Herding Generalizes Diverse $M$-Best Solutions

Abstract

We show that the algorithm by Batra et al. (2012) to extract diverse $M$-solutions from a Conditional Random Field (called divMbest) takes exactly the form of a Herding procedure (Welling, 2009b), i.e., a deterministic dynamical system that produces a sequence of solutions that respect a set of observed moment constraints. This generalization enables us to invoke properties of Herding that show that divMbest enforces implausible constraints, and to consider better alternatives for these constraints. Our experiments in image segmentation demonstrate that DivMbest is outperformed by a substantial margin when Herding is configured to enforce plausible constraints.

1 Introduction

Conditional Random Fields (CRFs) are probabilistic graphical models for structured prediction (Lafferty et al., 2001), which have been successfully applied in Computational Biology, Computer Vision, and Language Processing. Most CRF-based methods deliver the maximum a posteriori (MAP) of the CRF probability function, using an off-the-shelf optimizer that delivers the MAP labeling in an efficient way, e.g. (Frey and MacKay, 1997; Boykov and Kolmogorov, 2004).

Another strand of research stressed the need to exploit diverse, highly likely hypotheses, sampled from a CRF, rather than only delivering the MAP labeling (Batra et al., 2012; Chen et al., 2013; Guzman-Rivera et al., 2014). Batra et al. (2012) showed that in a pool of diverse and likely hypotheses sampled from a CRF, there may be some hypotheses with much higher accuracy than the MAP labeling. Also, applications that interact with a human or with other algorithms may benefit from delivering multiple hypotheses, since the human or the algorithms can select among them (Yadollahpour et al., 2013; Premachandran et al., 2014).

In this paper we analyze the successful algorithm by Batra et al. (2012) to extract diverse $M$-best solutions from a CRF. This algorithm has been extensively analyzed in the literature, e.g. (Chen et al., 2013; Guzman-Rivera et al., 2014; Prasad et al., 2014), and it is representative of the state-of-the-art for extracting diverse hypotheses. We will refer to this method as divMbest in the rest of the paper.

Our analysis shows that divMbest is a particular case of the Herding procedure by Welling (2009b). Herding is a deterministic dynamical system that generates samples given a set of statistical moments (Welling, 2009b,a; Gelfand et al., 2010; Chen and Welling, 2010; Chen et al., 2011, 2012; Bach et al., 2012). DivMbest is exactly a particular case of Herding, with the parameters set in a specific way; namely, divMbest sets the statistical moments to enforce that the set of delivered hypotheses have equiprobable marginals. Note that this is a rather implausible constraint, since enforcing equiprobable marginals produces hypotheses with a proportion of labels independent on the input.

This theoretical insight gained from the generalization allows for designing better algorithms for extracting diverse hypotheses in applications, such as image segmentation. We analyze Herding with Potts-based moments as an alternative to the constraints of the equiprobable marginals of divMbest, but we show that the Potts-based moments enforce different labels in dissimilar neighboring regions that belong to the same object category. This motivates us to introduce the sub-label Potts-based moments, which overcomes the problems of the Potts-based moments and divMbest.

We demonstrate the capabilities of sub-label Potts-based moments on VOC11 dataset (Everingham et al., 2011), in semantic segmentation, and also, in interactive image segmentation (Roig et al., 2013; Maji et al., 2014), where the user provides the ground-truth labeling for a subset of regions in the image. The experiments show that the hypothesis selected by an oracle is
significantly more accurate when the sub-label Potts-based moments are used than with other constraints.

**Contributions:** In summary, the contributions of this paper are: (i) we show that divMbest is a particular case of Herding, (ii) we invoke the properties of Herding to show that divMbest enforces implausible constraints, and (iii) we introduce alternative constraints (the sub-label Potts-based moments) that allow for significant improvements over divMbest in various scenarios in image segmentation.

# 2 Diverse M-best Solutions

This section revisits methods for extracting samples from a CRF, putting special emphasis in the diverse M-best solution procedure by Batra et al. (2012), which is the main focus of this paper.

**Notation:** We use $X = \{X_i\}$ to denote the set of random variables or nodes that represent the labeling of the entities in the CRF. Let $L$ be the set of discrete labels, and $x = \{x_i\}$ an instance of $X$, where $x_i \in L$. We denote $P(x|\theta)$ as the probability density distribution of a labeling $X = x$ modeled with the CRF, in which $\theta$ are the parameters of the model. We represent as $x^*$ the most probable labeling of $P(x|\theta)$.

