Tracking Approximate Solutions of Parameterized Optimization Problems over Multi-Dimensional (Hyper-)Parameter Domains

Katharina Blechschmidt KATHI.JENA@WEB.DE
Joachim Giesen JOACHIM.GIESEN@UNI-JENA.DE
Sören Laue SOEREN.LAUE@UNI-JENA.DE
Friedrich-Schiller-Universität Jena, Germany

Abstract

Many machine learning methods are given as parameterized optimization problems. Important examples of such parameters are regularization- and kernel hyperparameters. These parameters have to be tuned carefully since the choice of their values can have a significant impact on the statistical performance of the learning methods. In most cases the parameter space does not carry much structure and parameter tuning essentially boils down to exploring the whole parameter space. The case when there is only one parameter received quite some attention over the years. First, algorithms for tracking an optimal solution for several machine learning optimization problems over regularization- and hyperparameter intervals had been developed, but since these algorithms can suffer from numerical problems more robust and efficient approximate path tracking algorithms have been devised and analyzed recently. By now approximate path tracking algorithms are known for regularization- and kernel hyperparameter paths with optimal path complexities that depend only on the prescribed approximation error. Here we extend the work on approximate path tracking algorithms with approximation guarantees to multi-dimensional parameter domains. We show a lower bound on the complexity of approximately exploring a multi-dimensional parameter domain that is the product of the corresponding path complexities. We also show a matching upper bound that can be turned into a theoretically and practically efficient algorithm. Experimental results for kernelized support vector machines and the elastic net confirm the theoretical complexity analysis.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

We consider parameterized optimization problems of the form

$$\min_{x \in F_t} f_t(x), \tag{1}$$

where $t \in \Omega \subseteq \mathbb{R}^p$ is a parameter vector, Ω is the parameter domain whose dimension is p, $f_t : \mathbb{R}^d \to \mathbb{R}$ is some function depending on t, and $F_t \subseteq \mathbb{R}^d$ is the feasible region of the optimization problem at parameter value $t \in \Omega$.
The parameter vector t is typically tuned by minimizing some measure of generalization error on test data while an optimal solution to Problem (1) at a given parameter vector t is computed from training data. Other criteria like the sparsity of the solution can also be relevant for the choice of t. In any case, for optimizing t it is necessary to track an optimal or approximately optimal solution of Problem (1) over the whole parameter domain Ω.

The one-dimensional case. The case $p = 1$, i.e., one-dimensional parameter domains, has been extensively studied mostly in the context of regularization paths, i.e., parameterized optimization problems of the form

$$f_t(x) = r(x) + t \cdot l(x),$$

where $l : \mathbb{R}^d \to \mathbb{R}$ is a loss function and $r : \mathbb{R}^d \to \mathbb{R}$ is some regularizer, e.g., $r(x) = \|x\|_2^2$ that enables the kernel trick, or $r(x) = \|x\|_1$ that promotes sparse solutions. The work on regularization paths started with the work by (Efron et al., 2004) who observed that the regularization path of the LASSO is piecewise linear. In (Rosset & Zhu, 2007) a fairly general theory of piecewise linear regularization paths has been developed and exact path following algorithms have been devised. Important special cases are support vector machines whose regularization paths have been studied in (Zhu et al., 2003; Hastie et al., 2004), support vector regression, where also the loss-sensitivity parameter can be tracked (Wang et al., 2006b), and the generalized LASSO (Tibshirani & Taylor, 2011). From the beginning it was known, see for example (Allgower &

Georg, 1993; Hastie et al., 2004; Bach et al., 2004), that exact regularization path following algorithms suffer from numerical instabilities as they repeatedly need to invert a matrix whose condition number can be poor, especially when using kernels. It also turned out (Gärtner et al., 2012; Mairal & Yu, 2012) that the combinatorial- and thus also computational complexity of exact regularization paths can be exponential in the number of data points. This triggered the interest in approximate path algorithms (Rosset, 2004; Friedman et al., 2007). By now numerically robust, approximate regularization path tracking algorithms are known for many problems including support vector machines (Giesen et al., 2012b;c), the LASSO (Mairal & Yu, 2012), and regularized matrix factorization- and completion problems (Giesen et al., 2012a;c). These algorithms compute a piecewise constant approximation with $O\big(1/\sqrt{\varepsilon}\big)$ segments, where $\varepsilon > 0$ is the guaranteed approximation error. Notably, the complexity is independent of the number of data points and even matching lower bounds are known (Giesen et al., 2012c).

Another important example that involves a one-dimensional parameter domain is when f_t is given as a function $f : \mathbb{R}^d \to \mathbb{R}$ that is parameterized by a positive kernel function $k_t : X \times X \to \mathbb{R}$ that itself is parameterized by $t \in \mathbb{R}$. This leads to the kernel hyperparameter path tracking problem that has been first studied by (Wang et al., 2007b) for kernelized support vector machines, by (Wang et al., 2007a) for the kernelized LASSO, and by (Wang et al., 2006a; 2012) for Laplacian-regularized semi-supervised classification. All this work addresses the exact path tracking problem which is also prone to numerical problems. A numerically robust and efficient algorithm for approximate kernel path tracking has been designed and analyzed by (Giesen et al., 2014). The path complexity of this algorithm is in $O(1/\varepsilon)$, where $\varepsilon > 0$ is again the guaranteed approximation error. A matching lower bound shows that this is optimal.

The multi-dimensional case. In contrast to the one-dimensional case, most methods for the multi-dimensional case are heuristics that do not come with guarantees.

Still the most commonly used method for multi-parameter tuning is a grid or manual search over the parameter domain. As (Bergstra & Bengio, 2012) have shown, a simple random search can yield better results than grid search, when the different parameters are not independent or not equally important since this can lower the effective dimension of the parameter domain.

Recently global optimization techniques, especially Bayesian optimization, have been used successfully for parameter tuning over large continuous, discrete and mixed parameter domains for various machine learning problems, see for example (Hutter et al., 2011; Bergstra et al., 2011; Snoek et al., 2012) and the references therein.

Contributions. Here we address the multi-dimensional case for continuous parameter domains. The complexity of the parameter domain exploration task can be measured in the number of near optimal solutions that need to be computed for different parameter vectors such that the gamut of these solutions is sufficient to provide an approximate solution with prescribed error bound on the whole parameter domain. We show matching upper and lower bounds on this complexity for multi-parameter domains. We also turn the upper bound construction into a numerically stable and practically efficient algorithm for low dimensional problems.

2. Definitions and problem set-up

Our results apply to a fairly general class of parameterized convex optimization problems, namely problems of the form

$$\min_{x \in \mathbb{R}^d} f_t(x) \quad \text{s.t.} \quad c_t(x) \leq 0, \tag{2}$$

where $f_t : \mathbb{R}^d \to \mathbb{R}$ is convex and $c_t : \mathbb{R}^d \to \mathbb{R}^n$ is convex in every component $c_t^i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, n$, for all parameter vectors $t \in \Omega \subseteq \mathbb{R}^p$. We assume that $f_t(x)$ and $c_t(x)$ are Lipschitz continuous in t at any feasible point x, but we do not require convexity (or concavity) of these functions in t. The feasible region at t is given as

$$F_t = \big\{x \in \mathbb{R}^d \,|\, c_t(x) \leq 0\big\},$$

with componentwise inequalities.

Lagrangian duality. The *Lagrangian* of the parameterized convex optimization problem (2) is the function

$$\ell_t : \mathbb{R}^d \times \mathbb{R}^n_{\geq 0} \to \mathbb{R},\ (x, \alpha) \mapsto f_t(x) + \alpha^T c_t(x),$$

from which we derive a *dual optimization problem* as

$$\max_{\alpha \in \mathbb{R}^n} \min_{x \in \mathbb{R}^d} \ell_t(x, \alpha) \quad \text{s.t.} \quad \alpha \geq 0.$$

We call

$$\hat{\varphi}_t : \mathbb{R}^n \to \mathbb{R},\ \alpha \mapsto \min_{x \in \mathbb{R}^d} \ell_t(x, \alpha).$$

the *dual objective function*. From the Lagrangian we can also derive an alternative expression for the primal objective function, namely

$$\varphi_t : \mathbb{R}^d \to \mathbb{R},\ x \mapsto \max_{\alpha \geq 0} \ell_t(x, \alpha)$$

Note that $f_t(x) = \varphi_t(x)$ for all $x \in F_t$ since $\alpha^T c_t(x) \leq 0$ and thus $\max_{\alpha \geq 0} \alpha^T c_t(x) = 0$ (which can always be obtained by setting $\alpha = 0$) for all $x \in F_t$.

Weak and strong duality. At a fixed parameter vector t we have the following well known *weak duality* property

$$\hat{\varphi}_t(\alpha) \leq \varphi_t(x)$$

for any $x \in \mathbb{R}^d$ and any $\alpha \in \mathbb{R}^n_{\geq 0}$. In particular, we have $\hat{\varphi}_t(\alpha_t^*) \leq \varphi_t(x_t^*)$, where

$$\alpha_t^* = \operatorname{argmax}_{\alpha \geq 0} \hat{\varphi}_t(\alpha) \text{ and } x_t^* = \operatorname{argmin}_{x \in F_t} \varphi_t(x)$$

are the dual and primal optimal solutions, respectively. We say that *strong duality* holds if $\hat{\varphi}_t(\alpha_t^*) = \varphi_t(x_t^*)$ for all $t \in \Omega$. In the following we assume that strong duality holds.

Duality gap and approximate solution. At parameter vector t we call

$$g_t(x, \alpha) = \varphi_t(x) - \hat{\varphi}_t(\alpha)$$

the *duality gap* at $(x, \alpha) \in F_t \times \mathbb{R}^n_{\geq 0}$. For $\varepsilon > 0$, we call $x \in F_t$ an *ε-approximate solution* of the parameterized optimization problem (2) at parameter vector t, if

$$f_t(x) - f_t(x_t^*) \leq \varepsilon.$$

Assume that $g_t(x, \alpha) \leq \varepsilon$, then we have

$$\begin{aligned} f_t(x) - f_t(x_t^*) &= \varphi_t(x) - \varphi_t(x_t^*) \\ &= \varphi_t(x) - \hat{\varphi}_t(\alpha) + \hat{\varphi}_t(\alpha) - \varphi_t(x_t^*) \\ &= g_t(x, \alpha) - \big(\varphi_t(x_t^*) - \hat{\varphi}_t(\alpha)\big) \\ &\leq g_t(x, \alpha) \leq \varepsilon. \end{aligned}$$

Approximate solution gamut. Let

$$Q := \prod_{i=1}^{p} [t_{(i,\min)}, t_{(i,\max)}] \subset \mathbb{R}^p$$

be a compact parameter cuboid and $\varepsilon > 0$. We call a function

$$x : Q \to \mathbb{R}^d,\ t \mapsto x(t)$$

an *ε-approximate solution gamut* of the parameterized optimization problem (2), if for all $t \in Q$

$$1.\quad x(t) \in F_t \qquad \text{and} \qquad 2.\quad f_t(x(t)) - f_t(x^*(t)) \leq \varepsilon.$$

We say that the function $x : Q \to \mathbb{R}^d$ has a combinatorial *complexity* $k \in \mathbb{N}$, if x can be computed from k primal-dual pairs $(x(t_i), \alpha(t_i))$ with $t_i \in Q$, $i = 1, \ldots, k$.

The goal of this paper is to give upper and lower bounds on the complexity of ε-approximate solution gamuts, and to devise efficient algorithms for computing them.

3. Complexity of solution gamuts

We show matching upper and lower bounds on the combinatorial complexity of approximate solution gamuts by providing lower and upper bounds, respectively, on the size of the regions where near optimal primal-dual pairs remain good approximate solutions. The latter bounds are derived from the corresponding complexity analysis for the one-dimensional case, i.e., the complexity of solution paths. It turns out that the bounds for the multi-dimensional case are the product of the corresponding path complexities, i.e., the complexity along the paths where all but one parameter are fixed.

