Least Squares Estimation of Weakly Convex Functions

Sun Sun **Yaoliang Yu**
University of Waterloo, Canada
{sun.sun, yaoliang.yu}@uwaterloo.ca

Abstract

Function estimation under shape restrictions, such as convexity, has many practical applications and has drawn a lot of recent interests. In this work we argue that convexity, as a global property, is too strict and prone to outliers. Instead, we propose to use weakly convex functions as a simple alternative to quantify "approximate convexity"—a notion that is perhaps more relevant in practice. We prove that, unlike convex functions, weakly convex functions can exactly interpolate any finite dataset and they are universal approximators. Through regularizing the modulus of convexity, we show that weakly convex functions can be efficiently estimated both statistically and algorithmically, requiring minimal modifications to existing algorithms and theory for estimating convex functions. Our numerical experiments confirm the class of weakly convex functions as another competitive alternative for nonparametric estimation.

1 Introduction

Much of machine learning is about estimating an unknown function f from limited data. Key to the function estimation problem is our *a priori* knowledge about the function f, without which the problem is clearly hopeless. For example, a standard assumption in machine learning is that f can be described by a finite number of parameters (or weights). Estimating the function can thus be reduced to estimating its parameters, a setting known as parametric estimation. Widely used methods such as (linear) support vector machines, (linear) logistic regression, Lasso, deep neural networks, etc., all belong to this category. Parametric methods are popular because they are conceptually simple, their optimization algorithms scale well, and because their statistical properties are relatively well-understood.

Nonparametric methods, on the other hand, put less stringent assumptions on the underlying unknown function. The candidate functions in a nonparametric method cannot be described by finitely many parameters, i.e., they consist of an infinite dimensional space. Thus, nonparametric methods are more flexible when we have imprecise (often qualitative instead of quantitative) *a priori* information. Nonparametric function estimation can be roughly divided into two categories: those with *smoothness* restriction and those with *shape* restriction. Splines and kernels are typical examples for the former, see e.g. [Györfi et al., 2002], whereas the latter, starting from the pioneering work in [Hildreth, 1954; Ayer et al., 1955], has drawn a lot of recent interest as well [Hannah & Dunson, 2012; Hannah et al., 2014; Balázs et al., 2015; Yin & Yu, 2017; Lim & Glynn, 2012; Seijo & Sen, 2011; Xu et al., 2016].

The practical relevance of estimating shape-restricted functions such as convex functions has been well articulated in econometrics [Varian, 1982, 1984], geometric programming [Magnani & Boyd, 2009; Hannah & Dunson, 2012], operations research [Shapiro et al., 2009], and finance [Grenander, 1956; Hannah et al., 2014], just to name a few. For example, the optimal value function of a partially observable Markov decision process is convex [Sondik, 1978]. Convexity is also important from an operational perspective: if we were to find the minimum of an estimated function, see Hannah et al. [2014] for applications, then having convexity is certainly useful for *efficiently* finding the global minimum.

Convexity, however, is a global property. As we show in §2, even changing a single data point can completely destroy convexity and arbitrarily mislead the least-squares convex estimate. In this work we consider estimating an unknown regression function under *weak* convexity conditions, which belongs to the above nonparametric *shape* restricted category. *Our main contribution is to point out a surprisingly simple way to quantify (approximate) convexity, along with its algorithmic and*

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

statistical consequences. The advantage of our new class of weakly convex functions is: (1) it is universal, unlike the existing class of convex functions that is being used in the literature; (2) it can enforce approximate convexity, unlike existing methods based on smoothness assumptions. Moreover, we can achieve this advantage by modifying the existing convex estimator in a minimal way, retaining most of its appealing properties and adding universality. Each of these properties ideally should be possessed by any sensible method and yet *no existing one* could achieve them all.

In §3 we recall the definition of weakly convex functions, and we discuss a few key properties of them. In particular, it is known [Rockafellar, 1982; Shapiro & Yomdin, 1981] that on a compact domain weakly convex functions are simply piecewise quadratic (with a bounded Hessian and possibly infinitely many pieces). These properties help us identify weakly convex functions in applications. In fact, many (if not all) functions used in practice are weakly convex (but not necessarily convex or concave). Then, in §4 we prove two surprising facts about weakly convex functions: (a) they can *exactly* interpolate any finite dataset, hence bound to overfit, no matter how large the sample size is; (b) they can approximate any continuous function on a compact domain arbitrarily well. Note that *both properties do not hold for convex (or concave) functions.* The latter property, known as universal approximators, strongly motivates us to consider the class of weakly convex functions in nonparametric *shape-restricted* estimation, if we can address the first overfitting property.

Indeed, in §5 we show that the modulus of convexity can be used as a natural regularizer to alleviate the overfitting problem. The resulting optimization problem turns out to be quite convenient: it requires very minimal modification to existing algorithms for estimating convex functions. Moreover, almost all statistical properties of the least-squares convex estimate can be easily carried over to the much larger class of weakly convex functions. In §6 we discuss two appealing properties of the weakly convex estimates: adaptation to the underlying geometry and amenable to efficient minimization, and we show how to adapt existing scalable algorithms. Finally, we validate our results in §7 through numerical experiments, and we conclude in §8.

2 Motivating Example

In this section, we consider function estimation on a synthetic dataset. We show that the convexity constraint is too strict and is prone to outliers, which motivates our consideration of weakly convex functions.

Let us consider the following simple dataset $\mathbb{D}$ in $\mathbb{R}^2$: For $i = 1, \ldots, 2n+1$, let $x_i = -n - 1 + i$ and $y_i \equiv 0$. Then, we perturb the point in the middle, i.e., $y_{n+1} = t$ with $t \geq 0$, see Figure 1 for an illustration. We are interested in fitting a convex function to $\mathbb{D}$ in the least-squares sense:

$$\min_{f:\mathbb{R}\to\mathbb{R}} \quad \sum_{i=1}^{2n+1} \big(f(x_i) - y_i\big)^2, \tag{1}$$

where f is restricted to be convex. As mentioned before, in many applications (see [Groeneboom & Jongbloed, 2014] for many inspiring examples), the true function does have a convex shape (but not necessarily smooth), at least approximately. For example, the insurance risk is roughly a convex function of the age. We remark that for simplicity we have chosen to perturb a single point. More generally, we can perturb a small interval around the middle point $y_{n+1} = t$ and the conclusion would be the same.

