12 Aggregation

1 Setup

In aggregation problems, we are given a set of k functions

$$\{f_1, f_2, ..., f_k\}$$

and data $(X_i, Y_i) \overset{iid}{\sim} P$, $i = 1, ..., n$, and the goal is to find an aggregate estimator $\widehat{f} = \widehat{f}(x)$ of the regression function

$$\mu(x) = \mathbb{E}(Y \mid X = x)$$

that combines the functions and is close to an oracle, where we measure risk via integrated MSE

$$R(\widehat{f}) = \|\widehat{f} - \mu\|^2 = \int \left\{\widehat{f}(x) - \mu(x)\right\}^2 d\mathbb{P}(x).$$

We should usually think of $\{f_1, ..., f_k\}$ as a set of estimators built from a separate independent sample (e.g., based on parametric models, neural networks, local smoothers, series estimators, random forests, lasso, etc., which capture a wide variety of different kinds of structure). More specifically one can consider three forms of aggregation, as in Tsybakov [2003]:

1. (Linear aggregation) In linear aggregation we consider linear combinations

$$\left\{ f_\theta : f_\theta(x) = \sum_{j=1}^k \theta_j f_j(x) \ , \ \theta \in \mathbb{R}^k \right\}$$

and the goal is to find an estimator $\widehat{f}$ close to the best linear combination

$$\mathbb{E}R(\widehat{f}) \leq \inf_{\theta \in \mathbb{R}^k} R(f_\theta) + \Delta_l(n, k)$$

where $\Delta_l(n, k)$ is a remainder term that is as small as possible.

2. (Convex aggregation) In convex aggregation we consider convex combinations

$$\left\{ f_\theta : f_\theta(x) = \sum_{j=1}^k \theta_j f_j(x) \ , \ \theta_j \geq 0 \ , \ \sum_{j=1}^k \theta_j = 1 \right\}$$

and the goal is to find an estimator $\widehat{f}$ close to the best convex combination

$$\mathbb{E}R(\widehat{f}) \leq \inf_{\theta \in \Theta} R(f_\theta) + \Delta_c(n, k)$$

for the simplex $\Theta = \{\theta \in \mathbb{R}^k : \theta_j \geq 0, \sum_j \theta_j = 1\}$, and $\Delta_c(n, k)$ is a remainder term.

3. (Model selection) In model selection the goal is to find an estimator $\widehat{f}$ close to the best

$$\mathbb{E}R(\widehat{f}) \leq \min_j R(f_j) + \Delta_m(n, k)$$

for $\Delta_m(n, k)$ a remainder as small as possible.

Note the slight abuse of notation where we write $f_\theta = \sum_j \theta_j f_j(x)$ but $f_j = f_j(x) = f_{e_j}(x)$, for $e_j = (0, ..., 0, 1, 0, ..., 0)$ the vectors of the canonical basis in $\mathbb{R}^k$.

Clearly since the considered function classes are nested we have

$$\inf_{\theta \in \mathbb{R}^k} R(f_\theta) \leq \inf_{\theta \in \Theta} R(f_\theta) \leq \min_j R(f_j).$$

In fact, Tsybakov (2003) showed that the smallest possible remainder terms satisfy an inverse relation. Let $\mathcal{P}$ denote the set of distributions with bounded regression functions, so

$$\mathcal{M}_n(H) \equiv \inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P R_P(\widehat{f}) - \inf_{\theta \in H} R_P(f_\theta) \right\}$$

is the minimax rate of aggregation over a set H (where H will be $\mathbb{R}^k$ for linear aggregation, Θ for convex aggregation, and $\{1, ..., k\}$ for model selection). Then Tsybakov [2003] showed:

1. (Minimax linear aggregation)
$$\mathcal{M}_n(\mathbb{R}^k) \asymp \frac{k}{n}$$
2. (Minimax convex aggregation)
$$\mathcal{M}_n(\Theta) \asymp \begin{cases} \frac{k}{n} & \text{if } k \leq \sqrt{n} \\ \sqrt{\frac{\log(1+k/\sqrt{n})}{n}} & \text{if } k > \sqrt{n} \end{cases}$$
3. (Minimax model selection)
$$\mathcal{M}_n(\{1, ..., k\}) \asymp \frac{\log k}{n}$$

Therefore, linear aggregation has the smallest oracle risk but is statistically most difficult to achieve while, conversely, model selection has the largest oracle risk and is easiest to achieve. In what follows we mostly focus on the model selection problem.

