Information Geometry
https://doi.org/10.1007/s41884-024-00140-5

RESEARCH PAPER

Iterative minimization algorithm on a mixture family

Masahito Hayashi[1,2,3]

Received: 9 May 2022 / Revised: 10 May 2024 / Accepted: 13 July 2024

Abstract
Iterative minimization algorithms appear in various areas including machine learning, neural networks, and information theory. The em algorithm is one of the famous iterative minimization algorithms in the area of machine learning, and the Arimoto–Blahut algorithm is a typical iterative algorithm in the area of information theory. However, these two topics had been separately studied for a long time. In this paper, we generalize an algorithm that was recently proposed in the context of the Arimoto–Blahut algorithm. Then, we show various convergence theorems, one of which covers the case when each iterative step is done approximately. Also, we apply this algorithm to the target problem of the em algorithm, and propose its improvement. In addition, we apply it to other various problems in information theory.

Keywords Minimization · Em algorithm · Mixture family · Channel capacity · Divergence

1 Introduction

Optimization over distributions is an important topic in various areas. For example, the minimum divergence between a mixture family and an exponential family has been studied by using the em algorithm in the areas of machine learning and neural networks [1–4]. The em algorithm is an iterative algorithm to calculate the above minimization and it is rooted in the study of Boltzmann machines [5]. In particular, the paper [3] formulated the em algorithm under the framework with Bregman divergence [6, 7]. The topic of the em algorithm has been mainly studied in the community of machine

Communicated by Nihat Ay.

✉ Masahito Hayashi
hmasahito@cuhk.edu.cn; hayashi@iqasz.cn; masahito@math.nagoya-u.ac.jp

[1] School of Data Science, The Chinese University of Hong Kong, Shenzhen, Longgang District, Shenzhen 518172, Shenzhen, China

[2] International Quantum Academy (SIQA), Futian District, Shenzhen 518048, China

[3] Graduate School of Mathematics, Nagoya University, Chikusa-ku 464-8602, Nagoya, Japan

Published online: 13 August 2024

learning, neural networks, and information geometry. As another iterative algorithm, the Arimoto–Blahut algorithm is known as an algorithm to maximize the mutual information by changing the distribution on the input system [8, 9]. This maximization is needed to calculate the channel capacity [10]. This algorithm has been generalized to various settings including the rate distortion theory [9, 11–13], the capacity of a wire-tap channel [14], and their quantum extensions [15–19]. In particular, the two papers [13, 19] made very useful generalizations to cover various topics in information theory. The Arimoto–Blahut algorithm and its variants have been mainly studied in the community of information theory.

However, only a limited number of studies have discussed the relation between the two topics, the em algorithm and the Arimoto–Blahut algorithm, as follows. The papers [23, 24] pointed out that the Arimoto–Blahut algorithm can be considered as an alternating algorithm in a similar way to the EM and the em algorithms. Recently, the paper [20] pointed out that the maximization of the mutual information can be considered to be the maximization of the projected divergence to an exponential family by changing an element of a mixture family. The paper [21] generalized this maximization to the framework with Bregman divergence [6, 7] and applied this setting to various problems in information theory. Also, the recent paper [22] applied the em algorithm to the rate-distortion theory, which is a key topic in information theory.

In this paper, we focus on a generalized problem setting proposed in [19], which is given as an optimization over the set of input quantum states. As the difference from the former algorithm, their algorithm [19] has an acceleration parameter. Changing this parameter, we can enhance the convergence speed under a certain condition. To obtain wider applicability, we extend their problem setting to the minimization over a general mixture family. Although they discussed the convergence speed only when there is no local minimizer, our analysis covers the convergence speed to a local minimizer even when there exist several local minimizers. Further, since our setting covers a general mixture family as the set of input variables, our method can be applied to the minimum divergence between a mixture family and an exponential family, which is the objective problem in the em algorithm. That is, this paper presents a general algorithm including the em algorithm as well as the Arimoto–Blahut algorithm. This type of relation between the em algorithm and the Arimoto–Blahut algorithm is different from the relation pointed by the papers [23, 24].

There is a possibility that each iteration can be calculated only approximately. To cover such an approximated case, we evaluate the error of our algorithm with approximated iterations. Since the em algorithm has local minimizers in general, it is essential to cover the convergence to a local minimizer. Since our algorithm has the acceleration parameter, our application to the minimum divergence gives a generalization of the em algorithm. Also, our algorithm can be applied to the maximization of the projected divergence to an exponential family by changing an element of a mixture family.

In addition, our algorithm has various applications that were not discussed in the preceding study [19]. In channel coding, the decoding block error probability goes to zero exponentially under the proper random coding when the transmission rate is smaller than the capacity [25]. Also, the probability of correct decoding goes to zero exponentially when the transmission rate is greater than the capacity [26]. These exponential rates are written with the optimization of the so-called Gallager function.

Recently, the paper [27] showed that the Gallager function can be written as the minimization of the Rényi divergence. Using this fact, we apply our method to these optimizations. Further, we apply our algorithm to the capacity of a wiretap channel. In addition, since our problem setting allows a general mixture family as the range of input, we apply the channel capacity with cost constraint. Also, we point out that the calculation of the commitment capacity is given as the minimization of the divergence between a mixture family and an exponential family. Hence, we discuss this application as well.

The remaining part of this paper is organized as follows. Section 2 formulates our minimization problem for a general mixture family. Then, we proposed several algorithms to solve the minimization problem. We derive various convergence theorems including the case with approximated iterations. The remaining sections apply our algorithm to various examples. In these sections, examples of objective functions are discussed. Section 3 applies our algorithm to various information theoretical problems. Then, Sect. 4 demonstrates the application to the minimum divergence between a mixture family and an exponential family. Section 5 shows how to apply our algorithm to the commitment capacity. Section 6 discusses the application of our algorithm to the maximization of the projected divergence to an exponential family by changing an element of a mixture family. Section 7 considers the application to information bottleneck, which is a powerful method for machine learning. Appendices are devoted to the proofs of the theorems presented in Sect. 2.

2 General setting

2.1 Algorithm with exact iteration

We consider a finite sample space $\mathcal{X}$ and focus on the set $\mathcal{P}(\mathcal{X})$ of distributions whose support is $\mathcal{X}$. Using k linearly independent functions $f_1, \ldots, f_k$ on $\mathcal{X}$ and constants $a = (a_1, \ldots, a_k)$, we define the mixture family $\mathcal{M}_a$ as follows

$$\mathcal{M}_a := \{P \in \mathcal{P}(\mathcal{X}) | P[f_i] = a_i \text{ for } i = 1, \ldots, k\}, \tag{1}$$

where $P[f] := \sum_{x \in \mathcal{X}} P(x) f(x)$. We add additional $l - k$ linearly independent functions $f_{k+1}, \ldots f_l$ and $|\mathcal{X}| = l + 1$ such that the l functions $f_1, \ldots, f_l$ are linearly independent. Then, the distribution P can be parameterized by the mixture parameter $\eta = (\eta_1, \ldots, \eta_l)$ as $\eta_i = P[f_i]$. That is, the above distribution is denoted by P_η. Then, we denote the e-projection of P to $\mathcal{M}_a$ by $\Gamma^{(e)}_{\mathcal{M}_a}[P]$. That is, $\Gamma^{(e)}_{\mathcal{M}_a}[P]$ is defined as follows [1, 2].

$$\Gamma^{(e)}_{\mathcal{M}_a}[P] := \operatorname*{argmin}_{Q \in \mathcal{M}_a} D(Q\|P), \tag{2}$$

where the Kullback–Leibler divergence $D(Q\|P)$ is defined as

$$D(Q\|P) := \sum_{x \in \mathcal{X}} Q(x)(\log Q(x) - \log P(x)). \tag{3}$$

Using the e-projection, we have the following equation for an element of $Q \in \mathcal{M}_a$, which is often called Pythagorean theorem.

$$D(Q\|P) = D(Q\|\Gamma^{(e)}_{\mathcal{M}_a}[P]) + D(\Gamma^{(e)}_{\mathcal{M}_a}[P]\|P). \tag{4}$$

Given a continuous function Ψ from $\mathcal{M}_a$ to the set of functions on $\mathcal{X}$, we consider the minimization $\min_{P\in\mathcal{M}_a} \mathcal{G}(P)$;

$$\mathcal{G}(P) := \sum_{x\in\mathcal{X}} P(x)\Psi[P](x). \tag{5}$$

This paper aims to find

$$\overline{\mathcal{G}}(a) := \min_{P\in\mathcal{M}_a} \mathcal{G}(P), \quad P_{*,a} := \operatorname*{argmin}_{P\in\mathcal{M}_a} \mathcal{G}(P). \tag{6}$$

For this aim, we propose an iterative algorithm based on a positive real number $\gamma > 0$. Since the above formulation (5) is very general, we can choose the function Ψ dependently on our objective function. That is, different choices of Ψ lead to different objective functions.

For a distribution $Q \in \mathcal{P}(\mathcal{X})$, we define the distribution $\Phi[Q]$ as

$$\Phi[Q](x) := \frac{1}{\kappa[Q]} Q(x) \exp(-\frac{1}{\gamma}\Psi[Q](x)), \tag{7}$$

where $\kappa[Q]$ is the normalization factor $\sum_{x\in\mathcal{X}} Q(x)\exp(-\frac{1}{\gamma}\Psi[Q](x))$. Then, depending on $\gamma > 0$, we propose Algorithm 1. When the calculation of $\Psi[P]$ and the e-projection is feasible, Algorithm 1 is feasible.

Algorithm 1 Minimization of $\mathcal{G}(P)$

As inputs, we prepare the function Ψ, l linearly independent functions $f_1, \ldots, f_l$, constraints $a_1, \ldots, a_k$, a positive number $\gamma > 0$, and the initial value $P^{(1)} \in \mathcal{M}_a$;
repeat
Calculate $P^{(t+1)} := \Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]]$;
until convergence. We denote the convergent by $P^{(\infty)}$. The convergence of this algorithm is guaranteed by Theorem 1.
Output $P^{(\infty)}$.

Indeed, Algorithm 1 is characterized as the iterative minimization of the following two-variable function, i.e., the extended objective function;

$$J_\gamma(P, Q) := \gamma D(P\|Q) + \sum_{x\in\mathcal{X}} P(x)\Psi[Q](x). \tag{8}$$

To see this fact, we define

$$\mathcal{F}_1[P] := \operatorname*{argmin}_{Q\in\mathcal{M}_a} J_\gamma(P, Q), \quad \mathcal{F}_2[Q] := \operatorname*{argmin}_{P\in\mathcal{M}_a} J_\gamma(P, Q). \tag{9}$$

Then, $\mathcal{F}_2[Q]$ is calculated as follows.

Lemma 1 *Under the above definitions, for any positive value $\gamma > 0$, we have $\mathcal{F}_2[Q] = \Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]$, i.e.,*

$$\begin{aligned}
\min_{P\in\mathcal{M}_a} J_\gamma(P, Q) &= J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]], Q) \\
&= \gamma D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|\Phi[Q]) - \gamma \log\kappa[Q], &(10)\\
J_\gamma(P, Q) &= \min_{P'\in\mathcal{M}_a} J_\gamma(P', Q) + \gamma D(P\|\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]) &(11)\\
&= J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]], Q) + \gamma D(P\|\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]). &(12)
\end{aligned}$$

Proof We have the following relations.

$$\begin{aligned}
J_\gamma(P, Q) &= \gamma \sum_{x\in\mathcal{X}} P(x)(\log P(x) - \log Q(x) + \frac{1}{\gamma}\Psi[Q](x)) \\
&= \gamma \sum_{x\in\mathcal{X}} P(x)(\log P(x) - \log \Phi[Q](x) - \log\kappa[Q]) \\
&= \gamma D(P\|\Phi[Q]) - \gamma \log\kappa[Q] \\
&= \gamma D(P\|\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]) + \gamma D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|\Phi[Q]) - \gamma \log\kappa[Q],
\end{aligned} \tag{13}$$

where the final equation follows from (4). Then, the minimum is given as (10), and it is realized with $\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]$.

Applying (10) to the final line of (13), we obtain (11). Since the minimum in (11) is realized when $P' = \Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]$, we obtain (12). □

We calculate $\mathcal{F}_1[Q]$. For this aim, we define

$$D_\Psi(P\|Q) := \sum_{x\in\mathcal{X}} P(x)(\Psi[P](x) - \Psi[Q](x)). \tag{14}$$

Lemma 2 *Assume that two distributions $P, Q \in \mathcal{M}_a$ satisfy the following condition,*

$$D_\Psi(P\|Q) \le \gamma D(P\|Q). \tag{15}$$

Then, we have $\mathcal{F}_1[P] = P$, i.e.,

$$J_\gamma(P, Q) \ge J_\gamma(P, P). \tag{16}$$

Proof Eq. (15) guarantees that

$$\begin{aligned} & J_\gamma(P, Q) - J_\gamma(P, P) \\ &= \gamma D(P\|Q) - \sum_{x\in\mathcal{X}} P(x)(\Psi[P](x) - \Psi[Q](x)) \ge 0. \end{aligned} \tag{17}$$

□

Remark 1 The preceding study [19] discussed the minimization of a function defined over the set of density matrices, i.e., the set of quantum states. When the function is given as a function only of the diagonal part of the density matrix, the function is given as a function of probability distribution composed of the diagonal part. That is, the preceding study [19] covers the case when the function is optimized over a set of probability distributions as a special case. The obtained result of this paper covers the case when the function is optimized over a mixture family. That is, the preceding study [19] does not consider the case with linear constraints. In this sense, the obtained result of this paper generalizes the above special case of the result of [19], and Algorithm 1 is a generalization of the algorithm given in [19].

Lemma 3.2 [19] is composed of several statements. The combination of Lemmas 1 and 2 is a generalization of the above special case of [19, Lemma 3.2]. That is, the classical restriction of [19, Lemma 3.2] is equivalent to the combination of Lemmas 1 and 2 without linear constraints.

Due to Lemmas 1 and 2, when all pairs $(P^{(t+1)}, P^{(t)})$ satisfy (15), the relations

$$\begin{aligned} \mathcal{G}(P^{(t)}) &= J_\gamma(P^{(t)}, P^{(t)}) \ge J_\gamma(P^{(t+1)}, P^{(t)}) \\ &\ge J_\gamma(P^{(t+1)}, P^{(t+1)}) = \mathcal{G}(P^{(t+1)}) \end{aligned} \tag{18}$$

hold under Algorithm 1. In addition, we have the following theorem.

Theorem 1 *When all pairs $(P^{(t+1)}, P^{(t)})$ satisfy (15), i.e., the positive number γ is sufficiently large, Algorithm 1 converges to a local minimum.*

Proof Since $\{\mathcal{G}(P^{(t)})\}$ is monotonically decreasing for t, we have

$$\lim_{n\to\infty} \mathcal{G}(P^{(t)}) - \mathcal{G}(P^{(t+1)}) = 0. \tag{19}$$

Using (12), we have

$$\begin{aligned} \mathcal{G}(P^{(t)}) &= J_\gamma(P^{(t)}, P^{(t)}) \\ &= \gamma D(P^{(t)}\|P^{(t+1)}) + J_\gamma(P^{(t+1)}, P^{(t)}) \\ &\ge \gamma D(P^{(t)}\|P^{(t+1)}) + \mathcal{G}(P^{(t+1)}). \end{aligned} \tag{20}$$

Thus, we have

$$\gamma D(P^{(t)}\|P^{(t+1)}) \le \mathcal{G}(P^{(t)}) - \mathcal{G}(P^{(t+1)}). \tag{21}$$

Since due to (19) and (21), the sequence $\{\mathcal{G}(P^{(t)})\}$ is a Cauchy sequence, it converges. □

To discuss the details of Algorithm 1, we focus on the δ-neighborhood $U(P^0, \delta)$ of P^0 defined as

$$U(P^0, \delta) := \{P \in \mathcal{M}_a | D(P^0 \| P) \leq \delta\}. \tag{22}$$

In particular, we denote $\mathcal{M}_a$ by $U(P^0, \infty)$. Then, we address the following conditions for the δ-neighborhood $U(P^0, \delta)$ of P^0;

(A0) Any distribution $Q \in U(P^0, \delta)$ satisfies the inequality

$$\mathcal{G}(\mathcal{F}_2[Q]) \geq \mathcal{G}(P^0). \tag{23}$$

(A1) Any distribution $Q \in U(P^0, \delta)$ satisfies

$$D_\Psi(\mathcal{F}_2[Q] \| Q) \leq \gamma D(\mathcal{F}_2[Q] \| Q). \tag{24}$$

(A2) Any distribution $Q \in U(P^0, \delta)$ satisfies

$$D_\Psi(P^0 \| Q) \geq 0. \tag{25}$$

(A3) There exists a positive number $\beta > 0$ such that any distribution $Q \in U(P^0, \delta)$ satisfies

$$D_\Psi(P^0 \| Q) = \sum_{x \in \mathcal{X}} P^0(x)(\Psi[P^0](x) - \Psi[Q](x)) \geq \beta D(P^0 \| Q). \tag{26}$$

The condition (A3) is a stronger version of (A2).

