Supplementary materials for "Nearly second-order optimality of online joint detection and estimation via one-sample update schemes"

1 Proofs

Proof of Theorem 1. In the proof, for the simplicity of notation we use N to denote $\tau(b)$. Recall θ is the true parameter. Define that

$$S_t^\theta = \sum_{i=1}^{t} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}.$$

Then under the measure $\mathbb{P}_{\theta,0}$, S_t is a random walk with i.i.d. increment. Then, by Wald's identity (e.g., Siegmund [1985]) we have that

$$\mathbb{E}_{\theta,0}[S_N^\theta] = \mathbb{E}_{\theta,0}[N] \cdot I(\theta,\theta_0). \tag{1}$$

On the other hand, let θ_N^* denote the MLE based on $(X_1,\ldots,X_N)$. The key to the proof is to decompose the stopped process S_N^θ as a summation of three terms as follows:

$$\begin{aligned} S_N^\theta =& \sum_{i=1}^{N} \log \frac{f_\theta(X_i)}{f_{\theta_N^*}(X_i)} \\ &+ \sum_{i=1}^{N} \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat\theta_{i-1}}(X_i)} \\ &+ \sum_{i=1}^{N} \log \frac{f_{\hat\theta_{i-1}}(X_i)}{f_{\theta_0}(X_i)}, \end{aligned} \tag{2}$$

Note that the first term of the decomposition on the right-hand side of (2) is always non-positive since

$$\sum_{i=1}^{N} \log \frac{f_\theta(X_i)}{f_{\theta_N^*}(X_i)} = \sum_{i=1}^{N} \log f_\theta(X_i) - \sup_{\tilde\theta\in\Theta} \sum_{i=1}^{N} \log f_{\tilde\theta}(X_i) \le 0.$$

Therefore we have

$$\begin{aligned} &\mathbb{E}_{\theta,0}[S_N^\theta] \\ \le& \mathbb{E}_{\theta,0}[\sum_{i=1}^{N} \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat\theta_{i-1}}(X_i)}] + \mathbb{E}_{\theta,0}[\sum_{i=1}^{N} \log \frac{f_{\hat\theta_{i-1}}(X_i)}{f_{\theta_0}(X_i)}]. \end{aligned}$$

Now consider the third term in the decomposition (2). Similar to the proof of equation (5.109) in Tartakovsky et al. [2014], we obtain that under the condition (11), its expectation under measure $\mathbb{P}_{\theta,0}$ is upper bounded by $b/I(\theta,\theta_0) + O(1)$ as $b\to\infty$. Then, for any positive integer n, we may further decompose the third term in (2) as

$$\sum_{i=1}^{n} \log \frac{f_{\hat\theta_{i-1}}(X_i)}{f_{\theta_0}(X_i)} = M_n(\theta) - R_n(\theta) + m_n(\theta,\theta_0) + nI(\theta,\theta_0), \tag{3}$$

where

$$M_n(\theta) = \sum_{i=1}^{n} \log \frac{f_{\hat\theta_{i-1}}(X_i)}{f_\theta(X_i)} + R_n(\theta),$$

$$R_n(\theta) = \sum_{i=1}^{n} I(\theta,\hat\theta_{i-1}),$$

and

$$m_n(\theta,\theta_0) = \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} - nI(\theta,\theta_0).$$

The decomposition of (3) consists of stochastic processes $\{M_n(\theta)\}$ and $\{m_n(\theta,\theta_0)\}$, which are both $\mathbb{P}_{\theta,0}$-martingales with zero expectation, i.e., $\mathbb{E}_{\theta,0}[M_n(\theta)] = \mathbb{E}_{\theta,0}[m_n(\theta,\theta_0)] = 0$ for any positive integer n. Since for exponential family, the log-partition function $\Phi(\theta)$ is bounded, by the inequalities for martingales Lipster and Shiryayev [1989] we have that

$$\mathbb{E}_{\theta,0}|M_n(\theta)| \le C_1\sqrt{n}, \quad \mathbb{E}_{\theta,0}|m_n(\theta,\theta_0)| \le C_2\sqrt{n}, \tag{4}$$

where C_1 and C_2 are two absolute constants that do not depend on n. Applying (4), together with condition (11), we have that $n^{-1}R_n(\theta), n^{-1}M_n(\theta)$ and $n^{-1}m_n(\theta,\theta_0)$ converge to 0 almost surely. Moreover, the convergence is $\mathbb{P}_{\theta,0}$-r-quickly for $r=1$ (For the definition of r-quick convergence, refer to Section 2.4.3 in Tartakovsky et al. [2014]). Therefore, dividing both sides of (3) by n, we obtain $n^{-1}\sum_{i=1}^{n}\log(f_{\hat\theta_{i-1}}(X_i)/f_{\theta_0}(X_i))$ converges 1-quickly to $I(\theta,\theta_0)$.

