On Convergence of Discrete Schemes for Computing the Rate-Distortion Function of Continuous Source

Lingyi Chen[1], Shitong Wu[1], Wenyi Zhang[3†], Huihui Wu[2], and Hao Wu[1]

[1]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
[2]Yangtze Delta Region Institute (Huzhou),
University of Electronic Science and Technology of China, Huzhou, Zhejiang, 313000, P.R. China.
[3]Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei, Anhui 230027, China
Email: wenyizha@ustc.edu.cn

Abstract

Computing the rate-distortion function for continuous sources is commonly regarded as a standard continuous optimization problem. When numerically addressing this problem, a typical approach involves discretizing the source space and subsequently solving the associated discrete problem. However, existing literature has predominantly concentrated on the convergence analysis of solving discrete problems, usually neglecting the convergence relationship between the original continuous optimization and its associated discrete counterpart. This neglect is not rigorous, since the solution of a discrete problem does not necessarily imply convergence to the solution of the original continuous problem, especially for non-linear problems. To address this gap, our study employs rigorous mathematical analysis, which constructs a series of finite-dimensional spaces approximating the infinite-dimensional space of the probability measure, establishing that solutions from discrete schemes converge to those from the continuous problems.

I. Introduction

The lossy source coding theory [1] and the rate-distortion function (RDF) for discrete memoryless sources were first established by Shannon [2], [3]. Until the end of 1960s, major developments of the classical setup had been summarized in [4], including the lossy source coding theorem and associated RDF for sources and reproductions defined in abstract spaces. The RDF that characterizes the fundamental tradeoff between compression rate and

† Corresponding author.

This work was partially supported by National Key Research and Development Program of China (2018YFA0701603) and National Natural Science Foundation of China Grant (12271289 and 62231022).

distortion, and plays a fundamental role in information theory. Several recent problems, including the rate-distortion-perception (RDP) function [5] and the information bottleneck (IB) [6], have been formulated based on the concept of RDF. The RDF, along with its inspired extensions, has found extensive applications in fields such as image and video compression [7], [8] as well as machine learning [9], [10] problems.

The computation of RDF for discrete source and reproduction alphabets can be solved by the well-known Blahut-Arimoto (BA) algorithm [11], [12], which has been widely applied due to its conceptual simplicity. Recently, an intriguing connection has been identified between optimal transport (OT) and rate-distortion (RD) [13], [14]. Leveraging insights from the OT perspective, more efficient algorithms have been developed for the computation of RDF; see, *e.g.*, [15], [16].

For RD problems where the source alphabet is continuous, these algorithms are not directly applicable, and a general approach is to first discretize the continuous source space, and then solve the resulting discrete RD problem. The solution of the discrete RD problem can be ensured by these algorithms. However, one issue still remains and has usually been neglected: do the solutions of the discrete problems actually converge to those of the original continuous problems? To the best of our knowledge, we have not seen discussions of this issue in the relevant literature on the RD theory and related topics. From a mathematical perspective, the continuous RD problem is a non-linear optimization problem. As is well-known, non-linear problems do not necessarily yield the solutions of the original continuous problems after discretization [17]. Moreover, similar discussion is also presented for OT problems, and the convergence of solutions for discrete problems to those for continuous problems [18], [19] is also specialized. Hence, it is necessary to give a rigorous analysis for this issue.

Motivated by the discussion above, we propose a general framework for continuous RD problems and establish the convergence of their associated discrete problems. The main idea in our proof is considering a proper form of the RDF and constructing a series of finite-dimensional spaces to approximate the infinite-dimensional space of the probability measure for the reproduction. Leveraging functional space analysis and estimation techniques [20] and inspired by [21], we rigorously prove the desired convergence property. Notably, this convergence result is independent of specific numerical algorithms for discrete problems.

Furthermore, based on our framework, we provide a proof that $O(\frac{m|\log\varepsilon|}{\varepsilon^{d+1}})$ arithmetic operations are needed to achieve ε accuracy using the BA algorithm [11], [12] and that $O\big(\frac{m|\log\varepsilon|}{\varepsilon^{d+1}}(1+\log|\log\varepsilon|)\big)$ arithmetic operations while using the recently proposed Constrained BA (CBA) algorithm [15], which solves the RDF directly rather than the RD Lagrangian in the BA algorithm. Here m is the number of discretized nodes of the source X when conducting numerical integration and d is the dimension of the reproduction space $\mathcal{Y}$. It is worth mentioning that our framework may also be applied to other information theory problems including the IB problem and the RDP problem, due to their close connections in discrete formats.

For continuous RD problems, there is a considerable interest in determining the support set of reproduction, since in general, the optimal reproduction of a continuous source, under a mean squared error distortion, is discrete [22]–[24]. To address this issue, a mapping approach has been proposed [24] to optimize both the locations and the probability masses of the reproduction symbols. However, as remarked in [25], there has not been theoretical guarantee for the mapping approach because taking into account the locations of reproduction symbols results in a non-convex optimization problem. In contrast, our method considers a series of discrete problems with their discretization step tending to zero, and the convergence is guaranteed since this does not involve determining the locations of reproduction symbols. While our convergence proof focuses on uniform grids of discretization, it is noteworthy that, with some technical modifications [26], our approach remains applicable to non-uniform grids. This adaptability is crucial for designing more efficient discretization schemes, and is left for future research.

The remaining of this article is organised as follows. In Section II, the continuous and discrete RD problems are introduced and some mathematical notations are given. In Section III, the main theorem showing the convergence of discrete schemes is established. Then, in Section IV, the convergence rate and the complexity estimation are further provided. In Section V, numerical experiments are given to corroborate the analysis. The conclusion is drawn in Section VI.

II. Preliminaries

A. The Continuous RD Problem

Consider a continuous memoryless source $X \in \mathcal{X}$ with reproduction $Y \in \mathcal{Y}$. Here, $\mathcal{X}, \mathcal{Y}$ are finite-dimensional Euclidean spaces. We denote the dimension of $\mathcal{Y}$ as d. In the sequel, denote the probability distributions of X, Y as p, $\boldsymbol{r}$ respectively and the distortion measure as $\rho(x, y)$.

The original RD problem can be expressed as a max-min form (see [27], equation (7)):

$$\max_{\beta \geq 0} \min_{\boldsymbol{r} \in W} - \int \log[\int \exp(-\beta\rho(x,y)) d\boldsymbol{r}(y)] dp(x) - \beta D, \tag{1}$$

where $W = \{\boldsymbol{r} : \int d\boldsymbol{r}(y) = 1\}$ and D is the average distortion threshold. In the sequel, we denote the objective function in (1) as $F(\boldsymbol{r}, \beta)$. Here, $\boldsymbol{r}$ should be understood as a distribution and the topology on W is the weak convergence topology [28].

By fixing the multiplier β, we have another form of the RD function (see [24], equation (3)), which has been used to develop the BA algorithm [11], [12]

$$\min_{\boldsymbol{r} \in W} f(\boldsymbol{r}) \triangleq - \int \log[\int \exp(-\beta\rho(x,y)) d\boldsymbol{r}(y)] dp(x). \tag{2}$$

B. Discretization of the Continuous RD Problem

Since $\boldsymbol{r}(y)$ is a continuous distribution, we discretize it at discrete points $\{y_j^n\}_{j=1}^n$, the equidistant nodes of $[-n^{\frac{1}{2d}}, n^{\frac{1}{2d}}]^d$. We denote the discretization step size as h, *i.e.*, $h = 2n^{-\frac{1}{2d}}$ and the interval $[y_j^n - h/2, y_j^n + h/2]$ containing y_j as I_j and note that these intervals are disjoint. For (2), the following discrete form can be given:

$$\min_{\boldsymbol{r}: \sum_{j=1}^n r_j = 1} - \int \log[\sum_{j=1}^n \exp(-\beta\rho(x, y_j^n)) r_j] dp(x). \tag{3}$$

Correspondingly, (1) has the following discrete form:

$$\max_{\beta \geq 0} \min_{\boldsymbol{r}: \sum_{j=1}^n r_j = 1} - \int \log[\sum_{j=1}^n \exp(-\beta\rho(x, y_j^n)) r_j] dp(x) - \beta D. \tag{4}$$

We can express their solution as $\boldsymbol{r}^n = \sum_j r_j \delta_{y_j^n}$, where $\delta_{y_j^n}$ is the δ distribution at y_j^n.

C. Notations from Mathematical Analysis

In this paper, $\|\cdot\|$ represents the infinity norm and $\to$ means convergence in $\mathbb{R}$, while $\rightrightarrows$ means uniform convergence [20], *i.e.*, $f_n \rightrightarrows f$ means that $\forall \varepsilon > 0$, $|f_n(x) - f(x)| \leq \varepsilon, \forall x$, when n is large enough. We also denote the set that contains all the limit points of the sequence $\{\boldsymbol{r}^n\}_{n=1}^{\infty}$ as $L(\{\boldsymbol{r}^n\}_{n=1}^{\infty})$. The limit points [28] of $\{\boldsymbol{r}^n\}_{n=1}^{\infty}$ are all the points satisfying that every punctured neighbourhood contains at least one point in $\{\boldsymbol{r}^n\}_{n=1}^{\infty}$. Moreover, $[-M, M]^d$ means the Cartesian product $[-M, M] \times [-M, M] \times \cdots [-M, M]$, where d intervals are involved. In the sequel, $\log \int \exp(-\beta\rho(x, y))d\boldsymbol{r}(y)$ appears frequently and for simplicity, we denote it as $\mathcal{H}_{\boldsymbol{r}(y)}(x)$.

III. Convergence on Discrete Schemes

To prove the convergence of discrete schemes, we need some mild assumptions:

$$\forall \beta > 0, \forall \varepsilon > 0, \exists \delta > 0, \text{ s.t. } \forall x, \forall \|y_1 - y_2\| < \delta, \quad \Big|\exp(-\beta\rho(x, y_1)) - \exp(-\beta\rho(x, y_2))\Big| < \varepsilon, \tag{5a}$$

$$\int \max_{\|y\| \leq A} \rho(x, y)dp(x) < \infty, \forall A > 0, \tag{5b}$$

$$\begin{gathered}\forall \beta > 0, \forall \varepsilon > 0, \exists \delta > 0, \text{ s.t. } \forall x, \forall \|y_1 - y_2\| < \delta, \\ \Big|\exp(-\beta\rho(x, y_1))\rho(x, y_1) - \exp(-\beta\rho(x, y_2))\rho(x, y_2)\Big| < \varepsilon.\end{gathered} \tag{6}$$

These are reasonable assumptions since for most cases including the mean absolute and mean squared error distortions, they hold.

