We thank all the reviewers for their valuable feedback. In response we'll include Tab. 1, which gives the average epoch compute time (i.e., compute inference objective and update model weights) during training for our models for each task.

To Reviewer #2:

Re: Fig 1a. Yes, this should be 'iterations' – we will fix this.

Re: Variability of different optimization approaches. It is not the case that choosing an inference method is a "black art." The larger variance of some approaches arises due to an incompatibility of the inference objective and the optimization algorithm used to run inference. For example, unboundedness of the entropy at the boundaries of the domain is known to hurt convergence of Frank-Wolfe for objectives which contain it (see [*]). Furthermore, we present settings where the Struct model was trained with entropy while the T function in GSPEN was trained without entropy or vice-versa. Because the two different model components are trained using different objectives, final performance expectedly suffers. We include those settings for completeness sake, but we do not recommend to use them in practice.

	Struct	SPEN	NLStruct	GSPEN
OCR (size 1000)	0.40 s	0.60 s	68.56 s	8.41 s s
Tagging	18.85 s	30.49 s	208.96 s	171.65 s
Bibtex	0.36 s	11.75 s	–	13.87 s
Bookmarks	6.05 s	94.44 s	–	234.33 s
NER	29.16 s	–	–	99.83 s

Table 1: Average time to compute inference objective and complete a weight update for one pass through the training data. We show all models trained for the submission.

To Reviewer #3:

Re: Relaxation of marginal polytope. The formulation of inference for GSPEN allows the practitioner to select a structured prediction algorithm (and whatever relaxation of $\mathcal{M}$ this entails). Obviously this has consequences for both computational complexity and solution quality. For all of the experiments used in this paper, we use the local marginal polytope $\mathcal{M}_L$ described starting on line 90. We use the structured inference procedure employed in [9, 16, 30]. We think the selected approach provides a good tradeoff between computational complexity and solution quality.

To Reviewer #4:

Re: Theoretical analysis. We discuss the conditions under which convergence guarantees are available for inference: (1) for Frank-Wolfe starting in line 166 and (2) for structured entropic mirror descent starting in line 188. Unfortunately, convergence guarantees in the settings used for experimentation have not been proved due to non-concavity/non-convexity of the objectives and the employed update steps. However, we think these settings are useful for practitioners.

Re: Intuition behind GSPEN. We discuss the motivations for this model starting at line 17: this model permits to use a structured score function (which SPEN does not) while also enabling to use an energy function that scores entire predictions jointly (which is not supported in classic structured prediction).

Re: SPEN vs. GSPEN. GSPEN allows a practitioner to augment a structured score function with an additional energy function that jointly scores the prediction vector. Therefore, any setting where structured score functions are used (e.g., for NER – see [1, 29] for examples) can benefit from the GSPEN formulation. Our results demonstrate that for a variety of tasks, having both an energy function and a structured score function results in better performance than having either individually. Specifically, the datasets used for the OCR experiments were designed to demonstrate this: the per-variable information (i.e., predicting from images alone) is noisy, but due to the limited vocabulary used to generate the words, there is much useful information within the structure of the labels that GSPEN is able to exploit. Furthermore, Fig. 3 demonstrates that GSPEN is able to exploit this structure better than either Struct or SPEN. The tradeoff of GSPEN's ability to include structured score functions is its increased computational complexity, which is a consequence of maintaining the structural constraints during inference. Hence, in settings where structure is unknown or strictly higher-order, adding a structured score function to SPEN may not provide benefits performance-wise and will be slower due to the additional cost of structured inference.

Re: Complexity of inference. The inference algorithm does not scale based on the number of data points, but rather the number of variables in the problem, the number of regions being modeled in the structured score function, and the number of states each variable takes. The computational complexity of GSPEN depends on the chosen classic structured inference algorithm and its complexity, since it will be called once per iteration of inference. These algorithms scale with the size of the largest region, which is why pairwise structured models are commonly used. The complexity of the structured inference algorithm we use is $O(|\mathcal{R}| \cdot \max_r |P(r)| \cdot \max_r |\mathcal{Y}_r|)$, where $\mathcal{R}$ is the set of graph regions and $P(r)$ is the set of "parents" of region r (in a pairwise model, $P(i)$ is the set of pairs containing variable i). Depending on the employed model, computing the gradients of the model may also be costly. For all the problems we consider this cost is lower than the one of computing the structured inference objective.

References:

[*] R. G. Krishnan, S. Lacoste-Julien, and D. Sontag. "Barrier Frank-Wolfe for marginal inference." NIPS 2015.