https://creativecommons.org/licenses/by/4.0/
http://jmlr.org/papers/v23/21-0354.html

$\mathcal{X}$

$\mathcal{A}$

$\mathcal{D}$ $\mathcal{X}$ $\mathcal{A}$

$\mathcal{X}$ $\mathcal{D}$

$\mathcal{D}$ $\mathcal{D}$

$\mathcal{D}$ $\mathcal{D}$ $\mathcal{D}$

$\mathcal{D}$ $\mathcal{D}$

$\mathcal{D}$ $\mathbb{1}$ $\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$ $\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$ $\mathcal{D}$

$\mathcal{D}$

$\mathcal{A}$

$\mathcal{A}$

$\mathcal{A}$ $\mathcal{X}$

$\mathcal{F}$ $\mathcal{X}$

$\mathcal{F}$

$\mathcal{F}$

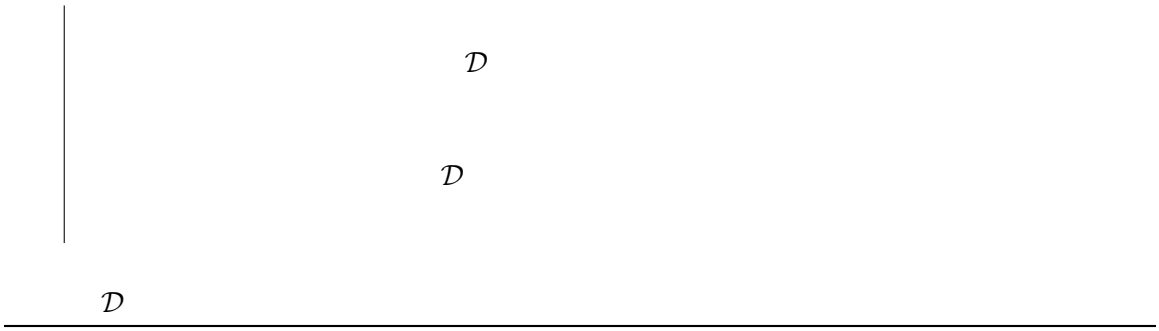
$\mathcal{F}$

$\mathcal{V}$ $\mathcal{V}$

$\mathcal{V}$ $\mathcal{V}$ $\mathcal{V}$

$\mathcal{V}$ $\mathcal{V}$

$\mathcal{F}$



$\mathcal{F}$

$\mathcal{F}$

$\mathcal{F}$

$\mathcal{F}$ $\mathcal{F}$

$\mathcal{F}$ $\sqrt{\mathcal{F}}$

$\mathcal{F}$ $\mathcal{X}$ $\mathcal{X}$ $\mathcal{F}$

$_{\mathcal{F}}$ $\mathcal{X}$

$\mathcal{F}$

$\sqrt{\mathcal{F}}$

$\mathcal{F}$

$\mathcal{V}$ $\mathcal{V}$ $\frac{\qquad}{\mathcal{F}}$ $\frac{-}{\mathcal{F}}$ $\frac{-}{}$ $\frac{}{\sqrt{\qquad}}$

