A. Theoretical Properties of Gibbs-OT

We develop quantitative concentration bounds for Gibbs-OT in a finite number of iterations in order to understand the relationship between the temperature schedule and the concentration progress. The analysis also guides us to adjust cooling schedule on-the-fly, as will be shown. Proofs are provided in Supplement.

Preliminaries. Before characterizing the properties of Gibbs-OT by Definition 1, we first give the analytic expression for $p(\mathbf{z}^{n+1}|\mathbf{z}^n)$. Let $G(\cdot) : [-\infty, \infty] \mapsto [0, 1]$ be the c.d.f. of standard exponential distribution. Because $L_j^{(t+1)} < x$ by definition $\Leftrightarrow \forall i, \quad g_i^{(t)} - M_{i,j} < x$, the c.d.f. of $L_j^{(t+1)}|\mathbf{U}^{(t)}$ reads

$$\Pr\left(L_j^{(t+1)} < x \,\middle|\, \mathbf{U}^{(t)}\right) = \prod_{i=1}^{m_1} \left(1 - G\left(\frac{-x - M_{i,j} + U_i^{(t)}}{T^{(2t)}/p_i}\right)\right).$$

Likewise, the c.d.f. of $U_i^{(t)}|\mathbf{L}^{(t)}$ reads

$$\Pr\left(U_i^{(t)} < x \,\middle|\, \mathbf{L}^{(t)}\right) = \prod_{j=1}^{m_2} G\left(\frac{x - M_{i,j} - L_j^{(t)}}{T^{(2t-1)}/q_j}\right).$$

With some calculation, the following can be shown. As a note, this lemma provides an intermediate result whose main purpose is to lay down the definition of $\phi_j^{(t)}$ and $\varphi_i^{(t)}$, which are then used in defining $O(z, T)$ (Eq. (21)) and r^n (Eq. (23)) and in Theorem A.2.

Lemma A.1. *(i) Given $1 \le j \le m_2$ and $1 \le t \le N$, let the sorted index of $\{U_i^{(t)} - M_{i,j}\}_{i=1}^{m_1}$ be permutation $\{\sigma(i)\}_{i=1}^{m_1}$ such that sequence $\{U_{\sigma(i)}^{(t)} - M_{\sigma(i),j}\}_{i=1}^{m_1}$ are monotonically non-increasing. Define the auxiliary quantity*

$$\phi_j^{(t)} \overset{\text{def.}}{=} \sum_{k=1}^{m_1} \frac{(1-\mu_k)\prod_{i=1}^{k-1}\mu_i}{\sum_{i=1}^{k} p_{\sigma(k)}}, \tag{19}$$

where

$$1 \ge \mu_i \overset{\text{def.}}{=} \exp\left\{\frac{\sum_{i=1}^{k} p_{\sigma(k)}}{T^{(2t)}}\Big[\left(U_{\sigma(i+1)} - M_{\sigma(i+1),j}\right) - \left(U_{\sigma(i)} - M_{\sigma(i),j}\right)\Big]\right\}$$

for $i = 1, \ldots, m_1 - 1$, and $\mu_{m_1} \overset{\text{def.}}{=} 0$. Then, the conditional expectation

$$\mathbb{E}\left[L_j^{(t+1)} \,\middle|\, \mathbf{U}^{(t)}\right] = U_{\sigma(1)}^{(t)} - M_{\sigma(1),j} - \phi_j^{(t)} T^{(2t)}.$$

In particular, we denote $\sigma(1)$ by I_j^t or $I(j,t)$.