However, the convergence to the global minimum is not guaranteed. As a generalization of [19, Theorem 3.3], the following theorem discusses the convergence to the global minimum and the convergence speed.

Theorem 2 *Assume that the δ-neighborhood $U(P^0, \delta)$ of P^0 satisfies the conditions (A1) and (A2) with γ, and $P^{(1)} \in U(P^0, \delta)$. Then, Algorithm 1 with t_0 iterations has one of the following two behaviors.*

(i) There exists an integer $t_1 \leq t_0 + 1$ such that

$$\mathcal{G}(P^{(t_1)}) < \mathcal{G}(P^0). \tag{27}$$

(ii) Algorithm 1 satisfies the conditions $\{P^{(t)}\}_{t=1}^{t_0+1} \subset U(P^0, \delta)$ and

$$\mathcal{G}(P^{(t_0+1)}) - \mathcal{G}(P^0) \leq \frac{\gamma D(P^0 \| P^{(1)})}{t_0}. \tag{28}$$

When the condition (A0) holds additionally, Algorithm 1 with t_0 iterations satisfies (ii).

The above theorem is shown in Appendix 10. Now, we choose an element $P^* \in \mathcal{M}_a$ to satisfy $\mathcal{G}(P^*) = \min_{P \in \mathcal{M}_a} \mathcal{G}(P)$. Then, the condition (A0) holds with $U(P^*, \infty) = \mathcal{M}_a$ and the choice $P^0 = P^*$. When the conditions (A1) and (A2) hold with $U(P^*, \infty) = \mathcal{M}_a$ and the choice $P^0 = P^*$, Theorem 2 guarantees the convergence to the minimizer P^* in Algorithm 1. Although Theorem 2 requires the conditions (A1) and (A2), the condition (A2) is essential due to the following reason. When we choose $\gamma > 0$ to be sufficiently large, the condition (A1) holds with the δ-neighborhood $U(P^*, \delta)$ of P^* because $U(P^*, \delta)$ is a compact set. Hence, it is essential to check the condition (A2) for Theorem 2.

However, as seen in (28), a larger γ makes the convergence speed slower. Therefore, it is important to choose γ to be small under the condition (A1). Practically, it is better to change γ to be smaller when the point $P^{(t)}$ is closer to the minimizer P^*. In fact, as a generalization of [19, Proposition 3.6], we have the following exponential convergence under a stronger condition dependent on γ. In this sense, the parameter is called an acceleration parameter [19, Remark 3.4].

Theorem 3 *Assume that the δ-neighborhood $U(P^0, \delta)$ of P^0 satisfies the conditions (A1) and (A3) with γ, and $P^{(1)} \in U(P^0, \delta)$. Then, Algorithm 1 with t_0 iterations has one of the following two behaviors.*

(i) There exists an integer $t_1 \le t_0 + 1$ such that

$$\mathcal{G}(P^{(t_1)}) < \mathcal{G}(P^0). \tag{29}$$

(ii) Algorithm 1 satisfies the conditions $\{P^{(t)}\}_{t=1}^{t_0+1} \subset U(P^0, \delta)$ and

$$\mathcal{G}(P^{(t_0+1)}) - \mathcal{G}(P^0) \le (1 - \frac{\beta}{\gamma})^{t_0} D(P^0 \| P^{(1)}). \tag{30}$$

When the condition (A0) holds additionally, Algorithm 1 with t_0 iterations satisfies (ii).

The above theorem is shown in Appendix 11. Next, we consider the case when there are several local minimizers $P_1^*, \ldots, P_n^* \in \mathcal{M}_a$ while the true minimizer is P^*. These local minimizers are characterized by the following corollary, which is shown in Appendix 10 as a corollary of Theorem 2.

Corollary 1

$$D_\Psi(P^* \| P_i^*) = \sum_{x \in \mathcal{X}} P^*(x)(\Psi[P^*](x) - \Psi[P_i^*](x)) = \mathcal{G}(P^*) - \mathcal{G}(P_i^*) < 0. \tag{31}$$

Hence, if there exist local minimizers, the condition (A2) does not hold with $U(P^*, \infty) = \mathcal{M}_a$ and the choice $P^0 = P^*$. In this case, when the δ-neighborhood

$U(P_i^*, \delta)$ of P_i^* satisfies the conditions (A0), (A1), and (A2), Algorithm 1 converges to the local minimizer P_i^* with the speed (28) except for the case (i). Since P_i^* is a local minimizer, the δ-neighborhood $U(P_i^*, \delta)$ of P_i^* satisfies the conditions (A0) and (A1) with sufficiently small $\delta > 0$. When the following condition (A4) holds, as shown below, the δ-neighborhood $U(P_i^*, \delta)$ of P_i^* satisfies the condition (A2) with sufficiently small $\delta > 0$. That is, when the initial point belongs to the δ-neighborhood $U(P_i^*, \delta)$, Algorithm 1 converges to P_i^*.

(A4) The function $\eta \mapsto \Psi[P_\eta](x)$ is differentiable, and the relation

$$\sum_{x \in \mathcal{X}} P_\eta(x) \Big(\frac{\partial}{\partial \eta_i} \Psi[P_\eta](x) \Big) = 0 \tag{32}$$

holds for $i = k+1, \ldots, l$ and $P_\eta \in \mathcal{M}_a$.

Lemma 3 *We consider the following two conditions for a convex subset $\mathcal{K} \subset \mathcal{M}_a$.*

(B1) The relation

$$D_\Psi(P \| Q) = \sum_{x \in \mathcal{X}} P(x)(\Psi[P](x) - \Psi[Q](x)) \geq 0 \tag{33}$$

holds for $P, Q \in \mathcal{K}$.

(B2) $\mathcal{G}(P)$ is convex for the mixture parameter in $\mathcal{K}$.

The condition (B1) implies the condition (B2). In addition, when the condition (A4) holds, the condition (B2) implies the condition (B1).

We consider two kinds of mixture parameters. These parametrizations can be converted to each other via affine conversion, which preserves the convexity. Therefore, the condition (B2) does not depend on the choice of mixture parameter.

When the function $\eta \mapsto \Psi[P_\eta](x)$ is twice-differentiable, and the Hessian of $\mathcal{G}(P_\eta)$ is strictly positive semi-definite at a local minimizer P_i^*, this function is convex in the δ-neighborhood $U(P_i^*, \delta)$ of P_i^* with a sufficiently small $\delta > 0$ because the Hessian of $\mathcal{G}(P_\eta)$ is strictly positive semi-definite in the neighborhood due to the continuity.

Then, Lemma 3 guarantees the condition (A2) for the δ-neighborhood $U(P_i^*, \delta)$. Algorithm 1 converges to the local minimizer P_i^* with the speed (28) except for the case (i). The mathematical symbols introduced in Sect. 2.1 are summarized in Table 1.

Proof of Lemma 3 Assume the condition (B1). Then, for $\lambda \in [0, 1]$, we have

$$\begin{aligned}
\varphi(\lambda) &:= \lambda \mathcal{G}(P) + (1-\lambda)\mathcal{G}(Q) - \mathcal{G}(\lambda P + (1-\lambda)Q) \\
&= \lambda \sum_{x \in \mathcal{X}} P(x)(\Psi[P](x) - \Psi[\lambda P + (1-\lambda)Q](x)) \\
&\quad + (1-\lambda) \sum_{x \in \mathcal{X}} Q(x)(\Psi[Q](x) - \Psi[\lambda P + (1-\lambda)Q](x)) \\
&\geq 0,
\end{aligned} \tag{34}$$

Table 1 List of mathematical symbols for Sect. 2.1

Symbol	Description	Eq. number
$\mathcal{P}(\mathcal{X})$	Set of probability distributions over $\mathcal{X}$	
$\mathcal{M}_a$	Mixture family	(1)
$\Gamma^{(e)}_{\mathcal{M}_a}$	e-projection to $\mathcal{M}_a$	(2)
$\mathcal{G}(P)$	Objective function	(5)
$\Psi[P]$	Functional of P used for objective function	(5)
$\overline{\mathcal{G}}(a)$	Minimum value of $\mathcal{G}(P)$	(6)
$P_{*,a}$	Minimizer of $\mathcal{G}(P)$	(6)
$\Phi[Q]$	Functional of Q	(7)
$J_\gamma(P, Q)$	Extended objective function	(8)
$\mathcal{F}_1[P]$	Minimizer of $J_\gamma(P, Q)$ for second argument	(9)
$\mathcal{F}_2[Q]$	Minimizer of $J_\gamma(P, Q)$ for first argument	(9)
$D_\Psi(P\|Q)$	Function of P and Q related to Ψ	(14)
$U(P^0, \delta)$	δ-neighborhood of P^0	(22)

which implies (B2).

Assume the conditions (A4) and (B2). Since $\varphi(\lambda) \ge 0$ for $\lambda \in [0, 1]$, we have

$$\begin{aligned} 0 \le & \frac{d\varphi(\lambda)}{d\lambda}|_{\lambda=0} \\ = & \mathcal{G}(P) - \mathcal{G}(Q) - \sum_{x\in\mathcal{X}} (P(x) - Q(x))\Psi[Q](x) \\ & - \sum_{x\in\mathcal{X}} Q(x) \frac{d\Psi[\lambda P + (1-\lambda)Q](x)}{d\lambda}|_{\lambda=0} \\ \overset{(a)}{=} & \mathcal{G}(P) - \mathcal{G}(Q) - \sum_{x\in\mathcal{X}} (P(x) - Q(x))\Psi[Q](x), \end{aligned} \tag{35}$$

which implies (B1), where (a) follows from the condition (A4). □

Remark 2 The preceding study [19, Theorem 3.3 and Proposition 3.6] consider similar statements as Theorems 2 and 3. As mentioned in Remark 1, the preceding study [19] covers the case when $\mathcal{M}_a$ is given as $\mathcal{P}(\mathcal{X})$, and does not cover the case with a general mixture family $\mathcal{M}_a$. In addition, the preceding study [19, Theorem 3.3 and Proposition 3.6] covers only the case when P^0 and $U(P^0, \delta)$ are P^* and $\mathcal{P}(\mathcal{X})$, respectively. That is, the preceding study does not cover the case with local minimizers. In this sense, Theorems 2 and 3 are more general under the classical setting.

2.2 Algorithm with approximated iteration

In general, it is not so easy to calculate the e-projection $\Gamma^{(e)}_{\mathcal{M}_a}(\Phi[Q])$. We consider the case when it is approximately calculated. There are two methods to calculate

the e-projection. One is the method based on the minimization in the given mixture family, and the other is the method based on the minimization in the exponential family orthogonal to the mixture family. In the first method, the e-projection $\Gamma^{(e)}_{\mathcal{M}_a}(\Phi[Q])$ is the minimizer of the following minimization;

$$\min_{P\in\mathcal{M}_a} D(P\|\Phi[Q]). \tag{36}$$

To describe the second method, using the functions f_j used in (1), we define the exponential family

$$Q_\theta(x) := \Phi[Q](x)e^{\sum_{j=1}^k \theta^j f_j(x)-\phi[Q](\theta)}, \tag{37}$$

where

$$\phi[Q](\theta) := \log \sum_{x\in\mathcal{X}} \Phi[Q](x)e^{\sum_{j=1}^k \theta^j f_j(x)}. \tag{38}$$

The projected element $\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]$ is the unique element of the intersection $\{Q_\theta\}\cap \mathcal{M}_a$. For example, for this fact, see [22, Lemma 3]. Then, the e-projection $\Gamma^{(e)}_{\mathcal{M}_a}(\Phi[Q])$ is given as the solution of the following equation;

$$\frac{\partial\phi[Q]}{\partial\theta^j}(\theta) = \sum_{x\in\mathcal{X}} Q_\theta(x) f_j(x) = a_j \tag{39}$$

for $j = 1,\dots,k$. The solution of (39) is given as the minimizer of the following minimization;

$$\min_{\theta\in\mathbb{R}^k} \phi[Q](\theta) - \sum_{j=1}^k \theta^j a_j. \tag{40}$$

We discuss the precision of our algorithm when each step in the above minimization has a certain error. Allowing certain errors in the first method, we propose Algorithm 2 instead of Algorithm 1.

However, the first method requires the minimization with the same number of parameters as the original minimization $\min_{P\in\mathcal{M}_a}\mathcal{G}(P)$. Hence, it is better to employ the second method. In fact, when $\mathcal{M}_a$ is given as a subset of $\mathcal{P}(\mathcal{X})$ with one linear constraint, the minimization (40) is written as a one-parameter convex minimization. Since any one-parameter convex minimization is performed by the bisection method, which needs $O(-\log\epsilon)$ iterations [35] to achieve a smaller error of the minimum of the objective function than ϵ, the cost of this minimization is much smaller than that of the original minimization $\min_{P\in\mathcal{M}_a}\mathcal{G}(P)$. To consider an algorithm based on the minimization (40), we assume that Ψ is defined in $\mathcal{P}(\mathcal{X})$. In the multi-parameter case, we can use the gradient method and the accelerated proximal gradient method [53–58].

Algorithm 2 Minimization of $\mathcal{G}(P)$ with an error in (36)

As inputs, we prepare the function Ψ, l linearly independent functions $f_1, \ldots, f_l$, constraints $a_1, \ldots, a_k$, positive numbers $\gamma, \epsilon_1, \epsilon_2 > 0$, an integer t_1 (the number of iterations), and the initial value $P^{(1)} \in \mathcal{M}_a$;

repeat

Calculate the pair of $P^{(t+1)} \in \mathcal{M}_a$ and $\bar{P}^{(t+1)} = Q_\theta$ with $Q = P^{(t)}$ in (37) to satisfy

$$\phi[\bar{P}^{(t+1)}](\theta) - \sum_{j=1}^k \theta^j a_j \le \min_{\theta' \in \mathbb{R}^k} \phi[\bar{P}^{(t+1)}](\theta') - \sum_{j=1}^k \theta'^j a_j + \epsilon_1 \tag{41}$$

$$D(\bar{P}^{(t+1)} \| P^{(t+1)}) \le \epsilon_2. \tag{42}$$

until $t = t_1 - 1$.

final step: We output the final estimate $P_f^{(t_1)} := P^{(t_2)} \in \mathcal{M}$ by using $t_2 := \operatorname{argmin}_{t=2,\ldots,t_1} \mathcal{G}(P^{(t)}) - \gamma D(P^{(t)} \| \bar{P}^{(t)})$.

To consider the convergence of Algorithm 2, we extend the conditions (A1) and (A2). For this aim, we focus on the δ-neighborhood $\bar{U}(P^0, \delta)$ of $P^0 \in \mathcal{M}_a$ defined as

$$\bar{U}(P^0, \delta) := \{P \in \mathcal{P}(\mathcal{X}) | D(P^0 \| P) \le \delta\}. \tag{43}$$

Then, we introduce the following conditions for the δ-neighborhood $\bar{U}(P^0, \delta)$ of P^0 as follows.

(A1+) Any distribution $Q \in \bar{U}(P^0, \delta) \cap \mathcal{M}_a = U(P^0, \delta)$ satisfies the following condition with a positive real number $\epsilon_2 > 0$. When a distribution $P \in \mathcal{M}_a$ satisfies $D(P \| \mathcal{F}_2[Q]) \le \epsilon_2$, we have

$$\sum_{x \in \mathcal{X}} P(x)(\Psi[P](x) - \Psi[Q](x)) \le \gamma D(P \| Q). \tag{44}$$

(A2+) Any distribution $Q \in \bar{U}(P^0, \delta)$ satisfies (25).

The convergence of Algorithm 2 is guaranteed in the following theorem.

