MIN-MAX OPTIMIZATION WITH A LOW-DIMENSIONAL BLOCK WITHOUT ACCURATE INNER SOLVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many learning problems couple a high-dimensional block of parameters with a few adversarial or dual variables, e.g., worst-group risk over a few groups or learning with a few constraints. We study $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y)$ with $\mathcal{X} \subseteq \mathbb{R}^n$, a compact convex set $\mathcal{Y} \subset \mathbb{R}^m$, and $m \ll n$, where $f(\cdot, y)$ is smooth and strongly convex with condition number κ, and count oracle calls for x and for y separately. The standard approach runs a cutting-plane method over y and solves each inner problem to accuracy ε with an accelerated method, costing $O(m\log(1/\varepsilon))$ oracle calls for y but $O(m\sqrt{\kappa}\log^2(1/\varepsilon))$ oracle calls for x. We show that accurate inner solves are unnecessary. Our method, certificate transport, maintains a strongly convex lower model of one slice $f(\cdot, y)$ and uses it to warm-start a short accelerated run on the next. Each oracle call for y then either cuts the localizer or, by concavity in y, transports the lower model to a mixture of the two slices with a certified increase. For $m = 1$, this yields an ε-saddle point with $O(\sqrt{\kappa}\log(1/\varepsilon))$ oracle calls for x, up to an initialization term, and $O(\log(1/\varepsilon))$ oracle calls for y, which is optimal both in the number of oracle calls for y and in the total number of oracle calls, even when $f(x,\cdot)$ is merely concave and Lipschitz. For general m, assuming the center of gravity of the localizer can be computed efficiently, the method uses $O((m+\sqrt{m\kappa})\log(1/\varepsilon))$ oracle calls for x and an optimal $O(m\log(1/\varepsilon))$ oracle calls for y.

1 INTRODUCTION

We consider the saddle-point problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y), \qquad x\in\mathbb{R}^n,\ \ y\in\mathbb{R}^m,\ \ m \ll n, \tag{1}$$

in which the two blocks of variables differ greatly in size: n is large, while m is small. For every y, the function $f(\cdot, y)$ is μ-strongly convex and L-smooth with condition number $\kappa = L/\mu$. For every x, the function $f(x,\cdot)$ is concave, with supergradients bounded in norm by M_y on a convex compact set $\mathcal{Y}$ of diameter R_y. No smoothness in y is assumed.

Problems of this form arise, for example, in learning under a few constraints (Agarwal et al., 2018; Cotter et al., 2019). There f is the Lagrangian, y is the vector of Lagrange multipliers, bounded as usual under Slater's condition, and m is the number of constraints. Another example is group distributionally robust learning (Sagawa et al., 2020), where y weights a few groups. In both examples f is linear in y, so the dual function is differentiable under strong convexity in x, but its smoothness constant, of order G^2/μ with G a bound on the Jacobian of the losses, is huge for small ridge parameters, and nonsmooth regularizers of y (e.g., ℓ_1 or total-variation penalties on the group weights) make it nonsmooth. Cutting planes are insensitive to both, so we assume no smoothness in y.

Each block calls for its own method. For the large block, accelerated gradient descent is optimal and dimension-free (Nesterov, 1983). For the small block, the center-of-gravity method needs only $O(m\log\frac{1}{\varepsilon})$ queries and requires no smoothness at all (Levin, 1965; Newman, 1965; Nemirovsky & Yudin, 1983). The two blocks are also accessed through different computations: a gradient in x costs a backward pass through the model, whereas a query in y returns only a function value and a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

supergradient in a low-dimensional space. We therefore count the calls to the two oracles separately, as T_x and T_y. The question is how to combine the two methods so that both counts are small.

The obvious approach and its cost. The textbook solution nests one method inside the other. Cutting planes are run on the dual function $\phi(y) = \min_x f(x, y)$, and every cut is computed by running accelerated gradient descent in x to high accuracy. With $B = M_y R_y$, this gives $T_y = O(m \log \frac{B}{\varepsilon})$ but $T_x = O(m\sqrt{\kappa} \log \frac{B}{\varepsilon} \log \frac{\mu R_x^2 + B}{\varepsilon})$, where R_x is the diameter of $\mathcal{X}$ (Alkousa et al., 2023; see (4)). The waste is plain: each inner problem is solved from scratch, and its solution is discarded as soon as y moves.

Our idea. The inner problems at different y are closely related, so the work spent on one of them can be reused. Suppose we hold a quadratic lower model $q_b \le f(\cdot, y_b)$, which certifies $\phi(y_b) \ge b$. We run a block of K accelerated gradient steps on $f(\cdot, y_c)$ at the current center of the localizer y_c, using q_b as a prior, and make a single y-query. By concavity in y, this yields a lower model of the same form at $y^+ = (1 - \vartheta) y_b + \vartheta y_c$, with a bound b^+ that increases whenever the queried value exceeds the current threshold. We call this *certificate transport*. With it, no inner problem is ever solved to accuracy ε: every y-query either shrinks the localizer or raises the certified lower bound.

1.1 Contributions

1. **Certificate transport.** We introduce a mechanism that couples cutting planes in y with blocks of K accelerated gradient steps in x, without ever solving an inner problem to high accuracy. It requires only concavity and Lipschitz continuity in y. The method maintains a computable bound on the saddle-point gap and returns a pair $(\bar{x}, y_b)$ certified by it (Theorem 1). Its correctness does not depend on K: the block length affects only the running time.
2. **Complexity for any block length** (Theorem 1). For every $K \ge 1$ the method reaches gap ε with

$$T_y = O\Big(\big(m + \tfrac{\kappa}{K^2}\big) \log \tfrac{B}{\varepsilon}\Big), \qquad T_x = O\Big(\sqrt{\kappa}\,\big[1 + \log_+ \tfrac{\mu R_x^2}{B}\big]\Big) + O\Big(\big(mK + \tfrac{\kappa}{K}\big) \log \tfrac{B}{\varepsilon}\Big),$$

   where the first term in T_x is a one-time initialization cost. The choice $K = \sqrt{\kappa/m}$ gives $T_y = O(m \log \frac{B}{\varepsilon})$, the optimal count for cutting planes, and $T_x = O(\sqrt{m\kappa} \log \frac{B}{\varepsilon})$ plus initialization. The product $m\sqrt{\kappa}$ of the nested scheme is thus replaced by the geometric mean $\sqrt{m\kappa}$, and the squared logarithm disappears. For $m = 1$, both counts match the known lower bounds for each block separately. The same choice of K minimizes our upper bound on any weighted cost $c_x T_x + c_y T_y$ up to a factor of three (Proposition 3).
3. **Independence of m under strong concavity** (Section 4). If f is also strongly concave in y and the coupling between the blocks is bounded by χ, a two-point accelerated method uses $T_x = O(\sqrt{\kappa + \chi}\,\Lambda)$ gradients, independently of m, and $T_y = O(m\sqrt{1 + \chi}\,\Lambda^2)$, with $\Lambda = 1 + \log \frac{B(1+\chi)}{\varepsilon}$. An accelerated method on the primal function has the same T_x but solves $O(\sqrt{\kappa + \chi})$ instead of $O(\sqrt{1 + \chi})$ problems in y per halving of the error.

Like the classical center-of-gravity method, our analysis treats the computation of centers of gravity in $\mathbb{R}^m$ as free and counts only oracle calls. Replacing them with implementable centers, as in Vaidya's method (Vaidya, 1996) or random-walk centroids (Bertsimas & Vempala, 2004), is a natural next step.

Related work. Nested schemes for problems with a low-dimensional block were analyzed by Gladin et al. (2021a); Alkousa et al. (2023); Gladin et al. (2022), and inexact oracles and accuracy certificates in such schemes by Gladin et al. (2021b; 2025). Our localizer $\{U \ge h\}$ is the level set of the level method (Lemaréchal et al., 1995). Bundle methods with inexact oracles (Kiwiel, 2006; de Oliveira et al., 2014; de Oliveira & Sagastizábal, 2014; van Ackooij & de Oliveira, 2014) and inexact dual decomposition (Necoara & Nedelcu, 2014) are the classical answer to Lagrangians with inexact inner solutions; they control the error of every inner solve, whereas certificate transport never solves an inner problem to a prescribed accuracy and reuses its partial solution instead. Warm starts across outer iterations are standard in bilevel optimization (Ji et al., 2021), but their analyses need the inner solution map to be Lipschitz, which we do not assume. Nested minimax methods that remove logarithmic factors (Lin et al., 2020; Kovalev & Gasnikov, 2022b) require smoothness

in both blocks. Algorithm 1 is the method of similar triangles (Gasnikov & Nesterov, 2018) with an arbitrary quadratic prior.

Organization. Section 2 states the assumptions, Sections 3 and 4 present the two methods, and Section 5 reports experiments. Proofs are in Sections A to C, experimental details in Section D, and a detailed discussion of related work in Section E.

2 Setting

Problem and accuracy. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set and $\mathcal{Y} \subset \mathbb{R}^m$ a convex body, that is, a compact convex set with nonempty interior in its affine hull; we assume $m = \dim \operatorname{aff} \mathcal{Y}$. We consider the problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y),$$

where $f(\cdot, y)$ is convex for every y and $f(x, \cdot)$ is concave for every x. Let

$$P(x) \coloneqq \max_{y \in \mathcal{Y}} f(x, y), \qquad \phi(y) \coloneqq \min_{x \in \mathcal{X}} f(x, y).$$

A candidate pair $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ is measured by the saddle-point gap

$$\operatorname{Gap}(\hat{x}, \hat{y}) \coloneqq P(\hat{x}) - \phi(\hat{y}) \ \geq 0, \tag{2}$$

which bounds the suboptimality of both $\hat{x}$ in the primal problem and $\hat{y}$ in the dual problem.

Split oracles. An algorithm accesses f only through two oracles, queried at points of $\mathcal{X} \times \mathcal{Y}$:

$$\mathsf{O}_x(x, y) = \nabla_x f(x, y), \qquad \mathsf{O}_y(x, y) = \big(f(x, y),\, g\big), \quad g \in \partial_y f(x, y), \tag{3}$$

where $\partial_y f$ denotes the superdifferential in y. We denote by T_x and T_y the numbers of calls to O_x and O_y. As in information-based complexity (Nemirovsky & Yudin, 1983), all other computations are free: projections onto $\mathcal{X}$, operations with the finitely many planes and quadratics collected so far, and centers of gravity of convex bodies in $\mathbb{R}^m$.

Assumptions. Our main results hold under the following assumption.

Assumption 1. The set $\mathcal{X}$ has diameter at most R_x, and $\mathcal{Y}$ has diameter at most R_y. For every $y \in \mathcal{Y}$, the function $f(\cdot, y)$ is μ-strongly convex and L-smooth on $\mathcal{X}$ with $\mu > 0$. For every $x \in \mathcal{X}$, the function $f(x, \cdot)$ is concave on $\mathcal{Y}$, and the supergradients returned by O_y have norm at most M_y.

We write $\kappa \coloneqq L/\mu$ and $B \coloneqq M_y R_y$. The constant B bounds the variation of $f(x, \cdot)$ over $\mathcal{Y}$ and sets the scale of the gap. Assumption 1 requires neither smoothness nor strong concavity in y, and nothing links the two blocks. The bound R_x is used only in the initialization (Lemma 8); for $\mathcal{X} = \mathbb{R}^n$ it suffices that R_x bounds the initial distance to $\arg\min f(\cdot, y_0)$.

In Section 4 we use the following additional assumption.

Assumption 2 (strong concavity and cross-Lipschitz coupling)**.** For every $x \in \mathcal{X}$, the function $f(x, \cdot)$ is ν-strongly concave on $\mathcal{Y}$ with $\nu > 0$, and for all $x \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,

$$\|\nabla_x f(x, y) - \nabla_x f(x, y')\| \leq \ell \, \|y - y'\|.$$

We set $\chi \coloneqq \ell^2/(\mu\nu)$. This quantity is scale-invariant and plays the role of a condition number for the coupling between the blocks.

Baseline. Running a cutting-plane method on the concave function ϕ and computing each cut by an accelerated method that solves the problem in x to accuracy ε (Alkousa et al., 2023) gives

$$T_x^{\mathrm{nest}} = O\Big(m\sqrt{\kappa}\, \log \frac{B}{\varepsilon}\, \log \frac{\mu R_x^2 + B}{\varepsilon}\Big), \qquad T_y^{\mathrm{nest}} = O\Big(m \log \frac{B}{\varepsilon}\Big). \tag{4}$$

In the rest of the paper we show how to replace the factor $m\sqrt{\kappa}$ by $\sqrt{m\kappa}$ and remove the second logarithm.

Algorithm 1 A block of accelerated steps with a prior model: $\textsc{Block}(F, \pi, K)$

Require: function F, μ_F-strongly convex and L_F-smooth on $\mathcal{X}$; prior model $\pi(x) = c + \frac{\gamma}{2}\|x - \bar{x}\|^2$; block length $K \geq 1$

$A_0 \leftarrow 0,\ \Gamma_0 \leftarrow \gamma,\ u_0 \leftarrow \bar{x},\ z_0 \leftarrow \bar{x},\ r_0 \leftarrow \gamma\bar{x}$
for $i = 1, \dots, K$ **do**
  $a_i \leftarrow \frac{(i+1)\gamma}{2L_F},\ A_i \leftarrow A_{i-1} + a_i,\ \Gamma_i \leftarrow \Gamma_{i-1} + \mu_F a_i$
  $x_i \leftarrow \frac{A_{i-1}u_{i-1} + a_i z_{i-1}}{A_i}$
  $g_i \leftarrow \nabla F(x_i)$ ▷ *one call to* O_x *if* $F = f(\cdot, y)$
  $r_i \leftarrow r_{i-1} + a_i(\mu_F x_i - g_i),\ z_i \leftarrow \Pi_{\mathcal{X}}\big(r_i/\Gamma_i\big)$
  $u_i \leftarrow \frac{A_{i-1}u_{i-1} + a_i z_i}{A_i}$
return (u_K, z_K, A_K), where $A_K = \frac{\gamma K(K+3)}{4L_F}$

3 Certificate transport

In this section we assume only Assumption 1. Nothing is known about $f(x, \cdot)$ beyond concavity and a Lipschitz bound: it may be piecewise linear, and its maximizers may lie on the boundary of $\mathcal{Y}$.

What the algorithm stores. The algorithm keeps two kinds of information. The first is a collection of *upper planes* $\ell_i(y) = v_i + \langle g_i, y - y_i\rangle$, each obtained from one call $\mathsf{O}_y(u_i, y_i)$. Since $f(u_i, \cdot)$ is concave, $\ell_i \geq f(u_i, \cdot) \geq \phi$ on $\mathcal{Y}$, and therefore

$$U := \min_i \ell_i \ \geq\ \phi, \qquad a := \max_{\mathcal{Y}} U \ \geq\ \max_{\mathcal{Y}} \phi.$$

The second is a single *lower model*

$$f(x, y_b) \ \geq\ q_b(x) := b + \tfrac{\sigma}{2}\|x - x_b\|^2 \quad \text{for all } x \in \mathcal{X}, \qquad \sigma := \mu/2, \tag{5}$$

which certifies $\phi(y_b) \geq b$. By the minimax theorem, some convex combination $\bar{x}$ of the points u_i satisfies $P(\bar{x}) \leq a$ (Lemma 12), hence $\mathrm{Gap}(\bar{x}, y_b) \leq a - b$. The algorithm reduces this *certified width* $a - b$ to ε.

The transport step. The key ingredient is the following observation (Lemma 6, proved in Section A). Run K steps of an accelerated method on $F = f(\cdot, y_c)$, using the lower model q_b of a *different* slice as a prior (Algorithm 1). The method returns two points $u_K, z_K \in \mathcal{X}$ and, with $A = K(K+3)/(8\kappa)$ and $v = F(u_K)$, guarantees

$$f\big(x, \underbrace{(1-\vartheta)y_b + \vartheta y_c}_{y^+}\big) \ \geq\ \underbrace{\frac{b + Av}{1 + A}}_{b^+} + \frac{\sigma}{2}\|x - z_K\|^2 \quad \forall x \in \mathcal{X}, \qquad \vartheta = \frac{A}{1+A}. \tag{6}$$

The proof combines the estimate-sequence invariant $b + A\,F(u_K) \leq \min_{\mathcal{X}} \Psi_K$ with concavity in y, namely $f(x, y^+) \geq (1-\vartheta)f(x, y_b) + \vartheta F(x)$. Only one function value, $v = F(u_K)$, is needed, and it comes together with a new upper plane from the same call to O_y.

The run is divided into epochs. In an epoch with initial width W, the threshold is $h = b + W/2$, and the algorithm queries the center of gravity y_c of the localizer $\{y \in \mathcal{Y} : U(y) \geq h\}$. Every query makes progress in one of two ways. If $v \leq h$, the new plane satisfies $\ell(y_c) = v \leq h$ and cuts off a half-space whose boundary passes through the center of gravity of the localizer; by Grünbaum's lemma, the volume of the localizer shrinks by a factor of at least $(1 - e^{-1})^{-1}$ (Lemma 10). If $v > h$, the certificate is transported, and b grows by $\vartheta(v - b) > \vartheta W/6$. Within an epoch b can grow by at most $W/3$, so each epoch contains $O(1/\vartheta) = O(1 + \kappa/K^2)$ such queries. In both cases the new plane can only decrease U, so the volume of the localizer does not grow within an epoch.

The width shrinks by a factor of at least $\frac{3}{2}$ per epoch, so there are $O(\log \frac{B}{\varepsilon})$ epochs in total. At a change of epoch, the volume of the localizer grows by at most a factor 6^m, and since U is Lipschitz, the volume never drops below $(\varepsilon/6B)^m \operatorname{vol}\mathcal{Y}$. Comparing these bounds shows that the total number of queries with $v \leq h$ is $O(m \log \frac{B}{\varepsilon})$ (Section B). Each query costs K gradients, so the total number of gradients is

$$\Big(\underbrace{m \cdot K}_{\text{cuts}} + \underbrace{\Big(1 + \frac{\kappa}{K^2}\Big) \cdot K}_{\text{transports}}\Big) \cdot O\Big(\log \frac{B}{\varepsilon}\Big), \tag{7}$$

Algorithm 2 Certificate transport (CT)

Require: $\varepsilon > 0$; block length $K \geq 1$ (default $K = \lceil \sqrt{8\kappa/m} \rceil$)
Build a lower model (b, x_b, y_b) and one plane with $a - b \leq 2B$ (Corollary 9)
$W \leftarrow a - b$
$h \leftarrow b + W/2$
while $a - b > \varepsilon$ **do**
  if $a - b \leq 2W/3$ **then**
    $W \leftarrow a - b$, $h \leftarrow b + W/2$ ▷ *new epoch*
  $y_c \leftarrow$ center of gravity of $\{y \in \mathcal{Y} : U(y) \geq h\}$
  $(u, z, A) \leftarrow \textsc{Block}(f(\cdot, y_c), q_b, K)$ ▷ K *calls to* O_x
  $(v, g) \leftarrow \mathsf{O}_y(u, y_c)$; add the plane $\ell(y) = v + \langle g, y - y_c \rangle$ ▷ *one call to* O_y
  if $v > h$ **then** ▷ *certificate transport*
    $b \leftarrow \frac{b + Av}{1+A}$, $x_b \leftarrow z$, $y_b \leftarrow \frac{y_b + Ay_c}{1+A}$
  $a \leftarrow \max_{\mathcal{Y}} U$
return $(\bar{x}, y_b)$, where $\bar{x} = \sum_i \lambda_i u_i$ and $\lambda \in \arg\min_{\lambda \in \Delta} \max_{y \in \mathcal{Y}} \sum_i \lambda_i \ell_i(y)$

and the choice $K = \sqrt{\kappa/m}$ balances the two terms at $O(\sqrt{m\kappa} \log \frac{B}{\varepsilon})$.

Theorem 1. *Let Assumption 1 hold and $0 < \varepsilon < 2B$. For every $K \geq 1$, Algorithm 2 returns $(\bar{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ with $\mathrm{Gap}(\bar{x}, \hat{y}) \leq \varepsilon$ after*

$$T_y = O\Big(\Big(m + \frac{\kappa}{K^2}\Big) \log \frac{B}{\varepsilon}\Big), \qquad T_x = I_0 + O\Big(\Big(mK + \frac{\kappa}{K}\Big) \log \frac{B}{\varepsilon}\Big)$$

oracle calls, where $I_0 = O\big(\sqrt{\kappa}\,[1 + \log_+ \frac{\mu R_x^2}{B}]\big)$ is the cost of initialization. In particular, for $K = \lceil \sqrt{8\kappa/m} \rceil$,

$$T_x + T_y = I_0 + O\Big(\big(m + \sqrt{m\kappa}\big) \log \frac{B}{\varepsilon}\Big), \qquad T_y = O\Big(m \log \frac{B}{\varepsilon}\Big).$$

The proof is given in Section B. The correctness of the output does not depend on K: the block length affects only the number of iterations.

