Lecturer: Kiran Vodrahalli

Contents

1 Introduction 1

2 Rate-Distortion Basics 2
2.1 Definitions . . . 2
2.2 Gaussian Source . . . 3
2.2.1 Sphere-Packing Intuition . . . 4
2.3 Proving Information Rate Distortion = Rate Distortion . . . 4
2.3.1 Convexity of $R(D)$ . . . 4
2.3.2 Converse Statement . . . 4

3 Optimizing the Rate-Distortion Function 5
3.1 Blahut-Arimoto Algorithm . . . 6

4 Connecting Rate-Distortion Theory and PCA 6

5 Application to Nonparametric Statistics 7
5.1 Minimax Framework . . . 7
5.2 Space Resource Constraints as Rate-Distortion . . . 8
5.3 Data Generation Model and Assumptions . . . 9
5.4 Main Result . . . 11

1 Introduction

What is rate-distortion theory? Essentially, it is the idea that drives lossy compression. A finite representation of a continuous random variable can never be perfect. It is thus necessary to define the "goodness" of representations of an information source. The basic problem can be stated as follows: Given a source distribution and a distortion measure (distance between random variable and its representation), what is the minimum expected distortion achievable at any given rate, the number of bits of information per symbol sent.

We can think of rate-distortion theory in terms of quantizing a random variable. For instance, given two points, we can optimally represent a Gaussian by assigning each point to the conditional mean of each half-plane. With more (R) points the problem becomes non-trivial, and Lloyd's algorithm must be applied.

An interesting result of rate distortion theory shows that if you try to represent n i.i.d. random variables with nR bits, it turns out that representing all the variables together is more efficient than treating each representation problem independently: We represent such a sequence with a single index taking on 2^{nR} values. This surprising result is due to the intrinsic geometry of the situation. It turns out to be easier to "pack information" when using symbols to express information from multiple independent variables together rather than separately.

Note: A good deal of the remaining discussion is taken from Elements of Information Theory, by Cover and Thomas (2006).

2 Rate-Distortion Basics

2.1 Definitions

Suppose we have a source producing a sequence $X_1, \cdots, X_n$ i.i.d. according to $p(X)$. We have an encoder which describes the source sequence X^n with an index $f_n(X^n) \subset \{1, \cdots, 2^{nR}\}$. Then the decoder maps back to X^n space with an estimate $\hat{X}^n$.

Definition 2.1. Distortion.
A distortion function is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$$

It is the cost of representing symbol x by $\hat{x}$. We often restrict consideration to distortion measures which are bounded. The **Hamming** distortion is $0 - 1$ indicator error, and the **squared-error** is $d(x, \hat{x}) = (x - \hat{x})^2$. Distortion between a sequence of symbols is just the average distortion per symbol.

Definition 2.2. $(2^{nR}, n)$-rate distortion code.
We have an encoding function

$$f_n : \mathcal{X}^n \to \{1, \cdots, 2^{nR}\}$$

and a decoding function

$$g_n : \{1, \cdots, 2^{nR}\} \to \hat{X}^n$$

with distortion

$$D = \mathbf{E}_{p(x)}[d(X^n, g_n(f_n(X^n)))]$$

We can think of g as partitioning the output space into a "codebook" — this is the discretization of the output. By looking at what maps to each value of the codebook (i.e. $f^{-1}(1), f^{-2}(2), \cdots, f^{-1}(2^{nR})$, we can find the "assignment regions": what things do we approximate with each code word?

A rate distortion pair (R, D) is **achievable** if there exists a sequence of $(2^{nR}, n)$-rate distortion codes (f_n, g_n) such that as $n \to \infty$, the expected distortion is $\leq D$.

If we plot R as a function of D, this describes both a region called the rate-distortion region (everything above the R-D function) as well as the function itself. The distortion-rate function describes D as a function of R. It turns out these approaches are the same, and also that this function is convex.