Upper bound on the gamut complexity. The known algorithms for computing approximate solution paths for one-dimensional parameterized optimization problems (Giesen et al., 2014; Mairal & Yu, 2012; Giesen et al., 2012c) essentially make use of two problem dependent families of functions (shift functions):

$$\begin{aligned} x_t &: [t_{\min}, t_{\max}] \to \mathbb{R}^d,\ \tau \mapsto x_t(\tau) \\ \alpha_t &: [t_{\min}, t_{\max}] \to \mathbb{R}^n_{\geq 0},\ \tau \mapsto \alpha_t(\tau) \end{aligned}$$

for $t \in [t_{\min}, t_{\max}]$. The functions x_t are such that a primal feasible solution $x \in \mathbb{R}^d$ at parameter value t is mapped to a feasible solution $x_t(\tau)$ at parameter value τ, and $x_t(t) = x$. Analogously, the α_t are such that a dual feasible solution $\alpha \in \mathbb{R}^n_{\geq 0}$ at parameter value t is mapped to a feasible solution $\alpha_t(\tau)$ at parameter value τ, and $\alpha_t(t) = \alpha$. For the approximate path algorithms to be efficient, the functions x_t and α_t need to satisfy some continuity conditions and need to be efficiently computable.

The crucial property that allows an efficient computation of approximate solution paths for one-dimensional parameterized optimization problems is that the duality gap for the primal-dual pair $(x_t(\tau), \alpha_t(\tau))$ at parameter value $\tau \in [t, t + \Delta t]$ can be bounded by the duality gap at parameter value $\tau = t$ as

$$g_\tau(x_t(\tau), \alpha_t(\tau)) \leq g_t(x_t(t), \alpha_t(t)) + e(\Delta t),$$

where $e : [0, t_{\max} - t_{\min}] \to \mathbb{R}$ is some (error) function that depends on the specific optimization problem and the shift functions, but not on t. For a large class of regularization path problems it was shown in (Mairal & Yu, 2012; Giesen et al., 2012c) that there exist shift functions such that $e(\Delta t) = L^2(\Delta t)^2$, where L is some problem dependent constant that can be computed explicitly for many problems and the appropriate shift functions. Thus any given primal-dual pair $(x_t(\tau), \alpha_t(\tau))$ that is an ϵ/γ-approximate solution for $\gamma > 1$ at parameter value $\tau = t$, i.e., $g_t(x_t(t), \alpha_t(t)) \leq \varepsilon/\gamma$, is still at least an ε-approximation on the whole interval $[t, t + \Delta t]$ for $\Delta t \leq \sqrt{\varepsilon}/L$. For several kernel-hyperparameter path problems

it was shown in (Giesen et al., 2014) that there exist shift functions such that $e(\Delta t) = L\Delta t$, where L is again some problem dependent constant. Thus any given primal-dual pair $(x_t(\tau), \alpha_t(\tau))$ that is an ϵ/γ-approximate solution at parameter value $\tau = t$ is still at least an ε-approximation on the whole interval $[t, t + \Delta t]$ for $\Delta t \leq \varepsilon/L$.

The approach from above can be generalized to p parameters if we already have shift functions for the corresponding one-dimensional problems, i.e., keeping all but one parameter fixed. Similarly as in the one-dimensional case, we assume that there exist error functions e_i such that at any parameter vector $t = (t_1, \ldots, t_p)$,

$$g_{\tau_i}\big(x_t(\tau_i), \alpha_t(\tau_i)\big) \leq g_t(x_t(t), \alpha_t(t)) + e_i(\Delta t_i) \quad (i = 1, \ldots, p)$$

for all $\tau_i = (t_1, \ldots, \hat{\tau}_i, \ldots, t_p)$ with $\hat{\tau}_i \in [t_i, t_i + \Delta t_i]$. Here the p-dimensional shift function $x_t(\cdot)$ is defined such that $x_t(\tau_i)$ is the i-th one-dimensional shift function applied to τ_i for fixed $t_j, j \neq i$, and similarly for $\alpha_t(\cdot)$. Combining these inequalities for the duality gap iteratively gives

$$g_{\tau_i}(x_t(\tau_i), \alpha_t(\tau_i)) \leq g_t(x_t(t), \alpha_t(t)) + \sum_{j=1}^{i} e_i(\Delta t_j)$$

for all $\tau_i \in \prod_{j=1}^{i}[t_j, t_j + \Delta t_j]$, $i = 1, \ldots, p$. Let Δt_i be such that if $(x_s(s), \alpha_s(s))$ is an $\varepsilon/\gamma^{p-i+1}$-approximation at some parameter vector $s = (s_1, \ldots, s_p)$, then $(x_s(\tau_i), \alpha_s(\tau_i))$ is at least an ε/γ^{p-i}-approximation for all τ_i in the interval $\big[s, (s_1, \ldots, s_i + \Delta t_i, \ldots, s_p)\big]$. It follows inductively that any primal-dual pair $(x_t(t), \alpha_t(t))$ that is an ϵ/γ^p-approximate solution for $\gamma > 1$ at parameter vector $t = (t_1, \ldots, t_p)$ is at least an ε-approximation on the whole cuboid $\prod_{i=1}^{p}[t_i, t_i + \Delta t_i]$. This results in a ε-approximate solution gamut complexity for a parameter cuboid $Q = \prod_{i=1}^{p}[t_{(i,\min)}, t_{(i,\max)}]$ of at most

$$\prod_{i=1}^{p} \frac{t_{(i,\max)} - t_{(i,\min)}}{\Delta t_i},$$

i.e., the solution gamut complexity can be upper bounded by the product of the corresponding path complexities. For instance, if all the Δt_i are in $\Omega(\sqrt{\varepsilon})$, i.e., as for regularization paths, then the solution gamut complexity is in $O\big(\varepsilon^{-p/2}\big)$.