For $\epsilon>0$, we now define the last entry time

$$\begin{aligned} &L(\epsilon) \\ =& \sup\left\{ n\ge 1 : \left| \frac{1}{I(\theta,\theta_0)} \sum_{i=1}^{n} \log \frac{f_{\hat\theta_{i-1}}(X_i)}{f_{\theta_0}(X_i)} - n \right| > \epsilon n \right\}. \end{aligned}$$

By the definition of 1-quickly convergence, we have that $\mathbb{E}_{\theta,0}[L(\epsilon)] < +\infty$ for all $\epsilon > 0$. In the following, define a scaled threshold $\tilde{b} = b/I(\theta, \theta_0)$. Observe that conditioning on the event $\{L(\epsilon) + 1 < N < +\infty\}$, we have that

$$(1-\epsilon)(N-1)I(\theta,\theta_0) < \sum_{i=1}^{N-1} \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)} < b.$$

Therefore, conditioning on the event $\{L(\epsilon) + 1 < N < +\infty\}$, we have that $N < 1 + b/(1-\epsilon)$. Hence, for any $0 < \epsilon < 1$, we have

$$\begin{aligned} N \leq & 1 + \mathbb{I}(\{N > L(\epsilon)+1\}) \cdot \frac{\tilde{b}}{1-\epsilon} \\ & + \mathbb{I}(\{N \leq L(\epsilon)+1\}) \cdot L(\epsilon) \leq 1 + \frac{\tilde{b}}{1-\epsilon} + L(\epsilon). \end{aligned} \tag{5}$$

Since $\mathbb{E}_{\theta,0}[L(\epsilon)] < \infty$ for any $\epsilon > 0$, from (5) above, we have that the third term in (4) is upper bounded by $\tilde{b} + \mathsf{O}(1)$.

Finally, the second term in (2) can be written as

$$\begin{aligned} & \sum_{i=1}^{N} \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat{\theta}_{i-1}}(X_i)} \\ = & \sum_{i=1}^{N} -\log f_{\hat{\theta}_{i-1}}(X_i) - \inf_{\tilde{\theta}\in\Theta} \sum_{i=1}^{N} -\log f_{\tilde{\theta}}(X_i), \end{aligned}$$

which is just the regret defined in (10) for the online estimators: $\mathcal{R}_t$, when the loss function is defined to be the negative likelihood function. Then, the theorem is proven by combining the above analysis for the three terms in (4) and (1). □

Proof of Corollary 1. Let $\alpha = (b + O(1))/I(\theta,\theta_0)$, $\beta = C/I(\theta,\theta_0)$ and $x = \mathbb{E}_{\theta,0}[\tau(b)]$. Applying Jensen's inequality, the upper bound in equation (12) becomes $x \leq \alpha + \beta \log(x)$. From this, we have $x \leq O(\alpha)$. Taking logarithm on both sides and using the fact that $\max\{a_1 + a_2\} \leq a_1 + a_2 \leq 2\max\{a_1, a_2\}$ for $a_1, a_2 \geq 0$, $\log(x) \leq \max\{\log(2\alpha), \log(2\beta \log x)\} \leq \log(\alpha) + o(\log b)$. Therefore, we have that $x \leq \alpha + \beta(\log(\alpha) + o(\log b))$. Using this argument, we obtain

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta,\theta_0)} + \frac{C\log b}{I(\theta,\theta_0)}(1 + o(1)). \tag{6}$$

□

Next we will establish a few Lemmas useful for proving theorem 2 for sequential detection procedures. Define a measure $\mathbb{Q}$ on $(\mathcal{X}^\infty, \mathcal{B}^\infty)$ under which the probability density of X_i conditional on $\mathcal{F}_{i-1}$ is $f_{\hat{\theta}_{i-1}}$. Then for any event $A \in \mathcal{F}_i$, we have that $\mathbb{Q}(A) = \int_A \Lambda_i d\mathbb{P}_\infty$. The following lemma shows that the restriction of $\mathbb{Q}$ to $\mathcal{F}_i$ is well defined.

