Mathematics for Data Sciences **09/23/2011**

Lecture 05. Diffusion Map, Convergence theory

Instructor: Xiuyuan Cheng, Princeton University *Scribe: Jun Yin, Ya'ning Liu*

W is positive definite if using Gaussian Kernel

This is left by previous lecture.

One can check that, when

$$Q(x) = \int_{\mathbb{R}} e^{-ix\xi} d\mu(\xi),$$

for some positive finite Borel measure $d\mu$ on $\mathbb{R}$, then the (symmetric/Hermitian) integral kernel

$$k(x, y) = Q(x - y)$$

is positive definite, that is, for any function $\phi(x)$ on $\mathbb{R}$,

$$\int\int \bar{\phi}(x)\phi(y)k(x, y) \geq 0.$$

Proof omitted. The reverse is also true, which is Bochner theorem. High dimensional case is similar.

Take 1-dimensional as an example. Since the Gaussian distribution $e^{-\xi^2/2}d\xi$ is a positive finite Borel measure, and the Fourier transform of Gaussian kernel is itself, we know that $k(x, y) = e^{-|x-y|^2/2}$ is a positive definite integral kernel. The matrix W as an discretized version of $k(x, y)$ keeps the positive-definiteness (make this rigorous? Hint: take $\phi(x)$ as a linear combination of n delta functions).

1 Main Result

In this lecture, we will study the bias and variance decomposition for sample graph Laplacians and their asymptotic convergence to Laplacian-Beltrami operators on manifolds.

Let $\mathcal{M}$ be a smooth manifold without boundary in $\mathbb{R}^p$ (*e.g.* a d-dimensional sphere). Randomly draw a set of n data points, $x_1, ..., x_n \in M \subset \mathbb{R}^p$, according to distribution $p(x)$ in an independent and identically distributed (i.i.d.) way. We can extract an $n \times n$ weight matrix W_{ij} as follows:

$$W_{ij} = k(x_i, x_j)$$

where $k(x, y)$ is a symmetric $k(x, y) = k(y, x)$ and positivity-preserving kernel $k(x, y) \geq 0$. As an example, it can be the *heat kernel* (or Gaussian kernel),

$$k_\epsilon(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\epsilon}\right),$$

where $||\centerdot||^2$is the Euclidean distance in space R^p and ϵ is the bandwidth of the kernel. W_{ij} stands for similarity function between x_i and x_j. A diagonal matrix D is defined with diagonal elements are the row sums of W:

$$D_{ii} = \sum_{j=1}^{n} W_{ij}.$$

Let's consider a family of re-weighted similarity matrix, with superscript (α),

$$W^{(\alpha)} = D^{-\alpha} W D^{-\alpha}$$

and

$$D_{ii}^{(\alpha)} = \sum_{j=1}^{n} W_{ij}^{(\alpha)}.$$

Denote $A^{(\alpha)} = (D^{(\alpha)})^{-1} W$, and we can verify that $\sum_{j=1}^{n} A_{ij}^{(\alpha)} = 1$, *i.e.* a row Markov matrix. Now define $L^{(\alpha)} = A^{(\alpha)} - I = (D^{(\alpha)})^{-1} W^{(\alpha)} - I$; and

$$L_{\epsilon,\alpha} = \frac{1}{\epsilon}(A_{\epsilon}^{(\alpha)} - I)$$

when $k_\epsilon(x,y)$ is used in constructing W. In general, $L^{(\alpha)}$ and $L_{\epsilon,\alpha}$ are both called *graph Laplacians*. In particular $L^{(0)}$ is the unnormalized graph Laplacian in literature.

The target is to show that graph Laplacian $L_{\epsilon,\alpha}$ converges to continuous differential operators acting on smooth functions on $\mathcal{M}$ the manifold. The convergence can be roughly understood as: we say a sequence of n-by-n matrix $L^{(n)}$ as $n \to \infty$ converges to a limiting operator $\mathcal{L}$, if for $\mathcal{L}$'s eigenfunction $f(x)$ (a smooth function on $\mathcal{M}$) with eigenvalue λ, that is

$$\mathcal{L} f = \lambda f,$$

the length-n vector $f^{(n)} = (f(x_i)), (i = 1, \cdots, n)$ is approximately an eigenvector of $L^{(n)}$ with eigenvalue λ, that is

$$L^{(n)} f^{(n)} = \lambda f^{(n)} + o(1),$$

where $o(1)$ goes to zero as $n \to \infty$.

