# Near-optimal max-affine estimators for convex regression

**Gábor Balázs** **András György** **Csaba Szepesvári**
Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8

## Abstract

This paper considers least squares estimators for regression problems over convex, uniformly bounded, uniformly Lipschitz function classes minimizing the empirical risk over max-affine functions (the maximum of finitely many affine functions). Based on new results on nonlinear nonparametric regression and on the approximation accuracy of max-affine functions, these estimators are proved to achieve the optimal rate of convergence up to logarithmic factors. Preliminary experiments indicate that a simple randomized approximation to the optimal estimator is competitive with state-of-the-art alternatives.

## 1 INTRODUCTION

In this paper we consider the problem of estimating an unknown regression function that is known to be convex based on independent, identically distributed (i.i.d.) samples. We also restrict the estimates to be convex. Such *convex regression problems* arise in various contexts, such as econometrics (Varian, 1982, 1984; Merton, 1992), geometric programming (Magnani and Boyd, 2009; Hannah and Dunson, 2012), or operations research/reinforcement learning (Shapiro et al., 2009; Hannah et al., 2014), just to name a few. While early papers on convex regression explored the single-variable case, recently attention shifted towards multivariate problems (see, e.g., Seijo and Sen, 2011; Hannah and Dunson, 2012). Despite all the effort in designing new algorithms and proving theoretical guarantees for them, even basic questions such as whether least-squares estimators can achieve the optimal minimax rate for bounded convex regression problems remained unknown (for a bounded regression problem, the regression domain, the function and its Lipchitz factor are all bounded). In this paper we resolve this open question and answer many related ones.

Our results are built on a new, technical theorem that bounds the expected risk of nonlinear least-squares estimators (LSEs), which might be of interest on its own. This theorem extends the "classical" chaining argument and achieves the same rates as techniques based on local Rademacher complexity (Bartlett et al., 2005; Koltchinskii, 2008) for problems with subgaussian noise and function classes having finite supremum-norm entropy (Section 3). Note that the methods based on local Rademacher complexity can only handle bounded noise, but work with more general norms and loss functions. Additionally, our treatment improves the constants in the bounds by orders of magnitude, and also includes the case of nonzero approximation error, which turns out to be a crucial detail to prove near-optimal rates for max-affine estimators (Section 4.4).

After establishing a lower bound on the minimax expected risk for a wide class of convex regression problems (Theorem 4.1), we show that LSEs, up to logarithmic factors, achieve the optimal rate on bounded regression problems when the dimension $d$ is not higher than four, while we obtain a suboptimal rate for $d > 4$. A similar "phase-transition" is expected to happen because the function space becomes "massive" when the dimension is larger than four in the sense that its entropy integral diverges (see, e.g., van de Geer, 2000). To prove these results, Section 4.1 builds on the work of Bronshteyn and Ivanov (1975), and shows how well convex functions can be approximated by the maximum of finitely many affine functions (in short: max-affine functions). The same section also provides a new covering bound for the class of bounded convex Lipschitz functions. This new result removes the exponential dependence of a constant in the minimax risk bounds on the dimension.

It is important to point out that in the above mentioned results, LSEs search the class of convex functions having a bounded range and Lipschitz factor. Earlier works considered the problem without the boundedness conditions (see, e.g., Seijo and Sen, 2011;

Appearing in Proceedings of the $18^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. 

Lim and Glynn, 2012); however, we demonstrate that the expected risk of their estimator may become infinite for any sample size even for perfectly benign data. In Section 4.3, we show that the efficient (polynomial) computation methods available for the case without constraints carries over to our situation, and the resulting optimization problem still belongs to the family of quadratic programs (QPs). Since the number of constraints grows quadratically in the sample size, standard QP solvers become impractical even for moderate sample sizes. To alleviate this problem, we propose a specialized cutting plane solver, which is observed to provide significant speedups compared to other variants.

Finally, Section 4.4 provides a new LSE class whose rate (up to logarithmic factors) matches the lower bound (Theorem 4.2), thus resolving the open question mentioned earlier. The idea of this new LSE comes from the observation that the class of convex LSEs (restricted to a bounded range and Lipchitz factor) contains max-affine functions using at most as many hyperplanes as the sample size $n$. The new LSE class is formed by taking these max-affine LSEs and further restricting their complexity (the number of used hyperplanes) to balance their estimation and approximation errors (by using the approximation results of Section 4.1). We also propose a heuristic approach to compute these estimators, whose performance is studied empirically in Section 5.

## 2 REGRESSION PROBLEMS

We consider regression problems defined by some class of probability distributions $\mathcal{D}$ over some set $\mathbb{X} \times \mathbb{R}$, where $\mathbb{X} \subseteq \mathbb{R}^d$ is a subset of the $d$-dimensional Euclidean space.[1] An instance of a regression problem is defined by a distribution $\mu \in \mathcal{D}$. The regression estimator's job is to produce a function $f : \mathbb{X} \to \mathbb{R}$ based on a training set $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of $n \in \mathbb{N}$ pairs $(X_i, Y_i)$, independently sampled from $\mu$ (in short: $D_n \sim \mu^n$), such that on a new instance $(X, Y) \sim \mu$, the prediction error, $|f(X) - Y|^2$ is small. Formally, an estimator is a sequence $(h_n)_{n\in\mathbb{N}}$ of mappings $h_n : (\mathbb{X} \times \mathbb{R})^n \to \{\mathbb{X} \to \mathbb{R}\}$, where $\{\mathbb{X} \to \mathbb{R}\}$ denotes the set of functions mapping $\mathbb{X}$ to $\mathbb{R}$.

For a fixed $\mu$, the (expected) cost of using a fixed function $f : \mathbb{X} \to \mathbb{R}$ is equivalently measured by either its expected squared prediction error (or $L_2$-risk) defined by $\mathbb{E}[|f(X) - Y|^2]$, or its squared $L_2$-error, $\|f - f_*\|_\mu^2 \doteq \mathbb{E}[|f(X) - f_*(X)|^2]$, where $f_* = f_*^\mu = \arg\min_{f:\mathbb{X}\to\mathbb{R}} \mathbb{E}[|f(X) - Y|^2]$ denotes the regression function, which also satisfies $f_*(X) = \mathbb{E}[Y|X]$ almost surely (a.s.). The said equivalence follows from the well known identity $\mathbb{E}[|f(X) - Y|^2] = \|f - f_*^\mu\|_\mu^2 + \mathbb{E}[|f_*^\mu(X) - Y|^2]$.

The cost of an estimator on a regression problem specified by a distribution $\mu$ is defined as its expected squared $L_2$-error, $L_n(h_n, \mu) = \mathbb{E}[\|h_n(D_n) - f_*^\mu\|_\mu^2]$, where the data is generated i.i.d. from $\mu$, $D_n \sim \mu^n$. The worst-case cost of $h_n$ over $\mathcal{D}$ is

$$L_n(h_n, \mathcal{D}) \doteq \sup_{\mu\in\mathcal{D}} L_n(h_n, \mu)\,.$$

As a baseline for comparing estimators, we use the minimax error over $\mathcal{D}$,

$$L_n(\mathcal{D}) \doteq \inf_{h_n} \sup_{\mu\in\mathcal{D}} \mathbb{E}\Big[\, \|h_n(D_n) - f_*\|_\mu^2 \,\Big]\,,$$

where the infimum is taken over all $(\mathbb{X} \times \mathbb{R})^n \to \{f \mid f : \mathbb{X} \to \mathbb{R}\}$ mappings (including non-convex ones). We say that the estimator $(h_n)_{n\in\mathbb{N}}$ is near-optimal if it is suboptimal only up to poly-logarithmic factors, that is, if for some $p \geq 0$, $\limsup_{n\to\infty} \frac{L_n(h_n,\mathcal{D})}{L_n(\mathcal{D})\ln^p(n)} < \infty$.

In this paper we will be concerned with *convex regression* problems when the domain $\mathbb{X}$ is convex and the regression functions $f_*^\mu$ are convex for all $\mu \in \mathcal{D}$. Furthermore, we are interested in finding estimators that produce convex functions as estimates.

We study least squares estimators which minimize the empirical squared prediction error, or empirical $L_2$-risk. Precisely, an estimator is called an $\alpha$-approximate least-squares estimator with respect to some function set $\mathcal{F} \subseteq \{\mathbb{X} \to \mathbb{R}\}$, or $\alpha$-LSE($\mathcal{F}$) for short, if its estimate $f_n = h_n(D_n)$ satisfies

$$\frac{1}{n}\sum_{i=1}^n |f_n(X_i) - Y_i|^2 \leq \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 + \alpha\,, \quad (1)$$

where $\alpha \in [0, \infty)$ is a constant.

We will need covering numbers[2] for our results, hence we give the basic definitions here. Let $(\mathcal{F}, \ell)$ be a metric space and $\epsilon \geq 0$. The set $\{f_1, \dots, f_k\} \subseteq \mathcal{F}$ is called an $\epsilon$-net of $\mathcal{F}$ with respect to $\ell$ if the $\ell$-balls of centers $\{f_1, \dots, f_k\}$ and radius $\epsilon$ cover $\mathcal{F}$: for any $f \in \mathcal{F}$, $\min_{i=1,\dots,k} \ell(f, f_i) \leq \epsilon$. The *$\epsilon$-covering number of $\mathcal{F}$* with respect to $\ell$, denoted by $\mathcal{N}(\epsilon, \mathcal{F}, \ell)$, is the cardinality of the $\epsilon$-net with the fewest elements:

$$\mathcal{N}(\epsilon, \mathcal{F}, \ell) \doteq \inf\Big\{k \in \mathbb{N} \,\Big|\, \exists f_1, \dots, f_k \in \mathcal{F} : \sup_{f\in\mathcal{F}} \min_{i=1,\dots,k} \ell(f, f_i) \leq \epsilon\Big\}$$

[1] All sets and functions considered are assumed to be measurable as necessary. To simplify the presentation, we omit these conditions in the rest of the paper by noting here that all the measurability issues can be overcome using standard techniques as we work only with bounded domains and functions over Euclidean spaces.

[2] More precisely, we use internal covering numbers, where the net is restricted to lie inside the covered set.

with $\inf \emptyset = \infty$. The *$\epsilon$-entropy* of $\mathcal{F}$ with respect to $\ell$ is defined as $\mathcal{H}(\epsilon, \mathcal{F}, \ell) \doteq \ln \mathcal{N}(\epsilon, \mathcal{F}, \ell)$. Furthermore, for any function space $\mathcal{F} \subseteq \{\mathbb{X} \to \mathbb{R}\}$, $\mathcal{N}_\infty(\epsilon, \mathcal{F}) \doteq \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ and $\mathcal{H}_\infty(\epsilon, \mathcal{F}) \doteq \mathcal{H}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ will denote the sup-norm covering number and entropy, resp., where $\|f\|_\infty \doteq \sup_{x \in \mathbb{X}} |f(x)|$ for any $f : \mathbb{X} \to \mathbb{R}$.

## 3 REGRESSION ERROR BOUNDS

We start with a general result, which bounds the expected squared $L_2$-error for LSEs over general function sets $\mathcal{F}$ in terms of the entropy of $\mathcal{F}$. This result allows us to obtain sharp (sometimes optimal) rates and so competes with the techniques based on local Rademacher complexity, such as Corollary 5.3 of Bartlett et al. (2005) or Theorem 5.1 of Koltchinskii (2008). These theorems also provide similar results when combined with an upper bound on the Rademacher complexity such as Lemma A.3 of Srebro et al. (2012). However, as opposed to all results based on local Rademacher complexities, which, as pointed out recently by Mendelson (2014), require that the range of the response variable $Y$ be bounded, our result allows unbounded $Y$ as long as its tail is sufficiently well-behaving.

Recently, Lecué and Mendelson (2013); Mendelson (2014) also developed a new technique to deal with subgaussian noise (or with even weaker assumptions on the noise) on the price of making stronger assumptions on the function class $\mathcal{F}$ (e.g., "Bernstein" and "star-shaped"). Furthermore, their bounds contain some quantities which are not straightforward to compute; in contrast, our result uses only sup-norm entropies, which are readily available for many standard function classes of interest. Additionally, we only require boundedness of the function class $\mathcal{F}$, making our result directly and easily applicable for analyzing LSEs in these cases. In particular, as we shall show, this bound will imply that some max-affine least squares estimators can achieve near-optimal rates for convex regression (Sections 4.2 and 4.4). The proof of the bound directly extends the "classical" chaining argument making it capable to deliver fast rates, while it provides significantly sharper constants than those previously reported in the literature (e.g., Bartlett et al., 2005, Corollary 5.3).

The promised result is as follows:

**Theorem 3.1.** *Assume that the regression function and the function class $\mathcal{F}$ are bounded by some positive real $B$, i.e. $\|f\|_\infty \le B$ for all $f \in \{f_*\} \cup \mathcal{F}$, and the noise is uniformly $\sigma$-subgaussian with some $\sigma \ge 0$, i.e.*

$$\sup_{s \in \mathbb{R}} \mathbb{E}\left[ e^{s(Y - f_*(X)) - s^2\sigma^2/2} \,\middle|\, X \right] \le 1 \quad a.s.$$

*Let $f_n = h_n(D_n)$ be an $\alpha$-LSE($\mathcal{F}$) estimate (1) and set $B_\sigma \doteq \max\{B, \sigma\}$. Then for all $\delta \in [0, B]$,*

$$\mathbb{E}\left[ \|f_n - f_*\|_\mu^2 \right] \le \frac{26 B_\sigma}{\sqrt{n}} \int_\delta^B \sqrt{\mathcal{H}_\infty(s, \mathcal{F})}\, ds + 40 B_\sigma \delta + \inf_{f \in \mathcal{F}} \|f - f_*\|_\mu^2 + \alpha\,, \tag{A}$$

*and for all $\epsilon \in [0, \infty)$, $\delta \in [0, \epsilon]$,*

$$\mathbb{E}\left[ \|f_n - f_*\|_\mu^2 \right] \le c_1 B_\sigma^2 \frac{\mathcal{H}_\infty(\epsilon, \mathcal{F})}{n} + \frac{c_2 B_\sigma}{\sqrt{n}} \int_\delta^\epsilon \sqrt{\mathcal{H}_\infty(s, \mathcal{F})}\, ds + c_3 B_\sigma \delta + c_4 \left( \inf_{f \in \mathcal{F}} \|f - f_*\|_\mu^2 + \alpha \right), \tag{B}$$

*where $(c_1, c_2, c_3, c_4)$ is an element of $\{(12, 153, 171, 9/4), (16, 108, 120, 3/2), (43, 80, 89, 10/9)\}$.*

*Proof.* See Appendix A.1 and Appendix A.2. □

The general connection among the constants $c_1, c_2, c_3, c_4$ can be found in the proofs of Theorem 3.1 and Lemma A.6. Here we simply listed the values we use in order to optimize different terms in our bounds.