It is easy to verify the best convex fit is a constant, i.e., $f(x) \equiv \frac{t}{2n+1}$, with the optimal least-squares objective $\frac{2n}{2n+1}t^2$. The important observation we immediately make is that even changing a single data point can completely destroy convexity: there is no longer any convex function that can fit the data *exactly*, and the optimal least squares convex fit can incur an arbitrarily large loss (as t goes to ∞).[1] We remark that on this toy dataset it is tempting to try to detect and remove the outlying data point $y_{n+1} = t$ (simply by say inspecting Figure 1) and then fit a convex function on the remaining "clean" data points. This approach, while of some value, can become misleading again in higher dimensions where no data point "stands out."

The lesson we learn here is that convexity is a very stringent condition and perhaps not exactly what we aim for. An *approximately* convex function would serve our purpose equally well, if not better. Our main contribution in this work is to point out a very natural quantification of approximate convexity, and a statistically consistent and computationally efficient procedure for estimating approximately convex functions.

Before delving into definitions, let us point out that for univariate functions, Yin & Yu [2017] recently proposed a natural regularizer for estimating a *univariate* function that is the difference of two convex functions (`DC`). Figure 1 illustrates this `DC` approach on our toy dataset. We observe that when the regularization constant λ is small, there is (almost) no restriction on choosing any difference of convex function. As a result, `DC` fits our dataset exactly. On the other hand, when λ is very big, `DC` degenerates to a linear fit, which happens to be the best convex fit here. For moderate λ, `DC` gives a piece-wise linear (but not necessarily convex) fit.

[1] While it is tempting to attribute the large loss to the non-robustness of the least squares loss, we note that if the underlying function that our dataset $\mathbb{D}$ is sampled from is say, a downwards parabola, then any convex estimate is bound to fit poorly and incur a large loss.

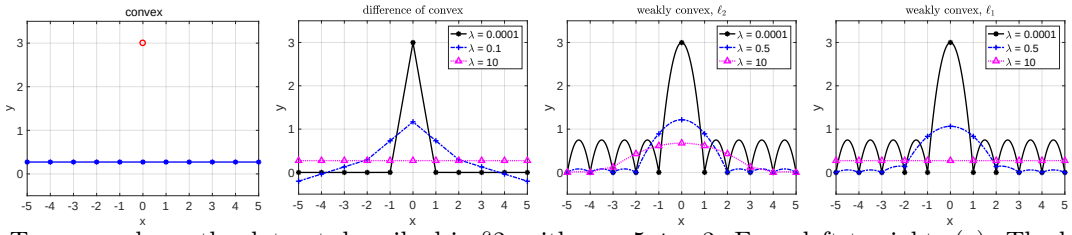

Figure 1: Toy example on the dataset described in §2, with $n = 5, t = 3$. From left to right: (a). The best convex fit is constant $f(x) \equiv \frac{t}{2n+1}$; note the big gap at $x = 0, y = t = 3$. (b). The difference of convex approach in Yin & Yu [2017]. (c). Our weakly convex approach; with ℓ_2^2 regularizer, c.f. (4) below. (d). Our weakly convex approach; with ℓ_1 regularizer. The oscillations between the marked examples in (c) and (d) are due to cancellation of the convex fit (piece-wise linear) and the global quadratic fit.

Despite the appealing performance of `DC` on our toy dataset, it has a serious limitation: it cannot be extended to multivariate non-additive functions. In contrast, our proposed weakly convex (`WC`) approach works for all multivariate functions. The last two subplots in Figure 1 illustrate the fits of `WC` (with different regularizations). Again, when the regularization constant λ is small, there is little restriction on `WC` and as a result it fits the training data (marked points) exactly. As λ gets bigger, `WC` gets closer and closer to the best convex fit. For ℓ_1 regularization (last subplot), a sufficiently large λ will recover the best convex fit while for ℓ_2^2 regularization the convergence only happens in the limit. We will discuss more about regularization in Section 5.

3 Weakly Convex Functions

In this section we recall some key definitions. Let our universe $\mathbb{X}$ be a (convex) subset of $\mathbb{R}^p$. A real-valued function $f : \mathbb{X} \to \mathbb{R}$ is called weakly convex w.r.t. an *arbitrary* norm $\|\cdot\|$ if there exists some $\sigma \in \mathbb{R}$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{X}$ and $\lambda \in [0, 1]$:

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y})+\sigma\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2 \le f(\mathbf{x})+f(\mathbf{y}). \quad (2)$$

Of particular interest are convex functions with $\sigma = 0$ and strongly convex functions with $\sigma > 0$. The largest σ so that (2) holds is called the modulus (of convexity) of f. We remark that convexity itself is an algebraic property that does not rely on the norm while in contrast, the modulus of convexity does depend on the norm. Weakly convex functions are systematically studied first by Vial [1983] under the Euclidean norm.

