A Technical proofs

Findings presented in the body of this paper rely upon a number of supporting results, some of which borrow heavily from earlier works. The main results of this section are as follows. Lemma 4 shows that the posterior variance becomes arbitrarily small as $\mathbf{X}_t$ becomes increasingly dense in $\mathcal{X}$. Lemma 6 upper bounds the expected supremum of $f \sim \mathcal{GP}(0, k)$ based on its maximum variance. Proposition 1 and Corollary 7 prove that the PRB stopping criterion (**??**) almost surely converges. Finally, Proposition 2 shows that Bayesian optimization using PRB terminates and returns an (ϵ, δ)-optimal solution.

Where relevant, we will attribute credit at the beginning of the corresponding proof.

Proposition 3. *Let $\mathcal{X} \subseteq \mathbb{R}^D$ be convex and suppose that $\mathbf{X} \subseteq \mathcal{X}$ generates an ε-cover of $\mathcal{X}$. For every $\boldsymbol{x} \in \mathcal{X}$ and $\rho \geq \varepsilon$, the intersection of the set $\mathbf{X}$ and the ball $B(\boldsymbol{x}, \rho) = \{\boldsymbol{x}' \in \mathcal{X} : \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \rho\}$ generates a 2ε-cover of $B(\boldsymbol{x}, \rho)$.*

Proof. Consider the ball $B(\boldsymbol{x}, r)$ with radius $r = \rho - \varepsilon$. Since $\mathcal{X}$ is convex, for every point $\boldsymbol{a} \in B(\boldsymbol{x}, \rho)$ there exists a $\boldsymbol{b} \in B(\boldsymbol{x}, r)$ such that $\|\boldsymbol{a} - \boldsymbol{b}\|_\infty \leq \varepsilon$. Moreover, because $\mathbf{X}$ generates an ε-cover of $\mathcal{X}$, for every point $\boldsymbol{b} \in B(\boldsymbol{x}, r)$ there exists a $\boldsymbol{c} \in \mathbf{X}$ so that $\|\boldsymbol{b} - \boldsymbol{c}\|_\infty \leq \varepsilon$, which implies that $\boldsymbol{c} \in B(\boldsymbol{x}, \rho)$. It follows by the triangle inequality that for every point $\boldsymbol{a} \in B(\boldsymbol{x}, \rho)$ there exists a pair of points $\boldsymbol{b}, \boldsymbol{c} \in B(\boldsymbol{x}, r) \times [B(\boldsymbol{x}, \rho) \cap \mathbf{X}]$ such that

$$\|\boldsymbol{a} - \boldsymbol{c}\|_\infty \leq \|\boldsymbol{a} - \boldsymbol{b}\|_\infty + \|\boldsymbol{b} - \boldsymbol{c}\|_\infty \leq \varepsilon + \varepsilon = 2\varepsilon, \tag{12}$$

which completes the proof. □

Lemma 4. *Under assumptions A1 and A2, if $y(\cdot) \sim \mathcal{N}\big(f(\cdot), \gamma^2\big)$ is observed on a set of points $\mathbf{X} \subseteq \mathcal{X}$ that generates an ε-cover of $\mathcal{X}$, $0 \leq \varepsilon \leq \min\{1, {}^{k(\boldsymbol{x},\boldsymbol{x})}/_{L_k}\}$, then*

$$\mathrm{Var}[f(\boldsymbol{x}) \mid y(\mathbf{X})] \leq \kappa_\varepsilon(\boldsymbol{x}), \tag{13}$$

where

$$\kappa_\varepsilon(\boldsymbol{x}) = \frac{\big[4L_k\rho(\varepsilon)k(\boldsymbol{x},\boldsymbol{x}) - L_k^2\rho(\varepsilon)^2\big]\eta(\varepsilon) + \gamma^2 k(\boldsymbol{x},\boldsymbol{x})}{[k(\boldsymbol{x},\boldsymbol{x}) + 2L_k\rho(\varepsilon)]\eta(\varepsilon) + \gamma^2}$$

is given in terms of $\eta(\varepsilon) = \max\{1, {}^{\rho(\varepsilon)}/_{4\varepsilon}\}^D$ and $\rho(\varepsilon) = \varepsilon^\varepsilon$ for any $0 < \varepsilon < 1$.