Lemma 1. *Let $\mathbb{Q}_i$ be the restriction of $\mathbb{Q}$ to $\mathcal{F}_i$. Then for any $A \in \mathcal{F}_k$ and any $i \geq k$, $\mathbb{Q}_i(A) = \mathbb{Q}_k(A)$.*

Proof of Lemma 1. To bound the term $\mathbb{P}_\infty(\tau(b) < \infty)$, we need take advantage of the martingale property of Λ_t in (2). The major technique is the combination of change of measure and Wald's likelihood ratio identity Siegmund [1985]. The proof is based on the method presented in Lai [2004] and Lorden and Pollak [2005].

Define the $L_i = d\mathbb{P}_i/d\mathbb{Q}_i$ as the Radon-Nikodym derivative, where $\mathbb{P}_i$ and $\mathbb{Q}_i$ are the restriction of $\mathbb{P}_\infty$ and $\mathbb{Q}$ to $\mathcal{F}_i$, respectively. Then we have that $L_i = (\Lambda_i)^{-1}$ for any $i \geq 1$ (note that Λ_i is defined in (2)). Combining the Lemma 1 and the Wald's likelihood ratio identity, we have that

$$\begin{aligned} & \mathbb{P}_\infty(A \cap \{\tau(b) < \infty\}) \\ = & \mathbb{E}_Q\left[\mathbb{I}(\{\tau(b) < \infty\}) \cdot L_{\tau(b)}\right], \forall A \in \mathcal{F}_{\tau(b)}, \end{aligned} \tag{7}$$

where $\mathbb{I}(E)$ is an indicator function that is equal to 1 for any $\omega \in E$ and is equal to 0 otherwise. By the definition of $\tau(b)$ we have that $L_{\tau(b)} \leq \exp(-b)$. Taking $A = \mathcal{X}^\infty$ in (7) we prove that $\mathbb{P}_\infty(\tau(b) < \infty) \leq \exp(-b)$. □

Proof of Corollary 2. Using (5.180) and (5.188) in Tartakovsky et al. [2014], which are about asymptotic performance of open-ended tests. Since our problem is a special case of the problem in Tartakovsky et al. [2014], we can obtain

$$\inf_{T\in C(\alpha)} \mathbb{E}_{\theta,0}[T] = \frac{\log\alpha}{I(\theta,\theta_0)} + \frac{\log(\log(1/\alpha))}{2I(\theta,\theta_0)}(1 + o(1)).$$

Combing the above result and the right-hand side of (13), we prove the corollary. □

Proof of Theorem 2. From (9), we have that for any $\nu \geq 1$,

$$\begin{aligned} & \mathbb{E}_{\theta,\nu}[T_{ASR}(b) - \nu \mid T_{ASR}(b) > \nu] \\ \leq & \mathbb{E}_{\theta,\nu}[T_{ACM}(b) - \nu \mid T_{ACM}(b) > \nu]. \end{aligned}$$

Therefore, to prove the theorem, using Theorem 1, it suffices to show that

$$\sup_{\nu\geq 0} \mathbb{E}_{\theta,\nu}[T_{ACM}(b) - \nu \mid T_{ACM}(b) > \nu] \leq \mathbb{E}_{\theta,0}[\tau(b)].$$

Using an argument similar to the remarks in Lorden and Pollak [2005], we have that the supreme of detection delay over all change locations is achieved by the case when change occurs at the first instance.

$$\sup_{\nu\geq 0} \mathbb{E}_{\theta,\nu}[T_{ACM}(b) - \nu \mid T_{ACM}(b) > \nu] = \mathbb{E}_{\theta,0}[T_{ACM}(b)]. \tag{8}$$

Notice that since θ_0 is known, for any $j \geq 1$, the distribution of $\{\max_{j+1\leq k\leq t} \Lambda_{k,t}\}_{t=j+1}^{\infty}$ under $\mathbb{P}_{\theta,j}$ conditional on $\mathcal{F}_j$ is the same as the distribution of $\{\max_{1\leq k\leq t} \Lambda_{k,t}\}_{t=1}^{\infty}$ under $\mathbb{P}_{\theta,0}$. Below, we use a renewal property of the ACM procedure. Define

$$T_{ACM}^{(j)}(b) = \inf\{t > j : \max_{j+1\leq k\leq t} \log \Lambda_{k,t} > b\}.$$

