A. Proofs

A.1. Proof of Prop. 1

Proof. For a discrete state space, the total variation is equivalent to half of L_1 distance between two probability vectors. Denote by $\hat{p}(X=i|\boldsymbol{\varepsilon})$ the distribution of the output of the approximate algorithm conditioned on the vector of Gumbel variables $\boldsymbol{\varepsilon}$, and $x(\boldsymbol{\varepsilon})$ the solution of Eq. 2 as a function of $\boldsymbol{\varepsilon}$. According to the premise of Prop. 1, $\hat{p}(X=x(\boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}) \geq 1-\delta, \forall \boldsymbol{\varepsilon}$. We can bound the L_1 error of the conditional probability as

$$\begin{aligned}&\sum_{i\in\mathcal{X}} \left|\hat{p}(X=i|\boldsymbol{\varepsilon}) - \delta_{i,x(\boldsymbol{\varepsilon})}\right| \\ &= |\hat{p}(X=x(\boldsymbol{\varepsilon})|\boldsymbol{\varepsilon}) - 1| + \sum_{i\neq x(\boldsymbol{\varepsilon})} |\hat{p}(X=i|\boldsymbol{\varepsilon})| \leq 2\delta, \forall\boldsymbol{\varepsilon}\end{aligned} \tag{13}$$

where $\delta_{i,j}$ is the Kronecker delta function. Then we can show

$$\begin{aligned}&\|\hat{p}(X) - p(X)\|_{\mathrm{TV}} \\ &= \frac{1}{2}\sum_{i\in\mathcal{X}} |\tilde{p}(X=i) - p(X=i)| \\ &= \frac{1}{2}\sum_{i\in\mathcal{X}} \left|\int_{\boldsymbol{\varepsilon}} \left(\hat{p}(X=i|\boldsymbol{\varepsilon}) - \delta_{i,x(\boldsymbol{\varepsilon})}\right) \mathrm{d}P(\boldsymbol{\varepsilon})\right| \\ &\leq \frac{1}{2}\sum_{i\in\mathcal{X}} \int_{\boldsymbol{\varepsilon}} \left|\hat{p}(X=i|\boldsymbol{\varepsilon}) - \delta_{i,x(\boldsymbol{\varepsilon})}\right| \mathrm{d}P(\boldsymbol{\varepsilon}) \\ &= \frac{1}{2}\int_{\boldsymbol{\varepsilon}} \left(\sum_{i\in\mathcal{X}} \left|\hat{p}(X=i|\boldsymbol{\varepsilon}) - \delta_{i,x(\boldsymbol{\varepsilon})}\right|\right) \mathrm{d}P(\boldsymbol{\varepsilon}) \\ &\leq \delta\end{aligned} \tag{14}$$

□

A.2. Sketch of the proof of Prop. 2

Proof. As the proof of this proposition is almost identical to the proof of Jamieson et al. (2014), we only outlines the difference due to the adaptation. In the proof of Thm. 2 in Jamieson et al. (2014), the i.i.d. assumption for rewards from each arm was used only in Lemma 3 to provide Chernoff's bound and Hoeffding's bound. As noted in Sec. 6 of Hoeffding (1963) those bounds would still hold when rewards are sampled from a finite population without replacement. Therefore, when $T^{(t)} < N$ all the bounds hold for adapted lil'UCB.

When $T_i^{(t)} = N$, the second modification sets the upper bound of the mean estimate to $\hat{\mu}^{(t)}$. That is a valid upper bound of μ_i, in fact much tighter than the bound in the original algorithm because $\hat{\mu}_i^{(t)} = \mu_i$ exactly when the entire population is observed. Therefore, as long as $T_i^{(t)} \leq N, \forall i$, Theorem 2 in Jamieson et al. (2014) applies to adapted lil'UCB with modification 1 and 2 only.

With the third modification, $T^{(t)}$ could never be bigger than N at the stopping time, which proves the second part of Prop 2. The proof can then be concluded if we can show modification 3 does not change the output of adapted lil'UCB with the first two modifications only. This is true because if we do not stop when the selected arm i satisfies $T_i^{(t)} = N$, we do not need to update the upper bound of i because the estimated mean is already exact. Since no upper bound is changed, the arm i will always be chosen for now on and eventually the original stopping criterion of $T_i^{(t)} \geq 1 + \lambda \sum_{j\neq i} T_j(t)$ is met and the same arm i will be returned. □

A.3. Proof of Prop. 3

Proof. Denote by $x^{(t)}$ the arm with the highest estimated mean at iteration t and x^* the optimal arm with the highest true mean, $\mu_{x^*} > \mu_i, \forall i \neq x^*$. If Alg. 1 does not stop in the first $t^* - 1$ iterations, the estimated means of all the survived arms become exact at the last iteration t^*, $\hat{\mu}_i^{(t^*)} = \mu_i$ because we require $T^{(t^*)} = N$. Then $x^{(t^*)} = x^*$. As we require $G(\delta, T=N, \hat{\sigma}, C) = 0, \forall \delta, \hat{\sigma}, C$, all the sub-optimal arms will be eliminated by the last iteration and the algorithm always returns the correct best arm. This proves the upper bound of the sample size of ND.

