BILEVEL OPTIMIZATION WITH DISCRETE SUBMODULAR LOWER LEVELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bilevel optimization (BLO) models hierarchical decision-making in applications such as electronic design automation, network planning, and learning systems. While recent advances in continuous settings have substantially expanded the scale and scope of solvable problems, many practical discrete BLO problems remain computationally difficult. This work studies BLO problems with discrete submodular lower levels (LLs) and coupled constraints, and introduces BiGAP, a randomized-smoothed value-function formulation that regularizes discrete response switches while preserving exact discrete oracles. We establish finite penalty exactness and show that the smoothed formulation agrees locally with the smoothed optimistic objective on positive-gap neighborhoods. The resulting first-order algorithms reduce all discrete computations to tractable unconstrained submodular function minimization (SFM), with $\mathcal{O}(\varepsilon^{-4})$ SFM complexity without coupled constraints and $\mathcal{O}(\varepsilon^{-6})$ with coupled constraints under exact dual support and uniform saddle-system conditioning. Experiments on challenging large-scale discrete optimization problems demonstrate the effectiveness and scalability of the proposed approach.

1 INTRODUCTION

BLO models hierarchical decisions where an upper-level (UL) variable anticipates a lower-level (LL) solver's response. Continuous BLO is well studied, underlying hyperparameter optimization, meta-learning, and related ML pipelines (Pedregosa, 2016; Franceschi et al., 2018). Yet many applications instead couple a continuous design with a combinatorial LL response—image segmentation (Boykov & Jolly, 2001; Rother et al., 2004), VLSI partitioning and placement-driven cuts (Fiduccia & Mattheyses, 1982; Alpert, 1998), and other structured discrete decisions (Sinha et al., 2018), and methods for these settings remain limited. Motivated by these applications, we study:

$$\min_{x\in\mathcal{X}} \phi(x) := \min_{y\in\mathcal{Y}^\star(x)} f(x,y), \quad \text{s.t.} \quad \mathcal{Y}^\star(x) := \arg\min_{y\in\mathcal{Y}} \big\{g(x,y) : c(x,y) \le 0\big\}. \tag{1}$$

Here $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ and $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ are the UL and LL objectives, $x \in \mathcal{X} \subset \mathbb{R}^p$ is continuous, $\mathcal{Y} = \{0,1\}^d$ is discrete, and $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^m$ defines the coupled constraints. The optimistic objective $\phi(x)$ selects the smallest UL cost among all LL minimizers.

Existing first-order BLO methods primarily follow two routes, both facing severe technical obstacles in problem (1). The first relies on *implicit differentiation or unrolling* to differentiate a smooth LL solution or its iterative approximation (Pedregosa, 2016; Franceschi et al., 2018). This approach fails here due to **T1) non-differentiability and discontinuity:** because $\mathcal{Y}$ is discrete, the optimal response set $\mathcal{Y}^\star(x)$ is piecewise constant in x. Small perturbations leave the discrete response unchanged, yielding zero informative gradients, while switching points cause abrupt changes that make $\phi(x)$ nonsmooth or discontinuous. Continuous LL relaxations restore differentiability, but introduce **T2) suboptimality of relaxations:** replacing $\mathcal{Y}$ with its convex hull $[0,1]^d$ alters the discrete structure because rounding need not recover an integer optimum and can change optimistic tie-breaking. Generic discrete BLO methods avoid relaxations but rely on exact combinatorial or mixed-integer machinery (Sinha et al., 2018; Lozano et al., 2026), which scales poorly with d.

The second route reformulates BLO into a single-level problem by encoding LL optimality via a value function, followed by penalty or primal-dual optimization (Ye & Zhu, 1995; Shen et al., 2025;

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.

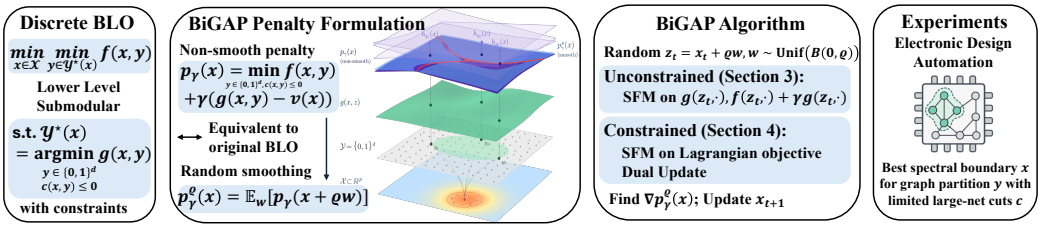


Figure 1: Overview of BiGAP. It replaces the discrete LL response with an exact value-gap penalty, smooths the penalty by random perturbations, and optimizes with SFM-based first-order updates. It covers unconstrained and coupled-constraint cases and is evaluated on EDA graph partitioning.

Jiang et al., 2024). Since it never differentiates the LL solution, it avoids **T1** and **T2**. We focus on a structured regime where $g(x,\cdot)$ is submodular. Without coupled constraints $c(x,y) \le 0$, such subproblems are polynomial-time solvable via unconstrained SFM (Fujishige, 2005; Schrijver, 2000; Orlin, 2009), with pairwise instances further reducible to fast graph cuts (Boykov et al., 2001; Kolmogorov & Zabih, 2004). However, coupled constraints introduce **T3) intractability under LL coupling:** enforcing $c(x,y) \le 0$ destroys submodularity, and even modular constraints make constrained SFM nondeterministic polynomial-time (NP)-hard (Garey et al., 1976). Consequently, the constrained value functions cannot be directly evaluated by SFM. This leads to our central question:

(Q) Can we efficiently optimize (1) *without differentiating or relaxing the discrete LL response?*

We answer this question affirmatively with BiGAP, which combines a value-gap penalty with randomized smoothing. Building on value-function penalty methods (Kwon et al., 2024; Jiang et al., 2024), we penalize LL suboptimality through

$$p_\gamma(x) := \min_{y \in \mathcal{Y}(x)} \{ f(x,y) + \gamma\big(g(x,y) - v(x)\big) \}, \tag{2}$$

$$v(x) := \min_{y \in \mathcal{Y}(x)} g(x,y), \quad \mathcal{Y}(x) = \{ y \in \mathcal{Y} : c(x,y) \le 0 \}.$$

Rather than differentiating or relaxing the discrete LL response, BiGAP smooths this scalar

$$p_\gamma^\varrho(x) := \mathbb{E}_{w \sim \mathrm{Unif}(\mathbb{B}_2^p)} \big[p_\gamma(x + \varrho w) \big], \quad x \in \mathcal{X}. \tag{3}$$

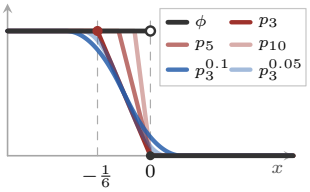


Figure 2: Increasing γ narrows the penalty ramp toward the discontinuous ϕ, while decreasing ϱ brings p_3^ϱ closer to p_3. Appendix B.3 gives the construction.

Construction (3) regularizes the penalty (Figure 2), enabling first-order optimization through sampled discrete oracles. Our contributions are:

C1) Exactness and stationarity. Section 2 shows that p_γ converges to ϕ and is finitely exact on positive-gap neighborhoods. Its smoothing p_γ^ϱ then yields finite-radius Goldstein stationarity of ϕ.

C2) Two-SFM optimization. Section 3 develops BiGAP-SFM for $c \equiv 0$, using two exact unconstrained SFM calls to recover first-order information and attain $\mathcal{O}(\varepsilon^{-4})$ SFM complexity.

C3) Coupled-constraint optimization. Section 4 extends BiGAP through exact bounded Lagrangian support, reducing the constrained values to unconstrained SFM-based dual solves. BiGAP-PD attains $\mathcal{O}(\varepsilon^{-6})$ SFM complexity under the stated sensitivity regularity.

C4) Large-scale experiments. Section 5 studies inverse cut design and EDA problems, demonstrating effective optimization on discrete LL problems with over 12k LL variables.

Related work. Existing BLO methods mainly rely on implicit differentiation or unrolling for continuous LL problems (Pedregosa, 2016; Franceschi et al., 2018), or value-function and penalty reformulations that avoid differentiating the LL response (Ye & Zhu, 1995; Liu et al., 2022; Shen et al., 2025; Jiang et al., 2024). Discrete and mixed-integer BLO instead uses combinatorial, mixed-integer, or discretization-based methods (Sinha et al., 2018; Kleinert et al., 2021; Lozano et al., 2026; Jiang et al., 2025), while differentiable combinatorial optimization constructs surrogate gradients through discrete solvers (Vlastelica et al., 2020; Berthet et al., 2020). BiGAP instead preserves the exact binary LL problem and exploits submodularity, reducing LL computation to exact unconstrained SFM (Fujishige, 2005; Schrijver, 2000; Orlin, 2009), with pairwise cases reducible to graph cuts (Boykov

Method	Coupled constraints	LL derivatives	LL primitive	Discrete LL
Implicit / unrolling	Limited	Required	Gradient / Hessian	No
Value-function methods	Limited	Not required	Continuous optimization	No
BLOCC	Yes	Not required	Primal–dual gradients	No
DIVIDE-BLO	Yes	Not required	Finite sampling	Low-dimensional
Discrete / mixed-integer BLO	Yes	Not required	Combinatorial / mixed-integer	Exact, NP-hard
BiGAP	**Yes**	**Not required**	**Unconstrained SFM**	**Exact,** $\mathcal{O}(\varepsilon^{-4})$ **/** $\mathcal{O}(\varepsilon^{-6})$

Table 1: Comparison of representative bilevel optimization approaches: implicit differentiation and unrolling (Pedregosa, 2016; Franceschi et al., 2018), value-function and penalty methods (Ye & Zhu, 1995; Liu et al., 2022; Shen et al., 2025), BLOCC (Jiang et al., 2024), DIVIDE-BLO (Jiang et al., 2025), and discrete or mixed-integer methods (Sinha et al., 2018; Lozano et al., 2026). For BiGAP, the final column reports SFM complexity for unconstrained and coupled settings, respectively.

et al., 2001; Kolmogorov & Zabih, 2004). Coupled constraints are handled through supported Lagrangian SFM subproblems. Table 1 summarizes these distinctions, with a more comprehensive discussion in Appendix A.

Notation. For $q \in \mathbb{N}$, let $[q] := \{1, \dots, q\}$. The dimensions p, d, m denote the continuous variables, binary variables, and coupled constraints. Vector inequalities are componentwise, $\mathbf{1}$ is the all-ones vector, and $\|\cdot\|_2$, $|\cdot|$, $\mathrm{conv}(\cdot)$, and $\mathrm{dist}(\cdot,\cdot)$ denote the Euclidean norm, absolute value, convex hull, and Euclidean distance. Let $\mathbb{B}_2^p$ be the unit Euclidean ball, $B(x,\varrho) := x + \varrho\mathbb{B}_2^p$, and $\Pi_{\mathcal{X}}$ the Euclidean projection onto $\mathcal{X}$. We use $\partial_C h(x)$ and $\partial_\varrho h(x)$ for the Clarke and radius-ϱ Goldstein subdifferentials, respectively, $\mathbb{E}$ for expectation, $\mathrm{Unif}(S)$ for the uniform distribution on S, and $\mathcal{O}(\cdot)$ for order notation. Gradients are with respect to x unless stated otherwise.

2 The value-gap penalty for an unconstrained LL

Section 2.1 establishes penalty and smoothing equivalence, while Section 2.2 introduces the submodular structure and SFM machinery underlying the penalty.

2.1 The Value-Gap Penalty in Discrete LL Settings

When $c \equiv 0$, the penalized LL problem $p_\gamma(x)$ reduces to a discrete minimization over $\mathcal{Y}$. Exactness follows directly from the combinatorial separation of discrete spaces, obviating smoothness or continuous relaxation, and is formalized via regularity conditions ensuring boundedness.

Assumption 1 (Regularity)**.** *Suppose $\mathcal{X} \subset \mathbb{R}^p$ is convex, nonempty, and compact, $\mathcal{Y}(x)$ is a finite discrete set, and, for every $y \in \mathcal{Y}(x)$, the functions $f(\cdot, y)$, $g(\cdot, y)$, and $c(\cdot, y)$ are continuously differentiable on some open set U containing $\mathcal{X}$.*

This is the standard compact finite-branch regularity setting for marginal-value and nonsmooth analysis (Clarke, 1983; Bonnans & Shapiro, 2000). Since $\mathcal{X}$ is compact and U is open, there exists $\varrho_0 > 0$ such that $\mathcal{X}^\varrho := \mathcal{X} + \varrho\mathbb{B}_2^p \subset U$ for all $\varrho \in (0, \varrho_0)$. Thus, randomized queries below remain in U. Finiteness of $\mathcal{Y}$ implies a positive LL value gap between optimal and suboptimal labels whenever the latter exist, yielding the following.

Proposition 1 (Value Equivalence)**.** *Define the LL gap $\delta(x) := \min_{y \in \mathcal{Y}(x) \setminus \mathcal{Y}^\star(x)} (g(x,y) - v(x))$, and the global range $\Delta_f := \max_{(x,y) \in \mathcal{X} \times \mathcal{Y}(x)} f(x,y) - \min_{(x,y) \in \mathcal{X} \times \mathcal{Y}(x)} f(x,y)$ of the UL objective. Under Assumption 1, $\delta(x) \in (0, +\infty]$ and $\Delta_f < \infty$. Then,*

(i) $p_\gamma(x) \le \phi(x)$, the mapping $\gamma \mapsto p_\gamma(x)$ is nondecreasing, and $p_\gamma(x) \uparrow \phi(x)$ as $\gamma \to \infty$.

(ii) Pointwise exactness $p_\gamma(x) = \phi(x)$ holds whenever $\gamma\delta(x) \ge \Delta_f$.

The proof of Proposition 1 is provided in Appendix B.1. The displayed decomposition resolves **T1)** because the value gap vanishes on $\mathcal{Y}^\star(x)$, so the penalty selects the LL optimal labels with the smallest upper cost. When exactness holds on an open set, local minimality and subdifferential structure also transfer between p_γ and ϕ.

Corollary 1 (Local Optimality Equivalence). *Let $K \subseteq \mathcal{X}$ be an open neighborhood of $\bar{x} \in \mathrm{int}(\mathcal{X})$. For any $\gamma > \sup_{x\in K,\, y\in\mathcal{Y}(x)\setminus\mathcal{Y}^\star(x)} \frac{\left[\inf_{z\in K}\phi(z)-f(x,y)\right]_+}{g(x,y)-v(x)}$, if $\bar{x}$ minimizes either objective over K, then $\bar{x}$ is a local minimizer of both p_γ and ϕ. If, in addition, p_γ is locally Lipschitz near $\bar{x}$ and $\inf_{x\in B(\bar{x},r)}\delta(x) > 0$ for some $r > 0$ with $\gamma \inf_{x\in B(\bar{x},r)}\delta(x) > \Delta_f$, then Clarke stationarity at $\bar{x}$ and radius-ϱ Goldstein stationarity for every $\varrho < r$ are equivalent between p_γ and ϕ.*

The proof of Corollary 1 is given in Appendix B.2. The penalty threshold is local and compares the possible UL improvement of an LL-suboptimal label with its LL value gap. It remains finite across response switches when these two quantities vanish at comparable rates, providing a discrete counterpart to the growth or error-bound conditions associated with Polyak–Łojasiewicz assumptions in continuous BLO (Shen et al., 2025). Lipschitzness of p_γ is established by Proposition 3(i) for $c \equiv 0$, and by Proposition 4 for coupled constraints case. Figure 2 illustrates this local penalty geometry, with the exact construction given in Appendix B.3.

2.2 Submodular Lower Levels: Properties, Solvers, and Examples

While Section 2.1 establishes penalty exactness for any finite set $\mathcal{Y}$, evaluating $p_\gamma(x)$ efficiently requires tractability in the LL subproblem. Submodularity over the hypercube $\mathcal{Y} = \{0,1\}d$ provides the combinatorial structure needed for exact, polynomial-time minimization.

Definition 1 (Submodularity). A binary function $h : \{0,1\}^d \to \mathbb{R}$ is submodular if it exhibits diminishing marginal returns: for all $y \in \{0,1\}^d$ and distinct $i, j \in [d]$ with $y_i = y_j = 0$,

$$\Delta_{ij}h(y) := h(y_{i=1,j=1}) - h(y_{i=1,j=0}) - h(y_{i=0,j=1}) + h(y_{i=0,j=0}) \le 0, \tag{4}$$

where $y_{i=a,j=b}$ denotes y with its i-th and j-th coordinates set to $a, b \in \{0,1\}$, leaving all other components unchanged. A function is *modular* if equality holds for all such y, i, j.

To guarantee that penalized combinations of UL and LL objectives remain submodular, we impose a relative submodularity condition.

Assumption 2 (LL Submodularity). *For every $x \in U$, $g(x,\cdot)$ is submodular, $c(x,\cdot)$ is componentwise submodular, and there exists $\beta \ge 0$ such that $f(x,\cdot) + \beta g(x,\cdot)$ is submodular in $\mathcal{Y} = \{0,1\}^d$.*

This condition is checkable through the discrete second differences in Proposition 2, and Appendix C gives graph-cut and structured submodular examples (Fujishige, 2005; Kolmogorov & Zabih, 2004). It allows supermodular directions in $f(x,\cdot)$ when they are offset by submodularity in $g(x,\cdot)$.

Proposition 2 (Checkable Relative-Submodularity). *Suppose $g(x,\cdot)$ is submodular and $c(x,\cdot)$ is componentwise submodular. Define $\beta_\star(x) := \max\left\{0,\ \max_{\Delta_{ij}g(x,y)<0} \frac{\Delta_{ij}f(x,y)}{-\Delta_{ij}g(x,y)}\right\}$, where the inner maximum is $-\infty$ if no strictly submodular direction exists. Then $f(x,\cdot)+\beta g(x,\cdot)$ is submodular if and only if $\beta \ge \beta_\star(x)$ and $\Delta_{ij}f(x,y) \le 0$ whenever $\Delta_{ij}g(x,y) = 0$. Consequently, Assumption 2 holds uniformly whenever this zero-slack condition holds and $\sup_{x\in U}\beta_\star(x) < \infty$. Moreover, adding any nonnegative linear combination of the components of $c(x,\cdot)$ to an already submodular objective preserves submodularity.*

The proof of Proposition 2 is in Appendix B.5, while Appendix C gives graph-cut and structured submodular examples. Assumption 2 enables exact polynomial-time function evaluations via SFM (Grötschel et al., 1981; Orlin, 2009), which reduces to an exact minimum s–t cut for graph-representable binary energies (Boykov et al., 2001; Kolmogorov & Zabih, 2004). We write

$$(y^\star, \mathsf{S}^+) = \mathrm{SFM}(g(x,\cdot), \mathsf{S}), \quad y^\star \in \arg\min_{y\in\mathcal{Y}} g(x,y), \tag{5}$$

where S and S^+ denote the incoming and updated solver states, such as residual graph capacities and search trees (Kohli & Torr, 2007). Warm-starting from S can accelerate wall-clock performance without altering exact oracle semantics.

When $c \equiv 0$, the penalized objective reduces to two unconstrained discrete minima. For $\gamma \ge \beta$ with $\gamma > 0$, define $v_\gamma(x) := \min_{y\in\mathcal{Y}}(\gamma^{-1}f(x,y) + g(x,y))$, so $p_\gamma(x) = \gamma(v_\gamma(x) - v(x))$. Each term is a finite minimum of C^1 branches and requires one exact SFM call under Assumption 2. Thus p_γ is continuous, Proposition 1 gives $p_\gamma \uparrow \phi$, and compactness of $\mathcal{X}$ yields convergence of global

minimizers to $\mathcal{X}^\star := \arg\min_{x\in\mathcal{X}} \phi(x)$ as $\gamma \to \infty$. Proposition 7 in Appendix B.6 formalizes this recovery. The same SFM solutions provide first-order information. After shrinking U to a bounded open convex neighborhood of $\mathcal{X}$ whose closure remains in the regularity domain, continuity of the gradients and finiteness of $\mathcal{Y}$ give constants $L_f, L_g, L_c < \infty$ such that

$$\|\nabla_x f(x,y)\|_2 \le L_f,\ \|\nabla_x g(x,y)\|_2 \le L_g,\ \|\nabla_x c(x,y)\|_F \le L_c, \quad \forall x \in U, y \in \mathcal{Y}. \tag{6}$$

These bounds turn the two minimizing labels into the following almost-everywhere oracle.

Proposition 3 (First-Order Oracle)**.** *Suppose Assumptions 1, 2 hold, $c \equiv 0$ and $\gamma > \beta$. Then,*

(i) p_γ is L_γ-Lipschitz with $L_\gamma \le L_f + 2\gamma L_g$, and v_γ and v are differentiable a.e.

(ii) Outside null sets, for any $y_1 \in \mathcal{Y}^\star_\gamma(x) := \arg\min_{y\in\mathcal{Y}}(\frac{1}{\gamma}f(x,y)+g(x,y))$ and $y_2 \in \mathcal{Y}^\star(x)$,

$$u(x,y_1,y_2) := \nabla_x f(x,y_1) + \gamma\big(\nabla_x g(x,y_1) - \nabla_x g(x,y_2)\big) = \nabla p_\gamma(x) \tag{7}$$

(iii) For all $x \in \mathcal{X}$, $\partial_C p_\gamma(x) \subseteq \gamma\partial_C v_\gamma(x) - \gamma\partial_C v(x)$, and $u(x,y_1,y_2)$ is in the right-hand set.

The proof of Proposition 3 is provided in Appendix B.7. The inclusion in Proposition 3(iii) can be strict at non-differentiable points, meaning an arbitrary oracle selection at an exceptional point yields a difference-of-convex subgradient estimate rather than a true Clarke subgradient of p_γ. For instance, setting $d = 1$, $f \equiv 0$, $g(x,0) = 0$, and $g(x,1) = x$ gives $v_\gamma(x) = v(x) = \min\{0,x\}$, so $p_\gamma \equiv 0$ and $\partial_C p_\gamma(0) = \{0\}$. At $x = 0$, selecting $y_1 = 1$ and $y_2 = 0$ yields $u(0,1,0) = \gamma \neq 0$.

3 BiGAP-SFM: RANDOMIZED FIRST-ORDER DESCENT FOR $c \equiv 0$

BiGAP-SFM is the SFM implementation of BiGAP for $c \equiv 0$. It minimizes the smoothed penalty p^ϱ_γ in (3) using exact stochastic gradients of p_γ at randomized queries. Throughout the algorithms, the subscript F denotes the penalized objective containing $f + \gamma g$, or its equivalent positive rescaling.

At iteration t, BiGAP samples $w^{t,j} \sim \mathrm{Unif}(\mathbb{B}^p_2)$ and queries $z^{t,j} := x^t + \varrho w^{t,j}$ for $j \in [M]$. Each query evaluates (7) through two exact stateful SFM calls of the form (5), producing $y_1^{t,j} \in \arg\min_{y\in\mathcal{Y}}(g(z^{t,j},y) + \gamma^{-1}f(z^{t,j},y))$ and $y_2^{t,j} \in \arg\min_{y\in\mathcal{Y}} g(z^{t,j},y)$. The perturbation is not projected, and $\varrho \in (0,\varrho_0)$ keeps every query inside U.

The method carries persistent solver states S_1^t and S_2^t for the two SFM families. It forms $G^{t,j} := u(z^{t,j}, y_1^{t,j}, y_2^{t,j})$, averages $\bar{G}^t := M^{-1}\sum_{j=1}^M G^{t,j}$, and updates $x^{t+1} := \Pi_{\mathcal{X}}(x^t - \eta\bar{G}^t)$. Algorithm 1 gives the complete procedure.

