Graph-based Composite Local Bregman Divergences on Discrete Sample Spaces

Takafumi Kanamori[1] and Takashi Takenouchi[2]

[1]Nagoya University
[2]Future University Hakodate

Abstract

One of the most common methods for statistical inference is the maximum likelihood estimator (MLE). The MLE needs to compute the normalization constant in statistical models, and it is often intractable. Using unnormalized statistical models and replacing the likelihood with the other scoring rule are a good way to circumvent such high computation cost, where the scoring rule measures the goodness of fit of the model to observed samples. The scoring rule is closely related to the Bregman divergence, which is a discrepancy measure between two probability distributions. In this paper, the purpose is to provide a general framework of statistical inference using unnormalized statistical models on discrete sample spaces. A localized version of scoring rules is important to obtain computationally efficient estimators. We show that the local scoring rules are related to the localized version of Bregman divergences. Through the localized Bregman divergence, we investigate the statistical consistency of local scoring rules. We show that the consistency is determined by the structure of neighborhood system defined on discrete sample spaces. In addition, we show a way of applying local scoring rules to classification problems. In numerical experiments, we investigated the relation between the neighborhood system and the estimation accuracy.

1 Introduction

Our purpose is to provide a general framework of statistical inference with unnormalized statistical models on discrete sample spaces. For statistical inference, one of the most common methods is the maximum likelihood estimator (MLE), which is obtained by maximizing the empirical mean of the log-likelihood for the statistical model. The MLE has some nice properties such as the statistical consistency and efficiency. When the dimension of the sample domain is large, however, the computation of the normalization constant in the statistical model is intractable.

Several approaches have been proposed to deal with the normalization constant. One way is to approximate the normalization constant by means of the Monte Carlo method, which is a generic framework to compute integrals and total sums by using random sampling [13, 17, 30]. The other approach is to replace the log-likelihood with other scoring rules that measure the goodness of fit of the model to observed samples. Using scoring rules that do not depend on the normalization constant is thought to be computationally efficient. Such scoring rules include pseudo-likelihood, composite likelihood, ratio matching, and so forth [2, 5, 15, 18, 19, 21, 23, 24, 25, 29]. As a whole, the locality of the scoring rule over the sample space is the key to reduce

the computational cost. Dawid et al. argued the theoretical properties of scoring rules that can be expressed by the sum of localized scoring rules [11].

In this paper, we study the statistical consistency of local scoring rules. In general, the scoring rule is closely related to the Bregman divergence $D(p, q)$, which is a discrepancy measure between two probability distributions, p, q [14, 16, 26]. The Bregman divergence takes non-negative real numbers and $D(p, p) = 0$ holds for any probability distribution. The coincidence axiom means that $D(p, q) = 0$ leads to $p = q$ [8, 27]. When the Bregman divergence satisfies the coincidence axiom, the associated scoring rule will have the statistical consistency under a mild assumption. We show that the local scoring rules are also related to the localized version of Bregman divergences. Through the localized Bregman divergence, we investigate the statistical consistency of local scoring rules. We show that the consistency is determined by the structure of neighborhood system defined on discrete sample spaces.

The rest of the paper is organized as follows. In Section 2, we introduce basic concepts such as scoring rules, Bregman divergences, and show some examples. Then, we show the importance of the locality for scoring rules in order to reduce the computational cost. Section 3 is devoted to define composite local Bregman divergences which is the key concept in our paper. In Section 4, We study how the neighborhood system on sample space relates to theoretical properties of composite local Bregman divergences. On the basis of the results in Section 4, we propose some new scoring rules such as an extension of composite likelihood and a localized version of the pseudo-spherical scoring rule in Section 5. Then, we investigate the relation between the neighborhood system and the statistical consistency of existing scoring rules and some newly proposed ones. In Section 6, we use local scoring rules to classification problems. Numerical experiments are presented in Section 7. In the experiments, mainly we focus on investigating the relation among the neighborhood system and the estimation accuracy. Finally, Section 8 concludes the paper with discussions.

2   Preliminaries

Let us introduce a scoring rule that is a basic concept in statistical inference. The scoring rule measures a loss suffered for inaccurate prediction. We show that scoring rules are related to convex functions and Bregman divergences. For computationally efficient statistical inference, homogeneous scoring rules have attracted research attention recently [11, 28, 19]. The following subsections are devoted to define some concepts to describe scoring rules and homogeneous scoring rules over discrete sample spaces.

Let us summarize the notations to be used throughout the paper. Let $\mathbb{R}$ be the set of all real numbers. The non-negative numbers and positive numbers are denoted as $\mathbb{R}_+ = \{x \in \mathbb{R} \,|\, x \geq 0\}$ and $\mathbb{R}_{++} = \{x \in \mathbb{R} \,|\, x > 0\}$, respectively. A discrete sample space is denoted as $\mathcal{Y}$, and the set of functions from $\mathcal{Y}$ to $\mathbb{R}$ is denoted as $\mathbb{R}^{\mathcal{Y}}$. Likewise, the notations, $\mathbb{R}^{\mathcal{Y}}_+$, $\mathbb{R}^{\mathcal{Y}}_{++}$, are used. An element of $\mathbb{R}^{\mathcal{Y}}$ is expressed as $a = (a_y)_{y \in \mathcal{Y}}$ like a numerical vector having the index set $\mathcal{Y}$. For a subset $A \subset \mathcal{Y}$, the sub-vector $a_A \in \mathbb{R}^A$ of $a \in \mathbb{R}^{\mathcal{Y}}$ denotes $a_A = (a_y)_{y \in A}$. The derivative of a function $\phi : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}$ with respect to the variable $f_y \, (y \in \mathcal{Y})$ of $\phi(f)$ is expressed as $\partial_y \phi(f)$ instead of $\frac{\partial \phi}{\partial f_y}(f)$. The indicator function is denoted as $\mathbf{1}[A]$ that takes 1 if A is true and 0 otherwise.

2.1 Scoring Rules and Bregman Divergences

A probability function p on the discrete sample space $\mathcal{Y}$ corresponds to an element in $\mathbb{R}_+^{\mathcal{Y}}$ such that $\sum_{y \in \mathcal{Y}} p_y = 1$. In this paper, we consider the probability $p = (p_y)_{y \in \mathcal{Y}}$ such that all p_y's are positive in order to avoid the difficulty concerning the boundary effect. The set of all non-degenerate probability functions is denoted as $\mathcal{P} \subset \mathbb{R}_{++}^{\mathcal{Y}}$, i.e., $\mathcal{P} = \{ p \in \mathbb{R}_{++}^{\mathcal{Y}} \mid \sum_{y \in \mathcal{Y}} p_y = 1 \}$. Let $\mathcal{F}$ be $\mathcal{F} = \mathbb{R}_{++}^{\mathcal{Y}}$ for simplicity.

The scoring rule $S(y, q) \in \mathbb{R}$, or the score for short, is a loss of the prediction by using the probability function $q \in \mathcal{P}$ for a given sample $y \in \mathcal{Y}$. Suppose that samples obeys the probability function $p \in \mathcal{P}$. Then, the expected score of $S(y, q)$ is denotes as

$$S(p, q) = \sum_{y \in \mathcal{Y}} p_y S(y, q)$$

with some abuse of notation.

Definition 1 (Proper score [14, 16]). *The score $S : \mathcal{Y} \times \mathcal{P} \to \mathbb{R}$ is called proper when*

$$S(p, q) \geq S(p, p) \tag{1}$$

holds for all $p, q \in \mathcal{P}$. It is strictly proper when the equality $S(p, q) = S(p, p)$ for $p, q \in \mathcal{P}$ means $p = q$.

Strictly proper scores are used to estimate the probability function of observed samples. Suppose that i.i.d. samples $y_1, \ldots, y_n \in \mathcal{Y}$ are generated from p in $\mathcal{P}$. The law of large numbers guarantees that the expected score $S(p, q)$ is approximated by the empirical mean of the score over the observed samples. The minimizer of the empirical score over a statistical model $\mathcal{Q} \subset \mathcal{P}$, i.e., the optimal solution of

$$\min_{q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} S(y_i, q)$$

is expected to provide a good estimator of p if $\mathcal{Q}$ includes p.

Here, let us define the Bregman divergence. We show that scores are closely related to Bregman divergences.

Definition 2 (Bregman divergence [8, 14]). *Let $\phi : \mathcal{F} \to \mathbb{R}$ be a convex function. The Bregman divergence $D_\phi : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is defined as*

$$D_\phi(f, g) = \phi(f) - \phi(g) - \sum_{y \in \mathcal{Y}} \partial_y \phi(g)(f_y - g_y),$$

for $f, g \in \mathcal{F}$. If ϕ is not differentiable, $(\partial_y \phi(g))_{y \in \mathcal{Y}}$ denotes a subgradient of ϕ at $g \in \mathcal{F}$. The function ϕ is called the potential of the Bregman divergence D_ϕ.

The convexity of ϕ guarantees the non-negativity of D_ϕ [8]. When ϕ is strictly convex, the equality $D_\phi(f, g) = 0$ leads to $f = g$. Thus, the Bregman divergence is regarded as a discrepancy measure on $\mathcal{F}$. The Bregman divergence has been used in wide range of problems in statistics and machine leaning [3, 9, 27]. In this paper, we focus on the differentiable potential in order to avoid technical difficulties. In practical problems, usually we use Bregman divergences defined from differentiable potentials.

A remarkable feature of proper scores is shown in the following theorem [16, 26]. A refined version of the theorem is presented by [14].

Theorem 1. *The score S is proper if and only if there exists a convex function $\phi : \mathcal{F} \to \mathbb{R}$ such that*

$$S(y, q) = -\phi(q) - \partial_y \phi(q) + \sum_{z \in \mathcal{Y}} q_z \partial_z \phi(q) \tag{2}$$

holds for $y \in \mathcal{Y}$ and $q \in \mathcal{P}$. When ϕ is non differentiable, $\partial_y \phi(q)$ denotes a subgradient of ϕ at q. Moreover, S is strictly proper if and only if ϕ is strictly convex on $\mathcal{P}$.

The potential $\phi(p)$ is expressed as $-S(p, p)$ on $\mathcal{P}$. The proper score of the form (2) can be defined on $\mathcal{F}$, and we have

$$S(f, g) - S(f, f) = D_\phi(f, g), \quad f, g \in \mathcal{F}.$$

Thus, the minimization of the strictly proper score is interpreted as the minimization of the corresponding Bregman divergence from the observed distribution to the statistical model.

Example 1 (Brier score)**.** *The score defined as*

$$S(y, q) = -2q_y + \sum_{z \in \mathcal{Y}} q_z^2, \quad q \in \mathcal{P}$$

is called Brier score. The divergence $D_\phi(p, q) = S(p, q) - S(p, p)$ is equal to the squared Euclidean distance between p and q, where the potential is given as $\phi(f) = \sum_{y \in \mathcal{Y}} f_y^2$ for $f \in \mathcal{F}$. Since ϕ is strictly convex, the Brier score is strictly proper.

Example 2 (Logarithmic score and Kullback-Leibler divergence)**.** *For the logarithmic score*

$$S(y, q) = -\log q_y,$$

we have

$$S(p, q) - S(p, p) = \sum_{y \in \mathcal{Y}} p_y \log \frac{p_y}{q_y},$$

that is nothing but the Kullback-Leibler divergence between $p, q \in \mathcal{P}$ [10]. The potential is the negative Shannon entropy $\phi(p) = \sum_{y \in \mathcal{Y}} p_y \log p_y$. The domain of ϕ can be directly extended to $\mathcal{F}$ by defining $\phi(f) = \sum_{y \in \mathcal{Y}} f_y \log f_y$ for $f \in \mathcal{F}$, which is strictly convex on $\mathcal{F}$. Hence, the logarithmic score is strictly proper. The estimator defined from the logarithmic score is the maximum likelihood estimator.