2 Why not cross validation?

In practice it is very common to use sample-splitting/cross-validation/ERM for model selection, e.g., split the sample, use one half to fit numerous estimators $\{f_1, ..., f_k\}$, use the other half to estimate the risk for each, and pick the minimizer

$$\widehat{f} = \arg\min_{f_j} \widehat{R}(f_j) \equiv \arg\min_{f_j} \mathbb{P}_n \Big[\{Y - f_j(X)\}^2 \Big].$$

However it turns out this approach is suboptimal for model selection, as formulated above.

There are two flavors of guarantees that come with cross-validation. One is of the form

$$\mathbb{E}R(\widehat{f}) \leq (1 + \delta) \min_j R(f_j) + \frac{C}{\delta} \left(\frac{1 + \log k}{n} \right)$$

for any $\delta > 0$, which obtains optimal $\log k/n$ remainders, but at the expense of giving up on the exact oracle risk $\min_j R(f_j)$, and instead settling for some factor $(1+\delta)$ times this. Recall the way to derive this kind of result is to write, for $\ell(f) = (Y-f)^2$,

$$\begin{aligned} R(\widehat{f}) = P\{\ell(\widehat{f}) - \ell(\mu)\} &= (1+\delta)\mathbb{P}_n\{\ell(\widehat{f}) - \ell(\mu)\} + T_n \\ &\leq (1+\delta)\mathbb{P}_n\{\ell(f^*) - \ell(\mu)\} + T_n \end{aligned}$$

where

$$T_n = (1+\delta)(P - \mathbb{P}_n)\{\ell(\widehat{f}) - \ell(\mu)\} - \delta P\{\ell(\widehat{f}) - \ell(\mu)\}$$

is a decentered empirical process term. Then $\mathbb{E}\mathbb{P}_n\{\ell(f^*) - \ell(\mu)\} = \|f^* - \mu\|^2 = \min_j R(f_j)$, and tail probabilities of T_n can be appropriately controlled via a union bound and Bernstein's inequality to obtain $\mathbb{P}(T_n \geq t) \leq k\exp(-nt/C)$, yielding the desired bound on $\mathbb{E}(T_n)$ of the order $C\log k/(n\delta)$. The decentering by $\delta P\{\ell(\widehat{f}) - \ell(\mu)\}$ is crucial here to avoid usual $1/\sqrt{n}$ bounds.

The other flavor of guarantee is of the form

$$\mathbb{E}R(\widehat{f}) \leq \min_j R(f_j) + C\sqrt{\frac{1+\log 2k}{n}}$$

which sticks with the exact oracle risk but has a suboptimal $\sqrt{\log k/n}$ remainder. The way to derive this second kind of result is to write

$$\begin{aligned} R(\widehat{f}) = P\{\ell(\widehat{f}) - \ell(\mu)\} &= \mathbb{P}_n\{\ell(\widehat{f}) - \ell(\mu)\} + T_n \\ &\leq \mathbb{P}_n\{\ell(f^*) - \ell(\mu)\} + T_n \end{aligned}$$

where now

$$T_n = (P - \mathbb{P}_n)\{\ell(\widehat{f}) - \ell(\mu)\}$$

is instead a centered empirical process term. As before, $\mathbb{E}\mathbb{P}_n\{\ell(f^*) - \ell(\mu)\} = \|f^* - \mu\|^2 = \min_j R(f_j)$, and tail probabilities of T_n can be controlled via a union bound and Hoeffding's inequality or sub-Gaussian maximal inequality. This slower rate again comes from the fact that T_n is no longer uncentered, so its tail probabilities and expectation is slightly larger.

