Differential Dynamic Programming for Structured Prediction and Attention

A short report on [Mensch and Blondel, 2018]

Nazarov Ivan

Skoltech

August 3, 2018

Introduction

Mina idea

- cast a dynamic programming problem as a linear problem
- somehow make the linear problem smooth by relaxation
- recast the smoothed problem back into dynamic programming setting

This enables

- take gradients w.r.t. parameters
- compute hessian-vector products w.r.t parameters

Dynamic Programming Problem

Typical setting for a dynamic programming problem

- $G = (V, E)$ DAG with unique source v_* and sink v^* nodes
- weights $\theta \in \mathbb{R}^{V \times V}$ with $\theta_{v_* v_*} = 1$ and $\theta_{uv} = -\infty$ if $uv \notin E$

Goal

Get a path with the highest score among all paths $v_* \to v^*$

Solution

Identify each $v \in V$ with its number $i = 1 ... m$ in topological order

$$
\begin{aligned}
F_j(\theta) &\leftarrow \max_{i\,:\; j \in G_i} \theta_{ij} + F_j(\theta)\,, \quad F_1(\theta) \leftarrow 0\,,\\
DP(\theta) &\leftarrow F_m(\theta)\,.
\end{aligned}
$$

Find the optimal path by *backtracking* through $(F_j(\theta))_{j=1}^m$

Dynamic Programming Problems as Linear Problems

[Bellman, 1952] showed that $DP(\theta) = LP(\theta)$

$$LP(\theta) = \max_{\pi} \sum_{u \in V} \sum_{v \in V} \theta_{uv} 1_{uv \in \pi} = \max_{\mathbf{y} \in \mathcal{Y}} \langle \theta, \mathbf{y} \rangle$$

where $\mathcal{Y}$ is the set of binary matrices representing paths $v_* \to v^*$

However ...

- $LP(\theta)$ is not differentiable unless its solution is unique
- the optimal solution $\mathbf{y}^*(\theta) = \arg LP(\theta)$ is a discontinuous map

What is a "maximum"?

The largest element of $\theta \in \mathbb{R}^d$ is $\max(\theta)$

$$\begin{aligned} \max\colon \quad & \mathbb{R}^d \longrightarrow \mathbb{R}\,, \\ & \theta \longmapsto \max_{i=1}^{d} \theta_i = \sup_{x \in \Delta^d} \langle x, \theta \rangle\,, \end{aligned} \tag{Max}$$

where Δ^d in the *unit simplex* in $\mathbb{R}^d$, $\Delta^d = \{x\colon \|x\|_1 = 1\,, x \geq 0\}$.

- used in every optimization problem
- differentiable almost everywhere (except on negligible sets)
- non-differentiable solution

Making maxima smooth

Let $\Omega \colon \mathbb{R}^d \to \mathbb{R}$ be a strongly convex regularizer on Δ^d

$$\begin{aligned} \max_\Omega \colon \quad \mathbb{R}^d &\longrightarrow \mathbb{R}\,, \\ \theta &\longmapsto \sup_{x \in \Delta^d} \langle x, \theta \rangle - \Omega(x)\,, \end{aligned} \tag{Smooth-Max}$$

Properties from strong convexity of Ω

- $x^*(\theta) = \arg\sup_{x \in \Delta^d} \langle x, \theta \rangle - \Omega(x)$ exists and unique
- $\nabla \max_\Omega(\theta) = x^*(\theta)$ and is Lipschitz-continuous
- $\nabla^2 \max_\Omega(\theta)$ exists almost everywhere

Natural generalization of max

Table 1: Various types of $\max_\Omega$

	regular-max	*soft*-max	*sparse soft*-max
Ω	0	$-\sum_i x_i \log x_i$	$\frac{1}{2}\|\cdot\|^2$

- $\max_\Omega(\theta_1) \leq \max_\Omega(\theta_2)$ whenever $\theta_1 \leq \theta_2$
- $\max_\Omega(\mathbf{1}c + \theta) = c + \max_\Omega(\theta)$ for any $c \in \mathbb{R}$
- $\max_\Omega(\pi\theta) = \max_\Omega(\theta)$ for any permutation P with $\Omega \circ P = \Omega$
- if $\theta_j = -\infty$ then $(\nabla \max_\Omega(\theta))_j = 0$
- $\max_\Omega(\theta)$ is not far from $\max(\theta)$

Smoothed LP and DP

For any $f : \mathcal{Y} \to \mathbb{R}$ *denote*

$$\max_{\mathbf{y}\in\mathcal{Y}}{}_{\Omega} f(\mathbf{y}) \triangleq \max{}_{\Omega}\big(f(\mathcal{Y})\big)\,, \quad f(\mathcal{Y}) = \big(f(\mathbf{y})\big)_{\mathbf{y}\in\mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}|}\,.$$

