Asymptotic Minimaxity of False Discovery Rate Thresholding for Sparse Exponential Data

David Donoho[1] and Jiashun Jin[2]
[1]Statistics Department, Stanford University
[2]Statistics Department, Purdue University

May 19, 2004

Abstract

Control of the *False Discovery Rate* (FDR) is a recent innovation in multiple hypothesis testing, allowing the user to limit the fraction of rejected null hypotheses which correspond to false rejections (i.e. false discoveries). The FDR principle also can be used in multiparameter estimation problems to set thresholds for separating signal from noise when the signal is sparse. Success has been proven when the noise is Gaussian; see [3].

In this paper, we consider the application of FDR thresholding to a non-Gaussian setting, in hopes of learning whether the good asymptotic properties of FDR thresholding as an estimation tool hold more broadly than just at the standard Gaussian model. We consider a vector X_i, $i = 1, \ldots, n$ whose coordinates are independent exponential with individual means μ_i. The vector μ is thought to be sparse, with most coordinates 1 and a small fraction significantly larger than 1. This models a situation where most coordinates are simply 'noise', but a small fraction of the coordinates contain 'signal'.

We develop an estimation theory working with $\log(\mu_i)$ as the estimand, and use the per-coordinate mean-squared error in recovering $\log(\mu_i)$ to measure risk. We consider minimax estimation over parameter spaces defined by constraints on the per-coordinate ℓ^p norm of $\log(\mu_i)$: $Ave_i \log^p(\mu_i) \leq \eta^p$. Members of such spaces are vectors (μ_i) which are sparsely heterogeneous.

We find that, for large n and small η, FDR thresholding is nearly minimax, increasingly so as η decreases. The FDR control parameter $0 < q < 1$ plays an important role: when $q \leq \frac{1}{2}$, the FDR estimator is nearly minimax, while choosing a fixed $q > \frac{1}{2}$ prevents near minimaxity. These conclusions mirror those found by [3] in the Gaussian case.

The techniques developed here seem applicable to a wide range of other distributional assumptions, other loss measures, and non-i.i.d. dependency structures.

Keywords: Minimax Decision theory, Minimax Bayes estimation, Mixtures of exponential model, Sparsity, False Discovery Rate (FDR), Multiple Comparisons, Threshold rules.

AMS 1991 subject classifications: Primary-62H12, 62C20. Secondary-62G20, 62C10, 62C12.

Acknowledgments: The authors would like to thank Yoav Benjamini and Iain Johnstone for extensive discussions, references and encouragement. These results were reported in JJ's Ph.D Thesis; JJ thanks committee members for valuable suggestions and encouragement. This work has been partially supported by National Science Foundation grants DMS 00-77261 and DMS 95-05151.

1 Introduction

Suppose we have n measurements X_i which are exponentially distributed, with possibly different means μ_i:

$$X_i \sim \mathrm{Exp}(\mu_i), \qquad \mu_i \geq 1, \quad i = 1, \ldots, n. \tag{1.1}$$

The unknown $\mu_i's$ exhibit *sparse heterogeneity*: most take the common value 1, but a small fraction take different values > 1.

There are various ways to define sparsity precisely; see [3] for example. In our setting of exponential means, the most intuitive notion of sparsity is simply that there is a *relatively small proportion of μ_i's which are strictly larger than* 1:

$$\frac{\#\{i : \mu_i \neq 1\}}{n} \leq \epsilon \approx 0. \tag{1.2}$$

Such situations arise in several application areas.

- *Multiple Lifetime Analysis.* Suppose the X_i represent failure times of many comparable independent systems, where a small fraction of the systems – we don't know which ones – may have significantly higher expected lifetimes than the typical system.
- *Multiple Testing.* Suppose that we conduct many independent statistical hypothesis tests, each yielding a p-value p_i say, and that the vast majority of those tests correspond to cases where the null distribution is true, while a small fraction correspond to cases where a Lehmann alternative [11] is true. Then $X_i \equiv \log(1/p_i) \sim \mathrm{Exp}(\mu_i)$ where most of the μ_i are 1 – corresponding to true null hypotheses, while a few are greater than 1, corresponding to Lehmann alternatives.
- *Signal Analysis.* A common model (e.g. in spread-spectrum communications) for a discrete-time signal $(Y_t)_{t=1}^n$ takes the form $Y_t = \sum_j W_j \exp\{\sqrt{-1}\,\lambda_j t\} + Z_t$, where Z_t is a white Gaussian noise, and the λ_j index a small number of unknown frequencies with white Gaussian noise coefficients W_j. In spectral analysis of such signals it is common to compute the periodogram $I(\omega) = n^{-1/2} \sum_t Y_t \exp(\sqrt{-1}\omega t)$, and consider as primary data the periodogram ordinates $X_i \equiv I(\frac{2\pi i}{n})$, $i = 1, \ldots, n/2 - 1$. These can be modeled as independently exponentially distributed with means μ_i, say; here most of the $\mu_i = 1$, meaning that there is only noise at those frequencies, while some of the $\mu_i > 1$, meaning that there is signal at those frequencies. (That is, certain frequencies $\omega_i = \frac{2\pi i}{n}$ happen to match some λ_j). In an incoherent or noncooperative setting, we wouldn't know the λ_j and hence we wouldn't know which $\mu_i > 1$.

The simple sparsity model (1.2) is merely a first pass at the problem, in applications we may also need to consider situations with a large number of means which are close to, but not exactly 1. A more general assumption (adapted from [6, 3]) is that for some $0 < p < 2$, the log means obey an ℓ^p constraint,

$$\frac{1}{n}(\sum_{i=1}^n \log^p \mu_i) \leq \eta^p, \qquad \eta \text{ small}, \qquad 0 < p < 2.$$

Working on the log-scale turns out to be useful because of the 'multiplicative' nature of the exponential data. The parameter p measures the degree of sparsity of μ. As $p \to 0$,

$$\sum_{i=1}^n \log^p(\mu_i) \quad \longrightarrow \quad \#\{i : \mu_i \neq 1\}.$$

1.1 Minimax Estimation of Sparse Exponential Means

We now turn to simultaneous estimation of the means μ_i. Let $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$, and suppose we use the squared ℓ^2-norm on the log-scale to measure loss

$$\|\log\hat{\mu} - \log\mu\|_2^2 = \sum_{i=1}^n (\log\hat{\mu}_i - \log\mu_i)^2.$$

Motivated by situations of sparsity, we consider restricted parameter spaces – ℓ^p-balls with radius η:

$$M_{n,p}(\eta) = \{\mu : \frac{1}{n}\sum_{i=1}^{n} \log^p(\mu_i) \leq \eta^p\}.$$

We quantify performance by the expected coordinatewise loss:

$$R_n(\hat{\mu}, \mu) = \mathcal{E}\big[\frac{1}{n}\sum_{i=1}^{n}(\log\hat{\mu}_i - \log\mu_i)^2\big].$$

We are interested in the minimax risk, the optimal risk which any estimator can guarantee to hold uniformly over the parameter space:

$$R_n^* = R_n^*(M_{n,p}(\eta)) = \inf_{\hat{\mu}} \sup_{M_{n,p}(\eta)} R_n(\hat{\mu}, \mu). \tag{1.3}$$

This quantity has been studied before in a related Gaussian noise setting [3], but not, to our knowledge, in an exponential noise setting.

Theorem 1.1

$$\lim_{\eta\to 0}\left[\frac{\lim_{n\to\infty} R_n^*(M_{n,p}(\eta))}{\eta^p \log^{2-p}\log\frac{1}{\eta}}\right] = 1.$$

A simple nearly-minimax estimator can be based on simple thresholding. In detail, set $\hat{\mu}_t \equiv (\hat{\mu}_{t,i})_{i=1}^n$, where

$$\hat{\mu}_{t,i} = \begin{cases} X_i, & X_i \geq t, \\ 1, & \text{otherwise.} \end{cases} \tag{1.4}$$

For an appropriate choice of threshold t (which depends in principle on p and η, but not on n), this can be highly effective:

Theorem 1.2

$$\lim_{\eta\to 0}\inf_t\left[\lim_{n\to\infty}\frac{\sup_{M_{n,p}(\eta)} R_n(\hat{\mu}_t, \mu)}{R_n^*(M_{n,p}(\eta))}\right] = 1.$$

The minimizing threshold $t_0 = t_0(p,\eta)$ referred to in this theorem behaves as

$$t_0(p,\eta) \sim p\log(1/\eta) + p\log\log(1/\eta)\cdot(1+o(1)), \qquad \eta \to 0.$$

In order to have asymptotic minimaxity, it is important to choose the threshold adapted to the sparsity parameters (p, η).

1.2 FDR Thresholding

FDR-controlling methods were first proposed in a multiple hypothesis testing situation in [1, 2]. For the exponential model we are considering, we suppose there are n independent tests of unrelated hypotheses, $H_{0,i}$ vs $H_{1,i}$, where the test statistics X_i obey

$$\text{under } H_{0,i}: \qquad X_i \sim \text{Exp}(1), \tag{1.5}$$

$$\text{under } H_{1,i}: \qquad X_i \sim \text{Exp}(\mu_i), \qquad \mu_i > 1, \tag{1.6}$$

and it is unknown how many of the alternative hypotheses are likely to be true. Pick a number q, $0 < q < 1$, which Benjamini et al. [1, 2], called the *FDR control parameter*. If we call a 'discovery' any case where $H_{0,i}$ is rejected in favor of $H_{1,i}$, then a 'false discovery' is a situation where $H_{0,i}$ is falsely rejected. An FDR-controlling procedure controls

$$\mathcal{E}\big[\frac{\#\{\text{False Discoveries}\}}{\#\{\text{Total Discoveries}\}}\big] \;\leq\; q.$$

Simes' procedure [14] was shown by [4] to be FDR controlling, and is easy to describe. We begin by, sorting all the observations in the descending order,

$$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}.$$

Next compare the sorted values with quantiles of Exp(1); more specifically, if $E(t)$ denotes the standard exponential distribution function, and $\bar{E} = 1 - E$ the corresponding survival function, compare $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ with $(t_1, t_2, \ldots, t_n)$, where

$$t_k = \bar{E}^{-1}(q \cdot \frac{k}{n}) = -\log(q \cdot \frac{k}{n}), \qquad 1 \leq k \leq n,$$

and let $t_0 = \infty$. Finally, let $k = k_{FDR}$ be the largest index $k \geq 1$ for which $X_{(k)} \geq t_k$, with $k = 0$ if there is no such index. The FDR thresholding estimator $\hat{\mu}^{FDR}_{q,n}$ uses the (data-dependent) threshold $\hat{t}^{FDR} \equiv t_{k_{FDR}}$, and has components $(\hat{\mu}_i)_{i=1}^n$, where

$$\hat{\mu}_i = \begin{cases} X_i, & X_i \geq \hat{t}^{FDR}, \\ 1, & \text{otherwise.} \end{cases} \tag{1.7}$$

In particular, if $k_{FDR} = 0$, $\hat{\mu}_i = 1$ for all i. We think of the observations exceeding t^{FDR} as *discoveries*; the FDR property guarantees relatively few false discoveries.

An attractive property of the procedure is its simplicity and definiteness. Another attractive property is its good performance in an estimation context. Our main result in this paper:

Theorem 1.3 *1. When $0 < q \leq \frac{1}{2}$, the FDR estimator $\hat{\mu}^{FDR}_{q,n}$ is asymptotically minimax:*

$$\lim_{\eta \to 0} \left[\lim_{n \to \infty} \frac{\sup_{\mu \in M_{n,p}(\eta)} R_n(\hat{\mu}^{FDR}_{q,n}, \mu)}{R^*_n(M_{n,p}(\eta))} \right] = 1.$$

2. When $q > \frac{1}{2}$, the FDR estimator $\hat{\mu}^{FDR}_{q,n}$ is not asymptotically minimax:

$$\lim_{\eta \to 0} \left[\lim_{n \to \infty} \frac{\sup_{\mu \in M_{n,p}(\eta)} R_n(\hat{\mu}^{FDR}_{q,n}, \mu)}{R^*_n(M_{n,p}(\eta))} \right] = \frac{q}{1-q} > 1.$$

1.3 Interpretation

By controlling the FDR so *there are at least as many 'true' discoveries above threshold as 'false' ones* we get an estimator that, with increasing sparsity $\eta \to 0$, asymptotically attains the minimax risk. This is so across a wide range of measures of sparsity.

The same general conclusion was found in a model of Gaussian observations by [3]. In that setting, we have $X_i \sim N(\mu_i, 1)$ and the μ_i are mostly close to zero, so that $Ave_i|\mu_i|^p \leq \eta_n^p$. (Note that the sparsity parameter η was replaced by a sequence $\eta_n \to 0$ as $n \to \infty$ in [3]). In that setting, it was shown that FDR thresholding again gave asymptotically minimax estimators. Hence, the results in our paper show that FDR thresholding, known previously to be successful in the Gaussian case, is also successful in an interesting non-Gaussian case.

It appears to us that there may be a wide range of non-Gaussian cases where the vector of means is sparse and FDR gives nearly-minimax results. Elsewhere, Jin will report results showing that similar conclusions are possible in the case of Poisson data. In that setting we have, for large n, n Poisson observations $N_i \sim \text{Poisson}(\mu_i)$ with μ_i mostly 1, with perhaps a small fraction significantly greater than 1. In that setting as well, it seems that FDR thresholding gives near-minimax risk.

In fact, the approach developed here seems applicable to a wide range of non-Gaussian distributions and loss functions. At the same time, it seems able to cover a wide range of dependence structures as well.

1.4 Contents

The paper is organized as follows. Theorems 1.1 (on minimax risk) and 1.2 (on thresholding risk) are developed and proved in Sections 2 and 3, respectively. These sections also introduce a model in which the parameter μ is realized by i.i.d. random sampling rather than as a fixed vector; this model is very useful for computations.

Sections 4-7 develop our technical approach for analyzing FDR thresholding. This starts, in Section 4, with a definition and analysis of the so-called FDR functional, establishing various boundedness and continuity properties. The FDR functional allows us to articulate the idea that, in a Bayesian setting where both the mean vector μ and the subordinate data X are drawn i.i.d. at random, there is a 'large-sample threshold' which FDR thresholding is consistently 'estimating'. Section 5 discusses the performance of an idealized pseudo-estimator which thresholds at this large-sample threshold even in finite samples; it shows that the idealized 'estimator' achieves risk performance approaching the minimax risk. Section 6 shows that, in large samples, the risk of FDR thresholding is well-approximated by the risk of idealized FDR thresholding. Section 7 ties together the pieces by showing that the results of Sections 4-6 for the Bayesian model have close parallels in the original frequentist setting of this introduction, implying Theorem 1.3.

Section 8 ends the paper by graphically illustrating two keys points about FDR thresholding, by comparing our results to recent work of Genovese and Wasserman and of Abramovich et al., and by describing generalizations to a variety of non-Gaussian and dependent data structures.

1.5 Notation

In this paper, we let E denote the cdf of Exp(1), while, to avoid confusion, we use $\mathcal{E}$ for the expectation operator applied to random variables; we also let $\bar{E}$ denote the survival function of Exp(1), and we extend this notation to all cdf's; that is for any cdf G, we let $\bar{G} = 1 - G$ denote the survival function.

We let '$\#'$ denote the scale mixture operator, mapping any (marginal) distribution F on $[1, \infty)$ to a corresponding $G = E\#F$ on $[0, \infty)$ according to :

$$F \overset{E\#}{\longmapsto} G: \qquad G(t) = \int E(t/\mu) dF(\mu).$$

We let $\mathcal{F}$ denote the set of all eligible cdf's:

$$\mathcal{F} = \{F: \ \ P_F\{\mu \geq 1\} = 1\},$$

and $\mathcal{F}_p(\eta)$ denote the convex set of p-th moment-constrained cdf's:

$$\mathcal{F}_p(\eta) = \{F \in \mathcal{F}: \ \ \int \log^p(\mu) dF(\mu) \leq \eta^p\}, \qquad 0 < p < 2. \tag{1.8}$$

We also let $\mathcal{G}$ denote the collection of all scale mixtures of exponentials:

$$\mathcal{G} = \{G: \ \ G = E\#F, \ \ F \in \mathcal{F}\},$$

and let $\mathcal{G}_p(\eta)$ denote the subclass where the mixing distributions obey the moment condition $\mathcal{E}|\mu|^p \leq \eta^p$:

$$\mathcal{G}_p(\eta) = E\#\mathcal{F}_p(\eta) = \{G: \ G = E\#F, \ F \in \mathcal{F}_p(\eta)\}, \qquad 0 < p < 2. \tag{1.9}$$

In this paper, except where we explicitly state otherwise, the cdf's F and G are always related by scale mixing, so

$$G = E\#F.$$

(The relation $F \mapsto E\#F$ is one-to-one.) We often use G and G_n together, always implicitly assuming they are related as the theoretical and empirical CDF of the same underlying samples, so that G_n is the empirical distribution for n iid samples $X_i \sim G$, where

$$G_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i < t\}}.$$

2 Asymptotics of Minimax Risk

In this section, we prove Theorem 1.1. As usual, $R_n^*(M) = \sup_{\pi\in\Pi} \rho_n(\pi)$, where $\rho_n(\pi)$ denotes the Bayes risk $\mathcal{E}_\pi\mathcal{E}_\mu\big[\frac{1}{n}\|\log\hat{\mu}_\pi - \log\mu\|_2^2\big]$ with μ random, $\mu\sim\pi$; $\hat{\mu}_\pi$ denotes the Bayes estimator corresponding to prior π and ℓ^2 loss, and Π denotes the set of all priors supported on M. Throughout this paper, we always implicitly assume that $P_{\pi_i}\{\mu_i \geq 1\} = 1$, where π_i is the i^{th} entry of π.

As in [6], we get a simple approximation to R_n^* by considering a minimax-Bayes problem in which μ is a random vector that is only required to belong to M *on average*. Define the minimax-Bayes risk

$$\bar{R}_n^*(M_{p,n}(\eta)) = \inf_{\hat{\mu}}\sup_{\pi}\big\{\mathcal{E}_\pi\mathcal{E}_\mu\big[\frac{1}{n}\|\log\hat{\mu} - \log\mu\|_2^2\big] :\ \mathcal{E}_\pi\big[\frac{1}{n}\sum_{i=1}^n \log^p\mu_i\big] \leq \eta^p\big\}. \tag{2.1}$$

Since a degenerate prior distribution concentrated at a single point $\mu\in M_{p,n}(\eta)$ trivially satisfies the moment constraint, the minimax-Bayes risk is an upper bound for the minimax risk:

$$R_n^*(M_{n,p}(\eta)) \leq \bar{R}_n^*(M_{n,p}(\eta)). \tag{2.2}$$

In fact, for large n we have asymptotic equality; in Section 2.1 below we prove:

Theorem 2.1

$$\lim_{n\to\infty}\frac{R_n^*(M_{n,p}(\eta))}{\bar{R}_n^*(M_{n,p}(\eta))} = 1.$$

Consider a univariate decision problem with data X a scalar random variable, with μ a random scalar $\mu\sim F$ and $X|\mu\sim \mathrm{Exp}(\mu)$. The corresponding *univariate* minimax-Bayes risk is

$$\bar{\rho}(\eta) = \bar{\rho}_p(\eta) = \inf_\delta \sup_{F\in\mathcal{F}_p(\eta)} \mathcal{E}_F\mathcal{E}_\mu(\log\delta(X) - \log\mu)^2. \tag{2.3}$$

The univariate and n-variate minimax risks are closely connected; in Section 2.2 we prove:

Theorem 2.2

$$\bar{R}_n^*(M_{n,p}(\eta)) = \bar{\rho}_p(\eta).$$

The univariate minimax-Bayes risk has a simple asymptotic expression:

Theorem 2.3 *For* $0 < p < 2$,

$$\lim_{\eta\to 0}\left(\frac{\bar{\rho}_p(\eta)}{\eta^p\log^{2-p}\log\frac{1}{\eta}}\right) = 1.$$

Theorem 1.1 follows immediately by combining Theorems 2.1-2.3. □

2.1 Proof of Theorem 2.1

Because (2.2) gives half of what we need, our task is to establish an asymptotic inequality in the other direction. We use a strategy similar to [6].

