Information Theoretic Minimax Lower Bounds

Jennifer Rogers

December 2018

1 Introduction

Suppose we are given n observations $X_1, \ldots, X_n$, and our goal is to determine whether the samples were drawn from distribution P_0 or P_1. This corresponds to performing a test of two hypotheses,

$$H_0 : X_1, \ldots, X_n \sim P_0$$
$$H_1 : X_1, \ldots, X_n \sim P_1$$

For example, suppose we know our samples come from a normal distribution with standard deviation σ, and we are testing whether the mean is θ_0 or θ_1. Then $P_0 = \mathcal{N}(\theta_0, \sigma^2)$ and $P_1 = \mathcal{N}(\theta_1, \sigma^2)$. We will use this example throughout this note.

What is the relationship between the number of samples and the error of the hypothesis test? If we had a specific hypothesis test in mind, we could answer this question by computing its type 1 and type 2 errors for a given n. In this note, we will be interested in lower bounds, which are statements of the form "every hypothesis test that sees n samples gets the answer wrong at least δ fraction of the time," where the lower bound δ is a function of n. In order to answer this question, we bound the *minimax probability of error* on the hypothesis test of distinguishing between the two populations. In this note, we describe three different types of minimax lower bounds: bounding the chance of error in a hypothesis test with a given number of samples, bounding the number of samples a δ-PAC correct hypothesis test must take, and bounding the risk of an estimator.

2 Minimax Probability of Error in Hypothesis Testing

We begin with the first type of lower bound: given n samples i.i.d. from either P_0 or P_1, we lower bound the probability that any hypothesis test makes an error. There are two types of errors: choosing H_1 if the truth is H_0, and vice versa. We are interested in bounding the probability of *any* error occurring, which means we want to bound the maximum of the two types. If $\Psi : \{X_1, \ldots, X_n\} \to \{0, 1\}$ is a hypothesis test that decides between H_0 and H_1, then we have the following lower bound on the minimax probability of error.

Theorem 2.1. *Any hypothesis test that takes n samples and distinguishes between $H_0 : X_1, \ldots, X_n \sim P_0$ and $H_1 : X_1, \ldots, X_n \sim P_1$ has probability of error lower bounded by*

$$\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) \geq \frac{1}{4} e^{-nKL(P_0||P_1)}$$

where $KL(P||Q)$ is the KL divergence.

Proof. To begin, we note that the maximum error is bounded below by the average error.

$$\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) \geq \frac{1}{2}\mathbb{P}_0(\Psi = 1) + \frac{1}{2}\mathbb{P}_1(\Psi = 0)$$

Define the product distribution $\widetilde{P}_i = P_i^{\otimes n}$ (i.e., P_i *i.i.d.* on each of the n samples), which is a measure on the domain of Ψ. Then we can rewrite the probability of each type of error as an integral,

$$\begin{aligned}
\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) &\geq \frac{1}{2}\left(\int_{\Psi=1} d\widetilde{P}_0 + \int_{\Psi=0} d\widetilde{P}_1\right) \\
&\geq \frac{1}{2}\left(\int_{\Psi=1} \min(d\widetilde{P}_0, d\widetilde{P}_1) + \int_{\Psi=0} \min(d\widetilde{P}_0, d\widetilde{P}_1)\right) \\
&= \frac{1}{2}\int \min(d\widetilde{P}_0, d\widetilde{P}_1)
\end{aligned}$$

Next, multiply by the integral of $\max(d\widetilde{P}_0, d\widetilde{P}_1)$. This integral is always less than 2, since the maximum of two quantities is less than their sum, and $\int dP_i = 1$. This lets us introduce another inequality,

$$\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) \geq \frac{1}{4} \int \min(d\widetilde{P}_0, d\widetilde{P}_1) \int \max(d\widetilde{P}_0, d\widetilde{P}_1)$$

Now, we use the Cauchy-Schwarz inequality, $\langle a, b\rangle^2 \leq ||a||^2||b||^2$, to bound this expression by a single integral. If we apply the Cauchy-Schwarz inequality with $a = \sqrt{\min(d\widetilde{P}_0, d\widetilde{P}_1)}$ and $b = \sqrt{\max(d\widetilde{P}_0, d\widetilde{P}_1)}$, we get

$$\begin{aligned}
\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) &\geq \frac{1}{4} \left(\int \sqrt{\min(d\widetilde{P}_0, d\widetilde{P}_1) \max(d\widetilde{P}_0, d\widetilde{P}_1)} \right)^2 \\
&= \frac{1}{4} \left(\int \sqrt{d\widetilde{P}_0 d\widetilde{P}_1} \right)^2 \\
&= \frac{1}{4} \left(\int d\widetilde{P}_0 \sqrt{\frac{d\widetilde{P}_1}{d\widetilde{P}_0}} \right)^2 \\
&= \frac{1}{4} \left(\mathbb{E}_0 \left[\sqrt{\frac{d\widetilde{P}_1}{d\widetilde{P}_0}} \right] \right)^2
\end{aligned}$$

