Convex Relaxation Regression: Black-Box Optimization of Smooth Functions by Learning Their Convex Envelopes

Mohammad Gheshlaghi Azar
Rehabilitation Institute of Chicago
Northwestern University
Chicago, IL 60611

Eva L. Dyer
Rehabilitation Institute of Chicago
Northwestern University
Chicago, IL 60611

Konrad P. Körding
Rehabilitation Institute of Chicago
Northwestern University
Chicago, IL 60611

Abstract

Finding efficient and provable methods to solve non-convex optimization problems is an outstanding challenge in machine learning and optimization theory. A popular approach used to tackle non-convex problems is to use convex relaxation techniques to find a convex surrogate for the problem. Unfortunately, convex relaxations typically must be found on a problem-by-problem basis. Thus, providing a general-purpose strategy to estimate a convex relaxation would have a wide reaching impact. Here, we introduce *Convex Relaxation Regression* (CoRR), an approach for learning convex relaxations for a class of smooth functions. The idea behind our approach is to estimate the convex envelope of a function f by evaluating f at a set of T random points and then fitting a convex function to these function evaluations. We prove that with probability greater than $1-\delta$, the solution of our algorithm converges to the global optimizer of f with error $\mathcal{O}\big(\big(\frac{\log(1/\delta)}{T}\big)^{\alpha}\big)$ for some $\alpha > 0$. Our approach enables the use of convex optimization tools to solve non-convex optimization problems.

1 Introduction

Modern machine learning relies heavily on optimization techniques to extract information from large and noisy datasets (Friedman et al., 2001). Convex optimization methods are widely used in machine learning applications, due to fact that convex problems can be solved efficiently, often with a first order method such as gradient descent (Shalev-Shwartz and Ben-David, 2014; Sra et al., 2012; Boyd and Vandenberghe, 2004). A wide class of problems can be cast as convex optimization problems; however, many important learning problems, including binary classification with 0-1 loss, sparse and low-rank matrix recovery, and training multi-layer neural networks, are non-convex.

In many cases, non-convex optimization problems can be solved by first relaxing the problem: *convex relaxation* techniques find a convex function that approximates the original objective function (Tropp, 2006; Candès and Tao, 2010; Chandrasekaran et al., 2012). A convex relaxation is considered tight when it provides a tight lower bound to the original objective function. Examples of problems for which tight convex relaxations are known include binary classification (Cox, 1958), sparse and low-rank approximation (Tibshirani, 1996; Recht et al., 2010). The recent success of both sparse and low rank matrix recovery has demonstrated the power of convex relaxation for solving high-dimensional machine learning problems.

When a tight convex relaxation is known, then the underlying non-convex problem can often be solved by optimizing its convex surrogate in lieu of the non-convex problem. However, there are important classes of machine learning problems for which no such relaxation is known. These include a wide range of machine learning problems such as training deep neural nets, estimating latent variable models (mixture density models), optimal control, reinforcement learning, and hyper-parameter optimization. Thus, methods for finding convex relaxations of arbitrary non-convex functions would have wide reaching impacts throughout machine learning and the computational sciences.

Here we introduce a principled approach for black-box (zero-order) global optimization that is based on learning a convex relaxation to a non-convex function of interest (Sec. 3). To motivate our approach, consider the problem of estimating the convex envelope of the function f, i.e., the tightest convex lower bound of the function (Grotzinger, 1985; Falk, 1969; Kleibohm, 1967). In this case, we know that the envelope's minimum coincides with the minimum of the original non-convex function (Kleibohm, 1967). Unfortunately, finding the *exact* convex envelope of a non-convex function can be at least as hard as solving the original optimization problem. This is due to the fact that the problem of finding the convex envelope of a function is equivalent to the problem of computing its Legendre-Fenchel bi-conjugate (Rockafellar, 1997; Falk, 1969), which is in general as hard as optimizing f. Despite this result, we show that for a class of smooth (non-convex) functions, it is possible to accurately and efficiently *estimate* the convex envelope from a set of function evaluations.

The main idea behind our approach, *Convex Relaxation Regression* (CoRR), is to estimate the convex envelope of f and then optimize the resulting empirical convex envelope. We do this by solving a constrained ℓ_1 regression problem which estimates the convex envelope by a linear combination of a set of convex functions (basis vectors). As our approach only requires samples from the function, it can be used to solve optimization problems where gradient information is unknown. Whereas most methods for global optimization rely on local search strategies which find a new search direction to explore, CoRR takes a global perspective: it aims to form a global estimate of the function to "fill in the gaps" between samples. Thus CoRR provides an efficient strategy for global minimization through the use of convex optimization tools.

One of the main theoretical contributions of this work is the development of guarantees that CoRR can find accurate convex relaxations for a broad class of non-convex functions (Sec. 4). We prove in Thm. 1 that with probability greater than $1-\delta$, we can approximate the global minimizer with error of $\mathcal{O}\Big(\big(\frac{\log(1/\delta)}{T}\big)^{\alpha}\Big)$, where T is the number of function evaluations and $\alpha > 0$ depends upon the exponent of the Hölder-continuity bound on $f(x) - f^*$. This result assumes that the true convex envelope lies in the function class used to form a convex approximation. In Thm. 2, we extend this result for the case where the convex envelope is in the proximity of this set of functions. Our results may also translated to a bound with polynomial dependence on the dimension (Sec. 4.2.4).

The main contributions of this work are as follows. We introduce CoRR, a method for black-box optimization that learns a convex relaxation of a function from a set of random function evaluations (Sec. 3). Following this, we provide performance guarantees which show that as the number of function evaluations T grows, the error decreases polynomially in T (Sec. 4). In Thm. 1 we provide a general result for the case where the true convex envelope f_c lies in the function class $\mathcal{H}$ and extend this result to the approximate setting where $f_c \notin \mathcal{H}$ in Thm. 2. Finally, we study the performance of CoRR on several multi-modal test functions and compare it with a number of widely used approaches for global optimization (Sec. 5). These results suggest that CoRR can accurately find a tight convex lower bound for a wide class of non-convex functions.

2 Problem Setup

We now introduce relevant notation, setup our problem, and then provide background on global optimization of non-convex functions.

