MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Geometric Characterization of Local Estimator on Manifold

Taiji SUZUKI

METR 2009–50 November 2009

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

Geometric Characterization of Local Estimator on Manifold

Taiji Suzuki
Department of Mathematical Informatics,
Graduate School of Information Science and Technology,
The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
t-suzuki@mist.i.u-tokyo.ac.jp

November 16, 2009

Abstract

We investigate a problem of estimating distributions that are aligned on a manifold Θ embedded in Euclidean space. To estimate the distributions, we utilize *local data aggregation* that aggregates samples around a target point on the manifold Θ. We investigate a weighted maximal likelihood estimator on a locally aggregated data where the weight is given by a kernel function defined by the distance of the Euclidean space where Θ is embedded. We show that the asymptotic risk of the estimator is characterized by geometric quantities such as Laplacian and Riemannian metric. We also give the optimal kernel width that balances bias and variance trade off induced by the kernel width.

1 Introduction

In this article we consider a problem of estimating probability distributions equipped on each point of a manifold Θ. Here Θ is embedded in higher dimensional Euclidean space $\mathbf{R}^m$ as an ℓ-dimensional compact smooth regular submanifold with boundary (possibly $\partial\Theta = \emptyset$ where $\partial\Theta$ is the boundary of Θ). Instead of considering an abstract manifold Θ, it might be helpful to consider a setting where $\Theta = [0,1]^\ell$ and there exists a smooth embedding map $\varpi : \Theta \to \mathbf{R}^m$, and identify $\varpi(\Theta)$ as Θ. We equip Θ with a metric $\tilde{g} = (\tilde{g}_{ij})$ which is induced from the Euclidean metric on $\mathbf{R}^m$ so that Θ has a structure of Riemannian manifold with boundary. $\tilde{g}^{ij}$ denotes the (i,j)-component of the inverse matrix of $(\tilde{g}_{ij})$.

At each $\theta \in \Theta$, a probability density $q_\theta(X) = q(X|\theta)$ is equipped with. The task we consider here is to estimate q_θ from the sample observations $D^N = \{(\theta_1, x_1), \dots, (\theta_N, x_N)\}$ where $(\theta_n, x_n) \in \Theta \times \mathcal{X}$ is distributed independent identically from the following model:

$$\theta_n \sim \pi(\theta)$$
$$x_n \sim q_{\theta_n}(X).$$

Here $\pi(\theta)$ is a probability density on Θ with respect to the volume element $\sqrt{|\tilde{g}|}\mathrm{d}\theta$, i.e., $\int_\Theta \pi(\theta)\sqrt{|\tilde{g}|}\mathrm{d}\theta = 1$. We suppose both $\pi(\theta)$ and q_θ are unknown. A key geometric quantity is the Laplacian operator on $(\Theta, \tilde{g})$ defined by

$$\Delta_\Theta f(\theta) := \frac{1}{\sqrt{|\tilde{g}|}} \partial_i \left(\sqrt{|\tilde{g}|} \tilde{g}^{ij} \partial_j f(\theta) \right)$$

where ∂_i is a partial derivative with respect to θ_i a local coordinate of Θ. As seen later, the Laplacian operator gives a geometric interpretation of the asymptotic risk of estimators on locally aggregated data.

We assume all q_θ $(\theta \in \Theta)$ are contained in a parametric model $\mathcal{M}$:

$$\mathcal{M} = \{p_\mu(X) = p(X|\mu) \mid \mu \in \mathcal{U}\},$$

where $\mathcal{U}$ is a d-dimensional C^∞ manifold. We use the same notation for the local coordinate of $\mathcal{U}$ with $\mathcal{U}$ itself, thus we will deal with $\mathcal{U}$ as if $\mathbf{R}^d$. We also write a partial derivative with respect to μ_i as $\frac{\partial}{\partial \mu_i} = \bar{\partial}_i$. We assume there exists a smooth mapping $\iota : \Theta \to \mathcal{U}$ such that

$$q_\theta(X) = p_{\iota(\theta)}(X) = p(X|\iota(\theta)) \in \mathcal{M}. \tag{1}$$

From now on we fix θ which is an interior point of Θ and consider to estimate q_θ by *local data aggregation*. Corresponding to θ we define μ as

$$\mu = \mu(\theta) := \iota(\theta), \quad \theta \in \mathrm{int}(\Theta).$$

Thus (1) is equivalent to

$$q_\theta(X) = p_\mu(X) = p(X|\mu).$$

We consider an estimator which maximizes weighted log-likelihood on aggregated data around θ. To do this, we introduce window width $h_N \in \mathbf{R}^+$, which depends on the sample size N, and a weight kernel $K : \mathbf{R}^+ \to \mathbf{R}^+$[*1]. Define a scaled weight kernel as

$$K_h(\|y\|) = \frac{1}{h^\ell} K\left(\frac{\|y\|^2}{h^2}\right).$$

To estimate q_θ we employ maximum likelihood estimator $\hat{\mu}$ for weighted log-likelihood:

$$\hat{\mu} = \hat{\mu}(\theta) := \operatorname*{arg\,max}_{\mu' \in \mathcal{U}} \frac{1}{N} \sum_{n=1}^{N} K_{h_N}(\|\theta_n - \theta\|) \log p(x_n|\mu'), \tag{2}$$

where $\|\theta' - \theta\|$ is Euclidean distance between θ' and θ in $\mathbf{R}^m$. We consider a class of weight kernels $K_{h_N}(x)$ that decay exponentially as $x \to \infty$ (Assumption 1). Therefore only information around θ contributes to the estimation of $q(x|\theta)$. In other words, the weighted maximum likelihood is performed on samples locally aggregated around θ. We say aggregating samples around the target point θ as *local data aggregation.*

The main purpose of this paper is to investigate properties of $\hat{\mu}$ under a condition that the window width goes to 0 ($h_N \searrow 0$), and give geometric interpretations to the results. We employ the KL-divergence as a risk measure. An important point is that the prediction performance is controlled by the window width h_N, so the main result will be presented in the context of the optimal h_N that balances bias and variance trade off induced by the window width h_N.

[*1] $\mathbf{R}^+ := [0, \infty)$

Our analysis is closely related to the work of Eguchi, Kim, and Park (2003) which investigated a local regression problem on exponential families, i.e., a problem to estimate $\mu = \mu(\theta)$ where $\Theta = \mathbf{R}^m$ and $\mathcal{M}$ is an exponential family in our terminology. Our setting is more general than that of Eguchi et al. (2003) in the sense that Θ is generalized to a manifold, the model $\mathcal{M}$ is not restricted to an exponential family and a geometric interpretation will be given. On the other hand, they also (locally) model the "regression function" $\mu(\theta)$ by a parametric model while our analysis deals with the pointwise estimation of $\mu(\theta)$, thus their analysis is more general than ours in that aspect. In that direction, Tibshirani and Hastie (1987) also considered local regression on some exponential families, and Yu and Jones (2004) dealt with a regression problem where $\mathcal{M}$ is a class of normal distributions.

Estimation by local data aggregation is closely related to local likelihood density estimation which has been studied by many authors (Copas, 1995; Hjort & Jones, 1996; Loader, 1996; Eguchi & Copas, 1998; Park, Kim, & Jones, 2002). Local likelihood density estimation is a semiparametric estimation method that combines nonparametric approach and parametric one to density estimation in such a way that it fits a parametric model to the sample density locally around a certain target point. If the parametric model offers a good representation of the underlying distribution, it is efficient with large window width, otherwise, it flexibly fit the density with small window width as usual nonparametric density estimation method does.

In Section 2, we present some assumptions for our analysis and prepare basic lemmas that are needed for the main result. In Section 3, we show the main results concerning the optimal window width and the asymptotic risk under the optimal window width, and give their geometric interpretation. In Section 4, we derive an asymptotic expansion of geometric quantities that appear in the asymptotic risk of the estimator.

2 Preliminaries

To analyze the behavior of $\hat{\mu}$, we put some assumptions on K and show a key theorem that is useful for our analysis. We impose the following assumptions on the kernel function K that corresponds to Assumption 20 and 21 of Hein, Audibert, and Luxburg (2007).

The assumption for the kernel K is as follows.

