Extrapolating expected accuracies for multi-class classification

Charles Zheng, Rakesh Achanta and Yuval Benjamini

November 15, 2016

Abstract

The difficulty of multi-class classification generally increases with the number of classes. Using data from a subset of the classes, can we predict how well a classifier will scale with an increased number of classes? Under the assumption that the classes are sampled exchangeably, and under the assumption that the classifier is generative (e.g. QDA or Naive Bayes), we show that the expected accuracy when the classifier is trained on k classes is the $k-1$st moment of a *conditional accuracy distribution*, which can be estimated from data. This provides the theoretical foundation for performance extrapolation based on pseudolikelihood, unbiased estimation, and high-dimensional asymptotics. We investigate the robustness of our methods to non-generative classifiers in simulations and one optical character recognition example.

1 Introduction

Machine learning models are becoming increasingly employed in scientific and industrial applications. A common problem in these settings in *multi-class classification*, where the goal is to label objects (e.g. images, sentences, etc.) from a set of finitely many labels. Example applications:

- In biology, labeling images of cancerous cells by the type of cancer.
- In language detection, labeling a sentence by the language of the sentence.

- In face recognition, labeling a photograph of a person with their name.

In the most traditional point of view, the actual implementation of machine learning models can be divided into two stages: *development* and *deployment*. In the development stage, engineers collect an initial dataset for the purpose of "training" a good model. Various competing models may be evaluated using a subset of the initial dataset. A final model may be selected on the basis of empirical performance. Then, in the deployment stage, the model is "deployed" in the real world: it is implemented in a larger decision-making system, and at this stage errors in the classification have real consequences.

The goal of the engineers in the development stage is to choose a model that will perform well in the deployment stage. For this purpose, it would be ideal if the initial "training" data was sufficient for obtaining a good estimate of the performance of the model in the deployment phase. However, often there is a mismatch between the training set and the population of post-deployment instances. One, the problem of "concept drift," is the issue where the training set is sampled from a different population than the deployment population: perhaps the deployment population also changes over time. Secondly, the training set may not be large enough to adequately estimate the error in the test set. In *multi-class classification* in particular, there arises a unique problem wherein the training set may not include all of the labels that exist in the deployment population. For example:

- In the initial set, only three types of cancer have been labelled, but the model is to deployed for the purpose of classifying 10 types of cancer.
- The language model is trained on data taken from 20 different languages, but after deployment, more languages will be addded.
- The face recognition model is initially trained to distinguish 200 individuals, but after deployment, it will be applied to classify a new, much larger set of individuals.

This is the problem that we address in this paper: suppose we have trained a classification model on one classification task (involving a small number of classes): can we estimate its performance on a *different* classification task, possibly involving a larger number of classes?

Of course, a number of specifying assumptions are needed to make sense of the problem. What do we mean by a "classification model," and how do

we define a model so that it makes sense to consider the same model being applied to different classification tasks? What kind of classification tasks are we considering? And what kind of assumptions are needed in linking the smaller classification task to the larger classification task? These assumptions are outlined in section **??**

1.1 Multi-class classification

- More details about examples like image recognition, face recognition, etc.
- Examples of learning algorithms? OVA, OVO
- Important: talking about LOSS functions. 0-1 loss, hierarchical loss functions
- Mention multi-label classification, but we don't address it in the paper.
- Introduce some of the formalism: we can think of a dataset as an empirical joint distribution.

2 Framework

2.1 Problem Formulation

This section lays out the basic framework which is necessary to *formulate* the problem. However, in the rest of the paper, we will also adopt some additional assumptions in order to solve the problem we pose in this section.

Let $\mathcal{Y}$ be a collection of labels and $\mathcal{X}$ be a space of feature vectors. For each label $y \in \mathcal{Y}$, there exists a distribution F_y supported on $\mathcal{X}$. Also suppose that there exists a *cost function* $C(\hat{y}, y)$ which measures the cost of incorrectly labelling an instance as $\hat{y}$ when the correct label is y. Further suppose that $0 \leq C(y, y') \leq \infty$ and $C(y, y) = 0$ for all $y, y' \in \mathcal{Y}$.

