Generative Local Metric Learning for Nadaraya-Watson Kernel Estimation

Yung-Kyun Noh[*]
Seoul National University
nohyung@snu.ac.kr

Masashi Sugiyama
The University of Tokyo
sugi@k.u-tokyo.ac.jp

Kee-Eung Kim
KAIST
kekim@cs.kaist.ac.kr

Frank C. Park
Seoul National University
fcp@snu.ac.kr

Daniel D. Lee
University of Pennsylvania
ddlee@seas.upenn.edu

1 Introduction

In kernel learning, people have been using kernel parameters such as bandwidth in Gaussian for learning representations. In this work, we consider the Nadaraya-Watson (NW) estimation, and we present, instead of selecting a simple kernel parameter, learning a metric is possible from the configuration of data, and it drastically increases the estimation performance.

In this work, we show the followings:

- For Gaussian data, the regression needs to calculate kernel with only two-dimensional information regardless of the original dimensionality, and it always obtains mean square error (MSE) close to zero with finite samples.
- Previously a well-known extension of the NW estimator, locally linear regression (LLR), is known to alleviate the asymptotic bias. The proposed metric acts differently from LLR to reduce the bias, and the proposed method highly outperforms LLR in the empirical experiments.
- In order to reduce MSE, reducing variance with metric is not important. The asymptotic contribution of the variance shows little dependency on the metric once appropriate bandwidth is selected in a high dimensional space.
- A simple and coarse generative model captures appropriate metric for nonparametric NW estimation.

The remainder shows how we derive the metric dependency from the asymptotic bias of NW estimator, a brief description of the proofs of the above results, and the experimental results with synthetic and real data.

1.1 Nadaraya-Watson Estimation and Metric Dependency

The Nadaraya-Watson (NW) estimator has been widely applied for nonparametric classification and regression, using the average output of local information around $\mathbf{x}$. Local information, typically determined by a kernel function $K(\cdot,\cdot)$: $\mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, calculates the output using the following equation:

$$\widehat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x}) y_i}{\sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x})} \tag{1}$$

[*]Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

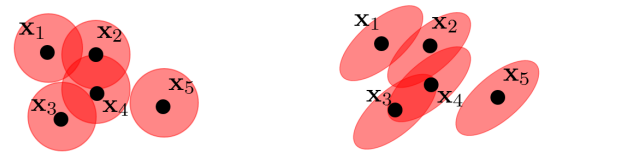


Figure 1: Metric dependency of kernels. The level curves of kernels are hyper-spheres for isotropic kernels (left), while they are hyper-ellipsoids for kernels with Mahalanobis metric as shown on the right. The principal directions of hyper-ellipsoid are the eigenvectors of the symmetric positive definite matrix A which is used in the Mahalanobis distance.

for data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{1, \ldots, C\}$ for classification and $y_i \in \mathbb{R}$ for regression. The kernel is a translation-invariant kernel $K(\mathbf{x}_i, \mathbf{x}) = K(||\mathbf{x} - \mathbf{x}_i||^2)$. This estimator provides one of the most fundamental equations for supervised learning utilizing local information, and the objective is to predict $\mathbb{E}_{p(y|\mathbf{x})}[y]$ achieving the minimum mean square error (MSE).

With many theoretical analysis of this well-known method, the Nadaraya-Watson estimator predicts the optimal result with infinite data [12, 13, 26, 28], however the estimator inevitably suffers a finite sampling effect associated with the metric dependency. We consider a Mahalanobis-type distance for metric learning. The distance between two data points $\mathbf{x}_i \in \mathbb{R}^D$ and $\mathbf{x}_j \in \mathbb{R}^D$ is defined using a symmetric positive definite matrix $A \in \mathbb{R}^{D \times D}$ as follows:

$$||\mathbf{x}_i - \mathbf{x}_j||_A = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top A(\mathbf{x}_i - \mathbf{x}_j)}, \tag{2}$$

$$A^\top = A, \quad |A| = 1, \quad A \succ 0 \tag{3}$$

where $|A|$ is the determinant of the matrix A.

