Perturb-and-MAP Random Fields: Using Discrete Optimization to Learn and Sample from Energy Models – ICCV 2011 paper supplementary material –

George Papandreou and Alan Yuille
Department of Statistics, University of California, Los Angeles
[gpapan,yuille]@stat.ucla.edu

Abstract

In the supplementary material we give proofs of the propositions stated in the main paper and provide additional experimental results.

1. Perturb-and-MAP random fields

Recall the definitions of the polyhedra $\mathcal{P}_{\mathbf{x}}$ in the space of perturbations

$$\mathcal{P}_{\mathbf{x}} = \{\boldsymbol{\theta} \in \mathbb{R}^M : \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{q}) \rangle \leq 0, \forall \mathbf{q} \in \mathcal{L}^N\}, \quad (1)$$

as well as the density

$$f_{PM}(\mathbf{x}; \boldsymbol{\theta}) = \int_{\mathcal{P}_{\mathbf{x}} - \boldsymbol{\theta}} f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}\,, \quad (2)$$

and log-likelihood

$$L_{PM}(\boldsymbol{\theta}) = (1/K) \sum_{k=1}^{K} \log f_{PM}(\mathbf{x}_k; \boldsymbol{\theta}) \quad (3)$$

under the Perturb-and-MAP model (Section 3 of the main paper [7]).

1.1. Concavity of Perturb-MAP log-likelihood

We will use an important property of log-concave functions (i.e., functions whose logarithm is concave) arising in stochastic programming [8]; also see [3, Sec.3.5]:

Lemma 0. *If $h : \mathbb{R}^{M_1} \times \mathbb{R}^{M_2} \to \mathbb{R}$ is log-concave in $(\boldsymbol{\theta}, \boldsymbol{\epsilon})$, then $g(\boldsymbol{\theta}) = \int h(\boldsymbol{\theta}, \boldsymbol{\epsilon}) d\boldsymbol{\epsilon}$ is log-concave in $\boldsymbol{\theta}$.*

Proposition 1. *If the perturbations $\boldsymbol{\epsilon}$ are drawn from a log-concave density $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$, the log-likelihood $L_{PM}(\boldsymbol{\theta})$ is a concave function of the energy parameters $\boldsymbol{\theta}$.*

Proof. We will prove that $f_{PM}(\mathbf{x}; \boldsymbol{\theta})$ of Eq. (2) is log-concave in $\boldsymbol{\theta}$ for any $\mathbf{x} \in \mathcal{L}^N$; the log-likelihood $L_{PM}(\boldsymbol{\theta})$ of Eq. (3) will then be concave as the sum of concave functions. We write $f_{PM}(\mathbf{x}; \boldsymbol{\theta}) = \int \Psi_{\mathbf{x}}(\boldsymbol{\theta}, \boldsymbol{\epsilon}) f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}$, where $\Psi_{\mathbf{x}}(\boldsymbol{\theta}, \boldsymbol{\epsilon})$ equals 1, if $\boldsymbol{\theta} + \boldsymbol{\epsilon} \in \mathcal{P}_{\mathbf{x}}$ and 0, otherwise. As a function jointly of $(\boldsymbol{\theta}, \boldsymbol{\epsilon})$, $\Psi_{\mathbf{x}}$ is the indicator function of the convex set (polyhedron) $\{(\boldsymbol{\theta}, \boldsymbol{\epsilon}) : \langle \boldsymbol{\theta} + \boldsymbol{\epsilon}, \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{q}) \rangle \leq 0, \forall \mathbf{q} \in \mathcal{L}^N\}$ and is thus log-concave. Since $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ is also log-concave, the integrand is log-concave in $(\boldsymbol{\theta}, \boldsymbol{\epsilon})$, as the product of log-concave functions. By invoking Lemma 0, we conclude that $f_{PM}(\mathbf{x}; \boldsymbol{\theta})$ is log-concave in $\boldsymbol{\theta}$. □