Now to prove the confidence level, all we need to show is that with at least a probability $1-\delta$ arm x^* survived all the iterations $t < t^*$.

Let us first consider the case when Alg. 1 uses the marginal variance estimate $\hat{\sigma}_i^{(t)}$. Let the events

$$\begin{aligned}E_i &= \left\{\exists t < t^*, \hat{\mu}_i^{(t)} - \mu_i > G\left(\frac{\delta}{D}, T^{(t)}, \hat{\sigma}_i^{(t)}, C_i\right)\right\}, \forall i \neq x^* \\ E_{x^*} &= \left\{\exists t < t^*, -\hat{\mu}_{x^*}^{(t)} - (-\mu_{x^*}) > G\left(\frac{\delta}{D}, T^{(t)}, \hat{\sigma}_i^{(t)}, C_i\right)\right\}\end{aligned} \tag{15}$$

Applying condition Eq. 5 and the union bound, we get $P(\cup_{i\in\mathcal{X}} E_i) \leq \sum_{i\in\mathcal{X}} E_i = \delta$. So with a probability at least $1-\delta$, none of those events will happen. In that case for any iteration $t < t^*$,

$$\begin{aligned}&\hat{\mu}_x - \hat{\mu}_{x^*} = (\hat{\mu}_x - \mu_x) - (\hat{\mu}_{x^*} - \mu_{x^*}) + (\mu_x - \mu_{x^*}) \\ &< G\left(\frac{\delta}{D}, T^{(t)}, \hat{\sigma}_x^{(t)}, C_x\right) + G\left(\frac{\delta}{D}, T^{(t)}, \hat{\sigma}_{x^*}^{(t)}, C_{x^*}\right)\end{aligned} \tag{16}$$

So arm x^* won't be eliminated at iteration t.

Similarly, for the case when Alg. 1 uses the pairwise variance estimate $\hat\sigma^{(t)}_{x,i}$, let the events

$$E_{i,x} = \Bigg\{\exists t < t^*, (\hat\mu_i^{(t)} - \hat\mu_{x^*}^{(t)}) - (\mu_i - \mu_{x^*}) > G\left(\frac{\delta}{D-1}, T^{(t)}, \hat\sigma_i^{(t)}, C_i + C_{x^*}\right)\Bigg\}, \forall i \neq x^* \tag{17}$$

Applying condition Eq. 5 and the union bound, we get $P(\cup_{i\in\mathcal{X}\backslash\{x^*\}} E_{i,x}) \leq \sum_{i\in\mathcal{X}\backslash\{x^*\}} E_{i,x} = \delta$. So with a probability at least $1-\delta$ for any iteration $t < t^*$,

$$\begin{aligned}\hat\mu_x - \hat\mu_{x^*} &= (\hat\mu_x - \hat\mu_{x^*}) - (\mu_x - \mu_{x^*}) + (\mu_x - \mu_{x^*}) \\ &< G\left(\frac{\delta}{D-1}, T^{(t)}, \hat\sigma^{(t)}_{x,x^*}, C_x + C_{x^*}\right)\end{aligned} \tag{18}$$

Therefore arm x^* won't be eliminated at iteration t. □

A.4. Proof of Prop. 7

Proof. Denote by $x^{(t)}$ the arm with the highest estimated mean at iteration t. First consider the case when Alg. 1 uses the marginal variance estimate $\hat\sigma_i^{(t)}$. With the condition in Eq. 5, it follows that $P(\cup_{i\in\mathcal{X}} E_i) \leq \sum_{i\in\mathcal{X}} P(E_i) \leq \delta$ where E_i is defined in Eq. 15. So with a probability at least $1-\delta$,

$$\begin{aligned}\hat\mu^{(t)}_{x^*} - \hat\mu^{(t)}_i >&\mu_{x^*} - \mu_i - G\left(\frac{\delta}{D}, T^{(t)}, \hat\sigma^{(t)}_{x^*}, C_{x^*}\right) \\ &- G\left(\frac{\delta}{D}, T^{(t)}, \hat\sigma^{(t)}_i, C_i\right), \forall i \neq x^*\end{aligned} \tag{19}$$

Alg. 1 will stop by iteration t if the RHS of the equation above satisfies the stopping criterion for all $i \neq x^*$, that is,

$$\begin{aligned}\mu_{x^*} - \mu_i > 2\Bigg(&G\left(\frac{\delta}{D}, T^{(t)}, \hat\sigma^{(t)}_{x^*}, C_{x^*}\right) \\ &+ G\left(\frac{\delta}{D}, T^{(t)}, \hat\sigma^{(t)}_i, C_i\right)\Bigg), \forall i \neq x^*\end{aligned} \tag{20}$$

