Loss is more

Exploring the network weight space to harvest information on its training dataset

ALPACA

Margherita Mele

Motivation

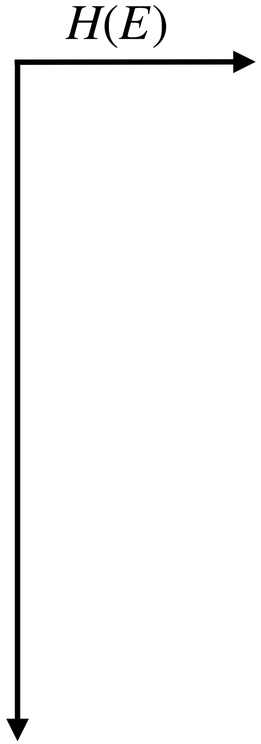
How does learning take place?

Motivation

How does learning take place?

- Exhaustively **explore** the energy profile of learning problems
- **Identify** the configurations minimising energy
- **Recover** the thermodynamic properties of the entire system
- **Study** real learning problems
- **Characterise** the impact of the input-output correlation in learning problems

Wang-Landau algorithm



Proposal

$$g(w, \tilde{w})$$

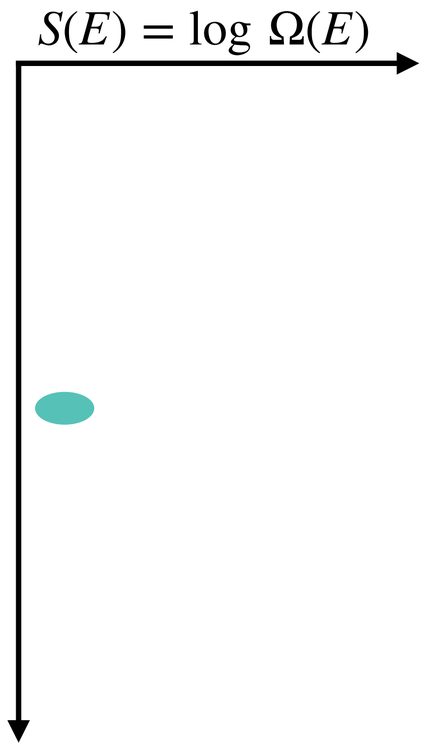
Acceptance

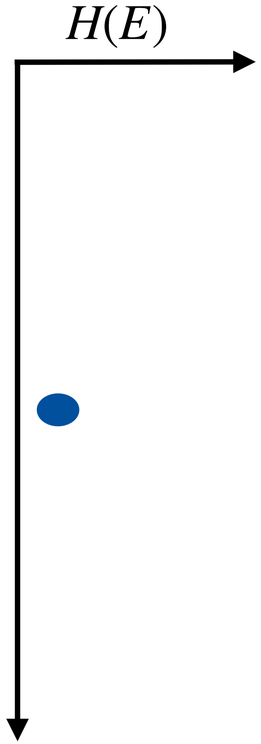
$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm





Proposal

$$g(w, \tilde{w})$$

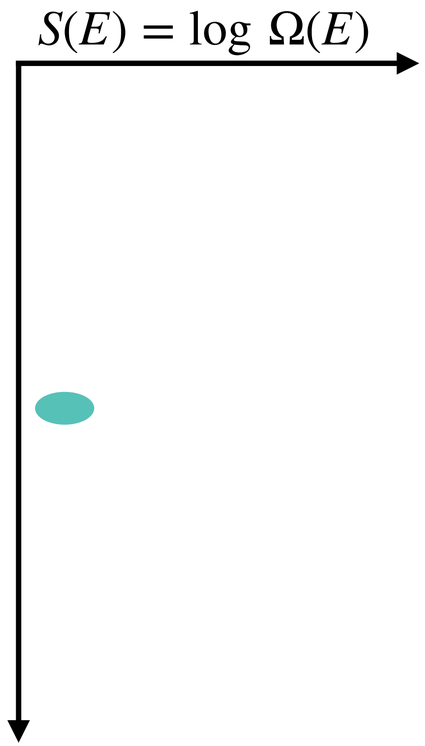
Acceptance

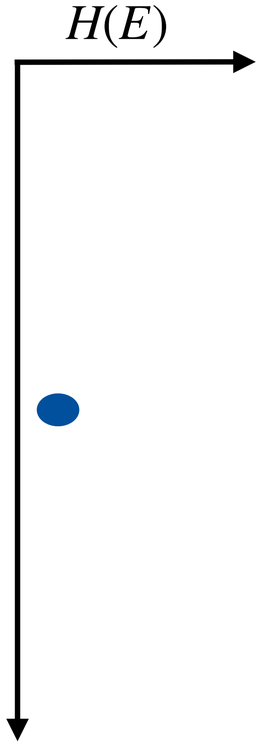
$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm





Proposal

$$g(w, \tilde{w})$$

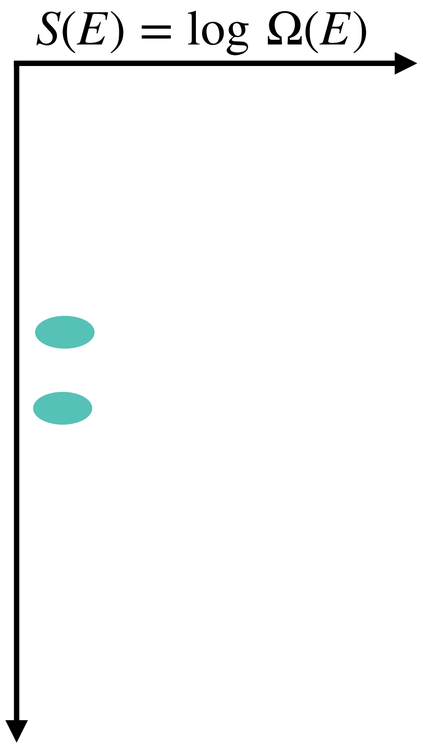
Acceptance

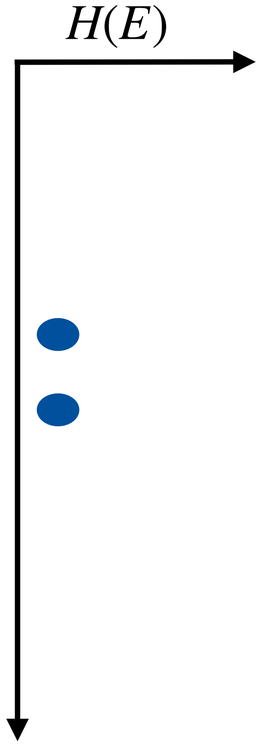
$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm





Proposal

$$g(w, \tilde{w})$$

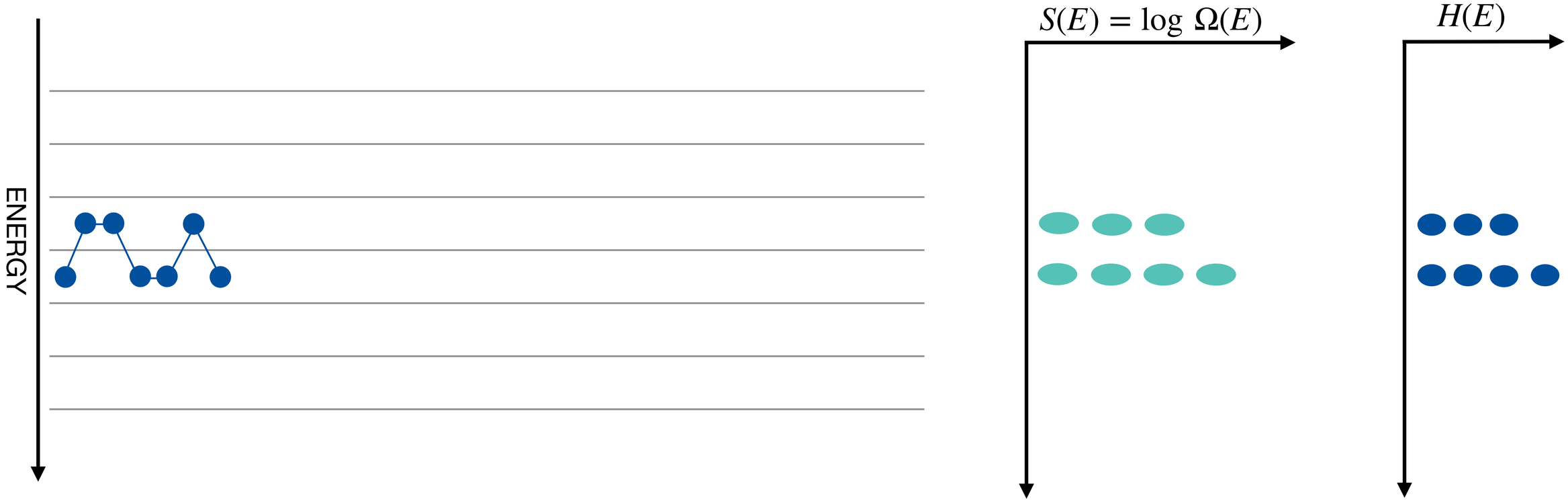
Acceptance

$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm



Proposal

$$g(w, \tilde{w})$$

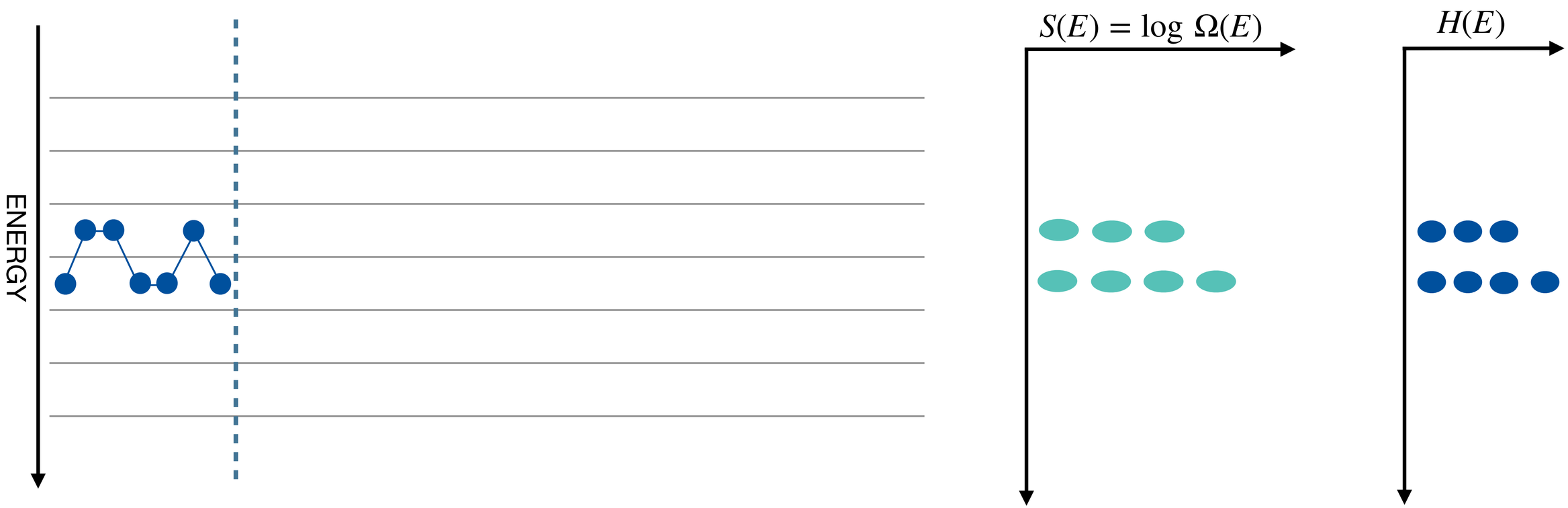
Acceptance

$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm



Proposal

$$g(w, \tilde{w})$$

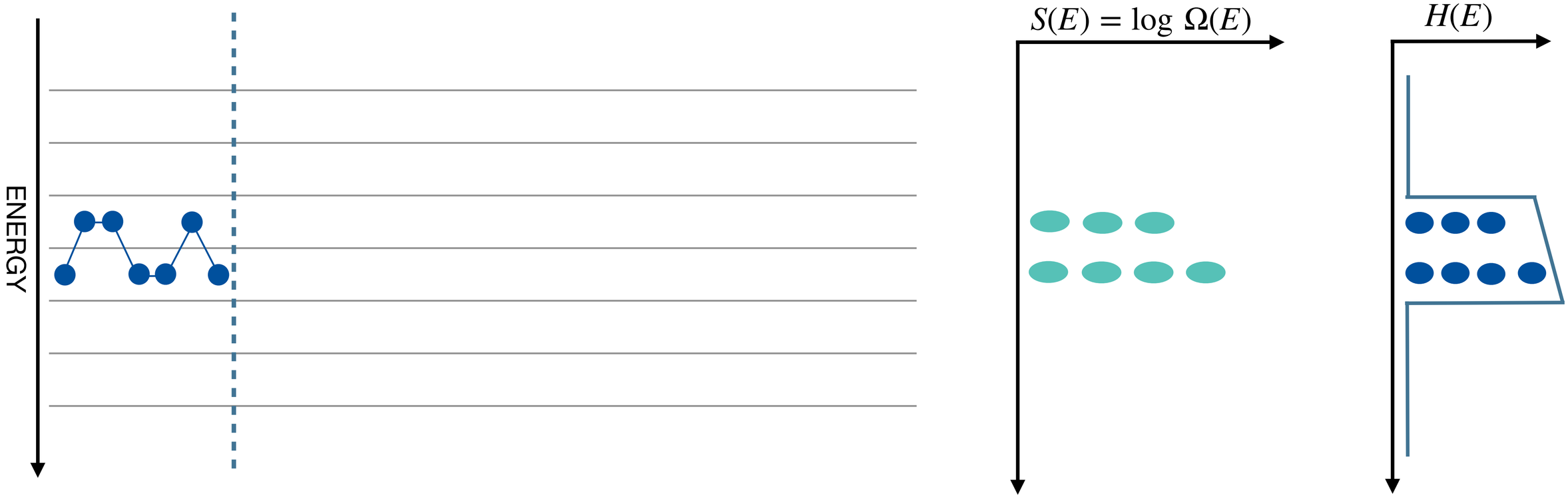
Acceptance

$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm



Proposal

$$g(w, \tilde{w})$$

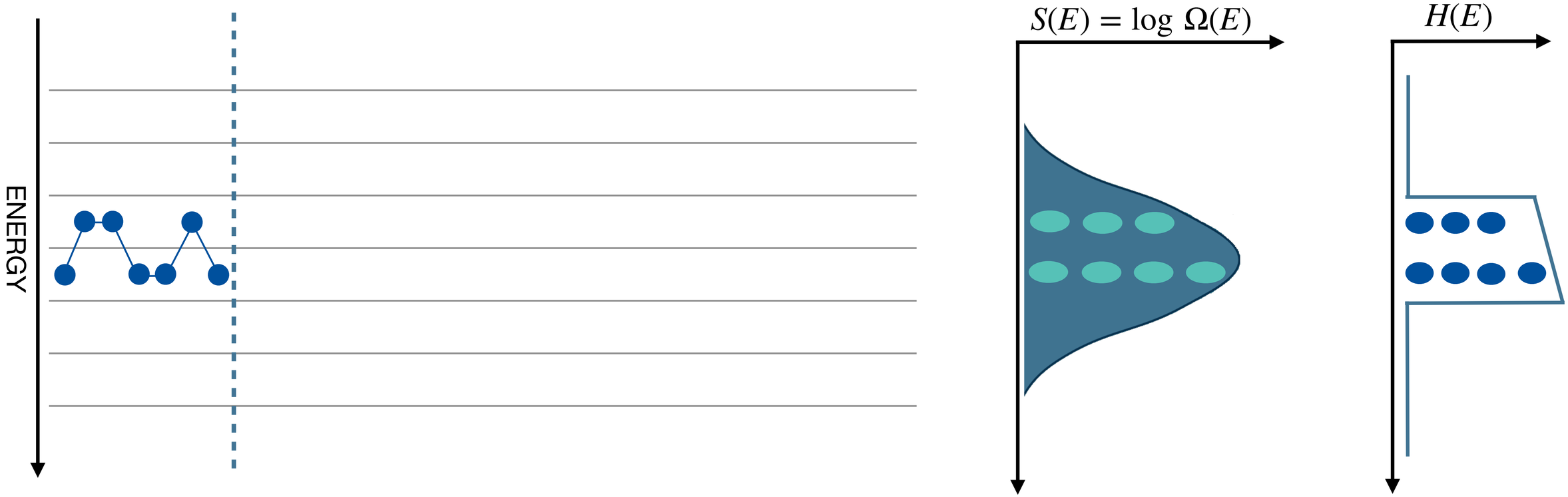
Acceptance

$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm



Proposal

$$g(w, \tilde{w})$$

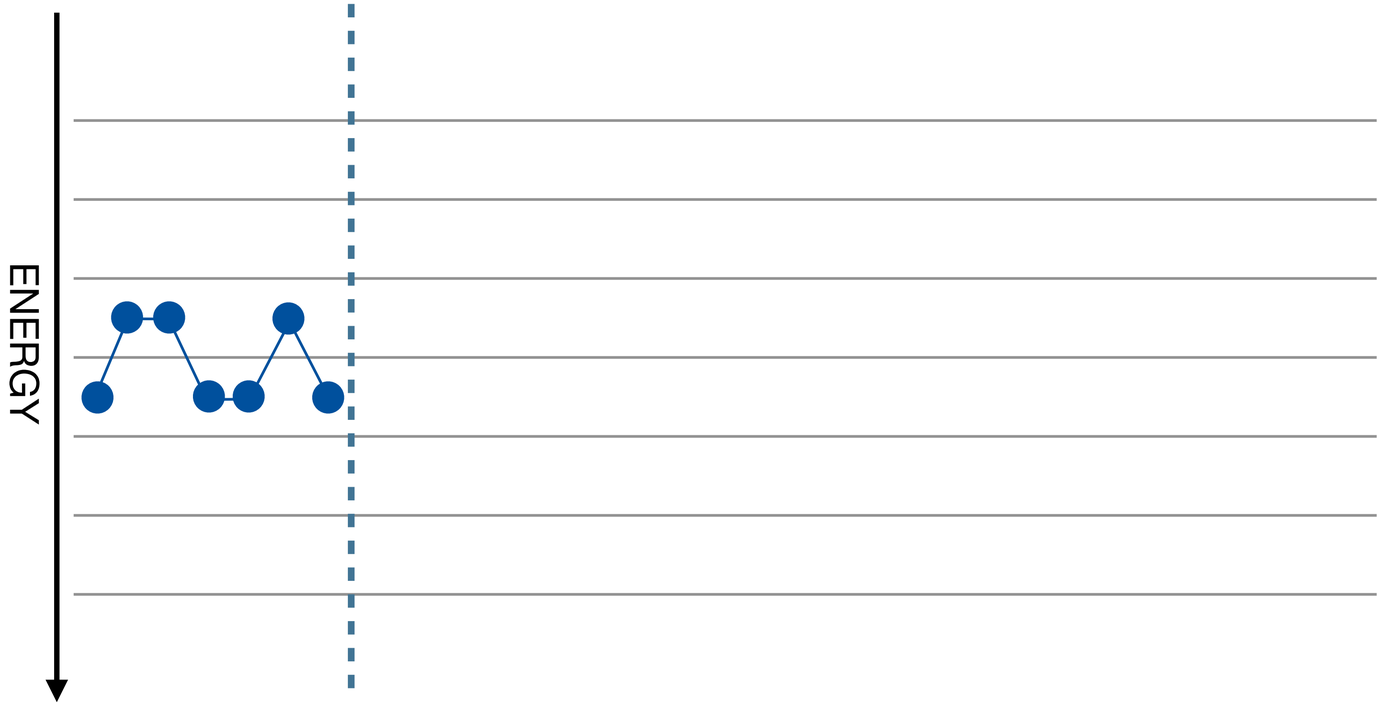
Acceptance

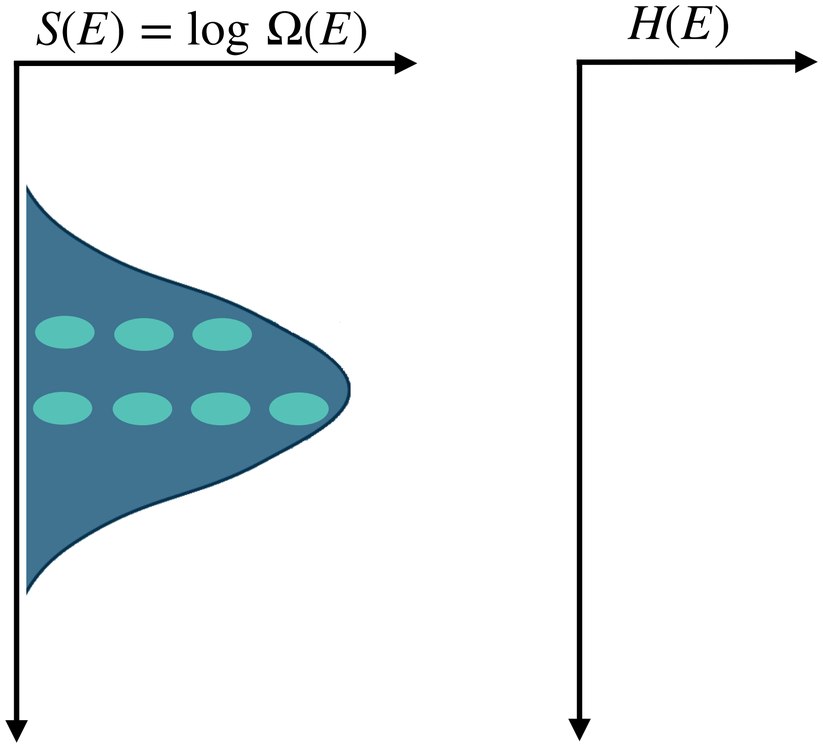
$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm





Proposal

$$g(w, \tilde{w})$$

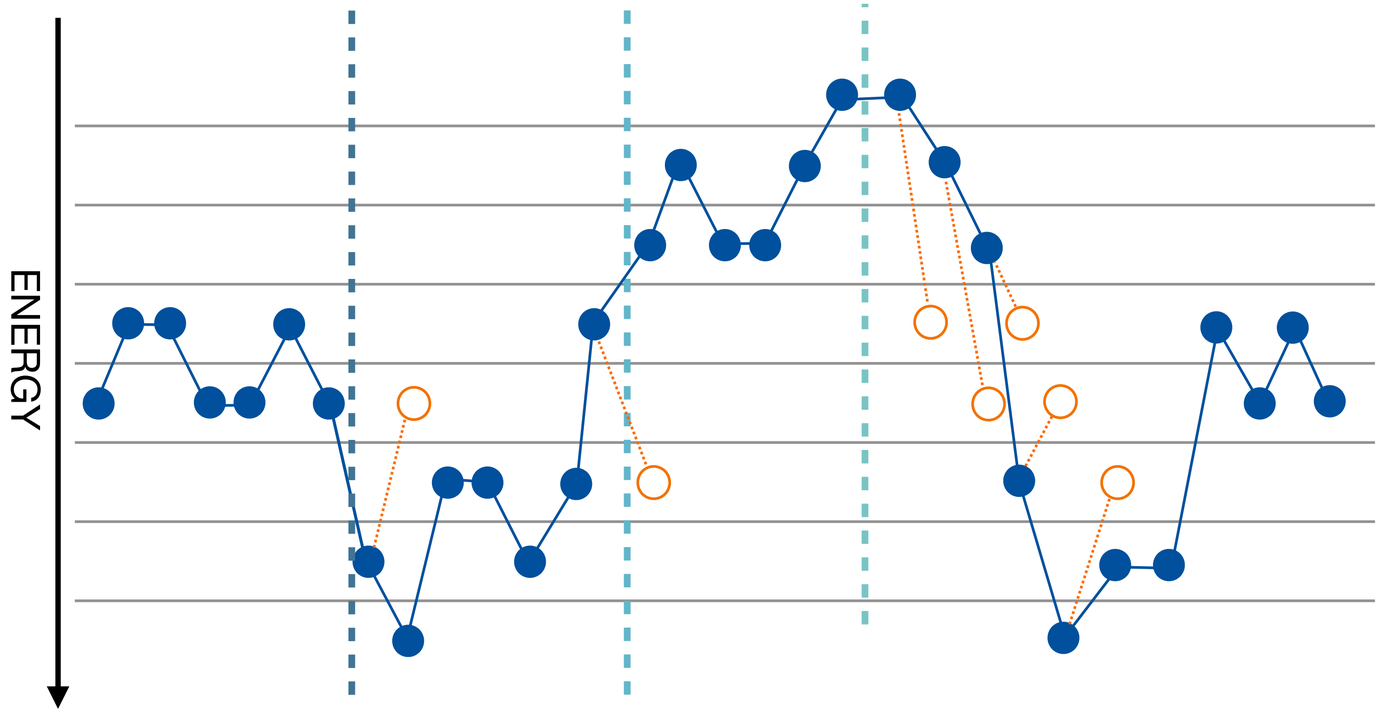
Acceptance

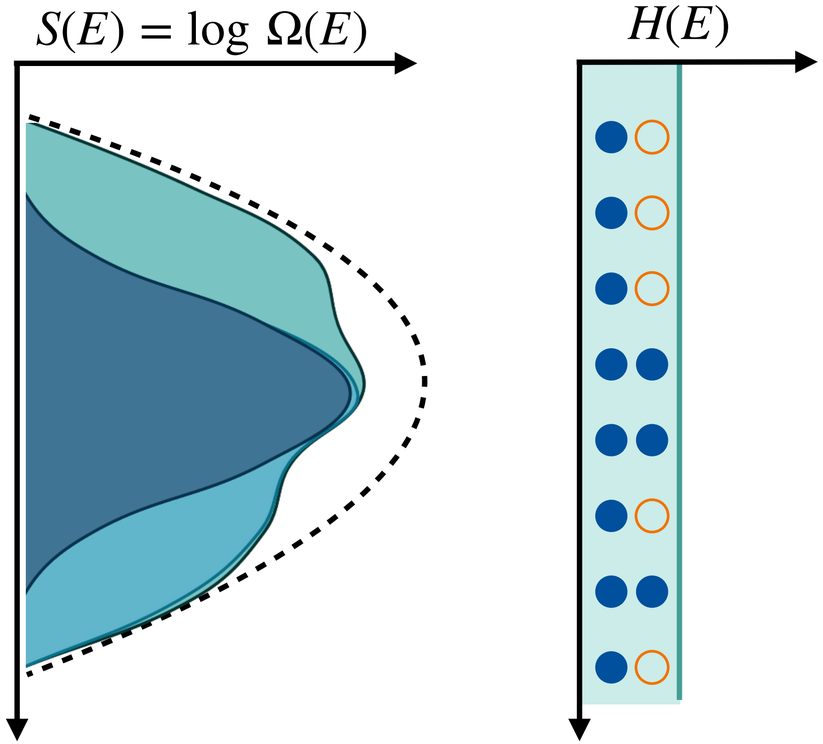
$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm





Proposal

$$g(w, \tilde{w})$$

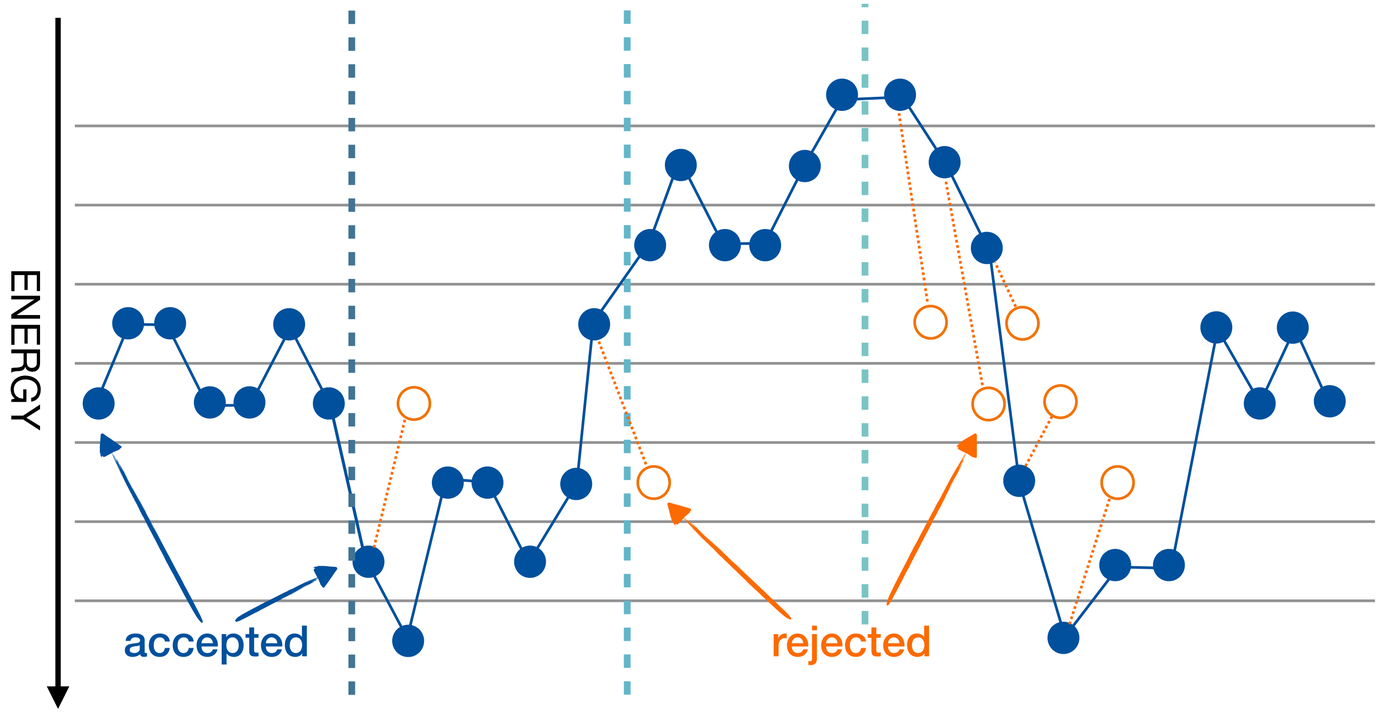
Acceptance

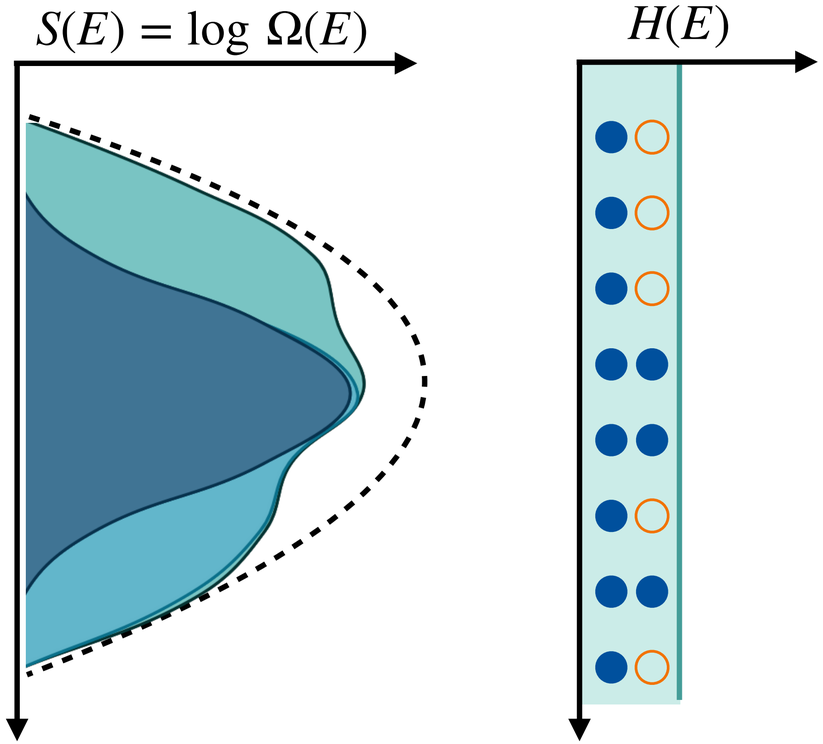
$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm





Proposal

$$g(w, \tilde{w})$$

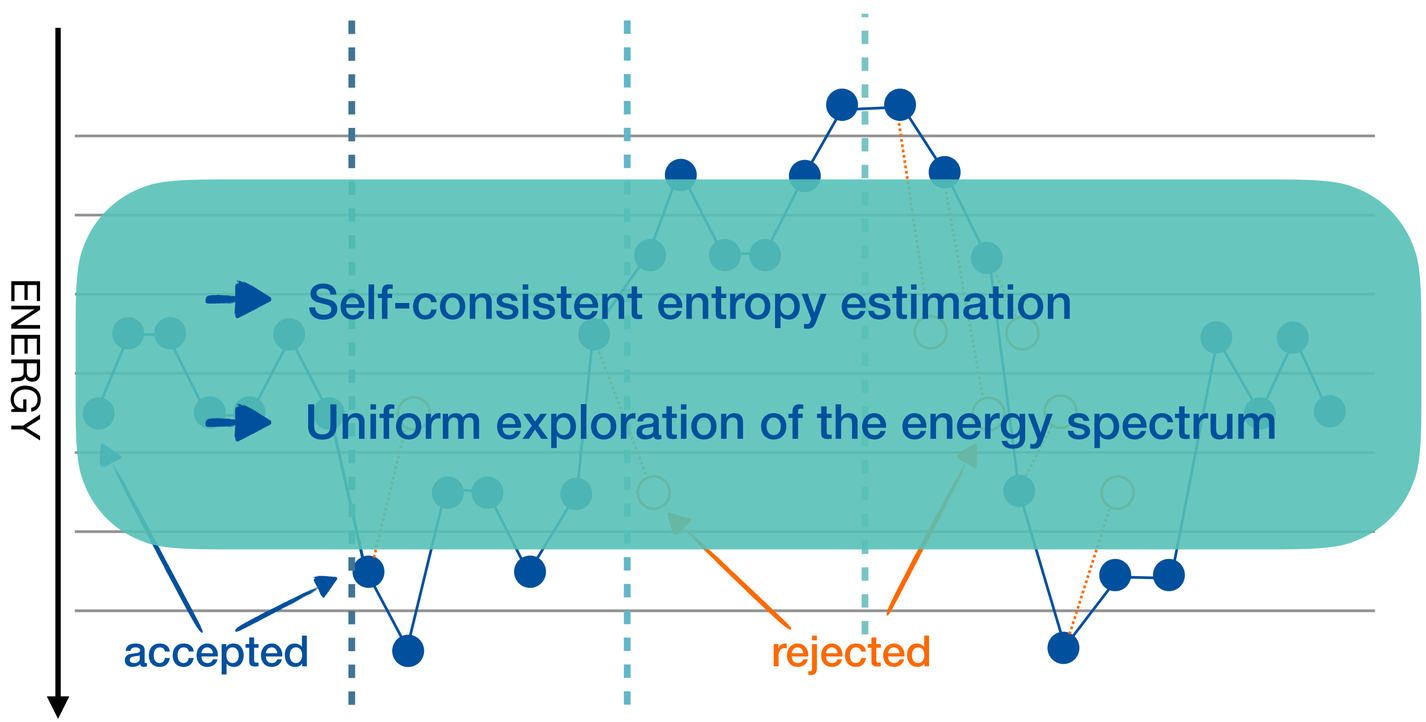
Acceptance

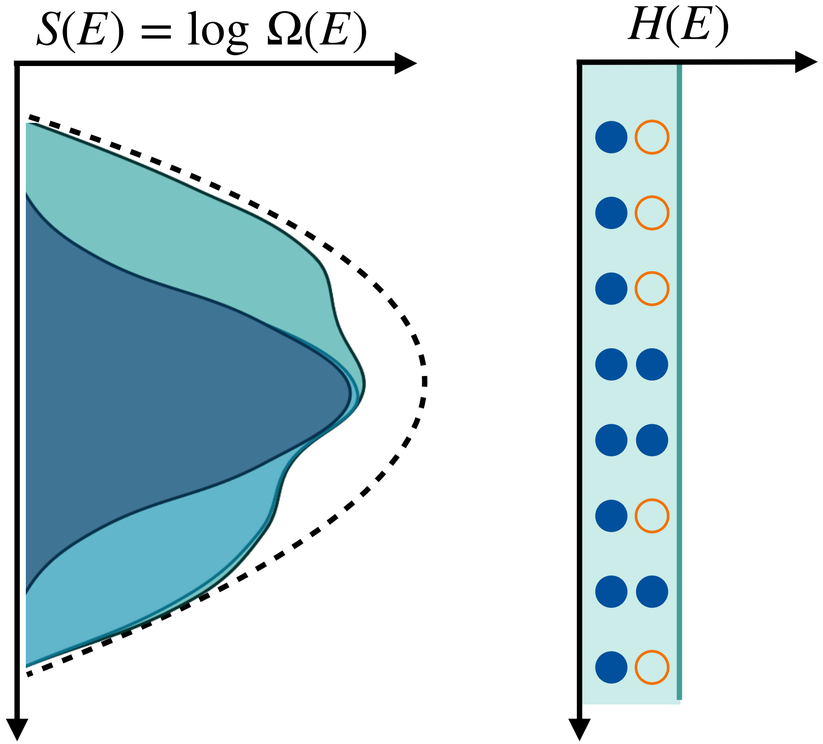
$$A(w, \tilde{w}) = \min \left\{ 1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})} \right\}$$

Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Wang-Landau algorithm





Proposal

$$g(w, \tilde{w})$$

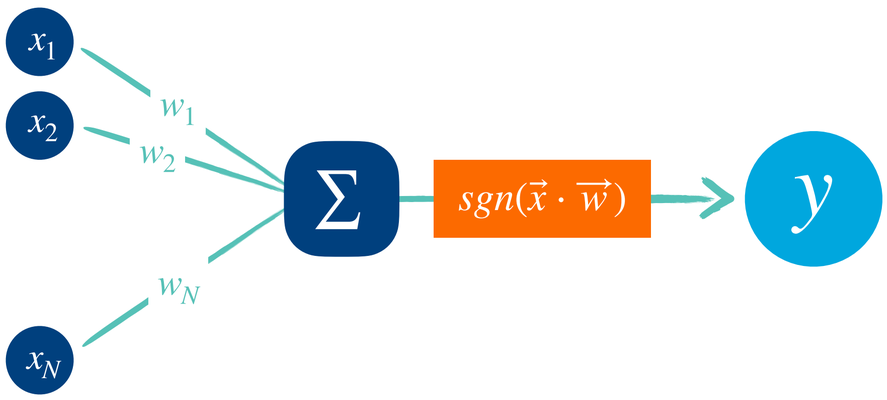
Acceptance

$$A(w, \tilde{w}) = \min\left\{1,\ e^{S_w - S_{\tilde{w}}} \frac{g(\tilde{w}, w)}{g(w, \tilde{w})}\right\}$$

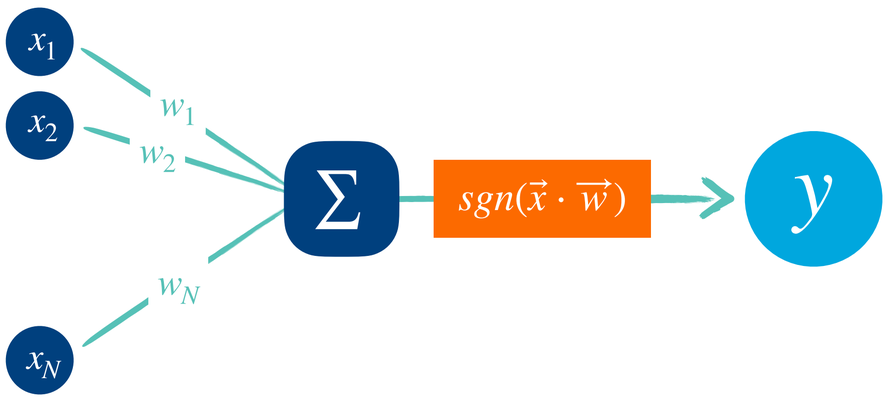
Updates

$$S_w \leftarrow S_w + f \qquad H_w \leftarrow H_w + 1$$

Validation



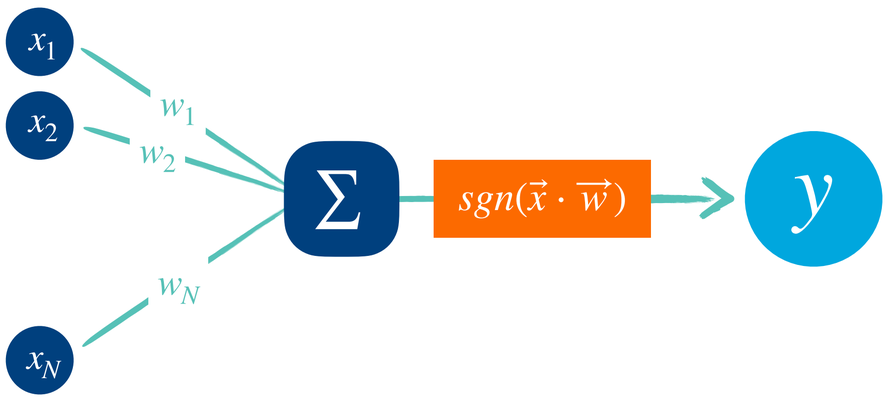
Validation



Random Inputs $\{x_i\}$

$$[1,1,-1,\ldots,1,-1,-1]$$
$$[-1,1,1,\ \ldots,-1,1,-1]$$
$$\vdots$$
$$[-1,-1,1,\ \ldots,-1,1,1]$$

Validation

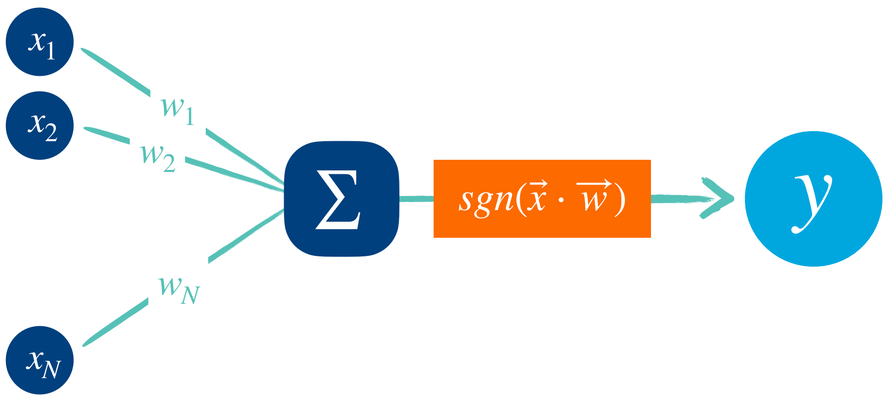


Random Inputs $\{x_i\}$

$$\begin{bmatrix} 1,1,-1,\ldots,1,-1,-1 \\ [-1,1,1,\ldots,-1,1,-1] \\ \vdots \\ [-1,-1,1,\ldots,-1,1,1] \end{bmatrix}$$

$\longleftarrow$ N $\longrightarrow$

Validation



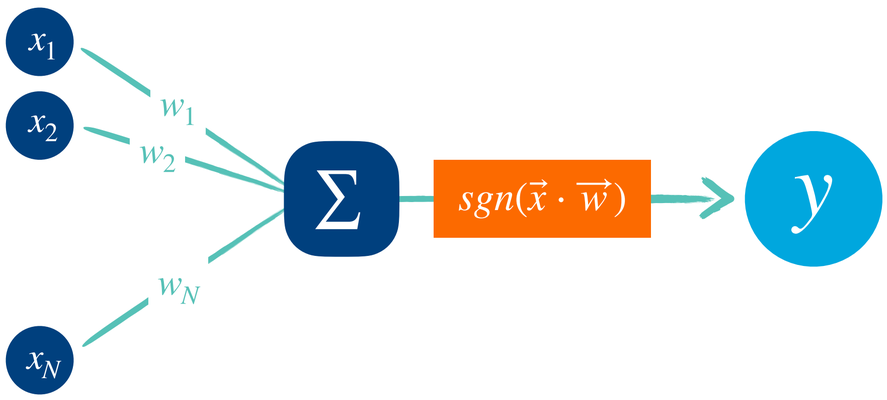
Random Inputs $\{x_i\}$

$$\begin{bmatrix} 1,1,-1,\ldots,1,-1,-1 \\ [-1,1,1,\ldots,-1,1,-1] \\ \vdots \\ [-1,-1,1,\ldots,-1,1,1] \end{bmatrix}$$

P

N

Validation



Random Inputs $\{x_i\}$

$$\begin{bmatrix} 1,1,-1,\ldots,1,-1,-1 \\ -1,1,1,\ldots,-1,1,-1 \\ \vdots \\ -1,-1,1,\ldots,-1,1,1 \end{bmatrix}$$

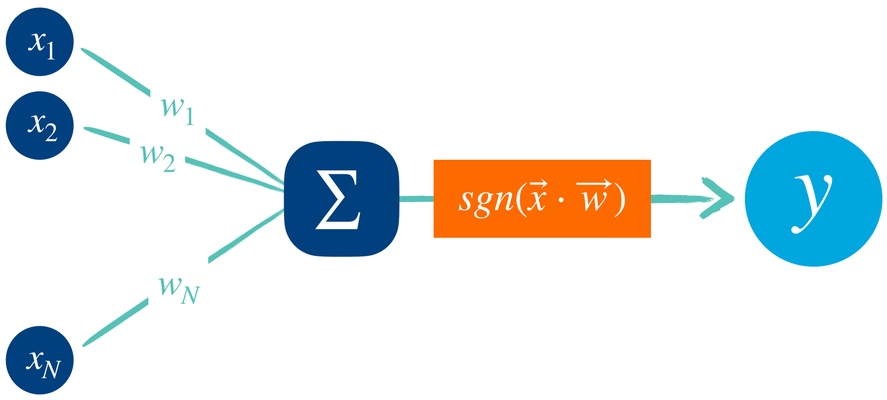
P

N

Teacher vector $\overline{w}$

$$[-1,1,-1\ \ldots,-1,-1,1]$$

Validation



Random Inputs $\{x_i\}$

$$\begin{bmatrix} 1,1,-1,\ldots,1,-1,-1 \\ [-1,1,1,\ \ldots,-1,1,-1] \\ \vdots \\ [-1,-1,1,\ \ldots,-1,1,1] \end{bmatrix}$$

P

N

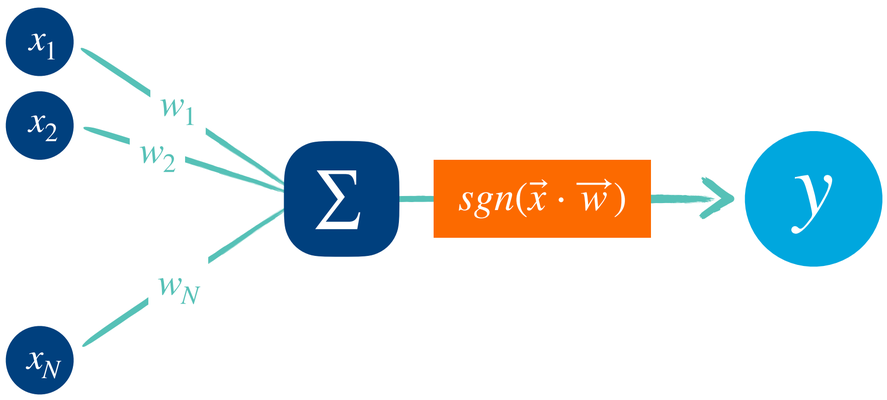
Teacher vector $\overline{w}$

$$[-1,1,-1\ \ldots,-1,-1,1]$$

Labels $\{y_i\}$

$$y_i = \mathrm{sgn}\left(\overline{w}\, x_i\right)$$

Validation



Random Inputs $\{x_i\}$

$$\begin{bmatrix} 1,1,-1,\ldots,1,-1,-1 \\ -1,1,1,\ldots,-1,1,-1 \\ \vdots \\ -1,-1,1,\ldots,-1,1,1 \end{bmatrix}$$

P

N

Teacher vector $\overline{w}$

$$[-1,1,-1\ \ldots,-1,-1,1]$$

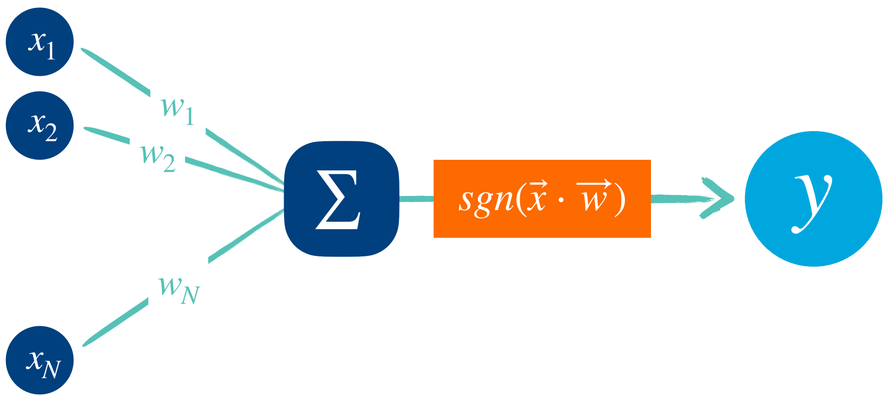
Labels $\{y_i\}$

$$y_i = \mathrm{sgn}\left(\overline{w}\, x_i\right)$$

Energy

$$E_w = \sum_i^P \Theta\left(-y_i \cdot \mathrm{sgn}\left(w\, x_i\right)\right)$$

Validation



Random Inputs $\{x_i\}$

$$\begin{array}{c} [1,1,-1,\ldots,1,-1,-1] \\ [-1,1,1,\ldots,-1,1,-1] \\ \vdots \\ [-1,-1,1,\ldots,-1,1,1] \end{array}$$

P

N

Teacher vector $\overline{w}$

$$[-1,1,-1\ \ldots,-1,-1,1]$$

Labels $\{y_i\}$

$$y_i = \mathrm{sgn}\left(\overline{w}\, x_i\right)$$

Energy

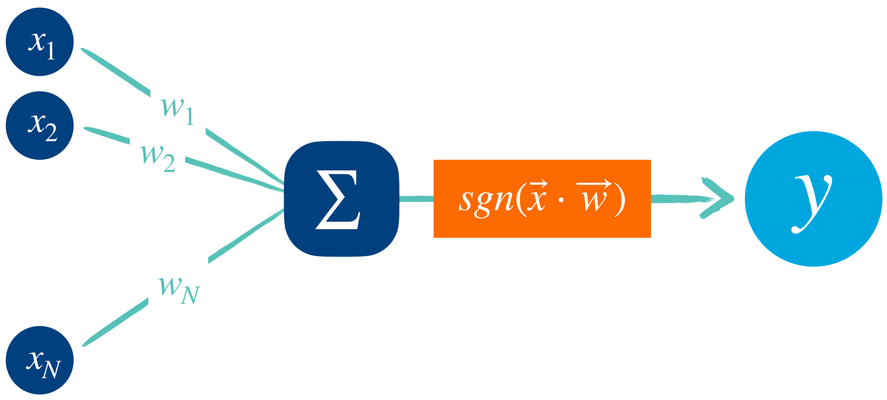
$$E_w = \sum_i^P \Theta\left(-y_i \cdot \mathrm{sgn}\left(w\, x_i\right)\right)$$