2.1 Preliminaries

Let n be a positive integer. For every $x \in \mathbb{R}^n$, its ℓ_2-norm is denoted by $\|x\|$, where $\|x\|^2 := \langle x, x\rangle$ and $\langle x, y\rangle$ denotes the inner product between two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$. We denote the ℓ_2 metric by d_2 and the set of ℓ_2-normed bounded vectors in $\mathbb{R}^n$ by $\mathcal{B}(\mathbb{R}^n)$, where for every $x \in \mathcal{B}(\mathbb{R}^n)$ we assume that there exists some finite scalar C such that $\|x\| < C$. Let $(\mathcal{X}, d)$ be a metric space, where $\mathcal{X} \in \mathcal{B}(\mathbb{R}^n)$ is a convex set of bounded vectors and $d(., x)$ is convex w.r.t. its first argument for every $x \in \mathcal{B}(\mathbb{R}^n)$.[1] We denote the set of all bounded functions on $\mathcal{X}$ by $\mathcal{B}(\mathcal{X}, \mathbb{R})$, such that for every $f \in \mathcal{B}(\mathcal{X}, \mathbb{R})$ and $x \in \mathcal{X}$ there exists some finite scalar $C > 0$ such that $|f(x)| \leq C$. Finally, we denote the set of all convex bounded functions on $\mathcal{X}$ by $\mathcal{C}(\mathcal{X}, \mathbb{R}) \subset \mathcal{B}(\mathcal{X}, \mathbb{R})$. Also for every $\mathcal{Y} \subseteq \mathcal{B}(\mathbb{R}^n)$, we denote the convex hull of $\mathcal{Y}$ by $\mathrm{conv}(\mathcal{Y})$. Let $\mathcal{B}(x_0, r)$ denote an open ball of radius r centered at x_0. Let $\mathbf{1}$ denote a vector of ones.

The convex envelope of function $f : \mathcal{X} \to \mathbb{R}$ is denoted by $f_c : \mathcal{X} \to \mathbb{R}$. Let $\widetilde{\mathcal{H}}$ be the set of all convex functions defined over $\mathcal{X}$ such that $h(x) \leq f(x)$ for all $x \in \mathcal{X}$. The function f_c is the convex envelope of f if for every $x \in \mathcal{X}$ **(a)** $f_c(x) \leq f(x)$, **(b)** for every $h \in \widetilde{\mathcal{H}}$ the inequality $h(x) \leq f_c(x)$ holds. Convex envelopes are also related to the concepts of the convex hull and the epigraph of a function. For every function $f : \mathcal{X} \to \mathbb{R}$ the epigraph is defined as $\mathrm{epi}f = \{(\xi, x) : \xi \geq f(x), x \in \mathcal{X}\}$. One can then show that the convex envelope of f is obtained by $f_c(x) = \inf\{\xi : (\xi, x) \in \mathrm{conv}(\mathrm{epi}f)\}, \ \forall x \in \mathcal{X}$.

In the sequel, we will generate a set of function evaluations from f by evaluating the function over i.i.d. samples from ρ, where ρ denotes a probability distribution on $\mathcal{X}$ such that $\rho(x) > 0$ for all $x \in \mathcal{X}$. In addition, we approximate the convex envelope using a function class $\mathcal{H}$ that contains a set of convex functions $h(\cdot; \theta) \in \mathcal{H}$ parametrized by $\theta \in \Theta \subseteq \mathcal{B}(\mathbb{R}^p)$. We also assume that every $h \in \mathcal{H}$ can be expressed as a linear combination of a set of basis $\phi : \mathcal{X} \to \mathcal{B}(\mathbb{R}^p)$, that is, $h(x; \theta) = \langle \theta, \phi(x)\rangle$ for every $h(\cdot; \theta) \in \mathcal{H}$ and $x \in \mathcal{X}$.

2.2 Black-box Global Optimization Setting

We consider a *black-box* (zero-order) global optimization setting, where we assume that we do not have access to

[1] This also implies that $d(x, .)$ is convex w.r.t. its second argument argument for every $x \in \mathcal{B}(\mathbb{R}^n)$ due to the fact that the metric d by definition is symmetric.

information about the gradient of the function that we want to optimize. More formally, let $\mathcal{F} \subseteq \mathcal{B}(\mathcal{X}, \mathbb{R})$ be a class of bounded functions, where the image of every $f \in \mathcal{F}$ is bounded by R and $\mathcal{X}$ is a convex set. We consider the problem of finding the global minimum of the function f,

$$f^* := \min_{x \in \mathcal{X}} f(x). \tag{1}$$

We denote the set of minimizers of f by $\mathcal{X}_f^* \subseteq \mathcal{X}$.

In the black-box setting, the optimizer has only access to the inputs and outputs of the function f. In this case, we assume that our optimization algorithm is provided with a set of input points $\widehat{\mathcal{X}} = \{x_1, x_2, \ldots, x_T\}$ in $\mathcal{X}$ and a sequence of outputs $[f]_{\widehat{\mathcal{X}}} = \{f(x_1), f(x_2), \ldots, f(x_T)\}$. Based upon this information, the goal is to find an estimate $\widehat{x} \in \mathcal{X}$, such that the error $f(\widehat{x}) - f^*$ becomes as small as possible.

2.3 Methods for Black-box Optimization

Standard tools that are used in convex optimization, cannot be readily applied to solve non-convex problems as they only converge to local minimizers of the function. Thus, effective global optimization approaches must have a mechanism to avoid getting trapped in local minima. In low-dimensional settings, performing an exhaustive grid search or drawing random samples from the function can be sufficient (Bergstra and Bengio, 2012). However, as the dimension grows, smarter methods for searching for the global minimizer are required.

Non-adaptive search strategies. A wide range of global optimization methods are build upon the idea of iteratively creating a deterministic set (pattern) of points at each iteration, evaluating the function over all points in the set, and selecting the point with the minimum value as the next seed for the following iteration (Hooke and Jeeves, 1961; Lewis and Torczon, 1999). Deterministic pattern search strategies can be extended by introducing some randomness into the pattern generation step. For instance, simulated annealing (Kirkpatrick et al., 1983) (SA) and genetic algorithms (Bäck, 1996) both use randomized search directions to determine the next place that they will search. The idea behind introducing some noise into the pattern, is that the method can jump out of local minima that deterministic pattern search methods can get stuck in. While many of these search methods work well in low dimensions, as the dimension of problem grows, these algorithms often become extremely slow due to the curse of dimensionality.

Adaptive and model-based search. In higher dimensions, adaptive and model-based search strategies can be used to further steer the optimizer in good search directions (Mockus et al., 1978; Hutter, 2009). For instance, recent results in Sequential Model-Based Global Optimization (SMBO) have shown that Gaussian processes are useful priors for global optimization (Mockus et al., 1978; Bergstra et al., 2011). In these settings, each search direction is driven by a model (Gaussian process) and updated based upon the local structure of the function. These techniques, while useful in low-dimension problems, become inefficient in high-dimensional settings.

Hierarchical search methods take a different approach in exploiting the structure of the data to find the global minimizer (Munos, 2014; Bubeck et al., 2011; Azar et al., 2014; Munos, 2011). The idea behind hierarchical search methods is to identify regions of the space with small function evaluations to sample further (exploitation), as well as generate new samples in unexplored regions (exploration). One can show that it is possible to find the global optimum with a finite number of function evaluations using hierarchical search; however, the number of samples needed to achieve a small error increases exponentially with the dimension. For this reason, hierarchical search methods are often not efficient for high-dimensional problems.