Compared with the nested scheme (4), Theorem 1 improves T_x by a factor of order $\min\{\sqrt{m}, \sqrt{\kappa}\} \cdot \log \frac{1}{\varepsilon}$ and keeps the optimal T_y.

The one-dimensional case. For $m = 1$, the set $\mathcal{Y}$ is a segment and its center of gravity is the midpoint. Theorem 1 then gives the following.

Corollary 2. *Let Assumption 1 hold with $m = 1$, and let $0 < \varepsilon < 2B$. Then Algorithm 2 with $K = \lceil \sqrt{8\kappa} \rceil$ returns $(\bar{x}, \hat{y})$ with $\mathrm{Gap}(\bar{x}, \hat{y}) \leq \varepsilon$ after*

$$T_x = O\Big(\sqrt{\kappa}\,\Big[1 + \log_+ \frac{\mu R_x^2}{B} + \log \frac{B}{\varepsilon}\Big]\Big), \qquad T_y = O\Big(\log \frac{B}{\varepsilon}\Big)$$

oracle calls.

Both bounds match, in order, the known lower bounds for each block separately (Nemirovsky & Yudin, 1983; Nesterov, 2018) (for T_x, when n is large enough). Thus a merely concave, nonsmooth one-dimensional y-block costs nothing beyond its own complexity.

Choosing K for a weighted cost. Suppose a call to O_x costs c_x, a call to O_y costs c_y, and let $r = c_y/c_x$. Up to the initialization term, Theorem 1 bounds the total cost $c_x T_x + c_y T_y$ by $c_x\, g(K) \log \frac{B}{\varepsilon}$, where

$$g(K) = \Big(\alpha + \frac{\beta}{K^2}\Big)(K + r), \qquad \alpha \asymp m, \quad \beta \asymp \kappa.$$

Proposition 3. *For all $K > 0$, $g(K) \geq \max\{\alpha r,\, 2\sqrt{\alpha\beta}\}$. For $K_0 = \sqrt{\beta/\alpha}$,*

$$g(K_0) = 2\sqrt{\alpha\beta} + 2\alpha r \;\leq\; 3 \min_{K > 0} g(K).$$

Proof. First, $g(K) \geq \alpha(K + r) \geq \alpha r$. Second, $g(K) \geq (\alpha + \beta/K^2)K = \alpha K + \beta/K \geq 2\sqrt{\alpha\beta}$. Substituting K_0 gives $\alpha + \beta/K_0^2 = 2\alpha$ and $g(K_0) = 2\alpha(K_0 + r)$. □

Table 1. Oracle calls up to constants and the initialization term I_0, with Λ as in Theorem 4. Second row: reverse nested scheme, up to logarithms.

Method	T_x	T_y
Certificate transport (Theorem 1)	$(m+\sqrt{m\kappa})\,\Lambda$	$m\Lambda$
Accelerated method on P + cutting planes in y	$\sqrt{\kappa+\chi}\,\Lambda$	$m\sqrt{\kappa+\chi}\,\Lambda^2$
Two-point method (Algorithm 3, Theorem 4)	$\sqrt{\kappa+\chi}\,\Lambda$	$m\sqrt{1+\chi}\,\Lambda^2$

Thus the default block length $K \asymp \sqrt{\kappa/m}$ minimizes the upper bound g up to a factor of three, whatever the relative cost of the two oracles; the guarantee concerns the bound, not the actual cost on a given problem. If $\kappa < m$, then $K_0 < 1$ and one takes $K = 1$.

4 Strong concavity: a gradient count independent of m

Under Assumption 2, one can reverse the roles of the blocks: accelerate on the primal function P and use the small block only to build inexact models of it. The key inequality is a two-sided model of P. If y_i is a δ-maximizer of $f(v_i,\cdot)$, then for all $x \in \mathcal{X}$

$$f(x,y_i) \;\le\; P(x) \;\le\; f(x,y_i) + \gamma_\chi \|x - v_i\|^2 + 2\delta, \qquad \gamma_\chi = \ell^2/\nu = \mu\chi, \tag{8}$$

so P admits a smooth inexact model in the sense of Devolder et al. (2014); Stonyakin et al. (2021). An outer accelerated method with parameter $\Gamma = \mu + \gamma_\chi$ needs $O(\sqrt{1+\chi})$ steps to halve the error. Each step solves a proximal subproblem with condition number $1 + L/\Gamma$, which costs $O(\sqrt{1+\kappa/(1+\chi)})$ gradients. The product, $O(\sqrt{\kappa+\chi})$, does not depend on m. The subproblems are *not* solved to a prescribed accuracy: the inner run of Algorithm 1 returns two points that satisfy an exact inequality, which is plugged directly into the outer potential. The method is stated as Algorithm 3 in Section C, together with the proof of the following theorem.

Theorem 4. *Let Assumptions 1 and 2 hold, and let $0 < \varepsilon < B$. Put $\bar\chi = 1+\chi$ and $\Lambda = 1 + \log \frac{B\bar\chi}{\varepsilon}$. Algorithm 3, with the problems in y solved by the center-of-gravity method, returns $(\hat x, \hat y)$ with $\mathrm{Gap}(\hat x, \hat y) \le \varepsilon$ after*

$$T_x = I_0 + O\big(\sqrt{\kappa+\chi}\,\Lambda\big), \qquad T_y = O\big(m\sqrt{1+\chi}\,\Lambda^2\big)$$

oracle calls. If the problems in y are solved by another method that returns a δ-maximizer after $N_y(\delta)$ calls, then $T_y = O\big(\sqrt{1+\chi}\,\Lambda \cdot N_y(\varepsilon/(C\bar\chi^{5/2}))\big)$ for an absolute constant C.

Corollary 5. *Under the assumptions of Theorem 4, running the better of Algorithms 2 and 3 gives*

$$T_x + T_y = I_0 + O\Big(\min\Big\{(m+\sqrt{m\kappa})\,\Lambda,\;\; \sqrt{\kappa}\,\Lambda + m\sqrt{1+\chi}\,\Lambda^2\Big\}\Big).$$

In particular, if $\chi = O(1)$, then $T_x = I_0 + O(\sqrt{\kappa}\,\Lambda)$ and $T_y = O(m\Lambda^2)$, so the complexity is additive up to one logarithm in the y-term. Up to logarithms, Algorithm 3 is preferable to Algorithm 2 exactly when $\chi \lesssim \kappa/m$. If only T_x counts, which is natural when a gradient in x is a backward pass through a model, Algorithm 3 is preferable whenever $\chi \lesssim m\kappa$.

What the two-point method buys. A gradient count independent of m alone is not new: by (8), an accelerated method on P with the inexact gradient $\nabla_x f(v,\tilde y)$, where $\tilde y$ is a δ-maximizer of $f(v,\cdot)$, is an inexact first-order method in the sense of Devolder et al. (2014) and also gives $T_x = O(\sqrt{\kappa+\chi}\,\Lambda)$ up to logarithms, but it solves a problem in y at each of its $O(\sqrt{\kappa+\chi})$ steps per halving of the error. Algorithm 3 solves only $O(\sqrt{1+\chi})$ of them (Table 1); $\sqrt{\kappa+\chi}$ is the x-part of the lower bound of Zhang et al. (2022). The independence of m is for a fixed χ, which in natural examples may grow with m through the Jacobian of the group losses.

5 Experiments

The analysis predicts where certificate transport helps. A nested scheme pays for every inner solve, and even with warm starts this cost grows with κ and with the distance between consecutive slice solutions; CT pays a fixed $K_0 + 1 \approx \sqrt{8\kappa/m}$ gradients per query. The advantage should therefore appear with *many dual variables* and *weak regularization*.

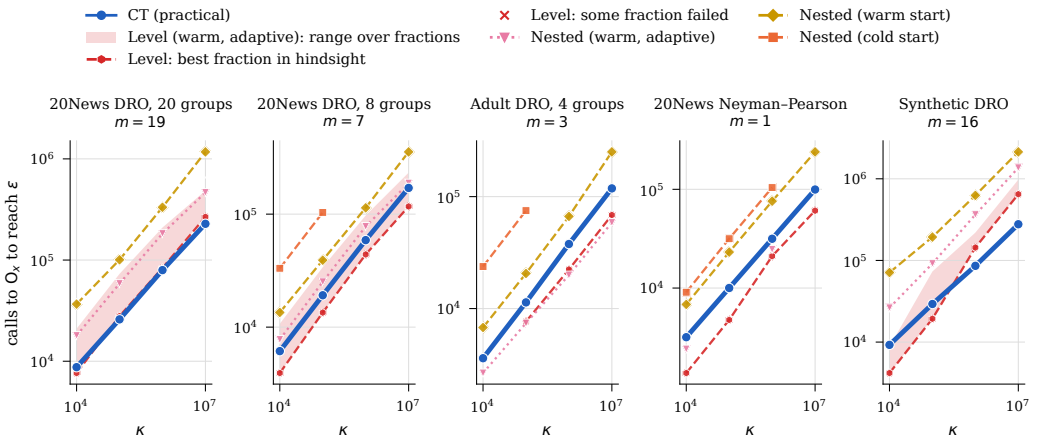

Figure 1. Calls to O_x to certify Gap $\leq 10^{-6}$ as κ grows, panels ordered by decreasing m. Band: the level baseline over its fractions c; dashed: its best c for each κ. CT has no tuned parameter.

Setup. We use group DRO, learning with a few constraints and Neyman–Pearson classification on Adult and 20 Newsgroups, and synthetic least-squares group DRO; $f(x, y) = \sum_j w_j(y)L_j(x) + \frac{\mu}{2}\|x\|^2 + \psi(y)$ with w affine in y, and the ridge parameter sets κ. *CT* is Algorithm 2 with the default K_0 and three changes that keep Theorem 1 valid (Section D.2): the largest step sizes allowed by the proof of Lemma 6, $\sigma = \mu$, and a direct certificate on the queried slice (one extra gradient per query). Baselines: the textbook nested scheme (4) (cold start, inner accuracy $\varepsilon/2$); warm-started inner runs; warm starts with inner accuracy proportional to the certified width; and, the strongest competitor, the latter with the localizer and epochs of Algorithm 2 ("level", an inexact level method (Lemaréchal et al., 1995)), whose fraction c is tuned over $\{0.03, 0.1, 0.3, 0.5, 1\}$ in hindsight. Extragradient and GDA with multiplicative weights are tuned on a grid. Reported gaps are computed independently and rigorously; the true gap never exceeded the certified width (Section D).

Many groups, weak regularization (Figure 1). For $m \geq 16$ and $\kappa \geq 10^6$ CT beats every configuration of the nested and level baselines (Table 2 and Figure 4). On the 20-group problem at $\kappa = 10^7$ it needs 228k calls to O_x and 88 to O_y; the level baseline needs 267k and 1,082 with its best fraction and 402k with the default one, and the nested schemes 465k–1.2M. On the synthetic instance with $m = 16$ the numbers are 278k for CT and 647k for the best level fraction. For smaller κ or $m \leq 7$ the tuned level baseline is faster, by 1.3–2.3×: CT's cost per query decays like $1/\sqrt{m}$, that of a warm-started solve does not. No fixed fraction is both fast and reliable (Table 3): relative to CT the geometric-mean ratio of T_x is 2.0, 1.2, 1.1, 0.91, 1.07 for $c = 0.03, \ldots, 1$, but $c = 0.5$ failed to certify the target once and $c = 1$ in 8 of 17 cells.

Strong regularization, few groups (Figure 2). At $\kappa = 10^4$ CT needs 2.7–7.0k gradients, 3.3–13× fewer than the textbook nested scheme and 2.5–6× fewer than the warm-started one; the tuned level baseline is 1.2–2.2× faster. Single-loop methods need 8–21× more gradients and over 10^3× more y-queries; on the nonsmooth ℓ_1 problem extragradient does not reach 10^{-5} in 10^5 calls. The second logarithm of (4) is visible directly: the cost of CT per digit of accuracy is constant, while for the textbook scheme it grows from 1.6k to 3.6k (Figure 5); in m the fitted exponent of T_x is 0.61 for CT versus 1.16 for the nested scheme (Figure 6).

Strong concavity (Figure 3). On the bilinear problem Algorithm 3 needs 3.6k calls to O_x for every $m \leq 32$, as Theorem 4 predicts, while CT grows as $m^{0.39}$. The independence of m is not specific to Algorithm 3: AGD on P with inexact gradients needs 0.9k calls. What Algorithm 3 saves is y-subproblems, 96 versus 1,900, in line with $\sqrt{1+\chi}$ versus $\sqrt{\kappa+\chi}$ per halving of the error. On χ^2-DRO the large χ makes Algorithm 3 cost 10–53k calls, while CT, which ignores strong concavity, needs 1.2–2.2k, as few as AGD on P with 70–170× fewer y-queries.

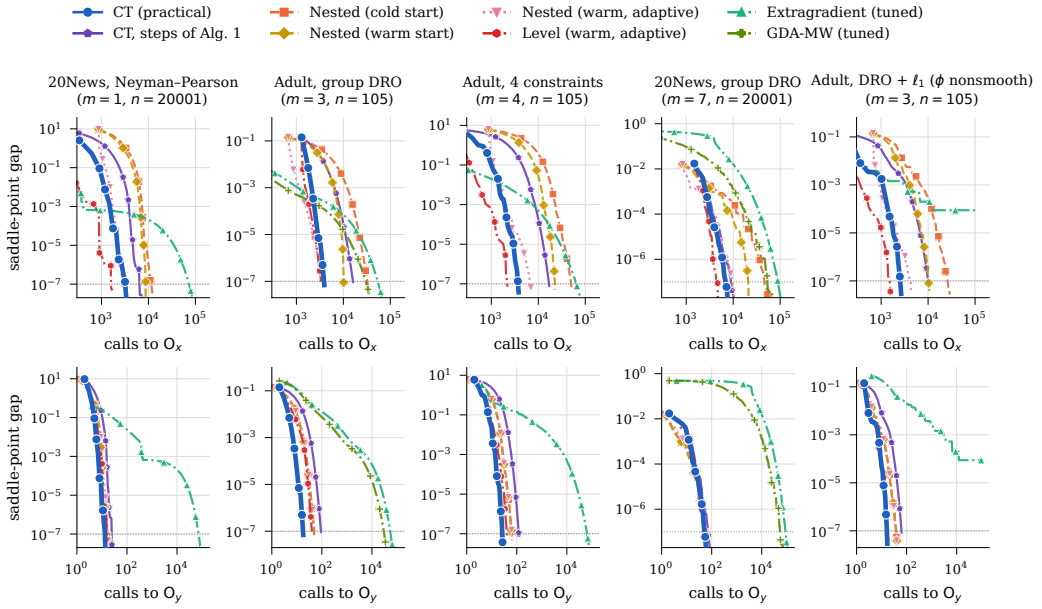

Figure 2. Gap versus calls to O_x (top) and O_y (bottom) on five real problems, $\kappa = 10^4$, target $\varepsilon = 10^{-7}$ (dotted). The last problem has an ℓ_1 penalty on the group weights, which makes ϕ nonsmooth.

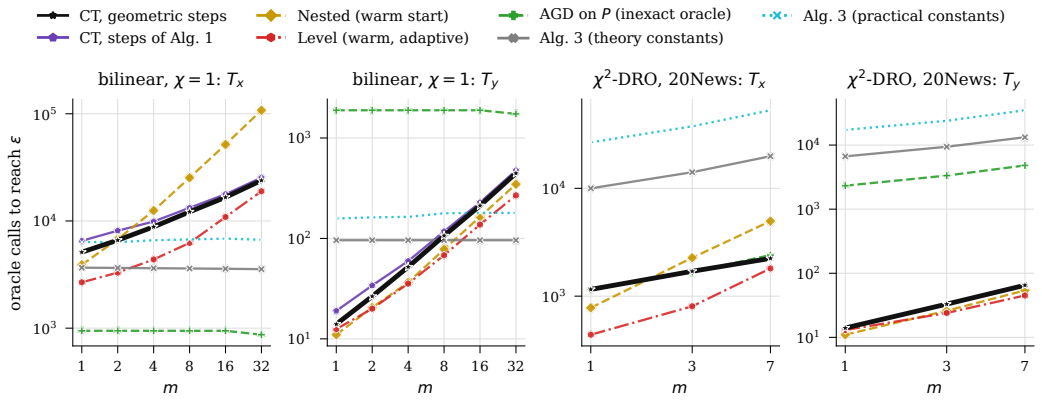

Figure 3. Strongly concave problems, $\varepsilon = 10^{-6}$. Left: bilinear problem with coupling $\chi = 1$ independent of m. Right: χ^2-regularized group DRO on 20 Newsgroups, where the certified $\chi = 1.5$–$6.2 \cdot 10^4$ grows with the number of groups. T_y counts y-subproblems.

Takeaway. With many dual variables and weak regularization CT is the fastest method, by up to $2.3\times$ in gradients and $12\times$ in y-queries over the best-tuned heuristic and 2–$8\times$ over nested schemes, with no tuning and a certified gap. With few dual variables and strong regularization a tuned warm-started level method is faster.

6 Conclusion

Certificate transport removes the multiplicative cost of nested schemes while asking nothing of the small block beyond concavity and Lipschitz continuity: an inexact inner solution either cuts the localizer or transports the lower model. The product $m\sqrt{\kappa}$ becomes $\sqrt{m\kappa}$, the second logarithm disappears, and for $m = 1$ both counts are optimal; under strong concavity, a two-point method makes the gradient count independent of m.

What the certificate gives. For group DRO, $a - b \le \varepsilon$ certifies the regularized worst-group risk $P(\bar{x}) = \max_j L_j(\bar{x}) + \frac{\mu}{2}\|\bar{x}\|^2$ up to ε; for constraints with multipliers in $[0, \bar{y}]^m$, it certifies the problem with the exact penalty $\bar{y} \sum_j \max\{0, g_j(x)\}$ on the violations $g_j(x)$.

Limitations. We assume exact oracles, full gradients in x and known L, μ, and treat centers of gravity as free; for larger m they must be approximated (Remark 1), as exact computation is hard (Rademacher, 2007). If O_y returns values with error δ_y, shifting the planes up and the transported value down by δ_y keeps all certificates valid, with attainable width of order δ_y. Optimality of $\sqrt{m\kappa}$ for $m > 1$ is open; the separate lower bounds give $\Omega((\sqrt{\kappa} + m) \log \frac{1}{\varepsilon})$.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation (parts of the text were translated from the authors draft in another language into English), to implement methods (the experimental code, including the preprocessing of the public datasets), to generate synthetic data sets (the synthetic instances of Section D, whose construction and properties were checked by the authors). Supporting qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to polish the writing, to create the figures from our experimental results, to find related work, and to check that bibliographic entries exist and are cited correctly. We have reviewed all AI-assisted work: the code and the all text was checked line by line by the authors. We take responsibility for the final content of this work, including text, claims produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical: it studies the oracle complexity of saddle-point problems and proposes an algorithm with provable guarantees. It involves no human subjects and no new data collection. The experiments use two public benchmark datasets, Adult and 20 Newsgroups. The Adult dataset contains sensitive attributes (sex and race), which we use only to define groups for the worst-group and constrained learning benchmarks that motivate the problem class. Our method makes such worst-group and constrained formulations cheaper to solve, but it does not decide which groups or constraints are appropriate in a given application, and a faster solver does not by itself make a model fair. We have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 2 (Assumptions 1 and 2). The algorithms are given as pseudocode in Algorithms 1 to 3. Complete proofs of all results are in the appendices: the technical tools in Section A, the proofs of Theorem 1 and Corollary 2 in Section B, and the proof of Theorem 4 in Section C. Section D describes the problems, datasets, preprocessing, baselines, parameter choices and evaluation protocol. The code is included in the supplementary material (directory `experiments/`); it produces every reported number from raw JSON logs. Both datasets are publicly available.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69, 2018.

M. S. Alkousa, A. V. Gasnikov, D. M. Dvinskikh, D. A. Kovalev, and F. S. Stonyakin. Accelerated methods for saddle-point problem. *Computational Mathematics and Mathematical Physics*, 60 (11):1787–1809, 2020.

Mohammad Alkousa, Alexander Gasnikov, Egor Gladin, Ilya Kuruzov, Dmitry Pasechnyuk, and Fedor Stonyakin. Solving strongly convex-concave composite saddle-point problems with low dimension of one group of variable. *Sbornik: Mathematics*, 214(3):285–333, 2023.

Dimitris Bertsimas and Santosh Vempala. Solving convex programs by random walks. *Journal of the ACM*, 51(4):540–556, 2004.

Ekaterina Borodich, Alexander Gasnikov, and Dmitry Kovalev. On linear convergence in smooth convex-concave bilinearly-coupled saddle-point optimization: Lower bounds and optimal algorithms. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5045–5100, 2025a.

Ekaterina Borodich, Georgiy Kormakov, Dmitry Kovalev, Aleksandr Beznosikov, and Alexander Gasnikov. Near-optimal algorithm with complexity separation for strongly convex-strongly concave composite saddle point problems. *Optimization Methods and Software*, 2025b. doi: 10.1080/10556788.2025.2545846.

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19052–19063, 2020.

Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Thinking inside the ball: Near-optimal minimization of the maximal loss. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pp. 866–882, 2021.

Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Optimal and adaptive Monteiro-Svaiter acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory (ALT)*, volume 98 of *Proceedings of Machine Learning Research*, pp. 300–332, 2019.

Welington de Oliveira and Claudia Sagastizábal. Level bundle methods for oracles with on-demand accuracy. *Optimization Methods and Software*, 29(6):1180–1209, 2014.

Welington de Oliveira, Claudia Sagastizábal, and Claude Lemaréchal. Convex proximal bundle methods in depth: a unified analysis for inexact oracles. *Mathematical Programming*, 148(1–2): 241–277, 2014.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1–2):37–75, 2014.

A. V. Gasnikov and Yu. E. Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

E. Gladin, M. Alkousa, and A. Gasnikov. Solving convex min-min problems with smoothness and strong convexity in one group of variables and low dimension in the other. *Automation and Remote Control*, 82(10):1679–1691, 2021a.

E. L. Gladin, A. V. Gasnikov, and E. S. Ermakova. Vaidya's method for convex stochastic optimization problems in small dimension. *Mathematical Notes*, 112(1–2):183–190, 2022.

Egor Gladin, Abdurakhmon Sadiev, Alexander Gasnikov, Pavel Dvurechensky, Aleksandr Beznosikov, and Mohammad Alkousa. Solving smooth min-min and min-max problems by mixed oracle algorithms. In *Mathematical Optimization Theory and Operations Research: Recent Trends (MOTOR 2021)*, Communications in Computer and Information Science, pp. 19–40. Springer, 2021b. doi: 10.1007/978-3-030-86433-0_2.

Egor Gladin, Maksim Lavrik-Karmazin, Karina Zainullina, Varvara Rudenko, Alexander Gasnikov, and Martin Takáč. Algorithm for constrained Markov decision process with linear convergence. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11506–11533, 2023.

Egor Gladin, Alexander Gasnikov, and Pavel Dvurechensky. Accuracy certificates for convex minimization with inexact oracle. *Journal of Optimization Theory and Applications*, 204(1), 2025. doi: 10.1007/s10957-024-02599-9.

B. Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4882–4892, 2021.

Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu-wai Wong. An improved cutting plane method for convex optimization, convex-concave games, and its applications. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 944–953, 2020.

Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4362–4415, 2022.

Krzysztof C. Kiwiel. A proximal bundle method with approximate subgradient linearizations. *SIAM Journal on Optimization*, 16(4):1007–1023, 2006.

Ron Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Dmitry Kovalev and Alexander Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022a.

Dmitry Kovalev and Alexander Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 14691–14703, 2022b.

Dmitry Kovalev, Alexander Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Guanghui Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159: 201–235, 2016.

Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1049–1065, 2015.

Yin Tat Lee, Aaron Sidford, and Santosh S. Vempala. Efficient convex optimization with membership oracles. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1292–1294, 2018.

Claude Lemaréchal, Arkadii Nemirovskii, and Yurii Nesterov. New variants of bundle methods. *Mathematical Programming*, 69(1–3):111–147, 1995.

A. Yu. Levin. On an algorithm for the minimization of convex functions. *Soviet Mathematics Doklady*, 6:286–290, 1965.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2738–2779, 2020.

Ion Necoara and Valentin Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *IEEE Transactions on Automatic Control*, 59(5):1232–1243, 2014.

Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Arkadi Nemirovski, Shmuel Onn, and Uriel G. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005.

Yu. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2 edition, 2018.

Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186:157–183, 2021.

Donald J. Newman. Location of the maximum on unimodal surfaces. *Journal of the ACM*, 12(3): 395–398, 1965.

D. A. Pasechnyuk and F. S. Stonyakin. One method for minimization a convex Lipschitz-continuous function of two variables on a fixed square. *Computer Research and Modeling*, 11(3):379–395, 2019.

Luis A. Rademacher. Approximating the centroid is hard. In *Proceedings of the 23rd Annual Symposium on Computational Geometry (SoCG)*, pp. 302–305, 2007.

Anton Rodomanov and Yurii Nesterov. Subgradient ellipsoid method for nonsmooth convex problems. *Mathematical Programming*, 199:305–341, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

N. Z. Shor. Cut-off method with space extension in convex programming problems. *Cybernetics*, 13(1):94–96, 1977.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Mohammad Alkousa, Dmitry Pasechnyuk, Sergei Artamonov, and Victoria Piskunova. Inexact model: A framework for optimization and variational inequalities. *Optimization Methods and Software*, 36(6):1155–1201, 2021.

Kiran Koshy Thekumparampil, Niao He, and Sewoong Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4281–4308, 2022.

Pravin M. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical Programming*, 73(3):291–341, 1996.

Wim van Ackooij and Welington de Oliveira. Level bundle methods for constrained convex optimization with various oracles. *Computational Optimization and Applications*, 57(3):555–597, 2014.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194:901–935, 2022.

A TOOLS: ACCELERATED BLOCKS, INITIALIZATION, CENTERS OF GRAVITY, AND RECOVERY

Throughout the appendices, $\mathcal{X} \subseteq \mathbb{R}^n$ is closed and convex, $\mathrm{proj}_{\mathcal{X}}$ is the Euclidean projection onto $\mathcal{X}$, and $\mathcal{Y}$ is a convex body. All statements about a function F on $\mathcal{X}$ use only its values and gradients at points of $\mathcal{X}$.

We will repeatedly use the following property of the projection: if $p \in \mathbb{R}^n$ and $z = \mathrm{proj}_{\mathcal{X}}(p)$, then $\langle p - z, x - z\rangle \le 0$ for all $x \in \mathcal{X}$, and therefore

$$\|x - p\|^2 = \|x - z\|^2 + 2\langle x - z, z - p\rangle + \|z - p\|^2 \;\ge\; \|x - z\|^2 + \|z - p\|^2 \qquad \forall x \in \mathcal{X}. \tag{9}$$

A.1 AN ACCELERATED BLOCK WITH AN ARBITRARY QUADRATIC PRIOR

The block of Algorithm 1 is an estimate-sequence method in which the usual starting quadratic is replaced by an arbitrary quadratic *prior* π. The prior enters the analysis only through the inequality $\pi + A_k F \ge c + A_k F(u_k) + \dots$ below; in Algorithm 2 it is the lower model of a *different* slice $f(\cdot, y_b)$.

Setting. Let $F : \mathcal{X} \to \mathbb{R}$ be μ_F-strongly convex and L_F-smooth on $\mathcal{X}$ with $0 \le \mu_F \le L_F$, and let

$$\pi(x) = c + \frac{\gamma}{2}\|x - \bar{x}\|^2, \qquad c \in \mathbb{R},\ \ \gamma > 0,\ \ \bar{x} \in \mathcal{X}.$$

Every gradient $g = \nabla F(t)$ at a point $t \in \mathcal{X}$ gives a quadratic lower bound on F by strong convexity:

$$F(x) \;\ge\; \omega_t(x) \coloneqq F(t) + \langle g, x - t\rangle + \frac{\mu_F}{2}\|x - t\|^2 \qquad \forall x \in \mathcal{X}. \tag{10}$$

Given nonnegative weights $a_1, a_2, \dots$ and query points $t_1, t_2, \dots \in \mathcal{X}$, put $A_0 = 0$, $A_i = A_{i-1} + a_i$, and

$$\Psi_i(x) \coloneqq \pi(x) + \sum_{j=1}^{i} a_j\, \omega_{t_j}(x), \qquad \Psi_0 = \pi.$$

Summing (10) over $j \le i$ with weights $a_j \ge 0$ and adding π gives, for every i,

$$\Psi_i(x) \;\le\; \pi(x) + A_i F(x) \qquad \forall x \in \mathcal{X}. \tag{11}$$

The iteration. Start from $u_0 = z_0 = \bar{x}$ and, for $i = 1, \dots, k$, compute

$$\begin{aligned} a_i &= \frac{(i+1)\gamma}{2L_F}, \qquad A_i = A_{i-1} + a_i, \qquad t_i = \frac{A_{i-1}u_{i-1} + a_i z_{i-1}}{A_i}, \qquad g_i = \nabla F(t_i), \\ z_i &= \arg\min_{x \in \mathcal{X}} \Psi_i(x), \qquad u_i = \frac{A_{i-1}u_{i-1} + a_i z_i}{A_i}. \end{aligned} \tag{12}$$

The points t_i are the points x_i of Algorithm 1; we rename them to keep x free for a generic point of $\mathcal{X}$. By induction, $t_i, u_i, z_i \in \mathcal{X}$: $z_i \in \mathcal{X}$ by definition, and t_i, u_i are convex combinations of points of $\mathcal{X}$.

Lemma 6 (accelerated block)**.** *In the setting above, the following hold.*

(i) $z_i = \mathrm{proj}_{\mathcal{X}}(r_i/\Gamma_i)$, *where* $\Gamma_i = \gamma + \mu_F A_i$ *and* $r_i = \gamma\bar{x} + \sum_{j \le i} a_j(\mu_F t_j - g_j)$. *Hence each step uses exactly one gradient and no function values.*

(ii) $A_k = \gamma k(k+3)/(4L_F)$.

(iii) *For all* $x \in \mathcal{X}$,

$$\pi(x) + A_k F(x) \;\ge\; \Psi_k(x) \;\ge\; c + A_k F(u_k) + \frac{\gamma + \mu_F A_k}{2}\|x - z_k\|^2. \tag{13}$$

Proof. Step 1: Ψ_i is an isotropic quadratic. Expanding the squares in π and in each ω_{t_j} and collecting the terms of degree two, one and zero in x,

$$\Psi_i(x) = c + \frac{\gamma}{2}\|x-\bar{x}\|^2 + \sum_{j\le i} a_j\Big[F(t_j) + \langle g_j, x - t_j\rangle + \frac{\mu_F}{2}\|x-t_j\|^2\Big]$$
$$= \frac{\gamma+\mu_F A_i}{2}\|x\|^2 - \Big\langle \gamma\bar{x} + \sum_{j\le i} a_j(\mu_F t_j - g_j),\, x\Big\rangle + \mathrm{const}_i = \frac{\Gamma_i}{2}\|x\|^2 - \langle r_i, x\rangle + \mathrm{const}_i,$$

where const_i does not depend on x (it contains the unknown values $F(t_j)$). Completing the square,

$$\Psi_i(x) = \Psi_i^\star + \frac{\Gamma_i}{2}\Big\|x - \frac{r_i}{\Gamma_i}\Big\|^2 \tag{14}$$

for some constant $\Psi_i^\star$. The minimizer of Ψ_i over $\mathcal{X}$ is therefore the projection of r_i/Γ_i onto $\mathcal{X}$, which proves (i). For $i=0$ we get $r_0/\Gamma_0 = \bar{x} \in \mathcal{X}$, so $z_0 = \bar{x}$, in agreement with the initialization.

Step 2: quadratic growth around z_i. Apply (9) with $p = r_i/\Gamma_i$ and $z = z_i$ to (14):

$$\Psi_i(x) \;\ge\; \Psi_i(z_i) + \frac{\Gamma_i}{2}\|x-z_i\|^2 = \min_{\mathcal{X}}\Psi_i + \frac{\Gamma_i}{2}\|x-z_i\|^2 \qquad \forall x\in\mathcal{X}. \tag{15}$$

Step 3: the weights. Since $\sum_{j=1}^{i}(j+1) = \frac{i(i+1)}{2}+i = \frac{i(i+3)}{2}$, we have $A_i = \frac{\gamma}{2L_F}\cdot\frac{i(i+3)}{2} = \frac{\gamma i(i+3)}{4L_F}$, which proves (ii). Moreover, for $i\ge 1$,

$$L_F a_i^2 = \frac{\gamma^2(i+1)^2}{4L_F} \;\le\; \frac{\gamma^2 i(i+3)}{4L_F} = \gamma A_i \;\le\; \Gamma_{i-1}A_i, \tag{16}$$

because $(i+1)^2 = i^2+2i+1 \le i^2+3i = i(i+3)$ when $i\ge 1$, and $\Gamma_{i-1} = \gamma + \mu_F A_{i-1} \ge \gamma$.

Step 4: the invariant $c + A_i F(u_i) \le \min_{\mathcal{X}}\Psi_i$. We argue by induction on i. For $i=0$, both sides equal c, since $A_0 = 0$ and $\min_{\mathcal{X}}\Psi_0 = \pi(\bar{x}) = c$. Assume the invariant for $i-1$ and evaluate $\Psi_i = \Psi_{i-1} + a_i\omega_{t_i}$ at its minimizer z_i:

$$\min_{\mathcal{X}}\Psi_i = \Psi_i(z_i) = \Psi_{i-1}(z_i) + a_i\Big[F(t_i) + \langle g_i, z_i - t_i\rangle + \frac{\mu_F}{2}\|z_i-t_i\|^2\Big].$$

We bound the two terms on the right. By (15) for $i-1$ and then the induction hypothesis,

$$\Psi_{i-1}(z_i) \;\ge\; \min_{\mathcal{X}}\Psi_{i-1} + \frac{\Gamma_{i-1}}{2}\|z_i - z_{i-1}\|^2 \;\ge\; c + A_{i-1}F(u_{i-1}) + \frac{\Gamma_{i-1}}{2}\|z_i-z_{i-1}\|^2.$$

In the bracket we drop the nonnegative term $\frac{\mu_F}{2}\|z_i-t_i\|^2$ (recall $a_i\ge 0$). Hence

$$\min_{\mathcal{X}}\Psi_i \;\ge\; c + A_{i-1}F(u_{i-1}) + a_i\big[F(t_i) + \langle g_i, z_i - t_i\rangle\big] + \frac{\Gamma_{i-1}}{2}\|z_i-z_{i-1}\|^2.$$

Next, by convexity of F, $F(u_{i-1}) \ge F(t_i) + \langle g_i, u_{i-1} - t_i\rangle$. Substituting and using $A_{i-1}+a_i = A_i$,

$$\min_{\mathcal{X}}\Psi_i \;\ge\; c + A_iF(t_i) + \big\langle g_i,\; A_{i-1}u_{i-1} + a_iz_i - A_it_i\big\rangle + \frac{\Gamma_{i-1}}{2}\|z_i-z_{i-1}\|^2.$$

By the definition of u_i, $A_{i-1}u_{i-1} + a_iz_i = A_iu_i$, so the inner product equals $A_i\langle g_i, u_i - t_i\rangle$ and

$$\min_{\mathcal{X}}\Psi_i \;\ge\; c + A_i\big[F(t_i) + \langle g_i, u_i - t_i\rangle\big] + \frac{\Gamma_{i-1}}{2}\|z_i-z_{i-1}\|^2. \tag{17}$$

Now we use smoothness. Subtracting the definitions of t_i and u_i gives $u_i - t_i = \frac{a_i}{A_i}(z_i - z_{i-1})$. By L_F-smoothness,

$$F(u_i) \;\le\; F(t_i) + \langle g_i, u_i - t_i\rangle + \frac{L_F}{2}\|u_i-t_i\|^2 = F(t_i) + \langle g_i, u_i - t_i\rangle + \frac{L_Fa_i^2}{2A_i^2}\|z_i-z_{i-1}\|^2.$$

Multiplying by A_i and inserting the result into (17),

$$\min_{\mathcal{X}}\Psi_i \;\ge\; c + A_iF(u_i) + \frac{1}{2}\Big[\Gamma_{i-1} - \frac{L_Fa_i^2}{A_i}\Big]\|z_i-z_{i-1}\|^2 \;\ge\; c + A_iF(u_i),$$

where the bracket is nonnegative by (16). This completes the induction.

Step 5: conclusion. The first inequality in (13) is (11) with $i=k$. For the second, combine (15) for $i=k$ with the invariant of Step 4: $\Psi_k(x) \ge \min_{\mathcal{X}}\Psi_k + \frac{\Gamma_k}{2}\|x-z_k\|^2 \ge c + A_kF(u_k) + \frac{\Gamma_k}{2}\|x-z_k\|^2$, and $\Gamma_k = \gamma + \mu_F A_k$. □

Algorithm 1 is Lemma 6 with $F = f(\cdot, y)$, $\mu_F = \mu$, $L_F = L$, and prior $\pi = q_b$, that is, $c = b$, $\gamma = \sigma = \mu/2$ and $\bar{x} = x_b$. In this case

$$A_K = \frac{\sigma K(K+3)}{4L} = \frac{K(K+3)}{8\kappa}.$$

A.2 Certificate transport

Corollary 7 (transport). *Let Assumption 1 hold, and let (b, x_b, y_b) be a lower model, that is, $x_b \in \mathcal{X}$, $y_b \in \mathcal{Y}$, and*

$$f(x, y_b) \;\geq\; q_b(x) = b + \frac{\sigma}{2}\|x - x_b\|^2 \quad \forall x \in \mathcal{X}, \qquad \sigma = \mu/2.$$

Let $y_c \in \mathcal{Y}$, let $(u, z, A) = \textsc{Block}(f(\cdot, y_c), q_b, K)$, and let $v = f(u, y_c)$. Put $\vartheta = A/(1+A)$. Then the triple

$$b^+ = \frac{b + Av}{1+A}, \qquad x_b^+ = z, \qquad y_b^+ = (1-\vartheta)y_b + \vartheta y_c = \frac{y_b + Ay_c}{1+A}$$

is again a lower model: $x_b^+ \in \mathcal{X}$, $y_b^+ \in \mathcal{Y}$, and $f(x, y_b^+) \geq b^+ + \frac{\sigma}{2}\|x - x_b^+\|^2$ for all $x \in \mathcal{X}$. This is inequality (6).

Proof. First, $z \in \mathcal{X}$ by Lemma 6(i), and $y_b^+ \in \mathcal{Y}$ as a convex combination of $y_b, y_c \in \mathcal{Y}$. Fix $x \in \mathcal{X}$ and apply (13) to $F = f(\cdot, y_c)$ with prior $\pi = q_b$ (so $c = b$, $\gamma = \sigma$, $\mu_F = \mu$) and $F(u) = v$:

$$q_b(x) + A\, f(x, y_c) \;\geq\; b + Av + \frac{\sigma + \mu A}{2}\|x - z\|^2.$$

Since $f(x, y_b) \geq q_b(x)$, the left-hand side is at most $f(x, y_b) + Af(x, y_c)$. By concavity of $f(x, \cdot)$,

$$f(x, y_b^+) \;\geq\; (1-\vartheta)f(x, y_b) + \vartheta f(x, y_c) = \frac{f(x, y_b) + Af(x, y_c)}{1+A}.$$

Combining the two displays and dividing by $1 + A$,

$$f(x, y_b^+) \;\geq\; \frac{b + Av}{1+A} + \frac{\sigma + \mu A}{2(1+A)}\|x - z\|^2 \;\geq\; b^+ + \frac{\sigma}{2}\|x - z\|^2,$$

where the last step uses $\frac{\sigma + \mu A}{1+A} \geq \sigma$, which holds because $\mu \geq \sigma$. □

The conclusion holds for *every* value of v: the transported triple is always a valid lower model. The comparison of v with the threshold in Algorithm 2 matters only for progress, because $b^+ - b = \vartheta(v - b)$ is positive exactly when $v > b$. The proof also shows where the curvature $\sigma = \mu/2$ of the lower model comes from: the transported curvature is $\frac{\sigma + \mu A}{1+A} \geq \sigma$ for every $\sigma \leq \mu$, so transport works equally with $\sigma = \mu$, and the factor $\frac{1}{2}$ is needed only by the initialization of Lemma 8.

A.3 Initialization of the lower model

Lemma 8 (initialization). *Let F be μ-strongly convex and L-smooth on $\mathcal{X}$, let $\operatorname{diam}\mathcal{X} \leq R_x$, let $x^\star = \arg\min_{\mathcal{X}} F$, and let $E_{\text{target}} > 0$. Using at most*

$$\lceil\sqrt{8\kappa}\,\rceil\left(2 + \left\lceil \log_2^+ \frac{\mu R_x^2}{4E_{\text{target}}}\right\rceil\right)$$

gradients of F and no function values, one can find $u, z \in \mathcal{X}$ such that

$$F(u) - F(x^\star) \leq \frac{E_{\text{target}}}{2} \qquad \text{and} \qquad F(x) \;\geq\; F(u) - 3E_{\text{target}} + \frac{\sigma}{2}\|x - z\|^2 \quad \forall x \in \mathcal{X},$$

where $\sigma = \mu/2$.

