Information-Based Multi-Fidelity Bayesian Optimization

Yehong Zhang[†], **Trong Nghia Hoang**[§], **Bryan Kian Hsiang Low**[†] **and Mohan Kankanhalli**[†]
Department of Computer Science, National University of Singapore, Republic of Singapore[†]
Laboratory of Information and Decision Systems, Massachusetts Institute of Technology, USA[§]
{yehong, lowkh, mohan}@comp.nus.edu.sg[†], nghiaht@mit.edu[§]

Abstract

This paper presents a novel generalization of *predictive entropy search* (PES) for multi-fidelity *Bayesian optimization* (BO) called *multi-fidelity PES* (MF-PES). In contrast to existing multi-fidelity BO algorithms, our proposed MF-PES algorithm can naturally trade off between exploitation vs. exploration over the target and auxiliary functions with varying fidelities without needing to manually tune any such parameters or input discretization. To achieve this, we first model the unknown target and auxiliary functions jointly as a *convolved multi-output Gaussian process* (CMOGP) whose convolutional structure is then exploited for deriving an efficient approximation of MF-PES. Empirical evaluation on synthetic and real-world experiments shows that MF-PES outperforms the state-of-the-art multi-fidelity BO algorithms.

1 Introduction

Bayesian optimization (BO) has recently demonstrated to be highly effective in optimizing an unknown (possibly non-convex and with no closed-form derivative) target function using a finite budget of often expensive function evaluations [13]. In practice, this expensive-to-evaluate target function often correlates well with some auxiliary function(s) of varying fidelities (i.e., degrees of accuracy in reproducing the target function) that may be less noisy and/or cheaper to evaluate and can thus be exploited to boost the BO performance. For example, to accelerate the hyperparameters tuning of a *machine learning* (ML) model [15], one may consider a low-fidelity auxiliary function with the same inputs (i.e., hyperparameters) and output (i.e., validation accuracy) as the target function except that its validation accuracy is evaluated by training the ML model with a small subset of the dataset, hence incurring less time [16]. Similarly, the parameter setting/configuration of a real robot [10, 18] can be calibrated faster by simulating its motion in a low-fidelity but low-cost and noise-free simulation environment [4]. The above practical examples motivate the need to design and develop a *multi-fidelity* BO algorithm that selects not just the most informative inputs but also the target or auxiliary function(s) with varying fidelities and costs to be evaluated at each selected input for finding or improving the belief of the global target maximizer, which is the focus of our work here.

To do this, a number of multi-fidelity BO algorithms have been proposed [7, 8, 9, 11, 16]. However, their performance are either highly sensitive (and hence not robust) to the manual/heuristic selection of parameters to trade off between exploration vs. exploitation over the target and auxiliary functions with varying fidelities [7, 8, 9, 16] or dependent on input discretization [11, 16], both of which are undesirable in practice especially if there is no prior knowledge about how to optimize them for a specific application. In this paper, we present a novel generalization of *predictive entropy search* (PES) for multi-fidelity BO, which, in contrast to the state-of-the-art multi-fidelity BO algorithms, does not suffer the above limitations: Our proposed *multi-fidelity PES* (MF-PES) does not require input discretization and can jointly and naturally optimize the non-trivial exploration-exploitation trade-off without needing to manually tune any such parameters to perform well in different real-world applications.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

To achieve this, we model the unknown target and auxiliary functions jointly as a *convolved multi-output Gaussian process* (CMOGP) [1] whose convolutional structure is exploited to formally characterize the fidelity of each auxiliary function through its cross-correlation with the target function (Section 2). Then, we derive an efficient approximation of MF-PES by exploiting (a) a novel *multi-output random features* (MRF) approximation of the CMOGP model whose cross-correlation structure between the target and auxiliary functions can be exploited for improving the belief of the target maximizer (Section 3.1), and (b) new practical constraints relating the global target maximizer to that of the auxiliary functions (Section 3.2). We empirically evaluate the performance of our MF-PES algorithm in Section 4.

