Personalized playlist modeling

Brian McFee

November 16, 2014

1 Preliminaries

Let $\mathcal{U}$ denote the set of m users, $\mathcal{X}$ denote the set of n songs, and $\mathcal{Y}$ denote the set of playlists. Let $\mathcal{H}$ denote an undirected hypergraph over $(\mathcal{X}, \mathcal{E})$, where $\mathcal{E} \subseteq 2^{\mathcal{X}}$ is the collection of edges (attribute coincidence).

We require the following conditions on $\mathcal{H}$:

- No edge is empty: $\forall e \in \mathcal{E} : |e| \geq 1$,
- Each vertex is contained: $\forall x \in \mathcal{X}, \exists e \in \mathcal{E} \,\mathrm{s.\,t.}\, x \in e$,
- Every pair of vertices are connected: $\forall x_1, x_2 \in \mathcal{E}, \exists e \in \mathcal{E} \,\mathrm{s.\,t.}\{x_1, x_2\} \subseteq e$.

The first condition is trivially satisfied by preprocessing. The second and third conditions are both satisfied by including a uniform edge $\mathcal{E} \ni e_0 = \mathcal{X}$.

e_i will denote the ith standard basis vector.

$\gtrsim$ and $\eqsim$ will denote (in)equality modulo an additive constant.

2 Personalized model

The hypergraph random walk model proceeds as follows:

- select an initial subset $e \in \mathcal{E}$
- select a song x from e
- select a new subset e' containing x
- go to step 2

We add two features to the previous model. First, a bias term b_i is included to model the global popularity of each song. Next, we incorporate a latent factor model to capture individual user preferences for songs.

2.1 Model equations

Let $y = (x_0, x_1, \ldots, x_T) \in \mathcal{Y}$ denote a playlist, and let $i \in [m]$ index the corresponding user. The probability of generating y given u_i and the model parameters $\theta := \{u, v, b, w\}$ is defined as follows:

$$\mathbf{P}\left[Y = y \middle|\, U = i, \Theta = \theta\right] = \mathbf{P}\left[X = x_0 \middle|\, U = i, \Theta = \theta\right] \prod_{t=1}^{T} \mathbf{P}\left[X_t = x_t \middle|\, X_{t-1} = x_{t-1}, U = i, \Theta = \theta\right]$$

The initial edge distribution is characterized as

$$\mathbf{P}\left[E = e \middle|\, \Theta = \theta\right] := \frac{\exp\{w_e\}}{\sum_{f \in \mathcal{E}} \exp\{w_f\}}$$

The probability of drawing a song from a given subset is characterized as

$$\mathbf{P}\left[X_t = x_t \middle|\ E = e, U = i, \Theta = \theta\right] := \frac{[\![x_t \in e]\!] \exp\left\{u_i^\mathsf{T} v_t + b_t\right\}}{\sum_{j \in \mathcal{X}} [\![x_j \in e]\!] \exp\left\{u_i^\mathsf{T} v_j + b_j\right\}},$$

The bigram transition probability is defined by marginalizing over the edge set $\mathcal{E}$, as follows:

$$\mathbf{P}\left[X_t = x_t \middle|\ X_{t-1} = x_{t-1}, U = i, \Theta = \theta\right] := \sum_{e \in \mathcal{E}} \mathbf{P}\left[X_t = x_t \middle|\ E = e, U = i, \Theta = \theta\right] \cdot \mathbf{P}\left[E = e \middle|\ X_{t-1} = x_{t-1}, \Theta = \theta\right]$$

$$\mathbf{P}\left[E = e \middle|\ X_{t-1} = x_{t-1}, \Theta = \theta\right] := \frac{[\![x_{t-1} \in e]\!] \exp\{w_e\}}{\sum_{f \in \mathcal{E}} [\![x_{t-1} \in f]\!] \exp\{w_f\}}$$

Finally, the model parameters are defined by the following prior distributions

$$\begin{aligned} u_i &\sim \mathcal{N}\left(0, \sigma_u^2 I\right) \\ v_j &\sim \mathcal{N}\left(0, \sigma_v^2 I\right) \\ b_j &\sim \mathcal{N}\left(0, \sigma_b^2\right) \\ w_e &\sim \mathcal{N}\left(0, \sigma_w^2\right). \end{aligned}$$

This differs from the previous hypergraph random walk model in that the edge weights are log-normal instead of exponentially distributed.

