Machine Learning Notes

This note is based on Machine Learning at DIKU

7 December 2021

Chapter 2. Concentration of Measure Inequalities

2.5.1 Relaxations of the kl-inequality: Pinsker's and refined Pinsker's in equations

By denoting the right hand side of kl inequality (2.98) by δ, we obtain that with probability greater than $(1-\delta)$,

$$\mathbf{kl}(\hat{p}\|p) \leqslant \frac{\ln \frac{n+1}{\delta}}{n}\,. \tag{2.100}$$

This leads to an implicit bound on p, which is not very intuitive and not always convenient to work with. In order to understand the behavior of the kl inequality better we use a couple of its relaxations. The first relaxation is known as Pinsker's inequality, see (Cover and Thomas 2006, Lemma 11.6.1).

NB. According to the right side of kl inequality (2.98)[14], and let $\delta \overset{\text{def}}{=} (n+1)e^{n\varepsilon}$ [15], therefore, (2.98) will become

$$\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \tfrac{1}{n}\ln\tfrac{n+1}{\delta}\right) \leqslant \delta\,, \tag{2.101}$$

that is to say, with probability greater than $(1-\delta)$, we have $\mathbf{kl}(\hat{p}\|p) \leqslant \frac{1}{n}\ln\frac{n+1}{\delta}$.

Lemma 2.16 (Pinsker's inequality)**.**

$$\mathbf{KL}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2\,, \tag{2.102}$$

where $\|p-q\|_1 = \sum_{x\in\mathcal{X}}|p(x)-q(x)|$ *is the* L_1*-norm.*

Corollary 2.17 (Pinsker's inequality for the binary kl divergence)**.**

$$\mathbf{kl}(p\|q) \geqslant \tfrac{1}{2}\big(|p-q|+|(1-p)-(1-q)|\big)^2 = 2(p-q)^2\,. \tag{2.103}$$

NB. We have known that by definition,

$$\mathbf{KL}(p\|q) = \mathbf{E}_p\left[\ln\frac{p(X)}{q(X)}\right] = \begin{cases}\sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}\,, & \text{if } \mathcal{X} \text{ is discrete;}\\ \int_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}dx\,, & \text{if } \mathcal{X} \text{ is continuous,}\end{cases} \tag{2.104a}$$

$$\mathbf{kl}(p\|q) = \mathbf{KL}([1-p,p]\|[1-q,q]) = p\ln\tfrac{p}{q} + (1-p)\ln\tfrac{1-p}{1-q}\,. \tag{2.104b}$$

Back to the inequality in Lemma 2.16, therefore, we have: for the binary kl divergence,

$$\mathbf{KL}(p\|q) = \textstyle\sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)} = \sum_{x\in\mathcal{X}} p(x)\left(\ln p(x) - \ln q(x)\right) \overset{\text{def}}{=} \text{LHS}\,, \tag{2.105a}$$

$$\tfrac{1}{2}\|p-q\|_1^2 = \tfrac{1}{2}\big(\textstyle\sum_{x\in\mathcal{X}}|p(x)-q(x)|\big)^2 \overset{\text{def}}{=} \text{RHS}\,, \tag{2.105b}$$

$$\mathbf{kl}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2 = \tfrac{1}{2}\left(|p-q|+|(1-p)-(1-q)|\right)^2 = \tfrac{1}{2}\left(2|p-q|\right)^2 = 2|p-q|^2 = 2(p-q)^2\,. \tag{2.105c}$$

[14] that is, $\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) \leqslant (n+1)e^{-n\varepsilon}$.

[15] then we obtain that $-n\varepsilon = \ln\frac{\delta}{n+1}$ and $\varepsilon = \frac{1}{n}\ln\frac{n+1}{\delta}$.

Qustion: Unlike Corollary 2.17, I'm not very sure how to demonstrate Lemma 2.16.

By applying Corollary 2.17 to (2.100) we obtain that with probability greater than $(1-\delta)$,

$$|p-\hat{p}| \leqslant \sqrt{\frac{\mathbf{kl}(\hat{p}\|p)}{2}} \leqslant \sqrt{\frac{\ln\frac{n+1}{\delta}}{2n}}\,. \tag{2.106}$$

Recall that Hoeffding's inequality assures that with probability greater than $(1-\delta)$,

$$p \leqslant \hat{p} + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,. \tag{2.107}$$

Thus, in the worst case the kl inequality is only weaker by the $\ln(n+1)$ factor and in fact the $\ln(n+1)$ factor can be reduced by a more careful analysis, see (Maurer 2004; Langford and Schapire 2005). Next we show that the kl inequality can actually be significantly tighter than Hoeffding's inequality. For this we use refined Pinsker's inequality, see (Marton 1996; Samson 2000), (Boucheron, Lugosi, and Massart 2013, Lemma 8.4).