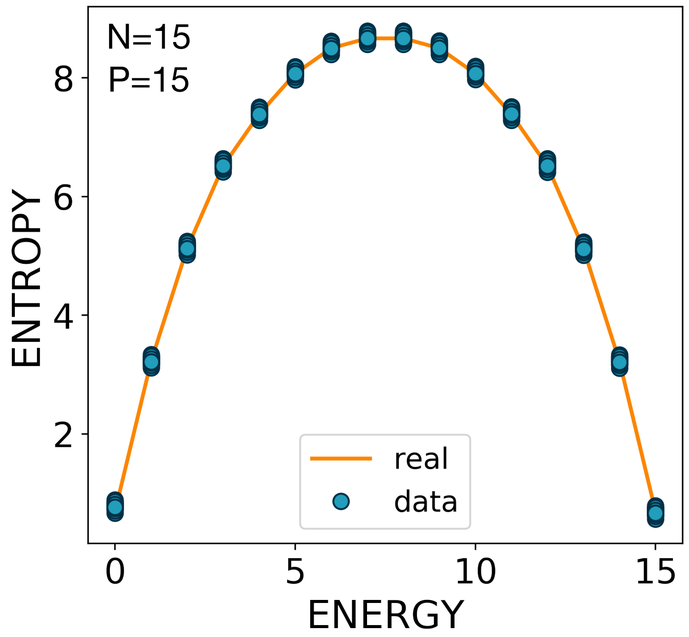
Entropy

$$S(\overline{E}) = \log\left(\sum_{\{w\}} \delta(E_w - \overline{E})\right)$$

Validation

100 realizations
on the same data-set





Random Inputs $\{x_i\}$

$$\begin{array}{c} [1,1,-1,\ldots,1,-1,-1] \\ [-1,1,1,\ \ldots,-1,1,-1] \\ \vdots \\ [-1,-1,1,\ \ldots,-1,1,1] \end{array}$$

P

N

Teacher vector $\overline{w}$

$$[-1,1,-1\ \ldots,-1,-1,1]$$

Labels $\{y_i\}$

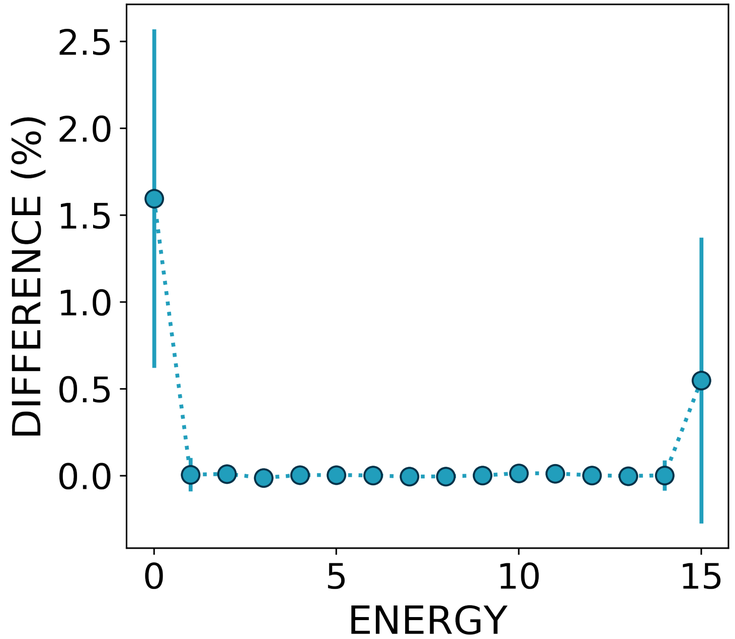
$$y_i = \mathrm{sgn}\left(\overline{w}\, x_i\right)$$

Energy

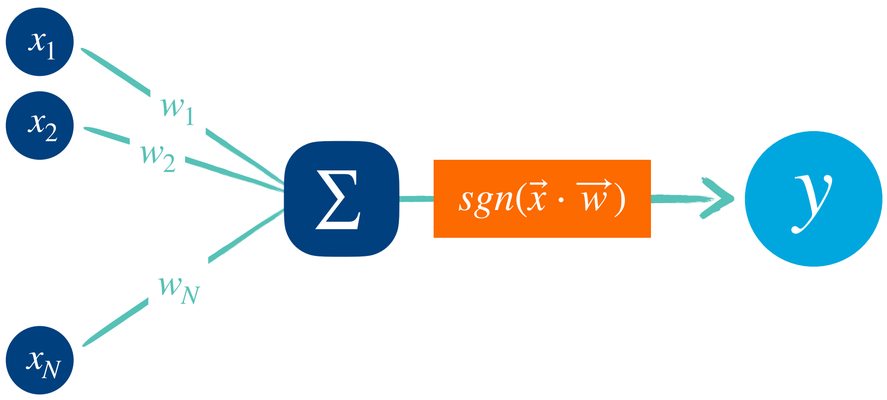
$$E_w = \sum_i^P \Theta\left(-y_i \cdot \mathrm{sgn}\left(w\, x_i\right)\right)$$

Entropy

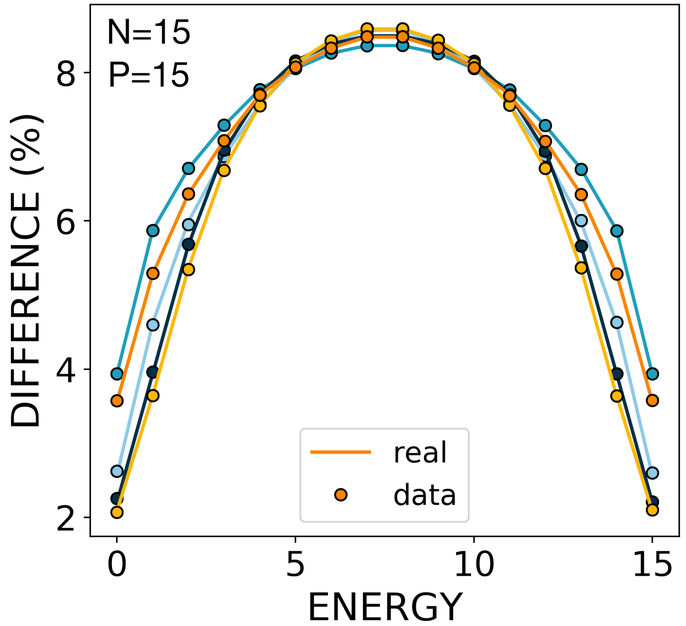
$$S(\overline{E}) = \log\left(\sum_{\{w\}} \delta(E_w - \overline{E})\right)$$



Validation



Multiple realizations on different data-sets



Random Inputs $\{x_i\}$

$$\begin{array}{c} [1,1,-1,\ldots,1,-1,-1] \\ [-1,1,1,\ \ldots,-1,1,-1] \\ \vdots \\ [-1,-1,1,\ \ldots,-1,1,1] \end{array}$$

P

N

Teacher vector $\overline{w}$

$$[-1,1,-1\ \ldots,-1,-1,1]$$

Labels $\{y_i\}$

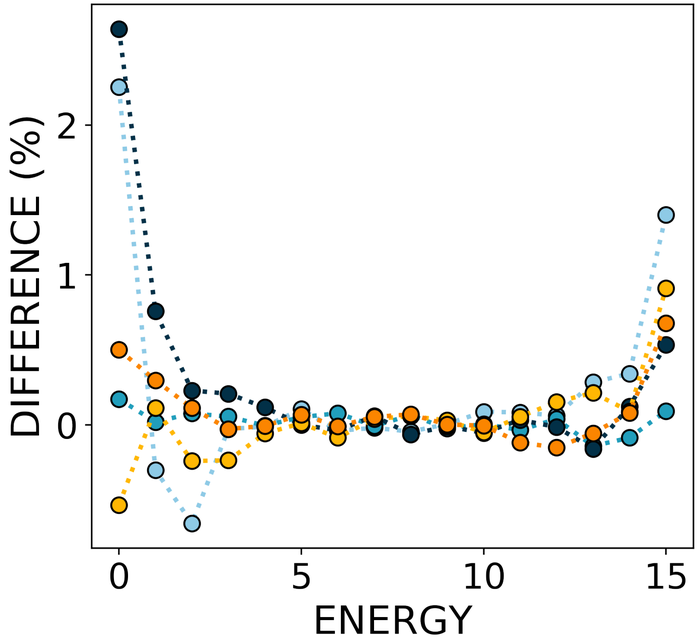
$$y_i = \mathrm{sgn}\left(\overline{w}\, x_i\right)$$

Energy

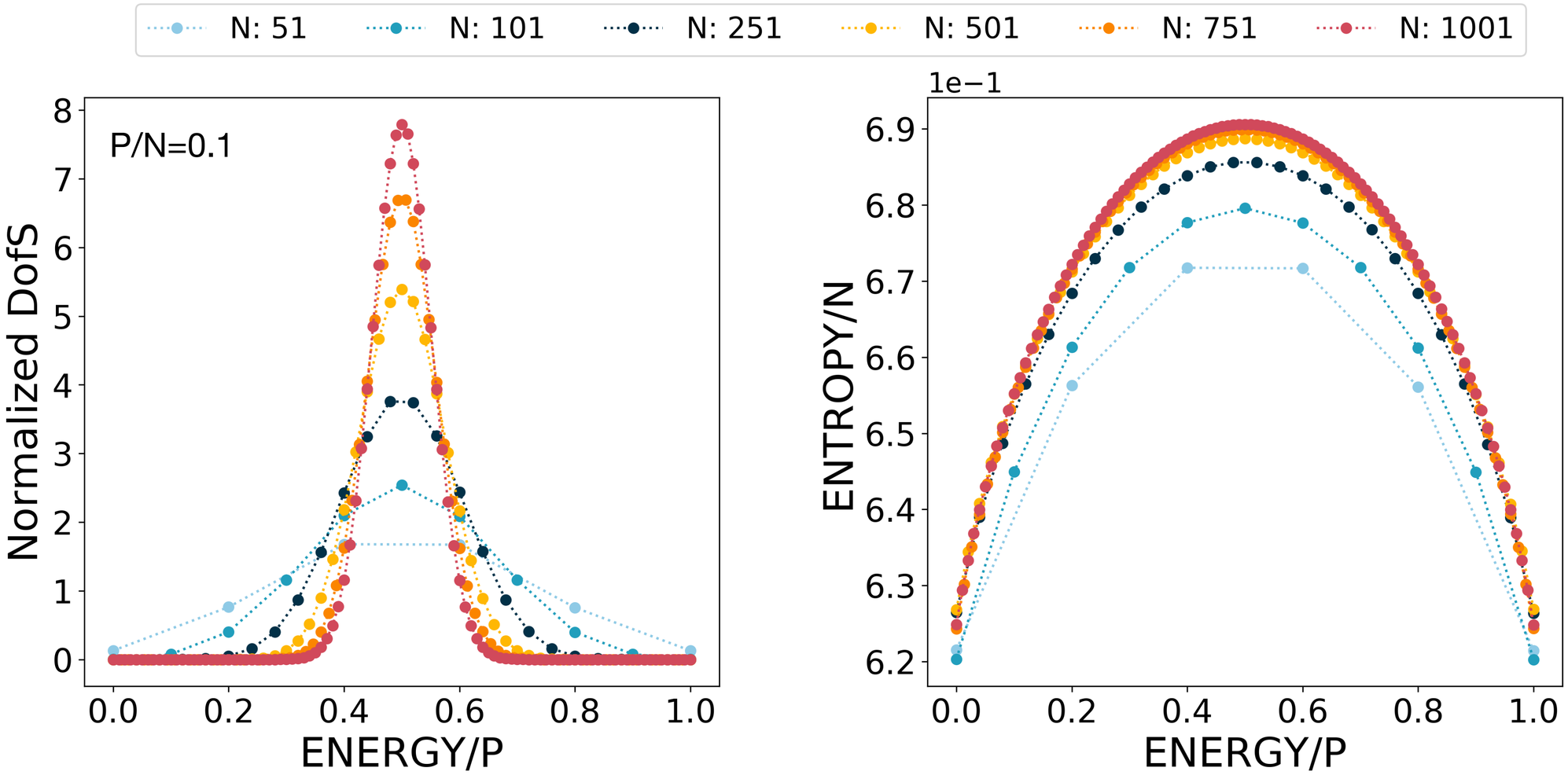
$$E_w = \sum_i^P \Theta\left(-y_i \cdot \mathrm{sgn}\left(w\, x_i\right)\right)$$

Entropy

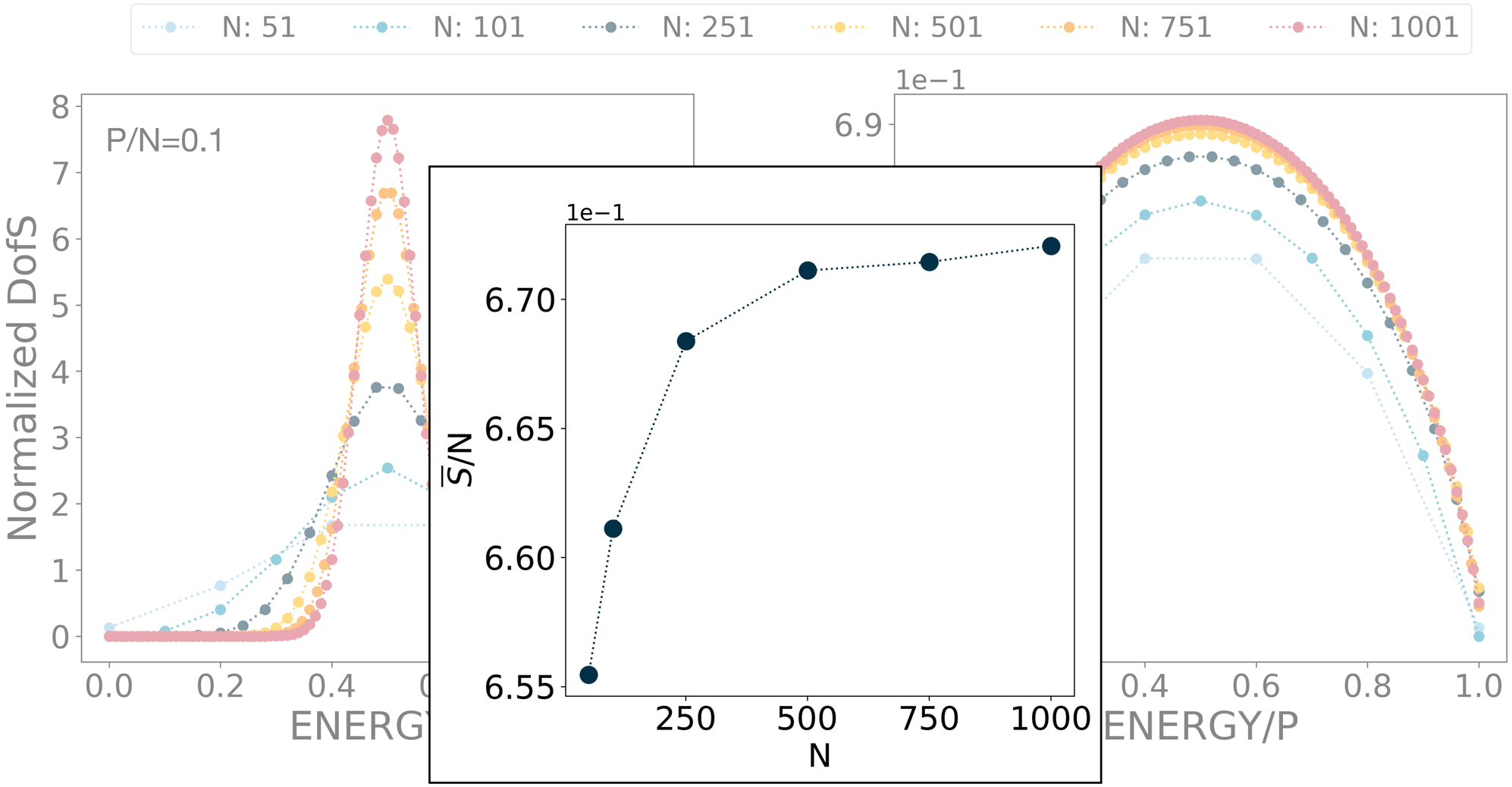
$$S(\overline{E}) = \log\left(\sum_{\{w\}} \delta(E_w - \overline{E})\right)$$



Random Data


N: 51
N: 101
N: 251
N: 501
N: 751
N: 1001
P/N=0.1
Normalized DofS
ENERGY/P
1e−1
ENTROPY/N
ENERGY/P

Random Data


N: 51
N: 101
N: 251
N: 501
N: 751
N: 1001
P/N=0.1
Normalized DofS
ENERGY
ENERGY/P
1e−1
6.9
0.0
0.2
0.4
0.6
0.8
1.0
1e−1
6.70
6.65
6.60
6.55
$\overline{S}/N$
250
500
750
1000
N