(ii) Given $1 \le i \le m_1$ and $1 \le t \le N$, let the sorted index of $\{M_{i,j} + L_j\}_{j=1}^{m_2}$ be permutation $\{\sigma(j)\}_{j=1}^{m_2}$ such that the sequence $\{M_{i,\sigma(j)} + L_{\sigma(j)}^{(t)}\}_{j=1}^{m_2}$ are monotonically non-decreasing. Define the auxiliary quantity

$$\psi_i^{(t)} \overset{\text{def.}}{=} \sum_{k=1}^{m_2} \frac{(1-\lambda_k)\prod_{j=1}^{k-1}\lambda_k}{\sum_{j=1}^{k} q_{\sigma(j)}}, \tag{20}$$

where

$$1 \ge \lambda_j \overset{\text{def.}}{=} \exp\left\{\frac{\sum_{j=1}^{k} q_{\sigma(j)}}{T^{(2t-1)}}\Big[\left(M_{i,\sigma(j)} + L_{\sigma(j)}^{(t)}\right) - \left(M_{i,\sigma(j+1)} + L_{\sigma(j)}^{(t+1)}\right)\Big]\right\}$$

for $i = 1, \ldots, m_2 - 1$ and $\lambda_{m_2} = 0$. Then, the conditional expectation

$$\mathbb{E}\left[U_i^{(t)} \,\middle|\, \mathbf{L}^{(t)}\right] = M_{i,\sigma(1)} + L_{\sigma(1)}^{(t)} + \psi_i^{(t)} T^{(2t-1)}.$$

In particular, we denote $\sigma(1)$ by J_i^t or $J(i,t)$.

We note that the calculation of Eq. (19) and Eq. (20) needs $O(m_1 \log m_1)$ and $O(m_2 \log m_2)$ time respectively. By a few additional calculations, we introduce the notation $\mathcal{O}(\cdot,\cdot)$:

$$\begin{aligned}
&\mathcal{O}(\mathbf{z}^{2t}, T^{(2t)}) \\
&\overset{\text{def.}}{=} \mathbb{E}\left[\langle \mathbf{q}, \mathbf{L}^{(t)}\rangle - \langle \mathbf{q}, \mathbf{L}^{(t+1)}\rangle \,\middle|\, \mathbf{U}^{(t)}, \mathbf{L}^{(t)}\right] \\
&= \sum_{j=1}^{m_2} \left(M_{I_j^t,j} + L_j^{(t)} - U_{I_j^t}^{(t)} + \phi_j^{(t)} T^{(2t)}\right) q_j \\
&\mathcal{O}(\mathbf{z}^{2t-1}, T^{(2t-1)}) \\
&\overset{\text{def.}}{=} \mathbb{E}\left[\langle \mathbf{p}, \mathbf{U}^{(t)}\rangle - \langle \mathbf{p}, \mathbf{U}^{(t-1)}\rangle \,\middle|\, \mathbf{U}^{(t-1)}, \mathbf{L}^{(t)}\right] \\
&= \sum_{i=1}^{m_1} \left(M_{i,J_i^t} + L_{J_i^t}^{(t)} - U_i^{(t-1)} + \psi_i^{(t)} T^{(2t-1)}\right) p_i
\end{aligned} \tag{21}$$

Note that $\mathcal{O}(\mathbf{z}^n, T^n) = \mathbb{E}\left[V(\mathbf{z}^{n+1}) - V(\mathbf{z}^n)|\mathbf{z}^n\right]$.

Recovery of Approximate Primal Solution. An approximate $(m_1 + m_2)$-sparse primal solution[4] can be recovered from $\mathbf{z}^n$ at $n = 2t$ by

$$Z \approx \frac{1}{2}\text{sparse}(1 : m_1, J(1 : m_1, t), \mathbf{p}) + \frac{1}{2}\text{sparse}(I(1 : m_2, t), 1 : m_2, \mathbf{q}) \in \mathbb{R}^{m_1 \times m_2}. \tag{22}$$

Concentration Bounds. We are interested in the concentration bound related to $V(\mathbf{z}^n)$ because it replaces the true

[4] The notation of sparse$(\cdot,\cdot,\cdot)$ function is introduced under the syntax of MATLAB: http://www.mathworks.com/help/matlab/ref/sparse.html