1.2. Moment matching update takes steps in the right direction

Proposition 2. *If $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ differ only in the j-element, with $\theta'_j > \theta_j$, then $E^{PM}_{\boldsymbol{\theta}'}\{\phi_j(\mathbf{x})\} \leq E^{PM}_{\boldsymbol{\theta}}\{\phi_j(\mathbf{x})\}$. The inequality will be strict if the the perturbation density $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ is "rich enough".*

Proof. The j-moment under the Perturb-and-MAP model is $E^{PM}_{\boldsymbol{\theta}}\{\phi_j(\mathbf{x})\} = \sum_{\mathbf{x}} \phi_j(\mathbf{x}) f_{PM}(\mathbf{x}; \boldsymbol{\theta})$, where from Eq. (2) $f_{PM}(\mathbf{x}; \boldsymbol{\theta}) = \int_{\mathcal{P}_{\mathbf{x}} - \boldsymbol{\theta}} f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}$ is the measure of the shifted polyhedron $\mathcal{P}_{\mathbf{x}} - \boldsymbol{\theta}$ under the perturbation density $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$.

Let as re-write the linear inequalities defining the shifted polyhedron as $\mathcal{P}_{\mathbf{x}} - \boldsymbol{\theta} = \{\boldsymbol{\epsilon} \in \mathbb{R}^M : \sum_{j'=1}^{M} \epsilon_{j'}(\phi_{j'}(\mathbf{x}) - \phi_{j'}(\mathbf{q})) \leq -\sum_{j'=1}^{M} \theta_{j'}(\phi_{j'}(\mathbf{x}) - \phi_{j'}(\mathbf{q})), \forall \mathbf{q} \in \mathcal{L}^N\}$. Since $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ are identical except for $\theta'_j > \theta_j$, any of the inequalities above gets tighter if $\phi_j(\mathbf{x}) \geq \phi_j(\mathbf{q})$, looser if $\phi_j(\mathbf{x}) \leq \phi_j(\mathbf{q})$, and stays the same if $\phi_j(\mathbf{x}) = \phi_j(\mathbf{q})$. Thus, the effect in $\boldsymbol{\epsilon}$-space of increasing θ_j is that if a state $\mathbf{x}$ has a neighbor $\mathbf{q}$ for which $\phi_j(\mathbf{x}) > \phi_j(\mathbf{q})$, then the polyhedron $\mathcal{P}_{\mathbf{x}} - \boldsymbol{\theta}$ of state $\mathbf{x}$ shrinks in favor of the polyhedron $\mathcal{P}_{\mathbf{q}} - \boldsymbol{\theta}$ of state $\mathbf{q}$, whereas if $\phi_j(\mathbf{x}) < \phi_j(\mathbf{q})$ then $\mathcal{P}_{\mathbf{x}} - \boldsymbol{\theta}$ expands over $\mathcal{P}_{\mathbf{q}} - \boldsymbol{\theta}$; in both cases, the j-moment will decrease $E^{PM}_{\boldsymbol{\theta}'}\{\phi_j(\mathbf{x})\} \leq E^{PM}_{\boldsymbol{\theta}}\{\phi_j(\mathbf{x})\}$.

The inequality will be strict if *any* of the $\boldsymbol{\epsilon}$-space exchanges among neighboring polyhedra has strictly positive measure under $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$. This condition is satisfied even if we only perturb the unary terms with a strictly positive density. In particular, the Gumbel perturbation of any order (even

the order-1) satisfies this condition. □

Note that if we change many elements of $\boldsymbol{\theta}$ at once, there can be interference due to correlation among different features. This implies that the convergence of the model sufficient statistics towards their observed values needs not be monotonic, which is also the case with the gradient ascent learning rule for Gibbs MRFs.

1.3. Gumbel perturbations

The Gumbel continuous univariate distribution[1] with mode μ has probability density function (PDF) $g(z;\mu) = \exp((z-\mu) - e^{z-\mu}))$ and cumulative distribution function (CDF) $G(z;\mu) = 1 - \exp(-e^{z-\mu})$. Note that the Gumbel density is log-concave, since $z - e^z$ is concave in z. Also, if u is drawn from the standard uniform distribution, then $z = \mu + \log(-\log(u)) \sim \mathcal{G}(\mu)$, i.e., z follows a Gumbel distribution with mode μ – see, e.g., [9]. We plot the Gumbel PDF and CDF for $\mu = 0$ in Fig. 1.