In fact Lecué [2007] and others have shown that the above result is not an artifact of the proof technique: ERM cannot be optimal, and must suffer losses greater than the optimal remainder, as it does not take advantage of the convexity of the loss. Further, it turns out that any estimator/selector restricted to be *a single element* f_j cannot be an optimal model selector, and must have a remainder term of the order $\sqrt{\log k/n}$ or larger [Rigollet and Tsybakov, 2012]. Therefore for optimal model selection we actually have to average multiple estimators f_j, instead of only picking one.

One natural idea for remedying ERM is to convexify the set $\{f_1, ..., f_k\}$ and instead minimize over its convex hull; however, Lecué and Mendelson [2009] and Dai et al. [2012] prove this fix is also suboptimal, since enlarging to the convex hull makes the set too complex. Audibert [2007] and Lecué and Mendelson [2009] show that ERM over a carefully chosen star-shaped subset of the convex hull does actually solve the problem, yielding optimal rates for model selection.

3 Exponential weights

Another way to improve upon cross-validation for model selection (rather than sticking with ERM but over a carefully chosen subset of the convex hull of $\{f_1, ..., f_k\}$) is to average the estimators using so-called exponential weights. Exponential weights are related to progressive mixture rules, which were introduced in various forms by Vovk [1990] and Catoni [2004], for example. The exponential weights aggregation is given by $\widehat{f} = f_{\widehat{\theta}} = \sum_j \widehat{\theta}_j f_j(x)$ where the weights can be motivated from penalized risk minimization via

$$\widehat{\theta} = \operatorname*{arg\,min}_{\theta \in \Theta} \left\{ \sum_j \theta_j \widehat{R}(f_j) + \frac{\beta}{n} \sum_j \theta_j \log\left(\frac{\theta_j}{\pi_j}\right) \right\}$$

for β a tuning parameter and π_j a prior distribution (e.g., $\pi_j = 1/k$).

Note that the quantity minimized above can be viewed as a penalized version of the risk $\sum_j \theta_j \widehat{R}(f_j)$, which when minimized over θ would amount to just model selection, and where the penalty $\beta/n \sum_j \theta_j \log(\theta_j/\pi_j)$ is the (β/n-weighted) KL divergence between θ_j and π_j. From the KKT conditions there is actually a closed-form solution for the exponential weights given by

$$\widehat{\theta}_j = \frac{\exp(-n\widehat{R}(f_j)/\beta)\pi_j}{\sum_\ell \exp(-n\widehat{R}(f_\ell)/\beta)\pi_\ell},$$

hence the exponential weights name.

The first guarantees proving optimality of the exponential weights procedure are from Catoni [2004] and Dalalyan and Tsybakov [2008, 2007], and show that

$$\mathbb{E}R(f_{\widehat{\theta}}) \leq \min_j R(f_j) + \frac{\beta \log k}{n}$$

for β large enough, when $\pi_j = 1/k$. Note this matches the desired $\log k/n$ optimal rate for model selection in expectation.

One might further hope a similar result holds with high probability, e.g., that say

$$R(f_{\widehat{\theta}}) \leq \min_j R(f_j) + \frac{C \log k}{n} + \frac{C' \log(1/\delta)}{n}$$

with probability at least $1 - \delta$. In fact, Audibert [2007], Dai et al. [2012], and Lecué and Mendelson [2013] have shown that this high-probability bound provably *does not hold* for exponential weights, i.e., that

$$R(f_{\widehat{\theta}}) \geq \min_j R(f_j) + \frac{C}{\sqrt{n}}$$

with some constant probability. Remedying this was a primary motivation for the next approach, called Q-aggregation.

4 Q-aggregation

Q-aggregation was developed by Rigollet [2012], Dai et al. [2012] and Lecué and Rigollet [2014] as an alternative to exponential weights and other procedures, which gives error guarantees with high probability as well as in expectation, and applies in both fixed/random design settings. The Q in Q-aggregation stands for "quadratic", as in quadratic loss; however what really matters is that the loss is strongly convex.

