HAL
open science

On the Equivalence between Herding and Conditional Gradient Algorithms

Francis Bach, Simon Lacoste-Julien, Guillaume Obozinski

To cite this version:

Francis Bach, Simon Lacoste-Julien, Guillaume Obozinski. On the Equivalence between Herding and Conditional Gradient Algorithms. ICML 2012 International Conference on Machine Learning, Jun 2012, Edimburgh, United Kingdom. hal-00681128v2

HAL Id: hal-00681128

https://hal.science/hal-00681128v2

Submitted on 11 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the Equivalence between Herding and Conditional Gradient Algorithms

Francis Bach, Simon Lacoste-Julien, Guillaume Obozinski FIRSTNAME.LASTNAME@INRIA.FR
Sierra project-team, INRIA, Département d'Informatique de l'Ecole Normale Supérieure, Paris, France

Abstract

We show that the herding procedure of Welling (2009b) takes exactly the form of a standard convex optimization algorithm—namely a conditional gradient algorithm minimizing a quadratic moment discrepancy. This link enables us to invoke convergence results from convex optimization and to consider faster alternatives for the task of approximating integrals in a reproducing kernel Hilbert space. We study the behavior of the different variants through numerical simulations. Our experiments shed more light on the learning bias of herding: they indicate that while we can improve over herding on the task of approximating integrals, the original herding algorithm approaches more often the maximum entropy distribution.

1. Introduction

The herding algorithm has recently been presented by Welling (2009b) as a computationally attractive alternative method for learning in intractable Markov random fields models (MRF). Instead of first estimating the parameters of the MRF by maximum likelihood / maximum entropy (which requires approximate inference to estimate the gradient of the partition function), and then sampling from the learned MRF to answer queries, herding directly generates *pseudo-samples* in a deterministic fashion with the property of asymptotically matching the empirical moments of the data (akin to maximum entropy). The herding algorithm generates pseudo-samples x_t with the following simple recursion:

$$\begin{aligned} x_{t+1} &\in \arg\max_{x \in \mathcal{X}} \langle w_t, \Phi(x) \rangle \\ w_{t+1} &= w_t + \mu - \Phi(x_{t+1}), \end{aligned} \tag{1}$$

where $\mathcal{X}$ is the observation space; Φ is a feature map from $\mathcal{X}$ to $\mathcal{F}$, which could be viewed as the vector of sufficient statistics for some exponential family, and μ is a mean vector to match (the empirical moment vector of the same family). Unlike in frequentist learning of MRFs, the parameter w_t never converges to a point in herding and actually follows a "weakly chaotic" walk (Welling & Chen, 2010).

The herding updates can be motivated from two different perspectives. From the *learning perspective*, the herding updates can be derived by performing fixed-step-size subgradient ascent on the zero-temperature limit of the annealed likelihood function of the MRF—called the "tipi function" by Welling (2009b). From this perspective, herding was later generalized to MRFs with latent variables (Welling, 2009a) as well as discriminative MRFs (Gelfand et al., 2010).

From the *moment matching* perspective, which has been explored more in details by Chen et al. (2010), the herding updates can be derived as an effective way to choose *greedily* pseudo-samples x_t in order to quickly decrease the moment discrepancy $\mathcal{E}_t \doteq \|\mu - \frac{1}{t}\sum_{i=1}^{t} \Phi(x_i)\|$ (Chen et al., 2010). Under suitable regularity conditions, $\mathcal{E}_t$ decreases as $O(1/t)$ for the herding updates—this is faster than i.i.d. sampling from the distribution generating μ (e.g., the training data) which would yield the slower $O(1/\sqrt{t})$ rate. This faster rate has been explained by *negative* autocorrelations amongst the pseudo-samples and was used by Chen et al. (2010) to sub-select a small collection of representative "super-samples" from a much larger set of i.i.d. samples. We make the following contributions:

- We show that herding as described by Eq. (1) is equivalent to a specific type of conditional gradient algorithm (a.k.a. Frank-Wolfe algorithm) for the problem of estimating the mean μ. This provides a novel understanding and another explicit cost function that herding is minimizing.
- This interpretation yields improvements, *for the task of estimating means*, with other faster variants of the conditional gradient algorithm, which lead to

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.
Copyright 2012 by the author(s)/owner(s).

non-uniform weights, one based on line-search, one based on an active-set algorithm.

- Based on existing results from convex optimization, we extend and improve the convergence results of herding. In particular, we provide a linear convergence rate for the line-search variant in finite-dimensional settings and show how the conditions assumed by Chen et al. (2010) in fact never hold in the infinite-dimensional setting.
- We run experiments that show that algorithms which estimate faster the mean than herding generate samples that are not better (and typically worse) than the ones obtained with herding when evaluated in terms of the ability to approximate a sample with large entropy, a property which (if or when satisfied by herding) could be the basis for an interpretation of herding as a learning algorithm (Welling, 2009b).

2. Mean estimation

We start with a similar setup as Chen et al. (2010), where herding can be interpreted as a way to approximate integrals of functions in a reproducing kernel Hilbert space (RKHS). We consider a set $\mathcal{X}$ and a mapping Φ from $\mathcal{X}$ to a RKHS $\mathcal{F}$. Through this mapping, all elements of $\mathcal{F}$ may be identified with real functions f on $\mathcal{X}$ defined by $f(x) = \langle f, \Phi(x)\rangle$, for $x \in \mathcal{X}$. We denote by $k : (x, y) \mapsto k(x, y)$ the associated positive definite kernel. Note that the mapping Φ may be explicit (classically in low-dimensional settings) or implicit—where the kernel trick can be used, see Section 4.3 and Chen et al. (2010).