Wasserstein loss in WLMs. Given $\mathbf{U}^{(0)}$ (*i.e.*, $\mathbf{z}^1$ is implied), for $n = 1, \ldots, 2N$, we let

$$r^n = V(\mathbf{z}^n) - \sum_{s=1}^{n-1} \mathcal{O}(\mathbf{z}^s, T^{(s)}) \ . \tag{23}$$

This is crucial for one who wants to know whether the cooling schedule is too fast to secure the suboptimality within a finite budget of iterations. The following Theorem A.2 gives a possible route to approximately realize this goal. It bounds the difference between

$$V(\mathbf{z}^n) - V(\mathbf{z}^1) \text{ and } \sum_{s=1}^{n-1} \mathbb{E}\left[V(\mathbf{z}^{s+1}) - V(\mathbf{z}^s)|\mathbf{z}^s\right],$$

the second of which is a quantitative term representing sum of a sequence. We see that $\mathcal{O}(\mathbf{z}^s, T^{(s)}) = \mathbb{E}\left[V(\mathbf{z}^{s+1}) - V(\mathbf{z}^s)|\mathbf{z}^s\right] = 0$ if and only if $T^{(s)} = \mathcal{T}(\mathbf{z}^s) \stackrel{\text{def.}}{=}$

$$\begin{cases} -\dfrac{1}{\langle \phi^{(t)}, \mathbf{q}\rangle} \sum\limits_{j=1}^{m_2} q_j \left[M_{I_j^t, j} + L_j^{(t)} - U_{I_j^t}^{(t)}\right] & \text{if } s = 2t \\ -\dfrac{1}{\langle \psi^{(t)}, \mathbf{p}\rangle} \sum\limits_{i=1}^{m_1} p_i \left[M_{i, J_i^t} + L_{J_i^t}^{(t)} - U_i^{(t-1)}\right] & \text{if } s = 2t-1 \end{cases} \tag{24}$$

In the practice of Gibbs-OT, choosing the proper cooling schedule for a specific WLM needs trial-and-error. Here we present a heuristics that the temperature $T^{(s)}$ is often chosen and adapted around $\eta\mathcal{T}(z^s)$, where $\eta \in [0.1, 0.9]$. We have two concerns regarding the choice of temperature T: First, in a WLM, the cost $V(\mathbf{z})$ is to be gradually minimized, hence a temperature T smaller than $\mathcal{T}(\mathbf{z}^s)$ at every iteration ensures that the cost is actually decreased by expectation, *i.e.*, $\mathbb{E}[V(\mathbf{z}^n) - V(\mathbf{z}^1)] < 0$; second, if T is too small, it takes many iterations to reach a highly accurate equilibrium, which might not be necessary for a single outer level step of parameter update.

Theorem A.2 (Concentration bounds for finite time Gibbs-OT)**.** *First, r^n (by definition) is a martingale subject to the filtration of $\mathbf{z}_1, \ldots, \mathbf{z}_n$. Second, given a $\varepsilon \in (0, 1)$, for $n = 1, \ldots, 2N-1$ if we choose the temperature schedule $T^{(1)}, \ldots, T^{(2N)}$ such that (i) $C^n \cdot T^{(n)} \le a_n$, or (ii) $\exists \gamma > 0$, $\log\left(\frac{2N\max\{m_1, m_2\}}{\varepsilon}\right) \cdot T^{(n)} + D^n \le \gamma a_n$, where $\{a_n \ge 0\}$ is a pre-determined array. Here for $t = 1, \ldots, N$,*

$$\begin{aligned} C^{2t-1} &\stackrel{\text{def.}}{=} \langle \psi^{(t)}, \mathbf{p}\rangle \ , \\ C^{2t} &\stackrel{\text{def.}}{=} \langle \phi^{(t)}, \mathbf{q}\rangle \ , \\ D^{2t-1} &\stackrel{\text{def.}}{=} \sum_{i=1}^{m_1} p_i \mathcal{R}\left(M_{i,\cdot}^T + \mathbf{L}^{(t)}; \mathbf{q}\right) \ , \\ D^{2t} &\stackrel{\text{def.}}{=} \sum_{i=1}^{m_2} q_j \mathcal{R}\left(M_{\cdot,j} - \mathbf{U}^{(t)}; \mathbf{p}\right) \ , \end{aligned}$$

