Random Sampling with Removal

Bernd Gärtner * Johannes Lengler * May Szedlák *

Abstract

Random sampling is a classical tool in constrained optimization. Under favorable conditions, the optimal solution subject to a small subset of randomly chosen constraints violates only a small subset of the remaining constraints. Here we study the following variant that we call random sampling with removal: suppose that after sampling the subset, we remove a fixed number of constraints from the sample, according to an arbitrary rule. Is it still true that the optimal solution of the reduced sample violates only a small subset of the constraints? The question naturally comes up in situations where the solution subject to the sampled constraints is used as an approximate solution to the original problem.

We study random sampling with removal in a generalized, completely abstract setting where we assign to each subset R of the constraints an arbitrary set $V(R)$ of constraints disjoint from R; in applications, $V(R)$ corresponds to the constraints violated by the optimal solution subject to only the constraints in R. Furthermore, our results are parametrized by the dimension δ, i.e., we assume that every set R has a subset B of size at most δ with the same set of violated constraints. This is the first time this generalized setting is studied.

In this setting, we prove matching upper and lower bounds for the expected number of constraints violated by a random sample, after the removal of k elements. For a large range of values of k, the new upper bounds improve the previously best bounds for LP-type problems, which moreover had only been known in special cases. We show that this bound on special LP-type problems can be derived in the much more general setting of violator spaces, and with very elementary proofs.

1 Introduction

On a high level, random sampling can be described as an efficient way of learning something about a problem, by first solving a subproblem of much smaller size. A classical example is the problem of finding the smallest element in a *sorted compact list* [2, Problem 11-3]. Such a list stores its elements in an array, but in arbitrary order. Additional pointers are used to link each element to the next smaller one in the list. Given a sorted compact list of size n, the smallest element can be found in expected time $O(\sqrt{n})$ as follows: sample a set of $\lfloor\sqrt{n}\rfloor$ array elements at random. Starting from their minimum, follow the predecessor pointers to the global minimum. The key fact is that the expected number of pointers to be followed is bounded by $\sqrt{n}$, and this yields the expected runtime.

On an abstract level, the situation can be modeled as follows. Let H be a set of size n that we can think of as the set of constraints in an optimization problem, for example the elements in a sorted compact list. Let $V : 2^H \to 2^H$ be a function that assigns to each subset $R \subseteq H$ of constraints a set $V(R) \subseteq H \setminus R$. We can think of $V(R)$ as the set of constraints violated by the optimal solution subject to only the constraints in R. In the sorted compact list example, $V(R)$ is the set of elements that are smaller than the minimum of R.

In this setting, the above "key fact" is a concrete answer to the following abstract question: Suppose that we sample a set $R \subseteq H$ of size $r \leq n$ uniformly at random. What can we say about the quantity v_r, the expected size of $V(R)$? What are conditions on V under which v_r is small?

The main workhorse in this context is the *Sampling Lemma* [6]. It states that $v_r = \frac{n-r}{r+1} \cdot x_{r+1}$, where x_r is the expected size of $X(R) = \{h \in R : h \in V(R\backslash\{h\})\}$. In other words, $h \in X(R)$ is a constraint that is not automatically satisfied if the problem is solved without enforcing it. In the sorted compact list example, every nonempty set R has one such "extreme" constraint, namely its minimum. Consequently, we have $x_{r+1} = 1$, and hence $v_r = (n-r)/(r+1)$. With $r = \lfloor\sqrt{n}\rfloor$, $v_r < \sqrt{n}$ follows. The Sampling Lemma has many other applications in computational geometry when x_{r+1} can be bounded [6].

In this paper, we address the following more general question in the abstract setting: Suppose that we sample a set $R \subseteq H$ of size $r \leq n$ uniformly at random, but then we remove a subset $K_R \subseteq R$ of a fixed size k, according to an arbitrary but fixed rule. What can we still say about the expected size of $V(R\setminus K_R)$? If K_R is a random subset of R, the expectation is v_{r-k}, but if K_R is chosen by another (deterministic) rule, then $R \setminus K_R$ is no longer a uniformly random subset, and the Sampling Lemma does not apply.