Throughout the paper, we assume that the data is uniformly bounded in feature space, that is, for all $x \in \mathcal{X}$, $\|\Phi(x)\| \leqslant R$, for some $R > 0$; this condition is needed for the updates of Eq. (1) to be well-defined.

We denote by $\mathcal{M} \subset \mathcal{F}$ the *marginal polytope* (Wainwright & Jordan, 2008; Chen et al., 2010), i.e., the convex-hull of all vectors $\Phi(x)$ for $x \in \mathcal{X}$. Note that for any $f \in \mathcal{F}$, we have

$$\sup_{x\in\mathcal{X}} f(x) = \sup_{g\in\mathcal{M}} \langle f, g\rangle,$$

and that $|f(x)| = |\langle f, \Phi(x)\rangle| \leqslant \|f\|R$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$ (i.e., all functions with finite norm are bounded).

Extreme points of the marginal polytope. In all the cases we consider in Section 5, it turns out that all points of the form $\Phi(x)$, $x \in \mathcal{X}$ are extreme points of $\mathcal{M}$ (see an illustration in Figure 1). This is indeed always true when $\|\Phi(x)\|$ is constant for all $x \in \mathcal{X}$ (for example for our infinite-dimensional kernels on $[0, 1]$ in Section 5.1); it is also true if $\Phi(x)$ contains both an injective feature map $\widetilde{\Phi}(x)$ and its self-tensor-product $\widetilde{\Phi}(x) \otimes \widetilde{\Phi}(x)$, which is the case in the graphical model examples in Section 5.2.

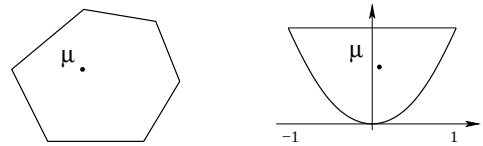

Figure 1. Marginal polytope in two situations: (left) finite number of extreme points (typical with discrete data), (right) polynomial kernel in one dimension, with $\Phi(x) = (x, x^2)$ for $x \in [-1, 1]$.

Mean element and expectation. We consider a fixed probability distribution $p(x)$ over $\mathcal{X}$. Following Chen et al. (2010), we denote by μ the mean element (see, e.g., Smola et al., 2007):

$$\mu = \mathbb{E}_{p(x)}\Phi(x) \in \mathcal{M}.$$

Note that in the learning perspective, p is the empirical distribution on the data and so μ is the corresponding empirical moment vector to match. To approximate this mean, we consider n points $x_1, \dots, x_n \in \mathcal{X}$ combined linearly with positive weights $w_1, \dots, w_n$ that sum to one. These define $\hat{p}$, the associated weighted empirical distribution, and $\hat{\mu}$ the approximating mean:

$$\hat{\mu} = \mathbb{E}_{\hat{p}(x)}\Phi(x) = \textstyle\sum_{i=1}^n w_i\Phi(x_i) \in \mathcal{M}. \qquad (2)$$

For all functions $f \in \mathcal{F}$, we then have

$$\mathbb{E}_{p(x)} f(x) = \mathbb{E}_{p(x)}\langle f, \Phi(x)\rangle = \langle \mu, f\rangle,$$

and similarly $\mathbb{E}_{\hat{p}(x)} f(x) = \langle \hat{\mu}, f\rangle$. We thus get, using Cauchy-Schwarz inequality,

$$\sup_{f\in\mathcal{F},\ \|f\|=1} |\mathbb{E}_{p(x)} f(x) - \mathbb{E}_{\hat{p}(x)} f(x)| = \|\mu - \hat{\mu}\|,$$

and controlling $\mu - \hat{\mu}$ is enough to control the error in computing the expectation for all $f \in \mathcal{F}$ with finite norm. Note that a random i.i.d. sample from $p(x)$ would lead to an expected worst-case error which is less than $\frac{4R}{\sqrt{n}}$—a classical result based on Rademacher averages (see, e.g. Boucheron et al., 2005).

In this paper, we will try to find a good estimate $\hat{\mu}$ of μ based on a *weighted* set of points from $\{\Phi(x), x \in \mathcal{X}\}$, generalizing Chen et al. (2010), and show how this relates to herding.

3. Related work

This paper brings together three lines of work, namely the approximation of integrals, herding and convex optimization. The links between the first two were clearly outlined by Chen et al. (2010), while the present paper provides the novel interpretation of herding as a well-established convex optimization algorithm.

3.1. Quadrature/cubature formulas

The evaluation of expectation, or equivalently of integrals, is a classical problem in numerical analysis. When the input space $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$ and $p(x)$ is the density of the distribution with respect to the Lebesgue measure, then this is equivalent to evaluating the integral $\int_{\mathcal{X}} f(x)p(x)dx$. Quadrature formulas are aimed at computing such integrals as a weighted combinations of values of f at certain points, which is exactly the problem we consider in Section 2.

Although a thorough review of quadrature formulas is outside of the scope of this paper, we mention two methods which are related to our approach. First, given a positive definite kernel and a given set of points (typically sampled i.i.d. from a given distribution), the Bayes-Hermite quadrature of O'Hagan (1991) essentially computes an orthogonal projection of μ onto the affine hull of this set of points. This does not lead to positive quadrature weights, and one may thus replace the affine hull by the convex hull to obtain such non-negative weights, which we do in our experiments in Section 5.

Moreover, quasi-Monte Carlo methods consider sequences of so-called "quasi-random" quadrature points so that the empirical average approaches the integral. These quasi-random sequences are such that the approximation error goes down as $O(1/n)$ (up to logarithmic terms) for functions of bounded variation, as opposed to $O(1/\sqrt{n})$ for a random sequence. In simulations, we used a Sobol sequence (see, e.g., Morokoff & Caflisch, 1994).

3.2. Franke-Wolfe algorithms

Given a smooth (twice continuously differentiable) *convex* function J on a *compact convex set* $\mathcal{M}$ with gradient J', Frank-Wolfe algorithms are a class of iterative optimization algorithms that only require (in addition to the computation of the gradient J') to be able to optimize linear functions on $\mathcal{M}$. The first class of such algorithms is often referred to as *conditional gradient* algorithms: given an iterate g_t, the minimum of $\langle J'(g_t), g\rangle$ over $g \in \mathcal{M}$ is computed, and the next iterate is taken on the segment between g_t and g, i.e., $g_{t+1} = \rho_t g_t + (1-\rho_t)g$, where $\rho_t \in [0,1]$. There are two natural choices for ρ_t, (a) simply taking $\rho_t = 1/(t+1)$ and (b) performing a line search to find the point in the segment with optimal value (either for J or for a quadratic upper-bound of J). These are illustrated in Figure 2, and convergence rates are detailed in Section 4.2. Moreover, for quadratic functions, the conditional gradient algorithm with step sizes $\rho_t = 1/(t+1)$ has a dual interpretation as subgradient ascent (see, e.g., Bach, 2011), which we outline in Section 4.1.

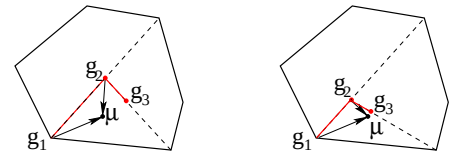

Figure 2. Two versions of two iterations of conditional gradient after moving to an initial corner g_1; left: $\rho_t = \frac{1}{t+1}$, right: line search. The minimum-norm-point algorithm would have converged to μ after two iterations.

Finally, in order to minimize the number of iterations, a variant known as the *minimum-norm-point* algorithm will find g_{t+1} that minimizes J on the convex hull of *all* previously visited points, using a specific active-set algorithm (see Bach, 2011, Sec. 6, for details). For convex sets with finitely many extreme points, it converges in a finite number of iterations with higher (though still polynomial) iteration computational cost (Wolfe, 1976).

4. Herding as a Frank-Wolfe algorithm

To relate herding with conditional gradient algorithms, we consider the following optimization problem:

$$\min_{g \in \mathcal{M}} J(g) = \frac{1}{2}\|g-\mu\|^2, \tag{3}$$

with the trivial unique solution μ. A conditional gradient algorithm to solve this optimization problem with stepsize $\rho_t = 1/(t+1)$ use the iterates

$$\begin{aligned} \bar{g}_{t+1} &\in \arg\min_{g \in \mathcal{M}} \langle g_t - \mu, g\rangle, \\ g_{t+1} &= (1-\rho_t)\, g_t + \rho_t\, \bar{g}_{t+1}. \end{aligned} \tag{4}$$

But these updates are exactly the same as herding via the change of variable $g_t = \mu - w_t/t$. Indeed, the minimizer of a linear function of a convex set $\bar{g}_{t+1}$ can be restricted to be an extreme point of $\mathcal{M}$; this implies that $\bar{g}_{t+1} = \Phi(x_{t+1})$ for a certain x_{t+1}. The herding updates in Eq. (1) are thus equivalent to the conditional gradient minimization of J with step size given by $\rho_t = 1/(t+1)$.

With this choice of step size, we get $(t+1)g_{t+1} = tg_t + \Phi(x_{t+1})$, that is $g_t = \frac{1}{t}\sum_{u=1}^t \Phi(x_u)$, and we thus get uniform weights in Eq. (2).

For general step-sizes $\rho_t \in [0,1]$, if we assume that we start the algorithm with $\rho_0 = 1$ (which implies $g_1 = \Phi(x_1)$), then we get that $g_t = \sum_{u=1}^t \big(\prod_{v=u}^{t-1}(1-\rho_{v-1})\rho_{u-1}\big)\Phi(x_u)$, which thus leads to non-uniform weights in Eq. (2), though they still sum to one. The conditional gradient updates in Eq. (4) can thus be seen as a generic algorithm to obtain a weighted set of points to approximate μ, and traditional herding is the $\rho_t = 1/(t+1)$ step-size case.

A second choice of step-size for $t \geq 1$ is to use a line search, which leads in this setting (where the global

unconstrained minimum happens to belong to $\mathcal{M}$) to $\rho_t = \frac{\langle g_t - \mu, g_t - \bar{g}_{t+1}\rangle}{\|\bar{g}_{t+1} - g_t\|^2} \in [0,1]$. This leads to a variant of herding with non-uniform weights.

We finally comment on the initialization g_0 for the updates in Eq. (4). In the kernel herding algorithm of Chen et al. (2010), the authors use $w_0 = \mu$ as this is required to interpret the herding updates as greedily minimizing $\mathcal{E}_t$ (with the additional assumptions that $\|\Phi(x)\|$ is constant). In our setting, this corresponds to choosing $g_0 = 0$ (which might be outside of $\mathcal{M}$, though this is not problematic in practice). Another standard choice (for MRFs in particular) is to use $w_0 = 0$ ($g_0 = \mu$), which means that the first point x_1 is chosen randomly from the extreme points of $\mathcal{M}$—this is the scheme we used. As is common in convex optimization, we didn't see any qualitative difference in our experiments between the two types of initialization.

4.1. Dual problem and subgradient descent

Welling (2009b) proposed originally an algorithmic interpretation of herding as performing subgradient ascent with constant step size on a function obtained as the zero temperature limit of the log-likelihood of an exponential model that he called the "tipi function". Our interpretation of the procedure as a Frank-Wolfe algorithm might therefore appear as a conflicting interpretation at first sight. To establish a natural connection between these two interpretations, we can compute the Fenchel-dual optimization problem to Eq. (3). Indeed, we have (with standard arguments for swapping the min and max operations):

$$\begin{aligned}\min_{g\in\mathcal{M}} \frac{1}{2}\|g-\mu\|^2 &= \min_{g\in\mathcal{M}} \max_{f\in\mathcal{F}} f^\top(g-\mu) - \frac{1}{2}\|f\|^2 \\ &= \max_{f\in\mathcal{F}} \min_{g\in\mathcal{M}} f^\top(g-\mu) - \frac{1}{2}\|f\|^2 \\ &= \max_{f\in\mathcal{F}} \big\{ \min_{x\in\mathcal{X}} f(x) - \langle f,\mu\rangle - \frac{1}{2}\|f\|^2 \big\}.\end{aligned}$$

The dual function $f \mapsto \min_{x\in\mathcal{X}} f(x) - \langle f,\mu\rangle - \frac{1}{2}\|f\|^2$ is 1-strongly concave and non-differentiable, and a natural algorithm to maximize it is thus subgradient ascent with a step size equal to $\frac{1}{t+1}$, which is known to be equivalent to the primal conditional gradient algorithm with step sizes $\rho_t = 1/(t+1)$ (Bach, 2011, App. A). It is therefore not surprising that herding is equivalent to subgradient ascent with a decreasing stepsize on this function (with the identification $f_t = g_t - \mu = -w_t/t$). The presence of the squared norm which is added to the "tipi function" merely reflects the change of scaling between g_t and w_t. It is worthwhile mentioning that other functions, like Bregman divergences, would have led to a different dual function hence adding a different term than a squared norm to the "tipi function".

4.2. Convergence analysis

Without further assumptions on the problem, then the two choices of step sizes lead to a convergence rate of the form (Dunn, 1980; Bach, 2011):

$$\frac{1}{2}\|g_t - \mu\|^2 \leqslant 4\frac{R^2}{t},$$

where R is diameter of the marginal polytope. Note that the convergence in $O(1/t)$ does not lead to improved estimation of integrals over random sampling. Moreover, without further assumptions, current theoretical analysis does not allow to distinguish between the two forms of conditional gradient algorithms (although they differ a bit in practice, see Section 5).

However, if we assume that within the affine hull of $\mathcal{M}$, there exists a ball of center μ and radius $d > 0$ that is included in $\mathcal{M}$ (i.e., μ is in the relative interior of $\mathcal{M}$), then both choices of step sizes yield faster rates than random sampling. For the version with line search, we actually obtain a linear convergence rate (Beck & Teboulle, 2004):

$$\frac{1}{2}\|g_t - \mu\|^2 \leqslant R^2 \exp\big(-\frac{d^2 t}{R^2}\big).$$

For the version without line search ($\rho_t = 1/(t+1)$), Chen et al. (2010) shows the slower convergence rate in $O(1/t^2)$:

$$\frac{1}{2}\|g_t - \mu\|^2 \leqslant \frac{2R^4}{d^2 t^2}.$$

Concerning the assumption that μ is in the relative interior of $\mathcal{M}$, we now show that in finite-dimensional settings, this assumption is always satisfied under reasonable conditions, while it is *never* satisfied in a large class of infinite-dimensional settings (namely for Mercer kernels).

We first provide an equivalent definition of this condition which is used in the proofs. Let $\mathcal{A}$ be the affine hull of $\mathcal{M}$, μ_0 the orthogonal projection of 0 on $\mathcal{A}$, and $\widetilde{\mathcal{F}}$ the space of directions (or difference space) of $\mathcal{A}$ (i.e., $\widetilde{\mathcal{F}} = \mathcal{A} - \mu_0$).[1] Now there exists $d > 0$ so that any element of $\mathcal{A}$ which is at distance less than d of μ is in $\mathcal{M}$ if and only if $\forall f \in \mathcal{F}$, the maximum of $f^\top g$ over $g \in \mathcal{A}$ and $\|g - \mu\| \leqslant d$ is less than the maximum of $f^\top g$ over $g \in \mathcal{M}$. Given the properties of $\mathcal{A}$ and $\widetilde{\mathcal{F}}$, this is equivalent to:

$$\begin{aligned}&\forall f \in \widetilde{\mathcal{F}},\ \max_{\|g-\mu\|\leqslant d} f^\top g \leqslant \max_{g\in\mathcal{M}} f^\top g \\ \Leftrightarrow\ &\forall f \in \widetilde{\mathcal{F}},\ \langle \mu, f\rangle + d\|f\| \leqslant \max_{x\in\mathcal{X}} f(x). \qquad (5)\end{aligned}$$

Proposition 1 *Assume that $\mathcal{F}$ is finite-dimensional, that $\mathcal{X}$ is a compact topological measurable space with*

[1] It turns out that μ_0 is a function taking a constant value since the orthogonality condition yields $\langle \mu_0, \Phi(x) - \Phi(y)\rangle = 0$, i.e., $\mu_0(x) = \mu_0(y)$ for all $x, y \in \mathcal{X}$, by the reproducing property of $\mathcal{F}$.

a continuous kernel function, and that the distribution p has full support on $\mathcal{X}$. Then $\exists d > 0$ so that Eq. (5) is true.