Theorem 4 *Assume that the δ-neighborhood $\bar{U}(P^0, \delta)$ of P^0 satisfies the conditions (A1+) and (A2+) with two positive real numbers $\gamma > 0$, $\epsilon_2 > 0$, and $P^{(1)} \in U(P^0, \delta)$. Then, for a positive real number $\epsilon_1 > 0$, Algorithm 2 satisfies the conditions*

$$D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t)}]] \| \bar{P}^{(t+1)}) \le \epsilon_1 \tag{45}$$

$$\mathcal{G}(P_f^{(t_1)}) - \mathcal{G}(P^*) \le \frac{\gamma D(P^* \| P^{(1)})}{t_1 - 1} + \epsilon_1 + \gamma \epsilon_2. \tag{46}$$

The above theorem is shown in Appendix 12. We discussed the convergences of Algorithms 1 and 2 under several conditions. When these conditions do not hold,

Table 2 List of mathematical symbols for Sect. 2.2

Symbol	Description	Eq. number
Q_θ	Exponential family	(37)
$\phi[Q](\theta)$	Potential function	(38)

we cannot guarantee its global convergence but, the algorithms achieve a local minimum. Hence, we need to repeat these algorithms by changing the initial value. The mathematical symbols introduced in Sect. 2.2 are summarized in Table 2.

Remark 3 To address the minimization with a cost constraint, the paper [13] added a linear penalty term to the objective function. However, this method does not guarantee that the obtained result satisfies the required cost constraint. Our method can be applied to any mixture family including the distribution family with cost constraint(s). Hence, our method can be applied directly without the above modification while we need to calculate the e-projection. As explained in this subsection, this e-projection can be obtained with the convex minimization whose number of variables is the number of the constraint to define the mixture family. If the number of the constraints is not so large, still the e-projection is feasible.

2.3 Combination of the gradient method and the Algorithm 1

Although we can use the gradient method to calculate (40) for a general mixture family $\mathcal{M}_a$, in order to calculate $\overline{\mathcal{G}}(a) := \min_{P\in\mathcal{M}_a} \mathcal{G}(P)$ with $a \in \mathbb{R}^k$, we propose another algorithm to combine the gradient method and Algorithm 1. This algorithm avoids the calculation of the e-projection $\Gamma^{(e)}_{\mathcal{M}_a}$. For simplicity, we assume that the function $\overline{\mathcal{G}}(a)$ is convex and $\mathcal{M}_a$ is not empty. Then, we replace $f_i(x)$ by $f_i(x) - a_i$, which implies that $\overline{\mathcal{G}}(a)$ is changed to $\overline{\mathcal{G}}(0)$. In the following, we aim to calculate $\overline{\mathcal{G}}(0)$, and we denote the expectation of the function f under the distribution P by $P[f]$.

Then, we consider the following functions by using Legendre transform; For $b = (b^1, \ldots, b^k) \in \mathbb{R}^k$ and $c = (c^1, \ldots, c^{l-k}) \in \mathbb{R}^{l-k}$, we define

$$\mathcal{G}_*(b, c) := \sup_{P\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^{k} b^i P[f_i] + \sum_{j=1}^{l-k} c^i P[f_{k+i}] - \mathcal{G}(P), \tag{47}$$

and

$$\overline{\mathcal{G}}_*(b) := \mathcal{G}_*(b, 0) = \sup_{P\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^{k} b^i P[f_i] - \mathcal{G}(P) = \sup_{a\in\mathbb{R}^k} \sum_{i=1}^{k} b^i a_i - \overline{\mathcal{G}}(a). \tag{48}$$

In the following, we consider the calculation of $\overline{\mathcal{G}}(0)$ by assuming that the function $\eta \mapsto \mathcal{G}(P_\eta)$ is C^2-continuous and convex. Since Legendre transform of $\overline{\mathcal{G}}_*(b)$ is $\overline{\mathcal{G}}(a)$ due to the convexity of $\overline{\mathcal{G}}(a)$, we have $\sup_{b\in\mathbb{R}^k} \sum_{i=1}^k b^i a_i - \overline{\mathcal{G}}_*(b) = \overline{\mathcal{G}}(a)$. As a

special case, we have

$$- \inf_{b \in \mathbb{R}^k} \overline{\mathcal{G}}_*(b) = \overline{\mathcal{G}}(0). \tag{49}$$

That is, when we find the minimizer $b_* := \operatorname*{argmin}_{a \in \mathbb{R}^k} \overline{\mathcal{G}}_*(b)$, we can calculate $\overline{\mathcal{G}}(0)$ as $\overline{\mathcal{G}}(0) = -\sup_{P \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^k b_*^i P[f_i] - \mathcal{G}(P) = \inf_{P \in \mathcal{P}(\mathcal{X})} \mathcal{G}(P) - \sum_{i=1}^k b_*^i P[f_i]$.

To find it, we denote the gradient vector of a function f on $\mathbb{R}^k$ by ∇f. That is, ∇f is the vector $(\frac{\partial}{\partial x^1} f, \ldots, \frac{\partial}{\partial x^k} f)$. Then, we choose a real number L that is larger than the matrix norm of the Hessian of $\overline{\mathcal{G}}_*$, which implies the uniform Lipschitz condition;

$$\|\nabla \overline{\mathcal{G}}_*(b) - \nabla \overline{\mathcal{G}}_*(b')\| \le L \|b - b'\|. \tag{50}$$

Then, we apply the following update rule for the minimization of $\overline{\mathcal{G}}_*(b)$;

$$b_{t+1} := b_t - \frac{1}{L} \nabla \overline{\mathcal{G}}_*(b_t). \tag{51}$$

The following precision is guaranteed [52, Chapter 10] [53, 54];

$$|\overline{\mathcal{G}}_*(b_k) - \overline{\mathcal{G}}_*(b_*)| \le \frac{L}{2k} \|b_* - b_0\|^2. \tag{52}$$

We notice that

$$\nabla \overline{\mathcal{G}}_*(b) = \operatorname*{argmax}_{a \in \mathbb{R}^k} \sum_{i=1}^k b^i a_i - \overline{\mathcal{G}}(a) = (Q_b[f_i])_{i=1}^k, \tag{53}$$

where

$$\begin{aligned} Q_b &:= \operatorname*{argmax}_{P \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^k b^i P[f_i] - \mathcal{G}(P) \\ &= \operatorname*{argmin}_{P \in \mathcal{P}(\mathcal{X})} \sum_{x \in \mathcal{X}} P(x) \Big(\Psi[P](x) - \sum_{i=1}^k b^i f_i(x) \Big). \end{aligned} \tag{54}$$

However, the calculation of (54) requires a large calculation amount. Hence, replacing the update rule (51) by a one-step iteration in Algorithm 1, we propose another algorithm.

Using $\Phi^b[Q](x) := \frac{1}{\kappa} Q(x) \exp(-\frac{1}{\gamma} \Big(\Psi[P](x) - \sum_{i=1}^k b^i f_i(x) \Big)$ with the normalizing constant κ, we propose Algorithm 3.

It is not so easy to evaluate the convergence speed of Algorithm 3. But, when it converges, the convergent point is the true minimizer.

Theorem 5 *When the pair (b, P) is a convergence point, we have $b = b_*$ and $P = P_*$.*

Algorithm 3 Minimization of $\mathcal{G}(P)$

As inputs, we prepare the function Ψ, l linearly independent functions $f_1, \ldots, f_l$, a positive number $\gamma > 0$, the maximum iteration number t_1, and the initial values $P^{(1)} \in \mathcal{M}_0$, $b_1 \in \mathbb{R}^k$;
repeat
Calculate $P^{(t+1)} := \Phi^{b_t}[P^{(t)}]$ and $b_{t+1} := b_t - \frac{1}{L}(P^{(t+1)}[f_i])_{i=1}^k$;
until convergence if it converges. If it does not converge, we stop the algorithm at $t = t_1$. We denote the convergent by $(P^{(\infty)}, b_\infty)$.
Output $P^{(\infty)}$ and $\mathcal{G}(P^{(\infty)})$.

Proof Since the pair (b, P) is a convergence point, we have $P = \Phi^b[P]$, which implies

$$\sum_{i=1}^k b^i P[f_i] - \mathcal{G}(P) = \sup_{P' \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^k b^i P'[f_i] - \mathcal{G}(P') = \overline{\mathcal{G}}_*(b). \tag{55}$$

Since the pair (b, P) is a convergence point, we have $P[f_i] = 0$ for $i = 1, \ldots, k$, i.e., the distribution P satisfies the required condition in (1). The relation (53) implies $\nabla \overline{\mathcal{G}}_*(b) = 0$. Hence, (49) yields $\overline{\mathcal{G}}_*(b) = \overline{\mathcal{G}}(0)$, which implies $b = b_*$. Therefore, the relation (55) is rewritten as $\mathcal{G}(P) = \overline{\mathcal{G}}(0)$, which implies $P = P_*$. □

Remark 4 We compare our algorithm with a general algorithm proposed in [13]. The input of the objective function in [13] forms a mixture family. The function f given in [13, (6)] satisfies the condition of $\mathcal{G}$ by considering the second line of [13, (6)] as Ψ. Their algorithm is the same as Algorithm 1 with $\gamma = 1$ when there is no constraint because their extended objective function g defined in [13, (16)] can be considered as $D(P\|Q) + \sum_{x \in \mathcal{X}} P(x)\Psi[Q](x)$, where the choice of q in [13] corresponds to the choice of P and the choice of $Q_1, \ldots, Q_K$ in [13] does to the choice of Q.

Also, we can show that the function f given in [13, (6)] satisfies the condition (A4). Since the condition (A4) holds, the convexity of f is equivalent to the condition (B1). This equivalence, in this case, was shown as [13, Proposition 4.1]. They showed the convergence of their algorithm as [13, Theorem 4.1], which can be considered as a special case of our Theorem 2.

However, our treatment for the constraint is different from theirs. They consider the minimization $\min_{P \in \mathcal{P}(\mathcal{X})} \mathcal{G}(P) - \sum_{i=1}^k b^i P[f_i]$ without updating the parameter b. Hence, their algorithm cannot achieve the minimum with the desired constraint while Algorithms 1, 2, and 3 achieve the minimum with the desired constraint. Although their algorithm is similar to Algorithm 3, Algorithm 3 updates the parameter b to achieve the minimum with the desired constraint.

3 Application to information theoretical problems

3.1 Channel capacity

In the same way as the reference [19], we apply our problem setting to the channel coding. A channel is given as a conditional distribution $W_{Y|X}$ on the sample space

$\mathcal{Y}$ with conditions on the sample space $\mathcal{X}$, where $\mathcal{Y}$ is a general sample space with a measure μ and $\mathcal{X}$ is a finite sample space. For two absolutely continuous distributions P_Y and Q_Y with respect to μ on $\mathcal{Y}$, the Kullback–Leibler divergence $D(P_Y\|Q_Y)$ is given as

$$D(P_Y\|Q_Y) := \int_{\mathcal{Y}} p_Y(y)(\log p_Y(y) - \log q_Y(y))\mu(dy), \tag{56}$$

where p_Y and q_Y are the probability density functions of P_Y and Q_Y with respect to μ. This quantity is generalized to the Renyi divergence with order $\alpha > 0$ as

$$D_\alpha(P_Y\|Q_Y) := \frac{1}{\alpha-1}\log \int_{\mathcal{Y}} \Big(\frac{p_Y(y)}{q_Y(y)}\Big)^{\alpha-1} p_Y(y)\mu(dy). \tag{57}$$

The channel capacity $C(W_{Y|X})$ is given as the maximization of the mutual information $I(P_X, W_{Y|X})$ as [10]

$$C(W_{Y|X}) := \max_{P_X} I(P_X, W_{Y|X}) \tag{58}$$

$$\begin{aligned} I(P_X, W_{Y|X}) &:= \sum_{x\in\mathcal{X}} P_X(x) D(W_{Y|X=x}\|W_{Y|X}\cdot P_X) \\ &= D(W_{Y|X}\times P_X\|(W_{Y|X}\cdot P_X)\times P_X), \end{aligned} \tag{59}$$

where $W_{Y|X}\cdot P_X$ and $W_{Y|X}\times P_X$ are defined as the following probability density functions $w_{Y|X}\cdot P_X$ and $w_{Y|X}\times P_X$;

$$(w_{Y|X}\cdot P_X)(y) := \sum_{x\in\mathcal{X}} P_X(x) w_{Y|X=x}(y) \tag{60}$$

$$(w_{Y|X}\times P_X)(x,y) := P_X(x) w_{Y|X=x}(y). \tag{61}$$

However, the mutual information $I(P_X, W_{Y|X})$ has another form as

$$I(P_X, W_{Y|X}) = \min_{Q_Y} \sum_{x\in\mathcal{X}} P_X(x) D(W_{Y|X=x}\|Q_Y). \tag{62}$$

When we choose $\mathcal{M}_a$ and Ψ as $\mathcal{P}(\mathcal{X})$ and

$$\Psi_{W_{Y|X}}[P_X](x) := -D(W_{Y|X=x}\|W_{Y|X}\cdot P_X), \tag{63}$$

$-I(P_X, W_{Y|X})$ coincides with $\mathcal{G}(P_X)$ [19]. Since

$$D_\Psi(P_X\|Q_X) = D(W_{Y|X}\cdot P_X\|W_{Y|X}\cdot Q_X) \ge 0, \tag{64}$$

the condition (A2) holds with $\mathcal{P}(\mathcal{X})$. In addition, since the information processing inequality guarantees that

$$D(W_{Y|X} \cdot P_X \| W_{Y|X} \cdot Q_X) \le D(P_X \| Q_X), \tag{65}$$

the condition (A1) holds with $\gamma = 1$ and $\mathcal{P}(\mathcal{X})$. In this case, Φ is given as

$$\Phi[Q_X](x) = \frac{1}{\kappa_{W_{Y|X}}[Q_X]} Q_X(x) \exp(\frac{1}{\gamma} D(W_{Y|X=x} \| W_{Y|X} \cdot Q_X)), \tag{66}$$

where the normalizing constant $\kappa_{W_{Y|X}}[Q_X]$ is given as

$$\kappa_{W_{Y|X}}[Q_X] = \sum_{x \in \mathcal{X}} Q_X(x) \exp(\frac{1}{\gamma} D(W_{Y|X=x} \| W_{Y|X} \cdot Q_X)). \tag{67}$$

When $\gamma = 1$, it coincides with the Arimoto–Blahut algorithm [8, 9]. Since $\Phi[Q_X] \in \mathcal{P}(\mathcal{X})$, $P_X^{(t+1)}$ is given as $\Phi[P_X^{(t)}]$.

Remark 5 The reference [19] covers the case when $\mathcal{M}_a$ is given as $\mathcal{P}(\mathcal{X})$, and the reference [19] presented the algorithms presented in this subsection in a more general form. Also, they proposed an adaptive choice of γ in this case [19, (22)]. In addition, they numerically compared their adaptive choice with the case of $\gamma = 1$ [19, Figs. 1,..., 6]. These comparisons show a significant improvement by their adaptive choice.

3.2 Reliability function in channel coding

In channel coding, we consider the reliability function, which was originally introduced by Gallager [25] and expresses the exponential decreasing rate of an upper bound of the decoding block error probability under the random coding. To achieve this aim, for $\alpha > 0$, we define

$$I_\alpha(P_X, W_{Y|X}) := \frac{\alpha}{\alpha - 1} \log \Big(\int_{\mathcal{Y}} \Big(\sum_{x \in \mathcal{X}} P_X(x) w_{Y|X=x}(y)^\alpha \Big)^{\frac{1}{\alpha}} \mu(dy) \Big). \tag{68}$$

Then, when the code is generated with the random coding based on the distribution P_X, the decoding block error probability with coding rate R is upper bounded by the following quantity;

$$e^{n \min_{\rho \in [0,1]} \left(\rho R - \rho I_{\frac{1}{1+\rho}}(P_X, W_{Y|X}) \right)} \tag{69}$$

when we use the channel $W_{Y|X}$ with n times. Notice that $e^{-\rho I_{\frac{1}{1+\rho}}(P_X, W_{Y|X})} = \int_{\mathcal{Y}} \Big(\sum_{x \in \mathcal{X}} P_X(x) w_{Y|X=x}(y)^{\frac{1}{1+\rho}} \Big)^{1+\rho} \mu(dy)$. That is, the Gallager function [25] is

given as $\rho I_{\frac{1}{1+\rho}}(P_X, W_{Y|X})$, i.e., the parameter α is different from the parameter ρ in the Gallager function. Taking the minimum for the choice of P_X, we have

$$\min_{P_X} e^{\min_{\rho\in[0,1]}\left(\rho R-\rho I_{\frac{1}{1+\rho}}(P_X,W_{Y|X})\right)} = \min_{\alpha\in[1/2,1]}\left(e^{-\frac{\alpha-1}{\alpha}R}\min_{P_X} e^{\frac{\alpha-1}{\alpha}I_\alpha(P_X,W_{Y|X})}\right) \quad (70)$$

with $\alpha = \frac{1}{1-\rho} \in [1/2, 1]$. Therefore, we consider the following minimization;

$$\min_{P_X} e^{\frac{\alpha-1}{\alpha}I_\alpha(P_X,W_{Y|X})} = \min_{P_X}\int_{\mathcal{Y}}\Big(\sum_{x\in\mathcal{X}} P_X(x)w_{Y|X=x}(y)^\alpha\Big)^{\frac{1}{\alpha}}\mu(dy). \quad (71)$$

In the following, we discuss the RHS of (71) with $\alpha \in [1/2, 1]$.