$\mathcal{F}$ $\mathcal{X}$ $\mathcal{F}$

$\mathcal{F}$ $\mathcal{B}$

$\mathcal{F}$ $\mathcal{B}$

$\mathcal{B}$ $\mathcal{F}$

$\mathcal{X}$

$\mathcal{G}$

$\mathcal{G}$

$\mathcal{F}$

$-$

$-$

$\mathcal{G}$ $\mathcal{G}$

$\mathcal{G}$

$-$

$\mathcal{F}$

$\mathcal{F}$ $\mathcal{F}$ $\mathcal{F}$

$\mathcal{V}$ $\mathcal{V}$

$\mathcal{D}$ $\mathcal{D}$ $\mathcal{D}$

$\mathcal{V}$ $\sqrt{}$

$-$

$\mathcal{D}$ $\mathcal{D}$ $\mathcal{D}$

$\mathcal{V}$

$\mathcal{V}$

$\mathcal{V}$ $\sqrt{}$ $-$ $\sqrt{}$ $-$

$\sqrt{}$

$\mathcal{V}$ $\mathcal{V}$

$\mathcal{V}$

$\mathcal{D}$

$\mathcal{D}$ $\mathcal{D}$

$\mathcal{G}$

$\mathcal{D}$ $\mathcal{D}$

$\mathcal{G}$

$\mathcal{A}$

$\mathcal{A}$

$\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$ $\mathbb{1}$

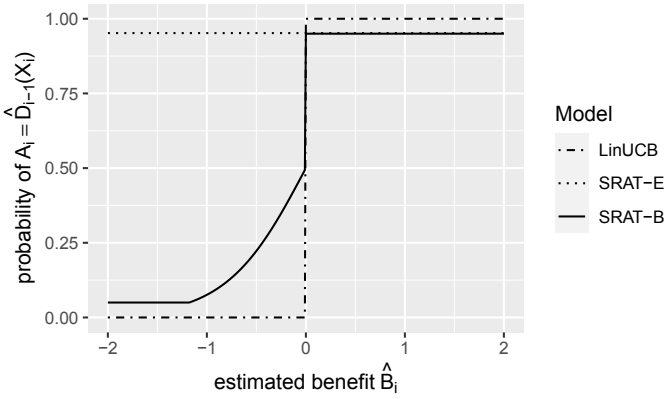
$\mathcal{D}$

$\mathcal{D}$

$\mathbb{1}$

$\mathcal{D}$

$\mathbb{1}$



$\mathcal{D}$

$\mathcal{D}$ $\mathcal{D}$

$\mathcal{D}$

$\mathcal{D}$

DTRlearn2

$\mathcal{D}$

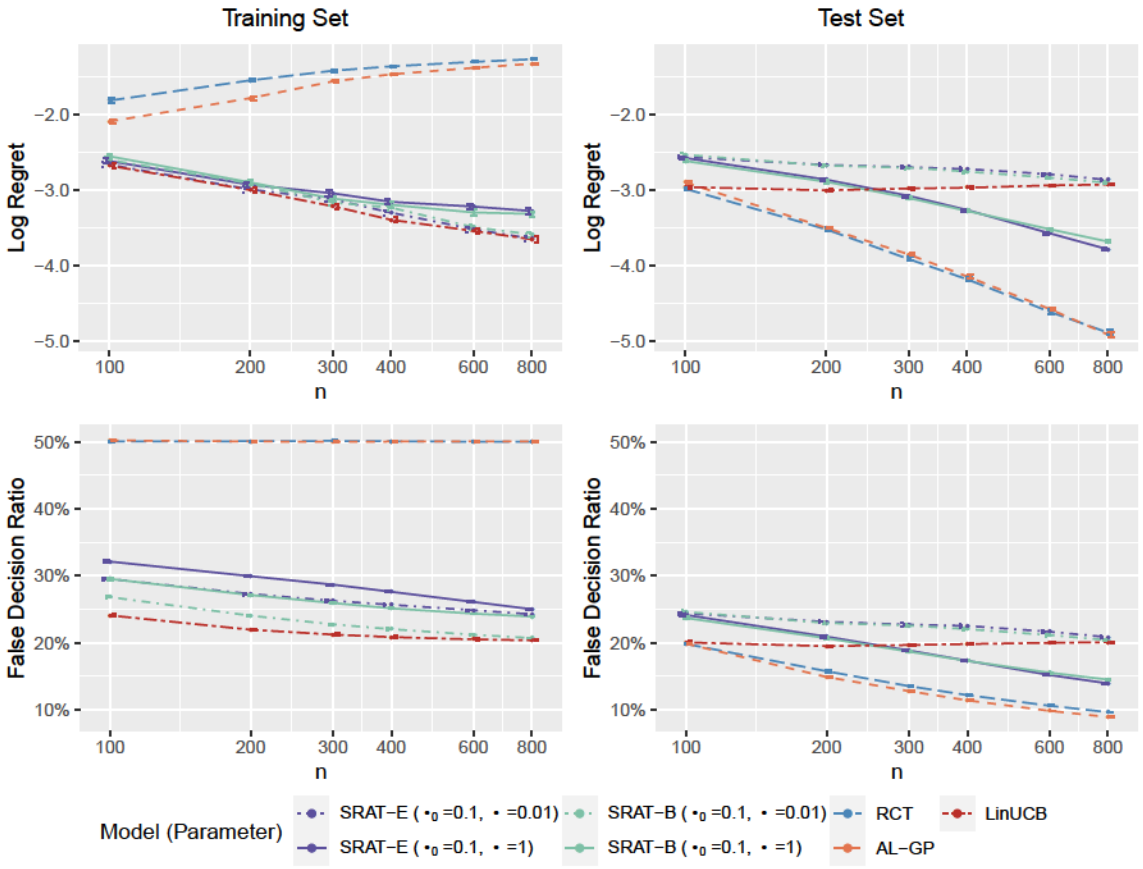


Figure 2: Scenario 1. The regret (logarithmic scale) and the false decision ratio on the training or test set against sample size n.

We have proved the properties of training and test regrets of SRAT in theoretical analysis and they will be used here as an indication of training and test performance of each algorithm. Each value function $\mathcal{V}$ is computed numerically using a sample of size $100,000$ randomly drew out of an independent population. The value function is estimated using the mean reward on this set.

We first compare the convergence rate of regret for different algorithms. SRAT-E and SRAT-B are implemented with $\epsilon_0 = 0.1$ and $\theta = 0.01$ or 1. As will be discussed later in Figure 4, the training and test regrets are monotone in the parameters ϵ_0 and θ. Therefore, to save space, we only show two possible combinations of parameters here. The scheduling parameter γ_i for SRAT-B is taken as 0.999^i so that it will not decay too fast to zero. RCT is a special case of SRAT with $\epsilon_0 = 0.5$ and $\theta = 1$. According to Li et al. (2010), the click-through rate (mean reward) of LinUCB in news article recommendation does not change much on the deployment bucket (test set) when $\alpha \geq 0.2$, while it decreases quickly on the learning bucket (training set) as α increases from 0.2. In our experiment settings, α does not affect training and test regrets significantly. Therefore, we will fix $\alpha_i = 0.2$ for all i for LinUCB and SRAT-B in our following experiments. The process is repeated $1,000$

Figure 3: The weighted sum of training and test regrets in scenario 1 when $n = 800$.

times and the resulting values are averaged across all iterations. To better illustrate the polynomial relationship between training or test regret and the sample size n, we plot the regret values and the sample sizes on the logarithmic scale. The false decision ratio, or $1 -$ accuracy in classification literature, is also displayed against n. One standard error of the mean regret or the mean false decision ratio across the $1,000$ iterations is reported on each point. The result of scenario 1 is plotted in Figure 2. The plot of scenario 2, Figure 7, is included in the supplementary material since it shows a similar conclusion as scenario 1.

According to Figure 2, LinUCB is the greediest on the training process, with the least regret and false decision ratio. As discussed in Section 4, LinUCB can actually be viewed as a limiting case of SRAT-B on the training set. Indeed, our proposed greediest algorithms, SRAT-E and SRAT-B with parameters $\epsilon_0 = 0.1$ and $\theta = 0.01$, perform similarly as LinUCB in terms of training regret. AL-GP and RCT take purely randomized treatments on the training set, so they have the largest training regret and a 50% training accuracy. Since the training regret is calculated based on $\mathcal{V}(\hat{f}_{n-1})$ which is increasing as n grows, the training regret actually increases for largely randomized methods. In theory, the training regret of RCT is bounded by a constant that does not rely on n when the ϵ-sequence is constant. SRAT-E and SRAT-B perform similarly in terms of regrets on both training and test sets, but SRAT-B has a lower false decision ratio on the training set. The logarithms of their training and test regrets are approximately linear in $\log n$, which is consistent with our theory.

On the test set, AL-GP and RCT perform the best due to their full exploration in the training process. LinUCB needs to fit the regression model of rewards and thus relies on both the main effect and the treatment effect model. In addition, to estimate the upper confidence bound, it needs an assumption on the inference model. With these limitations, the regret or false decision ratio of LinUCB on the test set does not decrease. When n is small, the final ITR estimated by LinUCB can sometimes be optimal since the true ITR is linear. However, the ITR converges to the projection onto that of the linear total reward space when n is large and thus the average regret gets pulled up. On the other hand,

