Supplementary material

A. Proofs from Section 2

A.1. Proof of Proposition 1

Proposition 1. *Let f be L-smooth, and let y_0^t and x_0^t be the sequence of iterates generated by* AGD-UNTIL-GUILTY*$(f, y_0, L, \varepsilon, \sigma)$ for some $\varepsilon > 0$ and $0 < \sigma \leq L$. Fix $w \in \mathbb{R}^d$. If for $s = 0, 1, \ldots, t-1$ we have*

$$f(u) \geq f(x_s) + \nabla f(x_s)^T (u - x_s) + \frac{\sigma}{2}\|u - x_s\|^2 \tag{5}$$

for both $u = w$ and $u = y_s$, then

$$f(y_t) - f(w) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \psi(w), \tag{6}$$

where $\kappa = \frac{L}{\sigma}$ and $\psi(w) = f(y_0) - f(w) + \frac{\sigma}{2}\|w - y_0\|^2$.

Proof. The proof is closely based on the proof of Theorem 3.18 of (Bubeck, 2014), which itself is based on the estimate sequence technique of Nesterov (2004). We modify the proof slightly to avoid arguments that depend on the global minimum of f. This enables using inequalities (5) to prove the result, instead of σ-strong convexity of the function f.

We define σ-strongly convex quadratic functions Φ_s by induction as

$$\Phi_0(z) = f(x_0) + \frac{\sigma}{2}\|z - x_0\|^2,$$

and, for $s = 0, ..., t-1$,

$$\Phi_{s+1}(z) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s(z) + \frac{1}{\sqrt{\kappa}}\left(f(x_s) + \nabla f(x_s)^T(z - x_s) + \frac{\sigma}{2}\|z - x_s\|^2\right). \tag{20}$$

Using (5) with $u = w$, straightforward induction shows that

$$\Phi_s(w) \leq f(w) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^s \psi(w) \text{ for } s = 0, 1, ..., t. \tag{21}$$

Let $\Phi_s^* = \min_{x \in \mathbb{R}^n} \Phi_s(x)$. If

$$f(y_s) \leq \Phi_s^* \text{ for } s = 0, 1, ..., t \tag{22}$$

then (6) follows immediately, since

$$f(y_t) - f(w) \leq \Phi_t^* - f(w) \leq \Phi_t(w) - f(w) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \psi(w)$$

We now prove (22) by induction. Note that it is true at $s = 0$ since $x_0 = y_0$ is the global minimizer of Φ_0. We have,

$$\begin{aligned}
f(y_{s+1}) &\overset{(a)}{\leq} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|^2 \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right) f(y_s) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(f(x_s) - f(y_s)) + \frac{1}{\sqrt{\kappa}} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|^2 \\
&\overset{(b)}{\leq} \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(f(x_s) - f(y_s)) + \frac{1}{\sqrt{\kappa}} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|^2 \\
&\overset{(c)}{\leq} \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(x_s)^T(x_s - y_s) + \frac{1}{\sqrt{\kappa}} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|^2,
\end{aligned}$$

where inequality (a) follows from the definition $y_{s+1} = x_s - \frac{1}{L}\nabla f(x_s)$ and the L-smoothness of f, inequality (b) is the induction hypothesis and inequality (c) is assumption (5) with $u = y_s$.

Past this point the proof is identical to the proof of Theorem 3.18 of (Bubeck, 2014), but we continue for sake of completeness.

To complete the induction argument we now need to show that:

$$\left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(x_s)^T(x_s - y_s) + \frac{1}{\sqrt{\kappa}}f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|^2 \leq \Phi_{s+1}^*. \tag{23}$$

Note that $\nabla^2\Phi_s = \sigma I_n$ (immediate by induction) and therefore

$$\Phi_s(x) = \Phi_s^* + \frac{\sigma}{2}\|x - v_s\|^2,$$

for some $v_s \in \mathbb{R}^n$. By differentiating (20) and using the above form of Φ_s we obtain

$$\nabla\Phi_{s+1}(x) = \sigma\left(1 - \frac{1}{\sqrt{\kappa}}\right)(x - v_s) + \frac{1}{\sqrt{\kappa}}\nabla f(x_s) + \frac{\sigma}{\sqrt{\kappa}}(x - x_s).$$

Since by definition $\Phi_{s+1}(v_{s+1}) = 0$, we have

$$v_{s+1} = \left(1 - \frac{1}{\sqrt{\kappa}}\right)v_s + \frac{1}{\sqrt{\kappa}}x_s - \frac{1}{\sigma\sqrt{\kappa}}\nabla f(x_s) \tag{24}$$

Using (20), evaluating evaluating Φ_{s+1} at x_s gives,

$$\Phi_{s+1}(x_s) = \Phi_{s+1}^* + \frac{\sigma}{2}\|x_s - v_{s+1}\|^2 = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left[\Phi_s^* + \frac{\sigma}{2}\|x_s - v_s\|^2\right] + \frac{1}{\sqrt{\kappa}}f(x_s). \tag{25}$$

Substituting (24) gives

$$\|x_s - v_{s+1}\|^2 = \left(1 - \frac{1}{\sqrt{\kappa}}\right)^2\|x_s - v_s\|^2 + \frac{1}{\sigma^2\kappa}\|\nabla f(x_s)\|^2 - \frac{2}{\sigma\sqrt{\kappa}}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(x_s)^T(v_s - x_s)$$

which combined with (25) yields

$$\Phi_{s+1}^* = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s^* + \frac{1}{\sqrt{\kappa}}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(x_s)^T(v_s - x_s) + \frac{1}{\sqrt{\kappa}}f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|^2 + \frac{\sigma}{2\sqrt{\kappa}}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\|x_s - v_s\|^2.$$

Examining this equation, it is seen that $v_s - x_s = \sqrt{\kappa}(x_s - y_s)$ implies (23) and therefore concludes the proof of Proposition 1. We establish the relation $v_s - x_s = \sqrt{\kappa}(x_s - y_s)$ by induction,

$$\begin{aligned} v_{s+1} - x_{s+1} &= \left(1 - \frac{1}{\sqrt{\kappa}}\right)v_s + \frac{1}{\sqrt{\kappa}}x_s - \frac{1}{\sigma\sqrt{\kappa}}\nabla f(x_s) - x_{s+1} \\ &= \sqrt{\kappa}x_s - (\sqrt{\kappa} - 1)y_s - \frac{\sqrt{\kappa}}{L}\nabla f(x_s) - x_{s+1} \\ &= \sqrt{\kappa}y_{s+1} - (\sqrt{\kappa} - 1)y_s - x_{s+1} = \sqrt{\kappa}(x_{s+1} - y_{s+1}). \end{aligned}$$

where the first equality comes from (24), the second from the induction hypothesis, the third from the definition of y_{s+1} and the last one from the definition of x_{s+1}. □

A.2. Proof of Lemma 1

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ have L_2-Lipschitz Hessian. Let $\alpha > 0$ and let u and v satisfy (10). If $\|u - v\| \leq \frac{\alpha}{2L_2}$, then for every $\eta \leq \frac{\alpha}{L_2}$,* EXPLOIT-NC-PAIR*$(f, u, v, \eta)$ finds a point z such that*

$$f(z) \leq f(u) - \frac{\alpha\eta^2}{12}. \tag{11}$$

Proof. We proceed in two parts; in the first part, we show that f has negative curvature of at least $\alpha/2$ in the direction of $u - v$ at the point u. In the second part we show that such negative curvature guarantees that a step with magnitude η produces the required progress in function value.

