Supplementary Material for SAFFRON: an Adaptive Algorithm for Online Control of the False Discovery Rate

Aaditya Ramdas [1] **Tijana Zrnic** [2] **Martin J. Wainwright** [1] **Michael I. Jordan** [1]

1. Relationship to Storey-BH

Here, we provide details of the Benjamini-Hochberg (BH) procedure (1995), and of the relationship of its adaptive improvement, which we refer to as Storey-BH (Storey, 2002; Storey et al., 2004), to SAFFRON.

The Benjamini-Hochberg procedure is a classical method for guaranteeing FDR control in an offline setting, i.e. when all p-values are available before testing. Although the initial motivation for the BH method was different, it was reinterpreted by Storey et al. in the following manner. Since the small p-values are more likely to be non-null, suppose that one rejects all p-values below some fixed threshold $s \in (0,1)$, meaning that $\mathcal{R}(s) = \{i : P_i \leq s\}$. Then, an oracle estimate for the FDP is given by:

$$\text{FDP}^*_{\text{BH}}(s) := \frac{|\mathcal{H}^0| \cdot s}{|\mathcal{R}(s)|}.$$

The numerator is a sensible estimate because the nulls are uniformly distributed, and hence we would expect about $|\mathcal{H}^0| \cdot s$ many nulls to be below s. This is an "oracle" estimate because the scientist does not know $|\mathcal{H}^0|$. Ideally, one would like to choose a data-dependent s using the rule:

$$s^* := \max\{s : \text{FDP}^*_{\text{BH}}(s) \leq \alpha\},$$

and then reject the set $\mathcal{R}(s^*)$. Given n p-values, the BH procedure overestimates the oracle FDP by the empirically computable quantity:

$$\widehat{\text{FDP}}_{\text{BH}}(s) := \frac{n \cdot s}{|\mathcal{R}(s)|},$$

and then rejects the set $\mathcal{R}(\widehat{s}_{\text{BH}})$, where $\widehat{s}_{\text{BH}} := \max\{s : \widehat{\text{FDP}}_{\text{BH}}(s) \leq \alpha\}$. On interpreting the BH procedure in terms of an estimated FDP, Storey et al. (2002; 2004) noted that when the p-values are independent, the estimate $\widehat{\text{FDP}}_{\text{BH}}$ underutilizes the available FDR budget α. Indeed, when the p-values are exactly uniform, it is known (Benjamini & Yekutieli, 2001; Ramdas et al., 2017) to satisfy the stronger bound FDR $= \alpha|\mathcal{H}^0|/n$, which demonstrates that BH underutilizes the FDR budget of α provided to it. Instead, Storey et al. pick a constant $\lambda \in (0,1)$, and calculate:

$$\widehat{\text{FDP}}_{\text{StBH}}(s) := \frac{n \cdot s \cdot \widehat{\pi_0}}{|\mathcal{R}(s)|},$$

where the unknown proportion of nulls $\pi_0 = |\mathcal{H}^0|/n$ is estimated as:

$$\widehat{\pi_0} := \frac{1 + \sum_{i=1}^n \mathbf{1}\{P_i > \lambda\}}{n(1-\lambda)}.$$

This procedure then calculates $\widehat{s}_{\text{StBH}} := \max\{s : \widehat{\text{FDP}}_{\text{StBH}}(s) \leq \alpha\}$ and rejects the set $\mathcal{R}(\widehat{s}_{\text{StBH}})$ which satisfies the bound FDR $\leq \alpha$. We refer to this improved method as Storey-BH. Storey et al. demonstrated via simulations that the Storey-BH procedure is typically more powerful than the BH procedure, the improvement increasing with the fraction of non-nulls, and the strength of underlying signal. Since procedures such as Storey-BH adapt to the unknown proportion of nulls, they are known in the multiple testing literature as adaptive procedures.

Both BH and LORD result from a trivial upper bound on an oracle estimate of the FDP. On the other hand, Storey-BH and SAFFRON *adapt* to the unknown amount of the provided FDR budget spent on testing nulls. In the particular setting of online FDR, this corresponds to keeping a running estimate of the amount of alpha-wealth that was spent on testing nulls thus far, and not the proportion of nulls π_0, like in the case of Storey-BH; unlike the offline setting where all p-values are compared to the same level $\widehat{s}$, different p-values have to pass different thresholds α_i. In light of the above analysis, and additionally comparing the derivation of Storey-BH and SAFFRON, it is clear that Storey-BH is to BH as SAFFRON is to LORD.