A *classification task* consists of a subset of labels, $\mathcal{S} \subset \mathcal{Y}$, and a prior distribution π over the label subset. A *classification rule* for the task consists of a function f which maps feature vectors $x \in \mathcal{X}$ to labels in $\mathcal{S}$:

$$f : \mathcal{X} \to \mathcal{S}.$$

The classification task defines the *risk* of a classification rule. Under the classification task, a label y is drawn from the distribution π. Then, we draw $x \sim F_y$. The label assigned by the classification rule is $\hat{y} = f(x)$. The *loss* incurred is $C(\hat{y}, y)$. The *risk* of the classification rule is the expected loss under the class distribution π:

$$\text{Risk}(f) = \mathbf{E}_\pi[C(\hat{y}, y)] = \int_{\mathcal{S}} d\pi(y) \int_{\mathcal{X}} C(f(x), y) dF_y(x).$$

A *classification model* $\mathcal{F}$ is an algorithm or procedure for producing classification rules given an empirical distributions $\hat{F}_y$ for each $y \in \mathcal{S}$, and a vector of prior probabilities π. The model maps a distribution G and a vector π to a classification rule f.

We consider *datasets* of size $|\mathcal{S}|r$ for a given classification task consisting r i.i.d. observations $x_i^{(y)} \sim F_y$ for each $y \in \mathcal{S}$. Then define

$$\hat{F}_y = \frac{1}{r} \sum_{i=1}^{r} \delta_{x_i^{(y)}}.$$

The sampling distribution of $\hat{F}_y$ is referred to as the size-r sampling distribution $\Pi_{y,r}$.

The n-sample *risk* of the classification model $\mathcal{F}$ is the expected risk of a classification rule $\hat{f} = \mathcal{F}(\hat{G})$ for a random dataset of size n for the classification task, $\hat{G} \sim \Pi_n$. That is,

$$\text{Risk}_n(\mathcal{F}; \pi) = \int \text{Risk}(\mathcal{F}(\{\hat{F}_y\}_{y \in \mathcal{S}}; \pi)) \prod_{y \in \mathcal{S}} d\Pi_{y,r}(\hat{F}_y).$$

The problem of *performance extrapolation* is as follows. Suppose we have two classification tasks: the ith classification task is specified by label subset $\mathcal{S}_i$, prior distribution π_i. We observe data from the first classification task consisting of a dataset of size n_1. The goal is to estimate the n_2-sample risk of a $\mathcal{F}$ on the second classification task, $\text{Risk}_{n_2}(\mathcal{F}; \pi_2)$.

2.2 Additional assumptions

In order to obtain a tractable solution to the problem of performance extrapolation, we make a number of special assumptions on the nature of the

classification tasks, and the classifiers themselves, which make the problem much easier.

Firstly, we assume that the label space $\mathcal{Y}$ is a continuum: in fact, that $\mathcal{Y}$ is a subset of d-dimensional Euclidean space. Note this is not such a strong assumption as it might seem, since cases where there are k discrete labels can be equivalently formulated as continous models where the the continuum can be partitioned into k equivalence classes, and in which the cost between two label y, y' is a function only of their equivalence classes.

We work with bounded cost functions. Without loss of generality, assume that

$$\sup_{y,y' \in \mathcal{Y}} C(y, y') \leq 1.$$

With regards to the classification tasks, we assume that there exists some prior density ν_0 over $\mathcal{Y}$, and that the label subsets $\mathcal{S}_i = \{y^{(1)}, \ldots, y^{(k_i)}\}$ are obtained by iid samples with replacement from the density ν_0. (An alternative assumption would be that $\mathcal{S}_1 \subset \mathcal{S}_2$ with $\mathcal{S}_1$ being a subsample of $\mathcal{S}_2$: this assumption can also be addressed, as we wil discuss later.)

Next, suppose there exists some other density ν_1 over $\mathcal{Y}$, and that the prior probabilities for each classification task are given by

$$\pi_i(y) = \frac{\nu_1(y)}{\sum_{y' \in \mathcal{S}_i} \nu_1(y')}.$$

Define π_0 as the distribution over $\mathcal{Y}$ with density proportional to $\nu_0 \nu_1$. Then the marginal distribution of any element of $\mathcal{S}$ is given by π_0.

Further, let us assume that we have more repeats per class in the first classification task than in the second, $r_1 > r_2$. We discuss the possibility of relaxing this condition in the Discussion.