Our work is originally motivated by *chance-*

*Department of Computer Science, Institute of Theoretical Computer Science, ETH Zürich, CH-8092 Zürich, Switzerland, {gaertner, johannes.lengler, may.szedlak}@inf.ethz.ch.
Research supported by the Swiss National Science Foundation (SNF Project 200021_150055 / 1)

constrained optimization, see [6] and the explanations and references therein, but we also believe that the question is natural and interesting in itself.

A first bound on the change of the expected number of violated constraints was given in [3] in the case where (H, V) is a *nondegenerate LP-type problem*. The results are parametrized by the dimension δ (for definition of dimension see Definition 3 below). LP-type problems have been introduced and analyzed by Matoušek, Sharir and Welzl as a combinatorial framework that encompasses linear programming and other geometric optimization problems [9, 7]. The quantitative result was that under removal of k elements, the expected number of violated constraints increases by a factor of δ^k at most, which is constant if both δ and k are constant. It was left open whether this factor can be improved for interesting sample sizes (for very specific and rather irrelevant values of δ, r, k, it was shown to be best possible).

In this paper, we improve over the results in [3] in several respects. In Theorem 6 we show that the increase factor δ^k can be replaced by $\log n + k$, which is a vast improvement for a large range of values of k. Moreover, the new bound neither requires the machinery of LP-type problems, nor nondegeneracy. It holds in the completely abstract setting considered above. In this setting, we can also show that the bound is best possible for all sample sizes of the form $r = n^\alpha, 0 < \alpha < 1$. We also show that this bound is best possible for violator spaces, in the case where $k = \Omega(\delta \log n)$. In general, for violator spaces the gap to the lower bound is $\log n$.

Hence, if anything can be gained over the new bound, additional properties of the violator function V have to be used. Indeed, for small values of k, the increase factor in [3] is better than our new bound for nondegenerate LP-type problems, and most notably, it does not depend on the problem size n. We show in Theorem 9 that the same factor can be derived under the much weaker conditions of a *nondegenerate violator space*, and with a much simpler proof, based on a "removal version" of the Sampling Lemma (Lemma 8). Furthermore the proof of [3] is given for a specific rule to remove k, whereas our proof works for any rule.

Intuitively, violator spaces are LP-type problems without objective function, and they were introduced to show that many combinatorial properties of LP-type problems and algorithms for LP-type problems do not require the objective function at all [5, 1].

In Section 3, Theorem 10 we show tight upper and lower bounds for the case $\delta = 1$, which shows that the improved bound for nondegenerate violator spaces is best possible for *all* violator spaces. For smaller (and in particular constant) k, the quest for the best bound on the increase factor remains open.

What also remains open is the role of nondegeneracy. In many geometric situations, nondegeneracy can be attained through symbolic perturbation and can therefore be assumed without loss of generality for most purposes. In the abstract setting, this is not necessarily true, as there are examples of LP-type problems for which any "combinatorial perturbation" increases the dimension [8].

2 Basics and Definitions

Throughout the paper we will work with three combinatorial concepts, the LP-type problem, the violator space and the consistent space.

Definition 1 (LP-type Problems) *An* LP-type *problem is a triple $\mathcal{P} = (H, \Omega, \omega)$ that satisfies the following. H is a finite set (the constraints), Ω a totally ordered set with a smallest element $-\infty$ and $\omega : 2^H \to \Omega$ a function that assigns an objective function value to $G \subseteq H$, such that $\omega(\emptyset) = -\infty$. For all $F \subseteq G \subseteq H$ and $h \in H$, it holds that (1) $\omega(F) \leq \omega(G)$, and (2) if $\omega(F) = \omega(G) > -\infty$, then $\omega(G \cup \{h\}) > \omega(G) \Rightarrow \omega(F \cup \{h\}) > \omega(F)$. The first condition is called* monotonicity*, the second* locality*.*

A constraint $h \in H \setminus G$ is violated by G if $\omega(G \cup \{h\}) > \omega(G)$. We denote the set of violated constraints by $V(G)$. The classic example of an LP-type problem is the problem of computing the smallest enclosing ball (SEB) of a finite set of points P in $\mathbb{R}^d$ [10]. For SEB, the violated constraints of G are exactly the points lying outside the smallest enclosing ball of G.