Example 3 (Density-power score)**.** *For a positive constant $\gamma > 0$, the density-power score is defined as*

$$S(y, q) = -\frac{1 + \gamma}{\gamma} q_y^\gamma + \sum_{z \in \mathcal{Y}} q_z^{1+\gamma}$$

for $q \in \mathcal{P}$. The Brier score is obtained by setting $\gamma = 1$. The density-power score is expressed as the form of (2) by using the strictly convex potential $\phi(f) = \sum_{y \in \mathcal{Y}} f_y^{1+\gamma} / \gamma$. Hence, the density-power score is strictly proper. The corresponding Bregman divergence is

$$D_\phi(f, g) = \sum_{y \in \mathcal{Y}} \left(\frac{1}{\gamma} f_y^{1+\gamma} - \frac{1 + \gamma}{\gamma} f_y g_y^\gamma + g_y^{1+\gamma} \right), \quad f, g \in \mathcal{F}.$$

Since the potential is strictly convex on $\mathcal{F}$, $D_\phi(f, g) = 0$ leads to $f = g$ on $\mathcal{F}$. The density-power score is used for robust estimation [4].

Example 4 (Pseudo-spherical score). *For a positive constant $\gamma > 0$, the pseudo-spherical score is defined as*

$$S(y, q) = -\frac{q_y^\gamma}{(\sum_{z \in \mathcal{Y}} q_z^{1+\gamma})^{\gamma/(1+\gamma)}}$$

for $q \in \mathcal{P}$. The corresponding potential is $\phi(f) = (\sum_{y \in \mathcal{Y}} f_y^{1+\gamma})^{1/(1+\gamma)}$ for $f \in \mathcal{F}$, i.e., $(1+\gamma)$-norm of the vector f. Though the potential is not strictly convex on $\mathcal{F}$, it is strictly convex on $\mathcal{P}$. Hence, the density-power score is strictly proper. The Bregman divergence for $f, g \in \mathcal{F}$ is

$$D_\phi(f, g) = \left(\sum_{y \in \mathcal{Y}} f_y^{1+\gamma}\right)^{1/(1+\gamma)} - \frac{\sum_{y \in \mathcal{Y}} f_y g_y^\gamma}{(\sum_{y \in \mathcal{Y}} g_y^{1+\gamma})^{\gamma/(1+\gamma)}}.$$

The inequality $D_\phi(f, g) \geq 0$ is equivalent with the Hölder's inequality. For $f, g \in \mathcal{F}$, the equality $D_\phi(f, g) = 0$ means that f and g are linearly dependent. When p and q in $\mathcal{P}$ are linearly dependent, they should be the same. The pseudo-spherical score is used for robust estimation [12, 20].

2.2 Homogeneous Scoring Rules and Locality

On a large sample space $\mathcal{Y}$ such as the high dimensional binary variables $\mathcal{Y} = \{+1, -1\}^D$, finding the normalization constant of statistical models is often computationally intractable. Suppose that the statistical model $q_\theta = (q_{\theta,y})_{y \in \mathcal{Y}} \in \mathcal{P}$ defined as

$$q_{\theta,y} = \frac{f_{\theta,y}}{Z_\theta}, \quad Z_\theta = \sum_{z \in \mathcal{Y}} f_{\theta,z}, \tag{3}$$

is used to estimate the probability of observed samples, where $f_\theta = (f_{\theta,y})_{y \in \mathcal{Y}} \in \mathcal{F}$ is an unnormalized model having the parameter θ. The logarithmic score needs to compute $\log Z_\theta$, and the Brier score requires the normalization constant Z_θ and $\sum_{y \in \mathcal{Y}} f_y^2$, though the summing over $\mathcal{Y}$ is computationally prohibitive.

In such a case, proper local homogeneous scores are useful to greatly reduce the computation cost. To begin with, let us define a homogeneous score as the score $S(y, f)$ defined for $y \in \mathcal{Y}$ and $f \in \mathcal{F}$ such that $S(y, \lambda f) = S(y, f)$ holds for all $\lambda > 0$. The score $S(y, q)$ for $q \in \mathcal{P}$ can be extended to the homogeneous score by

$$S(y, f) = S\left(y, f / \sum_{z \in \mathcal{Y}} f_z\right), \quad f \in \mathcal{F}.$$

The homogeneous score is called proper if the score in the right-hand side of the above expression is proper on $\mathcal{P}$. The function $\phi : \mathcal{F} \to \mathbb{R}$ is called 1-homogeneous if $\phi(\lambda f) = \lambda \phi(f)$ holds for all $\lambda > 0$ and all $f \in \mathcal{F}$. Let us introduce the relation between homogeneous scores and 1-homogeneous functions.

Theorem 2 ([16, 26]). *Suppose that $\phi : \mathcal{F} \to \mathbb{R}$ is a convex and 1-homogeneous function. Let $\partial \phi$ be a subgradient of ϕ. Define $S(y, q)$ as $-\partial_y \phi(q)$ for $q \in \mathcal{P}$. Then, S is a proper homogeneous score, and ϕ is the potential of S. Conversely, suppose that $S(y, f), f \in \mathcal{F}$ is a proper homogeneous score, i.e., $S(y, \lambda f) = S(y, f)$ holds for $\lambda > 0$ and $S(y, q)$ is proper for $q \in \mathcal{P}$. Then, $\phi(f) = -\sum_{y \in \mathcal{Y}} f_y S(y, f), f \in \mathcal{F}$ is a 1-homogeneous and convex function and S is expressed as the subgradient of ϕ.*

2.3 Neighborhood Systems of Sample Spaces

Let us consider the computation of scores. If the score $S(y, f)$ depends on all components of $(f_y)_{y \in \mathcal{Y}}$, the computation of the score will be intractable. Below we define localized scores for efficient comptuation.

The locality of the score is determined from a neighborhood system on the sample space. Suppose that a neighborhood of y in the sample space $\mathcal{Y}$ is defined as a subset $n(y) \subset \mathcal{Y}$ that contains y. The neighborhood $n(y)$ is regarded as the set of points that are close to y. The size of $n(y)$ is supposed to be small comparing to that of $\mathcal{Y}$. If the proper homogeneous score $S(y, f)$ depends only on $(f_z)_{z \in n(y)}$, the computation of $S(y, f)$ will be tractable. Theorem 2 immediately reveals that $z \in n(y)$ denotes $y \in n(z)$, i.e., the neighborhood system $\{n(z) \,|\, z \in \mathcal{Y}\}$ can be expressed as the adjacents of y in an undirected graph with the vertex set $\mathcal{Y}$. The detail is shown in [11]. When the neighborhood system of S is determined by the undirected graph G, the score is called G-local [11]. Some examples of proper G-local homogeneous scores are presented in Section 5.

3 Composite Local Bregman Divergences

In this section, we define a composite local Bregman divergence that is determined from a set of localized potentials. We show the relation between proper local homogeneous scores and composite local Bregman divergences.

We start from the 1-homogeneous convex function $\phi : \mathcal{F} \to \mathbb{R}$. Since ϕ is 1-homogeneous, ϕ is expressed as

$$\phi(f) = \sum_{y \in \mathcal{Y}} f_y \frac{\phi(f/f_y)}{|\mathcal{Y}|},$$

where $f/f_y = (f_z/f_y)_{z \in \mathcal{Y}} \in \mathcal{F}$. The domain of the function $\phi(f/f_y)$ can be thought of $\mathbb{R}_{++}^{\mathcal{Y} \setminus \{y\}}$, because the component $(f/f_y)_y = 1$ can be removed. As a result, 1-homogeneous convex function $\phi : \mathcal{F} \to \mathbb{R}$ can be constructed by the collection of convex functions defined on the domain $\mathbb{R}_{++}^{\mathcal{Y} \setminus \{y\}}$, $y \in \mathcal{Y}$, that is properly included in $\mathcal{F}$.

Conversely, we start from a collection of convex functions in order to obtain the potential, Bregman divergence and score. Let us define the neighborhood system on the sample space $\mathcal{Y}$ from the undirected graph $G = (\mathcal{Y}, E)$, where $E \subset \mathcal{Y} \times \mathcal{Y}$ is the set of edges. Here, $(y, z) \in E$ and $(z, y) \in E$ denote the same edge in the graph. We assume that E does not contain the loop such as (y, y). The set of adjacents of $y \in \mathcal{Y}$ is denoted as $b(y) = \{z \in \mathcal{Y} \,|\, (y, z) \in E\}$ and let $n(y) = b(y) \cup \{y\}$, where $b(y)$ does not have the component y. Suppose that a convex function $\phi_y : \mathbb{R}_{++}^{b(y)} \to \mathbb{R}$ is assigned to each $y \in \mathcal{Y}$. Then, the set of convex functions $\{\phi_y\}_{y \in \mathcal{Y}}$ produces the 1-homogeneous convex function

$$\phi(f) = \sum_{y \in \mathcal{Y}} f_y \phi_y(f_{b(y)}/f_y), \tag{4}$$

for $f \in \mathcal{F}$, where $f_{b(y)}/f_y$ is the sub-vector $(f_z/f_y)_{z \in b(y)} \in \mathbb{R}_{++}^{b(y)}$. It is clear that ϕ is 1-homogeneous. The convexity of ϕ is confirmed from the fact that $f_y \phi_y(f_{b(y)}/f_y)$ is the perspective
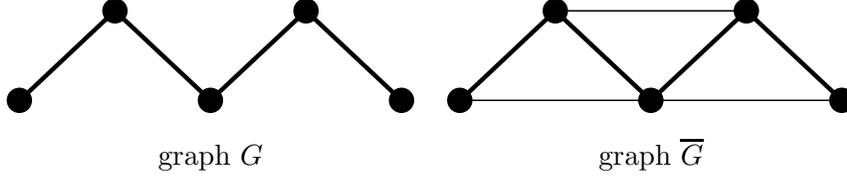
Figure 1: Graph G and its extension $\overline{G}$.

operation of the convex function ϕ_y; see Section 3.2.6 of [6]. The potential ϕ in (4) is referred to as the G-local potential, and ϕ_y is called the local potential.

The corresponding proper homogeneous score is given by

$$
\begin{aligned}
S(y, f) &= -\frac{\partial}{\partial f_y}\phi(f) = -\phi_y(f_{b(y)}/f_y) + \sum_{z \in b(y)} \frac{f_z}{f_y}\partial_z\phi_y(f_{b(y)}/f_y) - \sum_{z:y \in b(z)} \partial_y\phi_z(f_{b(z)}/f_z) \\
&= -\phi_y(f_{b(y)}/f_y) + \sum_{z \in b(y)} \frac{f_z}{f_y}\partial_z\phi_y(f_{b(y)}/f_y) - \sum_{z \in b(y)} \partial_y\phi_z(f_{b(z)}/f_z),
\end{aligned}
\tag{5}
$$

where we used the equality $\{z \in \mathcal{Y} \,|\, y \in b(z)\} = b(y)$. Since the neighborhoods $b(z)$, $z \in b(y)$ appear in $S(y, f)$, it is not necessarily the G-local score. Let us define the extended graph $\overline{G} = (\mathcal{Y}, \overline{E})$ of $G = (\mathcal{Y}, E)$ as

$$
\overline{E} = E \cup \{(z, z') \in \mathcal{Y} \times \mathcal{Y} \,|\, z \neq z' \text{ and there exists } y \in \mathcal{Y} \text{ s.t. } z, z' \in b(y)\},
$$

where $b(y)$ is the set of adjacents of the graph G. Figure 1 shows the graph G and its extension $\overline{G}$. Eq (5) means that the G-local potential leads to $\overline{G}$-local score.

Suppose that the convex function $\phi_y : \mathbb{R}_{++}^{b(y)} \to \mathbb{R}$ is expressed as

$$
\phi_y(f) = \sum_{z \in b(y)} \phi_{yz}(f_z),
\tag{6}
$$

where $\phi_{yz} : \mathbb{R}_{++} \to \mathbb{R}$ is a one-dimensional function. The potential of the form (6) is called additive. Then, the associated score is G-local. Indeed, the score obtained from (6) is

$$
S(y, f) = \sum_{z \in b(y)} \left\{ -\phi_{yz}(f_z/f_y) + \frac{f_z}{f_y}\phi'_{yz}(f_z/f_y) - \phi'_{zy}(f_y/f_z) \right\},
\tag{7}
$$

which depends on $f \in \mathcal{F}$ through $(f_z)_{z \in b(y)}$.

