Super-Samples from Kernel Herding

Chen, Welling, Smola, ICML 2010

(Arthur Gretton's notes)

September 5, 2012

What is herding?

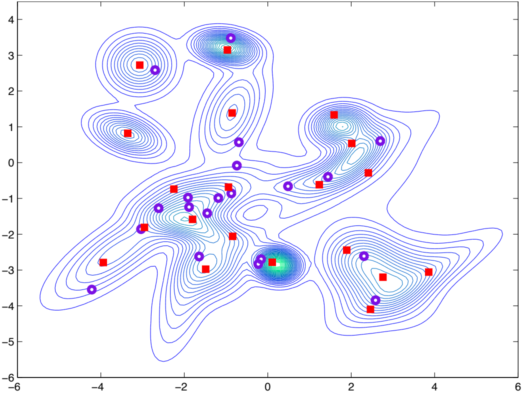

Figure: Herding example: 20 points from Herding vs 20 i.i.d. samples. Contour is density, red squares are from Herding, purple circles are i.i.d. samples.

What is herding?

Herding in an RKHS $\mathcal{F}$ is the following iteration:

1. $x_{T+1} = \operatorname*{argmax}_{x \in \mathcal{X}} \langle w_T, \phi(x) \rangle$
2. $w_{T+1} = w_T + \mathbb{E}_{x \sim P}(\phi(x)) - \phi(x_{T+1})$

Recall: mean embedding:

$$\mu_P := \mathbb{E}_{x \sim P}\phi(X)$$

which has the property

$$\mathbb{E}_x f(x) = \langle f, \mu_p \rangle \quad \forall f \in \mathcal{F}.$$

Hence 2nd step is: $w_{T+1} = w_T + \mu_p - \phi(x_{T+1})$

What does it do?

Define $w_0 := \mu_P$.
Then Herding becomes:

$$
\begin{aligned}
x_{T+1} &= \underset{x \in \mathcal{X}}{\operatorname{argmax}} \langle w_T, \phi(x) \rangle \\
&= \underset{x \in \mathcal{X}}{\operatorname{argmax}} \left\langle w_0 + T\mu_P - \sum_{t=1}^{T} \phi(x_t), \phi(x) \right\rangle \\
&= \underset{x \in \mathcal{X}}{\operatorname{argmax}} \left\langle (T+1)\mu_P - \sum_{t=1}^{T} \phi(x_t), \phi(x) \right\rangle \\
&= \underset{x \in \mathcal{X}}{\operatorname{argmax}} \left((T+1)\mathbb{E}_{x'} k(x, x') - \sum_{t=1}^{T} k(x_t, x) \right)
\end{aligned}
$$

What does it do? (2)

Let's say we want to choose x_{T+1} greedily to minimize:

$$\mathcal{E}_{T+1} := \left\| \mu_P - \frac{1}{T+1} \sum_{t=1}^{T+1} \phi(x_t) \right\|$$
$$= \mathbb{E}_{x,x'} k(x,x') - \frac{2}{T+1} \sum_{t=1}^{T+1} \mathbb{E}_x k(x, x_t) + \frac{1}{(T+1)^2} \sum_{t,t'}^{T} k(x_t, x_{t'}).$$

Keep terms that are a function of x_{T+1}:

$$\frac{-2}{T+1} \mathbb{E}_x k(x, x_{T+1}) + \frac{2}{(T+1)^2} \sum_{t=1}^{T} k(x_t, x_{T+1}) + \frac{1}{(T+1)^2} \underbrace{k(x_{t+1}, x_{t+1})}_{\text{constant}}$$

Why might it be useful?