The following theorem provides a convenient characterization of weakly convex functions:

Theorem 1 (Nikodem & Pàles [2011])**.** *The following are equivalent:*

- *The norm $\|\cdot\|$ is induced by some inner product;*
- *The function $q(\mathbf{x}) = \|\mathbf{x}\|^2$ is weakly convex (w.r.t. norm $\|\cdot\|$) with modulus 1;*
- *f is weakly convex with modulus σ iff $f - \sigma q$ is convex.*

From here on, the norm $\|\cdot\|$ is always induced by an inner product. We denote the set of weakly convex functions on $\mathbb{X}$ with modulus at least σ as $\mathcal{WC}_\sigma = \mathcal{WC}_\sigma(\mathbb{X})$. We will often omit the domain $\mathbb{X}$ in our notations as it is clear from context. In particular, $\mathcal{WC}_+ := \bigcup_{\sigma\ge 0} \mathcal{WC}_\sigma$ is the set of all convex functions, and $\mathcal{WC} := \bigcup_{\sigma\in\mathbb{R}} \mathcal{WC}_\sigma$ denotes the set of all weakly convex functions. We remark that while $\mathcal{WC}_\sigma$ depends on the underlying norm $\|\cdot\|$ (since the modulus does), $\mathcal{WC}_+$ and $\mathcal{WC}$ remain the same for all norms.

Weakly convex functions inherit many nice properties from convex functions. For instance, weakly convex functions are locally Lipschitz hence they admit generalized gradients in the sense of Clarke [1990]. It is easy to verify that $\mathcal{WC}$ is a convex cone, i.e., $f, g \in \mathcal{WC}$ implies $\alpha f + \beta g \in \mathcal{WC}$ for all $\alpha, \beta \ge 0$. However, $\mathcal{WC}$ is not closed under negation (consider for instance the function $-x^4$). It immediately follows that multiplication or composition does not preserve weak convexity in general.

When the domain $\mathbb{X}$ is compact convex, which is perhaps the most relevant in practice, we can say much more about weakly convex functions. Indeed, let us recall the set of locally weakly convex functions $\mathcal{LWC}(\mathbb{X})$, i.e., for each $\mathbf{x} \in \mathbb{X}$ there exists a neighborhood $\mathcal{N}$ of $\mathbf{x}$ so that f is weakly convex on $\mathcal{N} \cap \mathbb{X}$.

Theorem 2. *For compact convex domain $\mathbb{X} \subseteq \mathbb{R}^p$, $\mathcal{LWC}(\mathbb{X}) = \mathcal{WC}(\mathbb{X})$.*

Rockafellar [1982] and Shapiro & Yomdin [1981] characterized locally weakly convex functions, through Clarke's generalized gradient. Many of their results carry over to the weakly convex case, either by restricting the domain $\mathbb{X}$ to be compact or by strengthening the conditions into a global sense. We mention the following convenient results on (locally) weakly convex functions, and defer more to the appendix.

Theorem 3. *A (closed) function f is σ-weakly convex iff $f(\mathbf{x}) = \sup_{t\in T}\langle \mathbf{a}_t, \mathbf{x}\rangle + b_t + \sigma\|\mathbf{x}\|^2$ for some index set T.*

Theorem 4 (Shapiro & Yomdin [1981])**.** *Let $f(\mathbf{x}) =$*

$\sup_{t\in T} f_t(\mathbf{x})$ *for some index set* T*, where each* f_t *is twice continuously differentiable with uniformly bounded Hessian. Then,* f *is weakly convex.*

Thus, weakly convex functions are simply piecewise quadratic, with a bounded Hessian everywhere in the domain (and with possibly *infinitely* many pieces). In the next section we prove that weakly convex functions are in some sense "universal."

4 Universal Approximator

As is apparent from the definition, every weakly convex function is a difference of two convex functions (with the subtrahend being convex quadratic). Perhaps most surprisingly, weakly convex functions can interpolate any finite dataset *exactly*.

Theorem 5. *Let* $\{(\mathbf{x}_i, y_i)\} \subseteq \mathbb{X}\times\mathbb{R}$ *be a finite dataset with* $\mathbf{x}_i = \mathbf{x}_j \implies y_i = y_j$*. Then, there exists* $f \in \mathcal{WC}$ *such that* $f(\mathbf{x}_i) = y_i$ *for all* i*.*

In contrast, as shown in §2, convex functions cannot interpolate certain finite dataset exactly. Moreover, convex functions cannot approximate certain functions, e.g. concave ones, well. On the other hand, weakly convex functions, in addition to accommodating approximate convexity, also enjoy the following *universal approximation* property, putting itself in the same category as deep neural networks (with an unbounded number of neurons) and kernel machines (with universal kernels). Note that the usual class of smooth functions (such as polynomials), while being universal, cannot enforce (approximate) convexity.

Recall that $\mathcal{C}_0(\mathbb{X})$ denotes the set of continuous functions f on $\mathbb{X}$ that vanishes at infinity, i.e., for all $\epsilon > 0$ there exists a compact (convex) set $K \subseteq \mathbb{X}$ such that $|f(\mathbf{x})| < \epsilon$ if $\mathbf{x} \notin K$. As usual, we equip $\mathcal{C}_0(\mathbb{X})$ with the uniform metric $\|f-g\|_\infty := \sup_{\mathbf{x}\in\mathbb{X}} |f(\mathbf{x}) - g(\mathbf{x})|$.

Theorem 6. *Let* $\mathbb{X} \subseteq \mathbb{R}^p$ *be closed (or open) convex. Then,* $\mathcal{WC}(\mathbb{X})\cap\mathcal{C}_0(\mathbb{X})$ *is dense in* $\mathcal{C}_0(\mathbb{X})$*, i.e., for all* $\epsilon > 0$*, for all* $g \in \mathcal{C}_0(\mathbb{X})$*, we can find* $f \in \mathcal{WC}(\mathbb{X}) \cap \mathcal{C}_0(\mathbb{X})$ *so that* $\|f - g\|_\infty < \epsilon$*.*

Using standard results in real analysis (e.g. [Rudin, 1987, Theorem 3.14]), it is immediate that $\mathcal{WC}(\mathbb{X}) \cap \mathcal{L}^p(\mathbb{X})$ is also dense in $\mathcal{L}^p(\mathbb{X})$, the set of functions f with $|f|^p$ integrable w.r.t. say the Lebesgue measure, as long as $1 \le p < \infty$. The surprising aspect of Theorem 6 is that neither convex functions nor quadratic (convex) functions are universal, yet by subtracting the two classes we get a universal approximator[2]. More pleasantly, as we show next, the resulting weakly convex functions inherit many nice statistical and algorithmic properties from the two parent classes, in addition to the analytical ones mentioned in §3.