A kernel function, capturing the local information, typically decays rapidly outside a certain distance; conventionally a bandwidth parameter h is used to control the effective number of data within a range of interests. If we use the Gaussian kernel as an example, with the aforementioned metric and bandwidth, the kernel function can be written as

$$K(\mathbf{x}_i, \mathbf{x}) = K\left(\frac{||\mathbf{x}_i - \mathbf{x}||_A}{h}\right) = \tag{4}$$

$$\frac{1}{\sqrt{2\pi}^D h^D} \exp\left(-\frac{1}{2h^2}(\mathbf{x}_i - \mathbf{x})^\top A (\mathbf{x}_i - \mathbf{x})\right),$$

where the relative bandwidths along individual directions are determined by A, and the overall size of the kernel is determined by h. The change of nonparametric density by adopting Mahalanobis-type distance is depicted in Fig. 1

Conventionally, adapting the size of local data has been achieved mostly by the bandwidth parameter h only. Under the MSE criterion, bandwidth selection has an intuitive explanation of the tradeoff between the bias and the variance; the variance is high with small bandwidth due to the small number of data used, while with large bandwidth, the variance becomes small but at the cost of increasing bias caused by the corruption of information at a distance.

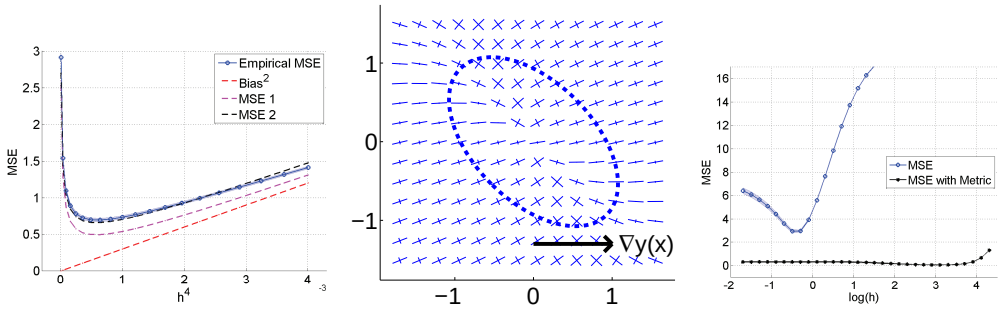
With the use of metric selection, the directions with more relevant information can be emphasized, resulting in the decrease of MSE without any increase in the amount of data. Still, the overall size of the kernel can be determined using the bandwidth selection methods in the previous studies based on the conventional bias-variance analysis.

2 Metric Effects on Bias

The bias is the expected deviation of the estimator from the true mean of the target variable $y(\mathbf{x})$:

$$\text{Bias} = \mathbb{E}\left[\widehat{y}(\mathbf{x}) - y(\mathbf{x})\right] \tag{5}$$

$$= \mathbb{E}\left[\frac{\sum_{i=1}^N K(\mathbf{x}_i, \mathbf{x}) y_i}{\sum_{i=1}^N K(\mathbf{x}_i, \mathbf{x})} - y(\mathbf{x})\right]. \tag{6}$$



(a) (b) (c)

Figure 2: Empirical results with synthetic data (a) Empirical MSE and calculated bias^2. MSE 1 is the sum of bias^2 and the leading term of the derived variance. MSE 2 is the MSE with variance up to the second leading order of the variance (order h^{2-D}). (b) Metric calculation for a Gaussian (c) Empirical MSEs with and without metric.

First, the bias can be calculated for classification ($y \in \{1, 2\}$), and it is noteworthy that the bias is minimized by the same metric minimizing the bias in nearest neighbor (NN) classification [21, 29]. Further comparison is difficult because the MSE analysis for metric learning is different from that of nearest neighbor classification [7, 8]. For classification, we do not show the analysis and empirical experiments due to the limitation of space, but similar analysis was presented in the analysis for NN classification [21].

