BMML: Deep Structured Models Supplementary Material

Ashuha Arseniy

March 31, 2016

There is a Supplementary Material for my talk, probably it's contain some mistakes and confused moments. I tried to prove most difficult moments from article "Learning Deep Structured Models". Maybe I am writing this text in order to structure my knowledge and improve understanding of this article.

Let's Introduce some definitions:

1. $(x, y) \in D$ – dataset
2. $Y \subset \{y_1, y_2, ..., y_n\}$ – set of all possible configurations
3. $F(x, y; w)$ – get score of configuration y for object x given w
4. $p_{(x,y)}(\hat{y}|w) \propto exp(F(x, \hat{y}; w))$ – distribution over configurations
5. $Z(x, w) = \sum_y exp(F(x, y; w)$ – normalization constant
6. $P(r) = \{p \in Y : r \subset p\}$ – subset of parent configurations, which subsumes those regions for which we want the marginalization constraint to hold
7. $C(r) = \{c \in Y : r \in P(c)\}$ – subset of children configurations
8. $\delta(y' = y) = [y' = y]$

1 Deep Structured Models

Statement 1. *Define L as* $\sum_{(x,y)\in D} \left(log \sum_{y'\in Y} exp\ F(x, y'; w) - F(x, y; w) \right)$ *then gradient* $\frac{\partial}{\partial w} L$ *is given by transformed difference between model and target distribution* $\sum_{(x,y)\in D} \sum_{y'\in Y} (p_{(x,y)}(y'|w) - \delta(y' = y)) \frac{\partial}{\partial w} F(x, y'; w)$.

Proof

$$\frac{\partial}{\partial w} \sum_{(x,y)\in D} \left(log \sum_{y'\in Y} \exp\ F(x, y'; w) - F(x, y; w) \right) =$$

$$= \sum_{(x,y)\in D} \left(\frac{1}{\sum_{y''\in Y} \exp\ F(x, y''; w)} \frac{\partial}{\partial w} \sum_{y'\in Y} \exp\ F(x, y'; w) - \frac{\partial}{\partial w} F(x, y; w) \right) =$$

$$= \sum_{(x,y)\in D} \left(\sum_{y'\in Y} \frac{\exp(F(x, y'; w))}{\sum_{y''\in Y} exp\ F(x, y''; w)} \frac{\partial}{\partial w} F(x, y'; w) - \frac{\partial}{\partial w} F(x, y; w) \right) =$$

$$= \sum_{(x,y)\in D} \left(\sum_{y'\in Y} p_{(x,y)}(y'|w) \frac{\partial}{\partial w} F(x, y'; w) - \frac{\partial}{\partial w} F(x, y; w) \right) = \sum_{(x,y)\in D, y'} (p_{(x,y)}(y'|w) - \delta(y' = y)) \frac{\partial}{\partial w} F(x, y'; w)$$

end

Statement 2. *If we assume that is decomposed as a* $F(x, y; w) = \sum_r f_r(x, y; w)$ *that the gradient of previous one is equal*

$$\nabla_w \sum_{(x,y)\in D} \left(log \sum_{y'\in Y} exp\ F(x, y', w) - F(x, y, w) \right) = \sum_{(x,y)\in D, y'_r, r} (p_{(x,y),r}(y'_r|w) - \delta(y'_r = y_r)) \frac{\partial}{\partial w} f_r(x, y'_r, w)$$

Proof It can be proved like a [Statement 1.], in this case $p_{(x,y),r}(y'_r|w)$ means marginal distribution over subset of configurations y_r. **end**

2 Efficient Approximate Learning of DSM

Statement 3. *We can represent ln Z as*

$$ln\ Z(x, w) = \sum_y exp\ F(x, y, w) = \max_{p_{(x,y)}(\hat{y})} \mathbb{E}_{p_{(x,y)}(\hat{y})} F(x, \hat{y}; w) + H(p_{(x,y)}(\hat{y}))$$

