Multi-Target Optimisation via Bayesian Optimisation and Linear Programming

Alistair Shilton **Santu Rana** **Sunil Gupta** **Svetha Venkatesh**
Deakin University, Geelong, Australia,
Centre for Pattern Recognition and Data Analytics
{alistair.shilton,santu.rana,sunil.gupta,svetha.venkatesh}@deakin.edu.au

Abstract

In Bayesian Multi-Objective optimisation, expected hypervolume improvement is often used to measure the goodness of candidate solutions. However when there are many objectives the calculation of expected hypervolume improvement can become computationally prohibitive. An alternative approach measures the goodness of a candidate based on the distance of that candidate from the Pareto front in objective space. In this paper we present a novel distance-based Bayesian Many-Objective optimisation algorithm. We demonstrate the efficacy of our algorithm on three problems, namely the DTLZ2 benchmark problem, a hyper-parameter selection problem, and high-temperature creep-resistant alloy design.

1 INTRODUCTION

Bayesian optimisation (Brochu et al., 2010) is a method for maximising black-box functions that are expensive to evaluate either in terms of time or cost. Bayesian optimisation works by modelling the objective function (typically) using a Gaussian process (GP) (Rasmussen and Williams, 2006). At each iteration a point (called a recommendation) is selected to maximise an acquisition function, where the acquisition function is a measure of the *goodness* of a proposed point. Unlike the black-box function, the acquisition function is cheap to evaluate and therefore amenable to global optimisation.

In the context of multi-objective optimisation, Bayesian optimisation is typically applied using an acquisition function based on expected hypervolume improvement (EHI) (Ponweiser et al., 2008; Emmerich and Klinkenberg, 2008; Shir et al., 2007; Zaefferer et al., 2013; Shimoyama et al., 2013), which is the expected change in the hypervolume dominated by the estimated Pareto front (the set of dominant evaluations of prior recommendations in objective space - see figure 1). However this can be expensive to evaluate, particularly in the many-objective case where the number of objectives is large (Wagner et al., 2010; Zaefferer et al., 2013). While optimised algorithms have been developed for calculating EHI for up to 3 dimensions (Hupkens et al., 2015) the general (many-objective (Ishibuchi et al., 2008)) case remains computationally challenging.

An alternative approach is to use a distance-based acquisition function (or score function) (Miranda and Von Zuben, 2015; Yun et al., 2004). Distance-based acquisition functions seek to maximise the signed distance of a point from the estimated Pareto front, as shown in figure 1. Unlike EHI this acquisition function is cheap to evaluate, making its global optimisation (and hence Bayesian optimisation) practical in the many-objective case. While the underlying concept is old (e.g. (Yun et al., 2004)) it has only recently been formalised in a rigorous manner (Miranda and Von Zuben, 2015) in the form of conditions that must be met by a distance (score) function measuring the signed distance *in advance of (dominating)* the Pareto front; whereas (Yun et al., 2004) for example defines the signed distance from the estimated feasible region in any direction, dominating or otherwise. However, while (Miranda and Von Zuben, 2015) defines the conditions that must be met by such a score function, the method implemented therein - namely a GP model with a probability distribution over the gradient - only approximately meets these requirements.

In the present paper we introduce an alternative model based on a modified 1-norm support vector machine (SVM) that is able to *exactly* satisfy the conditions laid down in (Miranda and Von Zuben, 2015). To be precise, we use a restricted 1-norm, 1-class SVM to define a signed distance function which is strictly positive for points that dominate the estimated Pareto front, strictly negative for points dominated by the estimated Pareto

front, and for which signed distance and the dominance relation are congruent.[1]. This distance function forms the basis for an acquisition function that is computationally cheap to evaluate and scales well with the number of objectives, thereby making feasible Bayesian many-objective optimisation.

To test our proposed algorithm we have applied it to one benchmark problem and two practical problems. The benchmark problem used is taken from the DTLZ suite of benchmarks (Deb et al., 2005). For practical problems we have chosen a hyper-parameter selection problem and an experimental problem involving high-temperature, creep resistant alloy design.

The first practical problem considered is hyperparameter selection for a multi-class classifier where the relative weights (importance) of the various classes is unknown. While the default assumption often made for such problems is that all classes should have equal weight (or alternatively that their weight should be proportional to their class density) this will not be valid in general. Instead the accuracy of the classifier with respect to each class of training data forms an independent objective, and the problem of hyper-parameter selection in the absence of additional information regarding relative weight is one of multi-objective optimisation.

The second practical problem considered is the design of high-temperature, creep-resistant alloys. High-temperature creep resistant Ni-superalloy is used for making boilers of super-critical thermal power plants. In a joint project with metallurgists we were asked to optimize the current alloy recipe to obtain superior creep resistance than the industry standard. This involves using phase simulation (via ThermoCalc) to design an alloy with maximum good phases (those that improved creep-resistance) and minimum bad phases (those that made the alloy less creep-resistant) over a range of temperatures. The total number of objectives for this experiment is 12, each corresponding to a particular phase and temperature, which leads to recommendation times of up to 1 day/recommendation if EHI is used. We demonstrate that our approach is able to provide a range of potential alloys, each Pareto-optimal in terms of phase contents, for further assessment by the metallurgist.

We note that there exists an abundance of such many-objective optimisation problems in physical systems - for example advanced fibre production (Li et al., 2017). By making many-objective Bayesian optimisation feasible we envisage that such problems will be able to be formulated and solved.

[1]That is, if $\mathbf{y}$ dominates $\mathbf{y}'$ then the distance of $\mathbf{y}$ from the estimated Pareto front, as measured by the score function, is greater than the distance of $\mathbf{y}'$ from the estimated Pareto front.

2 NOTATION

Column vectors are written $\mathbf{a}, \mathbf{b}, \ldots$ with elements $a_i, b_i, \ldots$. Matrices are written $\mathbf{A}, \mathbf{B}, \ldots$, with elements $A_{i,j}, B_{i,j}, \ldots$. If $\mathbf{f} : \mathbb{X} \subset \mathbb{R}^m \to \mathbb{R}^n$ is a map from design to objective space then $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{X}$ we say $\mathbf{x}$ dominates $\mathbf{x}'$, written $\mathbf{x} \succeq_{\mathbf{f}} \mathbf{x}'$, if $f_i(\mathbf{x}) \geq f_i(\mathbf{x}')\ \forall i$; and $\mathbf{x}$ strongly dominates $\mathbf{x}'$, written $\mathbf{x} \succ_{\mathbf{f}} \mathbf{x}'$, if $\mathbf{x} \succeq_{\mathbf{f}} \mathbf{x}' \wedge \mathbf{f}(\mathbf{x}) \neq \mathbf{f}(\mathbf{x}')$. Analogously, $\forall \mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$ we say $\mathbf{y}$ dominates $\mathbf{y}'$, written $\mathbf{y} \succeq \mathbf{y}'$, if $y_i \geq y_i'\ \forall i$; and $\mathbf{y}$ strongly dominates $\mathbf{y}'$, written $\mathbf{y} \succ \mathbf{y}'$, if $\mathbf{y} \succeq \mathbf{y}' \wedge \mathbf{y} \neq \mathbf{y}'$.

