Lower Bounds on Estimation Error

Andrew D. McRae

1. Why Lower Bounds?

Consider a probabilistic model

$$X \sim P_\theta$$

where X takes values in some set S, and θ is some unknown parameter in a set Θ of possible parameter values (we use parametric notation for its familiarity and simplicity, but much of the material in these notes applies to more general models as well). We try to come up with an estimator of θ, which is a function $\hat{\theta}(X)$ that we hope is close to the true parameter θ.

Many results in statistics literature yield an upper bound on error: given some nonnegative error function w on $\Theta \times \Theta$, we try to show that, given that θ is the true parameter in our model, the quantity $w(\hat{\theta}(X), \theta)$ can be bounded from above with high probability and/or in expectation.

The upper bounds on error which we can derive depend, in general, on the parametric distribution (i.e., what is the form of P_θ) and on the set Θ of valid parameters under consideration.

An upper bound tells us that the error cannot be too large: when applying statistics to the real world, this is the most interesting kind of result, since it gives us some assurance that our results are approximately correct.

However, merely knowing that our estimation procedure does not perform too badly does not tell us whether or not it is possible to do better. It is therefore

of great interest (both to the theory of statistics and to the development of practical algorithms) to find *lower* bounds on error. If we can find a lower bound of error in a model that (approximately) matches an upper bound for an estimator that we or somebody else has developed, we have proved that we cannot come up with a more accurate estimation procedure. On the other hand, a failure to find matching bounds suggests that we have more work to do, either in coming up with better estimators or in improving our theoretical bounds!

2. Review of Bayesian Estimation and Bayes Risk

The simplest situation for finding lower bounds is Bayesian statistics. Here, we assume that the parameter of interest is itself random, and we can exactly calculate the expected error of the optimal estimator.

Suppose we have a probability distribution Π on our parameter space Θ; we call this the "prior" distribution of θ. We must now write our original model as a factored model with conditional distributions: θ has density $\pi(\theta)$,[1] and, conditioned on θ, X has density $p(x \mid \theta) = p_\theta(x)$. The risk of an estimator for a given parameter θ is defined as

$$R(\hat{\theta}, \theta) = \mathbf{E}[w(\hat{\theta}(X), \theta) \mid \theta]$$

and the average risk with respect to Π is

$$\begin{aligned} R_\Pi(\hat{\theta}) &= \mathbf{E}_\Pi \, R(\hat{\theta}, \theta) \\ &= \int_\Theta \pi(\theta) \int_S w(\hat{\theta}(x), \theta) p(x \mid \theta) \; dx \; d\theta \end{aligned}$$

The Bayes risk (the minimum possible average risk of the model) is defined as $R_\Pi = \inf_{\hat{\theta}} R_\Pi(\hat{\theta})$.

We define the standard "posterior" conditional density by Bayes' formula:

$$q(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{\int_\Theta p(x \mid \theta)\pi(\theta) \; d\theta} = \frac{p(x \mid \theta)\pi(\theta)}{r(x)}$$

where r is the marginal density of x.

[1] All densities are with respect to some appropriate measure. We will abuse notation somewhat by writing, for example, $\int f(x) \; dx$ instead of $\int f(x) \; \mu(dx)$.

By reversing the order of integration and applying Bayes' formula to the expression for the average risk, we can calculate

$$R_\Pi(\hat{\theta}) = \int_S r(x) \int_\Theta w(\hat{\theta}, \theta) q(\theta \mid x) \, d\theta \, dx$$

Then, the optimal ("Bayes") estimator $\hat{\theta}_B$ is given by[2]

$$\hat{\theta}_B(x) = \operatorname*{argmin}_{\theta' \in \Theta} \int_\Theta w(\theta', \theta) q(\theta \mid x) \, d\theta$$

and then $R_\Pi = R_\Pi(\hat{\theta}_B)$.