Lower bound on the gamut complexity. In the one-dimensional case matching lower bounds for path complexities are known. These lower bounds result from upper bounds on Δt. For regularization paths it was shown that $\Delta t \in O(\sqrt{\varepsilon})$ and for kernel-hyperparameter paths it was shown that $\Delta t \in O(\varepsilon)$. Hence, the path complexity is in $\Omega(1/\sqrt{\varepsilon})$ for regularization paths and in $\Omega(1/\varepsilon)$ for kernel-hyperparameter paths.

For constructing a matching lower bound example in the multi-parameter case we consider p problems of the form $\min f_{t_i}(x)$, with $f_{t_i}(x) \geq 0$ for all $x \in \mathbb{R}^d$, that are each parameterized by a single parameter $t_i, i = 1, \ldots, p$. Assume that the ε-approximate path complexity of the i-th problem is in $\Omega(\omega_i(\varepsilon))$. Then the problem

$$\min_{x \in \mathbb{R}^{pd}} \sum_{i=1}^{p} f_{t_i}(x_{[i]}), \tag{3}$$

where $x_{[i]} = \big(x_{(i-1)d+1}, \ldots, x_{id}\big)$, has a solution gamut complexity in $\Omega\Big(\prod_{i=1}^{p} \omega_i(\varepsilon)\Big)$. To see this, let (x_t^*, α_t^*) be an optimal primal-dual pair at some parameter vector t. The region where this pair remains an ε-approximation must be contained in a cuboid Q with side lengths $2\Delta t_i \in O(1/\omega_i(\varepsilon))$ since all the terms $f_{t_i}(x_{[i]})$ need to be optimized independently. The volume of the cuboid Q is

$$2^p \prod_{i=1}^{p} \Delta t_i \in \prod_{i=1}^{p} O(1/\omega_i(\varepsilon)) = O\left(\Big(\prod_{i=1}^{p} \omega_i(\varepsilon)\Big)^{-1}\right).$$

Thus we need at least $\Omega\Big(\prod_{i=1}^{p} \omega_i(\varepsilon)\Big)$ such cuboids to cover the whole parameter domain whose volume is independent of ε. Hence, the solution gamut complexity for Problem (3) can be lower bounded by the product of the corresponding path complexities.

4. Computing solution gamuts adaptively

Here we turn the upper bound construction from the previous section into an algorithm for computing an approximate solution gamut that inherits the theoretical complexity guarantee and is also practically efficient. The algorithm is based on two simple observations. First, the lower bound on Δt in the upper bound construction can be too pessimistic locally, and second, it is computationally much cheaper to evaluate the duality gap for a given primal-dual pair than to compute such a pair.

The upper bound construction from the previous section shows, that a lower bound $\sigma_i(\varepsilon, \gamma)$ on Δt_i such that an $\varepsilon/\gamma^{p-i+1}$-approximation at some parameter vector $t = (t_1, \ldots, t_p)$ remains at least an ε/γ^{p-i}-approximation on the whole interval $\big[t, (t_1, \ldots, t_i + \Delta t_i, \ldots, t_p)\big]$ guarantees that a grid search, i.e., computing ϵ/γ^p-approximate solutions at the vertices of the grid, on a grid with spacing $\sigma_i(\varepsilon, \gamma)$ in the i-th parameter direction (that only depends on the often explicitly known error function e_i) provides an ε-approximate solution gamut.

The idea now is to keep the grid, but trade the computation of primal-dual pairs for the evaluation of duality gaps at grid vertices. The adaptive algorithm works iteratively and stores at every grid vertex the primal-dual pair that has the

smallest duality gap so far. Once the duality gap at a grid vertex is smaller than the prescribed error bound $\varepsilon/\gamma^p > 0$ the grid vertex does not have to be considered anymore. More formally, the algorithm comprises the following initialization and iteration phases:

Initialization. Compute an optimal primal-dual pair

$$(x^*, \alpha^*) = \big(x_{t_{\min}}(t_{\min}), \alpha_{t_{\min}}(t_{\min})\big)$$

at the grid vertex $t_{\min} = \big(t_{1,\min}, \ldots, t_{p,\min}\big)$ and compute the duality gap of the pairs $\big(x_{t_{\min}}(t), \alpha_{t_{\min}}(t)\big)$ at all grid vertices t. Here $x_{t_{\min}}(\cdot)$ and $\alpha_{t_{\min}}(\cdot)$ are shift functions as defined in the previous section.

Iteration. While there is a grid vertex at which the stored duality gap is still larger than ε/γ^p: compute an optimal primal-dual pair at the grid vertex $t_{\max}$ at which the stored duality gap is maximal, and update the duality gap at all the grid vertices t, where the stored duality gap is larger than ε/γ^p, using the primal-dual pairs $\big(x_{t_{\max}}(t), \alpha_{t_{\max}}(t)\big)$, if the resulting duality gap is smaller than the stored gap.

5. Experiments

We consider two examples with a two-dimensional parameter domain each, namely kernelized support vector machines, that are parameterized by a regularization- and a kernel hyperparameter, and elastic net regularization, a regression method that has two regularization parameters.

5.1. Kernelized support vector machines (SVMs)

Primal and dual problem. We consider the standard hinge loss SVM with a kernel. The primal SVM optimization problem reads as

$$\begin{aligned} \min_{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \quad & \frac{1}{2} w^T K_\gamma w + c \cdot \|\xi\|_1 \\ \text{s.t.} \quad & y \odot (K_\gamma w + b) \geq 1 - \xi \text{ and } \xi \geq 0, \end{aligned}$$

where $c > 0$ is a regularization parameter, $y \in \mathbb{R}^n$ is a label vector with entries in $\{-1, +1\}$, $\odot$ is the element-wise multiplication, and K_γ is some kernel matrix that is parameterized by $\gamma > 0$. In our experiments we use the Gaussian kernel with bandwidth parameter γ, i.e.,

$$K_\gamma = \big(k_\gamma(x, x')\big) = \big(\exp(-\gamma \|x - x'\|_2^2)\big).$$

Hence, the two parameters to consider for kernelized SVMs are the regularization parameter c and the kernel hyperparameter γ.