where $M_{i,\cdot}$ and $M_{\cdot,j}$ represents the i-th rows and j-th columns of matrix M respectively, $\psi^{(t)}$ and $\phi^{(t)}$ are defined in Lemma A.1, and regret function $\mathcal{R}(\mathbf{x}; \mathbf{w}) \stackrel{\text{def.}}{=} \sum_{i=1}^m w_i x_i - \min_{1\le i\le m} x_i$ for any $\mathbf{w} \in \Delta_m$ and $\mathbf{x} \in \mathbb{R}^m$. Then for any $K > 0$, we have

$$Pr\left(r^{2N} < r^1 - K\right) \le \exp\left[-\frac{K^2}{2\sum_{i=1}^{2N-1} a_n^2}\right] \ , \tag{25}$$

or

$$Pr\left(r^{2N} > r^1 + \gamma K\right) \le \exp\left[-\frac{K^2}{2\sum_{i=1}^{2N-1} a_n^2}\right] + \varepsilon \ . \tag{26}$$

Remark 5. The bound obtained is a quantitative Hoeffding bound, not a bound that guarantees contraction around the true solution of dual OT. Nevertheless, we argue that this bound is still useful in investigating the proposed Gibbs sampler when the temperature is not annealed to zero. Particularly, the bound is for cooling schedules in general, *i.e.*, it is more applicable than a bound for a specific schedule. There has long been a gap between the practice and theory of SA despite of its wide usage. Our result likewise falls short of firm theoretical guarantee from the optimization perspective, as with the usual application of SA.

B. Proof of Lemmas and Theorem

The minimum of n independent exponential random variables with different parameters has computatable formula for its expectation. The result immediately lays out the proof of Lemma A.1.

Lemma B.1. *Suppose we have n independent exponential random variables e_i whose c.d.f. is by $f_i(x) = \min\{\exp(\omega_i(x - z_i)), 1\}$. Without lose of generality, we assume $z_1 \ge z_2 \ge \ldots \ge z_n$, then let $z_{n+1} = -\infty, h_i = \exp\left[\sum_{j=1}^i \omega_j (z_{i+1} - z_i)\right] \le 1$ (with $h_n = 0, z_{n+1} h_n = 0$), we have*

$$\mathbb{E}\left[\max\{e_1, \ldots, e_n\}\right] = z_1 - \sum_{i=1}^n \frac{(1 - h_i)\prod_{j=1}^{i-1} h_i}{\sum_{j=1}^i \omega_j} \ .$$

Proof. The c.d.f. of $\max\{e_1, \ldots, e_n\}$ is $F(x) = \prod_{i=1}^n f_i(x)$ which is piece-wise smooth with interval

(z_{i+1}, z_i), we want to calculate $\int_{-\infty}^{\infty} x dF(x)$.