Intuitively a violator space is an LP-type problem without an objective function.

Definition 2 (Violator Space) *A* violator space *is a pair (H, V), $|H| = n$ finite and $V : 2^H \to 2^H$ such that for all $F \subseteq G \subseteq H$, it holds that (1) $G \cap V(G) = \emptyset$ and (2) if $G \cap V(F) = \emptyset$, then $V(G) = V(F)$. The first condition is called* consistency*, the second* locality*.*

The notion of a violator space is more general than the LP-type problem, since every LP-type problem can naturally be converted into a violator space. On the other hand, not every violator space can be converted into an LP-type problem [5].

Definition 3 *Let (H, V) be a violator space.*

1. *$B \subseteq H$ is called a* basis *in (H, V), if for all $F \subsetneq B$, $B \cap V(F) \neq \emptyset$ (or equivalently, $V(F) \neq V(B)$).*
2. *A basis of $G \subseteq H$ is a basis B in (H, V) such that $B \subseteq G$ and $V(B) = V(G)$.*
3. *The* combinatorial dimension *of (H, V), denoted $\delta := \delta(H, V)$ is defined by the size of the largest basis in (H, V).*

For SEB, a basis of G is a minimal subset of points with the same enclosing ball of G. In particular, all points of the basis are on the ball's boundary. In d-dimensional space, the combinatorial dimension of any SEB-instance is at most $d+1$, since any enclosing ball can be defined by at most $d+1$ points on its boundary. However, a basis can be smaller than the combinatorial dimension, and a point set can have more than one basis: in $\mathbb{R}^2$ the set of four corners of a square has two bases, the two pairs of diagonally opposite points.

The set of *extreme constraints* $X(G) \subseteq G$ is defined by $r \in X(G) \Leftrightarrow r \in V(G \setminus \{r\})$.

In the SEB case, h is extreme in G if its removal allows for a smaller enclosing ball. Therefore h is necessarily on the boundary of the smallest enclosing ball, but this is not sufficient. For the case $\mathbb{R}^2$, if G consists of the four points on a circle, then G has no extreme point.

It is not hard to see that $X(G)$ is the intersection of all bases of G, hence $|X(G)| \leq \delta$. To bound the expected number of violators, the following result from [6] is known.

Lemma 4 *[Sampling Lemma] Let (H, V) be a violator space with combinatorial dimension δ. Let $R \subseteq H$ a u.a.r. set of size r, $v_r = \mathbb{E}[|V(R)|]$ and $x_r = \mathbb{E}[|X(R)|]$. Then $v_r = \frac{n-r}{r+1} \cdot x_{r+1} \leq \frac{n-r}{r+1} \cdot \delta$.*

The Sampling Lemma can be used to argue that v_r is small if the expected number x_{r+1} of extreme constraints of a random sample of size $r+1$ is small.

Hence in the SEB case every set has at most $d+1$ extreme points and therefore $v_r \leq \frac{n-r}{r+1} \cdot (d+1)$. If $d = 2$, then the smallest enclosing ball of a random sample of size $\sqrt{n}$ has in expectation at most $3\sqrt{n}$ points outside.

A violator space (H, V) is called *nondegenerate* if every set $G \subseteq H$ has a unique basis. Note that SEB it not nondegenerate, since as mentioned in $\mathbb{R}^2$, four points on a circle have two bases.

A consistent space is a violator space without the locality condition.

Definition 5 (Consistent Spaces) *A consistent space is a pair (H, V), $|H| = n$ finite and V a function $2^H \to 2^H$ such that for all $G \subseteq H$ it holds that $G \cap V(G) = \emptyset$.*

The basis, combinatorial dimension and extreme constraints of a consistent space can be defined equivalently as in the violator space.

In consistent spaces the first equality $v_r = \frac{n-r}{r+1} \cdot x_{r+1}$ of the Sampling Lemma 4 still holds. However, in general it does not hold that $|X(R)| \leq \delta$ for all $R \subseteq H$. One can construct examples where $X(R) = R$ [4].