Proof. This result extends Lederer et al. [21, Theorem 3.1], who showed that, for all $0 \leq \rho \leq {}^{k(\boldsymbol{x},\boldsymbol{x})}/_{L_k}$,

$$\mathrm{Var}[f(\boldsymbol{x}) \mid y(\mathbf{B}_\rho(\boldsymbol{x}))] \leq \frac{\big(4L_k\rho k(\boldsymbol{x},\boldsymbol{x}) - L_k^2\rho^2\big)|\mathbf{B}_\rho(\boldsymbol{x})| + \gamma^2 k(\boldsymbol{x},\boldsymbol{x})}{(k(\boldsymbol{x},\boldsymbol{x}) + 2L_k\rho)|\mathbf{B}_\rho(\boldsymbol{x})| + \gamma^2}, \tag{14}$$

where $|\mathbf{B}_\rho(\boldsymbol{x})|$ is the cardinality of the set $\mathbf{B}_\rho(\boldsymbol{x}) = B(\boldsymbol{x}, \rho) \cap \mathbf{X}$. We would like to convert this upper bound into a function of $0 \leq \varepsilon \leq 1$. To this end, begin by noticing that the bound (14) increases monotonically on $0 \leq \rho \leq {}^{k(\boldsymbol{x},\boldsymbol{x})}/_{L_k}$ and decreases monotonically on $n = |\mathbf{B}_\rho(\boldsymbol{x})| \in \mathbb{N}_0$. Substituting $\rho(\varepsilon)$ for ρ and $\eta(\varepsilon)$ for n therefore yields a valid bound so long as $\rho \leq \rho(\varepsilon) \leq {}^{k(\boldsymbol{x},\boldsymbol{x})}/_{L_k}$ and $0 \leq \eta(\varepsilon) \leq n$. For clarity, note that $\rho(\varepsilon)$ defines the radius of a ball around $\boldsymbol{x}$ and $\eta(\varepsilon)$ denotes the minimum possible number of elements from $\mathbf{X}$ that lie within this ball.

Starting with the latter, lower bounds on the cardinality of $\mathbf{B}_\rho(\boldsymbol{x})$ may be obtained from the fact that $\mathbf{X}$ is assumed to generate an ε-cover of $\mathcal{X}$. By Proposition 3, it follows that $\mathbf{B}_\rho(\boldsymbol{x})$ generates a 2ε-cover of $B(\boldsymbol{x}, \rho)$. Accordingly, $|\mathbf{B}_\rho(\boldsymbol{x})|$ must be greater-equal to the minimum number of points required to construct such a cover. Under the $\|\cdot\|_\infty$ norm, the ε-covering number of a ball

$$B(\boldsymbol{x}, \rho) = \prod_{d=1}^{D} [\max(x_d - \rho, 0), \min(x_d + \rho, 1)] \tag{15}$$

is given by

$$M\big(B(\boldsymbol{x}, \rho), \|\cdot\|_\infty, \varepsilon\big) = \prod_{d=1}^{D} \left\lceil \frac{\min(x_d + \rho, 1) - \max(0, x_d - \rho)}{2\varepsilon} \right\rceil. \tag{16}$$

This number is minimized when $B(\cdot, \rho)$ is placed in a corner, such as $B(\mathbf{0}, \rho) = [0, \rho]^D$. Choosing

$$\eta(\varepsilon) = \max\left\{1, \left(\frac{\rho(\varepsilon)}{4\varepsilon}\right)^D\right\} \leq \left\lceil\frac{\rho(\varepsilon)}{4\varepsilon}\right\rceil^D \tag{17}$$

therefore ensures that $\eta(\varepsilon)$ lower bounds the cardinality of every $\mathbf{B}_\rho(\cdot)$. Note that there are two factors of two at play here: one accounts for the fact that $\mathbf{B}_\rho(\cdot)$ is only guaranteed to provide a 2ε-cover of $B(\cdot, \rho)$, and the other accounts for the fact that the corner balls are up to 2^D times smaller than other balls with the same radius.

Turning our attention to the choice of function $\rho(\varepsilon)$, some desiderata come into focus. First, we require $\rho(\varepsilon) \geq \varepsilon$ so that every $\mathbf{B}_\rho(\cdot)$ is nonempty. Second, we desire $\lim_{\varepsilon\to 0^+} \rho(\varepsilon) = 0$ because the resulting posterior variance bound will increase monotonically in $\rho(\varepsilon)$. Lastly, we want the ratio of $\rho(\varepsilon)$ to ε to diverge to infinity as ε approaches zero from above so that $\lim_{\varepsilon\to 0^+} \eta(\varepsilon) = \infty$. Based on these criteria, a convenient choice when $\mathcal{X} = [0, 1]^D$ is

$$\rho(\varepsilon) = \varepsilon^\alpha \qquad\qquad 0 < \alpha < 1. \tag{18}$$

In summary, the claim follows by expressing ρ as a function of ε and using it to lower bound $|\mathbf{B}_\rho(\cdot)|$ with $\eta(\varepsilon)$:

$$\operatorname{Var}[f(\boldsymbol{x}) \mid y(\mathbf{X})] \leq \frac{\big(4L_k\rho(\varepsilon)k(\boldsymbol{x}, \boldsymbol{x}) - L_k^2\rho(\varepsilon)^2\big)\eta(\varepsilon) + \gamma^2 k(\boldsymbol{x}, \boldsymbol{x})}{(k(\boldsymbol{x}, \boldsymbol{x}) + 2L_k\rho(\varepsilon))\eta(\varepsilon) + \gamma^2}. \tag{19}$$