Definition 2.3. Information Rate-Distortion.

$$R^{(I)}(D) = \min_{p(\hat{x}|x) : \sum_{(\hat{x},x)} p(\hat{x},x) d(\hat{x},x) \leq D} I(X; \hat{X})$$

recalling that the **mutual information** is defined as

$$I(X;\hat{X}) = \sum_{x,\hat{x}} p(x,\hat{x}) \log \frac{p(x,\hat{x})}{p(x)p(\hat{x})} = \mathcal{D}\left(p(x,\hat{x}) \| p(x)p(\hat{x})\right)$$

where $\mathcal{D}(\cdot\|\cdot)$ is the KL-divergence. We can express the rate-distortion notion a different way with this formulation. It is not obvious how this notion relates to the first description of rate-distortion, but it turns out it is equivalent. The idea is that while staying above some threshold of error, we want the input and output sequences to be far apart as possible, i.e. stretch the limits of lossiness to get as efficient a rate as possible.

Henceforth, we will assume that $R^{(I)}(D) = R(D)$. We will prove it a bit later today.

2.2 Gaussian Source

Theorem 2.4. *The rate-distortion function for $\mathcal{N}(0,\sigma^2)$ source with squared-error distortion is given by*

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & \text{if } 0 \le D \le \sigma^2 \\ 0 & \text{if } D > \sigma^2 \end{cases}$$

Proof. We have

$$R(D) = \min_{f(\hat{x}|x):\mathbf{E}[(\hat{X}-X)^2]\le D} I(X;\hat{X})$$

We proceed by lower-bounding and then demonstrating achievablility. Recall that $h(X) = -\int_X f(x)\log f(x)dx$.

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) \\ &= \frac{1}{2}\log(2\pi e)\sigma^2 - h(X-\hat{X}|\hat{X}) \\ &\ge \frac{1}{2}\log(2\pi e)\sigma^2 - h(X-\hat{X}) \\ &\ge \frac{1}{2}\log(2\pi e)\sigma^2 - h(\mathcal{N}(0,\mathbf{E}[(X-\hat{X})^2])) \\ &= \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)\mathbf{E}[(X-\hat{X})^2] \\ &\ge \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)D \\ &= \frac{1}{2}\log\frac{\sigma^2}{D} \end{aligned} \tag{1}$$

where we have used the fact that the maximum entropy distribution for fixed second moment is a Gaussian. It turns out we can achieve this lower bound with the following setup. If $D > \sigma^2$, we choose $\hat{X} = 0$ and achieve $R(D) = 0$. Otherwise, we have $X = \hat{X} + Z$ where $\hat{X} \sim \mathcal{N}(0, \sigma^2 - D)$ and $Z \sim \mathcal{N}(0, D)$, with $\hat{X}$ and Z independent. The mutual information of this channel matches the lower bound. The rate distortion function decreases steeply until $D = 1$, where it becomes 0. □

2.2.1 Sphere-Packing Intuition

Another way to intuitively see this result is in terms of sphere-packing.

Consider a Gaussian source with variance σ^2. Then a $(2^{nR}, n)$ rate distortion code with distortion D is a set of $M = 2^{nR}$ sequences in $\mathbb{R}^n$. Source sequences of length n lie within a ball of radius $\sqrt{n\sigma^2}$, since the total variance is $n\sigma^2$. Then, by the definition of distortion, each sequence is within $\sqrt{D}$ of some codewords. We want to know how many codewords we can use without intersection in the decoding sphere – i.e., a sphere packing problem! We look at the volume ratio:

$$M = \frac{A_n\left(\sqrt{n\sigma^2}\right)^n}{A_n\left(\sqrt{nD}\right)^n} = \left(\frac{\sigma^2}{D}\right)^{n/2}$$

and we again recover the rate

$$\frac{1}{n}\log M = \frac{1}{2}\log\frac{\sigma^2}{D}$$

2.3 Proving Information Rate Distortion = Rate Distortion

2.3.1 Convexity of $R(D)$

In this section we show that $R(D)$ is a nonincreasing convex function of D.

Proof. First note that $R(D)$ is a minimum of mutual information over increasingly large sets as D increases. The minimum over A is less than or equal to the minimum over B if $B \subseteq A$ since there are more options. Thus $R(D)$ is nonincreasing.