Then we have that $\mathbb{E}_{\theta,0}[T_{ACM}(b)] = \mathbb{E}_{\theta,j}[T_{ACM}^{(j)}(b) - j \mid T_{ACM}^{(j)}(b) > j]$. However, $\max_{1\leq k\leq t} \log \Lambda_{k,t} \geq \max_{j+1\leq k\leq t} \Lambda_{k,t}$ for any $t > j$. Therefore, $T_{ACM}^{(j)}(b) \geq T_{ACM}(b)$ conditioning on $\{T_{ACM}(b) > j\}$. So that for all $j \geq 1$,

$$\begin{aligned}\mathbb{E}_{\theta,0}[T_{ACM}(b)] =& \mathbb{E}_{\theta,j}[T_{ACM}^{(j)}(b) - j \mid T_{ACM}(b) > j] \\ \geq& \mathbb{E}_{\theta,j}[T_{ACM}(b) - j \mid T_{ACM}(b) > j].\end{aligned}$$

Thus, to prove (8), it suffices to show that $\mathbb{E}_{\theta,0}[T_{ACM}(b)] \leq \mathbb{E}_{\theta,0}[\tau(b)]$. To show this, define $\tau(b)^{(t)}$ as the new stopping time that applies the sequential hypothesis testing procedure $\tau(b)$ to data $\{X_i\}_{i=t}^{\infty}$. Then we have that in fact $T_{ACM}(b) = \min_{t\geq 1}\{\tau(b)^{(t)} + t - 1\}$, this relationship was developed in Lorden [1971]. Thus, $T_{ACM}(b) \leq \tau(b)^{(1)} + 1 - 1 = \tau(b)$, and $\mathbb{E}_{\theta,0}[T_{ACM}(b)] \leq \mathbb{E}_{\theta,0}[\tau(b)]$. □

Proof of Lemma 2. First, rewrite $T_{ASR}(b)$ as

$$T_{ASR}(b) = \inf\left\{t \geq 1 : \log\left(\sum_{k=1}^{t} \Lambda_{k,t}\right) > b\right\}.$$

Next, since

$$\log\left(\sum_{k=1}^{t} \Lambda_{k,t}\right) > \log\left(\max_{1\leq k\leq t} \Lambda_{k,t}\right) = \max_{1\leq k\leq t} \log \Lambda_{k,t}, \tag{9}$$

we have $\mathbb{E}_{\infty}[T_{ACM}(b)] \geq \mathbb{E}_{\infty}[T_{ASR}(b)]$. So it suffices to show that $\mathbb{E}_{\infty}[T_{ASR}(b)] \geq \gamma$, if $b \geq \log \gamma$. Define $R_t = \sum_{k=1}^{t} \Lambda_{k,t}$. Direct computation shows that

$$\begin{aligned}&\mathbb{E}_{\infty}[R_t \mid \mathcal{F}_{t-1}] \\ =&\mathbb{E}_{\infty}\left[\Lambda_{t,t} + \sum_{k=1}^{t-1} \Lambda_{k,t} \mid \mathcal{F}_{t-1}\right] \\ =&\mathbb{E}_{\infty}\left[1 + \sum_{k=1}^{t-1} \Lambda_{k,t-1} \cdot \log \frac{f_{\hat{\theta}_{t-1}}(X_t)}{f_{\theta_0}(X_t)} \mid \mathcal{F}_{t-1}\right] \\ =&1 + \sum_{k=1}^{t-1} \Lambda_{k,t-1} \cdot \mathbb{E}_{\infty}\left[\log \frac{f_{\hat{\theta}_{t-1}}(X_t)}{f_{\theta_0}(X_t)} \mid \mathcal{F}_{t-1}\right] \\ =&1 + R_{t-1}.\end{aligned}$$

Therefore, $\{R_t - t\}_{t\geq 1}$ is a $(\mathbb{P}_{\infty}, \mathcal{F}_t)$-martingale with zero mean. Suppose that $\mathbb{E}_{\infty}[T_{ASR}(b)] < \infty$ (otherwise the statement of proposition is trivial), then we have that

$$\sum_{t=1}^{\infty} \mathbb{P}_{\infty}(T_{ASR}(b) \geq t) < \infty. \tag{10}$$

