IV. Statistical Estimation and Classification

Probability: An Extremely Concise Review

1. A scalar-valued random variable X is completely characterized by its distribution function
$$F_X(u) = \mathrm{P}\left(X \leq u\right).$$
This is also called the **cumulative distribution function** (cdf). $F_X(u)$ is monotonically increasing in u; it goes to one as $u \to \infty$ and goes to zero as $u \to -\infty$.

2. If F_X is differentiable, then we can also characterize X using its **probability density function** (pdf)
$$f_X(x) = \left.\frac{dF_X(u)}{du}\right|_{u=x}$$
The density has the properties $f_X(x) \geq 0$ and
$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$
Events of interest are subsets[1] of the real line — given such an event/subset $\mathcal{E}$, we can compute the probability of $\mathcal{E}$ occurring as
$$\mathrm{P}\left(\mathcal{E}\right) = \int_{x\in\mathcal{E}} f_X(x)\, dx.$$

[1]Technically, it must be a subset of the real line that can be written as some combination of countable unions, countable intersections, and complements of intervals. You really have to know something about real analysis to construct a set that does not meet this criteria.

It is possible that a pdf exists even if F_X is not differentiable everywhere, for example:

$$F_X(u) = \begin{cases} 0, & u < 0, \\ u, & 0 \le u \le 1, \\ 1, & u \ge 1 \end{cases} \quad \text{has pdf} \quad f_X(x) = \begin{cases} 0, & x < 0, \\ 1, & 0 \le x \le 1, \\ 0, & x > 1. \end{cases}$$

3. The **expectation** of a function $g(X)$ of a random variable is

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

This is the "average value" of $g(X)$ in that given a series of realizations $X = x_1, X = x_2, \ldots,$ of X,

$$\frac{1}{M} \sum_{m=1}^{M} g(x_m) \to \mathrm{E}[g(X)], \quad \text{as } M \to \infty.$$

This fact is known as the **(weak) law of large numbers**.

4. The **moment** of X of degree p is the expectation of the monomial $g(x) = x^p$. The zeroth moment is always 1:

$$\mathrm{E}[X^0] = \mathrm{E}[1] = \int_{-\infty}^{\infty} f_X(x)\, dx = 1,$$

and the first moment is the **mean**:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$$

The **variance** is the second moment minus the mean squared:

$$\mathrm{var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{E}[(X - \mathrm{E}[X])^2].$$

This is sometime referred to as the "variation around the mean". Aside from the zeroth moment, there is nothing that says that the integrals above must converge; it is easy to construct examples of well-defined random variables where $\mathrm{E}[X] = \infty$.

5. A pair of random variables (X, Y) are completely described by their **joint distribution function** (joint cdf)[2]

$$F_{X,Y}(u, v) = \mathrm{P}\left(X \leq u, Y \leq v\right).$$

Again, if $F_{X,Y}$ is continuously differentiable, (X, Y) is also characterize by the density

$$f_{X,Y}(x, y) = \left.\frac{\partial F_{X,Y}(u, v)}{\partial u\, \partial v}\right|_{(u,v)=(x,y)}.$$

In this case, events of interest correspond to regions in the plane $\mathbb{R}^2$, and the probability of an event occurring is the integral of the density over this region.

6. From the joint pdf $f_{X,Y}(x, y)$, we can recover the individual **marginal pdfs** for X and Y using

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\; dy, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\; dx. \end{aligned}$$

The pair of densities $f_X(x)$, $f_Y(y)$ tell us how X and Y behave individually, but not how they *interact*.

[2]For fixed $u, v \in \mathbb{R}$, the notation $\mathrm{P}\left(X \leq u, Y \leq v\right)$ should be read as "the probability that X is $\leq u$ and Y is $\leq v$.

7. If X and Y do interact in a meaningful way, then observing one of them affects the distribution of the other. If we observe $X = x$, then with this knowledge, the density for Y becomes

$$f_Y(y|X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

This is a density over y; it is easy to check that it is positive everywhere and that it integrates to one. $f_X(y|X = x)$ is called the **conditional density** for Y given $X = x$.