The Bregman divergence $D_\phi(f, g)$ for $f, g \in \mathcal{F}$ associated with the G-local potential (4) defined from the collection of local potentials $\{\phi_y\}_{y \in \mathcal{Y}}$ is expressed as

$$
D_\phi(f, g) = \sum_{y \in \mathcal{Y}} f_y S(y, g) + \phi(f) = \sum_{y \in \mathcal{Y}} f_y D_{\phi_y}(f_{b(y)}/f_y, g_{b(y)}/g_y).
\tag{8}
$$

The derivation is shown in the Appendix A. In the derivation of the above equality, we use the formula

$$
\sum_{x \in \mathcal{Y}} \sum_{y \in b(x)} A_{xy} = \sum_{x \in \mathcal{Y}} \sum_{y \in b(x)} A_{yx}
$$

7

for any $(A_{xy})_{x,y\in\mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$.

The undirected graph determining the neighborhood system and the collection of local potentials lead to the Bregman divergence as the sum of localized Bregman divergences $D_{\phi_y}, y \in \mathcal{Y}$. Here, let us define the composite local Bregman divergence.

Definition 3 (Composite local Bregman divergence). *Let $\mathcal{Y}_0$ be a subset of $\mathcal{Y}$ and $G = (\mathcal{Y}, E)$ be an undirected graph that determines the neighborhood $b(y), y \in \mathcal{Y}$. Each point $y \in \mathcal{Y}_0$ has a local potential $\phi_y : \mathbb{R}_{++}^{b(y)} \to \mathbb{R}$. For a collection of local potentials $\Phi = \{\phi_y \,|\, y \in \mathcal{Y}_0\}$, the composite local Bregman divergence is defined as*

$$D_\Phi(f, g) = \sum_{y\in\mathcal{Y}_0} f_y D_{\phi_y}(f_{b(y)}/f_y,\, g_{b(y)}/g_y), \quad f, g \in \mathcal{F}. \tag{9}$$

In the above definition, the subset $\mathcal{Y}_0$ rather than the whole sample space $\mathcal{Y}$ is introduced as the general expression.

The 1-homogeneous potential of $D_\Phi(f, g)$ is given by

$$\phi(f) = \sum_{y\in\mathcal{Y}_0} f_y \phi_y(f_{b(y)}/f_y).$$

The gradient of the above potential leads to the proper homogeneous score,

$$S_\Phi(y, f)$$
$$= \mathbf{1}[y \in \mathcal{Y}_0]\left\{ -\phi_y(f_{b(y)}/f_y) + \sum_{z\in b(y)} \frac{f_z}{f_y}\partial_z\phi_y(f_{b(y)}/f_y) \right\} - \sum_{z\in b(y)} \mathbf{1}[z \in \mathcal{Y}_0]\,\partial_y\phi_z(f_{b(z)}/f_z) \tag{10}$$

for $f \in \mathcal{F}$. The derivation of (10) is shown in the Appendix B. Conversely, any proper homogeneous score $S(y, f)$ is expressed by using the (sub)gradient of a G-local potential with some graph G. A trivial expression is given by using a singleton $\mathcal{Y}_0 = \{y_0\}$ and the complete graph G.

In the sequel sections, we investigate the condition of the graph G and the collection of local potentials Φ such that the proper homogeneous score $S_\Phi(y, f)$ becomes to be strictly proper.

4 Coincidence Axiom and Neighborhood Systems

In order to use the score to statistical inference, strictly proper scores are favorable rather than just proper ones. The Bregman divergence associated with the strictly proper score satisfies the coincidence axiom on $\mathcal{P}$, i.e.,

$$D_\Phi(p, q) = 0 \implies p = q \text{ for } p, q \in \mathcal{P}.$$

Firstly, we consider a sufficient condition of the coincidence axiom when the function set

$$\Phi = \{\phi_y : \mathbb{R}_{++}^{b(y)} \to \mathbb{R} \,|\, y \in \mathcal{Y}_0\}$$

includes only strictly convex functions. Secondly, we consider the case that ϕ_y's are convex but not strictly convex. In particular, we consider the case that ϕ_y is the potential defined as the localized variant of the pseudo-spherical divergence in Example 4.

4.1 Strictly Convex Local Potentials

Suppose that all local potentials in $\Phi = \{\phi_y \mid y \in \mathcal{Y}_0\}$ are strictly convex. We assume that the composite local Bregman divergence (9) vanishes, i.e., $D_\Phi(f,g) = 0$ for $f, g \in \mathcal{F}$. Then, due to the positivity of f_y for all $y \in \mathcal{Y}_0$, each term $D_{\phi_y}(f_{b(y)}/f_y, g_{b(y)}/g_y)$, $y \in \mathcal{Y}_0$ should be zero. The strict convexity of the potential ϕ_y leads to $f_{b(y)}/f_y = g_{b(y)}/g_y$ for all $y \in \mathcal{Y}_0$. Hence, we have

$$\lambda_y := f_y/g_y = f_z/g_z$$

for all $z \in b(y)$ and all $y \in \mathcal{Y}_0$. If there exists a point $z \in n(y) \cap n(y')$ for some $y, y' \in \mathcal{Y}_0$, we have $f_z/g_z = \lambda_y = \lambda_{y'}$. This implies that all f_z/g_z's are the same for $z \in n(y) \cup n(y')$ if $n(y) \cap n(y') \neq \emptyset$ for $y, y' \in \mathcal{Y}_0$.

On the basis of the above result, we give a sufficient condition that the composite local Bregman divergence satisfies the coincidence axiom on $\mathcal{P}$.

Theorem 3. *Suppose that the neighborhood system of the sample space $\mathcal{Y}$ is determined by the undirected graph $G = (\mathcal{Y}, E)$. For a subset $\mathcal{Y}_0$ of $\mathcal{Y}$, we assume*

$$\bigcup_{y \in \mathcal{Y}_0} n(y) = \mathcal{Y}$$

Let us define the graph $G_0 = (\mathcal{Y}_0, E_0)$ as

$$(y, y') \in E_0 \iff y \neq y' \text{ and } n(y) \cap n(y') \neq \emptyset.$$

We assume that the graph G_0 is connected and all functions in $\Phi = \{\phi_y \mid y \in \mathcal{Y}_0\}$ are strictly convex. Then, the composite local Bregman divergence D_Φ satisfies the coincidence axiom on $\mathcal{P}$.

Proof. Assume that $D_\Phi(p, q) = 0$ for $p, q \in \mathcal{P}$. The assumption of the theorem guarantees that for any $z, z' \in \mathcal{Y}$, there exist $y, y' \in \mathcal{Y}_0$ satisfying $z \in n(y)$ and $z' \in n(y')$. Thus, we have $p_z/q_z = p_y/q_y$ and $p_{z'}/q_{z'} = p_{y'}/q_{y'}$ as shown above. Since G_0 is connected, there is a path in G_0 such that $(y, y_1), (y_1, y_2), \ldots, (y_k, y') \in E_0$. The argument just before Theorem 3 guarantees the equations,

$$p_z/q_z = p_y/q_y = p_{y_1}/q_{y_1} = \cdots = p_{y_k}/q_{y_k} = p_{y'}/q_{y'} = p_{z'}/q_{z'}.$$

This implies the probabilities p and q are linearly dependent. Hence, they should be the same. $\square$

The similar argument of the above proof was presented in [19] to prove that the ratio matching score is strictly proper. Note that $D_\Phi(f, g) = 0$ for $f, g \in \mathcal{F}$ leads to the linear dependency of f and g. Even if strict convexity is locally assumed, the coincidence axiom on $\mathcal{F}$ is not guaranteed in general.

The following theorem shows that the connectedness of G_0 is closely related to that of G.

Theorem 4. *If $G_0 = (\mathcal{Y}_0, E_0)$ is connected, $G = (\mathcal{Y}, E)$ is connected. If $\mathcal{Y}_0 = \mathcal{Y}$ holds, the connectedness of G is equivalent with that of G_0.*

Proof. Let us assume that G_0 is connected. Then, for any $y, y' \in \mathcal{Y}_0$, there exists a sequence of points $y_0 = y, y_1, \ldots, y_{K-1}, y_K = y'$ such that $n(y_k) \cap n(y_{k+1}) \neq \emptyset$ holds for all $k = 0, 1, \ldots, K-1$. Hence, for the point $v_k \in n(y_k) \cap n(y_{k+1})$, we have $(y_k, v_k), (v_k, y_{k+1}) \in E$. Since the sequence of the points $y_0 = y, v_0, y_1, v_1, \ldots, y_{K-1}, v_{K-1}, y_K = y'$ connects y and y', we find that G is connected.

Let us prove the later part of the theorem. Suppose that G is connected. Then, for any $y, y' \in \mathcal{Y}_0 = \mathcal{Y}$, there exist a sequence of points $y_0 = y, y_1, \ldots, y_{K-1}, y_K = y'$ such that $(y_k, y_{k+1}) \in E$ holds for all $k = 0, 1, \ldots, K-1$. Hence, we have $n(y_k) \cap n(y_{k+1}) \neq \emptyset$ for $k = 0, 1, \ldots, K-1$, meaning that G_0 is connected. $\square$

4.2 Local Pseudo-spherical Potential

Here, let us consider the case that the convex function ϕ_y is not strictly convex. In particular, we assume that all ϕ_y's are the localized version of the pseudo-spherical potential shown in Example 4. More precisely, the local pseudo-spherical potential $\phi_y : \mathbb{R}_{++}^{b(y)} \to \mathbb{R}$ is defined as the $(1+\gamma)$ norm on $b(y)$, i.e.,

$$\phi_y(f) = \left(\sum_{z \in b(y)} f_z^{1+\gamma} \right)^{1/(1+\gamma)}$$

for a positive constant $\gamma > 0$. Since $\phi_y(\lambda f) = \lambda \phi_y(f)$ holds for $\lambda > 0$, ϕ_y is not strictly convex. Let D_{ϕ_y} on $\mathbb{R}_{++}^{b(y)}$ be the Bregman divergence defined from ϕ_y. Suppose that $D_{\phi_y}(f, g) = 0$ holds for $f, g \in \mathcal{F}$. Then, f and g are linearly dependent as shown in Example 4.

Let D_Φ be the composite local Bregman divergence defined from the local pseudo-spherical potentials.

Lemma 1. *Suppose that $D_\Phi(f, g) = 0$ holds for $f, g \in \mathcal{F}$. If $b(y) \cap b(y') \neq \emptyset$ for some $y, y' \in \mathcal{Y}_0$, all f_z/g_z's are the same for $z \in b(y) \cup b(y')$.*

In this case, $f_y/g_y = f_{y'}/g_{y'}$ is not guaranteed.

Proof. When $D_\Phi(f, g) = 0$ holds for $f, g \in \mathcal{F}$, the positivity of f leads to $D_{\phi_y}(f_{b(y)}/f_y, g_{b(y)}/g_y) = 0$ for all $y \in \mathcal{Y}_0$. Thus, $f_{b(y)}/f_y$ and $g_{b(y)}/g_y$ are linearly dependent, meaning that there exists λ_y such that $f_z/g_z = \lambda_y$ for all $z \in b(y)$ (not $n(y)$) and all $y \in \mathcal{Y}_0$. For $z \in b(y) \cap b(y')$, we have $f_z/g_z = \lambda_y = \lambda_{y'}$. This implies that all f_z/g_z's are the same for $z \in b(y) \cup b(y')$ if $b(y) \cap b(y') \neq \emptyset$ for some $y, y' \in \mathcal{Y}_0$. $\square$

On the basis of the above Lemma, we give a sufficient condition that $D_\Phi(p, q)$ leads to $p = q$ on $\mathcal{P}$.