3 BACKGROUND

Multi-objective optimisation (Deb, 2001; Coello et al., 2002; Miettinen, 1999) extends standard single-objective optimisation to the case where there are multiple, potentially conflicting objectives. The multi-objective optimisation problem is:

$$\underset{\mathbf{x} \in \mathbb{X}}{\operatorname{argmax}}\, \mathbf{f}(\mathbf{x}) \tag{1}$$

where $\mathbf{f} : \mathbb{X} \to \mathbb{R}^n$ maps from design space to objective space; $\mathbb{X} \subset [0, r]^m \subset \mathbb{R}^m$ is the feasible region; and argmax is defined in the Pareto sense described below. This is known as a *many-objective optimisation problem* (Ishibuchi et al., 2008) if the number of objectives is sufficiently large to cause difficulties with standard multi-objective optimisation algorithms (as shown in our experiments, as few as 6 objectives can cause difficulties).

Our aim is to find a representation of the Pareto set:

$$\mathbb{X}^\star = \{\mathbf{x}^\star \in \mathbb{X} | \nexists \mathbf{x} \in \mathbb{X} : \mathbf{x} \succ_{\mathbf{f}} \mathbf{x}^\star\}$$

where $\succ_{\mathbf{f}}$ is the dominance relation as defined in section 2 ($\mathbf{x}$ strongly dominates $\mathbf{x}'$, written $\mathbf{x} \succ_{\mathbf{f}} \mathbf{x}'$, if $f_i(\mathbf{x}) \geq f_i(\mathbf{x}')\ \forall i$ and $\mathbf{f}(\mathbf{x}) \neq \mathbf{f}(\mathbf{x}')$). This is the set of all Pareto-optimal $\mathbf{x} \in \mathbb{X}$, where a vector is Pareto-optimal if it cannot be changed without causing a decrease in at least one objective $f_i : \mathbb{X} \to \mathbb{R}$. The Pareto front is the image of the Pareto set in objective space:

$$\mathbb{Y}^\star = \{\mathbf{y}^\star \in \mathbb{Y} | \nexists \mathbf{y} \in \mathbb{Y} : \mathbf{y} \succ \mathbf{y}^\star\}$$

where $\mathbb{Y} = \mathbf{f}(\mathbb{X})$ and $\succ$ is the dominance relation in objective space ($\mathbf{y}$ strongly dominates $\mathbf{y}'$, written $\mathbf{y} \succ \mathbf{y}'$, if $y_i \geq y_i'\ \forall i$ and $\mathbf{y} \neq \mathbf{y}'$). The solution to (1) is a finite set of Pareto-optimal solutions $\mathcal{X}^\star \subset \mathbb{X}^\star$.

3.1 GAUSSIAN PROCESSES

We assume the many-objective case where, for all i, $f_i(\mathbf{x}) \sim \mathrm{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ is a sample from a zero-mean Gaussian process (Rasmussen and Williams, 2006;

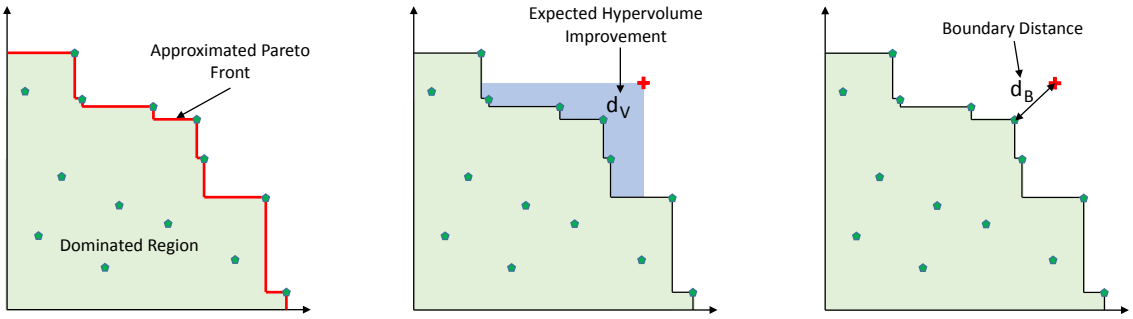

Figure 1: Pareto front (left), expected hypervolume improvement (EHI, middle) and boundary distance (right) for a simple two-objective problem.

MacKay, 1998) (we assume the objectives are non-correlated) that is costly to evaluate. Evaluations of $\mathbf{f}$ are presumed noisy, so $y_i = f_i(\mathbf{x}) + \epsilon$, where $\epsilon \in \mathcal{N}(0, \sigma^2)$. Given observations $\mathcal{D}_t = \{(\mathbf{x}, \mathbf{y}) | \mathbf{y} = \mathbf{f}(\mathbf{x}) + \epsilon\}$ we have $\mathbf{f}(\mathbf{x})|\mathcal{D}_t \sim \mathcal{N}(\boldsymbol{\mu}_t(\mathbf{x}), \mathbf{I}\sigma_t(\mathbf{x}))$, where:

$$\begin{aligned} \boldsymbol{\mu}_t(\mathbf{x}) &= \mathbf{Y}_t^{\mathrm{T}} \left(\mathbf{K}_t + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{k}_t(\mathbf{x}) \\ \sigma_t^2(\mathbf{x}) &= k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t^{\mathrm{T}}(\mathbf{x}) \left(\mathbf{K}_t + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{k}_t(\mathbf{x}) \end{aligned} \tag{2}$$

where $\mathbf{Y}_t = [\mathbf{y}^{\mathrm{T}}]_{(-,\mathbf{y}) \in \mathcal{D}_t}$, $\mathbf{k}_t(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}')]_{(\mathbf{x}',-) \in \mathcal{D}_t}$ and $\mathbf{K}_t = [k(\mathbf{x}, \mathbf{x}')]_{(\mathbf{x},-),(\mathbf{x}',-) \in \mathcal{D}_t}$.[2] Given $\mathcal{D}_t$ the estimated Pareto set $\mathcal{X}_t^\star$ and Pareto front $\mathcal{Y}_t^\star$ at iteration t are the dominant subsets of $\mathcal{D}_t$:

$$\begin{aligned} \text{where:} \quad \mathcal{X}_t^\star &= \{\mathbf{x}^\star \in \mathcal{X}_t | \nexists \mathbf{x} \in \mathcal{X}_t : \mathbf{x} \succ_{\mathbf{f}} \mathbf{x}^\star\} \\ \mathcal{Y}_t^\star &= \{\mathbf{y}^\star \in \mathcal{Y}_t | \nexists \mathbf{y} \in \mathcal{Y}_t : \mathbf{y} \succ \mathbf{y}^\star\} \\ \mathcal{X}_t &= \{\mathbf{x} \in \mathbb{X} | (\mathbf{x}, -) \in \mathcal{D}_t\} \\ \mathcal{Y}_t &= \{\mathbf{y} \in \mathbb{Y} | (-, \mathbf{y}) \in \mathcal{D}_t\} \end{aligned} \tag{3}$$

3.2 BAYESIAN OPTIMISATION

Bayesian optimisation (Brochu et al., 2010) is an optimisation method designed for problems where the function being optimised is expensive to evaluate in terms of time or monetary cost. A typical Bayesian optimisation algorithm is presented in algorithm 1. For each iteration t we maximise a (cheap) acquisition function $a_t : \mathbb{X} \to \mathbb{R}$ based on μ_{t-1} and σ_{t-1}, and the resulting recommendation is evaluated to obtain $y_t = f(\mathbf{x}_t) + \epsilon$. GP models are updated, and the algorithm continues. Standard acquisition functions include expected improvement (EI) (Mockus et al., 1978), probability of improvement (PI) (Kushner, 1964), and GP upper confidence bound (GP-UCB) (Jones et al., 1998; Srinivas et al., 2012; Brochu et al., 2010).

