Locally Kernel Regression Adapting with Data Distribution in Prediction of Traffic Flow

Lei Han, Meng Shuai, Kunqing Xie, Guojie Song, Xiujun Ma
Key Laboratory of Machine Perception (Ministry of Education), Peking University
Beijing, China
Email: hanlei@cis.pku.edu.cn

***Abstract*—Prognosis of traffic flow is a basic part of intelligent transportation research. Due to the extremely complexity of vehicular traffic, efficient models should be constructed to do accurate simulation and prediction of real traffic, such as locally kernel models. However, locally kernel regression fails when the traffic data points are sparse, and the data distribution should be considered seriously. Moreover, the spatiotemporal features of real traffic make pure locally kernel regression inapplicable. This paper proposes a locally kernel regression mechanism adapting with data distribution for the prediction of traffic flow. This mechanism is also explained by Three-Phase Traffic Theory. Experimental studies show the feasibility and efficiency of our approach.**

***Index Terms*—Locally kernel regression; Density; Adaptive ; Three-Phase Traffic Theory**

I. Introduction

Vehicular traffic, due to its spatiotemporal properties, is an extremely complex dynamic system. Appropriate and effectual models are required to help human do traffic control, management and other transportation applications, and these pertinent methods associated with traffic optimization aim to minimize the traffic costs, which refer to travel time, fuel consumption, etc. In particular, traffic congestion is the most important problem for researchers to resolve. In recent years, many methodologies are raised for traffic system, and one of the basic functions is short-time traffic prediction which helps to predict congestion and guide traffic control before the jam formation.

To construct an architectonical traffic theory system, Boris S.Kerner's Three-Phase Traffic Theory has explained the spatiotemporal features of traffic rigorously, and it has been used as a basic traffic theory by many other researchers associated with their own concerned fields.

Prognosis of traffic flow is a learning problem within traffic background. Many typical learning methods have been introduced into traffic prediction, such as locally kernel regression, neural network, etc. However, because of the complex spatiotemporal feature of vehicular traffic, these original learning methods are less capable. Shuai Meng, etc. advocate an adaptive locally weighted learning mechanism based on Kerner's Three-Phase Traffic Theory that calculates a bandwidth parameter in local models with the three phases separately and then emerges the contribution of the neighbors of points located at phase boundary together by averaged weighting function as a final prediction result [1], but they did not explain the reason and feasibility of this adaptive mechanism.

Locally kernel regression, which is effective and flexible for prediction problems, is a kind of lazy learning and focus on the local neighborhood of a given query. However, locally kernel regression fails when the neighborhood is sparse, and few surveys concern about adapting parameters with the data distribution [2]. Within traffic background, when there are some traffic states with extreme odd traffic variables that create a sparse traffic pattern, locally kernel regression can hardly find similar neighbors for a given query. This difficulty can be solved by adapting the kernel model parameters with traffic data distribution in the regression. The feature of traffic can be represented by flow rate and vehicle occupancy measured at a road location. In a flow-occupancy plane, every data point represents a traffic state. When we do adaption with data density, it means we do adaption with traffic state density. This paper discusses a prediction mechanism that tuning the model parameters with traffic state density based on locally kernel regression. The relationship of model parameters and traffic state density is derived from a theorem which can be explained by Three-Phase Traffic Theory. Experiments show this mechanism's accuracy by comparing with Shuai Meng, etc.'s work and Global Bandwidth Selection [3].

The paper is organized as follows: following this introduction is an preliminary. Next, a theorem about the relationship between model parameters and data density are described and explained by Three-Phase Traffic Theory. Experiments are carried out in the fourth part. Section five gives a conclusion.

II. Preliminaries

A. Locally Kernel Regression

Locally kernel regression is a kind of lazy learning which has a standard regression formation:

$$y = f(x) + \varepsilon \tag{1}$$

where *x* is the input, and *y* the output. *f()* is a unknown function with various forms such as constant, linear and quadratic model. A kernel function, involved in *f()* and includes two parameters known as bandwidth h and number of neighbors

This work was supported by the National Natural Science Foundation of China under Grant No. 60703066, No. 60874082 and No. 40801046; supported by the National High Technology Research and Development Program of China (863 Key Program) No.2007AA120502.