To apply our method, as a generalization of (62), we consider another expression of $I_\alpha(P_X, W_{Y|X})$;

$$I_\alpha(P_X, W_{Y|X}) = \min_{Q_Y} D_\alpha(W_{Y|X}\times P_X\|Q_Y\times P_X), \quad (72)$$

which was shown in [27, Lemma 2]. Using

$$\begin{aligned} Q_{Y|\alpha,P_X} &:= \operatorname*{argmin}_{Q_Y} D_\alpha(W_{Y|X}\times P_X\|Q_Y\times P_X) \\ &= \operatorname*{argmax}_{Q_Y}\sum_{x\in\mathcal{X}} P_X(x)e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_Y)}, \end{aligned} \quad (73)$$

we have

$$\Big(\min_{P_X} e^{\frac{\alpha-1}{\alpha}I_\alpha(P_X,W_{Y|X})}\Big)^\alpha = \min_{P_X}\sum_{x\in\mathcal{X}} P_X(x)e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})}. \quad (74)$$

The probability density function $q_{Y|\alpha,P_X}$ of the minimizer $Q_{Y|\alpha,P_X}$ is calculated as

$$q_{Y|\alpha,P_X}(y) = C\Big(\sum_{x\in\mathcal{X}} P_X(x)w_{Y|X=x}(y)^\alpha\Big)^{\frac{1}{\alpha}}, \quad (75)$$

where C is the normalizing constant [27, Lemma 2].

To solve the minimization (74), we apply our method to the case when we choose $\mathcal{M}_a$ and Ψ as $\mathcal{P}(\mathcal{X})$ and

$$\Psi_{\alpha,W_{Y|X}}[P_X](x) := e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})}. \quad (76)$$

Since (73) guarantees that

$$\sum_{x\in\mathcal{X}} P_X(x)(\Psi_{\alpha,W_{Y|X}}[P_X](x) - \Psi_{\alpha,W_{Y|X}}[Q_X](x)) = \sum_{x\in\mathcal{X}} P_X(x)\Big(e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})} - e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,Q_X})}\Big) \ge 0, \quad (77)$$

the condition (A2) holds with $\mathcal{M}_a = \mathcal{P}(\mathcal{X})$. The condition (A1) can be satisfied with sufficiently large γ. In this case, Φ is given as

$$\Phi_\alpha[Q_X](x) = \frac{1}{\kappa_{\alpha,W_{Y|X}}[Q_X]} Q_X(x) \exp(-\frac{1}{\gamma} e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})}), \quad (78)$$

where the normalizing constant $\kappa_{\alpha,W_{Y|X}}[Q_X]$ is given as $\kappa_{\alpha,W_{Y|X}}[Q_X] = \sum_{x\in\mathcal{X}} Q_X(x)$ $\exp(-\frac{1}{\gamma} e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})})$. Since $\Phi_\alpha[Q_X] \in \mathcal{M}_a$, $P_X^{(t+1)}$ is given as $\Phi_\alpha[P_X^{(t)}]$.

3.3 Strong converse exponent in channel coding

In channel coding, we discuss an upper bound of the probability of correct decoding. This probability is upper bounded by the following quantity;

$$\max_{P_X} e^{n \min_{\rho\in[0,1]} \left(-\rho R + \rho I_{\frac{1}{1-\rho}}(P_X, W_{Y|X})\right)} \quad (79)$$

when we use the channel $P_{Y|X}$ with n times and the coding rate is R [26]. Therefore, we consider the following maximization;

$$\max_{P_X} e^{\rho I_{\frac{1}{1-\rho}}(P_X, W_{Y|X})} = \max_{P_X} e^{\frac{\alpha-1}{\alpha} I_\alpha(P_X, W_{Y|X})} \quad (80)$$

with $\alpha = \frac{1}{1-\rho} > 1$. In the following, we discuss the RHS of (80) with $\alpha > 1$.

To apply our method, we consider another expression (72) of $I_\alpha(P_X, W_{Y|X})$. Using (73), we have

$$\Big(\max_{P_X} e^{\frac{\alpha-1}{\alpha} I_\alpha(P_X, W_{Y|X})}\Big)^\alpha = \max_{P_X} \sum_{x\in\mathcal{X}} P_X(x) e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})}. \quad (81)$$

The maximization (81) can be solved by choosing $\mathcal{M}_a$ and Ψ as $\mathcal{P}(\mathcal{X})$ and

$$\Psi_{\alpha,W_{Y|X}}[P_X](x) := -e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})}. \quad (82)$$

Since (73) guarantees that

$$\sum_{x\in\mathcal{X}} P_X(x)(\Psi_{\alpha,W_{Y|X}}[P_X](x) - \Psi_{\alpha,W_{Y|X}}[Q_X](x))$$

$$= \sum_{x\in\mathcal{X}} P_X(x)\Big(- e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})} + e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,Q_X})}\Big) \ge 0, \tag{83}$$

the condition (A2) holds with $\mathcal{M}_a = \mathcal{P}(\mathcal{X})$. Similarly, the condition (A1) can be satisfied with sufficiently large γ. In this case, Φ is given as

$$\Phi_\alpha[Q_X](x) = \frac{1}{\kappa_{\alpha,W_{Y|X}}[Q_X]} Q_X(x) \exp(\frac{1}{\gamma} e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})}), \tag{84}$$

where the normalizing constant $\kappa_{\alpha,W_{Y|X}}[Q_X]$ is given as $\kappa_{\alpha,W_{Y|X}}[Q_X] = \sum_{x\in\mathcal{X}} Q_X(x) \exp(\frac{1}{\gamma} e^{(\alpha-1)D_\alpha(W_{Y|X=x}\|Q_{Y|\alpha,P_X})})$. Since $\Phi_\alpha[Q_X] \in \mathcal{M}_a$, $P_X^{(t+1)}$ is given as $\Phi_\alpha[P_X^{(t)}]$.

3.4 Wiretap channel capacity

3.4.1 General case

Given a pair of a channel $W_{Y|X}$ from $\mathcal{X}$ to a legitimate user $\mathcal{Y}$ and a channel $W_{Z|X}$ from $\mathcal{X}$ to a malicious user $\mathcal{Z}$, the wiretap channel capacity is given as [28, 29]

$$C(W_{Y|X}, W_{Z|X}) := \max_{P_{VX}} I(P_V, W_{Y|X}\cdot P_{X|V}) - I(P_V, W_{Z|X}\cdot P_{X|V}) \tag{85}$$

with a sufficiently large discrete set $\mathcal{V}$. The recent papers showed that the above rate can be achieved even with the strong security [30–32] and the semantic security [33, 34].[1] Furthermore, the paper [34, Appendix D] showed the above even when the output systems are general continuous systems including Gaussian channels. The wiretap capacity (85) can be calculated via the minimization;

$$\min_{P_{VX}} -I(P_V, W_{Y|X}\cdot P_{X|V}) + I(P_V, W_{Z|X}\cdot P_{X|V}). \tag{86}$$

Here, $\mathcal{V}$ is an additional discrete sample space. When we choose $\mathcal{M}_a$ and Ψ as $\mathcal{P}(\mathcal{X}\times\mathcal{V})$ and

$$\begin{aligned}&\Psi_{W_{Y|X},W_{Z|X}}[P_{VX}](v,x)\\ &:= D(W_{Z|X}\cdot P_{X|V=v}\|W_{Z|X}\cdot P_X) - D(W_{Y|X}\cdot P_{X|V=v}\|W_{Y|X}\cdot P_X),\end{aligned} \tag{87}$$

$-I(P_V, W_{Y|X}\cdot P_{X|V}) + I(P_V, W_{Z|X}\cdot P_{X|V})$ coincides with $\mathcal{G}(P_{VX})$. Here, although $\Psi_{W_{Y|X},W_{Z|X}}[P_{VX}]$ is a function of (v,x), the function value depends only on v. Hence, the general theory in Sect. 2 can be used for the minimization of (86). In this case, it

[1] The strong security means that the mutual information between the message and the eavesdropper's information goes to zero as the number of uses of the channel goes to zero when the message is subject to the uniform distribution. The semantic security means that the maximum of the mutual information by changing the distribution of the message goes to zero as the number of uses of the channel goes to zero.

is difficult to clarify whether the conditions (A1) and (A2) hold in general. Φ is given as

$$\begin{aligned}&\Phi[Q_{VX}](v,x)\\&= \frac{1}{\kappa_{W_{Y|X},W_{Z|X}}[Q_{VX}]} Q_{VX}(v,x) \exp\Big(\frac{1}{\gamma}\Big(D(W_{Y|X}\cdot P_{X|V=v}\|W_{Y|X}\cdot P_X)\\&\quad - D(W_{Z|X}\cdot P_{X|V=v}\|W_{Z|X}\cdot P_X)\Big)\Big),\end{aligned} \tag{88}$$

where $\kappa_{W_{Y|X},W_{Z|X}}[Q_{VX}]$ is the normalizing constant. Since $\Phi[Q_X] \in \mathcal{M}_a$, $P_X^{(t+1)}$ is given as $\Phi[P_X^{(t)}]$.

3.4.2 Degraded case

However, when there exists a channel $W_{Z|Y}$ from $\mathcal{Y}$ to $\mathcal{Z}$ such that $W_{Z|Y}\cdot W_{Y|X} = W_{Z|X}$, i.e., the channel $W_{Z|X}$ is a degraded channel of $W_{Y|X}$, we can define the joint channel $W_{YZ|X}$ with the following conditional probability density function

$$w_{YZ|X}(yz|x) := w_{Z|Y}(z|y) w_{Y|X}(y|x). \tag{89}$$

Then, the maximization (85) is simplified as

$$C(W_{YZ|X}) := \max_{P_X} I(X;Y|Z)[P_X, W_{YZ|X}] \tag{90}$$

where the conditional mutual information is given as

$$I(X;Y|Z)[P_X, W_{YZ|X}] := \sum_{x,z} P_{XZ}(x,z) D(P_{Y|X=x,Z=z}\|P_{Y|Z=z}), \tag{91}$$

where the conditional distributions $P_{Y|XZ}$ and $P_{Y|Z}$ are defined from the joint distribution $W_{YZ|X}\times P_X$. To consider (90), we consider the following minimization with a general two-output channel $W_{YZ|X}$;

$$\min_{P_X} -I(X;Y|Z)[P_X, W_{YZ|X}]. \tag{92}$$

When we choose $\mathcal{M}_a$ and Ψ as $\mathcal{P}(\mathcal{X})$ and

$$\Psi_{W_{YZ|X}}[P_X](x) := -\sum_z P_{Z|X=x}(z) D(P_{Y|X=x,Z=z}\|P_{Y|Z=z}). \tag{93}$$

$-I(X;Y|Z)[P_X, W_{YZ|X}]$ coincides with $\mathcal{G}(P_X)$. Hence, the general theory in Sect. 2 can be used for the minimization of (92). In this case, as shown in Sect. 6.2, the conditions (A1) with $\gamma = 1$ and (A2) hold. Φ is given as

$$\Phi[Q_X](x) = \frac{1}{\kappa_{W_{YZ|X}}[Q_X]} Q_X(x) \exp\Big(\frac{1}{\gamma}\Big(\sum_z P_{Z|X=x}(z) D(P_{Y|X=x,Z=z} \| P_{Y|Z=z})\Big)\Big), \tag{94}$$

where $\kappa_{W_{YZ|X}}[Q_X]$ is the normalizing constant. Since $\Phi[Q_X] \in \mathcal{M}_a$, $P_X^{(t+1)}$ is given as $\Phi[P_X^{(t)}]$. The above algorithm with $\gamma = 1$ coincides with the algorithm proposed by [14].

3.5 Capacities with cost constraint

Next, we consider the case when a cost constraint is imposed. Consider a function f on $\mathcal{X}$ and the following constraint for a distribution $P_X \in \mathcal{X}$;

$$P_X[f] = a. \tag{95}$$

We define $\mathcal{M}_a$ by imposing the condition (95) as a special case of (1). The capacity of the channel $W_{Y|X}$ under the cost constraint is given as $\max_{P_X \in \mathcal{M}_a} I(P_X, W_{Y|X})$. That is, we need to solve the minimization $\min_{P_X \in \mathcal{M}_a} -I(P_X, W_{Y|X})$. In this case, the $t+1$-th distribution $P^{(t+1)}$ is given as $\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P_X^{(t)}]]$. Since $\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P_X^{(t)}]]$ cannot be calculated analytically, we can use Algorithm 2 instead of Algorithm 1. Since conditions (A1) with $\gamma = 1$ and (A2) hold, Theorem 4 guarantees the global convergence to the minimum in Algorithm 2.

We can consider the cost constraint (95) for the problems (74) and (81). In these cases, we have a similar modification by considering $\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P_X^{(t)}]]$.

4 em problem

We apply our algorithm to the problem setting with the em algorithm [2–4], which is a generalization of Boltzmann machines [5]. The em algorithm is implemented by iterative applications of the projection to an exponential family (the m-projection) and the projection to a mixture family (the e-projection). Hence, this algorithm is called the em algorithm. On the other hand, the EM algorithm is implemented by iterative applications of expectation and maximization. Their relation is summarized as follows. In particular, the expectation in the EM algorithm, which is often called E-step, corresponds to the e-projection to a mixture family, which is often called e-step in the em algorithm. Also, the maximization in the EM algorithm, which is often called M-step, corresponds to the m-projection to an exponential family, which is often called m-step in the em algorithm. In this reason, they are essentially the same [2].

For this aim, we consider a pair of an exponential family $\mathcal{E}$ and a mixture family $\mathcal{M}_a$ on $\mathcal{X}$. We denote the m-projection to $\mathcal{E}$ of P by $\Gamma^{(m)}_{\mathcal{E}}[P]$, which is defined as [1, 2]

$$\Gamma_{\mathcal{E}}^{(m)}[P] := \operatorname*{argmin}_{Q\in\mathcal{E}} D(P\|Q). \tag{96}$$

We consider the following minimization;

$$\begin{aligned}\min_{P\in\mathcal{M}_a}\min_{Q\in\mathcal{E}} D(P\|Q) &= \min_{P\in\mathcal{M}_a} D(P\|\Gamma_{\mathcal{E}}^{(m)}[P])\\ &= \min_{P\in\mathcal{M}_a}\sum_{x\in\mathcal{X}} P(x)(\log P(x) - \log\Gamma_{\mathcal{E}}^{(m)}[P](x)).\end{aligned} \tag{97}$$

We choose the function Ψ as

$$\Psi_{\mathrm{em}}[P](x) := (\log P(x) - \log\Gamma_{\mathcal{E}}^{(m)}[P](x)), \tag{98}$$

and apply the discussion in Sect. 2. Then, we have

$$\begin{aligned}&\sum_{x\in\mathcal{X}} P^0(x)(\Psi_{\mathrm{em}}[P^0](x) - \Psi_{\mathrm{em}}[Q](x))\\ &= \sum_{x\in\mathcal{X}} P^0(x)\Big((\log P^0(x) - \log\Gamma_{\mathcal{E}}^{(m)}[P^0](x)) - (\log Q(x) - \log\Gamma_{\mathcal{E}}^{(m)}[Q](x))\Big)\\ &= \sum_{x\in\mathcal{X}} P^0(x)(\log P^0(x) - \log Q(x))\\ &\quad + \sum_{x\in\mathcal{X}} P^0(x)\Big(\log\Gamma_{\mathcal{E}}^{(m)}[Q](x)) - \log\Gamma_{\mathcal{E}}^{(m)}[P^0](x))\Big)\\ &= D(P^0\|Q) + D(P^0\|\Gamma_{\mathcal{E}}^{(m)}[P^0]) - D(P^0\|\Gamma_{\mathcal{E}}^{(m)}[Q])\\ &= D(P^0\|Q) - D(\Gamma_{\mathcal{E}}^{(m)}[P^0]\|\Gamma_{\mathcal{E}}^{(m)}[Q]),\end{aligned} \tag{99}$$

where the final equation follows from (4). The condition (A1) holds with $U(P^0,\infty) = \mathcal{M}_a$ and $\gamma = 1$. There is a possibility that the condition (A1) holds with a smaller γ. Therefore, with $\gamma = 1$, Theorem 1 guarantees that Algorithm 1 converges to a local minimum. In addition, when the relation

$$D(P^0\|Q) \geq D(\Gamma_{\mathcal{E}}^{(m)}[P^0]\|\Gamma_{\mathcal{E}}^{(m)}[Q]) \tag{100}$$

holds for $Q \in U(P^0,\delta)$, the condition (A2) holds with $U(P^0,\delta)$. That is, if the condition (100) holds, Algorithm 1 has the global convergence to the minimizer. The condition (100) is a condition similar to the condition given in [22].