Plugging in the definition of G_{Normal} in Eq. 9 and applying the assumption $\hat\sigma^{(t)}_i = \sigma_i$, we will get

$$\frac{\mu_{x^*} - \mu_i}{(\sigma_{x^*} + \sigma_i)} > \frac{2}{\sqrt{T^{(t)}}}\left(1 - \frac{T^{(t)} - 1}{N-1}\right)^{1/2} B_{\text{Normal}}, \forall i \neq x^* \tag{21}$$

Solve the above inequality for $T^{(t)}$ and use the definition of the gap Δ we get

$$T^{(t)} > \frac{N}{(N-1)\frac{\Delta^2}{4B^2_{\text{Normal}}(\delta/D)} + 1} \overset{\text{def}}{=} \tilde T \tag{22}$$

Since we use a doubling schedule $T^{(t)} = 2T^{(t-1)}$ with $T^{(1)} = m^{(1)}$ and $T^{(t^*)} = N$, Alg. 1 stops at an iteration no later than

$$t = \lceil \log_2(\tilde T / m^{(0)}) \rceil + 1 \tag{23}$$

And the total number of samples drawn by t is upper bounded by $D(m^{(0)} 2^{t-1} \wedge N) = T^*(\Delta)$.

Now consider the case when Alg. 1 uses the pairwise variance estimate $\hat\sigma^{(t)}_{x,i}$. With the condition in Eq. 5, it follows with the union bound that $P(\cup_{i\in\mathcal{X}\backslash\{x^*\}} E_i) \leq \sum_{i\in\mathcal{X}\backslash\{x^*\}} P(E_i) \leq \delta$ where E_i is defined in Eq. 17. So with a probability at least $1-\delta$,

$$\begin{aligned}&\hat\mu^{(t)}_{x^*} - \hat\mu^{(t)}_i \\ &> \mu_{x^*} - \mu_i - G\left(\frac{\delta}{D-1}, T^{(t)}, \hat\sigma^{(t)}_{x^*,i}, C_{x^*} + C_i\right), \forall i \neq x^*\end{aligned} \tag{24}$$

Now we can follow a similar argument as in the case with marginal variance estimate and prove the proposition. □

B. Table and Figure of $B_{\text{Normal}}(\delta, \pi_{T^{(1)}})$

Table 1 shows $B_{\text{Normal}}(\delta, \pi_{T^{(1)}})$ with δ varying in $[10^{-6}, 0.49]$, and the proportion of the first mini-batch $\pi_{T^{(1)}} = m^{(1)}/N \in \{5\times10^{-5}, 10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$. $\Phi(B)$ can be interpreted as the marginal confidence level for one iteration. The function is also shown in Fig. 4 for visualization. We will release the code to generate the table and to compute $B_{\text{Normal}}(\delta, \pi_{T^{(1)}})$ numerically.

C. Experiment Detailed Setting and Extra Results

C.1. More Results of the Synthetic Data Experiment

The results with the marginal variance estimate $\hat\sigma_i$ for Racing are shown in Fig. 5. The Racing algorithms (both EBS and Normal) performs more conservatively compared to the plots when using pairwise variance estimate $\hat\sigma_{i,j}$ in Fig. 1, but the relative performance of all the algorithms are very similar to Fig. 1.

We also provide the results with $D = 2$ and $D = 100$ when Racing algorithms use pairwise variance estimate in Fig. 6 and 7 respectively. Racing-Normal performs the best in all situations and the empirical error never exceeds the provided bound δ with a statistical significance of 0.05.

Notice that the error of adaptive lil'UCB exceeds the error tolerance in the experiment with $D = 100$ and $l_{i,n} \sim \text{Uniform}[0,1]$. This is because we use the recommended heuristic setting of parameters in Jamieson et al. (2014) that unfortunately does not satisfy the theoretical guarantee of Thm. 2 in Jamieson et al. (2014). lil'UCB (heuristic)