Notice that $c_4 > 1$ in all cases, so the second inequality provides a convergence rate only if the approximation error is zero, or converges to zero fast enough by expanding $\mathcal{F}$ appropriately as $n$ grows.

One might wonder whether a similar result could be proved in general when $\mathcal{F}$ is not uniformly bounded.[3] In Section 4.3 we answer this question negatively by means of giving an example where the risk of LSEs is infinite when used with max-affine functions even when $Y$ is bounded.

## 4 CONVEX REGRESSION

Let $\mathbb{X} \subseteq \mathbb{R}^d$ be a convex set with a nonzero, finite diameter $\operatorname{diam}(\mathbb{X}) \doteq \sup_{x,z \in \mathbb{X}} \|x - z\|_\infty$ with respect to the supremum norm $\|\cdot\|_\infty$ on $\mathbb{R}^d$. Let

$$\partial f(x) \doteq \left\{ s \in \mathbb{R}^d \,\middle|\, \forall z \in \mathbb{X} : f(z) \ge f(x) + s^\top (z - x) \right\}$$

denote the set of subgradients of a convex function $f : \mathbb{X} \to \mathbb{R}$ at $x \in \mathbb{X}$. Define the class of convex, uniformly bounded, subdifferentiable, uniformly Lipschitz functions on $\mathbb{X}$,

$$\mathcal{C}_{\mathbb{X},B,L} \doteq \Big\{ f : \mathbb{X} \to \mathbb{R} \,\Big|\, f \text{ is convex}, \|f\|_\infty \le B, \forall x \in \mathbb{X} : \partial f(x) \ne \emptyset, \forall s \in \partial f(x) : \|s\|_\infty \le L \Big\}$$

[3] Boundedness might be dropped in some special cases. See, e.g., Mendelson (2014) for hyperplane estimation.

with $B, L > 0$. In what follows we consider convex regression problems, where the regression function satisfies $f_*^\mu \in \mathcal{C}_{\mathbb{X},B,L}$.

First, we give a lower bound on the minimax rate for these problems. For this, define

$$\begin{aligned}\mathcal{D}_\sigma \doteq \big\{&\text{distribution } \mu \text{ over } \mathbb{X}\times\mathbb{R} \,\big|\, X_0 \times Y \sim \mu,\\ &X_0 \text{ is uniformly distributed on } \mathbb{X},\\ &Y = f_*(X_0) + \xi,\ f_* \in \mathcal{C}_{\mathbb{X},B,L},\ \xi \sim \mathcal{N}(0,\sigma^2),\\ &X_0 \text{ and } \xi \text{ are independent}\big\},\end{aligned}$$

where $\sigma \in (0,\infty)$ and $\mathcal{N}(0,\sigma^2)$ denotes the zero-mean normal distribution with variance $\sigma^2$. Define the Euclidean distance for square integrable functions, $\|f-g\|_{X_0}^2 \doteq \mathbb{E}\big[\big(f(X_0)-g(X_0)\big)^2\big]$, where $X_0$ is uniformly distributed on $\mathbb{X}$. Guntuboyina and Sen (2013) proved[4] that

$$c_{l,d}\,(B/\epsilon)^{d/2} \le \mathcal{H}(\epsilon, \mathcal{C}_{\mathbb{X},B,L}, \|\cdot\|_{X_0}) \le c_{u,d}\,(B/\epsilon)^{d/2}\,,$$

for $\mathbb{X} = [0,1]^d$, $L \ge 2/d$ and $\epsilon \in (0, \epsilon_{0,d}B]$, where $c_{l,d}, c_{u,d}, \epsilon_{0,d}$ are positive constants depending on the domain dimension $d$. Unfortunately, this dependence is not known precisely, so we provide the lower bound only asymptotically for $n$, treating $d$ as a constant. Combining the above result with Theorem 6 of Yang and Barron (1999), we get that the minimax error with a sample size of $n$ is lower bounded by $\Omega(\epsilon^2)$, where $\epsilon$ is the solution of the equation $(B/\epsilon)^{d/2} = n\,\epsilon^2$, thus implying the following result:

**Theorem 4.1.** *For any class of distributions $\mathcal{D} \supseteq \mathcal{D}_\sigma$, $L_n(\mathcal{D}) = \Omega\big(n^{-4/(d+4)}\big)$.*

Next, we wish to study the $L_2$-error of LSEs with various choices of $\mathcal{F}$. The simplest, reasonable choice is $\mathcal{F} = \mathcal{C}_{\mathbb{X},B,L}$, and the plan is to use Theorem 3.1 to bound the $L_2$-error of this estimator. For this, we need an upper bound on $\mathcal{H}_\infty(\epsilon, \mathcal{C}_{\mathbb{X},B,L})$. Such a bound was given by Bronshtein (1976, Theorem 6) and further improved by Guntuboyina and Sen (2013, Theorem 3.2). However, their results are valid only for a small range of $\epsilon$. As such, these results would provide an upper bound where the "constant" term (i.e., the term independent of $n$) would be exponentially large in $d$. To avoid this, in the next section we prove a new bound for $\mathcal{H}_\infty(\epsilon, \mathcal{C}_{\mathbb{X},B,L})$ which improves the dependence on $d$ at the price of an extra $\ln(n)$ factor.

Besides proving this entropy upper bound, the next section provides approximation and covering results for $\mathcal{C}_{\mathbb{X},B,L}$ based on max-affine functions, which can be of independent interest. These results will be used later to prove near-optimality of some max-affine estimators for convex regression problems over $\mathcal{C}_{\mathbb{X},B,L}$.

[4] Guntuboyina and Sen (2013) proved the lower bound without the Lipschitz bound, which is a larger function class. However, in the proof of their Theorem 3.3, they construct a packing subset by functions having $(2/d)$-bounded Lipschitz constants. For the upper bound, simply consider the sup-norm result, Theorem 3.2 in their paper.

### 4.1 Approximation and covering results

Define the class of max-affine, uniformly bounded, uniformly Lipschitz functions on $\mathbb{X}$ having at most $K \in \mathbb{N}$ hyperplanes as

$$\begin{aligned}\mathcal{M}_{\mathbb{X},B,L}^K \doteq \Big\{h : \mathbb{X}\to\mathbb{R} \,\Big|\, h(x) = \max_{k=1,\dots,K} p_k^\top x + q_k,\\ p_k, q_k \in \mathbb{R}^d,\ \|p_k\|_\infty \le L,\ h(x) \in [-B_d, B]\Big\}\end{aligned}$$

with $B_d \doteq B + L_d$ and $L_d \doteq dL\mathrm{diam}\,(\mathbb{X})$. Now consider the following result bounding the approximation error of $\mathcal{M}_{\mathbb{X},B,L}^K$ to functions in $\mathcal{C}_{\mathbb{X},B,L}$ with respect to $\|\cdot\|_\infty$.

**Lemma 4.1.** *For all $f \in \mathcal{C}_{\mathbb{X},B,L}$ and $K \in \mathbb{N}$,*

$$\inf_{h \in \mathcal{M}_{\mathbb{X},B,L}^K} \|f-h\|_\infty \le 72 L_d\, K^{-2/d}\,.$$

*Proof.* The proof is based on ideas of Bronshteyn and Ivanov (1975). For any $x \in \mathbb{X}$, let $\nabla f(x) \in \partial f(x)$ be an arbitrary fixed subgradient of $f$ at $x$; recall that $\|\nabla f(x)\|_\infty \le L$.

For any $t > 0$, define $R \doteq \mathrm{diam}\,(\mathbb{X}) + 2tL$, $\nu(x) \doteq x + t\,\nabla f(x)$ for any $x \in \mathbb{X}$, and $\mathcal{K} \doteq \{\nu(x) : x \in \mathbb{X}\} \subseteq \mathbb{R}^d$. Notice that $R \ge \mathrm{diam}\,(\mathcal{K})$ and $\nu(x) \ne \nu(y)$ for any $x \ne y$ by the convexity of $f$. Furthermore, let $\mathcal{K}_\epsilon \subseteq \mathcal{K}$ be a $\sqrt{\epsilon}$-net of $\mathcal{K}$ with respect to the Euclidean norm $\|\cdot\|$ and $\mathbb{X}_\epsilon \doteq \{\nu^{-1}(z) \in \mathbb{X} : z \in \mathcal{K}_\epsilon\}$. Then by $\|\cdot\| \le \sqrt{d}\,\|\cdot\|_\infty$ and Lemma A.7, we have $|\mathbb{X}_\epsilon| = |\mathcal{K}_\epsilon| = \mathcal{N}(\sqrt{\epsilon}, \mathcal{K}, \|\cdot\|) \le \mathcal{N}(\sqrt{\epsilon/d}, \mathcal{K}, \|\cdot\|_\infty) \le (9dR^2/\epsilon)^{d/2}$ for all $\epsilon \in (0, 9dR^2]$.

Now, for any $x, z \in \mathbb{X}$, the convexity of $f$ implies

$$\begin{aligned}&\big(\nabla f(x) - \nabla f(z)\big)^\top (x-z)\\ &\quad= \nabla f(x)^\top (x-z) + \nabla f(z)^\top (z-x)\\ &\quad\ge f(x) - f(z) + f(z) - f(x) = 0\,.\end{aligned} \tag{2}$$

By definition, for any $x \in \mathbb{X}$ there exists $\hat{x} \in \mathbb{X}_\epsilon$ such that $\|\nu(x) - \nu(\hat{x})\| \le \sqrt{\epsilon}$. Then by (2) we have

$$\begin{aligned}&\|x-\hat{x}\|^2 + t^2\,\|\nabla f(x) - \nabla f(\hat{x})\|^2\\ &\quad\le \|x-\hat{x}\|^2 + 2t\big(\nabla f(x) - \nabla f(\hat{x})\big)^\top (x-\hat{x})\\ &\qquad\qquad + t^2\,\|\nabla f(x) - \nabla f(\hat{x})\|^2\\ &\quad= \|\nu(x) - \nu(\hat{x})\|^2 \le \epsilon\,.\end{aligned}$$

Hence, $\|x-\hat{x}\| \le \sqrt{\epsilon}$ and $t\,\|\nabla f(x) - \nabla f(\hat{x})\| \le \sqrt{\epsilon}$.

Choose $\epsilon$ satisfying $K = (9dR^2/\epsilon)^{d/2} \ge |\mathbb{X}_\epsilon|$ and a set $\mathbb{X}_K \doteq \{\hat{x}_1, \dots, \hat{x}_K\} \subseteq \mathbb{X}$ such that $\mathbb{X}_\epsilon \subseteq \mathbb{X}_K$. Note

that $\hat{x} \in \mathbb{X}_K$ for any $x \in \mathbb{X}$. Consider the following max-affine function $h : \mathbb{X} \to \mathbb{R}$:

$$h(x) \doteq \max_{k=1,\dots,K} f(\hat{x}_k) + \nabla f(\hat{x}_k)^\top (x - \hat{x}_k)\,.$$

By the convexity of $f$, we have $h(x) \le f(x) \le B$; furthermore, by the Cauchy-Schwartz inequality and $\|\cdot\| \le \sqrt{d}\,\|\cdot\|_\infty$,

$$\begin{aligned} h(x) &\ge f(\hat{x}) + \nabla f(\hat{x})^\top (x - \hat{x}) \\ &\ge -B - d\,\|\nabla f(\hat{x})\|_\infty \,\|x - \hat{x}\|_\infty \ge -B_d\,, \end{aligned}$$

so $h \in \mathcal{M}^K_{\mathbb{X},B,L}$. Moreover, the convexity of $f$ and the Cauchy-Schwartz inequality also imply

$$\begin{aligned} 0 &\le f(x) - h(x) \\ &\le f(x) - f(\hat{x}) - \nabla f(\hat{x})^\top (x - \hat{x}) \\ &\le \nabla f(x)^\top (x - \hat{x}) - \nabla f(\hat{x})^\top (x - \hat{x}) \\ &= \big(\nabla f(x) - \nabla f(\hat{x})\big)^\top (x - \hat{x}) \\ &\le \frac{1}{t}\Big(t\,\|\nabla f(x) - \nabla f(\hat{x})\|\,\|x - \hat{x}\|\Big) \le \epsilon/t\,. \end{aligned}$$

Finally, rearranging $K = (9dR^2/\epsilon)^{d/2}$, we get the claim by $\epsilon = 9dR^2K^{-2/d}$ and $\|f - h\|_\infty \le \epsilon/t = 9d(R^2/t)K^{-2/d} = 72L_dK^{-2/d}$ by setting $t = \operatorname{diam}(\mathbb{X})/(2L)$. □

Soon, we shall use this approximation bound on $\mathcal{M}^K_{\mathbb{X},B,L}$ to construct an $\epsilon$-net of $\mathcal{C}_{\mathbb{X},B,L}$. For this, we need a cover of $\mathcal{M}^K_{\mathbb{X},B,L}$ first.