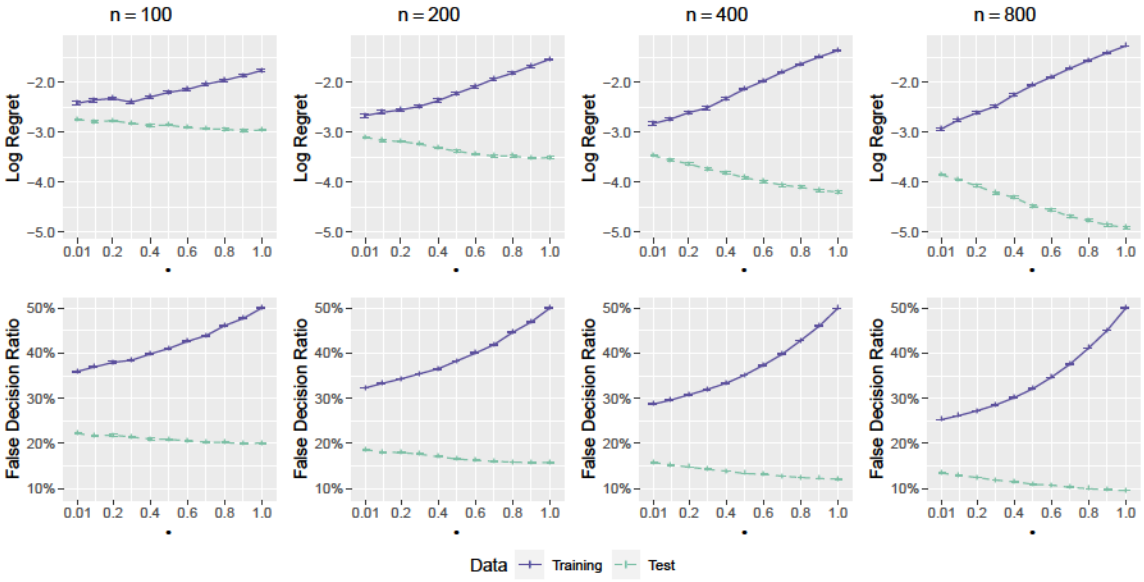


Figure 4: Scenario 1 with $\epsilon_0 = 0.5$. The regret (logarithmic scale) and the false decision ratio on the training or test set against parameter θ.

OWL tries to find the decision function that maximizes the reward directly. It only requires a correct model of the treatment effect for consistency, without any assumption on the main effect or the distribution of the error term. Therefore, SRATs with $\epsilon_0 = 0.1, \theta = 1$ outperform LinUCB on the test set when n is larger than 200.

We plot a weighted sum of training and test regrets in Figure 3 to show their balance. Specifically, the weighted sum is defined as

$$\lambda \text{Regret}_{test} + (1-\lambda)\text{Regret}_{train} = \lambda \frac{1}{n} \sum_{i=1}^{n} [\mathcal{V}(\hat{f}_{i-1}) - R_i] + (1-\lambda)[\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n)]$$

for $\lambda \in [0, 1]$, so that it equals the training regret when $\lambda = 0$ and equals the test regret when $\lambda = 1$. The sample size is fixed at 800. The initial value of truncation parameter ϵ_0 equals 0.1 and the decay parameter θ takes values in $0.01, 1$ for SRAT-E and SRAT-B. The plot shows that we should choose LinUCB when we consider the training regret only, and should choose AL-GP or RCT when we consider the test regret only. However, if we want to consider the performance on both the training and the test sets, we should choose SRAT-E or SRAT-B with $\theta = 1$.

The change of SRAT-E with different parameters θ and sample size n is demonstrated in Figure 4 for scenario 1. Since SRAT-B performs quite similarly to SRAT-E as shown in Figures 2 and 7, we omit it here to save space. The parameter θ can take values from $0.01, 0.1, 0.2, \ldots, 1$ and n can take values from $100, 200, 400, 800$. Note that only when $\epsilon_0 = 0.5$ and $\theta = 1$, our algorithm represents pure RCT. Thus we only illustrate our findings with $\epsilon_0 = 0.5$ here. Other ϵ_0's give similar conclusion, and smaller ϵ_0 means better training performance and worse test performance. The values and standard errors of the

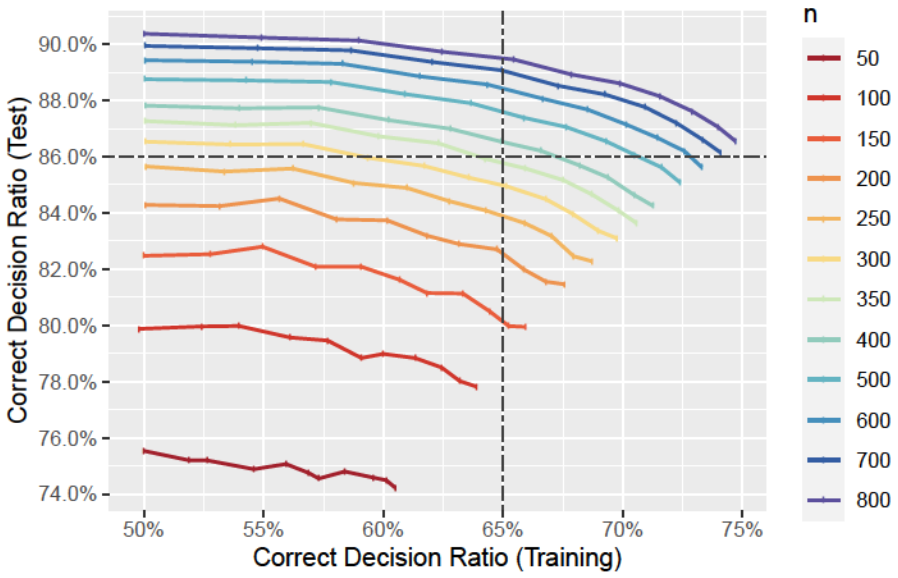


Figure 5: Sample size consideration for SRAT-E in scenario 1 with $\epsilon_0 = 0.5$. Correct decision ratios on the test set against that on the training set. Each line represents a sample size n and each point on the line represents a value of θ. Points to the right correspond to smaller θ, and thus lead to higher correct decision ratio on the training set and lower ratio on the test set.

mean regret and mean false decision ratio are shown. For all sample sizes, the plots clearly show the tradeoff between training and test performance. Note that when θ increases, ϵ_i increases for all i and the treatments are more randomized in the training process. While the training regret increases with more randomization, the test regret decreases. The false decision ratio shows a similar tendency. All the points with $\theta = 1$ have an accuracy of 50% on the training set, which indeed illustrates the pure randomization. In accordance with the theory, the logarithm of training and test regrets are approximately linear in θ. In practice, the training regret is more affected than the test regret by θ. As shown in Figure 4, when $n = 800$, the training regret increases by $e^{-1.27} - e^{-2.93} = 0.227$ while the test regret decreases by $e^{-3.85} - e^{-4.91} = 0.014$ when θ increases from 0.01 to 1.

Using this simulation example, we can also illustrate how to find the sample size needed for a clinical trial of certain purposes. Given different requirements for the trial and the population, we need different sample sizes. Here we illustrate the situation when the proportion of patients assigned the better treatments is required to reach a certain level in Figure 5 for SRAT-E in scenario 1. Note that the variation trends of correct decision ratios against θ are opposite for the training and test data. In particular, θ should be small enough so that the decision process is greedy on the training set, and in the meanwhile it should be large enough so that the final ITR is efficient on the test set. It is clear that the two accuracies are negatively correlated. For example, when we need the training ratio to be greater than 65%, $\theta \leq 0.1$ for $n = 150$, $\theta \leq 0.2$ for $n = 200$, $\theta \leq 0.3$ for $n = 250$, $\theta \leq 0.4$

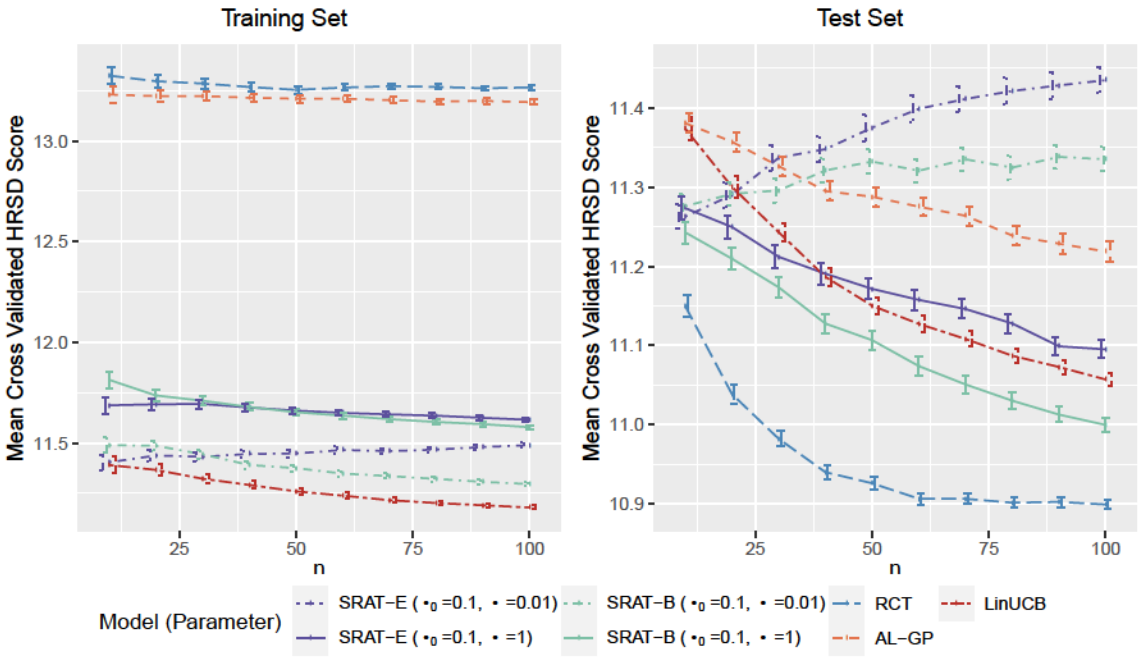


Figure 6: Mean cross-validated HRSD scores against the sample size n.

To simulate an adaptive clinical trial, we first generate a treatment suggestion based on the tailoring variables of the next patient using our algorithm. If the actual treatment taken is consistent with our suggestion, we take down the whole record of this patient, including feature variables, the treatment and the reward; otherwise, we drop this record and move on to the next. Note that the first n_0 suggestions are given with equal probabilities on each treatment. Five-fold cross validation is used here to avoid overfitting. Specifically, the data set is partitioned into five parts randomly. Four of the five parts are used iteratively as training data to apply our algorithm in generating the treatment suggestion. The last part is used as the test set to evaluate the ITR. The performance on the test data is evaluated using an unbiased estimator of the value function $\mathcal{V}(f)$ (Qian and Murphy, 2011; Minsker et al., 2016)

$$\sum_{i=1}^{n} \frac{R_i \mathbb{1}\left[A_i = \text{sign}\{f(\boldsymbol{X}_i)\}\right]}{\pi_i(A_i; X_i)} \Bigg/ \sum_{i=1}^{n} \frac{\mathbb{1}\left[A_i = \text{sign}\{f(\boldsymbol{X}_i)\}\right]}{\pi_i(A_i; X_i)} .$$

Here the rewards R_i's are defined as the negative HRSD scores.

The initial sample size n_0 is fixed at 50. The recruitment stops when the sample size n reaches 100, or the training data run out. We average the mean reward on each test fold for $n = 10, 20, \ldots, 100$. The process is repeated 1,000 times. Finally, the means and standard errors of means across all iterations are reported. From Section 5, we know that the training and test values are monotone in ϵ_0 and θ. Therefore, we only demonstrate the situation when $\epsilon_0 = 0.1$ and $\theta = 0.01, 1$. The contextual bandit algorithm LinUCB and the active clinical trial method AL-GP are also compared here. Figure 6 displays the negative mean rewards, that is, the mean cross validated HRSD scores, against the sample size n. Lower scores are more satisfactory.