Table 1. $B_{\mathrm{Normal}}(\delta, \pi_{T^{(1)}})$

	$\pi_{T^{(1)}}$					
δ	5×10^{-5}	10^{-4}	5×10^{-4}	10^{-3}	5×10^{-3}	10^{-2}
1.0e-06	5.27250	5.25978	5.21523	5.19704	5.15638	5.12982
3.0e-06	5.06504	5.05294	5.00570	4.98839	4.94490	4.91964
5.0e-06	4.96669	4.95260	4.90571	4.88735	4.84311	4.81818
7.0e-06	4.89969	4.88715	4.83793	4.82079	4.77535	4.75037
9.0e-06	4.85078	4.83613	4.78840	4.76941	4.72447	4.69877
1.0e-05	4.82952	4.81667	4.76734	4.74894	4.70377	4.67696
3.0e-05	4.60397	4.58943	4.53827	4.51911	4.47119	4.44485
5.0e-05	4.49660	4.48108	4.42961	4.40734	4.36137	4.33158
7.0e-05	4.42331	4.40694	4.35512	4.33353	4.28573	4.25692
9.0e-05	4.36853	4.35265	4.29963	4.27682	4.22961	4.19891
1.0e-04	4.34343	4.32914	4.27380	4.25455	4.20386	4.17608
3.0e-04	4.09380	4.07655	4.02027	3.99632	3.94601	3.91438
5.0e-04	3.97189	3.95539	3.89641	3.87263	3.82038	3.78605
7.0e-04	3.88945	3.87195	3.81223	3.78698	3.73467	3.70026
9.0e-04	3.82665	3.80955	3.74833	3.72365	3.66977	3.63422
1.0e-03	3.79932	3.78066	3.72003	3.69596	3.64066	3.60812
3.0e-03	3.51044	3.49128	3.42498	3.39721	3.34023	3.30253
5.0e-03	3.36685	3.34814	3.27812	3.25096	3.19048	3.15168
7.0e-03	3.26922	3.24913	3.17763	3.14844	3.08769	3.04691
9.0e-03	3.19383	3.17396	3.10034	3.07142	3.00871	2.96758
1.0e-02	3.16117	3.13913	3.06612	3.03755	2.97349	2.93484
3.0e-02	2.80261	2.77885	2.69625	2.66350	2.59450	2.55058
5.0e-02	2.61646	2.59217	2.50369	2.46819	2.39672	2.34862
7.0e-02	2.48285	2.45761	2.36449	2.33100	2.25369	2.20744
9.0e-02	2.37768	2.35127	2.25533	2.22026	2.14145	2.09317
1.0e-01	2.33161	2.30704	2.20851	2.17274	2.09292	2.04351
1.3e-01	2.21073	2.18499	2.08270	2.04536	1.96346	1.91214
1.6e-01	2.10639	2.08030	1.97430	1.93665	1.85177	1.80027
1.9e-01	2.01355	1.98592	1.87702	1.83878	1.75267	1.69949
2.2e-01	1.92898	1.90035	1.78969	1.74854	1.66259	1.60660
2.5e-01	1.84734	1.81893	1.70515	1.66472	1.57552	1.52056
2.8e-01	1.76920	1.73957	1.62421	1.58220	1.49310	1.43584
3.1e-01	1.69110	1.66145	1.54360	1.50171	1.41066	1.35354
3.4e-01	1.61302	1.58274	1.46319	1.42011	1.32819	1.27094
3.7e-01	1.52953	1.49919	1.37749	1.33482	1.24221	1.18303
4.0e-01	1.44411	1.41048	1.28960	1.24393	1.15002	1.09455
4.3e-01	1.33819	1.30896	1.18163	1.14025	1.04396	0.98381
4.6e-01	1.20662	1.17447	1.05191	1.00383	0.91939	0.85273
4.9e-01	0.97014	0.94399	0.81030	0.76485	0.69587	0.61783