It is in the above sense that SAFFRON is an adaptive online FDR method. As mentioned in Section 2.4, Foster and Stine's alpha-investing procedure is a special case of

[1]Departments of Statistics and Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, USA [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, USA. Correspondence to: Aaditya Ramdas <aramdas@eecs.berkeley.edu>, Tijana Zrnic <tijana@eecs.berkeley.edu>, Martin J. Wainwright <wainwrig@eecs.berkeley.edu>, Michael I. Jordan <jordan@eecs.berkeley.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

SAFFRON; hence, strictly speaking, alpha-investing would count as the first adaptive online FDR procedure (even though the motivation for alpha-investing in the original paper was entirely different, and did not mention estimating the FDP, or adaptivity). However, as noted in simulations by Javanmard and Montanari (2017), and re-confirmed in our simulations, alpha-investing seems *less* powerful than the non-adaptive algorithm LORD (and LORD++). As shown by simulations in Section 4, SAFFRON with constant $\lambda = 1/2$ is more powerful than LORD across a variety of signal proportions and strengths, and hence is arguably the first adaptive algorithm in the online FDR setting that can compete with the non-adaptive algorithms.

2. Additional Simulation Results

Here we provide plots demonstrating the comparison of achieved power and FDR of SAFFRON and LORD, depending on the chosen sequence $\{\gamma_j\}$. More precisely, we vary the aggressiveness of the sequence, meaning that more aggressive sequences have a higher proportion of wealth concentrated around the beginning of the sequence.

Recall that, in the setting with Gaussian observations, null p-values are computed from samples of the form $N(0,1)$, and p-values coming from the alternative are of the form $N(F_1, 1)$, where $F_1 = (\mu_c, 1)$ for some constant μ_c. The sequences considered for SAFFRON are of the form $\gamma_j \propto j^{-s}$, where the parameter $s > 1$ controls the aggressiveness of the procedure; for LORD, in addition to considering these sequences, we also consider $\gamma_j \propto \frac{\log(j \vee 2)}{j e^{\sqrt{\log j}}}$, which was shown to be the asymptotically optimal sequence for testing Gaussian means via the LORD method (Javanmard & Montanari, 2017). In Figure 1 and Figure 2 we consider $F_1 = N(2,1)$, and show how the level of aggressiveness of the sequence $\{\gamma_j\}$ affects the power and FDR of SAFFRON and LORD respectively. Figure 3 and Figure 4 demonstrate the same results in a similar, however easier, testing problem, with $F_1 = N(3,1)$.

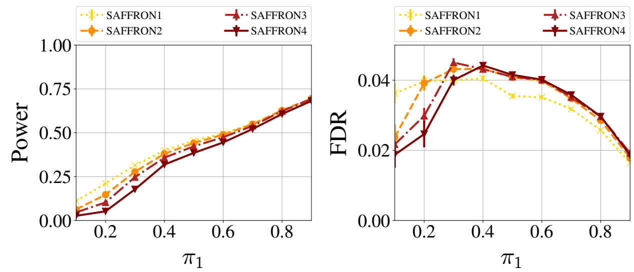

Figure 1. Statistical power and FDR versus fraction of non-null hypotheses π_1 for SAFFRON (at target FDR level $\alpha = 0.05$) using four different sequences $\{\gamma_j\}$ of increasing aggressiveness. The observations under the alternative are Gaussian with $\mu_i \sim N(2,1)$ and standard deviation 1, and are converted into one-sided p-values as $P_i = \Phi(-Z_i)$.

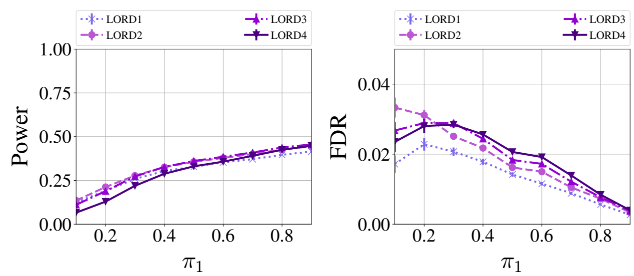

Figure 2. Statistical power and FDR versus fraction of non-null hypotheses π_1 for LORD (at target FDR level $\alpha = 0.05$) using four different sequences $\{\gamma_j\}$ of increasing aggressiveness. The LORD1 method uses the sequence proposed in the paper (Javanmard & Montanari, 2017). The observations under the alternative are Gaussian with $\mu_i \sim N(2,1)$ and standard deviation 1, and are converted into one-sided p-values as $P_i = \Phi(-Z_i)$.

