Fast Stochastic AUC Maximization with $O(1/n)$ Convergence Rate

Mingrui Liu†, Xiaoxuan Zhang†, Zaiyi Chen‡, Xiaoyu Wang♮, Tianbao Yang†

†Department of Computer Science, The University of Iowa, USA ‡The University of Science and Technology of China, China ♮ Intellifusion

Problem of Interest

1. **Problem**: We consider statistical learning with AUC (Area under the ROC curve) maximization where one random data is received at one iteration for updating the model

2. **Definition**: $\text{AUC}(h) = \Pr(h(\mathbf{x}) \geq h(\mathbf{x}')|y = 1, y' = -1)$, where $(\mathbf{x}, y), (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^d \times \{1, -1\}$, $h : \mathbb{R}^d \to \mathbb{R}$ is score function

3. **Challenges**: Online AUC Optimization is challenging since AUC loss depends on pairs of examples. The pairwise nature in the definition of AUC makes it difficult to design algorithms suitable for the classical stochastic or online setting

4. **Main Contribution**: Building on a saddle point formulation of AUC, we developed a novel stochastic (online) algorithm with $O(1/n)$ convergence rate for AUC maximization, where n is the number of received examples (this is the first result of fast rate for AUC)

Related Work

1. Zhao et al. [1] designed an online algorithm which keeps a large buffer (with size $O(\sqrt{n})$) and then utilizes the data in the buffer to update the classifier

2. Gao et al. [2] designed an online algorithm which needs to keep the first and second order statistics

3. Ying et al. [3] reformulates the original problem into a **convex-concave saddle point problem**, and then utilize the standard stochastic gradient method to solve it

Saddle Point Reformulation of AUC Maximization [3]:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ (a,b)\in\mathbb{R}^2}} \max_{\alpha\in\mathbb{R}} \{f(\mathbf{w}, a, b, \alpha) := \mathbb{E}_{\mathbf{z}}[F(\mathbf{w}, a, b, \alpha; \mathbf{z})]\},$$

where $\mathbf{z} = (\mathbf{x}, y)$, $F(\mathbf{w}, a, b, \alpha; \mathbf{z}) = (1-p)(\mathbf{w}^\top\mathbf{x} - a)^2\mathbb{I}_{[y=1]} + p(\mathbf{w}^\top\mathbf{x} - b)^2\mathbb{I}_{[y=-1]} - p(1-p)\alpha^2 + 2(1+\alpha)(p\mathbf{w}^\top\mathbf{x}\mathbb{I}_{[y=-1]} - (1-p)\mathbf{w}^\top\mathbf{x}\mathbb{I}_{[y=1]})$

Key Observation

Define $\mathbf{v} = (\mathbf{w}, a, b)$, assume $\sup_{\mathbf{x}} \|\mathbf{x}\|_2 \leq \kappa, \Omega_1 = \{(\mathbf{w}, a, b) : \|\mathbf{w}\|_1 \leq R, |a| \leq R\kappa, |b| \leq R\kappa\}, \Omega_2 = \{\alpha \in \mathbb{R} : |\alpha| \leq 2R\kappa\}$. $f_1(\mathbf{v}) = \max_{\alpha\in\Omega_2} f(\mathbf{v}, \alpha)$ restricted on the set Ω_1 satisfies a quadratic growth condition, i.e., for any $\mathbf{v} \in \Omega_1$, there exists $c > 0$ such that

$$\|\mathbf{v} - \mathbf{v}_*\|_2 \leq c(f_1(\mathbf{v}) - \min_{\mathbf{v}\in\Omega_1} f_1(\mathbf{v}))^{1/2}$$