The Q-aggregation estimator is given by $\widehat{f} = f_{\widehat{\theta}} = \sum_j \widehat{\theta}_j f_j(x)$, where

$$\widehat{\theta} = \operatorname*{arg\,min}_{\theta \in \Theta} \widehat{Q}(\theta)$$

for the Q-risk estimated by

$$\widehat{Q}(\theta) = (1-\nu)\mathbb{P}_n\Big\{\ell(Y, f_\theta)\Big\} + \nu \sum_{j=1}^k \theta_j \mathbb{P}_n\Big\{\ell(Y, f_j)\Big\} + \frac{\beta}{n}\sum_{j=1}^k \theta_j \log(1/\pi_j)$$

where $\nu \in (0,1)$ is a weight, β is a temperature parameter, controlling how much weight to assign to a prior π_j over the functions f_j, and ℓ is some generic convex loss.

For simplicity we will focus on the special case of a quadratic loss $\ell(f) = (Y-f)^2$ with a uniform prior $\pi_j = 1/k$, yielding

$$\widehat{Q}(\theta) \equiv (1-\nu)\mathbb{P}_n\Big\{(Y-f_\theta)^2\Big\} + \nu \sum_{j=1}^k \theta_j \mathbb{P}_n\Big\{(Y-f_j)^2\Big\} = (1-\nu)\widehat{R}(f_\theta) + \nu \sum_{j=1}^k \theta_j \widehat{R}(f_j)$$

Note when the weight parameter $\nu = 0$ this is just convex ERM, since then

$$\widehat{\theta} = \operatorname*{arg\,min}_{\theta \in \Theta} \widehat{R}(f_\theta),$$

and similarly when the weight is $\nu = 1$ this is just model selection, since then

$$\min_{\theta \in \Theta} \sum_{j=1}^k \theta_j \widehat{R}(f_j) = \min_j \widehat{R}(f_j).$$

So Q-aggregation can be viewed as an interpolation between convex ERM and model selection. Following Bellec [2017], one can also view Q-aggregation through a penalization lens. Specifically, we can write

$$\widehat{Q}(\theta) = \mathbb{P}_n\Big\{(Y-f_\theta)^2\Big\} + \widehat{\mathrm{pen}}(\theta)\ , \quad \text{where } \widehat{\mathrm{pen}}(\theta) = \nu \mathbb{P}_n\left\{\sum_j \theta_j (f_j - f_\theta)^2\right\}.$$

Note this penalty is minimized when $f_\theta = f_j$ for any j, and is larger farther from $\theta = e_j$. It can also be viewed as the (averaged) variance of a random function chosen from $\{f_1, ..., f_k\}$ with probabilities θ.

It should be noted that computing $\widehat{\theta}$ is a convex optimization problem, and approximate greedy solutions were studied in Dia et al. (2012), showing similar statistical behavior to the full solution.

The Q-aggregation guarantee

Here we prove the main Q-aggregation guarantee, that

$$R(f_{\widehat{\theta}}) \le \min_j R(f_j) + \frac{C \log M}{n}$$

with high probability. In order to avoid some extra complexity in the analysis, we assume the covariate distribution $d\mathbb{P}(x)$ is known. When this distribution is unknown, the proof is logically very similar, but requires some extra steps using the contraction principle (Theorem 4.12 of Ledoux and Talagrand [2013]). Below we follow the proof sketch in Section 3.4 of Bellec [2017].