In terms of regression with $y \in \mathbb{R}$, the bias and the variance of the NW estimator can be calculated.

$$E\left[\frac{\sum_{i=1}^N K(\mathbf{x}, \mathbf{x}_i) y_i}{\sum_{i=1}^N K(\mathbf{x}, \mathbf{x}_i)} - y(\mathbf{x})\right] \tag{7}$$
$$= h^2 \left(\frac{\nabla^\top p(\mathbf{x}) \nabla y(\mathbf{x})}{p(\mathbf{x})} + \frac{\nabla^2 y(\mathbf{x})}{2}\right) + o(h^4)$$

The derived bias along with the variance constitutes the MSE, and the sum of these two for leading orders of h fits nicely with the empirical result in a low dimensional space as shown in Fig. 2(a).

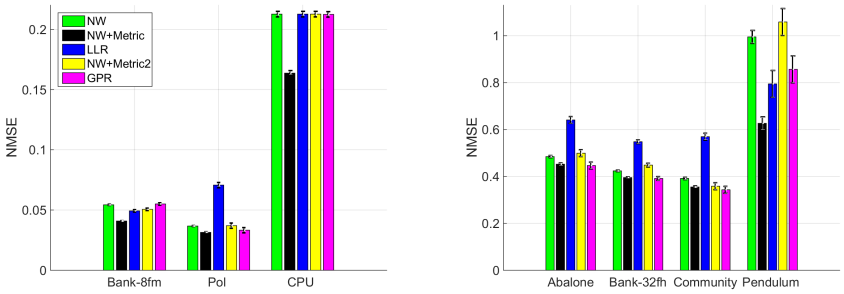
Proposition 1: *There exists a symmetric positive definite matrix A that eliminates the first term $\frac{\nabla^\top p(\mathbf{x}) \nabla y(\mathbf{x})}{p(\mathbf{x})}$ of the bias in Eq. (7), when used with the metric in Eq. (2), and when there exist two linearly independent gradients of $p(\mathbf{x})$ and $y(\mathbf{x})$, and $p(\mathbf{x})$ is away from zero.*

Proposition 1 comes from the proof that a symmetric rank-two matrix $B = \frac{1}{p(\mathbf{x})}\left[\nabla y(\mathbf{x}) \nabla^\top p(\mathbf{x}) + \nabla p(\mathbf{x}) \nabla^\top y(\mathbf{x})\right]$, with linearly independent $\nabla y(\mathbf{x})$ and $\frac{\nabla p(\mathbf{x})}{p(\mathbf{x})}$, has two eigenvalues with different signs. The metric A_{NW} is obtained from the estimation of $\nabla y(\mathbf{x})$ and $\frac{\nabla p(\mathbf{x})}{p(\mathbf{x})}$, followed by the calculation of two eigenvalues $\lambda_+ > 0$ and $\lambda_- < 0$, and their corresponding eigenvectors $\mathbf{u}_+$ and $\mathbf{u}_-$:

Algorithm 1 Metric for NW regression

$$A_{NW} = \beta [\mathbf{u}_+ \mathbf{u}_-] \begin{pmatrix} \lambda_+ & 0 \\ 0 & -\lambda_- \end{pmatrix} [\mathbf{u}_+ \mathbf{u}_-]^\top + \gamma I, \tag{8}$$
$$\text{for} \quad |A_{NW}| = 1 \tag{9}$$

Here we can choose constants β and γ for appropriate regularization. From the following corollary, the proposed metric produces low bias with target function $y(\mathbf{x})$ close to linear.



(a) (b)

Figure 3: Regression with real-world datasets. NW is the NW regression with conventionial kernels, NW+Metric is the NW regression with our metric, LLR is the locally linear regression, NW+Metric2 is the NW regression in [33], and GPR is the Gaussian process regression. Normalized MSE (NMSE) is provided.