Our goal is to express this in terms of the KL divergence, so we need to introduce the log-likelihood ratio. We do this by rewriting the inner term as the exponential of a log, and then applying Jensen's inequality.

$$\begin{aligned}
\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) &\geq \frac{1}{4} \exp\left(2 \log \left(\mathbb{E}_0 \left[\sqrt{\frac{d\widetilde{P}_1}{d\widetilde{P}_0}} \right] \right) \right) \\
&\geq \frac{1}{4} \exp\left(2\mathbb{E}_0 \left[\log \left(\sqrt{\frac{d\widetilde{P}_1}{d\widetilde{P}_0}} \right) \right] \right) \\
&= \frac{1}{4} \exp\left(-\mathbb{E}_0 \left[\log \left(\frac{d\widetilde{P}_0}{d\widetilde{P}_1} \right) \right] \right) \\
&= \frac{1}{4} \exp\left(-KL(\widetilde{P}_0 || \widetilde{P}_1) \right)
\end{aligned}$$

We have derived a lower bound on the minimax probability of error in terms of the KL divergence between $\widetilde{P}_0$ and $\widetilde{P}_1$. The KL divergence sums over product distributions, giving us $KL(\widetilde{P}_0||\widetilde{P}_1) = nKL(P_0||P_1)$, and we recover the desired lower bound.

$$\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) \geq \frac{1}{4} \exp\left(-nKL(P_0||P_1)\right)$$

□

We can apply this bound to our running example of distinguishing between two Gaussian hypotheses with different means and equal variances. The KL divergence between Gaussians $\mathcal{N}(\theta_0, \sigma^2)$ and $\mathcal{N}(\theta_1, \sigma^2)$ is given by $(\theta_1 - \theta_0)^2/(2\sigma^2)$. This implies the following corollary.

Corollary 2.2. *Any hypothesis test that takes n samples and distinguishes between hypotheses $H_0 : X_1, \ldots, X_n \sim \mathcal{N}(\theta_0, \sigma^2)$ and $H_1 : X_1, \ldots, X_2 \sim \mathcal{N}(\theta_1, \sigma^2)$ has probability of error bounded below by*

$$\max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) \geq \frac{1}{4} e^{-n(\theta_1 - \theta_0)^2/(2\sigma^2)}$$

This result matches our understanding of what would make this problem hard. If the means are close together, then the chance of error increases because the hypothesis distributions look more similar. If the variance is small, then the means are better separated, and the minimax probability of error is lower.

So far, we have seen how to compute a lower bound on the probability that any hypothesis test is in error after taking n samples. We might instead want to set a desired level of correctness, and prove a lower bound on the number of samples needed to guarantee that correctness. A δ-PAC guarantee on a hypothesis test states that its type 1 and type 2 error rates are both bounded by δ. We can use Theorem 2.1 to lower bound the number of samples required by any δ-PAC hypothesis test.

Corollary 2.3. *Any hypothesis that uses n samples to distinguish between $H_0 : X_1, \ldots, X_n \sim P_0$ and $H_1 : X_1, \ldots, X_n \sim P_1$ and is δ-PAC correct must satisfy*

$$n \geq \frac{\log(\frac{1}{8\delta})}{KL(P_0||P_1)}$$

Proof. We begin with Theorem 2.1, and then apply the fact that the maximum is upper bounded by the sum of its arguments.

$$\begin{aligned}\frac{1}{4}e^{-nKL(P_0||P_1)} &\leq \max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0)) \\ &\leq \mathbb{P}_0(\Psi = 1) + \mathbb{P}_1(\Psi = 0) \\ &\leq 2\delta\end{aligned}$$

where the last line follows from the δ-PAC guarantee. We can now solve for n in terms of δ, to get

$$n \geq \frac{\log(\frac{1}{8\delta})}{KL(P_0||P_1)}$$

□

We can use the preceding corollary to get a lower bound on the number of samples required by any δ-PAC hypothesis test to distinguish between two Gaussians of equal variance.

