CSE 291B 1/20

Non-negative matrix factorization (NMF)

How to approximate $X \simeq VY$ for $X, V, Y \geq 0$?

- Types of approximation error:

$$\mathcal{E}_{LS} = \| X - VY \|^2 = \sum_{in} (X_{in} - (VY)_{in})^2$$

$$\mathcal{E}_{KL} = \sum_{in} \left[X_{in} \log \frac{X_{in}}{(VY)_{in}} - X_{in} + (VY)_{in} \right]$$

Minimization of $\mathcal{E}_{LS}$

- Decomposition

$$\mathcal{E}^+_{LS} = \|X\|^2 + \|VY\|^2$$

$$\mathcal{E}^-_{LS} = 2 \text{ trace } [XVY]$$

$$\mathcal{E}_{LS} = \mathcal{E}^+_{LS} - \mathcal{E}^-_{LS}$$

- Multiplicative update

$$V_{i\alpha} \leftarrow V_{i\alpha} \left[\frac{\left(\frac{\partial \mathcal{E}^-_{LS}}{\partial V_{i\alpha}} \right)}{\left(\frac{\partial \mathcal{E}^+_{LS}}{\partial V_{i\alpha}} \right)} \right]$$

- Properties of V-update
  (1) preserves non-negativity because $V_{i\alpha}, \frac{\partial \mathcal{E}^{\pm}_{LS}}{\partial V_{i\alpha}} \geq 0$

  (2) increases $V_{i\alpha}$ when $\frac{\partial \mathcal{E}_{LS}}{\partial V_{i\alpha}} \leq 0$ because

  (3) decreases $V_{i\alpha}$ when $\frac{\partial \mathcal{E}_{LS}}{\partial V_{i\alpha}} \geq 0$

  $$\frac{\partial \mathcal{E}_{LS}}{\partial V_{i\alpha}} = \frac{\partial \mathcal{E}^+_{LS}}{\partial V_{i\alpha}} - \frac{\partial \mathcal{E}^-_{LS}}{\partial V_{i\alpha}}$$

  (4) fixed points at $V_{i\alpha} = 0$ or $\frac{\partial \mathcal{E}_{LS}}{\partial V_{i\alpha}} = 0$

  (5) no learning rate

  (6) monotonically decreases $\mathcal{E}_{LS}$ (prove later)

- Multiplicative update of Y

$$Y_{\alpha n} \leftarrow Y_{\alpha n} \left[\frac{\left(\frac{\partial \mathcal{E}^-_{LS}}{\partial Y_{\alpha n}} \right)}{\left(\frac{\partial \mathcal{E}^+_{LS}}{\partial Y_{\alpha n}} \right)} \right]$$

Analogous properties hold by symmetry.

Minimization of $\mathcal{E}_{KL}$

- Decomposition

$$\mathcal{E}_{KL}^{+} = \sum_{in} \left[X_{in} \log X_{in} + (VY)_{in} \right]$$

$$\mathcal{E}_{KL}^{-} = \sum_{in} \left[X_{in} \log (VY)_{in} + X_{in} \right]$$

$$\mathcal{E}_{KL} = \mathcal{E}_{KL}^{+} - \mathcal{E}_{KL}^{-}$$

- Non-negative gradients

$$\frac{\partial \mathcal{E}_{KL}^{+}}{\partial V_{i\alpha}} = \sum_{n} Y_{\alpha n} \geq 0$$

$$\frac{\partial \mathcal{E}_{KL}^{-}}{\partial V_{i\alpha}} = \sum_{n} X_{in} \frac{Y_{\alpha n}}{(VY)_{in}} \geq 0$$

similar for $\frac{\partial \mathcal{E}_{KL}^{\pm}}{\partial Y_{\alpha n}}$ ...