5 Estimating Weakly Convex Functions

In this section we turn to our main nonparametric *shape-restricted* function estimation problem. Consider the following statistical model: $Y = f(X) + \xi$, where f is an unknown function, and ξ is the random noise with $\mathsf{E}(\xi|X) = 0$. Given a finite dataset $\mathbb{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{X}\times\mathbb{R} : i = 1, \dots, n\}$, we are interested in estimating the unknown function f that generates our dataset $\mathbb{D}$.

We consider the popular least squares[3] estimate:

$$\min_{f:\mathbb{X}\to\mathbb{R}} \ \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 + \lambda\cdot\mathrm{reg}(f), \qquad (3)$$

where $\mathrm{reg}(f)$ is an appropriate regularization term that controls the complexity of f. In particular, we propose to restrict f to the class of weakly convex functions:

$$\min_{f\in\mathcal{WC}_\sigma,\sigma\le 0} \ \frac{1}{n}\sum\nolimits_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 + \lambda\sigma^2, \qquad (4)$$

where $\mathcal{WC}_\sigma$ denotes the set of weakly convex functions with negative[4] modulus σ, and the ℓ_2^2 regularization term on the modulus σ is employed to encourage an (approximately) convex fit. As shown in Theorem 5, without such regularization we are bound to overfit (achieving 0 training error). Clearly, if $\lambda = 0$, then we are searching in the entire class of weakly convex functions (hence we can fit exactly any finite dataset), whereas if $\lambda = \infty$, then we reduce to the familiar class of convex functions. An intermediate λ allows us to avoid overfitting but also to accommodate estimates that are *approximately* convex.

ℓ_2 regularization is also appealing in resolving identifiability issues. Indeed, a weakly convex function, by definition, can be decomposed into the sum of a convex function and a quadratic function. However, the decomposition is not unique. With ℓ_2 regularization, among all decompositions we choose the one with a modulus closest to 0. Of course, the minimizer f in (4) still need not be unique, after all a finite dataset can only determine the value of f on a finite number of points. Alternatively, we can also use ℓ_1 regularization, by replacing $\lambda\sigma^2$ with $\lambda|\sigma|$. In this case, σ may no longer be unique, but for sufficiently large λ, σ will

[2] As pointed out by an anonymous reviewer, many of our results still hold if we replace the quadratic function q with any strongly convex function. It would be interesting to study the effect of some other choice in concrete settings.

[3] It is straightforward to extend all of our results to other reasonable losses, such as the absolute loss.

[4] Strongly convex functions with positive modulus are automatically included, since the class $\mathcal{WC}_\sigma$ increases as σ decreases. The seemingly unnecessary constraint $\sigma \le 0$ is introduced for later developments in §6.

be exactly 0, i.e., we reduce to the convex case with a sufficiently large but finite λ, which is not possible for the ℓ_2 regularization. See Figure 1 for an illustration.

Using Theorem 1 and well-known results in estimating convex functions [Kuosmanen, 2008; Seijo & Sen, 2011; Lim & Glynn, 2012], for example [Boyd & Vandenberghe, 2004, p. 338], we can simplify (4) as the following finite dimensional convex problem:

$$\min_{\mathbf{z},\sigma\le 0, G\in\mathbb{R}^{n\times p}} \frac{1}{n}\sum\nolimits_{i=1}^{n}(z_i+\sigma\|\mathbf{x}_i\|^2-y_i)^2+\lambda\sigma^2 \quad (5)$$

$$\text{s.t.}\;\; z_i \ge z_j + \langle G_{j:}, \mathbf{x}_i-\mathbf{x}_j\rangle, \forall i\ne j. \quad (6)$$

Here we have the decomposition $f=g+\sigma\|\cdot\|^2$ for some convex function g, and $z_i=g(\mathbf{x}_i)$ are the function values on the training samples. Note that (6) is basically the subgradient condition for the convex function g, with the vector $G_{j:}$ being the subgradient of g at $\mathbf{x}_j$. Based on a finite dataset, (5)-(6) returns the estimated function values $f(\mathbf{x}_i)=z_i+\sigma\|\mathbf{x}_i\|^2$ on the training set.

While there are generally infinitely many weakly convex functions f that all fit equally well on the training set, there is a natural *piece-wise quadratic* representative (that is "simple" according to Occam's razor principle):

$$\hat{f}_\lambda(\mathbf{x}) = \hat{g}_\lambda(\mathbf{x})+\sigma\|\mathbf{x}\|^2, \quad (7)$$

$$\hat{g}_\lambda(\mathbf{x}) = \max_{j=1,\dots,n} z_j+\langle G_{j:}, \mathbf{x}-\mathbf{x}_j\rangle. \quad (8)$$

This choice accords to the common practice in previous works [Kuosmanen, 2008; Balázs et al., 2015; Seijo & Sen, 2011; Lim & Glynn, 2012], where $\sigma=0$. Using $\hat{f}_\lambda$ we can predict the function value at any test point $\mathbf{x}$. We remark that the non-uniqueness of $\hat{f}_\lambda$ is inherently common for all *rich* nonparametric function classes: Outside of the convex hull of the training data, we simply do not have any information to extrapolate the value of f (and shape or smoothness constraints would not help). On the other hand, within the convex hull of the training data, different solutions lead to similar (usually identical) interpolations. The statistical and algorithmic properties that we discuss in the next section apply to any choices of the solution, which all converge to the true function as the sample size increases.

6 Why Weakly Convex Functions?

In this section we elaborate on the gained advantages of framing nonparametric function estimation over the class of weakly convex functions. Most saliently, we can extend almost all known results for estimating convex functions to weakly convex functions, and yet gain the universal approximation property discussed in §3, a seemingly "free" lunch.