The dual SVM problem is given as

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^n} \quad & -\frac{1}{2} (y \odot \alpha)^T K_\gamma (y \odot \alpha) + \|\alpha\|_1 \\ \text{s.t.} \quad & y^T \alpha = 0 \text{ and } 0 \leq \alpha \leq c. \end{aligned}$$

Shift functions. A dual solution α that is feasible for some parameter pair (c, γ) is also feasible for any parameter pair $(\hat{c}, \hat{\gamma})$ as long as $\hat{c} \geq c$. Whenever $\hat{c} < c$, we can obtain a dual feasible solution by scaling α appropriately. Hence, an easily computable shift function $\alpha_{(c,\gamma)}(\cdot)$ is given as

$$\alpha_{(c,\gamma)}\big(\hat{c}, \hat{\gamma}\big) = \begin{cases} \alpha & : \ \hat{c} \geq c \\ \alpha \cdot (\hat{c}/c) & : \ \hat{c} < c. \end{cases}$$

The corresponding one-dimensional shift functions are the identity function for the parameter γ and the shift function $\alpha \mapsto \alpha \cdot \max\{1, \hat{c}/c\}$ for the parameter c.
For primal solutions (w, b, ξ) we do not need explicit shift functions, because feasible primal solutions can be computed from feasible dual solutions. A primal solution w can be computed as $w = y \odot \alpha$ from a solution α to the dual problem. If α is an optimal dual solution, then the bias b can be computed as $b = y_i - K_\gamma(i, :)w$ for a support vector index i, i.e., where $0 < \alpha_i < c$ holds true. Here $K_\gamma(i, :)$ is the i-th row of K_γ. In the case that α is not an optimal dual solution, the bias is chosen such that the primal objective function value becomes minimal. This can be accomplished by a linear scan over the sorted vector $y \odot (K_\gamma w)$. Once w and b are given also ξ can be computed.

Computing the duality gap. From $\alpha_{(c,\gamma)}(\hat{c}, \hat{\gamma})$ and the corresponding feasible primal solutions (w, b, ξ) at $(\hat{c}, \hat{\gamma})$ we can directly compute the duality gap of the resulting primal-dual pair at any grid vertex $(\hat{c}, \hat{\gamma})$.

Experiments. In our implementation of the adaptive algorithm from Section 4 we used the LIBSVM package, see (Fan et al., 2005), to compute a near optimal dual solution at a given grid vertex, i.e., parameter pair (c, γ). We considered the two-dimensional parameter space (c, γ) with $c \in [2^{-10}, 2^{10}]$ and $\gamma \in [2^{-10}, 2^{10}]$, and a uniform grid with vertices at $(2^i, 2^j)$, where i and j were incremented in steps of 0.05, i.e., the grid had $400 \times 400 = 160,000$ vertices.
The data sets that have been used in our experiments were obtained from the LIBSVM website, see (Lin) for a description.

Discussion. From the upper bound analysis in Section 3 we know that there exists an ε-approximate solution gamut for the kernelized SVM problem whose complexity is at most the product of the regularization path complexity, which is in $O\big(1/\sqrt{\varepsilon}\big)$, and the kernel hyperparameter path complexity, which is in $O(1/\varepsilon)$. That is, there exists a solution gamut with complexity in $O\big(\varepsilon^{-3/2}\big)$. Such a solution gamut is indeed computed by our adaptive algorithm as can be seen from Table 1 and Figure 1. Notably, also the lower bound holds experimentally, i.e., the computed gamut has a complexity in $\Theta\big(\varepsilon^{-3/2}\big)$.

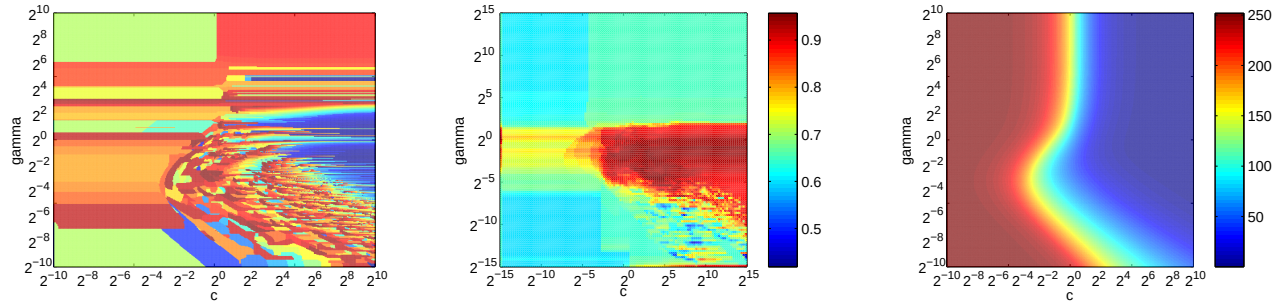

Figure 2. IONOSPHERE data set. Left: connected parameter regions that are covered by the same primal-dual pair by the adaptive algorithm are shown in the same color. Middle: 10-fold cross-validation values over the parameter domain. Right: optimal values for the primal kernelized SVM objective function over the parameter domain (remark: here the objective function value was scaled by $1/c$).

Table 1. Kernelized SVM: ε-solution gamut complexity for various data sets.