Corollary 2: *If the regression function $y(\mathbf{x})$ is locally linear, and $\nabla y(\mathbf{x})$ and ∇p are linearly independent, the second order approximation of the bias Eq. (7) is almost zero with small γ, because $tr[A^{-1}B] = \frac{1}{\beta_2}\left(\frac{\lambda_+}{\lambda_+ + \gamma} - \frac{\lambda_-}{\lambda_- + \gamma}\right) = \frac{\gamma}{\beta_2}\left(\frac{1}{\lambda_-} - \frac{1}{\lambda_+}\right) + o(\gamma^2)$ with the metric A_{NW}.*

If $\nabla^2 y(\mathbf{x}) \simeq 0$ for all space, the NW estimator produce almost unbiased estimation with small γ by Proposition 1 and Corollary 3. The metric A_{NW} considers a two-dimensional subspace spanned by $\nabla p(\mathbf{x})$ and $\nabla y(\mathbf{x})$ as shown in Fig. 2(b). Only components in this two-dimensional space affect the bias, and as long as the points are close within this subspace, the bias can still be very *small* even when the information is used from distant points. In this way, we can include more statistically relevant data than using the conventional isotropic kernel which extracts data uniformly in every direction.

The proposed method is in contrast to LLR which eliminates the second term of Eq. (7) in the bias analysis [26]. LLR is a straightforward well-known extension of Nadaraya-Watson estimation adopting more parameters for MSE minimization. LLR eliminates the second Laplacian term in the bias, whereas our model assumes a prediction function close to linear, resulting in a small Laplacian, and we try to minimize the first term. The empirical result is provided in Fig. 3.

In particular, when the joint probability density of $\mathbf{x}$ and y is Gaussian, the prediction function $y(\mathbf{x})$ is linear, and simply the regression with Nadaraya-Watson estimator achieves a very small bias. In this case, we can use a large bandwidth for balancing the bias and the variance as in Fig. 2. In addition , the following Proposition 3 shows that only minimizing the bias is important in high dimensional space with optimal bandwidth.

Proposition 3: *The asymptotic sum of bias and variance can be simplified as follows where C_1 and C_2 are the terms including no h:*

$$f(h) = h^4 C_1 + \frac{1}{Nh^D} C_2. \tag{10}$$

By plugging in the asymptotic optimal bandwidth, $h_* = N^{-\frac{1}{D+4}}\left(\frac{D \cdot C_2}{4 \cdot C_1}\right)^{\frac{1}{D+4}}$, the sum has the minimum $f(h_*) = C_1$ in the limit with infinite dimensionality D.

In the infinite dimensional space with $D \gg N$, minimizing bias only reduces the MSE. In practice, reducing the bias let the optimal bandwidth increases, and the variance can be reduced together along with bias with the large optimal bandwidth. In real experiments, variance itself is also reduced with our metric, and the result is robust to bandwidth selection as shown in Fig. 2(c).

For metric calculation, we used $\nabla y(\mathbf{x}) = \widehat{\Sigma}_{\mathbf{x}}^{-1}\widehat{\Sigma}_{\mathbf{x}y}$ and $\frac{\nabla p(\mathbf{x})}{p(\mathbf{x})} = -\widehat{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x} - \widehat{\mu}_{\mathbf{x}})$ from the estimated covariance parameters $\widehat{\Sigma}_{\mathbf{x}}$, $\widehat{\Sigma}_{\mathbf{x}y}$, and the estimated mean $\widehat{\mu}_{\mathbf{x}}$ with all data. With this simple and coarse estimation of the gradients, the experiments show substantial decrease of errors as in Fig. 3.

References

[1] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709, 2013.

[2] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:790–799, 1995.

[3] E. Choi, P. Hall, and V. Rousson1. Data sharpening methods for bias reduction in nonparametric regression. *Annals of Statistics*, 28(5):1339–1355, 2000.

[4] T. Cover. Estimation by the nearest neighbor rule. *Information Theory, IEEE Transactions on*, 14(1):50–55, January 1968.

