Sample average approximation with heavier tails I: non-asymptotic bounds with weak assumptions and stochastic constraints

Roberto I. Oliveira · Philip Thompson

Received: 21 November 2017 / Accepted: 20 March 2022 / Published online: 20 April 2022
© Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society 2022

Abstract
We derive new and improved non-asymptotic deviation inequalities for the sample average approximation (SAA) of an optimization problem. Our results give strong error probability bounds that are “sub-Gaussian” even when the randomness of the problem is fairly heavy tailed. Additionally, we obtain good (often optimal) dependence on the sample size and geometrical parameters of the problem. Finally, we allow for random constraints on the SAA and unbounded feasible sets, which also do not seem to have been considered before in the non-asymptotic literature. Our proofs combine different ideas of potential independent interest: an adaptation of Talagrand’s “generic chaining” bound for sub-Gaussian processes; “localization” ideas from the Statistical Learning literature; and the use of standard conditions in Optimization (metric regularity, Slater-type conditions) to control fluctuations of the feasible set.

Mathematics Subject Classification 90C15 · 90C31 · 60E15 · 60F10

1 Introduction
Understanding sample average approximations is a fundamental problem in Stochastic Programming [39, 43]. Suppose we are given an optimization problem:

---

Roberto I. Oliveira is supported by CNPq grants 432310/2018-5 (Universal) and 304475/2019-0 (Produtividade em Pesquisa), and FAPERJ grants 202.668/2019 (Cientista do Nosso Estado) and 290.024/2021 (Edital Inteligência Artificial).

Philip Thompson
thomsp@purdue.edu

Roberto I. Oliveira
rimfo@impa.br

1 Instituto de Matemática Pura e Aplicada IMPA, Rio de Janeiro, RJ, Brazil
2 Purdue University and Krannert School of Management, West Lafayette, USA
\[ f^* := \min_{x \in Y} f_0(x) \]
\[ \text{s.t. } f_i(x) \leq 0, \quad \forall i \in I, \] (1)

with \( Y \subset \mathbb{R}^d \) and a nonempty feasible set

\[ X := \{ x \in Y : f_i(x) \leq 0, \forall i \in I \}. \] (2)

In this paper, each of the functions \( f_i \) is given by an expectation

\[ f_i(x) = \mathbb{E} F_i(x, \cdot) := \int_{\mathcal{E}} F_i(x, \xi) \mathbb{P}(d\xi) \] (3)

where \( \mathbb{P} \) is a probability measure over a set \( \mathcal{E} \) and the \( F_i : Y \times \mathcal{E} \to \mathbb{R} \) are Carathéodory functions. In typical settings, the measure \( \mathbb{P} \) and the functions \( f_i \) are not directly accessible. It may be, however, that a random sample \( \{\xi_k\}_{k=1}^N \) from \( \mathbb{P} \) is available. If that is the case, it is natural to consider the sample-average approximation (or SAA) to (1), where the \( f_i \) are replaced by sample averages:

\[ \hat{F}_i(x) := \hat{\mathbb{E}} F_i(x, \cdot) = \frac{1}{N} \sum_{k=1}^N F_i(x, \xi_k) \] (4)

This leads to some natural questions considered in numerous works in stochastic optimization:

1. Are (nearly) optimal solutions to the SAA also nearly feasible and nearly optimal for the original problem (1)?
2. Are the values of the two problems typically close?

Asymptotic analyses of the SAA assume the sample size \( N \) diverges whereas the functions \( F_i, f_i \), the set \( Y \) and the measure \( \mathbb{P} \) remain fixed. Using tools such as uniform Strong Law of Large Numbers and Central Limit Theorems, these analyses obtain precise answers to the above questions. This program has been carried out in numerous works, e.g., [1, 13, 24, 25, 33–35, 40–43]. See [17, 23, 43] for extensive reviews.

Another type of analysis, which we pursue in this work, is non-asymptotic in nature. It consists of proving explicit bounds for the value and quality of SAA solutions with explicit dependence on the sample size \( N \) and other problem parameters. For instance, letting \( f^* \) and \( \hat{f}^* \) be the values of the original problem (1) and its SAA (respectively), a recent non-asymptotic result by Guigues, Juditsky and Nemirovski [15] gives guarantees of the form:

\[ \forall t \geq 0 : \mathbb{P} \left\{ |\hat{f}^* - f^*| \leq \frac{A + B \sqrt{t}}{\sqrt{N}} \right\} \geq 1 - e^{-t}, \] (5)

where \( A \) and \( B \) do not depend on \( N \) or \( t \) (but do depend on other problem parameters). Guarantees of this kind are called “sub-Gaussian”\(^1\) because they imply that the tail

\(^1\) Another typical light-tail condition is to assume an sub-exponential tail.
Sample average approximation with heavier tails I…

decay $\sqrt{N}|\hat{p}^* - f^*|$ roughly matches that of a Gaussian distribution with standard deviation $B$. This sort of asymptotic behavior is what one expects from asymptotic statements such as those found in [40].

With few exceptions, non-asymptotic guarantees in the literature require that the random variables $F_i(x, \xi)$ be very light-tailed: that is $|F_i(x, \xi) - f_i(x)|$ has finite $p$-th moments for all $p \geq 1$. In the rare cases where this is avoided [14, 20], the dependence on $N$ is suboptimal (as we shall see). Other limitations to current finite-sample analyses of SAA include requiring the feasible set $X$ to be bounded, and avoiding expected value constraints. Even in an equation like (5), it is often not clear if the dependence of “constants” like $A$ and $B$ on other problem parameters (such as the dimension) is reasonable.

There is thus a gap between what one may expect SAA to do on the basis of asymptotic analyses, and what has been proven to do non-asymptotically. Is this a technical issue, or does it point to underlying limitations of SAA? This question is especially pressing in high-dimensional problems, where asymptotic theory is not expected to give good results even for fairly large $N$. Luckily, one can prove significantly better finite-sample guarantees for SAA, as we explain below.

1.1 Our contribution

Our goal in this paper is to obtain new and improved non-asymptotic bounds for the sample average approximation. Our probabilistic assumptions are significantly weaker than in previous work, and our bounds often improve on other results by making better use of the geometry of our problem. We highlight some salient features of our approach. Finite-moment assumptions. We do not require infinitely many moments of any of the random variables involved in our problem. Our main assumption is that, given norm $\| \cdot \|$ over $\mathbb{R}^d$, the $F_i: Y \times \Xi \to \mathbb{R}$ are stochastically Hölder over sets $Z \subset Y$, in the sense that inequalities of the kind

$$\forall x, x' \in Z : |F_i(x, \xi) - F_i(x', \xi)| \leq L_i(\xi)\|x - x'\|^\alpha$$

hold in suitable $Z \subset Y$, with $0 < \alpha \leq 1$ and $L_i(\xi)$ satisfying weak conditions. See Assumptions 1 and 2 below for details. Conditions of the kind of (6) have often appeared in the literature [19, 20], but either with much stronger moment assumptions on the $L_i$ or with suboptimal error bounds in the sample size.

Joint guarantees for values, feasibility and optimality. An inequality such as (5) bounds the difference in values between the SAA and the original problem. Our results also quantify how good the extent to which SAA is close to being feasible and optimal for the original problem. In all cases, we obtain optimal dependence on the same size $N$, as well as “sub-Gaussianity” for a relevant set of parameters. In Sect. 1.2, we comment on what parameters we consider. Their precise definitions are discussed in more detail in Sects. 4 and 5 (see in particular the discussion in Sect. 5.1). In Sect. 4, Theorem 2 considers general (possibly non-convex) problems. In Sect. 5, Theorem 3 and Propositions 3–4 state sharper “localized” bounds for convex optimization.

Generic chaining without light tails. A key step in our proofs will be to obtain concentration inequalities for $\mathbb{E}F_i(x, \cdot) - f_i(x)$ under assumptions such as (6). For this
purpose, we adapt to our setting Talagrand’s generic chaining method for empirical processes [46, 47], as improved by Dirksen [12]. Generic chaining is an optimal method for taking problem geometry into account, and gives good problem-dependent bounds on “constants” like $A$ and $B$ in (5) under sub-Gaussian assumptions. We obtain novel concentration generic chaining inequalities (Theorem 4) that do not require light tails, which are of independent interest.

Localization, convexity and unbounded sets. “Localization” is a key idea developed by researchers in Statistical Learning, especially Koltchinskii, Mendelson and their collaborators [4, 5, 21, 22, 28, 29]. For convex problems, it means that “failure” for an SAA solution must originate from “bad behavior” of the SAA in a (often small) sublevel set around the minimum. We will show that this idea often leads to faster convergence rates for SAA. It also allows us to only require the Hölder condition (6) in a potentially “small” subset $Z \subset Y$. In some cases, this allows us to consider unbounded convex feasible sets and functions $F_i$ with superlinear growth. Theorem 3 presents general localized rates. Propositions 3 and 4 exemplify Theorem 3 when typical regularity assumptions hold.

Constraints in expectation. We deal systematically with constraints in expectation. These mean that the feasible set of the SAA is a perturbation of a deterministic set. We control these perturbations by combining tools from Optimization theory—metric regularity and Slater-type conditions—with our “localization toolbox”. A key result will be to show that, when constraints are perturbed, this does not change much the “generic chaining” parameters of relevant sub-level sets of the objective function. We remark that we do not make detailed reference to the large literature on optimization with chance constraints. This challenging problem is out of scope of this paper as the continuity assumption in (6) is not satisfied.

Examples. Finally, we present four different applications of our theory. The first two examples is treated in detail in this work as a proof of concept of Theorems 2 and 3. The other examples, which require finer analyses, are presented in a dedicated companion paper [30].

**Example 1** (Regular convex optimization problems; Sect. 5.3) We consider SAA with convex objective and constraints satisfying two typical regularity conditions: (1) a local Slater constraint qualification (Assumption 4) and (2) a local regular solution set (Assumption 5). The first is typical while the latter is satisfied, e.g., for objectives that are locally strongly convex or with local weak sharp minima [10, 11]. We consider constraint-free problems (Proposition 3) or problems with random constraints (Proposition 4). They offer concrete localized rates implied by the general Theorem 3. In particular, unlike Theorem 2, the obtained rates depend only on the diameter and a complexity measure of a neighbourhood of the solution set. See further discussions in Sects. 5.1 and 5.3.

**Example 2** (Metric projection problems; Sect. 6) We consider the special case of problem 1 where the feasible set is convex and $f_0(x) := \|x - x_0\|_2^2$ with $x_0 \in \mathbb{R}^d$ fixed and $\| \cdot \|$ the standard Euclidean norm. In this case, the unique optimal solution of our problem is the metric projection of $x_0$ onto the feasible set $X$. We provide finite-sample guarantees for SAA that make strong use of localization. One particular difficulty of
this problem is the fact that the Lipschitz modulus of the objective varies along the feasible set.

**Example 3** (Risk-averse portfolio optimization; in companion paper [30]) Here,

\[ \xi = (\xi[1], \ldots, \xi[d])^T \]

is a random vector whose coordinates correspond to losses of \( d \) distinct financial assets. If \( x = (x[1], \ldots, x[d])^T \) is a vector whose coordinates describe the fractions of the initial capital invested in assets \( 1, \ldots, d \), then the total loss is proportional to \( \langle x, \xi \rangle \). We wish to minimize the expectation of \( \langle x, \xi \rangle \) subject to a constraint on the conditional value-at-risk of the solution [37]. In this problem, the case of light-tailed \( \xi \) would be of little interest. In a companion paper [30], we describe specific assumptions that allow for heavy tails. We show that the localization toolbox obtained in this paper implies that “risk inflation” only affects a lower dimensional space.

**Example 4** (The Lasso estimator; in companion paper [30]) In Least-Squares-type problems, the loss function to be minimized is \( f(x) = \mathbb{E} F(x, \cdot) \), with \( F(x, \xi) := [y(\xi) - (x(\xi), x)]^2 \). Here, \( y(\xi) \in \mathbb{R} \) and the random vector \( x(\xi) \in \mathbb{R}^d \). Minimizing the empirical function \( \hat{F}(x) := \mathbb{E} F(x, \cdot) \) tends to work when \( N \gg d \), but not when \( N \ll d \), as the problem is undetermined. Tibshirani [48] proposed the Lasso estimator given by the problem \( \min_{x \in Y} \hat{F}(x) \), with \( Y := \{ x \in \mathbb{R}^d : \|x\|_1 \leq R \} \), where \( R > 0 \) is a tuning parameter and \( \| \cdot \|_1 \) denotes the \( \ell_1 \)-norm. Inspired by Bickel, Ritov and Tsybakov [8], we analyse in our companion paper [30] the least squares problem subjected to \( \|D_2x\|_1 \leq R \), where \( D_2 \) is a data-driven matrix. We obtain improved “persistence” bounds [6] for a least-squares Lasso-type estimator. Our proof is based on localization techniques established in this paper.

### 1.2 Discussion and comparison with previous work

Of the numerous papers on the topic of SAA, we highlight [2, 3, 15, 19, 20, 26, 34, 35, 39, 42–45, 49–52] as relevant to our findings. Except for [19, 20, 38], all of the non-asymptotic papers papers assume light-tailed data. This restriction is lifted in references [19, 20], but at the cost of worse dependence on \( N \) in (5): the error \( \hat{F}^* - f^* \) is stochastically bounded by a quantity that decays like \( N^{-\beta} \) for some \( \beta < 1/2 \). By contrast, the paper [38] makes weak probabilistic assumptions on the data, and obtains distributional results in a asymptotic setting for a reformulation of the SAA. Our results assume heavy-tails, are nonasymptotic, do not use reformulations and achieve the optimal rate \( N^{-\frac{1}{2}} \) in terms of the sample size, with joint guarantees of feasibility and optimality. In addition, our bounds explicitly account for the geometry of the feasible set.

To understand our improvements, it is necessary to take a step back and understand how light tails were used in previous analysis. For the moment, consider the case where \( \mathcal{I} = \emptyset \), i.e. the feasible set of our original problem (1) is \( X = Y \) and there are no constraints in expectation. The easiest way to bound the difference between \( \hat{F}^* - f^* \)
(say) is via a uniform bound:

\[ |\hat{F}^* - f^*| \leq \sup_{x \in Y} |\hat{F}(x) - f(x)| = \sup_{x \in X} \left| \frac{1}{N} \sum_{k=1}^{N} (F(x, \xi_k) - EF(x, \cdot)) \right|. \tag{7} \]

To bound the right hand side (RHS), a typical approach uses two steps. The first one is to discretize the feasible set \( X \); this reduces the problem of controlling the supremum over \( X \) to controlling the supremum over finite subsets. The next step is to use concentration-of-measure inequalities \cite{9} to deal with the finite subsets. For this it is essential to have strong concentration bounds, which typically require light tails. Our approach uses ideas that seem new in this setting. We discretize via Talagrand’s generic chaining method \cite{46}, which is optimal in Gaussian processes and gives better dependence on the geometry of the problem. We do this via a novel concentration inequality (Theorem 4) that separates the fluctuations of the RHS into two components: one that is \emph{always sub-Gaussian}, and another that depends on the fluctuations of \( L^2(\cdot) \). This will give us sub-Gaussian results in certain probability regimes.

To continue with our approach, we note that the bound (7) is often too pessimistic. Oftentimes, one can show that the minimizer of the SAA is usually quite close to the minimizer of the original problem. If that is the case, then \( |\hat{F}^* - f^*| \leq \sup_{x \in Z} |\hat{F}(x) - f(x)| \) for a potentially much smaller set \( Z \subset Y \). This localization idea goes back at least to the work of Koltchinskii and Panchenko \cite{21} and was more fully developed in Koltchinskii’s IMS Medallion Lecture \cite{22}. Mendelson has also greatly contributed to this approach, starting with joint work with Bousquet and Bartlett \cite{4} and continuing with his papers \cite{28, 29}. These works employed localization in a somewhat different form from \cite{21, 22} in convex settings. In this paper, we apply and extend those ideas to the setting where there are constraints in expectation. In Proposition 5, we give a “localized bound” for perturbations of the original problem. These include perturbations of the constraints. We then show in Lemma 4 that one can control the effect of those perturbations on the feasible sets via Slater-type conditions.

Before concluding this section, it is instructive to discuss beforehand what are the relevant parameters appearing in the improved rates of Theorems 2–3 and Propositions 3–4. Let \( c > 0 \) denote an absolute constant. For light-tailed Hölder functions with exponent \( \alpha \) and modulus \( \sigma \), a reanalysis of the arguments in \cite{15} shows that the parameters in bound (5) are typically of the form

\[ A(5) = c \sigma \text{diam}^\alpha(Y) \sqrt{d} \quad \text{and} \quad B(5) = c \sigma \text{diam}^\alpha(Y) \]

where \( d \) is the dimension and \( \text{diam}(Y) \) denotes the diameter of \( Y \).

The main improvement of Theorem 2 is to allow heavier tails and give joint guarantees for optimality and feasibility, with bounds that are of the form

\[ A = c \sigma \gamma^{(\alpha)}(Y) \quad \text{and} \quad B = c \sigma \text{diam}^\alpha(Y) + c \sigma_* \]
where $\sigma_*$ is the variance at a solution. Here $\gamma^{(\alpha)}(Z)$ denotes a complexity measure of a set $Z \subset \mathbb{R}^d$ coming from the theory of Gaussian processes, which we discuss in 2.2. A conservative upper bound $c \text{diam}^{\alpha}(Z)\sqrt{d}$ is possible. In case of random constraints, the probability bounds depend logarithmically on the number of constraints and (implicitly) on the metric regularity constant $c$ of the feasible set (Assumption 3).

Theorem 3 and Propositions 3–4 give sharper localized bounds for convex problems satisfying a Slater condition (Assumption 4). The statement of Theorem 3 is more involved. Qualitatively, the rates depend on factors of the form

$$A = c\sigma(\epsilon, \delta) \gamma^{(\alpha)}(X_*^{\epsilon, \delta}) \quad \text{and} \quad B = c\sigma(\epsilon, \delta) \text{diam}^{\alpha}(X_*^{\epsilon, \delta}) + c\sigma_*$$

where $\sigma(\epsilon, \delta)$ denotes the Hölder modulus variance over the set $X_*^{\epsilon, \delta}$ of approximate solutions having feasibility slackness $\delta > 0$ and optimality slackness $\epsilon > 0$. Hence we allow the Hölder modulus to vary across bounded regions. “Localization” stems from the fact that the diameter and complexity of $X_*^{\epsilon, \delta}$ are typically much smaller than the ones for $X$ or $Y$. In case of random constraints, the range of $(\epsilon, \delta)$ for which this bound holds depend on the parameters of the Slater condition (Assumption 4). We refer to Sect. 5.1 for a qualitative discussion on these points before Theorem 3 is presented formally. Technical rate statements also appear in the literature on localization in Statistics and Machine Learning [22, 29] (in this setting without random constraints). The difficulty lies in the fact that a precise rate depends on “solving” a fixed-point on $(\epsilon, \delta)$. In our case, an additional difficulty is that $(\epsilon, \delta)$ are coupled: feasibility affects optimality.

Propositions 3–4 presents specific rates implied by the general Theorem 3 assuming, besides Assumption 4, a typical local regularity assumption on the solution set (Assumption 5). This includes, e.g., cases when the objective is locally strongly convex or it has locally weakly sharp minima [10, 11]. For simplicity, we assume the Hölder modulus’s variance $\sigma$ is constant. For the sake of comparison with the literature on localization in Statistical Learning, Assumption 5 is an analog (with proper differences) of the so called local Bernstein condition on the loss function [29]. For strong regular sets ($\kappa = 1/2$), Proposition 3 presents “fast-1/$N$-rates” for the constraint-free case of the form $c\sigma^{-2}(C_\alpha + t)/N$ where $c$ is a condition number (Assumption 5). Here, $C_\alpha$ is the ratio comparing the complexity and diameter of the approximate solution set. A pessimistic bound for $C_\alpha$ is of order $d$. Proposition 4, allowing random constraints, presents “slower-1/$\sqrt{N}$-rates” of the form $C\sqrt{C_\alpha + \log m + t}/N$. Here, $m$ is the number of constraints and $C$ is a constant depending polynomially on the regularity constants of Assumptions 4 and 5, $\alpha, (\sigma, \sigma_*)$ and the diameter of the approximate solution set $X_*^{\epsilon, \delta}$ for a small slack $\delta > 0$. A notable fact is that these rates are localized in that they do not depend on the size and complexity of the entire feasible set $X$, just of an approximate solution set. Another notable fact is the deterioration of order $\sqrt{N}$ in the rate when random constraints are present. As explained for the metric projection problem, this feature is in general unavoidable.