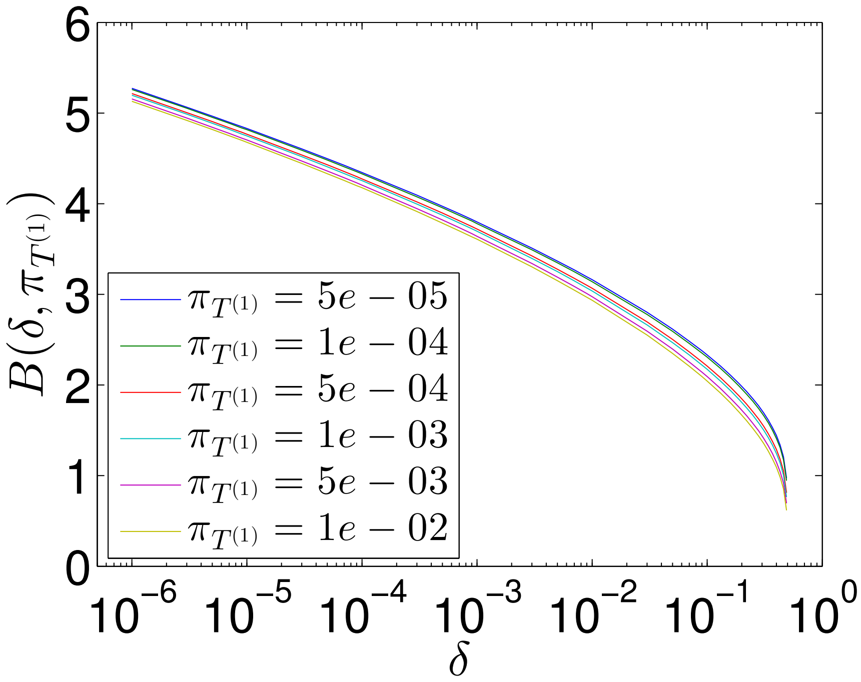


Figure 4. $B_{\text{Normal}}(\delta, \pi_{T^{(1)}})$

performed significantly better than the setting with guarantees in Jamieson et al. (2014). So we expect that adaptive lil'UCB with parameters satisfying Thm. 2 of Jamieson et al. (2014) will perform significantly worse than adaptive lil'UCB (heuristic) and Racing-Normal in terms of the reward sample complexity.

C.2. Details of the Bayesian ARCH Model Selection Experiment

An ARCH model is commonly used to model the stochastic volatility of financial times series. Let $r_t \overset{\text{def}}{=} \log(p_t/p_{t-1})$ be the logarithm return of some asset price p_t at time t. We assume a constant mean process for the return and remove the estimated mean in a pre-process step. An important problem in applying ARCH for financial data is to choose the complexity, the order q of the auto-regressive model. We treat the model selection problem as a Bayesian inference problem for the random variable q. We use a uniform prior distribution, $\pi(q) = 1/|\mathbb{Q}|$.

An MCMC algorithm was introduced in Carlin & Chib (1995) to infer the posterior model distribution by augmenting the parameter space to a complete parameter set for all models $((\alpha_i^{(j)})_{i=0}^j, \nu^{(j)}), j \in \mathbb{Q}$, then assigning the regular prior for the selected model $j = q$ and pseudopriors for those models that are not selected $j \neq q$. Then regular MCMC algorithms can be applied to sample all the random variables $q, ((\alpha_i^{(j)})_i, \nu^{(j)})_j$ without the problem of transdimensional moves as in reversible jump MCMC.

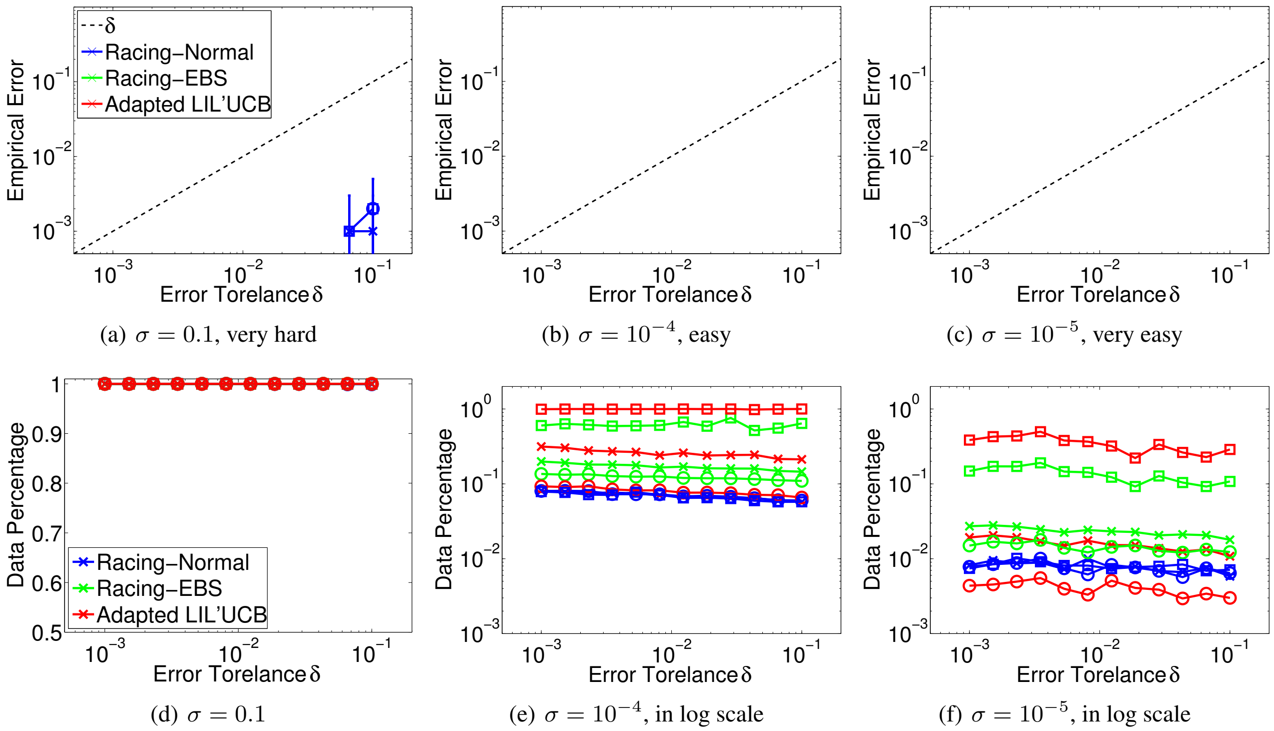
The mixing rate of Carlin & Chib (1995) depends on a proper choice of the pseudoprior for $(\alpha_i^{(j)}, \nu^{(j)})$. Ideally it should be similar to the parameter posterior when the model is chosen $p(\alpha_i^{(j)}, \nu^{(j)})|q = j, \mathbf{r})$. We first reparameterize $(\alpha_i^{(j)}, \nu^{(j)})$ with a softplus function $x = \log(1 + \exp(x'))$ to allow a full support along the real axis and then take the Laplace approximation at the MAP of transformed parameters as the pseudoprior for each model separately.