[5] J.V. Davis, B. Kulis, P. Jain, S. Sra, and I.S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, 2007.

[6] J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87:998–1004, 1992.

[7] K. Fukunaga and T.E. Flick. The optimal distance measure for nearest neighbour classification. *IEEE Transactions on Information Theory*, 27(5):622–627, 1981.

[8] K. Fukunaga and T.E. Flick. An optimal global nearest neighbour measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(3):314–318, 1984.

[9] K. Fukunaga and Larry D. H. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21:32–40, 1975.

[10] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. 2005.

[11] P. Hall, S. J. Sheather, M. C. Jones, and J. S. Marron. On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78:263–269, 1991.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[13] S. Haykin. *Neural Networks and Learning Machines (3rd Edition)*. Prentice Hall, 3 edition, 2008.

[14] R. Huang and S. Sun. Kernel regression with sparse metric learning. *Journal of Intelligent and Fuzzy Systems*, 24(4):775–787, 2013.

[15] M. A. Jacome, I. Gijbels, and R. Cao. Comparison of the nadaraya-watson and local linear methods in presmoothed density estimation under censoring. *Computational Statistics*, 23(3):381–406, 2008.

[16] P. W. Keller, S. Mannor, and D. Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 449–456, New York, NY, USA, 2006. ACM.

[17] S. Kpotufe. k-nn regression adapts to local intrinsic dimension. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 729–737. Curran Associates, Inc., 2011.

[18] S. Kpotufe and Abdeslam B. Gradient weights help nonparametric regressors. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2861–2869. Curran Associates, Inc., 2012.

[19] M. Lazaro-Gredilla and A. R. Figueiras-Vidal. Marginalized neural network mixtures for large-scale regression. *IEEE Transactions on Neural Networks*, 21(8):1345–1351, 2010.

[20] Y.-K. Noh, M. Sugiyama, S. Liu, M. C. du Plessis, F. C. Park, and D. D. Lee. Bias reduction and metric learning for nearest-neighbor estimation of kullback-leibler divergence. In *Proceedings of Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS2014), JMLR Workshop and Conference Proceedings*, pages 669–677, 2014.

[21] Y.-K. Noh, B.-T. Zhang, and D. D. Lee. Generative local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems 23*, pages 1822–1830. 2010.

[22] B. U. Park and J. S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85:66–72, 1990.

[23] B. U. Park and B. A. Turlach. Practical performance of several data driven bandwidth selectors. *Computational Statistics*, 7:251–270, 1992.

[24] E. Parzen. On estimation ofa probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[25] J. Roll, A. Nazin, and L. Ljung. A non-asymptotic approach to local modelling. In *Proceedings of the 41st IEEE Conference on Decision and Control*, volume 1, pages 638 – 643, 2002.

[26] D. Ruppert and M. P. Wand. Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, 22(3):1346–1370, September 1994.

[27] W. R. Schucany and John P. Sommers. Improvement of kernel type density estimators. *Journal of the American Statistical Association*, 72:420–423, 1977.

[28] L. Shi, T. L. Griffiths, N. H. Feldman, and A. N. Sanborn. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic bulletin & review*, 17(4):443–464, 2010.

[29] R. R. Snapp and S. S. Venkatesh. Asymptotic expansions of the k nearest neighbor risk. *Annals of Statistics*, 26(3):850–878, 1998.

[30] H. Takeda, S. Member, S. Farsiu, P. Milanfar, and S. Member. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16:349–366, 2007.

[31] M. Titsias and M. Lazaro-Gredilla. Variational inference for mahalanobis distance metrics in gaussian process regression. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 279–287. Curran Associates, Inc., 2013.

[32] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 18*, pages 1473–1480. 2006.

[33] K.Q. Weinberger and G. Tesauro. Metric learning for kernel regression. In *Eleventh international conference on artificial intelligence and statistics*, pages 608–615, 2007.