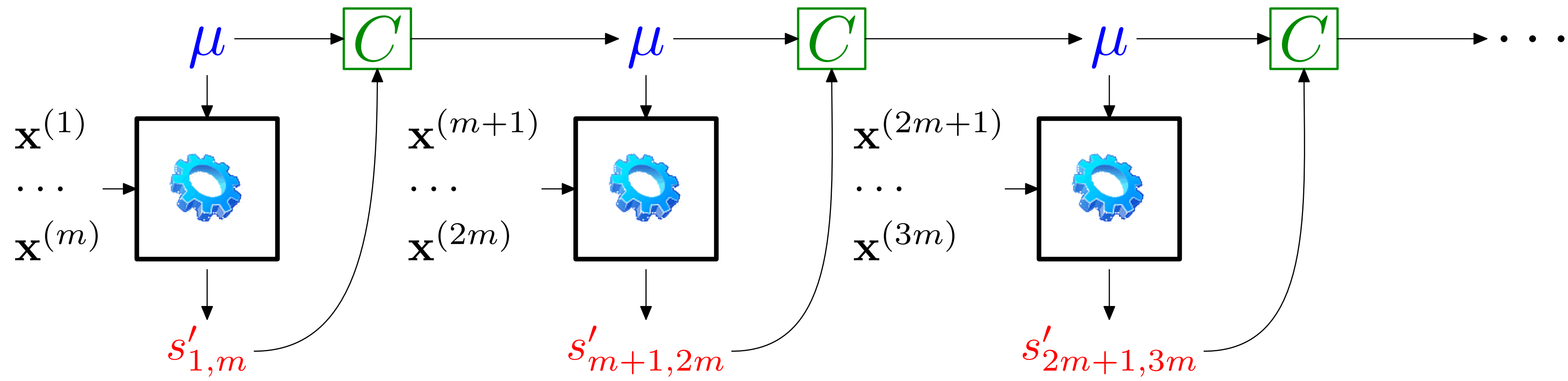
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



4 data points processed

Optimization parameter 2 of 2: minibatch size

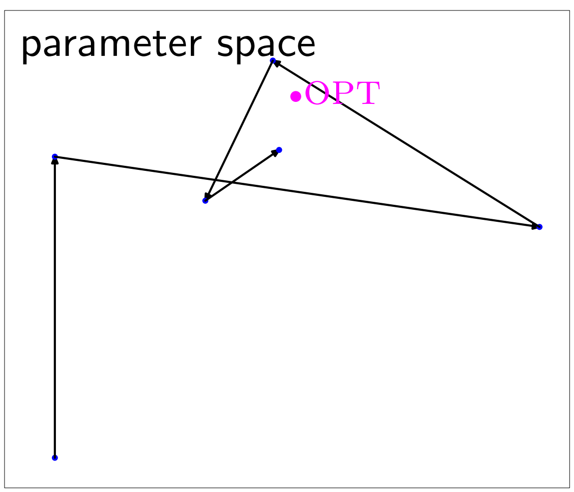


m = size of a mini-batch

$m = 1 \longleftrightarrow m = n$

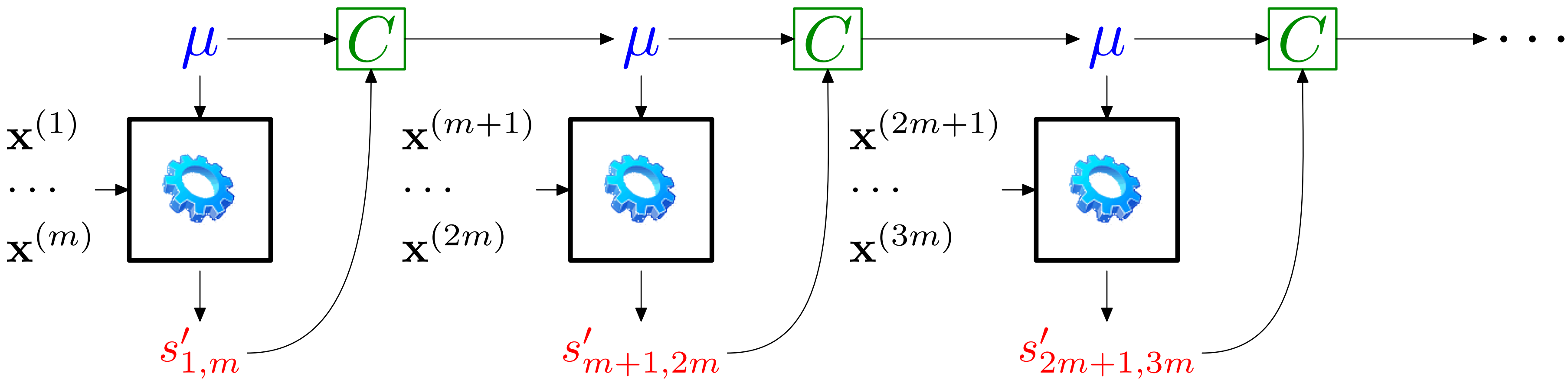
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



5 data points processed

Optimization parameter 2 of 2: minibatch size

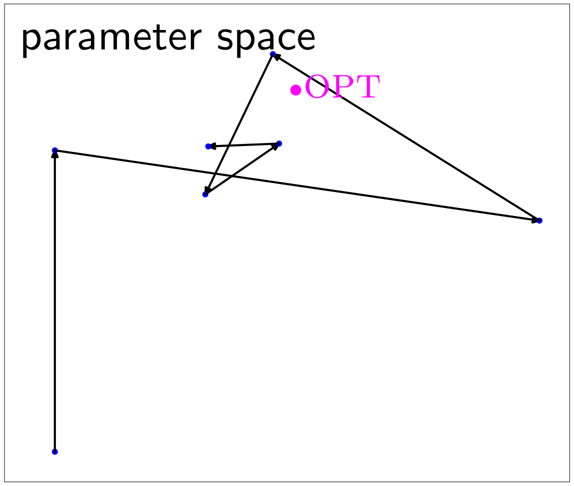


m = size of a mini-batch

$m = 1$ ⟷ $m = n$

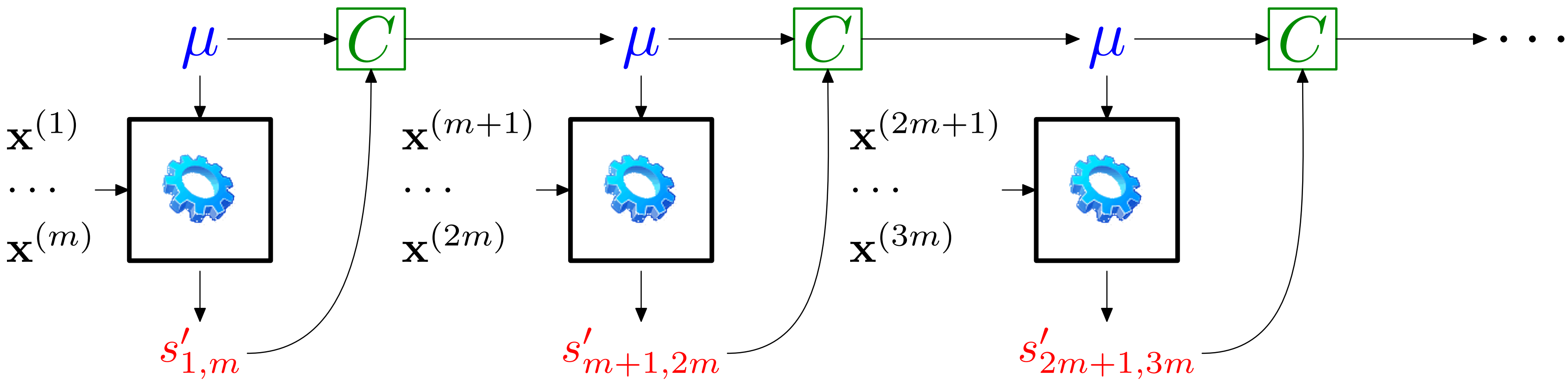
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