Now consider (R_1, D_1) and (R_2, D_2) which lie on the R-D curve. Let $p_1(x, \hat{x})$ and $p_2(x, \hat{x})$ be the joint distributions which achieve each of these pairs and consider $p_\lambda = \lambda p_1 + (1-\lambda)p_2$. Since distortion is a linear function in terms of the distribution, we have $D(p_\lambda) = \lambda D(p_1) + (1-\lambda)D(p_2)$. Mutual information is a convex function of the conditional distribution, thus

$$I_{p_\lambda}(X;\hat{X}) \leq \lambda I_{p_1}(X;\hat{X}) + (1-\lambda)I_{p_2}(X;\hat{X})$$

Thus by definition of R-D:

$$\begin{aligned} R(D_\lambda) &\leq I_{p_\lambda}(X;\hat{X}) \\ &\leq \lambda I_{p_1}(X;\hat{X}) + (1-\lambda)I_{p_2}(X;\hat{X}) \\ &= \lambda R(D_1) + (1-\lambda)R(D_2) \end{aligned} \tag{2}$$

as desired. □

2.3.2 Converse Statement

We will now prove that we cannot achieve distortion $\leq D$ if we describe X at rate less than $R(D)$. This is the converse of the statement that

Theorem 2.5. *The rate distortion function for an i.i.d. source X with distribution $p(x)$ and bounded distortion $d(x, \hat{x})$ is equal to the associated information rate distortion function. That is,*

$$R(D) = R^{(I)}(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X;\hat{X})$$

is the ***minimum achievable rate*** *at distortion* D.

Proof. We must show for any X drawn according to $p(x)$ with $d(x, \hat{x})$ and any $(2^{nR}, n)$ rate distortion code with distortion $\leq D$ that $R \geq R(D)$. Consider any $(2^{nR}, n)$ code defined by f_n, g_n. Then $\hat{X}^n = \hat{X}^n(X^n) = g_n(f_n(X^n))$ be the reproduced sequence according to X^n, and assume that the expected distortion is $\geq D$. Then we have

$$\begin{aligned}
nR &\geq H(f_n(X^n)) \text{ by the range of } f_n \leq 2^{nR} \\
&\geq H(f_n(X^n)) - H(f_n(X^n)|X^n) = I(X^n; f_n(X^n)) \\
&\geq I(X^n; \hat{X}^n) \text{ by data-processing inequality} \\
&= H(X^n) - H(X^n|\hat{X}^n) \\
&= \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \text{ by indep.} \\
&= \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n, X_{i-1}, \cdots, X_1) \text{ by chain rule of entropy} \\
&\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) \text{ since conditioning reduces entropy} \\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i) \\
&\geq \sum_{i=1}^{n} R(\mathbf{E}[d(X_i, \hat{X}_i)]) \text{ by definition of rate-distortion} \\
&= n\left(\frac{1}{n}\sum_{i=1}^{n} R(\mathbf{E}[d(X_i, \hat{X}_i)])\right) \\
&\geq n \cdot R\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[d(X_i, \hat{X}_i)]\right) \text{ by convexity of R and Jensen} \\
&= nR(D)
\end{aligned} \tag{3}$$

thus we get $R \geq R(D)$, as desired.

□

It also turns out that this rate is achievable, but we will not prove that here.

3 Optimizing the Rate-Distortion Function

We want to optimize

$$R(D) = \min_{q(\hat{x}|x): \sum_{(x,\hat{x})} p(x)q(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X})$$

over conditional distributions $q(\hat{x}|x)$. By applying the method of Lagrange multipliers, we write down the functional

$$J(q) = \sum_{x}\sum_{\hat{x}} p(x)q(\hat{x}|x)\log\frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda\sum_{x}\sum_{\hat{x}} p(x)q(\hat{x}|x)d(x,\hat{x}) + \sum_{x}\nu(x)\sum_{\hat{x}} q(\hat{x}|x)$$

where the first constraint is the distortion constraint and the second constraint enforces that for each x, $q(\hat{x}|x)$ is a probability distribution. Taking the derivative and setting equal to zero, we get

$$p(x)\left[\log\frac{q(\hat{x}|x)}{q(\hat{x})}+\lambda d(x,\hat{x})+\frac{\nu(x)}{p(x)}\right]=0$$

Let $\log\mu(x)=\nu(x)/p(x)$. Then

$$q(\hat{x}|x)=\frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)}=\frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\tilde{x}}q(\tilde{x})e^{-\lambda d(x,\tilde{x})}}$$

Since $q(\cdot|\cdot)$ is a probability distribution, the normalization $\mu(x)=\sum_{\hat{x}}q(\hat{x})e^{-\lambda d(x,\hat{x})}$. Since $q(\hat{x})=\sum_x p(x)q(\hat{x}|x)$, and $q(\hat{x})>0$, we can obtain conditions for the minimum:

$$\sum_x\frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\tilde{x}}q(\tilde{x})e^{-\lambda d(x,\tilde{x})}}=1 \text{ if } q(\hat{x})>0$$

and ≤ 1 if $q(\hat{x})=0$.

3.1 Blahut-Arimoto Algorithm

We want to rewrite the optimization problem as a minimization of the distance between two convex sets which correspond to probability distributions. It turns out that an alternating minimization procedure will in fact converge to the minimum relative entropy between two sets of distributions.

This alternating minimization, when applied to rate-distortion theory, goes by the name of the Blahut-Arimoto algorithm. We can rewrite

$$R(D)=\min_{r(\hat{x})}\min_{q(\hat{x}|x):\sum_{(x,\hat{x})}p(x)q(\hat{x}|x)d(x,\hat{x})\leq D}\sum_x\sum_{\hat{x}}p(x)q(\hat{x}|x)\log\frac{q(\hat{x}|x)}{r(\hat{x})}$$

Then we calculate $r(\hat{x})$ which minimizes mutual information, which is given by

$$r^*(\hat{x})=\sum_x p(x)q(\hat{x}|x)$$

In the next iteration, we minimize over $q(\cdot|\cdot)$:

$$q^*(\hat{x}|x)=\frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\tilde{x}}r(\tilde{x})e^{-\lambda d(x,\tilde{x})}}$$

and then minimize over r again, etc. It turns out that the limit of the process is $R(D)$.

4 Connecting Rate-Distortion Theory and PCA

We first show the connection between rate-distortion and PCA. Suppose we have d independent Gaussian random variables with square-distance distortion, where we have $X_i\sim$

$\mathcal{N}(0, \sigma_i^2)$. Consider that for a random Gaussian vector $X \in \mathbb{R}^d$ with zero mean and covariance Σ that we can find a similarity transformation U to a diagonal Gaussian Z with covariance Γ with $U^T \Sigma U = \Gamma$. Note that $h(X) = h(Z)$.

The rate-distortion in this case is

$$\begin{aligned} I(X^d; \hat{X}^d) &= h(X^d) - h(X^d | \hat{x}^d) \\ &= \sum_{i=1}^{d} h(X_i) - \sum_{i=1}^{d} h(X_i | X^{i-1}, \hat{X}^d) \\ &\geq \sum_{i=1}^{d} h(X_i) - \sum_{i=1}^{d} h(X_i | \hat{X}_i) \\ &= \sum_{i=1}^{d} I(X_i; \hat{X}_i) \\ &\geq \sum_{i=1}^{d} \frac{1}{2} \log \frac{\sigma_i^2}{D_i} \end{aligned} \tag{4}$$

where $\sum_i D_i = D$. After writing down the Lagrangian with constraint term $\lambda \sum_i D_i$, we get that minimizing with respect to the D_i gives $D_i = \min(\lambda, \sigma_i^2)$ where λ is chosen to ensure that $\sum_i D_i = D$.

First, we note that using PCA on the sequence of data points as a method of dimensionality reduction takes into account the information of the whole sequence, as we have noted rate-distortion theory suggests we do. In rate distortion theory, despite the independence of separate random variables, it is advantageous to represent a whole sequence at once.

The second connection comes from the methodology. In rate-distortion theory, the optimal rate given a distortion comes from thresholding the eigenvalues: We essentially zero out the rate for components with small variance. This methodology is similar to the way components are selected in PCA when we enforce low-dimensional constraints. The low-dimensional contraints of PCA are equivalent to setting a variance threshold equal to the sum of a subset of the variances in sorted order. We completely zero out all other directions to get the optimal PCA subspace.

5 Application to Nonparametric Statistics

The paper "Quantized Minimax Estimation over Sobolev Ellipsoids" by Zhu and Lafferty (https://arxiv.org/pdf/1503.07368.pdf) introduces a minimax framework for nonparametric estimation under storage constraints. First we briefly describe the minimax framework.

5.1 Minimax Framework

First, we describe the parametric setting. Suppose we are trying to estimate a parameter $\theta \in \Theta$ given some noisy data $x \in X$ using $\mathbf{P}\{X|\theta\}$. We would like to find a good estimator $h : X \to \Theta$ which estimates θ, which minimizes some risk function $R(h, \theta) = \mathbf{E}_{X|\theta}[L(h(X), \theta)]$

for some loss function L. Then, the **minimax** estimator h^M is the one which minimizes risk for the worst possible θ.

Definition 5.1. Minimax Estimator.
We define the minimax estimator $h^M : X \to \Theta$ as the estimator which satisfies

$$\sup_{\theta\in\Theta} R(h^M, \theta) = \inf_h \sup_{\tilde{\theta}\in\Theta} R(h, \tilde{\theta})$$

Notably, if we take a prior on θ, the minimax estimator can be thought of as the best estimator under the "most difficult" prior on θ.

Additionally, this definition need not assume a parametrization of the hypothesis class. We simply replace all θ with $f \in \mathcal{F}$, a function class. We will use this nonparametric formulation of minimax risk throughout the rest of the section.

5.2 Space Resource Constraints as Rate-Distortion

It is interesting to consider adding additional constraints to this problem, for instance, on the amount of space (number of bits) available for representing a given estimator $\hat{f}(x)$ given data x. The paper focuses on the setting of nonparametric regression under smoothness assumptions and characterizes how the excess risk depends on the storage constraint B_n, where n is the number of data points in the data x. Let us now make this formal.

Suppose $(X_1, \cdots, X_n) \in \mathcal{X}^n$ is a random vector drawn from distribution P_n. Consider estimating functional $\theta \in \Theta$ which is evaluated on the data. Let $\left\{\hat{\theta}_n : \mathcal{X}^n \to \Theta\right\}$ be the set of possible estimators. Then the minimax L_2 risk is given by

$$R_n = \inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbf{E}_\theta[\|\theta - \hat{\theta}_n\|^2]$$

Now suppose there is a prior $\pi(\theta)$ supported on a subset of Θ. We have

$$\begin{aligned}\sup_{\theta\in\Theta} \mathbf{E}_\theta[\|\theta - \hat{\theta}_n\|^2] &\geq \int_\Theta \mathbf{E}_\theta[\|\theta - \hat{\theta}_n\|^2] d\pi(\theta) \\ &\geq \int_\Theta \mathbf{E}_\theta[\|\theta - \mathbf{E}_{\pi(\theta)}[\theta|X_{1:n}]\|^2] d\pi(\theta)\end{aligned} \tag{5}$$

where it is well known that the posterior mean is the minimizer. Since the best estimator $\hat{\theta}_n$ is the posterior mean, let us define this lower bound as

$$R_n(\Theta; \pi) = \int_\Theta \mathbf{E}_\theta[\|\theta - \mathbf{E}_{\pi(\theta)}[\theta|X_{1:n}]\|^2] d\pi(\theta)$$

We have

$$R_n(\Theta) = \sup_\pi R_n(\Theta; \pi)$$

So in other words, our tightest lower bound comes from the "worst" prior distribution.

Now rate-distortion comes into the picture, as we desire to add a constraint B_n on the storage of our estimate $\hat{\theta}_n$. We allow for an encoder-decoder pair (ϕ_n, ψ_n) with $\phi_n : \mathcal{X}^n \to$

$\{1, 2, \cdots, 2^{B_n}\}$ and $\psi_n : \{1, 2, \cdots, 2^{B_n}\} \to \Theta$ where $\{1, 2, \cdots, 2^{B_n}\}$ are the indices of our codebook.