6.1 Adaptation to geometry

So far we have assumed the inner product induced norm $\|\cdot\|$ is given to us. Conceptually, it almost takes no extra effort to actually learn the underlying geometry induced by the norm and the function f simultaneously and tractably. Indeed, recall that any inner product induced norm can be represented by a symmetric and positive semidefinite matrix $0\preceq Q\in\mathbb{R}^{p\times p}$, with $\langle\mathbf{x},\mathbf{z}\rangle=\mathbf{x}^\top Q\mathbf{z}$ and $\|\mathbf{x}\|=\sqrt{\mathbf{x}^\top Q\mathbf{x}}$. To learn Q simultaneously with f, we modify (5) as follows:

$$\min_{\mathbf{z},Q\preceq 0, G} \frac{1}{n}\sum\nolimits_{i=1}^{n}(z_i+\mathbf{x}_i^\top Q\mathbf{x}_i-y_i)^2+\lambda\|Q\|_F^2 \quad (9)$$

$$\text{s.t.}\;\; z_i \ge z_j + \langle G_{j:}, \mathbf{x}_i-\mathbf{x}_j\rangle, \forall i\ne j, \quad (10)$$

where we have absorbed the (negative) modulus σ into Q. It is clear that (9) is an instance of semidefinite programming (SDP) hence can be solved in polynomial time.

For large datasets, the SDP formulation (9) might take a long time to solve, in which case we can consider restricting Q to be diagonal. The resulting quadratic program can be solved much more efficiently. In fact, its runtime complexity is on par with existing convex estimation methods [Seijo & Sen, 2011; Lim & Glynn, 2012; Balázs et al., 2015], when $p=O(n^2)$. Another alternative might be simply dropping the positive semidefinite constraint $Q\succeq 0$ in (9), without leaving the class of weakly convex functions.

6.2 Computational convenience

Next, we present an efficient meta-algorithm for solving the problem (5)-(6). We observe that the variables $(\mathbf{z},G)$ and σ do not constrain each other so we can simply minimize them alternatively, i.e., we fix $(\mathbf{z},G)$ and minimize σ in closed-form, and then we fix σ and minimize $(\mathbf{z},G)$ using any existing algorithm for estimating convex functions. This meta-algorithm is very intuitive: in each iteration, based on the current convex estimate $(\mathbf{z},G)$ we fit an "optimal" quadratic function to the residual, and then we repeat by estimating the convex component $(\mathbf{z},G)$ again with the adjusted, "more convex" dataset $\{(\mathbf{x}_i, y_i-\sigma\|\mathbf{x}_i\|^2): i=1,\dots,n\}$. We summarize the procedure in Algorithm 1.

Algorithm 1 Alternating Minimization

input : $X=[\mathbf{x}_1,\dots,\mathbf{x}_n]\in\mathbb{R}^{p\times n}$, $\mathbf{y}\in\mathbb{R}^n$, $\lambda,\mu\ge 0$

1 **while** *not converged* **do**

  `// solve the convex component` g

2 $(\mathbf{z},G)\leftarrow$ `cvx_est`$\big(X,\mathbf{y}-\mathrm{diag}(X^\top QX),\mu\big)$

  `// solve the quadratic component` Q

3 $Q\leftarrow$ `quad_est`$\big(X,\mathbf{y}-\mathbf{z},\lambda\big)$

In Algorithm 1, by caching the magnitudes $\|\mathbf{x}_i\|^2$, the step for estimating σ costs only $O(n)$, and is negligible compared to the cost for estimating $(\mathbf{z},G)$. It is straightforward to modify Algorithm 1 for solving (9)

(with a full Q or a diagonal Q), although the step for solving Q no longer admits a closed-form solution. Nevertheless, we can use (accelerated) projected gradient to solve Q where each step costs $O(p^2(p+n))$ for the full case (9) and $O(np)$ for the diagonal case. We omit the obvious details.

Lastly, let us mention how to solve the convex component g, i.e., solving $(\mathbf{z}, G)$. This is an instance of standard convex quadratic programs, with $(p+1)n$ variables and $n(n-1)$ linear constraints. For small p and n we can simply use standard convex optimization toolboxes. For large p or n, we can turn to first order methods such as the cutting-plane method in [Balázs et al., 2015] or the clustering approach in [Magnani & Boyd, 2009]. In fact, a more direct approach for solving (9) or (5) is also possible. For instance, we can easily adapt the multi-block ADMM algorithm of Mazumder et al. [2018].

The convergence of Algorithm 1 follows easily from the general result of Beck & Tetruashvili [2013, §5].

6.3 Statistical properties

The key to establish statistical consistency and rates of convergence of the least-squares weakly convex estimate in (7) is the notion of metric entropy. Let our domain $\mathbb{X}$ be compact convex and $\mathcal{F}$ be a class of uniformly bounded continuous functions on $\mathbb{X}$, equipped with the uniform metric. An ϵ-cover of $\mathcal{F}$ is a subset $\mathcal{N} \subseteq \mathcal{F}$ such that for all $f \in \mathcal{F}$, there exists some $g \in \mathcal{N}$ with $\|f - g\|_\infty \leq \epsilon$. The metric entropy of $\mathcal{F}$ (w.r.t. the uniform metric), defined as

$$H_\epsilon(\mathcal{F}) := \inf\{\log|\mathcal{N}| : \mathcal{N} \text{ is an } \epsilon\text{-cover of } \mathcal{F}\}, \quad (11)$$

is a natural measure of the size of $\mathcal{F}$, which is usually infinite dimensional in nonparametric estimation. The following result is obvious from the definitions:

Theorem 7. $H_{2\epsilon}(\mathcal{F}+\mathcal{G}) \leq H_\epsilon(\mathcal{F}) + H_\epsilon(\mathcal{G})$.

Now, let $\mathcal{F} = \mathcal{C}_{\mathbb{X},L,B}$ be the class of convex functions on $\mathbb{X}$ that are uniformly bounded by B and that are Lipschitz continuous with Lipschitz constant at most L. The recent work of Balázs et al. [2015] proved that $H_\epsilon = O\left((\frac{1}{\epsilon})^{d/2}\log\frac{1}{\epsilon}\right)$. On the other hand, it is well-known that a compact convex body in a d-dimensional space has metric entropy $O(\log\frac{1}{\epsilon})$. Therefore, if we let $\mathcal{M}$ be the class of positive semidefinite matrices Q such that (say) $\|Q\|_F \leq C$, then $H_\epsilon(\mathcal{M}) = O(\log\frac{1}{\epsilon})$. Let $\mathcal{G}$ be the class of quadratic functions $\mathbf{x}^\top Q\mathbf{x}$ with $\|Q\|_F \leq C$. Note that for two quadratic functions $\mathbf{x}^\top Q\mathbf{x}$ and $\mathbf{x}^\top P\mathbf{x}$, their uniform distance $\sup_{\mathbf{x}\in\mathbb{X}}|\mathbf{x}^\top Q\mathbf{x} - \mathbf{x}^\top P\mathbf{x}| \propto \|Q-P\|_2$, whence follows that $H_\epsilon(\mathcal{G}) \propto H_\epsilon(\mathcal{M}) = O(\log\frac{1}{\epsilon})$. Using Theorem 7 we conclude that $H_{2\epsilon}(\mathcal{F}+\mathcal{G}) \leq H_\epsilon(\mathcal{F}) + H_\epsilon(\mathcal{G}) = O\left((\frac{1}{\epsilon})^{d/2}\log\frac{1}{\epsilon}\right)$. Observe that the set $\mathcal{F}+\mathcal{G}$ consists of weakly convex functions whose modulus of convexity is uniformly bounded.

Equipped with the above result, we can now invoke [Balázs et al., 2015, Theorem 3.1]: assuming the noise ϵ is *iid* Gaussian (or more generally, subgaussian) with constant variance, and the regression function f belongs to the class $\mathcal{F}+\mathcal{G}$, then the least-squares estimate in (7) enjoys the following statistical guarantees: $\hat{f}_n \to f$ in mean square at the rate $n^{-2/d}\log n$ when say $d > 4$. In particular, the least-squares estimate is consistent. We remark that it is possible to improve the above bound, by choosing a suitable number of pieces in (7) (instead of n), see [Balázs et al., 2015, Theorem 4.2] for details.

6.4 Minimization of the estimate

Unlike many other function classes used in nonparametric estimation, weakly convex functions are particularly amenable to minimization [Spingarn, 1982]. In some applications, we are interested in finding the minimizer or minimum value of an unknown function f. The so-called meta-modeling approach, as studied in Hannah et al. [2014], first estimates f based on a finite sample, and then finds the minimizer of the estimated function. Apparently, the meta-modeling approach is appealing only when the estimated function is "easy" to minimize, which is true if we restrict ourselves to convex functions.

The same is true for weakly convex functions, with the additional advantage of being a universal approximator. Indeed, we can apply the celebrated proximal point algorithm [Rockafellar, 1976] to minimize the weakly convex estimate $\hat{f}$, which amounts to repeatedly computing the following proximity operator:

$$\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{x}} \hat{f}(\mathbf{x}) + \tfrac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|^2. \quad (12)$$

For sufficiently small η_t (in particular, when η_t is smaller than the absolute modulus of convexity $|\sigma|$ of $\hat{f}$), the proximity operator (12) is a well-defined *convex* problem hence can be solved using standard convex optimization techniques. In fact, using the convention in (7), (12) is a simple (convex) quadratic program. It follows then from the result of Attouch & Bolte [2009] that $\mathbf{x}_t$ converges to a critical point of $\hat{f}$, whose quality is usually found to be quite reasonable in practice.

7 Experiments

In this section, we conduct experiments to compare the following function estimation schemes:

- `convex` [Balázs et al., 2015; Mazumder et al., 2018; Lim & Glynn, 2012; Seijo & Sen, 2011]: assuming the unknown regression function is convex;
- `concave`: a straightforward "negation" of `convex`;
- `linear`: assuming the regression function is linear—a parametric estimation procedure;

Table 1: Mean square test error (standard deviation) on 4 synthetic datasets over 10 random repetitions.

	$\|\mathbf{x}\|_2^2(\lambda=0.1)$	$-\|\mathbf{x}\|_2(\lambda=0.1)$	$-\|\mathbf{x}\|_2^2(\lambda=10^{-5})$	$10\sin(\|\mathbf{x}\|_2)(\lambda=10^{-5})$
ℓ_1-wc	0.5829 (0.0246)	0.2374 (0.0071)	**0.0329** (**0.0035**)	**0.1825** (**0.0175**)
ℓ_2^2-wc	0.5829 (0.0246)	**0.0376** (**0.0068**)	**0.0328** (**0.0035**)	**0.1825** (**0.0175**)
convex	0.5829 (0.0246)	0.2374 (0.0071)	1.3061 (0.0464)	0.7626 (0.0489)
concave	1.3340 (0.0520)	**0.0371** (**0.0038**)	0.5730 (0.0196)	0.3434 (0.0248)
linear	0.9426 (0.0463)	0.0913 (0.0068)	0.9245 (0.0281)	0.5216 (0.0439)
dc	**0.1014** (**0.0173**)	0.0429 (0.0068)	0.1043 (0.0138)	0.2519 (0.0256)

- `DC` [Yin & Yu, 2017]: assuming the regression function is the sum of univariate difference of convex functions;
- `WC` (proposed): with ℓ_2^2 and ℓ_1 regularization on the modulus of convexity σ.