- Multiplicative update

$$V_{i\alpha} \longleftarrow V_{i\alpha} \left[\frac{\left(\frac{\partial \mathcal{E}_{KL}^{-}}{\partial V_{i\alpha}} \right)}{\left(\frac{\partial \mathcal{E}_{KL}^{+}}{\partial V_{i\alpha}} \right)} \right]$$

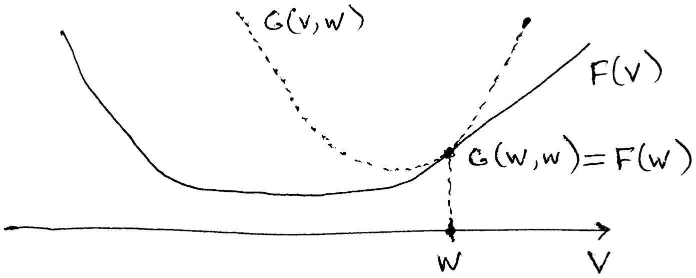
As before, updates converge monotonically to local minimum of $\mathcal{E}_{KL}$

Proofs of convergence

- Def: $G(v,w)$ is an <u>auxiliary function</u> for $F(v)$ if:

  (i) $G(v,v) = F(v)$ for all v, ~~w~~

  (ii) $G(v,w) \geq F(v)$ for all v, w

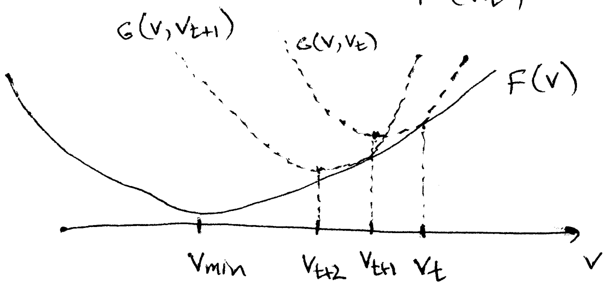


- Thm: if $G(v,w)$ is an auxiliary function for $F(v)$, then $F(v)$ is non-increasing under the update

$$\boxed{v_{t+1} = \underset{v}{\operatorname{argmin}}\ G(v, v_t)}$$

<u>Proof:</u>

$$\begin{aligned} F(v_{t+1}) &\leq G(v_{t+1}, v_t) && \text{by property (ii)} \\ &\leq G(v_t, v_t) && \text{b/c } v_{t+1} = \underset{v}{\operatorname{argmin}}\ G(v, v_t) \\ &= F(v_t) && \text{by property (i)} \end{aligned}$$

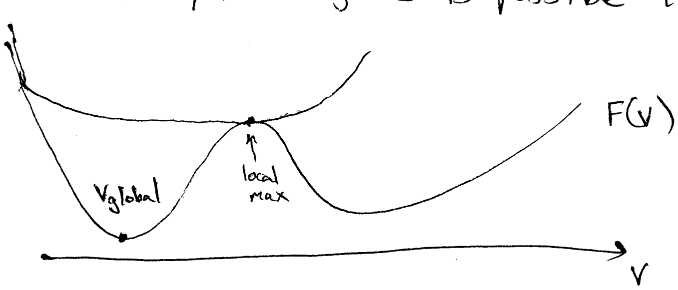


By iteration:

$$F(v_0) \geq F(v_1) \cdots\cdots F(v_t) \geq F(v_{t+1}) \geq \cdots \geq F(v_{final})$$

In practice, for smooth $F(v)$ and $G(v,w)$, updates converge to local (not necessarily global) minimum.

In theory, convergence is possible to any stationary point.



converging to local maximum... this is rare

- Auxiliary function for $\mathcal{E}_{KL}$

Let $F(V) = \sum_{in} \left[X_{in} \log \frac{X_{in}}{(VY)_{in}} - X_{in} + (VY)_{in} \right]$

Let $G(V,W) = \sum_{in} \left[X_{in} \log \frac{X_{in}}{(WY)_{in}} - X_{in} + (VY)_{in} \underbrace{- X_{in} \sum_{\alpha} \frac{W_{i\alpha} Y_{\alpha n}}{(WY)_{in}} \log \left(\frac{V_{i\alpha}}{W_{i\alpha}} \right)}_{\text{last term on RHS vanishes if } V=W} \right]$

Trivially: (i) $G(V,V) = F(V)$

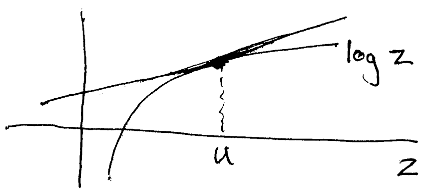
How to prove $G(V,W) \geq F(V)$?