**Lemma 4.2.** *For all $k \in \mathbb{N}$, $R_d \ge 2B + 4L_d$, and $\epsilon \in (0, R_d]$,*

$$\mathcal{H}_\infty(\epsilon, \mathcal{M}^K_{\mathbb{X},B,L}) \le (d+1)K \ln(R_d/\epsilon)\,.$$

*Proof.* Consider any $h \in \mathcal{M}^K_{\mathbb{X},B,L}$, and recall that $h(x) = \max_{k=1\dots,K} p_k^\top x + q_k$. Fix any $x_0 \in \mathbb{X}$ and define $r_k \doteq q_k + p_k^\top x_0$. Then $h(x) = \max_{k=1,\dots,K} p_k^\top (x - x_0) + r_k$, where $r_k \le \max_j r_j = h(x_0) \le B$. Furthermore, without loss of generality, we assume that for every $1 \le k \le K$, there is an $x_k \in \mathbb{X}$ such that $h(x_k) = p_k^\top (x_k - x_0) + r_k$. Then, by $h(x_k) \ge -B_d$ and $|p_k^\top (x_k - x_0)| \le d\,\|p_k\|_\infty \,\|x_k - x_0\|_\infty \le L_d$, we also have $r_k = h(x_k) - p_k^\top (x_k - x_0) \ge -B_d - L_d$.

Using Lemma A.7 for rectangular sets, we take an $\epsilon_1$-cover of $[-L, L]^d$, an $\epsilon_2$-cover of $[-B_d - L_d, B]$ with cardinalities no more than $(2L/\epsilon_1)^d$, $(2B + 2L_d)/\epsilon_2$, respectively, with $\epsilon_1 \in (0, 2L]$, $\epsilon_2 \in (0, 2B + 2L_d]$. Now let $\mathcal{M}^K_{\mathbb{X},B,L}(\epsilon_1, \epsilon_2)$ denote the set of functions $\hat{h}(x) = \max_{j=1,\dots,K} \hat{p}_j^\top (x - x_0) + \hat{r}_j$, where $\hat{p}_j$ and $\hat{r}_j$ belong to the aforementioned two nets, respectively. Then $|\mathcal{M}^K_{\mathbb{X},B,L}(\epsilon_1, \epsilon_2)| \le (2L/\epsilon_1)^d (2B + 2L_d)/\epsilon_2$. In what follows we show that $|\mathcal{M}^K_{\mathbb{X},B,L}(\epsilon_1, \epsilon_2)|$ provides a good cover for $\mathcal{M}^K_{\mathbb{X},B,L}$.

For any $h \in \mathcal{M}^K_{\mathbb{X},B,L}$ and $k = 1, \dots, K$, let $\hat{p}_k$, $\hat{r}_k$ be the closest elements in the nets to $p_k$ and $r_k$, respectively, and define $\hat{h}(x) \doteq \max_{j=1,\dots,K} \hat{p}_j^\top (x - x_0) + \hat{r}_j$. If $h(x) \ge \hat{h}(x)$, we have

$$\begin{aligned} &h(x) - \hat{h}(x) \\ &\quad\le \max_{k=1,\dots,K} p_k^\top (x - x_0) + r_k - \big(\hat{p}_k^\top (x - x_0) + \hat{r}_k\big) \\ &\quad\le \max_{k=1,\dots,K} d\,\operatorname{diam}(\mathbb{X})\,\|p_k - \hat{p}_k\|_\infty + |r_k - \hat{r}_k| \\ &\quad\le d\,\operatorname{diam}(\mathbb{X})\,\epsilon_1 + \epsilon_2\,. \end{aligned}$$

If $h(x) < \hat{h}(x)$, an analogous proof gives the same bound for $\hat{h}(x) - h(x)$. Hence, setting $\epsilon_1 \doteq \eta\,\epsilon/(d\operatorname{diam}(\mathbb{X}))$ and $\epsilon_2 \doteq (1 - \eta)\,\epsilon$ for some $\eta \in (0, 1)$, we get $|h(x) - \hat{h}(x)| \le \epsilon$. Finally, setting $\eta \doteq 2L_d/R_d$, we get the claim. □

Notice that providing a tight enough covering to a well approximating $\mathcal{M}^K_{\mathbb{X},B,L}$ (i.e., having large enough $K$) gives us a cover of $\mathcal{C}_{\mathbb{X},B,L}$ formed by only max-affine functions. The details are presented in the next result.

**Lemma 4.3.** *Let $R^*_d \doteq \max\{8L_d, 2B + 4L_d\}$. Then for all $\epsilon \in (0, 80L_d]$,*

$$\mathcal{H}_\infty(\epsilon, \mathcal{C}_{\mathbb{X},B,L}) \le 2(d+1)\left(\frac{80L_d}{\epsilon}\right)^{d/2} \ln\left(\frac{10R^*_d}{\epsilon}\right)\,.$$

*Proof.* Set $\lambda \doteq 9/10$ and take any $\epsilon \in (0, 80L_d]$. Having $\lambda\epsilon \in (0, 72L_d]$, we can pick $K \in \mathbb{N}$ such that $K \ge \big(72L_d/(\lambda\epsilon)\big)^{d/2} \ge K/2$. Then for any $f \in \mathcal{C}_{\mathbb{X},B,L}$, let $h_f \in \mathcal{M}^K_{\mathbb{X},B,L}$ be the best approximation of $f$. Furthermore, let $\hat{h}_f$ be the best approximation of $h_f$ in the $(1 - \lambda)\epsilon$-cover of $\mathcal{M}^K_{\mathbb{X},B,L}$. By using Lemma 4.1, we get

$$\begin{aligned} \|f - \hat{h}_f\|_\infty &\le \|f - h_f\|_\infty + \|h_f - \hat{h}_f\|_\infty \\ &\le 72L_d\,K^{-2/d} + (1 - \lambda)\epsilon \le \epsilon\,. \end{aligned}$$

Finally, Lemma 4.2 (extended to $R^*_d \ge R_d$) provides the bound with $(1 - \lambda)\epsilon = \epsilon/10 \in (0, 8L_d]$, $8L_d \le R^*_d$ and $K \le 2\big(80L_d/\epsilon\big)^{d/2}$. □

### 4.2 Least squares estimators over $\mathcal{C}_{\mathbb{X},B,L}$

In this section we consider the class of $\alpha$-LSE($\mathcal{C}_{\mathbb{X},B,L}$). We prove that any such estimator is near-optimal for $d \in \{1, 2, 3, 4\}$, while we get suboptimal bounds for $d > 4$. A similar "phase-transition" was previously noted for many other regression problems (e.g., van de Geer, 2000), and the usual recommendation is to "regularize" the LSE for larger dimensions. We shall consider this problem in Section 4.4.

To get the upper bound for an $\alpha$-LSE($\mathcal{C}_{\mathbb{X},B,L}$), we simply plug the covering number result of $\mathcal{C}_{\mathbb{X},B,L}$

(Lemma 4.3) into our regression bound (Theorem 3.1) and set the $(\epsilon,\delta)$ parameters properly to balance $n$ in the terms. For the case $d < 4$, the optimal rate for $n$ comes from (B) choosing $\epsilon$ to balance

$$\frac{\mathcal{H}_\infty(\epsilon, \mathcal{C}_{\mathbb{X},B,L})}{n} \approx \int_0^\epsilon \sqrt{\frac{\mathcal{H}_\infty(s, \mathcal{C}_{\mathbb{X},B,L})}{n}}\, ds\,.$$

For the cases $d \geq 4$, the entropy integral diverges and becomes the dominant term in the upper bound. Then the best ratio for $n$ is obtained by (A) choosing $\delta$ that solves

$$\delta \approx \int_\delta^B \sqrt{\frac{\mathcal{H}_\infty(s, \mathcal{C}_{\mathbb{X},B,L})}{n}}\, ds\,.$$

The balancing factors and the corresponding asymptotic formulas are provided for any $f_n$ being an $\alpha$-LSE$(\mathcal{C}_{\mathbb{X},B,L})$ and $n > 1$. More precise formulas are presented in Lemma A.8.

**Case $d < 4$:** Use Theorem 3.1 (B) and balance for $n$ with $\epsilon = 80L_d\, n^{-2/(d+4)}$ and $\delta = 0$. Then

$$\mathbb{E}\Big[\, \|f_n - f_*\|_\mu^2 \,\Big] = O\Big(n^{-4/(d+4)}\, \ln(n)\Big) + 2\alpha\,.$$

**Case $d = 4$:** Use Theorem 3.1 (A) and balance for $n$ with $\delta = 80L_4\, n^{-1/2}$. Then

$$\mathbb{E}\Big[\, \|f_n - f_*\|_\mu^2 \,\Big] = O\Big(n^{-1/2}\, \ln^{3/2}(n)\Big) + \alpha\,.$$

**Case $d > 4$:** Use Theorem 3.1 (A) and balance for $n$ with $\delta = 80L_d\, n^{-2/d}$. Then

$$\mathbb{E}\Big[\, \|f_n - f_*\|_\mu^2 \,\Big] = O\Big(n^{-2/d}\sqrt{d}\big(1 + \ln(n)/d\big)\Big) + \alpha\,.$$

Notice that we do not have any "constant" scaling exponentially in $d$. This is due to the wide range of $\epsilon$ in Lemma 4.3, and would not be possible with a limited $\epsilon \in (0, \epsilon_0]$ with $\epsilon_0 < 80L_d$.

### 4.3 Max-affine least squares estimators

It is widely known that one can find LSEs over all convex functions by considering only max-affine functions having $n$ hyperplanes (see, e.g., Holloway, 1979; Boyd and Vandenberghe, 2004, Section 6.5.5; Kuosmanen, 2008; Seijo and Sen, 2011; Lim and Glynn, 2012). However, such an estimator is not bounded and can easily overfit the data close to the domain boundary. As a result, this estimator has infinite expected $L_2$-error in many cases. Next we demonstrate this on a simple example, similar to Example 3.5 of Huang and Szepesvári (2014).

Let $\mathbb{X} = [0,1]$, $X \in \mathbb{X}$, $Y \in \{-1,+1\}$ be independent uniform random variables (so that $f_* \equiv 0$), $n \geq 2$, $D_n$ be the random sample as before, $f_n$ be a LSE$(\{f : \mathbb{X} \to \mathbb{R} \,|\, f \text{ is convex}\})$. Define the event

$$\begin{aligned} A = \big\{ & X_1 \in [1/4, 1/2], X_2 \in [1/2, 3/4], \\ & X_3, \ldots, X_n \geq 3/4,\ X \leq 1/4, \\ & Y_1 = +1, Y_2 = \ldots = Y_n = -1 \big\}\,. \end{aligned}$$

Then $\mathbb{P}\{A\} = (1/4)^{n+1}(1/2)^n > 0$, $f_* \equiv 0$, and so, using the observation that the LSE minimizing the test error on $[0, 1/4]$ is linear in $[0, X_2]$, we get

$$\begin{aligned} \mathbb{E}\big[\, \|f_n - f_*\|_\mu^2 \,\big] &\geq \mathbb{E}\left[\left(\frac{2X - X_1 - X_2}{X_1 - X_2}\right)^2 \,\middle|\, A\right] \mathbb{P}\{A\} \\ &\geq \mathbb{E}\left[\frac{1}{4(X_1 - X_2)^2} \,\middle|\, A\right] \mathbb{P}\{A\} = \infty\,. \end{aligned}$$

Although event $A$ is quite unlikely, empirically one can also observe that the test error of this LSE is quite often very large.

The main attraction of LSE$(\{f \mid f : \mathbb{X} \to \mathbb{R} \text{ convex}\})$ is that it leads to a convex optimization problem, which can be solved in polynomial time. To prevent the unbounded expected $L_2$-error, one idea is to consider max-affine estimates in LSE$(\mathcal{C}_{\mathbb{X},B,L})$ as these were shown to enjoy controlled expected $L_2$-error in Section 4.2. To see that there are indeed max-affine estimates in LSE$(\mathcal{C}_{\mathbb{X},B,L})$, take any estimate $f_n^*$ in LSE$(\mathcal{C}_{\mathbb{X},B,L})$ and construct

$$\hat{f}_n(x) \doteq \max_{i=1,\ldots,n} f_n^*(X_i) + g_i^\top (x - X_i)\,,$$

where $g_i \in \partial f_n^*(X_i)$. Then define the estimator as $f_n(x) \doteq \max\{-B, \hat{f}_n(x)\}$ being the lower truncated version of $\hat{f}_n$. Now notice that $\hat{f}_n \in \mathcal{M}^n_{\mathbb{X},B,L}$, $f_n \in \mathcal{C}_{\mathbb{X},B,L}$ and $f_n(X_i) = \hat{f}_n(X_i) = f_n^*(X_i) \in [-B, B]$. So the empirical risks of $f_n^*$ and $f_n$ must be equal, and so $f_n$ belongs to LSE$(\mathcal{C}_{\mathbb{X},B,L})$.

