



Probability Collectives

An uniformed overview

David S. Leslie
28 November 2007





Disclaimer

I knew nothing about this topic two weeks ago.

Talk based almost entirely on

Wolpert, Strauss and Rajnarayan (2006). Advances in optimization using probability collectives. Advances in Complex Systems.



Objective

$G : \mathcal{X} \rightarrow \mathcal{R}$.

Find $\min_x G(x)$.

Find distribution $q \in \mathcal{Q}$ to minimize

$$\mathbb{E}_q(G(X)).$$

Thought: $\operatorname{argmin}_{q \in \mathcal{Q}} \mathbb{E}_q[G(X)] = \delta_{\operatorname{argmin}_{x \in \mathcal{X}} G(x)}$.



Distributed optimization

Assume $x = (x_1, x_2, \dots, x_N)$.

Each variable x_i is controlled by a different agent.

The collective of agents must optimize G . Assume limited communication.

$q(x) = \prod_{i=1}^N q_i(x_i)$ where each q_i is a probability distribution.

Maxent Lagrangian

$$\text{Find } \min_{\{q_i\}} \left\{ \int G(x) \prod_i q_i(x_i) dx \right\}$$

such that $\int q_i(x_i) dx_i = 1$ and $q_i(x_i) \geq 0 \forall i, x_i$.



Maxent Lagrangian

$$\text{Find } \min_{\{q_i\}} \left\{ \int G(x) \prod_i q_i(x_i) dx + \sum_i T_i \int \phi_i(q_i(x_i)) dx_i \right\}$$

such that $\int q_i(x_i) dx_i = 1$.

[$\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ is a barrier function which is non-negative, and infinite for negative values of the argument.]



Maxent Lagrangian

$$\text{Find } \min_{\{q_i\}} \left\{ \int G(x) \prod_i q_i(x_i) dx + \sum_i T_i \int q_i(x_i) \log(q_i(x_i)) dx_i \right\}$$

such that $\int q_i(x_i) dx_i = 1$.

$$[\phi(q) = q \log(q), \int \phi(q(x)) dx = -S(q).]$$



Maxent Lagrangian

$$\text{Find } \min_{\{q_i, T_i\}} \left\{ \mathbb{E}_q[G(X)] - \sum_i T_i S(q_i) \right\}$$

such that $\int q^i(x^i) dx_i = 1, T_i \geq 0$.

[Let $T = (T_1, \dots, T_N)$ be an independent variable.]

Solution: $T_i = \infty \dots!$



Maxent Lagrangian

$$\text{Find } \min_{\{q_i, T\}} \left\{ \mathbb{E}_q[G(X)] - T \sum_i S(q_i) + \sum_i \lambda_i \left(\int q_i(x_i) dx_i - 1 \right) \right\}$$

such that $T \geq 0$.

[Set all $T_i = T$. Introduce Lagrange multipliers λ_i .]

This is the Maxent Lagrangian.



Vanilla approach

1. Fix T
2. Find optimizing q
3. Decrease T and return to 2.



Solution for q

q is a solution of the equations

$$q_i(x_i) \propto \exp \left\{ -\mathbb{E}_{q_{-i}}[G \mid x_i]/T \right\}.$$

where

$$\mathbb{E}_{q_{-i}}[G \mid x_i] = \int G(x_i, x_{-i}) \prod_{j \neq i} q_j(x_j) dx_{-i}$$

with $x_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$.

Brouwer's fixed point theorem shows existence of solution.



“Brouwer updating”

Iterative updates. Let q^n be the estimate at iteration n .

Three options:

1. $q_i^{n+1}(x_i) \propto \exp \left\{ -\mathbb{E}_{q_{-i}^n} [G | x_i] / T \right\}$ for each i . “Thrashing”.

2. Update one q_i (chosen sequentially/randomly) at each n .

3. $q_i^{n+1} = (1 - \alpha^n) q_i^n + \alpha^n b^i(q^n)$

Can prove 2 and 3 converge to a fixed point.



Gibbs, Kullback and Leibler

Choose q to approximate the Gibbs distribution

$$p_T \propto \exp\{-G/T\}.$$

The Kullback–Liebler divergence of p from q is

$$KL(q||p) = \int q(x) \log \left(\frac{q(x)}{p(x)} \right) dx.$$

$$KL(q||p^T) \propto L(q, T).$$

Minimizing the Lagrangian is minimizing the KL divergence of p from q .



KL is asymmetric

We could instead minimize the KL divergence of q from p .

This corresponds to choosing q_i to minimize

$$- \int p_i^T(x_i) \log(q_i(x_i)) dx_i$$

which implies

$$q_i(x_i) = p_i^T(x_i) = \int p^T(x) dx_{-i},$$

the marginal of p^T .

NB The optimal q_i doesn't depend on q_{-i} .



Why minimize $\mathbb{E}_q(G)$?

Replace $\mathbb{E}_q(G)$ with $\mathbb{P}_q(G < K)$ for some constant K .

All theory goes through as before.

Now need to update K as we go as well.



 Etc



Automatic annealing

$$L(q, T) = \mathbb{E}_q[G(X)] - T \sum_i S(q_i) + \sum_i \lambda_i \left(\int q_i(x_i) dx_i - 1 \right)$$

Simultaneous gradient descent in both q and T : choose stepsize α and set $q^{n+1} = q^n - \alpha(\partial_q L)$, $T^{n+1} = T^n - \alpha(\partial_T L)$.

$$\begin{aligned} \partial_{q_i(x_i)} L &= \{ \mathbb{E}[G | x_i] + \log(q_i(x_i)) + \eta_i \} \\ &\text{where } \eta_i \text{ is chosen so that } \sum_{x_i} \delta q_i(x_i) = 0. \end{aligned}$$

$$\partial_T L = -S(q) \quad (\text{in paper they use } +S(q)).$$

Estimation of expectations

Individuals do not know q_{-i} (and often G)

Expectations estimated by sampling.

Repeat M times:

- Each individual samples x_i from $q_i(x_i)$.
- $G(x)$ is calculated.
- All agents told of $(x, G(x))$.



Estimation of expectations

Brouwer updates

$\mathbb{E}_{q_{-i}}[G \mid x_i] \approx$ average of G values when x_i was sampled.

Estimation of the marginals

$$p_i^T(x_i) = \int \exp \{ -G(x_i, x_{-i}) / T \} dx_{-i}$$

\approx average of $\frac{\exp\{-G(x_i, x_{-i})/T\}}{q_{-i}(x_{-i})}$ values.

(Need to share $q_i(x_i)$ values too.)



Countable \mathcal{X}

If M large enough compared to finite \mathcal{X} should see all x_i in the sample

\Rightarrow can estimate each necessary expectation.

If \mathcal{X} infinite, or just large, will be some x_i which you don't get sampled.

“Shrink-wrapping” essentially doesn't bother changing $q_i(x_i)$ if x_i not seen.



Uncountable \mathcal{X}

Estimate $\mathbb{E}[G | x_i]$ using nonparametric regression.

Use the estimate to update q in the usual way.



Sampling from q_i

\mathcal{X} discrete is easy.

\mathcal{X} continuous is not.

Paper suggests “subsampling”. This is rejection sampling.

Proposal distribution either q_i^0 or an approximation to q_i^k .

Sequential Monte Carlo is probably directly relevant.