Now for fixed η, choose $0 < \delta \ll \eta$, and construct the product distribution $\Pi_{\eta-\delta}^{(n)} = \Pi_{i=1}^n \pi_{\eta-\delta}^*$, where $\mu_i \overset{iid}{\sim} \pi_{\eta-\delta}^*$, $\int \log^p(\mu)d\pi^* = (\eta-\delta)^p$, $1\leq i\leq n$, and π^* is least favorable for univariate Bayes Minimax problem (2.3), so $\Pi_{\eta-\delta}^{(n)}$ is least-favorable for the n-variate Bayes Minimax problem (2.1). Let $A_n = \{\frac{1}{n}\sum_{i=1}^n \log^p\mu_i \leq \eta^p\}$, then we construct a new prior $\tilde{\Pi}_{\eta-\delta}^{(n)} = \Pi_{\eta-\delta}^{(n)}(\cdot|A_n)$. By the Law of Large Numbers (LLN),

$$P(A_n) \to 1; \tag{2.4}$$

while under $\Pi_{\eta-\delta}^{(n)}$, $\mu\in M_{n,p}(\eta)$, i.e. supp $\Pi_{\eta-\delta}^{(n)} \subset M_{n,p}(\eta)$. As the minimax risk is the supremum of Bayes risks,

$$R_n^* \geq \rho_n(\tilde{\Pi}_{\eta-\delta}^{(n)}). \tag{2.5}$$

Now for any constant $w > 1$ and with $L(\cdot,\cdot)$ the loss function

$$L(\hat{\mu},\mu) = \frac{1}{n}\sum_{i=1}^{n}(\log\hat{\mu}_i - \log\mu_i)^2,$$

define the w-truncated loss function,

$$L^{(w)}(\hat{\mu},\mu) = \frac{1}{n}\sum_{i=1}^{n}\min\{(\log\hat{\mu}_i - \log\mu_i)^2, w\}.$$

Clearly,

$$\rho_n(\tilde{\Pi}^{(n)}_{\eta-\delta}, L) \geq \rho_n(\tilde{\Pi}^{(n)}_{\eta-\delta}, L^{(w)}), \tag{2.6}$$

where $\rho_n(\pi, L)$ denotes the Bayes risk with respect to loss function L. With $\|\cdot\|_{TV}$ the variation distance, the definition of $\tilde{\Pi}^{(n)}_{\eta-\delta}$ and (2.4) give

$$\|\tilde{\Pi}^{(n)}_{\eta-\delta} - \Pi^{(n)}_{\eta-\delta}\|_{TV} \leq 1 - P(A_n) \to 0.$$

For variation distance, $\|\mathcal{E}_P f - \mathcal{E}_Q f\| \leq \|f\|_\infty \cdot \|P - Q\|_{TV}$; thus for any fixed w, the Bayes risk

$$|\rho_n(\tilde{\Pi}^{(n)}_{\eta-\delta}, L^{(w)}) - \rho_n(\Pi^{(n)}_{\eta-\delta}, L^{(w)})| \leq w \cdot (1 - P(A_n)) \to 0, \qquad n \to \infty.$$

On the other hand, for L or $L^{(w)}$, the coordinatewise separability of the loss and the independence of the coordinates give that the per-coordinate Bayes risk does not depend on the number of coordinates:

$$\rho_n(\Pi^{(n)}_{\eta-\delta}, L) = \rho_1(\pi^*_{\eta-\delta}, L), \qquad \rho_n(\Pi^{(n)}_{\eta-\delta}, L^{(w)}) = \rho_1(\pi^*_{\eta-\delta}, L^{(w)}),$$

we conclude that, for each $w > 0$,

$$\rho_n(\tilde{\Pi}^{(n)}_{\eta-\delta}, L^{(w)}) \to \rho_1(\pi^*_{\eta-\delta}, L^{(w)}), \quad n \to \infty.$$

Using monotone convergence of $L^{(w)} \to L$, as $w \to \infty$,

$$\rho_1(\pi^*_{\eta-\delta}, L^{(w)}) \to \rho_1(\pi^*_{\eta-\delta}, L) = \bar{\rho}(\eta - \delta),$$

so from (2.5)-(2.6),

$$R^*_n \geq \bar{\rho}(\eta - \delta).$$

Now $\bar{\rho}(\eta)$ is monotone and continuous as a function of η; thus, by letting $\delta \to 0$, we have:

$$R^*_n \geq \bar{\rho}(\eta) = \bar{R}^*_n.$$

□

2.2 Proof of Theorem 2.2

First, observe that by the coordinatewise-separable nature of δ^n, and the i.i.d structure of the X_i/μ_i,

$$\frac{1}{n}\mathcal{E}_\pi\mathcal{E}_\mu\|\log\delta^n - \log\mu\|_2^2 = \frac{1}{n}\sum_i\int\mathcal{E}_{\mu_i}[\log\delta(X_i) - \log\mu_i]^2\pi_i(d\mu_i) \tag{2.7}$$

$$= \frac{1}{n}\int\mathcal{E}_{\mu_1}[\log\delta(X_1) - \log\mu_1]^2(\sum_i\pi_i)(d\mu_i) \tag{2.8}$$

$$= \mathcal{E}_{F_\pi}\mathcal{E}_{\mu_1}[\log\delta(X_1) - \mu_1]^2, \tag{2.9}$$

where $F_\pi = \frac{1}{n}\sum \pi_i(d\mu_1)$ is a univariate prior. Second, observe that the moment condition on π can also be expressed in terms of F_π, since

$$\frac{1}{n}\mathcal{E}_\pi \sum \log^p \mu_i = \frac{1}{n}\sum_i \int \log^p(\mu_i)\pi_i(d\mu_i) = \int \log^p(\mu_1)F_\pi(d\mu_1), \tag{2.10}$$

thus $\mathcal{E}_{F_\pi} \log^p \mu_1 \leq \eta^p$. Theorem 2.2 derives easily from (2.7) - (2.10). Indeed, let (F^0, δ^0) be a saddlepoint for the univariate problem (2.3): that is, δ^0 is a minimax rule, F^0 is a least favorable prior distribution and δ^0 is Bayes for F^0. Let $F^{0,n}$ denote the n-fold Cartesian product measure derived from F^0: from (2.10) and (2.7), it satisfies the moment constraint for $\bar{R}_n^*(M_{n,p}(\eta))$, and

$$\frac{1}{n}\mathcal{E}_{F^{0,n}}\mathcal{E}_\mu \|\log\delta^{0,n} - \log\mu\|_2^2 = \bar{\rho}_p(\eta).$$

To establish the Theorem, it is enough to verify that $(F^{0,n}, \delta^{0,n})$ is a saddlepoint for the minimax problem $\bar{R}_n^*(M_{n,p}(\eta))$, which would follow if for every π obeying the moment constraint for $\bar{R}_n^*(M_{n,p}(\eta))$,

$$\mathcal{E}_\pi\mathcal{E}_\mu\|\log\delta^{0,n} - \log\mu\|_2^2 \leq \mathcal{E}_{F^{0,n}}\mathcal{E}_\mu\|\log\delta^{0,n} - \log\mu\|_2^2.$$

But (2.7) - (2.10) reduce this to the saddlepoint property of (F^0, δ^0) in the 1-dimensional minimax problem $\bar{\rho}_p(\eta)$. □

2.3 Proof of Theorem 2.3

The following is proved in [10, Chapter 6].

Lemma 2.1 *For functions $a = a(\eta)$ and $d = d(\eta)$ such that $\lim_{\eta\to 0} a(\eta) = 0$, $\lim_{\eta\to 0} d(\eta) = \infty$, and $\lim_{\eta\to 0}[a(\eta)/d(\eta)]^{1/(d(\eta)-1)} = 0$, then:*

$$\int_0^1 \left[(a/d) + y^{1-1/d}\right]^{-1} dy = d\cdot\left(1 + O((a/d)^{1/(d-1)})\right), \qquad \eta \to 0.$$

We now describe lower and upper bounds for $\bar{\rho}(\eta)$, both asymptotically equivalent to $\eta^p\log^{2-p}(\log(1/\eta))$ as $\eta \to 0$. First, consider a lower bound for $\bar{\rho}(\eta)$. A natural lower bound uses 2-point priors:

$$\bar{\rho}(\eta) \equiv \sup_{F\in\mathcal{F}_p(\eta)} \rho_1(F) \geq \sup_{\{(\epsilon,\mu):\, \epsilon\log^p(\mu)=\eta^p\}} \rho_1(F_{\epsilon,\mu}), \tag{2.11}$$

where $F_{\epsilon,\mu} \in \mathcal{F}_p(\eta)$ denotes 2-point mixture $(1-\epsilon)\nu_1 + \epsilon\nu_\mu$, the Bayes rule $\delta_B(X; F_{\epsilon,\mu})$ obeys

$$\log(\delta_B(X; F_{\epsilon,\mu})) = \frac{\frac{\epsilon}{\mu}e^{-X/\mu}}{(1-\epsilon)e^{-X} + \frac{\epsilon}{\mu}e^{-X/\mu}}\log\mu, \tag{2.12}$$

and the Bayes risk is

$$\rho_1(F_{\epsilon,\mu}) = (\log\mu)^2\int_0^\infty \frac{(1-\epsilon)e^{-x}\frac{\epsilon}{\mu}e^{-\frac{x}{\mu}}}{(1-\epsilon)e^{-x} + \frac{\epsilon}{\mu}e^{-\frac{x}{\mu}}}dx = \frac{\epsilon\log^2(\mu)}{\mu}\int_0^1 (\frac{\epsilon}{(1-\epsilon)\mu} + y^{1-\frac{1}{\mu}})^{-1}dy; \tag{2.13}$$

particularly, if we let $\mu^* = \mu^*(\eta) = \log(\frac{1}{\eta})/(\log\log\frac{1}{\eta})$, $\epsilon^* = \epsilon^*(\eta) = \eta^p/\log^p(\mu^*)$, applying Lemma 2.1 with $a = a(\eta) = \epsilon^*/(1-\epsilon^*)$, and $d = d(\eta) = \mu^*$:

$$\rho_1(F_{\epsilon^*(\eta),\mu^*(\eta)}) = (\eta^p\log^{2-p}\log\frac{1}{\eta})\cdot(1+o(1)),$$

and we obtain the desired lower bound:

$$\bar{\rho}(\eta) \geq \rho_1(F_{\epsilon^*(\eta),\mu^*(\eta)}) = (\eta^p\log^{2-p}\log\frac{1}{\eta})\cdot(1+o(1)). \tag{2.14}$$

We get an upper bound by considering the risk of thresholding. Define the univariate thresholding nonlinearity

$$\delta_t(x) = \begin{cases} x, & x \geq t, \\ 1, & \text{otherwise.} \end{cases} \tag{2.15}$$

Then with thresholding estimator $\delta_t(X)$ based on scalar data X obeying $X|\mu \sim \text{Exp}(\mu)$, where scalar μ is distributed according to a prior $F \in \mathcal{F}_p(\eta)$, the univariate Bayes thresholding risk is:

$$\rho_T(t, F) = \mathcal{E}(\log(\delta_t(X)) - \log(\mu))^2.$$

We are particularly interested in the specific threshold

$$t_0 = t_0(p, \eta) = p\log(\frac{1}{\eta}) + p\log\log(\frac{1}{\eta}) + \sqrt{\log\log(\frac{1}{\eta})}.$$

The worst case univariate Bayes risk for this rule is

$$\bar{\rho}_T(t_0, \eta) = \bar{\rho}(t_0, \eta; p) \equiv \sup_{F \in \mathcal{F}_p(\eta)} \rho_T(t_0, F). \tag{2.16}$$

As the minimax rule is at least as good as any specific rule,

$$\bar{\rho}(\eta) \leq \bar{\rho}_T(t_0, \eta). \tag{2.17}$$

Now in the proof of Theorem 1.2 below, we show that the thresholding risk obeys:

$$\bar{\rho}_T(t_0, \eta; p) \leq \eta^p \log^{2-p} \log\frac{1}{\eta}(1 + o(1)), \qquad \eta \to 0. \tag{2.18}$$

Combining the lower bound given by (2.14) and the upper bounds given by (2.17)-(2.18), we obtain Theorem 2.3. □

3 Asymptotic Minimaxity of Thresholding

We now prove Theorem 1.2, showing that thresholding estimates can asymptotically approach the minimax risk.

3.1 Reduction to Univariate Thresholding

In effect, we only have to prove (2.18). We first remind the reader why this establishes Theorem 1.2. Let again $\hat{\mu}_t$ denote the thresholding procedure on samples of size n. Trivially, for any t and n, the risk of thresholding at t exceeds the minimax risk:

$$\sup_{M_{n,p}(\eta)} R_n(\hat{\mu}_t, \mu) \geq R_n^*(M_{n,p}(\eta)).$$

Theorem 1.2 thus follows from an asymptotic inequality in the other direction:

$$\limsup_{\eta \to 0} \inf_t \left[\limsup_{n \to \infty} \frac{\sup_{M_{n,p}(\eta)} R_n(\hat{\mu}_t, \mu)}{R_n^*(M_{n,p}(\eta))} \right] \leq 1. \tag{3.1}$$

Take

$$t_0 = t_0(p, \eta) = p\log(1/\eta) + p\log\log(1/\eta) + \sqrt{\log\log(1/\eta)}; \tag{3.2}$$

by Theorem 2.1 and Theorem 2.2, (3.1) reduces to:

$$\limsup_{\eta \to 0} \left[\frac{\limsup_{n \to \infty} \sup_{M_{n,p}(\eta)} R_n(\hat{\mu}_{t_0}, \mu)}{\bar{\rho}(\eta)} \right] \leq 1. \tag{3.3}$$

Consider the worst Bayes risk of $\hat{\mu}_{t_0}$ with respect to any prior $\mu \sim \pi$, where π is the distribution of a random vector which is only required to belong to $M_{n,p}$ *on average*:

$$\bar{R}_n^*(\hat{\mu}_{t_0}, \eta) = \bar{R}_n^*(\hat{\mu}_{t_0}, \eta; p) = \sup\{\mathcal{E}_\pi \mathcal{E}_\mu [\frac{1}{n} \| \log \hat{\mu}_{t_0} - \log \mu \|_2^2], \text{ for } \pi : \mathcal{E}_\pi \frac{1}{n} \sum_{i=1}^n \log^p \mu_i \leq \eta^p \}. \tag{3.4}$$

Now since degenerate prior distributions concentrated at points $\mu \in M_{p,n}(\eta)$ trivially satisfy the moment constraint $\mathcal{F}_p(\eta)$, we have:

$$\sup_{M_{n,p}(\eta)} R_n(\hat{\mu}_{t_0}, \mu) \leq \bar{R}_n^*(\hat{\mu}_{t_0}, \eta). \tag{3.5}$$

Consider also the worst univariate Bayes risk (2.16) of the scalar rule $\delta_{t_0}(X)$ as in (2.15) with respect to univariate prior $F \in \mathcal{F}_p(\eta)$. As in the proof of Theorem 2.2, it is not hard to show that the minimax multivariate Bayes risk is the same as the minimax univariate Bayes risk:

$$\bar{R}_n^*(\hat{\mu}_{t_0}, \eta) = \bar{\rho}_T(t_0, \eta). \tag{3.6}$$

Hence, we now see that given (2.14), the matching upper bound (2.18) implies

$$\lim_{\eta \to 0} \frac{\bar{\rho}_T(t_0, \eta)}{\bar{\rho}(\eta)} = 1. \tag{3.7}$$

Combining (3.5) – (3.7), yields (3.3), and Theorem 1.2. We thus turn to (2.18).

The univariate Bayes risk for thresholding at t can be decomposed into a *bias proxy* and a *variance proxy*:

$$\begin{aligned} \bar{\rho}_T(t, F) &= \int (\log \mu)^2 (1 - e^{-\frac{t}{\mu}}) dF(\mu) + \int [\int_{\frac{t}{\mu}}^\infty \log^2(x) e^{-x} dx] dF(\mu), \\ &= \int b(t, \mu) dF(\mu) + \int v(t, \mu) dF(\mu), \end{aligned}$$

say. We now proceed to show that, as $\eta \to 0$,

$$\sup_{F \in \mathcal{F}_p(\eta)} \int b(t_0, \mu) dF(\mu) \leq \eta^p \log^{2-p} \log \frac{1}{\eta}; \tag{3.8}$$

and

$$\sup_{F \in \mathcal{F}_p(\eta)} \int v(t_0, \mu) dF(\mu) = o(\eta^p \log^{2-p} \log \frac{1}{\eta}), \tag{3.9}$$

together these imply (2.18).

3.2 Maximizing Linear Functionals over $\mathcal{F}_p(\eta)$

The relations (3.8) - (3.9) concern maximization of functionals over classes of cdf's of moment-constrained scale mixtures. We now approach this problem from a general viewpoint, looking ahead to maximization problems in later sections.

Consider two functions $\psi(\mu), \phi(\mu)$ defined on $C[1, \infty) \cap C^2(1, \infty)$; suppose

(a) ϕ is strictly increasing and $\phi(1) = 0$;

(b) ψ is bounded, $\psi \geq 0$, $\psi \neq 0$, and $\psi(1) = 0$;

(c) $\lim_{\mu \to \infty} [\psi(\mu)/\phi(\mu)] = 0$.

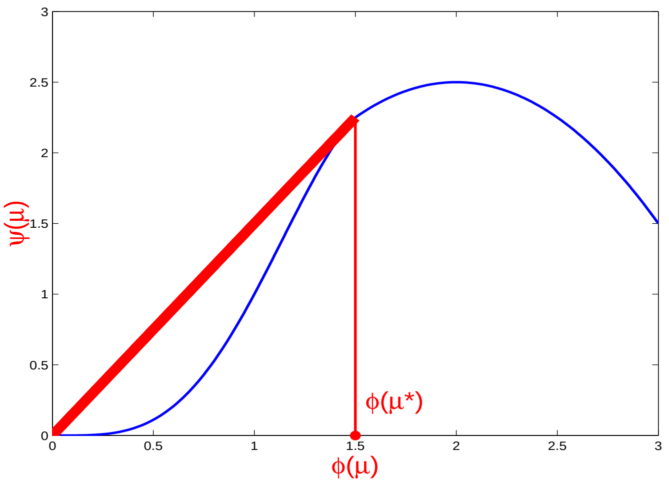

Figure 1: Illustration of $\Psi(z)$ for the case $\lim_{\mu\to 1+}[\psi(\mu)/\phi(\mu)] < \infty$ in the $\phi - \psi$ plane, the example showed here with $\lim_{\mu\to 1+}[\psi(\mu)/\phi(\mu)] = 0$, the blue curve is $\{(\phi(\mu), \psi(\mu)) : \mu \geq 1\}$. When $0 \leq z \leq \phi(\mu^*)$, $\Psi(z)$ is a linear function of z and is illustrated by the red line segment. The case $z > \phi(\mu^*)$ is not discussed.