NB. Let $X_1, ..., X_n$ be i.i.d. Bernoulli with bias p and q and let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias, then $\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) \leqslant (n+1)e^{-n\varepsilon}$. If we let $\delta \overset{\text{def}}{=} (n+1)e^{-n\varepsilon}$, we will get $\varepsilon = -\frac{1}{n}\ln\frac{\delta}{n+1}$, which means, with probability greater than $(1-\delta)$, it holds $\mathbf{kl}(\hat{p}\|p) \leqslant \frac{1}{n}\ln\frac{n+1}{\delta}$. According to the combination of Corollary 2.17 and (2.100), we have: with probability greater than $(1-\delta)$,

$$2(\hat{p}-p)^2 \leqslant \mathbf{kl}(\hat{p}\|p) \leqslant \tfrac{1}{n}\ln\tfrac{n+1}{\delta}\,, \tag{2.108a}$$

$$|\hat{p}-p| \leqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{n+1}{\delta}}\,, \tag{2.108b}$$

$$p \leqslant \hat{p} + \sqrt{\tfrac{1}{2n}\ln\tfrac{n+1}{\delta}} = \hat{p} + \sqrt{\tfrac{1}{2}\varepsilon}\,. \tag{2.108c}$$

Recall that Hoeffding's inequality[16] states that: Let $X_1, ..., X_n$ be independent real-valued random variables, then for every $\varepsilon > 0$, that is, $\delta = e^{-2n\varepsilon^2} \in (0,1)$,

$$\mathbf{Pr}\left(\hat{p} - p \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.112a}$$

$$\mathbf{Pr}\left(\hat{p} - p \leqslant -\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.112b}$$

[16] According to Theorem 2.3, Corollary 2.4–2.5, we have: Let $X_i, ..., X_n$ be independent random variables, such that $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$, it holds that

$$\mathbf{Pr}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)\,, \tag{2.109a}$$

$$\mathbf{Pr}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu \leqslant -\varepsilon\right) = \mathbf{Pr}\left(\mu - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)\,, \tag{2.109b}$$

$$\mathbf{Pr}\left(\left|\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(-2n\varepsilon^2\right)\,. \tag{2.109c}$$

If we let $\delta \overset{\text{def}}{=} e^{-2n\varepsilon^2}$, we will have $n\varepsilon^2 = -\frac{1}{2}\ln\delta = \frac{1}{2}\ln\frac{1}{\delta}$. In this case, the Hoeffding's inequalities will become

$$\mathbf{Pr}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.110a}$$

$$\mathbf{Pr}\left(\mu - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,. \tag{2.110b}$$

If we let $\delta \overset{\text{def}}{=} 2e^{-2n\varepsilon^2}$, we will have $n\varepsilon^2 = -\frac{1}{2}\ln\frac{\delta}{2} = \frac{1}{2}\ln\frac{2}{\delta}$, and then the Hoeffding's inequality will hold

$$\mathbf{Pr}\left(\left|\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\right) \leqslant \delta\,. \tag{2.111}$$

which means, it assures that with probability greater than $(1-\delta)$, it holds

$$-\sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}} < -\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}} \leqslant \hat{p}-p \leqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}} < \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\,, \tag{2.113a}$$

$$\hat{p}-\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}} \leqslant \quad p \quad \leqslant \hat{p}+\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\,. \tag{2.113b}$$

Therefore, the only difference between Hoeffding's inequality and the kl inequality is the $\ln(n+1)$ factor, in the worst case. To demonstrate that the kl inequality can actually be significantly tighter than Hoeffding's inequality, we also present some details about refined Pinsker's inequality (Marton 1996; Samson 2000) (Boucheron, Lugosi, and Massart 2013, Lemma 8.4).

As presented in (Marton 1996),[17] it gives a proof of Talagrand's inequality, which admits an extension to contracting Markov chains. The proof is based on a new asymmetric notion of distance between probability measures, and bounding this distance by informational divergence. The author also analyses the bin packing problem for Markov chains as an application. Note that the distance that the author proposed is $d_2(p,r)$, and the d_2-distance can be extended to a distance between probability measures p^n and r^n defined on the product space $\mathcal{X}^n$. (We get an analogue of the $\bar{d}$-distance.)