Let us now consider the problem of efficiently computing a max-affine estimate in LSE$(\mathcal{C}_{\mathbb{X},B,L})$. For a rectangular domain $\mathbb{X}$, we will show below that one can compute an $\hat{f}_n \in \mathcal{M}^n_{\mathbb{X},B,L}$ estimate by solving a quadratic program (QP) similar to the one computing unbounded LSEs. Then this function $\hat{f}_n$ is converted to an LSE$(\mathcal{C}_{\mathbb{X},B,L})$ by lower truncation. To see this, let $\mathbb{X} \doteq \times_{i=1}^d [\mathbf{l}_i, \mathbf{u}_i]$ with some $\mathbf{l}, \mathbf{u} \in \mathbb{R}^d$, $\mathbf{l} \leq \mathbf{u}$, and split the subgradients, $g_i = g_i^+ - g_i^-$, $g_i^+, g_i^- \geq 0$. Consider max-affine estimators given as

$$\hat{f}_n(x) = \max_{i=1,\ldots,n} y_i + (g_i^+ - g_i^-)^\top (x - X_i)\,,$$

with some $y_i \in [-B, B]$ and $g_i^+, g_i^- \in [0, L]^d$. Then notice that we can rewrite $\max_{x \in \mathbb{X}} \hat{f}_n(x) \leq B$ as linear constraints, and $\{y_i \geq -B : i = 1, \ldots, n\}$ implies

$f_n(x) \geq -B_d$. So we can compute a LSE($\mathcal{M}^n_{\mathbb{X},B,L}$) by the following QP:

$$\min_{\substack{y\in\mathbb{R}^n,\\ g_1^+,\dots,g_n^+\in\mathbb{R}^d,\\ g_1^-,\dots,g_n^-\in\mathbb{R}^d}} \sum_{i=1}^n (Y_i - y_i)^2 \quad \text{subject to}$$
$$\begin{aligned} & y_k \geq y_i + (g_i^+ - g_i^-)^\top (X_k - X_i)\,, \\ & B \geq y_i + (g_i^+)^\top(\mathbf{u} - X_i) + (g_i^-)^\top(X_i - \mathbf{l})\,, \\ & 0 \leq g_{ij}^+,\, g_{ij}^- \leq L\,, \quad -B \leq y_i\,, \\ & i,k = 1,\dots,n\,, \quad j = 1,\dots,d\,, \end{aligned} \tag{3}$$

where $y = [y_1 \dots y_n]^\top$ and $\{(X_i, Y_i) : i = 1,\dots,n\}$ is the data as above.

This QP has $n(1+2d)$ variables and $n^2$ non-box constraints, which makes it expensive to solve directly. Instead, we propose to apply cutting plane techniques similar to the CNLS$^+$-G algorithm given by Lee et al. (2013). These methods solve the QP problem using a relaxed constraint set and introduce new ones iteratively as necessary. We present our method in Appendix A.3.

Before discussing empirical studies in Section 5, we consider another class of max-affine estimators.

### 4.4 Near-optimal max-affine estimators

Section 4.2 provides only a suboptimal upper bound for LSE($\mathcal{M}^n_{\mathbb{X},B,L}$) estimators for $d > 4$. Here, we show that max-affine LSEs using no more than $\lceil n^{d/(d+4)} \rceil$ hyperplanes (instead of $n$), enjoy an optimal minimax rate up to a logarithmic factor. By reducing the number of planes, we decrease the estimation and increase the approximation error. The optimal rate is achieved when these two effects are balanced.

To show this formally, we combine Theorem 3.1 (B) with the covering number bound of max-affine functions (Lemma 4.2) and their approximation accuracy (Lemma 4.1). Let $f_n$ be an $\alpha$-LSE($\mathcal{M}^K_{\mathbb{X},B,L}$) with at most $K \doteq \lceil n^{d/(d+4)} \rceil$ hyperplanes and $\alpha \leq n^{-4/(d+4)}$. Then setting $\epsilon = \delta = R_d\, n^{-4/(d+4)}$ and $(c_1, c_2, c_3, c_4) = (12, 153, 171, 9/4)$, we get

$$\begin{aligned} \mathbb{E}\Big[\, \|f_n - f_*\|_\mu^2 \,\Big] &\leq 12 B_\sigma^2 \frac{(d+1)K \ln\left(R_d/\epsilon\right)}{n} \\ &\quad + 171 B_\sigma \delta + \frac{9}{4}\Big((72 L_d)^2 K^{-4/d} + \alpha\Big) \\ &< 96\Big(B_\sigma^2 \ln(n) + 2 B_\sigma R_d + 122 L_d^2 + 1\Big)\, n^{-4/(d+4)}\,, \end{aligned}$$

which provides the following result:[5]

[5] A similar rate was shown for convex set estimation by Guntuboyina (2012).

**Theorem 4.2.** *Suppose that the conditions of Theorem 3.1 hold for $\mu$ with some $f_* \in \mathcal{C}_{\mathbb{X},B,L}$, and $f_n$ is an $\alpha$-LSE($\mathcal{M}^K_{\mathbb{X},B,L}$) with $K = \lceil n^{d/(d+4)} \rceil$ and $\alpha = \mathcal{O}\big(n^{-4/(d+4)}\big)$. Then*

$$\mathbb{E}\Big[\, \|f_n - f_*\|_\mu^2 \,\Big] = \mathcal{O}\Big(n^{-4/(d+4)}\big(\ln(n) + d^2\big)\Big)\,.$$

Notice that LSEs with $n$ planes use a "degenerate" partitioning, where each data point $X_i$ forms a partition. Reducing the number of planes below $n$ induces a partition where some of the data points are grouped. When such a partition is given, the LSE problem can be solved, similar to (3).

Let $P_n \doteq \{C_1, \dots, C_K\}$ be a partition of $\{1,\dots,n\}$, that is, for each cell $C_k \subseteq \mathbb{N}$ with $n = \sum_{k=1}^K |C_k|$, and $k \neq l \iff C_k \cap C_l = \emptyset$ for all $k, l \in \{1,\dots,K\}$. Furthermore, let $\mathbb{X} \doteq [\mathbf{l}, \mathbf{u}] \subseteq \mathbb{R}^d$ as before and denote the centroid of cell $k$ by $\bar{X}_k \doteq |C_k|^{-1} \sum_{k=1}^K X_k$. Then a (non-truncated) LSE over $P_n$ is formed as

$$\hat{f}_K(x) = \max_{k=1,\dots,K} v_k + (g_k^+ - g_k^-)^\top (x - \bar{X}_k)\,,$$

where the variables $v \doteq [v_1 \dots v_K]^\top \in [-B, B]^K$, $g_k^+, g_k^- \in [0, L]^d$ $(k = 1,\dots,K)$ can be computed by the following QP,

$$\begin{aligned} &\min_{\substack{v\in\mathbb{R}^K,\\ g_1^+,\dots,g_K^+\in\mathbb{R}^d,\\ g_1^-,\dots,g_K^-\in\mathbb{R}^d}} \sum_{k=1}^K \sum_{i\in C_k} (Y_i - y_i)^2 \\ &\text{with } \; y_i = v_k + (g_k^+ - g_k^-)^\top (X_i - \bar{X}_k) \\ &\text{subject to} \\ &\quad v_k + (g_k^+ - g_k^-)^\top (X_i - \bar{X}_k) \\ &\qquad \geq v_l + (g_l^+ - g_l^-)^\top (X_i - \bar{X}_l)\,, \quad i \in C_k\,, \\ &\quad B \geq v_k + (g_k^+)^\top(\mathbf{u} - \bar{X}_k) + (g_k^-)^\top(\bar{X}_k - \mathbf{l})\,, \\ &\quad 0 \leq g_{kj}^+,\, g_{kj}^- \leq L\,, \quad -B \leq y_k\,, \\ &\quad k,l = 1,\dots,K\,, \quad j = 1,\dots,d\,. \end{aligned} \tag{4}$$

Finally, the estimator is given by lower truncation, $f_K(x) = \max\{-B, \hat{f}_K(x)\}$, $x \in \mathbb{X}$. Notice that (4) reduces to (3) when $K = n$ and all $|C_k| = 1$. Furthermore, the computation of this QP can also be improved by using cutting plane methods as mentioned before.

As finding the best partition (yielding the smallest training error) is too difficult, in our experiments we simply draw one uniformly from the data. We draw the index set $\{i_1,\dots,i_K\} \subseteq \{1,\dots,n\}$ selecting the center points $X_{i_k}$ of the cells $C_k$, which form a Voronoi partition, $C_k \doteq \big\{ j \in \{1,\dots,n\} \,\big|\, \|X_{i_k} - X_j\| = \min_{l=1,\dots,K} \|X_{i_l} - X_j\| \big\}$, for all $k = 1,\dots,K$, and the unlikely ties are broken arbitrarily. Surprisingly, this simple partitioning technique worked quite well in our experiments.

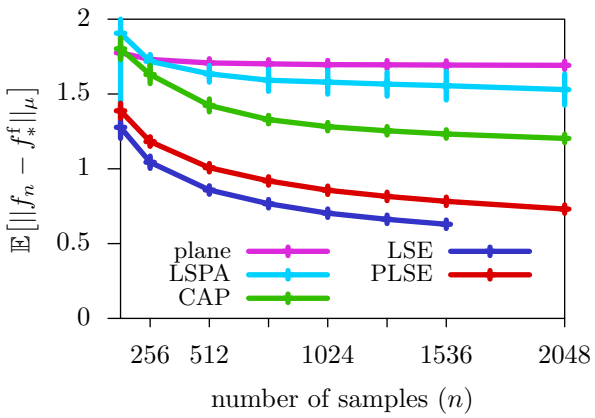


Figure 1: Full quadratic problem ($f_*^{\mathrm{f}}$) with $d = 8$.

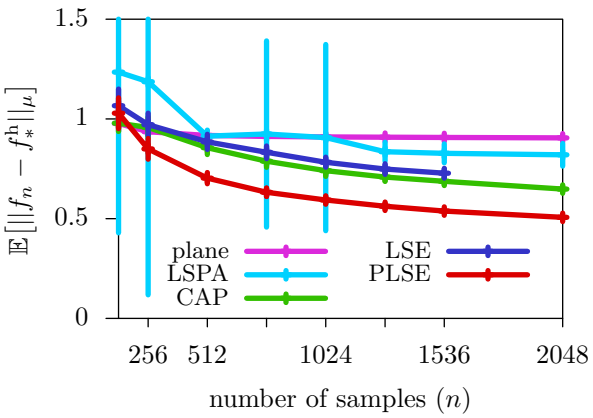


Figure 2: Half quadratic problem ($f_*^{\mathrm{h}}$) with $d = 8$.

## 5 EXPERIMENTS

To illustrate the behavior of the algorithms proposed and to compare them to some state-of-the-art alternatives, we present experiments with synthetic data. Let $\mathbb{X} \doteq [-2, 2]^d$, $X$ be a uniform random variable on $\mathbb{X}$ and $Y = f_*(X) + \xi$, where $\xi \sim \mathcal{N}(0, 1)$ is a standard normal random variable, independent from $X$. Consider the following two problems:

$$f_*^{\mathrm{f}}(x) \doteq -B + \frac{B\,\|x\|^2}{2d}\,,\; f_*^{\mathrm{h}}(x) \doteq -B + \frac{B\,\|(x)_+\|^2}{2d}\,, \quad (5)$$

where $y = (x)_+$ is the positive part, $y_i = \max(0, x_i)$, and $B \doteq 8$. We refer to these problems as "full quadratic" and "half quadratic", respectively.

We point out that our choice, the quadratic function, is a difficult target for max-affine estimators. Based on the remark of Bronshteyn and Ivanov (1975) about the approximation lower bound of polyherdal sets, it seems that Lemma 4.1 is tight for quadratic functions (up to some constants).

We performed the experiments with the max-affine LSE using at most $n$ planes (LSE, (3)), its partitioned version using at most $K = \lceil n^{d/(d+4)} \rceil$ planes (PLSE, (4)), where the Voronoi partition was chosen by uniformly drawing $K$ points from the training data. For a simple benchmark, we included a single fitted hyperplane (plane), which is a LSE($\{x \mapsto a^\top x + b\}$). Furthermore, we included two state of the art max-affine estimators, the convex adaptive partitioning (CAP, Hannah and Dunson 2013) and the least-squares partition algorithm[6] (LSPA, Magnani and Boyd 2009). We repeated each experiment 100 times and computed the test errors using $10^6$ samples.

[6]We initialized LSPA randomly using $n$ and $\lceil n^{d/(d+4)} \rceil$ cells, resp., ran the algorithm for 10000 iterations and returned the solution having the smallest training error found over these iterations. The figures show results only for the $\lceil n^{d/(d+4)} \rceil$ case, but the results were similar in the other case, as well.

Figure 1 shows the test error against the sample size for the full quadratic problem, while the same data is shown for the half quadratic problem on Figure 2. The error bars show standard deviation. While LSE performs really well on the full quadratic which grows at the boundaries, its performance, relative to the other methods is much worse on the half quadratic.[7] This is because LSE tends to overfit the noise on the flat side, occasionally creating hyperplanes with a large growth. This is prevented by the point subsampling scheme in PLSE. Overall, we find that PLSE is a competitive algorithm, at least on these examples.

## 6 CONCLUSIONS

We have given new results for bounded convex regression problems, resolving the open question of designing least-squares estimators with near-optimal rates. This is achieved by proving new results both in nonlinear least-squares estimation and convex approximations. Probably the most interesting open question is to design a computationally efficient, provably optimal estimator for this case, and perhaps our sampling based approximation to the near-optimal LSE can be used as the basis of such a method. Our preliminary experimental results indicate that, despite its simplicity, this method can be beneficial as compared to either the max-affine LSEs with $n$ hyperplanes, or its more advanced alternatives.

### Acknowledgements

This work was supported by the Alberta Innovates Technology Futures and the National Science and Engineering Research Council of Canada.

[7]The reason for stopping the LSE curves is that their calculations for 100 repetitions would have taken more than a month. PLSE runs much faster, at least on this problem size ($d = 8$). Comparison of LSE and PLSE running times for these experiments can be found in Appendix A.4.

## References


Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Boucheron, S., Lugosi, G., and Massart, P. (2012). *Concentration Inequalities: A nonasymptotic theory of independence*. Clarendon Press.

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

Bronshtein, E. M. (1976). $\epsilon$-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17(3):508–514.

Bronshteyn, E. M. and Ivanov, L. D. (1975). The approximation of convex sets by polyhedra. *Siberian Mathematical Journal*, 16(5):852–853.

Buldygin, V. V. and Kozachenko, Y. V. (2000). *Metric characterization of random variables and random processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society.