[2] We write $(\mathbf{x}, -) \in \mathcal{D}_t$ if $\exists \mathbf{y} \in \mathbb{Y} : (\mathbf{x}, \mathbf{y}) \in \mathcal{D}_t$; and likewise $(-, \mathbf{y}) \in \mathcal{D}_t$ if $\exists \mathbf{x} \in \mathbb{X} : (\mathbf{x}, \mathbf{y}) \in \mathcal{D}_t$.

Algorithm 1 Generic Bayesian Optimisation

input $\mathcal{D}_0 := \{(\mathbf{x}_i, y_i) | y_i = f(\mathbf{x}_i) + \epsilon, i = 1, 2, \ldots\}$.
  for $t = 1, 2, \ldots, T$ **do**
    Select test point $\mathbf{x}_t = \mathrm{argmax}_{\mathbf{x}}\, a_t(\mathbf{x})$.
    Perform Experiment $y_t = f(\mathbf{x}_t) + \epsilon$.
    Update $\mathcal{D}_t := \mathcal{D}_{t-1} \cup \{(\mathbf{x}_t, y_t)\}$.
  end for

3.3 MULTI-OBJECTIVE BAYESIAN OPTIMISATION

Adding an observation $\mathbf{y}_t$ to $\mathcal{Y}_{t-1}$ will either cause no change to the estimated Pareto front $\mathcal{Y}_{t-1}^\star$ (if $\exists \mathbf{y} \in \mathcal{Y}_{t-1}^\star : \mathbf{y} \succ \mathbf{y}_t$) or push it closer to the actual Pareto front $\mathbb{Y}^\star$ (if $\nexists \mathbf{y} \in \mathcal{Y}_{t-1}^\star : \mathbf{y} \succ \mathbf{y}_t$). The acquisition function $a_t(\mathbf{x})$ is designed to measure this expected change. Two popular measures used, as shown in figure 1, are:

- Expected hypervolume improvement (EHI):

$$a_t(\mathbf{x}) = \mathbb{E}\left[\mathrm{S}\left(\mathcal{Y}_{t-1}^\star \cup \{\mathbf{f}(\mathbf{x})\}\right) - \mathrm{S}\left(\mathcal{Y}_{t-1}^\star\right)\right]$$

  (Shir et al., 2007; Zaefferer et al., 2013; Shimoyama et al., 2013). This is the expected change in hypervolume dominated by $\mathcal{Y}_{t-1}^\star$, where $\mathrm{S}(\mathcal{Y})$ is the hypervolume dominated by $\mathcal{Y}$ (Zitzler, 1999; Huband et al., 2003; Purshouse, 2003; Laumanns et al., 2000; Fleischer, 2000).

- Boundary distance:

$$a_t(\mathbf{x}) = \mathbb{E}\left(d\left(\mathcal{Y}_{t-1}^\star, \mathbf{f}(\mathbf{x})\right)\right)$$

  (Yun et al., 2004; Keane, 2006). This is the expected (signed) distance between the the estimated Pareto front and $\mathbf{f}(\mathbf{x})$.

It has been noted that calculating the EHI is non-trivial (Wagner et al., 2010; Zaefferer et al., 2013) and, while heavily optimised algorithms are available for up to 3

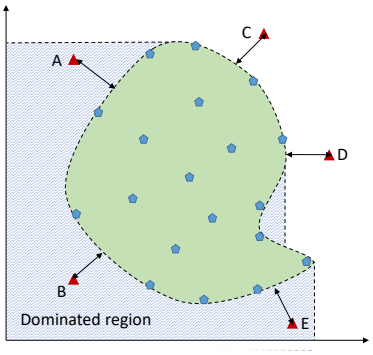

Figure 2: Distance maximisation of 1-class SVM (Yun et al., 2004). The set of observations (blue) are used to train a 1-class SVM, giving the boundary of the green region. Points A-E all give a positive boundary distance, but only points C and D dominate the observations as the boundary does not satisfy consistancy requirements.

objectives (Hupkens et al., 2015), the computational cost in the many-objective case remains prohibitive, making EHI unsuitable in a many-objectives context. Similarly, calculating the precise distance to the Pareto front is often computationally intractable, particularly in the many-objective case. Hence when calculating the boundary distance the estimated Pareto front is usually approximated (smoothed) using a score function such as the 1-class SVM (Yun et al., 2004), and the signed distance to this front used. We note that the Pareto front approximation used by (Yun et al., 2004) is in fact a hypersurface surrounding the set of observations and may contain pairs of points where one dominates the other (i.e. it does not satisfy the consistency requirements discussed in section 4). Thus, as shown in figure 2, maximising this measure will not necessarily maximise change to the Pareto front as points may be selected that are not in advance of (dominating) the set of observations.

4 PROPOSED METHOD

In the present paper we will be using an acquisition function based on GP-UCB (Srinivas et al., 2012) that we call AD-GP-UCB (approximated distance GP-UCB):

$$a_t(\mathbf{x}) = g_t\left(\boldsymbol{\mu}_{t-1}(\mathbf{x})\right) + \sqrt{\beta_t}\eta_t\left(\boldsymbol{\mu}_{t-1}(\mathbf{x}), \sigma_{t-1}(\mathbf{x})\right) \quad (4)$$

In this expression $g_t(\boldsymbol{\mu}_{t-1}(\mathbf{x}))$ is the approximate mean distance of $\mathbf{f}(\mathbf{x})$ from the estimated Pareto set and $\eta_t(\boldsymbol{\mu}_{t-1}(\mathbf{x}), \sigma_{t-1}(\mathbf{x}))$ the approximate variance. The constants β_t control the trade-off between exploitation (selecting recommendations with high predicted distance from the estimated Pareto set) and exploration (exploring unexplored regions of the feasible set $\mathbb{X}$) as per the GP-UCB method (Srinivas et al., 2012):

$$\beta_t = \begin{cases} 2\log\left(\frac{\pi^2 t^2}{6\delta}|\mathbb{X}|\right) & \text{if } |\mathbb{X}| < \infty \\ 2\log\left(\frac{2\pi^2 t^2}{3\delta}\left(t^2 mbr\left(\log\left(\frac{2ma}{\delta}\right)\right)^{\frac{1}{2}}\right)^{2m}\right) & \text{otherwise} \end{cases} \quad (5)$$

where $0 < \delta \ll 1$ and in the infinite case we assume $\mathbf{f}$ satisfies $\Pr\{\sup_{\mathbf{x}\in\mathbb{X}} |\partial f_i/\partial x_j| > L\} \leq ae^{-(L/b)^2}\ \forall i, L$. We have chosen GP-UCB here as it is explicitly designed to balance exploration and exploitation; and because, in the single objective case, there exist convergence bounds to show that, with probability $1-\delta$, the optimisation procedure is guaranteed to converge (as measured by cumulative risk) sub-linearly as $T \to \infty$. While our method is not a "true" GP-UCB method (g_t and η_t are only approximations of the mean and variance of the predicted distance from the estimated Pareto front; and moreover the function approximated by g_t changes over time) our experimental results demonstrate its efficacy.