K, calculates the weights of neighbors of input *x*, and there is an emphasized discussion about it in section III.

B. Three-Phase Traffic Theory

Three-Phase Traffic Theory claims that there are three traffic phases in traffic: free flow, synchronized flow and wide moving jam. Free flow is usually observed when the vehicle density is small enough that the intersection between vehicles can be neglected. When there is a bottleneck on a road location, such as on-ramp or decrease of lane, a synchronized flow usually formed, because vehicles arriving here have to decelerate. A wide moving jam emerges usually in a synchronized flow. In a wide moving jam, the vehicle speed is very low, while the vehicle density reaches the maximum. The synchronized flow and wide moving jam compose the traffic congestion [4].

Based on locally kernel regression and Three-Phase Traffic Theory above, the next section aims to introduce traffic state density into the locally kernel regression. When a given query is located in sparseness, the barren neighborhood makes the local regression model powerless because of small weights of neighbors if model parameters do not take any corresponding adjustments. However, there is no variable which refers to data distribution in locally regression model. In order to find the correspondences between model parameters and data distribution, the following part makes a derivation for the detection.

III. Locally Kernel Regression Adapting with Data Distribution within Traffic Background

A. Relation of Data Density and Parameters Theorem

The error criterion of regression is defined as follow:

$$C(\mathbf{q}) = \sum_{i}^{K} [(f(x_i,\beta) - y_i)^2 \cdot G(\frac{d(x_i,\mathbf{q})}{h})] \tag{2}$$

where *G()* is a kernel function to calculate the weights of neighbor points from the distance. A typical kernel function is Gaussian $G(d) = e^{-d^2}$. *d()* is the distance function. *K* is the number of neighbors that takes part in locally kernel regression. *h* is the bandwidth that determines the local regression area. β denotes a set of parameters and also includes the former two parameters *K* and *h*. **q** is a given query.

Another definition of error criterion is the directly measurement reflected in error as:

$$Error = |y - f(\mathbf{q},\beta)| \tag{3}$$

Optimal β is obtained by minimizing either of the two criterions, because the two definitions of error criterion are equivalent for planar local models [5].

Within traffic background, the regression objects are the traffic states measured by flow-occupancy points, $q_{veh} \sim \rho_{veh}$ points for short, where the occupancy denotes the vehicle density in a road. We find that when we use locally kernel regression with a set of global optimal parameters in kernel function, the prediction shows inappropriate in different traffic state distribution. For example, when we fit the parameters as a global optimal one, the locally kernel regression will perform well in free flow pattern while fails in wide moving jam pattern in which the traffic $q_{veh} \sim \rho_{veh}$ state points are very few. Thus, it requests an adaptive mechanism that can adjust the parameters in kernel function dynamically with the distribution of traffic states.

We aim to find a relationship between the model parameters and distribution of local traffic states. Data density ρ, which is an important property of data distribution, can be considered as a key value (note that this data density ρ is different from vehicle occupancy ρ_{veh}). Especially in traffic background, the state density in free flow is very high while wide moving jam contains fewer $q_{veh} \sim \rho_{veh}$ points.

In a two-dimension planar, data density can be represented as:

$$\rho = \frac{K}{S} \tag{4}$$

K is the number of neighbors in the neighborhood of a given query, assumed as a circular region, and S is the area of this circle. Then, the *K*th nearest neighbor of the query locates on the ring of this circular region whose radius is equal to the distance between the query and the *K*th nearest neighbor:

$$r = d(x_k,\mathbf{q}) \tag{5}$$

Combine equation 4 and equation 5, we have already introduced data density into locally kernel regression:

$$d(x_k,\mathbf{q}) = \sqrt{\frac{K}{\rho\pi}} \tag{6}$$

In this paper, we mainly talk about the constant model:

$$\hat{y} = f(x) = \frac{\sum_i^K y_i \cdot G(\frac{d(x_i,\mathbf{q})}{h})}{\sum_i^K G(\frac{d(x_i,\mathbf{q})}{h})} \tag{7}$$