Finally, we emphasize a few points about our approach. In most cases, we expect our results to be of optimal or nearly optimal order of magnitude in terms of problem geometry and/or sample size. Like with most non-asymptotic analyses, we do not expect our results to be as tight as asymptotic results when it comes to constants. The
goal of our paper is not to give bounds that can be directly used in practice, but rather to better understand the fundamental properties of SAA with finite samples, in settings where other problem parameters (such as the dimension and diameter) can be large.

1.3 Organization

The remainder of the paper is organized as follows. Section 2 fixes notation and recalls some notions from Probability theory, most notably “generic chaining”. Section 3 presents the setup for our problem and the assumptions we require on the random variables and feasible sets involved. Section 4 contains the statement of our main results for possibly non-convex problems (Theorem 2). Section 5 states our main results for convex problems (Theorem 3 and Propositions 3–4). These are immediately applied to a simple example in Sect. 6.

The next three sections presents our main technical tools separately, as we believe they might be applied or combined in different ways. The concentration inequality for heavy-tailed distributions is presented in Sect. 7. Section 8 describes the relationship between good approximation properties of the SAA and differences \( \widehat{F}_i(x) - f_i(x) \). In particular, this is where we prove our localization results. It will be clear that we need to understand how the feasible set \( X \) changes when constraints are slightly relaxed. We present our geometrical tools for that purpose in Sect. 9. Sections 8–9 are deterministic results and may be useful elsewhere.

The paper ends with Sect. 10, where the main results are proven. An Appendix presents a few technical proofs left over from the main text.

2 Preliminaries

2.1 Basic notation

Given a set \( S \), we denote its (potentially infinite) cardinality by \( |S| \). The complement of an event \( E \) in a probability space is \( E^c \). For \( m \in \mathbb{N} \), we write \([m] := \{1, \ldots, m\}\).

Elements of \( \mathbb{R}^d \) are column vectors. Given \( x \in \mathbb{R}^d \), its coordinates are denoted by \( x[i], 1 \leq i \leq d \). A superscript \( T \) is used to denote transposition of a vector, so \( x \in \mathbb{R}^d \) is given by \((x[1], \ldots, x[d])^T\). The inner product of \( x, y \in \mathbb{R}^d \) is denoted by \( \langle x, y \rangle \) or \( x^T y \). Norms are denoted by \( \| \cdot \| \) and the unit ball around 0 in that norm is \( \mathbb{B} \). Given \( a \in \mathbb{R}, a_+ := \max\{a, 0\} \).

Let \((\mathcal{M}, d)\) be a metric space. We let \( \text{diam}(A) \) denote the (potentially infinite) diameter of \( A \subset \mathcal{M} \). Given \( x \in \mathcal{M} \) and \( A \subset \mathcal{M} \) nonempty, \( d(x, A) := \inf_{a \in A} d(x, a) \).

We fix from now on a probability space \((\mathcal{F}, \mathcal{A}, \mathbb{P})\) and assume all random variables we consider are defined on it. Given a random variable \( Z \), we let \( \mathbb{E}[Z] \) denote its mean, \( \mathbb{V}[Z] \) denote its variance and \( \|Z\|_p := (\mathbb{E}[|Z|^p])^{1/p} \) denotes \( L^p \) norm (for \( p \geq 1 \)).
2.2 Complexity parameters for sets

We review in this section some definitions and results about “generic chaining”. Talagrand’s book [46] is the best reference for these concepts.

The “generic chaining” functional of a metric space \((\mathcal{M}, d)\) is a measure of the “complexity” of discretizing \(\mathcal{M}\) at different scales. To define it, we need the following concept. A sequence \(\{A_j\}_{j=0}^{\infty}\) is admissible if each \(A_j\) is a partition of \(\mathcal{M}\), with \(|A_0| = 1\) and \(|A_j| \leq 2^{2j}\) for each \(j \geq 1\). For each \(j\), let \(\text{diam}(A_j)\) denote the largest diameter of a set in partition \(A_j\).

Given \(0 < \alpha \leq 1\), \(\gamma_2^{(\alpha)}(\mathcal{M}, d)\) is defined as:

\[
\gamma_2^{(\alpha)}(\mathcal{M}, d) := \inf_{\{A_j\}_{j \geq 0} \text{admissible}} \left\{ \sum_{j \geq 0} 2^j \text{diam}(A_j)^{\alpha} \right\}. \tag{8}
\]

**Remark 1** In the usual definition of the \(\gamma_2\) functional, one takes \(\alpha = 1\). \(\gamma_2^{(\alpha)}\) is the functional obtained when the metric \(d\) is replaced by the equivalent metric \(d^{\alpha}\). We will omit \(d\) from the notation when it is clear which metric we are referring to. This remark should be kept in mind when reading Theorem 1 and Eq. (11) below.

Talagrand’s celebrated majorizing measures theorem [46, 47] shows that:

\[
c \mathbb{E} \left[ \sup_{x \in \mathcal{M}} |Y_x - Y_{x_0}| \right] \leq \gamma_2^{(\alpha)}(\mathcal{M}, d) \leq C \mathbb{E} \left[ \sup_{x \in \mathcal{M}} |Y_x - Y_{x_0}| \right], \tag{9}
\]

with \(c, C > 0\) universal, when the \(Y_x\) are mean-zero Gaussian and \(\mathbb{V}[Y_x - Y_{x'}] = d(x, x')^{2\alpha}\).

In fact, the upper bound in (9) does not require that the \(Y_x\) be truly Gaussian, only that they have sub-Gaussian tails. The next theorem, which we will use later, illustrates this point. It follows from Talagrand’s work [46, 47] with an improvement due to Dirksen [12].

**Theorem 1** (Generic chaining tail bound [12, 46, 47]) Suppose \(\gamma_2^{(\alpha)}(\mathcal{M}) < +\infty\) (in particular, \(\mathcal{M}\) is totally bounded). Let \(\{Y_x\}_{x \in \mathcal{M}}\) be a family of random variables indexed by the points of \(\mathcal{M}\), which depend almost surely continuously on \(x\). Assume further that the \(Y_x\) satisfy the following sub-Gaussian assumption.

\[
\forall t \geq 0 \forall x, x' \in \mathcal{M} : \mathbb{P}\{Y_x - Y_{x'} \geq d(x, x')^{\alpha} \sqrt{2(1 + t)}\} \leq e^{-t}.
\]

Then for any \(t \geq 0\) and \(x_0 \in \mathcal{M}\)

\[
\mathbb{P}\left\{ \sup_{x \in \mathcal{M}} |Y_x - Y_{x_0}| \leq 3\sqrt{2}\text{diam}(\mathcal{M})^{\alpha} \sqrt{1 + t} + 2\sqrt{2} \gamma_2^{(\alpha)}(\mathcal{M}) \right\} \leq e^{-t}.
\]

\(^2\) The constants appearing in our Theorem 1 are not the same as in [12], but can be easily obtained via the same method.
The functional $\gamma_2^{(\alpha)}(M, d)$ is somewhat mysterious, and can be quite difficult to compute. In the case $M \subset \mathbb{R}^d$, $\alpha = 1$ and $d$ is given by the standard Euclidean norm, Talagrand’s general theory connects $\gamma_2^{(1)}(M)$ to a parameter called the Gaussian width. Letting $g \in \mathbb{R}^d$ denote a standard Gaussian random vector, the Gaussian width of $M$ is defined as

$$w(M) := \mathbb{E} \sup_{x \in M} \langle x, g \rangle.$$ 

It follows from (9) that the ratio $\gamma_2^{(1)}(M)/w(M)$ is upper and lower bounded by absolute constants $c, C > 0$. One consequence of this fact is that, if $M$ is the convex hull of a finite set $F$ of points, then:

$$\gamma_2^{(1)}(M) \leq C' \sqrt{\log |F|} \max_{x \in F} \|x\|,$$ (10)

for an absolute constant $C' > 0$.

A more general upper bound for $\gamma_2^{(\alpha)}(M)$ comes from Dudley’s entropy integral [46]. Recall that an $r$-net in $M$ is a set $A \subset M$ such that $d(x, A) \leq r$ for all $x \in M$. The $r$-covering number of $M$ is the size of the smallest $r$-net. The $r$-entropy number of $M$, $H(M, r)$, is the natural log of the $r$-covering number. It is known that

$$\gamma_2^{(\alpha)}(M) \leq C \int_0^{\text{diam}(M)} \sqrt{H(M, r^{\frac{1}{\alpha}})} \, dr,$$ (11)

with $C > 0$ is a universal constant. An important special case is when $M \subset \mathbb{R}^d$ and $d$ is given by a norm, in which case the entropy integral bound is upper bounded by $\text{diam}(M)^{\alpha} \sqrt{d}$ up to a universal constant. In particular, we obtain,

$$\gamma_2^{(\alpha)}(M) \leq C_{\alpha} \sqrt{d} \text{diam}(M)^{\alpha}$$ (12)

with $C_{\alpha} > 0$ only depends on $\alpha$. However, this bound can be very loose, as the next example shows.

**Example 5** Let $M \subset \mathbb{R}^d$ denote the standard simplex in $d$ dimensions, that is, the convex hull of the $d$ canonical basis vectors. Let $d$ denote the standard Euclidean metric. In this case, (12) bounds $\gamma_2^{(1)}(M) = O(\sqrt{d})$. By contrast, (10) shows that $\gamma_2^{(1)}(M)$ is of the order of $\sqrt{\log d}$, which is sharp.

### 3 Setup and assumptions for main results

We now present the general setup and assumptions we will use in the analysis of SAA.
3.1 Ideal optimization versus SAA

Functions and sets. As in the introduction, \( \mathcal{I} \) is a finite set which will index the constraints of our problem. We use \( 0 \notin \mathcal{I} \) to index the objective function and set \( \mathcal{I}_0 := \mathcal{I} \cup \{0\} \).

We are given a set \( Y \subset \mathbb{R}^d \) and functions \( f_i : Y \rightarrow \mathbb{R} \), for \( i \in \mathcal{I}_0 \). We will also write \( f := f_0 \). Given \( \delta \in \mathbb{R} \), we define:

\[
X_\delta := \{ x \in Y : \forall i \in \mathcal{I}, f_i(x) \leq \delta \}.
\]

We also write \( X \) instead of \( X_0 \). Note that \( X = X_\delta = Y \) for all \( \delta > 0 \) when \( \mathcal{I} = \emptyset \). The “ideal” optimization problem we consider is:

\[
f^* := \min_{x \in Y} f(x) \text{ s.t. } f_i(x) \leq 0, \ i \in \mathcal{I}.
\]

(13)

In other words, the feasible set is \( X \), the objective function is \( f = f_0 \) and the value of the problem is \( f^* \). We will always assume implicitly that \( X \neq \emptyset \). We let

\[
x^* \in \arg\min_{x \in X} f(x) \text{ so that } f^* := f(x^*).
\]

In particular, we assume implicitly that our problem always has minimizers. We also use the symbols:

\[
f^*_\delta := \inf_{x \in X_\delta} f(x) \text{ and } \text{gap}(\delta) := |f^*_\delta - f^*| \text{ (when } X_\delta \neq \emptyset).\]

In case the above infimum is attained, we let

\[
x^*_\delta \in \arg\min_{x \in X_\delta} f(x) \text{ so that } f^*_\delta := f(x^*_\delta).
\]

We will need some additional notation. We write \( X_{\delta, \text{act}(i)} \) for the subset of \( X_\delta \) where constraint \( i \) is active:

\[
X_{\delta, \text{act}(i)} := \{ x \in X_\delta : f_i(x) = \delta \}.
\]

We also define the set of points \( x \in X_\delta \) that achieve \( f(x) \leq f^* + \vartheta \):

\[
X^*_{\delta, \vartheta} := \{ x \in X_\delta : f(x) \leq f^* + \vartheta \}.
\]

We set

\[
X^*_{\delta, =\vartheta} := \{ x \in X_\delta : f(x) = f^* + \vartheta \}
\]

and finally

\[
X^*_{\delta, \text{act}(i), \vartheta} := X^*_{\delta, \vartheta} \cap X_{\delta, \text{act}(i)}.
\]
We emphasize that $\delta$ will be omitted from our notation when it is equal to zero. With few exceptions which are clear from context, we reserve the symbols $\delta, \eta$ for feasibility deviations and $\epsilon, \vartheta$ for optimality deviations.

**Randomness.** Let $(\mathcal{X}, \sigma(\mathcal{X}), P)$ denote a probability space. We write $\xi \sim P$ to denote a random element of $\mathcal{X}$ with law $P$. In this paper, $(\xi_k)_{k=1}^N \subset \mathcal{X}$ is an i.i.d. random sample of size $N$ from the probability measure $P$. The $\xi_k$ are defined over a common probability space $(\Omega, \mathcal{A}, P)$ that will be always kept implicit. $\hat{P}$ denotes the empirical measure of the sample:

$$\hat{P} := \frac{1}{N} \sum_{k=1}^N \delta \xi_k.$$ 

Given a measurable function $H : Y \times \mathcal{X} \to \mathbb{R}$ and $y$, we define:

$$E H(y, \cdot) := \int_{\mathcal{X}} H(y, \xi) \, P(d\xi) \text{ and } \hat{E} H(y, \cdot) := \frac{1}{N} \sum_{k=1}^N H(y, \xi_k)$$

to denote the expectation and sample average (respectively) of $H(y, \cdot)$ with $y$ fixed. Our assumptions will be such that the integral over $P$ will always be well defined.

**Sample average approximation.** We are given measurable functions $F_i : Y \times \mathcal{X} \to \mathbb{R}$ for $i \in I_0$. We assume that

$$\forall i \in I_0 \forall x \in Y : E|F_i(x, \cdot)| < +\infty \text{ and } EF_i(x, \cdot) = f_i(x). \quad (14)$$

Write:

$$\hat{F}_i(x) := \hat{E} F_i(x, \cdot) = \frac{1}{N} \sum_{k=1}^N F_i(x, \xi_k)$$

to denote the sample average of $F_i$. Formally, $\hat{F}_i(x)$ is a function of $x$ and the sample, but we omit the sample from our notation. We sometimes write $\hat{F} := \hat{F}_0$. The sample average approximation to problem (13) is:

$$\hat{F}^* := \min_{x \in Y} \hat{F}(x) \text{ s.t. } \hat{F}_i(x) \leq 0, \ i \in I. \quad (15)$$

Intuitively, the $\hat{F}_i$ should give random approximations to the $f_i$ for large $N$, and optimization problems with the $\hat{F}_i$ should be similar to the “ideal” problems involving the $f_i$. Quantifying the extent to which this is true is the goal of this paper. We will need the following analogues of the notation introduced above:

$$\hat{X} := \{x \in Y : \forall i \in I, \hat{F}_i(x) \leq 0\};$$
$$\hat{X}^* : = \{x \in \hat{X} : \hat{F}(x) \leq \hat{F}^* + \vartheta\}.$$
\[ \hat{x}^* \in \arg\min_{x \in \hat{X}} \hat{F}(x). \]

Again, we implicitly assume that the SAA always has solutions.

### 3.2 Assumptions on the random functions

To state our general theorems, we will need some probabilistic assumptions on the random functions \( F_i \). We start with a definition.

**Definition 1** (Good and great random variables) Given \((\sigma^2, \rho) \in \mathbb{R}_+\), a function \( h : \Xi \to \mathbb{R}_+ \) is said to be \((\sigma^2, \rho)\)-good if \( E h(\cdot) \leq \sigma^2 \) and

\[
P \left\{ \hat{E} h(\cdot) > 2\sigma^2 \right\} \leq \rho.
\]

Given \( \sigma^2 > 0, p \geq 2 \) and \( \kappa_p > 0 \), we say that \( h \) is \((\sigma^2, p, \kappa_p)\)-great if \( E h(\cdot) \leq \sigma^2 \) and in addition we have the \( L^p \) norm bound:

\[
\| h(\xi) - E h(\cdot) \|_p \leq \kappa_p \sigma^2.
\]

Any fixed integrable function \( h \geq 0 \) with \( E h(\cdot) \leq \sigma^2 \) is \((\sigma^2, \rho)\)-good when \( N \) is large enough due to the Law of Large Numbers. The point of our definition is to have finite-\( N \) results. The next proposition says that great random variables satisfy a quantitative form of goodness.

**Proposition 1** (Proof in the Appendix) If \( h \) as above is \((\sigma^2, p, \kappa_p)\)-great, it is also \((\sigma^2, \rho)\)-good, with

\[
\rho := \left( c_{bdg} \kappa_p \sqrt{\frac{p}{N}} \right)^p
\]

and \( c_{bdg} \) is a universal constant.

The kind of assumption we will make on the \( F_i \) is described below. In what follows, \( Z \subset Y \) is a subset of \( Y \) containing \( x^*, \sigma^2, \sigma_*^2, \rho > 0, \alpha \in (0, 1], \kappa_p \geq 1 \) and \( p \geq 2 \). Also, \( \| \cdot \| \) is a norm over \( \mathbb{R}^d \).

**Assumption 1** \(((\sigma_*^2, \sigma^2, \alpha, \rho)\)-goodness over \( Z \)) The functions \( \{ f_i \}_{i \in \mathcal{I}_0} \) and \( \{ F_i \}_{i \in \mathcal{I}_0} \) are continuous in \( x \in Y \). Moreover,

1. The maps \( \xi \mapsto (F_i(x^*, \xi) - f_i(x^*))^2 \in \mathbb{R}_+ \) are \((\sigma_*^2, \rho)\)-good for each \( i \in \mathcal{I}_0 \);
2. For each map \( F_i \) with \( i \in \mathcal{I}_0 \), there exists \( L_i : \Xi \to \mathbb{R} \) such that \( L_i^2 \) is \((\sigma^2, \rho)\)-good and:

\[
\forall x, x' \in Z, \forall \xi \in \Xi : |F_i(x, \xi) - F_i(x', \xi)| \leq L_i(\xi) \| x - x' \|^\alpha.
\]

**Assumption 2** \(((\sigma_*^2, \sigma^2, \alpha, p, \kappa_p)\)-greatness over \( Z \)) The functions \( \{ f_i \}_{i \in \mathcal{I}_0} \) and \( \{ F_i \}_{i \in \mathcal{I}_0} \) are continuous in \( x \in Y \). Moreover,
1. The maps $\xi \mapsto (F_i(x^*, \xi) - f_i(x^*))^2 \in \mathbb{R}_+$ are $(\sigma^2, p, \kappa_p)$-great for each $i \in I_0$;
2. For each map $F_i$ with $i \in I_0$, there exists $L_i : \Xi \to \mathbb{R}$ such that $L_i^2$ is $(\sigma^2, p, \kappa_p)$-great and:

$$\forall x, x' \in Z, \forall \xi \in \Xi : |F_i(x, \xi) - F_i(x', \xi)| \leq L_i(\xi) \|x - x'\|^\alpha.$$  

In our main results, we will make one of these two assumptions. For general problems, without a convexity assumption, we will take $Z = Y$. In convex settings, we will take potentially much smaller sets $X_{\delta}^{*:\gamma} \subset X$. Notice that each of the above assumptions implies:

$$\forall i \in I_0, \forall x, x' \in Z : |f_i(x) - f_i(x')| \leq (EL_i(\cdot)) \|x - x'\|^\alpha \leq \sigma \|x - x'\|^\alpha \leq \sigma \|x - x'\|^\alpha,$$  

(16)

that is, the functions $f_i$ are $\alpha$-Hölder continuous over $Z$.

We note the following simple consequence of Proposition 1.

**Proposition 2** (Great implies good; proof omitted) **Assumption 2** implies **Assumption 1** with the same set $Z$, the same parameters $\sigma^2, \sigma^2$, and

$$\rho := \left( c b d g p \sqrt{\frac{p}{N}} \right)^p.$$  

In particular, our assumptions may be satisfied with $\rho$ polynomially small in $N$, even if the random variables involved do not have light tails.