6 data points processed

Optimization parameter 2 of 2: minibatch size

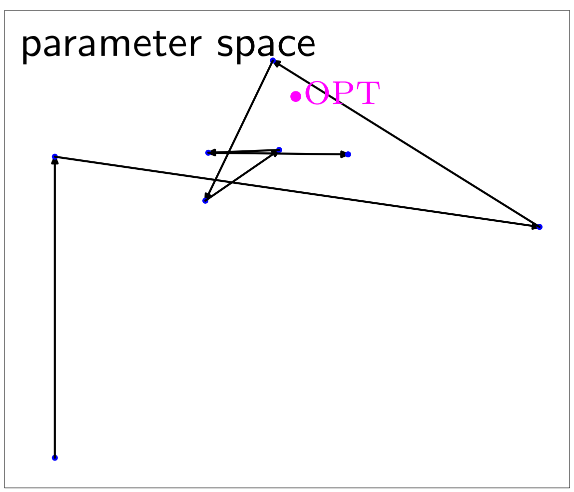


m = size of a mini-batch

$m = 1 \longleftrightarrow m = n$

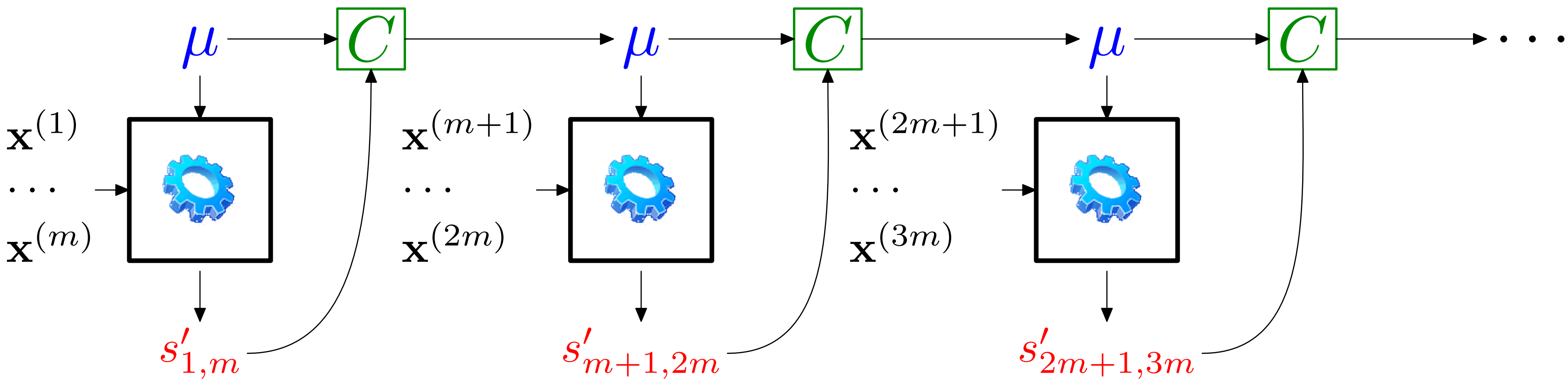
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



7 data points processed

Optimization parameter 2 of 2: minibatch size

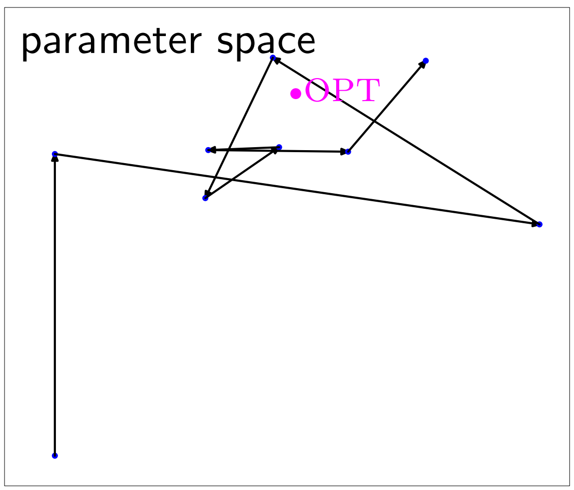


$m =$ size of a mini-batch

$m = 1 \longleftrightarrow m = n$

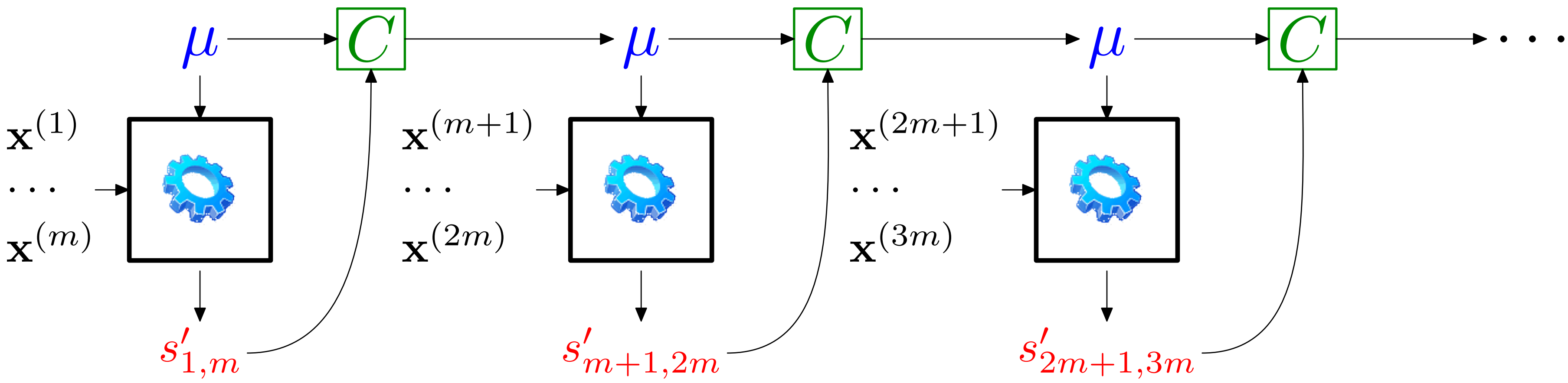
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



8 data points processed

Optimization parameter 2 of 2: minibatch size

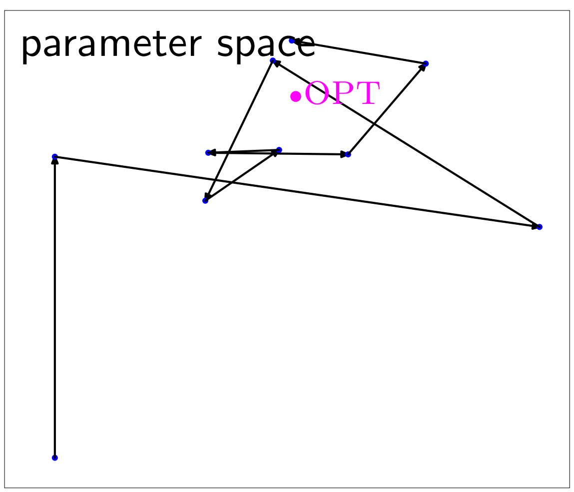


m = size of a mini-batch

$$m = 1 \longleftrightarrow m = n$$

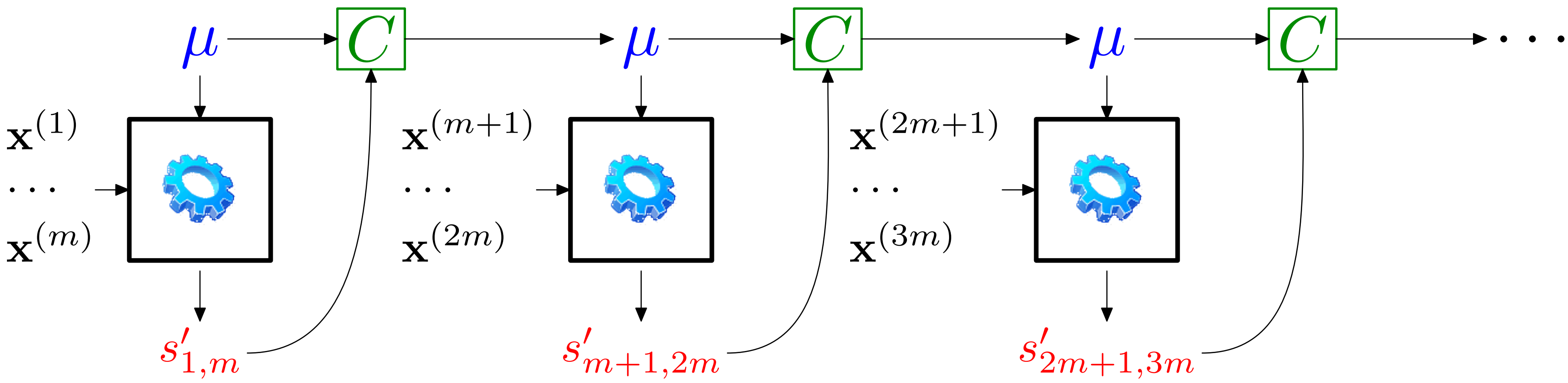
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



