Bregman divergence based em-algorithm and its application to classical and quantum rate distortion theory

arXiv:2201.02447

Masahito Hayashi

Shenzhen Institute for Quantum Science and Engineering,

Southern University of Science and Technology

Graduate School of Mathematics, Nagoya University

Em-algorithm

- em-algorithm is similar to EM (Expectation and Maximization) algorithm, but it is different from EM algorithm.
- em-algorithm is a generalization of Boltzmann machine.
- Generally, em-algorithm is an algorithm to minimize KL-divergence between exponential family and mixture family, which are key concepts of information geometry.

Rate distortion theory

- Data compression method for analogue data
- We can consider its quantum analogue.
- To make this, we need to solve minimization of mutual information under certain cost constraint.
- Arimoto-Blahut algorithm is known for this aim. But, it minimizes a different quantity, which is a modification of original target function.
- No efficient algorithm exists.

Task of rate distortion theory

Data X^n is generated subject to $P_X{}^n$

Receiver does not need to recover full information X^n.

It is sufficient to recover Y^n such that

$$d(X^n, Y^n) = \sum_{i=1}^{n} d(X_i, Y_i) \leq C$$

$d(x, y)$: error function

Rate distortion theory

Given P_X : distribution on $\mathcal{X}$

Cost function: $d(x,y)$

Conditional distribution: $W \in \mathcal{P}_{Y|X}$

$$\mathcal{P}_{Y|X,c} := \{W \in \mathcal{P}_{Y|X} \mid \sum_{x,y} P_X(x) W(y|x) d(x,y) = c\}$$

Minimum compression rate

$$\min\{I(X;Y)_{W\times P_X} \mid W \in \mathcal{P}_{Y|X,c}\}$$

$$W \times P_X(y,x) := W(y|x) P_X(x)$$

$I(X;Y)_{W\times P_X}$: Mutual information

Mutual information

Mutual information

$$I(X;Y)_{W\times P_X}$$

$$:= H(X)_{W\times P_X} + H(Y)_{W\times P_X} - H(X,Y)_{W\times P_X}$$

Entropy

$$H(X)_{W\times P_X} := -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x)$$

KL-divergence

$$D(P\|Q) := \sum_{x} P(x)(\log P(x) - \log Q(x))$$

Another expression for mutual information

$$I(X;Y)_{W\times P_X} = D(W\times P_X \| W_{Y|P_X}\times P_X)$$

$$W_{Y|P_X}(y) := \sum_{x\in\mathcal{X}} W(y|x)P_X(x)$$

Protocol by rate distortion theory

Assumption: Data X^n obeys $P_X{}^n$.

Code construction (Random coding method)

We randomly choose $M = e^{nI(X;Y)_{W\times P_X}}$ elements $y(1),\ldots,y(M)$ from $\mathcal{Y}^n$.

Encoding

For data X^n, encoder choose K as

$$K := \arg\min_k d(X^n, y(k))$$

Encoder sends K to receiver.

Decoding

Receiver converts K to $y(K)$.

Existing method for rate distortion

Minimization (Arimoto-Blahut)

$$\min_{W \in \mathcal{P}_{Y|X}} I(X;Y)_{W \times P_X} + s \sum_{x,y} P_X(x) W(y|x) d(x,y)$$

If s is a suitable value, the minimizer satisfies the condition;

$$\sum_{x,y} P_X(x) W(y|x) d(x,y) = c$$

However, it is not so easy to find such *s*.

Information geometry for probability distributions

Exponential family

$$P_\theta(x) := P_0(x)\exp\Big(\sum_{i=1}^{d}\theta^i f_i(x) - \mu(\theta)\Big)$$

$$\mathcal{E} := \{P_\theta \mid \theta \in \Theta\}$$

$$\mu(\theta) := \log \sum_{x\in\mathcal{X}} P_0(x)\exp\Big(\sum_{i=1}^{d}\theta^i f_i(x)\Big)$$

Mixture family

$$\mathcal{M} := \{P \mid \sum_{x\in\mathcal{X}} P(x) f_i(x) = c_i\}$$

with constants $c_1, \ldots, c_d$

Information geometry based on Bregman divergence

Information geometry structure can be recovered only by a convex function $\mu(z)$ defined on a convex set $\Theta \subset \mathbb{R}^m$.