We are interested in the maximization problem:

$$\Psi(z) = \sup_{F\in\mathcal{F}}\{\int \psi(\mu)dF(\mu) : \quad \int \phi(\mu)dF(\mu) \leq z\}. \tag{3.10}$$

In the case $\phi(\mu) = \mu$, $\Psi(z)$ is the usual convex envelope of ψ, i.e. $\Psi(z)$ traces out the least concave majorant of the graph of Ψ. The next two lemmas describe the computation of the envelope.

Lemma 3.1 *Suppose* $\lim_{\mu\to 1+}[\psi(\mu)/\phi(\mu)]$ *exists and the limit is strictly smaller than* $\Psi^* \equiv \sup_{\mu>1}\{\psi(\mu)/\phi(\mu)\}$. *Set*

$$\mu^* = \mu^*(\psi, \phi) \equiv \max\{\mu > 1 : \ \psi(\mu)/\phi(\mu) = \Psi^*\},$$

then for any $0 \leq z \leq \phi(\mu^*)$, $\Psi(z) = \Psi^* \cdot z$ *and is attained by the mixture of point masses at* 1 *and* μ^*, *with masses* $(1 - \epsilon(z))$ *and* $\epsilon(z)$ *respectively, where* $\epsilon(z) = \epsilon(z; \psi, \phi) = z/\phi(\mu^*)$.

See Figure 1.

Lemma 3.2 *If* $\lim_{\mu\to 1+}[\psi(\mu)/\phi(\mu)] = \infty$ *and there is a constant* $\bar{\mu} = \bar{\mu}(\psi, \phi) > 1$ *such that* $(\psi'(\mu)/\phi'(\mu))$ *is strictly decreasing in the interval* $(1, \bar{\mu}]$, *and that* $\psi'(\bar{\mu})/\phi'(\bar{\mu}) < \Psi^{**}(\bar{\mu})$, *where*

$$\Psi^{**}(\mu) = \Psi^{**}(\mu; \bar{\mu}, \phi, \psi) \equiv \sup_{\mu'>\bar{\mu}} \frac{\psi(\mu') - \psi(\mu)}{\phi(\mu') - \phi(\mu)}, \qquad 1 \leq \mu < \bar{\mu}, \tag{3.11}$$

then there is a unique solution $\mu_* = \mu_*(\psi, \phi)$ *to the equation*

$$\Psi^{**}(\mu) = \psi'(\mu)/\phi'(\mu), \qquad 1 < \mu \leq \bar{\mu};$$

moreover, letting

$$\mu^* = \max\{\mu \geq \bar{\mu} : \ \frac{\psi(\mu) - \psi(\mu_*)}{\phi(\mu) - \phi(\mu_*)} = \Psi^{**}(\mu_*)\},$$

then when $0 < z \leq \phi(\mu_*)$, $\Psi(z) = \psi(\phi^{-1}(z))$ *and is attained by the single point mass* ν_{μ_z} *with* $\mu_z = \phi^{-1}(z)$, *and when* $\phi(\mu_*) < z \leq \phi(\mu^*)$, $\Psi(z) = \psi(\mu_*) + \Psi^{**}(\mu_*)[z - \phi(\mu_*)]$ *and is attained by the mixture of point masses at* μ_* *and* μ^* *with masses* $(1 - \epsilon(z))$ *and* $\epsilon(z)$ *respectively, where* $\epsilon(z) = \epsilon(z; \phi, \psi) = [z - \phi(\mu_*)]/[\phi(\mu^*) - \phi(\mu_*)]$.

Notice here that the strict monotonicity of $\psi'(\mu)/\phi'(\mu)$ over $(1, \bar{\mu}]$ is equivalent to concavity of the curve $\{(\phi(\mu), \psi(\mu)) : \ 1 < \mu \leq \bar{\mu}\}$ in the $(\phi(\mu), \psi(\mu))$ plane. See Figure 2.

Lemmas 3.1 and 3.2 are proved in the Appendix.

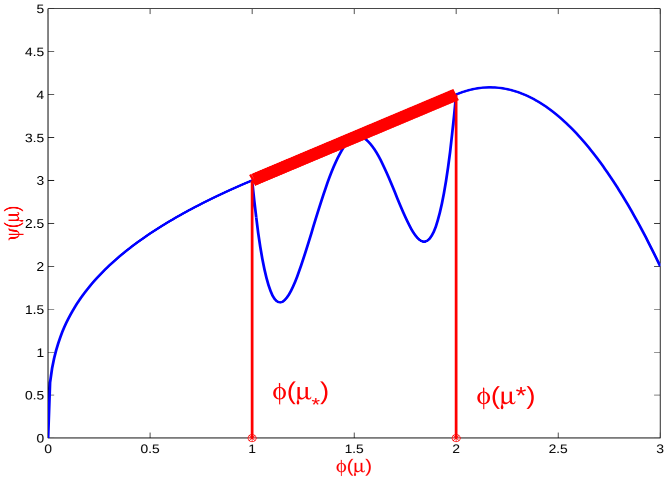

Figure 2: Illustration of $\Psi(z)$ for the case $\lim_{\mu\to 1+}[\psi(\mu)/\phi(\mu)] = \infty$ in the $\phi - \psi$ plane. The blue curve is $\{(\phi(\mu), \psi(\mu)) : \mu \geq 1\}$. When $0 < \mu < \mu_*$, $\{(\phi(\mu), \Psi(\mu)) : 0 < \mu < \mu_*\}$ traces out the same curve as that of $\{(\phi(\mu), \psi(\mu)) : 0 < \mu < \mu_*\}$, and when $\mu_* \leq \mu \leq \mu^*$, $\Psi(z)$ is a linear function of $z = \phi(\mu)$ which is illustrated by the red line segment. The slope of the line segment equals to the tangent at μ_* of the curve $\{(\phi(\mu), \psi(\mu)) : \mu \geq 1\}$. The case $z > \phi(\mu^*)$ is not discussed.

3.3 Maximizing Bias and Variance

To apply Lemma 3.1 to the bias proxy, set $\psi = \psi_\eta(\mu) = b(t_0, \mu) = \log^2(\mu)(1 - e^{-\frac{t_0}{\mu}})$, $\phi(\mu) = \log^p(\mu)$, and $\Psi(z)$ as in (3.10). Then the worst bias $\sup_{\mathcal{F}_p(\eta)} \int b(t_0, \mu) dF \equiv \Psi(\eta^p)$. Direct calculation shows that for large t_0:

$$\mu^* \equiv \mathrm{argmax}[\psi(\mu)/\phi(\mu)] \sim \frac{t_0}{\log\log t_0 - \log(2-p)},$$

and

$$\Psi^* = \bar{\Psi}_{p,\eta} \equiv \frac{\psi(\mu^*)}{\log^p(\mu^*)} \sim \log^{2-p} t_0 \sim \log^{2-p}\log(\frac{1}{\eta}).$$

It is obvious that for sufficiently small η, $\eta^p < \phi(\mu^*)$, thus by Lemma 3.1, $\Psi(\eta^p) = \Psi^* \cdot \eta^p$, and relation (3.8) follows directly.

Now consider the variance proxy. Letting $\psi(\mu) = \psi_\eta(\mu) \equiv v(t_0, \mu) - v(t_0, 1)$, $\phi(\mu) = \log^p(\mu)$, and again with $\Psi(z)$ as in (3.10), the maximal variance proxy $\sup_{\mathcal{F}_p(\eta)} \int v(t_0, \mu) dF = \Psi(\eta^p) + v(t_0, 1)$. Notice here that $v(t_0, 1) = o(\eta^2 \log^2(\log \frac{1}{\eta}))$, so to show relation (3.9), all we need to show is:

$$\Psi(\eta^p) = O(\eta^p). \tag{3.12}$$

Direct calculations show that:

$$\lim_{\mu\to 1+}[\frac{\psi(\mu)}{\phi(\mu)}] = \begin{cases} 0, & 0 < p < 1, \\ t_0 \log^2(t_0) e^{-t_0}, & p = 1, \\ \infty, & 1 < p < 2; \end{cases} \tag{3.13}$$

so we will calculate $\Psi(z)$ for the cases $0 < p \leq 1$ and $1 < p < 2$ separately.

When $0 < p \leq 1$, letting $c = \int_1^\infty \log^2(x) e^{-x} dx$, notice that for sufficiently large t_0, the condition of Lemma 3.1 is satisfied; moreover, direct calculations show that:

$$\mu^* = \mathrm{argmax}_{\mu > 1}\{\psi(\mu)/\phi(\mu)\} \sim t_0, \qquad \Psi^* = \psi(\mu^*)/\log^p(\mu^*) \sim \frac{c}{\log^p(t_0)};$$

for sufficiently small η, $\eta^p < \phi(\mu^*)$, so by Lemma 3.1, $\Psi(\eta^p) = \Psi^* \cdot \eta^p$, and (3.12) follows directly.

When $1 < p < 2$, letting $\bar{\mu}$ be the smaller solution of the equation $\frac{t_0}{\mu}\log(\mu) = (p-1)$, then for large t_0, $\bar{\mu} \sim 1 + \frac{p-1}{t_0}$; moreover, by elementary analysis, $[\psi'(\mu)/\phi'(\mu)]$ is strictly decreasing in $(1, \bar{\mu}]$ and $\psi'(\bar{\mu})/\phi'(\bar{\mu}) < \Psi^{**}(\bar{\mu})$, and the condition of Lemma 3.2 is satisfied; moreover, for large t_0,

$$\Psi^{**}(\mu) \sim \frac{c}{\log^p t_0}, \qquad \forall\, 1 < \mu \leq \bar{\mu}. \tag{3.14}$$

More elementary analysis shows that:

$$\mu^* = \mathrm{argmax}_{\mu \geq \bar{\mu}} \frac{\psi(\mu) - \psi(\mu_*)}{\phi(\mu) - \phi(\mu_*)} \sim \mathrm{argmax}_{\mu \geq \bar{\mu}} \frac{\psi(\mu)}{\phi(\mu)} \sim t_0,$$

and

$$\mu_* = \exp([ct_0 \log^{2+p} t_0 e^{-t_0}/p]^{1/(p-1)}), \qquad \phi(\mu_*) = [ct_0 \log^{2+p} t_0 e^{-t_0}/p]^{p/(p-1)};$$

it is now clear that for sufficiently small $\eta > 0$, $\phi(\mu_*) < \eta^p < \phi(\mu^*)$, thus by Lemma 3.2,

$$\Psi(\eta^p) = \psi(\mu_*) + \Psi^{**}(\mu_*)(\eta^p - \log(\mu_*)); \tag{3.15}$$

taking $\mu = \mu_*$ in (3.14) and insert it into (3.15) gives (3.12):

$$\Psi(\eta^p) = \psi(\mu_*) + \Psi^{**}(\mu_*)[\eta^p - \phi(\mu_*)] \sim \eta^p \frac{c}{\log^p t_0} = o(\eta^p).$$

4 The FDR Functional

We now come to the central idea in our analysis of FDR thresholding. We view the FDR threshold as a functional of the underlying cumulative distribution (cdf). For any fixed $0 < q < 1$, the *FDR functional* $T_q(\cdot)$ is defined as:

$$T_q(G) = \inf\{t : \bar{G}(t) \geq \frac{1}{q}\bar{E}(t)\}, \tag{4.1}$$

where G is any cdf.

Our terminology can be justified by the following observation. If G_n is the empirical distribution of $X_1, X_2, \ldots, X_n$, then $T_q(G_n)$ is effectively the same as the FDR threshold $\hat{t}^{FDR}(X_1, \ldots, X_n)$. In more detail – see Section 6.2 below – thresholding at $T_q(G_n)$ and at $\hat{t}^{FDR}(X_1, \ldots, X_n)$ always gives numerically the exact same estimate $\hat{\mu}_{q,n}$.

In this section, we expose several key properties of this functional.

4.1 Definition, Boundedness, Continuity

We first observe that $T_q(G)$ is well defined at nontrivial scale mixtures of exponentials.

Lemma 4.1 (Uniqueness) *For fixed* $0 < q < 1$ *and* $\forall G \in \mathcal{G}$, $G \neq E$, *the equation*

$$\bar{G}(t) = \frac{1}{q}\bar{E}(t) \tag{4.2}$$

has a unique solution on $[0, \infty)$ *which we call* $T_q(G)$.

Proof. Indeed, with μ a random variable ≥ 1, $\bar{G}(t) = \mathcal{E}[\bar{E}(t/\mu)]$. Hence if $\mu \neq 1$ a.s. then, for some $\mu_0 > 1$ and some $\epsilon > 0$, we have that for all $t \geq 0$, $\bar{G}(t) > \epsilon\bar{E}(t/\mu_0)$. Now $\bar{G}(0) < \bar{E}(0)/q$ while, for sufficiently large t, $\bar{E}(t)/q < \epsilon\bar{E}(t/\mu_0)$. Hence, for some $t = t_0$ on $[0, \infty)$, (4.2) holds. Now look at the slope of $\bar{G}(t)$

$$-\frac{d}{dt}\bar{G}(t) = \mathcal{E}[\bar{E}(t/\mu)/\mu] < \mathcal{E}[\bar{E}(t/\mu)] = \bar{G}(t).$$

Compare this with the slope of $\bar{E}(t)/q$. We have

$$-\frac{d}{dt}\frac{1}{q}\bar{E}(t) = \frac{1}{q}\bar{E}(t).$$

At $t = t_0$, $\frac{1}{q}\bar{E}(t) = \bar{G}(t)$, so

$$\frac{d}{dt}\big(\bar{G}(t_0) - \frac{1}{q}\bar{E}(t)\big)\,|_{t=t_0} > 0.$$

In short, at any crossing of $\bar{G} - \frac{1}{q}\bar{E}$ the slope is positive. Downcrossings being impossible, there is only one upcrossing, so the solution (4.2) is unique. □

The ideas of the proof immediately give two other important properties of T_q.

Lemma 4.2 (Quasi-Concavity) *The collection of distributions $G \in \mathcal{G}$ satisfying $T_q(G) = t$ is convex. The collection of distributions satisfying $T_q(G) \geq t$ is convex.*

Proof. The uniqueness lemma shows that the set $T_q(G) = t$ consists precisely of those cdf's G obeying $\bar{G}(t) = e^{-t}/q$; this is a linear equality constraint over the convex set $\mathcal{G}$ and defines a convex subset of $\mathcal{G}$. The set $T_q(G) \geq t$ consists precisely of those cdf's G obeying $\bar{G}(t) \leq e^{-t}/q$; this is a linear inequality constraint over the convex set $\mathcal{G}$ and generates a convex subset. □

We also immediately have

Lemma 4.3 (Stochastic Ordering) *Say that the cdf $G_0 \lesssim G_1$ if $\bar{G}_1(t) \geq \bar{G}_0(t)$ $\forall t > 0$. Then*

$$G_0 \lesssim G_1 \Longrightarrow T_q(G_0) \geq T_q(G_1).$$

We now turn to boundedness and continuity of T_q. Recall the Kolmogorov-Smirnov distance between cdf's G, G' is defined by

$$\|G - G'\| = \sup_t |G(t) - G'(t)|.$$

Viewing the collection of cdf's as a convex set in a Banach space equipped with this metric, the FDR functional $T_q(\cdot)$ is in fact locally bounded over neighbourhoods of nontrivial scale mixture of exponentials.

Lemma 4.4 (Boundedness). *For $G \in \mathcal{G}$, $G \neq E$,*

$$-\log(\frac{q}{1-q}\|G - E\|) \leq T_q(G) \leq \frac{1-q}{q}\frac{1}{\|G - E\|}.$$

Proof. Put for short $\tau = T_q(G)$. The left-hand inequality follows from $\bar{G}(\tau) = \bar{E}(\tau)/q$, which gives

$$\|G - E\| = \sup_t |G(t) - E(t)| \geq \bar{G}(\tau) - e^{-\tau} = \frac{1-q}{q}e^{-\tau}.$$

For the right-hand inequality, use again $\bar{G}(\tau) = \bar{E}(\tau)/q$ and convexity of e^t to get

$$\frac{1}{q} = \int e^{(1-\frac{1}{\mu})\tau}dF \geq 1 + \tau \cdot \int (1 - \frac{1}{\mu})dF.$$

At the same time, since $E \lesssim G$, $\|G - E\| = \sup_{t>0} \int [e^{-\frac{t}{\mu}} - e^{-t}]dF$; observe that as a function of t, $\int [e^{-\frac{t}{\mu}} - e^{-t}]dF$ has a unique maximum point $t = \bar{t}$ satisfying $\int \frac{1}{\mu}e^{-\frac{\bar{t}}{\mu}}dF = e^{-\bar{t}}$, so

$$\|G - E\| = \int [e^{-\frac{\bar{t}}{\mu}} - e^{-\bar{t}}]dF = \int (1 - \frac{1}{\mu})e^{-\frac{\bar{t}}{\mu}}dF \leq \int (1 - \frac{1}{\mu})dF,$$

and we have: $\tau \leq \frac{1-q}{q}\frac{1}{\|G-E\|}$. □

In fact the FDR functional is even locally Lipschitz away from $G = E$.

Lemma 4.5 (Modulus of Continuity) *Define*

$$\omega^*(\epsilon; t_0) \equiv \sup\{|T_q(G') - t_0| : \ T_q(G) = t_0, \ \|G - G'\| \leq \epsilon, \ G \in \mathcal{G}\};$$

then, for each fixed t_0,

$$\omega^*(\epsilon; t_0) \leq \frac{q}{\log(1/q)} t_0 e^{t_0} \epsilon \cdot (1 + o(1)), \qquad \epsilon \to 0. \tag{4.3}$$

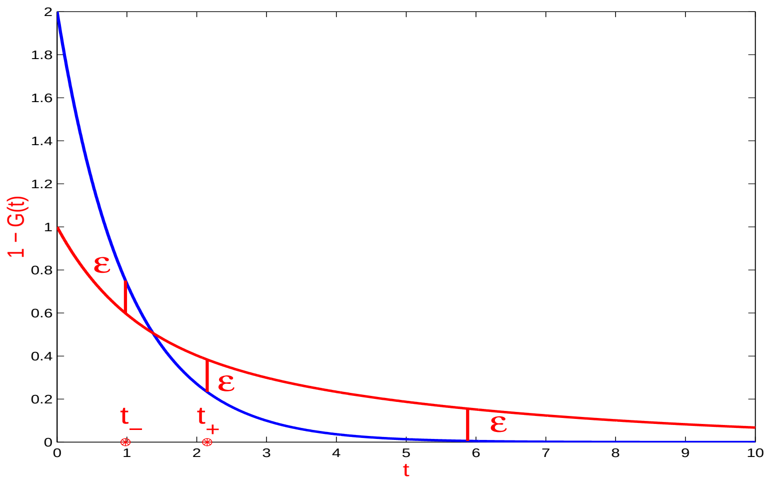

Figure 3: The blue curve is $(1/q)\bar{E}(t)$ with $q = 1/2$, and the red curve is $\bar{G}^*_{t_0}(t)$. In the plot, t_- is the solution of $\bar{G}^*_{t_0}(t) + \epsilon = (1/q)\bar{E}(t)$, and t_+ is the smallest solution to the equation of $\bar{G}^*_{t_0}(t) - \epsilon = (1/q)\bar{E}(t)$. For any other G with $T_q(G) = t_0$, $\bar{G}(t)$ is bounded above by $\bar{G}^*_{t_0}(t)$ when $0 < t < t_0$, and is bounded below by $\bar{G}^*_{t_0}(t)$ when $t > t_0$; moreover, for any G' with $\|G' - G\| \le \epsilon$, $t_- \le T_q(G') \le t_+$.

