Qifan's Notes in Statistics

Qifan Zhang

December 2023

Chapter 1

Statistical Theory

1.1 Asymptotic Properties of Maximal Likelihood Estimation

1.2 Probabilistic Inequalities

This section includes some useful probabilistic inequalities and their proofs.

1.2.1 Hoeffding's Inequality

Theorem 1.2.1 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \cdots, X_n$ be independent random variables and $S_n = \sum_{i=1}^n X_i$. If $a_i \leq X_i \leq b_i$, we have for any $t \in \mathbb{R}^+$,*

$$\begin{aligned} \Pr(S_n - \mathbb{E}[S_n] \geq t) &\leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}, \\ \Pr(|S_n - \mathbb{E}[S_n]| \geq t) &\leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}. \end{aligned} \tag{1.1}$$

Lemma 1.2.2 (Hoeffding's Lemma)**.** *Let X be a random variable with $\mathbb{E}[X] = 0$ and $a \leq X \leq b$. Then for any $t \in \mathbb{R}$ we have*

$$\mathbb{E}[e^{tX}] \leq e^{\frac{1}{8}t^2(b-a)^2}. \tag{1.2}$$

Proof. By the convexity of e^{tx}, we have

$$\begin{aligned} e^{tX} &\leq \frac{b-X}{b-a}e^{ta} + \frac{X-a}{b-a}e^{tb} \\ \Longrightarrow \mathbb{E}[e^{tX}] &\leq \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} \\ &= e^{ta}[\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}]. \end{aligned} \tag{1.3}$$

Let

$$L(h) = \frac{ha}{b-a} + \ln[\frac{b}{b-a} - \frac{a}{b-a}e^h]. \tag{1.4}$$

Then

$$e^{ta}[\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}] = e^{L(t(b-a))}. \tag{1.5}$$

Note that

$$\begin{aligned} L(0) &= 0, \\ L'(0) &= \frac{a}{b-a} - \frac{a}{b-a} = 0, \\ L''(0) &= -\frac{abe^h}{(b-ae^h)^2} \leq \frac{1}{4}. \end{aligned} \tag{1.6}$$

Thus by Taylor expansion, we have

$$\begin{aligned} L(t(b-a)) &\leq \frac{1}{8}t^2(b-a)^2 \\ \implies \mathbb{E}[e^{tX}] &\leq e^{\frac{1}{8}t^2(b-a)^2}. \end{aligned} \tag{1.7}$$

□

Proof of 1.2.1. Using 1.2.2, we have

$$\begin{aligned} \Pr(S_n - \mathbb{E}[S_n] \geq t) &= \Pr(e^{s(S_n - \mathbb{E}[S_n])} \geq e^{st}) \\ &= \Pr(e^{s(S_n - \mathbb{E}[S_n])}e^{-st} \geq 1) \\ &\leq \mathbb{E}[e^{s(S_n - \mathbb{E}[S_n])}e^{-st}] \\ &\leq e^{\frac{1}{8}s^2\sum_{i=1}^n (b_i-a_i)^2 - st} \\ &\leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}}. \end{aligned} \tag{1.8}$$

Note that 1.2.2 holds for all $t \in \mathbb{R}$. Similarly, we have

$$\Pr(\mathbb{E}[S_n] - S_n \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}}. \tag{1.9}$$

Thus

$$\Pr(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}}. \tag{1.10}$$

□

Chapter 2

Machine Learning Theory

2.1 Theory of Hidden Markov Model

2.2 Local Polynomial Regression

2.3 Score Matching

This section is my notes on Hyvärinen and Dayan [2005]. Consider that we observe a sequence of data $X_1, X_2, \cdots, X_n$ and we want to estimate the generative model $p(X_1, X_2, \cdots, X_n)$ behind it. Assume that our observations are i.i.d. and our candidate models can be parametrized i.e. $p(X) = p_\theta(X),\ \theta \in \Theta$. Furthermore, we assume that $X_i \in \mathbb{R}^p$. Under this framework, the de facto standard approach is Maximal Likelihood Estimation (MLE), which is to consider the optimization problem