Proof. One run. Fix $v \in \mathcal{X}$ and apply Lemma 6 with $\mu_F = \mu$, $L_F = L$, prior $\pi(x) = \frac{\gamma}{2}\|x - v\|^2$ (so $c = 0$, $\bar{x} = v$), $\gamma = \mu/2$, and $k = \lceil\sqrt{8\kappa}\,\rceil$ steps. Let u, z be its output and

$$A := A_k = \frac{\mu k(k+3)}{8L} \;\geq\; \frac{k^2}{8\kappa} \;\geq\; 1.$$

Dividing (13) by A gives

$$F(x) \;\geq\; F(u) + \frac{\mu/2 + \mu A}{2A}\|x - z\|^2 - \frac{\mu}{4A}\|x - v\|^2 \qquad \forall x \in \mathcal{X}. \tag{18}$$

We extract three consequences.

(a) *Accuracy of u.* Take $x = x^\star$ in (18) and drop the nonnegative term with $\|x^\star - z\|^2$: $F(u) - F(x^\star) \leq \frac{\mu}{4A}\|x^\star - v\|^2$.

(b) *Position of z.* Take $x = x^\star$ again and use $F(x^\star) \leq F(u)$: $\frac{\mu/2+\mu A}{2A}\|x^\star - z\|^2 \leq \frac{\mu}{4A}\|x^\star - v\|^2$, that is, $\|x^\star - z\|^2 \leq \frac{\|x^\star - v\|^2}{1+2A} \leq \frac{\|x^\star - v\|^2}{3}$. By the triangle inequality,

$$\|z - v\|^2 \leq \big(\|z - x^\star\| + \|x^\star - v\|\big)^2 \leq \big(1 + 3^{-1/2}\big)^2\|x^\star - v\|^2 \leq 3\|x^\star - v\|^2.$$

(c) *Lower model centered at z.* Insert $\|x - v\|^2 \leq 2\|x - z\|^2 + 2\|z - v\|^2$ into (18). The coefficient of $\|x - z\|^2$ becomes $\frac{\mu/2+\mu A}{2A} - \frac{\mu}{2A} = \frac{\mu}{2} - \frac{\mu}{4A} \geq \frac{\mu}{4} = \frac{\sigma}{2}$ (since $A \geq 1$), and by (b) the remaining term is at least $-\frac{\mu}{2A} \cdot 3\|x^\star - v\|^2 \geq -\frac{3\mu}{2}\|x^\star - v\|^2$. Thus

$$F(x) \;\geq\; F(u) + \frac{\sigma}{2}\|x - z\|^2 - \frac{3\mu}{2}\|x^\star - v\|^2 \qquad \forall x \in \mathcal{X}.$$

Suppose now that $F(v) - F(x^\star) \leq E$. Since $x^\star$ minimizes the μ-strongly convex F over $\mathcal{X}$, we have $F(v) - F(x^\star) \geq \frac{\mu}{2}\|v - x^\star\|^2$, hence $\|x^\star - v\|^2 \leq 2E/\mu$. Then (a) gives $F(u) - F(x^\star) \leq \frac{E}{2A} \leq \frac{E}{2}$, and (c) gives $F(x) \geq F(u) - 3E + \frac{\sigma}{2}\|x - z\|^2$. In words: *one run started from a point with accuracy E halves the accuracy and produces a lower model with constant $F(u) - 3E$.*

Restarts. Start from an arbitrary $v_0 \in \mathcal{X}$. Since $\|x^\star - v_0\| \leq R_x$, (a) shows that the first run returns u_1 with $F(u_1) - F(x^\star) \leq \frac{\mu R_x^2}{4} =: E_1$. Restarting from u_s gives $E_{s+1} = E_s/2$, so $E_s = 2^{1-s}E_1 \leq E_{\text{target}}$ as soon as $s \geq 1 + \log_2 \frac{\mu R_x^2}{4E_{\text{target}}}$. After $s = 1 + \lceil \log_2^+ \frac{\mu R_x^2}{4E_{\text{target}}} \rceil$ runs we have a point v with $F(v) - F(x^\star) \leq E_{\text{target}}$, and one more run from v gives the claim with $E = E_{\text{target}}$. Each run costs $k = \lceil\sqrt{8\kappa}\,\rceil$ gradients, and no run uses function values. □

The lemma produces a lower model whose constant $F(u) - 3E_{\text{target}}$ involves the value $F(u)$, which the lemma itself does not compute. In Algorithm 2 this value comes for free with the first query in y.

Corollary 9 (start of Algorithm 2)**.** *Let Assumption 1 hold, let $y_0 \in \mathcal{Y}$ be arbitrary, and apply Lemma 8 to $F = f(\cdot, y_0)$ with $E_{\text{target}} = B/3$. Make one call $(v_0, g_0) = \mathsf{O}_y(u, y_0)$, so that $v_0 = f(u, y_0)$. Then*

$$b = v_0 - B, \qquad x_b = z, \qquad y_b = y_0$$

is a lower model, the plane $\ell_0(y) = v_0 + \langle g_0, y - y_0\rangle$ satisfies $\ell_0 \geq f(u, \cdot)$ on $\mathcal{Y}$, and $a = \max_{\mathcal{Y}} \ell_0 \leq v_0 + B$. Hence $a - b \leq 2B$. The cost is

$$I_0 = \lceil\sqrt{8\kappa}\,\rceil\left(2 + \left\lceil \log_2^+ \frac{3\mu R_x^2}{4B} \right\rceil\right) = O\left(\sqrt{\kappa}\left[1 + \log_+ \frac{\mu R_x^2}{B}\right]\right)$$

calls to O_x and one call to O_y.

Proof. By Lemma 8 with $E_{\text{target}} = B/3$, $f(x, y_0) \geq v_0 - B + \frac{\sigma}{2}\|x - z\|^2$ for all $x \in \mathcal{X}$, which is the lower model. The plane is an upper bound on $f(u, \cdot)$ by concavity. For every $y \in \mathcal{Y}$, $\ell_0(y) \leq v_0 + \|g_0\|\|y - y_0\| \leq v_0 + M_y R_y = v_0 + B$. □

A.4 CENTERS OF GRAVITY

Lemma 10 (Grünbaum). *Let $S \subset \mathbb{R}^m$ be a convex body with center of gravity c, and let H be a closed half-space with $c \notin \operatorname{int} H$. Then $\operatorname{vol}(S \cap H) \le (1 - e^{-1}) \operatorname{vol}(S)$ (Grünbaum, 1960). For $m = 1$ the constant $1 - e^{-1}$ can be replaced by $1/2$.*

In our use, $H = \{y : \langle g, y - c\rangle \ge 0\}$ is a half-space whose boundary passes through c, so $c \notin \operatorname{int} H$.

The next lemma is the center-of-gravity method with inexact function values; it is used in Section C to solve the problems in y.

Lemma 11 (center of gravity with inexact values). *Let ψ be concave on $\mathcal{Y}$. Put $\mathcal{Y}_0 = \mathcal{Y}$ and, for $t = 1, 2, \ldots$, let the query y_t be the center of gravity of $\mathcal{Y}_{t-1}$ and*

$$\mathcal{Y}_t = \{y \in \mathcal{Y}_{t-1} : \ \langle g_t, y - y_t\rangle \ge 0\}.$$

Suppose that each query returns a plane $\ell_t(y) = w_t + \langle g_t, y - y_t\rangle$ with $\ell_t \ge \psi$ on $\mathcal{Y}$ and $\|g_t\| \le M_y$, and a number $\beta_t \le \psi(y_t)$ with $w_t - \beta_t \le \delta$. Let $a_t = \max_{\mathcal{Y}} \min_{s \le t} \ell_s$. Then

$$a_t - \max_{s \le t} \beta_s \ \le \ B(1 - e^{-1})^{t/m} + \delta.$$

In particular, a δ-maximizer of a concave M_y-Lipschitz function on $\mathcal{Y}$ can be found with $O(m \log \frac{B}{\delta})$ exact queries of its value and a supergradient.

Proof. Degenerate cases. If $g_s = 0$ for some $s \le t$, then $\ell_s \equiv w_s$, so $a_t \le \max_{\mathcal{Y}} \ell_s = w_s \le \beta_s + \delta$, and the claim holds. If $\operatorname{vol} \mathcal{Y}_s = 0$ for some $s < t$, the argument below applied at time s (with any $\lambda \in (0, 1]$) gives $a_s - \max_{s' \le s} \beta_{s'} \le \delta$, and the claim holds at all later times because $a_t \le a_s$. So assume neither happens.

Volume. Each $\mathcal{Y}_t$ is the intersection of $\mathcal{Y}_{t-1}$ with a half-space whose boundary passes through the center of gravity y_t of $\mathcal{Y}_{t-1}$. By Lemma 10, $\operatorname{vol} \mathcal{Y}_t \le (1 - e^{-1}) \operatorname{vol} \mathcal{Y}_{t-1}$, so $\operatorname{vol} \mathcal{Y}_t \le (1 - e^{-1})^t \operatorname{vol} \mathcal{Y}$.

A point outside $\mathcal{Y}_t$ with a large value of the model. Let $U = \min_{s \le t} \ell_s$, let $y^\circ \in \arg\max_{\mathcal{Y}} U$, so that $U(y^\circ) = a_t$, and fix $\lambda \in \big((1 - e^{-1})^{t/m}, 1\big]$. The set $S = y^\circ + \lambda(\mathcal{Y} - y^\circ)$ is contained in $\mathcal{Y}$ by convexity, and $\operatorname{vol} S = \lambda^m \operatorname{vol} \mathcal{Y} > \operatorname{vol} \mathcal{Y}_t$. Hence there is a point $y' \in S \setminus \mathcal{Y}_t$, and we can write $y' = y^\circ + \lambda(y'' - y^\circ)$ with $y'' \in \mathcal{Y}$. On the one hand, U is concave and M_y-Lipschitz (a minimum of affine functions with slopes of norm at most M_y), and $U(y'') \ge U(y^\circ) - M_y \|y'' - y^\circ\| \ge a_t - B$. Therefore

$$U(y') \ \ge \ (1 - \lambda) U(y^\circ) + \lambda U(y'') \ \ge \ a_t - \lambda B.$$

On the other hand, $y' \in \mathcal{Y} \setminus \mathcal{Y}_t$, so some cut excluded it: $\langle g_s, y' - y_s\rangle < 0$ for some $s \le t$. For this s,

$$U(y') \ \le \ \ell_s(y') = w_s + \langle g_s, y' - y_s\rangle \ < \ w_s \ \le \ \beta_s + \delta.$$

Comparing the two bounds gives $a_t - \beta_s < \lambda B + \delta$. Letting $\lambda \downarrow (1 - e^{-1})^{t/m}$ proves the claim.

Exact values. If the queries are exact, take $w_t = \beta_t = \psi(y_t)$ and $\delta = 0$, with g_t a supergradient of ψ at y_t, so that $\ell_t \ge \psi$ by concavity. Since $\max_{\mathcal{Y}} \psi \le a_t$, the best query point y_s satisfies $\max_{\mathcal{Y}} \psi - \psi(y_s) \le B(1 - e^{-1})^{t/m}$, which is at most δ after $t = \lceil m \log(B/\delta) / \log \frac{e}{e-1} \rceil$ queries. □

Remark 1 (implementing centers of gravity). In the oracle model, centers of gravity are free. If the exact center of gravity is replaced by an approximate one computed by a random walk (Bertsimas & Vempala, 2004), every central cut still removes a constant fraction of the volume with high probability, which changes only the absolute constants in Theorem 1. In general, computing the center of gravity exactly is hard (Rademacher, 2007). For $m \le 3$, exact centers of gravity of polytopes are cheap to compute, and for $m = 1$ the center of gravity is the midpoint of a segment.

A.5 RECOVERY OF A PRIMAL POINT

Lemma 12 (recovery). *Let $f(\cdot, y)$ be convex on $\mathcal{X}$ for every $y \in \mathcal{Y}$. Let $u_1, \ldots, u_N \in \mathcal{X}$, and let the planes $\ell_i(y) = v_i + \langle g_i, y - y_i\rangle$ satisfy $\ell_i \ge f(u_i, \cdot)$ on $\mathcal{Y}$. Put $a = \max_{\mathcal{Y}} \min_i \ell_i$. Then there is $\lambda \in \Delta_N$ such that $\bar{x} = \sum_i \lambda_i u_i$ satisfies $P(\bar{x}) \le a$. Moreover, λ can be computed from the planes alone, without further oracle calls.*

Proof. For fixed y, the minimum of the linear function $\lambda \mapsto \sum_i \lambda_i \ell_i(y)$ over the simplex Δ_N is attained at a vertex, so $\min_i \ell_i(y) = \min_{\lambda \in \Delta_N} \sum_i \lambda_i \ell_i(y)$. The function $(\lambda, y) \mapsto \sum_i \lambda_i \ell_i(y)$ is continuous and affine in each argument, and Δ_N and $\mathcal{Y}$ are compact and convex. By the minimax theorem (Sion, 1958),

$$a = \max_{y \in \mathcal{Y}} \min_{\lambda \in \Delta_N} \sum_i \lambda_i \ell_i(y) = \min_{\lambda \in \Delta_N} \max_{y \in \mathcal{Y}} \sum_i \lambda_i \ell_i(y).$$

Let λ attain the minimum on the right; it depends only on the planes. For $\bar{x} = \sum_i \lambda_i u_i \in \mathcal{X}$ and every $y \in \mathcal{Y}$, convexity of $f(\cdot, y)$ and $\ell_i \geq f(u_i, \cdot)$ give

$$f(\bar{x}, y) \;\leq\; \sum_i \lambda_i f(u_i, y) \;\leq\; \sum_i \lambda_i \ell_i(y) \;\leq\; a.$$

Taking the maximum over $y \in \mathcal{Y}$ gives $P(\bar{x}) \leq a$. □

B Proof of Theorem 1 and Corollary 2

Throughout this appendix Assumption 1 holds and $0 < \varepsilon < 2B$. Volumes of subsets of $\mathcal{Y}$ are taken in $\operatorname{aff} \mathcal{Y}$, which has dimension m. As usual, bounds of the form $O(\log \frac{B}{\varepsilon})$ absorb additive constants, that is, they should be read as $O(1 + \log \frac{B}{\varepsilon})$; note that $1 + \log \frac{B}{\varepsilon} > 1 - \log 2 > 0$ because $\varepsilon < 2B$.

Notation for one run. We number the queries of the main loop of Algorithm 2 by $t = 1, 2, \ldots, N$. In query t the algorithm

1. takes the center of gravity y_t of the localizer $S_t = \{y \in \mathcal{Y} : U_t(y) \geq h\}$, where U_t is the upper model before query t and h is the current threshold;
2. runs $(u_t, z_t, A) = \textsc{Block}(f(\cdot, y_t), q_b, K)$, which costs K calls to O_x;
3. calls $(v_t, g_t) = \mathsf{O}_y(u_t, y_t)$ and adds the plane $\ell_t(y) = v_t + \langle g_t, y - y_t \rangle$ to the upper model;
4. if $v_t > h$, replaces the lower model by the transported one from Corollary 7, and otherwise leaves it unchanged.

Here $A = K(K+3)/(8\kappa)$ is the same in every query, and we write $\vartheta = A/(1+A)$. Following (7), we call query t a *cut* if $v_t \leq h$ and a *transport* if $v_t > h$. The quantities $a = \max_{\mathcal{Y}} U$ and b always refer to their current values.

The proof has four parts. Lemma 13 shows that the output is always certified by the width $a - b$. Lemmas 14 and 15 control the epochs and the position of the threshold. Lemma 16 bounds the number of transports, and Lemmas 17 and 18 bound the number of cuts. The count of oracle calls then follows.

B.1 Invariants and correctness

Lemma 13 (invariants)**.** *At every moment of the run the following hold.*

(i) *Every plane satisfies $\ell_i \geq f(u_i, \cdot) \geq \phi$ on $\mathcal{Y}$, and $U = \min_i \ell_i$ is concave and M_y-Lipschitz on $\mathcal{Y}$. In particular, $a \geq \max_{\mathcal{Y}} \phi$.*

(ii) *(b, x_b, y_b) is a lower model: $x_b \in \mathcal{X}$, $y_b \in \mathcal{Y}$, and $f(x, y_b) \geq b + \frac{\sigma}{2}\|x - x_b\|^2$ for all $x \in \mathcal{X}$. In particular, $\phi(y_b) \geq b$.*

(iii) *For the point $\bar{x}$ of Lemma 12, $\operatorname{Gap}(\bar{x}, y_b) \leq a - b$.*

(iv) *a never increases.*

In particular, when the algorithm stops, $a - b \leq \varepsilon$ and the output satisfies $\operatorname{Gap}(\bar{x}, y_b) \leq \varepsilon$.

Proof. (i) Each plane comes from a call $\mathsf{O}_y(u_i, y_i)$, so $\ell_i \geq f(u_i, \cdot)$ on $\mathcal{Y}$ by concavity of $f(u_i, \cdot)$, and $f(u_i, \cdot) \geq \min_x f(x, \cdot) = \phi$. Each ℓ_i is affine with slope g_i, $\|g_i\| \leq M_y$, hence M_y-Lipschitz; a minimum of concave M_y-Lipschitz functions is concave and M_y-Lipschitz. Finally, $U \geq \phi$ gives $a = \max_{\mathcal{Y}} U \geq \max_{\mathcal{Y}} \phi$.

(ii) The initial triple is a lower model by Corollary 9. The lower model changes only in transports, and then the new triple is a lower model by Corollary 7. The bound on $\phi(y_b)$ follows by minimizing both sides over $x \in \mathcal{X}$: $\phi(y_b) = \min_x f(x, y_b) \geq \min_x \left(b + \frac{\sigma}{2}\|x - x_b\|^2\right) = b$.

(iii) By Lemma 12, applied to all planes collected so far (including ℓ_0 from Corollary 9), $P(\bar{x}) \leq a$. Together with (ii), $\mathrm{Gap}(\bar{x}, y_b) = P(\bar{x}) - \phi(y_b) \leq a - b$.

(iv) Adding a plane can only decrease $U = \min_i \ell_i$, and therefore $a = \max_{\mathcal{Y}} U$.

The last claim follows from (iii) and the stopping rule $a - b \leq \varepsilon$. □

Note that b is not monotone for an arbitrary transport, since $b^+ - b = \vartheta(v_t - b)$ is negative when $v_t < b$. It becomes monotone because transports happen only above the threshold, which lies above b (Lemma 14).

B.2 Epochs and the threshold

The run is divided into epochs. Epoch $j = 0, 1, \ldots$ starts at a moment when the algorithm sets $W \leftarrow a - b$ and $h \leftarrow b + W/2$; let a_j and b_j be the values of a and b at that moment, and let

$$W_j = a_j - b_j, \qquad h_j = b_j + \frac{W_j}{2}.$$

A new epoch starts before a query as soon as $a - b \leq \frac{2}{3}W_j$, and the algorithm stops as soon as $a - b \leq \varepsilon$. Consequently, immediately before every query of epoch j,

$$a - b \;>\; \max\left\{\varepsilon,\; \frac{2}{3}W_j\right\}. \tag{19}$$

For the first query of the epoch this holds because $a - b = W_j > \varepsilon$.

Lemma 14 (position of the threshold)**.** *Immediately before every query of epoch j,*

$$h_j - b \;>\; \frac{W_j}{6} \qquad \textit{and} \qquad a - h_j \;>\; \frac{W_j}{6}. \tag{20}$$

Moreover, b never decreases, and $a \leq a_j$, $b \geq b_j$ during epoch j.

Proof. We first prove (20) under the assumption that $a \leq a_j$ and $b \geq b_j$ at the moment in question, and then show that these inequalities indeed hold.