Theorem 5. *Suppose that the neighborhood system of the sample space $\mathcal{Y}$ is determined by the undirected graph $G = (\mathcal{Y}, E)$. For a subset $\mathcal{Y}_0$ of $\mathcal{Y}$, we assume*

$$\bigcup_{y \in \mathcal{Y}_0} b(y) = \mathcal{Y}$$

10

Let us define the graph $G_0' = (\mathcal{Y}_0, E_0')$ as

$$(y, y') \in E_0' \iff y \neq y' \text{ and } b(y) \cap b(y') \neq \emptyset.$$

We assume that the graph G_0' is connected and all functions in $\Phi = \{\phi_y \,|\, y \in \mathcal{Y}_0\}$ are pseudo-spherical potentials. Then, the composite local Bregman divergence D_Φ satisfies the coincidence axiom on $\mathcal{P}$.

Proof. Assume that $D_\Phi(p, q) = 0$ for $p, q \in \mathcal{P}$. For any $z, z' \in \mathcal{Y}$, there exist $y, y' \in \mathcal{Y}_0$ satisfying $z \in b(y)$ and $z' \in b(y')$. Since G_0' is connected, there is a path in G_0' such that $(y, y_1), (y_1, y_2), \ldots, (y_k, y') \in E_0'$. Choose a vertex $z_i \in b(y_i)$ for each $i = 1, \ldots, k$. Lemma 1 guarantees

$$p_z / q_z = p_{z_1}/q_{z_1} = \cdots = p_{z_k}/q_{z_k} = p_{z'}/q_{z'}.$$

This implies the p and q are linearly dependent. For probability vectors, they should be the same. $\qquad\square$

In Theorem 5, the neighborhood $n(y)$ in Theorem 3 is replaced with $b(y)$.

Let us show a simple example that illustrates how the sufficient condition affects the strict convexity of the composite local Bregman divergence. The sample space is given as the binary variables $\mathcal{Y} = \{\pm 1\}^D$, and the Hamming distance d_H on $\mathcal{Y}$ is defined as the number of components at which the corresponding symbols are different, i.e.,

$$d_H(y, y') = |\{i \,|\, y_i \neq y_i'\}|$$

for $y = (y_1, \ldots, y_D)$ and $y' = (y_1', \ldots, y_D')$. Let us define the neighborhood system $G = (\mathcal{Y}, E)$ as

$$E = \{(y, y') \in \mathcal{Y} \times \mathcal{Y} \,|\, d_H(y, y') = 1\}.$$

For $\mathcal{Y}_0 = \mathcal{Y}$, the pseudo-spherical penitential with this neighborhood system does not produce the strictly proper score. For $D = 2$, the probabilities

$$p = (p_{(+1,+1)}, p_{(-1,-1)}, p_{(-1,+1)}, p_{(+1,-1)}) = (0.1, 0.1, 0.4, 0.4), \text{ and}$$
$$q = (q_{(+1,+1)}, q_{(-1,-1)}, q_{(-1,+1)}, q_{(+1,-1)}) = (0.2, 0.2, 0.3, 0.3)$$

satisfy $p_{(+1,+1)}/p_{(-1,-1)} = q_{(+1,+1)}/q_{(-1,-1)}$ and $p_{(+1,-1)}/p_{(-1,+1)} = q_{(+1,-1)}/q_{(-1,+1)}$. Thus, we have $D_\Phi(p, q) = 0$ though $p \neq q$. If the neighborhood system of $\mathcal{Y}$ is replaced with

$$E = \{(y, y') \in \mathcal{Y} \times \mathcal{Y} \,|\, d_H(y, y') = 1 \text{ or } 2\},$$

the corresponding local homogeneous score is strictly proper. Indeed, when $D_\Phi(p, q) = 0$ holds for two dimensional binary variables in $\mathcal{Y} = \{\pm 1\}^2$, the edge connecting $(+1, +1)$ and $(-1, -1)$ leads to the additional constraint $p_{(+1,+1)}/p_{(-1,+1)} = q_{(+1,+1)}/q_{(-1,+1)}$, etc. As a result, we obtain $p = q$.

5 Examples

We show pseudo-likelihood [5], composite likelihood [23], ratio matching [19], local density-power score and local pseudo-spherical score as examples of proper local homogeneous scores. Usually, the pseudo-likelihood, composite likelihood and ratio matching are defined for the probability over the multidimensional binary vector space. In our formation, they can be defined on any discrete sample space endowed with the graph-based neighborhood system.

In the following examples, we assume $\mathcal{Y}_0 = \mathcal{Y}$.

5.1 Pseudo-likelihood

The pseudo-likelihood is a common estimation method that does not require the computation of normalization constant [5]. Usually, the pseudo-likelihood is used for multidimensional discrete samples such as $\mathcal{Y} = \{\pm 1\}^D$. For the sample $y = (y_1, \ldots, y_D) \in \{\pm 1\}^D$, we define $y_{\backslash i}$ as the $D - 1$ dimensional vector $(y_j)_{j \neq i} \in \{\pm 1\}^{D-1}$. For the random variable Y taking a value in $\{\pm 1\}^D$, the sub-vector $Y_{\backslash i}$ is defined in the same way. The pseudo-likelihood for the sample y is defined as

$$\prod_{i=1}^{D} \Pr(Y = y \mid Y_{\backslash i} = y_{\backslash i}),$$

where $\Pr(A|B)$ denotes the conditional probability of A given B.

When the statistical model (3) on $\mathcal{Y} = \{\pm 1\}^D$ is used, the normalization constant Z_θ is not needed to compute the pseudo-likelihood. The set of edges of the graph $G = (\mathcal{Y}, E)$ is given as

$$E = \{(y, y') \in \mathcal{Y} \times \mathcal{Y} \mid d_H(y, y') = 1\},$$

where d_H is the Hamming distance on $\mathcal{Y}$. Then, the adjacents of y is

$$b(y) = \{\bar{y}^{(i)} = (\bar{y}_1^{(i)}, \ldots, \bar{y}_D^{(i)}) \in \mathcal{Y} \mid i = 1, \ldots, D\},$$

$$\bar{y}_j^{(i)} = \begin{cases} y_j, & j \neq i, \\ -y_i, & j = i, \end{cases} \quad (j = 1, \ldots, D).$$

The event, $Y_{\backslash i} = y_{\backslash i}$, is the same as $Y \in \{y, \bar{y}^{(i)}\}$. The statistical model q_θ is defined by

$$q_\theta = \frac{f_\theta}{Z_\theta}, \quad q_\theta \in \mathcal{P}, \ f_\theta \in \mathcal{F}, \ Z_\theta = \sum_{y \in \mathcal{Y}} f_{\theta, y},$$

where θ is the parameter to specify the probability. The logarithm of the pseudo-likelihood is expressed as

$$\sum_{i=1}^{D} \log \frac{q_{\theta, y}}{q_{\theta, y} + q_{\theta, \bar{y}^{(i)}}} = -\sum_{i=1}^{D} \log(1 + f_{\theta, \bar{y}^{(i)}} / f_{\theta, y}) = -\sum_{z \in b(y)} \log(1 + f_{\theta, z} / f_{\theta, y}).$$

Hence, the score of the pseudo likelihood is given as

$$S(y, f) = \sum_{z \in b(y)} \log(1 + f_z / f_y), \quad f \in \mathcal{F}. \tag{11}$$

12

The corresponding local potential is

$$\phi_y(g) = -\sum_{z \in b(y)} \log(1 + g_z), \quad g \in \mathbb{R}_{++}^{b(y)}$$

which is G-local, additive and strictly convex on $\mathbb{R}_{++}^{b(y)}$. One can confirm that the equality (7) holds. The composite local Bregman divergence is

$$D_\Phi(p, q) = \sum_{y \in \mathcal{Y}} p_y \sum_{z \in b(y)} \sum_{a \in \{y,z\}} \frac{p_a}{p_y + p_z} \log \frac{p_a/(p_y + p_z)}{q_a/(q_y + q_z)}, \quad p, q \in \mathcal{P}.$$

Theorem 3 guarantees that the above composite local Bregman divergence satisfies the coincidence axiom on $\mathcal{P}$. The set $\mathcal{Y}_0 = \mathcal{Y}$ can be replaced with $\mathcal{Y}_0 = \{(y_1, \ldots, y_D) \in \{\pm 1\}^D \mid \sum_{i=1}^D y_i = \text{even}\}$, while keeping the score (11) to be strictly proper.

Even for the general discrete sample space $\mathcal{Y}$, the pseudo-likelihood defined by the score (11) is strictly proper when the neighborhood system is properly defined.

5.2   Composite Likelihood and its Extension

Composite likelihood (CL) was proposed as an extension of the pseudo-likelihood in order to improve the efficiency [23]. Let the sample space be $\mathcal{Y} = \{\pm 1\}^D$ and the sample is denoted as $y = (y_1, \ldots, y_D) \in \{\pm 1\}^D$. For a collection of subsets $A_\ell \subset \{1, \ldots, D\}$, $\ell = 1, \ldots, m$, we define $y_{A_\ell} = (y_i)_{i \in A_\ell}$ as the sub-vector of $y = (y_1, \ldots, y_D) \in \mathcal{Y}$. The CL for the sample y is defined as

$$\prod_{i=1}^m \Pr(Y_{A_\ell} = y_{A_\ell} \mid Y_{A_\ell^c} = y_{A_\ell^c}), \tag{12}$$

where A^c is the complement of the set A. The standard pseudo-likelihood is obtained by setting $A_\ell = \{\ell\}$ for $\ell = 1, \ldots, D$. The condition $Y_{A_\ell^c} = y_{A_\ell^c}$ can be expressed as the neighborhood system on $\mathcal{Y}$. Let us define the graph $G = (\mathcal{Y}, E)$ as

$$(y, z) \in E \Leftrightarrow y \neq z \text{ and there exists } \ell \text{ such that } y_{A_\ell^c} = z_{A_\ell^c}.$$

Then, we have the following theorem.

Theorem 6. *Suppose $\mathcal{Y}_0 = \mathcal{Y}$. Then, the undirected graph $G_0 = (\mathcal{Y}_0, E_0)$ determined from G is connected if and only if*

$$\bigcup_{\ell=1}^m A_\ell = \{1, \ldots, D\} \tag{13}$$

holds.

Proof. Theorem 4 guarantees that the connectedness of G_0 is equivalent with that of G when $\mathcal{Y}_0 = \mathcal{Y}$ holds. Firstly, we assume (13). Let $y = (y_1, \ldots, y_D)$ and $y' = (y_1', \ldots, y_D')$ be arbitrary points in $\mathcal{Y}$, and suppose $y_i \neq y_i'$. There exists an index set A_ℓ that includes i, meaning that the point $y'' = (y_1, \ldots, y_{i-1}, y_i', y_{i+1}, \ldots, y_D)$ is an adjacents of y. By repeating this process, we can confirm that there exists a path connecting y and y' in the graph G.

Conversely, let us assume that $1 \notin \bigcup_{\ell=1}^m A_\ell$. Then, we have $1 \in \bigcap_{\ell=1}^m A_\ell^c$. This means that the two points, $y = (y_1, \ldots, y_D)$ and $y' = (y_1', \ldots, y_D')$, can be adjacents to each other only when $y_1 = y_1'$. If $y_1 \neq y_1'$, a path connecting y and y' does not exist. $\qquad\square$

We define the subset $b_\ell(y) \subset \{\pm 1\}^D$ by

$$b_\ell(y) = \{z \in \{\pm 1\}^D \mid z \neq y,\ y_{A_\ell^c} = z_{A_\ell^c}\}.$$

Then, we have $b(y) = \cup_{\ell=1}^m b_\ell(y)$. The score of the CL is expressed as

$$S(y, f) = -\sum_{\ell=1}^m \log \frac{f_y}{f_y + \sum_{z \in b_\ell(y)} f_z} = \sum_{\ell=1}^m \log \left(1 + \sum_{z \in b_\ell(y)} f_z / f_y \right), \quad f \in \mathcal{F}.$$

The potential is obtained by $\phi(p) = -S(p, p)$ on $\mathcal{P}$. The set of local potentials, $\Phi = \{\phi_y \mid y \in \mathcal{Y}\}$, is given as

$$\phi_y(g) = -\sum_{\ell=1}^m \log \left(1 + \sum_{z \in b_\ell(y)} g_z \right), \quad g \in \mathbb{R}_{++}^{b(y)}. \tag{14}$$

The above ϕ_y may not be strictly convex. We show a sufficient condition such that ϕ_y is strictly convex on $\mathbb{R}_{++}^{b(y)}$. For the subset $b_\ell(y) \subset b(y)$, let us define the vector $\mathbf{1}_{b_\ell(y)} \in \mathbb{R}_{++}^{b(y)}$ as $(\mathbf{1}_{b_\ell(y)})_z = 1$ for $z \in b_\ell(y)$ and $(\mathbf{1}_{b_\ell(y)})_z = 0$ for $z \in b(y) \setminus b_\ell(y)$.