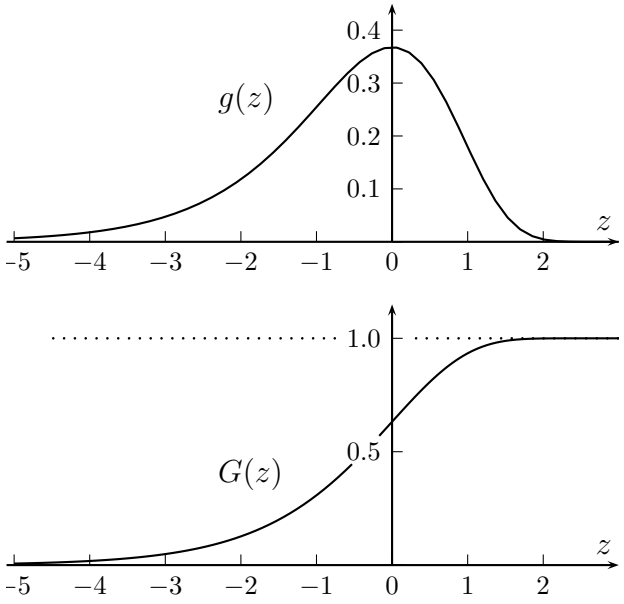


Figure 1. Probability density function and cumulative distribution function of the Gumbel distribution with mode $\mu = 0$.

The Gumbel density fits naturally into the Perturb-and-MAP model. We have the following Lemma on the minimum of independent Gumbel random variables:

Lemma 1. *Let $(\theta_1, \ldots, \theta_m)$, with $\theta_n \in \mathbb{R}$, $n = 1, \ldots, m$. We additively perturb them by $\tilde{\theta_n} = \theta_n + \epsilon_n$, with ϵ_n IID zero-mode Gumbel samples. Then:*

(a) The minimum of the perturbed parameters $\tilde{\theta}_{min} \triangleq \min_{n=1:m}\{\tilde{\theta}_n\}$ follows a Gumbel distribution with mode θ_0, where $e^{-\theta_0} = \sum_{n=1}^m e^{-\theta_n}$.

(b) The probability that $\tilde{\theta}_n$ attains the minimum value is $\Pr\{\text{argmin}(\tilde{\theta}_1, \ldots, \tilde{\theta}_m) = n\} = e^{-\theta_n}/e^{-\theta_0}$.

Proof. (a) The CDF of the minimum $\tilde{\theta}_{min}$ is

$$\begin{aligned} F_{min}(\theta) &= \Pr\{\tilde{\theta}_{min} \leq \theta\} \\ &= 1 - \Pr\{\tilde{\theta}_n > \theta, 1 \leq n \leq m\} \\ &= 1 - \prod_{n=1}^m \Big(1 - G(\theta;\theta_n)\Big) \\ &= 1 - \prod_{n=1}^m \exp(-e^{\theta-\theta_n}) \\ &= 1 - \exp\Big(-e^{\theta}(\sum_{n=1}^m e^{-\theta_n})\Big) \\ &= 1 - \exp(-e^{\theta-\theta_0}), \end{aligned}$$

which is the CDF of a Gumbel distribution with mode $\theta_0 = -\log(\sum_{n=1}^m e^{-\theta_n})$.