Theorem 1: Under Assumptions (5), the solutions $\boldsymbol{r}^n$ to the discrete problem (3) satisfy both value convergence and sequence convergence, *i.e.*,

$$f(\boldsymbol{r}^n) \to f^*, \text{ and } L(\{\boldsymbol{r}^n\}_{n=1}^{\infty}) \text{ is the solution set of (2),}$$

where f^* is the optimal value of the continuous problem (2).

Proof: We denote the actual optimal solution as $\boldsymbol{r}^*$ and $W_n = \{\sum_{j=1}^n r_j \delta_{y_j^n} | r_j \geq 0, \sum_{j=1}^n r_j = 1\}$. As shown in Section II-B, the continuous RD problem has the discrete form (3), which is in fact minimizing $f(\boldsymbol{r})$ as defined in (2), over W_n. Denote the optimal solution of the discrete problem as $\boldsymbol{r}^n$. Then $f(\boldsymbol{r}^n) \leq f(\boldsymbol{r}), \forall \boldsymbol{r} \in W_n$. We observe that if we can find $\tilde{\boldsymbol{q}}^n \in W_n$ satisfying $f(\tilde{\boldsymbol{q}}^n) \to f(\boldsymbol{r}^*)$, then we have

$$f(\boldsymbol{r}^*) \leq f(\boldsymbol{r}^n) \leq f(\tilde{\boldsymbol{q}}^n) \to f(\boldsymbol{r}^*). \tag{7}$$

The first inequality is due to the fact $W_n \subseteq W$. The relationship (7) would then lead to the convergence $f(\boldsymbol{r}^n) \to f(\boldsymbol{r}^*)$. Now, we let

$$\boldsymbol{q}^n = \sum_{j=1}^n A_j^n \delta_{y_j^n},$$

where $A_j^n = \int_{I_j} d\boldsymbol{r}^*$ and I_j is defined in Section II-B. Let

$$\tilde{\boldsymbol{q}}^n = \boldsymbol{q}^n \Big/ \int d\boldsymbol{q}^n.$$

Since

$$\int d\boldsymbol{q}^n = \sum_j \int_{I_j} d\boldsymbol{r}^* \leq 1,$$

we have

$$f(\tilde{\boldsymbol{q}}^n) = f(\boldsymbol{q}^n) + \log(\int d\boldsymbol{q}^n) \leq f(\boldsymbol{q}^n).$$

Since $\tilde{\boldsymbol{q}}^n \in W_n$ and $f(\tilde{\boldsymbol{q}}^n) \leq f(\boldsymbol{q}^n)$, we only need to prove $f(\boldsymbol{q}^n) \to f(\boldsymbol{r}^*)$. Before the proof of $f(\boldsymbol{q}^n) \to f(\boldsymbol{r}^*)$, we need to show the convergence after the truncation of $\mathcal{X}$ first. We let $A > 1$ satisfy

$$\int_{\|y\| \leq A-1} d\boldsymbol{r}^*(y) \geq 1/2,$$

and $M \geq A$ be given. We first do the following estimation for $x \in [-M, M]^d$

$$\begin{aligned}
&\Big| \int \exp(-\beta\rho(x,y)) d\boldsymbol{q}^n(y) - \int \exp(-\beta\rho(x,y)) d\boldsymbol{r}^*(y) \Big| \\
&\leq \sum_i \int_{I_i} \big| \exp(-\beta\rho(x,y)) - \exp(-\beta\rho(x,y_i^n)) \big| d\boldsymbol{r}^*(y) + \int_{|y| > n^{\frac{1}{2d}}} \exp(-\beta\rho(x,y)) d\boldsymbol{r}^*(y) \\
&\leq \sum_i \int_{I_i} \varepsilon d\boldsymbol{r}^*(y) + \int_{|y| > n^{\frac{1}{2d}}} d\boldsymbol{r}^*(y) \\
&\leq 2\varepsilon, \text{ for sufficiently large } n.
\end{aligned}$$

Here, we have used Assumptions (5) in the second inequality and used the fact

$$\boldsymbol{q}^n = \sum_j \int_{I_j} d\boldsymbol{r}^* \ \delta_{y_j^n}.$$

Next we establish the uniform convergence. Note that for $x \in [-M, M]^d$,

$$\begin{aligned}
&\int \exp(-\beta\rho(x,y)) d\boldsymbol{q}^n(y) \geq \int_{\|y\| \leq M} \exp(-\beta\rho(x,y)) d\boldsymbol{q}^n(y) \\
&\geq \int_{\|y\| \leq M} \exp(-\beta\rho^*) d\boldsymbol{q}^n(y) \geq \frac{1}{2} \exp(-\beta\rho^*) \triangleq \delta.
\end{aligned} \tag{8}$$

Here $\rho^* = \max_{x,y \in [-M,M]^d} \rho(x,y)$ and $\int_{\|y\| \leq M} d\boldsymbol{q}^n \geq \frac{1}{2}$ is due to

$$\int_{\|y\| \leq A} d\boldsymbol{q}^n(y) = \sum_{i: y_i^n \in [-A,A]^d} \int_{I_i} d\boldsymbol{r}^*(y) \geq \int_{\|y\| \leq A-1} d\boldsymbol{r}^*(y) \geq \frac{1}{2}. \tag{9}$$

Since $\log x$ is uniformly continuous in $[\delta, +\infty)$ and $\int \exp(-\beta\rho(x,y)) d\boldsymbol{q}^n(y)$ uniformly converges, we have

$$\log\Big(\int \exp(-\beta\rho(x,y)) d\boldsymbol{q}^n(y) \Big) \quad \rightrightarrows \quad \log\Big(\int \exp(-\beta\rho(x,y)) d\boldsymbol{r}^*(y) \Big), \quad x \quad \in \quad [-M, M]^d.$$

Then

$$\begin{aligned}
&\Big| \int_{\|x\| \leq M} \log\Big[\int \exp(-\beta\rho(x,y)) d\boldsymbol{q}^n(y) \Big] dp(x) - \int_{\|x\| \leq M} \log\Big[\int \exp(-\beta\rho(x,y)) d\boldsymbol{r}^*(y) \Big] dp(x) \Big| \\
&\leq \int_{\|x\| \leq M} \varepsilon dp(x) \leq \varepsilon, \text{ when } n \text{ is sufficiently large.}
\end{aligned}$$

Then as $n \to +\infty$, we have

$$\int_{\|x\| \leq M} \log\Big[\int \exp(-\beta\rho(x,y)) d\boldsymbol{q}^n(y) \Big] dp(x) \quad \to \quad \int_{\|x\| \leq M} \log\Big[\int \exp(-\beta\rho(x,y)) d\boldsymbol{r}^*(y) \Big] dp(x).$$

Thus, we obtain the convergence after truncation. In the following, based on this result, we prove the convergence $f(\boldsymbol{q}^n) \to f(\boldsymbol{r}^*)$. For $M \geq A$,

$$\begin{aligned} |f(\boldsymbol{q}^n) - f(\boldsymbol{r}^*)| &= \left| \int \mathcal{H}_{\boldsymbol{q}^n(y)}(x) dp(x) - \int \mathcal{H}_{\boldsymbol{r}^*(y)}(x) dp(x) \right| \\ &\leq \left| \int_{\|x\| \leq M} [\mathcal{H}_{\boldsymbol{q}^n(y)}(x) - \mathcal{H}_{\boldsymbol{r}^*(y)}(x)] dp(x) \right| + \int_{\|x\| > M} \Big(- \mathcal{H}_{\boldsymbol{q}^n(y)}(x) - \mathcal{H}_{\boldsymbol{r}^*(y)}(x) \Big) dp(x), \end{aligned} \tag{10}$$

where

$$\mathcal{H}_{\boldsymbol{r}(y)}(x) = \log \int \exp(-\beta \rho(x, y)) d\boldsymbol{r}(y).$$

We denote $g(x) = \max_{\|y\| \leq A} \rho(x, y)$, and note that

$$\int \exp(-\beta \rho(x, y)) d\boldsymbol{q}^n(y) \geq \exp(-\beta g(x)) \int_{\|y\| \leq A} d\boldsymbol{q}^n(y) \geq \frac{1}{2} \exp(-\beta g(x)).$$

Here the second inequality is similar to (9). Similarly,

$$\int \exp(-\beta \rho(x, y)) d\boldsymbol{r}^*(y) \geq \frac{1}{2} \exp(-\beta g(x)).$$

We thus obtain the following evaluation

$$\int_{\|x\| > M} \Big(- \mathcal{H}_{\boldsymbol{q}^n}(x) - \mathcal{H}_{\boldsymbol{r}^*}(x) \Big) dp(x) \leq \int_{\|x\| > M} (2 \log 2 + 2\beta g(x)) dp(x) \to 0, \text{ as } M \to +\infty.$$

Here Assumption (5b), *i.e.*, $\int g(x) dp(x) < \infty$ has been used. Thus, by (10), we have

$$|f(\boldsymbol{q}^n) - f(\boldsymbol{r}^*)| \leq \left| \int_{\|x\| \leq M} [\mathcal{H}_{\boldsymbol{q}^n(y)}(x) - \mathcal{H}_{\boldsymbol{r}^*(y)}(x)] dp(x) \right| + \int_{\|x\| > M} (2 \log 2 + 2\beta g(x)) dp(x).$$

By taking the upper limit of n for this inequality, we obtain

$$\limsup_n \Big| f(\boldsymbol{q}^n) - f(\boldsymbol{r}^*) \Big| \leq \int_{\|x\| > M} (2 \log 2 + 2\beta g(x)) dp(x).$$

By letting $M \to +\infty$, we obtain $f(\boldsymbol{q}^n) \to f(\boldsymbol{r}^*)$.

To prove the convergence of the solutions, we let $\tilde{\boldsymbol{r}}$ be a limit point of the solution sequence $\{\boldsymbol{r}^n\}_{n=1}^{\infty}$. Then there exists a subsequence $\{\boldsymbol{r}^{n_k}\}_{k=1}^{\infty}$ satisfying $\boldsymbol{r}^{n_k} \to \tilde{\boldsymbol{r}}$. Next, we have

$$f(\tilde{\boldsymbol{r}}) = \lim_k f(\boldsymbol{r}^{n_k}) = f(\boldsymbol{r}^*),$$

since we have proven $\lim_n f(\boldsymbol{r}^n) = f(\boldsymbol{r}^*)$. Therefore, we have shown that $f(\tilde{\boldsymbol{r}})$ is equal to the optimal value, and consequently $\tilde{\boldsymbol{r}}$ is an optimal solution.