9 data points processed

Optimization parameter 2 of 2: minibatch size

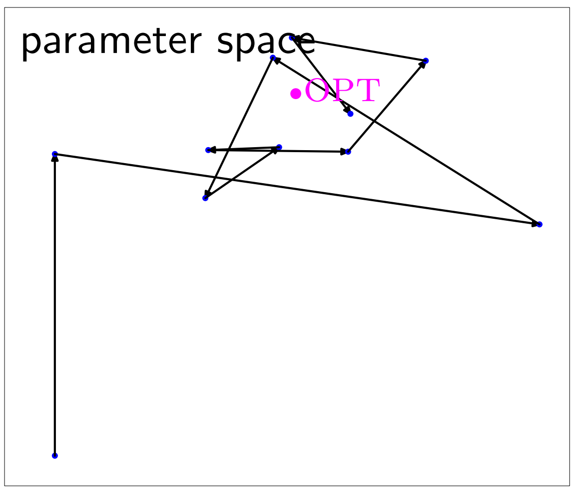


$$m = \text{size of a mini-batch}$$

$m = 1 \longleftrightarrow m = n$

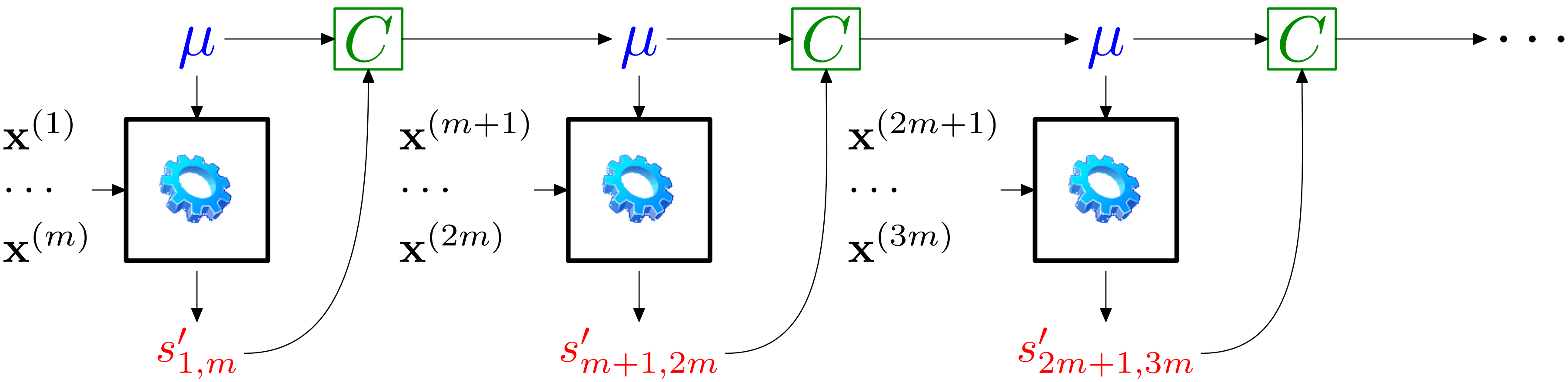
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



10 data points processed

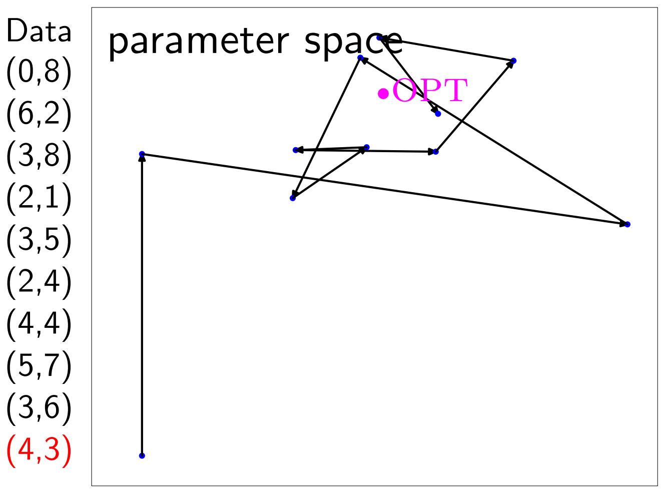
Optimization parameter 2 of 2: minibatch size



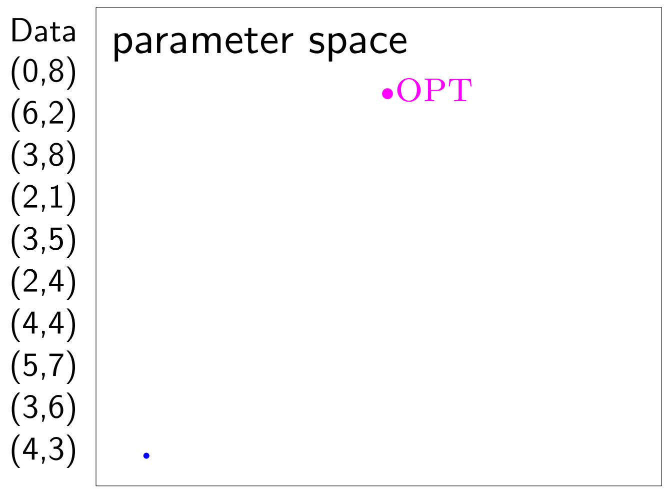
m = size of a mini-batch

$m = 1 \longleftrightarrow m = n$

frequent updates, unstable — infrequent updates, stable

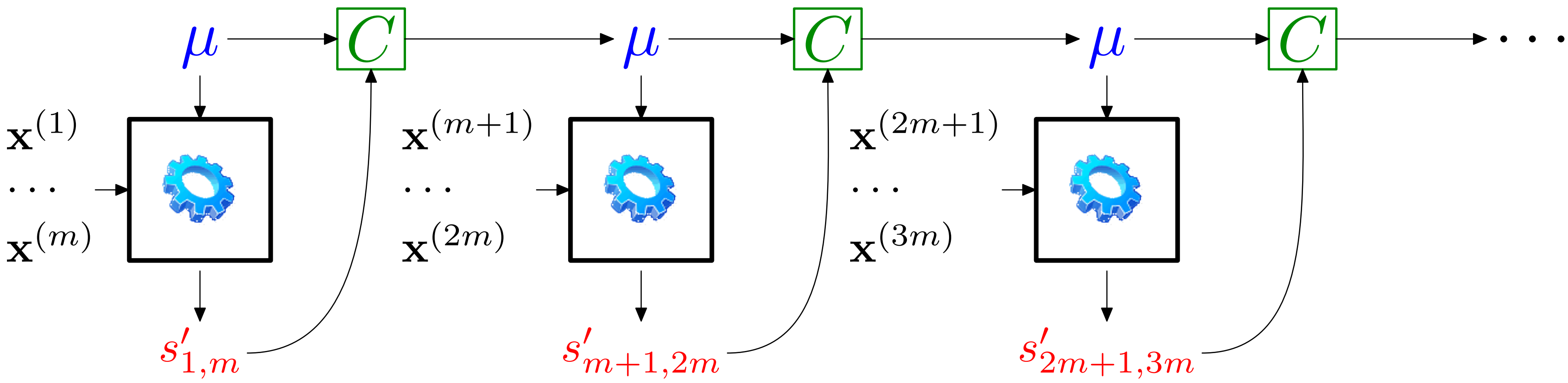


10 data points processed



0 data points processed

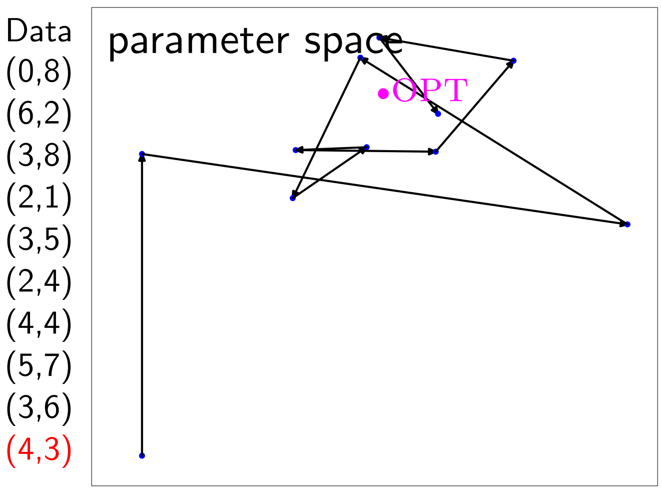
Optimization parameter 2 of 2: minibatch size



