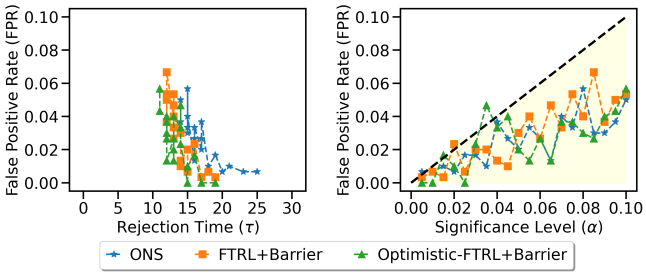


*Figure 1.* **(One-dimensional Setting.)** Comparisons of three different OAlg methods in **1D setting** (the results for the first synthetic experiment: distributions with disjoint supports). The left plot shows the average time for Algorithm 1 to correctly reject $H_0$ versus the false positive rates (FPRs) under each value of the significance level ($\alpha$) over 300 runs. The plots closer to the bottom left are more desirable. The right plot shows FPRs under each $\alpha$ when the null $H_0$ holds, with the dashed line and shaded area representing the desired significance levels. In the 1D setting, all three OAlg methods reject $H_0$ within 30 rounds under $H_1$. The two proposed methods consistently outperform ONS. For example, at $\alpha = 0.05$, FTRL+Barrier and Optimistic-FTRL+Barrier achieve around 20% faster rejection than ONS.

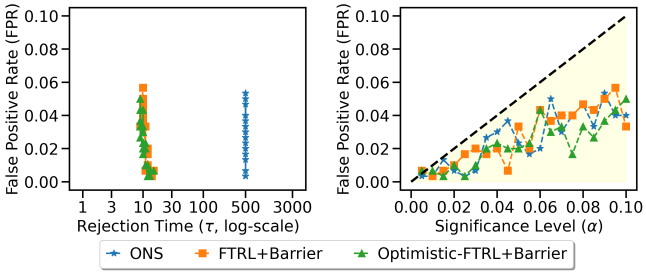


*Figure 2.* **(Multi-dimensional Setting.)** Comparisons of three different OAlg methods in **2D setting**. Specifically, we revisit the first synthetic experiment (distributions with disjoint supports.), and extend it to the 2-dimensional case. In the 2D setting, the two proposed OAlg methods still reject $H_0$ within 20 rounds when $H_1$ holds, maintaining similarly strong performance as in the 1D case. In contrast, the ONS curve appears as a vertical line, which indicates that the null hypothesis is not rejected within the time budget $T = 500$.

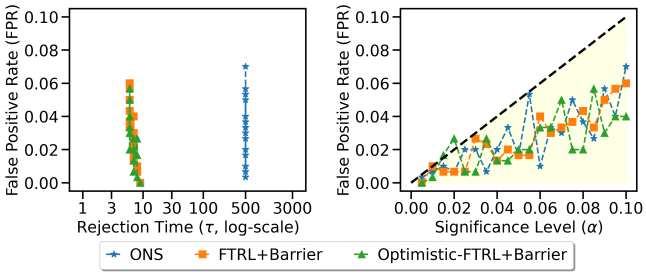


*Figure 3.* **(Multi-dimensional Setting.)** Comparisons of three different OAlg methods in **10D setting**. Specifically, we revisit the first synthetic experiment (distributions with disjoint supports.), and extend it to the 10-dimensional case. In the 10D setting, the two proposed OAlg methods maintain their effectiveness, and even exhibit slightly faster rejection compared to their 1D and 2D counterparts. In contrast, the ONS curve appears as a vertical line, indicating that the null hypothesis $H_0$ could not be rejected before the time budget $T = 500$.

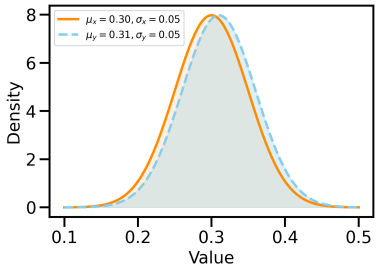


*Figure 4.* Normal distributions with overlapping supports; **low signal-to-noise ratio** ($H_1$ scenario).

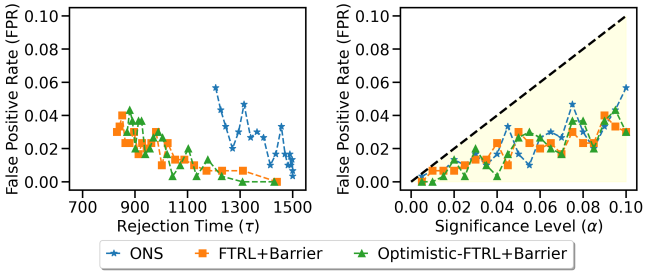


*Figure 5.* **(Low Signal-to-Noise Ratio (SNR) Scenario.)** Comparisons of OAlg methods in **low SNR** setting. In this experiment, each sequence contains 1500 samples (i.e., time budget $T = 1500$). The left plot shows the average time for Algorithm 1 to correctly reject $H_0$ versus the false positive rates (FPRs) under each value of the significance level ($\alpha$) over 300 runs. The plots closer to the bottom left are more desirable. The right plot shows FPRs under each $\alpha$ when the null $H_0$ holds, with the dashed line and shaded area representing the desired significance levels. Compared to the high SNR scenario (Example 2 - Figure 10 in Appendix E), three OAlg methods show **longer rejection time** under $H_1$, while our methods **outperform ONS** by rejecting the null hypothesis more quickly while maintaining control of the type-I error below the significance level $\alpha$. For example, when the significance level $\alpha$ is set to 0.05, the null hypothesis is rejected at step 1436 by ONS, at step 980 by FTRL, and at step 1005 by Optimistic-FTRL. Compared to ONS, Algorithm 1 using FTRL+Barrier or Optimistic-FTRL+Barrier as the online update method rejects the null under $H_1$ **approximately 30.0% faster**.