3 Results

As already introduced in [3] for LP-type problems, we are interested in sampling with removal. We define the concept here for the most general case of consistent spaces. All results will then naturally extend for violator spaces and LP-type problems. Suppose we sample uniformly at random $R \subseteq H$ of size r. By some fixed rule P_k, we remove $k < r$ elements of R and obtain a set R_{P_k} of size $r - k$. We define $V_{P_k}(R) := V(R_{P_k})$. Note that in general (H, V_{P_k}) is not a consistent space. We are interested in $\mathbb{E}[|V_{P_k}(R)|]$, for which we will give several bounds. In Theorem 6 we give a tight bound for consistent spaces. In Theorem 9 we give a tight bound for nondegenerate violator spaces, which is an improvement to the result given in [3]. It depends on the values of δ and k whether the bound of Theorem 6 or Theorem 9 is stronger. Finally, in Theorem 10 we give a tight bound for violator spaces for the case where $\delta = 1$.

Tight Bounds on Consistent Spaces. The following result is proven by counting, the main argument is, that very few sets can have a large set of violators, i.e., $Pr[|V_{P_k}(R)| \geq x] \leq n^{-1}$ for x and n as defined below. For a full version of the proof see [4, Theorem 10].

Theorem 6 *Let (H, V), with $|H| = n$, a consistent space, δ, k, P_k and R with $|R| = r \leq n$ as above.*

$$\mathbb{E}[|V_{P_k}(R)|] \leq c \cdot \max\left\{\frac{n}{r}\delta \log n, \frac{n}{r}k\right\} =: x$$

where c is some suitable constant (e.g. $c = 33$).

For $\delta \log n = \Omega(k)$ tightness of the bound can be shown by choosing for every set of size at most δ, a set of violators of size $\Theta(\frac{n}{r}\delta \log n)$ independently and u.a.r. [4, Lemma 15]. For $\delta \log n = o(k)$ the bound is even tight for violator spaces [4, Lemma 17].

Extreme Constraints after Removal. Let (H, V) be a violator space of combinatorial dimension δ. In particular, every set has at most δ extreme constraints. For a given natural number k, we want to understand the following quantity:

$$\Delta_k(H, V) := \max_{R \subseteq X} |\{X(R \setminus K) : K \subseteq R, |K| = k\}| .$$

In other words, how many sets of extreme constraints can we get by removing k elements from some set R?

We obviously have $\Delta_0(H, V) = 1$ for any violator space (H, V). Moreover for (H, V) *nondegenerate* we have $\Delta_1(H, V) \leq \delta + 1$. Indeed one can show that if we remove a non-extreme element x from R, we end up with the same set $X(R \setminus \{x\}) = X(R)$ of extreme elements, so only in at most δ cases, we will get a

different set. Note that this does not hold in general [4]. Continuing with the same argument the following bound follows (for a full proof see [4]).

Lemma 7 *Let (H, V) be a nondegenerate violator space. Then $\Delta_k(H, V) \leq \sum_{i=0}^{k} \delta^i$.*

Sampling Lemma after Removal. Let (H, V) be a violator space. For $R \subseteq H$ and a natural number k, we define the following two quantities: $V_k(R) = \{x \in H \setminus R : x \in V(R \setminus K)$ for some $K \subseteq R, |K| = k\}$ and $X_k(R) = \{x \in R : x \in X(R \setminus K)$ for some $K \subseteq R, |K| = k\}$. Clearly, $V(R) = V_0(R)$ and $X(R) = X_0(R)$. Furthermore, we let $v_{r,k} = \mathbb{E}[V_k(R)]$ and similarly $x_{r,k} = \mathbb{E}[X_k(R)]$.

Lemma 8 *[Sampling Lemma after Removal]*

$$v_{r,k} = \frac{n-r}{r+1} x_{r+1,k}.$$

The proof goes like the one for the "normal" Sampling Lemma 4 [6]. The main idea is to define a bipartite graph on the vertex set $\binom{X}{r} \cup \binom{X}{r+1}$, where we connect R and $R \cup \{x\}$ with an edge if and only if $x \in V_k(R)$. By counting the outgoing edges on both sides the lemma follows [4]. Again this equality holds for consistent spaces as well.