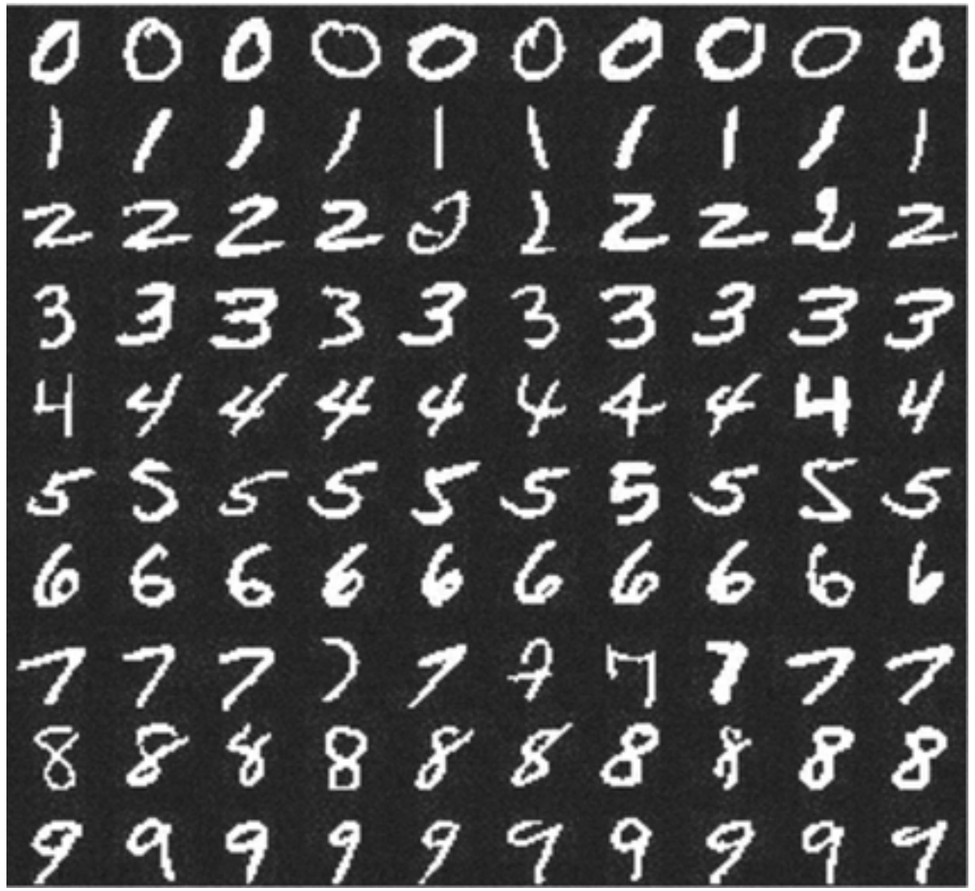
Real Data

MNIST

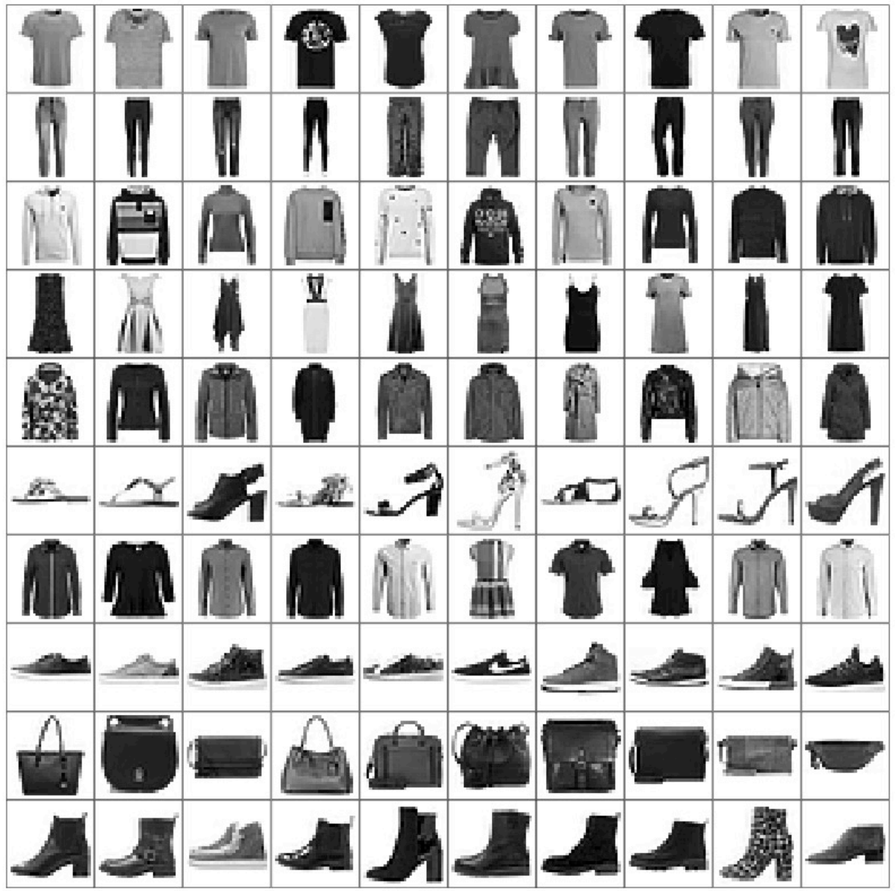
Fashion-MNIST

Real Data

MNIST



Fashion-MNIST



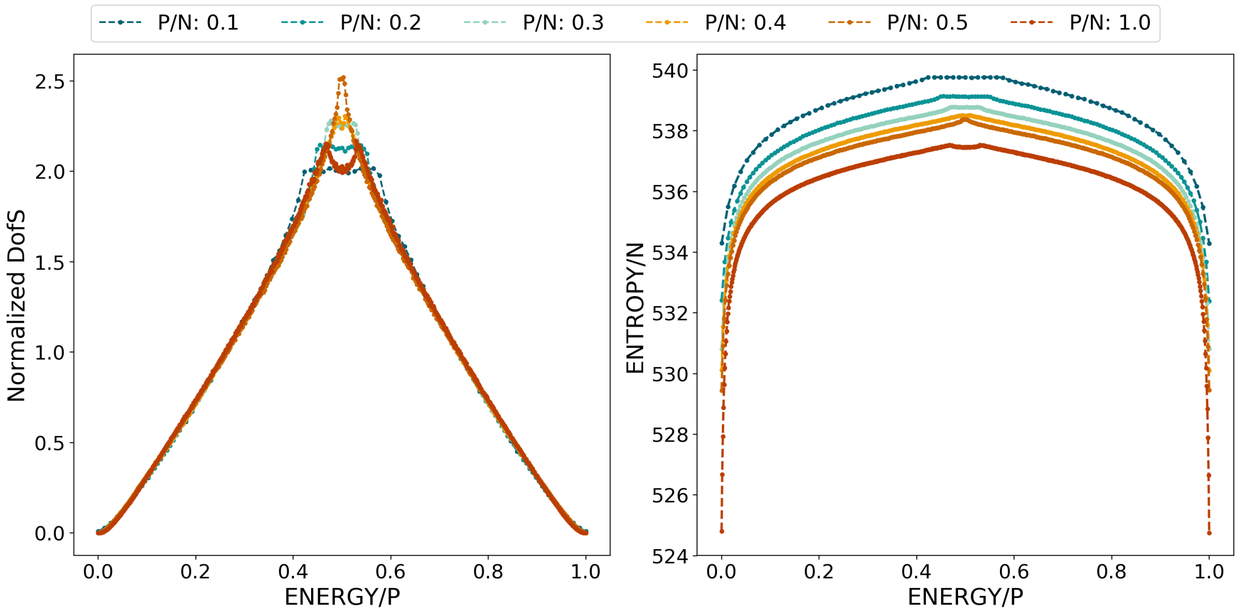
Real Data

MNIST

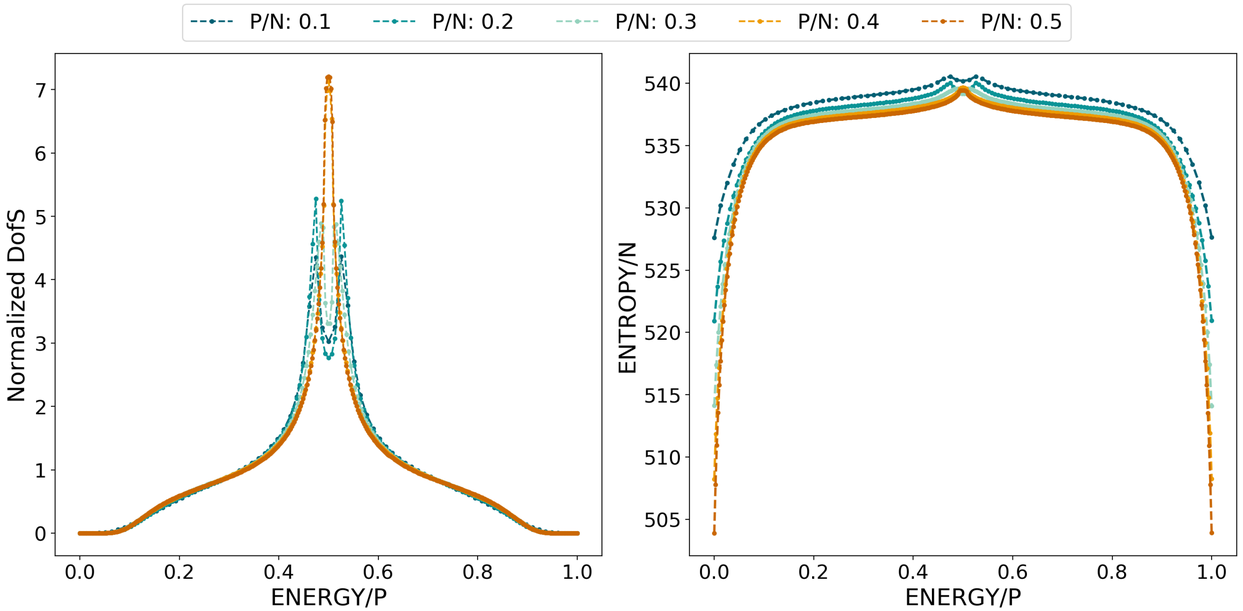
Fashion-MNIST

Real Data

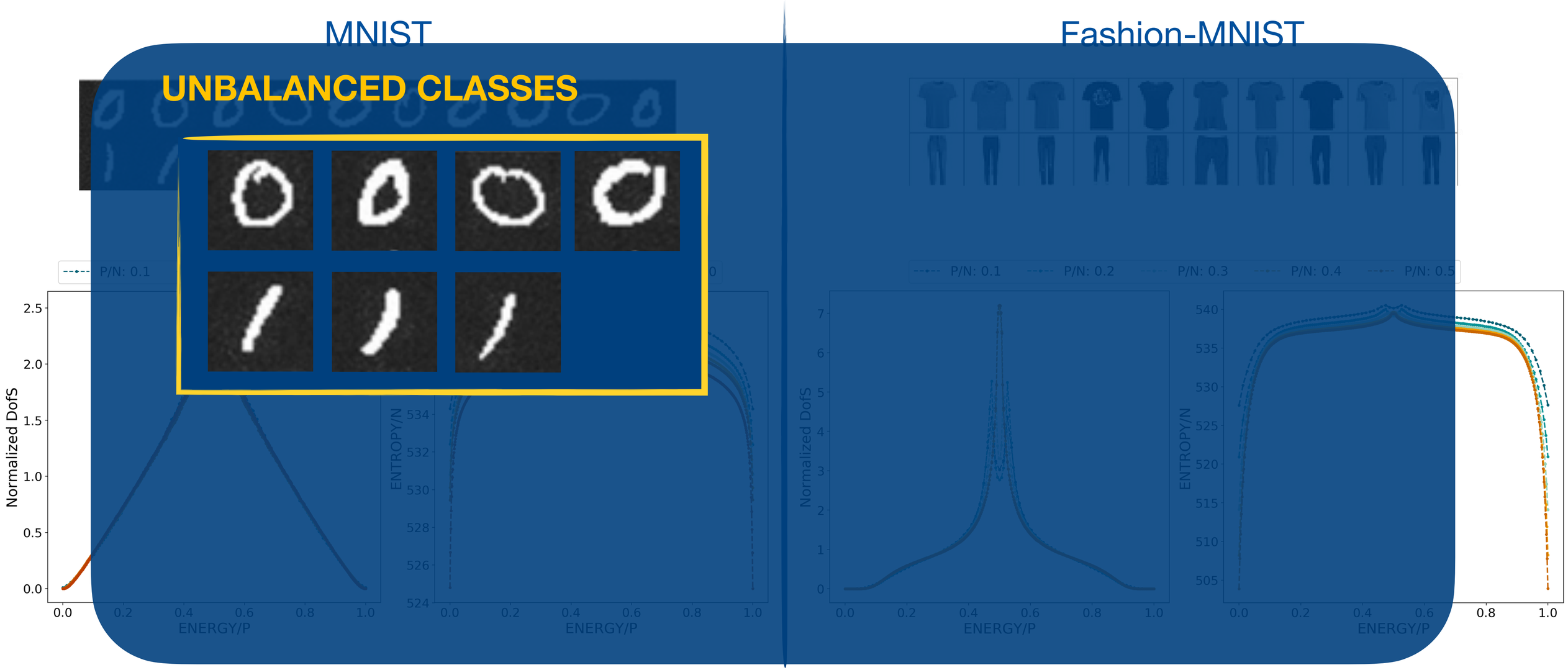
MNIST



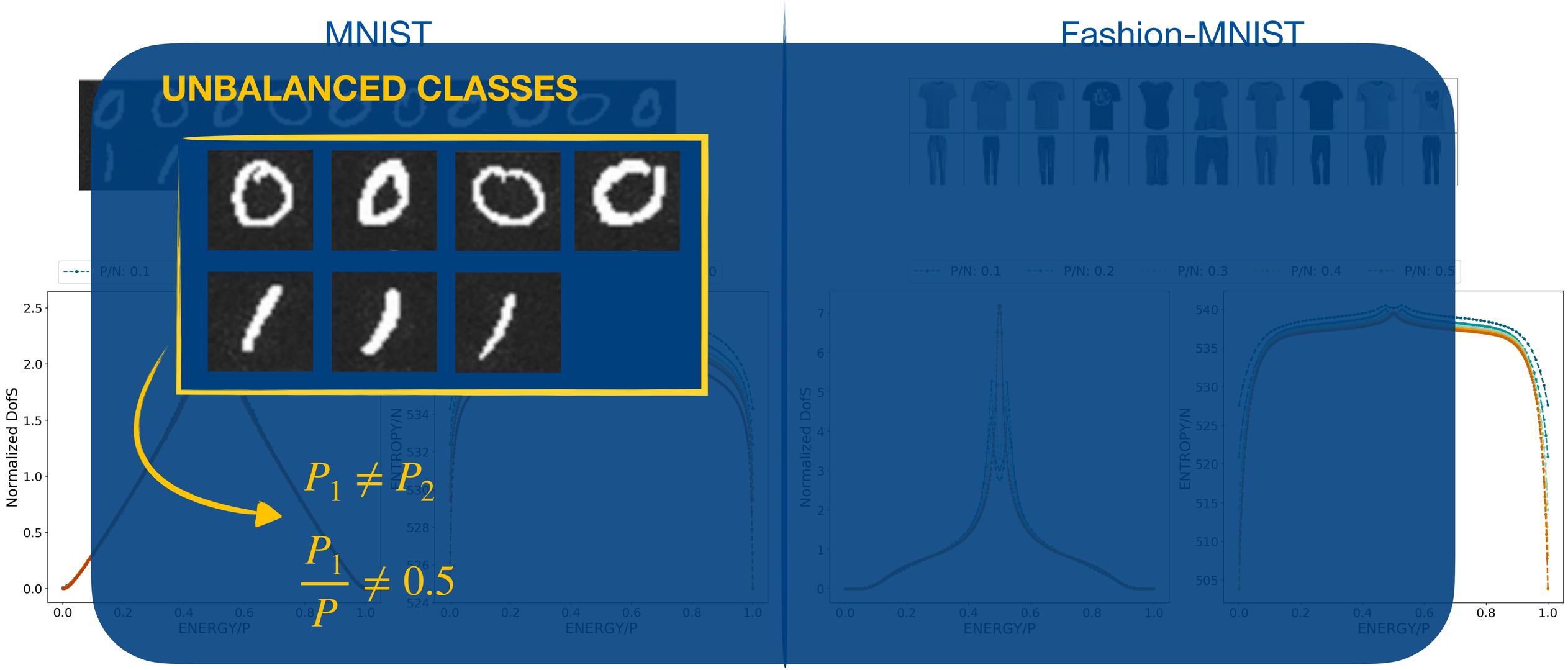
Fashion-MNIST



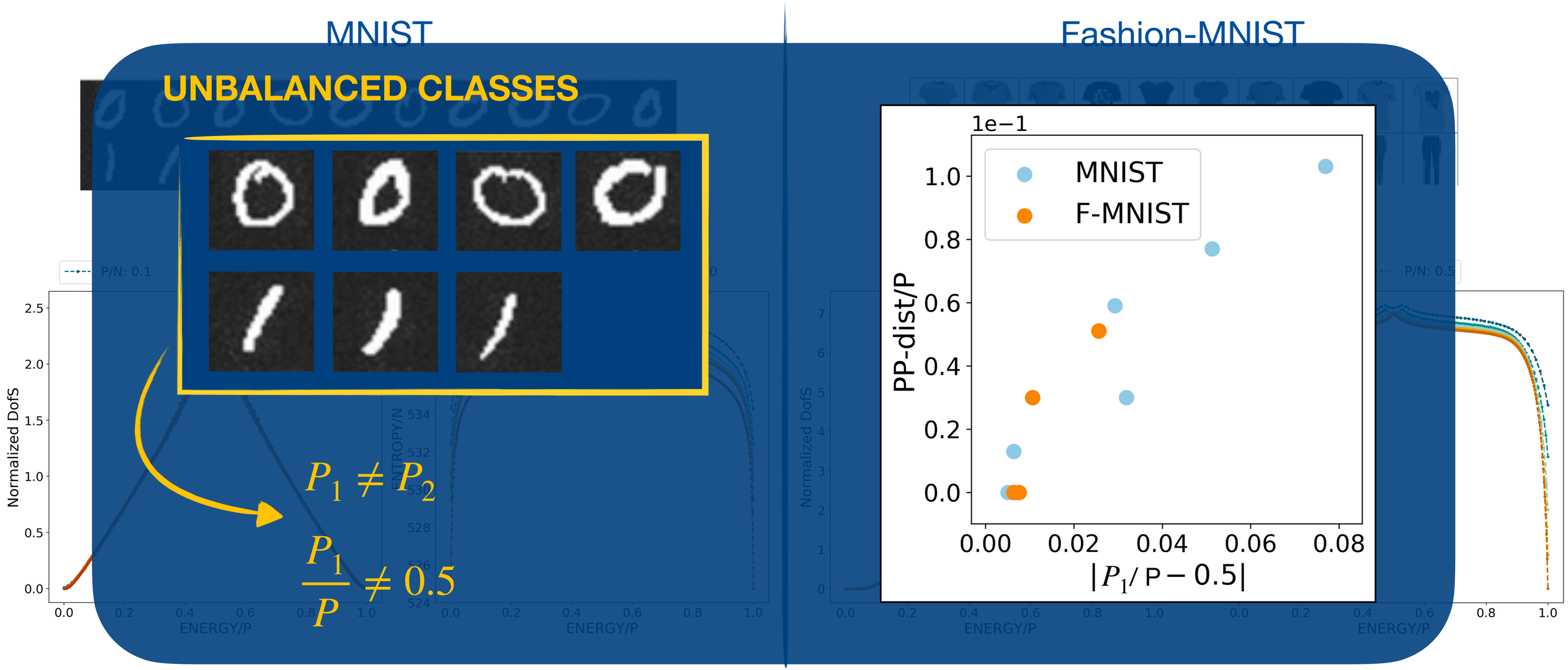
Real Data


MNIST
Fashion-MNIST
UNBALANCED CLASSES
P/N: 0.1
P/N: 0.2
P/N: 0.3
P/N: 0.4
P/N: 0.5
Normalized DofS
ENTROPY/N
ENERGY/P

Real Data


MNIST
Fashion-MNIST
UNBALANCED CLASSES
P/N: 0.1
P/N: 0.2
P/N: 0.3
P/N: 0.4
P/N: 0.5
Normalized DofS
ENTROPY/N
ENERGY/P
$P_1 \neq P_2$
$\frac{P_1}{P} \neq 0.5$

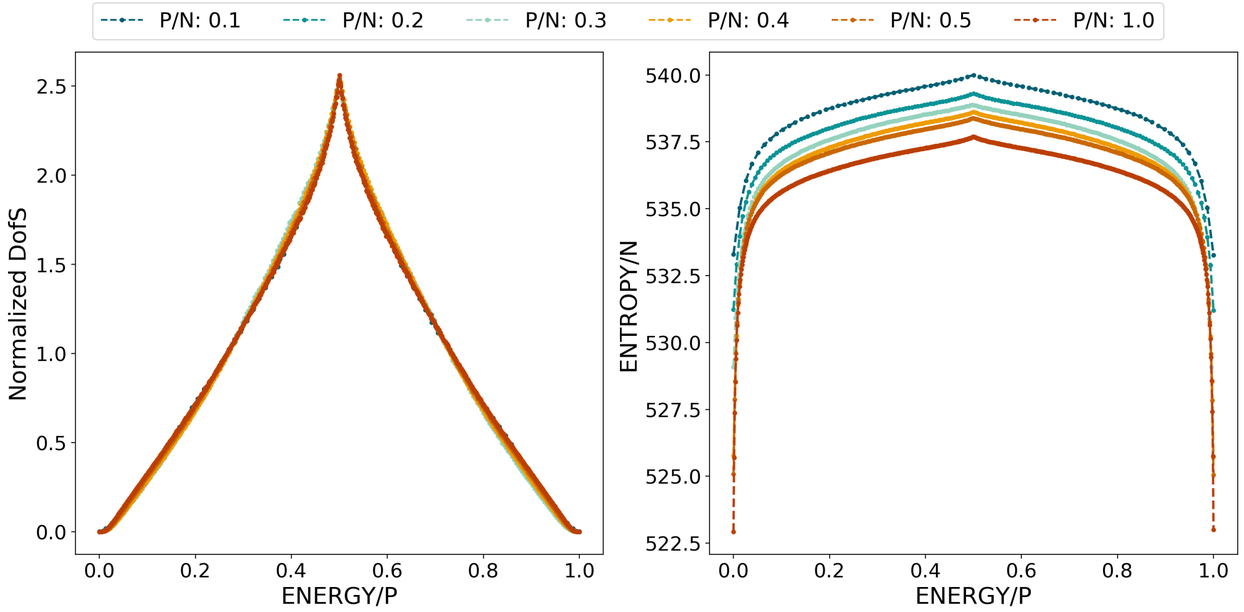
Real Data


MNIST
Fashion-MNIST
UNBALANCED CLASSES
P/N: 0.1
P/N: 0.5
Normalized DofS
ENTROPY/N
ENERGY/P
$P_1 \neq P_2$
$\frac{P_1}{P} \neq 0.5$
1e−1
MNIST
F-MNIST
PP-dist/P
$|P_1/P - 0.5|$

Real Data

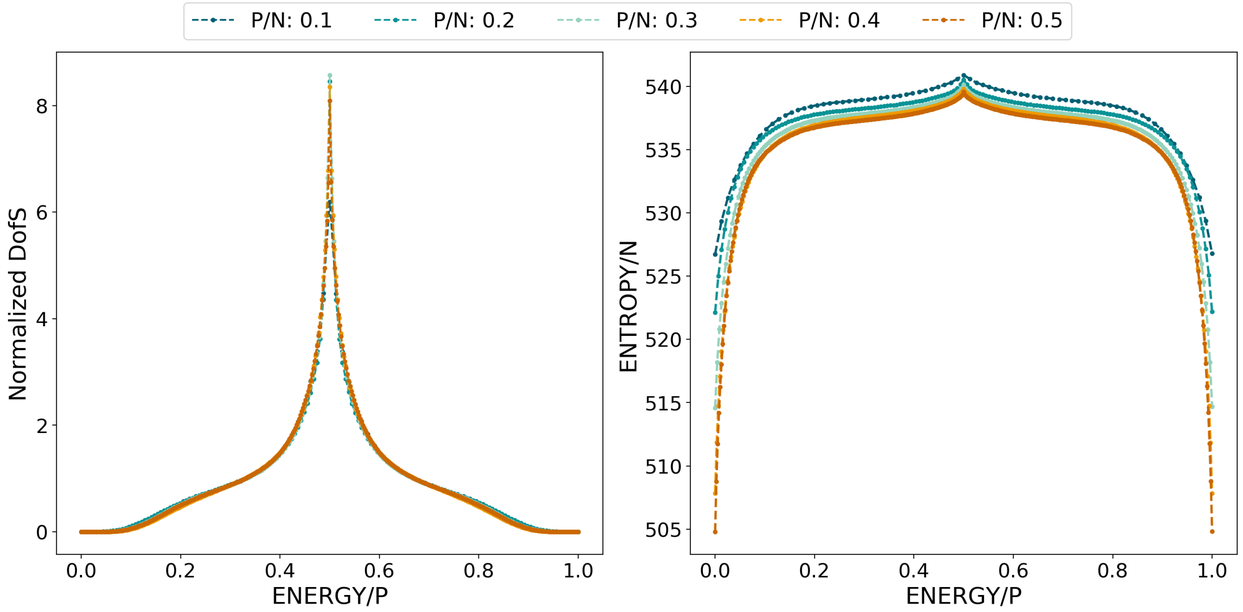
MNIST

BALANCED CLASSES



Fashion-MNIST

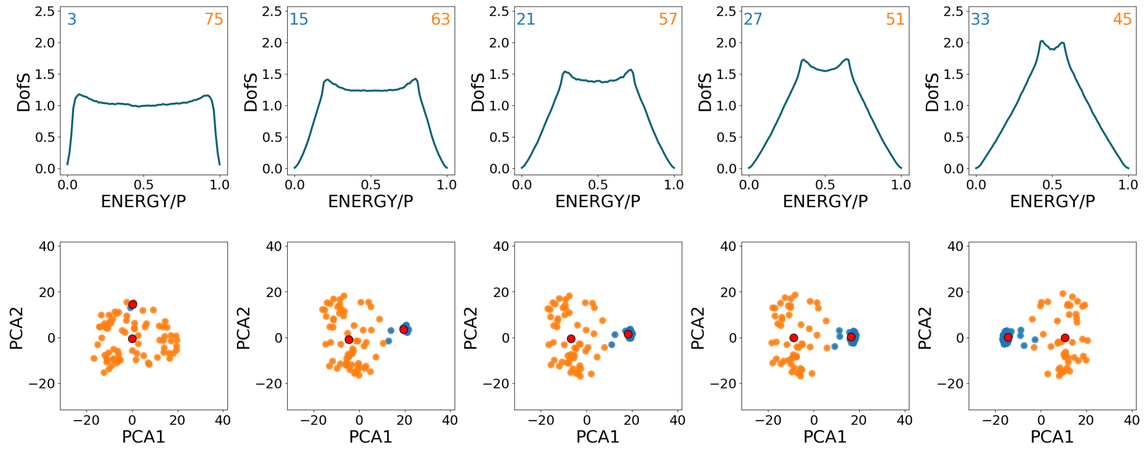
BALANCED CLASSES



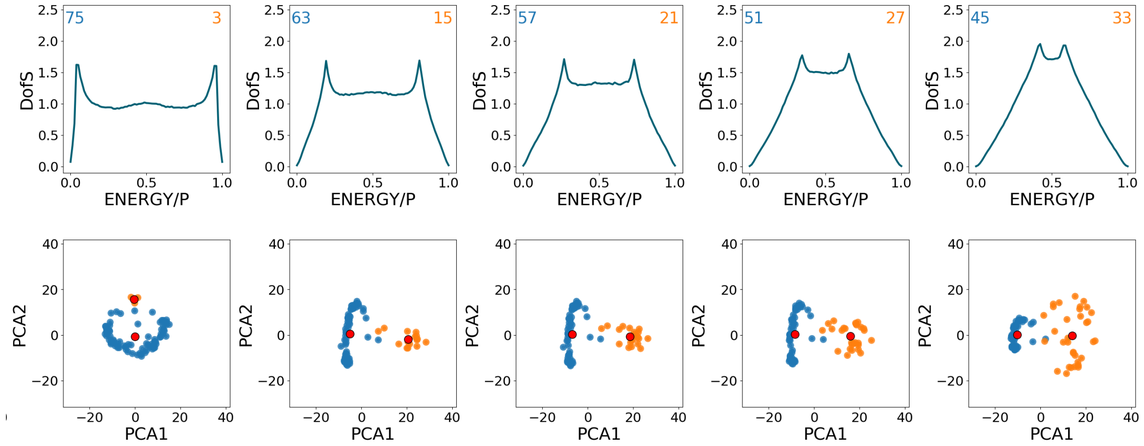
Real Data

MNIST (P/N=0.1)

P1<P2



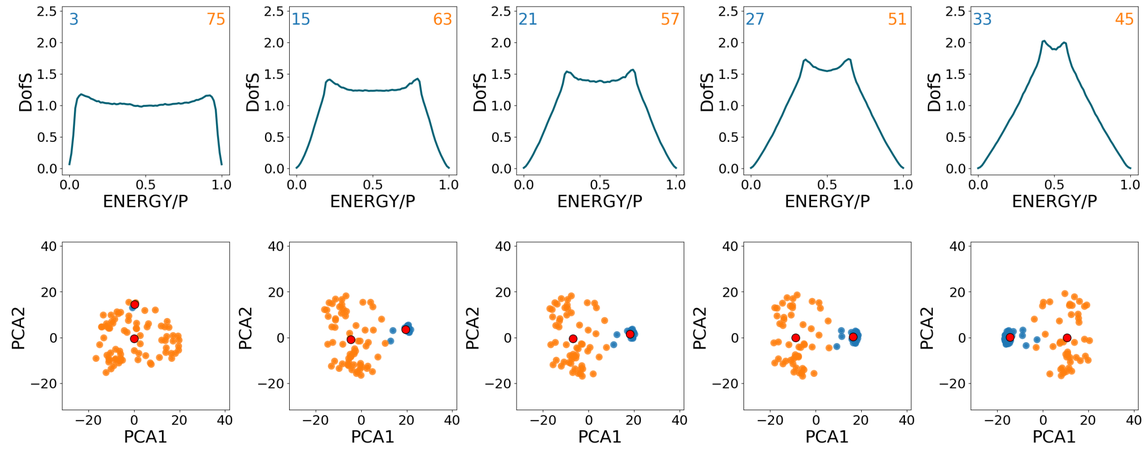
P1>P2



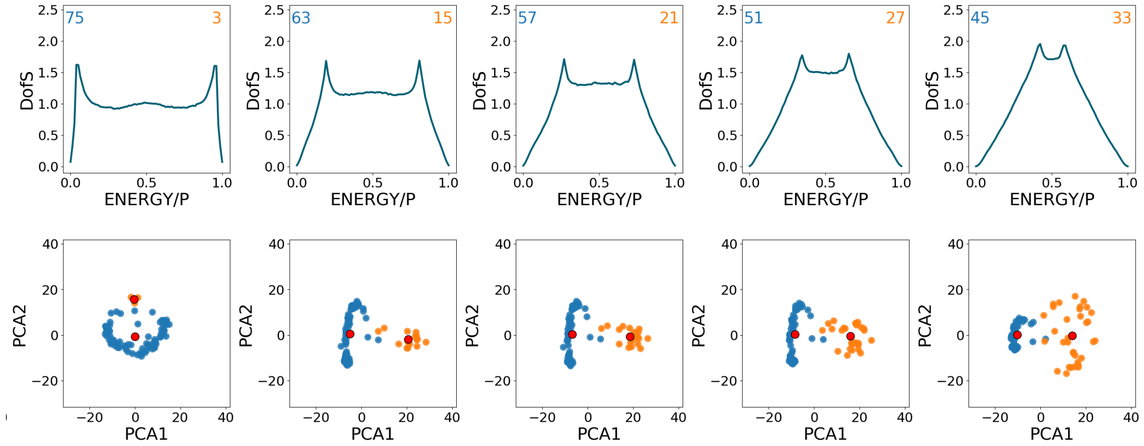
Real Data

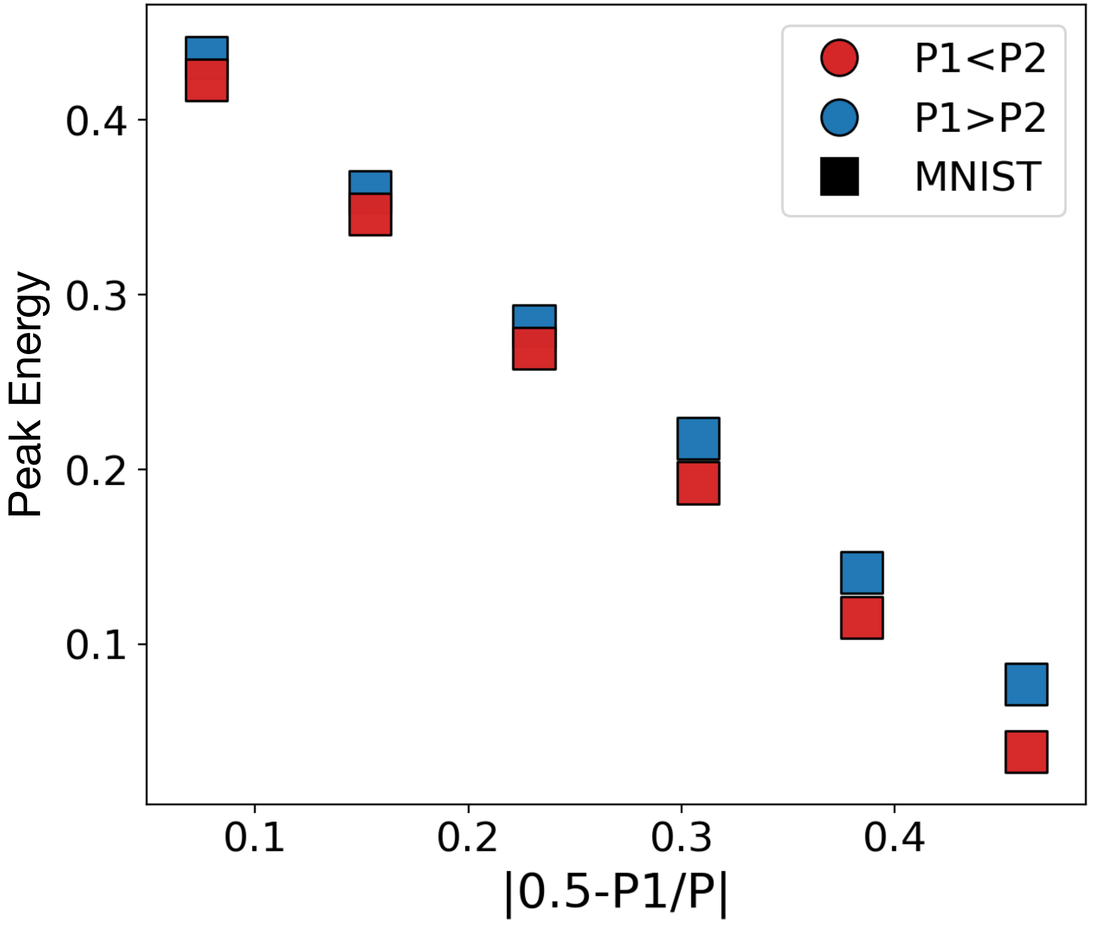
MNIST (P/N=0.1)