Proof sketch. It is sufficient to show that $\Omega : f \mapsto \max_{x \in \mathcal{X}} f(x) - \langle \mu, f \rangle$ is a norm on $\widetilde{\mathcal{F}}$: as all norms are equivalent in finite dimensions, we get $d\|f\| \leq \Omega(f)$ for some $d > 0$, yielding Eq. (5). Ω is convex and positively homogeneous by construction. Now $\Omega(f) = 0$ implies that $\mathbb{E}_{p(x)}[f(x) - \max_y f(y)] = 0$, and thus $f(x) = \max_y f(y)$ for x in the support of p (assumed to be $\mathcal{X}$) using the fact that f is continuous (since the kernel is continuous), and so f is a constant function. We then have two possibilities: either $\mu_0 = 0$, in which case one can show that there is no non-zero constant functions in $\mathcal{F}$; otherwise $f = \alpha\mu_0$ for some α and thus the orthogonality condition $\langle f, \mu_0 \rangle = 0$ implies that $\alpha = 0$. Both cases imply $f = 0$, hence Ω is a norm. ■

Proposition 2 *Assume $\mathcal{X}$ is a compact subspace of $\mathbb{R}^d$, and that the kernel k is a continuous function on $\mathcal{X} \times \mathcal{X}$. If $\mathcal{F}$ is infinite-dimensional, then there exists no $d > 0$ so that Eq. (5) is true.*

Proof sketch. We can apply Mercer theorem to the kernel $\tilde{k}(x, y)$ of the projection onto the orthogonal of $\{\mu, \mu_0\}$. This kernel is also a Mercer kernel, and we get an orthonormal basis $(e_k)_{k \geqslant 1}$ of $L^2(\mathcal{X})$ with *countably many* eigenvalues λ_k that are summable. Moreover, $(\lambda_k^{1/2} e_k)_{k \geqslant 1}$ is known to be an orthonormal basis of the associated feature space $\mathcal{F}$ (Cucker & Smale, 2002), and for all $x, y \in \mathcal{X}$, $\tilde{k}(x, y) = \sum_{k \geqslant 1} \lambda_k e_k(x) e_k(y)$, with uniform convergence. This implies that for $f_k = \lambda_k^{1/2} e_k$, we have $\|f_k\| = 1$, and $\langle f_k, \mu_0 \rangle = \langle f_k, \mu \rangle = 0$.

If we assume that there exists $d > 0$ so that Eq. (5) is true, then we have for all $k \geqslant 1$, $\max_{x \in \mathcal{X}} |f_k(x)| \geqslant d$. Since $\mathcal{X}$ is compact, there exists a cover of $\mathcal{F}$ with finitely many balls of radius $d/4R$. Let $\mathcal{Y}$ be the finite set of centers. Since all functions f_k are Lipschitz-continuous with constant $2R$, then for all $k \geqslant 1$, $\max_{x \in \mathcal{Y}} |f_k(x)| \geqslant d - 2R \times d/4R = d/2$. Since $\mathcal{Y}$ is finite, there exists $x \in \mathcal{Y}$ so that $|f_k(x)| \geqslant d/2 > 0$ for infinitely many values of k; this contradicts the summability of $\sum_{k \geqslant 1} f_k(x)^2$. Hence the result. ■

The last proposition shows that the current theory only supports the slower rates of $O(1/t)$ for the two conditional gradient algorithms in infinite-dimensional settings. On the other hand, we note that, in some cases, traditional herding performs empirically better without known theoretical justification (see Section 5).

4.3. Computational issues

In order to run a herding algorithm, there are two potential computational bottlenecks:

Computing expectations $\langle \mu, \Phi(x) \rangle$: in a learning context (empirical moment matching), these are done through an empirical average. In an integral evaluation context, in finite-dimensional settings, one needs to compute $\mathbb{E}_{p(x)} \Phi(x)$; while in an infinite-dimensional setting, following Chen et al. (2010), expectations of the form $\mathbb{E}_{p(x)} k(x, y)$ need to be computed. This can be done for some pairs of kernels/distributions, like the ones we choose in Section 5, but not in general.

Minimizing $\langle g_t - \mu, \Phi(x) \rangle$ **with respect to** $x \in \mathcal{X}$: in general, this computation may be relatively hard (it is for example NP-hard in the context of the graphical models we consider in Section 5). In practice, Chen et al. (2010) and Welling (2009a) perform local search, while another possibility is to perform the minimization through exhaustive search in a finite sample. Note that a convex relaxation through variational methods (Wainwright & Jordan, 2008) could provide an interesting alternative.

5. Experiments

The goals of these simulations are (a) to compare the different algorithms aimed at estimating integrals, i.e., assess herding for the task of mean estimation (Section 5.1 and Section 5.2), and (b) to briefly study the empirical relationship with maximum entropy estimation in a learning context (Section 5.3).

5.1. Kernel herding on $\mathcal{X} = [0, 1]$

Problem set-up. In this section, we consider $\mathcal{X} = [0, 1]$ and the norm $\|f\|^2 = \int_0^1 [f^{(\nu)}(x)]^2 dx$ on the infinite-dimensional space of functions with zero mean and whose ν-th derivative exists and is in $L^2([0, 1])$. As shown by Wahba (1990), the associated kernel is equal to $\frac{B_{2\nu}(x - y - \lfloor x - y \rfloor)}{(2\nu)!}$, where $B_{2\nu}$ is the (2ν)-th Bernoulli polynomial, with $B_2(x) = x^2 - x + \frac{1}{6}$ and $B_6(x) = x^6 - 3x^5 + \frac{5}{2}x^4 - \frac{1}{2}x^2 + \frac{1}{42}$.