$$
\begin{aligned}
&\int_{-\infty}^{\infty} x dF(x) \\
=\ & \sum_{i=1}^{n} \int_{z_{i+1}}^{z_i} x dF(x) + 0 \\
=\ & \sum_{i=1}^{n} \int_{z_{i+1}}^{z_i} x d \exp\left[\sum_{j=1}^{i} \omega_j (x - z_j)\right] \\
=\ & \sum_{i=1}^{n} \int_{z_{i+1}}^{z_i} \left[\sum_{j=1}^{i} \omega_j\right] x \exp\left[\sum_{j=1}^{i} \omega_j (x - z_j)\right] dx \\
=\ & \sum_{i=1}^{n} \Bigg\{ \left(z_i - \frac{1}{\sum_{j=1}^{i} \omega_j}\right) \exp\left[\sum_{j=1}^{i} \omega_j (z_i - z_j)\right] \\
& - \left(z_{i+1} - \frac{1}{\sum_{j=1}^{i} \omega_j}\right) \exp\left[\sum_{j=1}^{i} \omega_j (z_{i+1} - z_j)\right] \Bigg\} \\
=\ & \sum_{i=1}^{n} \left[(z_i - z_{i+1} h_i) - \frac{1 - h_i}{\sum_{j=1}^{i} \omega_j}\right] \prod_{j=1}^{i-1} h_i \\
=\ & \sum_{i=1}^{n} \left[z_i \prod_{j=1}^{i-1} h_i - z_{i+1} \prod_{j=1}^{i} h_i\right] \\
& - \sum_{i=1}^{n} \frac{(1 - h_i) \prod_{j=1}^{i-1} h_i}{\sum_{j=1}^{i} \omega_j} \\
=\ & z_1 - \sum_{i=1}^{n} \frac{(1 - h_i) \prod_{j=1}^{i-1} h_i}{\sum_{j=1}^{i} \omega_j} \ .
\end{aligned}
$$

□

Therefore Lemma A.1 is proved up to trivial calculation using the above Lemma B.1. In order to further prove Lemma B.3, we also have (by definition of $F(x)$).

Lemma B.2. *Subject to the setup of Lemma B.1, we also have*

$$\max\{e_1, \ldots, e_n\} \le z_1 \ ,$$

and

$$F(x) \le \min\left\{\exp\left[\sum_{i=1}^{n} \omega_i (x - z^*)\right], 1\right\}, -\infty < x < \infty,$$

where $z^* = \dfrac{\sum_{i=1}^{n} \omega_i z_i}{\sum_{i=1}^{n} \omega_i}$.

Therefore, based on the observation of Lemma B.2, the tail probability $Pr(\max\{e_1, \ldots, e_n\} < x)$ is upper bounded by the probability of an exponential random variable, which lead us to the proof of Lemma B.3.

Lemma B.3. *Note that Eq. (21) implies* $\mathbb{E}\left[r^{n+1} - r^n | \mathbf{z}^1, \ldots, \mathbf{z}^n\right] = 0$ *for* $t = 1, \ldots, 2N$. *Therefore,* $\{r^n\}$ *is a (discrete time) martingale subject to the filtration of* $\{\mathbf{z}^n\}$. *(Recall the notation by Eq. (14).) Moreover, we have the following two bounds. First, we can establish the left hand side bound for* $\{r^{n+1} - r^n\}_{n=1}^{2N-1}$*:*

$$r^n - r^{n+1} \le C^n \cdot T^{(n)},$$

where for $t = 1, \ldots, N$

$$C^{2t-1} \overset{\text{def.}}{=} \langle \psi^{(t)}, \mathbf{p} \rangle \text{ and } C^{2t} \overset{\text{def.}}{=} \langle \phi^{(t)}, \mathbf{q} \rangle. \tag{27}$$

Second, we also bound on the right hand side. That said, for any $1 > \varepsilon > 0$*, we have*

$$
\begin{aligned}
&Pr\Big(\exists n \in \{1, \ldots, 2N\}, \text{ s.t. } r^{n+1} - r^n \\
&\ge \log\left(\frac{2N \max\{m_1, m_2\}}{\varepsilon}\right) \cdot T^{(n)} + D^n \Big| \mathbf{z}^1, \ldots, \mathbf{z}^n\Big) \le \varepsilon,
\end{aligned} \tag{28}
$$

where for $t = 1, \ldots, N$

$$D^{2t-1} \overset{\text{def.}}{=} \sum_{i=1}^{m_1} p_i \mathcal{R}\left(M_{i,\cdot}^T + \mathbf{L}^{(t)}; \mathbf{q}\right) \tag{29}$$

$$D^{2t} \overset{\text{def.}}{=} \sum_{i=1}^{m_2} q_j \mathcal{R}\left(M_{\cdot,j} - \mathbf{U}^{(t)}; \mathbf{p}\right), \tag{30}$$

where $M_{i,\cdot}$ *and* $M_{\cdot,j}$ *represents the* i*-th rows and* j*-th columns of matrix* M *respectively.*