(10) leads to the fact that $\mathbb{P}_{\infty}(T_{ASR}(b)) \geq t = o(t^{-1})$ and the fact that $0 \leq R_t \leq \exp(b)$ conditioning on the event $\{T_{ASR}(b) > t\}$, we have that

$$\begin{aligned}&\liminf_{t\to\infty} \int_{\{T_{ASR}(b)>t\}} |R_t - t| d\mathbb{P}_{\infty} \\ \leq& \liminf_{t\to\infty} (\exp(b) + t)\mathbb{P}_{\infty}(T_{ASR}(b) \geq t) = 0.\end{aligned} \tag{11}$$

Therefore, we can apply the optional stopping theorem for martingale, to obtain that $\mathbb{E}_{\infty}[R_{T_{ASR}(b)}] = \mathbb{E}_{\infty}[T_{ASR}(b)]$. By the definition of $T_{ASR}(b)$, $R_{T_{ASR}(b)} > \exp(b)$ we have that $\mathbb{E}_{\infty}[T_{ASR}(b)] > \exp(b)$. Therefore, if $b \geq \log \gamma$, we have that $\mathbb{E}_{\infty}[T_{ACM}(b)] \geq \mathbb{E}_{\infty}[T_{ASR}(b)] \geq \gamma$. □

Proof of Corollary 3. Our Theorem 1 and the remarks in Siegmund and Yakir [2008] show that the minimum worst-case detection delay, given a fixed ARL level γ, is given by

$$\begin{aligned}&\inf_{T(b)\in S(\gamma)} \sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] \\ =&\frac{\log \gamma}{I(\theta,\theta_0)} + \frac{d \log\log \gamma}{2I(\theta,\theta_0)}(1 + o(1)).\end{aligned} \tag{12}$$

It can be shown that the infimum is attained by choosing $T(b)$ as a weighted Shiryayev-Roberts detection procedure, with a careful choice of the weight over the parameter space Θ. Combing (12) with the right-hand side of (13), we prove the corollary. □

2 Regret bound for OMD

In this subsection, we show that the regret bound $\mathcal{R}_t$ can be expressed as a weighted sum of Bregman divergences between two consecutive estimators. This form of $\mathcal{R}_t$ is useful in the showing of the logarithmic expected regret property. This is also useful in showing how the assumptions required by Corollary 1 are satisfied. The following result comes as a modification of Azoury and Warmuth [2001].

Theorem 1. *Assume that $X_1, X_2, \ldots$ are i.i.d. random variables with density function $f_\theta(x)$. Let $\eta_i = 1/i$ in Algorithm 1. Assume that $\{\hat{\theta}_i\}_{i\geq 1}, \{\hat{\mu}_i\}_{i\geq 1}$ are obtained using Algorithm 1 and $\hat{\theta}_i = \tilde{\theta}_i$ for any $i \geq 1$.*

Then for any $\theta_0 \in \Theta$ and $t \geq 1$,

$$\begin{aligned}\mathcal{R}_t =& \sum_{i=1}^{t} i \cdot B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) \\ =& \frac{1}{2}\sum_{i=1}^{t} i \cdot (\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal [\nabla^2 \Phi^*(\tilde{\mu}_i)](\hat{\mu}_i - \hat{\mu}_{i-1}),\end{aligned}$$

where $\tilde{\mu}_i = \lambda\hat{\mu}_i + (1-\lambda)\hat{\mu}_{i-1}$, for some $\lambda \in (0,1)$.

Let us delay the proof for Theorem 1 a bit and first see how to use Theorem 1 by a concrete example with multivariate normal distribution, $\{\mathcal{P}_\theta, \theta \in \Theta\}$ with unknown mean parameter θ, and known covariance matrix I_d (I_d is a $d \times d$ identity matrix), denoted by $\mathcal{N}(\theta, I_d)$. Here $\phi(x) = x$, $dH(x) = (1/\sqrt{|2\pi I_d|}) \cdot \exp(-x^\intercal x/2)$, $\Theta = \Theta_\sigma = \mathbb{R}^d$ for any $\sigma < 2$, $\Phi(\theta) = (1/2)\theta^\intercal\theta$, $\mu = \theta$ and $\Phi^*(\mu) = (1/2)\mu^\intercal\mu$, where $|\cdot|$ denotes the determinant of a matrix, and H is a probability measure under which the sample follows $\mathcal{N}(0, I_d)$). When the covariance matrix is known to be some $\Sigma \neq I_d$, one can "whiten" the vectors by multiplying $\Sigma^{-1/2}$ to obtain the situation here.