Then the error criterion can be represented as:

$$C(\mathbf{q}) = \sum_{i}^{K}[(f(x_i,\beta) - y_i)^2 \cdot G(\sqrt{\frac{i}{\rho\pi h^2}})] \tag{8}$$

$$or$$

$$Error = |\frac{\sum_i^K (y - y_i) \cdot G(\sqrt{\frac{i}{\rho\pi h^2}})}{\sum_i^K G(\sqrt{\frac{i}{\rho\pi h^2}})}| \tag{9}$$

Theorem: By minimizing the constant model's criterion (equation.9 used here), we can finally get a theorem called Relation of Data Density and Parameters Theorem (RDPT), which has been proved in [2] or Appendix:

$$\rho = \frac{K}{\pi h^2} + \varepsilon \tag{10}$$

ε is a fluctuate value. From the theorem and a traffic background view, it is obvious that when the traffic state density decrease, in order to maintain the minimization of the criterion, we should adapt *h* to be much larger or set *K* to be much smaller to keep the equivalent of equation.10. Because parameter *K* influences the efficiency of kernel regression, too large to be time consuming and too small to be less accurate, we fix K as a constant and change h dynamically. Then, here comes a key conception that data density determines its own optimal bandwidth.

B. *Explained by Three-Phase Traffic Theory*

The theorem mentioned above is generated entirely from a mathematical derivation, and whether it is helpful for the locally kernel regression in prediction of traffic flow have to be testified by experimental studies. Before proved by real traffic data, we first use Three-Phase Traffic Theory to explain the principle of this theorem.

Kerner's Three-Phase Traffic Theory separates traffic into two general parts: free flow and congested traffic, and there are two traffic phases in congested traffic known as synchronized flow and wide moving jam. However, this does not mean that all the three traffic phases are equivalent. Traffic variables observed by sensors at road locations in each day vary every minute, and most of the time the traffic variables denote a free flow pattern. Congestion happens only at rush hours or special road locations. Therefore, a $q_{veh} \sim \rho_{veh}$ planar can clear show an obvious feature that the state points in free flow are dense while the congested state points are sparse, shown in Figure. 1. Thus, the RDPT adaptive mechanism is spontaneously suitable for traffic background.

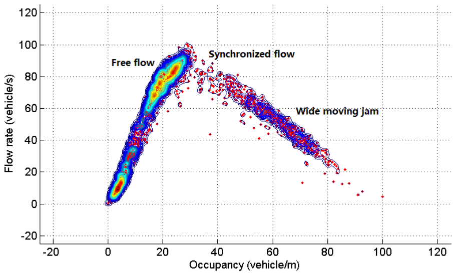

Fig. 1. $q_{veh} \sim \rho_{veh}$ planar highlighted by data density contour.

Then, it can be explained why selecting bandwidth separately in each traffic phases in Shuai Meng etc.'s work shows a good prediction result. As mentioned above, the distribution of traffic state points in each phase is different from each other. The density of points in free flow is usually much bigger than data density in congested traffic. So, the dissimilar distribution of three phases leads to different optimal bandwidth selection. Actually, to calculate optimal bandwidth separately and combine the prediction result of boundary points (traffic state points located in an intersection of different traffic phases) by weighted average is exactly a specific example of the RDPT theory. It impliedly consents the optimal bandwidths in three phases are different and does adaptive selection with traffic phases which is essentially a process adapting with density of traffic state points. The method substantially assumes that each traffic phase has a symmetrical distribution with a constant density. To adapt parameters with different phases is actually to adapt parameters with data density, and it separates the whole traffic data into three parts with different densities discretely, while RDPT theorem give a consecutive function between the data density and model parameters. Thus, adapting parameters by RDPT theorem takes more advantages than the above mechanism adapting with traffic phases theoretically.

Moreover, Global Bandwidth Selection (GBS) is a frequently-used method which calculates a global optimal bandwidth for locally kernel regression. Based on the explanation above, GBS is also a specific sample of RDPT, because it treats the whole data set with an average density and calculates the optimal bandwidth with it. Similarly, Query-Based Bandwidth Selection (QBS) and Point-Based Bandwidth Selection (PBS) are also followed RDPT theorem.

Therefore, the RDPT theorem, formalized in equation.10, is a general type for model parameters adaption in locally kernel regression.