### 3.3 Assumptions on the geometry of the problem

When there are constraints in expectation, the SAA will unavoidably have a different feasible set than the ideal problem. In this section, we present standard assumptions that allow us to bound the difference between the two sets. The first assumption is often used in the analysis of perturbations and algorithms for problems in Optimization and Variational Analysis [7, 18, 32]. In what follows, $\| \cdot \|$ is a norm over $\mathbb{R}^d$ and $d$ is the corresponding set-to-point distance.

The first assumption is of Metric Regularity.

**Assumption 3** (Metric regular feasible (MRF) set) There exists $c > 0$ such that for all $x \in Y$,

$$d(x, X) \leq c \sup_{i \in I} f_i(x)_+.$$  

This assumption is trivially satisfied when $I = \emptyset$.

MRF is related to standard constraint qualifications, e.g. the Slater constraint qualification (SCQ) which ensures that $X$ has a strictly feasible point. For instance, Robinson [36] proved that if the set $Y$ and the functions $f_i$ are convex, then, for some $\eta > 0$,

$$X - 2\eta \neq \emptyset \Rightarrow \text{Assumption 3 holds with } c := \frac{\text{diam}(X)}{\eta}.$$  

\[ \square \text{ Springer} \]
The MRF condition is also true for a larger class of sets which are neither strictly feasible nor convex. One fundamental instance is of a polyhedron, as implied by Hoffmann’s Lemma [16]. We remark here that, in Assumption 3, we restrict our analysis for the case of “Lipschitzian” bounds. Our results can be easily extended to the case of “Hölderian” bounds: for some $\beta > 0$, $d(\cdot, X) \leq \epsilon \sup_{i \in \mathcal{I}} [f_i(x)]^\beta$ (see Section 4.2 in [32]). In that case, MRF holds true for any compact nonconvex $X$ whose constraints are polynomial or real-analytic functions, a deep result implied by Lojasiewicz’s inequality [27]. We refer to Section 4.2 in [32] and references therein.

For convex problems, we will also consider a localized version of the Slater CQ condition. Here, we only require that the set $X^{*, \vartheta}$ be bounded and has an “interior point”.

**Assumption 4** (Localized Slater CQ with convexity (LSCQ)) The set $Y$ is convex and closed, and the functions $\{f_i\}_{i \in \mathcal{I}_0}$ are continuous and convex. Moreover, there exist $\eta_* > 0$ and $\vartheta_*$ such that $X^{*, \vartheta_*}$ is bounded and $X^{-\eta_*} \neq \emptyset$ (that is, there exists $x \in Y$ with $f(x) \leq f^* + \vartheta_*$ and $f_i(x) \leq -\eta_*$ for all $i \in \mathcal{I}$).

Boundedness of $X^{*, \vartheta_*}$ may be guaranteed by usual assumptions. In that case, the Slater CQ, i.e., $X^{-\eta_*} \neq \emptyset$ for some $\eta_* > 0$, implies Assumption 4 with $\vartheta_* \geq \inf_{x \in X^{-\eta_*}} f(x) - f^* = \text{gap}(-\eta_*)$. Assumption 4 allows us to control the complexity of $X^{*, \vartheta}$ in terms of $X^{*, \vartheta}$, for suitable $\vartheta$ and $\delta$; see Lemma 4 below for details.

We conclude this section by noting that in the next Sects. 4–5, the functional $\gamma_2^{(\alpha)}$ is defined with respect to $d$, i.e., the set-to-point distance associated to the norm $\| \cdot \|$ over $\mathbb{R}^d$. See Assumptions 1, 2 and 3.

### 4 Main result for not-necessarily convex problems

In this section we state formally and discuss our main result for SAA where we do not assume convexity. More precisely, we only make continuity and metric regularity assumptions on the functions we consider. Theorem 2 is closely related to previous results in the area. Our main contribution here is to obtain stronger bounds under light-tailedness assumptions, through the use of “generic chaining” and our novel concentration arguments.

**Theorem 2** (General functions and sets; proof in Sect. 10.1) Assume $Y$ is bounded. Additionally, make the assumption of $(\sigma^2_*, \sigma^2, \alpha, \rho)$-goodness over $Y$ (cf. Assumption 1). Given $t \geq 0$, define:

$$\hat{r}_N(t) := \sigma \sqrt{3} \gamma_2^{(\alpha)}(Y) + 6 \sqrt{3} \text{diam}^\alpha(Y) \sqrt{1 + \log(2|\mathcal{I}| + 2) + t} \cdot \sqrt{N}.$$  

Also define:

$$\hat{\delta}_N(t) := \begin{cases} \sigma_* \sqrt{\frac{6(1+\log(2|\mathcal{I}|+2)+t}{N}} + \hat{r}_N(t), & \mathcal{I} \neq \emptyset, \\ 0, & \mathcal{I} = \emptyset. \end{cases}$$
Let $\text{GoodThm}_2(\epsilon_0, t)$ denote the event where the following properties hold:

(a) $\hat{X} \subset X_{\tilde{\delta}_N(t)}$, that is, feasible points of the SAA violate ideal constraints by at most $\tilde{\delta}_N(t)$:

$$\hat{X}^* \subset X_{\tilde{\delta}_N(t)},$$

that is, for any $x \in \hat{X}$ with $\hat{F}(x) \leq \hat{F}^* + \epsilon_0$, we have $f(x) \leq f^* + \epsilon_0 + \text{gap}(-\tilde{\delta}_N(t)) + 2\tilde{\gamma}_N(t)$ and $\max_{i \in \mathcal{I}} f_i(x) \leq \tilde{\delta}_N(t)$ (recall that $\text{gap}(\delta) := |f^* - f^*|$, cf. Sect. 3.1);

(b) $|\hat{F}^* - f^*| \leq \tilde{\delta}_N(t) + \max \{2\tilde{\gamma}_N(t) + \text{gap}(-\tilde{\delta}_N(t)), \text{gap}(\tilde{\delta}_N(t))\}.$

Then $\mathbb{P}(\text{GoodThm}_2(\epsilon_0, t)) \geq 1 - e^{-t} - 2(|\mathcal{I}| + 1)\rho$. If we assume $(\sigma^2_\alpha, \sigma^2, \alpha, p, \kappa_p)$-goodness over $Y$ (cf. Assumption 2) instead of $(\sigma^2_\alpha, \sigma^2, \alpha, p)$-goodness, then one may take $\rho = (c_{bdg}\kappa_p\sqrt{p/N})^p$ above. Finally, if we additionally make the metric regularity assumption (Assumption 3), we have the following inequality whenever $\text{GoodThm}_2(\epsilon_0, t)$ occurs:

$$\forall x \in \hat{X}, \text{d}(X, x) \leq c \tilde{\delta}_N(t).$$

Let us parse this theorem. The error parameter $\tilde{\gamma}_N(t) + \text{gap}(-\tilde{\delta}_N(t))$ controls how good SAA solutions are for the original problem. The related parameter $\tilde{\delta}_N(t) + \tilde{\gamma}_N(t) + \max\{\text{gap}(-\tilde{\delta}_N(t)), \text{gap}(\tilde{\delta}_N(t))\}$ bounds the difference between values of the SAA and the ideal problem. Finally, $\tilde{\delta}_N(t)$ controls just how much SAA feasible points violate the constraints of the original problem, and also (under Assumption 3) how far feasible points of the SAA are from the ideal feasible set. Note that if Assumption 3 holds on the set $X - \delta_\cdot$ for some small $\delta_\cdot > 0$, $\max\{\text{gap}(\delta), \text{gap}(-\delta)\}$ is of the order of $c_\delta\alpha$ for any $\delta \in [0, \delta_\cdot]$.

The main features of these parameters is their dependence on the sample size $N$, the geometry of the problem and the desired probability level. The dependence on $N$ is always of the form $N^{-1/2}$, in contrast with previous analyses of SAA not requiring light tails [14, 20]. The geometry of the set $Y$ comes into play via the diameter of $Y$ and the Gaussian complexity parameter $\gamma_2(\alpha)(Y)$. These parameters are optimal for controlling fluctuations of Gaussian processes, and we show that they may still be used in heavier-tailed settings. Finally, the error bounds depend in a sub-Gaussian fashion on the desired probability level $e^{-t}$, at least when $(2|\mathcal{I}| + 1)\rho \ll e^{-t}$. In that connection, we note $\rho$ decays polynomially with $N$ under the $(\sigma^2_\alpha, \sigma^2, \alpha, p, \kappa_p)$-goodness assumption, if $p$ and $\kappa_p$ is treated as a constant. Therefore, our Theorem 2 does give sub-Gaussian-type error probabilities if the number of constraints satisfies $|\mathcal{I}| \leq N^{p/2-c}$ (with $c > 0$) and $t \leq c \log N$. We expect this to be usually the case in applications. Still, we observe that our assumptions for Theorem 2 are somewhat limiting, as they do not allow for unbounded feasible sets (for example).

Remark 2 A natural question is if there are advantages in considering the SAA feasible set $\hat{X} = \{x \in Y : \hat{F}_i(x) \leq \delta, \forall i \in \mathcal{I}\}$ with a positive slack $\delta > 0$. A corollary of the proof of Theorem 2 is that by choosing $\delta := O(\tilde{\delta}_N(t))$, we can remove $\text{gap}(-\tilde{\delta}_N(t))$ in the bounds of item (b) and (c) above under essentially the same assumptions. Of course, this is of theoretical interest only as the constants in the rate $\tilde{\delta}_N(t)$ are typically unknown.
5 Main result in the convex case

We now consider a situation where Theorem 2 can be improved upon. By assuming that the set $Y$ and the functions $F_i$ are convex and the feasible set satisfies a localized Slater-type condition (Assumption 4), we will see that we can obtain a stronger result, Theorem 3 below. Before we present it, we first discuss some geometrical aspects of the problem, which will explain the somewhat convoluted form of the theorem.

5.1 A preliminary discussion

Throughout this section, we make Assumption 4 that the set $X^*_{-\eta_*} \neq \emptyset$. This implies that there exists a point $x \in X$ in the feasible set of the ideal problem that satisfies the following properties:

1. $\max_{i \in I} f_i(x) \leq -\eta_*$ for some $\eta_* > 0$, that is, all constraints are “far” from being active on $x$;
2. $f(x) \leq f^* + \vartheta_*$ for some $\vartheta_* \geq 0$, that is, $x$ is a near optimizer of the ideal problem.

As noted in the discussion after Assumption 4, we can assume that $\vartheta_* \geq \text{gap}(-\eta*)$. We are especially interested in situations where $X^*_{-\eta_*\vartheta_*}$ is bounded; this is the case for instance if $f$ has a unique minimizer $x^*$, satisfies a growth condition $f(x) - f^* \geq c \min\{|x - x^*|^\beta, 1\}$, for some constants $\beta, c > 0$ and $\eta_*$ small enough. Notice that $X^*_{-\eta_*\vartheta_*}$ can be bounded while the whole set $X$ is unbounded.

We now consider the role of convexity. Recall that $\hat{x}$ is a solution to the SAA. Given $\delta \in (0, \eta_*)$ and $\epsilon \in [\text{gap}(-\delta), \vartheta_*]$. Say that $\hat{x}$ is $(\delta, \epsilon)$-good if:

- no constraint of the original problem is violated by more than $\delta$:
  $$\max_{i \in I} f_i(\hat{x}) \leq \delta;$$
- the objective function at $\hat{x}$ satisfies $f(\hat{x}) \leq f^* + \epsilon$.

We say $\hat{x}$ is $(\delta, \epsilon)$-bad if it is not $(\delta, \epsilon)$-good. What could cause $\hat{x}$ to be bad? Proposition 5, a deterministic result, shows that, if $\hat{x}$ is $(\delta, \epsilon)$-bad, then there exists a point $x \in X^*_{-\delta}$ where $\hat{F}_i(x) - f_i(x)$ is “large” for some $i \in I \cup \{0\}$ (recall that $i = 0$ corresponds to the objective function). That is, if the SAA solution is bad, this is due to a failure of concentration of the SAA functions around their ideal counterparts. Most importantly, this failure must happen in the set $X^*_{-\delta}$, which will often be much smaller than $X$ (it is at most as large as $X^*_{-\eta_*\vartheta_*}$). This is what we mean by localization: failure of the SAA manifests itself at “small scales”.

As a second step, we further analyze the set $X^*_{-\delta}$. It will follow from Lemma 4 that

$$X^*_{-\delta} \subset 2X^*_{-\epsilon} - x_{-\delta},$$

for some point $x_{-\delta} \in X^*_{-\delta}$. This means that $X^*_{-\delta}$ is contained in a homothetic copy of $X^*_{-\epsilon}$. As noted above, if $f$ satisfies a growth assumption, the diameter of $X^*_{-\epsilon}$ goes to 0 as $\epsilon \searrow 0$. In particular, this will mean that $X^*_{-\delta}$ is also small.
The upshot of our discussion so far is this. Suppose we can suitably guarantee that, with high probability, we have that the sample averages $\hat{F}_i(x)$ are uniformly close to $f_i(x)$ for all $i \in I \cap \{0\}$, for all $x \in X^{*,e}_δ \subset 2X^{*,e} - x_{-δ}$. Then it follows that the $\hat{x}$ is $(δ, ε)$-good.

How does one choose $δ$ and $ε$ that are as small as possible, while ensuring that $\hat{x}$ is $(δ, ε)$-good with high probability? As it turns out, this is somewhat tricky. To a first approximation, we should expect that:

$$\sup_{x \in X^{*,ε}_δ} |\hat{F}_i(x) - f_i(x) - \hat{F}_i(x_*) + f(x_*)| \approx σ(ε, δ) γ(α) 2^{\sqrt{N}}, \tag{17}$$

where $σ(ε, δ)$ is a term pertaining to the Lipschitz or Hölder constants of the functions $\hat{F}_i$ over the set $X^{*,ε}_δ$. The conditions we need are that these and other random quantities are smaller than both $δ$ and $ε$, so that the “noise” terms do not overwhelm the “signal” in the SAA. Such difficulties also appear in the literature on localization in Statistics and Machine Learning [22, 29], and lead to somewhat convoluted statements. This literature however assume fixed constraints. Our setting study localization with random constraints and one has to account for the fact that $δ$ and $ε$ are coupled via $ε \in [\text{gap}(−δ), ϑ_\ast]$. In any case, the parameter choices in Theorem 3 will be derived from variants of the above reasoning. In most typical situations, one has available upper bounds on the “local complexities” defined in the right hand side of (17). See e.g. (12). In this case, the above reasoning leads to solving a “fixed-point” equation in $(δ, ε)$. While difficult to solve in general, sufficient upper bounds can be obtained by solving inequalities in $(δ, ε)$. We exemplify this reasoning in Sects. 5.3 and 6.

5.2 The theorem

We can now state the main result of this section.

**Theorem 3** (Convex sets and functions; proof in Sect. 10.2) Make Assumption 4 with constants $η_\ast, ϑ_\ast$. Also assume $(σ^2, ϑ, δ), α, ρ$-goodness over the set $Z = X^{*,δ}$ for every choice of $(ϑ, δ) \in [0, ϑ_\ast] \times [0, η_\ast]$ (cf. Assumption 1), where $σ^2(ϑ, δ)$ depends continuously on $ϑ$ and $δ$. Fix parameter $t \geq 0$. For every $0 < ε \leq ϑ_\ast$ and $0 < δ < η_\ast$ satisfying $ε + \text{gap}(−δ) \leq ϑ_\ast$, set:

$$\hat{w}_N(t; δ; ε) := σ(ε + \text{gap}(−δ); δ) \left\{ 4\sqrt{3}^{(α)}(X^{*,e+\text{gap}(−δ)}) \right\} \sqrt{N} + 6\sqrt{3}^{(α)}(X^{*,e+\text{gap}(−δ)}) \sqrt{1 + \log(2|I| + 2) + t} \right\},$$

For $(ε, δ)$ as above, we define parameters $δ(t; ε)$ and $\hat{w}(t; ε)$ as follows.
1. If \( I = \emptyset \) (there are no constraints in expectation), then \( \tilde{\delta}(t; \epsilon) := 0 \) and \( \tilde{w}(t; \epsilon) = \hat{w}_N(t; 0; \epsilon) \).

2. Otherwise, assume that

\[
S_{N, \eta_\ast}(t; \epsilon) := \left\{ \delta \in (0, \eta_\ast) : \epsilon + \text{gap}(-\delta) \leq \vartheta_\ast, \right.
\]
\[
\hat{w}_N(t; \delta; \epsilon) + \sigma_\ast \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}} < \delta \}
\]

is nonempty, and define

\[
\tilde{\delta}(t; \epsilon) := \inf S_{N, \eta_\ast}(t; \epsilon) \quad \text{and} \quad \tilde{w}(t; \epsilon) := \hat{w}_N(t; \tilde{\delta}(t; \epsilon); \epsilon).
\]

Now, fix \( \epsilon_0 \in [0, \vartheta_\ast) \) and assume the set

\[
R_{N, \eta_\ast}(t; \epsilon_0) := \{ \epsilon \in (\epsilon_0, \vartheta_\ast] : \epsilon > \epsilon_0 + \text{gap}(-\tilde{\delta}(t; \epsilon)) + 2\tilde{w}(t; \epsilon) \}
\]

is nonempty so that

\[
\tilde{r}(t; \epsilon_0) := \inf R_{N, \eta_\ast}(t; \epsilon_0)
\]

is well defined. Also set

\[
\tilde{\delta}(t) := \lim_{\epsilon \searrow \tilde{r}(t; \epsilon_0)} \tilde{\delta}(t; \epsilon).
\]

Now define

\[
\text{GoodThm.3}(t, \epsilon_0)
\]

as the event where the following properties all hold.

(a) \( \hat{X}^{*, \epsilon_0} \subset X^{*, \tilde{r}(t; \epsilon_0)} \); that is, all \( x \in \hat{X} \) with \( \hat{F}(x) \leq \hat{F} + \epsilon_0 \) also satisfy \( f(x) \leq f^* + \tilde{r}(t; \epsilon_0) \) and \( \max_{i \in I} f_i(x) \leq \tilde{\delta}(t) \);

(b) the values of the SAA and the ideal problem satisfy:

\[
|\hat{F}^* - f^*| \leq \sigma_\ast \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}} + \frac{\tilde{r}(t; \epsilon_0)}{2} + \max\{\tilde{r}(t; \epsilon_0), \text{gap}(\tilde{\delta}(t))\}.
\]

(c) for all \( x \in \hat{X}^{*, \epsilon_0} \),

\[
d(x, X) \leq \min \left\{ \frac{\text{diam}(X^{*, \vartheta_\ast}) \tilde{\delta}(t)}{\eta_\ast}, 2\text{diam}(X^{*, \tilde{r}(t; \epsilon_0)}) \right\}.
\]

Then

\[
P(\text{GoodThm.3}(t, \epsilon_0)) \geq 1 - e^{-\epsilon} - 2(|I| + 1)\rho.
\]

If we assume instead \( (\sigma_\ast^2, \sigma^2(\vartheta, \delta), \alpha, p, \kappa_p) \)-greatness of the functions \( F_i \) (cf. Assumption 2) instead of \( (\sigma_\ast^2, \sigma^2(\vartheta, \delta), \alpha, \rho) \)-goodness, then one may take \( \rho = (c_{\text{bgd}} \kappa_p \sqrt{p/N})^p \) above.
The comments we made on Theorem 2 on probabilities of error still apply. However, the statement of Theorem 3 is more convoluted. To begin with, the sets \( S_{N, \eta_\alpha}(\cdot; \cdot) \) and \( R_{N, \eta_\alpha}(\cdot; \cdot) \) essentially constrain the choices of \( \delta \) and \( \epsilon \) so that (in the parlance of the preliminary discussion) the “signal” terms are always larger than the stochastic “noise” in the SAA. Nonemptyness of these sets, which is assumed in Theorem 3, is a consequence of a lower bound on the sample size \( N \). The infima taken over these sets in Theorem 3 correspond to trying to find the smallest possible \( \delta \) and \( \epsilon \) to which our reasoning applies, which are given by \( \breve{\delta}(t) \) and \( \breve{\epsilon}(t; \epsilon_0) \) (respectively).