3. Minimax Risk

Although the Bayes risk derived in the previous section has an elegant expression, it is not always quite what we want. There is not usually any particular reason to think that any problem we find out in the woods is "average" in any sense; we are often more interested in the worst-case performance, which can be extremely different from an averaged risk. This motivates another definition of error, which is minimax risk:

$$R = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta)$$

In words, R is the smallest quantity such that no estimator has better (average) error than R for all possible parameter values.

3.1. Relation to Bayes Risk

Given that we are (at least at present) more interested in the minimax risk, why did we bother talking about Bayes risk? The reason is the following simple yet wonderful fact:

$$R \geq \sup_\Pi R_\Pi \tag{1}$$

Clearly, an average risk cannot be greater than a maximal risk. In fact, equality holds in many interesting situations (see Example 1 below), but, since we are primarily interested in *lower* bounds on minimax risk, we will not explore this here.

[2] Although the expression for $\hat{\theta}_B$ is somewhat intimidating at first glance, it often has a simple formula; for example, if w is the squared Euclidean norm, $\hat{\theta}_B(x)$ is simply the posterior mean of the conditional distribution $q(\theta \mid x)$.

Example 1. For a very simple situation where (1) is useful (and optimal), suppose $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$ for $\theta \in \mathbf{R}^d$. What is the minimax risk for squared Euclidean error? We can get an easy upper bound on the minimax risk by taking $\hat{\theta}(X) = X$, in which case we can easily compute the average error to be $R(\hat{\theta}, \theta) = \mathbf{E}_\theta \|X - \theta\|^2 = d\sigma^2$ for all $\theta \in \Theta$.

Now, consider a prior distribution $\theta \sim \mathcal{N}(0, \tau^2 I_d)$. It is easily calculated that

$$\theta \mid X \sim \mathcal{N}\left(\frac{1}{1+\frac{\sigma^2}{\tau^2}} X, \frac{\sigma^2}{1+\frac{\sigma^2}{\tau^2}} I_d\right)$$

so the Bayes estimator (which as noted before, is the posterior mean for this error function) is $\hat{\theta}_B(X) = \frac{1}{1+\frac{\sigma^2}{\tau^2}} X$, which has risk $\frac{d\sigma^2}{1+\frac{\sigma^2}{\tau^2}}$ for every X (so the average risk is the same).

Thus we see that $R \geq \frac{d\sigma^2}{1+\frac{\sigma^2}{\tau^2}}$; this holds for every $\tau > 0$, so we can take $\tau \to \infty$ to get that $R \geq d\sigma^2$.

There are other interesting (and much less trivial examples) where one can construct a prior (or sequence of priors) to get a bound on minimax risk. For example, Donoho and Johnstone (1994) computed asymptotic error rates for Gaussian means in ℓ_p balls with such a method. We will see another method for computing the rate in the case of $p = 1$ later.

4. Packing, Hypothesis Testing, and Fano's Method

In this section, we assume that Θ is a metric space, and our error function is the metric, which we denote d.

4.1. A Reduction to Multiple Hypothesis Testing

Suppose we have a finite collection of points $A \subset \Theta$ (of size N) such that for some $\delta > 0$, and for every distinct $\theta, \theta' \in A$, we have $d(\theta, \theta') \geq 2\delta$. We call A a 2δ-*packing set* of Θ. We know that any $\theta \in \Theta$ can be within distance δ of at most one $\theta_i \in A$.

We can now find a lower bound on the risk of the estimation problem by finding a lower bound on the probability of error for the multiple-hypothesis

testing problem on Θ_N. Indeed, if $\hat{\theta}$ is any estimator, let $\phi_{\hat{\theta}} : S \to A$ be the test defined by

$$\phi_{\hat{\theta}}(X) = \operatorname*{argmin}_{\theta \in A} d(\theta, \hat{\theta}(X))$$

Then, for any $\theta \in A$,

$$\begin{aligned}\mathbf{P}_\theta(d(\theta, \hat{\theta}) \geq \delta) &\geq \mathbf{P}_\theta(\phi_{\hat{\theta}}(X) \neq \theta) \\ &\geq \inf_{\phi} \mathbf{P}_\theta(\phi(X) \neq \theta)\end{aligned}$$

where the last infimum is over all tests $\phi : S \to A$. If we can find a minimax lower bound on the probability of testing error (say, $\inf_\phi \sup_{\theta \in A} \mathbf{P}_\theta(\phi(X) \neq \theta) \geq \epsilon$), then we have shown that $R \geq \delta\epsilon$.[3]

4.2. An Information-Theoretic Lower Bound for Testing Risk

It remains to find a minimax lower bound on the probability of testing error

$$R' = \inf_{\phi} \sup_{\theta \in A} \mathbf{P}_\theta(\phi(X) \neq \theta)$$

We note that this is another kind of minimax risk: our parameter space is the finite set A, and the error function is $w(\theta, \theta') = \mathbf{1}_{\{\theta \neq \theta'\}}$. We will therefore lower bound this risk by lower bounding the Bayes risk of a particular prior over A.

Specifically, we will chose the uniform prior on A: $\pi(\theta) = 1/N$. Note that for any test ϕ, $\theta \to X \to \phi(X)$ forms a Markov chain. Fano's inequality from information theory (Cover & Thomas, 2006) states that

$$\mathbf{P}(\phi(X) \neq \theta) \geq \frac{H(\theta \mid X) - \log 2}{\log N}$$

where $H(\theta \mid X)$ is the conditional entropy of θ given X. Noting the identity $H(\theta \mid X) = H(\theta) - I(X; \theta)$, where $I(X; \theta)$ is the mutual information between X and θ, and $H(\theta) = \log|A|$ is the entropy of θ, we have

$$\mathbf{P}(\phi(X) \neq \theta) \geq 1 - \frac{I(X;\, \theta) + \log 2}{\log N}$$

[3] We can usually choose A such that $\epsilon \geq 1/2$. Note that this method proves something a bit stronger than a lower bound on average error (which could be the mean of a heavy-tailed error distribution): we show that for certain $\theta \in \Theta$, the estimator *will* make a large error with non-negligible probability.

If we can choose A such that $I(X;\,\theta)$ is much smaller than $\log N$, we will have our desired lower bound on testing error. We will next show a couple of simple ways to bound the mutual information $I(X;\,\theta)$; we will later see an example of how to find a large packing set A.

4.3. Upper Bounds on Mutual Information

We can calculate the mutual information between the parameter θ and the observed random variable X by first noting that X has marginal density

$$g(x) = \sum_{\theta \in A} \pi(\theta) p(x \mid \theta) = \frac{1}{N} \sum_{\theta \in A} p_\theta(x)$$

The mutual information is then

$$\begin{aligned} I(X;\,\theta) &= \sum_{\theta \in A} \int_S p(x \mid \theta)\pi(\theta) \log \frac{p(x \mid \theta)\pi(\theta)}{g(x)\pi(\theta)}\, dx \\ &= \frac{1}{N} \sum_{\theta \in A} \int_S p_\theta(x) \log \frac{p_\theta(x)}{g(x)}\, dx \\ &= \frac{1}{N} \sum_{\theta \in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, G) \end{aligned}$$

where G is the distribution with density g.

Even if one has a convenient expression for the Kullback-Leibler divergence between members of the model, it is rare to have a tractable formula for $\mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, G)$. Therefore, we will try find some tractable upper bounds.

One simple bound is the following: for any distribution Q with density q,

$$\begin{aligned}
I(X;\,\theta) &= \frac{1}{N}\sum_{\theta\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, G) \\
&= \frac{1}{N}\sum_{\theta\in A} \int_S p_\theta \log\frac{p_\theta}{g} \\
&= \frac{1}{N}\sum_{\theta\in A} \int_S \left(p_\theta \log\frac{p_\theta}{q} - p_\theta \log\frac{g}{q}\right) \\
&= \frac{1}{N}\sum_{\theta\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, Q) - \int_S \left(\frac{1}{N}\sum_{\theta\in A} p_\theta\right) \log\frac{g}{q} \\
&= \frac{1}{N}\sum_{\theta\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, Q) - \mathrm{D}_{\mathrm{KL}}(G \,\|\, Q) \\
&\le \frac{1}{N}\sum_{\theta\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, Q) \\
&\le \max_{\theta\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, Q)
\end{aligned}$$

In other words, replacing G in the mutual information expression with an arbitrary (possibly more convenient, computationally) distribution gives us an upper bound on the mutual information. If we don't want to compute an average KL divergence over all of the distributions P_θ, we can upper bound by the largest divergence.

A weaker but commonly-used bound chooses Q to be one of the distributions $P_\theta, \theta \in A$ to get:[4]

$$I(X;\,\theta) \le \max_{\theta,\theta'\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, P_{\theta'})$$

In the case of Gaussian distributions, this weakening often does not hurt us too much. However, in situations (e.g., Poisson distributions) where the P_θ's can be singular with respect to one another, being able to choose a distribution Q such that no P_θ is singular with respect to Q is very valuable.

[4]This is often derived in much more direct fashion by noting that KL divergence is jointly convex in its arguments, so $\mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, G) \le \frac{1}{N}\sum_{\theta'\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, P_{\theta'})$, and therefore $I(X;\,\theta) \le \frac{1}{N^2}\sum_{\theta,\theta'\in A} \mathrm{D}_{\mathrm{KL}}(P_\theta \,\|\, P_{\theta'})$.

4.4. Example: Gaussian Mean in ℓ_1 Ball

We consider again the Gaussian shift model $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$, but this time we constrain the mean θ to be in an ℓ_1 ball in $\mathbf{R}^d$, i.e.

$$\Theta = \{\theta \in \mathbf{R}^d \mid \|\theta\|_1 \leq r\}$$

for some $r > 0$. Consider the maximum likelihood estimator

$$\hat{\theta}(X) = \operatorname*{argmin}_{\theta \in \Theta} \|X - \theta\|_2^2$$

If the true parameter is $\theta_0 \in \Theta$, we clearly must have $\|X - \hat{\theta}\|_2^2 \leq \|X - \theta_0\|_2^2$. With some algebra, we obtain that

$$\begin{aligned}
\|\hat{\theta} - \theta_0\|_2^2 &\leq 2\langle X - \theta_0, \hat{\theta} - \theta_0 \rangle \\
&\leq 2 \sup_{\theta \in \Theta} \langle X - \theta_0, \theta - \theta_0 \rangle \\
&\leq 2 \sup_{\theta \in \Theta} \|X - \theta_0\|_\infty \|\theta - \theta_0\|_1 \\
&\leq 4r \|X - \theta_0\|_\infty
\end{aligned}$$

If $Z = X - \theta_0$, we know, from standard Gaussian tail bounds and the union bound, that

$$\mathbf{P}(\|Z\|_\infty > t) \leq 2de^{-t^2/2\sigma^2}$$

so

$$\begin{aligned}
\mathbf{E}\|Z\|_\infty &= \int_0^\infty \mathbf{P}(\|Z\|_\infty > t)\ dt \\
&= \sigma \int_0^\infty \mathbf{P}(\|Z\|_\infty > \sigma t)\ dt \\
&\leq \sigma \left(\sqrt{2 \log 2d} + \int_{\sqrt{2 \log 2d}}^\infty 2de^{-t^2/2}\ dt \right) \\
&\leq \sigma \left(\sqrt{2 \log 2d} + 2d \int_{\sqrt{2 \log 2d}}^\infty \frac{t}{\sqrt{2 \log 2d}} e^{-t^2/2}\ dt \right) \\
&= \sigma \left(\sqrt{2 \log 2d} + 2d \frac{1}{2d\sqrt{2 \log 2d}} \right) \\
&\leq 2\sigma \sqrt{2 \log 2d}
\end{aligned}$$

Therefore

$$\mathbf{E}_{\theta_0}\|\hat{\theta} - \theta_0\|_2^2 \leq 8\sqrt{2}\sigma r\sqrt{\log 2d}$$

We want to show that, within a constant, this error bound is optimal (at least, for certain values[5] of R, σ, and d). In particular, we will show that it is optimal when $R \approx \sigma\sqrt{\log d}$, i.e., $c_1\sigma\sqrt{\log d} \leq r \leq c_2\sigma\sqrt{\log d}$ for some $c_2 \geq c_1 > 0$.