Note that all sums over edge membership indicators can be implemented as a dot product against the (sparse, constant) song-edge incidence matrix $H \in \{0,1\}^{|\mathcal{X}| \times |\mathcal{E}|}$. If we overload notation, and let $U \in \mathbb{R}^{d \times m}$, $V \in \mathbb{R}^{d \times n}$, $b \in \mathbb{R}^n$ and $w \in \mathbb{R}^{|\mathcal{E}|}$, then the probabilities can be expressed compactly as follows.

$$\begin{aligned} \mathbf{P}\left[E \middle|\ \Theta = \theta\right] &= \frac{\exp\{w\}}{\mathbf{1}^\mathsf{T} \exp\{w\}} \\ \mathbf{P}\left[E \middle|\ X_{t-1} = x_{t-1}, \Theta = \theta\right] &= \frac{H_{t-1,\cdot} \odot \exp\{w\}}{H_{t-1,\cdot}^\mathsf{T} \exp\{w\}} \\ \mathbf{P}\left[X \middle|\ E = e, U = i, \Theta = \theta\right] &= \frac{H_{\cdot,e} \odot \exp\left\{U_{\cdot,i}^\mathsf{T} V + b\right\}}{H_{\cdot,e}^\mathsf{T} \exp\left\{U_{\cdot,i}^\mathsf{T} V + b\right\}} \end{aligned}$$

2.2 Special cases

The model above generalizes several standard(ish) models directly:

- Fixing v_j, b_j to zero recovers the original hypergraph model, not counting the repetition constraint and change of prior.
- Fixing v_j to zero enables item bias without personalization.
- Restricting $\mathcal{E} = \{\mathcal{X}\}$ (*i.e.*, contain only the uniform edge) recovers a simple stochastic latent factor recommender.

Note that sampling is relatively efficient if the edge sets are small (on average) and H is sparse.

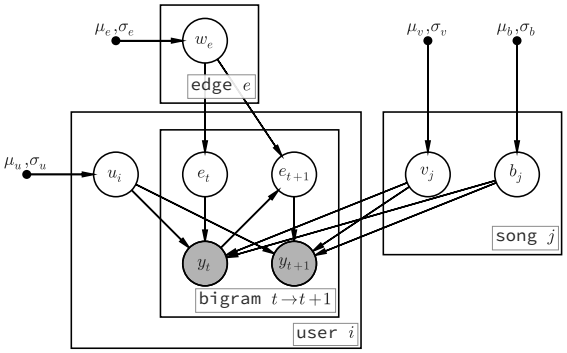

Figure 1: Model structure

3 Parameter estimation

Given a sample collection of independent playlists $\mathcal{Y} = \{y\}$, we seek the MAP parameters

$$\begin{aligned}\overline{\Theta} \in \operatorname*{argmax}_{\Theta} \mathbf{P}\left[\Theta|\ \mathcal{Y}\right] &= \operatorname*{argmax}_{\Theta} \mathbf{P}\left[\mathcal{Y}|\ \Theta\right] \mathbf{P}\left[\Theta\right] \\ &= \operatorname*{argmax}_{\Theta} \log \mathbf{P}\left[\mathcal{Y}|\ \Theta\right] + \log \mathbf{P}[\Theta] \\ &= \operatorname*{argmax}_{\Theta} \sum_{y \in \mathcal{Y}} \log \mathbf{P}\left[y|\ \Theta\right] + \log \mathbf{P}[\Theta]\end{aligned}$$

The MAP objective is not jointly concave in all parameters U, V, b, w. We will optimize parameters by block coordinate ascent.

3.1 U-step — V, b, w fixed

The U matrix can be decomposed into its columns, $U = [u_1, u_2, \cdots, u_m]$, where u_i corresponds to the factor for the ith user. Since users are independent, the columns of u can be optimized independently. Note also that each u_i depends only on the playlists for user i.

Let $\mathcal{Y}_i$ denote the playlists of user i. The data term for the objective over u_i looks as follows:[1]

$$\begin{aligned}f(u_i) &:= \sum_{y \in \mathcal{Y}_i} \log \mathbf{P}\left[y|\ u_i, V, b, w\right] \\ &= \sum_{y \in \mathcal{Y}_i} \left(\log \mathbf{P}\left[x_0|\ u_i, V, b, w\right] + \sum_{t=1}^{T_y} \log \mathbf{P}\left[x_t|\ x_{t-1}, u_i, V, b, w\right] \right) \end{aligned} \tag{1}$$

[1] Throughout this document, canceled terms (slashed in red) are constant with respect to the variable of optimization.