Assumption 1

1. *$K : \mathbf{R}^+ \to \mathbf{R}^+$ is measurable, non-negative and non-increasing on $\mathbf{R}^+$,*

2. $\mathrm{supp}(K)$*, the support of K, has its interior, and K is twice continuously differentiable on the interior of the support, that is in particular $\int_0^\infty K(x)\mathrm{d}x > 0$ and $\frac{\mathrm{d}K}{\mathrm{d}x}$ and $\frac{\mathrm{d}^2K}{\mathrm{d}x^2}$ exist and are bounded on* $\mathrm{supp}(K)$.

3. *K, $|\frac{\mathrm{d}K}{\mathrm{d}x}|$ and $|\frac{\mathrm{d}^2K}{\mathrm{d}x^2}|$ have exponential decay: there exist $\alpha, c > 0$ such that for any $t \geq 0$ in the interior of* $\mathrm{supp}(K)$, $\max\{K(t), |\frac{\mathrm{d}K}{\mathrm{d}x}(t)|, |\frac{\mathrm{d}^2K}{\mathrm{d}x^2}(t)|\} \leq ce^{-\alpha t}$.

Because of Assumption 1, the following two integrals converge:

$$C_1 := \int_{\mathbf{R}^\ell} K(\|y\|^2)\mathrm{d}y < \infty, \quad C_2 := \int_{\mathbf{R}^\ell} K(\|y\|^2)y_1^2\mathrm{d}y < \infty,$$

where y_1 is the first element of $y \in \mathbf{R}^\ell$. We also impose the following differentiability assumptions on $\pi(\theta)$ and q_θ.

Assumption 2

1. *$\pi \in C^2(\Theta)^{*2}$ and $\pi(\theta') > 0$ for all $\theta' \in \Theta$,*

2. *For all $x \in \mathcal{X}$, $q_{\theta'}(x) \in C^2(\Theta)$ as a function of θ'.*

We define an "averaged" probability density

$$\bar{q}_\theta(X) := \frac{\int_\Theta K_{h_N}(\|\theta - \theta'\|) q_{\theta'}(X) \pi(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta'}{\int_\Theta K_{h_N}(\|\theta - \theta'\|) \pi(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta'}.$$

The maximum weighted log-likelihood estimator corresponds to estimating $\bar{q}_\theta$ because the weighted log-likelihood gives an estimator of KL-divergence from $\bar{q}_\theta$ except constant multiplication and addition. Actually as $N \to \infty$, $\hat{\mu}$ converges to the "closest" point of $\mathcal{U}$ from $\bar{q}_\theta$ in probability (see Section 4).

$\bar{q}_\theta$ can be approximated by q_θ plus higher order terms as the following theorem.

Theorem 1 *Under Assumption 1 and 2, the density of the distribution of aggregated data is expressed by*

$$\bar{q}_\theta = q_\theta + \frac{C_2 {h_N}^2}{2C_1} \left[\Delta_\Theta(q_\theta) + 2\tilde{g}^{ij} (\partial_i \log \pi(\theta))(\partial_j q_\theta) \right] + o({h_N}^2).$$

□

Proof The proof utilizes Proposition 22 of Hein et al. (2007). Proposition 22 of Hein et al. (2007) and its proof indicate that the weighted average of probability density $q_{\theta'}$ can be expressed as follows:

$$\begin{aligned} \bar{r}_\theta &:= \int_\Theta K_{h_N}(\|\theta - \theta'\|) q_{\theta'} \pi(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta' \\ &= C_1 q_\theta \pi(\theta) + \frac{{h_N}^2}{2} C_2 \left(\Delta_\Theta(\pi(\theta) q_\theta) + \pi(\theta) q_\theta S(\theta) \right) + o({h_N}^2), \end{aligned} \tag{3}$$

where $S(\theta)$ is a function of θ defined by

$$S(\theta) = \frac{1}{2} \left[-R|_\theta + \frac{1}{2} \left\| \sum_i \Pi(\partial_i, \partial_i) \right\|^2 \right],$$

where R is the scalar curvature and Π is the second fundamental from of Θ. We omit detailed explanations of $S(\theta)$ because it is not related to the later discussions. See Hein et al. (2007) for details. This is proven by substituting $\pi(\theta)$, q_θ, and $\tilde{g}$ to p, f, and g in Proposition 22 of Hein et al. (2007) respectively. It should be noted that Assumption 19 of Hein et al. (2007) is satisfied because Θ is a smooth compact submanifold (see, remarks following after Assumption 19 of Hein et al. (2007)). In particular, self-approaching does not occur, namely there is no two distinct points which are far away in Θ with respect

[*2] $C^k(\mathcal{X})$ denotes the set of functions with k continuous partial derivatives on $\mathcal{X}$.

to the geodesic distance defined by $\tilde{g}$ but too close in $\mathbf{R}^m$ with respect to the Euclidean distance. Assumption 20 of Hein et al. (2007) assumes stronger condition of K than Assumption 1 of this article, namely K is twice continuously differentiable on whole $\mathbf{R}^+$. However that difference does not induce any difficulty to derive the required consequence. Therefore the result of Proposition 22 of Hein et al. (2007) is still valid under our setting. In addition, Proposition 22 of Hein et al. (2007) assumes $q_\theta, \pi(\theta) \in C^3(\Theta)$ as a function of θ and the direct consequence under the stronger condition indicates that the residual term appears in the asymptotic expansion (3) can be $O({h_N}^3)$ instead of $o({h_N}^2)$. However it is easy to check that by relaxing their assumption to $q_\theta, \pi(\theta) \in C^2(\Theta)$ a similar (but a little bit weak) result with the residual term $o({h_N}^2)$ is obtained.

Since $\bar{r}_\theta$ may not be a probability density, normalization is needed to obtain $\bar{q}_\theta$. Note that

$$\begin{aligned}\frac{1}{\int \bar{r}_\theta \mathrm{d}x} &= \frac{1}{C_1\pi(\theta) + \int \frac{{h_N}^2}{2} C_2 \left(\Delta_\Theta(\pi(\theta)q_\theta) + \pi(\theta)q_\theta S(\theta)\right)\mathrm{d}x + o({h_N}^2)} \\ &= \frac{1}{C_1\pi(\theta)} - \frac{{h_N}^2}{2(C_1\pi(\theta))^2} C_2 \int \Delta_\Theta(\pi(\theta)q_\theta) + \pi(\theta)q_\theta S(\theta)\mathrm{d}x + o({h_N}^2).\end{aligned}$$

Thus dividing $\bar{r}_\theta$ by normalizing constant, we have

$$\begin{aligned}\bar{q}_\theta &= \frac{\bar{r}_\theta}{\int \bar{r}_\theta \mathrm{d}x} \\ &= \left[C_1 q_\theta \pi(\theta) + \frac{{h_N}^2}{2} C_2 \left(\Delta_\Theta(\pi(\theta)q_\theta) + \pi(\theta)q_\theta S(\theta)\right) + o({h_N}^2)\right] \times \\ &\qquad \left[\frac{1}{C_1\pi(\theta)} - \frac{{h_N}^2}{2(C_1\pi(\theta))^2} C_2 \int \Delta_\Theta(\pi(\theta)q_\theta) + \pi(\theta)q_\theta S(\theta)\mathrm{d}x + o({h_N}^2)\right] \\ &= q_\theta - \frac{{h_N}^2 q_\theta}{2C_1\pi(\theta)} C_2 \int \Delta_\Theta(\pi(\theta)q_\theta) + \pi(\theta)q_\theta S(\theta)\mathrm{d}x \\ &\qquad + \frac{{h_N}^2}{2C_1\pi(\theta)} C_2 \left(\Delta_\Theta(\pi(\theta)q_\theta) + \pi(\theta)q_\theta S(\theta)\right) + o({h_N}^2) \\ &= q_\theta - \frac{{h_N}^2 q_\theta}{2C_1\pi(\theta)} C_2 \Delta_\Theta(\pi(\theta)) + \frac{{h_N}^2}{2C_1\pi(\theta)} C_2 \Delta_\Theta(\pi(\theta)q_\theta) + o({h_N}^2). \qquad (4)\end{aligned}$$