Graduated optimization. Graduated optimization methods (Blake and Zisserman, 1987; Yuille, 1989), are another class of methods for non-convex optimization which have received much attention in recent years (Chapelle and Wu, 2010; Dvijotham et al., 2014; Hazan et al., 2015; Mobahi and III, 2015). These methods work by locally smoothing the problem, descending along this smoothed objective, and then gradually sharpening the resolution to hone in on the true global minimizer. Recently Hazan et al. (2015) introduced a graduated optimization approach that can be applied in the black-box optimization setting. In this case, they prove that for a class of functions referred to as σ-nice functions, their approach is guaranteed to converge to an ε-accurate estimate of the global minimizer at a rate of $\mathcal{O}(n^2/\varepsilon^4)$. To the best of our knowledge, this result represents the state-of-the-art in terms of theoretical results for global black-box optimization.

3 Algorithm

In this section, we introduce *Convex Relaxation Regression* (CoRR), a black-box optimization approach for global minimization of a bounded function f.

3.1 Overview

The main idea behind our approach is to estimate the convex envelope f_c of a function and minimize this surrogate in place of our original function. The following result guarantees that the minimizer of f coincides with the minimizer of f_c.

Proposition 1 (Kleibohm 1967)**.** *Let f_c be the convex envelope of $f : \mathcal{X} \to \mathbb{R}$. Then (a) $\min_{x \in \mathcal{X}} f_c(x) = f^*$ and (b) $\mathcal{X}_f^* \subseteq \mathcal{X}_{f_c}^*$.*

This result suggests that one can find the minimizer of f by optimizing its convex envelope. Unfortunately, finding the exact convex envelope of a function is difficult in general. However, we will show that, for a certain class of functions, it is possible to estimate the convex envelope accurately from a set of function evaluations. Our aim is to estimate the convex envelope by fitting a convex function to these function evaluations.

The idea of fitting a convex approximation to samples from f is quite simple and intuitive. However, the best unconstrained convex fit to f does not necessarily coincide with f_c. Determining whether there exists a set of convex constraints under which the best convex fit to f coincides with f_c is an open problem. The following lemma, which is key to efficient optimization of f with CoRR, provides a solution. This lemma transforms our original non-convex optimization problem to a least-absolute-error regression problem with a convex constraint, which can be solved using convex optimization tools.

Lemma 1. *Let every $h \in \mathcal{H}$ and f be λ-Lipschitz for some $\lambda > 0$. Let $L(\theta) = \mathbb{E}[|h(x;\theta) - f(x)|]$ be the expected loss, where the expectation is taken with respect to the distribution ρ. Assume that there exists $\Theta_c \subseteq \Theta$ such that for every $\theta \in \Theta_c$, $h(x;\theta) = f_c(x)$ for all $x \in \mathcal{X}$. Consider the following optimization problem:*

$$\theta_\mu = \arg\min_{\theta\in\Theta} \; L(\theta) \;\; s.t. \;\; \mathbb{E}[h(x;\theta)] = \mu. \tag{2}$$

Then there exists a scalar $\mu \in [-R, R]$ for which $\theta_c \in \Theta_c$. In particular, $\theta_c \in \Theta_c$ when $\mu = \mathbb{E}(f_c(x))$.

The formal proof of this lemma is provided in the Supp. Materials. We prove this lemma by showing that for every $\theta \in \Theta$ where $\mathbb{E}[h(x;\theta)] = \mathbb{E}[f_c(x)]$, and for every $\theta_c \in \Theta_c$, the loss $L(\theta) \geq L(\theta_c)$. Equality is attained only when $\theta \in \Theta_c$. Thus, f_c is the only minimizer of $L(\theta)$ that satisfies the constraint $\mathbb{E}[h(x;\theta)] = \mathbb{E}[f_c(x)]$.

Optimizing μ. Lem. 1 implies that, for a certain choice of μ, Eqn. 2 provides us with the convex envelope f_c. However, finding the exact value of μ for which this result holds is difficult, as it requires knowledge of the envelope not available to the learner. Here we use an alternative approach to find μ which guarantees that the optimizer of $h(\cdot;\theta_\mu)$ lies in the set of true optimizers $\mathcal{X}_f^*$. Let x_μ denote the minimizer of $h(\cdot;\theta_\mu)$. We find a μ which minimizes $f(x_\mu)$:

$$\mu^* = \arg\min_{\mu\in[-R,R]} f(x_\mu). \tag{3}$$

Interestingly, one can show that x_{μ^*} lies in the set $\mathcal{X}_f^*$. To prove this, we use the fact that the minimizers of the convex envelope f_c and f coincide.This implies that $f(x_{\mu_c}) = f^*$, where $\mu_c := \mathbb{E}(f_c(x))$. It then follows that $f^* = f(x_{\mu_c}) \geq \min_{\mu\in[-R,R]} f(x_\mu) = f(x_{\mu^*})$. This combined with the fact that f^* is the minimizer of f implies that $f(x_{\mu^*}) = f^*$ and thus $x_{\mu^*} \in \mathcal{X}^*$.

3.2 Optimization Protocol

We now describe how we use the ideas presented in Sec. 3.1 to implement CoRR (see Alg. 1 for pseudocode). Our approach for black-box optimization requires two main ingredients: (1) samples from the function f and (2) a function class $\mathcal{H}$ from which we can form a convex approximation h. In practice, CoRR is initialized by first drawing two sets of T samples $\widehat{\mathcal{X}}_1$ and $\widehat{\mathcal{X}}_2$ from the domain $\mathcal{X} \subseteq \mathcal{B}(\mathbb{R}^n)$ and evaluating f over both of these sets. With these sets of function evaluations (samples) and a function class $\mathcal{H}$ in hand, our aim is to learn an approximation $h(x;\theta)$ to the convex envelope of f. Thus for a fixed value of μ, we solve the following constrained optimization problem (see the `OPT` procedure in Alg. 1):

$$\widehat{\theta}_c = \arg\min_{\theta\in\Theta} \widehat{\mathbb{E}}_1\big[|h(x;\theta) - f(x)|\big] \text{ s.t. } \widehat{\mathbb{E}}_2\big[h(x;\theta)\big] = \mu, \tag{4}$$

where the empirical expectation $\widehat{\mathbb{E}}_i[g(x)] := 1/T \sum_{x\in\widehat{\mathcal{X}}_i} g(x)$, for every $g \in \mathcal{B}(\mathcal{X}, \mathbb{R})$ and $i \in \{1, 2\}$. We provide pseudocode for optimizing Eqn. 4 in the `OPT` procedure of Alg. 1.

The optimization problem of Eqn. 4 is an empirical approximation of the optimization problem in Eqn. 2. However, unlike Eqn. 2, in which $L(\theta)$ is not easy to evaluate and optimize, the empirical loss can be optimized efficiently using standard convex optimization techniques. In addition, one can establish bounds on the error $|L(\widehat{\theta}_c) - L(\theta_c)|$ in terms of the sample size T using standard results from the literature on stochastic convex optimization (see, e.g., Thm. 1 in Shalev-Shwartz et al., 2009). Optimizing the empirical loss provides us with an accurate estimate of the convex envelope as the number of function evaluations increases.