P1<P2



P1>P2

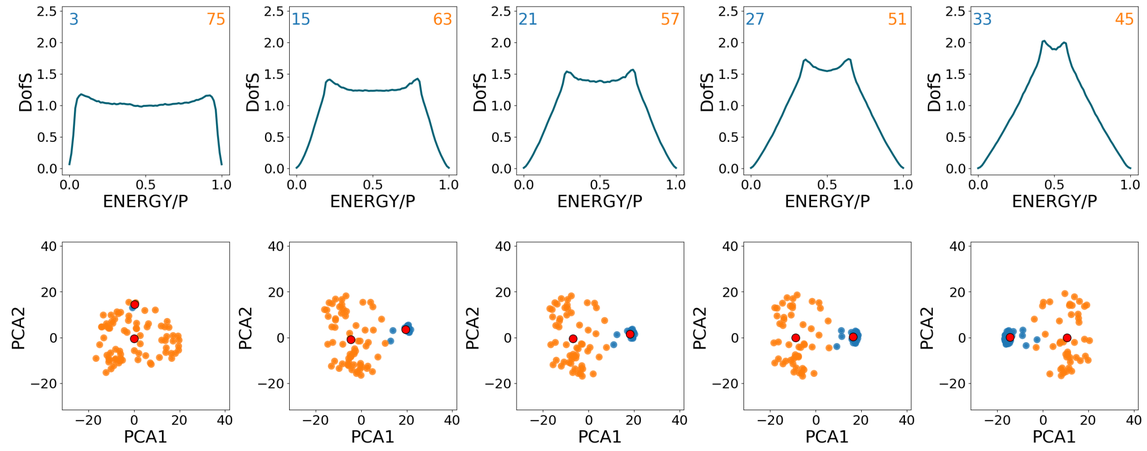




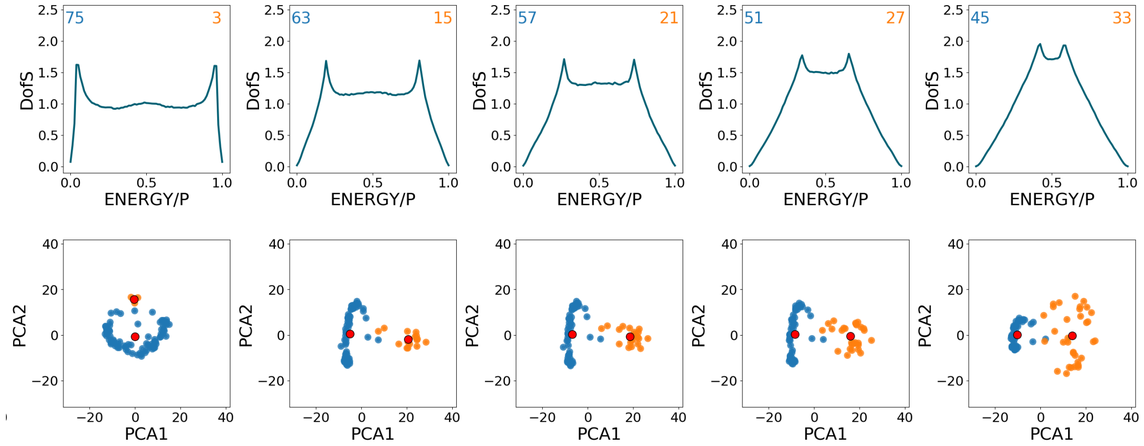
Real Data

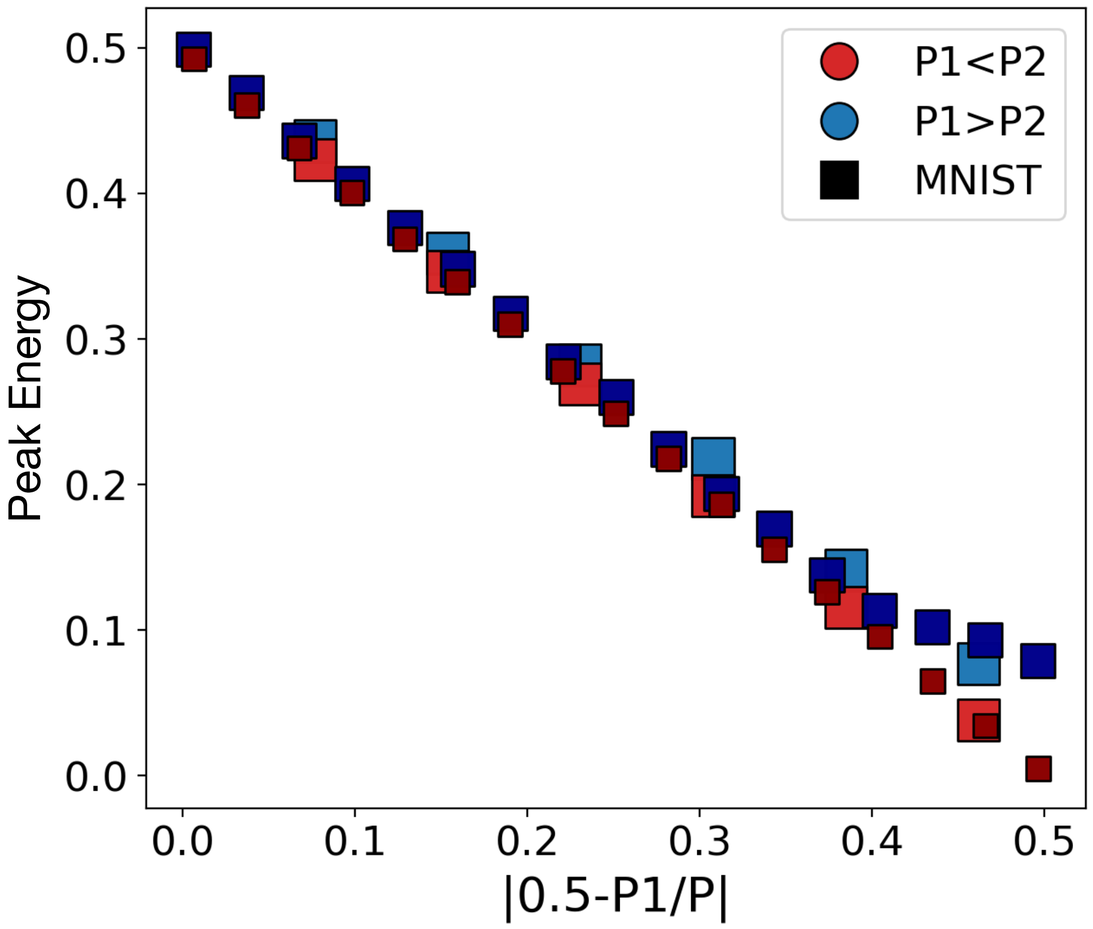
MNIST (P/N=0.1)

P1<P2



P1>P2

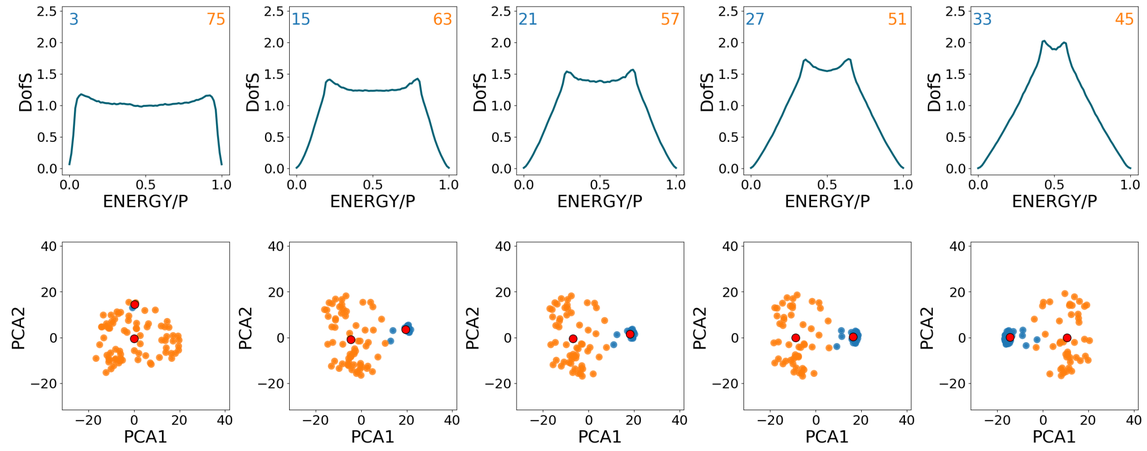




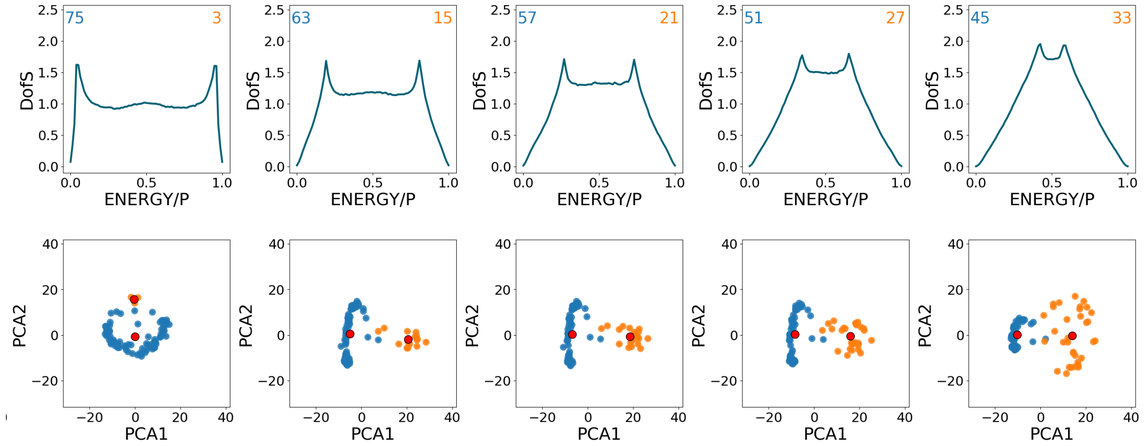
Real Data

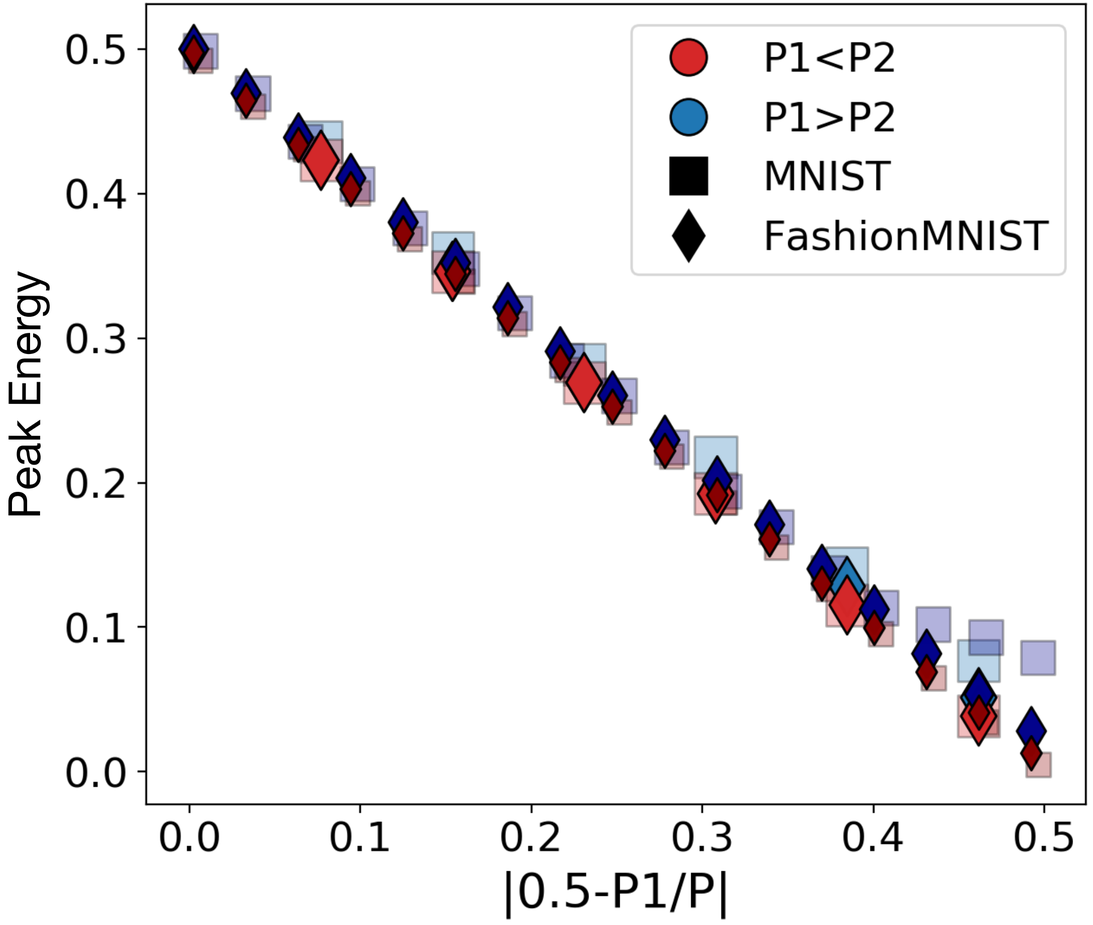
MNIST (P/N=0.1)

P1<P2



P1>P2

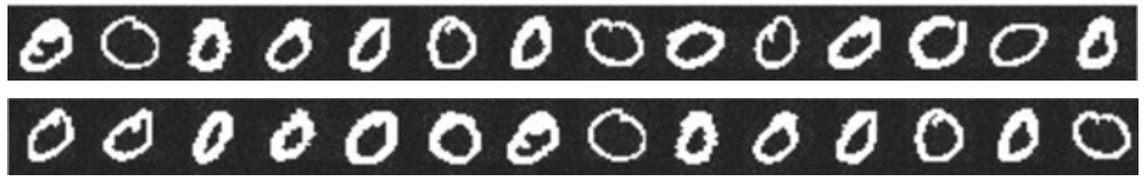


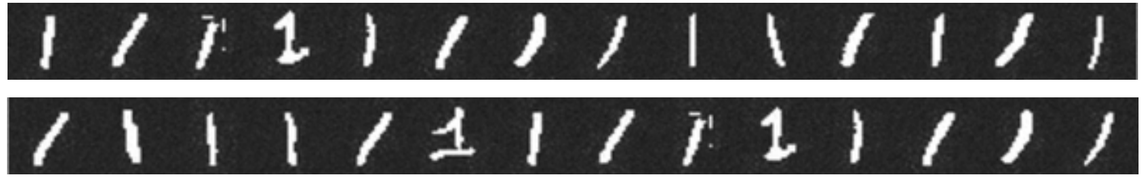


Gaussian Clones

Multivariate Normal Distribution

$$N(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$



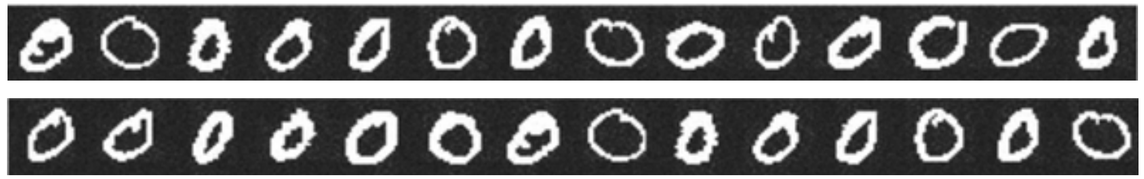


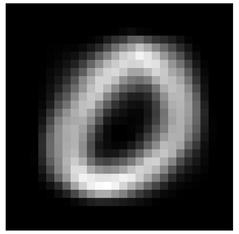
Gaussian Clones

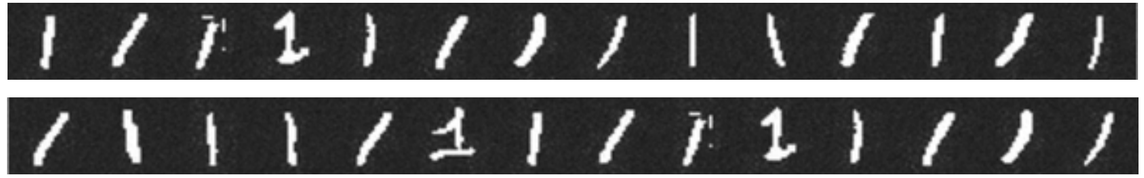
Multivariate Normal Distribution

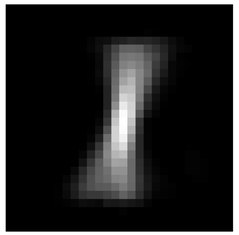
$$N(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Mean - μ









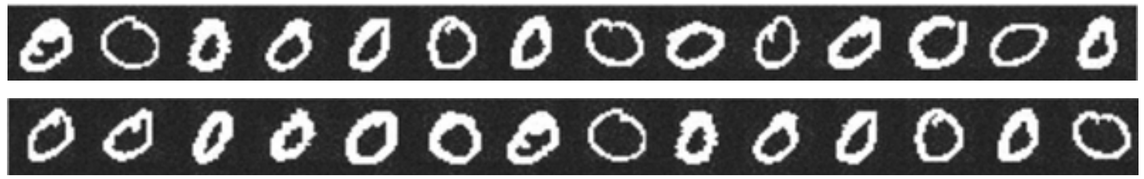
Gaussian Clones

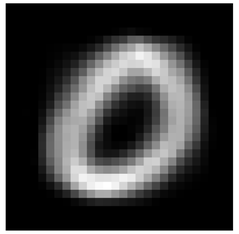
Multivariate Normal Distribution

$$N(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

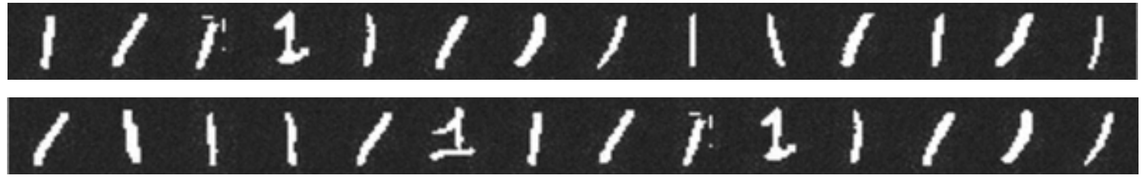
Mean - μ

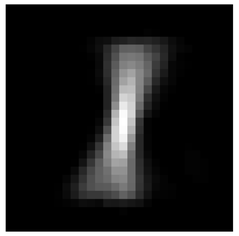
Covariance - Σ





1	0.16	0.79	-0.0019	0.09	-0.058	0.3	0.16
0.16	1	0.093	0.0067	-0.037	0.01	-0.025	-0.041
0.79	0.093	1	0.059	0.16	0.028	0.35	0.23
-0.0019	0.0067	0.059	1	0.48	0.14	0.16	0.35
0.09	-0.037	0.16	0.48	1	0.11	0.32	0.54
-0.058	0.01	0.028	0.14	0.11	1	-0.15	0.16
0.3	-0.025	0.35	0.16	0.32	-0.15	1	0.35
0.16	-0.041	0.23	0.35	0.54	0.16	0.35	1



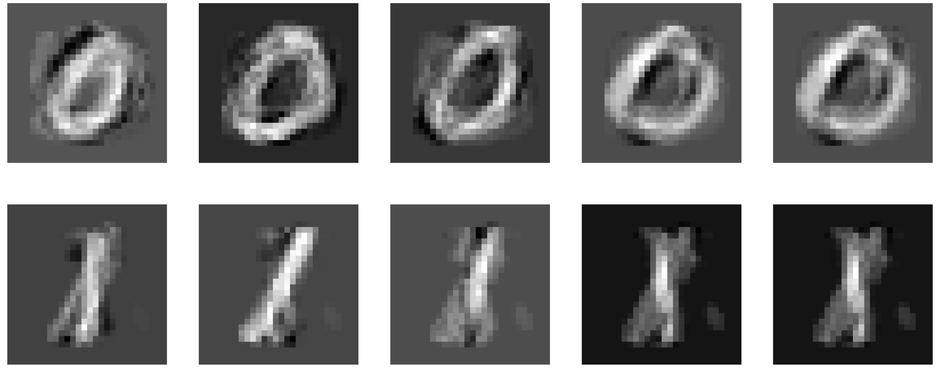


1	0.02	0.33	0.051	0.025	0.018	0.0074	0.086
0.02	1	0.3	0.3	0.18	0.26	0.3	0.27
0.33	0.3	1	0.19	0.15	0.3	0.19	0.33
0.051	0.3	0.19	1	0.47	0.35	0.88	0.53
0.025	0.18	0.15	0.47	1	0.33	0.33	0.75
0.018	0.26	0.3	0.35	0.33	1	0.28	0.76
0.0074	0.3	0.19	0.88	0.33	0.28	1	0.4
0.086	0.27	0.33	0.53	0.75	0.76	0.4	1

Gaussian Clones

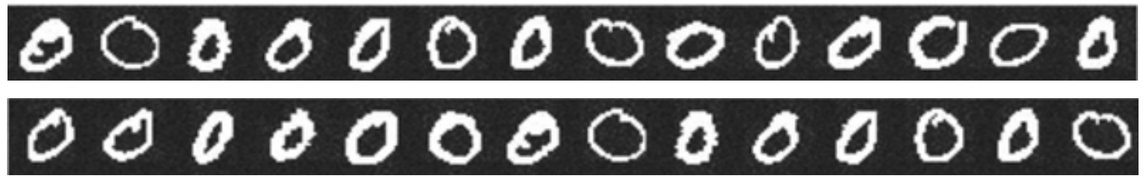
Multivariate Normal Distribution

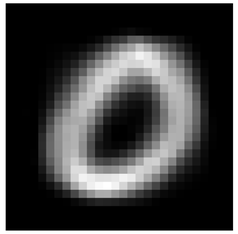
$$N(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$



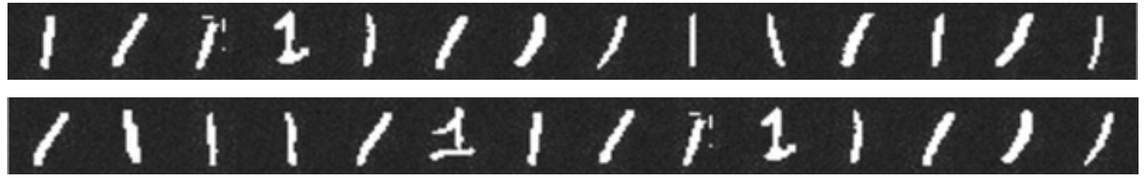
Mean - μ

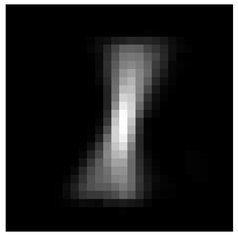
Covariance - Σ





1	0.16	0.79	-0.0019	0.09	-0.058	0.3	0.16
0.16	1	0.093	0.0067	-0.037	0.01	-0.025	-0.041
0.79	0.093	1	0.059	0.16	0.028	0.35	0.23
-0.0019	0.0067	0.059	1	0.48	0.14	0.16	0.35
0.09	-0.037	0.16	0.48	1	0.11	0.32	0.54
-0.058	0.01	0.028	0.14	0.11	1	-0.15	0.16
0.3	-0.025	0.35	0.16	0.32	-0.15	1	0.35
0.16	-0.041	0.23	0.35	0.54	0.16	0.35	1