2 Multi-Fidelity Modeling with Convolved Multi-Output Gaussian Process

Among the various models [2, 3, 14, 17] that can jointly model target and auxiliary functions, CMOGP [1] is used in this work due to its convolutional structure which can be exploited for deriving an efficient approximation of MF-PES. Let M unknown functions $f_1, \ldots, f_M$ with varying fidelities be jointly modeled as a CMOGP over a bounded input domain $D \subset \mathbb{R}^d$ such that each input $x \in D$ is associated with a noisy output $y_i(x) \sim \mathcal{N}(f_i(x), \sigma^2_{n_i})$ for $i = 1, \ldots, M$. CMOGP defines each i-th function f_i as a convolution between a smoothing kernel K_i and a latent function L:

$$f_i(x) \triangleq \int_{x' \in D} K_i(x - x')\, L(x')\, \mathrm{d}x' \,. \tag{1}$$

Let $D_i^+ \triangleq \{\langle x, i\rangle\}_{x \in D}$ and $D^+ \triangleq \bigcup_{i=1}^M D_i^+$. As shown by [1], if $\{L(x)\}_{x \in D}$ is a GP with prior covariance $\sigma(x, x') \triangleq \mathcal{N}(x - x' | \underline{0}, \Lambda^{-1})$ and $K_i(x) \triangleq \sigma_{s_i} \mathcal{N}(x | \underline{0}, P_i^{-1})$. Then, $\{f_i(x)\}_{\langle x,i\rangle \in D^+}$ is also a GP whose covariance function can be computed using $\sigma_{ij}(x, x') = \sigma_{s_i}\sigma_{s_j}\mathcal{N}(x - x' | \underline{0}, \Lambda^{-1} + P_i^{-1} + P_j^{-1})$ which characterizes both the correlation structure within each function (i.e., $i = j$) and the cross-correlation between different functions (i.e., $i \neq j$).

Let t be the index of the target function and x_{*_i} be the maximizer of function f_i. Interestingly, the fidelity of an auxiliary function f_i with respect to target function f_t in the context of BO can naturally be characterized by the following normalized covariance between $f_i(x_{*_i})$ and $f_t(x_{*_t})$:

$$\rho_i \triangleq \sigma_{ij}(x_{*_i}, x_{*_t})/(\sigma'_{s_i}\sigma'_{s_t}) \in [0, 1] \tag{2}$$

where $\sigma'_{s_i} \triangleq \sigma_{s_i}/(2\pi|\Lambda^{-1} + 2P_i^{-1}|)^{1/4}$. Note that our defined fidelity measure ρ_i tends to 1 (i.e., higher fidelity of f_i) when (a) the convolutional structure of f_i parametrized by P_i becomes more similar to that of f_t (i.e., P_t) and (b) the maximizer x_{*_i} of f_i is closer to the target maximizer x_{*_t}.

Given a vector $y_X \triangleq (y_i(x))^\top_{\langle x,i\rangle \in X}$ of observed noisy outputs, a CMOGP can predict the *posterior* distribution of $f_Z \triangleq (f_i(x))^\top_{\langle x,i\rangle \in Z}$ for any set $Z \subseteq D^+$ of input tuples as $\mathcal{N}(\mu_{Z|X}, \Sigma_{ZZ|X})$ with:

$$\mu_{Z|X} \triangleq \mu_Z + \Sigma_{ZX}(\Sigma_{XX} + \Sigma_\epsilon)^{-1}(y_X - \mu_X), \quad \Sigma_{ZZ|X} \triangleq \Sigma_{ZZ} - \Sigma_{ZX}(\Sigma_{XX} + \Sigma_\epsilon)^{-1}\Sigma_{XZ} \tag{3}$$

where $\Sigma_{AA'} \triangleq (\sigma_{ij}(x, x'))_{\langle x,i\rangle \in A, \langle x',j\rangle \in A'}$ and $\mu_A \triangleq (\mu_i(x))^\top_{\langle x,i\rangle \in A}$ for any $A, A' \subseteq D^+$.