Cesa-Bianchi, N. and Lugosi, G. (1999). Minimax regret under log loss for general classes of experts. *COLT*, pages 12–18.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Guntuboyina, A. (2012). Optimal rates of convergence for convex set estimation from support functions. *The Annals of Statistics*, 40(1):385–411.

Guntuboyina, A. and Sen, B. (2013). Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag.

Hannah, L. A. and Dunson, D. B. (2012). Ensemble methods for convex regression with applications to geometric programming based circuit design. *International Conference on Machine Learning*.

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

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

Holloway, C. A. (1979). On the estimation of convex functions. *Operations Research*, 27(2):401–407.

Huang, R. and Szepesvári, C. (2014). A finite-sample generalization bound for semiparametric regression: Partially linear models. *AISTATS*.

Koltchinskii, V. (2008). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *The Econometrics Journal*, 11(2):308–325.

Lecué, G. and Mendelson, S. (2013). Learning subgaussian classes: Upper and minimax bounds. http://arxiv.org/abs/1305.4825.

Lee, C.-Y., Johnson, A. L., Moreno-Centeno, E., and Kuosmanen, T. (2013). A more efficient algorithm for convex nonparametric least squares. *European Journal of Operational Research*, 227(2):391–400.

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

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

Mendelson, S. (2014). Learning without concentration. *JMLR: Workshop and Conference Proceedings*, 35:1–15.

Merton, R. C. (1992). *Continuous-Time Finance*. Wiley-Blackwell.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics.

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

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on Stochastic Programming, Modeling and Theory*. Society for Industrial and Applied Mathematics and the Mathematical Programming Society.

Srebro, N., Sridharan, K., and Tewari, A. (2012). Smoothness, low-noise and fast rates. *Advances in Neural Information Processing Systems*, 23.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

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

Varian, H. R. (1984). The nonparametric approach to production analysis. *Econometrica*, 52(3):579–598.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

# A Appendix

## A.1 Concentration inequalities

Let $\sigma \geq 0$. A random variable $W$ is called $\sigma$-subgaussian if $\sup_{s\in\mathbb{R}} \mathbb{E}\big[e^{s(W-\mathbb{E}[W])-s^2\sigma^2/2}\big] \leq 1$.

**Lemma A.1.** *Let $\mathcal{F}$ be a finite, nonempty set (i.e. $1 \leq |\mathcal{F}| < \infty$), $\sigma \in [0,\infty)$ and $W_f$ be a centered $\sigma$-subgaussian random variable for all $f \in \mathcal{F}$. Then $\mathbb{E}\big[\max_{f\in\mathcal{F}} W_f\big] \leq \sigma\sqrt{2\ln|\mathcal{F}|}$.*

*Proof.* The proof of the lemma is available in the literature (see, e.g., Cesa-Bianchi and Lugosi 1999, Lemma 7 or Boucheron et al. 2012, Theorem 2.5), and is provided for completeness. The claim is trivial if $|\mathcal{F}| = 1$ or $\sigma = 0$. When $|\mathcal{F}| > 1$ and $\sigma > 0$, applying Jensen's inequality, replacing the maximum of non-negative elements by their sum and using the condition on $W_f$, for any $t \in \mathbb{R}$, we get

$$\exp\left(t\,\mathbb{E}\left[\max_{f\in\mathcal{F}} W_f\right]\right) \leq \mathbb{E}\left[\max_{f\in\mathcal{F}} e^{t\,W_f}\right] \leq \sum_{f\in\mathcal{F}} \mathbb{E}\left[e^{t\,W_f}\right] \leq |\mathcal{F}|\, e^{t^2\sigma^2/2}\,.$$

Taking logarithm and dividing both sides by $t = \sqrt{2\ln|\mathcal{F}|}/\sigma > 0$, we get the claim. □

Let $(\mathcal{F},\ell)$ be a separable metric space, $W$ be a random variable taking values in the set $\mathbb{W}$ and $\phi : \mathcal{F}\times\mathbb{W} \to \mathbb{R}$ be a function. The following definitions will be useful for our purposes:

**Definition A.1** (Subgaussian Process)**.** *Let $\sigma \geq 0$. We call the random process $(\phi(f,W))_{f\in\mathcal{F}}$* $\sigma\ell$-subgaussian *if $\phi(f,W) - \phi(g,W)$ is a centered $(\sigma\,\ell(f,g))$-subgaussian random variable for all $f,g \in \mathcal{F}$.*

**Definition A.2** (Uniformly Lipschitz Process)**.** *We call the random process $(\phi(f,W))_{f\in\mathcal{F}}$ uniformly Lipschitz with respect to $\ell$ and (Lipschitz) modulus $\tau : \mathbb{W} \to [0,\infty)$ if $\phi(f,W) - \phi(g,W) \leq \ell(f,g)\,\tau(W)$ holds a.s. for all $f,g \in \mathcal{F}$.*

The following lemma gives a bound on the expectation of the supremum of the process $(\phi(f,W))_f$ over $f \in \mathcal{F}$ in terms of its entropy integral. The development is a modification of the proof of Lemma 3.4 of Pollard (1990) by replacing the packing numbers with internal covering numbers (for better numerical constants) and the sample continuity condition by Lipschitzness (for truncating the entropy integral at $\delta$). The result also improves upon Proposition 3 of Cesa-Bianchi and Lugosi (1999), which uses external covering numbers and a slightly different chaining argument.

**Lemma A.2.** *Let $(\mathcal{F},\ell)$ be a separable metric space, $W$ be a random variable taking values in the set $\mathbb{W}$ and $\phi : \mathcal{F}\times\mathbb{W} \to \mathbb{R}$ be a function such that:*

*(a) there exist $\beta \in [0,\infty)$ and $f_0 \in \mathcal{F}$ such that $\beta \geq \sup_{f\in\mathcal{F}} \ell(f,f_0)$ and $\mathbb{E}\big[\phi(f_0,W)\big] = 0$;*

*(b) $(\phi(f,W))_{f\in\mathcal{F}}$ is $\sigma\ell$-subgaussian for some $\sigma \geq 0$;*

*(c) $(\phi(f,W))_{f\in\mathcal{F}}$ is uniformly Lipschitz with respect to $\ell$ and modulus $\tau$ for some function $\tau : \mathbb{W} \to [0,\infty)$.*

*Then, for all $\delta \in [0,\beta/2]$,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \phi(f,W)\right] \leq 4\sqrt{2}\,\sigma \int_\delta^{\beta/2} \sqrt{\mathcal{H}(s,\mathcal{F},\ell)}\,ds + 4\delta\,\mathbb{E}[\tau(W)]\,.$$

*Proof.* Let $g \in \mathcal{F}$ and notice that by condition (c), we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \Big\{\phi(f,W) - \phi(g,W)\Big\}\right] \leq \left(\sup_{f\in\mathcal{F}} \ell(f,g)\right)\mathbb{E}[\tau(W)]\,. \tag{6}$$

Hence, by condition (a), the claim holds for $\delta = \beta/2$ (with $g = f_0$). Furthermore, if there exists $s \in (\delta,\beta/2]$ such that $\mathcal{N}(s,\mathcal{F},\ell) = \infty$, then the integral is infinite (since $\mathcal{N}(s,\mathcal{F},\ell)$ is a non-increasing function of $s$) and so the claim is trivial. Hence we can assume that $0 < \beta$, $\delta \in [0,\beta/2)$ and $\mathcal{N}(s,\mathcal{F},\ell) < \infty$ for all $s \in (\delta,\beta/2]$.

First consider the $\delta > 0$ case. Then there exists some $m \in \mathbb{N}\cup\{0\}$ such that $2\delta \leq \beta/2^m < 4\delta$. Now let $\mathcal{F}_0 = \{f_0\}$, $\epsilon_0 = \beta$, $\epsilon_k = \beta/2^k$ and $\mathcal{F}_k$ be an $\epsilon_k$-cover of $\mathcal{F}$ with respect to $\ell$ having minimal cardinality for all $k \in \{1,\ldots,m\}$.

Furthermore, let $g_k(f) \in \operatorname{argmin}_{g \in \mathcal{F}_k} \ell(f,g)$ be the closest element (or one of the closest elements if there are multiple ones) to $f \in \mathcal{F}$ in $\mathcal{F}_k$ for all $k \in \{0,\dots,m\}$.

Fix some $k \in \{0,\dots,m-1\}$ and $f \in \mathcal{F}_{k+1}$. When $k=0$, we have $\ell(f,g_k(f)) = \ell(f,f_0) \le \beta = \epsilon_0$, while for $k>0$, the definition of $\mathcal{F}_k$ implies that $\ell(f,g_k(f)) \le \epsilon_k$. So by condition (b), $\phi(f,W) - \phi(g_k(f),W)$ is a centered $\epsilon_k\sigma$-subgaussian random variable. Combining this with Lemma A.1, we can chain maximal inequalities for all $k \in \{0,\dots,m-1\}$ as

$$\begin{aligned}
\mathbb{E}\left[\max_{f\in\mathcal{F}_{k+1}} \phi(f,W)\right] &= \mathbb{E}\left[\max_{f\in\mathcal{F}_{k+1}} \Big\{\phi(g_k(f),W) + \phi(f,W) - \phi(g_k(f),W)\Big\}\right] \\
&\le \mathbb{E}\left[\max_{f\in\mathcal{F}_k} \phi(f,W)\right] + \mathbb{E}\left[\max_{f\in\mathcal{F}_{k+1}} \Big\{\phi(f,W) - \phi(g_k(f),W)\Big\}\right] \\
&\le \mathbb{E}\left[\max_{f\in\mathcal{F}_k} \phi(f,W)\right] + \epsilon_k\,\sigma\sqrt{2\ln|\mathcal{F}_{k+1}|} \\
&= \mathbb{E}\left[\max_{f\in\mathcal{F}_k} \phi(f,W)\right] + \epsilon_k\,\sigma\sqrt{2\ln\mathcal{N}(\epsilon_{k+1},\mathcal{F},\ell)}\,.
\end{aligned} \tag{7}$$

We further have

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \phi(f,W)\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \Big\{\phi\big(g_m(f),W\big) + \phi(f,W) - \phi\big(g_m(f),W\big)\Big\}\right] \\
&\le \mathbb{E}\left[\max_{f\in\mathcal{F}_m} \phi(f,W)\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \Big\{\phi(f,W) - \phi\big(g_m(f),W\big)\Big\}\right].
\end{aligned}$$

Using (6) with $\ell(f,g_m(f)) \le \varepsilon_m < 4\delta$, the second term can be bounded by $4\delta\mathbb{E}[\tau(W)]$. To bound the first term, we use (7) repeatedly with $k = m-1, m-2, \dots, 0$ and $\mathbb{E}\left[\max_{f\in\mathcal{F}_0}\phi(f,W)\right] = \mathbb{E}\left[\phi(f_0,W)\right] = 0$ for the last step (implied by condition (a) and the definition of $\mathcal{F}_0$), to get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \phi(f,W)\right] < \sigma \sum_{k=0}^{m-1} \epsilon_k \sqrt{2\ln\mathcal{N}(\epsilon_{k+1},\mathcal{F},\ell)} + 4\delta\,\mathbb{E}[\tau(W)]\,.$$

Now notice that the non-decreasing property of the covering number implies

$$\epsilon_k\sqrt{\ln\mathcal{N}\big(\epsilon_{k+1},\mathcal{F},\ell\big)} = 4\,\frac{\beta}{2^{k+2}}\sqrt{2\ln\mathcal{N}\big(\beta/2^{k+1},\mathcal{F},\ell\big)} \le 4\int_{\beta/2^{k+2}}^{\beta/2^{k+1}} \sqrt{2\ln\mathcal{N}\big(s,\mathcal{F},\ell\big)}\,ds\,,$$

for all $k \in \{0,\dots,m-1\}$. Combining this with $\delta \le \beta/2^{m+1}$ proves the claim for all $\delta \in (0,\beta/2)$.