$$\theta^* \in \arg\max_\theta \sum_{i=1}^n \log p_\theta(X_i). \tag{2.1}$$

The limitation of this method in practice is that it requires us to know the exact form of $p_\theta(x)$ beforehand and thus it is hard to generalize it to some complex data such as images, natural language and so on. Nevertheless, we can still make some assumptions on the shape of p_θ without normalizing it. Let

$$p_\theta(x) = \frac{f_\theta(x)}{Z_\theta}, \tag{2.2}$$

where Z_θ is the normalization constant[1]. Then we can rewrite the log-likelihood as

$$\ell_\theta(X_1, X_2, \cdots, X_n) = \sum_{i=1}^n \log f_\theta(X_i) - n \log Z_\theta. \tag{2.3}$$

[1]The expression here is not unique.

In the above equation, $f_\theta(x)$ is the term we can control. We can model it as any types of functions such as neural networks based on our prior knowledge on our datasets. But the normalization term is intractable and hard to compute for most $f_\theta(x)$. Instead of approximating the Z_θ term using techniques like MCMC or variational inference, Hyvärinen and Dayan [2005] proposes to consider the following optimization problem, the *score matching* process,

$$\min_\theta J(\theta) = \mathbb{E}_0[||\nabla_x \log p_\theta(x) - \nabla_x \log p_0(x)||^2]. \tag{2.4}$$

Here, for simplicity, we let θ_0 to be the true parameter and denote $\mathbb{E}_0$ and p_0 as the expectation of X under θ_0 and the true p.d.f. respectively. By considering the gradient of $\log p_\theta(x)$, we bypass the unknown term Z_θ since

$$\nabla_x \log p_\theta(x) = \nabla_x \log f_\theta(x). \tag{2.5}$$

The next step is to cancel the $\nabla_x \log p_0(x)$ term, using integration by parts under some regularity assumptions. Note that for $i = 1, 2, \cdots, p$,

$$\begin{aligned}
&\int_{\mathbb{R}^p} (\partial_i \log p_\theta(x) - \partial_i \log p_0(x))^2 p_0(x) dx \\
&= \int_{\mathbb{R}^p} [(\partial_i \log p_\theta(x))^2 - \frac{2\partial_i p_0(x)}{p_0(x)} \partial_i \log p_\theta(x) + (\partial_i \log p_0(x))^2] p_0(x) dx \\
&= \int_{\mathbb{R}^p} (\partial_i \log p_\theta(x))^2 p_0(x) dx - 2 \int_{\mathbb{R}^p} \partial_i p_0(x) \partial_i \log p_\theta(x) dx + \text{const} \\
&= \mathbb{E}_0[(\partial_i \log p_\theta(x))^2 + \partial_{ii} \log p_\theta(x)] + \text{const},
\end{aligned} \tag{2.6}$$

where we use ∂_i to denote the first-order derivative w.r.t. x_i and

$$\int_{\mathbb{R}^p} \partial_i p_0(x) \partial_i \log p_\theta(x) dx = -\int_{\mathbb{R}^p} p_0(x) \partial_i \log p_\theta(x) dx. \tag{2.7}$$

Thus we can rewrite (2.4) as

$$\min_\theta \mathbb{E}_0[\text{tr}(\nabla_{xx} \log p_\theta(x)) + \frac{1}{2} ||\nabla_x \log p_\theta(x)||^2]. \tag{2.8}$$

In practice, we consider its empirical form.

$$\min_\theta \hat{J}_n(\theta) = \frac{1}{n} \sum_{i=0}^{n} ||\nabla_x \log p_\theta(x_i) - \nabla_x \log p_0(x_i)||^2, \tag{2.9}$$

or

$$\min_\theta \frac{1}{n} \sum_{i=0}^{n} \text{tr}(\nabla_{xx} \log p_\theta(x_i)) + \frac{1}{2} ||\nabla_x \log p_\theta(x_i)||^2. \tag{2.10}$$

We can further show that the score matching problem can indeed recover the true θ_0 if our parametric model is not degenerated[2] i.e. for $\theta_1, \theta_2 \in \Theta$, if $\theta_1 \neq \theta_2$, $p_{\theta_1}(x) \neq p_{\theta_2}(x)$ on a positive-measure set. This also establishes the consistency of the empirical estimator.