(b) We have

$$\begin{aligned} &\Pr\{\text{argmin}(\tilde{\theta}_1, \ldots, \tilde{\theta}_m) = n\} = \\ &= \Pr\{\tilde{\theta}_n \leq \min_{j\neq n}\{\tilde{\theta}_j\}\} \\ &= \int_{-\infty}^{+\infty} g(t;\theta_n) \prod_{j\neq n} \Big(1 - G(t;\theta_j)\Big) dt \\ &= \int_{-\infty}^{+\infty} e^{t-\theta_n} \exp(-e^{t-\theta_n}) \prod_{j\neq n} \exp(-e^{t-\theta_j}) dt \\ &= \int_0^1 \prod_{j\neq n} z^{\exp(\theta_n-\theta_j)} dz \qquad [\text{we set } z \triangleq \exp(-e^{t-\theta_n})] \\ &= \int_0^1 z^{\sum_{j\neq n}\exp(\theta_n-\theta_j)} dz \\ &= \frac{1}{1 + \sum_{j\neq n} e^{\theta_n-\theta_j}} \\ &= \frac{e^{-\theta_n}}{\sum_{j=1}^m e^{-\theta_j}} \end{aligned}$$

□

A result related to Lemma 1 has appeared before in the context of online learning [6, 10].

Gumbel perturbation on fully-expanded potential table The Gibbs random field on N sites x_i, $i = 1, \ldots, N$, each allowed to take a value from the discrete label set $\mathcal{L}$ can be considered as a discrete distribution with $|\mathcal{L}|^N$ states. This

[1] We use the min-Gumbel form of the density which arises in the study of minima of random variables, as is the case with our energy minimization problem. In the study of maxima of random variables, the max-Gumbel form of the density is encountered, whose density is just the mirror $g(-z+\mu;\mu)$ of the min-Gumbel density.

can be made explicit if we enumerate $\{\mathbf{x}_j, j = 1, \ldots, \bar{M} = |\mathcal{L}|^N\}$ all the states and consider the maximal equivalent re-parameterization of Eq. (1) of the main paper

$$\bar{e}(\mathbf{x}; \bar{\boldsymbol{\theta}}) \triangleq \langle \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}}(\mathbf{x}) \rangle = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle \,, \tag{4}$$

where $\bar{\theta}_j = e(\mathbf{x}_j; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}_j) \rangle$, $j = 1, \ldots, \bar{M}$, is the *fully-expanded* potential table and $\bar{\phi}_j(\mathbf{x})$ is the indicator function of $\mathbf{x}_j$ (i.e., equals 1, if $\mathbf{x} = \mathbf{x}_j$ and 0 otherwise).

Using Lemma 1, we can then prove the following:

Proposition 3. *If we perturb each entry of the fully expanded $\mathcal{L}^N$ potential table with IID Gumbel noise samples ϵ_j, $j = 1, \ldots, \bar{M}$, then the Perturb-and-MAP and Gibbs models coincide, i.e., $f_{PM}(\mathbf{x}; \boldsymbol{\theta}) = f_G(\mathbf{x}; \boldsymbol{\theta})$.*

Proof. We have:

$$\begin{aligned}
f_{PM}(\mathbf{x}_j; \boldsymbol{\theta}) &= \\
&= \Pr\{\operatorname*{argmin}_{\mathbf{q}} \bar{e}(\mathbf{q}; \bar{\boldsymbol{\theta}}) = \mathbf{x}_j\} \\
&= \Pr\{\operatorname{argmin}(\bar{\theta}_1 + \epsilon_1, \ldots, \bar{\theta}_{\bar{M}} + \epsilon_{\bar{M}}) = j\} \\
&= \frac{\exp(-\bar{\theta}_j)}{\sum_{j'=1}^{\bar{M}} \exp(-\bar{\theta}_{j'})} \qquad \text{[by Lemma 1(b)]} \\
&= \frac{\exp(-e(\mathbf{x}_j; \theta)}{\sum_{j'=1}^{\bar{M}} \exp(-e(\mathbf{x}_{j'}; \theta))} \\
&= \frac{\exp(-e(\mathbf{x}_j; \theta)}{Z(\boldsymbol{\theta})} \\
&= f_G(\mathbf{x}_j; \boldsymbol{\theta})
\end{aligned}$$

□

Probability that a perturbed Ising link is submodular The order-2 Gumbel perturbation can yield non-submodular functions even when our original energy is submodular. It turns out that for the Ising model we can compute in closed-form (using a result of [1] on the distribution of the sum of two logistic random variables) the probability that a single pairwise link of strength λ will remain submodular after order-2 Gumbel perturbation. The formula is $\Pr\{\tilde{\lambda} \geq 0\} = e^{2\lambda}(e^{2\lambda} - 2\lambda - 1)/(e^{2\lambda} - 1)^2$, plotted in Fig. 2; for example, for $\lambda = 4$, $\Pr\{\tilde{\lambda} \geq 0\} \approx 0.998$.