3 Multi-Fidelity Bayesian Optimization with Predictive Entropy Search

A multi-fidelity BO algorithm repeatedly selects the next input tuple $\langle x, i\rangle$ for evaluating the i-th function f_i at x that maximizes an acquisition function $\alpha(y_X, \langle x, i\rangle)$ given the past observations (X, y_X): $\langle x, i\rangle^+ \triangleq \arg\max_{\langle x,i\rangle \in D^+ \setminus X} \alpha(y_X, \langle x, i\rangle)$ and updates $X \leftarrow X \cup \{\langle x, i\rangle^+\}$ until the budget is expended. Intuitively, the multi-fidelity acquisition function α should be constructed to enable the multi-fidelity BO algorithm to jointly and naturally optimize the non-trivial trade-off between exploitation vs. exploration over the target and auxiliary functions with varying fidelities for finding or improving the belief of the global target maximizer x_{*_t}. To do this, we follow the idea of information-based acquisition functions [5, 6] in conventional BO and try to maximize information gain of *only* the target maximizer x_{*_t} from observing the next input tuple $\langle x, i\rangle$:

$$\alpha(y_X, \langle x, i\rangle) \triangleq H(x_{*_t}|y_X) - \mathbb{E}_{p(y_i(x)|y_X)}[H(x_{*_t}|y_X, y_i(x))]. \tag{4}$$

Unfortunately, the approximation of (4) is very expensive and sometimes inaccurate since both Monte Carlo sampling and a small set of well selected input candidates are required [16]. To circumvent this issue, we can exploit the symmetric property of conditional mutual information and rewrite (4) as

$$\alpha(y_X, \langle x, i\rangle) = H(y_i(x)|y_X) - \mathbb{E}_{p(x_{*_t}|y_X)}[H(y_i(x)|y_X, x_{*_t})] \tag{5}$$

which we call *multi-fidelity PES* (MF-PES).

Due to (3), the first Gaussian predictive/posterior entropy term in (5) can be computed analytically: $H(y_i(x)|y_X) \triangleq 0.5\log(2\pi e(\sigma^2_{\langle x,i\rangle|X} + \sigma^2_{n_i}))$ where $\sigma^2_{\langle x,i\rangle|X} \triangleq \Sigma_{\{\langle x,i\rangle\}\{\langle x,i\rangle\}|X}$. Although the second term appears to resemble that in PES [6], their approximation method, however, cannot be applied straightforwardly since it cannot account for the cross-correlation structure between the target and auxiliary functions. To achieve this, we will first propose a novel MRF approximation of the CMOGP model whose cross-correlation (i.e., multi-fidelity) structure between the target and auxiliary functions can be exploited for sampling the target maximizer x_{*_t} from $p(x_{*_t}|y_X)$ more accurately, which is in turn used to approximate the expectation in (5). Then, we will formalize some practical constraints relating the global target maximizer to that of the auxiliary functions, which are used to approximate the second entropy term within the expectation in (5).

3.1 Multi-output random features (MRF) for sampling the target maximizer

Using the results of *single-output random features* (SRF) [12], the latent function L in (1) modeled using GP can be approximated by a linear model $L(x) \approx \phi(x)^\top\theta$ where $\phi(x)$ is a random vector of an m-dimensional feature mapping of the input x for $L(x)$ and $\theta \sim \mathcal{N}(\underline{0}, I)$ is an m-dimensional vector of weights. Then, interestingly, by exploiting the convolutional structure of the CMOGP model in (1), $f_i(x)$ can also be approximated analytically by a linear model:

$$f_i(x) = \int_{x'\in D} K_i(x-x')L(x)\,\mathrm{d}x' \approx \int_{x'\in D} K_i(x-x')\phi(x)^\top\theta\,\mathrm{d}x' \approx \phi_i(x)^\top\theta$$

where $\phi_i(x) \triangleq \sigma_{s_i}\,\mathrm{diag}(e^{-\frac{1}{2}W^\top P_i^{-1}W})\,\phi(x)$ can be interpreted as input features of $f_i(x)$, W is a $d\times m$ random matrix which is used to map $x \to \phi(x)$ in SRF [12] and function $\mathrm{diag}(A)$ returns a diagonal matrix with the same diagonal components as A.