■

Next, we provide the convergence proof of the RD problem (1). Due to its max-min form and additional variable β, the proof is more complicated than that of Theorem 1.

***Theorem** 2:* Under Assumptions (5) and (6), the solutions $(\boldsymbol{r}^n, \beta^n)$ to the discrete problem (4) satisfy both value convergence and sequence convergence, *i.e.*,

$$F(\boldsymbol{r}^n, \beta^n) \to F^*, \text{ and } L(\{(\boldsymbol{r}^n, \beta^n)\}_{n=1}^{\infty}) \text{ are solutions of (1)},$$

where F^* is the optimal value of the continuous problem (1).

Proof: We denote the optimal solution of the continuous problem (1) and the discrete problem (4) as $(\boldsymbol{r}^*, \beta^*)$, $(\boldsymbol{r}^n, \beta^n)$ respectively. Denote $W_n = \{\sum_{j=1}^n c_j \delta_{y_j^n} | c_j \geq 0, \sum_{j=1}^n c_j = 1\}$, where $\delta_{y_j^n}$ is the δ distribution at y_j^n. Similar to Theorem 1, the discrete problem (4) is equivalent to optimizing $F(\boldsymbol{r}, \beta)$ over W_n. By the property of max-min problems (4) and (1), we have $\forall \boldsymbol{r} \in W_n, \forall \beta \geq 0$,

$$F(\boldsymbol{r}^n, \beta^n) \leq F(\boldsymbol{r}, \beta^n),\ F(\boldsymbol{r}^n, \beta^n) \geq F(\boldsymbol{r}^n, \beta).$$

Furthermore, $\forall \boldsymbol{r} \in W, \forall \beta \geq 0$,

$$F(\boldsymbol{r}^*, \beta^*) \leq F(\boldsymbol{r}, \beta^*),\ F(\boldsymbol{r}^*, \beta^*) \geq F(\boldsymbol{r}^*, \beta).$$

Next, we construct $\boldsymbol{q}^n = \sum_{j=1}^n A_j^n \delta_{y_j^n}$, here

$$A_j^n = \int_{I_j} d\boldsymbol{r}^*,$$

I_j is the interval $[y_j^n - h/2, y_j^n + h/2]$ containing y_j. Let

$$\tilde{\boldsymbol{q}}^n = \boldsymbol{q}^n \Big/ \int d\boldsymbol{q}^n \in W_n.$$

Since $\int d\boldsymbol{q}^n \leq 1$, we have

$$F(\tilde{\boldsymbol{q}}^n, \beta) = F(\boldsymbol{q}^n, \beta) + \log(\int d\boldsymbol{q}^n) \leq F(\boldsymbol{q}^n, \beta), \quad \forall \beta \geq 0.$$

Then we have the following chain of inequalities:

$$F(\boldsymbol{r}^*, \beta^*) \leq F(\boldsymbol{r}^n, \beta^*) \leq F(\boldsymbol{r}^n, \beta^n) \leq F(\tilde{q}^n, \beta^n) \leq F(\boldsymbol{q}^n, \beta^n) \leq F(\boldsymbol{q}^n, \tilde{\beta}^n),$$

where $\tilde{\beta}^n = \text{argmax}_\beta F(\boldsymbol{q}^n, \beta)$. We observe that if $F(\boldsymbol{q}^n, \tilde{\beta}^n) \to F(\boldsymbol{r}^*, \beta^*)$, then we have

$$F(\boldsymbol{r}^*, \beta^*) \leq F(\boldsymbol{r}^n, \beta^n) \leq F(\boldsymbol{q}^n, \tilde{\beta}^n) \to F(\boldsymbol{r}^*, \beta^*). \tag{11}$$

Therefore, we obtain the convergence $F(\boldsymbol{r}^n, \beta^n) \to F(\boldsymbol{r}^*, \beta^*)$. So in summary, we only need to prove $F(\boldsymbol{q}^n, \tilde{\beta}^n) \to F(\boldsymbol{r}^*, \beta^*)$.

From the optimality of $\tilde{\beta}^n$, we know $\tilde{\beta}^n$ should satisfy the first order condition and is the root of a monotone function

$$G_{\boldsymbol{q}^n}(\beta) = \int \left(\int e^{-\beta \rho(x,y)} \rho(x,y) d\boldsymbol{q}^n(y) \right) \Big/ \left(\int e^{-\beta \rho(x,y)} d\boldsymbol{q}^n(y) \right) dp(x) - D.$$

The monotone property is due to Cauchy inequality

$$G'_{\boldsymbol{q}^n}(\beta) = -\int \frac{\left[\left(\int e^{-\beta\rho(x,y)} d\boldsymbol{q}^n(y) \right) \left(\int e^{-\beta\rho(x,y)} \rho(x,y)^2 d\boldsymbol{q}^n(y) \right) - \left(\int e^{-\beta\rho(x,y)} \rho(x,y) d\boldsymbol{q}^n(y) \right)^2 \right]}{\left(\int e^{-\beta\rho(x,y)} d\boldsymbol{q}^n(y) \right)^2} dp(x) < 0.$$

It does not equal to 0, otherwise, by the condition of equality for Cauchy inequality, $\rho(x, y)$ is a function only with respect to x, and this degenerate case can be disregarded.

We first show $\tilde{\beta}^n$ is lower bounded for n, *i.e.*, $\tilde{\beta}^n \geq B_1 > 0, \forall n$. Otherwise, there exists a subsequence $\tilde{\beta}^{n_k} \to 0$. We will show it is a contradiction. Since $\tilde{\beta}^{n_k} \to 0$, we have

$$\exists k_0, \forall k > k_0,\ \tilde{\beta}^{n_k} < \beta^*/2.$$

Here, we assume $\beta^* > 0$ and the degenerate case $\beta^* = 0$ can be disregarded, in which the rate equals to 0. Then by the monotone property of $G_{\boldsymbol{q}^{n_k}}$, we have

$$0 = G_{\boldsymbol{q}^{n_k}}(\tilde{\beta}^{n_k}) > G_{\boldsymbol{q}^{n_k}}(\beta^*/2).$$

Thus,

$$\begin{aligned} D &> \int \Big(\int e^{-\beta^* \rho(x,y)/2} \rho(x,y) d\boldsymbol{q}^{n_k}(y) \Big) \Big/ \Big(\int e^{-\beta^* \rho(x,y)/2} d\boldsymbol{q}^{n_k}(y) \Big) dp(x) \\ &\geq \int_{\|x\| \leq M} \Big(\int e^{-\beta^* \rho(x,y)/2} \rho(x,y) d\boldsymbol{q}^{n_k}(y) \Big) \Big/ \Big(\int e^{-\beta^* \rho(x,y)/2} d\boldsymbol{q}^{n_k}(y) \Big) dp(x). \end{aligned}$$

Here, M is a sufficiently large parameter. By the proof of Theorem 1, we have

$$\int e^{-\beta^* \rho(x,y)/2} d\boldsymbol{q}^{n_k}(y) \rightrightarrows \int e^{-\beta^* \rho(x,y)/2} d\boldsymbol{r}^*(y).$$

Similarly,

$$\int e^{-\beta^* \rho(x,y)/2} \rho(x,y) d\boldsymbol{q}^{n_k}(y) \rightrightarrows \int e^{-\beta^* \rho(x,y)/2} \rho(x,y) d\boldsymbol{r}^*(y).$$

And similar to (8) in the proof of Theorem 1, we have

$$\int e^{-\beta^* \rho(x,y)/2} d\boldsymbol{q}^{n_k}(y) \geq \delta_0.$$

Meanwhile, we have an upper bound estimation

$$\int e^{-\beta^* \rho(x,y)/2} \rho(x,y) d\boldsymbol{q}^{n_k}(y) \leq \int B \ d\boldsymbol{q}^{n_k}(y) \leq B.$$

Here, B is the bound of the function $e^{-\beta^* x/2} x, x \geq 0$. Next, since $g(x,y) = x/y$ is uniformly continuous in $[0,B] \times [\delta_0, \infty)$, we obtain

$$\frac{\Big(\int \exp(-\beta^* \rho(x,y)/2) \rho(x,y) d\boldsymbol{q}^{n_k}(y) \Big)}{\Big(\int \exp(-\beta^* \rho(x,y)/2) d\boldsymbol{q}^{n_k}(y) \Big)} \rightrightarrows \frac{\Big(\int \exp(-\beta^* \rho(x,y)/2) \rho(x,y) d\boldsymbol{r}^*(y) \Big)}{\Big(\int \exp(-\beta^* \rho(x,y)/2) d\boldsymbol{r}^*(y) \Big)}$$

Thus,

$$\begin{aligned} D &\geq \int_{\|x\| \leq M} \Big(\int e^{-\beta^* \rho(x,y)/2} \rho(x,y) d\boldsymbol{q}^{n_k}(y) \Big) \Big/ \Big(\int e^{-\beta^* \rho(x,y)/2} d\boldsymbol{q}^{n_k}(y) \Big) dp(x) \\ &\to \int_{\|x\| \leq M} \Big(\int e^{-\beta^* \rho(x,y)/2} \rho(x,y) d\boldsymbol{r}^*(y) \Big) \Big/ \Big(\int e^{-\beta^* \rho(x,y)/2} d\boldsymbol{r}^*(y) \Big) dp(x) \end{aligned} \tag{12}$$

We let $M \to \infty$, and get

$$G_{\boldsymbol{r}^*}(\beta^*/2) \leq 0 = G_{\boldsymbol{r}^*}(\beta^*).$$

Using the monotonicity of $G_{\boldsymbol{r}^*}$, we have $\beta^*/2 \geq \beta^*$, which is a contradiction.