4.1 APPROXIMATING THE MEAN DISTANCE

The score function $g_t(\boldsymbol{\mu}_{t-1}(\mathbf{x}))$ is used to approximate the signed distance between the estimated Pareto set $\mathcal{Y}^\star_{t-1}$ and the sample evaluation $\mathbf{f}(\mathbf{x})$ for a given $\mathbf{x} \in \mathbb{X}$. We use the GP posterior mean $\boldsymbol{\mu}_{t-1}(\mathbf{x})$ to estimate the mean of $\mathbf{f}(\mathbf{x})$ and g_t to approximate the distance of this from the Pareto front $\mathcal{Y}^\star_{t-1}$. Motivated by the "standard form" of the trained SVM in dual form, the score function g_t is defined as:

$$g_t(\mathbf{y}) = 1 - 2\sum_{i=1}^{N_t} \alpha_i^t L(\mathbf{y}_i, \mathbf{y}) \quad (6)$$

where the indices i applied to all $\mathbf{y}_i \in \mathcal{Y}_{t-1}$ correspond to the indices i on α_i^t, $N_t = |\mathcal{Y}_{t-1}|$, $\boldsymbol{\alpha}^t \geq \mathbf{0}$, and L is defined to ensure that g_t satisfies consistency conditions:

1. Observational Consistency:

$$g_t(\mathbf{y}) \leq 0\ \forall \mathbf{y} \in \mathcal{Y}^\star_{t-1}$$

2. Dominance Consistency:

$$g_t(\mathbf{y}) > g_t(\mathbf{y}')\ \forall \mathbf{y}, \mathbf{y}' \in \mathbb{Y} : \mathbf{y} \succ \mathbf{y}'$$

Observational consistency is required to ensure that the reported distance is never positive for existing observations that are by definition dominated by the current estimated Pareto set $\mathcal{Y}^\star_{t-1}$. Dominance consistency ensures that, $\forall \mathbf{y}, \mathbf{y}' \in \mathbb{Y}$, the dominant vector will receive the higher "score". Thus g_t is a score function in the sense of (Miranda and Von Zuben, 2015) and may be said to define an estimated Pareto set:

$$\mathbb{Y}^{g_t\star} = \{\mathbf{y} \in \mathbb{R}^m | \, g(\mathbf{y}) = 0\} \quad (7)$$

that dominates all points in $\mathcal{Y}_{t-1}$ (that is, $\forall \mathbf{y} \in \mathcal{Y}_{t-1}\ \exists \mathbf{y}' \in \mathbb{Y}^{g_t\star} : \mathbf{y}' \succeq \mathbf{y}$). The distance reported by g_t is the signed distance from $\mathbb{Y}^{g_t\star}$ as measured by some metric. Motivated by this we let $L(\mathbf{y}, \mathbf{y}') = \kappa(\min_q(y_q - y'_q))$, where:

$$\begin{array}{ll} \kappa(0) = \frac{1}{2} & \text{(centred)} \\ \kappa(y+\delta) > \kappa(y)\ \forall y \in \mathbb{R}, \delta \in \mathbb{R}^+ & \text{(increasing)} \end{array} \quad (8)$$

Many standard neural activation functions are suitable choices (e.g. the logistic function $\kappa(y) = 1/(1+\exp(-\upsilon y))$). It is straightforward to see that g_t defined by (6) satisfies dominance consistency if $\boldsymbol{\alpha}^t \neq \mathbf{0}$. To satisfy observational consistency $\boldsymbol{\alpha}^t$ is selected to solve the linear programming problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}^t\|_1 = \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha}^t & \\ \text{such that:} & \sum\limits_{i=1}^{N_t} \alpha_i^t L(\mathbf{y}_i, \mathbf{y}_j) \geq \frac{1}{2}\ \forall 1 \leq j \leq N_t \\ & \boldsymbol{\alpha}^t \geq \mathbf{0} \end{array} \quad (9)$$

which will be referred to this as the score-function optimisation problem. It may be noted that the score-function optimisation problem, and the form of the score function g_t, are closely related to the 1-norm SVM (Bradley and Mangasarian, 1998; Zhu et al., 2004), which is a variant of the standard SVM that retains the standard (dual) form of the trained machine but minimises $\|\boldsymbol{\alpha}\|_1$ rather than $\|\boldsymbol{\alpha}\|_{\mathcal{H}_L}$. This form has two distinct advantages: the kernel[3] L may be any function (not just positive definite) and the solution tends to be more sparse than the standard form. Our approach also borrows from the 1-class SVM (Schölkopf et al., 1999), but rather than using a *bias-forcing* term to achieve margin minimalisation (rather than maximising the margin of separation, the 1-class SVM seeks to minimise the margin) we instead use a fixed bias ($b = -1$) and restrict L using (8) so that the margin minimalisation occurs as a direct result from minimising the regularisation term $\|\boldsymbol{\alpha}\|_1$.

4.2 APPROXIMATING THE DISTANCE VARIANCE

The function $\eta_t(\boldsymbol{\mu}_{t-1}(\mathbf{x}), \sigma_{t-1}(\mathbf{x}))$ in the acquisition function (4) approximates the variance in the estimate $g_t(\boldsymbol{\mu}_{t-1}(\mathbf{x}))$ of the distance between $\mathbf{f}(\mathbf{x})$ and the estimated Pareto front $\mathcal{Y}^\star_{t-1}$. We approximate this using a simple first-order Taylor approximation of the second moment about $\boldsymbol{\mu}_{t-1}(\mathbf{x})$ - that is:

$$\eta_t\big(\boldsymbol{\mu}_{t-1}(\mathbf{x}), \sigma_{t-1}(\mathbf{x})\big) = \big\|\nabla_{\mathbf{x}} g_t\big(\boldsymbol{\mu}_{t-1}(\mathbf{x})\big)\big\| \sigma_{t-1}(\mathbf{x})$$

where, defining $q_i = \operatorname{argmin}_q(y_{i,q} - \mu_{t-1,q}(\mathbf{x}))\ \forall i$:

$$\begin{array}{l} \frac{\partial}{\partial x_l} g_t\left(\boldsymbol{\mu}_{t-1}(\mathbf{x})\right) = \ldots \\ \quad -2\sum_i \alpha_i^t \kappa'\left(y_{i,q_i} - \mu_{t-1,k_i}(\mathbf{x})\right) \frac{\partial}{\partial x_l}\mu_{t-1,q_i}(\mathbf{x}) \end{array}$$

where $\kappa'(y) = \partial\kappa(y)/\partial y$ and:

$$\frac{\partial}{\partial x_l}\boldsymbol{\mu}_t(\mathbf{x}) = \mathbf{Y}_t^{\mathrm{T}}\left(\mathbf{K}_t + \sigma^2\mathbf{I}\right)^{-1} \frac{\partial}{\partial x_l}\mathbf{k}_t(\mathbf{x})$$

We note that the accuracy of this approximation degrades as the non-linearity of $g_t(\boldsymbol{\mu}_{t-1}(\bullet))$ increases. This is a necessary trade-off as calculating the actual variance is not feasible. Assuming an squared-exponential kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/l)$ for the GP model and a sigmoid function for the score function kernel $\kappa(y) = 1/(1+\exp(-\upsilon y))$ we see that the variance approximation is best when the length scale l of k is large and the constant υ of κ is small.