The search for the best μ (Step 2 in Alg. 1) can be done by solving Eqn. 3. As μ is a scalar with known upper and lower bounds, we can employ a number of hyper-parameter search algorithms (Munos, 2011; Bergstra et al., 2011) to solve this 1D optimization problem. These algorithms guarantee fast convergence to the global minimizer in low dimensions and thus can be used to efficiently search for the solution to Eqn. 3. Let $\widehat{\mu}$ denote the final estimate of μ obtained in Step 2 of Alg. 1 and let $h(\cdot;\theta_{\widehat{\mu}})$ denote our final convex approximation to f_c. The final solution $\widehat{x}_{\widehat{\mu}}$ is then obtained by optimizing $h(\cdot;\theta_{\widehat{\mu}})$ (Step 2 of `OPT`).

To provide further insight into how CoRR works, we point the reader to Fig. 1. Here, we show examples of the convex surrogate obtained by `OPT` for different values of μ. We observe that as we vary μ, the minimum error is attained for $\mu \approx 0.47$. However, when we analytically compute the empirical expectation of convex envelope ($\widehat{E}_2[f_c(x)] = 0.33$) and use this value for μ, this produces a larger function evaluation. This may seem surprising, as we know that if we set $\mu = \mathbb{E}(f_c(x))$, then the solution of Eqn. 2 should provide us the exact convex envelope with the same opti-

Algorithm 1 Convex Relaxation Regression (CoRR)

Input: A black-box function f which returns a sample $f(x)$ when evaluated at a point x. The number of samples N to draw from f. A class $\mathcal{H} \subseteq \mathcal{B}(\mathcal{X}, \mathbb{R})$ of convex functions in $\mathcal{X}$ (parametrized by θ), a scalar R for which $\|f\|_\infty \leq R$, a sampling distribution ρ supported over $\mathcal{X}$.

Random function evaluations. Draw $2N$ i.i.d. samples according to the distribution ρ and partition them into two sets, $\widehat{\mathcal{X}} = \{\widehat{\mathcal{X}}_1, \widehat{\mathcal{X}}_2\}$. Generate samples $[f]_{\widehat{\mathcal{X}}_1}$ and $[f]_{\widehat{\mathcal{X}}_2}$, where $[f]_{\widehat{\mathcal{X}}_i} = \{f(x) : x \in \widehat{\mathcal{X}}_i\}, \quad i = \{1, 2\}\}$. Denote $[f]_{\widehat{\mathcal{X}}} = \{[f]_{\widehat{\mathcal{X}}_1}, [f]_{\widehat{\mathcal{X}}_2}\}$

Optimize for μ. Solve the 1D optimization problem

$$\widehat{\mu} = \underset{\mu \in [-R,R]}{\arg\min}\, f(\texttt{OPT}(\mu, [f]_{\widehat{\mathcal{X}}})),$$

Output: $\widehat{x}_{\widehat{\mu}}$ $=\texttt{OPT}(\widehat{\mu}, [f]_{\widehat{\mathcal{X}}})$.

Procedure $\texttt{OPT}(\mu, [f]_{\widehat{\mathcal{X}}})$

Estimate the convex envelope. Estimate $\widehat{f}_c = h(\cdot; \widehat{\theta}_\mu)$ by solving Eqn. 4.

Optimize the empirical convex envelope. Find an optimizer $\widehat{x}_\mu$ for $\widehat{f}_c$ by solving

$$\widehat{x}_\mu = \min_{x \in \mathcal{X}} \widehat{f}_c(x),$$

return $\widehat{x}_\mu$

mizer as f. This discrepancy can be explained by the approximation error introduced through solving the empirical version of Eqn. 2. This figure also highlights the stability of our approach for different values of μ. Our results suggest that our method is robust to the choice of μ, as a wide range of values of μ produce minimizers close to the true global minimum. Thus CoRR provides an accurate and robust approach for finding the global optimizer of f.

4 Theoretical Results

In this section, we provide our main theoretical results. We show that as the number of function evaluations T grows, the solution of CoRR converges to the global minimum of f with a polynomial rate. We also discuss the scalability of our result to high-dimensional settings.

4.1 Assumptions

We begin by introducing the assumptions required to state our results. The first assumption provides the necessary constraint on the candidate function class $\mathcal{H}$ and the set of all points in $\mathcal{X}$ that are minimizers for the function f.

Assumption 1 (Convexity). *Let $\mathcal{X}_f^*$ denote the set of minimizers of f. We assume that the following three convexity assumptions hold with regard to every $h(\cdot;\theta) \in \mathcal{H}$ and $\mathcal{X}_f^*$: (a) $h(x;\theta)$ is a convex function for all $x \in \mathcal{X}$, (b) h is a affine function of $\theta \in \Theta$ for all $x \in \mathcal{X}$, and (c) $\mathcal{X}_f^*$ is a convex set.*

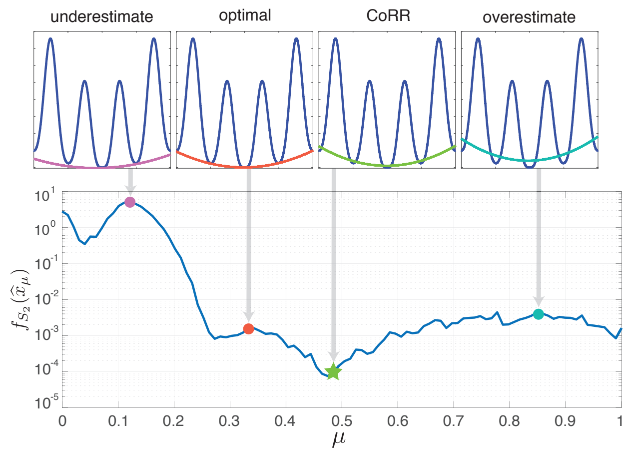

Figure 1: *Estimating the convex envelope of f with CoRR.* Here we demonstrate how CoRR learns a convex envelope by solving Eqn. 3. Along the top, we plot the test function f_{S_2} (see Sec. 5) and examples of the convex surrogates obtained for different values of μ. From left to right, we display the surrogate obtained for: an underestimate of μ, the empirical estimate of the convex envelope where $\mu \approx \widehat{\mathbb{E}}_2[f_c(x)]$, the result obtained by CoRR, and an overestimate of μ. Below, we display the value of the function f_{S_2} as we vary μ (solid blue).

Remark. Assumption 1c does not impose convexity on the function f. Rather, it requires that the set $\mathcal{X}_f^*$ is convex. This is needed to guarantee that both f_c and f have the same minimizers (see Prop. 1). Assumption 1c holds for a large class of non-convex functions. For instance, every continuous function with a unique minimizer satisfies this assumption (see, e.g., our example functions in Sect. 5).

Assumption 2 establishes the necessary smoothness assumption on the function f and the function class $\mathcal{H}$.

Assumption 2 (Lipschitz continuity). *We assume that f and h are Lipschitz continuous. That is for every $(x_1, x_2) \in \mathcal{X}^2$ we have that $|f(x_1) - f(x_2)| \leq d(x_1, x_2)$. Also for every $x \in \mathcal{X}$ and $(\theta_1, \theta_2) \in \Theta^2$ we have that $|h(x;\theta_1) - h(x;\theta_2)| \leq U d_2(\theta_1, \theta_2)$. We also assume that every $h \in \mathcal{H}$ is λ-Lipschitz on $\mathcal{X}$ w.r.t. the metric d for some $\lambda > 0$.*

We show that the optimization problem of Eqn. 1 provides us with the convex envelope f_c when the candidate class $\mathcal{H}$ contains f_c (see Lem. 1). The following assumption formalizes this condition.

Assumption 3 (Capacity of $\mathcal{H}$). *We assume that $f_c \in \mathcal{H}$, that is, there exist some $h \in \mathcal{H}$ and $\Theta \subseteq \Theta_c$ such that*

$h(x;\theta) = f_c(x)$ *for every* $x \in \mathcal{X}$ *and* $\theta \in \Theta_c$.

We also require that the following Hölder-type error bounds hold for the distances of our empirical estimates $\widehat{x}$ and $\widehat{\theta}$ from $\mathcal{X}_f^*$ and Θ_c, respectively.

Assumption 4 (Hölder-type error bounds)**.** *Let* $\Theta^e := \{\theta | \theta \in \Theta, \mathbb{E}(h(x;\theta)) = f_c(x)\}$. *Also denote* $L^* := \min_{\theta\in\Theta^e} L(\theta)$. *We assume that there exists some finite positive scalars* γ_1, γ_2, β_1 *and* β_2 *such that for every* $x \in \mathcal{X}$ *and* $\theta \in \Theta^e$*: (a)* $f(x) - f^* \geq \gamma_1 d(x, \mathcal{X}_f^*)^{1/\beta_1}$. *(b)* $L(\theta) - L^* \geq \gamma_2 d_2(\theta, \Theta_c)^{1/\beta_2}$.

Assumption 4 implies that whenever the error terms $f(x)-f^*$ and $L(\theta) - L^*$ are small, the distances $d(x, \mathcal{X}_f^*)$ and $d_2(\theta, \Theta_c)$ are small as well. To see why Assumption 4 is required for the analysis of CoRR, we note that the combination of Assumption 4 with Assumption 2 leads to the following *local* bi-Hölder inequalities for every $x \in \mathcal{X}$ and $\theta \in \Theta^e$:

$$\begin{aligned} \gamma_1 d(x, \mathcal{X}_f^*)^{1/\beta_1} &\leq f(x) - f^* \leq d(x, \mathcal{X}_f^*) \\ \gamma_2 d_2(\theta, \Theta_c)^{1/\beta_2} &\leq L(\theta) - L^* \leq U d_2(\theta, \Theta_c) \end{aligned} \tag{5}$$

These inequalities determine the behavior of function f and L around their minimums as they establish upper and lower bounds on the errors $f(x) - f^*$ and $L(\theta) - L^*$. Essentially, Eqn. 5 implies that there is a direct relationship between $d(x, \mathcal{X}_f^*)$ ($d_2(\theta, \Theta_c)$) and $f(x)-f^*$ ($L(\theta)-L(\Theta_c)$). Thus, bounds on $d(x, \mathcal{X}_f^*)$ and $d_2(\theta, \Theta_c)$, respectively, imply bounds on $f(x)-f^*$ and $L(\theta)-L(\Theta_c)$ and vice versa. These bi-directional bounds are needed due to the fact that CoRR doest not directly optimize the function. Instead it optimizes the surrogate loss $L(\theta)$ to find the convex envelope and then it optimizes this empirical convex envelope to estimate the global minima. This implies that the standard result of optimization theory can only be applied to bound the error $L(\widehat{\theta}) - L^*$. The inequalities of Eqn. 5 are then required to convert the bound on $L(\widehat{\theta}) - L^*$ to a bound on $f(\widehat{x}_{\widehat{\mu}}) - f^*$, which ensures that the solution of CoRR converges to a global minimum as $L(\widehat{\theta}) - L^* \to 0$.

It is noteworthy that global error bounds such as those in Assumption 4 have been extensively analyzed in the literature of approximation theory and variational analysis (see, e.g., Azé, 2003; Corvellec and Motreanu, 2008; Azé and Corvellec, 2004; Fabian et al., 2010). Much of this body of work can be applied to study convex functions such as $L(\theta)$, where one can make use of the basic properties of convex functions to prove lower bounds on $L(\theta) - L^*$ in terms of the distance between θ and Θ_c (see, e.g., Thm. 1.16 in Azé, 2003). While these results are useful to further study the class of functions that satisfy Assumption 4, providing a direct link between these results and the error bounds of Assumption 4 is outside the scope of this paper.

Assumptions 3-4 can not be applied directly when $f_c \notin \mathcal{H}$. When $f_c \notin \mathcal{H}$, we make use of the following generalized version of these assumptions. We first consider a relaxed version of Assumption 3, which assumes that f_c can be approximated by some $h \in \mathcal{H}$.

Assumption 5 (υ-approachability of f_c by $\mathcal{H}$)**.** *Let* υ *be a positive scalar. Define the distance between the function class* $\mathcal{H}$ *and* f_c *as* $\mathrm{dist}(f_c, \mathcal{H}) := \inf_{h\in\mathcal{H}} \mathbb{E}[|h(x;\theta) - f_c(x)|]$, *where the expectation is taken w.r.t. the distribution* ρ. *We then assume that the following inequality holds:* $\mathrm{dist}(f_c, \mathcal{H}) \leq \upsilon$.

The next assumption generalizes Assumption 4b to the case where $f_c \notin \mathcal{H}$:

Assumption 6. *Let* $\widetilde{p}$ *be a positive scalar. Assume that there exists a class of convex functions* $\widetilde{\mathcal{H}} \subseteq \mathcal{C}(\mathcal{X}, \mathbb{R})$ *parametrized by* $\theta \in \widetilde{\Theta} \subset \mathcal{B}(\mathbb{R}^{\widetilde{p}})$ *such that: (a)* $f_c \in \widetilde{\mathcal{H}}$, *(b) every* $h \in \widetilde{\mathcal{H}}$ *is linear in* θ *and (c)* $\mathcal{H} \subseteq \widetilde{\mathcal{H}}$. *Let* $\Theta_c \subseteq \widetilde{\Theta}$ *be the set of parameters for which* $h(x;\theta) = f_c(x)$ *for every* $x \in \mathcal{X}$ *and* $\theta \in \Theta_c$. *Also define* $\widetilde{\Theta}_e := \{\theta | \theta \in \widetilde{\Theta}, \mathbb{E}(h(x;\theta)) = f_c(x)\}$. *We assume that there exists some finite positive scalars* γ_2 *and* β_2 *such that for every* $x \in \mathcal{X}$ *and* $\theta \in \widetilde{\Theta}_e$

$$L(\theta) - L^* \geq \gamma_2 d_2(\theta, \widetilde{\Theta}_c)^{1/\beta_2}.$$

Intuitively speaking, Assumption 6 implies that the function class $\mathcal{H}$ is a subset of a larger unknown function class $\widetilde{\mathcal{H}}$ which satisfies the global error bound of Assumption 4b. Note that we do not require access to the class $\widetilde{\mathcal{H}}$, but we need that such a function class exists.