Next, we will prove the convergence of $\tilde{\beta}^n$. To prove $\tilde{\beta}^n \to \beta^*$, we only need to show every convergent subsequence converges to β^*, *i.e.*, if a convergent subsequence $\tilde{\beta}^{n_k} \to \bar{\beta}$, then $\bar{\beta} = \beta^*$. Now, let $\tilde{\beta}^{n_k}$ be a convergent subsequence and $\tilde{\beta}^{n_k} \to \bar{\beta}$. Since $\tilde{\beta}^{n_k}$ is lower bounded, we have $\bar{\beta} > 0$. Let A_0 satisfies $e^{-\bar{\beta} t/2} t \leq \varepsilon$ and $e^{-\bar{\beta} t/2} \leq \varepsilon, \ \forall t \geq A_0$.

$$\begin{aligned} &\left| \int e^{-\tilde{\beta}^{n_k} \rho(x,y)} \rho(x,y) d\boldsymbol{q}^{n_k}(y) - \int e^{-\bar{\beta} \rho(x,y)} \rho(x,y) d\boldsymbol{q}^{n_k}(y) \right| \leq \int |e^{-\tilde{\beta}^{n_k} \rho(x,y)} - e^{-\bar{\beta} \rho(x,y)}| \rho(x,y) d\boldsymbol{q}^{n_k}(y) \\ &\leq \int e^{-\bar{\beta} \rho(x,y)} |e^{(\bar{\beta} - \tilde{\beta}^{n_k}) \rho(x,y)} - 1| \rho(x,y) d\boldsymbol{q}^{n_k}(y) \end{aligned}$$

Next, we divide it into two parts and estimate each part.

$$\begin{aligned}&\int_{y:\rho(x,y)>A_0} e^{-\bar{\beta}\rho(x,y)}|e^{(\bar{\beta}-\tilde{\beta}^{n_k})\rho(x,y)}-1|\rho(x,y)d\boldsymbol{q}^{n_k}(y)\\ &\le \int_{y:\rho(x,y)>A_0} e^{-\bar{\beta}\rho(x,y)}e^{|\bar{\beta}-\tilde{\beta}^{n_k}|\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\\ &\le \int_{y:\rho(x,y)>A_0} e^{-\bar{\beta}\rho(x,y)/2}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\\ &\le \int_{y:\rho(x,y)>A_0} \varepsilon d\boldsymbol{q}^{n_k}(y)\le\varepsilon,\ \text{when k is sufficiently large}\end{aligned}$$

here, the second inequality is due to $|\bar{\beta}-\tilde{\beta}^{n_k}|\le\bar{\beta}/2$, when k is sufficiently large.

$$\begin{aligned}&\int_{y:\rho(x,y)\le A_0} e^{-\bar{\beta}\rho(x,y)}|e^{(\bar{\beta}-\tilde{\beta}^{n_k})\rho(x,y)}-1|\rho(x,y)d\boldsymbol{q}^{n_k}(y)\\ &\le \int_{y:\rho(x,y)\le A_0} |e^{(\bar{\beta}-\tilde{\beta}^{n_k})\rho(x,y)}-1|A_0\ d\boldsymbol{q}^{n_k}(y)\\ &\le \int_{y:\rho(x,y)\le A_0} (e^{|\bar{\beta}-\tilde{\beta}^{n_k}|\rho(x,y)}-1)A_0\ d\boldsymbol{q}^{n_k}(y)\\ &\le \int_{y:\rho(x,y)\le A_0} (e^{|\bar{\beta}-\tilde{\beta}^{n_k}|A_0}-1)A_0\ d\boldsymbol{q}^{n_k}(y)\\ &\le (e^{|\bar{\beta}-\tilde{\beta}^{n_k}|A_0}-1)A_0\le\varepsilon,\ \text{when k is sufficiently large}\end{aligned}$$

Combining the two parts, we obtain

$$\left|\int e^{-\tilde{\beta}^{n_k}\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)-\int e^{-\bar{\beta}\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\right|\le 2\varepsilon,\ \text{when k is sufficiently large.}$$

Thus,

$$\int e^{-\tilde{\beta}^{n_k}\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)-\int e^{-\bar{\beta}\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\rightrightarrows 0,\ \forall x. \tag{13}$$

And by the proof of Theorem 1, we have

$$\int e^{-\bar{\beta}\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\rightrightarrows\int e^{-\bar{\beta}\rho(x,y)}\rho(x,y)d\boldsymbol{r}^*(y),\ \forall x.$$

Thus,

$$\int e^{-\tilde{\beta}^{n_k}\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\rightrightarrows\int e^{-\bar{\beta}\rho(x,y)}\rho(x,y)d\boldsymbol{r}^*(y),\ \forall x.$$

Similarly, we have

$$\int e^{-\tilde{\beta}^{n_k}\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\rightrightarrows\int e^{-\bar{\beta}\rho(x,y)}d\boldsymbol{r}^*(y),\ \forall x.$$

Next, we will prove $G_{\boldsymbol{q}^{n_k}}(\tilde{\beta}^{n_k})\to G_{\boldsymbol{r}^*}(\bar{\beta})$. Since $\tilde{\beta}^{n_k}\to\bar{\beta}$, we have $|\tilde{\beta}^{n_k}-\bar{\beta}|\le\varepsilon$, when k is sufficiently large. Using the monotonicity of $G_{\boldsymbol{q}^{n_k}}$, we obtain

$$G_{\boldsymbol{q}^{n_k}}(\bar{\beta}+\varepsilon)\le G_{\boldsymbol{q}^{n_k}}(\tilde{\beta}^{n_k})\le G_{\boldsymbol{q}^{n_k}}(\bar{\beta}-\varepsilon). \tag{14}$$

Next, We divide $G_{\boldsymbol{q}^{n_k}}(\bar{\beta}+\varepsilon)$ into two parts, and estimate each part. For simplicity, we denote $\bar{\beta}+\varepsilon$ as β. We first estimate the part

$$\int_{\|x\|>M}\Big(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\Big)\Big/\Big(\int e^{-\beta\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\Big)dp(x).$$

We denote $A_x = -\frac{1}{\beta}\ \log \int e^{-\beta\rho(x,y)} d\boldsymbol{q}^{n_k}(y)$, then

$$\begin{aligned}
&\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\\
&\leq \int_{y:\rho(x,y)\leq A_x} e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y) + \int_{y:\rho(x,y)>A_x} e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\\
&\leq \int_{y:\rho(x,y)\leq A_x} e^{-\beta\rho(x,y)}A_x\ d\boldsymbol{q}^{n_k}(y) + \int_{y:\rho(x,y)>A_x} e^{-\beta\max(A_x,1/\beta)}\max(A_x,1/\beta)d\boldsymbol{q}^{n_k}(y)\\
&\leq A_x\int e^{-\beta\rho(x,y)}\ d\boldsymbol{q}^{n_k}(y) + e^{-\beta\max(A_x,1/\beta)}\max(A_x,1/\beta)\\
&= A_x\exp(-\beta A_x) + e^{-\beta\max(A_x,1/\beta)}\max(A_x,1/\beta)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\int_{\|x\|>M}\Big(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\Big)\Big/\Big(\int e^{-\beta\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\Big)dp(x)\\
&\leq \int_{\|x\|>M}\Big(A_x\exp(-\beta A_x) + e^{-\beta\max(A_x,1/\beta)}\max(A_x,1/\beta)\Big)\Big/\exp(-\beta A_x)dp(x)\\
&\leq \int_{\|x\|>M}\Big(A_x + \max(A_x,1/\beta)\Big)dp(x)\\
&\leq \int_{\|x\|>M}\Big(2A_x + 1/\beta\Big)dp(x)
\end{aligned}$$

Note that

$$\begin{aligned}
A_x &= -\frac{1}{\beta}\ \log\int e^{-\beta\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\\
&\leq -\frac{1}{\beta}\ \log\int_{\|y\|\leq A} e^{-\beta\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\\
&\leq -\frac{1}{\beta}\ \log\int_{\|y\|\leq A} e^{-\beta g(x)}d\boldsymbol{q}^{n_k}(y)\\
&\leq -\frac{1}{\beta}\ \log(e^{-\beta g(x)}/2)\\
&= g(x) + (\log 2)/\beta,\ \text{ when k is sufficiently large.}
\end{aligned}$$

Here, A is the constant in Theorem 1 and $g(x) = \max_{\|y\|\leq A}\rho(x,y)$. And we get

$$\int\Big(2A_x + 1/\beta\Big)dp(x) < \infty.$$

Similar to (12), the other part

$$\begin{aligned}
&\int_{\|x\|\leq M}\Big(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\Big)\Big/\Big(\int e^{-\beta\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\Big)dp(x)\\
&\to \int_{\|x\|\leq M}\Big(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{r}^*(y)\Big)\Big/\Big(\int e^{-\beta\rho(x,y)}d\boldsymbol{r}^*(y)\Big)dp(x), \text{as } k\to\infty.
\end{aligned}$$

Finally, combining the estimate of the two parts, we have

$$\begin{aligned}
&G_{\boldsymbol{q}^{n_k}}(\beta) = \int\Big(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\Big)\Big/\Big(\int e^{-\beta\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\Big)dp(x) - D\\
&\leq \int_{\|x\|\leq M}\Big(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y)\Big)\Big/\Big(\int e^{-\beta\rho(x,y)}d\boldsymbol{q}^{n_k}(y)\Big)dp(x) + \int_{\|x\|>M}\Big(2A_x + 1/\beta\Big)dp(x) - D
\end{aligned}$$

We first let k goes to ∞,

$$\limsup_k G_{\boldsymbol{q}^{n_k}}(\beta) \le \int_{\|x\|\le M} \left(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{r}^*(y) \right) \Big/ \left(\int e^{-\beta\rho(x,y)} d\boldsymbol{r}^*(y) \right) dp(x) + \int_{\|x\|>M} \left(2A_x + 1/\beta \right) dp(x) - D$$

Then let $M \to \infty$, we obtain

$$\limsup_k G_{\boldsymbol{q}^{n_k}}(\beta) \le \int \left(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{r}^*(y) \right) \Big/ \left(\int e^{-\beta\rho(x,y)} d\boldsymbol{r}^*(y) \right) dp(x) - D = G_{\boldsymbol{r}^*}(\beta),$$

i.e., $\limsup_k G_{\boldsymbol{q}^{n_k}}(\bar{\beta}+\varepsilon) \le G_{\boldsymbol{r}^*}(\bar{\beta}+\varepsilon)$.