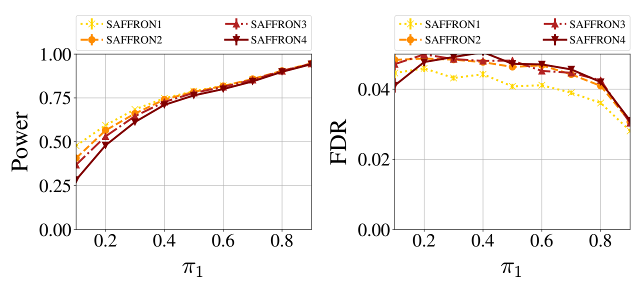

Figure 3. Statistical power and FDR versus fraction of non-null hypotheses π_1 for SAFFRON (at target FDR level $\alpha = 0.05$) using four different sequences $\{\gamma_j\}$ of increasing aggressiveness. The observations under the alternative are Gaussian with $\mu_i \sim N(3,1)$ and standard deviation 1, and are converted into one-sided p-values as $P_i = \Phi(-Z_i)$.

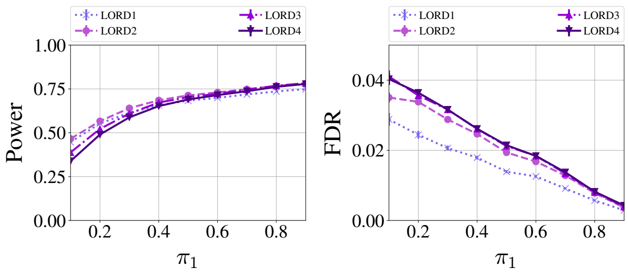

Figure 4. Statistical power and FDR versus fraction of non-null hypotheses π_1 for LORD (at target FDR level $\alpha = 0.05$) using four different sequences $\{\gamma_j\}$ of increasing aggressiveness. The LORD1 method uses the sequence proposed in the paper (Javanmard & Montanari, 2017). The observations under the alternative are Gaussian with $\mu_i \sim N(3, 1)$ and standard deviation 1, and are converted into one-sided p-values as $P_i = \Phi(-Z_i)$.

In the setting with beta alternatives, null p-values are uniformly distributed, and p-values coming from the alternative are distributed as Beta(m, n). For SAFFRON we again consider sequences $\gamma_j \propto j^{-s}$, where we vary $s > 1$, and for LORD we additionally consider $\gamma_j \propto (\frac{1}{j}\log j)^{1/m}$, which was shown to be asymptotically optimal or this testing setting (Javanmard & Montanari, 2017). Please refer to the Supplementary Material for plots of achieved power and FDR of SAFFRON and LORD obtained by varying the sequence. Figure 5 and Figure 6 show the changes in performance of SAFFRON and LORD respectively with increasing s; i.e., increasing aggressiveness of the sequence $\{\gamma_j\}$, where for the particular distribution of the observed p-values we choose $m = 0.5$ and $n = 5$.

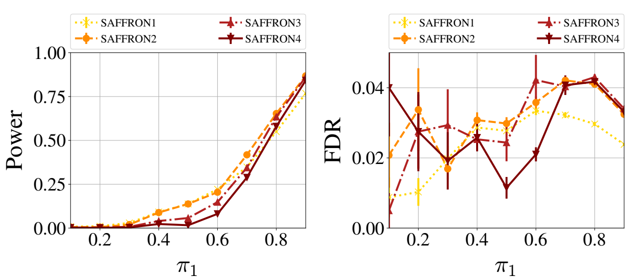

Figure 5. Statistical power and FDR versus fraction of non-null hypotheses π_1 for SAFFRON using four different sequences $\{\gamma_j\}$ of increasing aggressiveness. Under the alternative the p-values are distributed as Beta$(0.5, 5)$.