In this case, Φ is given as

$$\begin{aligned}\Phi[Q](x) &= \frac{1}{\kappa_{\rm em}[Q]} Q(x) \exp\Big(-\frac{1}{\gamma}(\log Q(x) - \log \Gamma_{\mathcal{E}}^{(m)}[Q](x))\Big) \\ &= \frac{1}{\kappa_{\rm em}[Q]} Q(x)^{\frac{\gamma-1}{\gamma}} \Gamma_{\mathcal{E}}^{(m)}[Q](x)^{\frac{1}{\gamma}}, \end{aligned} \tag{101}$$

where the normalizing constant $\kappa_{\rm em}[Q_X]$ is given as $\kappa_{\rm em}[Q_X] = \sum_{x\in\mathcal{X}} Q(x)^{\frac{\gamma-1}{\gamma}} \Gamma_{\mathcal{E}}^{(m)}[Q](x)^{\frac{1}{\gamma}}$. Since $\Phi[Q] \in \mathcal{M}_a$, $P^{(t+1)}$ is given as $\Gamma_{\mathcal{M}_a}^{(e)}[\Phi[P^{(t)}]]$. When $\gamma = 1$, it coincides with the conventional em-algorithm [2–4] because $\Phi[P^{(t)}] = \Gamma_{\mathcal{E}}^{(m)}[P^{(t)}]$. The above analysis suggests the choice of γ as a smaller value than 1. That is, there is a possibility that a smaller γ improves the conventional em-algorithm. In addition, we may use Algorithm 2 instead of Algorithm 1 when the calculation of e-projection is difficult.

Lemma 4 *When $\Psi_{\rm em}$ is given as (98), the condition (A4) holds.*

Therefore, by combining Lemmas 3 and 4, the assumption of Theorem 2 holds in the δ neighborhood of a local minimizer with sufficiently small $\delta > 0$. That is, the convergence speed can be evaluated by Theorem 2.

Proof Pythagorean theorem guarantees

$$\begin{aligned}&\sum_{x\in\mathcal{X}} P(x)\Big(\log \Gamma_{\mathcal{E}}^{(m)}[Q](x) - \log \Gamma_{\mathcal{E}}^{(m)}[P](x)\Big) \\ &\quad = D(P\|\Gamma_{\mathcal{E}}^{(m)}[P]) - D(P\|\Gamma_{\mathcal{E}}^{(m)}[Q]) = D(\Gamma_{\mathcal{E}}^{(m)}[P]\|\Gamma_{\mathcal{E}}^{(m)}[Q]). \end{aligned} \tag{102}$$

We make the parameterization $P_\eta \in \mathcal{M}_a$ with mixture parameter η. We denote $\eta(h,i) := (\eta(0)_1, \ldots, \eta(0)_{i-1}, \eta(0)_i + h, \eta(0)_{i+1}, \ldots, \eta(0)_k)$.

$$\begin{aligned}&\sum_{x\in\mathcal{X}} P_{\eta(0)}(x)\Big(\frac{\partial}{\partial \eta_i}\Psi[P_\eta](x)|_{\eta=\eta(0)}\Big) \\ &= \sum_{x\in\mathcal{X}} P_{\eta(0)}(x)\Big(\lim_{h\to 0} \frac{\Psi[P_{\eta(h,i)}](x) - \Psi[P_{\eta(0)}](x)}{h}\Big) \\ &= \sum_{x\in\mathcal{X}} P_{\eta(0)}(x)\Big(\lim_{h\to 0} \frac{\log P_{\eta(h,i)}(x) - \log P_{\eta(0)}(x)}{h} \\ &\quad - \lim_{h\to 0} \frac{\log \Gamma_{\mathcal{E}}^{(m)}[P_{\eta(h,i)}](x) - \log \Gamma_{\mathcal{E}}^{(m)}[P_{\eta(0)}](x)}{h}\Big) \\ &\overset{(a)}{=} \sum_{x\in\mathcal{X}} P_{\eta(0)}(x)\Big(\lim_{h\to 0} \frac{\log P_{\eta(h,i)}(x) - \log P_{\eta(0)}(x)}{h}\Big) \\ &\quad - \sum_{x\in\mathcal{X}} \Gamma_{\mathcal{E}}^{(m)}[P_{\eta(0)}](x)\Big(\lim_{h\to 0} \frac{\log \Gamma_{\mathcal{E}}^{(m)}[P_{\eta(h,i)}](x) - \log \Gamma_{\mathcal{E}}^{(m)}[P_{\eta(0)}](x)}{h}\Big) \end{aligned}$$

$$= \sum_{x\in\mathcal{X}} \frac{\partial}{\partial \eta_i} P_\eta(x)|_{\eta=\eta(0)} - \sum_{x\in\mathcal{X}} \frac{\partial}{\partial \eta_i} \Gamma_{\mathcal{E}}^{(m)}[P_\eta](x)|_{\eta=\eta(0)} = 0, \tag{103}$$

which implies the condition (A4). Here, (a) follows from (102). □

5 Commitment capacity

Using the same notations as Sect. 3, we address the bit commitment via a noisy channel $W_{Y|X}$. When we can guarantee the communication channel is given by $W_{Y|X}$, bit commitment is possible. In this topic, we are interested in the secure transmission rate in the sense of bit commitment per a single use of the noisy channel $W_{Y|X}$. The maximum value of this rate is called the commitment capacity. Given a distribution P_X, the Shannon entropy is given as

$$H(X)_{P_X} := -\sum_{x\in\mathcal{X}} P_X(x) \log P_X(x). \tag{104}$$

Given a joint distribution P_{XY}, the conditional entropy is defined as

$$H(X|Y)_{P_{XY}} := \int_{\mathcal{Y}} H(X)_{W_{X|Y=y}} p_Y(y)\mu(dy). \tag{105}$$

The commitment capacity is given as

$$C_c(W_{Y|X}) := \max_{P_X} H(X|Y)_{W_{Y|X}\times P_X}. \tag{106}$$

This problem setting has several versions. To achieve the bit commitment, the papers [36–38] considered interactive protocols with multiple rounds, where each round has one use of the given noisy channel $W_{Y|X}$ and free noiseless communications in both directions. Then, it derived the commitment capacity (106). Basically, the proof is composed of two parts. One is the achievability part, which is often called the direct part and shows the existence of the code to achieve the capacity. The other is the impossibility part, which is often called the converse part and shows the non-existence of the code to exceed the capacity. As the achievability part, they showed that the commitment capacity can be achieved with non-interactive protocol, which has no free noiseless communication during multiple uses of the given noisy channel $W_{Y|X}$. However, as explained in [39], their proof of the impossibility part skips so many steps that it cannot be followed. Later, the paper [40] showed the impossibility part only for non-interactive protocols by applying the wiretap channel. Recently, the paper [41] constructed a code to achieve the commitment capacity by using a specific type of list decoding. Further, the paper showed the achievability of the commitment capacity even in the quantum setting. In addition, the paper [39] showed the impossibility part for interactive protocols by completing the proof by [39]. The proof in [39] covers the impossibility part for a certain class even in the quantum setting.

5.1 Algorithm based on the em-algorithm problem

To calculate the commitment capacity, we consider the following mixture and exponential families;

$$\begin{aligned}\mathcal{M}_a &:= \{W_{Y|X} \times P_X | P_X \in \mathcal{P}(\mathcal{X})\} \\ &= \left\{ P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \middle| \sum_{y \in \mathcal{Y}} P_{XY}(x, y) = P_X(x) \right\} \qquad (107)\\ \mathcal{E} &:= \{Q_Y \times P_{X,Uni} | Q_Y \in \mathcal{P}(\mathcal{Y})\}, \qquad (108)\end{aligned}$$

where $P_{X,Uni}$ is the uniform distribution on $\mathcal{X}$. In (107), the integer k is chosen to be $|\mathcal{X}| - 1$, the linear functions $f_1, \ldots, f_k$ are chosen to be $\sum_{y \in \mathcal{Y}} P_{XY}(x_1, y), \ldots, \sum_{y \in \mathcal{Y}} P_{XY}(x_{k-1}, y)$, and the vector a is chosen to be $P_X(x_1), \ldots, P_X(x_k)$. These choices clarify that (107) gives a mixture family.

Since $\Gamma_{\mathcal{E}}^{(m)}[W_{Y|X} \times P_X] = (W_{Y|X} \cdot P_X) \times P_{X,Uni}$, the commitment capacity is rewritten as

$$\begin{aligned}\log |\mathcal{X}| - C_c(W_{Y|X}) &= \min_{P_X} H(X)_{P_{X,Uni}} + H(X)_{W_{Y|X} \cdot P_X} - H(XY)_{W_{Y|X} \times P_X} \\ &= \min_{P_X} D(W_{Y|X} \times P_X \| (W_{Y|X} \cdot P_X) \times P_{X,Uni}) \\ &= \min_{P_X} D(W_{Y|X} \times P_X \| \Gamma_{\mathcal{E}}^{(m)}[W_{Y|X} \times P_X]). \qquad (109)\end{aligned}$$

Hence, the minimization (109) is a special case of the minimization (97). Since $\Gamma_{\mathcal{E}}^{(m)}[W_{Y|X} \times Q_X](x, y) = (W_{Y|X} \cdot Q_X) \times P_{X,Uni}$,

$$D(\Gamma_{\mathcal{E}}^{(m)}[P^*] \| \Gamma_{\mathcal{E}}^{(m)}[Q_X]) = D(W_{Y|X} \cdot P_X \| W_{Y|X} \cdot Q_X) \le D(P^* \| Q_X), \qquad (110)$$

which yields the condition (100). Hence, the global convergence is guaranteed.

By applying (101), Φ is calculated as

$$\begin{aligned}&\Phi[W_{Y|X} \times Q_X](x, y) \\ &= \frac{1}{\kappa^1_{W_{Y|X}}[Q_X]} w_{Y|X}(y|x)^{\frac{\gamma-1}{\gamma}} Q_X(x)^{\frac{\gamma-1}{\gamma}} (w_{Y|X} \cdot Q_X)(y)^{\frac{1}{\gamma}} P_{X,Uni}(x)^{\frac{1}{\gamma}}, \qquad (111)\end{aligned}$$

where $\kappa^1_{W_{Y|X}}[Q_X]$ is the normalizer. Then, after a complicated calculation, the marginal distribution of its projection to $\mathcal{M}_a$ is given as

$$\begin{aligned}&\int_{\mathcal{Y}} \Gamma^{(e)}_{\mathcal{M}_a}[\Phi[W_{Y|X} \times Q_X]](x, y) \mu(dy) \\ &= \frac{1}{\kappa^2_{W_{Y|X}}[Q_X]} Q_X(x)^{1-\frac{1}{\gamma}} \exp(-\frac{1}{\gamma} D(W_{Y|X=x} \| W_{Y|X} \cdot Q_X)), \qquad (112)\end{aligned}$$

where $\kappa^2_{W_{Y|X}}[Q_X]$ is the normalizer. In the algorithm, we update $P_X^{(t+1)}$ as $P_X^{(t+1)}(x) := \int_{\mathcal{Y}} \Gamma^{(e)}_{\mathcal{M}_a}[\Phi[W_{Y|X} \times P_X^{(t)}]](x, y)\mu(dy)$.

5.2 Direct application

The update formula (112) requires a complicated calculation, we can derive the same update rule by a simpler derivation as follows. The commitment capacity is rewritten as

$$\begin{aligned} -C_c(W_{Y|X}) &= \min_{P_X} I(P_X, W_{Y|X}) - H(X)_{P_X} \\ &= \min_{P_X} \sum_{x\in\mathcal{X}} P_X(x)(D(W_{Y|X=x}\|W_{Y|X}\cdot P_X) + \log P_X(x)). \end{aligned} \tag{113}$$

We choose $\mathcal{M}_a$ and Ψ as $\mathcal{P}(\mathcal{X})$ and

$$\Psi_{c,W_{Y|X}}[P_X](x) := D(W_{Y|X=x}\|W_{Y|X}\cdot P_X) + \log P_X(x). \tag{114}$$

Then, we have

$$\begin{aligned} &\sum_{x\in\mathcal{X}} P_X(x)(\Psi[P_X](x) - \Psi[Q_X](x)) \\ &\quad = D(P_X\|Q_X) - D(W_{Y|X}\cdot P_X\|W_{Y|X}\cdot Q_X) \ge 0 \end{aligned} \tag{115}$$

and

$$D(P_X\|Q_X) \ge D(P_X\|Q_X) - D(W_{Y|X}\cdot P_X\|W_{Y|X}\cdot Q_X). \tag{116}$$

Since the condition (A1) with $\gamma = 1$ and the condition (A2) hold, Algorithm 1 converges with $\gamma = 1$. In this case, Φ is given as

$$\begin{aligned} &\Phi[Q_X](x) \\ &= \frac{1}{\kappa^3_{W_{Y|X}}[Q_X]} Q_X(x)\exp(-\frac{1}{\gamma}(\log Q_X(x) + D(W_{Y|X=x}\|W_{Y|X}\cdot Q_X))) \\ &= \frac{1}{\kappa^3_{W_{Y|X}}[Q_X]} Q_X(x)^{1-\frac{1}{\gamma}}\exp(-\frac{1}{\gamma}D(W_{Y|X=x}\|W_{Y|X}\cdot Q_X)), \end{aligned} \tag{117}$$

where the normalizing constant $\kappa^3_{W_{Y|X}}[Q_X]$ is given as $\kappa^3_{W_{Y|X}}[Q_X] := \sum_{x\in\mathcal{X}} Q_X(x)^{1-\frac{1}{\gamma}}\exp(-\frac{1}{\gamma}D(W_{Y|X=x}\|W_{Y|X}\cdot Q_X))$. Since $\Phi[Q_X] \in \mathcal{M}_a$, $P_X^{(t+1)}$ is given as $\Phi[P_X^{(t)}]$.

To consider the effect of the acceleration parameter γ, we made a numerical comparison when the channel with $\mathcal{X} = \{1, 2, 3, 4\}$ and $\mathcal{Y} = \{1, 2, 3, 4\}$ is given as

follows.

$$\begin{aligned}
&W_{Y|X}(1,1)=0.6, W_{Y|X}(2,1)=0.2, W_{Y|X}(3,1)=0.1, W_{Y|X}(4,1)=0.1,\\
&W_{Y|X}(1,2)=0.1, W_{Y|X}(2,2)=0.2, W_{Y|X}(3,2)=0.1, W_{Y|X}(4,2)=0.6,\\
&W_{Y|X}(1,3)=0.1, W_{Y|X}(2,3)=0.2, W_{Y|X}(3,3)=0.15, W_{Y|X}(4,3)=0.55,\\
&W_{Y|X}(1,4)=0.05, W_{Y|X}(2,4)=0.85, W_{Y|X}(3,4)=0.05, W_{Y|X}(4,4)=0.05.
\end{aligned} \tag{118}$$

We choose γ to be 1, 0.95, and 0.9. Figure 1 shows the numerical result for the iteration of our algorithm when the channel input is limited into $\{1, 2, 3\}$. A smaller γ does not improve the convergence in this case. Figure 2 shows the same numerical result when all elements of $\{1, 2, 3, 4\}$ are allowed as the channel input. In this case, a smaller γ improves the convergence.