By (19) and $a \leq a_j$,

$$b \;<\; a - \frac{2}{3}W_j \;\leq\; a_j - \frac{2}{3}W_j = b_j + \frac{1}{3}W_j,$$

hence

$$h_j - b \;>\; \left(b_j + \frac{W_j}{2}\right) - \left(b_j + \frac{W_j}{3}\right) = \frac{W_j}{6}.$$

Since $b \geq b_j$, we also have $h_j - b \leq h_j - b_j = \frac{W_j}{2}$. Together with (19),

$$a - h_j = (a - b) - (h_j - b) \;>\; \frac{2}{3}W_j - \frac{1}{2}W_j = \frac{W_j}{6}.$$

It remains to check $a \leq a_j$ and $b \geq b_j$. The first holds because a never increases (Lemma 13(iv)). For the second, we show by induction over the queries of the epoch that b does not decrease. At the start of the epoch $b = b_j$. If $b \geq b_j$ before a query, then (20) holds before it, so $h_j > b$. If the query is a transport, then $v_t > h_j > b$ and $b^+ - b = \vartheta(v_t - b) > 0$. If it is a cut, b does not change. Hence $b \geq b_j$ before the next query as well. The value of b does not change between epochs, so b never decreases over the whole run. □

Lemma 15 (number of epochs)**.** $W_{j+1} \leq \frac{2}{3}W_j$*, and the number of epochs satisfies*

$$N_{\mathrm{ep}} \;\leq\; 1 + \left\lceil \frac{\log(2B/\varepsilon)}{\log(3/2)} \right\rceil = O\Big(\log \frac{B}{\varepsilon}\Big).$$

Proof. Epoch $j+1$ starts when $a-b \le \frac{2}{3}W_j$, and at that moment $W_{j+1} = a-b$. Hence $W_j \le (\frac{2}{3})^j W_0 \le (\frac{2}{3})^j \cdot 2B$, where $W_0 \le 2B$ by Corollary 9. Every epoch starts with $W_j = a-b > \varepsilon$, since otherwise the algorithm would have stopped. Therefore the last epoch $N_{\mathrm{ep}}-1$ satisfies $\varepsilon < (\frac{2}{3})^{N_{\mathrm{ep}}-1} \cdot 2B$, that is, $N_{\mathrm{ep}} - 1 < \log(2B/\varepsilon)/\log(3/2)$. □

B.3 Transports

Lemma 16 (number of transports)**.** *Every epoch contains at most $\lceil 2/\vartheta \rceil$ transports, and*

$$\frac{2}{\vartheta} = 2\Big(1+\frac{1}{A}\Big) \;\le\; 2 + \frac{16\kappa}{K^2}.$$

Proof. Each transport increases b by a fixed fraction of W_j. In a transport of epoch j, $v_t > h_j$, so by Corollary 7 and (20)

$$b^+ - b = \vartheta(v_t - b) \;>\; \vartheta(h_j - b) \;>\; \frac{\vartheta W_j}{6}.$$

The total increase of b within an epoch is limited. Before every query of epoch j we showed in the proof of Lemma 14 that $b < b_j + \frac{1}{3}W_j$. Suppose epoch j contains n transports. Before the last of them, b has already grown by more than $(n-1)\frac{\vartheta W_j}{6}$ and is still below $b_j + \frac{1}{3}W_j$. Hence $(n-1)\frac{\vartheta}{6} < \frac{1}{3}$, that is, $n < 1 + \frac{2}{\vartheta}$, and so $n \le \lceil 2/\vartheta \rceil$.

The bound on $2/\vartheta$. By definition, $\frac{1}{\vartheta} = \frac{1+A}{A} = 1 + \frac{1}{A}$, and $A = \frac{K(K+3)}{8\kappa} \ge \frac{K^2}{8\kappa}$, so $\frac{1}{A} \le \frac{8\kappa}{K^2}$. □

B.4 Cuts

The number of cuts is controlled by the volume of the localizer. Let

$$V(t) = \operatorname{vol} S_t, \qquad S_t = \{y \in \mathcal{Y} :\; U_t(y) \ge h_{j(t)}\},$$

where $j(t)$ is the epoch of query t, and U_t is the upper model immediately before query t.

Lemma 17 (volume of the localizer)**.** *The following hold.*

(a) Lower bound. $V(t) \ge \big(\frac{\varepsilon}{6B}\big)^m \operatorname{vol}\mathcal{Y}$ *for every t. In particular, S_t is a convex body in* $\operatorname{aff}\mathcal{Y}$*, and its center of gravity y_t is well defined and lies in $\mathcal{Y}$.*

(b) Cuts shrink the localizer. *If query t is a cut and query $t+1$ belongs to the same epoch, then $V(t+1) \le (1-e^{-1})V(t)$.*

(c) Transports do not enlarge it. *If query t is a transport and query $t+1$ belongs to the same epoch, then $V(t+1) \le V(t)$.*

(d) A change of epoch enlarges it by at most 6^m. *If query t is the last query of epoch j and query $t+1$ is the first query of epoch $j+1$, then $V(t+1) \le 6^m V(t)$.*

Proof. Write $j = j(t)$, $U = U_t$, and let a be the value of $\max_{\mathcal{Y}} U_t$.

(a) Let $y^\circ \in \arg\max_{\mathcal{Y}} U$ and $\lambda = \min\{1, (a-h_j)/B\}$. Take any $y'' \in \mathcal{Y}$ and put $y' = y^\circ + \lambda(y'' - y^\circ)$, which lies in $\mathcal{Y}$ by convexity. By Lemma 13(i), U is M_y-Lipschitz, so $U(y'') \ge a - M_y\|y'' - y^\circ\| \ge a - B$. By concavity of U,

$$U(y') \;\ge\; (1-\lambda)U(y^\circ) + \lambda U(y'') \;\ge\; (1-\lambda)a + \lambda(a-B) = a - \lambda B \;\ge\; h_j,$$

where the last step uses $\lambda \le (a-h_j)/B$. Hence $y' \in S_t$, that is, S_t contains the homothetic copy $y^\circ + \lambda(\mathcal{Y} - y^\circ)$ of $\mathcal{Y}$, and $V(t) \ge \lambda^m \operatorname{vol}\mathcal{Y}$. By (20), $a - h_j > W_j/6$, and $W_j > \varepsilon$ by (19); hence $a - h_j > \varepsilon/6$. Since also $\varepsilon/(6B) < 1$, we get $\lambda \ge \varepsilon/(6B)$. Finally, S_t is convex (a superlevel set of a concave function intersected with $\mathcal{Y}$), closed, and has positive volume, so it is a convex body.

(b) The threshold is the same for queries t and $t+1$, and $U_{t+1} = \min\{U_t, \ell_t\}$. Hence

$$S_{t+1} = S_t \cap \{y \in \mathcal{Y} :\; \ell_t(y) \ge h_j\}.$$

If $\ell_t(y) \geq h_j$, then $\langle g_t, y - y_t \rangle = \ell_t(y) - v_t \geq h_j - v_t \geq 0$, because query t is a cut. So $S_{t+1} \subseteq S_t \cap H$ with $H = \{y : \langle g_t, y - y_t \rangle \geq 0\}$. The boundary of H passes through the center of gravity y_t of S_t, and Lemma 10 applied in $\operatorname{aff} \mathcal{Y}$ gives $V(t+1) \leq (1 - e^{-1})V(t)$.

This argument needs $H \cap \operatorname{aff} \mathcal{Y}$ to be a proper half-space of $\operatorname{aff} \mathcal{Y}$, that is, g_t must not be orthogonal to $\operatorname{aff} \mathcal{Y}$. If it is, then ℓ_t is constant on $\mathcal{Y}$ and equal to $v_t \leq h_j$, so after query t we have $a \leq v_t \leq h_j$ and $a - b \leq h_j - b \leq h_j - b_j = \frac{W_j}{2} < \frac{2}{3} W_j$. Then query t is the last query of its epoch, and (b) does not apply to it.

(c) Again the threshold is unchanged and $U_{t+1} \leq U_t$, so $S_{t+1} \subseteq S_t$.

(d) Let a' be the value of a before query t and $y^\circ \in \arg\max_{\mathcal{Y}} U_t$. Since $U_{t+1} \leq U_t$,

$$S_{t+1} = \{U_{t+1} \geq h_{j+1}\} \subseteq \{y \in \mathcal{Y} : \, U_t(y) \geq h_{j+1}\}.$$

If $h_{j+1} \geq h_j$, the right-hand side is contained in S_t and we are done. Otherwise $h_{j+1} < h_j < a'$, where the last inequality is (20). Put

$$\lambda = \frac{a' - h_j}{a' - h_{j+1}} \in (0, 1].$$

Take $y \in \mathcal{Y}$ with $U_t(y) \geq h_{j+1}$. By concavity,

$$U_t\big(y^\circ + \lambda(y - y^\circ)\big) \;\geq\; (1 - \lambda)a' + \lambda h_{j+1} = a' - \lambda(a' - h_{j+1}) = h_j,$$

so $y^\circ + \lambda(y - y^\circ) \in S_t$, that is, $y \in y^\circ + \lambda^{-1}(S_t - y^\circ)$. Hence $\{U_t \geq h_{j+1}\}$ is contained in a homothetic copy of S_t with ratio λ^{-1}, and $V(t+1) \leq \lambda^{-m} V(t)$. It remains to bound λ^{-1}. For the numerator, $a' \leq a_j$ and $h_{j+1} \geq b_{j+1} \geq b_j$ (Lemma 14), so $a' - h_{j+1} \leq a_j - b_j = W_j$. For the denominator, $a' - h_j > W_j/6$ by (20). Therefore $\lambda^{-1} < 6$. □

Lemma 18 (number of cuts)**.** *The total number C of cuts satisfies*

$$C \;\leq\; N_{\mathrm{ep}} + \frac{m \log \frac{6B}{\varepsilon} + m N_{\mathrm{ep}} \log 6}{\log \frac{e}{e-1}} = O\Big(m \log \frac{B}{\varepsilon}\Big).$$

Proof. Follow $V(t)$ from $t = 1$ to $t = N$. At the start, $V(1) \leq \operatorname{vol} \mathcal{Y}$. For each $t < N$, Lemma 17(b)–(d) gives:

- a factor at most $1 - e^{-1}$ if query t is a cut and not the last query of its epoch;
- a factor at most 1 if query t is a transport and not the last query of its epoch;
- a factor at most 6^m if query t is the last query of its epoch.

Every epoch has exactly one last query, and query N is the last query of the last epoch. Hence at most N_{ep} cuts are last queries of their epochs, at least $C - N_{\mathrm{ep}}$ cuts contribute the factor $1 - e^{-1}$, and at most $N_{\mathrm{ep}} - 1 \leq N_{\mathrm{ep}}$ changes of epoch contribute the factor 6^m. Combining this with the lower bound of Lemma 17(a) at $t = N$,

$$\Big(\frac{\varepsilon}{6B}\Big)^m \operatorname{vol} \mathcal{Y} \;\leq\; V(N) \;\leq\; (1 - e^{-1})^{C - N_{\mathrm{ep}}} \, 6^{m N_{\mathrm{ep}}} \operatorname{vol} \mathcal{Y}.$$

Dividing by $\operatorname{vol} \mathcal{Y}$ and taking logarithms,

$$(C - N_{\mathrm{ep}}) \log \frac{e}{e-1} \;\leq\; m \log \frac{6B}{\varepsilon} + m N_{\mathrm{ep}} \log 6,$$

which is the claim. The $O(\cdot)$ bound follows from $N_{\mathrm{ep}} = O(\log \frac{B}{\varepsilon})$ (Lemma 15). □

B.5 Proof of Theorem 1

Correctness. By Lemma 13, the output satisfies $\mathrm{Gap}(\bar{x}, y_b) \le \varepsilon$. This holds for every $K \ge 1$: the block length enters only the number of queries.

Number of queries. Every query is a cut or a transport. By Lemmas 15, 16 and 18,

$$N \;\le\; C + N_{\mathrm{ep}}\Big\lceil \frac{2}{\vartheta} \Big\rceil \;\le\; O\Big(m \log \frac{B}{\varepsilon}\Big) + N_{\mathrm{ep}}\Big(3 + \frac{16\kappa}{K^2}\Big) \;=\; O\Big(\Big(m + \frac{\kappa}{K^2}\Big) \log \frac{B}{\varepsilon}\Big),$$

where the constant term $3N_{\mathrm{ep}}$ is absorbed into mN_{ep} because $m \ge 1$. In particular, the algorithm terminates.

Oracle calls. Initialization costs I_0 calls to O_x and one call to O_y (Corollary 9), and every query costs K calls to O_x and one call to O_y. Therefore

$$T_y = 1 + N = O\Big(\Big(m + \frac{\kappa}{K^2}\Big) \log \frac{B}{\varepsilon}\Big), \qquad T_x = I_0 + KN = I_0 + O\Big(\Big(mK + \frac{\kappa}{K}\Big) \log \frac{B}{\varepsilon}\Big),$$

where we used $K \cdot \frac{\kappa}{K^2} = \frac{\kappa}{K}$. This is the first part of Theorem 1.

The default block length. Let $K = \lceil \sqrt{8\kappa/m} \rceil$. Then $K \ge \sqrt{8\kappa/m}$, so $\frac{\kappa}{K^2} \le \frac{m}{8}$ and $\frac{\kappa}{K} \le \sqrt{\frac{m\kappa}{8}}$. Also $K \le \sqrt{8\kappa/m} + 1$, so $mK \le \sqrt{8m\kappa} + m$. Substituting,

$$T_y = O\Big(m \log \frac{B}{\varepsilon}\Big), \qquad T_x = I_0 + O\Big(\big(m + \sqrt{m\kappa}\big) \log \frac{B}{\varepsilon}\Big),$$

which is the second part of Theorem 1. □

B.6 Proof of Corollary 2

For $m = 1$ the default block length is $K = \lceil \sqrt{8\kappa} \rceil$. Since $\kappa \ge 1$, we have $m + \sqrt{m\kappa} = 1 + \sqrt{\kappa} \le 2\sqrt{\kappa}$, and the second part of Theorem 1 gives

$$T_y = O\Big(\log \frac{B}{\varepsilon}\Big), \qquad T_x = I_0 + O\Big(\sqrt{\kappa} \log \frac{B}{\varepsilon}\Big) = O\Big(\sqrt{\kappa}\Big[1 + \log_+ \frac{\mu R_x^2}{B} + \log \frac{B}{\varepsilon}\Big]\Big),$$

where we inserted I_0 from Corollary 9. □

Remark 2 (the block length and the limit of this analysis). No correctness argument above uses the value of K: the transported triple is a lower model for every K (Corollary 7), and the certificate $\mathrm{Gap} \le a - b$ can be checked at any moment (Lemma 13). The block length affects only the number of queries, through the trade-off between the mK gradients spent on cuts and the κ/K gradients spent on transports. The choice $K \asymp \sqrt{\kappa/m}$ balances these two terms, so to go below $\sqrt{m\kappa}$ in general one needs a different measure of progress. Under strong concavity in y, the method of Section 4 provides one.

C Proof of Theorem 4: Accelerated two-point method

Throughout this appendix Assumptions 1 and 2 hold. Recall that $f(x, \cdot)$ is ν-strongly concave, $\nabla_x f$ is ℓ-Lipschitz in y, and $\chi = \ell^2/(\mu\nu)$. We use the constants

$$\gamma_\chi := \frac{\ell^2}{\nu} = \mu\chi, \qquad \bar{\chi} := 1 + \chi, \qquad \Gamma := \mu + \gamma_\chi = \mu\bar{\chi}.$$

Since $\mathcal{X}$ and $\mathcal{Y}$ are compact and f is strongly convex in x and strongly concave in y, the saddle point $(x^\star, y^\star)$ exists and is unique, and

$$P^\star := P(x^\star) = f(x^\star, y^\star) = \phi(y^\star).$$

For $x \in \mathcal{X}$, let $y^\star(x) = \arg\max_{\mathcal{Y}} f(x, \cdot)$; the maximizer is unique by strong concavity.

We will use two standard consequences of strong convexity and concavity on a convex set. If x' minimizes a μ-strongly convex function h over $\mathcal{X}$, then first-order optimality gives $\langle \nabla h(x'), x - x' \rangle \ge 0$ for $x \in \mathcal{X}$, and therefore

$$h(x) - h(x') \;\ge\; \frac{\mu}{2}\|x - x'\|^2 \qquad \forall x \in \mathcal{X}. \tag{21}$$

Algorithm 3 Two-point acceleration under strong concavity

Require: $\varepsilon > 0$; constants L, μ, χ, B; a solver $\mathsf{MaxY}(x, \delta)$ for the problem in y
Ensure: $(\hat{x}, \hat{y})$ with $\mathrm{Gap}(\hat{x}, \hat{y}) \le \varepsilon$

$\bar{\chi} \leftarrow 1 + \chi$; $\Gamma \leftarrow \mu\bar{\chi}$; $L_\Phi \leftarrow 1 + L/(4\Gamma)$
$k \leftarrow \lceil 2\sqrt{L_\Phi} \rceil$; $A_\Phi \leftarrow k(k+3)/(4L_\Phi)$; $\theta \leftarrow A_\Phi/(1 + A_\Phi)$
$K \leftarrow \lceil 8\sqrt{\bar{\chi}/\theta} \rceil$; $c_\delta \leftarrow \bar{\chi}/(4K^3)$; $E \leftarrow 2B$
Find w with $P(w) - P^\star \le E$ (Lemma 8)
while $E > \varepsilon/(6\bar{\chi}^2)$ **do**
  $A \leftarrow 0$; $u, z \leftarrow w$; $\delta \leftarrow c_\delta E$
  for $i = 1, \dots, K$ **do**
    Find $a > 0$ from $a^2 = \theta(A + a)/(4\Gamma)$; $A^+ \leftarrow A + a$
    $w_i(p) \leftarrow (Au + ap)/A^+$; $v \leftarrow w_i(z)$
    $y_i \leftarrow \mathsf{MaxY}(v, \delta)$ ▷ $P(v) - f(v, y_i) \le \delta$
    $\Phi_i(p) \leftarrow (A^+/\theta)\, f(w_i(p), y_i) + \frac{1}{2}\|p - z\|^2$
    $(p_i, z^+, _) \leftarrow \textsc{Block}\big(\Phi_i,\ \frac{1}{2}\|\cdot - z\|^2,\ k\big)$ ▷ $\mu_F = 1 + \frac{\mu}{4\Gamma}$, $L_F = L_\Phi$
    $u \leftarrow w_i(p_i)$; $z \leftarrow z^+$; $A \leftarrow A^+$
  $w \leftarrow u$; $E \leftarrow E/2$
$\hat{x} \leftarrow w$; $\hat{y} \leftarrow \mathsf{MaxY}(\hat{x}, c_\delta E)$; **return** $(\hat{x}, \hat{y})$

Symmetrically, if y' maximizes a ν-strongly concave function h over $\mathcal{Y}$, then $h(y') - h(y) \ge \frac{\nu}{2}\|y - y'\|^2$ for all $y \in \mathcal{Y}$. In particular, $P = \max_y f(\cdot, y)$ is μ-strongly convex as a maximum of μ-strongly convex functions, and $P(x) - P^\star \ge \frac{\mu}{2}\|x - x^\star\|^2$.

The method analyzed in this appendix is Algorithm 3. Its outer loop halves an upper bound E on the primal error $P(w) - P^\star$; each halving is one *outer block* of K steps, and each step makes one call to a solver MaxY of the problem in y and one run of Algorithm 1 on a proximal subproblem.

C.1 Cross-Lipschitz estimate

The coupling assumption enters the proof only through the following inequality. For $x, v \in \mathcal{X}$ put $\psi_{x,v}(y) = f(x, y) - f(v, y)$. Then

$$\big|\psi_{x,v}(y) - \psi_{x,v}(y')\big| \;\le\; \ell\,\|x - v\|\,\|y - y'\| \qquad \forall y, y' \in \mathcal{Y}. \tag{22}$$

Indeed, let $v_t = v + t(x - v) \in \mathcal{X}$ for $t \in [0, 1]$. By the fundamental theorem of calculus applied to $t \mapsto f(v_t, y)$,

$$\psi_{x,v}(y) = f(x, y) - f(v, y) = \int_0^1 \langle \nabla_x f(v_t, y), x - v\rangle\, dt,$$

and similarly for y'. Subtracting and using the Cauchy–Schwarz inequality and Assumption 2,

$$\big|\psi_{x,v}(y) - \psi_{x,v}(y')\big| = \Big|\int_0^1 \langle \nabla_x f(v_t, y) - \nabla_x f(v_t, y'), x - v\rangle\, dt\Big| \le \ell\|y - y'\|\|x - v\|.$$

C.2 A two-sided model of the primal function

Lemma 19 (two-sided model)**.** *Let $v \in \mathcal{X}$, and let $\tilde{y} \in \mathcal{Y}$ satisfy $P(v) - f(v, \tilde{y}) \le \delta$. Then for all $x \in \mathcal{X}$*

$$f(x, \tilde{y}) \;\le\; P(x) \;\le\; f(x, \tilde{y}) + \gamma_\chi\|x - v\|^2 + 2\delta.$$

Proof. The left inequality is the definition of P. For the right one, fix $x \in \mathcal{X}$ and write $d = \|x - v\|$, $y_x = y^\star(x)$, $y_v = y^\star(v)$, and $\psi = \psi_{x,v}$.