$\mathcal{X}$

$\mathcal{F}$ $\mathcal{X}$

$\mathcal{F}$

$\mathcal{F}$

$-$ $\mathcal{F}$

$\mathcal{Z}$

$\mathcal{Z}$

$\mathcal{Z}$

$\mathcal{F}$ $\mathcal{Z}$

$\mathcal{Z}$

$\mathcal{R}$ $\mathcal{F}$ $-$

$\mathcal{F}$

$\mathcal{R}$ $\mathcal{F}$ $\mathcal{R}$ $\mathcal{F}$

$-$ $\mathcal{R}$ $\mathcal{F}$ $=$

$\mathcal{F}$ $\mathcal{F}$

$\mathcal{Z}$ $\mathcal{F}$ $\mathcal{R}$ $\mathcal{F}$

$\mathcal{F}$

$\mathcal{F}$ $\mathcal{Z}$

$\mathcal{F}$

$\bar{\mathcal{F}}$

$\mathcal{N}$ $\mathcal{F}$ $\mathcal{F}$

$\mathcal{F}$ $\mathcal{N}$ $\mathcal{F}$

$\mathcal{N}$ $\mathcal{F}$

$\mathcal{F}$ $\mathcal{Z}$

$\bar{\mathcal{F}}$ $\overline{\mathcal{R}\ \mathcal{F}}$

$\mathcal{R}$ $\mathcal{F}$ $\mathcal{N}$ $\mathcal{F}$

$\mathcal{N}$ $\mathcal{F}$

$\mathcal{R}$ $\mathcal{F}$

$\mathcal{F}$

$\mathcal{G}$

$\bar{\ }$

$\mathcal{G}$

$\bar{\ }$

$$\sqrt{\qquad}$$

$\mathcal{X}$ $\mathcal{G}$

$\mathcal{F}$ $\mathcal{X}$

$\mathcal{F}$ $\mathcal{F}$ $\mathcal{F}$ $\sqrt{\qquad}$ $\sqrt{\mathcal{F}}$

$\mathcal{G}$

$\mathcal{G}$

$\mathcal{G}$

$\mathcal{G}$

$$\frac{}{}$$

$\mathcal{G}$

■

$\mathcal{G}$

$\mathcal{X}$

$\mathcal{G}$