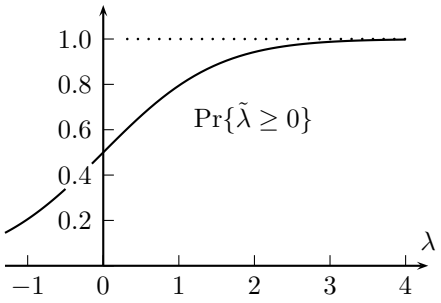


Figure 2. Probability that a perturbed Ising link is submodular.

2. Additional experimental results

2.1. Interactive image segmentation

In Fig. 3 we show further results obtained by the Perturb-and-MAP model on the interactive image segmentation task, produced as described in Sec. 5.1 of the main paper. We show the original image, the ground-truth hand-annotated segmentation, the least energy MAP solution (i.e., the result obtained by the standard Grabcut algorithm using the weights learned by Perturb-and-MAP moment matching), and the soft Perturb-and-MAP segmentation (average over 20 posterior samples) along with the corresponding alpha mask.

2.2. Scene layout labeling

We report additional results on the evaluation of the Perturb-and-MAP model on the scene layout labeling task, described in Sec. 5.2 of the main paper.

First, for completeness, we give in Table 1 the baseline mean accuracy and full row-normalized confusion matrices for confidence-only [5] and MAP [4] (the result obtained by the standard tiered algorithm using the hand set potentials of [4], as distinct from the learned potentials – the MAP results with our learned potentials are reported in Table 1 of the main paper.) In both cases we use the classifiers we trained using the dataset and software of [5] (same classifiers as the ones we used in conjunction with the Perturb-and-MAP experiments reported in the main paper).

Confidence only (acc 82.1%)