IV. Experiments

Experimental studies described in this section are based on the Freeway Performance Measurement System (PeMS) from Berkeley University of California [6]. Traffic data are selected from Los Angeles mainline I5, segment 759700. The sensors take a record of the traffic variables every 5 minutes, thus there are 288 traffic state points in each day. Training set concludes a week's traffic data from 2008-7-23 to 2008-7-29. Query data are from 2008-7-30. Experiments are implemented in Matlab 7, and the calculation hardware is Intel Core II 2.26G CPU and 2G memory.

A. *Good matched RDPT curve*

We first verify the theorem mentioned above with real traffic data. Kernel Density Estimation (KDE) is used here to calculate the density of traffic state points [7]. Due to the noise generated by single point, training sample is separated into small groups. After calculation of the density of each point, small groups have their own average density respectively. To testify the RDPT theory, optimal parameters of each small group are calculated by a round of locally kernel regression with constant model first. Then we plot the relationship curve of optimal bandwidths and small groups' densities, $h_{opt} \sim \rho$ for short, in Figure. 2.

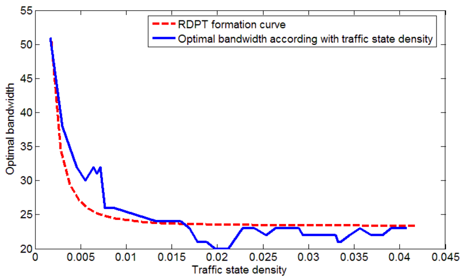

Fig. 2. $h_{opt} \sim \rho$ relation.

Figure. 2 shows a good match of the actual $h_{opt} \sim \rho$ relation and RDPT curve. K is fixed as a constant explained in section III. The blue curve denotes the relationship between the optimal bandwidth and traffic state density. Each optimal bandwidth, according with each density interval, is calculated by an ideal error minimum obtained from the ”prescient” knowledge of real output. The red dashed curve is a plot of RDPT function about bandwidths and traffic state density. The good match gives a sufficient proof of the availability of RDPT theorem.

B. Prediction using RDPT

In this part, prediction of real traffic data is required to be done by locally kernel regression adapting with RDPT rule. We first calculate the traffic state density of each point in the $q_{veh} \sim \rho_{veh}$ planar by KDE. Then, optimal bandwidths are obtained using the method mentioned in section IV.A. After the preprocessing, when a query comes, its location will be found in the planar as well as the density of the neighborhood region of this query. Consequently, each query is equipped with an according optimal bandwidth with its location neighborhood, and so the locally kernel regression can be done adaptively. This adaptive process is called RDPT Adaptive Mechanism. A comparison among RDPT Adaptive Mechanism, adaptive prediction based on Three-Phase Traffic Theory (Shuai Meng, etc.’s work, ATP for short) and Global Bandwidth Selection (GBS) is shown as follows. All of the three bandwidth selections are based on local constant model regression.