Meanwhile,

$$\begin{aligned} G_{\boldsymbol{q}^{n_k}}(\beta) &= \int \left(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y) \right) \Big/ \left(\int e^{-\beta\rho(x,y)} d\boldsymbol{q}^{n_k}(y) \right) dp(x) - D \\ &\ge \int_{\|x\|\le M} \left(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{q}^{n_k}(y) \right) \Big/ \left(\int e^{-\beta\rho(x,y)} d\boldsymbol{q}^{n_k}(y) \right) dp(x) - D \end{aligned}$$

Let $k \to \infty$ and then $M \to \infty$, we have

$$\liminf_k G_{\boldsymbol{q}^{n_k}}(\beta) \ge \lim_{M\to\infty} \int_{\|x\|\le M} \left(\int e^{-\beta\rho(x,y)}\rho(x,y)d\boldsymbol{r}^*(y) \right) \Big/ \left(\int e^{-\beta\rho(x,y)} d\boldsymbol{r}^*(y) \right) dp(x) - D = G_{\boldsymbol{r}^*}(\beta),$$

i.e., $\liminf_k G_{\boldsymbol{q}^{n_k}}(\bar{\beta}+\varepsilon) \ge G_{\boldsymbol{r}^*}(\bar{\beta}+\varepsilon)$. Thus, we obtain

$$\lim_k G_{\boldsymbol{q}^{n_k}}(\bar{\beta}+\varepsilon) = G_{\boldsymbol{r}^*}(\bar{\beta}+\varepsilon).$$

Similarly,

$$\lim_k G_{\boldsymbol{q}^{n_k}}(\bar{\beta}-\varepsilon) = G_{\boldsymbol{r}^*}(\bar{\beta}-\varepsilon).$$

Using (14), we have

$$G_{\boldsymbol{r}^*}(\bar{\beta}+\varepsilon) = \lim_k G_{\boldsymbol{q}^{n_k}}(\bar{\beta}+\varepsilon) \le \liminf_k G_{\boldsymbol{q}^{n_k}}(\tilde{\beta}^{n_k}) \le \limsup_k G_{\boldsymbol{q}^{n_k}}(\tilde{\beta}^{n_k}) \le \lim_k G_{\boldsymbol{q}^{n_k}}(\bar{\beta}-\varepsilon) = G_{\boldsymbol{r}^*}(\bar{\beta}-\varepsilon).$$

Then, let $\varepsilon \to 0$, we obtain $0 = \lim_k G_{\boldsymbol{q}^{n_k}}(\tilde{\beta}^{n_k}) = G_{\boldsymbol{r}^*}(\bar{\beta})$. Therefore, $G_{\boldsymbol{r}^*}(\bar{\beta}) = 0 = G_{\boldsymbol{r}^*}(\beta^*)$. Then by the strict monotonicity, we have $\bar{\beta} = \beta^*$. Thus $\tilde{\beta}^n \to \beta^*$ holds.

Now, we will show the convergence $F(\boldsymbol{q}^n, \tilde{\beta}^n) \to F(\boldsymbol{r}^*, \beta^*)$. By Theorem 1, we have $F(\boldsymbol{q}^n, \beta^*) \to F(\boldsymbol{r}^*, \beta^*)$. Thus we only need to show $F(\boldsymbol{q}^n, \tilde{\beta}^n) - F(\boldsymbol{q}^n, \beta^*) \to 0$.

$$0 \le F(\boldsymbol{q}^n, \tilde{\beta}^n) - F(\boldsymbol{q}^n, \beta^*) = -\int \log \frac{\int \exp(-\tilde{\beta}^n \rho(x,y)) d\boldsymbol{q}^n(y)}{\int \exp(-\beta^* \rho(x,y)) d\boldsymbol{q}^n(y)} dp(x) - (\tilde{\beta}^n - \beta^*)D$$

We divide the integration into two parts and estimate each part.

$$\begin{aligned} &- \int_{\|x\|\le M} \log \frac{\int \exp(-\tilde{\beta}^n \rho(x,y)) d\boldsymbol{q}^n(y)}{\int \exp(-\beta^* \rho(x,y)) d\boldsymbol{q}^n(y)} dp(x) \\ &= \int_{\|x\|\le M} \log \left(\frac{\int \exp(-\beta^* \rho(x,y)) d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n \rho(x,y)) d\boldsymbol{q}^n(y)}{\int \exp(-\tilde{\beta}^n \rho(x,y)) d\boldsymbol{q}^n(y)} + 1 \right) dp(x) \\ &\le \int_{\|x\|\le M} \frac{|\int \exp(-\beta^* \rho(x,y)) d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n \rho(x,y)) d\boldsymbol{q}^n(y)|}{\int \exp(-\tilde{\beta}^n \rho(x,y)) d\boldsymbol{q}^n(y)} dp(x) \\ &\le \int_{\|x\|\le M} \frac{|\int \exp(-\beta^* \rho(x,y)) d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n \rho(x,y)) d\boldsymbol{q}^n(y)|}{\delta_0} dp(x) \end{aligned}$$

Here, δ_0 is a lower bound

$$\begin{aligned}\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y) &\geq \int_{\|y\|\leq M} \exp(-B_0\rho(x,y))d\boldsymbol{q}^n(y)\\ &\geq \int_{\|y\|\leq M} \exp(-B_0\rho^*)d\boldsymbol{q}^n(y)\\ &\geq \frac{1}{2}\exp(-B_0\rho^*) \triangleq \delta_0.\end{aligned}$$

Here, B_0 is the upper bound of $\tilde{\beta}^n$, since it is convergent. And M is larger than the constant A in Theorem 1. ρ^* is a constant equals to $\max_{x,y\in[-M,M]^d}\rho(x,y)$. Using the same derivation as (13), we have

$$\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y) \rightrightarrows 0.$$

Thus,

$$\int_{\|x\|\leq M} \frac{|\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)|}{\delta_0}dp(x) \leq \varepsilon, \text{ when n is sufficiently large}$$

Meanwhile, we denote $g(x) = \max_{\|y\|\leq A}\rho(x,y)$, then

$$\begin{aligned}&-\int_{\|x\|>M}\log\frac{\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)}{\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y)}dp(x)\\ &= -\int_{\|x\|>M}\log\Big(\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)\Big)dp(x) - \int_{\|x\|>M}\log\Big(\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y)\Big)dp(x)\\ &\leq \int_{\|x\|>M}\Big[-\log\Big(\int_{\|y\|\leq A}\exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)\Big) - \log\Big(\int_{\|y\|\leq A}\exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y)\Big)\Big]dp(x)\\ &\leq \int_{\|x\|>M}\Big[-\log\Big(\int_{\|y\|\leq A}\exp(-\tilde{\beta}^n g(x))d\boldsymbol{q}^n(y)\Big) - \log\Big(\int_{\|y\|\leq A}\exp(-\beta^* g(x))d\boldsymbol{q}^n(y)\Big)\Big]dp(x)\\ &\leq \int_{\|x\|>M}\Big[-\log\Big(1/2\ \exp(-\tilde{\beta}^n g(x))\Big) - \log\Big(1/2\ \exp(-\beta^* g(x))\Big)\Big]dp(x)\\ &= \int_{\|x\|>M}\Big[2\log 2 + \tilde{\beta}^n g(x) + \beta^* g(x)\Big]dp(x)\\ &\leq \int_{\|x\|>M}\Big[2\log 2 + B_0 g(x) + \beta^* g(x)\Big]dp(x)\end{aligned}$$

Combining the two parts and taking the limit $n\to\infty$, we obtain

$$\begin{aligned}&\limsup_n[F(\boldsymbol{q}^n,\tilde{\beta}^n) - F(\boldsymbol{q}^n,\beta^*)] = \limsup_n -\int\log\frac{\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)}{\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y)}dp(x)\\ &= \limsup_n\Big(-\int_{\|y\|\leq M}\log\frac{\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)}{\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y)}dp(x) - \int_{\|y\|>M}\log\frac{\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)}{\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y)}dp(x)\Big)\\ &\leq 0 + \int_{\|x\|>M}\Big[2\log 2 + B_0 g(x) + \beta^* g(x)\Big]dp(x)\end{aligned}$$

Then by taking the limit of $M\to\infty$, we obtain the convergence. Here we used Assumption (5b) $\int g(x)dp(x) < \infty$.

To prove the convergence of the solutions, we let $(\tilde{\boldsymbol{r}},\tilde{\beta})$ be a accumulation point of the solution sequence $\{(\boldsymbol{r}^n,\beta^n)\}_{n=1}^{\infty}$. Then there is a subsequence $\{(\boldsymbol{r}^{n_k},\beta^{n_k})\}_{k=1}^{\infty}$ satisfying $\boldsymbol{r}^{n_k}\to\tilde{\boldsymbol{r}}$ and $\beta^{n_k}\to\tilde{\beta}$. Next, we have

$$F(\tilde{\boldsymbol{r}},\tilde{\beta}) = \lim_k F(\boldsymbol{r}^{n_k},\beta^{n_k}) = F(\boldsymbol{r}^*,\beta^*),$$

since we have proven $\lim_n F(\boldsymbol{r}^n, \beta^n) = F(\boldsymbol{r}^*, \beta^*)$. By the optimal property of $(\boldsymbol{r}^*, \beta^*), (\boldsymbol{r}^{n_k}, \beta^{n_k})$, we have

$$F(\tilde{\boldsymbol{r}}, \tilde{\beta}) = F(\boldsymbol{r}^*, \beta^*) \leq F(r^{n_k}, \beta^{*}) \leq F(\boldsymbol{r}^{n_k}, \beta^{n_k}) \leq F(\tilde{q}^{n_k}, \beta^{n_k}).$$

Here, $\tilde{q}^{n_k} \in W_{n_k}$ is a discrete version of $\bar{r} = \text{argmin}_{r \in W} F(r, \tilde{\beta})$. Similar to the value convergence analysis above, we have $F(\tilde{q}^{n_k}, \beta^{n_k}) \to F(\bar{r}, \tilde{\beta})$. Thus, we have

$$F(\tilde{\boldsymbol{r}}, \tilde{\beta}) \leq F(\bar{r}, \tilde{\beta}) \leq F(r, \tilde{\beta}), \ \forall r \in W.$$

This means $\tilde{\boldsymbol{r}} = \text{argmin}_{r \in W} F(r, \tilde{\beta})$ and

$$F(\tilde{\boldsymbol{r}}, \tilde{\beta}) = \min_{r \in W} F(r, \tilde{\beta}) \triangleq h(\tilde{\beta}).$$

However, $F(\tilde{\boldsymbol{r}}, \tilde{\beta}) = F(\boldsymbol{r}^*, \beta^*) = h(\beta^*)$, and $h(\beta^*) = \max_{\beta \geq 0} h(\beta)$, due to the optimal property of $(\boldsymbol{r}^*, \beta^*)$. So, we obtain $\tilde{\beta}$ is optimal, *i.e.*, $\tilde{\beta} = \text{argmax}_{\beta \geq 0} \ h(\beta)$. And combining $\tilde{\boldsymbol{r}} = \text{argmin}_{r \in W} F(r, \tilde{\beta})$, we obtain $(\tilde{\boldsymbol{r}}, \tilde{\beta})$ is optimal. ■

IV. Convergence Rate and Algorithm Complexity

First, we will establish some convergence rate results of problem (3). Before the analysis of the convergence rate, we need a lemma first.