Since the classification tasks are randomly generated, we will aim to develop a method for estimating the *average risk*. In the case where the classification tasks are independently generated, the average risk is the best predictor (in mean-squared error) for the (random) risk.

We make some rather strong assumptions with regards to the classifiers. The classifier $\mathcal{F}$ produces classification rules f which depend on *marginal scoring rules*, m_y for $y \in \mathcal{S}$. Each marginal scoring rule m_i is a mapping

$$m_y : \mathcal{X} \to \mathbb{R}.$$

The classification rule chooses the class with the highest marginal score,

$$f(x) = \text{argmax}_{y \in \mathcal{S}} m_y(x).$$

The marginal scoring rules m_i, in turn, are generated by a marginal model $\mathcal{M}$. The marginal model converts empirical distributions $\hat{F}$ over $\mathcal{X}$, and an (empirical) prior class probability, into a marginal scoring function $m : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$. For example, one could take

$$m(x, p) = \log(p) + \log(\hat{f}(x)).$$

where $\hat{f}$ is a density estimate obtained from $\hat{F}$. We call such a classification model $\mathcal{F}$ a *marginal classifier*, and such marginal classifiers are completely specified by the marginal model $\mathcal{M}$.

Quadratic discriminant analysis and Naive Bayes are two examples of marginal classification models. The *marginal* property allows us to prove strong results about the accuracy of the classifier under the exchangeable sampling assumption, as we see in Section [].

2.3 Local polynomial regression

Explain background.

Introduce the notation $\{(w_i, x_i, y_i)\}_{i=1}^n$: ordered triples of weight, predictor and response.

2.4 Measurement error models

Explain background.

3 Performance extrapolation for marginal classification models

Having outlined our assumption for randomized label subsets, the focus of our theory moves towards understanding the k-class average risk: that is, the expected risk of $\mathcal{F}$ when a random subset $\mathcal{S}$ of size k is drawn.

We obtain a method for estimating the risk in the second classification task using data from the first. The insight behind our estimation method is obtained via an analysis of the average risk of the classification task.

3.1 Easy special cases

Let us first mention two easy special cases, which can be handled using existing machine learning methodology.

In the special case where $k_1 = k_2 = k$: that is, where the label subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ are the same size, it is clear to see that any unbiased estimate of the risk of the classifier $\mathcal{F}$ for the first classification problem is an unbiased estimate of the average k-class risk. Since various methods, such as cross-validation can be used to obtain close-to-unbiased estimates of the risk in a given classification problem, the problem is essentially solved for this special case.

Meanwhile, in the case where $k_2 < k_1$, the problem can be solved by repeatedly subsampling label sets of size k_2 from $\mathcal{S}_1$ and averaging unbiased estimates of the risk of each subsampled classification task. Aside from computational issues with respect to computing or approximating the average of $\binom{k_1}{k_2}$ empirical accuracies, the problem is again more or less solved by using existing methods.

Therefore, the challenging case is when $k_2 > k_1$: we want to predict the performance of the classification model in a setting with more labels than we currently see in the training set.

3.2 Analysis of the average risk

The average risk is obtained by averaging over four randomizations:

1. Drawing the label subset $\mathcal{S}$.
2. Drawing the training dataset.
3. Drawing Y^* from $\mathcal{S}$ according to π.
4. Drawing X^* from F_{X^*}.

In other words, one can define a random variable L

$$L = C(\mathcal{F}(\{\hat{F}_y\}_{y \in \mathcal{S}}, \pi)(X^*), Y^*)$$

which clearly depends on all of the random quantities $\mathcal{S}, \hat{F}_y, \pi, Y^*, X^*$, and where

$$\text{Average Risk}_k(\mathcal{F}) = \mathbf{E}[L]$$

where the expectation is taken over all four randomization steps.

As we pointed out in the previous section, the challenging case for the analysis is the "undersampled" regime where we wish to predict the loss on a larger label set. Given data with k_1 classes, we already have means to estimate the average risk for all $k \leq k_1$, so the challenge is to understand how the risk will "extrapolate" to $k > k_1$. Hence, the goal of the current analysis is to isolate the effect of k, the size of the label subset, on the average risk.