We consider either the uniform density on $[0, 1]$, or a randomly selected density of the form $p(x) \propto \big(\sum_{i=1}^d a_i \cos(2i\pi x) + b_i \sin(2i\pi x)\big)^2$, for which all required expectations may be computed in closed form. In particular, the mean element is computed as $\mu : x \mapsto \mathbb{E}[k(Y, x)]$ which may be computed in closed form by expanding all terms in the Fourier basis. As for the optimization step, it consists in minimizing $g_t(x) - \mu(x)$ over the interval $[0, 1]$ which can be done efficiently with exhaustive search.

Comparison of mean estimation procedures. We compare in Figure 3 two kernels, i.e., with $\nu = 1$ (left and middle plots) and $\nu = 3$ (right plot), the following mean estimation procedures, and plot $\log_{10} \|\hat{\mu} - \mu\|$, for two densities, the uniform density (middle and right) and a randomly selected non-uniform density (left). We compare the following:

- `cg-1/(t+1)`: conditional gradient procedure with $\rho_t = \frac{1}{t+1}$, which is the original herding procedure of Welling (2009a), leading to uniform weights.
- `cg-l.search`: conditional gradient procedure with line search (with non-uniform weights).
- `min-norm-point`: Minimum-norm point algorithm. This leads to non-uniform weights.
- `random`: Random selection of points (from $p(x)$), averaged over 10 replications.
- `sobol`: a classical quasi-random sequence, with uniform weights. For non-uniform densities, we first apply the inverse cumulative distribution function.

For all of these (except for min-norm-point), we also consider an extra *a posteriori* projection step (denoted by the `-proj` suffix), which computes the optimal set of non-uniform weights by finding the best approximation of μ in the convex hull of the points selected by the algorithm. We can draw the following conclusions:

- Min-norm-point algorithms always perform best.
- Conditional gradient with line search is performing slightly worse than regular herding. (Note that we are in the infinite-dimensional setting and so they both have $O(t^{-1})$ as theoretical rate.)
- The extra projection step always significantly improves performance, and sometimes enough that random selection of point combined with a reprojection outperforms regular herding (at least for $\nu = 3$, i.e., with a smaller feature space).
- On the right plot, it turns out that the Sobol sequence is known to achieve the optimal rate of $O(t^{-2})$ for $\|\mu - \hat{\mu}\|^2$ for the associated Sobolev space (Wahba, 1990). In this situation, regular herding empirically achieves the same rate; however, the theoretical analysis provided in the present paper or by Chen et al. (2010) does not allow to explain or support this statement theoretically.

Estimation from a finite sample. In Figure 4, we compare three of the previously mentioned herding procedures when all quantities are computed from a random sample of size 1000. In plain, testing errors are computed (using exact expectations) while in dashed, the training errors are computed. All methods eventually fit the empirical mean, with no further progress on the testing error, this behavior happening faster with the min-norm-point algorithm.

5.2. Estimation on graphical models

We consider $\mathcal{X} = \{-1, 1\}^d$ and a random variable computed as the sign (in $\{-1, 1\}$) of a Gaussian random vector in $\mathbb{R}^d$, together with $\Phi(x)$ composed of x and of all of its pairwise products $xx^\top$. In this set-up, we can compute the expectation $\mathbb{E}_{p(x)}\Phi(x)$ in closed form, as the mass assigned to the positive orthant by a bivariate Gaussian distribution with correlation ρ, which is

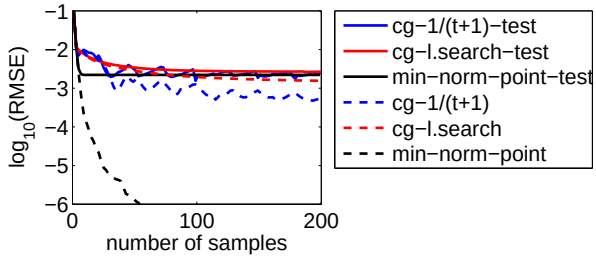

Figure 4. Comparison of *estimated* (from a finite sample) herding procedures for kernels on $[0, 1]$ with $\nu = 3$ and *non-uniform* density.

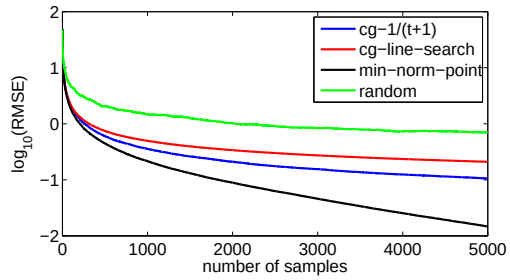

Figure 5. Comparison of herding procedures on graphical models with 100 binary variables. See Section 5.2.

equal to $\frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho$ (Abramowitz & Stegun, 1964). We are here in the *finite-dimensional* setting and the faster rates derived in Section 4.2 apply.

We generated 10000 samples from such a distribution and performed herding with exact expectations but with minima with respect to x computed over this sample (by exhaustive search over the sample). We plot results in Figure 5, where we see the superiority of the min-norm-point procedure over the two other procedures (which include regular herding). Note that the line-search algorithm is slower than the $1/(t+1)$-rule, which seems to contradict the bounds. The bounds depend on the distance d between the mean and the boundary of the marginal polytope. If this is too small (much like if the strong convexity constant is too small for gradient descent), the linear convergence rate can only be seen for larger numbers of iterations.