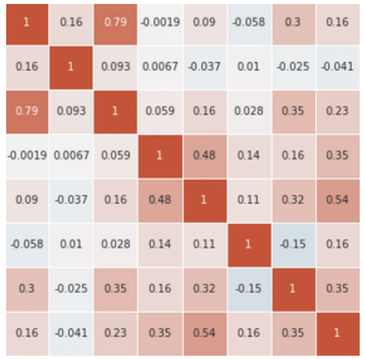
1	0.02	0.33	0.051	0.025	0.018	0.0074	0.086
0.02	1	0.3	0.3	0.18	0.26	0.3	0.27
0.33	0.3	1	0.19	0.15	0.3	0.19	0.33
0.051	0.3	0.19	1	0.47	0.35	0.88	0.53
0.025	0.18	0.15	0.47	1	0.33	0.33	0.75
0.018	0.26	0.3	0.35	0.33	1	0.28	0.76
0.0074	0.3	0.19	0.88	0.33	0.28	1	0.4
0.086	0.27	0.33	0.53	0.75	0.76	0.4	1

Gaussian Clones

Multivariate Normal Distribution

$$N(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

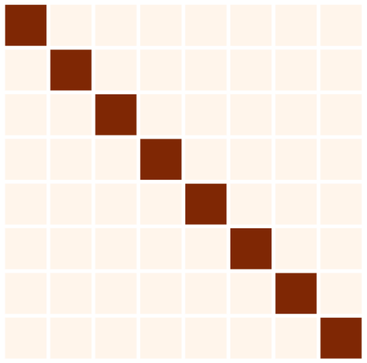
GM clone



1	0.16	0.79	-0.0019	0.09	-0.058	0.3	0.16
0.16	1	0.093	0.0067	-0.037	0.01	-0.025	-0.041
0.79	0.093	1	0.059	0.16	0.028	0.35	0.23
-0.0019	0.0067	0.059	1	0.48	0.14	0.16	0.35
0.09	-0.037	0.16	0.48	1	0.11	0.32	0.54
-0.058	0.01	0.028	0.14	0.11	1	-0.15	0.16
0.3	-0.025	0.35	0.16	0.32	-0.15	1	0.35
0.16	-0.041	0.23	0.35	0.54	0.16	0.35	1

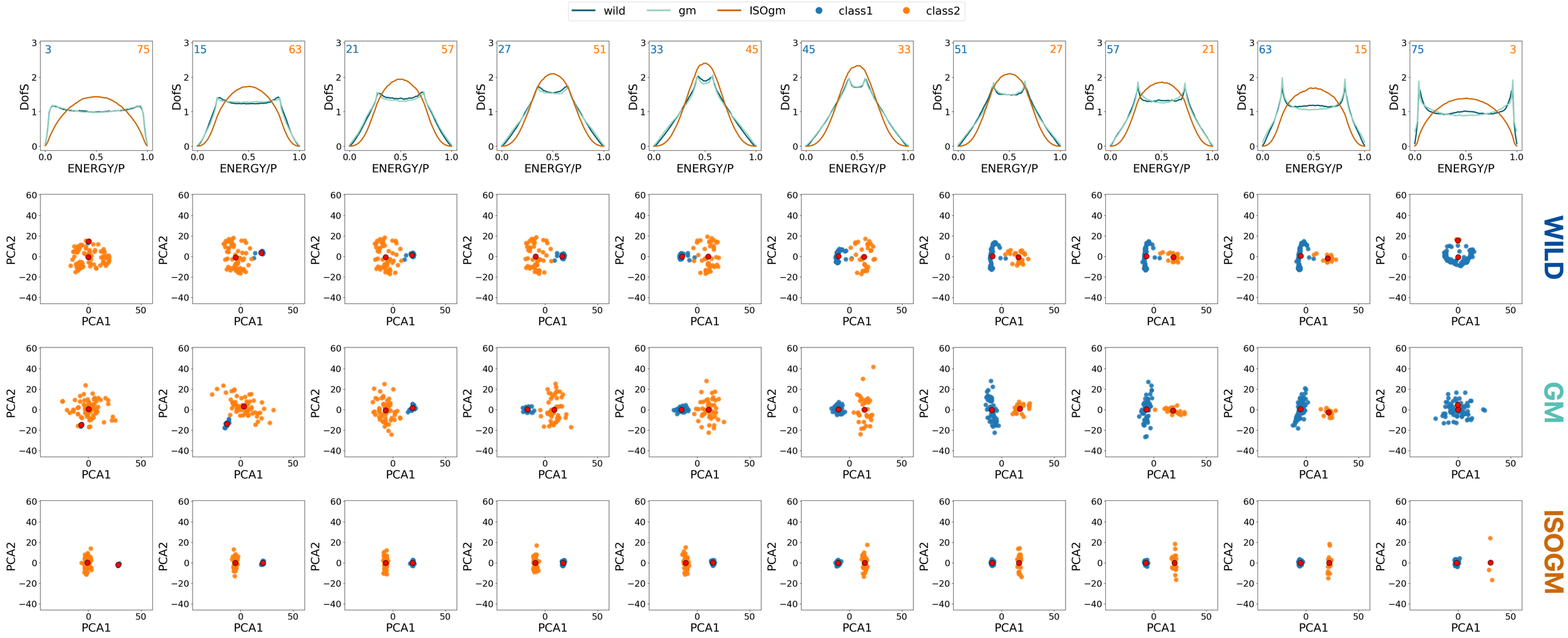
- **Mean vector**
- **Covariance matrix**

ISOGM clone

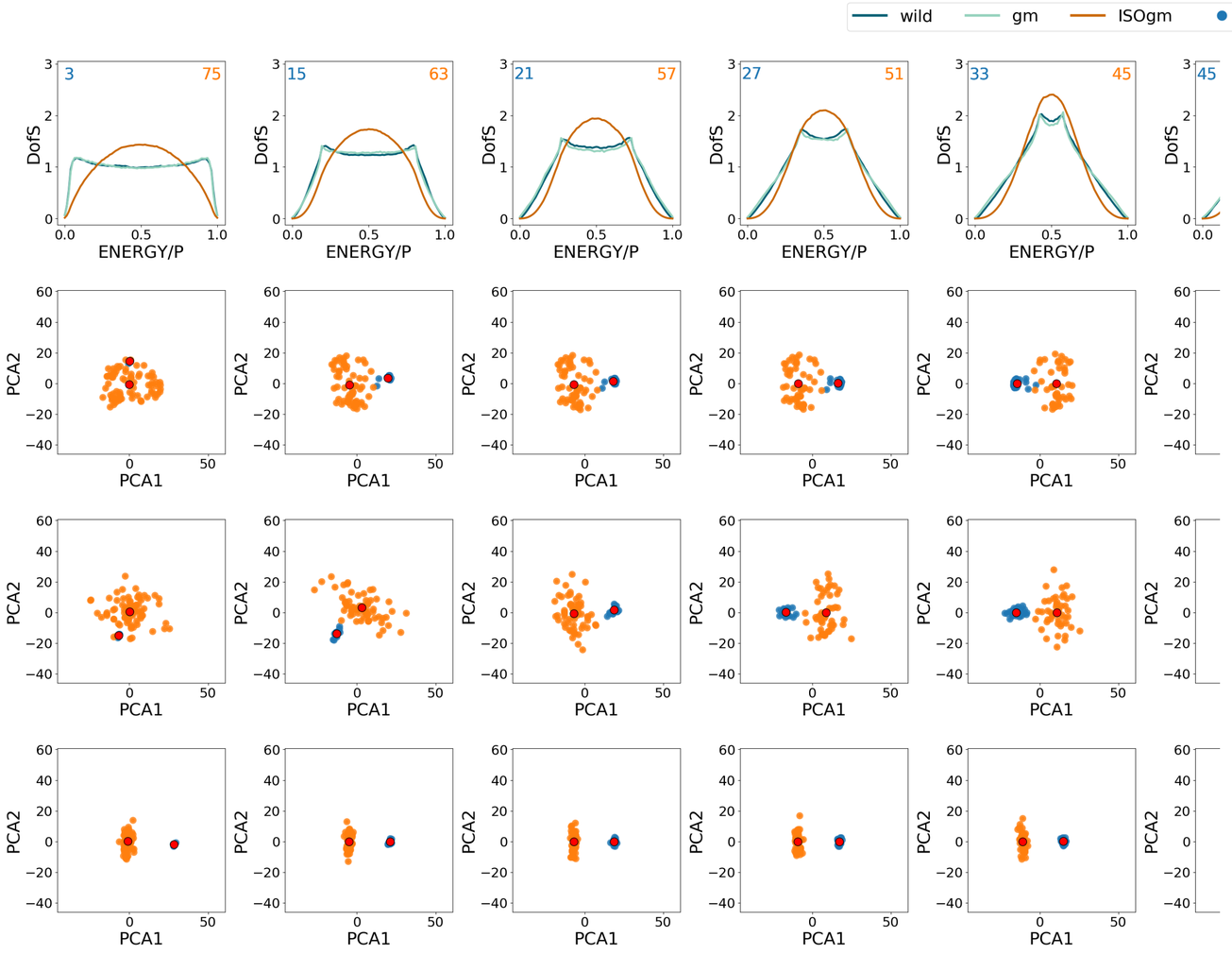


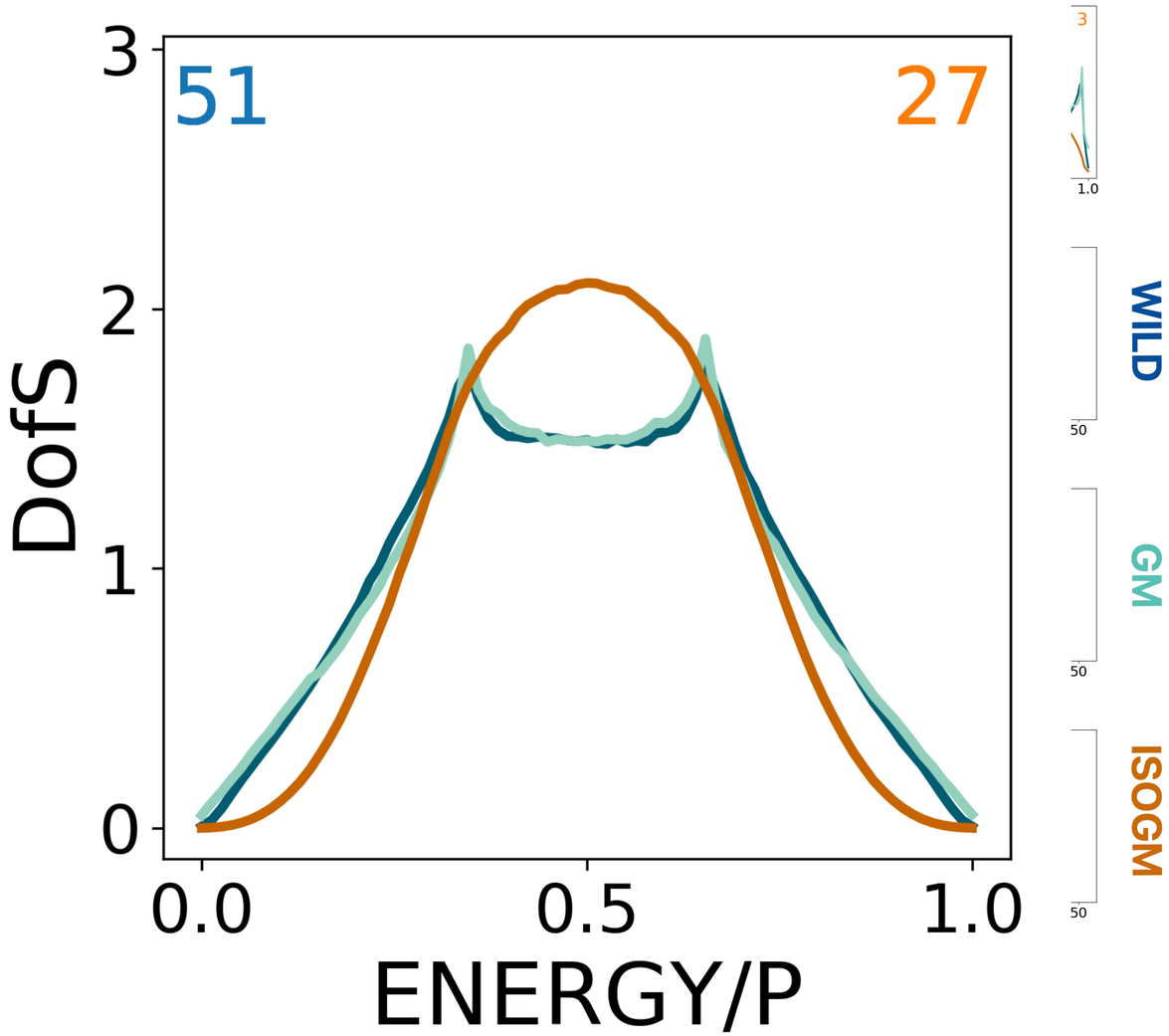
- **Mean vector**
- **Variance : diagonal matrix**

Gaussian Clones - MNIST (P/N=0.1)

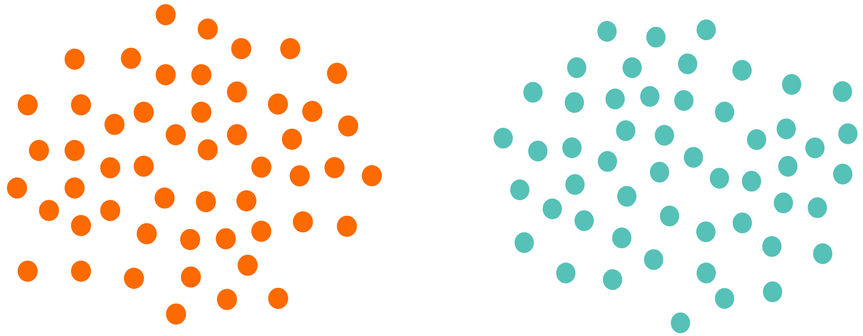


Gaussian Clones - MNIST (P/N=0.1)