7.1 Synthetic dataset

Our goal here is to confirm the usefulness of weakly convex functions in nonparametric estimation, through well-controlled numerical simulations (where we know the actual true regression functions). We generate our data $X \in \mathbb{R}^{n\times p}$ and $\mathbf{y} \in \mathbb{R}^n$ in the following way: each entry in X is an *iid* sample from the uniform distribution on $[-1,1]^p$ and each entry in $\mathbf{y}$ is obtained by $f(\mathbf{x})+\xi$, where f is the unknown regression function and $\xi \sim \mathcal{N}(0,\gamma^2 I)$ is an *iid* sample from the standard Gaussian distribution. We consider 4 different regression functions:

- Scenario 1: $f_1(\mathbf{x}) = \|\mathbf{x}\|_2^2$, which is additive and convex (favorable to `DC` and `convex`);
- Scenario 2: $f_2(\mathbf{x}) = -\|\mathbf{x}\|_2$, which is non-additive and concave (favorable to `concave`);
- Scenario 3: $f_3(\mathbf{x}) = -\|\mathbf{x}\|_2^2$, which is additive and concave (favorable to `concave`, `DC` and `WC`);
- Scenario 4: $f_4(\mathbf{x}) = 10\sin(\|\mathbf{x}\|_2)$, which is neither additive nor convex nor concave.

We set $\gamma = 1$ and $p = 10$ and we tune hyper-parameters on a validation set, after which we fix the best parameters and repeat each experiment 10 times. For each experiment, we generate 500 and 1000 training and test points, respectively. The averaged mean square test errors (standard deviations) are reported in Table 1, from which we make the following observations.

For Scenario 1, `DC` performed best, followed by `WC` and `convex`. The reason why `DC` performed the best here is that the true function f_1 is indeed additive hence `DC` can exploit this information for statistical efficiency. `WC` performed the same as `convex` since on this dataset the best estimate of the modulus of convexity σ is obviously 0. `concave` and `linear` performed significantly worse, since the true function is indeed quite far from being concave or linear. For Scenario 2, instead, we see that when the true regression function is not additive, the performance of `DC` suffered, while `WC` (with ℓ_2^2 regularization) correctly estimated the concave function f_2. Not surprisingly, `concave` also estimated f_2 well while `convex` and `linear` incurred a large error. Scenario 3 is the "negation" of Scenario 1. As a consequence, `DC` performed very similarly in these two scenarios while `WC` performed *surprisingly* much better in Scenario 3. The reason here is because in Scenario 1 `WC` had to estimate the convex quadratic function f_1 using piecewise linear functions (the nonnegative constraint $\sigma \le 0$ forces σ to be roughly 0) while in Scenario 3 `WC` benefited from employing an explicit concave quadratic term $\sigma\|\mathbf{x}\|_2^2$ in its estimate, which would have incurred a large loss if approximated by piecewise linear functions, like in Scenario 1. Lastly, Scenario 4 shows an example where the true regression function f_4 is neither additive nor convex nor concave. All methods except `WC` seemed to suffer significantly on this dataset, hence demonstrating the flexibility of weakly convex functions.

Figure 2 illustrates the overfitting effect of `WC` and how regularization helps alleviate this issue. We observe that without regularization (middle column), `WC` reproduces the training set (left column) almost exactly, confirming Theorem 5 and signaling overfitting. With regularization (right column), `WC` achieves much better test performance instead.

7.2 Real datasets

We also conduct experiments on three real datasets below. Each experiment is repeated 10 times with cross-validation (for selecting hyper-parameters), and we normalize the features and the output.

- **R-1:** Based on years of education and experience, we predict mean weekly wages using the dataset ex1029 in the `Sleuth2` package in R. It has also been used previously in Hannah & Dunson [2013] for convex estimation. Empirically, mean weekly wages are approximately concave w.r.t. years of experience, but not concave or convex w.r.t. years of education. A common trick is to conduct an exponential transform, e.g. $1.2^{\text{years education}}$, to induce concavity [Hannah &

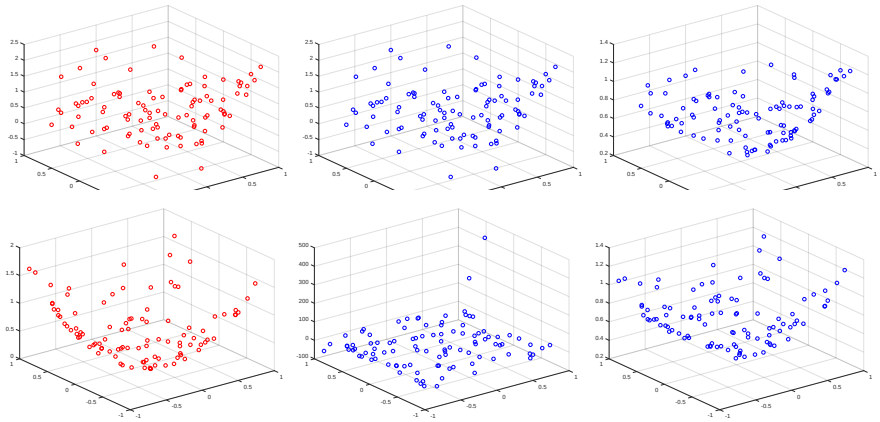

Figure 2: The effect of overfitting and regularization. First column: training (above) and test (below). Second column: estimated responses without regularization. Third column: estimated responses with regularization.

Table 2: Mean square test error (standard deviation) on 4 real datasets over 10 random repetitions.

	R-1(a)	**R-1(b)**	**R-2**	**R-3**
ℓ_1-wc	**0.0010** (**0.0002**)	**8.867**e-04 (**1.406**e-04)	**0.0016** (**0.0012**)	**0.0010** (**0.0001**)
ℓ_2^2-wc	**0.0010** (**0.0002**)	**8.867**e-04 (**1.406**e-04)	**0.0016** (**0.0012**)	**0.0010** (**0.0001**)
convex	0.0014 (0.0005)	0.0019 (0.0011)	0.0341 (0.0315)	3.2240 (3.5396)
concave	**0.0010** (**0.0004**)	0.0014 (0.0005)	0.1023 (0.1141)	0.5322 (0.1679)
linear	0.0012 (0.0002)	0.0012 (0.0002)	0.0030 (0.0020)	0.0014 (0.0002)
dc	0.0020 (0.0006)	0.0022 (0.0007)	0.0040 (0.0026)	0.0023 (0.0007)

Dunson, 2013]. We perform experiment for both the original input (denoted as **R-1(b)**) and the exponentially transformed input (denoted as **R-1(a)**). This dataset has 858 valid data points and we randomly choose 300 of them for training.