This encoding achieves the required space constraints since for a given $X_{1:n}$ we only need to store a B_n-bit-long index, and $\hat{\theta}_n = \psi_n \circ \phi_n$. Then the minimax risk including quantization is given by

$$R_n(\Theta, B_n) = \inf_{\hat{\theta}_n, C(\hat{\theta}_n) \le B_n} \sup_{\theta \in \Theta} \mathbf{E}_\theta[\|\theta - \hat{\theta}_n\|^2]$$

The added constraint on $\hat{\theta}_n$ means that we will not be able to optimize as well for a given θ, and thus we will experience some extra loss. We have

$$\begin{aligned}\int_\Theta \mathbf{E}_\theta[\|\theta - \hat{\theta}_n\|^2] d\pi(\theta) &\ge \mathbf{E}[\|\theta - \mathbf{E}_\theta[\theta|X_{1:n}] + \mathbf{E}_\theta[\theta|X_{1:n}] - \hat{\theta}_n\|^2] \\ &= \mathbf{E}[\|\theta - \mathbf{E}_\theta[\theta|X_{1:n}]\|^2] + \mathbf{E}[\|\mathbf{E}_\theta[\theta|X_{1:n}] - \hat{\theta}_n\|^2] \\ &= R_n(\Theta; \pi) + \text{RD-risk}(\Theta, B_n)\end{aligned} \tag{6}$$

where the expectations are over the joint distribution of $\theta \sim \pi(\theta)$ and the data distribution $X_{1:n}|\theta$. The given result follows since the inner product between the two terms is zero because θ and $\hat{\theta}_n$ are conditionally independent given $X_{1:n}$ and $\mathbf{E}_\theta[\theta - \mathbf{E}_\theta[\theta|X_{1:n}]|X_{1:n}] = 0$. Note that the first term is the Bayes risk and the second term is the **rate-distortion risk**, the error due to quantization.

Now we seek to control RD-risk(Θ, B_n). Since our constrained estimator $\hat{\theta}_n$ uses at most B_n bits, we now have the following chain of inequalities:

$$B_n \ge H(\hat{\theta}_n) \ge H(\hat{\theta}_n) - H(\hat{\theta}_n|\theta_n^*) = I(\hat{\theta}_n; \theta_n^*) \tag{7}$$

where we use θ_n^* to denote the optimal estimate of θ if there were no storage constraint (which is just the posterior mean). Therefore, to minimize the rate-distortion risk, we need to solve the following optimization problem:

$$\inf_{P(\tilde{\theta}_n|\theta_n^*)} \mathbf{E}_P[\|\theta_n^* - \tilde{\theta}_n\|^2]$$

$$\text{such that } I(\tilde{\theta}_n; \theta_n^*) \le B_n$$

This optimization problem is the distortion-rate formulation of the problem: We want the smallest possible distortion given a bound on the rate. This formulation is equivalent to the rate-distortion formulation since the R-D function is both convex and non-increasing (inversion is both well-defined and does not change the optima).

Thus, we can view this as rate-distortion where we are trying to learn the best lossy compression of the optimal **parameter**, rather than first compressing the data and only afterwards minimizing risk. In fact, this detail is a unique feature of this paper: They minimize risk first, and then compress the solution.

We can write the result of the optimization problem as $Q_n(\Theta, B_n; \pi)$. Then we have that the quantized risk is lower bounded by

$$R_n(\Theta, B_n) \ge R_n(\Theta; \pi) + Q_n(\Theta, B_n; \pi)$$

and taking the worst prior gives

$$R_n(\Theta, B_n) \ge \sup_\pi \{R_n(\Theta; \pi) + Q_n(\Theta, B_n; \pi)\}$$

5.3 Data Generation Model and Assumptions

Now, we define the assumptions about the data generation process as well as the function class the paper tries to learn.

Specifically, the paper works in the Gaussian white noise setting defined by

$$dX(t) = f(t)dt + \epsilon dB(t) \text{ for } t \in [0,1]$$

where B is Brownian motion on $[0,1]$, ϵ is the standard deviation, and $f \in \tilde{W}(m,c)$, the periodic Sobolev space of order m and radius c.