As noted in the previous section, the initial song likelihood is computed as:

$$\begin{aligned}
\log \mathbf{P}\left[x_0 \middle|\ u_i, V, b, w\right] &= \log \sum_{e \in \mathcal{E}} \mathbf{P}\left[e \middle|\ w\right] \mathbf{P}\left[x_0 \middle|\ e, u_i, V, b\right] \\
&= \log \sum_{e \in \mathcal{E}} \frac{\exp\{w_e\}}{\sum_f \exp\{w_f\}} \frac{H_{0,e} \exp\{u_i^\mathsf{T} v_0 + b_0\}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&= \log \exp\{u_i^\mathsf{T} v_0 + b_0\} \sum_{e \in \mathcal{E}} \frac{\exp\{w_e\}}{\sum_f \exp\{w_f\}} \frac{H_{0,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&= u_i^\mathsf{T} v_0 + \cancel{b_0} + \log \sum_{e \in \mathcal{E}} \frac{\exp\{w_e\}}{\sum_f \exp\{w_f\}} \frac{H_{0,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&\approx u_i^\mathsf{T} v_0 + \log \sum_{e \in \mathcal{E}} \frac{\exp\{w_e\}}{\cancel{\sum_f \exp\{w_f\}}} \frac{H_{0,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&\approx u_i^\mathsf{T} v_0 + \log \sum_{e \in \mathcal{E}} \exp\{w_e\} \frac{H_{0,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \qquad (2)
\end{aligned}$$

The bigram transition likelihoods are similarly computed:

$$\begin{aligned}
\log \mathbf{P}\left[x_t \middle|\ x_{t-1}, u_i, V, b, w\right] &= \log \sum_{e \in \mathcal{E}} \mathbf{P}\left[e \middle|\ x_{t-1}, w\right] \mathbf{P}\left[x_t \middle|\ e, u_i, V, b\right] \\
&= \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{H_{t-1,\cdot}^\mathsf{T} \exp\{w\}} \frac{H_{t,e} \exp\{u_i^\mathsf{T} v_t + b_t\}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&= u_i^\mathsf{T} v_t + \cancel{b_t} + \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{H_{t-1,\cdot}^\mathsf{T} \exp\{w\}} \frac{H_{t,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&\approx u_i^\mathsf{T} v_t + \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{\cancel{H_{t-1,\cdot}^\mathsf{T} \exp\{w\}}} \frac{H_{t,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&\approx u_i^\mathsf{T} v_t + \log \sum_{e \in \mathcal{E}} \exp\{w_e\} \frac{H_{t-1,e} H_{t,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}}, \qquad (3)
\end{aligned}$$

which generalizes the initial-state likelihood (with $t = 0$) by including the binary transition factor $H_{t-1,e}$. (Note, if we hallucinate a song at $t = -1$ with $H_{-1,\cdot} = \mathbf{1}^\mathsf{T}$, then these definitions are equivalent.)

Equation (1) is a sum over terms of the form of eq. (3), which are not concave. However, we can instead maximize a lower bound by applying Jensen's inequality to the summation. To ease presentation,

let $\mathcal{E}_{t-1,t} = \{e |\ e \in \mathcal{E},\ \{x_{t-1}, x_t\} \subseteq e\}$ denote the set of feasible edges containing the bigram (x_{t-1}, x_t).