5 THEORETICAL ANALYSIS

It is useful at this point to analyse the theoretical properties of our proposed algorithm. We have already noted that the score function g_t satisfies both observational and dominance consistency and so provides a sensible approximation of distance from the estimated Pareto front. Applying SVM techniques we find the following properties (all proofs presented in the supplementary material):

Theorem 1 (Non-triviality) *Let $\boldsymbol{\alpha}^t$ be the solution to the score-function optimisation problem (9). Then $\boldsymbol{\alpha}^t \neq \mathbf{0}$.*

Theorem 2 (Margin Minimisation) *Let $\boldsymbol{\alpha}^t$ be the solution to the score-function optimisation problem (9). Let $\mathbb{Y}^{g_t\star}$ be the estimated Pareto front defined by g_t. The minimum distance between $\mathcal{Y}_{t-1}$ and the estimated Pareto front $\mathbb{Y}^{g_t\star}$ is zero:*

$$\min_{\mathbf{y}\in\mathcal{Y}_{t-1}, \mathbf{y}'\in\mathbb{Y}^{g_t\star}} \|\mathbf{y} - \mathbf{y}'\| = 0$$

Theorem 3 (Sparsity) *Let $\boldsymbol{\alpha}^t$ be the solution to the score-function optimisation problem (9). Then $\alpha_i^t = 0\ \forall i : \mathbf{y}_i \notin \mathcal{Y}^\star_{t-1}$ (ie. points not in the estimated Pareto front cannot be support vectors).*

Theorem 4 (Heaviside Limit) *Let $\kappa = \kappa_\perp$, where*

$$\kappa_\perp(y) = \lim_{\upsilon\to\infty} \frac{1}{1+\exp(-\upsilon y)} = \frac{1}{2}\left(1 + \operatorname{sgn}(y)\right),$$

and $\nexists i \neq j : \mathbf{y}_i = \mathbf{y}_j$. Then $\alpha_i^t = 1\ \forall i : \mathbf{y}_i \in \mathcal{Y}^\star_{t-1}$, $\alpha_i^t = 0$ otherwise (ie. in the limiting case the support vectors are precisely the estimated Pareto set).

[3] In general we have tried to avoid using the word kernel to refer to L to avoid potential confusion with the covariance function (kernel) k used for Gaussian Processes.

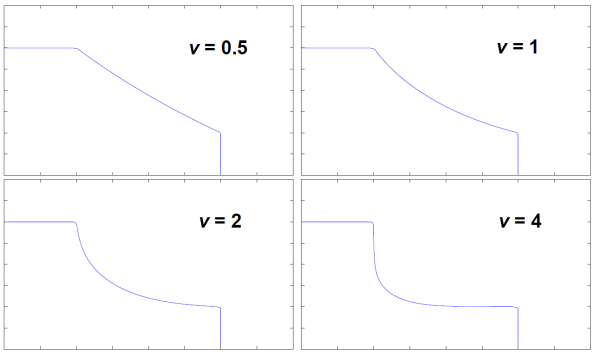

Figure 3: Estimated front $\mathbb{Y}^{g_t\star}$ (objective space) given $\mathbf{y}_1 = [\ -1; 1\ ]$, $\mathbf{y}_2 = [\ 1; -1\ ]$, where $\kappa(y) = \frac{1}{1+\exp(-\upsilon y)}$ and $\upsilon = 0.5, 1, 2, 4$, respectively.

It follows that solving (9) will define a score function, and hence an estimated Pareto front $\mathbb{Y}^{g_t\star}$, that is as tight as possible (insofar as it lies as close to $\mathcal{Y}_t$ as possible while maintaining observational consistency) and sparsely represented. As shown in figure 3 the parameter υ in the κ function acts as a smoothing parameter on the estimated Pareto front, where smaller υ will tend to favour smoother fronts while larger υ will attempt to achieve a tighter "fit" to the observations $\mathcal{Y}_t$. In the limiting case $\upsilon \to \infty$ the Pareto front becomes stepwise, as shown by theorem 4.

6 EXPERIMENTS

We consider three experiments in the this section: standard test function optimisation, hyper-parameter selection in multi-class SVM classification in the absence of relative class weighting, and high-temperature alloy design. All SVM and related code was written in C++ with linking to ThermoCalc via a Matlab interface. Where relevant EHI estimation was performed using the IRS algorithm (Hupkens et al., 2015). Global optimisation on our acquisition function was carried out using the DIRECT algorithm (Jones et al., 1993). The objective function $\mathbf{f}$ was modelled using a GP with a squared-exponential kernel. For the score function we use $\kappa(y) = \frac{1}{1+\exp(-\upsilon y)}$.

6.1 STANDARD TEST FUNCTION

In our first experiment we have evaluated the performance of AD-GP-UCB on the standard DTLZ2 multi-objective test function (Deb et al., 2005). We have run simulations for Bayesian optimisation using both EHI and AD-GP-UCB acquisition functions over a budget of $T = 200$ iterations for $n = 2, 3, \ldots, 10$ objectives.

We have evaluated performance on four criteria:

1. How close the elements of the estimated Pareto front $\mathcal{Y}_t^\star$ are to the actual Pareto front $\mathbb{Y}^\star$ (how optimal the elements of the $\mathcal{Y}_t^\star$ are):
$$d_{\mathbb{Y}^\star} = \sup_{\mathbf{y}\in\mathcal{Y}_T^\star} d(\mathbf{y}, \mathbb{Y}^\star) = \sup_{\mathbf{y}\in\mathcal{Y}_T^\star} \left(\inf_{\mathbf{y}'\in\mathbb{Y}^\star} \|\mathbf{y} - \mathbf{y}'\| \right)$$
2. The maximum distance between any point on the actual Pareto front $\mathbb{Y}^\star$ and the closest point to it in the estimated Pareto front (how well $\mathcal{Y}_t^\star$ approximates $\mathbb{Y}^\star$):
$$d_{\mathcal{Y}_T^\star} = \sup_{\mathbf{y}'\in\mathbb{Y}^\star} d(\mathcal{Y}_t^\star, \mathbf{y}') = \sup_{\mathbf{y}'\in\mathbb{Y}^\star} \left(\inf_{\mathbf{y}\in\mathcal{Y}_t^\star} \|\mathbf{y} - \mathbf{y}'\| \right)$$
3. The Hausdorff distance between the estimated Pareto front $\mathcal{Y}_t^\star$ and the actual Pareto front $\mathbb{Y}^\star$:
$$d_{\mathcal{H}} = d(\mathcal{Y}_T^\star, \mathbb{Y}^\star) = \max\left(d_{\mathbb{Y}^\star}, d_{\mathcal{Y}_T^\star}\right)$$
4. Simulation time per recommendation produced.