Here noticing

$$\begin{aligned}\Delta_\Theta(\pi(\theta)q_\theta) &= \frac{1}{\sqrt{|\tilde{g}|}} \partial_i \left(\sqrt{|\tilde{g}|}\tilde{g}^{ij} \partial_j (\pi(\theta)q_\theta)\right) \\ &= \frac{1}{\sqrt{|\tilde{g}|}} \partial_i \left(\sqrt{|\tilde{g}|}\tilde{g}^{ij} \left((\partial_j \pi(\theta))q_\theta + \pi(\theta)(\partial_j q_\theta)\right)\right) \\ &= \Delta_\Theta(\pi(\theta))q_\theta + \pi(\theta)\Delta_\Theta(q_\theta) + 2\tilde{g}^{ij}(\partial_i \pi(\theta))(\partial_j q_\theta),\end{aligned}$$

we see that the RHS of (4) is equivalent to

$$\begin{aligned}&q_\theta + \frac{C_2 {h_N}^2}{2C_1\pi(\theta)} \left[\pi(\theta)\Delta_\Theta(q_\theta) + 2\tilde{g}^{ij}(\partial_i \pi(\theta))(\partial_j q_\theta)\right] + o({h_N}^2) \\ =&q_\theta + \frac{C_2 {h_N}^2}{2C_1} \left[\Delta_\Theta(q_\theta) + 2\tilde{g}^{ij}(\partial_i \log \pi(\theta))(\partial_j q_\theta)\right] + o({h_N}^2).\end{aligned}$$

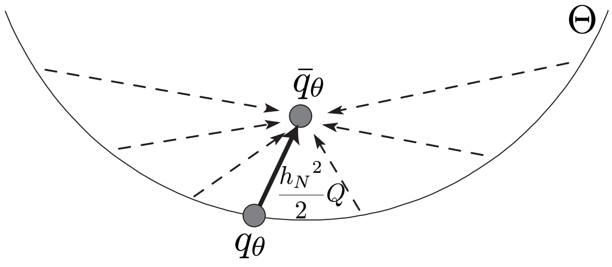

Figure 1: Relation between $\bar{q}_\theta$ and q_θ.

□

We define

$$Q := \frac{C_2}{C_1}\left[\Delta_\Theta(q_\theta) + 2\tilde{g}^{ij}(\partial_i \log \pi(\theta))(\partial_j q_\theta)\right],$$

$$T := \bar{q}_\theta - p_\mu - \frac{h_N{}^2}{2}Q.$$

Then by Theorem 1, $T = o(h_N{}^2)$. Moreover note that $\int Q\mathrm{d}x = 0$, thus $\int \bar{q}_\theta \mathrm{d}x = \int p_\mu \mathrm{d}x = 1$ gives $\int T\mathrm{d}x = 0$.

Remark 1 *Q can be characterized by the weighted Laplace-Beltrami operator. The t-th weighted Laplace-Beltrami operator with respect to density $\pi(\theta)$ is defined by*

$$\Delta_t := \Delta_\Theta + \frac{t}{\pi(\theta)}\tilde{g}^{ij}\partial_i(\pi(\theta))\partial_j = \frac{1}{\pi(\theta)^t}\mathrm{div}\left(\pi(\theta)^t \mathrm{grad}\right),$$

where div *and* grad *are divergence and gradient respectively corresponding to the metric $\tilde{g}$, thus Q is obtained by operating the 2nd weighted Laplace-Beltrami operator to q_θ except constant multiplication because*

$$\begin{aligned}\Delta_\Theta(q_\theta) + 2\tilde{g}^{ij}(\partial_i \log \pi(\theta))(\partial_j q_\theta) &= \frac{1}{\pi(\theta)^2}\frac{1}{\sqrt{|\tilde{g}|}}\partial_i\left(\sqrt{|\tilde{g}|}\tilde{g}^{ij}\pi(\theta)^2\partial_j(q_\theta)\right) \\ &= \frac{1}{\pi(\theta)^2}\mathrm{div}\left(\pi(\theta)^2\mathrm{grad}(q_\theta)\right) = \Delta_2(q_\theta).\end{aligned}$$

Details of the weighted Laplacian can be found in (Grigor'yan, 2006; Hein et al., 2007).

It should be noted that the Laplacian has an interpretation that it expresses the difference between the value of an argument function at a given point (say θ) and the average value taken over the neighborhood of θ. This interpretation matches the statement of Theorem 1 because Theorem 1 says the difference between the density at θ and the averaged density around θ is expressed by the weighted Laplacian (Figure 1).

Before stating the main results of this article, we define some notations. Let $\bar{\mu}$ be the "closest" point in $\mathcal{U}$ to $\bar{q}_\theta$, the distribution of locally aggregated data, and v be the difference between $\bar{\mu}$ and μ:

$$\bar{\mu} := \arg\min_{\mu'} D(\bar{q}_\theta || p_{\mu'}),$$

$$v := \bar{\mu} - \mu.$$

The Fisher metric g on the tangent space $T_\mu \mathcal{U}$ at μ is defined by

$$g_{ij} := -\int p_\mu \bar{\partial}_i \bar{\partial}_j \log p_\mu \mathrm{d}x = \int p_\mu \bar{\partial}_i \log p_\mu \bar{\partial}_j \log p_\mu \mathrm{d}x.$$

g^{ij} denotes the (i,j)-component of the inverse matrix of (g_{ij}) (not $g^{ij} = g_{kl}\tilde{g}^{ik}\tilde{g}^{lj}$). We define s_{ij}, a variant of Fisher metric, as

$$s_{ij} = \int -\bar{q}_\theta \bar{\partial}_i \bar{\partial}_j \log p_{\bar{\mu}} \mathrm{d}x.$$

Also we denote by s^{ij} the (i,j)-component of the inverse matrix of (s_{ij}).

3 Optimal window width and asymptotic risk

In this section, we state the main result of this article. The result gives the asymptotic behavior of KL-divergence between the true q_θ and the estimated one $p_{\hat{\mu}}$. The asymptotic risk is expressed by sum of risks induced by bias and variance. The optimal window width that balances the bias and variance trade off will be given. To state the main theorem we prepare two lemmas (Lemma 1, 2). The first one is about the bias $v = \bar{\mu} - \mu$ and the risk difference induced by the bias. The second lemma gives an asymptotic expansion of the risk of the estimator $\hat{\mu}$. The proofs of the lemmas are given in Appendix A.

Lemma 1 *Under Assumption 1 and 2, the bias v has the following asymptotic expansion:*

$$v^i = \frac{{h_N}^2}{2} g^{ik} \int Q \bar{\partial}_k \log p_\mu \mathrm{d}x + o({h_N}^2).$$

The KL-divergence between $\bar{q}_\theta$ and $p_{\bar{\mu}}$ can be expanded as

$$D(\bar{q}_\theta || p_{\bar{\mu}}) = -\frac{{h_N}^4}{8} g^{ij} \int Q \bar{\partial}_i \log p_\mu \mathrm{d}x \int Q \bar{\partial}_j \log p_\mu \mathrm{d}x + \frac{{h_N}^4}{8} \int \frac{Q^2}{p_\mu} \mathrm{d}x + o({h_N}^4).$$

Remark 2 *We can show that*

$$-g^{ij} \int Q \bar{\partial}_i \log p_\mu \mathrm{d}x \int Q \bar{\partial}_j \log p_\mu \mathrm{d}x + \int \frac{Q^2}{p_\mu} \mathrm{d}x \geq 0.$$

The proof is as follows. Decompose $\frac{Q}{p_\mu}$ into the part parallel to $\{\bar{\partial}_l \log p_\mu\}_{l=1}^d$ and the one perpendicular to those:

$$\frac{Q}{p_\mu} = \sum_{l=1}^d c^l \bar{\partial}_l \log p_\mu + r,$$

where $\int r(\bar{\partial}_l \log p_\mu) p_\mu \mathrm{d}x = 0$ for all $1 \leq l \leq d$. Then

$$\begin{aligned} g^{ij} \int Q \bar{\partial}_i \log p_\mu \mathrm{d}x \int Q \bar{\partial}_j \log p_\mu \mathrm{d}x &= g^{ij} c^k g_{ki} c^{k'} g_{k'j} = c^i c^j g_{ij} \\ = \int \left(\frac{Q}{p_\mu} - r \right)^2 p_\mu \mathrm{d}x = \int \frac{Q^2}{p_\mu} - 2Qr + r^2 p_\mu \mathrm{d}x &= \int \frac{Q^2}{p_\mu} - r^2 p_\mu \mathrm{d}x \leq \int \frac{Q^2}{p_\mu} \mathrm{d}x. \end{aligned}$$

Lemma 1 says that the primary term of the bias v^i is expressed by Q and the inner product of the basis vectors of the tangent space at μ. Thus if Q is orthogonal to the tangent space, the order of the bias v^i becomes smaller than ${h_N}^2$.