We express the energy function with the canonical over-complete representation. This is, $E_\theta(x) = \theta^T \phi(x)$, in which $\theta$ are the parameters of the energy function. $\phi(x)$ is the vector of potentials, or the so-called sufficient statistics, and it describes the samples using indicator functions that take values 1 or 0, depending on $x$. The potentials can involve one or two random variables, unary and pairwise potentials, respectively. We use $\phi_u(x)$ to denote unary potentials, and $\phi_p(x)$ for pairwise. There can also be potentials involving more than two random variables, which we omit for simplicity.

We use $\phi_u(x) = (\phi_u(x_1), \ldots, \phi_u(x_i), \ldots)^T$, where each component $\phi_u(x_i)$ is also a vector, the elements of which correspond to an entry of the labels in $L$, i.e. $\phi_u(x_i) = (\phi_{u1}(x_i), \ldots, \phi_{ui}(x_i), \ldots)^T$, and $|L|$ is the cardinality of $L$. Each entry of the vector $\phi_u(x)$ is equal to $\phi_{ui}(x_i) = I[x_i = i]$, where $I[a]$ is an indicator function that is 1 if $a$ is true and 0 otherwise. Note that only a single entry in $\phi_u(x_i)$ will be equal to 1. The pairwise potentials, $\phi_p$, use an *a priori* assumption that depends on the application.

Methods that deliver an estimate of the most probable labeling infer the maximum a posteriori (MAP) of $P(x|\theta)$, or equivalently, maximize the energy function, i.e. $x^* = \arg\max_{x \in L^N} E_\theta(x)$.

**Sampling from $P(x|\theta)$:** Approaches to sample from $P(x|\theta)$ include the Monte-Carlo methods, e.g. (Porreway and Zhu, 2011; Barbu and Zhu, 2005). These methods usually take long periods of sampling to extract different modes of the distribution. There are also methods to extract the $M$-best MAP modes in a CRF (Chen et al., 2013; Yanover and Weiss, 2003; Fromer and Globerson, 2009; Batra, 2012). Yet, the $M$-most probable modes may be very similar to each other, and hence, they may not represent a diverse set of samples.

As a result, several authors stressed the need to incorporate diversity constraints among the $M$-best solutions (Batra et al., 2012; Guzman-Rivera et al., 2014; Prasad et al., 2014). Among these algorithms, divMbest (Batra et al., 2012) has gained a lot of popularity for its simple implementation and high effectiveness. In this paper, we analyze divMbest, which is introduced below.

Recent approaches to extract diverse samples consist of learning multiple models to generate one sample from each model (Guzman-Rivera et al., 2012, 2014). These models are jointly learned using a loss function to yield diverse and accurate samples. However, this comes at the cost of increasing the training time, since multiple models are learned. DivMbest extracts samples from one CRF model rather than multiple models, which does not affect the training time.

**DivMbest algorithm:** Let $x_m^*$ be the sample extracted at the $m$-th iteration of the algorithm. $x_m^*$ is obtained by adding constraints to the MAP inference algorithm, in order to enforce diversity among samples. Concretely, the following constraints are added:

$$\forall q < m : \phi_u(x_m^*)^T \phi_u(x_q^*) < K,$$

where $K$ is a parameter that controls the diversity between samples. It evaluates the similarity between the mappings of the unary potentials, $\phi_u(x)$. This is equivalent to counting the number of nodes that take different labels.

A gradient descent algorithm is used to infer the MAP labeling with the diversity constraints in Eq. (1). To implement this in practice, the diversity constraints in Eq. (1) are relaxed. This relaxation allows for extracting a sample by running MAP inference only one time. The algorithm modifies the energy function at each iteration in order to reduce the probability of the samples obtained previously. This is done by subtracting $\lambda \phi_u(x_m^*)$ from the parameters of the unary potentials, $\theta_u(m)$, where $\lambda$ controls the diversity. The divMbest algorithm is shown in Alg. 1.
3 Herding

In this section, we revisit Herding, which is a deterministic dynamical system that generates samples that follow a given set of statistical moments. This will serve as the basis to show in the following section that divMbest is a particular instance of Herding. Herding has been extensively analyzed in the literature, e.g. (Welling, 2009b,a; Gelfand et al., 2010; Chen and Welling, 2010; Chen et al., 2011, 2012; Bach et al., 2012), see the references for further details on Herding.