Corollary 1 (Upper bound for expected regret bound, Gaussian)**.** *Assume $X_1, X_2, \ldots$ are i.i.d. following $\mathcal{N}(\theta, I_d)$ with some $\theta \in \mathbb{R}^d$. Assume that $\{\hat{\theta}_i\}_{i\geq 1}, \{\hat{\mu}_i\}_{i\geq 1}$ are obtained using Algorithm 1 with $\eta_i = 1/i$ and $\Gamma = \mathbb{R}^d$. For any $t > 0$, we have that for some constant $C_1 > 0$ that depends on θ,*

$$\mathbb{E}_{\theta,0}[\mathcal{R}_t] \leq C_1 d \log t/2.$$

The following calculations justify Corollary 1, which also serve as an example of how to use regret bound. First, the assumption $\hat{\theta}_t = \tilde{\theta}_t$ in Theorem 1 is satisfied for the following reasons. Consider $\Gamma = \mathbb{R}^d$ is the full space. According to Algorithm 1, using the non-negativity of the Bregman divergence, we have $\hat{\theta}_t = \arg\min_{u\in\Gamma} B_\Phi(u, \tilde{\theta}_t) = \tilde{\theta}_t$. The the regret bound can be written as

$$\begin{aligned}\mathcal{R}_t =& \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^\intercal(\hat{\mu}_1 - \hat{\mu}_0) \\ &+ \frac{1}{2}\sum_{i=2}^{t}[i \cdot (\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal(\hat{\mu}_i - \hat{\mu}_{i-1})] \\ =& \frac{1}{2}(X_1 - \theta_0)^\intercal(X_1 - \theta_0) \\ &+ \frac{1}{2}\sum_{i=2}^{t}(\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal(\phi(X_i) - \hat{\mu}_{i-1}).\end{aligned}$$

Since the step-size $\eta_i = 1/i$, the second term in the above equation can be written as:

$$\begin{aligned}&\frac{1}{2}\sum_{i=2}^{t}(\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal(\phi(X_i) - \hat{\mu}_{i-1}) \\ =& \frac{1}{2}\sum_{i=2}^{t}(\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal(\phi(X_i) + \hat{\mu}_i) \\ &- \sum_{i=2}^{t}\frac{1}{2}(\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal(\hat{\mu}_{i-1} + \hat{\mu}_i) \\ =& \sum_{i=2}^{t}\frac{1}{2(i-1)}(\phi(X_i) - \hat{\mu}_i)^\intercal(\phi(X_i) + \hat{\mu}_i) \\ &+ \sum_{i=2}^{t}\frac{1}{2}(\|\hat{\mu}_{i-1}\|^2 - \|\hat{\mu}_i\|^2) \\ =& \sum_{i=2}^{t}\frac{1}{2(i-1)}\|X_i\|^2 - \sum_{i=2}^{t}\frac{1}{2(i-1)}\|\hat{\mu}_i\|^2 \\ &+ \frac{1}{2}\|\hat{\mu}_1\|^2 - \frac{1}{2}\|\hat{\mu}_t\|^2.\end{aligned}$$

Combining above, we have

$$\begin{aligned}\mathbb{E}_{\theta,0}[\mathcal{R}_t] \leq& \frac{1}{2}\mathbb{E}_{\theta,0}[(X_1 - \theta_0)^\intercal(X_1 - \theta_0)] \\ &+ \frac{1}{2}\sum_{i=2}^{t}\frac{1}{i-1}\mathbb{E}_{\theta,0}[\|X_i\|^2] + \frac{1}{2}\mathbb{E}_{\theta,0}[\|X_1\|^2].\end{aligned}$$

Finally, since $\mathbb{E}_{\theta,0}[\|X_i\|^2] = d(1+\theta^2)$ for any $i \geq 1$, we obtain desired result. Thus, with i.i.d. multivariate normal samples, the expected regret grows logarithmically with the number of observations.