[2]Even the parametrization is degenerated, it can still be useful since we only care about the p not θ itself.

Theorem 2.3.1. *If our parametric model is non-degenerated i.e. for $\theta_1, \theta_2 \in \Theta$, if $\theta_1 \neq \theta_2$, $p_{\theta_1}(x) \neq p_{\theta_2}(x)$ on a positive-measure set and $p_0(x) > 0$ a.s. on its support, $J(\theta) = 0$ if and only if $\theta = \theta_0$.*

Proof. The sufficiency part is trivial. For the necessity part, note that if $p_\theta(x) \neq p_0(x)$ on a positive-measure set, we can always find a closed set with positive measure such that $||\nabla_x \log p_\theta - \nabla_x \log p_0||^2 > 0$ on it. Since it is a closed set, we have

$$\mathbb{E}_0[||\nabla_x \log p_\theta - \nabla_x \log p_0||^2] \geq \delta > 0. \tag{2.11}$$

□

Corollary 2.3.2 (Consistency)**.** *Let $\hat{\theta}_n = \arg\min_\theta \hat{J}_n(\theta)$. Under the non-degeneration condition and some conditions for the Law of Large Number (LLN) to hold, we have $\hat{\theta}_n \xrightarrow{\Pr} \theta_0$.*

Proof. It suffices to show that $J(\hat{\theta}_n) \xrightarrow{\Pr} 0$ since if $\hat{\theta}_n \not\xrightarrow{\Pr} \theta_0$, we have $\exists \delta > 0, \exists \epsilon > 0, \forall N > 0, \exists n > N$ such that

$$\Pr(||\hat{\theta}_n - \theta_0|| > \delta) > \epsilon, \tag{2.12}$$

which implies that

$$\Pr(||J(\hat{\theta}_n)|| > \delta) > \epsilon. \tag{2.13}$$

Note that for any $\theta \in \Theta$,

$$J(\theta) - J(\hat{\theta}_n) = (J(\theta) - \hat{J}(\theta)) + (\hat{J}(\theta) - \hat{J}(\hat{\theta}_n)) + (J(\hat{\theta}_n) - \hat{J}(\hat{\theta}_n)). \tag{2.14}$$

Thus it is easy to see that $J(\hat{\theta}_n) \xrightarrow{\Pr} 0$. □

Example 2.3.3 (Multivariate Gaussian)**.**

2.4 Fast Learning Rates for Plug-in Estimators

This section is my notes on Audibert and Tsybakov [2007]. The binary classification problem can be formally written as follows. Given a set of i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathcal{X} \subset \mathbb{R}^d$ and $Y_i \in \mathcal{Y} = \{0, 1\}$, we would like to find a function $f(X)$ such that it minimizes a type of classification loss $\mathcal{L}(f)$. In practice, we usually choose 0-1 loss or misclassification loss as our loss function i.e.

$$\mathcal{L}(f) := \Pr(f(X) \neq Y). \tag{2.15}$$

Let $\eta(X) = \Pr(Y = 1|X)$ and $f^*(X) = \mathbb{1}_{\{\eta(X) \geq 1/2\}}$. Then we have

$$f^* = \min_{f: \mathcal{X} \to \mathcal{Y}} \mathcal{L}(f). \tag{2.16}$$

Proof. Note that

$$\begin{aligned}\Pr(f(X) \neq Y) &= \mathbb{E}[\mathbb{1}_{\{f(X)\neq Y\}}] \\ &= \mathbb{E}_X[\eta(X)\mathbb{1}_{\{f(X)=1\}} + (1-\eta(X))\mathbb{1}_{\{f(X)=0\}}].\end{aligned} \tag{2.17}$$

Thus to minimize $\mathcal{L}(f)$, it suffices to minimize $\eta(x)\mathbb{1}_{\{f(x)=1\}}+(1-\eta(x))\mathbb{1}_{\{f(x)=0\}}$ for each $x \in \mathcal{X}$. Thus $f(x)$ should be equal to 1 when $\eta(x) \geq 1-\eta(x)$ i.e. $f(x) = \mathbb{1}_{\{\eta(x)\geq 1/2\}} = f^*(x)$. □