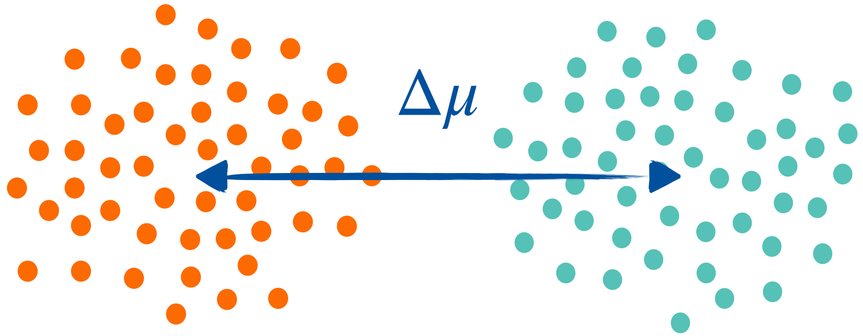


Silico dataset



Silico dataset

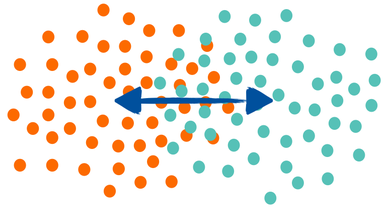
Class separation

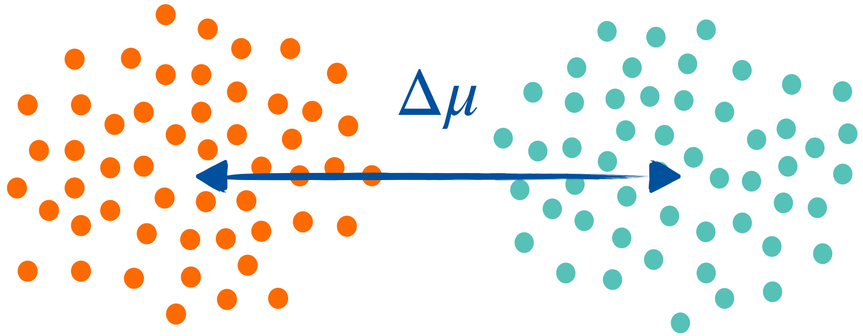


Silico dataset

Class separation

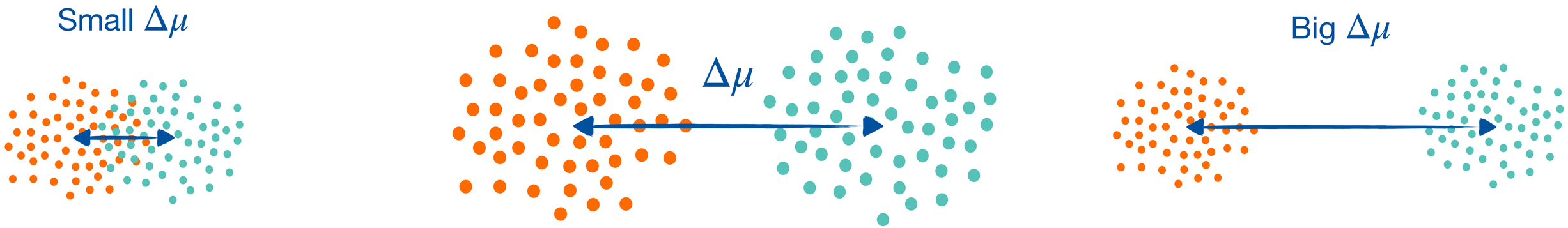
Small $\Delta\mu$





Silico dataset

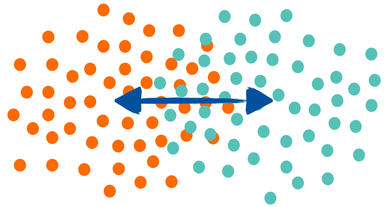
Class separation

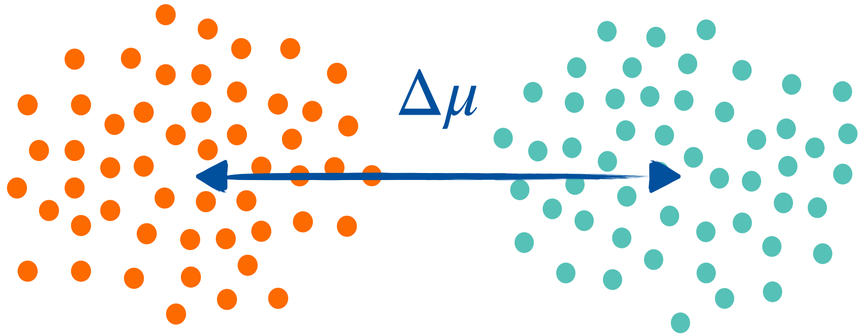


Silico dataset

Class separation

Small $\Delta\mu$

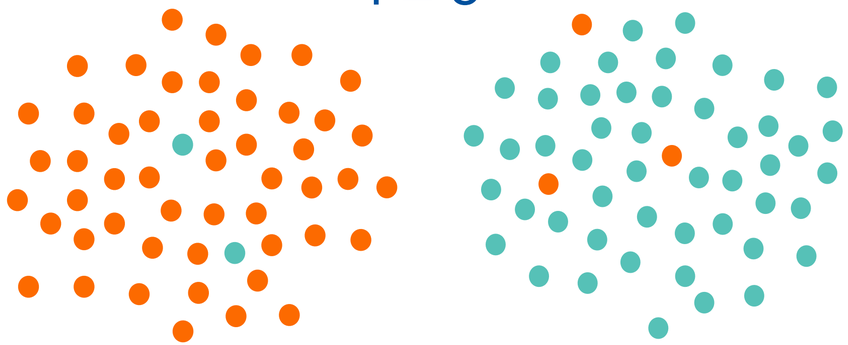




Big $\Delta\mu$

Misclassification

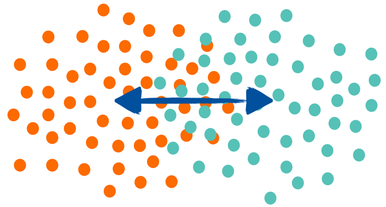
f = 5

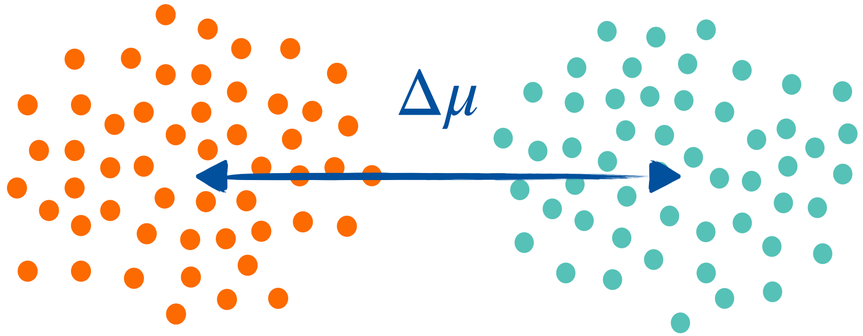


Silico dataset

Class separation

Small $\Delta\mu$

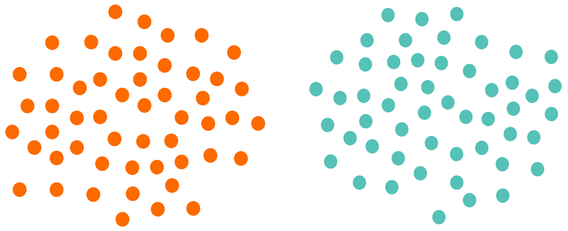




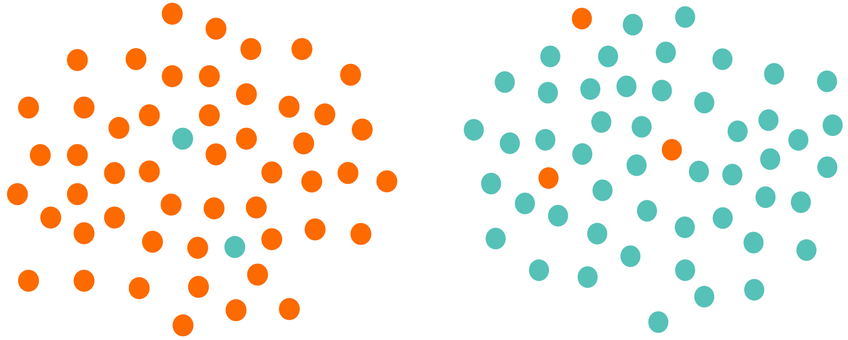
Big $\Delta\mu$

Misclassification

f = 0



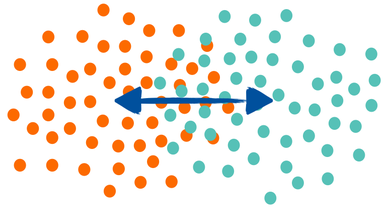
f = 5

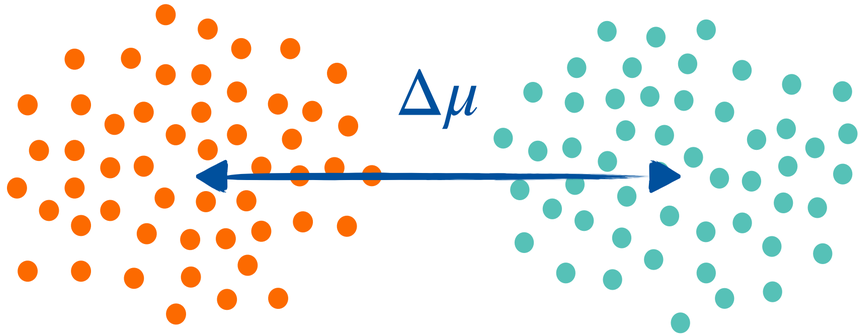


Silico dataset

Class separation

Small $\Delta\mu$

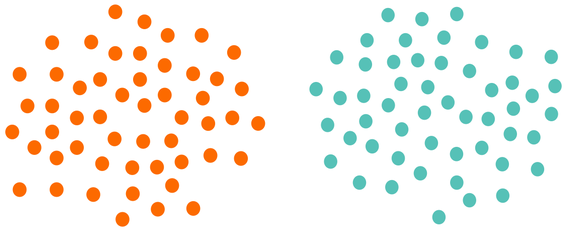




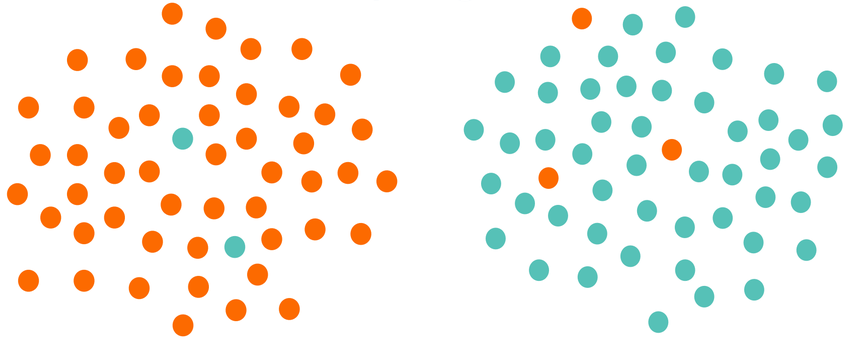
Big $\Delta\mu$

Misclassification

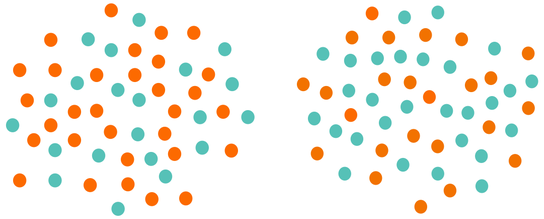
f = 0



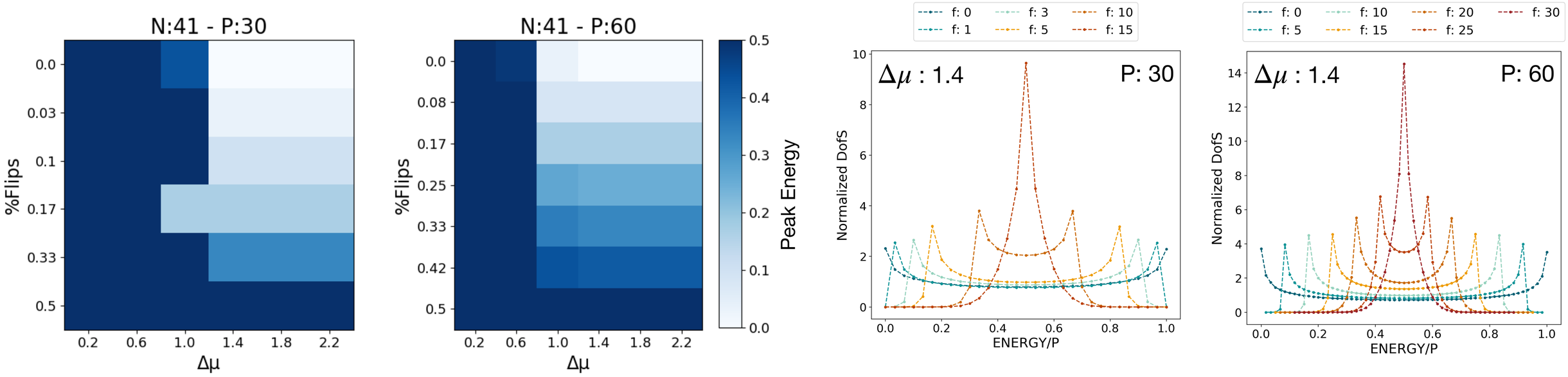
f = 5



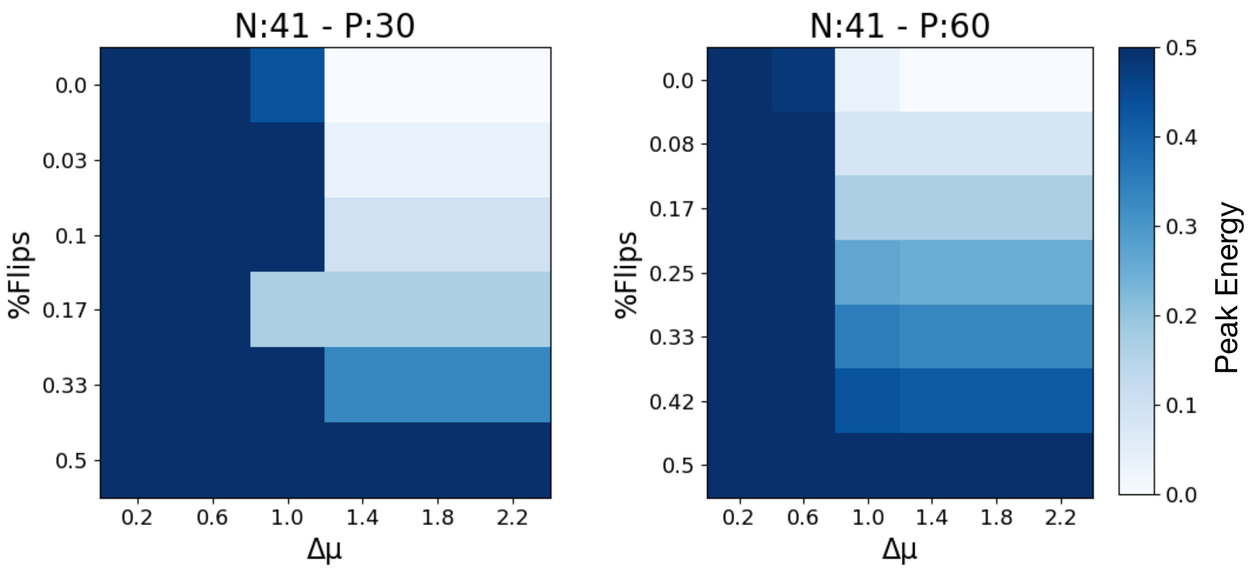
f = P/2

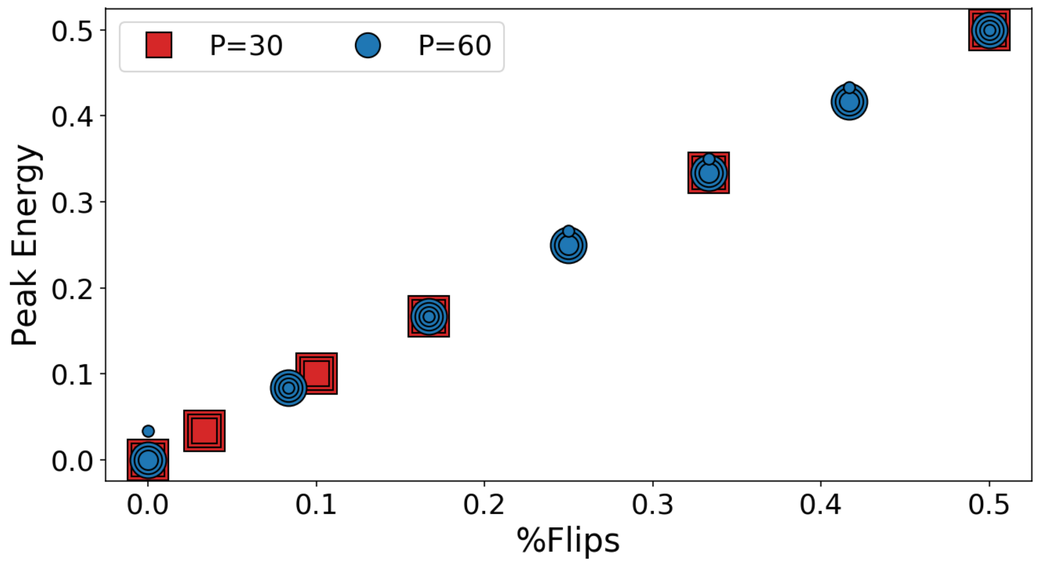


Silico dataset

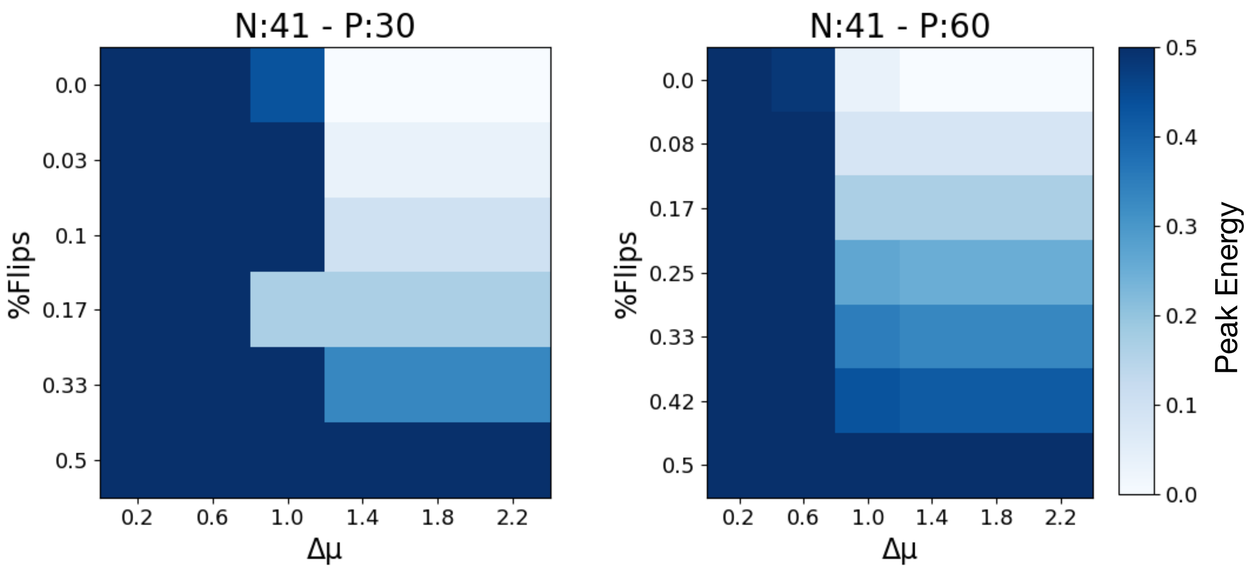


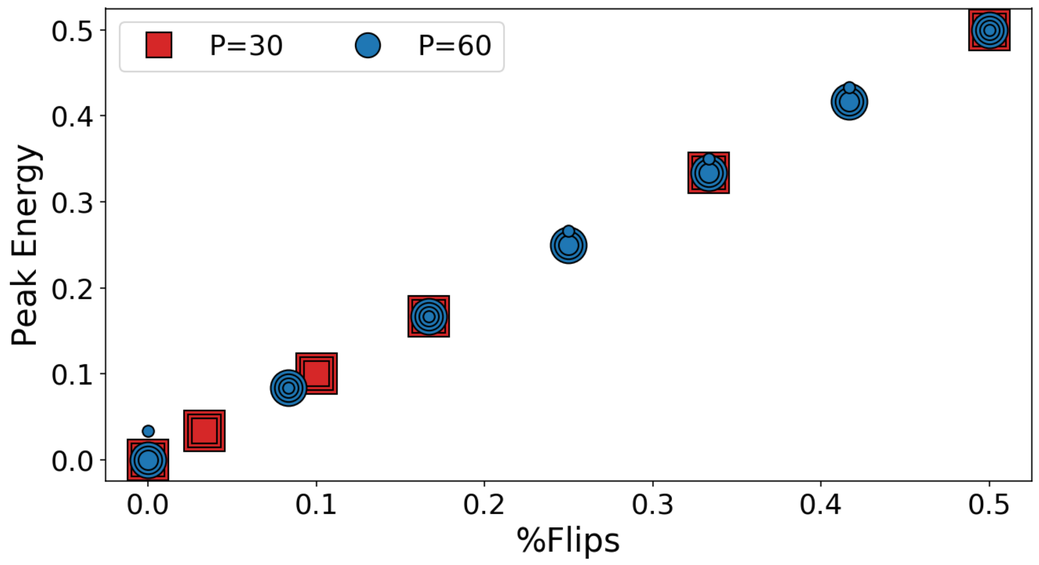
Silico dataset



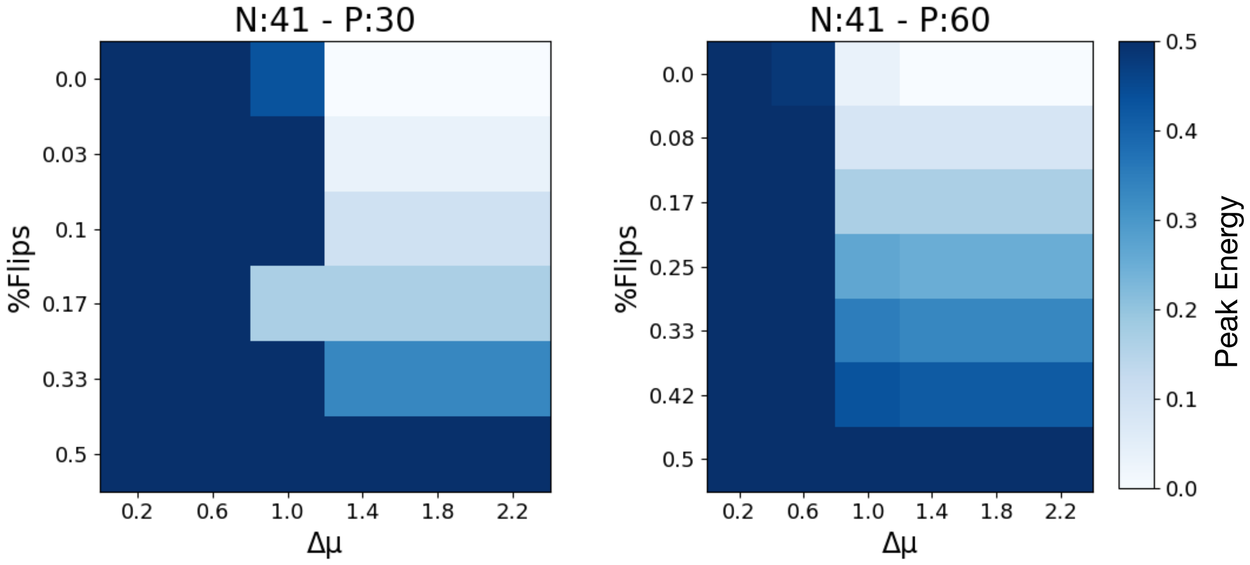


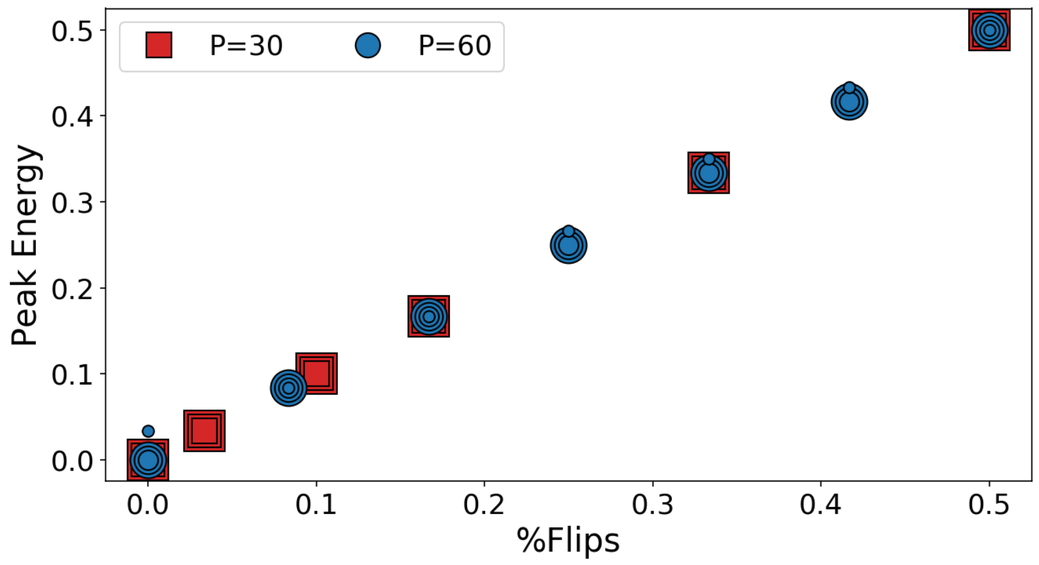
Silico dataset