Given a finite sample estimate $\hat{p}_T$ of p, then for all $f \in \mathcal{F}$, by Cauchy-Schwarz:

$$|\mathbb{E}_{x \sim p} f(x) - \mathbb{E}_{x \sim \hat{p}_T} f(x)| \leq \|f\| \, \|\mu_p - \mu_{\hat{p}_T}\| \, .$$

If $x_T \sim p$ i.i.d., then

$$\|\mu_p - \mu_{\hat{p}_T}\| = O_P(T^{-1/2}).$$

The claim for Herding:

$$\|\mu_p - \mu_{\hat{p}_T}\| = O(T^{-1})$$

Proof of fast convergence

Assume:

1. $\|\phi(x)\| \leq R$.
2. There exists an ϵ-ball around μ_p contained in $\mathcal{M} = \mathrm{conv}\{\phi(x)\}$ (this will cause problems).

If we can show $\|w_t\|$ is bounded, then we can show the result.

Proof of fast convergence

Assume:

1. $\|\phi(x)\| \leq R$.
2. There exists an ϵ-ball around μ_p contained in $\mathcal{M} = \mathrm{conv}\{\phi(x)\}$ (this will cause problems).

If we can show $\|w_t\|$ is bounded, then we can show the result.

Proof as follows:

1. Show why bounded $\|w_t\|$ gives the result we need
2. Show that $\|w_t\|$ is bounded under the ϵ-ball assumption

Proof that bounded $\|w_t\|$ gives fast convergence

Let's say that $\|w_T\|$ is bounded. Then

$$\|w_T\| = \left\| w_0 + T\mu_p - \sum_{t=1}^{T} \phi(x_t) \right\| \leq C$$

So, dividing by T:

$$\left\| \mu_p - \frac{1}{T} \sum_{t=1}^{T} \phi(x_t) \right\| \leq T^{-1}(\|w_0\| + C).$$

Proof that $\|w_t\|$ bounded:

We will prove there exists a constant C: which satisfies two properties:

1. If $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$.
2. If $\|w_t\| < C$, then $\|w_{t+1}\|^2 < C^2 + (2R)^2$

Interpretation:

1. The first result guarantees that if $\|w_t\|$ exceeds the limit C, it will shrink until it falls under the limit.
2. The second result guarantees that $\|w_t\|$ cannot grow too much in one time step (straightforward since the update has bounded norm).

The net effect is that $\|w_t\| \leq C + 2R$ for all t.
Note that $\|w_t\|$ never coverges!

Proof that if $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$

First: express the update in terms of a diference wrt μ_p:

$$\mathcal{C} := \mathcal{M} - \mu_p = \operatorname{conv}\left\{\phi(x) - \mu_p \middle| x \in \mathcal{X}\right\}$$

Then update equations are:

$$\begin{aligned} w_{t+1} &= w_t + \mathbb{E}_{x\sim P}(\phi(x)) - \phi(x_{t+1}) \\ &= w_t - c_t \end{aligned}$$

where $c_t = \operatorname*{argmax}_{c\in\mathcal{C}} \langle w_t, c_t\rangle$.

Proof if $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$ (2)

$$\begin{aligned}
\|w_t\|^2 - \|w_{t+1}\|^2 &= \|w_t\|^2 - \|w_t - c_t\|^2 \\
&= 2\langle w_t, c_t\rangle - \|c_t\|^2 \\
&= \|c_t\| \left[2\|w_t\| \left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle - \|c_t\| \right]
\end{aligned}$$

Proof if $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$ (2)

$$\begin{aligned}
\|w_t\|^2 - \|w_{t+1}\|^2 &= \|w_t\|^2 - \|w_t - c_t\|^2 \\
&= 2\langle w_t, c_t\rangle - \|c_t\|^2 \\
&= \|c_t\| \left[2\|w_t\| \left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle - \|c_t\|\right]
\end{aligned}$$

Next: since $\|\phi(x)\| \leq R$, then $\|\mu_p\| \leq R$ and hence $\|c_t\| \leq 2R$. So

$$\|w_t\|^2 - \|w_{t+1}\|^2 \geq 2\|c_t\| \left[\|w_t\| \left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle - R\right].$$

Proof if $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$ (3)

Is it the case that $\left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle \geq \gamma^* > 0$?