Exponential family: $\mathcal{E} := \left\{ e_0 + \sum_{i=1}^{d} \theta^i e_i \right\} \subset \Theta$

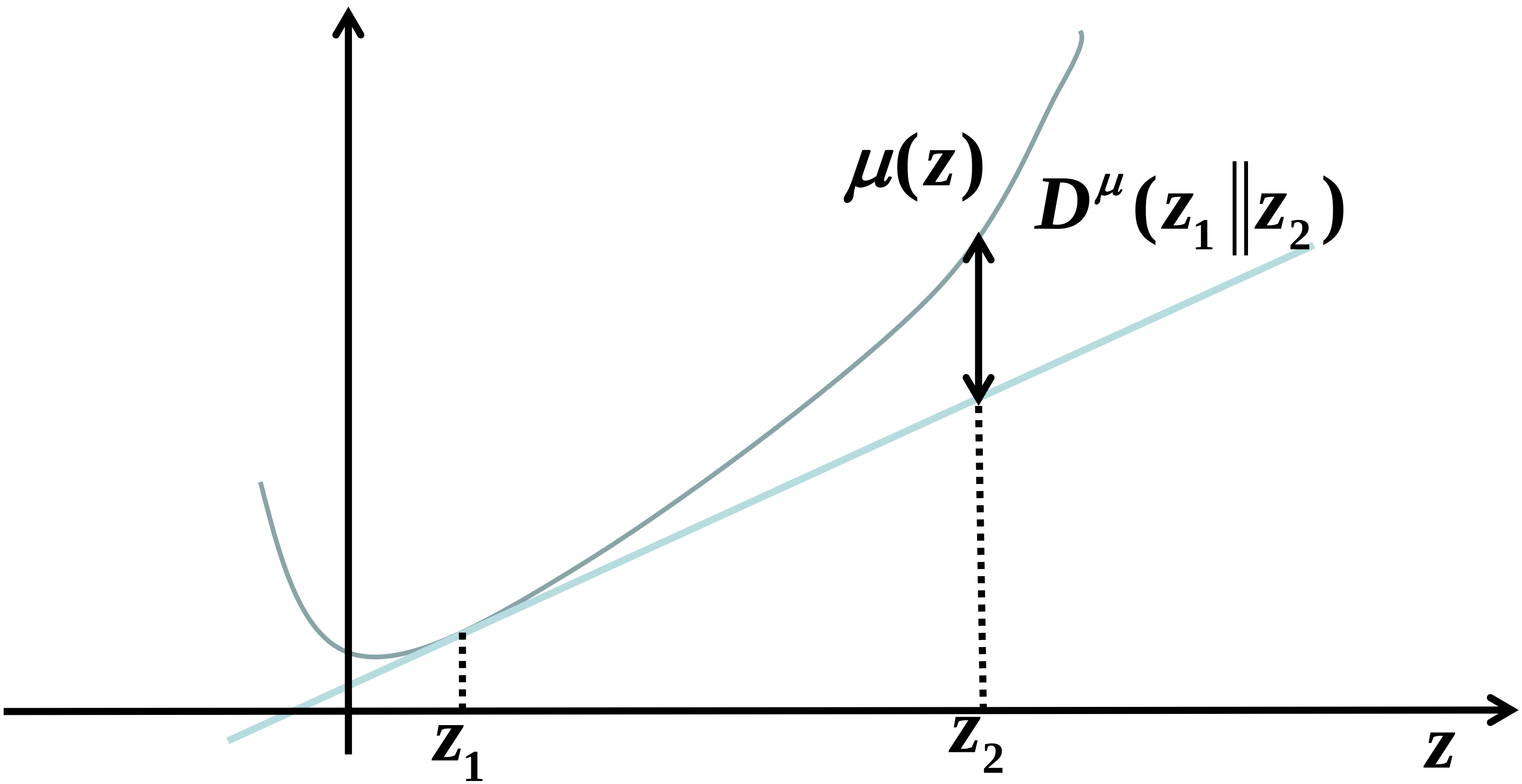
Mixture family $\mathcal{M} := \left\{ z_0 \in \Theta \,\middle|\, \sum_{j=1}^{m} e_i^j \frac{\partial \mu}{\partial z^j} = c_i \right\}$