Theorem 7. *The local potential (14) is strictly convex on $\mathbb{R}_{++}^{b(y)}$ if the rank of the $|b(y)|$ by m matrix $(\mathbf{1}_{b_1(y)}, \dots, \mathbf{1}_{b_m(y)})$ is $|b(y)|$.*

Proof. A simple calculation yields that the Hessian matrix of ϕ_y is given as the $|b(y)|$ by $|b(y)|$ matrix

$$\nabla^2 \phi_y(f) = \sum_{\ell=1}^m \frac{1}{(1 + \sum_{z \in b_\ell(y)} f_z)^2} \mathbf{1}_{b_\ell(y)} \mathbf{1}_{b_\ell(y)}^T.$$

For a column vector $a \in \mathbb{R}^{b(y)}$, we suppose that $a^T (\nabla^2 \phi_y(f)) a = 0$. Then, we have

$$\sum_{z \in b_\ell(y)} a_z = 0, \quad \ell = 1, \dots, m.$$

If the condition of the theorem holds, we have $a = 0 \in \mathbb{R}^{b(y)}$. Hence, the Hessian matrix of ϕ_y is positive definite and the function ϕ_y is strictly convex. □

The rank condition in Theorem 7 and (13) is not directly related to each other as shown in the following example.

Example 5 (Rank condition)**.** *Suppose $D = 3$ for $\mathcal{Y} = \{\pm 1\}^D$.*

1. *For $A_1 = \{1\}$ and $A_2 = \{2\}$, the condition (13) does not hold. We have $|b(y)| = 2$. Since the matrix $(\mathbf{1}_{b_1(y)}, \mathbf{1}_{b_2(y)})$ is equal to the 2 by 2 identity matrix, the rank condition is satisfied.*

2. *For $A_1 = \{1\}$ and $A_2 = \{2, 3\}$, the condition (13) holds. Because of $|b(y)| = 4$, the 4 by 2 matrix $(\mathbf{1}_{b_1(y)}, \mathbf{1}_{b_2(y)})$ does not satisfy the rank condition.*

The homogeneous score given by the gradient of the potential $\phi(f) = \sum_{y \in \mathcal{Y}} f_y \phi_y(f_{b(y)}/f_y)$ is expressed as

$$S(y, q) = \sum_{\ell=1}^{m} \left\{ \log \left(1 + \sum_{z \in b_\ell(y)} q_z/q_y \right) - \sum_{z \in b_\ell(y)} \frac{q_z}{\sum_{z' \in n_\ell(y)} q_{z'}} + \sum_{z: y \in b_\ell(z)} \frac{q_z}{\sum_{z' \in n_\ell(z)} q_{z'}} \right\}$$

$$= \sum_{\ell=1}^{m} \left\{ - \log q(y|n_\ell(y)) + \sum_{z: y \in n_\ell(z)} q(z|n_\ell(z)) - 1 \right\} \tag{15}$$

at $q \in \mathcal{P}$, where $n_\ell(y) = b_\ell(y) \cup \{y\}$ and $q(y|A)$ for $y \in A \subset \mathcal{Y}$ denotes the conditional probability $q_y / \sum_{z \in A} q_z$ of the probability $q \in \mathcal{P}$. For the multidimensional binary space $\{\pm 1\}^D$ with the above neighborhood system, we have

$$z \in n_\ell(y) \iff n_\ell(y) = n_\ell(z) \iff y \in n_\ell(z).$$

Thus, we obtain

$$\sum_{z: y \in n_\ell(z)} q(z|n_\ell(z)) = \sum_{z: z \in n_\ell(y)} q(z|n_\ell(z)) = \sum_{z: z \in n_\ell(y)} q(z|n_\ell(y)) = 1.$$

Therefore, the second and third terms in (15) cancel out, and the standard CL is obtained. Though the local potentials ϕ_y, $y \in \mathcal{Y}$ are not additive, the score becomes G-local.

Let us define an extension of the CL on the general sample space $\mathcal{Y}$ that is not necessarily the multidimensional binary space. Suppose that the neighborhood system $n(y)$, $y \in \mathcal{Y}$ is determined from an undirected graph $G = (\mathcal{Y}, E)$, and $n_\ell(y)$, $\ell = 1, \ldots, m$ be subsets of $n(y)$ such that $\cup_{\ell=1}^{m} n_\ell(y) = n_\ell(y)$ holds. Then, (15) is a G-local homogeneous proper score on the general sample space $\mathcal{Y}$. Note that the second and third terms in (15) will not cancel out. Thus, we need to use the proper $\overline{G}$-local score (15) instead of the standard G-local CL.

Let us summarize the theoretical properties of the extended CL. Suppose that the rank condition in Theorem 7 is satisfied and that the undirected graph $G = (\mathcal{Y}, E)$ is connected. Then, Theorems 3, 4, and 7 guarantees that the score (15) is strictly proper, $\overline{G}$-local, and homogeneous. Note that Theorem 7 works for any discrete sample space $\mathcal{Y}$.

Example 6. *Suppose $\mathcal{Y} = \{\pm 1\}^D$. For the pseudo-likelihood, the sets $A_1, \ldots, A_m$ in (12) is given by $\{1\}, \ldots, \{D\}$ and the rank condition in Theorem 7 is satisfied. As the another example, let us define the sets $A_1, \ldots, A_m$ as $\{1\}, \{3\}, \ldots, \{D-1\}$ and $\{1, 2\}, \{3, 4\}, \ldots, \{D-1, D\}$ with $m = D$ if D is even. Then, the rank condition in Theorem 7 holds. Similar result holds for the odd D. As shown in Proposition 7.1 of [7], the standard CL (12) is strictly proper if*

$$\bigcup_\ell A_\ell = \{1, \ldots, D\}, \quad A_\ell \cap A_{\ell'} = \emptyset, \ \ell \neq \ell' \tag{16}$$

holds. The rank condition in Theorem 7 is more strict than (16). Under the condition (16), the convexity of local potentials is not guaranteed, but the score becomes strictly proper. On the other hand, the rank condition is a sufficient condition for the local potential to be strictly convex. An advantage of the rank condition is that it is available for any discrete sample space.

We consider the sufficient condition under which (15) on the general sample space $\mathcal{Y}$ is reduced to the form of the standard CL. When $n_\ell(y) = n_\ell(z)$ holds for any $z \in n_\ell(y)$, the

neighborhood system n_ℓ determined from G is called the equivalence function [21]. Suppose that n_ℓ is the equivalence function for all ℓ. In the same way as the case of $\mathcal{Y} = \{\pm 1\}^D$, one can prove that (15) is reduced to the G-local CL

$$S(y, q) = -\sum_{\ell=1}^{m} \log q(y|n_\ell(y))$$

on the general sample space $\mathcal{Y}$. In [21], the equivalence function is used without referring to the relation between G-local and $\overline{G}$-local scores.

5.3 Ratio Matching

Suppose that the sample space is defined as $\mathcal{Y} = \{\pm 1\}^D$ and that two points in $\mathcal{Y}$ are adjacents to each other if the Hamming distance between them is equal to one. The neighborhood system is determined by such $G = (\mathcal{Y}, E)$. The score of the ratio matching is defined as

$$\begin{aligned} S(y, q) &= \sum_{i=1}^{D} \frac{1}{(1 + q_y/q_{\bar{y}^{(i)}})^2} \\ &= \sum_{z \in b(y)} \frac{1}{(1 + q_y/q_z)^2}, \end{aligned} \tag{17}$$

which is G-local. The normalization constant of the probability q is not needed to compute $S(y, q)$ due to the ratio q_y/q_z. The corresponding local potential is

$$\phi_y(g) = -\frac{1}{2} \sum_{z \in b(y)} \frac{g_z}{1 + g_z},$$

which is G-local, additive and strictly convex on $\mathbb{R}_{++}^{b(y)}$. One can confirm that the equality (7) holds. The composite local Bregman divergence is

$$D_\Phi(p, q) = \sum_{y \in \mathcal{Y}} p_y \sum_{z \in b(y)} \frac{(p_z/p_y - q_z/q_y)^2}{(1 + p_z/p_y)(1 + q_z/q_y)^2} = \sum_{y \in \mathcal{Y}} \sum_{z \in b(y)} (p_y + p_z) \left(\frac{p_z}{p_y + p_z} - \frac{q_z}{q_y + q_z} \right)^2.$$

The strict convexity of ϕ_y and $\mathcal{Y}_0 = \mathcal{Y}$ ensure that the assumptions of Theorem 3 hold. Hence, the above composite local Bregman divergence satisfies the coincidence axiom on $\mathcal{P}$. Even for the general discrete sample space $\mathcal{Y}$, the ratio matching using (17) is strictly proper whenever the graph $G = (\mathcal{Y}, E)$ is connected.

5.4 Local Density-Power Score

Below we present the strictly proper local homogeneous score defined from the density-power potentials in Example 3. For the graph $G = (\mathcal{Y}, E)$ that determines the neighborhood system of the discrete sample space $\mathcal{Y}$, let us define the G-local additive potential $\phi_y : \mathbb{R}_{++}^{b(y)} \to \mathbb{R}$ as

$$\phi_y(f) = \frac{1}{1 + \gamma} \sum_{z \in b(y)} f_z^{1+\gamma},$$

16

for a fixed positive constant γ. The potential is strictly convex. The local homogeneous score defined from $\{\phi_y \,|\, y \in \mathcal{Y}_0\}$ with $\mathcal{Y}_0 = \mathcal{Y}$ is

$$S(y, f) = \sum_{z \in b(y)} \left\{ \frac{\gamma}{1+\gamma} \left(\frac{f_z}{f_y} \right)^{1+\gamma} - \left(\frac{f_y}{f_z} \right)^{\gamma} \right\}, \quad f \in \mathcal{F}$$

which is G-local. The composite local Bregman divergence is given as

$$D_\Phi(f, g) = \sum_{y \in \mathcal{Y}} f_y \sum_{z \in b(y)} \left\{ \frac{1}{1+\gamma} \left(\frac{f_z}{f_y} \right)^{1+\gamma} + \frac{\gamma}{1+\gamma} \left(\frac{g_z}{g_y} \right)^{1+\gamma} - \left(\frac{g_z}{g_y} \right)^{\gamma} \frac{f_z}{f_y} \right\}, \quad f, g \in \mathcal{F}.$$

Theorem 3 guarantees that the coincidence axiom on $\mathcal{P}$ holds if G is connected. Accordingly, the above homogeneous score is strictly proper.

5.5 Local Pseudo-spherical Score

We show the explicit expression of the local pseudo-spherical score. For the graph $G = (\mathcal{Y}, E)$ that determines the neighborhood system of a discrete sample space $\mathcal{Y}$, let us define the G-local potential $\phi_y : \mathbb{R}_{++}^{b(y)} \to \mathbb{R}$ as

$$\phi_y(f) = \|f_{b(y)}\|_{1+\gamma} = \left(\sum_{z \in b(y)} f_z^{1+\gamma} \right)^{1/(1+\gamma)}.$$

This potential is convex but not strictly convex on $\mathbb{R}_{++}^{b(y)}$. Moreover, ϕ_y is not additive. Hence, the corresponding score is not G-local but $\overline{G}$-local in general. The score derived from the above potential is given as

$$S(y, f) = -\sum_{z \in b(y)} \frac{(f_y/f_z)^{\gamma}}{\|f_{b(z)}/f_z\|_{1+\gamma}^{\gamma}} = -\sum_{z \in b(y)} \|f_{b(z)}/f_y\|_{1+\gamma}^{-\gamma}, \quad f \in \mathcal{F}$$

and the composite local Bregman divergence is

$$D_\Phi(f, g) = \sum_{y \in \mathcal{Y}} f_y \sum_{z \in b(y)} \left(\|f_{b(z)}/f_y\|_{1+\gamma}^{-\gamma} - \|g_{b(z)}/g_y\|_{1+\gamma}^{-\gamma} \right), \quad f, g \in \mathcal{F}.$$

Whenever the neighborhood system is properly defined, $D_\Phi(p, q)$ satisfies the coincidence axiom on $\mathcal{P}$ and the corresponding local homogeneous score is strictly proper.