The following lemma concerns an asymptotic expansion of the risk of the estimator $\hat{\mu}$.

Lemma 2 *Under Assumption 1 and 2, the risk of $\hat{\mu}$ is decomposed as follows:*

$$\begin{aligned} & D(p_\mu || p_{\hat{\mu}}) \\ & = D(\bar{q}_\theta || p_{\bar{\mu}}) + \frac{1}{2}(\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j s_{ij} - \frac{{h_N}^2}{2} \int Q \left(\log \bar{q}_\theta - \log p_{\hat{\mu}}\right) \mathrm{d}x \\ & \quad + \frac{{h_N}^4}{8} \int \frac{Q^2}{\bar{q}_\theta} \mathrm{d}x - \int T \log \frac{\bar{q}_\theta}{p_{\hat{\mu}}} \mathrm{d}x + o({h_N}^4) + o_p(\|\hat{\mu} - \bar{\mu}\|^2). \end{aligned} \tag{5}$$

Combining the two lemmas (Lemma 1, 2) we obtain the following theorem.

Theorem 2 *The KL-divergence between p_μ and $p_{\hat{\mu}}$ has the following asymptotic property:*

$$\begin{aligned} & E_{D^N}[D(p_\mu || p_{\hat{\mu}})] \\ & = \frac{{h_N}^4}{8} \left(\int Q \bar{\partial}_i \log p_\mu \mathrm{d}x \int Q \bar{\partial}_j \log p_\mu \mathrm{d}x \right) g^{ij} + \frac{1}{2} \mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j] s_{ij} \\ & \quad + o\left({h_N}^4 + \mathrm{E}_{D^N}[\|\hat{\mu} - \bar{\mu}\|^2] + \|\mathrm{E}_{D^N}[\hat{\mu} - \bar{\mu}]\|\right). \end{aligned}$$

Proof

Taking expectation of (5) with respect to sample data D^N, we obtain

$$\begin{aligned} & E_{D^N}[D(q_\theta || p_{\hat{\mu}})] \\ & = D(\bar{q}_\theta || p_{\bar{\mu}}) + \frac{1}{2} \mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j] s_{ij} - \frac{{h_N}^2}{2} \mathrm{E}_{D_N} \left[\int Q \left(\log \bar{q}_\theta - \log p_{\hat{\mu}}\right) \mathrm{d}x \right] \\ & \quad + \frac{{h_N}^4}{8} \int \frac{Q^2}{\bar{q}_\theta} \mathrm{d}x - \mathrm{E}_{D_N} \left[\int T \log \frac{\bar{q}_\theta}{p_{\hat{\mu}}} \mathrm{d}x \right] + o({h_N}^4) + o(\mathrm{E}[\|\hat{\mu} - \bar{\mu}\|^2]). \end{aligned}$$

First we evaluate the third term of the RHS of the above equation. Since

$$\begin{aligned} & \int Q \left(\log \bar{q}_\theta - \log p_{\hat{\mu}}\right) \mathrm{d}x \\ & = \int Q \left[\log q_\theta + \frac{{h_N}^2 Q}{2 q_\theta} - \left(\log p_{\bar{\mu}} + (\hat{\mu} - \bar{\mu})^i \bar{\partial}_i \log p_{\bar{\mu}} + O_p(\|\hat{\mu} - \bar{\mu}\|^2 \right) \right] \mathrm{d}x + o({h_N}^2) \\ & = \int Q \left[\log p_\mu + \frac{{h_N}^2 Q}{2 q_\theta} - \left(\log p_\mu + v^i \bar{\partial}_i \log p_\mu \right) - (\hat{\mu} - \bar{\mu})^i \bar{\partial}_i \log p_{\bar{\mu}} + O_p(\|\hat{\mu} - \bar{\mu}\|^2 \right) \right] \mathrm{d}x \\ & \quad + o({h_N}^2) \\ & = \int \frac{{h_N}^2 Q^2}{2 q_\theta} \mathrm{d}x - [v^i + (\hat{\mu} - \bar{\mu})^i] \int Q \bar{\partial}_i \log p_\mu \mathrm{d}x + o_p({h_N}^2) + O_p(\|\hat{\mu} - \bar{\mu}\|^2), \end{aligned}$$

we have

$$\frac{{h_N}^2}{2} \mathrm{E}_{D_N} \left[\int Q \left(\log \bar{q}_\theta - \log p_{\hat{\mu}}\right) \mathrm{d}x \right]$$

$$
\begin{aligned}
&= \frac{{h_N}^4}{4}\int \frac{Q^2}{q_\theta}\mathrm{d}x - \frac{{h_N}^2 v^i}{2}\int Q\bar{\partial}_i \log q_\theta \mathrm{d}x - \frac{{h_N}^2 \mathrm{E}_{D^N}[(\hat{\mu}-\bar{\mu})^i]}{2}\int Q\bar{\partial}_i \log q_\theta \mathrm{d}x \\
&\qquad + o({h_N}^4 + \mathrm{E}[\|\hat{\mu}-\bar{\mu}\|^2]) \\
&= \frac{{h_N}^4}{4}\int \frac{Q^2}{q_\theta}\mathrm{d}x - \frac{{h_N}^4}{4} g^{ij}\int Q\bar{\partial}_i \log q_\theta \mathrm{d}x \int Q\bar{\partial}_j \log q_\theta \mathrm{d}x \\
&\qquad + o({h_N}^4 + \|\mathrm{E}_{D^N}[\hat{\mu}-\bar{\mu}]\| + \mathrm{E}_{D^N}[\|\hat{\mu}-\bar{\mu}\|^2]).
\end{aligned}
$$

Moreover by Theorem 1 we have

$$
\begin{aligned}
\mathrm{E}_{D_N}\left[\int T \log \frac{\bar{q}_\theta}{p_{\hat{\mu}}}\mathrm{d}x\right] &= \mathrm{E}_{D_N}\left[(\mu-\hat{\mu})^i\right]\int T\bar{\partial}_i \log p_\mu \mathrm{d}x + o({h_N}^4) \\
&= \mathrm{E}_{D_N}\left[(\mu-\bar{\mu}+\bar{\mu}-\hat{\mu})^i\right]\int T\bar{\partial}_i \log p_\mu \mathrm{d}x + o({h_N}^4) \\
&= o({h_N}^4 + \|\mathrm{E}_{D_N}\left[\hat{\mu}-\bar{\mu}\right]\|).
\end{aligned}
$$

Therefore applying Theorem 1 to expand $D(\bar{q}_\theta||p_{\bar{\mu}})$, we have

$$
\begin{aligned}
&E_{D^N}[D(q_\theta||p_{\hat{\mu}})] \\
=& -\frac{{h_N}^4}{8} g^{ij}\int Q\bar{\partial}_i \log p_\mu \mathrm{d}x \int Q\bar{\partial}_j \log p_\mu \mathrm{d}x + \frac{{h_N}^4}{8}\int \frac{Q^2}{p_\mu}\mathrm{d}x + \frac{1}{2}\mathrm{E}_{D^N}[(\hat{\mu}-\bar{\mu})^i(\hat{\mu}-\bar{\mu})^j]s_{ij} \\
&-\frac{{h_N}^4}{4}\int \frac{Q^2}{p_\mu}\mathrm{d}x + \frac{{h_N}^4}{4} g^{ij}\int Q\bar{\partial}_i \log p_\mu \mathrm{d}x \int Q\bar{\partial}_j \log p_\mu \mathrm{d}x + \frac{{h_N}^4}{8}\int \frac{Q^2}{\bar{q}_\theta}\mathrm{d}x \\
&+ o({h_N}^4 + \|\mathrm{E}_{D_N}\left[\hat{\mu}-\bar{\mu}\right]\| + \mathrm{E}_{D^N}[\|\hat{\mu}-\bar{\mu}\|^2]) \\
=& \frac{{h_N}^4}{8} g^{ij}\int Q\bar{\partial}_i \log p_\mu \mathrm{d}x \int Q\bar{\partial}_j \log p_\mu \mathrm{d}x + \frac{1}{2}\mathrm{E}_{D^N}[(\hat{\mu}-\bar{\mu})^i(\hat{\mu}-\bar{\mu})^j]s_{ij} \\
&+ \frac{{h_N}^4}{8}\int Q^2\left(\frac{1}{\bar{q}_\theta}-\frac{1}{p_\mu}\right)\mathrm{d}x + o({h_N}^4 + \|\mathrm{E}_{D_N}\left[\hat{\mu}-\bar{\mu}\right]\| + \mathrm{E}_{D^N}[\|\hat{\mu}-\bar{\mu}\|^2]).
\end{aligned}
$$