[17](Marton 1996) Let $\{X_i\}_{i=-\infty}^{\infty}$ be a Markov chain (not necessarily homogeneous) taking values in a complete separable metric space $\mathcal{X}$, endowed with the Borel σ-algebra $\mathcal{B}$. We assume that the diagonal of the product space $\mathcal{X}\times\mathcal{X}$ is measurable with respect to the product σ-algebra $\mathcal{B}\times\mathcal{B}$. We introduce an asymmetric notion of how much a probability distribution p differs from another one, r. The Radon-Nikodym derivative $\frac{dr}{dp}$ is assumed to be ∞ on the set of singularity of r with respect to p.

Remark (Definition). *If p and r are probability measures on $\mathcal{X}$ then*

$$d_2(p,r) = \left[\int \left|1-\tfrac{dr}{dp}(y)\right|_+^2 dp(y)\right]^{1/2}. \tag{2.114}$$

Notice that if μ is a probability measure on $\mathcal{X}$, and both p and r are absolutely continuous with respect to μ with Radon-Nikodym derivatives $f=\frac{dp}{d\mu}, g=\frac{dr}{d\mu}$, respectively, then $d_2(p,r)$ can be written as

$$d_2(p,r) = \left[\int \left|1-\tfrac{g}{f}\right|_+^2 f d\mu\right]^{1/2}. \tag{2.115}$$

The functional $d_2(p,r)$ is analogous to variational distance (divided by 2), which is defined as

$$|p-r| = \frac{1}{2}\int \left|1-\tfrac{dr}{dp}(y)\right| dp(y) = \int \left|1-\tfrac{dr}{dp}(y)\right|_+ dp(y)\,. \tag{2.116}$$

Obviously, $|p-r| \leqslant d_2(p,r) \leqslant |p-r|^{1/2}$. It is easy to see that $d_2(p,r) = \inf\left[\int \mathbf{Pr}(Z\neq y|Y=y)^2 dp(y)\right]^{1/2}$, where the infinum is taken over all joint distributions $\mathrm{dist}(Y,Z)$ with marginals $p=\mathrm{dist}(Y), r=\mathrm{dist}(Z)$. This is again analogous to the fact that $|p-r| = \inf \mathbf{Pr}(Z\neq Y)$, with the infimum taken over the same set of joint distributions.

Remark (Lemma 3.2 (Marton 1996)).

$$\begin{aligned} &(i) \quad d_2(p,r) \leqslant [2\mathbf{D}(p\|r)]^{1/2}\,,\\ &(ii) \quad d_2(r,p) \leqslant [2\mathbf{D}(p\|r)]^{1/2}\,. \end{aligned} \tag{2.117}$$

There is a simple inequality by Pinsker between variational distance and information divergence. Our next move is to prove a similar inequality for d_2. We shall have two different inequalities, since both d_2 and informational divergence are asymmetric. Pinsker's inequality says that $|p-r| \leqslant [\mathbf{D}(p\|r)/2]^{1/2}$. For d_2 we have the above bounds. On the right-hand side of these formulae we have a factor 2 instead of the $1/2$ of Pinsker's inequality. (It can be shown that the factor 2 cannot be improved.) Note, however, that the d_2-distance cannot be overbounded by a constant multiple of variational distance.

As presented in (Samson 2000),[18,19,20] it proves concentration inequalities for some classes of Markov chains and Φ-mixing processes, with constants independent of the size of the sample, that extend the inequalities for product measures of Talagrand. This work is based on information inequalities put forward by Marton in case of contracting Markov chains. Using a simple duality argument on entropy, the results in this paper also include the family of logarithmic Sobolev inequalities for convex functions. Additionally, applications to bounds on supremum

[18](Samson 2000) In a recent series of striking papers, Talagrand deeply analysed the concentration of measure phenomenon in product space, with applications to various areas of probability theory. A first result at the origin of his investigation concerns deviation inequalities for product measures $P = \mu_1 \otimes \cdots \otimes \mu_n$ on $[0,1]^n$. Namely, for every convex function f on $[0,1]^n$, with Lipschitz constant $\|f\|_{\mathrm{Lip}} \leqslant 1$, and for every $t \geqslant 0$,

$$P(|f - M| \geqslant t) \leqslant 4\exp(-t^2/4)\,, \tag{2.118}$$

where M is a median of f for P. This Gaussian-type bound may be considered as an important generalisation of the classical inequalities for sums of independent random variables. The deviation inequality (2.118) is a consequence of a concentration inequality on sets which takes the following form. To measure the "distance" of a point $\boldsymbol{x} \in \mathbb{R}^n$ to a set A, consider the functional (see "convex hull", (Talagrand 1995), Chapter 4),

$$f_{\mathrm{conv}}(A, \boldsymbol{x}) = \sup_{\alpha} \inf_{y \in A} \left(\sum_{i=1}^{n} \alpha_i \mathbb{I}(x_i \neq y_i) \right), \tag{2.119}$$