To apply Fano's method to this problem, we first find a bound on the mutual information: taking $Q = \mathcal{N}(0, \sigma^2 I_d)$, we have

$$\sup_{\theta\in\Theta} \mathrm{D}_{\mathrm{KL}}(P_\theta \parallel Q) \leq \frac{r^2}{2\sigma^2}$$

To get an error probability of at least $1/2$, we therefore want to find ϵ such that there is an 2ϵ-packing of Θ if size at least 2^{k+1}, where k is an integer such that $k \geq \frac{r^2}{2\sigma^2} + \log 2$. It is shown in (Kühn, 2001) that there is a universal constant c such that[6]

$$\begin{aligned} 2\epsilon &\geq cr\sqrt{\frac{\log\frac{d}{k+2} + 1}{k+2}} \\ &\geq cr\sqrt{\frac{\log d}{r^2/2\sigma^2 + 2 + \log 2}} \\ &\geq \sqrt{2}c\sigma\sqrt{\log d} \end{aligned}$$

Then, Fano's method gives us an expected minimax risk of at least

$$\begin{aligned} R &= \sup_{\theta\in\Theta}\inf_{\hat{\theta}} \mathbf{E}_\theta\|\theta - \hat{\theta}\|_2^2 \\ &\geq \frac{1}{\sqrt{2}}c\sigma^2\log d \\ &\geq \frac{1}{\sqrt{2}}\frac{c}{c_2}r\sigma\sqrt{\log d} \end{aligned}$$

[5]For example, if $\sigma\sqrt{\log d} \gg r$, we could get a better error bound of r^2 by simply taking $\hat{\theta} = 0$.

[6]The result in (Kühn, 2001) requires $\log d \leq k + 2 \leq d$, which holds for the values of R, σ, and d that are of interest to us. Furthermore, the result is for covering numbers, but it is easily shown that the packing number for distance 2ϵ is at least the covering number for distance ϵ, so this distinction comes out in the constant.

Note that the rate (and, to some extent, the constraints on R, σ, and d) match the (asymptotic) results found by Donoho and Johnstone (1994).

4.5. Literature on Fano's Method

Most of the material on Fano's method (as well as historical references) can be found in the following sources.

Good overviews of Fano's method (as well as Assouad's method, which is another class of minimax risk bounding methods that uses binary hypothesis testing—an interesting topic for another day) is found in (Huber, 1997; Yu, 1997).

Many similar methods as those presented here, as well as the application of Fano's method to finding minimax rates for density estimation, can be found in (Yang & Barron, 1999).

Gushchin (2003), inspired by Birgé (2005, 4), develops some very interesting extensions to Fano's method in which one can replace Kullback-Leibler divergences by arbitrary f-divergences. I am still investigating the usefulness of these more general approaches. This source also contains another nice summary of the state-of-the-art bounds that can be attained from Fano's method.

References

Birgé, L. (2005). A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, *51*, 1611–1615.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. New York, NY, USA: Wiley-Interscience.

Donoho, D. L., & Johnstone, I. M. (1994). Minimax risk over ℓ_p-balls for ℓ_p-error. *Probability Theory and Related Fields*, *99*(2), 277–303.

Gushchin, A. A. (2003). On Fano's lemma and similar inequalities for the minimax risk. *Theory of Probability and Mathematical Statistics*, *67*, 29–41.

Huber, C. (1997). Lower bounds for function estimation. In *Festschrift for Lucien Le Cam* (Chap. 15, pp. 245–258). Springer.

Kühn, T. (2001). A lower estimate for entropy numbers. *J. Approx. Theory*, *110*(1), 120–124.

Yang, Y., & Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Stat.*, *27*, 1564–1599.

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* (Chap. 29, pp. 423–435). Springer.