Now noticing the relation

$$
\begin{aligned}
\frac{{h_N}^4}{8}\int Q^2\left(\frac{1}{\bar{q}_\theta}-\frac{1}{p_\mu}\right)\mathrm{d}x &= \frac{{h_N}^4}{8}\int Q^2\left(\frac{1}{p_\mu + \frac{Q{h_N}^2}{2} + O({h_N}^4)} - \frac{1}{p_\mu}\right)\mathrm{d}x \\
&= \frac{{h_N}^4}{8}\int Q^2\left(\frac{1}{p_\mu} - \frac{Q{h_N}^2}{2} + O({h_N}^4) - \frac{1}{p_\mu}\right)\mathrm{d}x \\
&= o({h_N}^4),
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&E_{D^N}[D(q_\theta||p_{\hat{\mu}})] \\
=& \frac{{h_N}^4}{8} g^{ij}\int Q\bar{\partial}_i \log p_\mu \mathrm{d}x \int Q\bar{\partial}_j \log p_\mu \mathrm{d}x + \frac{1}{2}\mathrm{E}_{D^N}[(\hat{\mu}-\bar{\mu})^i(\hat{\mu}-\bar{\mu})^j]s_{ij} \\
&+ o\left({h_N}^4 + \|\mathrm{E}_{D^N}\left[\hat{\mu}-\bar{\mu}\right]\| + \mathrm{E}_{D^N}[\|\hat{\mu}-\bar{\mu}\|^2]\right).
\end{aligned}
$$

This yields the desired formula. □

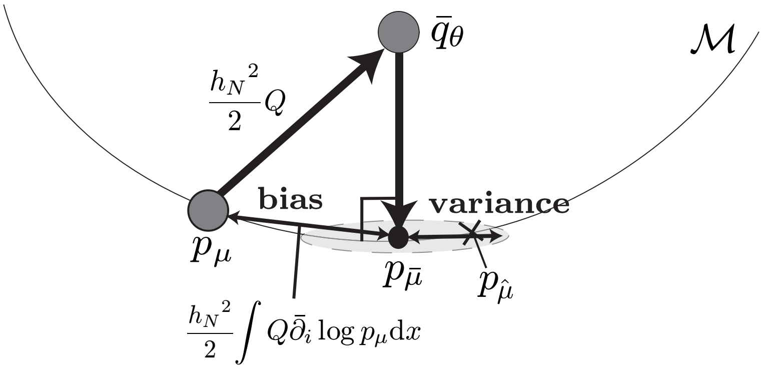

Figure 2: Geometric relation between p_μ, $p_{\bar{\mu}}$ and $p_{\hat{\mu}}$, and the bias and variance.

It can be shown that there exists a constant B depending only on the kernel K such that, for

$$N(h) = BNh^\ell \pi(\theta),$$

the maximum likelihood estimator $\hat{\mu}$ on locally aggregated data behaves as if it is estimated from $N(h_N)$ samples distributed from $\bar{q}_\theta$:

$$\frac{1}{2}\mathrm{E}_{D^N}[(\hat{\mu}-\bar{\mu})^i(\hat{\mu}-\bar{\mu})^j]s_{ij} = \frac{d}{2N(h_N)} + o(N(h_N)^{-1}), \tag{6}$$

(see next section for details). Thus the (asymptotic) optimal window width is obtained by minimizing over h_N the primary term of the asymptotic risk expressed by the following quantity:

$$(\text{asymptotic risk}) \simeq \frac{{h_N}^4}{8} g^{ij}\int Q\bar{\partial}_i \log p_\mu \mathrm{d}x \int Q\bar{\partial}_j \log p_\mu \mathrm{d}x + \frac{d}{2N(h_N)}.$$

This expression explicitly illustrates the bias and variance trade off. If we take too small h_N, the substantial sample size for estimation becomes too small and the second term, variance term, becomes large. On the other hand too large h_N induces large bias, i.e., the first term, bias term, becomes large. Thus we should choose the optimal window size that balances the bias-variance trade off. It should be noted again the bias term is controlled by the geometric quantity Q characterized by the weighted Laplace-Beltrami operator. Figure 2 illustrates how the bias and variance are induced by geometric relations between p_μ, $p_{\bar{\mu}}$ and $p_{\hat{\mu}}$.

If $\exists i$, $\int Q\bar{\partial}_i \log p_\mu \mathrm{d}x \neq 0$, the minimum of the asymptotic risk is achieved at

$$h_N^* = \left(\frac{d\ell}{NB\pi(\theta)\gamma_{ij}g^{ij}}\right)^{1/(\ell+4)},$$

where

$$\gamma_{ij} = \int Q\bar{\partial}_i \log p_\mu \mathrm{d}x \int Q\bar{\partial}_j \log p_\mu \mathrm{d}x.$$

Thus under the optimal window width the risk can be expanded as

$$E_{D^N}[D(q_\theta||p_{\hat{\mu}})] = E_{D^N}[D(p_\mu||p_{\hat{\mu}})]$$

$$= (g^{ij}\gamma_{ij})^{\frac{\ell}{4+\ell}} \left(\frac{d}{B\pi(\theta)}\right)^{\frac{4}{4+\ell}} \left(\frac{\ell^{-\frac{\ell}{\ell+4}}}{2} + \ell^{\frac{4}{\ell+4}}\right) N^{-\frac{4}{4+\ell}} + o(N^{-\frac{4}{\ell+4}}).$$

On the other hand if $\forall i$, $\int Q\bar{\partial}_i \log p_\mu \mathrm{d}x = 0$, i.e., Q is perpendicular to the tangent space spanned by $\{\bar{\partial}_i \log p_\mu\}_{i=1}^d$, the optimal asymptotic risk is

$$E_{D^N}[D(q_\theta||p_{\hat{\mu}})] = E_{D^N}[D(p_\mu||p_{\hat{\mu}})] = o(N^{-\frac{4}{4+\ell}}).$$

Remark 3 *In the context of nonparametric regression, the convergence rate $N^{-\frac{4}{4+\ell}}$ is known as mini-max rate (Györfi, Kohler, Kryżak, & Walk, 2002). Namely, if the true regression function is taken from a class of twice-differentiable functions as in our setting (Assumption 2), estimation accuracy of any estimator is at most $O_p(N^{-\frac{4}{4+\ell}})$ for a certain choice of regression function. The problem setting in this article includes regression with Gaussian noise that corresponds to a situation where $\{p_\mu \mid \mu \in \mathcal{U}\}$ is a set of Gaussian distributions with different mean and fixed variance. Therefore the local data aggregation achieves the optimal rate in a sense of mini-maxity.*

What is remaining is to prove (6). In the next section we give the proof of (6).

4 Asymptotic behavior of maximum weighted log-likelihood estimator

In this section we prove (6). Instead of considering $\hat{\mu}$ defined in (2), we consider a simpler formulation defined as follows:

$$\hat{\mu} := \mathop{\arg\max}_{\mu' \in \mathcal{U}} \frac{1}{N}\sum_{n=1}^{N} w(z_n) \log p(x_n|\mu'), \tag{7}$$

where $\{z_n\}_{n=1}^N = \{(\theta_n, x_n)\}_{n=1}^N$ are i.i.d. samples from a probability density $q(z)$, and $w(z)$ is a non-negative weight function depending on z. If we set $q(z) \leftarrow \pi(\theta)q_\theta(x)$, and $w(z_n) \leftarrow K_{h_N}(\|\theta_n - \theta\|)$, then the estimator $\hat{\mu}$ defined by (7) is reduced to that of (2).

We denote by $\bar{\mu}$ the "closest" point of $\mathcal{U}$ to $q(Z)$ with respect to the expectation of the weighted log-likelihood:

$$\bar{\mu} := \mathop{\arg\max}_{\mu' \in \mathcal{U}} \int q(Z)w(Z) \log p(X|\mu')\mathrm{d}Z.$$

We write $D^N = \{z_1, \ldots, z_N\}$. Let ϱ and ς be matrices the (i, j)-th elements of which are defined by

$$\begin{aligned}\varrho_{ij} &= \mathrm{E}_q[w(Z)^2 \bar{\partial}_i \log p(Z|\bar{\mu})\bar{\partial}_j \log p(Z|\bar{\mu})],\\ \varsigma_{ij} &= -\mathrm{E}_q[w(Z)\bar{\partial}_i\bar{\partial}_j \log p(Z|\bar{\mu})],\end{aligned}$$

where $\bar{\partial}_i$ is partial derivative with respect to μ^i. We denote by ς^{ij} the (i, j)-th element of ς^{-1} (inverse of ς).