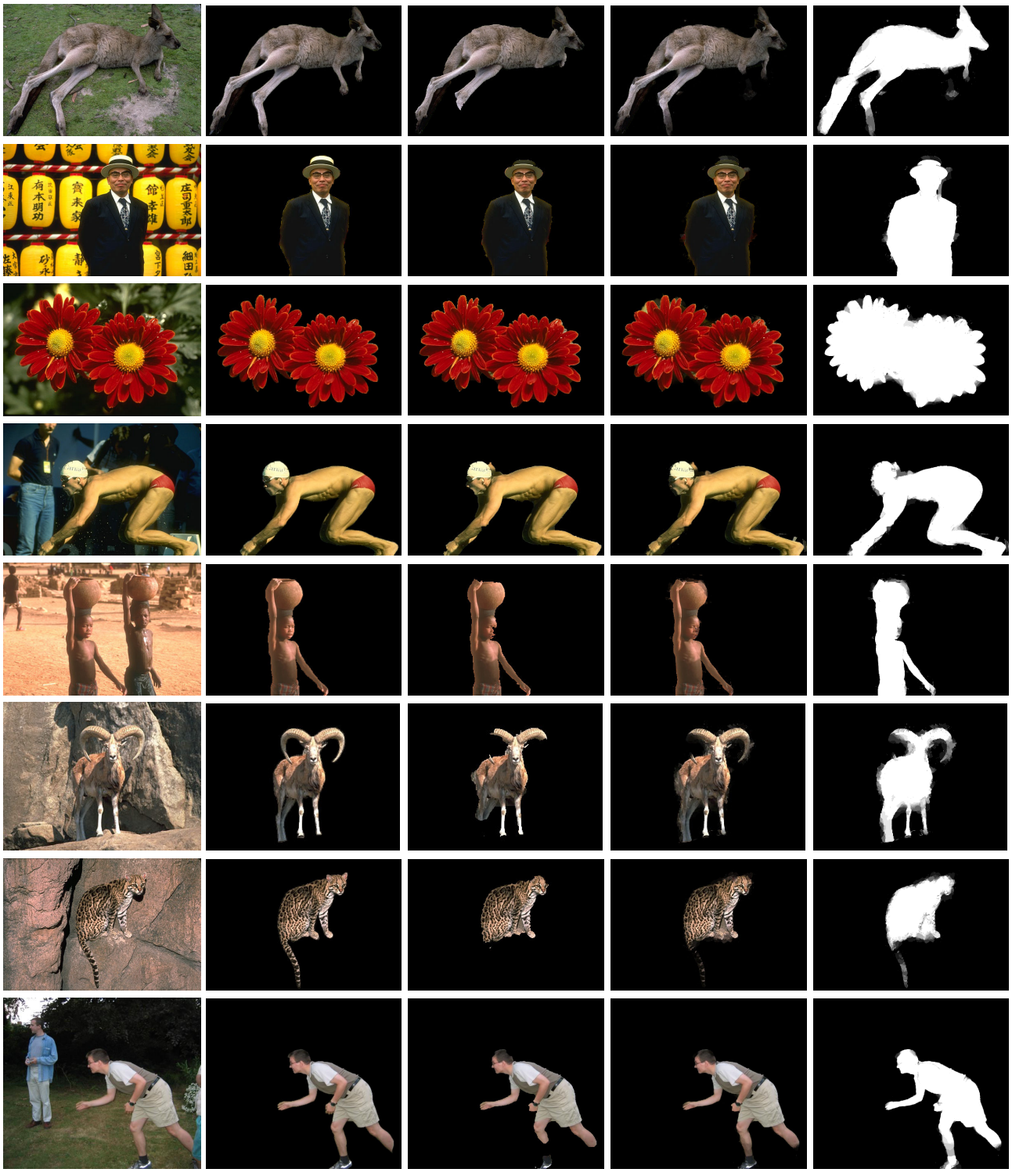
	B	L	C	R	T
B	93.6	0.4	4.7	0.8	0.5
L	20.7	38.2	31.7	8.5	1.0
C	21.3	5.2	55.6	12.9	5.0
R	13.5	2.8	26.6	51.8	5.3
T	0.5	0.3	3.2	1.0	95.0

MAP (hand set w) (acc 82.1%)

	B	L	C	R	T
B	93.8	0.3	4.6	0.8	0.5
L	20.8	35.3	34.3	8.0	1.7
C	21.5	4.3	57.5	11.8	4.9
R	13.9	1.2	28.4	51.1	5.4
T	0.7	0.3	3.6	0.8	94.6

Table 1. *Baseline scene labeling confusion matrices.*

In Fig. 4 we show further examples obtained by the Perturb-and-MAP model on the scene layout labeling task with the tiered model. We show the original image, the confidence-only result, the least energy MAP solution (i.e., the result obtained by the standard tiered algorithm using the potentials learned by Perturb-and-MAP moment matching), the per-pixel most probable labeling (mode on the labels over 20 posterior Perturb-and-MAP samples), one Perturb-and-MAP sample, and the per-pixel entropy of the labeling (computed on the per-pixel labeling histogram of the 20 posterior Perturb-and-MAP samples). In all Perturb-and-MAP results we have used order-1 Gumbel perturbations.



(a) (b) (c) (d) (e)

Figure 3. Interactive image segmentation results. We show: (a) the original image, (b) the ground-truth segmentation, (c) the least energy MAP solution (i.e., the result obtained by the standard Grabcut algorithm using the weights learned by Perturb-and-MAP moment matching), (d & e) the soft Perturb-and-MAP segmentation (average over 20 posterior samples), and the corresponding alpha mask.

(a) (b) (c) (d) (e) (f)

Figure 4. Tiered scene labeling results. We show (a) original image, (b) confidence-only result, (c) least energy MAP solution, (d) indicative sample, (e) per-pixel Perturb-and-MAP marginal mode, (f) per-pixel marginal entropy map. The results (c-f) use the potentials learned by our Perturb-and-MAP moment matching algorithm. Marginal densities used in (e-f) are averages over 20 posterior samples.