In practice, we need to estimate the f^* using our observations (training data). Let $\hat{f}$ be our estimation and thus the excess risk can be defined as

$$\mathcal{E}(\hat{f}) = \mathbb{E}[\mathcal{L}(\hat{f})] - \mathcal{L}(f^*). \tag{2.18}$$

Here, we take expectation on $\mathcal{L}(\hat{f})$ since $\hat{f}$ is a function of training data. Thus the expectation here should be on $\{(X_i, Y_i)\}_{i=1}^n$. For convenience, we use $\mathbb{E}$ to denote the expectation on all random variables. Intuitively, we should find a procedure to reduce the excess risk[3]. First, we can rewrite the excess risk as follows.

$$\begin{aligned}\mathcal{E}(\hat{f}) &= \mathbb{E}[\mathbb{1}_{\{\hat{f}(X)\neq Y\}} - \mathbb{1}_{\{f^*(X)\neq Y\}}] \\ &= \mathbb{E}[\mathbb{1}_{\{\hat{f}\neq f^*\}}(\mathbb{1}_{\{\hat{f}\neq Y\}} - \mathbb{1}_{\{f^*\neq Y\}})].\end{aligned} \tag{2.20}$$

We can further decompose the set $\hat{f} \neq f^*$ into four parts such that

$$\begin{aligned}\mathbb{1}_{\{\hat{f}\neq Y\}} - \mathbb{1}_{\{f^*\neq Y\}} &= \mathbb{1}_{\{\eta(X)\geq 1/2\}} \quad \text{when } f^*(X)=1, Y=1, \\ \mathbb{1}_{\{\hat{f}\neq Y\}} - \mathbb{1}_{\{f^*\neq Y\}} &= -\mathbb{1}_{\{\eta(X)\geq 1/2\}} \quad \text{when } f^*(X)=1, Y=0, \\ \mathbb{1}_{\{\hat{f}\neq Y\}} - \mathbb{1}_{\{f^*\neq Y\}} &= -\mathbb{1}_{\{\eta(X)<1/2\}} \quad \text{when } f^*(X)=0, Y=1, \\ \mathbb{1}_{\{\hat{f}\neq Y\}} - \mathbb{1}_{\{f^*\neq Y\}} &= \mathbb{1}_{\{\eta(X)<1/2\}} \quad \text{when } f^*(X)=0, Y=0.\end{aligned} \tag{2.21}$$

Thus

$$\begin{aligned}\mathcal{E}(\hat{f}) &= \mathbb{E}[\mathbb{1}_{\{\hat{f}\neq f^*\}}[(2\eta(X)-1)\mathbb{1}_{\{\eta(X)\geq 1/2\}} + (1-2\eta(X))\mathbb{1}_{\{\eta(X)<1/2\}}]] \\ &= \mathbb{E}[\mathbb{1}_{\{\hat{f}\neq f^*\}}|2\eta(X)-1|].\end{aligned} \tag{2.22}$$

This expression shows that the excess rate mainly depends on the part where $\hat{f} \neq f^*$ and $\eta(X)$ is away from $1/2$. This is intuitive since when $\eta(X)$ is away from $1/2$, it means that we have strong confidence on which class X belongs to. Thus if $\hat{f}$ and f^* give different answers, it should incur more risk difference.

To estimate the excess risk of a estimator $\hat{f}$, there are three types of assumption on $\eta(X)$ mentioned in Audibert and Tsybakov [2007]. The first one is a *complexity assumption on the regression function* (CAR), which is of the following form. Under the CAR, the rates of convergence of plug-in estimators can not be faster than $n^{-1/2}$.

[3]For any class of functions $\mathcal{F}$ and for any function $f \in \mathcal{F}$, we have decomposition of its risk as

$$\mathcal{L}(f) - \mathcal{L}(f^*) = (\mathcal{L}(f) - \min_{f'\in\mathcal{F}} \mathcal{L}(f')) + (\min_{f'\in\mathcal{F}} \mathcal{L}(f') - \mathcal{L}(f^*)). \tag{2.19}$$

Here our excess risk only refers to the second part.