Bregman divergence

$$D^{\mu}(z_1 \| z_2) := \sum_{i=1}^{m} \frac{\partial \mu}{\partial z^i}(z_1)(z_1{}^i - z_2{}^i) - \mu(z_1) + \mu(z_2)$$

Information geometry based on Bregman divergence

$$D^{\mu}(z_1 \| z_2) := \sum_{i=1}^{m} \frac{\partial \mu}{\partial z^i}(z_1)(z_1{}^i - z_2{}^i) - \mu(z_1) + \mu(z_2)$$



Information geometry for probability distributions

Exponential family

$$P_\theta(x) := P_0(x)\exp\left(\sum_{i=1}^{d}\theta^i f_i(x) - \mu(\theta)\right)$$

$$\mathcal{E} := \{P_\theta \mid \theta \in \Theta\}$$

$$\mu(\theta) := \log \sum_{x\in\mathcal{X}} P_0(x)\exp\left(\sum_{i=1}^{d}\theta^i f_i(x)\right)$$

Mixture family

$$\mathcal{M} := \{P \mid \sum_{x\in\mathcal{X}} P(x) f_i(x) = c_i\}$$

with constants $c_1,\ldots,c_d$

The above is recovered by Bregman divergence system. $D^\mu(\theta_1 \| \theta_2) = D(P_{\theta_1} \| P_{\theta_2})$

Information geometry for quantum states

Exponential family

$$\rho_\theta := \exp(X_0 + \sum_{i=1}^d \theta^i X_i - \mu(\theta))$$

$$\mathcal{E} := \{\rho_\theta \mid \theta \in \Theta\}$$

$$\mu(\theta) := \log \mathrm{Tr} \exp(X_0 + \sum_{i=1}^d \theta^i X_i)$$

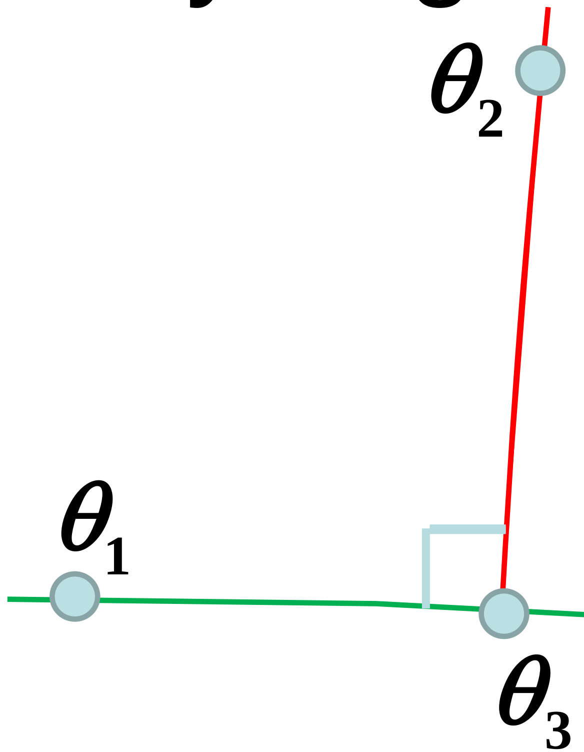
Mixture family $\mathcal{M} := \{\rho \mid \mathrm{Tr}\rho X_i = c_i\}$

with constants $c_1, \ldots, c_d$

The above is recovered by Bregman divergence system.