5.3. Herding and maximum/minimum entropy

Given a moment vector μ obtained from the empirical mean of $\Phi(x)$ on data, the goal of herding is to produce a pseudo-sample whose moments match μ without having to estimate the canonical parameters of the corresponding model. A natural candidate for such a distribution is the maximum entropy distribution and we will compare it to the results of herding in cases where it can be easily computed, namely for $\mathcal{X} = \{-1, 1\}^d$ (with $d \leqslant 10$) and either $\Phi(x) = x \in [-1, 1]^d$ or $\Phi(x) = (x, xx^\top)$. In this setup, following Welling (2009a), the distribution on $x \in \mathcal{X}$ is estimated by the empirical distribution $\sum_{i=1}^t w_i \delta(x = x_i)$.

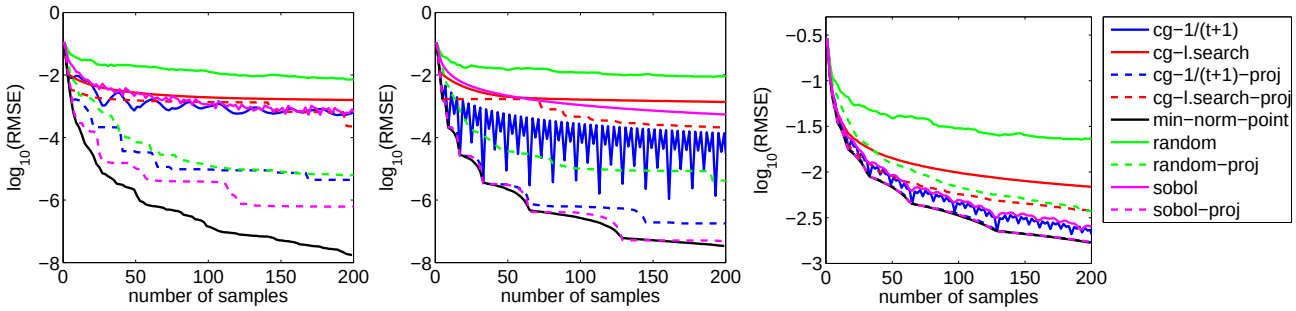

Figure 3. Comparison of population herding procedures for kernels on $[0, 1]$. From left to right: $\nu = 3$ and *non-uniform* density, $\nu = 3$ and *uniform* density, $\nu = 1$ decay of eigenvalues and *uniform* density. Best seen in color.

Learning independent bits. We first consider $\Phi(x) = x$ and some specific feasible moment $\mu \in \mathcal{M}$. It is well-known that the maximum entropy distribution is the one with independent bits. In the top panels of Figure 6, we compare the norm between the maximum entropy probability vector and the one estimated by two versions of herding, namely conditional gradient with stepsize $\rho_t = 1/(t+1)$ (regular herding with uniform weights) and with line search (with non-uniform weights)—the min-norm-point algorithm leads to quantitatively similar results. We show results in Figure 6 for a mean vector μ which is a random uniform vector in $[-1, 1]^d$ (left plots), and for a mean μ which is random with uniform $(\mu_i + 1)/2$ values in $\{1, 2, 3, 4, 5\} \times \frac{2\sqrt{2}}{3}$ (middle plots), and for mean values μ which is are random with uniform $(\mu_i + 1)/2$ values in $\{1, 2, 3, 4, 5\}/6$ (right plots). Note that the difference between rational and irrational means was already brought up by Welling & Chen (2010) through the link between herding and Sturmian sequences.

For each of the mean vector, we plot in the top plots, the error in estimating the full maximum entropy distribution (a vector of size 2^d), and in the bottom plots, the error in estimating the feature means (a vector of size d). We can draw the following conclusions:

- For a random vector μ (left plots), then regular herding (with no line search) empirically converges to the maximum entropy distribution.
- For rational ratios between the means (but irrational means, middle plots), then there is no convergence to the maximum entropy distribution.
- For rational means (right), there is no convergence either, but the behavior is more erratic.
- The line-search procedure never converges to the maximum entropy procedure. On the opposite, it happens to be close to a minimum entropy solution, where many states have probability zero.

Experiments considered in Figure 6 considered a single draw of the mean vector μ, but similar empirical conclusions may be drawn from other random samples, and we conjecture that for *almost surely* all random vectors $\mu \in [-1, 1]^d$ (which would avoid rational ratios between mean values), then regular herding converges to the maximum entropy distribution. The next experiment shows that this is not the case in general.

Learning non-independent bits. We now consider $\Phi(x) = (x, xx^\top)$, and a certain random feasible moment $\mu \in \mathcal{M}$. As before, we compare the norm between the maximum entropy probability vector and the one estimated by the two versions of herding. We present results in Figure 7 for a mean vector obtained by the corresponding exponential family distribution with zero weights for the features x and constant weights on the features $xx^\top$. We see that the herding procedures, while leading to a consistent estimation of the mean vector, does not converge to the maximum entropy distribution and other unreported experiments have led to similar results.

6. Conclusion

We showed that herding generates a sequence of points which give in sequence the descent directions of a conditional gradient algorithm minimizing the quadratic error on the moment vector. Therefore, if herding is only assessed in terms of its ability to approximate the moment vector, it is outperformed by other more efficient algorithms. Clearly, herding was originally defined with another goal, which was to generate a pseudo-sample whose distribution could approach the maximum entropy distribution with a given moment vector. Our experiments suggest empirically, that while this is the case in certain cases, herding fails in other case, which are not chosen to be particularly pathological. This probably prompts for a further study of herding.