As with the obtained rates, a localized lower bound on \( N \) can be obtained by using the control on the quantities \( \text{gap}(\cdot; \cdot) \), \( \sigma(\epsilon + \text{gap}(\cdot; \cdot); \cdot) \), \( \gamma_2^{(\alpha)}(X^*, \epsilon + \text{gap}(\cdot; \cdot)) \) and \( \text{diam}(X^*, \epsilon + \text{gap}(\cdot; \cdot)) \) and solving the inequalities defining \( S_{N, \eta_\alpha}(\cdot; \cdot) \) and \( R_{N, \eta_\alpha}(\cdot; \cdot) \). This will be exemplified in Sects. 5.3 and 6.

Let us now discuss \( \breve{\delta}(t) \). Basically, this parameter controls fluctuations in the random constraints of the SAA. On the one hand, if we assume

\[
\mathcal{I} = \emptyset \quad \text{or} \quad \max_{i \in \mathcal{I}} f_i(x^*) \leq -\eta_0, \quad \text{where} \quad \eta_0 := \sigma \sqrt{6(1+\log(4|\mathcal{I}|+2)+t)/N},
\]

we may then take \( \eta_* := \eta_0 \) and note \( x^* \in X^*_{-\eta_*} \), so that \( \text{gap}(\cdot; \cdot) = 0 \) for all \( 0 \leq \delta \leq \eta_* \). Intuitively, what this means is that \( x^* \) satisfies the constraints with enough slack that it is nearly certain to be feasible for the SAA, in which case the random constraints do not matter much.

Now assume (18) does not hold. This means that there are random constraints and \( x^* \) is on or near the boundary of the feasible set of the ideal problem. In particular, it may not be feasible for the SAA. However, the existence of a point \( x_{-\eta_*} \in X^*_{-\eta_*} \) gives stability results. Lemma 4 below implies:

\[
\forall 0 \leq \delta < \eta_* : \exists x_{-\delta} \in X^*_{-\delta} : \|x_{-\delta} - x^*\| \leq \frac{\text{diam}(X^*_{-\delta}) \delta}{\eta_* - \delta}. \tag{19}
\]

The goodness assumption over \( Z = X^*_{-\delta} \) in Theorem 3, i.e., (16) with \( \sigma := \sigma(\bar{\theta}_*, \cdot) \), gives

\[
\text{gap}(\cdot; \cdot) \leq f(x_{-\delta}) - f(x^*) \leq \sigma \text{diam}^{\alpha}(X^*_{-\delta}) \left( \frac{\delta}{\eta_* - \delta} \right)^{\alpha}. \tag{20}
\]

One can then use this bound on \( \text{gap}(\cdot; \cdot) \) and the regularity conditions of the objective function and constraints to obtain upper bounds on \( \tilde{r}(t; \epsilon_0) \) and \( \tilde{\delta}(t) \). In general, this may lead to bounds that can be significantly larger than when (18) holds. We will see in Sect. 6.2 (especially in Remark 5) that such larger bounds are unavoidable in general even for simple metric projection problems.

**Remark 3** As in Remark 2, there are advantages in considering \( \widehat{X} = \{x \in Y : \widehat{f}_i(x) \leq -\delta, \forall i \in \mathcal{I} \} \) with a slack \( \delta > 0 \). A corollary of the proof of Theorem 3 is that by
taking \( \delta := O(\hat{\delta}(t)) \) with similar assumptions, it is possible to improve item (a) to 
\( \hat{X}_{\delta, \epsilon_0} \subset X^*, \hat{f}(t; \epsilon_0) \) and remove gap(\( \hat{\delta}(t) \)) in the bound of item (b). Again, tuning \( \delta \) to the order of \( \hat{\delta}(t) \) is of theoretical interest only as the latter is typically unknown.

5.3 Two instructive particular cases

We finish this section with an application of the general Theorem 3 when the solution set satisfies a local regularity condition. The purpose here is to further clarify the usefulness of “localization” (as discussed in Sect. 5.1) in a typical setting in stochastic convex optimization.

**Assumption 5** (Locally regular solution set) Suppose that there exist \( c > 0, \kappa \in (0, 1] \) and \( \vartheta_* > 0 \) such that for all \( \epsilon \in (0, \vartheta_*] \),

\[
\text{diam}(X^{*, \epsilon}) \leq c\epsilon^\kappa + \text{diam}(X^*).
\]

In the following,

\[
\mathcal{C}_\alpha := \sup_{\epsilon \in [0, \vartheta_*]} \left( \frac{\gamma_2^{(\alpha)}(X^{*, \epsilon})}{\text{diam}^{\alpha}(X^{*, \epsilon})} \right)^2.
\]

A conservative bound is \( \mathcal{C}_\alpha \leq C_\alpha \) for a constant \( C_\alpha \) that depends only on \( \alpha \). See Sect. 2.2.

Assumption 5 deserves some discussion. One typical instance of Assumption 5 is when \( f \) is strongly convex on \( Y \) (in this case, \( \kappa = 1/2 \)). More generally, Assumption 5 is implied when \( f \) is locally strongly convex on an open neighbourhood of \( X^* \).\(^3\)

Other important instance when Assumption 5 holds is when the problem has (local) weak sharp minima \([10, 11]\) (in this case with \( \kappa = 1 \)).

We present two results, one when the feasible set \( X \) is fixed and the second when \( X \) has random constraints.

**Proposition 3** (Fixed feasible set) Assume that \( \mathcal{I} = \emptyset \) (that is, there are no random constraints). Grant Assumption 5 with constants \( c, \kappa \) and \( \vartheta_* \). The set \( Y \) is convex and closed, and the functions \( \{f_i\}_{i \in \mathcal{I}_0} \) are continuous and convex. Assume \((\sigma_2^2, \sigma^2, \alpha, \rho)\)-goodness over the set \( Z = X^{*, \vartheta_*} \) (cf. Assumption 1). Let \( t > 0 \) and \( \epsilon_0 \in [0, \vartheta_/2] \).

Define \( \phi_N(t) := \sqrt{(\mathcal{C}_\alpha + t)/N} \), and, suppose \( N \) large enough so that, for an absolute constant \( C > 0 \),

\[
\epsilon_0 + 2C\sigma \left( c^-\vartheta_* + \text{diam}^{\alpha}(X^*) \right) \cdot \phi_N(t) \leq \vartheta_/2. \tag{21}
\]

Define

\[
\mathfrak{R}_N := \begin{cases} 2\epsilon_0 + 4C\sigma \text{diam}^{\alpha}(X^*) \cdot \phi_N(t), & \text{if } \alpha \kappa = 1, \\ \max \left\{ 2\epsilon_0 + 4C\sigma \text{diam}^{\alpha}(X^*) \cdot \phi_N(t), [4C\sigma \phi_N(t)]^{\frac{1}{1-\alpha \kappa}} \right\}, & \text{if } \alpha \kappa \in (0, 1). \end{cases}
\]

\(^3\) When \( \mathcal{I} \neq \emptyset \), we assume without too much loss in generality in Assumption 5 that \( \epsilon \in (0, \vartheta_/] \) with \( \vartheta_/ \) as in Assumption 4. For instance, in case \( f \) is locally strongly convex on a neighbourhood \( U \) of \( X^* \) and Assumption 4 holds, the existence of a \( x \in X_{\vartheta_/ \eta_*} \cap U \) is a mild requirement.
Finally, define \( \text{GoodProp.3}(t, \epsilon_0) \) as the event where the following properties all hold.

(a) \( \hat{X}^{*,\epsilon_0} \subseteq X^*, \mathcal{R}_N \); that is, all \( x \in \hat{X} \) with \( \hat{F}(x) \leq \hat{F} + \epsilon_0 \) also satisfy \( f(x) \leq f^* + \mathcal{R}_N \) and \( x \in X \);
(b) the values of the SAA and the ideal problem satisfy:

\[
|\hat{F}^* - f^*| \leq \sigma_* \sqrt{\frac{6(1 + \log 2 + t)}{N}} + \frac{3}{2} \mathcal{R}_N.
\]

Then \( \mathbb{P}(\text{GoodProp.3}(t, \epsilon_0)) \geq 1 - e^{-t} - 2\rho \).

If we assume instead \( (\sigma_*^2, \sigma^2, \alpha, \rho, \kappa) \)-greatness of the functions \( F_i \) (cf. Assumption 2) instead of \( (\sigma_*^2, \sigma^2, \alpha, \rho) \)-goodness, then one may take \( \rho = (c_{bdg}\kappa \sqrt{p/N})^p \) above.

**Proof** We only present a proof sketch. Denote \( \phi := \phi_N(t) \). As there is no random constraints, \( \text{gap}(\cdot; \delta) \equiv 0 \) and we may set \( \delta(t, \epsilon) \equiv 0 \). In particular, \( \delta(t) = 0 \) and, by the definitions of \( C_{\alpha}, \phi_N(t) \) and \( \hat{w}(t; \epsilon) := \hat{w}(t; 0; \epsilon) \),

\[
\hat{w}(t; \epsilon) = C\sigma\hat{\phi}\text{diam}^\alpha(X^*; \epsilon) \leq C\sigma\hat{\phi}\text{diam}^\alpha(X^*),
\]

for some constant \( C > 0 \) and for all \( \epsilon \in (0, \vartheta_*) \) by Assumption 5. Let us define \( A_N(\epsilon) := C\sigma\epsilon^\alpha\phi\hat{\epsilon}^\alpha \) and \( B_N := C\sigma\text{diam}^\alpha(X^*)\phi \).

Recall that \( \hat{r}(t; \epsilon_0) \) is the infimum over \( \epsilon \) with constraints \( \epsilon_0 + 2\hat{w}(t; \epsilon) < \epsilon \leq \vartheta_* \). An upper bound on \( \hat{r}(t; \epsilon_0) \) is obtained by considering the infimum over the smaller set \( R \) defined by \( \epsilon \) such that \( \epsilon_0 + 2B_N + 2A_N(\epsilon) < \epsilon \leq \vartheta_* \). If (21) holds then \( R \neq \emptyset \). Suppose first \( \kappa\alpha \in (0, 1) \). A simple calculation yields \( \hat{r}(t; \epsilon_0) \leq \max\{2\epsilon_0 + 4B_N, (4C\epsilon^\alpha\phi)(1-\alpha/\epsilon)\} \).

Suppose now \( \kappa\alpha = 1 \). Again using (21), a simple calculation shows that \( \hat{r}(t; \epsilon_0) \leq 2\epsilon_0 + 4B_N \). This finishes the proof.

For instance, in case of quadratic growth (\( \kappa = 1/2 \)) and Lipschitz continuity (\( \alpha = 1 \)), one has the optimality slackness \( \mathcal{R}_N \) of the order \( \epsilon_0 + \sigma^2(\mathcal{C}_\alpha + t)/N \). A notable feature of Proposition 3 is that the “rate” \( \mathcal{R}_N \) on the sample size \( N \) is independent of \( \text{diam}(X) \). In particular, it allows unbounded \( X \). This is in large contrast with the bounds obtained in Theorem 2 in the general non-convex case. “Localization”, implied by convexity, is the technique allowing for such sharper rates. When \( \kappa\alpha = 1 \) and the solution is unique, so that \( \text{diam}(X^*) = 0 \), Proposition 3 implies that for large enough \( N \), the SAA solution is an exact solution of the original problem with high probability. For convex piece-wise linear programs, this was been observed in [44].

We now consider the case of random constraints.

**Proposition 4** (Random feasible set) Make Assumption 4 with constants \( \eta_*, \vartheta_* \). Also assume \( (\sigma_*^2, \sigma^2, \alpha, \rho) \)-goodness over the set \( Z = X^*_{\delta_*} \) (cf. Assumption 1). Grant Assumption 5 with constants \( c, \kappa \) and \( \vartheta_* \). Assume (for simplicity) that \( 0 < \alpha\kappa < 1 \).

Let \( t > 0 \) and \( 0 \leq \epsilon_0 < \min\{\vartheta_*/4, 1/2\} \).

\( \mathfrak{C} \) Springer
Then there is constant $\mathcal{C} > 1$ depending only on $\eta_\ast, \vartheta_\ast, \sigma, \sigma_\ast, \alpha, \kappa, \epsilon$ and $\text{diam}(X^*)$ for which the following statement holds. Let

$$\phi_N(t) := \sqrt{(C_\alpha + \log |\mathcal{I}| + t)/N},$$

and assume that $N$ is large enough so that $\mathcal{C}\phi_N(t) \leq 1$. Let

$$\mathcal{R}_N := \max \left\{ 2\epsilon_0, C\phi_N^\alpha(t) \right\}, \quad \text{and} \quad \mathcal{D}_N := C\phi_N(t) + C\phi_N(t) \max \left\{ \epsilon_0^{\alpha\kappa}, \phi_N^{\alpha^2\kappa}(t) \right\}. $$

Finally, define $\text{GoodProp}_4(t, \epsilon_0)$ as the event where the following properties all hold.

(a) $\hat{X}_{e_0} \subset X_{\mathcal{R}_N}$; that is, all $x \in \hat{X}$ with $\hat{F}(x) \leq \hat{F} + \epsilon_0$ also satisfy $f(x) \leq f^* + \mathcal{R}_N$ and $\max_i \{ \hat{f}_i(x) \} \leq \mathcal{D}_N$;

(b) the values of the SAA and the ideal problem satisfy:

$$|\hat{F}^* - f^*| \leq \sigma_\ast \sqrt{6(1 + \log(2|\mathcal{I}| + 2)) + \frac{\mathcal{R}_N}{2} + \max\{\mathcal{R}_N, \text{gap}(\mathcal{D}_N)\}}.$$

(c) for all $x \in \hat{X}_{e_0},$

$$d(x, X) \leq \min \left\{ \frac{\text{diam}(X^*_{\vartheta_\ast}) \mathcal{D}_N}{\eta_\ast}, 2\text{diam}(X^*, \mathcal{R}_N) \right\}.$$

Then

$$\mathbb{P}(\text{GoodProp}_4(t, \epsilon_0)) \geq 1 - e^{-t} - 2(|\mathcal{I}| + 1)\rho.$$

If we assume instead $(\sigma^2_\ast, \sigma^2, \alpha, \rho, \kappa)$-greatness of the functions $F_i$ over $Y$ (cf. Assumption 2) instead of $(\sigma^2_\ast, \sigma^2, \alpha, \rho)$-goodness, then one may take $\rho = (c_{bdg}\kappa\rho \sqrt{p}/N)^p$ above.

**Proof** Let $0 < \epsilon \leq \min\{\vartheta_\ast/2, 1\}$ and $0 < \delta \leq \min\{\eta_\ast/2, 1\}$ with $\epsilon_1 := \epsilon + \text{gap}(-\delta) \leq \vartheta_\ast$. We first need to bound the quantity $\hat{\omega}_N(t; \delta; \epsilon)$ which is tantamount bounding the quantities $\text{gap}(-\delta), \text{diam}(X^*_{\epsilon_1})$ and $\gamma^{(\alpha)}(X^*_{\epsilon_1})$. In the following, $C > 0$ is an absolute constant and $\mathcal{C} > 0$ is a constant depending on $\eta_\ast, \vartheta_\ast, \sigma, \sigma_\ast, \alpha, \kappa, \epsilon$ and $\text{diam}(X^*)$ that may change from line to line. We use the abbreviation $\phi := \phi_N(t)$.

From Assumption 5, (19), (20) and $\delta \leq \eta_\ast/2$,

$$\text{gap}(-\delta) \leq \mathcal{C}\delta^\alpha. \quad (22)$$

By Assumption 5 and $\epsilon_1 = \epsilon + \text{gap}(-\delta) \leq \vartheta_\ast$,

$$\text{diam}(X^*_{\epsilon_1}) \leq c(\epsilon + \text{gap}(-\delta))^k + \text{diam}(X^*) \leq c(\epsilon^k + \mathcal{C}\delta^\alpha k) + \text{diam}(X^*).$$
From definition of $\mathcal{C}_\alpha$, $\phi_N(t)$ and $\tilde{w}_N(t; \delta; \epsilon)$,

$$\tilde{w}_N(t; \delta; \epsilon) \leq C \sigma \text{diam}^\alpha(X^*, \epsilon) \phi \leq \mathcal{C} \left( \epsilon^{k\alpha} + \delta^{\alpha^2\kappa} + \text{diam}^\alpha(X^*) \right) \phi. \quad (23)$$

**Upper bound on $\tilde{\delta}(t; \epsilon)$.** From (23), the last inequality defining $S_{N, \eta_*}(t; \epsilon)$ is satisfied if $\delta > A_N \delta^{\alpha^2\kappa} + B_N(\epsilon)$, with $A_N := \mathcal{C} \phi$ and $B_N(\epsilon) := \mathcal{C}(1 + \text{diam}^\alpha(X^*) + \epsilon^{k\alpha}) \phi$. Let

$$\delta_N(\epsilon) := \max\{2B_N(\epsilon), (2A_N)^{1/(1-\alpha^2\kappa)}\}.$$  

First, $\delta_N(\epsilon)$ belongs to the set $\{\delta : \delta > A_N \delta^{\alpha^2\kappa} + B_N(\epsilon)\}$. Second, one has $0 < \delta_N(\epsilon) \leq \eta_* / 2$ if $\mathcal{C} \phi \leq 1$ for large enough $\mathcal{C} \geq 1$. Thirdly, from (22) and $\epsilon \leq \theta_* / 2$, the constraint $\epsilon + \text{gap}(\delta_N(\epsilon)) \leq \theta_*$ is satisfied asking for $\mathcal{C} \phi_N(t) \leq 1$ for possibly larger $\mathcal{C}$. We thus conclude that $\delta_N(\epsilon) \in S_{N, \eta_*}(t; \epsilon)$ implying that $\tilde{\delta}(t; \epsilon) = \inf S_{N, \eta_*}(t; \epsilon) \leq \delta_N(\epsilon)$.

**Upper bound on $\tilde{r}(t; \epsilon_0)$.** Fix $0 \leq \epsilon_0 \leq \min\{\theta_* / 4, 1 / 2\}$. Since $\tilde{\delta}(t; \epsilon) \leq \delta_N(\epsilon)$ and the fact that $\delta \mapsto \tilde{w}_N(t; \delta; \epsilon)$ is nondecreasing, we get from (23) and the facts that $\delta_N(\epsilon) \leq \eta_* / 2$ and $\epsilon + \text{gap}(\delta_N(\epsilon)) \leq \theta_*$,

$$\tilde{w}(t; \epsilon) \leq \tilde{w}_N(t; \delta_N(\epsilon); \epsilon) \leq C_N + D_N \epsilon^{k\alpha} + E_N \epsilon^{\alpha^2\kappa^3}, \quad (24)$$

with the definitions

$$C_N := \mathcal{C}(1 + \phi^{\alpha^2\kappa} + A_N^{\alpha^2\kappa/(1-\alpha^2\kappa)}) \phi,$$

as well as $D_N := \mathcal{C} \alpha$ and $E_N := \mathcal{C} \phi^{1-\alpha^2\kappa}$. Moreover, from $\tilde{\delta}(t; \epsilon) \leq \delta_N(\epsilon)$, $\delta_N(\epsilon) \leq \eta_* / 2$, (22) and the fact that $\delta \mapsto \text{gap}(\delta)$ is nondecreasing, one has

$$\text{gap}(\tilde{\delta}(t; \epsilon)) \leq \mathcal{C} \delta_N^\alpha(\epsilon) \leq F_N + G_N \epsilon^{\alpha^2\kappa}. \quad (25)$$

with the definitions $F_N := \mathcal{C} \alpha^{\alpha^2\kappa} \phi^{\alpha} + \mathcal{C} A_N^{\alpha/(1-\alpha^2\kappa)}$ and $G_N := \mathcal{C} \phi^{\alpha}$. From (24)–(25), in upper bounding $\inf R_{N, \eta_*}(t; \epsilon_0)$ it is enough to take the infimum over $\epsilon$ belonging to the set

$$R := \left\{ \epsilon : \epsilon_0 + F_N + G_N \epsilon^{\alpha^2\kappa} + 2(C_N + D_N \epsilon^{k\alpha} + E_N \epsilon^{\alpha^2\kappa^3}) < \epsilon \leq \min\left\{ \frac{\theta_*}{2}, 1 \right\} \right\}.$$  

Define

$$\epsilon_N := \max\left\{ 2\epsilon_0, 8F_N + 16C_N, (16D_N)^{1/(1-\alpha^2)}, (8G_N)^{1/(1-\alpha^2)}, (16E_N)^{1/(1-\alpha^2^2)} \right\}.$$  

It is straightforward to check that one gets $\epsilon_N \in R$ as long as $\epsilon_N < \min\{\theta_* / 2, 1\}$. This requirement follows from $\epsilon_0 < \min\{\theta_* / 4, 1 / 2\}$ and the fact that $\mathcal{C} \phi \leq 1$ for enough large $\mathcal{C}$. We conclude that $\tilde{r}(t; \epsilon_0) = \inf R_{N, \eta_*}(t; \epsilon_0) \leq \epsilon_N$. 