Let us consider the local pseudo-spherical score on $\mathcal{Y} = \{\pm 1\}^D$. The undirected graph $G = (\mathcal{Y}, E)$ is defined such that $(y, y') \in E$ if and only if $d_H(y, y') = 1$, meaning that the neighborhood system of $\mathcal{Y}$ is expressed as the D-dimensional hyper-cube. Suppose $\mathcal{Y}_0 = \mathcal{Y}$. Then, it is clear that the assumptions on the graph in Theorem 3 are satisfied. The first assumption in Theorem 5, $\cup_{y \in \mathcal{Y}_0} b(y) = \mathcal{Y}$, also holds. Let us consider the second condition of Theorem 5, i.e., the connectedness of G'_0. We see that $(y, y') \in E'_0$ if and only if $d_H(y, y')$ is equal to 2. Suppose that $y, y' \in \mathcal{Y}_0 = \mathcal{Y}$ are connected by a path $(y, y_1), (y_1, y_2), \ldots, (y_k, y')$ in G'_0. Then, we have $d_H(y, y_1) = d_H(y_1, y_2) = \cdots = d_H(y_k, y') = 2$. If $d_H(y, y')$ is odd, there does not exist a path connecting y and y' in G'_0. As a result, G'_0 is not connected. Indeed, G'_0 has two disjoint components. The two components are expressed by

$$C_{\mathrm{odd}} = \{y \in \mathcal{Y}_0 \,|\, d_H(y_0, y) = \mathrm{odd}\}, \quad C_{\mathrm{even}} = \{y \in \mathcal{Y}_0 \,|\, d_H(y_0, y) = \mathrm{even}\},$$

where $y_0 \in \mathcal{Y}_0$ is a fixed point. By adding an edge in between C_{odd} and C_{even}, we obtain a connected graph. In order to obtain strictly proper local homogeneous scores from local pseudo-spherical potentials, we need to define the neighborhood system carefully.

6 Classification

Let us consider classification problems. Suppose that i.i.d. samples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ are observed from the probability distribution $p(x)p(y|x)$, where $\mathcal{X}$ is a domain of feature x and $\mathcal{Y}$ is a finite label set. Our goal is to estimate the conditional probability $p(y|x)$ or to predict the label $y \in \mathcal{Y}$ of a newly observed x. We assume the statistical model of the conditional probability

$$q(y|x;\theta) = \frac{f(y|x;\theta)}{Z_\theta(x)}, \quad Z_\theta(x) = \sum_{y \in \mathcal{Y}} f(y|x;\theta),$$

where $q(\cdot|x;\theta) \in \mathcal{P}$, $f(\cdot|x;\theta) \in \mathcal{F}$ and θ is the parameter of the model. When the size of the label set $\mathcal{Y}$ is large, the computation of $Z_\theta(x)$ is intractable. The proper local homogeneous score is used in order to avoid the computation of the normalization constant.

The score is applicable for the estimation of the conditional probability. Let $\mathbb{E}_x[\cdot]$ be the expectation with respect to the distribution of the feature x. When $q(\cdot|x) \in \mathcal{F}$ is used to fit the data, the empirical mean of the proper homogeneous score satisfies

$$\frac{1}{n}\sum_{i=1}^{n} S(y_i, q(\cdot|x_i)) \longrightarrow \mathbb{E}_x\left[\sum_{y \in \mathcal{Y}} p(y|x)S(y, q(\cdot|x))\right]$$

$$= \mathbb{E}_x\left[S(p(\cdot|x), q(\cdot|x))\right] \geq \mathbb{E}_x\left[S(p(\cdot|x), \lambda\, p(\cdot|x))\right], \quad \lambda > 0.$$

The minimization of the empirical score with respect to the unnormalized model is expected to provide a good estimator of $p(y|x)$ up to a constant factor.

7 Simulations

We performed numerical experiments to investigate how scores and neighborhood systems affect estimation accuracy. In the first experiment, We used the fully-visible Boltzmann machine as the statistical model to estimate probability functions. Next, we consider the classification problems with a small label set to study the relation between neighborhood system and prediction accuracy.

7.1 Boltzmann Machines

In the experiments, the sample space was the multidimensional binary variables $\mathcal{Y} = \{\pm 1\}^D$. The fully-visible and fully connected Boltzmann Machine (BM) on $\mathcal{Y}$ is the statistical model defined as

$$q_W(y) = \frac{\exp(y^T W y)}{Z_W}, \quad Z_W = \sum_{z \in \mathcal{Y}} \exp(z^T W z),$$

where $y \in \mathcal{Y}$ is the D-dimensional column vector. The model parameter W is the D by D symmetric matrix. For large D, the computation of Z_W is intractable. When D-dimensional i.i.d. samples $y(1), \ldots, y(n) \in \mathcal{Y}$ are observed, our task is to estimate the parameter W such that $q_W(y)$ approximates the probability distribution of the samples.

To estimate the parameter W, the unnormalized model

$$f_W(y) = \exp(y^T W y)$$

is used with proper local homogeneous scores in order to circumvent the computation of the normalization constant. The estimator $\widehat{W}$ of W is obtained as the minimum solution of empirical score,

$$\min_W \frac{1}{n} \sum_{t=1}^{n} S(y(t), f_W).$$

In the experiment, we used the pseudo-likelihood (PL), ratio matching (RM), local pseudo-spherical (PS) score as proper local homogeneous scores. The parameter γ in the pseudo-spherical score was set to $\gamma = 1$ and 3. We examined two neighborhood systems determined by the Hamming distance d_H, one is $n_1(y) = \{y' \mid d_H(y, y') \le 1\}$ and the other is $n_2(y) = \{y' \mid d_H(y, y') \le 2\}$.

We estimated the parameter W for $D = 8, 16$ and 32. The sample size was set to $n = 1000$ and 3000. When $D = 8$, the maximum likelihood estimator was also computed. In order to evaluate the accuracy of the estimator $\widehat{W}$, the negative log-loss, $-\frac{1}{N} \sum_{t=1}^{N} \log q_{\widehat{W}}(\widetilde{y}(t))$, was computed on the test samples $\widetilde{y}(1), \ldots, \widetilde{y}(N) \in \mathcal{Y}$. Here, the logarithm of the normalization constant $Z_{\widehat{W}}$ was approximated by using the annealed importance sampling(AIS) [30]. Though the AIS is available to obtain an approximation of $\log Z_W$ of a single parameter W, using the AIS in the optimization process is computationally impractical.

In the first setup, the training and test samples were generated from the probability distribution defined by the Boltzmann machine, meaning that the statistical model was correct. We generated the random matrix $W = (\widetilde{W} + \widetilde{W}^T)/2$, where each element of $\widetilde{W}$ was independently distributed from the standard normal distribution. Then, the training and test samples were sampled from the probability $q_W(y)$ by using the Markov chain Monte Carlo (MCMC) method [1] with the transition probability $q_W(y_i|(y_j)_{j\neq i})$ for $y = (y_i)_{i=1,\ldots,D} \in \mathcal{Y}$. The burn-in period was set to $100D$. The size of the test samples was fixed to $N = 5000$. The experiments were repeated 20 times. In each repetition, the different matrix W was used. The averaged negative log-loss was calculated to evaluate the accuracy of the estimator $\widehat{W}$ for each estimation method.

The mean and standard deviation (sd) (resp. median and median absolute deviation (mad)) of the averaged negative log-loss was shown in the top (resp. bottom) panel of Table 1. The negative log-loss for the uniform distribution, i.e., $D \log 2$, was also presented in the bottom line. When $n = 1000$ and $D = 16, 32$, the mean value of the averaged loss was extremely large and the median stays small. It means that the averaged loss can become extremely large for a certain matrix W. The estimator using 3000 samples took into account the wider range of the sample space than the estimator using 1000 samples. As a result, the accuracy was much improved. About the neighborhood system, the estimator using the neighborhood n_2 was superior to that using the neighborhood n_1. For the pseudo-likelihood, Liang and Jordan proved that the wider neighborhood leads to more accurate estimator under some conditions [21]. We numerically

verified that not only the pseudo-likelihood but also the ratio matching score has the same tendency.

The estimation accuracy of the local PS score was comparable to the other methods with n_1. Though the local PS-score with the neighborhood system n_1 is not strictly proper on the set of all probabilities over $\{\pm1\}^D$, it might be strictly proper on smaller models such as Boltzmann machines. In addition, the estimation accuracy of the local PS score was not significantly affected by the choice of γ. This will be because the statistical model was well-specified, and the local PS score with any γ provides a consistent estimator for Boltzmann machines.

In the second setup, the handwritten data called "Optical Recognition of Handwritten Digits Data Set" in UCI repository [22] was used as training and test datasets. The handwritten data has 5620 samples, 64 features, and each sample has one label in $\{0, \ldots, 9\}$. Each feature takes an integer ranging from 0 to 16. In the experiment, the feature vector in $\{0, \ldots, 16\}^{64}$ was converted into the binary sample in $\{\pm1\}^{64}$ by changing 0 to -1 and $1, \ldots, 16$ to $+1$, and the Boltzmann machine q_W was used as the statistical model. Here, the class label was ignored and the probability distribution of the converted features was estimated. The training sample size was set to $n = 1000$ and $n = 3000$, and the rest were used as test samples. We chose $D = 8, 16, 32$ features out of 64 features. In the same way as the first setup, we examined the PS, RM, and local PS score to estimate the parameters, and the estimation accuracy was evaluated by the negative log-loss on the test samples. The AIS was used for the computation of the test loss.

Table 2 shows the mean value of the averaged negative log-loss for each estimator with the standard deviation. In this case, the median was almost the same as the mean. By comparing the result for n_1 and n_2, we find that the estimator using the wider neighborhood achieved a higher accuracy particularly in the case of $D = 32$. In the local PS score, the choice of γ significantly affected the estimation accuracy. In the second setup, the statistical model will not be well specified. Thus, the estimation bias depended on the parameter γ. In practical data analysis, one needs to choose the parameter γ carefully.

7.2 Classification

In this section, we used the scores to classification problems. The handwritten data in the previous section was again used. The data had 5620 samples, 64 features, and 10 labels. This time, the features were not reduced to the binary, i.e., the original features were used. Since the data has only 10 labels, the local homogeneous score is not needed to reduce the computational cost. Here, we investigate the influence of the neighborhood system to the prediction accuracy in classification problems. Using the training data $(x_1, y_1), \ldots, (x_n, y_n)$, the conditional probability of the label $y \in \mathcal{Y} = \{0, 1, \ldots, 9\}$ given the input vector $x \in \{0, 1, \ldots, 16\}^{64}$ was estimated. The statistical model for the conditional probability was defined as

$$q(y|x; \theta) = \frac{f(y|x; \theta)}{Z_\theta(x)}, \quad Z_\theta(x) = \sum_{y \in \mathcal{Y}} f(y|x; \theta),$$

where

$$f(y|x; \theta) = \exp\{\theta_y^T t(x)\}, \quad \theta_y \in \mathbb{R}^{64}, \ t(x) = (x_1, \ldots, x_{64})^T.$$

The dimension of the model parameter $\theta = (\theta_y)_{y \in \mathcal{Y}}$ is 640. The model is overparametrized, i.e., the different parameter can specify the same conditional probability. The estimator was evaluated by the prediction error rate and negative log-loss on the test samples.

