Neural networks

Training CRFs - unary log-factor gradient

MACHINE LEARNING

Topics: stochastic gradient descent (SGD)

- Algorithm that performs updates after each example
  - initialize $\boldsymbol{\theta}$
  - for N iterations
    - for each training example $(\mathbf{X}^{(t)}, \mathbf{y}^{(t)})$
      - ✓ $\Delta = -\nabla_{\boldsymbol{\theta}} l(\mathbf{f}(\mathbf{X}^{(t)}; \boldsymbol{\theta}), \mathbf{y}^{(t)}) - \lambda \nabla_{\boldsymbol{\theta}} \Omega(\boldsymbol{\theta})$
      - ✓ $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\, \Delta$

training epoch
=
iteration over **all** examples

- To apply this algorithm to a CRF, we need
  - the loss function $l(\mathbf{f}(\mathbf{X}^{(t)}; \boldsymbol{\theta}), \mathbf{y}^{(t)})$
  - a procedure to compute the parameter gradients $\nabla_{\boldsymbol{\theta}} l(\mathbf{f}(\mathbf{X}^{(t)}; \boldsymbol{\theta}), \mathbf{y}^{(t)})$
  - the regularizer $\Omega(\boldsymbol{\theta})$ (and the gradient $\nabla_{\boldsymbol{\theta}} \Omega(\boldsymbol{\theta})$)
  - initialization method

LOSS FUNCTION

Topics: loss function for sequential classification with CRF

- CRF estimates $p(\mathbf{y}|\mathbf{X})$
  - we could maximize the probabilities of $\mathbf{y}^{(t)}$ given $\mathbf{X}^{(t)}$ in the training set

- To frame as minimization, we minimize the negative log-likelihood

$$l(\mathbf{f}(\mathbf{X}), \mathbf{y}) = -\log p(\mathbf{y}|\mathbf{X})$$

  - unlike for non-sequential classification, we never explicitly compute the value of $p(\mathbf{y}|\mathbf{X})$ for all values of $\mathbf{y}$

PARAMETER GRADIENTS

Topics: loss gradient at unary log-factors

- Partial derivative wrt $a_u(y_k')$:

$$\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y_k')} = -(1_{y_k=y_k'} - p(y_k'|\mathbf{X}))$$

- Gradient for each unary (log-)factors:

$$\begin{aligned}
\nabla_{\mathbf{a}^{(L+1,0)}(\mathbf{x}_k)} - \log p(\mathbf{y}|\mathbf{X}) &= -(\mathbf{e}(y_k) - \mathbf{p}(y_k|\mathbf{X})) \\
\nabla_{\mathbf{a}^{(L+1,-1)}(\mathbf{x}_{k-1})} - \log p(\mathbf{y}|\mathbf{X}) &= -1_{k>1}\,(\mathbf{e}(y_k) - \mathbf{p}(y_k|\mathbf{X})) \\
\nabla_{\mathbf{a}^{(L+1,+1)}(\mathbf{x}_{k+1})} - \log p(\mathbf{y}|\mathbf{X}) &= -1_{k<K}\,(\mathbf{e}(y_k) - \underbrace{\mathbf{p}(y_k|\mathbf{X})}_{\text{vector of all marginal probabilities}})
\end{aligned}$$

$$\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} \quad = \quad \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right)$$

$$\begin{aligned}
\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} &= \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right) \\
&= -\left(1_{y_k = y'_k} - \frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) \right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} &= \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right) \\
&= - \left(1_{y_k = y'_k} - \frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) \right)
\end{aligned}$$

$$\frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} Z(\mathbf{X})$$

$$\begin{aligned}
\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} &= \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right) \\
&= - \left(1_{y_k = y'_k} - \frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) \right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) &= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} Z(\mathbf{X}) \\
&= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} &= \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right) \\
&= - \left(1_{y_k = y'_k} - \frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) &= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} Z(\mathbf{X}) \\
&= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \frac{1}{Z(\mathbf{X})} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \frac{\partial}{\partial a_u(y'_k)} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} &= \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right) \\
&= - \left(1_{y_k = y'_k} - \frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) &= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} Z(\mathbf{X}) \\
&= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \frac{1}{Z(\mathbf{X})} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \frac{\partial}{\partial a_u(y'_k)} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \frac{1}{Z(\mathbf{X})} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} 1_{y'_k = y''_k} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} &= \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right) \\
&= - \left(1_{y_k = y'_k} - \frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) &= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} Z(\mathbf{X}) \\
&= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \exp\left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \frac{1}{Z(\mathbf{X})} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \frac{\partial}{\partial a_u(y'_k)} \exp\left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \frac{1}{Z(\mathbf{X})} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} 1_{y'_k = y''_k} \exp\left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} 1_{y'_k = y''_k} \; p(y''_1, \ldots, y''_K | \mathbf{X})
\end{aligned}
$$