6 Reverse em problem

6.1 General problem description

In this section, given a pair of an exponential family $\mathcal{E}$ and a mixture family $\mathcal{M}_a$ on $\mathcal{X}$, we consider the following maximization;

$$\begin{aligned}
&\max_{P\in\mathcal{M}_a}\min_{Q\in\mathcal{E}} D(P\|Q) = \max_{P\in\mathcal{M}_a} D(P\|\Gamma_{\mathcal{E}}^{(m)}[P])\\
&= \max_{P\in\mathcal{M}_a}\sum_{x\in\mathcal{X}} P(x)(\log P(x) - \log \Gamma_{\mathcal{E}}^{(m)}[P](x))
\end{aligned} \tag{119}$$

while Sect. 4 considers the minimization of the same value. When $\mathcal{M}_a$ is given as (107) and $\mathcal{E}$ is given as $\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$, this problem coincides with the channel capacity (58). This problem was firstly studied for the channel capacity in [20], and was discussed with a general form in [21]. To discuss this problem, we choose the function Ψ as $\Psi_{\rm rem} := -\Psi_{\rm em}$, and apply the discussion in Sect. 2. Due to (99), (24) in the condition (A1) is written as

$$(\gamma+1)D(P^0\|Q) \ge D(\Gamma_{\mathcal{E}}^{(m)}[P^0]\|\Gamma_{\mathcal{E}}^{(m)}[Q]), \tag{120}$$

and (25) in the condition (A2) is written as

$$D(\Gamma_{\mathcal{E}}^{(m)}[P^0]\|\Gamma_{\mathcal{E}}^{(m)}[Q]) \ge D(P^0\|Q). \tag{121}$$

Further, due to Lemma 4, the condition (A4) holds.

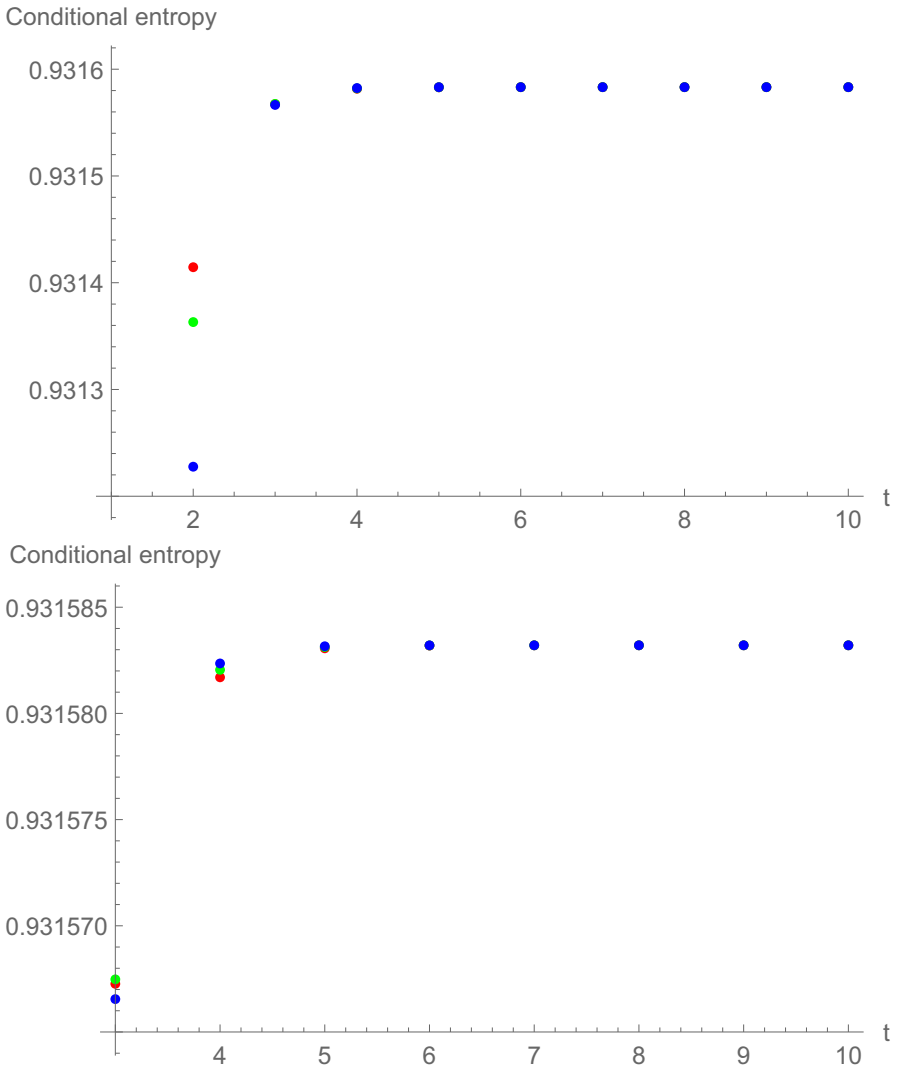

Fig. 1 Calculation of commitment capacity for the channel given in (118) with $\mathcal{X} = \{1, 2, 3\}$. The lower plot shows an enlarged plot of the upper plot. The horizontal axis shows the number of iterations. The vertical axis shows the conditional entropy. Red points show the case with $\gamma = 1$. Green points show the case with $\gamma = 0.95$. Blue points show the case with $\gamma = 0.9$. For $t = 5, 6, \ldots, 10$, these cases have almost the same value. Hence, these plots cannot be distinguished for $t = 5, 6, 7, 8, 9, 10$. At $t = 2, 3$, the case with $\gamma = 1$ is better than other cases. However, in this case, a smaller γ does not improve the convergence

6.2 Application to wiretap channel

Now, we apply this problem setting to wiretap channel with the degraded case discussed in Sect. 3.4.2. We choose $\mathcal{M}_a$ as $\{W_{YZ|X} \times P_X | P_X \in \mathcal{P}(\mathcal{X})\}$ and $\mathcal{E}$ as the set of distributions with the Markov chain condition $X - Z - Y$ [42]. Then, the conditional mutual

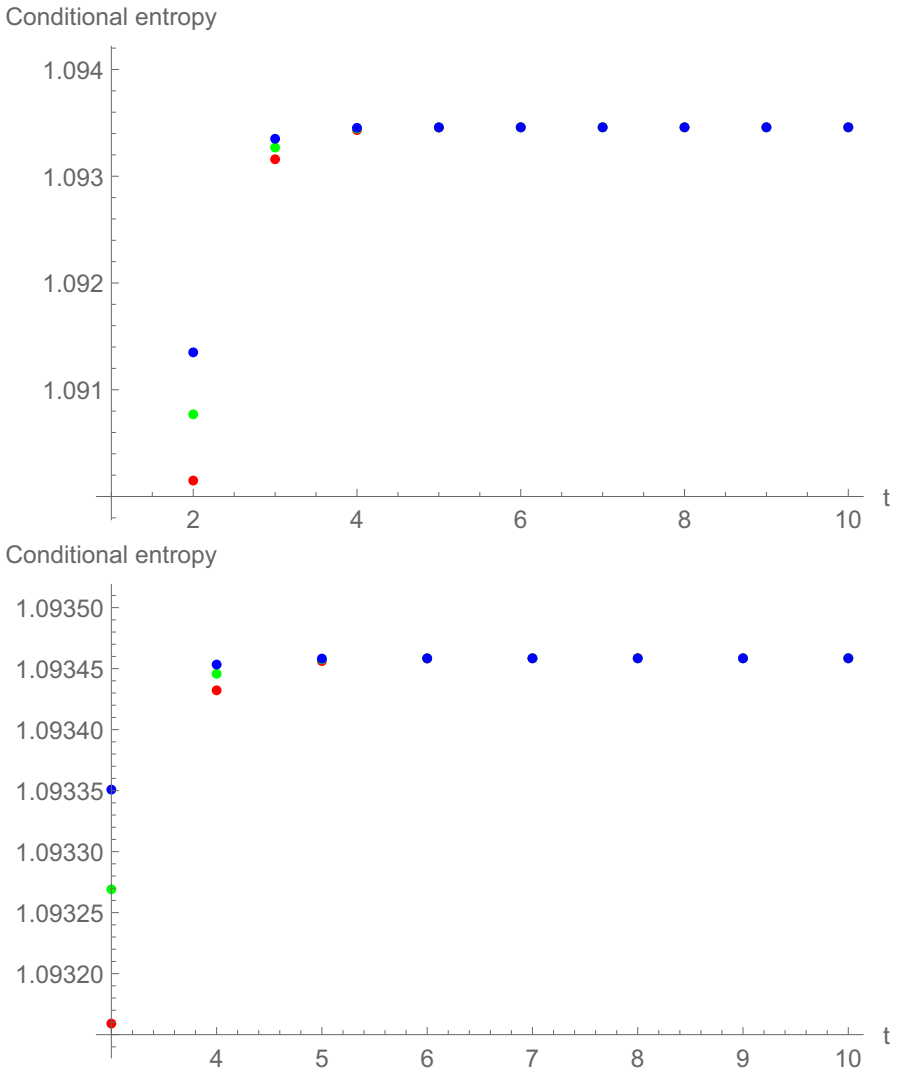

Fig. 2 Calculation of commitment capacity for the channel given in (118) with $\mathcal{X} = \{1, 2, 3, 4\}$. The role of color is the same as Fig. 1. In this case, a smaller γ improves the convergence

information $I(X; Y|Z)[P_X, W_{YZ|X}]$ is given as $D(W_{YZ|X} \times P_X \| \Gamma_{\mathcal{E}}^{(m)}[W_{YZ|X} \times P_X])$. In this application, we have

$$D_{\Psi_{\mathrm{rem}}}(W_{YZ|X} \times P_X \| W_{YZ|X} \times Q_X) = D_{\Psi_{W_{YZ|X}}} D(P_X \| Q_X) \tag{122}$$

$$D(W_{YZ|X} \times P_X \| W_{YZ|X} \times Q_X) = D(P_X \| Q_X). \tag{123}$$

To check the conditions (A1) and (A2) for $\Psi_{W_{YZ|X}}$, it is sufficient to check them for Ψ_{rem} in this application. Since we have

$$\begin{aligned} &D(\Gamma_{\mathcal{E}}^{(m)}[W_{YZ|X} \times P_X] \| \Gamma_{\mathcal{E}}^{(m)}[W_{YZ|X} \times Q_X]) \\ &\quad = D(W_{Z|X} \times P_X \| Q_{XZ}) + D(W_{YZ|X} \cdot P_X \| W_{YZ|X} \cdot Q_X) \\ &\qquad - D(W_{Z|X} \cdot P_X \| W_{Z|X} \cdot Q_X) \\ &\quad = D(P_X \| Q_X) + D(W_{YZ|X} \cdot P_X \| W_{YZ|X} \cdot Q_X) \\ &\qquad - D(W_{Z|X} \cdot P_X \| W_{Z|X} \cdot Q_X) \\ &\quad \le 2D(P_X \| Q_X), \end{aligned} \tag{124}$$

LHS of (24) in the condition (A1) is written as

$$\begin{aligned} &\gamma D(P_X^0 \| Q_X) - D(W_{YZ|X} \cdot P_X \| W_{YZ|X} \cdot Q_X) \\ &\qquad + D(W_{Z|X} \cdot P_X \| W_{Z|X} \cdot Q_X) \\ &\quad \ge \gamma D(P^0 \| Q) - D(P_X^0 \| Q_X). \end{aligned} \tag{125}$$

It is not negative when $\gamma \ge 1$. Also, RHS of (25) in the condition (A2) is written as

$$D(W_{YZ|X} \cdot P_X \| W_{YZ|X} \cdot Q_X) - D(W_{Z|X} \cdot P_X \| W_{Z|X} \cdot Q_X) \ge 0. \tag{126}$$

Hence, the conditions (A1) and (A2) hold with $\gamma \ge 1$.

7 Information bottleneck

As a method for information-theoretical machine learning, we often consider information bottleneck [43]. Consider two correlated systems $\mathcal{X}$ and $\mathcal{Y}$ and a joint distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. The task is to extract an essential information from the space $\mathcal{X}$ to $\mathcal{T}$ with respect to $\mathcal{Y}$. Here, we discuss a generalized problem setting proposed in [44]. For this information extraction, given parameters $\alpha \in [0, 1]$ and $\beta \ge \alpha$, we choose a conditional distribution $P_{T|X}^*$ as

$$P_{T|X}^* := \operatorname*{argmin}_{P_{T|X}} \alpha I(T; X) + (1 - \alpha) H(T) - \beta I(T; Y). \tag{127}$$

This method is called information bottleneck. To apply our method to this problem, we set $\mathcal{M}_a$ to be $\mathcal{P}(\mathcal{X})$, and define

$$\begin{aligned} \Psi_{\alpha,\beta}[P_{TX}](t, x) := &\alpha \log P_{TX}(t, x) - \alpha \log P_X(x) + (\beta - 1) \log P_T(t) \\ &+ \beta \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x)(\log P_Y(y) - \log P_{TY}(t, y)). \end{aligned} \tag{128}$$

Then, when the joint distribution P_{TX} is chosen to be $P_{T|X} \times P_X$, the objective function is written as

$$\mathcal{G}_{\alpha,\beta}(P_{TX}) := \sum_{t\in\mathcal{T},x\in\mathcal{X}} P_{TX}(t,x)\Psi_{\alpha,\beta}[P_{TX}](t,x). \tag{129}$$

That is, our problem is reduced to the minimization

$$\min_{P_{TX}\in\mathcal{M}(P_X)} \mathcal{G}_{\alpha,\beta}(P_{TX}), \tag{130}$$

where $\mathcal{M}(P_X)$ is the set of joint distributions on $\mathcal{X}\times\mathcal{Y}$ whose marginal distribution on $\mathcal{X}$ is P_X. The e-projection $\Gamma^{(e)}_{\mathcal{M}(P_X)}$ to $\mathcal{M}(P_X)$ is written as

$$\Gamma^{(e)}_{\mathcal{M}(P_X)}[Q_{TX}] = Q_{T|X} \times P_X \tag{131}$$

because the relation $D(P_{TX}\|Q_{TX}) = D(P_X\|Q_X) + D(P_{TX}\|Q_{T|X}\times P_X)$ holds for a distribution $P_{TX}\in\mathcal{M}(P_X)$.

When Algorithm 1 is applied to this problem, due to (131), the update rule for the conditional distribution is given as

$$P^{(t)}_{T|X} \mapsto P^{(t+1)}_{T|X}(t|x) := \kappa_x P^{(t)}_{T|X}(t|x)\exp(-\frac{1}{\gamma}\Psi_{\alpha,\beta}[P^{(t)}_{T|X}\times P_X](t,x)), \tag{132}$$

where κ_x is a normalizing constant. This update rule is the same as the update rule proposed in Sect. 3 of [45] when the states $\{\rho_{Y|x}\}$ are given as diagonal density matrices, i.e., a conditional distribution $P_{Y|X}$. Also, as shown in [45, (22)], we have the relation

$$D_{\Psi_{\alpha,\beta}}(P_{TX}\|Q_{TX}) \le \alpha D(P_{TX}\|Q_{TX}) \tag{133}$$

for P_{TX}, $Q_{TX}\in\mathcal{M}(P_X)$ as follows. First, we have

$$D(P_{T|X}\cdot P_{XY}\|Q_{T|X}\cdot P_{XY}) \ge D(P_T\|Q_T), \tag{134}$$

where $P_{T|X}\cdot P_{XY}(t,y) := \sum_{x\in\mathcal{X}} P_{T|X}(t|x)P_{XY}(x,y)$. Then, we have

$$\begin{aligned}
&D_{\Psi_{\alpha,\beta}}(P_{TX}\|Q_{TX})\\
&= (\beta-1)D(P_T\|Q_T) + \alpha\sum_{x\in\mathcal{X}} P_X(x)D(P_{T|X=x}\|Q_{T|X=x})\\
&- \beta D(P_{T|X}\cdot P_{XY}\|Q_{T|X}\cdot P_{XY})\\
&\le -D(P_T\|Q_T) + \alpha\sum_{x\in\mathcal{X}} P_X(x)D(P_{T|X=x}\|Q_{T|X=x})\\
&\le \alpha\sum_{x\in\mathcal{X}} P_X(x)D(P_{T|X=x}\|Q_{T|X=x}) = \alpha D(P_{TX}\|Q_{TX}).
\end{aligned} \tag{135}$$

That is, the condition (A1) holds with $\gamma = \alpha$. Therefore, with $\gamma = \alpha$, Theorem 1 guarantees that Algorithm 1 converges to a local minimum, which was shown as [45, Theorem 3]. This fact shows the importance of the choice of γ dependently on the problem setting. That is, it shows the necessity of our problem setting with a general positive real number $\gamma > 0$.

The paper [45] also discussed the case when $\mathcal{Y}$ and $\mathcal{T}$ are quantum systems. It numerically compared these algorithms depending on γ [45, Fig. 2]. This numerical calculation indicates the following behavior. When γ is larger than a certain threshold, a smaller γ realizes faster convergence. But, when γ is smaller than a certain threshold, the algorithm does not converge.