Table 1: Negative log-loss on test data for maximum likelihood estimator (MLE), pseudo-likelihood (PS), ratio matching (RM), and local pseudo-spherical score (PS). The mean with standard deviation and median with median absolute deviation are shown.

	dim $x = 8$		dim $x = 16$		dim $x = 32$	
	$n = 1000$	$n = 3000$	$n = 1000$	$n = 3000$	$n = 1000$	$n = 3000$
			mean (sd)			
MLE	1.491 (0.642)	1.624 (0.667)	–	–	–	–
PL: n_1	4.273 (5.713)	1.752 (0.779)	40.577 (86.027)	4.480 (5.226)	41.839 (93.027)	6.393 (3.897)
PL: n_2	1.700 (0.888)	1.646 (0.664)	26.409 (66.334)	2.682 (2.639)	36.205 (89.569)	4.295 (2.773)
RM: n_1	4.277 (5.715)	1.755 (0.787)	39.707 (84.838)	4.528 (5.740)	40.917 (89.778)	6.365 (4.144)
RM: n_2	3.481 (5.275)	1.652 (0.660)	38.264 (85.522)	3.940 (4.964)	40.464 (93.362)	5.521 (3.620)
PS($\gamma = 1$): n_1	3.744 (5.482)	1.675 (0.687)	38.993 (83.623)	4.471 (5.849)	38.478 (87.264)	5.497 (3.159)
PS($\gamma = 3$): n_1	4.014 (5.698)	1.756 (0.754)	39.670 (84.260)	4.384 (5.409)	40.370 (95.292)	6.243 (3.835)
			median (mad)			
MLE	1.413 (0.720)	1.604 (0.890)	–	–	–	–
PL: n_1	1.669 (1.154)	1.667 (0.958)	5.614 (5.451)	2.435 (1.050)	11.578 (9.743)	4.934 (1.799)
PL: n_2	1.445 (0.901)	1.628 (0.874)	3.083 (1.566)	2.056 (0.679)	6.108 (2.719)	3.735 (1.626)
RM: n_1	1.657 (1.179)	1.683 (0.921)	4.390 (3.787)	2.502 (0.805)	10.544 (8.494)	5.606 (2.109)
RM: n_2	1.459 (0.907)	1.623 (0.877)	4.361 (3.776)	2.251 (0.547)	8.364 (5.477)	4.721 (3.263)
PS($\gamma = 1$): n_1	1.452 (0.937)	1.666 (0.898)	3.903 (3.091)	2.360 (1.117)	9.438 (7.153)	4.752 (2.900)
PS($\gamma = 3$): n_1	1.453 (1.026)	1.663 (0.948)	4.761 (4.173)	2.430 (1.451)	7.452 (5.648)	4.795 (2.326)
uniform dist.	5.545		11.090		22.181	

The neighborhood system introduced on $\mathcal{Y} = \{0, \ldots, 9\}$ is defined as

$$n_k(y) = \{z \in \mathcal{Y} \,|\, |y - z| \le k\},$$

where $|y - z|$ for labels is computed as the difference of integers. We used the neighborhood system with $k = 1$ and 2. We examined the MLE, PL, RM, local PS score, composite likelihood (CL) and modified composite likelihood (mCL) in (15). The CL with the above neighborhood system does not leads to proper score on $\mathcal{Y}$. In order to guarantee the consistency of the estimator, we need to use the mCL instead of the CL.

The signal to noise ratio of the original data is very low and the estimator of the conditional probability tends to overfit to training data. In this experiment, we added artificial noise to the data. Concretely, 10% or 20% samples were picked up, and their labels were replaced according to the uniform distribution on $\mathcal{Y}$. The training sample size was set to $n = 2000$ and $n = 4000$, and the rest were used as test samples. We randomly split the dataset into training and test sets, and repeated the experiments 100 times.

The results are presented in Figures 2 and 3. Overall, the MLE achieved the highest accuracy. The estimators with the wider neighborhood n_2 were superior to those with n_1. Under the test error of the label, the local PS score with n_2 (PS:2) provided the better classifier than the other local scores. The negative log-loss of the PL, CL and mCL with n_2 (PL:2, CL:2, and mCL:2) was smaller than the other local homogeneous scores. This is because, the empirical loss to be minimized in the learning is close to the negative log-loss for the evaluation. In this experiment, the CL and mCL showed almost the same accuracy. The efficiency of the local PS score was less than that of the PL, CL and mCL.

Table 2: The negative log-loss of each estimator for handwritten data.

	dim $x = 8$		dim $x = 16$		dim $x = 32$	
	$n = 1000$	$n = 3000$	$n = 1000$	$n = 3000$	$n = 1000$	$n = 3000$
MLE	2.986 (0.029)	2.954 (0.021)	–	–	–	–
PL: n_1	3.009 (0.175)	2.959 (0.024)	4.703 (0.568)	4.467 (0.057)	25.938 (13.716)	21.163 (7.706)
PL: n_2	2.997 (0.035)	2.963 (0.027)	4.620 (0.110)	4.473 (0.062)	19.937 (11.421)	13.104 (6.164)
RM: n_1	3.018 (0.191)	2.969 (0.025)	4.969 (1.012)	4.505 (0.067)	23.640 (13.171)	18.503 (7.623)
RM: n_2	2.999 (0.038)	2.963 (0.026)	4.625 (0.123)	4.475 (0.057)	21.051 (11.180)	14.188 (6.026)
PS($\gamma = 1$): n_1	3.062 (0.127)	2.998 (0.069)	4.665 (0.116)	4.488 (0.061)	19.091 (9.907)	13.296 (5.443)
PS($\gamma = 3$): n_1	3.046 (0.198)	3.013 (0.067)	6.345 (1.070)	5.487 (0.929)	84.557 (23.407)	63.777 (21.563)

8   Conclusion

In this paper, we study strict properness of local homogeneous scores through the coincidence axiom for composite local Bregman divergences on discrete sample space. The connectedness of the neighborhood graph is important to derive strictly proper local homogeneous scores. The consistency of some existing scores and newly proposed ones is investigated. Some numerical experiments are conducted to investigate how neighborhood systems affect the estimation accuracy.

There are several directions to be further explored in the future. It is of great interest to study the efficiency and robustness of local scores. More precisely, how the estimation accuracy and robustness of the estimator relates to the neighborhood system, local potential and statistical models. In the numerical experiments in Section 7.2, the efficiency of the local PS score was inferior to that of the pseudo-likelihood and composite likelihood. However, the local PS score may provide a robust estimator as well as the standard PS score used in robust statistics [12] instead of providing an efficient estimator. Another important research direction is to improve the computational efficiency. Though the computation cost of local scores are much smaller than the MLE, the standard optimization solver is still insufficient for the statistical analysis of extremely massive data using huge models. Developing efficient optimization algorithms is needed in the big-data era.

A   Derivation of (8)

Lemma 2. *The equality*

$$\sum_{x \in \mathcal{Y}} \sum_{y \in b(x)} A_{xy} = \sum_{x \in \mathcal{Y}} \sum_{y \in b(x)} A_{yx}$$

holds for any $(A_{xy})_{x,y \in \mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$.

Proof. Since $y \in b_x$ is equivalent with $x \in b_y$, we have

$$\sum_{x \in \mathcal{Y}} \sum_{y \in b(x)} A_{xy} = \sum_{y \in \mathcal{Y}} \sum_{x \in b(y)} A_{yx} = \sum_{x,y \in \mathcal{Y}} \mathbf{1}[x \in b(y)] A_{yx} = \sum_{x,y \in \mathcal{Y}} \mathbf{1}[y \in b(x)] A_{yx} = \sum_{x \in \mathcal{Y}} \sum_{y \in b(x)} A_{yx}.$$

$\square$
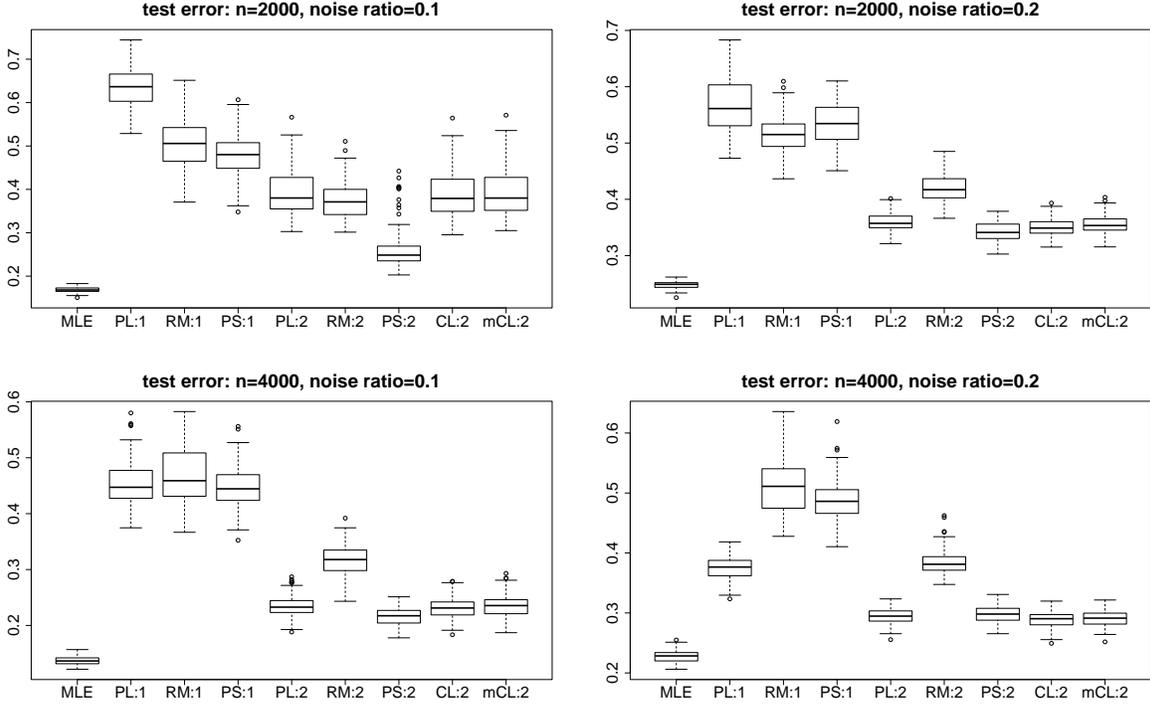
Figure 2: Test errors of the maximum likelihood estimator (MLE), pseudo-likelihood (PL:k), ratio-matching (RM:k), local pseudo-spherical score (PS:k), composite likelihood (CL:k), and modified composite likelihood (mCL:k) with neighborhood systems $n_k, k = 1, 2$. Top panels: $n = 2000$ and the noise ratio 0.1 and 0.2. Bottom panels: $n = 4000$ and the noise ratio 0.1 and 0.2.