Step 1: two consequences of strong concavity at v. Since y_v maximizes the ν-strongly concave function $f(v, \cdot)$,

$$\frac{\nu}{2}\|\tilde{y} - y_v\|^2 \le P(v) - f(v, \tilde{y}) \le \delta \quad\Longrightarrow\quad \|\tilde{y} - y_v\|^2 \le \frac{2\delta}{\nu},$$

and, with $y = y_x$,

$$f(v, y_x) \;\le\; P(v) - \frac{\nu}{2}\|y_x - y_v\|^2.$$

Step 2: splitting the error. Since $f(x,\cdot) = f(v,\cdot) + \psi$,

$$P(x) - f(x,\tilde{y}) = f(x,y_x) - f(x,\tilde{y}) = \big[\psi(y_x) - \psi(\tilde{y})\big] + \big[f(v,y_x) - f(v,\tilde{y})\big].$$

By (22) and the triangle inequality, the first bracket is at most $\ell d\,\|y_x - \tilde{y}\| \le \ell d\big(\|y_x - y_v\| + \|y_v - \tilde{y}\|\big)$. By Step 1 and $P(v) - f(v,\tilde{y}) \le \delta$, the second bracket is at most $\delta - \frac{\nu}{2}\|y_x - y_v\|^2$. Hence, with $s = \|y_x - y_v\|$,

$$P(x) - f(x,\tilde{y}) \;\le\; \Big[\ell d s - \frac{\nu}{2}s^2\Big] + \ell d\,\|y_v - \tilde{y}\| + \delta.$$

Step 3: bounding the two terms. The quadratic in s is maximized at $s = \ell d/\nu$, so $\ell d s - \frac{\nu}{2}s^2 \le \frac{\ell^2 d^2}{2\nu}$. By the inequality $\alpha\beta \le \frac{\alpha^2}{2\nu} + \frac{\nu\beta^2}{2}$ and Step 1,

$$\ell d\,\|y_v - \tilde{y}\| \;\le\; \frac{\ell^2 d^2}{2\nu} + \frac{\nu}{2}\|y_v - \tilde{y}\|^2 \;\le\; \frac{\ell^2 d^2}{2\nu} + \delta.$$

Adding up, $P(x) - f(x,\tilde{y}) \le \frac{\ell^2}{\nu}d^2 + 2\delta = \gamma_\chi\|x - v\|^2 + 2\delta$. □

The lemma says that one approximate maximizer $\tilde{y}$ computed at v gives a model of P that is exact from below and γ_χ-smooth from above up to 2δ. This is an inexact model of P in the sense of Devolder et al. (2014); Stonyakin et al. (2021).

C.3 From primal accuracy to the full gap

Lemma 20 (from primal accuracy to the gap)**.** *If $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ satisfy $P(x) - P^\star \le E$ and $P(x) - f(x,y) \le \delta$, then*

$$\mathrm{Gap}(x,y) \;\le\; E + \bar{\chi}(2\delta + 4\chi E).$$

Proof. Split the gap as $\mathrm{Gap}(x,y) = \big(P(x) - P^\star\big) + \big(P^\star - \phi(y)\big)$. The first term is at most E, so we bound the second one.

Step 1: how far y is from $y^\star$. Let $D = f(x^\star,y^\star) - f(x^\star,y)$. Since $y^\star$ maximizes the ν-strongly concave function $f(x^\star,\cdot)$, we have $D \ge \frac{\nu}{2}\|y - y^\star\|^2$. To bound D from above, add and subtract $f(x,y^\star)$ and $f(x,y)$:

$$D = \underbrace{\big[f(x,y^\star) - f(x,y)\big]}_{\le P(x) - f(x,y) \le \delta} + \underbrace{\big[\psi_{x,x^\star}(y) - \psi_{x,x^\star}(y^\star)\big]}_{\le \ell\|y - y^\star\|\,\|x - x^\star\| \text{ by (22)}}.$$

Now $\|y - y^\star\|^2 \le 2D/\nu$ and $\|x - x^\star\|^2 \le 2E/\mu$ (strong convexity of P). Hence

$$\ell\|y - y^\star\|\|x - x^\star\| \;\le\; \ell\sqrt{\frac{2D}{\nu}\cdot\frac{2E}{\mu}} = 2\sqrt{\chi D E} \;\le\; \frac{D}{2} + 2\chi E,$$

where the last step is $2\sqrt{\alpha\beta} \le \alpha + \beta$ with $\alpha = D/2$ and $\beta = 2\chi E$. Thus $D \le \delta + \frac{D}{2} + 2\chi E$, that is,

$$D \;\le\; 2\delta + 4\chi E.$$

Step 2: from D to $P^\star - \phi(y)$. Let $x_y = \arg\min_{\mathcal{X}} f(\cdot,y)$, so that $\phi(y) = f(x_y,y)$. Then

$$P^\star - \phi(y) = f(x^\star,y^\star) - f(x_y,y) = D + \big[f(x^\star,y) - f(x_y,y)\big].$$

By μ-strong convexity of $f(\cdot,y)$,

$$f(x^\star,y) - f(x_y,y) \;\le\; \langle\nabla_x f(x^\star,y), x^\star - x_y\rangle - \frac{\mu}{2}\|x^\star - x_y\|^2.$$

Since $x^\star$ minimizes $f(\cdot,y^\star)$ over $\mathcal{X}$, first-order optimality gives $\langle\nabla_x f(x^\star,y^\star), x^\star - x_y\rangle \le 0$, and we may subtract this term:

$$f(x^\star,y) - f(x_y,y) \;\le\; \langle\nabla_x f(x^\star,y) - \nabla_x f(x^\star,y^\star), x^\star - x_y\rangle - \frac{\mu}{2}\|x^\star - x_y\|^2 \;\le\; \ell\|y - y^\star\|s - \frac{\mu}{2}s^2,$$

with $s = \|x^\star - x_y\|$. Maximizing over s and using Step 1,

$$f(x^\star,y) - f(x_y,y) \;\le\; \frac{\ell^2\|y - y^\star\|^2}{2\mu} \;\le\; \frac{\ell^2}{2\mu}\cdot\frac{2D}{\nu} = \chi D.$$

Therefore $P^\star - \phi(y) \le (1+\chi)D = \bar{\chi}D \le \bar{\chi}(2\delta + 4\chi E)$. □

C.4 THE INNER TWO-POINT INEQUALITY

Each step of the outer method minimizes a proximal subproblem only approximately, by one run of Algorithm 1. The next lemma turns the guarantee of Lemma 6 into the form needed by the outer analysis. It returns two points: p, at which the objective is evaluated, and z, around which the quadratic lower bound is centered.

Lemma 21 (two-point inequality)**.** *Let g be convex on $\mathcal{X}$, let $z_- \in \mathcal{X}$ and $\theta_0 > 0$, and suppose that*

$$\Phi(z) = \frac{1}{\theta_0} g(z) + \frac{1}{2}\|z - z_-\|^2$$

is σ_Φ-strongly convex and L_Φ-smooth on $\mathcal{X}$. Run the method of Lemma 6 on $F = \Phi$ with prior $\pi = \frac{1}{2}\|\cdot - z_-\|^2$ (that is, $c = 0$, $\gamma = 1$, $\bar{x} = z_-$) for $k = \lceil 2\sqrt{L_\Phi} \rceil$ steps, and let $p = u_k$ and $z = z_k$ be its output. Then

$$A_\Phi := A_k = \frac{k(k+3)}{4L_\Phi} \;\geq\; 1, \qquad \theta := \frac{A_\Phi}{1 + A_\Phi} \in \Big[\frac{1}{2}, 1\Big),$$

and for all $w \in \mathcal{X}$

$$\frac{\theta}{\theta_0} g(p) + \frac{\theta}{2}\|p - z_-\|^2 + \frac{1 + \theta(\sigma_\Phi - 1)}{2}\|w - z\|^2 \;\leq\; \frac{\theta}{\theta_0} g(w) + \frac{1}{2}\|w - z_-\|^2. \tag{23}$$

Proof. The bound $A_\Phi \geq k^2/(4L_\Phi) \geq 1$ follows from $k \geq 2\sqrt{L_\Phi}$, and then $\theta = A_\Phi/(1 + A_\Phi) \geq \frac{1}{2}$. By (13) with $\mu_F = \sigma_\Phi$, for all $w \in \mathcal{X}$,

$$\frac{1}{2}\|w - z_-\|^2 + A_\Phi\, \Phi(w) \;\geq\; A_\Phi\, \Phi(p) + \frac{1 + \sigma_\Phi A_\Phi}{2}\|w - z\|^2.$$

Substitute the definition of Φ on both sides. On the left, the two quadratic terms combine into $\frac{1+A_\Phi}{2}\|w - z_-\|^2$:

$$\frac{1 + A_\Phi}{2}\|w - z_-\|^2 + \frac{A_\Phi}{\theta_0} g(w) \;\geq\; \frac{A_\Phi}{\theta_0} g(p) + \frac{A_\Phi}{2}\|p - z_-\|^2 + \frac{1 + \sigma_\Phi A_\Phi}{2}\|w - z\|^2.$$

Divide by $1 + A_\Phi$ and use $\frac{A_\Phi}{1+A_\Phi} = \theta$ and

$$\frac{1 + \sigma_\Phi A_\Phi}{1 + A_\Phi} = \frac{1}{1 + A_\Phi} + \sigma_\Phi \frac{A_\Phi}{1 + A_\Phi} = (1 - \theta) + \sigma_\Phi \theta = 1 + \theta(\sigma_\Phi - 1).$$

This is (23). □

We apply Lemma 21 with $\theta_0 = \theta$. This is not circular: θ is determined by L_Φ alone, and below L_Φ is a fixed constant that does not depend on θ_0 (see (25)). So we first compute L_Φ, then k, A_Φ and θ, and only then define the subproblems with $\theta_0 = \theta$. With $\theta_0 = \theta$, (23) reads

$$g(p) + \frac{\theta}{2}\|p - z_-\|^2 + \frac{1 + \theta(\sigma_\Phi - 1)}{2}\|w - z\|^2 \;\leq\; g(w) + \frac{1}{2}\|w - z_-\|^2 \qquad \forall w \in \mathcal{X}. \tag{24}$$

C.5 ONE OUTER BLOCK

An outer block takes a point $w \in \mathcal{X}$ with $P(w) - P^\star \leq E$ and returns a point u_K with $P(u_K) - P^\star \leq \frac{9}{32}E$, that is, it more than halves the primal error.

Parameters. Set

$$L_\Phi = 1 + \frac{L}{4\Gamma}, \qquad \sigma_\Phi = 1 + \frac{\mu}{4\Gamma}, \qquad k = \lceil 2\sqrt{L_\Phi} \rceil, \qquad A_\Phi = \frac{k(k+3)}{4L_\Phi}, \qquad \theta = \frac{A_\Phi}{1 + A_\Phi},$$
$$K = \Big\lceil 8\sqrt{\bar{\chi}/\theta} \Big\rceil, \qquad c_\delta = \frac{\bar{\chi}}{4K^3}, \qquad \delta = c_\delta E. \tag{25}$$

Weights. Let $A_0 = 0$ and, for $i = 1, \dots, K$, let $a_i > 0$ be the positive root of

$$a_i^2 = \frac{\theta}{4\Gamma}\,(A_{i-1} + a_i), \qquad A_i = A_{i-1} + a_i,$$

so that $a_i^2/A_i = \theta/(4\Gamma)$.

Steps. Start from $u_0 = z_0 = w$. For $z \in \mathcal{X}$ put $w_i(z) = \frac{A_{i-1}u_{i-1} + a_i z}{A_i} \in \mathcal{X}$, and let $v_i = w_i(z_{i-1})$. For $i = 1, \dots, K$:

1. Using O_y, find $y_i \in \mathcal{Y}$ with $P(v_i) - f(v_i, y_i) \le \delta$, for example by Lemma 11 applied to the concave function $f(v_i, \cdot)$.
2. Apply Lemma 21 to $g_i(z) = A_i f(w_i(z), y_i)$ with $\theta_0 = \theta$ and $z_- = z_{i-1}$, that is, to

$$\Phi_i(z) = \frac{A_i}{\theta} f(w_i(z), y_i) + \frac{1}{2}\|z - z_{i-1}\|^2.$$

Let p_i and z_i be its output, and put $u_i = w_i(p_i)$.

The function g_i is convex as a composition of the convex function $f(\cdot, y_i)$ with an affine map. The gradient $\nabla\Phi_i(z) = \frac{a_i}{\theta}\nabla_x f(w_i(z), y_i) + z - z_{i-1}$ costs one call to O_x.

The constants of Φ_i. The map $z \mapsto w_i(z)$ is affine with linear part $\frac{a_i}{A_i} I$. Hence $z \mapsto \frac{A_i}{\theta} f(w_i(z), y_i)$ is strongly convex and smooth with constants

$$\frac{A_i}{\theta} \cdot \frac{a_i^2}{A_i^2} \cdot \mu = \frac{a_i^2}{\theta A_i}\,\mu = \frac{\mu}{4\Gamma} \qquad \text{and} \qquad \frac{a_i^2}{\theta A_i}\,L = \frac{L}{4\Gamma},$$

by the choice of a_i. Adding the 1-strongly convex and 1-smooth term $\frac{1}{2}\|z - z_{i-1}\|^2$ shows that Φ_i is σ_Φ-strongly convex and L_Φ-smooth, with the constants in (25). These constants do not depend on i or θ.

Lemma 22 (growth of the weights)**.** *For $i \ge 1$,*

$$\frac{\sqrt{\theta}\,(i+1)}{4\sqrt{\Gamma}} \;\le\; \sqrt{A_i} \;\le\; \frac{\sqrt{\theta}\,i}{2\sqrt{\Gamma}}.$$

Consequently, $\sum_{i=1}^K A_i \le \frac{\theta K^3}{4\Gamma}$ and $A_K \ge \frac{\theta(K+1)^2}{16\Gamma} \ge \frac{4}{\mu}$.

Proof. Write $\rho = \sqrt{\theta/(4\Gamma)} = \frac{\sqrt{\theta}}{2\sqrt{\Gamma}}$, so $a_i = \rho\sqrt{A_i}$. For $i = 1$, $A_1 = a_1 = \rho\sqrt{A_1}$ gives $\sqrt{A_1} = \rho$, which matches both bounds. For $i \ge 2$,

$$\sqrt{A_i} - \sqrt{A_{i-1}} = \frac{A_i - A_{i-1}}{\sqrt{A_i} + \sqrt{A_{i-1}}} = \frac{\rho\sqrt{A_i}}{\sqrt{A_i} + \sqrt{A_{i-1}}} \in \Big[\frac{\rho}{2},\, \rho\Big],$$

because $0 \le A_{i-1} \le A_i$. Summing from 2 to i gives $\rho\big(1 + \frac{i-1}{2}\big) \le \sqrt{A_i} \le \rho i$, which is the claim since $\rho(1 + \frac{i-1}{2}) = \frac{\rho(i+1)}{2}$.

For the sum, $A_i \le \rho^2 i^2$ and $\sum_{i=1}^K i^2 = \frac{K(K+1)(2K+1)}{6} \le K^3$; the last inequality is equivalent to $(4K+1)(K-1) \ge 0$. Hence $\sum_i A_i \le \rho^2 K^3 = \frac{\theta K^3}{4\Gamma}$. For the last claim, $A_K \ge \frac{\rho^2(K+1)^2}{4} = \frac{\theta(K+1)^2}{16\Gamma}$, and $K \ge 8\sqrt{\bar{\chi}/\theta}$ gives $(K+1)^2 \ge K^2 \ge \frac{64\bar{\chi}}{\theta} = \frac{64\Gamma}{\theta\mu}$, so $A_K \ge \frac{4}{\mu}$. □

Lemma 23 (potential)**.** *Let $V_i = A_i\big(P(u_i) - P^\star\big) + \frac{1}{2}\|z_i - x^\star\|^2$. Then for $i = 1, \dots, K$*

$$V_i + \frac{\theta}{4}\|p_i - z_{i-1}\|^2 + \frac{\theta\mu}{8\Gamma}\|z_i - x^\star\|^2 \;\le\; V_{i-1} + 2A_i\delta, \tag{26}$$

and consequently $P(u_K) - P^\star \le \frac{9}{32}E$.

Proof. Fix i. We combine three inequalities.

(1) The inner run. Apply (24) with $g = g_i$, $z_- = z_{i-1}$, $p = p_i$, $z = z_i$ and $w = x^\star$. Since $g_i(p_i) = A_i f(u_i, y_i)$ and $\sigma_\Phi - 1 = \frac{\mu}{4\Gamma}$,

$$A_i f(u_i, y_i) + \frac{\theta}{2}\|p_i - z_{i-1}\|^2 + \frac{1}{2}\Big(1 + \frac{\theta\mu}{4\Gamma}\Big)\|x^\star - z_i\|^2 \;\le\; A_i f(w_i(x^\star), y_i) + \frac{1}{2}\|x^\star - z_{i-1}\|^2.$$

(2) From the model to P at u_i. Apply Lemma 19 with $v = v_i$, $\tilde{y} = y_i$ and $x = u_i$. Since $u_i - v_i = w_i(p_i) - w_i(z_{i-1}) = \frac{a_i}{A_i}(p_i - z_{i-1})$,

$$A_i P(u_i) \;\le\; A_i f(u_i, y_i) + \frac{\gamma_\chi a_i^2}{A_i}\|p_i - z_{i-1}\|^2 + 2A_i\delta \;\le\; A_i f(u_i, y_i) + \frac{\theta}{4}\|p_i - z_{i-1}\|^2 + 2A_i\delta,$$

where we used $\frac{a_i^2}{A_i} = \frac{\theta}{4\Gamma}$ and $\gamma_\chi \le \Gamma$.

(3) The right-hand side at $x^\star$. By convexity of $f(\cdot, y_i)$ and $f(\cdot, y_i) \le P$,

$$A_i f(w_i(x^\star), y_i) \;\le\; A_{i-1} f(u_{i-1}, y_i) + a_i f(x^\star, y_i) \;\le\; A_{i-1} P(u_{i-1}) + a_i P^\star.$$

Combining. Use (2) to replace $A_i f(u_i, y_i)$ in (1) by the smaller quantity $A_i P(u_i) - \frac{\theta}{4}\|p_i - z_{i-1}\|^2 - 2A_i\delta$, and use (3) on the right-hand side of (1):

$$A_i P(u_i) + \frac{\theta}{4}\|p_i - z_{i-1}\|^2 + \frac{1}{2}\Big(1 + \frac{\theta\mu}{4\Gamma}\Big)\|z_i - x^\star\|^2 \;\le\; A_{i-1} P(u_{i-1}) + a_i P^\star + \frac{1}{2}\|z_{i-1} - x^\star\|^2 + 2A_i\delta.$$

Subtracting $A_i P^\star = A_{i-1} P^\star + a_i P^\star$ from both sides gives (26).

Consequence. Drop the nonnegative extra terms on the left of (26) and sum over $i = 1, \dots, K$. Using Lemma 22 and $\delta = c_\delta E = \frac{\bar{\chi} E}{4K^3}$,

$$V_K \;\le\; V_0 + 2\delta \sum_{i=1}^{K} A_i \;\le\; V_0 + 2 \cdot \frac{\bar{\chi} E}{4K^3} \cdot \frac{\theta K^3}{4\Gamma} = V_0 + \frac{\theta E}{8\mu},$$

where $\Gamma = \mu\bar{\chi}$. Since $A_0 = 0$ and $z_0 = w$, strong convexity of P gives $V_0 = \frac{1}{2}\|w - x^\star\|^2 \le \frac{E}{\mu}$; with $\theta < 1$ we get $V_K \le \frac{9E}{8\mu}$. Finally, $A_K(P(u_K) - P^\star) \le V_K$ and $A_K \ge \frac{4}{\mu}$ by Lemma 22, so

$$P(u_K) - P^\star \;\le\; \frac{V_K}{A_K} \;\le\; \frac{9E}{8\mu} \cdot \frac{\mu}{4} = \frac{9}{32}E.$$

□

C.6 Proof of Theorem 4

Put $\Lambda = 1 + \log\frac{B\bar{\chi}}{\varepsilon}$. Since $\varepsilon < B$ and $\bar{\chi} \ge 1$, both $\log\frac{B}{\varepsilon}$ and $\log\bar{\chi}$ are at most $\log\frac{B\bar{\chi}}{\varepsilon}$.

Initialization. Fix $y_0 \in \mathcal{Y}$ and apply Lemma 8 to $f(\cdot, y_0)$ with $E_{\text{target}} = B$. The output w_0 satisfies $f(w_0, y_0) - \phi(y_0) \le B/2$. Moreover, for a supergradient g_0 of $f(w_0, \cdot)$ at y_0 with $\|g_0\| \le M_y$, concavity gives $f(w_0, y) \le f(w_0, y_0) + \langle g_0, y - y_0\rangle \le f(w_0, y_0) + B$ for all $y \in \mathcal{Y}$, so $P(w_0) - f(w_0, y_0) \le B$. Since also $\phi(y_0) \le \max_{\mathcal{Y}} \phi = P^\star$,

$$P(w_0) - P^\star = \big[P(w_0) - f(w_0, y_0)\big] + \big[f(w_0, y_0) - \phi(y_0)\big] + \big[\phi(y_0) - P^\star\big] \;\le\; B + \frac{B}{2} + 0 \;\le\; 2B.$$

This costs $O(I_0)$ gradients and no calls to O_y.