where the supremum is over all vectors $\alpha = (\alpha_i)_{1 \leqslant i \leqslant n}, \alpha_i \geqslant 0, \sum_{i=1}^n \alpha_i^2 = 1$. If we let $A_t^{\mathrm{conv}} = \{\boldsymbol{x} \in \mathbb{R}^n, f_{\mathrm{conv}}(A, \boldsymbol{x}) \leqslant t\}$, Talagrand shows that for every $t \geqslant \sqrt{2\log(1/P(A))}$,

$$P(A_t^{\mathrm{conv}}) \geqslant 1 - \exp\left[-\tfrac{1}{2} \left(t - \sqrt{2\log \tfrac{1}{\mathbf{Pr}(A)}} \right)^2 \right]. \tag{2.120}$$

Besides the convex hull approximation, Talagrand considers two other approximations on product spaces for which he proves similar concentration properties. One of the main features of these inequalities is that they are independent of the dimension of the product space, that is, of the size of the sample. We will be mainly concerned with extensions of the convex hull approximation in this work.

[19](Samson 2000) Recently, an alternative, simpler, approach to some of Talagrand's inequalities was suggested by Ledoux (Ledoux 1997) on the basis of log-Sobolev inequalities. Introduce, for every function g on $\mathbb{R}^n$, the entropy functional,

$$\mathrm{Ent}_P(g^2) = \int g^2 \log g^2 dP - \int g^2 dP \log \int g^2 dP\,. \tag{2.121}$$

Then, it can easily be shown that, for every product measure P on $[0,1]^n$ and for every separately convex function f, $\mathrm{Ent}_P(e^f) \leqslant \frac{1}{2}\int |\nabla f|^2 e^f dP$, where ∇f denotes the usual gradient of f on $\mathbb{R}^n$ and $|\nabla f|$ its Euclidean length. This inequality easily implies deviation inequalities of the type of (2.118). Indeed, the preceding log-Sobolev inequality may be turned into a differential inequality on the Laplace transform of convex Lipschitz functions, which then yields tail estimates by Chebyshev's inequality. This type of argument may be pushed further to recover most of Talagrand's deviation inequalities for functions (Ledoux 1997). It however does not seem to succeed for deviations under the median (or for concave functions).

[20](Samson 2000) Let now P denote the law of the sample X on $\mathbb{R}^n$. For every probability measures Q and R on $\mathbb{R}^n$, let $\mathcal{U}(Q,R)$ denote the set of all probability measures on $\mathbb{R}^n \otimes \mathbb{R}^n$ with marginals Q and R. Define

$$d_2(Q,R) = \inf_{\Pi \in \mathcal{U}(Q,R)} \sup_{\alpha} \iint \sum_{i=1}^{n} \alpha_i(y) \mathbb{I}(x_i \neq y_i)\, d\Pi(x,y)\,, \tag{2.122}$$

where the $\sup_\alpha$ is over all vectors of positive functions $\alpha = (\alpha_1, ..., \alpha_n)$, with $\int \sum_{i=1}^n \alpha_i^2(y) dR(y) \leqslant 1$. As a main result, we show in Theorem 1 below that, for every probability measure Q on $\mathbb{R}^n$ with Radon-Nikodym derivative dQ/dP with respect to the measure P, $d_2(Q,P) \leqslant \|\Gamma\| \sqrt{2\mathrm{Ent}_P\left(dQ/dP\right)}$. Furthermore, $d_2(P,Q) \leqslant \|\Gamma\| \sqrt{2\mathrm{Ent}_P\left(dQ/dP\right)}$. Such Pinsker type inequalities have already been investigated by Marton for contracting Markov chains, and then by Dembo in the independent case (Marton 1996). Recently, Marton also obtained related bounds with a parameter readily comparable to $\|\Gamma\|$ (**marton1999measure**). Following these works, we could easily derive concentration in the form of (2.120) [and thus (2.118)] from these information inequalities. We however

of dependent empirical processes complete this work.