Corollary 2.4. *Any hypothesis test that uses n samples to distinguish between $H_0 : X_1, \ldots, X_n \sim \mathcal{N}(\theta_0, \sigma^2)$ and $H_1 : X_1, \ldots, X_n \sim \mathcal{N}(\theta_1, \sigma^2)$ and is δ-PAC correct must satisfy*

$$n \geq \frac{2\sigma^2 \log(\frac{1}{8\delta})}{(\theta_1 - \theta_0)^2}$$

We see that more samples are required when the variance is large or the means are close together. The number of samples also increases as the error tolerance δ decreases.

In Section 4, we will use the result of Theorem 2.1 to bound a different quantity of interest: the minimax risk of an estimator. In the next section, we will define this quantity.

3 Defining the Minimax Risk

So far, we have been working in the setting where we must distinguish between two known hypotheses, H_0 and H_1. We might instead be faced with a setting where we observe n samples $X_1, \ldots, X_n$ from some distribution parameterized by $\theta \in \Theta$, and we want to estimate θ with some estimator $\widehat{\theta}_n$. For example, we might want to estimate the average height of a set of individuals with heights distributed $\mathcal{N}(\theta, \sigma^2)$.

A natural question we might ask is: how well can we estimate θ given these n samples? To answer this question, we first need some notion of what it means to approximate θ "well." In this note, we will restrict our discussion to the mean squared error, $||\widehat{\theta} - \theta||^2$.

Risk is defined relative to the truth θ, but in general we don't know θ (after all, we're trying to estimate it). How can we evaluate the quality of our estimator in general, over all $\theta \in \Theta$?

We could imagine trying to describe the performance of our estimator in terms of its average risk, perhaps taking an expectation of the risk over all parameters $\theta \in \Theta$. However, in order to take this expectation, we need to define a distribution over Θ. While we may sometimes have access to such a distribution, in general we will not. In order to avoid a dependence on this knowledge, we instead describe the performance of an estimator in terms of its worst-case risk, R_{max}.

$$R_{max}(\widehat{\theta}_n) = \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right]$$

If we want to measure the performance of our estimator in terms of the worst-case risk, then we would like to find the estimator that minimizes this worst-case risk. The *minimax risk* is the smallest risk attainable by any estimator $\widehat{\theta}_n$ on the most difficult $\theta \in \Theta$. This can be written as

$$R_{minimax}(\widehat{\theta}_n) = \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right]$$

The minimax risk can be read, from the outside in, as an adversarial optimization. Beginning with the outer inf, the player chooses some estimation procedure on the set of n samples. Next, in the inner sup, the adversary chooses the value of θ to maximize the player's risk. The adversary may use knowledge of the player's estimation procedure, as well as the number of samples n. The player's goal is to choose an estimator $\widehat{\theta}_n$ such that this worst-case risk is minimized. The minimax risk is the worst-case risk under this most conservative choice of estimator.

We are interested in computing lower bounds on the minimax risk. Such a bound allows us to make statements of the following form: no procedure which takes n samples from a distribution with parameter $\theta \in \Theta$, and returns an estimator $\widehat{\theta}_n$, achieves a risk lower than the minimax risk on every distribution.

4 Lower Bounds on Hypothesis Testing Imply Lower Bounds on Estimation Error

Theorem 4.1. *Suppose that P is a probability distribution parameterized by θ, with $P_0 = P(\theta_0)$ and $P_1 = P(\theta_1)$. Additionally, define $s = \frac{1}{2}||\theta_0 - \theta_1||$. Then, the minimax risk of any estimator $\widehat{\theta}_n$ that estimates θ with n samples from $P(\theta)$ is lower bounded by*

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq \frac{1}{4} s^2 e^{-nKL(P_0||P_1)}$$

for all $\theta_0, \theta_1 \in \Theta$, where $KL(P||Q)$ is the KL divergence.

Proof. In order to prove minimax lower bounds, we will transform our problem into one of hypothesis testing. We begin with Markov's inequality, which allows us to relate the probability of a nonnegative random variable exceeding some threshold, and the expected value of that random variable. Let s be nonnegative. Then,

$$\begin{aligned} \mathbb{P}(||\widehat{\theta}_n - \theta|| \geq s) &= \mathbb{P}(||\widehat{\theta}_n - \theta||^2 \geq s^2) \\ &\leq \frac{\mathbb{E}\left[||\widehat{\theta}_n - \theta||^2 \right]}{s^2} \end{aligned}$$

This gives us a lower bound on the minimax risk,

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq s^2 \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{P}(||\widehat{\theta}_n - \theta|| \geq s)$$

We have managed to express the minimax risk in terms of the probability that our estimator deviates from the truth by more than some threshold s. This isn't quite sufficient, because we still have no way to determine how often the best estimator will violate that threshold. In order to control this quantity, we switch from a supremum over all of Θ to a maximum over two hypotheses. Additionally, we separate our hypotheses by a distance of $2s$. This lets us establish a connection between the test choosing the wrong hypothesis, and the estimator being far from the truth.