3. Monotonicity of SAFFRON

In applying the reverse super-uniformity lemma in Section 3 to prove that SAFFRON controls the FDR, it is assumed that SAFFRON is a monotone rule, meaning that

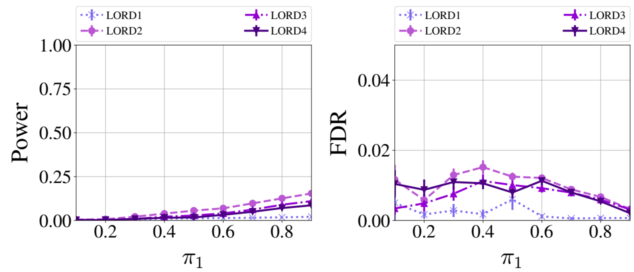

Figure 6. Statistical power and FDR versus fraction of non-null hypotheses π_1 for LORD using four different sequences $\{\gamma_j\}$ of increasing aggressiveness. The LORD1 method uses the sequence proposed in the paper (Javanmard & Montanari, 2017). Under the alternative the p-values are distributed as Beta$(0.5, 5)$.

$f_t : (R_{1:T}, C_{1:T}) \mapsto \alpha_t$ is a coordinatewise non-decreasing function. Here we provide a proof of this claim. We prove it assuming λ is constant, however the same arguments can be applied if it changes at every step, i.e. if it is predictable as stated in Section 3.

Consider some $(R_{1:T}, C_{1:T})$ and $(\tilde{R}_{1:T}, \tilde{C}_{1:T})$ for a fixed T. We will accordingly denote all relevant variables in the SAFFRON procedures which result in $(R_{1:T}, C_{1:T})$ and $(\tilde{R}_{1:T}, \tilde{C}_{1:T})$, e.g. α_t and $\tilde{\alpha}_t$, respectively. Taking into account the possible relations between indicators for rejection and candidacy, $(\tilde{R}_{1:T}, \tilde{C}_{1:T}) \succeq (R_{1:T}, C_{1:T})$ if and only if, for every $t \leq T$, one of the following holds:
(i) $R_t = \tilde{R}_t$ and $C_t = \tilde{C}_t$,
(ii) $R_t = 0$, $C_t = 1$ and $\tilde{R}_t = 1$, $\tilde{C}_t = 1$,
(iii) $R_t = 0$, $C_t = 0$ and $\tilde{R}_t = 0$, $\tilde{C}_t = 1$,
(iv) $R_t = 0$, $C_t = 0$ and $\tilde{R}_t = 1$, $\tilde{C}_t = 1$.
From this it is clear that the procedure which generated $(R_{1:T}, C_{1:T})$ up to time T could not have made more rejections or encountered more candidate p-values. Further, at each time that it made a rejection, the procedure that generated $(\tilde{R}_{1:T}, \tilde{C}_{1:T})$ also made a rejection. Looking into the SAFFRON update rule for the rejection thresholds, recall that α_t is computed as:

$$\alpha_t := \min\{\lambda, \overline{\alpha}_t\}, \;\; \text{where} \;\; \overline{\alpha}_t := W_0\gamma_{t-C_{0+}} + ((1-\lambda)\alpha - W_0)\gamma_{t-\tau_1-C_{1+}} + \sum_{j\geq 2}(1-\lambda)\alpha\gamma_{t-\tau_j-C_{j+}}.$$

Note that, by construction, the terms $((1-\lambda)\alpha - W_0)$ and $(1-\lambda)\alpha$ are strictly positive. Therefore, since the sequence $\{\gamma_j\}$ is non-increasing, the sum of the terms $(1-\lambda)\alpha\gamma_{t-\tau_j-C_{j+}}$ contributing to α_t is at most as great as the the sum of the terms $(1-\lambda)\alpha\gamma_{t-\tilde{\tau}_j-\tilde{C}_{j+}}$, because $\tilde{\alpha}_t$ considers at least all the rejection times in α_t, and has $\tilde{C}_{j+} \geq C_{j+}$ (the same holds for the term $((1-\lambda)\alpha - W_0)$).

References

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995.

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Javanmard, A. and Montanari, A. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, to appear, 2017.

Ramdas, A., Barber, R. F., Wainwright, M., and Jordan, M. A unified treatment of multiple testing with prior knowledge. *arXiv preprint arXiv:1703.06222*, 2017.

Storey, J. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64(3):479–498, 2002.

Storey, J., Taylor, J., and Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66(1):187–205, 2004.