We find that we can do this by *conditioning on* the pair (x^*, y^*) while *averaging* over the first two steps. Define the *conditional risk* $R_k(y^*, x^*)$ as

$$R_k(y^*, x^*) = \mathbf{E}[L|Y^* = y, X^* = x].$$

Let G denote the joint distribution over (X, Y) obtained by drawing $Y \sim \pi_0$ and $X \sim F_Y$. Since Y^* has the marginal distribution π_0, it follows that

$$\text{Average Risk}_k(\mathcal{F}) = \int\int R_k(y^*, x^*) dG(x^*, y^*). \tag{1}$$

What we have done is to rewrite the average risk as the expectation of R_k, which depends on k, according to a measure G which does *not* depend on k.

However, we will further decompose R_k into k-dependent and k-independent components.

Additional technical assumptions:

- Scaling property of margins, if $\mathcal{M}(\hat{F}_1, \pi_1)(x) > \mathcal{M}(\hat{F}_2, \pi_2)(x)$ then also $\mathcal{M}(\hat{F}_1, c\pi_1)(x) > \mathcal{M}(\hat{F}_2, c\pi_2)(x)$.
- Tie-breaking condition, for all $x \in \mathcal{X}$, $\mathcal{M}(\hat{F}_1, \pi_1)(x) = \mathcal{M}(\hat{F}_2, \pi_2)(x)$ with zero probability.

Define the U-functions

$$U_x(y) = \Pr[\mathcal{F}(\hat{F}_y, \pi_0(y))(x) > \mathcal{F}(\hat{F}_Y, \pi_0(Y))(x)] \tag{2}$$

$$= \int_{\mathcal{Y}} I\{\mathcal{F}(\hat{F}_y, \pi_0(y))(x) > \mathcal{F}(\hat{F}_{y'}, \pi_0(y'))(x)\} d\Pi_{y,r}(\hat{F}_y) d\Pi_{y',r}(\hat{F}_{y'}) d\pi_0(y'). \tag{3}$$

Under the scaling property of margins, we have

$$U_x(y) = \Pr[\mathcal{F}(\hat{F}_y, \pi(y))(x) > \mathcal{F}(\hat{F}_Y, \pi(Y))(x)]$$

for the *random* π corresponding to $\mathcal{S}$. Hence, $U_x(y)$ gives the probability that an observation x would be assigned to class y, supposing that the only two choices for labels were y and Y', with Y' drawn uniformly from π_0.

Note that the random variable $U_x(Y)$ for $Y \sim \pi_0$ is uniformly distributed for all $x \in \mathcal{X}$ (hence the name "U-function").

Define the *conditional cost function* $K(y^*, x^*, u)$ by

$$K(y^*, x^*, u) = \mathbb{E}[C(Y, y^*) I\{U_{x^*}(Y) > U_{x^*}(y^*)\} | U_{x^*}(y) = u]. \tag{4}$$

The conditional cost function gives the expected cost conditional on x^*, y^*, and the U_{x^*}-value of the incorrect label with the largest margin.

Obtaining the conditional risk $R_k(y^*, x^*)$ from the conditional cost function requires the following observation. Let the $(k-1)$ incorrect labels in $\mathcal{S}$ be denoted by $y^{(1)}, \ldots, y^{(k-1)}$, and define $U_i = U_{X^*}(y^{(i)})$. Let U_{max} denote the U_{x^*}-value of the incorrect label: we have

$$U_{max} = \max_{i=1}^{k-1} U_i.$$

Meanwhile, by definition, we have

$$R_k(y^*, x^*) = \mathbf{E}[K(y^*, x^*, U_{max})].$$

But we know the density of U_{max}! Recall that U_i are iid uniform, and therefore U_{max} has density $p(u) = ku^{k-1}$. We therefore have

$$R_k(y^*, x^*) = k \int K(y^*, x^*, u) u^{k-1} du.$$

Returning to equation (1), we obtain

$$\text{Average Risk}_k(\mathcal{F}) = k \int u^{k-1} du \int K(y^*, x^*, u) dG(x^*, y^*) = k \int u^{k-1} \bar{K}(u) du,$$

where

$$\bar{K}(u) = \int K(y^*, x^*, u) dG(x^*, y^*). \tag{5}$$

The average risk is expressed as a weighted integral of a certain function $\bar{K}(u)$ defined on $u \in [0, 1]$. We have clearly isolated the part of the average risk which is independent of k–the univariate function $\bar{K}(u)$, and the part which is dependent on k–which is the weighting density ku^{k-1} (which is the Beta(k, 1) density.)