- Lemma: <u>Jensen's inequality</u>

Let z be a nonnegative random variable.

Then: $E[\log z] \leq \log E[z]$

- Proof of lemma:



log is a concave function

$\log z \leq \log u + \frac{1}{u}(z-u)$ FOR ALL $z, u \geq 0$

$$
\begin{aligned}
E[\log z] &\leq E\left[\log u + \frac{1}{u}(z-u)\right] && \text{for all constant } u \\
&= \log u + \frac{1}{u}(E[z]-u) && \text{by linearity of } E[\,] \\
&= \log E[z] && \text{by setting } u = E[z]
\end{aligned}
$$

Equivalently: if $p_\alpha \geq 0$ and $\sum_\alpha p_\alpha = 1$,

then $\sum_\alpha p_\alpha \log z_\alpha \leq \log \sum_\alpha p_\alpha z_\alpha$

- Proof of auxiliary function bound

$$G(v,w) = \sum_{in}\left[X_{in}\log\frac{X_{in}}{(WY)_{in}} - X_{in} + (VY)_{in} - X_{in}\sum_{\alpha}\frac{W_{i\alpha}Y_{\alpha n}}{(WY)_{in}}\log\left(\frac{V_{i\alpha}}{W_{i\alpha}}\right)\right]$$

Apply Jensen's inequality with $p_\alpha = \frac{W_{i\alpha}Y_{\alpha n}}{(WY)_{in}}$ and $z_\alpha = \frac{V_{i\alpha}}{W_{i\alpha}}$

$$G(v,w) \geq \sum_{in}\left[X_{in}\log\frac{X_{in}}{(WY)_{in}} - X_{in} + (VY)_{in} - X_{in}\log\sum_{\alpha}\frac{\cancel{W_{i\alpha}}Y_{\alpha n}V_{i\alpha}}{(WY)_{in}\cancel{W_{i\alpha}}}\right]$$

$$= \sum_{in}\left[X_{in}\log\frac{X_{in}}{(WY)_{in}} - X_{in}\log\frac{(VY)_{in}}{(WY)_{in}} - X_{in} + (VY)_{in}\right]$$

$$= \sum_{in}\left[X_{in}\log\frac{X_{in}}{(VY)_{in}} - X_{in} + (VY)_{in}\right]$$

$$= F(v)$$

- Update from $G(v,w)$

How to minimize $G(v,w)$ for fixed w?

$$\frac{\partial G}{\partial V_{i\alpha}} = \sum_n Y_{\alpha n} - \sum_n X_{in}\frac{W_{i\alpha}Y_{\alpha n}}{(WY)_{in}}\cdot\frac{1}{V_{i\alpha}}$$

$$\frac{\partial G}{\partial V_{i\alpha}} = 0 \longrightarrow V_{i\alpha} = W_{i\alpha}\left[\frac{\sum_n X_{in}Y_{\alpha n}/(WY)_{in}}{\sum_n Y_{\alpha n}}\right]$$

Update rule:

$$v_{t+1} = \operatorname*{argmin}_v G(v, v_t)$$

$$V_{i\alpha} \leftarrow V_{i\alpha}\left[\frac{\sum_n \frac{X_{in}Y_{\alpha n}}{(VY)_{in}}}{\sum_n Y_{\alpha n}}\right]$$

<u>To think about</u>

- Given $G(v,w)$ and $F(v)$, it is easy to verify that (i) $G(v,v) = F(v)$ and (ii) $G(v,w) \geq F(v)$
- Given a cost function $F(v)$, how to derive an auxiliary function $G(v,w)$?
- In addition to properties (i)–(ii), the update $v_{t+1} = \operatorname*{argmin}_v G(v, v_t)$ is easy to compute (in closed form).