For $0 \le \theta \le \|u - v\|$, let

$$\lambda(\theta) = v + \theta\delta, \text{ where } \delta = \frac{u - v}{\|u - v\|}.$$

Then,

$$\int_0^{\|u-v\|} d\tau \int_0^{\tau} \left[\delta^T \nabla^2 f(\lambda(\theta))\delta\right] d\theta = f(u) - f(v) - \nabla f(v)^T (u - v) < -\frac{\alpha}{2}\|u - v\|^2,$$

where the equality follows from basic calculus, and the inequality is assumption (10). Substituting $\kappa = \min_{0\le\theta\le\|u-v\|} \left\{\delta^T \nabla^2 f(\lambda(\theta))\delta\right\}$ for the integrand, we find that $\kappa < -\alpha$.

By Lipschitz continuity of $\nabla^2 f$ and $\|u - v\| \le \alpha/(2L_2)$ we thus have,

$$\delta^T \nabla^2 f(u)\delta \le \kappa + L_2\|u - v\| < -\alpha + L_2\alpha/(2L_2) \le -\frac{\alpha}{2}, \tag{26}$$

which concludes the first part of the proof.

The Lipschitz continuity of $\nabla^2 f$ also implies that it is upper bounded by its quadratic approximation with a cubic residual term, *i.e.*

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + \frac{L_2}{6}\|y - x\|^3$$

for any $y, x \in \mathbb{R}^d$. Applying this to $u_\pm = u \pm \eta\delta$, we have

$$f(u_\pm) \le f(u) \pm \eta \nabla f(u)^T \delta + \frac{\eta^2}{2}\delta^T \nabla^2 f(u)\delta + \frac{L_2}{6}|\eta|^3.$$

We note that the first order term must be negative for either u_+ or u_-. Therefore, using (26) and $\eta \le \alpha/L_2$, we have that

$$f(z) = \min\{f(u_+), f(u_-)\} \le f(u) - \frac{\alpha\eta^2}{12}.$$

□

B. Proofs from Section 3

B.1. Proof of Lemma 2

Lemma 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L_1-smooth and have L_2-Lipschitz continuous Hessian, let $\epsilon, \alpha > 0$ and $p_0 \in \mathbb{R}^d$. Let $p_1, \ldots, p_K$ be the iterates* GUARDED-NON-CONVEX-AGD*$(f, p_0, L_1, \epsilon, \alpha, \frac{\alpha}{L_2})$ generates. Then for each $k \in \{1, \ldots, K-1\}$,*

$$f(p_k) \le f(p_{k-1}) - \min\left\{\frac{\epsilon^2}{5\alpha}, \frac{\alpha^3}{64L_2^2}\right\}. \tag{12}$$

Proof. Fix an iterate index $1 \le k < K$; throughout the proof we let y_0^t, x_0^t and u, v refer to outputs of AGD-UNTIL-GUILTY in the kth iteration. We consider the cases $u, v = \text{NULL}$ and $u, v \ne \text{NULL}$ separately.

The simpler case is $u, v = \text{NULL}$ (no convexity violation detected), in which $p_k = y_t$ and $\|\nabla\hat{f}(p_k)\| \le \epsilon/10$ (since AGD-UNTIL-GUILTY terminated on line 9). Moreover, $k < K$ implies that GUARDED-NON-CONVEX-AGD does not terminate at iteration k, and therefore $\|\nabla f(p_k)\| > \epsilon$. Consequently,

$$2\alpha\|p_k - p_{k-1}\| = \|\nabla f(p_k) - \nabla\hat{f}(p_k)\| \ge \|\nabla f(p_k)\| - \|\nabla\hat{f}(p_k)\| \ge 9\epsilon/10.$$

The case $u, v = \text{NULL}$ also implies $\hat{f}(p_k) = \hat{f}(y_t) \le \hat{f}(y_0) = f(p_{k-1})$, as the condition in line 2 of CERTIFY-PROGRESS never holds, and therefore

$$f(p_k) = \hat{f}(p_k) - \alpha \|p_k - p_{k-1}\|^2 \le f(p_{k-1}) - \alpha \left(\frac{9\epsilon}{20\alpha}\right)^2 \le f(p_{k-1}) - \frac{\epsilon^2}{5\alpha},$$

which establishes the claim in the case $u, v = \text{NULL}$.

Next we consider the case $u, v \neq \text{NULL}$ (non-convexity detected). By Corollary 1,

$$\hat{f}(u) < \hat{f}(v) + \nabla \hat{f}(v)^T (u - v) + \frac{\alpha}{2} \|u - v\|^2 \tag{27}$$

By definition of $\hat{f}(x) = f(x) + \frac{\alpha}{2} \|x - y_0\|^2$, we have that,

$$f(v) + \nabla f(v)^T (u - v) - \frac{\alpha}{2} \|u - v\|^2 - f(u) = \hat{f}(v) + \nabla \hat{f}(v)^T (u - v) + \frac{\alpha}{2} \|u - v\|^2 - \hat{f}(u) > 0$$

for every $u, v \in \mathbb{R}^d$. Therefore, we conclude that (10) must hold.

We now set $\tau := \frac{\alpha}{8L_2}$ and consider two cases. First, if $f(b^{(1)}) \le f(y_0) - \alpha\tau^2 = f(p_{k-1}) - \frac{\alpha^3}{64L_2^2}$ then we are done, since $f(p_k) \le f(b^{(1)})$. Second, if $f(b^{(1)}) \ge f(y_0) - \alpha\tau^2$, by Lemma 3 we have that

$$\|v - u\| \le 4\tau \le \frac{\alpha}{2L_2}.$$

Therefore, we can use Lemma 1 (with $\eta = \frac{\alpha}{L_2}$) to show that

$$f(b^{(2)}) \le f(u) - \frac{\alpha^3}{12L_2^2} \le f(p_{k-1}) - \frac{\alpha^3}{12L_2^2},$$

where the last transition uses again $f(u) \le \hat{f}(u) \le \hat{f}(y_0) = f(p_{k-1})$, due to Corollary 1. This implies (12) holds and concludes the case $u, v \neq \text{NULL}$. □

B.2. Proof of Lemma 3

Lemma 3. *Let f be L_1-smooth, and $\tau \ge 0$. At any iteration of* GUARDED-NON-CONVEX-AGD*, if $u, v \neq \text{NULL}$ and the best iterate $b^{(1)}$ satisfies $f(b^{(1)}) \ge f(y_0) - \alpha\tau^2$ then for $1 \le i < t$,*

$$\|y_i - y_0\| \le \tau, \quad \textit{and} \quad \|x_i - y_0\| \le 3\tau.$$

Consequently, $\|u - v\| \le 4\tau$.