$$D^\mu(\theta_1 \| \theta_2) = D(\rho_{\theta_1} \| \rho_{\theta_2}) := \mathrm{Tr}\rho_{\theta_1}(\log \rho_{\theta_1} - \log \rho_{\theta_2})$$

Pythagorean theorem



Exponential family

$$\mathcal{E} := \left\{ e_0 + \sum_{i=1}^{d} \theta^i e_i \right\} \subset \Theta$$

$$\mathcal{M} := \left\{ z_0 \in \Theta \,\middle|\, \sum_{j=1}^{m} e_i^j \frac{\partial \mu}{\partial z^j} = c_i \right\}$$

Mixture family

$$D^\mu(\theta_1 \| \theta_2) = D^\mu(\theta_1 \| \theta_3) + D^\mu(\theta_3 \| \theta_2)$$

e-Projection

$$\Gamma_{\mathcal{E}}^{(e)}(\theta_1) := \theta_3$$

m-Projection

$$\Gamma_{\mathcal{M}}^{(m)}(\theta_2) := \theta_3$$

em-algorithm

$$\min_{\theta_2 \in \mathcal{E}} \min_{\theta_1 \in \mathcal{M}} D^{\mu}(\theta_1 \| \theta_2)$$

em-algorithm is an iterative algorithm.

We set initial point $\theta_{2(1)} \in \mathcal{E}$

m-step $\theta_{1(t+1)} := \arg\min_{\theta_1 \in \mathcal{M}} D^{\mu}(\theta_1 \| \theta_{2(t)})$

e-step $\theta_{2(t+1)} := \arg\min_{\theta_2 \in \mathcal{E}} D^{\mu}(\theta_{1(t+1)} \| \theta_2)$

However, the convergence to the global minimum has not been discussed.

em-algorithm

$$\min_{\theta_2 \in \mathcal{E}} \min_{\theta_1 \in \mathcal{M}} D^{\mu}(\theta_1 \| \theta_2)$$

em-algorithm is an iterative algorithm.

We set initial point $\theta_{2(1)} \in \mathcal{E}$

m-step $\theta_{1(t+1)} := \arg\min_{\theta_1 \in \mathcal{M}} D^{\mu}(\theta_1 \| \theta_{2(t)})$