Algorithm and Theoretical Results

1. Standard Primal-Dual Stochastic Gradient Algorithm (PDSG)

Algorithm 1 PDSG$(\mathbf{v}_1, \alpha_1, r, D, T, \eta)$

```
Initialize variables $\widehat{A}_+ \in \mathbb{R}^{d+2}, \widehat{A}_- \in \mathbb{R}^{d+2}$, $T_+, T_-, \widehat{p} \in \mathbb{R}$ as zeros
for $t = 1, \ldots, T$ do
    Receive a sample $\mathbf{z}_t = (\mathbf{x}_t, y_t)$
    Update $\widehat{A}_\pm, T_\pm, \widehat{p}$ using the data $\mathbf{z}_t$
    $\mathbf{v}_{t+1} = \Pi_{\Omega_1\cap\mathcal{B}(\mathbf{v}_1,r)}(\mathbf{v}_t - \eta\partial_{\mathbf{v}}\widehat{F}_t(\mathbf{v}_t, \alpha_t, \mathbf{z}_t))$
    $\alpha_{t+1} = \Pi_{\Omega_2\cap\mathcal{B}(\alpha_1,D)}(\alpha_t + \eta\partial_\alpha\widehat{F}_t(\mathbf{v}_t, \alpha_t, \mathbf{z}_t))$
end for
Compute $\bar{\mathbf{v}}_T = \frac{\sum_{t=1}^T \mathbf{v}_t}{T}$ and $\widehat{\alpha} = (\frac{\widehat{A}_-}{T_-} - \frac{\widehat{A}_+}{T_+})^\top\bar{\mathbf{v}}_T$
{The closed form of $\alpha$ is $\alpha = \mathbf{w}^\top[\mathbb{E}(\mathbf{x}|y = -1) - \mathbb{E}(\mathbf{x}|y = 1)]$}
Let $r = r/2$, update $\beta, D$
return $(\bar{\mathbf{v}}_T, \widehat{\alpha}, \beta, r, D)$
```

Remark: $A_\pm$ store the summation of feature vectors for positive (negative) examples, $T_\pm$ stand for the number of received positive (negative) examples until the current iteration

Theoretical Gurantee of PDSG

Suppose $\|\mathbf{v}_1 - \mathbf{v}_*\|_2 \leq r$, where $\mathbf{v}_* \in \Omega_1$ is the optimal solution closest to $\mathbf{v}_1$, run the PDSG for T iterations. Then with probability at least $1-\delta$,

$$\max_{\alpha\in\Omega_2} f(\bar{\mathbf{v}}_T, \alpha) - \min_{\mathbf{v}\in\Omega_1}\max_{\alpha\in\Omega_2} f(\mathbf{v}, \alpha) \leq O\left(\frac{\ln(T/\delta)}{\sqrt{T}}\right)$$

2. Motivated by Juditsky and Nesterov [4], we design Fast Stochastic AUC Maximization Algorithm (FSAUC)

Algorithm 2 FSAUC

```
Set $m = \lfloor\frac{1}{2}\log_2\frac{2n}{\log_2 n}\rfloor - 1$, $n_0 = \lfloor n/m\rfloor$, $R_0 = 2\sqrt{1+2\kappa^2}R$, $\beta_0 = 1+8\kappa^2$
Initialize $\widehat{\mathbf{v}}_0 = 0 \in \mathbb{R}^{d+2}$, $\widehat{\alpha}_0 = 0$,
for $k = 1, \ldots, m$ do
    Set $\eta_k = \frac{\sqrt{\beta_{k-1}}}{\sqrt{3n_0G}}R_{k-1}$
    $(\widehat{\mathbf{v}}_k, \widehat{\alpha}_k, \beta_k, R_k, D_k)$ = PDSG$(\widehat{\mathbf{v}}_{k-1}, \widehat{\alpha}_{k-1}, R_{k-1}, D_{k-1}, n_0, \eta_k)$
end for
return $\widehat{\mathbf{v}}_m$
```

Main Theorem

Theoretical Gurantee of FSAUC

When $n > \max(100, m\frac{32\ln(\frac{12}{\delta})}{(\min(p,1-p))^2})$, then with probability at least $1-\delta$,

$$\max_{\alpha\in\Omega_2} f(\widehat{\mathbf{v}}_m, \alpha) - \min_{\mathbf{v}\in\Omega_1}\max_{\alpha\in\Omega_2} f(\mathbf{v}, \alpha) \leq \tilde{O}\left(\ln(\frac{1}{\delta})/n\right)$$