Violators after Removal. For $R \subseteq H$, let K_R be the k-element set removed by P_k, i.e., $R_{P_k} = R \setminus K_R$. Then $E[|V_{P_k}(R)|] \leq v_{r,k} + k$. This follows since $v_{r,k}$ counts the expected number of violators in $H \setminus R$ that we can possibly get by removing *any* set of k elements and the removed elements in K_R can also be in $V(R_{P_k})$.

Theorem 9 *Let (H, V) be a nondegenerate violator space, δ, k, P_k and R with $|R| = r \leq n$ as above. Then*

$$\mathbb{E}[|V_{P_k}(R)|] \leq v_{r,k} + k \leq \sum_{i=1}^{k+1} \delta^i \cdot \frac{n-r}{r+1} + k.$$

Proof. By Lemma 8 it suffices to show that $|X_k(R)| \leq \sum_{i=1}^{k+1} \delta^i$. This holds, since by Lemma 7, at most $\sum_{i=0}^{k} \delta^i$ many *sets* of extreme elements can be obtained by removing k elements from R, and each of these sets has at most δ elements. □

By [3, Section 7.2], there exists an LP-type problem and a rule P_k, such that $|X_k(R)| = \Theta(\delta^{k+1})$, for $|R| = n-1$. However, the behavior of the bound is unknown for general r.

Combinatorial Dimension 1. In the case of violator spaces it is open whether (or when) the upper bound of Theorem 6 is tight for $k < \delta \log n$. In this case, there is a gap of up to $\log n$ between upper and lower bounds [4, Lemma 17]. For $k = 0$ we know a stronger upper bound of $O(\frac{n-r}{r+1}\delta)$ by the Sampling Lemma 4.

For the case $\delta = 1$ one can show that there exists only one class of violator spaces of dimension 1 [4, Lemma 21], namely the class of the *smallest number with repetitions* violator space, which is defined as follows: Let $|H| = n$ and H a multiset of $[n]$, i.e., every element of H is in $[n]$ and there might be repetitions. For $R \subseteq H$, let $V(R) = \{x \in H \mid x < \min_{i \in R} i\}$. Finally we require that either $V(\emptyset) = H$ or $V(\emptyset) = V(i)$ for some $i \in H$. In this setting one can prove that $\mathbb{E}[|V_{P_k}(R)|] = O(\frac{n}{r}k)$ and that this bound is tight [4, Theorem 18]. The theorem below follows immediately.

Theorem 10 *Let (H, V) be a violator space with dimension $\delta = 1$. Let k, P_k and R with $|R| = r \leq n$ as above. Then $\mathbb{E}[|V_{P_k}(R)|] = O(\frac{n}{r}k)$, and this bound is tight.*

Acknowledgments. The authors are grateful to Kenneth Clarkson and Emo Welzl for sharing important insights. Furthermore we thank Luis Barba for useful discussions.

References

[1] Y. Brise and B. Gärtner. Clarkson's algorithm for violator spaces. *Comp. Geom.*, 44(2):70 – 81, 2011.

[2] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. The MIT Press, Cambridge, MA., 1990.

[3] B. Gärtner. Sampling with removal in LP-type problems. *Journal of Comp. Geom.*, 6(2):93–112, 2015.

[4] B. Gärtner, L. Lengler, and M. Szedlák. Random sampling with removal. Preprint arXiv:1512.04226.

[5] B. Gärtner, J. Matoušek, L. Rüst, and P. Škovroň. Violator spaces: Structure and algorithms. *Discrete Appl. Math.*, 156(11):2124–2141, 2008.

[6] B. Gärtner and E. Welzl. A simple sampling lemma: Analysis and applications in geometric optimization. *Discrete & Comp. Geom.*, 25(4):569–590, 2001.

[7] J. Matoušek, M. Sharir, and E. Welzl. A subexponential bound for linear programming. *Algorithmica*, 16:498–516, 1996.

[8] J. Matoušek. Removing degeneracy in LP-type problems revisited. *Discrete & Comp. Geom.*, 42(4):517–526, 2009.

[9] M. Sharir and E. Welzl. A combinatorial bound for linear programming and related problems. In *Proc. of STACS*, pages 569–579, 1992.

[10] E. Welzl. Smallest enclosing disks (balls and ellipsoids). In *Results and New Trends in Comp. Science*, pages 359–370. Springer-Verlag, 1991.