4.1 Fixed weight $w(Z)$ against N

First we consider a situation where $w(Z)$ is independent of the sample size N. Although $K_{h_N}(\|\theta_n - \theta\|)$ depends on the sample size, to state asymptotic property of the estimator under $w(Z)$ independent of N is instructive and helpful to consider the weight kernel depending on the sample size N.

Proposition 1 *The expectations of the first and second moment of $\hat{\mu} - \bar{\mu}$ are given by*

$$\mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i] = O\left(\frac{1}{N}\right), \tag{8}$$

and

$$\mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i(\hat{\mu} - \bar{\mu})^j] = \frac{\varsigma^{ik}\varrho_{kl}\varsigma^{lj}}{N} + O\left(\frac{1}{N\sqrt{N}}\right). \tag{9}$$

Moreover

$$\sqrt{N}(\hat{\mu} - \bar{\mu}) \rightsquigarrow \mathcal{N}(0, \varsigma^{-1}\varrho\varsigma^{-1}),$$

where $\mathcal{N}(\mu, \Sigma)$ is normal distribution with mean μ and covariance Σ.

Proof Since $\hat{\mu}$ is an *M-estimator* (Huber, 1964) with respect to a risk function $\rho(Z|\mu) = w(Z)\log p(X|\mu)$, the proof is given by standard asymptotic analysis of M-Estimators, see for example (van der Vaart, 1998). We only give brief proofs of assertions (8) and (9). Since $\hat{\mu}$ is the maximizer of the weighted log-likelihood, we have

$$\frac{1}{N}\sum_{n=1}^{N} w(z_n)\bar{\partial}_i \log p(x_n|\hat{\mu}) = 0 \qquad (\forall i).$$

Thus we have

$$0 = \frac{1}{N}\sum_{n=1}^{N} w(z_n)\left(\bar{\partial}_i \log p(x_n|\bar{\mu}) + (\hat{\mu} - \bar{\mu})^j \bar{\partial}_i\bar{\partial}_j \log p(x_n|\hat{\mu})\right) + O_p(\|\hat{\mu} - \bar{\mu}\|^2).$$

This yields

$$\varsigma_{ij}(\hat{\mu} - \bar{\mu})^j = \frac{1}{N}\sum_{n=1}^{N} w(z_n)\bar{\partial}_i \log p(x_n|\bar{\mu}) + O_p\left(\|\hat{\mu} - \bar{\mu}\|^2\right). \tag{10}$$

Since $\bar{\mu}$ maximizes the expectation of the weighted log-likelihood, we have

$$\mathrm{E}_q[w(Z)\bar{\partial}_i \log p(X|\bar{\mu})] = 0.$$

Therefore taking expectation of (10), we obtain the assertion (8). (9) is also proven from (10). □

4.2 Varying weight $w(Z)$ against N

Here we fix $\theta \in \Theta$ which is an interior point of Θ. If w depends on the sample size N as in the case of $w(z_n) = K_{h_N}(\|\theta_n - \theta\|)$, the above proposition should be modified because g and h are not constants and $\varsigma^{-1}\varrho\varsigma^{-1}$ may converge to 0 or diverge to ∞. However for the settings

$$\begin{aligned} &\text{SettingA :} \\ &\quad q(z) \leftarrow \pi(\theta')q_{\theta'}(x),\ w(z) \leftarrow K_{h_N}(\|\theta' - \theta\|) \quad (z = (\theta', x)) \end{aligned}$$

(we call this setting Setting A), we have a proposition which is analogous to Proposition 1. Before stating the proposition we remark the following lemma.

Lemma 3 *Let C_3 be*

$$C_3 := \int_{\mathbf{R}^\ell} K(\|y\|^2)^2 \mathrm{d}y,$$

which is finite because of Assumption 3. Then for all continuous function $f : \Theta \to \mathbf{R}$, we have

$$\lim_{h_N \to 0} \int_\Theta K_{h_N}(\|\theta' - \theta\|) f(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta' \to C_1 f(\theta), \tag{11}$$

and

$$\lim_{h_N \to 0} \int_\Theta {h_N}^\ell K_{h_N}(\|\theta' - \theta\|)^2 f(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta' \to C_3 f(\theta). \tag{12}$$

The proof is given in A. Then we obtain the following proposition.

Proposition 2 *Under Setting A and Assumption 1 and 2, we have*

$$\varsigma_{ij} - C_1 \pi(\theta) s_{ij} \to 0,$$

with $h_N \to 0$. Moreover, if $N{h_N}^\ell \to \infty$, the first and second moment of $\hat{\mu} - \bar{\mu}$ are given by

$$\mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i] = O\left(\frac{1}{N{h_N}^\ell}\right), \tag{13}$$

and

$$\mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j] = \frac{s^{ij}}{B\pi(\theta)N{h_N}^\ell} + o\left(\frac{1}{N{h_N}^\ell}\right), \tag{14}$$

where $B = \frac{C_1^2}{C_3}$. Moreover

$$\sqrt{B\pi(\theta)N{h_N}^\ell}(\hat{\mu} - \bar{\mu}) \rightsquigarrow \mathcal{N}(0, s^{-1}). \tag{15}$$

Proof

By (11) we have

$$\varsigma_{ij} = -\mathrm{E}_q[K_{h_N}(\|\theta' - \theta\|)\bar{\partial}_i \bar{\partial}_j \log p(X|\bar{\mu})]$$

$$
\begin{aligned}
&= \int -K_{h_N}(\|\theta' - \theta\|)\pi(\theta') \left(\int \bar{\partial}_i\bar{\partial}_j \log p(X|\bar{\mu}) q_{\theta'}(X)\mathrm{d}X \right) \sqrt{|\tilde{g}|}\mathrm{d}\theta' \\
&\to C_1\pi(\theta) \int -\bar{\partial}_i\bar{\partial}_j \log p(X|\mu) q_\theta(X)\mathrm{d}X.
\end{aligned}
$$

Also by (12) we have

$$
\begin{aligned}
{h_N}^\ell \varrho_{ij} &= \mathrm{E}_q[{h_N}^\ell K_{h_N}(\|\theta' - \theta\|)^2 \bar{\partial}_i \log p(X|\bar{\mu}) \bar{\partial}_j \log p(X|\bar{\mu})] \\
&= \int {h_N}^\ell K_{h_N}(\|\theta' - \theta\|)^2 \pi(\theta') \left(\int \bar{\partial}_i \log p(X|\bar{\mu}) \bar{\partial}_j \log p(X|\bar{\mu}) q_{\theta'}(X)\mathrm{d}X \right) \sqrt{|\tilde{g}|}\mathrm{d}\theta' \\
&\to C_3\pi(\theta) \int \bar{\partial}_i \log p(X|\mu) \bar{\partial}_j \log p(X|\mu) q_\theta(X)\mathrm{d}X
\end{aligned}
$$

where we used $\bar{\mu} \to \mu$ in the last line. Since it is easy to prove $s_{ij} \to \int -\bar{\partial}_i\bar{\partial}_j \log p(X|\mu) q_\theta(X)\mathrm{d}X$, we obtain the first assertion.