$\mathcal{G}$

$\mathcal{G}$

$\mathbb{1}$ $\mathbb{1}$ $\mathcal{G}$ $\mathcal{G}$

$\mathbb{1}$

$\mathcal{G}$

$\mathcal{G}$

$\mathbb{1}$ $\mathbb{1}$ $\mathcal{G}$ $\mathcal{G}$

$\mathcal{G}$

■

$\sqrt{\mathcal{F}}$ $\sqrt{\mathcal{F}}$

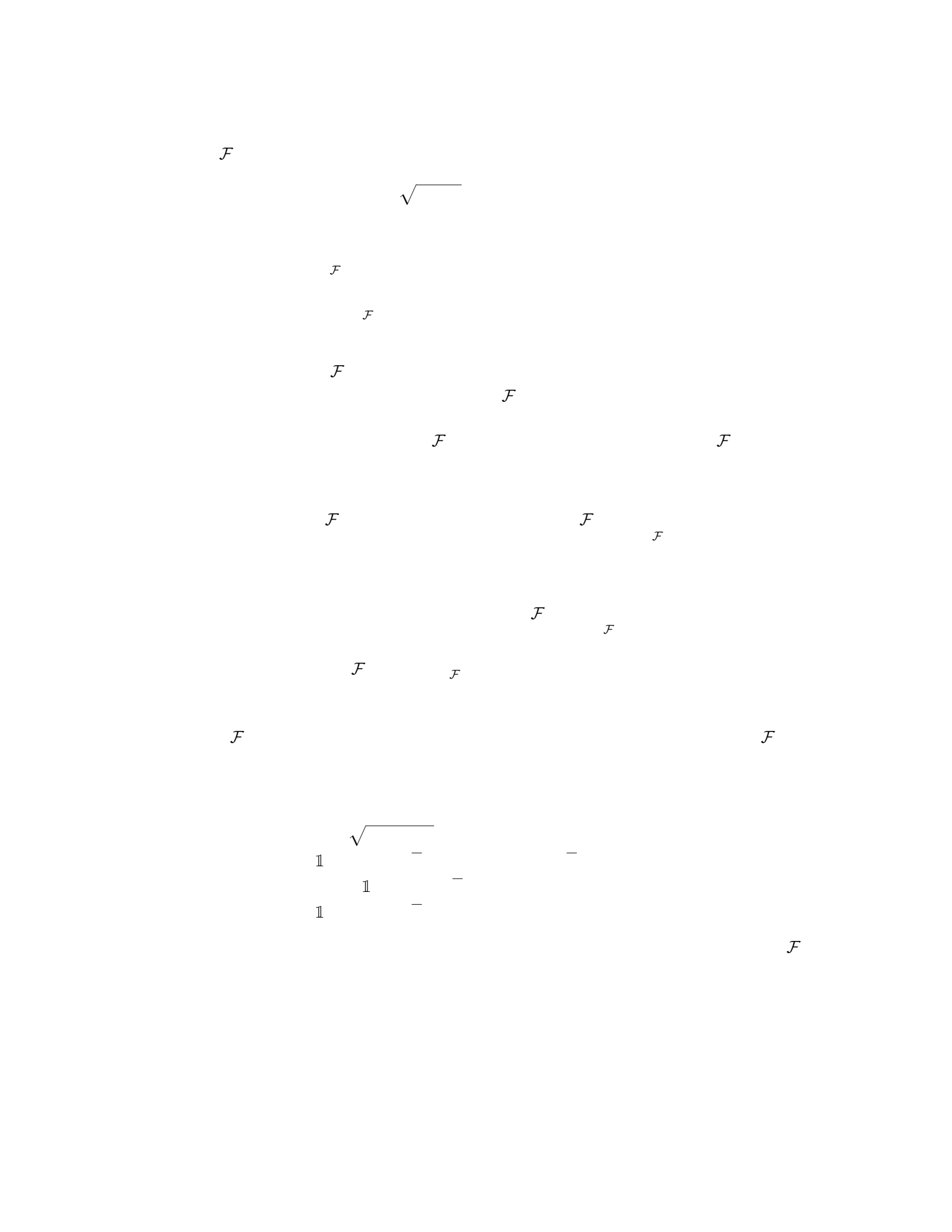
$\mathcal{F}$

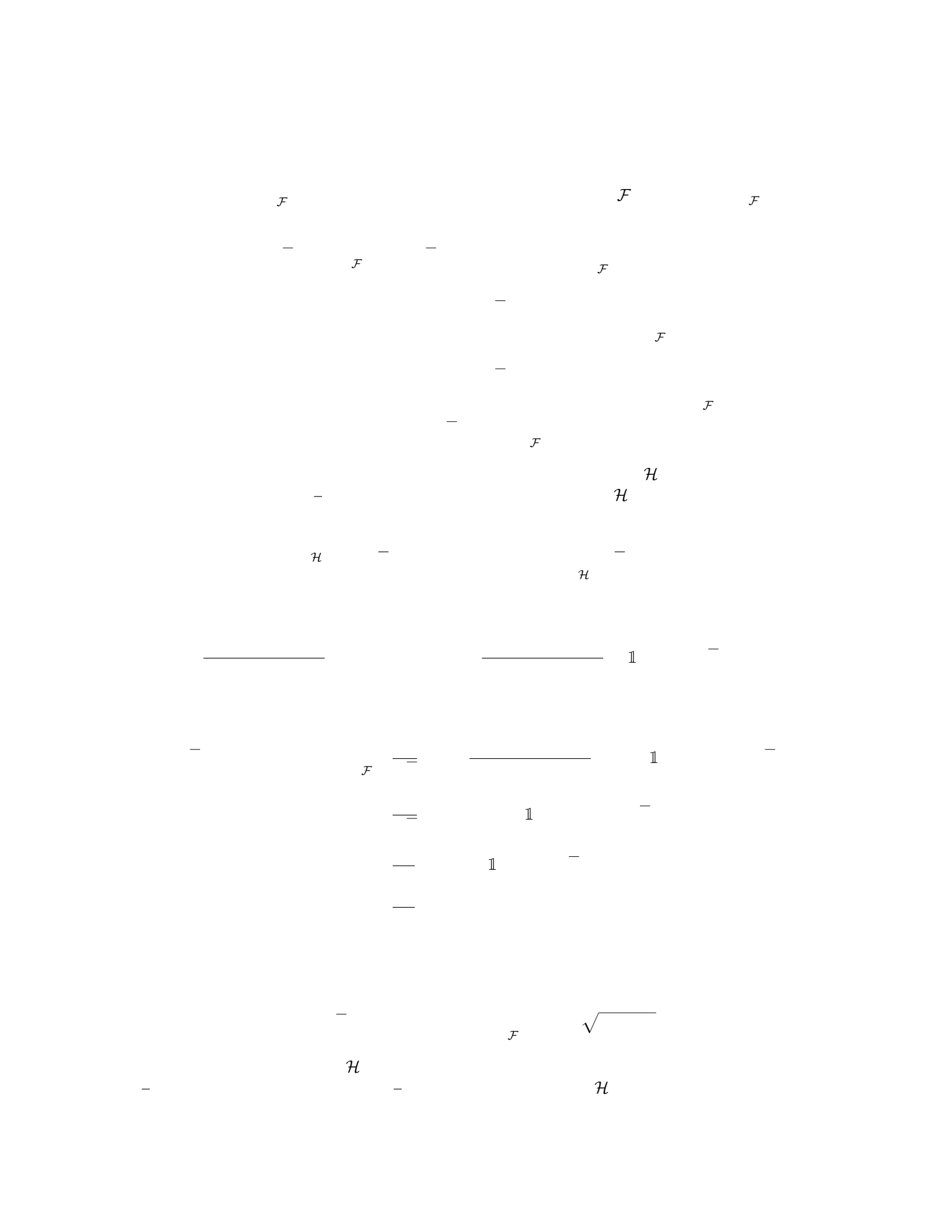
$\mathcal{F}$ $\mathcal{F}$

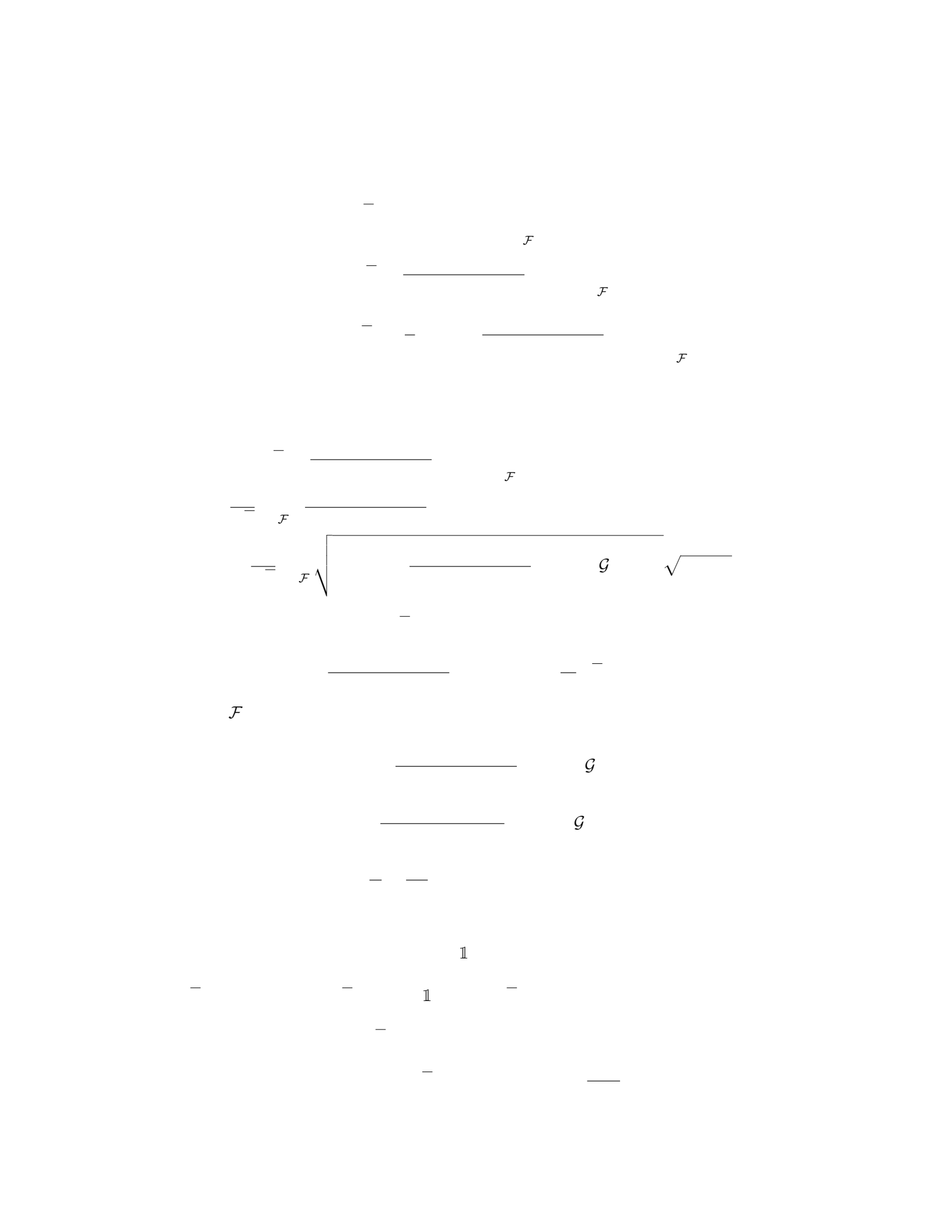
$\mathcal{F}$

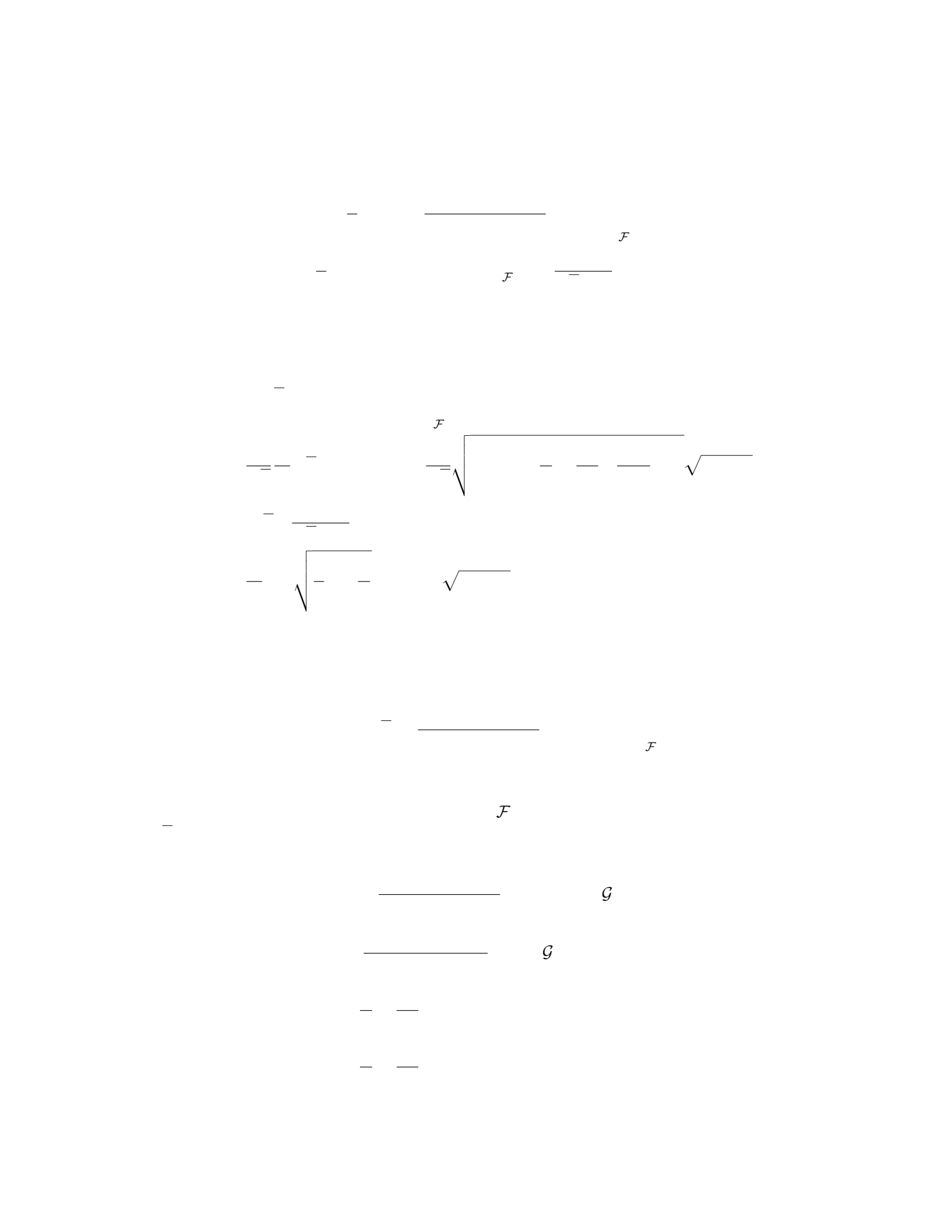
$\mathcal{F}$ $\mathcal{F}$

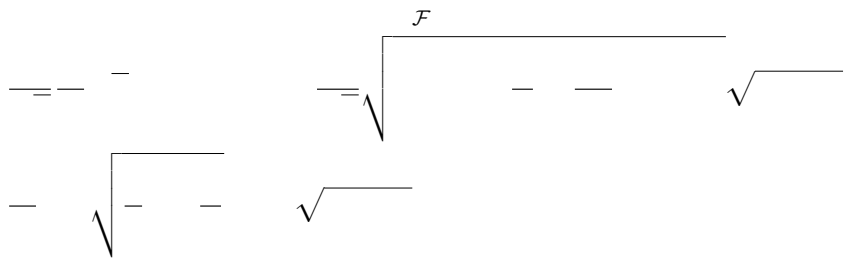
$\mathcal{F}$ $\mathcal{F}$ $\sqrt{}$ $\sqrt{}$