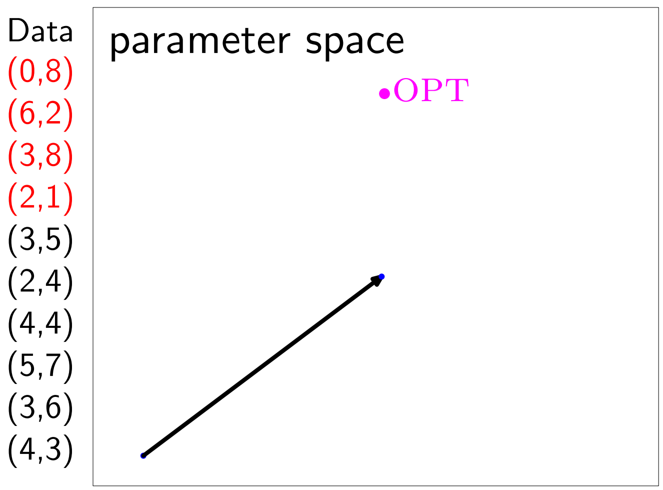
$m =$ size of a mini-batch

$m = 1 \longleftrightarrow m = n$

frequent updates, unstable — infrequent updates, stable

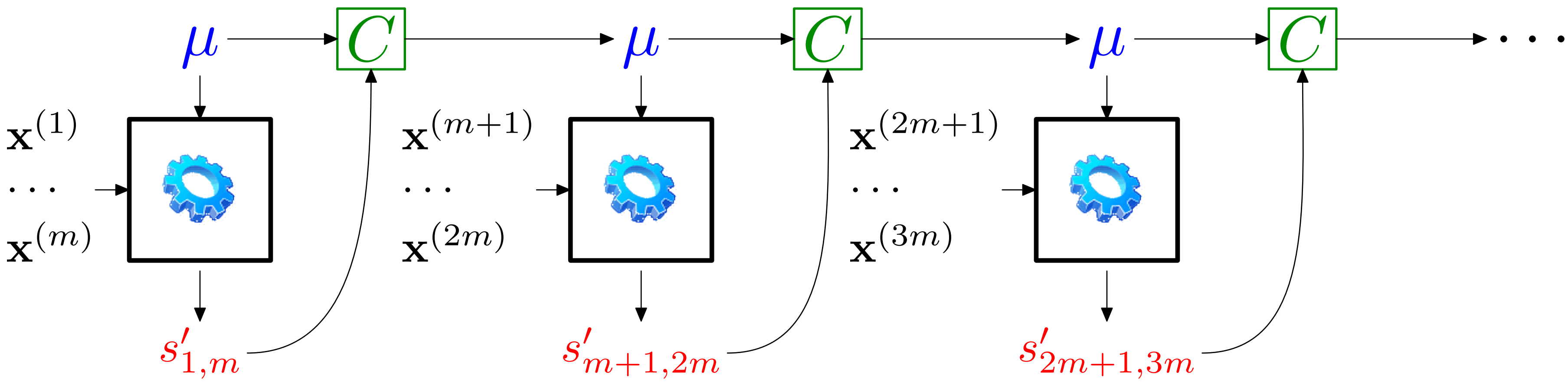


10 data points processed



4 data points processed

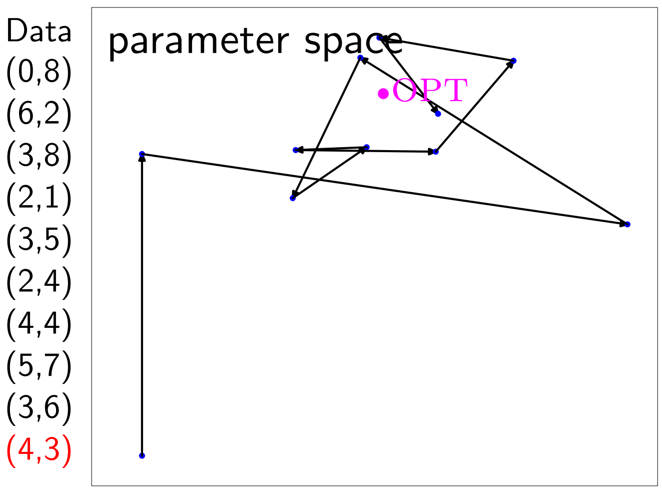
Optimization parameter 2 of 2: minibatch size