The formula (8) is obtained by using Lemma 2 as follows.

$$
\begin{aligned}
&D_\phi(f, g) \\
&= \sum_{y \in \mathcal{Y}} f_y S(y, g) + \phi(f) \\
&= \sum_{y \in \mathcal{Y}} f_y \bigg\{ -\phi_y(g_{b(y)}/g_y) + \sum_{z \in b(y)} \frac{g_z}{g_y} \partial_z \phi_y(g_{b(y)}/g_y) - \sum_{z \in b(y)} \partial_y \phi_z(g_{b(z)}/g_z) \bigg\} + \sum_{y \in \mathcal{Y}} f_y \phi_y(f_{b(y)}/f_y) \\
&= \sum_{y \in \mathcal{Y}} f_y \bigg\{ -\phi_y(g_{b(y)}/g_y) + \phi_y(f_{b(y)}/f_y) + \sum_{z \in b(y)} \frac{g_z}{g_y} \partial_z \phi_y(g_{b(y)}/g_y) \bigg\} - \sum_{y \in \mathcal{Y}} f_y \sum_{z \in b(y)} \partial_y \phi_z(g_{b(z)}/g_z) \\
&= \sum_{y \in \mathcal{Y}} f_y \bigg\{ -\phi_y(g_{b(y)}/g_y) + \phi_y(f_{b(y)}/f_y) + \sum_{z \in b(y)} \frac{g_z}{g_y} \partial_z \phi_y(g_{b(y)}/g_y) \bigg\} - \sum_{y \in \mathcal{Y}} f_y \sum_{z \in b(y)} \frac{f_z}{f_y} \partial_z \phi_y(g_{b(y)}/g_y) \\
&= \sum_{y \in \mathcal{Y}} f_y \bigg\{ \phi_y(f_{b(y)}/f_y) - \phi_y(g_{b(y)}/g_y) - \sum_{z \in b(y)} \partial_z \phi_y(g_{b(y)}/g_y) \bigg(\frac{f_z}{f_y} - \frac{g_z}{g_y} \bigg) \bigg\} \\
&= \sum_{y \in \mathcal{Y}} f_y D_{\phi_y}(f_{b(y)}/f_y, g_{b(y)}/g_y).
\end{aligned}
$$
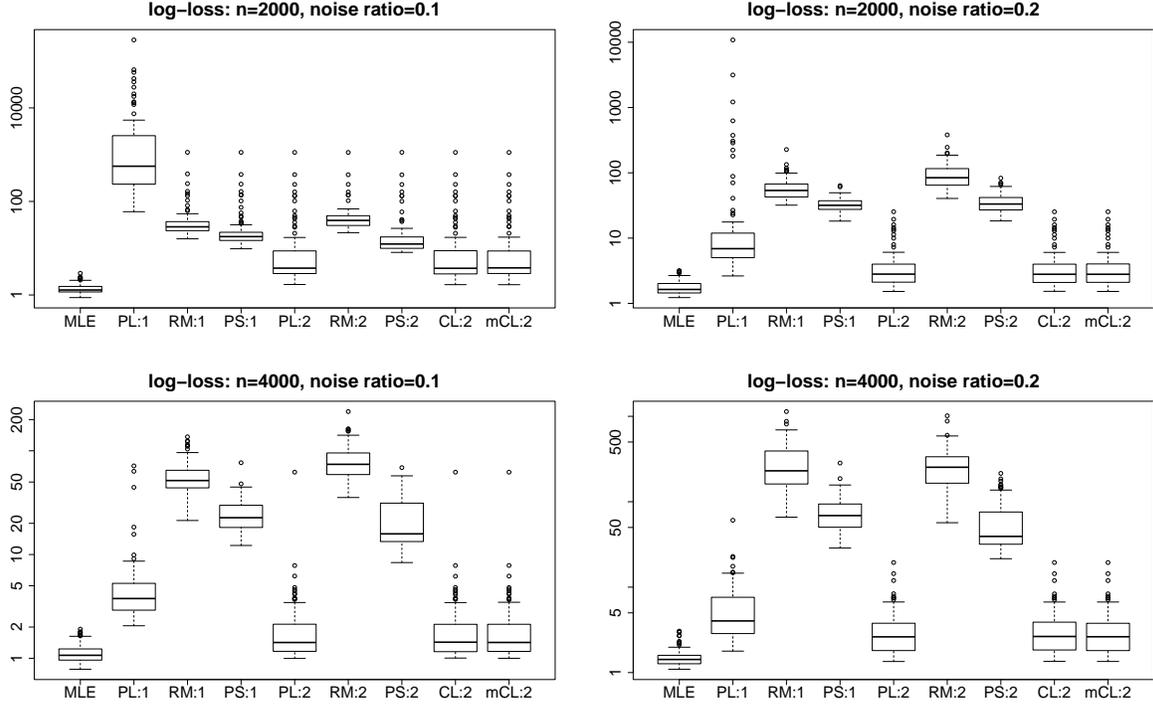
Figure 3: Negative log-loss of the maximum likelihood estimator (MLE), pseudo-likelihood (PL:k), ratio-matching (RM:k), local pseudo-spherical score (PS:k), composite likelihood (CL:k), and modified composite likelihood (mCL:k) with neighborhood systems $n_k, k = 1, 2$. Top panels: $n = 2000$ and the noise ratio 0.1 and 0.2. Bottom panels: $n = 4000$ and the noise ratio 0.1 and 0.2.

B Score of Composite Local Bregman Divergence

Let $\{\phi_y\}_{y \in \mathcal{Y}_0}$ be a set of local potentials. The associated composite local Bregman divergence is given as follows.

$$
\begin{aligned}
& D_\phi(p, q) \\
&= \sum_{y \in \mathcal{Y}_0} p_y D_{\phi_y}(p_{b(y)}/p_y, q_{b(y)}/q_y) \\
&= \sum_{y \in \mathcal{Y}_0} p_y \left\{ \phi_y(p_{b(y)}/p_y) - \phi_y(q_{b(y)}/q_y) - \sum_{z \in b(y)} \partial_z \phi_y(q_{b(y)}/q_y) \left(\frac{p_z}{p_y} - \frac{q_z}{q_y} \right) \right\} \\
&= \sum_{y \in \mathcal{Y}_0} p_y \phi_y(p_{b(y)}/p_y) - \sum_{y \in \mathcal{Y}_0} p_y \phi_x(q_{b(y)}/q_y) - \sum_{\substack{y \in \mathcal{Y}_0 \\ z \in b(y)}} \partial_z \phi_y(q_{b(y)}/q_y) \left(p_z - p_y \frac{q_z}{q_y} \right) \\
&= - \sum_{y \in \mathcal{Y}_0} p_y \phi_y(q_{b(y)}/q_y) - \sum_{\substack{y \in \mathcal{Y}_0 \\ z \in b(y)}} p_z \partial_z \phi_y(q_{b(y)}/q_y) + \sum_{\substack{y \in \mathcal{Y}_0 \\ z \in b(y)}} p_y \frac{q_z}{q_y} \partial_z \phi_y(q_{b(y)}/q_y) + \phi(p).
\end{aligned}
$$

Therefore, the expected score is given by

$$S(p,q) = -\sum_{y \in \mathcal{Y}_0} p_y \phi_y(q_{b(y)}/q_y) - \sum_{\substack{y \in \mathcal{Y}_0 \\ z \in b(y)}} p_z \partial_z \phi_y(q_{b(y)}/q_y) + \sum_{\substack{y \in \mathcal{Y}_0 \\ z \in b(y)}} p_y \frac{q_z}{q_y} \partial_z \phi_y(q_{b(y)}/q_y).$$

Then, we obtain

$$
\begin{aligned}
&S(p,q) \\
&= -\sum_{y \in \mathcal{Y}} p_y \mathbf{1}[y \in \mathcal{Y}_0] \phi_y(q_{b(y)}/q_y) - \sum_{y \in \mathcal{Y}} \sum_{z \in b(y)} \mathbf{1}[y \in \mathcal{Y}_0] p_z \partial_z \phi_y(q_{b(y)}/q_y) \\
&\quad + \sum_{y \in \mathcal{Y}} p_y \mathbf{1}[y \in \mathcal{Y}_0] \sum_{z \in b(y)} \frac{q_z}{q_y} \partial_z \phi_y(q_{b(y)}/q_y) \\
&= \sum_{y \in \mathcal{Y}} p_y \left\{ \mathbf{1}[y \in \mathcal{Y}_0] \left(\sum_{z \in b(y)} \frac{q_z}{q_y} \partial_z \phi_y(q_{b(y)}/q_y) - \phi_y(q_{b(y)}/q_y) \right) - \sum_{z \in b(y) \cap \mathcal{Y}_0} \partial_y \phi_z(q_{b(z)}/q_z) \right\}.
\end{aligned}
$$

Therefore, we have

$$S(y,q) = \mathbf{1}[y \in \mathcal{Y}_0] \left(\sum_{z \in b(y)} \frac{q_z}{q_y} \partial_z \phi_y(q_{b(y)}/q_y) - \phi_y(q_{b(y)}/q_y) \right) - \sum_{z \in b(y) \cap \mathcal{Y}_0} \partial_y \phi_z(q_{b(z)}/q_z).$$

When the local potential is expressed as the additive form $\phi_y(f) = \sum_{z \in b(y)} \phi(f_z)$, we have

$$
\begin{aligned}
S(y,q) &= \mathbf{1}[y \in \mathcal{Y}_0] \cdot \left(\sum_{z \in b(y)} \frac{q_z}{q_y} \partial_z \sum_{w \in b(y)} \phi(q_w/q_y) - \sum_{z \in b(y)} \phi(q_z/q_y) \right) - \sum_{z \in b(y) \cap \mathcal{Y}_0} \partial_y \sum_{w \in b(z)} \phi(q_w/q_z) \\
&= \mathbf{1}[y \in \mathcal{Y}_0] \cdot \left(\sum_{z \in b(y)} \frac{q_z}{q_y} \phi'(q_z/q_y) - \sum_{w \in b(y)} \phi(q_w/q_y) \right) - \sum_{z \in b(y) \cap \mathcal{Y}_0} \phi'(q_y/q_z) \\
&= \sum_{z \in b(y)} \left\{ \mathbf{1}[y \in \mathcal{Y}_0] \left(\frac{q_z}{q_y} \phi'(q_z/q_y) - \phi(q_z/q_y) \right) - \mathbf{1}[z \in \mathcal{Y}_0] \phi'(q_y/q_z) \right\}.
\end{aligned}
$$

For $\mathcal{Y}_0 = \mathcal{Y}$, we have

$$S(y,q) = \sum_{z \in b(y)} \psi(q_z/q_y),$$

where $\psi(r) = r\phi'(r) - \phi(r) - \phi'(1/r)$, $r \in \mathbb{R}_{++}$.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[2] A. U. Asuncion, Q. Liu, A. T. Ihler, and P. Smyth. Learning with blocks: Composite likelihood and contrastive divergence. In Yee W. Teh and D. M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, volume 9, pages 33–40, 2010.

[3] A. Banerjee, S. Merugu, and J. Ghosh I. S. Dhillon. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[4] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[5] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[7] J. K. Bradley and C. Guestrin. Sample complexity of composite likelihood. In *In Artificial Intelligence and Statistics (AISTATS)*, La Palma, Canary Islands, April 2012.

[8] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex prog ramming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[9] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and Bregman distances. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 158–169, 2000.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[11] A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *Annals of Statistics*, 40:593–608, 2012.

[12] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.*, 99(9):2053–2081, 2008.

[13] C. J. Geyer. Markov chain monte carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

[14] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[15] M. Gutmann and J. Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI2011)*, pages 283–290, 2011.

[16] A. D. Hendrickson and R. J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42:19161921, 1971.

[17] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Compututation*, 14(8):1771–1800, 2002.

[18] A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531, 2007.

[19] A. Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51:2499–2512, 2007.

[20] T. Kanamori and H. Fujisawa. Affine invariant divergences associated with composite scores and its applications. *Bernoulli*, 20(4):2278–2304, 2014.

[21] P. Liang and M. I. Jordan . An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 584–591, New York, NY, USA, 2008. ACM.

[22] M. Lichman. UCI machine learning repository, 2013.

[23] B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1), 1988.

[24] S. Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 359–366, Arlington, Virginia, United States, 2009. AUAI Press.

[25] B. Marlin and N. de Freitas. Asymptotic efficiency of deterministic estimators for discrete energy-based models: Ratio matching and pseudolikelihood. In *Uncertainty in Artificial Intelligence (UAI)*, Corvallis, Oregon, 2011. AUAI Press.

[26] J. McCarthy. Measures of the value of information. *Proc Natl Acad Sci U S A.*, 42(9):654–655, 1956.

[27] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, 2004.

[28] M. Parry, A. P. Dawid, and S. Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40:561–592, 2012.

[29] M. Pihlaja, M. Gutmann, and A. Hyvärinen. A Family of Computationally Efficient and Simple Estimators for Unnormalized Statistical Models. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 442–449, Corvallis, Oregon, 2010. AUAI Press.

[30] R. Salakhutdinov. Learning and evaluating boltzmann machines. Technical report, UTML TR 2008-002, Dept. of Computer Science, University of Toronto, 2008.