Assumption 1 (CAR)**.** *The regression function $\eta(X)$ belongs to a class of function Σ such that*

$$\mathcal{H}(\epsilon, \Sigma, L_p) \leq A_* \epsilon^{-\rho}, \ \forall \epsilon > 0, \tag{2.23}$$

for some constant $\rho > 0$ and $A_* > 0$. Here, $\mathcal{H}(\epsilon, \Sigma, L_p)$ denotes the ϵ-entropy of Σ w.r.t. a L_p norm with $1 \leq p \leq \infty$[4]

The second one is a *complexity assumption on the decision set* (CAD), which is of the following form. This assumption is usually used in empirical risk minimization (ERM) type estimators. However, there is no clear connection between CAR and CAD.

Assumption 2 (CAD)**.** *The decision set $\{x : \eta(x) \geq 1/2\}$ belongs to a class of decision set $\mathcal{G}$ such that*

$$\mathcal{H}(\epsilon, \mathcal{G}, d_\Delta) \leq A_* \epsilon^{-\rho}, \ \forall \epsilon > 0, \tag{2.24}$$

for some constant $\rho > 0$ and $A_* > 0$. Here, $\mathcal{H}(\epsilon, \mathcal{G}, d_\Delta)$ denotes the ϵ-entropy of Σ w.r.t. $d_\Delta(G_1, G_2) = \Pr_X(G_1 \Delta G_2)$.

Audibert and Tsybakov [2007] provides a detailed discussion on how the two assumptions can lead to different rates of convergence for plug-in and ERM type estimators. One simple conclusion is that the two assumptions are in disfavor of the plug-in estimators. The goal of Audibert and Tsybakov [2007] is to show that the plug-in estimator can actually achieve fast rates of convergence i.e. $O(n^{-1})$ or even $O(e^{-n})$ with a new margin assumption (MA).

Assumption 3 (MA)**.** *The regression function $\eta(X)$ satisfies that*

$$\Pr_X(0 < |\eta(X) - 1/2| \leq t) \leq C t^\alpha, \ \forall 0 < t, \tag{2.25}$$

with some constants $\alpha \geq 0$ and $C > 0$.

One trivial case is when $\alpha = 0$. When $\alpha = \infty$, $\eta(X)$ is bounded away from $1/2$, which is advantageous in classification. I think MA indicates that the two classes should be well separated in $\mathcal{X}$. Here, I also calculate some examples to better understand MA.

Example 2.4.1 (Gaussian Mixture Model)**.** *Consider the following mixture model*

$$\begin{aligned} Y &\sim \textit{Bernoulli}(p), \\ X &\sim Y\mathcal{N}(\mu, I_d) + (1 - Y)\mathcal{N}(0, I_d), \end{aligned} \tag{2.26}$$

where we assume $||\mu||_2 = 1$ for simplicity. We use $\phi_\mu(x)$ to denote the p.d.f. of $\mathcal{N}(\mu, I_d)$. Then we have

$$\begin{aligned} \Pr(Y = 1|X) &= \frac{p\phi_\mu(x)}{p\phi_\mu(x) + (1-p)\phi_0(x)} \\ &= \frac{p}{p + (1-p)e^{-\mu^T x + 1/2}}. \end{aligned} \tag{2.27}$$

[4]Currently, I don't really understand the intuition behind this assumption. Why do we put an assumption on the set that η belongs to?