This is the key result behind our estimation method, and we restate it in the following theorem.

Theorem 3.1 *Suppose* π, $\{F_y\}_{y\in\mathcal{Y}}$ *and marginal classifier* $\mathcal{F}$ *satisfy the marginal scaling condition tie-breaking condition. Then, under the definitions* (2), (4), *and* (5), *we have*

$$\textit{Average Risk}_k(\mathcal{F}) = k\int u^{k-1}\bar{K}(u)du.$$

The proof is given in the appendix.

Having this theoretical result allows us to understand how the expected k-class risk scales with k in problems where all the relevant densities are known. However, applying this result in practice to estimate Average Risk$_k$ requires some means of estimating the unknown function $\bar{K}$–which we discuss in the following.

3.3 Estimation

Now we address the problem of estimating Average Risk$_{k_2}$ from data. Recall that the data consists of $k_1 < k_2$ classes, with r_1 repeats per class. The set of class labels is $\mathcal{S}_1 = \{y^{(1)}, \ldots, y^{(k_1)}\}$. For each class $i = 1, \ldots, k_1$, we have repeats $x_j^{(i)}$ for $j = 1, \ldots, r_1$.

Define $r_{test} = r_1 - r_2$. Let us take the first r_{test} repeats per class, and form the *test set* $\{x_j^{(i)}\}_{j=1,i=1}^{r_{test},k_1}$. From the remaining r_2 repeats, we form the empirical distributions

$$\hat{F}_{y^{(i)}} = \frac{1}{r_2}\sum_{j=r_{test}+1}^{r_1}\delta_{x_j^{(i)}}.$$

The marginal model $\mathcal{M}$ yields margins for each point in the test set for each label in $\mathcal{S}_1$. Define the margins

$$M_{i,j}^{\ell} = \mathcal{M}(\hat{F}_{y^{(\ell)}}; \pi_1(y^{(\ell)}))(x_j^{(i)}).$$

The predicted label for each test point is

$$\hat{y}_{i,j} = y^{(\text{argmax}_{\ell\in\{1,\ldots,k\}} M_{i,j}^{\ell})}.$$

Therefore, an unbiased estimate of the risk (which is also an unbiased estimate of the k_1-class average risk) is

$$\text{Test Risk} = \frac{1}{r_{test}}\sum_{i=1}^{k}\sum_{j=1}^{r_{test}}\eta_1(y^{(i)})C(\hat{y}_{i,j}, y^{(i)}).$$

Now we turn to the question of estimating the function $\bar{K}(u)$. Suppose that hypothetically, we could have observed the quantities $U_{i,j}^{\ell}$, defined

$$U_{i,j}^{\ell} = U_{x_j^{(i)}}(y^{(\ell)}).$$

Also define

$$C_i^{\ell} = C(y^{(\ell)}, y^{(i)})$$

and

$$w_i = \eta_1(y^{(i)}).$$

Then $\bar{K}(u)$ can be estimated via a local regression for the dataset $\{(w_i, U_{i,j}^{\ell}, C_i^{\ell})\}_{i=1,j=1,\ell=1}^{r_{test},k_1,k_1}$.

However, the issue is that $U_{i,j}^{\ell}$ are not directly observed, but must be inferred. Define

$$\hat{U}_{i,j}^{\ell} = \frac{\sum_{m \neq \ell} \eta_1(y^{(m)}) I\{M_{i,j}^{\ell} > M_{i,j}^{m}\}}{\sum_{m \neq \ell} \eta_1(y^{(m)})}.$$

Marginally, we have

$$(k-1)\hat{U}_{i,j}^{\ell} \sim \text{Binomial}(k-1, U_{i,j}^{\ell}).$$

Therefore, $\hat{U}_{i,j}^{\ell}$ is an unbiased estimate of $U_{i,j}^{\ell}$. However, if we are to apply local polynomial regression to the dataset

$$\{(w_i, \hat{U}_{i,j}^{\ell}, C_i^{\ell})\}_{i=1,j=1,\ell=1}^{r_{test},k_1,k_1},$$

the estimated $\widehat{\bar{K}(u)}$ will be biased, since we have *errors in the covariates*. It is necessary to make use of the *covariate adjustment* technique. Covariate adjustment is justified since the error in the covariates is conditionally independent of the response given the ture covariates:

$$\hat{U}_{i,j}^{\ell} \perp C_i^{\ell} | U_{i,j}^{\ell}.$$

Define

$$\bar{C}_m = \frac{1}{r_{test}} \sum_{j=1}^{r_{test}} \sum_{j=1,i=1,\ell=1}^{r_{test},k_1,k_1} C_i^{\ell} I\{\hat{U}_{ij}^{\ell} = \frac{m-1}{k_1 - 1}\}$$

for $m = 1, \ldots, k_1$, and let $\vec{C} = (\bar{C}_1, \ldots, \bar{C}_k)$.

We have

$$\mathbf{E}[\bar{C}_m] = \int_0^1 K(u) \frac{k_1!}{(m-1)!(k_2-m)!} u^{m-1}(1-u)^{k_1-m} du,$$

and the approximate variance upper bound

$$v_m = \mathrm{Var}[\bar{C}_m] \lessapprox \frac{\bar{C}_m(1-\bar{C})_m}{r}.$$

under the assumption that $\sup_{\mathcal{Y}^2} C(y, y') \leq 1$.

Now suppose we adopt a linear model

$$\bar{K}(u) = \sum_{\ell=1}^{d} \beta_\ell h_\ell(u)$$

where $h_\ell(u)$ are some basis functions. If we knew β_ℓ, then we could determine the average risk for any k, since defining

$$W_\ell = \int_0^1 h_\ell(u) K u^{K-1} du.$$

we have

$$\mathrm{AvRisk}_k(\mathcal{F}) = \int_0^1 \bar{K}(u) k u^{k-1} du = \sum_{\ell=1}^{d} \beta_\ell W_\ell. \tag{6}$$

Then it follows that defining

$$Z_{m,\ell} = \int_0^1 h_\ell(u) \frac{k!}{(m-1)!(k-m)!} u^{m-1}(1-u)^{k-m} du$$

for $\ell = 1, \ldots, d$, we have

$$\mathbf{E}[\bar{C}_m] = \sum_{\ell=1}^{d} \beta_\ell Z_{m,\ell}.$$

Therefore, we can apply linear regression to the dataset

$$\{(v_m, \vec{Z}_m, \bar{C}_m)\}$$

where $\vec{Z}_m = (Z_{m,1}, \ldots, Z_{m,d})$, obtaining the coefficient estimate

$$\hat{\beta} = (\boldsymbol{Z}^T V^{-1} \boldsymbol{Z})^{-1} \boldsymbol{Z}^T V^{-1} \vec{C},$$

where the matrices $\boldsymbol{Z} = (Z_{m,\ell})$ and $V = \mathrm{diag}(v_1, \ldots, v_k)$.

Now we can consider prediction of the average risk. Plugging in the estimates of β into (6), we get the unbiased risk estimate

$$\widehat{AvRisk}_k(\mathcal{F}) = \hat{\beta}^T \vec{W},$$

where $\vec{W} = (W_1, \ldots, W_d)$. The variance of the estimate is approximately bounded by

$$\mathrm{Var}[\widehat{AvRisk}_k(\mathcal{F})] \lessapprox \vec{W}^T (\boldsymbol{Z}^T V^{-1} \boldsymbol{Z})^{-1} \vec{W}.$$

Convergence properties can be considered for specific models. For instance, we examine a polynomial model in the next section.

3.4 Convergence analysis

Let us take a d-th order polynomial model

$$\bar{K}(u) = \sum_{\ell=0}^{d} \beta_\ell u^d.$$

Taking a fairly conservative analysis, we will use the universal variance bound

$$\mathrm{Var}[\bar{C}_m] \leq \frac{1}{4r}$$

which holds given the condition $\sup_{\mathcal{Y}^2} C(y, y') \leq 1$.

We have the explicit formulas

$$W_\ell = \frac{\ell!(K-\ell)!}{K!}$$

and

$$Z_{m,\ell} = \frac{(m+\ell-1)!(k-1)!}{(m-1)!(k+\ell-1)!}.$$

The variance of the unbiased estimate of average risk is bounded by

$$\mathrm{Var}[\widehat{AvRisk}_k(\mathcal{F})] \leq \frac{1}{4r} \vec{W}^T (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \vec{W}.$$