DATA SET	$\varepsilon = 2^2$	2^1	2^0	2^{-1}	2^{-2}	2^{-3}
A1A	29	62	155	659	2027	3643
A2A	21	45	118	444	1463	2752
A3A	23	45	114	434	1770	3333
A4A	21	42	93	329	1332	2706
DIABETES	1222	2389	3710	5030	6136	7420
HEART	602	1743	3273	4918	6868	9239
IONOSPHERE	909	1842	3021	5105	7420	9958

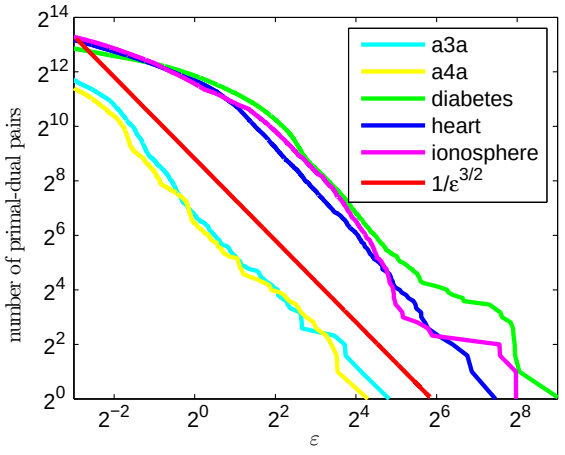

Figure 1. ε-solution gamut complexity for various data sets (log-log plot).

The adaptivity and practical efficiency of our algorithm can be seen in Figure 2. In Figure 2(left) grid regions are shown for the IONOSPHERE data set that are covered by one primal-dual pair. Note that many primal-dual pairs are sufficiently good solutions for wide ranges of parameter values which renders our adaptive algorithm much more efficient than a simple grid search. In Figure 2(middle) a 10-fold cross-validation plot is shown for the same data set. It can be seen that in regions where the cross-validation accuracy does not change much only very few (or even a single) primal-dual pairs are sufficient to cover the region, while in regions where the cross-validation accuracy changes a lot many primal-dual pairs are necessary. That is, the most primal-dual pairs are computed in statistically interesting regions. These regions cannot be determined by looking just at the optimal primal objective function values over the parameter domain that are shown in Figure 2(right). That is, the adaptive algorithm indeed adapts to the statistically interesting regions but not to regions with similar optimal objective function values.

5.2. Elastic Net regularization

Primal and dual problem. Elastic net regularization combines ℓ_2- and ℓ_1-regularization for linear regression. It is given as the following unconstrained optimization problem, see (Zou & Hastie, 2005),

$$\min_{x\in\mathbb{R}^d} \frac{1}{2n}\|Ax-y\|_2^2 + c\left(\frac{1-\lambda}{2}\|x\|_2^2 + \lambda\|x\|_1\right),$$

where $A \in \mathbb{R}^{n\times d}$ is the data matrix for n data points in $\mathbb{R}^d$ and $y \in \mathbb{R}^n$ are the corresponding responses. The problem is parameterized by $c \geq 0$ and $0 \leq \lambda \leq 1$. Special cases of the elastic net are ridge regression (for $\lambda = 0$) and the Lasso (for $\lambda = 1$), see (Tibshirani, 1996).

A standard calculation shows that the dual of the elastic net is the following constrained optimization problem

$$\begin{aligned}\max_{u\in\mathbb{R}^d} \quad & -\frac{1}{8n}(u+2A^Ty)^TQ(u+2A^Ty) + \frac{1}{2n}y^Ty \\ \text{s.t.} \quad & 0 \leq u \leq 2nc\lambda,\end{aligned}$$

where $Q \in \mathbb{R}^{d\times d}$ is the pseudoinverse of $A^TA+nc(1-\lambda)\mathbb{I}$ and $\mathbb{I} \in \mathbb{R}^{d\times d}$ is the identity matrix.

Shift functions. We do not need a shift function for the primal elastic net since it is an unconstrained problem, i.e.,

Table 2. Elastic net: ε-solution gamut complexity for various data sets.

DATA SET	$\varepsilon = 2^{-1}$	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}
ABALONE	9	14	17	32	75	136	251	438	750	1274
BODYFAT	2	3	5	7	12	22	38	69	137	280
CPUSMALL	6	16	26	35	56	85	145	215	376	670
PYRIM	2	2	3	5	7	14	29	52	99	202
SYNTHETIC ($n = 50$, $d = 40$)	3	3	6	10	17	37	62	152	365	771
SYNTHETIC ($n = 500$, $d = 100$)	2	3	5	7	10	22	52	106	220	408
SYNTHETIC ($n = 5000$, $d = 100$)	2	3	4	6	10	17	35	71	132	262
SYNTHETIC ($n = 5000$, $d = 1000$)	3	6	8	13	24	48	105	249	517	929

any $x \in \mathbb{R}^d$ is feasible for all admissible parameter pairs (c, λ). Hence, we only need shift functions for the dual problem. Note first, that an optimal solution u for the dual problem can be computed from an optimal solution x for the primal problem as

$$u \;=\; 2(A^T A + nc(1-\lambda)\mathbb{I})x - 2A^T y,$$

which follows from duality theory and some straightforward calculations. An optimal dual solution u at some parameter pair (c, λ) is a feasible solution for the dual problem at some other parameter pair $(\hat{c}, \hat{\lambda})$ whenever $\|u\|_\infty \leq 2n\hat{c}\hat{\lambda}$. Otherwise, we can scale u such that it becomes feasible. Thus, an easily computable shift function is given as

$$u_{(c,\lambda)}\big(\hat{c}, \hat{\lambda}\big) \;=\; \begin{cases} u & : \quad \|u\|_\infty \leq 2n\hat{c}\hat{\lambda} \\ u \cdot \frac{\hat{c}\hat{\lambda}}{c\lambda} & : \quad \|u\|_\infty > 2n\hat{c}\hat{\lambda}. \end{cases}$$

The corresponding one-dimensional shift function for the parameter c is $u \mapsto u \cdot \hat{c}/c$ if $\|u\|_\infty > 2n\hat{c}\lambda$, and the identity function otherwise. Analogously, the shift function for the parameter λ is $u \mapsto u \cdot \hat{\lambda}/\lambda$ if $\|u\|_\infty > 2nc\hat{\lambda}$, and the identity function otherwise.