Then, a sample of f_i can be constructed using $f_i^{(s)}(x) \triangleq \phi_i^{(s)}(x)^\top\theta^{(s)}$ where $\phi_i^{(s)}(x)$ and $\theta^{(s)}$ are vectors of features and weights sampled, respectively, from the random vector $\phi_i(x)$ and the posterior distribution of weights θ given the past observations (X, y_X), the latter of which is derived to be Gaussian by exploiting the conditionally independent property of MRF: $p(\theta|y_X) = \mathcal{N}(\theta|A^{-1}\Phi\Sigma_\epsilon^{-1}y_X, A^{-1})$ where $A = \Phi\Sigma_\epsilon^{-1}\Phi^\top + I$ and $\Phi \triangleq (\phi_j(x))_{\langle x,j\rangle\in X}$. Consequently, the expectation in (5) can be approximated by averaging over S samples of the target maximizer $x_{*_t}^{(s)}$ which can be achieved by optimizing $f_t^{(s)}$ with any existing gradient-based optimization method.

3.2 Approximating the predictive entropy conditioned on the target maximizer

Next, we will discuss how the second entropy term in (5) is approximated. Firstly, the posterior distribution $p(y_i(x)|y_X, x_{*_t}) = \int p(y_i(x)|f_i(x))\; p(f_i(x)|y_X, x_{*_t})\;\mathrm{d}f_i(x)$ where $p(y_i(x)|f_i(x))$ is Gaussian and $p(f_i(x)|y_X, x_{*_t})$ can be approximated using *expectation propagation* (EP) by considering it as a constrained version of $p(f_i(x)|y_X)$, as detailed later.

It is intuitive that the posterior distribution of $f_i(x)$ is constrained by $f_i(x) \le f_i(x_{*_i}), \forall\langle x,i\rangle \in D^+$. However, since only the target maximizer x_{*_t} is of interest, how should $f_i(x)$ be constrained by x_{*_t} instead of x_{*_i} if $i \neq t$? To resolve this, we introduce a slack variable c_i to formalize the relationship between maximizers of the target and auxiliary functions:

$$f_i(x) \le f_i(x_{*_t}) + c_i \quad \forall x \in D, i \neq t \tag{6}$$

where $c_i \triangleq \mathbb{E}_{p(x_{*_i}|y_X)}[f_i(x_{*_i})] - \mathbb{E}_{p(x_{*_t}|y_X)}[f_i(x_{*_t})]$ measures the gap between the expected maximum of f_i and the expected output of f_i evaluated at x_{*_t}. Consequently, the following simplified constraints instead of (6) will be used to approximate $p(f_i(x)|y_X, x_{*_t})$:

- $C1$. $f_i(x) \le f_i(x_{*_t}) + \delta_i c_i$ for a given $\langle x,i\rangle \in D^+$ where δ_i equals to 0 if $i = t$, and 1 otherwise.
- $C2$. $f_j(x_{*_t}) + \delta_j c_j \ge y_{\max_j} + \epsilon_j$ for $j = 1,\ldots,M$ where $y_{\max_j} \triangleq \max_{\langle x,i\rangle\in X_j} y_i(x)$ is the largest among the noisy outputs observed by evaluating f_j at X_j.

Similar as in [6], we can use an indicator function and the cdf of a standard Gaussian distribution to represent the probability of C1 and C2, respectively. Let $f_j^* \triangleq f_j(x_{*_t})$ for $j = 1, ..., M$,

$$p(f_i(x)|y_X, x_{*_t}) \approx p(f_i(x)|y_X, C1, C2) \propto \int p(f_i(x)|y_X, f_i^*)\; p(f_i^*|y_X, C2)\; \mathbb{I}(f_i(x) \le f_i^* + \delta_i c_i)\;\mathrm{d}f_i^* \tag{7}$$