Algorithm 1 BiGAP without coupled LL constraints

Require: $x^0 \in \mathcal{X}$, $\gamma \ge \beta$, $\varrho, \eta > 0$, batch size M, horizon T, states $\mathsf{S}_1, \mathsf{S}_2 \leftarrow \varnothing$
for $t = 0, 1, \ldots, T-1$ **do**
  for $j = 1, \ldots, M$ **do**
    Sample $w^{t,j} \sim \mathrm{Unif}(\mathbb{B}^p_2)$ and set $z^{t,j} \leftarrow x^t + \varrho w^{t,j}$
    $(y_1^{t,j}, \mathsf{S}_1) \leftarrow \mathrm{SFM}\big(g(z^{t,j},\cdot) + \gamma^{-1}f(z^{t,j},\cdot), \mathsf{S}_1\big)$
    $(y_2^{t,j}, \mathsf{S}_2) \leftarrow \mathrm{SFM}\big(g(z^{t,j},\cdot), \mathsf{S}_2\big)$
    $G^{t,j} \leftarrow u(z^{t,j}, y_1^{t,j}, y_2^{t,j})$ by (7)
  end for
  $\bar{G}^t \leftarrow M^{-1}\sum_{j=1}^M G^{t,j}$
  $x^{t+1} \leftarrow \Pi_{\mathcal{X}}\big(x^t - \eta\bar{G}^t\big)$
end for
Sample $x^{t_{\mathrm{out}}}$ uniformly from $\{x^0, \ldots, x^{T-1}\}$
$(y_F, \mathsf{S}_1) \leftarrow \mathrm{SFM}\big(g(x^{t_{\mathrm{out}}},\cdot) + \gamma^{-1}f(x^{t_{\mathrm{out}}},\cdot), \mathsf{S}_1\big)$
Ensure: $(x^{t_{\mathrm{out}}}, y_F)$

Convergence analysis. We next establish the convergence and SFM complexity of Algorithm 1. By construction, every query $z^{t,j} = x^t + \varrho w^{t,j}$ lies in $\mathcal{X}^\varrho \subset U$, where Proposition 3(i) makes p_γ L_γ-Lipschitz. For a fixed penalty parameter $\gamma \ge \beta$ with $\gamma > 0$, initialization $x^0 \in \mathcal{X}$, and smoothing radius $\varrho \in (0,\varrho_0)$, define the initial objective gap $\Delta_{\gamma,0} := p_\gamma(x^0) - \inf_{x\in\mathcal{X}^\varrho} p_\gamma(x) < \infty$. We characterize the randomized search direction through the smoothed penalty p^ϱ_γ in (3) and measure first-order progress using the generalized gradient mapping (Lin et al., 2022):

$$\mathcal{G}_{\eta,\varrho}(x) := \eta^{-1}\left(x - \Pi_{\mathcal{X}}\big(x - \eta\nabla p^\varrho_\gamma(x)\big)\right). \tag{8}$$

By Proposition 3, the nondifferentiability set of p_γ is Lebesgue-null. Since each randomized query $z^{t,j} = x^t + \varrho w^{t,j}$ has an absolutely continuous distribution conditional on x^t, we have $G^{t,j} =$

$\nabla p_\gamma(z^{t,j})$ almost surely. Moreover, $\|G^{t,j}\|_2 \le L_\gamma$ almost surely, and random sampling gives

$$\mathbb{E}(\bar{G}^t \mid x^t) = \nabla p_\gamma^\varrho(x^t), \quad \text{and} \quad \mathbb{E}\left(\|\bar{G}^t - \nabla p_\gamma^\varrho(x^t)\|_2^2 \mid x^t\right) \le \frac{L_\gamma^2}{M}. \tag{9}$$

These properties provide the stochastic-gradient ingredients used below.

Theorem 1 (Convergence of Algorithm 1)**.** *Suppose Assumptions 1, 2 hold with $c \equiv 0$. Fix $\gamma \ge \beta$ with $\gamma > 0$ and $\varrho \in (0, \varrho_0)$. Then, ∇p_γ^ϱ is H-Lipschitz on $\mathcal{X}$ with $H \le \mathcal{O}(L_\gamma\sqrt{p}/\varrho)$. Algorithm 1 with stepsize $\eta = 1/(2H)$, batch size $M = \lceil 12L_\gamma^2\varepsilon^{-2}\rceil$, and outer round*

$$T = \mathcal{O}\left(H(\Delta_{\gamma,0} + \varrho L_\gamma)\varepsilon^{-2}\right) = \mathcal{O}\left(\sqrt{p}\left(L_\gamma^2\varepsilon^{-2} + \Delta_{\gamma,0}L_\gamma\varrho^{-1}\varepsilon^{-2}\right)\right), \tag{10}$$

gives output $(x^{t_{\text{out}}}, y_F)$ and guarantees $T^{-1}\sum_{t=0}^{T-1}\mathbb{E}\big(\|\mathcal{G}_{\eta,\varrho}(x^t)\|_2^2\big) \le \varepsilon^2$. Uniform output sampling also gives $\mathbb{E}\big(\|\mathcal{G}_{\eta,\varrho}(x^{t_{\text{out}}})\|_2^2\big) \le \varepsilon^2$.

The proof of Theorem 1 is provided in Appendix B.8. The algorithm makes exactly $2MT + 1$ SFM calls, including the final call that returns y_F, hence the total number of SFM calls is $\mathcal{O}\left(\sqrt{p}\left(L_\gamma^4\varepsilon^{-4} + \Delta_{\gamma,0}L_\gamma^3\varrho^{-1}\varepsilon^{-4}\right)\right)$. The following corollary connects this guarantee to finite-radius Goldstein stationarity of the original nonsmooth penalty and, on positive-gap neighborhoods, the optimistic objective.

Corollary 2 (Goldstein stationarity of the nonsmooth penalty)**.** *Suppose the conditions of Theorem 1 hold and let $x^{t_{\text{out}}}$ be the output of Algorithm 1. If $\eta L_\gamma \le \varrho$ and $B(x^{t_{\text{out}}}, \varrho) \subset \operatorname{int}\mathcal{X}$, then $\operatorname{dist}\big(0, \partial_\varrho p_\gamma(x^{t_{\text{out}}})\big) \le \|\mathcal{G}_{\eta,\varrho}(x^{t_{\text{out}}})\|_2$. Consequently, if the ball condition holds almost surely, then $\mathbb{E}\big(\operatorname{dist}(0, \partial_\varrho p_\gamma(x^{t_{\text{out}}}))\big) \le \varepsilon$. Whenever the local-equivalence conditions of Corollary 1 hold on the same ball, both conclusions hold with ϕ in place of p_γ.*

The proof of Corollary 2 is provided in Appendix B.9. We next restore coupled LL constraints, where the penalty connection remains valid but evaluating its constrained discrete values requires an additional supported-dual construction.

4 Extension: coupled LL constraints

This section restores the coupled LL constraints in (1). We retain randomized smoothing p_γ^ϱ to address **T1)** and keep the LL variables binary, $y \in \{0,1\}^d$, to avoid the relaxation errors in **T2)**. The new difficulty is **T3)**: although unconstrained SFM is polynomial-time solvable (Fujishige, 2005; Schrijver, 2000), evaluating p_γ requires two constrained submodular minimizations, to which unconstrained SFM cannot be directly applied. We address this through exact Lagrangian support.

To evaluate the constrained values in (2), we use Lagrangian duality with dual functions $d_h(x,\nu) := min_{y\in Y}\{h(x,y) + \nu^\top c(x,y)\}$, where $\nu \ge 0$ and $h \in \{g, f+\gamma g\}$. Weak duality guarantees only a lower bound on each constrained value. We therefore assume that both dual problems attain the corresponding constrained values at uniformly bounded multipliers. Appendix D gives checkable conditions for exact duality in Assumption 3.

Assumption 3 (Exact duality and strict feasibility)**.** *Fix $\gamma \ge \beta$, $\gamma > 0$, and $\varrho \in (0, \varrho_0)$. Assume (i) there exists $\sigma > 0$ such that, for any $x \in X^\varrho$, there is $\bar{y} \in \mathcal{Y} : c(x,\bar{y}) \le -\sigma\mathbf{1}$. (ii) For any $h \in \{g, f+\gamma g\}$, there exist multipliers $\nu_h^\star(x) \ge 0 : d_h(x, \nu_h^\star(x)) = \min_{y\in\mathcal{Y}(x)} h(x,y),\ x \in X^\varrho$.*

Under Assumption 3, both constrained problems satisfy strong duality with attained dual optima. We now approximate these optima using projected supergradient ascent with exact SFM subproblems.

Proposition 4 (Exact dual representation and Lipschitz continuity)**.** *Under Assumptions 1, 2, 3, there exist finite $R, Q \ge 0$ such that, with $\Lambda = [0,R]^m$ and $\mathcal{M} = [0,Q]^m$,*

$$p_\gamma(x) = \max_{\lambda\in\Lambda}\min_{y\in\mathcal{Y}}\{f(x,y) + \gamma g(x,y) + \lambda^\top c(x,y)\} - \gamma\max_{\mu\in\mathcal{M}}\min_{z\in\mathcal{Y}}\{g(x,z) + \mu^\top c(x,z)\}, \tag{11}$$

for all $x \in \mathcal{X}^\varrho$. p_γ is L_γ-Lipschitz on $\mathcal{X}^\varrho$, where $L_\gamma := L_f + mRL_c + 2\gamma L_g + \gamma mQL_c < \infty$.

The proof of Proposition 4 is in Appendix E.1. Algorithm 2 runs projected supergradient ascent on each dual problem in (11) (Rockafellar, 1970; Bertsekas, 2016). Each step is one exact SFM call

Algorithm 2 DUAL

Require: objective h, constraint c, UL variable x, budget K, stepsize θ, multiplier ν^0, SFM state S

Initialize $s_\nu \leftarrow 0$, $s_h \leftarrow 0$, $S_c \leftarrow 0$
for $k = 0, 1, \dots, K-1$ **do**
  $(y^k, \mathsf{S}) \leftarrow \mathrm{SFM}\big(h(x,\cdot) + (\nu^k)^\top c(x,\cdot), \mathsf{S}\big)$
  $s_\nu \mathrel{+}= \nu^k$
  $s_h \mathrel{+}= \nabla_x h(x, y^k)$
  $S_c \mathrel{+}= \nabla_x c(x, y^k)$
  $\nu^{k+1} \leftarrow \Pi_{\mathbb{R}^m_+}\big(\nu^k + \theta c(x, y^k)\big)$
end for
$\widehat{G}_h \leftarrow K^{-1} s_h + K^{-2} s_\nu^\top S_c$

Ensure: $\widehat{G}_h$, $y_h \leftarrow y^{K-1}$, ν^K, and S

Algorithm 3 BiGAP-PD

Require: $x^1 \in \mathcal{X}$, $\gamma \geq \beta$, $\varrho \in (0, \varrho_0)$, stepsizes $\eta, \theta_\lambda, \theta_\mu > 0$, batch size $M \geq 1$, budgets T, K_λ, K_μ. $\mu, \lambda \leftarrow 0$, $\mathsf{S}_g, \mathsf{S}_F \leftarrow \varnothing$

for $t = 1, 2, \dots, T$ **do**
  for $j = 1, \dots, M$ **do**
    Sample $w^{t,j} \sim \mathrm{Unif}(\mathbb{B}_2^p)$, set $z^{t,j} \leftarrow x^t + \varrho w^{t,j}$
    $(\widehat{G}_g^{t,j}, y_g^{t,j}, \mu, \mathsf{S}_g) \leftarrow \textsc{Dual}\big(g, c, z^{t,j}, K_\mu, \theta_\mu, \mu, \mathsf{S}_g\big)$
    $(\widehat{G}_F^{t,j}, y_F^{t,j}, \lambda, \mathsf{S}_F) \leftarrow \textsc{Dual}\big(f + \gamma g, c, z^{t,j}, K_\lambda, \theta_\lambda, \lambda, \mathsf{S}_F\big)$
  end for
  $\bar{G}^t \leftarrow M^{-1} \sum_{j=1}^M (\widehat{G}_F^{t,j} - \gamma \widehat{G}_g^{t,j})$
  $x^{t+1} \leftarrow \Pi_{\mathcal{X}}\big(x^t - \eta \bar{G}^t\big)$
end for
Sample $x^{t_{\mathrm{out}}}$ uniformly from $\{x^1, \dots, x^T\}$
$(\widehat{G}_F^{\mathrm{out}}, y_F, \lambda, \mathsf{S}_F) \leftarrow \textsc{Dual}\big(f + \gamma g, c, x^{t_{\mathrm{out}}}, K_\lambda, \theta_\lambda, \lambda, \mathsf{S}_F\big)$

Ensure: $(x^{t_{\mathrm{out}}}, y_F)$

(5). The routine takes and returns the multiplier and the SFM state. So Algorithm 3 warm-starts one trajectory for g and one for $f+\gamma g$. For the analysis, let $\bar{\nu}^K$ be the average multiplier and $\bar{a}^K \in \Delta(\mathcal{Y})$ the empirical distribution of $y^0, \dots, y^{K-1}$. Let $L_h(x, a, \nu) := \sum_y a_y \big(h(x,y) + \nu^\top c(x,y)\big)$, and let $\mathcal{B} \in \{\Lambda, \mathcal{M}\}$ be the box of h. The saddle gap and the saddle set are

$$\begin{aligned} \mathrm{Gap}_h(x, a, \nu) &:= \max_{\nu' \in \mathcal{B}} L_h(x, a, \nu') - \min_{a' \in \Delta(\mathcal{Y})} L_h(x, a', \nu), \\ \mathcal{S}_h(x) &:= \big\{(\nu, a) : \mathrm{Gap}_h(x, a, \nu) = 0\big\}. \end{aligned} \tag{12}$$

The running sums return $\widehat{G}_h = G_h(x, \bar{\nu}^K, \bar{a}^K)$, where $G_h(x, \nu, a) := \nabla_x L_h(x, a, \nu) = \sum_y a_y \big(\nabla_x h(x,y) + \nu^\top \nabla_x c(x,y)\big)$. Algorithm 3 uses $\widehat{G}_F - \gamma \widehat{G}_g$. If $c \equiv 0$, the multiplier term is zero, and this is Proposition 3 at differentiability points. With constraints, Proposition 10 gives an $\mathcal{O}(K^{-1/2})$ saddle gap. The following condition prevents the active linear systems of the two supported problems from becoming singular as the query varies.

Assumption 4 (Uniform supported-basis conditioning)**.** *For each supported objective $h \in \{f + \gamma g, g\}$, every maximal linearly independent collection of rows active at a supported saddle point has smallest singular value at least $\kappa_{\mathrm{reg}}^{-1} > 0$, uniformly over $x \in \mathcal{X}^\varrho$ and the compact multiplier region reached by Algorithm 2.*

Under this condition, which Appendix E.2 spells out as an explicit coefficient system checkable from the branch values and constraint vectors, Hoffman's lemma gives $\mathrm{dist}((\nu, a), \mathcal{S}_h(x)) \leq \kappa_{\mathrm{reg}} \mathrm{Gap}_h(x, a, \nu)$ (Hoffman, 1952; Robinson, 1981). Consequently, V_h is differentiable on $\mathrm{int}(\mathcal{X}^\varrho)$ with $\nabla V_h(x) = G_h(x, \nu^\star, a^\star)$ for every saddle pair. Since G_h is Lipschitz in (ν, a), the approximation bound of Proposition 10 (with supporting lemmas in Appendix E.2) gives

$$\|\widehat{G}_h(x) - \nabla V_h(x)\|_2 \leq (B_h + \sqrt{m} L_c)\, \kappa_{\mathrm{reg}}\, \mathrm{Gap}_h(x, \bar{a}^K, \bar{\nu}^K) \leq (B_h + \sqrt{m} L_c)\, \kappa_{\mathrm{reg}} \Gamma_h / \sqrt{K}, \tag{13}$$

with B_h a bound on the branch gradients. The condition is not standard: it rules out kinks of V_h. Appendix E.2 gives a subclass where it holds and a switching example where it fails. Smoothing turns (13) into the bias bound below.

Proposition 5 (Smoothed first-order accuracy)**.** *Suppose Assumptions 1, 2, 3, and 4 hold. There exist finite constants C_F and C_g such that, conditional on the incoming multiplier and SFM states, the direction constructed from the two independent outputs of* DUAL *satisfies*

$$\big\|\mathbb{E}_w\big[\widehat{G}(x + \varrho w)\big] - \nabla p_\gamma^\varrho(x)\big\|_2 \leq C_F K_\lambda^{-1/2} + C_g K_\mu^{-1/2}, \quad x \in \mathcal{X}. \tag{14}$$

Appendix E.3 derives the required saddle-set error bound and proves Proposition 5. Set $C_{\mathrm{or}} := \max\{C_F, C_g\}$. The constants $B_{f+\gamma g}$ and B_g defined in the setup of Appendix E uniformly bound the two supported directions over the multiplier trajectories.

Theorem 2 (Convergence of Algorithm 3). *Suppose Assumptions 1, 2, 3, and 4 hold. Fix* $\varrho \in (0, \varrho_0)$ *and* $\varepsilon > 0$*, and define* $\Delta_{\gamma,1} := p_\gamma(x^1) - \inf_{x \in \mathcal{X}^\varrho} p_\gamma(x)$*. Choose* $H > 0$ *satisfying* $\mathrm{Lip}(\nabla p_\gamma^\varrho) \le H$ *and* $H = \mathcal{O}(L_\gamma \sqrt{p}/\varrho)$ *whenever* $L_\gamma > 0$*. Run Algorithm 3 with* $\eta = 1/(2H)$*. It suffices to take* $M = \max\{1, \lceil 36(B_{f+\gamma g} + \gamma B_g)^2 \varepsilon^{-2} \rceil\}$*,* $K_\lambda = K_\mu = \max\{1, \lceil 144 C_{\mathrm{or}}^2 \varepsilon^{-2} \rceil\}$*, and* $T = \max\{1, \lceil 24H(\Delta_{\gamma,1} + \varrho L_\gamma)\varepsilon^{-2} \rceil\}$ *to obtain* $T^{-1} \sum_{t=1}^T \mathbb{E}\|\mathcal{G}_{1/(2H),\varrho}(x^t)\|_2^2 \le \varepsilon^2$.

Corollary 3 (Single-query constrained convergence). *Suppose Assumptions 1, 2, 3, 4 hold. Set* $M = 1$*,* $\eta = \varepsilon^2/[27H \max\{(B_{f+\gamma g} + \gamma B_g)^2, \varepsilon^2\}]$*,* $K_\lambda = K_\mu = \max\{1, \lceil 108 C_{\mathrm{or}}^2 \varepsilon^{-2} \rceil\}$*, and* $T = \max\{1, \lceil 27(\Delta_{\gamma,1} + \varrho L_\gamma)/(\eta \varepsilon^2) \rceil\}$*. Then Algorithm 3 satisfies the stationarity guarantee of Theorem 2 with reference stepsize* $1/(2H)$*. Its SFM complexity is* $T(K_\lambda + K_\mu) + K_\lambda = \mathcal{O}(\varepsilon^{-6})$.

The proofs of Theorem 2 and Corollary 3 are provided in Appendix E.4. The corollary trades the batch size for additional outer iterations without changing the SFM complexity order.

Remark 1 (Smaller inner budget). For $M = 1$ and $K_\lambda = K_\mu = K$, a horizon-dependent outer stepsize of order $T^{-1/2}$ gives $T^{-1} \sum_{t=1}^T \mathbb{E}\|\mathcal{G}_{1/(2H),\varrho}(x^t)\|_2^2 = \mathcal{O}(T^{-1/2}) + \mathcal{O}(K^{-1})$. The choice $K = 1$ is the existing algorithm with two SFM calls per outer iteration and $2T + 1$ calls including final recovery. By (RATE′), the stationarity bound depends on $T^{-1} \sum_{t=1}^T \mathbb{E}\|\mathbb{E}[\bar{G}^t \mid \mathcal{F}_t] - \nabla p_\gamma^\varrho(x^t)\|_2^2$ along this persistent trajectory. If this average vanishes under a chosen warm-start schedule, the corresponding stationarity bound vanishes.

The finite-radius interpretation follows Corollary 2, with transfer to ϕ supplied by Corollary 1 on positive-gap neighborhoods. BiGAP-PD uses projected ascent with exact SFM subproblems, and Assumption 4 yields the uniform polyhedral error bound used to recover first-order directions.

5 Experiments

We first illustrate BiGAP on an inverse cut-design problem with an exact reference solution, then evaluate its constrained extension on a large-scale electronic design automation task.

5.1 Toy example: Inverse cut design

We tune edge lengths on a 4×7 grid strip so that a target cut becomes the minimum cut. Column 0 is the source side, column 6 the sink side, and the 20 vertices in between carry binary labels, so $\mathcal{Y} = \{0,1\}^{20}$. Three column gaps are corridors, with edge lengths $1 + \frac{x_1}{2}$, $1 - \frac{x_1}{2}$, and $1 + \frac{x_2}{2}$; all other edges have length 10. Let $\mathcal{X} := [-0.6, 0.6] \times [-0.6, -0.2]$, y_{tar} the cut through corridor 1, $g(x,y) = \sum_{(i,j)\in E} w_{ij}(x)|y_i - y_j|$, and $f(x,y) = \frac{1}{2}\|x - x_0\|_2^2 + \frac{1}{8}\|y - y_{\mathrm{tar}}\|_1$. Then g is a graph cut with positive weights and f is modular in y, so Assumptions 1 and 2 hold with $\beta = 0$ and each lower call in Algorithm 1 is one max-flow (Appendix G).

The exact solution is known. The minimum cut is one of the three corridor cuts, with lengths $L_1 = 4 + 2x_1$, $L_2 = 4 - 2x_1$, $L_3 = 4 + 2x_2$ (Appendix G). The target wins iff $x_1 \le 0$ and $x_1 \le x_2$. For $x_{0,1} > x_{0,2}$, the closest such point of $\mathcal{X}$ is $x^\star = (m, m)$, $m = \min\{\frac{x_{0,1}+x_{0,2}}{2}, -0.2\}$; a wrong cut costs at least 1 while $\frac{1}{2}\|x^\star - x_0\|_2^2 \le 0.4$, so $x^\star$ solves (1). With $x_0 = (0.3, -0.5)$, $\gamma = 0.35$, $M = 1$, and $\eta_k = (k+10)^{-1}$, Algorithm 1 moves from x_0, where corridor 3 wins, to $x^\star = (-0.2, -0.2)$ (Figure 3). With $\varrho = 0$ the errors are 9.2×10^{-4}, 2.0×10^{-4}, 6.6×10^{-5} at $k = 10^2, 10^3, 3 \times 10^3$; with $\varrho = 0.01$ the iterates stay within 5×10^{-3} of $x^\star$, the smoothing bias. At $x^\star$, $L_1 = L_3 = 3.6 < L_2$.

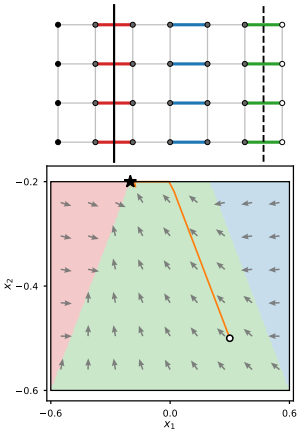


Figure 3: Top: corridors 1–3 (red, blue, green), minimum cut at x_0 (dashed), target cut (solid). Bottom: winning corridor, descent field, iterates from x_0 (○) to $x^\star$ (⋆), $\varrho = 0.01$ orange.

5.2 Electronic design automation: die-boundary calibration

We evaluate Algorithm 3 on electronic design automation (EDA) using the International Symposium on Physical Design 1998 (ISPD98) `ibm01` benchmark (Alpert, 1998), which contains 12,752 cells

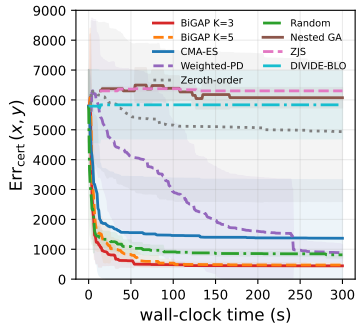


Figure 4: $\mathrm{Err}_{\mathrm{cert}}$ v.s. time. Curves and bands are mean $\pm$ standard deviation over 8 initializations.

Method	Hamming error	Certified error
BiGAP-PD, $K = 3$	$\mathbf{442.6 \pm 3.1}$	$\mathbf{445.0 \pm 8.6}$
BiGAP-PD, $K = 5$	470.6 ± 74.2	472.1 ± 74.5
CMA-ES	1360.6 ± 1956.5	1369.1 ± 1954.0
Weighted-PD	868.4 ± 275.3	874.8 ± 268.2
Zeroth-order	4935.2 ± 2333.9	4935.9 ± 2333.0
Random search	805.1 ± 110.9	813.6 ± 101.3
Nested GA + MILP LL	6073.2 ± 337.4	6073.2 ± 337.4
ZJS binary restriction	6302.0 ± 0.0	6302.0 ± 0.0
DIVIDE-BLO	5835.2 ± 1133.6	5835.2 ± 1133.6

Table 2: Best constrained partition after 300 seconds, reported as mean $\pm$ standard deviation over 8 initializations.

and 14,111 nets. The reference partition y^{ref} is generated by K-SpecPart (Bustany et al., 2024). The goal is to identify boundary parameters x whose optimal constrained graph-cut partition best matches y^{ref}, yielding a natural BLO formulation. Given spectral coordinates $r_i \in \mathbb{R}^2$, the UL variable $x = (n_1, n_2, b) \in [-1.5, 1.5]^2 \times [-0.8, 0.8]$ defines the signed boundary score $s_i(x) = (n_1, n_2)^\top r_i - b$, which parameterizes the unary costs of the LL problem. For each x, the LL variable $y \in \{0, 1\}12752$ is the resulting optimal constrained partition. This yields an instance of (1), where the UL selects x to minimize the discrepancy between y and y^{ref}.

$$f(y) = \sum_i |y_i - y_i^{\mathrm{ref}}|, \quad g(x, y) = \sum_i q_i^{y_i}(x) + \sum_{(i,j)\in E} w_{ij}|y_i - y_j|, \tag{15a}$$

$$c(y) = \sum_{(i,j)\in E_{\geq 5}} \mathbf{1}\{y_i \neq y_j\} - 1300 \leq 0. \tag{15b}$$

Here, $q_i^1(x) = 3\,\mathrm{softplus}(-30 s_i(x))$ and $q_i^0(x) = 3\,\mathrm{softplus}(30 s_i(x))$. The set E denotes the distance-weighted clique expansion over all nets, while $E_{\geq 5}$ is the unweighted clique expansion restricted to nets of degree at least five. Consequently, both the LL energy and constraint are submodular binary pairwise graph-cut functions (Boykov et al., 2001; Kolmogorov & Zabih, 2004), while the UL calibrates x based on the induced optimal partition.

We compare BiGAP-PD with CMA-ES (Hansen & Ostermeier, 2001), Weighted-PD, zeroth-order and random search, nested GA with a MILP LL solver (Sinha et al., 2018), the ZJS binary restriction, and DIVIDE-BLO (Jiang et al., 2025). Experiments were run on a dual-socket Intel Xeon Gold 6448Y server with 64 physical cores. We report certified error $\mathrm{Err}_{\mathrm{cert}}(x, y) := f(y) + [g(x, y) - \widehat{v}(x)]_+ + [c(y)]_+$, where $\widehat{v}(x)$ is a supported-dual lower bound on the constrained LL value. Results are averaged over eight random initializations under a 300-second budget. BiGAP-PD uses $\gamma = 0.01$, $M = 1$, $\eta = 5 \times 10^{-5}$, $\theta_\lambda = \theta_\mu = 5 \times 10^{-4}$, and $\varrho = 0.01$. We write $K := K_\lambda = K_\mu$ in plots and tables. Figure 4 shows that BiGAP-PD achieves the fastest reduction in certified error and best final performance. As summarized in Table 2, a small inner budget, $K_\lambda = K_\mu = 3$, reaches a Hamming error of 442.6 ± 3.1 with only a small certification gap, whereas a larger inner budget slows wall-clock convergence. DIVIDE-BLO makes little progress at partition dimension $d = 12{,}752$, where its discretization is expensive. The nested GA and zeroth-order method (Lin et al., 2022) also incur high search costs over the large discrete LL space, while CMA-ES and Weighted-PD have higher final errors and variability. Appendix F provides additional ablations.

6 Conclusions and Future Work

This paper develops BiGAP for bilevel optimization with discrete submodular lower levels, combining a value-gap penalty with randomized smoothing while preserving the original discrete LL structure. The resulting methods reduce discrete computation to unconstrained SFM and provide convergence guarantees for both unconstrained and coupled-constraint settings, with experiments demonstrating scalability to large discrete problems. Appendix H discusses exploratory extensions to fully discrete BLO, where exact and tractable optimization at both levels remains open. Another promising direction is the development of parallel and GPU-friendly SFM or graph-cut oracles for large-scale accelerator-based computation.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs for all theoretical claims are provided in the appendix. Algorithms 1, 2, and 3 specify the oracle calls and update rules. Section 5.2 and Appendix F document the data construction, hyperparameters, computing environment, evaluation protocol, and ablations. Experimental code is provided in the supplementary material.

AI USE STATEMENT

Generative AI tools were used to assist with theoretical checking, proof and methodology feedback, scientific figure and table preparation, literature search, paper formatting, and language and presentation editing. The authors reviewed and validated all AI-assisted mathematical arguments, citations, code, experimental results, and prose. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Eitaro Aiyoshi and Kiyotaka Shimizu. A solution method for the static constrained Stackelberg problem via penalty method. *IEEE Transactions on Automatic Control*, 29(12):1111–1114, 1984.

Charles J. Alpert. The ISPD98 circuit benchmark suite. In *Proceedings of the 1998 International Symposium on Physical Design*, pp. 80–85, 1998.

G. Anandalingam and D. J. White. A solution method for the linear static Stackelberg problem using penalty functions. *IEEE Transactions on Automatic Control*, 35(10):1170–1173, 1990.

Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2–3):145–373, 2013.

Kuang Bai and Jane J. Ye. Directional necessary optimality conditions for bilevel programs. *Mathematics of Operations Research*, 47(2):1169–1191, 2022.

Kuang Bai, Jane J. Ye, and Shangzhi Zeng. Optimality conditions for bilevel programmes via Moreau envelope reformulation. *Optimization*, 74(12):2685–2719, 2025.

Jonathan F. Bard and James T. Moore. An algorithm for the discrete bilevel programming problem. *Naval Research Logistics*, 39(3):419–435, 1992.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9508–9519, 2020.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3rd edition, 2016.

J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, New York, 2000.

Stephenwco Boyd. Subgradient methods. Lecture notes for EE364b, Stanford University, 2014.

Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 1, pp. 105–112, 2001.

Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

Ismail Bustany, Andrew B. Kahng, Ioannis Koutis, Bodhisatta Pramanik, and Zhiang Wang. K-SpecPart: Supervised embedding algorithms and cut overlay for improved hypergraph partitioning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(4): 1232–1245, 2024.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2466–2488. PMLR, 2022.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley-Interscience, New York, 1983.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

Stephan Dempe. *Foundations of Bilevel Programming*, volume 61 of *Nonconvex Optimization and Its Applications*. Kluwer Academic Publishers, Dordrecht, 2002.

Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer, New York, 2nd edition, 2014.

Charles M Fiduccia and Robert M Mattheyses. A linear-time heuristic for improving network partitions. In *Proceedings of the 19th Design Automation Conference*, pp. 175–181. IEEE, 1982.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1165–1173. PMLR, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 2018.

Satoru Fujishige. *Submodular Functions and Optimization*, volume 58 of *Annals of Discrete Mathematics*. Elsevier, 2nd edition, 2005.

Michael R. Garey, David S. Johnson, and Larry Stockmeyer. Some simplified NP-complete graph problems. *Theoretical Computer Science*, 1(3):237–267, 1976.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3748–3758. PMLR, 2020.

Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Yo Ishizuka and Eitaro Aiyoshi. Double penalty method for bilevel optimization problems. *Annals of Operations Research*, 34(1):73–88, 1992.

Satoru Iwata. A fully combinatorial algorithm for submodular function minimization. *Journal of Combinatorial Theory, Series B*, 84(2):203–212, 2002.

Satoru Iwata. Submodular function minimization. *Mathematical Programming*, 112(1):45–64, 2008.

Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

Rishabh Iyer, Stefanie Jegelka, and Jeff Bilmes. Fast semidifferential-based submodular function optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 855–863. PMLR, 2013.

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

Liuyuan Jiang, Quan Xiao, Victor M. Tenorio, Fernando Real-Rojas, Antonio G. Marques, and Tianyi Chen. A primal-dual-assisted penalty approach to bilevel optimization with coupled constraints. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Xiaotian Jiang, Ioannis Tsaknakis, Prashant Khanduri, and Mingyi Hong. A discretization approach for bilevel optimization with low-dimensional and non-convex lower-level. *arXiv preprint arXiv:2505.10830*, 2025.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 137–146, 2003.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30271–30283, 2021.

Thomas Kleinert, Martine Labbé, Ivana Ljubić, and Martin Schmidt. A survey on mixed-integer programming techniques in bilevel optimization. *EURO Journal on Computational Optimization*, 9:100007, 2021.

Pushmeet Kohli and Philip H. S. Torr. Dynamic graph cuts for efficient inference in Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2079–2088, 2007.

Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser, 2nd edition, 2002.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D. Nowak. A fully first-order method for stochastic bilevel optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18083–18113. PMLR, 2023.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D. Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

Tianyi Lin, Zeyu Zheng, and Michael I. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26160–26175, 2022.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. BOME! Bilevel optimization made easy: A simple first-order approach. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17248–17262, 2022.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6305–6315. PMLR, 2020.

Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6882–6892. PMLR, 2021a.

Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8662–8675, 2021b.

Risheng Liu, Xuan Liu, Shangzhi Zeng, Jin Zhang, and Yixuan Zhang. Value-function-based sequential minimization for bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15930–15948, 2023a.

Risheng Liu, Yaohua Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21839–21866. PMLR, 2023b.

Risheng Liu, Zhu Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Moreau envelope for nonconvex bi-level optimization: A single-loop and Hessian-free solution strategy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31566–31596. PMLR, 2024.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552. PMLR, 2020.

Leonardo Lozano, David Bergman, and Andre A. Cire. Network relaxations for combinatorial bilevel optimization under linear interactions. *Operations Research*, 2026. doi: 10.1287/opre.2025.2457. Articles in Advance.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2113–2122. PMLR, 2015.

Akshay Mehra and Jihun Hamm. Penalty method for inversion-free deep bilevel optimization. In *Proceedings of the 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pp. 347–362. PMLR, 2021.

James T. Moore and Jonathan F. Bard. The mixed integer linear bilevel programming problem. *Operations Research*, 38(5):911–921, 1990.

Martin Nägele, Benny Sudakov, and Rico Zenklusen. Submodular minimization under congruency constraints. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 849–866. SIAM, 2018.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

James B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 737–746. PMLR, 2016.

Stephen M. Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980.

Stephen M. Robinson. Some continuity properties of polyhedral multifunctions. In *Mathematical Programming at Oberwolfach*, volume 14 of *Mathematical Programming Studies*, pp. 206–214. North-Holland, 1981.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, Berlin, 1998.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

Alexander Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1723–1732. PMLR, 2019.

Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. *Mathematical Programming*, 214:539–589, 2025.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22 (2):276–295, 2018.

Daouda Sow, Kaiyi Ji, Ziwei Guan, and Yingbin Liang. A primal-dual approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

Zoya Svitkina and Lisa Fleischer. Submodular approximation: Sampling-based algorithms and lower bounds. *SIAM Journal on Computing*, 40(6):1715–1737, 2011.

Sahar Tahernejad, Ted K. Ralphs, and Scott T. DeNegre. A branch-and-cut algorithm for mixed integer bilevel linear optimization problems and its implementation. *Mathematical Programming Computation*, 12(4):529–568, 2020.

Luís N. Vicente, Gilles Savard, and Joaquim J. Júdice. Discrete linear bilevel programming problem. *Journal of Optimization Theory and Applications*, 89(3):597–614, 1996.

Marin Vlastelica, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolínek. Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*, 2020.

D. J. White and G. Anandalingam. A penalty function approach for solving bi-level linear programs. *Journal of Global Optimization*, 3(4):397–419, 1993.

Jane J. Ye and Daoli Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33(1):9–27, 1995.

Jane J. Ye and Daoli Zhu. New necessary optimality conditions for bilevel programs by combining the MPEC and value function approaches. *SIAM Journal on Optimization*, 20(4):1885–1905, 2010.

Jiajin Yu and Shabbir Ahmed. Polyhedral results for a class of cardinality constrained submodular minimization problems. *Discrete Optimization*, 24:87–102, 2017.

SUPPLEMENTARY MATERIAL: THEORY, APPLICATIONS, AND EXTENSIONS

A Extended Related Work **15**

B Proofs for the Penalty and Unconstrained Method **17**

B.1 Proof of Proposition 1 17

B.2 Proof of Corollary 1 17

B.3 Penalty Geometry in a Finite Lower Problem 17

B.4 Additional Technical Discussion 18

B.5 Proof of Proposition 2 19

B.6 Proof of Global Recovery 19

B.7 Proof of Proposition 3 20

B.8 Proof of Theorem 1 20

B.9 Proof of Corollary 2 21

C Detailed Application Checks **21**

C.1 Scope and Two Reusable Checks 22

C.2 Pairwise Graph-Cut Energies 22

C.3 Concave Costs of Aggregate Load 22

D Verification of Exact Lagrangian Support and Bounded Multipliers **23**

D.1 Exact-Support Characterization and Structured Sufficient Conditions 23

D.2 Uniform Multiplier Bounds 24

E Proofs for the Constrained Method **25**

E.1 Proof of Proposition 4 26

E.2 Supported-Basis Conditioning and First-Order Recovery 27

E.3 Proof of Proposition 5 35

E.4 Proofs of Theorem 2 and Corollary 3 37

F Constrained EDA Details and Ablations **39**

G Toy Instance: Assumptions, Reference Solution, and the Systolic Illustration **40**

G.1 Assumptions and Oracle Calls 40

G.2 Exact Reference Solution 41

G.3 The Inverse Systolic Example 41

H Exploratory Fully Discrete Extensions **42**

H.1 Measure Lifting and the Tie-Breaking Knob 42

H.2 Entropy-Smoothed Penalty 42

H.3 Moment Lifting 43

A EXTENDED RELATED WORK

Bilevel optimization. Bilevel optimization has a long history in hierarchical decision-making and mathematical programming (Dempe, 2002; Colson et al., 2007; Sinha et al., 2018). Modern learning-oriented methods predominantly differentiate through a continuous LL problem, either by unrolling its optimization trajectory (Maclaurin et al., 2015; Franceschi et al., 2017; 2018; Shaban et al., 2019; Grazzi et al., 2020) or by implicit differentiation and approximations of the inverse LL Hessian (Pe-

dregosa, 2016; Ghadimi & Wang, 2018; Lorraine et al., 2020; Ji et al., 2021). Subsequent stochastic and single-loop methods improve oracle and sample complexity while retaining this hypergradient viewpoint (Khanduri et al., 2021; Chen et al., 2022; Hong et al., 2023). Beyond the usual strongly convex LL setting, first-order methods have also been developed for multiple LL solutions and nonconvex followers (Liu et al., 2020; 2021b; Sow et al., 2022; Liu et al., 2023b). Discrete and mixed-integer BLO instead relies on enumeration, reformulation, decomposition, branch-and-bound, or branch-and-cut techniques (Moore & Bard, 1990; Bard & Moore, 1992; Vicente et al., 1996; Tahernejad et al., 2020; Kleinert et al., 2021; Lozano et al., 2026; Jiang et al., 2025). These methods cover substantially broader combinatorial followers than considered here, but generally do not exploit exact polynomial-time submodular oracles.

Value-function and penalty methods. A second major line replaces LL optimality by a single-level value-function or penalty formulation. Early Stackelberg and bilevel methods already used penalty functions to eliminate the nested problem (Aiyoshi & Shimizu, 1984; Anandalingam & White, 1990; Ishizuka & Aiyoshi, 1992; White & Anandalingam, 1993). The lower-level value-function reformulation later provided an equivalent constraint and connected its degeneracy to calmness and partial exact penalization (Ye & Zhu, 1995; 2010; Bai & Ye, 2022). This viewpoint has recently become central to scalable first-order BLO. BVFIM penalizes a regularized LL value function through an interior-point construction (Liu et al., 2021a), with related approximation schemes addressing nonconvex followers (Liu et al., 2021b) and sequential value-function minimization developed in BVFSM (Liu et al., 2023a). Mehra and Hamm use a classical penalty formulation to avoid Hessian inversion (Mehra & Hamm, 2021), while BOME applies a dynamic-barrier method to the value-function constraint (Liu et al., 2022). F2SA instead uses an increasing multiplier on the value-function gap to obtain fully first-order stochastic algorithms (Kwon et al., 2023), and first-order penalty methods have also been analyzed through structured minimax reformulations for convex LL problems (Lu & Mei, 2024). Related multiplier and primal–dual approaches include constrained formulations for multiple LL minimizers and averaged multiplier methods without LL strong convexity (Sow et al., 2022; Liu et al., 2023b). More recently, penalty formulations have been extended to bilevel reinforcement learning and RLHF (Shen et al., 2025), while Moreau-envelope reformulations provide another way to regularize LL value functions (Liu et al., 2024; Bai et al., 2025). BLOCC uses convex duality to handle coupled constraints while retaining a value-function formulation (Jiang et al., 2024). BiGAP belongs to this value-function family but differs in that the LL variable is discrete: it penalizes the exact discrete value gap and smooths the resulting scalar function, rather than differentiating or relaxing the discrete solution mapping.

Submodular optimization. Submodularity is a fundamental tractable structure in discrete optimization (Fujishige, 2005; Iwata, 2008; Bach, 2013). Unconstrained submodular function minimization (SFM) admits polynomial-time algorithms through polyhedral and combinatorial methods (Grötschel et al., 1981; Schrijver, 2000; Iwata et al., 2001; Iwata, 2002; Orlin, 2009). For binary pairwise submodular energies, SFM specializes to minimum-cut problems, with graph-representability characterized for broad classes of binary energies (Boykov et al., 2001; Kolmogorov & Zabih, 2004); dynamic graph-cut methods can further reuse residual states across related instances while retaining exact solutions (Kohli & Torr, 2007). In contrast, adding side constraints generally destroys the tractability of unconstrained SFM: cardinality, balance, and other constrained variants can become NP-hard and require approximation or specialized algorithms (Svitkina & Fleischer, 2011; Yu & Ahmed, 2017; Iyer et al., 2013), although exact algorithms remain available for particular constraint families such as congruency constraints (Nägele et al., 2018). This distinction is central to BiGAP. For $c \equiv 0$, each required discrete minimization is an exact unconstrained SFM problem. With coupled constraints, BiGAP does not assume direct tractability of constrained SFM; instead, under Assumption 3, each fixed nonnegative Lagrangian multiplier preserves submodularity and reduces the computation back to unconstrained SFM. Appendix D characterizes when this supported representation is exact.

B PROOFS FOR THE PENALTY AND UNCONSTRAINED METHOD

B.1 PROOF OF PROPOSITION 1

We prove the result directly for the general constrained penalty (2). The statement in Section 2.1 follows by setting $c \equiv 0$.

Proof. Fix $x \in \mathcal{X}$. Since $\mathcal{Y}$ is finite, the minimum defining $\delta(x)$ over feasible labels outside $\mathcal{Y}^\star(x)$ is either over finitely many strictly positive gaps or over the empty set. Hence $\delta(x) \in (0, +\infty]$. Compactness of $\mathcal{X}$, continuity of $f(\cdot, y)$, and finiteness of $\mathcal{Y}$ imply $\Delta_f < \infty$. Since $g(x, y) - v(x) = 0$ for every $y \in \mathcal{Y}^\star(x)$, splitting the feasible minimum in (2) over $\mathcal{Y}^\star(x)$ and the remaining feasible labels gives

$$p_\gamma(x) = \min \left\{ \underbrace{\min_{y \in \mathcal{Y}^\star(x)} f(x, y)}_{\phi(x)}, \min_{y \in \mathcal{Y}(x) \setminus \mathcal{Y}^\star(x)} \big[f(x, y) + \gamma\big(g(x, y) - v(x)\big) \big] \right\}. \tag{16}$$

By (16), the branch corresponding to $\mathcal{Y}^\star(x)$ equals $\phi(x)$, whereas every other feasible branch is affine and nondecreasing in γ. Hence $p_\gamma(x) \le \phi(x)$ and $\gamma \mapsto p_\gamma(x)$ is nondecreasing.

Moreover, for every feasible $y \notin \mathcal{Y}^\star(x)$,

$$\begin{aligned} f(x, y) + \gamma\big(g(x, y) - v(x)\big) &\ge \min_{\bar{y} \in \mathcal{Y}} f(x, \bar{y}) + \gamma\delta(x) \\ &\ge \min_{\bar{y} \in \mathcal{Y}} f(x, \bar{y}) + \Delta_f \\ &\ge \max_{\bar{y} \in \mathcal{Y}} f(x, \bar{y}) \\ &\ge \phi(x) \end{aligned}$$

whenever $\gamma\delta(x) \ge \Delta_f$. Thus no feasible LL-suboptimal branch can decrease the minimum in (16) below $\phi(x)$, and therefore $p_\gamma(x) = \phi(x)$. Finally, since $\delta(x) > 0$, the condition $\gamma\delta(x) \ge \Delta_f$ holds for all sufficiently large γ. Together with monotonicity, this yields $p_\gamma(x) \uparrow \phi(x)$ as $\gamma \to \infty$. □

B.2 PROOF OF COROLLARY 1

Proof. Let $m := \inf_{z \in K} \phi(z)$. Proposition 1(i) gives $p_\gamma(x) \le \phi(x)$ for every $x \in K$. For any $y \in \mathcal{Y}(x) \setminus \mathcal{Y}^\star(x)$, the choice of γ gives

$$f(x, y) + \gamma\big(g(x, y) - v(x)\big) > f(x, y) + [m - f(x, y)]_+ \ge m.$$

For $y \in \mathcal{Y}^\star(x)$, the penalized branch equals $f(x, y) \ge \phi(x) \ge m$. Hence $m \le p_\gamma(x) \le \phi(x)$ on K. Moreover, $p_\gamma(x) = m$ if and only if $\phi(x) = m$, since every LL-suboptimal branch is strictly larger than m. Thus p_γ and ϕ have the same minimizers over K, proving the local optimality statement.

For the stationarity statement, let $\delta_r := \inf_{x \in B(\bar{x}, r)} \delta(x) > 0$. Since $\gamma\delta_r > \Delta_f$, Proposition 1(ii) gives $p_\gamma = \phi$ on $B(\bar{x}, r)$. Their local Lipschitzness therefore gives $\partial_C p_\gamma(z) = \partial_C \phi(z)$ for every $z \in B(\bar{x}, r)$. In particular, $\partial_C p_\gamma(\bar{x}) = \partial_C \phi(\bar{x})$, proving Clarke-stationarity equivalence. For every $\varrho < r$, taking the convex hull over $B(\bar{x}, \varrho)$ gives

$$\partial_\varrho p_\gamma(\bar{x}) = \operatorname{conv} \bigcup_{z \in B(\bar{x}, \varrho)} \partial_C p_\gamma(z) = \operatorname{conv} \bigcup_{z \in B(\bar{x}, \varrho)} \partial_C \phi(z) = \partial_\varrho \phi(\bar{x}),$$

which proves radius-ϱ Goldstein-stationarity equivalence. □

B.3 PENALTY GEOMETRY IN A FINITE LOWER PROBLEM

Finiteness ensures that the feasible-label gap $\delta(x)$ in Proposition 1 is positive pointwise, with $\delta(x) = +\infty$ when every feasible label is lower optimal. Uniform exactness on a region K follows from $\delta_K := \inf_{x \in K} \delta(x) > 0$ by Corollary 1.

Proposition 6 (Uniform gap in the unconstrained-lower specialization)**.** *Under Assumption 1, suppose $c \equiv 0$ and $\mathcal{X}$ is connected. Then $\inf_{x\in\mathcal{X}} \delta(x) > 0$ if and only if $\mathcal{Y}^\star(x)$ is constant on $\mathcal{X}$.*

Proof. If $\mathcal{Y}^\star(x)$ is constant, $\delta(x)$ is either identically $+\infty$ or the minimum of finitely many continuous, strictly positive functions. Compactness gives a positive uniform gap.

Conversely, if $\mathcal{Y}^\star(x)$ changes, some label y is optimal at one point and nonoptimal at another. The continuous function $g(x, y) - v(x)$ takes both zero and positive values, hence arbitrarily small positive values by connectedness. At these points, $\delta(x) \le g(x, y) - v(x)$, so $\inf_{x\in\mathcal{X}} \delta(x) = 0$. □

Persistent ties are therefore compatible with a uniform gap. With coupled constraints, feasibility changes can alter the solution collection without shrinking the objective gap, so the uniform feasible-label gap must be checked directly.

Example 1 (One-sided penalty geometry and smoothing)**.** Take $c \equiv 0$, $\mathcal{X} = [-1, 1]$, $\mathcal{Y} = \{0, 1\}$, $f(x, y) = 1 - y$, and $g(x, y) = (1 - 2y)x$. Both objectives are modular in y. The lower value is $v(x) = -|x|$, and optimistic selection gives $\phi(x) = 1$ for $x < 0$ and $\phi(x) = 0$ otherwise. For every $\gamma > 0$,

$$p_\gamma(x) = \begin{cases} 1, & x \le -\frac{1}{2\gamma}, \\ -2\gamma x, & -\frac{1}{2\gamma} < x < 0, \\ 0, & x \ge 0. \end{cases} \tag{17}$$

Thus p_γ replaces the jump by a left-sided ramp. Exactness holds for sufficiently large γ at each fixed x, but no finite threshold works throughout a neighborhood of $x = 0$.

For $\gamma = 3$, the ramp endpoint $x^0 = -1/6$ exposes an oracle degeneracy. The lower SFM uniquely returns $y_2 = 0$, while the penalized SFM minimizes $f(x^0, y) + 3g(x^0, y)$ and ties between both labels. Deterministically choosing $y_1 = 0$ yields

$$G = \nabla_x f(x^0, y_1) + 3\nabla_x g(x^0, y_1) - 3\nabla_x g(x^0, y_2) = 0,$$

so every iterate remains at x^0.

Ball smoothing instead gives

$$p_3^\varrho(x) = \frac{1}{2\varrho} \int_{x-\varrho}^{x+\varrho} p_3(z)\, dz,$$
$$\nabla p_3^\varrho(-1/6) = \frac{p_3(-1/6 + \varrho) - p_3(-1/6 - \varrho)}{2\varrho} = -3, \quad 0 < \varrho \le \frac{1}{6}.$$

Uniformly perturbed queries avoid the ties almost surely and yield $\mathbb{E}[G] = -3$ for both $\varrho = 0.1$ and $\varrho = 0.05$. Individual queries may return zero, but the initial point is no longer absorbing.