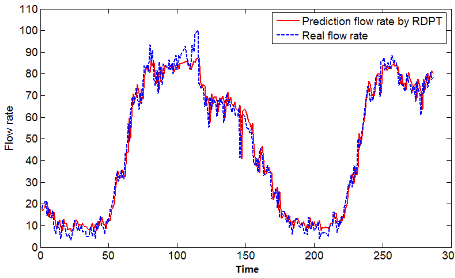

Fig. 3. Prediction of traffic flow by RDPT

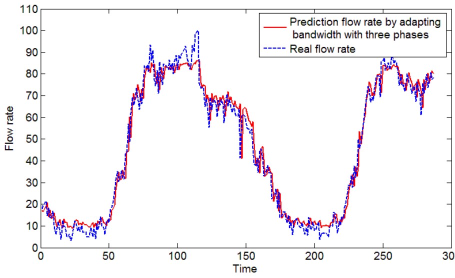

Fig. 4. Prediction of traffic flow by ATP

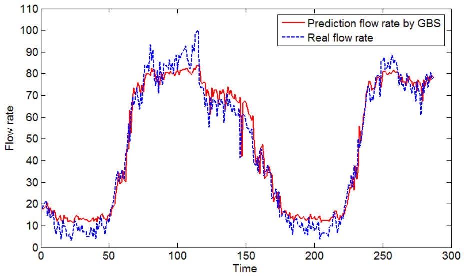

Fig. 5. Prediction of traffic flow by GBS

TABLE I
PREDICTION ERROR COMPARISON WITH CONSTANT MODEL

Error estimation	RDPT	ATP	GBS
ME	4.146	4.282	5.091
MRE	8.991%	9.287%	11.041%

The results of three prediction mechanisms are shown in figure.3, figure.4 and figure.5 separately. The blue dashed curves denote the actual traffic flow, and the red lines are the prediction values. We can find that the prediction curve done by RDPT matches the real traffic flow more accurate than the other two curves, especially in critical fluctuant phases. Table. I gives the prediction Mean Error (ME) and Mean Relative Error (MRE) comparison according with the figures. Table. II compares the time consuming. From the tables, it is obvious that RDPT Adaptive Mechanism is more superior than adaptive prediction based on Three-Phase Traffic Theory and Global Bandwidth Selection, which is corresponding to the analysis in Section III.B. The preprocessing consumed time of the three mechanism are close to each other, while the online regression consumed time in RDPT Adaptive Mechanism is considerably less than ATP mechanism when a query is given, because RDPT Adaptive Mechanism just uses the parameters according to the query’s neighborhood density obtained in the preprocessing, but the ATP mechanism has to search which neighbors of the query belong to which traffic phase, and then choose the corresponding bandwidth.

The RDPT theorem is completely derived from a constant model, so the experiments above are all based on constant model regression. Nevertheless, we find that this theorem also suits for linear model in kernel regression. The relationship curve of optimal bandwidth and traffic state density is very similar with the curves described in figure. 2. Therefore, we equip linear model with this theorem and implement regression on the same traffic data set. Table. III and table. IV give the comparison with linear model regression. Although the error differences for linear model of RDPT and ATP are slight, RDPT is less time consuming than ATP mechanism that with the same results as constant model.

From a model selection view, it can be easily found that linear model regression is more suitable for this traffic data set, while when we focus on the adaptive mechanism, both constant and linear model have the same properties for the comparison of RDPT, ATP and GBS.

TABLE II
PREDICTION TIME CONSUMING COMPARISON WITH CONSTANT MODEL

Time Consuming	Data Set Scale	RDPT	ATP	GBS
Off-line preprocessing	2016	3.05m	3.20m	2.95m
On-line regression	288	0.95s	39.01s	0.88s

TABLE III
PREDICTION ERROR COMPARISON WITH LINEAR MODEL

Error estimation	RDPT	ATP	GBS
ME	3.945	4.045	4.705
MRE	8.554%	8.773%	10.204%

V. CONCLUSION

This paper discusses a new adaptive locally kernel regression method for traffic flow prediction. The core conception of adaptation focuses on the relationship between parameters in locally kernel regression and traffic state density (traffic data point density). A theorem about this relationship is generated by mathematical derivation. Furthermore, the essential implication of this theorem is also explained by Three-Phases Traffic Theory. Experiments are based on real traffic data. The experiments first give a verification of the correctness and applicability of this theorem. Then, predictions of traffic flow based on locally kernel regression with RDPT Adaptive Mechanism, adaptive prediction based on Three-Phase Traffic Theory and Global Bandwidth Selection are implemented, and the prediction results show the superiority of our theory.

Because the theorem is derived from constant model, we mainly concentrate on the comparison of mechanisms based on constant model regression. However, by sufficient experiments, we find that the same relationship between parameters in locally kernel regression and data density maintains in linear model. The future work is to detect the characteristics of the parameter-density relationship in different models, and expect a uniform theory.

REFERENCES

[1] M. Shuai, L. Han, K.Q. Xie, G.J. Song, X.J. Ma, G.H. Chen, "An Adaptive Traffic Flow Prediction Mechanism Based on Locally Weighted Learning," ACTA SCIENTIARUM NATURALIUM UNIVERSITATIS PEKINENSIS, vol. 46(1), pp. 64-68, 2010.

[2] L. Han, K.Q. Xie, G.J. Song, "Adaptive Fit Parameters Tuning with Data Density Changes in Locally Weighted Learning," in International Symposium on Neural Networks (ISNN 2010). 2010. (Accepted to be published).