8. We call X and Y **independent** if observing X tells us nothing about Y (and vice versa). This means

$$f_Y(y|X = x) = f_Y(y), \quad \text{for all } x \in \mathbb{R},$$

and

$$f_X(x|Y = y) = f_X(x), \quad \text{for all } y \in \mathbb{R}.$$

(If one of the statements above is true, then the other follows automatically.) Equivalently, independence means that the joint pdf is *separable*:

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y).$$

9. We can always factor the joint pdf in two different ways:

$$f_X(x) f_Y(y|X = x) = f_{X,Y}(x, y) = f_Y(y) f_X(x|Y = y).$$

At this point, we should be comfortable enough with what is going on that we can use $f_Y(y|x)$ as short-hand notation for $f_Y(y|X = x)$. Then we can rewrite the above in its more common form as

$$f_X(x) f_Y(y|x) = f_{X,Y}(x, y) = f_Y(y) f_X(x|y).$$

This factorization also gives us a handy way to compute the marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_X(x|y)\ dy.$$

It also yields **Bayes' equation**

$$f_X(x|y) = \frac{f_Y(y|x) f_X(x)}{f_Y(y)},$$

which is a fundamental relation for statistical inference.

10. All of the above extends in the obvious way to more than two random variables. A **random vector**

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}$$

is completely characterized by the density $f_X(\boldsymbol{x}) = f_X(x_1, \ldots, x_D)$ on $\mathbb{R}^D$. In general, we can factor the joint pdf as

$$f_X(\boldsymbol{x}) = f_{X_1}(x_1)\, f_{X_2}(x_2|x_1)\, f_{X_3}(x_3|x_2, x_1) \cdots f_{X_D}(x_D|x_1, \ldots, x_{D-1}).$$

11. The pth moment of a random vector X that maps into $\mathbb{R}^D$ is the collection of expectations of all monomials of order p. The mean of a random vector is a vector of length D:

$$\mathrm{E}[X] = \begin{bmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_D] \end{bmatrix},$$

the second moment is the $D \times D$ matrix of all correlations between entries:

$$\mathrm{E}[XX^{\mathrm{T}}] = \begin{bmatrix} \mathrm{E}[X_1^2] & \mathrm{E}[X_1X_2] & \cdots & \mathrm{E}[X_1X_D] \\ \vdots & & \ddots & \vdots \\ \mathrm{E}[X_DX_1] & \cdots & & \mathrm{E}[X_D^2] \end{bmatrix},$$

the third moment is the $D \times D \times D$ tensor $\mathrm{E}[X \otimes X \otimes X]$, where

$$\left(\mathrm{E}[X \otimes X \otimes X]\right)(i,j,k) = \mathrm{E}[X_iX_jX_k],$$

and so on. The **covariance** matrix contains all the pairs of second moments:

$$R_{i,j} = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])].$$

If $\boldsymbol{\mu}_X = \mathrm{E}[X]$ is the mean vector, we can write the covariance matrix succinctly in terms of the second moment as

$$\boldsymbol{R} = \mathrm{E}[XX^{\mathrm{T}}] - \boldsymbol{\mu}_X\boldsymbol{\mu}_X^{\mathrm{T}}$$

12. Given independent observations $X = \boldsymbol{x}_1,\ X = \boldsymbol{x}_2, \ldots,\ X = \boldsymbol{x}_M$ of a random vector X with unknown (or partially known) distribution, a completely reasonable way to estimate the mean vector is using

$$\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{x}_m.$$

If the mean $\boldsymbol{\mu}_X = \mathrm{E}[X]$ is known but the covariance is not, we can estimate the covariance using

$$\hat{\boldsymbol{R}} = \left(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{x}_m\boldsymbol{x}_m^{\mathrm{T}}\right) - \boldsymbol{\mu}_X\boldsymbol{\mu}_X^{\mathrm{T}}.$$

If both the mean and covariance are unknown, we first estimate the mean vector as above, then take

$$\hat{\boldsymbol{R}} = \left(\frac{1}{M-1}\sum_{m=1}^{M} \boldsymbol{x}_m \boldsymbol{x}_m^{\mathrm{T}}\right) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathrm{T}}.$$

The difference in the scaling is to ensure that $\mathrm{E}[\hat{\boldsymbol{R}}] = \boldsymbol{R}$ in both cases.