In order to avoid accessing the entire dataset each iteration, we use subsampling-based algorithms to sample all the conditionals except the pseudoprior as follows

$$q|(\boldsymbol{\alpha}^{(j)}, \nu^{(j)})_j \sim \pi(q) \prod_t p(r_t|\boldsymbol{\alpha}^{(q)}, \mathbf{r}_{t-q:t-1}, \nu^{(q)}),$$
$$(\boldsymbol{\alpha}^{(q)}, \nu^{(q)})|q \sim p(\boldsymbol{\alpha}^{(q)})p(\nu^{(q)}) \prod_t p(r_t|\boldsymbol{\alpha}^{(q)}, \mathbf{r}_{t-q:t-1}, \nu^{(q)}),$$
$$(\boldsymbol{\alpha}^{(j)}, \nu^{(j)})|q \overset{iid}{\sim} p_{\text{pseudoprior}}(\boldsymbol{\alpha}^{(j)}, \nu^{(j)}), \forall j \neq q, \quad (25)$$

where we sample q with Racing-Normal Gibbs and sample $\boldsymbol{\alpha}^{(q)}, \nu^{(q)}$ using MH with a proposal from SGLD and a rejection step provided by Racing-Normal MH. The rejection step controls the error introduced in SGLD when the step size is large.

As the marginal likelihood for each model could be differed by a few orders of magnitudes, to make sure every model is sampled sufficiently often, we first adjust the prior distribu-



(a) $\sigma = 0.1$, very hard (b) $\sigma = 10^{-4}$, easy (c) $\sigma = 10^{-5}$, very easy

(d) $\sigma = 0.1$ (e) $\sigma = 10^{-4}$, in log scale (f) $\sigma = 10^{-5}$, in log scale

Figure 5. Synthetic data. $D = 10$. Racing uses marginal variance estimate $\hat{\sigma}_i$. ((a),(b),(c)) Estimated error with 95% confidence interval. Plots not shown if no error occured. ((d),(e),(f)) proportion of sampled data. $\log f_n(i)$ is sampled from Normal (×), Uniform (◯) and LogNormal (□) distributions. Plots of Racing-Normal overlap in ((f),(g),(h)).

tion $\tilde{\pi}$ with the Wang-Landau algorithm with an annealing adaptation on $\log \tilde{\pi}$, $1/(1 + t/100)$, so that the posterior distribution $\tilde{p}(q|\mathbf{r})$ is approximately uniform. We then fix $\tilde{\pi}$ and compare the exact and approximate MCMC algorithms. The real posterior distribution can be computed as $p(q|\mathbf{r}) \propto \tilde{p}(q|\mathbf{r})/\tilde{\pi}(q)$.

We choose the step size separately for the exact and stochastic gradient Langevin dynamics (Welling & Teh, 2011) so that the acceptance rate is about 36%.

We apply the control variates by first segmenting the 2-D space of $\mathbf{z}_{j,t} \overset{\text{def}}{=} (r_t, \alpha_0^{(j)} + (\boldsymbol{\alpha}_{1:j}^{(j)})^T \mathbf{r}_{t-j:t-1})$, where $\boldsymbol{\alpha}^{(j)}$ takes the MAP value, equally into 100 bins according to marginal quantiles and then taking the reference points at the mean of each bin. We also notice that some data points have large residual reward $l_{i,n} - h_{i,n}$ when $\mathbf{z}_{j,t}$ is far from the reference point. We take 20% of the points with the largest distance in $\mathbf{z}$ as outliers, always compute them every iteration and apply the subsampling algorithm for the rest data.

C.3. Details of the Author Coreference Experiment

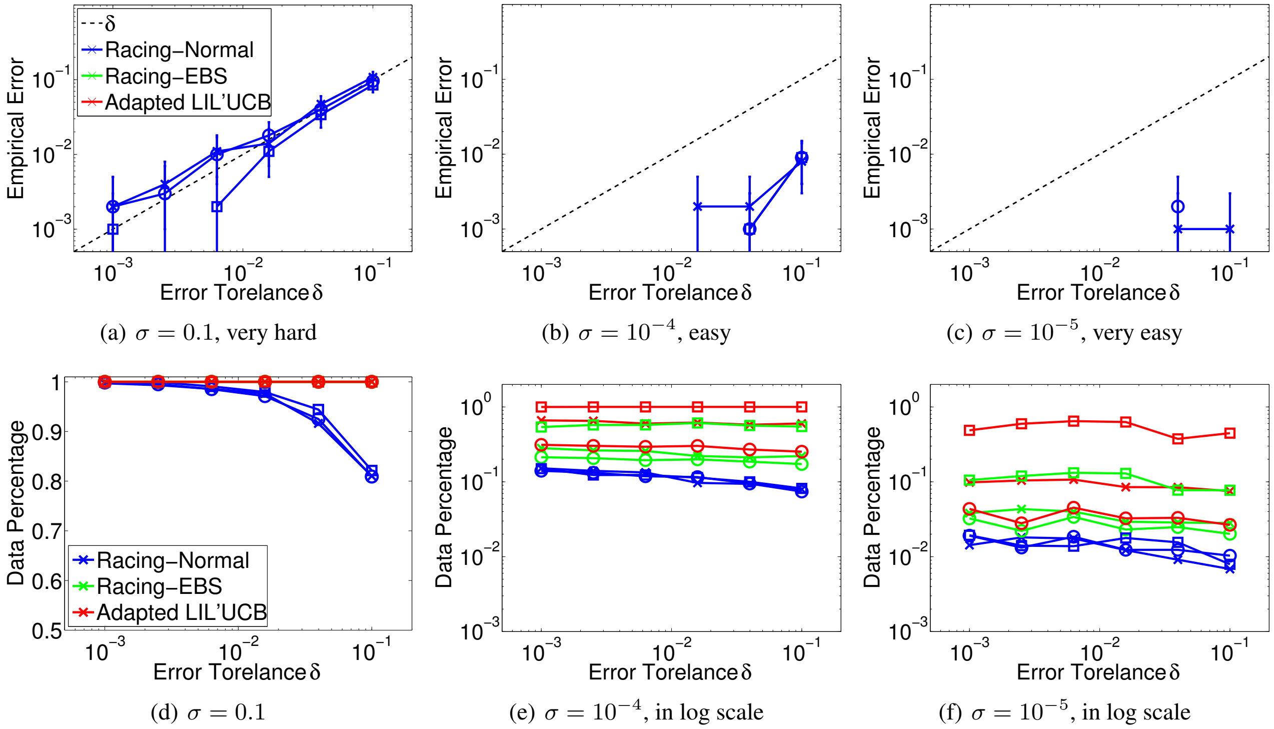
The main differences of this sampling problem from Eq. 1 are that

1. $|C_y| \neq |C_{y'}|$ and the distribution of the cluster size follows approximately a power law with the value varying from as small as 1 to thousands. If we set $m^{(1)} = 50$ as usual, we already draw about 33% of all the rewards in the first mini-batch. So we slightly abuse the Normal assumption and use a small size for $m^{(1)} = 3$ and use doubling scheme for the rest with $m_y^{(2)} = (|C_y| - 3)/10 \wedge 1$. The experiment shows an empirical error 0.045 of mis-identification of the best arm with the provided bound $\delta = 0.05$.

2. The distribution of $\{f_\theta(x_i, x_j) : j \in C_y\}$ is independent from different clusters/arms. We exploit the independence of rewards and choose the bound

$$G_{\text{Normal}}(\delta, T_i, T_j, \hat{\sigma}_i, \hat{\sigma}_j) = \left(\frac{\hat{\sigma}_i}{T_i} \left(1 - \frac{T_i - 1}{N_i - 1} \right) + \frac{\hat{\sigma}_j}{T_j} \left(1 - \frac{T_j - 1}{N_j - 1} \right) \right)^{-1/2} B_{\text{Normal}}. \tag{26}$$

This modification has the same performance as with the pairwise variance estimate and has the same computational complexity as with the marginal variance estimate $\mathcal{O}(DN)$. We compute B_{Normal} with a sub-



(a) $\sigma = 0.1$, very hard

(b) $\sigma = 10^{-4}$, easy

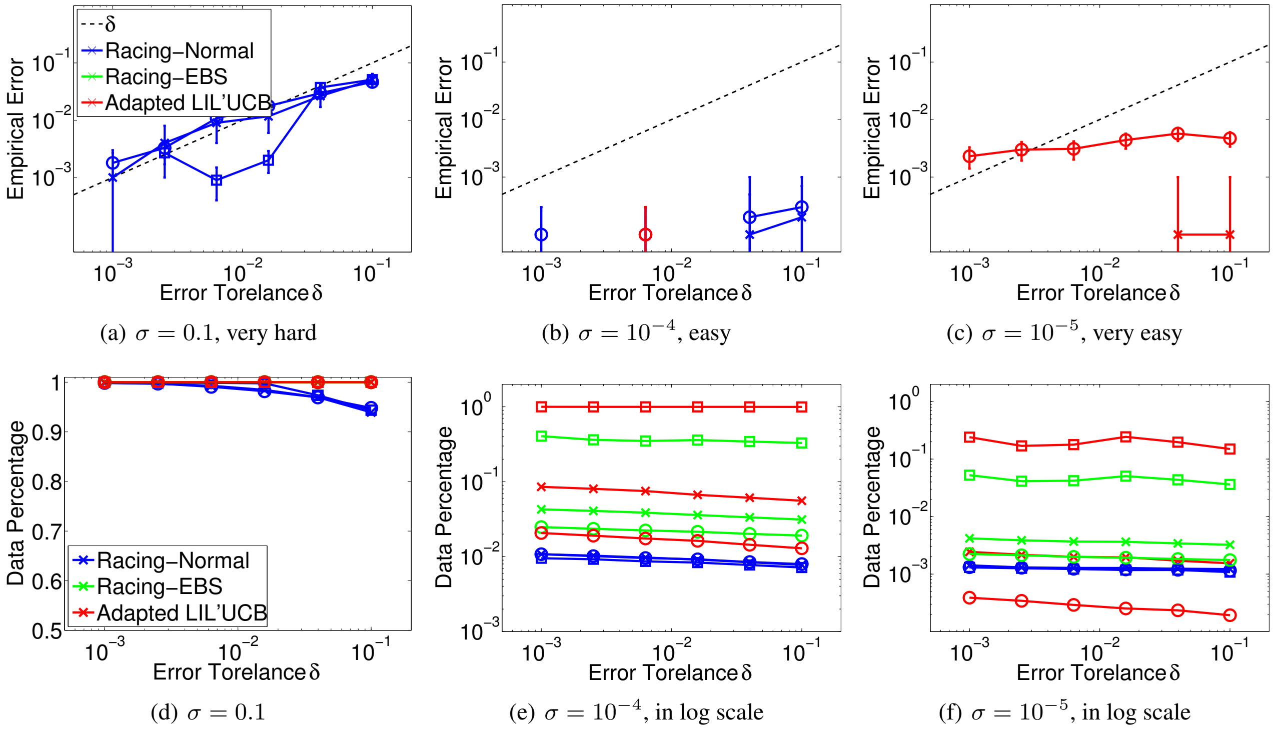
(c) $\sigma = 10^{-5}$, very easy

(d) $\sigma = 0.1$

(e) $\sigma = 10^{-4}$, in log scale

(f) $\sigma = 10^{-5}$, in log scale

Figure 6. Synthetic data. $D = 2$. Racing uses pairwise variance estimate $\hat{\sigma}_{i,j}$. ((a),(b),(c)) Estimated error with 95% confidence interval. Plots not shown if no error occured. ((d),(e),(f)) proportion of sampled data. $\log f_n(i)$ is sampled from Normal (×), Uniform (◯) and LogNormal (□) distributions. Plots of Racing-Normal overlap in ((f),(g),(h)).



(a) $\sigma = 0.1$, very hard

(b) $\sigma = 10^{-4}$, easy

(c) $\sigma = 10^{-5}$, very easy

(d) $\sigma = 0.1$

(e) $\sigma = 10^{-4}$, in log scale

(f) $\sigma = 10^{-5}$, in log scale

Figure 7. Synthetic data. $D = 100$. Racing uses pairwise variance estimate $\hat{\sigma}_{i,j}$. ((a),(b),(c)) Estimated error with 95% confidence interval. Plots not shown if no error occured. ((d),(e),(f)) proportion of sampled data. $\log f_n(i)$ is sampled from Normal (×), Uniform (◯) and LogNormal (□) distributions. Plots of Racing-Normal overlap in ((f),(g),(h)).

optimal but simpler choice as

$$B_{\mathrm{Normal}}(\delta) = \Phi^{-1}\left(1 - \frac{\delta}{t^* - 1}\right). \tag{27}$$

It is easy to show that Eq. 5 still holds in this case using a union bound across t. The bound in Eq. 27 is strictly looser than $B_{\mathrm{Normal}} = \mathcal{E}^{-1}(\delta)$ but the difference is small when $\delta \ll 1$ and diminishes to 0 as $\delta \to 0$.

We obtained the dataset from the authors of Singh et al. (2012) but it is different from what is used in Singh et al. (2012) with more difficult citations. The best B^3 F-1 score reported in this paper is a reasonable value for this data set according to personal communications with the authors of Singh et al. (2012).