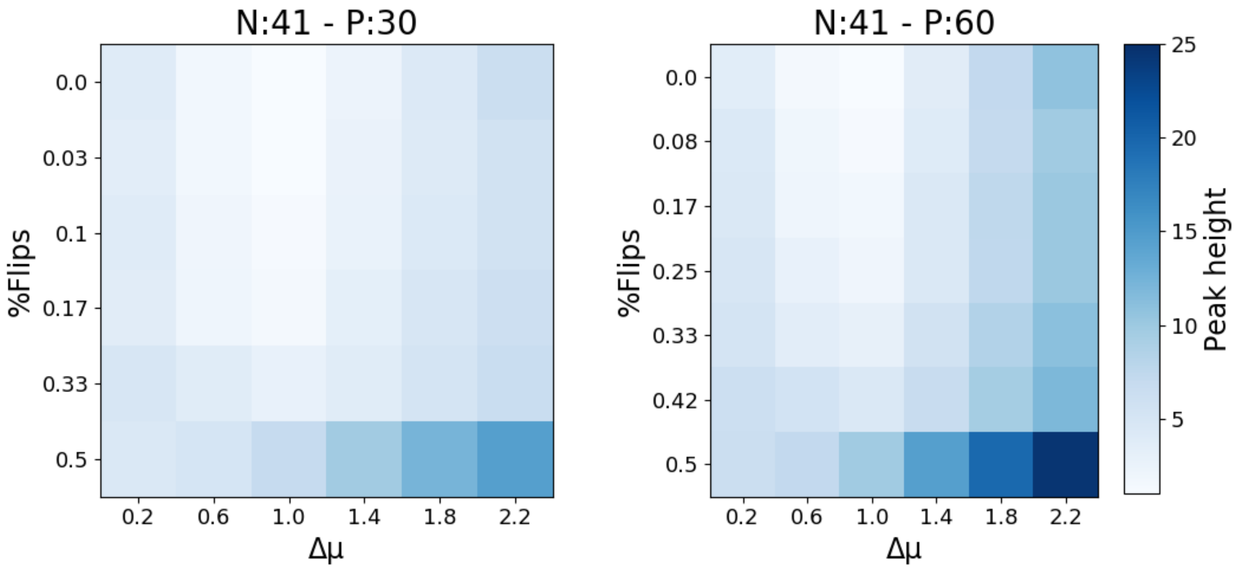




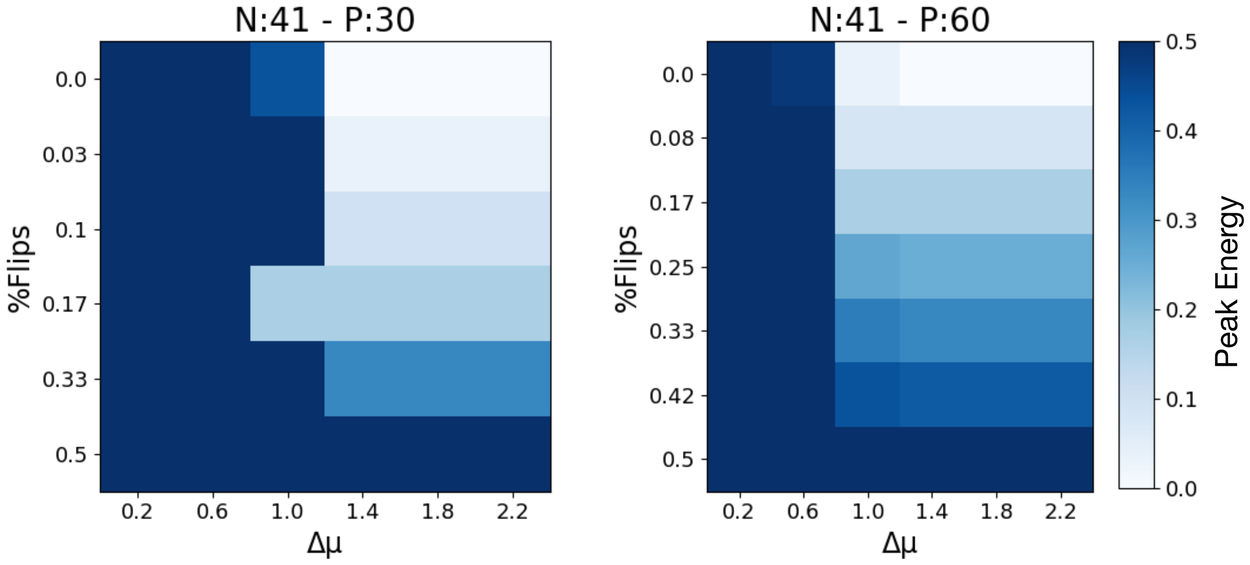
Silico dataset

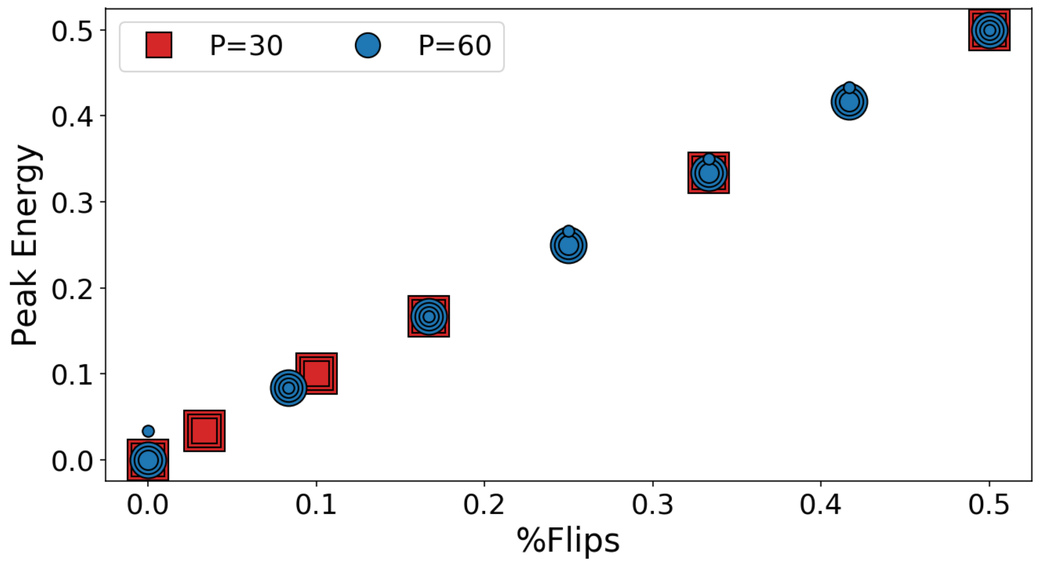


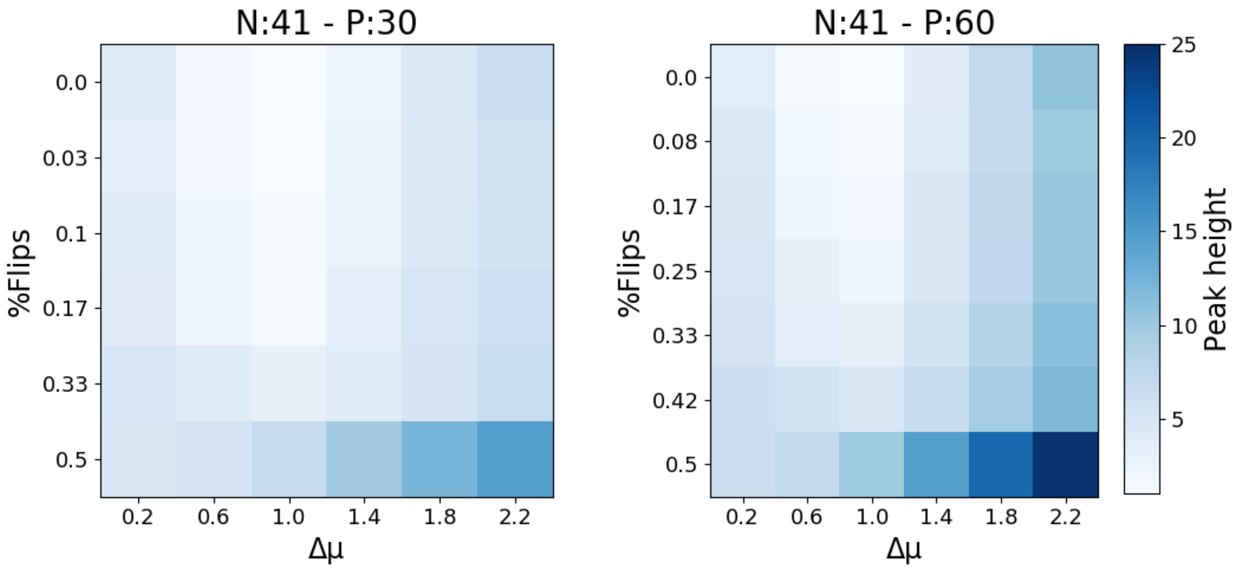


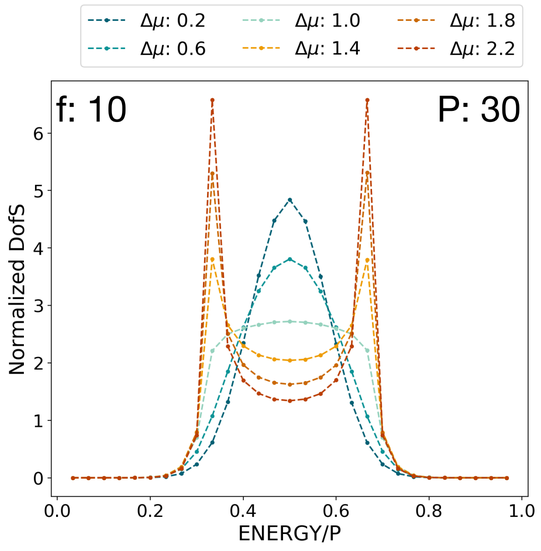


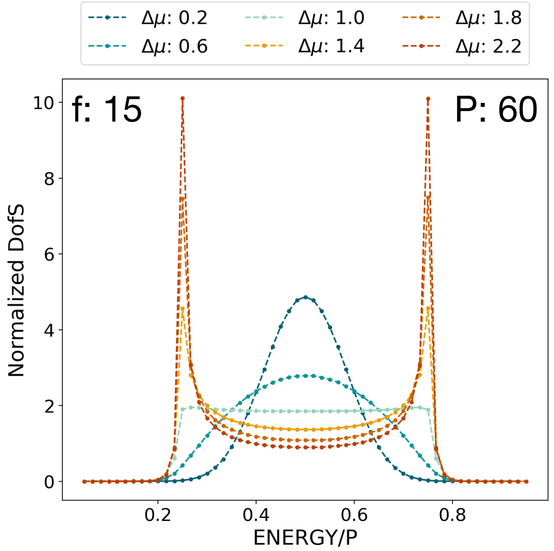
Silico dataset



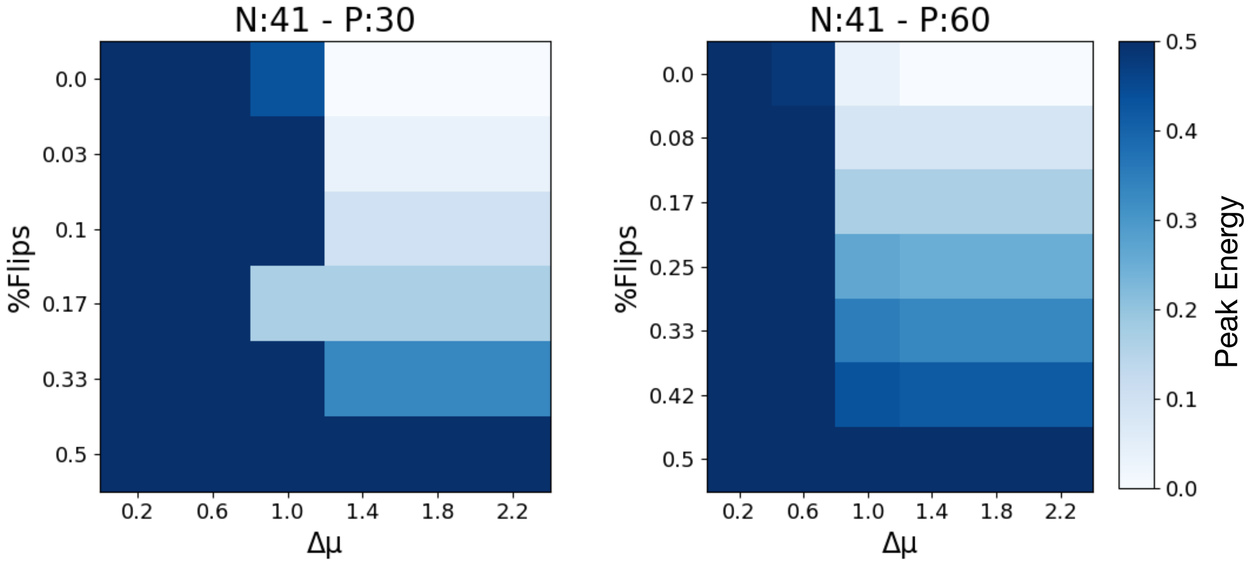


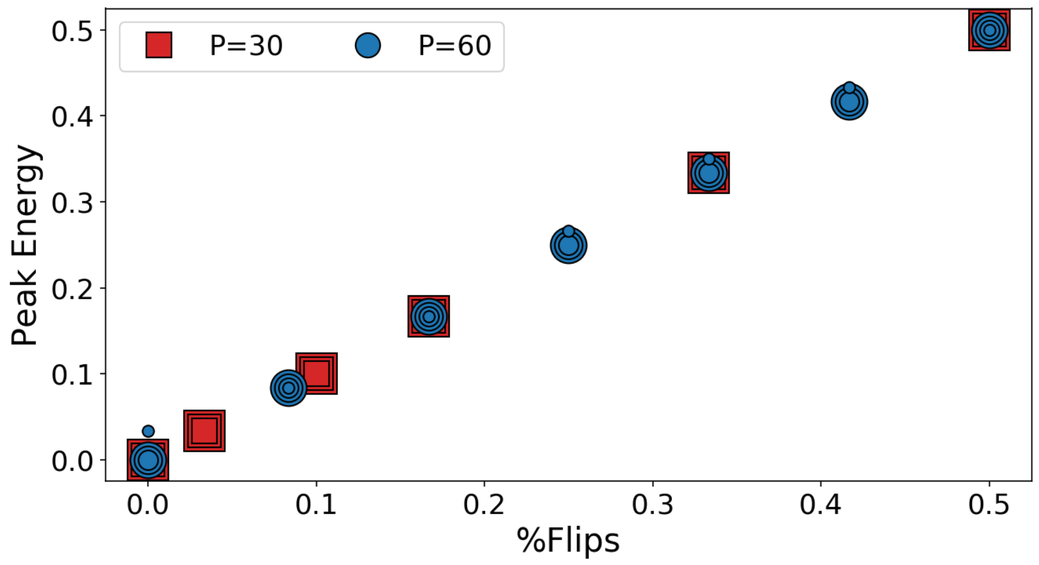


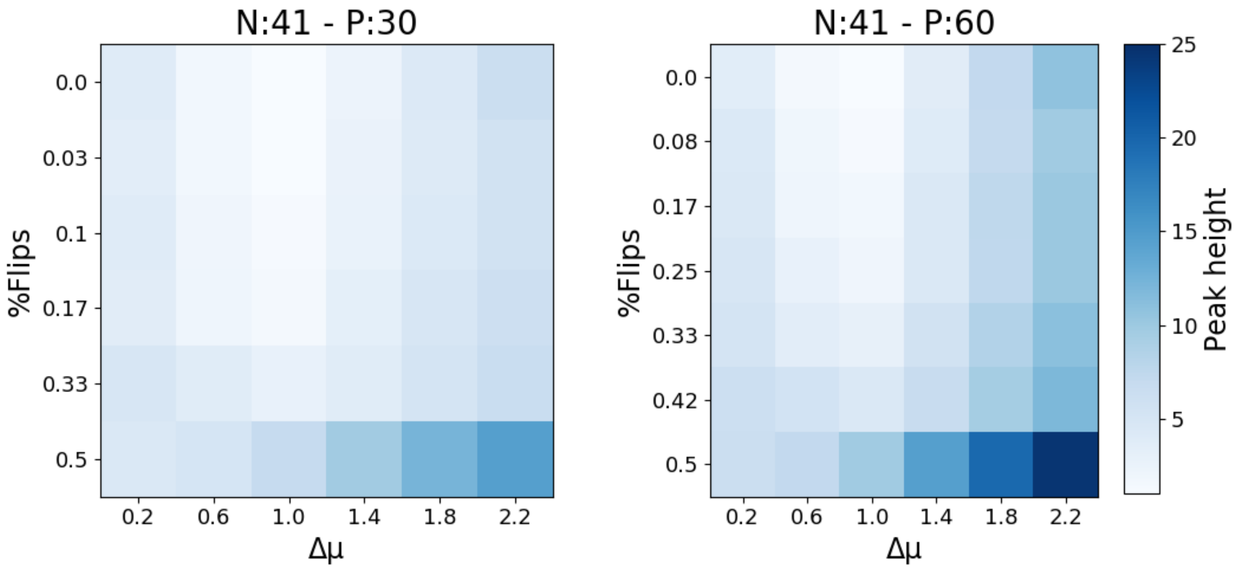


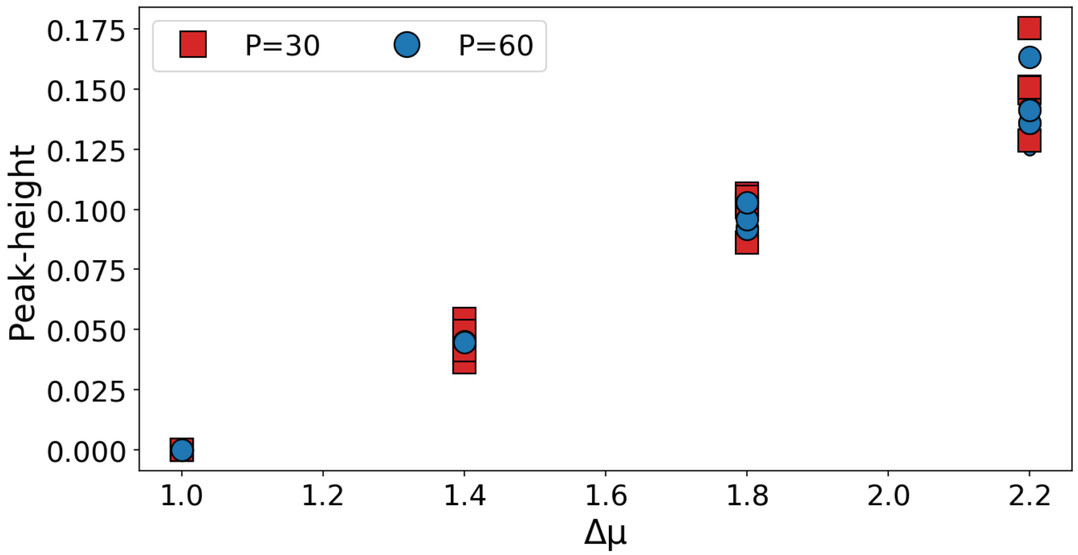


Silico dataset









Conclusions

- Thermodynamic characterisation of the system at any energy level
- Possibility to study real learning problems
- Input-output correlation structure directly impacts the density of states of learning problems
- Insight into the learning process due to a large pull of solutions and higher energy states

UNIVERSITÀ DI TRENTO

INFN | TIFPA

FARE RICERCA IN ITALIA

MIUR MINISTERO DELL'ISTRUZIONE, DELL'UNIVERSITA' E DELLA RICERCA

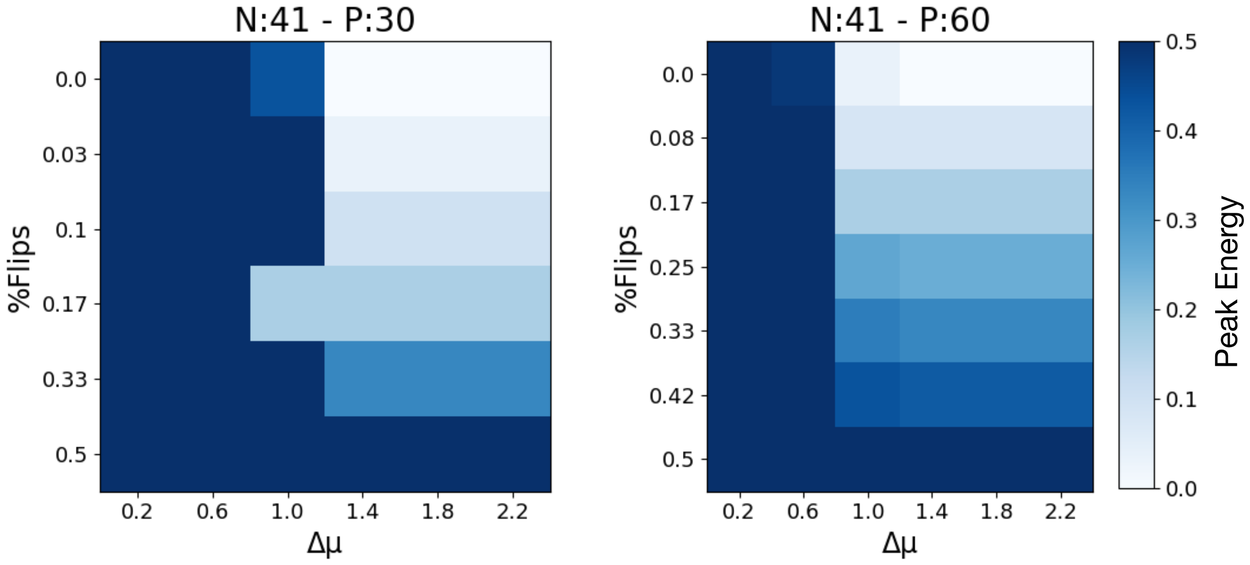
HAMOCK

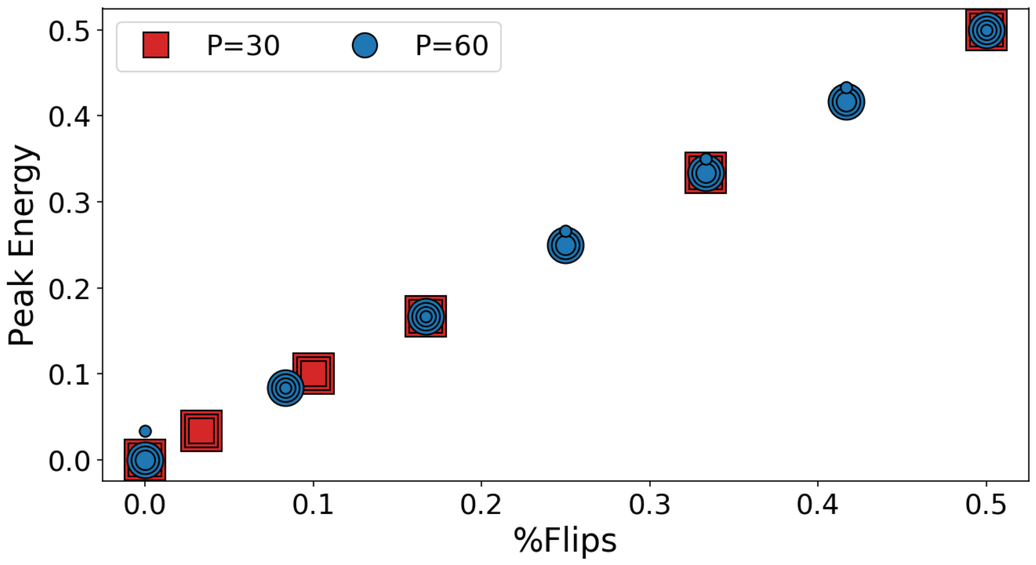
FONDAZIONE CASSA RURALE DI TRENTO

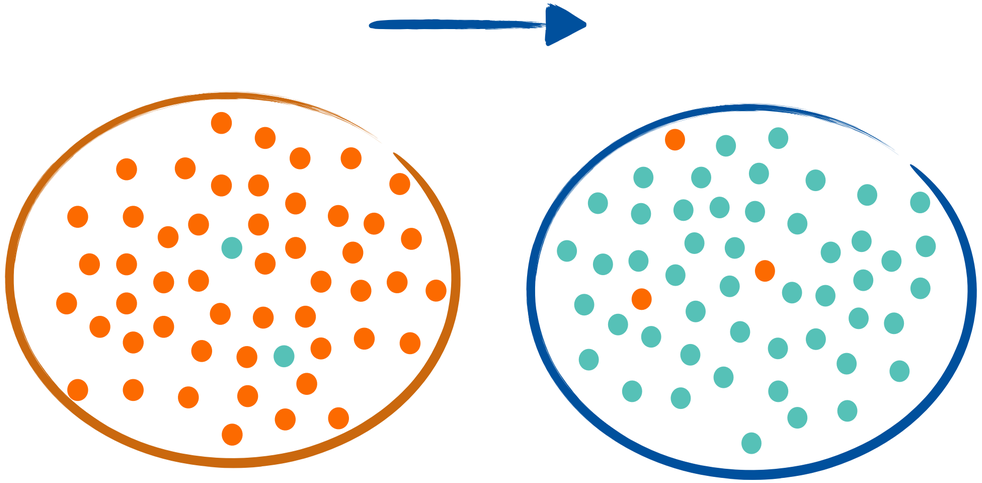
ICTP The Abdus Salam International Centre for Theoretical Physics

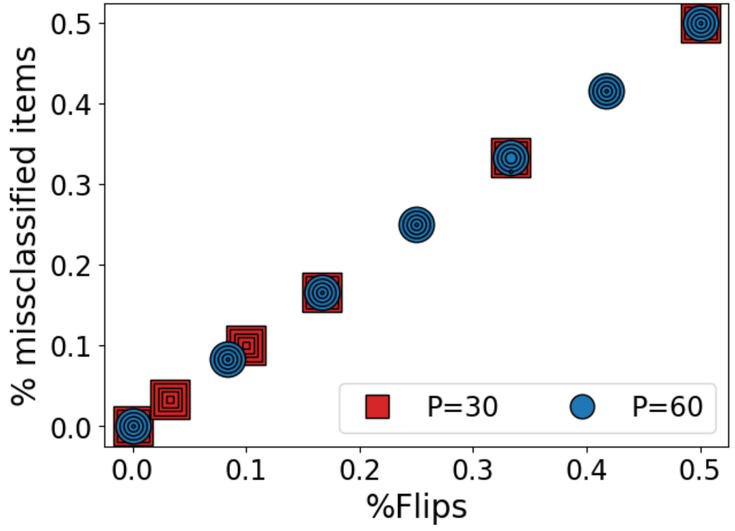
IAEA

Silico dataset

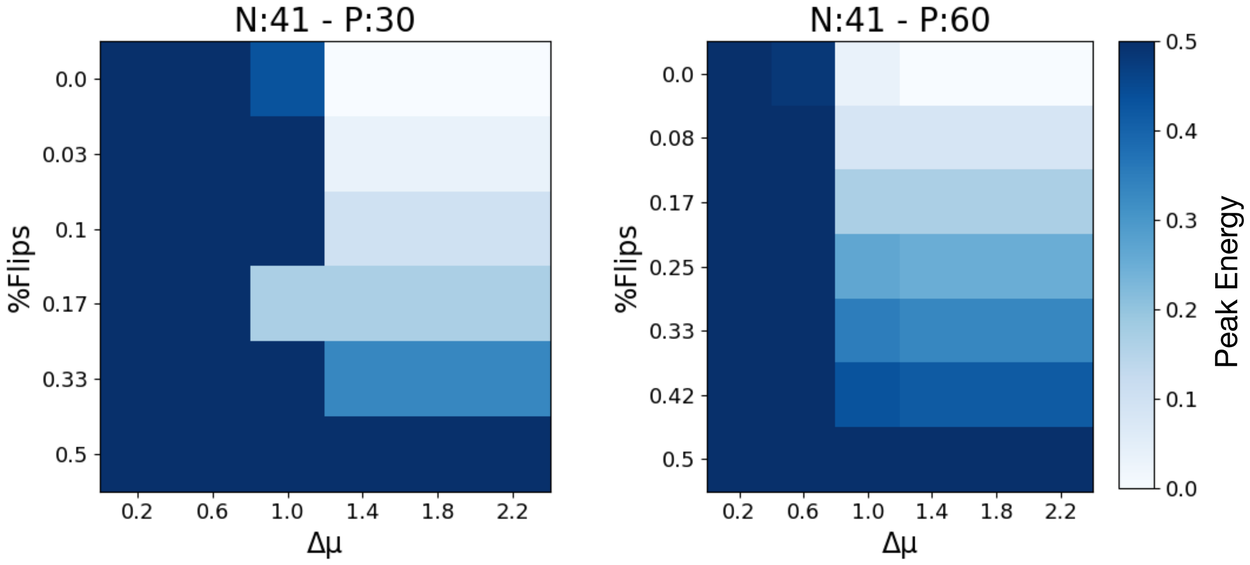


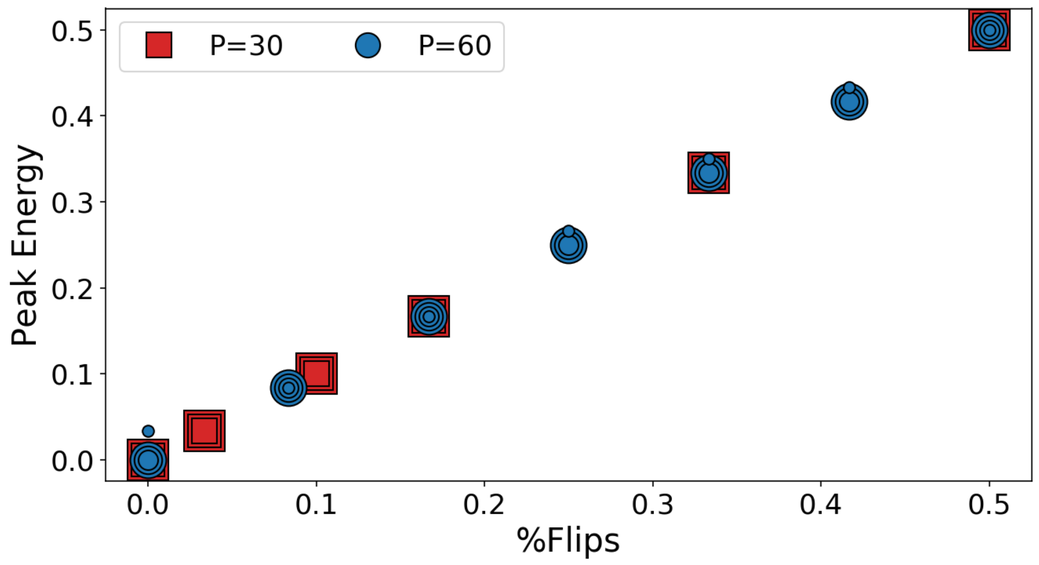


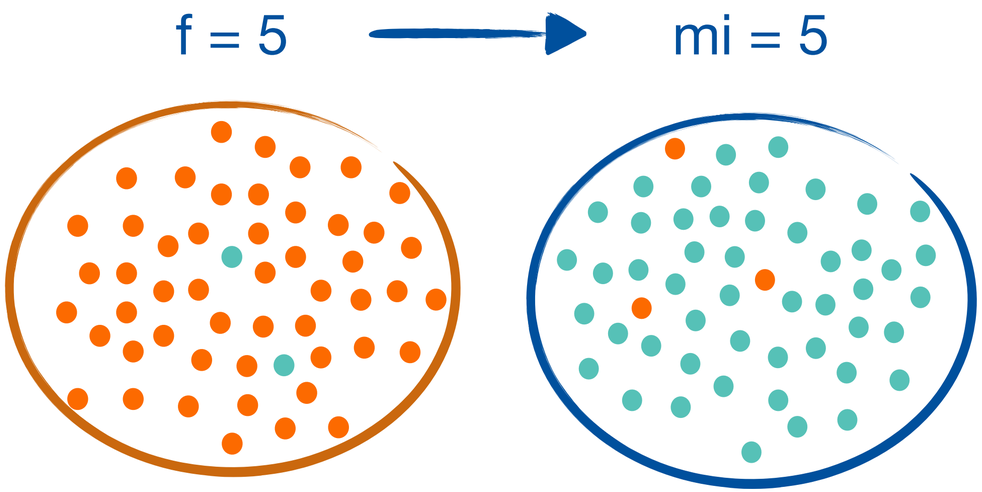


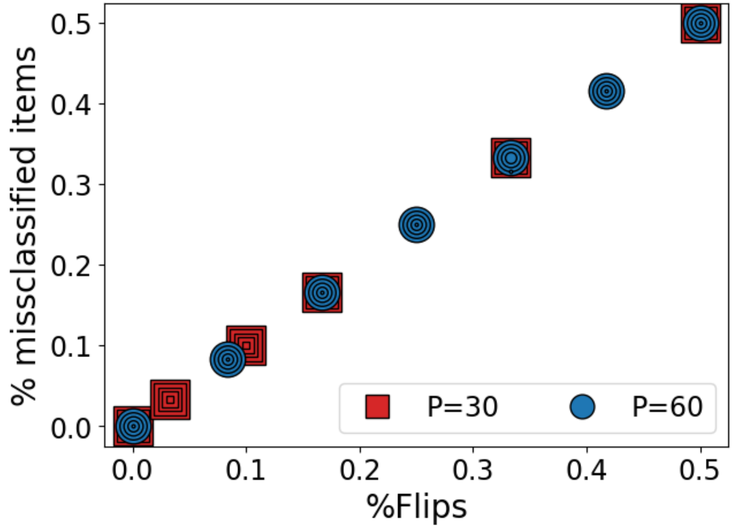
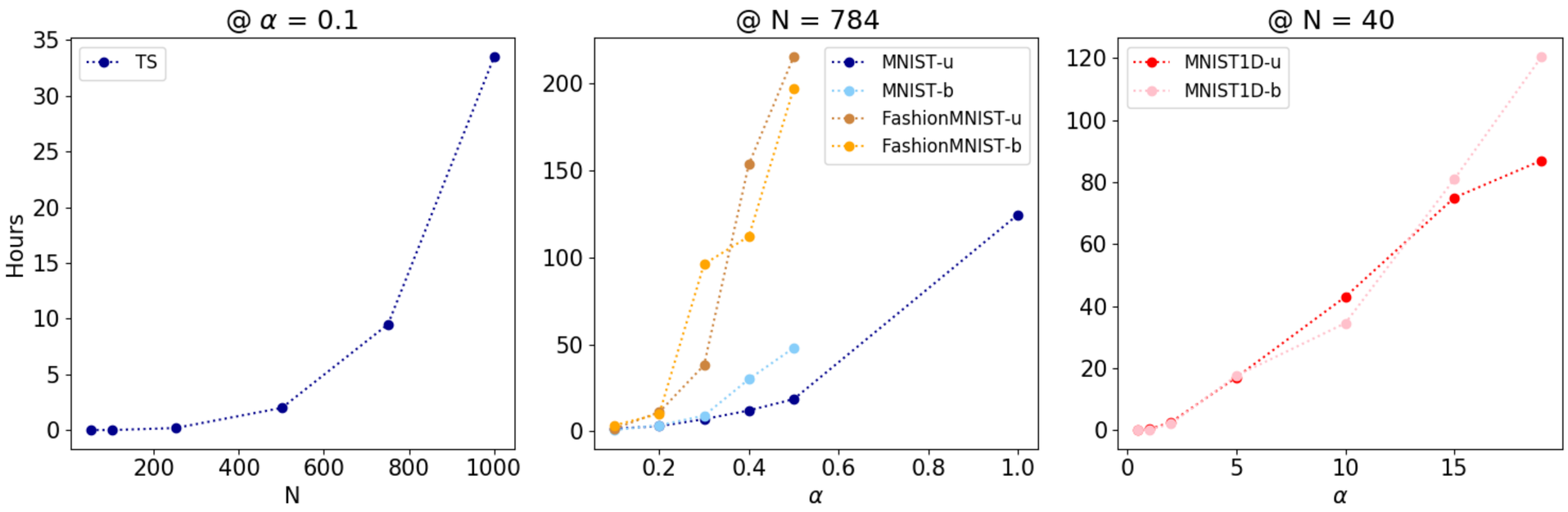


Silico dataset









Convergence time