Proof There exist really true distribution $p_{(x,y)}$, we can note that

$$\min_{p_{(x,y)}} D_{KL} \left(p_{(x,y)}(\hat{y}) || \frac{1}{Z(x, w)} \exp(F(x, \hat{y}; w)) \right) = 0$$

$$D_{KL} \left(p_{(x,y)}(\hat{y}) || \frac{1}{Z(x, w)} \exp(F(x, \hat{y}; w)) \right) = \sum_{\hat{y}_i} p_{(x,y)}(\hat{y}_i) \log \frac{p_{(x,y)}(\hat{y}_i) \cdot Z}{\exp(F(x, \hat{y}_i; w))} =$$

$$= \sum_{\hat{y}_i} p_{(x,y)}(\hat{y}_i)(\log p_{(x,y)}(\hat{y}_i) + \log Z - F(x, \hat{y}_i; w)) =$$

$$= \sum_{\hat{y}_i} p_{(x,y)}(\hat{y}_i) \log p_{(x,y)}(\hat{y}_i) + p_{(x,y)}(\hat{y}_i) \log Z - p_{(x,y)}(\hat{y}_i) F(x, \hat{y}_i; w)$$

$$= -H(p_{(x,y)}(\hat{y})) + 1 \cdot \log Z - \mathbb{E}_{p_{(x,y)}(\hat{y})} F(x, \hat{y}; w)$$

$$\min_{p_{(x,y)}} (-H(p_{(x,y)}(\hat{y})) + 1 \cdot \log Z - \mathbb{E}_{p_{(x,y)}(\hat{y})} F(x, \hat{y}; w)) = 0$$

$$\log Z = \max_{p_{(x,y)}} (\mathbb{E}_{p_{(x,y)}(\hat{y})} F(x, \hat{y}; w) + H(p_{(x,y)}(\hat{y})))$$

end

Statement 4. *Function* $\sum_{\hat{y}_r, r} b_{(x,y),r}(\hat{y}_r) f_r(x, \hat{y}_r; w) + \sum_r H(b_{(x,y),r}(\hat{y}_r))$ *is concave by* $b_{(x,y)}$ *that lay in convex set.*

$$b_{(x,y)} \in C_{(x,y)} = \begin{cases} b_{(x,y),r}(y_r) \geq 0 \quad \sum_{y_r} b_{(x,y),r}(y_r) = 1 \quad \forall r \\ b_{(x,y),r}(y_r) = \sum_{\hat{y}_p \setminus \hat{y}_r} p_{(x,y),p}(\hat{y}_p) \qquad \forall r, \hat{y}_r, p \in P(r) \end{cases}$$

Proof Fist part is a linear combination of $b_{(x,y),r}$ therefor we can don't hesitate about that, Entropy is concave function and $b_{(x,y)}$ is given from convex set $C_{(x,y)}$. $C_{(x,y)}$ is convex because constrained from linear restrictions. Fist restrictions is just a discrete distribution constraint, second ones is a marginalization constraint. **end**

Statement 5. *Regularity conditions are satisfied for*

$$\max_{b_{(x,y)} \in C_{(x,y)}} \left\{ \sum_{\hat{y}_r, r} b_{(x,y),r}(\hat{y}) f_r(x, y; w) + \sum_r H(b_{(x,y),r}(\hat{y}_r)) \right\}$$

Proof We have optimization concave function in convex set that constrain from linear restrictions, then no other condition is needed. **end**

Statement 6. *The Lagrangian of*

$$\max_{b_{(x,y)} \in C_{(x,y)}} \left\{ \sum_{\hat{y}_r, r} b_{(x,y),r}(\hat{y}) f_r(x, \hat{y}_r; w) + \sum_r H(b_{(x,y),r}(\hat{y}_r)) \right\}$$

is

$$L_{(x,y)} = \sum_{r, \hat{y}_r} b_{(x,y),r}(\hat{y}_r) \cdot \left(f_r(x, \hat{y}_r; w) + \sum_{c \in C(r)} \lambda_{(x,y),c \to r}(\hat{y}_c) - \sum_{p \in P(r)} \lambda_{(x,y),r \to p}(\hat{y}_r) \right) + \sum_r H(b_{(x,y),r}(\hat{y}_r))$$