Figure 2 compares p_3, p_5, p_{10} with $p_3^{0.1}$ and $p_3^{0.05}$. Increasing γ narrows the ramp, while decreasing ϱ localizes the smoothing. For $\varrho < 1/12$, smoothing preserves the central ramp on $[-1/6 + \varrho, -\varrho]$.

More generally, consider an unconstrained two-label switch from y_- to y_+ at x_0, with $f(x_0, y_+) < f(x_0, y_-)$. Continuity preserves this upper-cost ordering locally. On the right, y_+ minimizes both objectives, so $p_\gamma = \phi$. On the left, $\delta(x) = g(x, y_+) - g(x, y_-) > 0$, and

$$\phi(x) - p_\gamma(x) = \max\{f(x, y_-) - f(x, y_+) - \gamma\delta(x), 0\}.$$

The discrepancy is therefore confined to the side where the lower optimal label has the larger upper cost.

B.4 Additional technical discussion

This appendix explains why continuous implicit systems do not extend to a binary response and why arbitrary SFM tie-breaking is harmless at differentiability points.

In the smooth strongly convex unconstrained setting, the usual hypergradient is

$$0 = \nabla_y g\big(x, y^\star(x)\big), \tag{18a}$$

$$D_x y^\star(x) = -\big(\nabla^2_{yy} g\big(x, y^\star(x)\big)\big)^{-1} \nabla^2_{yx} g\big(x, y^\star(x)\big), \tag{18b}$$

$$\nabla \phi(x) = \nabla_x f\big(x, y^\star(x)\big) + \nabla_y f\big(x, y^\star(x)\big) D_x y^\star(x). \tag{18c}$$

For a smooth continuous lower problem with active constraint set $\mathcal{A}$, define locally $\mathcal{L}(x, y, \nu) := g(x, y) + \nu^\top c(x, y)$. If the active set is stable and the KKT system is strongly regular, differentiating stationarity and the active constraints gives

$$\begin{bmatrix} \nabla^2_{yy} \mathcal{L} & \nabla_y c_{\mathcal{A}}^\top \\ \nabla_y c_{\mathcal{A}} & 0 \end{bmatrix} \begin{bmatrix} D_x y^\star \\ D_x \nu^\star \end{bmatrix} = - \begin{bmatrix} \nabla^2_{yx} \mathcal{L} \\ \nabla_x c_{\mathcal{A}} \end{bmatrix}, \tag{19}$$

where all derivatives are evaluated at the local primal–dual solution. The response derivative is obtained by solving this linear system and then inserted into the chain rule in (18c).

The binary lower problem has no counterpart of (19). There is no infinitesimal feasible motion in the y-coordinates, no differentiable lower stationarity equation in y, and no KKT Jacobian whose inverse tracks changes between discrete labels. Indeed, a locally stable label has derivative zero, which misses the entire effect of a later branch switch. At the switch, the solution map is set-valued. An optimistic tie creates a second mismatch because the relevant branch is chosen by f, not necessarily by the tie-breaking rule of the lower solver. Continuous, multilinear, or entropy-regularized relaxations restore differentiability only after changing the lower problem. Their gradients therefore optimize the relaxed value unless one separately proves integrality and preservation of the optimistic selection. The fully discrete examples in Appendix H demonstrate why those conclusions cannot be assumed. The finite-minimum and SFM calculations used here replace the unavailable implicit-function mechanism while retaining the original binary feasible set.

The same finite-minimum viewpoint resolves tie-breaking. The calculus used in the unconstrained oracle does not assert that discrete tie sets have measure zero. Under C^1 branch regularity, a tie set may have positive measure. Rademacher's theorem instead makes the nondifferentiability sets of the finite-minimum values null. At a point where a value is differentiable, Lemma 1 forces all active branch gradients to agree, so arbitrary SFM tie-breaking is harmless.

B.5 Proof of Proposition 2

Proof. Fix x. Since $g(x, \cdot)$ is submodular, $-\Delta_{ij} g(x, y) \geq 0$. The inequality $\Delta_{ij}(f + \beta g)(x, y) \leq 0$ is therefore equivalent, direction by direction, to

$$\beta \geq \frac{\Delta_{ij} f(x, y)}{-\Delta_{ij} g(x, y)} \quad \text{when } \Delta_{ij} g(x, y) < 0,$$

and to $\Delta_{ij} f(x, y) \leq 0$ when $\Delta_{ij} g(x, y) = 0$. Taking the largest required ratio proves the first claim. Taking the supremum over x gives the uniform statement. Finally, if $c(x, \cdot)$ is componentwise submodular and $\lambda \geq 0$, then $\lambda^\top c(x, \cdot)$ is submodular. Adding it to an already submodular objective therefore preserves submodularity. □

B.6 Proof of global recovery

Proposition 7 (Global recovery of penalized minimizers)**.** *Suppose Assumption 1 holds with $c \equiv 0$. Let $\phi^\star := \min_{x \in \mathcal{X}} \phi(x)$ and $\mathcal{X}^\star := \arg\min_{x \in \mathcal{X}} \phi(x)$. Then $\mathcal{X}^\star$ is nonempty and*

$$\min_{x \in \mathcal{X}} p_\gamma(x) \uparrow \phi^\star \quad \text{as } \gamma \to \infty.$$

Moreover, for every sequence $\gamma_j \to \infty$ and every choice $x^j \in \arg\min_{x \in \mathcal{X}} p_{\gamma_j}(x)$,

$$\operatorname{dist}(x^j, \mathcal{X}^\star) \to 0.$$

Equivalently, every cluster point of $\{x^j\}$ belongs to $\mathcal{X}^\star$.

Proof. Fix $\gamma > 0$. Since $p_\gamma = \gamma(v_\gamma - v)$, it is the scaled difference of two finite minima of continuous branches and is therefore continuous on $\mathcal{X}$. Proposition 1 gives $\phi = \sup_{\gamma>0} p_\gamma$, so ϕ is lower semicontinuous. Compactness of $\mathcal{X}$ then implies that ϕ attains its infimum and $\mathcal{X}^\star$ is nonempty.

Let $m_\gamma = \min_{x\in\mathcal{X}} p_\gamma(x)$. These values increase and are bounded above by $\phi^\star$. Let m be their limit, and let $x_{\gamma_j} \to \bar{x}$ be a convergent subsequence of global minimizers. For every fixed $\gamma_0 > 0$, monotonicity and continuity give

$$p_{\gamma_0}(\bar{x}) = \lim_j p_{\gamma_0}(x_{\gamma_j}) \le \lim_j m_{\gamma_j} = m.$$

Taking the supremum over γ_0 yields $\phi(\bar{x}) \le m \le \phi^\star$. Since $\phi^\star \le \phi(\bar{x})$ by definition, all three quantities are equal, so $\bar{x} \in \mathcal{X}^\star$ and $m = \phi^\star$.

For the distance statement, suppose there were an $\varepsilon_0 > 0$, a sequence $\gamma_j \to \infty$, and global minimizers x_{γ_j} with $\operatorname{dist}(x_{\gamma_j}, \mathcal{X}^\star) > \varepsilon_0$. Compactness would produce a cluster point, and the preceding argument would place that point in $\mathcal{X}^\star$, a contradiction. □

B.7 Proof of Proposition 3

We first establish an auxiliary finite-minimum calculus lemma used in the proof of Proposition 3.

Lemma 1 (Finite-minimum calculus)**.** *Let $h(x) = \min_{i\in[q]} H_i(x)$, where every H_i is continuously differentiable near x. Then h is locally Lipschitz and*

$$\partial_C h(x) = \operatorname{conv}\{\nabla H_i(x) : i \in I(x)\}, \quad I(x) := \arg\min_i H_i(x).$$

Moreover, h is differentiable at x if and only if all active gradients coincide. In that case every active branch has gradient $\nabla h(x)$.

Proof. The identity follows by applying the standard Clarke formula for a finite maximum to $-h = \max_i(-H_i)$ (Clarke, 1983). The directional derivative is $h'(x,d) = \min_{i\in I(x)} \nabla H_i(x)^\top d$. This finite minimum of linear forms is linear in d exactly when all active coefficient vectors coincide, which proves the differentiability claim. □

Using this, we can proceed by applying the lemma.

Proof of Proposition 3. Lemma 1 applies to both v_γ and v. The branch-gradient bounds in (6) give $\operatorname{Lip}(v_\gamma) \le L_f/\gamma + L_g$ and $\operatorname{Lip}(v) \le L_g$, so $p_\gamma = \gamma(v_\gamma - v)$ is L_γ-Lipschitz with $L_\gamma \le L_f + 2\gamma L_g$. Rademacher's theorem then implies that the nondifferentiability sets of v_γ and v are Lebesgue-null (Clarke, 1983). At any point outside their union, Lemma 1 implies that all active branch gradients within each minimum coincide, so differentiating $p_\gamma = \gamma(v_\gamma - v)$ gives (7).

Finally, the Clarke sum rule (Clarke, 1983) gives $\partial_C p_\gamma(x) \subseteq \gamma\partial_C v_\gamma(x) - \gamma\partial_C v(x)$. For any $y_1 \in \mathcal{Y}^\star_\gamma(x)$ and $y_2 \in \mathcal{Y}^\star(x)$, Lemma 1 gives $\gamma^{-1}\nabla_x f(x,y_1) + \nabla_x g(x,y_1) \in \partial_C v_\gamma(x)$ and $\nabla_x g(x,y_2) \in \partial_C v(x)$, which proves that $u(x,y_1,y_2)$ belongs to the right-hand side. □

B.8 Proof of Theorem 1

Proof. **Randomized oracle.** Conditionally on x^t, each $w^{t,j}$ has a density on $\mathbb{B}_2^p$, so $z^{t,j} = x^t + \varrho w^{t,j}$ has a density on $\mathbb{R}^p$. Proposition 3(i) makes the nondifferentiability sets of v and v_γ Lebesgue-null. Hence the nondifferentiability set of $p_\gamma = \gamma(v_\gamma - v)$ is Lebesgue-null, and each $z^{t,j}$ is almost surely a differentiability point of p_γ. Proposition 3(ii) then gives $G^{t,j} = \nabla p_\gamma(z^{t,j})$ almost surely. For $\varrho \in (0, \varrho_0)$, the definitions of ϱ_0 and $\mathcal{X}^\varrho$ confine every query to $\mathcal{X}^\varrho \subset U$, where Proposition 3(i) makes p_γ L_γ-Lipschitz, so $\|G^{t,j}\|_2 \le L_\gamma$ almost surely. Because p_γ is Lipschitz on $\mathcal{X}^\varrho$, the uniform smoothing (3) is differentiable with $\nabla p^\varrho_\gamma(x) = \mathbb{E}_w(\nabla p_\gamma(x + \varrho w))$, and $\nabla p^\varrho_\gamma(x) \in \partial_\varrho p_\gamma(x)$ (Lin et al., 2022). Conditional independent sampling therefore gives the first identity in (9). Independence together with $\mathbb{E}(\|G^{t,j}\|_2^2 \mid x^t) \le L_\gamma^2$ gives

$$\mathbb{E}\big(\|\bar{G}^t - \nabla p^\varrho_\gamma(x^t)\|_2^2 \mid x^t\big) = \frac{1}{M}\,\mathbb{E}\big(\|G^{t,1} - \nabla p^\varrho_\gamma(x^t)\|_2^2 \mid x^t\big) \le \frac{L_\gamma^2}{M}.$$

This proves the second inequality in (9).

Convergence and complexity. By Lin et al. (2022), there is a universal constant $c_0 > 0$ such that ∇p_γ^ϱ is H-Lipschitz on $\mathcal{X}$ with $H \le c_0 L_\gamma \sqrt{p}/\varrho$. Since p_γ is L_γ-Lipschitz on $\mathcal{X}^\varrho$, Jensen's inequality gives $p_\gamma^\varrho(x^0) \le p_\gamma(x^0) + \varrho L_\gamma$, and averaging preserves the lower bound $\inf_{x\in\mathcal{X}} p_\gamma^\varrho(x) \ge \inf_{x\in\mathcal{X}^\varrho} p_\gamma(x)$. Hence $p_\gamma^\varrho(x^0) - \inf_{x\in\mathcal{X}} p_\gamma^\varrho(x) \le \Delta_{\gamma,0} + \varrho L_\gamma$. Let $\xi^t := \bar{G}^t - \nabla p_\gamma^\varrho(x^t)$ and $\widehat{\mathcal{G}}^t := \eta^{-1}(x^t - x^{t+1})$. Then $\mathbb{E}(\xi^t \mid x^t) = 0$, $\mathbb{E}(\|\xi^t\|_2^2 \mid x^t) \le L_\gamma^2/M$, and nonexpansiveness of $\Pi_\mathcal{X}$ gives $\|\widehat{\mathcal{G}}^t - \mathcal{G}_{\eta,\varrho}(x^t)\|_2 \le \|\xi^t\|_2$. Projection optimality yields $\langle \bar{G}^t, x^{t+1} - x^t\rangle \le -\eta^{-1}\|x^{t+1} - x^t\|_2^2$. The H-descent lemma and $\eta \le 1/(2H)$ then give

$$\begin{aligned} p_\gamma^\varrho(x^{t+1}) &\le p_\gamma^\varrho(x^t) + \langle \nabla p_\gamma^\varrho(x^t), x^{t+1} - x^t\rangle + \tfrac{H}{2}\|x^{t+1} - x^t\|_2^2 \\ &\le p_\gamma^\varrho(x^t) - \eta\big(1 - \tfrac{H\eta}{2}\big)\|\widehat{\mathcal{G}}^t\|_2^2 + \eta\langle \xi^t, \widehat{\mathcal{G}}^t\rangle \\ &\le p_\gamma^\varrho(x^t) - \tfrac{3\eta}{4}\|\widehat{\mathcal{G}}^t\|_2^2 + \eta\langle \xi^t, \widehat{\mathcal{G}}^t\rangle. \end{aligned}$$

Young's inequality gives

$$p_\gamma^\varrho(x^{t+1}) \le p_\gamma^\varrho(x^t) - \tfrac{\eta}{2}\|\widehat{\mathcal{G}}^t\|_2^2 + \eta\|\xi^t\|_2^2.$$

Using $\|\mathcal{G}_{\eta,\varrho}(x^t)\|_2^2 \le 2\|\widehat{\mathcal{G}}^t\|_2^2 + 2\|\xi^t\|_2^2$ then yields

$$p_\gamma^\varrho(x^{t+1}) \le p_\gamma^\varrho(x^t) - \tfrac{\eta}{4}\|\mathcal{G}_{\eta,\varrho}(x^t)\|_2^2 + \tfrac{3\eta}{2}\|\xi^t\|_2^2.$$

Taking conditional expectations, telescoping, and using $\eta = 1/(2H)$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big(\|\mathcal{G}_{\eta,\varrho}(x^t)\|_2^2\big) \le \frac{8H(\Delta_{\gamma,0} + \varrho L_\gamma)}{T} + \frac{6L_\gamma^2}{M}. \tag{20}$$

Choosing $M = \lceil 12 L_\gamma^2 \varepsilon^{-2}\rceil$ and $T \ge \lceil 16H(\Delta_{\gamma,0} + \varrho L_\gamma)\varepsilon^{-2}\rceil$ bounds the right-hand side by ε^2. Substituting $H \le c_0 L_\gamma\sqrt{p}/\varrho$ gives (10). Each outer iteration uses exactly $2M$ SFM calls, and the final recovery of y_F uses one additional SFM call, proving the total SFM complexity stated in Theorem 1. □

B.9 Proof of Corollary 2

Proof. If $B(x^{t_{\mathrm{out}}}, \varrho) \subset \operatorname{int}\mathcal{X}$ and $\eta L_\gamma \le \varrho$, then $\|\eta \nabla p_\gamma^\varrho(x^{t_{\mathrm{out}}})\|_2 \le \eta L_\gamma \le \varrho$, so the deterministic smoothed step lies in $\mathcal{X}$ and projection is inactive. Therefore $\mathcal{G}_{\eta,\varrho}(x^{t_{\mathrm{out}}}) = \nabla p_\gamma^\varrho(x^{t_{\mathrm{out}}})$. The smoothing identity and the definition of the Goldstein subdifferential give $\nabla p_\gamma^\varrho(x^{t_{\mathrm{out}}}) \in \partial_\varrho p_\gamma(x^{t_{\mathrm{out}}})$, which proves the claimed inequality.

If the ball condition holds almost surely, uniform sampling of t_{out}, the preceding inequality, Cauchy–Schwarz, and Theorem 1 give

$$\mathbb{E}\big(\operatorname{dist}(0, \partial_\varrho p_\gamma(x^{t_{\mathrm{out}}}))\big) \le \left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big(\|\mathcal{G}_{\eta,\varrho}(x^t)\|_2^2\big)\right)^{1/2} \le \varepsilon.$$

Finally, Corollary 1 identifies the two Goldstein subdifferentials whenever its local-equivalence conditions hold on the query ball, so the same inequalities apply to ϕ. □

Remark 2 (Vanishing-radius limit). If p_γ is locally Lipschitz near $\bar{x}$, $\varrho_j \downarrow 0$, $\varepsilon_j \downarrow 0$, $x^j \to \bar{x}$, and $\operatorname{dist}(0, \partial_{\varrho_j} p_\gamma(x^j)) \le \varepsilon_j$, then outer semicontinuity of the Clarke subdifferential and Carathéodory's theorem give $0 \in \partial_C p_\gamma(\bar{x})$. If $p_\gamma = \phi$ near $\bar{x}$, then $0 \in \partial_C p_\gamma(\bar{x}) = \partial_C \phi(\bar{x})$.

C Detailed application checks

The examples below verify relative submodularity and branch regularity. Submodular constraints preserve the fixed-multiplier SFM structure under nonnegative multipliers, but they do *not* by themselves verify the binary-optimum condition of Proposition 8. Exact dual support must still be proved for the application class or measured on the instances.

C.1 Scope and two reusable checks

The framework minimizes submodular *costs*. It does not cover the more familiar problem of maximizing a monotone submodular utility, such as coverage or influence (Nemhauser et al., 1978; Kempe et al., 2003). Taking a complement does not repair the mismatch: the negative or uncovered version is generally supermodular, so the resulting minimization remains hard.

In the examples, the threshold β is often the supremum of a ratio such as $a_{ij}(x)/w_{ij}(x)$. Assumption 2 requires this bound on the neighborhood U, not only on $\mathcal{X}$. Accordingly, all application suprema below are taken after choosing U bounded with compact closure in the branch smoothness domain. Continuity then makes a ratio bounded when its denominator is uniformly separated from zero, or when its numerator vanishes sufficiently fast wherever the denominator does. Pointwise positivity without compact uniform separation is not enough.

The Rademacher argument for the unconstrained finite-minimum oracle does not require tie sets to have measure zero. If their geometry is nevertheless of interest, a sufficient condition is that every relevant branch difference is real-analytic and is not identically zero on any connected component (Krantz & Parks, 2002). Real analyticity alone does not verify the second condition.

For graph models, connectedness of the graph or the presence of one nonzero unary term does not show that every compared branch gap for both g and $f + \gamma g$ is non-identical. That optional analytic check must be carried out pair by pair. It is deliberately not part of the main oracle assumption.

C.2 Pairwise graph-cut energies

For $i \neq j$ and a base label with $y_i = y_j = 0$, unary terms and terms not involving both coordinates have zero discrete second difference. The three pairwise primitives needed below satisfy

$$w\mathbf{1}\{y_i = 1, y_j = 0\} \quad \Longrightarrow \quad \Delta_{ij} = -w, \tag{21}$$

$$w\mathbf{1}\{y_i \neq y_j\} \quad \Longrightarrow \quad \Delta_{ij} = -2w, \tag{22}$$

$$a\mathbf{1}\{y_i = y_j\} \quad \Longrightarrow \quad \Delta_{ij} = 2a. \tag{23}$$

Thus the directed and symmetric separation costs are submodular, whereas a positive agreement cost is supermodular. Sums of the first two forms with nonnegative weights, together with arbitrary unary terms, are the standard binary graph-representable energies minimized by an s–t cut (Boykov et al., 2001; Kolmogorov & Zabih, 2004).

Under the interface (5), each such cut is one exact SFM call. For graph-representable objectives and constraints, successive multiplier values induce graph-capacity updates, so a compatible dynamic min-cut solver may reuse residual state, repair the affected capacities, and reaugment to an exact cut (Kohli & Torr, 2007). This reuse is implementation dependent because consecutive objective differences need not be submodular. It can reduce wall-clock cost without changing the exact SFM semantics or the asymptotic SFM-call counts of Theorems 1 and 2.

Electronic-design automation. The constrained EDA experiment in Section 5.2 uses symmetric cut terms in both the LL objective and its constraint. Equation (22) supplies the reusable submodularity check, while Appendix F records the instance construction, branch regularity, and experimental details.

Binary image segmentation. Consider a Potts lower energy with edge costs $\lambda_{ij}|y_i - y_j|$ and a competing upper cost with agreement penalties $\mu_{ij}(x)\mathbf{1}\{y_i = y_j\}$. The lower energy is submodular, and the upper cost has relative-submodularity threshold $\beta = \sup_{x \in U,(i,j)\in N} \mu_{ij}(x)/\lambda_{ij} < \infty$. Quadratic color terms and smooth $\kappa_i^0, \kappa_i^1, \mu_{ij}$ give the required C^1 branch regularity. Optional size constraints remain modular but again require discrete-duality verification.

C.3 Concave costs of aggregate load

Pairwise energies are not the only source of relative submodularity. Let $W(x, y) = \sum_i w_i(x)y_i$, where $w_i(x) \geq 0$, and let $h \in C^2$. For $y_i = y_j = 0$, two applications of the fundamental theorem

of calculus give the exact identity

$$\Delta_{ij}(h \circ W)(x, y) = \int_0^{w_i(x)} \int_0^{w_j(x)} h''(W(x, y) + s + t)\, dt\, ds. \tag{24}$$

Consequently, $h \circ W$ is submodular whenever $h'' \leq 0$. If another C^2 function $\widetilde{h}$ obeys $\widetilde{h}''(t) \leq \beta|h''(t)|$ throughout the attainable range of W, then integrating over the same rectangle in (24) proves $\Delta_{ij}(\widetilde{h} \circ W) \leq -\beta\Delta_{ij}(h \circ W)$.

For example, let $g(x, y) = \log(1 + W(x, y))$ and $f(x, y) = W(x, y)^2$. The lower cost models economies of scale and is submodular because $h''(t) = -(1 + t)^{-2}$. If $W_{\max} := \sup_{x \in U} \sum_i w_i(x) < \infty$, then $\widetilde{h}''(t) = 2 \leq \beta/(1+t)^2$ for all $t \in [0, W_{\max}]$ when $\beta = 2(1+W_{\max})^2$. Thus $f + \beta g$ is submodular. Affine nonnegative $w_i(x)$ make all branches smooth and Lipschitz on the compact closure of U.

D Verification of exact Lagrangian support and bounded multipliers

D.1 Exact-support characterization and structured sufficient conditions

For fixed x, suppose $h(x, \cdot)$ is submodular and $c(x, \cdot)$ is componentwise submodular. Let $\widehat{h}_x$ and $\widehat{c}_x$ denote their scalar and componentwise Lovász extensions in the binary argument. These extensions are convex polyhedral. For this subsection, define

$$P_h(x) := \min_{y \in \mathcal{Y}(x)} h(x, y), \quad V_h(x) := \sup_{\nu \geq 0} \min_{y \in \mathcal{Y}} \big(h(x, y) + \nu^\top c(x, y)\big). \tag{25}$$

Proposition 8 (Exact discrete-duality characterization)**.** *Fix x, suppose $h(x, \cdot)$ is submodular and $c(x, \cdot)$ is componentwise submodular, and suppose the discrete feasible set is nonempty. Then the supremum in $V_h(x)$ is attained and*

$$V_h(x) = \min_{u \in [0,1]^d, \widehat{c}_x(u) \leq 0} \widehat{h}_x(u) \leq P_h(x). \tag{26}$$

Moreover, the following statements are equivalent:

(i) $P_h(x) = V_h(x)$.

(ii) The convex problem in (26) *has a binary minimizer.*

(iii) there exist $y^\star \in \mathcal{Y}$ and $\nu^\star \geq 0$ such that

$$\begin{aligned} c(x, y^\star) \leq 0, \quad y^\star \in \arg\min_{y \in \mathcal{Y}} \big(h(x, y) + (\nu^\star)^\top c(x, y)\big), \\ (\nu^\star)^\top c(x, y^\star) = 0. \end{aligned} \tag{27}$$

Thus $P_h(x)$ admits an *exact Lagrangian support* precisely when any, and hence all, of the equivalent conditions above hold. This proposition answers when exact discrete strong duality holds. It does not require strict feasibility or a multiplier bound.