Using similar calculation, we can also bound the expected regret in the general case. As shown in the proof above for Corollary 1, the dominating term for $\mathcal{R}_t$ can be rewritten as

$$\sum_{i=2}^{t}\frac{1}{2(i-1)}(\phi(X_i) - \hat{\mu}_i)^\intercal[\nabla^2\Phi^*(\tilde{\mu}_i)](\phi(X_i) + \hat{\mu}_i),$$

where $\tilde{\mu}_i$ is a convex combination of $\hat{\mu}_{i-1}$ and $\hat{\mu}_i$. For an arbitrary distribution, the term $(\phi(X_i) - \hat{\mu}_i)^\intercal[\nabla^2\Phi^*(\tilde{\mu}_i)](\phi(X_i)+\hat{\mu}_i)$ can be viewed as a local normal distribution with the changing curvature $\nabla^2\Phi^*(\tilde{\mu}_i)$. Thus, it is possible to prove case-by-case the $O(\log t)$-style bounds. Proofs for Bernoulli distribution and Gamma distribution can be found in Azoury and Warmuth [2001]. Proof of OCM for covariance matrix in multivariate normal can be found in Dasgupta and Hsu [2007]. A more general solution can be found in the Theorem 3 in Raginsky et al. [2012], which however requires stronger conditions.

The following derivation borrows ideas from [Azoury and Warmuth, 2001]. First, we derive concise forms of the two terms in the definition of R_t in (10).

Lemma 2. *Assume that $X_1, X_2, \dots$ are i.i.d. random variables with density function $f_\theta(x)$, and assume decreasing step-size $\eta_i = 1/i$ in Algorithm 1. Given $\{\hat{\theta}_i\}_{i\geq 1}, \{\hat{\mu}_i\}_{i\geq 1}$ generated by Algorithm 1. If $\hat{\theta}_i = \tilde{\theta}_i$ for any $i \geq 1$, then for any null distribution parameter $\theta_0 \in \Theta$ and $t \geq 1$,*

$$\sum_{i=1}^{t}\{-\log f_{\hat{\theta}_{i-1}}(X_i)\} = \sum_{i=1}^{t} iB_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t). \tag{13}$$

Moreover, for any $t \geq 1$,

$$\inf_{\tilde{\theta}\in\Theta}\sum_{i=1}^{t}\{-\log f_{\tilde{\theta}}(X_i)\} = -t\Phi^*(\hat{\mu}), \tag{14}$$

where $\hat{\mu} = (1/t)\cdot\sum_{i=1}^{t}\phi(X_i)$.

By subtracting the expressions in (13) and (14), we obtain the following result which shows that the regret can be represented by a weighted sum of the Bregman divergences between two consecutive estimators.

Proof of Lemma 2. By the definition of the Legendre-Fenchel dual function we have that $\Phi^*(\mu) = \theta^\intercal\mu - \Phi(\theta)$ for any $\theta \in \Theta$. By this definition, and choosing $\eta_i = 1/i$, we have that for any $i \geq 1$

$$\begin{aligned}
&-\log f_{\hat{\theta}_{i-1}}(X_i)\\
=&\Phi(\hat{\theta}_{i-1}) - \hat{\theta}_{t-1}^\intercal\phi(X_i)\\
=&\hat{\theta}_{i-1}^\intercal(\hat{\mu}_{t-1} - \phi(X_i)) - \Phi^*(\hat{\mu}_{i-1})\\
=&\frac{1}{\eta_i}\hat{\theta}_{i-1}^\intercal(\hat{\mu}_{i-1} - \hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1})\\
=&\frac{1}{\eta_i}(\Phi^*(\hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1})) - \hat{\theta}_{i-1}^\intercal(\hat{\mu}_i - \hat{\mu}_{i-1})\\
&-\frac{1}{\eta_i}\Phi^*(\hat{\mu}_i) + \left(\frac{1}{\eta_i} - 1\right)\Phi^*(\hat{\mu}_{i-1})\\
=&\frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) + \frac{1}{\eta_{i-1}}\Phi^*(\hat{\mu}_{i-1}) - \frac{1}{\eta_i}\Phi^*(\hat{\mu}_i),
\end{aligned} \tag{15}$$