Outer loop. Let $E_0 = 2B$ and $E_s = 2^{-s}E_0$. Block $s = 1, 2, \dots$ starts from w_{s-1} with $P(w_{s-1}) - P^\star \le E_{s-1}$, uses $E = E_{s-1}$ and $\delta = c_\delta E_{s-1}$, and returns $w_s = u_K$. By Lemma 23, $P(w_s) - P^\star \le \frac{9}{32}E_{s-1} \le E_s$, so the invariant is maintained. Algorithm 3 stops after the first block S with $E_S \le \varepsilon/(6\bar{\chi}^2)$, that is,

$$S = \left\lceil \log_2 \frac{12B\bar{\chi}^2}{\varepsilon} \right\rceil = O(\Lambda),$$

and by minimality of S we also have $E_S > \varepsilon/(12\bar{\chi}^2)$. Finally, it computes $\hat{y}$ with $P(\hat{x}) - f(\hat{x}, \hat{y}) \le c_\delta E_S$ at $\hat{x} = w_S$.

Accuracy. By Lemma 20 with $E = E_S$ and $\delta = c_\delta E_S$,

$$\mathrm{Gap}(\hat{x}, \hat{y}) \;\le\; E_S\big(1 + 2\bar{\chi}c_\delta + 4\bar{\chi}\chi\big).$$

Each of the three terms in parentheses is at most a multiple of $\bar{\chi}^2$. First, $1 \le \bar{\chi}^2$. Second, $K \ge 8\sqrt{\bar{\chi}/\theta} \ge 8\sqrt{\bar{\chi}}$ gives $c_\delta = \frac{\bar{\chi}}{4K^3} \le \frac{1}{2048\sqrt{\bar{\chi}}}$, so $2\bar{\chi}c_\delta \le \frac{\sqrt{\bar{\chi}}}{1024} \le \bar{\chi}^2$. Third, $\chi \le \bar{\chi}$ gives $4\bar{\chi}\chi \le 4\bar{\chi}^2$. Hence $\mathrm{Gap}(\hat{x}, \hat{y}) \le 6\bar{\chi}^2 E_S \le \varepsilon$.

Gradient count. Each outer block consists of K inner runs of k steps, one gradient each, so it uses Kk calls to O_x. Since $\theta \ge \frac{1}{2}$, $K \le 8\sqrt{2\bar{\chi}} + 1 = O(\sqrt{\bar{\chi}})$, and $k = \lceil 2\sqrt{1 + \kappa/(4\bar{\chi})}\,\rceil = O(\sqrt{1 + \kappa/\bar{\chi}})$. Hence

$$Kk = O\Big(\sqrt{\bar{\chi}}\sqrt{1 + \kappa/\bar{\chi}}\Big) = O\big(\sqrt{\bar{\chi} + \kappa}\big) = O\big(\sqrt{\kappa + \chi}\big),$$

where we used $\kappa \ge 1$. With $S = O(\Lambda)$ blocks and the initialization, $T_x = I_0 + O(\sqrt{\kappa + \chi}\,\Lambda)$.

Query count in y. The algorithm solves $SK + 1 = O(\sqrt{\bar{\chi}}\,\Lambda)$ problems in y, each to accuracy at least

$$\delta_{\min} = c_\delta E_S \;>\; \frac{c_\delta\,\varepsilon}{12\bar{\chi}^2}.$$

Since $K \le 8\sqrt{2\bar{\chi}} + 1 \le 13\sqrt{\bar{\chi}}$, we have $\frac{1}{c_\delta} = \frac{4K^3}{\bar{\chi}} \le 4 \cdot 13^3\sqrt{\bar{\chi}} = 8788\sqrt{\bar{\chi}}$, so $\frac{B}{\delta_{\min}} \le \frac{12 \cdot 8788\, B\bar{\chi}^{5/2}}{\varepsilon}$ and $\log \frac{B}{\delta_{\min}} = O(\Lambda)$. By Lemma 11, each problem in y is solved by the center-of-gravity method with $O(m\Lambda)$ calls to O_y. Therefore

$$T_y = O\big(\sqrt{\bar{\chi}}\,\Lambda\big) \cdot O(m\Lambda) = O\big(m\sqrt{1 + \chi}\,\Lambda^2\big).$$

If instead a solver returns a δ-maximizer after $N_y(\delta)$ calls, with N_y nonincreasing, the same count gives $T_y = O\big(\sqrt{1 + \chi}\,\Lambda \cdot N_y(\varepsilon/(Cr^{5/2}))\big)$ with $C = 12 \cdot 8788$. □

D Experimental details and additional results

All code is in the supplementary material (directory `experiments/`); every number in Section 5 is produced by `run.py` and `plots.py` from raw JSON logs.

D.1 Problems

All problems have the form

$$f(x, y) = \sum_{j=0}^{J} w_j(y)\, L_j(x) + \frac{\mu}{2}\|x\|^2 + \langle c, y\rangle - \frac{\nu}{2}\|y\|^2, \tag{27}$$

with convex losses L_j, weights $w_j(y)$ that are affine in y and nonnegative on $\mathcal{Y}$, and $\mathcal{X} = \mathbb{R}^n$. One call to O_x returns $\sum_j w_j(y)\nabla L_j(x) + \mu x$ (one weighted backward pass over the data); one call to O_y returns $f(x, y)$ and $\nabla_y f(x, y)$, i.e., the vector $(L_j(x))_j$ (one forward pass). In our NumPy implementation both calls cost one pass over the data; the measured wall-clock ratio of a call to O_y to a call to O_x is between 0.6 and 1.9 across the four real problems (in deep models a forward pass typically costs about a third of a gradient). We therefore report T_x and T_y separately, and weighted costs $T_x + rT_y$ over a range of r. We set $\mu = L_{\max}/(\kappa - 1)$, where $L_{\max}$ is a certified upper bound on the smoothness constant of $f(\cdot, y)$ over $y \in \mathcal{Y}$ (power iteration on the per-group data matrices), so κ is the condition number seen by the methods. The strength of the ridge term is thus the knob that sets κ; $\kappa = 10^4$ corresponds to $\mu \approx 10^{-5}$ on unit-norm features, a typical amount of regularization.

Group DRO (Adult; 20 Newsgroups). L_j is the mean logistic loss of group j, $w(y) = y \in \Delta_G$, and $m = G - 1$ (the simplex is parametrized in its affine hull by an orthonormal basis of $\mathbf{1}^\perp$, so that $\|\cdot\|$ is preserved). *Adult* (Kohavi, 1996): 45,222 examples after removing missing values; $n = 105$

(standardized numerical features, one-hot categorical features, intercept); $G = 4$ groups given by sex $\times$ {White, non-White}. *20 Newsgroups*: binary task `sci.*` vs. `talk.*` (7,205 documents); TF-IDF with $2 \cdot 10^4$ terms plus intercept, rows normalized; the $G = 8$ groups are the eight newsgroups, so $m = 7$ and $n = 20{,}001 > N$.

Learning with a few constraints (Adult). L_0 is the mean logistic loss over all examples and L_j, $j = 1, \dots, 4$, is the mean logistic loss over the *positive* examples of group j (a convex surrogate of the group's false-negative rate). The constraints $L_j(x) \le \tau$ are imposed with τ equal to 0.9 times the average of L_j at the unconstrained ridge solution, so that several constraints are active. The Lagrangian is (27) with $w(y) = (1, y)$, $c = -\tau\mathbf{1}$, $\mathcal{Y} = [0, \Lambda]^4$, $\Lambda = 4$.

Neyman–Pearson classification (20 Newsgroups, $m = 1$). All 18,846 documents; positives are the four `sci.*` groups. We minimize the logistic loss on negatives subject to the logistic loss on positives being at most 0.7 times its value at the unconstrained solution; $\mathcal{Y} = [0, 4]$, $n = 20{,}001$.

Synthetic group DRO (scaling experiments). $L_j(x) = \frac{1}{2}\|A_j x - b_j\|^2$, $j = 0, \dots, m$, where $A_j = G_j S / \|G_j S\|$ with a Gaussian $G_j \in \mathbb{R}^{200\times 400}$, a shared scaling $S = \mathrm{diag}(i^{-1})$ (so that every $f(\cdot, y)$ is genuinely ill-conditioned) and $b_j = A_j c_j + 0.1\xi_j$ with random centers c_j; $L_{\max} = 1 + \mu$. Three independent instances (seeds) per configuration.

Group DRO over all 20 newsgroups ($m = 19$). All 18,846 documents; the label is technology/science/recreation (`comp.*`, `sci.*`, `rec.*`, `misc.forsale`) versus society/politics/religion (`talk.*`, `alt.atheism`, `soc.religion.christian`); the $G = 20$ groups are the newsgroups, so $m = 19$ and $n = 20{,}001$. The same TF-IDF features as above.

Weak regularization. For the experiments of Figure 1 the ridge parameter is decreased so that κ ranges over 10^4–10^7 (10^8 on the synthetic instance), i.e., $\mu/L_{\max}$ between 10^{-4} and 10^{-7}. On 20 Newsgroups $n > N$, so the data term is singular and μ is the actual strong convexity constant.

Group DRO with an ℓ_1 penalty (nonsmooth dual). On Adult with the four groups above, $f(x, y) = \sum_g w_g L_g(x) + \frac{\mu}{2}\|x\|^2 - \rho\|w - p\|_1$, where p is the vector of group proportions and $\rho = 0.1$. The dual function is then nonsmooth; at the solution one group weight sits exactly at its kink $w_g = p_g$ and two are at 0.

χ^2-regularized group DRO (Section 4). On 20 Newsgroups (`sci.*` vs. `talk.*`) with $G \in \{2, 4, 8\}$ groups (the two topics, four pairs of newsgroups, the eight newsgroups), $f = \sum_g w_g L_g(x) + \frac{\mu}{2}\|x\|^2 - \frac{\nu}{2}\|w - p\|^2$ with $\nu = 1$ and $\kappa = 10^3$. Since $\|\nabla L_g\| \le \max_i \|a_i\| = 1$, the coupling constant is certified as $\ell = \sqrt{G}$, so $\chi = \ell^2/(\mu\nu) = G/\mu$ grows with the number of groups (see the discussion after Corollary 5); it equals $1.5 \cdot 10^4$, $3.1 \cdot 10^4$ and $6.2 \cdot 10^4$ for $G = 2, 4, 8$.

Strongly concave problem with bounded coupling (Section 4). $f(x, y) = \frac{1}{2}x^\top H x - \langle h, x\rangle + \langle y, Cx\rangle - \frac{\nu}{2}\|y\|^2$ on $\mathcal{Y} = [-1, 1]^m$, $n = 400$, spec H log-uniform in $[\mu, 1]$, and $C = \ell Q^\top$ with Q having orthonormal columns. Hence $\|\nabla_x f(x, y) - \nabla_x f(x, y')\| \le \ell\|y - y'\|$ with the same ℓ for every m; we take $\kappa = 10^4$ and $\chi = \ell^2/(\mu\nu) = 1$.

D.2 Methods and implementation

Certificate transport. Algorithm 2 exactly as stated, with the default $K_0 = \lceil\sqrt{8\kappa/m}\,\rceil$ unless a sweep over K is reported. The lower model is initialized on the slice $y_0 = \mathrm{cg}(\mathcal{Y})$ by accelerated gradient descent stopped by the gradient certificate $f(x, y_0) \ge f(u, y_0) - \frac{\|g\|^2}{2\mu} + \frac{\mu}{2}\|x - (u - g/\mu)\|^2$, which is a valid model (5) and replaces the restarts of Lemma 8 (whose only role is to handle an unknown R_x). $a = \max_{\mathcal{Y}} U$ and the weights λ of the output $\bar{x}$ (Lemma 12) are obtained from one LP whose dual variables are λ. We also report a variant, *CT with geometric steps*, which differs only in the step sizes of Algorithm 1: instead of $a_i = (i+1)\gamma/(2L)$ it takes the largest a_i admitted by the proof of Lemma 6, i.e., the root of $La_i^2 = \Gamma_{i-1}A_i$ (the step rule of accelerated gradient descent for strongly convex functions). The proof of Lemma 6 uses the step sizes only through $La_i^2 \le \Gamma_{i-1}A_i$,

so Lemma 6, Corollary 7 and Theorem 1 hold verbatim; the only difference is that A_K becomes larger (exponentially in $K/\sqrt{\kappa}$ once $K \gtrsim \sqrt{\kappa}$), which makes ϑ closer to 1.

Practical certificate transport. In addition to geometric steps, *CT (practical)* uses two further modifications, neither of which changes Theorem 1. (a) *Prior curvature* $\sigma = \mu$. The initial model has curvature μ, and Corollary 7 maps a model of curvature σ to one of curvature $(\sigma+\mu A)/(1+A)$, which equals μ when $\sigma = \mu$. Hence one can run Algorithm 1 with $\gamma = \mu$ throughout ($\sigma = \mu/2$ is needed only in Lemma 8), which doubles A and therefore increases ϑ. (b) *Direct certificate.* After the block, one extra call $g_u = \mathsf{O}_x(u, y_c)$ gives the model $f(x, y_c) \geq v - \frac{\|g_u\|^2}{2\mu} + \frac{\mu}{2}\|x-(u-g_u/\mu)\|^2$ on the queried slice itself. The algorithm keeps whichever of this model and the transported one has the larger constant, and only if it exceeds the current b. Thus b never decreases and still grows by at least $\vartheta(v-b)$ after every transport, so the counting in Section B is unchanged; each query costs $K+1$ calls to O_x. When the block has essentially solved the slice, (b) makes CT jump to the slice like a nested scheme; when it has not, the transported bound is used. The default block length K_0 is unchanged.

Nested cutting planes. The textbook scheme of (4): the center-of-gravity method on ϕ; at each center y_c, accelerated gradient descent on $f(\cdot, y_c)$ stopped as soon as $\|\nabla_x f\|^2/(2\mu) \leq \delta$, which certifies $f(\tilde{x}, y_c) - \phi(y_c) \leq \delta$; the cut is the plane of $f(\tilde{x}, \cdot)$ and the lower bound is $f(\tilde{x}, y_c) - \delta$ (Lemma 11). *Cold start*: every inner run starts from $x = 0$ and $\delta = \varepsilon/2$, as in the analysis of Alkousa et al. (2023). *Warm start*: the inner run starts from the previous inner solution (same δ); this heuristic has no better worst-case guarantee but is what one would implement in practice. *Warm, adaptive*: in addition $\delta = \max\{\varepsilon/2,\ 0.1 \cdot \text{current certified width}\}$. All nested variants use the same center-of-gravity routine, LPs and output recovery as CT.

Level-localizer ablation. This baseline is the warm-started nested scheme with adaptive inner accuracy, but with the localizer and epochs of Algorithm 2. Queries are centers of gravity of $\{y : U(y) \geq h\}$ with $h = \beta + W/2$ updated in epochs as in Algorithm 2, and each query is an inner accelerated run from the previous inner point stopped at accuracy $\max\{\varepsilon/2,\ c\,(a-\beta)\}$. Because $U \geq \phi$, a cut never removes a point with $\phi \geq h$, which repairs the failure mode of the plain adaptive scheme (Section D.3); the certified width is valid as for all nested schemes. This is an inexact level method in the sense of Lemaréchal et al. (1995). We are not aware of a bound of order $\sqrt{m\kappa}\log\frac{1}{\varepsilon}$ for it: the cost of each warm-started inner run grows with the distance between consecutive slice solutions. We ran $c \in \{0.03, 0.1, 0.3, 0.5, 1\}$ on 20 Newsgroups ($m = 7, 19$) and on the synthetic instances ($m = 4, 16$, seed 0), $c \in \{0.1, 0.3, 0.5\}$ on the real problems at $\kappa = 10^4$, and the default $c = 0.1$ elsewhere, and we report the best c in hindsight. Large fractions make each inner solve cheap but the lower bound β loose. With $c = 1$ the method failed to certify ε on the synthetic instance with $m = 4$ for every κ, on the synthetic instance with $m = 16$ for $\kappa \geq 10^6$, and on 20News with 20 groups at $\kappa = 10^7$ (budget: $2 \cdot 10^4$ calls to O_y or 5 hours), and it needed thousands of y-queries on the problems with $m \geq 16$ when it did converge.

AGD on the primal function (Section 4). Nesterov's method for strongly convex functions on P with step $1/L_P$, $L_P = L + 2\ell^2/\nu$, and the inexact gradient $\nabla_x f(v, \tilde{y})$ at a δ-maximizer $\tilde{y}$ of $f(v, \cdot)$ with $\delta = \varepsilon/(4\sqrt{L_P/\mu})$. By Lemma 19 this is a $(2\delta, L_P)$-inexact oracle (Devolder et al., 2014). It uses one call to O_x per step and one y-subproblem per step. For both this baseline and Algorithm 3, the y-subproblems of the strongly concave problems are solved by projected gradient ascent with step $1/\nu$ (Theorem 4 allows any δ-maximizer), so the number of calls to O_y is proportional to the number of y-subproblems.

Single-loop baselines. Extragradient (Korpelevich, 1976) on the joint variable with Euclidean projection onto $\mathcal{Y}$ (two calls to each oracle per iteration) and, for group DRO, simultaneous gradient descent in x with exponentiated-gradient ascent on the group weights (deterministic version of Sagawa et al., 2020). Both have separate step sizes for the two blocks, which we tune over a grid ($\eta_x \in \{0.5, 1\}/L_{\max}$, $\eta_y \in \{0.1, 1, 10, 100\}$) and report the best configuration *in hindsight*, i.e., the one reaching the target accuracy with the fewest calls to O_x; both the last and the averaged iterate are evaluated and the better one is reported. These choices favor the baselines. CT and the nested schemes have no tuned parameters.

Centers of gravity. For $m \leq 4$ we compute the exact centroid of the localizer (vertex enumeration with Qhull and triangulation). For $m \geq 5$ we use the approximate centroid given by the average of $\max(4000, 150m^2)$ hit-and-run samples with a rounding matrix carried over from the previous call, in the spirit of Bertsimas & Vempala (2004). This affects only the number of iterations, never the correctness of the certificates, which hold for every choice of query points.

Evaluation. The reported gap is $\mathrm{Gap}(\hat{x}, \hat{y}) = P(\hat{x}) - \phi(\hat{y})$ of each method's output, computed independently of the method and without counting oracle calls: $P(\hat{x})$ is computed exactly (a vertex of the simplex or the box, or a projection when $\nu > 0$), and $\phi(\hat{y})$ is bounded from below by $f(\tilde{x}, \hat{y}) - \|\nabla_x f(\tilde{x}, \hat{y})\|^2/(2\mu)$ after running Newton-CG on $f(\cdot, \hat{y})$ to relative accuracy 10^{-4}. The reported numbers are therefore rigorous upper bounds on the gap. For CT and the nested schemes the output is $(\bar{x}, y_b)$ (resp. $(\bar{x}, y_{\mathrm{best}})$), exactly as returned by the algorithms; in every run we checked that the true gap never exceeds the certified width. The evaluation solves for $\phi(\hat{y})$ until the gradient bound $\|\nabla_x f(\tilde{x}, \hat{y})\|^2/(2\mu)$ is below 10^{-4} times the gap being measured (a relative accuracy *of the gap*, not of ϕ), so the reported gap may exceed the true gap by at most this relative amount; in 24 of the 4,068 recorded points (from 12 runs) this made the reported gap exceed the certified width, by at most $6.4 \cdot 10^{-5}$ in relative terms. We re-ran these 12 runs and re-evaluated all their outputs with relative accuracy 10^{-7} or better (script `recheck.py`); the true gap was then below the certified width everywhere.

Compute. All experiments ran on a single 22-core CPU machine, one process and one BLAS thread per run; the whole suite (about 900 runs) takes several hours on this machine and is dominated by the tuned single-loop baselines and by hit-and-run centers for $m \geq 16$.