As presented in (Boucheron, Lugosi, and Massart 2013),[21] the authors take a step further and relax the bounded differences condition. They assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies $f(y) - f(x) \leqslant \sum_{i=1}^n c_i(x)\mathbb{I}(x_i \neq y_i)$, for some functions $c_i : \mathcal{X}^n \to [0,\infty), i = 1,...,n$. Instead of forcing the c_i to be bounded we only assume that they are bounded in "quadratic mean" in the sense that $v \stackrel{\text{def}}{=} \mathbf{E}\left[\sum_{i=1}^n c_i^2(X)\right]$, is finite. Under this assumption, the transportation method may be used as follows. Let Q be a probability distribution, absolutely continuous with respect to P, the distribution of X. Let $\mathbf{P}$ be a coupling of P and Q. Then

$$\mathbf{E}_Q[f] - \mathbf{E}_P[f] \leqslant \sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[c_i(X)\mathbf{P}(X_i \neq Y_i|X)\right], \tag{2.127}$$

which implies, by applying the Cauchy-Schwarz inequality twice,

$$\mathbf{E}_Q[f] - \mathbf{E}_P[f] \leqslant \left(\mathbf{E}_{\mathbf{P}}[c_i^2(X)]\right)^{1/2}\left(\mathbf{E}_{\mathbf{P}}[\mathbf{P}^2(X_i \neq Y_i|X)]\right)^{1/2} \tag{2.128a}$$

$$\leqslant \left(\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}[c_i^2(X)]\right)^{1/2}\left(\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[\mathbf{P}^2(X_i \neq Y_i|X)\right]\right)^{1/2}. \tag{2.128b}$$

Using their assumption on f, this implies

$$\mathbf{E}_Q[f] - \mathbf{E}_P[f] \leqslant \sqrt{v}\left(\inf_{\mathbf{P}\in\mathcal{P}(P,Q)}\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[\mathbf{P}^2(X_i \neq Y_i|X)\right]\right)^{1/2}. \tag{2.129}$$

Thus, by the road map laid down in the introduction of this chapter, if the authors can prove the inequality

$$\inf_{\mathbf{P}\in\mathcal{P}(P,Q)}\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[\mathbf{P}^2(X_i \neq Y_i|X)\right] \leqslant 2\mathbf{D}(Q\|P), \tag{2.130}$$

then Lemma 4.18 implies $\psi_{Z-\mathbf{E}[Z]}(\lambda) \leqslant v\lambda^2/2$ and the resulting sub-Gaussian tail inequality with variance factor v. Since Lemma 8.13 is applicable, it suffices to prove the transportation inequality above for $n = 1$. They solve the corresponding transportation cost problem first.

---

take a somewhat different route related to exponential integrability and log-Sobolev inequalities. TBC.

[21] (Boucheron, Lugosi, and Massart 2013) The next step is an analog of Pinsker's inequality in which d_2 plays the role of the total variation distance.

Remark (Lemma 8.4 (Boucheron, Lugosi, and Massart 2013)). *Let P and Q be probability distributions on a common measure space $(\Omega, \mathcal{A})$. If Q is absolutely continuous with respect to P, then*

$$d_2^2(Q,P) + d_2^2(P,Q) \leqslant 2\mathbf{D}(Q\|P). \tag{2.123}$$

Proof. Since $Q \ll P$, setting $q = dQ/dP$ we may write

$$d_2^2(Q,P) + d_2^2(P,Q) = \mathbf{E}_P\left[(1-q(X))_+^2\right] + \mathbf{E}_P\left[\tfrac{(q(X)-1)_+^2}{q(X)}\right]. \tag{2.124}$$

Moreover, defining $h(t) = (1-t)\log(1-t) + t$ for $t < 1$ and $h(1) = 1$, we may write

$$\mathbf{D}(Q\|P) = \mathbf{E}_P\left[h\left(1-q(X)\right)\right] = \mathbf{E}_P\left[h\left((1-q(X))_+\right)\right] + \mathbf{E}_P\left[h\left(-(q(X)-1)_+\right)\right], \tag{2.125}$$

and the result follows by the inequalities

$$h(t) \geqslant t^2/2 \quad \text{for } t \in [0,1], \quad \text{and} \quad h(-t) \geqslant \tfrac{t^2}{2(1+t)} \quad \text{for } t \geqslant 0. \tag{2.126}$$

(recall Exercise 2.8). ∎

Lemma 2.18 (Refined Pinsker's inequality)**.**

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}}\,. \tag{2.131}$$

Corollary 2.19 (Refined Pinsker's inequality)**.** *If $q > p$ then*

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2q}\,. \tag{2.132}$$

Corollary 2.20 (Refined Pinsker's inequality)**.** *If $\mathbf{kl}(p\|q) \leqslant \varepsilon$ then*

$$q \leqslant p + \sqrt{2p\varepsilon} + 2\varepsilon\,. \tag{2.133}$$

NB. **Qustion:** I'm not sure how to demonstrate Lemma 2.18.