The Weak Law of Large Numbers

The WLLN is absolutely fundamental to machine learning (and really to all of probability and statistics). It basically formalizes the notion that given a series of independent samples of a random variable X, we can approximate $E[X]$ by averaging the samples. The WLLN states that if $X_1, X_2, \ldots$ are independent copies of a random variable X,

$$\frac{1}{N}\sum_{n=1}^{N} X_n \to \mathrm{E}[X] \quad \text{as } N \to \infty.$$

The only condition for this convergence is that X has finite variance.

We start by stating the main result precisely. Let X be a random variable with pdf $f_X(x)$, mean $\mathrm{E}[X] = \mu$, and variance $\mathrm{var}(X) = \sigma^2 < \infty$. We observe *samples* of X labeled $X_1, X_2, \ldots, X_N$. The X_i are independent of one another, and they all have the same distribution as X. We will show that the sample mean formed from a sample of size N:

$$M_N = \frac{1}{N}(X_1 + X_2 + \cdots + X_N),$$

obeys[3]

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) \;\leq\; \frac{\sigma^2}{N\epsilon^2},$$

where $\epsilon > 0$ is an arbitrarily small number. In the expression above, M_N is the only thing which is random; μ and σ^2 are fixed underlying properties of the distribution, N is the amount of data we see, and ϵ is something we can choose arbitrarily.

[3] This is a simple example of a *concentration bound*. It is not that tight; we will later counter inequalities of this type that are much more precise. But it is relatively simple and will serve our purpose here.

Notice that no matter how small ϵ is, the probability on the right hand side above goes to zero as $N \to \infty$. That is, for any fixed $\epsilon > 0$,

$$\lim_{N\to\infty} \mathrm{P}\left(|M_N - \mu| > \epsilon\right) = 0.$$

This result is follows from two simple but important tools known as the *Markov* and *Chebyshev* inequalities.

Markov inequality

Let X be a random variable that only takes positive values:

$$f_X(x) = 0, \quad \text{for } \; x < 0, \quad \text{or} \quad F_X(0) = 0.$$

Then

$$\boxed{\mathrm{P}\left(X \geq a\right) \;\; \leq \;\; \frac{\mathrm{E}[X]}{a} \quad \text{for all } a > 0.}$$

For example, the probability that X is more than 5 times its mean is 1/5, 10 times the mean is 1/10, etc. And this holds for **any distribution**.

The Markov inequality is easy to prove:

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x)\; dx \\
&\geq \int_a^\infty x f_X(x)\; dx \\
&\geq \int_a^\infty a f_X(x)\; dx \\
&= a \cdot \mathrm{P}\left(X \geq a\right)
\end{aligned}$$

and so $P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$.

Again, this is a very general statement in that we have assumed nothing about X other than it is positive. The price for the generality is that the bound is typically very loose, and does not usually capture the behavior of $\mathrm{P}(X \geq a)$. We can, however, cleverly apply the Markov inequality to get something slightly more useful.

Chebyshev inequality

The main use of the Markov inequality turns out to be its use in deriving other, more accurate deviation inequalities. Here we will use it to derive the **Chebyshev inequality**, from which the weak law of large numbers will follow immediately.