Results are summarised in table 1. It may be seen from this that in terms of the Hausdorff distance between estimated and actual Pareto fronts AD-GP-UCB consistently outperformed EHI in this experiment. Moreover although the EHI estimated Pareto front was closer to the actual Pareto front (measure 1) the AD-GP-UCB estimated Pareto front better approximated the actual Pareto front in terms of coverage (measure 2). We note that the hypervolume dominated by the AD-GP-UCB estimated Pareto front was consistently higher than the hypervolume dominated by the EHI estimated Pareto front. Figure 5 shows the average time required per recommendation for each of the algorithms. We have chosen this measure to factor out extraneous fluctuations observed in the estimated Pareto set size generated by the EHI method as n varied.

To aid visualisation the estimated Pareto fronts for AD-GP-UCB and EHI in the case $n = 3$ are shown in figure 4, where the Pareto front for DTLZ2 consists of a first-quadrant unit sphere in objective space (Deb et al., 2005). From this figure it may be seen that EHI constructs a cluster of recommendations that are close to a fragment of the actual Pareto front; whereas AD-GP-UCB creates a more diverse coverage of the Pareto front. We postulate that this results from the fact that AD-GP-UCB explicitly incorporates an exploration term $\sqrt{\beta_t}\eta_t$ in the acquisition function, encouraging greater exploration and hence more diversity in $\mathcal{Y}_T^\star$.

As noted previously, and as may be seen from table 1, the size of the estimated Pareto front found by EHI was surprisingly small for larger n. It is unclear why this occurs; however it appears to contribute to the significant fluctuation in the total time τ for required EHI to complete $T = 200$ iterations.

	EHI-based Bayesian Optimisation							AD-GP-UCB						
n	$N^\star$	$d_{\mathbb{Y}^\star} \downarrow$	$d_{\mathcal{Y}_T^\star} \downarrow$	$d_{\mathcal{H}} \downarrow$	HV $\uparrow$	$\tau \downarrow$	$\frac{\tau}{N^\star} \downarrow$	$N^\star$	$d_{\mathbb{Y}^\star} \downarrow$	$d_{\mathcal{Y}_T^\star} \downarrow$	$d_{\mathcal{H}} \downarrow$	HV $\uparrow$	$\tau \downarrow$	$\frac{\tau}{N^\star} \downarrow$
2	124	**0.007**	0.48	0.48	2.52	**162**	**1.31**	66	0.25	**0.13**	**0.25**	**3.18**	815	12.35
3	185	**0.027**	0.66	0.66	5.67	**174**	**0.94**	107	0.25	**0.25**	**0.25**	**7.33**	824	7.70
4	198	**0.25**	0.90	0.90	11.1	**358**	**1.81**	117	**0.25**	**0.35**	**0.35**	**15.5**	986	8.43
5	14	**0.25**	1.26	1.26	25.2	**243**	17.4	121	**0.25**	**0.49**	**0.49**	**31.5**	1012	**8.36**
6	187	**0.25**	1.16	1.16	51.5	3950	21.1	126	**0.25**	**0.71**	**0.71**	**63.3**	**992**	**7.87**
7	167	**0.25**	1.32	1.32	90.3	27546	165	141	**0.25**	**0.92**	**0.92**	**127**	**1024**	**7.26**
8	60	**0.24**	1.39	1.39	207	1483	24.7	180	0.25	**0.99**	**0.99**	**253**	**1227**	**6.82**
9	46	**0.20**	1.38	1.38	378	2025	44.0	174	0.24	**1.15**	**1.15**	**499**	**1444**	**8.30**
10	32	**0.20**	1.41	1.41	818	**973**	30.4	200	0.23	**1.13**	**1.13**	**978**	2795	**13.98**

Table 1: Results summary for DTLZ2 optimisation over range of n. In this table $d_{\mathbb{Y}^\star}$ measures how close the estimated Pareto front is to the actual Pareto front (optimality); $d_{\mathcal{Y}_T^\star}$ measures the maximum distance from any point on the actual Pareto front to any point in the estimated Pareto front (coverage) (for calculation $\mathbb{Y}^\star$ is approximated as a projected grid); $d_{\mathcal{H}}$ is the Hausdorff distance between the estimated and actual Pareto fronts; and HV is the dominated hypervolume. $T = 200$ iterations were used, producing an estimated Pareto front $\mathcal{Y}_T^\star$ containing $N^\star = |\mathcal{Y}_T^\star|$ recommendations in τ seconds - ie. $\tau/N^\star$ recommendations per second. $\uparrow$ indicates that larger values are preferable and $\downarrow$ that smaller values are preferable.

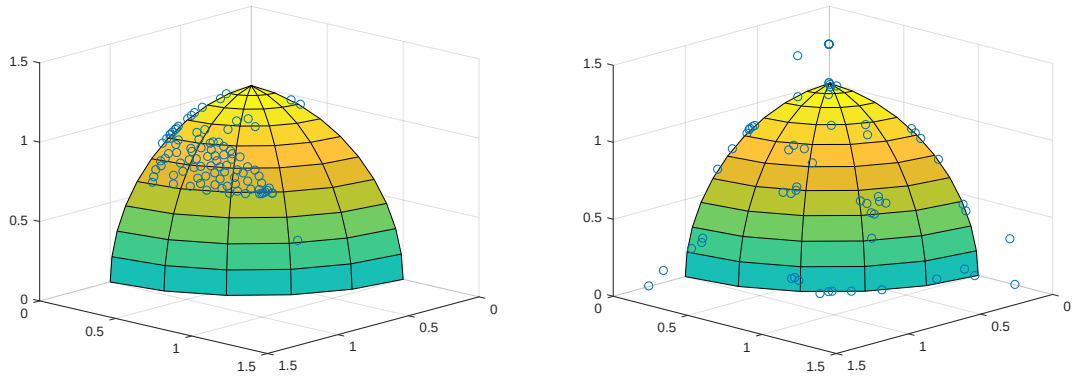

Figure 4: Estimated Pareto fronts for EHI (left) and AD-GP-UCB (right) for $n = 3$ objectives. Note that DTLZ2 is a minimisation problem, so the BO maximises its negative.

6.2 HYPERPARAMETER SELECTION

In this experiment we have compared our algorithm to Bayesian multi-objective optimisation with an expected hypervolume improvement (EHI) based acquisition function. We consider hyper-parameter selection for multi-class classifiers in the absence of information about the relative importance of classes. As there is no objective way to compare (weight) the cost of misclassification for the different classes this is an example of a multi-objective optimisation, where the classification accuracy with respect to each class is a single objective.