Our experiments also show that algorithms that are more efficient than herding at approximating the moment vector fail more blatantly to approach a maximum entropy distribution and they present characteristics which would rather suggest a minimization of the entropy. This suggests the question of whether there is a tradeoff between approximating most efficiently the

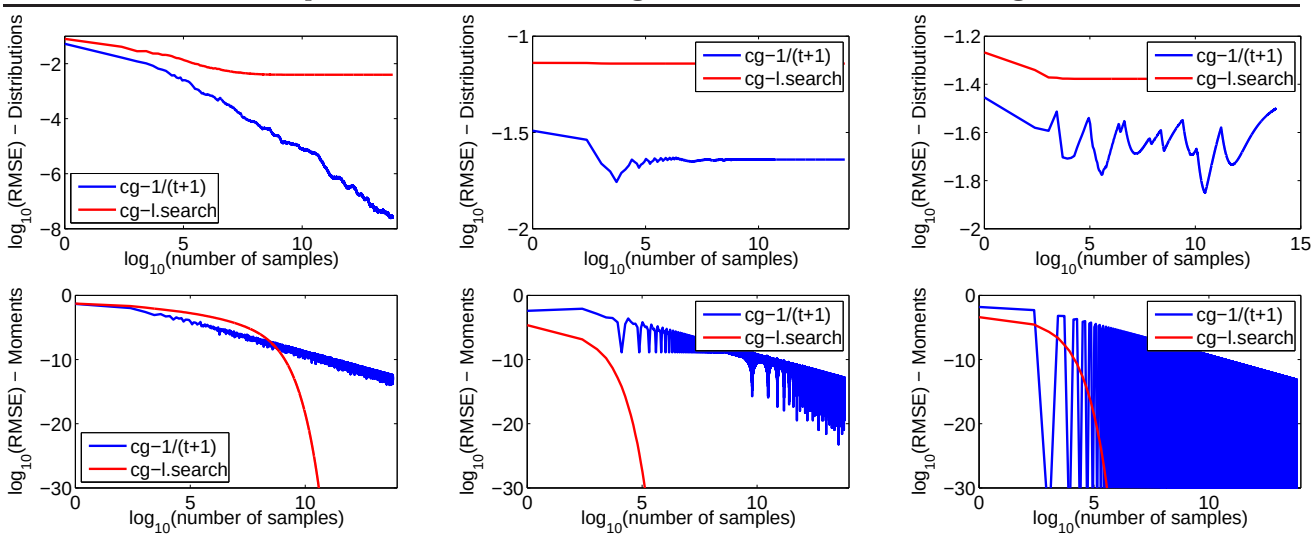

Figure 6. Comparison of herding procedures on the independent bit problem with $d = 10$ binary variables. Top: estimation of the maximum entropy distribution, bottom: estimation of the mean of the features $\Phi(x)$. From left to right: Mean values are selected uniformly at random on $[-1, 1]$, mean values are equal to $\sqrt{2}$ times random rational numbers in $[-1, 1]$, mean values are equal to random rational numbers in $[-1, 1]$.

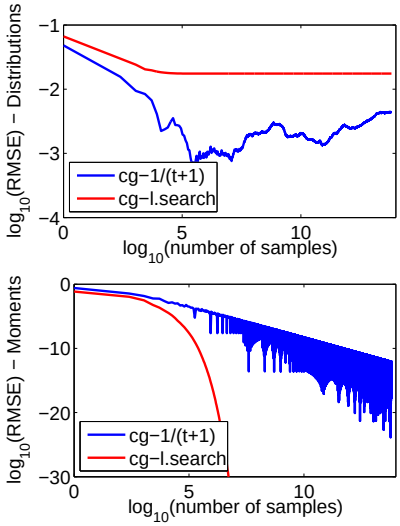

Figure 7. Comparison of herding procedures on graphical models with 10 binary variables. Top: estimation of the maximum entropy distribution, bottom: estimation of the mean of the features $\Phi(x)$.

mean vector and approximating well the maximum entropy distribution, or if the two goals are in fact rather aligned? In any case, we hope that formulating herding as an optimization problem can help form a better understanding of its goals and its properties.

Acknowledgements. We thank Ferenc Huszár and Zoubin Ghahramani for helpful discussions. This work was supported by the European Research Council (SIERRA Project) and the city of Paris ("Research in Paris" program).

References

Abramowitz, M. and Stegun, I.A. *Handbook of mathematical functions*. Dover publications, 1964.

Bach, F. Learning with submodular functions: A convex optimization perspective. Technical Report 1111.6453, Arxiv, 2011.

Beck, A. and Teboulle, M. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Math. Meth. Op. Res.*, 59(2):235–247, 2004.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: A survey of some recent advances. *ESAIM Probability and statistics*, 9:323–375, 2005.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. In *Proc. UAI*, 2010.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *Bull. AMS*, 39(1), 2002.

Dunn, J. C. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM J. Control & Opt.*, 18:473–487, 1980.

Gelfand, A., van der Maaten, L., Chen, Y., and Welling, M. On herding and the perceptron cycling theorem. In *Adv. NIPS*, 2010.

Morokoff, W.J. and Caflisch, R.E. Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, 15(6):1251–1279, 1994.

O'Hagan, A. Bayes-Hermite quadrature. *J. Stat. Planning & Inference*, 29(3):245–260, 1991.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Wahba, Grace. *Spline Models for Observational Data*. SIAM, 1990.

Wainwright, M.J. and Jordan, M.I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

Welling, M. Herding dynamic weights for partially observed random field models. In *Proc. UAI*, 2009a.

Welling, M. Herding dynamical weights to learn. In *Proc. ICML*, 2009b.

Welling, M. and Chen, Y. Statistical inference using weak chaos and infinite memory. *J. Physics: Conf. Series*, 233, 2010.

Wolfe, P. Finding the nearest point in a polytope. *Math. Progr.*, 11(1):128–149, 1976.