Proof. We begin by noting that $\hat{f}(y_i) \le \hat{f}(y_0) = f(y_0)$ for $i = 1, ..., t-1$, as guaranteed by Corollary 1. Using $f(y_i) \ge f(b^{(1)}) \ge f(y_0) - \alpha\tau^2$ we therefore have

$$\alpha \|y_i - y_0\|^2 = \hat{f}(y_i) - f(y_i) \le f(y_0) - f(y_i) \le \alpha\tau^2,$$

which implies $\|y_i - y_0\| \le \tau$. Since by Corollary 1 we also have $\hat{f}(u) \le \hat{f}(y_0)$, we similarly obtain $\|u - y_0\| \le \tau$.

Using $x_i = (1 + \omega)y_i - \omega y_{i-1}$ we have, by the triangle inequality and $0 \le \omega < 1$,

$$\|x_i - y_0\| \le (1 + \omega) \|y_i - y_0\| + \omega \|y_{i-1} - y_0\| \le 3\tau$$

for every $i = 1, ..., t-1$. Finally, since $v = x_j$ for some $0 \le j \le t-1$, this gives

$$\|u - v\| \le \|u - y_0\| + \|x_j - y_0\| \le 4\tau.$$

□

B.3. Proof of Theorem 1

Theorem 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L_1-smooth and have L_2-Lipschitz continuous Hessian. Let $p_0 \in \mathbb{R}^d$, $\Delta_f = f(p_0) - \inf_{z\in\mathbb{R}^d} f(z)$ and $0 < \epsilon \le \min\{\Delta_f^{2/3} L_2^{1/3}, L_1^2/(64L_2)\}$. Set*

$$\alpha = 2\sqrt{L_2\epsilon} \tag{13}$$

then GUARDED-NON-CONVEX-AGD*(f, p_0, L_1, ϵ, α, $\frac{\alpha}{L_2}$) finds a point p_K such that $\|\nabla f(p_K)\| \le \epsilon$ with at most*

$$20 \cdot \frac{\Delta_f L_1^{1/2} L_2^{1/4}}{\epsilon^{7/4}} \log \frac{500 L_1 \Delta_f}{\epsilon^2} \tag{14}$$

gradient evaluations.

Proof. We bound two quantities: the number of calls to AGD-UNTIL-GUILTY, which we denote by K, and the maximum number of steps AGD-UNTIL-GUILTY performs when it is called, which we denote by T. The overall number gradient evaluations is $2KT$, as we compute at most $2T$ gradients per iterations (at the points $x_0, \dots, x_{t-1}$ and $y_1, \dots, y_t$).

The upper bound on K is immediate from Lemma 2, as by telescoping the progress guarantee (12) we obtain

$$\Delta_f \ge f(p_0) - f(p_{K-1}) = \sum_{k=1}^{K-1} (f(p_{k-1}) - f(p_k)) \ge (K-1) \cdot \min\left\{\frac{\epsilon^2}{5\alpha}, \frac{\alpha^3}{64 L_2^2}\right\} \ge (K-1)\frac{\epsilon^{3/2}}{10 L_2^{1/2}},$$

where the final inequality follows by substituting our choice (13), of α. We conclude that

$$K \le 1 + 10 \Delta_f L_2^{1/2} \epsilon^{-3/2}. \tag{28}$$

To bound the number T of steps of AGD-UNTIL-GUILTY, note that for every $z \in \mathbb{R}^d$

$$\psi(z) = \hat{f}(y_0) - \hat{f}(z) + \frac{\alpha}{2}\|z - y_0\|^2 = f(y_0) - f(z) - \frac{\alpha}{2}\|z - y_0\|^2 \le \Delta_f.$$

Therefore, substituting $\varepsilon = \epsilon/10$, $L = L_1 + 2\alpha$ and $L = \alpha = 2\sqrt{L_2\epsilon}$ into the guarantee (7) of Corollary 1,

$$T \le 1 + \sqrt{2 + \frac{L_1}{2\sqrt{L_2\epsilon}}} \log_+ \left(\frac{200(L_1 + 4\sqrt{L_2\epsilon})\Delta_f}{\epsilon^2}\right). \tag{29}$$

We use $\epsilon \le \min\{\Delta_f^{2/3} L_2^{1/3}, L_1^2/(12L_2)\}$ to simplify the bounds on K and T. Using $1 \le \Delta_f L_2^{1/2} \epsilon^{-3/2}$ simplifies the bound (28) to

$$K \le 11 \Delta_f L_2^{1/2} \epsilon^{-3/2}.$$

Applying $1 \le L_1/(8\sqrt{L_2\epsilon})$ in the bound (29) gives

$$T \le \sqrt{\frac{3}{4}} \frac{L_1^{1/2}}{L_2^{1/4}\epsilon^{1/4}} \log\left(\frac{500 L_1 \Delta_f}{\epsilon^2}\right),$$

where $\Delta_f L_1 \epsilon^{-2} \ge 8$ allows us to drop the subscript from the log. Multiplying the product of the above bounds by 2 gives the theorem. □

C. Proofs from Section 4

C.1. Proof of Lemma 5

We begin by proving the following normalized version of Lemma 5.

Lemma 8. *Let $h : \mathbb{R} \to \mathbb{R}$ be thrice differentiable, h''' be L-Lipschitz continuous for some $L > 0$ and let*

$$h(1) - h(-1) - 2h'(-1) = \int_{-1}^{1} d\nu \int_{-1}^{\nu} h''(\xi) d\xi \leq -A, \tag{30}$$

for some $A \geq 0$. Then for any $\rho \geq 4$

$$\min\{h(-1-\rho), h(1+\rho')\} \leq \max\left\{h(-1) - \frac{A}{4}\rho^2, h(1) - \frac{A}{6}\rho^2\right\} + \frac{L}{8}\rho^4, \tag{31}$$

where $\rho' = \sqrt{\rho(\rho+2)} - 2$.