Specifically, (the convergence is in the sense of multiplying a positive constant)

(I) $L_{\epsilon,0} = \frac{1}{\epsilon}(A_\epsilon - I) \to \frac{1}{2}(\Delta_{\mathcal{M}} + 2\frac{\nabla p}{p} \cdot \nabla)$ as $\epsilon \to 0$ and $n \to \infty$. $\Delta_{\mathcal{M}}$ is the Laplace-Beltrami operator of manifold M. At a point on M which is d-dimensional, in local (orthogonal) geodesic coordinate $s_1, \cdots, s_d$, the Laplace-Beltrami operator has the same form as the laplace in calculus

$$\Delta_{\mathcal{M}} f = \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2} f;$$

∇ denotes the gradient of a function on M, and $\cdot$ denotes the inner product on tangent spaces of $\mathcal{M}$. Note that $p = e^{-V}$, so $\frac{\nabla p}{p} = -\nabla V$.

(Ignore this part if you don't know stochastic process) Suppose we have the following diffusion process

$$dX_t = -\nabla V(X_t)dt + \sigma dW_t^{(M)},$$

where $W_t^{(M)}$ is the Brownian motion on M, and σ is the volatility, say a positive constant, then the backward Kolmogorov operator/Fokker-Plank operator/infinitesimal generator of the process is

$$\frac{\sigma^2}{2}\Delta_{\mathcal{M}} - \nabla V \cdot \nabla,$$

so we say in (I) the limiting operator is the Fokker-Plank operator. Notice that in Lafon '06 paper they differ the case of $\alpha = 0$ and $\alpha = 1/2$, and argue that only in the later case the limiting operator is the Fokker-Plank. However the difference between $\alpha = 0$ and $\alpha = 1/2$ is a $1/2$ factor in front of $-\nabla V$, and that can be unified by changing the volatility σ to another number. (Actually, according to Thm 2. on Page 15 of Lafon'06, one can check that $\sigma^2 = \frac{1}{1-\alpha}$.) So here we say for $\alpha = 0$ the limiting operator is also Fokker-Plank. (not talked in class, open to discussion...)

(II) $L_{\epsilon,1} = \frac{1}{\epsilon}(A_\epsilon^{(1)} - I) \to \frac{1}{2}\Delta_{\mathcal{M}}$ as $\epsilon \to 0$ and $n \to \infty$. Notice that this case is of important application value: whatever the density $p(x)$ is, the Laplacian-Beltrami operator of $\mathcal{M}$ is approximated, so the geometry of the manifold can be understood.

A special case is that samples x_i are uniformly distributed on $\mathcal{M}$, whence $\nabla p = 0$. Then (I) and (II) are the same up to multiplying a positive constant, due to that D's diagonal entries are almost the same number and the re-weight does not do anything.

Convergence results like these can be found in Coifman and Lafon (2006), *Diffusion maps, Applied and Computational Harmonic Analysis.*

We also refer Singer (2006) *From graph to manifold Laplacian: The convergence rate, Applied and Computational Harmonic Analysis* for a complete analysis of the variance error, while the analysis of bias is very brief in this paper.

2 Proof

For a smooth function $f(x)$ on $\mathcal{M}$, let $f = (f_i) \in \mathbb{R}^n$ as a vector defined by $f_i = f(x_i)$. At a given fixed point x_i, we have the formula:

$$\begin{aligned}(Lf)^i &= \frac{1}{\epsilon}\left(\frac{\sum_{j=1}^n W_{ij} f_j}{\sum_{j=1}^n W_{ij}} - f_i\right) = \frac{1}{\epsilon}\left(\frac{\frac{1}{n}\sum_{j=1}^n W_{ij} f_j}{\frac{1}{n}\sum_{j=1}^n W_{ij}} - f_i\right) \\ &= \frac{1}{\epsilon}\left(\frac{\frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j).f(x_j)}{\frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j)} - f(x_i) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right)\end{aligned}$$

where in the last step the diagonal terms $j = i$ are excluded from the sums resulting in an $O(n^{-1}\epsilon^{-\frac{d}{2}})$ error. Later we will see that compared to the variance error, this term is negligible.