**Moments vs Potentials:** Herding uses the same vector $\phi(x)$ to describe a labeling $x$, that we introduced above for the energy function of a CRF. Yet, Herding uses a vector that describes the statistical moments of the samples to generate. Let $\mu$ be the vector of statistical moments that describe a set of samples $\{x\}$, which we divide into unary and pairwise moments, denoted as $\mu_u$ and $\mu_p$, respectively. $\mu$ is computed by averaging the vector of sufficient statistics $\phi(x)$ over $\{x\}$, i.e.

$$
\mu = \mathbb{E}_{x \sim p}[\phi(x)], \quad (2)
$$

where $p$ is the distribution that generated the samples $\{x\}$. In Herding, instead of $p$, a vector of moments $\mu$ is provided, from which it generates a set of samples that follow $\mu$.

Note that since $\phi(x)$ describes the samples using indicator functions that take values 1 or 0, we can see that the moments $\mu$ indicate the proportion of samples that each entry in $\phi(x)$ is equal to 1. This is different from methods that sample from a CRF: Herding defines the properties of the samples using the statistical moments $\mu$, while CRFs define the parameters of the probability distribution of the samples, i.e. the parameters of the potentials, $\theta$.

Note that the unary moments, $\mu_u$, indicate the proportion of samples that $x$ is equal to each label in the set of samples. Let $\mu^i_u$ be an element of $\mu_u$, which corresponds to the marginal distribution of $x_i$. Thus, it is constrained to sum up to 1 ($\|\mu^i_u\|_1 = 1$), where each entry in $\mu^i_u$ is higher or equal to 0. Also, note that the elements of the pairwise moments, denoted as $\mu^i_p$, correspond to the pairwise marginal distribution, and hence, the sum of the entries in $\mu^i_p$ is equal to 1.

**Herding Reconstructs the Moments:** Herding is summarized in Alg. 2. Each iteration of the algorithm consists of two steps, which collectively generate one sample. The first step is a maximization of an energy function, which is done with an off-the-shelf MAP inference algorithm. $x^*_m$ denotes the optimal value obtained from this maximization at the $m$-th iteration of Herding. The second step is the update of parameters of the energy function $\theta_m$, which we use in the optimization in the first step. The update rate is denoted as $\eta$ and remains fixed at a constant value. We initialize $\theta$ equal to $\mu$, in accordance with the literature (Chen et al., 2011, 2012). Thus, Herding does not specify the analytical form of the probability distribution that generates $\{x^*_m\}$, i.e. $p$ is unknown. Yet, Welling (2009b) showed that the samples generated from Herding greedily minimize the following reconstruction error of the moments when $\theta_m$ does not diverge (which is often the case in practical scenarios):

$$
\|\mu - \frac{1}{M} \sum_{m=1}^{M} \phi(x^*_m)\|_2^2. \quad (3)
$$

Herding directly generates the samples $\{x^*_m\}$ that reconstruct the moments $\mu$, without specifying the probability distribution, $p$. In practice, this probability distribution has been shown to be close to the Gibbs distribution with moments $\mu$ (Bach et al., 2012).

**Herding for Diverse Sampling:** Herding can be rewritten in an equivalent form which intuitively shows that Herding generates diverse samples. As shown by Chen et al. (2012), a new sample of Herding, $x^*_{m+1}$, is obtained by inferring the MAP labeling of the fol-
The first term in Eq. (4) encourages the correlation of \( \phi(x) \) with the moments \( \mu \), and hence, facilitates the set of samples to reconstruct the moments. The second term is the similarity between a new sample and the samples that have been generated previously. Thus, Herding encourages that the samples are as different as possible among them, which may help to quickly converge reconstruct the moments. Note that the similarity used in Eq. (4) is the same as in Eq. (1) of divMbest, with the difference that in Eq. (4) all the potentials are taken into account, whereas in Eq. (1) only the unary potentials, \( \phi_u(x) \), are used. In Eq. (4) there is no parameter that controls the diversity constrain. In practice, \( \mu \) is set in different ways to generate different degrees of diversity, as shown after.