Experimental Results

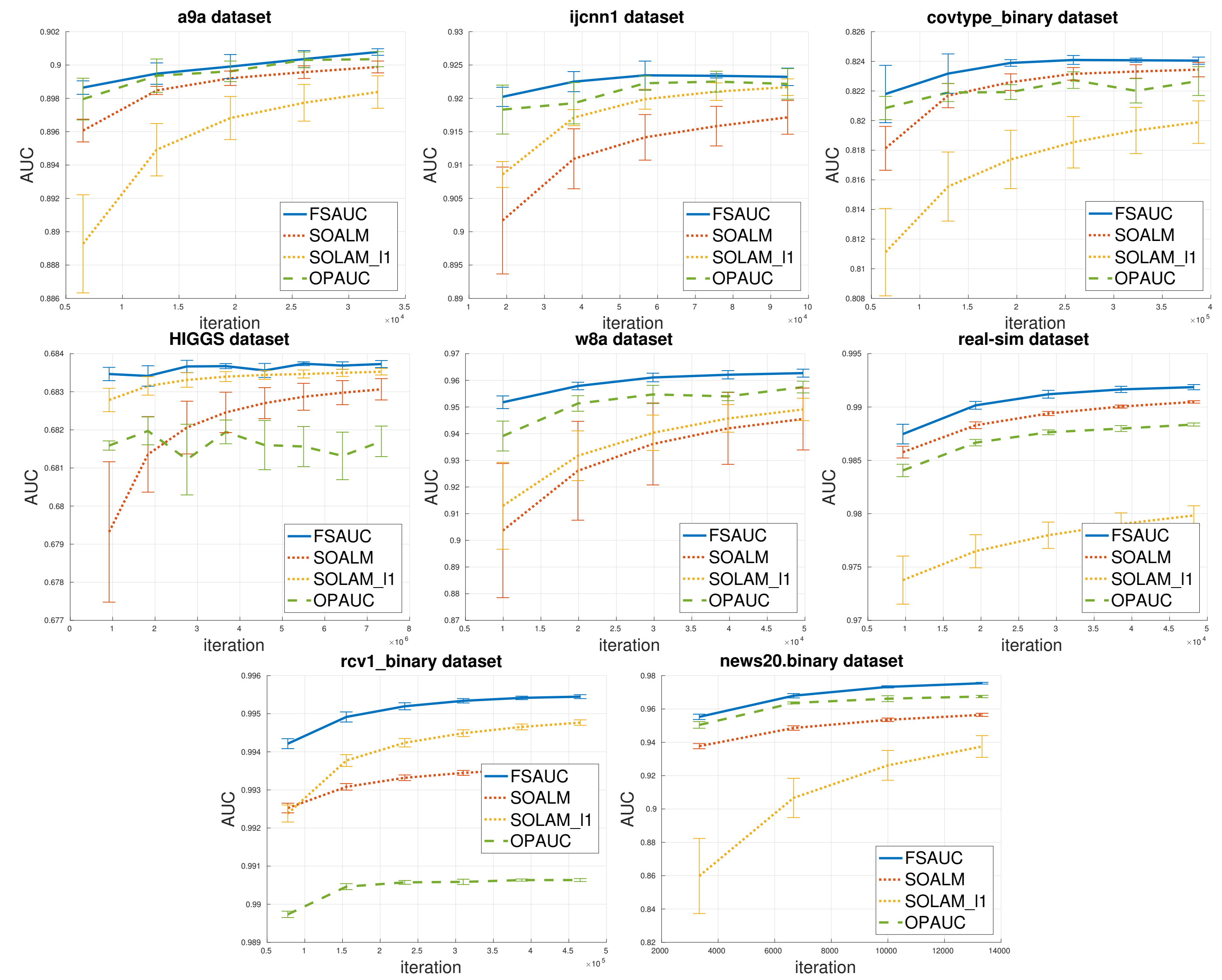

Figure 1: AUC-Iteration curves of FSAUC and the baselines

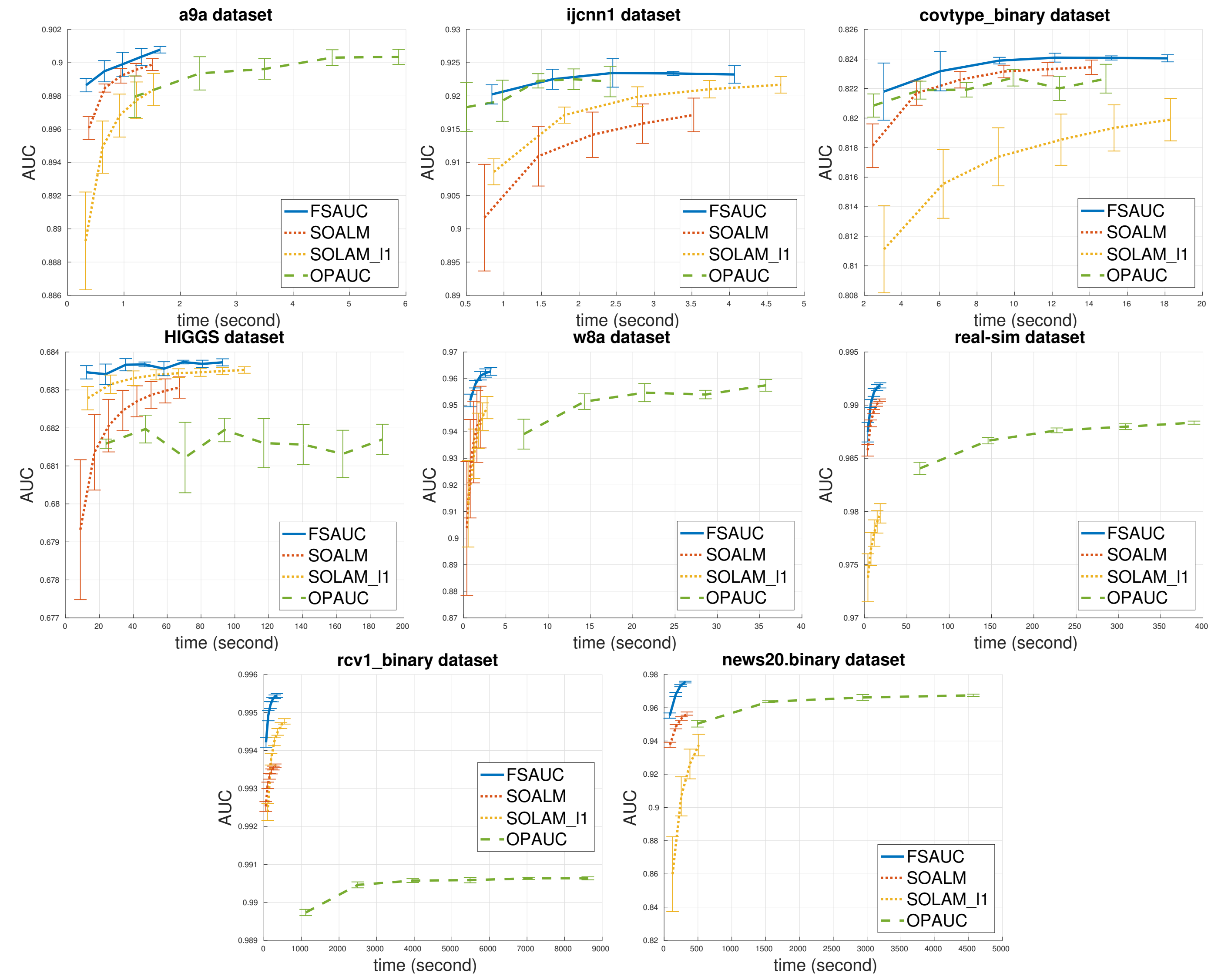

Figure 2: AUC-Time curves of FSAUC and the baselines

References:

[1] P. Zhao, R. Jin, T. Yang, and S. C. Hoi. Online AUC maximization. ICML 2011.

[2] W. Gao, R. Jin, S. Zhu, and Z. Zhou. One-pass AUC optimization. ICML 2013.

[3] Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. NIPS 2016.

[4] A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. Stochastic Systems, 2014.