Linear Program

$$\mathrm{LP}_{\Omega}(\theta) = \max_{\mathbf{y}\in\mathcal{Y}}{}_{\Omega}\langle\theta, \mathbf{y}\rangle\,.$$

Bellman Iterations

In topological order of G do

$$F_j(\theta) \leftarrow \max_{i\,:\,j\in G_i}{}_{\Omega}\big(\theta_{ij} + F_j(\theta)\big)\,, \quad F_1(\theta) \leftarrow 0\,,$$

$$\mathrm{DP}_{\Omega}(\theta) \leftarrow F_m(\theta)\,.$$

Behaviour of smoothed LP and DP

Both $\mathrm{LP}_\Omega(\theta)$ and $\mathrm{DP}_\Omega(\theta)$ are well-behaved

- convex and differentiable everywhere
- have Lipschitz continuous gradients
- gradients are differentiable almost everywhere

However ...

- $\mathrm{LP}_\Omega(\theta)$ is intractable due to exponential size of $\mathcal{Y}$
- $\mathrm{DP}_\Omega(\theta)$ is tractable and has complexity $\mathcal{O}\big(|E|\big)$

Properties of smoothed LP and DP

[Mensch and Blondel, 2018] propose and prove:

- $\mathrm{DP}_\Omega(\theta)$ is convex w.r.t θ
- $\mathrm{DP}_\Omega(\theta) = \mathrm{LP}_\Omega(\theta)$ **if and only if** $\Omega(x) = -\gamma \sum_i x_i \log x_i$
- is "close" to LP: $\left|\mathrm{LP}(\theta) - \mathrm{DP}_\Omega(\theta)\right| \leq m\, M(\Omega, m)$ and

$$\lim_{\gamma \to 0} \mathrm{DP}_{\gamma\Omega(\theta)} = \mathrm{LP}(\theta)\,,$$

- efficient recursion to compute $\nabla_\theta \mathrm{DP}_\Omega(\theta) \in \mathbb{R}^{V \times V}$ and hessian-vector products $\nabla^2 \mathrm{DP}_\Omega(\theta) Z$

Computing $\nabla_\theta \mathrm{DP}_\Omega(\theta)$

Simply backpropagate along the reverse-topological order of G

- **forward-pass:** while computing $F_i(\theta)$ get

$$q_i(\theta) = \nabla \max_\Omega\big(\theta_i + F(\theta)\big) \in \Delta^m\,,$$

  assuming $F_k(\theta) = -\infty$ for all k **after** i

- **backward-pass:** in reverse-topological order $j = m \dots 1$ do

$$\begin{aligned} \bar{w}_j &\leftarrow \sum_{i \in G_j} w_{ij} \;\text{ if } j \neq m \text{ else } 1\,, \\ w_{ij} &\leftarrow \bar{w}_i q_{ij} \;\text{ if } i \in G_j \text{ else } 0\,, \\ \nabla_\theta \mathrm{DP}_\Omega(\theta) &\leftarrow (w_{ij})_{i,j=1\dots m} \in \mathbb{R}^{V \times V}\,. \end{aligned}$$

Interpretaion of $\nabla_\theta \mathrm{DP}_\Omega(\theta)$

The matrix $Q(\theta) = \big(q_i(\theta)\big)_{i=1}^m$

- a transition matrix for backward random walks from $v^* = m$ back to $v_* = 1$
- $\mathbb{P}(i \to j) = q_{ij}(\theta)$ if $i \in G_j$.

[Mensch and Blondel, 2018] demonstrate

- the gradient is the expected path of the random walk

$$\nabla_\theta \mathrm{DP}_\Omega = \mathbb{E}_{\mathbf{y} \sim Q(\theta)} \mathbf{y}\,.$$

- convergence to the optimal solution

$$\nabla_\theta \mathrm{DP}_{\gamma\Omega}(\theta) \xrightarrow[\gamma \to 0]{} \mathbf{y}^*(\theta) \in \partial \mathrm{LP}(\theta)\,.$$

Conclusion

[Mensch and Blondel, 2018] propose

- a theoretical framework for turning dynamic programs into convex, differentiable and tractable operators
- efficient way to embed the smoothed programs into models learnt by gradient descent

Applications: learning optimal cost parameters for end-to-end training in

- sequence prediction in part-of-speech tagging
- time series alignment in audio transcription
- attention mechanism in machine translation

References

Bellman, R. (1952).
On the theory of dynamic programming.
Proceedings of the National Academy of Sciences, 38(8):716–719.

Mensch, A. and Blondel, M. (2018).
Differentiable Dynamic Programming for Structured Prediction and Attention.
In *35th International Conference on Machine Learning*, volume 80 of *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden.