ImpatientCapsAndRuns: Approximately Optimal Algorithm Configuration from an Infinite Pool

Anonymous Author(s)
Affiliation
Address
e-mail

Abstract
Algorithm configuration procedures optimize parameters of a given algorithm to perform well over a distribution of inputs. Recent theoretical work focused on the case of selecting between a small number of alternatives. In practice, parameter spaces are often very large or infinite, and so successful heuristic procedures discard parameters “impatiently”, based on very few observations. Inspired by this idea, we introduce IMPATIENTCAPSANDRUNS, which quickly discards less promising configurations, significantly speeding up the search procedure compared to previous algorithms with theoretical guarantees, while still achieving optimal runtime up to logarithmic factors under mild assumptions. Experimental results demonstrate a practical improvement.

1 Introduction

Solvers for computationally hard problems (e.g., SAT, MIP) often expose many parameters that only affect runtime rather than solution quality. Choosing values for these parameters is seldom easy or intuitive, and different settings can lead to drastically different runtimes—days versus seconds—for a given input instance. Such parameters are exposed in the first place because they do not have known, globally optimal settings, instead typically expressing tradeoffs between different heuristic mechanisms or implicit assumptions about problem structure. In practice, solver end-users typically need to repeatedly solve similar problems: e.g., integer programs modeling airline crew scheduling problems; or SAT formulae used to formally verify a sequence of related hardware or software designs. This gives rise to the problem of algorithm configuration: finding a joint setting of parameters for a given algorithm so that it performs well on input instances drawn from a given distribution. We make no restrictions on the space of possible parameters or its structure: they may be continuous, categorical, subject to arbitrary constraints, and may contain jump discontinuities. We refer to a joint setting of all the algorithm’s parameters as a configuration to stress this generality. A common metric of performance for a configuration, and the one we consider in this work, is mean runtime: we prefer configurations that are faster, on average, on the problems we care about solving. An algorithm configuration method can sample instances from the distribution underlying an application and can run any configuration (possibly also sampled from the set possible configurations) on any sampled instance until a timeout of its choice, and the goal is to find a configuration with nearly optimal mean runtime while using the least amount of time during the search.\footnote{As usual, we treat the cumulative runtime of all the configurations tried as the total search time. One could also consider including the overhead imposed by the configuration algorithm itself. However, beyond being difficult to model, this cost is typically negligible compared to the runtime of the configurations.}

Heuristic methods for algorithm configuration such as ParamILS \cite{17,18}, GGA \cite{22}, irace \cite{11,28} and SMAC \cite{20,21} have been used with great success for more than a decade, but they do not come with any rigorous performance guarantees. More recently, algorithm configuration has also

been considered from a theoretical perspective. Kleinberg et al. \[23\] introduced a framework to analyze algorithm configuration methods theoretically, and presented the first configuration procedure, STRUCTURED PROCRAS TINATION (SP), which is guaranteed to find an approximately optimal solution with a non-trivial worst-case runtime bound. Since then algorithms with better theoretical guarantees have been developed \[24\], \[25\] \[34\]. Overall, these theoretically-motivated configuration procedures have nice properties, such as achieving near-optimal asymptotic worst-case running times. However, none of them yet achieves competitive performance on practical problem benchmarks, for two key reasons: (i) heuristic methods usually iteratively select candidate configurations that appear likely to perform well given previous samples from the configuration space (e.g., leveraging structure in the parameter space, such as smoothness or low pseudodimension \[19\], \[29\]), whereas the theoretical algorithms select configurations randomly; and (ii) heuristic methods often impatiently discard less promising configurations based on just a few runtime observations, while the theoretical algorithms are more conservative and continue evaluating them until they demonstrate, with high probability, that another configuration is better. Such early discard strategies are particularly effective when the configuration space contains one or a few configurations that drastically outperform all others. This “needle-in-a-haystack” scenario is common in practice, perhaps in part explaining the success of these heuristic methods.

In this paper we take a significant step towards theoretically grounded and practical algorithm configuration by addressing the second problem. We build on CAPSANDRUNS (CAR) \[35\], a simple and intuitive algorithm that continuously discards configurations that perform poorly relative to a global upper bound on the best achievable mean runtime. Here we introduce IMPATIENTCAPSANDRUNS (ICAR), which equips CAR with the ability to quickly discard less-promising configurations by applying an initial “precheck” mechanism that allows poorly performing configurations to be discarded quickly. Additionally, via a more careful analysis we are able speed up a key subroutine from CAR. While ICAR retains the favorable optimality and runtime guarantees of CAR under mild assumptions, it is also provably faster in needle-in-a-haystack scenarios where most configurations are considerably weaker than the best ones (these are the cases where good algorithm configuration procedures are the most useful, because identifying a good configuration is the most consequential.) Because of its precheck procedure, ICAR is able to examine more configurations than CAR, and hence finds configurations with better mean runtime. Furthermore, not wasting time on examining bad configurations, the total runtime of ICAR is significantly smaller than that of CAR and any other existing procedure with theoretical guarantees, making a step towards closing the performance gap relative to heuristic procedures.

Finally, we briefly survey some less closely related work. Gupta & Roughgarden \[13\] initiated the study of algorithm configuration from a learning-theoretic perspective. Rather than seek general purpose configuration procedures, as we do in this work, this and subsequent approaches seek to bound the number of training samples required to guarantee good generalization for specific classes of problems. Examples include combinatorial partitioning problems such as max-cut and clustering \[6\], branching strategies in tree search algorithms \[7\], and general algorithm configuration when the runtime is piecewise-constant over its parameter space \[8\]. Hyperparameter-search methods based on multi-armed bandit algorithms are also related. The main difference is that this literature focuses on settings where every configuration run costs the same amount or where there is a tradeoff between how long each configuration is run and the accuracy with which its performance is estimated \[5\], \[27\]; thus, these methods do not face questions like how many instances to consider and how to cap runs.

The rest of the paper is organized as follows. The formal model of algorithm configuration is given in Section 2. The ICAR algorithm is presented and analyzed in Section 3. Experiments on some algorithm configuration benchmarks are given in Section 4. Proofs and additional experimental results are deferred to the appendix.

2 The Model

Following Kleinberg et al. \[23\], the algorithm configuration problem is defined by a triplet \((\Pi, \Gamma, R)\), where \(\Pi\) is a distribution over possible configurations, \(\Gamma\) is a distribution over input instances, and \(R(i, j)\) is the runtime of a configuration \(i\) on a problem instance \(j\). For example \(\Pi\) and \(\Gamma\) may simply be uniform distributions, respectively over the space of hyperparameters and the set of past problem instances seen. The mean runtime of a configuration \(i\) is defined as \(R(i) = \mathbb{E}_{j \sim \Gamma} [R(i, j)]\), and the ultimate goal of an algorithm configuration method is to find a configuration \(i\) minimizing \(R(i)\).
During this search the configuration method needs to explore new configurations, which can be sampled from $\Pi$. The configuration method can also sample problem instances from $\Gamma$ and run a configuration $i$ on an instance $j$ until it finishes, or the execution time exceeds a specified timeout $\tau \geq 0$. The use of such a timeout allows for a tradeoff between learning more about the runtime of a single configuration–instance pair and considering a larger number of such pairs.

To this end, for any configuration $i$ we consider the $\tau$-capped expected runtime $R_\tau(i) = \mathbb{E}_{j \sim \Gamma} [\min\{R(i,j), \tau\}]$. Furthermore, for any $\delta \in (0, 1)$, let $t_\delta(i) = \inf \{ t : \Pr_{j \sim \Gamma} (R(i,j) > t) \leq \delta \}$ denote the $\delta$-quantile of $i$'s runtime, and define $R_\delta(i) = R_{t_\delta(i)}(i)$ the $\delta$-capped expected runtime of $i$. That is, $R_\delta(i)$ is the mean runtime of $i$ if we cap the slowest $\delta$-fraction of its runtimes.

Since a globally optimal configuration may be arbitrarily hard to find, we instead seek a solution that is competitive with the performance of the top $\gamma$-fraction of the configurations for $\gamma \in (0, 1)$. That is, instead of finding a configuration close to $\operatorname{OPT} = \min_i \{ R(i) \}$, we search for one close to $\operatorname{OPT}_\gamma = \inf \{ x : \Pr_{i \sim \Pi} (R(i) \leq x) \geq \gamma \}$. Additionally, since the average runtime of any configuration, including the optimal one, could be totally dominated by a few incredibly unlikely but arbitrarily large runtime values, we seek solutions whose expected $\delta$-capped runtime is close to the $\delta$-capped optimum. However, it turns out that this relaxed property is still impossible to verify [34].

Following Weisz et al. [34], we address this by adding a small amount of slack to the benchmark, comparing to the $(\delta/2)$-capped optimum rather than the $\delta$-capped optimum. Putting this together, we seek solutions whose expected $\delta$-capped runtime is close to the $(\delta/2)$-capped optimum, after excluding the best $\gamma$-fraction of configurations: $\operatorname{OPT}_{\delta/2}^\gamma = \inf \{ x : \Pr_{i \sim \Pi} (R_{x}(i) \leq x) \geq \gamma \}$.

**Definition 1** $(\varepsilon, \delta, \gamma)$-optimality. A configuration $i$ is $(\varepsilon, \delta, \gamma)$-optimal if $R_\delta(i) \leq (1 + \varepsilon)OPT_{\delta/2}^\gamma$.

This definition generalizes the notion of $(\varepsilon, \delta)$-optimality of Weisz et al. [35] for a finite set of configurations, where instead of the top-$\gamma$ portion, we aim to achieve the performance of the best configuration (up to $\varepsilon$): for a finite set of $N$ configurations, configuration $i$ is $(\varepsilon, \delta)$-optimal if it is $(\varepsilon, \delta, 1/N)$-optimal when $\Pi$ is the uniform distribution over the $N$ configurations.

### 3 The Algorithm

Recent theoretically-sound algorithm configuration procedures make several runtime measurements for every configuration in a finite pool $N$, and stop when they can confirm, with high probability, that one configuration is close enough to the best one. The main challenge is to avoid wasting time on (a) hard input instances with large runtimes; and (b) bad configurations that will be eliminated later. To this end, **Structured Procrastination (SP)** [23] and its improved version **Structured Procrastination with Confidence (SPC)** [24] gradually increase the runtime cap for every configuration-instance pair, while carefully determining an order to evaluate these pairs, depending on the configurations’ empirical average runtime (SP) or empirical confidence bounds on the mean runtimes (SPC). **LeapsAndBounds (LAB)** [34], which introduced empirical confidence bounds to the algorithm configuration problem, works with a much simpler schedule, and tests all configurations for a given time budget, which is increased gradually.