4.2 Performance Guarantees

We now present the two main theoretical results of our work and provide sketches of their proofs (the complete proofs of our results is provided in the Supp. Material).

4.2.1 Exact Setting

Our first result considers the case where the convex envelope $f_c \in \mathcal{H}$. In this case, we can guarantee that as the number of function evaluations grows, the solution of Alg. 1 converges to the optimal solution with a polynomial rate.

Theorem 1. *Let* δ *be a positive scalar. Let Assumptions 1, 2 ,3, and 4 hold. Then Alg. 1 returns* $\widehat{x}$ *such that with probability* $1 - \delta$

$$f(\widehat{x}) - f^* = \mathcal{O}\left[\xi_s \left(\frac{\log(1/\delta)}{T}\right)^{\beta_1\beta_2/2}\right],$$

where the smoothness coefficient $\xi_s := (\frac{1}{\gamma_1})^{\beta_2} (\frac{1}{\gamma_2})^{\beta_1\beta_2} U^{(1+\beta_2)\beta_1} (RB)^{\beta_2\beta_1}$.

Sketch of proof. To prove this result, we first prove bound on the error $L(\widehat{\theta}) - \min_{\theta\in\Theta_e} L(\theta)$ for which we rely on standard results from stochastic convex optimization. This combined with the result of Lem. 1 leads to a bound on $L(\widehat{\theta}) - L^*$. The bound on $L(\widehat{\theta}) - L^*$ combined with Assumption 4 translates to a bound on $d(\widehat{x}, \mathcal{X}_f^*)$. The result then follows by applying the Lipschitz continuity assumption (Assumption 2). □

Thm. 1 guarantees that as the number of function evaluations T grows, the solution of CoRR converges to f^* with a polynomial rate. The order of polynomial depends on the constants β_1 and β_2. The following corollary, which is an immediate result of Thm. 1, quantifies the number of function evaluations T needed to achieve an ε-optimal solution.

Corollary 1. *Let Assumptions 1, 2, 3, and 4 hold. Let ε and δ be some positive scalars. Then Alg. 1 needs $T = (\frac{\xi_s}{\varepsilon})^{2/(\beta_1\beta_2)} \log(1/\delta)$ function evaluations to return $\widehat{x}$ such that with probability $1-\delta$, $f(\widehat{x}) - f^* \leq \varepsilon$.*

This result implies that one can achieve an ε-accurate approximation of the global optimizer with CoRR with a polynomial number of function evaluations.

4.2.2 Approximate Setting

Thm. 1 relies on the assumption that the convex envelope f_c lies in the function class $\mathcal{H}$. However, in general, there is no guarantee that f_c belongs to $\mathcal{H}$. When the convex envelope $f_c \notin \mathcal{H}$, the result of Thm. 1 cannot be applied. However, one may expect that Alg. 1 still may find a close approximation of the global minimum as long as the distance between f_c and $\mathcal{H}$ is small. To prove that CoRR finds a near optimal solution in this case, we must show that $f(\widehat{x}) - f^*$ remains small when the distance between f_c and $\mathcal{H}$ is small. We now generalize Thm. 1 to the setting where the convex envelope f_c does not lie in $\mathcal{H}$ but is close to it.

Theorem 2. *Let Assumptions 1, 2, 5, and 6 hold. Then Alg. 1 returns $\widehat{x}$ such that for every $\zeta > 0$ with probability (w.p.) $1-\delta$*

$$f(\widehat{x}) - f^* = \mathcal{O}\left[\xi_s\left(\sqrt{\frac{\log(1/\delta)}{T}} + \zeta + \upsilon\right)^{\beta_1\beta_2}\right].$$

Sketch of proof. To prove this result, we rely on standard results from stochastic convex optimization to first prove a bound on the error $L(\widehat{\theta}) - \min_{\theta\in\Theta^c} L(\theta)$ when we set μ the empirical mean of the convex envelope. We then make use of Assumption 5 as well as Lem. 1 to transform this bound to a bound on $L(\widehat{\theta}) - L^*$. The bound on $f(\widehat{x}) - f^*$ then follows by combining this result with Assumptions 2 and 6. □

4.2.3 Approximation Error υ vs. Complexity of $\mathcal{H}$

From function approximation theory, it is known that for a sufficiently smooth function g, one can achieve an υ-accurate approximation of g by a linear combination of $p = \mathcal{O}(n/\upsilon)$ bases (Mhaskar, 1996; Girosi and Anzellotti, 1992). These results imply that one can make the error υ in Thm. 2 arbitrary small by increasing the complexity of function class $\mathcal{H}$ (i.e., increasing the number of convex bases p). Similar *shape preserving* results have been established for the case when the function and bases are both convex (see, e.g., Gal, 2010; Konovalový et al., 2010; Shvedov, 1981) under some mild assumptions on g. In particular, Konovalový et al. (2010) have proven that for a rather general class of $\mathcal{H}$, the approximation error between a convex function g and class $\mathcal{H}$, can be bounded in terms of the approximation error between g and $\mathcal{H}$ when no convexity constraint is imposed on $\mathcal{H}$. This implies that existing results in the approximation theory literature can be used to bound the approximation error υ in terms of the complexity of function class $\mathcal{H}$.

4.2.4 Dependence on Dimension

The results of Thm. 1 and Thm. 2 have no explicit dependence on the dimension n. However, the Lipschitz constant U can, in the worst-case scenario, be of $\mathcal{O}(\sqrt{p})$ (due to the Cauchy-Schwarz inequality). On the other hand to achieve an approximation error of υ the number of bases p needs be of $\mathcal{O}(n/\upsilon)$ (see Sect. 4.2.3). When we plug this result in the bound of Thm. 2, this leads to a dependency of $\mathcal{O}(n^{(1+\beta_2)\beta_1/2})$ on the dimension n due to the Lipschitz constant U. In the special case where $\beta_2 = \beta_1 = 1$, i.e., when the error bounds of Assumption 4 are linear, the dependency on n becomes linear. The linear dependency on n in this case matches the results of the black-box (zero-order) convex optimization (see, e.g., Duchi et al., 2015).

5 Numerical Results

In this section, we evaluate the performance of CoRR on several multi-dimensional test functions used for benchmarking non-convex optimization methods (Jamil and Yang, 2013).