By definition, $q_\theta(X) = p(X|\mu)$. This gives the relation

$$
\int -\bar{\partial}_i\bar{\partial}_j \log p(X|\mu) q_\theta(X)\mathrm{d}X = \int \bar{\partial}_i \log p(X|\mu) \bar{\partial}_j \log p(X|\mu) q_\theta(X)\mathrm{d}X
$$

so that we obtain

$$
\begin{aligned}
\frac{\varsigma^{ik}\varrho_{kl}\varsigma^{lj}}{N} &= \frac{\varsigma^{ij} C_3}{N{h_N}^\ell C_1} + o(1/N{h_N}^\ell) \\
&= \frac{s^{ij} C_3}{N{h_N}^\ell C_1^2 \pi(\theta)} + o(1/N{h_N}^\ell). \qquad (16)
\end{aligned}
$$

Here noticing that (10) gives

$$
(\hat{\mu} - \bar{\mu})^i = \varsigma^{ij} \frac{1}{N} \sum_{n=1}^{N} K_{h_N}(\|\theta_n - \theta\|) \bar{\partial}_j \log p(x_n|\bar{\mu}) + O_p\left(\|\hat{\mu} - \bar{\mu}\|^2\right), \qquad (17)
$$

(16) yields

$$
\mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j] = \frac{s^{ij}}{BN{h_N}^\ell \pi(\theta)} + o\left(\frac{1}{N{h_N}^\ell}\right).
$$

Hence (14) is shown. Also taking the expectation of (17), we have the assertion (13) because

$$
\mathrm{E}_{D^N}\|\hat{\mu} - \bar{\mu}\|^2 = O\left(\frac{1}{N{h_N}^\ell}\right).
$$

Finally (15) is proven by the central limit theorem and Slutsky's lemma. □

Proposition 2 indicates that

$$
\begin{aligned}
\frac{1}{2}\mathrm{E}_{D^N}[(\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j] s_{ij} &= \frac{s^{ij} s_{ij}}{2N(h_N)} + o(N(h_N)^{-1}) \\
&= \frac{d}{2N(h_N)} + o(N(h_N)^{-1}),
\end{aligned}
$$

with $N(h_N) = NB{h_N}^{\ell}\pi(\theta)$. This gives (6).

We remark that the substantial sample size $BN{h_N}^{\ell}\pi(\theta)$ of local aggregated data is proportional to $\pi(\theta)$. This implies that for θ with small $\pi(\theta)$, the substantial sample size around θ is small. This is intuitively appealing because the number of samples observed around θ is approximately proportional to $\pi(\theta)$.

Finally we show two examples which satisfy Assumption 1.

Example 1 $K(\|\theta' - \theta\|) = \exp(-\|\theta' - \theta\|^2)$.

The first relation (11) *of Lemma 3 holds with* $C_1 = \sqrt{\boldsymbol{\pi}^{\ell}}$. *The second relation* (12) *of Lemma 3 is given by* $C_3 = \sqrt{(\boldsymbol{\pi}/2)^{\ell}}$ *because*

$$\int_{\mathbf{R}^{\ell}} \frac{1}{h^{2\ell}} \exp\left(-2\frac{\|\theta'\|^2}{h^2}\right) \mathrm{d}\theta' = \frac{1}{h^{2\ell}} \left(\sqrt{2\boldsymbol{\pi}\frac{h^2}{4}}\right)^{\ell} = \frac{1}{h^{\ell}} \left(\sqrt{\frac{\boldsymbol{\pi}}{2}}\right)^{\ell}.$$

Example 2 $K(\|\theta' - \theta\|) = \mathbf{1}\{\|\theta' - \theta\| \leq 1\}$.

In this example, Lemma 3 holds with $C_1 = C_3 = \mathrm{Leb}(\{\theta' \in \mathbf{R}^{\ell} \mid \|\theta'\| \leq 1\})$, *where* Leb *is the Lebesgue measure.*

5 Conclusion and discussion

We investigated the maximum weighted log-likelihood estimator on locally aggregated data. Asymptotic properties including its asymptotic risk were shown. We observed there appears the bias-variance trade off induced by the window width. In particular, it was seen that the bias term is characterized by a geometric quantity, weighted Laplacian, which gives an intuitive explanation that the bias v is determined by the "parallel" component of Q (the difference between locally averaged distribution and the true one) to the model. Optimal window width that minimizes the asymptotic risk of the estimator was also given.

As stated in the introduction, the maximum weighted log-likelihood estimator of our settings is a simple "pointwise" estimator for $\mu(\theta)$ while Eguchi et al. (2003) locally modeled the regression function $\mu(\theta)$ in a exponential family. The (local) modelling of the regression function exploits some smoothness property of the regression function so that faster convergence of the generalization performance will be expected under a regression function with enough smoothness. That direction of extension of our analysis might be interesting. However the geometric interpretation would be lost in higher order convergence analysis.

Another interesting (and even challenging) future work is to construct a window width selection protocol to select the optimal window width h_N^*. Ideally AIC-type information criterion is preferable. However a problem is that the primary term of the asymptotic risk contains $g^{ij}\gamma_{ij}$ which might be hard to be known beforehand. In practice, cross validation or bootstrap might be helpful for the determination of h_N, and they would achieve the optimal window width asymptotically.

Acknowledgements

The author thanks Prof. Fumiyasu Komaki for their valuable comments and suggestions. This work was supported in part by Global COE Program "The research and training center for new development in mathematics", MEXT, Japan.

A Proof of Lemmas

Proof of Lemma 1

By Theorem 1 and definitions following the theorem, $\bar{q}_\theta$ is expressed by

$$\bar{q}_\theta = p_\mu + \frac{{h_N}^2}{2} Q + T$$

where $T = o({h_N}^2)$ and $\int T \mathrm{d}x = 0$. Since

$$\log\left(p_\mu + \frac{{h_N}^2}{2}Q + T\right) = \log p_\mu + \frac{Q{h_N}^2}{2p_\mu} + \frac{T}{p_\mu} - \frac{{h_N}^4 Q^2}{8p_\mu^2} + o({h_N}^4),$$

we observe

$$\begin{aligned} D(\bar{q}_\theta || p_{\bar\mu}) &= \int \left(p_\mu + \frac{{h_N}^2}{2}Q + T\right)\left(\log p_\mu + \frac{{h_N}^2}{2p_\mu}Q + \frac{T}{p_\mu} - \frac{{h_N}^4 Q^2}{8p_\mu^2} - \log p_{\bar\mu}\right)\mathrm{d}x + o({h_N}^4) \\ &= \int p_\mu \log\frac{p_\mu}{p_{\bar\mu}}\mathrm{d}x + \frac{{h_N}^2}{2}\left[\int Q\mathrm{d}x + \int Q\log\frac{p_\mu}{p_{\bar\mu}}\mathrm{d}x\right] + \int T\mathrm{d}x \\ &\quad - \frac{{h_N}^4}{8}\int\frac{Q^2}{p_\mu}\mathrm{d}x + \frac{{h_N}^4}{4}\int\frac{Q^2}{p_\mu}\mathrm{d}x + \int T\log\frac{p_\mu}{p_{\bar\mu}}\mathrm{d}x + o({h_N}^4). \end{aligned} \tag{18}$$

Now noticing that

$$\begin{aligned} \int p_\mu \log\frac{p_\mu}{p_{\bar\mu}}\mathrm{d}x &= \int p_\mu\left[\log p_\mu - \left(\log p_\mu + v^i\bar\partial_i \log p_\mu + \frac{v^iv^j}{2}\bar\partial_i\bar\partial_j\log p_\mu\right)\right]\mathrm{d}x + O(\|v\|^3) \\ &= \frac{v^iv^j}{2}g_{ij} + O(\|v\|^3), \end{aligned}$$

$$\int Q\mathrm{d}x = 0, \qquad \int T\mathrm{d}x = 0,$$

and

$$\int Q\log\frac{p_\mu}{p_{\bar\mu}}\mathrm{d}x = \int Q\left(-v^i\bar\partial_i\log p_\mu - \frac{v^iv^j}{2}\bar\partial_i\bar\partial_j\log p_\mu\right)\mathrm{d}x + O(\|v\|^3),$$

the terms with orders not higher than ${h_N}^2$ is expressed as

$$\begin{aligned} &\int p_\mu\log\frac{p_\mu}{p_{\bar\mu}} + \frac{{h_N}^2}{2}\left[\int Q\mathrm{d}x + \int Q\log\frac{p_\mu}{p_{\bar\mu}}\mathrm{d}x\right] \\ =&\frac{v^iv^j}{2}g_{ij} - \frac{{h_N}^2}{2}v^i\int Q\bar\partial_i\log p_\mu\mathrm{d}x - \frac{{h_N}^2}{4}v^iv^j\int Q\bar\partial_i\bar\partial_j\log p_\mu\mathrm{d}x + O(\|v\|^3) \\ =&\frac{g_{ij}}{2}\left(v^i - \frac{{h_N}^2}{2}\int Q\bar\partial_k\log p_\mu\mathrm{d}x g^{ki}\right)\left(v^j - \frac{{h_N}^2}{2}\int Q\bar\partial_k\log p_\mu\mathrm{d}x g^{kj}\right) \\ &- \frac{{h_N}^4}{8}g^{ij}\int Q\bar\partial_i\log p_\mu\mathrm{d}x\int Q\bar\partial_j\log p_\mu\mathrm{d}x - \frac{{h_N}^2}{4}v^iv^j\int Q\bar\partial_i\bar\partial_j\log p_\mu\mathrm{d}x + O(\|v\|^3). \end{aligned}$$