8 Conclusion

We have proposed iterative algorithms with an acceleration parameter for a general minimization problem over a mixture family. For these algorithms, we have shown convergence theorems in various forms, one of which covers the case with approximated iterations. Then, we have applied our algorithms to various problem settings including the em algorithm and several information theoretical problem settings.

There are two existing studies to numerically evaluate the effect of the acceleration parameter γ [19, 45]. They reported improvement in the convergence by modifying the acceleration parameter γ. For example, in the numerical calculation for information bottleneck in [45, Fig. 2], the case with $\gamma = 0.55$ improves the convergence. Our numerical calculation for the commitment capacity has two cases. In one case, the choices with $\gamma = 0.95, 0.9$ do not improve the convergence. In another case, the choices with $\gamma = 0.95, 0.9$ improve the convergence. These facts show that the effect of the acceleration parameter γ depends on the parameters of the problem setting. The commitment capacity is considered as a special case of the divergence between a mixture family and an exponential family.

There are several future research directions. The first direction is the evaluation of the convergence speed of Algorithm 3 because we could not derive its evaluation. The second direction is to find various applications of our methods. Although this paper studied several examples, in order to clarify the usefulness of our algorithm, it is needed to find more useful examples for our algorithm. The third direction is the extensions of our results. A typical extension is the extension to the quantum setting [46–48]. As a further extension, it is an interesting topic to extend our result to the setting with Bregman divergence. Recently, the Bregman proximal gradient algorithm has been studied for the minimization of a convex function [49–51]. Since this algorithm uses Bregman divergence, it might have an interesting relation with the above-extended algorithm. Therefore, it is an interesting study to investigate this relation.

9 Useful lemma

To show various theorems, we prepare the following lemma.

Lemma 5 *For any two distributions $Q, Q' \in \mathcal{M}_a$, we have*

$$\begin{aligned} & D(P^0\|Q) - D(P^0\|Q') \\ &= \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]], Q) - \frac{1}{\gamma}\mathcal{G}(P^0) + \frac{1}{\gamma} D_\Psi(P^0\|Q) \\ &\quad - D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q') \qquad (136) \\ &= \frac{1}{\gamma}\mathcal{G}(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]) - \frac{1}{\gamma}\mathcal{G}(P^0) + D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q) \\ &\quad - \frac{1}{\gamma} D_\Psi(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q) \\ &\quad + \frac{1}{\gamma} D_\Psi(P^0\|Q) - D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q'). \qquad (137) \end{aligned}$$

In addition, when Ψ is defined for any distribution in $\mathcal{P}(\mathcal{X})$, the above relations holds for any distribution $Q \in \mathcal{P}(\mathcal{X})$.

Proof We have

$$\begin{aligned} \mathcal{G}(P^0) &= \sum_{x\in\mathcal{X}} P^0(x)\Psi[P^0](x) = J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^0]], P^0) \\ &= \gamma(D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^0]]\|\Phi[P^0]) - \log\kappa[P^0]). \qquad (138) \end{aligned}$$

Using (138), we have

$$\begin{aligned} & D(P^0\|Q) - D(P^0\|Q') = \sum_{x\in\mathcal{X}} P^0(x)(\log Q'(x) - \log Q(x)) \\ &= \sum_{x\in\mathcal{X}} P^0(x)\Big(\log Q'(x) - \log\Phi[Q](x) + \log\Phi[Q](x) - \log Q(x)\Big) \\ &\overset{(a)}{=} D(P^0\|\Phi[Q]) - D(P^0\|Q') + \sum_{x\in\mathcal{X}} P^0(x)\Big(-\frac{1}{\gamma}\Psi[Q](x) - \log\kappa[Q]\Big) \\ &\overset{(b)}{=} D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|\Phi[Q]) - D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q') \\ &\quad - \log\kappa[Q] - \frac{1}{\gamma}\sum_{x\in\mathcal{X}} P^0(x)\Psi[Q](x) \\ &\overset{(c)}{=} \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]], Q) - \frac{1}{\gamma}\mathcal{G}(P^0) + \frac{1}{\gamma}\sum_{x\in\mathcal{X}} P^0(x)\Big(\Psi[P^0](x) - \Psi[Q](x)\Big) \\ &\quad - D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q') \\ &\overset{(d)}{=} \frac{1}{\gamma}\mathcal{G}(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]) - \frac{1}{\gamma}\mathcal{G}(P^0) + D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q) \\ &\quad - \frac{1}{\gamma}\sum_{x\in\mathcal{X}} \Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]](x)(\Psi[\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]](x) - \Psi[Q](x)) \end{aligned}$$

$$+\frac{1}{\gamma}\sum_{x\in\mathcal{X}}P^0(x)\Big(\Psi[P^0](x)-\Psi[Q](x)\Big)-D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q'), \tag{139}$$

where each step is shown as follows. (a) follows from the definition of $\Phi(Q)$. (c) follows from (10) and (138). (d) follows from (17). (b) follows from the relations

$$D(P^0\|\Phi[Q])=D(P^0\|\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]])+D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|\Phi[Q]) \tag{140}$$

$$D(P^0\|Q')=D(P^0\|\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]])+D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[Q]]\|Q'), \tag{141}$$

which are shown by the Pythagorean equation. Therefore, considering the definition of $D_\Psi(P\|Q)$, we obtain (136) and (137). □

10 Proof of Theorem 2 and Corollary 1

Step 1: This step aims to show the following inequalities by assuming that item (i) does not hold and the conditions (A1) and (A2) hold.

$$D(P^0\|P^{(t+1)})\le\delta \tag{142}$$

$$D(P^0\|P^{(t)})-D(P^0\|P^{(t+1)})\ge\frac{1}{\gamma}\mathcal{G}(P^{(t+1)})-\frac{1}{\gamma}\mathcal{G}(P^0) \tag{143}$$

for $t=1,\dots,t_0-1$. We show these relations by induction.

For any t, by using the relation $\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]]=P^{(t+1)}$, the application of (137) of Lemma 5 to the case with $Q=P^{(t)}$ and $Q'=P^{(t+1)}$ yields

$$\begin{aligned}&D(P^0\|P^{(t)})-D(P^0\|P^{(t+1)})\\&=\frac{1}{\gamma}\mathcal{G}(P^{(t+1)})-\frac{1}{\gamma}\mathcal{G}(P^0)+D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]]\|P^{(t)})\\&\quad-\frac{1}{\gamma}D_\Psi(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]]\|\Psi[P^{(t)}])+\frac{1}{\gamma}D_\Psi(P^0\|P^{(t)}).\end{aligned} \tag{144}$$

First, we show the relations (142) and (143) with $t=1$. Since $D(P^0\|P^{(1)})\le\delta$, $P^{(1)}$ belongs to $U(P^0,\delta)$. Hence, the conditions (A1) and (A2) guarantee the following inequality with $t=1$;

$$(\text{RHS of } 144)\ge\frac{1}{\gamma}\mathcal{G}(P^{(t+1)})-\frac{1}{\gamma}\mathcal{G}(P^0). \tag{145}$$

The combination of (144) and (145) implies (143). Since item (i) does not hold, we have

$$\frac{1}{\gamma}\mathcal{G}(P^{(t+1)})-\frac{1}{\gamma}\mathcal{G}(P^0)\ge 0. \tag{146}$$

The combination of (144), (145), and (146) implies (142).

Next, we show the relations (142) and (143) with $t = t'$ by assuming the relations (142) and (143) with $t = t' - 1$. Since the assumption guarantees $D(P^0 \| P^{(t')}) \le \delta$, the conditions (A1) and (A2) guarantee (145) with $t = t'$. We obtain (142) and (143) in the same way as $t = 1$.

Step 2: This step aims to show (28) by assuming that item (i) does not hold and the conditions (A1) and (A2) hold. Due to (142), the condition (A1) and Lemmas 1 and 2 guarantee that

$$\mathcal{G}(P^{(t+1)}) \le \mathcal{G}(P^{(t)}). \tag{147}$$

We have

$$\begin{aligned}
&\frac{t_0}{\gamma}\Big(\mathcal{G}(P^{(t_0+1)}) - \mathcal{G}(P^0)\Big) \overset{(a)}{\le} \frac{1}{\gamma}\sum_{t=1}^{t_0} \mathcal{G}(P^{(t+1)}) - \mathcal{G}(P^0) \\
&\overset{(b)}{\le} \sum_{t=1}^{t_0} D(P^0 \| P^{(t)}) - D(P^0 \| P^{(t+1)}) \\
&= D(P^0 \| P^{(1)}) - D(P^0 \| P^{(t_0+1)}) \\
&\le D(P^0 \| P^{(1)}),
\end{aligned} \tag{148}$$

where (a) and (b) follow from (147) and (143), respectively.

Step 3: This step aims to show item (ii) by assuming the conditions (A0) as well as (A1) and (A2). In the discussion of Step 1, since $D(P^0 \| P^{(t)}) \le \delta$, the condition (A0) guarantees (146). We can show item (ii) with assuming that item (i) does not hold. Hence, we obtain Theorem 2.

Step 4: To show Corollary 1, we apply (144) to the case when $P^0 = P^*$ and $P^{(t)} = P_i^*$. Then, we have

$$0 = \mathcal{G}(P_i^*) - \mathcal{G}(P^*) + \sum_{x \in \mathcal{X}} P^*(x)(\Psi[P^*](x) - \Psi[P_i^*](x)), \tag{149}$$

which implies (31).

11 Proof of Theorem 3

We have already shown that $\{P^{(t)}\}_{t=1}^{t_0+1} \subset U(P^0, \delta)$ when item (i) does not hold. Hence, in the following, we show only (30) by using (A1), (A3), and $\{P^{(t)}\}_{t=1}^{t_0+1} \subset U(P^0, \delta)$ when item (i) does not hold.

We have

$$\begin{aligned}
&J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]], P^{(t)}) - \mathcal{G}(P^0) \\
&\overset{(a)}{=} \mathcal{G}(P^{(t+1)}) - \mathcal{G}(P^0) + \gamma D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]] \| P^{(t)})
\end{aligned} \tag{150}$$

$$
\begin{aligned}
&- D_\Psi(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]] \| \Psi[P^{(t)}]) \\
&\overset{(b)}{\geq} \mathcal{G}(P^{(t+1)}) - \mathcal{G}(P^0) \overset{(c)}{\geq} 0, \qquad (151)
\end{aligned}
$$

where (a) follows from Lemma 2, (b) follows from the condition (A1) and $P^{(t)} \in U(P^0, \delta)$, and (c) holds because item (i) does not hold.

Since $\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]] = P^{(t+1)}$, the application of (136) of Lemma 5 to the case with $Q = P^{(t)}$ and $Q' = P^{(t+1)}$ yields

$$
\begin{aligned}
&D(P^0 \| P^{(t)}) - D(P^0 \| P^{(t+1)}) \\
&= \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]], P^{(t)}) - \frac{1}{\gamma}\mathcal{G}(P^0) + \frac{1}{\gamma} D_\Psi(P^0 \| P^{(t)}) \qquad (152) \\
&\overset{(a)}{\geq} \frac{1}{\gamma} D_\Psi(P^0 \| P^{(t)}) \overset{(b)}{\geq} \frac{\beta}{\gamma} D(P^0 \| P^{(t)}), \qquad (153)
\end{aligned}
$$

where (a) follows from (151), and (b) follows from (26) in the condition (A3) and $P^{(t)} \in U(P^0, \delta)$. Hence, we have

$$
D(P^0 \| P^{(t+1)}) \leq (1 - \frac{\beta}{\gamma}) D(P^0 \| P^{(t)}). \qquad (154)
$$

Using the above relations, we have

$$
\begin{aligned}
\mathcal{G}(P^{(t+1)}) - \mathcal{G}(P^0) &\overset{(a)}{\leq} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[P^{(t)}]], P^{(t)}) - \mathcal{G}(P^0) \\
&\overset{(b)}{=} D(P^0 \| P^{(t)}) - D(P^0 \| P^{(t+1)}) - \frac{1}{\gamma} D_\Psi(P^0 \| P^{(t)}) \\
&\overset{(c)}{\leq} D(P^0 \| P^{(t)}) - D(P^0 \| P^{(t+1)}) - \frac{\beta}{\gamma} D(P^0 \| P^{(t)}) \\
&\overset{(d)}{\leq} (1 - \frac{\beta}{\gamma}) D(P^0 \| P^{(t)}) \overset{(e)}{\leq} (1 - \frac{\beta}{\gamma})^t D(P^0 \| P^{(1)}), \qquad (155)
\end{aligned}
$$

where each step is derived as follows. Step (a) follows from (151). Step (b) follows from (152). Step (c) follows from (26) in the condition (A3) and $P^{(t)} \in U(P^0, \delta)$. Step (d) follows from (153). Step (e) follows from (154). Hence, we obtain (30). Therefore, we have shown item (ii) under the conditions (A1) and (A3) when item (i) does not hold.

When (A0) holds in addition to (A1) and (A3), as shown in Step 1 of the proof of Theorem 2, the relation $\{P^{(t)}\}_{t=1}^{t_0+1} \subset U(P^0, \delta)$ holds. Hence, item (ii) holds.

12 Proof of Theorem 4

In this proof, we choose $\bar{P}^{(1)}$ to be $P^{(1)}$.

Step 1: This step aims to show the inequality (45). We denote the maximizer in (41) by θ'. The condition (41) implies that

$$\phi[\bar{P}^{(t)}](\theta) - \sum_{j=1}^{k} \theta^j a_j \le \phi[\bar{P}^{(t)}](\theta') - \sum_{j=1}^{k} \theta'^j a_j + \epsilon_1. \tag{156}$$

The divergence in the exponential family $\{Q_\theta\}$ can be considered as the Bregmann divergence of the potential function $\phi[\bar{P}^{(t)}](\theta)$. For example, for this fact, see [22, Section III-A]. Hence, we have

$$\begin{aligned} D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t)}]] \| \bar{P}^{(t+1)}) &= \phi[\bar{P}^{(t)}](\theta) - \sum_{j=1}^{k} \theta^j a_j \\ &\quad - \Big(\phi[\bar{P}^{(t)}](\theta') - \sum_{j=1}^{k} \theta'^j a_j\Big) \\ &\le \epsilon_1. \end{aligned} \tag{157}$$

Step 2: This step aims to show Eq. (46) when the following inequality

$$\mathcal{G}(P^{(t_2)}) - \gamma D(P^{(t_2)} \| \bar{P}^{(t_2)}) \le \frac{\gamma}{t_1 - 1} D(P^0 \| \bar{P}^{(1)}) + \epsilon_1 + \mathcal{G}(P^0) \tag{158}$$

holds. Eq. (46) is shown as follows;

$$\begin{aligned} \mathcal{G}(P_f^{(t_1)}) - \mathcal{G}(P^0) &\overset{(a)}{=} \mathcal{G}(P^{(t_2)}) - \mathcal{G}(P^0) \\ &\overset{(b)}{\le} \frac{\gamma}{t_1 - 1} D(P^0 \| \bar{P}^{(1)}) + \gamma D(P^{(t_2)} \| \bar{P}^{(t_2)}) + \epsilon_1 \\ &\overset{(b)}{\le} \frac{\gamma}{t_1 - 1} D(P^0 \| \bar{P}^{(1)}) + \gamma \epsilon_2 + \epsilon_1, \end{aligned} \tag{159}$$

where Steps (a), (b), and (c) follow from the definition of $P_f^{(t_1)}$, (158), and (42), respectively. Therefore, the remaining task is the proof of (158).

Step 3: We choose $t_4 \in [1, t_1 - 1]$ as the minimum integer $t \in [1, t_1 - 1]$ to satisfy the following inequality

$$\frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t)}]], \bar{P}^{(t)}) \le \frac{1}{\gamma} \mathcal{G}(P^0) + \epsilon_1. \tag{160}$$

If no integer $t \in [1, t_1 - 1]$ satisfies (160), we set t_4 to be t_1. This step aims to show the following two facts for $t = 1, \ldots, t_4 - 1$. (i) $D(P^0 \| \bar{P}^{(t+1)}) \le \delta$. (ii) The inequality

$$D(P^0 \| \bar{P}^{(t)}) - D(P^0 \| \bar{P}^{(t+1)}) \ge \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t)}]], \bar{P}^{(t)})$$

$$- \frac{1}{\gamma}\mathcal{G}(P^0) - \epsilon_1 \tag{161}$$

holds. The above two items are shown by induction for t as follows. It is sufficient to show the case when $t_4 \geq 2$.