We rewrite the Laplacian above as

$$(Lf)^i = \frac{1}{\epsilon}\left(\frac{F(x_i)}{G(x_i)} - f(x_i) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right) \tag{1}$$

where

$$F(x_i) = \frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j)f(x_j), \quad G(x_i) = \frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j).$$

depends only on the other $n-1$ data points than x_i. In what follows we treat x_i as a fixed chosen point and write as x.

Bias-Variance Decomposition. The points $x_j, j \neq i$ are independent identically distributed (i.i.d), therefore every term in the summation of $F(x)$ ($G(x)$) are i.i.d., and by the Law of Large Numbers (LLN) one should expect $F(x) \approx \mathbb{E}_{x_1}[k(x,x_1)f(x_1)] = \int_{\mathcal{M}} k(x,y)f(y)p(y)dy$ (and $G(x) \approx \mathbb{E}k(x,x_1) = \int_{\mathcal{M}} k(x,y)p(y)dy$). Recall that given a random variable x, and a sample estimator $\hat{\theta}$ (*e.g.* sample mean), the bias-variance decomposition is given by

$$\mathbb{E}\|x - \hat{\theta}\|^2 = \mathbb{E}\|x - \mathbb{E}x\|^2 + \mathbb{E}\|\mathbb{E}x - \hat{\theta}\|^2.$$

If we use the same strategy here (though not exactly the same, since $\mathbb{E}[\frac{F}{G}] \neq \frac{\mathbb{E}[F]}{\mathbb{E}[G]}$!), we can decompose Eqn. (1) as

$$\begin{aligned}(Lf)^i &= \frac{1}{\epsilon}\left(\frac{\mathbb{E}[F]}{\mathbb{E}[G]} - f(x_i) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right) + \frac{1}{\epsilon}\left(\frac{F(x_i)}{G(x_i)} - \frac{\mathbb{E}[F]}{\mathbb{E}[G]}\right) \\ &= bias + variance.\end{aligned}$$ span

In the below we shall show that for case (I) the estimates are

$$bias = \frac{1}{\epsilon}\left(\frac{\mathbb{E}[F]}{\mathbb{E}[G]} - f(x) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right) = \frac{m_2}{2}(\Delta_{\mathcal{M}} f + 2\nabla f \cdot \frac{\nabla p}{p}) + O(\epsilon) + O\left(n^{-1}\epsilon^{-\frac{d}{2}}\right). \quad (2)$$

$$variance = \frac{1}{\epsilon}\left(\frac{F(x_i)}{G(x_i)} - \frac{\mathbb{E}[F]}{\mathbb{E}[G]}\right) = O(n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}), \quad (3)$$

whence

$$bias + variance = O(\epsilon, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}) = C_1\epsilon + C_2 n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}.$$

As the bias is a monotone increasing function of ϵ while the variance is decreasing w.r.t. ϵ, the optimal choice of ϵ is to balance the two terms by taking derivative of the right hand side equal to zero (or equivalently setting $\epsilon \sim n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}$) whose solution gives the optimal rates

$$\epsilon^* \sim n^{-1/(2+d/2)}.$$

Lafon'06 gives the bias and Hein'05 contains the variance parts, which are further improved by Singer'06 in both bias and variance.

2.1 The Bias Term

Now focus on $\mathbb{E}[F]$

$$\mathbb{E}[F] = \mathbb{E}\left[\frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i, x_j) f(x_j)\right] = \frac{n-1}{n}\int_{\mathcal{M}} k_\epsilon(x,y) f(y) p(y) dy$$

$\frac{n-1}{n}$ is close to 1 and is treated as 1.

1. the case of one-dimensional and flat (which means the manifold $\mathcal{M}$ is just a real line, *i.e.* $\mathcal{M} = \mathbb{R}$)

   Let $\tilde{f}(y) = f(y)p(y)$, and $k_\epsilon(x,y) = \frac{1}{\sqrt{\epsilon}} e^{-\frac{(x-y)^2}{2\epsilon}}$, by change of variable

   $$y = x + \sqrt{\epsilon} z,$$

   we have

   $$\square = \int_{\mathbb{R}} \tilde{f}(x + \sqrt{\epsilon}z) e^{-\frac{\epsilon^2}{2}} dz = m_0 \tilde{f}(x) + \frac{1}{2} m_2 f''(x)\epsilon + O(\epsilon^2)$$

   where $m_0 = \int_{\mathbb{R}} e^{-\frac{\epsilon^2}{2}} dz$, and $m_2 = \int_{\mathbb{R}} z^2 e^{-\frac{\epsilon^2}{2}} dz$.

