From Kernel Regression to Attention Mechanisms:
A Six Decade Journey (1960–2020)

Peyman Milanfar

Based on the presentation dated October 2025

Abstract

This document outlines the conceptual evolution from classical kernel regression methods in statistics to the attention mechanisms foundational to modern machine learning. It traces a path through three major periods: Kernel Regression (1960s–1990s), Data-Adaptive Filters in signal processing (1990s–2010s), and Attention Mechanisms in machine learning (2010s–2020s) [1].

1 Kernel Regression (Nadaraya-Watson, 1964)

The journey begins with kernel regression, a non-parametric method to fit a smooth curve to data points by modeling a nonlinear relationship [2]. At any query position x, the value $\hat{y}(x)$ is estimated as a weighted average of observed data y_i. The weights are determined by a kernel function, $K(x, x_i)$, which measures the proximity between the query point x and each data point x_i. The Nadaraya-Watson estimator is given by:

$$\hat{y}(x) = \frac{\sum_i K(x, x_i) y_i}{\sum_i K(x, x_i)} \tag{1}$$

A common choice for the kernel is the Gaussian (or Radial Basis Function) kernel:

$$K(x_i, x_j) = \exp\left\{ \frac{-\|x_i - x_j\|^2}{h_x^2} \right\} \tag{2}$$

where h_x is a bandwidth parameter controlling smoothness.

2 Evolution in Signal Processing: Data-Adaptive Filters

The concepts from kernel regression were extended in signal and image processing to create powerful data-adaptive filters. These filters generalize the kernel to consider not just spatial distance but also similarity in data values (e.g., pixel intensities).

2.1 Bilateral Filter (Tomasi & Manduchi, 1998)

The Bilateral Filter extends the kernel to operate on both the positions (domain) and the values (range) of pixels, allowing it to smooth images while preserving sharp edges [3]. The kernel includes a term for value similarity:

$$K(x_i, x_j, y_i, y_j) = \exp\left\{ \frac{-\|x_i - x_j\|^2}{h_x^2} - \frac{(y_i - y_j)^2}{h_y^2} \right\} \tag{3}$$

2.2 Non-local Means (Buades, Coll, & Morel, 2005)

Non-local Means further generalizes this idea by comparing entire patches of pixels rather than individual pixel values, making it more robust to noise [4]. The kernel is defined as:

$$K(x_i, x_j, \mathbf{p}_i, \mathbf{p}_j) = \exp\left\{ \frac{-\|x_i - x_j\|^2}{h_x^2} - \frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{h_y^2} \right\} \tag{4}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ represent the image patches centered at positions x_i and x_j.

2.3 Locally Adaptive Regression Kernel (LARK)

LARK, developed by Takeda et al. [5] and related to the work of Sochen et al. [6], introduced a "Learned Metric" where the kernel's shape adapts to the local data structure:

$$K(x_i, x_j, y) = \exp\left\{-(x_i - x_j)^T \mathbf{C}_{ij}(y)(x_i - x_j)\right\} \tag{5}$$

Here, $\mathbf{C}_{ij}(y)$ is a matrix capturing the local data geometry.

3 A Unified Framework

These methods can be viewed under a single framework by defining an augmented variable t combining position and value information. The generalized kernel is written in a quadratic form:

$$K(t_i, t_j) = \exp\left\{-(t_i - t_j)^T \mathbf{Q}_{i,j}(t_i - t_j)\right\} \tag{6}$$

where $\mathbf{Q}$ is a block matrix $\mathbf{Q} = \mathrm{diag}(\mathbf{Q}_x, \mathbf{Q}_y)$ that changes form for each method. This unified perspective is detailed in Milanfar (2013) [7].

4 From Kernels to Attention Mechanisms

The final leap in this evolution is the attention mechanism, which can be seen as a learned, generalized form of the Nadaraya-Watson estimator [1]. The core components are vectors representing a **Query** (Q), a **Key** (K), and a **Value** (V).

The mechanism computes an output for each query by taking a weighted sum of all values. The weights are derived from a compatibility function, or similarity score, between the query and each key. The structure illustrated in the presentation is a direct parallel to the kernel regression formula (1). The output $\hat{y}_i$ for a given query q_i is computed as a normalized weighted sum over all values v_j:

$$\hat{y}_i = \frac{\sum_j \exp(\mathrm{score}(q_i, k_j)) \cdot v_j}{\sum_j \exp(\mathrm{score}(q_i, k_j))} \tag{7}$$

This is the structure of the softmax function applied to the raw similarity scores. Here, the kernel $K(x, x_i)$ from Equation (1) is effectively replaced by a learned similarity function passed through an exponential, $\exp(\mathrm{score}(q_i, k_j))$. In modern architectures like the Transformer, this score is typically the scaled dot-product: $\mathrm{score}(q_i, k_j) = q_i^T k_j / \sqrt{d_k}$. The query, key, and value vectors are themselves learned projections of input data, allowing the model to dynamically "attend" to the most relevant information.

References

[1] P. Milanfar, "From Kernel Regression to Attention Mechanisms," Google, October 2025. (Source Presentation)

[2] E. A. Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141-142, 1964.
G. S. Watson, "Smooth regression analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359-372, 1964.

[3] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Sixth International Conference on Computer Vision*, pp. 839-846, 1998.

[4] A. Buades, B. Coll, and J-M. Morel, "A non-local algorithm for image denoising," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 60-65, 2005.

[5] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349-366, 2007.

[6] N. Sochen, R. Kimmel, and R. Malladi, "A general framework for low level vision," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 310-318, 1998.

[7] P. Milanfar, "A Tour of Modern Image Filtering," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 106-128, Jan. 2013.