Crucially, the estimate (4.3) is uniform over $\{G \in \mathcal{G},\ T_q(G) \le t_0\}$ for fixed $t_0 > 0$. The proof even shows that

$$\omega^*(\epsilon; t_0) \le C \cdot \epsilon \qquad \text{for } 0 < \epsilon < \epsilon_{t_0}; \tag{4.4}$$

where $C = C_{t_0,q} < \infty$ if $t_0 < \infty$; this implies the local Lipschitz property.
Proof. Consider the optimization problem of finding the cdf $G^* \in \mathcal{G}$ which (a): satisfies $T_q(G^*) = t_0$ and (b): subject to that constraint, is as 'flat' as possible at t_0:

$$\frac{\partial}{\partial t}\bar{G}^*(t)|_{t=t_0} \quad = \quad \inf\{\frac{\partial}{\partial t}\bar{G}(t)|_{t=t_0} : \ \bar{G}(t_0) = \frac{1}{q}\bar{E}(t_0),\ G \in \mathcal{G}\}.$$

The solution is $\bar{G}^*_{t_0}(t) = e^{-t/\mu^*}$ for $\mu^* = 1/(1+\log(q)/t_0)$. It satisfies a more remarkable property as well: if $T_q(G) = t_0$,

$$\bar{G}(t) \le \bar{G}^*_{t_0}(t), \ \ 0 < t < t_0, \qquad \bar{G}(t) \ge \bar{G}^*_{t_0}(t), \qquad t > t_0. \tag{4.5}$$

Indeed, letting

$$h(t) \equiv [\bar{G}(t)/\bar{G}^*_{t_0}(t)] - 1 = \int e^{(\frac{1}{\mu^*} - \frac{1}{\mu})t} dF(\mu) - 1;$$

direct calculation shows that $h(t)$ is strictly convex as long as $P_F\{\mu = \mu^*\} \ne 1$ (otherwise $h \equiv 0$), (4.5) follows by noticing that $h(0) = h(t_0) = 0$.

For sufficiently small ϵ, define t_- by

$$\bar{G}^*_{t_0}(t_-) + \epsilon = \bar{E}(t_-)/q, \tag{4.6}$$

and define t_+ be the smallest solution to the equation

$$\bar{G}^*_{t_0}(t) - \epsilon = \bar{E}(t)/q, \tag{4.7}$$

see Figure 3. Now if $\|G' - G\| \le \epsilon$, then by (4.5) and (4.7):

$$\bar{G}'(t_+) \ge \bar{G}(t_+) - \epsilon \ge \bar{G}^*_{t_0}(t_+) - \epsilon = \bar{E}(t_+)/q,$$

hence $T_q(G') \le t_+$; similarly, by (4.5) and (4.6):

$$\bar{G}'(t_-) \le \bar{G}(t_-) + \epsilon \le \bar{G}^*_{t_0}(t_-) + \epsilon = \bar{E}(t_-)/q; \tag{4.8}$$

observe that the function $(\bar{G}^*_{t_0}(t) - \bar{E}(t)/q)$ is strictly decreasing in the interval $[0, t_0]$, (4.8) can be strengthened into:

$$\bar{G}'(t) \leq \bar{G}(t) + \epsilon \leq \bar{G}^*_{t_0}(t) + \epsilon < \bar{E}(t)/q, \qquad 0 < t < t_-,$$

hence $T_q(G') \geq t_-$. It follows that

$$\omega(\epsilon; t_0) \leq \max\{t_0 - t_-(\epsilon), \;\; t_+(\epsilon) - t_0\}. \tag{4.9}$$

Last, setting $w = t_+ - t_0$, (4.6) can be rewritten as

$$e^{-w/\mu^*} - e^{-w} = \epsilon q e^{t_0};$$

letting $w(\delta)$ denote the smaller one of the two solutions to

$$e^{-w/\mu^*} - e^{-w} = \delta, \qquad \delta > 0,$$

elementary analysis shows that for small $\delta > 0$, $w(\delta) \sim \delta/(1 - 1/\mu^*) = \delta t_0/\log(1/q)$, so

$$t_+ - t_0 \quad \sim \quad (q/\log(1/q)) \cdot t_0 e^{t_0} \epsilon, \qquad \epsilon \to 0;$$

similarly,

$$t_0 - t_-(\epsilon) \quad \sim \quad (q/\log(1/q)) \cdot t_0 e^{t_0} \epsilon, \qquad \epsilon \to 0.$$

Inserting these into (4.9) gives the Lemma. □

4.2 Behavior under the Bayesian Model

The continuity of T_q established in Lemma 4.5, and the role of minimax Bayes risk in solving for the Minimax risk in Sections 2 and 3 immediately suggests a fruitful change of viewpoint. Instead of viewing the $X_i \sim \text{Exp}(\mu_i)$ with fixed constants μ_i, $i = 1, \dots, n$, we instead view the μ_i as themselves sampled i.i.d. from a distribution F, and so the X_i are sampled i.i.d. from a mixture of exponentials $G = E\#F$. Starting now, and continuing through Sections 5 and 6, we adopt this viewpoint exclusively. Moreover, for our sparsity constraint, instead of assuming that $Ave_i \log^p(\mu_i) \leq \eta^p$, we assume that this happens *in expectation*, so that F obeys $\mathcal{E}_F \log(\mu_1)^p \leq \eta^p$. We call this viewpoint the *Bayesian model* because now the estimands are random. Although it seems a digression from our original purposes, it is interesting in its own right, and will be connected back to the original model in Section 7.

The motivation for this model is of course the ease of analysis. We get immediately the asymptotic consistency of FDR thresholding:

Corollary 4.1 *For $G \in \mathcal{G}$ and $G \neq E$, the empirical FDR threshold $T_q(G_n)$ converges to $T_q(G)$:*

$$\lim_{n\to\infty} T_q(G_n) = T_q(G) \qquad a.s.$$

In a natural sense, the FDR functional $T_q(G)$ can be considered as the *ideal FDR* threshold: the threshold that FDR is 'trying" to estimate and use.
Proof. The 'Fundamental Theorem of Statistics', e.g. [13, Page 1], tells us that if G_n is the empirical cdf of $X_1, X_2, \dots, X_n$ *i.i.d.* G, then

$$\|G_n - G\| \to 0 \qquad \text{a.s.} \tag{4.10}$$

Simply combine this with continuity of $T_q(G)$ at $G \neq E$. □

Of course, we can sharpen our conclusions to rates. Under i.i.d. sampling $X_i \sim G$, $\|G_n - G\| = O_P(n^{-1/2})$. Matching this, we have a root-n rate of convergence for the FDR functional.

Corollary 4.2 *If $G \in \mathcal{G}$ and $G \neq E$,*

$$|T_q(G_n) - T_q(G)| = O_P(n^{-1/2}),$$

where the $O_P()$ is locally uniform in G.

Proof. Indeed,

$$|T_q(G_n) - T_q(G)| \leq \omega^*(\|G_n - G\|; T_q(G)) = \omega^*(O_P(n^{-1/2}); T_q(G)).$$

By (4.4), for small $\epsilon > 0$, $\omega^*(\epsilon; T_q(G)) \leq C_G \epsilon$ where C_G locally bounded where $G \neq E$. So this last term is locally uniformly $O_P(n^{-1/2})$ at each $G \in \mathcal{G}$ where $G \neq E$. □

More is of course true: by Massart's work on the DKW constant [12], we have

$$P\{\|G_n - G\| \geq s/\sqrt{n}\} \leq 2e^{-2s^2}, \qquad \forall s \geq 0, \tag{4.11}$$

which combines with estimates of ω^* to control probabilities of deviations $T_q(G_n) - T_q(G)$.

5 Ideal FDR Thresholding

Continuing now in the Bayesian model just defined, we define the *ideal FDR thresholding* pseudo-estimate $\tilde{\mu}_{q,n}$, with coordinates $(\tilde{\mu}_i)$ given by

$$\tilde{\mu}_i = \begin{cases} X_i, & X_i \geq T_q(G), \\ 1, & \text{otherwise.} \end{cases} \tag{5.1}$$

In words, we are thresholding at the large-sample limit of the FDR procedure.

Notice that $T_q(G)$ depends on the underlying cdf G, which is actually unknown in any realistic situation; $\tilde{\mu}_{q,n}$ is not a *true* estimator; it could only be applied in a setting where we had side information supplied by an *oracle*, which told us $T_q(G)$. We view $\tilde{\mu}_{q,n}$ as an *ideal procedure*, and the risk for $\tilde{\mu}_{q,n}$ as an *ideal risk*: the risk we would achieve if we could use the threshold that FDR is 'trying' to 'estimate'. Despite the gap between 'true' and 'ideal', $\tilde{\mu}_{q,n}$ plays an important role in studying the true risk for (true) FDR thresholding; in fact, asymptotically there is only a negligible difference between the ideal risk for $\tilde{\mu}_{q,n}$ and the (true) risk for the FDR thresholding estimator $\hat{\mu}_{q,n}$. Let $\tilde{\mathcal{R}}_n(T_q, G)$ denote the ideal risk for $\tilde{\mu}_{q,n}$ in the Bayesian model:

$$\tilde{\mathcal{R}}_n(T_q, G) \equiv \frac{1}{n}\mathcal{E}\big[\sum_{i=1}^n (\log(\tilde{\mu}_{q,n})_i - \log \mu_i)^2\big].$$

Arguing much as in Sections 2 and 3 above we also have in the Bayesian model an identity with univariate thresholding risk:

$$\tilde{\mathcal{R}}_n(T_q, G) = \rho_T(T_q(G), F). \tag{5.2}$$

Since this ideal risk only depends on an univariate random variable $X_1 \sim G$ and $T_q(G)$ is non-stochastic, its analysis is relatively straightforward. Also, we can now drop the subscript n from $\tilde{\mathcal{R}}_n$.

Theorem 5.1 *Fix $0 < q < 1$ and $0 < p < 2$.*

1. *Worst-Case Ideal Risk. We have*

$$\lim_{\eta \to 0}\left[\frac{\sup_{G \in \mathcal{G}_p(\eta)} \tilde{\mathcal{R}}(T_q, G)}{\eta^p \log^{2-p} \log \frac{1}{\eta}}\right] = \begin{cases} 1, & 0 < q \leq \frac{1}{2}, \\ \frac{q}{1-q}, & \frac{1}{2} < q < 1. \end{cases} \tag{5.3}$$

2. *Least-Favorable Scale Mixture. Fix $0 \leq s \leq 1$. Set*

$$\mu_b^* = \mu_b^*(\eta) = \log(\frac{1}{\eta})/\log\log(\frac{1}{\eta}), \qquad \mu_v^* = \mu_v^*(\eta) = \log(\frac{1}{\eta}) \cdot \log\log(\frac{1}{\eta}),$$

and

$$G_{\epsilon,\mu} = (1-\epsilon)E(\cdot) + \epsilon E(\cdot/\mu), \qquad \epsilon \cdot \log^p(\mu) = \eta^p;$$

define

$$\tilde{\mu} = \tilde{\mu}(\eta; q, s) = \begin{cases} \mu_b^*(\eta), & 0 < q < \frac{1}{2}, \\ \mu_v^*(\eta), & \frac{1}{2} < q < 1, \\ (1-s) \cdot \mu_b^*(\eta) + s \cdot \mu_v^*(\eta), \ 0 \leq s \leq 1, & q = \frac{1}{2}, \end{cases}$$

then $G_{\epsilon,\tilde{\mu}}$ is asymptotically least-favorable for T_q:

$$\lim_{\eta\to 0}\Big[\frac{\tilde{\mathcal{R}}(T_q, G_{\epsilon,\tilde{\mu}})}{\sup_{G\in\mathcal{G}_p(\eta)}\tilde{\mathcal{R}}(T_q, G)}\Big] = 1.$$

By Theorems 2.1-2.3, the denominator on the left-hand side of (5.3) is asymptotically equivalent to the minimax risk in the original model of Section 1. In words, the worst-case ideal risk for the i.i.d. sampling model is asymptotically equivalent to the minimax risk (1.3) as $\eta \to 0$. This of course is no accident; it is a key step towards Theorem 1.3.

5.1 Proof of Theorem 5.1

We now describe the ideas for proving Theorem 5.1, in a series of lemmas. In further subsections we prove the individual lemmas.

Since the ideal risk $\tilde{\mathcal{R}}(T_q, G)$ is, by (5.2), reducible to the univariate thresholding Bayes risk which we studied in Section 3, we know to split ideal risk $\tilde{\mathcal{R}}(T_q, G)$ into two terms, the *bias* proxy and the *variance* proxy:

$$\tilde{B}^2(T_q, G) \equiv \int b(T_q(G), \mu) dF(\mu), \qquad \tilde{V}(T_q, G) \equiv \int v(T_q(G), \mu) dF(\mu).$$

Consider $\tilde{V}(T_q, G)$. Asymptotically, as $\eta \to 0$, every eligible $F \in \mathcal{F}_p(\eta)$ puts almost all mass in the vicinity of 1, and so

$$\tilde{V}(T_q, G) \approx v(T_q(G), 1) \approx \log^2(T_q(G)) e^{-T_q(G)}. \tag{5.4}$$

We set $\tilde{v}(t) \equiv \log^2(t) e^{-t}$. The following formal approximation result is proved in [10, Chapter 6].

Lemma 5.1

$$\sup_{G\in\mathcal{G}_p(\eta)} |\tilde{V}(T_q, G) - \tilde{v}(T_q(G))| = o(\eta^p \log^{2-p} \log \frac{1}{\eta}).$$

Notice that as $T_q(G) \to \infty$, $\tilde{v}(T_q(G))$ decreases rapidly, so the key for majorizing the variance is to keep $T_q(G)$ small, motivating study of:

$$T_q^* = T_q^*(\eta; p) = \inf_{G\in\mathcal{G}_p(\eta)} T_q(G). \tag{5.5}$$

Lemma 5.2 *As $\eta \to 0$,*

$$T_q^* = T_q^*(\eta; p) = p(\log \frac{1}{\eta} + \log\log\log \frac{1}{\eta}) + \log(\frac{1-q}{q}) + o(1).$$

The proof is given in Section 5.2 below. As a direct result, we get

$$\log^2(T_q^*) e^{-T_q^*} = [\frac{q}{1-q} \eta^p \log^{2-p} \log \frac{1}{\eta}] \cdot (1 + o(1));$$

moreover, when $T_q(G)$ exceeds T_q^*, the variance proxy $\tilde{v}(T_q^*)$ drops; we obtain:

Lemma 5.3 *As $\eta \to 0$, we have:*

$$\sup_{G\in\mathcal{G}_p(\eta)} \tilde{V}(T_q, G) = [\frac{q}{1-q} \eta^p \log^{2-p} \log \frac{1}{\eta}] \cdot (1 + o(1)),$$

and

$$\sup_{G\in\mathcal{G}_p(\eta), T_q(G)\geq T_q^* + \sqrt{T_q^*}} \tilde{V}(T_q, G) = o(\eta^p \log^{2-p} \log \frac{1}{\eta}).$$

We now study the bias proxy. The key observation:

$$b(t,\mu) \approx \begin{cases} \log^2 \mu, & \mu \ll t, \\ \frac{t}{\mu} \log^2 \mu, & \mu \gg t. \end{cases} \tag{5.6}$$

To develop intuition, consider the family of 2-point mixtures:

$$\mathcal{G}_p^{2,0}(\eta) = \{G_{\epsilon,\mu} = (1-\epsilon)E(\cdot) + \epsilon E(\cdot/\mu), \epsilon \log^p \mu = \eta^p\}.$$

Now (5.6) tells us that the maximum of the bias functional over this family is obtained by taking μ as large as possible while avoiding

$$\frac{T_q(G_{\epsilon,\mu})}{\mu} \ll 1;$$

moreover, direct calculations show that:

$$\frac{T_q(G_{\epsilon,\mu})}{\mu} = \frac{\log(1 + p(\frac{1}{q} - 1)\frac{1}{\eta^p}\log(\mu))}{\mu - 1}, \tag{5.7}$$

so the value of μ causing the worst bias proxy should be close to the solution of the following equation:

$$\frac{\log(1 + p(\frac{1}{q} - 1)\frac{1}{\eta^p}\log(\mu))}{\mu - 1} = 1.$$

Elaborating this idea leads to the following result, proven in Section 5.3 below.

Lemma 5.4 *As $\eta \to 0$,*

$$\sup_{G \in \mathcal{G}_p(\eta)} \tilde{B}^2(T_q, G) = (\eta^p \log^{2-p} \log \frac{1}{\eta}) \cdot (1 + o(1)).$$

Combine the above analysis for bias and variance proxies, giving

$$1 + o(1) \le \frac{\sup_{G \in \mathcal{G}_p(\eta)} \tilde{\mathcal{R}}(T_q, G)}{\eta^p \log^{2-p} \log \frac{1}{\eta}} \le \frac{1}{1-q} + o(1), \qquad \eta \to 0.$$

Compare to the conclusion of Theorem 5.1; we have obtained the correct rate, but not yet the precise constant. To refine our analysis, observe that the worst bias and the worst variance are obtained at different values μ within the family $\mathcal{G}_p^{2,0}(\eta)$. Label the μ's causing the worst bias and the worst variance by μ_b^* and μ_v^*; then

$$\mu_b^* \sim \frac{\log \frac{1}{\eta}}{\log\log \frac{1}{\eta}}, \qquad \mu_v^* \sim \log \frac{1}{\eta} \cdot \log\log \frac{1}{\eta}, \qquad \eta \to 0.$$

Divide $\mathcal{G}_p(\eta)$ into two subsets,

$$\mathcal{G}_1 \equiv \{G \in \mathcal{G}_p(\eta), T_q(G) \ge T_q^* + \sqrt{T_q^*}\}, \qquad \mathcal{G}_2 \equiv \{G \in \mathcal{G}_p(\eta), T_q(G) < T_q^* + \sqrt{T_q^*}\},$$

and consider each separately. (Note that $G_{\mu_b^*} \in \mathcal{G}_1$, while $G_{\mu_v^*} \in \mathcal{G}_2$). Over the first subset, the variance is uniformly $O(\eta^p)$, and we immediately obtain

$$\sup_{\mathcal{G}_1} \tilde{\mathcal{R}}(T_q, G) \approx \sup_{\mathcal{G}_1} \tilde{B}^2(T_q, G) \approx \eta^p \log^{2-p} \log \frac{1}{\eta}, \qquad \eta \to 0.$$

For the second subset,

Lemma 5.5

$$\sup_{\mathcal{G}_2} \tilde{\mathcal{R}}(T_q, G) = \begin{cases} (\eta^p \log^{2-p} \log \frac{1}{\eta}) \cdot (1 + o(1)), & 0 < q \le \frac{1}{2}, \\ \frac{q}{1-q} \cdot (\eta^p \log^{2-p} \log \frac{1}{\eta}) \cdot (1 + o(1)), & \frac{1}{2} < q < 1. \end{cases}$$

Theorem 5.1 follows once Lemmas 5.2 – 5.5 are proved. □

5.2 Proof of Lemma 5.2

Consider the upper envelope of the survivor function among moment-constrained scale mixtures:

$$\bar{G}^*_t(\eta; p) = \sup\{\bar{G}(t),\ G \in \mathcal{G}_p(\eta)\}.$$

The quantity of interest is the crossing point where this envelope meets the FDR boundary:

$$T^*_q = \inf\{t :\ \bar{G}^*_t \ge \bar{E}(t)/q\}.$$

Equivalently:

$$T^*_q = \inf\{t :\ [(\bar{G}^*_t/\bar{E}(t)) - 1] \ge (1-q)/q\}; \tag{5.8}$$

letting

$$h^*(t; \eta, p) = [(\bar{G}^*_t/\bar{E}(t)) - 1],$$

the key for calculating T^*_q is to explicitly express $h^*(t)$ as a function of t, asymptotically for small η.