Finally, taking the limit $\delta \downarrow 0$, we get the claim for $\delta = 0$ as well. □

**Lemma A.3.** *Let $\mathcal{F}$ be a finite, nonempty set (i.e. $1 \le |\mathcal{F}| < \infty$), and $W_f$ be a random variable such that $\sup_{f\in\mathcal{F}} \mathbb{E}\big[\exp(W_f/\theta)\big] \le 1$ holds for some $\theta > 0$. Then $\mathbb{E}\big[\max_{f\in\mathcal{F}} W_f\big] \le \theta \ln|\mathcal{F}|$.*

*Proof.* Applying Jensen's inequality, replacing the maximum of non-negative elements by their sum and using the condition on $W_f$, we have

$$\exp\left(\mathbb{E}\left[\max_{f\in\mathcal{F}} W_f/\,\theta\right]\right) \le \mathbb{E}\left[\max_{f\in\mathcal{F}} e^{W_f/\,\theta}\right] \le \sum_{f\in\mathcal{F}} \mathbb{E}\left[e^{W_f/\,\theta}\right] \le |\mathcal{F}|\,.$$

Taking logarithm and multiplying both sides by $\theta$, we get the claim. □

**Lemma A.4.** *Let $(\mathcal{F},\ell)$ be a separable metric space, $W$ be a random variable on a set $\mathbb{W}$, $\lambda : \mathcal{F}\times\mathbb{W} \to \mathbb{R}$ be a function and $\phi(f,W) \doteq \lambda(f,W) - \mathbb{E}[\lambda(f,W)]$ for all $f \in \mathcal{F}$. Furthermore, assume that the following conditions hold:*

*(a) there exists $\theta \in (0,\infty)$ such that $\mathbb{E}\big[\exp\big(\lambda(f,W)/\theta\big)\big] \le 1$ holds for all $f \in \mathcal{F}$;*

*(b)* $(\phi(f,W))_{f\in\mathcal{F}}$ *is* $\sigma\ell$*-subgaussian for some* $\sigma \geq 0$*;*

*(c)* $(\phi(f,W))_{f\in\mathcal{F}}$ *is uniformly Lipschitz with respect to* $\ell$ *and modulus* $\tau$ *for some function* $\tau : \mathbb{W} \to [0,\infty)$*.*

*Then for all* $0 \leq \delta \leq \epsilon$*,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \lambda(f,W)\right] \leq \theta \ln \mathcal{N}(\epsilon,\mathcal{F},\ell) + 16\sigma \int_{\delta}^{\epsilon} \sqrt{\ln \mathcal{N}(s,\mathcal{F},\ell)}\, ds + 8\delta\, \mathbb{E}[\tau(W)]\,.$$

*Proof.* When $\mathcal{N}(\delta,\mathcal{F},\ell) = \infty$ for some $\delta \in (0,\epsilon]$, the claim is trivial. So we can assume that $\mathcal{N}(\delta,\mathcal{F},\ell) < \infty$ for all $\delta \in (0,\epsilon]$. Let $\mathcal{F}_\epsilon$ be an $\epsilon$-net of $\mathcal{F}$ with respect to $\ell$ with minimal cardinality and define $g_f \in \operatorname{argmin}_{g\in\mathcal{F}_\epsilon} \ell(f,g)$, the closest element to $f \in \mathcal{F}$ in $\mathcal{F}_\epsilon$. Due to Jensen's inequality and condition (a), $\mathbb{E}[\lambda(f,W)] \leq 0$ holds for all $f \in \mathcal{F}$. Define[8] $g_f^* \in \operatorname{argmax}_{g\in\mathcal{F}:\ell(g,g_f)\leq\epsilon} \mathbb{E}[\lambda(g,W)]$. Then, for all $f \in \mathcal{F}$, $\ell(f,g_f^*) \leq \ell(f,g_f) + \ell(g_f,g_f^*) \leq 2\epsilon$ and, due to $\ell(g_f,f) \leq \epsilon$, $\mathbb{E}[\lambda(f,W)] \leq \mathbb{E}[\lambda(g_f^*,W)]$. Consequently, $\mathcal{F}_\epsilon^* \doteq \{g_f^* : g_f \in \mathcal{F}_\epsilon\}$ is a $2\epsilon$-cover of $\mathcal{F}$ with respect to $\ell$ with $|\mathcal{F}_\epsilon^*| \leq |\mathcal{F}_\epsilon| = \mathcal{N}(\epsilon,\mathcal{F},\ell)$.

Now consider the following decomposition,

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \lambda(f,W) &= \sup_{f\in\mathcal{F}} \Big\{\lambda(g_f^*,W) + \lambda(f,W) - \lambda(g_f^*,W)\Big\} \\
&\leq \max_{g\in\mathcal{F}_\epsilon^*} \lambda(g,W) + \sup_{f\in\mathcal{F}} \Big\{\phi(f,W) - \phi(g_f^*,W) + \mathbb{E}\Big[\lambda(f,W) - \lambda(g_f^*,W)\Big]\Big\} \\
&\leq \max_{g\in\mathcal{F}_\epsilon^*} \lambda(g,W) + \sup_{f\in\mathcal{F}} \Big\{\phi(f,W) - \phi\big(g_f^*,W\big)\Big\}\,.
\end{aligned} \tag{8}$$

Then by Lemma A.3 and condition (a), we obtain

$$\mathbb{E}\left[\max_{g\in\mathcal{F}_\epsilon^*} \lambda(g,W)\right] \leq \theta \ln |\mathcal{F}_\epsilon^*| \leq \theta \ln \mathcal{N}(\epsilon,\mathcal{F},\ell)\,. \tag{9}$$

For all $(f,g_1^*),(h,g_2^*) \in \mathcal{F}\times\mathcal{F}_\epsilon^*$ and $w \in \mathbb{W}$, define

$$\begin{aligned}
\tilde{\phi}\big((f,g_1^*),w\big) &\doteq \phi(f,w) - \phi(g_1^*,w)\,, \\
\tilde{\ell}\big((f,g_1^*),(h,g_2^*)\big) &\doteq \min\big\{\ell(f,h) + \ell(g_1^*,g_2^*),\, 4\epsilon\big\}\,,
\end{aligned}$$

Notice that $(\mathcal{F}\times\mathcal{F}_\epsilon^*,\tilde{\ell})$ is a metric space[9], and recall that $\ell(f,g_f^*) \leq 2\epsilon$ for all $(f,g_f^*) \in \mathcal{K}$. Furthermore, let $\mathcal{K} \doteq \{(f,g_f^*) : f \in \mathcal{F}\} \subseteq \mathcal{F}\times\mathcal{F}_\epsilon^*$ and $f_0 \in \operatorname{argmax}_{f\in\mathcal{F}_\epsilon^*} \mathbb{E}[\lambda(f,W)]$ arbitrary; then $f_0 = g_{f_0}^*$, and so $(f_0,f_0) \in \mathcal{K}$. Now for all $(f,g_f^*),(h,g_h^*) \in \mathcal{K}$, conditions (b) and (c) imply that

$$\tilde{\phi}\big((f_0,f_0),W\big) = 0 \text{ a.s.}\,, \quad \tilde{\ell}\big((f_0,f_0),(f,g_f^*)\big) \leq 4\epsilon\,.$$

Moreover, the process $(\tilde{\phi}(\kappa,W))_{\kappa\in\mathcal{K}}$ is $\sigma\tilde{\ell}$-subgauss, since

$$\begin{aligned}
\tilde{\phi}\big((f,g_f^*),W\big) - \tilde{\phi}\big((h,g_h^*),W\big) &= \phi(f,W) - \phi(h,W) + \phi(g_h^*,W) - \phi(g_f^*,W) \\
&\qquad \text{is subgaussian with parameter} \big(\ell(f,h) + \ell(g_h^*,g_f^*)\big)\sigma\,, \text{ and} \\
\tilde{\phi}\big((f,g_f^*),W\big) - \tilde{\phi}\big((h,g_h^*),W\big) &= \phi(f,W) - \phi(g_f^*,W) + \phi(g_h^*,W) - \phi(h,W) \\
&\qquad \text{is subgaussian with parameter} \big(\ell(f,g_f^*) + \ell(g_h^*,h)\big)\sigma \leq 4\epsilon\,\sigma;
\end{aligned}$$

and $(\tilde{\phi}(\kappa,W))_{\kappa\in\mathcal{K}}$ is also uniformly Lipschitz with respect to $\tilde{\ell}$ and modulus $\tau$, since

$$\begin{aligned}
\tilde{\phi}\big((f,g_f^*),W\big) - \tilde{\phi}\big((h,g_h^*),W\big) &= \phi(f,W) - \phi(h,W) + \phi(g_h^*,W) - \phi(g_f^*,W) \\
&\leq \big(\ell(f,h) + \ell(g_h^*,g_f^*)\big)\tau(W) \text{ a.s., and} \\
\tilde{\phi}\big((f,g_f^*),W\big) - \tilde{\phi}\big((h,g_h^*),W\big) &= \phi(f,W) - \phi(g_f^*,W) + \phi(g_h^*,W) - \phi(h,W) \\
&\leq \big(\ell(f,g_f^*) + \ell(g_h^*,h)\big)\tau(W) \leq 4\epsilon\,\tau(W) \text{ a.s.}
\end{aligned}$$

[8] If such $g_f^*$ element does not exist, one can choose an element which is arbitrary close to the supremum and shrink the gap to zero at the end of the analysis.

[9] To prove the triangle inequality, use $\min\{a+b,c\} \leq \min\{a,c\} + \min\{b,c\}$ for $a,b,c \geq 0$.

Thus, the requirements of Lemma A.2 hold for the function $\tilde{\phi}$ defined on the metric space $(\mathcal{K}, \tilde{\ell})$ with the centering element being $(f_0, f_0)$. Letting $\beta = 4\epsilon \geq \sup_{\kappa,\kappa' \in \mathcal{K}} \tilde{\ell}(\kappa, \kappa')$, we get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\phi(f,W) - \phi\big(g_f^*, W\big)\right\}\right] = \mathbb{E}\left[\sup_{\kappa\in\mathcal{K}} \tilde{\phi}(\kappa, W)\right] \leq 4\sigma \int_{2\delta}^{2\epsilon} \sqrt{2\ln \mathcal{N}(s, \mathcal{K}, \tilde{\ell})}\, ds + 8\delta\, \mathbb{E}[\tau(W)]\,. \tag{10}$$

It remains to bound the entropy of $(\mathcal{K}, \tilde{\ell})$. For any $s \in (\delta, 2\epsilon]$, let $\mathcal{F}_s$ be an $s$-cover of $\mathcal{F}$ with respect to $\ell$ with minimal cardinality and define $\mathcal{K}_s \doteq \mathcal{F}_s \times \mathcal{F}_\epsilon^*$. Then $\mathcal{K}_s$ is an external $s$-cover of $\mathcal{K}$ in the metric space $(\mathcal{F}\times\mathcal{F}_\epsilon^*, \tilde{\ell})$, which means that $\mathcal{K}_s$ might not be a subset of $\mathcal{K}$, but for any $\kappa \in \mathcal{K}$ there exists $\hat{\kappa} \in \mathcal{K}_s$ for which $\tilde{\ell}(\kappa, \hat{\kappa}) \leq s$. Notice that $|\mathcal{K}_{s/2}| = |\mathcal{F}_{s/2}| \cdot |\mathcal{F}_\epsilon^*| \leq \mathcal{N}(s/2, \mathcal{F}, \ell)^2$, so using the relation between internal and external covering numbers (Dudley, 1999, Theorem 1.2.1), we have

$$\sqrt{2\ln \mathcal{N}(s, \mathcal{K}, \tilde{\ell})} \leq \sqrt{2\ln |\mathcal{K}_{s/2}|} \leq 2\sqrt{\ln \mathcal{N}(s/2, \mathcal{F}, \ell)}\,. \tag{11}$$

Taking expectation of (8) and plugging in (9,10,11), we get the claim. □

### A.2 Proof of Theorem 3.1

Recall that $\mu$ denotes the distribution of $(X, Y)$ and that $\mathbb{E}[Y^2] < \infty$ by the assumptions of the theorem. For a function $f : \mathbb{X} \to \mathbb{R}$, define the prediction error ($L_2$-risk) of $f$ to be

$$\|f - y\|_\mu^2 \doteq \mathbb{E}\Big[\big|f(X) - Y\big|^2\Big] = \int_{\mathbb{X}\times\mathbb{R}} |f(x) - y|^2\, \mu(dx, dy)\,,$$

where we slightly abused notation, treating $y$ as the function $y : \mathbb{X} \times \mathbb{R} \to \mathbb{R}$, $y(\hat{x}, \hat{y}) = \hat{y}$ and $f$ as the two argument function $f(\hat{x}, \hat{y}) = f(\hat{x})$. With this notation we have

$$\|f - f_*\|_\mu^2 = \|f - y\|_\mu^2 - \|f_* - y\|_\mu^2\,. \tag{12}$$

Furthermore, denote the empirical risk of $f : \mathbb{X} \to \mathbb{R}$ by $\|f - y\|_n^2 \doteq \frac{1}{n}\sum_{i=1}^n \big|f(X_i) - Y_i\big|^2$. Then the condition for $f_n$ to be an $\alpha$-LSE($\mathcal{F}$), given by (1), can be rewritten as $\|f_n - y\|_n^2 \leq \inf_{f\in\mathcal{F}} \|f - y\|_n^2 + \alpha$.

Following the derivation of Equation 11.12 in the book of Györfi et al. (2002),

$$\begin{aligned}
\mathbb{E}\Big[\,\|f_n - y\|_n^2\,\Big] - \|f_* - y\|_\mu^2 &\leq \mathbb{E}\Big[\inf_{f\in\mathcal{F}} \|f - y\|_n^2\Big] - \|f_* - y\|_\mu^2 + \alpha && \text{(by the definition of } f_n\text{)}\\
&\leq \inf_{f\in\mathcal{F}} \mathbb{E}\Big[\,\|f - y\|_n^2\,\Big] - \|f_* - y\|_\mu^2 + \alpha\\
&= \inf_{f\in\mathcal{F}} \Big(\|f - y\|_\mu^2 - \|f_* - y\|_\mu^2\Big) + \alpha && \big(\text{because } \mathbb{E}[\|f - y\|_n^2] = \|f - y\|_\mu^2\big)\\
&= \inf_{f\in\mathcal{F}} \|f - f_*\|_\mu^2 + \alpha\,. && \text{(by (12))}
\end{aligned} \tag{13}$$

Now for $f : \mathbb{X} \to \mathbb{R}$, $d_n = \{(x_i, y_i) \in \mathbb{X} \times \mathbb{R} : i = 1, \ldots, n\}$, $r \in (0, 1]$ and $(x, y) \in \mathbb{X} \times \mathbb{R}$ define

$$\lambda_r(f, d_n) \doteq r\,\mathbb{E}[\psi(f, X, Y)] - \frac{1}{n}\sum_{i=1}^n \psi(f, x_i, y_i) \quad \text{where } \psi(f, x, y) \doteq |f(x) - y|^2 - |f_*(x) - y|^2\,,$$

so that $\frac{1}{r}\lambda_r(f, D_n) = \|f - y\|_\mu^2 - \|f_* - y\|_\mu^2 - \frac{1}{r}(\|f - y\|_n^2 - \|f_* - y\|_n^2)$. Together with (13), this gives

$$\begin{aligned}
\mathbb{E}\Big[\,\|f_n - f_*\|_\mu^2\,\Big] &= \mathbb{E}\Big[\,\|f_n - y\|_\mu^2 - \|f_* - y\|_\mu^2\,\Big]\\
&= \mathbb{E}\left[\|f_n - y\|_\mu^2 - \|f_* - y\|_\mu^2 - \frac{1}{r}\Big(\,\|f_n - y\|_n^2 - \|f_* - y\|_n^2\,\Big)\right] + \frac{1}{r}\,\mathbb{E}\Big[\,\|f_n - y\|_n^2 - \|f_* - y\|_n^2\,\Big]\\
&= \frac{1}{r}\mathbb{E}\Big[\lambda_r(f_n, D_n)\Big] + \frac{1}{r}\left(\mathbb{E}\Big[\,\|f_n - y\|_n^2\,\Big] - \|f_* - y\|_\mu^2\right)\\
&\leq \frac{1}{r}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \lambda_r(f, D_n)\right] + \frac{1}{r}\left(\inf_{f\in\mathcal{F}} \|f - f_*\|_\mu^2 + \alpha\right)\,.
\end{aligned} \tag{14}$$