4 Herding Generalizes DivMbest

Now we show that divMbest is an instance of Herding. Observe that the differences between the algorithm of divMbest (Alg. 1) and Herding (Alg. 2) are small. Recall that \( \eta \) is the update rate in Herding. Let \( \eta_u \) be the update rate of the unary terms, and \( \eta_p \) for the pairwise. Except for the initialization, we can recover divMbest from Herding by setting the unary moments of Herding and the pairwise update rate to zero, i.e., DivMbest is equal to Herding with the following parameters:

\[
\mu_u = 0, \quad \eta_p = 0.
\]

Observe that this allows to recover the update step of divMbest in Alg. 1 from Herding in Alg. 2. Then, note that \( \eta_u \) is the parameter \( \lambda \) and the pairwise parameters \( \theta_p \) do not change in divMbest because \( \eta_p = 0 \), and hence, the pairwise moments, \( \mu_p \), do not need to be defined in divMbest. The initialization of divMbest differs from the typical initialization in Herding (\( \theta_{(0)} = \mu \)), and \( \theta \) in divMbest are initialized in the same way as the CRF potentials.

Thus, the formulation of divMbest as Herding gives an interpretation of the samples delivered by divMbest as being generated by a deterministic dynamical system that is reconstructing unary moments equal to 0, i.e., \( \mu_u = 0 \). This observation begs the question whether always enforcing \( \mu_u = 0 \) independently on the input is a plausible constraint.

Consequences of \( \mu_u = 0 \) in divMbest: For the sufficient statistics we use, the closest reconstructable moments to \( \mu_u = 0 \) are the unary moments the labels of which have equiprobable probability of occurrence. This can be seen by substituting \( \mu_u = 0 \) and \( \eta_p = 0 \) in Eq. (3), which yields \( \| \frac{1}{M} \sum_{m=1}^{M} \phi_u(x_m^*) \|_2^2 \), and finding the set of samples that minimize this expression, in the same way as Herding. Since the pairwise terms have been cancelled (\( \eta_p = 0 \)), we can minimize Eq. (3) for each unary term independently, i.e., \( \| \frac{1}{M} \sum_{m=1}^{M} \phi_u(x_m^*) \|_2^2 \). Recall that \( \frac{1}{M} \sum_{m=1}^{M} \phi_u(x_m^*) \) lies in the marginal polytope, i.e., \( \| \frac{1}{M} \sum_{m=1}^{M} \phi_u(x_m^*) \|_1 = 1 \). Since Eq. (3) minimizes the \( \ell_2 \) norm, and the marginal polytope lies in the \( \ell_1 \) ball of radius 1, we can see by simple geometry that the point in this \( \ell_1 \) ball closest to 0 in terms of the \( \ell_2 \) distance is the point of equiprobable unary marginals.

Thus, the samples of divMbest greedy converge to equiprobable unary marginals, independently of the parameters of the problem and the input. As a consequence, there are no guarantees that the samples that recover \( \mu_u = 0 \) are representative of the problem.

How does divMbest alleviates enforcing implausible constraints (\( \mu_u = 0 \))? The way that divMbest avoids to recover the moments \( \mu_u = 0 \) is to use a slow update rate, \( \lambda \). This allows to remain far from the attractor that enforces \( \mu_u = 0 \). We can see this by observing that the parameter that controls the diversity, \( \lambda \), is in fact the update rate of Herding, \( \eta_u \). A higher update rate implies that the convergence to the equiprobability of the unary marginals is faster, and hence, the samples are more diverse. Thus, the samples of divMbest avoid \( \mu_u = 0 \) by slowly moving to the direction of an attractor that yields \( \mu_u = 0 \), and remaining close to the initialization. In the experiments we show empirical evidence that support these observations.

5 Herding for Image Segmentation

In this section, we analyze different set-ups of Herding to extract multiple hypotheses for image segmentation. Note that Herding has already been applied in image segmentation to extract one single hypothesis, but not multiple hypotheses (Chen et al., 2011). We use two different modalities of image segmentation: \( (i) \) semantic image segmentation, and \( (ii) \) interactive image segmentation. In both cases, \( X = \{ x_i \} \) represents a semantic labeling of the regions in the image, and \( x_i \) takes a value from a predefined set of labels.

In the case of interactive image segmentation, the user marks few pixels or superpixels providing their true semantic label, which are used to instantiate the corresponding unary terms. The rest of the unary terms remain as missing or unknown (Roig et al., 2013). The probabilistic inference propagates the information in
the known unary terms through the pairwise information. Since there is missing information, delivering several hypotheses of image labeling may help the user to pick the best segmentation hypothesis among all.  