Computing the duality gap. Given a primal solution x, the value of the primal objective function can be computed in constant time at any parameter pair $(\hat{c}, \hat{\lambda})$ from the value at (c, λ) since the computation boils down to evaluating a linear function in the product $c\lambda$. For computing the value of the dual objective function note that the matrix Q can be computed efficiently for varying values of c and λ from the singular value decomposition of $A^T A$. Let USU^T be the singular value decomposition of $A^T A$. We then have

$$Q \;=\; U(S + nc(1-\lambda)\mathbb{I})^{-1}U^T$$

in case that $c(1-\lambda) > 0$, and otherwise Q is simply the pseudoinverse of $A^T A$. Let $\delta = \frac{\hat{c}\hat{\lambda}}{c\lambda}$, computing the dual objective function value at some dual solution u now reduces to evaluating the expression

$$\begin{aligned}
&(\delta u + 2A^T y)^T Q(\delta u + 2A^T y) \\
&\quad = (\delta u + 2A^T y)^T \ldots \\
&\qquad \ldots U(S + n\hat{c}(1-\hat{\lambda})\mathbb{I})^{-1}U^T(\delta u + 2A^T y) \\
&\quad = (\delta U^T u + 2U^T A^T y)^T \ldots \\
&\qquad \ldots (S + n\hat{c}(1-\hat{\lambda})\mathbb{I})^{-1}(\delta U^T u + 2U^T A^T y) \\
&\quad = (\delta U^T u + 2U^T A^T y)^T \ldots \\
&\qquad \ldots \Big((\delta U^T u + 2U^T A^T y) \oslash (s + n\hat{c}(1-\hat{\lambda})\mathbf{1})\Big),
\end{aligned}$$

where $\oslash$ is the elementwise vector division, $s = \operatorname{diag}(S)$, and $\mathbf{1}$ is the all-ones vector. The last equality follows since S is a diagonal matrix. The values $U^T u$ and $2U^T A^T y$ can be precomputed for any optimal solution u. Hence, the dual objective function value for varying parameter pairs (c, λ) can be computed in time $O(d)$. Note that this is much faster than computing a primal-dual pair which amounts to a running time in $\Theta(d^{7/2})$.

Experiments. In our implementation of the adaptive algorithm from Section 4 we used GLMNET, see (Friedman et al., 2010), for solving the primal optimization problem at given parameter values for c and λ. Note that GLMNET allows to compute the exact regularization path for c. We considered parameter values $\lambda \in [0, 1]$ and $c \in [2^{-10}, 2^5]$. For the experiments we used standard data sets from the LIBSVM website and also generated synthetic data similarly as in (Friedman et al., 2010), i.e., the synthetic outcome values were generated as

$$Y \;=\; \sum_{i=1}^{k} X_i \beta_i + \alpha \cdot Z$$

where the X_i are Gaussian variables with d observations, the coefficients β_i are linearly decreasing, $Z \sim \mathcal{N}(0, 1)$, and α is chosen such that the signal-to-noise ratio is 3.

Discussion. From the upper bound analysis in Section 3 we know that there exists an ε-approximate solution gamut for the elastic net problem whose complexity is the product

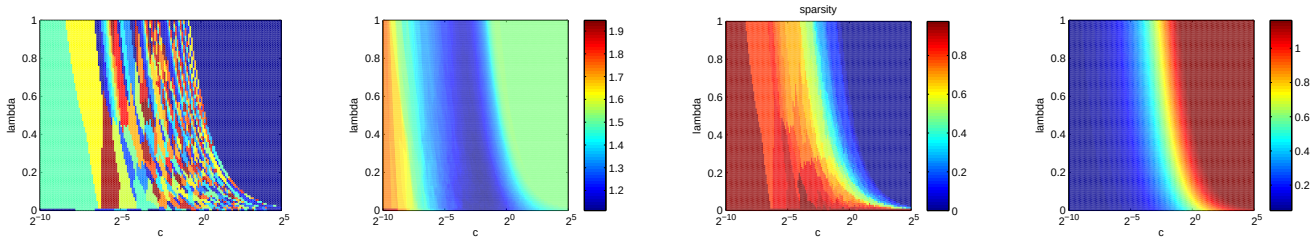

Figure 3. SYNTHETIC data set. Left: connected parameter regions that are covered by the same primal-dual pair by the adaptive algorithm are shown in the same color. Middle/left: 10-fold cross-validation RMSE values over the parameter domain. Middle/right: sparsity of the computed solution over the parameter domain. Right: optimal values for the primal elastic net function over the parameter domain.

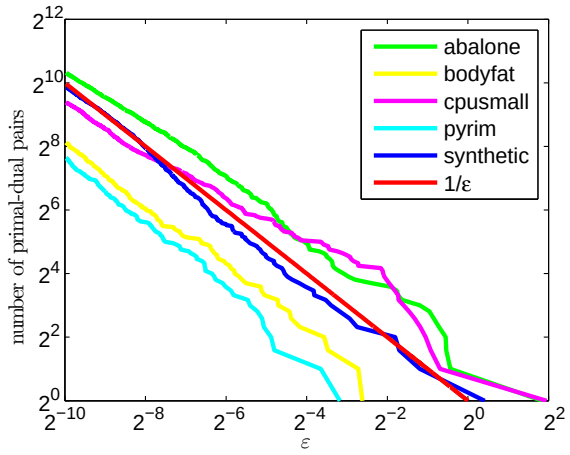

Figure 2. ε-solution gamut complexity for various data sets (log-log plot).

of two regularization path complexities each in $O\big(1/\sqrt{\varepsilon}\big)$. Hence, there exists a solution gamut with complexity in $O(1/\varepsilon)$. Again, such a solution gamut is computed by our adaptive algorithm as can be seen from Table 2 and Figure 2. Experimentally, the computed gamut also obeys the theoretical lower complexity bound in $\Omega\big(1/\varepsilon\big)$.