We first present Pinsker's inequality[22] here, that is to say,

$$\mathbf{KL}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2 = \tfrac{1}{2}\left(\textstyle\sum_{x\in\mathcal{X}}|p(x)-q(x)|\right)^2\,, \tag{2.138}$$

where $\|p-q\|_1 = \sum_{x\in\mathcal{X}}|p(x)-q(x)|$ is the L_1-norm.

According to Lemma 2.18, if $q > p$, then $1-q < 1-p$, and because of $p, q \in [0,1]$, it holds

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}} = \frac{(p-q)^2}{2q} + \frac{(p-q)^2}{2(1-p)} \geqslant \frac{(p-q)^2}{2q}\,. \tag{2.139}$$

---

[22] Wikipedia (Pinsker's inequality) In information theory, Pinsker's inequality, named after its inventor Mark Semenovich Pinsker, is an inequality that bounds the total variation distance (or statistical distance) in terms of the Kullback-Leibler divergence. The inequality is tight up to constant factors.

Formal statement. Pinsker's inequality states that, if P and Q are two probability distributions on a measurable space (X, Σ), then

$$\delta(P,Q) \leqslant \sqrt{\tfrac{1}{2}\mathbf{D}_{KL}(P\|Q)}\,, \tag{2.134}$$

where

$$\delta(P\|Q) = \sup\left\{|P(A)-Q(A)|\,\middle|\,A\in\Sigma \text{ is a measurable event}\right\}, \tag{2.135}$$

is the total variation distance (or statistical distance) between P and Q and

$$\mathbf{D}_{KL}(P\|Q) = \mathbf{E}_P\left(\log\tfrac{\mathrm{d}P}{\mathrm{d}Q}\right) = \int_X\left(\log\tfrac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P\,, \tag{2.136}$$

is the Kullback-Leibler divergence in nats. When the sample space X is a finite set, the Kullback-Leibler divergence is given by

$$\mathbf{D}_{KL}(P\|Q) = \sum_{i\in X}\left(\log\tfrac{P(i)}{Q(i)}\right)P(i)\,. \tag{2.137}$$

Note that in terms of the total variation norm $\|P-Q\|$ of the signed measure $(P-Q)$, Pinsker's inequality differs from the one given above by a factor of two, $\|P-Q\| \leqslant \sqrt{2\mathbf{D}_{KL}(P\|Q)}$. A proof of Pinsker's inequality uses the partition inequality for f-divergences.

Alternative version. There is an alternative statement of Pinsker's inequality in some literature that relates information divergence to variation distance, $\mathbf{D}(P\|Q) \geqslant \frac{1}{2\ln 2}\mathbf{V}^2(p,q)$, in which $\mathbf{V}(p,q) = \sum_{x\in\mathcal{X}}|p(x)-q(x)|$ is the variation distance between two probability density functions p and q on the same alphabet $\mathcal{X}$. This form of Pinsker's inequality shows that "convergence in divergence" is strong notion than "convergence in variation distance".

Inverse problem. A precise inverse of the inequality cannot hold: for every $\varepsilon > 0$, there are distributions P_ε, Q with $\delta(P_\varepsilon, Q) \leqslant \varepsilon$ but $\mathbf{D}_{KL}(P_\varepsilon\|Q) = \infty$. An easy example given by the two-point space $\{0,1\}$ with $Q(0) = 0, Q(1) = 1$ and $P_\varepsilon(0) = \varepsilon, P_\varepsilon(1) = 1-\varepsilon$. However, an inverse inequality holds on finite spaces X with a constant depending on Q. More specifically, it can be shown that with the definition $\alpha_Q \overset{\text{def}}{=} \min_{x\in X:Q(x)>0} Q(x)$ we have for any measure P which is absolutely continuous to Q, $\frac{1}{2}\mathbf{D}_{KL}(P\|Q) \leqslant \frac{1}{\alpha_Q}\delta(P,Q)^2$. As a consequence, if Q has full support (i.e., $Q(x) > 0$ for all $x \in X$), then $\delta(P,Q)^2 \leqslant \frac{1}{2}\mathbf{D}(P\|Q) \leqslant \delta(P,Q)^2/\alpha_Q$.

Therefore, Corollary 2.19 is demonstrated.

Question: It seems that we still require one more condition, which is $q = 1 - p$, before we reach the result in Corollary 2.19. Never mind, I was wrong, it doesn't have to.