On the other hand, **CapsAndRuns (CAR)** [35] first measures the runtime cap for each configuration guaranteeing that at least a $(1 - \delta)$-portion of the instances can be solved within that cap, then runs a racing algorithm (based on continuously recomputing confidence bounds on the mean runtimes) to select which capped configuration is the best. During the race, all configurations are run in parallel on more and more problem instances, and their mean runtime is continuously estimated. This makes it possible to maintain a high-probability upper bound $T$ on the optimal capped runtime, and any configuration with a runtime lower bound above $T$ can be eliminated. The algorithm stops when it can prove that a configuration is $(\varepsilon, \delta)$-optimal.

To apply any of the above methods to an infinite pool of configurations, one can simply select a pool of $\left\lceil \frac{\log(\zeta)}{\log(1 - \gamma)} \right\rceil$ configurations randomly from $\Pi$ to ensure that with probability at least $1 - \zeta$ it contains a configuration that belongs to the top $\gamma$-fraction of all the configurations. Thus the above

---

2We can see $\Pi$ as reflecting beliefs about the distribution of good configurations in the parameter space. This implicitly neglects any search procedure that leverages structural assumptions about the parameter space.

3With a slight abuse of terminology, throughout we use the same expression for capping with timeouts ($\tau$) and quantiles ($\delta$), when the interpretation is clear from the context; we specify the type of capping otherwise.
methods can select $(\varepsilon, \delta, \gamma)$-optimal configurations from an infinite pool, with attractive theoretical guarantees. Our focus in this paper is on extending CAR, due to its conceptual simplicity and good practical performance. However, in contrast to LAB and SPC, which try to assign little runtime to bad configurations from the very beginning, at the start CAR spends the same amount of time testing all configurations. This is because the estimation of the runtime caps is done in parallel, so every configuration is run for an equally long time until the first cap is found for any configuration (only after this can the algorithm start eliminating configurations with large mean runtimes). As a result, CAR spends more time testing the worst configurations than LAB or SPC.

**ImpatientCapsAndRuns (ICAR)** addresses this problem, introducing a “precheck” mechanism to ensure that bad configurations are eliminated early. The Precheck function estimates the mean capped runtime (up to a constant multiplicative factor) needed by a configuration to solve at least a constant fraction of the problem instances (less than $1 - \delta/2$). If this capped runtime is large compared to the upper bound $T$ on the $(\varepsilon, \delta, \gamma)$-optimal runtime (maintained similarly as in CAR), the configuration is rejected and eliminated from further analysis. This procedure is very similar to the CAR algorithm (with some fixed, constant $\varepsilon$ and $\delta$); only the specific rejection conditions differ mildly. Note that the runtime estimated by Precheck is a lower bound to the $\delta/2$-capped runtime, ensuring that good configurations are unlikely to be rejected. The efficiency of Precheck crucially depends on the quality of the bound $T$ on the optimal runtime. Therefore, similarly to SPC, ICAR gradually introduces more and more configurations in batches $N_k, k = K - 1, \ldots, 0$:

if a configuration passes Precheck, a (rough) estimate of its capped runtime is calculated (up to a multiplicative constant, for a cap slightly larger than the $\delta$ quantile), again by first measuring the runtime cap, then estimating the mean runtime using the measured cap. This runtime estimate is then used to reduce the bound $T$, which improves the performance of Precheck for the next batch of configurations, $N_{k-1}$. The size of batch $N_k$ is of order $1/\gamma_k$ with $\gamma_k = 2^k\gamma$, ensuring that with high probability it contains an $(\varepsilon, \delta, \gamma_k)$-optimal configuration (whose mean runtime is then bounded by $\text{OPT}^{\gamma_k}_{\delta/2}$). As a consequence, after batch $N_k$, $T$ is at most $2^k\text{OPT}^{\gamma_k}_{\delta/2}$, gradually reducing towards $2^k\text{OPT}^{\gamma_k}_{\delta/2}$. Finally, the racing part of CAR is run over all surviving configurations, further reducing $T$ towards $2^k\text{OPT}^{\gamma_k}_{\delta/2}$, and stopping when an $(\varepsilon, \delta, \gamma)$-optimal configuration is found.

Now we are ready to present the main theoretical result of the paper, a performance guarantee for ICAR. The components of the algorithm are presented in Algorithms 1–5. We then discuss each and present a proof sketch for the theorem (the detailed proof is given in Appendix A).

**Theorem 1.** For input parameters $\varepsilon \in (0, 1/3), \delta \in (0, 0.2), \gamma \in (0, 1)$, integer $K \geq 1$, and failure parameter $\zeta \in (0, 1/12)$, with probability at least $1 - 12\gamma$, ImpatientCapsAndRuns finds an $(\varepsilon, \delta, \gamma)$-optimal configuration with total work $\tilde{O}(\text{OPT})$ bounded by

$$
\tilde{O}\left(\frac{\text{OPT}^{\gamma}_{\delta/2}}{\varepsilon^2 \delta \gamma} \cdot F(38 \text{OPT}^{\gamma}_{\delta/2}) + \sum_{k=0}^{K-2} \frac{\text{OPT}^{\gamma_k}_{\delta/2}}{\gamma_k} \left(1 + \frac{F(38 \text{OPT}^{\gamma_{k+1}}_{\delta/2})}{\delta} \right) + \frac{\text{OPT}^{\gamma_{K-1}}_{\delta/2}}{\delta \gamma_{K-1}}\right),
$$

where $\gamma_k = 2^k\gamma$, and $F(x) = \Pr_{i \sim \Pi}(R_0^{0.35}(i) \leq x) + 4\zeta/K$.

**Discussion.** (i) To illustrate the advantages captured by the theorem, consider a situation where configuration runtimes are distributed exponentially, with their mean distributed uniformly over an interval $[A, A + B]$. When the number of near-optimal configurations is small (i.e., $B/A$ is large enough), the bound on the fraction of configurations surviving Precheck, $F(38 \text{OPT}^{\gamma}_{\delta/2})$, roughly scales with $\gamma$, resulting in a runtime $\text{OPT}^{\gamma}_{\delta/2}/(\varepsilon^2 \delta)$, providing a $\gamma$-factor speedup over typical bounds in other work (which scale with $\text{OPT}^{\gamma}_{\delta/2}/(\varepsilon^2 \delta \gamma)$). (Details are given in Appendix B.)

(ii) The first term in the bound corresponds to the work done in the final racing part of ICAR. The other terms correspond to the work done for each batch $N_k$ (except that the cost of the last precheck is included in the $k = 0$ term).

(iii) Kleinberg et al. [23] showed that to find an $(\varepsilon, \delta)$-optimal configuration out of a pool of size $n$, the worst-case minimum total runtime is $\Omega(\frac{\text{OPT}}{n^2 \varepsilon^2})$ since we need to test $\Omega(1/\gamma)$ configurations, in

We use “total work” and “total runtime” interchangeably; both sum over all parallel threads.

We use the standard $\mathcal{O}$ and $\tilde{O}$ notation, where the latter hides poly-logarithmic factors.

Essentially this holds since we need $\Omega(\frac{1}{n \varepsilon^2})$ sample runs to estimate the $\delta$-capped runtime of a configuration with accuracy $\varepsilon$, as the maximum runtime for configuration $i$ on some instance can be as large as $R_0(i)/\delta$.
Global variables
1: Instance distribution \( \Gamma \)
2: Phase I measurements count \( b \)
3: \( T \leftarrow \infty \) \( \triangleright \) Upper bound on \( \text{OPT}_{3/2}^\gamma \), updated continuously by all parallel processes
4: Set \( \mathcal{N}' \) of algorithm configurations

Algorithm 1 IMPATIENTCAPS\&RUNS
1: Inputs: Precision parameter \( \varepsilon \in (0, \frac{1}{3}) \), Quantile parameter \( \delta \in (0, \frac{1}{3}) \), Optimality quantile target parameter \( \zeta \in (0, \frac{1}{12}) \), Failure probability parameter \( \zeta' \in (0, \frac{1}{12}) \), Number of iterations \( K \), Instance distribution \( \Gamma \), Configuration distribution \( \mathcal{N}' \)
2: \( \mathcal{N}_k \leftarrow \text{Sample } \frac{\log(\zeta/K)}{\log(1-\delta)} - \frac{\log(\zeta/K)}{\log(1-\varepsilon)} \) many configurations from \( \mathcal{N}' \) for \( k \in [0, K-1] \)
3: \( b \leftarrow \left[ \frac{2n \log(2n/\varepsilon)}{\varepsilon^3 \log(n)} \right] \)
4: Reset \( T \leftarrow \infty \)
5: \( \mathcal{N} \leftarrow \bigcup_{k=0}^{K-1} \mathcal{N}_k \)
6: for \( k = K-1 \) downto 0 do
7: \( \mathcal{N}_k \leftarrow \text{PRECHECK}(\mathcal{N}_k, \zeta/K) \)
8: for configurations \( i \in \mathcal{N}_k \) in parallel do
9: \( P_i \leftarrow \text{CAPS\&RUNS}(i, \varepsilon, \delta, \zeta) \) thread
10: Start running \( P_i \)
11: Pause \( P_i \) when \( b \) runs of \( \text{RUNTIME} \) finished
12: end for
13: end for
14: \( \mathcal{N} \leftarrow \text{PRECHECK}(\mathcal{N}, \zeta/K) \)
15: Continue running \( P_i \) for \( i \in \mathcal{N} \)
16: // \text{CAPS\&RUNS} eliminates the threads
17: Wait until all threads finish, abort if \( |\mathcal{N}| = 1 \)
18: return \( \tau' = \arg\min_{i \in \mathcal{N}} \hat{Y}(i) \) and \( \tau' \).

Algorithm 2 CAPS\&RUNS thread
1: Inputs: Configuration \( i \), precision \( \varepsilon \), quantile parameter \( \delta \), failure probability parameter \( \zeta \)
2: // Phase I:
3: Run \( \tau_i \triangleleft \text{QUANTILE} \)(i, \( \delta \))
4: // Phase II:
5: if QUANTILE(i, \( \delta \)) aborted then
6: Remove \( i \) from \( \mathcal{N} \)
7: else
8: \( \hat{Y}(i) \leftarrow \text{RUNTIME} \)(i, \( \tau_i, \varepsilon, \delta, \zeta \))
9: if RUNTIME rejected \( i \) then
10: Remove \( i \) from \( \mathcal{N} \)
11: end if
12: end if

Algorithm 3 QUANTILE
1: Inputs: \( i, \delta \)
2: Initialize: \( m \leftarrow \left( 1 - \frac{3}{4} \delta \right) b \)
3: Run configuration \( i \) on \( b \) instances, in parallel, until \( m \) of these complete. Abort and return abort if total work \( > 1.5tb \).
4: \( \tau \leftarrow \text{runtime of } m\text{-th completed instance} \)
5: return \( \tau \)

*When running \text{CAPS\&RUNS} threads in parallel, we allocate the same amount of time for every running thread, regardless of the number of parallel tasks they themselves may be performing.

Algorithm 4 PRECHECK
1: Inputs: Configuration \( \mathcal{M} \), error parameter \( \zeta/K \)
2: \( \mathcal{M}' \leftarrow \{ \} \) \( \triangleright \) empty set
3: \( b' \leftarrow \left[ 32.1 \log \left( \frac{2K}{\zeta} \right) \right] \)
4: if \( T = \infty \) then
5: return \( \mathcal{M} \)
6: end if
7: for \( i \in \mathcal{M} \) do
8: if \( T \) last set when evaluating \( i \) then
9: append \( i \) to \( \mathcal{M}' \) \( \triangleright \) Add automatically
10: Continue
11: end if
12: // Phase I:
13: Run \( i \) on \( b' \) instances in parallel until \( \left[ 0.8b' \right] \) complete. Abort if total work \( \geq 1.9tb' \).
14: if not aborted then
15: \( \tau' \leftarrow \text{runtime of } \left[ 0.8b' \right] \text{-th completed instance} \)
16: // Phase II:
17: for \( l = 1, l \leq b' \) do
18: \( Y_l \leftarrow \text{runtime of configuration } i \) on instance \( j \sim \Gamma \), with timeout \( \tau' \)
19: if \( \sum_{l=1}^{b'} Y_l > 2.997b' \) then
20: // Stop measuring if total work too large
21: Break
22: end if
23: end for
24: Sample mean \( \bar{Y} \leftarrow \frac{1}{b'} \sum_{l \in \mathcal{Y}} Y_l \)
25: Sample variance \( \bar{\sigma}^2 \leftarrow \frac{1}{b'} \sum_{l \in \mathcal{Y}} (y - \bar{Y})^2 \)
26: Confidence \( C \leftarrow \sigma \sqrt{\frac{2 \log(\frac{2n}{\delta}) + 3\zeta \log \left( \frac{2n}{\delta} \right)}{b'}} \)
27: if \( \bar{Y} - C \leq \zeta \) then
28: append \( i \) to \( \mathcal{M}' \)
29: end if
30: end if
31: end for
32: return \( \mathcal{M}' \)

Algorithm 5 RUNTIME
1: Inputs: \( i, \tau, \varepsilon, \delta, \zeta \)
2: Initialize: \( j \leftarrow 0 \)
3: while True do
4: Sample \( j\text{th} \) instance \( J \) from \( \Gamma \)
5: \( Y_{i,j} \leftarrow \text{runtime of configuration } i \) on instance \( J \), with timeout \( \tau_i \)
6: Sample mean \( \bar{Y}(i) \leftarrow \frac{1}{j} \sum_{j=1}^{j} Y_{i,j} \)
7: Sample variance \( \bar{\sigma}^2 \leftarrow \frac{1}{j} \sum_{j=1}^{j} (Y_{i,j} - \bar{Y}(i))^2 \)
8: // Calculate confidence:
9: \( C_i \leftarrow \sigma \sqrt{\frac{2 \log(\frac{3n}{\delta / j}) + 3\tau_i \log \left( \frac{2n}{\delta} \right)}{j}} \)
10: if \( \bar{Y}(i) - C_i > T \) then
11: return reject \( i \)
12: end if
13: if \( j\text{th} \) then
14: \( T \leftarrow \min(T, 2\bar{Y}(i)) \).
15: end if
16: \( T \leftarrow \min(T, \bar{Y}(i) + C_i) \) \( \triangleright \) upper confidence
17: if \( C_i \leq \frac{\bar{Y}(i)}{2} \) then
18: return accept \( i \) with runtime estimate \( \bar{Y}(i) \).
19: end if
20: \( j \leftarrow j + 1 \)
21: end while
the worst case the total runtime needed to find an $(\varepsilon, \delta, \gamma)$-optimal configuration is about $\frac{OPT^7_{\delta/2}}{\gamma^2}$. The first term in our bound matches this, except that it is multiplied by (an upper bound on) the fraction of configurations surviving PRECHECK, $F(38OPT^7_{\delta/2})$. Under typical parameter settings, this is the main term of the bound—the only one scaling with $1/(\varepsilon^2 \delta \gamma)$—and the performance improvement of ICAR over CAR comes from this additional factor of $F(38OPT^7_{\delta/2})$. Note that this term, and all the others, scale with a bound on the optimal runtime for the set of configurations they correspond to (e.g., for batch $N_k$ they scale with $OPT^7_{\delta/2}$).

(iv) $F(38OPT^7_{\delta/2})$ is an upper bound on the number of configurations surviving PRECHECK from $N_k$. Due to the a worst-case nature of our analysis, the bound is conservative, and in practice the number of surviving configurations is much smaller. In essence, this term measures how many configurations are competitive with a very good ($OPT^7_{\delta/2}$-optimal) configuration. In other words, it measures the “needle-in-a-haystack” property of the configuration task.

(v) The first term can be replaced with the problem-dependent bound of Weisz et al. [35, Equation 1] for $n = F(38OPT^7_{\delta/2})^2$ configurations. This bound depends on the characteristics of the runtime distributions of the configurations, and show that the algorithm can run much faster if the problem is easy, e.g., adapting to the relative variance of the runtime distributions. However, for simplicity, we only present the worst-case form here.

(vi) The rest of the terms represent the cost of iteratively selecting only the best configurations to evaluate. None of these terms depends on $1/\varepsilon^2$. Note $1/\gamma_k$ is roughly the number of configurations in batch $N_k$, and each configuration is run essentially as long as the best configuration in that batch ($OPT^7_{\delta/2}$). Each of these configurations is run on constantly many instances in PRECHECK, and the surviving fraction of $F(38OPT^7_{\delta/2})$ configurations is also run on $1/\delta$ instances to measure an accurate cap and set the bound $T$. These terms scale with $OPT^7_{\delta/2}$ and $T$, and are only meaningful when $2^{-k}OPT^7_{\delta/2}$ is not too large. While in principle they can be infinite, in realistic scenarios this is not the case. Nevertheless, this requires the practitioner to choose $\gamma_k$ such that it guarantees a small-enough optimal runtime $OPT^7_{\delta/2}$, which is essentially the same task as choosing a proper $\gamma$. The terms also scale with $1/\delta$, but the effect of this is mitigated by the success of PRECHECK: for $k \neq K - 1$, each term is multiplied by the upper bound $F(38OPT^7_{\delta/2})$ on the fraction of configurations surviving PRECHECK.

(vii) Our analysis shows that CAR can be sped up significantly without sacrificing any of its guarantees from Weisz et al. [35], by measuring the runtime caps on fewer samples (i.e., replacing the original value of $b$ from Weisz et al. [35] with the one in Line 3 of Algorithm 1). We call this improved algorithm CAR++. This effect is also partly responsible for the improved performance of ICAR.

Insights into the algorithm and proof sketch We start with a brief description of the CAR algorithm, which runs parallel threads of Algorithm 2 for all configurations it considers. As described before, one thread, working on configuration $i$, has two phases: In the first phase, implemented in QUANTILEEST (Algorithm 3), a runtime cap $\tau_i$ is determined such that $i$ is guaranteed, with high probability, to solve a random instance with probability between $1 - \delta$ and $1 - \delta/2$ (i.e. $t_\delta(i) \leq \tau_i < t_{\delta/2}(i)$). This is achieved by solving sufficiently many instances in parallel, and $\tau_i$ is selected to be the time when a $(1 - 3\delta/4)$-fraction of the instances are solved. If measuring this cap takes too long, then QUANTILEEST stops measuring and eliminates configuration $i$. Unless this happens, in the second phase, the method RUNTIMEEST (Algorithm 4) is used to estimate the mean $\tau_i$-capped runtime $R_{\tau_i}(i)$ of $i$, by solving successively selected random instances and computing the average runtime $Y(i)$. Then the empirical Bernstein inequality [41] is used to guarantee that $R_{\tau_i}(i) \leq \hat{Y}(i) - C_{\varepsilon}(\varepsilon) \sqrt{\hat{Y}(i) + C_{\varepsilon}(\varepsilon)}$ for $C_{\varepsilon}$ calculated in Line 9 of Algorithm 5. This confidence interval is used continuously in multiple ways: (i) to reduce a global upper bound $T$ on the best possible runtime of all the configurations (Line 10); (ii) to eliminate a configuration if it shows that $R_{\tau_i}(i) > T$ (Line 10); and (iii) to check if $R_{\tau_i}(i)$ is estimated accurately enough (Line 17). The procedure (which is an instance of a so-called Bernstein race [30]) continues until each configuration is either measured accurately or eliminated. The continuous elimination (also in QUANTILEEST) and parallel execution

---

7Almost all guarantees provided in this paper are based on random sampling and hence hold with high probability. For brevity, when it is clear from the context, we often omit the ‘high-probability’ qualifier.
guarantees that when the procedure stops, every configuration is run for at most $\tilde{O}(\text{OPT}/(\varepsilon^2 \delta))$
time, and eventually an $(\varepsilon, \delta)$-optimal configuration is found, where \text{OPT} is the minimum mean
$\delta/2$-quantile capped runtime of the configurations.

As explained before, ICAR (Algorithm 1) starts to examine new configurations in batches. For
any batch $N_k$, first each configuration is quickly tested to see if it can be excluded from the set of
potentially optimal configurations. This is done by the \textsc{Precheck} function, given in Algorithm 4.
\textsc{Precheck} is very similar to \textsc{Car}, but works with constant accuracy and quantile parameters
instead of $\varepsilon$ and $\delta$, ensuring that it runs quickly, in time independent of these parameters. Also,
the conditions to reject configurations are slightly different. For a configuration $i$, \textsc{Precheck} first
estimates a cap $\tau'$ that guarantees solving random instances with constant probability $p_i \in \{0.1, 0.35\};$
then the mean $\tau'$-capped runtime is estimated roughly up to a constant multiplicative error. Since
$\delta/2 \leq 0.1$ (the lower bound on $p_i$), \textsc{Precheck} can compute multiplicative lower bounds on the
runtime $R_{\delta/2}(i)$. These are then used to set the rejection conditions such that at least one of the
best configurations from this batch $i$ with $R_{\delta/2}(i) \leq T$ is not rejected. Combining with the fact
that $\bigcup_{j=\delta/2}^{K-1} N_j$ contains a top-$\gamma_k$ configuration, such a configuration survives \textsc{Precheck} and the
corresponding \textsc{CapsAndRuns}-thread in ICAR (Algorithm 1) ensures that $T$ is set to at most
$2\text{OPT}_{\delta/2}$ in Line 11 of Algorithm 1, that is, $T$ is continuously refined as new batches are evaluated.
The number of configurations surviving \textsc{Precheck} can be bounded by looking at mean runtimes
capped at the 0.35-quantile (upper bound on $p_i$). Together with the setting of $T$, this implies that
at most a $\tilde{O}(F(380\text{OPT}^{\delta/2}))$ fraction of the $|N_k| = 3\text{OPT}/\gamma_k$ configurations survive \textsc{Precheck}.
Considering that the number of runs carried out for each configuration is constant in \textsc{Precheck},
$|\text{OPT}/\delta|$ in the loop of Algorithm 1 and $\tilde{O}(1/(\varepsilon^2 \delta))$ in the last full \textsc{Car} procedure, since the average
runtime per configuration for $N_k$ is $\text{OPT}_{\delta/2}$ (by the analysis of \textsc{Car}), the runtime bound of the
theorem follows. Correctness (i.e., the fact that the procedure finds an $(\varepsilon, \delta, \gamma)$-optimal configuration)
follows from that of \textsc{Car} and because \textsc{Precheck} retains good configurations, as just shown.

## 4 Experiments

The basic setup and main results of our experimental analysis of ICAR are given below, while details
are presented in Appendix C along with a synthetic experiment examining ICAR’s speedup as good
configurations become increasingly rare. We compared against the best available configurators that
come with theoretical guarantees. We used the improved version of \textsc{Car} (\textsc{Car++}), derived in
this paper, which uses a smaller $b$-value than the original version, thanks to our improved analysis
(see Section 3 and Appendix A for details). Including \textsc{Car++} in the experiments allowed us to
separately examine the effects of two improvements we introduced: (i) the smaller number of samples
$b$ needed in \textsc{Car}, and (ii) the main conceptual innovation of this paper, the impatient discarding of
configurations using \textsc{Precheck}. We attempted to compare against \textsc{Spc} [24] as well. However, in
the experiments presented in Table 1 although \textsc{Spc} identified good configurations, it usually was
not able to provide the required guarantees on $\varepsilon$ and $\delta$ even after running for twice as long as the
slowest alternative considered (\textsc{Car}): \textsc{Spc} did not provide guarantees for 7 out of the 9 scenarios
while also being the slowest in the other two cases (1.56 and 1.91 times slower than \textsc{Car}). Therefore,
we decided not to include \textsc{Spc} in our further comparisons.

**Datasets.** We looked at two datasets from MIP and one from SAT. We considered true runtime
data from the minisat SAT solver on instances generated by CNFuzzDD (http://fmv.jku.at/
cnfuzzdd), which was examined in past work [34][35][24]. For the MIP scenarios, we looked at the
Cplex integer program solver on combinatorial auction instances (Regions200 [26]) and problems
from wildlife conservation (RCW [11]). To generate sufficient MIP runtime data, following Hutter
et al. [22], we used an \textit{Empirical Performance Model (EPM)}—a random forest model trained on
existing runtime data—to predict the runtime of new configurations on new instances. EPMs can
do surprisingly well at predicting individual runtimes, particularly on the MIP datasets we consider.
More importantly for our purposes, Eggensperger et al. [12] showed that such EPMs are effective
surrogates for algorithm configuration, capturing key properties of runtime distributions such as the
relative quality of configurations. We note that similar surrogates have also been used to guide search
procedures [19][9][33][37], to build algorithm portfolios [31][36], to impute missing data [10], and to
optimize hyperparameters from limited observations [32].

**Main Results.** Table 1 shows the total CPU time needed to find a $(0.05, 0.1, \gamma)$-optimal configura-
tion on each dataset with the same total failure probability $(0.05)$ and with different values of \gamma. The
We do note that state-of-the-art heuristic methods also evaluate many random configurations. A key limitation is that our work focuses simply on evaluating randomly sampled configurations. The performance improvement was largest when the γ parameters were not specifically chosen; results for varying ε and δ are reported in Appendix C. ICAR consistently outperformed CAR in all cases; ICAR outperformed CAR++ on the MIP datasets and was competitive on the SAT one. The performance improvement was largest when the PRECHECK mechanism managed to discard the most configurations; the MIP datasets have relatively more weak configurations, enabling PRECHECK to filter out more configurations quickly (see Fig. 2 in the Appendix for the distribution of configuration means). When γ is relatively small, ICAR was more likely to sample a really good configuration, making it easier to discard weak ones. In this case its runtime was as little as half that of CAR++, a significant improvement. Despite taking less total CPU time, ICAR actually sampled more configurations than CAR did. To understand this phenomenon better, Fig. 1 shows the time spent running each configuration. For all datasets the plots nearly overlap for the very best few configurations, indicating that ICAR treated these good configurations in much the same way as CAR or CAR++. However, the effect of the PRECHECK mechanism is clear, as ICAR ran many bad configurations for near-zero time, discarding them quickly. In cases where a bad configuration made it past PRECHECK (largest spikes in the blue curve), ICAR ran it for an amount of time similar to CAR++. Finally, the empirical mean δ-capped runtime (Rδ) of the returned configuration is reported in Table 1. All configurators returned solutions with similar quality, but thanks to its ability to examine more configurations, ICAR often did slightly better.

5 Conclusions

This paper presented a novel algorithm configuration method, ICAR, that selects configurations from an infinite pool with optimal theoretical guarantees up to logarithmic factors under mild conditions. While earlier theoretically grounded methods thoroughly test all configurations, ICAR—like successful heuristic approaches—quickly discards less promising ones. As a result, ICAR achieves significant speedups, particularly in needle-in-a-haystack scenarios. It thus constitutes an important step towards closing the gap between theoretical and heuristic procedures.

A key limitation is that our work focuses simply on evaluating randomly sampled configurations. We do not note that state-of-the-art heuristic methods also evaluate many random configurations to avoid getting stuck in local optima, so analyzing such procedures is of obvious practical importance. Furthermore, ICAR can be understood as a way of weighing different candidate configurations against...
each other, which could be proposed by model- or gradient-based methods as well as by random sampling (see, e.g., an argument to this effect in [23 Theorem 7.1]).

Broader Impact

We expect that our theorems will guide the design of future algorithm configuration procedures. We note that speeding up computationally expensive algorithms saves time, money, and electricity, arguably reducing carbon emissions and yielding social benefit. The algorithms we study can be applied to a limitless range of problems and so could yield both positive and negative impacts; however, we do not foresee our work particularly amplifying such impacts beyond the computational speedups already discussed.

References


A Proof of Theorem 1

The first step of the proof improves the analysis of \( \text{CAPSANDRUNS} \) given in [35]. In [35], the value of \( b \) was \( \left\lceil \frac{4 \log \left( \frac{2n}{\varepsilon} \right)}{\varepsilon} \right\rceil \), which we replace here with \( \left\lceil \frac{2n \log \left( \frac{2n}{\varepsilon} \right)}{\varepsilon} \right\rceil \). This value is used in the original analysis of \( \text{CAPSANDRUNS} \) twice, in Lemma 2 and Lemma 3 of [35]. The analysis of Lemma 2 still holds with the new value without any change, while we give a new proof for Lemma 3 of [35]: the difference is that in the new proof we use the Bernstein inequality rather than its empirical version. The new version of the lemma, Lemma 3, is slightly stronger, which means we can replace \( 2Tb \) with \( 1.5Tb \) in Line 3 of the sub-routine \( \text{QUANTILE\textsc{Est}} \). Note that this change of the value of \( b \) itself improves the runtime of CAR, and we call the resulting algorithm CAR++, which will also be examined in the experiment section.

To prove Theorem 1, we need to (i) prove the correctness of \( \text{IMPATIENT\textsc{CapsAndRuns}} \), that is, the \((\varepsilon, \delta, \gamma)\)-optimality of the configuration returned by the algorithm; and (ii) give a bound on the total runtime. Starting with the correctness, we note that the algorithm proceeds in iterations from \( K - 1 \) to 0 in decreasing order, sampling bigger and bigger sets of configurations \( \mathcal{N}_k \). Each new set \( \mathcal{N}_k \), together with those configurations sampled before for \( k' > k \), contains an \( \text{OPT}_{\delta/2}^{\gamma/\varepsilon} \)-optimal configuration with high probability (Lemma 3), in other words, a configuration from an exponentially decreasing small quantile of the best configurations. The size of \( \mathcal{N}_k \), for all \( k \in [0, K - 1] \) is roughly \( \log(K/\varepsilon)/\gamma_k \) (Lemma 6). Next, we prove in Lemma 8 that \text{PRECHECK} does not reject a good configuration, and does reject a truly bad configuration. Unlike other parts of the proof, we do not guarantee this to hold with high probability for all configurations, instead we guarantee it to hold with high probability for any one configuration per each iteration \( k \); this will be chosen later to be one of the \( \text{OPT}_{\delta/2}^{\gamma/\varepsilon} \)-optimal configurations. Then, Lemma 10 shows that there remains an \( \text{OPT}_{\delta/2}^{\gamma/\varepsilon} \)-optimal configuration after each iteration \( k \) (Line 1 of Algorithm 1) that is not rejected by \( \text{QUANTILE\textsc{Est}} \) or \( \text{RUNTIME\textsc{Est}} \). This is because even if our designated configuration was rejected by \text{PRECHECK}, that means that there was an even better configuration, which from the proof of \( \text{CAPSANDRUNS} \), by Lemma 9 will not be rejected by \( \text{QUANTILE\textsc{Est}} \) or \( \text{RUNTIME\textsc{Est}} \). Several corollaries follow from this. Corollary 11 shows that with high probability, the configuration \( \text{IMPATIENT\textsc{CapsAndRuns}} \) returns in the end is \((\varepsilon, \delta, \gamma)\)-optimal, showing the correctness of the algorithm.

To prove the runtime bound, we start by showing that in every iteration \( k \), \( T \) is set to at most \( 2\text{OPT}_{\delta/2}^{\gamma/\varepsilon} \), after evaluating a configuration for no more than \( 4b\text{OPT}_{\delta/2}^{\gamma/\varepsilon} \) time (Corollary 12).

From this, Corollary 13 deduces a runtime bound for CAR in each iteration, which depends on the number of configurations surviving \text{PRECHECK}. Using the correctness analysis of \text{PRECHECK} (Lemma 3), Lemma 14 gives an upper bound on this number, essentially saying that roughly only \( F(38\text{OPT}_{\delta/2}^{\gamma/\varepsilon}+1) \) fraction of the \( \mathcal{N}_k \) configurations survive \text{PRECHECK} in round \( k \), where \( F(x) \) is roughly the probability of a random configuration sampled from II having a larger 0.35th quantile-capped runtime than \( x \). That is, essentially only those configurations survive which can solve at least 65% of the problem instances reasonably fast.

This is complemented by Lemma 15 which gives a runtime bound for \text{PRECHECK}, relying on Lines 13 and 21 of \text{PRECHECK} (Algorithm 4) stopping lengthy evaluations. To finish the proof, we combine the runtime bounds for all the components of ICAR discussed above. The lemmas above introduce various high-probability events under which their statements hold (by guaranteeing mostly that our bounds on the runtime caps and on the average runtimes hold), and a union bound over them proves that all those events hold simultaneously with probability at least \( 1 - 12\zeta \), proving Theorem 1.

**Lemma 2 (Improved version of Lemma 3 of [35]).** Let \( \tau \) be a constant satisfying 0 \( \leq \tau \leq t_{\delta/2}(i) \), and let \( \bar{Z}_r(i, j) \), \( j \in [1, b] \), be \( b \) runtime measurements of configuration \( i \) with timeout \( \tau \). Let \( \bar{Z}_r(i) \) be their average and \( R_r(i) \) their expectation. Then for \( c > 0 \), \( \Pr \left( \left| \bar{Z}_r(i) - R_r(i) \right| \geq cR_r(i) \right) \leq 2 \exp \left( \frac{-bh^2}{2(1+c/3)} \right) \). In particular, for \( S_i = \{ \frac{1}{2}R_r(i) \leq \bar{Z}_r(i) \leq 1.5R_r(i) \} \) and \( b = \left\lceil \frac{2n \log \left( \frac{2n}{\varepsilon} \right)}{\varepsilon} \right\rceil \), we have \( \Pr(S_i^c) \leq \frac{\zeta}{\pi} \) (by substituting \( c = \frac{1}{2} \)).

---

This constant 0.35 can be set arbitrarily, and it only affects other constants in the algorithm. It was set to 0.35 so that these constant do not increase beyond how large they have to be to guarantee other statements with high probability.
Proof. Since \( Z_\tau(i, j) \leq \tau \), \( \Var(Z_\tau(i)) = \Var_{j \sim R}[Z_\tau(i, j)] \leq \mathbb{E}_{j \sim R} Z_\tau^2(i, j) \leq \tau R_\tau(i) \). As at least \( \delta/2 \) fraction of instances run longer than \( \tau \), we have that \( R_\tau(i) \geq \frac{\delta}{2} \tau \), so \( \Var(Z_\tau(i)) \leq \frac{\delta}{2} R_\tau^2(i) \).

Using the Bernstein inequality,

\[
\Pr \left( |\bar{Z}_\tau(i) - R_\tau(i)| \geq c R_\tau(i) \right) \leq 2 \exp \left( -\frac{bc^2 R_\tau^2(i)}{\frac{\delta}{2} c R_\tau(i) + \Var(Z_\tau(i))} \right) 
\leq 2 \exp \left( -\frac{bc^2 R_\tau^2(i)}{\frac{\delta}{2} c R_\tau(i)} \right) 
= 2 \exp \left( -\frac{bc^2}{4(1 + c/3)} \right) .
\]

Remark 3. From Lemma 6 of [33], there is an event \( E_1 \) (with the notation of [33], this event is \( E_1 \cap E_2 \cap E_3 \) with \( \Pr(E_1) \geq 1 - 6\zeta \), under which all the high-probability statements in the analysis of \( \text{CAPS\_AND\_RUNS} \) hold for the algorithm with the constants improved as above. In particular, \( E_1 \) guarantees that the average runtime estimates of \( \text{CAR} \) are close to their expectations, and that \( \text{QUANTILE\_EST} \) measures an accurate cap for each configuration such that \( t_\delta(i) \leq \tau_i \leq t_{\delta/2}(i) \).

Lemma 4. There is an event \( E_2 \) with \( \Pr(E_2) \geq 1 - \zeta \) such that under \( E_2 \), for all integers \( k \in [0, K - 1] \), there is a configuration \( i \in \bigcup_{j=k}^{K-1} \mathcal{N}_j \) with \( R_\tau^2(i) \leq \text{OPT}_{\delta/2}^\gamma \) after Line 2 of \( \text{IMPATIENT\_CAPS\_AND\_RUNS} \) (Algorithm 1).

Proof. For any \( i \) chosen randomly from the distribution \( \Pi \), \( R_\tau^2(i) \leq \text{OPT}_{\delta/2}^\gamma \) with probability at least \( \gamma_k \). As the configurations are sampled independently, the probability that none of the sampled configurations are optimal for \( \gamma_k \) is at most \((1 - \gamma_k)^{\left| \bigcup_{j=0}^{K-1} \mathcal{N}_j \right|} = (1 - \gamma_k)^{\left[ \log(\gamma/K) \right]} \leq \zeta/K \). Applying the union bound over \( k \in [0, K - 1] \), with probability at least \( 1 - \zeta \), for all \( k \in [0, K - 1] \), there is a configuration \( i \) with \( R_\tau^2(i) \leq \text{OPT}_{\delta/2}^\gamma \) sampled into \( \bigcup_{j=k}^{K-1} \mathcal{N}_j \) at Line 2 of \( \text{IMPATIENT\_CAPS\_AND\_RUNS} \).

Remark 5. Noting that \( \gamma_0 = \gamma \), and \( \bigcup_{k=0}^{K-1} \mathcal{N}_k = \mathcal{N} \), the previous lemma with \( k = 0 \) states that under \( E_2 \), a configuration \( i \) with \( R_\tau^2(i) \leq \text{OPT}_{\delta/2}^\gamma \) is sampled into \( \mathcal{N} \).

In the following we refer to the last part of Algorithm 1 (Lines 14 to 18) iteration \(-1\), denote it with \( k = -1 \), and accordingly define \( \mathcal{N}_{-1} = \mathcal{N} \) and \( \mathcal{N}_{-1} = \mathcal{N} \).

Lemma 6. After Line 2 of Algorithm 1 for all \( k \in [-1, K - 1] \), \( |\mathcal{N}_k| \leq \log(K/\zeta)/\gamma_k + 1 \).

Proof. Using that for any \( x \in (0, 1) \), \( x \leq -\log(1 - x) \), we have for any \( k \in [-1, K - 1] \) that

\[
|\mathcal{N}_k| \leq \left\lfloor \frac{\log(\zeta/K)}{\log(1 - \gamma_k)} \right\rfloor \leq \frac{\log(K/\zeta)}{-\log(1 - \gamma_k)} + 1 \leq \log(K/\zeta)/\gamma_k + 1.
\]

By [33], Lemma 2], under \( E_1 \), and noting that \( T \) can only be set by \( \text{RUNTIME\_EST} \) evaluating a configuration after the cap \( \tau \) for that configuration has already been measured by \( \text{QUANTILE\_EST} \):

Lemma 7. If a configuration \( i \) sets \( T \), then \( t_\delta(i) \leq \tau_i \leq t_{\delta/2}(i) \).

Lemma 8. Suppose \( \delta \leq 0.2 \) and assume that we are in the \( \text{PRECHECK} \) call in iteration \(-1 \leq k < K - 1 \) of Algorithm 1 (recall that iteration \(-1 \) refers to the last part of the algorithm after the iteration loop is finished). Let \( \Pr_k \) denote the conditional probability conditioned on all the random events before the call to \( \text{PRECHECK} \). Let \( i' \) be the configuration that was last evaluated to set \( T \) by \( \text{RUNTIME\_EST} \) (in an iteration \( k' > k \)). For any \( i \in \mathcal{M} \), there is an event \( E_{3,k,i} \) with \( \Pr_k(E_{3,k,i}) \geq 1 - 4\zeta/K \) such that under \( E_1 \) and \( E_{3,k,i} \), (1) if \( \text{OPT}(i) \leq R_{\tau_2}(i') \), \( i \) won’t be rejected by \( \text{PRECHECK} \), and (2) if \( R_{0.35}(i) \geq 19T \), then \( i \) will be rejected by \( \text{PRECHECK} \).
Proof. Consider the evaluation of configuration $i$ in PRECHECK. Let $I_i$ be the indicator that the $i^{th}$ instance in Phase I of PRECHECK takes at least $t_{1/10}(i)$ time to complete (without capping). The $I_i$ are independent and identically distributed Bernoulli random variables with $Pr_k(I_i = 1) = 1/10$. We use the Chernoff bound to get that $Pr_k\left(\sum_{i=0}^{b'} I_i > 0.2b'\right) = Pr_k\left(\sum_{i=0}^{b'} I_i > \frac{1}{10}b'(1 + 1)\right) \leq \exp\left(-\frac{\delta^2}{2}\right) \leq \frac{\zeta}{(2K)}$. Let $E_{3,k,i,1}$ be the event that $\sum_{i=0}^{b'} I_i \leq 0.2b'$. Similarly, defining $J_i$ to be the indicator that the $i^{th}$ instance in Phase I of PRECHECK takes at least $t_{0.35}$ time to complete (without capping), noting that $Pr_k(J_i = 1) = 0.35$, the Chernoff bound implies that $Pr_k\left(\sum_{i=0}^{b'} J_i < 0.2b'\right) = Pr_k\left(\sum_{i=0}^{b'} J_i < 0.35b'(1 - \frac{0.35-0.2}{0.35})\right) \leq \frac{\zeta}{(2K)}$. Let $E_{3,k,i,2}$ be the event that $\sum_{i=0}^{b'} J_i \geq 0.2b'$.

So for any configuration $i \in \mathcal{M}$, under $E_{3,k,i,1}$, the number of samples from the 1/10-tail will be at most $\lceil 0.2b' \rceil$, and under $E_{3,k,i,2}$, the number of samples from the 0.35-tail will be at least $\lceil 0.2b' \rceil$, so picking the $\lceil 0.8b' \rceil^{th}$ finished run and denoting it by $\tau'$ (in Line 15 of Algorithm 3) ensures under and $E_{3,k,i,1}$ and $E_{3,k,i,2}$ that $t_{0.35}(i) \leq \tau' \leq t_{1/10}(i)$. For $\delta \leq 0.2$ (which we have by assumption), this implies that $\tau' \leq t_{0.35}(i)$.

By [35, Lemma 7], under $E_1$, $R_{\tau'}(i') \leq T$. Assuming that $R_{\hat{\tau}}(i) \leq R_{\tau'}(i')$ for (1), we have $R_{\hat{\tau}}(i) \leq T$. Then under $E_{3,k,i,1}$, $R_{\tau'}(i) \leq R_{1/10}(i) \leq T$ (as by assumption $\delta \leq 0.2$ and $\tau' \leq t_{1/10}(i)$, so $R_{\tau'}(i) \leq R_{1/10}(i) \leq R_{\hat{\tau}}(i) \leq T$). There are two cases in which we reject configuration $i$. First, if $\text{avg}(Y) - C \geq T$. By the empirical Bernstein bound [4], there is an event $E_{3,k,i,3}$ such that $Pr_k(E_{3,k,i,3}) \geq 1 - \frac{\zeta}{K}$, and under $E_{3,k,i,3}$, $\text{avg}(Y) - \text{avg}(\tau') \leq C$. As $E_1[\text{avg}(Y)] \cdot \text{avg}(\tau') = R_{\tau'}(i) \leq T$, we have that under $E_1$, $E_{3,k,i,1}$ and $E_{3,k,i,3}$, configuration $i$ in iteration $k$ will not be rejected in Line 27 of PRECHECK if (1) holds.

The second type of rejection happens in Line 13 of PRECHECK when Phase I of PRECHECK runs for at least $1.9Tb'$ time. For each run $i$ that is performed in Phase I, denote by $X_i$ the hypothetical runtime of that instance if the cap were $t_{1/10}(i)$, and by $Y_i$ the runtime if the runtime cap were $t_{0.35}(i)$.

From the above, under $E_{3,k,i,1}$ $\tau' \leq t_{1/10}(i)$, and if we had to abort then that means we haven’t run any instance for $\tau'$ time yet, so by denoting measurements performed so far by Phase I of PRECHECK by $\hat{X}_i$, we have $X_i \leq \hat{X}_i$, so when we abort we have that $\text{avg}(X) \geq 1.9T$.

Applying Lemma 2 with $c = 0.9$, $b = b' = \left\lceil 32.1 \log \left( \frac{2K}{\epsilon} \right) \right\rceil$, $\delta = 0.2$, and $\tau = t_{1/10}(i)$, we get that $Pr_k\left(\text{avg}(X) \geq R_{1/10}(i) \right) \leq 2 \exp\left( b' \frac{81}{5500} \right) \leq \frac{\zeta}{K}$. Denote by $E_{3,k,i,4}$ the event that $\text{avg}(X) \leq 1.9T$. Then, for (1), under $E_1$ and $E_{3,k,i,1}$, by the above $R_{1/10}(i) \leq T$, we have $Pr_k(\text{avg}(X) \geq 1.9T) \leq Pr_k(\text{avg}(X) - R_{1/10}(i) > 0.9R_{1/10}(i)) \leq \frac{\zeta}{K}$, so $Pr_k(E_{3,k,i,4}) \leq 1 - \frac{\zeta}{K}$, and under $E_1$ and $E_{3,k,i,4}$, this configuration won’t be rejected in Line 13 of PRECHECK if it satisfies (1).

For (2), let $E_{3,k,i,5}$ the event that $\text{avg}(Y) \geq 0.1R_{0.35}(i)$. Apply Lemma 2 with the same parameters except $\tau = t_{0.35}(i)$, to get that $Pr_k\left(\text{avg}(Y) \leq 0.1R_{0.35}(i) \right) = Pr_k\left(\text{avg}(Y) - R_{0.35}(i) \leq 0.9R_{0.35}(i) \right) \leq 2 \exp\left( b' \frac{81}{5500} \right) \leq \frac{\zeta}{K}$. For PRECHECK to not reject a configuration $i$, it measures a cap $\tau' \geq t_{0.35}(i)$ (under $E_{3,k,i,2}$), and so the measurements $Y_i$ satisfy $Y_i \geq \hat{Y}_i$, so we spend $b' \text{avg}(\hat{Y}_i) \geq b' \text{avg}(Y_i) \geq 0.1R_{0.35}(i)$ time for configuration $i$ under $E_{3,k,i,5}$. Thus, with probability at least $Pr_k(E_{3,k,i,4}) \leq 1 - \frac{\zeta}{K}$, a configuration where $R_{0.35}(i) \geq 10T$ is rejected in Line 13. Taking a union bound and letting $E_{3,k,i} = E_{3,k,i,1} \cap E_{3,k,i,2} \cap E_{3,k,i,3} \cap E_{3,k,i,4} \cap E_{3,k,i,5}$ (the event that all the high-probability statements above hold for configuration $i$ and iteration $k$), we have that $Pr_k(E_{3,k,i}) \geq 1 - 4\zeta/K$. From the proof of [35, Theorem 1] we can extract the following result:

Lemma 9. Let $\mathcal{N}$ be the set of configurations CAPSANDRUNS is called with, and $\mathcal{N}'$ the ones among these that are not rejected in QUANTILEEST. Let $i_* = \min_{i \in \mathcal{N}'} R_{\tau'}(i)$. Under $E_1$, $i_*$ is not rejected in RUNTIMEEST and CAPSANDRUNS returns a configuration $I$ for which $R_{\tau'}(I) \leq (1 + \epsilon)R_{\tau'}(i_*)$.

To proceed, we instantiate the events $E_{3,k,i}$ of Lemma 8 for one of the best configurations $i$ in $\mathcal{N}_k$.

By Lemma 4 and Remark 5, under $E_2$, for every iteration $k$ of IMPATIENTCAPSANDRUNS, there is a configuration $i^*_k \in \bigcup_{j=k}^{\infty} \mathcal{N}_j$ such that $R_{\hat{\tau}}(i^*_k) \leq \text{OPT}^{\text{best}}$. Furthermore, this guarantees that
\[ \hat{\gamma}_k \in \mathcal{N} \text{ satisfies } R^\hat{\gamma}_k(i_0^k) \leq \text{OPT}^\gamma_{\delta/2}, \] which also implies, through the first part of Lemma 9, that under \( E_1 \), there is a configuration \( i_{-1}^k \) satisfying \( R^\hat{\gamma}_k(i_{-1}^k) \leq \text{OPT}^\gamma_{\delta/2} \). Now we define the following event \( E_4 \subset E_1 \cap E_2 \) as \( E_4 = \cap_{k=1}^{K-2} E_{3,k} \cap E_1 \cap E_2 \).

**Lemma 10.** Under \( E_4 \), for all integer \( 0 \leq k \leq K - 1 \), there is a configuration \( i_k^* \) remaining in \( \bigcup_{j=k}^{K-1} \mathcal{N}_j \) at the end of the \( k \)th iteration (after Line 11 in Algorithm 1), that is not rejected by QUANTILE\( \text{EST} \) or RUNTIME\( \text{EST} \), for which \( R_{\gamma_k}(i_k^*) \leq \text{OPT}^\gamma_{\delta/2} \). Similarly, there is a configuration \( i_* \) remaining in \( \mathcal{N} \) at the end of the final \( \text{CAPSANDRUNS} \) call (after Line 17 in Algorithm 1), for which \( R_{\gamma_*}(i_*) \leq \text{OPT}^\gamma_{\delta/2} \).

**Proof.** Suppose \( E_4 \) holds (this also means that \( E_1 \) and \( E_2 \) hold). Let \( i^\prime \) denote the configuration that last set \( T \).

For \( k = K - 1 \) there is no \texttt{PRECHECK} as \( T = \infty \), in other words nothing is rejected by \texttt{PRECHECK}.

For iterations \( 0 \leq k \leq K - 2 \), and for the final \texttt{CAPSANDRUNS} call, either \( R^\hat{\gamma}_k(i_k^*) \leq R_{\gamma_k}(i^\prime) \), in which case by Lemma 8 under \( E_1 \) and \( E_{3,k} \), \( i_k^* \) is not rejected, or \( R^\hat{\gamma}_k(i_k^*) > R_{\gamma_k}(i^\prime) \). Thus under \( E_4 \), \( R^\hat{\gamma}_k(i_k^*) > R_{\gamma_k}(i^\prime) \) holds whenever \( i_k^* \) is rejected. We assume this for the rest of the proof.

The remainder of this proof handles iterations \( 0 \leq k \leq K - 1 \), but the arguments transfer for the final \texttt{CAPSANDRUNS} call case by writing \( \gamma \) and \( \hat{\gamma}_{-1} \) instead of \( \gamma_k \) and \( \hat{\gamma}_k \). We investigate the two possible cases:

- If \( i_k^* \) is not rejected by \texttt{PRECHECK}, then under \( E_1 \) by Lemma 8 in [35], there is an \( i \) in the set of configurations \texttt{CAPSANDRUNS} is called with, that will not be rejected by QUANTILE\( \text{EST} \), and for which \( R_{\gamma_k}(i) \leq R^\hat{\gamma}_k(i_k^*) \leq \text{OPT}^\gamma_{\delta/2} \).

- If \( i_k^* \) is rejected by \texttt{PRECHECK}, \( i^\prime \) is a configuration not rejected by QUANTILE\( \text{EST} \) (as it set \( T \), for which \( R_{\gamma_k}(i^\prime) \leq R^\hat{\gamma}_k(i_k^*) \leq \text{OPT}^\gamma_{\delta/2} \).

In either case, there is a configuration \( i \) not rejected by QUANTILE\( \text{EST} \), for which \( R_{\gamma_k}(i) \leq \text{OPT}^\gamma_{\delta/2} \).

Thus by Lemma 9 under \( E_1 \), there is a configuration \( i_k^* \) not rejected by QUANTILE\( \text{EST} \) or RUNTIME\( \text{EST} \) for which \( R_{\gamma_k}(i_k^*) \leq \text{OPT}^\gamma_{\delta/2} \).

**Corollary 11.** Under \( E_4 \), the configuration returned by \texttt{IMPATIENTCAPSANDRUNS} is \((\varepsilon, \delta, \gamma)\)-optimal.

**Proof.** By Lemma 9 and Lemma 10 under \( E_4 \), the final \texttt{CAPSANDRUNS} call returns with a configuration \( I \) for which \( R_{\gamma_k}(I) \leq (1 + \varepsilon)R_{\gamma_k}(i_*) \leq (1 + \varepsilon)\text{OPT}^\gamma_{\delta/2} \). Under \( E_1 \), \( R^\delta(I) \leq R_{\gamma_1}(I) \), so \( I \) is \((\varepsilon, \delta, \gamma)\)-optimal.

**Corollary 12.** Under \( E_4 \), for all iterations \( 0 \leq k \leq K - 1 \), \( T \) is set by QUANTILE\( \text{EST} \) to at most \( 2\text{OPT}^\gamma_{\delta/2} \), and the combined time spent by QUANTILE\( \text{EST} \) and RUNTIME\( \text{EST} \) evaluating the configuration that has set \( T \) is bounded by \( 4b\text{OPT}^\gamma_{\delta/2} \) when it sets \( T \).

**Proof.** Take \( i_k^* \) as in Lemma 10. Since \( i_k^* \) is not rejected in either QUANTILE\( \text{EST} \) or RUNTIME\( \text{EST} \), its \( b \) measurements in RUNTIME\( \text{EST} \) will complete, and by [35] Lemma 4, under \( E_1 \), this measurement will be at most \( 2R_{\gamma_k}(i_k^*) \leq 2\text{OPT}^\gamma_{\delta/2} \). \( T \) is thus set to at most this value. From the proof of [35] Lemma 5, the work spent by QUANTILE\( \text{EST} \) and RUNTIME\( \text{EST} \) evaluating \( i_k^* \) is bounded by \( 4b\text{OPT}^\gamma_{\delta/2} \) time.

**Corollary 13.** Suppose \( E_4 \) holds. Then for all iterations \( 0 \leq k \leq K - 1 \), \texttt{CAPSANDRUNS} performs at most \( \tilde{O}\left(b\sum_{k=1}^{K-1} \text{OPT}^\gamma_{\delta/2}\right) \) work.

**Proof.** By Corollary 12 under \( E_4 \), in iteration \( k \), \( T \) is set to at most \( 2\text{OPT}^\gamma_{\delta/2} \), after which by the proof of [35] Lemma 5, each configuration performs at most \( \tilde{O}\left(b\text{OPT}^\gamma_{\delta/2}\right) \) work. Also by
Corollary \[12\] the work performed by the configuration that set \( T \) to this value in iteration \( k \) is upper bounded by \( 4b\text{OPT}^{\gamma_k}_{\frac{\delta}{2}} \). Since configurations are run in parallel, all the other configurations performed at most this amount of work in the meantime. Thus in total \textsc{CapsAndRuns} performs at most \( \tilde{O}\left(b|\mathcal{N}_k|\text{OPT}^{\gamma_k}_{\frac{\delta}{2}}\right) \) work in iteration \( k \).

Lemma 14. There is an event \( E_5 \) such that \( \Pr(E_5) \geq 1 - \zeta \), and under \( E_5 \), \( E_1 \), and \( E_4 \), for all iterations \( k \in [-1, K - 2] \), the number of configurations not rejected by \textsc{Precheck} can be bounded as

\[
|\mathcal{N}_k| \leq (\log(K/\zeta) + 1) \left[ F(38\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}}) \frac{1}{\gamma_k} + \sqrt{2F(38\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}})} \frac{1}{\gamma_k} + 2 \right],
\]

where \( F(x) = \Pr_{i \sim \Pi}(R^{0.35}(i) \leq x) + 4\zeta/K \).

Proof. Note that for the first call of \textsc{Precheck}, with \( k = K - 1 \), \textsc{Precheck} returns its input without any modification, so \( |\mathcal{M}'| = |\mathcal{M}| \). For the rest of the calls, \(-1 \leq k < K - 1 \).

Denoting by \( B_i \) the indicator whether configuration \( i \in \mathcal{N}_k \) is accepted by \textsc{Precheck}. Since elements of \( \mathcal{N}_k \) are independent and identically distributed random variables, and there are no interactions between configurations being evaluated by \textsc{Precheck}, the outcomes \( B_i \) of \textsc{Precheck} are also independent and identically distributed. By Lemma \[8\] under \( E_1 \), \textsc{Precheck} rejects a configuration \( i \) if \( R^{0.35}(i) \geq 19T \) with probability at least \( 1 - 4\zeta/K \), so the probability of reject is at least \( \Pr_{i \sim \Pi}(R^{0.35}(i) \geq 19T) \geq \Pr_{i \sim \Pi}(R^{0.35}(i) \geq 19T | T) - 4\zeta/K = 1 - F(19T) \), so the conditional probability of accept is at least \( F(19T) \). The number of configurations not accepted is \( \sum_{i \in \mathcal{N}_k} B_i \). By the Bernstein inequality,

\[
\Pr \left[ \sum_{i \in \mathcal{N}_k} B_i \geq F(19T)|\mathcal{N}_k| + \sqrt{2F(19T)|\mathcal{N}_k| \log\frac{K}{\zeta} + \frac{2}{3} \log\left(\frac{K}{\zeta}\right) |T| \right] \leq \frac{\zeta}{K},
\]

so by a union bound over \( k \in [-1, K - 2] \), this is true for all iterations under an event \( E_5 \) with probability at least \( 1 - \zeta \). By Lemma \[6\] \( |\mathcal{N}_k| \leq (\log(K/\zeta))/\gamma_k + 1 \). By Corollary \[12\] under \( E_4 \), \( T \leq 2\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \) when \textsc{Precheck} is run for iteration \( k \). Making these substitutions and reordering the terms gives the result.

Lemma 15. For iterations \(-1 \leq k < K - 1 \), under \( E_4 \), \textsc{Precheck} runs for at most \( 10\text{OPT}^{\gamma_k}_{\frac{\delta}{2}} |32.1 \log\left(\frac{2K}{\zeta}\right)| (\log(K/\zeta)/\gamma_k + 1) \) time.

Proof. By Corollary \[12\] under \( E_4 \), \( T \leq 2\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \) when \textsc{Precheck} is run for iteration \( k \). Phase I of \textsc{Precheck} is aborted when the total runtime reaches \( 9.7T' \leq 3.8\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \). Phase II of \textsc{Precheck} is aborted when the total Phase II runtime exceeds \( 2.99T' \leq 5.98\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \). This abort only happens after the last run, which takes at most \( \tau' \) time, where \( \tau' \) is measured in Phase I of \textsc{Precheck}. Because of the way \( \tau' \) is calculated by Phase I, at least \( \lfloor 0.2b' \rfloor \) instances were running up until \( \tau' \) time, which took \( \lfloor 0.2b' \rfloor \tau' \leq 1.97T'b' \) time. For any valid setting of \( \zeta, |0.2b'| \geq 0.19b' \), so \( \tau' \leq 10T \leq 20\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \leq 0.21\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \), so the work of \textsc{Precheck} for each configuration is upper bounded by \( (3.8 + 5.98 + 0.21)\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \leq 10\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} \). Multiplying this by the number of configurations \( |\mathcal{N}_k| \) \textsc{Precheck} evaluates, and using Lemma \[6\] the total work of \textsc{Precheck} is bounded by \( 10\text{OPT}^{\gamma_k+1}_{\frac{\delta}{2}} |32.1 \log\left(\frac{2K}{\zeta}\right)| (\log(K/\zeta)/\gamma_k + 1) \).

Proof of Theorem \[7\] Suppose \( E_4 \) and \( E_5 \) hold. By a union bound, taking also into account the lower bounds on the probabilities of these events and for events \( E_1, E_2, \) and \( E_{3, k+1, j} \) (given by Remark \[5\], Lemma \[4\], Lemma \[8\], Lemma \[14\]), we have \( \Pr(E_1 \cap E_2 \cap E_4 \cap E_5) \geq 1 - 12\zeta \). By Corollary \[11\] under these events, the configuration returned by \textsc{ImpatientCapsAndRuns} is \((\delta, \epsilon, \gamma)\)-optimal.

Next we consider the runtime of \textsc{ImpatientCapsAndRuns}. For iteration \( k = K - 1 \), \( E_1, E_2, \) and \( E_4, \) by Corollary \[13\] and Lemma \[6\], the runtime of \textsc{CapsAndRuns} is \( \tilde{O}\left(b\text{OPT}^{\gamma_{K-1}}_{\frac{\delta}{2}}\right) \). For iterations \( 0 \leq k < K - 1 \), by Corollary \[13\], the runtime of \textsc{CapsAndRuns} in iteration \( k \) is upper bounded

16
as \( \tilde{O} \left( b|\mathcal{N}_k| \text{OPT}^\gamma_{\delta/2} \right) \). Using the bound \(|\mathcal{N}_k| = \tilde{O} \left( (38\text{OPT}^\gamma_{\delta/2})^2 / \gamma_k \right) \) given by Lemma 14, the work by \textsc{CapsAndRuns} in iteration \( k \) is bounded by \( \tilde{O} \left( bF(38\text{OPT}^\gamma_{\delta/2})^2 \text{OPT}^\gamma_{\delta/2} / \gamma_k \right) \).

For the final \textsc{CapsAndRuns} call, the total work performed by \textsc{CapsAndRuns} would only increase if we didn’t do any work on any configurations before, in other words, if we restarted \textsc{CapsAndRuns} with the input configurations \( \mathcal{N} \). By this idea we can upper bound the total work of the final \textsc{CapsAndRuns} call using [35] Theorem 1, which states that under \( E_1 \), the total work of a restarted \textsc{CapsAndRuns} with input configurations \( \mathcal{N} \) is at most \( \tilde{O} \left( |\mathcal{N}| \frac{1}{\varepsilon^2 \delta} \min_i \mathbb{E} R^2(i) \right) \), which is a simplified form of the problem-dependent bound (1) in [35]. By Lemma 10, \( \min_i \mathbb{E} R^2(i) \leq \text{OPT}^\gamma_{\delta/2} \), and by Lemma 14, \( |\mathcal{N}| = \tilde{O} \left( F(38\text{OPT}^\gamma_{\delta/2}) / \gamma \right) \). Plugging these in the bound we get that the final \textsc{CapsAndRuns} takes \( \tilde{O} \left( \text{OPT}^\gamma_{\delta/2} F(38\text{OPT}^\gamma_{\delta/2}) \frac{1}{\varepsilon^2 \delta} \right) \) time.

Now we turn our attention to bounding the work done in \textsc{Precheck}. By Lemma 15 under \( E_1 \), for all the iterations, and including the final \textsc{Precheck} call, the total work is \( \tilde{O} \left( \sum_{k=0}^{K-1} \text{OPT}^\gamma_{\delta/2} / \gamma_k \right) \).

Adding all this work up, noting that \( b = \tilde{O}(1/\delta) \), we get that under \( E_4 \) and \( E_5 \), the total work performed by \textsc{ImpatientCapsAndRuns} is

\[
\tilde{O} \left( \frac{1}{\varepsilon^2 \delta^2} \text{OPT}^\gamma_{\delta/2} F(38\text{OPT}^\gamma_{\delta/2}) + \sum_{k=0}^{K-2} \frac{1}{\gamma_k} \text{OPT}^\gamma_{\delta/2} \left( 1 + F(38\text{OPT}^\gamma_{\delta/2}) / \gamma \right) + \frac{1}{\delta^2 \gamma K-1} \text{OPT}^\gamma_{\delta/2} \right).
\]

\[ \square \]

### B Runtime bound analysis with exponential distributions

To better understand the runtime bound in Theorem 1, consider a scenario where the runtime of each configuration follows an exponential distribution. Such scenarios are realistic and motivated by practical applications: roughly speaking, many solvers for NP-hard problems (e.g., SAT) proceed by initializing with a random seed and, if they fail, try again with another random seed. The runtimes of such solvers will follow approximately exponential distributions, due to being essentially memoryless.