Evaluation setup. Here we study CoRR's effectiveness in finding the global minimizer of the following test functions (Fig. 2a). We assume that all functions are supported over $\mathcal{X} = \mathcal{B}(0,2) \subseteq \mathbb{R}^n$, and otherwise rescale them to lie within this set. **(S1)** Salomon function: $f_S(x) = 1 - \cos(2\pi\|x\|) + 0.5\|x\|$. **(S2)** Squared Salomon: $f_{S_2}(x) = 0.1 f_S(x)^2$. **(SL)** Salomon and Langerman combination: $f_{SL}(x) = f_S(x) + f_L(x)\ \forall x \in \mathcal{B}(0,10) \cap \mathcal{B}(0,0.2)$ and $f_{SL}(x) = 0$, otherwise (before rescaling the domain). **(L)** Langerman function: $f_L(x) = -\exp(\|x-\alpha\|_2^2/\pi)\cos(\pi\|x-\alpha\|_2^2) + 1$, $\forall x \in \mathcal{B}(0,5)$ (before rescaling the domain). **(G)** The Griewank function: $f_G(x) = 0.1\big[1 + \frac{1}{4000}\sum_{i=1}^N x(i)^2 - \prod_{i=1}^N \frac{\cos(x)}{\sqrt{i}}\big]$, $\forall x \in \mathcal{B}(0,200)$ (before rescaling the domain). All of these functions have their minimum at the origin, except for the Langerman function which has its minimum at $x^* = c\mathbf{1}$ for $c = 0.5$.

All of the aforementioned functions exhibit some amount of global structure for which the convex envelope can be approximated by a quadratic basis (Fig. 2a). We thus use a quadratic basis to construct our function class $\mathcal{H}$. The basis functions $h(x;\theta) \in \mathcal{H}$ are parameterized by a vector of coefficients $\theta = [\theta_1, \theta_2, \theta_3]$, and can be written as $h(x;\theta) = \langle \theta_1, x^2 \rangle + \langle \theta_2, x \rangle + \theta_3$. Thus, the number of parameters that we must estimate to find a convex approximation h equals $2n + 1$ (we drop the cross terms in our construction of the quadratic class). In practice, we impose a non-negativity constraint on all entries of the vector θ_1 to ensure that our approximation is convex.

Summary of results. To understand the difficulty of finding the minimizers for the test functions above, we compute the error $f(\widehat{x}) - f^*$ as we increase the number of function evaluations. Here, we show each of our five test functions (Fig. 2a) and their average scaling behavior in one dimension (Fig. 2b), where the error is averaged over 100 trials. We observe that CoRR quickly converges for all five test functions, with varying convergence rates. We observe the smallest initial error (for only 20 samples) for f_{SL} and the highest error for f_S. In addition, f_L achieves nearly perfect reconstruction of the global minimum after only 200 samples. The good scaling properties of f_L and f_{SL} is likely due the the fact that both of these functions have a wide basin around their global minimizer. This result provides nice insight into the scaling of CoRR in low dimensions.

Next, we study the approximation error as we vary the sample size and dimension for the Salomon function f_S (Fig. 2c-d). Just as our theory suggests, there is a clear dependence between the dimension and number of samples required to obtain small error. In Fig. 2c, we display the scaling behavior of CoRR as a function of both dimension and number of function evaluations T. In all of the tested dimensions, we obtain an error smaller than $1e^{-5}$ when we draw one million samples. In Fig. 2d, we compare the performance of CoRR (for fixed number of evaluations T) as we vary the dimension. In contrast, the quasi-Newton (QN) method and hybrid simulated annealing (SA) method (Hedar and Fukushima, 2004) recover the global minimizer for low dimensions but fail in dimensions greater than ten.[2] We posit that this is due to the fact the minimizer of the Salomon function lies at the center of its domain and as the dimension of the problem grows, drawing an initialization point (for QN) that is close to the global minimizer becomes extremely difficult.

[2]These methods are selected from a long list of candidates in MATLAB's global optimization toolbox. We report results for the methods that gave the best results for our test functions.

6 Discussion and Future Work

This paper introduced CoRR, an approach for learning a convex relaxation for a wide class of non-convex functions. The idea behind CoRR is to find an empirical estimate of the convex envelope of a function from a set of function evaluations. We demonstrate that CoRR is an efficient strategy for global optimization, both in theory and in practice. In particular, we provide theoretical results (Sec. 4) which show that CoRR is guaranteed to produce a convergent estimate of the convex envelope that exhibits polynomial dependence on the dimension. In numerical experiments (Sec. 5), we showed that CoRR provides accurate approximations to the global minimizer of multiple test functions and appears to scale well with dimension.

Our current instantiation of CoRR finds a convex surrogate for f based upon a set of samples that are drawn at random at the onset of the algorithm. In our evaluations, we draw i.i.d. samples from a uniform distribution over $\mathcal{X}$. However, the choice of the sampling distribution ρ has a significant impact on our estimation procedure. As such, selecting samples in an intelligent manner would significantly reduce the number of samples required to obtain an accurate estimate. A natural extension of CoRR is to the case where we can iteratively refine our distribution ρ based upon the output of the algorithm at previous steps.

An important factor in the success of our algorithm is the basis that we use to form our approximation. As discussed in Sec. 4.2.3, we know that a polynomial basis can be used to form a convex approximation to any convex function (Gal, 2010). However, finding a concise representation of the convex envelope using high-degree polynomials is not an easy task. Thus finding other well-suited bases for this approximation, such as the exponential basis, may improve the efficiency of CoRR by reducing the number of bases required. While outside the scope of this paper, exploring the use of constrained dictionary learning methods (Yaghoobi et al., 2009) for finding a good basis for our fitting procedure, is an interesting line for future work.

In our experiments, we observe that CoRR typically provides a good approximation to the global minimizer. However, in most cases, we do not obtain machine precision (like QN for low dimensions). Thus, we can combine CoRR with a local search method like QN by using the solution of CoRR as an initialization point for the local search. When using this hybrid approach, we obtain perfect reconstruction of the global minimum for the Salomon function for all of the dimensions we tested (Fig. 2d). This suggests that, as long the function does not fluctuate too rapidly around its global minimum (Asm. 2), CoRR can be coupled with other local search methods to quickly converge to the absolute global minimizer.