$$\begin{aligned}
\log \sum_{e\in\mathcal{E}} \exp\{w_e\} \frac{H_{t-1,e} H_{t,e}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} &= \log \sum_{e\in\mathcal{E}_{t-1,t}} \exp\{w_e\} \frac{1}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&= \log \left(\left(\frac{\sum_{f\in\mathcal{E}_{t-1,t}} \exp\{w_f\}}{\sum_{f\in\mathcal{E}_{t-1,t}} \exp\{w_f\}} \right) \sum_{e\in\mathcal{E}_{t-1,t}} \exp\{w_e\} \frac{1}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \right) \\
&= \log \sum_{e\in\mathcal{E}_{t-1,t}} \frac{\exp\{w_e\}}{\sum_{f\in\mathcal{E}_{t-1,t}} \exp\{w_f\}} \frac{1}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} + \log \cancel{\sum_{f\in\mathcal{E}_{t-1,t}} \exp\{w_f\}} \\
&\gtrsim \sum_{e\in\mathcal{E}_{t-1,t}} \frac{\exp\{w_e\}}{\sum_{f\in\mathcal{E}_{t-1,t}} \exp\{w_f\}} \log \frac{1}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \\
&= -\sum_{e\in\mathcal{E}_{t-1,t}} \frac{\exp\{w_e\}}{\sum_{f\in\mathcal{E}_{t-1,t}} \exp\{w_f\}} \log H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\} \\
&= -\sum_{e\in\mathcal{E}} \frac{H_{t-1,e} H_{t,e} \exp\{w_e\}}{(H_{t-1,\cdot} \odot H_{t,\cdot})^\mathsf{T} \exp\{w\}} \log H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}. \qquad (4)
\end{aligned}$$

Equation (4) is a convex combination of negative log-sum-exp terms, and is therefore concave in u_i.

Note that applying the same approximation to eq. (2) results in the same form as eq. (4), assuming a phantom initial transition point from $t = -1$ as described above. Combining terms, we achieve the lower-bound approximation:

$$f(u_i) \gtrsim \phi(u_i) = \sum_{y\in\mathcal{Y}_i} \left(\sum_{t=0}^{T_y} u_i^\mathsf{T} v_t - \sum_{e\in\mathcal{E}} \frac{H_{t-1,e} H_{t,e} \exp\{w_e\}}{(H_{t-1,\cdot} \odot H_{t,\cdot})^\mathsf{T} \exp\{w\}} \log \sum_{k\in e} \exp\{u_i^\mathsf{T} v_k + b_k\} \right) \qquad (5)$$

Finally, the gaussian prior on $u_i \sim \mathcal{N}(0, \sigma_u^2 I)$ results in a quadratic penalty of $g(u) = -\frac{1}{2\sigma^2} \|u_i\|^2$.

The total surrogate objective, $\phi(u) + g(u)$ is differentiable and concave, and can be easily optimized by quasi-newton methods such as L-BFGS. Note that due to independence among users, all user vectors can be optimized in parallel.

The optimization can be streamlined by pre-computing the edge weight distribution for each observed bigram $(t-1, t)$. Let

$$p_e^t := \frac{H_{t-1,e} H_{t,e} \exp\{w_e\}}{(H_{t-1,\cdot} \odot H_{t,\cdot})^\mathsf{T} \exp\{w\}} \quad \Rightarrow \quad p^t \propto H_{t-1,\cdot} \odot H_{t,\cdot} \odot \exp\{w\},$$

with $H_{-1,\cdot} = \mathbf{1}^\mathsf{T}$ so that $p^0 \propto H_{0,\cdot} \odot \exp\{w\}$. Then the optimization can be written as

$$\begin{aligned}
g(u_i) + \phi(u_i) &= -\frac{1}{2\sigma_u^2} \|u_i\|^2 + \sum_{y\in\mathcal{Y}_i} \left(\sum_{t=0}^{T_y} u_i^\mathsf{T} v_t - \sum_{e\in\mathcal{E}} p_e^t \log \sum_{k\in e} \exp\{u_i^\mathsf{T} v_k + b_k\} \right) \qquad (6) \\
&= -\frac{1}{2\sigma_u^2} \|u_i\|^2 + \sum_{y\in\mathcal{Y}_i} \left(\sum_{t=0}^{T_y} u_i^\mathsf{T} v_t - \left(p^t\right)^\mathsf{T} \log H^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\} \right) \qquad (7)
\end{aligned}$$

with gradient

$$\nabla_{u_i} = -\frac{1}{\sigma_u^2} u_i + \sum_{y \in \mathcal{Y}_i} \left(\sum_{t=0}^{T_y} v_t - \sum_{e \in \mathcal{E}} p_e^t \sum_{k \in e} v_k \frac{\exp\{u_i^\mathsf{T} v_k + b_k\}}{\sum_{j \in e} \exp\{u_i^\mathsf{T} v_j + b_j\}} \right)$$
$$= -\frac{1}{\sigma_u^2} u_i + \sum_{y \in \mathcal{Y}_i} \left(\sum_{t=0}^{T_y} v_t - \sum_{e \in \mathcal{E}} p_e^t \sum_{k \in e} v_k \frac{\exp\{u_i^\mathsf{T} v_k + b_k\}}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \right). \tag{8}$$