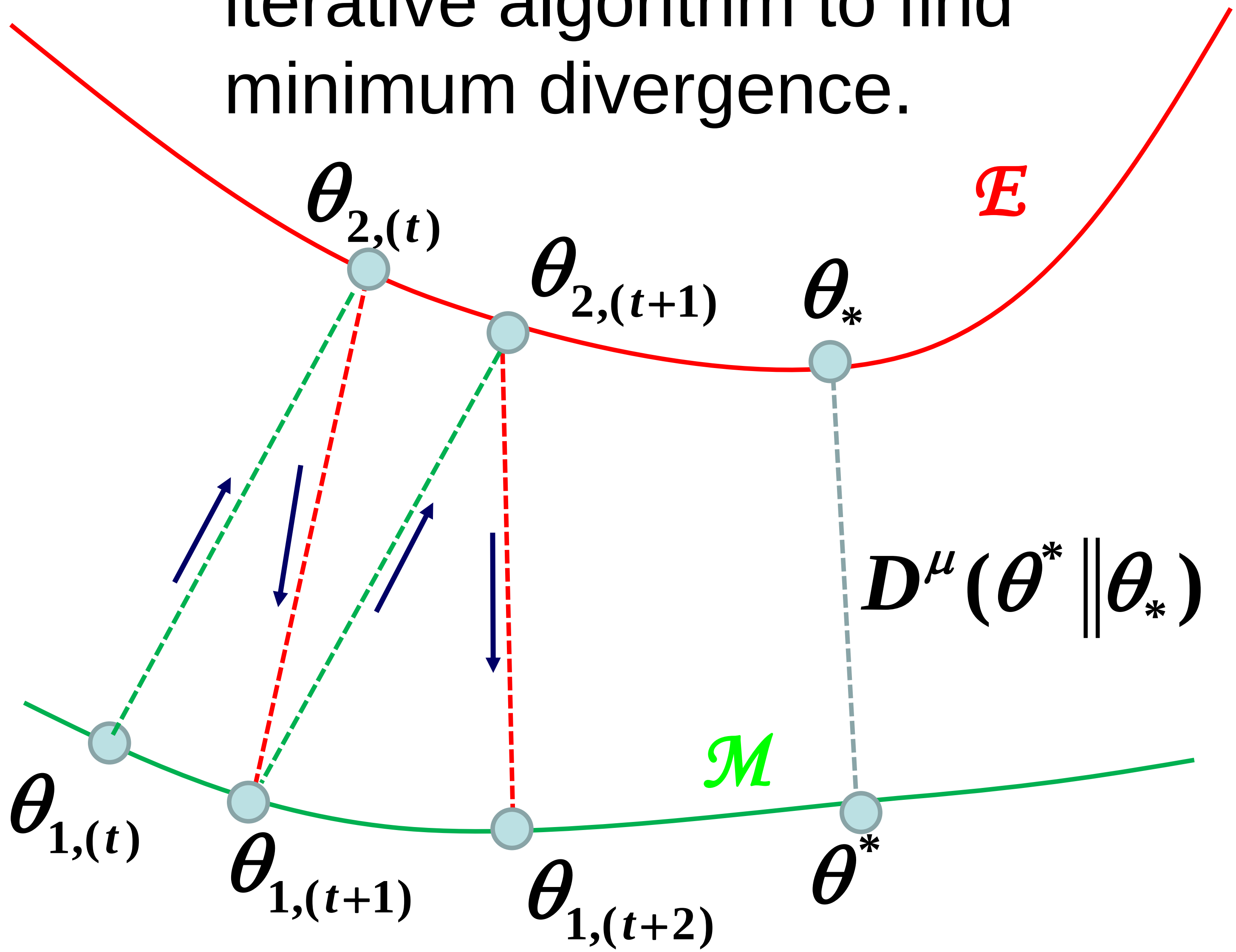
e-step $\theta_{2(t+1)} := \arg\min_{\theta_2 \in \mathcal{E}} D^{\mu}(\theta_{1(t+1)} \| \theta_2)$

$$D^{\mu}(\theta_{1(t+1)} \| \theta_{2(t+1)}) \le D^{\mu}(\theta_{1(t+1)} \| \theta_{2(t)}) \le D^{\mu}(\theta_{1(t)} \| \theta_{2(t)})$$

However, the convergence to the global minimum has not been discussed.

em-algorithm

iterative algorithm to find minimum divergence.



em-algorithm

Theorem

$$D^{\mu}(\theta_1 \| \theta_2) \geq D^{\mu}(\Gamma^{(e)}_{\mathcal{E}}(\theta_1) \| \Gamma^{(e)}_{\mathcal{E}}(\theta_2)) \qquad \theta_1, \theta_2 \in \mathcal{M}$$

$(\theta_{1(t)}, \theta_{2(t)})$ converges to global minimum.

Convergence speed

$$D^{\mu}(\theta_{1(t)} \| \Gamma^{(e)}_{\mathcal{E}}(\theta_{1,t})) - \min_{\theta \in \mathcal{M}} D^{\mu}(\theta \| \Gamma^{(e)}_{\mathcal{E}}(\theta))$$

$$\leq \frac{\sup_{\theta \in \mathcal{M}} D^{\mu}(\theta \| \theta_{2(1)})}{t-1}$$

Rate distortion theory

Given P_X : distribution on $\mathcal{X}$

Cost function: $d(x, y)$

Conditional distribution: $W \in \mathcal{P}_{Y|X}$

$$\mathcal{P}_{Y|X,c} := \{W \in \mathcal{P}_{Y|X} \mid \sum_{x,y} P_X(x) W(y|x) d(x,y) = c\}$$

