CS861: Theoretical Foundations of Machine Learning Lecture 13 - 10/02/2024

University of Wisconsin–Madison, Fall 2023

Lecture 13: Alternatives via Tight Packings

Lecturer: Kirthevasan Kandasamy *Scribed by: Ziyang Cai, Jakob Juergens*

Disclaimer 2: *These notes are an updated version of existing notes created in 2023 by Elliot Pickens and Yuya Shimizu to reflect the contents covered in 2024.*

This lecture provides further methods to derive a lower bound for the minimax risk. In addition to the previously introduced method using the `Varshamov-Gilbert lemma`, we introduce the technique of constructing alternatives via `tight packings`. Finally, we will discuss `nonprarametric regression` and apply the previously established ideas to find the corresponding minimax optimal rate.

1 Constructing Alternatives via Tight Packings

To construct alternatives using tight packings, we first introduce multiple definitions that will allow us a relatively intuitive construction of alternatives contained in a parametric family.

Definition 1 (ε-Packing & ε-Packing Number)**.**
Consider a metric space (M, ρ) and a set $\mathcal{X} \subset M$.

- *An ε-packing of $\mathcal{X}$ with respect to ρ is a finite subset of $\{x_i\}_{i=1}^n \subset \mathcal{X}$ such that every two points are separated by a distance of at least ε, i.e., $\forall i \neq j \in [n]: \quad \rho(x_i, x_j) \geq \varepsilon$.*
- *The ε-packing number of $\mathcal{X}$ with respect to ρ, denoted by $M(\varepsilon, \mathcal{X}, \rho)$ is the size of the largest ε-packing of $\mathcal{X}$ with respect to ρ.*
- *An ε-packing of size $M(\varepsilon, \mathcal{X}, \rho)$ is called a maximal ε-packing of $\mathcal{X}$.*

It is easy to see that for an L_2-ball of radius r in d-dimensional Euclidean space, $\mathcal{X} = \{x \in \mathbb{R}^d \,|\, \|x\| \leq r\}$, equipped with the L_2-norm, we have $M\left(\varepsilon, \mathcal{X}, \|\|_2\right) \geq \left(\frac{r}{\varepsilon}\right)^d$. Similarly, we define ε-covering numbers that serve as a kind of counterpart to ε-packing numbers.

Definition 2 (Covering Number, Metric Entropy)**.**
Consider a metric space (M, ρ) and a set $\mathcal{X} \subset M$.

- *An ε-**covering** of set $\mathcal{X}$ with respect to ρ is a set $\{x_i\}_{i=1}^n$ such that for all $x \in \mathcal{X}$, there exists some $x_j \in \{x_i\}_{i=1}^n$ s.t. $\rho(x, x_j) \leq \varepsilon$.*
- *The ε-**covering number** $N(\varepsilon, \mathcal{X}, \rho)$ is the size of the <u>smallest</u> ε-covering.*
- *The **metric entropy** is $\log(N(\varepsilon, \mathcal{X}, \rho))$.*

We have the following lemma that relates covering numbers and packing numbers. This relationship is useful since we can use existing work on both the covering number and packing number to bound the metric entropy $\log(N(\varepsilon, \mathcal{X}, \rho))$ in a number of applications.

Lemma 1. *A covering number $N(\cdot, \mathcal{X}, \rho)$ and a packing number $M(\cdot, \mathcal{X}, \rho)$ satisfy*

$$M(2\varepsilon, \mathcal{X}, \rho) \leq N(\varepsilon, \mathcal{X}, \rho) \leq M(\varepsilon, \mathcal{X}, \rho).$$

To explore the usefulness of constructing alternatives via tight packings, we consider the example of estimating the mean of a normal distribution with known variance in $\mathbb{R}^d$.

Example 1 (Normal mean estimation in $\mathbb{R}^d$)**.**
Say we are interested in estimating the mean of a d-dimensional normal distribution with known variance $\sigma^2 I_d$ using the squared L_2 loss. In other terms, our parameter of interest is given by the expectation functional $\theta(P) = \mathbb{E}_{X\sim P}[X]$ and we are using a loss that conforms to our usual structure $\phi \circ \rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2$.

As the variance is assumed to be known, the space of distributions we consider as alternatives is as follows.

$$\mathcal{P} = \left\{\mathcal{N}\left(\mu, \sigma^2 I_d\right) \mid \mu \in \mathbb{R}^d\right\}$$

For this exact estimation problem, we previously established an upper bound to the minimax risk.

$$R_n^* = \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^n}\left[\|\theta(P) - \hat{\theta}(S)\|_2^2\right] \leq \sigma^2 \frac{d}{n}$$

We are interested in establishing a lower bound of the same order, i.e., d/n, to establish the minimax rate optimality of the estimator considered previously. To do so, we will employ tools based on the ideas of tight-packings. As the minimax risk considers the supremum over all alternatives, we will first consider a reduced set of alternatives sufficient for constructing a lower bound. To this end, let U be a maximal δ-packing of the L_2-ball of radius 2δ in $\mathbb{R}^d$ and define the following reduced set of alternatives.

$$\mathcal{P}' = \left\{\mathcal{N}\left(u, \sigma^2 I_d\right) \mid u \in U\right\}$$

As U is a δ-packing, we can immediately deduce the following two properties.

$$\forall u \neq u' \in U : \|u - u'\| \geq \delta \quad \text{and} \quad N := |\mathcal{P}'| = |U| \geq 2^d$$

Furthermore, as the Kullback–Leibler divergence between two Gaussians of the form at hand has a simple closed-form expression, we can use the following.

$$\mathrm{KL}\left(P_u \,||\, P_{u'}\right) = \frac{\|u - u'\|_2^2}{2\sigma^2} \overset{\text{Radius: } 2\delta}{\leq} \frac{8\delta^2}{\sigma^2}$$

We want to use the local Fano method to achieve a lower bound of the desired rate. Considering the bound given in the local Fano method, it becomes apparent that we thus require $\mathrm{KL}\left(P_u \,||\, P_{u'}\right) \leq \frac{|\mathcal{P}'|}{4n}$ to achieve the desired rate of convergence, as this ensures that $\frac{8\delta^2}{\sigma^2} \leq \frac{d\log(2)}{4n}$. Solving for δ, we obtain $\delta = \sigma\sqrt{\frac{d\log(2)}{32n}}$, which, plugged into the local Fano lower bound, gives us the desired rate.

$$R^* \leq \left(\sigma\sqrt{\frac{d\log(2)}{32n}}\right)^2 \left(1 - \frac{\frac{N}{4n} + \log(2)}{\log(N)}\right) = \sigma^2 \frac{d\log(2)}{32n}\left(1 - \frac{N}{4n\log(N)} - \frac{\log(2)}{\log(N)}\right) \leq \frac{\log(2)}{32}\frac{\sigma^2 d}{n}$$

As the lower and upper bound share the same rate of convergence, this shows rate optimality of the previously considered estimator.