where we use the update rule in Line 6 of Algorithm 1 and the assumption $\hat{\theta}_i = \tilde{\theta}_i$ to have the third equation. We define $1/\eta_0 = 0$ in the last equation. Now summing the terms in (15), where the second term form a telescopic series, over i from 1 to t, we have that

$$\begin{aligned}
&\sum_{i=1}^{t}\{-\log f_{\hat{\theta}_{i-1}}(X_i)\}\\
=&\sum_{i=1}^{t}\frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) + \frac{1}{\eta_0}\Phi^*(\hat{\mu}_0) - \frac{1}{\eta_t}\Phi^*(\hat{\mu}_t)\\
=&\sum_{i=1}^{t}\frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t).
\end{aligned}$$

Moreover, from the definition we have that

$$\sum_{i=1}^{t}\{-\log f_\theta(X_i)\} = \sum_{i=1}^{t}\left[\Phi(\theta) - \theta^\intercal\phi(X_i)\right].$$

Taking the first derivative of $\sum_{i=1}^{t}\{-\log f_\theta(X_i)\}$ with respect to θ and setting it to 0, we find $\hat{\mu}$, the stationary point, given by

$$\hat{\mu} = \nabla\Phi(\theta) = \frac{1}{t}\sum_{i=1}^{t}\phi(X_i).$$

Similarly, using the expression of the dual function, and plugging $\hat{\mu}$ back into the equation, we have that

$$\begin{aligned}
&\inf_{\tilde{\theta}\in\Theta}\sum_{i=1}^{t}\{-\log f_{\tilde{\theta}}(X_i)\}\\
=&t\cdot\theta^\intercal\hat{\mu} - t\Phi^*(\hat{\mu}) - \sum_{i=1}^{t}\theta^\intercal\phi(X_i) = -t\Phi^*(\hat{\mu}).
\end{aligned}$$

□

Proof of Theorem 1. By choosing the step-size $\eta_i = 1/i$ for any $i \geq 1$ in Algorithm 1, and assuming $\hat{\theta}_i = \tilde{\theta}_i$ for any $i \geq 1$, we have by induction that

$$\hat{\mu}_t = \frac{1}{t}\sum_{i=1}^{t}\phi(X_i) = \hat{\mu}.$$

Subtracting (13) by (14), we obtain

$$\begin{aligned}
\mathcal{R}_t &= \sum_{i=1}^{t}\{-\log f_{\hat{\theta}_{i-1}}(X_i)\} - \inf_{\tilde{\theta}\in\Theta}\sum_{i=1}^{t}\{-\log f_{\tilde{\theta}}(X_i)\}\\
&= \sum_{i=1}^{t} iB_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t) + t\Phi^*(\hat{\mu})\\
&= \sum_{i=1}^{t} iB_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1})\\
&= \sum_{i=1}^{t} i[\Phi^*(\hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1}) - \langle\nabla\Phi^*(\hat{\mu}_{i-1}), \hat{\mu}_i - \hat{\mu}_{i-1}\rangle]\\
&= \frac{1}{2}\sum_{i=1}^{t} i\cdot(\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal[\nabla^2\Phi^*(\tilde{\mu}_i)](\hat{\mu}_i - \hat{\mu}_{i-1}).
\end{aligned}$$

The final equality is obtained by Taylor expansion. □

References

D. Siegmund. *Sequential analysis: tests and confidence intervals*. Springer-Verlag, 1985.

A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

R. Lipster and A. Shiryayev. Theory of martingales. 1989.

T.-Z. Lai. Likelihood ratio identities and their applications to sequential analysis. *Sequential Analysis*, 23 (4):467–497, 2004.

G. Lorden and M. Pollak. Nonanticipating estimation applied to sequential analysis and changepoint detection. *Annals of statistics*, pages 1422–1454, 2005.

G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908, 1971.

D. Siegmund and B. Yakir. Minimax optimality of the Shiryayev-Roberts change-point detection rule. *Journal of Statistical Planning and Inference*, 138(9): 2815–2825, 2008.

K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3): 211–246, 2001.

Sanjoy Dasgupta and Daniel Hsu. On-line estimation with the multivariate gaussian distribution. *Learning Theory*, pages 278–292, 2007.

M. Raginsky, R. Willet, C. Horn, J. Silva, and R. Marcia. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, 58(8):5544–5562, 2012.