Chebyshev inequality: If X is a random variable with mean μ and variance σ^2, then

$$\boxed{\mathrm{P}(|X - \mu| > c) \;\leq\; \frac{\sigma^2}{c^2} \quad \text{for all } c > 0.}$$

The Chebyshev inequality follows immediately from the Markov inequality in the following way. No matter what range of values X takes, the quantity $|X-\mu|^2$ is always positive. Thus

$$\mathrm{P}\left(|X-\mu|^2 > c^2\right) \;\leq\; \frac{\mathrm{E}[|X-\mu|^2]}{c^2} = \frac{\sigma^2}{c^2}.$$

Since squaring $(\cdot)^2$ is monotonic (invertible) over positive numbers,

$$\mathrm{P}\left(|X-\mu|^2 > c^2\right) = \mathrm{P}(|X-\mu| > c) \;\leq\; \frac{\sigma^2}{c^2}.$$

We now have a bound which depends on the mean and the variance of X; this leads to a more accurate approximation of the probability.

Simple proof of the weak law of large numbers

We now turn to the behavior of the the sample mean

$$M_N = \frac{X_1 + X_2 + \cdots + X_N}{N},$$

where again the X_i are iid random variables with $\mathrm{E}[X_i] = \mu$ and $\operatorname{var} X_i = \sigma^2$. We know that

$$\mathrm{E}[M_N] = \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_N]}{N} = \frac{N\mu}{N} = \mu,$$

and since the X_i are independent,

$$\operatorname{var}(M_N) = \frac{\operatorname{var}(X_1) + \operatorname{var}(X_2) + \cdots + \operatorname{var}(X_N)}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

For any $\epsilon > 0$, a direct application of the Chebyshev inequality tells us that

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) \;\leq\; \frac{\sigma^2}{N\epsilon^2}.$$

The point is that this gets arbitrarily small as $N \to \infty$ no matter what ϵ was chosen to be. We have established, in some sense, that even though $\{M_N\}$ is a sequence of random numbers, it converges to something deterministic, namely μ.

WLLN: Let $X_1, X_2, \ldots$ be iid random variables as above. For **every** $\epsilon > 0$, we have

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) = \mathrm{P}\left(\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| > \epsilon\right) \longrightarrow 0,$$

as $N \to \infty$.

One of the philosophical consequences of the WLLN is that it tells us that probabilities can be estimated through **empirical frequencies**. Suppose I want to estimate the probability of and event A occurring related to some probabilistic experiment. We run a series of (independent) experiments, and set $X_i = 1$ if A occurred in experiment i, and $X_i = 0$ otherwise. Then given $X_1, \ldots, X_N$, we estimate the probability of A in a completely reasonable way, by computing the percentage of times it occurred:

$$p_{\text{empirical}} = \frac{X_1 + \cdots + X_N}{N}.$$

The WLLN tells us that

$$p_{\text{empirical}} \to \mathrm{P}\left(A\right), \quad \text{as } N \to \infty.$$

This lends some mathematical weight to our interpretation of probabilities as *relative frequencies*.

All of the above of course applies to functions of random variables. That is, if X is a random variable, and $g(X)$ is a function of that random variable with

$$\mathrm{var}(g(X)) = \mathrm{E}[(g(X) - \mathrm{E}[g(X)])^2] < \infty,$$

then given independent realizations $X_1, \dots, X_N$, we have

$$\frac{1}{N}\sum_{n=1}^{N} g(X_n) \to \mathrm{E}[g(X)]$$

as $N \to \infty$.

Minimum Mean-Square Error Estimation

Now we will take our first look at estimating variables that are themselves random, subject to a known probability law.

We start our discussion with a very basic problem. Suppose Y is a scalar random variable with a known pdf $f_Y(y)$. Here is a fun game: you guess what Y is going to be, then I draw a realization of Y corresponding to its probability law, then we see how close you were with your guess.

What is your best guess?

Well, that of course depends on what exactly we mean by "best", i.e. what price I pay for being a certain amount off. But if we penalize the *mean-squared error*, we know exactly how to minimize it.