 Springer
Upper bound on \( \tilde{d}(t) \). Letting \( \epsilon \wedge \tilde{r}(t; \epsilon_0) \) and using monotonicity, we obtain from \( \tilde{d}(t; \epsilon) \leq \delta_N(\epsilon) \) that \( \tilde{d}(t) \leq \delta_N(\tilde{r}(t; \epsilon_0)) \leq \delta_N(\epsilon) \). Examining the expressions of \( \epsilon_N \) and \( \delta_N(\epsilon_N) \) one may check that \( \epsilon_N \leq \mathcal{R}_N \) and \( \delta_N(\epsilon_N) \leq \mathcal{C}\phi + \mathcal{C}\phi \max\{\epsilon_0^{\alpha}, \phi^{\alpha^2}\} =: \mathcal{D}_N \) by enlarging \( \mathcal{C} \) if necessary.

To finalize, Theorem 3 and \( \tilde{r}(t; \epsilon_0) \leq \mathcal{R}_N \) and \( \tilde{d}(t) \leq \mathcal{D}_N \) entail the claim. \( \Box \)

In case of Lipschitz continuity (\( \alpha = 1 \)), one has the optimality slackness \( \mathcal{R}_N \) of the order \( \epsilon_0 + \mathcal{C}\sqrt{(\mathcal{C}_\alpha + \log|\Omega| + t)/N} \) and the feasibility slackness \( \mathcal{D}_N \) of the order \( \mathcal{C}\sqrt{(\mathcal{C}_\alpha + \log|\Omega| + t)/N} \). These “localized rates” are independent of \( \text{diam}(X) \) allowing for an unbounded \( X \). They do depend however on the diameter and complexity of \( X^*, \partial_* \). Note that for \( \kappa = 1/2 \) these rates are worse than the case of a fixed feasible set (Proposition 3). This rate deterioration implied by random constraints is unavoidable in general (see Remark 5).

6 Application to metric projection problems

In Sect. 5.3 we presented in Propositions 3–4 an application of the localization technique (Theorem 3) in case the solution set satisfies Assumption 5. Still, in both of these applications, the Hölder modulus is assumed “uniform” in the sense that \( \sigma^2(\vartheta, \delta) \equiv \sigma^2 \) is constant. In this section we present another application where it is important to consider that the Hölder modulus \( \sigma^2(\vartheta, \delta) \) varies across the feasible set. The road map will be similar to the proof of Proposition 4 with some additional technicalities.

Specifically, we sketch the application of Theorems 2–3 to a simple problem illustrating the difference between the two results. Specifically, we consider a metric projection problem where \( X \subset Y \subset \mathbb{R}^d \) and:

\[
 f_0(x) := \|x - x_0\|^2, \quad \text{with}\|\cdot\| \text{ the Euclidean norm.}
\]

A minimizer \( x^* \) of \( f_0 \) over \( X \) corresponds to the metric projection of \( x_0 \) over \( X \). We set \( f^* := \mathbb{R}^2 \) to be the value of the problem.

As usual, we assume that \( f_0(x) = \mathbb{E} F_0(x, \cdot) = \int_{\mathcal{E}} F_0(x, \xi) \mathbf{P}(d\xi) \). Potential examples include:

1. \( \mathcal{E} = \mathbb{R}_+ \times \mathbb{R}, \xi = (\xi[1], \xi[2])^T \sim \mathbf{P} \) is a random vector with mean \((1, 0)^T \), and \( F_0(x, \xi) := \xi[1]\|x - x_0\|^2 + \xi[2] \);

2. \( \mathcal{E} = \mathbb{R}^d, \xi \sim \mathbf{P} \) is an isotropic random vector, that is, satisfying \( \mathbb{E}(\xi, x)^2 = \|x\|^2 \) for all \( x \in \mathbb{R}^d \). In our setting, \( F_0(x, \xi) := \langle \xi, x - x_0 \rangle^2 \).

In both examples, mild moment conditions on \( \mathbf{P} \) imply that the \((\sigma_2^2, \sigma^2, \alpha, \rho)\)-goodness assumption is satisfied with \( \alpha = 1 \) over any bounded set \( Z \subset Y \), with a value \( \sigma^2 = \sigma^2(Z) \) that will in general depend on the set \( Z \). Recalling that \( \mathbb{B} \) is the unit ball in \( \mathbb{R}^d \), we will assume the following condition:

\[
 \forall r > 0, (\sigma_2^2, \sigma^2, \alpha, \rho)\text{-goodness holds over } Z = x_0 + r\mathbb{B} \text{ with } \sigma = c_0 r, \quad (26)
\]
where \( c_0 > 0 \) is a constant. This condition is compatible with the quadratic growth of \( f_0 \) in our two examples, with a \( c_0 \) that depends on \( P \). For convenience, we assume \( \mathcal{T} \neq \emptyset \).

**Remark 4** The constant \( \sigma^2 \) in the second example above will inevitably depend on \( \|\xi\|^2 \). The expectation of \( \|\xi\|^2 \) is \( d \) under our assumptions, which implies \( c_0 \geq d \) in our assumptions. In this specific setting of \( F_0(x, \xi) = (x - x_0, \xi)^2 \), it has been noticed by Mendelson and others [28, 29] that one can bound the quadratic form \( \hat{F}_0(x) \) from below using very weak assumptions that lead to improved bounds. We will return to this issue in the companion paper [30].

Before continuing, we note the following direct consequence of strong convexity. Note that:

\[
\forall x \in X : f(x) = f^* + 2\langle x - x^*, x - x_0 \rangle + \|x - x^*\|^2 \text{ and } \langle x - x^*, x^* - x_0 \rangle \geq 0. 
\]

From strong-convexity,

\[
\forall \theta > 0 : \forall x \in X^{*, \theta}, \|x - x^*\| \leq \sqrt{\theta},
\]

and, in particular, for some universal constant \( C > 0 \),

\[
\forall \theta > 0 : \text{diam}(X^{*, \theta}) \leq 2\sqrt{\theta} \quad \text{and} \quad \gamma_2^{(1)}(X^{*, \theta}) \leq C\sqrt{d\theta},
\]

(27) using Dudley’s bound in (12). We note that the second bound might be far from sharp in several examples.

**6.1 Application of Theorem 2**

We make the \((\sigma^2_\alpha, \sigma^2, \alpha, \rho)\)-goodness assumption over \( Z = Y \) (cf. Assumption 1). We treat \( \sigma_\alpha \), the parameter \( c_0 \) in (26) as universal constants, and use \( C \) to denote a universal constant that might change from line to line. We set \( \epsilon_0 = 0 \), as we are interested in exact minimizers of the SAA problem.

Recalling that \( R^2 := \inf_{x \in X} \|x - x_0\|^2 \), we obtain that \( \sigma \leq C \left( R + \text{diam}(Y) \right) \) (cf. (26)). The parameters \( \hat{r}_N(t) \) and \( \hat{\delta}_N(t) \) satisfy:

\[
\hat{r}_N(t) \leq \hat{\delta}_N(t) \leq \frac{C \left( R + \text{diam}(Y) \right) \left( \text{diam}(Y) \sqrt{t + \log(|I| + 1)} + \gamma_2^{(1)}(Y) \right)}{\sqrt{N}}.
\]

If \( x^* \) belongs to the relative interior of \( X \) in \( Y \) (ie. there exists \( r > 0 \) with \((x^* + r\mathbb{B}) \cap Y \subset X \), then

\[
\text{gap}(\hat{\delta}_N(t)) \leq \frac{C \left( R + \text{diam}(Y) \right)^2 \left( \text{diam}(Y)^2 (t + \log(|I| + 1)) + (\gamma_2^{(1)}(Y))^2 \right)}{N}.
\]
In general, using (16), the metric regularity condition (Assumption 3) and the above estimate on \( \sigma \) to deduce that\(^4\)

\[
\text{gap}(\hat{\delta}_N(t)) \leq \frac{C \left( R + \text{diam}(Y) \right)^2 (\text{diam}(Y) \sqrt{t + \log(|\mathcal{I}| + 1)} + \gamma_2^{(1)}(Y))}{\sqrt{N}},
\]

where now \( C \) depends on \( c \) from Assumption 3 as well. In order to bound \( \text{gap}(-\hat{\delta}_N(t)) \), we assume a slightly stronger version of Assumption 3:

\[
\exists \eta_* > 0, \forall 0 \leq \delta \leq \eta_*, \forall x \in Y, \quad d(x, X_{-\delta}) \leq c \sup_{i \in \mathcal{I}} [f_i(x) + \delta]_+.
\]

Similarly,

\[
\text{gap}(-\hat{\delta}_N(t)) \leq \frac{C \left( R + \text{diam}(Y) \right)^2 (\text{diam}(Y) \sqrt{t + \log(|\mathcal{I}| + 1)} + \gamma_2^{(1)}(Y))}{\sqrt{N}},
\]

where now \( C \) depends on \( c \) and \( \eta_* \). By Theorem 2, we obtain that with probability

\[
\geq 1 - e^{-t} - (2|\mathcal{I}| + 1)|\rho|
\]

\[
\|\hat{x}^* - x_0\|^2 \leq R^2 + \frac{C \left( R + \text{diam}(Y) \right)^2 (\text{diam}(Y) \sqrt{t + \log(|\mathcal{I}| + 1)} + \gamma_2^{(1)}(Y))}{\sqrt{N}};
\]

\[
\max_{i \in \mathcal{I}} f_i(\hat{x}^*) \leq \frac{C \left( R + \text{diam}(Y) \right)^2 (\text{diam}(Y) \sqrt{t + \log(|\mathcal{I}| + 1)} + \gamma_2^{(1)}(Y))}{\sqrt{N}}.
\]

The bounds above are of the order \( N^{-1/2} \), which coincides with what comes from asymptotic analyses. Other interesting aspects of our results are the explicit dependence on \( \text{diam}(Y), \gamma_2^{(1)}(Y) \) and \( R \). In the next subsection, we show that these bounds can be refined significantly under the assumptions of Theorem 3.

### 6.2 Application of Theorem 3

We now work under the assumptions of Theorem 3 combined with our discussion in the beginning of the section. We treat \( \sigma_*, \eta_*, \rho_* \) and the constant \( c_0 \) in (26) as absolute constants, and use \( C, C_0 > 0 \) to denote generic constants depending only on \( \sigma_*, \eta_*, \rho_* \), \( \text{diam}(X_*, \rho_*) \) and \( c_0 \). In particular, their precise values may be different in each occurrence. We assume without loss on generality that \( R \geq 1 \). We set \( \epsilon_0 = 0 \), as we are interested in exact minimizers of the SAA problem. Fix also \( t > 0 \).

Let \( 0 < \epsilon \leq \min\{\rho_*/2, 1\} \) and \( 0 < \delta \leq \min\{\eta_*/2, 1\} \) with \( \epsilon + \text{gap}(-\delta) \leq \rho_* \). For simplicity let \( \epsilon_1 := \epsilon + \text{gap}(-\delta) \). We first need to bound the quantity \( \hat{w}_N(t; \delta; \epsilon) \) which is tantamount bounding the quantities \( \text{diam}(X_*, \epsilon_1^*) \), \( \gamma^{(1)}(X_*, \epsilon_1^*) \) and \( \sigma(\epsilon_1; \delta) \).

\(^4\) Indeed, if \( x \in X_\delta \) is the metric projection of \( x_\delta^* \) onto \( X \) for some \( \delta > 0 \), by (16) and Assumption 3, we have \( f_* - f_\delta^* \leq f(x) - f(x_\delta^*) \leq \sigma d(x_\delta^*, X) \leq \sigma c \delta \).
Bound on gap(−δ): If \( x_{−\delta} \) is the metric projection of \( x^* \) onto \( X^*_{−\delta} \), (19) above implies \( \|x_{−\delta} − x^*\| ≤ C\delta \), with \( C > 0 \) depending on \( \text{diam}(X^*_{0*}/\eta_*) \). Hence, as \( \|x_0 − x^*\| = R \), we have that \( x^*, x_{−\delta} \) lies in \( x_0 + rB \) with \( r ≤ C(R + \delta) \). From the goodness assumption in (26) with a Lipschitz modulus \( \sigma = c_0r \) over \( x_0 + rB \), we get

\[
gap(−\delta) ≤ f(x_{−\delta}) − f^* ≤ C(R + \delta)\|x_{−\delta} − x^*\| ≤ C(R\delta + \delta^2).
\]

(28)

Bound on \( \text{diam}(X^*,ε_1) \): By Lemma 4, (27) and the previous bound on \( \text{gap}(−\delta) \), we have

\[
\text{diam}(X^*_{δ,ε_1+\text{gap}(−\delta)}) ≤ 2\text{diam}(X^*,ε+\text{gap}(−\delta)) ≤ C\sqrt{\epsilon + \text{gap}(−\delta)} ≤ C(\sqrt{\epsilon} + \sqrt{R\delta} + \delta).
\]

Bound on \( \sigma(ε_1; δ) \): Recall that \( \sigma(ε + \text{gap}(−\delta); δ) \) is the Lipschitz constant over the set \( X^*_{δ,ε+\text{gap}(−\delta)} \). Of course, \( X^*_{δ,ε+\text{gap}(−\delta)} \subset X_δ \subset x_0 + rB \) with \( r = C(\sqrt{\epsilon} + R + \delta) \). The goodness assumption in (26) thus implies

\[
\sigma(ε + \text{gap}(−\delta); δ) ≤ C(\sqrt{\epsilon} + R + \delta).
\]

Now, let us define

\[
φ := \sqrt{\frac{d + \log |Z| + I}{N}}.
\]

Using the above bounds, Dudley’s bound (12), which yields

\[
γ^{(1)}(X^*,ε+\text{gap}(−\delta)) ≤ C\sqrt{d}(\sqrt{\epsilon} + \sqrt{R\delta} + \delta),
\]

and a simple but tedious computation,\(^5\) we obtain

\[
\tilde{w}_N(t; δ; ε) ≤ Cφσ(ε_1; δ)\text{diam}(X^*,ε_1) ≤ C(\sqrt{\epsilon} + R + \delta)(\sqrt{\epsilon} + \sqrt{R\delta} + \delta)
\]

\[
≤ Cφ(R^{3/2}\sqrt{\delta} + \delta^2) + Cφ(R\sqrt{\epsilon} + ε),
\]

(29)

for all \( 0 < ε ≤ \min\{δ_*/2, 1\} \) and \( 0 < δ ≤ \min\{η_*/2, 1\} \) with \( ε + \text{gap}(−\delta) ≤ \delta_* \).

Bound on \( \delta(t; ε) \). With a bound on \( \tilde{w}_N(t; δ; ε) \), we may obtain a sufficient upper bound on \( \delta(t; ε) \) using the definition of \( S_{N,η_*}(t; ε) \). Using that \( \delta^2 ≤ 1 \), for the constraint

\[
δ > \tilde{w}_N(t; δ; ε) + σ_*\sqrt{\frac{1 + \log(2(|Z| + 1)) + I}{N}},
\]

\(^5\) Using \( \sqrt{\epsilon R\delta} ≤ 2\epsilon + 2R\delta \) and \( \epsilon, δ ≤ 1 \).
to hold, it suffices

$$\delta > A\sqrt{\delta} + B(\epsilon),$$  \hfill (30)

with \(A := C\phi R^{3/2}\) and \(B(\epsilon) := C\phi(R\sqrt{\epsilon} + \epsilon + 1)\). First, for (30) to hold, it thus suffices to choose

$$\delta(\epsilon) := \max\left\{4A^2, 2B(\epsilon)\right\}.$$

Second, one has \(0 < \delta(\epsilon) \leq \eta_\epsilon/2\) if \(C_0R^{3/2}\phi \leq 1\) for large enough \(C_0 > 1\). Thirdly, from (28) and \(\epsilon \leq \vartheta_\epsilon/2\), the constraint \(\epsilon + \text{gap}(-\delta_N(\epsilon)) \leq \vartheta_\epsilon\) is satisfied asking for possibly larger \(C_0\). We thus conclude that \(\delta(\epsilon) \in S_{N, \eta_\epsilon}(t; \epsilon)\) implying that

$$\tilde{\delta}(t; \epsilon) = \inf S_{N, \eta_\epsilon}(t; \epsilon) \leq \delta(\epsilon) \leq C\phi^2 R^3 + C\phi(R\sqrt{\epsilon} + \epsilon + 1),$$  \hfill (31)

for all \(0 < \epsilon \leq \min\{\vartheta_\epsilon/2, 1\}\).

**Bound on \(r'(t; 0)\).** Let for all \(0 < \epsilon \leq \min\{\vartheta_\epsilon/2, 1\}\). Recall \(\hat{w}(t; \epsilon) = \hat{w}_N(t; \tilde{\delta}(t; \epsilon); \epsilon)\). We now pursue an upper bound on \(r'(t; 0)\) by checking the definition of \(R_{N, \eta_\epsilon}(t; 0)\). After some computations, using that \(0 < \delta(\epsilon) \leq \min\{\eta_\epsilon/2, 1\}\) and \(\epsilon + \text{gap}(-\delta(\epsilon)) \leq \vartheta_\epsilon\) it follows from monotonicity and (31) and (29) that

$$\hat{w}(t; \epsilon) \leq C\phi^{3/2} R^2 \epsilon^{1/4} + C(\phi^{3/2} R^{3/2} + \phi R)\sqrt{\epsilon} + \frac{\epsilon}{4} + C(\phi^3 + \phi^5 R^6 + \phi^{3/2} R^{3/2} + \phi^2 R^3),$$

where we used that \(0 < \epsilon \leq 1\) and \(C(\phi^3 R^2 + \phi)\epsilon + C\phi^3 \epsilon^2 \leq \epsilon/4\) by enlarging \(C_0\) if necessary. Moreover, by (28) and (31) we get

$$\text{gap}(-\tilde{\delta}(t; \epsilon)) \leq C\phi R^2 \sqrt{\epsilon} + \frac{\epsilon}{4} + C(\phi^4 R^6 + \phi^2 R^4 + R\phi + \phi^2),$$

using \(C(R\phi + \phi^2 R^2)\epsilon + C\phi^2 \epsilon^2 \leq \epsilon/4\) for large enough \(C_0\). From the two previous displays, in order to have

$$2\hat{w}(t; \epsilon) + \text{gap}(-\tilde{\delta}(t; \epsilon)) < \epsilon,$$  \hfill (32)

it is enough that

$$C\phi^{3/2} R^2 \epsilon^{1/4} < \epsilon/10,$$

$$C(\phi^{3/2} R^{3/2} + \phi R + \phi R^2) \sqrt{\epsilon} < \epsilon/10,$$

$$C(\phi^4 R^6 + \phi^5 R^6 + \phi^2 R^4 + \phi^2 R^3 + \phi^{3/2} R^{3/2} + R\phi + \phi^2 + \phi^2) < \epsilon/10.$$

For the above conditions to hold, one may check that it is enough to have

$$\epsilon_N := C[\phi^{3/2} R^{3/2} + (R^4 + 1)\phi^2].$$  \hfill (33)
By enlarging $C_0$ if necessary we may also guarantee the additional constraint $0 < \varepsilon_N \leq \delta_*/2$ as required in $R_{N, \eta_0}(t; 0)$. In conclusion, $\varepsilon_N \in R_{N, \eta_0}(t; 0)$ and hence

$$\tilde{r}(t; 0) = \inf R_{N, \eta_0}(t; 0) \leq \varepsilon_N \leq C[\phi^{3/2} R^{3/2} + (R^4 + 1)\phi^2].$$

**Bound on $\tilde{\delta}(t)$**. Setting $\varepsilon \searrow \tilde{r}(t; 0)$ in (31) and using (33), simple calculation yields

$$\tilde{\delta}(t) \leq C\phi^{3/4} R^{7/4} + C\phi^{5/2} R^{3/2} + C\phi^3 R^4 + C\phi^2 R + C\phi^2 R^3 + C(\phi + \phi^3)$$

$$\leq C(\phi^2 R^3 + \phi^2 R + \phi^{1.75} R^{1.75} + \phi),$$

where we used that $C_0\phi R^{3/2} \leq 1$ for large enough $C_0 \geq 1$.