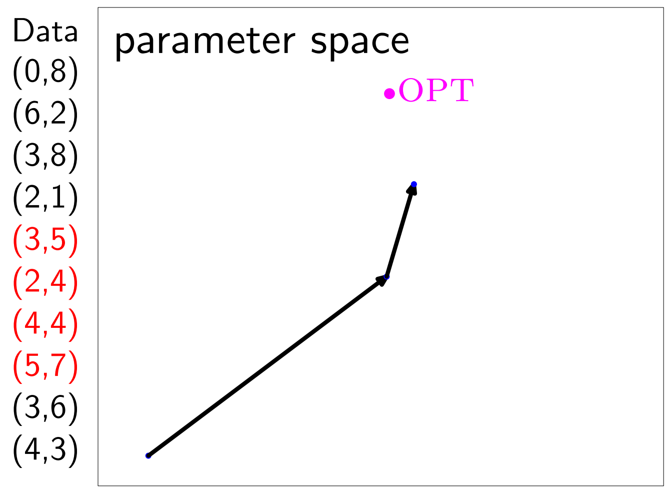
$m =$ size of a mini-batch

$m = 1 \longleftrightarrow m = n$

frequent updates, unstable — infrequent updates, stable

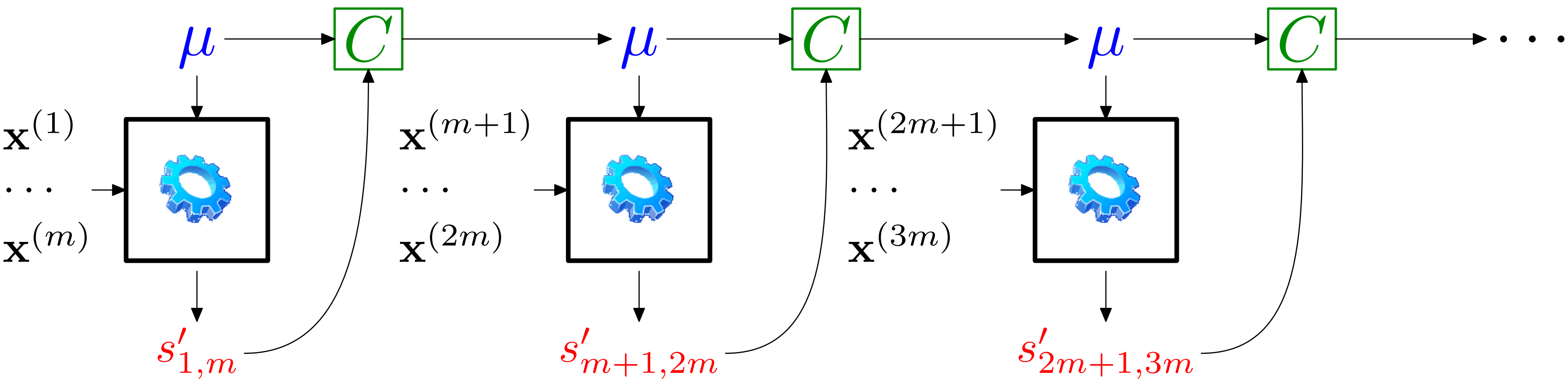


10 data points processed



8 data points processed

Optimization parameter 2 of 2: minibatch size

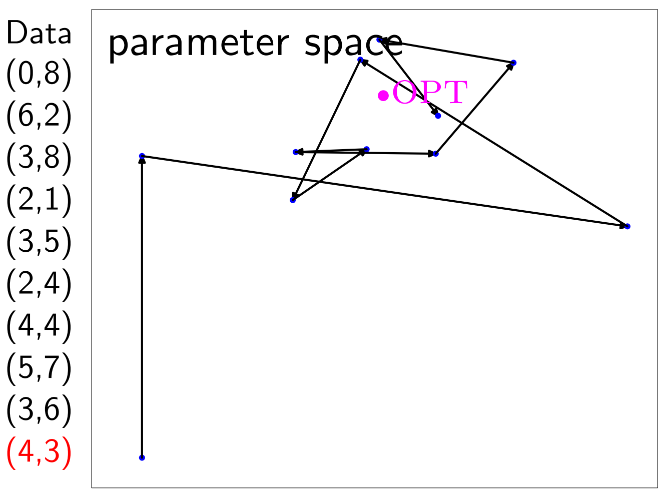


$m =$ size of a mini-batch

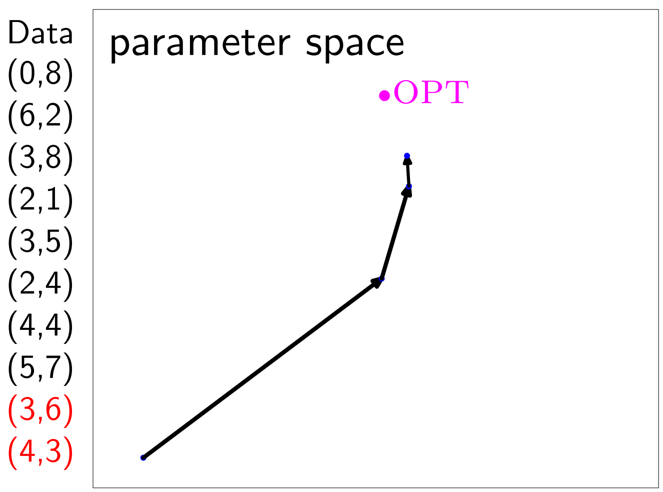
$m = 1 \longleftrightarrow m = n$

frequent updates, unstable

infrequent updates, stable

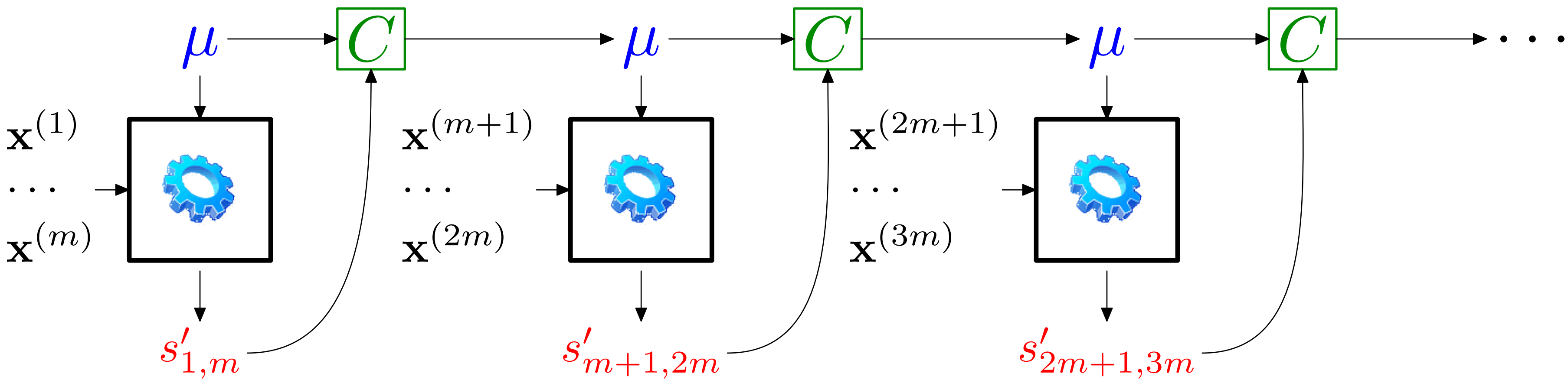


10 data points processed



10 data points processed

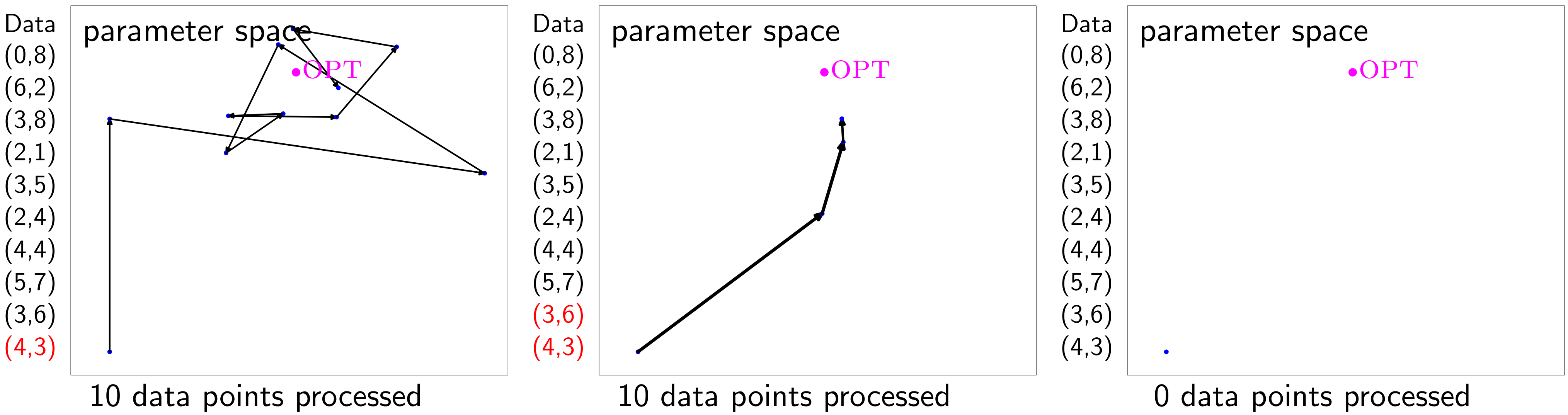
Optimization parameter 2 of 2: minibatch size



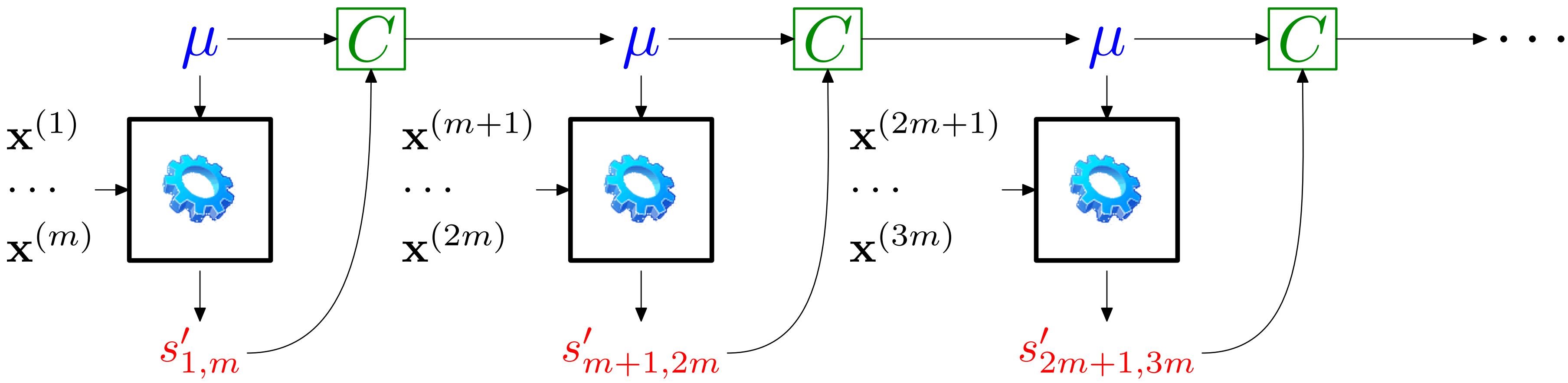
m = size of a mini-batch

$m = 1 \longleftrightarrow m = n$

frequent updates, unstable — infrequent updates, stable



Optimization parameter 2 of 2: minibatch size

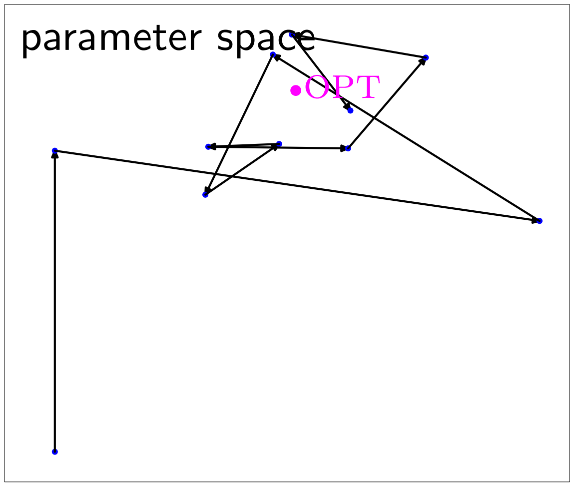


$m =$ size of a mini-batch

$m = 1 \longleftrightarrow m = n$

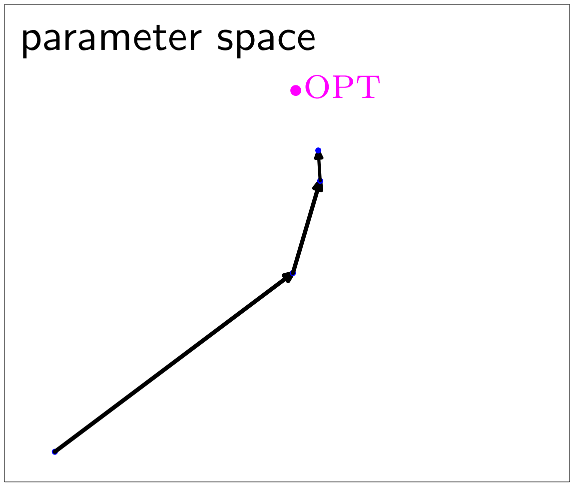
frequent updates, unstable — infrequent updates, stable

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



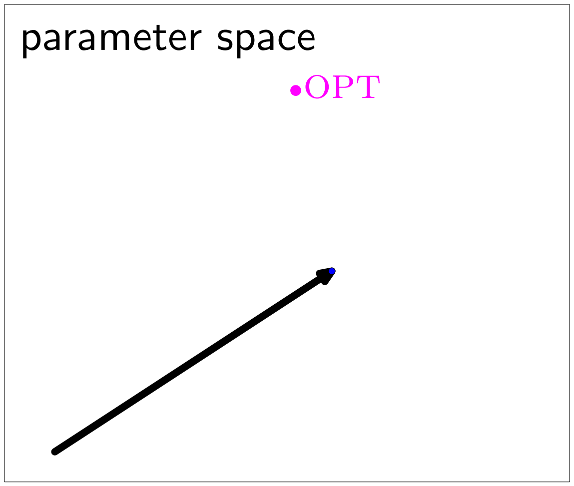
10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



10 data points processed

Data
(0,8)
(6,2)
(3,8)
(2,1)
(3,5)
(2,4)
(4,4)
(5,7)
(3,6)
(4,3)



10 data points processed

Setting optimization parameters

stepsize reduction power α

mini-batch size m

Setting optimization parameters

stepsize reduction power α mini-batch size m

Document classification:

[Likelihood]