In Fig. 5 we visualize the pairwise neighbor potential tables learned for the tiered model on the dataset of [5]. These correspond to the f_{pq} vertical and horizontal CRF weights of [4]. For training, we also added to the model log-likelihood a mild L_2 weight regularization term $\kappa \sum_{\alpha,b}(f_{pq}(\alpha,b))^2$, with $\kappa = 10^{-4}$ (and similarly for the unary term weights).

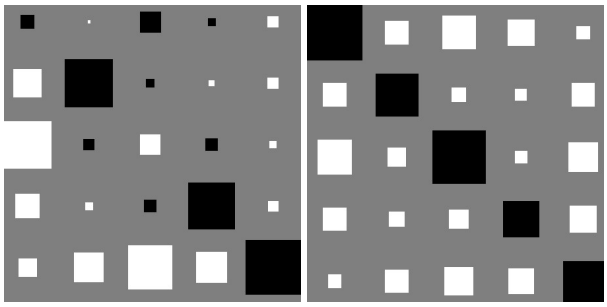


Figure 5. Vertical and horizontal pairwise potential tables of the tiered model learned on the dataset of [5]. Class labels are ordered as in [4]. Weights illustrated as Hinton diagrams.

3. Further discussion

This Section has been added after the conference camera-ready deadline and discusses issues that have come up from correspondence with colleagues.

3.1. Every state is reachable under the Perturb-and-MAP model

Similarly to the Gibbs-MRF, the Perturb-and-MAP model assigns strictly positive probability to every state[2]. Specifically:

Proposition 4. *The order-1 perturbation (i.e., adding noise to the unary potentials only) yields a Perturb-and-MAP model that assigns non-zero probability to every state, provided that the perturbation density is unbounded (i.e., its support is the whole real line), which holds for the order-1 Gumbel perturbation.*

Proof. We consider a general MRF with N nodes $x_i, i = 1, \ldots, N$ and L states per node. In the order-1 perturbation, the feature set $\boldsymbol{\phi}$ includes all $N{\times}L$ unary indicator functions $\phi_{i,l}(\mathbf{x}) = [x_i = l] = 1, \quad \text{if} \quad x_i = l, \quad \text{and} \quad 0,$ otherwise, plus any other features specific to the particular MRF (e.g., pairwise potentials etc.).

Using the notation of the main paper's Sec. 3.1, an arbitrary state $\mathbf{x}$ will be generated by the P-M sampler iff the perturbations $\{\epsilon_{i,l}\}$ satisfy the linear inequalities

$$\sum_{i,l}(\theta_{i,l}+\epsilon_{i,l})([x_i = l]-[q_i = l]) \leq -\sum_{\alpha}\theta_\alpha(\phi_\alpha(\mathbf{x})-\phi_\alpha(\mathbf{q})) \tag{5}$$

for all other states $\mathbf{q} \in \mathcal{L}^N$, with $\mathbf{q} \neq \mathbf{x}$, where α is an index to all the remaining features of order 2 or more. Note that since we are dealing with order-1 perturbations the corresponding parameters θ_α stay unperturbed.

Let $m_{\mathbf{x}} = \min_{\mathbf{q}} - \sum_\alpha \theta_\alpha(\phi_\alpha(\mathbf{x}) - \phi_\alpha(\mathbf{q}))$ be the minimum of the right hand side over all possible states $\mathbf{q} \in L^N$ and $\bar{m}_{\mathbf{x}} = \min(m_{\mathbf{x}}, 0) \leq 0$. Then it is sufficient that the perturbations $\{\epsilon_{i,l}\}$ satisfy the tighter set of linear inequalities

$$\sum_{i,l}(\theta_{i,l} + \epsilon_{i,l})([x_i = l] - [q_i = l]) \leq \bar{m}_{\mathbf{x}} \tag{6}$$

for all states $\mathbf{q} \in \mathcal{L}^N$, with $\mathbf{q} \neq \mathbf{x}$.