Proof. On one hand, because for each $i \in \{1, \ldots, m_1\}$, $U_i^{(t)} | \mathbf{L}^{(t)}$ is lower bounded by $M_{i,J(i,t)} + L_{J(i,t)}^{(t)}$ (Lemma B.2), and for each $j \in \{1, \ldots, m_2\}$, $L_j^{(t)} | \mathbf{U}^{(t-1)}$ is upper bounded by $U_{I(j,t)}^{(t-1)} - M_{I(j,t),j}$ (Lemma B.2), we easily (by definition) have $r^{n+1} | \mathbf{z}_1, \ldots, \mathbf{z}^n$ is lower bounded by $r^n - C^n \cdot T^{(n)}$.

On the other hand, we have if $r^{n+1} - r^n \ge \log(1/\varepsilon_0) \cdot T^{(n)} + D^n | \mathbf{z}_1, \ldots, \mathbf{z}_n$ for some $\varepsilon_0 > 0$, then at least one of $U_i^{(t)}$ (or $L_j^{(t)}$) violates the bound $\log(1/\varepsilon_0) \cdot T^{(n)} + \mathcal{R}(M_{i,\cdot}^T + \mathbf{L}^{(t)}; \mathbf{q})$ (or $\log(1/\varepsilon_0) \cdot T^{(n)} + \mathcal{R}(M_{\cdot,j} - \mathbf{U}^{(t)}; \mathbf{p})$), whose probability using Lemma B.2 is shown to be less than ε_0. Therefore, we have for each n

$$
\begin{aligned}
&Pr(r^{n+1} - r^n \ge \log(1/\varepsilon_0) \cdot T^{(n)} + D^n | \mathbf{z}_1, \ldots, \mathbf{z}_n) \\
&\qquad \le \max\{m_1, m_2\} \varepsilon_0 \ ,
\end{aligned} \tag{31}
$$

and

$$
\begin{aligned}
&Pr(\exists n, r^{n+1} - r^n \ge \log(1/\varepsilon_0) \cdot T^{(n)} + D^n | \mathbf{z}_1, \ldots, \mathbf{z}_n) \\
&\qquad \le 2N \max\{m_1, m_2\} \varepsilon_0 \ ,
\end{aligned} \tag{32}
$$

Let $\varepsilon = 2N \max\{m_1, m_2\} \varepsilon_0$, which concludes our result. □

Given Lemma B.3, we can prove Theorem A.2 by applying the classical Azuma's inequality for the left-hand side bound, and applying one of its extensions (Proposition 34 in (Tao and Vu, 2015)) for the right-hand side bound. Remark that Theorem A.2 is about a single OT. For multiple different OTs, which share the same temperature schedule, one can have asymptotic bounds using the Law of Large Numbers due to the fact that their Gibbs samplers are independent with each other. Let $R^n = \frac{1}{S}\sum_{k=1}^{S} r_k^n$, where r_k^n is defined by Eq. (23) for sample k. Since for any $\varepsilon > 0$, one has $P(\left|R^{n+1} - R^n\right| > \varepsilon) \to 0$, as $S \to \infty$, one can have the asymptotic concentration bound for R^{2N} that for any $\varepsilon_1, \varepsilon_2 > 0$, there exists S such that $P(\left|R^{2N} - R^1\right| > \varepsilon_1) \leq \exp\left(-\frac{1}{2N\varepsilon_2}\right)$.

Tao, Terence and Vu, Van. Random matrices: Universality of local spectral statistics of non-Hermitian matrices. *The Annals of Probability*, 43(2):782-874, 2015.