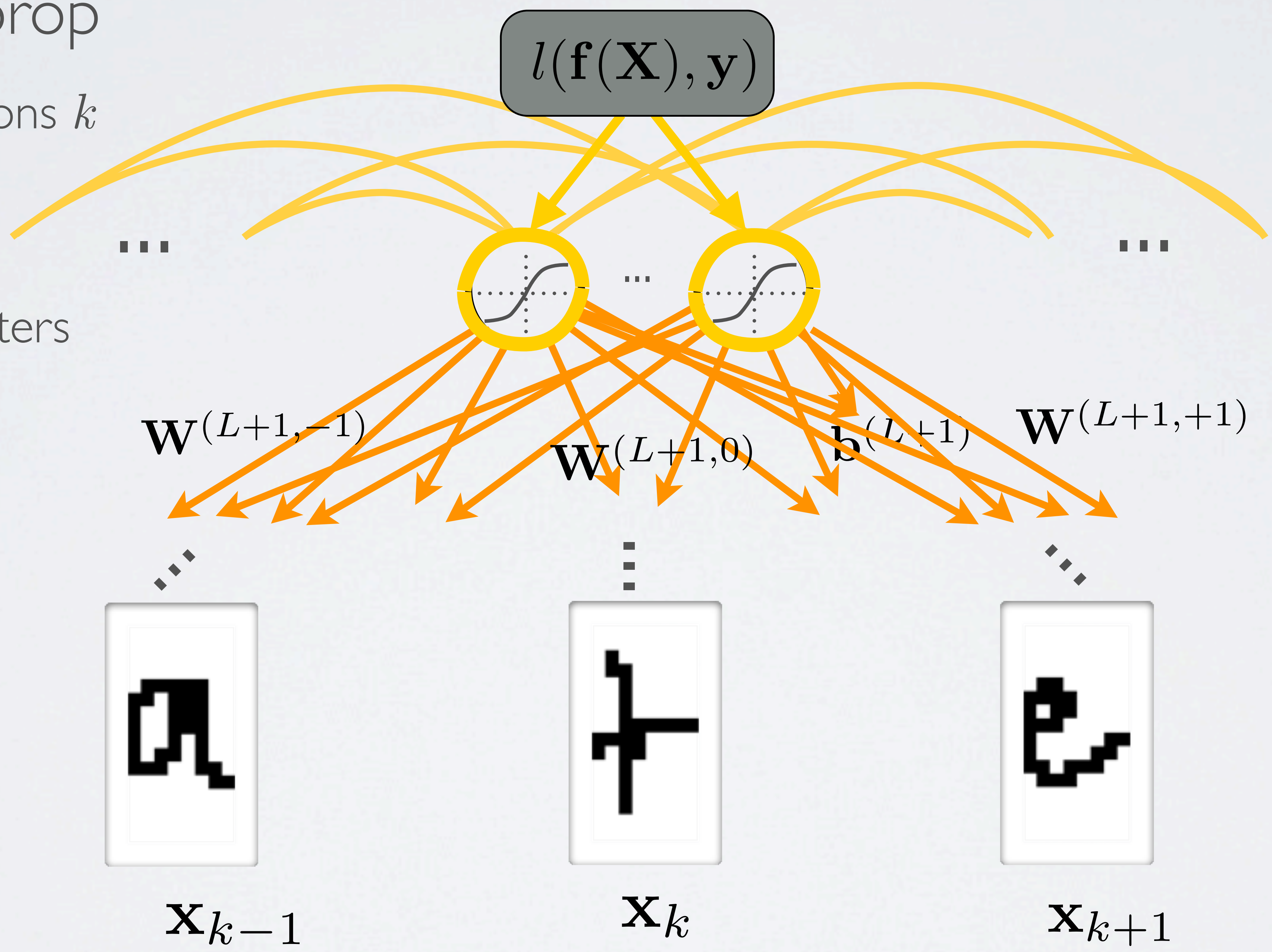
$$
\begin{aligned}
\frac{\partial - \log p(\mathbf{y}|\mathbf{X})}{\partial a_u(y'_k)} &= \frac{\partial}{\partial a_u(y'_k)} - \left(\left(\sum_{k=1}^{K} a_u(y_k) + \sum_{k=1}^{K-1} a_p(y_k, y_{k+1}) \right) - \log Z(\mathbf{X}) \right) \\
&= - \left(1_{y_k = y'_k} - \frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial a_u(y'_k)} \log Z(\mathbf{X}) &= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} Z(\mathbf{X}) \\
&= \frac{1}{Z(\mathbf{X})} \frac{\partial}{\partial a_u(y'_k)} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \frac{1}{Z(\mathbf{X})} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} \frac{\partial}{\partial a_u(y'_k)} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \frac{1}{Z(\mathbf{X})} \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} 1_{y'_k = y''_k} \exp \left(\sum_{k=1}^{K} a_u(y''_k) + \sum_{k=1}^{K-1} a_p(y''_k, y''_{k+1}) \right) \\
&= \sum_{y''_1} \sum_{y''_2} \cdots \sum_{y''_K} 1_{y'_k = y''_k} \; p(y''_1, \ldots, y''_K | \mathbf{X}) \\
&= p(y'_k | \mathbf{X})
\end{aligned}
$$