In the following, we analyze three different set-ups of Herding for the applications just described. Two of these set-ups of Herding are a configuration based on Potts moments, and divMBest. We show that these set-ups have different inconveniences for image segmentation. Then, we introduce a new set-up which improves over them, which we coin as sub-label Potts-based moments.

**Herding with Potts-based Moments:** In image segmentation, the marginal distribution of \( x_i \), \( \mu_u \), is obtained from the output of the classifiers for each region taking a certain label (or from the user manual labeling in interactive segmentation). \( \phi_p(x_i, x_j) \) is based on the smoothness of the labeling, and it is the following vector: \( \phi_p(x_i, x_j) = [I[x_i = x_j], I[x_i \neq x_j]]^T \), in which \( i \) and \( j \) index neighboring regions in the image.

We define the pairwise moments, \( \mu^{ij}_p \), using the common Potts formulation to enforce smoothness in the labeling, and we also take into account the color similarity between two neighboring regions. We use \( \mu^{ij}_p \) equal to \((C, 1-C)\), in which \( C \) is a constant value normalized in the range \([0, 1]\), that depends on the similarity between the regions (e.g. \( C \approx 1 \) if the two regions have the same color).

Recall that \( \mu^{ij}_p \) specifies the proportion of occurrences where two neighboring nodes take the same label. Thus, dissimilar regions are forced to take different labels, and similar regions take the same label. As a consequence, using the Potts-based moments may be problematic in objects that are composed of multiple dissimilar regions (e.g. objects with different parts), since these regions are erroneously enforced to be labeled differently. Note that in the case of Herding as divMBest this case is not problematic, because a CRF with color-modulated Potts potential does not introduce any penalty between dissimilar regions, as we show next.

**Herding configured as DivMBest:** Recall that from the generalization, divMBest is Herding with \( \mu_u = 0 \), and \( \mu_p^{ij} \) are not defined because the update rate \( \eta_p \) is 0. Also, \( \theta \) is initialized to the potentials of the CRF. Thus, the initialization of the unary parameters of divMBest algorithm, \( \theta_u(0) \), is the negative logarithm of the scores’ probability given by the local classifiers. \( \theta_p^{ij} \) is initialized to the color-modulated Potts potentials of the CRF, i.e. \( \theta_p^{ij} = (0, -C) \), where \( C \) is a constant that depends on the similarity between the regions. In this way, when the regions are similar (\( C \) with a high value), the Potts potential encourages them to have the same label; and when the regions are dissimilar (\( C \) with values close to 0), this potential does not enforce any constraint. In this way, divMBest does not suffer from the problem of dissimilar object parts that Herding with Potts-based moments has.

However, as described in sec. 4, divMBest enforces implausible constraints to the unary moments, \( \mu_u = 0 \), and this is alleviated by using a small update rate, which keeps the dynamical system close to the initialization. As we will show in the experiments, this strategy fails when there are few observed unary terms. Note that divMBest forces all the unary terms to recover the implausible moments \( \mu_u = 0 \), including the ones that we have no information. Thus, all unary moments, both known and unknown, have to be initialized. The unknown are initialized to an invented value (0 in the experiments) because we do not have information about them. Now it is not only the constraint \( \mu_u = 0 \) that is not representative of the input image, but also the initialization of the unknown unary terms. Using a small update rate to remain close the the initialization is not effective anymore, since the initialization may not take into account the input image.

Note that Herding with Potts-based moments does not suffer from having missing information as divMBest set-up, because in the Potts-based set-up the missing unary moments are not included in the vector of moments, and hence, are not updated nor initialized to invented values. This motivates us to introduce a new set-up based on Potts-based moments, but that alleviates the problem with the objects with dissimilar parts. In Table 1 we summarize the properties of the discussed set-ups of Herding.

**Sub-label Potts-based Moments:** We introduce a set-up built on the Potts-based moments but that does not add constraints between dissimilar regions. It al-

<table>
<thead>
<tr>
<th>Herding Moments</th>
<th>Not enforcing</th>
<th>Region Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DivMBest</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Potts</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Sub-label Potts</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: Properties of the Moments for Herding for Image Segmentation. We analyze three different moments for image segmentation.
Different labels. Now the nodes can bypass this constraint and enforce anymore that two neighboring nodes take different sub-labels, from the same set, two neighboring nodes can take two different sub-labels of same set $S_k$ The amount of sub-labels necessary per class, $S$, depends on the number of neighboring regions in the graph, and how different the parts are within the objects. We analyze the value of $S$ in the experiments.