Proof. Define

$$\tilde{h}(\xi) = h(0) + h'(0)\xi + \frac{1}{2}h''(0)\xi^2 + \frac{1}{6}h'''(0)\xi^3.$$

By the Lipschitz continuity of h''', we have that $|h(\xi) - \tilde{h}(\xi)| \leq L\xi^4/24$ for any $\xi \in \mathbb{R}$ (see Section 1.4). Similarly, viewing h''' as the first derivative of h'', we have and $|h''(\xi) - \tilde{h}''(\xi)| \leq L\xi^2/2$. The assumption (30) therefore implies,

$$4\left[\frac{1}{2}h''(0) - \frac{1}{6}h'''(0)\right] = \int_{-1}^{1} d\nu \int_{-1}^{\nu} \tilde{h}''(\xi) d\xi \leq -A + \frac{L}{2}\int_{-1}^{1} d\nu \int_{-1}^{\nu} \xi^2 d\xi = -A + \frac{1}{3}L. \tag{32}$$

It is also easy to verify that

$$h(0) = \frac{\tilde{h}(1) + \tilde{h}(-1)}{2} - \frac{1}{2}h''(0) \text{ and } h'(0) = \frac{\tilde{h}(1) - \tilde{h}(-1)}{2} - \frac{1}{6}h'''(0).$$

Substituting into the definition of $\tilde{h}$ and rearranging, this yields

$$\tilde{h}(1+\rho') = \tilde{h}(1) + \frac{\tilde{h}(1) - \tilde{h}(-1)}{2}\rho' + \left[\frac{1}{2}h''(0) - \frac{1}{6}h'''(0)\right]\rho'(2+\rho') + \frac{1}{6}h'''(0)\rho'(2+\rho')^2 \tag{33}$$

and

$$\tilde{h}(-1-\rho) = \tilde{h}(-1) - \frac{\tilde{h}(1) - \tilde{h}(-1)}{2}\rho + \left[\frac{1}{2}h''(0) - \frac{1}{6}h'''(0)\right]\rho(2+\rho) - \frac{1}{6}h'''(0)\rho^2(2+\rho). \tag{34}$$

Suppose that $\frac{\tilde{h}(1)-\tilde{h}(-1)}{2} + \frac{1}{6}h'''(0)\rho(2+\rho) \geq 0$. By (34), (32) and $\rho^2 \leq \rho(2+\rho) \leq 3\rho^2/2$ (since $\rho \geq 4$) we then have

$$\tilde{h}(-1-\rho) \leq \tilde{h}(-1) - \frac{A}{4}\rho^2 + \frac{L}{8}\rho^2,$$

and, using $|h(\xi) - \tilde{h}(\xi)| \leq L\xi^4/24$ for $\xi = -1$ and $\xi = -1-\rho$ along with $1 \leq \rho/4$, we get

$$h(-1-\rho) \leq h(-1) - \frac{A}{4}\rho^2 + \frac{L}{8}\rho^2 + \frac{L}{24}(1 + (1+\rho)^4) \leq h(-1) - \frac{A}{4}\rho^2 + \frac{L}{8}\rho^4. \tag{35}$$

Suppose now that $\frac{\tilde{h}(1)-\tilde{h}(-1)}{2} + \frac{1}{6}h'''(0)\rho(2+\rho) < 0$ holds instead. By (33) and (32) we then have

$$\tilde{h}(1+\rho') \leq \tilde{h}(1) - \left[\frac{A}{4} - \frac{L}{12}\right]\rho'(2+\rho') + \frac{1}{6}h'''(0)\rho'\left[(2+\rho')^2 - \rho(2+\rho)\right] = \tilde{h}(1) - \left[\frac{A}{4} - \frac{L}{12}\right]\rho'(2+\rho')$$

where the equality follows from the definition $(2+\rho')^2 = \rho(2+\rho)$. We lower bound $\rho'(2+\rho')$ as

$$\rho'(2+\rho') = \rho(2+\rho) - 2\sqrt{\rho(2+\rho)} \geq \rho\left(\frac{\rho}{2} + \rho\right) - \frac{\rho}{2}\sqrt{\rho\left(\frac{\rho}{2} + \rho\right)} \geq \frac{2\rho^2}{3},$$

where the first inequality follows from the fact that $\rho(\zeta+\rho) - \zeta\sqrt{\rho(\zeta+\rho)}$ is monotonically decreasing in $\zeta \geq 0$ and the assumption $2 \leq \rho/2$. Noting that $\rho' \leq \rho$, we have the upper bound $\rho'(2+\rho') \leq \rho(2+\rho) \leq 3\rho^2/2$. Combining these

bounds gives $\tilde{h}(1+\rho') \le \tilde{h}(1) - \frac{A}{6}\rho^2 + \frac{L}{8}\rho^2$. Applying $|h(\xi) - \tilde{h}(\xi)| \le L\xi^4/24$ at $\xi = 1$ and $\xi = 1+\rho'$, and using $\rho' \le \rho$ and $1 \le \rho/4$ once more, we obtain,

$$h(1+\rho') \le h(1) - \frac{A}{6}\rho^2 + \frac{L}{8}\rho^2 + \frac{L}{24}(1+(1+\rho)^4) \le h(1) - \frac{A}{6}\rho^2 + \frac{L}{8}\rho^4. \tag{36}$$

The fact that either (35) or (36) must hold implies (31).

□

With the auxiliary Lemma 8, we prove Lemma 5.

Lemma 5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ have L_3-Lipschitz third-order derivatives. Let $\alpha > 0$ and let u and v satisfy (10) and let $\eta \le \sqrt{2\alpha/L_3}$. Then for every $\|u - v\| \le \eta/2$, EXPLOIT-NC-PAIR$_3(f, u, v, \eta)$ finds a point z such that*

$$f(z) \le \max\left\{f(v) - \frac{\alpha}{4}\eta^2, f(u) - \frac{\alpha}{12}\eta^2\right\}. \tag{16}$$

Proof. Define

$$h(\theta) := f\left(\frac{1+\theta}{2}u + \frac{1-\theta}{2}v\right).$$

We have

$$h(1) - h(-1) - 2h'(-1) = f(u) - f(v) - \nabla f(v)^T(u-v) < -\frac{\alpha}{2}\|u-v\|^2 := -A.$$

Additionally, since f has L_3-Lipschitz third order derivatives, h''' is $\frac{1}{16}L_3\|u-v\|^4 := L$ Lipschitz continuous, so we may apply Lemma 8 at $\rho = 2\eta/\|u-v\| \ge 4$. Letting $\delta = (u-v)/\|u-v\|$, we note that $h(1-\rho) = f(v-\eta\delta)$. Similarly, for $2+\rho' = \sqrt{\rho(2+\rho)}$ we have $h(1+\rho') = f(u+\eta'\delta)$ with η' given in line 2 of EXPLOIT-NC-PAIR$_3$. The result is now immediate from (31), as

$$\begin{aligned} f(z) &= \min\{f(v-\eta\delta), f(u+\eta'\delta)\} = \min\{h(-1-\rho), h(1+\rho')\} \le \max\left\{h(-1) - \frac{A}{4}\rho^2, h(1) - \frac{A}{6}\rho^2\right\} + \frac{L}{8}\rho^4 \\ &= \max\left\{f(v) - \frac{\alpha}{2}\eta^2, f(u) - \frac{\alpha}{3}\eta^2\right\} + \frac{L_3}{8}\eta^4 \le \max\left\{f(v) - \frac{\alpha}{4}\eta^2, f(u) - \frac{\alpha}{12}\eta^2\right\}, \end{aligned}$$

where in the last transition we have used $\eta \le \sqrt{\frac{2\alpha}{L_3}}$. □

C.2. Proof of Lemma 6

We first state and prove a normalized version of the central argument in the proof of Lemma 6

Lemma 9. *Let $h : \mathbb{R} \to \mathbb{R}$ be thrice differentiable and let h''' be L-Lipschitz continuous for some $L > 0$. If*

$$h(0) \le A, h(-1/2) \ge -B, h(-1) \le C \text{ and } h(-3) \ge -D \tag{37}$$

for some $A, B, C, D \ge 0$, then

$$h(\theta) \le h(0) + 7A + 12.8B + 6C + 0.2D + L$$

for any $\theta \in [0, 1]$.

Proof. Define

$$\tilde{h}(\xi) = h(0) + h'(0)\xi + \frac{1}{2}h''(0)\xi^2 + \frac{1}{6}h'''(0)\xi^3.$$

By the Lipschitz continuity of h''', we have that $|h(\xi) - \tilde{h}(\xi)| \le L\xi^4/24$, for any $\xi \in \mathbb{R}$. Using the expressions for $\tilde{h}(x)$ at $\xi = -3, -1, -1/2$ to eliminate $h'(0), h''(0)$ and $h'''(0)$, we obtain:

$$\begin{aligned} \tilde{h}(\theta) = h(0) &- \tilde{h}(-3)\left[\frac{1}{30}\theta + \frac{1}{10}\theta^2 + \frac{1}{15}\theta^3\right] + \tilde{h}(-1)\left[\frac{3}{2}\theta + \frac{7}{2}\theta^2 + \theta^3\right] \\ &- \tilde{h}(-1/2)\left[\frac{24}{5}\theta + \frac{32}{5}\theta^2 + \frac{8}{5}\theta^3\right] + h(0)\left[\frac{10}{3}\theta + 3\theta^2 + \frac{2}{3}\theta^3\right]. \end{aligned}$$

Applying (37), $\theta \in [0,1]$ and $|h(\xi) - \tilde{h}(\xi)| \le L\xi^4/24$ gives the required bound:

$$\begin{aligned} h(\theta) &\le h(0) + 0.2D + 6C + 12.8B + 7A + \frac{L}{24}\left[\theta^4 + 0.2\cdot(-3)^4 + 6\cdot(-1)^4 + 12.8\cdot(-1/2)^4\right] \\ &\le h(0) + 7A + 12.8B + 6C + 0.2D + L \end{aligned}$$

□

We now prove Lemma 6 itself.

Lemma 6. *Let f be L_1-smooth and have L_3-Lipschitz continuous third-order derivatives, and let $\tau \le \sqrt{\alpha/(16L_3)}$ with $\tau, \alpha, L_1, L_3 > 0$. Consider* GUARDED-NON-CONVEX-AGD *with* FIND-BEST-ITERATE *replaced by* FIND-BEST-ITERATE$_3$. *At any iteration, if $u, v \neq$ NULL and the best iterate $b^{(1)}$ satisfies $f(b^{(1)}) \ge f(y_0) - \alpha\tau^2$ then,*

$$f(v) \le f(y_0) + 14\alpha\tau^2.$$

Proof. Let $0 \le j < t$ be such that $v = x_j$ (such j always exists by Corollary 1). If $j = 0$ then $x_j = y_0$ and the result is trivial, so we assume $j \ge 1$. Let

$$h(\theta) = f(y_j + \theta(y_j - y_{j-1})) - f(y_0) \text{ for } \theta \in \mathbb{R}$$

Note that

$$\begin{aligned} h(-3) &= f(q_j) - f(y_0) \ge f(b^{(1)}) - f(y_0) \ge -\alpha\tau^2, \\ h(-1) &= f(y_{j-1}) - f(y_0) \le 0, \\ h(-1/2) &= f(c_j) - f(y_0) \ge f(b^{(1)}) - f(y_0) \ge -\alpha\tau^2, \\ h(0) &= f(y_j) - f(y_0) \le 0 \text{ and} \\ h(\omega) &= f(x_j) - f(y_0), \end{aligned}$$

where $0 < \omega < 1$ is defined in line 1 of AGD-UNTIL-GUILTY, and we have used the guarantee $\max\{f(y_{j-1}), f(y_j)\} \le f(y_0)$ from Corollary 1. Moreover, by the Lipschitz continuity of the third derivatives of f, h''' is $L_3\|y_j - y_{j-1}\|^4$-Lipschitz continuous. Therefore, we can apply Lemma 9 with $A = C = 0$ and $B = D = \alpha\tau^2$ at $\theta = \omega$ and obtain

$$f(v) - f(y_0) = f(x_j) - f(y_0) \le f(y_j) - f(y_0) + 13\alpha\tau^2 + L_3\|y_j - y_{j-1}\|^4 \le 13\alpha\tau^2 + L_3\|y_j - y_{j-1}\|^4.$$

To complete the proof, we note that Lemma 3 guarantees $\|y_j - y_{j-1}\| \le \|y_j - y_0\| + \|y_{j-1} - y_0\| \le 2\tau$ and therefore

$$L_3\|y_j - y_{j-1}\|^4 \le 16L_3\tau^4 \le \alpha\tau^2,$$

where we have used $\tau^2 \le \alpha/(16L_3)$. □

C.3. Proof of Lemma 7

Lemma 7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L_1-smooth and have L_3-Lipschitz continuous third-order derivatives, let $\epsilon, \alpha > 0$ and $p_0 \in \mathbb{R}^d$. If p_0^K is the sequence of iterates produced by* GUARDED-NON-CONVEX-AGD*(f, p_0, L_1, ϵ, α, $\sqrt{\frac{2\alpha}{L_3}}$), then for every $1 \le k < K$,*

$$f(p_k) \le f(p_{k-1}) - \min\left\{\frac{\epsilon^2}{5\alpha}, \frac{\alpha^2}{32L_3}\right\}. \tag{17}$$

Proof. Fix an iterate index $1 \le k < K$; throughout the proof we let y_0^t, x_0^t and w refer to outputs of AGD-UNTIL-GUILTY in the kth iteration. We consider only the case $v, u \neq$ NULL, as the argument for $v, u =$ NULL is unchanged from Lemma 2.

As argued in the proof of Lemma 2, when $v, u \neq$ NULL, condition (10) holds. We set $\tau := \sqrt{\frac{\alpha}{32L_3}}$ and consider two cases. First, if $f(b^{(1)}) \le f(y_0) - \alpha\tau^2 = f(p_{k-1}) - \frac{\alpha^2}{32L_3}$ then we are done, since $f(p_k) \le f(b^{(1)})$. Second, if $f(b^{(1)}) \ge f(y_0) - \alpha\tau^2$, by Lemma 3 we have that

$$\|v - u\| \le 4\tau \le \sqrt{\frac{\alpha}{2L_3}} = \frac{\eta}{2},$$

Therefore, we can use Lemma 5 (with η as defined above) to show that

$$f(b^{(2)}) \leq \max\left\{f(v) - \frac{\alpha^2}{2L_3}, f(u) - \frac{\alpha^2}{6L_3}\right\}. \tag{38}$$

By Corollary 1, $f(u) \leq \hat{f}(u) \leq \hat{f}(y_0) = f(p_{k-1})$. Moreover, since $f(b^{(1)}) \geq f(y_0) - \alpha\tau^2$ and $\tau = \sqrt{\frac{\alpha}{32L_3}}$, we may apply Lemma 6 to obtain

$$f(v) \leq f(y_0) + 14\alpha\tau^2 \leq f(p_{k-1}) + \frac{7\alpha^2}{16L_3}.$$

Combining this with (38), we find that

$$f(p_k) \leq f(b^{(2)}) \leq f(p_{k-1}) - \min\left\{\frac{\alpha^2}{2L_3} - \frac{7\alpha^2}{16L_3}, \frac{\alpha^2}{6L_3}\right\} = f(p_{k-1}) - \frac{\alpha^2}{16L_3},$$

which concludes the case $v, u \neq \text{NULL}$ under third-order smoothness. □

C.4. Proof of Theorem 2

Theorem 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L_1-smooth and have L_3-Lipschitz continuous third-order derivatives. Let $p_0 \in \mathbb{R}^d$, $\Delta_f = f(p_0) - \inf_{z\in\mathbb{R}^d} f(z)$ and $0 < \epsilon^{2/3} \leq \min\{\Delta_f^{1/2} L_3^{1/6}, L_1/(8L_3^{1/3})\}$. If we set*

$$\alpha = 2L_3^{1/3}\epsilon^{2/3}, \tag{18}$$

GUARDED-NON-CONVEX-AGD*(f, p_0, L_1, ϵ, α, $\sqrt{\frac{2\alpha}{L_3}}$) finds a point p_K such that $\|\nabla f(p_K)\| \leq \epsilon$ and requires at most*

$$20 \cdot \frac{\Delta_f L_1^{1/2} L_3^{1/6}}{\epsilon^{5/3}} \log\left(\frac{500 L_1 \Delta_f}{\epsilon^2}\right) \tag{19}$$

gradient evaluations.

Proof. The proof proceeds exactly like the proof of Theorem 1. As argued there, the number of gradient evaluations is at most $2KT$, where K is number of iterations of GUARDED-NON-CONVEX-AGD and T is the maximum amount of steps performed in any call to AGD-UNTIL-GUILTY.

We derive the upper bound on K directly from Lemma 7, as by telescoping (12) we obtain

$$\Delta_f \geq f(p_0) - f(p_{K-1}) = \sum_{k=1}^{K-1} (f(p_{k-1}) - f(p_k)) \geq (K-1) \cdot \min\left\{\frac{\epsilon^2}{5\alpha}, \frac{\alpha^2}{32L_3}\right\} \geq (K-1)\frac{\epsilon^{4/3}}{10L_3^{1/3}},$$

where the last transition follows from substituting (18), our choice of α. We therefore conclude that

$$K \leq 1 + 10\Delta_f L_3^{1/3}\epsilon^{-4/3}. \tag{39}$$

To bound T, we recall that $\psi(z) \leq \Delta_f$ for every $z \in \mathbb{R}^d$, as argued in the proof Theorem 1. Therefore, substituting $\varepsilon = \epsilon/10$, $L = L_1 + 2\alpha$ and $\sigma = \alpha = 2L_3^{1/3}\epsilon^{2/3}$ into the guarantee (7) of Corollary 1 we obtain,

$$T \leq 1 + \sqrt{2 + \frac{L_1}{2L_3^{1/3}\epsilon^{2/3}}} \log_+\left(\frac{200(L_1 + 4L_3^{1/3}\epsilon^{2/3})\Delta_f}{\epsilon^2}\right), \tag{40}$$

where $\log_+(\cdot)$ is shorthand for $\max\{0, \log(\cdot)\}$.

Finally, we use $\epsilon^{2/3} \leq \min\{\Delta_f^{1/2} L_3^{1/6}, L_1/(8L_3^{1/3})\}$ to simplify the bounds on K and T. Using $1 \leq \Delta_f L_3^{1/3}\epsilon^{-4/3}$ reduces (28) to

$$K \leq 11\Delta_f L_3^{1/3}\epsilon^{-4/3}.$$

Applying $1 \leq L_1/(8L_3^{1/3}\epsilon^{2/3})$ on (29) gives

$$T \leq \sqrt{\frac{3}{4}} \frac{L_1^{1/2}}{L_3^{1/6}\epsilon^{1/3}} \log \frac{500 L_1 \Delta_f}{\epsilon^2},$$

where $\Delta_f L_1 \epsilon^{-2} \geq 8$ allows us to drop the subscript from the log. Multiplying the product of the above bounds by 2 gives the result. □

D. Adding a second-order guarantee

In this section, we sketch how to obtain simultaneous guarantees on the gradient and minimum eigenvalue of the Hessian. We use the $\widetilde{O}(\cdot)$ notation to hide logarithmic dependence on ϵ, Lipschitz constants Δ_f, L_1, L_2, L_3 and a high probability confidence parameter $\delta \in (0,1)$, as well as lower order polynomial terms in ϵ^{-1}.

Using approximate eigenvector computation, we can efficiently generate a direction of negative curvature, unless the Hessian is almost positive semi-definite. More explicitly, there exist methods of the form APPROX-EIG(f, x, L_1, α, δ) that require $\widetilde{O}(\sqrt{L_1/\alpha}\log d)$ Hessian-vector products to produce a unit vector v such that whenever $\nabla^2 f(x) \succeq -\alpha I$, with probability at least $1-\delta$ we have $v^T \nabla^2 f(x) v \leq -\alpha/2$, *e.g.* the Lanczos method (see additional discussion in (Carmon et al., 2016, §2.2)). Whenever a unit vector v satisfying $v^T \nabla^2 f(x) v \leq -\alpha/2$ is available, we can use it to make function progress. If $\nabla^2 f$ is L_2-Lipschitz continuous then by Lemma 1 $f(x \pm \frac{\alpha}{L_2} v) < f(x) - \frac{\alpha^3}{12 L_2^2}$ where by $f(x \pm z)$ we mean $\min\{f(x+z), f(x-z)\}$. If instead f has L_3-Lipschitz continuous third-order derivatives then by Lemma 4, $f(x \pm \sqrt{\frac{2\alpha}{L_3}} v) < f(x) - \frac{\alpha^2}{4L_3}$.

We can combine APPROX-EIG with Algorithm 3 that finds a point with a small gradient as follows:

$$\hat{z}_k \leftarrow \text{GUARDED-NON-CONVEX-AGD}(f, z_k, L_1, \epsilon, \alpha, \eta) \tag{41a}$$

$$v_k \leftarrow \text{APPROX-EIG}(f, \hat{z}_k, L_1, \alpha, \delta') \tag{41b}$$

$$z_{k+1} \leftarrow \operatorname*{arg\,min}_{x \in \{\hat{z}_k + \eta v_k, \hat{z}_k - \eta v_k\}} f(x) \tag{41c}$$

As discussed above, under third order smoothness , $\eta = \sqrt{2\alpha/L_3}$ guarantees that the step (41c) makes at least $\alpha^2/(4L_3)$ function progress whenever $v_k^T \nabla^2 f(\hat{z}_k) v_k \leq -\alpha/2$. Therefore the above iteration can run at most $\widetilde{O}(\Delta_f L_3/\alpha^2)$ times before $v_k^T \nabla^2 f(\hat{z}_k) v_k \geq -\alpha/2$ is satisfied. Whenever $v_k^T \nabla^2 f(\hat{z}_k) v_k \geq -\alpha/2$, with probability $1 - \delta' \cdot k$ we have the Hessian guarantee $\nabla^2 f(\hat{z}_k) \succeq -\alpha I$. Moreover, $\|\nabla f(\hat{z}_k)\| \leq \epsilon$ always holds. Thus, by setting $\alpha = L_3^{1/3}\epsilon^{2/3}$ we obtain the required second order stationarity guarantee upon termination of the iterations (41).

It remains to bound the computational cost of the method, with $\alpha = L_3^{1/3}\epsilon^{2/3}$. The total number of Hessian-vector products required by APPROX-EIG is,

$$\widetilde{O}\left(\Delta_f L_3/\alpha^2 \cdot \sqrt{\frac{L_1}{\alpha}} \log d\right) = \widetilde{O}\left(\Delta_f L_1^{1/2} L_3^{1/6} \epsilon^{-5/3} \log d\right).$$

Moreover, it is readily seen from the proof of Theorem 2 that every evaluation of (41a) requires at most

$$\widetilde{O}((f(x_k) - f(x_{k+1})) L_1^{1/2} L_3^{1/6} \epsilon^{-5/3} + L_1^{1/2} L_3^{-1/6} \epsilon^{-1/3}) \tag{42}$$

gradient and function evaluations. By telescoping the first term and multiplying the second by $\widetilde{O}(\Delta_f L_3/\alpha^2)$, we guarantee $\|\nabla f(x)\| \leq \epsilon$ and $\nabla^2 f(x) \succeq -L_3^{1/3}\epsilon^{2/3} I$ in at most $\widetilde{O}(\Delta_f L_1^{1/2} L_3^{1/6} \epsilon^{-5/3} \log d)$ function, gradient and Hessian-vector product evaluations.

The argument above is the same as the one used to prove Theorem 4.3 of (Carmon et al., 2016), but our improved guarantees under third order smoothness allows us get a better ϵ dependence for the complexity and lower bound on the Hessian in that regime. If instead we use the second order smoothness setting, we recover exactly the guarantees of (Carmon et al., 2016; Agarwal et al., 2016), namely $\|\nabla f(x)\| \leq \epsilon$ and $\nabla^2 f(x) \succeq -L_2^{1/2}\epsilon^{1/2} I$ in at most $\widetilde{O}(\Delta_f L_1^{1/2} L_2^{1/4} \epsilon^{-7/4} \log d)$ function, gradient and Hessian-vector product evaluations.

Finally, we remark that the above analysis would still apply if in (41a) we replace GUARDED-NON-CONVEX-AGD with any method with a run -time guarantee of the form (42). The resulting method will guarantee whatever the original method does, and also $\nabla^2 f(x) \succeq -\alpha I$. In particular, if the first method guarantees a small gradient, the combined method guarantees convergence to second-order stationary points.

E. Experiment details

E.1. Implementation details

Semi-adaptive gradient steps Both gradient descent and AGD are based on gradients steps of the form

$$y_{t+1} = x_t - \frac{1}{L_1}\nabla f(x_t). \tag{43}$$

In practice L_1 is often unknown and non-uniform, and therefore needs to be estimated adaptively. A common approach is backtracking line search, which we use for conjugate gradient. However, combining line search with AGD without invalidating its performance guarantees would involve non-trivial modification of the proposed method. Therefore, for the rest of the methods we keep an estimate of L_1, and double it whenever the gradient steps fails to make sufficient progress. That is, whenever

$$f\left(x_t - \frac{1}{L_1}\nabla f(x_t)\right) > f(x_t) - \frac{1}{2L_1}\left\|\nabla f(x_t)\right\|^2$$

we set $L_1 \leftarrow 2L_1$ and try again. In all experiments we start with $L_1 = 1$, which underestimates the actual smoothness of f by 2-3 orders of magnitude. We call our scheme for setting L_1 semi-adaptive, since we only increase L_1, and therefore do not adapt to situations where the function becomes more smooth as optimization progresses. Thus, we avoid painstaking tuning of L_1 while preserving the 'fixed step-size' nature of our approach, as L_1 is only doubled a small number of times.

Algorithm 3 We implement GUARDED-NON-CONVEX-AGD with the following modifications, indented to make it more practical without substantially compromising its theoretical properties.

1. We use the semi-adaptive scheme described above to set L. Specifically, whenever the gradient steps in lines 3 and 4 of AGD-UNTIL-GUILTY and CERTIFY-PROGRESS respectively fail, we double L until it succeeds, terminate AGD-UNTIL-GUILTY and multiply L_1 by the same factor.

2. We make the input parameters for AGD-UNTIL-GUILTY dynamic. In particular, we set $\epsilon' = \left\|\nabla f(p_{k-1})\right\|/10$ and use $\alpha = \sigma = C_1\left\|\nabla f(p_{k-1})\right\|^{2/3}$, where C_1 is a hyper-parameter. We use the same value of α to construct $\hat{f}$. This makes our implementation independent on the final desired accuracy ϵ.

3. In CERTIFY-PROGRESS we also test whether

   $$\hat{f}(x_t) + \nabla\hat{f}(x_t)^T(y_t - x_t) > \hat{f}(y_t).$$

   Since this inequality is a clear convexity violation, we return $w_t = y_t$ whenever it holds. We find that this substantially increases our method's capability of detecting negative curvature; most of the non-convexity detection in the first experiment is due to this check.

4. Whenever CERTIFY-PROGRESS produces a point $w_t \neq$ NULL (thereby proving non-convexity and stopping AGD-UNTIL-GUILTY), instead of finding a single pair (v, u) that violates strong convexity, we compute

   $$\alpha_{v,u} = 2\frac{f(v) - f(u) - \nabla f(v)^T(u - v)}{\left\|u - v\right\|^2}$$

   for the $2t$ points of the form $v = x_j$ and $u = y_j$ or $u = w_t$, with $0 \leq j < t$, where here we use the original f rather than $\hat{f}$ given to AGD-UNTIL-GUILTY. We discard all pairs with $\alpha_{v,u} < 0$ (no evidence of negative curvature), and select the 5 pairs with highest value of $\alpha_{v,u}$. For each selected pair v, u, we exploit negative curvature by testing all the points of the form $\{z \pm \eta\delta\}$ with $\delta = (u - v)/\left\|u - v\right\|$, $z \in \{v, u\}$ and η in a grid of 10 points log-uniformly spaced between $0.01\left\|u - v\right\|$ and $100(\left\|u\right\| + \left\|v\right\|)$.

5. In FIND-BEST-ITERATE$_3$ we compute c_j and q_j for every j such that $f(x_j) > f(y_j)$. Moreover, when $v, u = \text{NULL}$ (no non-convexity detected), we still set the next iterate p_k to be the output of FIND-BEST-ITERATE$_3$ rather than just the last AGD step.

The hyper-parameter C_1 was tuned separately for each experiment by searching on a small grid. For the regression experiment the tuning was performed on different problem instances (different seeds) than the ones reported in Fig. 1. For the neural network training problem the tuning was performed on a subsample of 10% of the data and a different random initialization than the one reported in Fig. 2. The specific parameters used were $C_1 = 0.01$ for regression and $C_1 = 0.1$ for neural network training.

Algorithm 3 without negative curvature exploitation This method is identical to the one described above, except that at every iteration p_k is set to $b^{(1)}$ produced by FIND-BEST-ITERATE$_3$ (*i.e.* the output of negative curvature exploitation is never used). We used the same hyper-parameters described above.

Gradient descent Gradient descent descent is simply (43), with $y_{t+1} = x_{t+1}$, where the semi-adaptive scheme is used to set L_1.

Adaptive restart accelerated gradient descent We use the accelerated gradient descent scheme of Beck and Teboulle (2009) with $\omega_t = t/(t+3)$. We use the restart scheme given by O'Donoghue and Candès (2015) where if $f(y_t) > f(y_{t-1})$ then we restart the algorithm from the point y_t. For the gradient steps we use the same semi-adaptive procedure described above and also restart the algorithm whenever the L_1 estimate changes (restarts performed for this reason are not shown in Fig. 1 and 2).

Non-linear conjugate gradient The method is given by the following recursion (Polak and Ribière, 1969),

$$\delta_t = -\nabla f(x_t) + \max\left\{\frac{\nabla f(x_t)^T(\nabla f(x_t) - \nabla f(x_{t-1}))}{\|\nabla f(x_{t-1})\|^2}, 0\right\}\delta_{t-1} \quad , \quad x_{t+1} = x_t + \eta_t \delta_t$$

where $\delta_0 = 0$ and η_t is found via backtracking line search, as follows. If $\delta^T \nabla f(x_t) \geq 0$ we set $\delta_t = -\nabla f(x_t)$ (truncating the recursion). We set $\eta_t = 2\eta_{t-1}$ and then check whether

$$f(x_t + \eta_t \delta_t) \leq f(x_t) + \frac{\eta_t \delta_t^T \nabla f(x_t)}{2}$$

holds. If it does we keep the value of η_t, and if it does not we set $\eta_t = \eta_t/2$ and repeat. The key difference from the semi-adaptive scheme used for the rest of the methods is the initialization $\eta_t = 2\eta_{t-1}$, that allows the step size to grow. Performing line search is crucial for conjugate gradient to succeed, as otherwise it cannot produce approximately conjugate directions. If instead we use the semi-adaptive step size scheme, performance becomes very similar to that of gradient descent.

Comparison of computational cost In the figures, the x-axis is set to the number of steps performed by the methods. We do this because it enables a one-to-one comparison between the steps of the restarted AGD and Algorithm 3. However, Algorithm 3 requires twice the number of gradient evaluations per step of the other algorithms. Furthermore, the number of function evaluations of Algorithm 3 increases substantially when we exploit negative curvature, due to our naive grid search procedure. Nonetheless, we believe it is possible to derive a variation of our approach that performs only one gradient computation per step, and yet maintains similar performance (see remark after Corollary 1, and that effective negative curvature exploitation can be carried out with only few function evaluations, using a line search.

While the rest of the methods tested require one gradient evaluation per step, the required number of function evaluations differs. GD requires only one function evaluation per step, while RAGD evaluates f twice per step (at x_t and y_t); the number of additional function evaluations due to the semi-adaptive scheme is negligible. NCG is expected to require more function evaluations due to its use of a backtracking line search. In the first experiment, NCG required 2 function evaluations per step on average, indicating that its L_1 estimate was stable for long durations. Alg. 3 required 5.3 function evaluations per step (on average over the 1,000 problem instances, with standard deviation 0.5), putting the amortized cost of our crude negative curvature exploitation scheme at 3.3 function evaluations per step.

E.2. Non-convex regression

The problem is to

$$\text{minimize } f(x) := \frac{1}{m}\sum_{i=1}^{m} \phi(a_i^T x - b_i)$$

where $\phi(\theta) = \theta^2/(1+\theta^2)$, $x \in \mathbb{R}^d$, $b \in \mathbb{R}^m$, and $a_i \in \mathbb{R}^d$. The function ϕ is a robust modification of the quadratic loss; it is approximately quadratic for small errors, but insensitive to larger errors.

To generate problem instances, we set $d = 30$, $m = 60$, and draw $a_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$. We draw b as follows. We first define draw a "ground truth" vector $z \sim \mathcal{N}(0, 4I_d)$. We then set $b = Az + 3\nu_1 + \nu_2$, where ν_1 is standard Gaussian and the elements of ν_2 are i.i.d. Bernoulli(0.3). The above parameters were manually chosen to make the problem substantially non-convex.

E.3. Neural network training

The function f is the average cross-entropy loss of 10-way prediction of class labels from input features. The prediction if formed by applying softmax on the output of a neural network with three hidden layers of 20, 10 and 5 units and $\tanh$ activations. To obtain data features we perform the following preprocessing, where the training examples are treated as 28^2 dimensional vectors. First, each example is separately normalized to zero mean and unit variance. Then, the $28^2 \times 28^2$ data covariance matrix is formed, and a projection to the 10 principle components is found via eigen-decomposition. The projection is then applied to the training set, and then each of the 10 resulting features is normalized to have zero mean and unit variance across the training set. The resulting model has $d = 545$ parameters and underfits the 60,000 examples training set. We randomly initialize the weights according the well-known scaling proposed by Glorot and Bengio (2010). We repeated the experiment for 10 different initializations of the weights, and all results were consistent with those reported in Fig. 2.