We show items (i) and (ii) for $t = 1$ as follows. The application of Lemma 5 to the case with $Q = \bar{P}^{(1)}$ and $Q' = \bar{P}^{(2)}$ yields

$$\begin{aligned}
&D(P^0\|\bar{P}^{(1)}) - D(P^0\|\bar{P}^{(2)}) \\
&= \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(1)}]], \bar{P}^{(1)}) - \frac{1}{\gamma}\mathcal{G}(P^0) \\
&\quad + \frac{1}{\gamma} D_\Psi(P^0\|P^{(1)}) - D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(1)}]]\|\bar{P}^{(2)}) \\
&\overset{(a)}{\geq} \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(1)}]], \bar{P}^{(1)}) - \frac{1}{\gamma}\mathcal{G}(P^0) - \epsilon_1 \overset{(b)}{\geq} 0,
\end{aligned} \tag{162}$$

where (a) follows from (A2+) and (45) because the relation $D(P^0\|\bar{P}^{(1)}) \leq \delta$ follows from the assumption of this theorem. (b) follows from the fact that $t = 1$ does not satisfy the condition (160). Hence, $D(P^0\|\bar{P}^{(2)}) \leq D(P^0\|\bar{P}^{(1)}) \leq \delta$.

Assume that items (i) and (ii) hold with $t = t' - 1$. Then, the application of Lemma 5 to the case with $Q = \bar{P}^{(t)}$ and $Q' = \bar{P}^{(t+1)}$ yields

$$\begin{aligned}
&D(P^0\|\bar{P}^{(t')}) - D(P^0\|\bar{P}^{(t'+1)}) \\
&= \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t')}]], \bar{P}^{(t')}) - \frac{1}{\gamma}\mathcal{G}(P^0) \\
&\quad + \frac{1}{\gamma} D_\Psi(P^0\|\bar{P}^{(t)}) - D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t')}]]\|\bar{P}^{(t'+1)}) \\
&\overset{(a)}{\geq} \frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t')}]], \bar{P}^{(t')}) - \frac{1}{\gamma}\mathcal{G}(P^0) - \epsilon_1 \overset{(b)}{\geq} 0,
\end{aligned} \tag{163}$$

where (a) follows from (A2+) and (45) because the relation $D(P^0\|\bar{P}^{(t')}) \leq \delta$ follows from the assumption of induction. (b) follows from the fact that $t = t'$ does not satisfy the condition (160). Hence, $D(P^0\|\bar{P}^{(t'+1)}) \leq D(P^0\|\bar{P}^{(t')}) \leq \delta$.

Step 4: This step aims to show the inequality (158) when $t_4 \leq t_1 - 1$, i.e., there exists an integer $t \in [1, t_1 - 1]$ to satisfy (160).

Pythagorean theorem guarantees

$$\begin{aligned}
&D(P^{(t_4+1)}\|\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_4)}]]) \\
&\leq D(P^{(t_4+1)}\|\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_4)}]]) + D(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_4)}]]\|\bar{P}^{(t_4+1)}) \\
&= D(P^{(t_4+1)}\|\bar{P}^{(t_4+1)}) \leq \epsilon_2.
\end{aligned} \tag{164}$$

Then, we have

$$\begin{aligned}
&\mathcal{G}(P^{(t_2)}) - \gamma D(P^{(t_2)} \| \bar{P}^{(t_2)}) \overset{(a)}{\le} \mathcal{G}(P^{(t_4+1)}) - \gamma D(P^{(t_4+1)} \| \bar{P}^{(t_4+1)}) \\
&\overset{(b)}{\le} J_\gamma(P^{(t_4+1)}, \bar{P}^{(t_4)}) - \gamma D(P^{(t_4+1)} \| \bar{P}^{(t_4+1)}) \\
&\overset{(c)}{=} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_4)}]], \bar{P}^{(t_4)}) + \gamma D(P^{(t_4+1)} \| \Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_4)}]]) \\
&\quad - \gamma D(P^{(t_4+1)} \| \bar{P}^{(t_4+1)}) \\
&\overset{(d)}{\le} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_4)}]], \bar{P}^{(t_4)}),
\end{aligned} \tag{165}$$

where each step is derived as follows. Step (a) follows from the relation $t_2 = \operatorname{argmin}_{t=2,\ldots,t_1} \mathcal{G}(P^{(t)}) - \gamma D(P^{(t)} \| \bar{P}^{(t)})$. Step (b) follows from Lemma 2 and the condition (A1+) because (164) holds, and the relation $D(P^0 \| \bar{P}^{(t_4)}) \le \delta$ follows from item (i) with $t = t_4 - 1$ shown in Step 3. Step (c) follows from (12). Step (d) follows from the equation (164).

Combining (165) and (160), we have

$$\mathcal{G}(P^{(t_2)}) - \gamma D(P^{(t_2)} \| \bar{P}^{(t_2)}) \le \epsilon_1 + \mathcal{G}(P^0), \tag{166}$$

which implies (158).

Step 5: This step aims to show

$$J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_3)}]], \bar{P}^{(t_3)}) - \mathcal{G}(P^0) - \epsilon_1 \le \frac{\gamma}{t_1 - 1} D(P^0 \| \bar{P}^{(1)}) \tag{167}$$

under the choice of $t_3 := \operatorname{argmin}_{1 \le t \le t_1 - 1} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t)}]], \bar{P}^{(t)})$ when $t_4 = t_1$, i.e., there exists no integer $t \in [1, t_1 - 1]$ to satisfy (160).

Using (161), we have

$$\begin{aligned}
&\frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_3)}]], \bar{P}^{(t_3)}) - \frac{1}{\gamma}\mathcal{G}(P^0) - \epsilon_1 \\
&\quad \le D(P^0 \| \bar{P}^{(t)}) - D(P^0 \| \bar{P}^{(t+1)})
\end{aligned} \tag{168}$$

for $t \le t_1 - 1$. Taking the sum for (168), we have

$$\begin{aligned}
&\frac{1}{\gamma} J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_3)}]], \bar{P}^{(t_3)}) - \frac{1}{\gamma}\mathcal{G}(P^0) - \epsilon_1 \\
&\quad \le \frac{1}{t_1 - 1} \sum_{t=1}^{t=t_1-1} D(P^0 \| \bar{P}^{(t)}) - D(P^0 \| \bar{P}^{(t+1)}) \\
&\quad = \frac{1}{t_1 - 1}(D(P^0 \| \bar{P}^{(1)}) - D(P^0 \| \bar{P}^{(t_1)})) \le \frac{1}{t_1 - 1} D(P^0 \| \bar{P}^{(1)}).
\end{aligned} \tag{169}$$

Therefore, we obtain (167).

Step 6: This step aims to show the inequality (158) when $t_4 = t_1$, i.e., there exists no integer $t \in [1, t_1 - 1]$ to satisfy (160). We obtain the following inequality

$$\mathcal{G}(P^{(t_2)}) - \gamma D(P^{(t_2)} \| \bar{P}^{(t_2)}) \le J_\gamma(\Gamma^{(e)}_{\mathcal{M}_a}[\Phi[\bar{P}^{(t_3)}]], \bar{P}^{(t_3)}) \tag{170}$$

in the same way as (165) in Step 4 by changing t_4 by t_3. Combining (170) and (167), we obtain (158).

Acknowledgements The author was supported in part by the National Natural Science Foundation of China (Grant No. 62171212) and Guangdong Provincial Key Laboratory (Grant No. 2019B121203002). The author is very grateful to Mr. Shoji Toyota for helpful discussions. In addition, he pointed out that the secrecy capacity can be written as the reverse em algorithm in a similar way as the channel capacity [42] under the degraded condition.

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The author states that there is no Conflict of interest.

References

1. Amari, S.: Information Geometry and Its Applications, Springer Japan (2016)
2. Amari, S.: Information geometry of the EM and em algorithms for neural networks. Neural Netw. **8**(9), 1379–1408 (1995)
3. Fujimoto, Y., Murata, N.: A modified EM algorithm for mixture models based on Bregman divergence. Ann. Inst. Stat. Math. **59**, 3–25 (2007)
4. Allassonnière, S., Chevallier, J.: A New Class of EM Algorithms. Escaping Local Minima and Handling Intractable Sampling, Computational Statistics & Data Analysis, Elsevier, vol. 159(C), (2019)
5. Amari, S., Kurata, K., Nagaoka, H.: Information geometry of Boltzmann machines. IEEE Trans. Neural Netw. **3**(2), 260–271 (1992)
6. Amari, S., Nagaoka, H.: Methods of Information Geometry (AMS and Oxford, 2000)
7. Amari, S.: α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes. IEEE Trans. Inform. Theory **55**, 4925–4931 (2009)
8. Arimoto, S.: An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Trans. Inform. Theory **18**(1), 14–20 (1972)
9. Blahut, R.: Computation of channel capacity and rate-distortion functions. IEEE Trans. Inform. Theory **18**(4), 460–473 (1972)
10. Shannon, C.E.: A Mathematical Theory of Communication, Bell Syst. Tech. J. 27, 379 – 423 and 623 – 656 (1948)
11. Csiszár, I.: On the computation of rate-distortion functions. IEEE Trans. Inform. Theory **20**(1), 122–124 (1974)
12. Cheng, S., Stankovic, V., Xiong, Z.: Computing the channel capacity and rate-distortion function with two-sided state information. IEEE Trans. Inform. Theory **51**(12), 4418–4425 (2005)
13. Yasui, K., Suko, T., Matsushima, T.: On the Global Convergence Property of Extended Arimoto-Blahut Algorithm, EICE Trans. Fundamentals, J91-A(9), 846-860, (2008). (**In Japanese**)
14. Yasui, K., Suko, T., Matsushima, T.: An Algorithm for Computing the Secrecy Capacity of Broadcast Channels with Confidential Messages. In: Proc. 2007 IEEE Int. Symp. Information Theory (ISIT 2007), Nice, France, 24-29 June 2007, pp. 936 – 940

15. Nagaoka, H.: Algorithms of Arimoto-Blahut type for computing quantum channel capacity. In: Proc. 1998 IEEE Int. Symp. Information Theory (ISIT 1998), Cambridge, MA, USA, 16-21 Aug. (1998), pp. 354
16. Dupuis, F., Yu, W., Willems, F.: Blahut-Arimoto algorithms for computing channel capacity and rate-distortion with side information. In: Proc. 2014 IEEE Int. Symp. Information Theory (ISIT 2014), Chicago, IL, USA, 27 June-2 July 2004, pp. 179
17. Sutter, D., Sutter, T., Esfahani, P.M., Renner, R.: Efficient approximation of quantum channel capacities. IEEE Trans. Inform. Theory **62**, 578–598 (2016)
18. Li, H., Cai, N.: A Blahut-Arimoto type algorithm for computing classical-quantum channel capacity. In: Proc. 2019 IEEE Int. Symp. Information Theory (ISIT 2019), Paris, France, 7-12 July (2019), pp. 255–259
19. Ramakrishnan, N., Iten, R., Scholz, V.B., Berta, M.: Computing quantum channel capacities. IEEE Trans. Inform. Theory **67**, 946–960 (2021)
20. Toyota, S.: Geometry of Arimoto algorithm. Inf. Geom. **3**, 183–198 (2020)
21. Hayashi, M.: Reverse em-problem based on Bregman divergence and its application to classical and quantum information theory, Submitted to Information Geometry; arXiv: 2201.02447 (2022)
22. Hayashi, M.: Bregman divergence based em algorithm and its application to classical and quantum rate distortion theory. IEEE Trans. Inform. Theory **69**, 3460–3492 (2023)
23. Csiszár, I., Tusnády, G.: Information geometry and alternating minimization procedures. Stat. Decis. Suppl. Issue **1**, 205–2377 (1984)
24. O'Sullivan, J.A.: Alternating minimization algorithms: From Blahut-Arimoto to expectation-maximization". In: Vardy, A. (ed.) Codes, Curves, and Signals, pp. 173–192. Kluwer Academic, Norwell (1998)
25. Gallager, R.G.: Information Theory and Reliable Communication. Wiley, New York (1968)
26. Arimoto, S.: On the converse to the coding theorem for discrete memoryless channels. IEEE Trans. Inform. Theory **19**, 357–359 (1973)
27. Hayashi, M.: Quantum wiretap channel with non-uniform random number and its exponent and equivocation rate of leaked information. IEEE Trans. Inform. Theory **61**(10), 5595–5622 (2015)
28. Wyner, A.D.: The wire-tap channel. Bell. Syst. Tech. J. **54**, 1355–1387 (1975)
29. Csiszár, I., Körner, J.: Broadcast channels with confidential messages. IEEE Trans. Inform. Theory **24**(3), 339–348 (1978)
30. Csiszár, I.: Almost independence and secrecy capacity. Probl. Inf. Transm. **32**(1), 40–47 (1996)
31. Hayashi, M.: General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel. IEEE Trans. Inform. Theory **52**(4), 1562–1575 (2006)
32. Hayashi, M.: Exponential decreasing rate of leaked information in universal random privacy amplification. IEEE Trans. Inform. Theory **57**(6), 3989–4001 (2011)
33. Bellare, M., Tessaro, S., Vardy, A.: Semantic security for the wiretap channel. In: Proc. 32nd Annu. Cryptol. Conf. 7417, 294-311 (2012)
34. Hayashi, M., Matsumoto, R.: Secure Multiplex Coding with Dependent and Non-Uniform Multiple Messages. IEEE Trans. Inform. Theory **62**(5), 2355–2409 (2016)
35. Boyd, S., Vandenberghe, L.: Convex Optimization, Cambridge University Press (2004)
36. Winter, A., Nascimento, A.C.A., Imai, H.: Commitment Capacity of Discrete Memoryless Channels. In: Proc. 9th IMA International Conferenece on Cryptography and Coding (Cirencester 16-18 December 2003), pp. 35-51, (2003)
37. Imai, H., Morozov, K., Nascimento, A.C.A., Winter, A.: Commitment Capacity of Discrete Memoryless Channels, https: arXiv:cs/0304014
38. Imai, H., Morozov, K., Nascimento, A.C.A., Winter, A.: Efficient protocols achieving the commitment capacity of noisy correlations. In: Proc. IEEE International Symposium on Information Theory (ISIT2006), Seattle, Washington, USA July 9 – 14, 1432-1436 (2006)
39. Hayashi, M., Warsi, N.: Commitment capacity of classical-quantum channels. IEEE Trans. Inform. Theory **69**(8), 5083–5099 (2023)
40. Yamamoto, H., Isami, D.: Multiplex Coding of Bit Commitment Based on a Discrete Memoryless Channel. In: Proc. IEEE ISIT 2007, pp. 721 – 725, June 24-29, (2007)
41. Hayashi, M.: Secure list decoding and its application to bit-string commitment. IEEE Trans. Inform. Theory **68**(6), 3620–3642 (2022)
42. Toyota, S.: Private communication (2019)

43. Tishby, N., Pereira, F.C., Bialek. W.: The information bottleneck method, In: The 37th annual Allerton Conference on Communication, Control, and Computing, pages 368 – 377. Univ. Illinois Press, 1999. https://doi.org/10.48550/arXiv.physics/0004057
44. Strouse, D.J., Schwab, D.J.: The deterministic information Bottleneck. Neural Comput. **29**(6), 1611–1630 (2017)
45. Hayashi, M., Yang, Y.: Efficient algorithms for quantum information bottleneck. Quantum **7**, 936 (2023)
46. Holevo, A.S.: The capacity of the quantum channel with general signal states. IEEE Trans. Inform. Theory **44**, 269 (1998)
47. Schumacher, B., Westmoreland, M.D.: Sending classical information via noisy quantum channels. Phys. Rev. A **56**, 131 (1997)
48. Hayashi, M.: Quantum Information Theory: Mathematical Foundation, Graduate Texts in Physics, Springer-Verlag, (2017)
49. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM J. Optim. **3**(3), 538–543 (1993)
50. Teboulle, M.: Convergence of proximal-like algorithms. SIAM J. Optim. **7**(4), 1069–1083 (1997)
51. Zhou, Y., Liang, Y., Shen, L.: A simple convergence analysis of Bregman proximal gradient algorithm. Comput. Optim. Appl. **73**(3), 903–912 (2019)
52. Beck, A.: First-Order Methods in Optimization. SIAM, MOS-SIAM Series on Optimization (2017)
53. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
54. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. Ser. B **140**, 125–161 (2013)
55. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. **16**(3), 697–725 (2006)
56. Nesterov, Y.: A method for solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. - Doklady **27**(2), 372–376 (1983)
57. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Boston (2004)
58. Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. Ser. B **170**, 67–96 (2018)