Proof. For every $\nu \geq 0$, the function $h(x, \cdot) + \nu^\top c(x, \cdot)$ is submodular and its Lovász extension is $\widehat{h}_x + \nu^\top \widehat{c}_x$. Exact unconstrained SFM through the Lovász extension therefore gives

$$\min_{y \in \mathcal{Y}} \big(h(x, y) + \nu^\top c(x, y)\big) = \min_{u \in [0,1]^d} \big(\widehat{h}_x(u) + \nu^\top \widehat{c}_x(u)\big).$$

Thus $V_h(x)$ is exactly the Lagrangian dual of the convex problem in (26). Since $\widehat{h}_x$ and every component of $\widehat{c}_x$ are convex polyhedral, epigraph representations give an equivalent finite-dimensional linear program. It is feasible because the discrete feasible set is nonempty and has a finite optimum because $[0, 1]^d$ is compact. Linear-programming strong duality and dual attainment therefore prove the equality in (26) and attainment of the supremum in $V_h(x)$, without an additional constraint qualification. The inequality follows because every feasible binary vector is feasible for the relaxation.

If the relaxation has a binary minimizer, its value is at least $P_h(x)$ and hence equals $P_h(x)$. This proves (ii)⇒(i). Conversely, if $P_h(x) = V_h(x)$, any discrete primal minimizer also attains the relaxed value, proving (i)⇒(ii).

Suppose (iii) holds. Lagrangian weak duality and (27) give

$$V_h(x) \geq \min_y \big(h(x,y) + (\nu^\star)^\top c(x,y)\big) = h(x,y^\star) \geq P_h(x),$$

so equality holds. Conversely, suppose (i) holds, choose a discrete primal minimizer $y^\star$, and let $\nu^\star$ be a dual optimizer. Then

$$P_h(x) = V_h(x) \leq h(x,y^\star) + (\nu^\star)^\top c(x,y^\star) \leq h(x,y^\star) = P_h(x).$$

Every inequality is therefore an equality, which gives both Lagrangian optimality of $y^\star$ and complementarity. □

Threshold preservation gives a useful sufficient condition for the binary-optimum property in Proposition 8.

Proposition 9 (Threshold-preserving constraints imply exact support)**.** *Fix x, suppose the discrete feasible set is nonempty, and suppose that every $u \in [0,1]^d$ satisfying $\widehat{c}_x(u) \leq 0$ has the following property: the threshold label $y^t \in \mathcal{Y}$ defined by $y^t_i := \mathbf{1}\{u_i \geq t\}$ satisfies $c(x,y^t) \leq 0$ for almost every $t \in [0,1]$. Then $P_h(x) = V_h(x)$ for every submodular binary cost $h(x,\cdot)$, and hence $P_h(x)$ admits an exact Lagrangian support.*

Proof. For $u \in [0,1]^d$, the threshold representation of the Lovász extension gives $\widehat{h}_x(u) = \int_0^1 h(x,y^t)\,dt$. If u is feasible for (26), then y^t is a feasible binary label for almost every t by hypothesis. Therefore $h(x,y^t) \geq P_h(x)$ almost everywhere and $\widehat{h}_x(u) \geq P_h(x)$. Minimizing over relaxed-feasible u gives the same lower bound for the Lovász relaxation, while the reverse inequality follows because every feasible binary label is also relaxed-feasible. Thus the relaxation value equals $P_h(x)$, and Proposition 8 yields $P_h(x) = V_h(x)$ and exact Lagrangian support. □

Closure constraints provide a standard threshold-preserving representation. Implications $y_i \leq y_j$ and coordinate fixings $y_i = 0$ or $y_i = 1$ are represented respectively by $y_i - y_j \leq 0$, $y_i \leq 0$, and $1 - y_i \leq 0$, or by positive rescalings of these inequalities. Indeed, $u_i \leq u_j$ implies $\mathbf{1}\{u_i \geq t\} \leq \mathbf{1}\{u_j \geq t\}$ for every t, and thresholding also preserves either fixing for almost every $t \in [0,1]$. Hence Proposition 9 verifies the pointwise exact-support condition used in Corollary 4 for these representations. The conclusion does not follow merely because the feasible binary family has a ring-family description under another representation. Uniform strict feasibility has the separate multiplier-boundedness role described by Lemma 2.

General submodular constraints need not preserve thresholds, and this can fail even for modular constraints. Let $\mathcal{Y} = \{0,1\}$, $h(y) = 1 - y$, and $c(y) = 2y - 1$. The binary label $y = 0$ is strictly feasible, but the Lovász relaxation satisfies

$$\min_{0 \leq u \leq 1/2} (1-u) = \frac{1}{2} < 1 = \min_{y \in \{0,1\}, 2y-1 \leq 0} (1-y).$$

Thus the Lagrangian dual value is $1/2$, although the objective and constraint are submodular, the constraint is even modular, and a strictly feasible binary label exists. Submodular constraints make every nonnegatively supported Lagrangian objective tractable by SFM, but neither submodularity nor strict feasibility establishes exact support.

D.2 Uniform multiplier bounds

Exact support identifies the correct Lagrangian dual value but does not by itself bound a supporting multiplier. Uniform strict feasibility supplies this second step through the following elementary estimate, which then yields common compact multiplier boxes for the two constrained values.

Lemma 2 (Multiplier bound from binary strict feasibility)**.** *Fix x, suppose the discrete feasible set is nonempty, and suppose $P_h(x) = V_h(x)$. If there exists $y \in \mathcal{Y}$ such that $c(x,y) \leq -\sigma\mathbf{1}$ for some $\sigma > 0$, then every dual optimizer $\nu^\star$ satisfies*

$$\|\nu^\star\|_1 \leq \frac{h(x,y) - \min_{\bar{y} \in \mathcal{Y}} h(x,\bar{y})}{\sigma} \leq \frac{\max_{\bar{y} \in \mathcal{Y}} h(x,\bar{y}) - \min_{\bar{y} \in \mathcal{Y}} h(x,\bar{y})}{\sigma}. \tag{28}$$

The lemma assumes that exact support already exists. Its strictly feasible binary label is used only to bound the magnitude of a supporting multiplier and, under a uniform margin, to obtain a common compact multiplier set.

Proof. Fix a label $y \in \mathcal{Y}$ satisfying $c(x, y) \leq -\sigma\mathbf{1}$. Any dual optimizer satisfies $\min_{\bar{y}\in\mathcal{Y}}\big(h(x,\bar{y}) + (\nu^\star)^\top c(x,\bar{y})\big) \geq \min_{\bar{y}\in\mathcal{Y}} h(x,\bar{y})$, while evaluation at this strictly feasible binary label gives

$$\min_{\bar{y}\in\mathcal{Y}}\big(h(x,\bar{y}) + (\nu^\star)^\top c(x,\bar{y})\big) \leq h(x,y) + (\nu^\star)^\top c(x,y) \leq h(x,y) - \sigma\|\nu^\star\|_1.$$

Rearranging proves the first inequality in (28). The second follows from $h(x, y) \leq \max_{\bar{y}\in\mathcal{Y}} h(x, \bar{y})$. □

Define the global range of g by

$$\Delta_g := \max_{x\in\mathcal{X},y\in\mathcal{Y}} g(x,y) - \min_{x\in\mathcal{X},y\in\mathcal{Y}} g(x,y). \tag{29}$$

Applying this bound to the two constrained values gives the compact multiplier boxes used by the main algorithm.

Corollary 4 (Two-value exact Lagrangian representation)**.** *Suppose Assumptions 1 and 3 hold. Then, with*

$$Q := \frac{\Delta_g + 2\varrho L_g}{\sigma}, \quad R := \frac{\Delta_f + 2\varrho L_f + \gamma(\Delta_g + 2\varrho L_g)}{\sigma}, \quad \Lambda := [0,R]^m, \quad \mathcal{M} := [0,Q]^m,$$

the following exact representations hold for every $x \in \mathcal{X}^\varrho$:

$$\min_{y\in\mathcal{Y}(x)}\big(f(x,y) + \gamma g(x,y)\big) = \max_{\lambda\in\Lambda}\min_{y\in\mathcal{Y}}\big(f(x,y) + \gamma g(x,y) + \lambda^\top c(x,y)\big), \tag{30a}$$

$$v(x) = \max_{\mu\in\mathcal{M}}\min_{z\in\mathcal{Y}}\big(g(x,z) + \mu^\top c(x,z)\big). \tag{30b}$$

Proof. Assumption 3 gives exact Lagrangian supports for the costs g and $f + \gamma g$ and a strictly feasible binary label with the common margin σ at every $x \in \mathcal{X}^\varrho$. For any $x, x' \in \mathcal{X}^\varrho$, choose $\bar{x}, \bar{x}' \in \mathcal{X}$ within distance ϱ of x, x'. The branch-gradient bounds from Assumption 1 give

$$\max_{x\in\mathcal{X}^\varrho,y\in\mathcal{Y}} g(x,y) - \min_{x\in\mathcal{X}^\varrho,y\in\mathcal{Y}} g(x,y) \leq \Delta_g + 2\varrho L_g,$$
$$\max_{x\in\mathcal{X}^\varrho,y\in\mathcal{Y}} f(x,y) - \min_{x\in\mathcal{X}^\varrho,y\in\mathcal{Y}} f(x,y) \leq \Delta_f + 2\varrho L_f.$$

Applying Lemma 2 to both supports therefore gives $\|\mu^\star(x)\|_1 \leq Q$ and $\|\lambda^\star(x)\|_1 \leq R$. Thus the supporting multipliers lie in $\mathcal{M}$ and Λ, respectively, and their support identities prove (30) throughout $\mathcal{X}^\varrho$. □

Propositions 8 and 9 verify the exact-support component of Assumption 3. Lemma 2 uses its uniform strict-feasibility margin only to bound the supporting multipliers, and Corollary 4 derives the compact representations used by the two supported SFM dual oracles.

E Proofs for the constrained method

Setup. Appendices E.1–E.4 use the following objects. Throughout, $h \in \{g, f + \gamma g\}$ denotes one of the two supported objectives, and every definition is for a fixed h; $x \in \mathcal{X}^\varrho$ is a fixed query point unless a supremum over x is written. Define the constrained value and dual functions by

$$V_h(x) := \min_{y\in\mathcal{Y}(x)} h(x,y). \tag{C}$$

$$d_h(x,\nu) := \min_{y\in\mathcal{Y}}\big\{h(x,y) + \nu^\top c(x,y)\big\}, \quad \nu \geq 0. \tag{D}$$

The associated Lagrangian dual problem is

$$\max_{\nu\geq 0} \quad d_h(x,\nu) = \max_{\nu\geq 0}\min_{y\in\mathcal{Y}}\{h(x,y) + \nu^\top c(x,y)\}. \tag{LD}$$

For any $\nu \geq 0$ and any feasible $y \in \mathcal{Y}$, we have $d_h(x,\nu) \leq h(x,y) + \nu^\top c(x,y) \leq h(x,y)$. Taking the minimum over feasible labels gives weak duality:

$$d_h(x,\nu) \leq V_h(x).$$

Assumption 3 guarantees strong duality and dual attainment: there exists $\nu_h^\star \geq 0$ such that

$$d_h(x,\nu_h^\star) = V_h(x).$$

Uniform strict feasibility and uniformly bounded objective ranges ensure that all optimal multipliers lie in fixed compact boxes independent of the query point (Lemma 2, Corollary 4). We therefore choose

$$\mathcal{B}_h = [0,r]^m, \quad \mathcal{B}_{f+\gamma g} = \Lambda = [0,R]^m, \quad \mathcal{B}_g = \mathcal{M} = [0,Q]^m,$$

containing the optimal multipliers; $\mathcal{B}_h$ is the box $\mathcal{B}$ of Section 4. Then

$$V_h(x) = d_h(x,\nu_h^\star) \leq \max_{\nu \in \mathcal{B}_h} d_h(x,\nu) \leq V_h(x).$$

Hence,

$$V_h(x) = \max_{\nu \in \mathcal{B}_h} d_h(x,\nu). \tag{SD}$$

In the notation of Appendix D, $V_h = P_h$ under Assumption 3.

By the definition of the constrained penalty (2), $p_\gamma(x) = V_{f+\gamma g}(x) - \gamma V_g(x)$.

To analyze the averaged outputs, define the probability simplex

$$\Delta(\mathcal{Y}) := \left\{ a \in \mathbb{R}^{|\mathcal{Y}|} : a_y \geq 0 \text{ for all } y \in \mathcal{Y}, \quad \sum_{y \in \mathcal{Y}} a_y = 1 \right\}$$

and the averaged Lagrangian

$$L_h(x,a,\nu) := \sum_{y \in \mathcal{Y}} a_y \big[h(x,y) + \nu^\top c(x,y) \big].$$

Fix one call of Algorithm 2 with budget K. At step k, the call holds the multiplier ν^k and receives the SFM response y^k. Let

$$a^k := \delta_{y^k}, \quad \bar{a}^K := \frac{1}{K} \sum_{k=0}^{K-1} a^k, \quad \bar{\nu}^K := \frac{1}{K} \sum_{k=0}^{K-1} \nu^k.$$

Here a^k is the point mass at y^k. So $\bar{a}^K$ is the empirical distribution of the K responses: $\bar{a}_y^K$ is the fraction of steps that returned y. And $\bar{\nu}^K$ is the average of the K multipliers. The algorithm stores neither $\bar{a}^K$ nor $\bar{\nu}^K$; they are for analysis only.

Three constants appear in the bounds below:

$$C_\varrho := \sup_{x \in \mathcal{X}^\varrho,\ y \in \mathcal{Y}} \|c(x,y)\|_2 < \infty, \quad \Delta_h^\varrho := \sup_{x \in \mathcal{X}^\varrho} \Big(\max_{y \in \mathcal{Y}} h(x,y) - \min_{y \in \mathcal{Y}} h(x,y) \Big) < \infty$$

$$B_h := \sup_{x \in \mathcal{X}^\varrho,\ y \in \mathcal{Y}, \nu \in \widetilde{\mathcal{B}}_h} \big\| \nabla_x h(x,y) + \nu^\top \nabla_x c(x,y) \big\|_2 < \infty$$

E.1 Proof of Proposition 4

Proof. By (SD) for $h = f + \gamma g$ and $h = g$, and $p_\gamma = V_{f+\gamma g} - \gamma V_g$ from the setup, substituting the two representations into (2) gives (11).

The first fixed-multiplier objective is submodular because $f + \gamma g + \lambda^\top c = (f + \beta g) + (\gamma - \beta) g + \lambda^\top c$ is a sum of submodular functions when $\gamma \geq \beta$, $\lambda \geq 0$, and c is componentwise submodular. The second objective is submodular because g is submodular, c is componentwise submodular, and $\mu \geq 0$. Hence each inner minimum is one exact SFM call. For fixed $\nu \in \mathcal{B}_h$, $d_h(\cdot,\nu)$ is a minimum of C^1 branches with gradient norm at most $L_h + \|\nu\|_1 L_c$ by (6) and Lemma 1, and a maximum of Lipschitz functions with a common modulus keeps that modulus. So the supported value for $f + \gamma g$ is $(L_f + \gamma L_g + mRL_c)$-Lipschitz, and the supported value for g is $(L_g + mQL_c)$-Lipschitz. Their difference in (11) therefore has modulus $L_f + mRL_c + 2\gamma L_g + \gamma mQL_c$. □

E.2 SUPPORTED-BASIS CONDITIONING AND FIRST-ORDER RECOVERY

Claim 1. *For every fixed (x, ν), consider the dual problem $\max_{\nu \in \mathcal{B}_h} \min_{a \in \Delta(\mathcal{Y})} L_h(x, a, \nu)$, which is an equivalent reformulation of* (LD)*, we have*

(i) $\min_{a \in \Delta(\mathcal{Y})} L_h(x, a, \nu) = \min_{y \in \mathcal{Y}} \{h(x, y) + \nu^\top c(x, y)\} = d_h(x, \nu)$

(ii) $\max_{q \in \mathcal{B}_h} L_h(x, a, q) \geq V_h(x) \geq d_h(x, \nu)$

Proof of the claim. (i) For fixed (x, ν), let $b_y := h(x, y) + \nu^\top c(x, y)$. Since $a \in \Delta(\mathcal{Y})$,

$$\sum_{y \in \mathcal{Y}} a_y b_y \geq \sum_{y \in \mathcal{Y}} a_y \min_{u \in \mathcal{Y}} b_u = \min_{u \in \mathcal{Y}} b_u.$$

Equality is attained by placing all mass on a minimizer: take $a = \delta_{y^\star}$, where $y^\star \in \arg\min_{y \in \mathcal{Y}} b_y$. Here, $\delta_{y^\star}$ denotes the probability vector with unit mass at $y^\star$. Thus,

$$\min_{a \in \Delta(\mathcal{Y})} L_h(x, a, \nu) = \min_{y \in \mathcal{Y}} b_y = d_h(x, \nu).$$

(ii) By (SD), there exists $\nu^\star \in \mathcal{B}_h$ such that $d_h(x, \nu^\star) = V_h(x)$. For any $a \in \Delta(\mathcal{Y})$, part (i) gives

$$\begin{aligned} \max_{q \in \mathcal{B}_h} L_h(x, a, q) &\geq L_h(x, a, \nu^\star) \\ &\geq \min_{b \in \Delta(\mathcal{Y})} L_h(x, b, \nu^\star) \\ &= d_h(x, \nu^\star) = V_h(x). \end{aligned}$$

Moreover, for any $\nu \in \mathcal{B}_h$,

$$V_h(x) \overset{\text{(SD)}}{=} \max_{q \in \mathcal{B}_h} d_h(x, q) \geq d_h(x, \nu).$$

△

Combining (SD) and Claim 1 gives the equivalent max–min representation

$$V_h(x) = \max_{\nu \in \mathcal{B}_h} \min_{a \in \Delta(\mathcal{Y})} L_h(x, a, \nu).$$

Fix h and x. Define

$$\mathrm{Gap}_h(x, a, \nu) := \max_{q \in \mathcal{B}_h} L_h(x, a, q) - \min_{b \in \Delta(\mathcal{Y})} L_h(x, b, \nu). \tag{Gap}$$

This is the saddle gap of (12), written with the box of h. By Claim 1, the gap function is nonnegative, i.e. $\mathrm{Gap}_h(x, a, \nu) \geq d_h(x, \nu) - d_h(x, \nu) = 0$.

At each iteration, Algorithm 2 computes

$$y^k \in \arg\min_{y \in \mathcal{Y}} \{h(x, y) + (\nu^k)^\top c(x, y)\}$$

through one exact SFM call (5) and updates

$$\nu^{k+1} = \Pi_{\mathbb{R}^m_+}(\nu^k + \theta c_k), \quad c_k := c(x, y^k).$$

Proposition 10 needs a bound on the multiplier trajectory that does not depend on the warm start.

Lemma 3 (Bounded warm-started DUAL trajectory)**.** *Suppose Assumptions 1, 2, and 3 hold. Consider all successive calls to Algorithm 2 for one supported objective h, with the final multiplier of each call used to initialize the next. Let $r = R$ for $h = f + \gamma g$ and $r = Q$ for $h = g$. If $C_\varrho > 0$, the first call starts from zero, and every call uses $\theta = r\sqrt{m}/(C_\varrho\sqrt{K})$, then every generated multiplier satisfies $\|\nu^k\|_2 \leq \widetilde{r}_h := \Delta_h^\varrho/\sigma + r\sqrt{m}C_\varrho/(2\sigma) + r\sqrt{m}$. Consequently, every within-call average belongs to $[0, \widetilde{r}_h]^m$. The same conclusion holds trivially when $C_\varrho = 0$ and $\theta = 0$.*

Proof. Write $c_k := c(x, y^k)$ and $\tau := \Delta_h^\varrho/\sigma + \theta C_\varrho^2/(2\sigma)$.

We first bound $\langle c_k, \nu^k \rangle$. Let $\bar{y}$ be the strictly feasible label with margin σ from Assumption 3, so $c(x, \bar{y}) \le -\sigma \mathbf{1}$. Since y^k minimizes $h(x, \cdot) + (\nu^k)^\top c(x, \cdot)$ and $\nu^k \ge 0$,

$$h(x, y^k) + \langle c_k, \nu^k \rangle \le h(x, \bar{y}) + (\nu^k)^\top c(x, \bar{y}) \le h(x, \bar{y}) - \sigma \|\nu^k\|_1.$$

Rearranging and using $\|\nu^k\|_1 \ge \|\nu^k\|_2$,

$$\langle c_k, \nu^k \rangle \le h(x, \bar{y}) - h(x, y^k) - \sigma \|\nu^k\|_1 \le \Delta_h^\varrho - \sigma \|\nu^k\|_2.$$

Projection onto $\mathbb{R}_+^m$ is nonexpansive and fixes ν^k and 0. Hence

$$\|\nu^{k+1} - \nu^k\|_2 = \|\Pi_{\mathbb{R}_+^m}(\nu^k + \theta c_k) - \Pi_{\mathbb{R}_+^m}(\nu^k)\|_2 \le \theta \|c_k\|_2 \le \theta C_\varrho,$$

and

$$\begin{aligned} \|\nu^{k+1}\|_2^2 &= \|\Pi_{\mathbb{R}_+^m}(\nu^k + \theta c_k) - \Pi_{\mathbb{R}_+^m}(0)\|_2^2 \le \|\nu^k + \theta c_k\|_2^2 \\ &= \|\nu^k\|_2^2 + 2\theta \langle c_k, \nu^k \rangle + \theta^2 \|c_k\|_2^2 \\ &\overset{\text{Step 1}}{\le} \|\nu^k\|_2^2 + 2\theta\big(\Delta_h^\varrho - \sigma \|\nu^k\|_2\big) + \theta^2 C_\varrho^2. \end{aligned}$$

If $\|\nu^k\|_2 \ge \tau$, then $2\theta(\Delta_h^\varrho - \sigma\|\nu^k\|_2) + \theta^2 C_\varrho^2 \le 0$, so Step 2 gives $\|\nu^{k+1}\|_2 \le \|\nu^k\|_2$. If $\|\nu^k\|_2 < \tau$, then Step 2 gives $\|\nu^{k+1}\|_2 \le \|\nu^k\|_2 + \theta C_\varrho < \tau + \theta C_\varrho$. In both cases,

$$\|\nu^{k+1}\|_2 \le \max\{\|\nu^k\|_2,\ \tau + \theta C_\varrho\}.$$

With $\theta = r\sqrt{m}/(C_\varrho \sqrt{K})$ and $K \ge 1$, $\theta \le r\sqrt{m}/C_\varrho$, so

$$\tau + \theta C_\varrho = \frac{\Delta_h^\varrho}{\sigma} + \frac{\theta C_\varrho^2}{2\sigma} + \theta C_\varrho \le \frac{\Delta_h^\varrho}{\sigma} + \frac{r\sqrt{m} C_\varrho}{2\sigma} + r\sqrt{m} = \widetilde{r}_h.$$

This bound does not depend on K, so it holds for every call in the chain even when K changes between calls. The first call starts from $\nu^0 = 0$, and each later call starts from the final multiplier of the previous one. Step 3 and induction over all steps of all calls give $\|\nu^k\|_2 \le \widetilde{r}_h$ for every generated multiplier.

Every $\nu^k \ge 0$, and $\|\nu^k\|_\infty \le \|\nu^k\|_2 \le \widetilde{r}_h$, so $\nu^k \in [0, \widetilde{r}_h]^m$. A within-call average is a convex combination of such points, and the box is convex.

If $C_\varrho = 0$, then $c_k = 0$ for every k, $\theta = 0$, and $\nu^k = \nu^0 = 0$ throughout. □

Proposition 10 (Finite-time dual accuracy)**.** *Suppose Assumptions 1, 2, and 3 hold. Fix $x \in \mathcal{X}^\varrho$. When $C_\varrho > 0$, run Algorithm 2 for $h = f + \gamma g$ with $\theta = R\sqrt{m}/(C_\varrho\sqrt{K_\lambda})$, and for $h = g$ with $\theta = Q\sqrt{m}/(C_\varrho\sqrt{K_\mu})$. When $C_\varrho = 0$, take $\theta = 0$ in both runs. Let $\widehat{d}_F$ and $\widehat{d}_g$ denote the largest exact dual values recorded along the two multiplier trajectories. There exist finite constants Γ_F and Γ_g, independent of K_λ and K_μ, such that*

$$0 \le V_{f+\gamma g}(x) - \widehat{d}_F \le \frac{\Gamma_F}{\sqrt{K_\lambda}}, \quad 0 \le v(x) - \widehat{d}_g \le \frac{\Gamma_g}{\sqrt{K_\mu}}, \tag{31}$$

and their averaged outputs satisfy

$$\mathrm{Gap}_{f+\gamma g}(x, \bar{a}^{K_\lambda}, \bar{\lambda}^{K_\lambda}) \le \frac{\Gamma_F}{\sqrt{K_\lambda}}, \quad \mathrm{Gap}_g(x, \bar{b}^{K_\mu}, \bar{\mu}^{K_\mu}) \le \frac{\Gamma_g}{\sqrt{K_\mu}}. \tag{32}$$

The two runs use exactly K_λ and K_μ SFM calls.

Proof. We first establish the following claim.