Calculating $h^*(t)$ is again an optimization problem of a linear functional over a class of moment-constrained cdf's, and we can apply the theory in Section 3.2. Set $\psi = \psi_t(\mu) = [e^{(1-\frac{1}{\mu})t} - 1]$ and $\phi(\mu) = \log^p(\mu)$, define $\Psi = \Psi_t$ as in (3.10), so that $h^*(t; \eta, p) = \Psi_t(\eta^p)$. Notice that

$$\lim_{\mu \to 1+} [\frac{\psi_t(\mu)}{\log^p(\mu)}] = \begin{cases} 0, & 0 < p < 1, \\ t^2, & p = 1, \\ \infty, & 1 < p < 2; \end{cases} \tag{5.9}$$

so we treat the cases $0 < p \le 1$ and $1 < p < 2$ separately.

When $0 < p \le 1$, elementary analysis shows that for large t:

$$\mu^* = \text{argmax}_{\mu \ge 1}\{\frac{e^{(1-\frac{1}{\mu})t} - 1}{\log^p(\mu)}\} \sim \frac{t}{p\log(t)}, \qquad \Psi^* = \frac{e^{(1-\frac{1}{\mu})t} - 1}{\log^p(\mu)} \sim e^t/[\log^p(t)],$$

so the condition of Lemma 3.1 is satisfied, and

$$\Psi_t(\eta^p) \sim \eta^p e^t / \log^p(t), \tag{5.10}$$

inserting (5.10) into (5.8) and solving for t gives the Lemma in case $0 < p \le 1$.

When $1 < p < 2$, direct calculations show that the function $\psi'(\mu)/\phi'(\mu)$ strictly increases in the interval $(1, \bar{\mu}]$ with $\log(\bar{\mu}) = \log(\bar{\mu}(t; p)) = (p-1)/t$; also that $[\psi'(\bar{\mu})/\phi'(\bar{\mu})] \le \Psi^{**}(\bar{\mu})$, so the condition of Lemma 3.2 is satisfied. More calculations show that, first,

$$\mu^* = \mu^*(t; p) \sim \text{argmax}_{\{\mu' \ge \bar{\mu}\}}\{\frac{\psi(\mu')}{\log^p(\mu')}\} \sim \frac{t}{p\log(t)};$$

secondly, for any $1 < \mu \le \bar{\mu}$,

$$\Psi^{**}(\mu) = \Psi^{**}(\mu; t) \equiv \max_{\{\mu' \ge \bar{\mu}\}}\{\frac{\psi(\mu') - \psi(\mu)}{\log^p(\mu') - \log^p(\mu)}\} \sim \max_{\{\mu' \ge \bar{\mu}\}}\{\frac{\psi(\mu')}{\log^p(\mu')}\} \sim \frac{e^t}{\log^p(t)};$$

and finally,

$$\log(\mu_*) = \log(\mu_*(t; p)) \sim (\frac{1}{p} t \log^p(t) e^{-t})^{1/(p-1)},$$

since $h^*(t, \eta, p) = \Psi_t(\eta^p)$. By Lemma 3.2,

$$h^*(t, \eta, p) = \begin{cases} e^{(1-e^{-\eta})t} - 1, & \eta^p \le \log^p(\mu_*), \\ e^{(1-\frac{1}{\mu_*})t} - 1 + \Psi^{**}(\mu_*)(\eta^p - \log(\mu_*)), & \log^p(\mu_*) < \eta^p \le \log^p(\mu^*); \end{cases} \tag{5.11}$$

moreover, by letting $t^* = t^*_p(\eta)$ be the solution of $\log^p(\mu_*(t, p)) = \eta^p$, then we can rewrite (5.11) as:

$$h^*(t; \eta, p) = \begin{cases} e^{(1-e^{-\eta})t} - 1, & t \le t^*, \\ e^{(1-\frac{1}{\mu_*})t} - 1 + \Psi^{**}(\mu_*)(\eta^p - \log(\mu_*)), & t \ge t^*, \end{cases} \tag{5.12}$$

here, noticing $t^* \sim (p-1)p\log(\frac{1}{\eta})$ for small η.

Inserting (5.12) into (5.8), clearly for sufficiently small η and $t \leq t^*$, $h(t;\eta,p) \sim 0$, thus T_q^* is obtained by equating

$$\frac{1-q}{q} = e^{(1-\frac{1}{\mu_*})t} - 1 + \Psi^{**}(\mu_*)(\eta^p - \log(\mu_*)) \sim \eta^p e^t/\log^p(t),$$

which gives the Lemma for the case $1 < p < 2$. □

5.3 Proof of Lemma 5.4

Lemma 5.6 *For a measurable function ψ defined on $[1,\infty)$, suppose $\psi \geq 0$, $\psi \neq 0$, and $\sup_{\mu\geq 1}\{\psi(\mu)/\mu\} < \infty$, then for $G \in \mathcal{G}$ and $0 < \tau < T_q(G)$,*

$$\int \psi(\mu)[e^{-\tau/\mu} - e^{-T_q(G)/\mu}]dF \leq (1/q) \sup_{\{\mu\geq 1\}} \{\psi(\mu)/\mu\} \cdot \tau e^{-\tau}/(1-e^{-\tau}).$$

Letting $\tau \to 0$, combining Lemma 5.6 with Fatou's Lemma, we have:

$$\int \psi(\mu)[1 - e^{-T_q(G)/\mu}]dF \leq (1/q) \sup_{\{\mu\geq 1\}} \{\psi(\mu)/\mu\}. \tag{5.13}$$

Proof. Let $k_0 = k_0(\tau;G) = \lfloor \frac{T_q(G)}{\tau} \rfloor$; since $T_q(G) > \tau$, $k_0 \geq 1$; moreover:

$$\int \psi(\mu)[e^{-\tau/\mu} - e^{-T_q(G)/\mu}]dF \leq \int \psi(\mu)[e^{-\tau/\mu} - e^{-(k_0+1)\tau/\mu}]dF \tag{5.14}$$

$$= \int \psi(\mu)(1 - e^{-\tau/\mu})[\sum_{j=1}^{k_0} e^{-j\cdot\tau/\mu}]dF. \tag{5.15}$$

Put for short $c = \max_{\mu\geq 1}\{\psi(\mu)/\mu\}$, recall that $1 - e^{-x/\mu} \leq x/\mu$ for all $x \geq 0$, so for $1 \leq j \leq k_0$:

$$\int \psi(\mu)(1-e^{-\tau/\mu})e^{-j\cdot\tau/\mu}dF \leq \tau \int (\psi(\mu)/\mu)e^{-j\cdot\tau/\mu}dF \leq \tau \cdot c \cdot \int e^{-j\cdot\tau/\mu}dF; \tag{5.16}$$

by definition of k_0 and the FDR functional,

$$\int e^{-j\cdot\tau/\mu}dF = \bar{G}(j\cdot\tau) \leq (1/q)e^{-j\cdot\tau}, \qquad 1 \leq j \leq k_0, \tag{5.17}$$

combining (5.14) - (5.17) gives:

$$\int \psi(\mu)[e^{-\tau/\mu} - e^{-T_q(G)/\mu}]dF \leq (c/q)\cdot\tau\cdot\sum_{j=1}^{k_0} e^{-j\cdot\tau} \leq (c/q)\cdot\tau\cdot e^{-\tau}/(1-e^{-\tau}). \tag{5.18}$$

□

We now prove Lemma 5.4. As in Section 3, let

$$t_0 = t_0(p,\eta) = p\log(1/\eta) + p\log\log(1/\eta) + \sqrt{\log\log(1/\eta)},$$

by the monotonicity of $b(t,\mu)$ and (3.8), for sufficiently small $\eta > 0$:

$$\sup_{G\in\mathcal{G}_p(\eta), T_q(G)\leq t_0} \tilde{B}^2(T_q,G) \leq \sup_{\mathcal{G}_p(\eta)} \int b(t_0,\mu)dF = \eta^p \log^{2-p}\log(1/\eta)(1+o(1)). \tag{5.19}$$

Moreover, for any G with $T_q(G) > t_0$, letting $\psi(\cdot) = \log^2(\cdot)$ and $\tau = t_0$ in Lemma 5.6,

$$0 \leq \tilde{B}^2(T_q,G) - \int b(t_0,\mu)dF = \int \log^2(\mu)[e^{-t_0/\mu} - e^{-T_q(G)/\mu}]dF \leq ct_0e^{-t_0}/(1-e^{-t_0}),$$

where $c = \max_{\mu\geq 1}\{\log^2(\mu)/\mu\}$, so it is clear

$$\sup_{\{G\in\mathcal{G}_p(\eta), T_q(G)>t_0\}} \tilde{B}^2(T_q,G) \leq \int b(t_0,\mu)dF + O(t_0e^{-t_0}), \tag{5.20}$$

Lemma 5.4 follows directly from (5.19) - (5.20) and by noticing $t_0e^{-t_0} = o(\eta^p\log^{2-p}\log(\frac{1}{\eta}))$.

□

5.4 Proof of Lemma 5.5

By Lemma 5.1, the difference between $\tilde{V}(T_q, G)$ and $\tilde{v}(T_q(G))$ is uniformly negligible over $\mathcal{G}_p(\eta)$, so it is sufficient to prove

$$\sup_{\mathcal{G}_2}[\tilde{B}^2(T_q, G) + \tilde{v}(T_q(G))] = \begin{cases} (\eta^p \log^{2-p} \log \frac{1}{\eta}) \cdot (1 + o(1)), & 0 < q \leq \frac{1}{2}, \\ \frac{q}{1-q} \cdot (\eta^p \log^{2-p} \log \frac{1}{\eta}) \cdot (1 + o(1)), & \frac{1}{2} < q < 1. \end{cases} \tag{5.21}$$

Let $\phi(\cdot) = \log^p(\cdot)$ and

$$\psi_t(\mu) = \log^2(\mu)(1 - e^{-\frac{t}{\mu}}) + \frac{q}{1-q} \log^2 t[e^{-\frac{t}{\mu}} - e^{-t}].$$

Put for short $\tau = T_q(G)$; by definition of the FDR functional, $\int [e^{-\frac{\tau}{\mu}} - e^{-\tau}]dF = \frac{1-q}{q} e^{-\tau}$, so

$$\int \psi_\tau(\mu) dF(\mu) \equiv \tilde{B}^2(T_q, G) + \tilde{v}(T_q(G)).$$

Define Ψ_t according to (3.10), so that

$$\sup_{\mathcal{G}_1} \int \psi_t(\mu) dF \leq \sup_{\{T_q^* \leq t \leq T_q^* + \sqrt{T_q^*}\}} \Psi_t(\eta^p).$$

Hence (5.21) follows from:

$$\sup_{\{T_q^* \leq t \leq T_q^* + \sqrt{T_q^*}\}} \Psi_t(\eta^p) = \begin{cases} \eta^p \log^{2-p} \log(\frac{1}{\eta})(1 + o(1)), & 0 < q < \frac{1}{2}, \\ \eta^p \frac{q}{1-q} \log^{2-p} \log(\frac{1}{\eta})(1 + o(1)), & 1/2 \leq q < 1. \end{cases} \tag{5.22}$$

Now for (5.22), applying again the theory of Section 3.2, notice that

$$\lim_{\mu \to 1+} [\frac{\psi_t(\mu)}{\log^p(\mu)}] = \begin{cases} 0, & 0 < p < 1, \\ te^{-t}, & p = 1, \\ \infty, & 1 < p < 2, \end{cases}$$

so we treat the cases $0 < p \leq 1$ and $1 < p < 2$ separately.

When $0 < p \leq 1$, for sufficiently large t, the condition of Lemma 3.1 is satisfied. Before we prove (5.22), we explain the key role of q.

An intuitive way to see the role of q is the following. Observe that $\psi(\mu)/\phi(\mu)$ splits into two parts, $r_1 + r_2$, where

$$r_1(\mu) \equiv \log^{2-p}(\mu)(1 - e^{-t/\mu}), \qquad r_2(\mu) \equiv \frac{q}{1-q} \log^2(t)[e^{-t/\mu} - e^{-t}]/\log^p(\mu).$$

Elementary analysis shows that

$$\mu_1^* \equiv \text{argmax}_{\{\mu > 1\}} r_1(\mu) \sim t/[\log\log(t) - \log(2-p)] \sim t/\log\log(t), \qquad r_1(\mu_1^*) \sim \log^{2-p}(t),$$

and

$$\mu_2^* \equiv \text{argmax}_{\{\mu > 1\}} r_2(\mu) \sim t\log(t)/p, \qquad r_2(\mu_2^*) = \frac{q}{1-q} \log^2(t)/\log^p(\mu_2^*) \sim \frac{q}{1-q} \log^{2-p}(t).$$

In comparison, asymptotically, $r_1(\mu_1^*) > r_2(\mu_2^*)$ when $0 < q < 1/2$ and $r_1(\mu_1^*) \leq r_2(\mu_2^*)$ otherwise; accordingly, the maximum point of $[\psi(\mu)/\phi(\mu)]$ is attained at μ_1^* and μ_2^*; in other words,

$$\mu^* \equiv \text{argmax}_{\mu \geq 1}\{\psi_t(\mu)/\phi(\mu)\} \sim \begin{cases} \mu_1^* \sim t/\log\log(t), & 0 < q < 1/2, \\ \mu_2^* \sim t\log(t)/p, & 1/2 \leq q < 1, \end{cases}$$

and

$$\Psi^* = \begin{cases} \psi(\mu_1^*)/\phi(\mu_1^*) \sim \log^{2-p}(t), & 0 < q < 1/2, \\ \psi(\mu_2^*)/\phi(\mu_2^*) \sim \frac{q}{1-q} \log^{2-p}(t), & 1/2 \leq q < 1, \end{cases} \tag{5.23}$$

this explains the role of q.

Back to (5.22), for sufficiently small η, it is clear $\eta^p \leq \phi(\mu^*)$, so by Lemma 3.1,

$$\Psi_t(\eta^p) = \Psi^* \cdot \eta^p; \tag{5.24}$$

inserting (5.23) to (5.24) gives (5.22).

When $1 < p < 2$, elementary analysis shows that for large t, the function $\psi'(\mu)/\phi'(\mu)$ strictly increases in the interval $(1, \bar{\mu}]$ with $\log(\bar{\mu}) = \log(\bar{\mu}(t;p)) = (p-1)e^{-t}$, and $[\psi'(\mu)/\phi'(\mu)]$ is strictly decreasing in $(1, \bar{\mu}]$ and $\psi'(\bar{\mu})/\phi'(\bar{\mu}) < \Psi^{**}(\bar{\mu})$, so the condition of Lemma 3.2 is satisfied; also for any $1 < \mu \leq \bar{\mu}$,

$$\Psi^{**}(\mu) \sim \begin{cases} \log^2(t), & 0 < q < 1/2, \\ \frac{q}{1-q}\log^{2-p}(t), & 1/2 \leq q < 1. \end{cases} \tag{5.25}$$

More calculations show that

$$\mu^* = \begin{cases} t/[\log\log(t) - \log(2-p)], & 0 < q < 1/2, \\ t\log(t)/p, & 1/2 \leq q < 1, \end{cases}$$

and

$$\log(\mu_*) = \begin{cases} (\frac{q}{1-q} pt \log^p(t) e^{-t})^{1/(p-1)}, & 0 < q < 1/2, \\ (pt\log^p(t)e^{-t})^{1/(p-1)}, & 1/2 \leq q < 1, \end{cases}$$

notice here we clearly have a similar phenomenon as in the case $0 < p \leq 1$, we omit for further discussion.

Now for sufficiently small η and $T_q^* \leq t \leq T_q^* + \sqrt{T_q^*}$, it is clear that $\phi(\mu_*) < \eta^p < \phi(\mu^*)$, so by Lemma 3.2:

$$\Psi_t(\eta^p) = \psi_t(\mu_*) + \Psi^{**}(\mu_*)(\eta^p - \log(\mu_*)) \sim \Psi^{**}(\mu_*)\eta^p; \tag{5.26}$$

taking $\mu = \mu_*$ in (5.25) and insert it to (5.26), gives (5.22). □

6 Asymptotic Risk Behavior for FDR Thresholding

Now we turn to $\hat{\mu}_{q,n}$, the true FDR thresholding estimator. For technical reasons, we define a threshold $\hat{T}_{q,n}$ slightly differently than $\hat{t}^{FDR}$. This difference does not affect the estimate. Thus we will have $\hat{\mu}_{q,n} \equiv \hat{\mu}_{\hat{T}_{q,n}} = (\hat{\mu}_i)$ with

$$\hat{\mu}_i = \begin{cases} X_i, & X_i \geq \hat{T}_{q,n}, \\ 1, & X_i < \hat{T}_{q,n}. \end{cases}$$

Our strategy is to show that ideal and true FDR behave similarly.

We are still in the Bayesian model, and let $\mathcal{R}_n(\hat{T}_{q,n}, G)$ denote the per-coordinate average risk for $\hat{\mu}_{q,n}$:

$$\mathcal{R}_n(\hat{T}_{q,n}, G) \equiv \frac{1}{n}\mathcal{E}\big[\sum_{i=1}^{n}(\log(\hat{\mu}_{q,n})_i - \log\mu_i)^2\big].$$

Here again the expectation is over (X_i, μ_i) pairs i.i.d. with bivariate structure $X_i|\mu_i \sim \text{Exp}(\mu_i)$.

We will show that as $n \to \infty$ the difference between the true risk $\mathcal{R}_n(\hat{T}_{q,n}, G)$ and the ideal risk $\tilde{\mathcal{R}}(T_q, G)$ is asymptotically negligible. We suppress the subscript n on $\mathcal{R}_n$; this is an abuse of notation.

Theorem 6.1

$$\lim_{n\to\infty}\Big[\sup_{G\in\mathcal{G}}\big|\mathcal{R}(\hat{T}_{q,n}, G) - \tilde{\mathcal{R}}(T_q, G)\big|\Big] = 0.$$

As a result:

$$\lim_{n\to\infty}\Big[\sup_{G\in\mathcal{G}_p(\eta)}\big|\mathcal{R}(\hat{T}_{q,n}, G) - \tilde{\mathcal{R}}(T_q, G)\big|\Big] = 0.$$

Combining Theorems 6.1 and 5.1 we have:

$$\lim_{\eta\to 0}\left[\lim_{n\to\infty}\frac{\sup_{G\in\mathcal{G}_p(\eta)}\mathcal{R}(\hat{T}_{q,n},G)}{\eta^p\log^{2-p}\log\frac{1}{\eta}}\right]=\begin{cases}1, & 0<q\le\frac{1}{2},\\ \frac{q}{1-q}, & \frac{1}{2}<q<1.\end{cases}$$

Hence, $\hat{T}_{q,n}$ asymptotically achieves the n-variate minimax Bayes risk, when $n\to\infty$ followed by $\eta\to\infty$.