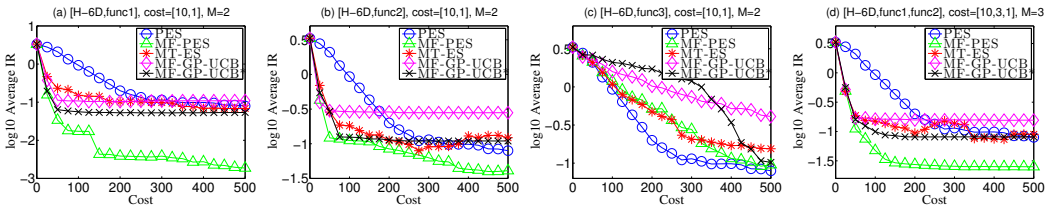

Figure 1: Graphs of $\log_{10}$(averaged IR) vs. cost incurred by tested algorithms for Hartmann-6D (H-6D) function and its auxiliary functions func1, func2, and func3 with the respective fidelities ρ^1, ρ^2, and ρ^3 computed using (2) where $\rho^1 > \rho^2 > \rho^3$.

where $p(f_i(x)|y_X, f_i^*)$ can be computed using (3), $\mathbb{I}(f_i(x) \leq f_i^* + \delta_i c_i)$ can be approximated as a Gaussian distribution using EP, and

$$p(f_i^*|y_X, C2) \propto \int p(f_1^*, ..., f_M^*|y_X) \prod_{j=1}^{M} \Phi_{\text{cdf}}\left((f_j^* + c_j - y_{\max_j})/\sigma_{n_j}\right) \mathrm{d}f^*_{\{1,...,M\}\setminus i} \,. \quad (8)$$

Then, $p(f_i^*|y_X, C2)$ can be approximated as a Gaussian distribution using EP by approximating all the non-Gaussian factors (i.e., Φ_{cdf}) in (8) to be a Gaussian. Consequently, the posterior distribution $p(f_i(x)|y_X, x_{*_t})$ can be approximated as $\mathcal{N}(f_i(x)|\mu_{f_i}, v_{f_i})$ due to (7) and (8). Note that (8) is independent of x such that it can be computed once and reused in (7).

Interestingly, by sampling the target and auxiliary maximizers x_{*_t} and x_{*_j} using the method proposed in Section 3.1, the value of c_j in (8) can be approximated in practice by Monte Carlo sampling:

$$c_j = \mathbb{E}_{p(x_{*_j}|y_X)}[f_j(x_{*_j})] - \mathbb{E}_{p(x_{*_t}|y_X)}[f_j(x_{*_t})] \approx S^{-1} \sum_{s=1}^{S} \left(f_j^{(s)}(x_{*_j}^{(s)}) - f_j^{(s)}(x_{*_t}^{(s)})\right).$$

Using the results in Section 3.1 and (7), it follows that MF-PES (5) can be approximated by

$$\alpha(y_X, \langle x, i\rangle) \approx \tfrac{1}{2}\log\left(\sigma^2_{\langle x,i\rangle|X} + \sigma^2_{n_i}\right) - \tfrac{1}{2S}\sum_{s=1}^{S}\log\left(v_{f_i}^{(s)} + \sigma^2_{n_i}\right).$$

When the costs of evaluating target vs. auxiliary functions differ, we use the following cost-sensitive MF-PES instead: $\alpha_{\text{cost}}(y_X, \langle x, i\rangle) \triangleq \alpha(y_X, \langle x, i\rangle)/\text{cost}(i)$ which can be interpreted as the information gain of the target maximizer per cost.

4 Experiments and Discussion

This section empirically evaluates the multi-fidelity BO performance of our MF-PES algorithm against that of (a) PES [6], (b) MT-ES [16], (c) MF-GP-UCB with all parameters trading off between exploitation vs. exploration set according to [8], and (d) MF-GP-UCB*: MF-GP-UCB with carefully fine-tuned parameters. For a fair comparison, CMOGP is used to model multiple functions in all tested algorithms since it has been empirically demonstrated by [1] to outperform the other MOGPs. The performance of the tested algorithms are evaluated using *immediate regret* (IR) $|f_t(x_{t_*}) - f_t(\tilde{x}_{t_*})|$ where $\tilde{x}_{t_*} \triangleq \arg\max_{x\in D} \mu_{\{\langle x,t\rangle\}|X}$ is their recommended target maximizer.