2. 1 Dimensional & Not flat:

   Divide the integral into 2 parts:

   $$\int_m k_\epsilon(x,y)\tilde{f}(y)p(y)dy = \int_{||x-y||>c\sqrt{\epsilon}} \cdot + \int_{||x-y||<c\sqrt{\epsilon}} \cdot$$

First part $= \circ$

$$|\circ| \le ||\tilde{f}||_\infty \frac{1}{\epsilon^{\frac{a}{2}}} e^{-\frac{\epsilon^2}{2\epsilon}},$$

due to $||x-y||^2 > c\sqrt{\epsilon}$

$$c \sim ln(\frac{1}{\epsilon}).$$

so this item is tiny and can be ignored.

Locally, that is $u \sim \sqrt{\epsilon}$, we have the curve in a plane and has the following parametrized equation

$$(x(u), y(u)) = (u, au^2 + qu^3 + \cdots),$$

then the chord length

$$\frac{1}{\epsilon}||x-y||^2 = \frac{1}{\epsilon}[u^2 + (au^2 + qu^3 + ...)^2] = \frac{1}{\epsilon}[u^2 + a^2u^4 + q_5(u) + \cdots],$$

where we mark $a^2u^4 + 2aqu^5 + ... = q_5(u)$. Next, change variable $\frac{u}{\sqrt{\epsilon}} = z$, then with $h(\xi) = e^{-\frac{\xi}{2}}$

$$h(\frac{||x-y||}{\epsilon})^2 = h(z^2) + h'(z^2)(\epsilon^2 az^4 + \epsilon^{\frac{3}{2}} q_5 + O(\epsilon^2)),$$

also

$$\tilde{f}(s) = \tilde{f}(x) + \frac{d\tilde{f}}{ds}(x)s + \frac{1}{2}\frac{d^2\tilde{f}}{ds^2}(x)s^2 + \cdots$$

and

$$s = \int_0^u \sqrt{1 + (2au + 3quu^2 + ...)^2 du} + \cdots$$

and

$$\frac{ds}{du} = 1 + 2a^2u^2 + q_2(u) + O(\epsilon^2), \quad s = u + \frac{2}{3}a^2u^3 + O(\epsilon^2).$$

Now come back to the intergral

$$\begin{aligned}
&\int_{|x-y|<c\sqrt{\epsilon}} \frac{1}{\sqrt{\epsilon}} h(\frac{x-y}{\epsilon})\tilde{f}(s)ds \\
\approx &\int_{-\infty}^{+\infty} [h(z^2) + h'(z^2)(\epsilon^2 az^4 + \epsilon^{\frac{3}{2}} q_5] \cdot [\tilde{f}(x) + \frac{d\tilde{f}}{ds}(x)(\sqrt{\epsilon}z + \frac{2}{3}a^2z^2\epsilon^{\frac{3}{2}}) \\
&+ \frac{1}{2}\frac{d^2\tilde{f}}{ds^2}(x)\epsilon z^2] \cdot [1 + 2a^2 + \epsilon^3 y_3(z)]dz \\
= &m_0\tilde{f}(x) + \epsilon\frac{m_2}{2}(\frac{d^2\tilde{f}}{ds^2}(x) + a^2\tilde{f}(x)) + O(\epsilon^2),
\end{aligned}$$

span

where the $O(\epsilon^2)$ tails are omitted in middle steps, and $m_0 = \int h(z^2)dz$,$m_2 = \int z^2h(z^2)dz$, are positive constants. In what follows we normalize both of them by m_0, so only m_2 appears as coefficient in the $O(\epsilon)$ term. Also the fact that $h(\xi) = e^{-\frac{\xi}{2}}$, and so $h'(\xi) = -\frac{1}{2}h(\xi)$, is used.