In order to reduce to two hypotheses, let us first define a pair of distributions $\{\theta_0, \theta_1\} \subset \Theta$, which satisfy $||\theta_0 - \theta_1|| \geq 2s$. If we restrict our true distribution θ to be either θ_0 or θ_1, then we have created an easier problem. In particular, the supremum over all $\theta \in \Theta$ certainly upper bounds the maximum over the subset $\{\theta_0, \theta_1\}$, which gives us the lower bound

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq s^2 \inf_{\widehat{\theta}_n} \max_{i \in \{0,1\}} \mathbb{P}_i(||\widehat{\theta}_n - \theta_i|| \geq s)$$

Next, we transform the infimum over all estimators into an infimum over all hypothesis tests. As before, let $\Psi : \{X_1, \ldots, X_n\} \to \{0, 1\}$ be a hypothesis test that selects either H_0 or H_1 based on the observed data. We want to claim that a lower bound on any hypothesis test's probability of error implies a lower bound on any estimator's

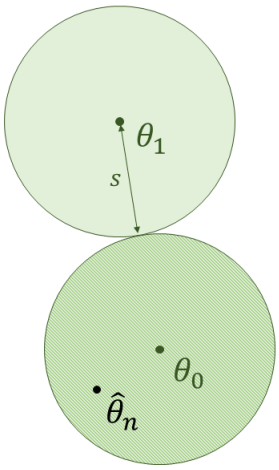

Figure 1: If the estimator $\widehat{\theta}_n$ is within distance s of the true hypothesis with probability at least p, then we could use it to design a hypothesis test with probability of error less than $1-p$. Therefore, a lower bound on the probability of error for the hypothesis test is also a lower bound on the probability of $||\widehat{\theta}_n - \theta||$ exceeding s.

probability of getting within s of the truth. To see this, suppose, to obtain a contradiction, that we had a lower bound on the performance of a hypothesis test, which said that no hypothesis test could correctly identify θ_i with probability greater than some p, for each hypothesis θ_i. Then, we cannot have an estimator which gets within distance s of the true distribution θ_i with probability p. If we did, then we would use that estimator to design a hypothesis test with error p: simply estimate $\widehat{\theta}_n$ and return the closest θ_i. With probability p, this estimator is within s of the true distribution. In this event, since our hypotheses are separated by $2s$, the closest hypothesis to $\widehat{\theta}_n$ is the true θ_i. We have used our estimator to create a hypothesis test with probability of success greater than p, which is a contradiction. This tells us that no estimator can be within s of the true parameter with probability p. The idea is illustrated in Figure 1.

We now formalize this argument in the following lemma.

Lemma 4.2. *Let $d(\theta_0, \theta_1) = ||\theta_0 - \theta_1||$. If $d(\theta_0, \theta_1) \geq 2s$, then for any estimator $\widehat{\theta}_n$, the probability that its distance from the truth exceeds s is bounded by*

$$\mathbb{P}_{\theta_j}(d(\widehat{\theta}_n, \theta_j) \geq s) \geq \mathbb{P}_{\theta_j}(\Psi^* \neq j)$$

where Ψ^* is the minimum distance test defined by

$$\Psi^* = \arg\min_{k \in \{0,1\}} d(\widehat{\theta}_n, \theta_k)$$

Proof. To prove the desired statement, it suffices to prove the implication

$$\Psi^* \neq j \implies d(\widehat{\theta}_n, \theta_j) \geq s$$

under the true distribution θ_j. We will show the proof for $j = 0$; the case for $j = 1$ is identical.

If $\Psi^* \neq 0$, then the minimum distance test chose $k = 1$. Since the hypotheses are separated by at least $2s$, we can use the triangle inequality to conclude

$$\begin{aligned} 2s &\leq d(\theta_1, \theta_0) \\ &\leq d(\theta_1, \widehat{\theta}_n) + d(\widehat{\theta}_n, \theta_0) \\ &\leq 2d(\widehat{\theta}_n, \theta_0) \end{aligned}$$

where the last inequality follows from the fact that the minimum distance test chose 1 instead of 0, so $d(\widehat{\theta}_n, \theta_1) < d(\widehat{\theta}_n, \theta_0)$. We have shown the desired implication for $j = 0$; this argument can be repeated for $j = 1$ to complete the proof of the lemma. □

We use this lemma to lower bound the risk by the maximum error of the best hypothesis test.