Recall $\phi = \sqrt{\frac{d + \log |I| + t}{N}}$. From Theorem 3 and the fact that gap($-\tilde{\delta}(t)$) $\leq \tilde{r}(t; 0)$ by (32), we conclude that, for $N$ large enough so that $C_0\phi R^{3/2} \leq 1$, with probability $\geq 1 - e^{-t} - 2|I| + 2), \rho$,

$$\|\tilde{x}^* - x_0\|^2 \leq R^2 + C[\phi^{3/2} R^{3/2} + (R^4 + 1)\phi^2];$$

$$d(x^*, X) \leq C\phi[1 + \phi(R + R^3)] + C\phi^{1.75} R^{1.75}.$$  

For large $d, N$, $f(\tilde{x}^*) - f(x^*)$ decays like $R^{3/2}(d/N)^{3/2} + (R^4 + 1)d/N$. Note that it depends on $\text{diam}(X^*, \tilde{\delta}_*)$ but not on the diameters of $X$ nor $Y$.

Now assume additionally that

$$x_0 = x^* \in X_{-\delta_*} \text{ with } \delta_* = C_* \sqrt{\frac{1 + t + \log(1 + |I|)}{N}} \quad (34)$$

with sufficiently large $C_*; \text{ i.e. } x_0 \text{ is "sufficiently interior" to } X.$ Then we have $R = 0$ and gap($-\delta$) $= 0$ for $0 \leq \delta \leq \delta_*$. One can see that, in this case, the dependence on $0 \leq \delta \leq \delta_*$ disappears in the bounds related to gap($-\delta$). Some calculations then improve our high-probability bound on $\|\tilde{x}^* - x_0\|^2$ to:

$$\|\tilde{x}^* - x_0\|^2 \leq C \left(\frac{d + t + \log(|I| + 1)}{N}\right). \quad (35)$$

whenever $N \geq C_0(d + t + \log(|I| + 1)).$ This is the kind of fast rate expected in strongly convex problem. However, such an improvement requires that $x_0$ be a "sufficiently interior" point of $X$; see Remark 5 below.

Finally, when $N < C_0(d + t + \log(|I| + 1))$ and (34) does not hold, we may still obtain a bound by using:

$$\gamma_2^{(1)}(X^*, \epsilon \text{+gap}(-\delta)) \leq \gamma_2(X^*, \delta_*)$$

for small enough $\epsilon, \delta$. This leads to nontrivial bounds whenever $\gamma_2(X^*, \delta_*) \ll \sqrt{d} \text{ (eg. if } X^*, \tilde{\delta}_* \text{ is contained in a small simplex).}$
Remark 5 We observe that even in one dimension we may expect fluctuations of order $R^2/\sqrt{N}$ on $||\hat{x}^* - x_0||^2 - R^2$ in metric projection problems, when $x_0$ lies outside the feasible set. Assume $\mathcal{Z} = Y = \mathbb{R}$, $\xi \sim P$ is exponential with parameter 1 (that is, $\mathbb{P}\{\xi > t\} = e^{-t}$ for all $t > 0$), $F_0(x, \xi) = f_0(x) = x^2$, $\mathcal{I} = \{1\}$ and:

$$F_1(x, \xi) = R - \xi x.$$  

The solutions to the ideal problem and SAA are $x^* = R$ and $\hat{x}^* = R/\bar{\xi}_N$, where $\bar{\xi}_N$ is the sample average of the $\xi_k$. Using the Central Limit Theorem for $\bar{\xi}_N$, one can show that $\sqrt{N}(\hat{x}^* - R)$ has a Gaussian limit with standard deviation $2R^2$ when $N \to +\infty$.

7 Concentration inequalities for sample averages

We present a novel concentration inequality. The proof of Theorems 2 and 3 rely on this tool which may be of independent interest in stochastic optimization.

Theorem 4 (Proof below) Suppose $(\mathcal{M}, d)$ is a totally bounded metric space. Assume $(\mathcal{Z}, \sigma(\mathcal{Z})), P, \{\xi_k\}_{k=1}^N$ are as in Sect. 3 and $G: \mathcal{M} \times \mathcal{Z} \to \mathbb{R}$ is a measurable function with $E|G(x_0, \cdot)| < +\infty$ for some $x_0 \in \mathcal{M}$. Assume additionally that there exists a measurable function $L: \mathcal{Z} \to \mathbb{R}_+$ with $\mathbb{E}L^2 \leq \nu^2 < +\infty$ and a constant $0 < \alpha \leq 1$ such that:

$$|G(x, \xi) - G(x', \xi)| \leq L(\xi) d(x, x')^\alpha.$$  

Write:

$$\Delta G := \sup_{x \in \mathcal{M}} |(\hat{E} - E)(G(x, \cdot) - G(x_0, \cdot))|$$  

and assume:

$$\mathbb{P}\{\hat{E}L^2(\cdot) > 2\nu^2\} \leq \rho \in [0, 1].$$  

Then, for any $t \geq 0$:

$$\mathbb{P}\left\{\hat{E}L^2(\cdot) \leq 2\nu^2, \Delta G > \frac{4\sqrt{3} \gamma_2^{(\alpha)}(\mathcal{M}, d) + 6\sqrt{3}\text{diam}(\mathcal{M})^\alpha \sqrt{1 + t}}{\sqrt{N}}\right\} \leq e^{-t}.$$  

Notice that, if $N$ grows, $\hat{E}L^2 \to E L^2 \leq \nu^2$ almost surely. Therefore, we expect the probability of $\hat{E}L^2(\cdot) \leq 2\nu^2$ to be large when $N$ is large. The above theorem shows that on the event that $\hat{E}L^2(\cdot) \leq 2\nu^2$, the likelihood of $\Delta G$ being large is exponentially small.

Springer
To prove this result, we will use the next lemma. It is a simple consequence of a much more general result of Panchenko [31].

**Lemma 1** (Proof in Appendix 1) Assume $Z_1, \ldots, Z_N$ are i.i.d. random variables with finite second moments. Then

$$
P \left\{ \frac{1}{N} \sum_{i=1}^{N} (Z_i - \mathbb{E}[Z_1]) > \sqrt{\frac{2(1 + t)}{N} \left( \mathbb{V}[Z_1] + \frac{1}{N} \sum_{i=1}^{N} (Z_i - \mathbb{E}[Z_1])^2 \right)} \right\} \leq e^{-t}.
$$

In particular, if $\mathbb{V}[Z_1] \leq \nu^2$,

$$
P \left\{ \frac{1}{N} \sum_{i=1}^{N} (Z_i - \mathbb{E}[Z_1]) > \nu \sqrt{\frac{6(1 + t)}{N}}, \frac{1}{N} \sum_{i=1}^{N} (Z_i - \mathbb{E}[Z_1])^2 \leq 2\nu^2 \right\} \leq e^{-t}.
$$

This remarkable inequality by Panchenko shows that averages of $Z_i$, when normalized by an empirical term, have sub-Gaussian tails under extremely weak assumptions. We will use this both to prove Theorem 4 and to control fluctuations of other random variables in the proofs of Theorems 2 to 3.

**Proof of Theorem 4** In this proof we use a combination of generic chaining (as encapsulated by Theorem 1) and Panchenko’s self-normalized concentration inequality (Lemma 1).

We begin by noting that, since $\mathbb{E}L^2(\cdot) \leq \nu^2$, for any $u > 0$,

$$
P \left\{ \widehat{\mathbb{E}}L^2(\cdot) \leq 2\nu^2, \Delta G > u \sqrt{\frac{6\nu^2}{N}} \right\} \leq P \left\{ \widehat{\mathbb{E}}L^2(\cdot) + \mathbb{E}L^2(\cdot) \leq 3\nu^2, \Delta G > u \sqrt{\frac{6\nu^2}{N}} \right\}
$$

$$
\leq P \left\{ \Delta G > u \sqrt{\frac{2(\widehat{\mathbb{E}} + \mathbb{E})L^2(\cdot)}{N}} \right\}.
$$

Letting $t \geq 0$ and

$$
u = u_t := 2\sqrt{2} y_2^{(\alpha)}(\mathcal{M}, d) + 3\sqrt{2} \text{diam}(\mathcal{M})^\alpha \sqrt{1 + t},
$$

we see that it suffices to show that:

Sufficient : $P \left\{ \Delta G > u_t \sqrt{\frac{2(\widehat{\mathbb{E}} + \mathbb{E})L^2(\cdot)}{N}} \right\} \leq e^{-t}$. (36)
To prove (36), we will use our “generic chaining” bound, Theorem 1. For each \( x \in \mathcal{M} \), define the random quantity:

\[
Y_x := \frac{(\hat{\mathbb{E}} - \mathbb{E}) G(x, \cdot)}{\sqrt{\frac{2(\hat{\mathbb{E}} + \mathbb{E}) L^2(\cdot)}{N}}},
\]

when the denominator is \( \neq 0 \), or \( Y_x = 0 \) otherwise. Note that:

\[
\Delta G = \sqrt{\frac{2(\hat{\mathbb{E}} + \mathbb{E}) L^2(\cdot)}{N}} \sup_{x \in \mathcal{M}} |Y_x - Y_{x_0}|.
\]

If we can show that:

**Goal**: \( \forall x, x' \in \mathcal{M}, \forall t \geq 0 : \mathbb{P}\{Y_x - Y_{x'} \geq \sqrt{2(1 + t) d(x, x')^\alpha}\} \leq e^{-t} \),

then Theorem 1 gives us (36).

To obtain our goal, we fix \( x, x' \) and \( t \). We will apply Panchenko’s inequality (Lemma 1) to the i.i.d. random variables:

\[
Z_k := G(x, \xi_k) - G(x', \xi_k) (k \in [N]),
\]

so that:

\[
(\hat{\mathbb{E}} - \mathbb{E}) (G(x, \cdot) - G(x', \cdot)) = \frac{1}{N} \sum_{k=1}^{N} (Z_k - \mathbb{E}[Z_k]). \tag{37}
\]

To apply Lemma 1, we will estimate the terms \( \mathbb{V}[Z_1] \) and \( (Z_k - \mathbb{E}[Z_k])^2 \) appearing in that bound. Note that:

\[
\mathbb{V}[Z_1] + (Z_k - \mathbb{E}[Z_k])^2 = \mathbb{E}[Z_k^2] + Z_k^2 - 2Z_k \mathbb{E}[Z_k] \\
\leq \mathbb{E}[Z_k^2] + Z_k^2 + 2 \sqrt{\mathbb{E}[Z_k^2]} \mathbb{E}[Z_k^2] \\
(2\sqrt{xy} \leq x + y \text{ for all } x, y \in \mathbb{R}_+) \leq 2(\mathbb{E}[Z_k^2] + Z_k^2).
\]

Now, by our assumptions:

\[
|Z_k| = |G(x, \xi_k) - G(x', \xi_k)| \leq L(\xi_k) d(x, x')^\alpha,
\]

therefore:

\[
\mathbb{V}[Z_1] + \frac{1}{N} \sum_{k=1}^{N} (Z_k - \mathbb{E}[Z_k])^2 \leq 2 [\hat{\mathbb{E}} + \mathbb{E}] L(\cdot)^2 d(x, x')^{2\alpha}.
\]
We may finally apply Panchenko’s inequality and deduce the following bound:

\[
\mathbb{P} \left\{ \frac{1}{N} \sum_{k=1}^{N} (Z_i - \mathbb{E}[Z_i]) \geq 2 \sqrt{\frac{(1 + t)(\hat{E} + \mathcal{E}) L(.)^2}{N}} d(x, x')^\alpha \right\} \leq e^{-t}. 
\]

This implies our goal once we combine it with (37) and the definition of \(Y_x\). \(\Box\)

### 8 Deviation and localization arguments

This section compiles a series of deterministic results on how the SAA differs from the ideal optimization problem. We consider general sets and functions in Sect. 8.1 and the convex case in Sect. 8.2. We are particularly careful to distinguish lower and upper tails in our bounds, as lower tails can be much better behaved than upper tails. One such setting will be explored in the companion paper [30].

**Remark 6** The results in this section are purely deterministic in the sense that we do not need (3) and (4) to hold. We simply need to assume that \(Y\) is as given in Sect. 3; that \(f_i, \hat{F}_i : Y \to \mathbb{R}\) are functions (with \(i \in \mathcal{I}_0\)) and that \(X, \hat{X}\), etc are defined in terms of the \(f_i\) and \(\hat{F}_i\) as prescribed in Sect. 3. Our results will be the most interesting when the \(\hat{F}_i\) are good approximations to the respective \(f_i\).

For convenience, we introduce the following notation. Given \(x, y \in Y, Z \subset Y\) and \(i \in \mathcal{I}_0\):

\[
\hat{\Delta}_i(x) := \hat{F}_i(x) - f_i(x); \\
\hat{\Delta}_i(y; x) := \hat{\Delta}_i(y) - \hat{\Delta}_i(x). 
\]

### 8.1 General sets and functions

The next lemma is quite straightforward.

**Lemma 2** Given \(\delta, \epsilon, \epsilon_0 \geq 0\), assume \(|\hat{\Delta}_i(y)| \leq \delta\) for all \(y \in Y\) and \(i \in \mathcal{I}\) and also that \(|\hat{\Delta}_0(y; x^*)| \leq \epsilon\) for all \(y \in Y\). Then we have the following.

1. \(X_{-\delta} \subset \hat{X} \subset X_\delta\).
2. \(\hat{X}^{*, \epsilon_0} \subset X^{*, \epsilon_0 + 2\epsilon + \text{gap}(-\delta)}\).
3. \(|\hat{F}^* - f^*| \leq \delta + \max\{2\epsilon + \text{gap}(-\delta), \text{gap}(\delta)\}\).

**Proof** The first item is an immediate consequence of the fact that \(-\delta \leq \hat{\Delta}_i(y) \leq \delta\) for each \(i \in \mathcal{I}\).

Let \(\hat{x} \in \hat{X}^{*, \epsilon_0}\) and \(x^*_\delta \in \text{argmin}_{x \in X_{-\delta}} f(x)\), so that \(f(x^*_\delta) = f^* + \text{gap}(-\delta)\). By item 1, \(x^*_\delta \in \hat{X}\), so

\[
f(\hat{x}) + \hat{\Delta}_0(\hat{x}) = \hat{F}(\hat{x}) \leq \hat{F}^* + \epsilon_0 \leq \hat{F}(x^*_\delta) + \epsilon_0 = f^*_\delta + \hat{\Delta}_0(x^*_\delta) + \epsilon_0.
\]
Therefore, for any \( \hat{x} \in \hat{X}^{*,\epsilon_0} \subset X_\delta \),

\[
    f(\hat{x}) - f^* \leq f^*_\delta - f^* + \Delta_0(x^*_\delta) + \epsilon_0 - \inf_{y \in X_\delta} \hat{\Delta}_0(y)
\]

\[
    \leq \text{gap}(\delta) + \epsilon_0 + \sup_{y \in X_\delta} |\hat{\Delta}_0(y; x^*_\delta)|
\]

\[
    \leq \text{gap}(\delta) + \epsilon_0 + 2 \sup_{y \in X_\delta} |\hat{\Delta}_0(y; x^*)|,
\]

(40)

which gives item 2.

For item 3, we assume for simplicity that some \( \hat{x} \in \hat{X} \) achieves the minimum of \( \hat{F} \): \( \hat{F}(\hat{x}) = \hat{F}^* \). In this case,

\[
    |\hat{F}^* - f^*| = |\hat{F}(\hat{x}) - f(x^*)|
\]

\[
    \leq |\hat{F}(\hat{x}) - f(\hat{x})| + |f(\hat{x}) - f(x^*)|
\]

\[
    \leq \sup_{x \in X_\delta} |\hat{\Delta}_0(x)| + |f(\hat{x}) - f(x^*)|.
\]

In one hand \( f(\hat{x}) - f^* \) is upper bounded by (40) (with \( \epsilon_0 = 0 \)). Since \( \hat{x} \in X_\delta \), a lower bound is given by \( f^* - f(\hat{x}) = \text{gap}(\delta) + f^*_\delta - f(\hat{x}) \leq \text{gap}(\delta) \), finishing the proof.

\[ \square \]

### 8.2 Convex sets and functions

We now consider the convex setting with “localized” bounds.

**Proposition 5** Assume that \( Y \) is convex and closed and that the functions \( \{ f_i \}_{i \in I} \) and \( \{ \hat{f}_i \}_{i \in I_0} \) are all convex and continuous. Given \( \delta^0, \delta > 0 \), assume \( x_{-\delta^0} \in X_{-\delta^0} \). Fix \( \epsilon \geq f(x_{-\delta^0}) - f^* \) and \( \epsilon_0 > 0 \). Let \( \hat{\epsilon} := \hat{F}(x_{-\delta^0}) - \hat{F}^* + \epsilon_0 \). If the following three conditions hold:

\[
    \forall i \in I, \quad \hat{\Delta}_i(x_{-\delta^0}) \leq \delta^0;
\]

(41)

\[
    \forall i \in I, \quad \inf_{x \in X_{\delta}^* \cap \tilde{X}_{\delta}^*} \hat{\Delta}_i(x) > -\delta;
\]

(42)

\[
    \inf_{x \in X_{\delta}^* \cap \tilde{X}_{\delta}^*} \hat{\Delta}_0(x; x_{-\delta^0}) > - (\epsilon - (f(x_{-\delta^0}) - f^*) - \epsilon_0),
\]

(43)

then:

1. \( \hat{X}^{*,\epsilon_0} \subset X_{\delta}^* \), or equivalently, any \( x \in \hat{X} \) with \( \hat{F}(x) \leq \hat{F}^* + \epsilon_0 \) satisfies \( x \in Y \),

   \[ \max_{i \in I} f_i(x) \leq \delta \] and \( f(x) \leq f^* + \epsilon \);

2. The values of the SAA and the ideal problem satisfy

\[
    |\hat{F}^* - f^*| \leq |\hat{\Delta}_0(x^*)| + \sup_{x \in X_{\delta}^*} |\hat{\Delta}_0(x; x^*)| + \max \{ \epsilon, \text{gap}(\delta) \}.
\]
**Proof** The proof consists of three main steps. In the first step, we show that assumption (41) implies \( x_{-\delta} \in \hat{X} \cap X^{*, \epsilon} \). In the second step, we show that if (41) holds and there exists a point \( z \in \hat{X}^{*, \epsilon_0} \setminus X^{*, \epsilon} \), then one of (42) or (43) cannot hold. In contrapositive form, the second step implies that, if we assume the three conditions (41), (42) and (43), then \( \hat{X}^{*, \epsilon_0} \subset X^{*, \epsilon} \). Finally, the third step proves the inequality for \( |\hat{F}^* - f^*| \).

**First step.** Assume (41). We argue that

\[
\hat{F}_i(x_{-\delta}) \leq -\delta + \hat{\Delta}_i(x_{-\delta}) \leq 0 \text{ by (41)}.
\]

We also have \( x_{-\delta} \in X^{*, \epsilon} \) because \( f(x_{-\delta}) - f^* \leq \epsilon \) by assumption.