Hartmann-6D function. The original Hartmann-6D function is used as target function and $M \triangleq 2$ or 3. Similar to that in [8], three auxiliary functions of varying degrees of fidelity are constructed by tweaking the Hartmann-6D function. The experiments are run with 10 different initializations.

Figs. 1 show results of all tested algorithms with a cost budget of 500. It can be observed from Figs. 1a-b and 1d that MF-PES can achieve a much lower averaged IR with considerably less cost than PES, which implies that the BO performance can be improved by auxiliary function(s) of sufficiently high fidelity and low evaluation cost. The Hartmann-6D function are difficult to optimize due to its multimodal nature (6 local and 1 global maxima) and the large input domain which cause MT-ES and MF-GP-UCB to be trapped easily in some local maximum and hence perform not as well. We have dedicated time to carefully fine-tune the parameters of MF-GP-UCB* such that it explores more to perform better than MF-GP-UCB but is still outperformed by MF-PES. In contrast, MF-PES is rarely trapped in a local maximum and performs significantly better than all the other tested algorithms by naturally exploring more over these multimodal functions. Finally, Fig. 1c shows that when the fidelity of the auxiliary function is very low ($\rho^3 = 0.0037$), MF-PES can achieve a comparable performance to PES, hence demonstrating its robustness to a low-fidelity auxiliary function.

Hyperparameters tunning. The tested algorithms are also used to automatically tune the hyperparameters of logistic regression and convolutional neural network (CNN) models in image classification

tasks. For both models, MF-PES converges much faster than other tested algorithms. Also, MF-PES improves the performance of CNN compared to the baseline achieved using the default hyperparameters in the online code, which shows that MF-PES is promising in finding more competitive hyperparameters of complex ML models.

Acknowledgments. This research is supported by Singapore Ministry of Education Academic Research Fund Tier 2, MOE2016-T2-2-156, and the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centre in Singapore Funding Initiative.

References

[1] M. A. Álvarez and N. D. Lawrence. Computationally efficient convolved multiple output Gaussian processes. *JMLR*, 12:1459–1500, 2011.

[2] E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In *Proc. NIPS*, pages 153–160, 2007.

[3] N. A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., 2nd edition, 1993.

[4] M. Cutler, T. J. Walsh, and J. P. How. Real-world reinforcement learning via multi-fidelity simulators. *IEEE Transactions on Robotics*, 31(3):655–671, 2015.

[5] P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *JMLR*, 13:1809–1837, 2012.

[6] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Proc. NIPS*, pages 918–926, 2014.

[7] D. Huang, T. T. Allen, W. I. Notz, and R. A. Miller. Sequential kriging optimization using multiple-fidelity evaluations. *Struct. Multidisc. Optim.*, 32(5):369–382, 2006.

[8] K. Kandasamy, G. Dasarathy, J. B. Oliva, J. Schneider, and B. Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Proc. NIPS*, pages 992–1000, 2016.

[9] K. Kandasamy, G. Dasarathy, J. Schneider, and B. Póczos. Multi-fidelity Bayesian optimisation with continuous approximations. In *Proc. ICML*, pages 1799 – 1808, 2017.

[10] D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans. Automatic gait optimization with Gaussian process regression. In *Proc. IJCAI*, pages 944–949, 2007.

[11] M. Poloczek, J. Wang, and P. I. Frazier. Multi-information source optimization. *arXiv preprint arXiv:1603.00389*, 2016.

[12] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Proc. NIPS*, pages 1177–1184, 2007.

[13] B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[14] G. Skolidis. *Transfer Learning with Gaussian Processes*. PhD thesis, University of Edinburgh, 2012.

[15] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Proc. NIPS*, pages 2951–2959, 2012.

[16] K. Swersky, J. Snoek, and R. P. Adams. Multi-task Bayesian optimization. In *Proc. NIPS*, pages 2004–2012, 2013.

[17] Y. W. Teh and M. Seeger. Semiparametric latent factor models. In *Proc. AISTATS*, pages 333–340, 2005.

[18] M. Tesch, J. Schneider, and H. Choset. Expensive function optimization with stochastic binary outcomes. In *Proc. ICML*, pages 1283–1291, 2013.