PARAMETER GRADIENTS

Topics: loss gradient at unary log-factor parameters

- Use regular backprop
  - backprop at all positions k
  - accumulate all gradients, from every position, into parameters



PARAMETER GRADIENTS

Topics: loss gradient at unary log-factor parameters

- For linear log-factors:
  - the log-factors are directly connected to the input:

$$\mathbf{a}^{(1,0)}(\mathbf{x}_k) = \mathbf{b}^{(1)} + \mathbf{W}^{(1,0)}\mathbf{x}_k$$

$$\mathbf{a}^{(1,-1)}(\mathbf{x}_k) = \mathbf{W}^{(1,-1)}\mathbf{x}_k$$

$$\mathbf{a}^{(1,+1)}(\mathbf{x}_k) = \mathbf{W}^{(1,+1)}\mathbf{x}_k$$

PARAMETER GRADIENTS

Topics: loss gradient at unary log-factor parameters

- For linear log-factors:
  - the gradients are:

$$\nabla_{\mathbf{b}^{(1)}} - \log p(\mathbf{y}|\mathbf{X}) = \sum_{k=1}^{K} \left(\nabla_{\mathbf{a}^{(1,0)}(\mathbf{x}_k)} - \log p(\mathbf{y}|\mathbf{X})\right) \quad = \quad \sum_{k=1}^{K} -(\mathbf{e}(y_k) - \mathbf{p}(y_k|\mathbf{X}))$$

$$\nabla_{\mathbf{W}^{(1,0)}} - \log p(\mathbf{y}|\mathbf{X}) = \sum_{k=1}^{K} \left(\nabla_{\mathbf{a}^{(1,0)}(\mathbf{x}_k)} - \log p(\mathbf{y}|\mathbf{X})\right)\ \mathbf{x}_k^\top \quad = \quad \sum_{k=1}^{K} -(\mathbf{e}(y_k) - \mathbf{p}(y_k|\mathbf{X}))\, \mathbf{x}_k^\top$$

$$\nabla_{\mathbf{W}^{(1,-1)}} - \log p(\mathbf{y}|\mathbf{X}) = \sum_{k=2}^{K} \left(\nabla_{\mathbf{a}^{(1,-1)}(\mathbf{x}_k)} - \log p(\mathbf{y}|\mathbf{X})\right)\ \mathbf{x}_{k-1}^\top \quad = \quad \sum_{k=2}^{K} -\left(\mathbf{e}(y_k) - \mathbf{p}(y_k|\mathbf{X})\right)\ \mathbf{x}_{k-1}^\top$$

$$\nabla_{\mathbf{W}^{(1,+1)}} - \log p(\mathbf{y}|\mathbf{X}) = \sum_{k=1}^{K-1} \left(\nabla_{\mathbf{a}^{(1,+1)}(\mathbf{x}_k)} - \log p(\mathbf{y}|\mathbf{X})\right)\ \mathbf{x}_{k+1}^\top \quad = \quad \sum_{k=1}^{K-1} -\left(\mathbf{e}(y_k) - \mathbf{p}(y_k|\mathbf{X})\right)\ \mathbf{x}_{k+1}^\top$$