Claim 2. *For every comparison multiplier $q \in \mathcal{B}_h$,*

$$\|\nu^{k+1} - q\|_2^2 \le \|\nu^k + \theta c_k - q\|_2^2.$$

Proof of the claim. Let $u := \nu^k + \theta c_k$. Since the projection onto $\mathbb{R}^m_+$ is componentwise, $\nu_i^{k+1} = \max\{u_i, 0\}$.
For $q \in \mathcal{B}_h \subseteq \mathbb{R}^m_+$, we have $q_i \geq 0$, and hence

$$|\nu_i^{k+1} - q_i| = \begin{cases} |u_i - q_i|, & u_i \geq 0, \\ q_i \leq q_i - u_i = |u_i - q_i|, & u_i < 0. \end{cases}$$

Squaring and summing over coordinates proves the claim. △

Expanding the squared norm and using $\|c_k\|_2 \leq C_\varrho$ gives

$$\|\nu^{k+1} - q\|_2^2 \leq \|\nu^k - q\|_2^2 + 2\theta\langle c_k, \nu^k - q\rangle + \theta^2 C_\varrho^2. \tag{①}$$

Since $a^k = \delta_{y^k}$,

$$\begin{aligned} \langle c_k, q - \nu^k\rangle &= h(x, y^k) + q^\top c(x, y^k) - \big[h(x, y^k) + (\nu^k)^\top c(x, y^k)\big] \\ &= L_h(x, a^k, q) - L_h(x, a^k, \nu^k). \end{aligned} \tag{②}$$

Substituting into the projection inequality and rearranging,

$$L_h(x, a^k, q) - L_h(x, a^k, \nu^k) \leq \frac{\|\nu^k - q\|_2^2 - \|\nu^{k+1} - q\|_2^2}{2\theta} + \frac{\theta C_\varrho^2}{2}.$$

Summing over $k = 0, \ldots, K-1$, the squared-distance terms telescope:

$$\sum_{k=0}^{K-1} \big(\|\nu^k - q\|_2^2 - \|\nu^{k+1} - q\|_2^2\big) = \|\nu^0 - q\|_2^2 - \|\nu^K - q\|_2^2.$$

Therefore,

$$\begin{aligned} \sum_{k=0}^{K-1} \big[L_h(x, a^k, q) - L_h(x, a^k, \nu^k)\big] &\leq \frac{\|\nu^0 - q\|_2^2 - \|\nu^K - q\|_2^2}{2\theta} + \frac{\theta K C_\varrho^2}{2} \\ &\leq \frac{\|\nu^0 - q\|_2^2}{2\theta} + \frac{\theta K C_\varrho^2}{2}. \end{aligned}$$

Let $\widetilde{\mathcal{B}}_h := [0, \widetilde{r}_h]^m$ be the fixed compact region containing the warm-started multiplier trajectories from Lemma 3, and define

$$D_h := \widetilde{r}_h + r\sqrt{m} \geq \sup_{\nu \in \widetilde{\mathcal{B}}_h, q \in \mathcal{B}_h} \|\nu - q\|_2.$$

The uniform trajectory bound makes D_h independent of the incoming warm start. Since $\nu^0 \in \widetilde{\mathcal{B}}_h$, dividing by K and maximizing over q yields

$$\max_{q \in \mathcal{B}_h} \frac{1}{K} \sum_{k=0}^{K-1} \big[L_h(x, a^k, q) - L_h(x, a^k, \nu^k)\big] \leq \frac{D_h^2}{2\theta K} + \frac{\theta C_\varrho^2}{2}. \tag{L-GAP}$$

We then look into the saddle gap of the averaged outputs.

Claim 3. *(i) If y^k is the optimal solution with multiplier ν^k, i.e., the exact SFM responses, then $L_h(x, a^k, \nu^k) = d_h(x, \nu^k)$, (ii) and their averages satisfy*

$$d_h(x, \bar{\nu}^K) \geq \frac{1}{K} \sum_{k=0}^{K-1} d_h(x, \nu^k) = \frac{1}{K} \sum_{k=0}^{K-1} L_h(x, a^k, \nu^k).$$

Proof of the claim. For (i), since y^k is an exact minimizer at ν^k,

$$\begin{aligned} d_h(x, \nu^k) &\overset{(D)}{=} \min_{y \in \mathcal{Y}} \big\{h(x, y) + (\nu^k)^\top c(x, y)\big\} \\ &= h(x, y^k) + (\nu^k)^\top c(x, y^k) \\ &= L_h(x, a^k, \nu^k). \end{aligned}$$

For (ii), using $\bar{\nu}^K = K^{-1}\sum_{k=0}^{K-1}\nu^k$,

$$\begin{aligned} d_h(x,\bar{\nu}^K) &\overset{\text{linearity of (D)}}{=} \min_{y\in\mathcal{Y}}\left\{\frac{1}{K}\sum_{k=0}^{K-1}\left[h(x,y)+(\nu^k)^\top c(x,y)\right]\right\} \\ &\geq \frac{1}{K}\sum_{k=0}^{K-1}\min_{y\in\mathcal{Y}}\left\{h(x,y)+(\nu^k)^\top c(x,y)\right\} \\ &\overset{\text{(D)}}{=} \frac{1}{K}\sum_{k=0}^{K-1} d_h(x,\nu^k) \\ &\overset{\text{Claim 3(i)}}{=} \frac{1}{K}\sum_{k=0}^{K-1} L_h(x,a^k,\nu^k). \end{aligned}$$

This is also Jensen's inequality for the concave function $d_h(x,\cdot)$, which is a pointwise minimum of affine functions.

△

Then we consider the gap of the average,

$$\begin{aligned} \mathrm{Gap}_h(x,\bar{a}^K,\bar{\nu}^K) &\overset{\text{(Gap)}}{=} \max_{q\in\mathcal{B}_h} L_h(x,\bar{a}^K,q) - \min_{b\in\Delta(\mathcal{Y})} L_h(x,b,\bar{\nu}^K) \\ &\overset{\text{Claim 1(i)}}{=} \max_{q\in\mathcal{B}_h} L_h(x,\bar{a}^K,q) - d_h(x,\bar{\nu}^K) \\ &\overset{\text{linearity of }\bar{a}^K}{=} \max_{q\in\mathcal{B}_h}\frac{1}{K}\sum_{k=0}^{K-1} L_h(x,a^k,q) - d_h(x,\bar{\nu}^K) \\ &\overset{\text{Claim 3(ii)}}{\leq} \max_{q\in\mathcal{B}_h}\frac{1}{K}\sum_{k=0}^{K-1}\left[L_h(x,a^k,q) - L_h(x,a^k,\nu^k)\right] \\ &\overset{\text{(L-GAP)}}{\leq} \frac{D_h^2}{2\theta K} + \frac{\theta C_\varrho^2}{2}. \end{aligned}$$

Choose $\theta = \alpha_h/\sqrt{K}$, where $\alpha_h > 0$ is independent of K. Then

$$\mathrm{Gap}_h(x,\bar{a}^K,\bar{\nu}^K) \leq \frac{\Gamma_h}{\sqrt{K}}, \quad \Gamma_h := \frac{D_h^2}{2\alpha_h} + \frac{\alpha_h C_\varrho^2}{2}.$$

The original stepsize choice corresponds to the constant stepsize of the subgradient method with a fixed number of steps (Boyd, 2014, §3.3),

$$\alpha_h = \frac{r\sqrt{m}}{C_\varrho}, \quad \mathcal{B}_h = [0,r]^m,$$

when $C_\varrho > 0$; the supporting box may be enlarged so that $r > 0$. If $C_\varrho = 0$, the multiplier term vanishes and one exact SFM call suffices.

This establishes a saddle-gap bound for the averaged outputs using K exact SFM calls for each objective h. Taking q to be a dual maximizer in the projection inequality and using the supergradient inequality gives the same bound for $V_h(x) - \widehat{d}$. Lemma 3 places every computed multiplier and its average in $[0,\widetilde{r}_h]^m$. Substituting $r = R$ and $r = Q$, with $\Gamma_F := \Gamma_{f+\gamma g}$ and Γ_g denoting the corresponding constant for $h = g$, proves (31) and (32). When $r = 0$ or $C_\varrho = 0$, both conclusions are immediate. Each ascent iteration invokes exactly one SFM call, proving both oracle counts. A bound on the distance to the saddle set, or on the sensitivity error, requires further analysis. □

The same sequence of inner problems may also admit exact solver-state reuse.

Remark 3 (Hierarchy of SFM state reuse). At fixed x, every supported objective $h(x,\cdot) + (\nu^k)^\top c(x,\cdot)$ is submodular because $\nu^k \geq 0$. The difference between consecutive objectives, $(\nu^{k+1} - \nu^k)^\top c(x,\cdot)$, need not be submodular because the multiplier difference can have negative entries, so solver-state reuse is implementation dependent. When the objective and constraints are graph representable, multiplier changes correspond to graph-capacity updates and dynamic min-cut state may still be reused while preserving exact oracle semantics (Kohli & Torr, 2007). Such reuse changes neither the required exactness nor the theoretical SFM-call count.

For $h \in \{f + \gamma g, g\}$, let $\mathcal{B}_h = [0,r]^m$ be its supporting box Λ or $\mathcal{M}$, and let $\widetilde{\mathcal{B}}_h := [0,\widetilde{r}_h]^m$ be the warm-started trajectory box from Lemma 3. Distances between multiplier-distribution pairs use the norm $\|\nu\|_2 + \|a\|_1$. Write $n := |\mathcal{Y}|$.

Lemma 4 (Fixed-multiplier finite-minimum calculus)**.** *Suppose Assumption 1 holds. For fixed (h,ν), $d_h(\cdot,\nu)$ is locally Lipschitz and*

$$\partial_C^x d_h(x,\nu) = \operatorname{conv}\left\{\nabla_x h(x,y) + \nu^\top \nabla_x c(x,y) : y \in \arg\min_{\bar{y}\in\mathcal{Y}} \big(h(x,\bar{y}) + \nu^\top c(x,\bar{y})\big)\right\}. \tag{33}$$

Moreover, $c(x,y)$ is a supergradient of $d_h(x,\cdot)$ at ν for every minimizing label y.

Proof. The first claim follows from Lemma 1 applied to the finitely many C^1 branches. The supergradient claim follows by evaluating the active affine branch at any other multiplier. □

Fix h and x. Recall that we minimize over a and maximize over ν, so for any $a \in \Delta(\mathcal{Y})$ and $\nu \in \mathcal{B}_h$,

$$\min_{b\in\Delta(\mathcal{Y})} L_h(x,b,\nu) \leq \min_{b\in\Delta(\mathcal{Y})} \max_{q\in\mathcal{B}_h} L_h(x,b,q) \leq \max_{q\in\mathcal{B}_h} L_h(x,a,q).$$

Let $\mathcal{S}_h(x)$ denote the set of exact saddle points of $L_h(x,\cdot,\cdot)$ on $\Delta(\mathcal{Y}) \times \mathcal{B}_h$. Thus, $(\nu^\star, a^\star) \in \mathcal{S}_h(x)$ if and only if

$$L_h(x,a^\star,\nu) \leq L_h(x,a^\star,\nu^\star) \leq L_h(x,a,\nu^\star), \quad a \in \Delta(\mathcal{Y}), \quad \nu \in \mathcal{B}_h. \tag{34}$$

Equivalently,

$$a^\star \in \arg\min_{a\in\Delta(\mathcal{Y})} L_h(x,a,\nu^\star), \quad \nu^\star \in \arg\max_{\nu\in\mathcal{B}_h} L_h(x,a^\star,\nu).$$

Claim 4. *For $(\nu,a) \in \mathcal{B}_h \times \Delta(\mathcal{Y})$,*

$$\operatorname{Gap}_h(x,a,\nu) = 0 \quad \Longleftrightarrow \quad (\nu,a) \in \mathcal{S}_h(x).$$

Proof of the claim. We can write

$$\operatorname{Gap}_h(x,a,\nu) = \underbrace{\max_{q\in\mathcal{B}_h} L_h(x,a,q) - L_h(x,a,\nu)}_{\text{(a) } \geq 0} + \underbrace{L_h(x,a,\nu) - \min_{b\in\Delta(\mathcal{Y})} L_h(x,b,\nu)}_{\text{(b) } \geq 0}.$$

Term (a) is the maximum increase obtained by changing the multiplier while holding a fixed. Term (b) is the maximum decrease obtained by changing the distribution while holding ν fixed.

Both terms are nonnegative, so the gap vanishes if and only if both terms vanish, i.e.,

$$\nu \in \arg\max_{q\in\mathcal{B}_h} L_h(x,a,q), \quad a \in \arg\min_{b\in\Delta(\mathcal{Y})} L_h(x,b,\nu).$$

which is precisely the saddle-point condition $(\nu,a) \in \mathcal{S}_h(x)$. △

Claim 5. *Let $\widetilde{\mathcal{B}}_h \subseteq \mathbb{R}^m_+$ be a compact box containing $\mathcal{B}_h$ and the multiplier trajectories. Under the multiplier bound in Assumption 3 (Lemma 2),*

$$\mathcal{S}_h(x) = \left\{(\nu,a) \in \widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y}) : \operatorname{Gap}_h(x,a,\nu) = 0\right\}.$$

This is the description of $\mathcal{S}_h(x)$ used in (12)*.*

Proof of the claim. For any $(\nu, a) \in \widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y})$, Claim 1 and weak duality give

$$\max_{q \in \mathcal{B}_h} L_h(x, a, q) \geq V_h(x) \geq d_h(x, \nu).$$

If $\mathrm{Gap}_h(x, a, \nu) = 0$, then

$$\max_{q \in \mathcal{B}_h} L_h(x, a, q) = V_h(x) = d_h(x, \nu).$$

Thus, ν is a globally optimal dual multiplier. The multiplier bound under Assumption 3 places every such multiplier in $\mathcal{B}_h$. Claim 4 therefore gives $(\nu, a) \in \mathcal{S}_h(x)$.

Conversely, if $(\nu, a) \in \mathcal{S}_h(x)$, then $(\nu, a) \in \mathcal{B}_h \times \Delta(\mathcal{Y}) \subseteq \widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y})$, and Claim 4 gives $\mathrm{Gap}_h(x, a, \nu) = 0$. △

Claim 6 (the saddle set $\mathcal{S}_h(x)$ can be represented by linear inequalities)**.** *. Fix h and x. For each $q \in \mathrm{vert}(\mathcal{B}_h)$ and $y \in \mathcal{Y}$, define $\ell_{q,y}(\nu, a) := \sum_{u \in \mathcal{Y}} a_u \big[h(x, u) + q^\top c(x, u)\big] - h(x, y) - \nu^\top c(x, y)$. Then $\mathcal{S}_h(x)$ is the polyhedron*

$$\mathcal{S}_h(x) = \Big\{(\nu, a) \in \widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y}) : \ell_{q,y}(\nu, a) \leq 0 \quad \forall\, q \in \mathrm{vert}(\mathcal{B}_h),\ y \in \mathcal{Y}\Big\}. \tag{35}$$

Moreover, for every $(\nu, a) \in \widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y})$,

$$\max_{q \in \mathrm{vert}(\mathcal{B}_h), y \in \mathcal{Y}} [\ell_{q,y}(\nu, a)]_+ = \mathrm{Gap}_h(x, a, \nu), \quad [t]_+ := \max\{t, 0\}.$$

Proof of the claim. Following Claim 1(i), there is

$$\begin{aligned}
\mathrm{Gap}_h(x, a, \nu) &= \max_{q \in \mathcal{B}_h} L_h(x, a, q) - \min_{y \in \mathcal{Y}} \big\{h(x, y) + \nu^\top c(x, y)\big\} \\
&= \max_{q \in \mathrm{vert}(\mathcal{B}_h), y \in \mathcal{Y}} \left\{\sum_{u \in \mathcal{Y}} a_u \big[h(x, u) + q^\top c(x, u)\big] - h(x, y) - \nu^\top c(x, y)\right\} \\
&= \max_{q \in \mathrm{vert}(\mathcal{B}_h), y \in \mathcal{Y}} \ell_{q,y}(\nu, a),
\end{aligned}$$

where the second equality holds using $-\min f = \max(-f)$, and the third as an affine function on a box attains its maximum at a vertex. Since the gap is nonnegative on $\widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y})$,

$$\mathrm{Gap}_h(x, a, \nu) = 0 \quad \Longleftrightarrow \quad \ell_{q,y}(\nu, a) \leq 0 \quad \forall\, q \in \mathrm{vert}(\mathcal{B}_h),\ y \in \mathcal{Y}.$$

Together with Claim 5, this gives the stated representation of $\mathcal{S}_h(x)$. For fixed x, each $\ell_{q,y}$ is affine in (ν, a). The box and simplex constraints are also linear, so $\mathcal{S}_h(x)$ is a polyhedron.

Finally,

$$\max_{q,y} [\ell_{q,y}(\nu, a)]_+ = \Big[\max_{q,y} \ell_{q,y}(\nu, a)\Big]_+ = [\mathrm{Gap}_h(x, a, \nu)]_+ = \mathrm{Gap}_h(x, a, \nu).$$

△

Claim 7. *Write $n := |\mathcal{Y}|$ and $v := (\nu, a) \in \mathbb{R}^{m+n}$. There exist a matrix A_x and a vector b_x such that $\mathcal{S}_h(x) = \{v : A_x v \leq b_x\}$, where the rows encode the inequalities $\ell_{q,y} \leq 0$ and the box and simplex constraints, with the simplex equality represented by two inequalities. This is the coefficient system of Assumption 4: a supported basis matrix at a point of $\mathcal{S}_h(x)$ is a maximal linearly independent collection of rows of A_x active there. Moreover, for every $v = (\nu, a) \in \widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y})$,*

$$A_x v - b_x \leq \varepsilon \mathbf{1}, \quad \varepsilon := \mathrm{Gap}_h(x, a, \nu).$$

Proof of the claim. The linear inequality representation follows from Claim 6. For $v \in \widetilde{\mathcal{B}}_h \times \Delta(\mathcal{Y})$, the box and simplex constraints have nonpositive residuals. Each remaining residual satisfies

$$\ell_{q,y}(\nu, a) \leq \max_{q', y'} \ell_{q', y'}(\nu, a) = \mathrm{Gap}_h(x, a, \nu) = \varepsilon.$$

Since $\varepsilon \geq 0$, every residual is at most ε, giving $A_x v - b_x \leq \varepsilon \mathbf{1}$. △

Lemma 5 (Active-basis error bound)**.** *Suppose Assumptions 3 and 4 hold. After enlarging κ_{reg} by a fixed factor depending only on m and $|\mathcal{Y}|$, the supported saddle systems satisfy*

$$\operatorname{dist}\big((\nu,a),\mathcal{S}_h(x)\big) \le \kappa_{\mathrm{reg}}\,\mathrm{Gap}_h(x,a,\nu), \quad (\nu,a)\in\widetilde{\mathcal{B}}_h\times\Delta(\mathcal{Y}). \tag{36}$$

Proof. By Hoffman's error bound (Hoffman, 1952; Robinson, 1981), the distance to $\mathcal{S}_h(x)$ is bounded by a constant times the maximum constraint violation. By Claim 6, this violation equals $\mathrm{Gap}_h(x,a,\nu)$ for $(\nu,a)\in\widetilde{\mathcal{B}}_h\times\Delta(\mathcal{Y})$. We make the constant explicit.

Fix $v\in\widetilde{\mathcal{B}}_h\times\Delta(\mathcal{Y})$ and let $v^\star := \Pi_{\mathcal{S}_h(x)}(v),\ d := v-v^\star,\ \varepsilon := \mathrm{Gap}_h(x,a,\nu)$, where $\Pi_{\mathcal{S}_h(x)}(v) := \arg\min_{u\in\mathcal{S}_h(x)} \frac{1}{2}\|u-v\|_2^2$. If $d=0$, the desired bound is immediate. Suppose $d\neq 0$.

The point $v^\star$ solves the projection problem $\min_u \frac{1}{2}\|u-v\|_2^2 \quad \text{s.t.} \quad A_x u \le b_x$.
Its optimality conditions give $v^\star - v + A_x^\top\lambda = 0,\ \lambda\ge 0,\ \lambda_i\big((A_x v^\star)_i - b_{x,i}\big)=0$ for all i.
Let $r_i^\top$ denote the ith row of A_x, and define $J := \{i : r_i^\top v^\star = b_{x,i}\}$. By complementary slackness, $d := v-v^\star = \sum_{j\in J}\lambda_j r_j = A_{x,J}^\top\lambda_J,\ A_{x,J}v^\star = b_{x,J}$. By removing linearly dependent generators from this nonnegative combination, we can choose a linearly independent subset $I\subseteq J$ and nonnegative coefficients, again denoted by λ, such that $d = A_{x,I}^\top\lambda,\ A_{x,I}v^\star = b_{x,I},\ \lambda\ge 0$.

Extend I to a maximal linearly independent subset $J_0\subseteq J$. Assumption 4 gives $\sigma_{\min}(A_{x,J_0})\ge\kappa_{\mathrm{reg}}^{-1}$. Let $\widetilde{\lambda}\in\mathbb{R}^{|J_0|}$ be the extension of λ by zero on $J_0\setminus I$. Then $A_{x,J_0}^\top\widetilde{\lambda} = A_{x,I}^\top\lambda, \|\widetilde{\lambda}\|_2 = \|\lambda\|_2$. Therefore,

$$\|d\|_2 = \|A_{x,I}^\top\lambda\|_2 = \|A_{x,J_0}^\top\widetilde{\lambda}\|_2 \ge \sigma_{\min}(A_{x,J_0})\|\widetilde{\lambda}\|_2 \ge \kappa_{\mathrm{reg}}^{-1}\|\lambda\|_2.$$

By Cauchy–Schwarz inequality,

$$\begin{aligned}\|\lambda\|_1 = \sum_{i\in I}|\lambda_i| &\le \left(\sum_{i\in I}1^2\right)^{1/2}\left(\sum_{i\in I}\lambda_i^2\right)^{1/2} = \sqrt{|I|}\,\|\lambda\|_2\\ &\le \sqrt{m+n}\,\|\lambda\|_2 \le \sqrt{m+n}\,\kappa_{\mathrm{reg}}\|d\|_2,\end{aligned}$$

where $|I|\le m+n$ because the selected rows are linearly independent in $\mathbb{R}^{m+n}$. Using $\lambda\ge 0$ and $d = A_{x,I}^\top\lambda = v-v^\star$, we obtain

$$\begin{aligned}\|d\|_2^2 = \langle A_{x,I}^\top\lambda, v-v^\star\rangle &= \lambda^\top(A_{x,I}v - A_{x,I}v^\star) = \lambda^\top(A_{x,I}v - b_{x,I}) = \sum_{i\in I}\lambda_i(A_x v - b_x)_i\\ &\overset{\text{Claim 7}}{\le} \varepsilon\sum_{i\in I}\lambda_i = \varepsilon\|\lambda\|_1 \le \sqrt{m+n}\,\kappa_{\mathrm{reg}}\,\varepsilon\|d\|_2.\end{aligned}$$

Write $v^\star = (\nu^\star, a^\star)$. Converting to the product norm,

$$\begin{aligned}\operatorname{dist}\big((\nu,a),\mathcal{S}_h(x)\big) &\le \|\nu-\nu^\star\|_2 + \|a-a^\star\|_1 \le \|\nu-\nu^\star\|_2 + \sqrt{n}\,\|a-a^\star\|_2\\ &\le \sqrt{n+1}\,\|v-v^\star\|_2 \le \kappa_{\mathrm{reg}}\,\mathrm{Gap}_h(x,a,\nu)\end{aligned} \tag{37}$$

where, as stated in the lemma, we absorb the fixed factor $\sqrt{(n+1)(m+n)}$ into κ_{reg}. □

Applying it to the averaged output gives

$$\operatorname{dist}\big((\bar\nu^K,\bar a^K),\mathcal{S}_h(x)\big) \le \kappa_{\mathrm{reg}}\,\mathrm{Gap}_h(x,\bar a^K,\bar\nu^K) \le \frac{\kappa_{\mathrm{reg}}\Gamma_h}{\sqrt{K}}.$$

Since $\mathcal{S}_h(x)$ is nonempty and compact, the distance in the product norm is attained. Thus, there exists $(\nu^\star,a^\star)\in\mathcal{S}_h(x)$ so that the distance bound is

$$\|\bar\nu^K-\nu^\star\|_2 + \|\bar a^K - a^\star\|_1 \le \frac{\kappa_{\mathrm{reg}}\Gamma_h}{\sqrt{K}}. \tag{38}$$

Robinson strong regularity is sufficient locally but is stronger than the solution-set bound in (36) (Robinson, 1980; Dontchev & Rockafellar, 2014; Rockafellar & Wets, 1998). The uniform condition may fail when an active basis degenerates at a supported-value switch.

For example, let $\mathcal{Y} = \{0, 1\}$, $h(x, y) = xy$, and $c(x, y) = -1$. Then $V_h(x) = \min\{0, x\}$. For $x > 0$, the exact saddle pair has $\nu = 0$ and all response mass on $y = 0$. After eliminating the simplex equality, two active rows have coefficient matrix $\begin{pmatrix} 1 & x \\ -1 & 0 \end{pmatrix}$, whose smallest singular value tends to zero as $x \downarrow 0$. Thus Assumption 4 need not cover an ordinary response switch, even when the SFM best response is exact in one call.