6.1 Proof of Theorem 6.1

We begin by defining $\hat{T}_{q,n}$. In applying the FDR functional to the empirical distribution, it is always possible that

$$\bar{G}_n(t)<\frac{1}{q}\bar{E}(t),\qquad\text{for all } t>0, \tag{6.1}$$

in which case $T_q(G_n)=\hat{t}^{FDR}=+\infty$. Letting W_n denote the event (6.1), define:

$$\hat{T}_{q,n}=\begin{cases}T_q(G_n), & \text{over } W_n^c,\\ \log(\frac{n}{q}), & \text{over } W_n.\end{cases} \tag{6.2}$$

The following lemma, proven in Section 6.2 below, shows that this definition of threshold gives the same estimator as $T_q(G_n)$, while obeying a bound which is convenient for analysis.

Lemma 6.1 *Suppose $X_i \overset{iid}{\sim} G$, $G\in\mathcal{G}$, $G\ne E$, and $\hat{T}_{q,n}$ is defined as in (4.1), then:*

1. *The FDR estimator is equivalently realized by thresholding at $\hat{T}_{q,n}$: $\hat{\mu}^{FDR}_{q,n}=\hat{\mu}_{\hat{T}_{q,n}}$.*

2. *$\hat{T}_{q,n}\le\log(\frac{n}{q})$.*

Next we study the risk for $\hat{T}_{q,n}$. We have:

$$\begin{aligned}\mathcal{R}(\hat{T}_{q,n},G)&=\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_F\mathcal{E}_\mu\big[\log^2(\mu_i)1_{\{X_i<\hat{T}_{q,n}\}}+\log^2(\frac{X_i}{\mu_i})1_{\{X_i\ge\hat{T}_{q,n}\}}\big]\\ &=\mathcal{E}_F\mathcal{E}_\mu\big[\log^2(\mu_1)1_{\{X_1<\hat{T}_{q,n}\}}+\log^2(X_1/\mu_1)1_{\{X_1\ge\hat{T}_{q,n}\}}\big],\end{aligned}$$

and $\mathcal{R}(\hat{T}_{q,n},G)$ naturally splits into a 'bias' proxy and the 'variance' proxy:

$$\begin{aligned}B^2(\hat{T}_{q,n},G)&=\mathcal{E}_F\mathcal{E}_\mu\big[\log^2(\mu_1)1_{\{X_1<\hat{T}_{q,n}\}}\big],\\ V(\hat{T}_{q,n},G)&=\mathcal{E}_F\mathcal{E}_\mu\big[\log^2(X_1/\mu_1)1_{\{X_1\ge\hat{T}_{q,n}\}}\big].\end{aligned}$$

The comparable notions in the ideal risk case were:

$$\begin{aligned}\tilde{B}^2(T_q,G)&=\mathcal{E}_F\mathcal{E}_\mu\big[\log^2(\mu_1)1_{\{X_1<T_q(G)\}}\big],\\ \tilde{V}(T_q,G)&=\mathcal{E}_F\mathcal{E}_\mu\big[\log^2(X_1/\mu_1)1_{\{X_1\ge T_q(G)\}}\big].\end{aligned}$$

Intuitively, we expect that $\tilde{B}^2$ is 'close' to B^2 and $\tilde{V}$ is 'close' to V; our next task is to validate these expectations. Observe that

$$|B^2(\hat{T}_{q,n},G)-\tilde{B}^2(T_q,G)|\le\mathcal{E}\big[\log^2(\mu_1)|1_{\{X_1<\hat{T}_{q,n}\}}-1_{\{X_1<T_q(G)\}}|\big], \tag{6.3}$$

$$|V(\hat{T}_{q,n},G)-\tilde{V}(T_q,G)|\le\mathcal{E}\big[\log^2(X_1/\mu_1)|1_{\{X_1<\hat{T}_{q,n}\}}-1_{\{X_1<T_q(G)\}}|\big], \tag{6.4}$$

it would not be hard to validate the expectations if $|\hat{T}_{q,n}-T_q(G)|$ is negligible for large n, uniformly for $G\in\mathcal{G}$. In Section 4, Lemma 4.5 tells us that $T_q(G)$ is locally $O_P(n^{-1/2})$, or more specifically,

$$|T_q(G)-T_q(G_n)|\sim\frac{q}{\log(1/q)}T_q(G)e^{T_q(G)}\|G-G_n\|,\qquad G\ne E. \tag{6.5}$$

Unfortunately, for any fixed n, G could get arbitrary close to E and as a result $T_q(G)$ could get arbitrary large, so the relationship in (6.5) could not hold *uniformly* over $G \in \mathcal{G}$.

A closer look reveals that those G's failing (6.5) would, roughly, satisfy:

$$T_q(G)e^{T_q(G)} \;\geq\; \sqrt{n}, \qquad \text{or} \qquad T_q(G) \;\geq\; \log(n)/2;$$

notice that, when n's increases from 1 to ∞, $\{G \in \mathcal{G} : \; T_q(G) \geq \log(n)/2\}$ defines a sequence of subsets, which is strictly decreasing to $\emptyset$; motivated by this, we look for a subsequence of subsets of $\mathcal{G}$ obeying:

(a) $\mathcal{G}^{(1)} \subset \mathcal{G}^{(2)} \subset \ldots \subset \mathcal{G}^{(n)} \subset \ldots$ and $\cup_1^\infty \mathcal{G}^{(n)} = \mathcal{G}$,

(b) $\mathcal{G}^{(n)}$ approaching $\mathcal{G}$ *slowly* enough such that

$$\sup_{\mathcal{G}^{(n)}} \big[\sqrt{n} T_q(G) e^{T_q(G)})\big] = o(1), \qquad n \to \infty,$$

(c) For large n, $|\mathcal{R}(\hat{T}_{q,n}) - \tilde{\mathcal{R}}(T_q, G)|$ is uniformly negligible over $\mathcal{G} \setminus \mathcal{G}^{(n)}$.

A convenient choice is:

$$\mathcal{G}_1^{(n)} \equiv \{G \in \mathcal{G} : \; T_q(G) \leq \log(n)/8\}. \qquad n \geq 1. \tag{6.6}$$

We expect that the difference between $T_q(G_n)$ and $T_q(G)$ is uniformly negligible over $\mathcal{G}_1^{(n)}$:

$$\sup_{\mathcal{G}_1^{(n)}} |T_q(G) - T_q(G_n)| = o_p(1).$$

Lemma 6.2 *Letting A_n denote the event $\{|\hat{T}_{q,n} - T_q(G)| \leq n^{-1/4}\}$, then for sufficiently large n,*

$$\sup_{G \in \mathcal{G}_1^{(n)}} P_G\{A_n^c\} \;\leq\; 3e^{-[32(1-q)^2/q^2] n^{1/4}/\log^2(n)}.$$

Based on Lemma 6.2, we expect that:

Lemma 6.3 *For sufficiently small* $0 < \delta < 1$,

1.

$$\lim_{n \to \infty} \sup_{G \in \mathcal{G}_1^{(n)}} \left| B^2(\hat{T}_{q,n}, G) - \tilde{B}^2(T_q, G) \right| = 0.$$

2.

$$\lim_{n \to \infty} \sup_{G \in \mathcal{G}_1^{(n)}} \left| V(\hat{T}_{q,n}, G) - \tilde{V}(T_q, G) \right| = 0.$$

As a result,

$$\lim_{n \to \infty} \sup_{G \in \mathcal{G}_1^{(n)}} \left| \mathcal{R}(\hat{T}_{q,n}, G) - \tilde{\mathcal{R}}(T_q, G) \right| = 0.$$

We now consider (c). Define

$$\mathcal{G}_0^{(n)} \equiv \mathcal{G} \setminus \mathcal{G}_1^{(n)}, \qquad n \geq 1. \tag{6.7}$$

Though it is no longer sensible to require that $|T_q(G_n) - T_q(G)|$ be uniformly negligible over $\mathcal{G}_0^{(n)}$, we still hope that $T_q(G_n)$ at least stays at the *same* magnitude as $T_q(G)$, or $T_q(G_n) = O_p(\log(n))$; this turns out to be true, and in fact is a direct result of Massart (4.11).

Lemma 6.4 *Letting D_n be the event $\{\hat{T}_{q,n} \geq \log(n)/16\}$,*

$$\sup_{G \in \mathcal{G}_0^{(n)}} P_G\{D_n^c\} = 2e^{-2[(1-\sqrt{q})^2/q^2] n^{7/8}}.$$

Proof. Put for short $\tau = T_q(G)$ and $\tau_n \equiv T_q(G_n)$. For any $G \in \mathcal{G}_0^{(n)}$, over event D_n^c, $\hat{T}_{q,n} \equiv T_q(G_n)$, and $\tau_n < 2\tau$, so by Hölder and definition of the FDR functional, $\bar{G}(\tau_n) \leq (\bar{G}(2\tau_n))^{1/2} \leq (1/\sqrt{q})e^{-\tau_n}$; but $\bar{G}_n(\tau_n) \geq \frac{1}{q}e^{-\tau_n}$ and $\tau_n \leq \log(n)/16$ over D_n^c:

$$\|G_n - G\| \geq \bar{G}_n(\tau_n) - \bar{G}(\tau_n) \geq \frac{1-\sqrt{q}}{q}e^{-\tau_n} \geq \frac{1-\sqrt{q}}{q}n^{-\frac{1}{16}};$$

this implies that $D_n^c \subset \{\|G_n - G\| \geq \frac{1-\sqrt{q}}{q}n^{-\frac{1}{16}}\}$. Now use (4.11). □

Combining this with Lemma 6.1, except for an event with negligible probability, we have:

$$\log(n)/16 \quad \leq \quad \hat{T}_{q,n} \quad \leq \quad \log(n/q).$$

Since $v(t, \mu)$ is monotone decreasing in t, it is now clear that both $V(\hat{T}_{q,n}, G)$ and $\tilde{V}(T_q, G)$ are uniformly negligible over $\mathcal{G}_0^{(n)}$:

Lemma 6.5

$$\lim_{n\to\infty}\Big[\sup_{G\in\mathcal{G}_0^{(n)}} \tilde{V}(T_q, G)\Big] = 0, \qquad \lim_{n\to\infty}\Big[\sup_{G\in\mathcal{G}_0^{(n)}} V(\hat{T}_{q,n}, G)\Big] = 0.$$

Last, notice that $b(t, \mu)$ is strictly increasing in t, so either $B^2(\hat{T}_{q,n}, G)$ or $\tilde{B}^2(T_q, G)$ wouldn't be uniformly negligible over $\mathcal{G}_0^{(n)}$; however, notice that $b(t, \mu)$ increases very *slowly* in t for large t, we expect that $|B^2(\hat{T}_{q,n}, G) - \tilde{B}^2(T_q, G)|$ is uniformly negligible over $\mathcal{G}_0^{(n)}$:

Lemma 6.6

$$\lim_{n\to\infty}\Big[\sup_{G\in\mathcal{G}_0^{(n)}} |B^2(\hat{T}_{q,n}, G) - \tilde{B}^2(T_q, G)|\Big] = 0.$$

The choice of $\log(n)/8$ is only for convenience, a similar result holds if we replace $\log(n)/8$ by $c\log(n)$ for $0 < c < 1/2$.

Combining the above Lemmas yields Theorem 6.1. □

6.2 Proof of Lemma 6.1

Consider Claim 1. Sort the X_i's in descending order, $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$. First, over event W_n, $\frac{1}{q}e^{-t} > \bar{G}_n(t)$ for all $t > 0$; thus $\frac{1}{q}e^{-X_{(k)}} > \bar{G}(X_{(k)}) = \frac{k}{n}$, or $X_{(k)} < -\log(q\frac{k}{n})$, $1 \leq k \leq n$; it then follows that $\hat{\mu}_{q,n}^{FDR} \equiv 1$. Moreover, $X_{(1)} < \log(\frac{n}{q})$; since $\hat{T}_{q,n} = \log(\frac{n}{q})$ over W_n, so $\hat{\mu}_{\hat{T}_{q,n}} \equiv 1$; this shows $\hat{\mu}_{\hat{T}_{q,n}} = \hat{\mu}_{q,n}^{FDR}$ over the event W_n. Second, over the event W_n^c, $\hat{\mu}_{q,n}^{FDR}$ uses the threshold $t^{FDR} = -\log(q\frac{k_{FDR}}{n})$, $X_{(k_{FDR}+1)} < t^{FDR} \leq X_{(k_{FDR})}$, where k_{FDR} is the largest k such that $X_{(k)} \geq -\log(q\frac{n}{k})$ or $\frac{1}{q}e^{-X_{(k)}} \leq \frac{k}{n}$; since $\bar{G}(X_{(k)}) \equiv \frac{k}{n}$, equivalently, k_{FDR} is the largest k such that $\bar{G}(X_{(k)}) \geq \frac{1}{q}e^{-X_{(k)}}$; by definition of the FDR functional, this implies: $X_{(k_{FDR}+1)} < T_q(G_n) \leq X_{(k_{FDR})}$. Since over W_n^c, $\hat{T}_{q,n} \equiv T_q(G_n)$, it then follows that $\hat{\mu}_{\hat{T}_{q,n}} = \hat{\mu}_{q,n}^{FDR}$. This shows that $\hat{\mu}_{\hat{T}_{q,n}} = \hat{\mu}_{q,n}^{FDR}$ over the event W_n^c.

Consider Claim 2. It is sufficient to prove that $T_q(G_n) \leq \log(n/q)$ over W_n^c. By definition of the FDR functional, $\bar{G}(T_q(G_n)) = \frac{1}{q}e^{-T_q(G_n)}$; since the smallest non-zero value of $\bar{G}(T_q(G_n))$ is $\frac{1}{n}$, $T_q(G_n) \leq \log(n/q)$. □

6.3 Proof of Lemma 6.2

Notice that $P_G\{A_n^c\} \leq P_G\{W_n\} + P_G\{A_n^c \cap W_n^c\}$, where W_n is defined in (6.1). First, we evaluate $P_G\{W_n\}$. By Hölder and definition of the FDR functional, $\bar{G}(2T_q(G)) \geq \bar{G}^2(T_q(G)) = \frac{1}{q^2}e^{-2T_q(G)}$; moreover, on W_n, $G_n(t) \leq \frac{1}{q}e^{-t}$ for all t, particular $\bar{G}_n(2T_q(G)) \leq \frac{1}{q}e^{-2T_q(G)}$. Hence, for any $G \in \mathcal{G}_1^{(n)}$,

$$\|G - G_n\| \geq \bar{G}(2T_q(G)) - G_n(2T_q(G)) \geq \frac{1-q}{q^2}e^{-2T_q(G)} \geq \frac{1-q}{q^2}n^{-1/4};$$

this implies: $W_n \subset \{\|G - G_n\| \geq (1-q)n^{-1/4}/q^2\}$. Using Massart (4.11), we claim:

$$\sup_{G \in \mathcal{G}_1^{(n)}} P_G\{W_n\} \leq \sup_{G \in \mathcal{G}} P_G\{\|G - G_n\| \geq (1-q)n^{-1/4}/q^2\} \leq 2e^{-2(1-q)^2\sqrt{n}/q^4}. \tag{6.8}$$

Next, we evaluate $A_n^c \cap W_n^c$. Noticing $T_q(G_n) \equiv \hat{T}_{q,n}$ over event W_n^c, for sufficiently large n and $G \in \mathcal{G}_1^{(n)}$, $T_q(G) \leq \log(n)/8$, so by Lemma 4.5:

$$\begin{aligned}
A_n^c \cap W_n^c &\subset \{|T_q(G) - T_q(G_n)| \geq n^{-1/4}\} \\
&\subset \{\frac{2q}{1-q} T_q(G) e^{T_q(G)} \|G - G_n\| \geq n^{-1/4}\} \\
&= \{\|G - G_n\| \geq [(1-q)/2q]n^{-1/4}e^{-T_q(G)}/T_q(G)\} \\
&\subset \{\|G - G_n\| \geq 4(1-q)/q]n^{-3/8}/\log(n)\};
\end{aligned}$$

using again Massart (4.11):

$$\sup_{G \in \mathcal{G}_1^{(n)}} P_G\{A_n^c \cap W_n^c\} \leq 2e^{-[32(1-q)^2/q^2]n^{1/4}/\log^2(n)}. \tag{6.9}$$

Notice that for sufficiently large n, $e^{-2(1-q)^2\sqrt{n}/q^4} \ll e^{-[32(1-q)^2/q^2]n^{1/4}/\log^2(n)}$, Lemma 6.2 follows by combining (6.8) and (6.9). □

6.4 Proof of Lemma 6.3

By (6.3) - (6.4), all we need to show is (for convenience, drop the subscript for X_1 and μ_1):

$$\lim_{n\to\infty} \sup_{\mathcal{G}_1^{(n)}} \mathcal{E}\big[\log^2(\mu)|1_{\{X<\hat{T}_{q,n}\}} - 1_{\{X<T_q(G)\}}| \cdot 1_{\{A_n^c\}}\big] = 0, \tag{6.10}$$

$$\lim_{n\to\infty} \sup_{\mathcal{G}_1^{(n)}} \mathcal{E}\big[\log^2(X/\mu)|1_{\{X<\hat{T}_{q,n}\}} - 1_{\{X<T_q(G)\}}| \cdot 1_{\{A_n^c\}}\big] = 0, \tag{6.11}$$

$$\lim_{n\to\infty} \sup_{\mathcal{G}_1^{(n)}} \mathcal{E}\big[\log^2(\mu)|1_{\{X<\hat{T}_{q,n}\}} - 1_{\{X<T_q(G)\}}| \cdot 1_{\{A_n\}}\big] = 0, \tag{6.12}$$

$$\lim_{n\to\infty} \sup_{\mathcal{G}_1^{(n)}} \mathcal{E}\big[\log^2(X/\mu)|1_{\{X<\hat{T}_{q,n}\}} - 1_{\{X<T_q(G)\}}| \cdot 1_{\{A_n\}}\big] = 0. \tag{6.13}$$

To show (6.10), first, for any $G \in \mathcal{G}_1^{(n)}$, $T_q(G) \leq \log(n)/8$, and by Lemma 6.1, $\hat{T}_{q,n} \leq \log(n/q)$, so:

$$\mathcal{E}\big[\log^2(\mu)|1_{\{X<\hat{T}_{q,n}\}} - 1_{\{X<T_q(G)\}}| \cdot 1_{\{A_n^c\}}\big] \leq \mathcal{E}\big[\log^2(\mu) \cdot 1_{\{X<\log(n/q)\}} \cdot 1_{\{A_n^c\}}\big], \tag{6.14}$$

second, by Hölder:

$$\mathcal{E}\big[\log^2(\mu) \cdot 1_{\{X<\log(n/q)\}} \cdot 1_{\{A_n^c\}}\big] \leq \big(\mathcal{E}\big[\log^4(\mu) \cdot 1_{\{X\leq\log(\frac{n}{q})\}}\big]\big)^{\frac{1}{2}} \cdot \big(P_G\{A_n^c\}\big)^{1/2}, \tag{6.15}$$

last, recall that $1 - e^{-\frac{x}{\mu}} \leq x/\mu$ for any $x > 0$,

$$\mathcal{E}\big[\log^4(\mu) \cdot 1_{\{X<\log(n/q)\}}\big] = \int \log^4(\mu)(1 - e^{-\log(n/q)/\mu})dF \leq \log(n/q) \int \frac{\log^4(\mu)}{\mu} dF, \tag{6.16}$$

combining (6.14) - (6.16) and Lemma 6.2 gives (6.10).