Proof Redefine entropy function as barrier function if $b_{(x,y)}$ is not a distribution

$$H(b) = \begin{cases} -\sum_{\hat{y}_i} b(\hat{y}_i) \log b(\hat{y}_i) & b_{(x,y),r}(y_r) \geq 0 \;\; \sum_{y_r} b_{(x,y),r}(y_r) = 1 \quad \forall r \\ -\infty & else \end{cases}$$

Therefore we need only to introduce Lagrangian multilayer for each marginalization constant.

$$L_{(x,y)} = \sum_{r, \hat{y}_r} b_{(x,y),r}(\hat{y}_r) f_r(x, \hat{y}_r; w) + \sum_r H(b_{(x,y),r}(\hat{y}_r)) + \sum_{r, \hat{y}_r, p \in P(r)} \lambda_{(x,y),r \to p}(\hat{y}_r) \left(\sum_{y_p \setminus y_r} b_{(x,y),p}(\hat{y}_p) - b_{(x,y),r}(\hat{y}_r) \right) =$$

$$= \sum_{r, \hat{y}_r} b_{(x,y),r}(\hat{y}_r) f_r(x, \hat{y}_r; w) + \sum_r H(b_{(x,y),r}(\hat{y}_r)) + \sum_{r, \hat{y}_r, p \in P(r)} \sum_{y_p \setminus y_r} \lambda_{(x,y),r \to p}(\hat{y}_r) b_{(x,y),p}(\hat{y}_p) -$$

$$- \sum_{r, \hat{y}_r} b_{(x,y),r}(\hat{y}_r) \sum_{p \in P(r)} \lambda_{(x,y),r \to p}(\hat{y}_r) =$$

$$\ldots TODO$$

$$= \sum_{r, \hat{y}_r} b_{(x,y),r}(\hat{y}_r) \cdot \left(f_r(x, y; w) + \sum_{c \in C(r)} \lambda_{(x,y),c \to r}(\hat{y}_c) - \sum_{p \in P(r)} \lambda_{(x,y),r \to p}(\hat{y}_r) \right) + \sum_r H(b_{(x,y),r}(\hat{y}_r))$$

end

Statement 7. *For problem*

$$\min_w \sum_{(x,y) \in D} \left(\max_{b_{(x,y)} \in C_{(x,y)}} \left\{ \sum_r b_{(x,y),r}(\hat{y}) f_r(x, y; w) + H(b_{(x,y)}) \right\} - \sum_r f_r(x, y; w) \right)$$

Duality task is

$$\min_{w, \lambda} \sum_{(x,y),r} \ln \sum_{\hat{y}_r} exp \left(f_r(x, y; w) + \sum_{c \in C(r)} \lambda_{(x,y),c \to r}(\hat{y}_c) - \sum_{p \in P(r)} \lambda_{(x,y),r \to p}(\hat{y}_r) \right) - \sum_{(x,y) \in D} F(x, y; w)$$

Proof

$$\min_{w,\lambda} \sum_{(x,y),r} \max_{b_{(x,y),r}} \left\{ \sum_{r,\hat{y}_r} b_{(x,y),r}(\hat{y}_r) \cdot \left(f_r(x,y;w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) \right) + \sum_r H(b_{(x,y),r}(\hat{y}_r)) \right\}$$

$$-\sum_r f_r(x,y;w) = \min_{w,\lambda} \sum_{(x,y),r} \left(ln\ Z - \sum_r f_r(x,y;w) \right) =$$

$$= \min_{w,\lambda} \sum_{(x,y),r} \ln \sum_{\hat{y}_r} exp \left(f_r(x,y;w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) \right) - \sum_{(x,y)\in D} F(x,y;w)$$

end

3 Blending Learning

$$D(\lambda,w) = \sum_{(x,y),r} \ln \sum_{\hat{y}_r} exp \left(f_r(x,y;w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) \right) - \sum_{(x,y)} F(x,y;w) \to \min_{w,\lambda}$$