$\square$

Proposition 5. *For any choice of constants $a > 0$, $b \geq 0$, $c \geq 0$,*

$$\int_0^c \sqrt{\log(1 + b\varepsilon^{-1/a})}d\varepsilon \leq c\sqrt{a^{-1} + \log\big(1 + bc^{-1/a}\big)}. \tag{20}$$

Proof. This proof ammends Grünewälder et al. [15, Appendix A]. Let $\xi = (1 + \sqrt[a]{c}b^{-1})^a$ so that

$$\int_0^c \sqrt{\log\big(1 + b\varepsilon^{-1/a}\big)}d\varepsilon \leq \int_0^c \sqrt{\log\big(\xi^{1/a}b\varepsilon^{-1/a}\big)}d\varepsilon. \tag{21}$$

Next, define auxiliary functions

$$f(u) = \sqrt{\log\big(u^{-1/a}\big)} \qquad\qquad g(\varepsilon) = \frac{\varepsilon}{\xi b^a} \tag{22}$$

such that $f(g(\varepsilon)) = \sqrt{\log\big(\xi^{1/a}b\varepsilon^{-1/a}\big)}$, and use them to integrate by substitution as

$$\int_0^c \sqrt{\log\big(\xi^{1/a}b\varepsilon^{-1/a}\big)}d\varepsilon = \xi b^a \int_0^{g(c)} \sqrt{\log\big(u^{-1/a}\big)}du = \frac{\xi b^a}{\sqrt{a}} \int_0^{g(c)} \sqrt{-\log(u)}du. \tag{23}$$

The Cauchy-Schwarz inequality now gives

$$\int_0^{g(c)} \sqrt{-\log(u)}du \leq \left(\int_0^{g(c)} du\right)^{1/2} \left(-\int_0^{g(c)} \log(u)du\right)^{1/2} = \frac{c}{\xi b^a}\sqrt{1 - \log\left(\frac{c}{\xi b^a}\right)}. \tag{24}$$

Hence, the claim follows

$$\int_0^c \sqrt{\log\big(1 + b\varepsilon^{-1/a}\big)}d\varepsilon \leq c\sqrt{\frac{1 + \log(\xi b^a c^{-1})}{a}} = c\sqrt{a^{-1} + \log\big(1 + bc^{-1/a}\big)}. \tag{25}$$

For comparison with Grünewälder et al. [15], when $b^a = 2c$ it follows that

$$\xi = \left(1 + 2^{-1/a}\right)^a \leq 2^a \implies c\sqrt{\frac{1 + \log(\xi b^a c^{-1})}{a}} = c\sqrt{\frac{1 + \log(2\xi)}{a}} \leq c\sqrt{\frac{\log(e2^{a+1})}{a}}. \tag{26}$$

$\square$

Lemma 6. *Let $f \sim \mathcal{GP}(0, k)$ be a Gaussian process with an L_k-Lipschitz continuous covariance function $k : \mathcal{X}^2 \to \mathbb{R}$ on $\mathcal{X} = [0, r]^D$ having maximum variance $\sigma^2 = \max_{\boldsymbol{x} \in \mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x})$. Then,*

$$\mathbb{E}\left[\sup_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})\right] \leq 12\sigma\sqrt{2D + D\log(1 + 4L_k r\sigma^{-2})}. \tag{27}$$

Proof. This proof paraphrases parts of Grünewälder et al. [15, Section 4.3].

Massart [24, Theorem 3.18] proved that the expected supremum of f is upper bounded by

$$\mathbb{E}\left[\sup_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})\right] \leq 12 \int_0^\sigma \sqrt{\log N(\mathcal{X}, d_k, \varepsilon)} d\varepsilon, \tag{28}$$

where $N(\mathcal{X}, d_k, \varepsilon)$ is defined as the ε-packing number—i.e. the largest number of points that can be "packed" inside of $\mathcal{X}$ without any two points being within ε of one another—under the canonical pseudo-metric[3]

$$d_k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\big[(f(\boldsymbol{x}) - f(\boldsymbol{x}'))^2\big]^{1/2} = \sqrt{k(\boldsymbol{x}, \boldsymbol{x}) - 2k(\boldsymbol{x}, \boldsymbol{x}') + k(\boldsymbol{x}', \boldsymbol{x}')}. \tag{29}$$

We may use (28) by upper bounding the right-hand side with a known quantity. We will bound the ϵ-packing number $N(\mathcal{X}, d_k, \varepsilon)$, translate this bound from the d_k pseudo-metric to the infinity norm, and then integrate the result.

The first step follows immediately from the fact that the ε-packing number is smaller than the $\frac{\varepsilon}{2}$-covering number—defined as the minimum number of balls $B(\cdot, \frac{\varepsilon}{2})$ required to cover $\mathcal{X}$. The second is accomplished by using Lipschitz continuity of k to show that the squared pseudo-metric $d_k(\cdot, \cdot)^2$ is $2L_k$-Lipschitz: for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$d_k(\boldsymbol{x}, \boldsymbol{x}')^2 = \big[k(\boldsymbol{x}, \boldsymbol{x}) - k(\boldsymbol{x}, \boldsymbol{x}')\big] + \big[k(\boldsymbol{x}', \boldsymbol{x}') - k(\boldsymbol{x}', \boldsymbol{x})\big] \leq 2L_k \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty. \tag{30}$$

It follows that, for any set $\mathbf{X} \subseteq \mathcal{X}$,

$$\max_{\boldsymbol{x} \in \mathcal{X}} \min_{\boldsymbol{x}' \in \mathbf{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq C \implies \max_{\boldsymbol{x} \in \mathcal{X}} \min_{\boldsymbol{x}' \in \mathbf{X}} d_k(\boldsymbol{x}, \boldsymbol{x}') \leq \sqrt{2L_k C}. \tag{31}$$

An $\varepsilon^2/8L_k$-cover under the infinity norm therefore guarantees an $\varepsilon/2$-cover under d_k. The former may be constructed from a grid of uniformly spaced points with elements at intervals of $\varepsilon^2/4L_k$. This grid will consist of $\lceil 4L_k r\varepsilon^{-2} \rceil^D$ points assuming $\mathcal{X} = [0, r]^D$, meaning that

$$N(\mathcal{X}, d_k, \varepsilon) < \big(1 + 4L_k r\varepsilon^{-2}\big)^D. \tag{32}$$

To complete the proof, use Proposition 5 with $a = \frac{1}{2}$, $b = 4L_k r$, and $c = \sigma$ to show that

$$\mathbb{E}\left[\sup_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})\right] \leq 12\sqrt{D} \int_0^\sigma \sqrt{\log(1 + 4L_k r\varepsilon^{-2})} d\varepsilon \leq 12\sigma\sqrt{2D + D\log(1 + 4L_k r\sigma^{-2})}. \tag{33}$$

□

Proposition 1. *Under assumptions A1–A3 and for all regret bounds $\epsilon > 0$ and risk tolerances $\delta > 0$, there almost surely exists $T \in \mathbb{N}_0$ so that, at each time $t \geq T$, every $\boldsymbol{s}_t \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} \mu_t(\boldsymbol{x})$ satisfies*

$$\Psi_t(\boldsymbol{x}; \epsilon) = \mathbb{P}(r_t(\boldsymbol{s}_t) \leq \epsilon) \geq 1 - \delta. \tag{9}$$

Proof. Consider the centered process

$$g_t(\cdot) = \big[f_t(\cdot) - \mu_t(\cdot)\big] - \big[f_t(\boldsymbol{s}_t) - \mu_t(\boldsymbol{s}_t)\big], \tag{34}$$

with covariance

$$c_t(\boldsymbol{x}, \boldsymbol{x}') = k_t(\boldsymbol{x}, \boldsymbol{x}') - k_t(\boldsymbol{x}, \boldsymbol{s}_t) - k_t(\boldsymbol{s}_t, \boldsymbol{x}') + k_t(\boldsymbol{s}_t, \boldsymbol{s}_t). \tag{35}$$

The term $\mu_t(\boldsymbol{s}_t) - \mu_t(\cdot)$ is nonnegative by construction such that

$$g_t^* = \sup_{\boldsymbol{x} \in \mathcal{X}} g_t(\boldsymbol{x}) \geq f_t^* - f_t(\boldsymbol{s}_t) \qquad f_t^* = \sup_{\boldsymbol{x} \in \mathcal{X}} f_t(\boldsymbol{x}) \tag{36}$$

[3] While d_k has most of the properties of a proper metric, $d_k(\boldsymbol{x}, \boldsymbol{x}') = 0$ does not always imply that $\boldsymbol{x} = \boldsymbol{x}'$ [2].

and, therefore,

$$\mathbb{P}(g_t^* \geq \epsilon) \geq \mathbb{P}(f_t^* - f_t(\boldsymbol{s}_t) \geq \epsilon). \tag{37}$$

We would now like to use the Borell-TIS inequality [8, 38] to show that: if $\epsilon > \mathbb{E}(g_t^*)$, then

$$\mathbb{P}(g_t^* \geq \epsilon) \leq \exp\left(-\frac{1}{2}\left[\frac{\epsilon - \mathbb{E}(g_t^*)}{2\sigma_t}\right]^2\right), \tag{38}$$

where $\sigma_t = \max_{\boldsymbol{x}\in\mathcal{X}} \sqrt{k_t(\boldsymbol{x}, \boldsymbol{x})}$ and $2\sigma_t$ appears in the denominator (rather than σ_t) because $\max_{\boldsymbol{x}\in\mathcal{X}} c_t(\boldsymbol{x}, \boldsymbol{x}) \leq 4\sigma_t^2$. Since (38) is an increasing, continuous function of both $\sigma_t \geq 0$ and $0 \leq \mathbb{E}(g_t^*) < \epsilon$, the claim will hold if these quantities vanish as the (global) fill distance $h_t = \max_{\boldsymbol{x}\in\mathcal{X}} \min_{1\leq i\leq t} \|\boldsymbol{x} - \boldsymbol{x}_i\|_\infty$ goes to zero.

The former result is an immediate consequence of Lemma 4. Regarding the latter, $f_t(\boldsymbol{s}_t) - \mu_t(\boldsymbol{s}_t)$ is a centered random variable. It follows by linearity of expectation that

$$\mathbb{E}(g_t^*) = \mathbb{E}\left[\sup_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x})\right]. \tag{39}$$

Next, denote the canonical pseudo-metric at time t by

$$d_{k_t}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\big[(f_t(\boldsymbol{x}) - f_t(\boldsymbol{x}'))^2\big]^{1/2} = \sqrt{k_t(\boldsymbol{x}, \boldsymbol{x}) - 2k_t(\boldsymbol{x}, \boldsymbol{x}') + k_t(\boldsymbol{x}', \boldsymbol{x}')}. \tag{40}$$

This pseudo-metric is non-increasing in t. To see this, let $\beta = k(\boldsymbol{x}_{t+1}, \boldsymbol{x}) - k(\boldsymbol{x}_{t+1}, \boldsymbol{x}')$ and write

$$\begin{aligned} d_{k_{t+1}}(\boldsymbol{x}, \boldsymbol{x}')^2 &= k_{t+1}(\boldsymbol{x}, \boldsymbol{x}) - 2k_{t+1}(\boldsymbol{x}, \boldsymbol{x}') + k_{t+1}(\boldsymbol{x}', \boldsymbol{x}') \\ &= \underbrace{k_t(\boldsymbol{x}, \boldsymbol{x}) - 2k_t(\boldsymbol{x}, \boldsymbol{x}') + k_t(\boldsymbol{x}', \boldsymbol{x}')}_{d_{k_t}(\boldsymbol{x}, \boldsymbol{x}')^2} - \underbrace{\beta^2\big[k_t(\boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1}) + \gamma^2\big]^{-1}}_{\geq 0}. \end{aligned} \tag{41}$$

As t increases, points therefore become closer together under the d_{k_t} pseudo-metric. For this reason, the posterior ε-packing number $N(\mathcal{X}, d_{k_t}, \varepsilon)$ is less-equal to the prior ε-packing number $N(\mathcal{X}, d_k, \varepsilon)$. By Lemma 6, we now have

$$\begin{aligned} \mathbb{E}(g_t^*) &\leq \int_0^{\sigma_t} \sqrt{\log N(\mathcal{X}, d_{k_t}, \varepsilon)}d\varepsilon \\ &\leq \int_0^{\sigma_t} \sqrt{\log N(\mathcal{X}, d_k, \varepsilon)}d\varepsilon \\ &\leq 12\sigma_t\sqrt{2D + D\log\big(1 + 4L_k\sigma_t^{-2}\big)}. \end{aligned} \tag{42}$$

From here, note that (42) is an increasing, continuous function of σ_t that vanishes as $\sigma_t \to \infty$. By Lemma 4, the same is true of σ_t as a function of h_t. As a result, (38) becomes arbitrarily small as $h_t \to 0$ and there exists a constant $h_* > 0$ such that this upper bound is less-equal to δ whenever $h_t \leq h_*$. Finally, since $(\boldsymbol{x}_t)$ is almost surely dense in $\mathcal{X}$, there almost surely exists a time $T \in \mathbb{N}_0$ such that

$$t \geq T \implies h_t \leq h_* \implies \mathbb{P}(g_t^* > \epsilon) \leq \delta \implies \mathbb{P}(f_t^* - f_t(\boldsymbol{s}_t) > \epsilon) \leq \delta. \tag{43}$$

□

Corollary 7. *Suppose assumptions A1–A3 hold and that there exists a constant $\epsilon' > 0$ so that*

$$\lim_{t\to\infty}\left[\max_{\boldsymbol{x}\in\mathcal{X}} \mu_t(\boldsymbol{x}) - \mu_t(\boldsymbol{s}_t)\right] \leq \epsilon' \tag{44}$$

with probability one, where $\boldsymbol{s}_t \in \arg\max_{\boldsymbol{x}\in\mathbf{X}_t} \mu_t(\boldsymbol{x})$. Then, for every $\epsilon > \epsilon'$ and $\delta \in (0, 1]$, there almost surely exists a time $T \in \mathbb{N}$ such that, for all $t \geq T$,

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x}) - f_t(\boldsymbol{s}_t) \leq \epsilon\right] \geq 1 - \delta. \tag{45}$$

Proof. Per (44), there almost surely exists an $S \in \mathbb{N}$ such that $t \geq S \implies \max_{\boldsymbol{x} \in \mathcal{X}} \mu_t(\boldsymbol{x}) - \mu_t(\boldsymbol{s}_t) \leq \epsilon'$. Proposition 1 therefore implies there almost surely exists a $T \geq S$ so that

$$t \geq T \implies \mathbb{P}\big[f_t^* - f_t(\boldsymbol{s}_t) \geq \epsilon - \epsilon'\big] \leq \delta, \tag{46}$$

which completes the proof. □

The assumption that the posterior mean approaches its maximum on $(\boldsymbol{x}_t)$ protects against adversarial cases where—no matter how densely we observe f—there is always an $\boldsymbol{x} \in \mathcal{X} \setminus \mathbf{X}_t$ whose expected value $\mu_t(\boldsymbol{x})$ exceeds $\mu_t(\boldsymbol{s}_t)$ by at least ϵ. Note that (44) becomes a necessary condition when $\delta < \frac{1}{2}$. Nevertheless, it is unclear how to ensure this condition without making stronger assumptions for f and $(\boldsymbol{x}_t)$. One can use A2 and the Cauchy-Schwarz inequality to show that the posterior mean is Lipschitz continuous [22]; but, its Lipschitz constant may continue to grow as $t \to \infty$, so (44) may not hold.

B Practical recommendations

B.1 Parameter schedules for Algorithm 1

We follow Mnih et al. [26] by setting $d_j = j^{-\alpha} \frac{(\alpha-1)}{\alpha} \delta_{\text{est}}$, where $\alpha = 1.1$. Given an initial sample size $N \in \mathbb{N}$, we similarly define $n_j = \lceil \beta^{j-1} N \rceil$ where $\beta = 1.5$. In our experiments, we set $N = 64$. Using a geometric schedule for (n_j) helps to avoid cases where performing many tests with very few samples causes d_j to become very small. In exchange, this schedule can result in nearly β times too many samples being requested.

B.2 Choosing where to evaluate the stopping rule

If solutions must belong to the set of previously evaluate points $\mathbf{X}_t$, then we may safely ignore any point $\boldsymbol{x} \in \mathbf{X}_t$ for which $\mathbb{P}(f_t(\boldsymbol{s}_t) - f_t(\boldsymbol{x}) \geq \epsilon) \geq \delta_{\text{mod}}$. Empirically, we found that this heuristic usually eliminates all but a few points. If solutions may be chosen freely on $\mathcal{X}$, we instead recommend using a fixed number of joint draws for f_t^* and $\boldsymbol{x}_t^* \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} f_t(\boldsymbol{x})$ to average over

$$\mathbb{P}(r_t(\boldsymbol{x}) \leq \epsilon \mid f_t(\boldsymbol{x}) \leq f_t^*, f_t(\boldsymbol{x}_t^*) = f_t^*) = \Phi\left(\frac{f_t^* - \mu_{t+1}(\boldsymbol{x})}{k_{t+1}(\boldsymbol{x}, \boldsymbol{x})^{1/2}}\right)^{-1} \Phi\left(\frac{\mu_{t+1}(\boldsymbol{x}) - f_t^* - \epsilon}{k_{t+1}(\boldsymbol{x}, \boldsymbol{x})^{1/2}}\right), \tag{47}$$

where μ_{t+1} and k_{t+1} are the posterior mean and variance of f_t given an additional observation $f(\boldsymbol{x}_t^*) = f_t^*$ and Φ denotes the standard normal cumulative distribution function. This estimator can be maximize using gradient information and the resulting points can then be tested with Algorithm 1.

C Experiment Details

Experiments were run using a combination of GPFlow [25] and Trieste [29]. Runtimes reported in Figure 3 were measured on an Apple M1 Pro Chip using an off-the-shelf version of TensorFlow [1].

C.1 Model specification

We employed Gaussian process priors $f \sim \mathcal{GP}(\mu, k)$ with constant mean functions $\mu(\cdot) = c$ and Matérn-$5/2$ covariance functions equipped with ARD lengthscales.

Synthetic When optimizing functions drawn from GP priors, we set the prior mean to zero and used unit variance kernels with lengthscales $\ell_i = \frac{1}{4}\sqrt{D}$. Noise variances are reported alongside results.

Real When optimizing black-box functions, we employed broad and uninformative hyperpriors. Let $[\mathcal{X}]_i = [a_i, b_i]$ be the range of the i-th design variable, $q_t : [0, 1] \to \mathbb{R}$ be the empirical quantile function of y at time t, and $\nu_t = \overline{\text{Var}}\big[\boldsymbol{y}_{t-1}\big]$ be the empirical variance of observations $\boldsymbol{y}_{t-1} = \{y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_{t-1})\}$. Our hyperpriors are then as follows:

Name	Distribution	Parameters	
Constant Mean	$\mathrm{Uniform}(a, b)$	$a = q_t(0.05)$	$b = q_t(0.95)$
Log Kernel Variance	$\mathrm{Uniform}(a, b)$	$a = \log(10^{-1}\nu_t)$	$b = \log(10\nu_t)$
Log Noise Variance	$\mathrm{Uniform}(a, b)$	$a = \log(10^{-9}\nu_t)$	$b = \log(10\nu_t)$
i-th Lengthscale	$\mathrm{LogNormal}(\mu, \sigma)$	$\mu = \frac{1}{2}(b_i - a_i)$	$\sigma = 1$

Note that we directly parameterize certain hyperparamters in log-space and that, e.g., $\log(\theta) \sim \mathrm{Uniform}(a, b)$ is not the same as $\theta \sim \mathrm{LogUniform}(e^a, e^b)$.

C.2 Acquisition function

In our experiments, we defined the set of feasible solutions at time $t \in \mathbb{N}$ as the set of previously evaluated points $\mathbf{X}_t$. Under these circumstances, one can show that the optimal one-step policy is given by an "in-sample" version of the Knowledge Gradient strategy [27, 12]. Let

$$\mu_{t+1}(\cdot; \boldsymbol{x}, z) = \mu_t(\cdot) + \frac{k_t(\cdot, \boldsymbol{x})z}{\sqrt{k_t(\boldsymbol{x}, \boldsymbol{x}) + \gamma^2}} \tag{48}$$

be the posterior mean of f at time $t+1$ if we observe $y_{t+1} = \mu_t(\boldsymbol{x}) + \sqrt{k_t(\boldsymbol{x}, \boldsymbol{x}) + \gamma^2}z$, where $z \sim \mathcal{N}(0, 1)$. Then, the aforementioned acquisition function is given by

$$\mathrm{ISKG}_t(\boldsymbol{x}) = \mathbb{E}_z\Big[\max \mu_{t+1}(\mathbf{X}_t \cup \{\boldsymbol{x}\}; \boldsymbol{x}, z)\Big] - \max \mu_t(\mathbf{X}_t) \qquad z \sim \mathcal{N}(0, 1). \tag{49}$$

ISKG is identical to the Expected Improvement function when $\gamma^2 = 0$ [32], but avoids pathologies (such as re-evaluated previously observed points) when $\gamma^2 > 0$.

In practice, we estimated (49) with Gauss-Hermite quadrature and maximized it using multi-start gradient ascent [41, 5]. Starting positions we obtained by running CMA-ES [16] several times to partial convergence. The best point from each run was then combined with a large number of random points and the top 16 points were fine-tuned using L-BFGS-B [9].