The proof of (6.11) is similar. In fact,

$$\begin{aligned}
\mathcal{E}\big[\log^2(X/\mu)|1_{\{X<\hat{T}_{q,n}\}} - 1_{\{X<T_q(G)\}}| \cdot 1_{\{A_n^c\}}\big] &\leq \mathcal{E}\big[\log^2(X/\mu) \cdot 1_{\{A_n^c\}}\big] \\
&\leq \big(\mathcal{E}\big[\log^4(X/\mu)\big)^{\frac{1}{2}} \cdot \big(P_G\{A_n^c\}\big)^{\frac{1}{2}};
\end{aligned}$$

notice that $\mathcal{E}[\log^4(X/\mu)] = \int_0^\infty \log^4(x)e^{-x}dx < \infty$, (6.11) follows by using Lemma 6.2.

To show (6.12), recall that $e^{-\frac{t-\delta}{\mu}} - e^{-\frac{t+\delta}{\mu}} \le \frac{2\delta}{\mu}$, $0 < \delta < t$; write $\tau = T_q(G)$ for short, by the definition of A_n, (6.12) follows directly from:

$$\begin{aligned}\mathcal{E}\big[(\log^2\mu)\cdot|1_{\{X<\hat{T}_{q,n}\}} - 1_{\{X<T_q(G)\}}|\cdot 1_{\{A_n\}}\big] &\le \mathcal{E}\big[(\log^2\mu)\cdot 1_{\{\tau-n^{-1/4}\le X\le\tau+n^{-1/4}\}}\big]\\ &= \int \log^2(\mu)[e^{-[\tau-n^{-1/4}]/\mu} - e^{-[\tau+n^{-1/4}]/\mu}dF\\ &\le 2\cdot n^{-1/4}\cdot\int[\log^2(\mu)/\mu]dF.\end{aligned}$$

Similarly, for (6.13), write $\tau = T_q(G)$ for short, again by $e^{-\frac{t-\delta}{\mu}} - e^{-\frac{t+\delta}{\mu}} \le \frac{2\delta}{\mu}$, $0 < \delta < t$, (6.13) follows from:

$$\begin{aligned}\mathcal{E}\big[\log^2(X/\mu)|1_{\{X_1<\hat{T}_{q,n}\}} - 1_{\{X<\tau\}}|\cdot 1_{\{A_n\}}\big] &\le \mathcal{E}\big[(\log^2(X/\mu)\cdot 1_{\{\tau-n^{-1/4}\le X\le\tau+n^{-1/4}\}}\big]\\ &\le [\mathcal{E}\log^4(X/\mu)]^{1/2}\cdot[P_G\{|X-\tau|\le n^{-1/4}\}]^{1/2}\\ &= [\mathcal{E}\log^4(X/\mu)]^{1/2}\cdot[\int[e^{-[\tau-n^{-1/4}]/\mu} - e^{-[\tau+n^{-1/4}]/\mu}]dF]^{1/2}\\ &\le 2n^{-1/4}\cdot[\mathcal{E}\log^4(X/\mu)]^{1/2}\cdot\int(1/\mu)dF.\end{aligned}$$

□

6.5 Proof of Lemma 6.5

First, we show

$$\lim_{n\to\infty}\big[\sup_{\mathcal{G}_0^{(n)}}\tilde{V}(T_q, G)\big] = 0. \tag{6.17}$$

By monotonicity and Hölder, it is clear that when $T_q(G) > \log(n)/8$,

$$\tilde{V}(T_q, G) \le \mathcal{E}\big[\log^2(X/\mu)1_{\{X\ge\log(n)/8\}}\big] \le [\mathcal{E}\log^4(X/\mu)]^{1/2}[P_G(X\ge\log(n)/8]^{1/2};$$

by definition of the FDR functional, $P_G\{X\ge\log(n)/8\} = \bar{G}(\log(n)/8) \le \frac{1}{q}e^{-\log(n)/8}$, so (6.17) follows directly by recalling that $\mathcal{E}[\log^4(X/\mu)] = \int_0^\infty \log^4(x)e^{-x}dx < \infty$.

Second, we show

$$\lim_{n\to\infty}\big[\sup_{\mathcal{G}_0^{(n)}}V(\hat{T}_{q,n}, G)\big] = 0, \tag{6.18}$$

which is equivalent to (drop the subscript of X_1 and μ_1 for convenience):

$$\lim_{n\to\infty}\big[\sup_{\mathcal{G}_0^{(n)}}\mathcal{E}(\log^2(X/\mu)\cdot 1_{\{X\ge\hat{T}_{q,n}\}}\cdot 1_{\{B_n^c\}})\big] = 0, \tag{6.19}$$

$$\lim_{n\to\infty}\big[\sup_{\mathcal{G}_0^{(n)}}\mathcal{E}(\log^2(X/\mu)\cdot 1_{\{X\ge\hat{T}_{q,n}\}}\cdot 1_{\{B_n\}})\big] = 0. \tag{6.20}$$

First, (6.19) is the direct result of Lemma 6.4 and Hölder:

$$\mathcal{E}\big[\log^2(X/\mu)\cdot 1_{\{X\ge\hat{T}_{q,n}\}}\cdot 1_{\{B_n^c\}})\big] \le [\mathcal{E}\log^4(X/\mu)]^{1/2}\cdot[P_G\{B_n^c\}]^{1/2}.$$

Second, the proof of (6.20) is very similar to that of (6.17); in fact, since over B_n, $\hat{T}_{q,n} \ge \log(n)/16$, by monotonicity

$$\mathcal{E}\big[\log^2(X/\mu)\cdot 1_{\{X\ge\hat{T}_{q,n}\}}\cdot 1_{\{B_n\}})\big] \le \mathcal{E}\big[\log^2(X/\mu)\cdot 1_{\{X\ge\log(n)/16\}}\big],$$

and (6.20) follows by similar arguments. □

6.6 Proof of Lemma 6.6

By (6.3) - (6.4), all we need to show is (for convenience, drop the subscript for X_1 and μ_1):

$$\lim_{n\to\infty} \sup_{\mathcal{G}_0^{(n)}} \mathcal{E}\big[\log^2(\mu) |1_{\{X < \hat{T}_{q,n}\}} - 1_{\{X < T_q(G)\}}| \cdot 1_{\{B_n\}}\big] = 0, \tag{6.21}$$

$$\lim_{n\to\infty} \sup_{\mathcal{G}_0^{(n)}} \mathcal{E}\big[\log^2(\mu) |1_{\{X < \hat{T}_{q,n}\}} - 1_{\{X < T_q(G)\}}| \cdot 1_{\{B_n^c\}}\big] = 0. \tag{6.22}$$

To show (6.21), we consider the case $T_q(G) \le \hat{T}_{q,n}$ and the case $T_q(G) > \hat{T}_{q,n}$ separately.

For the case $T_q(G) \le \hat{T}_{q,n}$, recall that by Lemma 6.1, $\hat{T}_{q,n} \le \log(n/q)$, so:

$$\begin{aligned}\mathcal{E}\big[\log^2(\mu) |1_{\{X < \hat{T}_{q,n}\}} - 1_{\{X < T_q(G)\}}| \cdot 1_{\{B_n\}}\big] &\le \mathcal{E}[(\log^2(\mu) \cdot 1_{\{T_q(G) \le X \le \log(n/q)\}}] \\ &\le \mathcal{E}[(\log^2(\mu) \cdot 1_{\{T_q(G) \le X \le \log(n/q)\} + T_q(G)}] \\ &= \int e^{-\frac{T_q(G)}{\mu}} [\log^2(\mu)(1 - e^{-\frac{[\log(n/q)]}{\mu}})] dF(\mu);\end{aligned}$$

similarly, $\log^2(\mu)(1 - e^{-\log(n/q)/\mu}) \le \log(n/q)\log^2(\mu)/\mu$, so:

$$\int e^{-\frac{T_q(G)}{\mu}} [\log^2(\mu)(1 - e^{-\log(n/q)/\mu})] dF(\mu) \le \log(n/q) \cdot \max_{\{\mu \ge 1\}} \{\log^2(\mu)/\mu\} \cdot \int e^{-T_q(G)/\mu} dF,$$

but by definition of the FDR functional, for any $G \in \mathcal{G}_0^{(n)}$, $\int e^{-\frac{T_q(G)}{\mu}} dF = (1/q) e^{-T_q(G)} \le (1/q) n^{-1/8}$, (6.21) follows directly.

For the case $T_q(G) > \hat{T}_{q,n}$, let $\tau_n = \log(n)/16$ for short, since $\hat{T}_{q,n} \ge \tau_n$ over B_n:

$$\begin{aligned}\mathcal{E}\big[\log^2(\mu) |1_{\{X < \hat{T}_{q,n}\}} - 1_{\{X < T_q(G)\}}| \cdot 1_{\{B_n\}}\big] &\le \mathcal{E}[\log^2(\mu) \cdot 1_{\{\tau_n \le X \le T_q(G)\}}] \\ &= \int \log^2(\mu) [e^{-\tau_n/\mu} - e^{-T_q(G)/\mu}] dF;\end{aligned}$$

in Lemma 5.6, letting $\psi(\cdot) = \log^2(\cdot)$, $\tau = \tau_n$, it is clear that $T_q(G) > \tau$ for any $G \in \mathcal{G}_0^{(n)}$, so:

$$\int \log^2(\mu) [e^{-\tau_n/\mu} - e^{-T_q(G)/\mu}] dF \le \frac{1}{q} \cdot \max_{\{\mu \ge 1\}} \{\log^2(\mu)/\mu\} \cdot \tau_n e^{-\tau_n}/(1 - e^{-\tau_n}),$$

(6.21) follows directly for this case.

For (6.22), we also discuss the case $T_q(G) \le \hat{T}_{q,n}$ and the case $T_q(G) > \hat{T}_{q,n}$ separately. For the case $T_q(G) \le \hat{T}_{q,n}$, again by Lemma 6.1 and Hölder,

$$\begin{aligned}\mathcal{E}\big[\log^2(\mu) |1_{\{X < \hat{T}_{q,n}\}} - 1_{\{X < T_q(G)\}}| \cdot 1_{\{B_n^c\}}\big] &\le \mathcal{E}\big[\log^2(\mu) \cdot 1_{\{X \le \log(n/q)\}} \cdot 1_{\{B_n^c\}}\big] \\ &\le [\mathcal{E}\log^4(\mu) \cdot 1_{\{X \le \log(n/q)\}}]^{1/2} \cdot [P_G\{B_n^c\}]^{1/2},\end{aligned}$$

again by $(1 - e^{-x/\mu}) \le x/\mu$ for any $x \ge 0$,

$$\mathcal{E}\big[\log^4(\mu) \cdot 1_{\{X \le \log(n/q)\}}\big] = \int \log^4(\mu)(1 - e^{-\log(n/q)/\mu}) dF \le \log(n/q) \int (\log^4(\mu)/\mu) dF, \tag{6.23}$$

and (6.22) follows for this case by using Lemma 6.4.

For the case $T_q(G) > \hat{T}_{q,n}$, similarly by Hölder:

$$\begin{aligned}\mathcal{E}\big[\log^2(\mu) |1_{\{X < \hat{T}_{q,n}\}} - 1_{\{X < T_q(G)\}}| \cdot 1_{\{B_n^c\}}\big] &\le \mathcal{E}\big[\log^2(\mu) 1_{\{X < T_q(G)\}} \cdot 1_{\{B_n^c\}}\big] \\ &\le [\int \log^4(\mu)(1 - e^{-T_q(G)/\mu}) dF]^{1/2} \cdot [P_G\{B_n^c\}]^{1/2}.\end{aligned}$$

Letting $\psi(\cdot) = \log^4(\cdot)$ in 5.13, $\int \log^4(\mu)(1 - e^{-T_q(G)}) dF \le \frac{1}{q} \cdot \max_{\{\mu \ge 1\}} \{\log^4(\mu)/\mu\}$ for any $G \in \mathcal{G}$, so (6.22) follows using Lemma 6.4. □

7 Proof of Theorem 1.3

We now complete the proof of Theorem 1.3. The key point is to relate the Bayesian model of Sections 4-6 with the frequentist model of Section 1. In the frequentist model $X_i \sim \text{Exp}(\mu_i), 1 \leq i \leq n$, where $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ is an arbitrary deterministic vector $\mu \in M_{n,p}(\eta)$. Recall that $\mathcal{R}_n(\hat{T}_{q,n}, G)$ denotes the risk of FDR estimation in the Bayesian model, while $R_n(\hat{\mu}_{q,n}, \mu)$ denotes the risk in the frequentist model. Below we will show:

$$\lim_{\eta \to 0} \left[\lim_{n \to \infty} \frac{\sup_{G \in \mathcal{G}_p(\eta)} \mathcal{R}_n(\hat{T}_{q,n}, G)}{\sup_{\mu \in M_{n,p}(\eta)} R_n(\hat{\mu}_{q,n}, \mu)} \right] = 1. \tag{7.1}$$

Recall that by Theorems 1.1, 5.1, and 6.1, we have:

$$\lim_{\eta \to 0} \left[\lim_{n \to \infty} \frac{\sup_{G \in \mathcal{G}_p(\eta)} \mathcal{R}_n(\hat{T}_{q,n}, G)}{R_n^*(M_{n,p}(\eta))} \right] = \begin{cases} 1, & 0 < q \leq \frac{1}{2}, \\ \frac{q}{1-q}, & \frac{1}{2} < q < 1, \end{cases}$$

so Theorem 1.3 follows from (7.1). To prove (7.1), let now G_μ denote the mixture $G_\mu = \frac{1}{n} \sum_{i=1}^n E(\cdot/\mu_i)$. Let $\tilde{R}_n(\tilde{\mu}_{q,n}, \mu)$ denote the ideal risk for thresholding at $T_q(G_\mu)$ under the frequentist model. Let $\tilde{\mathcal{R}}(T_q, G)$ again denote the ideal risk for thresholding at $T_q(G)$ in the Bayesian model. We have the following crucial identity:

$$\tilde{R}_n(\tilde{\mu}_{q,n}, \mu) \equiv \tilde{\mathcal{R}}(T_q, G_\mu), \quad \forall \mu, n. \tag{7.2}$$

Also, note that the class of G_μ's arising from some $\mu \in M_{n,p}(\eta)$ is a subset of the class of all G's arising in $\mathcal{G}_p(\eta)$, for each $n > 0$. Hence,

$$\sup_{\mu \in M_{n,p}(\eta)} \tilde{\mathcal{R}}(T_q, G_\mu) \leq \sup_{G \in \mathcal{G}_p(\eta)} \tilde{\mathcal{R}}(T_q, G), \qquad \forall n.$$

However, notice that by Theorem 5.1, appropriately chosen 2-point priors can be asymptotically least-favorable for ideal risk in the Bayesian model. By picking μ containing entries with only the two underlying values in the least-favorable prior, and with appropriate underlying frequencies, we can obtain

$$\lim_{\eta \to 0} \left[\frac{\lim_{n \to \infty} \sup_{\mu \in M_{n,p}(\eta)} \tilde{\mathcal{R}}(T_q, G_\mu)}{\sup_{G \in \mathcal{G}_p(\eta)} \tilde{\mathcal{R}}(T_q, G)} \right] = 1. \tag{7.3}$$

Relating now the Bayesian with the frequentist model via (7.2),

$$\lim_{\eta \to 0} \left[\frac{\lim_{n \to \infty} \sup_{\mu \in M_{n,p}(\eta)} \tilde{R}_n(\hat{\mu}_{q,n}, \mu)}{\sup_{G \in \mathcal{G}_p(\eta)} \tilde{\mathcal{R}}(T_q, G)} \right] = 1. \tag{7.4}$$

Suppose we can next show that the ideal FDR risk in the frequentist model is equivalent to the true risk in the frequentist model, in the same sense as was proved in Theorem 6.1. Hence:

$$\lim_{\eta \to 0} \lim_{n \to \infty} \left[\frac{\sup_{\mu \in M_{n,p}(\eta)} R_n(\hat{\mu}_{q,n}, \mu)}{\sup_{\mu \in M_{n,p}(\eta)} \tilde{R}_n(\tilde{\mu}_{q,n}, \mu)} \right] = 1. \tag{7.5}$$

Then, that yields (7.1).

The key point is that (7.5) follows exactly as in Section 6. Indeed there is a precise analog of Theorem 6.1 for the relation between the frequentist risk and the frequentist ideal risk. This is based on two ideas.

First, if G_n now denotes the cdf of $X_1, \ldots, X_n$ *in the frequentist model*, we again have very strong convergence properties of G_n, this time to G_μ. This concerns convergence of the empirical cdf for *non-iid* samples, which is not well known, but can be found in [13, Chapter 25].

Lemma 7.1 *(Bretagnolle) Let $X_{n1}, X_{n2}, \ldots, X_{nn}$ be independent random variables with arbitrary df's F_{ni}, and $F_n(x)$ be the empirical cdf, and $\bar{F} = \text{Ave}_i\{F_{ni}\}$. Then for all $n \geq 1$, $s > 0$, there exists an absolute constant c such that*

$$Prob\{\sqrt{n}\|F_n - \bar{F}_n\| \geq s\} \leq 2ece^{-2s^2}.$$

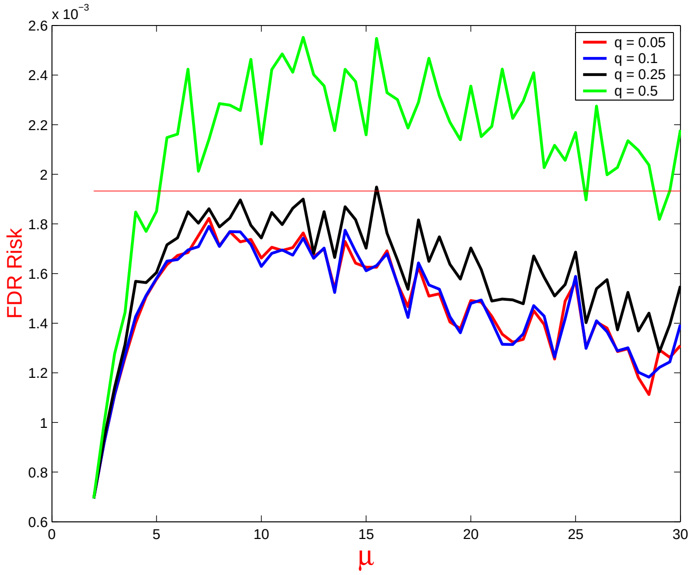

Figure 4: Simulation Results for FDR Thresholding. Curves describe per-coordinate loss of the FDR procedure with different q values at different two-point mixtures. Here the mixture concentrate at 1 and μ with mass $\epsilon = \eta/\log(\mu)$ at μ. The curves in red, green, blue, and black (displayed bottom to top) correspond to $q = 0.05, 0.1, 0.25$, and 0.5 respectively. The horizontal line corresponds to the asymptotic risk expression $\eta \log\log(\frac{1}{\eta})$.