***Lemma** 1:* $e^{-\lambda \rho(x,y)}$ is Lipschitz continuous, when $(x, y) \in [-M, M]^d \times [-M, M]^d$, *i.e.*, there exists a constant $L > 0$ satisfying:

$$|e^{-\lambda \rho(x_1,y_1)} - e^{-\lambda \rho(x_2,y_2)}| \leq L(\|x_1 - x_2\| + \|y_1 - y_2\|), \forall (x_1, y_1), (x_2, y_2) \in [-M, M]^d \times [-M, M]^d.$$

Proof: Since $e^{-\lambda \rho(x,y)}$ is continuously differentiable, its Jacobi matrix $J(x, y)$ is bounded in $[-M, M]^d \times [-M, M]^d$, *i.e.*, $\|J(x, y)\| \leq L$. Then by the Finite Increment Theorem, we have

$$|e^{-\lambda \rho(x_1,y_1)} - e^{-\lambda \rho(x_2,y_2)}| \leq \|J(x, y)\|(\|x_1 - x_2\| + \|y_1 - y_2\|),$$

for some $(x, y) \in [-M, M]^d \times [-M, M]^d$. Then by the bound on $\|J(x, y)\|$, we obtain the Lipchitz continuous property. ■

Next, we will prove the convergence rate of the discrete schemes (3).

***Theorem** 3:* When the distribution p of X is supported in $[-M, M]^d$, the optimal values $f(\boldsymbol{r}^n)$ of the discrete problem (3) satisfy the error estimate

$$|f(\boldsymbol{r}^n) - f^*| \leq Ch,$$

where C is a constant and $h = 2M/n^{\frac{1}{d}}$ is the discretization step size.

Proof: By Lemma 4.1 in [23], we have the distribution $\boldsymbol{r}$ of the reproduction variable is supported in $[-M, M]^d$ as long as $\rho(x, y)$ is a strictly increasing continuous difference distortion measure. Thus, we can just take the equidistant discretization nodes $\{y_j^n\}_{j=1}^n$ from $[-M, M]^d$. Due to the inequality (7) in Theorem 1, we have

$$0 \leq f(\boldsymbol{r}^n) - f(\boldsymbol{r}^*) \leq f(\boldsymbol{q}^n) - f(\boldsymbol{r}^*).$$

Thus, we only need to evaluate $f(\boldsymbol{q}^n) - f(\boldsymbol{r}^*)$.

$$\begin{aligned}
&\left|\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\right| \\
&= \left|\sum_i \int_{I_i} d\boldsymbol{r}^*(y)\exp(-\lambda\rho(x,y_i^n)) - \int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\right| \\
&= \left|\sum_i \int_{I_i} d\boldsymbol{r}^*(y)\exp(-\lambda\rho(x,y_i^n)) - \sum_i \int_{I_i} \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\right| \\
&\le \sum_i \int_{I_i} |\exp(-\lambda\rho(x,y)) - \exp(-\lambda\rho(x,y_i^n))| d\boldsymbol{r}^*(y) \\
&\le \sum_i \int_{I_i} L\|y - y_i^n\| d\boldsymbol{r}^*(y) \\
&\le \sum_i \int_{I_i} Lh/2 \; d\boldsymbol{r}^*(y) = Lh/2
\end{aligned}$$

The second equality is due to $\bigcup_{i=1}^n I_i \supseteq [-M,M]^d$ and we have used the result in Lemma 1.

Next, we have an evaluation on the lower bound of $\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y)$.

$$\begin{aligned}
\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y) &= \int_{\|y\|\le M} \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y) \\
&\ge e^{-\lambda\rho^*} \int_{\|y\|\le M} d\boldsymbol{q}^n(y) = e^{-\lambda\rho^*} \triangleq \delta_0, \forall x \in [-M,M]^d.
\end{aligned}$$

Here, $\rho^* = \max_{x,y\in[-M,M]^d} \rho(x,y)$. Similarly, we have

$$\int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y) \ge \delta_0.$$

Then,

$$\begin{aligned}
&\log\Big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y)\Big) - \log\Big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\Big) \\
&= \log\Big(\big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\big)\Big/\int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y) + 1\Big) \\
&\le \big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\big)\Big/\int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y) \\
&\le Lh/(2\delta_0)
\end{aligned}$$

Finally, we have an evaluation on the convergence rate

$$\begin{aligned}
f(\boldsymbol{q}^n) - f(\boldsymbol{r}^*) &= \int \log\Big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y)\Big)dp(x) - \int \log\Big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\Big)dp(x) \\
&= \int_{\|x\|\le M} \Big(\log\Big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{q}^n(y)\Big) - \log\Big(\int \exp(-\lambda\rho(x,y))d\boldsymbol{r}^*(y)\Big)\Big)dp(x) \\
&\le \int_{\|x\|\le M} Lh/(2\delta_0) \; dp(x) \\
&\le Lh/(2\delta_0) = O(h).
\end{aligned}$$

Thus,

$$f(\boldsymbol{r}^n) - f(\boldsymbol{r}^*) = O(h).$$

Since the total number of discrete nodes is n, there is $n^{\frac{1}{d}}$ nodes in each direction. So, the discretization distant is $\frac{2M}{n^{1/d}}$. Thus the convergence rate equals $O(1/n^{\frac{1}{d}})$ as well. ■

Next, based on the convergence rate analysis above, we can conduct a complexity analysis of solving the continuous RD problem (2) for achieving ε-accuracy via the BA algorithm.

***Theorem** 4:* To ensure ε-accuracy when computing the optimal value, the BA algorithm needs $O(\frac{m|\log \varepsilon|}{\varepsilon^{d+1}})$ arithmetic operations. Here, m is the number of discretizaion nodes of X when conducting numerical integration and d is the dimension of $\mathcal{Y}$.

Proof: By Theorem 3, to ensure $f(\boldsymbol{r}^n) - f(\boldsymbol{r}^*) \leq \varepsilon$, we need n to satisfy $1/n^{1/d} \sim \varepsilon$, *i.e.*, $n \sim 1/\varepsilon^d$. Then we use the BA algorithm to solve the associated discrete problem within ε tolerance. As shown in [16, Theorem 2, Equation (96)], the BA algorithm needs $O(\log n/\varepsilon)$ iterations to achieve ε-accuracy. In each iteration, the BA algorithm iterates between two variables $w(y_i^n|x)$ and $r(y_i^n)$ in the following way:

$$w(y_i^n|x) = \left(r(y_i^n)e^{-\beta\rho(x,y_i^n)}\right) \Big/ \sum_{i=1}^{n} e^{-\beta\rho(x,y_i^n)} r(y_i^n),$$

$$r(y_i^n) = \int p(x) w(y_i^n|x) dx.$$

Let $x_1, x_2 \cdots x_m$ be the nodes of numerical integration with respect to x. When computing $w(y_i^n|x)$, we need to perform matrix-vector multiplication

$$\sum_{i=1}^{n} e^{-\beta\rho(x_j,y_i^n)} r(y_i^n), \ j = 1, 2 \cdots m,$$

and it involves $O(mn)$ arithmetic operations. The total computation cost for $w(y_i^n|x)$ is $O(mn)$. When computing $r(y_i^n)$, we need to perform numerical integration,

$$\int p(x) w(y_i^n|x) dx \sim \sum_{j=1}^{m} A_j p(x_j) w(y_i^n|x_j), \ i = 1, 2 \cdots n,$$

where A_j are the numerical integration coefficients. This is a matrix-vector multiplication and it involves $O(mn)$ arithmetic operations. Thus, the computation cost for each iteration in the BA algorithm is $O(mn)$. Hence, the total computation cost is $O(mn \log n/\varepsilon) = O(m|\log \varepsilon|/\varepsilon^{d+1})$. ■

Correspondingly, the convergence rate of problem (4) can be obtained.

***Theorem** 5:* When the distribution p of X is supported in $[-M, M]^d$, the optimal values $F(\boldsymbol{r}^n, \beta^n)$ of discrete problem (4) satisfy the error estimate

$$|F(\boldsymbol{r}^n, \beta^n) - F^*| \leq Ch,$$

where C is a constant and $h = 2M/n^{\frac{1}{d}}$ is the discretization step size.

Before the analysis of the convergence rate, we need a lemma first.

***Lemma** 2:* $e^{-\beta\rho(x,y)}$, $e^{-\beta\rho(x,y)}\rho(x, y)$ is uniformly Lipschitz continuous for $0 < B_1 \leq \beta \leq B_0$, when $(x, y) \in [-M, M]^d \times [-M, M]^d$, *i.e.*, there exists a constant $L > 0$ satisfying:

$$|e^{-\beta\rho(x_1,y_1)} - e^{-\beta\rho(x_2,y_2)}| \leq L(\|x_1 - x_2\| + \|y_1 - y_2\|),$$

$$|e^{-\beta\rho(x_1,y_1)}\rho(x_1, y_1) - e^{-\beta\rho(x_2,y_2)}\rho(x_2, y_2)| \leq L(\|x_1 - x_2\| + \|y_1 - y_2\|),$$

$$\forall (x_1, y_1), (x_2, y_2) \in [-M, M]^d \times [-M, M]^d, \quad \forall 0 < B_1 \leq \beta \leq B_0,$$

Proof: Since $e^{-\beta\rho(x,y)}$ is continuously differentiable, its Jacobi matrix $J_\beta(x,y)$ with respect to (x,y) is continuous in $[-M,M]^d\times[-M,M]^d$. And taking into account β, $J_\beta(x,y)$ is continuous in $[-M,M]^d\times[-M,M]^d\times[B_1,B_0]$. Thus it is bounded, *i.e.*, $\|J_\beta(x,y)\|\le L$. Then by the Finite Increment Theorem, we have

$$|e^{-\beta\rho(x_1,y_1)}-e^{-\beta\rho(x_2,y_2)}|\le\|J_\beta(x,y)\|(\|x_1-x_2\|+\|y_1-y_2\|),$$

for some $(x,y)\in[-M,M]^d\times[-M,M]^d$. Then by the bound on $\|J_\beta(x,y)\|$, we obtain the Lipschitz continuous property. Similarly, we can prove the Lipschitz continuous property of $e^{-\beta\rho(x,y)}\rho(x,y)$. ■

Next, we will give the proof of Theorem 5.

Proof: Since r is supported in $[-M,M]^d$, we can just take the equidistant discretization nodes $\{y_j^n\}_{j=1}^n$ from $[-M,M]^d$. Due to the inequality (11) in Theorem 4, we have

$$0\le f(\boldsymbol{r}^n,\beta^n)-f(\boldsymbol{r}^*,\beta^*)\le f(\boldsymbol{q}^n,\tilde\beta^n)-f(\boldsymbol{r}^*,\beta^*).$$

Thus, we only need to evaluate $f(\boldsymbol{q}^n,\tilde\beta^n)-f(\boldsymbol{r}^*,\beta^*)$.

First, we estimate the convergence rate of $\tilde\beta^n\to\beta^*$.

$$\begin{aligned}
&\left|\int\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{q}^n(y)-\int\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{r}^*(y)\right|\\
&=\left|\sum_i\int_{I_i}d\boldsymbol{r}^*(y)\exp(-\tilde\beta^n\rho(x,y_i^n))-\int\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{r}^*(y)\right|\\
&=\left|\sum_i\int_{I_i}d\boldsymbol{r}^*(y)\exp(-\tilde\beta^n\rho(x,y_i^n))-\sum_i\int_{I_i}\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{r}^*(y)\right|\\
&\le\sum_i\int_{I_i}|\exp(-\tilde\beta^n\rho(x,y))-\exp(-\tilde\beta^n\rho(x,y_i^n))|d\boldsymbol{r}^*(y)\\
&\le\sum_i\int_{I_i}L\|y-y_i^n\|d\boldsymbol{r}^*(y)\\
&\le\sum_i\int_{I_i}Lh/2\ d\boldsymbol{r}^*(y)=Lh/2
\end{aligned}$$

The second equality is due to $\bigcup_{i=1}^n I_i\supseteq[-M,M]^d$. And the second inequality uses the bound on $\tilde\beta^n$, *i.e.*, $0<B_1\le\tilde\beta^n\le B_0$. Similarly, we have

$$\left|\int\exp(-\tilde\beta^n\rho(x,y))\rho(x,y)d\boldsymbol{q}^n(y)-\int\exp(-\tilde\beta^n\rho(x,y))\rho(x,y)d\boldsymbol{r}^*(y)\right|\le Lh/2.$$

Next, we have an evaluation on the lower bound of $\int\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{q}^n(y)$.

$$\begin{aligned}
\int\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{q}^n(y)&=\int_{\|y\|\le M}\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{q}^n(y)\\
&\ge e^{-\tilde\beta^n\rho^*}\int_{\|y\|\le M}d\boldsymbol{q}^n(y)=e^{-\tilde\beta^n\rho^*}\triangleq\delta_0,\forall x\in[-M,M]^d.
\end{aligned}\tag{15}$$

Here, $\rho^*=\max_{x,y\in[-M,M]^d}\rho(x,y)$. Similarly, we have

$$\int\exp(-\tilde\beta^n\rho(x,y))d\boldsymbol{r}^*(y)\ge\delta_0.$$

And we have an upper bound estimation on $\int \exp(-\tilde{\beta}^n\rho(x,y))\rho(x,y)d\boldsymbol{r}^*(y)$.

$$\begin{aligned}\int \exp(-\tilde{\beta}^n\rho(x,y))\rho(x,y)d\boldsymbol{r}^*(y) &\leq \int \exp(-B_1\rho(x,y))\rho(x,y)d\boldsymbol{r}^*(y)\\ &\leq \int M_1 d\boldsymbol{r}^*(y) = M_1.\end{aligned}$$

Here, M_1 is the upper bound on the function $\exp(-B_1 t)t, t\geq 0$. Now, since $G_{\boldsymbol{r}^*}(\beta^*) = 0 = G_{\boldsymbol{q}^n}(\tilde{\beta}^n)$, we have the following estimation

$$\begin{aligned}&|G_{\boldsymbol{r}^*}(\tilde{\beta}^n) - G_{\boldsymbol{r}^*}(\beta^*)| = |G_{\boldsymbol{r}^*}(\tilde{\beta}^n) - G_{\boldsymbol{q}^n}(\tilde{\beta}^n)|\\ &\leq \int_{\|x\|\leq M}\left|\frac{\left(\int \exp(-\tilde{\beta}^n\rho(x,y))\rho(x,y)d\boldsymbol{q}^n(y)\right)}{\left(\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)\right)} - \frac{\left(\int \exp(-\tilde{\beta}^n\rho(x,y))\rho(x,y)d\boldsymbol{r}^*(y)\right)}{\left(\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{r}^*(y)\right)}\right|dp(x)\\ &= \int_{\|x\|\leq M}\left|\frac{b_1}{a_1} - \frac{b_2}{a_2}\right|dp(x) = \int_{\|x\|\leq M}\frac{|a_2 b_1 - a_1 b_2|}{a_1 a_2}dp(x)\\ &\leq \int_{\|x\|\leq M}\frac{a_2|b_1-b_2| + b_2|a_2-a_1|}{a_1 a_2}dp(x) \leq \int_{\|x\|\leq M} Lh/(2\delta_0) + \frac{M_1 Lh/2}{\delta_0^2}dp(x)\\ &\leq Lh/(2\delta_0) + \frac{M_1 Lh/2}{\delta_0^2} = O(h).\end{aligned}$$

Here, for simplicity, we use a_1, a_2, b_1, b_2 to represent the corresponding integration.

Next, we will give the estimate of $\tilde{\beta}^n - \beta^*$. Let $-L_1 = G'_{\boldsymbol{r}^*}(\beta^*) < 0$, then in a neighborhood $(-\delta_2+\beta^*, \delta_2+\beta^*)$ of β^*, we have $G'_{\boldsymbol{r}^*}(\beta) \leq -L_1/2$ and $\tilde{\beta}^n$ is in the neighborhood when n is sufficiently large. Therefore, by the Lagrange Mean Value Theorem, we have

$$|G_{\boldsymbol{r}^*}(\tilde{\beta}^n) - G_{\boldsymbol{r}^*}(\beta^*)| = |G'_{\boldsymbol{r}^*}(\zeta)|\ |\tilde{\beta}^n - \beta^*| \geq \frac{L_1}{2}|\tilde{\beta}^n - \beta^*|.$$

Here, ζ is a real number between $\tilde{\beta}^n$ and β^*. Thus, we obtain

$$|\tilde{\beta}^n - \beta^*| \leq \frac{2}{L_1}|G_{\boldsymbol{r}^*}(\tilde{\beta}^n) - G_{\boldsymbol{r}^*}(\beta^*)| = O(h).$$

Next, we will give the convergence rate of $f(\boldsymbol{q}^n, \tilde{\beta}^n) - f(\boldsymbol{r}^*, \beta^*)$.

$$|f(\boldsymbol{q}^n, \tilde{\beta}^n) - f(\boldsymbol{r}^*, \beta^*)| \leq |f(\boldsymbol{q}^n, \tilde{\beta}^n) - f(\boldsymbol{q}^n, \beta^*)| + |f(\boldsymbol{q}^n, \beta^*) - f(\boldsymbol{r}^*, \beta^*)|.$$

For the first part, we have the following estimation

$$\begin{aligned}0 \leq f(\boldsymbol{q}^n, \tilde{\beta}^n) - f(\boldsymbol{q}^n, \beta^*) &= -\int_{\|x\|\leq M}\log\frac{\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)}{\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y)}dp(x) - (\tilde{\beta}^n - \beta^*)D\\ &\leq \int_{\|x\|\leq M}\log\left(\frac{\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)}{\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)} + 1\right)dp(x) + O(h)\\ &\leq \int_{\|x\|\leq M}\frac{|\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)|}{\int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)}dp(x) + O(h)\\ &\leq \int_{\|x\|\leq M}\frac{|\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)|}{\delta_0}dp(x) + O(h)\end{aligned}$$

Here, δ_0 is a lower bound as shown in (15).

$$
\begin{aligned}
&|\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)| \\
&\leq \int e^{-\beta^*\rho(x,y)}|e^{(\beta^*-\tilde{\beta}^n)\rho(x,y)} - 1|d\boldsymbol{q}^n(y) \\
&\leq \int_{\|y\|\leq M} |e^{(\beta^*-\tilde{\beta}^n)\rho(x,y)} - 1| \ d\boldsymbol{q}^n(y) \\
&\leq \int_{\|y\|\leq M} (e^{|\beta^*-\tilde{\beta}^n|\rho(x,y)} - 1) \ d\boldsymbol{q}^n(y) \\
&\leq \int_{\|y\|\leq M} (e^{|\beta^*-\tilde{\beta}^n|\rho^*} - 1) \ d\boldsymbol{q}^n(y) \\
&\leq (e^{|\beta^*-\tilde{\beta}^n|\rho^*} - 1) \leq 2|\beta^* - \tilde{\beta}^n|\rho^* = O(h), \ \text{when n is sufficiently large}
\end{aligned}
$$

Here, $\rho^* = \max_{x,y\in[-M,M]^d} \rho(x,y)$. Thus,

$$
\begin{aligned}
|f(\boldsymbol{q}^n,\tilde{\beta}^n) - f(\boldsymbol{q}^n,\beta^*)| &\leq \int_{\|x\|\leq M} \frac{|\int \exp(-\beta^*\rho(x,y))d\boldsymbol{q}^n(y) - \int \exp(-\tilde{\beta}^n\rho(x,y))d\boldsymbol{q}^n(y)|}{\delta_0} dp(x) + O(h) \\
&\leq \int_{\|x\|\leq M} O(h)/\delta_0 \ dp(x) + O(h) \leq O(h)/\delta_0 + O(h) = O(h).
\end{aligned}
$$

Meanwhile, due to Theorem 2, $|f(\boldsymbol{q}^n,\beta^*) - f(\boldsymbol{r}^*,\beta^*)| = O(h)$. Thus, we obtain the convergence rate $|f(\boldsymbol{r}^n,\beta^n) - f(\boldsymbol{r}^*,\beta^*)| = O(h) = O(1/n^{\frac{1}{d}})$.

■

Using the recently proposed CBA algorithm [15], we can solve the original RD problem (1) directly. The next theorem gives the complexity result of solving the continuous RD problem (1) for achieving ε-accuracy via the CBA algorithm.

***Theorem** 6:* To ensure ε-accuracy when computing the optimal value, the CBA algorithm needs $O\big(\frac{m|\log\varepsilon|}{\varepsilon^{d+1}}(1+\log|\log\varepsilon|)\big)$ arithmetic operations [1]. Here, m is the number of discretizaion nodes of X when conducting numerical integration and d is the dimension of $\mathcal{Y}$.

Proof: By Theorem 5, to ensure $F(\boldsymbol{r}^n,\beta^n) - F(\boldsymbol{r}^*,\beta^*) \leq \varepsilon$, we need n to satisfy $1/n^{1/d} \sim \varepsilon$, *i.e.*, $n \sim 1/\varepsilon^d$. Then we use the CBA algorithm to solve the discrete problem within ε tolerance. As shown in a recent work [15], we need $O(\frac{mn\log n}{\varepsilon}(1+\log|\log\varepsilon|))$ arithmetic operations to achieve ε-accuracy. Thus, the complexity is

$$
O\big(\frac{mn\log n}{\varepsilon}(1+\log|\log\varepsilon|)\big) = O\big(\frac{m|\log\varepsilon|}{\varepsilon^{d+1}}(1+\log|\log\varepsilon|)\big).
$$

■

V. Numerical Experiments

In this section, we conduct experiments on uniform source to confirm the convergence. We consider the uniform source on interval $[-8,8]$ and conduct experiments with different discretization parameters, namely the node number $n = 20, 40, 80, 160$ of Y, while we take the node number $m = 300$ of X to ensure computing the integral of x with

[1] The CBA algorithm takes an additional cost $\log|\log\varepsilon|$ to update the multiplier β, while the BA algorithm fixes β. Thus, the CBA algorithm is applicable for solving the original problem (1) directly.

high accuracy. Moreover, we use the BA algorithm and the CBA algorithm to solve the discrete RD problem (3) and (4) respectively with high accuracy. The corresponding results of the reproduction distribution are illustrated in Figure 1 and Figure 2.

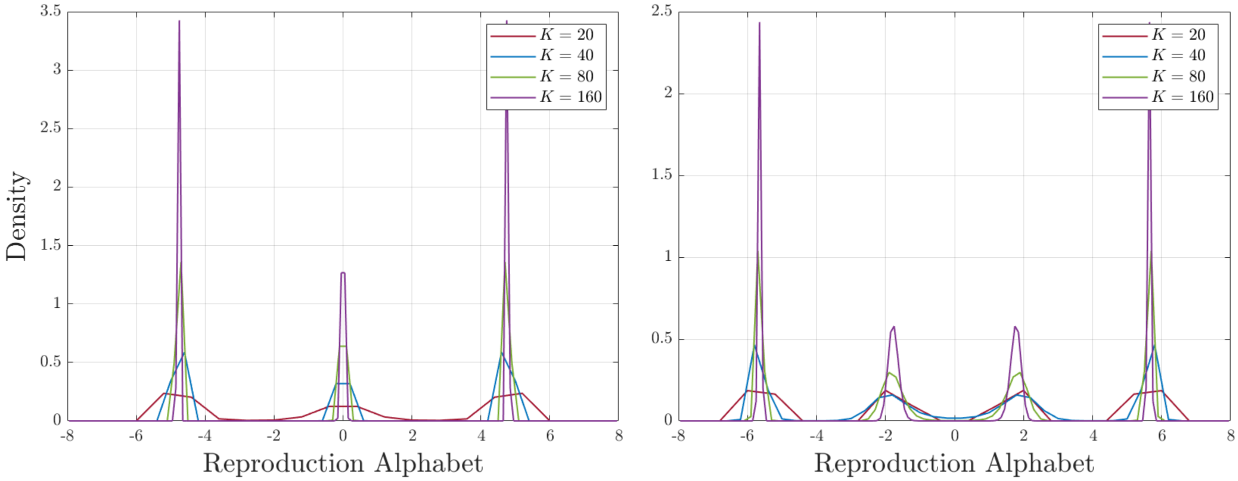

Fig. 1. The discrete optimal reproduction produced by the BA algorithm for the slope $\beta = 0.1$ (left) and $\beta = 0.2$ (right).

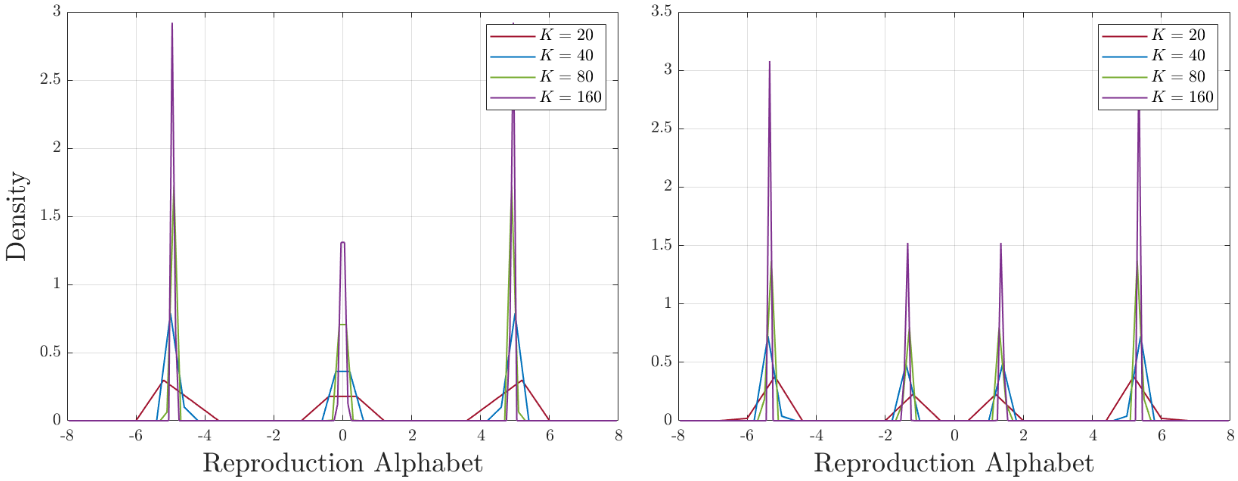

Fig. 2. The discrete optimal reproduction produced by the CBA algorithm for the distortion $D = 4$ (left) and $D = 3$ (right).

As shown in [24], the optimal reproduction of the uniform source is a discrete distribution. From the figures, the convergence is clearly demonstrated and the solutions of discrete problems converge to a discrete distribution as the grids become finer.

VI. CONCLUSION

In this paper, we prove the convergence of discrete schemes for computing the continuous RD problem and establish convergence rate results and complexity estimations. Numerical experiments confirm the convergence. Considering the fundamental role of the RD problem, it is envisioned that our method may lead to a series of applications to various information problems.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[2] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, July 1948.

[3] C. E. Shannon *et al.*, "Coding Theorems for a Discrete Source with a Fidelity Criterion," *Institute of Radio Engineers International Convention Record*, vol. 4, no. 142-163, p. 1, Mar. 1959.

[4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

[5] Y. Blau and T. Michaeli, "Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff," in *36th International Conference on Machine Learning (ICML)*, Long Beach, California, USA, Jun. 2019, pp. 675–685.

[6] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," *arXiv preprint physics/0004057*, 2000.

[7] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 Still Image Compression Standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, Sep. 2001.

[8] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[9] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[10] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep Variational Information Bottleneck," in *Proc. 5th International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017, pp. 1–5.

[11] R. Blahut, "Computation of Channel Capacity and Rate-Distortion Functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, July 1972.

[12] S. Arimoto, "An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.

[13] S. Wu, W. Ye, H. Wu, H. Wu, W. Zhang, and B. Bai, "A Communication Optimal Transport Approach to the Computation of Rate Distortion Functions," in Proc. *2023 IEEE Information Theory Workshop (ITW)*, Saint-Malo, France, Apr. 2023.

[14] Y. Yang, S. Eckstein, M. Nutz, and S. Mandt, "Estimating the rate-distortion function by Wasserstein gradient descent," *arXiv preprint arXiv:2310.18908*, 2023.

[15] L. Chen, S. Wu, W. Ye, H. Wu, W. Zhang, H. Wu, and B. Bai, "A constrained BA algorithm for rate-distortion and distortion-rate functions," *arXiv:2305.02650 [cs.IT]*, 2023, version 2. [Online]. Available: https://arxiv.org/abs/2305.02650

[16] M. Hayashi, "Bregman Divergence Based Em Algorithm and its Application to Classical and Quantum Rate Distortion Theory," *IEEE Transactions on Information Theory*, vol. 69, no. 6, pp. 3460–3492, Jan. 2023.

[17] J. W. Thomas, *Numerical Partial Differential Equations: Conservation Laws and Elliptic Equations*. Springer Science & Business Media, 2013, vol. 33.

[18] F. Santambrogio, "Optimal transport for applied mathematicians," *Birkäuser, NY*, vol. 55, no. 58-63, p. 94, 2015.

[19] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*. Springer, 2007.

[20] W. Rudin, *Principles of Mathematical Analysis*. McGraw-hill New York, 1976, vol. 3.

[21] I. E. Schochetman and R. L. Smith, "A finite algorithm for solving infinite dimensional optimization problems," *Annals of Operations Research*, vol. 101, pp. 119–142, 2001.

[22] S. L. Fix, *Rate Distortion Functions for Continuous Alphabet Memoryless Sources*. University of Michigan, sep 1977.

[23] S. L. Fix, "Rate distortion functions for squared error distortion measures," in *Annual Allerton Conference on Communication, Control and Computing, 16 th, Monticello, Ill*, 1978, pp. 704–711.

[24] K. Rose, "A Mapping Approach to Rate-Distortion Computation and Analysis," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1939–1952, Nov. 1994.

[25] M. Z. Mao, R. M. Gray, and T. Linder, "Rate-constrained simulation and source coding iid sources," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4516–4529, 2011.

[26] S. C. Brenner, *The Mathematical Theory of Finite Element Methods*. Springer, 2008.

[27] E. Lei, H. Hassani, and S. S. Bidokhti, "Neural estimation of the rate-distortion function with applications to operational source coding," *IEEE Journal on Selected Areas in Information Theory*, 2023.

[28] J. Munkres, *Topology*, 2nd ed. Prentice Hall, 1999.