C.3 Link function

When modeling classification rates for MNIST and Adult, we used a logit (i.e. inverse sigmoid) link function,

$$g(y) = \log\left(\frac{y}{1-y}\right) \qquad\qquad g^{-1}(x) = \frac{1}{1+e^{-x}}, \tag{50}$$

in order so that $g^{-1} \circ f : \mathcal{X} \to [0, 1]$. When evaluating stopping rules, we handled this link functions by pulling draws of, e.g., $f_t(\boldsymbol{x})$ backward through g and using te resulting values to estimate expectations and probabilities. This approach was used for all but ΔCB [23], where we instead computed $g^{-1} \circ \mathrm{UCB}_t$ and $g^{-1} \circ \mathrm{LCB}_t$.

C.4 Convolutional neural networks

When training convolutional neural networks (CNNs) on MNIST [11], we used a simple architecture consisting of two convolutional layers with 3×3 filters and ReLU activation functions [3] followed by max pooling layers with a poolsize of 2. The output of the final pooling layer was flattened and subjected to dropout before being passed to a dense classification layer consisting of ten neurons. Each model was trained using Adam [19], with batches of size 64. The search space for this problem was four-dimensional as seen on the right. Integer valued parameters were handled by rounding to the nearest value.

Name	Low	High
Number of filters	1	64
Number of epochs	1	25
Log learning rate	$\log(10^{-5})$	0
Dropout rate	0	1

C.5 XGBoost Classifiers

We used an off-the-shelf implementation of XGBoost [10] for the the adult income classification problem [7]. The search space was three-dimensional and is shown on the right. Integer valued parameters were handled by rounding to the nearest value.

Name	**Low**	**High**
Maximum tree depth	1	10
Log number of estimators	0	$\log(10^3)$
Log learning rate	$\log(10^{-3})$	0

D Extended Results

D.1 Results without adjusted cutoff values

Problem	D	T	**Oracle**†	**Budget**†	**Acq**	**ΔCB**	**ΔES**	**PRB**
GP† 10^{-6}	2	64	10 (100)	17 (96)	28 (100)	12 (89)	14 (94)	**17 (97)**
GP† 10^{-2}	2	128	11 (100)	22 (96)	78 (100)	82 (100)	18 (91)	**23 (99)**
GP† 10^{-6}	4	128	27 (100)	64 (95)	90 (100)	23 (66)	28 (74)	**64 (99)**
GP† 10^{-2}	4	256	30 (100)	94 (95)	106 (98)	256 (100)	36 (65)	**86 (96)**
GP† 10^{-6}	6	256	40 (99)	124 (95)	142 (98)	31 (50)	46 (65)	**134 (98)**
GP† 10^{-2}	6	512	65 (100)	227 (96)	**181 (96)**	512 (100)	45 (34)	235 (100)
GP 10^{-6}	4	128	35 (100)	79 (95)	**92 (100)**	18 (30)	22 (41)	61 (88)
GP 10^{-2}	4	256	51 (100)	157 (95)	**128 (97)**	224 (80)	27 (22)	100 (92)
Branin	2	128	19 (100)	25 (95)	64 (100)	**31 (99)**	32 (100)	33 (99)
Hartmann	3	64	14 (100)	22 (96)	26 (100)	15 (83)	17 (84)	**19 (100)**
Hartmann	6	64	36 (67)	256 (67)	40 (67)	26 (46)	30 (56)	40 (64)
Rosenbrock	4	96	34 (100)	46 (95)	95 (100)	**68 (99)**	71 (100)	84 (100)
CNN	4	256	5 (100)	11 (96)	64 (100)	8 (92)	14 (94)	**17 (100)**
XGBoost	3	128	4 (100)	8 (97)	128 (100)	**16 (97)**	19 (99)	28 (99)

Table 2: Same as Table 1, but where ϵ is used as the cutoff value for ΔCB and ΔES.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are either discussed in the main or in Appendix A.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The primary limitation of the proposed is that it relies on an underlying model being well-calibrated. This issue is clearly discussed at prominent locations in the text, such as the introduction and experiments section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are clearly stated throughout the paper (e.g., in Section 4) and detailed proofs are provided in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Details are provided in the text or in supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: No

Justification: Code will be made available as part of the final release of this work, but was not ready in time for the initial submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Details are provided in the text or in supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Quantiles for key metrics are reported in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: No.

   Justification: We ran thousands of experiments on heterogenous hardware at different points in time and did not keep track.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This question is not particular relevant to our submission, since our focus is on making existing optimization algorithms, e.g., more cost-efficient.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This question is not particular relevant to our submission, since our focus is on making existing optimization algorithms, e.g., more cost-efficient.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The proposed methods do not lend themselves to the this type of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The content of this paper was either created by the authors for use herein. Borrowed material has been cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Noe assets have been released at this time.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Not revelent.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not relevnat.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.