Then if $\mathbf{kl}(p\|q) \leqslant \varepsilon$, we will get that

$$\text{RHS} = p + \sqrt{2p\varepsilon} + 2\varepsilon = \left(\sqrt{p} + \sqrt{\tfrac{\varepsilon}{2}}\right)^2 + \tfrac{3}{2}\varepsilon^2 = \left(\sqrt{2\varepsilon} + \tfrac{1}{2}p\right)^2 + \tfrac{3}{4}p^2; \tag{2.140a}$$

$$\tfrac{(p-q)^2}{2q} \leqslant \mathbf{kl}(p\|q) \leqslant \varepsilon, \tag{2.140b}$$

$$(p-q)^2 = p^2 - 2pq + q^2 \leqslant 2q\varepsilon, \quad \because p, q \in [0,1], \tag{2.140c}$$

$$-\sqrt{2q\varepsilon} \leqslant p - q \leqslant \sqrt{2q\varepsilon}, \quad q - \sqrt{2q\varepsilon} \leqslant p \leqslant q + \sqrt{2q\varepsilon}, \tag{2.140d}$$

$$q^2 - 2(p+\varepsilon)q + p^2 \leqslant 0 \tag{2.140e}$$

$$[q - (p+\varepsilon)]^2 \leqslant (p+\varepsilon)^2 - p^2 = 2p\varepsilon + \varepsilon^2 \tag{2.140f}$$

$$p + \varepsilon - \sqrt{2p\varepsilon + \varepsilon^2} \leqslant q \leqslant p + \varepsilon + \sqrt{2p\varepsilon + \varepsilon^2} \tag{2.140g}$$

Question: I'm not sure how to demonstrate Corollary 2.20 as well.

By applying Corollary 2.20 to (2.100) we obtain that with probability greater than $(1-\delta)$,

$$p \leqslant \hat{p} + \sqrt{\frac{2\hat{p}\ln\frac{n+1}{\delta}}{n}} + \frac{2\ln\frac{n+1}{\delta}}{n}. \tag{2.141}$$

Note that when $\hat{p}$ is close to zero, then latter inequality is much tighter than Hoeffding's inequality. Finally, we note that although there is no analytic inversion of $\mathbf{kl}(\hat{p}\|p)$ it is possible to invert it numerically to obtain even tighter bounds than the relaxations above. Additionally, the bound in Theorem 2.15 can be improved slightly, see (Maurer 2004; Langford and Schapire 2005).

NB. Note that (2.100) means that: with probability greater than $(1-\delta)$, it holds that $\mathbf{kl}(\hat{p}\|p) \leqslant \varepsilon = \frac{1}{n}\ln\frac{n+1}{\delta}$. Besides, Corollary 2.20 gives that $q \leqslant p + \sqrt{2p\varepsilon} + 2\varepsilon$, if $\mathbf{kl}(p\|q) \leqslant \varepsilon$. Therefore,

$$p \leqslant \hat{p} + \sqrt{2\hat{p}\varepsilon} + 2\varepsilon = \hat{p} + \sqrt{2\hat{p}\tfrac{1}{n}\ln\tfrac{n+1}{\delta}} + 2\tfrac{1}{n}\ln\tfrac{n+1}{\delta} = \hat{p} + \sqrt{\tfrac{2\hat{p}}{n}\ln\tfrac{n+1}{\delta}} + \tfrac{2}{n}\ln\tfrac{n+1}{\delta}. \tag{2.142}$$

We already know that Hoeffding's inequality means that $p \leqslant \hat{p} + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}$ with probability greater than $(1-\delta)$. So when $\hat{p}$ is close to zero, the latter inequality is much tighter than Hoeffding's inequality.

Question: Doesn't it mean that smaller is tighter? Where am I wrong?

The bound in Theorem 2.15 [23] is $\mathbf{Pr}(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon) \leqslant (n+1)e^{-n\varepsilon}$, which can be improved slightly, see (Maurer 2004; Langford and Schapire 2005).

According to (Maurer 2004),[24] the author proves general exponential moment inequalities for averages of $[0,1]$-valued iid random variables and use them to tighten the PAC Bayesian Theo-

[23] where $X_1, ..., X_n$ are i.i.d. Bernoulli with bias p and $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ is the empirical bias.

[24] Throughout this note $X_1, ..., X_n$ are assumed to be IID random variables with values in $[0,1]$ and expectation $\mathbf{E}[X_i] = \mu$. We use $\mathbf{X}$ to denote the corresponding random vector $\mathbf{X} = (X_1, ..., X_n)$ with values in $[0,1]^n$ and $M(\mathbf{X})$ to denote its arithmetic mean $M(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$. For any $[0,1]$-valued random variables X use X' to denote the unique Bernoulli ($\{0,1\}$-valued) random variable with $\mathbf{Pr}(X' = 1) = \mathbf{E}[X'] = \mathbf{E}[X]$. Evidently $X'' = X', \forall X$. For $\mathbf{X} = (X_1, ..., X_n)$ we denote $\mathbf{X}' = (X_1', ..., X_n')$. We restate our principal bounds in a slightly more general way.