A direct consistency subclass has $c(x, y) = \bar{c}(y)$ and $h(x, y) = \alpha_h(x) + \bar{h}(y)$. After eliminating the simplex equality, the common term $\alpha_h(x)$ cancels from every row of (35). The finitely many supported basis matrices are therefore independent of x, and the minimum of their positive singular values verifies Assumption 4. In this subclass, $G_h(x, \nu, a) = \nabla\alpha_h(x) = \nabla V_h(x)$ for every admissible pair, so it checks consistency rather than nontrivial recovery across response switches.

Note that the algorithm returns $\widehat{G}_h(x) = G_h(x, \bar{\nu}^K, \bar{a}^K)$. By the Lipschitz continuity of G_h in (ν, a) and the distance bound (38), there exists $(\nu^\star, a^\star) \in \mathcal{S}_h(x)$ such that

$$\|\widehat{G}_h(x) - G_h(x, \nu^\star, a^\star)\|_2 \le L_{G,h} \left(\|\bar{\nu}^K - \nu^\star\|_2 + \|\bar{a}^K - a^\star\|_1\right) \le \frac{L_{G,h}\kappa_{\mathrm{reg}}\Gamma_h}{\sqrt{K}} = \mathcal{O}(K^{-1/2}).$$

However, the quantity we need to bound is $\|\widehat{G}_h(x) - \nabla V_h(x)\|_2$. It therefore remains to show that V_h is differentiable at x and that $\nabla V_h(x) = G_h(x, \nu^\star, a^\star)$. Define the sensitivity mapping

$$G_h(x, \nu, a) := \nabla_x L_h(x, a, \nu) = \sum_{y\in\mathcal{Y}} a_y \left[\nabla_x h(x, y) + \nu^\top \nabla_x c(x, y)\right], \tag{SM}$$

where a and ν are held fixed when differentiating, and $\nabla_x c(x, y) \in \mathbb{R}^{m\times p}$ is the Jacobian of c.

Lemma 6 (Error-bound envelope sensitivity)**.** *Suppose Assumptions 1, 3, and 4 hold. For each supported objective $h \in \{f+\gamma g, g\}$ and every $x \in \mathrm{int}(\mathcal{X}^\varrho)$, the supported value V_h is differentiable at x and*

$$\nabla V_h(x) = G_h(x, \nu, a), \quad (\nu, a) \in \mathcal{S}_h(x). \tag{39}$$

Thus all exact saddle pairs induce the same sensitivity direction.

Proof. Since $a \in \Delta(\mathcal{Y})$,

$$\begin{aligned}
\|G_h(x, \nu, a)\|_2 &\overset{\text{(SM)}}{=} \left\| \sum_{y\in\mathcal{Y}} a_y \left[\nabla_x h(x, y) + \nu^\top \nabla_x c(x, y)\right] \right\|_2 \\
&\overset{\text{triangle inequality}}{\le} \sum_{y\in\mathcal{Y}} a_y \left\|\nabla_x h(x, y) + \nu^\top \nabla_x c(x, y)\right\|_2 \\
&\le B_h \sum_{y\in\mathcal{Y}} a_y = B_h,
\end{aligned}$$

Consequently,

$$\begin{aligned}
&|\operatorname{Gap}_h(x', a, \nu) - \operatorname{Gap}_h(x, a, \nu)| \\
&\overset{\text{(Gap)}}{\le} \left|\max_{q\in\mathcal{B}_h} L_h(x', a, q) - \max_{q\in\mathcal{B}_h} L_h(x, a, q)\right| + \left|\min_{b\in\Delta(\mathcal{Y})} L_h(x', b, \nu) - \min_{b\in\Delta(\mathcal{Y})} L_h(x, b, \nu)\right| \\
&\le \max_{q\in\mathcal{B}_h} |L_h(x', a, q) - L_h(x, a, q)| + \max_{b\in\Delta(\mathcal{Y})} |L_h(x', b, \nu) - L_h(x, b, \nu)| \\
&\le B_h\|x' - x\|_2 + B_h\|x' - x\|_2 == 2B_h\|x' - x\|_2.
\end{aligned}$$

the third inequality holds by the mean value theorem along the segment $[x, x']$.

Fix an interior query point x and any $(\nu^\star, a^\star) \in \mathcal{S}_h(x)$. Since $\operatorname{Gap}_h(x, a^\star, \nu^\star) = 0$,

$$\begin{aligned}
\operatorname{dist}\big((\nu^\star, a^\star), \mathcal{S}_h(x')\big) &\overset{(37)}{\le} \kappa_{\mathrm{reg}} \operatorname{Gap}_h(x', a^\star, \nu^\star) \\
&= \kappa_{\mathrm{reg}} \left[\operatorname{Gap}_h(x', a^\star, \nu^\star) - \operatorname{Gap}_h(x, a^\star, \nu^\star)\right] \\
&\le 2\kappa_{\mathrm{reg}} B_h \|x' - x\|_2.
\end{aligned}$$

where the last inequality holds by Lipschitzness. The distance is attained, so we can choose $(\nu', a') \in \mathcal{S}_h(x')$ such that

$$\|\nu' - \nu^\star\|_2 + \|a' - a^\star\|_1 \le 2\kappa_{\mathrm{reg}} B_h \|x' - x\|_2. \tag{40}$$

In particular, $x' \to x$ gives $(\nu', a') \to (\nu^\star, a^\star)$.

Claim 8. *For $(\nu^\star, a^\star) \in \mathcal{S}_h(x)$ and $(\nu', a') \in \mathcal{S}_h(x')$,*

$$L_h(x', a', \nu^\star) - L_h(x, a', \nu^\star) \le V_h(x') - V_h(x) \le L_h(x', a^\star, \nu') - L_h(x, a^\star, \nu') \tag{SQZ}$$

Proof of the claim. The saddle inequalities at x and x' give

$$L_h(x, a^\star, \nu') \le V_h(x) \le L_h(x, a', \nu^\star),$$
$$L_h(x', a', \nu^\star) \le V_h(x') \le L_h(x', a^\star, \nu').$$

Subtracting the appropriate bounds gives

$$L_h(x', a', \nu^\star) - L_h(x, a', \nu^\star) \le V_h(x') - V_h(x),$$
$$V_h(x') - V_h(x) \le L_h(x', a^\star, \nu') - L_h(x, a^\star, \nu').$$

△

Write $\delta := x' - x$. Then

$$\begin{aligned}
&|L_h(x', a', \nu^\star) - L_h(x, a', \nu^\star) - \langle G_h(x, \nu^\star, a^\star), \delta\rangle| \\
\overset{\mathrm{FTC}}{=}& \left| \int_0^1 \langle G_h(x + t\delta, \nu^\star, a') - G_h(x, \nu^\star, a^\star), \delta\rangle \, dt \right| \\
\overset{\mathrm{CS}}{\le}& \|\delta\|_2 \int_0^1 \|G_h(x + t\delta, \nu^\star, a') - G_h(x, \nu^\star, a^\star)\|_2 \, dt \\
\le& \|\delta\|_2 \sup_{t\in[0,1]} \|G_h(x + t\delta, \nu^\star, a') - G_h(x, \nu^\star, a^\star)\|_2 = o(\|\delta\|_2),
\end{aligned}$$

where the last equality is from continuity and (40). Here, the last step follows from the continuity of G_h and $a' \to a^\star$, which imply $\sup_{t\in[0,1]} \|G_h(x + t\delta, \nu^\star, a') - G_h(x, \nu^\star, a^\star)\|_2 \to 0$. Similarly, since $\nu' \to \nu^\star$,

$$\begin{aligned}
&|L_h(x', a^\star, \nu') - L_h(x, a^\star, \nu') - \langle G_h(x, \nu^\star, a^\star), \delta\rangle| \\
\le& \|\delta\|_2 \sup_{t\in[0,1]} \|G_h(x + t\delta, \nu', a^\star) - G_h(x, \nu^\star, a^\star)\|_2 \\
=& o(\|\delta\|_2).
\end{aligned}$$

where the last equality follows continuity and (40). Consequently,

$$\begin{aligned}
&|V_h(x + \delta) - V_h(x) - \langle G_h(x, \nu^\star, a^\star), \delta\rangle| \\
\overset{\text{(SQZ)}}{\le}& \max \left\{ \begin{array}{l} \left|L_h(x', a', \nu^\star) - L_h(x, a', \nu^\star) - \langle G_h(x, \nu^\star, a^\star), \delta\rangle\right|, \\ \left|L_h(x', a^\star, \nu') - L_h(x, a^\star, \nu') - \langle G_h(x, \nu^\star, a^\star), \delta\rangle\right| \end{array} \right\} = o(\|\delta\|_2).
\end{aligned}$$

Thus, by the definition of differentiability, there is $\nabla V_h(x) = G_h(x, \nu^\star, a^\star)$. Since $(\nu^\star, a^\star) \in \mathcal{S}_h(x)$ was arbitrary, every exact saddle point at x gives the same sensitivity. This is the saddle-point envelope argument under a solution-set error bound (Bonnans & Shapiro, 2000). □

E.3 Proof of Proposition 5

We now combine Proposition 10, Lemma 5, and Lemma 6. Since $\bar{a}^K = K^{-1} \sum_{k=0}^{K-1} \delta_{y^k}$, for any function ϕ on $\mathcal{Y}$, $\sum_{y\in\mathcal{Y}} \bar{a}_y^K \phi(y) = \frac{1}{K} \sum_{k=0}^{K-1} \phi(y^k)$. Thus, the running sums in Algorithm 2 return

$$\begin{aligned}
\widehat{G}_h(x) &= \frac{1}{K} \sum_{k=0}^{K-1} \nabla_x h(x, y^k) + (\bar{\nu}^K)^\top \left(\frac{1}{K} \sum_{k=0}^{K-1} \nabla_x c(x, y^k) \right) \\
&= \sum_{y\in\mathcal{Y}} \bar{a}_y^K \left[\nabla_x h(x, y) + (\bar{\nu}^K)^\top \nabla_x c(x, y) \right] \\
&\overset{\text{(SM)}}{=} G_h(x, \bar{\nu}^K, \bar{a}^K).
\end{aligned} \tag{41}$$

The multiplier term is the averaged Jacobian multiplied by the averaged multiplier; in general, it differs from the average of the products $(\nu^k)^\top \nabla_x c(x, y^k)$.

Let L_c from (6) satisfy $\|\nabla_x c_i(x, y)\|_2 \le L_c$ for every $i = 1, \dots, m$. For any ν, ν',

$$\begin{aligned}\|(\nu - \nu')^\top \nabla_x c(x, y)\|_2 &= \left\| \sum_{i=1}^m (\nu_i - \nu_i') \nabla_x c_i(x, y) \right\|_2 \\ &\le L_c \sum_{i=1}^m |\nu_i - \nu_i'| \\ &\overset{\text{CS}}{\le} \sqrt{m}\, L_c \|\nu - \nu'\|_2.\end{aligned}$$

Choose $(\nu^\star, a^\star) \in \mathcal{S}_h(x)$ satisfying (38). Adding and subtracting $G_h(x, \bar\nu^K, a^\star)$ gives

$$\begin{aligned}&\|G_h(x, \bar\nu^K, \bar a^K) - G_h(x, \nu^\star, a^\star)\|_2 \\ &\le \|G_h(x, \bar\nu^K, \bar a^K) - G_h(x, \bar\nu^K, a^\star)\|_2 + \|G_h(x, \bar\nu^K, a^\star) - G_h(x, \nu^\star, a^\star)\|_2 \\ &= \left\| \sum_{y \in \mathcal{Y}} (\bar a_y^K - a_y^\star) \left[\nabla_x h(x, y) + (\bar\nu^K)^\top \nabla_x c(x, y) \right] \right\|_2 \\ &\quad + \left\| \sum_{y \in \mathcal{Y}} a_y^\star (\bar\nu^K - \nu^\star)^\top \nabla_x c(x, y) \right\|_2 \\ &\le B_h \sum_{y \in \mathcal{Y}} |\bar a_y^K - a_y^\star| + \sqrt{m}\, L_c \|\bar\nu^K - \nu^\star\|_2 \sum_{y \in \mathcal{Y}} a_y^\star \\ &= B_h \|\bar a^K - a^\star\|_1 + \sqrt{m}\, L_c \|\bar\nu^K - \nu^\star\|_2. \end{aligned} \tag{42}$$

Combining (41)–(42), Lemma 6, and (38), we obtain

$$\begin{aligned}\|\widehat{G}_h(x) - \nabla V_h(x)\|_2 &\overset{\text{Lemma 6}}{=} \|G_h(x, \bar\nu^K, \bar a^K) - G_h(x, \nu^\star, a^\star)\|_2 \\ &\le B_h \|\bar a^K - a^\star\|_1 + \sqrt{m}\, L_c \|\bar\nu^K - \nu^\star\|_2 \\ &\le (B_h + \sqrt{m}\, L_c) \left(\|\bar a^K - a^\star\|_1 + \|\bar\nu^K - \nu^\star\|_2 \right) \\ &\overset{(38)}{\le} \frac{C_h}{\sqrt{K}}, \quad C_h := \kappa_{\text{reg}} \Gamma_h (B_h + \sqrt{m}\, L_c). \end{aligned} \tag{43}$$

This bound holds at every interior query point, uniformly over the admissible incoming multiplier and SFM states. Run the inner routine separately for $h = f + \gamma g$ and $h = g$, using K_λ and K_μ SFM calls, respectively, and define $\widehat{G}(x) := \widehat{G}_{f+\gamma g}(x) - \gamma \widehat{G}_g(x)$. Since $p_\gamma = V_{f+\gamma g} - \gamma V_g$,

$$\begin{aligned}\|\widehat{G}(x) - \nabla p_\gamma(x)\|_2 &= \left\| \widehat{G}_{f+\gamma g}(x) - \nabla V_{f+\gamma g}(x) - \gamma \big(\widehat{G}_g(x) - \nabla V_g(x) \big) \right\|_2 \\ &\le \|\widehat{G}_{f+\gamma g}(x) - \nabla V_{f+\gamma g}(x)\|_2 + \gamma \|\widehat{G}_g(x) - \nabla V_g(x)\|_2 \\ &\overset{(43)}{\le} \frac{C_{f+\gamma g}}{\sqrt{K_\lambda}} + \frac{\gamma C_g}{\sqrt{K_\mu}}. \end{aligned} \tag{E}$$

For equal budgets $K_\lambda = K_\mu = K$, setting $C := C_{f+\gamma g} + \gamma C_g$ gives

$$\|\widehat{G}(x) - \nabla p_\gamma(x)\|_2 \le \frac{C}{\sqrt{K}}.$$

The total number of SFM calls is $K_\lambda + K_\mu$, or $2K$ for equal budgets.

Proof of Proposition 5. Let $\mathcal{F}_t$ denote the information available before sampling w, including x and the incoming multiplier and SFM states. Let

$$w \sim \text{Unif}(\mathbb{B}_2^p), \quad w \perp\!\!\!\perp \mathcal{F}_t, \quad z := x + \varrho w.$$

By (3),

$$p_\gamma^\varrho(x) := \mathbb{E}_w[p_\gamma(x+\varrho w)] = \frac{1}{\mathrm{vol}(\mathbb{B}_2^p)} \int_{\mathbb{B}_2^p} p_\gamma(x+\varrho u)\, du.$$

At interior query points,

$$\begin{aligned}\|\nabla p_\gamma(z)\|_2 &= \|\nabla V_{f+\gamma g}(z) - \gamma \nabla V_g(z)\|_2 \\ &\le \|\nabla V_{f+\gamma g}(z)\|_2 + \gamma \|\nabla V_g(z)\|_2 \\ &\overset{\text{Lemma 6}}{\le} B_{f+\gamma g} + \gamma B_g.\end{aligned}$$

Using the gradient identity for Lipschitz ball smoothing (Lin et al., 2022),

$$\begin{aligned}\nabla p_\gamma^\varrho(x) &= \nabla_x \left[\frac{1}{\mathrm{vol}(\mathbb{B}_2^p)} \int_{\mathbb{B}_2^p} p_\gamma(x+\varrho u)\, du \right] \\ &\overset{\text{smoothing}}{=} \frac{1}{\mathrm{vol}(\mathbb{B}_2^p)} \int_{\mathbb{B}_2^p} \nabla p_\gamma(x+\varrho u)\, du \\ &= \mathbb{E}_w[\nabla p_\gamma(x+\varrho w)] \\ &\overset{w \perp\!\!\!\perp \mathcal{F}_t}{=} \mathbb{E}[\nabla p_\gamma(x+\varrho w) \mid \mathcal{F}_t]. \end{aligned} \tag{44}$$

Since $\mathcal{X}^\varrho = \mathcal{X} + \varrho \mathbb{B}_2^p$, $x \in \mathcal{X}$, $\|w\|_2 < 1$ gives $x + \varrho w \in x + \varrho\, \mathrm{int}(\mathbb{B}_2^p) \subseteq \mathrm{int}(\mathcal{X}^\varrho)$. Thus, $\mathbb{P}(x+\varrho w \in \mathrm{int}(\mathcal{X}^\varrho) \,|\, \mathcal{F}_t) = 1$. Applying (E) at $z = x + \varrho w$ yields

$$\|\widehat{G}(x+\varrho w) - \nabla p_\gamma(x+\varrho w)\|_2 \le \frac{C_{f+\gamma g}}{\sqrt{K_\lambda}} + \frac{\gamma C_g}{\sqrt{K_\mu}} \quad \text{almost surely.}$$

Consequently,

$$\begin{aligned}&\left\| \mathbb{E}[\widehat{G}(x+\varrho w) \mid \mathcal{F}_t] - \nabla p_\gamma^\varrho(x) \right\|_2 \\ \overset{(44)}{=} &\left\| \mathbb{E}[\widehat{G}(x+\varrho w) \mid \mathcal{F}_t] - \mathbb{E}[\nabla p_\gamma(x+\varrho w) \mid \mathcal{F}_t] \right\|_2 \\ \overset{\text{linearity}}{=} &\left\| \mathbb{E}\left[\widehat{G}(x+\varrho w) - \nabla p_\gamma(x+\varrho w) \,\middle|\, \mathcal{F}_t \right] \right\|_2 \\ \overset{\text{conditional Jensen}}{\le} &\mathbb{E}\left[\|\widehat{G}(x+\varrho w) - \nabla p_\gamma(x+\varrho w)\|_2 \,\middle|\, \mathcal{F}_t \right] \\ \overset{\text{(E)}}{\le} &\mathbb{E}\left[\frac{C_{f+\gamma g}}{\sqrt{K_\lambda}} + \frac{\gamma C_g}{\sqrt{K_\mu}} \,\middle|\, \mathcal{F}_t \right] \\ = &\frac{C_{f+\gamma g}}{\sqrt{K_\lambda}} + \frac{\gamma C_g}{\sqrt{K_\mu}}\end{aligned}$$

where the budgets are fixed before sampling w and the constants are uniform over the incoming states. For $K_\lambda = K_\mu = K$,

$$\left\| \mathbb{E}[\widehat{G}(x+\varrho w) \mid \mathcal{F}_t] - \nabla p_\gamma^\varrho(x) \right\|_2 \le \frac{C_{f+\gamma g} + \gamma C_g}{\sqrt{K}} = \frac{C}{\sqrt{K}}, \quad C := C_{f+\gamma g} + \gamma C_g.$$

This is (14) with $C_F := C_{f+\gamma g}$ and the factor γ absorbed into C_g. □

E.4 Proofs of Theorem 2 and Corollary 3

Proof of Theorem 2. For $t = 1, \dots, T$, let $\mathcal{F}_t$ contain the outer history and both persistent solver states before the first query of iteration t. Let $\mathcal{F}_{t,j}$ additionally contain the first $j-1$ queries and their updated states. Define

$$m^{t,j} := \mathbb{E}[\widehat{G}^{t,j} \mid \mathcal{F}_{t,j}], \quad \zeta^{t,j} := \widehat{G}^{t,j} - m^{t,j}.$$

Proposition 5 at $x = x^t$, conditional on $\mathcal{F}_{t,j}$, gives

$$\|m^{t,j} - \nabla p_\gamma^\varrho(x^t)\|_2 \le C_F K_\lambda^{-1/2} + C_g K_\mu^{-1/2}. \tag{BIAS}$$

The averaged multiplier-response pairs lie in the trajectory boxes, so the definition of B_h gives $\|\widehat{G}^{t,j}\|_2 \le B_{f+\gamma g} + \gamma B_g$, and hence $\mathbb{E}[\|\zeta^{t,j}\|_2^2 \mid \mathcal{F}_{t,j}] \le (B_{f+\gamma g} + \gamma B_g)^2$. The variables $\zeta^{t,j}$ form a martingale-difference sequence within the batch. Orthogonality of its increments yields

$$\mathbb{E}\left[\|\frac{1}{M}\sum_{j=1}^M \zeta^{t,j}\|_2^2 \middle| \mathcal{F}_t\right] = \frac{1}{M^2}\sum_{j=1}^M \mathbb{E}\big[\|\zeta^{t,j}\|_2^2 \mid \mathcal{F}_t\big] \le \frac{(B_{f+\gamma g} + \gamma B_g)^2}{M}. \tag{VAR}$$

Now we control the mean-squared error of the direction. Write $\bar{G}^t - \nabla p_\gamma^\varrho(x^t) = \frac{1}{M}\sum_j \zeta^{t,j} + \frac{1}{M}\sum_j \big(m^{t,j} - \nabla p_\gamma^\varrho(x^t)\big)$. Then, by Young's inequality,

$$\begin{aligned}\mathbb{E}\left[\|\bar{G}^t - \nabla p_\gamma^\varrho(x^t)\|_2^2 \middle| \mathcal{F}_t\right] &\le 2\,\mathbb{E}\left[\|\frac{1}{M}\sum_{j=1}^M \zeta^{t,j}\|_2^2 \middle| \mathcal{F}_t\right] + 2\max_j \|m^{t,j} - \nabla p_\gamma^\varrho(x^t)\|_2^2 \\ &\overset{\text{(VAR),(BIAS)}}{\le} \frac{2(B_{f+\gamma g} + \gamma B_g)^2}{M} + 2\left(C_F K_\lambda^{-1/2} + C_g K_\mu^{-1/2}\right)^2.\end{aligned}$$

Use the positive smoothness upper bound H chosen in Theorem 2. The comparison used in Theorem 1 gives $p_\gamma^\varrho(x^1) - \inf_{x\in\mathcal{X}} p_\gamma^\varrho(x) \le \Delta_{\gamma,1} + \varrho L_\gamma$. Applying the projected-descent calculation from that proof with the full error $\bar{G}^t - \nabla p_\gamma^\varrho(x^t)$, the bound of Step 3, and $\eta = 1/(2H)$ gives

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\mathcal{G}_{1/(2H),\varrho}(x^t)\|_2^2 \le \frac{8H(\Delta_{\gamma,1} + \varrho L_\gamma)}{T} + \frac{12(B_{f+\gamma g} + \gamma B_g)^2}{M} \\ + 12\left(C_F K_\lambda^{-1/2} + C_g K_\mu^{-1/2}\right)^2.\end{aligned} \tag{RATE}$$

The choices in Theorem 2 make each term at most $\varepsilon^2/3$. Each outer iteration uses $M(K_\lambda + K_\mu)$ SFM calls and final recovery uses K_λ calls. Thus the exact count is $TM(K_\lambda + K_\mu) + K_\lambda$. For fixed problem parameters, γ, and ϱ, substitution gives $\mathcal{O}(\varepsilon^{-6})$. □

Proof of Corollary 3. Write η for the algorithmic stepsize in the corollary.