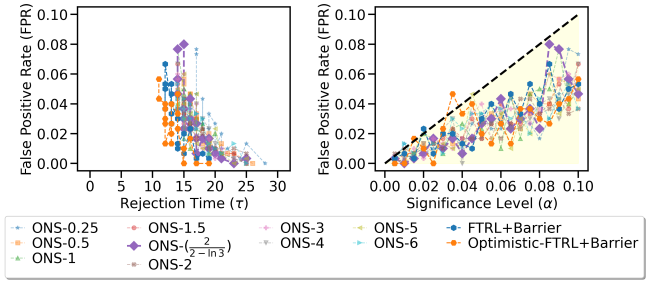


*Figure 6.* **(Tuning the Learning Rate in ONS.)** Comparisons of our proposed OAlg methods and ONS with different learning rates for distributions with disjoint supports **(synthetic dataset for Example 1)**. We consider 10 learning rate values for ONS, including $\{0.25, 0.5, 1, 1.5, 2, \frac{2}{2-\ln 3}, 3, 4, 5, 6\}$, where $\frac{2}{2-\ln 3}$ is the empirical value commonly used in all related literature to the best of our knowledge. The left subfigure illustrates the average time required to correctly reject the null $H_0$ versus the average false positive rates (FPRs) over 300 runs under different significance levels ($\alpha$). The right subfigure evaluates the performance of each online algorithm under the null hypothesis, with a dashed line and shaded area representing the expected significance levels.

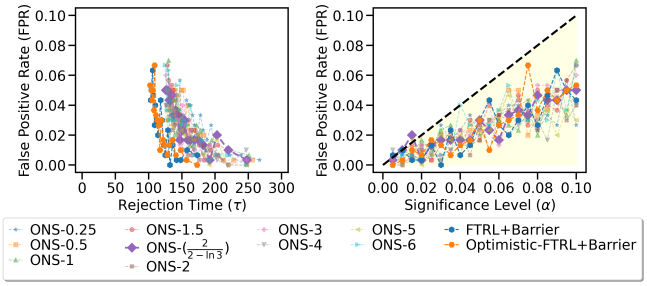


*Figure 7.* **(Tuning the Learning Rate in ONS.)** Comparisons of our proposed OAlg methods and ONS with different learning rates for distributions with overlapping supports; high signal-to-noise ratio **(synthetic dataset for Example 2)**. The left subfigure illustrates the average time required to correctly reject the null $H_0$ versus the average false positive rates (FPRs) over 300 runs under different significance levels ($\alpha$). The right subfigure evaluates the performance of each online algorithm under the null hypothesis, with a dashed line and shaded area representing the expected significance levels.

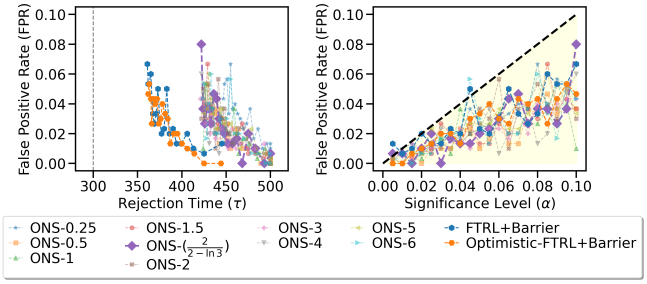


*Figure 8.* **(Tuning the Learning Rate in ONS.)** Comparisons of our proposed OAlg methods and ONS with different learning rates for time-varying distributions with mean shift **(synthetic dataset for Example 3)**. $\{x_t\}_{t\geq 1} \sim \mathcal{N}(0.30, 0.05^2)$. $H_0 : \{y_t\}_{t\geq 1} \sim \mathcal{N}(0.30, 0.05^2)$ versus. $H_1 : \{y_t\}_{1\leq t<300} \sim \mathcal{N}(0.30, 0.05^2), \{y_t\}_{t\geq 300} \sim \mathcal{N}(0.35, 0.05^2)$. The left subfigure illustrates the average time required to correctly reject the null $H_0$ versus the average false positive rates (FPRs) over 300 runs under different significance levels ($\alpha$). The right subfigure evaluates the performance of each online algorithm under the null hypothesis, with a dashed line and shaded area representing the expected significance levels.

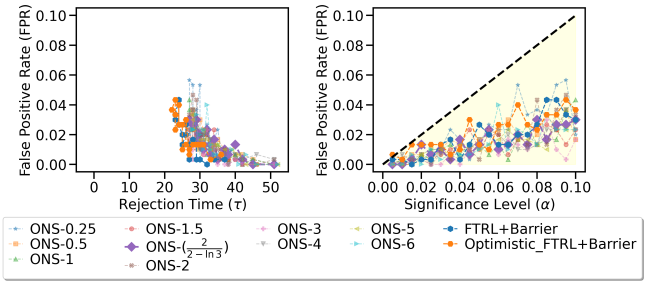


*Figure 9.* **(Tuning the Learning Rate in ONS.)** Comparisons of our proposed OAlg methods and ONS with different learning rates **(real-world dataset for detecting LLM-generated texts)**. The left subfigure illustrates the average time required to correctly identify texts generated by Gemini-1.5-Flash versus the average false positive rates (FPRs) over 300 runs under different significance levels ($\alpha$). The right subfigure evaluates the performance of each online algorithm under the null hypothesis, with a dashed line and shaded area representing the expected significance levels.