Minimum compression rate

$$\min\{I(X;Y)_{W \times P_X} \mid W \in \mathcal{P}_{Y|X,c}\}$$

$$W \times P_X(y, x) := W(y|x) P_X(x)$$

$I(X;Y)_{W \times P_X}$: Mutual information

Application of em-algorithm to rate-distortion theory

$$\min_{W \in \mathcal{P}_{Y|X,c}} I(X;Y)_{W \times P_X}$$

$$= \min_{W \in \mathcal{P}_{Y|X,c}} D(W \times P_X \| W_{Y|P_X} \times P_X)$$

$$= \min_{W \in \mathcal{P}_{Y|X,c}} \min_{Q_Y} D(W \times P_X \| Q_Y \times P_X)$$

Mixture family

Exponential family

Convergence condition holds.

$$D(W \times P_X \| W' \times P_X)$$

$$\geq D(W_{Y|P_X} \| W_{Y|P_X}') = D(W_{Y|P_X} \times P_X \| W_{Y|P_X}' \times P_X)$$

Application of em-algorithm to rate-distortion theory

E-step can be done by calculating the marginal distribution $P_Y^{(t)} := \sum_{x \in \mathcal{X}} P_{Y|X}^{(t-1)}(y \mid x) P_X(x)$

M-step needs to solve the following for τ

$$\frac{\partial}{\partial \tau} \underbrace{\sum_{x \in \mathcal{X}} P_X(x) \log\left(\sum_{y \in \mathcal{Y}} P_Y^{(t)}(y) e^{\tau d(x,y)} \right)}_{\text{convex function}} = c$$

$$P_{Y|X}^{(t)}(y \mid x) := P_Y^{(t)}(y) e^{\tau d(x,y)} \Big(\sum_{y \in \mathcal{Y}} P_Y^{(t)}(y) e^{\tau d(x,y)} \Big)^{-1}$$

We repeat this process.

Numerical calculation

$$\begin{pmatrix} P_X(1) \\ P_X(2) \\ P_X(3) \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \end{pmatrix} \qquad \begin{pmatrix} d(1,1) & d(1,2) & d(1,3) \\ d(2,1) & d(2,2) & d(2,3) \\ d(3,1) & d(3,2) & d(3,3) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