For this last set of inequalities to be satisfied, it suffices that the perturbations fall within the quadrant

$$\theta_{i,l} + \epsilon_{i,l} \leq \bar{m}_{\mathbf{x}}, \quad \text{if} \quad x_i = l \tag{7}$$

$$\theta_{i,l} + \epsilon_{i,l} \geq -\bar{m}_{\mathbf{x}}, \quad \text{if} \quad x_i \neq l \tag{8}$$

Since the support of the independent unary perturbations is the whole real line, the probability of the state $\mathbf{x}$ under the P-M model is bounded by $f_{PM}(\mathbf{x}) \geq \prod_{i,l:x_i=l} \Pr\{\epsilon_{i,l} \leq \bar{m}_{\mathbf{x}} - \theta_{i,l}\} \prod_{i,l:x_i\neq l} \Pr\{\epsilon_{i,l} \leq -\bar{m}_{\mathbf{x}} - \theta_{i,l}\} > 0$.

In other words, the order-1 perturbation is expressive enough to generate an arbitrary state $\mathbf{x}$ as a sample. For this to happen, it suffices that the unary term perturbation encourages a lot the particular state, Eq. (7), and discourages a lot its competitors, Eq. (8). □

For a geometric understanding of this property and with reference to Fig. 1 of the main paper, we are guaranteed to reach state $\mathbf{x} = (1, -1)$ if we select ϵ_1 sufficiently large and ϵ_2 sufficiently small, no matter what the values of the coupling strength $|\lambda|$ or the external field β_i are.

Perturbations of order 2 or larger are even more expressive and thus also guaranteed to assign strictly positive probability to all states.

3.2. Bibliographic note

In the related work Section of the main paper, we have missed referring to the article of Blum et al. [2][3]. They also propose adding noise to the weighted edges of a graph so as to randomize the minimum energy configuration found by mincuts. They deal with a submodular binary MRF problem arising in the context of semi-supervised learning. Their main arguments are: (a) In the presence of multiple cuts with the same minimum cost, randomization breaks the graph symmetries and allows the standard mincut algorithm produce a different mincut solution at each run. In particular, randomization, coupled with an extra post-processing pruning step, allows one to avoid highly unbalanced cuts

[2] Question raised by M. Welling of UC Irvine.

[3] We thank an anonymous reviewer and C. Lampert of IST Austria for pointing us to it.

which can be detrimental to semi-supervised learning classification performance [2]. (b) They interpret the relative frequency of each node receiving one or the other label as a confidence score for binary classification. However, beyond randomizing the deterministic mincut algorithm, they do not study the implied probabilistic model as a standalone object nor attempt to design the perturbation mechanism so as to approximate the corresponding Gibbs model. Indeed, the choice of perturbation distribution is not discussed at all in [2], presumably because adding small amounts of noise from any well-behaved continuous distribution suffices to break model symmetries. Their discussion is also limited to submodular energies on binary labels.

References

[1] D. Aldous. The $\zeta(2)$ limit in the random assignment problem. *Random Struc. Algor.*, 18(4):381–418, 2001.

[2] A. Blum, J. Lafferty, M. Rwebangira, and R. Reddy. Semi-supervised learning using randomized mincuts. In *Proc. ICML*, 2004.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[4] P. Felzenszwalb and O. Veksler. Tiered scene labeling with dynamic programming. In *Proc. CVPR*, 2010.

[5] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.

[6] D. Kuzmin and M. K. Warmuth. Optimum follow the leader algorithm. In *Proc. COLT*, pages 684–686, 2005.

[7] G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *Proc. ICCV*, 2011.

[8] A. Prékopa. *Stochastic Programming*. Kluwer, 1995.

[9] F. Steutel and K. Van Harn. *Infinite divisibility of probability distributions on the real line*. Dekker, 2004.

[10] M. K. Warmuth. A perturbation that makes "follow the leader" equivalent to "randomized weighted majority". unpublished note, 2009. Result attributed to A. Kalai.