Here we only consider the case when $t < 1/2$. For $t \geq 1/2$, clearly $\Pr(|\eta(X) - \frac{1}{2}| \leq t) = 1$. *Note that*

$$\begin{aligned}\Pr(|\eta(X) - \frac{1}{2}| \leq t) &= \Pr(|\frac{p}{p + (1-p)e^{-\mu^T x + 1/2}} - \frac{1}{2}| \leq t) \\ &= \Pr(C_1 \leq \mu^T X \leq C_2),\end{aligned} \tag{2.28}$$

where

$$\begin{aligned}C_1 &= \frac{1}{2} - \log\{\frac{p(1+2t)}{q(1-2t)}\}, \\ C_2 &= \frac{1}{2} - \log\{\frac{p(1-2t)}{q(1+2t)}\}.\end{aligned} \tag{2.29}$$

Let $\Phi(t)$ denotes the c.d.f. of $\mathcal{N}(0,1)$. Then we have

$$\Pr(|\eta(X) - \frac{1}{2}| \leq t) = p(\Phi(C_2 - 1) - \Phi(C_1 - 1)) + q(\Phi(C_2) - \Phi(C_1)). \tag{2.30}$$

One can easily find a polynomial bound for $\Pr(|\eta(X) - 1/2| \leq t)$ *using polynomial approximation of $\Phi(x)$*[5]*. My numerical results show that $\alpha = 0.7$ and $C = 2$ can be a good bound for most cases.*

Example 2.4.2 (Uniform Distribution on a unit ball)**.** *Suppose that $X \sim$ Uniform$(B_1(0))$, where $B_1(0)$ is the ball centered at the origin with radius 1 in $\mathbb{R}^d$. Let*

$$\eta(X) = \frac{1}{2} - \frac{1}{2}||X||_2. \tag{2.31}$$

Then we have

$$\begin{aligned}\Pr(|\eta(X) - \frac{1}{2}| \leq t) &= \Pr(||X||_2 \leq 2t) \\ &= 2^d t^d,\end{aligned} \tag{2.32}$$

which satisfies the MA condition.

Under the MA condition, (2.22) can be further expanded as follows.

$$\begin{aligned}\mathcal{E}(\hat{f}) &= \mathbb{E}[\mathbb{1}_{\{\hat{f} \neq f^*, |\eta(X) - 1/2| \leq \delta\}} |2\eta(X) - 1|] + \mathbb{E}[\mathbb{1}_{\{\hat{f} \neq f^*, |\eta(X) - 1/2| > \delta\}} |2\eta(X) - 1|] \\ &\leq 2\delta^{\alpha+1} + \mathbb{E}[\mathbb{1}_{\{\hat{f} \neq f^*, |\eta(X) - 1/2| > \delta\}} |2\eta(X) - 1|].\end{aligned} \tag{2.33}$$

For plug-in estimator $\hat{f}^{PI}(X) = \mathbb{1}_{\{\hat{\eta}(X) \geq 1/2\}}$, we further have

$$\mathcal{E}(\hat{f}) \leq 2\delta^{\alpha+1} + 2\mathbb{E}[\mathbb{1}_{\{|\eta(X) - \hat{\eta}(X)| > \delta\}} |\eta(X) - \hat{\eta}(X)|]. \tag{2.34}$$

This upper bound indicates that the rate of convergence of the excess risk can be controlled by the estimation error of regression function $\eta(X)$. Similar to the approach used in (2.33), the upper bound of the excess risk can be further refined if our estimation of the regression function can achieve exponential rate of convergence.

[5] See the Computation section in Wikipedia

Lemma 2.4.3. *Let $\mathcal{P}$ be a set of probability measure on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\hat{\eta}$ be an estimation of η such that there exist some constants $C_1 > 0$ and $C_2 > 0$ and a sequence $\{a_n\}$ s.t. $a_n > 0$, and for almost all x w.r.t. $\Pr_X$, we have*

$$\sup_{\Pr \in \mathcal{P}} \Pr^{\otimes n}(|\eta(x) - \hat{\eta}(x)| \geq \delta) \leq C_1 e^{-C_2 a_n \delta^2}, \ \forall \delta > 0. \tag{2.35}$$

Then the plug-in estimator $\hat{f}^{PI}$ satisfies that

$$\sup_{\Pr \in \mathcal{P}} \mathcal{E}_{\Pr}(\hat{f}^{PI}) \leq C a_n^{-(1+\alpha)/2}, \tag{2.36}$$

where $\mathcal{E}_{\Pr}$ indicates that the expectations in $\mathcal{E}$ is on measure $\Pr$.

Proof. □

Bibliography

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. ISSN 00905364. URL `http://www.jstor.org/stable/25463570`.