For each user, the set of playlists (and bigrams) is likely to be small. Similarly, the p^t weights can be encoded as a sparse matrix for greater efficiency. Note also that the factor $\log \sum_{k \in e} \exp\{u_i^\mathsf{T} v_k + b_k\}$ is constant over the playlist summation; each bigram simply changes the weighting factor p_e^t. These values can therefore be cached while computing the objective (and gradient) at each iterate $u_i^{(\delta)}$.

The terms in the objective can loosely be interpreted as pulling the user vector toward the centroid of each edge, weighted by the user's affinity for the items in that edge.

3.2 V-step — U, b, w fixed

For the V-step, note that all approximations applied in the derivation of the U-step can be applied just as well, since they only affect the edge weight factors w and bias terms b.

Note that each song factor v_i appears in all terms of the MAP objective, due to the normalization over song selection and requirement that each potential transition has nonzero probability. Consequently, there is no obvious independence structure here, nor implicit parallelism. Instead, song factors can be optimized by cyclic block coordinate ascent on each v_i with all $V_{\setminus i}$ held fixed.

Recall that the data log-likehood can be lower-bounded by the sum over all users, playlists, and transitions:

$$\mathcal{L}(\mathcal{Y}; \Theta) \gtrsim \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \left(\sum_{t=0}^{T_y} u_i^\mathsf{T} v_t - \sum_{e \in \mathcal{E}} p_e^t \log \sum_{k \in e} \exp\{u_i^\mathsf{T} v_k + b_k\} \right)$$

This leads to the lower-bounding objective

$$f(V) \gtrsim \phi(V) \coloneqq \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \left(\sum_{t=0}^{T_y} u_i^\mathsf{T} V \mathsf{e}_t - \sum_{e \in \mathcal{E}} p_e^t \log \sum_{k \in \mathcal{X}} H_{k,e} \exp\{u_i^\mathsf{T} V \mathsf{e}_k + b_k\} \right)$$
$$= \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \left(\sum_{t=0}^{T_y} \operatorname{tr}\left(\mathsf{e}_t u_i^\mathsf{T} V\right) - \sum_{e \in \mathcal{E}} p_e^t \log \sum_{k \in \mathcal{X}} H_{k,e} \exp\left\{\operatorname{tr}\left(\mathsf{e}_k u_i^\mathsf{T} V\right) + b_k\right\} \right)$$

The term $u_i^\mathsf{T} v_t$ only counts if a user selects song v_j, and the normalization term only counts for edges which contain v_j. Again, the approximation is concave and differentiable, and amenable to optimization via L-BFGS.

The gradient of the penalized objective can be computed as

$$\nabla_V = -\frac{1}{\sigma_v^2} V + \sum_{i \in \mathcal{U}} u_i \left(\sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \mathsf{e}_t^\mathsf{T} - \sum_{e \in \mathcal{E}} p_e^t \sum_{j \in \mathcal{X}} \mathsf{e}_j^\mathsf{T} \cdot \frac{H_{j,e} \exp\left\{\operatorname{tr}\left(\mathsf{e}_j u_i^\mathsf{T} V\right) + b_j\right\}}{\sum_{k \in \mathcal{X}} H_{k,e} \exp\left\{\operatorname{tr}\left(\mathsf{e}_k u_i^\mathsf{T} V\right) + b_k\right\}} \right). \tag{9}$$

3.3 b-step — U, V, w fixed

For the b-step, we follow the same general approach as the U-step, although the initial approximation differs.

$$\begin{aligned}
\log \mathbf{P}\left[x_t \middle|\, x_{t-1}, u_i, V, b, w\right] &= \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{H_{t-1,\cdot}^{\mathsf{T}} \exp\{w\}} \frac{H_{t,e} \exp\{u_i^{\mathsf{T}} v_t + b_t\}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} \\
&= \cancel{u_i^{\mathsf{T}} v_t} + b_t + \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{H_{t-1,\cdot}^{\mathsf{T}} \exp\{w\}} \frac{H_{t,e}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} \\
&\eqsim b^{\mathsf{T}} \mathsf{e}_t + \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{\cancel{H_{t-1,\cdot}^{\mathsf{T}} \exp\{w\}}} \frac{H_{t,e}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} \\
&\eqsim b^{\mathsf{T}} \mathsf{e}_t + \log \sum_{e \in \mathcal{E}} \exp\{w_e\} \frac{H_{t-1,e} H_{t,e}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}}
\end{aligned} \tag{10}$$