The key innovation behind CoRR is that one can efficiently

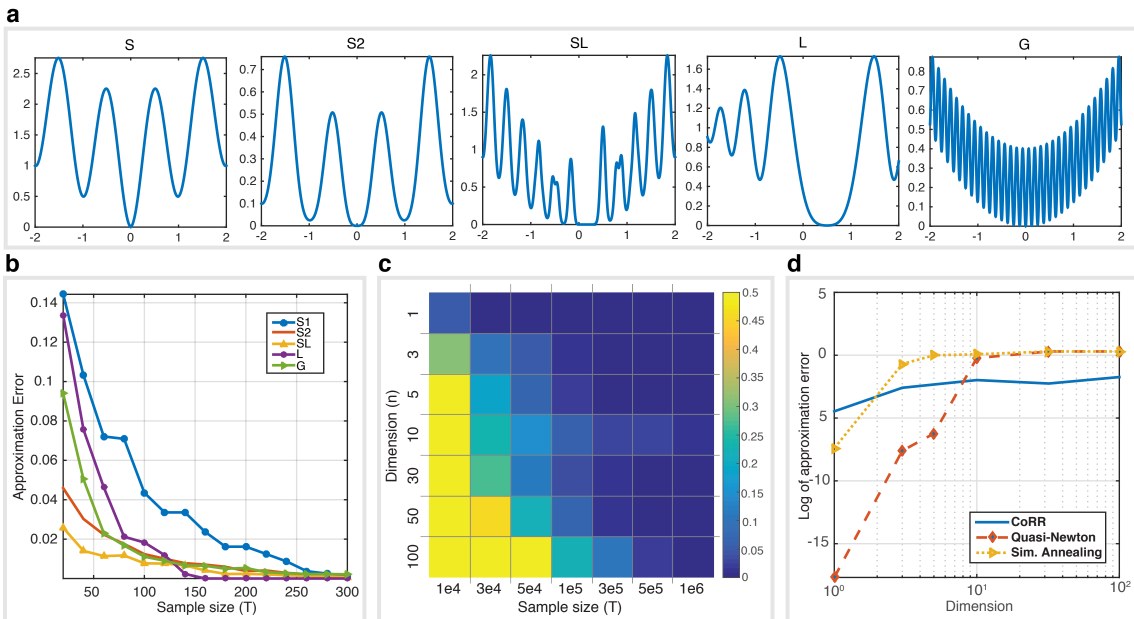

Figure 2: *Scaling behavior and performance of CoRR.* Along the top row in (a), we plot all five test functions studied in this paper. In (b), we display the mean approximation error between $f(\widehat{x}) - f^*$ as a function of the number of function evaluations T for all test functions in 1D. In (c), we display the mean approximation error as a function of the dimension and number of samples for the Salomon function. In (d), we compare CoRR's approximation error with other approaches for non-convex optimization, as we vary the dimension.

approximate the convex envelope of a non-convex function by solving a constrained regression problem which balances the approximation error with a constraint on the empirical expectation of the estimated convex surrogate. While our method could be improved by using a smart and adaptive sampling strategy, this paper provides a new way of thinking about how to relax non-convex problems. As such, our approach opens up the possibility of using the myriad of existing tools and solvers for convex optimization problems to efficiently solve non-convex problems.

References

Azar, M. G., Lazaric, A., and Brunskill, E. (2014). Stochastic optimization of a locally smooth function under correlated bandit feedback. In *ICML*.

Azé, D. (2003). A survey on error bounds for lower semicontinuous functions. In *ESAIM: ProcS*, volume 13, pages 1–17. EDP Sciences.

Azé, D. and Corvellec, J.-N. (2004). Characterizations of error bounds for lower semicontinuous functions on metric spaces. *ESAIM: Control, Optimisation and Calculus of Variations*, 10:409–425.

Bäck, T. (1996). *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(1):281–305.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *NIPS*, pages 2546–2554.

Blake, A. and Zisserman, A. (1987). *Visual reconstruction*, volume 2. MIT Press Cambridge.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvari, C. (2011). X-armed bandits. *J. Mach. Learn. Res.*, 12:1655–1695.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Found Comput Math*, 12(6):805–849.

Chapelle, O. and Wu, M. (2010). Gradient descent optimization of smoothed information retrieval metrics. *Inform Retrieval*, 13(3):216–235.

Corvellec, J.-N. and Motreanu, V. V. (2008). Nonlinear error bounds for lower semicontinuous functions on metric spaces. *Math Program*, 114(2):291–319.

Cox, D. R. (1958). The regression analysis of binary sequences. *J R Stat Soc*, pages 215–242.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Inf. Theory*, 61(5):2788–2806.

Dvijotham, K., Fazel, M., and Todorov, E. (2014). Universal convexification via risk-aversion. In *UAI*, pages 162–171.

Fabian, M. J., Henrion, R., Kruger, A. Y., and Outrata, J. V. (2010). Error bounds: Necessary and sufficient conditions. *Set-Valued and Variational Analysis*, 18(2):121–149.

Falk, J. E. (1969). Lagrange multipliers and nonconvex programs. *SIAM J Control Optim*, 7(4):534–545.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer.

Gal, S. (2010). *Shape-preserving approximation by real and complex polynomials*. Springer.

Girosi, F. and Anzellotti, G. (1992). Convergence rates of approximation by translates. Technical report, Massachusetts Inst. of Tech. Cambridge Artificial Intelligence Lab.

Grotzinger, S. J. (1985). Supports and convex envelopes. *Math Program*, 31(3):339–347.

Hazan, E., Levy, K. Y., and Shalev-Swartz, S. (2015). On graduated optimization for stochastic non-convex problems. *arXiv:1503.03712 [cs.LG]*.

Hedar, A.-R. and Fukushima, M. (2004). Heuristic pattern search and its hybridization with simulated annealing for nonlinear global optimization. *Optim Method and Softw*, 19(3-4):291–308.

Hooke, R. and Jeeves, T. A. (1961). "Direct search" solution of numerical and statistical problems. *J ACM*, 8(2):212–229.

Hutter, F. (2009). Automated configuration of algorithms for solving hard computational problems. *University of British Columbia*.

Jamil, M. and Yang, X.-S. (2013). A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2):150–194.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., et al. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kleibohm, K. (1967). Bemerkungen zum problem der nichtkonvexen programmierung. *Unternehmensforschung*, 11(1):49–60.

Konovalový, V., Kopotun, K., and Maiorov, V. (2010). Convex polynomial and ridge approximation of Lipschitz functions in $\mathbf{R}^d$. *Rocky Mountains Journal of Mathematics*, 40(3).

Lewis, R. M. and Torczon, V. (1999). Pattern search algorithms for bound constrained minimization. *SIAM J Optimiz*, 9(4):1082–1099.

Mhaskar, H. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput*, 8(1):164–177.

Mobahi, H. and III, J. W. F. (2015). A theoretical analysis of the optimization by gaussian continuation. In *AAAI*.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2.

Munos, R. (2011). Optimistic optimization of deterministic functions without the knowledge of its smoothness. In *NIPS*.

Munos, R. (2014). From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–129.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501.

Rockafellar, R. T. (1997). *Convex analysis*. Princeton University Press.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009). Stochastic convex optimization. In *COLT*.

Shvedov, A. S. (1981). Orders of coapproximation of functions by algebraic polynomials. *Mathematical Notes*, 29(1):63–70.

Sra, S., Nowozin, S., and Wright, S. J. (2012). *Optimization for machine learning*. MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc*, pages 267–288.

Tropp, J. A. (2006). Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Process*, 86(3):589–602.

Yaghoobi, M., Blumensath, T., and Davies, M. E. (2009). Dictionary learning for sparse approximations with the majorization method. *IEEE Trans. Signal Process.*, 57(6):2178–2191.

Yuille, A. (1989). Energy functions for early vision and analog networks. *Biol Cybern*, 61(2):115–123.