In Figure 3(left) grid regions are shown for the SYNTHETIC ($n = 50, d = 40$) data set that are covered by one primal-dual pair. As for the kernelized SVM many primal-dual pairs are again sufficiently good solutions for a wide range of parameter values not only for c but also for λ. This information is lost if one considers only the one-dimensional regularization path in c as it has been done previously. In Figure 3(middle/left) a 10-fold cross-validation RMSE plot is shown for the same data set. Also here it can be seen that in regions where the cross-validation accuracy does not change much only very few primal-dual pairs are sufficient to cover the region, while in regions where the cross-validation accuracy changes rapidly many primal-dual pairs are necessary. Also for the elastic net these statistically interesting regions cannot be determined by looking only at the optimal primal objective function values over the parameter domain that are shown in Figure 3(right). This holds also true for the sparsity of the solution that is shown in Figure 3(middle/right). Note that the sparsity of a solution is not necessarily a monotone function in c. A comparison of Figures 3(middle/left) and (middle/right) also shows that only exploring the whole parameter domain allows to make an informed trade-off between the two objectives of low RMSE and sparsity.

6. Conclusions

We addressed the problem of exploring multi-dimensional parameter domains of parameterized optimization problems that are frequently encountered in machine learning. We showed matching upper- and lower bounds on the complexity of this task in terms of a prescribed approximation error that are the product of the associated path complexities, i.e., the parameter tracking problems where all but one parameter are fixed. The path complexities for a fairly large class of problems had previously been shown to be in at least $\Omega\big(1/\sqrt{\varepsilon}\big)$ for a prescribed approximation error $\varepsilon > 0$. Under the assumption of this lower bound on the path complexities our lower bound construction shows that the complexity of the parameter space exploration problem grows exponentially with the number of parameters. Hence, parameter domain exploration with guarantees will only be practically feasible for low dimensional problems, if the domain does not possess an additional (dependence) structure. Identifying such structures could be an interesting direction of future research.

We have also turned the upper bound construction into an efficient and numerically robust algorithm for exploring low-dimensional parameter domains that adapts to the true problem complexity. Remarkably, the seemingly loose theoretical lower complexity bound is attained in both example problems that we have analyzed with an implementation of this algorithm.

Acknowledgments. This work has been supported by the DFG grant (GI-711/3-2).

References

Allgower, Eugene and Georg, Kurt. Continuation and path following. *Acta Numerica*, 2:1–64, 1993.

Bach, Francis R., Thibaux, Romain, and Jordan, Michael I. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Bergstra, James and Bengio, Yoshua. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

Bergstra, James, Bardenet, Rémi, Bengio, Yoshua, and Kégl, Balázs. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2546–2554, 2011.

Efron, Bradley, Hastie, Trevor, Johnstone, Iain, and Tibshirani, Robert. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Fan, Rong-En, Chen, Pai-Hsuen, and Lin, Chih-Jen. Working Set Selection Using Second Order Information for Training Support Vector Machines. *Journal of Machine Learning Research*, 6:1889–1918, 2005.

Friedman, Jerome, Hastie, Trevor, Höfling, Holger, and Tibshirani, Robert. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Regularized Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33 (1), 2010.

Gärtner, Bernd, Jaggi, Martin, and Maria, Clément. An Exponential Lower Bound on the Complexity of Regularization Paths. *Journal of Computational Geometry (JoCG)*, 3(1):168–195, 2012.

Giesen, Joachim, Jaggi, Martin, and Laue, Sören. Regularization Paths with Guarantees for Convex Semidefinite Optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 432–439, 2012a.

Giesen, Joachim, Jaggi, Martin, and Laue, Sören. Approximating parameterized convex optimization problems. *ACM Transactions on Algorithms*, 9(1):10, 2012b.

Giesen, Joachim, Müller, Jens K., Laue, Sören, and Swiercy, Sascha. Approximating Concavely Parameterized Optimization Problems. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2114–2122, 2012c.

Giesen, Joachim, Laue, Sören, and Wieschollek, Patrick. Robust and efficient kernel hyperparameter paths with guarantees. In *International Conference on Machine Learning (ICML)*, pp. 1296–1304, 2014.

Hastie, Trevor, Rosset, Saharon, Tibshirani, Robert, and Zhu, Ji. The Entire Regularization Path for the Support Vector Machine. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Hutter, Frank, Hoos, Holger H., and Leyton-Brown, Kevin. Sequential Model-Based Optimization for General Algorithm Configuration. In *Learning and Intelligent Optimization (LION)*, pp. 507–523, 2011.

Lin, Chih-Jen. LIBSVM Tools. Data sets available at `www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Mairal, Julien and Yu, Bin. Complexity analysis of the lasso regularization path. In *International Conference on Machine Learning (ICML)*, 2012.

Rosset, Saharon. Following curved regularized optimization solution paths. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Rosset, Saharon and Zhu, Ji. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030, 2007.

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2960–2968, 2012.

Tibshirani, Robert. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

Tibshirani, Ryan and Taylor, Jonathan. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3): 1335–1371, 2011.

Wang, Gang, Chen, Tao, Yeung, Dit-Yan, and Lochovsky, Frederick H. Solution path for semi-supervised classification with manifold regularization. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1124–1129, 2006a.

Wang, Gang, Yeung, Dit-Yan, and Lochovsky, Frederick H. Two-dimensional solution path for support vector regression. In *International Conference on Machine Learning (ICML)*, pp. 993–1000, 2006b.

Wang, Gang, Yeung, Dit-Yan, and Lochovsky, Frederick H. The Kernel Path in Kernelized LASSO. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 580–587, 2007a.

Wang, Gang, Yeung, Dit-Yan, and Lochovsky, Frederick H. A kernel path algorithm for support vector machines. In *International Conference on Machine Learning (ICML)*, pp. 951–958, 2007b.

Wang, Gang, Wang, Fei, Chen, Tao, Yeung, Dit-Yan, and Lochovsky, Frederick H. Solution Path for Manifold Regularized Semisupervised Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(2): 308–319, 2012.

Zhu, Ji, Rosset, Saharon, Hastie, Trevor, and Tibshirani, Robert. 1-norm Support Vector Machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

Zou, Hui and Hastie, Trevor. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, pp. 301–320, 2005.