$$c = 1.5$$

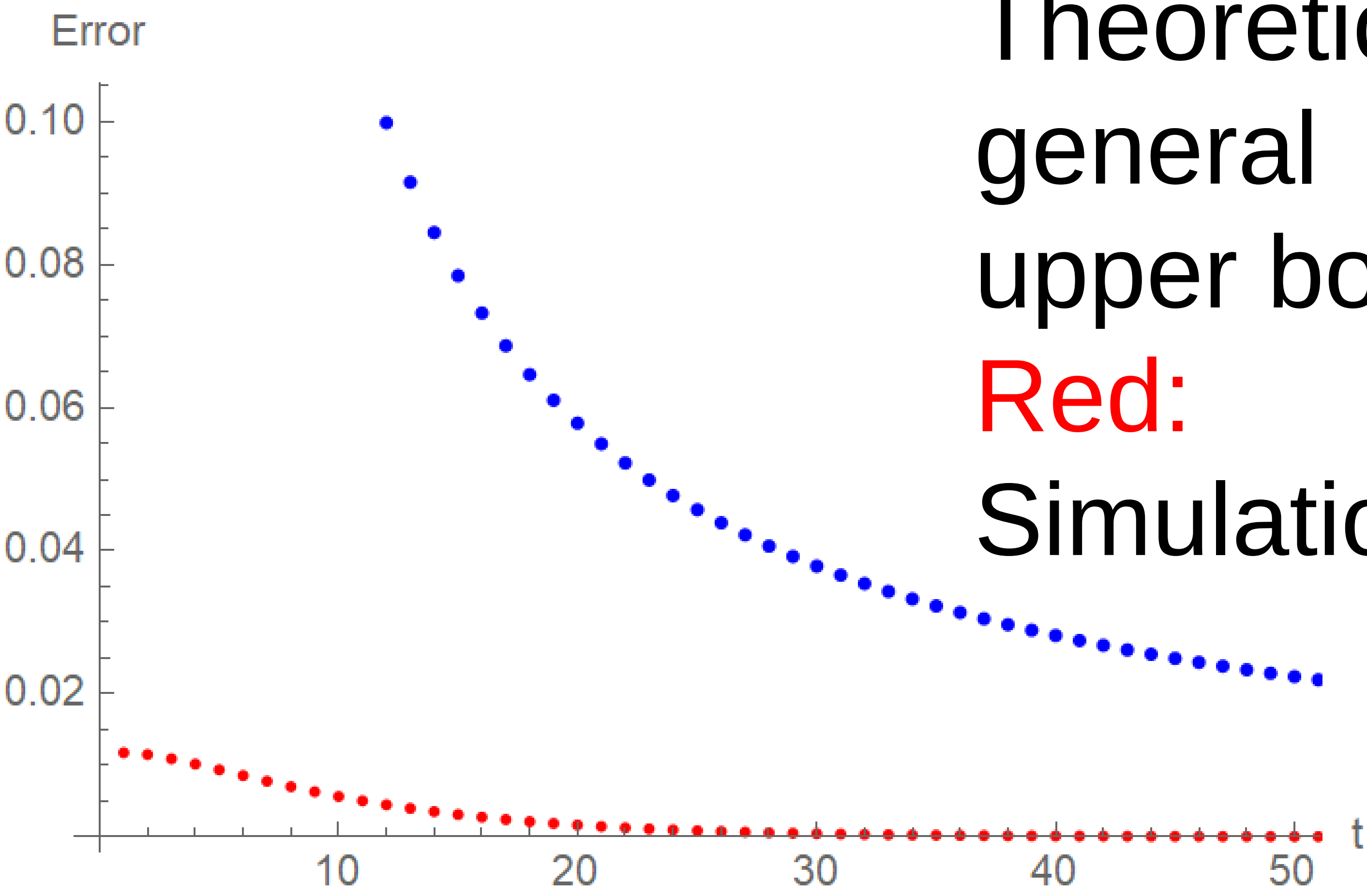
Solution

$$P^*_{Y|X} = \begin{pmatrix} 0.0856 & 0.1886 & 0.4310 \\ 0.2243 & 0.4944 & 0.1296 \\ 0.6901 & 0.3170 & 0.4294 \end{pmatrix}$$