Set

$$\phi(f, d_n) \doteq \lambda_1(f, d_n)\,, \quad \tau(d_n) \doteq 8B + 2\left(\frac{1}{n}\sum_{i=1}^{n} \left|f_*(x_i) - y_i\right|^2\right)^{1/2}\,,$$

and notice that the $\sigma$-subgaussian property of the noise ensures $\mathbb{E}[\tau(D_n)] \le 2(4B + \sigma)$. Let $f, g \in \mathcal{F}$ and recall the condition $\max\{\|f\|_\infty, \|g\|_\infty, \|f_*\|_\infty\} \le B$. By elementary algebra,

$$\psi(g, x, y) - \psi(f, x, y) = \big(g(x) - f(x)\big)\Big(g(x) + f(x) - 2f_*(x) + 2\big(f_*(x) - y\big)\Big)\,.$$

Using the tower rule, $\mathbb{E}[Y|X] = f_*(X)$ and Jensen's inequality we derive

$$\begin{aligned}
\mathbb{E}\Big[\psi(f, X, Y) - \psi(g, X, Y)\Big] &= \mathbb{E}\Big[\big(f(X) - g(X)\big)\Big(f(X) + g(X) - 2f_*(X)\Big)\Big] \\
&\le 4B\,\|f - g\|_\infty\,, \\
\frac{1}{n}\sum_{i=1}^{n}\Big(\psi(g, x_i, y_i) - \psi(f, x_i, y_i)\Big) &= \frac{1}{n}\sum_{i=1}^{n}\big(g(x_i) - f(x_i)\big)\Big(g(x_i) + f(x_i) - 2f_*(x_i) + 2\big(f_*(x_i) - y_i\big)\Big) \\
&\le \|f - g\|_\infty\,(4B + 2\,\|f_* - y\|_n)\,,
\end{aligned}$$

and so obtain $\phi(f, d_n) - \phi(g, d_n) \le \|f - g\|_\infty\,\tau(d_n)$. Using the same expansions, the $\sigma$-subgaussian property of the noise and Hoeffding's lemma (i.e., a bounded random variable $Z$, $|Z| \le K$ a.s., is $K$-subgaussian), for any $s \in \mathbb{R}$, we get

$$\begin{aligned}
&\mathbb{E}\left[\exp\Big(s\,\mathbb{E}[\psi(f, X, Y) - \psi(g, X, Y)] - s\big(\psi(f, X, Y) - \psi(g, X, Y)\big)\Big)\right] \\
&\quad= \mathbb{E}\Bigg[\exp\Big(s\,\mathbb{E}[\psi(f, X, Y) - \psi(g, X, Y)] - s\big(g(X) - f(X)\big)\big(g(X) + f(X) - 2f_*(X)\big)\Big)\times \\
&\qquad\qquad \mathbb{E}\Big[\exp\Big(2s\big(g(X) - f(X)\big)\big(f_*(X) - Y\big)\Big)\,\Big|\,X\Big]\Bigg] \\
&\quad\le \mathbb{E}\Bigg[\exp\Big(s\,\mathbb{E}\big[\big(f(X) - g(X)\big)\big(f(X) + g(X) - 2f_*(X)\big)\big] - s\big(f(X) - g(X)\big)\big(f(X) + g(X) - 2f_*(X)\big)\Big)\Bigg]\times \\
&\qquad\qquad \exp\Big(2\,s^2\,\|f - g\|_\infty^2\,\sigma^2\Big) \\
&\quad\le \exp\Big(s^2\,\|f - g\|_\infty^2\,(4B)^2/2\Big)\exp\Big(2\,s^2\,\|f - g\|_\infty^2\,\sigma^2\Big) = \exp\Big(s^2\,\|f - g\|_\infty^2\,\big((4B)^2 + 4\sigma^2\big)/2\Big)\,.
\end{aligned}$$

Then exploiting the independence of $(X_1, Y_1), \dots, (X_n, Y_n)$, Lemma 1.7 of Buldygin and Kozachenko (2000) implies that $\phi(f, D_n) - \phi(g, D_n)$ is $\hat{\sigma}\,\|\cdot\|_\infty$-subgaussian with $\hat{\sigma} = 2\sqrt{4B^2 + \sigma^2}/\sqrt{n}$.

Then the conditions of Lemma A.2 hold for the process $(\phi(f, D_n))_{f\in\mathcal{F}}$ with $\ell(f, g) = \|f - g\|_\infty$: the process is centered (any $f_0 \in \mathcal{F}$ works), it is $\hat{\sigma}\ell$-subgaussian and it is Lipschitz with modulus $\tau(D_n)$. Choosing $\beta = 2B \ge \sup_{f,g\in\mathcal{F}}\|f - g\|_\infty$, we get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\lambda_1(f, D_n)\right] \le \frac{8\sqrt{8B^2 + 2\sigma^2}}{\sqrt{n}}\int_\delta^B \sqrt{\mathcal{H}_\infty(s, \mathcal{F})}\,ds + 8(4B + \sigma)\,\delta\,. \tag{15}$$

Combining (14) using $r = 1$ with (15), we prove Theorem 3.1 (A).

For the second inequality, we need the following results.

**Lemma A.5** (Bernstein's lemma)**.** *Let $\beta \in (0, \infty)$ and $W \in [-\beta, \beta]$ be a bounded random variable. Then for all $s \in [0, 3/\beta)$,*

$$\mathbb{E}\Big[e^{s\,W}\Big] \le \exp\left(s\,\mathbb{E}[W] + \frac{s^2\,\mathbb{E}\big[W^2\big]}{2(1 - s\beta/3)}\right)\,.$$

*Proof.* See Boucheron et al. (2012, Theorem 2.10) using $n \leftarrow 1$, $v \leftarrow \mathbb{E}[W^2]$ and $c \leftarrow \beta/3$. □

**Lemma A.6.** *Let $h : \mathbb{X} \to \mathbb{R}$ be a function with $\|h\|_\infty \le K$ for some $K > 0$, $X \in \mathbb{X}$, $Y \in \mathbb{R}$ be random variables, $Z = h^2(X) + 2h(X)Y$ and assume that for some $\sigma \ge 0$, $Y|X$ is a centered $\sigma$-subgaussian random variable. Then $\mathbb{E}\big[\exp\big((r\mathbb{E}[Z] - Z)/\theta\big)\big] \le 1$ holds with $\theta = \rho\max\{\sigma^2, K^2/4\}$ and the following configurations,*

$$(r, \rho) \in \big\{(0.468, 5.6),\ (2/3, 10.5),\ (0.9, 38.6)\big\}\,. \tag{16}$$

*Proof.* Let $k > 0$, $M = \max\{\sigma^2, K^2/k\}$ and $s > 0$ such that $0 < s - 2s^2M < 3/K^2$. Then, using the subgaussian property of $Y|X$ with $\sigma^2 \le M$ and Lemma A.5, we get

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-sZ\right)\right] &= \mathbb{E}\Big[\exp\left(-sh^2(X)\right)\mathbb{E}\left[\exp\left(-2sh(X)Y\right)|X\right]\Big] \\
&\le \mathbb{E}\Big[\exp\Big(-\big(s - 2s^2M\big)h^2(X)\Big)\Big] \\
&\le \exp\left(-\big(s - 2s^2M\big)\mathbb{E}\big[h^2(X)\big] + \frac{(s - 2s^2M)^2\,\mathbb{E}\big[h^4(X)\big]}{2\big(1 - (s - 2s^2M)K^2/3\big)}\right).
\end{aligned} \tag{17}$$

Now let $c > 1$ to be chosen later, set $s = 1/(2cM)$ and notice that the $s - 2s^2M = (c-1)/(2c^2M) \in (0, 3/K^2)$ condition holds if $k < 6c^2/(c-1)$. Then by using $\mathbb{E}[Z] = \mathbb{E}\big[h^2(X)\big] + \mathbb{E}\big[h(X)\mathbb{E}[Y|X]\big] = \mathbb{E}\big[h^2(X)\big]$ and (17) with $\mathbb{E}\big[h^4(X)\big] \le K^2\,\mathbb{E}\big[h^2(X)\big]$, $sK^2 \le k/(2c)$, $1 - 2sM = (c-1)/c$ and $1 - s(1-2sM)K^2/3 \ge 1 - k(c-1)/(6c^2)$, we have

$$\mathbb{E}\Big[\exp\Big(s\big(r\,\mathbb{E}[Z] - Z\big)\Big)\Big] \le \exp\left(s\,\mathbb{E}\big[h(X)^2\big]\left(r + \frac{1-c}{c} + \frac{3k\,(c-1)^2}{2c\,(6c^2 - kc + k)}\right)\right) \le 1\,,$$

with $k = 4$, $c = \rho/2$ and any of the given $(r, \rho)$ configurations for the second inequality. This proves the claim with $\theta = 1/s = 2cM$. □

Then applying Lemma A.6 to

$$Z = \psi(f, X_i, Y_i) = \big(f(X_i) - f_*(X_i)\big)^2 + 2\big(f(X_i) - f_*(X_i)\big)\big(f_*(X_i) - Y_i\big)$$

with $h \leftarrow f - f_*$, $X \leftarrow X_i$, $Y \leftarrow f_*(X_i) - Y_i$, $K \leftarrow 2B$, $(r, \rho)$ chosen as in (16) and $\theta \leftarrow \rho\max\{\sigma^2, B^2\} = \rho B_\sigma^2$, we get

$$\mathbb{E}\Big[\exp\Big(\big(r\mathbb{E}[\psi(f, X_i, Y_i)] - \psi(f, X_i, Y_i)\big)/\theta\Big)\Big] \le 1\,,$$

for all $i = 1, \dots, n$. We will apply Lemma A.4 with the metric space $(\mathcal{F}, \ell)$, $\ell(f, g) = \|f - g\|_\infty$ and with the process $(\lambda_r(f, D_n))_{f\in\mathcal{F}}$. For any $(r, \rho)$ as in (16), we have

$$\mathbb{E}\Big[\exp\Big(\lambda_r(f, D_n)/(\theta/n)\Big)\Big] = \prod_{i=1}^n \mathbb{E}\Big[\exp\Big(\big(r\mathbb{E}[\psi(f, X_i, Y_i)] - \psi(f, X_i, Y_i)\big)/\theta\Big)\Big] \le 1\,,$$

showing that condition (a) is satisfied with the parameter $\rho B_\sigma^2/n$. Now, notice that $\lambda_r(f, D_n) - \mathbb{E}[\lambda_r(f, D_n)] = \lambda_1(f, D_n) = \phi(f, D_n)$. Thus, in (b) and (c) of Lemma A.4 we can use the same $\hat\sigma$ and $\tau$ as in the first part of the proof. Hence, the conclusion of Lemma A.4 gives

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \lambda_r(f, D_n)\right] \le \frac{\rho\,B_\sigma^2\,\mathcal{H}_\infty(\epsilon, \mathcal{F})}{n} + \frac{32\sqrt{4B^2 + \sigma^2}}{\sqrt{n}}\int_\delta^\epsilon \sqrt{\mathcal{H}_\infty(s, \mathcal{F})}\,ds + 16(4B + \sigma)\,\delta\,. \tag{18}$$

Combining (14) with (18) and using the values of (16), finishes the proof of Theorem 3.1 (B).

### A.3 Computation of LSEs (3)

In this section we present a cutting plane method to compute the QP of LSEs given in (3). We took the CNLS$^+$-G variant proposed by Lee et al. (2013) as a starting point. We kept its constraint updating step by adding the most violated constraints in each iteration, and replaced the initialization completely. We also use aggregate constraints for the ones left out, which can improve the quality of solutions on the price of a small additional computational cost.

To form an initial constraint set, we search for no more than $2d$ points around each $X_i$, which are close for some coordinate on both sides, to approximate a minimal set of points whose convex hull contains $X_i$. Then we form a group $G_i$ of these points and aggregate the corresponding constraints by their sum, $\sum_{k\in G_i} y_i - y_k + (g_i^+ - g_i^-)^\top (X_k - X_i) \le 0$. Then we start with these aggregate constraints ($n$ of them) and introduce the $y_i \le B$ box constraints to relax the one corresponding to the boundary. After each iteration, all the $n-1$ non-boundary constraints are verified for each point $X_i$ and the most violated ones are added to the set (if the corresponding point was a member of $G_i$, the related constraint is subtracted from the aggregate). So the overall constraint set is increased by at most $n$ constraints in each iteration. Finally, after there are no more violated non-boundary constraints, the boundary ones are also checked and introduced to the constraint set if necessary.

We observed that the aggregate constraints significantly reduce solution time. We compared the running time of this aggregate cutting plane (AGCP) method to the CNLS$^+$-G algorithm[10] on 8-dimensional full and half quadratic problems (5). Figure 3 shows the performance comparison of the two methods with standard deviation error bars.[11] Each experiment was repeated 50 times.