Define locally

$$\mathcal{E}(x) := \min_{u\in\mathcal{X}} \left\{p_\gamma^\varrho(u) + H\|u - x\|_2^2\right\}.$$

Since p_γ^ϱ has an H-Lipschitz gradient, the minimized function is strongly convex in u. Let $u(x)$ be its unique minimizer. Standard envelope calculus and the proximal optimality condition give $\nabla\mathcal{E}(x) = 2H(x - u(x))$. The optimality condition also implies

$$\left\langle \nabla p_\gamma^\varrho(x), x - u(x)\right\rangle \ge H\|x - u(x)\|_2^2 = \frac{1}{4H}\|\nabla\mathcal{E}(x)\|_2^2. \tag{ALIGN}$$

For $M = 1$, write $\bar{G}^t$ for the sampled direction and let $x^{t+1} = \Pi_\mathcal{X}(x^t - \eta\bar{G}^t)$. Projection nonexpansiveness, evaluated against $u(x^t)$, gives

$$\mathcal{E}(x^{t+1}) \le \mathcal{E}(x^t) - \eta\left\langle \bar{G}^t, \nabla\mathcal{E}(x^t)\right\rangle + H\eta^2\|\bar{G}^t\|_2^2.$$

Conditioning on $\mathcal{F}_t$, using (ALIGN), Young's inequality, and $\|\bar{G}^t\|_2 \le B_{f+\gamma g} + \gamma B_g$ yields

$$\begin{aligned}\mathbb{E}[\mathcal{E}(x^{t+1}) \mid \mathcal{F}_t] \le \mathcal{E}(x^t) - \frac{\eta}{4}\|\nabla\mathcal{E}(x^t)\|_2^2 + \eta\|\mathbb{E}[\bar{G}^t \mid \mathcal{F}_t] - \nabla p_\gamma^\varrho(x^t)\|_2^2 \\ + H\eta^2(B_{f+\gamma g} + \gamma B_g)^2.\end{aligned} \tag{DESC}$$

Let $q(x) := \Pi_\mathcal{X}(x - (2H)^{-1}\nabla p_\gamma^\varrho(x))$. Proximal optimality gives $u(x) = \Pi_\mathcal{X}(x - (2H)^{-1}\nabla p_\gamma^\varrho(u(x)))$. Hence nonexpansiveness and H-smoothness imply

$$\|q(x) - u(x)\|_2 \le \frac{1}{2H}\|\nabla p_\gamma^\varrho(x) - \nabla p_\gamma^\varrho(u(x))\|_2 \le \tfrac{1}{2}\|x - u(x)\|_2,$$

and therefore

$$\|\mathcal{G}_{1/(2H),\varrho}(x)\|_2 = 2H\|x-q(x)\|_2 \le 2H\big(\|x-u(x)\|_2+\|u(x)-q(x)\|_2\big) \le \frac{3}{2}\|\nabla\mathcal{E}(x)\|_2. \quad \text{(GM)}$$

Summing (DESC) over $t = 1, \dots, T$, using $\mathcal{E}(x^1) - \inf \mathcal{E} \le \Delta_{\gamma,1} + \varrho L_\gamma$, and applying (GM) give

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}_{1/(2H),\varrho}(x^t)\|_2^2 \le \frac{9(\Delta_{\gamma,1}+\varrho L_\gamma)}{\eta T} + 9H\eta(B_{f+\gamma g}+\gamma B_g)^2 + \frac{9}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathbb{E}[\bar{G}^t \mid \mathcal{F}_t] - \nabla p_\gamma^\varrho(x^t)\|_2^2. \quad \text{(RATE}')$$

Proposition 5 bounds the final average in (RATE′) by $(C_F K_\lambda^{-1/2} + C_g K_\mu^{-1/2})^2$. The parameter choices in Corollary 3 make the three resulting terms at most $\varepsilon^2/3$. The exact call count is $T(K_\lambda + K_\mu) + K_\lambda$, which is $\mathcal{O}(\varepsilon^{-6})$ for fixed problem parameters, γ, and ϱ. Taking $K_\lambda = K_\mu = K$ and η proportional to $T^{-1/2}$ gives the finite-budget rate in Remark 1. □

F Constrained EDA details and ablations

We use the constrained EDA formulation in Section 5.2. The K-SpecPart two-way reference for the ISPD98 `ibm01` hypergraph has block sizes $6{,}302/6{,}450$ and hypergraph cut 203. The hypergraph and reference partition come from the TILOS institute repository. Clique expansion with distance decay $\exp(-\|r_i - r_j\|_2^2/0.1^2)$ yields 109,082 weighted edges after removing weights at most 10^{-6}. The LL constraint uses a separate clique expansion of nets with degree at least five and imposes $c(y) = \sum_{(i,j)\in E_{\ge 5}} \mathbf{1}\{y_i \ne y_j\} - 1300 \le 0$. By the symmetric-cut calculation in (22), the pairwise terms in both g and c are submodular. The Hamming loss $f(y) = \sum_i |y_i - y_i^{\mathrm{ref}}|$ is modular, so Assumption 2 holds with $\beta = 0$, and nonnegative Lagrangian multipliers preserve the SFM structure. The softplus unary terms are smooth in x, and $q_i^1(x) - q_i^0(x) = -90 s_i(x)$, so the label-dependent unary term is affine in x and the common label-independent term cancels from the oracle gradient difference. These observations verify Assumptions 1 and 2. Exact support and uniform supported-basis conditioning are not certified instance-wide for this experiment. The reference has constraint value 1991 and is therefore infeasible for the threshold 1300.

For BiGAP-PD we set $\gamma = 0.01$, $K_\lambda = K_\mu = 3$, $\eta = 5 \times 10^{-5}$, $\varrho = 0.01$, and $\theta_\lambda = \theta_\mu = 5 \times 10^{-4}$. Each outer step uses one randomized query. The two supported chains retain independent graph, multiplier, and response states. Seeds $0, \dots, 7$ generate eight shared initial points uniformly from $0.8\mathcal{X}$, with one run per initialization and a 300-second optimization budget. The optimization clock includes graph construction, oracle solves, updates, and evaluations used for incumbent selection. A deferred audit with $K_{\mathrm{eval}} = 10$ reconstructs the responses and evaluates the certified error defined in Section 5.2. Audit time is excluded from the horizontal axis, while evaluations intrinsic to each baseline remain inside its budget.

The reported Hamming error is $f(y_g)$ at the checkpoint minimizing the certified score. The LL term $[g(x, y_g) - \widehat{v}(x)]_+$ measures the remaining certified LL gap and $[c(y_g)]_+$ measures constraint violation. For the retained $K = 3$ configuration, these terms increase the mean error from 442.6 to 445.0.

CMA-ES uses initial standard deviation 0.3 with box projection, random search samples uniformly from $\mathcal{X}$, and the zeroth-order baseline uses a symmetric two-point estimator with radius 0.01. Weighted-PD applies projected descent to a single weighted supported objective. The nested baseline uses a real-coded genetic algorithm with population 12 and a time-limited lexicographic MILP follower. The ZJS restriction evaluates the eight corners of the three-dimensional upper box, while DIVIDE-BLO uses 32 fixed feasible lower samples generated with 20 inner steps. These general-purpose baselines use the same initializations and time budget, but not identical oracle models.

Figure 5 and Table 3 isolate the penalty and inner-budget effects. The penalty sweep uses $K = 3$ and evaluates both returned response types. The budget sweep fixes $\gamma = 0.01$ and varies $K \in$

ϱ	Mean ± standard deviation	Median
0	449.7 ± 8.9	447.1
0.001	453.0 ± 9.9	450.3
0.01	$\mathbf{445.2 \pm 6.8}$	**442.7**
0.1	453.0 ± 5.0	451.5
0.15	484.8 ± 42.7	461.6
0.2	528.2 ± 27.5	537.8
0.3	588.1 ± 23.0	586.4
0.5	727.7 ± 86.6	739.5

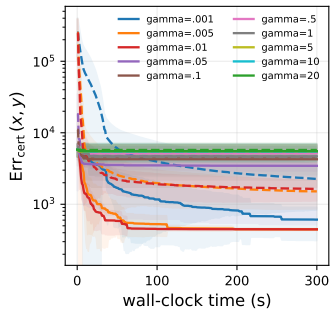


Figure 5: Penalty ablation at $K = 3$. Solid curves report y_g and dashed curves report y_F.

γ sweep, $K = 3$			K sweep, $\gamma = 0.01$	
γ	y_g	y_F	K	y_g
0.001	609.9 ± 300.1	2260.3 ± 1430.7	1	445.5 ± 2.3
0.005	$\mathbf{445.9 \pm 16.2}$	$\mathbf{1515.1 \pm 655.4}$	3	$\mathbf{445.0 \pm 8.6}$
0.01	446.4 ± 8.6	1632.1 ± 516.8	5	472.1 ± 74.5
0.05	3447.6 ± 2928.9	4644.6 ± 1734.4	10	523.2 ± 122.0
0.1	4262.8 ± 2536.2	4575.1 ± 2388.2	20	592.0 ± 184.7
1	4672.1 ± 2086.5	4600.6 ± 2152.9	30	661.2 ± 215.9

Table 3: Endpoint $\mathrm{Err}_{\mathrm{cert}}(x, y)$ after 300 seconds, reported as mean $\pm$ standard deviation over eight initializations.

$\{1, 3, 5, 10, 20, 30\}$. Small penalties preserve useful separation between the two supported responses, while large penalties make their certified errors similar and substantially worse. Increasing K improves the inner solve but reduces the number of outer updates available within the fixed time budget, which explains the best endpoint at $K = 3$.

Table 4 shows that the selected radius $\varrho = 0.01$ improves over $\varrho = 0$ in both mean certified error, 445.2 versus 449.7, and median certified error, 442.7 versus 447.1. The unsmoothed variant nevertheless remains competitive because smoothing is a safeguard against exceptional branch ties rather than a requirement for descent at every point. When $\varrho = 0$, the sampled direction is multiplied by zero and introduces no within-run query randomness. The observed variation then comes from the eight initializations, and the resulting trajectories generally avoid an absorbing tie. Larger radii increasingly average distinct discrete-response regions and degrade the finite-time solution, which is consistent with the smoothing bias controlled by ϱ.

G Toy instance: assumptions, reference solution, and the systolic illustration

G.1 Assumptions and oracle calls

The strip has 4 rows and 7 columns. Vertices in column 0 have label 0 and vertices in column 6 have label 1. The 20 free vertices are the binary variables. The horizontal edges in column gaps $1|2$, $3|4$, $5|6$ form corridors 1, 2, 3 with lengths $1 + \frac{x_1}{2}$, $1 - \frac{x_1}{2}$, $1 + \frac{x_2}{2}$. Every other edge has length 10. Let $U := (-0.8, 0.8)^2$. Then $\mathcal{X} \subset U$, $\varrho_0 = 0.2$, and every edge length is at least 0.6 on U.

Assumption 1. $\mathcal{X}$ is a box. $\mathcal{Y} = \{0, 1\}^{20}$ is finite. For fixed y, $g(\cdot, y)$ is affine and $f(\cdot, y)$ is quadratic, so both are C^1 on U.

Assumption 2. $g(x, \cdot)$ is a sum of terms $w_{ij}(x)\mathbf{1}\{y_i \neq y_j\}$ with $w_{ij}(x) > 0$ on U. Each term is submodular by (22). The term $\frac{1}{8}\|y - y_{\mathrm{tar}}\|_1$ is modular, so $\Delta_{ij} f \equiv 0$. The zero-slack condition of Proposition 2 holds, $\beta^\star(x) = 0$, and Assumption 2 holds with $\beta = 0$. Here $c \equiv 0$.

The two oracle calls. Algorithm 1 needs $y_2 \in \arg\min_y g(z, y)$ and $y_1 \in \arg\min_y \{\gamma^{-1} f(z, y) + g(z, y)\}$. The first is a minimum s–t cut. The second adds the unary term $\frac{1}{8\gamma}\sum_i |y_i - y_{\mathrm{tar},i}|$. It enters as terminal edges of capacity $\frac{1}{8\gamma}$: from the source to i if $y_{\mathrm{tar},i} = 0$, from i to the sink if $y_{\mathrm{tar},i} = 1$. The quadratic term of f does not depend on y. So each call is one max-flow on 28 vertices plus source and sink. The direction (7) is $u(z, y_1, y_2) = (z - x_0) + \gamma(\nabla_x g(z, y_1) - \nabla_x g(z, y_2))$ with

$\nabla_x g(z,y) = \sum_{y_i \neq y_j} a_{ij}$, where a_{ij} is the slope of w_{ij}. The code uses the Boykov–Kolmogorov solver (Boykov & Kolmogorov, 2004). No closed form enters the iteration.

G.2 Exact reference solution

Every cut weight is affine in x: $g(x,y) = b_y + a_y^\top x$. The target is a minimum cut at x iff $b_{\mathrm{tar}} + a_{\mathrm{tar}}^\top x \le b_y + a_y^\top x$ for all y. We enumerate all 2^{20} labelings, keep the tightest right-hand side for each direction $a_y - a_{\mathrm{tar}}$, and drop the inequalities that are never active on $\mathcal{X}$. Two remain: $4x_1 \le 0$ and $2x_1 - 2x_2 \le 0$, that is, $L_1 \le L_2$ and $L_1 \le L_3$. So $F := \{x \in \mathcal{X} : y_{\mathrm{tar}} \in \mathcal{Y}^\star(x)\} = \{x \in \mathcal{X} : x_1 \le 0,\ x_1 \le x_2\}$. On F, $\phi(x) = \frac{1}{2}\|x - x_0\|_2^2$. Off F, every minimum cut differs from y_{tar} on 8 or 16 vertices, so $\phi(x) \ge \frac{1}{2}\|x - x_0\|_2^2 + 1$. The projection $x^\star$ of x_0 onto F has $\frac{1}{2}\|x^\star - x_0\|_2^2 \le 0.4 < 1$, so it solves (1). For $x_{0,1} > x_{0,2}$ and $x_{0,2} \le -0.2$ it is $x^\star = (m, m)$ with $m = \min\{\frac{x_{0,1}+x_{0,2}}{2}, -0.2\}$. The code projects onto the enumerated constraints, not onto this formula; the two agree.

At $x_0 = (0.3, -0.5)$: $L_1 = 4.6$, $L_2 = 3.4$, $L_3 = 3.0$, so corridor 3 is the minimum cut. At $x^\star = (-0.2, -0.2)$: $L_1 = L_3 = 3.6 < L_2 = 4.4$. With $\gamma = 0.35$, $M = 1$, $\eta_k = (k+10)^{-1}$, $T = 3000$, the errors $\|x^k - x^\star\|_2$ at $k = 10^2, 10^3, 3 \times 10^3$ are 9.2×10^{-4}, 2.0×10^{-4}, 6.6×10^{-5} for $\varrho = 0$; 5.9×10^{-3}, 3.2×10^{-3}, 5.0×10^{-3} for $\varrho = 0.01$; and 3.3×10^{-2}, 1.8×10^{-2}, 1.9×10^{-2} for $\varrho = 0.05$. The floor grows linearly in ϱ, as the smoothing bias predicts. We also start from the 11 points of a 5×3 grid on $\mathcal{X}$ at which the target is not the minimum cut. With $\varrho = 0.01$ and $\gamma = 0.41$, all 11 runs end within 1.1×10^{-2} of their $x^\star$ after 1500 steps. With $\gamma = 0.35$, the three starts with $x_{0,1} = 0.55$ stop about 0.05 short; there 0.35 is below the penalty threshold of Corollary 1.

G.3 The inverse systolic example

An earlier version of Section 5.1 used a metric on a torus. We keep it as an illustration of the penalty geometry. It is outside the oracle model of Section 2.2: its lower level is a shortest-cycle problem, not an SFM instance.

Let K be the 4×4 periodic triangulation of T^2 with one diagonal of direction $(1,1)$ per square. Horizontal, vertical, and diagonal edges have lengths $w_H = 1 + \frac{x_1}{2}$, $w_V = 1 - \frac{x_1}{2}$, $w_D = 1 + \frac{x_2}{2}$, with $x \in \mathcal{X} = [-0.6, 0.6] \times [-0.6, -0.2]$. The lower level minimizes $g(x,y) = \sum_e w_e(x)|y_e|$ over nontrivial integer 1-cycles y. The upper level is $f(x,y) = \frac{1}{2}\|x - x_0\|_2^2 + \mathbf{1}\{[y] \neq (1,0)\}$.

Proposition 11 (Three-class reduction)**.** *For every $x \in \mathcal{X}$, the systole of $(K, w(x))$ lies in one of the classes $\pm(1,0)$, $\pm(0,1)$, $\pm(1,1)$. The shortest cycles in these classes have lengths $L_{10} = 4w_H$, $L_{01} = 4w_V$, $L_{11} = 4w_D$.*

Proof. On $\mathcal{X}$, $w_H, w_V \in [0.7, 1.3]$ and $w_D \in [0.7, 0.9]$. So every edge has length at least 0.7, and $L_{11} = 4w_D \le 3.6$. A cycle in class (a,b) lifts to a path in the universal cover with displacement $(4a, 4b)$. Each edge changes each coordinate by at most 1. So the cycle has at least $4\max\{|a|,|b|\}$ edges and length at least $2.8\max\{|a|,|b|\}$. For $\max\{|a|,|b|\} \ge 2$ this is at least $5.6 > 3.6$. For the class $(1,-1)$, a diagonal edge changes both coordinates by the same amount, so the difference 8 of the displacements must come from horizontal and vertical edges, each contributing at most 1; this needs at least 8 edges and length at least 5.6. So only $\pm(1,0)$, $\pm(0,1)$, $\pm(1,1)$ remain.

In class $(1,0)$, replacing one horizontal edge by a diagonal and a vertical edge keeps the displacement and costs $w_D + w_V \ge 1.4 > 1.3 \ge w_H$. So the straight cycle of four horizontal edges is shortest, $L_{10} = 4w_H$. The same argument gives $L_{01} = 4w_V$. In class $(1,1)$, each missing diagonal must be replaced by a horizontal and a vertical edge at cost $w_H + w_V = 2 > w_D$. So $L_{11} = 4w_D$. □

By Proposition 11, $v(x) = \min\{L_{10}, L_{01}, L_{11}\}$, and the target class is systolic iff $x_1 \le 0$ and $x_1 \le x_2$, the same region as in Section 5.1. The penalty is $p_\gamma(x) = \frac{1}{2}\|x - x_0\|_2^2 + \min\{1, \gamma[L_{10} - \min\{L_{01}, L_{11}\}]_+\}$. Projected descent with the closed-form three-class oracle reaches the same $x^\star$. The corridor instance of Section 5.1 realizes the same three lengths inside the SFM oracle model.

H EXPLORATORY FULLY DISCRETE EXTENSIONS

This appendix preserves the fully discrete analysis as a research direction, not as part of the proved BiGAP framework. The entropy lifting introduces a $\gamma\tau$ bias, while the moment lifting can have an integrality gap. No exact and tractable algorithm for a fully discrete upper variable is claimed.

For clarity, take $c \equiv 0$ and suppose both $\mathcal{X}$ and $\mathcal{Y}$ are finite. The optimistic problem is

$$\min_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}^\star(x)} f(x,y), \quad \mathcal{Y}^\star(x) = \arg\min_{y\in\mathcal{Y}} g(x,y). \tag{45}$$

Neither variable now admits a derivative. Two natural liftings replace finite decisions by distributions and then, when possible, compress those distributions to sufficient statistics.

H.1 MEASURE LIFTING AND THE TIE-BREAKING KNOB

Let $\Delta(\mathcal{Y}) = \{p \in \mathbb{R}_+^{\mathcal{Y}} : \sum_y p_y = 1\}$. Extend the two costs affinely by $F(x,p) = \sum_y p_y f(x,y)$ and $G(x,p) = \sum_y p_y g(x,y)$. For a continuous strongly convex Ω on $\Delta(\mathcal{Y})$ and $\tau \geq 0$, set $G_\tau(x,p) = G(x,p) + \tau\Omega(p)$ and consider

$$\min_{x\in\mathcal{X}} F(x,p) \quad \text{s.t.} \quad p \in \arg\min_{p'\in\Delta(\mathcal{Y})} G_\tau(x,p'). \tag{46}$$

Proposition 12 (Exactness versus unique tie-breaking)**.** *For every fixed x:*

1. *at $\tau = 0$, $\arg\min_{p\in\Delta(\mathcal{Y})} G(x,p) = \Delta(\mathcal{Y}^\star(x))$, so* (46) *has the same optimistic value and the same optimal upper decisions as* (45)*.*
2. *at $\tau > 0$, the lower minimizer $p_\tau^\star(x)$ is unique, but as $\tau \downarrow 0$ it converges to the unique minimizer of Ω over $\Delta(\mathcal{Y}^\star(x))$. For negative entropy $\Omega(p) = \sum_y p_y \log p_y$, this limit is uniform on $\mathcal{Y}^\star(x)$, and hence*

$$F(x, p_\tau^\star(x)) \longrightarrow \frac{1}{|\mathcal{Y}^\star(x)|} \sum_{y\in\mathcal{Y}^\star(x)} f(x,y),$$

which is generally an average rather than the optimistic minimum.

The proof of Proposition 12 follows immediately below.

Proof. At $\tau = 0$, a linear function over a simplex is minimized precisely on the convex hull of its minimum-cost vertices. For $\tau > 0$, strong convexity gives uniqueness. Compactness shows that every limit point as $\tau \downarrow 0$ minimizes G. Comparing the regularized objective against points in that optimal face then shows that it minimizes Ω there. Strong convexity makes this point unique. Negative entropy is minimized by the uniform distribution on a fixed face. □

Thus τ cannot simultaneously preserve optimistic tie-breaking and remove multiplicity: decreasing τ restores the original optimal face, while increasing it selects a unique point without consulting f.

H.2 ENTROPY-SMOOTHED PENALTY

For $\gamma > 0$, consider

$$\min_{x\in\mathcal{X},\, p\in\Delta(\mathcal{Y})} F(x,p) + \gamma\left(G_\tau(x,p) - \min_{p'\in\Delta(\mathcal{Y})} G_\tau(x,p')\right). \tag{47}$$

For a function $h : \mathcal{Y} \to \mathbb{R}$, write $\mathrm{sm}_\tau(h) := -\tau \log \sum_{y\in\mathcal{Y}} e^{-h(y)/\tau}$.

Proposition 13 (Closed form and smoothing bias)**.** *Let $\tau > 0$ and use negative entropy. The inner distribution in* (47) *is proportional to $\exp\big(-(f+\gamma g)/(\gamma\tau)\big)$, and the reduced value is*

$$p_{\gamma,\tau}(x) = \gamma\left(\mathrm{sm}_\tau\left(g + \frac{f}{\gamma}\right)(x) - \mathrm{sm}_\tau(g)(x)\right). \tag{48}$$

Moreover,

$$0 \leq \min_y h(y) - \mathrm{sm}_\tau(h) \leq \tau \log|\mathcal{Y}|, \quad |p_{\gamma,\tau}(x) - p_\gamma(x)| \leq \gamma\tau \log|\mathcal{Y}|.$$

The proof of Proposition 13 follows immediately below.

Proof. The entropy variational identity

$$\min_{p\in\Delta(\mathcal{Y})}\left\{\sum_y p_y h(y)+\tau\sum_y p_y\log p_y\right\}=-\tau\log\sum_y e^{-h(y)/\tau}$$

gives both the Gibbs distribution and (48). Bounding the log-sum-exp between its largest term and $|\mathcal{Y}|$ times that term gives the soft-min inequality. The two soft-min errors both lie in the same interval $[0,\tau\log|\mathcal{Y}|]$, so their difference has absolute value at most $\tau\log|\mathcal{Y}|$. Multiplying by γ proves the final bound. □

The sign in the soft-min inequality is important: $\mathrm{sm}_\tau(h)\le\min h$. The bias is controlled by $\gamma\tau$, not τ alone. Therefore a regime with $\gamma\to\infty$ requires at least $\gamma\tau\to 0$ to recover the hard penalty from this bound.

H.3 Moment lifting

Suppose there are feature maps $\psi:\mathcal{X}\to\mathbb{R}^n$ and $\varphi:\mathcal{Y}\to\mathbb{R}^q$ such that $f(x,y)=\widetilde{f}(\psi(x),\varphi(y))$ and $g(x,y)=\widetilde{g}(\psi(x),\varphi(y))$, where $\widetilde{f}$ and $\widetilde{g}$ are bilinear. Under a product distribution, the expectations depend only on $w=\mathbb{E}(\psi(x))$ and $z=\mathbb{E}(\varphi(y))$, which range over $\mathcal{W}=\operatorname{conv}\psi(\mathcal{X})$ and $\mathcal{Z}=\operatorname{conv}\varphi(\mathcal{Y})$. This suggests the continuous relaxation

$$\min_{w\in\mathcal{W}}\widetilde{f}(w,z)\quad\text{s.t.}\quad z\in\arg\min_{z'\in\mathcal{Z}}\widetilde{g}(w,z'). \tag{49}$$

Proposition 14 (Exact lower faces, potentially lossy upper relaxation)**.** *At every vertex $w=\psi(x)$, the lower solution of* (49) *is* $\operatorname{conv}\varphi(\mathcal{Y}^\star(x))$*, and the optimistic value agrees with* (45) *at x. The upper relaxation can nevertheless be strict. For $\mathcal{X}=\mathcal{Y}=\{0,1\}$ and identity maps $\psi=\varphi$, let $\widetilde{g}(w,z)=(1-2w)z$ and $\widetilde{f}(w,z)=-(1-w)z$. Both vertices of $\mathcal{W}=[0,1]$ have optimistic value zero, whereas $w=1/2$ makes the lower problem indifferent and attains value $-1/2$.*

The proof of Proposition 14 follows immediately below.

Proof. At a vertex $w=\psi(x)$, bilinearity makes lower minimization over $\mathcal{Z}$ the convexification of minimization over $\varphi(\mathcal{Y})$, proving the first statement. In the displayed example, the lower minimizer is $z=0$ at $w=0$ and $z=1$ at $w=1$, giving upper value zero at both vertices. At $w=1/2$, every $z\in[0,1]$ is lower optimal, and optimistic selection chooses $z=1$, giving $-1/2$. □

The gap again appears at a tie: the discontinuity of the original optimistic value becomes an integrality gap of the moment relaxation. Entropy regularization makes the lower response unique but does not prove that the relaxed upper solution is a vertex. A rounding or exactness analysis remains open.