Theorem (1). *For all $n \geqslant 2$, $\mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}), \mu)\right)\right] \leqslant \mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}'), \mu)\right)\right] \leqslant e^{1/(12n)}\left(\frac{\pi n}{2}\right)^{1/2} + 2$. If the X_i are nontrivial Bernoulli variables (i.e., if $\mu \in (0,1)$) then there is a sequence c_n such that $1 \leqslant c_n \to \pi$ as $n \to \infty$ and $e^{-1/6}\left(\frac{n}{2\pi}\right)^{1/2} c_n + 2 \leqslant \mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}), \mu)\right)\right]$. In this case the expectation on the right is independent of μ.*

rem. The logarithmic dependence on the sample count in the enumerator of the PAC Bayesian bound is halved.

According to (Langford and Schapire 2005), the author discusses basic prediction theory and its impact on classification success evaluation, implications for learning algorithm design, and uses in learning algorithm execution. This tutorial is meant to be a comprehensive compilation of results which are both theoretically rigorous and quantitatively useful.

References

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart (2013). *Concentration inequalities: A non-asymptotic theory of independence*. Oxford university press.

Cover, Thomas M and Joy A Thomas (2006). *Elements of Information Theory*. Second Edition. Wiley Series in Telecommunications and Signal Processing.

Csiszár, Imre and János Körner (2011). *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press.

Feller, William (1957). *An Introduction to Probability Theory and Its Applications*. Second Edition. New York: John Wiley & Sons.

Langford, John and Robert Schapire (2005). "Tutorial on Practical Prediction Theory for Classification." In: *Journal of machine learning research* 6.3.

Ledoux, Michel (1997). "On Talagrand's deviation inequalities for product measures". In: *ESAIM: Probability and statistics* 1, pp. 63–87.

Marton, Katalin (1996). "A measure concentration inequality for contracting Markov chains". In: *Geometric & Functional Analysis GAFA* 6.3, pp. 556–571.

Maurer, Andreas (2004). "A note on the PAC Bayesian theorem". In: *arXiv preprint cs/0411099*.

Samson, Paul-Marie (2000). "Concentration of measure inequalities for markov chains and ϕ-mixing processes". In: *The Annals of Probability* 28.1, pp. 416–461.

Talagrand, Michel (1995). "Concentration of measure and isoperimetric inequalities in product spaces". In: *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques* 81.1, pp. 73–205.

The right side (i.e., $e^{1/(12n)}(\frac{\pi n}{2})^{1/2}+2$) is bounded above by $2\sqrt{n}$ for $n \geqslant 8$ and the left side (i.e., $e^{-1/6}(\frac{n}{2\pi})^{1/2}c_n+2$) is bounded below by $\sqrt{n}$ for $n \geqslant 2$, thus giving the simpler bounds ($\mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}),\mu)\right)\right] \leqslant 2\sqrt{n}$) and ($\sqrt{n} \leqslant \mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}),\mu)\right)\right]$) of the introduction. To prove Theorem 1 we need some auxilliary results. The first is Stirling's Formula:

Theorem (2). *For $n \in \mathbb{N}$, $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(g(n)/(12n)\right)$, with $0 < g(n) < 1$.*

We will use Theorem 2 in form of the following inequalities $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n}\right)$. The simple Lemma 3 shows that the expectation of a convex function of iid variables can always be bounded by the expectation of the corresponding Bernoulli variables. The next Lemma 4 is concerned with a series which can be viewed as a Rieman sum approximation an instance of the Beta-function.

Remark (Lemma 3). *Suppose that $f : [0,1]^n \to \mathbb{R}$ is convex. Then $\mathbf{E}[f(\mathbf{X})] \leqslant \mathbf{E}[f(\mathbf{X}')]$. If f is permutation symmetric in its arguments and $\boldsymbol{\theta}(k)$ denotes the vector $\boldsymbol{\theta}(k) = (1,...,1,0,...,0)$ in $\{0,1\}^n$, whose first k coordinates are 1 and whose remaining $(n-k)$ coordinates are zero, we also have $\mathbf{E}[f(\mathbf{X}')] = \sum_{k=0}^{n}\binom{n}{k}(1-\mu)^{n-k}\mu^k f\left(\boldsymbol{\theta}(k)\right)$.*

Remark (Lemma 4). *For $n \geqslant 2$ the sequence $c_n = \sum_{k=1}^{n-1}\frac{1}{\sqrt{k(n-k)}}$, satisfies $1 \leqslant c_n \leqslant \pi$, and $c_n \to \pi$ as $n \to \infty$.*