Let g be your guess. The error in your guess is of course random (since the realization of Y is random), and so is the squared-error $(Y-g)^2$. We want to choose g so that the mean of the squared error is as small as possible:

$$\underset{g}{\text{minimize}} \quad \mathrm{E}[(Y-g)^2].$$

Expanding the squared error makes it clear how to do this:

$$\mathrm{E}[(Y-g)^2] = \mathrm{E}[Y^2] - 2g\,\mathrm{E}[Y] + g^2.$$

No matter what the first moment $\mathrm{E}[Y]$ and second moment $\mathrm{E}[Y^2]$ are (as long as they are finite), the expression above is a convex quadratic function in g, and hence is minimized when its first derivative (w.r.t. g) is zero, i.e. when

$$-2\,\mathrm{E}[Y] + 2g = 0 \quad \Rightarrow \hat{g} = \mathrm{E}[Y].$$

The mean squared error for this choice $\hat{g}$ is of course exactly the variance of Y,

$$\mathrm{E}[(Y-\hat{g})^2] = \mathrm{E}[(Y-\mathrm{E}[Y])^2] = \mathrm{var}(Y).$$

The story gets more interesting (and relevant) when we have multiple random variables, some of which we observe, some of which we do not. Suppose that two random variables (Y, Z) have joint pdf $f_{Y,Z}(y, z)$. Suppose that a realization of (Y, Z) is drawn, and I get to observe Z. What have I learned about Y?

If Y and Z are independent, then the answer is of course nothing. But if they are not independent, then the marginal distribution of Y changes. In particular, before the random variables were drawn, the (marginal) pdf for Y was

$$f_Y(y) = \int f_{Y,Z}(y, z)\; dz.$$

After we observe $Z = z$, we have

$$f_Y(y|Z = z) = \frac{f_{Y,Z}(y, z)}{f_Z(z)} = \frac{f_{Y,Z}(y, z)}{\int f_{Y,Z}(y, z)\; dy}.$$

Y is still a random variable, but its distribution depends on the value z that was observed for Z.

Now, given that I have observed $Z = z$, what is the best guess for Y? If by "best" we mean that which minimizes the mean squared error, it is the conditional mean. That is, the minimizer of

$$\underset{g}{\text{minimize}} \;\; \mathrm{E}[(Y-g)^2|Z = z]$$

is

$$\hat{g} = \mathrm{E}[Y|Z = z].$$

Notice that unlike before, $\hat{g}$ is not pre-determined, it depends on the outcome $Z = z$. We might denote

$$\hat{g}(z) = \mathrm{E}[Y|Z = z].$$

For a particular choice of z, the mean-squared error is the conditional variance

$$\begin{aligned} \mathrm{E}[(Y - \hat{g}(z))^2|Z = z] &= \mathrm{E}\left[(Y - \mathrm{E}[Y|Z = z])^2|Z = z\right] \\ &= \mathrm{E}[Y^2|Z = z] - (E[Y|Z = z])^2 \\ &= \mathrm{var}(Y|Z = z). \end{aligned}$$

So in general, not only does $\hat{g}$ depend on z, but its performance (its mean-square error) also depends on z.

We can also average over the draw of Z. First, note that since Z is a random variable, $\hat{g}$ is a priori also random, we might say

$$\hat{g}(Z) = \mathrm{E}[Y|Z].$$

Let me pause here because this is the point where many people start to get confused. The quantities

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy, \quad \text{and} \quad \mathrm{E}[Y|Z = z] = \int_{-\infty}^{\infty} y f_Y(y|Z = z)\, dy$$

are deterministic, but we re-emphasize that

$$\mathrm{E}[Y|Z] = \int_{\infty}^{\infty} y f_Y(y|Z)\ dy$$

is *random*. The above integrates out the randomness of Y, but not that of Z. (Note that in $\mathrm{E}[Y|Z=z]$ the randomness in Z is removed through direct observation.)