Now, applying the same lower bounding technique, we get

$$\begin{aligned}
\log \sum_{e \in \mathcal{E}} \exp\{w_e\} \frac{H_{t-1,e} H_{t,e}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} &= \log \sum_{e \in \mathcal{E}_{t-1,t}} \frac{\exp\{w_e\}}{\sum_{f \in \mathcal{E}_{t-1,t}} \exp\{w_f\}} \frac{1}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} + \log \cancel{\sum_{f \in \mathcal{E}_{t-1,t}} \exp\{w_f\}} \\
&\gtrsim \sum_{e \in \mathcal{E}} p_e^t \log \sum_{k \in e} \exp\{u_i^{\mathsf{T}} v_k + b^{\mathsf{T}} \mathsf{e}_k\}.
\end{aligned} \tag{11}$$

This results in the surrogate objective function

$$\phi(b) := \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} b^{\mathsf{T}} \mathsf{e}_t - \sum_{e \in \mathcal{E}} p_e^t \log \sum_{k \in e} \exp\{u_i^{\mathsf{T}} v_k + b^{\mathsf{T}} \mathsf{e}_k\}. \tag{12}$$

This function, again, is concave in b. Including the prior penalty, the gradient is computed as

$$\begin{aligned}
\nabla_b &= -\frac{1}{\sigma_b^2} b + \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \left(\mathsf{e}_t - \sum_{e \in \mathcal{E}} p_e^t \sum_{k \in e} \mathsf{e}_k \frac{\exp\{u_i^{\mathsf{T}} v_k + b^{\mathsf{T}} \mathsf{e}_k\}}{\sum_{j \in e} \exp\{u_i^{\mathsf{T}} v_j + b^{\mathsf{T}} \mathsf{e}_j\}} \right) \\
&= -\frac{1}{\sigma_b^2} b + \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \left(\mathsf{e}_t - \sum_{e \in \mathcal{E}} p_e^t \frac{H_{\cdot,e} \odot \exp\{u_i^{\mathsf{T}} V + b\}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} \right).
\end{aligned} \tag{13}$$

3.4 w-step — U, V, b fixed

Holding U, V, b constant and updating w, the bigram likelihood looks as follows:

$$\begin{aligned}
\log \mathbf{P}\left[x_t \middle|\, x_{t-1}, u_i, V, b, w\right] &= \log \sum_{e \in \mathcal{E}} \mathbf{P}\left[e \middle|\, x_{t-1}, w\right] \mathbf{P}\left[x_t \middle|\, e, u_i, V, b\right] \\
&= \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{H_{t-1,\cdot}^{\mathsf{T}} \exp\{w\}} \frac{H_{t,e} \exp\{u_i^{\mathsf{T}} v_t + b_t\}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} \\
&= \log \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} \exp\{w_e\}}{H_{t-1,\cdot}^{\mathsf{T}} \exp\{w\}} \frac{H_{t,e} \cancel{\exp\{u_i^{\mathsf{T}} v_t + b_t\}}}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} \\
&\eqsim \log \sum_{e \in \mathcal{E}_{t-1,t}} \exp\{w_e\} \frac{1}{H_{\cdot,e}^{\mathsf{T}} \exp\{u_i^{\mathsf{T}} V + b\}} - \log H_{t-1,\cdot}^{\mathsf{T}} \exp\{w\}
\end{aligned} \tag{14}$$

The second term in eq. (14) is a negative log-sum-exp, and is therefore concave in w. However, the first term is convex. In principal, this can be solved by DC programming, or by direct optimization as in the previous work. Here, we derive a concave lower bound using the same techniques as before.