Let's define beliefs as

$$b_{(x,y),r}(\hat{y}_r) \propto exp \left(f_r(x,y;w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) \right)$$

Statement 8. *Gradient by* $\frac{\partial D}{\partial w}$ *is given by* $\sum_{(x,y),r,\hat{y}_r} b_{(x,y),r,\hat{y}_r} \frac{\partial}{\partial w} f_r(x,\hat{y}_r;w) + \sum_{(x,y)} \frac{\partial}{\partial w} F(x,y;w)$

Proof

$$\frac{\partial D}{\partial w} = \sum_{(x,y),r,\hat{y}_r} \frac{exp\left(f_r(x,y;w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) \right) \cdot \frac{\partial}{\partial w} f_r(x,y;w)}{\sum_{\hat{y}_r} exp\left(f_r(x,y;w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) \right)} - \sum_{(x,y)} \frac{\partial}{\partial w} F(x,y;w) =$$

$$= \sum_{(x,y),r,\hat{y}_r} b_{(x,y),r}(\hat{y}_r) \frac{\partial}{\partial w} f_r(x,\hat{y}_r;w) - \sum_{(x,y)} \frac{\partial}{\partial w} F(x,y;w)$$

end

Statement 9. *Update the* $\lambda_{(x,y),r\to p}(\hat{y}_r)$ *by flowing rule minimize D*

$$\mu_{(x,y),p\to r}(\hat{y}_r) = \ln \sum_{\hat{y}_p \setminus \hat{y}_r} \exp\left(f_p(x,\hat{y}_p;w) - \sum_{p'\in P(p)} \lambda_{(x,y),p\to p'}(\hat{y}_p) + \sum_{r'\in C(p)\setminus r} \lambda_{(x,y),r'\to p}(\hat{y}_{r'}) \right)$$

$$\lambda_{(x,y),r\to p}(\hat{y}_r) \propto \frac{1}{1+|P(r)|} \left(f_r(x,\hat{y}_r;w) - \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) + \sum_{p\in P(r)} \mu_{(x,y),p\to r}(\hat{y}_r) \right) - \mu_{(x,y),p\to r}(\hat{y}_r)$$

Proof

$$\frac{\partial D}{\partial \lambda_{(x,y),r\to p}(\hat{y}_r)} = \sum_{\hat{y_p}\setminus\hat{y}_r} b_{(x,y),p}(\hat{y}_p) - b_{(x,y),p}(\hat{y}_r) = 0$$

$$\sum_{\hat{y_p}\setminus\hat{y}_r} b_{(x,y),p}(\hat{y}_p) \propto \exp(\mu_{(x,y),p\to r}(\hat{y}_p) + \lambda_{(x,y),r\to p}(\hat{y}_r)) \propto exp \left(f_r(x,y;w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) \right)$$

then

$$\mu_{(x,y),p\to r}(\hat{y}_p) + \lambda_{(x,y),r\to p}(\hat{y}_r) = f_r(x, y; w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c) - \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r)$$

sum both sides with respect to p

$$(1 + |P(r)|) \cdot \sum_{p\in P(r)} \lambda_{(x,y),r\to p}(\hat{y}_r) = |P(r)|(f_r(x, y; w) + \sum_{c\in C(r)} \lambda_{(x,y),c\to r}(\hat{y}_c)) - \sum_{p\in P(r)} \mu_{(x,y),p\to r}(\hat{y}_p)$$

put that into above eq **end**

Statement 10. *The block-coordinate descent algorithm with follow updates w and λ is guaranteed to monotonically decrease the cost function.*

Proof When optimizing w.r.t. blocks of Lagrange multipliers we ensure that the cost function decreases since we minimize a convex function analytically. Similarly via the Armijo rule we ensure that a descent step w.r.t. w is taken, which concludes the proof. **end**