Definition 5.2. Sobolev Space, Periodic Sobolev Space, Sobolev Ellipsoid.
The Sobolev space $W(m,c)$ of order m and radius c is defined by

$$W(m,c) = \left\{ f \in [0,1] \to \mathbb{R} : f^{(m-1)} \text{ is absolutely continuous; } \int_0^1 \left(f^{(m-1)}(x)\right)^2 dx \leq c^2 \right\}$$

The periodic Sobolev space $\tilde{W}(m,c)$ of order m and radius c is a restriction of Sobolev space defined by

$$\tilde{W}(m,c) = \left\{ f \in W(m,c) : f^{(j)}(0) = f^{(j)}(1) \text{ for all } j \in [m-1] \right\}$$

That is, we require that all derivative orders from 0 to $m-1$ are equal at the start and end of the interval. Intuitively, if m were infinite, this requirement enforces the function to return to the same "position" and the end of the interval, allowing us to stack the $[0,1]$ intervals together to build a periodic function. In practice, we truncate at the $(m-1)^{th}$ derivative as an approximation.

First, we assume that we can express f in terms of its trigonometric basis and that the series converges:

$$f = \sum_{j=1}^{\infty} \left(\int_0^1 \varphi_j(t) f(t) dt \right) \varphi_j = \sum_{j=1}^{\infty} \theta_j \varphi_j$$

where $\{\varphi_j\}_{j=1}^{\infty}$ is an orthonormal basis for $L_2[0,1]$ defined as

$$\begin{aligned} \varphi_1(x) &= 1 \\ \varphi_{2k}(x) &= \sqrt{2}\cos(2\pi k x) \\ \varphi_{2k+1}(x) &= \sqrt{2}\sin(2\pi k x) \end{aligned} \tag{8}$$

for $k \in \mathbb{N}$.

It turns out that a function f belongs to $\tilde{W}(m,c)$ iff in the above trigonometric representation, we have that the Fourier coefficients θ satisfy

$$\sum_{j=1}^{\infty} j^{2m} \theta_j^2 \leq \frac{c^2}{\pi^{2m}}$$

This restriction on θ defines the **Sobolev ellipsoid** $\Theta(m,c)$, the subset of functions in L_2 with this condition.

Finally, the observed path $X(t)$ is converted into an infinite Gaussian sequence $Y_j = \int_0^1 \varphi_j(t) dX(t)$, with $Y_j \sim \mathcal{N}(\theta_j, \epsilon^2)$. Thus, for estimates $\left(\hat{\theta}_j\right)_{j=1}^{\infty}$ of $(Y_j)_{j=1}^{\infty}$, we can estimate $\hat{f} = \sum_{j=1}^{\infty} \hat{\theta}_j \varphi_j$. Additionally, we use the square loss

$$L(\hat{f}, f) = \|\hat{f} - f\|_2^2 = \|\hat{\theta} - \theta\|_2^2$$

5.4 Main Result

To summarize, the paper tries to characterize the asymptotic behavior of the minimax risk with storage constraints in perodic Sobolev space under Gaussian white noise:

$$R_\epsilon(m, c, B_\epsilon) = \inf_{\hat{f}_\epsilon, C(\hat{f}_\epsilon) \leq B_\epsilon} \sup_{f \in \tilde{W}(m,c)} \mathbf{E}[\|f - \hat{f}_\epsilon\|_2^2]$$

The main result shows that for given standard deviation ϵ, the risk behaves as

$$R_\epsilon(m, c, B_\epsilon) \approx P_{m,c} \epsilon^{\frac{4m}{2m+1}} + \frac{c^2 m^{2m}}{\pi^{2m}} B_\epsilon^{-2m}$$

where we have decomposed the risk into the traditional minimax estimation error when there are no constraints, and the rate-distortion error (or quantized error). The term $P_{m,c}$ is from the classic nonparametric statistics literature and is known as Pinsker's constant. The coefficient of the rate-distortion error is the main new result of this paper. The high-level gist of their approach is as follows: In order to solve the rate-distortion problem between the standard best estimator and the best estimator in the compressed setting, they make use of the quantization of a Gaussian source with varying noise levels; except here the Sobolev ellipsoid is an infinite Gaussian sequence and requires careful truncation. In particular, one of the main points of novelty of this paper is that they can view the uncompressed data as fully available to the user at the time of estimation, and constraints/quantization only happens after having found the best estimate. Future work in particular requires computationally efficient coding schemes to achieve the rates set out in this paper, since they show achievability only with random coding schemes.