Since the covariate distribution is known, we can directly compute the mean of $\mathbb{P}_n(f_\theta^2)$, which is the $L_2(\mathbb{P})$ norm $\|f_\theta\|^2 = \int f_\theta(x)^2 \, d\mathbb{P}(x)$. Therefore we consider using the modified Q-risk given by

$$\widetilde{Q}(\theta) \equiv \frac{1}{2}\left[\Big\{\|f_\theta\|^2 - \mathbb{P}_n(2Yf_\theta)\Big\} + \sum_{j=1}^k \theta_j \Big\{\|f_j\|^2 - \mathbb{P}_n(2Yf_j)\Big\}\right] = \frac{1}{2}\left\{\widetilde{R}(f_\theta) + \sum_{j=1}^k \theta_j \widetilde{R}(f_j)\right\}$$

which replaces the estimate $\mathbb{P}_n(f_\theta^2)$ with its expectation, and again for simplicity we focus on the special case of a quadratic loss $\ell(f) = (Y-f)^2$ with a uniform prior $\pi_j = 1/k$.

Remark 1. Note in the above that we take the weight parameter to be $\nu = 1/2$. When the covariate distribution is known, this is inconsequential, but in the standard setting where the covariate distribution is unknown, the choice of weight is non-trivial and affects the guarantee.

Note we can also write this Q-risk in penalized form as

$$\widetilde{Q}(\theta) = \widetilde{R}(f_\theta) + \mathrm{pen}(\theta)/2$$

for $\mathrm{pen}(\theta) = \sum_j \theta_j \|f_\theta - f_j\|^2$ since

$$\begin{aligned}\mathrm{pen}(\theta) &= \sum_{j=1}^k \theta_j \widetilde{R}(f_j) - \widetilde{R}(f_\theta) = \sum_{j=1}^k \theta_j \Big\{\|f_j\|^2 - \mathbb{P}_n(2Yf_j)\Big\} - \Big\{\|f_\theta\|^2 - \mathbb{P}_n(2Yf_\theta)\Big\} \\ &= \sum_{j=1}^k \theta_j \|f_j\|^2 - \|f_\theta\|^2 = \sum_{j=1}^k \theta_j \int (f_j^2 - f_\theta^2)\, d\mathbb{P} = \sum_{j=1}^k \theta_j \int (f_j - f_\theta)^2 \, d\mathbb{P} = \sum_{j=1}^k \theta_j \|f_\theta - f_j\|^2.\end{aligned}$$

Now as usual we minimize the Q-risk as

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \widetilde{Q}(\theta)$$

and construct the Q-aggregation estimator as $\widehat{f} = f_{\widehat{\theta}} = \sum_j \widehat{\theta}_j f_j(x)$.

Theorem 1. *Let $\widehat{f}$ be defined as above. Assume $\|f_j\|_\infty \le B$ and $\|\mu\|_\infty \le B$. Then*

$$R(\widehat{f}) \le \min_j R(f_j) + \frac{C \log k}{n} + \frac{C \log(1/\delta)}{n}$$

with probability at least $1-\delta$ for $C = 20B/3$.

Proof. We start with a lemma, which illustrates the strong convexity of the loss and improves on the usual $\widehat{R}(\widehat{\theta}) \leq \widehat{R}(\theta)$ guarantee provided by ERM.

Lemma 1. *Let $\widetilde{Q}(\theta)$ be defined as above. Then $\widetilde{Q}(\theta) - \widetilde{Q}(\theta') = \langle \nabla \widetilde{Q}(\theta'), \theta - \theta' \rangle + \frac{1}{2}\|f_\theta - f_{\theta'}\|^2$ for any (θ, θ'), and for $\widehat{\theta} \in \arg\min_{\theta \in \Theta} \widetilde{Q}(\theta)$ we have*

$$\widetilde{Q}(\widehat{\theta}) \leq \widetilde{Q}(\theta) - \frac{1}{2}\|f_\theta - f_{\widehat{\theta}}\|^2$$

so that $\widetilde{Q}$ is strongly convex.

Remark 2. The same result holds for the standard Q-risk $\widehat{Q}$ in the unknown covariate distribution setting, if we replace the $L_2(\mathbb{P})$ norm with the empirical analog $\|f_\theta - f_{\widehat{\theta}}\|_n^2 = \mathbb{P}_n\{(f_\theta - f_{\widehat{\theta}})^2\}$.