For multi-class classification we used the CS-SVM algorithm (Shilton et al., 2012) in SVMHeavy (Shilton, 2001) with an RBF kernel with length-scale g. Performance on each class was measured using 10-fold cross-validation. The hyper-parameters being tuned were the CS-SVM trade-off parameter $C \in [0.1, 10]$ and the kernel parameter $g \in [0.1, 10]$. Three datsets from the UCI collection (Dheeru and Karra Taniskidou, 2017) were used: SAT (6 classes, $N = 4435$ vectors), SEG (7 classes, $N = 2310$), and WAV (3 classes, $N = 5000$).

Results of simulations are shown in figure 6. These figures show both hypervolume as a function of iteration number (the hypervolume is used as a measure of the optimality of the Pareto set) and also the time required to recommend the next sample at each iteration. As may be seen from the graphs our proposed method is significantly faster than EHI. In fact, in the higher-dimensional

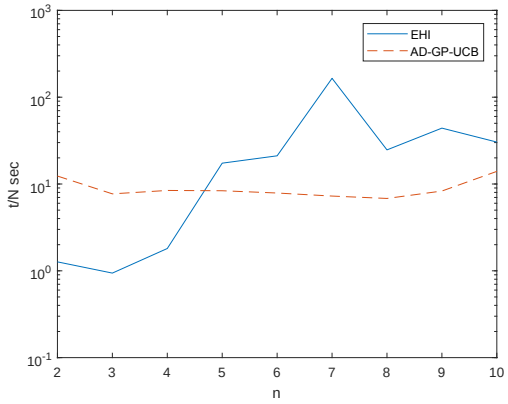

Figure 5: Average time taken to produce Pareto front in seconds/recommendation ($\tau/N^{\star}$).

cases - namely SAT (6 classes/objectives) and SEG (7 classes/objectives) - the EHI simulations had to be terminated early due to excessive computation time when computing the next recommendation (up to 1 day to produce a single recommendation). With regard to optimality (as measured by dominated hypervolume) it is somewhat difficult to say with certainty, but based on the WAV dataset at least our algorithm is certainly competitive (particularly given that the EHI alternative was unable to finish for either the SAT or SEG datasets due to excessive computational load resulting from EHI calculations made inside the global optimisation DIRECT call).

6.3 ALLOY DESIGN

High-temperature creep resistant Ni-superalloy is used for making boilers of super-critical thermal power plants. In a joint project with metallurgist we were asked to optimize the current alloy recipe to obtain superior creep resistance to the industry standard. The alloy consist of Ni, Cr, Co, Al, Ti, Mo, Ta, W and V. We use ThermoCalc software for phase simulation i.e. to predict what compounds (phases) get formed at a given temperature. Based on the existing knowledge, phases were clubbed into either good or bad for creep resistance. The phase simulation is performed at 6 different temperatures and a total of 12 objectives are created. Recommendation times for EHI were found to be excessive ($\sim$ 1 day), whereas our method was able to complete the task without difficulty.

Results for our simulation are shown in figure 7. In these figures dominated hypervolume has been used as a measure of convergence (Zitzler, 1999; Huband et al., 2003; Purshouse, 2003; Laumanns et al., 2000; Fleischer, 2000). The GP length scale in these results is 20 and was selected experimentally to optimise the rate of convergence; and the time budget $T = 100$ was chosen for practical reasons. A total of 21 Pareto-optimal alloys were found by our simulation. As may be seen our algorithm was able to calculate a set of Pareto-optimal recommendations within a reasonable time-frame despite the high number of objectives to provide the experimentalist with a good selection of options for further investigation.

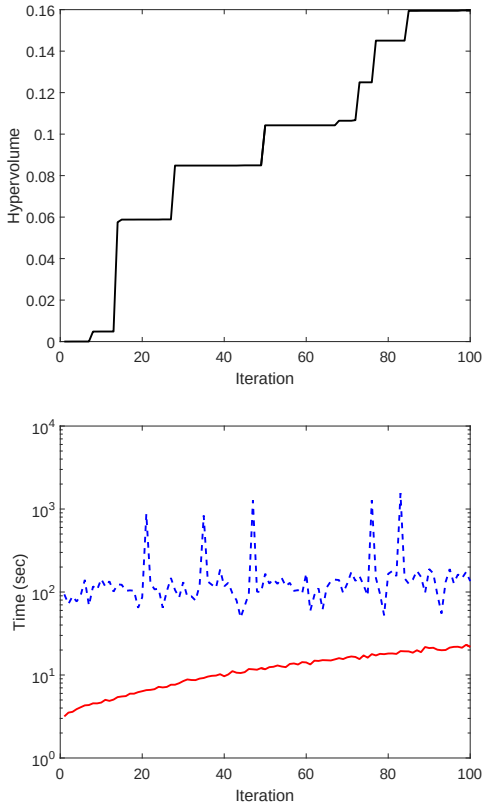

Figure 7: Simulation results for alloy design. Top: dominated hypervolume. Bottom: recommendation time (solid red), ThermoCalc simulation time (dashed blue).

7 CONCLUSIONS

In this paper we have proposed a method for Bayesian multi-objective optimisation based on score functions. Our proposed method is particularly well suited to the many-objective case where the number of objectives is significant and renders alternative methods such as EHI unsuitable due to reasons of computational infeasibility (for example, on 2 of our datasets we found that the EHI method failed early as the time required for a single recommendation grew to over 1 day). We have analysed the theoretical properties of our method and shown that it possesses properties such as sparseness, inherited from the 1-norm SVM, that make it well suited to the task. For