$\alpha \backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	-8.875	-8.710	-8.610	-8.555	-8.505	-8.172	-7.920	-7.906	-7.916
0.6	-8.604	-8.575	-8.540	-8.524	-8.235	-8.041	-7.898	-7.901	-7.916
0.7	-8.541	-8.533	-8.531	-8.354	-8.023	-7.943	-7.886	-7.896	-7.918
0.8	-8.519	-8.506	-8.493	-8.228	-7.933	-7.896	-7.883	-7.890	-7.922
0.9	-8.505	-8.486	-8.283	-8.106	-7.910	-7.889	-7.889	-7.891	-7.927
1.0	-8.471	-8.319	-8.204	-8.052	-7.919	-7.889	-7.892	-7.896	-7.937

Setting optimization parameters

stepsize reduction power α mini-batch size m

Document classification:

[Likelihood]

$\alpha \backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	-8.875	-8.710	-8.610	-8.555	-8.505	-8.172	-7.920	-7.906	-7.916
0.6	-8.604	-8.575	-8.540	-8.524	-8.235	-8.041	-7.898	-7.901	-7.916
0.7	-8.541	-8.533	-8.531	-8.354	-8.023	-7.943	-7.886	-7.896	-7.918
0.8	-8.519	-8.506	-8.493	-8.228	-7.933	-7.896	**-7.883**	-7.890	-7.922
0.9	-8.505	-8.486	-8.283	-8.106	-7.910	-7.889	-7.889	-7.891	-7.927
1.0	-8.471	-8.319	-8.204	-8.052	-7.919	-7.889	-7.892	-7.896	-7.937

Setting optimization parameters

stepsize reduction power α mini-batch size m

Document classification:

[Likelihood]

$\alpha\backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	-8.875	-8.710	-8.610	-8.555	-8.505	-8.172	-7.920	-7.906	-7.916
0.6	-8.604	-8.575	-8.540	-8.524	-8.235	-8.041	-7.898	-7.901	-7.916
0.7	-8.541	-8.533	-8.531	-8.354	-8.023	-7.943	-7.886	-7.896	-7.918
0.8	-8.519	-8.506	-8.493	-8.228	-7.933	-7.896	**-7.883**	-7.890	-7.922
0.9	-8.505	-8.486	-8.283	-8.106	-7.910	-7.889	-7.889	-7.891	-7.927
1.0	-8.471	-8.319	-8.204	-8.052	-7.919	-7.889	-7.892	-7.896	-7.937

[Accuracy]

$\alpha\backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	5.4	5.4	5.5	5.6	6.0	25.7	48.8	49.9	44.6
0.6	5.4	5.4	5.6	5.6	22.3	36.1	48.7	49.3	44.2
0.7	5.5	5.5	5.6	11.1	39.9	43.3	48.1	49.0	43.5
0.8	5.6	5.6	6.0	21.7	47.3	45.0	**47.8**	49.5	42.8
0.9	5.8	6.0	13.4	32.4	48.7	48.4	46.4	49.4	42.4
1.0	6.2	11.8	19.6	35.2	47.6	49.5	47.5	49.3	41.7

Setting optimization parameters

stepsize reduction power α mini-batch size m

Document classification:

[Likelihood]

$\alpha \backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	-8.875	-8.710	-8.610	-8.555	-8.505	-8.172	-7.920	-7.906	-7.916
0.6	-8.604	-8.575	-8.540	-8.524	-8.235	-8.041	-7.898	-7.901	-7.916
0.7	-8.541	-8.533	-8.531	-8.354	-8.023	-7.943	-7.886	-7.896	-7.918
0.8	-8.519	-8.506	-8.493	-8.228	-7.933	-7.896	**-7.883**	-7.890	-7.922
0.9	-8.505	-8.486	-8.283	-8.106	-7.910	-7.889	-7.889	-7.891	-7.927
1.0	-8.471	-8.319	-8.204	-8.052	-7.919	-7.889	-7.892	-7.896	-7.937

[Accuracy]

$\alpha \backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	5.4	5.4	5.5	5.6	6.0	25.7	48.8	**49.9**	44.6
0.6	5.4	5.4	5.6	5.6	22.3	36.1	48.7	49.3	44.2
0.7	5.5	5.5	5.6	11.1	39.9	43.3	48.1	49.0	43.5
0.8	5.6	5.6	6.0	21.7	47.3	45.0	**47.8**	49.5	42.8
0.9	5.8	6.0	13.4	32.4	48.7	48.4	46.4	49.4	42.4
1.0	6.2	11.8	19.6	35.2	47.6	49.5	47.5	49.3	41.7

Setting optimization parameters

stepsize reduction power α mini-batch size m

Document classification:

[Likelihood]

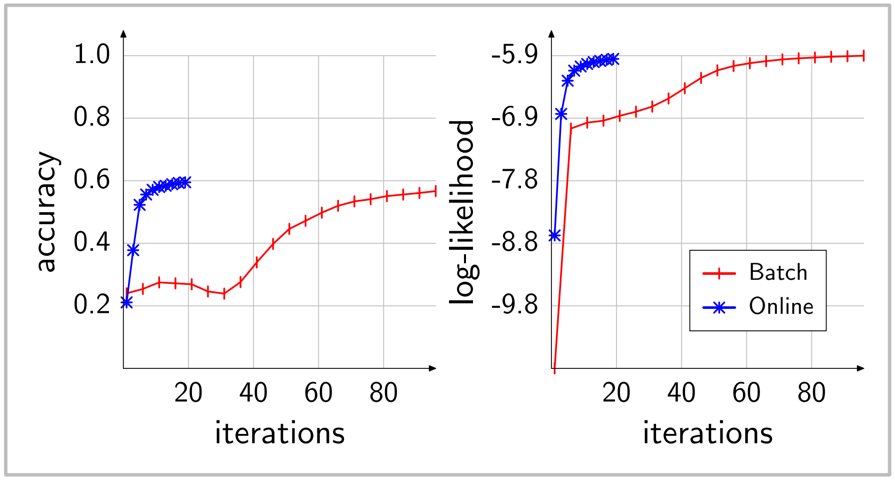
$\alpha \backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	-8.875	-8.710	-8.610	-8.555	-8.505	-8.172	-7.920	-7.906	-7.916
0.6	-8.604	-8.575	-8.540	-8.524	-8.235	-8.041	-7.898	-7.901	-7.916
0.7	-8.541	-8.533	-8.531	-8.354	-8.023	-7.943	-7.886	-7.896	-7.918
0.8	-8.519	-8.506	-8.493	-8.228	-7.933	-7.896	**-7.883**	-7.890	-7.922
0.9	-8.505	-8.486	-8.283	-8.106	-7.910	-7.889	-7.889	-7.891	-7.927
1.0	-8.471	-8.319	-8.204	-8.052	-7.919	-7.889	-7.892	-7.896	-7.937

[Accuracy]

$\alpha \backslash m$	1	3	10	30	100	300	1K	3K	10K
0.5	5.4	5.4	5.5	5.6	6.0	25.7	48.8	**49.9**	44.6
0.6	5.4	5.4	5.6	5.6	22.3	36.1	48.7	49.3	44.2
0.7	5.5	5.5	5.6	11.1	39.9	43.3	48.1	49.0	43.5
0.8	5.6	5.6	6.0	21.7	47.3	45.0	**47.8**	49.5	42.8
0.9	5.8	6.0	13.4	32.4	48.7	48.4	46.4	49.4	42.4
1.0	6.2	11.8	19.6	35.2	47.6	49.5	47.5	49.3	41.7

(α, m) important, but can set using likelihood (unsupervised)