[3] S. Vijayakumar, A.D. Souza, S. Schaal, "Incremental Online Learning in High Dimensions," in Neural Computation, vol. 17. MIT Press.(2005)

[4] B.S. Kerner, "Introduction to Modern Traffic Flow Theory and Control - The Long Road to Three-Phase Traffic Theory," Springer Heidelberg Dordrecht London New York, 2009.

[5] C. Atkeson, A. Moore, S. Schaal, "Locally Weighted Learning," Artificial Intelligence Review, 11–73(1997)

[6] Freeway Performance Measurement System, available online at: http://pems.eecs.berkeley.edu

[7] Kernel Density Estimation, available online at: http://en.wikipedia.org/wiki/

APPENDIX

Proof:

TABLE IV
PREDICTION TIME CONSUMING COMPARISON WITH LINEAR MODEL

Time Consuming	Data Set Scale	RDPT	ATP	GBS
Off-line preprocessing	2016	4.06m	6.28m	4.06m
On-line regression	288	1.39s	35.60s	1.38s

Use Gaussian $G(d) = e^{-d^2}$ as the kernel function here,

$$\sum_{i=1}^{K} e^{-\frac{i}{\rho\pi h^2}} > \int_1^K e^{-\frac{i}{\rho\pi h^2}} di$$

$$= -\rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}} + \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}}$$

$$\sum_{i=1}^{K} e^{-\frac{i}{\rho\pi h^2}} < e^{-\frac{1}{\rho\pi h^2}} + \int_1^K e^{-\frac{i}{\rho\pi h^2}} di$$

$$= -\rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}} + \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}} + e^{-\frac{1}{\rho\pi h^2}}$$

Propose a variable B_1, and set it to:

$$\rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}} < B_1 < \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}} + e^{-\frac{1}{\rho\pi h^2}}$$

$$\sum_{i=1}^{K} \frac{C \cdot i \cdot e^{-\frac{i}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} >$$

$$\frac{C \cdot \rho\pi h^2}{\pi\sqrt{\rho\rho_0}} \left(\left(-K - \rho\pi h^2\right) \cdot e^{-\frac{K}{\rho\pi h^2}} + \left(1 + \rho\pi h^2\right) \cdot e^{-\frac{1}{\rho\pi h^2}} \right)$$

$$\sum_{i=1}^{K} \frac{C \cdot i \cdot e^{-\frac{i}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} <$$

$$\frac{C \cdot \rho\pi h^2}{\pi\sqrt{\rho\rho_0}} \left(\left(-K - \rho\pi h^2\right) \cdot e^{-\frac{K}{\rho\pi h^2}} + \left(1 + \rho\pi h^2\right) \cdot e^{-\frac{1}{\rho\pi h^2}} \right)$$

$$+ \frac{C \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}}$$

B_2 is proposed here like B_1 and set to:

$$\frac{C \cdot \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} \cdot \left(1 + \rho\pi h^2\right) < B_2$$

$$< \frac{C \cdot \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} \cdot \left(1 + \rho\pi h^2\right) + \frac{C \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}}$$

Thus,

$$Error = \left| \frac{\frac{C \cdot \rho\pi h^2}{\pi\sqrt{\rho\rho_0}} \cdot e^{-\frac{K}{\rho\pi h^2}} \cdot \left(-K - \rho\pi h^2\right) + B_2}{-\rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}} + B_1} \right|$$

Because the most errors occur in sparse neighborhood, we focus on the low density situation. When data density is under a threshold ρ_{low}, both B_1 and B_2 approach zero. Then,

$$Error \approx \frac{CK}{\pi\sqrt{\rho\rho_0}} + \frac{C\sqrt{\rho}h^2}{\sqrt{\rho_0}}, \quad \rho \leq \rho_{low}$$

To minimize this function, its derivative about ρ is required to equal to zero:

$$\frac{\partial Error}{\partial \rho} = -\frac{1}{2} \cdot \frac{CK}{\pi\sqrt{\rho_0}} \cdot \rho^{-\frac{3}{2}} + \frac{1}{2} \cdot \frac{Ch^2}{\sqrt{\rho_0}} \cdot \rho^{-\frac{1}{2}} = 0$$

Finally,

$$\rho = \frac{K}{\pi h^2}$$

Because the final equation is derived from a approximation, we add a fluctuate value ε to it,

$$\rho = \frac{K}{\pi h^2} + \varepsilon$$