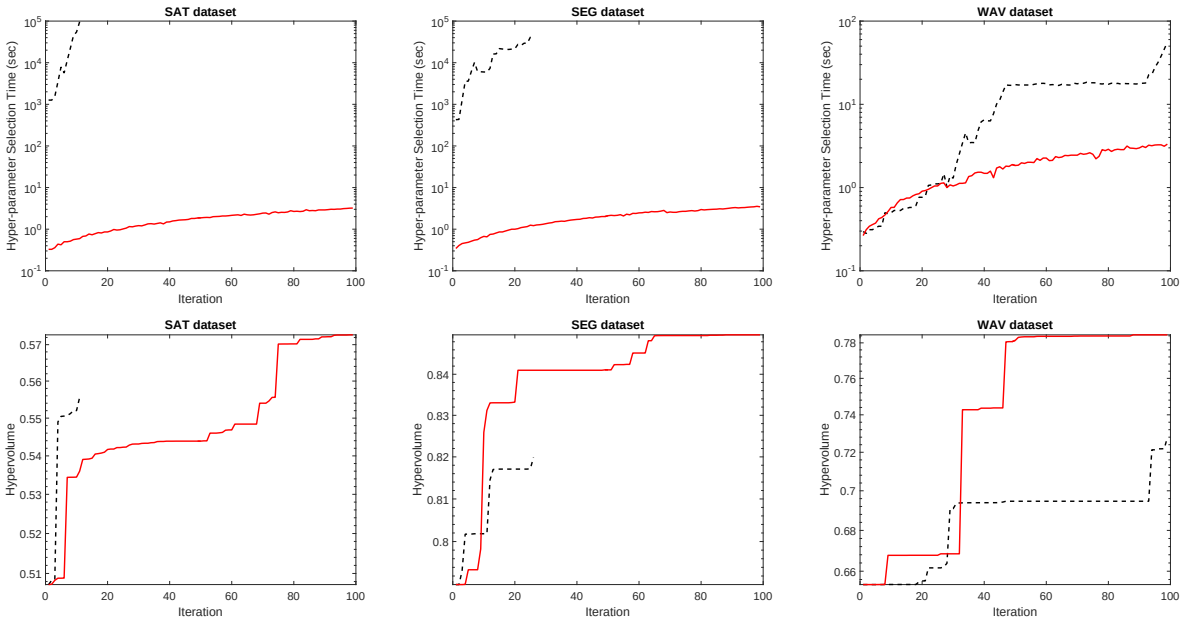

Figure 6: Hyper-parameter tuning results. Top row: recommendation times per iteration. Bottom row: enclosed hypervolume. EHI shown as black (dashed) line, our method as red (solid) line. EHI Simulations for SAT/SEG datasets had to be terminated early due to excessive recommendation times (we estimate for example that running the SAT dataset simulation to completion using the EHI method would have taken at least 3 months, which is clearly impractical).

experimental validation we have applied our proposed method to high-temperature alloy design and hyperparameter selection in the multi-class case where no information is provided with regard to the relative weight (or importance) of the classes. Our results clearly showed that our method is able to continue in cases where EHI breaks down due to computational complexity, and moreover that the results achieved by our method are competitive with those achieved by EHI.

ACKNOWLEDGEMENTS

This research was partially funded by the Australian Government through the Australian Research Council (ARC) and the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning. Prof Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

References

Paul S. Bradley and Olvi L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*, pages 82–90, San Fansisco, 1998. Kaufmann.

Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with applications to active user modeling and heirarchical reinforcement learning. eprint arXiv:1012.2599, arXiv.org, December 2010.

Coello A. Coello Coello, David A. Van Veldhuizen, and Gary B. Lamont. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishing, New York, 2002.

Kalyanmoy Deb. *Multi-Objective Optimization using Evoluationary Algorithms*. John Wiley & Sons, Chichester, UK, 2001.

Kalyanmoy Deb, Lothar Thiele, Marco Laumanns, and Eckart Zitzler. Scalable multi-objective optimization test problems. *Evolutionary Multiobjective Optimization. Theoretical Advances and Applications*, pages 105–145, 2005.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Michael Emmerich and Jan-Willem Klinkenberg. The computation of the expected improvement in dominated hypervolume of pareto front approximations. Technical report, Leiden University, 2008.

Mark Fleischer. The measure of pareto optima: Applications to multi-objective metaheuristics. In *Inter-*

national Conference on Evolutionary Multi-Criterion Optimization, pages 519–533, 2000.

Simon Huband, Phil Hingston, Lyndon While, and Luigi Barone. An evolution strategy with probabilistic mutation for multi-objective optimisation. In *Proceedings of the IEEE Congress on Evolutionary Computation*, volume 4, pages 2284–2291, 2003.

Iris Hupkens, André Deutz, Kaifeng Yang, and Michael Emmerich. Faster exact algorithms for computing expected hypervolume improvement. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 65–79. Springer, 2015.

Hisao Ishibuchi, Noritaka Tsukamoto, and Yusuke Nojima. Evolutionary many-objective optimization: A short review. In *Proceedings of the 2008 IEEE Confress on Evolutionary Computation*, pages 2424–2431, June 2008.

Donald R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993. ISSN 1573-2878. doi: 10.1007/BF00941892. URL `http://dx.doi.org/10.1007/BF00941892`.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Andy J. Keane. Statistical improvement criteria for use in multiobjective design optimization. *Journal of the American Institute of Aeronautics and Astronautics AIAA*, 44(4):879–891, 2006.

Harold J. Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1): 97–106, 1964.

Marco Laumanns, Eckart Zitzler, and Lothar Thiele. A unified model for multi-objective evolutionary algorithms with elitism. In *Proceedings of the 2000 Congress on Evolutionary Computation*, volume 1, pages 46–53, 2000.

Cheng Li, David Rubin de Celis, Santu Rana, Sunil Gupta, Alessandra Sutti, Stewart Greenhill, Teo Slezak, Murray Height, and Svetha Venkatesh. Rapid bayesian optimisation for synthesis of short polymer fiber materials. *Nature Scientific Reports*, 7, 2017.

David J. C. MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168, 1998.

Kaisa Miettinen. *Multi-Objective Optimization Using Evolutionary Algorithms*. Springer US, 1999.

Conrado S. Miranda and Fernando J. Von Zuben. Necessary and sufficient conditions for surrogate functions of pareto frontiers and their synthesis using gaussian processes. *IEEE Transactions on Evolutionary Computation*, PP(99):1–1, May 2015.

Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. In *Towards Global Optimization*, volume 2, pages 117–129. September 1978. ISBN 0-444-85171-2.

Wolfgang Ponweiser, Tobias Wagner, Dirk Biermann, and Markus Vincze. Multiobjective optimization on a limited budget of evaluations using model-assisted S-Metric selection. In *Proceedings of the 10th International Conference on Parallel Problem Solving from Nature: PPSN X*, pages 784–794, Berlin, Heidelberg, 2008. Springer-Verlag.

Robin Charles Purshouse. *On the Evolutionary Optimisation of Many Objectives*. PhD thesis, University of Sheffield, 2003.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, Redmond, 1999.

Alistair Shilton. SVMHeavy: a support vector machine optimiser, 2001.

Alistair Shilton, Daniel T. H. Lai, and Marimuthu Palaniswami. The conic-segmentation support vector machine - a target space method for multiclass classification. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8, June 2012.

Koji Shimoyama, Shinkyu Jeong, and Shigeru Obayashi. Kriging-surrogate-based optimization considering expected hypervolume improvement in non-constrained many-objective test problems. In *Proceedings of 2013 IEEE Congress on Evolutionary Computation*, 2013.

Ofer M. Shir, Michael Emmerich, Thomas Back, and Marc J. J. Vrakking. The application of evolutionary multi-criteria optimization to dynamic molecular aligment. In *Proceedings of 2007 IEEE Congress on Evolutionary Computation*, 2007.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58 (5):3250–3265, May 2012.

Tobias Wagner, Michael Emmerich, André Deutz, and Wolfgang Ponweiser. On expected-improvement criteria for model-based multi-objective optimization. In *Proceedings of the 2010 International Conference on Parallel Problem Solving from Nature*, pages 718–727, 2010.

Yeboon Yun, Hirotaka Nakayama, and Masao Arakava. Generation of pareto frontiers using support vector machines. In *International Converence on Multiple Criteria Decision Making*, 2004.

Martin Zaefferer, Thomax Bartz-Beielstein, Boris Naujoks, Tobias Wagner, and Michael Emmerich. A case study on multi-criteria optimization of an event detection software under limited budgets. In *Proceedings of the 2013 International Conference on Evolutionary Multi-Criterion Optimization*, pages 756–770. Springer, 2013.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *The Annual Conference on Neural Information Processing Systems* 16, 2004.

Eckart Zitzler. *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*. PhD thesis, Swiss Federal Institute of Technology Zurich, 1999.