To make the example concrete, suppose that the mean runtime for each configuration is distributed uniformly between \( A \) and \( A + B \), denoted by \( U(A, A + B) \), for some \( A, B > 0 \). Here \( A \) can be thought of as a small runtime associated with the cost of starting the run of any configuration on any problem instance, and \( B \) as the maximum “true” mean runtime of the configurations.

We can simplify the runtime bound of Theorem 1 with this assumption. The best \( \gamma_k \frac{1}{\gamma} \) fraction of the configurations have mean \( A + B \gamma_k \), so \( \text{OPT}^\gamma_{\delta/2} \leq A + B \gamma_k \). Furthermore, for a configuration \( i \) with mean \( \frac{1}{\gamma} \), \( R_k(i) = \frac{1}{\gamma} (1 - e^{-\lambda \tau}) \) for any runtime cap \( \tau \). Substituting \( \tau = t_k(i) \), noting that the probability of running over the cap is \( \delta e^{-\lambda \tau} = \delta \), we have \( R_k(i) = \frac{1}{\gamma} (1 - \delta) \). Similarly, \( \text{R}_{0.35}(i) = 0.65 \gamma \).

Then \( F(38\text{OPT}^\gamma_{\delta/2}) / (4\gamma K) = \max_{i \sim U(A, A + B)} (4\gamma K) / (4\gamma K) \leq \max_{i \sim U(A, A + B)} (4\gamma K) / (4\gamma K) \leq \frac{58.5(A + B \gamma_k)}{A + B \gamma_k} = O(\gamma_k + A \gamma_k) \).

This bounds \( F(38\text{OPT}^\gamma_{\delta/2}) / (4\gamma K) \) in inexpressible, as the failure probability \( \zeta \) can simply be chosen to be \( O(\varepsilon^2 \delta) \) with only additional logarithmic factors in the runtime as a result, and then any term multiplied by \( \zeta \) in the runtime bound disappears in the \( \tilde{O}(\cdot) \) notation. Substituting the bound on \( F \) and \( \text{OPT}^\gamma_{\delta/2} \), assuming a choice of \( \zeta \) as above, the runtime bound (Eq. 11) becomes

\[
\tilde{O} \left( \frac{A + B \gamma_k}{\varepsilon^2 \delta \gamma} \left( \gamma + \frac{A}{B} \right) + \sum_{k=0}^{K-2} (A + B \gamma_k) \frac{1}{\gamma_k} \left( 1 + \frac{(A + B \gamma_k)}{\delta} \right) + (A + B \gamma K-1) \frac{1}{\delta^2 \gamma K-1} \left( \frac{A}{\gamma} \right) \right).
\]

where the \( K \) term disappears in the \( \tilde{O}(\cdot) \) notation as \( K \leq \log_2(\frac{1}{\zeta}) \). Contrasting this with the typical runtime bound \( \tilde{O} \left( \frac{A + B}{\varepsilon^2 \delta} \text{OPT}^\gamma_{\delta/2} \right) = \tilde{O} \left( \frac{A}{\gamma} \right) \) of previous works, we see the main term...
Figure 2: Distribution of \( \delta \)-capped mean runtime of the sampled configurations, with \( \delta = 0.1 \). For Minisat/CNFuzzDD, many configurations are close to the optimal one, whereas for CPLEX/Regions200 and CPLEX/RCW, many configurations are significantly worse than the optimal one. Consequently, PRECHECK is able to discard more configurations in the latter two scenarios.

\[
\text{(the one multiplied by } \frac{1}{\varepsilon^2} \text{)} \text{ is reduced by a factor of } \max\{\gamma, \frac{A}{B}\}. \text{ The rest of the terms have no dependence on } \varepsilon \text{ and are indeed always much smaller than the typical runtime bound for other works: } \\
\frac{A}{\gamma K - 1} \text{ is smaller than the first term provided that } K \text{ is chosen to be large enough for } \gamma K - 1 \text{ not to be too small, } \frac{B}{\gamma} \text{ does not depend on the number of configurations evaluated, and } \frac{A}{\gamma} \text{ is associated with having to evaluate about } \frac{1}{\gamma} \text{ number of configurations, and this term could not scale better than with the minimum runtime } A. 
\]

C Details of Experiments

We followed the experimental setup of [35]. Runs were pre-computed and then queried from a simulation environment in which they can be stopped and resumed. In a scenario where this is not possible (e.g., due to memory constraints when performing real runs) the experiments can still be implemented by restarting runs from scratch with doubling cap times, resulting in at most a factor of 2 slowdown.

Parameter values Experiments on all datasets were done with \( (\varepsilon, \delta) = (0.05, 0.1) \) and varying \( \gamma \in \{0.01, 0.02, 0.05\} \). For each configurator, \( \zeta \) was set so that the total failure probability is 0.05. The hyperparameter \( K \) was set such that \( 0.25 < \gamma 2^{K-1} \leq 0.5 \). This is a somewhat arbitrary choice, but was made so that values of \( \gamma_K \) were neither too big to be trivial, nor too small to be computationally prohibitive.

EPM Setup We used the provided generators for Regions200 and RCW to produce as many new random instances as needed, which were pre-processed using the feature extractors provided with the EPM. Runtime-related features were dropped since they are machine-dependent. We then used the provided configurations and runtime data\footnote{http://www.cs.ubc.ca/labs/beta/Projects/EPMs/} to train the EPM model, using the parameters suggested in [12]. Finally, the trained EPM was used as a surrogate model to provide runtimes on future scenarios. We can query this model to produce a runtime estimate for any configuration-instance pair. New configurations were sampled by uniformly choosing a value for each parameter of CPLEX. A new instance was then generated and the pair was given to the EPM. Note that ICAR examined more configurations than CAR did. For consistency, sampling was done so that the configurations seen by CAR were a subset of those seen by ICAR.

Datasets Description

- Minisat/CNFuzzDD is a SAT scenario based on the minisat solver, with 6 parameters, applied to the CNFuzzDD\footnote{http://fmv.jku.at/cnfuzzdd/} instances. The benchmark dataset we use is the same as in [34, 35, 24].
Figure 3: As the proportion of configurations that are far from the optimal gets larger (i.e., as \(c\) gets larger), the CPU runtime of CAR was more dominated by the work spent on bad configurations, while ICAR was able to drop more bad configurations with its PRECHECK mechanism.

- CPLEX/Regions200 is a MIP scenario based on the CPLEX, an integer programming solver, applied to the combinatorial auction winner determination problem. There are 74 parameters for CPLEX. The benchmark dataset is generated with the EPM described above.

- CPLEX/RCW is a MIP scenario based on the CPLEX solver, applied to Red-cockaded Woodpecker conservation problem. The configuration space is the same as CPLEX/Regions200. The benchmark dataset is generated with the EPM described above.

**D IMPATIENTCAPSANDRUNS** **with varying parameters**

We also compared the performance of ICAR and CAR++ with varying values of \(\varepsilon\) and \(\delta\) (with fixed \(\gamma = 0.02\) and failure probability \(\zeta = 0.05\)). The speedup (computed as the ratio of the runtimes) achieved by ICAR over CAR++ is reported in Table 2. As we can see, the speedup was fairly stable across a range of \(\varepsilon\) and \(\delta\) that a user might be likely to care about.

**Table 2:** Speedup achieved by ICAR over CAR++ for various values of \(\varepsilon\) and \(\delta\). Values greater than one indicate ICAR is faster.

<table>
<thead>
<tr>
<th>(\delta)</th>
<th>Minisat/CNFuzzDD</th>
<th>CPLEX/Regions200</th>
<th>CPLEX/RCW</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.025</td>
<td>0.80</td>
<td>2.76</td>
<td>2.27</td>
</tr>
<tr>
<td>0.05</td>
<td>0.83</td>
<td>2.54</td>
<td>2.54</td>
</tr>
<tr>
<td>0.075</td>
<td>0.84</td>
<td>2.58</td>
<td>2.66</td>
</tr>
<tr>
<td>0.1</td>
<td>0.85</td>
<td>2.60</td>
<td>2.73</td>
</tr>
</tbody>
</table>

**E Synthetic Experiments**

To better understand how well ICAR can exploit a needle-in-a-haystack scenario, we examined its performance on synthetic data. In this way we could choose each configuration’s true mean, and thus control their distribution. The runtimes of each configuration were sampled from an exponential distribution, with the means being uniformly chosen from the interval \([\text{OPT}, c \cdot \text{OPT}]\). We tend to think that real algorithm runtimes do look somewhat exponential, and there is justification for this, at least in certain cases [14, 16, 15, 25].

We ran the simulation for \(c \in \{2, 5, 10, 25\}\). The larger the value of \(c\), the more configurations will tend to be far from the best one, creating more and more of a “needle-in-a-haystack” scenario. We used \((\varepsilon, \delta, \gamma) = (0.05, 0.1, 0.02)\) and failure probability of 0.05, as before. Table 3 shows the total CPU time to find a (0.05, 0.1, 0.02)-optimal configuration for the range of \(c\) values. The degree to which ICAR outperformed CAR increases as \(c\) increases, as expected. We see that PRECHECK was able to reject a greater proportion of configurations when many tended to be far from optimal (large \(c\)).

Figure 3 shows the CPU time spent on each configuration, sorted by the \(\delta\)-capped mean runtime. When \(c = 2\), ICAR rejected very few configurations in PRECHECK, but as \(c\) increases we can see a greater proportion of configurations were being run for minimal time compared to CAR++. Furthermore, the runtime of CAR and CAR++ became dominated by the runs on the bad configurations, as we can see from the area under the curve.
<table>
<thead>
<tr>
<th></th>
<th>Total CPU Time (days)</th>
<th>Number of Configurations Before/After Precheck</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c = 2$</td>
<td>$c = 5$</td>
</tr>
<tr>
<td>ICAR</td>
<td>505 (23)</td>
<td>187 (18)</td>
</tr>
<tr>
<td>CAR++</td>
<td>344 (24)</td>
<td>214 (12)</td>
</tr>
<tr>
<td>CAR</td>
<td>447 (21)</td>
<td>384 (15)</td>
</tr>
</tbody>
</table>

Table 3: Total CPU time in days to find a $(0.05, 0.1, 0.02)$-optimal configuration in the synthetic datasets, and the number of configurations before and after precheck. For CAR and CAR++, the number of configurations sampled is reported. CAR++ is the improved version CAR arising from more careful analysis. Error terms are standard deviations over five runs.