Thus v^i which minimizes the above equation is

$$v^i = \frac{{h_N}^2}{2} g^{ik} \int Q \bar{\partial}_k \log p_\mu \mathrm{d}x + o({h_N}^2).$$

This gives $\int T \log \frac{p_\mu}{p_{\bar{\mu}}} \mathrm{d}x = o({h_N}^4)$. Therefore returning to (18), we obtain

$$D(\bar{q}_\theta || p_{\bar{\mu}}) = -\frac{{h_N}^4}{8} g^{ij} \int Q \bar{\partial}_i \log p_\mu \mathrm{d}x \int Q \bar{\partial}_j \log p_\mu \mathrm{d}x + \frac{{h_N}^4}{8} \int \frac{Q^2}{p_\mu} \mathrm{d}x + o({h_N}^4).$$

□

Proof of Lemma 2

By Theorem 1 we have the following expansion:

$$\begin{aligned}
D(p_\mu || p_{\hat{\mu}}) &= \int \left(\bar{q}_\theta - \frac{{h_N}^2 Q}{2} - T \right) \log \left(\frac{\bar{q}_\theta - \frac{{h_N}^2 Q}{2} - T}{p_{\hat{\mu}}} \right) \mathrm{d}x \\
&= \int \left(\bar{q}_\theta - \frac{{h_N}^2}{2} Q - T \right) \left(\log \bar{q}_\theta - \frac{{h_N}^2 Q}{2\bar{q}_\theta} - \frac{T}{\bar{q}_\theta} - \frac{{h_N}^4 Q^2}{8\bar{q}_\theta^2} - \log p_{\hat{\mu}} \right) \mathrm{d}x + o({h_N}^4) \\
&= \int \bar{q}_\theta \left(\log \bar{q}_\theta - \frac{{h_N}^2 Q}{2\bar{q}_\theta} - \frac{T}{\bar{q}_\theta} - \frac{{h_N}^4 Q^2}{8\bar{q}_\theta^2} - \log p_{\hat{\mu}} \right) \mathrm{d}x \\
&\quad - \int \frac{{h_N}^2}{2} Q \left(\log \bar{q}_\theta - \frac{{h_N}^2 Q}{2\bar{q}_\theta} - \log p_{\hat{\mu}} \right) \mathrm{d}x - \int T \log \frac{\bar{q}_\theta}{p_{\hat{\mu}}} \mathrm{d}x + o({h_N}^4) \\
&= D(\bar{q}_\theta || p_{\hat{\mu}}) - \int \frac{{h_N}^4 Q^2}{8\bar{q}_\theta} \mathrm{d}x - \int \frac{{h_N}^2}{2} Q \left(\log \bar{q}_\theta - \log p_{\hat{\mu}} \right) \mathrm{d}x + \frac{{h_N}^4}{4} \int \frac{Q^2}{\bar{q}_\theta} \mathrm{d}x \\
&\quad - \int T \log \frac{\bar{q}_\theta}{p_{\hat{\mu}}} \mathrm{d}x + o({h_N}^4). \qquad (19)
\end{aligned}$$

Now $D(\bar{q}_\theta || p_{\hat{\mu}})$ is expanded as

$$\begin{aligned}
&D(\bar{q}_\theta || p_{\hat{\mu}}) \\
&= \int \bar{q}_\theta \log \left(\frac{\bar{q}_\theta}{p_{\bar{\mu}}} \right) \mathrm{d}x + \int \bar{q}_\theta \log \left(\frac{p_{\bar{\mu}}}{p_{\hat{\mu}}} \right) \mathrm{d}x \\
&= D(\bar{q}_\theta || p_{\bar{\mu}}) - \int \bar{q}_\theta \left[(\hat{\mu} - \bar{\mu})^i \bar{\partial}_i \log p_{\bar{\mu}} + \frac{(\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j}{2} \bar{\partial}_i \bar{\partial}_j \log p_{\bar{\mu}} \right] \mathrm{d}x + o_p(\|\hat{\mu} - \bar{\mu}\|^2).
\end{aligned}$$

Here $\int \bar{q}_\theta \bar{\partial}_i \log p_{\bar{\mu}} \mathrm{d}x = 0$ because $\bar{\mu}$ is the minimizer of $D(\bar{q}_\theta || p_{\mu'})$ over $\mu' \in \mathcal{M}$. Therefore we have

$$D(\bar{q}_\theta || p_{\hat{\mu}}) = D(\bar{q}_\theta || p_{\bar{\mu}}) + \frac{1}{2} (\hat{\mu} - \bar{\mu})^i (\hat{\mu} - \bar{\mu})^j s_{ij} + o_p(\|\hat{\mu} - \bar{\mu}\|^2).$$

This and (19) yields the assertion. □

Proof of Lemma 3

Let $B_\Theta(\theta, \epsilon)$ be a ball around θ of radius ϵ in Θ with respect to the geodesic distance. Since Θ is compact, it is not self-approaching. Thus for all $\epsilon > 0$ there exists $\delta > 0$ such that $\forall \theta' \in \Theta \setminus B_\Theta(\theta, \epsilon)$ satisfies $\|\theta' - \theta\| > \delta$. We decompose the integral of (11) to

$$\int_\Theta K_{h_N}(\|\theta' - \theta\|) f(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta' = \int_{\Theta \setminus B_\Theta(\theta, \epsilon)} + \int_{B_\Theta(\theta, \epsilon)} K_{h_N}(\|\theta' - \theta\|) f(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta'.$$

The first term decays exponentially because of Assumption 1. The second term is evaluated as

$$\frac{\int_{B_\Theta(\theta,\epsilon)} K_{h_N}(\|\theta' - \theta\|) f(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta'}{\int_{B_\Theta(\theta,\epsilon)} K_{h_N}(\|\theta' - \theta\|) \sqrt{|\tilde{g}|} \mathrm{d}\theta'} \int_{B_\Theta(\theta,\epsilon)} K_{h_N}(\|\theta' - \theta\|) \sqrt{|\tilde{g}|} \mathrm{d}\theta'$$

Since $f(\theta')$ is continuous, taking ϵ small enough the term

$$\frac{\int_{B_\Theta(\theta,\epsilon)} K_{h_N}(\|\theta' - \theta\|) f(\theta') \sqrt{|\tilde{g}|} \mathrm{d}\theta'}{\int_{B_\Theta(\theta,\epsilon)} K_{h_N}(\|\theta' - \theta\|) \sqrt{|\tilde{g}|} \mathrm{d}\theta'}$$

becomes arbitrary close to $f(\theta)$ because the above display is an average of $f(\theta')$ taken over ϵ-neighborhood of θ. Since $\tilde{g}$ is induced by Euclidean metric and K has an exponential tail decay, then we obtain

$$\int_{B_\Theta(\theta,\epsilon)} K_{h_N}(\|\theta' - \theta\|) \sqrt{|\tilde{g}|} \mathrm{d}\theta' \to \int_{\mathbf{R}^\ell} K(\|y\|^2) \mathrm{d}y \qquad (\text{as } h_N \to 0).$$

This concludes the proof of (11). (12) is also proven in a similar way. □

References

Copas, J. B. (1995). Local likelihood based on kernel censoring. *Journal of the Royal Statistical Society B* 57:221–235.

Eguchi, S., & Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society B* 60:709–724.

Eguchi, S., Kim, T. Y., & Park, B. U. (2003). Local likelihood method: A bridge over parametric and nonparametric regression. *Nonparametric Statistics* 15(6):665–683.

Grigor'yan, A. (2006). Heat kernels on weighted manifolds and applications. *Contemporary Mathematics* 398:93–191.

Györfi, L., Kohler, M., Kryżak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression.* Springer, New York.

Hein, M., Audibert, J.-Y., & Luxburg, U. von. (2007). Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research* 8:1325–1370.

Hjort, N. L., & Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics* 24(4):1619–1647.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35(1):73–101.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics* 24(4):1602–1618.

Park, B. U., Kim, W. C., & Jones, M. C. (2002). On local likelihood density estimation. *The Annals of Statistics* 30(5):1480–1495.

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* 82(398):559–567.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press.

Yu, K., & Jones, M. C. (2004). Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association* 99(465):139–144.