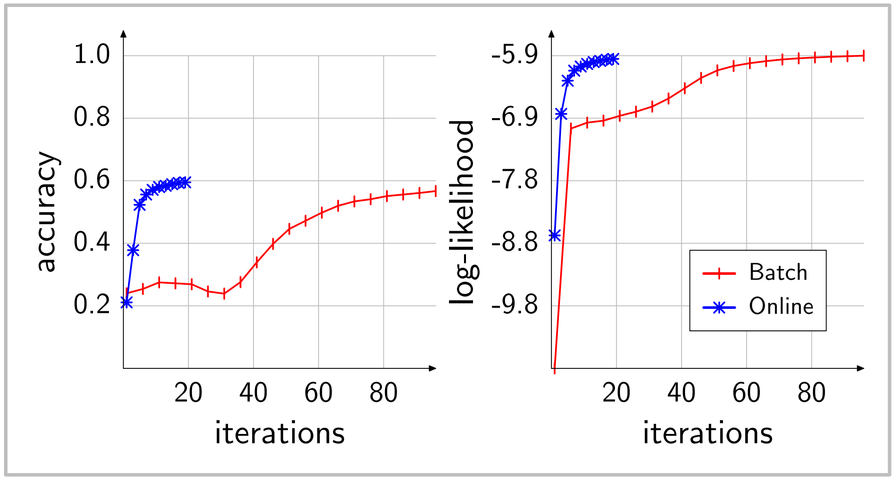
Results: speed



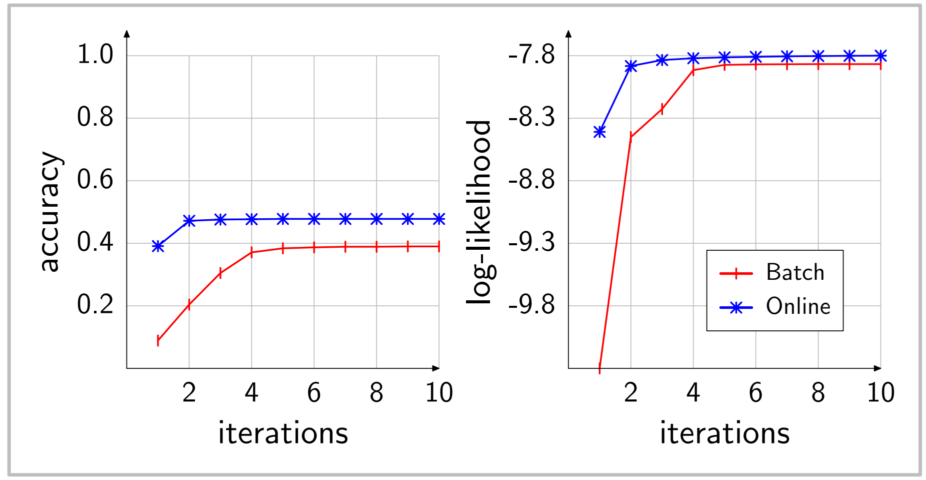
(a) POS tagging

Online converges faster than Batch

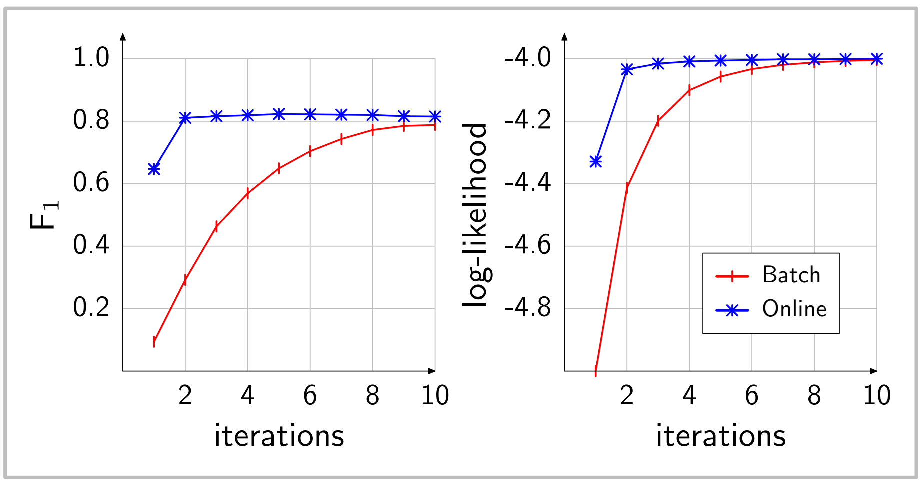
Results: speed



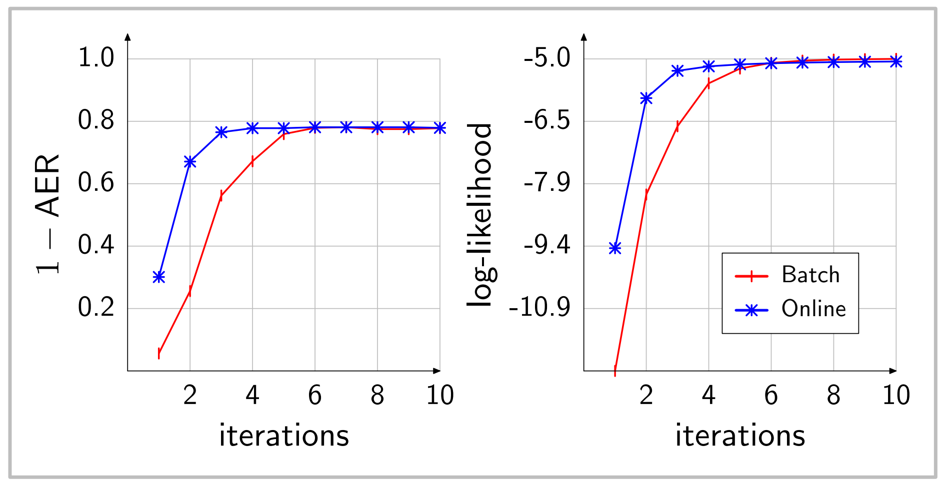
(a) POS tagging



(b) Document classification



(c) Word segmentation (English)



(d) Word alignment

Online converges faster than Batch

Results: final accuracy

	POS	DOC	SEG	ALIGN
Batch	57.3	39.1	80.5	78.8
Online	59.6	47.8	80.7	78.9

Online: choose (α, m) with highest likelihood

Results: final accuracy

	POS	DOC	SEG	ALIGN
Batch	57.3	39.1	80.5	78.8
Online	59.6	47.8	80.7	78.9
Online*	66.7	49.9	83.5	78.9

Online: choose (α, m) with highest likelihood

Online*: choose (α, m) with highest accuracy

Results: final accuracy

	POS	DOC	SEG	ALIGN
Batch	57.3	39.1	80.5	78.8
Online	59.6	47.8	80.7	78.9
Online*	66.7	49.9	83.5	78.9

Online: choose (α, m) with highest likelihood

Online*: choose (α, m) with highest accuracy

Mystery:

- Online EM obtains higher accuracy

Results: final accuracy

	POS	DOC	SEG	ALIGN
Batch	57.3	39.1	80.5	78.8
Online	59.6	47.8	80.7	78.9
Online*	66.7	49.9	83.5	78.9

Online: choose (α, m) with highest likelihood

Online*: choose (α, m) with highest accuracy

Mystery:

- Online EM obtains higher accuracy
- Batch EM and online EM optimize same objective function

Optimization intuitions

Two parts of optimizing non-convex objectives:

(1) Find a good peak

(2) Climb to the top of that peak

Optimization intuitions

Two parts of optimizing non-convex objectives:

(1) Find a good peak

(2) Climb to the top of that peak

Hypothesis:

Accuracy affected by (1), likelihood affected by (1) and (2)

Optimization intuitions

Two parts of optimizing non-convex objectives:

(1) Find a good peak

(2) Climb to the top of that peak

Hypothesis:

Accuracy affected by (1), likelihood affected by (1) and (2)

Batch EM does (2) well, online EM does (1) well

Optimization intuitions

Two parts of optimizing non-convex objectives:

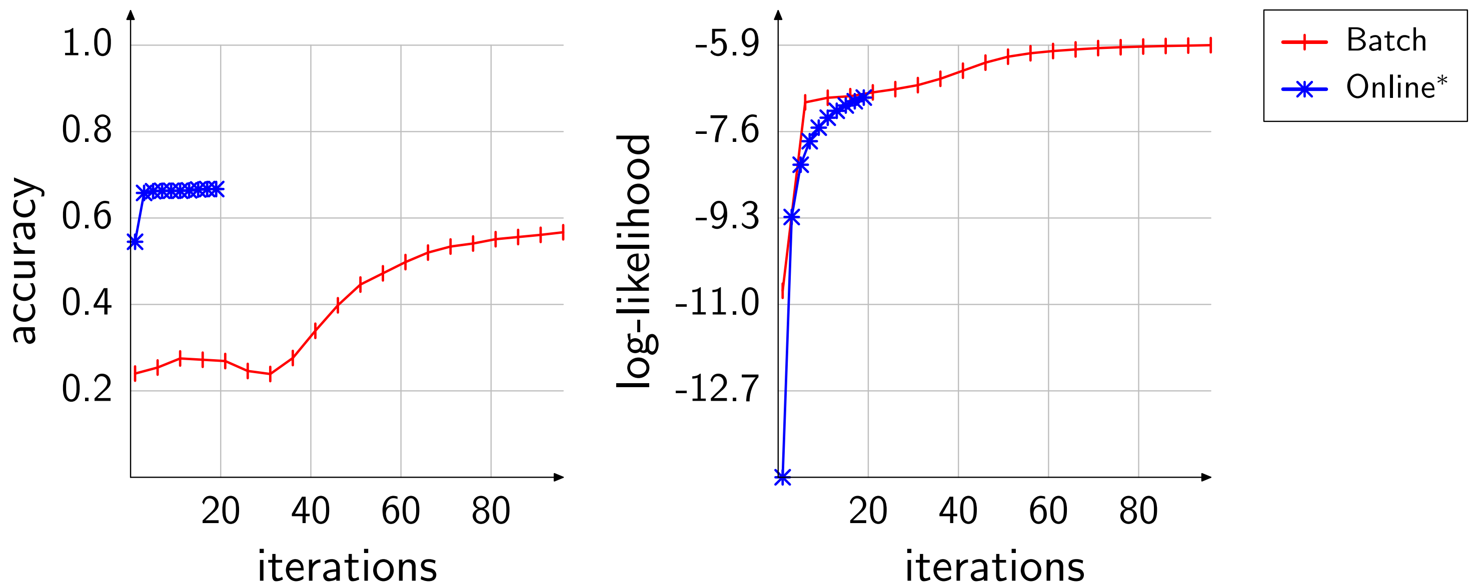
(1) Find a good peak

(2) Climb to the top of that peak

Hypothesis:

Accuracy affected by (1), likelihood affected by (1) and (2)

Batch EM does (2) well, online EM does (1) well



Optimization intuitions

Two parts of optimizing non-convex objectives:

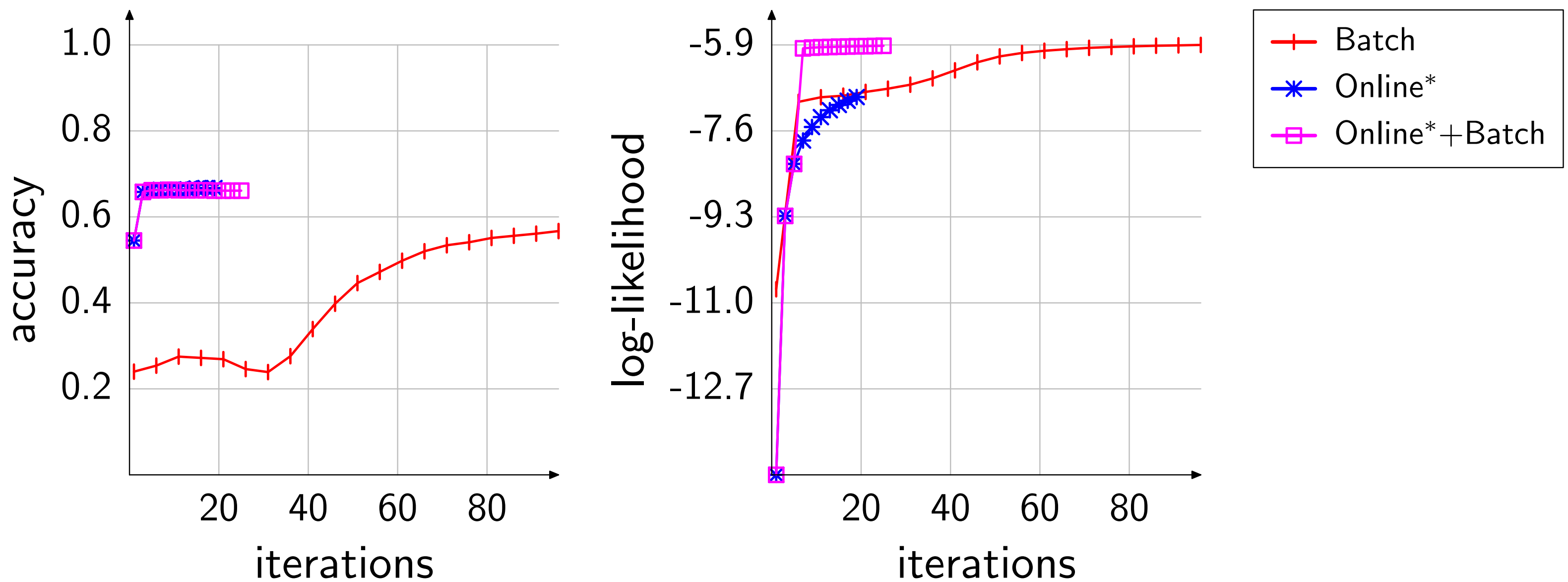
(1) Find a good peak

(2) Climb to the top of that peak

Hypothesis:

Accuracy affected by (1), likelihood affected by (1) and (2)

Batch EM does (2) well, online EM does (1) well



Online*+Batch: 5 iterations of Online* then Batch

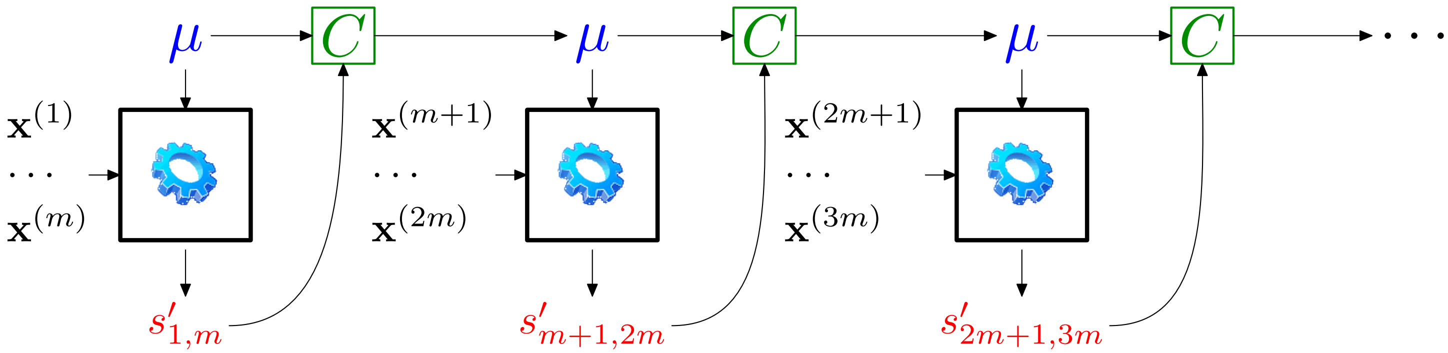
Summary

Goal: fast unsupervised learning

Summary

Goal: fast unsupervised learning

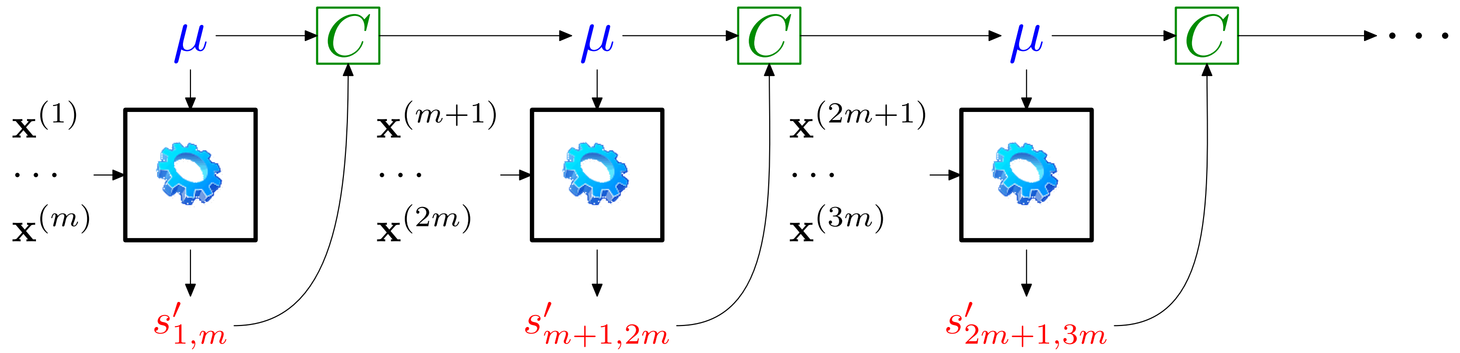
Online EM: update parameters more often



Summary

Goal: fast unsupervised learning

Online EM: update parameters more often



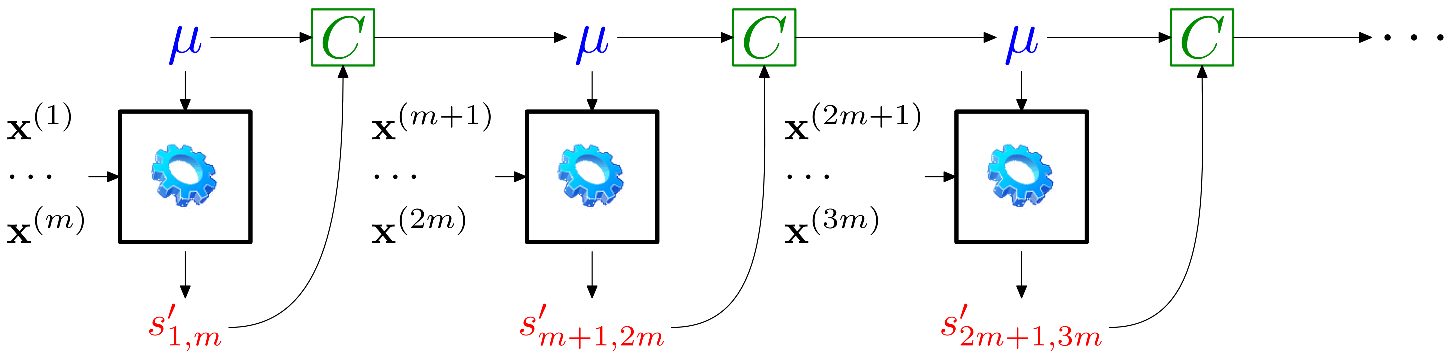
fast, unstable — slow, stable

Stepsize and minibatches balance this tradeoff

Summary

Goal: fast unsupervised learning

Online EM: update parameters more often



fast, unstable $\longleftrightarrow$ slow, stable

Stepsize and minibatches balance this tradeoff

Result: online EM is faster,
and sometimes more accurate than batch EM