Finally, we group the different sub-labels of the same set $S_k$ into the same unary term, and we use the parameter given by the label $l_k$ for all the sub-labels. Thus, we redefine the unary terms as: $\phi_k^l(x_i) = I[x_i = l^1_k \lor \ldots \lor x_i = l^{|S_k|}_k]$, in which $\lor$ is the OR operator. Observe that the unary term related to label $l_k$ is equal to 1 if $x_i \in S_k$. In this way, in the MAP inference of Herding, the same parameters of the unary terms are used for all sub-labels in $S_k$. In Fig. 1 we show a summary of the sub-label Potts-based moments.

6 Experiments

In this section, we evaluate Herding for generating hypotheses in the two aforementioned settings: image segmentation and interactive image segmentation with missing potentials. We report results on VOC 2011 (Everingham et al., 2011) which has pixel-wise annotations of 20 object classes plus background. We use VOC11 because there are publicly available pre-trained models (Carreira et al., 2012), which will allow the reproducibility of our experiments. We report results in the validation set, since the ground-truth for the testing set is not provided. We further extend this results in the suppl. material, where we evaluate a second dataset of medical interactive segmentation by Gass et al. (2014), from which similar conclusions as in VOC11 can be extracted.

We use the standard evaluation metrics provided with VOC11. We evaluate the hypotheses using the criteria of **model** and **oracle**. The mode is obtained by selecting the most frequent label among the hypotheses in each pixel, and gives an idea of the performance for the average of the hypotheses. The oracle selects the final labeling, i.e. the hypothesis that is most similar to the oracle (ground-truth). We also report results in the suppl. material using an uncertainty measure. For the rest of this section, we denote Herding with sub-label Potts moments as $s$Herd, and with Potts-based moments as Herd.

6.1 Implementation Details

**Learning the parameters:** We use the publicly available precomputed models by Carreira et al. (2012) in the VOC11 training set. The rest of the parameters are learned via two-fold cross-validation in the validation set of VOC11. The parameters are optimized for the oracle accuracy. In the following, we report the parameters obtained in the first fold, which do not differ significantly from the second fold.
Superpixels: We use non-overlapping superpixels that are provided in CPMC by Carreira and Sminchiescu (2012) for VOC11 (about 600 per image).

Unary Moments: We use classification scores computed with second-order-pooling on the CPMC regions as in (Carreira et al., 2012), using the publicly available code. We select for each superpixel the CPMC region with the highest score, and use the scores in this region as the scores of the superpixel. We use a sigmoid function to map the classification scores to probabilities, i.e. $(1 + \exp(-(a + b \cdot s)))^{-1}$, where $s$ is the output of the classifiers, and $a$ and $b$ are two parameters that we learn with cross-validation. In the suppl. material we show the impact of these parameters. The optimal values are $a = -7$ and $b = 15$. Finally, the output of the sigmoidals are normalized to yield probabilities, which correspond to $\mu_u$ in Herd and $sHerd$. In divMbest, the negative logarithm of the probabilities are used as initialization of the algorithm. The MAP labeling gives 44.8\% accuracy for the VOC11 validation set, which is similar to the one reported in (Carreira et al., 2012).

The unary moments for the interactive image segmentation application correspond to the true annotation for the locations where the human has provided information. We assume that the user assigns some random points of the image and assigning their ground-truth to their unary moments.

Pairwise Moments: To evaluate the similarity between neighbouring regions, we use the exponential of the normalized Euclidean distance between the mean of the RGB values of the connected superpixels, with decreasing factor equal to 10. The optimal weight of the pairwise term for Herd and $sHerd$ is 0.03. For divMbest, it is 0.08. For interactive segmentation, the decreasing factor is 1 and the weight of the pairwise term for Herd and $sHerd$ is 0.03 and for divMbest 0.15.