- **R-2:** We use the NBER-CES Manufacturing Industry Database. We predict the total value of shipment (vship) based on 4 features: total real capital stock (cap), production worker hours (prodh), non-production workers (emp-prode), and production workers (prode). The data is collected from 1958 to 2009, and we focus on the year 2000. There are totally 473 data points, and 300 of them are used for training. This dataset has been used in Mazumder et al. [2018] for convex estimation and we follow its suggestion to take a log-transform of all features.
- **R-3:** Concrete is the most important material in civil engineering. We use the UCI dataset to predict the concrete compressive strength based on 8 features: cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age. There are totally 1030 data points and we randomly choose 300 points for training.

The experiment results are reported in Table 2. For **R-1(a)**, since the exponential transform leads to an approximately concave relationship between the input and output, the performance of `WC` and `concave` are not surprisingly very similar. However, without the exponential transform, `WC` performed the best on **R-1(b)**, followed by `linear` and then `concave`, demonstrating its robustness. On **R-2**, we empirically verified each input and output exhibiting an increasing and linear trend at the beginning but a less clear trend at the end. In this case, `WC` performed the best with MSE nearly a half of that of `linear`, demonstrating its flexibility. For **R-3**, the concrete compressive strength is considered to be highly non-linear w.r.t. age and ingredients. Empirical visualization does not show any clear trend between individual features and the output. Nevertheless, `WC` again performed the best, thanks to its universal approximation property and regularization, followed by `linear`, while `convex` severely overfits and suffers a significantly large MSE.

8 Conclusions

We have proposed weakly convex regression to alleviate the stringent convexity constraint. Weakly convex functions can exactly interpolate any finite dataset, they are universal approximators and they can accommodate approximate convexity. To combat overfitting, we proposed to regularize the least-squares weakly convex estimate by the modulus of convexity. The resulting formulation inherits nice statistical and algorithmic properties from its convex counterpart. Our numerical experiments confirmed the competitiveness of weakly convex functions in nonparametric estimation.

Acknowledgements

We thank the reviewers for their critical comments and suggestions. We gratefully acknowledge the support from NSERC.

References

Attouch, H. and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

Ayer, Miriam, Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, Edward. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955.

Balázs, Gábor, György, András, and Szepesvári, Csaba. Near-optimal max-affine estimators for convex regression. In *AISTATS*, 2015.

Beck, Amir and Tetruashvili, Luba. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Clarke, Frank H. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

Grenander, U. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 39(2):125–153, 1956.

Groeneboom, Piet and Jongbloed, Geurt. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

Györfi, Làszlò, Kohler, Micael, Krzyzak, Adam, and Walk, Harro. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Hannah, L. A. and Dunson, D. B. Ensemble methods for convex regression with applications to geometric programming based circuit design. In *ICML*, 2012.

Hannah, L. A. and Dunson, D. B. Multivariate convex regression with adaptive partitioning. *Journal of Machine Learning Research*, 14:3261–3294, 2013.

Hannah, L. A., Powell, W. B., and Dunson, D. B. Semiconvex regression for metamodeling-based optimization. *SIAM Journal on Optimization*, 24(2): 573–597, 2014.

Hildreth, Clifford. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619, 1954.

Kuosmanen, Timo. Representation theorem for convex nonparametric least squares. *Econometrics Journal*, 11:308–325, 2008.

Lim, Eunji and Glynn, Peter W. Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208, 2012.

Magnani, Alessandro and Boyd, Stephen P. Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1):1–17, 2009.

Mazumder, Rahul, Choudhury, Arkopal, Iyengar, Garud, and Sen, Bodhisattva. A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 2018. to appear.

NBER-CES. NBER-CES manufacturing industry database. http://www.nber.org/data/nbprod2005.html.

Nikodem, Kazimierz and Pàles, Zsolt. Characterizations of inner product spaces by strongly convex functions. *Banach Journal of Mathematical Analysis*, 5(1):83–87, 2011.

Rockafellar, R. Tyrrell. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

Rockafellar, R. Tyrrell. Favorable classes of lipschitz continuous functions in subgradient optimization. In Nurminski, E. (ed.), *Progress in Nondifferentiable Optimization*, pp. 125–143. 1982.

Rudin, Walter. *Real and Complex Analysis*. McGraw-Hill, 3rd edition, 1987.

Seijo, Emilio and Sen, Bodhisattva. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3): 1633–1657, 2011.

Shapiro, A. and Yomdin, Y. On functions, representable as a difference of two convex functions, and necessary conditions in a constrained optimization. Technical report, 1981.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on Stochastic Programming, Modeling and Theory*. SIAM, 2009.

Sondik, Edward J. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, 1978.

Spingarn, Jonathan E. Submonotone mappings and the proximal algorithm. *Numerical Functional Analysis and Optimization*, 4(2):123–150, 1982.

UCI. Concrete compressive strength data set. https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength.

Varian, H. R. The nonparametric approach to demand analysis. *Econometrica*, 50(4):945–973, 1982.

Varian, H. R. The nonparametric approach to production analysis. *Econometrica*, 52(3):579–597, 1984.

Vial, Jean-Philippe. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

Xu, Min, Chen, Minhua, and Lafferty, John. Faithful variable screening for high-dimensional convex regression. *The Annals of Statistics*, 44(6):2624–2660, 2016.

Yin, J. and Yu, Y. Convex-constrained sparse additive modeling and its extensions. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.