3. For high dimension, $\mathcal{M}$ is of dimension d,

$$k_\epsilon(x, y) = \frac{1}{\epsilon^{\frac{d}{2}}} e^{-\frac{|x-y|^2}{2\epsilon}},$$

the corresponding result is (Lemma 8 in Appendix B of Lafon '06 paper)

$$\int_{\mathcal{M}} k_\epsilon(x,y)\tilde{f}(y)dy = \tilde{f}(x) + \epsilon\frac{m_2}{2}(\Delta_{\mathcal{M}}\tilde{f} + E(x)\tilde{f}(x)) + O(\epsilon^2), \tag{4}$$

where

$$E(x) = \sum_{i=1}^{d} a_i(x)^2 - \sum_{i_1 \neq i_2} a_{i_1}(x)a_{i_2}(x),$$

and $a_i(x)$ are the curvatures along coordinates s_i $(i = 1, \cdots, d)$ at point x.

Now we study the limiting operator and the bias error:

$$\begin{aligned}\frac{\mathbb{E}F}{\mathbb{E}G} &= \frac{\int k_\epsilon(x,y)f(y)p(y)dy}{\int k_\epsilon(x,y)p(y)dy} \approx \frac{f + \epsilon\frac{m_2}{2}(f'' + 2f'\frac{p'}{p} + f\frac{p^2}{p} + Ef) + O(\epsilon^2)}{1 + \epsilon\frac{m_2}{2}(\frac{p''}{p} + E) + O(\epsilon^2)} \\ &= f(x) + \epsilon\frac{m_2}{2}(f'' + 2f'\frac{p'}{p}) + o(\epsilon^2), \end{aligned} \tag{5}$$

and as a result, for generally d-dim case,

$$\frac{1}{\epsilon}\left(\frac{\mathbb{E}F}{\mathbb{E}G} - f(x)\right) = \frac{m_2}{2}(\Delta_{\mathcal{M}}f + 2\nabla f \cdot \frac{\nabla p}{p}) + O(\epsilon).$$

Using the same method and use Eqn. (4), one can show that for case (II) where $\alpha = 1$, the limiting operator is exactly the Laplace-Beltrami operator and the bias error is again $O(\epsilon)$ (homework).

About $\mathcal{M}$ with boundary: firstly the limiting differential operator bears Newmann/no-flux boundary condition. Secondly, the convergence at a belt of width $\sqrt{\epsilon}$ near $\partial\mathcal{M}$ is slower than the inner part of $\mathcal{M}$, see more in Lafon'06 paper.

2.2 Variance Term

Our purpose is to derive the large deviation bound for[1]

$$Prob\left(\frac{F}{G} - \frac{\mathbb{E}[F]}{\mathbb{E}[G]} \geq \alpha\right) \tag{6}$$

where $F = F(x_i) = \frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j)f(x_j)$ and $G = G(x_i) = \frac{1}{n}\sum_{j\neq i} k_\epsilon(x,x_j)$. With $x_1, x_2, ..., x_n$ as i.i.d random variables, F and G are sample means (up to a scaling constant). Define a new random variable

$$Y = \mathbb{E}[G]F - \mathbb{E}[F]G - \alpha\mathbb{E}[G](G - \mathbb{E}[G])$$

which is of mean zero and Eqn. (6) can be rewritten as

$$Prob(Y \geq \alpha\mathbb{E}[G]^2).$$

For simplicity by *Markov (Chebyshev) inequality*[2] ,

$$Prob(Y \geq \alpha\mathbb{E}[G]^2) \leq \frac{\mathbb{E}[Y^2]}{\alpha^2\mathbb{E}[G]^4}$$

[1] The opposite direction is omitted here.
[2] It means that $Prob(X > \alpha) \leq \mathbb{E}(X^2)/\alpha^2$. A Chernoff bound with exponential tail can be found in Singer'06.

and setting the right hand side to be $\delta \in (0,1)$, then with probability at least $1-\delta$ the following holds

$$\alpha \leq \frac{\sqrt{\mathbb{E}[Y^2]}}{\mathbb{E}[G]^2\sqrt{\delta}} \sim O\left(\frac{\sqrt{\mathbb{E}[Y^2]}}{\mathbb{E}[G]^2}\right).$$