Inference of the Hypotheses: We use loopy belief propagation by Frey and MacKay (1997) for MAP inference (it takes 10ms to converge in one CPU of a MacBook Air). Note that graph cuts Boykov and Kolmogorov (2004) can not be used due to the non-submodular functions that may arise during Herding updates. The update rate of Herding, $\eta$, is set to $\eta_p = 0.1$ for the pairwise, and $\eta_u = 1$ for the unary terms. In the divMbest we set $\lambda = 0.5$. For interactive segmentation, the update rate of Herd and $sHerd$, $\eta$, is set to $\eta_p = 0.25$ for the pairwise, and $\eta_u = 0.75$ for the unary terms. In divMbest we set $\lambda = 5$. Note that in divMbest the above optimal values of the update rate implies updates much smaller than for Herd and $sHerd$, since in divMbest the potentials are initialized to the logarithm of the probability scores of the marginals used to initialize $Herd$ and $sHerd$. This is in accordance to the observation that divMbest uses small update rates to compensate the constraint $\mu_u = 0$.

6.2 Results

We evaluate the performance of Herd and $sHerd$, both in image segmentation and interactive image segmentation and compare it with divMbest. In all the plots the accuracy is the average accuracy over all classes. We show the per-class accuracy in the suppl. material.

Sub-label Potts-based Moments in Image Segmentation: The oracle and mode performance of $sHerd$ is shown in Fig. 2a and 2b, for different number of sub-labels and hypotheses. We can see that the optimal amount of sub-labels is $S = 5$, since no improvement is achieved by increasing $S$ (we use $S = 5$ for the rest of the experiments). Also, we can observe that $sHerd$ outperforms Herd in both measures by a big margin, which confirms the benefits of using the sub-label Potts-based moments for objects with dissimilar parts. Note that in some cases the performance may decrease with more samples, since the oracle selects the best sample individually, and does not take into account the VOC evaluation over classes. The mode evaluation of Herd is better than the evaluation with oracle, because the pairwise moment in Herd produces erroneous hypotheses, but on average they recover the unary moments. We can see this interesting effect in the examples in Fig. 3.

Herd as divMbest: In Fig. 2a we also compare the oracle performance of divMbest for different number of hypotheses. We can see that $sHerd$ outperforms divMbest. Also, note that divMbest outperforms $Herd$, which means that it penalizes more having implausible constraints between dissimilar object parts than enforcing $\mu_u = 0$.

In Fig. 2b and c we show the mode accuracy, which is related to the unary marginals, and allows to analyze the constraint $\mu_u = 0$. We can see that under the optimal parameters for divMbest (Fig. 2b) the mode accuracy of divMbest is similar to $sHerd$ and Herd, suggesting that divMbest avoided enforcing the constraint $\mu_u = 0$. Yet, when we increase both the update rate of divMbest and we let the dynamical system run longer cycles by increasing $M$ (Fig. 2c), we can observe that the mode performance decreases, suggesting that divMbest is greedily enforcing $\mu_u = 0$. Observe that the mode accuracy sometimes enters in a plateau and does not further decrease. This is not surprising if we recall that in divMbest the potentials are initialized to the logarithm of the probabilities, and hence, there may be some labels which are very difficult to appear since their score may tend to infinity, and many more iterations may be needed for divMbest to recover.
\( \mu_u = 0 \).

In Fig. 3 some examples are depicted. We can see that the difference between the hypotheses of \textit{sHerd} are in the shape of the segmentation, and for the \textit{divMbest} also in the shape but mostly in the class labels. This is in accordance to the observation that \textit{divMbest} enforces to recover unary moments equal to zero.

**Interactive Segmentation with Missing Potentials:** In Fig. 2d we report the accuracy for the oracle criterion for the interactive image segmentation application with \( M = 20 \). We can see that with only 2\% observed unaries, \textit{sHerd} improves the accuracy of \textit{divMbest} more than 30\%! This provides more reassurance about the fact that when the initialization of \textit{divMbest} is implausible, even if we use a small update rate, the constraints enforced by \textit{divMbest} make the dynamical system easily go astray (sec. 5). Note that as expected, when more unary terms are observed, the performance of \textit{divMbest} recovers and becomes similar as in the previous experiment. Several qualitative results are shown in Fig. 3.

**7 Conclusions**

We showed that Herding generalizes the \textit{divMbest} algorithm by Batra et al. (2012), which enabled us to show that \textit{divMbest} enforces implausible constraints, and to introduce alternatives to them. Results on image segmentation benchmarks showed that we can design new set-ups for Herding to outperform \textit{divMbest} for generating segmentation hypotheses. We expect that in the future this generalization can bring Herding to other applications in which multiple hypotheses are used.

**References**


Premachandran, V., Tarlow, D., and Batra, D. (2014). Empirical minimum bayes risk prediction: How to extract an extra few% performance from vision models with just three more parameters. In *CVPR*.