D.3 Additional results

Table 2. Weak regularization ($\varepsilon = 10^{-6}$): calls T_x / T_y of CT, of the level baseline with its default fraction $c = 0.1$ and with its best fraction in hindsight (value of c in parentheses), and of the adaptive and warm-started nested schemes. Bold: fewest calls in each row (separately for T_x and T_y). "–": not run or did not certify ε within the budget.

problem	κ	CT	level, $c = 0.1$	level, best c	nested, adaptive	nested, warm
20News, 20 groups ($m = 19$)	10^4	8,735 / 112	16,523 / **98**	**7,644** / 111 (0.5)	17,941 / 122	36,684 / 125
	10^5	**25,933** / 104	57,399 / **98**	27,907 / 111 (0.5)	58,557 / 113	100,665 / 110
	10^6	**79,521** / 100	168,809 / **81**	83,557 / 139 (0.5)	182,646 / 107	332,526 / 111
	10^7	**228,277** / 88	401,969 / **61**	267,176 / 1,082 (0.5)	465,067 / 125	1,172,507 / 128
20News, 8 groups ($m = 7$)	10^4	6,131 / 45	8,367 / 46	**3,924** / 47 (0.5)	7,769 / 46	13,488 / **41**
	10^5	19,150 / 44	27,889 / 41	**13,453** / 42 (0.5)	25,186 / 43	39,142 / **40**
	10^6	58,891 / 42	79,268 / **35**	**43,979** / 413 (1)	77,770 / 39	113,717 / 37
	10^7	170,604 / 37	206,029 / **28**	**116,934** / 49 (0.5)	187,866 / 38	355,585 / 41
Adult, 4 groups ($m = 3$)	10^4	3,601 / **16**	–	–	**2,655** / 43	6,774 / 43
	10^5	11,356 / **16**	7,651 / 19	7,651 / 19 (0.1)	**7,400** / 43	20,590 / 43
	10^6	37,732 / **17**	22,354 / 19	22,354 / 19 (0.1)	**19,905** / 44	66,580 / 44
	10^7	118,464 / **17**	68,531 / 20	68,531 / 20 (0.1)	**58,671** / 44	251,010 / 44
20News NP ($m = 1$)	10^4	3,178 / **13**	**1,378** / 16	**1,378** / 16 (0.1)	2,410 / 16	6,828 / 15
	10^5	10,016 / **13**	**4,760** / 16	**4,760** / 16 (0.1)	–	23,144 / 16
	10^6	31,568 / **13**	**21,064** / 17	**21,064** / 17 (0.1)	24,615 / 19	75,881 / 17
	10^7	99,526 / **13**	**60,839** / 17	**60,839** / 17 (0.1)	–	239,422 / 18
Synthetic ($m = 16$)	10^4	9,248 / 130	6,992 / **57**	**4,177** / 66 (0.3)	26,285 / 222	71,113 / 224
	10^5	29,284 / 131	44,534 / **81**	**19,329** / 164 (0.5)	91,504 / 178	193,446 / 175
	10^6	**85,893** / 121	219,211 / 93	143,954 / **85** (0.3)	364,163 / 156	626,619 / 153
	10^7	**277,572** / 123	976,732 / **101**	647,344 / 214 (0.5)	1,364,564 / 141	2,136,345 / 139
Synthetic ($m = 4$)	10^4	4,568 / 33	4,235 / **30**	**3,196** / 35 (0.3)	5,610 / 34	12,075 / 32
	10^5	15,869 / 36	18,392 / **31**	**9,717** / 44 (0.5)	23,322 / 35	43,280 / 35
	10^6	47,935 / 34	76,521 / **32**	**30,944** / 35 (0.5)	97,874 / 36	163,584 / 36
	10^7	135,480 / **30**	304,671 / 31	**133,266** / 32 (0.5)	360,539 / 33	551,295 / 32
	10^8	**433,824** / **30**	1,121,064 / **30**	713,778 / 35 (0.5)	1,411,130 / 35	2,013,173 / 34

Cuts and transports. With few groups almost every query of CT raises the lower bound: f is linear in y and $f(\cdot, y)$ is strongly convex, so ϕ is differentiable (Danskin's theorem), a few planes

Table 3. The level baseline with a fixed accuracy fraction c, over all (problem, κ) cells of Table 2 where it was run: geometric mean of $T_x(\text{level}_c)/T_x(\text{CT})$ over the cells where it certified ε, and the number of cells where it did not. Bold: best value in each row.

c	0.03	0.1	0.3	0.5	1
geometric mean of the ratio	2.01	1.21	1.09	**0.91**	1.07
failures / cells	**0/13**	1/25	**0/17**	1/17	8/17

Table 4. Calls $T_x\,/\,T_y$ until the true gap is $\le 10^{-7}$ on the real problems ($\kappa = 10^4$). Tuned methods: best configuration in hindsight (the level baseline over $c \in \{0.1, 0.3, 0.5\}$). Single-loop methods are evaluated on a geometric grid of ratio 1.04. Bold: fewest calls in each column (separately for T_x and T_y). †No guarantee of the order of Theorem 1.

	NP, 20News $m=1$	DRO, Adult $m=3$	constr., Adult $m=4$	DRO, 20News $m=7$	DRO+ℓ_1, Adult $m=3$
CT (practical)	3,462 / **14**	3,931 / **18**	3,817 / **27**	6,995 / **53**	2,670 / **17**
CT, steps of Algorithm 1	6,563 / 25	16,051 / 92	17,140 / 121	8,977 / 72	10,199 / 63
Nested, cold start	11,575 / 16	33,127 / 49	50,231 / 63	46,941 / 56	26,933 / 39
Nested, warm start	8,687 / 16	10,158 / 51	22,808 / 65	20,862 / 55	10,279 / 38
Nested, warm+adaptive†	2,810 / 17	**3,262** / 51	6,939 / 65	9,691 / 55	4,289 / 40
Level, best c†	**1,603** / 16 ($c=0.1$)	3,344 / 40 ($c=0.3$)	**2,190** / 40 ($c=0.3$)	**4,562** / 54 ($c=0.5$)	**1,523** / 18 ($c=0.1$)
Extragradient (tuned)	74,154 / 74,154	54,008 / 54,008	66,718 / 66,718	91,606 / 91,606	>100,000
GDA-MW (tuned)	–	29,853 / 29,853	–	54,433 / 54,433	–

make the upper model accurate, and the certified width is dominated by b. With many groups the localizer matters: on 20News DRO with $m = 7$ about half of the queries of CT are cuts, and on the 20-group problem 28–41 of about 100. Without the direct certificate ("CT, steps of Algorithm 1" and "geometric steps" in the figures) almost all queries are transports, and larger blocks reduce T_x up to $K \approx (1.6\text{–}3.6)K_0$ (Figure 7).

Wall-clock time. All runs shared one machine with a varying load, so we compare oracle calls, not time. The share of wall-clock time that CT spends inside the two oracles is 82–100% for $m \le 4$ with exact centers, 66–92% for $m = 7$, and 17–89% for $m \in \{16, 19\}$, where hit-and-run centers are relatively expensive when κ (and hence the cost per query) is small. Computing implementable centers more efficiently (Bertsimas & Vempala, 2004; Vaidya, 1996) is therefore what matters in practice for $m \gtrsim 10$.

Seeds of the approximate centers. On 20News DRO ($m = 7$, hit-and-run centers) three seeds of the center oracle changed the number of calls to O_x of CT by at most 2% and of the nested and level schemes by at most 6%.

Table 5. Price of the y-block for $m = 1$ (Neyman–Pearson on 20News, $\varepsilon = 10^{-6}$): calls to O_x of each method divided by the calls of accelerated gradient descent on $f(\cdot, y^\star)$ with $y^\star$ known in advance (absolute counts of the latter in the last row).

κ	10^2	10^3	10^4	10^5
CT (Algorithm 2)	12.2	12.9	14.0	16.5
CT, geometric steps	7.6	8.6	9.2	11.6
Nested, cold start	14.3	15.9	17.1	21.4
Nested, warm start	9.6	11.6	12.9	15.7
AGD on $f(\cdot, y^\star)$ (calls)	50	169	529	1,477

Warm starts with adaptive inner accuracy: fast but not robust. Making the inner accuracy of the warm-started nested scheme proportional to the current width, $\delta = \max\{\varepsilon/2, 0.1(a-\beta)\}$, makes the nested scheme the fastest method on three of our four real problems, but it can break the method. A coarse early cut is the plane of $f(\tilde{x}, \cdot)$ with an inexact $\tilde{x}$ and can remove the maximizer of ϕ from the localizer. Lemma 11 then only guarantees $a - \beta \le B(1 - e^{-1})^{t/m} + \max_s \delta_s$. This happened in our synthetic experiments: the variant failed to certify the target accuracy in 4 of the 39 scaling runs, all with $m = 1$ (4 of 15 such runs), and on the instance of Figure 5 (synthetic, $m = 1$)

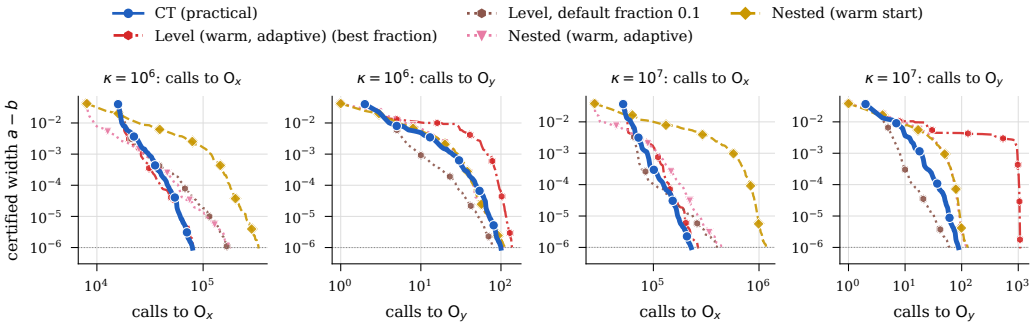

Figure 4. Group DRO over the 20 newsgroups ($m = 19$, $n = 20{,}001$): certified width versus calls to O_x and to O_y for $\kappa = 10^6$ and 10^7.

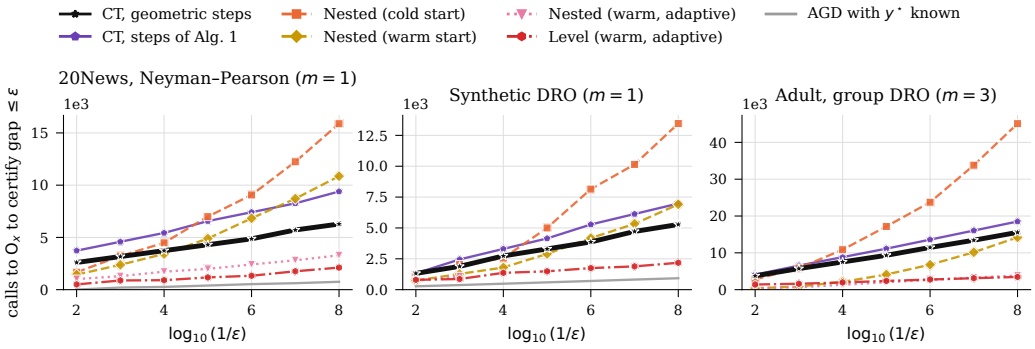

Figure 5. Calls to O_x needed to certify gap $\leq \varepsilon$ ($\kappa = 10^4$). CT is run once and read off at every ε (anytime); the nested schemes are rerun for every ε because their inner accuracy depends on ε. The gray curve is accelerated gradient descent on $f(\cdot, y^\star)$ with $y^\star$ given in advance.

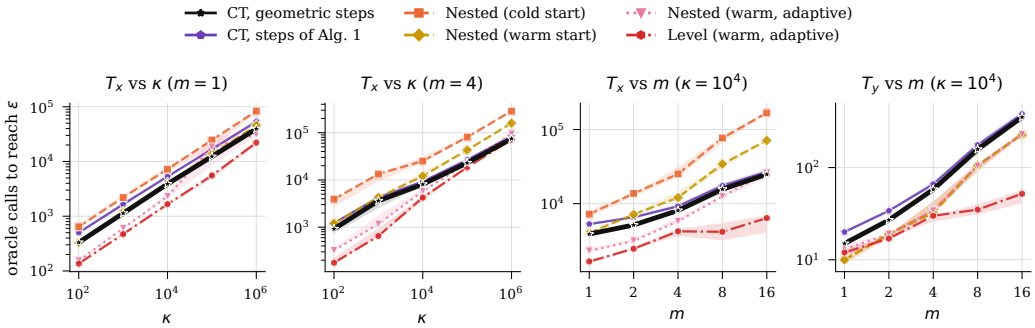

Figure 6. Synthetic group DRO: calls to reach $\varepsilon = 10^{-6}$ (median over three instances, band: min–max). Left: dependence on κ. Right: dependence on m at $\kappa = 10^4$.

for every target $\varepsilon \leq 10^{-4}$, and in 1 of the 12 runs of the strongly concave experiment. In each failure the certified width froze between $5 \cdot 10^{-6}$ and $3 \cdot 10^{-4}$ and did not move during $2 \cdot 10^4$ further calls to O_y. It never failed on the real problems. CT and the two nested schemes with guarantees certified the target accuracy in every one of their runs.

E Detailed discussion of related work

Information-based complexity and cutting-plane methods. The theory of oracle complexity was laid out in the monograph of Nemirovsky & Yudin (1983). It shows that m-dimensional convex problems in the high-accuracy regime require, and can be solved with, $\Theta(m \log \frac{1}{\varepsilon})$ first-order

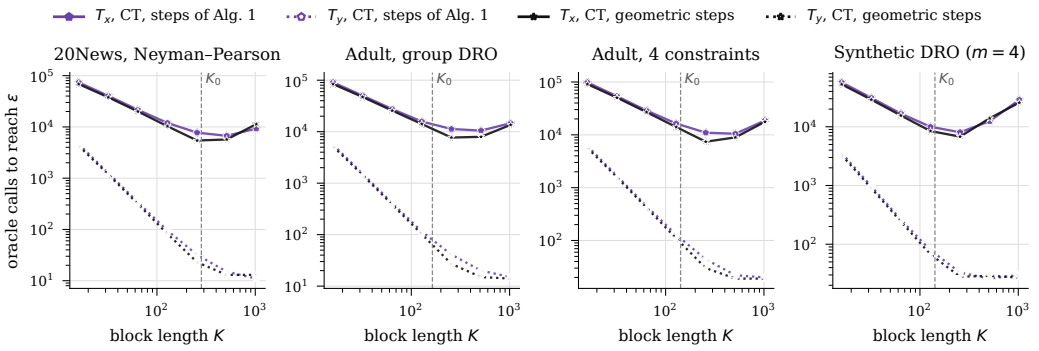

Figure 7. Calls to O_x (solid) and O_y (dotted) to reach $\varepsilon = 10^{-6}$ as a function of the block length K, for the step sizes of Algorithm 1 and for geometric steps ($\kappa = 10^4$). The dashed line marks the default $K_0 = \lceil \sqrt{8\kappa/m} \rceil$. Runs that did not reach ε within $2 \cdot 10^4$ calls to O_y are omitted.

queries. The upper bound is attained by the center-of-gravity method (Levin, 1965; Newman, 1965), which rests on Grünbaum's inequality (Grünbaum, 1960). The method itself is not implementable. The ellipsoid method (Shor, 1977; Nemirovsky & Yudin, 1983) makes it polynomial-time at the price of $O(m^2 \log \frac{1}{\varepsilon})$ queries, while Vaidya's method (Vaidya, 1996), approximate centers of gravity based on random walks (Bertsimas & Vempala, 2004) and modern cutting-plane methods (Lee et al., 2015; Jiang et al., 2020) keep the optimal linear dependence on m at polynomial arithmetic cost; the last of these works also treats convex-concave games. Using only function values, Lee et al. (2018) need $\widetilde{O}(m^2)$ evaluations. Accuracy certificates (Nemirovski et al., 2010) make the output of cutting-plane methods verifiable; we use a related, model-based notion of certificate. Rodomanov & Nesterov (2023) proposed a subgradient ellipsoid method that interpolates between the ellipsoid method and subgradient methods; we return to it in Section F. For a textbook exposition, see Nesterov (2018).

Problems with a low-dimensional block. Problems in which one group of variables has small dimension were studied systematically in a series of works (Gladin et al., 2021a; Alkousa et al., 2023; Gladin et al., 2022; 2023; Pasechnyuk & Stonyakin, 2019). Gladin et al. (2021a) consider min-min problems that are smooth and strongly convex in one block and low-dimensional in the other: an outer ellipsoid-type method is combined with an inner accelerated method, and the effect of the inexactness of the outer oracle is analyzed. Alkousa et al. (2023) use the same kind of nested scheme with inexact oracles for strongly convex-concave composite saddle-point problems with a low-dimensional block and obtain, in our notation, the bound (4), $T_x = O\big(m\sqrt{\kappa} \log \frac{B}{\varepsilon} \log \frac{\mu R_x^2 + B}{\varepsilon}\big)$, up to an additional $\log m$ inside the first logarithm that comes from their outer method. Gladin et al. (2022) extend Vaidya's method to stochastic low-dimensional problems, and Gladin et al. (2023) use a low-dimensional dual problem to build linearly convergent algorithms for constrained Markov decision processes. Pasechnyuk & Stonyakin (2019) develop a dichotomy-type method for functions of two variables.

Our work continues this line. We show that the multiplicative cost of nested schemes can be removed with no assumptions on the y-block beyond concavity and Lipschitz continuity: the product $m\sqrt{\kappa}$ is replaced by $\sqrt{m\kappa}$, and the second logarithm disappears. The key difference from nested schemes is that inner problems are never solved to a prescribed accuracy: an inexact answer is either used as a cut or transports the lower model to a new point.

Inexact oracles and models. The analysis of inexact oracles goes back to Devolder et al. (2014); Stonyakin et al. (2021) introduce the notion of an inexact model of a function. The method of Section 4 relies on exactly this technique: under strong concavity in y, the primal function P admits a smooth inexact model built from approximate solutions of the problem in y.

Complexity separation in saddle-point problems. Gradient sliding (Lan, 2016) separates the numbers of gradient evaluations for the smooth and nonsmooth parts of composite problems. Alkousa et al. (2020) separate the complexities in the blocks x and y for smooth strongly convex-concave problems; Borodich et al. (2025b) obtain complexity separation for strongly convex-strongly concave composite problems of the form $p(x) + R(x, y) - q(y)$; Kovalev et al. (2022) and Thekumparampil et al. (2022) propose optimal methods for bilinear coupling; and Borodich et al. (2025a) establish lower bounds and optimal algorithms that separate the numbers of gradient evaluations and matrix-vector products for bilinearly coupled problems. Jin et al. (2022) obtain sharper rates for separable minimax problems. For smooth convex-concave problems, lower complexity bounds are given by Zhang et al. (2022); classical methods include the extragradient method (Korpelevich, 1976), mirror-prox (Nemirovski, 2004) and smoothing (Nesterov, 2005).

All of these works assume smoothness in both blocks (or a nonsmooth part with a simple proximal step) and give dimension-free bounds. We instead allow a nonsmooth, merely concave y-block and exploit its small dimension. Closer in spirit, Carmon et al. (2021) show that minimizing the maximum of N losses admits near-optimal methods based on ball-oracle acceleration (Carmon et al., 2020). Their oracle, however, accesses one loss at a time, whereas our y-oracle returns a value and a supergradient for the whole small block at once, so the oracle models differ.

Other oracles for the small block. The small dimension of the y-block also makes other types of oracles attractive; we discuss them in Section F. In particular, a derivative of order p in dimension m has only m^p entries, so tensor methods (Nesterov, 2021), which attain optimal rates for functions with Lipschitz higher-order derivatives (Kovalev & Gasnikov, 2022a; Carmon et al., 2022), become affordable for the small block.

F Open questions

1. **Implementable centers.** Our analysis uses the exact center of gravity of the localizer and treats its computation as free. Approximate centers of gravity based on random walks (Bertsimas & Vempala, 2004) satisfy Grünbaum's lemma with a worse constant, and we expect our bounds to carry over with only the absolute constants changed. Vaidya's method (Vaidya, 1996) relies on a different potential, and combining it with certificate transport requires a separate analysis.
2. **A small block with function values only.** If O_y returns only $f(x, y)$, a separation oracle for a concave Lipschitz function in dimension m can be emulated with $\widetilde{O}(m)$ function values (Lee et al., 2018). Since certificate transport needs only the single value $v = f(u_K, y_c)$, one may expect $T_y = \widetilde{O}(m^2 \log \frac{B}{\varepsilon})$ with the same T_x. The difficulty is that planes built from approximate supergradients are no longer exact upper bounds on ϕ, so the bound a must be replaced by an approximate one with controlled error.
3. **Adaptive block length and adaptive steps in y.** Theorem 1 uses a fixed block length K. In practice, one can monitor the fraction of iterations that transport the certificate and adjust K on the fly. The center-of-gravity step could also be replaced by a step of the subgradient ellipsoid method (Rodomanov & Nesterov, 2023), which produces accuracy certificates (Nemirovski et al., 2010) and interpolates between cutting planes and subgradient methods. This would allow a smooth transition from the low-dimensional to the high-dimensional regime in y without switching algorithms.

Finally, the small dimension of the y-block also makes higher-order oracles attractive: a derivative of order p in dimension m has only m^p entries, and tensor methods (Nesterov, 2021; Kovalev & Gasnikov, 2022a; Carmon et al., 2022) could replace cutting planes when f is sufficiently smooth in y.