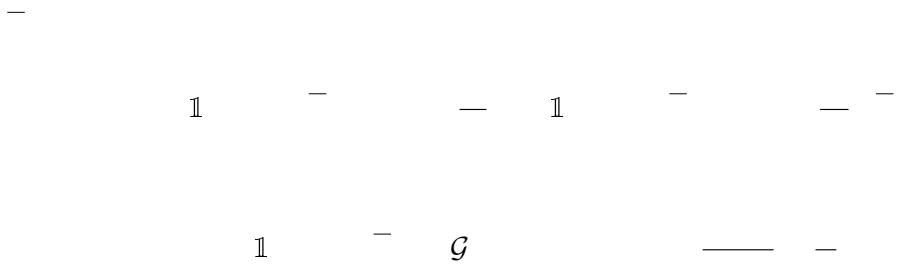


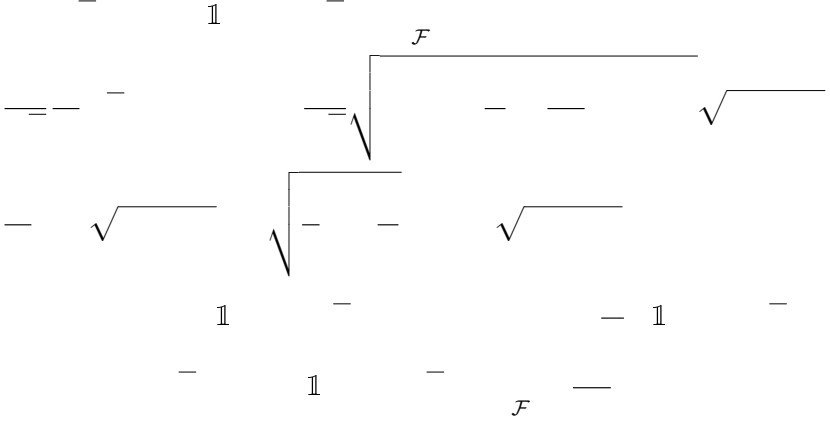


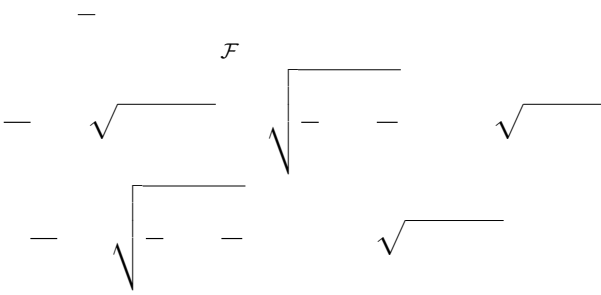












$\mathcal{F}$

$$- \qquad _{\mathcal{F}} \qquad - \qquad \sqrt{}^{-} \qquad - \qquad \sqrt{}$$

$$- \qquad \mathcal{F}$$

■

$\mathcal{V}$ $\mathcal{V}$

$\mathcal{G}$

$\mathcal{G}$ $\mathcal{G}$

$\mathbb{1}$ $\mathcal{G}$ $\mathbb{1}$ $\mathcal{G}$

$\mathcal{G}$

$\mathcal{V}$ $\mathcal{V}$

$\mathcal{V}$ $\mathcal{V}$

$\mathcal{F}$

$\mathcal{R}$ $\mathcal{F}$ $\mathcal{F}$

$\mathcal{F}$

$\mathcal{F}$

$\mathcal{H}$ $\mathcal{F}$ $\mathcal{F}$

$\mathcal{R}$ $\mathcal{H}$ $\mathcal{F}$ $\mathcal{F}$

$\mathcal{Z}$ $\mathcal{F}$

$\mathcal{R}$ $\mathcal{H}$ $\mathcal{F}$ $\mathcal{F}$

$\mathcal{F}$

$\mathcal{F}$

$\sqrt{}$

$\mathcal{F}$

$\mathcal{F}$

$\mathcal{F}$

$\mathcal{F}$

—

— $\sqrt{}$

$\mathcal{F}$

$\mathcal{F}$ $\overline{}$

$\overline{}$ — — $\overline{}\sqrt{}$

■

– $\mathcal{V}$ – – $\mathcal{V}$

$\mathcal{G}$ $\mathcal{G}$ $\mathcal{V}$

$\mathcal{G}$ $\mathcal{G}$ $\mathcal{G}$ $\mathcal{V}$ $\mathcal{G}$ $\mathcal{V}$

– $\mathcal{G}$ $\mathcal{V}$ $\mathcal{V}$ —

– – $\overline{}$

■

ν

$-$ $-$ ν $-$ ν ν

ν

ν

■

$\sqrt{}$

$-$ $-$

$-$ $-$

■

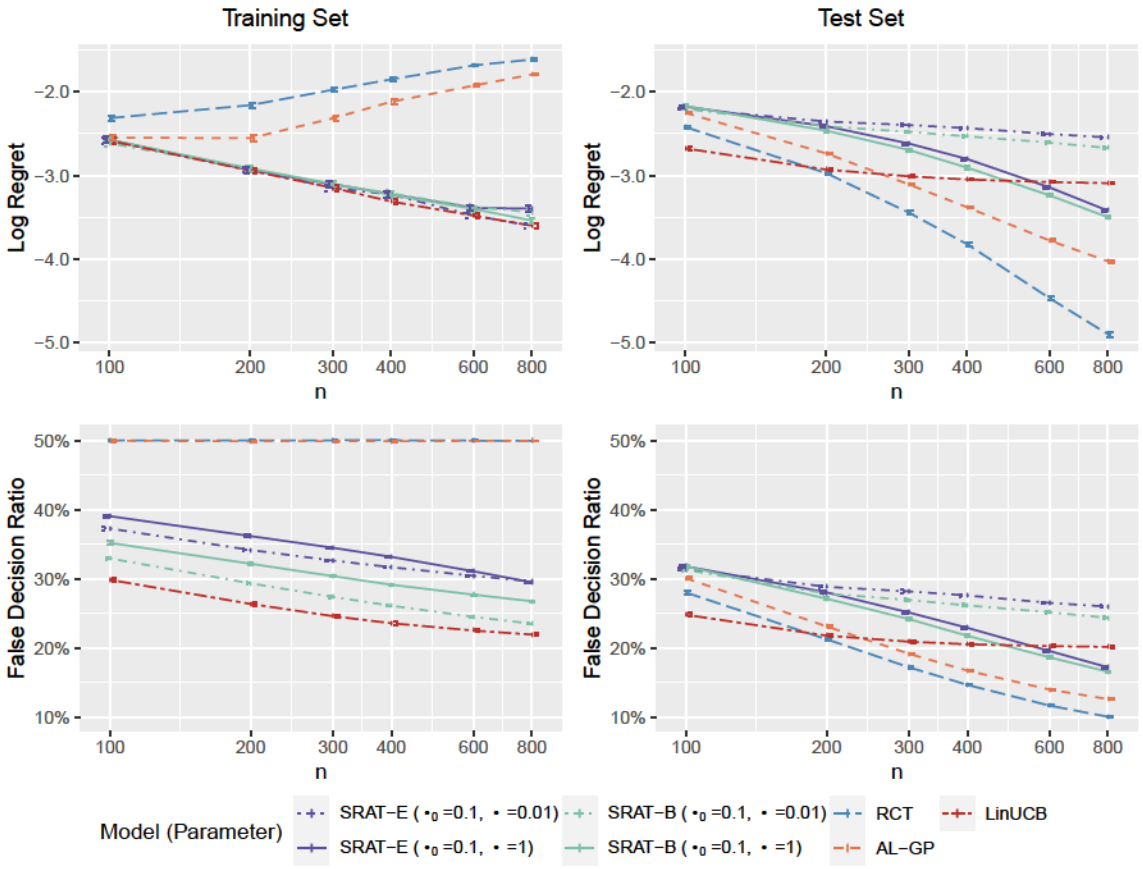


Figure 7: Scenario 2. The regret (logarithmic scale) and the false decision ratio on the training or test set against sample size n.

References

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Jongsig Bae and Shlomo Levental. Uniform CLT for Markov chains and its invariance principle: a martingale approach. *Journal of Theoretical Probability*, 8(3):549–570, 1995.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

https://CRAN.
R-project.org/package=DTRlearn2