Let $q_e := (H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\})^{-1}$. Then

$$\begin{aligned}
\log \sum_{e \in \mathcal{E}_{t-1,t}} \frac{1}{H_{\cdot,e}^\mathsf{T} \exp\{u_i^\mathsf{T} V + b\}} \exp\{w_e\} &= \log \sum_{e \in \mathcal{E}_{t-1,t}} q_e \exp\{w_e\} \\
&= \log \sum_{e \in \mathcal{E}_{t-1,t}} \frac{q_e}{\sum_{f \in \mathcal{E}_{t-1,t}} q_f} \exp\{w_e\} + \log \sum_{f \in \mathcal{E}_{t-1,t}} q_f \\
&\gtrsim \sum_{e \in \mathcal{E}_{t-1,t}} \frac{q_e}{\sum_{f \in \mathcal{E}_{t-1,t}} q_f} \log \exp\{w_e\} \\
&= \sum_{e \in \mathcal{E}_{t-1,t}} \frac{q_e}{\sum_{f \in \mathcal{E}_{t-1,t}} q_f} w_e \\
&= \sum_{e \in \mathcal{E}} \frac{H_{t-1,e} H_{t,e} q_e}{(H_{t-1,\cdot} \odot H_{t,\cdot})^\mathsf{T} q} w_e
\end{aligned} \tag{15}$$

So the final surrogate objective is

$$\phi(w) := \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \left(\sum_{e \in \mathcal{E}} \frac{H_{t-1,e} H_{t,e} q_e}{(H_{t-1,\cdot} \odot H_{t,\cdot})^\mathsf{T} q} w^\mathsf{T} \mathsf{e}_e - \log H_{t-1,\cdot}^\mathsf{T} \exp\{w\} \right), \tag{16}$$

which is concave in w.

Including the prior penalty, the gradient is computed as

$$\nabla_w = -\frac{1}{\sigma_w^2} w + \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \sum_{e \in \mathcal{E}} \mathsf{e}_e H_{t-1,e} \left(\frac{H_{t,e} q_e}{(H_{t-1,\cdot} \odot H_{t,\cdot})^\mathsf{T} q} - \frac{\exp\{w^\mathsf{T} \mathsf{e}_e\}}{\sum_{f \in \mathcal{E}} H_{t-1,f} \exp\{w^\mathsf{T} \mathsf{e}_f\}} \right) \tag{17}$$

4 Model restrictions

Here, we spell out the surrogate objectives for various special cases of the model.

4.1 No bias, no personalization

Fix $b = 0$, $u_i = 0$ for all users. The latent factor terms V disappear, and all that's left is the w term.

The w objective remains relatively unchanged, except that the q_e factors are now defined as

$$q_e := \frac{1}{|e|},$$

since each song in e has equal probability of being selected.

4.2 Bias, no personalization

Fix $u_i = 0$ for all users. Then the latent factor terms V all disappear from the model as well. The result is a popularity-adaptive version of the hypergraph model.

$$\phi(b) := \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} b^\mathsf{T} \mathsf{e}_t - \sum_{e \in \mathcal{E}} p_e^t \log \sum_{k \in e} \exp\{b^\mathsf{T} \mathsf{e}_k\} \tag{18}$$

The w updates are the same, but now with the q_e factors defined as

$$q_e := \frac{1}{H_{\cdot,e}^{\mathsf{T}} \exp\{b\}},$$

that is, the total amount of (unnormalized) song mass in edge e.

4.3 No features

If we eliminate all edges from the hypergraph except for e_0, then the result is a stochastic latent factor playlist model, along the lines of BPR-MF. All p_e^t terms become 1, and the edge weights w_e disappear from the model. Note that in this case, bigram transitions disappear completely, and the lower-bounding objective matches the data log-likelihood exactly (modulo constants).

$$\phi(u_i) := \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \left(u_i^{\mathsf{T}} v_t - \log \sum_{k \in \mathcal{X}} \exp\{u_i^{\mathsf{T}} v_k + b_k\} \right) \tag{19}$$

$$\phi(v_j) := \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \left(u_i^{\mathsf{T}} v_t \, [\![x_t = x_j]\!] - \log \sum_{k \in \mathcal{X}} \exp\{u_i^{\mathsf{T}} v_k + b_k\} \right) \tag{20}$$

$$\phi(b) := \sum_{i \in \mathcal{U}} \sum_{y \in \mathcal{Y}_i} \sum_{t=0}^{T_y} \left(b^{\mathsf{T}} \mathsf{e}_t - \log \sum_{k \in \mathcal{X}} \exp\{u_i^{\mathsf{T}} v_k + b^{\mathsf{T}} \mathsf{e}_k\} \right) \tag{21}$$