Proof. Differentiating shows the elements of the k-vector $\nabla \widetilde{Q}(\theta)$ are given by

$$\begin{aligned} 2\frac{\partial}{\partial \theta_j}\widetilde{Q}(\theta) &= \frac{\partial}{\partial \theta_j}\left\{\int \left(\sum_j \theta_j f_j\right)^2 d\mathbb{P} - \mathbb{P}_n\left(2Y\sum_j \theta_j f_j\right)\right\} + \widetilde{R}(f_j) \\ &= \int 2 f_\theta f_j \; d\mathbb{P} - \mathbb{P}_n(2Y f_j) + \left\{\int f_j^2 \; d\mathbb{P} - \mathbb{P}_n(2Y f_j)\right\} \end{aligned}$$

so that

$$\begin{aligned} \langle \nabla \widehat{Q}(\theta'), \theta - \theta' \rangle &= \frac{1}{2}\sum_j (\theta_j - \theta_j')\left\{\int 2 f_\theta f_j \; d\mathbb{P} - \mathbb{P}_n(2Y f_j) + \widetilde{R}(f_j)\right\} \\ &= \frac{1}{2}\sum_j (\theta_j - \theta_j')\widetilde{R}(f_j) - \mathbb{P}_n\left\{Y(f_\theta - f_{\theta'}) - \int f_\theta(f_\theta - f_{\theta'}) \; d\mathbb{P}\right\}. \end{aligned}$$

Therefore, rearranging terms,

$$\begin{aligned} &\widetilde{Q}(\theta) - \widetilde{Q}(\theta') - \langle \nabla \widetilde{Q}(\theta'), \theta - \theta' \rangle \\ &\quad = \frac{1}{2}\widetilde{R}(f_\theta) + \frac{1}{2}\sum_j \theta_j \widetilde{R}(f_j) - \frac{1}{2}\widetilde{R}(f_{\theta'}) - \frac{1}{2}\sum_j \theta_j' \widetilde{R}(f_j) - \frac{1}{2}\sum_j (\theta_j - \theta_j')\widetilde{R}(f_j) \\ &\qquad + \mathbb{P}_n\left\{Y(f_\theta - f_{\theta'}) - \int f_\theta(f_\theta - f_{\theta'}) \; d\mathbb{P}\right\} = \frac{1}{2}\int (f_\theta - f_{\theta'})^2 \; d\mathbb{P} \end{aligned}$$

which shows the first result. The second result follows from the fact that $\langle \nabla \widetilde{Q}(\widehat{\theta}), \theta - \widehat{\theta} \rangle \geq 0$ for any $\theta \in \Theta$, since $\widetilde{Q}$ is convex and differentiable and $\widehat{\theta}$ is the minimizer over the simplex (see, e.g., Sec 4.2.3 of Boyd and Vandenberghe [2004]). □

Now let $J^* = \arg\min_j \mathbb{E}\{\widetilde{Q}(e_j)\} = \arg\min_j \mathbb{E}\{\widetilde{R}(f_j)\}$ so by the above we have

$$-\frac{1}{2}\|f_{J^*} - f_{\widehat{\theta}}\|^2 \geq \widetilde{Q}(\widehat{\theta}) - \widetilde{Q}(J^*) = \widetilde{R}(f_{\widehat{\theta}}) + \text{pen}(\widehat{\theta})/2 - \widetilde{R}(f_{J^*}) - \text{pen}(e_{J^*})$$

which implies

$$\widetilde{R}(f_{\widehat{\theta}}) - \widetilde{R}(f_{J^*}) \leq -\frac{1}{2}\|f_{J^*} - f_{\widehat{\theta}}\|^2 - \text{pen}(\widehat{\theta})/2 = -\frac{1}{2}\sum_{j=1}^{k} \widehat{\theta}_j \|f_j - f_{J^*}\|^2 \tag{1}$$

where we used $\text{pen}(e_{J^*}) = 0$ and rearranged terms based on the definition of the penalty.