$$\begin{aligned}\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] &\geq s^2 \max_{j \in \{0,1\}} \mathbb{P}_j(\Psi^* \neq j) \\ &\geq s^2 \inf_{\Psi} \max_{j \in \{0,1\}} \mathbb{P}_j(\Psi \neq j) \\ &= s^2 \inf_{\Psi} \max(\mathbb{P}_0(\Psi = 1), \mathbb{P}_1(\Psi = 0))\end{aligned}$$

We have reduced the problem of bounding the minimax risk of an estimator to the problem of bounding the error rate of a hypothesis test. From here, we can use the bounds on the minimax probability of error from Theorem 2.1, completing the proof. □

Returning to our example, suppose we knew our data came from a Gaussian distribution, and we wanted to estimate its mean. We can use Theorem 4.1 to get a lower bound for the risk of any estimator for this problem.

Corollary 4.3. *Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$. The minimax risk of any estimator $\widehat{\theta}_n$ that estimates θ given $X_1, \ldots, X_n$ is bounded below by*

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq \frac{1}{8e} \cdot \frac{\sigma^2}{n}$$

Proof. Let $\delta = ||\theta_0 - \theta_1||$. Then, Theorem 4.1 gives us

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq \frac{1}{16} \delta^2 e^{-nKL(P_0||P_1)}$$

The KL divergence between two Gaussians is given by $(\theta_0 - \theta_1)^2/(2\sigma^2)$, which we substitute into the KL divergence.

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq \frac{1}{16} \delta^2 e^{-n\delta^2/(2\sigma^2)}$$

Theorem 4.1 holds for all choices of θ_0 and θ_1, so we maximize this lower bound over all positive values of δ.

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq \max_{\delta} \frac{1}{16} \delta^2 e^{-n\delta^2/(2\sigma^2)}$$

This maximum is achieved at $\delta = \sqrt{2\sigma^2/n}$, which gives a lower bound of

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] \geq \frac{1}{8e} \cdot \frac{\sigma^2}{n}$$

□

We have just seen that no estimator $\widehat{\theta}_n$ for the mean of a Gaussian has expected mean squared error less than $O\left(\frac{\sigma^2}{n}\right)$. We might ask whether that lower bound is tight, that is, whether there exists an estimator that achieves the lower bound. As the next remark shows, there is such an estimator.

Remark 4.4. *Let $X_i \sim \mathcal{N}(\theta, \sigma^2)$. The empirical average, $\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$, achieves the minimax risk of $O\left(\frac{\sigma^2}{n}\right)$ up to constant factors.*

Proof. To prove this claim, we will compute the MSE of this estimator. We begin with the bias-variance decomposition:

$$\mathbb{E}_\theta \left[||\widehat{\theta}_n - \theta||^2 \right] = ||\mathbb{E}[\widehat{\theta}_n] - \theta||^2 + \mathbb{E}[||\widehat{\theta}_n - \mathbb{E}[\widehat{\theta}_n]||^2]$$

The estimator is unbiased, so we are left with the variance term,

$$\begin{aligned}
\mathbb{E}_\theta\left[||\widehat{\theta}_n - \theta||^2\right] &= \mathbb{E}[||\widehat{\theta}_n - \mathbb{E}[\widehat{\theta}_n]||^2] \\
&= \mathrm{Var}(\widehat{\theta}_n) \\
&= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

We conclude that the empirical average has risk $\frac{\sigma^2}{n}$. By corollary 4.3, it is minimax optimal up to constant factors. □

There are many different types of information theoretic lower bounds. For example, if the KL divergence is difficult to calculate for a given pair of hypothesis distributions, there are bounds that employ the Hellinger distance, the total variation distance, or the likelihood ratio between the distributions. There are also situations in which two hypotheses are insufficient for a strong lower bound on the risk, in which case it is necessary to use bounds based on multiple hypotheses. Chapter 2 of [3] is a useful reference for a variety of information theoretic lower bounds.

The statement of Theorem 2.1 can be found in [3] (as Theorem 2.2). The proof presented here was synthesized from techniques found in [3] and the proof of Lemma 7 in [1]. The proof of Theorem 4.1 is based on Section 2.2 of [3]. Readers may also be interested in John Duchi's excellent set of notes on minimax lower bounds [2].

References

[1] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. "Bounded regret in stochastic multi-armed bandits". In: *Journal of Machine Learning Research* 30 (Feb. 2013).

[2] John Duchi. "Lecture Notes for Statistics 311/Electrical Engineering 377". In: http://stanford.edu/class/stats311/Lectures/lec-03.pdf. 2014. Chap. 2, pp. 10–33.

[3] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st ed. Springer-Verlag New York, 2009. ISBN: 978-0-387-79051-0.