By Massart's work ([13, Chapter 25] and [12]), we can take $c = 1$. Then, taking $F_{ni} = \mathrm{Exp}(\mu_i)$ and $\bar{F} = G_\mu$, we get

$$P_\mu\{\|G_n - G_\mu\| \geq s/\sqrt{n}\} \leq 6e^{-2s^2}, \quad \forall \mu.$$

This is completely parallel to the bound (4.11).

Second, it follows immediately from Section 4's analysis that there are frequentist fluctuation bounds for $T_q(G_n) - T_q(G_\mu)$ paralleling those in the Bayesian case. To apply this, we define:

$$M^1_{n,p}(\eta) \quad = \quad \{\mu \in M_{n,p}(\eta),\ T_q(G_\mu) \leq\ \log(n)/8\}, \tag{7.6}$$

and

$$M^0_{n,p}(\eta) \quad = \quad M_{n,p}(\eta) \setminus M^1_{n,p}(\eta). \tag{7.7}$$

Lemma 7.2 *For sufficiently small $\eta > 0$,*

1.

$$\lim_{n\to\infty}\left[\sup_{\mu \in M^1_{n,p}(\eta)} \left|R_n(\hat{\mu}_{q,n}, \mu) - \tilde{R}_n(\tilde{\mu}_{q,n}, \mu)\right|\right] = 0.$$

2.

$$\lim_{n\to\infty}\left[\sup_{\mu \in M^0_{n,p}(\eta)} \left|R_n(\hat{\mu}_{q,n}, \mu) - \tilde{R}_n(\tilde{\mu}_{q,n}, \mu)\right|\right] = 0.$$

The proof of this lemma is entirely parallel to that behind Theorem 6.1; we omit it. This completes the proof of (7.1).

8 Discussion

8.1 Illustrations

We briefly illustrate two key points.

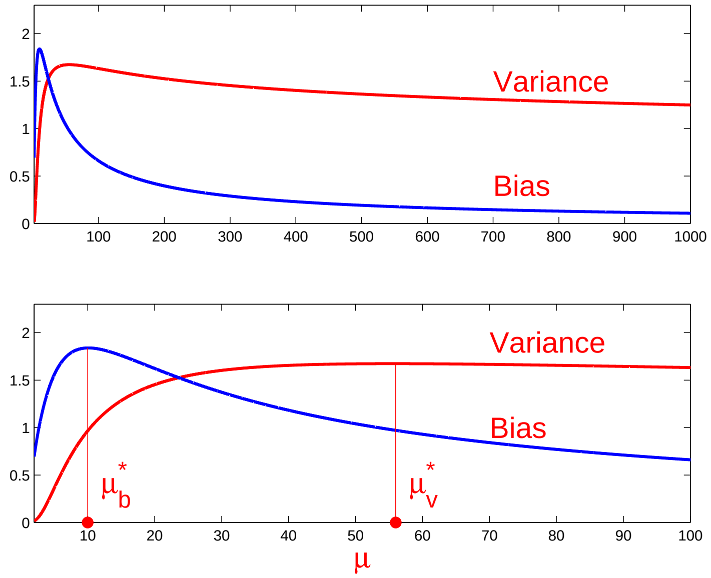

Figure 5: Panel (a): The 'bias proxy' $\tilde{B}^2(T_q, G_{\epsilon,\mu})$ and the 'variance proxy' $\tilde{V}(T_q, G_{\epsilon,1,\mu}))$. Panel (b): Enlargement of (a). The maxima of $\tilde{B}^2(T_q, G_{\epsilon,\mu}))$ and $\tilde{V}(T_q, G_{\epsilon,\mu}))$ are obtained roughly at μ_b^* and μ_v^* respectively, with $\mu_b^* = \log(\frac{1}{\eta})/\log\log\log(\frac{1}{\eta})$, $\mu_v^* = \log(\frac{1}{\eta}) \cdot \log\log(\frac{1}{\eta})$. For this figure, $\eta = 10^{-6}$.

First, we consider finite-sample performance of FDR Thresholding. Figure 4 shows the result of FDR thresholding with various values of q. It used a sample size $n = 10^6$, sparsity parameters $p = 1$, $\eta = 10^{-3}$, and a range of two-point mixtures of the kind discussed in Theorem 5.1. The figure compares the actual risk of the FDR procedure under a range of situations with the asymptotic limit given by Theorem 1.3. Clearly, the risk depends more strongly on q in finite samples than seems called for by the asymptotic expression in Theorem 1.3. In the simulations, the mixtures were based on various (ϵ, μ) pairs with μ ranging between 2 and 30, and for each μ, $\epsilon = \frac{\eta}{\log(\mu)}$.

For each $q \in \{0.05, 0.1, 0.25, 0.5\}$, we applied the FDR thresholding estimator $\hat{\mu}_{q,n}^{FDR}$, getting an empirical risk measure

$$\hat{R}(q, \mu) = \hat{R}(q, \mu; \eta, n) = \frac{1}{n}\|\log\hat{\mu}_{q,n} - \log\mu\|_2^2.$$

Figure 4 plots $\hat{R}(q, \mu; \eta, n)$ versus μ for each q. As μ varies between 2 and 30, the empirical FDR risk first increases to a maximum, then decreases; this fits well with our theory. We also notice that for q smaller than 1/2, the empirical FDR risk is not larger than $\eta\log\log(\frac{1}{\eta})$; and when q is close to 1/2, though the empirical FDR risk can be larger than $\eta\log\log(\frac{1}{\eta})$, it is rarely larger than, say, $1.3 \cdot \eta\log\log(\frac{1}{\eta})$.

Second, we illustrate the behavior of the ideal risk function indicated in the second part of Theorem 5.1. Figure 5 works out an example of the ideal risk decomposition into bias proxy and variance proxy, showing the maxima of each and the different ranges over which the two assume their large values.

8.2 Generalizations

The approach described here can be directly extended to other settings. Jin has recently derived by similar methods asymptotic minimaxity of FDR thresholding for sparse Poisson means obeying $\mu \geq 1$, with most $\mu_i = 1$. This could be useful in situations where we have a collection of 'cells' and expect one event per cell in typical cases, with occasional 'hot spots' containing more than one event per cell.

Preliminary calculations show that a wide range of non-Gaussian additive noises can also be handled by these methods. To see why, note that due to the use of $\log(\mu_i)$ in both loss measure and parameter set, results of this paper can be considered a study of FDR thresholding in a situation with *additive* noise having a standard Gumbel distribution. Thus, defining $Y_i = \log(X_i)$, the model of Section 1 posits effectively

$$Y_i = \theta_i + Z_i, \qquad i = 1, \ldots, n,$$

where $\theta_i > 0$,

$$\sum_i \theta_i^p \leq \eta^p,$$

we measure loss by $\sum_i (\hat{\theta}_i - \theta_i)^2$ and the noise Z_i obeys $e^{Z_i} \sim \mathrm{Exp}(1)$. Other additive non-Gaussian noises which have been considered include Double-Exponential. Of course, in considering non-Gaussian distributions, the effectiveness of thresholding depends on the tails of the noise distribution being sufficiently light. Thus, asymptotic minimaxity of thresholding would be doubtful for additive Cauchy noise.

Another generalization concerns dependent settings. In principle, FDR thresholding can still be 'estimating' the FDR functional in large samples even without i.i.d. stochastic disturbances. Suppose that the X_i are weakly dependent, in such a way that their empirical cdf still converges at a root-n rate. Then all the above analysis can be carried through in detail without essential change.

8.3 Relation to Other Work

There are two points of contact with earlier literature. The first of course is with the work of Abramovich, Benjamini, Donoho, and Johnstone [3]. Like the present work, [3] proves an asymptotic minimaxity property for the FDR thresholding estimator, only for Gaussian noise, and for a subtly different notion of sparsity. In [3], the sparsity parameter $\eta = \eta_n$, so that the sparsity is linked to sample size, which makes sense in a variety of nonparametric estimation applications, like wavelet denoising [1, 2, 7, 6]. In our work η goes to zero only *after* $n \to \infty$. This simplifies our analysis; the underlying tools in [3] – empirical processes, moderate deviations – are more delicate to deploy than ours. The advantage of our approach seems principally in the ease of generalization to a wider range of non-Gaussian and dependent situations.

The second connection is with the work of Genovese and Wasserman [9]. While they do not consider the estimation problem we do, they do use a Bayesian viewpoint related to Sections 4-6 in our paper. Our approach considers of course a different class of Bayesian examples, and a different notion of estimation risk. Their paper seems focused on developing intuition and broader understanding of the FDR approach, while ours uses it for solving a specific estimation problem.

9 Appendix

Proof of Lemma 3.1. Extend the function $\psi(\mu)/\phi(\mu)$ to $\mu = 1$ by defining $\psi(1)/\phi(1) = \lim_{\mu\to 1+}[\psi(\mu)/\phi(\mu)]$; notice that

$$\int \psi(\mu) dF(\mu) = \int [\frac{\psi(\mu)}{\phi(\mu)}] \cdot \phi(\mu) dF(\mu) \leq \Psi^* \cdot \int \phi(\mu) dF(\mu),$$

so it follows that $\Psi(z) \leq \Psi^* \cdot z$ for all $z \geq 0$. Moreover, for any $0 \leq z \leq z^*$, letting $F_{\{\epsilon(z),\mu^*\}}$ be the mixture of point masses at 1 and μ^* each with mass $(1-\epsilon(z))$ and $\epsilon(z)$, where $\epsilon(z) = \epsilon(z;\psi,\phi) = z/\phi(\mu_*)$, it is clear $\int \phi(\mu) dF_{\{\epsilon(z),\mu^*\}} = z$. By the assumptions that $\lim_{\mu\to\infty}[\psi(\mu)/\phi(\mu)] = 0$ and that $\lim_{\mu\to 1+}[\psi(\mu)/\phi(\mu)] < \Psi^*$, μ^* is well defined, $1 < \mu^* < \infty$, and $\psi(\mu^*)/\phi(\mu^*) = \Psi^*$; combining these:

$$\Psi(z) \geq \int \psi(\mu) dF_{\{\epsilon(z),\mu_*\}} = \epsilon(z)\psi(\mu^*) = \Psi^* \cdot z,$$

and Lemma 3.1 follows directly. □

Proof of Lemma 3.2. First, we check existence and uniqueness of μ_*. For existence: elementary calculus shows that $\Psi^{**}(\mu)$ is continuous over $(1,\bar{\mu}]$; moreover, by definition, $\Psi^{**}(\mu)$ is bounded for μ's bounded away from $\bar{\mu}$, and $\psi'(\mu)/\phi'(\mu) \to \infty$ as $\mu \to 1$, so for sufficiently small μ, $\Psi^{**}(\mu) - \psi'(\mu)/\phi'(\mu) < 0$; existence follows directly by the assumption $\Psi^{**}(\bar{\mu}) - \psi'(\bar{\mu})/\phi'(\bar{\mu}) > 0$. For uniqueness: suppose there were two solutions $1 < \mu_*^{(1)} < \mu_*^{(2)} < \bar{\mu}$. Consider $\mu_*^{(1)}$; by the assumption that $\lim_{\mu\to\infty}[\phi(\mu)/\phi(\mu)] = 0$ and ϕ is strictly increasing, it follows that

$$\lim_{\mu\to\infty} \frac{\psi(\mu) - \psi(\mu_*^{(1)})}{\phi(\mu) - \phi(\mu_*^{(1)})} \le \lim_{\mu\to\infty} \frac{\psi(\mu)}{\phi(\mu) - \phi(\mu_*^{(1)})} = 0, \tag{9.1}$$

so the supremum in the definition of $\Psi^{**}(\mu_*^{(1)})$ is attainable; thus for some $\tilde{\mu} > \bar{\mu}$ we have:

$$\Psi^{**}(\mu_*^{(1)}) = [\psi(\tilde{\mu}) - \psi(\mu_*^{(1)})]/[\phi(\tilde{\mu}) - \phi(\mu_*^{(1)})] = \psi'(\mu_*^{(1)})/\phi'(\mu_*^{(1)});$$

moreover, the strict monotonicity of $\psi'(\mu)/\phi'(\mu)$ over $(1,\bar{\mu}]$ implies that, the planar curve $\{(\phi(\mu),\psi(\mu)) : 1 \le \mu \le \bar{\mu}\}$ traces out the graph of a strictly concave function, so $[\psi(\mu_*^{(2)}) - \psi(\mu_*^{(1)})]/[\phi(\mu_*^{(2)}) - \phi(\mu_*^{(1)})] < \psi'(\mu_*^{(1)})/\phi'(\mu_*^{(1)})$; combining these we have:

$$\Psi^{**}(\mu_*^{(2)}) \ge \frac{\psi(\tilde{\mu}) - \psi(\mu_*^{(2)})}{\phi(\tilde{\mu}) - \phi(\mu_*^{(2)})} = \frac{[\psi(\tilde{\mu}) - \psi(\mu_*^{(1)})] - [\psi(\mu_*^{(2)}) - \psi(\mu_*^{(1)})]}{[\phi(\tilde{\mu}) - \phi(\mu_*^{(1)})] - [\phi(\mu_*^{(2)}) - \phi(\mu_*^{(1)})]} > \psi'(\mu_*^{(1)})/\phi'(\mu_*^{(1)}),$$

which implies $\Psi^{**}(\mu_*^{(2)}) = \psi'(\mu_*^{(2)})/\phi'(\mu_*^{(2)}) < \psi'(\mu_*^{(1)})/\phi'(\mu_*^{(1)})$; this contradicts the strict monotonicity of $\psi'(\mu)/\phi'(\mu)$ and uniqueness follows directly.

Notice here that, by similar argument in (9.1), the supremum in the definition of $\Psi^{**}(\mu_*)$ is attainable, so μ^* is well defined.

The remaining part of the claim would follow easily if we could show that for any fixed $1 < \mu_0 \le \mu_*$:

$$[\psi(\mu) - \psi(\mu_0)] \le (\psi'(\mu_0)/\phi'(\mu_0))[\phi(\mu) - \phi(\mu_0)], \qquad \forall \mu \ge 1. \tag{9.2}$$

In fact, for the case $0 < z \le \phi(\mu_*)$, taking $\mu_0 = \mu_z \equiv \phi^{-1}(z)$ in (9.2), for any F with $\int \phi(\mu)dF = z$,

$$\int [\psi(\mu) - \psi(\mu_z)]dF(\mu) \le (\psi'(\mu_z)/\phi'(\mu_z)) \int [\phi(\mu) - z]dF(\mu) = 0, \tag{9.3}$$

it then follows that $\int \psi(\mu)dF \le \psi(\mu_z)$ for any such F; this implies that $\Psi(z) \le \psi(\mu_z)$. At the same time, taking the point mass ν_{μ_z}, it is clear that $\int \phi(\mu)d\nu_{\mu_z} = z$, and $\int \psi(\mu)d\nu_{\mu_z} = \psi(\mu_z)$, so $\Psi(z) \ge \psi(\mu_z)$; combining these gives the claim in Lemma 3.2 in this case.

Second, for the case $\phi(\mu_*) < z \le \phi(\mu^*)$, take $\mu = \mu_*$ in (9.2), by the definition of Ψ^{**},

$$[\psi(\mu) - \psi(\mu_*)] \quad \le \quad \Psi^{**}(\mu_*) \cdot [\phi(\mu) - \phi(\mu_*)], \qquad 1 < \mu < \infty,$$

so for any F with $\int \phi(\mu)dF(\mu) = z$:

$$\int [\psi(\mu) - \psi(\mu_*)]dF(\mu) \le \Psi^{**}(\mu_*) \int [\phi(\mu) - \phi(\mu_*)]dF(\mu) = \Psi^{**}(\mu_*)[z - \phi(\mu_*)]; \tag{9.4}$$

this implies $\Psi(z) \le \psi(\mu_*) + \Psi^{**}(\mu_*)[z - \phi(\mu_*)]$. At the same time, let $F_{\{\epsilon(z),\mu_*,\mu^*\}}$ be the mixture of point masses at μ_* and μ^* with masses $(1-\epsilon(z))$ and $\epsilon(z)$. By direct calculation, it is clear that $\int \phi(\mu)dF_{\{\epsilon(z),\mu_*,\mu^*\}} = z$ and that $\int \psi(\mu)dF_{\{\epsilon(z),\mu_*,\mu^*\}} = \psi(\mu_*) + \Psi^{**}(\mu_*)[z - \phi(\mu_*)]$; so Lemma 3.2 also follows for the case that $\phi(\mu_*) < z \le \phi(\mu^*)$.

We now show (9.2). Fixed $1 < \mu_0 \le \mu_*$, by the definition of μ_*, $\Psi^{**}(\mu_0) \le (\psi'(\mu_0)/\phi'(\mu_0))$, combining this with the definition of $\Psi^{**}(\mu_0)$:

$$[\psi(\mu) - \psi(\mu_0)] \le \Psi^{**}(\mu_0) \cdot [\phi(\mu) - \phi(\mu_0)] \le (\psi'(\mu_0)/\phi'(\mu_0)) \cdot [\phi(\mu) - \phi(\mu_0)], \qquad \forall\ \mu \ge \bar{\mu}; \tag{9.5}$$

moreover, observe that the curve $\{(\phi(\mu),\psi(\mu)) : 1 < \mu \le \bar{\mu}\}$ in the $\phi - \psi$ plane is concave, also noticing the tangent of this curve at $(\phi(\mu_0),\psi(\mu_0))$ equals to $\psi'(\mu_0)/\phi'(\mu_0)$, we have:

$$[\psi(\mu) - \psi(\mu_0)] \le (\psi'(\mu_0)/\phi'(\mu_0)) \cdot [\phi(\mu) - \phi(\mu_0)], \qquad \forall\ 1 < \mu \le \bar{\mu}, \tag{9.6}$$

combining (9.5)-(9.6) gives (9.2). □

References

[1] Abramovich, F. and Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypothesis testing procedure. *in* A. Antoniaadis, ed., *Wavelets and Statistiscs*, **103**, Springer Verlag Lecture Notes in Statistics, pp. 5-14.

[2] Abramovich, F. and Benjamini, Y. (1995). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, **22**, 351-361.

[3] Abramovich, F. and Benjamini, Y. and Donoho, D. and Johnstone, I. (2000). Adapting to Unkown Sparsity by Controlling the False Discovery Rate. *Accepted for publication pending revision, Ann. Statist.*.

[4] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journ. Roy. Stat. Soc., Series B.*, **57**, 289-300.

[5] Bretagnolle, J. (1980). Statistique de Kolmogorov-Smirnov pour un enchantillon nonequireparti. *Colloq. internat.* CNRS, **307**, 39-44.

[6] Donoho, D. and Johnstone, I. (1994). Minimax Risk over ℓ_p-Balls for ℓ_q error. *Probability Theory and Related Fields*, 277–303.

[7] Donoho, D. and Johnstone, I. (1998). Minimax estimation via Wavelet Shrinkage. *Ann. Statist.*, **26**, 879-921.

[8] Dvoretzky, A., and Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution and of the classical multinomial estimator. *Ann. Statist.*, **27**, 642-669.

[9] Genovese, C., and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64*, **3**, 499–517.

[10] Jin, J. (2003). Detecting and Estimating Sparse Mixtures. Ph.D Thesis, Statistics Department, Stanford University.

[11] Lehmann, E. (1953). The Power of Rank Tests. *Ann. Math. Statist.* **24**, 23-43.

[12] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality, *Ann. Prob.*, **18**, 1269-1283.

[13] Shorack, G.R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* John Wiley & Sons.

[14] Sime, R. (1986). An improved Bonferroni procedure for multiple tests of significances. *Biometrika*, **73**, 751-754.