It remains to bound

$$\mathbb{E}[Y^2] = (\mathbb{E}G)^2\mathbb{E}(F^2) - 2(\mathbb{E}G)(\mathbb{E}F)\mathbb{E}(FG) + (\mathbb{E}F)^2\mathbb{E}(G^2) + ...$$
$$+2\alpha(\mathbb{E}G)[(\mathbb{E}F)\mathbb{E}(G^2) - (\mathbb{E}G)\mathbb{E}(FG)] + \alpha^2(\mathbb{E}G)^2(\mathbb{E}(G^2) - (\mathbb{E}G)^2).$$

So it suffices to give $\mathbb{E}(F)$, $\mathbb{E}(G)$, $\mathbb{E}(FG)$, $\mathbb{E}(F^2)$, and $\mathbb{E}(G^2)$. The former two are given in bias and for the variance parts in latter three, let's take one simple example with $\mathbb{E}(G^2)$.

Recall that $x_1, x_2, ..., x_n$ are distributed i.i.d according to density $p(x)$, and

$$G(x) = \frac{1}{n}\sum_{j\neq i} k_\epsilon(x, x_j),$$

so

$$Var(G) = \frac{1}{n^2}(n-1)\left[\int_{\mathcal{M}} k_\epsilon(x,y))^2 p(y)dy - (\mathbb{E}k_\epsilon(x,y))^2\right].$$

Look at the simplest case of 1-dimension flat $\mathcal{M}$ for an illustrative example:

$$\int_{\mathcal{M}} (k_\epsilon(x,y))^2 p(y)dy = \int_{\mathbb{R}} \frac{1}{\sqrt{\epsilon}} h^2(z^2)(p(x) + p'(x)(\sqrt{\epsilon}z + O(\epsilon)))dz,$$

let $M_2 = \int_{\mathbb{R}} h^2(z^2)dz$

$$\int_{\mathcal{M}} (k_\epsilon(x,y))^2 p(y)dy = p(x)\cdot\frac{1}{\sqrt{\epsilon}}M_2 + O(\sqrt{\epsilon}).$$

Recall that $\mathbb{E}k_\epsilon(x,y) = O(1)$, we finally have

$$Var(G) \sim \frac{1}{n}\left[\frac{p(x)M_2}{\sqrt{\epsilon}} + O(1)\right] \sim \frac{1}{n\sqrt{\epsilon}}.$$

Generally, for d-dimensional case, $Var(G) \sim n^{-1}\epsilon^{-\frac{d}{2}}$. Similarly one can derive estimates on $Var(F)$.

Ignoring the joint effect of $\mathbb{E}(FG)$, one can somehow get a rough estimate based on $F/G = [\mathbb{E}(F) + O(\sqrt{\mathbb{E}(F^2)})]/[\mathbb{E}(G) + O(\sqrt{\mathbb{E}(G^2)})]$ where we applied the Markov inequality on both the numerator and denominator. Combining those estimates together, we have the following,

$$\begin{aligned}\frac{F}{G} &= \frac{fp + \epsilon\frac{m_2}{2}(\Delta(fp) + \mathbb{E}[fp]) + O(\epsilon^2, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}})}{p + \epsilon\frac{m_2}{2}(\Delta p + \mathbb{E}[p]) + O(\epsilon^2, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}})} \\ &= f + \epsilon\frac{m_2}{2}(\Delta p + \mathbb{E}[p]) + O(\epsilon^2, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}}),\end{aligned}$$

here $O(B_1, B_2)$ denotes the dominating one of the two bounds B_1 and B_2 in the asymptotic limit. As a result, the error (bias + variance) of $L_{\epsilon,\alpha}$ (dividing another ϵ) is of the order

$$O(\epsilon, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}). \tag{7}$$

In Amit Singer '06 paper, the last term in the last line is improved to

$$O(\epsilon, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-\frac{1}{2}}), \tag{8}$$

where the improvement is by carefully analyzing the large deviation bound of $\frac{F}{G}$ around $\frac{\mathbb{E}F}{\mathbb{E}G}$ shown above, making use of the fact that F and G are correlated. Technical details are not discussed here.

In conclusion, we need to choose ϵ to balance bias error and variance error to be both small. For example, by setting the two bounds in Eqn. (8) to be of the same order we have

$$\epsilon \sim n^{-1/2}\epsilon^{-1/2-d/4},$$

that is

$$\epsilon \sim n^{-1/(3+d/2)},$$

so the total error is $O(n^{-1/(3+d/2)})$.