**Second step.** Assume (41) and also that there exists a point \( z \in \hat{X}^{*, \epsilon_0} \setminus X^{*, \epsilon} \). Since \( X^{*, \epsilon} \) is closed and convex and \( x_{-\delta} \in X^{*, \epsilon} \subset X^{*, \epsilon}_\delta \), the intersection of the line segment \( [x_{-\delta}, z] \) with \( X^{*, \epsilon}_\delta \) is also closed and convex. That is,

\[
[x_{-\delta}, z] \cap X^{*, \epsilon}_\delta = [x_{-\delta}, x] \text{ with } x \in X^{*, \epsilon}_\delta.
\]

In fact we have \( x \in \hat{X}^{*, \epsilon_0} + \hat{F}(x_{-\delta}) - \hat{F}^* \cap X^{*, \epsilon}_\delta \) as well. To see this, note that both \( x_{-\delta} \) and \( z \) belong to \( \hat{X} \), and this set is convex under our assumptions on \( Y \) and the \( f_i \), so \( x \in \hat{X} \). In addition, convexity of \( \hat{F} \) implies:

\[
\hat{F}(x) \leq \max\{\hat{F}(z), \hat{F}(x_{-\delta})\}
\]

(\text{use that } z \in \hat{X}^{*, \epsilon_0} \leq \max\{\hat{F}^* + \epsilon_0, \hat{F}(x_{-\delta})\})

(note that \( x_{-\delta} \in \hat{X} \Rightarrow \hat{F}(x_{-\delta}) \geq \hat{F}^* \) \( \leq \hat{F}(x_{-\delta}) + \epsilon_0 \).

Note that \( x \neq z \) and any point \( x' \in (x, z] \) cannot lie in \( X^{*, \epsilon}_\delta \). It follows that one of the restrictions defining \( X^{*, \epsilon}_\delta \) is active at \( x \). That is, one of the following properties holds:

\[
f(x) = f^* + \epsilon \quad \text{(that is, } x \in X^{*, \epsilon}_0 \cap \hat{X}^{*, \epsilon} \hat{F}(x_{-\delta}) = \hat{F}^* + \epsilon_0) \text{; or} \quad (44)
\]

\[
\exists i \in I : f_i(x) = \delta \quad \text{(that is, } x \in X^{*, \epsilon}_\delta, \hat{F}(x_{-\delta}) = \hat{F}^* + \epsilon_0) \text{.} \quad (45)
\]

If (44) holds, then

\[
f(x) - f(x_{-\delta}) = \epsilon - (f(x_{-\delta}) - f^*) \text{ and } \hat{F}(x) - \hat{F}(x_{-\delta}) \leq \epsilon_0.
\]

Therefore, if (44) holds, we obtain

\[
\hat{\Delta}_i(x; x_{-\delta}) = \hat{\Delta}_0(x) - \hat{\Delta}_0(x_{-\delta}) \leq \epsilon_0 - (\epsilon - (f(x_{-\delta}) - f^*))
\]

which means that (43) does not hold.

Now assume (45) holds. Fix an \( i \in I \) with \( f_i(x) = \delta \). Notice that \( x \in X^{*, \epsilon}_\delta \). Since \( x \in \hat{X} \) and \( \hat{F}_i(x) \leq 0 \), we deduce that \( \hat{\Delta}_i(x) = \hat{F}_i(x) - f_i(x) \leq -\delta \), which means that (42) cannot hold.

\( \Box \) Springer
Third step. We now assume that the three conditions in the Theorem hold. As shown above, this implies item 1 of the theorem. For simplicity, we prove item 2 assuming that some $\hat{x} \in \hat{X}$ achieves the minimum of $\hat{F}$: $\hat{F}(\hat{x}) = \hat{F}^{*}$. By item 1, $\hat{x} \in X^{*}_{\delta}$, so $f(\hat{x}) \leq f(x^{*}) + \epsilon$ and $f(x^{*}) - f(\hat{x}) \leq \text{gap}(\delta)$. Therefore:

$$|\hat{F}^{*} - f^{*}| = |\hat{F}(\hat{x}) - f(x^{*})|$$

$$\leq |\hat{F}(\hat{x}) - f(\hat{x})| + |f(\hat{x}) - f(x^{*})|$$

$$\leq \sup_{x \in X^{*}_{\delta,\epsilon}} |\hat{\Delta}_{0}(x)| + \max\{\epsilon, \text{gap}(\delta)\}$$

$$\leq \sup_{x \in X^{*}_{\delta,\epsilon}} |\hat{\Delta}_{0}(x^{*})| + |\hat{\Delta}_{0}(x^{*})| + \max\{\epsilon, \text{gap}(\delta)\}.$$

9 The effect of small changes in constraints on the feasible set

Our ideal optimization problem (13) naturally involves the feasible set $X$ and the sublevel sets $X^{*,\vartheta_{*}}$. However, it transpires from the previous section that we will need to consider the perturbed sets $X_{\delta}$ and $X^{*,\epsilon}_{\delta}$, where constraints are violated by a small amount. The goal of this section is to show how one can bound the geometry and complexity of the perturbed sets in terms of the corresponding sets for the ideal problem. For this, we will make use of the geometrical assumptions from Sect. 3.3. In what follows, $\| \cdot \|$ is a norm over $\mathbb{R}^{d}$ and $d$ is the corresponding set-to-point distance.

9.1 Small constraint violations under metric regularity conditions

The first result applies to general problems.

**Lemma 3** Make Assumption 3. Let $\mathcal{B}$ denote the unit ball of $\mathbb{R}^{d}$ under its norm $\| \cdot \|$. Then $X_{\delta} \subset X + c\delta \mathcal{B}$.

**Proof** This follows trivially from the Assumption, combined with the fact that $X_{\delta} \subset Y$ and the fact that $f_{i}(x) \leq \delta$ for all $x \in X_{\delta}$. □

9.2 Small constraint violations under convexity

The next lemma is a key contribution of this paper. It shows that, under Assumption 4, one can give a tight control of the relevant complexity parameters of $X^{*,\vartheta_{*}}$ in terms of $X^{*,\vartheta}$, for suitably small $\delta$ and $\vartheta$. Recall that $x^{*}_{\delta}$ minimizes $f$ over $X^{*}_{\delta}$.

**Lemma 4** Make Assumption 4. Then:

1. For all $x \in Y$ with $f(x) \leq f^{*} + \vartheta_{*}$ and all $\vartheta^{0} \in (-\eta_{*}, \eta_{*}]$,

$$d(x, X^{*,\vartheta^{0}}_{\delta}) \leq \frac{\text{diam}(X^{*,\vartheta^{0}}_{\delta})}{\eta_{*} + \delta^{0}} \max_{i \in \mathcal{I}} (f_{i}(x) - \vartheta^{0})_{+}.$$
2. For all $\delta \in [0, \eta_*]$ and all $\vartheta \geq \text{gap}(-\delta)$,

$$X_{\delta}^{*,\vartheta} \leq 2X_{\delta}^{*,\vartheta} - x_{-\delta}^*.$$  

3. For $\delta$ and $\vartheta$ as in item 2,

$$\gamma_2^{(\alpha)}(X_{\delta}^{*,\vartheta}) \leq 2^{\alpha} \gamma_2^{(\alpha)}(X_{\delta}^{*,\vartheta})$$

$$\text{diam}(X_{\delta}^{*,\vartheta}) \leq 2\text{diam}(X_{\delta}^{*,\vartheta}).$$

The Lemma deserves some comments. Item 1 is a translation of a result of Robinson [36] to our setting. Item 2 seems to be new: it states that $X_{\delta}^{*,\vartheta}$ is contained in an homothetic copy of $X_{\delta}^{*,\vartheta}$. This is important because, in principle, all we know from metric regularity is that $X_{\delta}^{*,\vartheta}$ is “close” to $X_{\delta}^{*,\vartheta}$, meaning that $X_{\delta}^{*,\vartheta} \subset X_{\delta}^{*,\vartheta} + c\delta B$ for the unit ball $B$ and some constant $c > 0$. By contrast, item 2 means that the actual shape of $X_{\delta}^{*,\vartheta}$ is controlled by $X_{\delta}^{*,\vartheta}$. As a result, we obtain item 3, which says that the size and complexity of $X_{\delta}^{*,\vartheta}$ are controlled by the intrinsic geometry of $X_{\delta}^{*,\vartheta}$. By contrast, one can show

$$\gamma_2^{(\alpha)}(X_{\delta}^{*,\vartheta} + c\delta B) \approx \gamma_2^{(\alpha)}(X_{\delta}^{*,\vartheta}) + c \delta^\alpha \sqrt{d}$$

for some $c > 0$ depending only on $c$. In other words, metric regularity alone cannot give intrinsic bounds on the complexity of $X_{\delta}^{*,\vartheta}$.

We now prove the Lemma.

**Proof of Lemma 4** We will need the following geometrical fact that essentially comes from Robinson’s paper [36].

**Claim 1** Take $\delta_2 \in (0, \eta_*)$ and $\delta^0 \in (-\delta_2, \delta_2)$ and $\vartheta \geq f_{-\delta_2}^* - f^*$. Consider $x \in Y$ with $f(x) \leq f^* + \vartheta$ and take $r \geq \max_{i \in I} (f_i(x) - \delta^0)_+ \geq 0$. Let $x_{-\delta_2}^* \in X_{-\delta_2}$ be a minimizer of $f$ over that set (which exists under our assumptions of convexity and $X_{-\eta_*} \neq \emptyset$) and take

$$\lambda := \frac{r}{\delta_2 + \delta^0 + r} \in [0, 1).$$

Then $x^{(\lambda)} := (1 - \lambda)x + \lambda x_{-\delta_2} \in X_{\delta^0}^{*,\vartheta}$.

Indeed, it is obvious that $x^{(\lambda)} \in Y$ because this set is convex. We also have that $f(x^{(\lambda)}) \leq f^* + \vartheta$ because $f$ is convex and both $x, x_{-\delta_2}$ satisfy this inequality. Finally, for each $i \in I$,

$$f_i(x^{(\lambda)}) - \delta^0 \leq (1 - \lambda)(f_i(x) - \delta^0) + \lambda(f_i(x_{-\delta_2}) - \delta^0) \leq (1 - \lambda)r - \lambda(\delta_2 + \delta^0) = 0.$$  

So $x^{(\lambda)} \in X_{\delta^0}^{*,\vartheta}$.

We now use this Claim to obtain parts 1 and 2 of the Lemma. We will then obtain part 3 from part 2.
Proof of Lemma 4, part 1. We apply the claim to $x$ as in item 1 with $\delta_2 = \eta_*$, $r := \max_{i \in \mathcal{I}} (f_i(x) - \delta^o)_+$ and $\vartheta = \vartheta_*$. In that case, we see that:

$$x^{(\lambda)} \in X^*_{\delta^o, \vartheta} \Rightarrow d(x, X^*_{\delta^o, \vartheta}) \leq \|x - x^{(\lambda)}\|.$$  

Since

$$x - x^{(\lambda)} = \lambda (x - x^*_{-\eta_*}) \text{ and } x^{(\lambda)} - x^*_{-\eta_*} = (1 - \lambda)(x - x^*_{-\eta_*}),$$

$$\|x - x^{(\lambda)}\| = \frac{\lambda}{1 - \lambda} \|x^{(\lambda)} - x^*_{-\eta_*}\| \leq \frac{r}{\delta_2 + \delta^o} \text{diam}(X^*_{\delta^o, \vartheta})$$

because both $x^*_{-\eta_*}$ and $x^{(\lambda)}$ belong to $X^*_{\delta^o, \vartheta}$. Noting that $\delta_2 = \eta_*$, gives the result.

Proof of Lemma 4, part 2. For $\delta = 0$ the claim is trivial. Suppose $\delta \neq 0$. Take an arbitrary $x \in X^*_{\delta, \vartheta}$; in particular, $f_0(x) \leq f^* + \vartheta$ and $f_i(x) \leq \delta$ for all $i \in \mathcal{I}$. Apply the claim with $\delta^o = 0$, $r = \delta$ and $\delta_2 = \delta$. In this case, $\lambda = 1/2$ and therefore,

$$\frac{x + x^*_{-\delta}}{2} = x^{(\lambda)} \in X^*_{\delta, \vartheta} \Rightarrow x = 2x^{(\lambda)} - x^*_{-\delta} \in 2X^*_{\delta, \vartheta} - x^*_{-\delta}.$$  

Proof of Lemma 4, part 3. We combine the previous item with the following simple facts. The first is that $\gamma_2^{(\alpha)}$ and diam are invariant under translations. Moreover, $\gamma_2^{(\alpha)}(\lambda S) = \lambda^2 \gamma_2^{(\alpha)}(S)$ and diam$(\lambda S) = \lambda$ diam$(S)$ for all $S \subset \mathbb{R}^d$ and $\lambda \geq 0$. \hfill \Box

10 Proofs of main results

We combine here the tools from the previous three sections to prove Theorems 2 (in Sect. 10.1) and 3 (in Sect. 10.2).

10.1 General sets and functions

Proof of Theorem 2 The strategy of the proof is as follows. We will use Lemma 2 to show that the event GoodThm$_2(t, \epsilon_0)$ contains the intersection of events $E_1$ and $E_2$ below. We then lower bound $\mathbb{P}(E_1)$ and $\mathbb{P}(E_2)$ to finish the proof.

The two events are defined as follows.

$$E_1 := \bigcap_{i \in \mathcal{I}_0} \left\{ \left| \hat{\Delta}_i(x^*) \right| \leq \sigma_* \sqrt{\frac{6(1 + \log(2|\mathcal{I}| + 2) + t)}{N}} \right\}; \quad (46)$$

$$E_2 := \bigcap_{i \in \mathcal{I}_0} \left\{ \sup_{x \in Y} |\hat{\Delta}_i(x; x^*)| \leq \hat{\tau}_N(t) \right\}. \quad (47)$$
First part: containment. Recalling the definitions of \( \hat{\Delta}_i \) in Sect. 8, we see that 
\[
\hat{\Delta}_i(y) = \hat{\Delta}_i(y; x^*) + \hat{\Delta}_i(x^*)
\]
for all \( y \in X \), so:
\[
\sup_{y \in Y, i \in I_0} |\hat{\Delta}_i(y)| \leq \sup_{y \in Y, i \in I_0} |\hat{\Delta}_i(y; x^*)| + |\hat{\Delta}_i(x^*)|.
\]
In particular,
\[
\text{if } E_1 \cap E_2 \text{ holds, } \sup_{y \in Y, i \in I_0} |\hat{\Delta}_i(y)| \leq \sigma^* \sqrt{\frac{6(1 + \log(2|I| + 2) + t}{N}} + r_N(t).
\]
So the assumptions of Lemma 2 are satisfied with 
\[
\delta := \sigma^* \sqrt{\frac{6(1 + \log(2|I| + 2) + t}{N}} + r_N(t) \text{ and } \epsilon := r_N(t).
\]
Applying the Lemma and inspecting the definitions shows that Good Thm. 2 \((t, \epsilon_0)\) holds. Indeed, We deduce that 
\[
E_1 \cap E_2 \subset \text{GoodThm} \cdot \text{Second part: probability bounds To finish, we must prove that } \mathbb{P}(E_1 \cap E_2) \geq 1 - e^{-t} - (2|I| + 1)\rho. \text{ Note that:}
\]
\[
1 - \mathbb{P}(E_1 \cap E_2) \leq \sum_{i \in I_0} \mathbb{P} \left\{ |\hat{\Delta}_i(x^*)| > \sigma^* \sqrt{\frac{6(1 + \log(2|I| + 2) + t}{N}} \right\} + \sum_{i \in I_0} \mathbb{P} \left\{ \sup_{x \in Y} |\hat{\Delta}_i(x; x^*)| > r_N(t) \right\}
\]
Therefore, it suffices to bound each term in (48) and (49) separately. For the terms in (48), we apply Lemma 1 with 
\[
v^2 = \sigma^2_* \text{ and } Z_k := \pm F_i(x^*, \xi_k),
\]
so that:
\[
\frac{1}{N} \sum_{k=1}^{N} (Z_k - \mathbb{E}[Z_1]) = \pm (\hat{\mathbb{E}} - \mathbb{E}) F_i(x^*, \cdot).
\]
Because of Assumption 1, we know that:
\[
\mathbb{P} \left\{ \frac{1}{N} \sum_{k=1}^{N} (Z_k - \mathbb{E}[Z_1])^2 \geq 2\sigma^2_* \right\} \leq \rho.
\]
Therefore Lemma 1 gives:
\[
\mathbb{P} \left\{ |(\hat{\mathbb{E}} - \mathbb{E}) F_i(x^*, \cdot)| > \sigma_* \sqrt{\frac{6(1 + \log(2|I| + 2) + t}{N}} \right\} \leq \frac{e^{-t}}{2(|I| + 2)} + \rho.
\]
To bound the terms in (49) we fix an \( i \in \mathcal{I}_0 \) and apply our concentration result, Theorem 4. In the language of that theorem, we have

\[
\hat{\Delta}_i(x; x^*) = (\hat{E} - E) (G(x, \cdot) - G(x^*, \cdot)) \quad \text{for } G := F_i.
\]

With these choices,

\[
\Delta G = \sup_{x \in Y} |\hat{\Delta}_i(x; x^*)|.
\]

Assumption 1 guarantees that:

\[
P \left\{ \hat{E} L_i(\cdot) > 2\sigma^2 \right\} \leq \rho.
\]

So Theorem 4 is applicable with \( \nu^2 = \sigma^2, L = L_i \). Checking the formula for \( \hat{r}_N(t) \), we may now use Theorem 4 to deduce:

\[
P \left\{ \sup_{x \in Y} |\hat{\Delta}_i(x; x^*)| > \hat{r}_N(t) \right\} \leq \frac{e^{-t}}{2(|\mathcal{I}| + 2)} + \rho.
\]

We have now bounded all the terms in the sums (48) and (49). Plugging the bounds back into these equations give the desired lower bound on \( P(E_1 \cap E_2) \).

\[ \square \]

10.2 Convex sets and functions

**Proof of Theorem 3** For convenience, we only consider the case where \( \mathcal{I} \neq \emptyset \), as the other case is simpler.

Our general proof strategy is similar to the one of Theorem 2. In the first step of the proof, we define decreasing sequences of events \( E_{1,k}, E_{2,k} \) and argue that \( \text{GoodThm}_3(t, \epsilon_0) \) contains \( \bigcap_k (E_{1,k} \cap E_{2,k}) \). We then bound the probability of the good event via bounds on \( P(E_{1,k}) \) and \( P(E_{2,k}) \).

Let us first define the events. Looking at the definition of \( \hat{\delta}(t; \epsilon_0) \), we see that one can find a decreasing sequence \( \{\epsilon_k\}_{k=1}^{+\infty} \).

\[ \forall k \geq 1 : \epsilon_k \in R_{N,\eta^*}(t; \epsilon_0) \text{ and moreover } \epsilon_k \downarrow \hat{\delta}(t; \epsilon_0). \]

For each \( k \), we have:

\[ 2\hat{w}(t; \epsilon_k) + \text{gap}(\hat{\delta}(t; \epsilon_k)) + \epsilon_0 < \epsilon_k. \]

Now, \( \hat{w}(t; \epsilon_k) = \hat{w}_N(t; \delta(t; \epsilon_k); \epsilon_k) \) where \( \delta(t; \epsilon_k) = \inf S_{N,\eta^*}(t; \epsilon_k) \). Given our assumptions, (16) and Lemma 4 above, it is easy to check that

\[ \delta \mapsto \hat{w}_N(t; \delta; \epsilon_k) \text{ and } \delta \mapsto \text{gap}(\delta). \]
are continuous nonincreasing functions of $\delta \in [0, \eta_*]$. Moreover, the sets $S_{N, \eta_0}(t; \epsilon_k)$ increase with $k$, so $\tilde{\delta}(t; \epsilon_k)$ decreases with $k$. Therefore, one can find a decreasing sequence $\{\delta_k\}_{k=1}^{+\infty}$ such that:

$$\forall k \geq 1 : \delta_k \in S_{N, \eta_0}(t; \epsilon_k),$$

$$2\tilde{w}_N(t; \delta_k; \epsilon_k) + \text{gap}(\delta_k) + \epsilon_0 < \epsilon_k,$$

and $\delta_k < \tilde{\delta}(t; \epsilon_k) + k^{-1}$. (50)

It follows in particular, that

$$\lim_{k} \delta_k = \lim_{k} \tilde{\delta}(t; \epsilon_k) = \lim_{\epsilon \searrow \tilde{w}_N(t; \epsilon_0)} \tilde{\delta}(t; \epsilon) = \tilde{\delta}(t).$$

The events we define are:

$$E_1 := \bigcap_{i \in I_0} \left\{ \left| \hat{\Delta}_i(x^*) \right| \leq \sigma_* \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}} \right\};$$

$$E_{2,k} := \bigcap_{i \in I_0} \left\{ \sup_{x \in X_{\delta_k}^{*, \epsilon_k + \text{gap}(\delta_k)}} \left| \hat{\Delta}_i(x; x^*) \right| \leq \tilde{w}_N(t; \delta_k; \epsilon_k) \right\}. \quad (52)$$

The fact that $\{\epsilon_k\}_k$ and $\{\delta_k\}$ are both decreasing implies that the events $E_{2,k}$ are decreasing.

First part: containment. We will argue that $\cap_k (E_1 \cap E_{2,k}) \subset \text{GoodThm.3}(t, \epsilon_0)$. To show this, we assume that the event $\cap_k (E_1 \cap E_{2,k})$ holds, and deduce that GoodThm.3($t, \epsilon_0$) must hold as well.

Fix an index $k$. We may assume that there exists $x_{-\delta_k}^*$ minimizing $f$ over $X_{-\delta_k}$ and:

$$f(x_{-\delta_k}^*) - f^* = \text{gap}(\delta_k).$$

In particular, $x_{-\delta_k}^* \in X_{\delta_k}^{*, \epsilon_k}$. Because $E_1 \cap E_{2,k}$ holds, we have that for all $i \in I_0$ and $x \in X_{\delta_k}^{*, \epsilon_k + \text{gap}(\delta_k)}$:

$$|\hat{\Delta}_i(x)| \leq |\hat{\Delta}_i(x; x^*)| + |\hat{\Delta}_i(x^*)|$$

$$\leq \tilde{w}_N(t; \delta_k; \epsilon_k) + \sigma_* \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}}$$

(use $\delta_k \in S_{N, \eta_0}(t; \epsilon_k) < \delta_k$).

In particular, $\hat{\Delta}_i(x_{-\delta_k}^*) < \delta_k$. For the same $x$’s, we also have:

$$|\hat{\Delta}_0(x; x_{-\delta_k}^*)| \leq |\hat{\Delta}_0(x_{-\delta_k}^*; x^*)| + |\hat{\Delta}_0(x; x^*)|$$

$$\leq 2\tilde{w}_N(t; \delta_k; \epsilon_k)$$

(use (50)) $< \epsilon_k - \epsilon_0 - \text{gap}(\delta_k)$. 

\[ Springer \]
We now apply Proposition 5 with \( \delta^o = \delta = \delta_k \), \( x_{-\delta^o} := x_{-\delta_k}^{*} \) and \( \epsilon := \epsilon_k \). The above calculations imply that the three conditions of such lemma, given by (41), (42) and (43), are satisfied. We conclude that

\[
\hat{X}^{*,\epsilon_0} \subset X_{\delta_k}^{*,\epsilon_k}
\]

and (by the same estimates)

\[
|\hat{F}^* - f^*| \leq |\hat{\Delta}_0(x^*)| + \sup_{x \in X_{\delta_k}^{*,\epsilon_k}} |\hat{\Delta}_0(x; x^*)| + \max\{\epsilon_k, \text{gap}(\delta_k)\}
\]

\((E_1 \text{ holds}) \leq \sigma^* \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}} + \sup_{x \in X_{\delta_k}^{*,\epsilon_k}} |\hat{\Delta}_0(x; x^*)| + \max\{\epsilon_k, \text{gap}(\delta_k)\} \)

\((E_{2,k} \text{ occurs}) \leq \sigma^* \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}} + \bar{w}_N(t; \delta_k, \epsilon_k) + \max\{\epsilon_k, \text{gap}(\delta_k)\} \)

\((\text{use (50)}) \leq \sigma^* \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}} + \epsilon_k^2 + \max\{\epsilon_k, \text{gap}(\delta_k)\} \).

Both (53) and (54) hold for all \( k \). Letting \( k \to +\infty \) and recalling that \( \delta_k \searrow \tilde{\delta}(t) \) and \( \epsilon_k \searrow \tilde{\rho}(t; \epsilon_0) \), we obtain:

\[
\hat{X}^{*,\epsilon_0} \subset X_{\tilde{\delta}(t)}^{*,\tilde{\rho}(t; \epsilon_0)},
\]

\[
|\hat{F}^* - f^*| \leq \sigma^* \sqrt{\frac{6(1 + \log(2|I| + 2) + t)}{N}} + \frac{\tilde{\rho}(t; \epsilon_0)}{2} + \max\{\tilde{\rho}(t; \epsilon_0), \text{gap}(\tilde{\delta}(t))\}.
\]

Going back to the statement of Theorem 3 (page 19), we see that the two properties above correspond to (a) and (b) in the definition of GoodThm.3 \( (t, \epsilon_0) \). The remaining property (c) that defines that event also holds due to Lemma 4. Therefore, by assuming that \( E_1 \cap E_{2,k} \) occurs for all \( k \), we have deduced that GoodThm.3 \( (t, \epsilon_0) \) also holds. Second step: probability bounds Recall that the events \( E_{2,k} \) are decreasing. By the first step,

\[
\mathbb{P}(\text{GoodThm.3} \ (t, \epsilon_0)) \geq \mathbb{P} \left( E_1 \cap \bigcap_{k=1}^{+\infty} E_{2,k} \right) = \lim_{k \to +\infty} \mathbb{P}(E_1 \cap E_{2,k}).
\]

Therefore, all that remains to show is that:

\[
\text{Goal} : \forall k \geq 1 : 1 - \mathbb{P}(E_1 \cap E_{2,k}) \leq e^{-t} + 2(|I| + 1)\rho.
\]

From this point on, the proof resembles the second step in the proof of Theorem 2, and we will be a bit briefer. Following (48) and (49), but with the definition of \( E_{2,k} \) in
(52), we obtain
\[
P(E_i \cup E_{2,k}) \leq \sum_{i \in \mathcal{I}_0} P \left\{ |\hat{\Delta}_i(x^*)| > \sigma^* \sqrt{\frac{6(1 + \log(2|\mathcal{I}| + 2) + t}{N}} \right\} + \sum_{i \in \mathcal{I}_0} P \left\{ \sup_{x \in X_{\epsilon_k + \text{gap}(-\delta_k)}^*} |\hat{\Delta}_i(x; x^*)| > \hat{w}_N(t; \delta_k; \epsilon_k) \right\}.
\]

As in the proof of Theorem 2, Lemma 1 gives:
\[
\forall i \in \mathcal{I}_0 : P \left\{ |\hat{\Delta}_i(x^*)| > \sigma^* \sqrt{\frac{6(1 + \log(2|\mathcal{I}| + 2) + t)}{N}} \right\} \leq \frac{e^{-t}}{2(|\mathcal{I}| + 1)} + \rho.
\]

On the other hand, the bound
\[
\forall i \in \mathcal{I}_0 : P \left\{ \sup_{x \in X_{\epsilon_k + \text{gap}(-\delta_k)}^*} |\hat{\Delta}_i(x; x^*)| > \hat{w}_N(t; \delta_k; \epsilon_k) \right\} \leq \frac{e^{-t}}{2(|\mathcal{I}| + 1)} + \rho
\]
follows from applying Theorem 4 as in the proof of Theorem 2, noting that this time we have Assumption 1 over \(Z = X_{\eta^*}^{*, \epsilon_k + \text{gap}(-\delta_k)}\), and also that
\[
\gamma_2(\alpha)(X_{\delta_k}^{*, \epsilon_k + \text{gap}(-\delta_k)}) \leq 2\gamma_2(\alpha)(X_{\delta_k}^{*, \epsilon_k + \text{gap}(-\delta_k)})
\]
and
\[
\text{diam}(X_{\epsilon_k + \text{gap}(-\delta_k)}^*) \leq 2\text{diam}(X_{\epsilon_k + \text{gap}(-\delta_k)})
\]
by Lemma 4.

\[
\square
\]

**Appendix**

**Proof of Lemma 1** The second statement in the Lemma is a direct consequence of the first. Therefore, we will only prove the first statement.

Assume that \(Z_1', \ldots, Z_n'\) are independent copies of the \(Z_1, \ldots, Z_n\). Also let \(Z = (Z_1, \ldots, Z_n)^T\). What we want to prove is that, for any \(t \geq 0\),
\[
\text{Want} : \mathbb{P} \left\{ \mathbb{E} \left[ \sum_{k=1}^{N} (Z_k - Z_k') \mid Z \right] \geq \sqrt{2(1 + t) \sum_{k=1}^{N} \mathbb{E}[(Z_k - Z_k')^2 \mid Z]} \right\} \leq e^{-t}.
\]
By [31, Corollary 1], it suffices to prove that, for any \( t \geq 0 \),

\[
\textbf{Sufficient : } \mathbb{P}\left\{ \sum_{k=1}^{N} (Z_k - Z'_k) \geq \sqrt{2t \sum_{k=1}^{N} (Z_k - Z'_k)^2} \right\} \leq e^{-t}.
\]

We will prove that the above inequality holds almost surely conditionally on values \( |Z_k - Z'_k| = a_k, 1 \leq k \leq N \). Notice that, conditionally on these values,

\[
Z_k - Z'_k = u_k a_k
\]

where the \( u_k \) are i.i.d. unbiased random signs. So what we must show is that:

\[
\forall t \geq 0 : \mathbb{P}\left\{ \sum_{k=1}^{N} u_k a_i \geq \sqrt{2t \sum_{k=1}^{N} a_k^2} \right\} \leq e^{-t},
\]

for any choice of \( a_k, 1 \leq k \leq N \). This follows easily from the standard inequalities:

\[
\forall \theta > 0 : \mathbb{E}[e^{\theta \sum_{k=1}^{N} u_i a_i}] = \prod_{k=1}^{N} \cosh(\theta a_k) \leq e^{\frac{\theta^2 \sum_{k=1}^{N} a_k^2}{2}},
\]

and Bernstein’s trick:

\[
\mathbb{P}\left\{ \sum_{k=1}^{N} u_i a_i \geq \sqrt{2t \sum_{k=1}^{N} a_k^2} \right\} \leq \inf_{\theta > 0} \mathbb{E}[e^{\theta \sum_{k=1}^{N} u_i a_i}]e^{-\theta \sqrt{2t \sum_{k=1}^{N} a_k^2}} \leq e^{-t}.
\]

\( \square \)

**Proof of Proposition 1** We will need the following Lemma.

**Lemma 5** There exists a constant \( c_{\text{bdg}} \) such that, for all \( p \geq 2 \) and all i.i.d. random variables \( Z_1, \ldots, Z_N \in L^p \) with \( \mathbb{E}[Z_i] = 0 \),

\[
\left\| \frac{Z_1 + \cdots + Z_N}{N} \right\|_p \leq c_{\text{bdg}} \sqrt{\frac{p}{N}} \|Z_1\|_p.
\]

**Proof of the Lemma** By the Burkholder–Davis–Gundy inequality and the subaditivity of the \( L^{p/2} \) norm:

\[
\|Z_1 + \cdots + Z_N\|_p \leq c_{\text{bdg}} \sqrt{p} \left\| Z_1^2 + \cdots + Z_N^2 \right\|_{p/2}^{1/2} \leq c_{\text{bdg}} \sqrt{p \sum_{i=1}^{N} \|Z_i^2\|_{p/2}}
\]

and the proof finishes when we note \( \|Z_i^2\|_{p/2} = \|Z_i\|_p^2 \) for each index \( i \). \( \square \)
Now note that the random variables
\[ H_k := \frac{h(\xi_k) - \mathbb{E}h(\cdot)}{\sigma^2} \quad (1 \leq k \leq N) \]
are i.i.d. and satisfy \( \mathbb{E}[H_k] = 0, \|H_k\|_p \leq \kappa_p. \) Markov’s inequality implies:
\[
\mathbb{P}\left\{ \frac{\hat{\mathbb{E}}h(\cdot)}{\sigma^2} > 2 \mathbb{E}^2 \right\} \leq \mathbb{P}\left\{ \frac{1}{N} \sum_{k=1}^{N} H_k > 1 \right\} \leq \frac{\|1/N \sum_{k=1}^{N} H_k\|_p}{\kappa_p}.
\]
Now use Lemma 5 to bound the RHS. \( \square \)

References

1. Artstein, Z., Wets, R.J.B.: Consistency of minimizers and the SLLN for stochastic programs. J. Convex Anal. 2, 1–17 (1995)
2. Atlason, J., Epelman, M.A., Henderson, S.G.: Call center staffing with simulation and cutting plane methods. Ann. Oper. Res. 127(1), 333–358 (2004)
3. Banholzer, D., Fliege, J., Werner, R.: On rates of convergence for sample average approximations in the almost sure sense and in mean. Math. Program. (2019). https://doi.org/10.1007/s10107-019-01400-4
4. Bartlett, P., Bousquet, O., Mendelson, S.: Local Rademacher complexities. Ann. Stat. 33, 1497–1537 (2005)
5. Bartlett, P., Mendelson, S.: Empirical minimization. Probab. Theory Relat. Fields 135(3), 311–334 (2006)
6. Barlett, P.L., Mendelson, S., Neeman, J.: \ell_1-regularized linear regression: persistence and oracle inequalities. Probab. Theory Relat. Fields 154, 193–224 (2012)
7. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Rev. 38(3), 367–426 (1996)
8. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of the Lasso and Dantzig selector. Ann. Stat. 37(4), 1705–1732 (2009)
9. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford (2013)
10. Burke, J.V., Deng, S.: Weak sharp minima revisited, part II: application to linear regularity and error bounds. Math. Program. 104, 235–261 (2005)
11. Burke, J.V., Ferris, M.C.: Weak sharp minima in mathematical programming. SIAM J. Control. Optim. 31, 1340–1359 (1993)
12. Dirksen, S.: Tail bounds via generic chaining. Electron. J. Probab. 20, 1–29 (2015)
13. Duvalcová, J., Wets, R.J.B.: Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. Ann. Stat. 16(4), 1517–1549 (1988)
14. Ermoliev, Y.M., Norkin, V.I.: Sample average approximation for compound stochastic optimization problems. SIAM J. Optim. 23(4), 2231–2263 (2013)
15. Guigues, V., Juditsky, A., Nemirovski, A.: Non-asymptotic confidence bounds for the optimal value of a stochastic program. Optim. Methods Softw. 32(5), 1033–1058 (2017)
16. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. J. Res. Natl. Bur. Stand. 49, 263–265 (1952)
17. Homem-de-Mello, T., Bayraksan, G.: Monte Carlo sampling-based methods for stochastic optimization. Surv. Oper. Res. Manag. Sci. 19, 56–85 (2014)
18. Iusem, A., Jofré, A., Thompson, P.: Incremental constraint projection methods for monotone stochastic variational inequalities. Math. Oper. Res. 44(1), 236–263 (2018)
19. Kanková, V., Houda, M.: Thin and heavy tails in stochastic programming. Kybernetika 51(3), 433–456 (2015)
20. Kanková, V., Omelchenko, V.: Empirical estimates in stochastic programs with probability and second order stochastic dominance constraints. Acta Math. Univ. Comenianae LXXXIV 2, 267–281 (2015)
21. Koltchinskii, V., Panchenko, D.: Complexities of convex combinations and bounding the generalization error in classification. Ann. Stat. 33, 1455–1496 (2005)
22. Koltchinskii, V.: Local Rademacher complexities and oracle inequalities in risk minimization. Ann. Stat. 34(6), 2593–2656 (2006)
23. Kim, S., Pasupathy, R., Henderson, S.G.: A guide to sample average approximation. In: Michael, Fu. (ed.) Handbook of Simulation Optimization, International Series in Operations Research & Management Science, vol. 216, pp. 207–243. Springer, New York (2015)
24. King, A.J., Rockafellar, R.T.: Asymptotic theory for solutions in statistical estimation and stochastic programming. Math. Oper. Res. 18, 148–162 (1993)
25. King, A.J., Wets, R.J.B.: Epi-consistency of convex stochastic programs. Stoch. Stoch. Rep. 34, 267–281 (1991)
26. Kleywegt, A.J., Shapiro, A., Homem-de-Mello, T.: The sample average approximation method for stochastic discrete optimization. SIAM J. Optim. 12(2), 479–502 (2001)
27. Lojasiewicz, M.S.: Sur le problème de la division. Stud. Math. 18, 87–136 (1959)
28. Mendelson, S.: Learning without concentration. J. ACM 62(3), 1–25 (2015)
29. Mendelson, S.: Local vs. global parameters—breaking the Gaussian complexity barrier. Ann. Stat. 45(5), 1835–1862 (2017)
30. Oliveira, R.I., Thompson, P.: Sample average approximation with heavier tails II: localization in stochastic convex optimization and persistence results for the lasso (2020)
31. Panchenko, D.: Symmetrization approach to concentration inequalities for empirical processes. Ann. Probab. 31, 2068–2081 (2003)
32. Pang, J.-S.: Error bounds in mathematical programming. Math. Program. Ser. B 79(1), 299–332 (1997)
33. Pflug, G.C.: Asymptotic stochastic programs. Math. Oper. Res. 20, 769–789 (1995)
34. Pflug, G.C.: Stochastic programs and statistical data. Ann. Oper. Res. 85, 59–78 (1999)
35. Pflug, G.C.: Stochastic optimization and statistical inference. In: Ruszczynski, A., Shapiro, A. (eds.) Handbooks in OR & MS, vol. 10, pp. 427–482. Elsevier, Amsterdam (2003)
36. Robinson, S.M.: An application of error bounds for convex programming in a linear space. SIAM J. Control 13, 271–273 (1975)
37. Rockafellar, R.T., Urysaev, S.: Optimization of conditional value-at-risk. J. Risk 2(3), 493–517 (2000)
38. Royset, J.O.: Optimality functions in stochastic programming. Math. Program. Ser. A 135, 293–321 (2012)
39. Römisch, W.: Stability of stochastic programming problems. In: Ruszczynski, A., Shapiro, A. (eds.) Handbooks in OR & MS, vol. 10, pp. 483–554. Elsevier, Amsterdam (2003)
40. Shapiro, A.: Asymptotic properties of statistical estimators in stochastic programming. Ann. Stat. 17, 841–858 (1989)
41. Shapiro, A.: Asymptotic analysis of stochastic programs. Ann. Oper. Res. 30, 169–186 (1991)
42. Shapiro, A.: Monte Carlo sampling methods. In: Ruszczynski, A., Shapiro, A. (eds.) Handbooks in OR & MS, vol. 10, pp. 353–425. Elsevier, Amsterdam (2003)
43. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory. MOS-SIAM Series on Optimization. SIAM, Philadelphia (2009)
44. Shapiro, A., Homem-de-Mello, T.: On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. SIAM J. Optim. 11(1), 70–86 (2000)
45. Shapiro, A., Nemirovski, A.: On the complexity of stochastic programming problems. In: Jeyakumar, V., Rubinov, A. (eds.) Continuous Optimization: Current Trends and Modern Applications, vol. 99, pp. 111–146. Springer, Boston (2005)
46. Talagrand, M.: Upper and Lower Bounds for Stochastic Processes. Springer, Berlin (2014)
47. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. 22, 28–76 (1994)
48. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288 (1996)
49. Vogel, S.: Stability results for stochastic programming problems. Optimization 19(2), 269–288 (1998)
50. Vogel, S.: Confidence Sets and Convergence of Random Functions (2008) preprint at https://www.tu-ilmenau.de/fileadmin/media/orsto/vogel/Publikationen/Vogel-Grecksch-Geb-korr-1.pdf
51. Vogel, S.: Universal confidence sets for solutions of optimization problems. SIAM J. Optim. 19(3), 1467–1488 (2008)

52. Wand, W., Ahmed, S.: Sample average approximation of expected value constrained stochastic programs. Oper. Res. Lett. 36, 515–519 (2008)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.