Reminder:

$$c_t = \operatorname*{argmax}_{c \in \mathcal{C}} \langle w_t, c_t \rangle$$
$$\mathcal{C} := \operatorname{conv}\left\{ \phi(x) - \mu_p \middle| x \in \mathcal{X} \right\}.$$

I.e. can c_t be chosen in the direction w_t?

Proof if $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$ (3)

Is it the case that $\left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle \geq \gamma^* > 0$?
Reminder:

$$c_t = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \langle w_t, c_t \rangle$$
$$\mathcal{C} := \operatorname{conv} \left\{ \phi(x) - \mu_p \middle| x \in \mathcal{X} \right\}.$$

I.e. can c_t be chosen in the direction w_t?
Yes, as long as μ_p is in the relative interior of $\mathcal{M}$ (the problematic assumption)

Proof if $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$ (4)

$$
\begin{aligned}
\|w_t\|^2 - \|w_{t+1}\|^2 &\geq 2\|c_t\| \left[\|w_t\| \left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle - R \right] \\
&\geq 2\|c_t\| \left[\|w_t\| \gamma^* - R \right]
\end{aligned}
$$

Proof if $\|w_t\| > C$, then $\|w_{t+1}\| < \|w_t\|$ (4)

$$\begin{aligned}\|w_t\|^2 - \|w_{t+1}\|^2 &\geq 2\|c_t\| \left[\|w_t\| \left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle - R\right] \\ &\geq 2\|c_t\| \left[\|w_t\| \gamma^* - R\right]\end{aligned}$$

What if $\|w_t\| > R/\gamma^* =: C$? Then

$$\begin{aligned}\|w_t\|^2 - \|w_{t+1}\|^2 &> 2\|c_t\| [R - R] \\ &= 0\end{aligned}$$

so $\|w_t\|^2 > \|w_{t+1}\|^2$.
QED

Proof that if $\|w_t\| < C$, then $\|w_{t+1}\|^2 < C^2 + (2R)^2$

Now prove the second result.
Recall $C = R/\gamma^*$. Then

$$\begin{aligned}\|w_{t+1}\|^2 &= \|w_t - c_t\|^2 \\ &= \|w_t\|^2 - 2\langle w_t, c_t\rangle + \|c_t\|^2 \\ &\leq \|w_t\|^2 - 2\|c_t\|\|w_t\|\gamma^* + \|c_t\|^2 \\ &\leq \left(\frac{R}{\gamma^*}\right)^2 + (2R)^2\end{aligned}$$

QED

...and a warning

$$\left\| \mu_p - \frac{1}{T}\sum_{t=1}^{T}\phi(x_t) \right\| \leq T^{-1}(\|w_0\| + R/\gamma^*).$$

so we need $\gamma^* > 0$ for fast rates.
From Bach, Lacoste-Julien, Obozinski, ICML2012: this never holds for Mercer kernels.

Does it work?

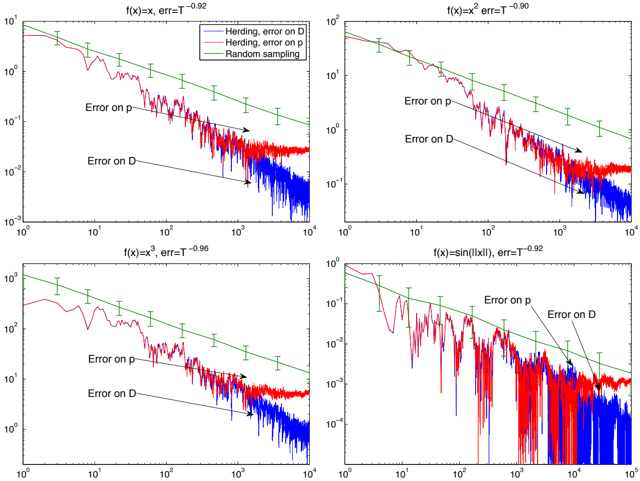

Figure: Herding results: empirical mean embeddings computed from 10^5 samples. Note that Herding uses fewer samples to get same accuracy.