Now note for true risk $R(g) = \|g\|^2 - 2\int gf \; d\mathbb{P}$ we have $R(g) - \widetilde{R}(g) = -2(\mathbb{P} - \mathbb{P}_n)(Yg)$, so that

$$\widetilde{R}(f_{\widehat{\theta}}) - \widetilde{R}(f_{J^*}) = R(f_{\widehat{\theta}}) - R(f_{J^*}) + 2(\mathbb{P} - \mathbb{P}_n)Y(f_{\widehat{\theta}} - f_{J^*}).$$

Remark 3. Note the above empirical process term is linear in $(f_{\widehat{\theta}} - f_{J^*})$; in the standard setting where the covariate distribution is unknown, the analogous empirical process term is instead quadratic, which is why the contraction principle is required for appropriate control.

Now for any C we have

$$\begin{aligned}
R(f_{\widehat{\theta}}) - R(f_{J^*}) - \frac{C\log k}{n} &= \widetilde{R}(f_{\widehat{\theta}}) - \widetilde{R}(f_{J^*}) - (\mathbb{P} - \mathbb{P}_n)2Y(f_{\widehat{\theta}} - f_{J^*}) - \frac{C\log k}{n} \\
&\leq -(\mathbb{P} - \mathbb{P}_n)2Y(f_{\widehat{\theta}} - f_{J^*}) - \frac{C\log k}{n} - \frac{1}{2}\sum_{j=1}^{k} \widehat{\theta}_j \|f_j - f_{J^*}\|^2 \\
&\equiv Z_n(f_{\widehat{\theta}}) \leq \max_{\theta \in \Theta} Z_n(f_\theta) = \max_j Z_n(f_j)
\end{aligned}$$

where the inequality in the second line follows from (1), and the last line since Z_n is affine in θ and an affine function over the simplex is maximized at a vertex. Now for a fixed j, by Bernstein's inequality as in Proposition 5.3 of Bellec [2017], we have

$$\mathbb{P}\left\{Z_n(f_j) > \frac{Ct}{n}\right\} = \mathbb{P}\left\{-(\mathbb{P} - \mathbb{P}_n)2Y(f_j - f_{J^*}) - \frac{C\log k}{n} - \frac{1}{2}\|f_j - f_{J^*}\|^2 > \frac{Ct}{n}\right\} \leq \frac{\exp(-t)}{k}$$

for $C = 20B/3$ as in the theorem statement. Therefore

$$\mathbb{P}\left\{R(f_{\widehat{\theta}}) - R(f_{J^*}) - \frac{C\log k}{n} > \frac{Ct}{n}\right\} \leq \mathbb{P}\left\{Z_n(f_{\widehat{\theta}}) > \frac{Ct}{n}\right\} \leq \sum_{j=1}^{k} \frac{\exp(-t)}{k} = \exp(-t).$$

where the last inequality uses a union bound. Rearranging yields

$$\mathbb{P}\left\{R(f_{\widehat{\theta}}) \leq R(f_{J^*}) + \frac{C\log k}{n} + \frac{Ct}{n}\right\} \geq 1 - \exp(-t)$$

which is the result. □

References

J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20, 2007.

P. C. Bellec. Optimal exponential bounds for aggregation of density estimators. *Bernoulli*, 23 (1):219–248, 2017.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

O. Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1):39–61, 2008.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *International Conference on Computational Learning Theory*, pages 97–111. Springer, 2007.

G. Lecué. Suboptimality of penalized empirical risk minimization in classification. *International Conference on Computational Learning Theory*, pages 142–156, 2007.

G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3):591–613, 2009.

G. Lecué and S. Mendelson. On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*, 19(2):646–675, 2013.

G. Lecué and P. Rigollet. Optimal learning with q-aggregation. *The Annals of Statistics*, pages 211–224, 2014.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

P. Rigollet. Kullback-leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, pages 639–665, 2012.

P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

A. B. Tsybakov. Optimal rates of aggregation. *Learning theory and kernel machines*, pages 303–313, 2003.

V. G. Vovk. Aggregating strategies. *Proceedings of the third annual workshop on Computational learning theory*, pages 371–386, 1990.