$$I(X;Y)_{P^*_{Y|X} \times P_X} = 0.100039$$

Numerical calculation

$$I(X;Y)_{W^{(t)}\times P_X} - I(X;Y)_{P^*_{Y|X}\times P_X}$$

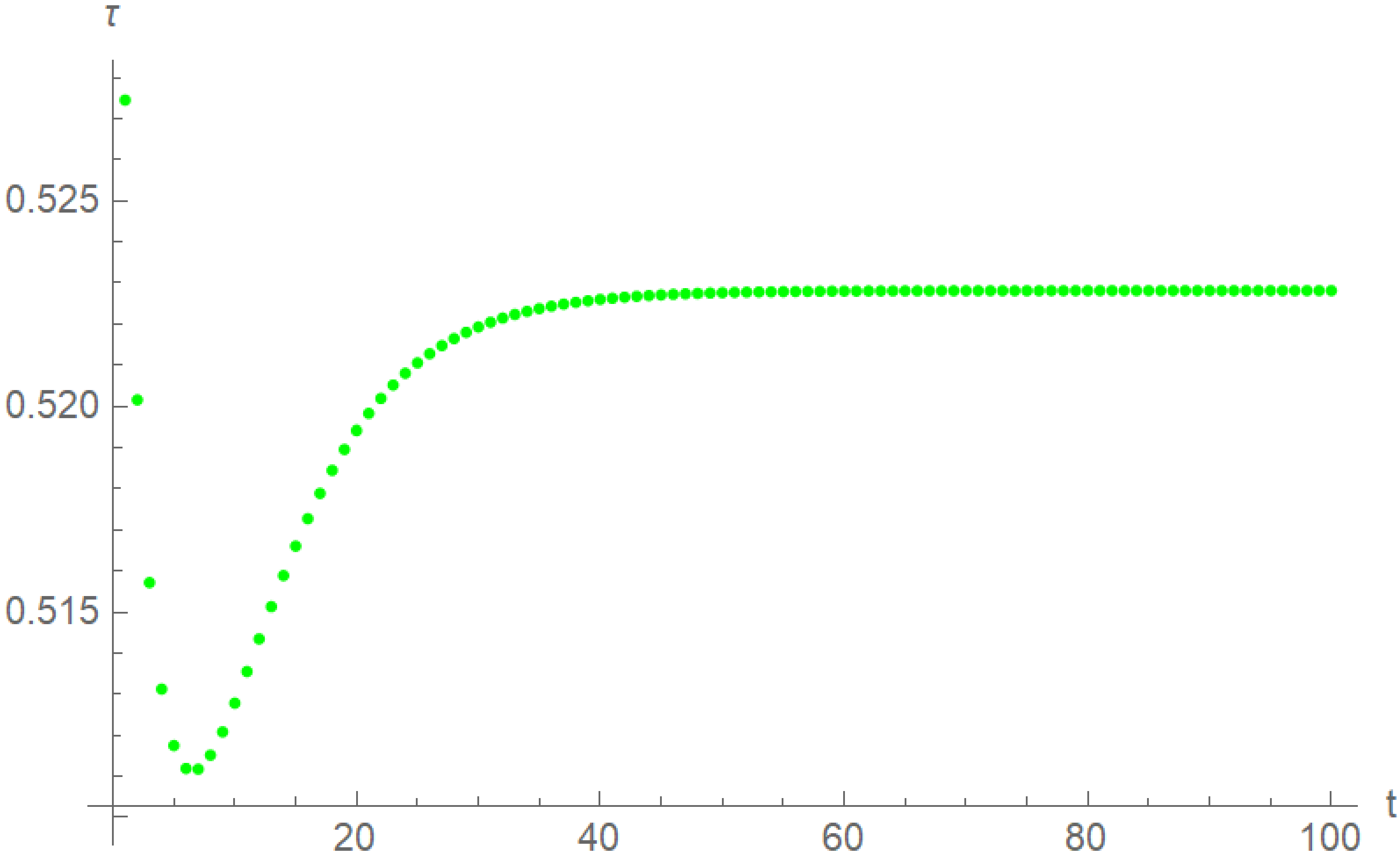


Blue: Theoretical general upper bound

Red: Simulation

Numerical calculation

Behavior of τ



Further results

arXiv:2201.02447

(1) We made various types of evaluations on em-algorithm.

(2) We applied em-algorithm to several variants of rate-distortion theory including the quantum setting.

Conclusion

- We have studied the convergence of em-algorithm under the framework of Bregman divergence.
- We have applied our result to the rate-distortion theory.
- Our algorithm rapidly converges to the true value.

References

[1] S. Amari, “Information geometry of the EM and em algorithms for neural networks,” Neural Networks **8**, 1379 – 1408 (1995).

[2] S. Amari, K. Kurata and H. Nagaoka, “Information geometry of Boltzmann machines,” IEEE Transactions on Neural Networks, **3**, 260 – 271, (1992).

[3] R. Blahut, “Computation of channel capacity and rate-distortion functions,” IEEE Trans. IT, **18**, 460 – 473 (1972).

[4] S. Arimoto, “An algorithm for computing the capacity of arbitrary discrete memoryless channels,” IEEE Trans. IT, **18**, 14 – 20 (1972).

[5] N. Datta, M.-H. Hsieh, and M. M. Wilde, “Quantum rate distortion, reverse Shannon theorems, and source-channel separation,” IEEE Trans. IT, **59**, 615 – 630 (2013).

[6] S. Toyota, “Geometry of Arimoto algorithm,” Information Geometry **3**, 183 (2020).