The "average estimate" is now

$$\begin{aligned}\mathrm{E}[g(Z)] &= \mathrm{E}[\mathrm{E}[Y|Z]] \\ &= \int \mathrm{E}[Y|Z=z] f_Z(z)\, dz \\ &= \mathrm{E}[Y].\end{aligned}$$

The identity $\mathrm{E}[\mathrm{E}[Y|Z]] = \mathrm{E}[Y]$ is known as the law of *iterated expectation* or *total expetation*. The inside E above integrates out the randomness in Y while the outside one integrates over Z, the result is a deterministic quantity.

Since $\mathrm{E}[g(Z)] = \mathrm{E}[Y]$, on average we are doing the same thing as if we didn't observe Z at all. But since we are adapting g to the draw of Z, we get better average performance. The mean square error (which is random with Z) is

$$\mathrm{E}[(Y - \hat{g}(Z))^2|Z] = \mathrm{E}[Y^2|Z] - (\mathrm{E}[Y|Z])^2 = \mathrm{var}(Y|Z).$$

The average performance is then

$$\mathrm{E}[\mathrm{E}[(Y - \hat{g}(Z))^2|Z]] = \mathrm{E}[\mathrm{var}(Y|Z)] \le \mathrm{var}(Y).$$

The last inequality, which follows from the law of total variance discussed below, means that on average, an (optimal) estimator using knowledge of Z will outperform an (optimal) estimator without knowledge of Z. Of course, when Y and Z are uncorrelated (meaning $\mathrm{E}[YZ] = \mathrm{E}[Y]\,\mathrm{E}[Z]$) we will have $\mathrm{E}[\mathrm{var}(Y|Z)] = \mathrm{var}(Y)$ and knowing Z makes no difference.

The law of total variance

Recall that for any random variable Y,

$$\mathrm{var}(Y) = \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2. \tag{1}$$

As we have seen above, $\mathrm{E}[Y|Z]$ is a random variable (where now the randomness is being caused by Z). Hence it also has a mean

$$\mathrm{E}[\mathrm{E}[Y|Z]] = \mathrm{E}[Y],$$

and a variance

$$\begin{aligned}\mathrm{var}(\mathrm{E}[Y|Z]) &= \mathrm{E}\left[(\mathrm{E}[Y|Z])^2\right] - \left(\mathrm{E}[\mathrm{E}[Y|Z]]\right)^2 \\ &= \mathrm{E}\left[(\mathrm{E}[Y|Z])^2\right] - \left(\mathrm{E}[Y]\right)^2. \end{aligned} \tag{2}$$

The quantity (again as we have seen above) $\mathrm{var}(Y|Z)$ is also a random variable; we can write its mean as

$$\begin{aligned}\mathrm{E}\left[\mathrm{var}(Y|Z)\right] &= \mathrm{E}\left[\ \mathrm{E}[(Y - \mathrm{E}[Y|Z])^2 \mid Z]\ \right] \\ &= \mathrm{E}\left[\mathrm{E}[Y^2|Z]\right] - \mathrm{E}\left[(\mathrm{E}[Y|Z])^2\right] \\ &= \mathrm{E}[Y^2] - \mathrm{E}\left[(\mathrm{E}[Y|Z])^2\right]. \end{aligned} \tag{3}$$

Adding together (2) and (3) and applying (1) gives us the cute expression

$$\mathrm{var}(Y) = \mathrm{E}\left[\mathrm{var}(Y|Z)\right] + \mathrm{var}(\mathrm{E}[Y|Z]).$$

This is known as the **law of total variance**. It basically says that you can decompose the variance in a random variable by the expected variance in "using Z to predict Y" and the variance of the prediction (condition expectation) itself. Note that all of the quantities above are non-negative, so we have the inequalities

$$\mathrm{E}[\mathrm{var}(Y|Z)] \le \mathrm{var}(Y) \quad \text{and} \quad \mathrm{var}(\mathrm{E}[Y|Z]) \le \mathrm{var}(Y).$$