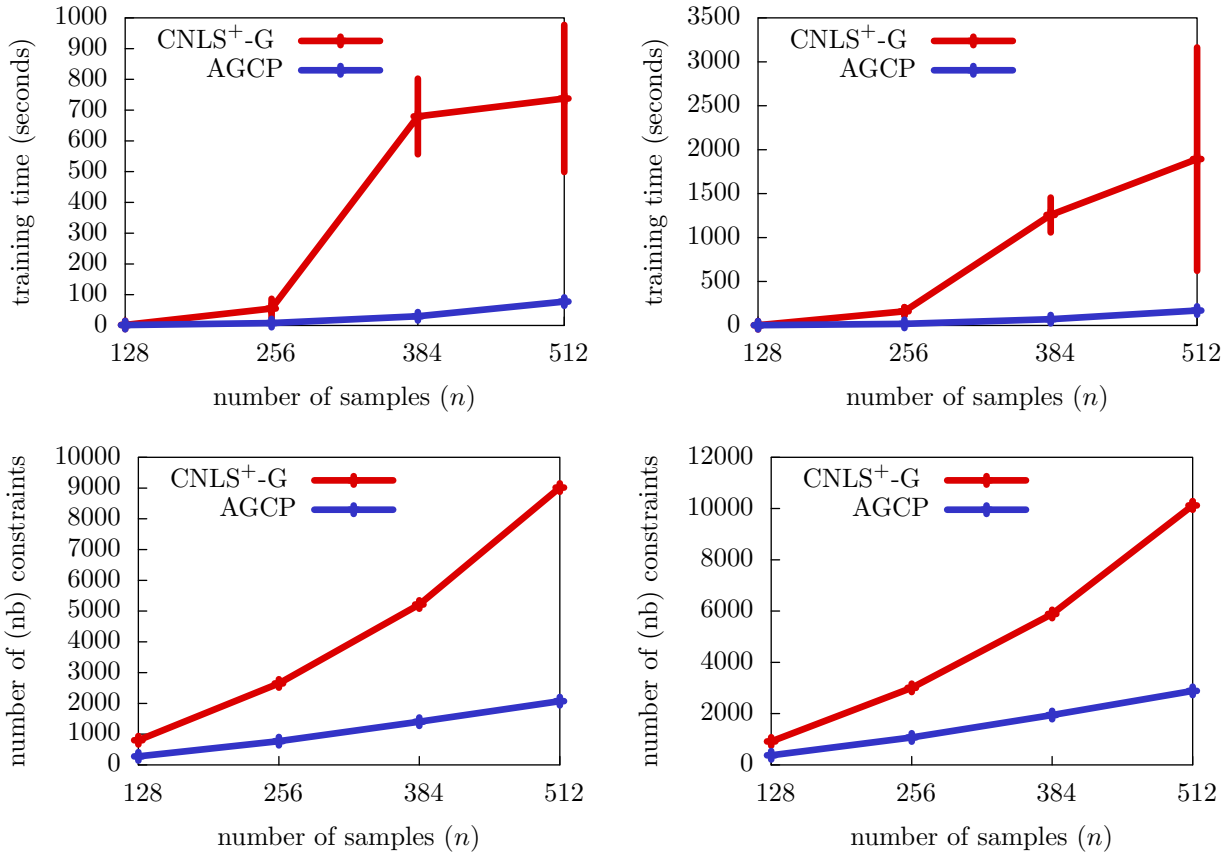


Figure 3: Performance comparison of CNLS$^+$-G and AGCP to find a LSE (3) on the full quadratic (left) and half quadratic (right) problems (5) with $d = 8$.

The lower figures show the number of non-boundary constraints (cardinality of $\bar{C}_{nb}$) used by the algorithms in the last iteration. In both cases, this number is far less than the quadratic $n(n-1)$ used by a direct method. Still, the CNLS$^+$-G algorithm is already too slow for sample size $n = 1024$. However, we could find LSEs with ACPG for $n = 1536$ in $4986 \pm 671$ and $15082 \pm 1972$ seconds (mean $\pm$ standard deviation over 100 repetitions) for the full and half quadratic problems, repsectively.

[10] As we did not find any available source code, we implemented the algorithm according to our best understanding.

[11] Hardware: Dual-Core AMD Opteron(tm) Processor 250 (1KB L1 Cache, 1MB L2 Cache, 2.4 GHz), 8GB RAM. Software: MATLAB R2010b (constraint matrices were built using C), MOSEK 7 Optimization Software (to solve QPs).

To present the algorithm, denote the set of gradient variables by $g \doteq \{g_1^+, \dots, g_n^+, g_1^-, \dots, g_n^-\}$ and the constraints by

$$c_{ik}(y, g) \doteq y_i - y_k + (g_i^+ - g_i^-)^\top (X_k - X_i) \le 0\,,$$
$$\bar{c}_i(y, g) \doteq y_i - B + (g_i^+)^\top (u - X_i) + (g_i^-)^\top (X_i - l) \le 0\,,$$

for all $i, k = 1, \dots, n$. Then the AGCP method is given in Algorithm 1.

**input:** $\{(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R} : i = 1, \dots, n\}$
$\bar{C}_{nb} \leftarrow \emptyset$ and $\bar{C}_b \leftarrow \emptyset$
**for all** $i = 1, \dots, n$ **do**
  $G_i \leftarrow \emptyset$
  **for all** $j = 1, \dots, d$ **do**
    $k_l \leftarrow \operatorname{argmin}_{k=1,\dots,n: X_{ij} > X_{kj}} (X_{ij} - X_{kj})$ {choose one randomly if tied}
    $k_u \leftarrow \operatorname{argmax}_{k=1,\dots,n: X_{ij} < X_{kj}} (X_{ij} - X_{kj})$ {choose one randomly if tied}
    $G_i \leftarrow G_i \cup \{k_l, k_u\}$
  **end for**
**end for**
**while** true **do**
  Solve the following QP:

$$\min_{y,\,g} \sum_{i=1}^{n} (Y_i - y_i)^2 \quad \text{subject to} \quad y_i \in [-B, B]\,, \quad g_i^+, g_i^- \in [0, L]^d\,, \quad i = 1, \dots, n\,,$$
$$\sum_{k \in G_i} c_{ik}(y, g) \le 0\,, \quad i = 1, \dots, n\,,$$
$$c_{ik}(y, g) \le 0\,, \quad (i, k) \in \bar{C}_{nb}\,, \quad \bar{c}_l \le 0\,, \quad l \in \bar{C}_b$$

  violated $\leftarrow$ false
  **for** $i = 1, \dots, n$ **do**
    $V_i \leftarrow \{k = 1, \dots, n : c_{ik}(y, g) > 0\}$
    **if** $|V_i| > 0$ **then**
      violated $\leftarrow$ true
      $k_* \leftarrow \operatorname{argmax}_{k \in V_i} c_{ik}(y, g)$ {choose one randomly if tied}
      **if** $k_* \in G_i$ **then**
        $G_i \leftarrow G_i \setminus \{k_*\}$
      **end if**
      $\bar{C}_{nb} \leftarrow \bar{C}_{nb} \cup \{(i, k_*)\}$
    **end if**
  **end for**
  **if** not violated **then**
    $\bar{V} \leftarrow \{l = 1, \dots, n : \bar{c}_l(y, g) > 0\}$
    **if** $|\bar{V}| = 0$ **then**
      **break**{a soultion is found, exit}
    **end if**
    $\bar{C}_b \leftarrow \{1, \dots, n\}$ {use all boundary constraints}
  **end if**
**end while**
**output:** $y \in \mathbb{R}^n, g_1^+, \dots, g_n^+, g_1^-, \dots, g_n^- \in \mathbb{R}^d$

**Algorithm 1:** AGCP

### A.4 Computation of PLSEs (4)

In this section we present the running time comparison of LSE and PLSE for the 8-dimensional full and half quadratic problems (5). Notice that the QP of LSE has $(2d+1)n$ variables and $n(n-1)$ constraints (not counting

the boundary ones), while PLSE has $(2d+1)K$ variables and $n(K-1)$ constrains with $K = \lceil n^{d/(d+4)} \rceil$. So it is not surprising that PLSE can be more efficiently computed as long as $d$ is not too large. In particular, for our settings with $d = 8$ and $K = \lceil n^{2/3} \rceil$, Figure 4 shows the running time statistics for LSE and PLSE computed by AGCP methods (for computing PLSE we use a slightly different method than for LSE to aggregate the constraints) averaged over 100 repetitions (error bars show standard deviation).

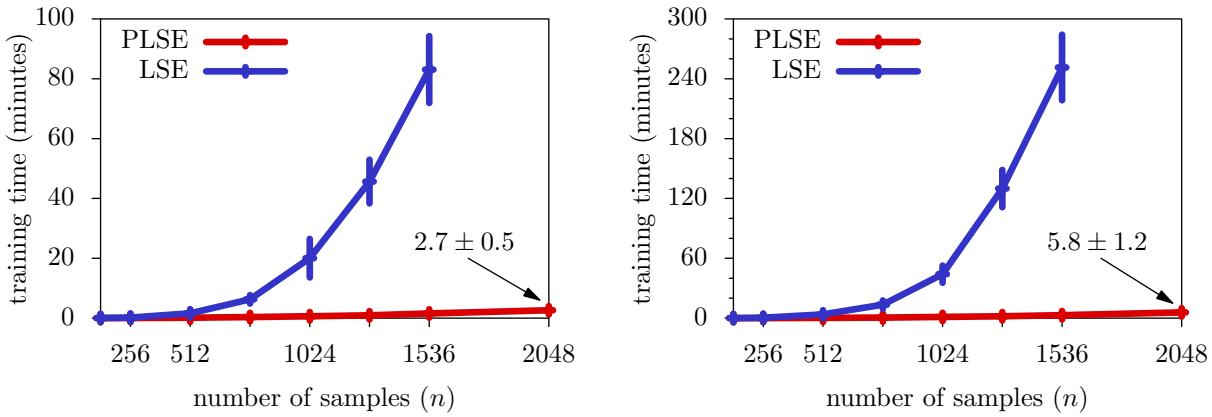


Figure 4: Training time comparison of LSE and PLSE using AGCP methods on the full quadratic (left) and half quadratic (right) problems (5) with $d = 8$.

### A.5 Additional material

**Lemma A.7.** *Let $p \in \mathbb{N} \cup \{\infty\}$ and $\mathbb{X} \subset \mathbb{R}^d$ such that $\mathbb{X}$ has a finite diameter with respect to $\|\cdot\|_p$. Let $R \in (0, \infty)$ such that $diam_p(\mathbb{X}) \doteq \sup_{x,z \in \mathbb{X}} \|x - z\|_p \leq R < \infty$. Then $\mathcal{N}(\epsilon, \mathbb{X}, \|\cdot\|_p) \leq (3R/\epsilon)^d$ for all $\epsilon \in (0, 3R]$. Furthermore, for a rectangular set $\mathbb{X} = \{x : \|x\|_\infty \leq L\}$ with some $L > 0$, we have $\mathcal{N}(\epsilon, \mathbb{X}, \|\cdot\|_\infty) \leq (R/\epsilon)^d$ for all $\epsilon \in (0, R]$.*

*Proof.* For the first claim, consider the volume argument as shown by Pollard (1990, Lemma 4.1) for $p = 2$. Then notice that the volumes of $\|\cdot\|_p$ balls scale proportionally to the $d$-th power of the radius, so the claim can be proved similarly for any $p \in \mathbb{N} \cup \{\infty\}$. Then relate the internal covering and packing numbers by Dudley (1999, Theorem 1.2.1) to get the result for $\epsilon \in (0, \mathrm{diam}_p(\mathbb{X})]$. Finally simply observe that $\mathcal{N}(\epsilon, \mathbb{X}, \|\cdot\|_p) = 1$ for all $\epsilon \geq \mathrm{diam}_p(\mathbb{X})$.

For the second claim, cover $\mathbb{X}$ by hypercubes with side length $2\epsilon$ having centers in $\mathbb{X}$. The number of such cubes is no more than $(1 + \mathrm{diam}_\infty(\mathbb{X})/(2\epsilon))^d \leq (\mathrm{diam}_\infty(\mathbb{X})/\epsilon)^d$ for all $\epsilon \in (0, \mathrm{diam}_\infty(\mathbb{X})/2]$. Finally, note that $\mathcal{N}(\epsilon, \mathbb{X}, \|\cdot\|_\infty) = 1$ for all $\epsilon \geq \mathrm{diam}_\infty(\mathbb{X})/2$. □

**Lemma A.8** (Detailed bounds for Section 4.2)**.** *Let $f_* \in \mathcal{C}_{\mathbb{X},B,L}$ and $f_n$ be an $\alpha$-LSE($\mathcal{C}_{\mathbb{X},B,L}$). Then the following cases hold:*

$$\mathbb{E}\Big[\|f_n - f_*\|_\mu^2\Big] < 7629(d+1)B_\sigma \max\{B_\sigma, L_d\}\, n^{-4/(d+4)} \left(\ln\left((R_d^*/(8L_d))^2\, n\right) + 4\right) + 2\alpha \qquad \text{, for } d < 4\,,$$

$$\mathbb{E}\Big[\|f_n - f_*\|_\mu^2\Big] < 1551\, B_\sigma L_4\, n^{-1/2} \left(\ln^{3/2}\left((R_4^*/(8L_4))^2\, n\right) + 3\right) + \alpha \qquad \text{, for } d = 4\,,$$

$$\mathbb{E}\Big[\|f_n - f_*\|_\mu^2\Big] < 11767 \frac{\sqrt{d+1}}{d-4} B_\sigma L_d\, n^{-2/d} \left(\ln\left((R_d^*/L_d)\, n^{2/d}\right) + 6\right) + \alpha \qquad \text{, for } d > 4\,.$$

*Proof.* To handle the integrals of Theorem 3.1 in the $d \neq 4$ cases, we used the $\sqrt{\ln(10R_d^*/s)} \leq \ln(10eR_d^*/s)$ approximation with $s \leq 80L_d \leq 10R_d^*$. For the case $d < 4$, we used the $(c_1, c_2, c_3, c_4) = (43, 80, 89, 10/9)$ constants. The rest is pure calculation using $\epsilon$, $\delta$ as given in Section 4.2. □